14

Riccati Equations and their Solution

Vladimír Kučera
Czech Technical University and Institute of Information Theory and Automation

14.1 Introduction

An ordinary differential equation of the form

$$\dot{x}(t) + f(t)x(t) - b(t)x^2(t) + c(t) = 0 \tag{14.1}$$

is known as a *Riccati equation,* deriving its name from Jacopo Francesco, Count Riccati (1676–1754) [1], who studied a particular case of this equation from 1719 to 1724.

For several reasons, a differential equation of the form of Equation 14.1, and generalizations thereof comprise a highly significant class of nonlinear ordinary differential equations. First, they are intimately related to ordinary linear homogeneous differential equations of the second order. Second, the solutions of Equation 14.1 possess a very particular structure in that the general solution is a fractional linear function in the constant of integration. In applications, Riccati differential equations appear in the classical problems of the calculus of variations and in the associated disciplines of optimal control and filtering.

The *matrix* Riccati differential equation refers to the equation

$$\dot{X}(t) + X(t)A(t) - D(t)X(t) - X(t)B(t)X(t) + C(t) = 0 \tag{14.2}$$

defined on the vector space of real $m \times n$ matrices. Here, A, B, C, and D are real matrix functions of the appropriate dimensions. Of particular interest are the matrix Riccati equations that arise in optimal control and filtering problems and that enjoy certain symmetry properties. This chapter is concerned

with these symmetric matrix Riccati differential equations and concentrates on the following four major topics:

- Basic properties of the solutions
- Existence and properties of constant solutions
- Asymptotic behavior of the solutions
- Methods for the numerical solution of the Riccati equations

14.2 Optimal Control and Filtering: Motivation

The following problems of optimal control and filtering are of great engineering importance and serve to motivate our study of the Riccati equations.

A linear-quadratic *optimal control* problem consists of the following. Given a linear system

$$\dot{x}(t) = Fx(t) + Gu(t), \quad x(t_0) = c, \quad y(t) = Hx(t), \tag{14.3}$$

where x is the n-vector state, u is the q-vector control input, y is the p-vector of regulated variables, and F, G, H are constant real matrices of the appropriate dimensions. One seeks to determine a control input function u over some fixed time interval $[t_1, t_2]$ such that a given quadratic cost functional of the form

$$\eta(t_1, t_2, T) = \int_{t_1}^{t_2} [y'(t)y(t) + u'(t)u(t)]\,\mathrm{d}t + x'(t_2)Tx(t_2), \tag{14.4}$$

with T being a constant real symmetric ($T = T'$) and nonnegative definite ($T \geq 0$) matrix, is afforded a minimum in the class of all solutions of Equation 14.3, for any initial state c.

A unique optimal control exists for all finite $t_2 - t_1 > 0$ and has the form

$$u(t) = -G'P(t, t_2, T)x(t),$$

where $P(t, t_2, T)$ is the solution of the matrix Riccati differential equation

$$-\dot{P}(t) = P(t)F + F'P(t) - P(t)GG'P(t) + H'H \tag{14.5}$$

subject to the terminal condition

$$P(t_2) = T.$$

The optimal control is a linear state feedback, which gives rise to the closed-loop system

$$\dot{x}(t) = [F - GG'P(t, t_2, T)]x(t)$$

and yields the minimum cost

$$\eta^*(t_1, t_2, T) = c'P(t_1, t_2, T)c. \tag{14.6}$$

A Gaussian *optimal filtering* problem consists of the following. Given the p-vector random process z modeled by the equations

$$\begin{aligned} \dot{x}(t) &= Fx(t) + Gv(t), \\ z(t) &= Hx(t) + w(t), \end{aligned} \tag{14.7}$$

where x is the n-vector state and v, w are independent Gaussian white random processes (respectively, q-vector and p-vector) with zero means and identity covariance matrices. The matrices F, G, and H are constant real ones of the appropriate dimensions.

Given known values of z over some fixed time interval $[t_1, t_2]$ and assuming that $x(t_1)$ is a Gaussian random vector, independent of v and w, with zero mean and covariance matrix S, one seeks to determine an estimate $\hat{x}(t_2)$ of $x(t_2)$ such that the variance

$$\sigma(S, t_1, t_2) = Ef'[x(t_2) - \hat{x}(t_2)][x(t_2) - \hat{x}(t_2)]'f \tag{14.8}$$

of the error encountered in estimating any real-valued linear function f of $x(t_2)$ is minimized.

A unique optimal estimate exists for all finite $t_2 - t_1 > 0$ and is generated by a linear system of the form

$$\dot{\hat{x}}(t) = F\hat{x}(t) + Q(S, t_1, t)H'e(t), \quad \hat{x}(t_0) = 0, \quad e(t) = z(t) - H\hat{x}(t),$$

where $Q(S, t_1, t)$ is the solution of the matrix Riccati differential equation

$$\dot{Q}(t) = Q(t)F' + FQ(t) - Q(t)H'HQ(t) + GG' \tag{14.9}$$

subject to the initial condition

$$Q(t_1) = S.$$

The minimum error variance is given by

$$\sigma^*(S, t_1, t_2) = f'Q(S, t_1, t_2)f. \tag{14.10}$$

Equations 14.5 and 14.9 are special cases of the matrix Riccati differential Equation 14.2 in that A, B, C, and D are constant real $n \times n$ matrices such that

$$B = B', \quad C = C', \quad D = -A'.$$

Therefore, symmetric solutions $X(t)$ are obtained in the optimal control and filtering problems.

We observe that the control Equation 14.5 is solved *backward* in time, while the filtering Equation 14.9 is solved *forward* in time. We also observe that the two equations are *dual* to each other in the sense that

$$P(t, t_2, T) = Q(S, t_1, t)$$

on replacing F, G, H, T, and $t_2 - t$ in Equation 14.5 respectively, by F', H', G', S, and $t - t_1$ or, vice versa, on replacing F, G, H, S, and $t - t_1$ in Equation 14.9 respectively, by F', H', G', T, and $t_2 - t$. This makes it possible to dispense with both cases by considering only one prototype equation.

14.3 Riccati Differential Equation

This section is concerned with the basic properties of the prototype matrix Riccati differential equation

$$\dot{X}(t) + X(t)A + A'X(t) - X(t)BX(t) + C = 0, \tag{14.11}$$

where A, B, and C are constant real $n \times n$ matrices with B and C being symmetric and nonnegative definite,

$$B = B', \ B \geq 0 \quad \text{and} \quad C = C', \ C \geq 0. \tag{14.12}$$

By definition, a *solution* of Equation 14.11 is a real $n \times n$ matrix function $X(t)$ that is absolutely continuous and satisfies Equation 14.11 for t on an interval on the real line R.

Generally, solutions of Riccati differential equations exist only locally. There is a phenomenon called finite escape time: the equation

$$\dot{x}(t) = x^2(t) + 1$$

has a solution $x(t) = \tan t$ in the interval $(-\frac{\pi}{2}, 0)$ that cannot be extended to include the point $t = -\frac{\pi}{2}$. However, Equation 14.11 with the sign-definite coefficients as shown in Equation14.12 does have global solutions.

Let $X(t, t_2, T)$ denote the solution of Equation 14.11 that passes through a constant real $n \times n$ matrix T at time t_2. We shall assume that

$$T = T' \quad \text{and} \quad T \geq 0. \tag{14.13}$$

Then the solution exists on every finite subinterval of R, is symmetric, nonnegative definite and enjoys certain monotone properties.

Theorem 14.1:

Under the assumptions of Equations 14.12 and 14.13 Equation 14.11 has a unique solution $X(t, t_2, T)$ satisfying

$$X(t, t_2, T) = X'(t, t_2, T), \quad X(t, t_2, T) \geq 0$$

for every T and every finite t, t_2, such that $t \geq t_2$.

This can most easily be seen by associating Equation 14.11 with the optimal control problem described in Equations 14.3 through 14.6. Indeed, using Equation 14.12, one can write $B = GG'$ and $C = H'H$ for some real matrices G and H. The quadratic cost functional η of Equation 14.4 exists and is nonnegative for every T satisfying Equation 14.13 and for every finite $t_2 - t$. Using Equation 14.6, the quadratic form $c'X(t, t_2, T)c$ can be interpreted as a particular value of η for every real vector c.

A further consequence of Equations 14.4 and 14.6 follows.

Theorem 14.2:

For every finite t_1, t_2 and τ_1, τ_2 such that $t_1 \leq \tau_1 \leq \tau_2 \leq t_2$,

$$X(t_1, \tau_1, 0) \leq X(t_1, \tau_2, 0)$$
$$X(\tau_2, t_2, 0) \leq X(\tau_1, t_2, 0)$$

and for every $T_1 \leq T_2$,

$$X(t_1, t_2, T_1) \leq X(t_1, t_2, T_2).$$

Thus, the solution of Equation 14.11 passing through $T = 0$ does not decrease as the length of the interval increases, and the solution passing through a larger T dominates that passing through a smaller T.

The Riccati Equation 14.11 is related in a very particular manner with linear Hamiltonian systems of differential equations.

Theorem 14.3:

Let

$$\Phi(t, t_2) = \begin{bmatrix} \Phi_{11} & \Phi_{12} \\ \Phi_{21} & \Phi_{22} \end{bmatrix}$$

be the fundamental matrix solution of the linear Hamiltonian matrix differential system

$$\begin{bmatrix} \dot{U}(t) \\ \dot{V}(t) \end{bmatrix} = \begin{bmatrix} A & -B \\ -C & -A' \end{bmatrix} \begin{bmatrix} U(t) \\ V(t) \end{bmatrix}$$

that satisfies the transversality condition

$$V(t_2) = TU(t_2).$$

If the matrix $\Phi_{11} + \Phi_{12}T$ is nonsingular on an interval $[t, t_2]$, then

$$X(t, t_2, T) = (\Phi_{21} + \Phi_{22}T)(\Phi_{11} + \Phi_{12}T)^{-1} \tag{14.14}$$

is a solution of the Riccati Equation 14.11.

Thus, if $V(t_2) = TU(t_2)$, then $V(t) = X(t, t_2, T)U(t)$ and the formula of Equation 14.14 follows.
Let us illustrate this with a simple example. The Riccati equation

$$\dot{x}(t) = x^2(t) - 1, \quad x(0) = T$$

satisfies the hypotheses of Equations 14.12 and 14.13. The associated linear Hamiltonian system of equations reads

$$\begin{bmatrix} \dot{u}(t) \\ \dot{v}(t) \end{bmatrix} = \begin{bmatrix} 0 & -1 \\ -1 & 0 \end{bmatrix} \begin{bmatrix} u(t) \\ v(t) \end{bmatrix}$$

and has the solution

$$\begin{bmatrix} u(t) \\ v(t) \end{bmatrix} = \begin{bmatrix} \cosh t & -\sinh t \\ -\sinh t & \cosh t \end{bmatrix} \begin{bmatrix} u(0) \\ v(0) \end{bmatrix},$$

where $v(0) = Tu(0)$. Then the Riccati equation has the solution

$$x(t, 0, T) = \frac{-\sinh t + T \cosh t}{\cosh t - T \sinh t}$$

for all $t \leq 0$. The monotone properties of the solution are best seen in Figure 14.1.

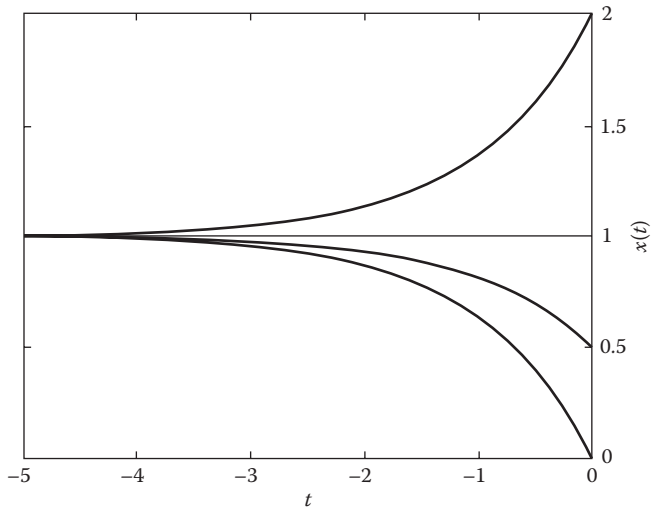


FIGURE 14.1 Graph of solutions.

14.4 Riccati Algebraic Equation

The constant solutions of Equation 14.11 are just the solutions of the quadratic equation

$$XA + A'X - XBX + C = 0, \tag{14.15}$$

called the *algebraic Riccati equation*. This equation can have real $n \times n$ matrix solutions X that are symmetric or nonsymmetric, sign definite or indefinite, and the set of solutions can be either finite or infinite. These solutions will be studied under the standing assumption of Equation 14.12, namely

$$B = B', \;\; B \geq 0 \quad \text{and} \quad C = C', \;\; C \geq 0.$$

14.4.1 General Solutions

The solution set of Equation 14.15 corresponds to a certain class of n-dimensional invariant subspaces of the associated $2n \times 2n$ matrix

$$H = \begin{bmatrix} A & -B \\ -C & -A' \end{bmatrix}. \tag{14.16}$$

This matrix has the *Hamiltonian* property

$$\begin{bmatrix} 0 & I \\ -I & 0 \end{bmatrix} H = -H' \begin{bmatrix} 0 & I \\ -I & 0 \end{bmatrix}.$$

It follows that H is similar to-H' and therefore, the spectrum of H is symmetrical with respect to the imaginary axis.

Now suppose that X is a solution of Equation 14.15. Then

$$H \begin{bmatrix} I \\ X \end{bmatrix} = \begin{bmatrix} I \\ X \end{bmatrix} (A - BX).$$

Denote $J = U^{-1}(A - BX)U$, the Jordan form of $A - BX$ and put $V = XU$. Then

$$H \begin{bmatrix} U \\ V \end{bmatrix} = \begin{bmatrix} U \\ V \end{bmatrix} J,$$

which shows that the columns of

$$\begin{bmatrix} U \\ V \end{bmatrix}$$

are Jordan chains for H, that is, sets of vectors $x_1, x_2, \ldots, x_r$ such that $x_1 \neq 0$ and for some eigenvalue λ of H

$$\begin{aligned} Hx_1 &= \lambda x_1 \\ Hx_j &= \lambda x_j + x_{j+1}, \quad j = 2, 3, \ldots, r. \end{aligned}$$

In particular, x_1 is an eigenvector of H. Thus, we have the following result.

Theorem 14.4:

Equation 14.15 has a solution X if and only if there is a set of vectors $x_1, x_2, \ldots, x_n$ forming a set of Jordan chains for H and if

$$x_i = \begin{bmatrix} u_i \\ v_i \end{bmatrix},$$

where u_i is an n-vector, then $u_1, u_2, \ldots, u_n$ are linearly independent.

Furthermore, if

$$U = [u_1 \dots u_n], \quad V = [v_1 \dots v_n],$$

every solution of Equation 14.15 has the form $X = VU^{-1}$ *for some set of Jordan chains* $x_1, x_2, \dots, x_n$ *for* H.

To illustrate, consider the scalar equation

$$X^2 + pX + q = 0,$$

where p, q are real numbers and $q \leq 0$. The Hamiltonian matrix

$$H = \begin{bmatrix} -\frac{p}{2} & -1 \\ q & \frac{p}{2} \end{bmatrix}$$

has eigenvalues λ and $-\lambda$, where

$$\lambda^2 = \left(\frac{p}{2}\right)^2 - q.$$

If $\lambda \neq 0$ there are two eigenvectors of H, namely

$$x_1 = \begin{bmatrix} 1 \\ -\frac{p}{2} + \lambda \end{bmatrix}, \quad x_2 = \begin{bmatrix} 1 \\ -\frac{p}{2} - \lambda \end{bmatrix},$$

which correspond to the solutions

$$X_1 = -\frac{p}{2} + \lambda, \quad X_2 = -\frac{p}{2} - \lambda.$$

If $\lambda = 0$ there exists one Jordan chain,

$$x_1 = \begin{bmatrix} 1 \\ -\frac{p}{2} \end{bmatrix}, \quad x_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix},$$

which yields the unique solution

$$X_1 = -\frac{p}{2}.$$

Theorem 14.4 suggests that, generically, the number of solutions of Equation 14.15 to be expected will not exceed the binomial coefficient $\binom{2n}{n}$, the number of ways in which the vectors $x_1, x_2, \dots, x_n$ can be chosen from a basis of $2n$ eigenvectors for H. The solution set is infinite if there is a continuous family of Jordan chains. To illustrate this point consider Equation 14.15 with

$$A = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad C = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}.$$

The Hamiltonian matrix

$$H = \begin{bmatrix} 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

has the eigenvalue 0, associated with two Jordan chains

$$x_1 = \begin{bmatrix} a \\ b \\ 0 \\ 0 \end{bmatrix}, \quad x_2 = \begin{bmatrix} c \\ d \\ -a \\ -b \end{bmatrix}, \quad \text{and} \quad x_3 = \begin{bmatrix} c \\ d \\ 0 \\ 0 \end{bmatrix}, \quad x_4 = \begin{bmatrix} a \\ b \\ -c \\ -d \end{bmatrix},$$

where a, b and c, d are real numbers such that $ad - bc = 1$. The solution set of Equation 14.15 consists of the matrix

$$X_{13} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$$

and two continuous families of matrices

$$X_{12}(a,b) = \begin{bmatrix} ab & -a^2 \\ b^2 & -ab \end{bmatrix} \quad \text{and} \quad X_{34}(c,d) = \begin{bmatrix} -cd & c^2 \\ -d^2 & cd \end{bmatrix}.$$

Having in mind the applications in optimal control and filtering, we shall be concerned with the solutions of Equation 14.15 that are symmetric and nonnegative definite.

14.4.2 Symmetric Solutions

In view of Theorem 14.4, each solution X of Equation 14.15 gives rise to a factorization of the characteristic polynomial χ_H of H as

$$\chi_H(s) = (-1)^n q(s) q_1(s),$$

where $q = \chi_{A-BX}$. If the solution is *symmetric,* $X = X'$, then $q_1(s) = q(-s)$. This follows from

$$\begin{bmatrix} I & 0 \\ X & I \end{bmatrix}^{-1} \begin{bmatrix} A & -B \\ -C & -A' \end{bmatrix} \begin{bmatrix} I & 0 \\ X & I \end{bmatrix} = \begin{bmatrix} A - BX & -B \\ 0 & -(A - BX)' \end{bmatrix}.$$

There are two symmetric solutions that are of particular importance. They correspond to a factorization

$$\chi_H(s) = (-1)^n q(s) q(-s)$$

in which q has all its roots with nonpositive real part; it follows that $q(-s)$ has all its roots with nonnegative real part. We shall designate these solutions X_+ and X_-.

One of the basic results concerns the existence of these particular solutions. To state the result, we recall some terminology. A pair of real $n \times n$ matrices (A, B) is said to be *controllable* (*stabilizable*) if the $n \times 2n$ matrix $[\lambda I - A \quad B]$ has linearly independent rows for every complex λ (respectively, for every complex λ such that $Re\ \lambda \geq 0$). The numbers λ for which $[\lambda I - A \quad B]$, loses rank are the eigenvalues of A that are not controllable (stabilizable) from B. A pair of real $n \times n$ matrices (A, C) is said to be *observable* (*detectable*) if the $2n \times n$ matrix $\begin{bmatrix} \lambda I - A \\ C \end{bmatrix}$ has linearly independent columns for every complex λ (respectively, for every complex λ such that $Re\ \lambda \geq 0$). The numbers λ for which $\begin{bmatrix} \lambda I - A \\ C \end{bmatrix}$ loses rank are the eigenvalues of A that are not observable (detectable) in C. Finally, let dim V denote the dimension of a linear space V and Im M, Ker M the image and the kernel of a matrix M, respectively.

Theorem 14.5:

There exists a unique symmetric solution X_+ of Equation 14.15 such that all eigenvalues of $A - BX_+$ have nonpositive real part if and only if (A, B) is stabilizable.

Theorem 14.6:

There exists a unique symmetric solution X_- of Equation 14.15 such that all eigenvalues of $A - BX_-$ have nonnegative real part if and only if $(-A, B)$ is stabilizable.

We observe that both (A, B) and $(-A, B)$ are stabilizable if and only if (A, B) is controllable. It follows that both solutions X_+ and X_- exist if and only if (A, B) is controllable.

For two real symmetric matrices X_1 and X_2, the notation $X_1 \geq X_2$ means that $X_1 - X_2$ is nonnegative definite. Since $A - BX_+$ has no eigenvalues with positive real part, neither has $X_+ - X$. Hence, $X_+ - X \geq 0$. Similarly, one can show that $X - X_- \geq 0$, thus introducing a partial order among the set of symmetric solutions of Equation 14.15.

Theorem 14.7:

Suppose that X_+ and X_- exist. If X is any symmetric solution of Equation 14.15, then

$$X_+ \geq X \geq X_-.$$

That is why X_+ and X_- are called the *extreme* solutions of Equation 14.15; X_+ is the maximal symmetric solution, while X_- is the minimal symmetric solution. The set of all symmetric solutions of Equation 14.15 can be related to a certain subset of the set of invariant subspaces of the matrix $A - BX_+$ or the matrix $A - BX_-$. Denote V_0 and V_+ the invariant subspaces of $A - BX_+$ that correspond, respectively, to the pure imaginary eigenvalues and to the eigenvalues having negative real part. Denote W_0 and W_- the invariant subspaces of $A - BX_-$ that correspond, respectively, to the pure imaginary eigenvalues and to the eigenvalues having positive real part. Then it can be shown that $V_0 = W_0$ is the kernel of $X_+ - X_-$ and the symmetric solution set corresponds to the set of all invariant subspaces of $A - BX_+$ contained in V_+ or, equivalently, to the set of all invariant subspaces of $A - BX_-$ contained in W_-.

Theorem 14.8:

Suppose that X_+ and X_- exist. Let X_1, X_2 be symmetric solutions of Equation 14.15 corresponding to the invariant subspaces $\mathcal{V}_1$, $\mathcal{V}_2$ of $\mathcal{V}_+$ (or $\mathcal{W}_1$, $\mathcal{W}_2$ of $\mathcal{W}_-$). Then $X_1 \geq X_2$ if and only if $\mathcal{V}_1 \supset \mathcal{V}_2$ (or if and only if $\mathcal{W}_1 \subset \mathcal{W}_2$).

This means that the symmetric solution set of Equation 14.15 is a complete *lattice* with respect to the usual ordering of symmetric matrices. The maximal solution X_+ corresponds to the invariant subspace $\mathcal{V}_+$ of $A - BX_+$ or to the invariant subspace $\mathcal{W} = 0$ of $A - BX_-$, whereas the minimal solution X_- corresponds to the invariant subspace $\mathcal{V} = 0$ of $A - BX_+$ or to the invariant subspace $\mathcal{W}_-$ of $A - BX_-$.

This result allows one to count the distinct symmetric solutions of Equation 14.15 in some cases. Thus, let α be the number of distinct eigenvalues of $A - BX_+$ having negative real part and let $m_1, m_2, \ldots, m_\alpha$ be the multiplicities of these eigenvalues. Owing to the symmetries in H, the matrix $A - BX_-$ exhibits the same structure of eigenvalues with positive real part.

Theorem 14.9:

Suppose that X_+ and X_- exist. Then the symmetric solution set of Equation 14.15 has finite cardinality if and only if $A - BX_+$ is cyclic on $\mathcal{V}_+$ (or if and only if $A - BX_-$ is cyclic on $\mathcal{W}_-$). In this case, the set contains exactly $(m_1 + 1) \ldots (m_\alpha + 1)$ solutions.

Simple examples are most illustrative. Consider Equation 14.15 with

$$A = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}, \quad B = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad C = \begin{bmatrix} 1 & 0 \\ 0 & 3 \end{bmatrix}$$

and determine the lattice of symmetric solutions. We have

$$\chi_H(s) = s^4 - 5s^2 + 4$$

and the following eigenvectors of H:

$$x_1 = \begin{bmatrix} 1 \\ 0 \\ -1 \\ 0 \end{bmatrix}, \quad x_2 = \begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \end{bmatrix}, \quad x_3 = \begin{bmatrix} 0 \\ 1 \\ 0 \\ -1 \end{bmatrix}, \quad x_4 = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 3 \end{bmatrix}$$

are associated with the eigenvalues 1, -1, 2, and -2, respectively. Hence, the pair of solutions

$$X_+ = \begin{bmatrix} 1 & 0 \\ 0 & 3 \end{bmatrix}, \quad X_- = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}$$

corresponds to the factorization

$$\chi_H(s) = (s^2 - 3s + 2)(s^2 + 3s + 2)$$

and the solutions

$$X_{2,3} = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}, \quad X_{1,4} = \begin{bmatrix} -1 & 0 \\ 0 & 3 \end{bmatrix}$$

correspond to the factorization

$$\chi_H(s) = (s^2 - s - 2)(s^2 + s - 2).$$

There are four subspaces invariant under the matrices

$$A - BX_+ = \begin{bmatrix} -1 & 0 \\ 0 & -2 \end{bmatrix}, \quad A - BX_- = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}$$

each corresponding to one of the four solutions above. The partial ordering

$$X_+ \geq X_{2,3} \geq X_-, \quad X_+ \geq X_{1,4} \geq X_-$$

defines the lattice visualized in Figure 14.2.

As another example, we consider Equation 14.15 where

$$A = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad C = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

and classify the symmetric solution set.

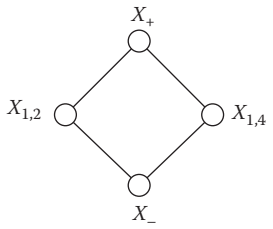


FIGURE 14.2 Lattice of solutions.

We have

$$\chi_H(s) = (s-1)^2(s+1)^2$$

and a choice of eigenvectors corresponding to the eigenvalues $1, -1$ of H is

$$x_1 = \begin{bmatrix} 1 \\ 0 \\ -1 \\ 0 \end{bmatrix}, \quad x_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \\ -1 \end{bmatrix}, \quad x_3 = \begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \end{bmatrix}, \quad x_4 = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 1 \end{bmatrix}.$$

Hence,

$$X_+ = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad X_- = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}$$

are the extreme solutions.

We calculate

$$A - BX_+ = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}, \quad A - BX_- = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

and observe that the set of subspaces invariant under $A - BX_+$ or $A - BX_-$ (other than the zero and the whole space, which correspond to X_+ and X_-) is the family of one-dimensional subspaces parameterized by their azimuth angle θ. These correspond to the solutions

$$X_\vartheta = \begin{bmatrix} \cos\vartheta & \sin\vartheta \\ \sin\vartheta & -\cos\vartheta \end{bmatrix}.$$

Therefore, the solution set consists of X_+, X_- and the continuous family of solutions X_θ. It is a complete lattice and $X_+ \geq X_\theta \geq X_-$ for every θ.

14.4.3 Definite Solutions

Under the standing assumption (Equation 14.12), namely

$$B = B', \ \ B \geq 0 \quad \text{and} \quad C = C', \ \ C \geq 0,$$

one can prove that $X_+ \geq 0$ and $X_- \leq 0$. The existence of X_+, however, excludes the existence of X_- and vice versa, unless (A, B) is controllable.

If X_+ does exist, any other solution $X \geq 0$ of Equation 14.15 corresponds to a subspace W of W_- that is invariant under $A - BX$. From Equation 14.15,

$$X(A - BX) + (A - BX)'X = -XBX - C.$$

The restriction of $A - BX$ to W has eigenvalues with positive real part. Since $-XBX - C \leq 0$, it follows from the Lyapunov theory that X restricted to W is nonpositive definite and hence zero. We conclude

that the solutions $X \geq 0$ of Equation 14.15 correspond to those subspaces $\mathcal{W}$ of $\mathcal{W}_-$ that are invariant under A and contained in Ker C.

The set of symmetric nonnegative definite solutions of Equation 14.15 is a sublattice of the lattice of all symmetric solutions. Clearly X_+ is the largest solution and it corresponds to the invariant subspace $\mathcal{W} = 0$ of A. The smallest nonnegative definite solution will be denoted by X_* and it corresponds to $\mathcal{W}_*$, the largest invariant subspace of A contained in Ker $C$and associated with eigenvalues having positive real part.

The nonnegative definite solution set of Equation 14.15 has finite cardinality if and only if A is cyclic on W_*. In this case, the set contains exactly $(p_1 + 1) \ldots (_\rho+1)$ solutions, where ρ is the number of distinct eigenvalues of A associated with $\mathcal{W}_*$ and $p_1, p_2, \ldots, p_\rho$ are the multiplicities of these eigenvalues.

Analogous results hold for the set of symmetric solutions of Equation 14.15 that are nonpositive definite. In particular, if X_- exists, then any other solution $X \leq 0$ of Equation 14.15 corresponds to a subspace V of $\mathcal{V}_+$ that is invariant under A and contained in Ker C. Clearly X_- is the smallest solution and it corresponds to the invariant subspace $\mathcal{V} = 0$ of A. The largest nonpositive definite solution is denoted by $X_\times$ and it corresponds to $\mathcal{W}_\times$, the largest invariant subspace of A contained in Ker C and associated with eigenvalues having negative real part.

Let us illustrate this with a simple example. Consider Equation 14.15 where

$$A = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}, \quad B = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}, \quad C = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$$

and classify the two sign-definite solution sets. We have

$$X_+ = \begin{bmatrix} 8 & 4 \\ 4 & 4 \end{bmatrix}, \quad X_- = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}.$$

The matrix A has one eigenvalue with positive real part, namely 1, and a basis for $\mathcal{W}_*$ is

$$x_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad x_2 = \begin{bmatrix} 0 \\ -1 \end{bmatrix}.$$

Thus, there are three invariant subspaces of $\mathcal{W}_*$ corresponding to the three nonnegative definite solutions of Equation 14.15

$$X_+ = \begin{bmatrix} 8 & 4 \\ 4 & 4 \end{bmatrix}, \quad X_1 = \begin{bmatrix} 0 & 0 \\ 0 & 2 \end{bmatrix}, \quad X_* = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}.$$

These solutions make a lattice and

$$X_+ \geq X_1 \geq X_*.$$

The matrix A has no eigenvalues with negative real part. Therefore, $\mathcal{V}_* = 0$ and X_- is the only nonpositive definite solution of Equation 14.15.

Another example for Equation 14.15 is provided by

$$A = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \quad C = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}.$$

It is seen that neither (A, B) nor $(-A, B)$ is stabilizable; hence, neither X_+ nor X_- exists. The symmetric solution set consists of one continuous family of solutions

$$X_\alpha = \begin{bmatrix} 0 & 0 \\ 0 & \alpha \end{bmatrix}$$

for any real α. Therefore, both sign-definite solution sets are infinite; the nonnegative solution set is unbounded from above while the nonpositive solution set is unbounded from below.

14.5 Limiting Behavior of Solutions

The length of the time interval $t_2 - t_1$ in the optimal control and filtering problems is rather artificial. For this reason, an infinite time interval is often considered. This brings in the question of the limiting behavior of the solution $X(t, t_2, T)$ for the Riccati differential Equation 14.11.

In applications to optimal control, it is customary to fix t and let t_2 approach $+\infty$. Since the coefficient matrices of Equation 14.11 are constant, the same result is obtained if t_2 is held fixed and t approaches $-\infty$. The limiting behavior of $X(t, t_2, T)$ strongly depends on the terminal matrix $T \geq 0$. For a suitable choice of T, the solution of Equation 14.11 may converge to a constant matrix $X \geq 0$, a solution of Equation 14.15. For some matrices T, however, the solution of Equation 14.11 may fail to converge to a constant matrix, but it may converge to a periodic matrix function.

Theorem 14.10:

Let (A, B) be stabilizable. If t and T are held fixed and $t_2 \to \infty$, then the solution $X(t, t_2, T)$ of Equation 14.11 is bounded on the interval $[t, \infty)$.

This result can be proved by associating an optimal control problem with Equation 14.11. Then stabilizability of (A, B) implies the existence of a stabilizing (not necessarily optimal) control. The consequent cost functional of Equation 14.4 is finite and dominates the optimal one.

If (A, B) is stabilizable, then X_+ exists and each real symmetric nonnegative definite solution X of Equation 14.15 corresponds to a subset $\mathcal{W}$ of $\mathcal{W}_*$, the set of A-invariant subspaces contained in Ker C and associated with eigenvalues having positive real part. The convergence of the solution $X(t, t_2, T)$ of Equation 14.11 to X depends on the properties of the image of $\mathcal{W}_*$ under T.

For simplicity, it is assumed that the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_\rho$ of A associated with $\mathcal{W}_*$ are simple and, except for pairs of complex conjugate eigenvalues, have different real parts. Let the corresponding eigenvectors be ordered according to decreasing real parts of the eigenvalue

$$v_1, v_2, \ldots, v_\rho,$$

and denote $\mathcal{W}_i$ the A-invariant subspace of $\mathcal{W}_*$ spanned, by $v_1, v_2, \ldots, v_i$.

Theorem 14.11:

Let (A, B) be stabilizable and the subspaces $\mathcal{W}_i$ of $\mathcal{W}_*$ satisfy the above assumptions. Then, for all fixed t and a given terminal condition $T \geq 0$, the solution $X(t, t_2, T)$ of Equation 14.11 converges to a constant solution of Equation 14.15 as $t_2 \to \infty$ if and only if the subspace $\mathcal{W}_{k+1}$ corresponding to any pair λ_k, λ_{k+1} of complex conjugate eigenvalues is such that dim $T\mathcal{W}_{k+1}$ equals either dim $T\mathcal{W}_{k-1}$ or dim $T\mathcal{W}_{k-1} + 2$.

Here is a simple example. Let

$$A = \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}, \quad B = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad C = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}.$$

The pair (A, B) is stabilizable and A has two eigenvalues $1 + j$ and $1 - j$. The corresponding eigenvectors

$$v_1 = \begin{bmatrix} j \\ 1 \end{bmatrix}, \quad v_2 = \begin{bmatrix} -j \\ 1 \end{bmatrix}$$

span $\mathcal{W}_*$. Now consider the terminal condition

$$T = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}.$$

Then,

$$T\mathcal{W}_0 = 0, \quad T\mathcal{W}_2 = \text{Im} \begin{bmatrix} 1 \\ 0 \end{bmatrix}.$$

Theorem 14.11 shows that $X(t, t_2, T)$ does not converge to a constant matrix; in fact,

$$X(t, t_2, T) = \frac{1}{1 + e^{2(t-t_2)}} \begin{bmatrix} 2\cos^2(t - t_2) & -\sin 2(t - t_2) \\ -\sin 2(t - t_2) & 2\sin^2(t - t_2) \end{bmatrix}$$

tends to a periodic solution if $t_2 \to \infty$. On the other hand, if we select

$$T_0 = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$$

we have

$$T_0\mathcal{W}_0 = 0, \quad T_0\mathcal{W}_2 = 0$$

and $X(t, t_2, T_0)$ does converge. Also, if we take

$$T_1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

we have

$$T_1\mathcal{W}_0 = 0, \quad T_1\mathcal{W}_2 = R^2$$

and $X(t, t_2, T_1)$ converges as well.

If the solution $X(t, t_2, T)$ of Equation 14.11 converges to a constant matrix X_T as $t_2 \to \infty$, then X_T is a real symmetric nonnegative definite solution of Equation 14.15. Which solution is attained for a particular terminal condition?

Theorem 14.12:

Let (A, B) be stabilizable. Let

$$X_T = \lim_{t_2 \to \infty} X(t, t_2, T)$$

for a fixed $T \geq 0$. Then $X_T \geq 0$ is the solution of Equation 14.15 corresponding to the subspace $\mathcal{W}_T$ of $\mathcal{W}_*$, defined as the span of the real vectors v_i such that $T\mathcal{W}_i = T\mathcal{W}_{i-1}$ and of the complex conjugate pairs v_k, v_{k+1} such that $T\mathcal{W}_{k+1} = T\mathcal{W}_{k-1}$.

The cases of special interest are the extreme solutions X_+ and X_*. The solution $X(t, t_2, T)$ of Equation 14.12 tends to X_+ if and only if the intersection of $\mathcal{W}_*$ with Ker T is zero, and to X_* if and only if $\mathcal{W}_*$ is contained in Ker T.

This is best illustrated in the previous example, where

$$A = \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}, \quad B = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad C = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$$

and $\mathcal{W}_* = R^2$. Then $X(t, t_2, T)$ converges to X_+ if and only if T is positive definite; for instance, the identity matrix T yields the solution

$$X(t, t_2, I) = \frac{2}{1 + e^{2(t-t_2)}} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix},$$

which tends to

$$X_+ = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}.$$

On the other hand, $X(t, t_2, T)$ converges to X_* if and only if T=0; then

$$X(t, t_2, 0) = 0$$

and X_*=0 is a fixed point of Equation 14.11.

14.6 Optimal Control and Filtering: Application

The problems of optimal control and filtering introduced in Section 14.2 are related to the matrix Riccati differential Equations 14.5 and 14.9, respectively. These problems are defined over a finite horizon $t_2 - t_1$. We now apply the convergence properties of the solutions to study the two optimal problems in case the horizon becomes large.

To fix ideas, we concentrate on the optimal control problem. The results can easily be interpreted in the filtering context owing to the duality between Equations 14.5 and 14.9.

We recall that the finite horizon optimal control problem is that of minimizing the cost functional of Equation 14.4,

$$\eta(t_2) = \int_{t_1}^{t_2} [y'(t)y(t) + u'(t)u(t)]\, dt + x'(t_2)Tx(t_2)$$

along the solutions of Equation 14.3,

$$\dot{x}(t) = Fx(t) + Gu(t)$$
$$y(t) = Hx(t).$$

The optimal control has the form

$$u_0(t) = -G'X(t, t_2, T)x(t),$$

where $X(t, t_2, T)$ is the solution of Equation 14.11,

$$\dot{X}(t) + X(t)A + A'X(t) - X(t)BX(t) + C = 0$$

subject to the terminal condition $X(t_2) = T$, and where

$$A = F, \quad B = GG', \quad C = H'H.$$

The optimal control can be implemented as a state feedback and the resulting closed-loop system is

$$\dot{x}(t) = [F - GG'X(t, t_2, T)]x(t)$$
$$= [A - BX(t, t_2, T)]x(t).$$

Hence, the relevance of the matrix $A - BX$, which plays a key role in the theory of the Riccati equation.

The *infinite horizon* optimal control problem then amounts to finding

$$\eta_* = \inf_{u(t)} \lim_{t_2 \to \infty} \eta(t_2) \tag{14.17}$$

and the corresponding optimal control $u_*(t)$, $t \geq t_1$ achieving this minimum cost.

The *receding horizon* optimal control problem is that of finding

$$\eta_{**} = \lim_{t_2 \to \infty} \inf_{u(t)} \eta(t_2) \tag{14.18}$$

and the limiting behavior $u_{**}(t)$, $t \geq t_1$ of the optimal control $u_o(t)$.

The question is whether η_* is equal to η_{**} and whether u_* coincides with u_{**}. If so, the optimal control for the infinite horizon can be approximated by the optimal control of the finite horizon problem for a sufficiently large time interval.

It turns out that these two control problems have different solutions corresponding to different solutions of the matrix Riccati algebraic Equation 14.15,

$$XA + A'X - XBX + C = 0.$$

Theorem 14.13:

Let (A, B) be stabilizable. Then the infinite horizon optimal control problem of Equation 14.17 has a solution

$$\eta_* = x'(t_1)X_o x(t_1), \quad u_*(t) = -G'X_o x(t)$$

where $X_o \geq 0$ is the solution of Equation 14.15 corresponding to $\mathcal{W}_o$, the largest A-invariant subspace contained in $\mathcal{W}_* \cap Ker\ T$.

Theorem 14.14:

Let (A, B) be stabilizable. Then the receding horizon optimal control problem of Equation 14.18 has a solution if and only if the criterion of Theorem 14.5 is satisfied and, in this case,

$$\eta_{**} = x'(t_1)X_T x(t_1), \quad u_{**}(t) = -G'X_T x(t)$$

where $X_T \geq 0$ is the solution of Equation 14.16 corresponding to $\mathcal{W}_T$ and defined in Theorem 14.5.

The equivalence result follows.

Theorem 14.15:

The solution of the infinite horizon optimal control problem is exactly the limiting case of the receding horizon optimal control problem if and only if the subspace $\mathcal{W}_* \cap KerT$ is invariant under A.

A simple example illustrates these points. Consider the finite horizon problem defined by

$$\dot{x}_1(t) = 2x_1(t) + u_1(t),$$
$$\dot{x}_2(t) = x_2(t) + u_2(t)$$

and

$$\eta(t_2) = [x_1(t_2) + x_2(t_2)]^2 + \int_{t_2}^{\int} t_1(\tau) + u_2^2(\tau)]\, d\tau,$$

which corresponds to the data

$$A = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}, \quad B = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad C = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$$

and

$$T = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}.$$

Clearly $\mathcal{W}_* = R^2$ and the subspace

$$\mathcal{W}_* \cap \text{Ker } T = \text{Im} \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

is not invariant under A. Hence, the infinite and receding horizon problems are not equivalent.

The lattice of symmetric nonnegative definite solutions of Equation 14.11 has the four elements

$$X_+ = \begin{bmatrix} 4 & 0 \\ 0 & 2 \end{bmatrix}, \quad X_1 = \begin{bmatrix} 0 & 0 \\ 0 & 2 \end{bmatrix}, \quad X_2 = \begin{bmatrix} 4 & 0 \\ 0 & 0 \end{bmatrix}, \quad X_* = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$$

depicted in Figure 14.3.

Since the largest A-invariant subspace of $\mathcal{W}_* \cap \text{Ker}T$ is zero, the optimal solution X_o of Equation 14.11 is the maximal element X_+. The infinite horizon optimal control reads

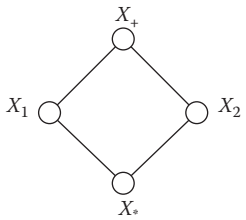
$$u_{1*}(t) = -4x_1(t),$$
$$u_{2*}(t) = -2x_2(t),$$

and affords the minimum cost

$$\eta_* = 4x_1^2(t_1) + 2x_2^2(t_1).$$

Now the eigenvectors of A spanning $\mathcal{W}_*$ are

$$v_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad v_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$



Q1 **FIGURE 14.3** The four elements of the lattice of solutions.

and their T-images

$$Tv_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad Tv_2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

are linearly dependent. Hence, $\mathcal{W}_T$ is spanned by v_2 only,

$$\mathcal{W}_T = \text{Im} \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

and the optimal limiting solution X_T of Equation 14.11 equals X_2. The receding horizon optimal control reads

$$u_{1**}(t) = -4x_1(t)$$
$$u_{2**}(t) = 0$$

and affords the minimum cost

$$\eta_{**}(t) = 4x_1^2(t_1).$$

The optimal control problems with large horizon are practically relevant if the optimal closed-loop system

$$\dot{x}(t) = (A - BX)x(t)$$

is stable. A real symmetric nonnegative definite solution X of Equation 14.15 is said to be *stabilizing* if the eigenvalues of $A - BX$ all have negative real part. It is clear that the stabilizing solution, if it exists, is the maximal solution X_+. Thus, the existence of a stabilizing solution depends on $A - BX_+$. having eigenvalues with only negative real part.

Theorem 14.16:

Equation 14.15 has a stabilizing solution if and only if (A, B) is stabilizable and the Hamiltonian matrix H of Equation 14.16 has no pure imaginary eigenvalue.

The optimal controls over large horizons have a certain stabilizing effect. Indeed, if $X \geq 0$ is a solution of Equation 14.15 that corresponds to an A-invariant subspace $\mathcal{W}$ of $\mathcal{W}_*$, then the control $u(t) = -G'Xx(t)$ leaves unstable in $A - BX$ just the eigenvalues of A associated with $\mathcal{W}$; all the remaining eigenvalues of A with positive real part are stabilized. Of course, the pure imaginary eigenvalues of A, if any, cannot be stabilized; they remain intact in $A - BX$ for any solution X of Equation 14.15.

In particular, the infinite horizon optimal control problem leaves unstable the eigenvalues of A associated with Ω_o, which are those not detectable either in C or in T, plus the pure imaginary eigenvalues. It follows that the infinite horizon optimal control results in a stable system if and only if X_o is the stabilizing solution of Equation 14.15. This is the case if and only if the hypotheses of Theorem 14.6 hold and $\mathcal{W}_o$, the largest A-invariant subspace contained in $\mathcal{W}_* \cap \text{Ker}T$, is zero. Equivalently, this corresponds to the pair

$$\left(\begin{bmatrix} C \\ T \end{bmatrix}, A \right)$$

being detectable.

The allocation of the closed-loop eigenvalues for the receding horizon optimal control problem is different, however. This control leaves unstable all eigenvalues of A associated with $\mathcal{W}_T$, where $\mathcal{W}_T$ is a subspace of $\mathcal{W}_*$ defined in Theorem 14.5. Therefore, the number of stabilized eigenvalues may be lower, equal to the dimension of $T\mathcal{W}_*$, whenever Ker T is not invariant under A. It follows that the receding horizon optimal control results in a stable system if and only if X_T is the stabilizing solution

of Equation 14.15. This is the case if and only if the hypotheses of Theorem 14.6 hold and $\mathcal{W}_T$ is zero. Equivalently, this corresponds to $\mathcal{W}_* \cap \mathrm{Ker} T = 0$. Note that this case occurs in particular if $T \geq X_+$.

It further follows that under the standard assumption, namely that

$$(A, B) \text{ stabilizable}$$
$$(A, C) \text{ detectable},$$

both infinite and receding horizon control problems have solutions; these solutions are equivalent for any terminal condition T; and the resulting optimal system is stable.

14.7 Numerical Solution

The matrix Riccati *differential* Equation 14.11 admits an analytic solution only in rare cases. A numerical integration is needed and the Runge–Kutta methods can be applied.

A number of techniques are available for the solution of the matrix Riccati *algebraic* Equation 14.15. These include invariant subspace methods and the matrix sign function iteration. We briefly outline these methods here with an eye on the calculation of the stabilizing solution to Equation 14.15.

14.7.1 Invariant Subspace Method

In view of Theorem 14.4, any solution X of Equation 14.15 can be computed from a Jordan form reduction of the associated $2n \times 2n$ Hamiltonian matrix

$$H = \begin{bmatrix} A & -B \\ -C & -A' \end{bmatrix}.$$

Specifically, compute a matrix of eigenvectors V to perform the following reduction:

$$V^{-1} H V = \begin{bmatrix} -J & 0 \\ 0 & J \end{bmatrix},$$

where $-J$ is composed of Jordan blocks corresponding to eigenvalues with negative real part only. If the stabilizing solution X exists, then H has no eigenvalues on the imaginary axis and J is indeed $n \times n$. Writing

$$V = \begin{bmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{bmatrix},$$

where each V_{ij} is $n \times n$, the solution sought is found by solving a system of linear equations,

$$X = V_{21} V_{11}^{-1}.$$

However, there are numerical difficulties with this approach when H has multiple or near-multiple eigenvalues. To ameliorate these difficulties, a method has been proposed in which a nonsingular matrix V of eigenvectors is replaced by an orthogonal matrix U of Schur vectors so that

$$U' H U = \begin{bmatrix} S_{11} & S_{12} \\ 0 & S_{22} \end{bmatrix}$$

where now S_{11} is a quasi-upper triangular matrix with eigenvalues having negative real part and S_{22} is a quasi-upper triangular matrix with eigenvalues having positive real part. When

$$U = \begin{bmatrix} U_{11} & U_{12} \\ U_{21} & U_{22} \end{bmatrix},$$

we observe that

$$\begin{bmatrix} V_{11} \\ V_{21} \end{bmatrix}, \quad \begin{bmatrix} U_{11} \\ U_{21} \end{bmatrix}$$

span the same invariant subspace and X can again be computed from

$$X = U_{21} U_{11}^{-1}.$$

14.7.2 Matrix Sign Function Iteration

Let M be a real $n \times n$ matrix with no pure imaginary eigenvalues. Let M have a Jordan decomposition $M = V\,J\,V^{-1}$ and let $\lambda_1, \lambda_2, \ldots, \lambda_n$ be the diagonal entries of J (the eigenvalues of M repeated according to their multiplicities). Then the *matrix sign function* of M is given by

$$\operatorname{sgn} M = V \begin{bmatrix} \operatorname{sgn} Re\lambda_1 & & \\ & \ddots & \\ & & \operatorname{sgn} Re\lambda_n \end{bmatrix} \mathrm{V}^{-1}$$

It follows that the matrix $Z = \operatorname{sgn} M$ is diagonalizable with eigenvalues ± 1 and $Z^2 = I$. The key observation is that the image of $Z + I$ is the M-invariant subspace of R^n corresponding to the eigenvalues of M with negative real part.

This property clearly provides the link to Riccati equations, and what we need is a reliable computation of the matrix sign. Let $Z_0 = M$ be an $n \times n$ matrix whose sign is desired. For $k = 0, 1$, perform the iteration

$$Z_{k+1} = \frac{1}{2c}(Z_k + c^2 Z_k^{-1}),$$

where $c = |\det Z_k|^{1/n}$. Then

$$\lim_{k \to \infty} Z_k = Z = \operatorname{sgn} M.$$

The constant c is chosen to enhance convergence of this iterative process. If $c = 1$, the iteration amounts to Newton's method for solving the equation

$$Z^2 - I = 0.$$

Naturally, it can be shown that the iteration is ultimately quadratically convergent.

Thus, to obtain the stabilizing solution X of Equation 14.15, provided it exists, we compute $Z = \operatorname{sgn} H$, where H is the Hamiltonian matrix of Equation 14.16. The existence of X guarantees that H has no eigenvalues on the imaginary axis.

Writing

$$Z = \begin{bmatrix} Z_{11} & Z_{12} \\ Z_{21} & Z_{22} \end{bmatrix},$$

where each Z_{ij} is $n \times n$, the solution sought is found by solving a system of linear equations

$$\begin{bmatrix} Z_{12} \\ Z_{22} + I \end{bmatrix} X = -\begin{bmatrix} Z_{11} + I \\ Z_{21} \end{bmatrix}.$$

14.7.3 Concluding Remarks

We have discussed two numerical methods for obtaining the stabilizing solution of the matrix Riccati algebraic Equation 14.15. They are both based on the intimate connection between the Riccati equation solutions and invariant subspaces of the associated Hamiltonian matrix. The method based on Schur vectors is a direct one, while the method based on the matrix sign function is iterative.

The Schur method is now considered one of the more reliable for Riccati equations and has the virtues of being simultaneously efficient and numerically robust. It is particularly suitable for Riccati equations with relatively small dense coefficient matrices, say, of the order of a few hundreds or less. The matrix sign function method is based on the Newton iteration and features global convergence, with ultimately quadratic order. Iteration formulas can be chosen to be of arbitrary order convergence in exchange for, naturally, an increased computational burden. The effect of this increased computation can, however, be ameliorated by parallelization.

The two methods are not limited to computing the stabilizing solution only. The matrix sign iteration can also be used to calculate X_-, the antistabilizing solution of Equation 14.15, by considering the matrix sgn $H - I$ instead of sgn $H + I$. The Schur approach can be used to calculate any, not necessarily symmetric, solution of Equation 14.15, by ordering the eigenvalues on the diagonal of S accordingly.

Acknowledgments

Acknowledgment to Project 1M0567 Ministry of Education of the Czech Republic.

References

Q2

Historical documents:

1. Riccati, J. F., Animadversationes in aequationes differentiales secundi gradus, *Acta Eruditorum Lipsiae*, 8, 67–73, 1724.
2. Boyer, C. B., *The History of Mathematics*, Wiley, New York, NY, 1974.

Tutorial textbooks:

3. Reid, W. T., *Riccati Differential Equations,* Academic Press, New York, NY, 1972.
4. Bittanti, S., Laub, A. J., and Willems, J. C., Eds., *The Riccati Equation,* Springer-Verlag, Berlin, 1991.

Survey paper:

5. Kuèera, V., A review of the matrix Riccati equation, *Kybernetika*, 9, 42–61, 1973.

Original sources on optimal control and filtering:

6. Kalman, R. E., Contributions to the theory of optimal control, *Bol. Soc. Mat. Mexicana*, 5, 102–119, 1960.
7. Kalman, R. E. and Bucy, R. S., New results in linear filtering and prediction theory, *J. Basic Eng. (ASME Trans.)*, 83D, 95–108, 1961.

Original sources on the algebraic equation:

8. Willems, J. C., Least squares stationary optimal control and the algebraic Riccati equation, *IEEE Trans. Autom. Control*, 16, 612–634, 1971.
9. Kuèera, V., A contribution to matrix quadratic equations, *IEEE Trans. Autom. Control*, 17, 344–347, 1972.

Original sources on the limiting behavior:

10. Callier, F. M. and Willems, J. L., Criterion for the convergence of the solution of the Riccati differential equation, *IEEE Trans. Autom. Control*, 26, 1232–1242, 1981.
11. Willems, J. L. and Callier, F. M., Large finite horizon and infinite horizon LQ-optimal control problems, *Optimal Control Appl. Methods*, 4, 31–45, 1983.

Original sources on the numerical methods:

12. Laub, A. J., A Schur method for solving algebraic Riccati equations, *IEEE Trans. Autom. Control*, 24, 913–921, 1979.
13. Roberts, J. D., Linear model reduction and solution of the algebraic Riccati equation by use of the sign function, *Int. J. Control*, 32, 677–687, 1980.

Cat: 73648 **Chapter: 14**

TO: CORRESPONDING AUTHOR

AUTHOR QUERIES - TO BE ANSWERED BY THE AUTHOR

The following queries have arisen during the typesetting of your manuscript. Please answer these queries by marking the required corrections at the appropriate point in the text.

Q1	We have modified the caption of Figure 14.3 as “The four elements of the lattice solutions” because the captions of Figures 14.2 and 14.3 are identical. Please confirm if this is OK.	
Q2	Only reference [1] has been cited in the text. Please confirm whether can we place this under “Reference” and the rest under “Further Reading”	