Properties of a Quadratic Matrix Equation and the Solution of the Continuous-Time Algebraic Riccati Equation

Hong-guo Xu
Department of Mathematics
Fudan University
Shanghai, 200433, P.R. China

and

Lin-zhang Lu
Department of Mathematics
Xiamen University
Xiamen, Fujian, P.R. China

Submitted by Ludwig Elsner

ABSTRACT

We discuss some properties of a quadratic matrix equation with some restrictions, then use these results on the algebraic Riccati equation to get a new algorithm. The algorithm sufficiently takes account of the structure of the associated matrix; hence it is very effective.

1. INTRODUCTION

It is a problem of great practical importance to solve the algebraic Riccati equation (ARE). There are no completely satisfactory methods; there are some efficient ones, but they are not stable. Since A. J. Laub [3] proposed a Schur method on a Hamiltonian matrix \( M \) from the ARE, there have been many attempts to design an efficient algorithm by using the Hamiltonian property. R. Byers [4] gave a very beautiful algorithm for the single-input and the single-output problems. But the more general cases still remain unsolved. It is very difficult to reduce a Hamiltonian matrix \( M \) to the Hamiltonian Schur form with symplectic unitary QR-like algorithms. See [2] and [6].

In this paper we propose a new method for solving the ARE which is based on the properties of a quadratic matrix equation. The method can take account of special matrix structures.
First, we introduce some basic definitions and properties which will be used throughout this paper.

The real continuous-time algebraic Riccati equation is defined as follows:

\[ A^T X + X A - X D X + G = 0, \]  
(1.1)

where all matrices are in \( \mathbb{R}^{n \times n} \) and \( D = B B^T, \ G = C^T C \) are full-rank factorizations of \( D \) and \( G \). If the matrices \( A, B, C \) satisfy the conditions

\[ (A, B) \text{ is stabilizable}, \]  
(1.2a)

\[ (C, A) \text{ is detectable}, \]  
(1.2b)

then \( X \), the solution of the ARE satisfying \( \text{Re} \lambda(A - DX) < 0 \) and \( X = X^T \geq 0 \), exists and is unique [3].

**Definition.** (i) Let

\[ J = \begin{bmatrix} 0 & I_n \\ -I_n & 0 \end{bmatrix}, \]

where \( I_n \) denotes the \( n \times n \) identity matrix. Note that \( J^T = J^{-1} = -J \).

(ii) \( M \in \mathbb{R}^{2n \times 2n} \) is Hamiltonian if

\[ (JM)^T = JM, \quad M = \begin{bmatrix} A & D \\ G & -A^T \end{bmatrix}, \]

where \( A, D, G \in \mathbb{R}^{n \times n}, \ D = D^T, \ G = G^T \).

(iii) \( N \in \mathbb{R}^{2n \times 2n} \) is skew-Hamiltonian if

\[ (JN)^T = -JN, \quad N = \begin{bmatrix} A & D \\ G & -A^T \end{bmatrix}, \]

where \( A, D, G \in \mathbb{R}^{n \times n}, \ D = -D^T, \ G = -G^T \).

(iv) \( V \in \mathbb{R}^{2n \times 2n} \) is real symplectic if \( V^T J V = J \). \( V \) is complex symplectic if \( V \in \mathbb{C}^{2n \times 2n}, \ V^H J V = J \).

(v) \( S \in \mathbb{R}^{2n \times 2n} \) is symplectic orthogonal if

\[ S^T J S = J, \quad S S^T = I_{2n}, \quad S = \begin{bmatrix} S_1 & S_2 \\ -S_2 & S_1 \end{bmatrix}. \]

If \( S \in \mathbb{C}^{2n \times 2n}, \ S^H J S = J, \ S^H S = S S^H = I \), then \( S \) is symplectic unitary.
Now consider the Hamiltonian matrix

\[ M = \begin{bmatrix} A & D \\ G & -A^T \end{bmatrix}, \]

(1.3)

where \( A, D, G \) are from (1.1). If the conditions (1.2) are satisfied, we have the following properties: (i) \( \sigma(M) = \{\lambda_1, \ldots, \lambda_n, -\lambda_1, \ldots, -\lambda_n\} \), \( \text{Re}\lambda_i < 0 \), \( i = 1, \ldots, n \), where \( \sigma(M) \) represents the set of all eigenvalues of \( M \).

(ii) There is a symplectic orthogonal matrix \( U = [U_1, U_2] \), \( U_1, U_2 \in \mathbb{R}^{2n \times n} \), such that

\[ U^T MU = \begin{bmatrix} W & R \\ 0 & -W^T \end{bmatrix} \]

is the real Hamiltonian Schur form, where \( W, R \in \mathbb{R}^{n \times n} \), \( R = R^T \), and \( W \) is quasi-upper-triangular with \( 2 \times 2 \) blocks on the block diagonal corresponding to the complex conjugate eigenvalues and \( 1 \times 1 \) blocks corresponding to the real eigenvalues. \( \sigma(W) = \sigma(W^T) = \{\lambda_1, \ldots, \lambda_n\} \).

(iii) There is a symplectic unitary matrix \( U = [U_1, U_2] \), \( U_1, U_2 \in \mathbb{C}^{2n \times n} \), such that

\[ U^H MU = \begin{bmatrix} W & R \\ 0 & -W^H \end{bmatrix} \]

is the complex Hamiltonian Schur form, where \( W, R \in \mathbb{C}^{n \times n} \), \( R = R^H \), \( W \) is upper triangular, and \( \sigma(W) = \sigma(W^H) = \{\lambda_1, \ldots, \lambda_n\} \).

(iv) If there is \( U_1 = [U_{11}^H, U_{12}^H]^H \in \mathbb{C}^{2n \times n} \) that is of full column rank, such that \( MU_1 = U_1 W \), where \( W \in \mathbb{C}^{n \times n} \), \( \text{Re}\lambda(W) < 0 \), then \( U_1^{-1} \) exists and the solution of (1.1) is \( X = -U_2 U_1^{-1} \). Clearly \( U_1 \) in (ii), (iii) satisfies the conditions. See [2] and [3].

Clearly \( M^2 \) is skew-Hamiltonian and we have

(v) \( \sigma(M^2) = \{\lambda_1^2, \ldots, \lambda_n^2, \lambda_1^2, \ldots, \lambda_n^2\} \); similarly, \( M^2 \) has the real and complex skew-Hamiltonian Schur forms.

In our discussions we assume that Equation (1.1) always satisfies the conditions in (1.2), i.e., all the properties above always hold. The matrix \( M \) is always real, and \( \lambda_i \), \( i = 1, \ldots, n \), are the \( n \) eigenvalues of \( M \), with \( \text{Re}\lambda_i < 0 \), \( i = 1, \ldots, n \). We will use these assumptions without explanation in the following sections.

We will give some properties of a quadratic matrix equation in Section 2. With these properties, we set up our results for the solution of the ARE in Section 3; then, in Sections 4 and 5, using Van Loan’s idea [1], we design an algorithm for the real ARE using only real operations, though we also can give one using complex operation. In Section 6 we show several numerical examples which are from Laub’s paper [3], and Section 7 is our conclusion.
2. SOLUTIONS OF A QUADRATIC MATRIX EQUATION

Now we consider the quadratic matrix equation
\[ X^2 - BX + C = 0 \quad \text{with } C \text{ nonsingular.} \quad (2.1) \]

**Theorem 2.1.** Given a matrix \( X \in \mathbb{C}^{n \times n} \), suppose that (a) \( X \) satisfies the equation \((2.1)\) and \( X^2 - C \) is nonsingular; (b) There exist two matrices \( X_1 \) and \( X_2 \) such that

\[
X_1 + X_2 = B, \quad X_1X_2 = X_2X_1 = C, 
\]

and \( \sigma(X_1) \cap \sigma(X_2) = \emptyset, \sigma(X_1) \cup \sigma(X_2) = \sigma(X) \); (c) \( X_1X = XX_1, \quad X_2X = XX_2 \). Then the columns of \( X - X_1 \) and \( X - X_2 \) span the invariant subspace of \( X \) associated with the set \( \sigma(X) \cap \sigma(X_2) \) and the set \( \sigma(X) \cap \sigma(X_1) \) respectively.

**Proof.** In view of the assumptions, we have

\[
(X - X_2)(X - X_1) = X^2 - (X_1 + X_2)X + X_1X_2 = X^2 - BX + C = 0; 
\]

hence

\[
X(X - X_1) - X_2(X - X_1) = (X - X_1)X_2. \quad (2.2)
\]

Similarly, we have

\[
X(X - X_2) = (X - X_2)X_1. \quad (2.3)
\]

Thus we know from \((2.2)\) and \((2.3)\) that \( X - X_1 \) and \( X - X_2 \) are in the invariant subspaces of \( X \) associated with the set \( \sigma(X) \cap \sigma(X_2) \) and \( \sigma(X) \cap \sigma(X_1) \) respectively.

Furthermore, let \( k \) and \( l \) be the dimension numbers of the invariant subspace of \( X \) associated with \( \sigma(X) \cap \sigma(X_1) \) and \( \sigma(X) \cap \sigma(X_2) \); then in view of assumption (b) of the theorem, \( k + l = n \), and \((2.2)\) and \((2.3)\) give that \( \text{rank}(X - X_1) \leq l, \text{rank}(X - X_2) \leq k \). Now we only have to prove that

\[
\text{rank}(X - X_1) + \text{rank}(X - X_2) = n.
\]

To see this, we use assumption (a). Because \( C \) is nonsingular and \( (X - B)X = -C \), it follows that \( X \) is also nonsingular. Because \( X^2 - C \) is nonsingular and

\[
X(X - X_1 + X - X_2) = X^2 + X^2 - BX = X^2 - C,
\]
so that \( X - X_1 + X - X_2 \) is also nonsingular, we have

\[
n = \text{rank}(X - X_1 + X - X_2) \leq \text{rank}(X - X_1) + \text{rank}(X - X_2) \leq k + l = n.
\]

**Theorem 2.2.** Suppose that there is a nonsingular matrix \( U \) such that 

\[
\tilde{B} = UBU^{-1} = (\tilde{b}_{ij}) \quad \text{and} \quad \tilde{C} = UCU^{-1} = (\tilde{c}_{ij}) \quad \text{are both upper triangular.}
\]

If there are \( t_1, t_2, \ldots, t_n \in \mathbb{C} \) satisfying

\[
t^2_i - \tilde{b}_{ii} t_i + \tilde{c}_{ii} = 0, \quad \text{but} \quad t^2_i - \tilde{c}_{ii} \neq 0, \quad i = 1, \ldots, n, \tag{2.4}
\]

\[
\text{if} \quad t_i \neq t_j, \quad \text{then} \quad t^2_i - \tilde{b}_{jj} t_i + \tilde{c}_{jj} \neq 0, \tag{2.5}
\]

then the equation (2.1) has solution \( \tilde{Y} = U^{-1} \tilde{Y} U \), where \( \tilde{Y} \) is upper triangular and \( \tilde{y}_{ii} = t_i, i = 1, \ldots, n \).

**Proof.** We only have to show that there exists an upper triangular \( \tilde{Y} \) satisfying

\[
\tilde{y}^2_{ii} - \tilde{b}_{ii} \tilde{y}_{ii} + \tilde{c}_{ii} = 0, \quad i = 1, 2, \ldots, n, \tag{2.7a}
\]

\[
\alpha_{ij} \tilde{y}_{ij} = \tilde{b}_{ij} \tilde{y}_{jj} + \sum_{k=i+1}^{j-1} (\tilde{b}_{ik} - \tilde{y}_{ik}) \tilde{y}_{kj} - \tilde{c}_{ij}, \quad i < j, \tag{2.7b}
\]

where \( \alpha_{ij} = \tilde{y}_{ii} + \tilde{y}_{jj} - \tilde{b}_{ij} \quad \text{for} \quad i < j \). Thus \( \tilde{y}_{11}, \tilde{y}_{22}, \ldots, \tilde{y}_{nn} \) exist if we choose \( \tilde{y}_{ii} = t_i, i = 1, \ldots, n \). Under this conditions we have to show that

\[
\alpha_{ij} = \tilde{y}_{ii} + \tilde{y}_{jj} - \tilde{b}_{ij} = \frac{\tilde{y}_{ii} \tilde{y}_{jj} - \tilde{c}_{ii}}{\tilde{y}_{ii}} \neq 0
\]

under the assumptions (2.4), (2.5).

In fact, when \( \tilde{y}_{ii} = \tilde{y}_{jj} \), if \( \alpha_{ij} = 0 \), then \( \tilde{y}_{ii}^2 - \tilde{c}_{ii} = 0 \); this is impossible because of (2.4). When \( \tilde{y}_{ii} \neq \tilde{y}_{jj} \), if \( \alpha_{ij} = 0 \), then

\[
\tilde{y}_{ii} + \tilde{y}_{jj} = \tilde{b}_{ii}, \quad \tilde{y}_{ii} \tilde{y}_{jj} = \tilde{c}_{ii},
\]

where \( \tilde{y}_{ii} \) and \( \tilde{y}_{jj} \) are two roots of \( y^2 - \tilde{b}_{ii} y + \tilde{c}_{ii} = 0 \); this is also impossible because of (2.5). Therefore, in any case, \( \alpha_{ij} \neq 0 \); hence \( \tilde{Y} = (\tilde{y}_{ij}) \) can
be got from (2.7), and finally $Y$ can be obtained in the form given in the theorem.

Theorem 2.2 shows that if matrices $B, C$ can be simultaneously triangularized and the spectra of $B, C$ have the properties (2.4) and (2.5), then a solution of (2.1) can be constructed.

In the following, we discuss the commutativity of two roots of (2.1) under the special case $B = 0$, which is enough for our application in solving the ARE. The following lemma is needed for the proof of the commutativity.

**Lemma 2.3.** Let

$$T(a_0, a_1, \ldots) = \begin{cases} T_k = \begin{bmatrix} a_0 & a_1 & \cdots & a_{k-1} \\ \vdots & \vdots & \ddots & \vdots \\ \vdots & \vdots & \ddots & a_1 \\ 0 & 0 & \cdots & a_0 \end{bmatrix} & k = 1, 2, \ldots \\ \end{cases} \in \mathbb{C}^{k \times k} \end{cases}$$

be the set of upper triangular Toeplitz matrices with parameters $a_0, a_1, \ldots$. If $\Lambda_n, \Lambda_m \subset T(a_0, a_1, \ldots), a_1 \neq 0$, and there is a matrix $B \in \mathbb{C}^{m \times n}$ such that

$$BA_n = A_m B,$$

then we have:

$m = n$: $B \subset \mathbb{C}^{n \times n}$, $B \subset T(b_0, b_1, \ldots)$ for arbitrary $b_0, b_1, \ldots$; (2.9a)

$m < n$: $B = \begin{bmatrix} n - m \\ 0 \end{bmatrix}, B_1 \in \mathbb{C}^{m \times m}$, and $B_1 \in T(b_0, b_1, \ldots); (2.9b)$

$m > n$: $B = \begin{bmatrix} B_1 \\ 0 \end{bmatrix}^n_{m-n}, B_1 \in \mathbb{C}^{n \times n}$, and $B_1 \in T(b_0, b_1, \ldots); (2.9c)$

and for arbitrary two matrices $T_m, T_n \in T(t_0, t_1, \ldots)$ with arbitrary
CONTINUOUS-TIME ALGEBRAIC RICCATI EQUATION

parameters $t_0, t_1, \ldots$,

$$BT_n = T_n B.$$ \hspace{1cm} (2.10)

**Proof.** We only consider the case of $m > n$. The other two cases are analogous. Then

$$BA_n = A_mB$$

is equivalent to

$$B = \begin{bmatrix} 0 & a_1 & \cdots & a_{n-1} \\ \vdots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ a_1 & \cdots & \cdots & 0 \end{bmatrix} \quad \text{for } i = m, m-1, \ldots, 1,$$

$$\text{for } j = 2, \ldots, n,$$

$$\sum_{k=1}^{j-1} a_k b_{i-k,j} = \sum_{k=1}^{m-i} a_k b_{i+k,j}.$$  

Combining this with the assumption $a_1 \neq 0$, we can get (2.9c).

We also consider the case of $m > n$ for (2.10). From (2.9c)

$$BT_n = \begin{bmatrix} \bar{B}_1 \\ 0 \end{bmatrix} T_n = \begin{bmatrix} \bar{B}_1 T_n \\ 0 \end{bmatrix},$$

$$T_n B = \begin{bmatrix} T_n & T_{n,m-n} \\ 0 & T_{m-n} \end{bmatrix} \begin{bmatrix} \bar{B}_1 \\ 0 \end{bmatrix} = \begin{bmatrix} T_n B_1 \\ 0 \end{bmatrix}.$$  

Thus we only prove $\bar{B}_1 T_n = T_n \bar{B}_1$.

Since $\bar{B}_1$, $T_n$ are both upper triangular, $\bar{B}_1 T_n$, $T_n \bar{B}_1$ are also both upper triangular. It is enough to compare the upper triangular elements.
For $i < j$,

$$(B_1 T_n)_{ij} = \sum_{k=i}^{j} b_{k-i} t_{j-k} = \sum_{k=0}^{j-i} b_k t_{j-i-k},$$

and

$$(T_n B_1)_{ij} = \sum_{u=i}^{j} t_{u-i} b_{j-u} = \sum_{u=0}^{j-i} t_u b_{j-i-u} = \sum_{k=0}^{j-i} b_k t_{j-i-k},$$

which implies that

$$(B_1 T_n)_{ij} = (T_n B_1)_{ij}, \quad i < j, \quad i, j = 1, \ldots, n.$$ 

Hence

$$B_1 T_n = T_n B_1$$

and

$$B T_n = T_n B.$$  

**Theorem 2.4.** Suppose that a matrix $Y$ satisfies (2.1) with $B = 0$; let

$$\sigma(Y) = \{\lambda_1, \lambda_2, \ldots, \lambda_n\}$$

where $\lambda_i (i = 1, \ldots, n)$ satisfy

$$\lambda_i + \lambda_j \neq 0, \quad i, j = 1, \ldots, n. \tag{2.11}$$

If there is a matrix $X$ such that $X$ and $C$ commute, then $X$ and $Y$ also commute.

**Proof.** Let $Y = U J_D U^{-1}$ be the Jordan canonical decomposition. Then

$$J_D = \text{diag}[J_1, \ldots, J_k],$$

$$\lambda_i \in \sigma(J_i), \quad \lambda_j \in \sigma(J_j), \quad \lambda_i \neq \lambda_j, \quad i \neq j, \quad i, j = 1, \ldots, k, \tag{2.12}$$

where

$$J_i = \text{diag}[J_{i,1}, \ldots, J_{i,p_i}], \quad i = 1, \ldots, k,$$

and

$$J_{i,l} = \begin{bmatrix}
\lambda_i & 1 & 0 \\
& \ddots & \ddots \\
& & \ddots & 1 \\
0 & \cdots & & \lambda_i
\end{bmatrix}, \quad l = 1, \ldots, p_i, \quad i = 1, \ldots, k.$$
From (2.1) we have
\[ C = -Y^2 = -UJ_d^2 U^{-1}, \]
where
\[ J_d^2 = \text{diag}[J_1^2, \ldots, J_k^2], \]
\[ J_i^2 = \text{diag}[J_{i,1}^2, \ldots, J_{i,p_i}^2], \quad i = 1, \ldots, k, \]
and
\[
J_{i,l}^2 = \begin{bmatrix}
\lambda_i^2 & 2\lambda_i & 1 & 0 \\
& \ddots & \ddots & \ddots \\
& & \lambda_i^2 & 2\lambda_i \\
0 & & & \lambda_i^2 
\end{bmatrix}, \quad l = 1, \ldots, p_i, \quad i = 1, \ldots, k.
\]

Because \( X \) and \( C \) commute, we get
\[ U^{-1}XUJ_d^2 = J_d^2 U^{-1}XU. \]

Let \( S = U^{-1}XU; \) we have
\[ SJ_d^2 = J_d^2 S. \]
Partition \( S \) conforming with \( J_d: S = [S_{ij}]_{k \times k}. \) We get
\[ S_{ij} J_i^2 = J_i^2 S_{ij}, \quad i, j = 1, \ldots, k. \quad (2.13) \]

From (2.11) and (2.12),
\[ i \neq j: \quad \lambda_i^2 \quad \lambda_j^2 = (\lambda_i + \lambda_j)(\lambda_i - \lambda_j) \neq 0, \quad \text{i.e., } \sigma(J_i^2) \cap \sigma(J_j^2) = 0. \]
Thus
\[ i \neq j: \quad S_{i,j} = 0. \]
Now \( S = \text{diag}[S_{11}, \ldots, S_{kk}], \) and \( SJ_d^2 = J_d^2 S \) is equivalent to
\[ S_{ii} J_i^2 = J_i^2 S_{ii}, \quad i = 1, \ldots, k. \quad (2.14) \]
Again we partition each $S_{ii}$ conforming with corresponding $J_i$: $S_{ii} = [S_{lm}^i]_{p_i \times p_i}$. From (2.14) we have

$$S_{lm}^i J_{lm}^{2i} = J_{il}^2 S_{lm}^i, \quad l, m = 1, \ldots, p_i, \quad i = 1, \ldots, k. \tag{2.15}$$

Now, according to the definitions in Lemma 2.3, $J_{il} \in T(\lambda_i, 1, 0, \ldots)$, and $J_{il}^2 \in T(\lambda_i^2, 2\lambda_i, 1, 0, \ldots)$; $J_{il}$ and $J_{il}^2$ have the same order. The condition (2.11) implies $2\lambda_i \neq 0$, and each equivalence of (2.15) corresponds to (2.8). Thus from (2.10) we have

$$S_{lm}^i J_{lm} = J_{il} S_{lm}^i, \quad l, m = 1, \ldots, p_i, \quad i = 1, \ldots, k,$$

which implies

$$S_{ii} J_i = J_i S_{ii}, \quad i = 1, \ldots, k.$$

Thus

$$SJ_D = J_D S.$$

Recalling $S = U^{-1}XU$, we have

$$U^{-1}XUJ_D = J_D U^{-1}XU,$$

which is equivalent to

$$XUJ_D U^{-1} = UJ_D U^{-1}X.$$

Thus

$$XY = YX. \quad \blacksquare$$

**Theorem 2.5.** Suppose all of the conditions of Theorem 2.4 are satisfied. Then $Y$, the solution of (2.1) with $B = 0$, is uniquely determined by the $n$ parameters $\lambda_1, \lambda_2, \ldots, \lambda_n$ if they satisfy (2.11).

**Proof.** When $B = 0$ in the equation (2.1), the inequality (2.11) is equivalent to (2.4) and (2.5); therefore the existence of $Y$ is obtained from Theorem 2.2. Now we prove the uniqueness.

Suppose there are two matrices $Y_1$ and $Y_2$ both satisfying the conditions of the theorem. Then from Theorem 2.4 we get that $Y_1$ and $Y_2$ commute, i.e.

$$Y_1 Y_2 = Y_2 Y_1. \tag{2.16}$$
From (2.1) and (2.16) we get

\[ Y_1^2 + C - (Y_2^2 + C) = Y_1^2 - Y_2^2 = (Y_1 + Y_2)(Y_1 - Y_2) = 0. \]  

(2.17)

Because \( Y_1 \) and \( Y_2 \) commute, they can be similar to the upper triangular forms simultaneously. So the eigenvalue of \( Y_1 + Y_2 \) takes the form \( \lambda_i + \lambda_j \), where \( \lambda_i \) (\( i = 1, \ldots, n \)) are defined in Theorem 2.4. Again by (2.11) we get that none of the eigenvalues of \( Y_1 + Y_2 \) is equal to zero, so \( Y_1 + Y_2 \) is nonsingular, and with (2.17) we get that\[ Y_1 = Y_2. \]

This proves the uniqueness.

3. APPLICATION IN SOLVING THE ARE

In this section, \( \mathbb{C}^+ (\mathbb{C}^-) \) denotes the open right (left) half plane. Let \( B = 0 \) and \( C = -M^2 \) in Equation (2.1), where \( M \) is the Hamiltonian matrix of (1.3), then \( M \) satisfies the equation (2.1) and \( M^2 - C = 2M^2 \) is nonsingular, since \( M \) is nonsingular under our assumption (see Section 1).

Suppose that \( C \) has Schur decomposition \( C = U \tilde{C} U^H \) with \( \tilde{C} = (\tilde{c}_{ij}) \) upper triangular and \( \tilde{c}_{ii} = \tilde{c}_{n+i,n+i} = -\lambda_i^2 \) (\( i = 1, \ldots, n \)), where \( \lambda_i \) (\( i = 1, \ldots, n \)) are defined as in Section 1, i.e., they are \( n \) eigenvalues of \( M \) and \( \text{Re} \lambda_i < 0 \) (\( i = 1, \ldots, n \)). Let \( Y = U \tilde{Y} U^H \), and take an upper triangular \( \tilde{Y} = (\tilde{y}_{ij}) \) with diagonals satisfying \( \tilde{y}_{ii} = \tilde{y}_{n+i,n+i} = \lambda_i \) (\( i = 1, \ldots, n \)). It is easy to verify that \( \tilde{y}_{ii} \) satisfies (2.4), (2.5), and (2.11) by using the properties \( \text{Re} \lambda_i < 0 \). In view of Theorem 2.2 (with \( B = 0 \)) and Theorem 2.5, the solution \( Y \) of \( Y^2 - M^2 = 0 \) satisfying \( \sigma(Y) \subseteq \mathbb{C}^- \) [i.e. \( \sigma(Y) = \{\lambda_1, \ldots, \lambda_n, \lambda_1, \ldots, \lambda_n\} \) exists and is unique. Furthermore, it is known from Theorem 2.4 that \( Y \) and \( M \) commute because that \( M \) and \( C (= -M^2) \) commute. Take \( \tilde{X}_1 = Y \) and \( \tilde{X}_2 = -Y \); by Theorem 2.1, the columns of \( M + Y \) span the invariant subspace of \( M \) associated with the eigenvalues in \( \sigma(M) \cap \mathbb{C}^- \). Thus solving the ARE can be reduced to finding the solution of equation

\[ Y^2 - M^2 = 0 \quad \text{with} \quad \sigma(Y) \subseteq \mathbb{C}^- . \]  

(3.1)

Now we do further search for the matrix \( Y \) in (3.1). Since the matrix \( M \) is Hamiltonian, we ask whether \( Y \) has any special structure. The answer is yes, as we see in the following discussion.
Lemma 3.1. If $M$ is Hamiltonian as given in (1.3), then there is a symplectic matrix $\tilde{U}$ with $\tilde{U}^{-1} M \tilde{U}$ of the form $\tilde{U}^{-1} M \tilde{U} = \text{diag}[W, -W^H]$ with $\text{Re}\lambda(W) < 0$. Further, if

\[
\tilde{U} = \begin{bmatrix} \tilde{U}_1 & \tilde{U}_2 \\ \tilde{U}_1 & \tilde{U}_2 \\ \end{bmatrix} = \begin{bmatrix} \tilde{U}_{11} & \tilde{U}_{12} \\ \tilde{U}_{21} & \tilde{U}_{22} \\ \end{bmatrix}, \quad \text{and thus} \quad \tilde{U}^{-1} = \begin{bmatrix} \tilde{U}_{22}^H & -\tilde{U}_{12}^H \\ -\tilde{U}_{21}^H & \tilde{U}_{11}^H \\ \end{bmatrix},
\]

(3.2)

then $\tilde{U}_{11}^{-1}, \tilde{U}_{22}^{-1}$ exist and the solution of (1.1) is $X = -\tilde{U}_{21} \tilde{U}_{11}^{-1}$.

Proof. From property (iii) in Section 1 we have

\[
U^H M U = \begin{bmatrix} W & R \\ 0 & -W^H \\ \end{bmatrix},
\]

where $U$ is symplectic unitary, $\text{Re}\lambda(W) < 0$, $R = R^H$.

The rest of the proof follows directly from Lyapunov’s theorem applied to the elimination of $R$ in

\[
\begin{bmatrix} W & R \\ 0 & -W^H \\ \end{bmatrix}.
\]

The invertibility of $\tilde{U}_{11}, \tilde{U}_{22}$ follows from a well-known result; see [2, 3].

Theorem 3.2. Let $M = \tilde{U} \text{diag}[W, -W^H] \tilde{U}^{-1}$, with $\tilde{U}$ symplectic and partitioned as in (3.2), $\sigma(W) = \{\lambda_1, \ldots, \lambda_n\}$, and $\text{Re}\lambda_i < 0$ ($i = 1, \ldots, n$). Then we have:

(i) There exists a unique real skew-Hamiltonian matrix $Y$, such that $Y^2 - M^2 = 0$, and $Y$ has $2n$ eigenvalues $\{\lambda_1, \ldots, \lambda_n, \lambda_1, \ldots, \lambda_n\}$.

(ii) With this $Y$ we have

\[
M + Y = \tilde{U}_{12} W [\tilde{U}_{22}^H - \tilde{U}_{12}^H],
\]

(3.3)

and if we let

\[
Y = \begin{bmatrix} Y_{11} & Y_{12} \\ Y_{21} & Y_{11}^T \\ \end{bmatrix},
\]

(3.4)

then $X$, the positive semidefinite solution of (1.1), (1.2), is

\[
X = -(G + Y_{21})(A + Y_{11})^{-1}.
\]

(3.5)
PROOF. (i) has already been obtained at the beginning of this section; we only need to prove that $Y$ is real and skew-Hamiltonian.

Since $M$ is real and $Y^2 - M^2 = 0$, then $\overline{Y}^2 - M^2 = 0$ and $\sigma(Y) = \{\overline{\lambda}_1, \ldots, \overline{\lambda}_n, \overline{\lambda}_1, \ldots, \overline{\lambda}_n\}$. By using the property that $M$ is real again we get $\sigma(\overline{Y}) = \sigma(Y)$. Hence $\overline{Y} = Y$ because of the uniqueness of $Y$, which means that $Y$ satisfying the conditions of the theorem is real.

From Lemma 3.1

$$M = \hat{U}\text{diag}[W, -W^H]\hat{U}^{-1},$$

thus

$$M^2 = \hat{U}\text{diag}[W^2, (W^H)^2]\hat{U}^{-1}.$$  

Let

$$\hat{Y} = \hat{U}\text{diag}[W, W^H]\hat{U}^{-1}. \quad (3.6)$$

Clearly $\hat{Y}^2 - M^2 = 0$, $\sigma(\hat{Y}) = \{\lambda_1, \ldots, \lambda_n, \lambda_1, \ldots, \lambda_n\}$, and $\hat{Y}$ is skew-Hamiltonian. Thus from the uniqueness we conclude that $Y$, which satisfies $Y^2 = M^2$, and $\sigma(Y) = \{\lambda_1, \ldots, \lambda_n, \lambda_1, \ldots, \lambda_n\}$ must be of the form (3.6) and thus is real skew-Hamiltonian.

Considering the forms of $U$ and $U^{-1}$ in (3.2), we have

$$M + Y = \hat{U}\text{diag}[2W, 0]\hat{U}^{-1} = \hat{U}_{11}2W[\hat{U}_{22}^H, -\hat{U}_{12}^H].$$

From the forms of $M$ and $Y$ mentioned above we get

$$A + Y_{11} = \hat{U}_{11}2W\hat{U}_{12}^H,$$
$$G + Y_{21} = \hat{U}_{21}2W\hat{U}_{22}^H.$$

$W$ is obviously nonsingular; from Lemma 3.1, $\hat{U}_{22}^H$ is nonsingular and

$$X = -\hat{U}_{21}\hat{U}_{12}^{-1} = -\hat{U}_{21}2W[\hat{U}_{11}2W\hat{U}_{22}^H]^{-1}$$
$$= -(G + Y_{21})(A + Y_{11})^{-1}. \quad \blacksquare$$

Theorem 3.2 shows the skew-Hamiltonian structure of the matrix $Y$ and the special form of $M + Y$. Therefore, if we can form $Y$ by some method, employing the first $n$ columns of $M + Y$ is enough to get the solution $X$ of (1.1). So the key is to find an efficient method to compute $Y$. We will use a method based on [1], because it reduces $M^2$ to the real skew-Hamiltonian Schur form with only symplectic orthogonal transformations and is quite stable.
4. METHOD FOR COMPUTING $Y$

The method is given in terms of real operations, though it also can be put in complex form.

Assume that $M^2$ has been put in real skew-Hamiltonian Schur form by the method in [1]:

$$V^T M^2 V = \begin{bmatrix} T & H \\ 0 & T^T \end{bmatrix},$$

(4.1)

where $V$ is symplectic orthogonal, $T$ is quasi-upper-triangular with $2 \times 2$ blocks on the block diagonal corresponding to the complex conjugate eigenvalues and $1 \times 1$ blocks corresponding to the real eigenvalues, and $H$ is skew-symmetric. Now we seek $Y_1$, such that

$$Y_1^2 = \begin{bmatrix} T & H \\ 0 & T^T \end{bmatrix}.$$  

Just let

$$Y_1 = \begin{bmatrix} W & K \\ 0 & W^T \end{bmatrix};$$

from $Y_1^2 = V^T M^2 V$ we get

$$W^2 = T,$$

(4.2)

$$WK + KW^T = H.$$  

(4.3)

Thus $W$ and $K$ can be produced by Higham's algorithm [14] and the numerical method for the solution of the Lyapunov equation presented in [5].

Thus we can produce $Y_1$ and let $Y = VY_1 V^T$; then $Y$ is the matrix we want.

**REMARK.** The method proposed above is only one way to form $Y$. Theoretically we can use other methods. Recently Lu [9] gave a new method to reduce $M^2$ by the Lanczos process. With his method, we get $M^2 = \hat{V} \text{diag} [\hat{T}, \hat{T}^T] \hat{V}^{-1}$, where $\hat{V}$ is real symplectic and $\hat{T}$ is real tridiagonal. We can reduce $T$ to the standard real Schur form with the QR algorithm. Thus we have

$$M^2 = V \text{diag} [T, T^T] V^{-1},$$

where $V = \hat{V} \text{diag} [Q, Q]$ is real symplectic, $Q$ is an orthogonal matrix, $\hat{T} = QTQ^T$, and $T$ is real quasi-upper-triangular.
Hence we can compute \( W: W^2 = T \) with \( \text{Re} \lambda(W) < 0 \), and we only compute the first \( n \) columns of \( Y \) according to Theorem 3.2:

\[
Y = V \text{diag}[W, W^T] V^{-1}.
\]

This method is quite simple for computing \( Y \). Unfortunately, the Lanczos process is usually unstable for computing the invariant subspace.

5. ALGORITHM

The algorithm consists of three steps. First we reduce \( M^2 \) to the real skew-Hamiltonian Schur form. Second we compute \( Y \), and finally we form the first \( n \) columns of \( M + Y \) to compute \( X \).

**ALGORITHM**

*Step 1:*

a. Form \( M^2 \).

b. Reduce \( M^2 \) to the real skew-Hamiltonian Hessenberg form \( \tilde{V}^T M^2 \tilde{V} \) with Householder symplectic matrices \( G_i, H_i \) and Givens symplectic matrices \( J_i \). See [1] for details.

c. Reduce \( \tilde{V}^T M^2 \tilde{V} \) to the real skew-Hamiltonian Schur form

\[
\tilde{V}^T M^2 \tilde{V} = \begin{bmatrix} T & H \\ 0 & T^T \end{bmatrix}
\]

with the QR algorithm, where \( \tilde{V} = \text{diag}[Q, Q] \), and \( Q \) is orthogonal. Store \( Q \).

*Step 2:*

a. Compute

\[
Y_1 = \begin{bmatrix} W & K \\ 0 & W^T \end{bmatrix},
\]

with the methods presented in [14, 5] according to Equations (4.2) and (4.3).

b. Form

\[
Y_2 = \text{diag}[Q, Q] Y_1 \text{diag}[Q^T, Q^T] = \begin{bmatrix} Q W Q^T & Q K Q^T \\ 0 & Q W^T Q^T \end{bmatrix}.
\]

Note that \( Y_2 \) is skew-Hamiltonian; we need only form \( Q W Q^T \) and the upper triangular part of \( Q K Q^T \).
c. Form

\[ Y = \hat{V} Y_2 \hat{V}^T = (H_1 J_1^T G_1 \cdots (H_{n-2} J_{n-2}^T G_{n-2}) J_{n-1}^T Y_2 J_{n-1} \times (G_{n-2} J_{n-2} H_{n-2}) \cdots (G_1 J_1 H_1). \]

Because \( G_i, H_i, J_i \) are symplectic orthogonal, in each step \( Y \) keeps the skew-Hamiltonian form, and we can decrease our work by making use of the skew-Hamiltonian property.

**Step 3:** Let

\[ Y = \begin{bmatrix} Y_{11} & Y_{12} \\ Y_{21} & Y_{22} \end{bmatrix}, \]

where \( Y_{22} = Y_{11}^T, \ Y_{12} = Y_{12}^T, \ Y_{21} = -Y_{21}^T \). Form \( A + Y_{11} \) and \( G + Y_{21} \). Solve the matrix equation \( X(A + Y_{11}) = -(G + Y_{21}) \) for \( X \).

The algorithm needs about \( 5n^2 \) storage locations.

Operation counts:

<table>
<thead>
<tr>
<th>Step</th>
<th>Description</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Form ( M^2 ):</td>
<td>( 4n^3 ) flops</td>
</tr>
<tr>
<td></td>
<td>Reduce to Schur form, form ( Q ):</td>
<td>( 20 \frac{1}{6} n^3 ) flops</td>
</tr>
<tr>
<td>2</td>
<td>Form ( Y_1 ):</td>
<td>( \frac{2}{3} n^3 ) flops</td>
</tr>
<tr>
<td></td>
<td>Form ( Y_2 ):</td>
<td>( 3 \frac{1}{2} n^3 ) flops</td>
</tr>
<tr>
<td></td>
<td>Form ( Y ):</td>
<td>( 8 n^3 ) flops</td>
</tr>
<tr>
<td>3</td>
<td>Solve ( X ):</td>
<td>( 2 \frac{1}{6} n^3 ) flops</td>
</tr>
<tr>
<td></td>
<td><strong>Total</strong></td>
<td>( 38 \frac{1}{2} n^3 ) flops</td>
</tr>
</tbody>
</table>

We make full use of the skew-Hamiltonian structure of \( Y \) in our algorithm; thus it also can be considered as a structure-preserving method. The only operation which causes instability is in the first part of step 2. We guess this may be related to the robustness of the original control system, but we can not give exact error analyses at present.

6. NUMERICAL TESTS

In this section we test some examples to illustrate the results of our algorithms. All examples are from [3] and done on a personal computer 286 with double-precision arithmetic and FORTRAN 77. Only the numerical results are displayed.
EXAMPLE 1[3]. The algorithm is excellent for solving this kind of low-order equation. The computed closed-loop spectrum is

\[-1.0000000000000000, \quad -1.0000000000000000,\]

and

\[X = \begin{bmatrix}
2.0000000000000000 & 1.0000000000000000 \\
1.0000000000000000 & 2.0000000000000000
\end{bmatrix}.\]

EXAMPLE 2[3]. The closed loop spectrum is

\[-0.14142135623731 \pm i, \quad -0.50000000000000 D + 00,\]

and

\[X = \begin{bmatrix}
0.21727922061358 D + 02 & 0.14485281374239 D + 02 \\
0.14485281374239 D + 02 & 0.96568542494924 D + 01
\end{bmatrix}.\]

EXAMPLE 3 (Example 4 in [3]). We solve this equation with 5 and 10 vehicles respectively. When \(N = 5\), the solution of the equation has at least six significant figures in agreement with the result of [3]. The diagonal elements of \(X\) are

\[0.13630206938090 D + 01, \quad 0.75925521954647 D + 01, \quad 0.17747816031505 D + 01,\]
\[0.82576995026609 D + 01, \quad 0.18056048615320 D + 01, \quad 0.82576995026609 D + 01,\]
\[0.17747816031505 D + 01, \quad 0.75925521954647 D + 01, \quad 0.13630206938090 D + 01,\]

while the closed-loop eigenvalues are

\[-0.14521501893058 D + 01 \pm i, \quad 0.12683612152305 D + 01,\]
\[-0.11077894826745 D + 01 \pm i, \quad 0.85275878061986 D + 00,\]
\[-0.10000000000000 D + 01, \quad 0.15193210220386 D + 01,\]
\[-0.16758091681359 D + 01 \pm i, \quad 0.18048558876092 D + 01 \pm i, \quad 0.16605736283097 D + 01;\]
when $N = 10$, we also have for the first row (column)

\[
\begin{align*}
0.14082559065178 \times 10^1 & \\
0.26676190896796 \times 10^1 & \\
-0.65821877192134 \times 10^0 & \\
0.10403124013015 \times 10^1 & \\
-0.24213305071508 \times 10^0 & \\
0.10345269956756 \times 10^0 & \\
-0.47208569311629 \times 10^{-1} & \\
0.50403617291430 \times 10^{-1} & \\
-0.45235212708864 \times 10^{-1} & \\
\end{align*}
\]

and for the fastest and slowest closed loop eigenvalues

\[
\begin{align*}
-0.18366667565655 \times 10^1 \pm i 0.16950943261874 \times 10^1, \\
-0.86295377285062 \times 10^0 \pm i 0.49466070601688 \times 10^0.
\end{align*}
\]

**EXAMPLE 4** (Example 5 of [3]). Our algorithm has an advantage in solving this kind of problem because for this equation the skew-Hamiltonian matrix $M^2$ is block-diagonal thus the square-root matrix $Y$ can be got more easily than usual. We do it with the orders $n = 5, 10, 20$. When $n = 5$, the solution has at least 14 significant figures, when $n = 10$ it has at least 13 significant figures, and when $n = 20$ it has at least 11 significant figures.

We also computed the very ill-conditioned problem of Example 6 in [3] with $n = 21, q = r = 1$. Unfortunately the algorithm fails. Far from the correct result $x_{1n} = 1$, the computed one is $0.17738736819855 \times 10^{-1}$.

7. CONCLUSION

We have discussed some properties of the quadratic matrix equation, then used them to set up theoretical results on the ARE, and finally constructed a structure-preserving algorithm. We have given a new way to reveal the property of a special invariant subspace of a Hamiltonian matrix.
We also note that the theory in Section 2 can be applied to the discrete-time algebraic Riccati equation with similar methods. Unfortunately the algorithm is not always numerically stable.

We would like to thank professor Jiang Er-xiong for his guidance, and my colleagues for their helpful discussions. We also thank Professor L. Elsner and the referee for providing many helpful suggestions for this paper.

REFERENCES

8 Er-xiong Jiang, Numerical Computations in Control Theory, teaching material.
9 Lin-zhang Lu, A new squared Hamiltonian reduced approach to the solution of the real algebraic Riccati equation, unpublished.

Received 29 January 1993; final manuscript accepted 11 October 1993