MATHEMATICS OF COMPUTATION
Volume 67, Number 223, July 1998, Pages 1089–1105
S 0025-5718(98)00947-8

ANALYSIS AND MODIFICATON OF NEWTON'S METHOD FOR ALGEBRAIC RICCATI EQUATIONS

CHUN-HUA GUO AND PETER LANCASTER

ABSTRACT. When Newton's method is applied to find the maximal symmetric solution of an algebraic Riccati equation, convergence can be guaranteed under moderate conditions. In particular, the initial guess need not be close to the solution. The convergence is quadratic if the Fréchet derivative is invertible at the solution. In this paper we examine the behaviour of the Newton iteration when the derivative is not invertible at the solution. We find that a simple modification can improve the performance of the Newton iteration dramatically.

1. INTRODUCTION AND REVIEW

Algebraic Riccati equations occur in many important applications [14], [16]. In this paper we consider algebraic Riccati equations of the form

$$\mathcal{R}(X) = XDX - XA - A^T X - C = 0, \tag{1.1}$$

where $A, D, C \in \mathbb{R}^{n\times n}$, and $D^T = D$, $C^T = C$. Let $\mathcal{S}$ be the set of symmetric matrices in $\mathbb{R}^{n\times n}$. For any matrix norm (not necessarily multiplicative) $\mathcal{S}$ is a Banach space, and $\mathcal{R}$ is a mapping from $\mathcal{S}$ into itself. The first Fréchet derivative of $\mathcal{R}$ at a matrix X is a linear map $\mathcal{R}'_X : \mathcal{S} \to \mathcal{S}$ given by

$$\mathcal{R}'_X(S) = -\big(S(A - DX) + (A - DX)^T S\big). \tag{1.2}$$

Also the second derivative at X, $\mathcal{R}''_X : \mathcal{S} \times \mathcal{S} \to \mathcal{S}$, is given by

$$\mathcal{R}''_X(S_1, S_2) = S_1 D S_2 + S_2 D S_1. \tag{1.3}$$

The Newton method for the solution of (1.1) is

$$X_{i+1} = X_i - (\mathcal{R}'_{X_i})^{-1}\mathcal{R}(X_i), \qquad i = 0, 1, \ldots, \tag{1.4}$$

given that the maps $\mathcal{R}'_{X_i}$ are all invertible. In view of (1.2), the iteration (1.4) is equivalent to

$$X_{i+1}(A - DX_i) + (A - DX_i)^T X_{i+1} = -X_i D X_i - C, \qquad i = 0, 1, \ldots. \tag{1.5}$$

It is readily seen that all the matrices X_i are symmetric if X_0 is so.

For $A, B \in \mathbb{R}^{n\times n}$, the pair (A, B) is said to be stabilizable if there is a $K \in \mathbb{R}^{n\times n}$ such that $A - BK$ is stable, i.e., all its eigenvalues are in the open left half-plane. The order relation on the set of symmetric matrices is the usual one: $X \geq Y$ if

Received by the editor February 18, 1997.
1991 *Mathematics Subject Classification.* Primary 65H10; Secondary 15A24, 93B40.
Research supported in part by a grant from the Natural Sciences and Engineering Research Council of Canada.

$X - Y$ is positive semidefinite. A symmetric solution X_+ of (1.1) is called maximal if $X_+ \geq X$ for every symmetric solution X. The following result is the real version of Theorem 9.1.1 in [14]. See also [4] and [8].

Theorem 1.1. *Assume that $D \geq 0$, $C^T = C$, (A, D) is stabilizable, and there exists a symmetric solution of the inequality $\mathcal{R}(X) \leq 0$. Then there exists a maximal symmetric solution X_+ of $\mathcal{R}(X) \doteq 0$. Moreover, all the eigenvalues of $A - DX_+$ are in the closed left half-plane.*

A symmetric solution X of (1.1) is called stabilizing (resp. almost stabilizing) if all the eigenvalues of $A - DX$ are in the open (resp. closed) left half-plane. Such solutions play important roles in applications. Theorem 1.1 tells us that, under the given conditions, the maximal solution is at least almost stabilizing. In fact (see [20] or [14, Theorem 7.9.3]), X_+ is the only symmetric solution that can be almost stabilizing. For this reason, the maximal solution is of particular interest.

Theorem 1.2. *Under the same conditions as in Theorem* 1.1, *starting with any symmetric matrix X_0 for which $A - DX_0$ is stable, the recursion* (1.5) *determines a sequence of symmetric matrices $\{X_i\}_{i=1}^{\infty}$ for which $A - DX_i$ is stable for $i = 1, 2, \ldots$, $X_1 \geq X_2 \geq \cdots$, and $\lim_{i\to\infty} X_i = X_+$.*

The maximal solution can thus be found by the Newton iteration without previous information about the solution. The proof of the above theorem can be found in [14, p. 232]. See also [4], [8] and [13]. There is no doubt about the existence of the matrix X_0. Since (A, D) is stabilizable and $D \geq 0$, we can find an $X_0 \geq 0$ such that $A - DX_0$ is stable. This is the real version of Lemma 4.5.4 in [14]. Moreover, a stabilizing symmetric matrix X_0 can be produced by automatic stabilizing procedures such as the one in [19], although the matrix X_0 so obtained may be far away from the solution X_+. We note that $X_0 \geq X_1$ is generally not true. In fact, the first Newton iteration is capable of making a big adjustment to the initial guess (see [2], for example). When $X_+ \geq 0$, we necessarily have $X_1 \geq 0$. But X_0 can be indefinite.

If X is an almost stabilizing solution of (1.1) (in the sense that $\sigma(A - DX)$ is in the *closed* left half-plane), then $\mathcal{R}'_X$ is invertible if and only if X is a stabilizing solution. This can be seen from the following classical result.

Theorem 1.3 (cf. [14, p. 100]). *For given matrices $A \in \mathbb{R}^{m\times m}, B \in \mathbb{R}^{n\times n}$ and $\Gamma \in \mathbb{R}^{n\times m}$ the Sylvester equation $SA - BS = \Gamma$ has a unique solution (necessarily real) if and only if A and B have no eigenvalues in common.*

It is readily seen that $\mathcal{R}'_X$, as a function of X, is Lipschitz continuous on $\mathcal{S}$. Thus the well known locally quadratic convergence of Newton's method [10], [17], in combination with Theorem 1.2, yields the following result.

Theorem 1.4. *If $A - DX_+$ is stable in Theorem* 1.1, *then for the sequence $\{X_i\}_{i=0}^{\infty}$ there is a constant $c > 0$ such that, for $i = 0, 1, \ldots$, $\|X_{i+1} - X_+\| \leq c\|X_i - X_+\|^2$, where $\|\cdot\|$ is any given matrix norm.*

We note that, because $\sigma(A - DX_+)$ is in the *open* left half-plane, $A - DX_0$ is necessarily stable if X_0 is close enough to X_+. A direct algebraic proof of the above theorem can be found in [14, p. 237].

In [2], an exact line search method is introduced which improves Newton's method for the numerical solution of the Riccati equation in several aspects. However, the theory established there does not cover the general situation described in Theorem 1.1, even when $A - DX_+$ has no eigenvalues on the imaginary axis.

When $A - DX_+$ has eigenvalues on the imaginary axis, $\mathcal{R}'_{X_+}$ is not invertible and the convergence behaviour of the Newton iteration is more complicated. In this paper we examine the behaviour of the Newton iteration for this case. The results we obtain suggest that a simple modification step can be introduced to improve the performance of the Newton iteration dramatically in many cases. Numerical results are also given to show the effectiveness of the modification.

The literature on Newton's method in the case of a non-invertible Jacobian at the solution ($\mathcal{R}'_{X_+}$ in our case) is considerable. Typically, one considers a smooth map F from a Banach space E into itself (see [5], [6], [7], [11], [12], [18]). Standard assumptions are that there is an $x^* \in E$ such that $F(x^*) = 0$ and the Fréchet derivative at x^*, $F'(x^*)$, has a null space N of dimension d with $0 < d < \infty$. Also, it is assumed that $F'(x^*)$ has closed range M and that there is a direct sum decomposition $E = N \oplus M$. Then we may define P_N to be the projection onto N parallel to M and let $P_M = I - P_N$. Assume that there is a $\phi_0 \in N$ such that the map B on N given by $B = P_N F''(x^*)(\phi_0, \cdot)$ is invertible. Linear convergence with common ratio $\frac{1}{2}$ is then predicted for Newton's method with an appropriate initial guess. The investigations of this paper began with Example 9.2.1 of [14] concerning Riccati equations in which this same constant appears.

These analyses establish *local* convergence results of course (in contrast with Theorem 1.2). Our main result will be an application of the following theorem.

Theorem 1.5 (cf. [11, Theorem 1.1]). *Let $E = N \oplus M$, let ϕ_0 be chosen so that B is invertible, and let $N = \operatorname{span}\{\phi_0\} \oplus N_1$ for some subspace N_1. Write $\tilde{x} = x - x^*$ and let*

(1.6)
$$W(\rho, \theta, \eta) = \{x \mid 0 < \|\tilde{x}\| < \rho, \|P_M \tilde{x}\| \le \theta \|P_N \tilde{x}\|, \|(P_N - P_0)\tilde{x}\| \le \eta \|P_N \tilde{x}\|\},$$

where P_0 is the projection onto $\operatorname{span}\{\phi_0\}$ parallel to $M \oplus N_1$. If $x_0 \in W(\rho_0, \theta_0, \eta_0)$ for ρ_0, θ_0, η_0 sufficiently small, then the Newton sequence $\{x_i\}$ is well defined and $\|F'(x_i)^{-1}\| \le c\|\tilde{x}_i\|^{-1}$ for all $i \ge 1$ and some constant $c > 0$. Moreover,

$$\lim_{i \to \infty} \frac{\|\tilde{x}_{i+1}\|}{\|\tilde{x}_i\|} = \frac{1}{2}, \quad \lim_{i \to \infty} \frac{\|P_M \tilde{x}_i\|}{\|P_N \tilde{x}_i\|^2} = 0.$$

Notice that the region $W(\rho, \theta, \eta)$ in which x_0 is required to lie is close to x^*, N, and ϕ_0 in the sense determined by the ρ, θ, η inequalities, respectively.

2. Further analysis of Newton's method applied to the Riccati equation

We now go back to the discussion of algebraic Riccati equations and assume throughout that the conditions of Theorem 1.1 are satisfied. Let X_+ be the maximal solution of (1.1) with $\mathcal{R}'_{X_+}$ not invertible. Let $\mathcal{N} = \operatorname{Ker} \mathcal{R}'_{X_+}$, $\mathcal{M} = \operatorname{Im} \mathcal{R}'_{X_+}$. We have the following characterization of the direct sum condition.

Theorem 2.1. *$\mathcal{S} = \mathcal{N} \oplus \mathcal{M}$ if and only if all elementary divisors of $A - DX_+$ corresponding to the eigenvalues on the imaginary axis are linear.*

Proof. Let J be the real Jordan canonical form for $A-DX_+$ with $P^{-1}(A-DX_+)P = J$ and a real matrix P. We find that $K \in \operatorname{Ker} \mathcal{R}'_{X_+}$ if and only if $K = P^{-T}QP^{-1}$ for some $Q \in \mathcal{S}$ satisfying $QJ + J^TQ = 0$. Also $W \in \operatorname{Im} \mathcal{R}'_{X_+}$ if and only if $W = P^{-T}RP^{-1}$ with $R = VJ + J^TV$ for some $V \in \mathcal{S}$. Therefore, we may assume without loss of generality that $A - DX_+$ is in real Jordan canonical form.

If all elementary divisors of $A - DX_+$ corresponding to the eigenvalues on the imaginary axis are linear, we gather the Jordan blocks of $A-DX_+$ in several groups:

$$A - DX_+ = \operatorname{diag}(G_1, G_2, \dots, G_{p-1}, G_p).$$

Here $G_1 = 0$, G_p consists of real Jordan blocks associated with eigenvalues in the open left half-plane, and for $i = 2, \dots, p-1$,

$$G_i = \operatorname{diag}\left(\begin{pmatrix} 0 & a_i \\ -a_i & 0 \end{pmatrix}, \dots, \begin{pmatrix} 0 & a_i \\ -a_i & 0 \end{pmatrix}\right),$$

where the a_i's are distinct positive numbers. Using block matrix multiplications and applying Theorem 1.3 repeatedly, we can show that $\mathcal{S} = \mathcal{N} \oplus \mathcal{M}$.

If $A - DX_+$ has nonlinear elementary divisors corresponding to eigenvalues on the imaginary axis, we can arrange the Jordan blocks so that the first Jordan block J_1 has one of the following two forms:

$$(1) \quad J_1 = \begin{pmatrix} 0 & 1 & & \\ & 0 & \ddots & \\ & & \ddots & 1 \\ & & & 0 \end{pmatrix}.$$

$$(2) \quad J_1 = \begin{pmatrix} B & I & & \\ & B & \ddots & \\ & & \ddots & I \\ & & & B \end{pmatrix}, \quad B = \begin{pmatrix} 0 & a \\ -a & 0 \end{pmatrix}, \ a \neq 0.$$

For the first case, $\operatorname{diag}(0, \dots, 0, 1, 0, \dots, 0) \in \mathcal{N} \cap \mathcal{M}$, where the element 1 appears at the same position as the last diagonal element of J_1. For the second case, $\operatorname{diag}(0, \dots, 0, I, 0, \dots, 0) \in \mathcal{N} \cap \mathcal{M}$, where the 2×2 identity matrix I appears at the same position as the last diagonal block of J_1. Therefore, $\mathcal{S} \neq \mathcal{N} \oplus \mathcal{M}$. □

When $\mathcal{S} = \mathcal{N} \oplus \mathcal{M}$, we let $P_{\mathcal{N}}$ denote the projection onto $\mathcal{N}$ parallel to $\mathcal{M}$ and let $P_{\mathcal{M}} = I - P_{\mathcal{N}}$. For the algebraic Riccati equation, we start the Newton iteration with a symmetric matrix X_0 for which $A - DX_0$ is stable. Although the Newton sequence is well-defined and converges to X_+, we do not know whether the iterates X_i will finally fall into a special region of the form (1.6). Therefore Theorem 1.5 cannot be applied directly. Instead, we have the following result.

Theorem 2.2. *For any fixed $\theta > 0$, let*

$$Q = \{i \mid \|P_{\mathcal{M}}(X_i - X_+)\| > \theta \|P_{\mathcal{N}}(X_i - X_+)\|\}.$$

Then there exist an integer i_0 and a constant $c > 0$ such that $\|X_i - X_+\| \leq c\|X_{i-1} - X_+\|^2$ for all i in Q for which $i \geq i_0$.

Proof. Let $\tilde{X}_i = X_i - X_+$. Using Taylor's Theorem with (1.3) and the fact that $\mathcal{R}'_{X_+}(P_{\mathcal{N}}\tilde{X}_i) = 0$,

$$\mathcal{R}(X_i) = \mathcal{R}(X_+) + \mathcal{R}'_{X_+}(\tilde{X}_i) + \frac{1}{2}\mathcal{R}''_{X_+}(\tilde{X}_i, \tilde{X}_i) = \mathcal{R}'_{X_+}(P_{\mathcal{M}}\tilde{X}_i) + \tilde{X}_i D \tilde{X}_i. \tag{2.1}$$

Since $\mathcal{R}'_{X_+}|_{\mathcal{M}} : \mathcal{M} \to \mathcal{M}$ is invertible, $\|\mathcal{R}'_{X_+}(P_{\mathcal{M}}\tilde{X}_i)\| \geq c_1 \|P_{\mathcal{M}}\tilde{X}_i\|$ for some constant $c_1 > 0$. For $i \in Q$, we have $\|\tilde{X}_i\| \leq \|P_{\mathcal{M}}\tilde{X}_i\| + \|P_{\mathcal{N}}\tilde{X}_i\| \leq \left(\theta^{-1} + 1\right) \|P_{\mathcal{M}}\tilde{X}_i\|$. Thus by (2.1),

$$\|\mathcal{R}(X_i)\| \geq c_1 \|P_{\mathcal{M}}\tilde{X}_i\| - c_2 \|\tilde{X}_i\|^2 \geq \left(c_1(\theta^{-1} + 1)^{-1} - c_2\|\tilde{X}_i\|\right)\|\tilde{X}_i\|. \tag{2.2}$$

On the other hand, we have by (1.5)

$$X_i(A - DX_{i-1}) + (A - DX_{i-1})^T X_i = -X_{i-1} D X_{i-1} - C,$$

and obviously,

$$X_+(A - DX_+) + (A - DX_+)^T X_+ = -X_+ D X_+ - C.$$

By subtraction, we obtain after some manipulations

$$\tilde{X}_i(A - DX_{i-1}) + (A - DX_{i-1})^T \tilde{X}_i = -\tilde{X}_{i-1} D \tilde{X}_{i-1}.$$

Writing $X_+ = X_{i-1} - \tilde{X}_{i-1}$ in (2.1) and using the last equation it is found that

$$\begin{aligned} \mathcal{R}(X_i) &= -\tilde{X}_i\big((A - DX_{i-1}) + D\tilde{X}_{i-1}\big) \\ &\quad - \big((A - DX_{i-1}) + D\tilde{X}_{i-1}\big)^T \tilde{X}_i + \tilde{X}_i D \tilde{X}_i \\ &= \tilde{X}_{i-1} D \tilde{X}_{i-1} - \tilde{X}_i D \tilde{X}_{i-1} - \tilde{X}_{i-1} D \tilde{X}_i + \tilde{X}_i D \tilde{X}_i. \end{aligned}$$

Since $\|\cdot\|$ is equivalent to a multiplicative matrix norm on $\mathbb{R}^{n\times n}$, we have

$$\|\mathcal{R}(X_i)\| \leq c_3 \|\tilde{X}_i\|^2 + c_4 \|\tilde{X}_i\| \|\tilde{X}_{i-1}\| + c_5 \|\tilde{X}_{i-1}\|^2. \tag{2.3}$$

In view of (2.2) and the fact that $X_i \neq X_+$ for any i, we have

$$c_1(\theta^{-1} + 1)^{-1} - c_2\|\tilde{X}_i\| \leq c_3\|\tilde{X}_i\| + c_4\|\tilde{X}_{i-1}\| + c_5\|\tilde{X}_{i-1}\|^2/\|\tilde{X}_i\|.$$

Since $\tilde{X}_i \to 0$ by Theorem 1.2, we can find an i_0 such that $\|\tilde{X}_i\| \leq c\|\tilde{X}_{i-1}\|^2$ for all $i \geq i_0$. □

Corollary 2.3. *Assume that, for given $\theta > 0$, $\|P_{\mathcal{M}}(X_i - X_+)\| > \theta \|P_{\mathcal{N}}(X_i - X_+)\|$ for all i large enough. Then $X_i \to X_+$ quadratically.*

The above result is somewhat surprising, since it is generally believed that linear convergence is the best we can expect when the derivative at the solution is not invertible (see [5], [7] and [12]). We cannot rejoice in the possibility of quadratic convergence, however, since the condition in the corollary is not easily satisfied. Nevertheless, we can conclude that, when the convergence is not quadratic, *the error will generally be dominated by its $\mathcal{N}$-component.* This will be the basis for a numerical strategy proposed in the next section. Meanwhile, the following theorem shows what happens in the generic case when convergence is not quadratic.

Theorem 2.4. *Assume $\mathcal{S} = \mathcal{N} \oplus \mathcal{M}$. If the convergence of the Newton sequence $\{X_i\}$ is not quadratic, then $\|(\mathcal{R}'_{X_i})^{-1}\| \leq c\|X_i - X_+\|^{-1}$ for all $i \geq 1$ and some constant $c > 0$. Moreover,*

$$\lim_{i\to\infty} \frac{\|X_{i+1} - X_+\|}{\|X_i - X_+\|} = \frac{1}{2}, \quad \lim_{i\to\infty} \frac{\|P_{\mathcal{M}}(X_i - X_+)\|}{\|P_{\mathcal{N}}(X_i - X_+)\|^2} = 0.$$

The proof of this theorem is an application of Theorem 1.5 and follows readily from the next lemma. The map B appearing in Theorem 1.5, when applied to the Riccati equation (at a fixed $Z \in \mathcal{N}$ instead of ϕ_0), takes the form

$$\mathcal{B}_Z = P_{\mathcal{N}} \mathcal{R}'_{X_+}(Z, \cdot) : \mathcal{N} \to \mathcal{N}. \tag{2.4}$$

Lemma 2.5. *If $\mathcal{S} = \mathcal{N} \oplus \mathcal{M}$, then.*

$$\mathcal{U} = \{Z \in \mathcal{N} \mid \mathcal{B}_Z : \mathcal{N} \to \mathcal{N} \text{ is not invertible}\}$$

has measure zero in $\mathcal{N}$.

The proof of this lemma is rather long and technical and is presented in the appendix. However, it has its own interest, and includes an explicit construction of the spaces $\mathcal{N}$ and $\mathcal{M}$.

Proof of Theorem 2.4. We apply Theorem 1.5, with some natural changes of notation. Let $\tilde{X}_i = X_i - X_+$ and $\tilde{X} = X - X_+$. We are to show that there is a Φ_0 such that $\mathcal{B}_{\Phi_0}$ is invertible and, if $\mathcal{N} = \operatorname{span}\{\Phi_0\} \oplus \mathcal{N}_1$ and P_0 is the projection on Φ_0 along $\mathcal{N}_1 \oplus \mathcal{M}$, then there is an i such that $X_i \in \mathcal{W}(\rho_0, \theta_0, \eta_0)$ where

$$\begin{aligned} \mathcal{W}(\rho_0, \theta_0, \eta_0) = \{X \mid 0 < \|\tilde{X}\| < \rho_0, \|P_{\mathcal{M}} \tilde{X}\| \le \theta_0 \|P_{\mathcal{N}} \tilde{X}\|, \\ \|(P_{\mathcal{N}} - P_0)\tilde{X}\| \le \eta_0 \|P_{\mathcal{N}} \tilde{X}\|\}. \end{aligned} \tag{2.5}$$

First, Theorem 1.2 shows that by choosing X_0 so that $A - DX_0$ is stable there is an i_1 such that $0 < \|\tilde{X}_i\| < \rho_0$ for all $i \ge i_1$. Then, since the convergence of the Newton sequence is not quadratic it follows from Corollary 2.3 that $\|P_{\mathcal{M}} \tilde{X}_{i_2}\| \le \theta_0 \|P_{\mathcal{N}} \tilde{X}_{i_2}\|$ for some $i_2 \ge i_1$. Note that $P_{\mathcal{N}} \tilde{X}_{i_2} \ne 0$, since otherwise we would have $\tilde{X}_{i_2} = 0$.

Finally, if we choose $\Phi_0 = P_{\mathcal{N}} \tilde{X}_{i_2}$, then the last inequality of (2.5) is trivially satisfied for X_{i_2}, but $\mathcal{B}_{\Phi_0}$ may not be invertible. However, when η_0 is given, it follows from Lemma 2.5 that a Φ_0 can be chosen arbitrarily close to $P_{\mathcal{N}} \tilde{X}_{i_2}$ in such a way that $\mathcal{B}_{\Phi_0}$ is invertible and $X_{i_2} \in \mathcal{W}(\rho_0, \theta_0, \eta_0)$. Now apply Theorem 1.5. □

In H_∞-control problems, we will indeed encounter algebraic Riccati equations where $A - DX_+$ has eigenvalues on the imaginary axis (see [15], for example). The direct sum condition $\mathcal{S} = \mathcal{N} \oplus \mathcal{M}$ is usually satisfied there. On the other hand, it is always possible to find examples where the eigenvalues on the imaginary axis have elementary divisors of arbitrary degree. It suffices to consider (1.1) with

$$D = \begin{pmatrix} 1 & & & \\ & 0 & & \\ & & \ddots & \\ & & & 0 \end{pmatrix}, \quad A = \begin{pmatrix} 0 & & & \\ 1 & 0 & & \\ & \ddots & \ddots & \\ & & 1 & 0 \end{pmatrix}, \quad C = 0. \tag{2.6}$$

For $X = (x_{ij})$ with $x_{1j} = C_n^j$ (the binomial coefficients), all the eigenvalues of $A - DX$ are -1. Thus (A, D) is stabilizable. We also have $X_+ = 0$ (0 is an almost stabilizing solution, and thus maximal), and λ^n is the only elementary divisor of $A - DX_+$. Numerical experiments suggest that the Newton sequence converges to 0 linearly with common ratio $2^{-1/n}$. When $n = 2$, we can exhibit the linear convergence with common ratio $1/\sqrt{2}$ by the following example.

Example 2.1. Consider the Riccati equation (1.1) with D, A, C given by (2.6) and $n = 2$. For Newton's method, the symmetric matrices $X_{k+1} = (x_{ij}^{k+1})$ and $X_k = (x_{ij}^k)$ are now related by

$$x_{11}^{k+1} = \frac{1}{2}x_{11}^k + \frac{1}{2}\frac{x_{12}^k}{x_{11}^k}, \quad x_{12}^{k+1} = \frac{1}{2}x_{12}^k, \quad x_{22}^{k+1} = \frac{1}{2}\frac{(x_{12}^k)^2}{x_{11}^k}.$$

We choose $x_{11}^0 = \sqrt{(\sqrt{2}+1)c}$, $x_{12}^0 = c$ for any $c > 0$, and x_{22}^0 can be arbitrary. Then $A - DX_0$ is stable, and we find that for $k = 1, 2, \ldots$

$$x_{11}^k = \left(\frac{1}{\sqrt{2}}\right)^k \sqrt{(\sqrt{2}+1)c}, \quad x_{12}^k = \left(\frac{1}{2}\right)^k c, \quad x_{22}^k = \left(\frac{1}{2\sqrt{2}}\right)^k \frac{\sqrt{2}c^2}{\sqrt{(\sqrt{2}+1)c}}.$$

Thus $\lim_{k\to\infty} \|X_{k+1}\|/\|X_k\| = 1/\sqrt{2}$ for any matrix norm.

The role of the eigenvalues of $A - DX_+$ on the imaginary axis is clearly critical. Information about these eigenvalues can also be obtained from the Hamiltonian matrix

$$H = \begin{pmatrix} -A & D \\ C & A^T \end{pmatrix},$$

associated with the Riccati equation (1.1). Thus:

Theorem 2.6. *The complex number λ is an eigenvalue of $A - DX_+$ on the imaginary axis if and only if λ is an eigenvalue of H on the imaginary axis. Moreover, the partial multiplicities (i.e. the degrees of elementary divisors) of λ as an eigenvalue of H are twice the partial multiplicities of λ as an eigenvalue of $A - DX_+$.*

Proof. This is an immediate consequence of [14, Theorem 7.3.1], with some changes in notation. The condition (7.3.1) in that theorem is satisfied because (A, D) is stabilizable. □

3. A modified Newton method

The Newton iteration can be used to find the maximal solution of (1.1) when the Hamiltonian matrix has eigenvalues on the imaginary axis, while most other algorithms are not applicable in this case (see [16]). However, the convergence of the Newton sequence in this case is usually linear although, as Theorem 2.2 suggests, we have not excluded the possibility of quadratic convergence. Since the Newton iteration is an expensive procedure, we cannot be satisfied with linear convergence alone.

For the general case described in Section 1, much work has been done on modifications of the Newton iteration with a view to accelerating convergence when the Jacobian is not invertible at the solution. See, for example, [5], [7] and [12]. The modified methods as described in [5] and [7] are, however, not applicable for the Riccati equations. Motivated by consideration of quadratic problems, Kelley and Suresh [12] proposed other modified methods which could be applied to the Riccati equations. Again, the initial guess must be in a special region of the form (1.6) in order that their modified methods are well-defined and give fast convergence. When we apply the Newton iteration to find the maximal symmetric solution of (1.1), we start with a symmetric matrix X_0 for which $A - DX_0$ is stable. It is not

clear whether and when the iterate X_k will fall into that special region. We therefore take a different approach. We are not going to recover quadratic convergence. Instead we will add a simple modification step to the Newton iteration so that the required accuracy can be achieved at an early stage. The following simple result is very instructive. Note statement 2, especially, and the possibility that it presents for stepping directly to the solution X_+.

Theorem 3.1. *In the setting of Theorems* 1.1 *and* 1.2, *and under the condition that* $X_k - X_+ \in \mathcal{N}$, *we have*

1. $X_{k+1} - X_+ = \frac{1}{2}(X_k - X_+)$.
2. $X_+ = X_k - 2(\mathcal{R}'_{X_k})^{-1}\mathcal{R}(X_k)$.

Proof. By Taylor's Theorem,

$$\mathcal{R}'_{X_k}(X_k - X_+) = \mathcal{R}'_{X_+}(X_k - X_+) + \mathcal{R}''_{X_+}(X_k - X_+, X_k - X_+).$$

Since $\mathcal{R}(X_+) = 0$ and $\mathcal{R}'_{X_+}(X_k - X_+) = 0$, we may also write

$$\begin{aligned}\mathcal{R}'_{X_k}(X_k - X_+) &= 2\{\mathcal{R}(X_+) + \mathcal{R}'_{X_+}(X_k - X_+) + \frac{1}{2}\mathcal{R}''_{X_+}(X_k - X_+, X_k - X_+)\} \\ &= 2\mathcal{R}(X_k).\end{aligned}$$

The second part of the theorem follows immediately. The first part follows easily from (1.4) and the second part. □

We remark that similar conclusions can be reached for any map F from a Banach space into itself, for which F'' is constant.

Example 3.1. It is instructive to revisit Example 9.2.1 of [14] at this stage. Let

$$D = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \quad A = \begin{pmatrix} 0 & 0 \\ 0 & -1 \end{pmatrix}, \quad C = \begin{pmatrix} 0 & 1 \\ 1 & 2 \end{pmatrix}.$$

It is easily verified that there is a unique solution of $\mathcal{R}(X) = 0$, namely,

$$X_+ = \begin{pmatrix} 0 & 1 \\ 1 & 1/2 \end{pmatrix}.$$

Thus,

$$A - DX_+ = \begin{pmatrix} 0 & -1 \\ 0 & -1 \end{pmatrix}$$

and is not stable.

If Newton iterations are started with $X_0 = \left(\begin{smallmatrix} 1 & 0 \\ 0 & 0 \end{smallmatrix}\right)$ (when $A - DX_0 = -I$ and is stable) then it can be proved by induction that, for $n = 1, 2, \ldots$,

$$X_n = \begin{pmatrix} 2^{-n} & 1 - 2^{-n} \\ 1 - 2^{-n} & \frac{1}{2} + 2^{-n} \end{pmatrix}.$$

Consequently, for $n = 1, 2, \ldots$,

$$\frac{\|X_{n+1} - X_+\|}{\|X_n - X_+\|} = \frac{1}{2}. \tag{3.1}$$

It can be seen that, in this case

$$\mathcal{N} = \operatorname{span}\left\{\begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}\right\}, \quad \mathcal{M} = \operatorname{span}\left\{\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}\right\}$$

so that $\mathcal{S} = \mathcal{N} \oplus \mathcal{M}$. Furthermore, $X_+ \in \mathcal{M}$ and, for $n = 1, 2, \ldots$,

$$X_n - X_+ = 2^{-n} \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}.$$

Fortuitously, X_0 is chosen in such a way that $X_n - X_+ \in \mathcal{N}$ for $n = 1, 2, \ldots$. Thus, Theorem 3.1 applies and (3.1) holds. Furthermore, it is clear that, by applying the modified Newton step at *any* $n \geq 1$, the exact solution X_+ is obtained.

The direct sum condition $\mathcal{S} = \mathcal{N} \oplus \mathcal{M}$ is not required in Theorem 3.1. However, the condition $X_k - X_+ \in \mathcal{N}$ can hardly be satisfied without the direct sum condition. In fact, for the Riccati equation considered in Example 2.1, $X_k - X_+ \in \mathcal{N}$ implies the non-invertibility of $\mathcal{R}'_{X_k}$. Thus for this simple example, starting with a symmetric matrix X_0 for which $A - DX_0$ is stable, we can never have $X_k - X_+ \in \mathcal{N}$.

When the direct sum condition is satisfied and the convergence of the Newton sequence $\{X_k\}$ is not quadratic, we have at some (hopefully early) stage $\|P_{\mathcal{M}}(X_k - X_+)\| \ll \|P_{\mathcal{N}}(X_k - X_+)\|$ (cf. Theorem 2.4). A very good approximate solution could then be obtained by applying the modification step in Theorem 3.1(2). More precisely, we have the following result.

Theorem 3.2. *Assume* $\mathcal{S} = \mathcal{N} \oplus \mathcal{M}$ *and* $\|(\mathcal{R}'_{X_i})^{-1}\| \leq c_1 \|X_i - X_+\|^{-1}$ *for all* $i \geq 1$. *If for some* k, $\|P_{\mathcal{M}}(X_k - X_+)\| \leq \epsilon \|P_{\mathcal{N}}(X_k - X_+)\|$ *with* ϵ *sufficiently small, and* $Y_{k+1} = X_k - 2(\mathcal{R}'_{X_k})^{-1}\mathcal{R}(X_k)$, *then* $\|Y_{k+1} - X_+\| \leq c\epsilon$ *for some constant* c *independent of* ϵ *and* k.

Proof. Let $\hat{X}_k = X_+ + P_{\mathcal{N}}(X_k - X_+)$. We have

$$\|X_k - \hat{X}_k\| = \|P_{\mathcal{M}}(X_k - X_+)\| \leq \epsilon \|P_{\mathcal{N}}(X_k - X_+)\| \leq c_2 \epsilon \|X_k - X_+\|,$$

and

$$\begin{aligned} \|I - (\mathcal{R}'_{X_k})^{-1}\mathcal{R}'_{\hat{X}_k}\| &\leq \|(\mathcal{R}'_{X_k})^{-1}\| \|\mathcal{R}'_{X_k} - \mathcal{R}'_{\hat{X}_k}\| \\ &\leq c_1 \|X_k - X_+\|^{-1} c_3 \|X_k - \hat{X}_k\| \\ &\leq c_4 \epsilon. \end{aligned}$$

If $c_4\epsilon < \frac{1}{2}$, we know from the Banach lemma that $\mathcal{R}'_{\hat{X}_k}$ is invertible and

$$\|(\mathcal{R}'_{\hat{X}_k})^{-1}\| \leq 2\|(\mathcal{R}'_{X_k})^{-1}\| \leq c_5 \|X_k - X_+\|^{-1}.$$

Since $\hat{X}_k - X_+ \in \mathcal{N}$, we have by Theorem 3.1

$$X_+ = \hat{X}_k - 2(\mathcal{R}'_{\hat{X}_k})^{-1}\mathcal{R}(\hat{X}_k).$$

Hence

$$\|Y_{k+1} - X_+\| \leq \|X_k - \hat{X}_k\| + 2\|(\mathcal{R}'_{\hat{X}_k})^{-1}\mathcal{R}(\hat{X}_k) - (\mathcal{R}'_{X_k})^{-1}\mathcal{R}(X_k)\|.$$

On writing

$$\begin{aligned} &(\mathcal{R}'_{\hat{X}_k})^{-1}\mathcal{R}(\hat{X}_k) - (\mathcal{R}'_{X_k})^{-1}\mathcal{R}(X_k) \\ &= (\mathcal{R}'_{X_k})^{-1} \left\{ (\mathcal{R}'_{X_k} - \mathcal{R}'_{\hat{X}_k})(\mathcal{R}'_{\hat{X}_k})^{-1}\big(\mathcal{R}(\hat{X}_k) - \mathcal{R}(X_+)\big) + \mathcal{R}(\hat{X}_k) - \mathcal{R}(X_k) \right\}, \end{aligned}$$

we obtain easily $\|Y_{k+1} - X_+\| \leq c\epsilon$.

The following algorithm is suggested by the results of this section.

Algorithm 3.3. Modified Newton method for algebraic Riccati equations:

1. Choose a symmetric matrix X_0 for which $A - DX_0$ is stable.
2. For $k = 0, 1, \ldots$ do:
   Solve $\mathcal{R}'_{X_k}(H) = \mathcal{R}(X_k)$;
   Compute $X_{k+1} = X_k - 2H$;
   If $\|\mathcal{R}(X_{k+1})\| < \epsilon$, stop;
   Otherwise, compute $X_{k+1} = X_k - H$;
   If $\|\mathcal{R}(X_{k+1})\| < \epsilon$, stop.

In the above algorithm, $\|\cdot\|$ is an easily computable matrix norm (e.g. 1-norm) and ϵ is a prescribed accuracy. The equation $\mathcal{R}'_{X_k}(H) = \mathcal{R}(X_k)$ can be rewritten as a Lyapunov equation $(A-DX_k)^T H+H(A-DX_k) = -\mathcal{R}(X_k)$, which can be solved efficiently by the algorithms described in [1] and [9]. In Algorithm 3.3, all iterates except the last one are identical to those produced by the original Newton method. Thus all good properties of the Newton method are retained. When $A - DX_+$ has eigenvalues on the imaginary axis, the last iterate is usually produced by the modified step. Algorithm 3.3 needs roughly 10% more computational work per iteration, since we systematically perform one additional Riccati function evaluation with a view to achieving the required accuracy in the modified step as early as possible.

4. Numerical examples

In this section we present some numerical examples to illustrate the effectiveness of the modified Newton step in Algorithm 3.3.

Example 4.1. Consider the algebraic Riccati equation (1.1) with $n = 2$ and

$$A = \begin{pmatrix} \epsilon+1 & 1 \\ 1 & \epsilon+1 \end{pmatrix}, \quad D = I_2, \quad C = \epsilon^2 I_2$$

(cf. Example 10 of [3]). The maximal solution $X_+ = (x_{ij})$ is given by

$$x_{11} = x_{22} = \frac{1}{2}\big(2(\epsilon+1) + \sqrt{2(\epsilon+1)^2+2} + \sqrt{2}\epsilon\big),$$
$$x_{12} = x_{21} = \frac{x_{11}}{x_{11} - (\epsilon+1)}.$$

For $\epsilon = 0$, the pair (A, D) is stabilizable, and

$$A - DX_+ = \begin{pmatrix} -1 & -1 \\ -1 & -1 \end{pmatrix}, \quad \mathcal{N} = \left\{ \begin{pmatrix} a & -a \\ -a & a \end{pmatrix}, \ a \in \mathbb{R} \right\}.$$

Observe that $\sigma(A - DX_+) = \{0, -2\}$. Starting with

$$X_0 = \begin{pmatrix} 18 & 16 \\ 16 & 18 \end{pmatrix},$$

we perform 8 steps of the ordinary Newton iteration and then perform a modification step. The results are recorded in Table 1. As usual we let $\tilde{X}_k = X_k - X_+$ and write $\tilde{X}_k = (\tilde{x}^k_{ij})$. For this problem the convergence of the Newton iteration is linear with common ratio $\frac{1}{2}$ (cf. Theorem 2.4). After 8 Newton iterations, X_8 is still not very close to X_+. However, $X_8 - X_+$ is very close to an element in $\mathcal{N}$. A modification step then produces a very accurate approximate solution (cf. Theorem 3.2).

TABLE 1. Performance of Algorithm 3.3 for Example 4.1

k	$\tilde{x}_{11}^k$	$\tilde{x}_{12}^k = \tilde{x}_{21}^k$	$\tilde{x}_{22}^k$	$\|\tilde{X}_k\|_1$
0	$0.1600D+02$	$0.1400D+02$	$0.1600D+02$	$0.3000D+02$
1	$0.7531D+01$	$0.6531D+01$	$0.7531D+01$	$0.1406D+02$
2	$0.3328D+01$	$0.2828D+01$	$0.3328D+01$	$0.6154D+01$
3	$0.1287D+01$	$0.1037D+01$	$0.1287D+01$	$0.2323D+01$
4	$0.3746D+00$	$0.2496D+00$	$0.3746D+00$	$0.6242D+00$
5	$0.6837D-01$	$0.5867D-02$	$0.6837D-01$	$0.7423D-01$
6	$0.1629D-01$	$-0.1496D-01$	$0.1629D-01$	$0.3125D-01$
7	$0.7813D-02$	$-0.7812D-02$	$0.7813D-02$	$0.1562D-01$
8	$0.3906D-02$	$-0.3906D-02$	$0.3906D-02$	$0.7812D-02$
9	$-0.3531D-13$	$-0.1354D-13$	$-0.3575D-13$	$0.4929D-13$

When ϵ is a small positive number, X_+ is a stabilizing solution. According to Theorem 1.4, the Newton sequence $\{X_k\}$ converges to X_+ quadratically. However, the constant c in Theorem 1.4 will be very large for very small ϵ. Thus the quadratic convergence could be exhibited only after X_k gets very close to the solution. On the other hand, as X_k gets close to the solution, the corresponding Lyapunov equation will be ill-conditioned. As a result, quadratic convergence can hardly be realized. For example, take $\epsilon = 10^{-8}$ and X_0 as before. If we perform 8 Newton iterations and then perform a modification step, we get $\|\tilde{X}_9\|_1 = 0.4142D-08$. Without the modification step, the error $\|\tilde{X}_k\|_1$ for the Newton iterate decreases monotonically untill the $26th$ iteration with $\|\tilde{X}_{26}\|_1 = 0.3738D-07$.

Example 4.2. Consider the algebraic Riccati equation (1.1) with $n = 2$ and

$$A = \begin{pmatrix} 3-\epsilon & 1 \\ 4 & 2-\epsilon \end{pmatrix}, \quad D = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}, \quad C = \begin{pmatrix} 4\epsilon - 11 & 2\epsilon - 5 \\ 2\epsilon - 5 & 2\epsilon - 2 \end{pmatrix}$$

(cf. Example 11 of [3]). The maximal solution is

$$X_+ = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}.$$

For $\epsilon = 0$, the pair (A, D) is stabilizable. And we have

$$A - DX_+ = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}, \quad \mathcal{N} = \left\{ \begin{pmatrix} a & 0 \\ 0 & a \end{pmatrix}, \ a \in \mathbb{R} \right\},$$

and observe that $\sigma(A - DX_+) = \{-i, i\}$.

Starting with

$$X_0 = \begin{pmatrix} 20 & 15 \\ 15 & 25 \end{pmatrix},$$

we perform 8 steps of the ordinary Newton iteration and then perform a modification step. The results are recorded in Table 2. The situation for this example is very similar to that for Example 4.1.

TABLE 2. Performance of Algorithm 3.3 for Example 4.2

k	$\tilde{x}_{11}^k$	$\tilde{x}_{12}^k = \tilde{x}_{21}^k$	$\tilde{x}_{22}^k$	$\|\tilde{X}_k\|_1$
0	$0.1800D+02$	$0.1400D+02$	$0.2400D+02$	$0.3800D+02$
1	$0.8788D+01$	$0.7445D+01$	$0.1136D+02$	$0.1880D+02$
2	$0.5173D+01$	$0.3136D+01$	$0.6099D+01$	$0.9235D+01$
3	$0.3230D+01$	$0.1051D+01$	$0.3452D+01$	$0.4504D+01$
4	$0.1988D+01$	$0.1997D+00$	$0.2008D+01$	$0.2208D+01$
5	$0.1088D+01$	$0.1090D-01$	$0.1088D+01$	$0.1099D+01$
6	$0.5493D+00$	$0.5518D-04$	$0.5493D+00$	$0.5494D+00$
7	$0.2747D+00$	$0.2772D-08$	$0.2747D+00$	$0.2747D+00$
8	$0.1373D+00$	$-0.4441D-15$	$0.1373D+00$	$0.1373D+00$
9	$-0.3997D-14$	$0.8882D-15$	$-0.5218D-14$	$0.6106D-14$

TABLE 3. Performance of Algorithm 3.3 for Example 4.3

k	$\|\tilde{X}_k\|_1$	$\|\mathcal{R}(X_k)\|_1$
0	$0.1000D+01$	$0.5000D+01$
1	$0.6245D+00$	$0.1398D+01$
2	$0.2783D+00$	$0.3203D+00$
3	$0.1378D+00$	$0.7232D-01$
4	$0.6503D-01$	$0.1696D-01$
5	$0.3167D-01$	$0.4030D-02$
6	$0.1575D-01$	$0.9859D-03$
7	$0.7872D-02$	$0.2459D-03$
8	$0.3936D-02$	$0.6147D-04$
9	$0.1968D-02$	$0.1537D-04$
10	$0.5215D-10$	$0.5207D-10$

For $\epsilon = 10^{-10}$ and the same initial guess, we perform 8 steps of Newton iteration and then perform a modification step. We get $\|\tilde{X}_9\|_1 = 0.1000D-09$. For the Newton iteration, the error decreases monotonically untill the $31st$ iteration with $\|\tilde{X}_{31}\| = 0.3663D-08$.

Example 4.3. We consider the algebraic Riccati equation (1.1) with $n = 8$ and a block-diagonal matrix A with 2×2 blocks:

$$A = \operatorname{diag}\left(0, \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 2 \\ -2 & 0 \end{pmatrix}, \begin{pmatrix} -1 & 1 \\ 0 & -1 \end{pmatrix}\right),$$

$$D = \begin{pmatrix} 2 & 1 & & 1 \\ 1 & 2 & \ddots & \\ & \ddots & \ddots & 1 \\ 1 & & 1 & 2 \end{pmatrix}, \quad C = 0.$$

It is readily seen that $X_+ = 0$ so that $\sigma(A - DX_+) = \{-1, 0, \pm i, \pm 2i\}$ and the purely imaginary eigenvalues have linear elementary divisors.

We apply Algorithm 3.3 with $X_0 = I$ and $\epsilon = 10^{-10}$. The results are recorded in Table 3. The first 9 steps are ordinary Newton iterations. The convergence of the Newton iteration is linear with common ratio $\frac{1}{2}$ (cf. Theorem 2.4). And by (2.3) we have $\|\mathcal{R}(X_k)\| \leq c\|\tilde{X}_k\|^2$ in this case, as verified by the numerical results. The last step is a modification step, which improves the accuracy dramatically.

We carried out many other numerical experiments. The results reported above are typical. In these experiments the convergence of the Newton method is always observed to be linear with common ratio $\frac{1}{2}$ whenever all elementary divisors of $A - DX_+$ are linear.

Appendix

This appendix is devoted to a sequence of results leading to a proof of Lemma 2.5. First, explicit representations for the subspaces $\mathcal{N} = \operatorname{Ker} \mathcal{R}'_{X_+}$ and $\mathcal{M} = \operatorname{Im} \mathcal{R}'_{X_+}$ are obtained. It is assumed throughout this appendix that the hypotheses of Theorems 1.1 and 2.1 hold.

It will be convenient to introduce the matrices

$$E_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \; E_2 = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}, \; E_3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \; E_4 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix},$$

and recall the real Jordan reduction used in the proof of Theorem 2.1:

$$P^{-1}(A - DX_+)P = J = \operatorname{diag}(G_1, \ldots, G_p),$$

and let G_j have size $r_j \times r_j, j = 1, 2, \ldots, p$.

Let $\mathcal{S}^k$ be the linear space of real symmetric matrices of order k. For $2 \leq j \leq p - 1$, define subspaces $\mathcal{S}_j, \mathcal{T}_j \subset \mathcal{S}^{r_j}$ by

$$\mathcal{S}_j = \{X \otimes E_1 + Y \otimes E_2 \mid X \text{ symmetric, } Y \text{ anti-symmetric; both have order } \frac{r_j}{2}\};$$

$$\mathcal{T}_j = \{X \otimes E_3 + Y \otimes E_4 \mid X, Y \text{ symmetric of order } \frac{r_j}{2}\}.$$

Here, $\otimes$ denotes the Kronecker product (see p. 97 of [14], for example). Note that $\dim \mathcal{S}_j = \frac{1}{4}r_j^2$ and $\dim \mathcal{T}_j = \frac{1}{4}r_j^2 + \frac{1}{2}r_j$ for $2 \leq j \leq p - 1$. Finally define

$$\mathcal{N}_J = \{N = \operatorname{diag}(N_1, \ldots, N_p) \mid N_i \in \mathbb{R}^{r_i \times r_i}, 1 \leq i \leq p; \\ N_1^T = N_1, N_p = 0, N_i \in \mathcal{S}_i, 2 \leq i \leq p - 1\},$$

$$\mathcal{M}_J = \{M = (M_{ij}) \mid M_{ij} \in \mathbb{R}^{r_i \times r_j}, M_{ij}^T = M_{ji}, 1 \leq i, j \leq p; \\ M_{11} = 0, M_{ii} \in \mathcal{T}_i, 2 \leq i \leq p - 1\}.$$

Lemma A.1. *If all purely imaginary eigenvalues of $A - DX_+$ have linear elementary divisors, then*

$$\mathcal{N} = \{P^{-T}NP^{-1} \mid N \in \mathcal{N}_J\},$$
$$\mathcal{M} = \{P^{-T}MP^{-1} \mid M \in \mathcal{M}_J\}.$$

Proof. The statement can be verified by block matrix multiplications and repeated use of Theorem 1.3. □

Lemma A.2. *For every complex number λ with non-negative real part ,*

$$\operatorname{rank}(\lambda I - J \quad P^{-1}DP^{-T}) = n.$$

Proof. Since (A, D) is stabilizable, there is a real X such that $A - DX$ is stable. Now

$$\begin{aligned} J - P^{-1}DP^{-T}P^T(X - X_+)P &= P^{-1}(A - DX_+)P - P^{-1}(DX - DX_+)P \\ &= P^{-1}(A - DX)P, \end{aligned}$$

which is stable. Thus $(J, P^{-1}DP^{-T})$ is a stabilizable pair. The result now follows from Theorem 4.5.6(a) of [14]. □

Lemma A.3. *Let W be a Hermitian positive semidefinite matrix. If the determinant of a principal submatrix of W is zero, then the rows of W containing this submatrix must be linearly dependent.*

Proof. Let $W = (w_{ij})_{i,j=1}^n$. We may assume without loss of generality that the principal submatrix is $W_1 = (w_{ij})_{i,j=1}^r = (\alpha_1^T \cdots \alpha_r^T)^T (r \le n)$ and that $\alpha_1 = c_2\alpha_2 + \cdots + c_r\alpha_r$ for some constants $c_2, \ldots, c_r$. Let $E(i, j(k))$ be the elementary matrix obtained from I by adding k times row j to row i. Let $U = E(1, r(-c_r)) \cdots E(1, 2(-c_2))$. Then UWU^H is Hermitian positive semidefinite and has zero in the $(1,1)$ position. Hence the first row of UWU^H is zero. This means that $\beta_1 = c_2\beta_2 + \cdots + c_r\beta_r$, where $\beta_1, \ldots, \beta_r$ are the first r rows of W. □

Now consider the map $\mathcal{B}_Z : \mathcal{N} \to \mathcal{N}$ of (2.4). By Lemma A.1, we can write $Y = P^{-T}Y_JP^{-1}, Z = P^{-T}Z_JP^{-1}$ with $Y_J, Z_J \in \mathcal{N}_J$. Therefore

$$\begin{aligned} \mathcal{B}_Z(Y) &= P_{\mathcal{N}}(ZDY + YDZ) \\ &= P^{-T}P_{\mathcal{N}_J}(Z_JP^{-1}DP^{-T}Y_J + Y_JP^{-1}DP^{-T}Z_J)P^{-1}, \end{aligned}$$

where $P_{\mathcal{N}_J}$ is the projection onto $\mathcal{N}_J$ parallel to $\mathcal{M}_J$. Let $Z_J = \operatorname{diag}(Z_1, \ldots, Z_p)$, $Y_J = \operatorname{diag}(Y_1, \ldots, Y_p)$ and $\operatorname{diag}(D_1, \ldots, D_p)$ be the block diagonal of $P^{-1}DP^{-T}$. Let $\mathcal{S}_1 = \mathcal{S}^{r_1}$. We have further

$$\mathcal{B}_Z(Y) = P^{-T}\operatorname{diag}\big(\mathcal{F}_{Z_1}(Y_1), \mathcal{F}_{Z_2}(Y_2), \ldots, \mathcal{F}_{Z_{p-1}}(Y_{p-1}), 0\big)P^{-1}, \tag{A.1}$$

where we define linear transformations $\mathcal{F}_{Z_i} : \mathcal{S}_i \to \mathcal{S}_i$ by

$$\begin{aligned} \mathcal{F}_{Z_1}(Y_1) &= Z_1D_1Y_1 + Y_1D_1Z_1, \\ \mathcal{F}_{Z_i}(Y_i) &= P_{\mathcal{S}_i}(Z_iD_iY_i + Y_iD_iZ_i), \qquad 2 \le i \le p-1, \end{aligned}$$

with $P_{\mathcal{S}_i}$ being the projection onto $\mathcal{S}_i$ parallel to $\mathcal{T}_i$.

For $i = 1, 2, \ldots, p-1$, let

$$\mathcal{U}_i = \{Z_i \in \mathcal{S}_i \mid \mathcal{F}_{Z_i} : \mathcal{S}_i \to \mathcal{S}_i \text{ is not invertible}\}.$$

Lemma A.4. *The set $\mathcal{U}_1$ has measure zero in $\mathcal{S}_1$.*

Proof. For $W_1 \in \mathcal{S}_1$, we can rewrite $Z_1D_1Y_1 + Y_1D_1Z_1 = W_1$ as

$$\big(I \otimes (Z_1D_1) + (Z_1D_1) \otimes I\big)\operatorname{vec} Y_1 = \operatorname{vec} W_1$$

(see [14, p. 99]). Thus

$$\mathcal{U}_1 \subset \big\{Z_1 \in \mathcal{S}_1 \mid \det\big(I \otimes (Z_1D_1) + (Z_1D_1) \otimes I\big) = 0\big\}.$$

Since $Z_1 \in \mathcal{S}^{r_1}$, the determinant is an algebraic polynomial in $r_1(r_1+1)/2$ variables. The set $\mathcal{U}_1$ has measure zero in $\mathcal{S}_1$ unless

$$\det\big(I \otimes (Z_1 D_1) + (Z_1 D_1) \otimes I\big) \equiv 0. \tag{A.2}$$

If (A.2) is true, we have in particular $\det(I \otimes D_1^2 + D_1^2 \otimes I) = 0$. Thus 0 is an eigenvalue of the matrix $I \otimes D_1^2 + D_1^2 \otimes I$. We can then find eigenvalues λ_i, λ_j of D_1 such that $\lambda_i^2 + \lambda_j^2 = 0$ (see [14, Theorem 5.1.1]). Hence 0 is an eigenvalue of D_1 and $\det D_1 = 0$. By Lemma A.3, the first r_1 rows of $P^{-1}DP^{-T}$ are linearly dependent. Thus $\operatorname{rank}(-J \quad P^{-1}DP^{-T}) < n$, which contradicts Lemma A.2. □

Lemma A.5. *For $k = 2, 3, \ldots, p-1$, the set $\mathcal{U}_k$ has measure zero in $\mathcal{S}_k$.*

Proof. We will first find a more explicit expression for $\mathcal{F}_{Z_k}(Y_k)$. By Lemma A.1, we can write

$$Y_k = M_s \otimes E_1 + M_a \otimes E_2, \quad Z_k = N_s \otimes E_1 + N_a \otimes E_2, \tag{A.3}$$

where M_s and N_s are symmetric; M_a and N_a are anti-symmetric. Let

$$D_k = (D_{ij})_{i,j=1}^{r_k/2} \text{ with } D_{ij} = \begin{pmatrix} d_1^{ij} & d_3^{ij} \\ d_4^{ij} & d_2^{ij} \end{pmatrix},$$

$$Q_s = (q_{ij}^s)_{i,j=1}^{r_k/2} \text{ with } q_{ij}^s = \frac{1}{2}(d_1^{ij} + d_2^{ij}),$$

$$Q_a = (q_{ij}^a)_{i,j=1}^{r_k/2} \text{ with } q_{ij}^a = \frac{1}{2}(d_3^{ij} - d_4^{ij}).$$

Then

$$D_k = Q_s \otimes E_1 + Q_a \otimes E_2 + R_s \otimes E_3 + T_s \otimes E_4, \tag{A.4}$$

where Q_s, R_s and T_s are symmetric; Q_a is anti-symmetric. Using (A.3) and (A.4) to expand $Z_k D_k Y_k + Y_k D_k Z_k$, we finally get

$$\begin{aligned}\mathcal{F}_{Z_k}(Y_k) = {} & (N_s Q_s M_s + M_s Q_s N_s - N_a Q_a M_s - M_s Q_a N_a \\ & - N_s Q_a M_a - M_a Q_a N_s - N_a Q_s M_a - M_a Q_s N_a) \otimes E_1 \\ & + (N_s Q_a M_s + M_s Q_a N_s + N_a Q_s M_s + M_s Q_s N_a \\ & + N_s Q_s M_a + M_a Q_s N_s - N_a Q_a M_a - M_a Q_a N_a) \otimes E_2.\end{aligned}$$

For $W_k \in \mathcal{S}_k$, we write $W_k = L_s \otimes E_1 + L_a \otimes E_2$ with L_s symmetric and L_a anti-symmetric. Thus $\mathcal{F}_{Z_k}(Y_k) = W_k$ if and only if

$$\begin{aligned}& N_s Q_s M_s + M_s Q_s N_s - N_a Q_a M_s - N_s Q_a N_a \\ & \quad - N_s Q_a M_a - M_a Q_a N_s - N_a Q_s M_a - M_a Q_s N_a = L_s,\end{aligned} \tag{A.5}$$

and

$$\begin{aligned}& N_s Q_a M_s + M_s Q_a N_s + N_a Q_s M_s + M_s Q_s N_a \\ & \quad + N_s Q_s M_a + M_a Q_s N_s - N_a Q_a M_a - M_a Q_a N_a = L_a.\end{aligned} \tag{A.6}$$

The above two equalities produce a system of linear equations. The determinant of the coefficient matrix, $\det(N_s, N_a)$, is an algebraic polynomial in $(\frac{r_k}{2})^2$ variables ($\frac{r_k}{4}(\frac{r_k}{2}+1)$ variables from N_s and $\frac{r_k}{4}(\frac{r_k}{2}-1)$ variables from N_a). Now $\mathcal{U}_k = \{Z_k \in \mathcal{S}_k \mid \det(N_s, N_a) = 0\}$. The set $\mathcal{U}_k$ has measure zero in $\mathcal{S}_k$ unless $\det(N_s, N_a) \equiv 0$.

The equalities (A.5) and (A.6) can be combined into a neat complex form

$$\begin{aligned}(N_s + iN_a)(Q_s + iQ_a)(M_s + iM_a) + (M_s + iM_a)(Q_s + iQ_a)(N_s + iN_a) \\ = L_s + iL_a.\end{aligned}$$

If $\det(N_s, N_a) \equiv 0$, we have in particular $\det(Q_s, Q_a) = 0$. Thus

$$(Q_s + iQ_a)^2(M_s + iM_a) + (M_s + iM_a)(Q_s + iQ_a)^2 = 0$$

has a nonzero solution $M_s + iM_a$, which implies that

$$\det\big(I \otimes (Q_s + iQ_a)^2 + (Q_s + iQ_a)^2 \otimes I\big) = 0.$$

Since $Q_s + iQ_a$ is Hermitian, its eigenvalues are all real. It follows as in the proof of Lemma A.4 that $\det(Q_s + iQ_a) = 0$.

To complete the proof, we need to show $\det(Q_s + iQ_a) \neq 0$. By Lemma A.2 we have $\operatorname{rank}(a_k iI - J \quad P^{-1}DP^{-T}) = n$. Let $t_k = r_1 + \cdots + r_{k-1}$ and

$$U = E\big(t_k + r_k - 1, (t_k + r_k)(-i)\big) \cdots E\big(t_k + 3, (t_k + 4)(-i)\big) E\big(t_k + 1, (t_k + 2)(-i)\big).$$

Then

$$\operatorname{rank}\big(U(a_k iI - J) \quad UP^{-1}DP^{-T}U^H\big) = n.$$

Since the $(s_k + 1)$th, $(s_k + 3)$th, $\ldots$, $(s_k + r_k - 1)$th rows of the matrix $U(a_k iI - J)$ are all zero, the corresponding rows of the Hermitian positive semidefinite matrix $UP^{-1}DP^{-T}U^H$ must be linearly independent. By Lemma A.3, the principal submatrix (of order $r_k/2$) of $UP^{-1}DP^{-T}U^H$ contained in these rows must have a nonzero determinant. The principal submatrix turns out to be precisely $2(Q_s + iQ_a)$. Therefore $\det(Q_s + iQ_a) \neq 0$. □

Proof of Lemma 2.5. From (A.1) we see that $\mathcal{B}_Z$ is invertible if and only if $\mathcal{F}_{Z_i}$ is invertible for each i. Thus

$$\mathcal{U} = \bigcup_{i=1}^{p-1} \mathcal{V}_i,$$

where

$$\mathcal{V}_i = \{P^{-T}\operatorname{diag}(X_1, \ldots, X_{p-1}, 0)P^{-1} \mid X_i \in \mathcal{U}_i, X_j \in \mathcal{S}_j, j \neq i\}.$$

By Lemmas A.4 and A.5, each $\mathcal{V}_i$ has measure zero in $\mathcal{N}$. Therefore $\mathcal{U}$ has measure zero in $\mathcal{N}$. □

References

1. R. H. Bartels and G. W. Stewart, *Solution of the matrix equation $AX + XB = C$*, Comm. ACM **15** (1972), 820–826.
2. P. Benner and R. Byers, *An exact line search method for solving generalized continuous-time algebraic Riccati equations*, IEEE Trans. Autom. Control (to appear).
3. P. Benner, A. J. Laub and V. Mehrmann, *A collection of benchmark examples for the numerical solution of algebraic Riccati equations* I: *continuous-time case,* Technical Report SPC 95-22, Fakultät für Mathematik, Technische Universität Chemnitz-Zwickau, FRG, 1995.
4. W. A. Coppel, *Matrix quadratic equations*, Bull. Austral. Math. Soc. **10** (1974), 377–401. MR **51:**3623
5. D. W. Decker, H. B. Keller and C. T. Kelley, *Convergence rates for Newton's method at singular points*, SIAM J. Numer. Anal. **20** (1983), 296–314. MR **84d:**65041
6. D. W. Decker and C. T. Kelley, *Newton's Method at singular points* I, SIAM J. Numer. Anal. **17** (1980), 66–70. MR **81k:**65065a
7. ———, *Convergence acceleration for Newton's method at singular points*, SIAM J. Numer. Anal. **19** (1982), 219–229. MR **83e:**65090
8. I. Gohberg, P. Lancaster and L. Rodman, *On Hermitian solutions of the symmetric algebraic Riccati equation*, SIAM J. Control Optimization **24** (1986), 1323–1334. MR **88f:**93041

9. G. H. Golub, S. Nash and C. Van Loan, *A Hessenberg-Schur method for the problem* $AX + XB = C$, IEEE Trans. Autom. Control **24** (1979), 909–913. MR **81a:**65046
10. L. V. Kantorovich and G. P. Akilov, *Functional analysis in normed spaces*, Pergamon, New York, 1964. MR **35:**4699
11. C. T. Kelley, *A Shamanskii-like acceleration scheme for nonlinear equations at singular roots*, Math. Comp. **47** (1986), 609–623. MR **87m:**65100
12. C. T. Kelley and R. Suresh, *A new acceleration method for Newton's method at singular points*, SIAM J. Numer. Anal. **20** (1983), 1001–1009. MR **85c:**65063
13. D. L. Kleinman, *On an iterative technique for Riccati equation computations*, IEEE Trans. Autom. Control **13** (1968), 114–115.
14. P. Lancaster and L. Rodman, *Algebraic Riccati equations*, Oxford University Press, 1995. MR **97b:**93003
15. A. Linnemann, *Numerische methoden für lineare regelungssysteme*, BI Wissenschafts Verlag, Mannheim, 1993. MR **94g:**93001
16. V. L. Mehrmann, *The autonomous linear quadratic control problem*, Lecture Notes in Control and Information Sciences, Vol. 163, Springer Verlag, Berlin, 1991. MR **93d:**93004
17. J. M. Ortega and W. C. Rheinboldt, *Iterative solutions of nonlinear equations in several variables*, Academic Press, New York, 1970. MR **42:**8686
18. G. W. Reddien, *On Newton's method for singular problems*, SIAM J. Numer. Anal. **15** (1978), 993–996. MR **80b:**65064
19. V. Sima, *An efficient Schur method to solve the stabilizing problem*, IEEE Trans. Autom. Control **26** (1981), 724–725. MR **82j:**93032
20. H. K. Wimmer, *Monotonicity of maximal solutions of algebraic Riccati equations*, Syst. Control Lett. **5** (1985), 317–319. MR **86f:**93083

Department of Mathematics and Statistics, University of Calgary, Calgary, Alberta, Canada T2N 1N4

E-mail address: guo@math.ucalgary.ca

E-mail address: lancaste@math.ucalgary.ca