# NEWTON'S METHOD FOR DISCRETE ALGEBRAIC RICCATI EQUATIONS WHEN THE CLOSED-LOOP MATRIX HAS EIGENVALUES ON THE UNIT CIRCLE

CHUN-HUA GUO*

**Abstract.** When Newton's method is applied to find the maximal symmetric solution of a discrete algebraic Riccati equation (DARE), convergence can be guaranteed under moderate conditions. In particular, the initial guess does not need to be close to the solution. The convergence is quadratic if the Fréchet derivative is invertible at the solution. When the closed-loop matrix has eigenvalues on the unit circle, the derivative at the solution is not invertible. The convergence of Newton's method is shown to be either quadratic or linear with common ratio $\frac{1}{2}$, provided that the eigenvalues on the unit circle are all semi-simple. The linear convergence appears to be dominant, and the efficiency of the Newton iteration can be improved significantly by applying a double Newton step at the right time.

**Key words.** discrete algebraic Riccati equations, Newton's method, maximal symmetric solution, convergence rate, matrix pencils

**AMS subject classifications.** 15A24, 65H10, 93B40

**1. Introduction.** Algebraic Riccati equations occur in many important applications [18], [20]. In a previous paper [11] we considered Newton's method for continuous algebraic Riccati equations (CARE). In this paper we consider discrete algebraic Riccati equations (DARE) of the form

$$-X + A^TXA + Q - (C + B^TXA)^T(R + B^TXB)^{-1}(C + B^TXA) = 0, \tag{1.1}$$

where $A, Q \in \mathbb{R}^{n\times n}, B \in \mathbb{R}^{n\times m}, C \in \mathbb{R}^{m\times n}, R \in \mathbb{R}^{m\times m}$, and $Q^T = Q, R^T = R$. We denote by $\mathcal{R}(X)$ the left-hand side of (1.1). The function $\mathcal{R}(X)$ and its derivatives are much more complicated than their CARE counterparts. Nevertheless, it will be shown that most analytical properties established in [11] for the CARE can be extended to the DARE. The analysis here is more involved, but the line of attack is the same.

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. Let $\mathcal{D} = \{X \in \mathcal{S} \mid R+B^TXB \text{ is invertible}\}$. We have $\mathcal{R} : \mathcal{D} \to \mathcal{S}$. The first Fréchet derivative of $\mathcal{R}$ at a matrix $X \in \mathcal{D}$ is a linear map $\mathcal{R}'_X : \mathcal{S} \to \mathcal{S}$ given by

$$\mathcal{R}'_X(S) = -S + \hat{A}^TS\hat{A}, \tag{1.2}$$

where $\hat{A} = A - B(R + B^TXB)^{-1}(C + B^TXA)$. Also the second derivative at $X \in \mathcal{D}$, $\mathcal{R}''_X : \mathcal{S} \times \mathcal{S} \to \mathcal{S}$, is given by

$$\mathcal{R}''_X(S_1, S_2) = -\hat{A}^TS_1HS_2\hat{A} - \hat{A}^TS_2HS_1\hat{A}, \tag{1.3}$$

where $H = B(R + B^TXB)^{-1}B^T$.

For $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times m}$, the pair $(A, B)$ is said to be d-stabilizable if there is a $K \in \mathbb{R}^{m\times n}$ such that $A - BK$ is d-stable, i.e., all its eigenvalues are in the open unit disk. For real symmetric matrices $X$ and $Y$, we write $X \geq Y (X > Y)$ if $X - Y$ is positive semidefinite (definite). A symmetric solution $X_+$ of (1.1) is called maximal if

*Department of Mathematics and Statistics, University of Calgary, Calgary, Alberta T2N 1N4, Canada (guo@math.ucalgary.ca).

$X_+ \geq X$ for every symmetric solution $X$. The following result is essentially the real version of Theorem 13.1.1 in [18]. See also [22].

THEOREM 1.1. *Let $(A, B)$ be a d-stabilizable pair and assume that there is a symmetric solution $\tilde{X}$ of the inequality $\mathcal{R}(X) \geq 0$ for which $R + B^T\tilde{X}B > 0$. Then there exists a maximal symmetric solution $X_+$ of $\mathcal{R}(X) = 0$. Moreover, $R + B^T X_+ B > 0$ and all the eigenvalues of $A - B(R + B^T X_+ B)^{-1}(C + B^T X_+ A)$ lie in the closed unit disk.*

*Remark* 1.1. In Theorem 13.1.1 of [18], the matrix $R$ is required to be invertible. This requirement is needed for some later developments in [18], but is not necessary for the conclusions of Theorem 13.1.1. The proof of that theorem should be slightly modified. We have only to replace expressions of the form $Q - C^T R^{-1} C + (L - R^{-1}C)^T R(L - R^{-1}C)$ by expressions of the form $Q + L^T RL - C^T L - L^T C$. That the invertibility of $R$ is not necessary for the conclusions of Theorem 1.1 has also been noted in [2]. As noted in [3], the matrix $R$ may well be singular in applications.

A symmetric solution $X$ of (1.1) is called stabilizing (resp. almost stabilizing) if all the eigenvalues of $A - B(R+B^T XB)^{-1}(C+B^T XA)$ are in the open (resp. closed) unit disk. 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.

The Newton method for the solution of (1.1) is

$$X_i = X_{i-1} - (\mathcal{R}'_{X_{i-1}})^{-1}\mathcal{R}(X_{i-1}), \quad i = 1, 2, \ldots, \tag{1.4}$$

given that the maps $\mathcal{R}'_{X_i}(i = 0, 1, \ldots)$ are all invertible. The iteration (1.4) is closely related to the solution of the Stein equation described in the following classical result.

THEOREM 1.2 (cf. [18, p. 100]). *For any given matrices $A, B, \Gamma \in \mathbb{R}^{n\times n}$ the Stein equation $S - BSA = \Gamma$ has a unique solution (necessarily real) if and only if $\lambda_r\mu_s \neq 1$ for any $\lambda_r \in \sigma(A), \mu_s \in \sigma(B)$.*

It follows from Theorem 1.2 that, under the conditions of Theorem 1.1, $\mathcal{R}'_{X_+}$ is invertible if and only if $A - B(R + B^T X_+ B)^{-1}(C + B^T X_+ A)$ is d-stable.

When we apply Newton's method to the DARE (1.1) with $(A, B)$ d-stabilizable, the initial matrix $X_0$ is taken such that $A - B(R + B^T X_0 B)^{-1}(C + B^T X_0 A)$ is d-stable. The usual way to generate such an $X_0$ is as follows. We choose $L_0 \in \mathbb{R}^{m\times n}$ such that $A_0 = A - BL_0$ is d-stable, and take $X_0$ to be the unique solution of the Stein equation

$$X_0 - A_0^T X_0 A_0 = Q + L_0^T RL_0 - C^T L_0 - L_0^T C. \tag{1.5}$$

In view of (1.2), the Newton iteration (1.4) can be rewritten as

$$X_i - A_i^T X_i A_i = Q + L_i^T RL_i - C^T L_i - L_i^T C, \quad i = 1, 2, \ldots, \tag{1.6}$$

where

$$L_i = (R + B^T X_{i-1} B)^{-1}(C + B^T X_{i-1} A) \tag{1.7}$$

and

$$A_i = A - BL_i. \tag{1.8}$$

THEOREM 1.3. *Under the same conditions as in Theorem* 1.1 *and for any $L_0 \in \mathbb{R}^{m\times n}$ such that $A_0 = A - BL_0$ is d-stable, starting with the symmetric matrix $X_0$ determined by* (1.5)*, the recursion* (1.6) *determines a sequence of symmetric matrices*

$\{X_i\}_{i=0}^{\infty}$ *for which* $A - B(R + B^T X_i B)^{-1}(C + B^T X_i A)$ *is d-stable for* $i = 0, 1, \ldots,$ $X_0 \geq X_1 \geq \cdots,$ *and* $\lim_{i\to\infty} X_i = X_+$.

The maximal solution can thus be found by the Newton iteration with an initial guess not necessarily close to the solution. The proof of the above theorem can be found in [18, pp. 308–311] (with some slight modifications as pointed out in Remark 1.1). See also [13] and [22]. Note that an $L_0$ can be produced by automatic stabilizing procedures such as the one in [24]. It should also be noted that $X_0 \geq X_1$ is generally not true, if $X_0$ is not obtained from (1.5).

It is readily seen that $\mathcal{R}'_X$, as a function of $X$, is Lipschitz continuous on a closed ball centered at $X_+$ and contained in $\mathcal{D}$. Thus the well known locally quadratic convergence of Newton's method (see [15], [21]), in combination with Theorem 1.3, yields the following result.

Theorem 1.4. *If* $A - B(R + B^T X_+ B)^{-1}(C + B^T X_+ A)$ *is d-stable in Theorem* 1.3*, 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.*

When the closed-loop matrix $A - B(R + B^T X_+ B)^{-1}(C + B^T X_+ A)$ has eigenvalues on the unit circle, $\mathcal{R}'_{X_+}$ is not invertible. This situation happens in some important applications (see [4], for example). We will show that the convergence of Newton's method is either quadratic or linear with common ratio $\frac{1}{2}$, provided that the eigenvalues on the unit circle are all semi-simple (i.e. all elementary divisors corresponding to these eigenvalues are linear). The linear convergence appears to be dominant and, when this is the case, the efficiency of the Newton iteration can be improved significantly by applying a double Newton step at the right time. Numerical results are also given to illustrate these phenomena.

As in [11] we apply the following general formulation of Newton's method (see [5], [6], [7], [16], [17], [23]). Let $F$ be a smooth map from a Banach space $E$ into itself. Assume 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 further that the following *regularity* condition holds: 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. These ideas can now be used to formulate sufficient conditions for *local* convergence.

Theorem 1.5 (cf. [16, Theorem 1.1]). *Let* $E = N \oplus M$*, let* $\phi_0$ *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*

$$W(\rho, \theta, \eta) = \{x \mid 0 < \|\tilde{x}\| < \rho, \|P_M \tilde{x}\| \leq \theta \|P_N \tilde{x}\|, \\ \|(P_N - P_0)\tilde{x}\| \leq \eta \|P_N \tilde{x}\|\}, \tag{1.9}$$

*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* $\rho_0, \theta_0, \eta_0$ *sufficiently small, then the Newton sequence* $\{x_i\}$ *is well defined and* $\|F'(x_i)^{-1}\| \leq c\|\tilde{x}_i\|^{-1}$ *for all* $i \geq 1$ *and some constant* $c > 0$*. Moreover,*

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

The regularity condition is very important for the above theorem. Without this condition, the behaviour of Newton's method can be very erratic (see, e.g., [10]). Before we can apply Theorem 1.5 to the DARE (1.1), we need to check the direct sum condition and the regularity condition for the DARE. The direct sum condition will be discussed in Sections 2 and 3. The regularity condition is satisfied for the DARE whenever the matrix pair $(A, B)$ is d-stabilizable. This will be discussed in Section 4.

If the matrix pair $(A, B)$ is not d-stabilizable, a generalized Newton's method may be used for the solution of the DARE (1.1). For differential periodic Riccati equations without the stability condition, the convergence of a generalized Newton's method has been established in [12]. The ideas used in that paper can also be used for CAREs or DAREs without the stabilizability condition. In this paper, however, we restrict ourselves to the standard Newton's method and assume that the matrix pair $(A, B)$ is d-stabilizable.

**2. Interpretation of the direct sum condition for the DARE.** We now go back to the discussion of the DARE (1.1) 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 $\mathcal{N} = \mathrm{Ker}\mathcal{R}'_{X_+}$, $\mathcal{M} = \mathrm{Im}\mathcal{R}'_{X_+}$. We have the following interpretation of the direct sum condition.

THEOREM 2.1. *$\mathcal{S} = \mathcal{N} \oplus \mathcal{M}$ if and only if all eigenvalues of*

$$A_+ = A - B(R + B^T X_+ B)^{-1}(C + B^T X_+ A)$$

*on the unit circle are semi-simple.*

*Proof.* Let $J$ be the real Jordan canonical form for $A_+$ with $P^{-1}A_+P = J$ and a real matrix $P$. We find that $K \in \mathcal{N}$ if and only if $K = P^{-T}LP^{-1}$ for some $L \in \mathcal{N}_J = \{Y \in \mathcal{S} \mid -Y + J^TYJ = 0\}$. Also $W \in \mathcal{M}$ if and only if $W = P^{-T}UP^{-1}$ for some $U \in \mathcal{M}_J = \{Y \in \mathcal{S} \mid Y = -V + J^TVJ \text{ for some } V \in \mathcal{S}\}$. Therefore, $\mathcal{S} = \mathcal{N} \oplus \mathcal{M}$ if and only if $\mathcal{S} = \mathcal{N}_J \oplus \mathcal{M}_J$.

If all eigenvalues of $A_+$ on the unit circle are semi-simple, we gather the Jordan blocks of $J$ in several groups:

$$J = \mathrm{diag}(G_1, G_2, G_3, \ldots, G_{p-1}, G_p). \tag{2.1}$$

Here $G_1 = -I_{r_1}, G_2 = I_{r_2}, G_p \in \mathbb{R}^{r_p \times r_p}$ consists of real Jordan blocks associated with eigenvalues in the open unit disk, and for $i = 3, \ldots, p-1$,

$$G_i = \mathrm{diag}\left(\begin{pmatrix} a_i & b_i \\ -b_i & a_i \end{pmatrix}, \ldots, \begin{pmatrix} a_i & b_i \\ -b_i & a_i \end{pmatrix}\right) \in \mathbb{R}^{r_i \times r_i}, \tag{2.2}$$

where $-1 < a_3 < \cdots < a_{p-1} < 1$, $b_i > 0$, and $a_i^2 + b_i^2 = 1$ for $i = 3, \ldots, p-1$. Using block matrix multiplications and applying Theorem 1.2 repeatedly, we can show that $\mathcal{S} = \mathcal{N}_J \oplus \mathcal{M}_J$. The detailed expressions for $\mathcal{N}_J$ and $\mathcal{M}_J$ will be given in Lemma 2.2 below and will be needed in the sequel.

If $A_+$ has nonlinear elementary divisors corresponding to eigenvalues on the unit circle, we can arrange the Jordan blocks so that the first Jordan block $J_1$ has one of the following two forms:

$$\text{(i)} \quad J_1 = \begin{pmatrix} a & 1 & & \\ & a & \ddots & \\ & & \ddots & 1 \\ & & & a \end{pmatrix}, \quad a = \pm 1.$$

$$\text{(ii)}\ \ J_1 = \begin{pmatrix} B & I & & \\ & B & \ddots & \\ & & \ddots & I \\ & & & B \end{pmatrix}, \quad B = \begin{pmatrix} a & b \\ -b & a \end{pmatrix}, \quad b > 0, \quad a^2 + b^2 = 1.$$

For the first case, $D_1 = \text{diag}(0, \dots, 0, 1, 0, \dots, 0) \in \mathcal{N}_J \cap \mathcal{M}_J$, where the element 1 appears at the same position as the last diagonal element of $J_1$. Note that $D_1 = -V_1 + J^T V_1 J$ for

$$V_1 = \frac{1}{2}\text{sign}(a) \begin{pmatrix} 0 & & & \\ & 0 & 1 & \\ & 1 & 0 & \\ & & & 0 \end{pmatrix},$$

where the $2 \times 2$ matrix in the center appears at the same position as the southeast corner of $J_1$. For the second case, $D_2 = \text{diag}(0, \dots, 0, I, 0, \dots, 0) \in \mathcal{N}_J \cap \mathcal{M}_J$, where the $2 \times 2$ identity matrix $I$ appears at the same position as the last diagonal block of $J_1$. Note that $D_2 = -V_2 + J^T V_2 J$ for

$$V_2 = \frac{1}{2b} \begin{pmatrix} 0 & & & \\ & 0 & T & \\ & -T & 0 & \\ & & & 0 \end{pmatrix} \text{with } T = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix},$$

where the $4 \times 4$ matrix in the center appears at the same position as the southeast corner of $J_1$. Therefore, $\mathcal{S} \neq \mathcal{N}_J \oplus \mathcal{M}_J$. ∎

In order to give an explicit construction of the spaces $\mathcal{N}_J$ and $\mathcal{M}_J$, we introduce, as in [11], the matrices

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

and let $\mathcal{S}^k$ be the linear space of real symmetric matrices of order $k$. For $3 \leq j \leq p-1$, we 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 [18], for example).

Lemma 2.2. *If all eigenvalues of $A_+$ on the unit circle are semi-simple, then*

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

*with*

$$\begin{aligned} \mathcal{N}_J = \{N = \text{diag}(N_1, \dots, N_p) \mid N_i \in \mathbb{R}^{r_i \times r_i}, 1 \leq i \leq p; \\ N_1^T = N_1, N_2^T = N_2, N_p = 0, N_i \in \mathcal{S}_i, 3 \leq i \leq p-1\}, \\ \mathcal{M}_J = \{M = (M_{ij}) \mid M_{ij} \in \mathbb{R}^{r_i \times r_j}, M_{ij}^T = M_{ji}, 1 \leq i, j \leq p; \\ M_{11} = 0, M_{22} = 0, M_{ii} \in \mathcal{T}_i, 3 \leq i \leq p-1\}. \end{aligned}$$

*Proof.* The statement can be verified using block matrix multiplications and Theorem 1.2. ∎

**3. Characterization of the direct sum condition via a matrix pencil.** We have given in §2 a characterization of the direct sum condition, in which the sought after solution $X_+$ appears. In order to give a characterization which is independent of $X_+$, we consider the matrix pencil $\lambda F_e - G_e$ with

$$F_e = \begin{pmatrix} I & 0 & 0 \\ 0 & A^T & 0 \\ 0 & -B^T & 0 \end{pmatrix}, \qquad G_e = \begin{pmatrix} A & 0 & B \\ -Q & I & -C^T \\ C & 0 & R \end{pmatrix}.$$

Matrix pencils of this type were first introduced in [8] and [25], but for a different purpose. See also [14].

LEMMA 3.1. *If* (1.1) *has a Hermitian solution* $X$*, then*

$$(\lambda F_e - G_e)\begin{pmatrix} I & 0 & 0 \\ X & I & 0 \\ Z & 0 & I \end{pmatrix} = \begin{pmatrix} I & 0 & 0 \\ A^T X & I & Z^T \\ -B^T X & 0 & I \end{pmatrix}(\lambda M_e - N_e), \tag{3.1}$$

*where* $Z = -(R + B^T XB)^{-1}(C + B^T XA)$ *and*

$$M_e = \begin{pmatrix} I & 0 & 0 \\ 0 & (A+BZ)^T & 0 \\ 0 & -B^T & 0 \end{pmatrix}, \qquad N_e = \begin{pmatrix} A+BZ & 0 & B \\ 0 & I & 0 \\ 0 & 0 & R + B^T XB \end{pmatrix}.$$

*Proof.* It can be easily verified by direct computation. □

Note that, in contrast with Proposition 15.2.1 of [18], the equality (3.1) does not require the invertibility of $R$.

COROLLARY 3.2. *If* (1.1) *has a Hermitian solution* $X$*, then* $\lambda F_e - G_e$ *is a regular pencil. Moreover,* $\alpha \neq 0$ *is an eigenvalue of* $A + BZ$ *if and only if* $\alpha$ *and* $\bar{\alpha}^{-1}$ *are eigenvalues of* $\lambda F_e - G_e$*. A unimodular* $\alpha$ *is an eigenvalue of* $A + BZ$ *with algebraic multiplicity* $k$ *if and only if it is an eigenvalue of* $\lambda F_e - G_e$ *with algebraic multiplicity* $2k$*.*

*Proof.* We have by Lemma 3.1

$$\det(\lambda F_e - G_e) = (-1)^m \det(R + B^T XB)\det(\lambda I - (A + BZ))\det(\lambda(A + BZ)^T - I).$$

If $\det(\lambda I - (A + BZ)) = (\lambda - \lambda_1)\cdots(\lambda - \lambda_n)$, we have $\det(\lambda(A + BZ)^T - I) = (\bar{\lambda}_1\lambda - 1)\cdots(\bar{\lambda}_n\lambda - 1)$. The conclusions in the corollary now follow easily. □

If all unimodular eigenvalues of $\lambda F_e - G_e$ are of algebraic multiplicity two, then all unimodular eigenvalues of $A+BZ$ are simple and the direct sum condition is satisfied. To give a complete characterization, we need to consider the relationship between the elementary divisors of $A + BZ$ and $\lambda F_e - G_e$.

THEOREM 3.3. *Let* $\alpha$ *be a complex number with* $|\alpha| = 1$ *and* $X$ *be a solution of* (1.1) *with* $R + B^T XB > 0$*. If*

$$\operatorname{rank}(\alpha I - A \;\; B) = n,$$

*then the elementary divisors of* $A+BZ$ *corresponding to* $\alpha$ *have degrees* $k_1, \ldots, k_s (1 \leq k_1 \leq \cdots \leq k_s \leq n)$ *if and only if the elementary divisors of* $\lambda F_e - G_e$ *corresponding to* $\alpha$ *have degrees* $2k_1, \ldots, 2k_s$*.*

*Proof.* Suppose the elementary divisors of $A + BZ$ corresponding to $\alpha$ have degrees $k_1, \ldots, k_s$. By the local Smith form (see [9], for example), we can find matrix polynomials $E_\alpha(\lambda)$ and $F_\alpha(\lambda)$ invertible at $\alpha$ such that

$$\lambda I - (A + BZ) = E_\alpha(\lambda)\begin{pmatrix} I & 0 \\ 0 & D \end{pmatrix} F_\alpha(\lambda), \tag{3.2}$$

where $D = \mathrm{diag}((\lambda-\alpha)^{k_1},\ldots,(\lambda-\alpha)^{k_s})$. Replacing $\lambda$ by $\bar{\lambda}^{-1}$ in (3.2), and then taking conjugate transpose (denoted by $*$), we get

$$(A+BZ)^T - \lambda^{-1}I = K_\alpha(\lambda)\begin{pmatrix} I & 0 \\ 0 & D \end{pmatrix} L_\alpha(\lambda), \tag{3.3}$$

where $K_\alpha(\lambda)$ and $L_\alpha(\lambda) = (E_\alpha(\bar{\lambda}^{-1}))^*$ are rational matrix functions invertible at $\alpha$. For any rational matrix functions $F(\lambda)$ and $G(\lambda)$, we will write $F(\lambda) \sim G(\lambda)$ if there are rational matrix functions $K(\lambda)$ and $L(\lambda)$ invertible at $\alpha$ such that $F(\lambda) = K(\lambda)G(\lambda)L(\lambda)$.

Now, in view of Lemma 3.1, we have

$$\lambda F_e - G_e \sim \begin{pmatrix} \lambda I - (A+BZ) & 0 & -B \\ 0 & (A+BZ)^T - \lambda^{-1}I & 0 \\ 0 & -B^T & -(R+B^TXB) \end{pmatrix}.$$

By (3.2) and (3.3) we have further (for $\lambda$ in a neighborhood of $\alpha$)

$$\lambda F_e - G_e \sim \begin{pmatrix} I & 0 & 0 & 0 & B_{11} & B_{12} \\ 0 & D & 0 & 0 & B_{21} & B_{22} \\ 0 & 0 & I & 0 & 0 & 0 \\ 0 & 0 & 0 & D & 0 & 0 \\ 0 & 0 & C_{11} & C_{12} & S_{11} & S_{12} \\ 0 & 0 & C_{21} & C_{22} & S_{21} & S_{22} \end{pmatrix},$$

where we have written

$$-(E_\alpha(\lambda))^{-1}B = (B_{ij}), \quad -((E_\alpha(\bar{\lambda}^{-1}))^{-1}B)^* = (C_{ij}), \quad -(R+B^TXB) = (S_{ij}).$$

Since $\mathrm{rank}(\alpha I - A \;\; B) = n$, $\mathrm{rank}(\lambda I - (A+BZ) \;\; -B) = n$ at $\lambda = \alpha$. Therefore, at $\lambda = \alpha$,

$$\mathrm{rank}\begin{pmatrix} I & 0 & B_{11} & B_{12} \\ 0 & D & B_{21} & B_{22} \end{pmatrix} = n$$

and thus $\mathrm{rank}(B_{21} \;\; B_{22}) = s$. Note also that $E_\alpha(\bar{\lambda}^{-1}) = E_\alpha(\lambda)$ at $\lambda = \alpha$. We may then assume that $B_{21}$ and $C_{12}$ are invertible in a neighborhood of $\alpha$. Now we obtain by block elimination

$$\lambda F_e - G_e \sim W(\lambda) = \begin{pmatrix} I & 0 & 0 & 0 & 0 & 0 \\ 0 & D & 0 & 0 & I & 0 \\ 0 & 0 & I & 0 & 0 & 0 \\ 0 & 0 & 0 & D & 0 & 0 \\ 0 & 0 & 0 & I & V_{11} & V_{12} \\ 0 & 0 & 0 & 0 & V_{21} & V_{22} \end{pmatrix},$$

where

$$V(\lambda) = (V_{ij}) = \begin{pmatrix} C_{12}^{-1} & 0 \\ -C_{22}C_{12}^{-1} & I \end{pmatrix} (S_{ij}) \begin{pmatrix} B_{21}^{-1} & -B_{21}^{-1}B_{22} \\ 0 & I \end{pmatrix}$$

is a rational matrix function with $-V(\alpha) > 0$ (we have used $R + B^TXB > 0$ here). It is clear that no principal minors of $V(\lambda)$ are zero at $\alpha$.

All nonzero minors of order $i$ for $W(\lambda)$ have the form $(\lambda-\alpha)^l q(\lambda)$, where $l \geq 0$ and $\alpha$ is neither a zero nor a pole of the rational function $q(\lambda)$. For $2n+m-s+1 \leq i \leq 2n+m$, the smallest $l$ turns out to be $l_i = \Sigma_{j=1}^{s+i-2n-m} 2k_j$. For $1 \leq i \leq 2n+m-s$, the smallest $l$ is $l_i = 0$. By the Binet-Cauchy formula (see [19], for example), we can see that $(\lambda-\alpha)^{l_i}$ is also the greatest common divisor (of the form $(\lambda-\alpha)^l$) of all minors of order $i$ for $\lambda F_e - G_e$. Thus the elementary divisors of $\lambda F_e - G_e$ corresponding to $\alpha$ are $(\lambda-\alpha)^{2k_1}, \ldots, (\lambda-\alpha)^{2k_s}$. This proves the "only if" part of the theorem. The "if" part follows readily from the "only if" part. ∎

Corollary 3.4. *If the conditions of Theorem* 1.1 *are satisfied and* $\mathcal{R}'_{X_+}$ *is not invertible, then* $\mathcal{S} = \mathcal{N} \oplus \mathcal{M}$ *if and only if all the elementary divisors of* $\lambda F_e - G_e$ *corresponding to the eigenvalues on the unit circle are of degree two.*

A previous result of the same nature as Theorem 3.3 can be found in [26]. That result is applicable to the DARE (1.1) with $C = 0$, $R > 0$, and $Q \geq 0$.

**4. Convergence rate of the Newton method.** 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 DARE (1.1), we start the Newton iteration with the symmetric matrix $X_0$ obtained from the Stein equation (1.5). By Theorem 1.3, the Newton sequence is well-defined and converges to $X_+$. The following result shows there is some possibility of quadratic convergence.

Lemma 4.1. *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 $\tilde{X}_i = X_i - X_+$, $i = 0, 1, \ldots$, and let $L_+ = (R + B^T X_+ B)^{-1}(C + B^T X_+ A)$ (thus $A_+ = A - BL_+$). We have (see [18, p. 314])

$$\tilde{X}_i - A_+^T \tilde{X}_i A_+ = (L_+ - L_i)^T (R + B^T X_i B)(L_+ - L_i)$$

and $\|L_+ - L_i\| = O(\|\tilde{X}_{i-1}\|)$. We also have

$$\begin{aligned} L_+ - L_{i+1} &= \{(R + B^T X_+ B)^{-1} - (R + B^T X_i B)^{-1}\}(C + B^T X_+ A) \\ &\quad -(R + B^T X_i B)^{-1} B^T \tilde{X}_i A \\ &= (R + B^T X_i B)^{-1} B^T \tilde{X}_i B L_+ - (R + B^T X_i B)^{-1} B^T \tilde{X}_i A \\ &= -(R + B^T X_+ B)^{-1} B^T \tilde{X}_i A_+ + O(\|\tilde{X}_i\|^2), \end{aligned}$$

where we have written $O(\|\tilde{X}_i\|^2)$ for a term $W(X_i)$ satisfying $\|W(X_i)\| = O(\|\tilde{X}_i\|^2)$. Now, in view of (1.1) and (1.7),

$$\begin{aligned} \mathcal{R}(X_i) &= \mathcal{R}(X_i) - \mathcal{R}(X_+) \\ &= -\tilde{X}_i + A^T \tilde{X}_i A - (C + B^T X_i A)^T L_{i+1} + (C + B^T X_+ A)^T L_+ \\ &= -\tilde{X}_i + A_+^T \tilde{X}_i A_+ - A_+^T \tilde{X}_i A_+ + A^T \tilde{X}_i A \\ &\quad -\{(C + B^T X_i A)^T - (C + B^T X_+ A)^T\} L_{i+1} \\ &\quad +(C + B^T X_+ A)^T (L_+ - L_{i+1}) \\ &= O(\|\tilde{X}_{i-1}\|^2) - A_+^T \tilde{X}_i A_+ + A^T \tilde{X}_i A \\ &\quad -A^T \tilde{X}_i B L_+ + O(\|\tilde{X}_i\|^2) - (BL_+)^T \tilde{X}_i A_+ \\ &= O(\|\tilde{X}_{i-1}\|^2) + O(\|\tilde{X}_i\|^2). \end{aligned}$$

Thus for $i$ large enough,

$$\|\mathcal{R}(X_i)\| \leq c_1 \|\tilde{X}_{i-1}\|^2 + c_2 \|\tilde{X}_i\|^2 \tag{4.1}$$

for some constants $c_1$ and $c_2$.

On the other hand, for $i$ in $Q$ and large enough, we have as in [23]

$$\|\mathcal{R}(X_i)\| \geq (c_3(\theta^{-1}+1)^{-1} - c_4\|\tilde{X}_i\|)\|\tilde{X}_i\| \tag{4.2}$$

for some constants $c_3$ and $c_4$. Since $X_i \neq X_+$ for any $i$, we have by (4.1) and (4.2)

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

Therefore, we can find an $i_0$ such that $\|\tilde{X}_i\| \leq c\|\tilde{X}_{i-1}\|^2$ for all $i$ in $Q$ for which $i \geq i_0$. ∎

COROLLARY 4.2. *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 condition in Corollary 4.2 appears to be not easily satisfied. In fact, quadratic convergence has never been observed in our numerical experiments. We do not know if there are any examples of quadratic convergence in our setting. The next result describes what will happen if the convergence of the Newton iteration is not quadratic.

THEOREM 4.3. *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 will be an application of Theorem 1.5. We first establish some preliminary results.

LEMMA 4.4. *Let* $J$ *and* $P$ *be as in the proof of Theorem* 2.1. *Then*

$$\operatorname{rank}(\lambda I - J \quad P^{-1}B(R + B^TX_+B)^{-1}B^TP^{-T}) = n$$

*for every complex number* $\lambda$ *with* $|\lambda| \geq 1$.

*Proof.* In view of Theorem 4.5.6(b) of [18], we need only to show that the pair $(J, P^{-1}B(R + B^TX_+B)^{-1}B^TP^{-T})$ is d-stabilizable, or equivalently,

$$(A - BL_+, B(R + B^TX_+B)^{-1}B^T) \text{ is d-stabilizable.} \tag{4.3}$$

Since $(A, B)$ is d-stabilizable and $\operatorname{Im}(B(R + B^TX_+B)^{-1}B^T) = \operatorname{Im}B$, (4.3) follows from Lemma 4.5.3 of [18]. ∎

LEMMA 4.5 ([11, 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.*

We now set out to check the regularity condition needed in Theorem 1.5. For fixed $Z \in \mathcal{N}$, we consider the map $\mathcal{B}_Z : \mathcal{N} \to \mathcal{N}$ defined by

$$\mathcal{B}_Z(Y) = P_{\mathcal{N}}\mathcal{R}''_{X_+}(Z, Y).$$

By Lemma 2.2, we can write $Y = P^{-T}Y_JP^{-1}, Z = P^{-T}Z_JP^{-1}$ with $Y_J, Z_J \in \mathcal{N}_J$. Let $H_+ = B(R + B^TX_+B)^{-1}B^T$. We have by (1.3)

$$\begin{aligned}\mathcal{B}_Z(Y) &= -P_{\mathcal{N}}(A_+^TZH_+YA_+ + A_+^TYH_+ZA_+)\\ &= -P^{-T}P_{\mathcal{N}_J}(J^TZ_JD_+Y_JJ + J^TY_JD_+Z_JJ)P^{-1},\end{aligned}$$

where $D_+ = P^{-1}B(R+B^TX_+B)^{-1}B^TP^{-T}$, and $P_{\mathcal{N}_J}$ is the projection onto $\mathcal{N}_J$ parallel to $\mathcal{M}_J$. Let $Z_J = \mathrm{diag}(Z_1, \ldots, Z_p), Y_J = \mathrm{diag}(Y_1, \ldots, Y_p)$ and $\mathrm{diag}(D_1, \ldots, D_p)$ be the block diagonal of $D_+$. Let $\mathcal{S}_i = \mathcal{S}^{r_i}$ for $i = 1, 2$. We have further

$$\mathcal{B}_Z(Y) = -P^{-T}\mathrm{diag}(\mathcal{F}_{Z_1}(Y_1), \mathcal{F}_{Z_2}(Y_2), \ldots, \mathcal{F}_{Z_{p-1}}(Y_{p-1}), 0)P^{-1}, \tag{4.4}$$

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

$$\mathcal{F}_{Z_i}(Y_i) = Z_iD_iY_i + Y_iD_iZ_i, \quad i = 1, 2,$$
$$\mathcal{F}_{Z_i}(Y_i) = P_{\mathcal{S}_i}(G_i^T(Z_iD_iY_i + Y_iD_iZ_i)G_i), \quad i = 3, \ldots, p-1$$

with $P_{\mathcal{S}_i}$ being the projection onto $\mathcal{S}_i$ parallel to $\mathcal{T}_i$. The matrices $G_i$ were defined in (2.1) and (2.2).

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

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

LEMMA 4.6. *For $k = 1, 2, \ldots, p-1$, the set $\mathcal{U}_k$ has measure zero in $\mathcal{S}_k$.*

*Proof.* *Case* 1: $k = 1, 2$. We prove the result for $k = 1$, since the proof for $k = 2$ is similar. As in [11], we can show that $\mathcal{U}_1$ has measure zero in $\mathcal{S}_1$ unless $\det D_1 = 0$. Note that $D_+ = P^{-1}B(R+B^TX_+B)^{-1}B^TP^{-T}$ is symmetric positive semidefinite. If $\det D_1 = 0$, the first $r_1$ rows of $D_+$ would be linearly dependent by Lemma 4.5. Thus $\mathrm{rank}(-I - J \ \ D_+) < n$ , which contradicts Lemma 4.4.

*Case* 2: $k = 3, \ldots, p-1$. We will first find a more explicit expression for $\mathcal{F}_{Z_k}(Y_k)$. It is easily seen that

$$G_k = a_kI \otimes E_1 + b_kI \otimes E_2. \tag{4.5}$$

By Lemma 2.2, we can write

$$Y_k = M_s \otimes E_1 + M_a \otimes E_2, \quad Z_k = N_s \otimes E_1 + N_a \otimes E_2, \tag{4.6}$$

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

$$D_k = Q_s \otimes E_1 + Q_a \otimes E_2 + R_s \otimes E_3 + T_s \otimes E_4, \tag{4.7}$$

where $Q_s, R_s$ and $T_s$ are symmetric; $Q_a$ is anti-symmetric. Using (4.5)–(4.7) to expand $G_k^T(Z_kD_kY_k + Y_kD_kZ_k)G_k$, we find that each map $\mathcal{F}_{Z_k}$ has the same form as in the CARE case (see [11]). Thus, as in [11], each $\mathcal{U}_k$ has measure zero in $\mathcal{S}_k$ unless $\det(Q_s + iQ_a) = 0$.

To complete the proof, we need to show $\det(Q_s + iQ_a) \neq 0$. By Lemma 4.4 we have $\mathrm{rank}((a_k+b_ki)I - J \ \ D_+) = n$. Let $E(i, j(m))$ be the elementary matrix obtained from $I$ by adding $m$ times row $j$ to row $i$. Let $t_k = r_1 + \cdots + r_{k-1}$ and

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

Then

$$\operatorname{rank}(U((a_k + b_k i)I - J) \quad UD_+U^*) = n.$$

Since the $(t_k+1)$th, $(t_k+3)$th, $\ldots$, $(t_k+r_k-1)$th rows of the matrix $U((a_k+b_k i)I-J)$ are all zero, the corresponding rows of the Hermitian positive semidefinite matrix $UD_+U^*$ must be linearly independent. By Lemma 4.5, the principal submatrix (of order $r_k/2$) of $UD_+U^*$ contained in these rows must have a nonzero determinant. The principal submatrix is exactly $2(Q_s + iQ_a)$. Therefore $\det(Q_s + iQ_a) \neq 0$. □

LEMMA 4.7. *If $\mathcal{S} = \mathcal{N} \oplus \mathcal{M}$ then*

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

*has measure zero in $\mathcal{N}$. In particular, the regularity condition holds.*

*Proof.* The result follows from (4.4) and Lemma 4.6, as in [11]. □

*Proof of Theorem* 4.3. Note that the map $\mathcal{R}$ can be extended to a smooth map on $\mathcal{S}$ without changing its values on a closed ball centered at $X_+$ and contained in $\mathcal{D}$. Now, as in [11], the proof can be completed by applying Theorem 1.3, Theorem 1.5, Corollary 4.2 and Lemma 4.7. □

When all elementary divisors of the closed-loop matrix corresponding to the eigenvalues on the unit circle are linear, we know from Theorem 4.3 that the convergence of the Newton iteration is either quadratic or linear with rate $\frac{1}{2}$. Quadratic convergence, however, has not been observed in numerical experiments when the closed-loop matrix has eigenvalues on the unit circle. The convergence has been observed to be linear with rate $\frac{1}{\sqrt[p]{2}}$, where $p$ is the highest degree of elementary divisors associated with eigenvalues on the unit circle. The next example gives a little theoretical support for the observation. A general theory for the case $p > 1$ would be a topic for future research.

*Example* 4.1. Consider the DARE (1.1) with $n = 2, m = 1$ and

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

Clearly $(A, B)$ is d-stabilizable and $X_+ = 0$ (0 is the unique almost stabilizing solution in this case. See Theorem 13.5.2 of [18], for example). Note that $(\lambda - 1)^2$ is the only elementary divisor of $A_+ = A$. The Newton sequence $\{X_i\}$ is well defined and we write for $i = 0, 1, \ldots,$

$$X_i = \begin{pmatrix} a_i & c_i \\ c_i & b_i \end{pmatrix}.$$

Since $A - B(R + B^T X_i B)^{-1}(C + B^T X_i A)$ is d-stable, we can deduce that $c_i \neq 0$. Since $X_i \geq 0$, we also have $a_i, b_i > 0$.

By (1.6)–(1.8), we find for $i = 0, 1, \ldots$

$$a_{i+1} = \frac{2a_i^2 + 3a_i c_i + 2c_i}{(2a_i - c_i + 4)a_i}, \tag{4.8}$$

$$b_{i+1} = \frac{((2 + a_i)a_{i+1} - a_i)c_i}{2(1 + a_i)^2}, \tag{4.9}$$

$$c_{i+1} = \frac{(1 + a_{i+1})c_i}{2(1 + a_i)}. \tag{4.10}$$

Since $X_i \to 0$, we get from (4.10)

$$\lim_{i\to\infty} \frac{c_{i+1}}{c_i} = \frac{1}{2}. \tag{4.11}$$

It follows from (4.8) that

$$\lim_{i\to\infty} \frac{c_i}{a_i} = 0. \tag{4.12}$$

It then follows from (4.9), (4.11) and (4.12) that $\lim_{i\to\infty} b_i/a_i = 0$. If the convergence of the Newton iteration is linear with rate $\mu$, then $\lim_{i\to\infty} a_{i+1}/a_i = \mu$. Now by (4.8) and (4.12),

$$\lim_{i\to\infty} \frac{a_{i+1}}{a_i} = \frac{1}{2}\left(1 + \lim_{i\to\infty} \frac{c_i}{a_i^2}\right). \tag{4.13}$$

If $\lim_{i\to\infty} c_i/a_i^2 = 0$, we would have $\lim_{i\to\infty} a_{i+1}/a_i = 1/2$ by (4.13) and further $\lim_{i\to\infty} c_i/a_i^2 = \infty$ by (4.11), which is a contradiction. Therefore, $\lim_{i\to\infty} c_i/a_i^2 \neq 0$. Thus we get from (4.11) that $\mu = 1/\sqrt{2}$.

The above example can also serve to show that $X_0 \geq X_1$ is generally not true if $X_0$ is not determined by (1.5). Take

$$X_0 = \begin{pmatrix} \epsilon^\alpha & \epsilon \\ \epsilon & \delta \end{pmatrix}$$

with $\alpha > 1, 0 < \epsilon < 1$, and $\delta$ real. It is easily checked that $A - B(R + B^T X_0 B)^{-1}(C + B^T X_0 A)$ is d-stable. We see from (4.8) that $a_1 \sim 0.5\epsilon^{1-\alpha}$ as $\epsilon \to 0$. Thus $X_0 \geq X_1$ cannot be true for small $\epsilon$. As $\epsilon$ and $\delta$ go to zero, we have $\|X_0 - X_+\| \to 0$, but $\|X_1 - X_+\| \to \infty$.

**5. Using the double Newton step.** We have shown that the convergence of Newton's method is either quadratic or linear with rate $\frac{1}{2}$, provided that the unimodular eigenvalues of the closed-loop matrix are all semi-simple. Quadratic convergence has not been observed in our numerical experiments. Therefore, we should always be prepared for linear convergence. In this section we will show that the efficiency of the Newton iteration (when it is linearly convergent) can be improved significantly if a double Newton step is used at the right time. However, since the second derivative of the Riccati function is no longer constant, the improvement will not be as dramatic as in the CARE case.

LEMMA 5.1. *In the setting of Theorems* 1.1 *and* 1.3*, assume that $X_k$ is close enough to $X_+$ with $X_k - X_+ \in \mathcal{N}$ and that $\|(\mathcal{R}'_{X_k})^{-1}\| \leq c\|X_k - X_+\|^{-1}$ with $c$ independent of $k$. If $Y_{k+1} = X_k - 2(\mathcal{R}'_{X_k})^{-1}\mathcal{R}(X_k)$, then $\|Y_{k+1} - X_+\| \leq c_1\|X_k - X_+\|^2$ for some constant $c_1$ independent of $k$.*

*Proof.* By Taylor's Theorem,

$$\mathcal{R}(X_k) = \frac{1}{2}\mathcal{R}''_{X_+}(X_k - X_+, X_k - X_+) + O(\|X_k - X_+\|^3),$$

and then

$$\begin{aligned}\mathcal{R}'_{X_k}(X_k - X_+) &= \mathcal{R}''_{X_+}(X_k - X_+, X_k - X_+) + O(\|X_k - X_+\|^3) \\ &= 2\mathcal{R}(X_k) + O(\|X_k - X_+\|^3).\end{aligned}$$

Thus

$$X_k - X_+ = 2(\mathcal{R}'_{X_k})^{-1}\mathcal{R}(X_k) + O(\|X_k - X_+\|^2). \qquad \square$$

When the direct sum condition is satisfied and the convergence of the Newton sequence $\{X_k\}$ is not quadratic, we have $\|(\mathcal{R}'_{X_k})^{-1}\| \le c\|X_k - X_+\|^{-1}$ for all $k$ (cf. Theorem 4.3). Moreover, the error $X_k - X_+$ will be dominated by its $\mathcal{N}$-component for large $k$. A much better approximate solution can then be obtained by applying the double Newton step. More precisely, we have the following result.

THEOREM 5.2. *Assume $\mathcal{S} = \mathcal{N} \oplus \mathcal{M}$ and the convergence of the Newton iteration is not quadratic. If for some $k$, $\|X_k - X_+\|$ is small enough and $\|P_{\mathcal{M}}(X_k - X_+)\| \le \epsilon\|P_{\mathcal{N}}(X_k - X_+)\|$ with $\epsilon$ sufficiently small, and $Y_{k+1} = X_k - 2(\mathcal{R}'_{X_k})^{-1}\mathcal{R}(X_k)$, then $\|Y_{k+1} - X_+\| \le c_1\epsilon + c_2\|X_k - X_+\|^2$ for some constants $c_1$ and $c_2$ independent of $\epsilon$ and $k$.*

*Proof.* The result follows from Lemma 5.1 and the argument used in the proof of [11, Theorem 3.2]. $\square$

In contrast to the CARE case, the error estimate for $Y_{k+1}$ contains the term $c_2\|X_k - X_+\|^2$. For a problem which produces a large $c_2$, the error $\|Y_{k+1} - X_+\|$ will be small only when $\|X_k - X_+\|$ is already sufficiently small. In this case the double Newton step will be useful only at a very late stage of the iteration.

In the CARE case (as described in [11]), the iterate produced by the double Newton step is at least almost stabilizing (see the discussions in [2]). For the DARE case, however, it can happen that the matrix $Y_{k+1}$ in Theorem 5.2 is neither stabilizing nor almost stabilizing.

*Example* 5.1 (cf. [18, Example 13.2.1]). Consider the DARE (1.1) with $Q = C = 0$ and $A = B = R = I$. Clearly $(A, B)$ is d-stabilizable and $X_+ = 0$. All eigenvalues of the closed-loop matrix are on the unit circle and semi-simple. For $L_0 = I$, the Newton iterates are found to be

$$X_k = \frac{1}{2^{k+1} - 1} I, \quad k = 0, 1, \ldots.$$

Thus, the convergence is linear with rate 1/2. If we compute $Y_{k+1}$ as in Theorem 5.2, we get

$$Y_{k+1} = -\frac{1}{(2^{k+1} - 1)(2^{k+2} - 1)} I.$$

Although $Y_{k+1}$ is much more accurate than $X_{k+1}$ for large $k$, it is neither stabilizing nor almost stabilizing.

The double Newton step is useful in that it can significantly improve the accuracy of the current Newton iterate and thus find more correct digits of the exact solution. The potential problem of getting a slightly non-stabilizing approximate solution is not our concern here. Even if an exact solution with infinite number of decimals is known, we will probably get a slightly non-stabilizing approximate solution by keeping only a finite number of decimals.

Theorem 5.2 suggests the following modification of the Newton method.

ALGORITHM (Modified Newton method for DARE).

1. Choose a matrix $L_0$ for which $A - BL_0$ is d-stable.
2. Find $X_0$ from (1.5).
3. 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 $\epsilon$ is a prescribed accuracy. The equation $\mathcal{R}'_{X_k}(H) = \mathcal{R}(X_k)$ can be rewritten as a Stein equation $H - A_{k+1}^T H A_{k+1} = -\mathcal{R}(X_k)$, which can be solved efficiently by a variation of the Bartels/Stewart algorithm [1]. See also [20]. According to Theorem 5.2, the double Newton step will be efficient only when the current iterate is already reasonably close to the solution. This is a major difference between the CARE case and the DARE case. We may try the double Newton step only when the norm of the residual is small enough (less than $\sqrt{\epsilon}$, say) and save a little more computational work. In the above algorithm, 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.

**6. Numerical results.** In this section we give two simple examples to illustrate the performance of the modified Newton method.

*Example* 6.1. We consider the DARE (1.1) with $n = m = 2$ and

$$A = \begin{pmatrix} 0 & -1 \\ 0 & 2 \end{pmatrix}, \ B = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}, \ C = 0, \ Q = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \ R = \begin{pmatrix} 4 & 2 \\ 2 & 1 \end{pmatrix}.$$

Note that $A$ and $R$ are both singular. It can be easily verified that $X_+ = \mathrm{diag}(1, 0)$ is the only solution of the DARE and the closed-loop eigenvalues are 0 and 1. We take $L_0 = \mathrm{diag}(0, 2)$ so that $A_0 = A - BL_0$ is d-stable, and apply the modified Newton method with $\epsilon = 10^{-10}$. The numerical results are recorded in Table 6.1. The last iterate is produced by the double Newton step.

TABLE 6.1
*Performance of the modified Newton method for Example* 6.1

| $k$ | $\|X_k - X_+\|_1$ | $\|\mathcal{R}(X_k)\|_1$ |
|---|---|---|
| 0 | $0.5000D+01$ | $0.4545D+01$ |
| 1 | $0.4167D+00$ | $0.1894D+00$ |
| 2 | $0.1471D+00$ | $0.3342D-01$ |
| 3 | $0.6410D-01$ | $0.7284D-02$ |
| 4 | $0.3012D-01$ | $0.1711D-02$ |
| 5 | $0.1462D-01$ | $0.4153D-03$ |
| 6 | $0.7205D-02$ | $0.1023D-03$ |
| 7 | $0.3577D-02$ | $0.2540D-04$ |
| 8 | $0.1782D-02$ | $0.6328D-05$ |
| 9 | $0.3170D-05$ | $0.2009D-10$ |

*Example* 6.2. We consider the DARE (1.1) with $n = m = 8$ and

$$A = \mathrm{diag}\left( \begin{pmatrix} -1 & & \\ & 1 & \\ & & 1 \end{pmatrix}, \begin{pmatrix} \frac{\sqrt{3}}{2} & \frac{1}{2} \\ -\frac{1}{2} & \frac{\sqrt{3}}{2} \end{pmatrix}, \begin{pmatrix} \frac{1}{2} & 1 & \\ & \frac{1}{2} & 1 \\ & & \frac{1}{2} \end{pmatrix} \right),$$

$$B = \begin{pmatrix} 1 & & & \\ 1 & 1 & & \\ & \ddots & \ddots & \\ & & 1 & 1 \end{pmatrix}, \quad C = 0, \quad Q = 0, \quad R = I.$$

For this example, $X_+ = 0$ and the closed-loop eigenvalues are those of $A$. The unimodular eigenvalues are all semi-simple. We take $L_0 = \mathrm{diag}(-1, 1, 1, 1, 1, 0.1, 0.1, 0.1)$ so that $A_0 = A - BL_0$ is d-stable, and apply the modified Newton method with $\epsilon = 10^{-10}$. The results are recorded in Table 6.2. Again, the last iterate is produced by the double Newton step.

TABLE 6.2
*Performance of the modified Newton method for Example* 6.2

| $k$ | $\|X_k - X_+\|_1$ | $\|\mathcal{R}(X_k)\|_1$ |
|---|---|---|
| 0 | $0.2344D+02$ | $0.2327D+02$ |
| 1 | $0.2273D+01$ | $0.1855D+01$ |
| 2 | $0.3733D+00$ | $0.1766D+00$ |
| 3 | $0.1419D+00$ | $0.2444D-01$ |
| 4 | $0.6291D-01$ | $0.6681D-02$ |
| 5 | $0.2987D-01$ | $0.1611D-02$ |
| 6 | $0.1458D-01$ | $0.3826D-03$ |
| 7 | $0.7204D-02$ | $0.9472D-04$ |
| 8 | $0.3581D-02$ | $0.2357D-04$ |
| 9 | $0.1785D-02$ | $0.5877D-05$ |
| 10 | $0.8914D-03$ | $0.1467D-05$ |
| 11 | $0.4454D-03$ | $0.3666D-06$ |
| 12 | $0.2226D-03$ | $0.9161D-07$ |
| 13 | $0.3986D-07$ | $0.1312D-10$ |

In both examples, the convergence of the Newton method is linear and the final double Newton step reduces the error significantly. We have by (4.1) that $\|\mathcal{R}(X_k)\| \leq c\|X_k - X_+\|^2$, where $X_k$ are the Newton iterates. The last iterate, $Y_l$, is produced by the double Newton step and $\|\mathcal{R}(Y_l)\| \leq c\|Y_l - X_+\|^2$ is not necessarily true. Typically, for $l$ large enough, the error $\|Y_l - X_+\|$ is comparable to $\|\mathcal{R}(X_{l-1})\|$.

**Acknowledgements.** The author wishes to thank Dr. Peter Lancaster for his many helpful suggestions. Careful comments by Dr. Peter Benner are also gratefully acknowledged.

## REFERENCES

[1] R. H. BARTELS AND G. W. STEWART, *Solution of the matrix equation* $AX + XB = C$, Comm. ACM, 15 (1972), pp. 820–826.

[2] P. BENNER, *Contributions to the Numerical Solution of Algebraic Riccati Equations and Related Eigenvalue Problems*, Logos-Verlag, Berlin, 1997.

[3] P. BENNER, A. J. LAUB, AND V. MEHRMANN, *A collection of benchmark examples for the numerical solution of algebraic Riccati equations* II: *discrete-time case*, Technical Report SPC 95-23, Fakultät für Mathematik, Technische Universität Chemnitz-Zwickau, FRG, 1995.

[4] S. W. CHAN, G. C. GOODWIN, AND K. S. SIN, *Convergence properties of the Riccati difference equation in optimal fitering of nonstabilizable systems*, IEEE Trans. Autom. Control, 29 (1984), pp. 110–118.

[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), pp. 296–314.

[6] D. W. DECKER AND C. T. KELLEY, *Newton's Method at singular points* I, SIAM J. Numer. Anal., 17 (1980), pp. 66–70.

[7] ———, *Convergence acceleration for Newton's method at singular points*, SIAM J. Numer. Anal., 19 (1982), pp. 219–229.

[8] A. EMAMI-NAEINI AND G. F. FRANKLIN, *Comments on "On the numerical solution of the discrete-time algebraic Riccati equation"*, IEEE Trans. Autom. Control, 25 (1980), pp. 1015–1016.

[9] I. Gohberg, P. Lancaster, and L. Rodman, *Matrix Polynomials*, Academic Press, New York, 1982.

[10] A. Griewank and M. R. Osborne, *Analysis of Newton's method at irregular singularities*, SIAM J. Numer. Anal., 20 (1983), pp. 747–773.

[11] C.-H. Guo and P. Lancaster, *Analysis and modification of Newton's method for algebraic Riccati equations*, Math. Comp., 67 (1998), to appear.

[12] V. Hernández and A. Pastor, *On the Kleinman iteration for periodic nonstabilizable systems*, in Proceedings of the European Control Conference ECC97, Paper 946, 1997.

[13] G. A. Hewer, *An iterative technique for the computation of the steady-state gains for the discrete optimal regulator*, IEEE Trans. Autom. Control, 16 (1971), pp. 382–384.

[14] V. Ionescu and M. Weiss, *On computing the stabilizing solution of the discrete-time Riccati equation*, Linear Algebra Appl., 174 (1992), pp. 229–238.

[15] L. V. Kantorovich and G. P. Akilov, *Functional Analysis in Normed Spaces*, Pergamon, New York, 1964.

[16] C. T. Kelley, *A Shamanskii-like acceleration scheme for nonlinear equations at singular roots*, Math. Comp., 47 (1986), pp. 609–623.

[17] C. T. Kelley and R. Suresh, *A new acceleration method for Newton's method at singular points*, SIAM J. Numer. Anal., 20 (1983), pp. 1001–1009.

[18] P. Lancaster and L. Rodman, *Algebraic Riccati Equations*, Clarendon Press, Oxford, 1995.

[19] P. Lancaster and M. Tismenetsky, *The Theory of Matrices*, Second Edition, Academic Press, Orlando, 1985.

[20] V. L. Mehrmann, *The Autonomous Linear Quadratic Control Problem*, Lecture Notes in Control and Information Sciences, Vol. 163, Springer-Verlag, Berlin, 1991.

[21] J. M. Ortega and W. C. Rheinboldt, *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York, 1970.

[22] A. C. M. Ran and R. Vreugdenhil, *Existence and comparison theorems for algebraic Riccati equations for continuous- and discrete-time systems*, Linear Algebra Appl., 99 (1988), pp. 63–83.

[23] G. W. Reddien, *On Newton's method for singular problems*, SIAM J. Numer. Anal., 15 (1978), pp. 993–996.

[24] V. Sima, *An efficient Schur method to solve the stabilizing problem*, IEEE Trans. Autom. Control, 26 (1981), pp. 724–725.

[25] P. Van Dooren, *A generalized eigenvalue approach for solving Riccati equations*, SIAM J. Sci. Comput., 2 (1981), pp. 121–135.

[26] H. K. Wimmer, *Normal forms of symplectic pencils and the discrete algebraic Riccati equation*, Linear Algabra Appl., 147 (1991), pp. 411–440.