Iterative Solution of a Nonsymmetric Algebraic Riccati Equation

Guo, Chun-Hua and Higham, Nicholas J.

2007

MIMS EPrint: **2005.48**

ISSN 1749-9097

SIAM J. MATRIX ANAL. APPL.
Vol. 29, No. 2, pp. 396–412

ITERATIVE SOLUTION OF A NONSYMMETRIC ALGEBRAIC RICCATI EQUATION*

CHUN-HUA GUO† AND NICHOLAS J. HIGHAM‡

Abstract. We study the nonsymmetric algebraic Riccati equation whose four coefficient matrices are the blocks of a nonsingular M-matrix or an irreducible singular M-matrix M. The solution of practical interest is the minimal nonnegative solution. We show that Newton's method with zero initial guess can be used to find this solution without any further assumptions. We also present a qualitative perturbation analysis for the minimal solution, which is instructive in designing algorithms for finding more accurate approximations. For the most practically important case, in which M is an irreducible singular M-matrix with zero row sums, the minimal solution is either stochastic or substochastic and the Riccati equation can be transformed into a unilateral matrix equation by a procedure of Ramaswami. The minimal solution of the Riccati equation can then be found by computing the minimal nonnegative solution of the unilateral equation using the Latouche–Ramaswami algorithm. When the minimal solution of the Riccati equation is stochastic, we show that the Latouche–Ramaswami algorithm, combined with a shift technique suggested by He, Meini, and Rhee, is breakdown-free and is able to find the minimal solution more efficiently and more accurately than the algorithm without a shift. When the minimal solution of the Riccati equation is substochastic, we show how the substochastic minimal solution can be found by computing the stochastic minimal solution of a related Riccati equation of the same type.

Key words. nonsymmetric algebraic Riccati equation, M-matrix, minimal nonnegative solution, perturbation analysis, Newton's method, Latouche–Ramaswami algorithm, shifts

AMS subject classifications. 15A24, 15A48, 65F30, 65H10

DOI. 10.1137/050647669

1. Introduction.

We consider the nonsymmetric algebraic Riccati equation (or NARE)

$$\mathcal{R}(X) = XCX - XD - AX + B = 0, \tag{1.1}$$

where A, B, C, D are real matrices of sizes $m \times m, m \times n, n \times m, n \times n$, respectively, and we assume throughout that

$$M = \begin{bmatrix} D & -C \\ -B & A \end{bmatrix} \tag{1.2}$$

is a nonsingular M-matrix or an irreducible singular M-matrix. Some relevant definitions are as follows. For any matrices $A, B \in \mathbb{R}^{m\times n}$, we write $A \geq B$ $(A > B)$ if $a_{ij} \geq b_{ij} (a_{ij} > b_{ij})$ for all i, j. A real square matrix A is called a Z-matrix if all its off-diagonal elements are nonpositive. It is clear that any Z-matrix A can be written as $sI - B$ with $B \geq 0$. A Z-matrix A is called an M-matrix if $s \geq \rho(B)$, where $\rho(\cdot)$ is the spectral radius; it is a singular M-matrix if $s = \rho(B)$ and a nonsingular M-matrix if $s > \rho(B)$.

*Received by the editors December 15, 2005; accepted for publication (in revised form) by Q. Ye October 3, 2006; published electronically March 22, 2007. This work was supported in part by a Royal Society-Wolfson Research Merit Award to the second author.

http://www.siam.org/journals/simax/29-2/64766.html

†Department of Mathematics and Statistics, University of Regina, Regina, SK S4S 0A2, Canada (chguo@math.uregina.ca, http://www.math.uregina.ca/∼chguo/). The work of this author was supported in part by a grant from the Natural Sciences and Engineering Research Council of Canada.

‡School of Mathematics, The University of Manchester, Sackville Street, Manchester, M60 1QD, UK (higham@ma.man.ac.uk, http://www.ma.man.ac.uk/∼higham/).

The NARE (1.1) has applications in transport theory and Markov models [20, 27, 28]. The solution of practical interest is the minimal nonnegative solution. The equation has attracted much attention recently [1, 4, 10, 11, 14, 16, 17, 18, 21, 24, 26].

For application to Markov models, the case of primary interest is the one where M is an irreducible singular M-matrix with zero row sums. When M is an irreducible singular M-matrix, we have $M = \rho(N)I - N$ for some irreducible nonnegative matrix N. Thus, by applying the Perron–Frobenius theorem to N, there are positive vectors $u_1, v_1 \in \mathbb{R}^n$ and $u_2, v_2 \in \mathbb{R}^m$ such that

$$M(v_1^T \;\, v_2^T)^T = 0, \quad (u_1^T \;\, u_2^T)M = 0, \tag{1.3}$$

and the vectors $(v_1^T \;\, v_2^T)$ and $(u_1^T \;\, u_2^T)$ are each unique up to a scalar multiple.

Since M is a nonsingular M-matrix or an irreducible singular M-matrix, we have $B, C \geq 0$, and A and D are nonsingular M-matrices (see [10], for example). Therefore, the matrix $I \otimes A + D^T \otimes I$ is also a nonsingular M-matrix, where $\otimes$ is the Kronecker product. Some properties of the NARE (1.1) are summarized below. See [10, 11, 13] for more details.

THEOREM 1.1. *Assume that M is a nonsingular M-matrix or an irreducible singular M-matrix. Then the NARE* (1.1) *has a minimal nonnegative solution S. If M is irreducible, then $S > 0$ and $A - SC$ and $D - CS$ are irreducible M-matrices. If M is a nonsingular M-matrix, then $A - SC$ and $D - CS$ are nonsingular M-matrices. If M is a nonsingular M-matrix or an irreducible singular M-matrix with $u_1^T v_1 \neq u_2^T v_2$, then*

$$M_S = I \otimes (A - SC) + (D - CS)^T \otimes I$$

is a nonsingular M-matrix. If M is an irreducible singular M-matrix with $u_1^T v_1 = u_2^T v_2$, then M_S is an irreducible singular M-matrix.

We will also need the dual equation of (1.1):

$$YBY - YA - DY + C = 0. \tag{1.4}$$

This equation has the same type as (1.1): the matrix

$$\begin{bmatrix} A & -B \\ -C & D \end{bmatrix}$$

is a nonsingular M-matrix or an irreducible singular M-matrix if and only if the matrix M has the same property. The minimal nonnegative solution of (1.4) is denoted by $\widehat{S}$.

A number of numerical methods have been studied for finding the minimal solution S, some of which require additional assumptions on the NARE (1.1). In particular, a class of basic fixed-point iterations has been studied in [10] and [16]. The Schur method has been studied in [10] and a modified Schur method is given in [14]. These methods are applicable without further assumptions on (1.1). Newton's method has also been studied in [10] and [16], where convergence of the Newton sequence $\{X_k\}$, with $X_0 = 0$, to the minimal solution S has been established under the additional assumption that

$$B, C \neq 0, \quad (I \otimes A + D^T \otimes I)^{-1}\text{vec}B > 0. \tag{1.5}$$

Here, the vec operator stacks the columns of a matrix into one long vector. When M is irreducible, we have $B, C \neq 0$. However, the condition $(I \otimes A + D^T \otimes I)^{-1}\text{vec}B > 0$

is not guaranteed by the irreducibility of M, as is shown in [10]. The question then arises as to whether (1.5) is necessary for the convergence of the Newton iteration. Our first contribution in this paper is a proof of convergence without this additional condition.

When M is an irreducible singular M-matrix and $u_1^T v_1 = u_2^T v_2$, the matrix M_S is a singular M-matrix. In this case, Newton's method has a singular Jacobian at the solution, and thus we cannot expect to find an accurate solution by the Newton iteration in finite precision arithmetic. A modified Schur method has been proposed in [14] to find a more accurate solution when $u_1^T v_1 \approx u_2^T v_2$. Another approach is to transform the bilateral equation (1.1) into a unilateral equation and use methods based on cyclic reduction, including the Latouche–Ramaswami (LR) algorithm [23], in combination with a shift technique proposed in [19].

The design of numerical methods for finding the minimal solution with higher accuracy is related to the perturbation behavior of the minimal solution. The minimal solution S is a function of M in (1.2). If the matrix M is perturbed to $\widetilde{M}$, which is always assumed to be again a nonsingular M-matrix or an irreducible singular M-matrix, and $\widetilde{S}$ is the new minimal solution, we would like to know the relation between $\|\widetilde{S} - S\|$ and $\|\widetilde{M} - M\|$, where $\|\cdot\|$ is any matrix norm. Our second contribution is to prove the following.

- If M is a nonsingular M-matrix or an irreducible singular M-matrix with $u_1^T v_1 \neq u_2^T v_2$, then there exist constants $\gamma > 0$ and $\epsilon > 0$ such that $\|\widetilde{S} - S\| \leq \gamma \|\widetilde{M} - M\|$ for all $\widetilde{M}$ with $\|\widetilde{M} - M\| < \epsilon$.
- If M is an irreducible singular M-matrix with $u_1^T v_1 = u_2^T v_2$, then there exist constants $\gamma > 0$ and $\epsilon > 0$ such that
  (a) $\|\widetilde{S} - S\| \leq \gamma \|\widetilde{M} - M\|^{1/2}$ for all $\widetilde{M}$ with $\|\widetilde{M} - M\| < \epsilon$;
  (b) $\|\widetilde{S} - S\| \leq \gamma \|\widetilde{M} - M\|$ for all *singular* $\widetilde{M}$ with $\|\widetilde{M} - M\| < \epsilon$.

This result tells us that to achieve high accuracy for S when M is an irreducible singular M-matrix with $u_1^T v_1 \approx u_2^T v_2$, it is necessary to use the singularity of M in the design of algorithms. Otherwise, we can only expect to achieve an accuracy of $O(\epsilon_m^{1/2})$, where ϵ_m is the machine epsilon. The modified Schur method in [14] and the methods using a shift technique in [4] and [14] all use the singularity of M. However, the use of the shift technique creates a new problem: it is not clear whether the resulting algorithm may break down, although quadratic convergence is guaranteed if no breakdown occurs. Our third contribution is to show that the (simplified) LR algorithm with a shift technique, presented in [14], is breakdown-free.

2. Convergence of Newton's method. The Riccati function $\mathcal{R}$ is a mapping from $\mathbb{R}^{m\times n}$ into itself. The Fréchet derivative of $\mathcal{R}$ at a matrix X is a linear map $\mathcal{R}'_X : \mathbb{R}^{m\times n} \to \mathbb{R}^{m\times n}$ given by

$$\mathcal{R}'_X(Z) = -\big((A - XC)Z + Z(D - CX)\big). \tag{2.1}$$

The Newton method for the solution of (1.1) is

$$X_{i+1} = X_i - (\mathcal{R}'_{X_i})^{-1}\mathcal{R}(X_i), \quad i = 0, 1, \dots, \tag{2.2}$$

where the maps $\mathcal{R}'_{X_i}$ all need to be nonsingular. In view of (2.1), the iteration (2.2) is equivalent to

$$(A - X_iC)X_{i+1} + X_{i+1}(D - CX_i) = B - X_iCX_i, \quad i = 0, 1, \dots . \tag{2.3}$$

We will need the following well-known result (see [2], for example).

THEOREM 2.1. *For a Z-matrix A, the following are equivalent:*

(a) *A is a nonsingular M-matrix.*
(b) $A^{-1} \geq 0$.
(c) *$Av > 0$ for some vector $v > 0$.*
(d) *All eigenvalues of A have positive real parts.*

The equivalence of (a) and (c) in Theorem 2.1 implies the next result.

LEMMA 2.2. *Let A be a nonsingular M-matrix. If $B \geq A$ is a Z-matrix, then B is also a nonsingular M-matrix.*

We can now give a proof of convergence of the Newton iteration that does not require the assumption (1.5) made in [10].

THEOREM 2.3. *Let S be the minimal nonnegative solution of* (1.1)*. Then for the Newton iteration* (2.3) *with $X_0 = 0$, the sequence $\{X_i\}$ is well defined, $X_k \leq X_{k+1} \leq S$ for all $k \geq 0$, and $\lim_{i\to\infty} X_i = S$.*

Proof. Throughout the proof, we use the notation

$$M_X = I \otimes (A - XC) + (D - CX)^T \otimes I$$

for a given matrix X (this notation is consistent with the notation M_S already used in Theorem 1.1). Since S is a solution of (1.1),

$$SCS - SD - AS + B = 0. \tag{2.4}$$

For the Newton iteration (2.3) with $X_0 = 0$, we have $AX_1 + X_1D = B$, which is equivalent to

$$(I \otimes A + D^T \otimes I)\mathrm{vec}X_1 = \mathrm{vec}B. \tag{2.5}$$

Since $I \otimes A + D^T \otimes I$ is a nonsingular M-matrix, Theorem 2.1(b) and (2.5) imply $\mathrm{vec}X_1 \geq 0$, i.e., $X_1 \geq 0$.

We first assume that M is a nonsingular M-matrix, and we will prove by induction that

$$X_k \leq X_{k+1}, \quad X_k \leq S, \quad M_{X_k} \text{ is a nonsingular } M\text{-matrix} \tag{2.6}$$

for $k \geq 0$. It is clear that (2.6) is true for $k = 0$. We now assume that (2.6) is true for $k = i \geq 0$. By (2.3) and (2.4) we have

$$\begin{aligned}(A - X_iC)(X_{i+1} - S) + (X_{i+1} - S)(D - CX_i) \\ = B - X_iCX_i - AS + X_iCS - SD + SCX_i \\ = -(S - X_i)C(S - X_i).\end{aligned} \tag{2.7}$$

Since $X_i \leq S$ and M_{X_i} is a nonsingular M-matrix, it follows from Theorem 2.1(b) and (2.7) that $X_{i+1} \leq S$. Since M_S is a nonsingular M-matrix by Theorem 1.1, it follows from Lemma 2.2 that $M_{X_{i+1}}$ is a nonsingular M-matrix. By (2.3),

$$\begin{aligned}&(A - X_{i+1}C)X_{i+1} + X_{i+1}(D - CX_{i+1}) \\ &= \big(A - X_iC - (X_{i+1} - X_i)C\big)X_{i+1} + X_{i+1}\big(D - CX_i - C(X_{i+1} - X_i)\big) \\ &= B - X_iCX_i - (X_{i+1} - X_i)CX_{i+1} - (X_i + X_{i+1} - X_i)C(X_{i+1} - X_i) \\ &= B - X_{i+1}CX_{i+1} - (X_{i+1} - X_i)C(X_{i+1} - X_i).\end{aligned} \tag{2.8}$$

By (2.8) and (2.3),

$$\begin{aligned}&(A - X_{i+1}C)(X_{i+1} - X_{i+2}) + (X_{i+1} - X_{i+2})(D - CX_{i+1})\\ &\quad = -(X_{i+1} - X_i)C(X_{i+1} - X_i) \le 0.\end{aligned}$$

Therefore, $X_{i+1} \le X_{i+2}$. We have thus proved that (2.6) is true for $k = i + 1$. Hence (2.6) is true for all $k \ge 0$ by induction.

We now assume that M is an irreducible singular M-matrix. Then $S > 0$ by Theorem 1.1. Thus, the statement

$$X_k \le X_{k+1}, \quad X_k < S, \quad M_{X_k} \text{ is a nonsingular } M\text{-matrix} \tag{2.9}$$

is true for $k = 0$. Assume that (2.9) is true for $k = i \ge 0$. Then, by (2.7) we get $X_{i+1} < S$. It follows from (2.8) and (2.4) that

$$\begin{aligned}&(A - X_{i+1}C)(X_{i+1} - S) + (X_{i+1} - S)(D - CX_{i+1})\\ &\quad = -(X_{i+1} - X_i)C(X_{i+1} - X_i) - (X_{i+1} - S)C(X_{i+1} - S) < 0.\end{aligned}$$

Therefore, $M_{X_{i+1}}\mathrm{vec}(S - X_{i+1}) > 0$. Thus $M_{X_{i+1}}$ is a nonsingular M-matrix by Theorem 2.1(c). It follows as before that $X_{i+1} \le X_{i+2}$. So (2.9) is true for $k = i + 1$, and hence for all $k \ge 0$ by induction.

Therefore, in both cases, the Newton sequence X_k is well defined, monotonically increasing, and bounded above by S. Let $\lim_{k\to\infty} X_k = X_*$. Then X_* is a nonnegative solution of (1.1) by (2.3). Since $X_* \le S$ and S is minimal, we have $X_* = S$. ∎

3. Perturbation analysis for the minimal solution. In this section we are interested in a qualitative description of the perturbation of the minimal nonnegative solution S of (1.1) as a function of M. The perturbation analysis of the minimal solution will be carried out through the perturbation analysis of a proper invariant subspace of the matrix

$$L = \begin{bmatrix} D & -C \\ B & -A \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & -I \end{bmatrix} M. \tag{3.1}$$

Let all eigenvalues of L be arranged in descending order of their real parts and be denoted by $\lambda_1, \ldots, \lambda_n, \lambda_{n+1}, \ldots, \lambda_{n+m}$. Then (see [10])

$$\sigma(D - CS) = \{\lambda_1, \ldots, \lambda_n\}$$

and

$$\sigma(A - SC) = \sigma(A - B\widehat{S}) = \{-\lambda_{n+1}, \ldots, -\lambda_{n+m}\}, \tag{3.2}$$

where $\widehat{S}$ is the minimal nonnegative solution of the dual equation (1.4). If M is a nonsingular M-matrix, then $\lambda_1, \ldots, \lambda_n \in \mathbb{C}^+$ (the open right half plane) and $\lambda_{n+1}, \ldots, \lambda_{n+m} \in \mathbb{C}^-$ (the open left half plane). If M is an irreducible singular M-matrix, then $\lambda_1, \ldots, \lambda_{n-1} \in \mathbb{C}^+$, $\lambda_{n+2}, \ldots, \lambda_{n+m} \in \mathbb{C}^-$. Moreover,

- if $u_1^T v_1 > u_2^T v_2$, then $\lambda_n = 0$ and $\lambda_{n+1} < 0$ are simple eigenvalues;
- if $u_1^T v_1 < u_2^T v_2$, then $\lambda_n > 0$ and $\lambda_{n+1} = 0$ are simple eigenvalues;
- if $u_1^T v_1 = u_2^T v_2$, then $\lambda_n = \lambda_{n+1} = 0$ is a double eigenvalue with only one linearly independent eigenvector.

Therefore, in all cases, there is a unique invariant subspace of L corresponding to the eigenvalues $\lambda_1, \ldots, \lambda_n$. Let the invariant subspace be $\mathrm{Im}\,[U_1^T\ U_2^T]^T$, where $U_1 \in \mathbb{C}^{n\times n}$, $U_2 \in \mathbb{C}^{m\times n}$ and $\mathrm{Im}\,U$ denotes the image (or range) of the matrix U. Then U_1 is nonsingular and $S = U_2U_1^{-1}$ (see [10]).

When M is an irreducible M-matrix, the matrices $D - CS$ and $A - SC$ are also irreducible M-matrices by Theorem 1.1. Since $A - SC$ and $(D - CS)^T$ can be written in the form $sI - N$, where $N \geq 0$ is irreducible, it follows from the Perron–Frobenius theorem that there exist unique positive vectors a and b with unit 1-norm such that

$$(A - SC)a = -\lambda_{n+1}a, \quad b^T(D - CS) = \lambda_n b^T. \tag{3.3}$$

Since M is irreducible, we have $C \neq 0$ and thus $b^TCa > 0$. We will need the following result [7] in the perturbation analysis below.

THEOREM 3.1. *Assume that M is an irreducible nonsingular M-matrix or an irreducible singular M-matrix with $u_1^Tv_1 \neq u_2^Tv_2$. Then there exists a second positive solution S_+ of* (1.1) *given by*

$$S_+ = S + \delta ab^T, \tag{3.4}$$

where the vectors a, b are specified in (3.3) *and $\delta = (\lambda_n - \lambda_{n+1})/b^TCa$. Moreover,*

$$\sigma(D - CS_+) = \{\lambda_1, \ldots, \lambda_{n-1}, \lambda_{n+1}\}. \tag{3.5}$$

Let $\mathcal{M}$ and $\mathcal{N}$ be any invariant subspaces of L. For any fixed norm $\|\cdot\|$ (for definiteness we use the spectral norm), let $\theta(\mathcal{M}, \mathcal{N})$ be the gap between $\mathcal{M}$ and $\mathcal{N}$, defined by

$$\theta(\mathcal{M}, \mathcal{N}) = \|P_{\mathcal{M}} - P_{\mathcal{N}}\|,$$

where P_M and $P_{\mathcal{N}}$ are the orthogonal projectors on $\mathcal{M}$ and $\mathcal{N}$, respectively, with orthogonality defined by the standard scalar product on $\mathbb{C}^{m+n}$. See [8] or [22] for properties of the gap metric.

We first consider the case where M is a nonsingular M-matrix or an irreducible singular M-matrix with $u_1^Tv_1 \neq u_2^Tv_2$. In this case, since the eigenvalues $\lambda_1, \ldots, \lambda_n$ are disjoint from the eigenvalues $\lambda_{n+1}, \ldots, \lambda_{n+m}$, the invariant subspace corresponding to the eigenvalues $\lambda_1, \ldots, \lambda_n$,

$$\mathcal{M} = \mathrm{Im}\begin{bmatrix} U_1 \\ U_2 \end{bmatrix} = \mathrm{Im}\begin{bmatrix} I \\ S \end{bmatrix},$$

is known to be Lipschitz stable [8], i.e., there exist constants $\gamma_1, \epsilon > 0$ such that every matrix K satisfying $\|K - L\| < \epsilon$ has an invariant subspace $\mathcal{N}$ for which $\theta(\mathcal{M}, \mathcal{N}) \leq \gamma_1\|K - L\|$. In particular, every $\widetilde{L} = \mathrm{diag}(I, -I)\widetilde{M}$ with $\|\widetilde{L} - L\| < \epsilon$ has an invariant subspace $\mathcal{N}$ for which $\theta(\mathcal{M}, \mathcal{N}) \leq \gamma_1\|\widetilde{L} - L\|$. Let $\mathcal{N} = \mathrm{Im}[V_1^T\ V_2^T]^T$. Then for ϵ small enough, V_1 is nonsingular and we let $T = V_2V_1^{-1}$. Then for $\|\widetilde{M} - M\| = \|\widetilde{L} - L\| < \epsilon$

$$\theta\left(\mathrm{Im}\begin{bmatrix} I \\ S \end{bmatrix}, \mathrm{Im}\begin{bmatrix} I \\ T \end{bmatrix}\right) \leq \gamma_1\|\widetilde{M} - M\|.$$

Note that there is a constant $\gamma_2 > 0$ such that [8]

$$\gamma_2^{-1}\|T - S\| \leq \theta\left(\mathrm{Im}\begin{bmatrix} I \\ S \end{bmatrix}, \mathrm{Im}\begin{bmatrix} I \\ T \end{bmatrix}\right) \leq \gamma_2\|T - S\|.$$

Thus

$$\|T - S\| \le \gamma_1\gamma_2\|\widetilde{M} - M\|.$$

For ϵ small enough, we know that the eigenvalues of $\widetilde{D} - \widetilde{C}T$ are individually close to the eigenvalues of $D - CS$, and hence they are the n eigenvalues of $\widetilde{L}$ with the largest real parts. It follows that $T = \widetilde{S}$, the minimal nonnegative solution of (1.1) with M replaced by $\widetilde{M}$.

We have thus proved the following result.

THEOREM 3.2. *If M is a nonsingular M-matrix or an irreducible singular M-matrix with $u_1^T v_1 \ne u_2^T v_2$, then there exist constants $\gamma > 0$ and $\epsilon > 0$ such that $\|\widetilde{S} - S\| \le \gamma\|\widetilde{M} - M\|$ for all $\widetilde{M}$ with $\|\widetilde{M} - M\| < \epsilon$.*

We now consider the case where M is an irreducible singular M-matrix with $u_1^T v_1 = u_2^T v_2$. Let $q_1, q_2, \dots, q_{n-1}$ be the eigenvectors and generalized eigenvectors corresponding to the eigenvalues $\lambda_1, \dots, \lambda_{n-1}$ and let v be the eigenvector corresponding to the zero eigenvalue. Now,

$$\mathrm{Im}\begin{bmatrix} I \\ S \end{bmatrix} = \mathrm{Im}[q_1\ q_2\ \dots\ q_{n-1}] \dotplus \mathrm{Im}[v].$$

As in the previous case, there exist constants $\gamma_1, \epsilon > 0$ such that for any $\widetilde{M}$ with $\|\widetilde{M} - M\| < \epsilon$, $\widetilde{L}$ has an invariant subspace $\mathcal{N}_1$ for which

$$\theta(\mathrm{Im}[q_1\ q_2\ \dots\ q_{n-1}], \mathcal{N}_1) \le \gamma_1\|\widetilde{M} - M\|.$$

We assume that ϵ is small enough such that the eigenvalues of $\widetilde{L}$ corresponding to $\mathcal{N}_1$ are the $n-1$ eigenvalues of $\widetilde{L}$ with the largest real parts. Note that when $\widetilde{M}$ is close enough to M, $\widetilde{M}$ is also irreducible. We consider two cases: (a) $\widetilde{M}$ is nonsingular and (b) $\widetilde{M}$ is singular.

For case (a), $\widetilde{L}$ has an eigenvalue $\widetilde{\lambda}_n > 0$ that is a perturbation of the zero eigenvalue (with index two) of L. The eigenvector $\widetilde{v}$ corresponding to $\widetilde{\lambda}_n$ is such that

$$\theta(\mathrm{Im}[v], \mathrm{Im}[\widetilde{v}]) \le \gamma_2\|\widetilde{M} - M\|^{1/2}$$

for some $\gamma_2 > 0$ (see section 16.5 of [8] or section 5 of [9]). Now, there are constants $\gamma_3, \gamma_4 > 0$ such that [8]

$$\begin{aligned}
&\theta(\mathrm{Im}[q_1\ q_2\ \dots\ q_{n-1}] \dotplus \mathrm{Im}[v], \mathcal{N}_1 \dotplus \mathrm{Im}[\widetilde{v}]) \\
&\quad \le \gamma_3[\theta(\mathrm{Im}[q_1\ q_2\ \dots\ q_{n-1}], \mathcal{N}_1) + \theta(\mathrm{Im}[v], \mathrm{Im}[\widetilde{v}])] \\
&\quad \le \gamma_4\|\widetilde{M} - M\|^{1/2}.
\end{aligned}$$

It then follows as before that $\|\widetilde{S} - S\| \le \gamma\|\widetilde{M} - M\|^{1/2}$ for some $\gamma > 0$.

For case (b), let $\widetilde{v}$ be the eigenvector corresponding to the zero eigenvalue of $\widetilde{L}$. Then v and $\widetilde{v}$ are also eigenvectors of M and $\widetilde{M}$ corresponding to its simple zero eigenvalue. It is known that

$$\theta(\mathrm{Im}[v], \mathrm{Im}[\widetilde{v}]) \le \gamma_2\|\widetilde{M} - M\|$$

for some $\gamma_2 > 0$. If $0 = \widetilde{\lambda}_n \geq \widetilde{\lambda}_{n+1}$ then as before $\|\widetilde{S} - S\| \leq \gamma\|\widetilde{M} - M\|$ for some $\gamma > 0$. If $0 = \widetilde{\lambda}_{n+1} < \widetilde{\lambda}_n$ then we use Theorem 3.1 with M replaced by $\widetilde{M}$ (so accordingly we have $\widetilde{S}, \widetilde{S}_+, \widetilde{a}, \widetilde{b}$, etc.) to get

$$\|\widetilde{S}_+ - S\| \leq \gamma_3\|\widetilde{M} - M\|$$

for some $\gamma_3 > 0$. Note that $\|\widetilde{S}_+ - \widetilde{S}\| \leq \|\widetilde{\delta}\ \widetilde{a}\ \widetilde{b}^T\| \leq \gamma_4|\widetilde{\lambda}_n|$ for some $\gamma_4 > 0$. The eigenvalues of $\widetilde{A} - \widetilde{S}_+\widetilde{C}$ are $-\widetilde{\lambda}_n, -\widetilde{\lambda}_{n+2}, \ldots, -\widetilde{\lambda}_{n+m}$. The simple eigenvalue $-\widetilde{\lambda}_n$ of $\widetilde{A} - \widetilde{S}_+\widetilde{C}$ is a perturbation of the simple eigenvalue $-\lambda_{n+1} = 0$ of $A - SC$. Thus $|\widetilde{\lambda}_n| \leq \gamma_5\|(\widetilde{A} - \widetilde{S}_+\widetilde{C}) - (A - SC)\| \leq \gamma_6\|\widetilde{M} - M\|$ for some $\gamma_5, \gamma_6 > 0$. Therefore $\|\widetilde{S} - S\| \leq \|\widetilde{S}_+ - S\| + \|\widetilde{S}_+ - \widetilde{S}\| \leq \gamma\|\widetilde{M} - M\|$ for some $\gamma > 0$.

In summary, we have shown the following.

THEOREM 3.3. *If M is an irreducible singular M-matrix with $u_1^T v_1 = u_2^T v_2$, then there exist constants $\gamma > 0$ and $\epsilon > 0$ such that*

(a) $\|\widetilde{S} - S\| \leq \gamma\|\widetilde{M} - M\|^{1/2}$ *for all $\widetilde{M}$ with $\|\widetilde{M} - M\| < \epsilon$;*

(b) $\|\widetilde{S} - S\| \leq \gamma\|\widetilde{M} - M\|$ *for all* singular $\widetilde{M}$ *with $\|\widetilde{M} - M\| < \epsilon$.*

We illustrate the results in Theorem 3.3 with a simple example.

Consider the matrix

$$M = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}$$

and its three different perturbations

$$M_1 = \begin{bmatrix} 1+\epsilon & -1 \\ -1 & 1 \end{bmatrix}, \quad M_2 = \begin{bmatrix} 1 & -(1+\epsilon) \\ -1 & 1+\epsilon \end{bmatrix}, \quad M_3 = \begin{bmatrix} 1 & -1 \\ -(1+\epsilon) & 1 \end{bmatrix},$$

where $0 < \epsilon < 1$. Note that M satisfies the condition in Theorem 3.3, and that $S = 1$ for the corresponding NARE (1.1). M_1 is a nonsingular M-matrix and the corresponding minimal solution is $S_1 = \frac{1}{2}(2 + \epsilon - \sqrt{4\epsilon + \epsilon^2}) \sim 1 - \epsilon^{1/2}$, which is the situation in Theorem 3.3(a). M_2 is an irreducible singular M-matrix and the corresponding minimal solution is $S_2 = 1/(1+\epsilon) \sim 1 - \epsilon$, which is the situation in Theorem 3.3(b). M_3 is not an M-matrix and the corresponding NARE does not have real solutions.

The continuity of the minimal solution shown in Theorem 3.3 can be used to prove the next result, where the statements are stronger than those given in [10, Thm. 4.8]. The result will be needed in section 4.

THEOREM 3.4. *Let M be an irreducible singular M-matrix.*

(a) *If $u_1^T v_1 = u_2^T v_2$, then $Sv_1 = v_2$ and $\widehat{S}v_2 = v_1$.*

(b) *If $u_1^T v_1 > u_2^T v_2$, then $Sv_1 = v_2$ and $\widehat{S}v_2 < v_1$.*

(c) *If $u_1^T v_1 < u_2^T v_2$, then $Sv_1 < v_2$ and $\widehat{S}v_2 = v_1$.*

Proof. We only need to prove the result for S since the result for $\widehat{S}$ follows immediately by duality. So we need to show $Sv_1 = v_2$ when $u_1^T v_1 \geq u_2^T v_2$ and $Sv_1 < v_2$ when $u_1^T v_1 < u_2^T v_2$. In fact,

$$\begin{aligned}(A - SC)(v_2 - Sv_1) &= Av_2 - SCv_2 + (SCS - AS)v_1 \\ &= Bv_1 - SDv_1 + (SD - B)v_1 = 0.\end{aligned}$$

If $u_1^T v_1 > u_2^T v_2$, then $A - SC$ is nonsingular and so $Sv_1 = v_2$. If $u_1^T v_1 < u_2^T v_2$, then $A - SC$ is an irreducible singular M-matrix and $v_2 - Sv_1 \geq 0$ is an eigenvector corresponding to the zero eigenvalue (it is already proved in [10] that $Sv_1 \le v_2$ and $Sv_1 \ne v_2$). By the Perron–Frobenius theorem, $v_2 - Sv_1 > 0$ and so $Sv_1 < v_2$. If $u_1^T v_1 = u_2^T v_2$, then for

$$M(\alpha) = \begin{bmatrix} D & -C \\ -\alpha B & \alpha A \end{bmatrix}$$

with $\alpha > 1$, we have

$$u_1(\alpha) = u_1, \quad u_2(\alpha) = \alpha^{-1} u_2, \quad v_1(\alpha) = v_1, \quad v_2(\alpha) = v_2.$$

So we have $u_1(\alpha)^T v_1(\alpha) > u_2(\alpha)^T v_2(\alpha)$. It follows that $S(\alpha)v_1(\alpha) = v_2(\alpha)$. However, $\lim_{\alpha \to 1^+} S(\alpha) = S$ by Theorem 3.3 and so $Sv_1 = v_2$. $\square$

4. Applicability of the shifted LR algorithm. In this section we assume that M is an irreducible singular M-matrix. For the NARE (1.1) arising in the study of Markov models, we have $Me = 0$, where e is the vector of ones. In that case, we may take $v_1 = e \in \mathbb{R}^n$ and $v_2 = e \in \mathbb{R}^m$ in (1.3).

If M is a general irreducible singular M-matrix, we can transform (1.1) into a new equation for which $v_1 = e$ and $v_2 = e$. More precisely, (1.1) can be rewritten as

$$W(V_1^{-1} C V_2) W - W(V_1^{-1} D V_1) - (V_2^{-1} A V_2) W + V_2^{-1} B V_1 = 0 \tag{4.1}$$

with $V_1 = \operatorname{diag}(v_1)$, $V_2 = \operatorname{diag}(v_2)$, and $W = V_2^{-1} X V_1$. Note that the minimal nonnegative solution of (4.1) is $\overline{S} = V_2^{-1} S V_1$ and that

$$\begin{bmatrix} V_1^{-1} D V_1 & -V_1^{-1} C V_2 \\ -V_2^{-1} B V_1 & V_2^{-1} A V_2 \end{bmatrix} \begin{bmatrix} e \\ e \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}. \tag{4.2}$$

It is clear that the leftmost matrix in (4.2) is still an irreducible singular M-matrix. From now on, we assume that M is an irreducible singular M-matrix with $Me = 0$.

Ramaswami [26] made the interesting observation that the matrix equation (1.1) is closely related to a quadratic matrix equation arising in quasi-birth-death processes. To see this connection, let

$$a_* = \max_{1 \le i \le m} a_{ii}, \quad d_* = \max_{1 \le i \le n} d_{ii}, \quad \theta_* = \max(a_*, d_*). \tag{4.3}$$

Choose a number $\theta \ge \theta_*$ and let $P = I - \frac{1}{\theta} M$. Then P is nonnegative with $Pe = e$, i.e., P is a stochastic matrix. Let

$$P = \begin{bmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{bmatrix},$$

where the partitioning is conformable with that for the matrix M. Thus

$$P_{11} = I - \frac{1}{\theta} D, \quad P_{12} = \frac{1}{\theta} C, \quad P_{21} = \frac{1}{\theta} B, \quad P_{22} = I - \frac{1}{\theta} A. \tag{4.4}$$

Ramaswami [26] constructed three nonnegative matrices from P:

$$A_0 = \begin{bmatrix} P_{11} & 0 \\ \frac{1}{2} P_{21} & 0 \end{bmatrix}, \quad A_1 = \begin{bmatrix} 0 & P_{12} \\ 0 & \frac{1}{2} P_{22} \end{bmatrix}, \quad A_2 = \begin{bmatrix} 0 & 0 \\ 0 & \frac{1}{2} I \end{bmatrix}. \tag{4.5}$$

Associated with the matrices A_0, A_1, A_2 are the matrix equation

$$G = A_0 + A_1 G + A_2 G^2 \tag{4.6}$$

and its dual equation

$$F = A_2 + A_1 F + A_0 F^2. \tag{4.7}$$

We let G and F be the minimal nonnegative solutions of (4.6) and (4.7), respectively.

The next two results are known (see [26, Thm. 4.1] and [14]).

PROPOSITION 4.1. *The minimal nonnegative solution of* (4.6) *is*

$$G = \begin{bmatrix} P_{11} + P_{12}S & 0 \\ S & 0 \end{bmatrix},$$

where S is the minimal nonnegative solution of (1.1).

PROPOSITION 4.2. *The minimal nonnegative solution of* (4.7) *is*

$$F = \begin{bmatrix} 0 & \widehat{S} \\ 0 & (2I - P_{22} - P_{21}\widehat{S})^{-1} \end{bmatrix},$$

where $\widehat{S}$ is the minimal nonnegative solution of (1.4).

Since $(2I - P_{22} - P_{21}\widehat{S})^{-1} = (I + \frac{1}{\theta}(A - B\widehat{S}))^{-1}$ is a nonnegative matrix, $\rho(F) = \rho((2I - P_{22} - P_{21}\widehat{S})^{-1})$ is the largest positive eigenvalue of $(I + \frac{1}{\theta}(A - B\widehat{S}))^{-1}$, which is $1/(1 - \frac{1}{\theta}\lambda_{n+1})$. Similarly, $\rho(G) = \rho(P_{11} + P_{12}S) = \rho(I - \frac{1}{\theta}(D - CS)) = 1 - \frac{1}{\theta}\lambda_n$.

The solution G can be computed by the LR algorithm [23], which is essentially the cyclic reduction algorithm combined with block-diagonal scaling (see [12]).

ALGORITHM 4.3. *Set*

$$\begin{aligned} L^{(0)} &= (I - A_1)^{-1} A_0, \\ H^{(0)} &= (I - A_1)^{-1} A_2, \\ G^{(0)} &= L^{(0)}, \\ T^{(0)} &= H^{(0)}. \end{aligned}$$

For $k = 0, 1, \ldots,$ compute

$$\begin{aligned} U^{(k)} &= H^{(k)} L^{(k)} + L^{(k)} H^{(k)}, \\ L^{(k+1)} &= (I - U^{(k)})^{-1} (L^{(k)})^2, \\ H^{(k+1)} &= (I - U^{(k)})^{-1} (H^{(k)})^2, \\ G^{(k+1)} &= G^{(k)} + T^{(k)} L^{(k+1)}, \\ T^{(k+1)} &= T^{(k)} H^{(k+1)}. \end{aligned}$$

It is shown in [23] that the matrices $H^{(k)}$ and $L^{(k)}$ are well defined and nonnegative and that the sequence $\{G^{(k)}\}$ converges quadratically to the matrix G, except for a critical case which corresponds to the case $u_1^T e = u_2^T e$ in the NARE (1.1). In the latter case, the convergence is expected to be linear with rate 1/2 (see [12] and [14]).

When $m = n$, the LR algorithm needs about $\frac{400}{3}n^3$ flops each iteration. Using the special structure of the matrices A_0, A_1, A_2, we can simplify the LR algorithm and the simplified algorithm requires about $\frac{124}{3}n^3$ flops each iteration [14]. The simplified LR algorithm is less expensive than Newton's method, which requires roughly $60n^3$ flops each iteration when $m = n$. However, there are examples [1] for which the (simplified) LR algorithm requires many more iterations than Newton's method, even though they both have quadratic convergence.

The matrix $G^{(k)}$ from Algorithm 4.3 has the form

$$G^{(k)} = \begin{bmatrix} G_1^{(k)} & 0 \\ G_2^{(k)} & 0 \end{bmatrix},$$

and the solution S is approximated by the matrices $S_k = G_2^{(k)}$. It is shown in [14] that

$$\limsup_{k\to\infty} \sqrt[2^{k+1}]{\|S_k - S\|} \le \rho(F)\rho(G), \tag{4.8}$$

so S_k converges to S quadratically when $\rho(F)\rho(G) < 1$ and the convergence will be fast if $\rho(F)\rho(G)$ is not close to 1.

Since

$$\rho(F) = 1/\left(1 - \frac{1}{\theta}\lambda_{n+1}\right), \quad \rho(G) = 1 - \frac{1}{\theta}\lambda_n$$

are nondecreasing functions of θ for $\theta \ge \theta_*$, we should take $\theta = \theta_*$ in (4.4) to have faster convergence for the (simplified) LR algorithm.

Note that when $u_1^T e = u_2^T e$, $Se = e$ and $\widehat{S}e = e$ by Theorem 3.4. So $Fe = Ge = e$, $\rho(F) = \rho(G) = 1$ and the convergence is expected to be linear with rate 1/2. To have faster convergence when $u_1^T e \ge u_2^T e$, we need to use a shift technique [19] for the (simplified) LR algorithm. The case $u_1^T e < u_2^T e$ for the NARE will be reduced to the case $u_1^T e > u_2^T e$ for a new NARE of the same type.

4.1. Case $\boldsymbol{u_1^T e \ge u_2^T e}$. In this subsection we assume $u_1^T e \ge u_2^T e$. In this case $Se = e$ and so G is stochastic. It is shown in [14] that the only eigenvalue of G on the unit circle is the simple eigenvalue 1.

The shift technique introduced in [19] is $H = G - ev^T$, where $v > 0$ and $v^T e = 1$. For our purposes here, we only require that $v \ge 0$ and $v^T e = 1$. Then the eigenvalues of H are those of G except that the eigenvalue 1 of G is replaced by 0, and H is a solution of the new equation

$$H = B_0 + B_1 H + B_2 H^2, \tag{4.9}$$

where

$$B_0 = A_0(I - ev^T), \quad B_1 = A_1 + A_2 ev^T, \quad B_2 = A_2. \tag{4.10}$$

It is shown in [14] that there is a matrix K with $\rho(K) = \rho(F)$ such that

$$K = B_2 + B_1 K + B_0 K^2. \tag{4.11}$$

To find the solution H of (4.9), we can apply Algorithm 4.3 with the triple (A_0, A_1, A_2) replaced by the triple (B_0, B_1, B_2). To avoid confusion, we will put a "hat" on each sequence generated. We take

$$v = \begin{bmatrix} p \\ 0 \end{bmatrix}, \tag{4.12}$$

where $p \in \mathbb{R}^n$ is positive and $p^T e = 1$. In this way we can get a simplified LR algorithm as before, with no increase in computational work for each iteration. Note that S is now approximated by $\widehat{S}_k = \widehat{G}_2^{(k)} + ep^T$.

It is shown in [14] that when Algorithm 4.3 is applied with (A_0, A_1, A_2) replaced by (B_0, B_1, B_2), the matrix $I - B_1$ in the initialization step is always invertible. Assuming that $I - \widehat{U}^{(k)}$ is invertible for each $k \geq 0$, it is shown in [14] that

$$\limsup_{k\to\infty} \sqrt[2^{k+1}]{\|\widehat{S}^{(k)} - S\|} \leq \rho(K)\rho(H) = \rho(F)\rho(H) < 1. \tag{4.13}$$

Since $\rho(H) < \rho(G)$, the shift technique has improved the speed of convergence. In particular, $\widehat{S}^{(k)}$ converges to S quadratically whenever $u_1^T e \geq u_2^T e$. It is also shown in [14] that $I - \widehat{U}^{(k)}$ converges to I quadratically, assuming that $I - \widehat{U}^{(k)}$ is nonsingular for all $k \geq 0$.

The problem as to whether the matrices $I - \widehat{U}^{(k)}$ could be singular for small k was unsolved in [14]. We will now solve this problem.

We proceed as in [6] but depart from [6] at some point. Let

$$T_k = \begin{bmatrix} I - A_1 & -A_2 & & \\ -A_0 & I - A_1 & \ddots & \\ & \ddots & \ddots & -A_2 \\ & & -A_0 & I - A_1 \end{bmatrix}$$

and

$$\widehat{T}_k = \begin{bmatrix} I - B_1 & -B_2 & & \\ -B_0 & I - B_1 & \ddots & \\ & \ddots & \ddots & -B_2 \\ & & -B_0 & I - B_1 \end{bmatrix}$$

be block $k \times k$ Toeplitz matrices. Since the LR algorithm is well defined if and only if the cyclic reduction (CR) algorithm is well defined [5], it follows from Theorem 13 of [3] that the matrices T_{2^j-1} are nonsingular for all $j \geq 1$ and that $I - \widehat{U}^{(k)}$ are nonsingular for all $k \geq 0$ if $\widehat{T}_{2^j-1}$ are nonsingular for all $j \geq 2$. The relation between T_k and $\widehat{T}_k$ (for $k \geq 3$) has been obtained in [6] as

$$\widehat{T}_k = T_k \begin{bmatrix} I & & & \\ V & I & & \\ \vdots & \ddots & \ddots & \\ V & \dots & V & I \end{bmatrix} + \begin{bmatrix} 0 \\ \vdots \\ 0 \\ -A_2 \end{bmatrix} \begin{bmatrix} V & V & \dots & V \end{bmatrix}, \tag{4.14}$$

where $V = ev^T$. Note that this relation can be obtained directly from (4.10). Let Q_k and P_k be the $(k, 1)$ block and (k, k) block of T_k^{-1}, respectively. From (4.14), it is shown in [6] that $\widehat{T}_k$ is nonsingular if and only if $v^T P_k A_2 e \neq 1$. From the proof of Theorem 9 in [6] we also know that

$$v^T Q_k A_0 e + v^T P_k A_2 e = 1. \tag{4.15}$$

In the case where v is taken to be positive and $u_1^T e > u_2^T e$, it has been shown in [6] that $v^T P_k A_2 e \neq 1$, using, among other things, the canonical factorizations of matrix polynomials and the so-called asymptotic applicability of the SCR (CR with a shift technique). So, the argument in [6] is very involved and it does not cover the case $u_1^T e = u_2^T e$. Suppose SCR were to break down for the case $u_1^T e = u_2^T e$. Then near-breakdown would happen to SCR with $u_1^T e > u_2^T e$, but $u_1^T e \approx u_2^T e$. Moreover, as we mentioned earlier, we need to take the vector v in the form (4.12) to avoid an increase in computational work when using the shift technique. Fortunately, we can prove the applicability of the LR algorithm, with a shift given by (4.12), for all cases with $u_1^T e \geq u_2^T e$ and $\theta > \theta_*$. Moreover, the proof is very simple.

In fact, what we need to prove is $v^T Q_k A_0 e > 0$, which implies $v^T P_k A_2 e \neq 1$ by (4.15). Note that

$$T_k^{-1} \geq \begin{bmatrix} I & & & \\ -A_0 & I & & \\ & \ddots & \ddots & \\ & & -A_0 & I \end{bmatrix}^{-1} = \begin{bmatrix} I & & & \\ A_0 & I & & \\ \vdots & \ddots & \ddots & \\ A_0^{k-1} & \cdots & A_0 & I \end{bmatrix}. \tag{4.16}$$

So $Q_k \geq A_0^{k-1}$ and hence $v^T Q_k A_0 e \geq v^T A_0^k e$. For A_0 given by (4.5), we have

$$A_0^k = \begin{bmatrix} P_{11}^k & 0 \\ \frac{1}{2} P_{21} P_{11}^{k-1} & 0 \end{bmatrix}.$$

Therefore, $v^T Q_k A_0 e \geq p^T P_{11}^k e$, by (4.12). Recall that the nonnegative matrix P_{11} is given by $P_{11} = I - \frac{1}{\theta} D$. If the diagonal elements d_{ii} of D are not all equal or a_* and d_* defined in (4.3) satisfy $d_* < a_*$, then P_{11} has at least one nonzero diagonal element and hence $p^T P_{11}^k e > 0$ for all $k \geq 1$ and for all $\theta \geq \theta_*$. If the elements d_{ii} are all equal and $d_* \geq a_*$, then $p^T P_{11}^k e > 0$ for all $k \geq 1$ and all $\theta > \theta_* = d_*$.

THEOREM 4.4. *Algorithm 4.3 can be applied with no breakdown when the shift technique is used, i.e., when the matrices A_0, A_1, A_2 in (4.5) are replaced by the matrices B_0, B_1, B_2 defined in (4.10), for all $\theta \geq \theta_*$ if the diagonal elements d_{ii} of D are not all equal or $d_* < a_*$, and for all $\theta > \theta_*$ if the elements d_{ii} are all equal and $d_* \geq a_*$.*

When the elements d_{ii} of D are all equal, it is possible for P_{11} to be nilpotent if we take $\theta = \theta_*$. One simple example is

$$M = \begin{bmatrix} 1 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & -1 \\ -1 & 0 & 0 & 1 \end{bmatrix}. \tag{4.17}$$

For this example with $\theta = 1$, $p^T P_{11}^k e = 0$ for $k \geq 2$. However, it is very likely that we still have $v^T Q_k A_0 e > 0$ since the lower bound in (4.16) is not tight.

For the LR algorithm without a shift, the number $\rho(F)\rho(G)$ in (4.8) in minimized for $\theta = \theta_*$. So $\theta = \theta_*$ is optimal in this sense and should be recommended. For the LR algorithm with a shift, however, the optimal θ should minimize $\rho(F)\rho(H)$ in (4.13).

When $u_1^T e = u_2^T e$, we have $\lambda_{n+1} = 0$ and $\rho(F) = 1$ for any θ. When $u_1^T e > u_2^T e$ but $u_1^T e \approx u_2^T e$, we have $\lambda_{n+1} \approx 0$ and hence the effect of θ on $\rho(F)$ is very limited. So one should try to minimize $\rho(H)$. Note that $\rho(H) = \max_{1 \le i \le n-1} |1 - \frac{1}{\theta}\lambda_i|$. For the matrix M given by (4.17), the corresponding matrix L has eigenvalues $\sqrt{2}, 0, 0, -\sqrt{2}$. So $\rho(F) = 1$ and $\rho(H)$ is minimized for $\theta = \sqrt{2}$ and the minimum is 0. This example shows that $\theta = \theta_*$ is not necessarily optimal when the shift technique is used. We can also give a necessary and sufficient condition for θ_* to be optimal. Let $D = \{z \in \mathbb{C} : |z - 1| < 1\}$. Then $\lambda_i/\theta_* \in D$ for $i = 1, \ldots, n-1$ since $\rho(H) < 1$. Let $D_1 = \{z \in \mathbb{C} : |z - 1/2| \le 1/2\}$, $D_2 = D \setminus D_1$, $I_1 = \{1 \le i \le n-1 : \lambda_i/\theta_* \in D_1\}$, and $I_2 = \{1 \le i \le n-1 : \lambda_i/\theta_* \in D_2\}$. Then we have the following result.

PROPOSITION 4.5. *For $\theta \in [\theta_*, \infty)$, $\rho(H)$ attains its minimum at $\theta = \theta_*$ if and only if*

$$\max_{i \in I_1} |1 - \lambda_i/\theta_*| \ge \max_{i \in I_2} |1 - \lambda_i/\theta_*|,$$

where the maximum over an empty set is defined to be zero.

Proof. Note that for any point (other than 0) on the circle $|z - 1/2| = 1/2$, the boundary of D_1, the line passing through z and 0 is perpendicular to the line passing through z and 1. If $\max_{i \in I_1} |1 - \lambda_i/\theta_*| \ge \max_{i \in I_2} |1 - \lambda_i/\theta_*|$, then for any $\theta > \theta_*$ and $i \in I_1$, which is nonempty, $|1 - \lambda_i/\theta| > |1 - \lambda_i/\theta_*|$ and thus $\rho(H)$ is minimized at θ_*. On the other hand, if $\max_{i \in I_1} |1 - \lambda_i/\theta_*| < \max_{i \in I_2} |1 - \lambda_i/\theta_*|$, we can take $\theta > \theta_*$ such that

$$\max_{i \in I_1} |1 - \lambda_i/\theta| < \max_{i \in I_2} |1 - \lambda_i/\theta| < \max_{i \in I_2} |1 - \lambda_i/\theta_*|.$$

(The first inequality holds when $\theta - \theta_*$ is small enough and the second inequality holds when $\theta - \theta_*$ is small enough so that $\lambda_i/\theta \in D_2$ for $i \in I_2$.) Thus $\rho(H)$ does not attain its minimum at θ_*. $\square$

In practice, we would not compute the eigenvalues $\lambda_1, \ldots, \lambda_{n-1}$ when we use the LR algorithm. However, the above result shows that $\theta = \theta_*$ is often not optimal when the shift technique is used. Therefore, when the diagonal elements d_{ii} of D are all equal and $d_* \ge a_*$, we can simply take $\theta > \theta_* = d_*$ (say $\theta = 1.1\theta_*$) to ensure the applicability of the LR algorithm with a shift.

4.2. Case $u_1^T e < u_2^T e$. We now assume $u_1^T e < u_2^T e$. Then $Se < e$ by Theorem 3.4. We will reduce this case to the case $u_1^T e > u_2^T e$ for a new NARE of the same type, and the substochastic minimal solution S of the original NARE will be obtained from the stochastic minimal solution of the new NARE. This reduction process is in essence similar to the one given in [25]. The difference is that the reduction here is given directly on the Riccati equation, rather than on the unilateral matrix equation obtained through the Ramaswami construction.

As in [15, Lem. 5.1] we note that the minimal nonnegative solution S of the NARE (1.1) is such that $S = Z^T$, where Z is the minimal nonnegative solution of the new NARE

$$ZC^T Z - ZA^T - D^T Z + B^T = 0. \tag{4.18}$$

As at the beginning of section 4, (4.18) can be rewritten as

$$W(U_2^{-1} C^T U_1) W - W(U_2^{-1} A^T U_2) - (U_1^{-1} D^T U_1) W + U_1^{-1} B^T U_2 = 0, \tag{4.19}$$

with $U_1 = \mathrm{diag}(u_1), U_2 = \mathrm{diag}(u_2)$, and $W = U_1^{-1} Z U_2$. Now the irreducible singular M-matrix corresponding to (4.19) is

$$\widehat{M} = \begin{bmatrix} U_2^{-1} A^T U_2 & -U_2^{-1} C^T U_1 \\ -U_1^{-1} B^T U_2 & U_1^{-1} D^T U_1 \end{bmatrix}.$$

It is easy to see that $(u_2^T \ u_1^T)\widehat{M} = 0$ and $\widehat{M}e = 0$. Since $u_2^T e > u_1^T e$, the new NARE (4.19) has a stochastic minimal solution W and it can be computed as in section 4.1. The substochastic minimal solution S of the original NARE is obtained through $S = U_2^{-1} W^T U_1$.

For the above procedure, we need to compute the vector $(u_1^T \ u_2^T)^T$ accurately since it determines the coefficient matrices of the NARE (4.19). This can be done by using the LU factorization of the irreducible singular M-matrix M^T, and the computational work is very minor compared with that required by each iteration for the simplified LR algorithm. So the shift technique is worthwhile as long as we can save one iteration. Moreover, as our perturbation analysis in section 3 suggests, the minimal solution computed by the LR algorithm without a shift is much more vulnerable to rounding errors when $u_1^T e \approx u_2^T e$.

We use one example to illustrate the usefulness of the above procedure. Consider the NARE (1.1) with $m = n = 100$ and

$$A = \begin{bmatrix} 3 & -1 & & \\ & \ddots & \ddots & \\ & & 3 & -1 \\ -1 & & & 1.9 \end{bmatrix}, \quad B = \begin{bmatrix} 1 & 1 & & \\ & \ddots & \ddots & \\ & & 1 & 1 \\ & & & 0.9 \end{bmatrix},$$

$$C = \begin{bmatrix} 1 & & & \\ 1 & 1 & & \\ & \ddots & \ddots & \\ & & 1 & 1 \end{bmatrix}, \quad D = \begin{bmatrix} 2 & -1 & & \\ & 3 & \ddots & \\ & & \ddots & -1 \\ -1 & & & 3 \end{bmatrix}.$$

It is easily verified that $Me = 0$ and $u_1^T e < u_2^T e$. We apply the (simplified) LR algorithm with a shift to the NARE (4.19) (so the matrices A, B, C, D in (4.4) are replaced accordingly), with $\theta = 3$ in (4.4) and $p = m^{-1}e$ in (4.12). After 6 iterations we find an approximation $\widetilde{W}$ to W with $\|\mathcal{R}(\widetilde{W})\|_\infty = 4.4 \times 10^{-11}$. We then use $\widetilde{W}$ to get an approximation $\widetilde{S}$ to S with $\|\mathcal{R}(\widetilde{S})\|_\infty = 6.1 \times 10^{-11}$. A very accurate approximation to S (with residual 2.3×10^{-14}) can be obtained by performing 7 iterations instead and we take it as the "exact" solution S. We now apply the (simplified) LR algorithm without a shift to the NARE (1.1), with $\theta = 3$ in (4.4). We find after 13 iterations an approximation $\widetilde{S}'$ to S, with $\|\mathcal{R}(\widetilde{S}')\|_\infty = 6.0 \times 10^{-10}$. However, the accuracy in this case is much lower than the residual suggests. Indeed, we find $\|\widetilde{S} - S\|_\infty = 1.4 \times 10^{-10}$ but $\|\widetilde{S}' - S\|_\infty = 4.2 \times 10^{-7}$. So the (simplified) LR algorithm with a shift is more efficient and more accurate.

5. Conclusions. In this further study of a class of NAREs, we have been able to relax the condition for the convergence of Newton's method to the minimal solution. The qualitative perturbation analysis for the minimal solution, while of independent interest, is instructive in designing algorithms for finding more accurate approximations. For the NAREs arising in Markov models, we have shown that the LR algorithm, combined with a shift technique, is breakdown-free in all cases and therefore is guaranteed to find the minimal solution more efficiently and more accurately.

Acknowledgments. This work was carried out while the first author visited MIMS in the School of Mathematics at the University of Manchester; he thanks the School for its hospitality. Both authors thank the referees for their helpful comments. In particular, comments from one referee led to a significant improvement of section 4.2.

REFERENCES

[1] N. G. Bean, M. M. O'Reilly, and P. G. Taylor, *Algorithms for return probabilities for stochastic fluid flows*, Stoch. Models, 21 (2005), pp. 149–184.
[2] A. Berman and R. J. Plemmons, *Nonnegative Matrices in the Mathematical Sciences*, revised reprint of the 1979 Academic Press original, SIAM, Philadelphia, PA, 1994.
[3] D. A. Bini, L. Gemignani, and B. Meini, *Computations with infinite Toeplitz matrices and polynomials*, Linear Algebra Appl., 343–344 (2002), pp. 21–61.
[4] D. A. Bini, B. Iannazzo, G. Latouche, and B. Meini, *On the solution of Riccati equations arising in fluid queues*, Linear Algebra Appl., 413 (2006), pp. 474–494.
[5] D. A. Bini, G. Latouche, and B. Meini, *Solving matrix polynomial equations arising in queueing problems*, Linear Algebra Appl., 340 (2002), pp. 225–244.
[6] D. A. Bini, B. Meini, and I. M. Spitkovsky, *Shift techniques and canonical factorizations in the solution of $M/G/1$-type Markov chains*, Stoch. Models, 21 (2005), pp. 279–302.
[7] S. Fital and C.-H. Guo, *Convergence of the solution of a nonsymmetric matrix Riccati differential equation to its stable equilibrium solution*, J. Math. Anal. Appl., 318 (2006), pp. 648–657.
[8] I. Gohberg, P. Lancaster, and L. Rodman, *Invariant Subspaces of Matrices with Applications*, John Wiley & Sons, New York, 1986.
[9] I. Gohberg and L. Rodman, *On the distance between lattices of invariant subspaces of matrices*, Linear Algebra Appl., 76 (1986), pp. 85–120.
[10] C.-H. Guo, *Nonsymmetric algebraic Riccati equations and Wiener–Hopf factorization for M-matrices*, SIAM J. Matrix Anal. Appl., 23 (2001), pp. 225–242.
[11] C.-H. Guo, *A note on the minimal nonnegative solution of a nonsymmetric algebraic Riccati equation*, Linear Algebra Appl., 357 (2002), pp. 299–302.
[12] C.-H. Guo, *Convergence analysis of the Latouche–Ramaswami algorithm for null recurrent quasi-birth-death processes*, SIAM J. Matrix Anal. Appl., 23 (2002), pp. 744–760.
[13] C.-H. Guo, *On a quadratic matrix equation associated with an M-matrix*, IMA J. Numer. Anal., 23 (2003), pp. 11–27.
[14] C.-H. Guo, *Efficient methods for solving a nonsymmetric algebraic Riccati equation arising in stochastic fluid models*, J. Comput. Appl. Math., 192 (2006), pp. 353–373.
[15] C.-H. Guo, B. Iannazzo, and B. Meini, *On the Doubling Algorithm for a (Shifted) Nonsymmetric Algebraic Riccati Equation*, Technical Report, 2006; available online at http://www.math.uregina.ca/~chguo/gim06.pdf.
[16] C.-H. Guo and A. J. Laub, *On the iterative solution of a class of nonsymmetric algebraic Riccati equations*, SIAM J. Matrix Anal. Appl., 22 (2000), pp. 376–391.
[17] X.-X. Guo and Z.-Z. Bai, *On the minimal nonnegative solution of nonsymmetric algebraic Riccati equation*, J. Comput. Math., 23 (2005), pp. 305–320.
[18] X.-X. Guo, W.-W. Lin, and S.-F. Xu, *A structure-preserving doubling algorithm for nonsymmetric algebraic Riccati equation*, Numer. Math., 103 (2006), pp. 393–412.
[19] C. He, B. Meini, and N. H. Rhee, *A shifted cyclic reduction algorithm for quasi-birth-death problems*, SIAM J. Matrix Anal. Appl., 23 (2001), pp. 673–691.
[20] J. Juang, *Existence of algebraic matrix Riccati equations arising in transport theory*, Linear Algebra Appl., 230 (1995), pp. 89–100.
[21] J. Juang and W.-W. Lin, *Nonsymmetric algebraic Riccati equations and Hamiltonian-like matrices*, SIAM J. Matrix Anal. Appl., 20 (1998), pp. 228–243.
[22] P. Lancaster and L. Rodman, *Algebraic Riccati Equations*, Clarendon Press, Oxford, 1995.
[23] G. Latouche and V. Ramaswami, *A logarithmic reduction algorithm for quasi-birth-death processes*, J. Appl. Probab., 30 (1993), pp. 650–674.

[24] L.-Z. Lu, *Solution form and simple iteration of a nonsymmetric algebraic Riccati equation arising in transport theory*, SIAM J. Matrix Anal. Appl., 26 (2005), pp. 679–685.
[25] V. Ramaswami, *A duality theorem for the matrix paradigms in queueing theory*, Comm. Statist. Stochastic Models, 6 (1990), pp. 151–161.
[26] V. Ramaswami, *Matrix analytic methods for stochastic fluid flows*, in Proceedings of the 16th International Teletraffic Congress, Elsevier Science B. V., Edinburgh, 1999, pp. 1019–1030.
[27] L. C. G. Rogers, *Fluid models in queueing theory and Wiener–Hopf factorization of Markov chains*, Ann. Appl. Probab., 4 (1994), pp. 390–413.
[28] L. C. G. Rogers and Z. Shi, *Computing the invariant law of a fluid model*, J. Appl. Probab., 31 (1994), pp. 885–896.