Nonsymmetric algebraic Riccati equations associated with an M-matrix: recent advances and algorithms *

Dario A. Bini, Bruno Iannazzo, Beatrice Meini, Federico Poloni

Abstract

We survey theoretical properties and algorithms concerning the problem of solving a nonsymmetric algebraic Riccati equation, and we report on some known methods and new algorithmic advances. In particular, some results on the number of positive solutions are proved and a careful convergence analysis of Newton's iteration is carried out in the cases of interest where some singularity conditions are encountered. From this analysis we determine initial approximations which still guarantee the quadratic convergence.

1 Introduction

Nonsymmetric Algebraic Riccati equations (NARE) are quadratic matrix equations of the kind

$$XCX - AX - XD + B = 0, \tag{1}$$

where the unknown X is an $m \times n$ matrix, and the coefficients A, B, C and D have sizes $m \times m$, $m \times n$, $n \times m$ and $n \times n$, respectively.

The term *nonsymmetric* distinguishes this case from the widely studied continuous-time algebraic Riccati equations, defined by the quadratic matrix equation $XCX - AX - XA^T + B = 0$, where B and C are symmetric. We refer the reader to the books [38, 44] for a comprehensive analysis of continuous-time algebraic Riccati equations.

The matrix coefficients of the NARE (1) define the $(m+n) \times (m+n)$ matrix

$$M = \begin{bmatrix} D & -C \\ -B & A \end{bmatrix}, \tag{2}$$

which, throughout the paper, we assume to be an M-matrix. This assumption is motivated by the increasing applicative interest of this kind of NAREs, and by the recent theoretical and algorithmic advances that have been achieved.

We recall that M is an M-matrix if it can be written as $M = \alpha I - N$ where N has nonnegative entries, $\alpha \geqslant \rho(N)$, and $\rho(N)$ is the spectral radius of N. We say that equation (1) is associated with an M-matrix if its coefficients form an M-matrix M.

*This work was supported by MIUR grant number 2004015437

There are two important applications where nonsymmetric algebraic Riccati equations associated with M-matrices play a very important role: the study of fluid queues models [47, 46, 48], and the analysis of transport equations [35, 34]. In both cases the solution of interest is the matrix S with nonnegative entries, which among all the nonnegative solutions is the one with componentwise minimal entries. We call any solution S sharing this property *minimal nonnegative solution*. These applications will be outlined in Sections 1.1 and 1.2.

The research activity concerning the analysis of NAREs associated with M-matrices and the design of numerical algorithms for their solution has had a strong acceleration in the last decade. Important progress has been made concerning theoretical properties of this class of matrix equations and new effective algorithms relying on the properties of M-matrices have been designed and analyzed [6, 10, 11, 13, 15, 18, 20, 21, 23, 24, 25, 26, 32, 35, 41, 42, 43].

In this paper we provide a survey of the most important results and of the most effective algorithms concerning the analysis and the numerical treatment of the NAREs associated with M-matrices together with some new results. We also provide a unifying framework where different techniques and properties can be described in a simpler form and where more insights into the properties of matrix equations are given.

In particular, we report on results concerning the existence of a minimal nonnegative solution S and prove some new results on the number of nonnegative solutions of the NARE (1). We analyze the spectral properties of the matrix

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

relate them to the eigenvalues of S and use this relation when H is singular to classify the problem into three classes according to the sign of the *drift* associated with the equation.

After reporting on perturbation results of the solution, we present, analyze and compare different algorithms for computing the minimal nonnegative solution S. Besides a "direct" method based on the Schur decomposition we consider functional iterations having linear convergence and then the Newton iteration which has a generally quadratic speed of convergence. The class of doubling algorithms is discussed. This class includes cyclic reduction and the Structure-preserving Doubling Algorithm (SDA). Here we report very recent results relating these two algorithms, in particular the proof that SDA turns out to be cyclic reduction applied to a specific problem.

We give a separate treatment on the case of interest where the associated matrix M is singular. Here we prove that a particular choice of the initial point in Newton's iteration can restore the quadratic speed of convergence which otherwise would be linear. In fact, we provide a simple but general result concerning the "structured" convergence of Newton's iteration.

Furthermore, we discuss the possibility of replacing the original equation with a different one, having the same solution S, but where the singularity is removed. The advantage that we get with this technique is twofold: on one hand we can accelerate the speed of iterative methods by switching from the linear to the quadratic convergence; on the other hand we may guarantee the full machine accuracy ε in the solution which otherwise would be $O(\sqrt{\varepsilon})$.

Numerical experiments which validate our theoretical analysis conclude the paper.

The paper is structured as follows: in Sections 1.1 and 1.2 we describe the applications of NAREs; in Section 2 we deal with theoretical properties and in Section 3 with algorithms; the case where M is singular is discussed in Section 4 while Section 5 reports the results of numerical experiments and the concluding remarks.

1.1 Application to fluid queues

In the analysis of two dimensional continuous-time Markov processes, called fluid queues, a crucial step is to compute the element-wise minimal nonnegative solution S of the NARE (1). In [3, 46, 47, 17, 1, 6], the fluid flow models are described in terms of a two-dimensional continuous-time Markov process denoted by $\{(X(t), \varphi(t)), \quad t \geqslant 0\}$ where $X(t)$ represents the level, while $\varphi(t)$ represents the phase. The phase process $\{\varphi(t) : \ t \geqslant 0\}$ is an irreducible Markov chain with space state $\mathcal{S}_1 \cup \mathcal{S}_2$, $\mathcal{S}_1 = \{1, 2, \ldots, m\}$, $\mathcal{S}_2 = \{m+1, m+2, \ldots, m+n\}$, and infinitesimal generator the opposite of (2). The minimal nonnegative solution $S = (s_{i,j})$ of (1) is such that $s_{i,j}$ is the probability that, starting from level x in phase $i \in \mathcal{S}_2$, the process $(X(t), \varphi(t))$ first returns to level x in finite time and does so in phase $j \in \mathcal{S}_1$, while avoiding levels below x. A detailed description of this kind of models can be found in [46].

1.2 Application to transport equation

Riccati equations associated with M-matrices also appear in a problem in neutron transport theory, a variation of the one-group neutron transport equation, described in [35] where the mathematical model consists in solving an integrodifferential equation. After discretization of this integrodifferential equation, the problem can be expressed as the following equation for an unknown matrix $X \in \mathbb{R}^{n \times n}$

$$\Delta X + X\widehat{\Delta} = (Xq + e)(e^T + q^T X), \tag{3}$$

with

$$\begin{aligned} &\Delta = \operatorname{diag}(\delta_1, \ldots, \delta_n), && \widehat{\Delta} = \operatorname{diag}(\hat{\delta}_1, \ldots, \hat{\delta}_n), \\ &\delta_i = \frac{1}{cx_i(1-\alpha)}, && \hat{\delta}_i = \frac{1}{cx_i(1+\alpha)}, \quad i = 1, \ldots, n, \\ &e = \begin{bmatrix} 1 & 1 & \cdots & 1 \end{bmatrix}^T, && q_i = \frac{w_i}{2x_i}, \quad i = 1, \ldots, n. \end{aligned}$$

The matrices and vectors above depend on the two parameters $0 < c \leqslant 1$, $0 \leqslant \alpha < 1$, and on the sequences $(x_i)_{i=1}^n$ and $(w_i)_{i=1}^n$, which are the nodes and weights of a Gaussian quadrature on $[0, 1]$, ordered such that (x_i) is decreasing. The solution of physical interest is the minimal nonnegative one, whose existence can be proved thanks to Theorem 2.7 that we report in Section 2.3.

Equation (3) coincides with the NARE (1) with

$$\begin{aligned} &A = \widehat{\Delta} - eq^T, && B = ee^T, \\ &C = qq^T, && D = \Delta - qe^T. \end{aligned}$$

Under these hypotheses it is easy to prove that M is a diagonal-plus-rank-1 M-matrix. Due to this additional structure, ad-hoc algorithms can be developed, such as the ones described in [43, 42, 11]. Moreover, the iterates appearing when implementing most of the traditional algorithms are structured and can be completely described with $O(n)$ parameters. Therefore, structured versions of these algorithms can be developed, resulting in quadratically convergent iterations for (3) that require only $O(n^2)$ operations per step, as shown in [11] for Newton's method.

2 Theoretical properties

Before analyzing the numerical methods for the effective solution of equation (1), it is worth giving some theoretical properties of the NARE.

A large amount of properties concerning equation (1) have been stated in [18, 20, 21, 25, 35]; we summarize some of them. These results concern algebraic Riccati equations associated with nonsingular or singular irreducible M-matrices. The case in which M is singular and reducible is of minor interest.

A nonzero matrix $A = (a_{ij})$ is said *nonnegative* (*nonpositive*) if $a_{ij} \geqslant 0$ ($a_{ij} \leqslant 0$). In this case one writes $A \geqslant 0$ ($A \leqslant 0$). A matrix $A = (a_{ij})$ is said *positive* (*negative*) if $a_{ij} > 0$ ($a_{ij} < 0$). In this case one writes $A > 0$ ($A < 0$).

A matrix B is called a Z-matrix, if there exists a nonnegative matrix A such that $B = sI - A$, for a suitable scalar s. In other words a Z-matrix is a matrix all whose off-diagonal elements are nonpositive.

A matrix B is called an M-matrix, if it can be written in the form $B = sI - A$, where A is nonnegative, $s > 0$ and $s \geqslant \rho(A)$. If $s = \rho(A)$ the M-matrix is singular. Observe that an M-matrix is a Z-matrix.

We denote the set of the eigenvalues of A by $\sigma(A)$. Throughout the paper, e will denote the vector with components equal to 1, whose length is specified by the context.

2.1 Some useful facts about nonnegative matrices

A nonnegative matrix maps the cone of nonnegative vectors into itself; this cone contains an eigenvector as stated by the following celebrated result [8].

Theorem 2.1 (Perron–Frobenius theorem)**.** *Any nonnegative matrix A has a real eigenvalue $\lambda \geqslant 0$ such that $|\mu| \leqslant \lambda$ for each $\mu \in \sigma(A)$. Moreover, there exists a vector $v \geqslant 0$ such that $Av = \lambda v$.*

Any irreducible nonnegative matrix A has a real eigenvalue $\lambda > 0$ such that $|\mu| \leqslant \lambda$ for each $\mu \in \sigma(A)$. Moreover, λ is simple and there exists a vector $v > 0$ such that $Av = \lambda v$.

If A is positive, then $|\mu| < \lambda$ for each $\mu \in \sigma(A) \backslash \{\lambda\}$.

We state a useful corollary of the Perron–Frobenius theorem.

Corollary 2.2. *Let A be an irreducible nonnegative matrix and let $v_1, \ldots, v_n$ be a set of Jordan chains of A. Then there exists only one positive or negative vector among the v_i's and it is a scalar multiple of v.*

From the Perron–Frobenius theorem many interesting properties of Z- and M-matrices follow. For instance, a Z-matrix has a "leftmost" (in the complex plane) real eigenvalue corresponding to a nonnegative eigenvector, for an

M-matrix this eigenvalue is nonnegative. In particular, one deduces that the eigenvalues of an M-matrix have nonnegative real part.

A very common problem is to check if a given Z-matrix is an M-matrix. The following result states many equivalent conditions for a Z-matrix to be a nonsingular M-matrix. The proofs can be found in [8].

Theorem 2.3. *For a Z-matrix A, the following conditions are equivalent:*

(a) *A is a nonsingular M-matrix;*

(b) *$A^{-1} \geqslant 0$;*

(c) *$Au > 0$ for some vector $u > 0$;*

(d) *All the eigenvalues of A have positive real parts.*

Theorem 2.4. *For a Z-matrix A it holds that: A is an M-matrix if and only if there exists a nonzero vector $v \geqslant 0$ such that $Av \geqslant 0$ or a nonzero vector $w \geqslant 0$ such that $w^T A \geqslant 0$.*

The equivalence of (a) and (c) in Theorem 2.3 implies the next result.

Lemma 2.5. *Let A be a nonsingular M-matrix. If $B \geqslant A$ is a Z-matrix, then B is also a nonsingular M-matrix.*

The following well-known result concerns the properties of Schur complements of M-matrices.

Lemma 2.6. *Let M be a nonsingular M-matrix or an irreducible singular M-matrix. Partition M as*

$$M = \begin{bmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{bmatrix},$$

where M_{11} and M_{22} are square matrices. Then M_{11} and M_{22} are nonsingular M-matrices. The Schur complement of M_{11} (or M_{22}) in M is also an M-matrix (singular or nonsingular according to M). Moreover, the Schur complement is irreducible if M is irreducible.

2.2 The dual equation

Reverting the coefficients of equation (1) yields the *dual equation*

$$YBY - YA - DY + C = 0, \tag{4}$$

which is still a NARE, associated with the matrix

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

that is a nonsingular M-matrix or an irreducible singular M-matrix if and only if the matrix M is so. In fact N is clearly a Z-matrix and $N = \Pi M \Pi$, where $\Pi = \Pi^{-1}$ is the matrix which permutes the blocks of M. So, if $Mv \geqslant 0$, for $v \geqslant 0$, then $N\Pi v \geqslant 0$ and by Theorem 2.4, N is an M-matrix.

2.3 Existence of nonnegative solutions

The special structure of the matrix M of (2) allows one to prove the existence of a minimal nonnegative solution S of (1), i.e., a solution $S \geqslant 0$ such that $X - S \geqslant 0$ for any solution $X \geqslant 0$ to (1). See [20] and [21] for more details.

Theorem 2.7. *Let M in (2) be an M-matrix. Then the NARE (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 nonsingular, then $A - SC$ and $D - CS$ are nonsingular M-matrices.*

Observe that the above theorem holds for the dual equation (4) and guarantees the existence of a minimal nonnegative solution of (4) which is denoted by T.

2.4 The eigenvalue problem associated with the matrix equation

A useful technique frequently encountered in the theory of matrix equations consists in relating the solutions to some invariant subspaces of a matrix polynomial.

In particular, the solutions of (1) can be described in terms of the invariant subspaces of the matrix

$$H = \begin{bmatrix} D & -C \\ B & -A \end{bmatrix}, \tag{5}$$

which is obtained premultiplying the matrix M by $J = \begin{bmatrix} I_n & 0 \\ 0 & -I_m \end{bmatrix}$.

In fact, if X is a solution of equation (1), then, by direct inspection,

$$H \begin{bmatrix} I_n \\ X \end{bmatrix} = \begin{bmatrix} I_n \\ X \end{bmatrix} R, \tag{6}$$

where $R = D - CX$. Moreover, the eigenvalues of the matrix R are a subset of the eigenvalues of H. Conversely, if the columns of the $(n+m) \times n$ matrix $\begin{bmatrix} Y \\ Z \end{bmatrix}$ span an invariant subspace of H, and Y is a nonsingular $n \times n$ matrix, then ZY^{-1} is a solution of the Riccati equation [38].

Similarly, for the solutions of the dual equation it holds that

$$H \begin{bmatrix} Y \\ I_m \end{bmatrix} = \begin{bmatrix} Y \\ I_m \end{bmatrix} U,$$

where $U = BY - A$. The eigenvalues of the matrix U are a subset of the eigenvalues of H.

2.5 The eigenvalues of H

We say that a set $\mathcal{A}$ of k complex numbers has a (k_1, k_2) splitting with respect to the unit circle if $k = k_1 + k_2$, and $\mathcal{A} = \mathcal{A}_1 \cup \mathcal{A}_2$, where $\mathcal{A}_1$ is formed by k_1 elements of modulus at most 1 and $\mathcal{A}_2$ is formed by k_2 elements of modulus at least 1. Similarly, we say that $\mathcal{A}$ has a (k_1, k_2) splitting with respect to

the imaginary axis if $k = k_1 + k_2$, and $A = \mathcal{A}_1 \cup \mathcal{A}_2$, where $\mathcal{A}_1$ is formed by k_1 elements with nonpositive real part and $\mathcal{A}_2$ is formed by k_2 elements with nonnegative real part. We say that the splitting is *complete* if at lest one set $\mathcal{A}_1$ or $\mathcal{A}_2$ has no eigenvalues in its boundary.

Since the eigenvalues of an M-matrix have nonnegative real part, it follows that the eigenvalues of H have an (m, n) splitting with respect to the imaginary axis. This property is proved in the next

Theorem 2.8. *Let M be an irreducible M-matrix. Then the eigenvalues of $H = JM$ have an (m, n) splitting with respect to the imaginary axis. Moreover, the only eigenvalue that can lie on the imaginary axis is* 0.

Proof. Let $v > 0$ be the only positive eigenvector of M, and let $\lambda \geqslant 0$ be the associate eigenvalue; define $D_v = \mathrm{diag}(v)$. The matrix $\overline{M} = D_v^{-1} M D_v$ has the same eigenvalues as M; moreover, it is an M-matrix such that $\overline{M}e = \lambda e$. Due to the sign structure of M-matrices, this means that $\overline{M}$ is diagonal dominant (strictly in the nonsingular case). Notice that $\overline{H} = D_v^{-1} H D_v = J\overline{M}$, thus $\overline{H}$ is diagonal dominant as well, with m negative and n positive diagonal entries. We apply Gershgorin's theorem [30, Sec. 14] to $\overline{H}$; due to the diagonal dominance, the Gershgorin circles never cross the imaginary axis (in the singular case, they are tangent in 0). Thus, by using a continuity argument we can say that m eigenvalues of $\overline{H}$ lie in the negative half-plane and n in the positive one, and the only eigenvalues on the imaginary axis are the zero ones. But since H and $\bar{H}$ are similar, they have the same eigenvalues. □

We can give a more precise result on the location of the eigenvalues of H, after defining the *drift* of the Riccati equation. Indeed, when M is a singular irreducible M-matrix, by the Perron–Frobenius theorem, the eigenvalue 0 is simple, there are positive vectors u and v such that

$$u^T M = 0, \quad Mv = 0, \tag{7}$$

and both the vectors u and v are unique up to a scalar factor.

Writing $u = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}$ and $v = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}$, with $u_1, v_1 \in \mathbb{R}^n$ and $u_2, v_2 \in \mathbb{R}^m$, one can define

$$\mu = u_2^T v_2 - u_1^T v_1 = -u^T J v. \tag{8}$$

The number μ determines some properties of the Riccati equation. Depending on the sign of μ and following a Markov chain terminology, one can call μ the *drift* as in [6], and can classify the Riccati equations associated with a singular irreducible M-matrix in three categories:

(a) *positive recurrent* if $\mu < 0$;

(b) *null recurrent* if $\mu = 0$;

(c) *transient* if $\mu > 0$.

In fluid queues problems, v coincides with the vector of ones. In general v and u can be computed by performing the LU factorization of the matrix M, say $M = LU$, and solving the two triangular linear systems $u^T L = [0, \ldots, 0, 1]$ and $Uv = 0$ (see [30, Sec. 54]).

The location of the eigenvalues of H is made precise in the following [20, 23]:

Theorem 2.9. *Let M be a nonsingular or a singular irreducible M-matrix, and let $\lambda_1, \ldots, \lambda_{m+n}$ be the eigenvalues of $H = JM$ ordered by nonincreasing real part. Then λ_n and λ_{n+1} are real and*

$$\mathrm{Re}\lambda_{n+m} \leqslant \ldots \leqslant \mathrm{Re}\lambda_{n+2} < \lambda_{n+1} \leqslant 0 \leqslant \lambda_n < \mathrm{Re}\lambda_{n-1} \leqslant \ldots \leqslant \mathrm{Re}\lambda_1.$$

The minimal nonnegative solutions S and T of the equation (1) *and of the dual equation* (4)*, respectively, are such that $\sigma(D - CS) = \{\lambda_1, \ldots, \lambda_n\}$ and $\sigma(A - SC) = \sigma(A - BT) = \{-\lambda_{n+1}, \ldots, -\lambda_{n+m}\}$.*

If M is nonsingular then $\lambda_{n+1} < 0 < \lambda_n$. If M is singular and irreducible then:

1. *if $\mu < 0$ then $\lambda_n = 0$ and $\lambda_{n+1} < 0$;*
2. *if $\mu = 0$ then $\lambda_n = \lambda_{n+1} = 0$ and there exists only one eigenvector, up to a scalar constant, for the eigenvalue 0;*
3. *if $\mu > 0$ then $\lambda_n > 0$ and $\lambda_{n+1} = 0$.*

We call λ_n and λ_{n+1} the *central eigenvalues* of H. If H (and thus M) is nonsingular, then the central eigenvalues lie on two different half planes so the splitting is complete. In the singular case the splitting is complete if and only if $\mu \neq 0$.

The *close to null recurrent* case, i.e., the case $\mu \approx 0$, deserves particular attention, since it corresponds to an ill-conditioned null eigenvalue for the matrix H. In fact, if u and v are normalized such that $\|u\|_2 = \|v\|_2 = 1$, then $1/|\mu|$ is the condition number of the null eigenvalue for the matrix H (see [19]).

When M is singular irreducible, for the Perron–Frobenius theorem the eigenvalue 0 is simple, therefore $H = JM$ has a one dimensional kernel and $u^T J$ and v are the unique (up to a scalar constant) left and right eigenvectors, respectively, corresponding to the eigenvalue 0. However the algebraic multiplicity of 0 as an eigenvalue of H can be 2; in that case, the Jordan form of H has a 2×2 Jordan block corresponding to the 0 eigenvalue and it holds $u^T J v = 0$ [31].

The next result, presented in [25], shows the reduction from the case $\mu < 0$ to the case $\mu > 0$ and conversely, when M is singular irreducible. This property enable us to restrict our interest only to the case $\mu \leqslant 0$.

Lemma 2.10. *The matrix S is the minimal nonnegative solution of* (1) *if and only if $Z = S^T$ is the minimal nonnegative solution of the equation*

$$XC^T X - XA^T - D^T X + B^T = 0. \tag{9}$$

Therefore, if M is singular and irreducible, the equation (1) *is transient if and only if the equation* (9) *is positive recurrent.*

Proof. The first part is easily shown by taking transpose on both sides of the equation (1). The M-matrix corresponding to (9) is

$$M_t = \begin{bmatrix} A^T & -C^T \\ -B^T & D^T \end{bmatrix}.$$

Since

$$\begin{bmatrix} v_2^T & v_1^T \end{bmatrix} M_t = 0, \quad M_t \begin{bmatrix} u_2 \\ u_1 \end{bmatrix} = 0,$$

the second part readily follows. □

2.6 The differential of the Riccati operator

The matrix equation (1) defines a Riccati operator

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

whose differential $d\mathcal{R}_X$ at a point X is

$$d\mathcal{R}_X[H] = HCX + XCH - AH - HD. \tag{10}$$

The differential $H \to d\mathcal{R}_X[H]$ is a linear operator which can be represented by the matrix

$$\Delta_X = (CX - D)^T \otimes I_m + I_n \otimes (XC - A), \tag{11}$$

where $\otimes$ denotes the Kronecker product (see [30, Sec. 10]).

We say that a solution X of the matrix equation (1) is critical if the matrix Δ_X is singular.

From the properties of Kronecker product [30, Sec. 10], it follows that the eigenvalues of Δ_X are the sums of those of $CX - D$ and $XC - A$. If $X = S$, where S is the minimal nonnegative solution, then $D - CX$ and $A - XC$ are M-matrices (compare Theorem 2.7), and thus all the eigenvalues of Δ_S have nonpositive real parts. Moreover, since $D - CS$ and $A - SC$ are M-matrices then $-\Delta_S$ is an M-matrix. The minimal nonnegative solution S is critical if and only if both M-matrices $D - CS$ and $A - SC$ are singular, thus, in view of Theorem 2.9, the minimal solution is critical if and only if M is irreducible singular and $\mu = 0$.

Moreover, if $0 \leqslant X \leqslant S$ then $D - CX \geqslant D - CS$ and $A - XC \geqslant A - SC$ are nonsingular M-matrices by lemma 2.5, thus $-\Delta_X$ is a nonsingular M-matrix.

2.7 The number of positive solutions

If the matrix M is irreducible, Theorem 2.7 states that there exists a minimal positive solution S of the NARE. In the study of nonsymmetric Riccati differential equations associated with an M-matrix [18, 34] one is interested in all the positive solutions.

In [18] it is shown that if M is nonsingular or singular irreducible with $\mu \neq 0$, then there exists a second solution S_+ such that $S_+ > S$ and S_+ is obtained by a rank one correction of the matrix S. More precisely, the following result holds [18].

Theorem 2.11. *If M is irreducible nonsingular or irreducible singular with $\mu \neq 0$, then there exists a second positive solution S_+ of (1) given by*

$$S_+ = S + kab^T,$$

where $k = (\lambda_n - \lambda_{n+1})/b^T Ca$, a is such that $(A - SC)a = -\lambda_{n+1}a$ and b is such that $b^T(D - CS) = \lambda_n b^T$.

We prove that there are exactly two nonnegative solutions in the noncritical case and only one in the critical case. In order to prove this result it is useful to study the form of the Jordan chains of an invariant subspace of H corresponding to a positive solution.

Lemma 2.12. *Let M be irreducible and let Σ be any positive solution of* (1). *Denote by $\eta_1, \ldots, \eta_n$ the eigenvalues of $D - C\Sigma$ ordered by nondecreasing real part. Then η_1 is real, and there exists a positive eigenvector v of H associated with η_1. Moreover, any other vector independent of v, belonging to Jordan chains of H corresponding to $\eta_1, \ldots, \eta_n$ cannot be positive or negative.*

Proof. Since Σ is a solution of (1), then from (6) one has

$$H \begin{bmatrix} I \\ \Sigma \end{bmatrix} = \begin{bmatrix} I \\ \Sigma \end{bmatrix} (D - C\Sigma).$$

Since $D - CS$ is an irreducible M-matrix for Theorem 2.7, and $\Sigma \geqslant S$ (S is the minimal positive solution), then $D - C\Sigma$ is an irreducible Z-matrix and thus can be written as $sI - N$ with N nonnegative and irreducible. Then by Theorem 2.1 and Corollary 2.2 η_1 is a simple real eigenvalue of $D - C\Sigma$, the corresponding eigenvector can be chosen positive and there are no other positive or negative eigenvectors or Jordan chains corresponding to any of the eigenvalues. Let $P^{-1}(D - C\Sigma)P = K$ be the Jordan canonical form of $D - C\Sigma$, where the first column of P is the positive eigenvector corresponding to η_1. Then we have

$$H \begin{bmatrix} P \\ \Sigma P \end{bmatrix} = \begin{bmatrix} P \\ \Sigma P \end{bmatrix} K.$$

Thus, the columns of $\begin{bmatrix} P \\ \Sigma P \end{bmatrix}$ are the Jordan chains of H corresponding to $\eta_1, \ldots, \eta_n$, and there are no positive or negative columns, except for the first one. □

Theorem 2.13. *If M is an irreducible nonsingular M-matrix or an irreducible singular M-matrix with $\mu \neq 0$, then* (1) *has exactly two positive solutions. If M is irreducible singular with $\mu = 0$, then* (1) *has a unique positive solution.*

Proof. From Lemma 2.12 applied to S it follows that H has a positive eigenvector corresponding to λ_n, and no other positive or negative eigenvectors or Jordan chains corresponding to $\lambda_1, \ldots, \lambda_n$. Let T be the minimal nonnegative solution of the dual equation (4). Then

$$H \begin{bmatrix} T \\ I \end{bmatrix} = \begin{bmatrix} T \\ I \end{bmatrix} (-(A - BT)).$$

As in the proof of Lemma 2.12, we can prove that H has a positive eigenvector corresponding to the eigenvalue λ_{n+1} and no other positive or negative eigenvectors or Jordan chains corresponding to $\lambda_{n+1}, \ldots, \lambda_{n+m}$.

If M is irreducible nonsingular, or irreducible singular with $\mu \neq 0$, then $\lambda_n > \lambda_{n+1}$, and there are only two linearly independent positive eigenvectors corresponding to real eigenvalues. By Lemma 2.12, there can be at most two solutions corresponding to $\lambda_n, \lambda_{n-1}, \ldots, \lambda_1$, and to $\lambda_{n+1}, \lambda_{n-1}, \ldots, \lambda_1$, respectively. Since it is know from Theorem 2.11 that there exist at least two positive solutions, thus (1) has exactly two positive solutions.

If M is irreducible singular with $\mu = 0$, there is only one positive eigenvector corresponding to $\lambda_n = \lambda_{n+1}$, and the unique solution of (1) is obtained by the Jordan chains corresponding to $\lambda_n, \lambda_{n-1}, \ldots, \lambda_1$. □

The next results provide a useful property of the minimal solutions which will be useful in Section 4.

Theorem 2.14. *Let M be singular and irreducible, and let S and T be the minimal nonnegative solutions of* (1) *and* (4)*, respectively. Then the following properties hold:*

(a) *if $\mu < 0$, then $Sv_1 = v_2$ and $Tv_2 < v_1$;*

(b) *if $\mu = 0$, then $Sv_1 = v_2$ and $Tv_2 = v_1$;*

(c) *if $\mu > 0$, then $Sv_1 < v_2$ and $Tv_2 = v_1$.*

Proof. From the proof of Theorem 2.13, it follows that if $\mu \neq 0$, there exist two independent positive eigenvectors a and b of H relative to the central eigenvalues λ_n and λ_{n+1}, respectively. We write $a = \begin{bmatrix} a_1 \\ a_2 \end{bmatrix}$ and $b = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}$, with $a_1, b_1 \in \mathbb{R}^n$ and $a_2, b_2 \in \mathbb{R}^m$.

Since the solution S is constructed from an invariant subspace containing a, then $Sa_1 = a_2$, since the solution S_+ is constructed from an invariant subspace containing b, then $S_+ b_1 = b_2$. Analogously, if T_+ is the second positive solution of the dual equation, then $Tb_2 = b_1$ and $T_+ a_2 = a_1$.

The statements (a) and (c) follow from the fact that if $\mu < 0$ then $v = a$ (compare Theorem 2.9), so $Sv_1 = v_2$ and $Tv_2 < T_+ v_2 = v_1$, since $T < T_+$; if $\mu > 0$ then $v = b$, so $Tv_2 = v_1$ and $Sv_1 < S_+ v_1 = v_2$, since $S < S_+$.

The statement (b) corresponding to the case $\mu = 0$ can be proved in a similar way. □

Remark 2.15. When $\mu \geqslant 0$, from Lemma 2.10 and Theorem 2.14 we deduce that the minimal nonnegative solution S of (1) is such that $u_2^T S = u_1^T$.

2.8 Perturbation analysis for the minimal solution

We conclude this section with a result of Guo and Higham [24] who perform a qualitative description of the perturbation of the minimal nonnegative solution S of a NARE (1) associated with an M-matrix.

The result is split in two theorems where an M-matrix $\widetilde{M}$ is considered which is obtained by means of a *small* perturbation of M. Here, we denote by $\widetilde{S}$ the minimal nonnegative solution of the perturbed Riccati equation associated with $\widetilde{M}$.

Theorem 2.16. *If M is a nonsingular M-matrix or an irreducible singular M-matrix with $\mu \neq 0$, then there exist constants $\gamma > 0$ and $\varepsilon > 0$ such that $\|\widetilde{S} - S\| \leqslant \gamma \|\widetilde{M} - M\|$ for all $\widetilde{M}$ with $\|\widetilde{M} - M\| < \varepsilon$.*

Theorem 2.17. *If M is an irreducible singular M-matrix with $\mu = 0$, then there exist constants $\gamma > 0$ and $\varepsilon > 0$ such that*

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

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

It is interesting to observe that in the critical case, where $\mu = 0$ or if $\mu \approx 0$, one has to expect poor numerical performances even if the algorithm used for approximating S is backward stable. Moreover, the rounding errors introduced to represent the input values of M in the floating point representation with precision ε may generate an error of the order $\sqrt{\varepsilon}$ in the solution S.

This kind of problems will be overcome in Section 4.1.

3 Numerical methods

We give a brief review of the numerical methods developed so far for computing the minimal nonnegative solution of the NARE (1) associated with an M-matrix. Here we consider the case where the M-matrix M is nonsingular or is singular, irreducible and $\mu \leqslant 0$. The case $\mu > 0$ can be reduced to the case $\mu < 0$ by means of Lemma 2.10. The critical case where $\mu = 0$ needs different techniques which will be treated in the next Section 4.

We start with a direct method based on the Schur form of the matrix H then we consider iterative methods based on fixed-point techniques, Newton's iteration and we conclude the section by analyzing a class of doubling algorithms.

The latter class includes methods based on Cyclic Reduction (CR) of [12], and on the Structure-preserving Doubling Algorithm (SDA) of [2].

3.1 Schur method

A classical approach for solving equation (1) is to use the (ordered) Schur decomposition of the matrix M to compute the invariant subspaces of H corresponding to the minimal solution S. This approach for the symmetric algebraic Riccati equation was first presented by Laub in 1979 [40]. Concerning the NARE, a study of that method in the singular and critical case was done by Guo [23] who presented a modified Schur method for the critical or near critical case ($\mu \approx 0$).

As explained in Section 2.4 from

$$H \begin{bmatrix} I_n \\ S \end{bmatrix} = \begin{bmatrix} I_n \\ S \end{bmatrix} (D - CS)$$

it follows that finding the minimal solution S of the NARE (1) is equivalent to finding a basis of the invariant subspace of H relative to the eigenvalues of $D - CS$, i.e., the eigenvalues of H with nonnegative real part.

A method for finding an invariant subspace is obtained by computing a semi-ordered Schur form of H, that is, computing an orthogonal matrix Q and a quasi upper-triangular matrix T such that $Q^*HQ = T$, where T is block upper triangular with diagonal blocks $T_{i,i}$ of size at most 2. The semi-ordering means that if $T_{i,i}$, $T_{j,j}$ and $T_{k,k}$ are diagonal blocks having eigenvalues with positive, null and negative real parts, respectively, then $i < j < k$.

A semi-ordered Schur form can be computed in two steps:

- Compute a real Schur form of H by the customary Hessenberg reduction followed by the application of the QR algorithm as described in [19].
- Swap the diagonal blocks by means of orthogonal transformations as described in [4].

The minimal solution of the NARE can be obtained from the first n columns of the matrix Q partitioned as $\begin{bmatrix} Q_1 \\ Q_2 \end{bmatrix}$ such that Q_1 is an $n \times n$ matrix, that is, $S = Q_2 Q_1^{-1}$.

In the critical case this method does not work, since there is no way to choose an invariant subspace relative to the first n eigenvalues, moreover in the near critical case where $\mu \approx 0$, there is lack of accuracy since the 0 eigenvalue is ill-conditioned. However, the modified Schur method given by C.-H. Guo [24] overcomes these problems.

The cost of this algorithm is $200n^3$ ops [23].

3.2 Functional iterations

In [20] a class of fixed-point methods for (1) is considered.

The fixed-point iterations are based on suitable splittings of A and D, that is $A = A_1 - A_2$ and $D = D_1 - D_2$, with A_1, D_1 chosen to be M-matrices and $A_2, D_2 \geqslant 0$. The form of the iterations is

$$A_1 X_{k+1} + X_{k+1} D_1 = X_k C X_k + X_k D_2 + A_2 X_k + B, \tag{12}$$

where at each step a Sylvester equation of the form $M_1 X + X M_2 = N$ must be solved.

Some possible choices for the splitting are:

1. A_1 and D_1 are the diagonal parts of A and D, respectively;
2. A_1 is the lower triangular part of A and D_1 the upper triangular part of D;
3. $A_1 = A$ and $D_1 = D$.

The solution X_{k+1} of the Sylvester equation can be computed, for instance, by using the Bartels and Stewart method [5], as in MATLAB's `sylvsol` function of the Nick Higham Matrix Function toolbox [28]

The cost of this computation is roughly $60n^3$ ops including the computation of the Schur form of the coefficients A_1 and D_1 [29]. However, observe that for the first splitting, A_1 and D_1 are diagonal matrices and the Sylvester equation can be solved with $O(n^2)$ ops; for the second splitting, the matrices A_1 and D_1 are already in the Schur form. This substantially reduces the cost of the application of the Bartels and Stewart method to $2n^3$. Concerning the third iteration, observe that the matrix coefficients A_1 and D_1 are independent of the iteration. Therefore, the computation of their Schur form must be performed only once.

A monotonic convergence result holds for the three iterations [20].

Theorem 3.1. *If $\mathcal{R}(X) \leqslant 0$ for some positive matrix X, then for the fixed-point iterations* (12) *with $X_0 = 0$, it holds that $X_k < X_{k+1} < X$ for $k \geqslant 0$. Moreover, $\lim X_k = S$.*

We have also an asymptotic convergence result [20].

Theorem 3.2. *For the fixed-point iterations* (12) *with $X_0 = 0$, it holds that*

$$\limsup \sqrt[k]{\|X_k - S\|} = \rho((I \otimes A_1 + D_1^T \otimes I)^{-1}(I \otimes (A_2 + SC) + (D_2 + CS)^T \otimes I).$$

These iterations have linear convergence which turns to sublinear in the critical case. The computational cost varies from $8n^3$ arithmetic operations per step for the first splitting, to $64n^3$ for the first step plus $10n^3$ for each subsequent step for the last splitting. The most expensive iteration is the third one which, on the other hand, has the highest (linear) convergence speed.

3.3 Newton's method

Newton's iteration was first applied to the symmetric algebraic Riccati equation by Kleinman in 1968 [37] and later on by various authors. In particular, Benner and Byers [7] complemented the method with an optimization technique (exact line search) in order to reduce the number of steps needed for arriving at convergence. The study of the Newton method for nonsymmetric algebraic Riccati equations was started by Guo and Laub in [26], and a nice convergence result was given by Guo and Higham in [24].

The convergence of the Newton method is generally quadratic except for the critical case where the convergence is observed to be linear with rate 1/2 [26]. At each step, a Sylvester matrix equation must be solved, so the computational cost is $O(n^3)$ ops per step, but with a large overhead constant.

The Newton method for a NARE [26] consists in the iteration

$$X_{k+1} = \mathcal{N}(X_k) = X_k - (d\mathcal{R}_{X_k})^{-1}\mathcal{R}(X_k), \quad k = 0, 1, \ldots \tag{13}$$

which, in view of (10), can be written explicitly as

$$(A - X_k C)X_{k+1} + X_{k+1}(D - CX_k) = B - X_k C X_k. \tag{14}$$

Therefore, the matrix X_{k+1} is obtained by solving a Sylvester equation. This linear equation is defined by the matrix

$$\Delta_{X_k} = (D - CX_k)^T \otimes I_m + I_n \otimes (A - X_k C)$$

which is nonsingular if $0 \leqslant X_k < S$, as shown in section 2.6. Thus, if $0 \leqslant X_k < S$ for any k, the sequence (13) is well-defined.

In the noncritical case, $d\mathcal{R}_S$ is nonsingular, and the iteration is quadratically convergent in a neighborhood of the minimal nonnegative solution S by the traditional results on Newton's method (see e.g. [36]). Moreover, the following monotonic convergence result holds [24]:

Theorem 3.3. *Consider Newton's method* (14) *starting from* $X_0 = 0$. *Then for each* $k = 0, 1, \ldots$, *we have* $0 \leqslant X_k \leqslant X_{k+1} < S$ *and* Δ_{X_k} *is a nonsingular M-matrix. Therefore, the sequence* (X_k) *is well-defined and converges monotonically to* S.

The same result holds when $0 \leqslant X_0 \leqslant S$; the proof in [24] can be easily adapted to this case.

In [26], a hybrid method was suggested, which consists in performing a certain number of iterations of a linearly convergent algorithm, such as the ones of Section 3.2, and then using the computed value as the starting point for Newton's method.

At each step of Newton's iteration, the largest computational work is given by the solution of the Sylvester equation (14). We recall that the solution X_{k+1},

computed by means of the Bartels and Stewart method [5] costs roughly $60n^3$ ops. Therefore the overall cost of newton's iteration is $66n^3$ ops.

It is worth noting that in the critical and near critical cases, the matrix Δ_k becomes almost singular as X_k approaches the solution S; therefore, some numerical instability is to be expected. Such instability can be removed by means of a suitable technique which we will describe in Section 4.1.

3.4 Doubling algorithms

In this section we report some quadratically convergent algorithms obtained in [10] for solving (1). Quadratically convergent methods for computing the extremal solution of the NARE can be obtained by transforming the NARE into a Unilateral Quadratic Matrix Equation (UQME) of the kind

$$A_2X^2 + A_1X + A_0 = 0 \tag{15}$$

where A_0, A_1, A_2 and X are $p \times p$ matrices. Equations of this kind can be solved efficiently by means of doubling algorithms like Cyclic Reduction (CR) [12, 9] or Logarithmic Reduction (LR) [39].

The first attempt to reduce a NARE to a UQME was performed by Ramaswami [46] in the framework of fluid queues. Subsequently, many contributions in this direction have been given by several authors [23, 13, 10, 33, 6] and different reduction techniques have been designed.

Concerning algorithms, Cyclic Reduction and SDA are the most effective computational techniques. The former was applied the first time in [12] by D. Bini and B. Meini to solve unilateral quadratic equations. The latter, was first presented by Anderson in 1978 [2] for the numerical solution of discrete-time algebraic Riccati equations. A new interpretation was given by Chu, Fan, Guo, Hwang, Lin, Xu [16, 32, 41], for other kinds of algebraic Riccati equations.

CR applied to (15) generates sequences of matrices defined by the following equations

$$\begin{aligned}
&V^{(k)} = (A_1^{(k)})^{-1} \\
&A_0^{(k+1)} = -A_0^{(k)}V^{(k)}A_0^{(k)} \\
&A_1^{(k+1)} = A_1^{(k)} - A_0^{(k)}V^{(k)}A_2^{(k)} - A_2^{(k)}V^{(k)}A_0^{(k)} \quad k = 0, 1, \dots \\
&A_2^{(k+1)} = -A_2^{(k)}V^{(k)}A_2^{(k)} \\
&\widehat{A}^{(k+1)} = \widehat{A}^{(k)} - A_2^{(k)}V^{(k)}A_0^{(k)}
\end{aligned} \tag{16}$$

where $A_i^{(0)} = A_i$, $i = 0, 1, 2$, $\widehat{A}^{(0)} = A_1$.

The following result provides convergence properties of CR [9].

Theorem 3.4. *Let $x_1, \dots, x_{2p}$ be the roots of $a(z) = \det(A_0 + zA_1 + z^2A_2)$, including roots at the infinity if $\deg a(z) < 2p$, ordered by increasing modulus. Suppose that $|x_p| \leqslant 1 \leqslant |x_{p+1}|$ and $|x_p| < |x_{p+1}|$, and that a solution G exists to* (15) *such that $\rho(G) = |x_p|$. Then, G is the unique solution to* (15) *with minimal spectral radius, moreover, if CR* (16) *can be carried out with no breakdown, the sequence*

$$G^{(k)} = -\left(\widehat{A}^{(k)}\right)^{-1} A_0$$

is such that for any norm

$$||G^{(k)} - G|| \leqslant \vartheta |x_p/x_{p+1}|^{2^k}$$

where $\vartheta > 0$ is a suitable constant. Moreover, it holds that $||A_0^{(k)}|| = O(|x_p|^{2^k})$, $||A_2^{(k)}|| = O(|x_{p+1}|^{-2^k})$.

Observe that, the convergence conditions of the above theorem require that the roots of $a(z)$ have a (p, p) complete splitting with respect to the unit circle. For this reason, before transforming the NARE into a UQME, it is convenient to transform the Hamiltonian H into a new matrix $\widehat{H}$ such that the eigenvalues of $\widehat{H}$ have an (n, m) splitting with respect to the unit circle, i.e., n eigenvalues belong to the closed unit disk and m are outside. This can be obtained by means of one of the two operators: the Cayley transform $\mathcal{C}_\gamma(z) = (z+\gamma)^{-1}(z-\gamma)$, where $\gamma > 0$, or the shrink-and-shift operator $\mathcal{S}_\tau(z) = 1 - \tau z$, where $\tau > 0$. In fact, the Cayley transform maps the right open half-plane into the open unit disk. Similarly, for suitable values of τ, the transformation $\mathcal{S}_\tau$ maps a suitable subset of the right half-plane inside the unit disk. This property is better explained in the following result which has been proved in [10].

Theorem 3.5. *Let $\gamma, \tau > 0$ and let*

$$H_\gamma = \mathcal{C}_\gamma(H) = (H + \gamma I)^{-1}(H - \gamma I), \quad H_\tau = \mathcal{S}_\tau(H) = I - \tau H.$$

Assume $\mu < 0$, then:

1. *H_γ has eigenvalues $\xi_i = \mathcal{C}_\gamma(\lambda_i)$, $i = 1, \ldots, m + n$, such that*

$$\max_{i=1,\ldots,n} |\xi_i| \leqslant 1 < \min_{i=1,\ldots,m} |\xi_{i+n}|;$$

2. *if $\tau^{-1} \geqslant \max\{\max_i(A)_{i,i}, \max_i(D)_{i,i}\}$, H_τ has eigenvalues $\mu_i = \mathcal{S}_\tau(\lambda_i)$, $i = 1, \ldots, m + n$, such that*

$$\max_{i=1,\ldots,n} |\mu_i| \leqslant 1 < \min_{i=1,\ldots,m} |\mu_{i+n}|.$$

Moreover, if X is any solution of (1) then

$$H_\gamma \begin{bmatrix} I \\ X \end{bmatrix} = \begin{bmatrix} I \\ X \end{bmatrix} R_\gamma, \quad H_\tau \begin{bmatrix} I \\ X \end{bmatrix} = \begin{bmatrix} I \\ X \end{bmatrix} R_\tau$$

where $R_\gamma = \mathcal{C}_\gamma(D - CX)$, $R_\tau = \mathcal{S}_\tau(D - CX)$.

In the following we will denote by $\widehat{H} = \begin{bmatrix} \widehat{D} & -\widehat{C} \\ \widehat{B} & -\widehat{A} \end{bmatrix}$ either H_γ or H_τ. Since the transformations $\mathcal{C}_\gamma$ and $\mathcal{S}_\tau$ are invertible, from the above theorem one has that X is a solution of the NARE (1) if and only if X is a solution of the NARE defined by $\widehat{H}$. In particular, the extremal solution S is the solution of the NARE associated with H_γ or H_τ corresponding to the n eigenvalues H_γ or H_τ, respectively, smallest in modulus.

The following result provides a means for reducing a NARE into a UQME:

Theorem 3.6. *Let X be a solution of the NARE* (1)*. Then:*

1. $Y = \begin{bmatrix} \widehat{D} - \widehat{C}X & 0 \\ X & 0 \end{bmatrix}$ *is a solution to*

$$\begin{bmatrix} \widehat{D} & 0 \\ \widehat{B} & 0 \end{bmatrix} + \begin{bmatrix} -I & -\widehat{C} \\ 0 & -\widehat{A} \end{bmatrix} Y + \begin{bmatrix} 0 & 0 \\ 0 & -I \end{bmatrix} Y^2 = 0; \tag{17}$$

2. $Y = \begin{bmatrix} \widehat{D} - \widehat{C}X & 0 \\ X(\widehat{D} - \widehat{C}X) & 0 \end{bmatrix}$ *is a solution to*

$$\begin{bmatrix} L_1 & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} -I & U_1 \\ L_2 & -I \end{bmatrix} Y + \begin{bmatrix} 0 & 0 \\ 0 & U_2 \end{bmatrix} Y^2 = 0, \tag{18}$$

where $U_1 = -\widehat{C}\widehat{A}^{-1}$, $U_2 = -\widehat{A}^{-1}$, $L_1 = \widehat{D} - \widehat{C}\widehat{A}^{-1}\widehat{B}$, $L_2 = -\widehat{A}^{-1}\widehat{B}$.

Conversely,

$$V = \begin{bmatrix} \widehat{D} - \widehat{C}S & 0 \\ S & 0 \end{bmatrix}, \quad W = \begin{bmatrix} \widehat{D} - \widehat{C}S & 0 \\ S(\widehat{D} - \widehat{C}S) & 0 \end{bmatrix}$$

are the unique solutions of UQME (17) *and* (18)*, respectively, with m eigenvalues equal to 0 and n eigenvalues in the closed unit disk.*

A reduction similar to the one provided in equation (17) was proved by Ramaswami in [46] by using probabilistic tools.

The following reduction holds for any NARE (1) provided that $m = n$ and $\det C \neq 0$

Theorem 3.7. *Let $m = n$ and $\det C \neq 0$. The matrix X is a solution of the NARE* (1) *if and only if $Y = C^{-1}(D - CX)\widehat{C}$ is a solution of the UQME*

$$Y^2 + (C^{-1}DC - A)Y + (B - AC^{-1}D)C = 0. \tag{19}$$

Similarly, X is a solution of the NARE (1) *if and only if $Y = D - CX$ is a solution of the UQME*

$$Y^2 + (D - CAC^{-1})Y + C(B - AC^{-1}D) = 0. \tag{20}$$

If we choose $H = \widehat{H}$, then $Y = \widehat{C}^{-1}(\widehat{D} - \widehat{C}X)\widehat{C}$ is the solution of the (19) with minimal spectral radius. Similarly, $Y = \widehat{D} - \widehat{C}S$ is the solution of (20) with minimal spectral radius.

Observe that if $\det C = 0$, we may replace (1) with a new equation defined by blocks $\widetilde{A}$, $\widetilde{B}$, $\widetilde{C}$, and $\widetilde{D}$ such that $\det \widetilde{C} \neq 0$ according to the following

Lemma 3.8. *The Riccati equation* (1) *has solution X if and only if the Riccati equation*

$$Y\widetilde{C}Y - \widetilde{A}Y - Y\widetilde{D} + \widetilde{B} = 0$$

where $\widetilde{A} = A - BK$, $\widetilde{B} = B$, $\widetilde{C} = \widetilde{\mathcal{R}}(K)$, $\widetilde{D} = D - KB$, has solution $\widetilde{X} = X(I - KX)^{-1}$ and K is such that $\det(I - K\widetilde{X}) \neq 0$ (or equivalently, $\det(I + XK) \neq 0$). Moreover, $\widetilde{D} - \widetilde{C}\widetilde{X} = (I - KX)(D - CX)(I - KX)^{-1}$.

It can be easily verified that if $\widehat{H} = H_\tau$ then

$$\widehat{A} = -I - \tau A, \quad \widehat{B} = -B, \quad \widehat{C} = -C, \quad \widehat{D} = I - \tau D.$$

If $\widehat{H} = H_\gamma$ then a direct calculation shows that

$$\begin{aligned} \widehat{A} &= -I + 2\gamma V^{-1}, & \widehat{B} &= 2\gamma(-A + \gamma I)^{-1} B W^{-1}, \\ \widehat{C} &= 2\gamma(D + \gamma I)^{-1} C V^{-1}, & \widehat{D} &= I - 2\gamma W^{-1}, \end{aligned}$$

with $V = -A + \gamma I + B(D + \gamma I)^{-1} C$ and $W = D + \gamma I + C(-A + \gamma I)^{-1} B$.

Equations (17), (18), (19) and (20) can be solved by means of CR (16), which provides a matrix sequence $G^{(k)}$ that converges, when applicable, to the solution with minimal spectral radius. In view of Theorem 3.5, and of the subsequent discussion, this solution is the one which is needed for computing the extremal solution S to the NARE (1).

The cost of CR applied to (19) and (20) is about $(38/3)n^3$ ops.

Concerning convergence it follows from Theorem 3.4 that the approximation error is $O(\sigma^{2^k})$, for $\sigma = \sigma_\gamma$ if $\widehat{H} = H_\gamma$, $\sigma = \sigma_\tau$ if $\widehat{H} = H_\tau$. Here we define

$$\begin{aligned} \sigma_\tau &= \max_{i=1,\ldots,n} |\mu_i| / \min_{i=1,\ldots,m} |\mu_{n+i}|, \\ \sigma_\gamma &= \max_{i=1,\ldots,n} |\xi_i| / \min_{i=1,\ldots,m} |\xi_{n+i}|, \end{aligned}$$

where $\sigma_\tau, \sigma_\gamma < 1$ if $\mu < 0$.

Applying CR to (19) and (20) generates blocks of size $m + n$. However, it is possible to verify that the structure of the blocks A_i, $i = 0, 1, 2$ given in equations (17) and (18) is maintained unchanged by the blocks $A_i^{(k)}$, $i = 0, 1, 2$. More precisely, it turns out that applying (16) to the equation (17) yields blocks of the kind

$$\begin{aligned} A_0^{(k)} &= \begin{bmatrix} R_1^{(k)} & 0 \\ R_2^{(k)} & 0 \end{bmatrix}, \quad A_1^{(k)} = \begin{bmatrix} -I & R_3^{(k)} \\ R_4^{(k)} & R_5^{(k)} \end{bmatrix}, \\ A_2^{(k)} &= \begin{bmatrix} 0 & 0 \\ 0 & R_6^{(k)} \end{bmatrix}, \quad \widehat{A}^{(k)} = \begin{bmatrix} -I & R_3^{(0)} \\ R_4^{(k)} & R_5^{(0)} \end{bmatrix}. \end{aligned}$$

It can be easily verified that the matrices $R_i^{(k)}$, $i = 1, \ldots, 6$ satisfy the following equations:

$$\begin{aligned} S^{(k)} &= R_5^{(k)} + R_4^{(k)} R_3^{(k)}, & R_1^{(k+1)} &= -R_1^{(k)} X^{(k)}, \\ Y^{(k)} &= \left(S^{(k)}\right)^{-1} \left(R_2^{(k)} + R_4^{(k)} R_1^{(k)}\right), & R_2^{(k+1)} &= -R_2^{(k)} X^{(k)}, \\ X^{(k)} &= R_3^{(k)} Y^{(k)} - R_1^{(k)}, & R_3^{(k+1)} &= R_3^{(k)} - R_1^{(k)} T^{(k)}, \\ Z^{(k)} &= \left(S^{(k)}\right)^{-1} R_6^{(k)}, & R_4^{(k+1)} &= R_4^{(k)} - R_6^{(k)} Y^{(k)}, \\ T^{(k)} &= R_3^{(k)} Z^{(k)}, & R_5^{(k+1)} &= R_5^{(k)} - R_2^{(k)} T^{(k)}, \\ & & R_6^{(k+1)} &= -R_6^{(k)} Z^{(k)}. \end{aligned} \tag{21}$$

for $k = 0, 1, \ldots,$ starting from the initial values $R_1^{(0)} = \widehat{D}$, $R_2^{(0)} = \widehat{B}$, $R_3^{(0)} = -\widehat{C}$, $R_4^{(0)} = 0$, $R_5^{(0)} = -\widehat{A}$, $R_6^{(0)} = -I$. From Theorem 3.4 it follows that

$$S = -\left(R_5^{(0)} + R_4^{(k)} R_3^{(0)}\right)^{-1} \left(R_2^{(0)} + R_4^{(k)} R_1^{(0)}\right) + O(\sigma^{2^k}),$$

where $\sigma = \sigma_\tau$ if $\widehat{H} = H_\tau$, while for $\widehat{H} = H_\gamma$ one has $\sigma = \sigma_\gamma$.

The computational cost of this algorithm is $(74/3)n^3$ per step, assuming $m = n$.

Similarly, it turns out that applying (16) to the equation (18) yields blocks of the kind

$$A_0^{(k)} = \begin{bmatrix} E^{(k)} & 0 \\ 0 & 0 \end{bmatrix}, \quad A_1^{(k)} = \begin{bmatrix} -I & G^{(k)} \\ H^{(k)} & -I \end{bmatrix}, \quad A_2^{(k)} = \begin{bmatrix} 0 & 0 \\ 0 & F^{(k)} \end{bmatrix},$$

where the sequences $E^{(k)}$, $F^{(k)}$, $G^{(k)}$, $H^{(k)}$ are given by

$$\begin{aligned} E^{(k+1)} &= E^{(k)}(I - G^{(k)}H^{(k)})^{-1}E^{(k)}, \\ F^{(k+1)} &= F^{(k)}(I - H^{(k)}G^{(k)})^{-1}F^{(k)}, \\ G^{(k+1)} &= G^{(k)} + E^{(k)}(I - G^{(k)}H^{(k)})^{-1}G^{(k)}F^{(k)}, \\ H^{(k+1)} &= H^{(k)} + F^{(k)}(I - H^{(k)}G^{(k)})^{-1}H^{(k)}E^{(k)}, \end{aligned} \tag{22}$$

for $k \geqslant 0$, starting from the initial values $E^{(0)} = L_1$, $F^{(0)} = U_2$, $G^{(0)} = U_1$, $H^{(0)} = L_2$. The following convergence result holds:

$$S = H^{(k)} + O(\sigma^{2^k})$$

in the noncritical case, where $\sigma = \sigma_\gamma, \sigma_\tau$. Observe that in this case the computation of $\widehat{A}^{(k)}$ is not required.

The cost per step of this iteration is $(64/3)n^3$ for $m = n$.

It is interesting to point out that (22), obtained by applying CR to the solution of the UQME (18), coincides with SDA of [16, 32, 41]. In the critical case where H is singular and $\mu = 0$, the convergence of the doubling algorithms is linear as shown in [15, 25].

4 Exploiting the singularity of H

In this section we assume that M is singular irreducible. Under this assumption, the matrix $H = JM$ has only one independent left and only one independent right eigenvector corresponding to the null eigenvalue. These vectors can be computed easily as already explained in Section 2.5. The knowledge of these eigenvectors can be used for improving the performances of the algorithms by means of two techniques: the shift technique which we deal in Section 4.1, and a suitable choice of the initial approximation in iterative methods which we treat in Section 4.2.

The advantage that one can obtain from these two techniques is twofold: on one hand it is possible to increase the accuracy in the (close to) critical case where the approximation error changes from $O(\sqrt{\varepsilon})$ to $O(\varepsilon)$; on the other hand one can accelerate the convergence speed from the linear convergence to the quadratic convergence in the critical case and improve the quadratic convergence in the close to critical case.

In the rest of the section we consider only the case $\mu \leqslant 0$ in view of Lemma 2.10.

4.1 The shift technique

The shift technique was introduced by He, Meini and Rhee for a quadratic matrix equation arising in the numerical solution of Markov chains modeling quasi-birth-and-death (QBD) processes [27].

For these problems, the interest is in the computation of the minimal nonnegative solution G of the matrix equation

$$X = A_2X^2 + A_1X + A_0,$$

where $A_i \geqslant 0$, $i = 0, 1, 2$, and $(A_2 + A_1 + A_0)e = e$.

A property generally satisfied in the applications is that the polynomial $\det \varphi(z)$, $\varphi(z) = A_2z^2 + (A_1 - I)z + A_0$, has degree at least $n + 1$ and roots $\xi_1, \xi_2, \ldots$, ordered by increasing modulus such that ξ_n and ξ_{n+1} are real and $|\xi_{n-1}| < \xi_n = 1 \leqslant \xi_{n+1}$. Moreover one has $\varphi(1)e = 0$ and $\sigma(G) = \{\xi_1, \ldots, \xi_n\}$ [9].

The conditioning of the minimal nonnegative solution G and the convergence of the available algorithms depend on the ratio $1/\xi_{n+1}$ [39, 12, 27]: the closer is this ratio to 1, the worse conditioned is the solution and the slower is the convergence of the iterative algorithms.

The idea of the shift technique is to consider a new quadratic matrix equation in which the convergence of numerical algorithms and the conditioning of the solution is better, and whose solution easily provides the matrix G. This can be achieved by using the available information of G, that is, $\rho(G) = 1$ and $Ge = e$.

The new UQME is

$$X = B_2X^2 + B_1X + B_0, \tag{23}$$

where

$$\begin{aligned} B_2 &= A_2, \\ B_1 &= A_1 + A_2eu^T, \\ B_0 &= A_0 + (A_1 + A_2 - I)eu^T = A_0 - A_0eu^T, \end{aligned}$$

and u is any positive vector such that $u^Te = 1$. An easy computation shows that the equation (23) has the solution $F = G - eu^T$.

The matrix F has the same eigenvalues as the matrix G except for the eigenvalue 1 of G that becomes the eigenvalue 0 in F, with the same eigenvector e. It can be said that an eigenvalue of G is shifted to 0, this fact gives the name to the technique. Observe that F is the solution with minimal spectral radius of (23).

Concerning the matrix polynomials $\varphi(z)$ and $\psi(z) = B_2z^2 + (B_1 - I)z + B_0$, it holds that

$$\begin{aligned} \varphi(z) &= (A_1 - I + A_2G - zA_2)(zI - G), \\ \psi(z) &= (B_1 - I + B_2F - zB_2)(zI - F) = (A_1 - I + A_2G - zA_2)(zI - F). \end{aligned} \tag{24}$$

The latter equality follows from the fact that $A_2 = B_2$ and $A_1 + A_2G = B_1 + B_2F$ and implies that the determinants of the two matrix polynomials have the same roots except for the root 1 that is replaced by 0. In this way, the ratio between the nth and the $(n + 1)$-st root is reduced from $1/\xi_{n+1}$ to $|\xi_{n-1}|/\xi_{n+1}$ (see [27, 22] for further details).

The important case where $\xi_n = \xi_{n+1} = 1$ is critical for the convergence of algorithms since the ratio $1/\xi_{n=1}$ is 1. In fact in this case the convergence

of algorithms turns from quadratic to linear or from linear to sublinear. The shift technique transforms this critical equation into another one where the ratio between the nth and the $(n+1)$-st root is $|\xi_{n-1}| < 1$. In this way, the quadratic convergence is preserved.

Even in the case where ξ_n is very close to ξ_{n+1} the shift technique allows one to improve the convergence speed since the ratio between the nth and the $(n+1)$-st root becomes $|\xi_{n-1}|/\xi_{n+1}$ which is smaller than $|\xi_n|/\xi_{n+1}$.

The shift technique has a nice functional interpretation: the matrix polynomial $\psi(z)$ of (24) is obtained by the polynomial $\varphi(z)$ by the simple relation [9]

$$\psi(z)(I - z^{-1}Q) = \varphi(z),$$

where $Q = eu^T$. This characterization has the advantage that the shift technique can be extended to matrix equations of any degree or even to matrix power series [9].

The shift technique can be applied to the UQMEs (17), (18), (19), (20) which derive from NAREs. In particular, in the case of equation (17) this technique has been analyzed in detail in [13]. The cases of (18), (19), (20) can be similarly treated.

Similarly to the case of the quadratic matrix equation, one can directly apply the shift technique to the singular matrix H associated with the NARE [25]. Here the goal is to construct a new matrix $\widetilde{H}$ having the same eigenvalues of H except for the eigenvalue 0 which is moved to a positive eigenvalue η of $\widetilde{H}$. In this way we obtain a new NARE associated with $\widetilde{H}$ having better computational feature and with the same solution S of the original NARE.

The construction of $\widetilde{H}$ is based on the following result of which we give a simpler proof. This result was proved by Brauer in 1952 [14] and it has been rediscovered several times (see [31]).

Theorem 4.1. *Let A be an $n \times n$ matrix with eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$ and let v be a nonnull vector such that $Av = \lambda_1 v$. For any nonnull vector x, set $Q = vx^T$. Then the eigenvalues of $A + Q$ are $\lambda_1 + x^T v, \lambda_2, \ldots, \lambda_n$.*

Proof. Since $AQ = \lambda_1 Q$, one has the following identity

$$(\lambda - \lambda_1)(A + Q - \lambda I) = (A - \lambda I)((\lambda - \lambda_1)I - Q).$$

Taking the determinant of both sides and using the formula for the characteristic polynomial of a rank-one matrix, $p_{vx^T}(\lambda) = \det(vx^T - \lambda I) = (\lambda - x^T v)\lambda^{n-1}$, it holds that

$$\begin{aligned} p_{A+Q}(\lambda)(\lambda - \lambda_1)^n &= (-1)^n p_A(\lambda) p_{vx^T}(\lambda - \lambda_1) \\ &= (-1)^n p_A(\lambda)(\lambda - \lambda_1)^{n-1}(\lambda - \lambda_1 - x^T v). \end{aligned}$$

The unique factorization of polynomials completes the proof. □

From the above theorem follows immediately a corollary that will be useful in the following.

Corollary 4.2. *Let A be a singular matrix and $Aw = 0$ for a nonzero vector w. Assume that p is a vector such that $p^T w = 1$ and η is a scalar. Then the eigenvalues of the matrix*

$$\widetilde{A} = A + \eta w p^A$$

are those of A except that one zero eigenvalue of A is replaced by η.

We now construct a rank-one modification of the matrix H:

$$\widehat{H} = H + \eta v p^T, \tag{25}$$

where, v is a positive vector such that $Hv = 0$, $\eta > 0$ is a scalar and $p \geqslant 0$ is a vector with $p^T v = 1$. From Lemma 4.2 the eigenvalues of $\widehat{H}$ are those of H except that one zero eigenvalue of H is replaced by η.

We write $p^T = \left(p_1^T, p_2^T\right)$ and

$$\widetilde{H} = \begin{bmatrix} \widetilde{D} & -\widetilde{C} \\ \widetilde{B} & -\widetilde{A} \end{bmatrix},$$

where

$$\widetilde{D} = D + \eta v_1 p_1^T, \quad \widetilde{C} = C - \eta v_1 p_2^T,$$
$$\widetilde{B} = B + \eta v_2 p_1^T, \quad \widetilde{A} = A - \eta v_2 p_2^T.$$

Corresponding to $\widetilde{H}$ we define the new NARE

$$X\widetilde{C}X - X\widetilde{D} - \widetilde{A}X + \widetilde{B} = 0, \tag{26}$$

which defines the Riccati operator

$$\widetilde{\mathcal{R}}(X) = X\widetilde{C}X - X\widetilde{D} - \widetilde{A}X + \widetilde{B}. \tag{27}$$

We have the following important property about the NARE (26).

Theorem 4.3. *If $\mu \leqslant 0$, then S is a solution of the NARE* (26) *and $\sigma(\widetilde{D} - \widetilde{C}S) = \{\lambda_1, \ldots, \lambda_{n-1}, \eta\}$, where S is the minimal nonnegative solution of the original NARE* (1).

Computing the minimal nonnegative solution S of the NARE (1) can be achieved by computing the solution S of the new NARE (26) corresponding to eigenvalues with positive real parts. Observe that equation (26) is not associated with an M-matrix, however the algorithms and the techniques of Section 3 can be applied and, if break-down is not encountered, convergence is much faster than for the original equation (1). In particular, in the critical case, the convergence of SDA applied to the new NARE (26) is again quadratic. A detailed convergence analysis of SDA is reported in [25].

When $\mu = 0$, the matrix H has two zero eigenvalues. The above shift technique moves one zero eigenvalue to a positive number. We may use a *double-shift* to move the other zero eigenvalue to a negative number. Recall that $Hv = 0$, where $v = \left[\begin{smallmatrix} v_1 \\ v_2 \end{smallmatrix}\right]$, and $w^T H = 0$, where $w = \left[\begin{smallmatrix} u_1 \\ -u_2 \end{smallmatrix}\right]$. We define the matrix

$$\overline{H} = H + \eta v p^T + \xi q w^T, \tag{28}$$

where $\eta > 0$, $\xi < 0$, p and q are such that $p^T v = q^T w = 1$. Since v and w are orthogonal vectors, the double-shift moves one zero eigenvalue to η and the other to ξ. Indeed, the eigenvalues of $\widetilde{H} = H + \xi q w^T$ are those of $\widetilde{H}^T = H^T + \xi w q^T$, which are the eigenvalues of H except that one zero eigenvalue is replaced by ξ, by Lemma 4.2. Also, the eigenvalues of $\overline{H} = \widetilde{H} + \eta v p^T$ are the eigenvalues of $\widetilde{H}$ except that the remaining zero eigenvalue is replaced by η, by Lemma 4.2 again.

From $\overline{H}$ we may define a new Riccati equation

$$X\overline{C}X - X\overline{D} - \overline{A}X + \overline{B} = 0. \tag{29}$$

As before, the minimal nonnegative solution S of (1) is a solution of (29) such that $\sigma(\overline{D} - \overline{C}S) = \{\eta, \lambda_1, \ldots, \lambda_{n-1}\}$. However, it seems very difficult to determine the existence of a solution $\overline{Y}$ of the dual equation of (29) such that $\sigma(\overline{A} - \overline{B}\,\overline{Y}) = \{-\xi, -\lambda_{n+2}, \ldots, -\lambda_{n+m}\}$.

4.2 Choosing a new initial value

If the right eigenvector of H relative to the null eigenvalue is partitioned as $v = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}$, from Theorem 2.14 it follows that for the minimal nonnegative solution S, it holds that $Sv_1 = v_2$ (and then $(D - CS)v_1 = 0$).

In the algorithms in which the initial value can be chosen, like Newton's method, the *usual* choice $X_0 = 0$ does not exploit this information, rather it relies only on the positivity of S. Note that in the Riccati equations modeling fluid queues, the condition $Xv_1 = v_2$ is equivalent to the stochasticity of S since $v_1 = v_2 = e$.

A possibly better convergence is expected if one could generate a sequence such that $X_k v_1 = v_2$ for any $k \geqslant 0$. More precisely, one must choose an iteration which preserves the affine subspace $\widehat{W} = \{A \in \mathbb{C}^{n\times n} : Av_1 = v_2\}$ and an initial value $X_0 \in \widehat{W}$ for which the sequence converges to the desired solution.

A similar idea has been used in [45] in order to improve the convergence speed of certain functional iterations for solving nonlinear matrix equations related to special Markov chains.

A nice property of Newton's method is that it is structure-preserving with respect to the affine subspace $\widehat{W}$. To prove this fact consider the following preliminary result which concerns the Newton iteration

Lemma 4.4. *The Newton method $X_{k+1} = \mathcal{N}(X_k)$,*

$$\mathcal{N}(X_k) = X_k - (d\mathcal{F}_{X_k})^{-1}\mathcal{F}(X_k)$$

applied to the matrix equation $\mathcal{F}(X) = 0$, when defined, preserves the affine structure $\widehat{V}$ if and only if $\mathcal{F}$ is a function from $\widehat{V}$ to its parallel linear subspace V.

Proof. Consider the matrix $X \in \widehat{V}$. The matrix $\mathcal{N}(X)$ belongs to $\widehat{V}$ if and only if $\mathcal{N}(X) - X = (d\mathcal{F}_X)^{-1}(-\mathcal{F}(X))$ belongs to V, and that occurs if and only if $\mathcal{F}(X)$ (and then $-\mathcal{F}(X)$) belongs to V. □

Now, we are ready to prove that the Newton method applied to the Riccati operator is structure-preserving with respect to $\widehat{W}$.

Proposition 4.5. *If X_0 is such that $X_0 v_1 = v_2$, and the Newton method applied to the Riccati equation $\mathcal{R}(X) = 0$ is well defined then $X_k v_1 = v_2$ for any $k \geqslant 0$. That is, the Newton method preserves the structure $\widehat{W}$.*

Proof. In view of Lemma 4.4, one needs to prove that $\mathcal{R}$ is a function from $\widehat{W}$ to the parallel linear subspace W.

If $X \in \widehat{W}$, then $\mathcal{R}(X)v_1 = 0$, in fact

$$\mathcal{R}(X)v_1 = XCXv_1 - AXv_1 - XDv_1 + Bv_1 = XCv_2 - Av_2 - XDv_1 + Bv_1$$

and the last term is 0 since $Cv_2 = Dv_1$ and $Av_2 = Bv_1$. □

A possible choice for the starting value is $(X_0)_{i,j} = (v_2)_i/s$ where $s = \sum_i v_1(i)$. It must be observed that the structured preserving convergence is not anymore monotonic. Since the approximation error has a null component along the subspace W, one should expect a better convergence speed for the sequences obtained with $X_0 \in \widehat{W}$. A proof of this fact and the convergence analysis of this approach is still work in place.

If $\mu = 0$, the differential of $\mathcal{R}$ is singular at the solution S as well as at any point $X \in \widehat{W}$. This makes the sequence X_k undefined. A way to overcome this drawback is considering the shifted Riccati equation described in Section 4.1.

The differential of the shifted Riccati equation (26) at a point X is represented by the matrix

$$\widetilde{\Delta}_X = \Delta_X + I \otimes (\eta(Xv_1 - v_2)p_2^T) + (\eta v_1(p_1^T + p_2^T X))^T \otimes I, \tag{30}$$

where the vector $p \neq 0$ partitioned as $p = \begin{bmatrix} p_1 \\ p_2 \end{bmatrix}$ is an arbitrary nonnegative vector such that $p^T v = 1$. Choosing $p_2 = 0$ provides a nice simplification of the problem, in fact

$$\widetilde{\Delta}_X = \Delta_X - Q^T \otimes I,$$

where $Q = \eta v_1 p_1^T$.

The next result gives more insights on the action of the Newton iteration on the structure $\widehat{V}$.

Proposition 4.6. *Assume that $p_2 = 0$. If $X \in \widehat{W}$ then $\mathcal{R}(X) = \widetilde{\mathcal{R}}(X)$, where $\widetilde{\mathcal{R}}$ is defined in (27). Moreover the sequences generated by Newton's method, when defined, applied to $\mathcal{R}(X) = 0$ and to $\widetilde{\mathcal{R}}(X) = 0$ with $X_0 \in \widehat{W}$ are the same.*

Proof. The fact $\mathcal{R}(X) = \widehat{\mathcal{R}}(X)$, in the assumption $p_2 = 0$, follows from

$$\widetilde{\mathcal{R}}(X) = \mathcal{R}(X) - \eta(Xv_1 - v_2)p_1^T.$$

Let $\mathcal{N}(X) = X - (d\mathcal{R}_X)^{-1}\mathcal{R}(X)$ and $\widetilde{\mathcal{N}}(X) = X - (d\widetilde{\mathcal{R}}_X)^{-1}\widetilde{\mathcal{R}}(X)$ denote the Newton operator for the original equation and for the shifted one, respectively. To prove that the sequences are the same, it must be shown that

$$(A - XC)\mathcal{N}(X) + \mathcal{N}(X)(\widetilde{D} - CX) = \widetilde{B} - XCX$$

holds for any $X \in \widehat{W}$ and for any η (for which the equation has a unique solution). One has

$$\begin{aligned}&(A - XC)\mathcal{N}(X) + \mathcal{N}(X)(\widetilde{D} - CX)\\ &= B - XCX + \mathcal{N}(X)\eta v_1 p_1^T = B - XCX + \eta v_2 p_1^T = \widetilde{B} - XCX,\end{aligned}$$

where we have used that $\mathcal{N}(X)v_1 = v_2$ since $\mathcal{N}(X) \in \widehat{W}$. This completes the proof. □

Since any starting value $X_0 \in \widehat{V}$ gives the same sequence for the Newton method applied either to the Riccati equation (1) or to the shifted Riccati equation (26), then, choosing such an initial value has the same effect of applying the shift technique.

For the applicability one needs that the matrix Δ_{X_k} is nonsingular at each step. Unfortunately the derivative might be singular for some singular M-matrix and some $X \in \widehat{W}_+ = \{X \in \widehat{W}, X \geqslant 0\}$.

If a breakdown occurs, it is always possible to perform the iteration by using the shifted iteration, with $p_2 = 0$ and for a suitable choice of the parameter η. In fact, the iteration is proved in Proposition 4.6 to be the same by any choice of p_1 and η.

The convergence is more subtle. Besides the loss of monotonic convergence, one may note that S is not the only solution belonging to $\widehat{W}$, even if it is the only belonging to $\widehat{W}_+$. In fact, in view of Theorem 2.13, there are at most two positive solutions, and only one of them has the property $Sv_1 = v_2$. The proof of convergence is still work in progress, we conjecture that for each $X_0 \in \widehat{V}_+$, the sequence generated by the Newton method, if defined, converges to S.

A possible improvement of the algorithm could be obtained by implementing the exact line search introduced in [7].

5 Numerical experiments and comparisons

We present some numerical experiments to illustrate the behavior of the algorithms presented in Section 3 and 4.1 in the critical and noncritical case. To compare the accuracy of the methods we have used the relative error $\mathrm{err} = \|X - \widehat{X}\|_1/\|X\|_1$ on the computed solution $\widehat{X}$, when the exact solution X was provided. Elsewhere, we have used the relative residual error

$$\mathrm{res} = \frac{\|\widehat{X}C\widehat{X} - \widehat{X}D - A\widehat{X} + B\|_1}{\|\widehat{X}C\widehat{X}\|_1 + \|\widehat{X}D\|_1 + \|A\widehat{X}\|_1 + \|B\|_1}.$$

The tests were performed using MATLAB 6 Release 12 on a processor AMD Athlon 64. The code for the diverse algorithms is available for download at the web page `http://bezout.dm.unipi.it/mriccati/`.

In these tests we consider three methods: the Newton method (N), the SDA, and the Cyclic Reduction (CR) algorithm applied to the UQME (17) (in both SDA and CR we have considered the matrix $\widehat{H}$ obtained by the Cayley transform of H and not the one relying on the shrink-and-shift operator).

We have also considered the improved version of these methods applied to the singular/critical case; we denoted them as IN, ISDA and ICR, respectively, where "I" stands for "Improved". The initial value for IN is chosen

as suggested in Section 4.1; the parameter for the shift is chosen as $\eta = \max\{\max(A)_{i,i}, \max(D)_{i,i}\}$ and the vector p is chosen to be $e/\sum_i v_i$.

The iterations are stopped when the relative residual/error ceases to decrease or becomes smaller than 10ε, where ε is the machine precision.

Test 5.1. *A null recurrent case* [6, Example 1]. Let

$$M = \left[\begin{array}{cc|cc} 0.003 & -0.001 & -0.001 & -0.001 \\ -0.001 & 0.003 & -0.001 & -0.001 \\ \hline -0.001 & -0.001 & 0.003 & -0.001 \\ -0.001 & -0.001 & -0.001 & 0.003 \end{array}\right]$$

where D is a 2×2 matrix. The minimal positive solution is $X = \frac{1}{2}\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$.

As suggested by the Theorem 2.17, the accuracy of the customary algorithms N, SDA and CR is poor in the critical case, and is near to $\sqrt{\varepsilon} \approx 10^{-8}$. We report in Table 1 the number of steps and the relative error for the three algorithms. If one uses the singularity, due to the particular structure of the problem, the solution is achieved in one step by IN, ISDA and ICR with full accuracy.

Algorithm	Steps	Relative error
N	21	$6.0 \cdot 10^{-7}$
SDA	36	$8.6 \cdot 10^{-7}$
CR	31	$4.7 \cdot 10^{-9}$

Table 1: Accuracy of the algorithms in the critical case, Test 5.1

Test 5.2. *Random choice of a singular M-matrix with* $Me = 0$ [20]. To construct M, we generated a 100×100 random matrix R, and set $M = \mathrm{diag}(Re) - R$. The matrices A, B, C and D are 50×50. We generated 5 different matrices M and computed the relative residuals and number of steps needed for the iterations to converge.

All the algorithms (N, IN, SDA, ISDA, CR and ICR) arrive at a relative residual less than 10ε. The number of steps needed by the algorithms are reported in Table 2. As one can see the basic algorithms require the same number of steps, whilst using the singularity the Newton method requires one or two steps less than ISDA and ICR, however, the cost per step of these two methods make their overall cost much lower than the Newton method.

The use of the singularity reduces dramatically the number of steps needed for the algorithms to converge.

Table 3 summarizes the spectral and computational properties of the solutions of the NARE (1).

Table 4 reports the computational cost of the algorithms for solving (1) with $m = n$, together with the convergence properties in the noncritical case .

References

[1] Soohan Ahn and V. Ramaswami. Transient analysis of fluid flow models via stochastic coupling to a queue. *Stoch. Models*, 20(1):71–101, 2004.

Algorithm	Steps needed
N	11–12
IN	3
SDA	11–12
ISDA	4-5
CR	11–13
ICR	4–5

Table 2: Minimum and maximum number of steps needed for algorithms to converge in Test 5.2

	M nonsingular	M singular irreducible		
		$\mu < 0$	$\mu = 0$	$\mu > 0$
splitting	complete	complete	incomplete	complete
	$\lambda_{n+1} < 0 < \lambda_n$	$\lambda_{n+1} < 0 = \lambda_n$	$\lambda_{n+1} = 0 = \lambda_n$	$\lambda_{n+1} = 0 < \lambda_n$
solutions $\geqslant 0$	2	2	1	2
Δ_S	nonsingular	nonsingular	singular	nonsingular
accuracy	ε	ε	$\sqrt{\varepsilon}$	ε

Table 3: Summary of the properties of the NARE

[2] Brian D. O. Anderson. Second-order convergent algorithms for the steady-state Riccati equation. *Internat. J. Control*, 28(2):295–306, 1978.

[3] Søren Asmussen. Stationary distributions for fluid flow models with or without Brownian noise. *Comm. Statist. Stochastic Models*, 11(1):21–49, 1995.

[4] Zhaojun Bai and James W. Demmel. On swapping diagonal blocks in real Schur form. *Linear Algebra Appl.*, 186:73–95, 1993.

[5] R. H. Bartels and G. W. Stewart. Solution of the matrix equation $AX + XB = C$. *Commun. ACM*, 15(9):820–826, 1972.

[6] Nigel G. Bean, Malgorzata M. O'Reilly, and Peter G. Taylor. Algorithms for return probabilities for stochastic fluid flows. *Stochastic Models*, 21(1):149–184, 2005.

[7] Peter Benner and Ralph Byers. An exact line search method for solving generalized continuous-time algebraic Riccati equations. *IEEE Trans. Automat. Control*, 43(1):101–107, 1998.

[8] Abraham Berman and Robert J. Plemmons. *Nonnegative matrices in the mathematical sciences*, volume 9 of *Classics in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1994. Revised reprint of the 1979 original.

[9] D. A. Bini, G. Latouche, and B. Meini. *Numerical methods for structured Markov chains*. Numerical Mathematics and Scientific Computation. Oxford University Press, New York, 2005. Oxford Science Publications.

Algorithm	Computational cost	Reference
Schur method	$200n^3$	[23, 40]
Functional iteration	$8n^3$—$14n^3$ (per step)	[20, 26]
Newton's method	$66n^3$ (per step)	[26, 24]
CR applied to (17)	$\frac{74}{3}n^3$ (per step)	[13, 10]
CR applied to (18) (SDA)	$\frac{64}{3}n^3$ (per step)	[16, 25, 10]
CR applied to (19), (20)	$\frac{38}{3}n^3$ (per step)	[33, 10]

Table 4: Comparison of the algorithms.

[10] D. A. Bini, B. Meini, and F. Poloni. From algebraic Riccati equations to unilateral quadratic matrix equations: old and new algorithms. Technical Report 1665, Dipartimento di Matematica, Università di Pisa, Italy, July 2007.

[11] Dario Bini, Bruno Iannazzo, and Federico Poloni. A fast Newton's method for a nonsymmetric algebraic Riccati equation. Technical report, Dipartimento di Matematica, Università di Pisa, Italy, January 2007. To appear in SIAM J. Matrix Anal. Appl.

[12] Dario Bini and Beatrice Meini. On the solution of a nonlinear matrix equation arising in queueing problems. *SIAM J. Matrix Anal. Appl.*, 17(4):906–926, 1996.

[13] Dario A. Bini, Bruno Iannazzo, Guy Latouche, and Beatrice Meini. On the solution of algebraic Riccati equations arising in fluid queues. *Linear Algebra Appl.*, 413(2-3):474–494, 2006.

[14] Alfred Brauer. Limits for the characteristic roots of a matrix. IV. Applications to stochastic matrices. *Duke Math. J.*, 19:75–91, 1952.

[15] Chun-Yuei Chiang and Wen-Wei Lin. A structured doubling algorithm for nonsymmetric algebraic Riccati equations (a singular case). Technical report, National Center for Theoretical Sciences, National Tsing Hua University, Taiwan R.O.C., July 2006.

[16] E. K.-W. Chu, H.-Y. Fan, and W.-W. Lin. A structure-preserving doubling algorithm for continuous-time algebraic Riccati equations. *Linear Algebra Appl.*, 396:55–80, 2005.

[17] Ana da Silva Soares and Guy Latouche. Further results on the similarity between fluid queues and QBDs. In *Matrix-analytic methods (Adelaide, 2002)*, pages 89–106. World Sci. Publ., River Edge, NJ, 2002.

[18] Sandra Fital and Chun-Hua Guo. Convergence of the solution of a nonsymmetric matrix Riccati differential equation to its stable equilibrium solution. *J. Math. Anal. Appl.*, 318(2):648–657, 2006.

[19] Gene H. Golub and Charles F. Van Loan. *Matrix computations*. Johns Hopkins Studies in the Mathematical Sciences. Johns Hopkins University Press, Baltimore, MD, third edition, 1996.

[20] Chun-Hua Guo. Nonsymmetric algebraic Riccati equations and Wiener-Hopf factorization for M-matrices. *SIAM J. Matrix Anal. Appl.*, 23(1):225–242, 2001.

[21] Chun-Hua Guo. A note on the minimal nonnegative solution of a nonsymmetric algebraic Riccati equation. *Linear Algebra Appl.*, 357:299–302, 2002.

[22] Chun-Hua Guo. Comments on a shifted cyclic reduction algorithm for quasi-birth-death problems. *SIAM J. Matrix Anal. Appl.*, 24(4):1161–1166, 2003.

[23] Chun-Hua Guo. Efficient methods for solving a nonsymmetric algebraic Riccati equation arising in stochastic fluid models. *J. Comput. Appl. Math.*, 192(2):353–373, 2006.

[24] Chun-Hua Guo and Nicholas J. Higham. Iterative Solution of a Nonsymmetric Algebraic Riccati Equation. *SIAM Journal on Matrix Analysis and Applications*, 29(2):396–412, 2007.

[25] Chun-Hua Guo, Bruno Iannazzo, and Beatrice Meini. On the Doubling Algorithm for a (Shifted) Nonsymmetric Algebraic Riccati Equation. *SIAM J. Matrix Anal. Appl.*, 29(4):1083–1100, 2007.

[26] Chun-Hua Guo and Alan J. Laub. On the iterative solution of a class of nonsymmetric algebraic Riccati equations. *SIAM J. Matrix Anal. Appl.*, 22(2):376–391, 2000.

[27] C. He, B. Meini, and N. H. Rhee. A shifted cyclic reduction algorithm for quasi-birth-death problems. *SIAM J. Matrix Anal. Appl.*, 23(3):673–691, 2001/02.

[28] Nicholas J. Higham. The Matrix Function Toolbox. `http://www.ma.man.ac.uk/~higham/mftoolbox`.

[29] Nicholas J. Higham. *Functions of Matrices: Theory and Computation.* In preparation, 2008.

[30] Leslie Hogben, editor. *Handbook of linear algebra*. Discrete Mathematics and its Applications (Boca Raton). Chapman & Hall/CRC, Boca Raton, FL, 2007. Associate editors: Richard Brualdi, Anne Greenbaum and Roy Mathias.

[31] R. A. Horn and S. Serra Capizzano. Canonical and standard forms for certain rank one perturbations and an application to the (complex) Google pageranking problem. To appear in Internet Mathematics, 2007.

[32] T.-M. Hwang, E. K.-W. Chu, and W.-W. Lin. A generalized structure-preserving doubling algorithm for generalized discrete-time algebraic Riccati equations. *Internat. J. Control*, 78(14):1063–1075, 2005.

[33] Bruno Iannazzo and Dario Bini. A cyclic reduction method for solving algebraic Riccati equations. Technical report, Dipartimento di Matematica, Università di Pisa, Italy, 2005.

[34] Jonq Juang. Global existence and stability of solutions of matrix Riccati equations. *J. Math. Anal. Appl.*, 258(1):1–12, 2001.

[35] Jonq Juang and Wen-Wei Lin. Nonsymmetric algebraic Riccati equations and Hamiltonian-like matrices. *SIAM J. Matrix Anal. Appl.*, 20(1):228–243, 1999.

[36] L. V. Kantorovich. *Functional analysis and applied mathematics*. NBS Rep. 1509. U. S. Department of Commerce National Bureau of Standards, Los Angeles, Calif., 1952. Translated by C. D. Benster.

[37] D. Kleinman. On an iterative technique for riccati equation computations. *IEEE Trans. Automat. Control*, 13(1):114–115, 1968.

[38] Peter Lancaster and Leiba Rodman. *Algebraic Riccati equations*. Oxford Science Publications. The Clarendon Press Oxford University Press, New York, 1995.

[39] Guy Latouche and V. Ramaswami. A logarithmic reduction algorithm for quasi-birth-death processes. *J. Appl. Probab.*, 30(3):650–674, 1993.

[40] Alan J. Laub. A Schur method for solving algebraic Riccati equations. *IEEE Trans. Automat. Control*, 24(6):913–921, 1979.

[41] Wen-Wei Lin and Shu-Fang Xu. Convergence analysis of structure-preserving doubling algorithms for Riccati-type matrix equations. *SIAM J. Matrix Anal. Appl.*, 28(1):26–39, 2006.

[42] Lin-Zhang Lu. Newton iterations for a non-symmetric algebraic Riccati equation. *Numer. Linear Algebra Appl.*, 12(2-3):191–200, 2005.

[43] Lin-Zhang Lu. Solution form and simple iteration of a nonsymmetric algebraic Riccati equation arising in transport theory. *SIAM J. Matrix Anal. Appl.*, 26(3):679–685, 2005.

[44] V. L. Mehrmann. *The autonomous linear quadratic control problem*, volume 163 of *Lecture Notes in Control and Information Sciences*. Springer-Verlag, Berlin, 1991. Theory and numerical solution.

[45] Beatrice Meini. New convergence results on functional iteration techniques for the numerical solution of $M/G/1$ type Markov chains. *Numer. Math.*, 78(1):39–58, 1997.

[46] V. Ramaswami. Matrix analytic methods for stochastic fluid flows. In D. Smith and P. Hey, editors, *Teletraffic Engineering in a Competitive World*, Proceedings of the 16th International Teletraffic Congress, Elsevier Science B.V., Edimburgh, UK, pages 1019–1030, 1999.

[47] L. C. G. Rogers. Fluid models in queueing theory and Wiener-Hopf factorization of Markov chains. *Ann. Appl. Probab.*, 4(2):390–413, 1994.

[48] David Williams. A "potential-theoretic" note on the quadratic Wiener-Hopf equation for Q-matrices. In *Seminar on Probability, XVI*, volume 920 of *Lecture Notes in Math.*, pages 91–94. Springer, Berlin, 1982.