From algebraic Riccati equations to unilateral quadratic matrix equations: old and new algorithms

Dario A. Bini* Beatrice Meini* Federico Poloni†

Abstract

The problem of reducing an algebraic Riccati equation $XCX - AX - XD + B = 0$ to a unilateral quadratic matrix equation (UQME) of the kind $PX^2 + QX + R = 0$ is analyzed. New reductions are introduced which enable one to prove some theoretical and computational properties. In particular we show that the structure preserving doubling algorithm of B.D.O. Anderson [Internat. J. Control, 1978] is in fact the cyclic reduction algorithm of Hockney [J. Assoc. Comput. Mach., 1965] and Buzbee, Golub, Nielson [SIAM J. Numer. Anal., 1970], applied to a suitable UQME. A new algorithm obtained by complementing our reductions with the shrink-and-shift technique of Ramaswami is presented. Finally, faster algorithms which require some non-singularity conditions, are designed. The non-singularity restriction is relaxed by introducing a suitable similarity transformation of the Hamiltonian.

1 Introduction

Given $A \in \mathbb{R}^{m\times m}$, $B \in \mathbb{R}^{m\times n}$, $C \in \mathbb{R}^{n\times m}$ and $D \in \mathbb{R}^{n\times n}$, consider the nonsymmetric Algebraic Riccati Equation (NARE)

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

where $X \in \mathbb{R}^{m\times n}$ is the unknown. If $m = n$, $C = C^T$, $B = B^T$ and $D = A^T$, equation (1) reduces to a Continuous Algebraic Riccati Equation (CARE), which has been extensively studied by several authors, see the books [23], [26].

Most recently, some attention has been devoted to the nonsymmetric case, under the hypothesis that

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

is either a nonsingular M-matrix or a singular irreducible M-matrix, see the papers [7, 6, 14, 15, 16].

*Dipartimento di Matematica, Università di Pisa, Italy
†Scuola Normale Superiore, Pisa, Italy

Dagstuhl Seminar Proceedings 07461
Numerical Methods for Structured Markov Chains
http://drops.dagstuhl.de/opus/volltexte/2008/1398

The solutions of (1) are related to the spectral properties of the matrix

$$\mathcal{H} = (h_{i,j}) = \begin{bmatrix} D & -C \\ B & -A \end{bmatrix} \tag{3}$$

which we call the *Hamiltonian* of the Riccati equation, in analogy with the symmetric (CARE) case. We say that (1) is the NARE associated with $\mathcal{H}$.

Both in the nonsymmetric and in the symmetric case, the interest is the computation of the solution S such that the eigenvalues of $D - CS$ lie in the right complex half-plane.

The problem of reducing an algebraic Riccati equation $XCX - AX - XD + B = 0$ to a unilateral quadratic matrix equation (UQME) of the kind

$$PX^2 + QX + R = 0 \tag{4}$$

has been considered in [6, 18, 21, 27]. In [18] the authors express the solutions of (1) in terms of the solution of the equation $Y^2 - M^2 = 0$, where Y, M are matrices of size $m+n$. This approach encounters problems of numerical stability. Also in [6] and [27] the authors reduce an ARE to a UQME of the kind (4) where the size of the matrix coefficients is $m+n$. The approach followed in [21] allows one to keep the blocks of (4) of size m, but it has the strong limitation to work only if $m = n$ and $\det(C) \neq 0$. Moreover, the cases where C is ill-conditioned generate numerical instability. This drawback is removed in [21] by doubling the size of the blocks with a consequent increase of the computational cost.

Here we introduce three classes of reductions of AREs to UQMEs. The first two reductions transform the ARE into a UQME where the matrix coefficients have size $m+n$, but have a strong structure. The third reduction can be applied if $m = n$ and provides a UQME with block coefficients of size m. The matrix C is required to be nonsingular; possible singularity of C can be removed by performing a preliminary similarity transformation of the Hamiltonian.

These new reductions enable us to prove some theoretical and computational properties. In particular they provide a unifying framework which includes apparently different algorithms like the algorithm of Ramaswami [27], the structure preserving doubling algorithm (SDA) of Anderson [1], and Guo, Lin, Xu [18], and the algorithm of Bini and Iannazzo [21]. We prove that SDA is in fact the cyclic reduction algorithm applied to a suitable UQME. Cyclic Reduction (CR) was originally introduced by Hockney, Buzbee, Golub, Nielson [9, 19] for the Poisson equation over the rectangle, and later adapted to solving matrix equations by Bini and Meini [3, 8]. This fact enables one to deduce the convergence properties of SDA directly from the theory of cyclic reduction which is well consolidated [8, 4]. The relationships between SDA and CR have been recently investigated by C.-H. Guo and W.-W. Lin in [17] in the case where SDA is applied to a UQME.

This unifying framework allows us to design some new algorithms whose performance is under investigation. In particular, by complementing the second reduction with the shrink-and-shift technique of Ramaswami we obtain a new

algorithm having the same cost per iteration of SDA but relying on a more simple initialization.

Relying on the third reduction, we arrive at the algorithm of [21] having a lower cost with respect to SDA. Here we introduce a transformation which makes this algorithm numerically stable for a wide class of cases.

The reductions are based on two fundamental ideas, namely, rewriting the matrix pencil $\mathcal{H} - zI$ as a quadratic matrix polynomial of the kind

$$\mathcal{P}(z) = z^2\mathcal{A}_2 + z\mathcal{A}_1 + \mathcal{A}_0$$

where the block coefficients $\mathcal{A}_0, \mathcal{A}_1, \mathcal{A}_2$ of size at most $m+n$ are suitably structured, and transforming the Hamiltonian (3) into a new one whose eigenvalues have a splitting w.r.t. the unit circle in the complex plane.

In this way we arrive at a UQME of the kind

$$\mathcal{A}_2X^2 + \mathcal{A}_1X + \mathcal{A}_0 = 0.$$

whose solutions are simply related to the solutions of (1) and where the roots of $\det \mathcal{P}(z)$ are split w.r.t. the unit circle. Under these conditions the algorithm of cyclic reduction has the best numerical performance in terms of numerical stability and convergence. Moreover, the structure of the blocks $\mathcal{A}_i$ enables us to implement CR with the lowest computational cost.

The paper is organized as follows. In Section 2 we recall the assumptions and the main properties concerning NAREs. In Section 3 we review the main known algorithms, among which SDA and cyclic reduction. Section 4 is devoted to the reductions of a NARE to a UQME, Section 5 to the eigenvalue transformations. Finally Section 6 analyzes the main reductions from the algorithmic point of view. Numerical experiments are reported in Section 7 while conclusions and open issues are summarized in Section 8.

2 Assumptions on Algebraic Riccati Equations

The solutions of the NARE (1) are related to the invariant subspaces of the matrix $\mathcal{H}$ in (3). More specifically, X is a solution of (1) if and only if

$$\begin{bmatrix} D & -C \\ B & -A \end{bmatrix} \begin{bmatrix} I_n \\ X \end{bmatrix} = \begin{bmatrix} I_n \\ X \end{bmatrix} (D - CX), \tag{5}$$

so that the span of $\begin{bmatrix} I_n \\ X \end{bmatrix}$ is an invariant subspace for the matrix $\mathcal{H}$. Moreover, the eigenvalues of $D - CX$ are a subset of the eigenvalues of $\mathcal{H}$. In the sequel we will denote by $\{\lambda_1, \lambda_2, \ldots, \lambda_{m+n}\}$ the eigenvalues of the matrix $\mathcal{H}$ of (3), and assume that they are ordered according to their real part, i.e.,

$$\operatorname{Re}\lambda_{m+n} \leqslant \operatorname{Re}\lambda_{m+n-1} \leqslant \cdots \leqslant \operatorname{Re}\lambda_2 \leqslant \operatorname{Re}\lambda_1. \tag{6}$$

Throughout, unless differently specified, we will make the following assumption on the matrix $\mathcal{H}$:

Assumption 1 (General case)**.** The eigenvalues of $\mathcal{H}$ are such that

$$\operatorname{Re}\lambda_{n+1} \leqslant 0 \leqslant \operatorname{Re}\lambda_n. \tag{7}$$

Moreover, the invariant subspace of $\mathcal{H}$ corresponding to the eigenvalues $\lambda_1, \ldots, \lambda_n$, is spanned by the matrix $\begin{bmatrix} I_n \\ S \end{bmatrix}$ for a suitable $m \times n$ matrix S.

Observe that under the above assumption, if $\lambda_n \neq \lambda_{n+1}$ the matrix S is the only solution of the Riccati equation (1) such that the eigenvalues of $D - CS$ are $\lambda_1, \ldots, \lambda_n$. The matrix S is the solution of interest in many applications and it is called *extremal solution*. The term extremal refers to the fact that the eigenvalues of $D - CS$ are the rightmost eigenvalues of $\mathcal{H}$.

Assumption 1 is satisfied in particular under the following stronger conditions which are encountered in diverse applications.

Assumption 2 (Symmetric case [23, 26])**.** The matrix $\mathcal{H}$ is such that $D = A^T$, $C = C^T$, $B = B^T$, $C = VV^T$ for $V \in \mathbb{R}^{n\times k}$ full rank matrix, and the pair (A, V) is c-stabilizable.

Assumption 3 (M-matrix case [2, 11, 27])**.** The matrix $\mathcal{H}$ is such that

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

is either a nonsingular M-matrix or a singular irreducible M-matrix.

Assumption 4 (Complex case [13])**.** The matrix $\mathcal{H}$ is complex and such that $\widetilde{\mathcal{H}} = (\widetilde{h}_{i,j})$, $\widetilde{h}_{i,i} = |h_{i,i}|$, $\widetilde{h}_{i,j} = -|h_{i,j}|$, if $i \neq j$, is an M-matrix, moreover the diagonal elements of $\mathcal{H}$ are either positive or negative.

In the following we restrict our analysis to the case where the condition $\operatorname{Re}\lambda_n = \operatorname{Re}\lambda_{n+1} = 0$ is not satisfied. We recall that if $\operatorname{Re}\lambda_n = \operatorname{Re}\lambda_{n+1} = 0$, then one can apply suitable techniques in order to overcome the difficulties encountered in this critical case. For more details we refer the reader to [7]. Without loss of generality [7, 15], in the sequel we assume the following

Assumption 5. The eigenvalues of $\mathcal{H}$, ordered as in (6), are such that $\operatorname{Re}\lambda_{n+1} < 0 \leq \operatorname{Re}\lambda_n$.

In fact, the case $\operatorname{Re}\lambda_{n+1} \leqslant 0 < \operatorname{Re}\lambda_n$ can be reduced to the case in Assumption 5, see [15, 7].

Here and hereafter we use the following notation for the Riccati operator and its dual

$$\begin{aligned} \mathscr{R}(X) &= XCX - AX - XD + B \\ \mathscr{D}(Y) &= YBY - DY - YA + C \end{aligned} \tag{8}$$

where X and Y^T are $m \times n$ matrices.

3 Review of known algorithms

In this section we will review the main algorithms for computing the unique extremal solution S corresponding to the rightmost invariant subspace.

3.1 Outline of SDA

One of the most efficient algorithms for calculating the extremal solution S of a NARE is the Structure-preserving Doubling Algorithm (SDA) [1, 18]. A preliminary step of the algorithm consists in applying to $\mathcal{H}$ the Cayley transform

$$\mathcal{C}_\gamma : z \to \frac{z-\gamma}{z+\gamma}, \tag{9}$$

for a positive constant $\gamma > 0$, so that (5) with $X = S$ is transformed into

$$(\mathcal{H} - \gamma I) \begin{bmatrix} I \\ S \end{bmatrix} = (\mathcal{H} + \gamma I) \begin{bmatrix} I \\ S \end{bmatrix} R_\gamma, \tag{10}$$

where

$$R_\gamma = \mathcal{C}_\gamma(R) = (R + \gamma I)^{-1}(R - \gamma I), \quad R = D - CS.$$

By premultiplying both sides of (10) with a suitable nonsingular matrix, one gets

$$\mathcal{L}^{(\gamma)} \begin{bmatrix} I \\ S \end{bmatrix} = \mathcal{U}^{(\gamma)} \begin{bmatrix} I \\ S \end{bmatrix} R_\gamma, \tag{11}$$

where

$$\mathcal{L}^{(\gamma)} = \begin{bmatrix} E^{(\gamma)} & 0 \\ -H^{(\gamma)} & I \end{bmatrix}, \quad \mathcal{U}^{(\gamma)} = \begin{bmatrix} I & -G^{(\gamma)} \\ 0 & F^{(\gamma)} \end{bmatrix},$$

and

$$\begin{aligned} &E^{(\gamma)} = I - 2\gamma(V^{(\gamma)})^{-1}, \quad F^{(\gamma)} = I - 2\gamma(W^{(\gamma)})^{-1}, \\ &G^{(\gamma)} = 2\gamma(D^{(\gamma)})^{-1}C(W^{(\gamma)})^{-1}, \quad H^{(\gamma)} = 2\gamma(W^{(\gamma)})^{-1}B(D^{(\gamma)})^{-1}, \\ &W^{(\gamma)} = A^{(\gamma)} - B(D^{(\gamma)})^{-1}C, \quad V^{(\gamma)} = D^{(\gamma)} - C(A^{(\gamma)})^{-1}B, \\ &A^{(\gamma)} = A + \gamma I, \quad D^{(\gamma)} = D + \gamma I. \end{aligned} \tag{12}$$

SDA consists in the following iteration

$$\begin{aligned} E_{k+1} &= E_k(I - G_k H_k)^{-1} E_k, \\ F_{k+1} &= F_k(I - H_k G_k)^{-1} F_k, \\ G_{k+1} &= G_k + E_k(I - G_k H_k)^{-1} G_k F_k, \\ H_{k+1} &= H_k + F_k(I - H_k G_k)^{-1} H_k E_k, \end{aligned} \tag{13}$$

for $k \geq 0$, starting from the initial values $E_0 = E^{(\gamma)}$, $F_0 = F^{(\gamma)}$, $G_0 = G^{(\gamma)}$, $H_0 = H^{(\gamma)}$.

For the matrix sequences (13) generated by SDA it holds

$$\mathcal{L}_k \begin{bmatrix} I \\ S \end{bmatrix} = \mathcal{U}_k \begin{bmatrix} I \\ S \end{bmatrix} R_\gamma^{2^k}, \tag{14}$$

where

$$\mathcal{L}_k := \begin{bmatrix} E_k & 0 \\ -H_k & I \end{bmatrix}, \quad \mathcal{U}_k = \begin{bmatrix} I & -G_k \\ 0 & F_k \end{bmatrix}.$$

The parameter γ is chosen in order to guarantee applicability and convergence of SDA. For instance, under Assumptions 3 and 5, if γ is such that

$$\gamma \geqslant \max\left\{ \max_{1\leqslant i\leqslant m} A_{ii}, \max_{1\leqslant j\leqslant n} D_{jj} \right\},$$

then the following convergence result holds [15]: $E_k \to 0$, $H_k \to S$, $G_k \to T$, where T is the extremal solution of the dual equation

$$YBY - YA - DY + C = 0. \tag{15}$$

Moreover, the convergence is quadratic, that is, $||H_k - S|| = O(\rho(R_\gamma)^{2^k})$ for any matrix norm, where $\rho(R_\gamma)$ is the spectral radius of R_γ.

Similar results hold under Assumptions 2 and 4 (see [13, 25]).

It is interesting to observe that SDA is closely related to a block UL factorization. In fact, with a small modification to the approach presented in [18], the following result can easily be proved by mathematical induction.

Theorem 1. *Let $H_\gamma = \mathcal{C}_\gamma(H)$; at each step k of the SDA algorithm it holds*

$$\mathcal{U}_k \mathcal{H}_\gamma^{2^k} = \mathcal{L}_k.$$

Therefore, SDA can be interpreted as a means to calculate a special block UL factorization of $\mathcal{H}_\gamma^{2^k}$ without dealing explicitly with large numbers and ill-conditioned matrices.

The operations required by SDA are 10 matrix products and two LU factorizations per step, for the overall cost of $\frac{64}{3}n^3$ ops per step (assuming $m = n$).

3.2 Outline of Cyclic Reduction

Cyclic reduction [9, 8, 3] is a classical algorithm to compute the solution of the unilateral quadratic matrix equation

$$\mathcal{A}_0 + \mathcal{A}_1 X + \mathcal{A}_2 X^2 = 0, \tag{16}$$

with $\mathcal{A}_0, \mathcal{A}_1, \mathcal{A}_2 \in \mathbb{R}^{N\times N}$. It is defined by the iteration

$$\begin{aligned}
\mathcal{A}_1^{(k+1)} &= \mathcal{A}_1^{(k)} - \mathcal{A}_0^{(k)}\mathcal{K}^{(k)}\mathcal{A}_2^{(k)} - \mathcal{A}_2^{(k)}\mathcal{K}^{(k)}\mathcal{A}_0^{(k)}, \quad \mathcal{K}^{(k)} = \left(\mathcal{A}_1^{(k)}\right)^{-1}, \\
\mathcal{A}_0^{(k+1)} &= -\mathcal{A}_0^{(k)}\mathcal{K}^{(k)}\mathcal{A}_0^{(k)}, \\
\mathcal{A}_2^{(k+1)} &= -\mathcal{A}_2^{(k)}\mathcal{K}^{(k)}\mathcal{A}_2^{(k)}, \\
\mathcal{B}^{(k+1)} &= \mathcal{B}^{(k)} - \mathcal{A}_2^{(k)}\mathcal{K}^{(k)}\mathcal{A}_0^{(k)},
\end{aligned} \tag{17}$$

for $k \geq 0$, starting from $\mathcal{A}_i^{(0)} = \mathcal{A}_i$, $i = 0, 1, 2$, and $\mathcal{B}^{(0)} = \mathcal{A}_1$. Here and hereafter, we use the expression "roots of the matrix polynomial $\mathcal{A}(z) = \mathcal{A}_0 + z\mathcal{A}_1 + z^2\mathcal{A}_2$" to denote the roots of the polynomial $\det \mathcal{A}(z)$. The following convergence result holds [5, 8].

Theorem 2. *Let $\xi_1, \ldots, \xi_{2N}$ be the roots of $\mathcal{A}(z) = \mathcal{A}_0 + z\mathcal{A}_1 + z^2\mathcal{A}_2$, including roots at infinity if $\deg a(z) < 2N$, ordered with nondecreasing modulus. Suppose that $|\xi_N| \leqslant 1 \leqslant |\xi_{N+1}|$ and $|\xi_N| < |\xi_{N+1}|$, and that a solution $\mathcal{S}$ exists to (16) such that $\rho(\mathcal{S}) = |\xi_N|$. Then, $\mathcal{S}$ is the solution with minimal spectral radius, moreover, if CR (17) can be carried out with no breakdown, the sequence*

$$\mathcal{S}^{(k)} = -\left(\mathcal{B}^{(k)}\right)^{-1} \mathcal{A}_0$$

is such that for any norm

$$||\mathcal{S}^{(k)} - \mathcal{S}|| \leqslant \theta |\xi_N/\xi_{N+1}|^{2^k},$$

where $\theta > 0$ is a suitable constant. Moreover, the function $\varphi(z) = z^{-1}\mathcal{A}(z)$ is analytic and invertible for $|\xi_N| < |z| < |\xi_{N+1}|$. If in addition there exists a solution to the equation $\mathcal{A}_0 X^2 + \mathcal{A}_1 X + \mathcal{A}_2 = 0$ with spectral radius $|\xi_N|$, then the constant coefficient ψ_0 of $\psi(z) = \sum_{i=-\infty}^{+\infty} z^i \psi_i = \varphi(z)^{-1}$ is nonsingular and

$$\lim_k \mathcal{A}_1^{(k)} = \psi_0^{-1}, \quad ||A_1^{(k)} - \psi_0^{-1}|| \leqslant \theta |\xi_N/\xi_{N+1}|^{2^k}.$$

The computational cost of the CR iteration amounts to 6 matrix products and one LU factorization per step, that is $\frac{38}{3}N^3$ ops per step. Several results [3, 5, 8] provide necessary and sufficient conditions for the applicability of CR.

Another useful formulation of the cyclic reduction is the functional formulation [8]. Let

$$\varphi_k(z) = z^{-1}\mathcal{A}_0^{(k)} + \mathcal{A}_1^{(k)} + z\mathcal{A}_2^{(k)},$$

and $\psi_k(z) = \varphi_k(z)^{-1}$, defined for all z for which $\varphi_k(z)$ is nonsingular. Then the CR iteration can be seen as

$$\psi_{k+1}(z^2) = \frac{1}{2}(\psi_k(z) + \psi_k(-z)), \quad k = 0, 1, \ldots. \tag{18}$$

4 Reductions to UQME

In this section we present some transformations of a NARE to an equivalent UQME which enable one to apply the known algorithms for UQME's in order to solve a NARE. These transformations are based on the idea of modifying the linear matrix pencil

$$\mathcal{H} - zI$$

into a new quadratic matrix pencil having the same eigenvalues of $\mathcal{H}$ plus some additional eigenvalues at zero and/or at infinity.

4.1 Ramaswami's reduction

Ramaswami [27] proposed a reduction method to transform the NARE into a UQME of the kind

$$\mathcal{A}_0 + \mathcal{A}_1 Y + \mathcal{A}_2 Y^2 = 0, \quad \mathcal{A}_0, \mathcal{A}_1, \mathcal{A}_2 \in \mathbb{R}^{(m+n)\times(m+n)}.$$

His reduction can be interpreted in the following way. The eigenvalue problem

$$0 = (\mathcal{H} - zI)\, u = \left(\begin{bmatrix} D & -C \\ B & -A \end{bmatrix} - z \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix} \right) u$$

originating from (5) is transformed into a quadratic eigenvalue problem by multiplying the second block column by z:

$$\left(\begin{bmatrix} D & 0 \\ B & 0 \end{bmatrix} + \begin{bmatrix} -I & -C \\ 0 & -A \end{bmatrix} z + \begin{bmatrix} 0 & 0 \\ 0 & -I \end{bmatrix} z^2 \right) u = 0.$$

With the latter quadratic pencil we may associate the UQME

$$\begin{bmatrix} D & 0 \\ B & 0 \end{bmatrix} + \begin{bmatrix} -I & -C \\ 0 & -A \end{bmatrix} Y + \begin{bmatrix} 0 & 0 \\ 0 & -I \end{bmatrix} Y^2 = 0, \tag{19}$$

defined by the matrix quadratic polynomial

$$\begin{aligned} &\mathcal{A}(z) = \mathcal{A}_0 + z\mathcal{A}_1 + z^2\mathcal{A}_2, \\ &\mathcal{A}_0 = \begin{bmatrix} D & 0 \\ B & 0 \end{bmatrix}, \quad \mathcal{A}_1 = \begin{bmatrix} -I & -C \\ 0 & -A \end{bmatrix}, \quad \mathcal{A}_2 = \begin{bmatrix} 0 & 0 \\ 0 & -I \end{bmatrix}. \end{aligned} \tag{20}$$

The roots of $\mathcal{A}(z)$ and the eigenvalues of $\mathcal{H}$, as well as the solutions of the NARE (1) and the UQME (19), are closely related as stated by the following

Theorem 3. *The roots of the matrix polynomial $\mathcal{A}(z)$ defined in* (20) *are:*

- *m equal to 0,*
- *the $m+n$ eigenvalues $\lambda_1, \ldots, \lambda_{m+n}$ of $\mathcal{H}$,*
- *n at infinity.*

Moreover, if X is a solution of the NARE (1)*, then*

$$Y = \begin{bmatrix} D - CX & 0 \\ X & 0 \end{bmatrix}$$

is a solution to the UQME (19)*; conversely,*

$$\mathcal{S} = \begin{bmatrix} D - CS & 0 \\ S & 0 \end{bmatrix} \tag{21}$$

where S is the extremal solution of (1)*, is the unique solution of the UQME* (19) *with m eigenvalues equal to zero and n eigenvalues equal to $\lambda_1, \ldots, \lambda_n$.*

Proof. By construction, one has

$$\mathcal{A}(z) = (\mathcal{H} - zI_{m+n}) \begin{bmatrix} I_n & 0 \\ 0 & zI_m \end{bmatrix}.$$

Therefore $\det \mathcal{A}(z) = z^m \det(\mathcal{H} - zI_{m+n})$ and the properties on the roots immediately follow. If X solves the NARE (1), one may verify by direct inspection that Y solves the UQME (19). In particular $\mathcal{S}$ is a solution of (19), and its eigenvalues are m equal to zero, and $\lambda_1, \ldots, \lambda_n$; for the properties of the roots of $\mathcal{A}(z)$, $\mathcal{S}$ is the unique solution with these eigenvalues. □

From the above result, it follows that if one is interested in computing the extremal solution S of the NARE, then the solution of the UQME having eigenvalues with nonnegative real part must be computed. In [27] this UQME is solved by means of Logarithmic Reduction [24] of Latouche and Ramaswami.

4.2 UL-based reduction

We introduce another reduction of a NARE to a UQME which relies on the block UL factorization of the matrix $\mathcal{H}$. As we will point out in Section 6.2, this reduction is implicitly used in the SDA, and allows to relate SDA to CR.

Let us consider the block UL factorization

$$\mathcal{H} = \mathcal{U}^{-1}\mathcal{L}, \quad \mathcal{U} = \begin{bmatrix} I & -U_1 \\ 0 & U_2 \end{bmatrix}, \quad \mathcal{L} = \begin{bmatrix} L_1 & 0 \\ -L_2 & I \end{bmatrix},$$

with

$$\begin{aligned} U_1 &= -CA^{-1}, & U_2 &= -A^{-1}, \\ L_1 &= D - CA^{-1}B, & L_2 &= -A^{-1}B, \end{aligned}$$

where we assume $\det A \neq 0$, and transform the eigenvalue problem

$$0 = (\mathcal{H} - zI)u = (\mathcal{U}^{-1}\mathcal{L} - zI)u$$

into the generalized one

$$(\mathcal{L} - z\mathcal{U})u = 0.$$

Now we multiply the second block row by $-z$ to get

$$\left(\begin{bmatrix} L_1 & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} -I & U_1 \\ L_2 & -I \end{bmatrix} z + \begin{bmatrix} 0 & 0 \\ 0 & U_2 \end{bmatrix} z^2 \right) u = 0.$$

As in the previous section, with the latter quadratic pencil we may associate the UQME

$$\begin{bmatrix} L_1 & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} -I & U_1 \\ L_2 & -I \end{bmatrix} Y + \begin{bmatrix} 0 & 0 \\ 0 & U_2 \end{bmatrix} Y^2 = 0, \tag{22}$$

defined by the matrix quadratic polynomial

$$\begin{aligned} &\mathcal{A}(z) = \mathcal{A}_0 + z\mathcal{A}_1 + z^2\mathcal{A}_2, \\ &\mathcal{A}_0 = \begin{bmatrix} L_1 & 0 \\ 0 & 0 \end{bmatrix}, \quad \mathcal{A}_1 = \begin{bmatrix} -I & U_1 \\ L_2 & -I \end{bmatrix}, \quad \mathcal{A}_2 = \begin{bmatrix} 0 & 0 \\ 0 & U_2 \end{bmatrix}. \end{aligned} \tag{23}$$

The following result holds.

Theorem 4. *The roots of the matrix polynomial $\mathcal{A}(z)$ defined in* (23) *are:*

- *m equal to 0,*
- *the $m+n$ eigenvalues $\lambda_1, \ldots, \lambda_{m+n}$ of $\mathcal{H}$,*
- *n at infinity.*

Moreover, if X is a solution of the NARE (1)*, then*

$$Y = \begin{bmatrix} D - CX & 0 \\ X(D - CX) & 0 \end{bmatrix}$$

is a solution to the UQME (22)*. Conversely,*

$$\mathcal{S} = \begin{bmatrix} D - CS & 0 \\ S(D - CS) & 0 \end{bmatrix},$$

where S is the extremal solution of (1)*, is the unique solution of the UQME* (19) *with m eigenvalues equal to zero and n eigenvalues equal to $\lambda_1, \ldots, \lambda_n$.*

Proof. By construction, one has

$$\mathcal{A}(z) = \begin{bmatrix} I_n & 0 \\ 0 & -zI_m \end{bmatrix} \mathcal{U}(\mathcal{H} - zI_{m+n}).$$

Therefore $\det \mathcal{A}(z) = (-1)^m z^m \det \mathcal{U} \det(\mathcal{H} - zI_{m+n})$ and the properties on the roots immediately follow. The remaining part can be proved as in Theorem 3. □

4.3 Reduction to a UQME of lower size

In this section we assume that $m = n$ and that in the Riccati equation (1) it holds $\det C \neq 0$. We may easily transform (1) so that the latter condition is satisfied by replacing $\mathcal{H}$ with the matrix

$$\widetilde{\mathcal{H}} = \begin{bmatrix} I & -M \\ 0 & I \end{bmatrix} \mathcal{H} \begin{bmatrix} I & M \\ 0 & I \end{bmatrix} = \begin{bmatrix} D - MB & -\mathscr{D}(M) \\ B & BM - A \end{bmatrix}, \tag{24}$$

where M is any $m \times m$ matrix, and $\mathscr{D}(M) = MBM - DM - MA + C$ is the dual operator defined in (8). In fact, we have the following

Lemma 5. *The Riccati equation* (1) *has solution X such that* $\det(I - MX) \neq 0$ *if and only if the Riccati equation*

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

where $\widetilde{A} = A - BM$, $\widetilde{B} = B$, $\widetilde{C} = \mathscr{D}(M)$, $\widetilde{D} = D - MB$, has solution $\widetilde{X} = X(I - MX)^{-1}$ such that $\det(I - MX) \neq 0$*. Moreover, $\widetilde{D} - \widetilde{C}\widetilde{X} = (I - MX)(D - CX)(I - MX)^{-1}$.*

Proof. From

$$\mathcal{H}\begin{bmatrix} I \\ X \end{bmatrix} = \begin{bmatrix} I \\ X \end{bmatrix}(D - CX)$$

one obtains that

$$\widetilde{\mathcal{H}}\begin{bmatrix} I \\ \widetilde{X} \end{bmatrix} = \begin{bmatrix} I \\ \widetilde{X} \end{bmatrix}(I - MX)(D - CX)(I - MX)^{-1}$$

where $(I - MX)(D - CX)(I - MX)^{-1} = \widetilde{D} - \widetilde{C}\widetilde{X}$. □

Under Assumption 3 we may easily choose M such that the $(1,2)$ block of $\widetilde{\mathcal{H}}$ is nonsingular. Indeed, let T be the minimal nonnegative solution of the equation $\mathscr{D}(X) = 0$. It is easy to prove that for $0 \leqslant X \leqslant T$ the matrix $I - MX$ is nonsingular and the derivative of the map $\mathscr{D}(X)$ is a nonsingular $mn \times mn$ M-matrix, therefore by the inverse function theorem $\mathscr{D}(X)$ is locally invertible. In particular its image is an open set, and thus we can expect that, for "nearly any choice" of M, $\mathscr{D}(M)$ is nonsingular.

Remark 1. *Under Assumption 3, when $\mathcal{M}$ is a nonsingular M-matrix, an effective choice of M is any nonnegative matrix such that $Mv = u$, where $u \in \mathbb{R}^n$ and $v \in \mathbb{R}^m$ are such that*

$$\mathcal{M}\begin{bmatrix} u \\ v \end{bmatrix} = e,$$

with e denoting any positive vector, say, the vector of all ones. In fact, in this case $\mathscr{D}(M)$ is a Z-matrix such that $\mathscr{D}(M)u > 0$, and therefore an M-matrix, for a well-known fact on Z-matrices. A possible choice for M is $M = diag(u_1/v_1, \ldots, u_m/v_m)$.

Now let

$$Q = \begin{bmatrix} I & 0 \\ -U & I \end{bmatrix}$$

and consider

$$\widetilde{\mathcal{H}}_U = Q\mathcal{H}Q^{-1} = \begin{bmatrix} D - CU & -C \\ \mathscr{R}(U) & UC - A \end{bmatrix}$$

with $\mathscr{R}(U) = UCU - AU - UD + B$ defined in (8). Choosing $U = C^{-1}D$ yields

$$\widetilde{\mathcal{H}}_{C^{-1}D} = \begin{bmatrix} 0 & -C \\ B - AC^{-1}D & C^{-1}DC - A \end{bmatrix} \tag{25}$$

and

$$\widetilde{\mathcal{H}}_{C^{-1}D}\begin{bmatrix} I \\ X - C^{-1}D \end{bmatrix} = \begin{bmatrix} I \\ X - C^{-1}D \end{bmatrix}(D - CX). \tag{26}$$

From the latter equation we may easily recover two UQMEs just by scaling the block rows of $\widetilde{H}$. This result is summarized by the following

Theorem 6. *Let*

$$\mathcal{A}_0 = (B - AC^{-1}D)C, \quad \mathcal{A}_1 = A - C^{-1}DC, \quad \mathcal{A}_2 = I \tag{27}$$

and let $\mathcal{A}(z) = \mathcal{A}_0 + z\mathcal{A}_1 + z^2\mathcal{A}_2$. *Then the roots of* $\mathcal{A}(z)$ *are the eigenvalues of* $\mathcal{H}$. *Moreover,* X *is a solution of the NARE* (1) *if and only if* $Y = C^{-1}(D - CX)C$ *is a solution of the UQME* $\mathcal{A}_0 + \mathcal{A}_1 Y + \mathcal{A}_2 Y^2 = 0$.

Proof. From (25) one obtains that

$$\mathcal{C} := \begin{bmatrix} -C^{-1} & 0 \\ 0 & I \end{bmatrix} \widetilde{\mathcal{H}}_{C^{-1}D} \begin{bmatrix} -C & 0 \\ 0 & I \end{bmatrix} = \begin{bmatrix} 0 & I \\ -(B - AC^{-1}D)C & -(A - C^{-1}DC) \end{bmatrix}.$$

From (26) it readily follows that

$$\mathcal{C} \begin{bmatrix} -C^{-1} & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} I \\ X - C^{-1}D \end{bmatrix} = \begin{bmatrix} -C^{-1} & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} I \\ X - C^{-1}D \end{bmatrix} (D - CX).$$

Multiplying on the right the above expression by $-C$ yields

$$\mathcal{C} \begin{bmatrix} I \\ C^{-1}(D - CX)C \end{bmatrix} = \begin{bmatrix} I \\ C^{-1}(D - CX)C \end{bmatrix} C^{-1}(D - CX)C$$

which is equivalent to the UQME

$$Y^2 + (A - C^{-1}DC)Y + (B - AC^{-1}D)C = 0, \qquad Y = C^{-1}(D - CX)C. \tag{28}$$

Since the eigenvalues of $\mathcal{H}$ coincide with the eigenvalues of $\mathcal{C}$, then the roots of $\mathcal{A}(z)$ are the eigenvalues of $\mathcal{H}$. □

Similarly, we may derive the following result, which was obtained in [21] in the case of symmetric algebraic Riccati equations.

Theorem 7. *Let*

$$\mathcal{A}_0 = C(B - AC^{-1}D), \quad \mathcal{A}_1 = CAC^{-1} - D, \quad \mathcal{A}_2 = I \tag{29}$$

and let $\mathcal{A}(z) = \mathcal{A}_0 + z\mathcal{A}_1 + z^2\mathcal{A}_2$. *Then the roots of* $\mathcal{A}(z)$ *are the eigenvalues of* $\mathcal{H}$. *Moreover,* X *is a solution of the NARE* (1) *if and only if* $Y = D - CX$ *is a solution of the UQME* $\mathcal{A}_0 + \mathcal{A}_1 Y + \mathcal{A}_2 Y^2 = 0$.

Proof. By following the same argument used in the proof of Theorem 6, one arrives at

$$\begin{bmatrix} 0 & I \\ CAC^{-1}D - CB & D - CAC^{-1} \end{bmatrix} \begin{bmatrix} I \\ D - CX \end{bmatrix} = \begin{bmatrix} I \\ D - CX \end{bmatrix} (D - CX)$$

which is equivalent to the UQME

$$Z^2 + (CAC^{-1} - D)Z + C(B - AC^{-1}D) = 0, \qquad Z = D - CX. \tag{30}$$

□

Observe that the solution Y of (28) is similar to $D - CX$ while the solution of (30) is $Y = D - CX$ so that the solutions of (28) and (30) share the same eigenvalues of $D - CX$. Therefore, if $D - CX$ has eigenvalues with nonnegative real parts, then also Y has eigenvalues with nonnegative real parts.

5 Eigenvalues transformation

The reductions presented in Section 4 can effectively be used to transform the NARE into an UQME. The solution of interest in NAREs is the extremal one, which is the one associated with the eigenvalues in the nonnegative half-plane $\operatorname{Re} z \geq 0$. Algorithms for solving UQME are usually designed to find the solution with eigenvalues inside the unit disc, in fact, in this case they have the best performance; therefore, we need to apply an eigenvalue transformation in order to transform the eigenvalues splitting with respect to the imaginary axis into a splitting with respect to the unit circle. Several strategies are available for this task; the basic idea is the following lemma.

Lemma 8. *Let $a(z) = a_0 + a_1 z + \ldots + a_h z^h$ and $b(z) = b_0 + b_1 z + \ldots + b_l z^l$ be polynomials with complex coefficients, and extend them to square matrices as $a(X) = a_0 + a_1 X + \ldots + a_h X^h$ and $b(X) = b_0 + b_1 X + \ldots + b_l X^l$. Let $f(z) = \frac{a(z)}{b(z)}$ and $f(X) = b(X)^{-1} a(X)$. Let λ_i, be the eigenvalues of the matrix $\mathcal{H}$. If $b(\lambda_i) \neq 0$ for each i, then the eigenvalues of*

$$\widetilde{\mathcal{H}} = f(\mathcal{H})$$

are $f(\lambda_i)$. Moreover, if

$$\mathcal{H} \begin{bmatrix} I \\ X \end{bmatrix} = \begin{bmatrix} I \\ X \end{bmatrix} R \tag{31}$$

holds, then

$$f(\mathcal{H}) \begin{bmatrix} I \\ X \end{bmatrix} = \begin{bmatrix} I \\ X \end{bmatrix} f(R) \tag{32}$$

holds as well.

Proof. The first part is well-known [20]. As for the last formula, by applying repeatedly (31) we get

$$\mathcal{H}^k \begin{bmatrix} I \\ X \end{bmatrix} = \begin{bmatrix} I \\ X \end{bmatrix} R^k \quad \text{for all } k \geqslant 0; \tag{33}$$

then we rewrite (32) as

$$a(\mathcal{H}) \begin{bmatrix} I \\ X \end{bmatrix} b(R) = b(\mathcal{H}) \begin{bmatrix} I \\ X \end{bmatrix} a(R)$$

(observe that $a(R)$ and $b(R)$ commute), and apply (33) to all the monomials appearing in the above expression. □

In particular, this result implies that the solutions of the NARE associated with $\mathcal{H}$ are the same as the ones of the NARE associated with $f(\mathcal{H})$, for each rational function f for which $b(\mathcal{H})$ is nonsingular.

5.1 Shrink and shift

Ramaswami's approach [27] for the eigenvalues transformation is based on a shrink-and-shift approach. Under Assumption 5, for a sufficiently large value of $t > 0$ it holds that $|t - \lambda_i| \leq t$ for $i = 1, \ldots, n$, that is, the eigenvalues of $\mathcal{H}$ with nonnegative real part are contained in a sufficiently large disc $\mathcal{D}$ of center t and radius t. The remaining eigenvalues clearly lie outside $\mathcal{D}$, since they are on the opposite side of the imaginary axis. The transformation $f : z \mapsto 1 - \frac{z}{t}$ maps $\mathcal{D}$ onto the unit disc $|z| \leq 1$. Applying to $\mathcal{H}$ the transformation

$$\mathcal{H} \mapsto f(\mathcal{H}) = I - \frac{1}{t}\mathcal{H},$$

yields the matrix

$$\widehat{\mathcal{H}} = f(\mathcal{H}) = \begin{bmatrix} \widehat{D} & -\widehat{C} \\ \widehat{B} & -\widehat{A} \end{bmatrix}$$

where

$$\widehat{A} = -I - (1/t)A, \quad \widehat{B} = -B, \quad \widehat{C} = -C, \quad \widehat{D} = I - (1/t)D. \tag{34}$$

By Lemma 8 we find that the eigenvalues of $f(\mathcal{H})$ are split with respect to the unit circle. More precisely, for the eigenvalues $\mu_i = f(\lambda_i)$ of $f(\mathcal{H})$ we have

$$|\mu_i| \leq 1 \quad \text{for } i = 1, \ldots, n, \quad |\mu_i| > 1 \quad \text{for } i = n+1, \ldots, n+m,$$

By the same lemma, the solutions to the NARE associated with $f(\mathcal{H})$ are the same as the solutions of the original NARE (1).

As to the choice of t, with Assumption 3 it is sufficient to take

$$t \geqslant \max_{1 \leqslant i \leqslant n} D_{ii}.$$

In fact, with this choice of t the M-matrix $D - CS$ can be put in the form $tI - P$, with P a nonnegative matrix; for some well-known facts on M-matrices then $\rho(P) \leqslant t$, with equality only when $D - CS$ is a singular matrix. Therefore all the eigenvalues λ of $D - CS$ satisfy $|\lambda - t| \leqslant t$.

5.2 Cayley transform

Another possible approach is applying the Cayley transform $\mathcal{C}_\gamma : z \mapsto \frac{z+\gamma}{z-\gamma}$, for $\gamma > 0$. It is easy to see that this function maps the imaginary axis onto the unit circle, the half-plane $\operatorname{Re} z > 0$ onto the open unit disc, and the half-plane $\operatorname{Re} z < 0$ onto the complement of the closed unit disc, for any choice of $\gamma > 0$. Therefore, as in the section above, transforming

$$\mathcal{H} \mapsto \mathcal{H}_\gamma = \mathcal{C}_\gamma(\mathcal{H}) = (\mathcal{H} + \gamma I)^{-1}(\mathcal{H} - \gamma I)$$

maps the n eigenvalues of $\mathcal{H}$ lying in the nonnegative half-plane into the closed unit disc, and the other m eigenvalues outside of it. More precisely, the eigenvalues of $\mathcal{H}_\gamma$ are $\xi_i = \mathcal{C}_\gamma(\lambda_i)$, $i = 1, \ldots, m+n$, and are such that

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

Moreover, the solutions of the Riccati equation (1) are solutions of the Riccati equation associated with $\mathcal{H}_\gamma$, according to Lemma 8.

Observe that the matrix $\mathcal{H}_\gamma$ is given by

$$\mathcal{H}_\gamma = \mathcal{C}_\gamma(\mathcal{H}) = \begin{bmatrix} \widehat{D} & -\widehat{C} \\ \widehat{B} & -\widehat{A} \end{bmatrix}.$$

where

$$\begin{array}{ll} \widehat{A} = -I + 2\gamma V^{-1}, & \widehat{B} = 2\gamma(-A+\gamma I)^{-1}BW^{-1}, \\ \widehat{C} = 2\gamma(D+\gamma I)^{-1}CV^{-1}, & \widehat{D} = I - 2\gamma W^{-1}, \end{array} \tag{35}$$

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

6 Old and new algorithms

The algorithms that we outline in this section are based on two main transformations.

In the first transformation, the Hamiltonian $\mathcal{H}$ is transformed into a matrix $\widetilde{\mathcal{H}}$ whose eigenvalues are split with respect to the unit circle. This can be obtained either by means of the shrink-and-shift technique of Section 5.1, or by means of the Cayley transform of Section 5.2. According to Lemma 8 the solutions of the Riccati equations associated with $\mathcal{H}$ and $\widetilde{\mathcal{H}}$ are the same; in particular, the extremal solution S of the NARE (1) is the solution of the NARE associated with $\widetilde{\mathcal{H}}$ corresponding to the eigenvalues in the unit disk.

In the second transformation, the NARE associated with $\widetilde{\mathcal{H}}$ is reduced into a UQME. This is achieved by means of one of the three techniques of section 4. The resulting UQME is solved by means of cyclic reduction.

According to the combination of the techniques used for performing the two above transformations we obtain known and new algorithms.

6.1 Shrink-and-shift with Ramaswami's reduction

Under Assumption 3, Ramaswami [27] and later Guo [12] proposed two similar algorithms based on logarithmic reduction [24], which is a variant of cyclic reduction. First, one performs the shrink-and-shift transformation defined in section 5.1, to get $\widetilde{\mathcal{H}} = I - (1/t)\mathcal{H}$. Then, one applies Ramaswami's reduction defined in Section 4.1 to the NARE associated with the Hamiltonian $\widetilde{\mathcal{H}}$, to get an $(n+m)\times(n+m)$ UQME whose coefficients have the following block sparsity pattern

$$\begin{bmatrix} * & 0 \\ * & 0 \end{bmatrix} + \begin{bmatrix} -I & * \\ 0 & * \end{bmatrix} Y + \begin{bmatrix} 0 & 0 \\ 0 & * \end{bmatrix} Y^2 = 0. \tag{36}$$

Ramaswami [27] and Guo [12] applied logarithmic reduction to the above UQME. Here we apply cyclic reduction, which has a slightly lower computational cost per step with respect to logarithmic reduction [8]. It is easy to see that the

sparsity pattern of the block coefficients in (36) is preserved during the iterations of cyclic reduction; therefore, CR can be accelerated by working only on "small" blocks and optimizing out the products involving zero blocks.

More precisely, it turns out that applying (17) to the equation (19), where $\mathcal{H}$ is replaced by $\widetilde{\mathcal{H}}$, yields blocks of the kind

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

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

$$\begin{aligned} S^{(k)} &= R_5^{(k)} + R_4^{(k)} R_3^{(k)}, & R_1^{(k+1)} &= -R_1^{(k)} X^{(k)}, \\ Y^{(k)} &= \left(S^{(k)}\right)^{-1} \left(R_2^{(k)} + R_4^{(k)} R_1^{(k)}\right), & R_2^{(k+1)} &= -R_2^{(k)} X^{(k)}, \\ X^{(k)} &= R_3^{(k)} Y^{(k)} - R_1^{(k)}, & R_3^{(k+1)} &= R_3^{(k)} - R_1^{(k)} T^{(k)}, \\ Z^{(k)} &= \left(S^{(k)}\right)^{-1} R_6^{(k)}, & R_4^{(k+1)} &= R_4^{(k)} - R_6^{(k)} Y^{(k)}, \\ T^{(k)} &= R_3^{(k)} Z^{(k)}, & R_5^{(k+1)} &= R_5^{(k)} - R_2^{(k)} T^{(k)}, \\ & & R_6^{(k+1)} &= -R_6^{(k)} Z^{(k)}. \end{aligned} \tag{37}$$

for $k = 0, 1, \ldots$, starting from the initial values $R_1^{(0)} = I - (1/t)D$, $R_2^{(0)} = -B$, $R_3^{(0)} = C$, $R_4^{(0)} = 0$, $R_5^{(0)} = I + (1/t)A$, $R_6^{(0)} = -I$. This way, the CR iteration requires 12 matrix products and one LU factorization per step, leading to a total cost of $\frac{74}{3}n^3$ ops per step (when $m = n$).

From Theorem 2 and from Theorem 3 applied to $\mathcal{H} = \widetilde{\mathcal{H}}$ 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 $\sigma = \max_{i=1,\ldots,n} |\mu_i| / \min_{i=1,\ldots,m} |\mu_{n+i}| < 1$, and $\mu_i = 1 - (1/t)\lambda_i$ for $i = 1, \ldots, m+n$.

6.2 Cayley transform with UL-based reduction

A different approach consists in applying the Cayley transform followed by the UL-based reduction of Section 4.2. The following result shows that SDA is in fact CR applied to the resulting UQME.

Theorem 9. *Let*

$$\begin{bmatrix} L_1 & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} -I & U_1 \\ L_2 & -I \end{bmatrix} Y + \begin{bmatrix} 0 & 0 \\ 0 & U_2 \end{bmatrix} Y^2 = 0, \tag{38}$$

be the UQME obtained by applying the Cayley transform to $\mathcal{H}$ followed by the UL-based transformation of an ARE to a UQME. Then,

$$L_1 = E^{(\gamma)}, \quad L_2 = H^{(\gamma)},$$
$$U_1 = G^{(\gamma)}, \quad U_2 = F^{(\gamma)}.$$

where the matrices $E^{(\gamma)}$, $F^{(\gamma)}$, $G^{(\gamma)}$, $H^{(\gamma)}$, are defined in (12)*. Moreover the matrix sequences $\mathcal{A}_i^{(k)}$, $i = 0, 1, 2$ generated by CR* (17) *applied to* (38) *are given by*

$$\mathcal{A}_0^{(k)} = \begin{bmatrix} E_k & 0 \\ 0 & 0 \end{bmatrix}, \quad \mathcal{A}_1^{(k)} = \begin{bmatrix} -I & G_k \\ H_k & -I \end{bmatrix}, \quad \mathcal{A}_2^{(k)} = \begin{bmatrix} 0 & 0 \\ 0 & F_k \end{bmatrix},$$

where $\{E_k\}$, $\{F_k\}$, $\{G_k\}$ and $\{H_k\}$ are the sequences generated by SDA (13)*.*

Proof. The application of the Cayley transform to $\mathcal{H}$ generates a matrix $\mathcal{H}_\gamma = \mathcal{C}_\gamma(\mathcal{H})$ which can be factored as

$$\mathcal{H}_\gamma = \mathcal{U}^{-1}\mathcal{L} = \begin{bmatrix} I & -G^{(\gamma)} \\ 0 & F^{(\gamma)} \end{bmatrix}^{-1} \begin{bmatrix} E^{(\gamma)} & 0 \\ -H^{(\gamma)} & I \end{bmatrix}, \tag{39}$$

where $E^{(\gamma)}$, $F^{(\gamma)}$, $G^{(\gamma)}$, $H^{(\gamma)}$ are the initial values of the SDA algorithm. Applying the reduction of section 4.2 leads to the UQME (38). We can see that applying CR to this equation preserves the sparsity pattern of the coefficients $\mathcal{A}_i$, that is, the iterates can be written as

$$\mathcal{A}_0^{(k)} = \begin{bmatrix} E_k & 0 \\ 0 & 0 \end{bmatrix}, \quad \mathcal{A}_1^{(k)} = \begin{bmatrix} -I & G_k \\ H_k & -I \end{bmatrix}, \quad \mathcal{A}_2^{(k)} = \begin{bmatrix} 0 & 0 \\ 0 & F_k \end{bmatrix},$$

for suitable matrices E_k, F_k, G_k, H_k. Carrying out the CR iteration (17) block by block leads to exactly the relations (13) that define the SDA algorithm. □

The following result shows that for this specific problem, $\lim_k \mathcal{A}_1^{(k)}$ exists and provides the extremal solution S of the original NARE (1) and the extremal solution T of its dual (15)

Theorem 10. *Let*

$$\mathcal{A}_0 = \begin{bmatrix} E^{(\gamma)} & 0 \\ 0 & 0 \end{bmatrix}, \quad \mathcal{A}_1 = \begin{bmatrix} -I & G^{(\gamma)} \\ H^{(\gamma)} & -I \end{bmatrix}, \quad \mathcal{A}_2 = \begin{bmatrix} 0 & 0 \\ 0 & F^{(\gamma)} \end{bmatrix},$$

where the matrices $E^{(\gamma)}$, $F^{(\gamma)}$, $G^{(\gamma)}$, $H^{(\gamma)}$ are defined in (12)*, and let*

$$\varphi(z) = z^{-1}\mathcal{A}_0 + \mathcal{A}_1 + z\mathcal{A}_1.$$

Define

$$\mathcal{S} = \begin{bmatrix} R_\gamma & 0 \\ SR_\gamma & 0 \end{bmatrix}, \quad \mathcal{T} = \begin{bmatrix} 0 & TU_\gamma \\ 0 & U_\gamma \end{bmatrix}, \quad \widehat{\mathcal{S}} = \begin{bmatrix} 0 & 0 \\ Q_\gamma S & Q_\gamma \end{bmatrix}, \quad \widehat{\mathcal{T}} = \begin{bmatrix} P_\gamma & P_\gamma T \\ 0 & 0 \end{bmatrix},$$

where $R_\gamma = \mathcal{C}_\gamma(R)$ is the Cayley transform of $R = D - CS$, $U_\gamma = \mathcal{C}_\gamma(U)$ is the Cayley transform of $U = BT - A$, $Q_\gamma = F^{(\gamma)}(I - SG^{(\gamma)})^{-1}$, $P_\gamma = E^{(\gamma)}(I - TH^{(\gamma)})^{-1}$. Then:

1. *the matrix* $\mathcal{S}$ *is the only solution to*

$$\mathcal{A}_0 + \mathcal{A}_1 Y + \mathcal{A}_2 Y^2 = 0 \tag{40}$$

such that $\rho(\mathcal{S}) \leq 1$*;*

2. *the matrix* $\widehat{\mathcal{S}}$ *is the only solution to*

$$\mathcal{A}_2 + Y\mathcal{A}_1 + Y^2\mathcal{A}_0 = 0 \tag{41}$$

such that $\rho(\widehat{\mathcal{S}}) < 1$*;*

3. *the matrix* $\mathcal{T}$ *is the only solution to*

$$\mathcal{A}_2 + \mathcal{A}_1 Y + \mathcal{A}_0 Y^2 = 0 \tag{42}$$

such that $\rho(\mathcal{T}) \leq 1$*;*

4. *the matrix* $\widehat{\mathcal{T}}$ *is the only solution to*

$$\mathcal{A}_0 + Y\mathcal{A}_1 + Y^2\mathcal{A}_2 = 0 \tag{43}$$

such that $\rho(\widehat{\mathcal{T}}) < 1$*;*

5. *the following* canonical factorizations *hold*

$$\begin{aligned} \varphi(z) &= (I - z\widehat{\mathcal{S}})\mathcal{W}(I - z^{-1}\mathcal{S}), \\ \varphi(z) &= (I - z^{-1}\widehat{\mathcal{T}})\mathcal{Z}(I - z\mathcal{T}), \end{aligned} \tag{44}$$

valid for $|z| = 1$*, where* $\mathcal{W} = \mathcal{A}_2\mathcal{S} + \mathcal{A}_1 = \begin{bmatrix} -I & G^{(\gamma)} \\ S & -I \end{bmatrix}$, $\mathcal{Z} = \mathcal{A}_0\mathcal{T} + \mathcal{A}_1 = \begin{bmatrix} -I & T \\ H^{(\gamma)} & -I \end{bmatrix}$.

Proof. The matrix $\mathcal{H}_\gamma = \mathcal{C}_\gamma(\mathcal{H})$ has eigenvalues $\xi_i = \mathcal{C}_\gamma(\lambda_i)$, which are split with respect to the unit circle. From Lemma 8 one has $\mathcal{H}_\gamma \begin{bmatrix} I \\ S \end{bmatrix} = \begin{bmatrix} I \\ S \end{bmatrix} R_\gamma$, which in view of (39) is equivalent to

$$\begin{bmatrix} E^{(\gamma)} & 0 \\ -H^{(\gamma)} & I \end{bmatrix} \begin{bmatrix} I \\ S \end{bmatrix} = \begin{bmatrix} I & -G^{(\gamma)} \\ 0 & F^{(\gamma)} \end{bmatrix} \begin{bmatrix} I \\ S \end{bmatrix} R_\gamma. \tag{45}$$

From Theorem 4 applied to $\mathcal{H}_\gamma$ it follows that the matrix $\mathcal{S}$ is the only solution to (40) with eigenvalues $\xi_1, \ldots, \xi_n$, so that $\rho(\mathcal{S}) \leq 1$. It can be easily verified by direct inspection that the matrix $\widehat{\mathcal{S}} = -\mathcal{A}_2\mathcal{W}^{-1}$, for $\mathcal{W} = \mathcal{A}_1 + \mathcal{A}_2\mathcal{S}$, solves equation (41). By using the structure of $\mathcal{A}_1$ and $\mathcal{A}_2$ we find that

$$\mathcal{W} = \begin{bmatrix} 0 & 0 \\ 0 & F^{(\gamma)} \end{bmatrix} \begin{bmatrix} R_\gamma & 0 \\ SR_\gamma & 0 \end{bmatrix} + \begin{bmatrix} -I & G^{(\gamma)} \\ H^{(\gamma)} & -I \end{bmatrix} = \begin{bmatrix} -I & G^{(\gamma)} \\ S & -I \end{bmatrix},$$

From the above representation of $\mathcal{W}$ it follows that

$$\widehat{\mathcal{S}} = -\begin{bmatrix} 0 & 0 \\ 0 & F^{(\gamma)} \end{bmatrix} \begin{bmatrix} -I & G^{(\gamma)} \\ S & -I \end{bmatrix}^{-1} = \begin{bmatrix} 0 & 0 \\ Q_\gamma S & Q_\gamma \end{bmatrix},$$

with $Q_\gamma = F^{(\gamma)}(I - SG^{(\gamma)})^{-1}$. The first factorization in (44) follows from the equation $\widehat{\mathcal{S}} = -\mathcal{A}_2\mathcal{W}^{-1}$ and from the fact that $\widehat{\mathcal{S}}$ solves the matrix equation (41). Since the roots of $\phi(z)$ are $\xi_1, \ldots, \xi_{m+n}$ and since the eigenvalues of $\mathcal{S}$ are $\xi_1, \ldots, \xi_n$, it follows that the eigenvalues of $\widehat{\mathcal{S}}$ are $\xi_{n+1}^{-1}, \ldots, \xi_{m+n}^{-1}$. Therefore $\rho(\widehat{\mathcal{S}}) < 1$.

The properties of the matrices $\mathcal{T}$ and $\widehat{\mathcal{T}}$ as well as the second factorization of (44) can be similarly proved by using the property

$$\begin{bmatrix} E^{(\gamma)} & 0 \\ -H^{(\gamma)} & I \end{bmatrix} \begin{bmatrix} T \\ I \end{bmatrix} = \begin{bmatrix} I & -G^{(\gamma)} \\ 0 & F^{(\gamma)} \end{bmatrix} \begin{bmatrix} T \\ I \end{bmatrix} U_\gamma,$$

with $U_\gamma = \mathcal{C}_\gamma(A - BT)$. □

The computational cost of this algorithm is the same as the cost of SDA, i.e., $64n^3/3$ ops per step.

The solutions S and T can be expressed in functional form by means of the matrix Laurent power series $\psi(z) = \sum_{i=-\infty}^{+\infty} z^i \psi_i = \varphi(z)^{-1}$, where $\varphi(z)$ is the matrix function defined in Theorem 10, as shown by the following

Theorem 11. *The constant coefficient ψ_0 of $\psi(z)$ is nonsingular and such that $\psi_0^{-1} = \begin{bmatrix} -I & T \\ S & -I \end{bmatrix}$. Moroever, for the sequence $\{\mathcal{A}_1^{(k)}\}_k$ generated by CR, it holds*

$$\lim_{k\to\infty} \mathcal{A}_1^{(k)} = \lim_{k\to\infty} \begin{bmatrix} -I & G_k \\ H_k & -I \end{bmatrix} = \begin{bmatrix} -I & T \\ S & -I \end{bmatrix}.$$

The convergence is quadratic, that is,

$$||G_k - T|| = O(\sigma^{2^k}), \ ||H_k - S|| = O(\sigma^{2^k}),$$

for any norm $||\cdot||$, where $\sigma = \max_{i=1,\ldots,n} |\xi_i| / \min_{i=1,\ldots,m} |\xi_{n+i}| < 1$.

Proof. We observe that the hypotheses of Theorem 2 hold since the roots of $\mathcal{A}(z)$ coincide with the eigenvalues ξ_i of $\mathcal{H}_\gamma$ which are split with respect to the unit disk. Moreover, ψ_0 is nonsingular in view of Theorems 2 and 10. The first equation in (44) can be written as

$$\varphi(z) = \left(I - z\begin{bmatrix} 0 & 0 \\ Q_\gamma S & Q_\gamma \end{bmatrix}\right) \begin{bmatrix} -I & G^{(\gamma)} \\ S & -I \end{bmatrix} \left(I - z^{-1}\begin{bmatrix} R_\gamma & 0 \\ SR_\gamma & 0 \end{bmatrix}\right).$$

For $|z| = 1$, we invert both sides of the last equation to get

$$\psi(z) := \varphi(z)^{-1} = \left(\sum_{j\geqslant 0} z^{-j} \begin{bmatrix} R_\gamma^j & 0 \\ SR_\gamma^j & 0 \end{bmatrix}\right) \mathcal{W}^{-1} \left(\sum_{j\geqslant 0} z^j \begin{bmatrix} 0 & 0 \\ Q_\gamma^j S & Q_\gamma^j \end{bmatrix}\right).$$

The constant term of $\psi(z) = \sum_{i\in\mathbb{Z}} \psi_i z^i$ is

$$\begin{aligned}\psi_0 =& \sum_{j\geqslant 0} \begin{bmatrix} R_\gamma^j & 0 \\ SR_\gamma^j & 0 \end{bmatrix} \mathcal{W}^{-1} \begin{bmatrix} 0 & 0 \\ Q_\gamma^j S & Q_\gamma^j \end{bmatrix} \\ =& \mathcal{W}^{-1} + \begin{bmatrix} I \\ S \end{bmatrix} \left(\sum_{j\geqslant 1} \begin{bmatrix} R_\gamma^j & 0 \end{bmatrix} \mathcal{W}^{-1} \begin{bmatrix} 0 \\ Q_\gamma^j \end{bmatrix} \right) \begin{bmatrix} S & I \end{bmatrix}\end{aligned} \tag{46}$$

For the convergence properties of CR, see Theorem 2, one has

$$\lim_{k\to\infty} \mathcal{A}_1^{(k)} = \psi_0^{-1}$$

that is,

$$\lim_{k\to\infty} \begin{bmatrix} -I & G_k \\ H_k & -I \end{bmatrix} = \psi_0^{-1}. \tag{47}$$

Using (46), we can say more on the structure of ψ_0^{-1}; defining

$$K = \sum_{j\geqslant 1} \begin{bmatrix} R_\gamma^j & 0 \end{bmatrix} \mathcal{W}^{-1} \begin{bmatrix} 0 \\ Q_\gamma^j \end{bmatrix},$$

we get

$$\psi_0^{-1} = \mathcal{W}^{-1} + \begin{bmatrix} I \\ S \end{bmatrix} K \begin{bmatrix} S & I \end{bmatrix}.$$

Applying the Sherman-Morrison-Woodbury formula (see e.g. [10]) yields

$$\psi_0 = \mathcal{W} + \mathcal{W} \begin{bmatrix} I \\ S \end{bmatrix} \widehat{K}^{-1} \begin{bmatrix} S & I \end{bmatrix} \mathcal{W}, \tag{48}$$

where $\widehat{K}$ is the auxiliary matrix to be inverted in the SMW formula. We are not concerned with the explicit value and structure of $\widehat{K}$ now. Since

$$\mathcal{W} \begin{bmatrix} I \\ S \end{bmatrix} = \begin{bmatrix} * \\ 0 \end{bmatrix}, \quad \begin{bmatrix} S & I \end{bmatrix} \mathcal{W} = \begin{bmatrix} 0 & * \end{bmatrix}$$

($*$ stands for an arbitrary block of the right size), the second summand in the right-hand side of (48) is nonzero only in its $(1,2)$ block, thus we get

$$\psi_0^{-1} = \begin{bmatrix} -I & * \\ S & -I \end{bmatrix}.$$

Combining the latter equation with (47) yields $\lim_{k\to\infty} H_k = S$.

Using the second factorization in (44), similarly we may show that

$$\psi_0^{-1} = \begin{bmatrix} -I & T \\ * & -I \end{bmatrix}.$$

Therefore, we conclude that $\lim_{k\to\infty} G_k = T$. □

6.3 Shrink-and-shift with UL-based reduction

Here, we combine the shrink-and-shift technique with the UL-based reduction in order to arrive at a UQME having a splitting with respect to the unit circle.

Formally, we start from $\widetilde{\mathcal{H}} = 1 - \frac{1}{t}\mathcal{H}$, factor it as

$$\widetilde{\mathcal{H}} = \begin{bmatrix} D_t & t^{-1}C \\ -t^{-1}B & A_t \end{bmatrix} = \begin{bmatrix} I & -t^{-1}CA_t^{-1} \\ 0 & A_t^{-1} \end{bmatrix}^{-1} \begin{bmatrix} D_t - t^{-2}CA_t^{-1}B & 0 \\ -t^{-1}A_t^{-1}B & I \end{bmatrix},$$

with $D_t := I - t^{-1}D$ and $A_t := I + t^{-1}A$ and reduce it to a UQME with the same structure as (22). Cyclic reduction applied to this UQME is defined by the same relations (13) as SDA with initial values

$$\begin{aligned} \widehat{D}_0 &= D_t - t^{-2}CA_t^{-1}B, \\ \widehat{F}_0 &= A_t^{-1}, \\ \widehat{G}_0 &= t^{-1}CA_t^{-1}, \\ \widehat{H}_0 &= -t^{-1}A_t^{-1}B. \end{aligned}$$

The advantage is that we get an algorithm with the same computational cost as SDA that is, $\frac{64}{3}n^3$ ops per step, having somewhat simpler initial values.

The sequences G_k and H_k converge to T and S, respectively; the convergence speed is given by the bounds

$$||G_k - T|| = O(\sigma^{2^k}), \quad ||H_k - S|| = O(\sigma^{2^k})$$

for $\sigma = \max_{i=1,\ldots,n} |\eta_i| / \min_{i=1,\ldots,m} |\eta_{n+i}| < 1$.

6.4 Algorithms based on the small size transform

The reductions described in Section 4.3 lead to a UQME whose coefficients are $n \times n$ matrices. In this case there is no block structure to exploit in the CR computation so that the computational cost is just $38n^3/3$ ops per step, independently of the eigenvalue transformation applied to $\mathcal{H}$.

The UQME that we have to solve is

$$Y^2 + (\widehat{C}\widehat{A}\widehat{C}^{-1} - \widehat{D})Y + \widehat{C}(\widehat{B} - \widehat{A}\widehat{C}^{-1}\widehat{D}) = 0 \tag{49}$$

where the matrices $\widehat{A}, \widehat{B}, \widehat{C}$ and $\widehat{D}$ are given by (34) if the shrink-and-shift technique is used for transforming the spectrum of $\mathcal{H}$ and by (35) if the Cayley transform is used.

The convergence of the algorithm is ruled by Theorem 2. More precisely, one has

$$||Y_k - Y|| = O(\sigma^{2^k})$$

where $\sigma = \max_{i=1,\ldots,n} |\xi_i| / \min_{i=1,\ldots,m} |\xi_{n+i}| < 1$ for the equation obtained with the shrink-and-shift technique, and $\sigma = \max_{i=1,\ldots,n} |\eta_i| / \min_{i=1,\ldots,m} |\eta_{n+i}| < 1$ for the equation obtained with the Cayley transform. Here $Y = \widehat{C}^{-1}(\widehat{D} - \widehat{C}S)\widehat{C}$ is the solution of the UQME and $Y_k = -\mathcal{B}_k^{-1}\mathcal{A}_0$ is the approximation to Y provided at the kth step of CR, see equation (17).

6.5 Applicability of CR

Observe that, if the original Riccati equation (1) is associated with an M-matrix, then the UQMEs obtained by means of the reduction of Sections 4.1 and 4.2, are still associated with M-matrices. This property guarantees that CR does not break down since all the blocks $\mathcal{A}_1^{(k)}$ are nonsingular M-matrices. This property does not hold in general for the reduction to UQME of Section 4.3. In this case, the applicability of CR requires more analysis.

With the transformation based on shrink-and-shift, both the UQMEs (28) and (30) are of the kind

$$Y^2 + Y(2I - (1/t)V) + I - (1/t)V + (1/t)^2 V' = 0 \tag{50}$$

for suitable matrices V and V'. In particular, one has

$$V = C^{-1}DC - A, \quad V' = BC - AC^{-1}DC, \quad \text{for equation (28)},$$

$$V = D - CAC^{-1}, \quad V' = CB - CAC^{-1}D, \quad \text{for equation (30)}.$$

Moreover, for a suitable choice of t, the polynomial $\det(z^2 I + z(2I - (1/t)V) + I - (1/t)V + (1/t)^2 V')$ has a root splitting w.r.t. the unit circle which is a necessary condition for the convergence of CR. This analysis enables us to state some properties concerning the applicability of CR.

We recall that if the $n \times n$ block tridiagonal matrices $\mathrm{tridiag}_n(I, 2I-(1/t)V, I-(1/t)P + (1/t)^2 U)$ are nonsingular for any n, then CR can be applied with no break-down [5]. A first-order analysis can be carried out by neglecting the $O(1/t^2)$ terms. In fact, we may consider $\mathcal{T}_n = \mathrm{tridiag}_n(I, 2I-(1/t)V, I-(1/t)V)$ and observe that the set of its eigenvalues is the union of the eigenvalues of the matrices $T_n^{(i)} = \mathrm{tridiag}_n(1, 2 - (1/t)v_i, 1 - (1/t)v_i)$ where v_i ranges in the set of the eigenvalues of V. Simple arguments show that the eigenvalues of $T_n^{(i)}$ are given by $\lambda_j^{(i)} = 2 - (1/t)v_i + 2\sqrt{1 - (1/t)v_i}\cos(\pi j/(n+1))$, $j = 1, \ldots, n$. In a first-order analysis we may approximate $\sqrt{1 - (1/t)v_i} \approx 1 - (1/t)v_i/2$ and get $\lambda_j^{(i)} = (2 - (1/t)v_i)(1 + \cos(\pi j/(n+1)) + O(1/t^2)$. This shows that the eigenvalues are power series in $1/t$ with the first two terms nonzero. Thus, for sufficiently large t, cyclic reduction can be applied with no breakdown.

7 Numerical experiments

We implemented and tested the following algorithms:

`sda`: the algorithm based on Cayley transform and UL factorization (SDA), as outlined in Sections 3.1 and 6.2;

`ss-ul`: the algorithm based on shrink-and-shift and UL factorization, described in Section 6.3;

`ss-ram`: the algorithm based on shrink-and-shift and on Ramaswami's reduction, described in Section 6.1;

n	`sda`	`ss-ul`	`ss-ram`	`nodoub`
8	0.045209	0.02735	0.030078	0.027061
16	0.039896	0.041282	0.046027	0.03845
32	0.14559	0.14666	0.18047	0.13432
64	0.92806	0.93415	1.1707	0.8448
128	7.2632	7.3491	9.1974	6.6499
256	60.841	61.926	76.835	55.03
512	499.95	504.37	625.06	448.46

Table 1: Running time in seconds for Test 1

`nodoub`: the algorithm based on shrink-and-shift and on the reduction to the UQME (49) of size n proposed in Section 6.4.

We applied the algorithms to two test cases, choosing for each one several different values of the size $m = n$ of the coefficients.

Test 1 The structured NARE presented in [22], deriving from a problem in neutron transport theory. The coefficients of this equation are in the form

$$A = \widehat{\Delta} - eq^T, \qquad B = ee^T,$$
$$C = qq^T, \qquad D = \Delta - qe^T,$$

with

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

$(x_i)_{i=1}^n$ and $(w_i)_{i=1}^n$ being the nodes and weights of a Gaussian discretization. Here we have chosen $\alpha = 10^{-8}$, $c = 1 - 10^{-6}$, which yields a close-to-null-recurrent Riccati equation.

Test 2 The equation associated with a randomly chosen singular M-matrix M, generated using Octave's commands `R=rand(2*n)` and `M=diag(R*ones(2*n,1))-R`. The reported values are the average of ten different choices of the random matrix.

The experiments were performed on a 2.8Ghz Xeon, implementing the algorithms with GNU Octave; in Tables 1 and 2 we list the running times and absolute residual for different choices of the size n of the matrices for Test 1.

Similarly we do in Tables 3 and 4 for Test 2.

The residual was calculated as $\|XCX + B - AX - XD\|_1$, with X being the computed solution.

The computational times reflect the difference in the costs per step of the different algorithms: `nodoub` is the fastest, followed by the two variants of SDA (which share the same cost per step), and `ss-ram` is the slowest.

n	sda	ss-ul	ss-ram	nodoub
8	1.654e-13	5.8367e-14	6.6482e-14	1.4294e-11
16	1.328e-12	2.4418e-13	2.7769e-13	1.6405e-10
32	3.4631e-12	1.964e-12	1.7786e-12	7.8717e-10
64	2.2679e-11	1.3598e-11	8.2769e-12	7.8282e-09
128	1.3316e-10	8.1521e-11	6.4269e-11	5.4047e-08
256	1.0096e-09	5.6852e-10	3.7115e-10	4.5315e-07
512	6.7923e-09	4.2861e-09	1.7767e-09	5.4083e-06

Table 2: Absolute residual for Test 1

n	sda	ss-ul	ss-ram	nodoub
8	0.016927	0.015696	0.017	0.015045
16	0.028276	0.028877	0.032625	0.026565
32	0.083346	0.084624	0.099644	0.07022
64	0.48015	0.48756	0.58651	0.38958
128	4.8408	4.8895	6.1907	3.9967
256	34.036	34.497	40.72	26.933
512	291.47	295.6	354.06	228.18

Table 3: Running time in seconds for Test 2

n	sda	ss-ul	ss-ram	nodoub
8	4.3812e-15	3.8386e-15	2.8644e-15	2.8628e-11
16	1.3656e-14	1.0136e-14	6.8251e-15	9.1426e-11
32	3.8594e-14	2.3889e-14	1.8441e-14	3.2387e-10
64	1.1038e-13	6.2969e-14	4.6679e-14	2.5328e-08
128	3.6836e-13	1.5803e-13	1.2221e-13	6.6213e-09
256	1.0805e-12	4.3243e-13	3.3097e-13	5.6768e-10
512	3.2239e-12	1.1668e-12	9.0803e-13	3.8776e-10

Table 4: Absolute residual for Test 2

The solution computed by algorithm `nodoub` has a fairly large residual, many times larger than the other ones. In order to overcome this problem, the obtained solution can be further refined with one or more iterations of the Newton algorithm, with a similar strategy to that proposed in [16].

We also note that the proposed variant to SDA, `ss-ul`, gives more accurate results than the ones of the original algorithm. Therefore, in most cases it is convenient to apply `ss-ul` instead of `sda` or `ss-ram`, since it combines the advantages of the two: the lower cost per step of the former and the better accuracy of the latter.

8 Conclusions and open issues

The interpretation provided in this paper casts new light on the SDA algorithm and on the relationship between UQMEs and NAREs. The new algorithms proposed need to be examined more extensively; moreover, several other approaches to the solution of the NARE can be developed with this new setting. Among the possible ideas, we propose here:

- using numerical integration and the Cauchy integral theorem for computing the matrix ψ_0 of Theorem 11;
- using functional iterations borrowed from stochastic processes (QBD) for solving the UQME;
- using Newton's iteration applied to the UQME trying to exploit the specific matrix structure.

An important issue which is worth being analyzed is the search for more general transformations which map a Hamiltonian matrix $\mathcal{H}$ to a new one $\widetilde{\mathcal{H}}$ where the block $\widetilde{\mathcal{H}}_{1,2}$ is not only nonsingular but numerically well conditioned.

Another important case concerns the application of our techniques to the equation coming from transport theory of [22].

It is also to be noted that the convergence speed of CR depends on the ratio

$$\frac{\rho(f(D-CS))}{\rho(f(A-SC)^{-1})},$$

where f is either the Cayley or the shrink-and-shift transform. Observe that this is the ratio between the largest (in modulus) eigenvalue of $f(\mathcal{H})$ inside the open unit disc and the smallest eigenvalue outside it, or, with our notation for the eigenvalues of $\mathcal{H}$,

$$\frac{\max_{i=1,\ldots,n}|f(\lambda_i)|}{\min_{j=n+1,\ldots,n+m}|f(\lambda_j)|}.$$

It follows that the Cayley transform provides a slightly better convergence ratio than the shrink-and-shift transform, though in most cases the difference is negligible, as in our previous examples. It would be interesting to test other functions f mapping $\lambda_1,\ldots,\lambda_n$ inside the unit circle and the other eigenvalues

outside, in order to determine which one yields the best convergence ratio and the best accuracy in the computed solution.

References

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

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

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

[4] D. Bini, B. Meini, and V. Ramaswami. A probabilistic interpretation of cyclic reduction and its relationships with logarithmic reduction. *Calcolo*, 2008. To appear.

[5] D. A. Bini, L. Gemignani, and B. Meini. Computations with infinite Toeplitz matrices and polynomials. *Linear Algebra Appl.*, 343/344:21–61, 2002. Special issue on structured and infinite systems of linear equations.

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

[7] D. A. Bini, B. Iannazzo, B. Meini, and F. Poloni. Nonsymmetric algebraic Riccati equations associated with an M-matrix: recent advances and algorithms. In *Dagstuhl Seminar Proceedings, "Numerical Methods for Structured Markov Chains", 07461*, 2007.

[8] 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.

[9] B. L. Buzbee, G. H. Golub, and C. W. Nielson. On direct methods for solving Poisson's equations. *SIAM J. Numer. Anal.*, 7:627–656, 1970.

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

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

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

[13] C.-H. Guo. A new class of nonsymmetric algebraic Riccati equations. *Linear Algebra Appl.*, 426(2–3):636–649, 2007.

[14] C.-H. Guo and N. J. Higham. Iterative solution of a nonsymmetric algebraic Riccati equation. *Numer. Linear Algebra Appl.*, 12(2-3):191–200, 2005.

[15] C.-H. Guo, B. Iannazzo, and B. Meini. On the doubling algorithm for a (shifted) nonsymmetric algebraic Riccati equation. Technical report, Dipartimento di Matematica, Università di Pisa, Pisa, Italy, May 2006. To appear in SIMAX.

[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(2):376–391 (electronic), 2000.

[17] C.-H. Guo and W.-W. Lin. Convergence rate of the cyclic reduction algorithm for null recurrent quasi-birth-death problems. Technical report, Personal communication, 2007.

[18] X.-X. Guo, W.-W. Lin, and S.-F. Xu. A structure-preserving doubling algorithm for nonsymmetric algebraic Riccati equation. *Numer. Math.*, 103(3):393–412, 2006.

[19] R. W. Hockney. A fast direct solution of Poisson's equation using Fourier analysis. *J. Assoc. Comput. Mach.*, 12:95–113, 1965.

[20] R. A. Horn and C. R. Johnson. *Topics in matrix analysis*. Cambridge University Press, Cambridge, 1994. Corrected reprint of the 1991 original.

[21] B. Iannazzo and D. Bini. A cyclic reduction method for solving algebraic Riccati equations. Technical report, Dipartimento di Matematica, Università di Pisa, 2003.

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

[23] P. Lancaster and L. Rodman. *Algebraic Riccati equations*. Oxford Science Publications. The Clarendon Press Oxford University Press, New York, 1995.

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

[25] W.-W. Lin and S.-F. Xu. Convergence analysis of structure-preserving doubling algorithms for Riccati-type matrix equations. *SIAM J. Matrix Anal. Appl.*, 28(1):26–39 (electronic), 2006.

[26] 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.

[27] V. Ramaswami. Matrix analytic methods for stochastic fluid flows. In *Proceedings of the 16th International Teletraffic Congress*, pages 19–30. Elsevier Science, Edinburg, 1999.