

Deflating Irreducible Singular M- Matrix Algebraic Riccati Equations

**Wei-Guo Wang
Wei-Chao Wang
Ren-Cang Li**

Technical Report 2011-20

Deflating Irreducible Singular M -Matrix Algebraic Riccati Equations

Wei-guo Wang*

Wei-chao Wang[†]

Ren-cang Li[‡]

September 7, 2011

Abstract

A deflation technique is presented for an irreducible singular M -matrix Algebraic Riccati Equation (MARE). The technique improves the rate of convergence of a doubling algorithm, especially for an MARE in the critical case for which without deflation the doubling algorithm converges linearly and with deflation it converges quadratically. The deflation also improves the conditioning of the MARE in the critical case and thus enables its minimal nonnegative solution to be computed more accurately.

1 Introduction

An M -Matrix Algebraic Riccati Equation¹ (MARE) is the matrix equation

$$XDX - AX - XB + C = 0, \quad (1.1)$$

in which A , B , C , and D are matrices whose sizes are determined by the partitioning

$$W = \begin{matrix} & m & n \\ m & \begin{pmatrix} B & -D \\ -C & A \end{pmatrix} \end{matrix}, \quad (1.2)$$

and W is a nonsingular or an irreducible singular M -matrix. Such Riccati equations arise in applied probability, transportation theory, and stochastic fluid models, and have been extensively

*School of Mathematical Sciences, Ocean University of China, Qingdao, 266100, P.R. China. Email: wgwang@ouc.edu.cn. Supported in part by the National Natural Science Foundation of China Grant 10971204 and 11071228, China Scholarship Council, Shandong Province Natural Science Foundation Grant Y2008A07, and the Fundamental Research Funds for the Central Universities Grant 201013048. This work was initiated while this author was a visiting scholar at Department of Mathematics, University of Texas at Arlington from September 2010 to August 2011.

[†]Department of Mathematics, University of Texas at Arlington, P.O. Box 19408, Arlington, TX 76019. Email: weichao.wang@mavs.uta.edu. Supported in part by the National Science Foundation Grants DMS-0810506 and DMS-1115817.

[‡]Department of Mathematics, University of Texas at Arlington, P.O. Box 19408, Arlington, TX 76019. E-mail: rcli@uta.edu. Supported in part by the National Science Foundation Grants DMS-0810506 and DMS-1115817.

¹MARE was recently coined in [22, 21] to better reflect its characteristics.

studied. See [11, 9, 13, 14, 15, 16, 18, 19] and the references therein. It is shown in [9, 13] that MARE(1.1) has a unique minimal nonnegative solution Φ , i.e., in the entrywise sense,

$$\Phi \leq X \quad \text{for any other nonnegative solution } X \text{ of (1.1).}$$

Recently several doubling algorithms have been proposed to compute Φ efficiently and accurately. They include the structure-preserving doubling algorithm (SDA) of Guo, Lin, and Xu [14], the doubling algorithm called *SDA-ss* of Bini, Meini, and Poloni [4] which combined a shrink-and-shift approach of Ramaswami [18], and the alternating-directional doubling algorithm (ADDA) of Wang, Wang, and Li [20]. We point out that the idea of using a doubling algorithm for Riccati-type equations traces back to 1970s (see [1] and references therein). Recent resurgence of interests in the idea, however, attributes to [7, 6] and has since led to efficient doubling algorithms for various nonlinear matrix equations. In particular, (the optimal) ADDA is the fastest among all doubling algorithms derivable from bilinear transformations [20].

These doubling algorithms are very fast and efficient as they are globally and quadratically convergent, except for the so-called critical case [5]. Specifically, suppose W is *irreducible* and *singular*, and let $u, x \in \mathbb{R}^m$ and $v, y \in \mathbb{R}^n$ be entrywise positive vectors such that

$$W \begin{pmatrix} x \\ y \end{pmatrix} = 0, \quad \begin{pmatrix} u \\ v \end{pmatrix}^T W = 0. \quad (1.3)$$

We call MARE (1.1) is in the *critical case* if $u^T x = v^T y$. For the critical case, the doubling algorithms converge linearly [5], and thus are slow compared to the non-critical case. Define

$$H \stackrel{\text{def}}{=} \begin{pmatrix} I_m & \\ & -I_n \end{pmatrix} W = \begin{pmatrix} B & -D \\ C & -A \end{pmatrix}. \quad (1.4)$$

H is singular if and only if W is singular, and (1.3) implies

$$H \begin{pmatrix} x \\ y \end{pmatrix} = 0, \quad \begin{pmatrix} u \\ -v \end{pmatrix}^T H = 0. \quad (1.5)$$

Since the necessary condition for being in the critical case is H being singular, to speed up the convergence, Guo, Iannazzo, and Meini [12] proposed to shift away its eigenvalue 0 to a properly chosen positive number η :

$$\hat{H} = H + \eta z w^T,$$

before SDA is applied, where $z = \begin{pmatrix} x \\ y \end{pmatrix}$, and $w \in \mathbb{R}^{m+n}$ is entrywise nonnegative such that $w^T z = 1$. Dramatic improvements in reducing the number of iterative steps required for convergence were witnessed. In this article, we propose an alternative approach – deflation – to deflate out the eigenvalue 0 of H , before a doubling algorithm, ADDA in this case, is applied. We also argue that this shifting idea of Guo, Iannazzo, and Meini should be combined with ADDA, instead of SDA, for better performance.

Throughout this article, A , B , C , and D , unless explicitly stated differently, are reserved for the coefficient matrices of MARE (1.1) for which

W defined by (1.2) is an irreducible singular M -matrix, and (1.3) holds, where $0 < u, x \in \mathbb{R}^m$ and $0 < v, y \in \mathbb{R}^n$.

(1.6)

The rest of this paper is organized as follows. Section 2 presents essential properties of an irreducible singular MARE to be used later. Section 3 outlines ADDA originally developed for an MARE but will be applied to certain Algebraic Riccati Equations (AREs) later in this article. Our main contributions are described in detail in sections 4 and 5, beginning by laying out our deflating framework and its convergent analysis in section 4 and then giving out two efficient numerical realizations of the framework. We outline the shifting approach of Guo, Iannazzo, and Meini [12] in section 6 for comparison purpose. Several numerical examples are presented in section 7 to demonstrate the effectiveness of our deflating approach as well as the shifting approach of Guo, Iannazzo, and Meini. Finally in section 8 we give our concluding remarks.

Notation. $\mathbb{R}^{n \times m}$ is the set of all $n \times m$ real matrices, $\mathbb{R}^n = \mathbb{R}^{n \times 1}$, and $\mathbb{R} = \mathbb{R}^1$. I_n (or simply I if its dimension is clear from the context) is the $n \times n$ identity matrix and e_j is its j th column. $\mathbf{1}_{n,m} \in \mathbb{R}^{n \times m}$ is the matrix of all ones, and $\mathbf{1}_n = \mathbf{1}_{n,1}$. The superscript “ \cdot^T ” takes the transpose of a matrix or a vector. For $X \in \mathbb{R}^{n \times m}$,

1. $X_{(i,j)}$ refers to its (i,j) th entry; $X_{(i,:)}$ refers to its i th row; $X_{(:,j)}$ refers to its j th column;
2. when $m = n$, $\rho(X)$ is the spectral radius of X , $\text{eig}(X)$ is the set of the eigenvalues of X , and

$$\mathcal{C}(X; \alpha, \beta) \stackrel{\text{def}}{=} (X - \alpha I)(X + \beta I)^{-1}$$

is the so-called *generalized Cayley transformation* of X ;

3. $\|X\|_1 = \max_{1 \leq j \leq m} \sum_{i=1}^n |X_{(i,j)}|$ is the ℓ_1 -operator norm of X .

Inequality $X \leq Y$ means $X_{(i,j)} \leq Y_{(i,j)}$ for all (i,j) , and similarly for $X < Y$, $X \geq Y$, and $X > Y$. In particular, $X \geq 0$ means that X is entrywise nonnegative.

2 Irreducible Singular MARE

$A \in \mathbb{R}^{n \times n}$ is called a *Z-matrix* if it has nonpositive off-diagonal entries [3, p.284]. Any *Z-matrix* A can be written as $sI - N$ with $N \geq 0$, and it is called an *M-matrix* if $s \geq \rho(N)$; it is a *singular M-matrix* if $s = \rho(N)$ and a *nonsingular M-matrix* if $s > \rho(N)$.

We call MARE (1.1) an *irreducible singular MARE* if its associated coefficient matrix W given by (1.2) is an irreducible singular *M-matrix*. An irreducible singular MARE (1.1) always has a unique minimal nonnegative solution Φ [11] and its *complementary M-Matrix Algebraic Riccati Equation* (cMARE)

$$YCY - YA - BY + D = 0 \tag{2.1}$$

is also an irreducible singular MARE and thus has a unique minimal nonnegative solution Ψ , too. Some properties of Φ and Ψ are summarized in Theorem 2.1 below.

Theorem 2.1 ([9, 10, 11, 12]). *Assume (1.6).*

- (a) *MARE (1.1) has a unique minimal nonnegative solution Φ , and its cMARE (2.1) has a unique minimal nonnegative solution Ψ ;*
- (b) *$\Phi > 0$ and $A - \Phi D$ and $B - D\Phi$ are irreducible M-matrices;*
- (c) *Let $\mu = u^T x - v^T y$. Then*

1. If $\mu > 0$, then $B - D\Phi$ is a singular M -matrix with² $(B - D\Phi)x = 0$ and $A - \Phi D$ is a nonsingular M -matrix, and $\Phi x = y$, $\Psi y < x$;
2. If $\mu = 0$, then both $B - D\Phi$ and $A - \Phi D$ are singular M -matrices, and $\Phi x = y$, $\Psi y = x$;
3. If $\mu < 0$, then $B - D\Phi$ is a nonsingular M -matrix and $A - \Phi D$ is a singular M -matrix, and $\Phi x < y$, $\Psi y = x$.

(d) Let $\text{eig}(H) = \{\lambda_1, \dots, \lambda_{m+n}\}$, where λ_i 's are ordered by their nonincreasing real parts, i.e., $\text{Re}\lambda_j \leq \text{Re}\lambda_i$ for $i < j$. Then λ_m and λ_{m+1} are real, and

$$\text{Re}\lambda_{m+n} \leq \dots \leq \text{Re}\lambda_{m+2} < \lambda_{m+1} \leq 0 \leq \lambda_m < \text{Re}\lambda_{m-1} \leq \dots \leq \text{Re}\lambda_1, \quad (2.2a)$$

$$\text{eig}(B - D\Phi) = \text{eig}(B - \Psi C) = \{\lambda_1, \dots, \lambda_m\}, \quad (2.2b)$$

$$\text{eig}(A - \Phi D) = \text{eig}(A - C\Psi) = \{-\lambda_{m+1}, \dots, -\lambda_{m+n}\}, \quad (2.2c)$$

and

$$\begin{cases} \lambda_m = 0, \lambda_{m+1} < 0, & \text{if } \mu > 0; \\ \lambda_m = \lambda_{m+1} = 0, & \text{if } \mu = 0; \\ \lambda_m > 0, \lambda_{m+1} = 0, & \text{if } \mu < 0. \end{cases} \quad (2.2d)$$

3 ADDA: Alternating-Directional Doubling Algorithm

In this section, we briefly review the Alternating-Directional Doubling Algorithm (ADDA). Although it was originally proposed for an MARE [20], ADDA in principle can be applied to any Algebraic Riccati Equation (ARE), just that for a general ARE the optimal parameter selection and analysis in [20] are no longer valid. Since later in this article we will apply ADDA to AREs that are not necessarily MAREs, in what follows we simply state ADDA for a general ARE. Without causing any confusion, in the rest of this section we still use

$$XDX - AX - XB + C = 0 \quad (3.1)$$

to represent a general ARE, while in the rest of this article it is always assumed to be an MARE satisfying (1.6).

Pick some scalars α and β (such that all involved inverses exist³) and set

$$A_\beta = A + \beta I, \quad B_\alpha = B + \alpha I, \quad (3.2)$$

$$U_{\alpha\beta} = A_\beta - CB_\alpha^{-1}D, \quad V_{\alpha\beta} = B_\alpha - DA_\beta^{-1}C, \quad (3.3)$$

and

$$E_0 = I - (\beta + \alpha)V_{\alpha\beta}^{-1}, \quad Y_0 = (\beta + \alpha)B_\alpha^{-1}DU_{\alpha\beta}^{-1}, \quad (3.4a)$$

$$F_0 = I - (\beta + \alpha)U_{\alpha\beta}^{-1}, \quad X_0 = (\beta + \alpha)U_{\alpha\beta}^{-1}CB_\alpha^{-1}. \quad (3.4b)$$

²[9, Theorem 4.8] says in this case $D\Phi x = Dy$ which leads to $(B - D\Phi)x = Bx - Dy = 0$.

³We know how to ensure this for an MARE [20].

ADDA computes sequences $\{X_k\}$ and $\{Y_k\}$ iteratively by

$$E_{k+1} = E_k(I_m - Y_k X_k)^{-1} E_k, \quad (3.5a)$$

$$F_{k+1} = F_k(I_n - X_k Y_k)^{-1} F_k, \quad (3.5b)$$

$$X_{k+1} = X_k + F_k(I_n - X_k Y_k)^{-1} X_k E_k, \quad (3.5c)$$

$$Y_{k+1} = Y_k + E_k(I_m - Y_k X_k)^{-1} Y_k F_k. \quad (3.5d)$$

In [20], it is derived that

$$E_k = (I - Y_k X) [\mathcal{C}(R; \beta, \alpha)]^{2^k}, \quad (3.6a)$$

$$X - X_k = F_k X [\mathcal{C}(R; \beta, \alpha)]^{2^k}, \quad (3.6b)$$

$$Y - Y_k = E_k Y [\mathcal{C}(S; \alpha, \beta)]^{2^k}, \quad (3.6c)$$

$$F_k = (I - X_k Y) [\mathcal{C}(S; \alpha, \beta)]^{2^k}, \quad (3.6d)$$

where X is a solution of ARE (3.1) and Y is a solution of its *complementary* ARE

$$YCY - YA - BY + D = 0 \quad (3.7)$$

and

$$S = A - CY, \quad R = B - DX. \quad (3.8)$$

Equations (3.6b) and (3.6c) give errors in X_k and Y_k as approximations to the solutions of (3.1) and (3.7), respectively. If their right-hand sides go to 0 as $k \rightarrow \infty$, X_k and Y_k converge to the solutions, respectively. Convergence in general is hard to guarantee, but for an MARE we have the following theorem primarily from [20], except the convergence for the critical case which was established in [12].

Theorem 3.1. *For MARE (1.1), i.e., W given by (1.2) is a nonsingular or an irreducible singular M -matrix, ADDA produces monotonically convergent sequences $\{X_k\}$ and $\{Y_k\}$:*

$$\begin{aligned} 0 \leq X_k \leq X_{k+1} \leq \Phi, \quad \lim_{k \rightarrow \infty} X_k &= \Phi, \\ 0 \leq Y_k \leq Y_{k+1} \leq \Psi, \quad \lim_{k \rightarrow \infty} Y_k &= \Psi, \end{aligned}$$

for all parameters α and β satisfying

$$\alpha \geq \alpha_{\text{opt}} \stackrel{\text{def}}{=} \max_i A_{(i,i)}, \quad \beta \geq \beta_{\text{opt}} \stackrel{\text{def}}{=} \max_j B_{(j,j)}, \quad (3.9)$$

where Φ and Ψ are the minimal nonnegative solutions of MARE (1.1) and its complementary MARE (2.1), respectively. Moreover, under (3.9),

$$\limsup_{k \rightarrow \infty} \|\Phi - X_k\|^{1/2^k}, \quad \limsup_{k \rightarrow \infty} \|\Psi - Y_k\|^{1/2^k} \leq \rho(\mathcal{C}(S; \alpha, \beta)) \cdot \rho(\mathcal{C}(R; \beta, \alpha)), \quad (3.10)$$

where $\|\cdot\|$ is any matrix norm, and

$$\rho(\mathcal{C}(S; \alpha, \beta)) \cdot \rho(\mathcal{C}(R; \beta, \alpha)) \begin{cases} < 1, & \text{if } W \text{ is nonsingular or singular with } \mu \neq 0, \\ \equiv 1, & \text{if } W \text{ is singular with } \mu = 0, \end{cases} \quad (3.11)$$

The optimal α and β that minimize $\rho(\mathcal{C}(S; \alpha, \beta)) \cdot \rho(\mathcal{C}(R; \beta, \alpha))$, subject to (3.9), are $\alpha = \alpha_{\text{opt}}$ and $\beta = \beta_{\text{opt}}$.

For the ease of future reference, we summarize ADDA as follows.

Algorithm 3.1.

ADDA for ARE $XDX - AX - XB + C = 0$ **and,**
as a by-product, for cARE $YCY - YA - BY + D = 0$.

- 1 Pick α and β ;
- 2 $A_\beta \stackrel{\text{def}}{=} A + \beta I$, $B_\alpha \stackrel{\text{def}}{=} B + \alpha I$;
- 3 Compute A_β^{-1} and B_α^{-1} ;
- 4 Compute $V_{\alpha\beta}$ and $U_{\alpha\beta}$ as in (3.3) and then their inverses;
- 5 Compute E_0 , F_0 , X_0 and Y_0 by (3.4);
- 6 Compute $(I - X_0 Y_0)^{-1}$ and $(I - Y_0 X_0)^{-1}$;
- 7 Compute X_1 and Y_1 by (3.5c) and (3.5d);
- 8 For $k = 1, 2, \dots$, until convergence
- 9 Compute E_k and F_k by (3.5a) and (3.5b) (after substituting $k + 1$ for k);
- 10 Compute $(I - X_k Y_k)^{-1}$ and $(I - Y_k X_k)^{-1}$;
- 11 Compute X_{k+1} and Y_{k+1} by (3.5c) and (3.5d);
- 12 Enddo

4 Deflating an Irreducible Singular MARE

Assume that (1.6) holds. We have three cases: $\mu = u^\top x - v^\top y > 0$, $\mu = 0$, and $\mu < 0$. The case $\mu < 0$ can be converted to the case $\mu > 0$ by transposing (1.1) to get

$$ZD^\top Z - ZA^\top - B^\top Z + C^\top = 0, \quad (4.1)$$

where $Z = X^\top$. This MARE has the unique minimal nonnegative solution Φ^\top , and

$$\begin{pmatrix} A^\top & -D^\top \\ -C^\top & B^\top \end{pmatrix} \begin{pmatrix} v \\ u \end{pmatrix} = 0, \quad \begin{pmatrix} y \\ x \end{pmatrix}^\top \begin{pmatrix} A^\top & -D^\top \\ -C^\top & B^\top \end{pmatrix} = 0$$

as the result of (1.3), and the new μ for (4.1) is positive. By Theorem 2.1, we have $\Phi^\top v = u$.

If $m = 1$ and $\mu \geq 0$, then $B - D\Phi = 0$ by Theorem 2.1(c). MARE (1.1) after setting $X = \Phi$ becomes $C - A\Phi = 0$ to give $\Phi = A^{-1}C$ because A is an nonsingular M -matrix.

In light of these considerations, without loss of generality, we assume from now on

$$\mu = u^\top x - v^\top y \geq 0, \quad m \geq 2. \quad (4.2)$$

By Theorem 2.1, $\Phi x = y$. In what follows, we will first present a general framework for deflating an irreducible singular MARE with (4.2), and then its convergence analysis. Two numerical realizations of the framework will be discussed in detail in Section 5.

4.1 General Framework

The framework starts with a nonsingular matrix $V \in \mathbb{R}^{(m+n) \times (m+n)}$ such that

$$V^{-1}z = \delta e_1, \quad z = \begin{pmatrix} x \\ y \end{pmatrix}. \quad (4.3)$$

Any numerical realization of this framework in section 5 is simply a way of constructing such a matrix V .

Φ satisfies MARE (1.1), or equivalently,

$$H \begin{pmatrix} I \\ \Phi \end{pmatrix} = \begin{pmatrix} I \\ \Phi \end{pmatrix} R, \quad R = B - D\Phi \quad (4.4)$$

which is equivalent to

$$V^{-1}HV V^{-1} \begin{pmatrix} I \\ \Phi \end{pmatrix} = V^{-1} \begin{pmatrix} I \\ \Phi \end{pmatrix} R. \quad (4.5)$$

Partition

$$V^{-1} = \begin{matrix} & m & n \\ m & \begin{pmatrix} U_{11} & U_{12} \\ U_{21} & U_{22} \end{pmatrix} & \end{matrix}. \quad (4.6)$$

Assuming that $(U_{11} + U_{12}\Phi)^{-1}$ exists, we have from (4.5)

$$\begin{aligned} V^{-1}HV V^{-1} \begin{pmatrix} I \\ \Phi \end{pmatrix} (U_{11} + U_{12}\Phi)^{-1} \\ = V^{-1} \begin{pmatrix} I \\ \Phi \end{pmatrix} (U_{11} + U_{12}\Phi)^{-1} \left[(U_{11} + U_{12}\Phi) R (U_{11} + U_{12}\Phi)^{-1} \right]. \end{aligned} \quad (4.7)$$

Since

$$V^{-1} \begin{pmatrix} I \\ \Phi \end{pmatrix} (U_{11} + U_{12}\Phi)^{-1} = \begin{pmatrix} I \\ (U_{21} + U_{22}\Phi) (U_{11} + U_{12}\Phi)^{-1} \end{pmatrix},$$

we rewrite (4.7) as

$$V^{-1}HV \begin{pmatrix} I \\ \tilde{\Phi} \end{pmatrix} = \begin{pmatrix} I \\ \tilde{\Phi} \end{pmatrix} \tilde{R}, \quad (4.8)$$

where

$$\tilde{\Phi} = (U_{21} + U_{22}\Phi) (U_{11} + U_{12}\Phi)^{-1}, \quad (4.9)$$

$$\tilde{R} = (U_{11} + U_{12}\Phi) R (U_{11} + U_{12}\Phi)^{-1}. \quad (4.10)$$

Lemma 4.1. *The first column of $V^{-1}HV$ is 0; so is that of $\tilde{\Phi}$.*

Proof. We have from (4.3) $V e_1 = \delta^{-1}z$. Thus $V^{-1}HV e_1 = \delta^{-1}V^{-1}Hz = 0$, i.e., the first column of $V^{-1}HV$ is 0. To show $\tilde{\Phi} e_1 = 0$, we notice

$$\delta e_1 = V^{-1}z = V^{-1} \begin{pmatrix} x \\ y \end{pmatrix} = V^{-1} \begin{pmatrix} x \\ \Phi x \end{pmatrix} = V^{-1} \begin{pmatrix} I \\ \Phi \end{pmatrix} x = \begin{pmatrix} (U_{11} + U_{12}\Phi) x \\ (U_{21} + U_{22}\Phi) x \end{pmatrix}$$

which gives

$$x = \delta (U_{11} + U_{12}\Phi)^{-1} e_1, \quad (U_{21} + U_{22}\Phi) x = 0. \quad (4.11)$$

Therefore $\delta \tilde{\Phi} e_1 = (U_{21} + U_{22}\Phi) x = 0$ yielding $\tilde{\Phi} e_1 = 0$, as claimed. \square

Keeping in mind Lemma 4.1, we define matrices \tilde{A} , \tilde{B} , \tilde{C} , \tilde{D} , and \hat{A} , \hat{B} , \hat{C} , \hat{D} by the following partitioning

$$V^{-1}HV = \begin{matrix} & m & n \\ & \tilde{B} & -\tilde{D} \\ n & \tilde{C} & -\tilde{A} \end{matrix} = \begin{matrix} & 1 & m-1 & n \\ 1 & 0 & b & -d \\ m-1 & 0 & \hat{B} & -\hat{D} \\ n & 0 & \hat{C} & -\hat{A} \end{matrix}. \quad (4.12)$$

In particular,

$$\tilde{A} = \hat{A}, \quad \tilde{B} = \begin{matrix} 1 & m-1 \\ 0 & b \\ m-1 & \hat{B} \end{matrix}, \quad \tilde{C} = \begin{matrix} 1 & m-1 \\ 0 & \hat{C} \end{matrix}, \quad \tilde{D} = \begin{matrix} 1 & n \\ m-1 & \hat{D} \end{matrix}. \quad (4.13)$$

Equation (4.8) says $\tilde{X} = \tilde{\Phi}$ satisfies the following ARE

$$\tilde{X}\tilde{D}\tilde{X} - \tilde{A}\tilde{X} - \tilde{X}\tilde{B} + \tilde{C} = 0. \quad (4.14)$$

This ARE may have many solutions, and $\tilde{X} = \tilde{\Phi}$ is just one of them. If this particular solution $\tilde{X} = \tilde{\Phi}$ is known, then the minimal nonnegative solution Φ of (1.1) can be recovered as follows:

$$\begin{aligned} (U_{21} + U_{22}\Phi)(U_{11} + U_{12}\Phi)^{-1} &= \tilde{\Phi}, \\ \Rightarrow U_{21} + U_{22}\Phi &= \tilde{\Phi}(U_{11} + U_{12}\Phi) \\ &= \tilde{\Phi}U_{11} + \tilde{\Phi}U_{12}\Phi, \\ \Rightarrow U_{21} - \tilde{\Phi}U_{11} &= (-U_{22} + \tilde{\Phi}U_{12})\Phi. \end{aligned}$$

Thus if $(-U_{22} + \tilde{\Phi}U_{12})^{-1}$ exists, then

$$\Phi = (-U_{22} + \tilde{\Phi}U_{12})^{-1}(U_{21} - \tilde{\Phi}U_{11}). \quad (4.15)$$

Lemma 4.1 allows us to write

$$\tilde{\Phi} = \begin{pmatrix} 0 & \hat{\Phi} \end{pmatrix}, \quad \hat{\Phi} = \tilde{\Phi}_{(:,2:m)}. \quad (4.16)$$

In what follows, we look for a determining ARE for $\hat{\Phi}$. To this end, we substitute $\tilde{\Phi} = \begin{pmatrix} 0 & \hat{\Phi} \end{pmatrix}$ and the expressions in (4.13) for \tilde{A} , \tilde{B} , \tilde{C} , \tilde{D} into (4.14) to get

$$\begin{aligned} &\begin{pmatrix} 0 & \hat{\Phi} \end{pmatrix} \begin{pmatrix} d \\ \hat{D} \end{pmatrix} \begin{pmatrix} 0 & \hat{\Phi} \end{pmatrix} - \tilde{A} \begin{pmatrix} 0 & \hat{\Phi} \end{pmatrix} - \begin{pmatrix} 0 & \hat{\Phi} \end{pmatrix} \begin{pmatrix} 0 & b \\ 0 & \hat{B} \end{pmatrix} + \begin{pmatrix} 0 & \hat{C} \end{pmatrix} = 0 \\ \Leftrightarrow &\begin{pmatrix} 0 & \hat{\Phi}\hat{D}\hat{\Phi} \end{pmatrix} - \begin{pmatrix} 0 & \tilde{A}\hat{\Phi} \end{pmatrix} - \begin{pmatrix} 0 & \hat{\Phi}\hat{B} \end{pmatrix} + \begin{pmatrix} 0 & \hat{C} \end{pmatrix} = 0 \\ \Leftrightarrow &\hat{\Phi}\hat{D}\hat{\Phi} - \hat{A}\hat{\Phi} - \hat{\Phi}\hat{B} + \hat{C} = 0. \end{aligned}$$

This says that $\hat{X} = \hat{\Phi}$ is a solution of the following ARE:

$$\hat{X}\hat{D}\hat{X} - \hat{A}\hat{X} - \hat{X}\hat{B} + \hat{C} = 0 \quad (4.17)$$

which is equivalent to

$$\widehat{H} \begin{pmatrix} I_{m-1} \\ \widehat{X} \end{pmatrix} = \begin{pmatrix} I_{m-1} \\ \widehat{X} \end{pmatrix} (\widehat{B} - \widehat{D}\widehat{X}), \quad \widehat{H} = \begin{matrix} m-1 & n \\ \widehat{B} & -\widehat{D} \\ \widehat{C} & -\widehat{A} \end{matrix}. \quad (4.18)$$

The *complementary algebraic Riccati equation* (cARE) of (4.17) is

$$\widehat{Y}\widehat{C}\widehat{Y} - \widehat{Y}\widehat{A} - \widehat{B}\widehat{Y} + \widehat{D} = 0, \quad (4.19)$$

or equivalently

$$\widehat{H} \begin{pmatrix} \widehat{Y} \\ I \end{pmatrix} = \begin{pmatrix} \widehat{Y} \\ I \end{pmatrix} [-(\widehat{A} - \widehat{C}\widehat{Y})].$$

In the above deflation framework, we assume that both

$$U_{11} + U_{12}\Phi, \quad -U_{22} + \widetilde{\Phi}U_{12}$$

are invertible. Later in section 5. This assumption will be verified for the two realizations of this framework there.

Theorem 4.1. *Assume (1.6) and (4.2). Suppose assume $U_{11} + U_{12}\Phi$ is nonsingular, and define $\widehat{\Phi}$ as in (4.16). Then*

$$\text{eig}(\widehat{H}) = \{\lambda_1, \dots, \lambda_{m-1}, \lambda_{m+1}, \dots, \lambda_{m+n}\}, \quad (4.20)$$

$$\text{eig}(\widehat{B} - \widehat{D}\widehat{\Phi}) = \{\lambda_1, \dots, \lambda_{m-1}\}, \quad (4.21)$$

and cARE (4.19) has a unique solution $\widehat{\Psi}$, if exists, satisfying

$$\text{eig}(\widehat{A} - \widehat{C}\widehat{\Psi}) = \{-\lambda_{m+1}, \dots, -\lambda_{m+n}\}, \quad (4.22)$$

where λ_i ($i = 1, \dots, m+n$) are H 's eigenvalues as specified in Theorem 2.1.

Proof. Equation (4.20) is a consequence of Theorem 2.1, the preceding reduction that leads to the definition of \widehat{H} in (4.18), and (4.12).

We have (4.8) – (4.10). Since $Rx = (B - D\Phi)x = 0$ by Theorem 2.1, using (4.11) we find

$$\widetilde{R}e_1 = (U_{11} + U_{12}\Phi)R(U_{11} + U_{12}\Phi)^{-1}e_1 = \delta^{-1}(U_{11} + U_{12}\Phi)Rx = 0$$

and thus the partitioning

$$\widetilde{R} = \widetilde{B} - \widetilde{D}\widetilde{\Phi} = \begin{matrix} 1 & m-1 \\ 0 & \widetilde{R}_{12} \\ m-1 & 0 \\ & \widetilde{R}_{22} \end{matrix} \quad (4.23)$$

which together with (4.13) and $\widetilde{\Phi} = \begin{pmatrix} 0 & \widehat{\Phi} \end{pmatrix}$ give $\widetilde{R}_{22} = \widehat{B} - \widehat{D}\widehat{\Phi}$. Since

$$\text{eig}(\widetilde{R}) = \text{eig}(R) = \{\lambda_1, \dots, \lambda_m\}$$

and $0 = \lambda_m < \text{Re}\lambda_{m-1} \leq \dots \leq \text{Re}\lambda_1$ by Theorem 2.1, we have (4.21).

Let $Z \in \mathbb{R}^{(m+n-1) \times n}$ be a basis matrix of \widehat{H} 's invariant subspace associated with the eigenvalues $\lambda_{m+1}, \dots, \lambda_{m+n}$. If $Z_{(m:m+n-1,:)}$ is invertible, then $\widehat{\Psi}$ exists and is unique, and moreover $\widehat{\Psi} = Z_{(1:m-1,:)}[Z_{(m:m+n-1,:)}]^{-1}$ and (4.22) holds [17]. \square

Theorem 4.2. *Assume (1.6) and (4.2). Suppose both $U_{11} + U_{12}\Phi$ and $-U_{22} + \tilde{\Phi}U_{12}$ are nonsingular. Then ARE (4.17) constructed as above has a particular solution $\hat{X} = \hat{\Phi}$ characterized uniquely by (4.21), and the minimal nonnegative solution Φ can be recovered by (4.15) with $\tilde{\Phi} = \begin{pmatrix} 0 & \hat{\Phi} \end{pmatrix}$.*

Proof. The existence of $\hat{\Phi}$ is a consequence of the constructive deflation procedure above, and $\hat{\Phi}$ satisfies (4.21) by Theorem 4.1. That this particular solution $\hat{X} = \hat{\Phi}$ is uniquely characterized by (4.21) follows from the relation between the solutions of ARE (4.17) and the invariant subspaces of \hat{H} [17]. \square

Theorem 4.2 suggests a natural way to compute Φ by first solving ARE (4.17) for $\hat{\Phi}$ by Algorithm 3.1 and then recovering Φ by (4.15). This leads to the following *deflated Alternating-Directional Doubling Algorithm* (dADDA).

Algorithm 4.1.

dADDA for MARE $XDX - AX - XB + C = 0$ **with** (1.6).

- 1 Compute $\mu = u^T x - v^T y$;
- 2 If $\mu \geq 0$, then
 - 3 compute \hat{A} , \hat{B} , \hat{C} , and \hat{D} as defined by (4.12) and (4.13);
 - 4 solve (4.17) by Algorithm 3.1 for $\hat{\Phi}$;
 - 5 recover Φ by (4.15) with $\tilde{\Phi} = \begin{pmatrix} 0 & \hat{\Phi} \end{pmatrix}$;
- 6 else
 - 7 compute Φ^T instead by working with (4.1);
- 8 Enddo

REMARK 4.1. There are a few practically important issues to resolve for this dADDA.

1. In building ARE (4.17), we need $U_{11} + U_{12}\Phi$ to be nonsingular, and in recovering Φ by (4.15), we need $-U_{22} + \tilde{\Phi}U_{12}$ to be nonsingular. These requirements are satisfied for each of the realizations in section 5.
2. Both (4.21) and (4.22) uniquely characterize the particular solution $\hat{\Phi}$ of (4.17) and the particular solution $\hat{\Psi}$, if exists, of (4.19), respectively. Specifically, $\hat{\Phi}$ is the unique solution of (4.17) such that all eigenvalues of $\hat{B} - \hat{D}\hat{\Phi}$ have positive real parts and $\hat{\Psi}$ is the unique solution of (4.19) such that all eigenvalues of $\hat{A} - \hat{C}\hat{\Psi}$ have nonpositive real parts. These characterizations in principle can be used to verify that the computed solution of (4.17) at Line 4 of Algorithm 4.1 is the right one. But such a verification can only be performed at the end of the iterative process. In the next subsection we will show that with a proper restriction on α and β , this kind of verification becomes unnecessary, i.e., Line 4 of Algorithm 4.1 will always produces the right $\hat{\Phi}$.
3. What should α and β be for fast convergence at Line 4 of Algorithm 4.1?

REMARK 4.2. So far, the existence of $\hat{\Psi}$ is assumed, not proven. If it exists, it is uniquely characterized by (4.22). One way to look into this existence issue, naturally, is to relate $\hat{\Psi}$ to the minimal nonnegative solution Ψ of the original cMARE (2.1). We shall do it now. Ψ satisfies cMARE (2.1), or equivalently,

$$H \begin{pmatrix} \Psi \\ I \end{pmatrix} = \begin{pmatrix} \Psi \\ I \end{pmatrix} (-S), \quad S = A - C\Psi. \quad (4.24)$$

In the same way as we gotten (4.8), we can get

$$V^{-1}HV \begin{pmatrix} \tilde{\Psi} \\ I \end{pmatrix} = \begin{pmatrix} \tilde{\Psi} \\ I \end{pmatrix} (-\tilde{S}), \quad \tilde{S} = \tilde{A} - \tilde{C}\tilde{\Psi}, \quad (4.25)$$

where

$$\tilde{\Psi} = (U_{11}\Psi + U_{12})(U_{21}\Psi + U_{22})^{-1}, \quad (4.26)$$

$$\tilde{S} = (U_{21}\Psi + U_{22})S(U_{21}\Psi + U_{22})^{-1}, \quad (4.27)$$

assuming $(U_{21}\Psi + U_{22})^{-1}$ exists. Equation (4.25) says $\tilde{Y} = \tilde{\Psi}$ satisfies the following ARE

$$\tilde{Y}\tilde{C}\tilde{Y} - \tilde{Y}\tilde{A} - \tilde{B}\tilde{Y} + \tilde{D} = 0 \quad (4.28)$$

which is the complementary ARE of (4.14). Partition

$$\tilde{\Psi} = \begin{matrix} 1 \\ m-1 \end{matrix} \begin{pmatrix} \psi \\ \hat{\Psi} \end{pmatrix} \quad (4.29)$$

and substitute this and (4.13) into (4.28) to get

$$\begin{aligned} & \begin{pmatrix} \psi \\ \hat{\Psi} \end{pmatrix} \begin{pmatrix} 0 & \hat{C} \end{pmatrix} \begin{pmatrix} \psi \\ \hat{\Psi} \end{pmatrix} - \begin{pmatrix} \psi \\ \hat{\Psi} \end{pmatrix} \hat{A} - \begin{pmatrix} 0 & b \\ 0 & \hat{B} \end{pmatrix} \begin{pmatrix} \psi \\ \hat{\Psi} \end{pmatrix} + \begin{pmatrix} d \\ \hat{D} \end{pmatrix} = 0 \\ \Leftrightarrow & \begin{pmatrix} \psi \hat{C} \hat{\Psi} \\ \hat{\Psi} \hat{C} \hat{\Psi} \end{pmatrix} - \begin{pmatrix} \psi \hat{A} \\ \hat{\Psi} \hat{A} \end{pmatrix} - \begin{pmatrix} b \hat{\Psi} \\ \hat{B} \hat{\Psi} \end{pmatrix} + \begin{pmatrix} d \\ \hat{D} \end{pmatrix} = 0 \\ \Leftrightarrow & \begin{cases} \psi(\hat{C}\hat{\Psi} - \hat{A}) - b\hat{\Psi} + d = 0, \\ \hat{\Psi}\hat{C}\hat{\Psi} - \hat{\Psi}\hat{A} - \hat{B}\hat{\Psi} + \hat{D} = 0. \end{cases} \end{aligned}$$

This says that $\hat{Y} = \hat{\Psi}$ is a solution of the complementary ARE (4.19) and ψ satisfies $\psi(\hat{A} - \hat{C}\hat{\Psi}) = -b\hat{\Psi} + d$. Thus $\hat{\Psi}$ exists, provided $U_{21}\Psi + U_{22}$ is nonsingular. Later we will show that if $\mu \neq 0$, then $U_{21}\Psi + U_{22}$ is nonsingular for the two realizations in section 5. Unfortunately it is always singular in the critical case as confirmed by the following lemma. But we emphasize that that $U_{21}\Psi + U_{22}$ is nonsingular is just a sufficient condition, not a necessary one, i.e., $\hat{\Psi}$ may still exist even if $U_{21}\Psi + U_{22}$ is singular. For example, $\hat{\Psi}$ still exists in all the critical case examples in section 7. \diamond

Lemma 4.2. *If $\mu = 0$, then $(U_{21}\Psi + U_{22})y = 0$ and thus $U_{21}\Psi + U_{22}$ is always singular in the critical case.*

Proof. In the critical case $\mu = 0$, $\Psi y = x$ by Theorem 2.1. Therefore

$$\delta e_1 = V^{-1}z = V^{-1} \begin{pmatrix} x \\ y \end{pmatrix} = V^{-1} \begin{pmatrix} \Psi y \\ y \end{pmatrix} = \begin{pmatrix} (U_{11}\Psi + U_{12})y \\ (U_{21}\Psi + U_{22})y \end{pmatrix}$$

which implies $(U_{21}\Psi + U_{22})y = 0$. \square

4.2 Convergence Analysis

Assume, as in ADDA for the original MARE (1.1), that

$$\alpha \geq \alpha_{\text{opt}} \stackrel{\text{def}}{=} \max_i A_{(i,i)}, \quad \beta \geq \beta_{\text{opt}} \stackrel{\text{def}}{=} \max_j B_{(j,j)}. \quad (3.9)$$

By Theorem 4.1, $\widehat{X} = \widehat{\Phi} = \widetilde{\Phi}_{(:,2:m)}$ and $\widehat{\Psi}$ are such that

$$\widehat{H} \begin{pmatrix} I \\ \widehat{\Phi} \end{pmatrix} = \begin{pmatrix} I \\ \widehat{\Phi} \end{pmatrix} \widehat{R}, \quad \widehat{R} = \widehat{B} - \widehat{D}\widehat{\Phi}, \quad \text{eig}(\widehat{R}) = \{\lambda_1, \dots, \lambda_{m-1}\}, \quad (4.30a)$$

$$\widehat{H} \begin{pmatrix} \widehat{\Psi} \\ I \end{pmatrix} = \begin{pmatrix} \widehat{\Psi} \\ I \end{pmatrix} (-\widehat{S}), \quad \widehat{S} = \widehat{A} - \widehat{C}\widehat{\Psi}, \quad \text{eig}(\widehat{S}) = \{-\lambda_{m+1}, \dots, -\lambda_{m+n}\}. \quad (4.30b)$$

Lemma 4.3. *Assume (1.6) and (3.9). Let $R = B - D\Phi$ and $S = A - C\Psi$, and \widehat{R} and \widehat{S} as given by (4.30). Then*

$$\rho(\mathcal{C}(\widehat{S}; \alpha, \beta)) = \rho(\mathcal{C}(S; \alpha, \beta)), \quad \rho(\mathcal{C}(\widehat{R}; \beta, \alpha)) < \rho(\mathcal{C}(R; \beta, \alpha)) \quad (4.31)$$

and in particular

$$\rho(\mathcal{C}(\widehat{S}; \alpha, \beta)) \cdot \rho(\mathcal{C}(\widehat{R}; \beta, \alpha)) < \rho(\mathcal{C}(S; \alpha, \beta)) \cdot \rho(\mathcal{C}(R; \beta, \alpha)) \leq 1. \quad (4.32)$$

Proof. By Theorem 2.1(b), both R and S are irreducible M -matrices. Since by (3.9)

$$\alpha \geq \max_i A_{(i,i)} \geq \max_i S_{(i,i)}, \quad \beta \geq \max_j B_{(j,j)} \geq \max_j R_{(j,j)},$$

we have $\rho(\mathcal{C}(S; \alpha, \beta)) \cdot \rho(\mathcal{C}(R; \beta, \alpha)) \leq 1$ by Theorem 3.1. This is the second inequality in (4.32). The first inequality is a consequence of (4.31) which we now prove. It follows from Theorem 2.1(d) and Theorem 4.1 that

$$\text{eig}(\widehat{R}) \subset \text{eig}(R), \quad 0 \in \text{eig}(R), \quad 0 \notin \text{eig}(\widehat{R}), \quad \text{and} \quad \text{eig}(\widehat{S}) = \text{eig}(S).$$

Thus $\rho(\mathcal{C}(\widehat{S}; \alpha, \beta)) = \rho(\mathcal{C}(S; \alpha, \beta))$. The proof of [20, Theorem 2.1] implies that

$$\rho(\mathcal{C}(R; \beta, \alpha)) = [\beta - \lambda_{\min}(R)][\lambda_{\min}(R) + \alpha]^{-1},$$

where $\lambda_{\min}(R) = 0$ is the eigenvalue of R with the smallest absolute value among all eigenvalues of R . Since $-\mathcal{C}(R; \beta, \alpha) = -(\beta I - R)(\alpha I + R)^{-1} > 0$, by the Perron-Frobenius theorem [3, p.27], we know $\rho(\mathcal{C}(R; \beta, \alpha))$ is a simple eigenvalue with the greatest magnitude among all eigenvalues of $-\mathcal{C}(R; \beta, \alpha)$, i.e., $\rho(\mathcal{C}(R; \beta, \alpha))$ is strictly larger than the absolute value of any other eigenvalue of $-\mathcal{C}(R; \beta, \alpha)$. Since $\lambda_{\min}(R) = 0 \notin \text{eig}(\widehat{R}) \subset \text{eig}(R)$, the eigenvalues of $-\mathcal{C}(\widehat{R}; \beta, \alpha)$ are precisely those of $-\mathcal{C}(R; \beta, \alpha)$, except $\rho(\mathcal{C}(R; \beta, \alpha))$. Thus $\rho(\mathcal{C}(R; \beta, \alpha))$ is bigger than the absolute value of any eigenvalue of $-\mathcal{C}(\widehat{R}; \beta, \alpha)$. Therefore

$$\rho(\mathcal{C}(\widehat{R}; \beta, \alpha)) < \rho(\mathcal{C}(R; \beta, \alpha)),$$

as was to be shown. \square

Theorem 4.3. *Assume (1.6) and (4.2). Suppose $U_{11} + U_{12}\Phi$ is nonsingular. Let $\{\widehat{E}_k\}$, $\{\widehat{F}_k\}$, $\{\widehat{X}_k\}$, $\{\widehat{Y}_k\}$ be the sequences generated by ADDA applied to (4.17) with no breakdowns, i.e., all involved inverses exist. If (3.9) holds, then \widehat{X}_k and \widehat{Y}_k converge quadratically to $\widehat{\Phi}$ and $\widehat{\Psi}$, respectively, and*

$$\limsup_{k \rightarrow \infty} \|\widehat{\Phi} - \widehat{X}_k\|^{1/2^k} \leq \rho(\mathcal{C}(\widehat{S}; \alpha, \beta)) \cdot \rho(\mathcal{C}(\widehat{R}; \beta, \alpha)) < 1, \quad (4.33a)$$

$$\limsup_{k \rightarrow \infty} \|\widehat{\Psi} - \widehat{Y}_k\|^{1/2^k} \leq \rho(\mathcal{C}(\widehat{R}; \beta, \alpha)) \cdot \rho(\mathcal{C}(\widehat{S}; \alpha, \beta)) < 1, \quad (4.33b)$$

where $\|\cdot\|$ is any matrix norm.

Proof. Inequalities in (4.33) are the consequences of

$$\widehat{\Phi} - \widehat{X}_k = (I - \widehat{X}_k \widehat{\Psi}) \left[\mathcal{C}(\widehat{S}; \alpha, \beta) \right]^{2^k} \widehat{\Phi} \left[\mathcal{C}(\widehat{R}; \beta, \alpha) \right]^{2^k}, \quad (4.34a)$$

$$\widehat{\Psi} - \widehat{Y}_k = (I - \widehat{Y}_k \widehat{\Phi}) \left[\mathcal{C}(\widehat{R}; \beta, \alpha) \right]^{2^k} \widehat{\Psi} \left[\mathcal{C}(\widehat{S}; \alpha, \beta) \right]^{2^k}. \quad (4.34b)$$

Take (4.33a) for example. We have by (4.34a)

$$\begin{aligned} (\widehat{\Phi} - \widehat{X}_k) \left(I - \widehat{\Psi} \left[\mathcal{C}(\widehat{S}; \alpha, \beta) \right]^{2^k} \widehat{\Phi} \left[\mathcal{C}(\widehat{R}; \beta, \alpha) \right]^{2^k} \right) \\ = (I - \widehat{\Phi} \widehat{\Psi}) \left[\mathcal{C}(\widehat{S}; \alpha, \beta) \right]^{2^k} \widehat{\Phi} \left[\mathcal{C}(\widehat{R}; \beta, \alpha) \right]^{2^k}. \end{aligned} \quad (4.35)$$

Since by Lemma 4.3

$$\left\| \left[\mathcal{C}(\widehat{S}; \alpha, \beta) \right]^{2^k} \right\|^{1/2^k} \left\| \left[\mathcal{C}(\widehat{R}; \beta, \alpha) \right]^{2^k} \right\|^{1/2^k} \rightarrow \rho(\mathcal{C}(\widehat{S}; \alpha, \beta)) \cdot \rho(\mathcal{C}(\widehat{R}; \beta, \alpha)) < 1,$$

$\Gamma \stackrel{\text{def}}{=} \widehat{\Psi} \left[\mathcal{C}(\widehat{S}; \alpha, \beta) \right]^{2^k} \widehat{\Phi} \left[\mathcal{C}(\widehat{R}; \beta, \alpha) \right]^{2^k} \rightarrow 0$ as $k \rightarrow \infty$. Therefore for sufficiently large k , $(I - \Gamma)^{-1}$ exists and⁴

$$\begin{aligned} \|\widehat{\Phi} - \widehat{X}_k\|^{1/2^k} &\leq \|(I - \Gamma)^{-1}\|^{1/2^k} \|I - \widehat{\Phi} \widehat{\Psi}\|^{1/2^k} \\ &\quad \times \left\| \left[\mathcal{C}(\widehat{S}; \alpha, \beta) \right]^{2^k} \right\|^{1/2^k} \|\widehat{\Phi}\|^{1/2^k} \left\| \left[\mathcal{C}(\widehat{R}; \beta, \alpha) \right]^{2^k} \right\|^{1/2^k}. \end{aligned} \quad (4.36)$$

Letting $k \rightarrow \infty$ in both sides of (4.36) leads to (4.33a) because as $k \rightarrow \infty$,

$$\begin{aligned} \|(I - \Gamma)^{-1}\|^{1/2^k} &\rightarrow 1, \quad \|I - \widehat{\Phi} \widehat{\Psi}\|^{1/2^k} \rightarrow 1, \quad \|\widehat{\Phi}\|^{1/2^k} \rightarrow 1, \\ \left\| \left[\mathcal{C}(\widehat{S}; \alpha, \beta) \right]^{2^k} \right\|^{1/2^k} &\rightarrow \rho(\mathcal{C}(\widehat{S}; \alpha, \beta)), \quad \left\| \left[\mathcal{C}(\widehat{R}; \beta, \alpha) \right]^{2^k} \right\|^{1/2^k} \rightarrow \rho(\mathcal{C}(\widehat{R}; \beta, \alpha)). \end{aligned}$$

That \widehat{X}_k and \widehat{Y}_k converge quadratically to $\widehat{\Phi}$ and $\widehat{\Psi}$, respectively, is a consequence of the inequalities in (4.33). \square

⁴We assume $\|\cdot\|$ is a consistent matrix norm. This does not lose any generality because all matrix norms are equivalent and thus $\limsup_{k \rightarrow \infty} \|\widehat{\Phi} - \widehat{X}_k\|^{1/2^k}$ does not change with the norm used.

REMARK 4.3. A few comments are in order:

1. If $\mu \neq 0$, ADDA applied to the original MARE (1.1) is already quadratically convergent [20]. But it is only linearly convergent if $\mu = 0$ [5]. Theorem 4.3 says that ADDA applied to the deflated ARE (4.17) is still quadratically convergent.
2. ADDA applied to the original MARE (1.1) generates monotonic sequences, under (3.9). But this monotonicity property is generally lost in the sequences $\{\widehat{X}_k\}$ and $\{\widehat{Y}_k\}$ generated by ADDA applied to (4.17).
3. Theorem 3.1 says that under (3.9) $\rho(\mathcal{C}(S; \alpha, \beta)) \cdot \rho(\mathcal{C}(R; \beta, \alpha))$ is minimized at $\alpha = \alpha_{\text{opt}}$ and $\beta = \beta_{\text{opt}}$, leading to the optimal ADDA in [20]. For the current case, for fast convergence we should pick α and β such that $\rho(\mathcal{C}(\widehat{S}; \alpha, \beta)) \cdot \rho(\mathcal{C}(\widehat{R}; \beta, \alpha))$ is minimized subject to (3.9). While it is not clear whether $\rho(\mathcal{C}(\widehat{S}; \alpha, \beta)) \cdot \rho(\mathcal{C}(\widehat{R}; \beta, \alpha))$ is also minimized at $\alpha = \alpha_{\text{opt}}$ and $\beta = \beta_{\text{opt}}$, intuitively selecting $\alpha = \alpha_{\text{opt}}$ and $\beta = \beta_{\text{opt}}$ should be good. This is what we will do in our numerical tests in section 7. \diamond

5 Realizations

Two numerical realizations of the deflating framework given in Subsection 4.1 will be discussed in detail. Assume, throughout this section, (1.6) and (4.2).

5.1 By Elimination

Given an integer i_0 ($1 \leq i_0 \leq m+n$), set

$$P^T = (e_{i_0}, e_2, \dots, e_{i_0-1}, e_1, e_{i_0+1}, \dots, e_{m+n}) \in \mathbb{R}^{(m+n) \times (m+n)}, \quad (5.1)$$

a permutation matrix. Pz swaps $z_{(1)}$ and $z_{(i_0)}$ and serves as a pivoting strategy (or without one when $i_0 = 1$), where z is given as in (4.3). Set

$$L^{-1} = \begin{pmatrix} 1 & & \\ & -\hat{z} & \\ & & I_{m+n-1} \end{pmatrix}, \quad L = \begin{pmatrix} 1 & & \\ \hat{z} & & \\ & & I_{m+n-1} \end{pmatrix}, \quad (5.2a)$$

$$V^{-1} = L^{-1}P, \quad V = P^T L, \quad (5.2b)$$

where

$$\hat{z}^T = z_{(i_0)}^{-1} (z_{(2)}, \dots, z_{(i_0-1)}, z_{(1)}, z_{(i_0+1)}, \dots, z_{(m+n)}).$$

Then $V^{-1}z = z_{(i_0)}e_1$. We just mentioned that Pz serves as a pivoting strategy. We call it a *complete pivoting* if $i_0 = \operatorname{argmax}_i z_{(i)}$, and a *partial pivoting* if $i_0 = \operatorname{argmax}_{1 \leq i \leq m} z_{(i)}$. Simply setting $i_0 = 1$ corresponds to no pivoting. For the complete pivoting, $\|V\|_1 \|V^{-1}\|_1 \leq (m+n)^2$; but otherwise $\|V\|_1 \|V^{-1}\|_1$ can be very large if $z_{(i_0)}$ is tiny relative to some other entries of z . The involved formulas can be substantially complicated when $i_0 > m$, but are much simpler when $i_0 \leq m$, especially so when $i_0 = 1$. In all of our examples in section 7, $z = \mathbf{1}_{m+n}$ and thus it makes no difference with or without a pivoting strategy for them.

We can write

$$P^T = P = I - ww^T, \quad w = e_1 - e_{i_0}. \quad (5.3)$$

Partition

$$L^{-1} = \begin{matrix} & m & n \\ m & \begin{pmatrix} L_{11} & 0 \\ L_{21} & I \end{pmatrix} \\ n & \end{matrix}, \quad w = \begin{matrix} m \\ n \end{matrix} \begin{pmatrix} w_1 \\ w_2 \end{pmatrix}, \quad P = \begin{matrix} & m & n \\ m & \begin{pmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{pmatrix} \\ n & \end{matrix}.$$

Use (5.2a) and (5.3) to see

$$L_{11} = \begin{pmatrix} 1 & \\ -\hat{z}_{(1:m-1)} & I_{m-1} \end{pmatrix}, \quad L_{21} = -\hat{z}_{(m::m+n-1)} e_1^T, \quad (5.4a)$$

$$L = \begin{matrix} & m & n \\ m & \begin{pmatrix} L_{11}^{-1} & 0 \\ -L_{21} & I \end{pmatrix} \\ n & \end{matrix}, \quad L_{11}^{-1} = \begin{pmatrix} 1 & \\ \hat{z}_{(1:m-1)} & I_{m-1} \end{pmatrix}, \quad (5.4b)$$

$$P_{ii} = I - w_i w_i^T, \quad P_{ij} = -w_i w_j^T \text{ for } i \neq j. \quad (5.4c)$$

So the four submatrices U_{ij} of $V^{-1} = L^{-1}P$ partitioned as in (4.6) are

$$U_{11} = L_{11}(I - w_1 w_1^T), \quad U_{12} = -L_{11} w_1 w_2^T, \quad (5.5a)$$

$$U_{21} = L_{21}(I - w_1 w_1^T) - w_2 w_1^T, \quad U_{22} = -L_{21} w_1 w_2^T + I - w_2 w_2^T. \quad (5.5b)$$

Equations (4.9) and (4.15) that relate Φ and $\tilde{\Phi}$ remain valid, provided that $U_{11} + U_{12}\Phi$ and $-U_{22} + \tilde{\Phi}U_{12}$ are invertible, as ensured by Theorem 5.1 below.

Lemma 5.1 (Sherman-Morrison-Woodbury). *Let $E, F \in \mathbb{R}^{p \times q}$. The matrix $I_p - EF^T$ is invertible if and only if $I_q - F^T E$ is nonsingular. Moreover*

$$(I_p - EF^T)^{-1} = I_p + E(I_q - F^T E)^{-1} F^T.$$

Theorem 5.1. *Let U_{ij} be defined by (5.1) – (5.5). Then both $U_{11} + U_{12}\Phi$ and $-U_{22} + \tilde{\Phi}U_{12}$ are invertible, where $\tilde{\Phi}$ relates to Φ by (4.9).*

Proof. We have by (5.5)

$$\begin{aligned} U_{11} + U_{12}\Phi &= L_{11}(I - w_1 w_1^T) - L_{11} w_1 w_2^T \Phi \\ &= L_{11} [I - w_1(w_1^T + w_2^T \Phi)]. \end{aligned}$$

Since L_{11} is invertible, $U_{11} + U_{12}\Phi$ is invertible if and only if $I - w_1(w_1^T + w_2^T \Phi)$ is. By Lemma 5.1, $I - w_1(w_1^T + w_2^T \Phi)$ is invertible if and only if

$$\zeta \stackrel{\text{def}}{=} 1 - (w_1^T + w_2^T \Phi) w_1 \neq 0.$$

There are three cases to consider:

1. If $i_0 = 1$, then $w_1 = 0$ and $w_2 = 0$ and thus $\zeta = 1 - (w_1^T + w_2^T \Phi) w_1 = 1 > 0$;
2. If $1 < i_0 \leq m$, then $w_1 = e_1 - e_{i_0}$ and $w_2 = 0$ and thus

$$\zeta = 1 - (w_1^T + w_2^T \Phi) w_1 = 1 - w_1^T w_1 = -1 < 0;$$

3. If $i_0 > m$, then $w_1 = e_1$ and $w_2 = -e_{i_0-m}$ and thus

$$\zeta = 1 - (w_1^T + w_2^T \Phi)w_1 = -w_2^T \Phi w_1 = \Phi_{(i_0-m,1)} > 0$$

since $\Phi > 0$ by Theorem 2.1.

Thus $U_{11} + U_{12}\Phi$ is invertible and moreover

$$\begin{aligned} (U_{11} + U_{12}\Phi)^{-1} &= [I + \zeta^{-1} w_1(w_1^T + w_2^T \Phi)] L_{11}^{-1} \\ &= \begin{cases} L_{11}^{-1}, & \text{for } i_0 = 1, \\ [I - w_1 w_1^T] L_{11}^{-1}, & \text{for } 1 < i_0 \leq m, \\ [I + \Phi_{(i_0-m,1)}^{-1} e_1(e_1^T - \Phi_{(i_0-m,:)})] L_{11}^{-1}, & \text{for } m < i_0. \end{cases} \end{aligned} \quad (5.6)$$

Getting to $-U_{22} + \tilde{\Phi}U_{12}$, we have

$$-U_{22} + \tilde{\Phi}U_{12} = L_{21}w_1w_2^T - I + w_2w_2^T - \tilde{\Phi}L_{11}w_1w_2^T, \quad (5.7)$$

$$U_{21} + U_{22}\Phi = L_{21}(I - w_1w_1^T) - w_2w_1^T + (-L_{21}w_1w_2^T + I - w_2w_2^T)\Phi, \quad (5.8)$$

$$(U_{11} + U_{12}\Phi)^{-1} L_{11}w_1w_2^T = [I + \zeta^{-1} w_1(w_1^T + w_2^T \Phi)] w_1w_2^T. \quad (5.9)$$

Again there are three cases to consider:

1. If $i_0 = 1$, then $w_1 = 0$ and $w_2 = 0$ and thus $-U_{22} + \tilde{\Phi}U_{12} = -I$;
2. If $1 < i_0 \leq m$, then $w_1 = e_1 - e_{i_0}$ and $w_2 = 0$ and thus also $-U_{22} + \tilde{\Phi}U_{12} = -I$;
3. If $i_0 > m$, then $w_1 = e_1$ and $w_2 = -e_{i_0-m}$ and thus

$$(U_{11} + U_{12}\Phi)^{-1} L_{11}w_1w_2^T = \zeta^{-1}w_1w_2^T. \quad (5.10)$$

Therefore by (5.8) and (5.10)

$$\begin{aligned} \tilde{\Phi}L_{11}w_1w_2^T &= (U_{21} + U_{22}\Phi)(U_{11} + U_{12}\Phi)^{-1} L_{11}w_1w_2^T \\ &= [L_{21}(I - w_1w_1^T) - w_2w_1^T + (-L_{21}w_1w_2^T + I - w_2w_2^T)\Phi] \zeta^{-1}w_1w_2^T \\ &= \zeta^{-1}[-w_2w_2^T + \zeta L_{21}w_1w_2^T + \Phi w_1w_2^T + \zeta w_2w_2^T] \\ &= (1 - \zeta^{-1})w_2w_2^T + L_{21}w_1w_2^T + \zeta^{-1}\Phi w_1w_2^T. \end{aligned}$$

Combine this with (5.7) to get

$$\begin{aligned} -U_{22} + \tilde{\Phi}U_{12} &= -I + \zeta^{-1}w_2w_2^T - \zeta^{-1}\Phi w_1w_2^T \\ &= -[I - \zeta^{-1}(w_2 - \Phi w_1)w_2^T] \end{aligned} \quad (5.11)$$

which, by Lemma 5.1, is invertible if

$$1 - \zeta^{-1}w_2^T(w_2 - \Phi w_1) = -\zeta^{-1} = -\Phi_{(i_0-m,1)}^{-1} \neq 0.$$

Thus $-U_{22} + \tilde{\Phi}U_{12}$ is invertible, too, and moreover

$$\left(-U_{22} + \tilde{\Phi}U_{12}\right)^{-1} = \begin{cases} -I, & \text{for } i_0 \leq m, \\ -[I - (e_{i_0-m} + \Phi_{(:,1)})e_{i_0-m}^T], & \text{for } i_0 > m. \end{cases} \quad (5.12)$$

This completes the proof. \square

Rewrite (5.7) as

$$-U_{22} + \tilde{\Phi}U_{12} = - \left[I - (w_2 + L_{21}w_1 - \tilde{\Phi}L_{11}w_1)w_2^T \right]$$

to get

$$\left(-U_{22} + \tilde{\Phi}U_{12} \right)^{-1} = - \left[I + \frac{(w_2 + L_{21}w_1 - \tilde{\Phi}L_{11}w_1)w_2^T}{1 - w_2^T(w_2 + L_{21}w_1 - \tilde{\Phi}L_{11}w_1)} \right]. \quad (5.13)$$

The inversion formulas (5.6) and (5.13), together with (5.4) and (5.5), lead to fast algorithms to recover one of Φ and $\tilde{\Phi}$ from the other.

It is rather straightforward to extract \hat{A} , \hat{B} , \hat{C} , and \hat{D} from

$$\begin{aligned} V^{-1}HV &= (I - \tilde{z}e_1^T)PHP(I + \tilde{z}e_1^T) \\ &= PHP - \tilde{z}(e_1^T PHP) + (PHP\tilde{z})e_1^T - (e_1^T PHP\tilde{z})\tilde{z}e_1^T. \end{aligned} \quad (5.14)$$

where $\tilde{z} = (0, \hat{z}^T)^T$. The right-hand side of (5.14) lends itself for a fast evaluation of $V^{-1}HV$. In the case $i_0 = 1$, we have⁵

$$\tilde{\Phi} = \begin{pmatrix} 0 & \Phi_{(:,2:m)} \end{pmatrix}, \quad \hat{\Phi} = \Phi_{(:,2:m)}, \quad \Phi_{(:,1)} = x_{(1)}^{-1} [y - \Phi_{(:,2:m)}x_{(2:m)}], \quad (5.15)$$

and

$$\hat{B} = B_{(2:m,2:m)} - x_{(1)}^{-1} x_{(2:m)} B_{(1,2:m)}, \quad \hat{D} = D_{(2:m,:)} - x_{(1)}^{-1} x_{(2:m)} D_{(1,:)}, \quad (5.16a)$$

$$\hat{C} = C_{(:,2:m)} - x_{(1)}^{-1} y B_{(1,2:m)}, \quad \hat{A} = A - x_{(1)}^{-1} y D_{(1,:)}. \quad (5.16b)$$

Note also in this case

$$\hat{A} - \hat{\Phi}\hat{D} = A - \Phi D. \quad (5.17)$$

This is because $\Phi_{(:,1)} = x_{(1)}^{-1} [y - \hat{\Phi}x_{(2:m)}]$, and thus

$$\begin{aligned} \hat{A} - \hat{\Phi}\hat{D} &= A - x_{(1)}^{-1} y D_{(1,:)} - \hat{\Phi}(D_{(2:m,:)} - x_{(1)}^{-1} x_{(2:m)} D_{(1,:)}) \\ &= A - x_{(1)}^{-1} [y - \hat{\Phi}x_{(2:m)}] D_{(1,:)} - \hat{\Phi}D_{(2:m,:)} \\ &= A - \Phi_{(:,1)} D_{(1,:)} - \hat{\Phi}D_{(2:m,:)} \\ &= A - \Phi D. \end{aligned}$$

In Remark 4.2, we show $\hat{\Psi}$ exists if $U_{21}\Psi + U_{22}$ is nonsingular, and in Lemma 4.2 we show $U_{21}\Psi + U_{22}$ is always singular if $\mu = 0$. Theorem 5.2 asserts that $U_{21}\Psi + U_{22}$ is guaranteed nonsingular if $\mu \neq 0$. Thus the existence of $\hat{\Psi}$ is unresolved for the case $\mu = 0$, but otherwise $\hat{\Psi}$ exists. We point out that $\hat{\Psi}$ does exist for all our critical case examples in section 7 though.

Theorem 5.2. $U_{21}\Psi + U_{22}$ is singular when and only when $\mu = 0$.

Proof. We already know that $U_{21}\Psi + U_{22}$ is singular when $\mu = 0$ by Lemma 4.2. But the conclusion of the theorem is stronger than this. The proof below uses the explicit expressions for U_{ij} given in (5.5) which gives

$$U_{21}\Psi + U_{22} = [L_{21}(I - w_1 w_1^T) - w_2 w_1^T] \Psi - L_{21} w_1 w_2^T + I - w_2 w_2^T. \quad (5.18)$$

There are three cases to consider.

⁵This is not a misprint: the last $m - 1$ columns of $\tilde{\Phi}$ are the same as those of Φ .

1. If $i_0 = 1$, then $w_1 = 0$ and $w_2 = 0$ and thus (5.18) becomes

$$L_{21}\Psi + I = -x_{(1)}^{-1}ye_1^T\Psi + I$$

which is nonsingular if and only if $1 - x_{(1)}^{-1}e_1^T\Psi y \neq 0$. Now for $\mu > 0$, $\Psi y < x$ by Theorem 2.1 and then $x_{(1)}^{-1}e_1^T\Psi y < x_{(1)}^{-1}e_1^T x < 1$ implying $1 - x_{(1)}^{-1}e_1^T\Psi y > 0$. But for $\mu = 0$, $\Psi y = x$ by Theorem 2.1 and then $x_{(1)}^{-1}e_1^T\Psi y = x_{(1)}^{-1}e_1^T x = 1$ implying $1 - x_{(1)}^{-1}e_1^T\Psi y = 0$.

2. If $1 < i_0 \leq m$, then $w_1 = e_1 - e_{i_0}$ and $w_2 = 0$. Write $P_1 = I - w_1w_1^T$ which is the permutation matrix that swaps the first entry and the i_0 th entry of x . (5.18) becomes

$$L_{21}(I - w_1w_1^T)\Psi + I = -x_{(i_0)}^{-1}ye_1^T P_1\Psi + I$$

which is nonsingular if and only if $1 - x_{(i_0)}^{-1}e_1^T P_1\Psi y \neq 0$. Now for $\mu > 0$, $\Psi y < x$ by Theorem 2.1 and then $x_{(i_0)}^{-1}e_1^T P_1\Psi y < x_{(i_0)}^{-1}e_1^T P_1 x < 1$ implying $1 - x_{(i_0)}^{-1}e_1^T P_1\Psi y > 0$. But for $\mu = 0$, $\Psi y = x$ by Theorem 2.1 and then $x_{(i_0)}^{-1}e_1^T P_1\Psi y = x_{(i_0)}^{-1}e_1^T P_1 x = 1$ implying $1 - x_{(i_0)}^{-1}e_1^T P_1\Psi y = 0$.

3. If $i_0 > m$, then $w_1 = e_1$ and $w_2 = -e_{j_0}$, where $j_0 = i_0 - m$. We have

$$L_{21} = -\hat{y}e_1^T, \quad \hat{y} = y_{(j_0)}^{-1}y - e_{j_0} + y_{(j_0)}^{-1}x_{(1)}e_{j_0}.$$

It can be verified that $L_{21}(I - w_1w_1^T) = 0$. Therefore

$$\begin{aligned} U_{21}\Psi + U_{22} &= -w_2w_1^T\Psi - L_{21}w_1w_2^T + I - w_2w_2^T \\ &= e_{j_0}e_1^T\Psi - \hat{y}e_{j_0}^T + I - e_{j_0}e_{j_0}^T \\ &= I - (\hat{y} + e_{j_0})e_{j_0}^T + e_{j_0}e_1^T\Psi \\ &= I - \begin{pmatrix} \hat{y} + e_{j_0} & -e_{j_0} \end{pmatrix} \begin{pmatrix} e_{j_0}^T \\ e_1^T\Psi \end{pmatrix} \end{aligned}$$

which, by Lemma 5.1, is invertible if and only if

$$I_2 - \begin{pmatrix} e_{j_0}^T \\ e_1^T\Psi \end{pmatrix} \begin{pmatrix} \hat{y} + e_{j_0} & -e_{j_0} \end{pmatrix} \quad (5.19)$$

is invertible. Use $\hat{y} + e_{j_0} = y_{(j_0)}^{-1}y + y_{(j_0)}^{-1}x_{(1)}e_{j_0}$ to simplify the matrix (5.19) to

$$\begin{pmatrix} -y_{(j_0)}^{-1}x_{(1)} & 1 \\ -y_{(j_0)}^{-1}[e_1^T\Psi y + x_{(1)}e_1^T\Psi e_{j_0}] & 1 + e_1^T\Psi e_{j_0} \end{pmatrix}$$

whose determinant is $y_{(j_0)}^{-1}[e_1^T\Psi y - x_{(1)}]$. Now if $\mu > 0$, then $\Psi y < x$ by Theorem 2.1 and thus $y_{(j_0)}^{-1}[e_1^T\Psi y - x_{(1)}] < 0$. If $\mu = 0$, then $\Psi y = x$ by Theorem 2.1 and thus $y_{(j_0)}^{-1}[e_1^T\Psi y - x_{(1)}] = 0$ implying $U_{21}\Psi + U_{22}$ is singular.

This completes the proof. \square

5.2 By Orthogonal Transformation

We take V to be an orthogonal matrix $Q \in \mathbb{R}^{(m+n) \times (m+n)}$ such that $Q^T z = \delta e_1$. Partition

$$Q = \begin{matrix} & m & n \\ \begin{matrix} m \\ n \end{matrix} & \begin{pmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{pmatrix} \end{matrix}. \quad (5.20)$$

Then $V^{-1} = Q^T$ gives $U_{ij} = Q_{ji}^T$ and consequently

$$\tilde{\Phi} = (Q_{12}^T + Q_{22}^T \Phi) (Q_{11}^T + Q_{21}^T \Phi)^{-1}, \quad (5.21a)$$

$$\Phi = (-Q_{22}^T + \tilde{\Phi} Q_{21}^T)^{-1} (Q_{12}^T - \tilde{\Phi} Q_{11}^T), \quad (5.21b)$$

assuming $Q_{11}^T + Q_{21}^T \Phi$ and $-Q_{22}^T + \tilde{\Phi} Q_{21}^T$ are invertible. We know $\tilde{\Phi} e_1 = 0$ by Lemma 4.1, and $\hat{\Phi} = \tilde{\Phi}_{(:,2:m)}$ satisfies ARE (4.17).

Possible candidates for Q include a product of $m+n-1$ Givens rotations or a Householder transformation [8]. In what follows, we will use $V = Q$, the Householder transformation such that $Qz = -\|z\|_2 e_1$, as an example, partly because then both $Q_{11}^T + Q_{21}^T \Phi$ and $-Q_{22}^T + \tilde{\Phi} Q_{21}^T$ are guaranteed invertible⁶ by Theorem 5.3 below.

The Householder transformation $V = Q$ such that $Qz = -\|z\|_2 e_1$ is given by

$$Q = I - 2ww^T, \quad w = \frac{z - \delta e_1}{\|z - \delta e_1\|_2} = \frac{z - \delta e_1}{\gamma}, \quad (5.22)$$

where

$$\delta = -\|z\|_2, \quad \gamma = \|z - \delta e_1\|_2 = \sqrt{2x_{(1)}\|z\|_2 + 2\|z\|_2^2}. \quad (5.23)$$

Partition $w = \begin{pmatrix} w_1 \\ w_2 \end{pmatrix}$, where

$$0 < w_1 = \gamma^{-1}(x - \delta e_1) \in \mathbb{R}^m, \quad 0 < w_2 = \gamma^{-1}y \in \mathbb{R}^n. \quad (5.24)$$

Then the four submatrices Q_{ij} as defined by (5.20) are

$$Q_{11} = I_m - 2w_1 w_1^T, \quad Q_{12} = -2w_1 w_2^T, \quad (5.25a)$$

$$Q_{22} = I_n - 2w_2 w_2^T, \quad Q_{21} = -2w_2 w_1^T. \quad (5.25b)$$

Theorem 5.3. *Let $Q \in \mathbb{R}^{(m+n) \times (m+n)}$ be the Householder transformation as given by (5.22) and (5.23). Then both $Q_{11}^T + Q_{21}^T \Phi$ and $-Q_{22}^T + \tilde{\Phi} Q_{21}^T$ are invertible, where $\tilde{\Phi}$ relates to Φ by (5.21a).*

Proof. We have (5.22) – (5.25), and thus

$$Q_{11}^T + Q_{21}^T \Phi = I_m - 2w_1 w_1^T - 2w_1 w_2^T \Phi = I_m - 2w_1 (w_1^T + w_2^T \Phi).$$

⁶This is not so for the Householder transformation such that $Qz = \|z\|_2 e_1$. For example in Example 7.3, $Q_{11}^T + Q_{21}^T \Phi = \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix}$ for the Householder transformation Q such that $Q\mathbf{1}_4 = 2e_1$, but $Q_{11}^T + Q_{21}^T \Phi = \begin{pmatrix} -1 & -1 \\ -2/3 & 2/3 \end{pmatrix}$ for the Householder transformation Q such that $Q\mathbf{1}_4 = -2e_1$.

By Lemma 5.1, $Q_{11}^T + Q_{21}^T \Phi$ is invertible if and only if $1 - 2(w_1^T + w_2^T \Phi)w_1 \neq 0$ which we will verify. We have

$$\begin{aligned} 1 - 2(w_1^T + w_2^T \Phi)w_1 &= 1 - 2w_1^T w_1 - 2w_2^T \Phi w_1 \\ &= 1 - 2 \frac{\|x - \delta e_1\|_2^2}{\gamma^2} - 2w_2^T \Phi w_1 \\ &= -\frac{x_{(1)}\|z\|_2 + \|x\|_2^2}{x_{(1)}\|z\|_2 + \|z\|_2^2} - 2w_2^T \Phi w_1 < 0 \end{aligned}$$

because $x > 0$, $y > 0$, $\Phi > 0$, and $w_i > 0$. So $Q_{11}^T + Q_{21}^T \Phi$ is invertible and

$$(Q_{11}^T + Q_{21}^T \Phi)^{-1} = I_m + \frac{2w_1(w_1^T + w_2^T \Phi)}{1 - 2w_1^T w_1 - 2w_2^T \Phi w_1}.$$

Next we have

$$-Q_{22}^T + \tilde{\Phi}Q_{21}^T = -I + 2w_2 w_2^T - 2\tilde{\Phi}w_1 w_2^T = -\left[I - 2(w_2 - \tilde{\Phi}w_1)w_2^T\right]$$

which is invertible if and only if

$$1 - 2w_2^T(w_2 - \tilde{\Phi}w_1) = 1 - 2(w_2^T w_2 - w_2^T \tilde{\Phi}w_1) \neq 0$$

which we will now verify. We have

$$\begin{aligned} w_2^T (Q_{12}^T + Q_{22}^T \Phi) &= w_2^T [-2w_2 w_1^T + (I - 2w_2 w_2^T)\Phi] \\ &= (-2w_2^T w_2)w_1^T + (1 - 2w_2^T w_2)w_2^T \Phi, \\ (Q_{11}^T + Q_{21}^T \Phi)^{-1} w_1 &= \left[1 + \frac{2w_1^T w_1 + 2w_2^T \Phi w_1}{1 - 2w_1^T w_1 - 2w_2^T \Phi w_1}\right] w_1 \\ &= \frac{1}{1 - 2w_1^T w_1 - 2w_2^T \Phi w_1} w_1, \\ w_2^T \tilde{\Phi}w_1 &= w_2^T (Q_{12}^T + Q_{22}^T \Phi) \cdot (Q_{11}^T + Q_{21}^T \Phi)^{-1} w_1 \\ &= \frac{(-2w_2^T w_2)w_1^T w_1 + (1 - 2w_2^T w_2)w_2^T \Phi w_1}{1 - 2w_1^T w_1 - 2w_2^T \Phi w_1}, \\ w_2^T w_2 - w_2^T \tilde{\Phi}w_1 &= \frac{w_2^T w_2 - w_2^T \Phi w_1}{1 - 2w_1^T w_1 - 2w_2^T \Phi w_1}, \\ 1 - 2(w_2^T w_2 - w_2^T \tilde{\Phi}w_1) &= \frac{1 - 2w_1^T w_1 - 2w_2^T w_2}{1 - 2w_1^T w_1 - 2w_2^T \Phi w_1} \\ &= -\frac{1}{1 - 2w_1^T w_1 - 2w_2^T \Phi w_1} > 0, \end{aligned}$$

as expected. \square

REMARK 5.1. Theorem 5.3 is proved under the inherited conditions $x > 0$, $y > 0$, $\Phi > 0$, and $\Phi x = y$. Carefully examining the proof, one finds that the condition of the theorem can be relaxed to

$$x \geq 0, \quad x \neq 0, \quad y \geq 0, \quad \Phi \geq 0,$$

and $\tilde{\Phi}$ relates to Φ by (5.21a). Since $\Phi x = y$ is never referenced, it is not required. \diamond

The above proof also yields

$$(Q_{11}^T + Q_{21}^T \Phi)^{-1} = I_m + 2\zeta^{-1}w_1(w_1^T + w_2^T \Phi), \quad (5.26a)$$

$$\left(-Q_{22}^T + \tilde{\Phi}Q_{21}^T\right)^{-1} = -\left[I_n - 2\zeta(w_2 - \tilde{\Phi}w_1)w_2^T\right], \quad (5.26b)$$

where

$$\zeta = 1 - 2w_1^T w_1 - 2w_2^T \Phi w_1 < 0. \quad (5.27)$$

With the help of (5.26), we can express any one of Φ and $\tilde{\Phi}$ in terms of the other via a rank-one update. Details are as follows. By (5.21), we have

$$\begin{aligned} \tilde{\Phi} &= [-2w_2w_1^T + (I - 2w_2w_2^T)\Phi] [I + 2\zeta^{-1}w_1(w_1^T + w_2^T \Phi)] \\ &= [\Phi - 2w_2(w_1^T + w_2^T \Phi)] [I + 2\zeta^{-1}w_1(w_1^T + w_2^T \Phi)] \\ &= \Phi + 2\zeta^{-1}\Phi w_1(w_1^T + w_2^T \Phi) \\ &\quad - 2w_2(w_1^T + w_2^T \Phi) - 4\zeta^{-1}w_2 \underbrace{(w_1^T + w_2^T \Phi)w_1(w_1^T + w_2^T \Phi)}_{\text{scalar}} \\ &= \Phi + 2\zeta^{-1}\Phi w_1(w_1^T + w_2^T \Phi) - 2[1 + 2\zeta^{-1}(w_1^T + w_2^T \Phi)w_1]w_2(w_1^T + w_2^T \Phi) \\ &= \Phi + 2\{\zeta^{-1}\Phi w_1 - [1 + 2\zeta^{-1}(w_1^T + w_2^T \Phi)w_1]w_2\}(w_1^T + w_2^T \Phi), \end{aligned} \quad (5.28a)$$

$$\begin{aligned} \Phi &= \left[-I_n + 2\zeta(w_2 - \tilde{\Phi}w_1)w_2^T\right] \left[-2w_2w_1^T - \tilde{\Phi}(I - 2w_1w_1^T)\right] \\ &= \left[-I_n + 2\zeta(w_2 - \tilde{\Phi}w_1)w_2^T\right] \left[-\tilde{\Phi} - 2(w_2 - \tilde{\Phi}w_1)w_1^T\right] \\ &= \tilde{\Phi} + 2(w_2 - \tilde{\Phi}w_1)w_1^T \\ &\quad - 2\zeta(w_2 - \tilde{\Phi}w_1)w_2^T \tilde{\Phi} - 4\zeta(w_2 - \tilde{\Phi}w_1) \underbrace{w_2^T(w_2 - \tilde{\Phi}w_1)w_1^T}_{\text{scalar}} \\ &= \tilde{\Phi} + 2\left[1 - 2\zeta w_2^T(w_2 - \tilde{\Phi}w_1)\right] (w_2 - \tilde{\Phi}w_1)w_1^T - 2\zeta(w_2 - \tilde{\Phi}w_1)w_2^T \tilde{\Phi} \\ &= \tilde{\Phi} + 2(w_2 - \tilde{\Phi}w_1) \left\{ \left[1 - 2\zeta w_2^T(w_2 - \tilde{\Phi}w_1)\right] w_1^T - \zeta w_2^T \tilde{\Phi} \right\}. \end{aligned} \quad (5.28b)$$

Equation (5.28b) will become handy in coding up Algorithm 4.1, where recovering Φ is needed from computed $\hat{\Phi}$ by (5.28b) with $\tilde{\Phi} = \begin{pmatrix} 0 & \hat{\Phi} \end{pmatrix}$.

Extractions of the coefficient matrices \hat{A} , \hat{B} , \hat{C} , and \hat{D} for ARE (4.17) can be easily done from the partitioning (4.12) for

$$\begin{aligned} V^{-1}HV &= (I - 2ww^T)H(I - 2ww^T) \\ &= H - 2ww^T H - 2Hww^T + 4(w^T Hw)ww^T, \end{aligned} \quad (5.29)$$

where the expression in the right-hand side of (5.29) suggests an economical way to numerically compute $V^{-1}HV$.

In Remark 4.2, we show $\hat{\Psi}$ exists if $U_{21}\Psi + U_{22}$ is nonsingular, and in Lemma 4.2 we show $U_{21}\Psi + U_{22}$ is always singular if $\mu = 0$. Theorem 5.4 asserts that $U_{21}\Psi + U_{22}$ is guaranteed nonsingular if $\mu \neq 0$. Thus the existence of $\hat{\Psi}$ is unresolved for the case $\mu = 0$, but otherwise $\hat{\Psi}$ exists. We point out that $\hat{\Psi}$ does exist for all our critical case examples in section 7 though.

Theorem 5.4. Let $Q \in \mathbb{R}^{(m+n) \times (m+n)}$ be the Householder transformation as given by (5.22) and (5.23). $U_{21}\Psi + U_{22}$ is singular when and only when $\mu = 0$.

Proof. We have by (5.25) and $U_{ij} = Q_{ji}^T$ that

$$U_{21}\Psi + U_{22} = -2w_2w_1^T\Psi + I - 2w_2w_2^T = I - 2w_2(w_2^T + w_1^T\Psi)$$

which is invertible if and only if $1 - 2(w_2^T + w_1^T\Psi)w_2 \neq 0$ which we now verify. Recall (5.23) and (5.24) and that $\Psi y < x$ for $\mu > 0$ and $\Psi y = x$ for $\mu = 0$. We have

$$\begin{aligned} 2(w_2^T + w_1^T\Psi)w_2 &= \frac{2y^T y + 2(x + \|z\|_2 e_1)^T \Psi y}{\gamma^2} \\ &\leq \frac{y^T y + (x + \|z\|_2 e_1)^T x}{x_{(1)}\|z\|_2 + \|z\|_2^2} \\ &= 1, \end{aligned}$$

where the equality occurs when and only when $\mu = 0$. Therefore $1 - 2(w_2^T + w_1^T\Psi)w_2 \geq 0$ with equality when and only when $\mu = 0$. \square

6 Shifting Approach of Guo, Iannazzo, and Meini

Having recognized slow convergence of SDA on irreducible singular MAREs in the critical case, Guo, Iannazzo, and Meini [12] proposed to perform a rank-one update on H to shift away one of H 's eigenvalue 0, and then apply SDA on the resulting ARE (which is no longer an MARE, however).

Suppose MARE (1.1) with (1.6) and $\mu = u^T x - v^T y \geq 0$. Pick $\eta \in \mathbb{R}$ to be specified in a moment, and let

$$\widehat{H} = H + \eta z w^T \equiv \begin{matrix} m & n \\ \widehat{B} & -\widehat{D} \\ \widehat{C} & -\widehat{A} \end{matrix}, \quad (6.1)$$

where $w \in \mathbb{R}^{m+n}$ is entrywise nonnegative such that $w^T z = 1$. This gives arise the following ARE

$$\widehat{X}\widehat{D}\widehat{X} - \widehat{A}\widehat{X} - \widehat{X}\widehat{B} + \widehat{C} = 0. \quad (6.2)$$

It is proved in [12] that $\widehat{X} = \Phi$ is the solution of (6.2) uniquely characterized by

$$\text{eig}(\widehat{R}) = \{\lambda_1, \dots, \lambda_{m-1}, \eta\},$$

and at the same time the complementary ARE of (6.2) has the solution $\widehat{\Psi}$ uniquely characterized by

$$\text{eig}(\widehat{S}) = \{-\lambda_{m+1}, \dots, -\lambda_{m+n}\},$$

where

$$\widehat{R} = \widehat{B} - \widehat{D}\Phi, \quad \widehat{S} = \widehat{A} - \widehat{C}\widehat{\Psi}.$$

In solving (6.2) by SDA [14], Guo, Iannazzo, and Meini [12] picked

$$w = \mathbf{1}_{m+n} / (\mathbf{1}_{m+n}^T z) \quad (6.3)$$

for simplicity, and

$$\alpha = \beta = \eta = \max\{\alpha_{\text{opt}}, \beta_{\text{opt}}\} \quad (6.4)$$

to ensure⁷ $\eta \in \text{eig}(\widehat{R})$ contributes nothing to $\rho(\mathcal{C}(\widehat{R}; \eta, \eta))$, where α_{opt} and β_{opt} are as in (3.9).

It has been noted [20] that compared to ADDA, SDA will experience slow convergence if α_{opt} and β_{opt} differ substantially. Naturally applying ADDA to (6.2) would likely lead to a faster algorithm for the same reason. The rate of convergence of ADDA on (6.2) is determined by $\rho(\mathcal{C}(\widehat{S}; \alpha, \beta)) \cdot \rho(\mathcal{C}(\widehat{R}; \beta, \alpha))$, and we will pick

$$\alpha = \alpha_{\text{opt}}, \quad \eta = \beta = \beta_{\text{opt}}, \quad \eta = \beta_{\text{opt}}, \quad (6.5)$$

as discussed in Remark 4.3 and to make sure $\eta \in \text{eig}(\widehat{R})$ contributes nothing to $\rho(\mathcal{C}(\widehat{R}; \beta, \alpha))$.

For their references in the next section, we denote these two methods for solving MARE (1.1) via ARE (6.2) by SDAs and ADDAs, respectively, with the suffix “s” standing for the shift in (6.1). We will use the parameters in (6.3) and (6.4) for SDAs and those in (6.3) and (6.5) for ADDAs.

7 Numerical Examples

In this section, we will present five numerical examples to test numerical effectiveness of dADDA, in comparison with ADDA, SDAs, and ADDAs. We will use the *normalized residual* (NRes) error to gauge accuracy in a computed solution Φ :

$$\text{NRes} = \frac{\|\Phi D \Phi - A \Phi - \Phi B + C\|_1}{\|\Phi\|_1 (\|\Phi\|_1 \|D\|_1 + \|A\|_1 + \|B\|_1) + \|C\|_1}, \quad (7.1)$$

a commonly used measure because it is readily available, and the *entrywise relative error* (ERErr) and *normalized error* (NErr),

$$\text{ERErr} = \max_{i,j} \frac{|(\Phi - \widehat{\Phi})_{(i,j)}|}{\Phi_{(i,j)}}, \quad \text{NErr} = \frac{\|\Phi - \widehat{\Phi}\|_1}{\|\Phi\|_1} \quad (7.2)$$

which are not available in actual computations but is made available here for testing purpose. The use of ℓ_1 -operator norm is inconsequential but for computational convenience, and any other matrix norm would be equally effective in demonstrating our points. In the case of ERErr, the indeterminate $0/0$ is treated as 0. These errors defined in (7.1) and (7.2) are 0 if Φ is exact, but numerically they can only be made as small as $O(\mathbf{u})$, where \mathbf{u} is the unit machine roundoff.

In [20, 21], it was argued that the doubling algorithms SDA [14, 12], SDA-ss [4], and ADDA [20] all can deliver computed minimal nonnegative solutions of an MARE with deserved entrywise relative accuracy, if properly implemented to avoid harmful cancelations. But both our deflated ARE (4.17) and the shifted ARE (6.2) are no longer MAREs and thus there is no guarantee that all harmful cancelations can be avoided when SDA or ADDA is applied to either one of them. This means that in general computed minimal nonnegative solutions Φ may not have deserved entrywise relative accuracy if some of the entries of Φ are very tiny relative to others, even though NRes is reduced to the level of $O(\mathbf{u})$. For this reason, we will use $\text{NRes} \leq 5 \times 10^{-14}$

Example		ADDA	SDAs	ADDAs	dADDAe	dADDAq
7.1 ($\xi = 1$)	NRes	$2.1 \cdot 10^{-14}$	$3.0 \cdot 10^{-15}$	$3.0 \cdot 10^{-15}$	$5.1 \cdot 10^{-15}$	$1.0 \cdot 10^{-15}$
	NErr	$3.6 \cdot 10^{-7}$	$3.5 \cdot 10^{-14}$	$3.5 \cdot 10^{-14}$	$6.3 \cdot 10^{-14}$	$7.5 \cdot 10^{-15}$
	RErr	$4.8 \cdot 10^{-6}$	$6.2 \cdot 10^{-13}$	$6.2 \cdot 10^{-13}$	$8.5 \cdot 10^{-13}$	$1.5 \cdot 10^{-13}$
7.1 ($\xi = 10$)	NRes	$2.4 \cdot 10^{-17}$	$8.4 \cdot 10^{-16}$	$4.3 \cdot 10^{-16}$	$5.3 \cdot 10^{-15}$	$1.0 \cdot 10^{-15}$
	NErr	$7.5 \cdot 10^{-17}$	$2.1 \cdot 10^{-15}$	$1.3 \cdot 10^{-15}$	$1.5 \cdot 10^{-14}$	$3.3 \cdot 10^{-15}$
	RErr	$2.0 \cdot 10^{-3}$	$2.4 \cdot 10^{12}$	$2.3 \cdot 10^{11}$	$5.8 \cdot 10^{13}$	$4.8 \cdot 10^{12}$
7.2	NRes	$4.9 \cdot 10^{-16}$	$3.6 \cdot 10^{-16}$	$2.6 \cdot 10^{-16}$	$9.9 \cdot 10^{-15}$	$7.5 \cdot 10^{-16}$
	NErr	$2.2 \cdot 10^{-15}$	$1.5 \cdot 10^{-15}$	$1.1 \cdot 10^{-14}$	$2.3 \cdot 10^{-14}$	$3.2 \cdot 10^{-15}$
	RErr	$4.4 \cdot 10^{-15}$	$2.8 \cdot 10^{-15}$	$2.3 \cdot 10^{-14}$	$1.3 \cdot 10^{-13}$	$8.5 \cdot 10^{-15}$
7.3	NRes	$2.1 \cdot 10^{-14}$	$7.4 \cdot 10^{-17}$	$7.4 \cdot 10^{-17}$	$1.0 \cdot 10^{-14}$	$5.0 \cdot 10^{-15}$
	NErr	$3.6 \cdot 10^{-7}$	$2.2 \cdot 10^{-16}$	$2.2 \cdot 10^{-16}$	$3.0 \cdot 10^{-14}$	$1.5 \cdot 10^{-14}$
	RErr	$3.6 \cdot 10^{-7}$	$2.2 \cdot 10^{-16}$	$2.2 \cdot 10^{-16}$	$3.0 \cdot 10^{-14}$	$1.5 \cdot 10^{-14}$
7.4	NRes	$2.1 \cdot 10^{-14}$	$3.2 \cdot 10^{-20}$	$2.8 \cdot 10^{-20}$	$4.6 \cdot 10^{-17}$	$9.0 \cdot 10^{-17}$
	NErr	$4.6 \cdot 10^{-5}$	$3.3 \cdot 10^{-16}$	$3.3 \cdot 10^{-12}$	$2.3 \cdot 10^{-12}$	$4.4 \cdot 10^{-12}$
	RErr	$4.6 \cdot 10^{-5}$	$3.3 \cdot 10^{-16}$	$3.3 \cdot 10^{-12}$	$2.3 \cdot 10^{-12}$	$4.4 \cdot 10^{-12}$
7.5	NRes	$1.8 \cdot 10^{-16}$	$1.3 \cdot 10^{-16}$	$1.3 \cdot 10^{-16}$	$7.9 \cdot 10^{-16}$	$2.5 \cdot 10^{-16}$
	NErr	$1.0 \cdot 10^{-12}$	$3.7 \cdot 10^{-16}$	$2.5 \cdot 10^{-16}$	$1.5 \cdot 10^{-15}$	$5.6 \cdot 10^{-16}$
	RErr	$1.5 \cdot 10^{-12}$	$3.7 \cdot 10^{-16}$	$2.5 \cdot 10^{-16}$	$1.5 \cdot 10^{-15}$	$1.0 \cdot 10^{-15}$

Table 7.1: NRes, NErr, and RErr at convergence for all examples. Boldfaced entries are worth paying attention to. For the critical cases (Examples 7.1 ($\xi = 1$), 7.3, and 7.4), ADDA on the original MAREs returns solutions with NErr about $O(\sqrt{\mathbf{u}})$, consistent with the error analysis in [11], even though the corresponding NRes is already $O(\mathbf{u})$. Examples 7.1 ($\xi = 10$) is special in that Φ 's entries varies greatly in magnitude and consequently SDAs, ADDAs, dADDAe, and dADDAq have trouble getting tiny entries of Φ correct, even though all NErr are already $O(\mathbf{u})$. ADDA would have computed Φ to nearly full entrywise relative accuracy if it had continued for two more iterations [20].

as the stopping criteria⁸ in our tests here, instead of Kahan's criteria [22, 20] designed to stop the iterations only when Φ is computed to its deserved entrywise relative accuracy.

All computations are performed in MATLAB with $\mathbf{u} = 1.11 \times 10^{-16}$. Five methods are tested, and they are

1. ADDA of [20]. We use it as a representative for all doubling algorithms derivable from bilinear transformations, including SDA [14, 12] and SDA-ss [4], since ADDA is the fastest among all [20].
2. SDAs of [12] (as outlined in section 6). It is the first method ever proposed to improve

⁷Recall that SDA is ADDA (Algorithm 3.1) after setting $\alpha = \beta$, and its rate of convergence is determined by $\rho(\mathcal{E}(\hat{S}; \alpha, \alpha)) \cdot \rho(\mathcal{E}(\hat{R}; \alpha, \alpha))$.

⁸In [12], $\min\{\|E_k\|_1, \|F_k\|_1\} < 10^{-15}$ was used to stop the SDA iteration on (6.2). It unnecessarily asked too much because $X_k - \Phi$ is proportional to the product $\|E_k\|_1 \|F_k\|_1$. For example, Tests 2 and 3 of [12] are essentially the same as our Examples 7.3 and 7.4. We notice on (6.2) for both examples, X_0 from SDA's initial setup is *exact* in exact arithmetic as confirmed by Maple, whereas Table 7.1 of [12] still reported that SDAs needed 5 and 4 iterations, respectively, before $\min\{\|E_k\|_1, \|F_k\|_1\} < 10^{-15}$ was met.

SDA for irreducible singular MAREs.

3. ADDAs (as outlined in section 6). Since ADDA improves SDA, naturally we would expect ADDAs improves SDAs.
4. dADD Ae which is Algorithm 4.1 combined with the elimination approach in subsection 5.1. For simplicity, all $i_0 = 1$. Actually in all examples, $z = \mathbf{1}_{m+n}$; so there is no need to do pivoting to control $\|V\|_1\|V^{-1}\|_1$.
5. dADD Aq which is Algorithm 4.1 combined with the Householder transformation approach in subsection 5.2.

Example 7.1 ([20, Example 7.2]).

$$B = \begin{pmatrix} 3 & -1 & & \\ & 3 & \ddots & \\ & & \ddots & -1 \\ -1 & & & 3 \end{pmatrix} \in \mathbb{R}^{n \times n}, \quad D = 2I_n, \quad A = \xi B, \quad C = \xi D.$$

W is an irreducible singular M -matrix:

$$W\mathbf{1}_{2n} = 0, \quad \begin{pmatrix} \mathbf{1}_n \\ \xi^{-1} \cdot \mathbf{1}_n \end{pmatrix}^\top W = 0, \quad \mu = (1 - \xi^{-1})n.$$

For testing purpose, we computed for $n = 100$ an “exact” solution Φ by the computerized algebra system *Maple* with 100 decimal digits. We find that

$$7.4339 \cdot 10^{-4} \leq \Phi_{(i,j)} \leq 3.8270 \cdot 10^{-1}, \quad \text{for } \xi = 1, \quad (7.3)$$

$$5.7251 \cdot 10^{-30} \leq \Phi_{(i,j)} \leq 6.3012 \cdot 10^{-1}, \quad \text{for } \xi = 10. \quad (7.4)$$

Large variations in magnitudes in Φ 's entries for $\xi = 10$ suggest that all methods, except ADDA, may have trouble getting Φ 's tiny entries right. Indeed, they do.

Figure 7.1 plots the convergence histories of the five methods. For $\xi = 1$, ADDA converges linearly because the case falls into the critical case [5]. All methods are able to reduce NRes to about $O(\mathbf{u})$ as they should. Since Φ 's entries vary in magnitude by a factor about 500, we would expect that ERErr for all be about $O(500\mathbf{u}) = O(10^{-12})$ which is true for all methods, except ADDA as shown in Table 7.1. It can be explained. ADDA is applied to the original MARE in the critical case for which case it is argued by Guo and Higham [11] that roughly speaking a perturbation of size ϵ to W will result in an error in Φ about $O(\sqrt{\epsilon})$. On the other hand, the shifting technique built into SDAs and ADDAs and the deflating technique built into dADD Ae and dAADAq make the resulting ARE (4.17) and (6.2) sufficiently well-conditioned to be solved accurately. Guo, Iannazzo, and Meini [12] have already reported that SDAs produces more accurate solutions than SDA. Our explanation here for ERErr applies to the rest of examples, too.

Also for $\xi = 1$, quadratic convergence is evident for all methods, except ADDA, as expected. It is no longer in the critical case for $\xi = 10$. That partially explains ADDA's superior performance. ADDA would have computed Φ to with almost full entrywise accuracy if it had not been stopped prematurely by one stopping criteria $\text{NRes} \leq 5 \times 10^{-14}$ used for all. In fact, this

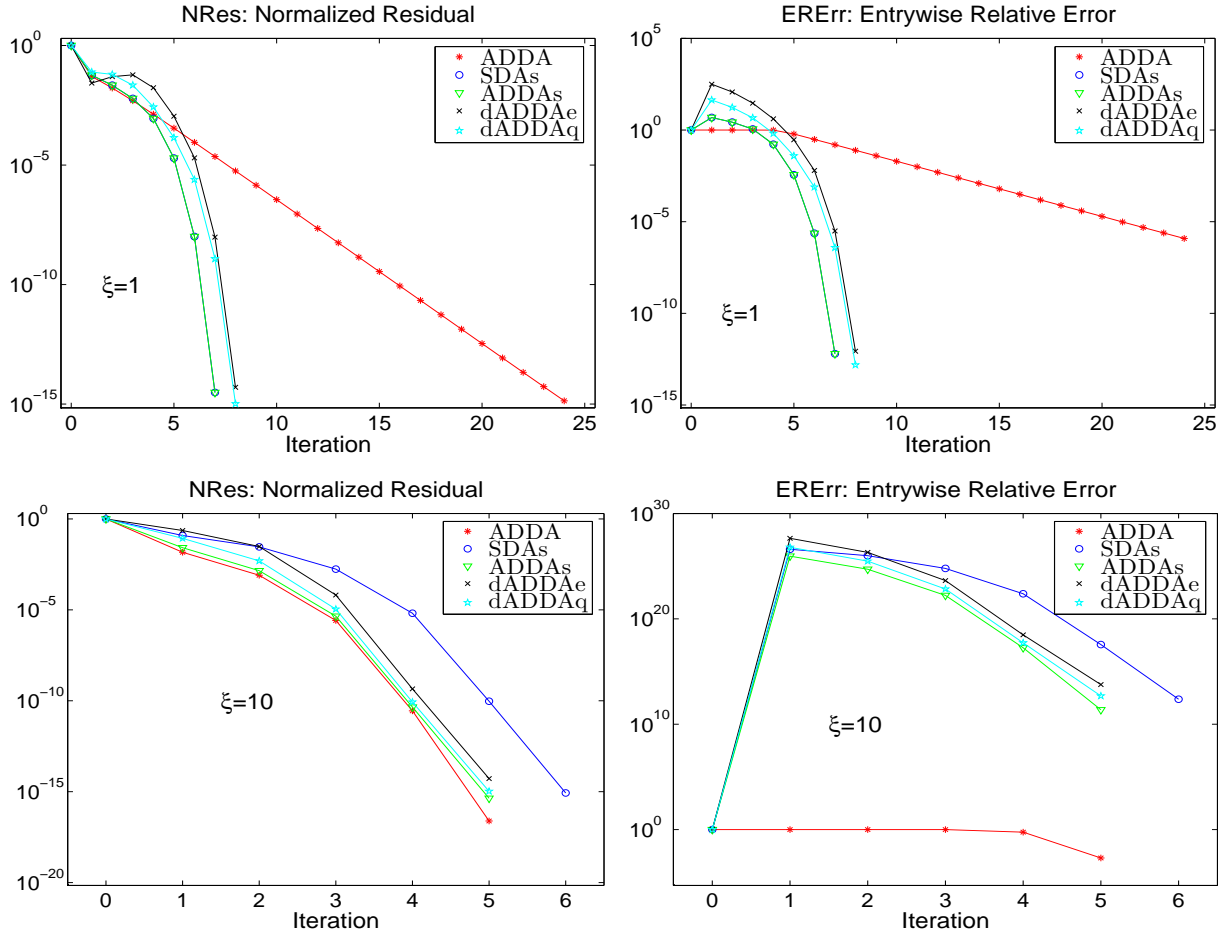


Figure 7.1: Example 7.1. For $\xi = 1$ ADDA converges linearly and for $\xi = 10$ ADDA performs the best. Also for $\xi = 10$, all methods, except ADDA (which took 7 iterations in [20], two more than here, to deliver Φ with about 15 correct decimal digits entrywise), fail to compute accurately Φ 's tiny entries.

example is the same as [20, Example 7.2], where ADDA delivered Φ to have almost 15 correct decimal digits entrywise in 7 iterations. The inability of the other methods to compute Φ 's tiny entries accurately is evident from the right-bottom plot in Figure 7.1 and Table 7.1, even though at the same time all methods are able to reduce NRes to about $O(\mathbf{u})$. \diamond

Example 7.2. W is an irreducible singular M -matrix, randomly generated by the following piece of MATLAB code:

```
n=100;
W=rand(2*n);    W(n+1:2*n,:)=10*W(n+1:2*n,:);
W=round(1000*W); W=diag(W*ones(2*n,1))-W;
```

In the end, $W\mathbf{1}_{2n} = 0$, and with $m = n$, the coefficient matrices A , B , C , and D for MARE (1.1) can be readily extracted. There are a couple of comments to make about constructing W this way. The factor 10 applied to the last n rows in the second line serves two purposes: (1) to make A and B differ in magnitude by a factor about 10, and (2) to make sure $\mu \geq 0$ (although

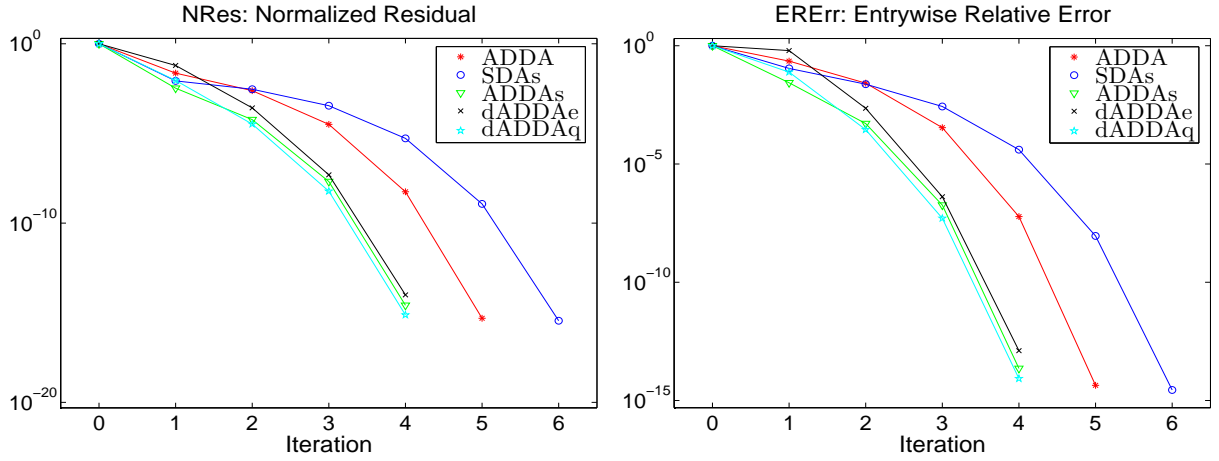


Figure 7.2: Example 7.2. ADDA is even faster than SDAs, and ADDAs, dADDAe, and dADDAq work about equally well.

Ex.		Elimination ($i_0 = 1$)	Householder ($\delta = -\ z\ _2$)
7.3	$V^{-1}HV$	$\begin{pmatrix} 0 & -1 & -1 & -1 \\ 0 & 4 & 0 & 0 \\ 0 & 2 & -2 & 2 \\ 0 & 2 & 2 & -2 \end{pmatrix}$	$\frac{1}{9} \begin{pmatrix} 0 & -12 & 24 & 24 \\ 0 & 32 & 8 & 8 \\ 0 & 8 & -16 & 20 \\ 0 & 8 & 20 & -16 \end{pmatrix}$
	$\hat{\Psi}$	$(0 \ 0)$	$-\frac{1}{4} \begin{pmatrix} 1 & 1 \end{pmatrix}$
7.4	$V^{-1}HV$	$\begin{pmatrix} 0 & -1 & -1 & -1 \\ 0 & 4 & 0 & 0 \\ 0 & 2 & -100001 & 100001 \\ 0 & 2 & 100001 & -100001 \end{pmatrix}$	$\frac{1}{9} \begin{pmatrix} 0 & -12 & 24 & 24 \\ 0 & 32 & 8 & 8 \\ 0 & 8 & -900007 & 900011 \\ 0 & 8 & 900011 & -900007 \end{pmatrix}$
	$\hat{\Psi}$	$(0 \ 0)$	$-\frac{1}{4} \begin{pmatrix} 1 & 1 \end{pmatrix}$

Table 7.2: Examples 7.3 and 7.4: $V^{-1}HV$ and $\hat{\Psi}$

not always guaranteed in theory but often it is). At the beginning of the third line, we multiply W by 1000 and round its entries to integers so that we can save one such a W and then move the generated W error-free to Maple to compute the “exact” Φ for testing purpose. For this saved W , we find that

$$4.7301 \cdot 10^{-3} \leq \Phi_{(i,j)} \leq 1.5684 \cdot 10^{-2}.$$

So all entries of Φ have about the same magnitude which suggests that tiny NRes implies tiny ERErr. This is clearly the case as shown in Figure 7.2. What is interesting to see is that SDAs is actually slower than ADDA. The reasons are twofold: (1) this is not a critical case example, and (2) A and B have different magnitudes which SDAs (and SDA) choose to ignore but ADDA doesn’t. ADDAs, dADDAe, and dADDAq work about equally well, with dADDAe a little worse in accuracy, however. \diamond

Example 7.3 ([20, Example 7.1]). In this example, $m = n = 2$ and

$$B = \begin{pmatrix} 3 & -1 \\ -1 & 3 \end{pmatrix}, \quad D = \mathbf{1}_{2,2}, \quad A = B, \quad C = D.$$

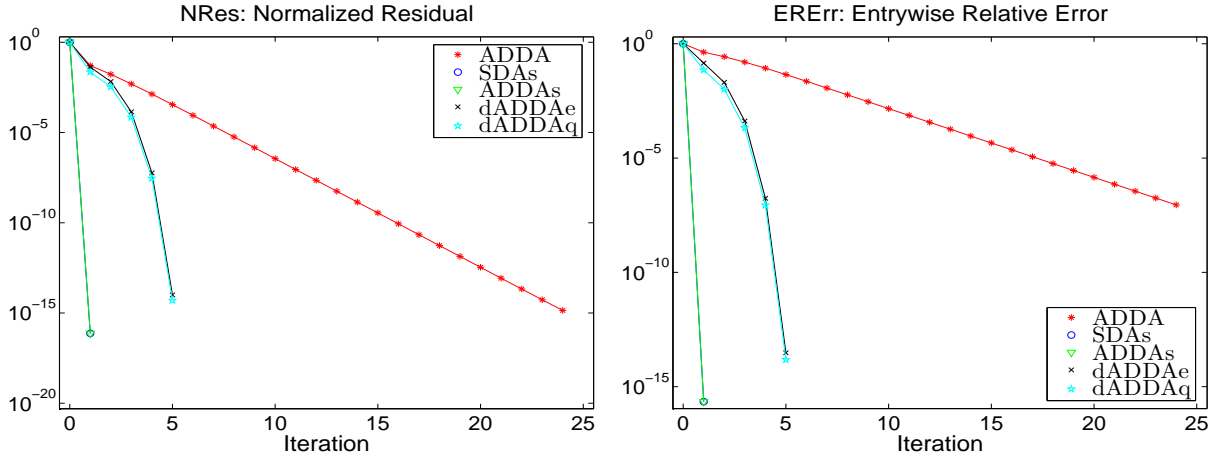


Figure 7.3: Example 7.3. SDAs and ADDAs get the solution in X_0 from their initial setup! In fact, $X_0 \equiv \Phi$ in exact arithmetic, independent of α and β , for ADDAs and thus SDAs. ADDA is linearly convergent, and dADDAe and dADDAq work about equally well.

Scaling W by 10^{-3} recovers a null recurrent case example in [2] (see also [12, Test 7.2]). It can be verified that

$$W\mathbf{1}_4 = 0, \quad \mathbf{1}_4^T W = 0, \quad \Phi = \frac{1}{2}\mathbf{1}_{2,2}, \quad \Psi = \frac{1}{2}\mathbf{1}_{2,2}, \quad \mu = 0. \quad (7.5)$$

This example is small enough to allow us to write out $V^{-1}HV$ and $\widehat{\Psi}$ in Table 7.2 for the realizations in section 5. (So is the next example.) From the table, the coefficient matrices for the deflated ARE (4.17) can be easily read off.

From Table 7.2, we see $\widehat{D} = 0$ for dADDAe, leading to a Sylvester equation $\widehat{A}\widehat{X} + \widehat{X}\widehat{B} = \widehat{C}$ which becomes a linear system $(\widehat{A} + 4I_2)\widehat{X} = \widehat{C}$:

$$\begin{pmatrix} 6 & -2 \\ -2 & 6 \end{pmatrix} \widehat{X} = \begin{pmatrix} 2 \\ 2 \end{pmatrix} \Rightarrow \widehat{\Phi} = \Phi_{(:,2)} = \widehat{X} = \frac{1}{2}\mathbf{1}_2,$$

and then $\Phi_{(:,1)} = y - \Phi_{(:,2)} = \frac{1}{2}\mathbf{1}_2$. But the deflated ARE (4.17) is still truly an ARE for dADDAq with the Householder transformation. Figure 7.3 displays the convergence histories. ADDA converges linearly since this is a critical case example [5]. It is interesting to note that both SDAs and ADDAs get the solution in X_0 , the initial setup for the doubling algorithms, rather unusual and atypical, to say the least. In fact, our Maple code for ADDAs with arbitrary α and β but $\eta = \beta$ gives⁹, in exact arithmetic,

$$X_0 \equiv \Phi, \quad Y_0 \equiv \frac{1}{2} \cdot \frac{4 - \beta}{4 + \beta}.$$

We did not see this phenomenon in Examples 7.1 and 7.2 both of which are nontrivial, relatively speaking. So this kind of pleasant surprise shouldn't be expected in general. Nonetheless, it comes up again in the next two examples both of which are, however, obtained from equivalently modifying examples in [2]. \diamond

⁹With $\beta = 0$, both $X_0 \equiv \Phi$ and $Y_0 \equiv \Psi$, independent of α . The same happens for Examples 7.4 and 7.5.

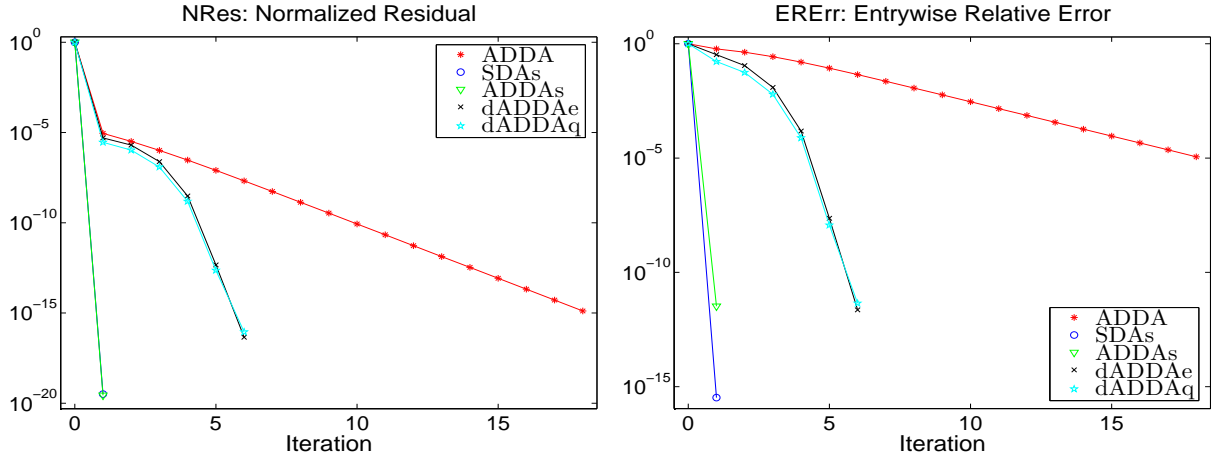


Figure 7.4: Example 7.4. SDAs and ADDAs get the solution in X_0 from their initial setup! In fact, $X_0 \equiv \Phi$ in exact arithmetic, independent of α and β , for ADDAs and thus SDAs. ADDA is linearly convergent, and dADDAe and dADDAq work about equally well.

Example 7.4. In this example $m = n = 2$, and

$$A = \begin{pmatrix} 100002 & -10^5 \\ -10^5 & 100002 \end{pmatrix}, \quad B = \begin{pmatrix} 3 & -1 \\ -1 & 3 \end{pmatrix}, \quad C = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}, \quad D = C.$$

Scaling W by 10^{-3} recovers a null recurrent case example in [2] (see also [4]). W is an irreducible singular M -matrix and (7.5) remains valid for this example. Again $\widehat{D} = 0$ for dADDAe as in Example 7.3. Figure 7.4 displays convergence histories for all tested methods. As in Example 7.3, $X_0 \equiv \Phi$ in exact arithmetic for ADDAs and thus SDAs, as confirmed by our Maple code (with arbitrary α and β but $\eta = \beta$):

$$X_0 \equiv \Phi, \quad Y_0 \equiv \frac{1}{2} \cdot \frac{4 - \beta}{4 + \beta}.$$

But it is interesting to note that SDAs' X_0 is much more accurate than ADDAs'. This is due to the conditioning of the involved matrices that have to be inverted. Specifically in (6.1),

$$\widehat{A} = \begin{pmatrix} 100001.25 & -100000.75 \\ -100000.75 & 100001.25 \end{pmatrix}, \quad \widehat{B} = \begin{pmatrix} 3.75 & -0.25 \\ -0.25 & 3.75 \end{pmatrix}.$$

For ADDA on (6.2), $\alpha = 100001.25$ and $\beta = 3.75$ which gives $\widehat{A} + \beta I_2$ whose condition number is 10^5 . Thus potentially 5 decimal digits could be lost in inverting $\widehat{A} + \beta I_2$. For SDA on (6.2), $\alpha = \beta = 100001.25$ which leads to very well-conditioned $\widehat{A} + \beta I_2$ and $\widehat{B} + \alpha I_2$. For the same reason, ADDAs, dADDAe, and dADDAq delivered less accurate solutions. \diamond

Example 7.5. This is essentially the example of a positive recurrent Markov chain with non-square coefficients, originally from [2]. Here

$$A = 18 \cdot I_2, \quad B = 180002 \cdot I_{18} - 10^4 \cdot \mathbf{1}_{18,18}, \quad C = \mathbf{1}_{2,18}, \quad D = C^T.$$

It is known $\Phi = \frac{1}{18} \cdot \mathbf{1}_{2,18} = \Psi^T$ and $\mu = 16 > 0$. All tested methods are quadratically convergent. As in Examples 7.3 and 7.4, $X_0 \equiv \Phi$ in exact arithmetic for ADDAs and thus SDAs

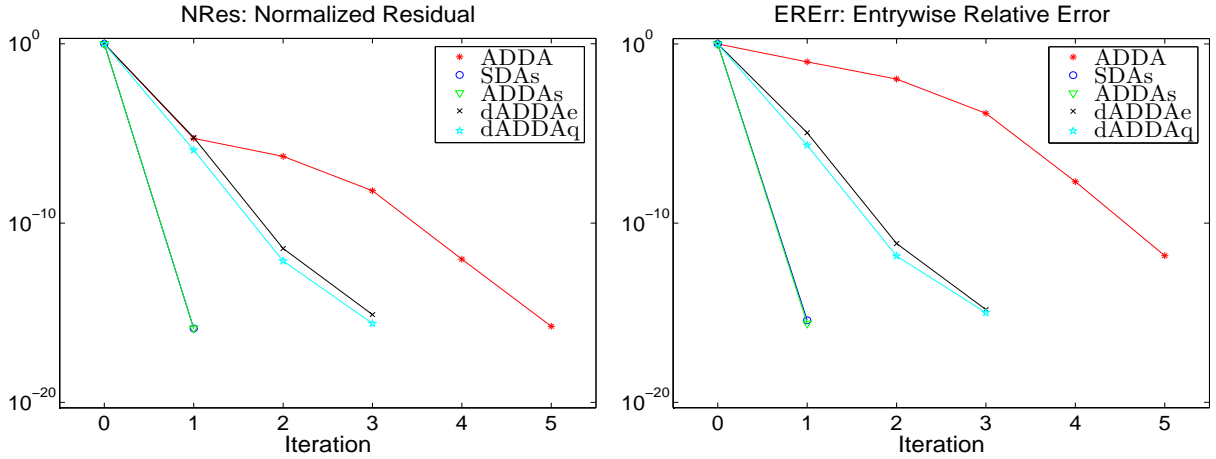


Figure 7.5: Example 7.5. Again SDAs and ADDAs get the solution in X_0 from their initial setup as they should because $X_0 \equiv \Phi$, independent of α and β . Both dADDAe and dADDAq take two iterations after the initial setup, while ADDA takes five iterations.

(with arbitrary α and β but $\eta = \beta$):

$$X_0 \equiv \Phi, \quad Y_0 \equiv \frac{1}{18} \cdot \frac{20 - \beta}{20 + \beta} \times \mathbf{1}_{18,2}.$$

Figure 7.5 displays convergence histories for all tested methods. That both NErr and ERErr for ADDA at convergence are about 10^{-12} can be explained by the relevant parameters in [20, Table 7.2]. \diamond

From these examples as well as many more random ones, we come to the following conclusions about speed and accuracy for the tested algorithms:

1. ADDA is linearly convergent for the critical case, but is able to deliver entrywise accurate approximations to Φ , even when some of the entries of Φ are extremely tiny relative to others. But entrywise accuracy in computed Φ is limited to about $O(\sqrt{u})$.
2. The shifting technique of Guo, Iannazzo, and Meini [12] and the deflating technique in this article can greatly improve the conditioning of an MARE in the critical case, enabling Φ to be computed much more accurately in the sense of making normalized error NErr to about $O(u)$. But when Φ 's entries vary too much in magnitude, tiny entries may lose some or even all significant digits. When that happens, ADDA should be used directly to the original MARE.
3. The last three examples are accidental for both ADDAs and SDAs in that $X_0 \equiv \Phi$, independent of the parameters α and β . In general, ADDAs is faster than SDAs as one might expect from the conclusion in [20] that ADDA is at least as good as SDA and can be faster if A and B are very different in magnitude.
4. One may have to monitor the conditioning of the matrices that have to be inverted in all doubling algorithms as Example 7.4 shows.

8 Concluding Remarks

Doubling algorithms converge linearly for MAREs in the critical case and quadratically for those that are not in the critical case. Guo, Iannazzo, and Meini [12] recognized it and proposed a shifting mechanism to still retain quadratical convergence. In this paper, We establish a general framework to deflate an irreducible singular MARE for the same purpose. Two particular numerical realizations of the framework are presented in detail. Numerical results demonstrate that our approach is effective and comparable to the shifting idea of Guo, Iannazzo, and Meini.

We also propose a natural improvement to the final algorithm in [12], namely ADDA instead of SDA should be used after an appropriate shift is performed on H . The worthiness of doing so is confirmed by our numerical tests.

The last three examples in section 7, all essentially from [2], are special in that X_0 in ADDAs and thus SDAs from their initial setup is exactly Φ . This is a pleasant surprise but should not be expected in general as it does not happen for the first two examples in the section.

The argument in Remark 4.2 about the existence of $\hat{\Psi}$ is inconclusive when $U_{21}\Psi + U_{22}$ is singular. Unfortunately, it is always singular in the critical case as guaranteed by Lemma 4.2. We conjecture that $\hat{\Psi}$ always exists, despite of the inconclusive argument, but a rigorous proof eludes us.

References

- [1] B. D. O. ANDERSON, *Second-order convergent algorithms for the steady-state Riccati equation*, Internat. J. Control, 28 (1978), pp. 295–306.
- [2] N. G. BEAN, M. M. O'REILLY, AND P. G. TAYLOR, *Algorithms for return probabilities for stochastic fluid flows*, Stoch. Models, 21 (2005), pp. 149–184.
- [3] A. BERMAN AND R. J. PLEMMONS, *Nonnegative Matrices in the Mathematical Sciences*, SIAM, Philadelphia, 1994. This SIAM edition is a corrected reproduction of the work first published in 1979 by Academic Press, San Diego, CA.
- [4] D. A. BINI, B. MEINI, AND F. POLONI, *Transforming algebraic Riccati equations into unilateral quadratic matrix equations*, Numer. Math., 116 (2010), pp. 553–578.
- [5] C.-Y. CHIANG, E. K.-W. CHU, C.-H. GUO, T.-M. HUANG, W.-W. LIN, AND S.-F. XU, *Convergence analysis of the doubling algorithm for several nonlinear matrix equations in the critical case*, SIAM J. Matrix Anal. Appl., 31 (2009), pp. 227–247.
- [6] 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 (2005), pp. 55 – 80.
- [7] E. K. W. CHU, H.-Y. FAN, W. W. LIN, AND C. S. WANG, *Structure-preserving algorithms for periodic discrete-time algebraic Riccati equations.*, Internat. J. Control, 77 (2004), pp. 767–788.
- [8] J. DEMMEL, *Applied Numerical Linear Algebra*, SIAM, Philadelphia, PA, 1997.
- [9] C.-H. GUO, *Nonsymmetric algebraic Riccati equations and Wiener-Hopf factorization for M -matrices*, SIAM J. Matrix Anal. Appl., 23 (2001), pp. 225–242.
- [10] ———, *A new class of nonsymmetric algebraic Riccati equations*, Linear Algebra Appl., 426 (2007), pp. 636–649.
- [11] C.-H. GUO AND N. HIGHAM, *Iterative solution of a nonsymmetric algebraic Riccati equation*, SIAM J. Matrix Anal. Appl., 29 (2007), pp. 396–412.

- [12] C.-H. GUO, B. IANNAZZO, AND B. MEINI, *On the doubling algorithm for a (shifted) nonsymmetric algebraic Riccati equation*, SIAM J. Matrix Anal. Appl., 29 (2007), pp. 1083–1100.
- [13] C.-H. GUO AND A. J. LAUB, *On the iterative solution of a class of nonsymmetric algebraic Riccati equations*, SIAM J. Matrix Anal. Appl., 22 (2000), pp. 376–391.
- [14] X. GUO, W. LIN, AND S. XU, *A structure-preserving doubling algorithm for nonsymmetric algebraic Riccati equation*, Numer. Math., 103 (2006), pp. 393–412.
- [15] J. JUANG, *Existence of algebraic matrix Riccati equations arising in transport theory*, Linear Algebra Appl., 230 (1995), pp. 89–100.
- [16] J. JUANG AND W.-W. LIN, *Nonsymmetric algebraic Riccati equations and Hamiltonian-like matrices*, SIAM J. Matrix Anal. Appl., 20 (1998), pp. 228–243.
- [17] P. LANCASTER AND L. RODMAN, *Algebraic Riccati Equations*, Oxford University Press, New York, USA, 1995.
- [18] V. RAMASWAMI, *Matrix analytic methods for stochastic fluid flows*, Proceedings of the 16th International Teletraffic Congress, Edinburg, 1999, Elsevier Science, pp. 19–30.
- [19] L. ROGERS, *Fluid models in queueing theory and Wiener-Hopf factorization of Markov chains*, Ann. Appl. Probab., 4 (1994), pp. 390–413.
- [20] W.-G. WANG, W.-C. WANG, AND R.-C. LI, *ADDA: Alternating-directional doubling algorithm for M-matrix algebraic Riccati equations*, Technical Report 2011-04, Department of Mathematics, University of Texas at Arlington, May 2011. Available at <http://www.uta.edu/math/preprint/>, submitted.
- [21] J. XUE, S. XU, AND R.-C. LI, *Accurate solutions of M-matrix algebraic Riccati equations*, Numer. Math. to appear.
- [22] ———, *Accurate solutions of M-matrix Sylvester equations*, Numer. Math. to appear.