DEFLECTING IRREDUCIBLE SINGULAR $M$-MATRIX ALGEBRAIC RICCATI EQUATIONS

WEI-GUO WANG
School of Mathematical Sciences, Ocean University of China
Qingdao, 266100, China

WEI-CHAO WANG AND REN-CANG LI
Department of Mathematics, University of Texas at Arlington
P.O. Box 19408, Arlington, TX 76019, USA

(Communicated by Xiaoqing Jin)

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 Equation (MARE) is the matrix equation

$$XD - AX - XB + C = 0,$$

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

$$W = \begin{pmatrix} B & -D \\ -C & A \end{pmatrix},$$

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 studied. See [12, 10, 14, 15, 16, 17, 19, 20] and the references therein. It is shown in [10, 14] that MARE (1.1) has a unique minimal nonnegative solution $\Phi$, i.e., in the entrywise sense,

$$\Phi \leq X$$

for any other nonnegative solution $X$ of (1.1).

Recently several doubling algorithms have been proposed to compute $\Phi$ efficiently and accurately. They include the structure-preserving doubling algorithm (SDA)

2010 Mathematics Subject Classification. Primary: 15A24, 65F30; Secondary: 65F10.

Key words and phrases. $M$-Matrix Algebraic Riccati equation, minimal nonnegative solution, deflation, doubling algorithm, convergence analysis.

W.-G. Wang is supported 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 W.-G. Wang was a visiting scholar at Department of Mathematics, University of Texas at Arlington from September 2010 to August 2011. W.-C. Wang and R.-C. Li are supported by the National Science Foundation Grants DMS-0810506 and DMS-1115817.
of Guo, Lin, and Xu [15], the doubling algorithm called SDA-ss of Bini, Meini, and Poloni [4] which combined a shrink-and-shift approach of Ramaswami [19], and the alternating-directional doubling algorithm (ADDA) of Wang, Wang, and Li [22]. 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 [22].

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. Then there exist \( u, x \in \mathbb{R}^m \) and \( v, y \in \mathbb{R}^n \), all entrywise positive vectors, such that
\[
W \left( \begin{array}{c} x \\ y \end{array} \right) = 0, \quad \left( \begin{array}{c} u \\ v \end{array} \right)^T W = 0.
\] (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
\[
H \overset{\text{def}}{=} \begin{pmatrix} I_m & -I_n \end{pmatrix} W = \begin{pmatrix} B & -D \\ C & -A \end{pmatrix}.
\] (1.4)
\( H \) is singular if and only if \( W \) is singular, and (1.3) implies
\[
H \left( \begin{array}{c} x \\ y \end{array} \right) = 0, \quad \left( \begin{array}{c} u \\ -v \end{array} \right)^T H = 0.
\] (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 [13] proposed to shift away its eigenvalue 0 to a properly chosen positive number \( \eta \):
\[
\tilde{H} = H + \eta zw^T,
\]
before SDA is applied, where \( z = \left( \begin{array}{c} x \\ y \end{array} \right) \), 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. The idea of shifting away and that of deflating out known eigenpairs are two common numerical techniques in eigenvalue computations, but often the deflation idea is preferred. 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
\[
W \text{ defined by (1.2) is an irreducible singular } M\text{-matrix, and (1.3) holds, where } 0 < u, x \in \mathbb{R}^m \text{ and } 0 < v, y \in \mathbb{R}^n.
\] (1.6)
Note that assuming (1.3) here is more for notational convenience later than a necessity because \( W \) being an irreducible singular \( M\)-matrix implies the existence of \( 0 < u, x \in \mathbb{R}^m \) and \( 0 < v, y \in \mathbb{R}^n \) that satisfy (1.3).
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 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. \( 1_{n,m} \in \mathbb{R}^{n \times m} \) is the matrix of all ones, and \( 1_n = 1_{n,1} \).

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

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

is the so-called generalized Cayley transformation of \( X \);
3. \( \|X\|_p = \max_{\|x\|_p = 1} \|Xx\|_p \) is the \( \ell_p \)-operator norm of \( X \), where \( \|x\|_p \) is the \( \ell_2 \)-norm of the vector \( x \).

Inequality \( X \preceq Y \) means \( X_{(i,j)} \leq Y_{(i,j)} \) for all \((i,j)\), and similarly for \( X \prec Y \), \( X \succeq Y \), and \( X \succ Y \). In particular, \( X \preceq 0 \) 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 non-positive off-diagonal entries \([3, \text{p.284}]\). 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 \( \Phi \) \([12]\) and its complementary \( M \)-Matrix Algebraic Riccati Equation (cMARE)

\[ YCY - YA - BY + D = 0 \quad \text{(2.1)} \]

is also an irreducible singular MARE and thus has a unique minimal nonnegative solution \( \Psi \), too. Some properties of \( \Phi \) and \( \Psi \) are summarized in Theorem 2.1 below.

**Theorem 2.1** \([10, 11, 12, 13]\). Assume (1.6).

(a): MARE (1.1) has a unique minimal nonnegative solution \( \Phi \), and its cMARE (2.1) has a unique minimal nonnegative solution \( \Psi \);

(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 with\(^1\) \((B - D \Phi)x = 0 \) and \( A - \Phi D \) is a nonsingular \( M \)-matrix, and \( \Phi x = y \), \( \Psi y < x \);

\(^1\)[10, Theorem 4.8] says in this case \( D \Phi x = Dy \) which leads to \((B - D \Phi)x = Bx - Dy = 0 \).
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 $\text{eig}(H) = \{\lambda_1, \ldots, \lambda_{m+n}\}$, where $\lambda_i$‘s are ordered by their nonincreasing real parts, i.e., $\text{Re} \lambda_j \leq \text{Re} \lambda_i$ for $i < j$. Then $\lambda_m$ and $\lambda_{m+1}$ are real, and

\[
\text{Re} \lambda_{m+n} \leq \cdots \leq \text{Re} \lambda_{m+2} < \lambda_{m+1} \leq 0 \leq \lambda_m < \text{Re} \lambda_{m-1} \leq \cdots \leq \text{Re} \lambda_1,
\]

\[
\text{eig}(B - D\Phi) = \text{eig}(B - \Psi C) = \{\lambda_1, \ldots, \lambda_m\},
\]

\[
\text{eig}(A - \Phi D) = \text{eig}(A - C\Psi) = \{-\lambda_{m+1}, \ldots, -\lambda_{m+n}\},
\]

and

\[
\begin{cases}
\lambda_m = 0, \lambda_{m+1} < 0, & \text{if } \mu > 0; \\
\lambda_m = \lambda_{m+1} = 0, & \text{if } \mu = 0; \\
\lambda_m > 0, \lambda_{m+1} = 0, & \text{if } \mu < 0.
\end{cases}
\]

3. **ADDA: Alternating-Direction Doubling Algorithm.** In this section, we briefly review the Alternating-Direction Doubling Algorithm (ADDA). Although it was originally proposed for an MARE \[22\], 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 \[22\] 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 $XDX - AX - XB + C = 0$ to represent a general ARE, while in the rest of this article it is always assumed to be a MARE satisfying (1.6).

Pick some scalars $\alpha$ and $\beta$ (such that all involved inverses exist\(^2\)) and set

\[
A_\beta = A + \beta I, \quad B_\alpha = B + \alpha I,
\]

\[
U_{\alpha\beta} = A_\beta - CB_\alpha^{-1}D, \quad V_{\alpha\beta} = B_\alpha - DA_\beta^{-1}C,
\]

and

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

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

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

\[
E_{k+1} = E_k(I_m - Y_kX_k)^{-1}E_k,
\]

\[
F_{k+1} = F_k(I_n - X_kY_k)^{-1}F_k,
\]

\[
X_{k+1} = X_k + F_k(I_n - X_kY_k)^{-1}X_kE_k,
\]

\[
Y_{k+1} = Y_k + E_k(I_m - Y_kX_k)^{-1}Y_kF_k.
\]

In \[22\], it is derived that

\[
E_k = (I - Y_kX) [\mathcal{E}(R; \beta, \alpha)]^{2^k},
\]

\[
X - X_k = F_kX [\mathcal{E}(R; \beta, \alpha)]^{2^k},
\]

\(^2\)We know how to ensure this for an MARE \[22\].
\[ Y - Y_k = E_k Y \left[ \mathcal{C}(S; \alpha, \beta) \right]^{2^k}, \quad (3.6c) \]
\[ F_k = (I - X_k Y) \left[ \mathcal{C}(S; \alpha, \beta) \right]^{2^k}, \quad (3.6d) \]
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 \( k \to \infty \), \( X_k \) 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 [22], except the convergence for the critical case which was established in [13].

**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 \( \{Y_k\} \):
\[
0 \leq X_k \leq X_{k+1} \leq \Phi, \quad \lim_{k \to \infty} X_k = \Phi,
\]
\[
0 \leq Y_k \leq Y_{k+1} \leq \Psi, \quad \lim_{k \to \infty} Y_k = \Psi,
\]
for all parameters \( \alpha \) and \( \beta \) satisfying
\[
\alpha \geq \alpha_{\text{opt}} \overset{\text{def}}{=} \max_i A_{i,i}, \quad \beta \geq \beta_{\text{opt}} \overset{\text{def}}{=} \max_j C_{j,j},
\quad (3.9)
\]
where \( \Phi \) and \( \Psi \) are the minimal nonnegative solutions of MARE (1.1) and its complementary MARE (2.1), respectively. Moreover, under (3.9),
\[
\limsup_{k \to \infty} \| \Phi - X_k \|^2 \leq \limsup_{k \to \infty} \| \Psi - Y_k \|^2 \leq \rho(\mathcal{C}(S; \alpha, \beta)) \cdot \rho(\mathcal{C}(R; \beta, \alpha)), \quad (3.10)
\]
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 \( \alpha \) and \( \beta \) 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.**
ADD for ARE \( XDX - AX - XB + C = 0 \) and, as a by-product, for cARE \( YCY - YA - BY + D = 0 \).
1. Pick \( \alpha \) and \( \beta \);
2. \( A_{\beta} \overset{\text{def}}{=} A + \beta I, \quad B_{\alpha} \overset{\text{def}}{=} B + \alpha I; \)
3. Compute \( A_{\beta}^{-1} \) and \( B_{\alpha}^{-1}; \)
4. Compute \( V_{\alpha,\beta} \) and \( U_{\alpha,\beta} \) as in (3.3) and then their inverses;
5. Compute \( E_0, \quad F_0, \quad X_0 \) 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 \( k = 1, 2, \ldots, \) until convergence
9. Compute \( E_k \) and \( F_k \) by (3.5a) and (3.5b) (after substituting \( k + 1 \) for \( k \)).
Compute \((I - X_k Y_k)^{-1}\) and \((I - Y_k X_k)^{-1}\);

Compute \(X_{k+1}\) and \(Y_{k+1}\) by (3.5c) and (3.5d);

Enddo

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

\[
ZD^T Z -ZA^T - B^T Z + C^T = 0,
\]

(4.1)

where \(Z = X^T.\) This MARE has the unique minimal nonnegative solution \(\Phi^T,\) and

\[
\begin{pmatrix}
A^T \\
-C^T \\
B^T
\end{pmatrix}
\begin{pmatrix}
v \\
y^T \\
x
\end{pmatrix} = 0,
\]

(4.4)

which is equivalent to

\[
V^{-1} HVV^{-1} \begin{pmatrix} I \\ \phi \end{pmatrix} = V^{-1} \begin{pmatrix} I \\ \phi \end{pmatrix} R.
\]

(4.5)

Partition

\[
V^{-1} = \begin{pmatrix}
m & n \\
m & n
\end{pmatrix}
\begin{pmatrix}
U_{11} & U_{12} \\
U_{21} & U_{22}
\end{pmatrix}
\]

(4.6)

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

\[
V^{-1} HVV^{-1} \begin{pmatrix} I \\ \phi \end{pmatrix} (U_{11} + U_{12}\phi)^{-1}
\]

\[
= V^{-1} \begin{pmatrix} I \\ \phi \end{pmatrix} (U_{11} + U_{12}\phi)^{-1} \left[ (U_{11} + U_{12}\phi) R (U_{11} + U_{12}\phi)^{-1} \right].
\]

(4.7)
Since
\[
V^{-1} \left( \begin{array}{c} I \\ \Phi \end{array} \right) (U_{11} + U_{12} \Phi)^{-1} = \left( \begin{array}{c} I \\ (U_{21} + U_{22} \Phi) (U_{11} + U_{12} \Phi)^{-1} \end{array} \right),
\]
we rewrite (4.7) as
\[
V^{-1} HV \left( \begin{array}{c} I \\ \Phi \end{array} \right) = \left( \begin{array}{c} I \\ \Phi \end{array} \right) \tilde{R},
\]
where
\[
\tilde{\Phi} = (U_{21} + U_{22} \Phi) (U_{11} + U_{12} \Phi)^{-1}, \quad \tilde{R} = (U_{11} + U_{12} \Phi) R (U_{11} + U_{12} \Phi)^{-1}.
\]

**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 $V^{-1} HV e_1 = \delta^{-1} V^{-1} H z = 0$, i.e., the first column of $V^{-1} HV$ is 0. To show $\tilde{\Phi} e_1 = 0$, we notice
\[
\delta e_1 = V^{-1} z = V^{-1} \left( \begin{array}{c} x \\ y \end{array} \right) = V^{-1} \left( \begin{array}{c} x \\ \Phi x \end{array} \right) = V^{-1} \left( \begin{array}{c} I \\ \Phi \end{array} \right) x = \left( \begin{array}{c} (U_{11} + U_{12} \Phi) x \\ (U_{21} + U_{22} \Phi) x \end{array} \right)
\]
which gives
\[
x = \delta (U_{11} + U_{12} \Phi)^{-1} e_1, \quad (U_{21} + U_{22} \Phi) x = 0.
\]
Therefore $\tilde{\delta} \tilde{\Phi} e_1 = (U_{21} + U_{22} \Phi) x = 0$ yielding $\tilde{\Phi} e_1 = 0$, as claimed.

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
\[
V^{-1} HV = \begin{pmatrix} m & n \\ n & m \end{pmatrix} \begin{pmatrix} \tilde{B} & -\tilde{D} \\ \tilde{C} & -\hat{A} \end{pmatrix} = \begin{pmatrix} 1 & m-1 & n \\ 0 & b & -d \\ 0 & \hat{B} & -\hat{D} \\ 0 & \hat{C} & -\hat{A} \end{pmatrix}. \tag{4.12}
\]
In particular,
\[
\tilde{A} = \hat{A}, \quad \tilde{B} = \begin{pmatrix} 1 & m-1 \\ 0 & b \\ 0 & \hat{B} \end{pmatrix}, \quad \tilde{C} = \begin{pmatrix} 1 & m-1 \\ 0 & \hat{C} \end{pmatrix}, \quad \tilde{D} = \begin{pmatrix} d \\ \hat{D} \end{pmatrix}. \tag{4.13}
\]
Equation (4.8) says $\tilde{X} = \tilde{\Phi}$ satisfies the following ARE
\[
\tilde{X} \tilde{D} \tilde{X} - \tilde{A} \tilde{X} - \tilde{X} \tilde{B} + \tilde{C} = 0. \tag{4.14}
\]
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 $\Phi$ of (1.1) can be recovered as follows:
\[
(U_{21} + U_{22} \Phi) (U_{11} + U_{12} \Phi)^{-1} = \tilde{\Phi},
\]
\[
\Rightarrow \quad U_{21} + U_{22} \Phi = \tilde{\Phi} (U_{11} + U_{12} \Phi)
\]
\[
= \tilde{\Phi} U_{11} + \tilde{\Phi} U_{12} \Phi,
\]
\[
\Rightarrow \quad U_{21} - \tilde{\Phi} U_{11} = (-U_{22} + \tilde{\Phi} U_{12}) \Phi.
\]
Thus if $(-U_{22} + \tilde{\Phi} U_{12})^{-1}$ exists, then
\[
\Phi = (-U_{22} + \tilde{\Phi} U_{12})^{-1} (U_{21} - \tilde{\Phi} U_{11}). \tag{4.15}
\]
While this formula suggests that it needs to do two matrix multiplications and to solve \( m \) linear systems of dimension \( n \) to recover \( \Phi \) from \( \tilde{\Phi} \) in general, later we will see for the two realizations in section 5 it actually costs negligibly \( O(m + n) \) and \( O(mn) \) flops (in comparison to the cost that will be incurred by Algorithm 4.1 later for computing \( \tilde{\Phi} \)), respectively.

Lemma 4.1 allows us to write
\[
\tilde{\Phi} = \begin{pmatrix} 0 & \hat{\Phi} \end{pmatrix}, \quad \hat{\Phi} = \tilde{\Phi}_{(:,1:m)}.
\] (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
\[
\begin{aligned}
& \begin{pmatrix} 0 & \hat{\Phi} \end{pmatrix} \begin{pmatrix} d & \hat{\Phi} \\ \hat{D} & 0 \end{pmatrix} \begin{pmatrix} 0 & \hat{\Phi} \\ \hat{D} & 0 \end{pmatrix} - \hat{A} \begin{pmatrix} 0 & \hat{\Phi} \\ \hat{D} & 0 \end{pmatrix} - \begin{pmatrix} 0 & \hat{\Phi} \\ \hat{D} & 0 \end{pmatrix} \begin{pmatrix} 0 & b \\ 0 & \hat{B} \end{pmatrix} + \begin{pmatrix} 0 & \hat{\Phi} \\ \hat{D} & 0 \end{pmatrix} \begin{pmatrix} 0 & \hat{\Phi} \\ \hat{D} & 0 \end{pmatrix} = 0 \\
\iff & \begin{pmatrix} 0 & \hat{\Phi} \end{pmatrix} \hat{D} \hat{\Phi} - \hat{\Phi} \hat{D} \hat{\Phi} - \hat{\Phi} \hat{B} + \hat{\Phi} \hat{C} = 0
\end{aligned}
\]

This says that \( \hat{X} = \hat{\Phi} \) is a solution of the following ARE:
\[
\hat{X} \hat{D} \hat{X} - \hat{A} \hat{X} - \hat{X} \hat{B} + \hat{C} = 0
\] (4.17)

which is equivalent to
\[
\hat{H} \begin{pmatrix} I_{m-1} \\ \hat{X} \end{pmatrix} = \begin{pmatrix} I_{m-1} \\ \hat{X} \end{pmatrix} \begin{pmatrix} \hat{B} - \hat{D} \hat{X} \end{pmatrix}, \quad \hat{H} = m^{-1} \begin{pmatrix} \hat{B} & -\hat{D} \\ \hat{C} & -\hat{A} \end{pmatrix}.
\] (4.18)

The complementary algebraic Riccati equation (cARE) of (4.17) is
\[
\hat{Y} \hat{C} \hat{Y} - \hat{Y} \hat{A} - \hat{B} \hat{Y} + \hat{D} = 0,
\] (4.19)

or equivalently
\[
\hat{H} \begin{pmatrix} \hat{Y} \\ I \end{pmatrix} = \begin{pmatrix} \hat{Y} \\ I \end{pmatrix} \begin{pmatrix} \hat{A} - \hat{C} \hat{Y} \end{pmatrix}.
\]

In the above deflation framework, we assume that both \( U_{11} + U_{12} \Phi, \quad -U_{22} + \hat{\Phi} U_{12} \) are invertible. Later in section 5, this assumption will be verified for the two realizations of this framework there.

**Theorem 4.2.** Assume (1.6) and (4.2). Suppose \( U_{11} + U_{12} \Phi \) is nonsingular, and define \( \Phi \) as in (4.16). Then
\[
\text{eig}(\hat{H}) = \{\lambda_1, \ldots, \lambda_{m-1}, \lambda_{m+1}, \ldots, \lambda_{m+n}\},
\] (4.20)
\[
\text{eig}(\hat{B} - \hat{D} \hat{\Phi}) = \{\lambda_1, \ldots, \lambda_{m-1}\},
\] (4.21)

and cARE (4.19) has a unique solution \( \hat{\Psi} \), if exists, satisfying
\[
\text{eig}(\hat{A} - \hat{C} \hat{\Psi}) = \{-\lambda_{m+1}, \ldots, -\lambda_{m+n}\},
\] (4.22)

where \( \lambda_i \) \((i = 1, \ldots, m + n)\) 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 $\tilde{R}$ 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
\[
\tilde{R}e_1 = (U_{11} + U_{12}\Phi)R(U_{11} + U_{12}\Phi)^{-1}e_1 = \delta^{-1}(U_{11} + U_{12}\Phi)Rx = 0
\]
and thus the partitioning
\[
\tilde{R} = \hat{B} - \hat{D}\hat{\Phi} = \begin{pmatrix} 1 & m-1 \\ 0 & \hat{R}_{12} \\ 0 & R_{22} \end{pmatrix}
\]
which together with (4.13) and $\hat{\Phi} = \begin{pmatrix} 0 & \hat{\Phi} \end{pmatrix}$ give $\tilde{R}_{22} = \hat{B} - \hat{D}\hat{\Phi}$. Since
\[
eig(\tilde{R}) = \eig(R) = \{\lambda_1, \ldots, \lambda_m\}
\]and $0 = \lambda_m < \text{Re}\lambda_{m-1} \leq \cdots \leq \text{Re}\lambda_1$ by Theorem 2.1, we have (4.21).

Let $Z \in \mathbb{R}^{(m+n-1) \times n}$ be a basis matrix of $\tilde{H}$’s invariant subspace associated with the eigenvalues $\lambda_{m+1}, \ldots, \lambda_{m+n}$. If $Z_{(m:m+n-1,:)}$ is invertible, then $\tilde{\Psi}$ exists and is unique, and moreover $\tilde{\Phi} = Z_{(1:m-1,:)}[Z_{(m:m+n-1,:)}]^{-1}$ and (4.22) holds [18].

Theorem 4.3. 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 $\tilde{X} = \tilde{\Phi}$ characterized uniquely by (4.21), and the minimal nonnegative solution $\Phi$ can be recovered by (4.15) with $\hat{\Phi} = \begin{pmatrix} 0 & \hat{\Phi} \end{pmatrix}$.

Proof. The existence of $\tilde{\Phi}$ is a consequence of the constructive deflation procedure above, and $\tilde{\Phi}$ satisfies (4.21) by Theorem 4.2. That this particular solution $\tilde{X} = \tilde{\Phi}$ is uniquely characterized by (4.21) follows from the relation between the solutions of ARE (4.17) and the invariant subspaces of $\tilde{H}$ [18].

Theorem 4.3 suggests a natural way to compute $\Phi$ by first solving ARE (4.17) for $\tilde{\Phi}$ by Algorithm 3.1 and then recovering $\Phi$ 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 $\Phi$ by (4.15) with $\hat{\Phi} = \begin{pmatrix} 0 & \hat{\Phi} \end{pmatrix}$;
6. else
7. compute $\Phi^T$ instead by working with (4.1);
8. Enddo

Remark 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 $\Phi$ 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, where we will also investigate the conditioning of both matrices.
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 $\hat{\Phi}$ have positive real parts and $\hat{\Psi}$ is the unique solution of (4.19) such that all eigenvalues of $\hat{\Phi} - \hat{\Psi}$ 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 $\alpha$ and $\beta$, this kind of verification becomes unnecessary, i.e., Line 4 of Algorithm 4.1 will always produces the right $\hat{\Phi}$.

Remark 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 $\Psi$ of the original cMARE (2.1). We shall do it now. $\Psi$ 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 $\tilde{\Psi} = \tilde{\Psi}$ satisfies the following ARE

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

which is the complementary ARE of (4.14). Partition

$$\tilde{\Psi} \equiv \frac{1}{m-1} \begin{pmatrix} \psi \\ \tilde{\Psi} \end{pmatrix} \quad (4.29)$$

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

$$\begin{pmatrix} \psi \\ \tilde{\Psi} \end{pmatrix} \begin{pmatrix} 0 & \tilde{C} \\ 0 & \tilde{A} \end{pmatrix} \begin{pmatrix} \psi \\ \tilde{\Psi} \end{pmatrix} - \begin{pmatrix} 0 & b \\ 0 & \tilde{B} \end{pmatrix} \begin{pmatrix} \psi \\ \tilde{\Psi} \end{pmatrix} + \begin{pmatrix} d \\ \tilde{D} \end{pmatrix} = 0 \quad (4.30)$$

$$\Leftrightarrow \begin{pmatrix} \psi \tilde{C} \tilde{\Psi} \\ \psi \tilde{C} \tilde{\Psi} \end{pmatrix} - \begin{pmatrix} \psi \tilde{A} \\ \psi \tilde{A} \end{pmatrix} - \begin{pmatrix} b \tilde{\Psi} \\ \tilde{B} \tilde{\Psi} \end{pmatrix} + \begin{pmatrix} d \\ \tilde{D} \end{pmatrix} = 0 \quad (4.31)$$

$$\Leftrightarrow \begin{pmatrix} \psi (\tilde{C} \tilde{\Psi} - \tilde{A}) - b \tilde{\Psi} + d = 0, \\ \psi \tilde{C} \tilde{\Psi} - \psi \tilde{A} - \tilde{B} \tilde{\Psi} + \tilde{D} = 0 \end{pmatrix} \quad (4.32)$$

This says that $\tilde{Y} = \tilde{\Psi}$ is a solution of the complementary ARE (4.19) and $\psi$ satisfies $\psi (\tilde{A} - \tilde{C} \psi) = -b \psi + d$. Thus $\psi$ 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 $U_{21}\Psi + U_{22}$ 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 and in [21].

\textbf{Lemma 4.4.} 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.

\textbf{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 $(U_{21}\Psi + U_{22})y = 0$. \hfill \Box

4.2. \textbf{Convergence Analysis.} Assume, as in ADDA for the original MARE (1.1), that

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

By Theorem 4.2, $\tilde{\mathcal{X}} = \hat{\Phi} = \hat{\Phi}_{(\alpha,2m)}$ and $\hat{\Psi}$ are such that

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

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

\textbf{Lemma 4.5.} Assume (1.6) and (3.9). Let $R = B - D\Phi$ and $S = A - C\Psi$, and $\hat{R}$ and $\hat{S}$ as given by (4.30). Then

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

and in particular

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

\textbf{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 $\rho(\mathcal{E}(S; \alpha, \beta)) \cdot \rho(\mathcal{E}(R; \beta, \alpha)) \leq 1$ 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 4.2 that

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

Thus $\rho(\mathcal{E}(\hat{S}; \alpha, \beta)) = \rho(\mathcal{E}(S; \alpha, \beta))$. The proof of [22, Theorem 2.1] implies that

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

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

\[
\rho(\mathcal{C}(\hat{R}; \beta, \alpha)) < \rho(\mathcal{C}(R; \beta, \alpha)),
\]

as was to be shown. \(\Box\)

**Theorem 4.6.** Assume (1.6) and (4.2). Suppose \(U_{11} + U_{12}\Phi\) is nonsingular. Let \(\{\hat{E}_k\}, \{\hat{F}_k\}, \{\hat{X}_k\}, \{\hat{Y}_k\}\) be the sequences generated by ADDA applied to (4.17) with no breakdowns, i.e., all involved inverses exist. If (3.9) holds, then \(\hat{X}_k\) and \(\hat{Y}_k\) converge quadratically to \(\Phi\) and \(\Psi\), respectively, and

\[
\limsup_{k \to \infty} \|\hat{\Phi} - \hat{X}_k\|^{1/2k} \leq \rho(\mathcal{C}(\hat{S}; \alpha, \beta)) \cdot \rho(\mathcal{C}(\hat{R}; \beta, \alpha)) < 1, \tag{4.33a}
\]

\[
\limsup_{k \to \infty} \|\hat{\Psi} - \hat{Y}_k\|^{1/2k} \leq \rho(\mathcal{C}(\hat{R}; \beta, \alpha)) \cdot \rho(\mathcal{C}(\hat{S}; \alpha, \beta)) < 1, \tag{4.33b}
\]

where \(\| \cdot \|\) is any matrix norm.

**Proof.** Inequalities in (4.33) are the consequences of

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

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

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

\[
(\hat{\Phi} - \hat{X}_k) \left(I - \hat{\Psi} \left[\mathcal{C}(\hat{S}; \alpha, \beta)\right]^{2k} \hat{\Phi} \left[\mathcal{C}(\hat{R}; \beta, \alpha)\right]^{2k}\right) = (I - \hat{\Phi}\hat{\Psi}) \left[\mathcal{C}(\hat{S}; \alpha, \beta)\right]^{2k} \hat{\Phi} \left[\mathcal{C}(\hat{R}; \beta, \alpha)\right]^{2k}. \tag{4.35}
\]

Since by Lemma 4.5

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

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

\[
\|\hat{\Phi} - \hat{X}_k\|^{1/2k} \leq \|(I - \Gamma)^{-1}\|^{1/2k} \|I - \hat{\Phi}\hat{\Psi}\|^{1/2k}
\times \left\|\mathcal{C}(\hat{S}; \alpha, \beta)\right\|^{2k} \left\|\hat{\Phi}\right\|^{1/2k} \left\|\mathcal{C}(\hat{R}; \beta, \alpha)\right\|^{2k} \left\|\hat{\Psi}\right\|^{1/2k}. \tag{4.36}
\]

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

\[
\|(I - \Gamma)^{-1}\|^{1/2k} \to 1, \quad \|I - \hat{\Phi}\hat{\Psi}\|^{1/2k} \to 1, \quad \|\hat{\Phi}\|^{1/2k} \to 1,
\]

\[
\left\|\mathcal{C}(\hat{S}; \alpha, \beta)\right\|^{2k} \rightarrow \rho(\mathcal{C}(\hat{S}; \alpha, \beta)), \quad \left\|\mathcal{C}(\hat{R}; \beta, \alpha)\right\|^{2k} \rightarrow \rho(\mathcal{C}(\hat{R}; \beta, \alpha)).
\]

\(\text{We assume } \| \cdot \| \text{ is a consistent matrix norm. This does not lose any generality because all matrix norms are equivalent and thus } \limsup_{k \to \infty} \|\hat{\Phi} - \hat{X}_k\|^{1/2k} \text{ does not change with the norm used.}\)
That $\hat{X}_k$ and $\hat{Y}_k$ converge quadratically to $\hat{\Phi}$ and $\hat{\Psi}$, respectively, is a consequence of the inequalities in (4.33).

Remark 3. A few comments are in order:

1. If $\mu \neq 0$, ADDA applied to the original MARE (1.1) is already quadratically convergent [22]. But it is only linearly convergent if $\mu = 0$ [5]. Theorem 4.6 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 $\{\hat{X}_k\}$ and $\{\hat{Y}_k\}$ generated by ADDA applied to (4.17).

3. Theorem 3.1 says that under (3.9) $\rho(\mathcal{C}(\hat{S};\alpha,\beta)) = \rho(\mathcal{C}(\hat{R};\beta,\alpha))$ is minimized at $\alpha = \alpha_{\text{opt}}$ and $\beta = \beta_{\text{opt}}$, leading to the optimal ADDA in [22]. For the current case, for fast convergence we should pick $\alpha$ and $\beta$ such that $\rho(\mathcal{C}(\hat{S};\alpha,\beta)) \cdot \rho(\mathcal{C}(\hat{R};\beta,\alpha))$ is minimized subject to (3.9). While it is not clear whether $\rho(\mathcal{C}(\hat{S};\alpha,\beta)) \cdot \rho(\mathcal{C}(\hat{R};\beta,\alpha))$ 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.

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

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

(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

$$ L^{-1} = \begin{pmatrix} 1 & 0 \\ -\hat{z} & I_{m+n-1} \end{pmatrix}, \quad L = \begin{pmatrix} 1 & 0 \\ \hat{z} & I_{m+n-1} \end{pmatrix}, $$

(5.2a)

$$ V^{-1} = L^{-1}P, \quad V = P^TL, $$

(5.2b)

where

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

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 $i_0 = \text{argmax}_i z(i)$, and a partial pivoting if $i_0 = \text{argmax}_{1 \leq i \leq m} z(i)$. 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 as well as those in the literature, $z = 1_{m+n}$ and thus it makes no difference with or without a pivoting strategy.

We can write

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

(5.3)

Partition

$$ L^{-1} = \begin{pmatrix} L_{11} & 0 \\ L_{21} & I \end{pmatrix}, \quad w = \begin{pmatrix} w_1 \\ w_2 \end{pmatrix}, \quad P = \begin{pmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{pmatrix}. $$
Use (5.2a) and (5.3) to see
\[ L_{11} = \begin{pmatrix} 1 & 0 \\ \hat{z}_{(1-m-1)} & I_{m-1} \end{pmatrix}, \quad L_{21} = -\hat{z}_{(m+n+1)}e_1^T, \quad \text{(5.4a)} \]
\[ L = \begin{pmatrix} L_{11}^{-1} & 0 \\ -L_{21}^{-1} & I \end{pmatrix}, \quad L_{11}^{-1} = \begin{pmatrix} 1 & 0 \\ \hat{z}_{(1-m-1)} & I_{m-1} \end{pmatrix}, \quad \text{(5.4b)} \]
\[ P_{ii} = I - w_1w_1^T, \quad P_{ij} = -w_iw_j^T \text{ for } i \neq j. \quad \text{(5.4c)} \]

So the four submatrices $U_{ij}$ of $V^{-1} = L^{-1}P$ partitioned as in (4.6) are
\[ U_{11} = L_{11}(I - w_1w_1^T), \quad U_{12} = -L_{11}w_1w_2^T, \quad \text{(5.5a)} \]
\[ U_{21} = L_{21}(I - w_1w_1^T) - w_2w_1^T, \quad U_{22} = -L_{21}w_1w_2^T + I - w_2w_2^T. \quad \text{(5.5b)} \]

Equations (4.9) and (4.15) that relate $\Phi$ and $\tilde{\Phi}$ remain valid, provided that $U_{11} + U_{12}\Phi$ and $-U_{22} + \Phi U_{12}$ are invertible, as ensured by Theorem 5.2 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 $I_q - F^TE$ is nonsingular. Moreover
\[ (I_p - EF^T)^{-1} = I_p + E(I_q - F^TE)^{-1}F^T. \]

**Theorem 5.2.** 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 $\Phi$ by (4.9).

**Proof.** We have by (5.5)
\[ U_{11} + U_{12}\Phi = L_{11}(I - w_1w_1^T) - L_{11}w_1w_2^T\Phi \]
\[ = L_{11}\left[ I - w_1(w_1^T + w_2^T\Phi) \right]. \]

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 and only if
\[ \zeta \overset{\text{def}}{=} 1 - (w_1^T + w_2^T\Phi)w_1 \neq 0. \]

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
   \[ \zeta = 1 - (w_1^T + w_2^T\Phi)w_1 = 1 - w_1^Tw_1 = -1 < 0; \]
3. If $i_0 > m$, then $w_1 = e_1$ and $w_2 = -e_1 - m$ 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
\[ (U_{11} + U_{12}\Phi)^{-1} = \left[ I + \zeta^{-1}w_1(w_1^T + w_2^T\Phi) \right] L_{11}^{-1} \]
\[ = \begin{cases} L_{11}^{-1}, & \text{for } i_0 = 1, \\ I - w_1w_1^T L_{11}^{-1}, & \text{for } 1 < i_0 \leq m, \\ \left[ I + \Phi_{(i_0-m,1)}^{-1}e_1(e_1^T - \Phi_{(i_0-m,1)}) \right] L_{11}^{-1}, & \text{for } m < i_0. \end{cases} \quad \text{(5.6)} \]
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

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

(5.9)

Therefore by (5.8) and (5.10)

\[
\tilde{\Phi} L_{11} w_1 w_2^T = (U_{21} + U_{22} \Phi) (U_{11} + U_{12} \Phi)^{-1} L_{11} w_1 w_2^T = [L_{21} (I - \tilde{\Phi}_1 w_1^T) - w_2 w_1^T + (-L_{21} w_1 w_2^T + I - w_2 w_2^T) \Phi] \zeta^{-1} w_1 w_2^T
\]

\[
= \zeta^{-1} [-w_2 w_2^T + \zeta L_{21} w_1 w_2^T + \Phi w_1 w_2^T + \zeta w_2 w_2^T]
\]

\[
= (1 - \zeta^{-1}) w_2 w_2^T + L_{21} w_1 w_2^T + \zeta^{-1} \Phi w_1 w_2^T.
\]

Combine this with (5.7) to get

\[
-U_{22} + \tilde{\Phi} U_{12} = -I + \zeta^{-1} w_2 w_2^T - \zeta^{-1} \Phi w_1 w_2^T = -[I - \zeta^{-1} (w_2 - \Phi w_1) w_2^T]
\]

(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

\[
(-U_{22} + \tilde{\Phi} U_{12})^{-1} = \begin{cases} 
I, & \text{for } i_0 \leq m, \\
[I - \phi_{(i_0 - m, e_1)}]^{-1}, & \text{for } i_0 > m.
\end{cases}
\]

(5.12)

This completes the proof.

The inversion formulas (5.6) and (5.12), together with (5.4) and (5.5), lead to fast algorithms via (4.9) and (4.15) to go from one of \(\Phi\) and \(\tilde{\Phi}\) to the other at the cost of \(O(m + n)\) flops. The numerical stability of going from \(\tilde{\Phi}\) to \(\Phi\) this way depends on \(\|U_{11} + U_{12} \Phi\|_1 \|(U_{11} + U_{12} \Phi)^{-1}\|_1\) and \(\|U_{22} + \tilde{\Phi} U_{12}\|_1 \|(U_{22} + \tilde{\Phi} U_{12})^{-1}\|_1\) for which we have, provided \(|\tilde{z}_{(i)}| \leq 1\) for \(1 \leq i \leq m + n - 1\),

\[
\|U_{11} + U_{12} \Phi\|_1 \leq \begin{cases} 
(m + 1), & \text{if } i_0 \leq m, \\
(m + 1) \left(1 + \max_{1 \leq j \leq m} \phi_{(i_0 - m, j)}\right), & \text{if } i_0 > m.
\end{cases}
\]

(5.13a)

\[
\|U_{11} + U_{12} \Phi\|^{-1}_1 \leq \begin{cases} 
(m + 1), & \text{if } i_0 \leq m, \\
(m + 1) \left(\phi_{(i_0 - m, 1)}^{-1} + \max_{2 \leq j \leq m} \phi_{(i_0 - m, j)}\right), & \text{if } i_0 > m.
\end{cases}
\]

(5.13b)
and
\[
\| - U_{22} + \Phi U_{12} \|_1 = \begin{cases} 
1, & \text{if } i_0 \leq m, \\
1 + \Phi_{(i_0-m,1)}(1 + \| \Phi_{(1,:)} \|_1), & \text{if } i_0 > m, 
\end{cases} \quad (5.14a)
\]
\[
\| ( - U_{22} + \Phi U_{12} )^{-1} \|_1 = \begin{cases} 
1, & \text{if } i_0 \leq m, \\
1 + \| \Phi_{(1,:)} \|_1, & \text{if } i_0 > m. 
\end{cases} \quad (5.14b)
\]

In particular, if \( i_0 \leq m \), all bounds by (5.13) and (5.14) are independent of \( \Phi \) and \( \Phi \), and thus calculating \( \Phi \) or \( \hat{\Phi} \) via (4.9) or (4.15) is numerically stable.

It is rather straightforward to extract \( A, B, C, \) and \( D \) from
\[
V^{-1} HV = (I - \hat{\varepsilon} e_1^T) PHP(I + \hat{\varepsilon} e_1^T) \\
\quad = PHP - \hat{\varepsilon}(e_1^T PHP)\hat{\varepsilon} + (e_1^T PHP\hat{\varepsilon})\hat{\varepsilon} e_1^T. \quad (5.15)
\]
where \( \hat{\varepsilon} = (0, \hat{z}^T)^T \). The right-hand side of (5.15) lends itself for a fast evaluation of \( V^{-1} HV \). In the case \( i_0 = 1 \), we have\(^4\)
\[
\tilde{\Phi} = (0 \Phi(:,2:m)), \quad \hat{\Phi} = \Phi(:,2:m), \quad \Phi_{(1,:)} = x_{(1)}^{-1} \left[ y - \Phi(:,2:m)x_{(2:m)} \right], \quad (5.16)
\]
and
\[
\tilde{B} = B(:,2:m,m) - x_{(1)}^{-1} x_{(2:m)}B(:,1,m), \quad \hat{D} = D(:,2:m,:) - x_{(1)}^{-1} x_{(2:m)}D(:,1,:), \quad (5.17a)
\]
\[
\tilde{C} = C(:,2:m) - x_{(1)}^{-1} yB(:,2:m), \quad \hat{A} = A - x_{(1)}^{-1} yD(:,1,:). \quad (5.17b)
\]

Note also in this case
\[
\hat{A} - \hat{\Phi} \hat{D} = A - \Phi D. \quad (5.18)
\]
This is because \( \Phi_{(1,:)} = x_{(1)}^{-1} \left[ y - \Phi x_{(2:m)} \right] \), and thus
\[
\hat{A} - \hat{\Phi} \hat{D} = A - x_{(1)}^{-1} yD(:,1,:) - \hat{\Phi}(D(:,2:m,:) - x_{(1)}^{-1} x_{(2:m)}D(:,1,:)) \\
= A - x_{(1)}^{-1} \left[ y - \Phi x_{(2:m)} \right] D(:,1,:) - \hat{\Phi} D(:,2:m,:) \\
= A - \Phi_{(1,:)} D(:,1,:) - \hat{\Phi} D(:,2:m,:) \\
= A - \Phi D.
\]

In Remark 2, we show \( \hat{\Psi} \) exists if \( U_{21}\Psi + U_{22} \) is nonsingular, and in Lemma 4.4 we show \( U_{21}\Psi + U_{22} \) is always singular if \( \mu = 0 \). Theorem 5.3 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 5.3.** Let \( U_{ij} \) be defined by (5.1) – (5.5). Then \( U_{21}\Psi + U_{22} \) 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 4.4. 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
\[
U_{21}\Psi + U_{22} = \left[ L_{21}(I - w_1 w_1^T) - w_2 w_1^T \right] \Psi - L_{21}w_1 w_2^T + I - w_2 w_2^T. \quad (5.19)
\]
There are three cases to consider.

---

\(^4\)This is not a misprint: the last \( m - 1 \) columns of \( \hat{\Phi} \) are the same as those of \( \Phi \).
1. If \( i_0 = 1 \), then \( w_1 = 0 \) and \( w_2 = 0 \) and thus (5.19) becomes

\[
L_{21} \Psi + I = -x_{(1)}^{-1} y e_1^T \Psi + I
\]

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 \( P_1 = I - w_1 w_1^T \) which is the permutation matrix that swaps the first entry and the \( i_0 \)th entry of \( x \). (5.19) becomes

\[
L_{21} (I - w_1 w_1^T) \Psi + I = -x_{(i_0)}^{-1} y e_1^T P_1 \Psi + I
\]

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 \( w_2 = -e_{i_0} \), where \( j_0 = i_0 - m \). We have

\[
L_{21} = -\hat{y} e_1^T, \quad \hat{y} = y_{(j_0)} - y_{(i_0)} + y_{(j_0)} x_{(1)} e_{j_0}.
\]

It can be verified that \( L_{21} (I - w_1 w_1^T) = 0 \). Therefore

\[
U_{21} \Psi + U_{22} = -w_2 w_1^T \Psi - L_{21} w_1 w_1^T + I - w_2 w_2^T = e_{j_0} e_1^T \Psi - \hat{y} e_{j_0} + I - e_{j_0} e_{j_0} = I - (\hat{y} + e_{j_0}) e_{j_0} + e_{j_0} e_1^T \Psi
\]

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

\[
I_2 - \begin{pmatrix} \hat{y} e_{j_0} & I_{j_0}^T \end{pmatrix} \begin{pmatrix} \hat{y} + e_{j_0} & -e_{j_0} \end{pmatrix}
\]

is invertible. Use \( \hat{y} + e_{j_0} = y_{(j_0)} y - y_{(j_0)} x_{(1)} e_{j_0} \) to simplify the matrix (5.20) to

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

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

This completes the proof. □

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

\[
Q = \begin{pmatrix} m \\ n \end{pmatrix} \begin{pmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{pmatrix}.
\]

(5.21)
Then $V^{-1} = Q^T$ gives $U_{ij} = Q_{ij}^T$, and consequently
\[ \tilde{\Phi} = (Q_{11}^T + Q_{22}^T \Phi)(Q_{11}^T + Q_{22}^T \Phi)^{-1}, \]  
\[ \Phi = (-Q_{22}^T + \tilde{\Phi} Q_{21}^T)^{-1}(Q_{12}^T - \tilde{\Phi} Q_{11}^T), \]  
(5.22a) (5.22b)
assuming $Q_{11}^T + Q_{22}^T \Phi$ and $-Q_{22}^T + \tilde{\Phi} Q_{21}^T$ are invertible. We know $\tilde{\Phi}e_1 = 0$ by Lemma 4.1, and $\tilde{\Phi} = \tilde{\Phi}(.,.,.)$ 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 invertible by Theorem 5.4 below.

The Householder transformation $V = Q$ such that $Qz = -\|z\|_2 e_1$ is given by
\[ Q = I - 2ww^T, \quad w = \frac{z - \delta e_1}{\|z - \delta e_1\|_2}, \]  
(5.23)
where
\[ \delta = -\|z\|_2, \quad \gamma = \|z - \delta e_1\|_2 = \sqrt{2x(1)\|z\|_2 + 2\|z\|_2^2}. \]  
(5.24)
Partition $w = (w_1 \ w_2)$, where
\[ 0 < w_1 = \gamma^{-1}(x - \delta e_1) \in \mathbb{R}^m, \quad 0 < w_2 = \gamma^{-1} y \in \mathbb{R}^n. \]  
(5.25)
Then the four submatrices $Q_{ij}$ as defined by (5.21) are
\[ Q_{11} = I_m - 2w_1 w_1^T, \quad Q_{12} = -2w_1 w_2^T, \]  
\[ Q_{22} = I_n - 2w_2 w_2^T, \quad Q_{21} = -2w_2 w_1^T. \]  
(5.26a) (5.26b)

**Theorem 5.4.** Let $Q \in \mathbb{R}^{(m+n)\times(m+n)}$ be the Householder transformation as given by (5.23) and (5.24). Then both $Q_{11}^T + Q_{21}^T \Phi$ and $-Q_{22}^T + \tilde{\Phi} Q_{21}^T$ are invertible, where $\Phi$ relates to $\Phi$ by (5.22a).

**Proof.** We have (5.23) = (5.26), and thus
\[ Q_{11}^T + Q_{21}^T \Phi = I_m - 2w_1 w_1^T - 2w_1 w_2^T \Phi = I_m - 2w_1(w_1^T + w_2^T \Phi). \]
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
\[ \zeta \overset{\text{def}}{=} 1 - 2(w_1^T + w_2^T \Phi)w_1 \]  
\[ = 1 - 2w_1^T w_1 - 2w_2^T \Phi w_1 \]  
\[ = 1 - 2 \frac{\|x - \delta e_1\|_2^2}{\gamma^2} - 2w_2^T \Phi w_1 \]  
(5.27)
\[ \overset{\text{def}}{=} 1 - 2(\|x\|_2^2 + \|\delta e_1\|_2^2). \]

---

\[ ^5 \text{This is not so for the Householder transformation such that } Qz = \|z\|_2 e_1. \text{ For example, } m = n = 2, B = \begin{pmatrix} 3 & -1 \\ -1 & 3 \end{pmatrix}, D = 1_{2,2}, A = B, \text{ and } C = D. \text{ For this example } W_{14} = 0, 1^T W = 0, \Phi = \frac{1}{2} 1_{2,2}, \psi = \frac{1}{2} 1_{2,2}, \text{ and thus } \mu = 0. \text{ We have } Q_{11}^T + Q_{21}^T \Phi = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \text{ for the Householder transformation } Q \text{ such that } Q_{14} = 2e_1, \text{ but } Q_{11}^T + Q_{21}^T \Phi = \begin{pmatrix} -1 & -1 \\ -2/3 & 2/3 \end{pmatrix} \text{ for the Householder transformation } Q \text{ such that } Q_{14} = -2e_1. \]
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

\[
w_2^T (Q_{12}^T + Q_{22}^T \Phi) = w_2^T \left[ -2w_2 w_2^T + (I - 2w_2 w_2^T) \Phi \right]
\]

\[
= (-2w_2^T w_2)w_2^T + (1 - 2w_2^T w_2)w_2^T \Phi,
\]

\[
(Q_{11}^T + Q_{21}^T \Phi)^{-1} w_1 = \left[ 1 + \frac{2w_1^T w_1 + 2w_1^T \Phi w_1}{1 - 2w_1^T w_1 - 2w_2^T \Phi w_1} \right] w_1
\]

\[
= \frac{1}{1 - 2w_1^T w_1 - 2w_2^T \Phi w_1} w_1,
\]

\[
w_2^T \tilde{\Phi} w_1 = w_2^T (Q_{12}^T + Q_{22}^T \Phi) \cdot (Q_{11}^T + Q_{21}^T \Phi)^{-1} w_1
\]

\[
= \frac{(-2w_2^T w_2)w_2^T w_1 + (1 - 2w_2^T w_2)w_2^T \Phi w_1}{1 - 2w_1^T w_1 - 2w_2^T \Phi w_1},
\]

\[
w_2^T w_2 - w_2^T \tilde{\Phi} w_1 = \frac{w_2^T w_2 - w_2^T \Phi w_1}{1 - 2w_1^T w_1 - 2w_2^T \Phi w_1}
\]

\[
1 - 2(w_2^T w_2 - w_2^T \tilde{\Phi} w_1) = \frac{1 - 2w_1^T w_1 - 2w_2^T w_2}{1 - 2w_1^T w_1 - 2w_2^T \Phi w_1}
\]

\[
= \frac{1}{1 - 2w_1^T w_1 - 2w_2^T \Phi w_1} > 0,
\]
as expected. \( \Box \)

**Remark 4.** Theorem 5.4 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 \( \Phi \) by (5.22a). Since \( \Phi x = y \) is never referenced, it is not required. \( \Diamond \)

The above proof also yields

\[
(Q_{11}^T + Q_{21}^T \Phi)^{-1} = I_m + 2\zeta^{-1} w_1 (w_1^T + w_2^T \Phi),
\]

(5.29a)

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

(5.29b)

where \( \zeta \) is defined by (5.27). With the help of (5.29), we can express any one of \( \Phi \) and \( \tilde{\Phi} \) in terms of the other via a rank-one update. Details are as follows. By (5.22), we have

\[
\tilde{\Phi} = \left[ -2w_2 w_1^T + (I - 2w_2 w_2^T) \Phi \right] \left[ I + 2\zeta^{-1} w_1 (w_1^T + w_2^T \Phi) \right]
\]
\[ = [\Phi - 2w_2(w_1^T + w_2^T \Phi)] \left[ I + 2\zeta^{-1}w_1(w_1^T + w_2^T \Phi) \right] \]
\[ = \Phi + 2\zeta^{-1}\Phi w_1(w_1^T + w_2^T \Phi) \]
\[ - 2w_2(w_1^T + w_2^T \Phi) - 4\zeta^{-1}w_2(w_1^T + w_2^T \Phi)w_1(w_1^T + w_2^T \Phi) \]
\[ = \Phi + 2 \{ \zeta^{-1}\Phi w_1 - [1 + 2\zeta^{-1}(w_1^T + w_2^T \Phi)w_1] w_2 \} (w_1^T + w_2^T \Phi) \]
\[ \Phi = \left[ -I_n + 2\zeta(w_2 - \Phi w_1)w_2^T \right] \left[ -2w_2w_1^T - \Phi(I - 2w_1w_1^T) \right] \]
\[ = \Phi + 2(w_2 - \Phi w_1)w_1^T \]
\[ = \Phi + 2 \{ 1 - 2\zeta w_2^T (w_2 - \Phi w_1) \} (w_2 - \Phi w_1)w_1^T - 2\zeta(w_2 - \Phi w_1)w_2 \]
\[ = \Phi + 2(w_2 - \Phi w_1) \left\{ 1 - 2\zeta w_2^T (w_2 - \Phi w_1) \right\} w_1^T - \zeta w_2 \Phi \]
\[ (5.30b) \]

Equation (5.30b) will become handy in coding up Algorithm 4.1, where recovering \( \Phi \) is needed from computed \( \tilde{\Phi} \) by (5.30b) with \( \Phi = \left[ \begin{array}{c} 0 \\ \tilde{\Phi} \end{array} \right] \). Equation (5.30a) expresses \( \Phi \) in terms of \( \Phi \). The cost of getting one of \( \Phi \) and \( \tilde{\Phi} \) from the other is only \( O(nm) \) flops. The numerical stability of doing so depends on \( \|Q_{11}^T + Q_{21}^T \Phi\|_2 \|Q_{11}^T + Q_{21}^T \Phi\|^{-1} \|Q_{22} + \Phi Q_{21}^T \|_2 \| - Q_{22} + \Phi Q_{21}^T \|^{-1} \|_2 \) for which we have, upon using \( \|w_i\|_2 \leq 1 \) for \( i = 1, 2 \),
\[ \|Q_{11}^T + Q_{21}^T \|_2 \leq 1 + 2(1 + \|\Phi\|_2), \quad (5.31a) \]
\[ \|Q_{11}^T + Q_{21}^T \Phi\|^{-1} \|_2 \leq 1 + 2(1 + \|\Phi\|_2), \quad (5.31b) \]
\[ \| - Q_{22} + \Phi Q_{21}^T \|_2 \leq 1 + 2(1 + \|\Phi\|_2), \quad (5.31c) \]
\[ \| - Q_{22} + \Phi Q_{21}^T \|^{-1} \|_2 \leq 1 + 2\zeta(1 + \|\Phi\|_2). \quad (5.31d) \]

Lower and upper bound on \( \|\zeta\| \) can be easily gotten from (5.28), for example
\[ \|x\|_2^2 / \|z\|_2^2 \leq \|\zeta\| \leq 1 + 2\|\Phi\|_2. \]

Thus calculating \( \Phi \) or \( \tilde{\Phi} \) via (5.22) is numerically stable unless \( \|x\|_2^2 \ll \|z\|_2^2 \).

Extractions of the coefficient matrices \( \tilde{A}, \tilde{B}, \tilde{C}, \) and \( \tilde{D} \) for ARE (4.17) can be easily done from the partitioning (4.12) for
\[ V^{-1}HV = (I - 2ww^T)H(I - 2ww^T) \]
\[ = H - 2ww^TH - 2HWw^T + 4(w^THw)ww^T, \quad (5.32) \]
where the expression in the right-hand side of (5.32) suggests an economical way to numerically compute \( V^{-1}HV \).

In Remark 2, we show \( \tilde{\Psi} \) exists if \( U_{21}\Psi + U_{22} \) is nonsingular, and in Lemma 4.4 we show \( U_{21}\Psi + U_{22} \) is always singular if \( \mu = 0 \). Theorem 5.5 asserts that \( U_{21}\Psi + U_{22} \) is guaranteed nonsingular if \( \mu \neq 0 \). Thus the existence of \( \tilde{\Psi} \) is unresolved for the case \( \mu = 0 \), but otherwise \( \tilde{\Psi} \) exists. We point out that \( \tilde{\Psi} \) does exist for all our critical case examples in section 7 though.
Proof. We have by (5.26) and $U_{ij} = Q_{ij}^{T}$ 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 (5.24) and (5.25) and that $\Psi y < x$ for $\mu > 0$ and $\Psi y = x$ for $\mu = 0$. We have

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

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 $\mu = 0$. \hfill \Box

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 [13] 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 $\mu = u^Tx - v^Ty \geq 0$. Pick $\eta \in \mathbb{R}$ to be specified in a moment, and let

$$\widehat{H} = H + \eta z w^T \equiv m \begin{pmatrix} \widehat{B} & -\widehat{D} \\ \widehat{C} & -\widehat{A} \end{pmatrix},$$

(6.1)

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

$$\widehat{X} \widehat{D} \widehat{X} - \widehat{A} \widehat{X} - \widehat{X} \widehat{B} + \widehat{C} = 0.$$  

(6.2)

It is proved in [13] that $\widehat{X} = \Phi$ is the solution of (6.2) uniquely characterized by

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

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

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

where

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

In solving (6.2) by SDA [15], Guo, Iannazzo, and Meini [13] picked

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

(6.3)

for simplicity, and

$$\alpha = \beta = \eta = \max\{\alpha_{\text{opt}}, \beta_{\text{opt}}\}$$  

(6.4)

to ensure $\eta \in \text{eig}(\widehat{R})$ contributes nothing to $\rho(\mathcal{E}(\widehat{R}; \eta, \eta))$, where $\alpha_{\text{opt}}$ and $\beta_{\text{opt}}$ are as in (3.9).

\footnote{Recall that SDA is ADDA (Algorithm 3.1) after setting $\alpha = \beta$, and its rate of convergence is determined by $\rho(\mathcal{E}(\widehat{S}; \alpha, \alpha)) \cdot \rho(\mathcal{E}(\widehat{R}; \alpha, \alpha))$.}
It has been noted [22] that compared to ADDA, SDA will experience slow convergence if $\alpha_{\text{opt}}$ and $\beta_{\text{opt}}$ 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 $\rho(\mathcal{C}(\hat{S}; \alpha, \beta)) \cdot \rho(\mathcal{C}(\hat{R}; \beta, \alpha))$, and we will pick $\alpha = \alpha_{\text{opt}}$, $\eta = \beta = \beta_{\text{opt}}$, $\eta_{\text{opt}}$, (6.5) as discussed in Remark 3 and to make sure $\eta \in \text{eig}(\hat{R})$ contributes nothing to $\rho(\mathcal{C}(\hat{R}; \beta, \alpha))$.

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 three 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 $\Phi$:

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

(7.1)

which are not available in actual computations but is made available here for testing purpose. The use of $\ell_1$-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 EREErr, the indeterminant $0/0$ is treated as $0$. These errors defined in (7.1) and (7.2) are 0 if $\Phi$ is exact, but numerically they can only be made as small as $O(u)$, where $u$ is the unit machine roundoff.

In [22, 23], it was argued that the doubling algorithms SDA [15, 13], SDA-ss [4], and ADDA [22] 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 $\Phi$ may not have deserved entrywise relative accuracy if some of the entries of $\Phi$ are very tiny relative to others, even though NRes is reduced to the level of $O(u)$. For this reason, we will use $\text{NRes} \leq 5 \times 10^{-14}$ as the stopping criteria in our tests here, instead of Kahan’s criteria [24, 22] designed to stop the iterations only when $\Phi$ is computed to its deserved entrywise relative accuracy.

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

1. ADDA of [22]. We use it as a representative for all doubling algorithms derivable from bilinear transformations, including SDA [15, 13] and SDA-ss [4], since ADDA is the fastest among all [22].
2. SDAs of [13] (as outlined in section 6). It is the first method ever proposed to improve SDA for irreducible singular MAREs.
For testing purpose, we computed for computerized algebra system $W(\lfloor x \rfloor)$ which is Algorithm 4.1 combined with the elimination approach in subsection 5.1. For simplicity, all $t_0 = 1$. Actually in all examples, $z = 1_{m+n}$; so there is no need to do pivoting to control $\|V\|_1\|V^{-1}\|_1$.

3. ADDAs (as outlined in section 6). Since ADDA improves SDA, naturally we would expect ADDAs improves SDAs.

4. dADDAe which is Algorithm 4.1 combined with the elimination approach in subsection 5.1. For simplicity, all $t_0 = 1$. Actually in all examples, $z = 1_{m+n}$; so there is no need to do pivoting to control $\|V\|_1\|V^{-1}\|_1$.

5. dADDAq which is Algorithm 4.1 combined with the Householder transformation approach in subsection 5.2.

Example 7.1 ([22, Example 7.2]).

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

$W$ is an irreducible singular $M$-matrix:

$$W1_{2n} = 0, \quad \left( \begin{pmatrix} 1 \end{pmatrix}_{n} \right)^T W = 0, \quad \mu = (1 - \xi^{-1})n.$$  

For testing purpose, we computed for $n = 100$ the “exact” solutions $\Phi$ by the computerized algebra system Maple with 100 decimal digits. We find that

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

(7.3)
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 [22], two more than here, to deliver $\Phi$ with about 15 correct decimal digits entrywise), fail to compute accurately $\Phi$'s tiny entries.

5.7251 $\cdot$ $10^{-30} \leq \Phi_{(i,j)} \leq 6.3012 \cdot 10^{-11}$, for $\xi = 10$.  

Large variations in magnitudes in $\Phi$'s entries for $\xi = 10$ suggest that all methods, except ADDA, may have trouble getting $\Phi$'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 $O(u)$ as they should. Since $\Phi$'s entries vary in magnitude by a factor about 500, we would expect that ERErr for all be about $O(500u) = O(10^{-12})$ 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 [12] that roughly speaking a perturbation of size $\epsilon$ to $W$ will result in an error in $\Phi$ about $O(\sqrt{\epsilon})$. On the other hand, the shifting technique built into SDAs and ADDAs and the deflating technique built into $dADDA$ and $dADDAq$ make the resulting ARE (4.17) and (6.2) sufficiently well-conditioned to be solved accurately. Guo, Iannazzo, and Meini [13]
Figure 7.2. Example 7.2. ADDA is even faster than SDAs. ADDAs, dADDAe, and dADDAq work about equally well.

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 $\Phi$ 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 example is the same as [22, Example 7.2], where ADDA delivered $\Phi$ to have almost 15 correct decimal digits entrywise in 7 iterations. The inability of the other methods to compute $\Phi$’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 $O(u)$.

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

```matlab
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, $W1_{2n} = 0$, 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 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” $\Phi$ 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 $\Phi$ 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.

Example 7.3. This is essentially the example of a positive recurrent Markov chain with nonsquare coefficients, originally from [2]. Here

$$A = 18 \cdot I_2, \quad B = 18000 \cdot I_{18} - 10^4 \cdot 1_{18,18}, \quad C = 1_{2,18}, \quad D = C^T.$$ 

It is known $\Phi = \frac{1}{18} \cdot 1_{2,18} = \Psi^T$ and $\mu = 16 > 0$. 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\(^7\), to say the least. In fact, our Maple code for ADDAs with arbitrary $\alpha$ and $\beta$ but $\eta = \beta$ gives, in exact arithmetic,

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

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. Figure 7.3 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 [22, Table 7.2].

From these examples as well as many more others, 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 $\Phi$, even when some of the entries of $\Phi$ are extremely tiny relative to others. But entrywise accuracy in computed $\Phi$ is limited to about $O(\sqrt{u})$.

\(^7\)More examples like this can be found in [21].
2. The shifting technique of Guo, Iannazzo, and Meini [13] and the deflating technique in this article can greatly improve the conditioning of an MARE in the critical case, enabling $\Phi$ to be computed much more accurately in the sense of making normalized error $NErr$ to about $O(u)$. But when $\Phi$'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 example is accidental for both ADDAs and SDAs in that $X_0 \equiv \Phi$, independent of the parameters $\alpha$ and $\beta$. In general, ADDAs is faster than SDAs as one might expect from the conclusion in [22] that ADDA is at least as good as SDA and can be faster if $A$ and $B$ are very different in magnitude.

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 [13] recognized it and proposed a shifting mechanism to still retain quadratical convergence. In this paper, we establish a general framework to deflate out 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. It is widely accepted that in computing eigen-decompositions, deflation approaches are often preferred to shifting mechanisms. But it is not clear whether this is still the case for an MARE.

The theoretical analysis given here for our deflating method is considerably more complicated than the one in [13] for the shifting method. Although the recovering formula of $\Phi$ from the one of the deflated AREs involves matrix multiplications and inversions in the general framework, in both realizations there are numerically stable and fast algorithms that cost negligibly in comparison to solving the deflated AREs.

We also propose a natural improvement to the final algorithm in [13], 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 argument in Remark 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 4.4. We conjecture that $\hat{\Psi}$ always exists, despite the inconclusive argument, but a rigorous proof eludes us.

**REFERENCES**


Received April 2012; 1st revision November 2012, final revision December 2012.

E-mail address: wgwang@ouc.edu.cn
E-mail address: weichao.wang@mavs.uta.edu
E-mail address: rcli@uta.edu