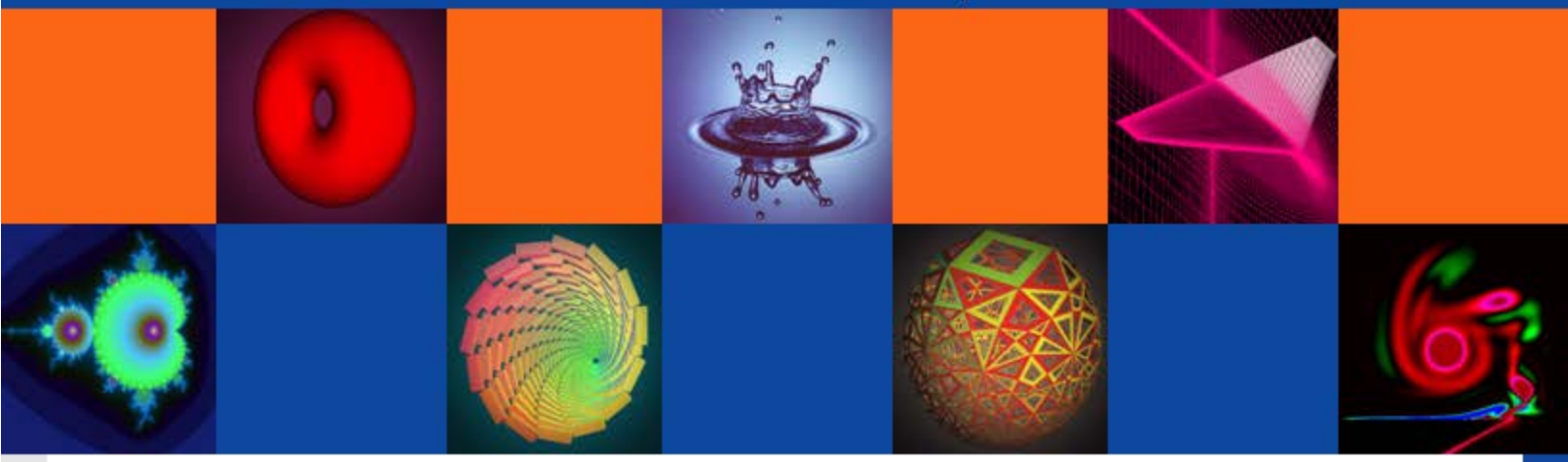
The University of Texas ARLINGTON

MATHEMATICS
Preprint Series



ADDA: Alternating-Directional Doubling Algorithm for M-Matrix Algebraic Riccati Equations

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

Technical Report 2011-04

http://www.uta.edu/math/preprint/

ADDA: Alternating-Directional Doubling Algorithm for M-Matrix Algebraic Riccati Equations

Wei-guo Wang* Wei-chao Wang† Ren-Cang Li‡

May 25, 2011
Revised September 2, 2011

Abstract

A new doubling algorithm – Alternating-Directional Doubling Algorithm (ADDA) – is developed for computing the unique minimal nonnegative solution of an M-Matrix Algebraic Riccati Equation (MARE). It is argued by both theoretical analysis and numerical experiments that ADDA is always faster than two existing doubling algorithms – SDA of Guo, Lin, and Xu (*Numer. Math.*, 103 (2006), pp. 393–412) and SDA-ss of Bini, Meini, and Poloni (*Numer. Math.*, 116 (2010), pp. 553–578) for the same purpose. Also demonstrated is that all three methods are capable of delivering minimal nonnegative solutions with entrywise relative accuracies as warranted by the defining coefficient matrices of an MARE.

2000 Mathematics Subject Classification. 15A24, 65F30, 65H10.

Key words and phrases. Matrix Riccati equation, M-Matrix, minimal nonnegative solution, doubling algorithm

1 Introduction

An *M-Matrix Algebraic Riccati Equation*[1] (MARE) is the matrix equation

$$XDX - AX - XB + C = 0, \tag{1.1}$$

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

$$W = \begin{array}{c} \\ m \\ n \end{array} \begin{array}{c} \begin{array}{cc} m & n \end{array} \\ \begin{pmatrix} B & -D \\ -C & A \end{pmatrix} \end{array}, \tag{1.2}$$

*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 author is currently a visiting scholar at Department of Mathematics, University of Texas at Arlington, Arlington, TX 76019.

†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 Grant DMS-0810506.

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

[1]Previously it was called a Nonsymmetric Algebraic Riccati Equation, a name that seems to be too broad to be descriptive. MARE was recently coined in [33] to better reflect its characteristics.

and W is a nonsingular or an irreducible singular M-matrix. This kind of Riccati equations arise in applied probability and transportation theory and have been attracting a lot of attention lately. See [18, 16, 20, 21, 22, 23, 28] and the references therein. It is shown in [16, 20] that (1.1) has a unique minimal nonnegative solution Φ, i.e.,

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

In [21], a structure-preserving doubling algorithm (SDA) was proposed and analyzed for an MARE with W being a nonsingular M-matrix by Guo, Lin, and Xu. SDA is very fast and efficient for small to medium size MAREs as it is globally and quadratically convergent. The algorithm has to select a parameter μ that is no smaller than the largest diagonal entries in both A and B. Such a choice of μ ensures the following:

1. An elegant theory of global and quadratic convergence [21] (except in the critical case for which only linear convergence is ensured [11]);

2. Computed Φ has an entrywise relative accuracy as the input data deserves, as argued recently in [33].

Consequently, SDA has since emerged as one of the most efficient algorithms.

But as we shall argue in this paper, SDA has room to improve. One situation is when A and B differs in magnitudes. But since SDA is blind to any difference between A and B, it still picks one parameter μ. Conceivably, if A and B could be treated differently with regard to their own characteristics, better algorithms would be possible. This is the motivational thought that drives our study in this paper. Specifically, we will propose a new doubling algorithm – *Alternating-Directional Doubling Algorithm* (ADDA) – that also imports the idea from the ADI (Alternating-Directional-Implicit) iteration for Sylvester equations [7, 32]. Our new doubling algorithm ADDA includes two parameters that can be tailored to reflect each individual characteristics of A and B, and consequently ADDA converges at least as fast as SDA and can be much faster when A and B have very different magnitudes.

We are not the first to notice that SDA often takes quite many iterations for an MARE with A and B having very different magnitudes. In fact, Bini, Meini, and Poloni [9] recently developed a doubling algorithm called *SDA-ss* using a shrink-and-shift approach of Ramaswami [27]. SDA-ss has shown dramatic improvements over SDA in some of the numerical tests in [9]. But it can happen that sometimes SDA-ss runs slower than SDA, although not by much. Later we will show our ADDA is always the fastest among the three methods.

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

$$\boxed{\begin{array}{l} W \text{ defined by (1.2) is a nonsingular } M\text{-matrix or an irre-} \\ \text{ducible singular } M\text{-matrix.} \end{array}} \tag{1.3}$$

The rest of this paper is organized as follows. Section 2 presents several known results about M-matrices, as well as a new result on optimizing the product of two spectral radii of the generalized Cayley transforms of two M-matrices. This new result which may be of independent interest of its own will be used to develop our optimal ADDA. Section 3 devotes to the development of ADDA whose application to M-matrix Sylvester equation leads to an improvement of the Smith method [29] in section 4. A detailed comparison on rates of convergence among

ADDA, SDA, and SDA-ss is given in section 5. Numerical results to demonstrate the efficiency of the three doubling methods are presented in section 6. Finally, we give our concluding remarks in section 7.

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 jth 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 "$\cdot^{\mathrm{T}}$" 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;
2. when $m = n$, $\mathrm{diag}(X)$ is the diagonal matrix with the same diagonal entries as X's, $\rho(X)$ is the spectral radius of X, and

$$\varrho(X) = \rho([\mathrm{diag}(X)]^{-1}[\mathrm{diag}(X) - X]).$$

Inequality $X \le Y$ means $X_{(i,j)} \le Y_{(i,j)}$ for all (i,j), and similarly for $X < Y$, $X \ge Y$, and $X > Y$.

2 Preliminary Results on M-Matrices

$A \in \mathbb{R}^{n\times n}$ is called a *Z-matrix* if it has nonpositive off-diagonal entries [8, p.284]. Any Z-matrix A can be written as $sI - N$ with $N \ge 0$, and it is called an *M-matrix* if $s \ge \rho(N)$, a *singular M-matrix* if $s = \rho(N)$, and a *nonsingular M-matrix* if $s > \rho(N)$.

In this section, we first collect a few well-known results about M-matrices in Lemmas 2.1 and 2.2 that are needed later in this paper. They can be found in, e.g., [8, 14, 31]. Then we establish a new result on optimizing the product of two spectral radii of the generalized Cayley transforms of two M-matrices.

Lemma 2.1 gives four equivalent statements about when a Z-matrix is an M-matrix.

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

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

(b) *$A^{-1} \ge 0$.*

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

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

Lemma 2.2 collects a few properties of M-matrices, important to our later analysis, where Item (d) can be found in [26].

Lemma 2.2. *Let $A, B \in \mathbb{R}^{n\times n}$, and suppose A is an M-matrix and B is a Z-matrix.*

(a) *If $B \ge A$, then B is an M-matrix. In particular, $\theta I + A$ is an M-matrix for $\theta \ge 0$ and a nonsingular M-matrix for $\theta > 0$.*

(b) *If $B \ge A$ and A is nonsingular, then B is a nonsingular M-matrix, and $A^{-1} \ge B^{-1}$.*

(c) *If A is nonsingular and irreducible, then $A^{-1} > 0$.*

(d) *The one with the smallest absolute value among all eigenvalues of A, denoted by $\lambda_{\min}(A)$, is nonnegative, and $\lambda_{\min}(A) \le \max_i A_{(i,i)}$.*

(e) *If A is a nonsingular M-matrix or an irreducible singular M-matrix, and is partitioned as*

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix},$$

where A_{11} and A_{22} are square matrices, then A_{11} and A_{22} are nonsingular M-matrices, and their Schur complements

$$A_{22} - A_{21}A_{11}^{-1}A_{12}, \quad A_{11} - A_{12}A_{22}^{-1}A_{21}$$

are nonsingular M-matrices if A is a nonsingular M-matrix or an irreducible singular M-matrices if A is an irreducible singular M-matrix.

Theorem 2.1 which may have independent interest of its own lays the foundation of our optimal ADDA in terms of its rate of convergence subject to certain nonnegativity condition. To the best of our knowledge, it is new. Define the following *generalized Cayley transformation*

$$\mathscr{C}(A;\alpha,\beta) \stackrel{\text{def}}{=} (A-\alpha I)(A+\beta I)^{-1} \tag{2.1}$$

of a square matrix A, where α, β are scalars such that $A+\beta I$ is assumed nonsingular.

Theorem 2.1. *For two M-matrices $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{m\times m}$, define*[2]

$$f(\alpha,\beta) \stackrel{\text{def}}{=} \rho(\mathscr{C}(A;\alpha,\beta)) \cdot \rho(\mathscr{C}(B;\beta,\alpha))$$

for

$$\alpha \ge \alpha_{\text{opt}} \stackrel{\text{def}}{=} \max_i A_{(i,i)}, \quad \beta \ge \beta_{\text{opt}} \stackrel{\text{def}}{=} \max_i B_{(i,i)}. \tag{2.2}$$

(a) *If both A and B are singular, then $f(\alpha,\beta) \equiv 1$;*

(b) *If one of A and B is nonsingular, then $f(\alpha,\beta)$ is strictly increasing in α and β and $f(\alpha,\beta) < 1$.*

In each case, we have

$$\min_{\alpha\ge\alpha_{\text{opt}},\beta\ge\beta_{\text{opt}}} f(\alpha,\beta) = f(\alpha_{\text{opt}},\beta_{\text{opt}}). \tag{2.3}$$

Proof. Assume for the moment that both A and B are irreducible M-matrices. Write $A = sI - N$, where $s \ge 0$ and $N \ge 0$, and N is irreducible. By the Perron-Frobenius theorem [8, p.27], there is a positive vector u such that $Nu = \rho(N)u$. It can be seen that $\lambda_{\min}(A) = s - \rho(N) \ge 0$, where $\lambda_{\min}(A)$ is as defined in Lemma 2.2(d). We have

$$-\mathscr{C}(A;\alpha,\beta)u = (\alpha I - A)(A+\beta I)^{-1}u = [\alpha - \lambda_{\min}(A)][\lambda_{\min}(A)+\beta]^{-1}u.$$

Since $-\mathscr{C}(A;\alpha,\beta)) \ge 0$ and irreducible for $\alpha > \alpha_{\text{opt}}$ and $\beta > 0$, it follows from the Perron-Frobenius theorem that

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

[2]It is possible that this function f may be undefined at $\alpha = \alpha_{\text{opt}}$ or $\beta = \beta_{\text{opt}}$. When that is the case, we take $f(\alpha,\beta)$ to be its right limits.

Similarly, we have for $\alpha > 0$ and $\beta > \beta_{\rm opt}$,

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

Finally for $\alpha > \alpha_{\rm opt}$ and $\beta > \beta_{\rm opt}$,

$$\begin{aligned} f(\alpha,\beta) &= \rho(\mathscr{C}(A;\alpha,\beta)) \cdot \rho(\mathscr{C}(B;\beta,\alpha)) \\ &= \frac{\alpha - \lambda_{\min}(A)}{\lambda_{\min}(A) + \beta} \cdot \frac{\beta - \lambda_{\min}(B)}{\lambda_{\min}(B) + \alpha} \\ &= g(\alpha)h(\beta), \end{aligned}$$

where

$$g(\alpha) = \frac{\alpha - \lambda_{\min}(A)}{\lambda_{\min}(B) + \alpha}, \quad h(\beta) = \frac{\beta - \lambda_{\min}(B)}{\lambda_{\min}(A) + \beta}.$$

Now if both A and B are singular, then $\lambda_{\min}(A) = \lambda_{\min}(B) = 0$ and thus $f(\alpha,\beta) \equiv 1$ which proves Item (a). If one of A and B is nonsingular, then $\lambda_{\min}(A) + \lambda_{\min}(B) > 0$ and thus

$$g'(\alpha) = \frac{\lambda_{\min}(A) + \lambda_{\min}(B)}{(\lambda_{\min}(B) + \alpha)^2} > 0, \quad h'(\beta) = \frac{\lambda_{\min}(A) + \lambda_{\min}(B)}{(\lambda_{\min}(A) + \beta)^2} > 0.$$

So $f(\alpha,\beta)$ is strictly increasing in α and β for $\alpha > \alpha_{\rm opt}$ and $\beta > \beta_{\rm opt}$ and

$$f(\alpha,\beta) < \lim_{\substack{\alpha\to\infty \\ \beta\to\infty}} f(\alpha,\beta) = 1.$$

This is Item (b).

Suppose now that A and B are possibly reducible. Let $\mathit{\Pi}_1 \in \mathbb{R}^{n\times n}$ and $\mathit{\Pi}_2 \in \mathbb{R}^{m\times m}$ be two permutation matrices such that

$$\mathit{\Pi}_1^{\rm T} A \mathit{\Pi}_1 = \begin{pmatrix} A_{11} & -A_{12} & \dots & -A_{1q} \\ & A_{22} & \dots & -A_{2q} \\ & & \ddots & \vdots \\ & & & A_{qq} \end{pmatrix}, \quad \mathit{\Pi}_2^{\rm T} B \mathit{\Pi}_2 = \begin{pmatrix} B_{11} & -B_{12} & \dots & -B_{1p} \\ & B_{22} & \dots & -B_{2p} \\ & & \ddots & \vdots \\ & & & B_{pp} \end{pmatrix},$$

where $A_{ij} \in \mathbb{R}^{n_i\times n_j}$, $B_{ij} \in \mathbb{R}^{m_i\times m_j}$, all A_{ii} and B_{jj} are irreducible M-matrices, and all $A_{ij} \ge 0$ and $B_{ij} \ge 0$ for $i \ne j$. It can be seen that

$$f(\alpha,\beta) = \max_{i,j} \rho(\mathscr{C}(A_{ii};\alpha,\beta)) \cdot \rho(\mathscr{C}(B_{jj};\beta,\alpha)).$$

If one of A and B is nonsingular, then one of A_{ii} and B_{jj} is nonsingular for each pair (A_{ii}, B_{jj}) and thus all $\rho(\mathscr{C}(A_{ii};\alpha,\beta)) \cdot \rho(\mathscr{C}(B_{jj};\beta,\alpha))$ are strictly increasing in α and β for $\alpha > \alpha_{\rm opt}$ and $\beta > \beta_{\rm opt}$; so is $f(\alpha,\beta)$. Now if both A and B are singular, then there is at least one pair (A_{ii}, B_{jj}) for which both A_{ii} and B_{jj} are singular and irreducible. By Item (a) we just proved for the irreducible case, for that pair $\rho(\mathscr{C}(A_{ii};\alpha,\beta)) \cdot \rho(\mathscr{C}(B_{jj};\beta,\alpha)) \equiv 1$ under (2.2). Since for all other pairs (A_{ii}, B_{jj}), $\rho(\mathscr{C}(A_{ii};\alpha,\beta)) \cdot \rho(\mathscr{C}(B_{jj};\beta,\alpha)) \le 1$ by Item (a). Thus $f(\alpha,\beta) \equiv 1$. $\square$

3 ADDA: Alternating-Directional Doubling Algorithm

The basic idea of doubling algorithms for an iterative scheme is to compute only the 2^kth approximations, instead of every approximation in the process. It traces back to 1970s (see [2] and references therein). Recent resurgence of interests in the idea has led to efficient doubling algorithms for various nonlinear matrix equations. The interested reader is referred to [11] for a more general presentation. The use of a structure-preserving doubling algorithm (SDA) to solve an MARE was first proposed and analyzed by Guo, Lin, and Xu [21]. For MARE (1.1), SDA simultaneously computes the minimal nonnegative solutions of (1.1) and its *complementary M-Matrix Algebraic Riccati Equation* (cMARE)

$$YCY - YA - BY + D = 0. \tag{3.1}$$

In what follows, we shall present our ADDA for MARE in this way: framework, analysis, and then optimal ADDA. We name it ADDA after taking into consideration that it is a doubling algorithm and relates to the Alternating-Directional-Implicit (ADI) iteration for Sylvester equations (see section 4).

3.1 Framework

The framework in this subsection actually works for any algebraic Riccati equation, provided all involved inverses exist. It is just that in general we may not have a similar convergence theory as for an MARE in the next subsection.

For any solution X of MARE (1.1) and Y of cMARE (3.1), it can be verified that

$$H\begin{pmatrix} I \\ X \end{pmatrix} = \begin{pmatrix} I \\ X \end{pmatrix} R, \quad H\begin{pmatrix} Y \\ I \end{pmatrix} = \begin{pmatrix} Y \\ I \end{pmatrix} (-S), \tag{3.2}$$

where

$$H = \begin{pmatrix} B & -D \\ C & -A \end{pmatrix}, \quad R = B - DX, \quad S = A - CY. \tag{3.3}$$

Given any scalars α and β, we have

$$\begin{aligned} (H - \beta I)\begin{pmatrix} I \\ X \end{pmatrix}(\ \ R + \alpha I) &= (H + \alpha I)\begin{pmatrix} I \\ X \end{pmatrix}(\ \ R - \beta I), \\ (H - \beta I)\begin{pmatrix} Y \\ I \end{pmatrix}(-S + \alpha I) &= (H + \alpha I)\begin{pmatrix} Y \\ I \end{pmatrix}(-S - \beta I). \end{aligned}$$

If $R + \alpha I$ and $S + \beta I$ are nonsingular, then

$$(H - \beta I)\begin{pmatrix} I \\ X \end{pmatrix} = (H + \alpha I)\begin{pmatrix} I \\ X \end{pmatrix}\mathscr{C}(R; \beta, \alpha), \tag{3.4a}$$

$$(H - \beta I)\begin{pmatrix} Y \\ I \end{pmatrix}\mathscr{C}(S; \alpha, \beta) = (H + \alpha I)\begin{pmatrix} Y \\ I \end{pmatrix}. \tag{3.4b}$$

Suppose for the moment that $A + \beta I$ and $B + \alpha I$ are nonsingular and set

$$A_\beta = A + \beta I, \qquad B_\alpha = B + \alpha I, \tag{3.5}$$

$$U_{\alpha\beta} = A_\beta - CB_\alpha^{-1}D, \quad V_{\alpha\beta} = B_\alpha - DA_\beta^{-1}C, \tag{3.6}$$

and

$$Z_1 = \begin{pmatrix} B_\alpha^{-1} & 0 \\ -CB_\alpha^{-1} & I \end{pmatrix},\ Z_2 = \begin{pmatrix} I & 0 \\ 0 & -U_{\alpha\beta}^{-1} \end{pmatrix},\ Z_3 = \begin{pmatrix} I & B_\alpha^{-1}D \\ 0 & I \end{pmatrix}.$$

It can be verified that

$$M_0 \overset{\text{def}}{=} Z_3Z_2Z_1(H - \beta I) = \begin{pmatrix} E_0 & 0 \\ -X_0 & I \end{pmatrix}, \tag{3.7a}$$

$$L_0 \overset{\text{def}}{=} Z_3Z_2Z_1(H + \alpha I) = \begin{pmatrix} I & -Y_0 \\ 0 & F_0 \end{pmatrix}, \tag{3.7b}$$

where

$$E_0 = I - (\beta+\alpha)V_{\alpha\beta}^{-1}, \quad Y_0 = (\beta+\alpha)B_\alpha^{-1}DU_{\alpha\beta}^{-1}, \tag{3.8a}$$

$$F_0 = I - (\beta+\alpha)U_{\alpha\beta}^{-1}, \quad X_0 = (\beta+\alpha)U_{\alpha\beta}^{-1}CB_\alpha^{-1}. \tag{3.8b}$$

Pre-multiply the equations in (3.4) by $Z_3Z_2Z_1$ to get

$$M_0 \begin{pmatrix} I \\ X \end{pmatrix} = L_0 \begin{pmatrix} I \\ X \end{pmatrix} \mathscr{C}(R;\beta,\alpha), \quad M_0 \begin{pmatrix} Y \\ I \end{pmatrix} \mathscr{C}(S;\alpha,\beta) = L_0 \begin{pmatrix} Y \\ I \end{pmatrix}. \tag{3.9}$$

Our development up to this point differs from SDA of [21] only in our inclusion of two parameters α and β. The significance of doing so will be demonstrated in our later comparisons on convergence rates in section 5 and numerical examples in section 6. From this point forward, ours is the same as in [21]. The idea is to construct a sequence of pairs $\{M_k, L_k\}$, $k = 0, 1, 2, \ldots$ such that

$$M_k \begin{pmatrix} I \\ X \end{pmatrix} = L_k \begin{pmatrix} I \\ X \end{pmatrix} [\mathscr{C}(R;\beta,\alpha)]^{2^k}, \quad M_k \begin{pmatrix} Y \\ I \end{pmatrix} [\mathscr{C}(S;\alpha,\beta)]^{2^k} = L_k \begin{pmatrix} Y \\ I \end{pmatrix}, \tag{3.10}$$

and at the same time M_k and L_k have the same forms as M_0 and L_0, respectively, i.e.,

$$M_k = \begin{pmatrix} E_k & 0 \\ -X_k & I \end{pmatrix}, \quad L_k = \begin{pmatrix} I & -Y_k \\ 0 & F_k \end{pmatrix}. \tag{3.11}$$

The technique for constructing $\{M_{k+1}, L_{k+1}\}$ from $\{M_k, L_k\}$ is not entirely new and can be traced back to 1980s in [10, 15, 25] and more recently in [3, 6, 30]. The idea is to seek suitable $\check{M}, \check{L} \in \mathbb{R}^{(m+n)\times(m+n)}$ such that

$$\operatorname{rank}\left((\check{M}, \check{L})\right) = m + n, \quad (\check{M}, \check{L}) \begin{pmatrix} L_k \\ -M_k \end{pmatrix} = 0 \tag{3.12}$$

and set $M_{k+1} = \check{M}M_k$ and $L_{k+1} = \check{L}L_k$. It is not hard to verify that if the equations in (3.10) hold, then they hold for k replaced by $k+1$, i.e., for the newly constructed M_{k+1} and L_{k+1}. The only problem is that not every pair $\{\check{M}, \check{L}\}$ satisfying (3.12) leads to $\{M_{k+1}, L_{k+1}\}$ having the forms of (3.11). For this, we turn to the constructions of $\{\check{M}, \check{L}\}$ in [12, 13, 21, 24]:

$$\check{M} = \begin{pmatrix} E_k(I_m - Y_kX_k)^{-1} & 0 \\ -F_k(I_n - X_kY_k)^{-1}X_k & I_m \end{pmatrix}, \quad \check{L} = \begin{pmatrix} I_n & -E_k(I_m - Y_kX_k)^{-1}Y_k \\ 0 & -F_k(I_n - X_kY_k)^{-1} \end{pmatrix},$$

with which $M_{k+1} = \check{M}M_k$ and $L_{k+1} = \check{L}L_k$ have the forms of (3.11) with

$$E_{k+1} = E_k(I_m - Y_kX_k)^{-1}E_k, \tag{3.13a}$$
$$F_{k+1} = F_k(I_n - X_kY_k)^{-1}F_k, \tag{3.13b}$$
$$X_{k+1} = X_k + F_k(I_n - X_kY_k)^{-1}X_kE_k, \tag{3.13c}$$
$$Y_{k+1} = Y_k + E_k(I_m - Y_kX_k)^{-1}Y_kF_k. \tag{3.13d}$$

By now we have presented the framework of ADDA:

1. Pick suitable α and β for (best) convergence rate;
2. Compute M_0 and L_0 of (3.7) by (3.5), (3.6), and (3.8);
3. Iteratively compute M_k and L_k by (3.13) until convergence.

Associated with this general framework arise a few questions:

1. Are the iterative formulas in (3.13) well-defined, i.e., do all the inverses exist?
2. How do we choose best parameters α and β for fast convergence?
3. What do X_k and Y_k converge to if they are convergent?
4. How much better is ADDA than the doubling algorithms: SDA of Guo, Lin, and Xu [21] and SDA-ss of Bini, Meini, and Poloni [9]?

The first three questions will be addressed in the next subsection while the last question will be answered in section 5.

3.2 Analysis

Recall that W defined by (1.2) is a nonsingular or an irreducible singular M-matrix. MARE (1.1) has a unique minimal nonnegative solution Φ [18] and cMARE (3.1) has a unique minimal nonnegative solution Ψ. Some properties of Φ and Ψ are summarized in Theorem 3.1 below. They are needed in order to answer the questions we posed at the end of the previous subsection.

Theorem 3.1 ([16, 17, 18])**.** *Assume* (1.3).

(a) *MARE* (1.1) *has a unique minimal nonnegative solution* Φ*, and and its cMARE* (3.1) *has a unique minimal nonnegative solution* Ψ*;*

(b) *If* W *is* irreducible*, then* $\Phi > 0$ *and* $A - \Phi D$ *and* $B - D\Phi$ *are* irreducible M*-matrices;*

(c) *If* W *is* nonsingular*, then* $A - \Phi D$ *and* $B - D\Phi$ *are* nonsingular M*-matrices;*

(d) *Suppose* W *is* irreducible *and* singular*. Let* $u_1, v_1 \in \mathbb{R}^m$ *and* $u_2, v_2 \in \mathbb{R}^n$ *be positive vectors such that*

$$W\begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = 0, \quad \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}^{\mathrm{T}} W = 0. \tag{3.14}$$

1. *If* $u_1^{\mathrm{T}}v_1 > u_2^{\mathrm{T}}v_2$*, then* $B - D\Phi$ *is a singular* M*-matrix with*[3] $(B - D\Phi)v_1 = 0$ *and* $A - \Phi D$ *is a nonsingular* M*-matrix, and* $\Phi v_1 = v_2$ *and* $\Psi v_2 < v_1$*;*
2. *If* $u_1^{\mathrm{T}}v_1 = u_2^{\mathrm{T}}v_2$ *(the so-called* critical case*), then both* $B - D\Phi$ *and* $A - \Phi D$ *are singular* M*-matrices, and* $\Phi v_1 = v_2$ *and* $\Psi v_2 = v_1$*;*
3. *If* $u_1^{\mathrm{T}}v_1 < u_2^{\mathrm{T}}v_2$*, then* $B - D\Phi$ *is a nonsingular* M*-matrix and* $A - \Phi D$ *is a singular* M*-matrix, and* $\Phi v_1 < v_2$ *and* $\Psi v_2 = v_1$*.*

(e) *$I - \Phi\Psi$ and $I - \Psi\Phi$ are M-matrices and they are nonsingular, except for the critical case in which both are singular.*

Recall our goal is to compute Φ as efficiently and accurately as possible and, as a by-product, Ψ, too. In view of this goal, we identify $X = \Phi$ and $Y = \Psi$ in all appearances of X and Y in subsection 3.1. In particular

$$S = A - C\Psi, \quad R = B - D\Phi, \tag{3.3$'$}$$

and (3.10) and (3.11) yield immediately

$$E_k = (I - Y_k\Phi)\ [\mathscr{C}(R;\beta,\alpha)]^{2^k}, \tag{3.15a}$$

$$\Phi - X_k = \quad F_k\Phi \quad [\mathscr{C}(R;\beta,\alpha)]^{2^k}, \tag{3.15b}$$

$$\Psi - Y_k = \quad E_k\Psi \quad [\mathscr{C}(S;\alpha,\beta)]^{2^k}, \tag{3.15c}$$

$$F_k = (I - X_k\Psi)\,[\mathscr{C}(S;\alpha,\beta)]^{2^k}. \tag{3.15d}$$

Examining (3.15), we see that ADDA will converge if

$$\rho(\mathscr{C}(R;\beta,\alpha)) < 1, \quad \rho(\mathscr{C}(S;\alpha,\beta)) < 1, \tag{3.16a}$$

because then

$$[\mathscr{C}(R;\beta,\alpha)]^{2^k} \to 0, \quad [\mathscr{C}(S;\alpha,\beta)]^{2^k} \to 0 \tag{3.16b}$$

as $k \to \infty$. This is one of the guiding principles in [21] which enforces

$$\alpha = \beta \geq \max_{i,j}\{A_{(i,i)}, B_{(j,j)}\} \tag{3.17}$$

which in turn ensures (3.16a) and thus (3.16b) because, by Theorem 3.1(c), both[4] S and R are nonsingular M-matrices if[5] W is a nonsingular M-matrix. Later Guo, Iannazzo, and Meini [19] proved SDA of [21] still converges even if W is a singular irreducible M-matrix. They also proved that taking

$$\alpha = \beta = \max_{i,j}\{A_{(i,i)}, B_{(j,j)}\} \tag{3.18}$$

makes the resulting SDA converge the fastest subject to (3.17) [19, Theorem 4.4]. Another critical implication of (3.17) is that it makes $-E_0$ and $-F_0$, E_k and F_k for $k \geq 1$, and X_k and Y_k for

[3] [16, Theorem 4.8] says in this case $D\Phi v_1 = Dv_2$ which leads to $(B - D\Phi)v_1 = Bv_1 - Dv_2 = 0$.

[4] That R is a nonsingular M-matrix is stated explicitly in Theorem 3.1(c). For S, we apply Theorem 3.1(c) to cMARE (3.1) identified as an MARE in the form of (1.1) with its coefficient W-matrix as $\begin{pmatrix} A & -C \\ -D & B \end{pmatrix}$.

[5] This is the case studied in [21].

$k \geq 0$ all nonnegative [21], a property that enables SDA of [21] (with some minor but crucial implementation changes [33]) to compute Φ with deserved entrywise relative accuracy as argued in [33].

We would like our ADDA to have such a capability as well, i.e., computing Φ with deserved entrywise relative accuracy. To this end, we require

$$\alpha \geq \alpha_{\rm opt} \stackrel{\rm def}{=} \max_i A_{(i,i)}, \quad \beta \geq \beta_{\rm opt} \stackrel{\rm def}{=} \max_j B_{(j,j)}, \tag{3.19}$$

allow α and β to be different, and seek to minimize the product of the spectral radii

$$\rho(\mathscr{C}(R;\beta,\alpha)) \cdot \rho(\mathscr{C}(S;\alpha,\beta)),$$

rather than each individual spectral radius. Later we will see that it is this product, not each individual spectral radius, that ultimately reflects the true rate of convergence. In particular, convergence is guaranteed if the product is less than 1, even if one of the spectral radii is bigger than 1. Moreover, the smaller the product, the faster the convergence.

The assumption (1.3) implies that A and B are nonsingular M-matrices by Lemma 2.2(e). Therefore both $\alpha_{\rm opt} > 0$ and $\beta_{\rm opt} > 0$.

Lemma 3.1. *Assume* (1.3). *If $\alpha > 0$ and $\beta > 0$, then A_β, B_α, $U_{\alpha\beta}$, and $V_{\alpha\beta}$ defined in* (3.5) *and* (3.6) *are nonsingular M-matrices. Furthermore, both $U_{\alpha\beta}$ and $V_{\alpha\beta}$ are irreducible if W is irreducible.*

Proof. If $\alpha > 0$ and $\beta > 0$,

$$\widehat{W} = W + \begin{pmatrix} \alpha I & 0 \\ 0 & \beta I \end{pmatrix} = \begin{pmatrix} B + \alpha I & -D \\ -C & A + \beta I \end{pmatrix} \geq \min\{\alpha, \beta\} \cdot I + W$$

is a nonsingular M-matrix. As the diagonal blocks of $\widehat{W}$, A_β and B_α are nonsingular M-matrices; so are their corresponding Schur complements $V_{\alpha\beta}$ and $U_{\alpha\beta}$ in $\widehat{W}$ by Lemma 2.2(e). If also W is irreducible, then $\widehat{W}$ is a nonsingular irreducible M-matrix, and thus both $U_{\alpha\beta}$ and $V_{\alpha\beta}$ are nonsingular irreducible M-matrices again by Lemma 2.2(e). □

Theorem 3.2. *Assume* (1.3) *and* (3.19).

(a) *We have*

$$E_0 \leq 0,\ F_0 \leq 0,\ \mathscr{C}(R;\beta,\alpha) \leq 0,\ \mathscr{C}(S;\alpha,\beta) \leq 0, \tag{3.20}$$
$$0 \leq X_0 \leq \Phi,\ 0 \leq Y_0 \leq \Psi. \tag{3.21}$$

If W is also irreducible, then

$$E_0 < 0,\ F_0 < 0,\ \mathscr{C}(R;\beta,\alpha) < 0,\ \mathscr{C}(S;\alpha,\beta) < 0, \tag{3.20$'$}$$
$$0 \leq X_0 < \Phi,\ 0 \leq Y_0 < \Psi. \tag{3.21$'$}$$

(b) *Both $I - Y_k X_k$ and $I - X_k Y_k$ are nonsingular M-matrices for all $k \geq 0$.*

(c) *We have*

$$E_k \geq 0,\ F_k \geq 0,\ 0 \leq X_{k-1} \leq X_k \leq \Phi,\ 0 \leq Y_{k-1} \leq Y_k \leq \Psi \quad \textit{for } k \geq 1. \tag{3.22}$$

If W is also irreducible, then

$$E_k > 0,\ F_k > 0,\ 0 \leq X_{k-1} < X_k < \Phi,\ 0 \leq Y_{k-1} < Y_k < \Psi \quad \textit{for } k \geq 1. \tag{3.22$'$}$$

Proof. Our proof is largely the same as in [19].

(a) That $\mathscr{C}(R;\beta,\alpha) \leq 0$ and $\mathscr{C}(S;\alpha,\beta) \leq 0$ is fairly straightforward because R and S are M-matrices and α and β are restricted by (3.19). For E_0 and F_0, we note

$$E_0 = V_{\alpha\beta}^{-1}[V_{\alpha\beta} - (\beta+\alpha)I] \tag{3.23a}$$
$$= V_{\alpha\beta}^{-1}(B - \beta I - DA_\beta^{-1}C), \tag{3.23b}$$
$$F_0 = U_{\alpha\beta}^{-1}[U_{\alpha\beta} - (\beta+\alpha)I] \tag{3.23c}$$
$$= U_{\alpha\beta}^{-1}(A - \alpha I - CB_\alpha^{-1}D). \tag{3.23d}$$

Since A_β, B_α, $V_{\alpha\beta}$, and $U_{\alpha\beta}$ are nonsingular M-matrices by Lemma 3.1, we have

$$A_\beta^{-1} \geq 0, \quad B_\alpha^{-1} \geq 0, \quad V_{\alpha\beta}^{-1} \geq 0, \quad U_{\alpha\beta}^{-1} \geq 0.$$

Therefore $E_0 \leq 0$, $F_0 \leq 0$, $X_0 \geq 0$, and $Y_0 \geq 0$. Equations (3.15b) and (3.15c) for $k = 0$ yields $\Phi - X_0 \geq 0$ and $\Psi - Y_0 \geq 0$, respectively.

Now suppose W is irreducible. By Lemma 3.1, both $U_{\alpha\beta}$ and $V_{\alpha\beta}$ are irreducible. So $U_{\alpha\beta}^{-1} > 0$, $V_{\alpha\beta}^{-1} > 0$, and no columns of $V_{\alpha\beta} - (\beta+\alpha)I \leq 0$ and $U_{\alpha\beta} - (\beta+\alpha)I \leq 0$ are zeros. There $E_0 < 0$ and $F_0 < 0$ by (3.23a) and (3.23c). Theorem 3.1(b) implies that $(S+\beta I)^{-1} > 0$, $(R+\alpha I)^{-1} > 0$, and no columns of $S - \alpha I \leq 0$ and $R - \beta I \leq 0$ are zeros, and thus

$$\mathscr{C}(S;\alpha,\beta) = (S+\beta I)^{-1}(S-\alpha I) < 0,\ \mathscr{C}(R;\beta,\alpha) = (R+\alpha I)^{-1}(R-\beta I) < 0.$$

Finally

$$\Phi - X_0 = F_0\Phi\mathscr{C}(R;\beta,\alpha) > 0, \quad \Psi - Y_0 = E_0\Psi\mathscr{C}(S;\alpha,\beta) > 0$$

because $\Phi > 0$ and $\Psi > 0$ by Theorem 3.1(b) and because (3.20$'$).

(b) and **(c)** We have $I - X_0Y_0 \geq I - \Phi\Psi$ and $I - Y_0X_0 \geq I - \Psi\Phi$. Suppose for the moment that W is nonsingular. Then both $I-\Phi\Psi$ and $I-\Psi\Phi$ are nonsingular M-matrices by Theorem 3.1(e), and thus $I - X_0Y_0$ and $I - Y_0X_0$ are nonsingular M-matrices, too, by Lemma 2.2(b).

Now suppose W is an irreducible singular matrix. By Theorem 3.1(d), we have $\Psi\Phi v_1 \leq v_1$, where $v_1 > 0$ is defined in Theorem 3.1(d). So $\rho(\Psi\Phi) \leq 1$ by [8, Theorem 1.11, p.28]. By part (a) of this theorem, $0 \leq X_0 < \Phi$ and $0 \leq Y_0 < \Psi$. Therefore $0 \leq X_0Y_0 < \Phi\Psi$. Since both X_0Y_0 and $\Phi\Psi$ are irreducible, we conclude by [8, Corollary 1.5, p.27]

$$\rho(Y_0X_0) = \rho(X_0Y_0) < \rho(\Psi\Phi) = \rho(\Phi\Psi) \leq 1,$$

and thus $I - Y_0X_0$ and $I - X_0Y_0$ are nonsingular M-matrices. This proves part (b) for $k = 0$.

Since $E_0 \leq 0$ and $F_0 \leq 0$, and $I-Y_0X_0$ and $I-X_0Y_0$ are nonsingular M-matrices, we deduce from (3.13)

$$E_1 \geq 0, \quad F_1 \geq 0, \quad X_1 \geq X_0, \quad Y_1 \geq Y_0.$$

By (3.15b) and (3.15c),

$$\Phi - X_1 = F_1\Phi\left[\mathscr{C}(R;\beta,\alpha)\right]^2, \quad \Psi - Y_1 = E_1\Psi\left[\mathscr{C}(S;\alpha,\beta)\right]^2, \tag{3.24}$$

yielding $\Phi - X_1 \geq 0$ and $\Psi - Y_1 \geq 0$, respectively. Consider now W is also irreducible. We have, by (3.20′) and (3.21′) and (3.13),

$$E_1 > 0, \quad F_1 > 0, \quad X_1 > X_0 \geq 0, \quad Y_1 > Y_0 \geq 0,$$

and then $X_1 < \Phi$ and $Y_1 < \Psi$ by (3.24). This proves part (c) for $k = 1$.

Part (b) for $k \geq 1$ and part (c) for $k \geq 2$ can be proved together through the induction argument. Detail is omitted. □

One important implication of Theorem 3.2 is that all formulas in subsection 3.1 for ADDA is well-defined under the assumptions (1.3) and (3.19).

Next we look into choosing α and β subject to (3.19) to optimize the convergence speed. We have (3.15) which yields

$$0 \leq \Phi - X_k = (I - X_k\Psi)\left[\mathscr{C}(S;\alpha,\beta)\right]^{2^k}\Phi\left[\mathscr{C}(R;\beta,\alpha)\right]^{2^k} \tag{3.25a}$$

$$\leq \left[\mathscr{C}(S;\alpha,\beta)\right]^{2^k}\Phi\left[\mathscr{C}(R;\beta,\alpha)\right]^{2^k}, \tag{3.25b}$$

$$0 \leq \Psi - Y_k = (I - Y_k\Phi)\left[\mathscr{C}(R;\beta,\alpha)\right]^{2^k}\Psi\left[\mathscr{C}(S;\alpha,\beta)\right]^{2^k} \tag{3.25c}$$

$$\leq \left[\mathscr{C}(R;\beta,\alpha)\right]^{2^k}\Psi\left[\mathscr{C}(S;\alpha,\beta)\right]^{2^k}. \tag{3.25d}$$

Now if W is a nonsingular M-matrix, then both R and S are nonsingular M-matrices, too, by Theorem 3.1(c). Therefore

$$\rho(\mathscr{C}(R;\beta,\alpha)) < 1, \ \rho(\mathscr{C}(S;\alpha,\beta)) < 1 \text{ under (3.17)}, \tag{3.26}$$

implying $X_k \to \Phi$ and $Y_k \to \Psi$ as $k \to \infty$. This is what was proved in [21]. But for irreducible singular M-matrix W with $u_1^{\mathrm{T}}v_1 \neq u_2^{\mathrm{T}}v_2$, it is proved in [19] that one of the spectral radii in (3.26) is less than 1 while the other one is equal to 1, still implying $X_k \to \Phi$ and $Y_k \to \Psi$ as $k \to \infty$. Furthermore, [19, Theorem 4.4] implies that the best choice is given by (3.18) in the sense that both spectral radii in $\rho(\mathscr{C}(R;\alpha,\alpha))$ and $\rho(\mathscr{C}(S;\alpha,\alpha))$ are minimized.

We can do better by allowing α and β to be different, with the help of Theorem 2.1. The main result is summarized in the following theorem.

Theorem 3.3. *Assume* (1.3) *and* (3.19). *We have*

$$\limsup_{k\to\infty} \|\Phi - X_k\|^{1/2^k} \leq \rho(\mathscr{C}(S;\alpha,\beta)) \cdot \rho(\mathscr{C}(R;\beta,\alpha)), \tag{3.27a}$$

$$\limsup_{k\to\infty} \|\Psi - Y_k\|^{1/2^k} \leq \rho(\mathscr{C}(R;\beta,\alpha)) \cdot \rho(\mathscr{C}(S;\alpha,\beta)), \tag{3.27b}$$

where $\|\cdot\|$ *is any matrix norm. The optimal* α *and* β *that minimize the right-hand sides of* (3.27) *are* $\alpha = \alpha_{\mathrm{opt}}$ *and* $\beta = \beta_{\mathrm{opt}}$.

Proof. By (3.25b), we have

$$\|\Phi - X_k\|^{1/2^k} \le \left\| [\mathscr{C}(S;\alpha,\beta)]^{2^k} \right\|^{1/2^k} \cdot \|\Phi\|^{1/2^k} \cdot \left\| [\mathscr{C}(R;\beta,\alpha)]^{2^k} \right\|^{1/2^k}.$$

which goes to $\rho(\mathscr{C}(S;\alpha,\beta)) \cdot \rho(\mathscr{C}(R;\beta,\alpha))$ as $k \to \infty$. Thus (3.27a) holds. Similarly we have (3.27b). Since $R = B - D\Phi$ and $S = A - C\Psi$ are M-matrices and $D\Phi \ge 0$ and $C\Psi \ge 0$,

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

By Theorem 2.1, $\rho(\mathscr{C}(R;\beta,\alpha)) \cdot \rho(\mathscr{C}(S;\alpha,\beta))$ is either strictly increasing if at least one of R and S is nonsingular or identically 1, subject to (3.19). So in any case, $\alpha = \alpha_{\text{opt}}$ and $\beta = \beta_{\text{opt}}$ minimize the product $\rho(\mathscr{C}(S;\alpha,\beta)) \cdot \rho(\mathscr{C}(R;\beta,\alpha))$. □

3.3 Optimal ADDA

We are now ready to present our ADDA, basing on the framework in subsection 3.1 and analysis in subsection 3.2.

Algorithm 3.1.
ADDA for MARE $XDX - AX - XB + C = 0$ **and,**
as a by-product, for cMARE $YCY - YA - BY + D = 0$.

1 Pick $\alpha \ge \alpha_{\text{opt}}$ and $\beta \ge \beta_{\text{opt}}$;
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 (3.6) and then their inverses;
5 Compute E_0 by (3.23b), F_0 by (3.23d), X_0 and Y_0 by (3.8);
6 Compute $(I - X_0Y_0)^{-1}$;
7 Compute X_1 and Y_1 by (3.13c) and (3.13d);
8 For $k = 1, 2, \ldots$, until convergence
9 Compute E_k and F_k by (3.13a) and (3.13b) (after substituting $k+1$ for k);
10 Compute $(I - X_kY_k)^{-1}$ and $(I - Y_kX_k)^{-1}$;
11 Compute X_{k+1} and Y_{k+1} by (3.13c) and (3.13d);
12 Enddo

REMARK **3.1.** ADDA differs from SDA of [21] only in its initial setup – Lines 1 – 5 that build two parameters α and β into the algorithm. In [33], we explained in detail how to make critical implementation changes to ensure computed Φ and Ψ by SDA to have entrywise relative accuracy as much as the input data deserves. The key is to use the GTH-like algorithm [1, 34] to invert all nonsingular M-matrices. Every comment in [33, Remark 4.1], except the selection of its sole parameter for SDA applies here. We shall not repeat most of those comments to save space.

About selecting the parameters α and β, Theorem 3.3 suggests $\alpha = \alpha_{\text{opt}}$ and $\beta = \beta_{\text{opt}}$ for the best convergence rate. But when the diagonal entries of A and B are not known exactly or not exactly floating point numbers, the diagonal entries of $A - \alpha I$ and $B - \beta I$ needed for computing E_0 by (3.23b) and F_0 by (3.23d) may suffer catastrophic cancelations. One remedy to avoid such possible catastrophic cancelations is to take $\alpha = \eta \cdot \alpha_{\text{opt}}$ and $\beta = \eta \cdot \beta_{\text{opt}}$ for some $\eta > 1$ but not too close to 1. This will slow down the convergence, but the gain is to ensure computed Φ and Ψ by ADDA have deserved entrywise relative accuracy. Usually ADDA converges so fast,

such a little degradation in the optimality of α and β does not increase the number of iteration steps needed for convergence.

Recall the convergence of ADDA does not depend on both spectral radii $\rho(\mathscr{C}(S;\alpha,\beta))$ and $\rho(\mathscr{C}(R;\beta,\alpha))$ being less than 1. In fact, often the larger one is bigger than 1 while the smaller one is less than 1 but the product is less than 1. It can happen that the larger one is so big that implemented as exactly given in Algorithm 3.1 ADDA can encounter overflown in E_k or F_k before X_k and Y_k converge to a desired accuracy. This happened in one of our test runs. To cure this, we notice that the scaling of E_k and F_k to ηE_k and $\eta^{-1}F_k$ for some $\eta > 0$ has no effect on X_{k+1} and Y_{k+1} and thereafter. In view of this, we devise the following strategy: at every iteration step after E_k and F_k are computed, we pick η such that $\|\eta E_k\| = \|\eta^{-1}F_k\|$, i.e., $\eta = \sqrt{\|F_k\|/\|E_k\|}$, to scale E_k and F_k to ηE_k and $\eta^{-1}F_k$, where $\|\cdot\|$ is some matrix norm. $\diamond$

The *optimal ADDA* is the one with $\alpha = \alpha_{\rm opt}$ and $\beta = \beta_{\rm opt}$. Since there is little reason not to use the optimal ADDA, except for the situation we mentioned in Remark 3.1 above, for the ease of presentation in what follows we always mean the optimal ADDA whenever we refer to an ADDA, unless explicitly stated differently.

4 Application to M-Matrix Sylvester Equation

When $D = 0$, MARE (1.1) degenerates to a Sylvester equation:

$$AX + XB = C. \tag{4.1}$$

The assumption (1.3) on its associated $\begin{pmatrix} B & 0 \\ -C & A \end{pmatrix}$ implies that A and B are nonsingular M-matrices and $C \geq 0$. Thus (4.1) is an *M-Matrix Sylvester Equation* (MSE) as defined in [34]: both A and B have positive diagonal entries and nonpositive off-diagonal entries and $P = I_m \otimes A + B^{\rm T} \otimes I_n$ is a nonsingular M-matrix, and $C \geq 0$.

MSE (4.1) has the unique solution $\Phi \geq 0$ and its cMARE has the solution $\Psi = 0$. Apply ADDA to (4.1) to get

$$E_0 = \mathscr{C}(B;\beta,\alpha) \equiv (B+\alpha I)^{-1}(B-\beta I), \tag{4.2a}$$
$$F_0 = \mathscr{C}(A;\alpha,\beta) \equiv (A+\beta I)^{-1}(A-\alpha I), \tag{4.2b}$$
$$X_0 = (\beta+\alpha)(A+\beta I)^{-1}C(B+\alpha I)^{-1}, \tag{4.2c}$$

and for $k \geq 0$

$$E_{k+1} = E_k^2, \qquad F_{k+1} = F_k^2, \tag{4.2d}$$
$$X_{k+1} = X_k + F_k X_k E_k. \tag{4.2e}$$

The associated error equation is

$$0 \leq \Phi - X_k = [\mathscr{C}(A;\alpha,\beta)]^{2^k}\, \Phi\, [\mathscr{C}(B;\beta,\alpha)]^{2^k}. \tag{4.3}$$

Smith's method [29, 34] is obtained after setting $\alpha = \beta$ in (4.2) always.

Alternatively, we can derive (4.2) through a combination of an *Alternating-Directional-Implicit* (ADI) iteration and Smith's idea in [29]. Given an approximation $\boldsymbol{X} \approx \Phi$, we compute next approximation $\boldsymbol{Z}$ by one step of ADI:

1. Solve $(A+\beta I)\boldsymbol{Y} = C - \boldsymbol{X}(B-\beta I)$ for $\boldsymbol{Y}$;

2. Solve $\boldsymbol{Z}(B+\alpha I) = C - (A-\alpha I)\boldsymbol{Y}$ for $\boldsymbol{Z}$.

Eliminate $\boldsymbol{Y}$ to get

$$\boldsymbol{Z} = X_0 + F_0 \boldsymbol{X} E_0, \tag{4.4}$$

where E_0, F_0, and X_0 are the same as in (4.2a) – (4.2c). With $\boldsymbol{Z}_0 = \boldsymbol{X} = 0$, keep iterate (4.4) to get

$$\boldsymbol{Z}_k = \sum_{i=0}^{k} F_0^i X_0 E_0^i.$$

If it converges, it converges to the solution of (4.1) $\Phi = \boldsymbol{Z}_\infty = \sum_{i=0}^{\infty} F_0^i X_0 E_0^i$. It can be verified that $\{\boldsymbol{Z}_i\}$ relates to $\{X_i\}$ by $X_k = \boldsymbol{Z}_{2^k}$. Namely, instead of computing every member in the sequence $\{\boldsymbol{Z}_i\}$, (4.2) computes only the 2^kth members. In view of its connection to ADI and Smith's method [29], we call (4.2) an *Alternating-Directional Smith Method* (ADSM) for MSE (4.1). This connection to ADI is also the reason for us to name our Algorithm 3.1 an *Alternating-Directional Doubling Algorithm* (ADDA).

Equation (4.3) gives

$$\limsup_{k\to\infty} \|\Phi - X_k\|^{1/2^k} \le \rho(\mathscr{C}(A;\alpha,\beta)) \cdot \rho(\mathscr{C}(B;\beta,\alpha)), \tag{4.5}$$

suggesting us to pick α and β to minimize the right-hand side of (4.5) for fastest convergence. Subject to again

$$\alpha \ge \alpha_{\rm opt} \stackrel{\rm def}{=} \max_i A_{(i,i)}, \quad \beta \ge \beta_{\rm opt} \stackrel{\rm def}{=} \max_j B_{(j,j)} \tag{3.19}$$

in order to ensure $F_0 \le 0$, $E_0 \le 0$ and all $F_k \ge 0$ and $E_k \ge 0$ for $k \ge 1$, we conclude by Theorem 2.1 that $\alpha = \alpha_{\rm opt}$ and $\beta = \beta_{\rm opt}$ minimize the right-hand side of (4.5).

5 Comparisons with Existing Doubling Algorithms

In this section, we will compare the rates of convergence among our ADDA, the structure-preserving doubling algorithm (SDA) of [21], and SDA combined with the shrink-and-shift technique (SDA-ss) of [9].

The right-hand sides in (3.27) provide an upper bound on convergence rate of ADDA. It is possible that the bound may overestimate the rate, but we expect in general it is tight. To facilitate our comparisons in what follows, we shall simply regard the upper bound as the *true* rate, and without loss of generality, assume

$$\alpha_{\rm opt} \stackrel{\rm def}{=} \max_i A_{(i,i)} \ge \beta_{\rm opt} \stackrel{\rm def}{=} \max_i B_{(i,i)}. \tag{5.1}$$

Let $\lambda_{\min}(S)$ be the eigenvalue of S in (3.3′) with the smallest real part among all its eigenvalues. We know $\lambda_{\min}(S) \ge 0$, and let $\lambda_{\min}(R)$ be the same for R also in (3.3′).

We have the convergence rate for the optimal ADDA

$$r_{\rm adda} = \frac{\alpha_{\rm opt} - \lambda_{\min}(S)}{\beta_{\rm opt} + \lambda_{\min}(S)} \cdot \frac{\beta_{\rm opt} - \lambda_{\min}(R)}{\alpha_{\rm opt} + \lambda_{\min}(R)}. \tag{5.2}$$

Estimates in (3.27) with $\alpha = \beta$ hold for SDA. Apply [19, Theorem 4.4] to conclude that the convergence rate for the optimal SDA is

$$r_{\text{sda}} = \frac{\alpha_{\text{opt}} - \lambda_{\min}(S)}{\alpha_{\text{opt}} + \lambda_{\min}(S)} \cdot \frac{\alpha_{\text{opt}} - \lambda_{\min}(R)}{\alpha_{\text{opt}} + \lambda_{\min}(R)} \tag{5.3}$$

upon noticing (5.1).

In order to see the convergence rate of the optimal SDA-ss, we outline the algorithm below. For

$$\beta \geq \beta_{\text{opt}} \stackrel{\text{def}}{=} \max_j B_{(j,j)}, \tag{5.4}$$

set

$$\widehat{H} = I - \beta^{-1}H, \quad \widehat{A} = I + \beta^{-1}A, \quad \widehat{B} = I - \beta^{-1}B, \tag{5.5}$$

where H is defined as in (3.3). With S and R given by (3.3′), we have

$$\widehat{H}\begin{pmatrix} I \\ \Phi \end{pmatrix} = \begin{pmatrix} I \\ \Phi \end{pmatrix}\widehat{R}, \qquad \widehat{H}\begin{pmatrix} \Psi \\ I \end{pmatrix}\widehat{S} = \begin{pmatrix} \Psi \\ I \end{pmatrix}, \tag{5.6a}$$

$$\widehat{R} = I - \beta^{-1}R, \qquad \widehat{S} = (I + \beta^{-1}S)^{-1}. \tag{5.6b}$$

Note that $\widehat{A}$ is a nonsingular M-matrix, and let

$$\widehat{M}_0 = \begin{pmatrix} \widehat{E}_0 & 0 \\ -\widehat{X}_0 & I \end{pmatrix}, \quad \widehat{L}_0 = \begin{pmatrix} I & -\widehat{Y}_0 \\ 0 & \widehat{F}_0 \end{pmatrix}, \tag{5.7}$$

where

$$\widehat{E}_0 = \widehat{B} + \beta^{-2}D\widehat{A}^{-1}C, \quad \widehat{Y}_0 = \beta^{-1}D\widehat{A}^{-1}, \tag{5.8a}$$

$$\widehat{F}_0 = \widehat{A}^{-1}, \qquad \widehat{X}_0 = \beta^{-1}\widehat{A}^{-1}C. \tag{5.8b}$$

It can be verified that $\widehat{H} = \widehat{L}_0^{-1}\widehat{M}_0$, substituting which into the equations in (5.6) to get

$$\widehat{M}_0\begin{pmatrix} I \\ \Phi \end{pmatrix} = \widehat{L}_0\begin{pmatrix} I \\ \Phi \end{pmatrix}\widehat{R}, \quad \widehat{M}_0\begin{pmatrix} \Psi \\ I \end{pmatrix}\widehat{S} = \widehat{L}_0\begin{pmatrix} \Psi \\ I \end{pmatrix}.$$

The rest follows the same idea in [21] (and also in section 3). SDA-ss seeks to construct a sequence of pairs $\{\widehat{M}_k, \widehat{L}_k\}$, $k = 0, 1, 2, \ldots$ such that

$$\widehat{M}_k\begin{pmatrix} I \\ \Phi \end{pmatrix} = \widehat{L}_k\begin{pmatrix} I \\ \Phi \end{pmatrix}\widehat{R}^{\,2^k}, \quad \widehat{M}_k\begin{pmatrix} \Psi \\ I \end{pmatrix}\widehat{S}^{\,2^k} = \widehat{L}_k\begin{pmatrix} \Psi \\ I \end{pmatrix}, \tag{5.9}$$

and at the same time $\widehat{M}_k$ and $\widehat{L}_k$ have the same forms as $\widehat{M}_0$ and $\widehat{L}_0$, respectively, i.e.,

$$\widehat{M}_k = \begin{pmatrix} \widehat{E}_k & 0 \\ -\widehat{X}_k & I \end{pmatrix}, \quad \widehat{L}_k = \begin{pmatrix} I & -\widehat{Y}_k \\ 0 & \widehat{F}_k \end{pmatrix}. \tag{5.10}$$

The formulas (3.13) for advancing from the kth approximations to the $(k+1)$st ones remain valid here after placing a "*hat*" over every occurrence of E, F, X, and Y there. At the end, we will have the following equations for errors in the approximations $\widehat{X}_k$ and $\widehat{Y}_k$:

$$\Phi - \widehat{X}_k = (I - \widehat{X}_k\Psi)\widehat{S}^{2^k}\Phi\widehat{R}^{2^k} \le \widehat{S}^{2^k}\Phi\widehat{R}^{2^k}, \tag{5.11}$$

$$\Psi - \widehat{Y}_k = (I - \widehat{Y}_k\Phi)\widehat{R}^{2^k}\Psi\widehat{S}^{2^k} \le \widehat{R}^{2^k}\Psi\widehat{S}^{2^k}. \tag{5.12}$$

Consequently

$$\limsup_{k\to\infty} \|\Phi - \widehat{X}_k\|^{1/2^k},\ \limsup_{k\to\infty} \|\Psi - \widehat{Y}_k\|^{1/2^k} \le \rho(\widehat{R})\cdot\rho(\widehat{S}). \tag{5.13}$$

In view of this inequality and (5.4), we conclude that the convergence rate of the optimal SDA-ss is

$$r_{\text{sda-ss}} = \frac{1-\beta_{\text{opt}}^{-1}\lambda_{\min}(R)}{1+\beta_{\text{opt}}^{-1}\lambda_{\min}(S)} = \frac{\beta_{\text{opt}}-\lambda_{\min}(R)}{\beta_{\text{opt}}+\lambda_{\min}(S)}. \tag{5.14}$$

Now we are ready to compare all three rates of convergence. To simplify notations, let us drop the subscript "opt" to α and β, and write $\lambda_S = \lambda_{\min}(S)$ and $\lambda_R = \lambda_{\min}(R)$. We have

$$\begin{aligned}
\frac{r_{\text{adda}}}{r_{\text{sda}}} &= \frac{\beta-\lambda_R}{\alpha-\lambda_R}\cdot\frac{\alpha+\lambda_S}{\beta+\lambda_S} \\
&= 1 - \frac{(\lambda_R+\lambda_S)(\alpha-\beta)}{(\alpha-\lambda_R)(\beta+\lambda_S)}, \qquad (5.15)\\
\frac{r_{\text{adda}}}{r_{\text{sda-ss}}} &= \frac{\alpha-\lambda_S}{\alpha+\lambda_R} \\
&= 1 - \frac{\lambda_R+\lambda_S}{\alpha+\lambda_R}, \qquad (5.16)\\
\frac{r_{\text{sda-ss}}}{r_{\text{sda}}} &= \frac{\beta-\lambda_R}{\beta+\lambda_S}\cdot\frac{\alpha+\lambda_S}{\alpha-\lambda_S}\cdot\frac{\alpha+\lambda_R}{\alpha-\lambda_R} \\
&= 1 - \frac{(\lambda_R+\lambda_S)[\alpha(\alpha-\beta)-\lambda_S(\alpha-\lambda_R)-\alpha(\beta-\lambda_R)]}{(\beta+\lambda_S)(\alpha-\lambda_S)(\alpha-\lambda_R)}. \qquad (5.17)
\end{aligned}$$

If $\lambda_R+\lambda_S = 0$ (which happens in the critical case), then all three ratios are 1. In fact, for the critical case $r_{\text{adda}} = r_{\text{sda}} = r_{\text{sda-ss}} = 1$ and thus the three doubling algorithms converge linearly [11]. Suppose, in what follows, that $\lambda_R+\lambda_S > 0$, and recall (5.1). The first ratio

$$r_{\text{adda}}/r_{\text{sda}} \le 1 \quad \text{always},$$

with equality for $\alpha = \beta$, as expected. The ratio can be made much less than 1 if $\alpha/\beta \gg 1$. The second ratio

$$r_{\text{adda}}/r_{\text{sda-ss}} < 1 \quad \text{always}.$$

There is no definitive word on the third ratio because the sign of

$$\zeta \stackrel{\text{def}}{=} \alpha(\alpha-\beta) - \lambda_S(\alpha-\lambda_R) - \alpha(\beta-\lambda_R)$$

can change, dependent on different cases. If $\zeta > 0$, then SDA-ss is faster than SDA; otherwise it is slower.

It is worth pointing out that for SDA-ss it is very important how the shift-and-shrink (5.5) is done. For example, instead of (5.1), if

$$\max_i A_{(i,i)} < \max_i B_{(i,i)}. \tag{5.18}$$

Then we still have (5.14), but, instead of (5.3),

$$r_{\rm sda} = \frac{\beta - \lambda_S}{\beta + \lambda_S} \cdot \frac{\beta - \lambda_R}{\beta + \lambda_R}. \tag{5.19}$$

Then

$$\frac{r_{\rm sda}}{r_{\text{sda-ss}}} = \frac{\beta - \lambda_S}{\beta + \lambda_R} = 1 - \frac{\lambda_R + \lambda_S}{\beta + \lambda_R} < 1$$

always, indicating SDA-ss is slower than SDA. To overcome this, when (5.18) holds, SDA-ss should be applied to cMARE (3.1), instead, and as a by-product, Φ is computed as the minimal nonnegative solution to the complementary MARE of cMARE (3.1).

6 Numerical Examples

In this section, we shall present a few numerical examples to test numerical effectiveness of ADDA, in comparison with SDA and SDA-ss, as well as their ability to deliver entrywise relative accurate numerical solutions as argued in [33]. We will use two error measures to gauge accuracy in a computed solution $\widehat{\Phi}$: the Normalized Residual (NRes)

$$\text{NRes} = \frac{\|\widehat{\Phi}D\widehat{\Phi} - A\widehat{\Phi} - \widehat{\Phi}B + C\|_1}{\|\widehat{\Phi}\|_1(\|\widehat{\Phi}\|_1\|D\|_1 + \|A\|_1 + \|B\|_1) + \|C\|_1}, \tag{6.1}$$

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

$$\text{ERErr} = \max_{i,j} |(\widehat{\Phi} - \Phi)_{(i,j)}|/\Phi_{(i,j)} \tag{6.2}$$

which is not available in actual computations but is made available here for testing purpose. In the case of ERErr, the indeterminant 0/0 is treated as 0. In (6.1), we use ℓ_1-operator norm as an example. For all practical purpose, any matrix norm should work just fine.

Both errors defined by (6.1) and (6.2) are 0 if $\widehat{\Phi}$ is exact, but numerically they can only be made as small as $O(\boldsymbol{u})$ in general, where $\boldsymbol{u}$ is the unit machine roundoff. As we will see, to achieve $\widehat{\Phi}$ with deserved entrywise relative accuracy, tiny NRes (as tiny as $O(\boldsymbol{u})$) is not sufficient. To get some idea about what deserved entrywise relative accuracy should be expected, we will first outline some of the main perturbation results in [33] and then present them along with our numerical results.

6.1 Deserved entrywise relative accuracy

Let[6] W be perturbed to $\widetilde{W}$ in such a way that

$$|\widetilde{A} - A| \le \epsilon|A|,\ |\widetilde{B} - B| \le \epsilon|B|,\ |\widetilde{C} - C| \le \epsilon C,\ |\widetilde{D} - D| \le \epsilon D, \tag{6.3}$$

[6]We'll denote each perturbed counterpart by the same symbol but with a *tilde*.

where $0 \le \epsilon < 1$. It has been shown [33] that $\widetilde{\Phi}_{(i,j)} = 0$ if and only if $\Phi_{(i,j)} = 0$, under (6.3) and the assumption that both W and $\widetilde{W}$ are M-matrices. This fact paves the way to investigate how much each entry changes relatively.

Split A and B as

$$A = D_1 - N_1, \quad D_1 = \operatorname{diag}(A), \tag{6.4a}$$
$$B = D_2 - N_2, \quad D_2 = \operatorname{diag}(B). \tag{6.4b}$$

Correspondingly

$$A - \Phi D = D_1 - N_1 - \Phi D, \quad B - D\Phi = D_2 - N_2 - D\Phi,$$

and set

$$\lambda_1 = \rho(D_1^{-1}(N_1 + \Phi D)), \quad \lambda_2 = \rho(D_2^{-1}(N_2 + D\Phi)), \quad \lambda = \max\{\lambda_1, \lambda_2\}, \tag{6.5}$$
$$\tau_1 = \frac{\min_i A_{(i,i)}}{\max_j B_{(j,j)}}, \qquad \tau_2 = \frac{\min_j B_{(j,j)}}{\max_i A_{(i,i)}}. \tag{6.6}$$

If W is nonsingular, then $A - \Phi D$ and $B - D\Phi$ are nonsingular M-matrices by Theorem 3.1; so $\lambda_1 < 1$ and $\lambda_2 < 1$ [31, Theorem 3.15 on p.90] and thus $0 \le \lambda < 1$. If W is an irreducible singular M-matrix, then by Theorem 3.1(d)

1. if $u_1^{\mathrm{T}} v_1 > u_2^{\mathrm{T}} v_2$, then $\lambda_1 < 1$ and $\lambda_2 = 1$;
2. if $u_1^{\mathrm{T}} v_1 < u_2^{\mathrm{T}} v_2$, then $\lambda_1 = 1$ and $\lambda_2 < 1$;
3. if $u_1^{\mathrm{T}} v_1 = u_2^{\mathrm{T}} v_2$, then $\lambda_1 = \lambda_2 = 1$.

The third case $u_1^{\mathrm{T}} v_1 = u_2^{\mathrm{T}} v_2$, the so-called *critical case*, is rather extreme. It is argued in [18] that for the critical case for sufficiently small $\|\widetilde{W} - W\|$ there exists a constant θ such that

1. $\|\widetilde{\Phi} - \Phi\| \le \theta \|\widetilde{W} - W\|^{1/2}$;
2. $\|\widetilde{\Phi} - \Phi\| \le \theta \|\widetilde{W} - W\|$ if $\widetilde{W}$ is also singular.

This θ is only known by its existence.

The following results are taken from [33]. They are more informative, but do not work for the critical case. Suppose that W is a nonsingular M-matrix or an irreducible singular M-matrix with $u_1^{\mathrm{T}} v_1 \ne u_2^{\mathrm{T}} v_2$, ϵ in (6.3) is sufficiently small, and $\widetilde{W}$ is an M-matrix. We have

1.
$$|\Phi - \widetilde{\Phi}| \le \left[2\gamma\epsilon\, \mathbf{1}_{n,m} + O\left(\epsilon^2\right)\right] \Phi, \tag{6.7}$$

where γ are given by

$$(A - \Phi D)\Upsilon + \Upsilon(B - D\Phi) = D_1\Phi + \Phi D_2, \quad \gamma = \max_{i,j} \Upsilon_{(i,j)}/\Phi_{(i,j)}. \tag{6.8}$$

2.

$$|\Phi - \widetilde{\Phi}| \le \left[2mn\,\kappa\chi\,\epsilon + O\left(\epsilon^2\right)\right]\Phi, \tag{6.9}$$

where κ is given by

$$(A - \Phi D)\Phi_1 + \Phi_1(B - D\Phi) = C, \quad \kappa = \max_{i,j} (\Phi_1)_{(i,j)}/\Phi_{(i,j)},$$

and dependent on different cases, χ is given by

(a) for nonsingular M-matrix W,

$$\chi = \max\left\{\frac{1+\lambda_1+(1+\lambda_2)\tau_1^{-1}}{1-\lambda_1+(1-\lambda_2)\tau_1^{-1}}, \frac{1+\lambda_2+(1+\lambda_1)\tau_2^{-1}}{1-\lambda_2+(1-\lambda_1)\tau_2^{-1}}\right\} \le \frac{1+\lambda}{1-\lambda}. \tag{6.10}$$

(b) for singular M-matrix W with $u_1^{\mathrm{T}}v_1 \ne u_2^{\mathrm{T}}v_2$,

$$\chi = 2 \times \begin{cases} \dfrac{1+\lambda_1+2\tau_1^{-1}}{1-\lambda_1}, & \text{if } u_1^{\mathrm{T}}v_1 > u_2^{\mathrm{T}}v_2, \\ \dfrac{1+\lambda_2+2\tau_2^{-1}}{1-\lambda_2}, & \text{if } u_1^{\mathrm{T}}v_1 < u_2^{\mathrm{T}}v_2. \end{cases} \tag{6.11}$$

It is proved both γ and κ are finite [33]. Between (6.7) and (6.9), the linear term in the former is sharp while the one in the latter is not. But (6.9) is more informative in that it reveals the critical role played by the spectral radii λ_i in Φ's sensitivity.

In view of these perturbation results under (6.3) with $\epsilon = O(\boldsymbol{u})$, it is reasonable to define the *deserved entrywise relative accuracy* in any computed $\widehat{\Phi}$ to be that the associated ERErr is about $O(\gamma\boldsymbol{u})$ or $O(\kappa\chi\boldsymbol{u})$. In our examples in the next subsection, we shall compare ERErr against $(m+n)\gamma\boldsymbol{u}$ to verify if all of our computed $\widehat{\Phi}$ at convergence have the deserved entrywise relative accuracy.

6.2 Examples

All computations are performed in MATLAB with $\boldsymbol{u} = 1.11 \times 10^{-16}$. Optimal parameters as specified in section 5 are used for ADDA, SDA, and SDA-ss. Kahan's stopping criteria [34]:

$$\frac{(X_{k+1} - X_k)_{(i,j)}^2}{(X_k - X_{k-1})_{(i,j)} - (X_{k+1} - X_k)_{(i,j)}} \le \epsilon \cdot (X_{k+1})_{(i,j)} \quad \text{for all } i \text{ and } j \tag{6.12}$$

is used to terminate iterations, unless explicitly stated differently, where ϵ is a pre-selected tolerance. After numerous numerical experiments, we find that ϵ about 10^{-10} to 10^{-12} works the best for computed $\widehat{\Phi}$ to achieve its deserved accuracy without wasting the last iteration step.

Since ADDA is SDA if $\alpha_{\mathrm{opt}} = \beta_{\mathrm{opt}}$ for which there are numerous tests in literature, our examples will mainly focus on the case:

$$\alpha_{\mathrm{opt}} \stackrel{\mathrm{def}}{=} \max_i A_{(i,i)} \ne \beta_{\mathrm{opt}} \stackrel{\mathrm{def}}{=} \max_i B_{(i,i)}.$$

We will present five examples here. Table 6.1 summarizes rates of convergence for ADDA, SDA-ss, and SDA for the examples. Also included in the table are quantities $\varrho(I - \Phi\Psi)$ and $\varrho(I - \Psi\Phi)$ which tell us how accurately all inverses of M-matrices $I - X_kY_k$ and $I - Y_kX_k$ arising from the methods may be computed [34]. Table 6.2 summarizes various stability parameters in the first order error bounds in subsection 6.1. They can and will be used to explain the entrywise relative accuracy in computed $\widehat{\Phi}$.

Example	r_{adda}	$r_{\text{sda-ss}}$	r_{sda}	$\varrho(I-\Phi\Psi)$	$\varrho(I-\Psi\Phi)$
6.1	0.11	0.17	0.49	$1.4\cdot10^{-2}$	$1.4\cdot10^{-2}$
6.2	0.58	0.75	0.64	0.5	0.5
6.3	0.96	0.96	$1-2\cdot10^{-6}$	0.89	0.89
6.4	0.06	0.14	0.25	$6.3\cdot10^{-2}$	$6.3\cdot10^{-2}$
6.5	0.11	0.11	$1-2\cdot10^{-4}$	$5.9\cdot10^{-2}$	$1.1\cdot10^{-1}$

Table 6.1: Rates of convergence of ADDA, SDA-ss, and SDA

Example	λ_1	λ_2	2γ	κ	$\kappa\chi$
6.1	0.70	0.68	6.9	1.01	5.36
6.2	0.78	1.0	15.0	3.0	84.0
6.3	$1-6.2\cdot10^{-7}$	0.98	$1.6\cdot10^{6}$	16.5	$2.6\cdot10^{7}$
6.4	1	0.4	$3.2\cdot10^{2}$	30.9	$1.6\cdot10^{2}$
6.5	0.11	1	$2.1\cdot10^{4}$	1.1	$4.8\cdot10^{4}$

Table 6.2: Parameters in the first order error bounds

Example 6.1. $A, C, D, B \in \mathbb{R}^{n\times n}$ are given by

$$\begin{aligned}
&n=\ell^2, \quad T_\ell = \operatorname{tridiag}\left(-1, 4+\frac{200}{(\ell+1)^2}, -1\right) \in \mathbb{R}^{\ell\times\ell},\\
&A = \operatorname{tridiag}(-I_\ell, T_\ell, -I_\ell), \quad B = \xi\cdot A,\\
&\Phi = \frac{1}{50}\mathbf{1}_{n,n}, \quad C = \Phi D\Phi - A\Phi - \Phi B, \quad D = \frac{1}{50}\operatorname{tridiag}(1,2,1),
\end{aligned}$$

where $\operatorname{tridiag}(\cdot,\cdot,\cdot)$ constructs a tridiagonal or block tridiagonal matrix with its three arguments in an evident way. Making $\xi = 1$ recovers one of the examples in [4, 21]. We use $0 < \xi \neq 1$ to make $A \neq B$. In this example W is a nonsingular M-matrix. All entries in Φ are the same and consequently tiny NRes does imply tiny ERErr. Figure 6.1 shows two plots: the *left* one for NRes and the *right* one for ERErr, for the three methods for $\ell = 10$ and $\xi = 10$. The horizontal dotted line in the right plot is $(m+n)\gamma\boldsymbol{u}$. If its ERErr is below the dotted line, we regard the computed $\widehat{\Phi}$ to have the deserved entrywise relative accuracy. We will follow this way of presenting iteration histories in the rest of examples.

In this example, ADDA is the fastest, SDA-ss comes in second, and SDA is the slowest. All three algorithms are able to compute $\widehat{\Phi}$ with the deserved entrywise relative accuracy at convergence. ◇

Example 6.2. In this example, $m = n = 2$ and

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

Making $\xi = 1$ and scaling W by 10^{-3} recovers a null recurrent case example in [5] (see also [19, Test 7.2]). It can be verified that

$$\Phi = \frac{1}{2}\mathbf{1}_{2,2}, \quad \Psi = \frac{1}{2\xi}\mathbf{1}_{2,2}.$$

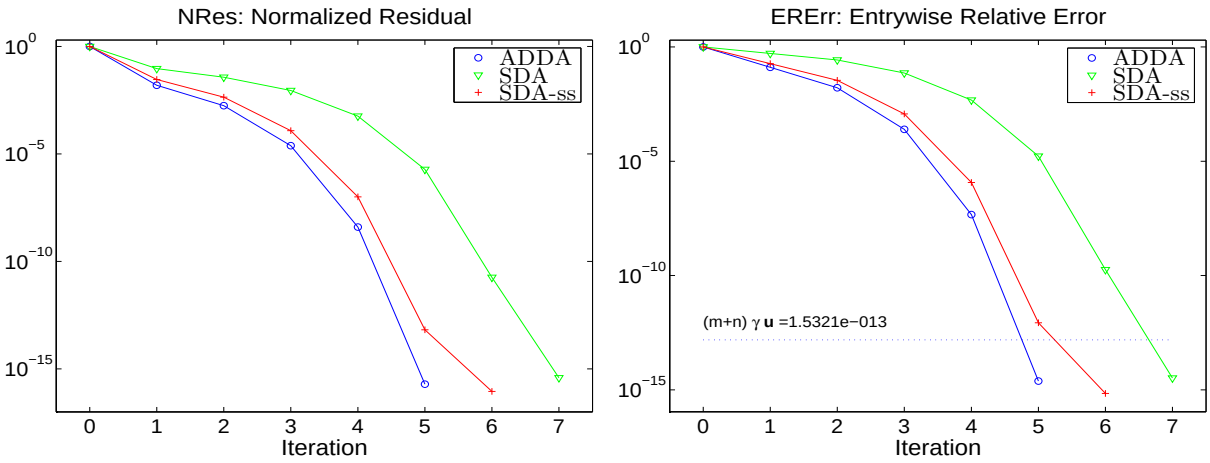


Figure 6.1: Example 6.1: $\ell = 10$ and $\xi = 10$.

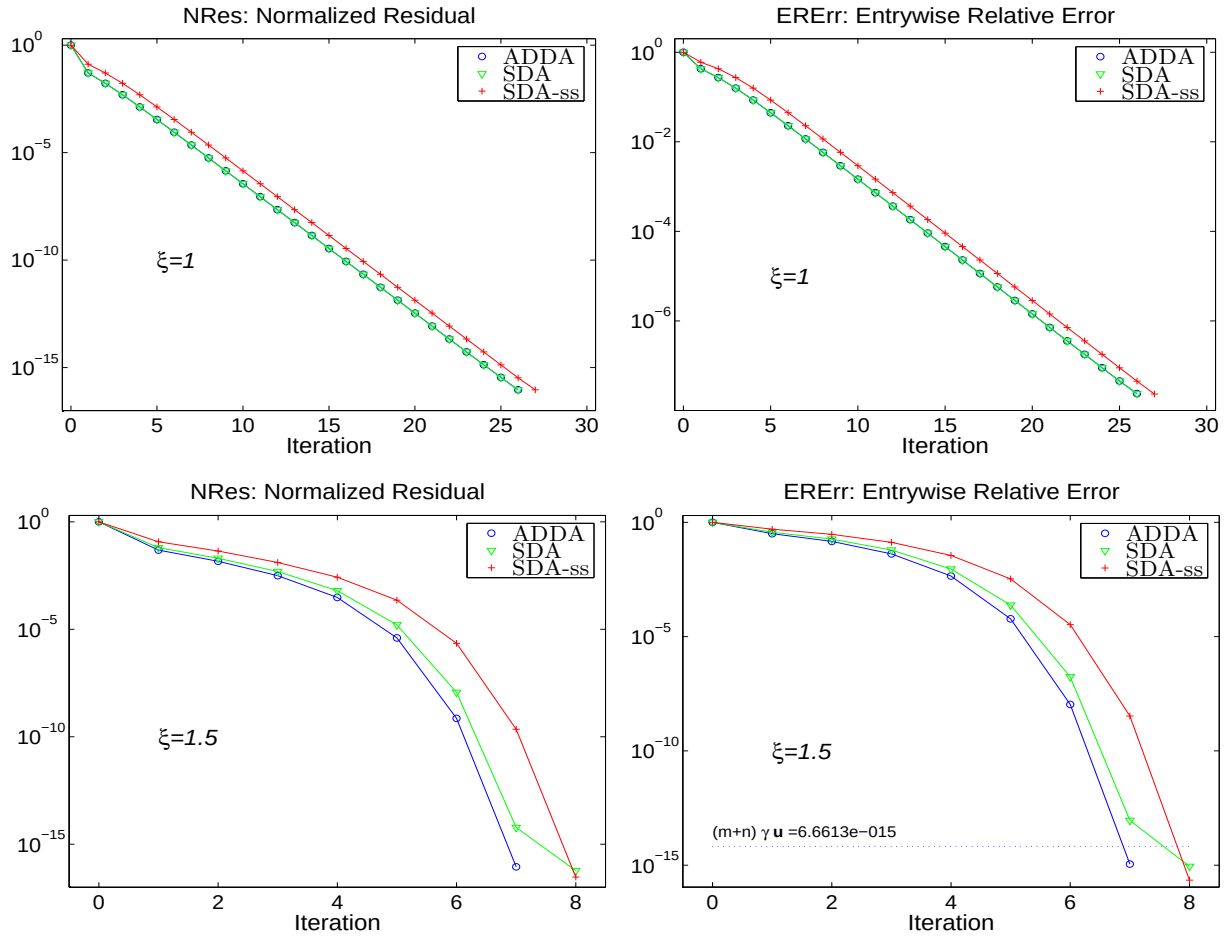


Figure 6.2: Example 6.2: *the top two plots* for $\xi = 1$ (the critical case) and the *bottom two plots* for $\xi = 1.5$. Stopping criteria for $\xi = 1$ is NRes $< 10^{-16}$ because Kahan's criteria (6.12) does not work well for linear convergent sequences [34]. Note SDA-ss is actually slower than SDA.

Note also W is an irreducible singular M-matrix:

$$W\mathbf{1}_4 = 0, \quad \begin{pmatrix} & \mathbf{1}_2 \\ \xi^{-1} \cdot \mathbf{1}_2 & \end{pmatrix}^{\mathrm{T}} W = 0.$$

The case in which $\xi = 1$ is the critical case. For the critical case, we know

1. $r_{\mathrm{adda}} = r_{\text{sda-ss}} = r_{\mathrm{sda}} = 1$ but all doubling algorithms still converge linearly [11];
2. $\varrho(I - \Phi\Psi) = \varrho(I - \Psi\Phi) = 1$, indicating $I - X_kY_k$ and $I - Y_kX_k$ are increasingly difficult to invert as they are becoming singular;
3. $\|\widetilde{\Phi} - \Phi\| \le \theta\|\widetilde{W} - W\|^{1/2}$ for some constant θ [18] provided $\|\widetilde{W} - W\|$ is sufficiently small. This means that we should expect errors no better than about $O(\sqrt{\boldsymbol{u}})$ in the computed Φ. The ERErr plot in Figure 6.2 for $\xi = 1$ certainly confirm this expectation.

But for $0 < \xi \neq 1$ all three methods converge quadratically. See Figure 6.2. Again we see that ADDA is the fastest, but this time SDA comes in second, and SDA-ss is the slowest. $\diamond$

Example 6.3. In this example $m = n = 2$, and

$$A = \begin{pmatrix} 100002 & -10^5 \\ -10^5 & 100002 \end{pmatrix}, \quad B = \begin{pmatrix} 3 & -1 \\ -1 & 3 \end{pmatrix}, \quad C = \begin{pmatrix} 1 & 1 - 2^{-p} \\ 1 - 2^{-p} & 1 \end{pmatrix}, \quad D = C.$$

Making $p = \infty$ and scaling W by 10^{-3} recovers a null recurrent case example in [5] (see also [9]). W is an irreducible singular M-matrix if $p = \infty$: $W\mathbf{1}_4 = 0$ and $\mathbf{1}_4^{\mathrm{T}}W = 0$ (the critical case), and a nonsingular M-matrix if $p > 0$. Thus the doubling algorithms converge linearly [11] for $p = \infty$ and quadratically for $p > 0$. See Figure 6.3. For both cases, there is little difference in performance for ADDA and SDA-ss with ADDA still being the faster one as expected, and both are much faster than SDA. Notice that for $p = \infty$, how much errors in going from X_0 to X_1 are suppressed for ADDA and SDA-ss but not so much for SDA. This is due to that both ADDA and SDA-ss are able to suppress the error components along eigenspace directions associated with nonzeros eigenvalues of R and S.

We notice from Figure 6.3 is that ERErr are not about $O(\boldsymbol{u})$ at convergence, but still below the dotted line $(m + n)\gamma\boldsymbol{u}$. $\diamond$

Example 6.4.

$$A = \begin{pmatrix} 3 & -1 & & \\ & 3 & \ddots & \\ & & \ddots & -1 \\ -1 & & & 3 \end{pmatrix} \in \mathbb{R}^{n\times n},\ C = 2I_n,\ B = 10A,\ D = 10C.$$

W is an irreducible singular M-matrix: $W\mathbf{1}_{2n} = 0$, but $u_1^{\mathrm{T}}v_1 \neq u_2^{\mathrm{T}}v_2$. For testing purpose, we have computed for $n = 100$ an "exact" solution Φ and Ψ by the computerized algebra system *Maple* with 100 decimal digits. This[7] "exact" solution Φ's entries range from $5.7 \cdot 10^{-31}$ to $6.3 \cdot 10^{-2}$ and Ψ's entries range from $5.7 \cdot 10^{-30}$ to $6.3 \cdot 10^{-1}$. Despite of this wide range

[7]The "exact" Φ and Ψ by *Maple* suggest $\Psi = 10\Phi$.

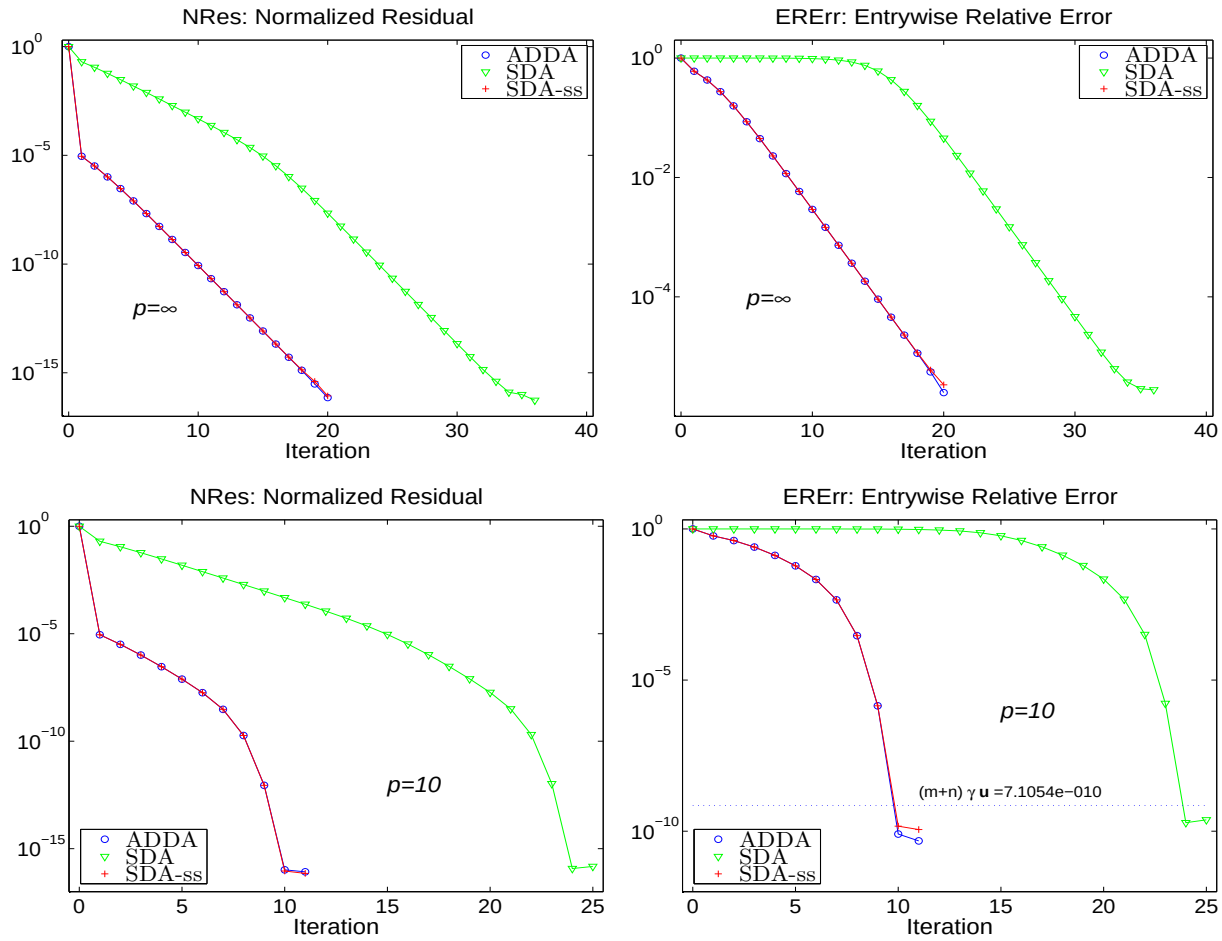


Figure 6.3: Example 6.3: *the top two plots* for $p = \infty$ (the critical case) and the *bottom two plots* for $p = 10$. Stopping criteria for $p = \infty$ is NRes $< 10^{-16}$ because Kahan's criteria (6.12) does not work well for linear convergent sequences [34]. In the plots, SDA and SDA-ss are mostly indistinguishable.

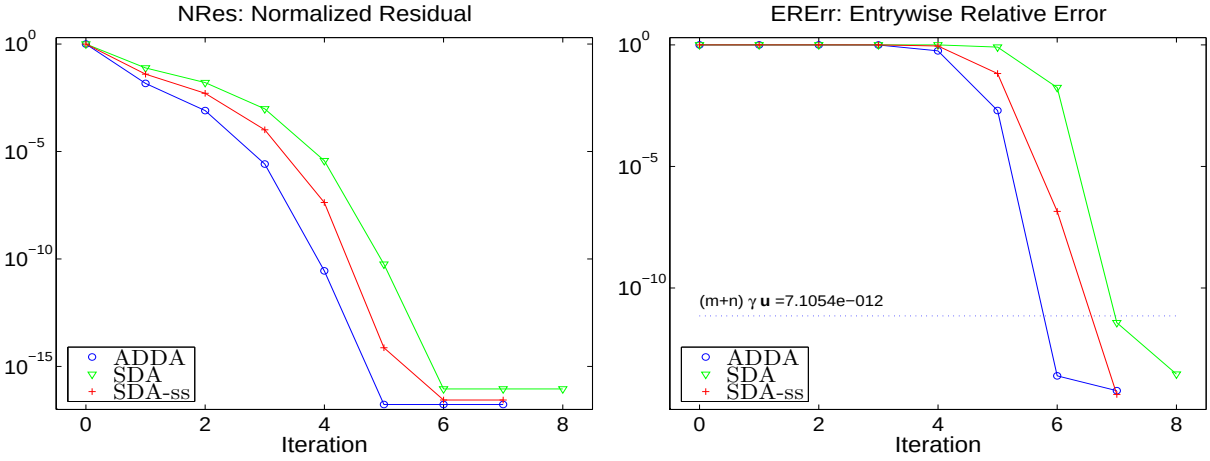


Figure 6.4: Example 6.4. Uneven convergence towards entries with widely different magnitudes. ERErr is still large even when NRes is already tiny before $\widehat{\Phi}$ is fully entrywise converged.

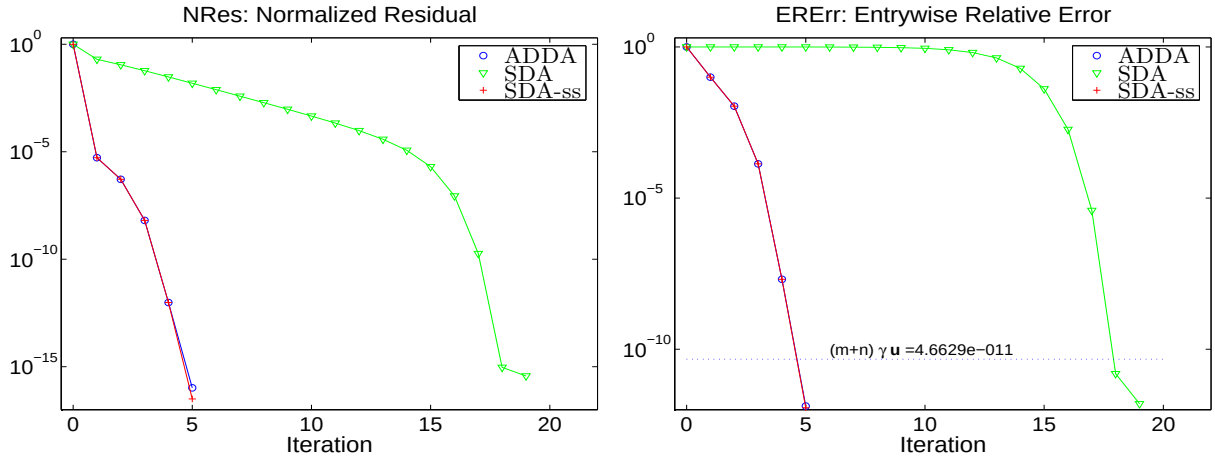


Figure 6.5: Example 6.5. ADDA and SDA-ss are barely distinguishable. Both are much faster than SDA.

of magnitudes in their entries, all three methods are able to deliver computed $\widehat{\Phi}$ and $\widehat{\Psi}$ with entrywise relative errors at the level of $O(\boldsymbol{u})$. See Figure 6.4. Notice how little improvements in ERErr for the first four iterations, even though NRes decrease substantially during the period. For example, at iteration 5,

	ADDA	SDA-ss	SDA
NRes	$1.6950 \cdot 10^{-17}$	$7.4124 \cdot 10^{-15}$	$5.7149 \cdot 10^{-11}$
ERErr	$2.0093 \cdot 10^{-3}$	$6.6470 \cdot 10^{-2}$	$8.1583 \cdot 10^{-1}$

This is because it takes a while for the tiny entries to gain some relative accuracy. ◇

Example 6.5 ([5, 19])**.** This is essentially the example of a positive recurrent Markov chain with nonsquare coefficients, originally from [5]. Here

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

It is is known $\Phi = \frac{1}{18} \cdot \mathbf{1}_{2,18} = \Psi^{\mathrm{T}}$. See Figure 6.5 for the performance of the three methods on this example. ADDA and SDA-ss are about the same, but both are much faster than SDA. ◇

Along with five examples above, we have conducted numerous other tests, including many random ones. We come up with the following two conclusions about speed and accuracy for the three doubling algorithms:

- ADDA is always the fastest among all three. SDA-ss can even run slower than SDA when $\max_i A_{(i,i)}$ and $\max_j B_{(j,j)}$ are about the same or differ within a fact of two. However, when $\max_i A_{(i,i)}$ and $\max_j B_{(j,j)}$ differ by a factor over, say 10 for example, ADDA and SDA-ss take about the same number of iterations to deliver fully converged $\widehat{\Phi}$ and both can be much faster than SDA.
- With the suggested optimal parameter selections in section 5, all three methods are capable of delivering computed $\widehat{\Phi}$ with the deserved entrywise relative accuracy as warranted by the input data.

7 Concluding Remarks

We have presented a new doubling algorithm for the unique minimal nonnegative solution Φ of MARE (1.1). It is the product of combining the alternating directional idea in ADI for the Sylvester equation (see [7, 32]) and the idea of SDA in [21]. For this reason, we name our new method ADDA (Alternating-Directional Doubling Algorithm). Compared with two existing double algorithms – SDA in [21] and SDA-ss in [9], *our ADDA is always the fastest* as we argued first through theoretical convergence analysis and then numerical tests. Finally, all three methods are able to compute Φ as entrywise accurately as the perturbation analysis in [33] suggests.

References

[1] A. S. Alfa, J. Xue, and Q. Ye, *Accurate computation of the smallest eigenvalue of a diagonally dominant M-matrix*, Math. Comp., 71 (2002), pp. 217–236.

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

[3] Z. Bai, J. Demmel, and M. Gu, *An inverse free parallel spectral divide and conquer algorithm for nonsymmetric eigenproblems*, Numer. Math., 76 (1997), pp. 279–308.

[4] Z.-Z. Bai, X.-X. Guo, and S.-F. Xu, *Alternately linearized implicit iteration methods for the minimal nonnegative solutions of the nonsymmetric algebraic Riccati equations*, Numer. Linear Algebra Appl., 13 (2006), pp. 655–674.

[5] 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.

[6] P. Benner, *Contributions to the Numerical Solution of Algebra Riccati Equations and Related Eigenvalue Problems*, Logos, Berlin, Germany, 1997.

[7] P. Benner, R.-C. Li, and N. Truhar, *On ADI method for Sylvester equations*, J. Comput. Appl. Math., 233 (2009), pp. 1035–1045.

[8] 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.

[9] D. A. Bini, B. Meini, and F. Poloni, *Transforming algebraic Riccati equations into unilateral quadratic matrix equations*, Numer. Math., 116 (2010), pp. 553–578.

[10] A. Y. Bulgakov and S. K. Godunov, *Circular dichotomy of the spectrum of a matrix*, Siberian Mathematical Journal, 29 (1988), pp. 734–744.

[11] 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.

[12] 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.

[13] 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.

[14] M. Fiedler, *Special Matrices and Their Applications in Numerical Mathematics*, Dover Publications, Inc., Mineola, New York, 2nd ed., 2008.

[15] S. K. Godunov, *Problem of the dichotomy of the spectrum of a matrix*, Siberian Mathematical Journal, 27 (1986), pp. 649–660.

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

[17] ——, *A new class of nonsymmetric algebraic Riccati equations*, Linear Algebra Appl., 426 (2007), pp. 636–649.

[18] C.-H. Guo and N. Higham, *Iterative solution of a nonsymmetric algebraic Riccati equation*, SIAM J. Matrix Anal. Appl., 29 (2007), pp. 396–412.

[19] 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.

[20] 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.

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

[22] J. Juang, *Existence of algebraic matrix Riccati equations arising in transport theory*, Linear Algebra Appl., 230 (1995), pp. 89–100.

[23] J. Juang and W.-W. Lin, *Nonsymmetric algebraic Riccati equations and Hamiltonian-like matrices*, SIAM J. Matrix Anal. Appl., 20 (1998), pp. 228–243.

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

[25] A. N. Malyshev, *Computing invariant subspaces of a regular linear pencil of matrices*, Siberian Mathematical Journal, 30 (1989), pp. 559–567.

[26] C. D. Meyer, *Stochastic complementation, uncoupling Markov chains, and the theory of nearly reducible systems*, SIAM Rev., 31 (1989), pp. 240–272.

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

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

[29] R. A. Smith, *Matrix equation $XA + BX = C$*, SIAM J. Appl. Math., 16 (1968), pp. 198–201.

[30] X. Sun and E. S. Quintana-Ortí, *Spectral division methods for block generalized schur decompositions*, Math. Comp., 73 (2004), pp. 1827–1847.

[31] R. Varga, *Matrix Iterative Analysis*, PrenticeHall, Englewood Cliffs, NJ, 1962.

[32] E. L. Wachspress, *The ADI Model Problem*, Windsor, CA, 1995. Self-published (`www.netlib.org/na-digest-html/96/v96n36.html`).

[33] J. Xue, S. Xu, and R.-C. Li, *Accurate solutions of M-matrix algebraic Riccati equations*, Numer. Math. to appear.

[34] ——, *Accurate solutions of M-matrix Sylvester equations*, Numer. Math. to appear.