Numerical Studies for M-Matrix Algebraic Riccati Equations

by

WEICHAO WANG

Presented to the Faculty of the Graduate School of
The University of Texas at Arlington in Partial Fulfillment
of the Requirements
for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT ARLINGTON
August 2013

To my mother Yan, my father Jinye who made me who I am
and my dearest grandma Guihua who set an example of my life.

ACKNOWLEDGEMENTS

Foremost, I would like to express my sincere gratitude to my advisor Prof. Rencang Li for the continuous support of my Ph.D study and research, for his patience, motivation, enthusiasm, and immense knowledge. His guidance helped me all the time in research and writing of this thesis. I could not have imagined having a better advisor and mentor for my Ph.D study.

I would like to thank the rest of my thesis committee: Prof. Chaoqun Liu, Prof. Gaik Ambartsoumian, Prof. Hristo Kojouharov and Dr. Jianzhong Su, for their encouragement, insightful comments, and hard questions.

My sincere thanks also goes to Dr.Wei-guo Wang, for helping me a lot with my research and giving me valuable advice on my PhD study.

Many thanks to UTA Mathematics departmental staffs, without whom I could not have such a comfortable environment for my PhD study.

I have to give my thanks to professors in UTA and USTC for their effort on teaching me.

Moreover, I would like to thank my lovely friends – Fei Xu for accompanying me for three years, Jinjiang He, Mengmeng Yuan, Yiguang Hou, Wei Li, Shan Gao, Yang Hu, Jingxuan Wang, Xi Chen, Xiao Wang, Zhengjie Wang and Lanjiang Zhou for encouraging me all the time.

Last but not least, I would like to thank my father Jinye and my mother Yan for giving birth to me in the first place and supporting me throughout my life.

July 18, 2013

ABSTRACT

Numerical Studies for M-Matrix Algebraic Riccati Equations

Weichao Wang, Ph.D.

The University of Texas at Arlington, 2013

Supervising Professor: Ren-cang Li

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.

A deflation technique is then presented for an irreducible singular 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.

TABLE OF CONTENTS

Appendix

LIST OF ILLUSTRATIONS

LIST OF TABLES

Table Page

CHAPTER 1

Introduction

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

$$\mathscr{R}(X) := XDX - AX - XB + C = 0, \tag{1.0.1}$$

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

$$W = \begin{array}{c c} & \begin{array}{cc} m & n \end{array} \\ \begin{array}{c} m \\ n \end{array} & \begin{pmatrix} B & -D \\ -C & A \end{pmatrix} \end{array}, \tag{1.0.2}$$

and W is a nonsingular or an irreducible singular M-matrix[2]. This kind of Riccati equations arise in applied probability and transportation theory and have been attracting a lot of attention lately. See [19, 21, 23, 24, 25, 26, 35] and the references therein.

In [24], 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. The idea of using a doubling algorithm for Riccati-type equations traces back to 1970s (see [2] and references therein). Recent resurgence of interests in the idea, however, attributes to [15, 14] and has since led to efficient doubling algorithms for various nonlinear matrix equations. SDA is very fast and efficient for small to medium size MAREs as it is globally and quadratically convergent. Later in [22], it was argued that SDA still works for the case in which W is an irreducible singular M-matrix.

[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 [42] to better reflect its characteristics.

[2] The definition of the M-matrix will be introduced in Chapter 2.

The algorithm has to select a parameter that is no smaller than the largest diagonal entries in both A and B. Such a choice of the parameter ensures the following:

1. An elegant theory of global and quadratic convergence [22, 24], except for the *null recurrent* or *critical case* [22, p.1085] (see also Theorem 6.3.1(d)) for which only linear convergence is ensured [13];
2. Computed Φ has an entrywise relative accuracy as the input data deserves, as argued recently in [42].

Consequently, SDA has since emerged as one of the most efficient algorithms. After that, the doubling algorithm called *SDA-ss* of Bini, Meini, and Poloni [11] which combined a shrink-and-shift approach of Ramaswami [34] came out to improve the convergence rate. One of the major contributions of this thesis is a new algorithm–ADDA, which is optimal among all the known doubling algorithms.

This thesis is devoted to studying algorithms for the MARE (1.0.1).

Chapter 2 will present some preliminary knowledge in numerical analysis and linear algebra that is necessary for this thesis. Especially, this chapter will give the definition and some basic properties of the M-matrix.

In Chapter 3, we will prove a theorem of the MARE which all these doubling algorithms base on. We will apply some general numerical methods in Chapter 2 to give a rigorous proof on the existence of the unique minimal nonnegative solution.

Next in Chapter 4, we will give a new alternating directional implicit(ADI) method and show its convergence rate on the MARE. Although the new ADI method is still linearly convergent as the previous one, it is the idea of using two parameters instead of one in the new method that inspires us to develop our ADDA which turns out to be optimal.

Chapter 5 will introduce the Smith method and explain why higher order algorithms are less efficient than the doubling algorithms.

Our main contributions are described in detail in Chapter 6 and Chapter 7.

Chapter 6 will first lay out the framework of ADDA and analyze its convergence properties. Next, we will compare the convergence rates of ADDA, SDA and SDA-ss to indicate that (the optimal) ADDA is the fastest among all doubling algorithms derivable from bilinear transformations.

In the so-called critical case [13] of the MARE, those doubling algorithms in Chapter 6 are linearly convergent. For this reason, in Chapter 7 we will propose a deflation approach before applying our ADDA method. The new method is called D-ADDA. As in Chapter 6, we will lay out framework of D-ADDA and the convergence properties. Two efficient numerical realizations of the framework will be given. Next we will compare our D-ADDA method with the shifting approach of Guo, Iannazzo, and Meini [22].

Chapter 8 will show some numerical examples to support the analytical results in Chapter 6 and Chapter 7.

Finally, in Chapter 9 we will give our conclusion on the whole thesis.

CHAPTER 2

Preliminaries

Before discussing numerical methods for matrix equations, let us recall some basic but very important concepts and general numerical methods that are quite useful to better understand this thesis.

2.1 Notation

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. $X_{(i,j)}$ or $X_{i,j}$ refers to its (i, j)th entry;
2. x_{ij} also refers to X's (i, j)th entry;
3. when $m = n$, $\operatorname{diag}(X)$ is the diagonal matrix with the same diagonal entries as X's, $\rho(X)$ is the spectral radius of X, and

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

Inequality $X \leq Y$ means $X_{(i,j)} \leq Y_{(i,j)}$ for all (i, j), and similarly for $X < Y$, $X \geq Y$, and $X > Y$. $\|X\|$ denotes some (general) matrix norm of X. Specifically, $\|X\|_p = \max\limits_{\|x\|_p=1} \|Xx\|_p$ is the l_p-operator norm of X, where $\|x\|_p$ is the l_p-norm of the vector x. Hence $\|X\|_\infty = \max\limits_{\|x\|_\infty=1} \|Xx\| = \max\limits_i \sum\limits_j |X_{ij}|$ is the maximum absolute row sum of matrix X.

2.2 Newton's Method

As we know, Newton's method is a general procedure that can be applied in many diverse situations. When specialized to the problem of locating a zero of a real-valued function of a real variable, it is often called the *Newton-Raphson iteration.*

Suppose r is a real solution of equation $f(x) = 0$, where $f \in \mathbf{C}^2(\mathbb{R})$ and x is a good[1] approximation to r such that $r = x + h$. By Taylor's Theorem,

$$0 = f(r) = f(x+h) = f(x) + hf'(x) + \frac{1}{2}h^2 f''(\xi_n), \tag{2.2.1}$$

where ξ_n is between x_n and $x_n + h$. If $|h|$ is small, then it is reasonable to ignore the last term of (2.2.1), under which condition we have

$$0 = f(x) + hf'(x).$$

It is hoped that

$$x + h = x - \frac{f(x)}{f'(x)}$$

is a better approximation to r. So if we write the iteration as

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} \quad n \geq 0, \tag{2.2.2}$$

then $\{x_n\}$ is a sequence of estimates of r.

Next theorem indicates quadratic convergence of Newton's method.

Theorem 2.2.1 ([27])**.** *Let f'' be continuous and let r be a simple zero of f. Then there is a neighborhood of r and a constant C such that if Newton's method is started in that neighborhood, then the successive points become quadratically convergent to r as*

$$|x_{n+1} - r| \leq C(x_n - r)^2, \quad n \geq 0.$$

[1] This good means x is near r, i.e. h is small.

Proof. Let $e_n = r - x_n$ and from (2.2.1),(2.2.2),

$$e_{n+1} = e_n + \frac{f(x_n)}{f'(x_n)} = -\frac{1}{2}\frac{f''(\xi_n)}{f'(x_n)}e_n^2.$$

If we can find a positive number C such that

$$|\frac{f''(\xi_n)}{2f'(x_n)}| \leq C,$$

then we will have

$$|e_{n+1}| \leq Ce_n^2,$$

which means quadratical convergence. Actually, if we define

$$c(\delta) = \frac{1}{2}\max_{|x-r|\leq\delta}|f''(x)|/\min_{|x-r|\leq\delta}|f'(x)|, \quad \delta > 0, \tag{2.2.3}$$

then we can pick δ small enough to make the denominator of (2.2.3) positive and bounded below because $f'(r) \neq 0$. Then if necessary, we could make δ less to ensure the numerator is also bounded above, which is possible since as δ converges to zero, $c(\delta)$ converges to $\frac{1}{2}|f''(r)|/|f'(r)|$. So we can pick a proper δ such that $c(\delta)$ in (2.2.3) is bounded above. □

We can generalize the Newton's method to solve nonlinear matrix equations. If we apply Newton's method to MARE (1.0.1). For any matrix norm, the Riccati function $\mathscr{R}$ is a mapping from $\mathbb{R}^{m\times n}$ into itself. $\mathscr{R}'_X$ is[2] a linear map from $\mathbb{R}^{m\times n}$ to $\mathbb{R}^{m\times n}$ given by

$$\mathscr{R}'_X(Z) = -(A - XD)Z + Z(DX - B). \tag{2.2.4}$$

Newton's method for an MARE (1.0.1)

$$X_{n+1} = X_n - (\mathscr{R}'_{X_n})^{-1}\mathscr{R}(X_n), \quad n = 0, 1, 2, \cdots, \tag{2.2.5}$$

[2] Here $\mathscr{R}'_X$ stands for first Fréchet derivative of $\mathscr{R}$.

is well-defined if all $\mathscr{R}'_{X_n}$ are invertible. Now with (2.2.4) and (2.2.5), we have the Newton's iteration as

$$(A - X_n D)X_{n+1} + X_{n+1}(B - DX_n) = C - X_n D X_n. \tag{2.2.6}$$

Actually those interested readers are referred to [33] to get more detail in the existence for $\mathscr{R}'_X$ and the convergence theory of (2.2.6).

2.3 Fixed-Point Iteration

As we saw in (2.2.5), the right-hand-side of the equation is a function of the matrix X_n, which is a point in $\mathbb{R}^{m\times n}$. Actually the Newton's method is an example of algorithms called *functional iteration* with form

$$X_{n+1} = F(X_n), \quad n \geq 0. \tag{2.3.1}$$

Suppose that

$$\lim_{n\to\infty} X_n = S.$$

If F is continuous, then

$$F(S) = F(\lim_{n\to\infty} X_n) = \lim_{n\to\infty} F(X_n) = \lim_{n\to\infty} X_{n+1} = S.$$

Thus, $F(S) = S$, and we call S a *fixed-point* of function F. We could think of a fixed-point as a matrix that the function "locks onto" in the iterative process.

Often a mathematical problem can be reduced to the problem of finding a fixed-point of a function. Very interesting applications occur in differential equations, optimization theory, and other areas. Usually the function F whose fixed-points are sought will be a mapping from one vector space into another. Here we extend F to be a map from a closed set $C \subset \mathbb{R}^{n\times n}$ into itself. The theorem to be proved concerns

Contractive mappings. A mapping (or function) F is said to be *contractive* if there exists a number $\lambda < 1$ such that

$$\|F(X) - F(Y)\| \leq \lambda \|X - Y\|, \tag{2.3.2}$$

for all points (matrices) X and Y in the domain of F, where $\| \cdot \|$ is some matrix norm. The *Contractive Mapping Theorem* indicates the uniqueness of fixed-point.

Theorem 2.3.1 ([27])**.** *Let F be a contractive mapping of a closed set $C \subset \mathbb{R}^{n\times n}$ into C. Then F has a unique fixed-point. Moreover, this fixed-point is the limit of every sequence obtained from* (2.3.1) *with any starting point $X_0 \in C$.*

Proof. From the property (2.3.2) and construct iteratively,

$$\begin{aligned}\|X_{n+1} - X_n\| &= \|F(X_n) - F(X_{n-1})\| \\ &\leq \lambda \|X_n - X_{n-1}\| \\ &\leq \lambda^2 \|X_{n-1} - X_{n-2}\| \\ &\leq \cdots \\ &\leq \lambda^n \|X_1 - X_0\|,\end{aligned}$$

where $\lambda < 1$. Since X_n can be written as

$$X_n = X_0 + (X_1 - X_0) + (X_2 - X_1) + \cdots + (X_n - X_{n-1}),$$

we see that the sequence $\{X_n\}$ converges if and only if the series

$$\sum_{n=1}^{\infty}(X_n - X_{n-1})$$

converges. To prove that this series converges, it suffices to show that the series

$$\sum_{n=1}^{\infty} \|X_n - X_{n-1}\|$$

converges. This is easy since

$$\sum_{n=1}^{\infty} \|X_n - X_{n-1}\| \leq \sum_{n=1}^{\infty} \lambda^{n-1} \|X_1 - X_0\| = \frac{1}{1-\lambda} \|X_1 - X_0\|.$$

Hence $\{X_n\}$ is convergent, say, with limit S. Moreover, we can see a contractive function is continuous from its definition. So S is a fixed-point.

Suppose there are two fixed-points X and Y, then

$$\|Y - X\| = \|F(Y) - F(X)\| \leq \lambda \|Y - X\| \Rightarrow |1 - \lambda| \cdot \|X - Y\| \leq 0 \Rightarrow X = Y.$$

Therefore, with $\lambda < 1$, we have $X = Y$, which proves the uniqueness of the fixed-point. □

2.4 Kronecker Products

Let X be m-by-n. Then vec(X) is defined to be a column vector of size $m \cdot n$ made of the columns of X stacked atop one another from left to right. Let A be an m-by-n matrix and B be a p-by-q matrix. Then $A \otimes B$, the *Kronecker Product* of A and B, is the $(m \cdot p)$-by-$(n \cdot q)$ matrix

$$\begin{pmatrix} a_{11} \cdot B & \cdots & a_{1n} \cdot B \\ \vdots & & \vdots \\ a_{m1} \cdot B & \cdots & a_{mn} \cdot B \end{pmatrix}$$

The following lemma is a useful way to express a matrix equation by Kronecker products and the vec$(\cdot)$ operator.

Lemma 2.4.1 ([16],[32])**.** *Let A be m-by-m, B be n-by-n, and X and C be m-by-n. Then the following properties hold:*

(a) $\text{vec}(AX) = (I_n \otimes A) \cdot \text{vec}(X)$.

(b) $\text{vec}(BX) = (B^T \otimes I_m) \cdot \text{vec}(X)$.

(c) $\text{vec}(AX + XB) = (I_n \otimes A + B^T \otimes I_m) \cdot \text{vec}(X)$.

The proof of this lemma is quite easy. The first two parts can be proved by comparing both sides of the equations and the last part directly follows from the first two.

2.5 Irreducible Matrix

Definition 2.5.1. *For $n \geq 2$, an $n \times n$ complex matrix A is* reducible *if there exists an $n \times n$ permutation matrix*[3] *P such that*

$$PAP^T = \begin{array}{c c} & \begin{array}{cc} r & n-r \end{array} \\ \begin{array}{c} r \\ n-r \end{array} & \begin{pmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{pmatrix} \end{array},$$

where $1 \leq r < n$. If no such permutation matrix exists, then A is called irreducible.

If A is nonnegative, there is another equivalent definition for irreducible matrix.

Definition 2.5.2. *Suppose A is nonnegative. A is called irreducible if for each pair of indices i and j, there exists an $m \in \mathbb{N}$ such that $(A^m)_{ij} \neq 0$.*

2.6 Definition for M-matrix.

Definition 2.6.1. *A matrix $A \in \mathbb{R}^{n\times n}$ is called a Z-matrix if $A_{(i,j)} \leq 0$ for all $i \neq j$. Any Z-matrix A can be written as $sI - N$ with $N \geq 0$. It is called an M-matrix if $s \geq \rho(N)$, a singular M-matrix if $s = \rho(N)$, and a nonsingular M-matrix if $s > \rho(N)$.*

2.7 Equivalent Definitions for M-matrix

With the definition of the M-matrix, we have the following equivalent statements for M-matrices which will be used in this thesis.

[3] A permutation matrix is a square matrix which in each row and each column has one and only one entry unity and all others zero.

Lemma 2.7.1 ([9, 17, 32, 38])**.** *The following are equivalent for a Z-matrix A:*

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

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

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

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

Proof. We will prove the lemma as follows: (a)$\Longleftrightarrow$(b), (b)$\Longleftrightarrow$(c), (a)$\Longleftrightarrow$(d).

(a)$\Longleftrightarrow$(b)

($\Longrightarrow$) Suppose A is a nonsingular M-matrix, by definition we have $A = sI - N$, where $N \geq 0$ and $s > \rho(N)$. So

$$\rho\left(\frac{N}{s}\right) < 1, \quad N \geq 0. \tag{2.7.1}$$

Consider

$$\left[I - \left(\frac{N}{s}\right)\right] \cdot \left[I + \left(\frac{N}{s}\right) + \left(\frac{N}{s}\right)^2 + \cdots + \left(\frac{N}{s}\right)^k\right] = I - \left(\frac{N}{s}\right)^{k+1}.$$

When $k \to \infty$, the right-hand side approaches I with condition (2.7.1). So

$$\left[I - \left(\frac{N}{s}\right)\right]^{-1} = \sum_{k=0}^{\infty} \left(\frac{N}{s}\right)^k \geq 0,$$

which means $A^{-1} \geq 0$.

($\Longleftarrow$) If $A^{-1} \geq 0$, then A is nonsingular. Suppose λ is any eigenvalue of N,

$$\begin{aligned}
&\lambda x = Nx \\
\Rightarrow &|\lambda||x| = N|x| \\
\Rightarrow &(sI - N)|x| \leq (s - |\lambda|)|x| \\
\Rightarrow &0 \leq |x| \leq (s - |\lambda|)(sI - N)^{-1}|x| \\
\Rightarrow &s - |\lambda| \geq 0 \\
\Rightarrow &s \geq |\lambda| \\
\Rightarrow &s \geq \rho(N).
\end{aligned}$$

Moreover, from $|x| \neq 0$, we have $s \neq |\lambda|$. Hence $s > \rho(N)$.

(b)$\Longleftrightarrow$(c)

($\Longrightarrow$) If $A^{-1} \geq 0$, then A^{-1} exists and $A^{-1} \neq 0$. There exists a vector $v > 0$, such that $A^{-1}v > 0$. Let $u = A^{-1}v$. It is obvious that $u > 0$. So $Au = v > 0$.

($\Longleftarrow$) If there exists a vector $x = (x_1, x_2, \cdots, x_n)^T > 0$ such that $Ax > 0$. Let D=diag$(x_1, x_2, \cdots, x_n)$, then $ADe > 0$, where $e = (1, 1, \cdots, 1)^T$. AD is diagonally dominant. Split $A = E - N$, where $E = \text{diag}(A)$. Then $AD = ED - ND$, $A = (ED - ND)D^{-1} = ED(I - D^{-1}E^{-1}ND)D^{-1}$. Let $H = D^{-1}E^{-1}ND$. Then we have $\rho(H) < 1$ since

$$\rho(H) \leq \|H\|_\infty = \max_i \sum_{j \neq i} \frac{|(AD)_{i,j}|}{|(AD)_{i,i}|} < 1.$$

Because $H \geq 0$, $A^{-1} = D(I - H)^{-1}D^{-1}E^{-1} \geq 0$.

(a)$\Longleftrightarrow$(d)

($\Longrightarrow$) From $A = sI - N$, we have $\lambda(A) = s - \lambda(N)$. Here $\lambda(A)$ stands for an eigenvalue of A and $\lambda(N)$ is the corresponding eigenvalue of N satisfies the equation. If $\lambda(N) = \alpha + \mathbf{i}\beta$, where α and β are real numbers and $\mathbf{i}$ is the imaginary unit satisfying $\mathbf{i}^2 = -1$. Hence $\lambda(A) = (s - \alpha) - \mathbf{i}\beta$. As assumed $s > \rho(N) \geq |\lambda(N)| = \sqrt{\alpha^2 + \beta^2}$, we have $\text{Re}(\lambda(A)) > 0$.

($\Longleftarrow$) Now suppose $\text{Re}(\lambda(A)) > 0$. This means that there is a real number γ such that the circle centered at γ with radius γ contains all eigenvalues of A. Let s be any real number satisfies $s > \max\{2\gamma, \max_i |a_{ii}|\}$, and set $N = sI - A$. Then $N \geq 0$ and by $\lambda(N) = s - \lambda(A)$, we have $|\lambda(N)| = |s - \lambda(A)| < s$, i.e. $s > \rho(N)$. Moreover, since $\text{Re}(\lambda(A)) > 0$, $\lambda(A) \neq 0$, which means A is nonsingular. □

REMARK **2.7.1.** *From the proof of (a)$\Longleftrightarrow$(b), we can similarly prove the following lemma.*

Lemma 2.7.2 ([17, 38])**.** *Given A, $\rho(A) < 1$ if and only if $(I - A)^{-1}$ exists and*

$$(I - A)^{-1} = \sum_{k=0}^{\infty} A^k.$$

2.8 Properties of M-matrices

Lemma 2.8.1 collects a few properties of M-matrices, important to our later analysis, where Item (e) can be found in [31].

Lemma 2.8.1 ([9, 17, 38])**.** *Let $A, B \in \mathbb{R}^{n \times n}$, and suppose A is an M-matrix and B is a Z-matrix.*

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

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

(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 $\lambda_{\min}(A) \leq \max_i A_{(i,i)}$.*

(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.

Proof. (a) The proof is essentially the same as the proof of Theorem 3.12 in [38].

Let D_A be a diagonal matrix whose diagonal entries are given by $d_{ii} = 1/A_{ii}$, and D_B be a diagonal matrix with $d_{ii} = 1/B_{ii}$. Q_A and Q_B are defined as

$$Q_A = I - D_A A, \quad Q_B = I - D_B B. \tag{2.8.1}$$

Since A is an M-matrix, we have

$$\rho(Q_A) < 1.$$

With the assumption $B \geq A$, we have $Q_B \leq Q_A$ and both of them are nonnegative. So

$$\rho(Q_B) \leq \rho(Q_A) < 1.$$

From the above inequality and Lemma 2.7.2, we have B is an M-matrix.

(b) From Lemma 2.7.1, A is a nonsingular M-matrix. Then there exists a vector $u > 0$, such that $Au > 0$. Since $B \geq A$, $(B - A)u \geq 0$. Hence $Bu \geq Au > 0$, which means B is also a nonsingular M-matrix.

(c) Suppose A is irreducible nonsingular and $A = sI - N$ as in the definition of the M-matrix. According to Definition 2.5.2, for any (i, j), there exists an $m \in \mathbb{N}$, such that $(N^m)_{ij} \neq 0$. So

$$\left[I - \left(\frac{N}{s}\right)\right]^{-1} = I + \left(\frac{N}{s}\right) + \left(\frac{N}{s}\right)^2 + \left(\frac{N}{s}\right)^3 + \cdots > 0,$$

i.e. $\frac{A^{-1}}{s} > 0$, $A^{-1} > 0$.

(d) Without loss of generality, suppose A is irreducible (If not, split it into irreducible blocks and work on submatrices). Let $A = sI - N$, where $s = \max_i(A_{ii})$ and $N \geq 0$. Apply Theorem 3.1.1 to N since N is also irreducible. We have an eigenvalue of N, say λ_N, such that

$$\lambda_N = \rho(N).$$

Thus from

$$\lambda_{\min}(A) = s - \lambda_{\max}(N) = s - \rho(N) \geq 0,$$

$\lambda_{\min}(A) \leq s = \max_i(A_{ii})$.

(e) The proof of (e) can be found in [31]. □

CHAPTER 3

A Fundamental Theorem on M-matrix Algebraic Riccati Equation

A fundamental result on the M-matrix Algebraic Riccati Equation is as follows.

Theorem 3.0.1 ([19, 23])**.** *An MARE* (1.0.1) *has a unique (entrywise)* minimal *nonnegative solution $\varPhi$, i.e.,*

$$\varPhi \le X \quad \textit{for any other nonnegative solution } X \textit{ of an MARE } (1.0.1).$$

The proof of Theorem 3.0.1 is very complicated. So we split it into two parts. Theorem 3.2.1 is about the existence of nonnegative solutions. Theorem 3.3.1 says that there is a unique minimal nonnegative solution, assuming the existence of a nonnegative solution of MARE (1.0.1).

3.1 Perron-Frobenius Theorem

The following theorem presents some important properties of nonnegative irreducible matrices.

Theorem 3.1.1 (Perron-Frobenius [32, 38])**.** *Let $A \ge 0$ be an irreducible $n \times n$ matrix. Then*

1. *A has a positive real eigenvalue equal to its spectral radius.*
2. *To $\rho(A)$ there corresponds a positive eigenvector.*
3. *$\rho(A)$ increases when any entry of A increases.*
4. *$\rho(A)$ is a simple eigenvalue of A.*

The interested reader is referred to [38] for the proof of Theorem 3.1.1 as well as some applications..

Remark **3.1.1.** *Iterative methods are studied for the numerical solution of an MARE* (1.0.1)*. The condition*

$$C \geq 0, \quad D \geq 0, \quad I \otimes A + B^T \otimes I \text{ is a nonsingular } M\text{-matrix,} \tag{3.1.1}$$

given in [23], where $\otimes$ is the Kronecker product, could be easily derived with Lemma 2.8.1 (e). It is shown in [23] that Newton's method and a class of basic fixed-point iterations can find its minimal nonnegative solution whenever it has a nonnegative solution. This conclusion is useful for the proofs in the following two sections.

3.2 The Existence of Nonnegative Solutions

As a part of proving Theorem 3.0.1, we will first prove the existence of nonnegative solutions.

Theorem 3.2.1 ([9, 38, 19, 23])**.** *If W in* (1.0.2) *is a nonsingular M-matrix, then the MARE* (1.0.1) *has a nonnegative solution X such that $B - DX$ is also a nonsingular M-matrix. If W in* (1.0.2) *is an irreducible singular M-matrix, then the MARE* (1.0.1) *has a nonnegative solution X such that $B - DX$ is also an M-matrix.*

Proof. Consider T=diag(B,A) $\geq W$ and Lemma 2.7.1. T is an M-matrix since W is an M-matrix. If W is nonsingular, then there exists a positive vector $u > 0$, such that $Wu > 0$. Thus

$$(T - W)u \geq 0 \Rightarrow Tu \geq Wu > 0 \Rightarrow Tu > 0.$$

So T is also a nonsingular M-matrix.

If W is an irreducible singular M-matrix, then with the definition of the M-matrix, there is a number s and two nonnegative matrices N_1 and N_2 such that $T = sI - N_1$ and $W = sI - N_2$. By Theorem 3.1.1,

$$\begin{aligned} T > W &\Rightarrow sI - N_1 > sI - N_2 \\ &\Rightarrow N_1 < N_2 \\ &\Rightarrow \rho(N_1) < \rho(N_2) \\ &\Rightarrow s - \rho(N_1) > s - \rho(N_2) \\ &\Rightarrow \rho(T) > \rho(W) \geq 0. \end{aligned}$$

Hence T is a nonsingular M-matrix. From, $\det(A)\cdot \det(B) = \det(T)$, A and B are nonsingular. Moreover we have $T^{-1} = \text{diag}(B^{-1}, A^{-1}) \geq 0$ from Lemma 2.7.1. Thus A^{-1} and B^{-1} are both nonnegative, which means A and B are both nonsingular M-matrices. So condition (3.1.1) is satisfied.

Next, take $X_0 = 0$ and use the Fixed-point Iteration:

$$A_1X_{i+1} + X_{i+1}B_1 = X_iDX_i + A_2X_i + X_2B_2 + C, \quad i = 1, 2, \cdots, \tag{3.2.1}$$

where $A = A_1 - A_2$, $B = B_1 - B_2$ that makes A_1 and B_1 both Z-matrices along with A_2, $B_2 \geq 0$. We need to prove that the sequence $\{X_i\}$ is bounded above and satisfies $0 \leq X_i \leq X_{i+1}$. If W is a nonsingular M-matrix, we can find v_1, v_2, such that

$$B_1v_1 - B_2v_1 - Dv_2 = u_1 > 0, \tag{3.2.2}$$

$$A_1v_2 - A_2v_2 - Cv_1 = u_2 > 0. \tag{3.2.3}$$

By induction, it is easy to get $0 \leq X_i \leq X_{i+1}$. For $k = 0$, $v_2 - A_1^{-1}u_2 = A_1^{-1}(A_2v_2 + Cv_1) \geq 0$ by (3.2.3). Now suppose $X_iv_1 \leq v_2 - A_1^{-1}u_2$ for some $i \geq 1$. Then

$$\begin{aligned}
A_1X_{i+1}v_1 + X_{i+1}B_1v_1 &= X_iDX_iv_1 + A_2X_iv_1 + X_iB_2v_1 + CV_1 \\
&\leq X_iDV_2 + A_2V_2 + X_iB_2V_1 + CV_1 \\
&= X_iB_1v_1 - X_iB_2v_1 - X_iu_1 + A_1v_2 - u_2 + X_IB_2v_1 \\
&= X_iB_1v_1 - X_iu_1 + A_1v_2 - u_2 \\
&\leq X_iB_1v_1 + A_1v_2 - u_2.
\end{aligned}$$

Since $X_{i+1} \geq X_i$, from above $A_1X_{i+1}v_2 \leq A_1v_2 - u_2$. So $\{X_i\}$ is increasing and bounded above. Hence it has a limit, say S, which is a nonnegative solution of MARE (1.0.1) and satisfies $Sv_1 \leq v_2 - A_1^{-1}u_2 < v_2$. Thus $(B - DS)v_1 \geq Bv_1 - Dv_2 = u_1 > 0$. Therefore $B - DS$ is a nonsingular M-matrix by Lemma 2.7.1. The proof for the irreducible singular case is almost the same. The only difference is at the initial step, where we use Theorem 3.1.1 to ensure that there are v_1, $v_2 > 0$ to make

$$\begin{aligned}
B_1v_1 - B_2v_1 - Dv_2 = 0, \\
A_1v_2 - A_2v_2 - Cv_1 = 0.
\end{aligned}$$

□

3.3 The Unique Minimal Nonnegative Solution

For the next lemma, we need to add the following assumption besides (3.1.1)

$$C, D \neq 0, \quad (I \otimes A + B^T \otimes I)^{-1}\text{vec}(C) > 0. \tag{3.3.1}$$

Lemma 3.3.1 ([23, 19, 38]). *Consider an MARE* (1.0.1) *with conditions* (3.1.1) *and* (3.3.1). *If there exists a positive matrix X such that $\mathscr{R}(X) \leq 0$, then the MARE* (1.0.1) *has a positive solution S satisfies $S \leq X$ for every positive matrix X for* *which $\mathscr{R}(X) \leq 0$. Furthermore, S is also the minimal positive solution of the MARE* (1.0.1).

Proof. Let X be an arbitrary positive matrix such that

$$\mathscr{R}(X) = XDX - AX - XB + C \leq 0. \tag{3.3.2}$$

We use induction to prove

$$X_k < X_{k+1}, \quad X_k < X, \quad I \otimes (A - X_k D) + (B - DX_k)^T \otimes I \text{ is an } M\text{-matrix.} \tag{3.3.3}$$

For $k = 0$, the Newton's iteration (2.2.6) is

$$AX_1 + X_1 B = C,$$

which is equivalent to

$$(I \otimes A + B^T \otimes I)\text{vec}(X_1) = \text{vec}(C).$$

In the above equation, $I \otimes A + B^T \otimes I$ is an M-matrix as assumed in (3.1.1). Thus $(I \otimes A + B^T \otimes I)^{-1} > 0$ by Lemma 2.7.1. With $C > 0$, we have $X_1 > 0$. Hence (3.3.3) is true for $k = 0$.

Now assuming that (3.3.3) is true for $k = i \geq 0$, by (2.2.6) and (3.3.2) we have

$$\begin{aligned} & (A - X_i D)(X_{i+1} - X) + (X_{i+1} - X)(B - DX_i) \\ = & \; C - X_i D X_i - AX + X_i DX - XB + XDX_i \\ \leq & - XDX - X_i D_X i + X_i DX + XDX_i \\ = & - (X - X_i)D(X - X_i). \end{aligned}$$

Since $I \otimes (A - X_i D) + (B - DX_i)^T \otimes I$ is an M-matrix and $D > 0$ by assumption, we can easily get $X_{i+1} - X < 0$, i.e. $X_{i+1} < X$.

Consider

$$\begin{aligned}
&(A - X_{i+1}D)X_{i+1} + X_{i+1}(B - DX_{i+1}) \\
&= (A - X_iD - (X_{i+1} - X_i)D)X_{i+1} + X_{i+1}[B - DX_i - D(X_{i+1} - X_i)] \\
&= (A - X_iD)X_{i+1} + X_{i+1}(B - DX_i) - (X_{i+1} - X_i)DX_{i+1} - X_{i+1}D(X_{i+1} - X_i) \\
&= C - (X_{i+1} - X_i)D(X_{i+1} - X_i) - X_{i+1}DX_{i+1},
\end{aligned}$$

i.e.

$$(A - X_{i+1}D)X_{i+1} + X_{i+1}(B - DX_{i+1}) = C - (X_{i+1} - X_i)D(X_{i+1} - X_i) - X_{i+1}DX_{i+1}. \tag{3.3.4}$$

Using the above equation, we have

$$\begin{aligned}
&(A - X_{i+1}D)(X_{i+1} - X) + (X_{i+1} - X)(B - DX_{i+1}) \\
=& - (X_{i+1} - X_i)D(X_{i+1} - X_i) - (X_{i+1} - X)D(X_{i+1} - X) < 0,
\end{aligned}$$

which means $(I \otimes (A - X_{i+1}D) + (B - DX_{i+1})^T \otimes I)\text{vec}(X - X_{i+1}) > 0$. From Lemma 2.7.1, $I \otimes (A - X_{i+1}D) + (B - DX_{i+1})^T \otimes I$ is an M-matrix.

Using (3.3.4) again, we get

$$\begin{aligned}
&(A - X_{i+1}D)(X_{i+1} - X_{i+2}) + (X_{i+1} - X_{i+2})(B - DX_{i+1}) \\
=& - (X_{i+1} - X_i)D(X_{i+1} - X_i) - (X_{i+1} - X)D(X_{i+1} - X) < 0,
\end{aligned}$$

by which we have $X_{i+1} < X_{i+2}$[1]. Thus, the proof by induction for (3.3.3) is done.

From (3.3.3), the sequence $\{X_i\}$ generated by the Newton's iteration (2.2.6) is monotonically increasing and bounded above. Thus the limit exists. If we write $S = \lim_{k\to\infty} X_k$, then S is a solution of $\mathscr{R}(X) = 0$. Moreover, since $S \leq X$ for

[1] Here we used $(I \otimes (A - X_{i+1}D) + (B - DX_{i+1})^T \otimes I)\text{vec}(X - X_{i+1}) > 0$ is an M-matrix as mentioned before.

any X satisfies $\mathscr{R}(X) \leq 0$, S is also the minimal nonnegative solution of the MARE (1.0.1). $\square$

Now with (3.2.1), we introduce an operator $\mathscr{L}$ defined as

$$\mathscr{L}(X) = A_1 X + X B_1. \tag{3.3.5}$$

Lemma 3.3.2 ([32, 38, 16]). *If $\mathscr{L}$ is defined as* (3.3.5)*, then*

(a) *$\mathscr{L}$ is a linear operator.*

(b) *If $I \otimes A_1 + B_1^T \otimes I$ is a nonsingular M-matrix, the operator $\mathscr{L}$ is invertible and $\mathscr{L}^{-1}(X) \geq 0$ for $X \geq 0$.*

Proof. Part (a) is trivial. For part (b), referring to Lemma 2.4.1 part (c), we have

$$\mathrm{vec}(A_1 X + X B_1) = (I_n \otimes A_1 + B_1^T \otimes I_m)\mathrm{vec}(X).$$

Thus, with the assumption that $I \otimes A_1 + B_1^T \otimes I$ is a nonsingular M-matrix and Lemma 2.7.1, $\mathscr{L}^{-1} \geq 0$. So $\mathscr{L}^{-1}(X) \geq 0$ for $X \geq 0$. $\square$

Consider an MARE (1.0.1) with condition (3.1.1), we have

Theorem 3.3.1 ([19, 23]). *For fixed-point iteration* (3.2.1) *and $X_0 = 0$, we have $X_k \leq X_{k+1}$ for any $k \geq 0$. If $\mathscr{R}(X) \leq 0$ for some nonnegative matrix X, then we also have $X_k \leq X$ for any $k \geq 0$. Moreover, $\{X_k\}$ converges to the minimal nonnegative solution of the MARE* (1.0.1)*.*

Proof. There are three steps for this proof.

(a) $X_k \leq X_{k+1}$ for any $k \geq 0$.

(b) If there exists $X \geq 0$ such that $\mathscr{R}(X) \leq 0$, then $X_k \leq X$ for any $k \geq 0$.

(c) $\{X_k\}$ converges to the minimal nonnegative solution of the MARE (1.0.1).

(a) can be easily proved with induction. For (b), we also apply induction. $X_0 \leq X$ is true. Suppose $X_k \leq X$, then

$$\begin{aligned}
X_{k+1} &= \mathscr{L}^{-1}(X_k D X_k + A_2 X_k + X_k + B_2 + C) \\
&= \mathscr{L}^{-1}(X_k D X_k - A X_k - X_k B + C + A_1 X_k + X_k B_1) \\
&= \mathscr{L}^{-1}\mathscr{R}(X_k) + \mathscr{L}^{-1}\mathscr{L}(X_k) \\
&= X_k - \mathscr{L}^{-1}[-\mathscr{R}(X_k)] \\
&\leq X_k \leq X.
\end{aligned}$$

The last but one inequality comes from Lemma 3.3.2.

Thus if $X^* = \lim_{k\to\infty} X_k$, then from Lemma 3.3.1 X^* is the minimal nonnegative solution of the MARE (1.0.1), which proved (c). □

Combining Theorem 3.2.1 with Theorem 3.3.1, we have proved Theorem 3.0.1.

CHAPTER 4

ADI Method

Bai, Guo, and Xu [4] proposed an alternating-directional-implicit[1] iteration method for the MARE (1.0.1):

$$X_{k+\frac{1}{2}}[\alpha I + (B - DX_k)] = (\alpha I - A)X_k + C, \tag{4.0.1a}$$

$$[\alpha I + (A - X_{k+\frac{1}{2}}D)]X_{k+1} = X_{k+\frac{1}{2}}(\alpha I - B) + C. \tag{4.0.1b}$$

They proved that with $X_0 = 0$,

$$0 \leq X_k \leq X_{k+\frac{1}{2}} \leq X_{k+1} \leq \varPhi, \quad \lim_{k\to\infty} X_k = \varPhi, \tag{4.0.2}$$

provided

$$\alpha \geq \max_{i,j}\{A_{(i,i)}, B_{(j,j)}\}. \tag{4.0.3}$$

While this theory reads beautifully, it does not tell what α value should be for the fastest convergence rate subjected to (4.0.3). Recently, Wang and Guo [40] essentially showed that under the constraint (4.0.3), $\alpha = \max_{i,j}\{A_{(i,i)}, B_{(j,j)}\}$ makes X_k converge to $\varPhi$ the fastest.

The method (4.0.1) is very much reminiscent of the well-known ADI (Alternating-Directional-Implicit) iteration for Sylvester equations [8, 39], except that it uses only one parameter α. Inspired by the effort in [39] on ADI parameter selections, it is conceivable that the method could be improved with more parameters. Added to the complexity here is the question how to still retain the nonnegativity of X_k and its monotonic convergence to $\varPhi$. So we are about to present a much improved version of the method (4.0.1) in terms of the speed of X_k approaching $\varPhi$.

[1] It is also called alternating-linearized-implicit(ALI) method.

4.1 ADI Method for M-Matrix Algebraic Riccati Equation

Consider an MARE (1.0.1). We reformulate it as the following two fixed-point equations:

$$X[\alpha I + (B - DX)] = (\alpha I - A)X + C, \tag{4.1.1}$$

$$[\beta I + (A - XD)]X = X(\beta I - B) + C, \tag{4.1.2}$$

where α and β are given positive parameters. In fact, this reformulation also provides a new technique for linearizing the nonlinear MARE (1.0.1). Now given X_k , by first solving $X_{k+\frac{1}{2}}$ from

$$X_{k+\frac{1}{2}}[\alpha I + (B - DX_k)] = (\alpha I - A)X_k + C, \tag{4.1.3}$$

and then solving X_{k+1} from

$$[\beta I + (A - X_{k+\frac{1}{2}}D)]X_{k+1} = X_{k+\frac{1}{2}}(\beta I - B) + C, \tag{4.1.4}$$

we can establish the following alternating directional implicit iteration method to solve the MARE (1.0.1).

Algorithm 4.1.1 (The new ADI iteration method)**.** *Set $X_0 = 0 \in \mathbb{R}^{n\times m}$. For $k = 0, 1, 2, \cdots$, compute X_{k+1} from X_k by solving the following two systems of linear matrix equations:*

$$X_{k+\frac{1}{2}}[\alpha I + (B - DX_k)] = (\alpha I - A)X_k + C, \tag{4.1.5a}$$

$$[\beta I + (A - X_{k+\frac{1}{2}}D)]X_{k+1} = X_{k+\frac{1}{2}}(\beta I - B) + C, \tag{4.1.5b}$$

where $\alpha > 0$ and $\beta > 0$ are given iteration parameters.

4.2 Convergence Theory

As a preparation, we first show several properties about the ADI iteration method, which are essential for us to establish its monotonic convergence theorem.

Lemma 4.2.1. *Let the matrix sequence $\{X_k\}$ be generated by Algorithm 4.1.1, and Φ be the minimal nonnegative solution of MARE* (1.0.1). *Then the following equalities hold true:*

(a) $(X_{k+\frac{1}{2}} - \Phi)[\alpha I + (B - DX_k)] = [\alpha I - (A - \Phi D)](X_k - \Phi)$;

(b) $(X_{k+\frac{1}{2}} - X_k)[\alpha I + (B - DX_k)] = \mathscr{R}(X_k)$;

(c) $\mathscr{R}(X_{k+\frac{1}{2}}) = [\alpha I - (A - X_{k+\frac{1}{2}}D)](X_{k+\frac{1}{2}} - X_k)$;

(d) $[\beta I + (A - X_{k+\frac{1}{2}}D)](X_{k+1} - \Phi) = (X_{k+\frac{1}{2}} - \Phi)[\beta I - (B - D\Phi)]$;

(e) $[\beta I + (A - X_{k+\frac{1}{2}}D)](X_{k+1} - X_{k+\frac{1}{2}}) = \mathscr{R}(X_{k+\frac{1}{2}})$;

(f) $\mathscr{R}(X_{k+1}) = (X_{k+1} - X_{k+\frac{1}{2}})[\beta I - (B - DX_{k+1})]$.

In Lemma 4.2.1, items (a), (b) and (c) are the same as Lemma 4.1 in [4]. Items (d), (e) and (f) are only different from Lemma 4.1 in [4] in the parameters.

4.3 Analysis

Based on Lemma 4.2.1 and Lemma 2.8.1, we are now ready to prove the monotonic convergence for the ADI iteration sequences.

Theorem 4.3.1. *Let Φ be the minimal nonnegative solution of the MARE* (1.0.1). *Let $X_0 = 0$ be the initial matrix, and α, β be the parameters such that*

$$\alpha \geq \max_{1\leq i\leq n} a_{ii}, \quad \beta \geq \max_{1\leq i\leq m} b_{ii},$$

where a_{ii} and b_{ii} are the ith diagonal elements of matrices A and B respectively. Then the matrix sequence $\{X_k\}$ generated by Algorithm 4.1.1 is well defined, and it holds that

(a) $\{X_k\}$ *is monotonically increasing and bounded, i.e.* $0 = X_0 \leq X_{\frac{1}{2}} \leq X_1 \leq \cdots \leq X_k \leq X_{k+\frac{1}{2}} \leq X_{k+1} \leq \cdots \leq \Phi$;

(b) $\{X_k\}$ *converges to Φ, i.e.* $\lim_{k\to\infty} X_k = \Phi$.

Proof. The proof is almost the same as the proof of Theorem 4.1 in [4]. The difference is in the iteration parameters. Because W is an M-matrix, its diagonal blocks A and B are M-matrices, too. Hence, when

$$\alpha \geq \max_{1 \leq i \leq n} a_{ii}, \quad \beta \geq \max_{1 \leq i \leq m} b_{ii},$$

the matrices $\alpha I - A$ and $\beta I - B$ are both nonnegative matrices. For the matrix sequence X_k generated by Algorithm 4.1.1, we can assert that the following facts hold true

(F1) $\{X_{k+\frac{1}{2}}\}$ and $\{X_{k+1}\}$ are bounded, i.e.

$$0 \leq X_{k+\frac{1}{2}} \leq \varPhi \;\text{ and }\; 0 \leq X_{k+1} \leq \varPhi, \quad k = 0, 1, 2, \cdots .$$

(F2) $A - X_{k+\frac{1}{2}} D$ and $B - DX_{k+1}$ are M-matrices, $\quad k = 0, 1, 2, \cdots$.

Facts (F1) and (F2) can be proved by induction. In fact, by substituting $X_0 = 0$ into (4.1.5) we get

$$X_{\frac{1}{2}}(\alpha I + B) = C.$$

As B is an M-matrix, by Lemma 2.8.1 the matrix $\alpha I + B$ is also an M-matrix. Hence,

$$X_{\frac{1}{2}} = C(\alpha I + B)^{-1} \geq 0.$$

In addition, by Lemma 4.2.1, we get

$$X_{\frac{1}{2}} - \varPhi = -[(\alpha I - A) + \varPhi D]\varPhi(\alpha I + B)^{-1} \leq 0.$$

This shows that $0 \leq X_{\frac{1}{2}} \leq \varPhi$. From $D \geq 0$, we have

$$A - \varPhi D \leq A - X_{\frac{1}{2}} D \leq A.$$

By Lemma 2.8.1, we know that $A - X_{\frac{1}{2}} D$ is an M-matrix. Analogously, by making use of Lemma 2.8.1, we know that $\beta I + (A - X_{\frac{1}{2}} D)$ is an M-matrix. From (4.1.5) and Lemma 4.2.1, we get

$$X_1 = [\beta I + (A - X_{\frac{1}{2}} D)]^{-1}(X_{\frac{1}{2}}(\beta I - B) + C) \geq 0,$$

and

$$X_1 - \mathit{\Phi} = [\beta I + (A - X_{\frac{1}{2}} D)]^{-1} (X_{\frac{1}{2}} - \mathit{\Phi})[(\beta I - B) + D\mathit{\Phi}] \leq 0.$$

This shows that $0 \leq X_1 \leq \mathit{\Phi}$. From $D \geq 0$, we get

$$B - D\mathit{\Phi} \leq B - DX_1 \leq B.$$

Thus we know that $B - DX_1$ is an M-matrix. The above proof shows that (F1) and (F2) are true for $k = 0$. Now assume that (F1) and (F2) are true for $k = l - 1$. Then from (4.1.5) and Lemma 4.2.1 (a) we get

$$X_{l+\frac{1}{2}} = [(\alpha I - A)X_l + C][\alpha I + (B - DX_l)]^{-1},$$

and

$$X_{l+\frac{1}{2}} - \mathit{\Phi} = [\alpha - (A - \mathit{\Phi} D)](X_l - \mathit{\Phi})[\alpha + (B - DX)]^{-1}.$$

Because $0 \leq X_l \leq \mathit{\Phi}$, $\alpha I - A \geq 0$ and $C \geq 0$, it holds that

$$(\alpha I - A)X_l + C \geq 0.$$

In addition, as $B - DX_l$ is an M-matrix, we know that $\alpha I + (B - DX_l)$ is also an M-matrix. Therefore, we have

$$X_{l+\frac{1}{2}} \geq 0 \quad \text{and} \quad X_{l+\frac{1}{2}} - \mathit{\Phi} \leq 0.$$

That is to say,

$$0 \leq X_{l+\frac{1}{2}} \leq \mathit{\Phi}.$$

It then follows from $D \geq 0$ that

$$A - \mathit{\Phi} D \leq A - X_{l+\frac{1}{2}} D \leq A.$$

Thus $A - X_{l+\frac{1}{2}} D$ is an M-matrix.

Similarly, from (4.1.5) and Lemma 4.2.1, we have

$$X_{l+1} = [\beta I + (A - X_{l+\frac{1}{2}}D)]^{-1}[X_{l+\frac{1}{2}}(\beta I - B) + C]$$

and

$$X_{l+1} - \Phi = [\beta I + (A - X_{l+\frac{1}{2}}D)]^{-1}(X_{l+\frac{1}{2}} - \Phi)[\beta I - (B - D\Phi)].$$

Because $0 \leq X_{l+\frac{1}{2}} \leq \Phi$, $\alpha I - B \geq 0$ and $C \geq 0$, it holds that

$$X_{l+\frac{1}{2}}(\beta I - B) + C \geq 0.$$

In addition, as $A - X_{l+\frac{1}{2}}D$ is an M-matrix, we know that $\beta I + (A - X_{l+1}D)$ is also an M-matrix. Therefore, we have

$$X_{l+1} \geq 0 \;\text{ and }\; X_{l+1} - \Phi \leq 0.$$

That is to say,

$$0 \leq X_{l+1} \leq \Phi.$$

It then follows from $D \geq 0$ that

$$B - D\Phi \leq B - DX_{l+1} \leq B.$$

Thus $B - DX_{l+1}$ is an M-matrix.

The above proof shows that (F1) and (F2) are true for $k = l$.

Hence, by induction, we have proved that facts (F1) and (F2) hold for all integers $k \geq 0$. Now, we begin to prove conclusions (a) and (b) of this theorem.

We first prove (a). What needs to be done is to inductively prove that the inequalities

$$X_k \leq X_{k+\frac{1}{2}} \leq X_{k+1},\ \mathscr{R}(X_k) \geq 0,\ \mathscr{R}(X_{k+\frac{1}{2}}) \geq 0,\ \text{and}\ \mathscr{R}(X_{k+1}) \geq 0 \qquad (4.3.1)$$

hold for all nonnegative integers k.

In fact, when $k = 0$, by substituting $X_0 = 0$ into (b), (c), (e) and (f) in Lemma 4.2.1, and making use of the facts (F1) and (F2), we can easily verify that the inequalities in (4.3.1) are all true. Assume that the inequalities in (4.3.1) hold for $k = l - 1$. Then by making use of the facts (F1) and (F2), from (b), (c), (e) and (f) in Lemma 4.2.1 again we can easily verify that the inequalities in (4.3.1) are also true for $k = l$. Hence, by induction, we have proved that (4.3.1) holds for all nonnegative integers k.

It follows straightforwardly from (4.3.1) and (F1) that (a) is true.

Because $\{X_k\}$ is nonnegative, monotonically increasing, and bounded from above, there exists a nonnegative matrix Φ^*, such that $\lim_{k\to\infty} X_k = \Phi^*$. Evidently, it also holds that $\lim_{k\to\infty} X_{k+\frac{1}{2}} = \Phi^*$. Obviously, (F1) implies $\Phi^* \leq \Phi$. On the other hand, by taking limits in both (4.1.5) and (4.3.1), we see that Φ^* is also a nonnegative solution of the MARE (1.0.1). Hence, it must hold that $\Phi \leq \Phi^*$ due to the minimal property of Φ. It then follows that $\Phi^* = \Phi$, and (b) is true. □

The following theorem shows that the iteration sequence is nonincreasing with respect to α and β, respectively.

Theorem 4.3.2. *Suppose that W in* (1.0.2) *is an M-matrix, and Φ is the minimal nonnegative solution of the MARE* (1.0.1). *Let $X_0 = \widetilde{X}_0 = 0$ be the initial matrices, and the matrix sequences $\{X_k\}$, $\{\widetilde{X}_k\}$ be generated by Algorithm 4.1.1 for which the corresponding iteration parameters are $(\alpha,\ \beta)$ and $(\alpha_1,\ \beta_1)$ respectively, satisfying*

$$\alpha_1 \geq \alpha \geq \max_{1\leq i\leq n} a_{ii}, \quad \beta_1 \geq \beta \geq \max_{1\leq i\leq m} b_{ii},$$

where a_{ii} and b_{ii} are the ith diagonal elements of the matrices A and B, respectively. Then

$$X_{k+\frac{1}{2}} \geq \widetilde{X}_{k+\frac{1}{2}}, \quad X_{k+1} \geq \widetilde{X}_{k+1}, \quad k = 0, 1, 2, \ldots. \tag{4.3.2}$$

Proof. The proof is almost the same as the proof of Theorem 3.1 in [40]. The difference is in the iteration parameters. It can be proved by induction. In fact, when $k = 0$,

$$X_0 = \widetilde{X}_0 = 0,\ X_{\frac{1}{2}} = C(\alpha I + B)^{-1},\ \widetilde{X}_{\frac{1}{2}} = C(\alpha_1 I + B)^{-1}.$$

As B is an M-matrix, by Lemma 2.8.1 the matrix $\alpha I + B$ and $\alpha_1 I + B$ are also M-matrices, and $\alpha I + B \leq \alpha_1 I + B$. By Lemma 2.8.1, we know

$$(\alpha I + B)^{-1} \geq (\alpha_1 I + B)^{-1}.$$

Therefore, $X_{\frac{1}{2}} \geq \widetilde{X}_{\frac{1}{2}}$.

From Lemma 4.2.1 (e), we have

$$X_{k+1} = [\beta I + (A - X_{k+\frac{1}{2}} D)]^{-1} \mathscr{R}(X_{k+\frac{1}{2}}) + X_{k+\frac{1}{2}}.$$

Therefore,

$$\begin{aligned} X_1 - \widetilde{X}_1 &= [\beta I + (A - X_{\frac{1}{2}} D)]^{-1} \mathscr{R}(X_{\frac{1}{2}}) + X_{\frac{1}{2}} \\ &\quad -[\beta_1 I + (A - \widetilde{X}_{\frac{1}{2}} D)]^{-1} \mathscr{R}(\widetilde{X}_{\frac{1}{2}}) - \widetilde{X}_{\frac{1}{2}} \\ &\geq [\beta I + (A - X_{\frac{1}{2}} D)]^{-1} [\mathscr{R}(X_{\frac{1}{2}}) - \mathscr{R}(\widetilde{X}_{\frac{1}{2}})] + X_{\frac{1}{2}} - \widetilde{X}_{\frac{1}{2}}. \end{aligned}$$

Since

$$\mathscr{R}(X_{\frac{1}{2}}) = X_{\frac{1}{2}} D X_{\frac{1}{2}} - A X_{\frac{1}{2}} - X_{\frac{1}{2}} B + C,$$

by substitution, we have

$$\begin{aligned} X_1 - \widetilde{X}_1 &\geq [\beta I + (A - X_{\frac{1}{2}} D)]^{-1} [X_{\frac{1}{2}} D X_{\frac{1}{2}} - \widetilde{X}_{\frac{1}{2}} D \widetilde{X}_{\frac{1}{2}} \\ &\quad -(X_{\frac{1}{2}} - \widetilde{X}_{\frac{1}{2}}) B - A(X_{\frac{1}{2}} - \widetilde{X}_{\frac{1}{2}})] + X_{\frac{1}{2}} - \widetilde{X}_{\frac{1}{2}} \\ &= [\beta I + (A - X_{\frac{1}{2}} D)]^{-1} \{(X_{\frac{1}{2}} D - A)(X_{\frac{1}{2}} - \widetilde{X}_{\frac{1}{2}}) + (X_{\frac{1}{2}} - \widetilde{X}_{\frac{1}{2}})(D \widetilde{X}_{\frac{1}{2}} - B) \\ &\quad +[\beta I + (A - X_{\frac{1}{2}} D)](X_{\frac{1}{2}} - \widetilde{X}_{\frac{1}{2}})\} \\ &= [\beta I + (A - X_{\frac{1}{2}} D)]^{-1} (X_{\frac{1}{2}} - \widetilde{X}_{\frac{1}{2}})(\beta I + D \widetilde{X}_{\frac{1}{2}} - B). \end{aligned}$$

By Lemma 2.8.1, we know that $\beta I + (A - X_{\frac{1}{2}} D)$ is an M-matrix. Hence $[\beta I + (A - X_{\frac{1}{2}} D)]^{-1} \geq 0$. From $\beta \geq \max_{1 \leq i \leq m} b_{ii}$, we get $\beta I + D\widetilde{X}_{\frac{1}{2}} - B \geq 0$. Hence $X_1 \geq \widetilde{X}_1$. The above proof shows that (4.3.2) is true for $k = 0$.

Assume that the inequalities in (4.3.2) hold for $k = l - 1$. Then from Lemma 4.2.1 (b), we get

$$X_{l+\frac{1}{2}} = \mathscr{R}(X_l)[\alpha I + (B - DX_l)]^{-1} + X_l.$$

Therefore,

$$\begin{aligned} X_{l+\frac{1}{2}} - \widetilde{X}_{l+\frac{1}{2}} &= \mathscr{R}(X_l)[\alpha I + (B - DX_l)]^{-1} + X_l \\ &\quad - \mathscr{R}(\widetilde{X}_l)[\alpha_1 I + (B - D\widetilde{X}_l)]^{-1} - \widetilde{X}_l \\ &\geq (\alpha_1 I - A + X_l D)(X_l - \widetilde{X}_l)[\alpha_1 I + (B - D\widetilde{X}_l)]^{-1}. \end{aligned}$$

By Lemma 2.8.1, we know that $\alpha_1 I + (B - D\widetilde{X}_l)$ is an M-matrix. Hence $[\alpha_1 I + (B - D\widetilde{X}_l)]^{-1} \geq 0$. From $\alpha_1 \geq \max_{1 \leq i \leq n} a_{ii}$, we get $(\alpha_1 I - A + X_l D) \geq 0$. Hence $X_{l+\frac{1}{2}} \geq \widetilde{X}_{l+\frac{1}{2}}$.

The inequality $X_{l+1} \geq \widetilde{X}_{l+1}$ can be derived analogously.

The above proof shows that (4.3.2) is also true for $k = l$. Hence, by induction, we have proved that (4.3.2) holds for all positive integers k. □

From this chapter, we find a way to improve ADI method. Although it is still linearly convergent, not as good as doubling algorithm SDA([24]), ADDA([41]) to be discussed in the next chapter. However, it is the analysis of two parameters in ADI method that inspires us to create our ADDA method.

CHAPTER 5

Doubling Verses High-Order Algorithms

In this chapter, we will use the Sylvester equation as an example to compare doubling algorithm with high-order algorithms.

5.1 The Smith Method

Consider the numerical solutions of a matrix equation

$$XA + BX = C, \tag{5.1.1}$$

in which X is an unknown $m \times n$ matrix, A, B and C are known matrices of sizes $n \times n$, $m \times m$ and $m \times n$ respectively.

Suppose I_m is the $m \times m$ identity matrix and α is a nonzero scalar, then (5.1.1) can be written as

$$(\alpha I_m - B)X(\alpha I_n - A) - (\alpha I_m + B)X(\alpha I_n + A) = -2\alpha C.$$

Pre-multiply by $(\alpha I_m - B)^{-1}$ and post-multiply by $(\alpha I_n - A)^{-1}$ to get

$$X - EXF = W, \tag{5.1.2}$$

where

$$\begin{aligned} E &= (\alpha I_m - B)^{-1}(\alpha I_m + B), \\ F &= (\alpha I_n + A)(\alpha I_n - A)^{-1}, \\ W &= -2\alpha(\alpha I_m - B)^{-1}C(\alpha I_n - A)^{-1}. \end{aligned}$$

It is easy to see, from (5.1.2) and with initial $X_0 = 0$, we can apply (5.1.2) iteratively to generate a series for X

$$X = \sum_{i=1}^{\infty} E^{i-1} W F^{i-1}. \tag{5.1.3}$$

Barnett and Storey in [5] have suggested (5.1.3) as a method for the practical solution of (5.1.1) while the rate of convergence is slow in general. Now if $\{Y_i\}$ is the sequence of matrices defined iteratively by

$$Y_0 = W, \qquad Y_{i+1} = Y_i + E^{2^i} Y_i F^{2^i}, \tag{5.1.4}$$

then it follows by induction that

$$Y_k = \sum_{i=1}^{2^k} E^{i-1} W F^{i-1} \tag{5.1.5}$$

for all i. Smith in [36] showed that Y_i converges to X very rapidly as $i \to \infty$. So we refer to (5.1.3) the Smith method. Since it only calculates the 2^ith terms, we also call it the doubling algorithm.

5.2 The Tripling Algorithm

It is natural to consider higher order algorithms. First let us think about a tripling algorithm. Suppose we have E and F, satisfy

$$X_{k+1} = X_k + F^{3^k} X_k E^{3^k} + F^{2\cdot 3^k} X_k E^{2\cdot 3^k},$$

then

Lemma 5.2.1. *We have a tripling algorithm,*

$$X_k = \sum_{i=0}^{3^k - 1} F^i X_0 E^i. \tag{5.2.1}$$

Proof.

$$
\begin{aligned}
X_0 &= X_0. \\
X_1 &= X_0 + FX_0E + F^2X_0E^2. \\
X_2 &= X_1 + F^3X_1E^3 + F^6X_1E^6 \\
&= (X_0 + FX_0E + F^2X_0E^2) + (F^3X_0E^3 + F^4X_0E^4 + F^5X_0E^5) \\
&+ (F^6X_0E^6 + F^7X_0E^7 + F^8X_0E^8) \\
&= \sum_{i=0}^{3^2-1} F^iX_0E^i.
\end{aligned}
$$

Assuming (5.2.1) is right for X_k, consider X_{k+1}

$$
\begin{aligned}
X_{k+1} &= X_k + F^{3^k}X_kE^{3^k} + F^{2\cdot 3^k}X_kE^{2\cdot 3^k} \\
&= \sum_{i=0}^{3^k-1} F^iX_0E^i + \sum_{i=0}^{3^k-1} F^{i+3^k}X_0E^{i+3^k} + \sum_{i=0}^{3^k-1} F^{i+2\cdot 3^k}X_0E^{i+2\cdot 3^k} \\
&= \sum_{i=0}^{3^k-1} F^iX_0E^i + \sum_{i=3^k}^{2\cdot 3^k-1} F^iX_0E^i + \sum_{i=2\cdot 3^k}^{3^{k+1}-1} F^iX_0E^i \\
&= \sum_{i=0}^{3^{k+1}-1} F^iX_0E^i.
\end{aligned}
$$

Thus (5.2.1) is true for $k+1$. The induction is completed. □

Next we calculate flops[1] of the doubling and tripling algorithms:

(Doubling) $X_{k+1} = X_k + F^{2^k}X_kE^{2^k}$.

(Tripling) $X_{k+1} = X_k + F^{3^k}X_kE^{3^k} + F^{2\cdot 3^k}X_kE^{2\cdot 3^k}$.

Note here F is of size $m \times m$, and E is $n \times n$.

[1] Count every + and × as 1 flop.

For the doubling algorithm,

$$\begin{aligned} \text{calculation}:&\quad \text{flops} \\ (F^{2^{k-1}})^2 \to F^{2^k} &: m^2(2m-1) \\ (E^{2^{k-1}})^2 \to E^{2^k} &: n^2(2n-1) \\ F^{2^k} X_k E^{2^k} &: mn(2m-1) + mn(2n-1) \\ + &: mn \end{aligned}$$

For tripling algorithm,

$$\begin{aligned} \text{calculation}:&\quad \text{flops} \\ (F^{2\cdot 3^{k-1}}) \cdot (F^{3^{k-1}}) \to F^{3^k} &: m^2(2m-1) \\ F^{3^k} \cdot F3^k \to F^{2\cdot 3^k} &: m^2(2m-1) \\ (E^{2\cdot 3^{k-1}}) \cdot (E^{3^{k-1}}) \to E^{3^k} &: n^2(2n-1) \\ E^{3^k} \cdot E3^k \to E^{2\cdot 3^k} &: n^2(2n-1) \\ F^{3^k} X_k E^{3^k}, F^{2\cdot 3^k} X_k E^{2\cdot 3^k} &: 2mn(2m+2n-2) \\ + &: 2mn \end{aligned}$$

Thus $Flop(Tripling) = 2 \cdot Flop(Doubling)$. Here function $Flop(\cdot)$ means the number of flops.

Analysis:

If we calculate k_0 steps with the tripling algorithm, we can get to 3^{k_0}th entry.

But if we use doubling algorithm, with the same number of flops, we can get to $2^{2\cdot k_0} = 4^{k_0}$th entry, which means the doubling algorithm is faster than the tripling since it goes further.

5.3 High Order Algorithms

For those algorithms with order higher than 3, we can compare these algorithms with doubling algorithm(the Smith method). We generalize higher order algorithms as follows

Lemma 5.3.1. *High-order algorithm:*

$$\begin{aligned} X_{k+1} &= X_k + F^{l^k} X_0 E^{l^k} + F^{2\cdot l^k} X_0 E^{2\cdot l^k} + \cdots + F^{(l-1)l^k} X_0 E^{(l-1)l^k} \\ &= \sum_{i=0}^{l-1} F^{i\cdot l^k} X_0 E^{i\cdot l^k} \end{aligned}$$

Proof. The proof is easily achieved by induction in the same way as for that for the tripling. □

When we count for flops, for example, the term F^{l^k}, we compute it as

$$F^{l^k} = F^{l^{k-1}\cdot l} = F^{(l-1)\cdot l^{k-1}} \cdot F^{l^{k-1}},$$

where both of the last two entries can be found as the results in the last iteration. As counted in Section 5.2, $Flop(high - order) = (l-1) \cdot Flop(Doubling)$.

Analysis:

If we calculate k_0 steps with lth order algorithm, we can get to l^{k_0}th term. But if we use doubling, with the same number of flops, we can get to $2^{(l-1)k_0}$th. And

$$\frac{l^{k_0}}{2^{(l-1)k_0}} = (\frac{l}{2^{l-1}})^{k_0} < 1, \quad \text{for} \;\; l \geq 3.$$

Here we can see higher order algorithms are slower than the doubling since doubling gets further. Thus we can conclude that *Doubling algorithm is the best of all*, which makes it unnecessary for us to seek tripling and high order algorithms.

CHAPTER 6

ADDA: Alternating Directional Doubling Algorithm

The basic idea of the doubling algorithm for an iterative scheme is to compute only the 2^kth approximations, instead of every approximation in the process as we mentioned in chapter 5. The idea 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 [13] 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 [24]. For an MARE (1.0.1), SDA simultaneously computes the minimal nonnegative solutions of an MARE (1.0.1) and its *complementary M-Matrix Algebraic Riccati Equation* (cMARE)

$$YCY - YA - BY + D = 0. \tag{6.0.1}$$

In this chapter, we shall present our ADDA for the 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 chapter 4). Interested readers can refer to Appendix A for the development of ADDA applied to M-matrix Sylvester equation which leads to an improvement of the Smith method. Some numerical examples will be given in section 8.1.

These doubling algorithms are very fast and efficient as they are globally and quadratically convergent, except for the so-called critical case [13], which is to be studied in detail in Chapter 7. 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 [9, 17]

$$W \begin{pmatrix} x \\ y \end{pmatrix} = 0, \quad \begin{pmatrix} u \\ v \end{pmatrix}^{\mathrm{T}} W = 0. \tag{6.0.2}$$

We call an MARE (1.0.1) is in the *critical case* if $u^{\mathrm{T}}x = v^{\mathrm{T}}y$. For the critical case, the doubling algorithms converge linearly [13], and thus are slow compared to the non-critical case. An improved method will be given in chapter 7. Define

$$H \stackrel{\text{def}}{=} \begin{pmatrix} I_m & \\ & -I_n \end{pmatrix} W = \begin{pmatrix} B & -D \\ C & -A \end{pmatrix}. \tag{6.0.3}$$

H is singular if and only if W is singular, and (6.0.2) implies

$$H \begin{pmatrix} x \\ y \end{pmatrix} = 0, \quad \begin{pmatrix} u \\ -v \end{pmatrix}^{\mathrm{T}} H = 0. \tag{6.0.4}$$

6.1 A Fundamental Theorem

Before we start introducing ADDA, let us look at a theorem, which may have independent interest of its own and 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 *generalized Cayley transformation*

$$\mathscr{C}(A; \alpha, \beta) \stackrel{\text{def}}{=} (A - \alpha I)(A + \beta I)^{-1} \tag{6.1.1}$$

of a square matrix A, where α, β are scalars such that $A + \beta I$ is nonsingular. Given square matrices A and B, define

$$f(\alpha, \beta) \stackrel{\text{def}}{=} \rho(\mathscr{C}(A; \alpha, \beta)) \cdot \rho(\mathscr{C}(B; \beta, \alpha)), \tag{6.1.2}$$

$$g(\beta) \stackrel{\text{def}}{=} \rho\big((A + \beta I)^{-1}\big) \cdot \rho(B - \beta I), \tag{6.1.3}$$

provided all involved inverses exist. It can be seen that $g(\beta) \equiv f(0, \beta)$.

Theorem 6.1.1 (Wang, Wang and Li)**.** *For two M-matrices $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{m\times m}$, define f and g by* (6.1.2) *and* (6.1.3)*, and set*

$$\alpha_{\text{opt}} \overset{\text{def}}{=} \max_i A_{(i,i)}, \quad \beta_{\text{opt}} \overset{\text{def}}{=} \max_i B_{(i,i)}. \tag{6.1.4}$$

(a) *If both A and B are singular, then $f(\alpha, \beta) \equiv 1$ for $\alpha > \alpha_{\text{opt}}$ and $\beta > \beta_{\text{opt}}$, and $g(\beta) \equiv 1$ for $\beta > \beta_{\text{opt}}$;*

(b) *If one of A and B is nonsingular, then $f(\alpha, \beta)$ for $\alpha > \alpha_{\text{opt}}$ and $\beta > \beta_{\text{opt}}$ is strictly increasing in α and β and $f(\alpha, \beta) < 1$, and $g(\beta)$ for $\beta > \beta_{\text{opt}}$ is strictly increasing in β and $g(\beta) < 1$.*

Consequently, f can be defined by continuity for all $\alpha \geq \alpha_{\text{opt}}$ and $\beta \geq \beta_{\text{opt}}$ and g can be defined by continuity for all $\beta \geq \beta_{\text{opt}}$. Moreover, we have

$$\min_{\alpha\geq\alpha_{\text{opt}},\beta\geq\beta_{\text{opt}}} f(\alpha, \beta) = f(\alpha_{\text{opt}}, \beta_{\text{opt}}), \quad \min_{\beta\geq\beta_{\text{opt}}} g(\beta) = g(\beta_{\text{opt}}). \tag{6.1.5}$$

Proof. Both $A + \beta I$ and $B + \alpha I$ are nonsingular M-matrices for $\alpha > 0$ and $\beta > 0$; thus f and g are well-defined for $\alpha > \alpha_{\text{opt}}$ and $\beta > \beta_{\text{opt}}$ since $\alpha_{\text{opt}} \geq 0$ and $\beta_{\text{opt}} \geq 0$. In what follows, we will prove the claims for f only. Similar arguments work for g and thus are omitted.

Assume for the moment that both A and B are irreducible M-matrices. Write $A = sI - N$, where $s \geq 0$ and $N \geq 0$, and N is irreducible. By the Perron-Frobenius theorem [Theorem 3.1.1], there is a positive vector u such that $Nu = \rho(N)u$. It can be seen that $\lambda_{\min}(A) = s - \rho(N) \geq 0$, where $\lambda_{\min}(A)$ is as defined in Lemma 2.8.1(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 $-\mathscr{C}(A; \alpha, \beta) \geq 0$ and irreducible for $\alpha > \alpha_{\text{opt}}$ and $\beta > 0$, it follows from the Perron-Frobenius theorem that

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

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

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

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

$$\begin{aligned} f(\alpha,\beta) &= \rho(\mathscr{C}(A;\alpha,\beta)) \cdot \rho(\mathscr{C}(B;\beta,\alpha)) \\ &= \frac{\alpha - \lambda_{\min}(A)}{\lambda_{\min}(A) + \beta} \cdot \frac{\beta - \lambda_{\min}(B)}{\lambda_{\min}(B) + \alpha} \\ &= h_1(\alpha) h_2(\beta), \end{aligned}$$

where

$$h_1(\alpha) = \frac{\alpha - \lambda_{\min}(A)}{\lambda_{\min}(B) + \alpha}, \quad h_2(\beta) = \frac{\beta - \lambda_{\min}(B)}{\lambda_{\min}(A) + \beta}.$$

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

$$h_1'(\alpha) = \frac{\lambda_{\min}(A) + \lambda_{\min}(B)}{(\lambda_{\min}(B) + \alpha)^2} > 0, \quad h_2'(\beta) = \frac{\lambda_{\min}(A) + \lambda_{\min}(B)}{(\lambda_{\min}(A) + \beta)^2} > 0.$$

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

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

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 $A_{ij} \geq 0$ and $B_{ij} \geq 0$ for $i \neq j$. 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 $\alpha > \alpha_{\text{opt}}$ and $\beta > \beta_{\text{opt}}$; 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 and singular case, for that pair $\rho(\mathscr{C}(A_{ii}; \alpha, \beta)) \cdot \rho(\mathscr{C}(B_{jj}; \beta, \alpha)) \equiv 1$ for $\alpha \geq \alpha_{\text{opt}}$ and $\beta \geq \beta_{\text{opt}}$. Since for all other pairs (A_{ii}, B_{jj}),

$$\rho(\mathscr{C}(A_{ii}; \alpha, \beta)) \cdot \rho(\mathscr{C}(B_{jj}; \beta, \alpha)) \leq 1$$

by Item (a). Thus $f(\alpha, \beta) \equiv 1$. □

6.2 Framework of ADDA

The framework in this section actually works for any algebraic Riccati equation, provided all involved inverses exist. It is just that in general we are not able to establish a convergence theory similar to the one to be given in the next section for an MARE.

For any solution X of an MARE (1.0.1) and Y of the cMARE (6.0.1), it can be verified that

$$H \begin{pmatrix} I \\ X \end{pmatrix} = \begin{pmatrix} I \\ X \end{pmatrix} R, \quad H \begin{pmatrix} Y \\ I \end{pmatrix} = \begin{pmatrix} Y \\ I \end{pmatrix} (-S), \tag{6.2.1}$$

where

$$H = \begin{pmatrix} B & -D \\ C & -A \end{pmatrix}, \quad R = B - DX, \quad S = A - CY. \tag{6.2.2}$$

Given any scalars α and β, we have

$$(H-\beta I)\begin{pmatrix} I \\ X \end{pmatrix}(\quad R+\alpha I) = (H+\alpha I)\begin{pmatrix} I \\ X \end{pmatrix}(\quad R-\beta I),$$

$$(H-\beta I)\begin{pmatrix} Y \\ I \end{pmatrix}(-S+\alpha I) = (H+\alpha I)\begin{pmatrix} Y \\ I \end{pmatrix}(-S-\beta I).$$

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

$$(H-\beta I)\begin{pmatrix} I \\ X \end{pmatrix} \qquad\qquad = (H+\alpha I)\begin{pmatrix} I \\ X \end{pmatrix}\mathscr{C}(R;\beta,\alpha), \tag{6.2.3a}$$

$$(H-\beta I)\begin{pmatrix} Y \\ I \end{pmatrix}\mathscr{C}(S;\alpha,\beta) = (H+\alpha I)\begin{pmatrix} Y \\ I \end{pmatrix}. \tag{6.2.3b}$$

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

$$A_\beta = A+\beta I, \qquad\qquad B_\alpha = B+\alpha I, \tag{6.2.4}$$

$$U_{\alpha\beta} = A_\beta - CB_\alpha^{-1}D, \quad V_{\alpha\beta} = B_\alpha - DA_\beta^{-1}C, \tag{6.2.5}$$

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

$$M_0 \overset{\text{def}}{=} Z_3Z_2Z_1(H-\beta I) = \begin{pmatrix} E_0 & 0 \\ -X_0 & I \end{pmatrix}, \tag{6.2.6a}$$

$$L_0 \overset{\text{def}}{=} Z_3Z_2Z_1(H+\alpha I) = \begin{pmatrix} I & -Y_0 \\ 0 & F_0 \end{pmatrix}, \tag{6.2.6b}$$

where

$$E_0 = I-(\beta+\alpha)V_{\alpha\beta}^{-1}, \quad Y_0 = (\beta+\alpha)B_\alpha^{-1}DU_{\alpha\beta}^{-1}, \tag{6.2.7a}$$

$$F_0 = I-(\beta+\alpha)U_{\alpha\beta}^{-1}, \quad X_0 = (\beta+\alpha)U_{\alpha\beta}^{-1}CB_\alpha^{-1}. \tag{6.2.7b}$$

Pre-multiply the equations in (6.2.3) by $Z_3 Z_2 Z_1$ to get

$$M_0 \begin{pmatrix} I \\ X \end{pmatrix} = L_0 \begin{pmatrix} I \\ X \end{pmatrix} \mathscr{C}(R;\beta,\alpha), \quad M_0 \begin{pmatrix} Y \\ I \end{pmatrix} \mathscr{C}(S;\alpha,\beta) = L_0 \begin{pmatrix} Y \\ I \end{pmatrix}. \tag{6.2.8}$$

Our development up to this point differs from SDA of [24] 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 6.5 and numerical examples in section 8.2. From this point forward, ours is the same as in [24]. The idea is to construct a sequence of pairs $\{M_k, L_k\}$, $k = 0, 1, 2, \dots$ such that

$$M_k \begin{pmatrix} I \\ X \end{pmatrix} = L_k \begin{pmatrix} I \\ X \end{pmatrix} \left[\mathscr{C}(R;\beta,\alpha)\right]^{2^k}, \quad M_k \begin{pmatrix} Y \\ I \end{pmatrix} \left[\mathscr{C}(S;\alpha,\beta)\right]^{2^k} = L_k \begin{pmatrix} Y \\ I \end{pmatrix}, \tag{6.2.9}$$

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

$$M_k = \begin{pmatrix} E_k & 0 \\ -X_k & I \end{pmatrix}, \quad L_k = \begin{pmatrix} I & -Y_k \\ 0 & F_k \end{pmatrix}. \tag{6.2.10}$$

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 [12, 18, 30] and more recently in [3, 7, 37]. The idea is to seek suitable $\check{M}, \check{L} \in \mathbb{R}^{(m+n)\times(m+n)}$ such that

$$\operatorname{rank}\big((\check{M}, \check{L})\big) = m + n, \quad (\check{M}, \check{L}) \begin{pmatrix} L_k \\ -M_k \end{pmatrix} = 0 \tag{6.2.11}$$

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 (6.2.9) 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 (6.2.11) leads to $\{M_{k+1}, L_{k+1}\}$ having the forms of (6.2.10). For this, we turn to the constructions of $\{\check{M}, \check{L}\}$ in [14, 15, 24, 29]:

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

with which $M_{k+1} = \check{M} M_k$ and $L_{k+1} = \check{L} L_k$ have the forms of (6.2.10) with

$$E_{k+1} = E_k(I_m - Y_k X_k)^{-1} E_k, \tag{6.2.12a}$$

$$F_{k+1} = F_k(I_n - X_k Y_k)^{-1} F_k, \tag{6.2.12b}$$

$$X_{k+1} = X_k + F_k(I_n - X_k Y_k)^{-1} X_k E_k, \tag{6.2.12c}$$

$$Y_{k+1} = Y_k + E_k(I_m - Y_k X_k)^{-1} Y_k F_k. \tag{6.2.12d}$$

By now I have presented the framework of ADDA:

1. Pick suitable α and β for (best) convergence rate;
2. Compute M_0 and L_0 of (6.2.6) by (6.2.4), (6.2.5), and (6.2.7);
3. Iteratively compute M_k and L_k by (6.2.12) until convergence.

Associated with this general framework arise a few questions:

1. Are the iterative formulas in (6.2.12) 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 [24] and SDA-ss of Bini, Meini, and Poloni [11]?

The first three questions will be addressed in the next section while the last question will be answered in section 6.5.

6.3 Analysis

Recall that W defined by (1.0.2) is a nonsingular or an irreducible singular M-matrix. An MARE (1.0.1) has a unique minimal nonnegative solution $\varPhi$ in chapter 3 and the cMARE (6.0.1) has a unique minimal nonnegative solution $\varPsi$. Some properties of $\varPhi$ and $\varPsi$ are summarized in Theorem 6.3.1 below. They are needed in order to answer the questions we posed at the end of the previous section.

Theorem 6.3.1 ([19, 20, 21])**.** *Assume* (7.0.1).

(a) *An MARE* (1.0.1) *has a unique minimal nonnegative solution $\varPhi$, and and its cMARE* (6.0.1) *has a unique minimal nonnegative solution $\varPsi$;*

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

(c) *If W is* nonsingular*, then $A - \varPhi D$ and $B - D\varPhi$ 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*

$$W \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = 0, \quad \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}^{\mathrm{T}} W = 0. \tag{6.3.1}$$

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

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

[1] [19, Theorem 4.8] says in this case $D\varPhi v_1 = Dv_2$ which leads to $(B - D\varPhi)v_1 = Bv_1 - Dv_2 = 0$.

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 section 6.2. In particular

$$S = A - C\Psi, \quad R = B - D\Phi, \tag{6.2.2$'$}$$

and (6.2.9) and (6.2.10) yield immediately

$$\begin{aligned}
E_k &= (I - Y_k\Phi)\ [\mathscr{C}(R;\beta,\alpha)]^{2^k}, && (6.3.2a)\\
\Phi - X_k &= \quad F_k\Phi \quad\ [\mathscr{C}(R;\beta,\alpha)]^{2^k}, && (6.3.2b)\\
\Psi - Y_k &= \quad E_k\Psi \quad\ [\mathscr{C}(S;\alpha,\beta)]^{2^k}, && (6.3.2c)\\
F_k &= (I - X_k\Psi)\,[\mathscr{C}(S;\alpha,\beta)]^{2^k}. && (6.3.2d)
\end{aligned}$$

Examining (6.3.2), we see that ADDA will converge if X_k and Y_k are uniformly bounded with respect to k, and if

$$\rho(\mathscr{C}(R;\beta,\alpha)) < 1, \quad \rho(\mathscr{C}(S;\alpha,\beta)) < 1, \tag{6.3.3a}$$

because then E_k and F_k are uniformly bounded with respect to k, and

$$[\mathscr{C}(R;\beta,\alpha)]^{2^k} \to 0, \quad [\mathscr{C}(S;\alpha,\beta)]^{2^k} \to 0 \tag{6.3.3b}$$

as $k \to \infty$, implying that $\Phi - X_k \to 0$ and $\Psi - Y_k \to 0$ as $k \to \infty$. This is one of the guiding principles in [24] which enforces

$$\alpha = \beta \geq \max_{i,j}\{A_{(i,i)}, B_{(j,j)}\} \tag{6.3.4}$$

which in turn ensures that X_k and Y_k are uniformly bounded and also ensures (6.3.3a) and thus (6.3.3b) because, by Theorem 6.3.1(c), both[2] S and R are nonsingular M-

[2] That R is a nonsingular M-matrix is stated explicitly in Theorem 6.3.1(c). For S, we apply Theorem 6.3.1(c) to the cMARE (6.0.1) identified as an MARE in the form of (1.0.1) with its coefficient matrix as $\begin{pmatrix} A & -C \\ -D & B \end{pmatrix}$.

matrices if[3] W is a nonsingular M-matrix. Later Guo, Iannazzo, and Meini [22] observed that SDA of [24] still converges even if W is a singular irreducible M-matrix. This observation was formally proved in [13]. Guo, Iannazzo, and Meini [22, Theorem 4.4] also proved that taking

$$\alpha = \beta = \max_{i,j}\{A_{(i,i)}, B_{(j,j)}\} \tag{6.3.5}$$

makes the resulting SDA converge the fastest subject to (6.3.4). Another critical implication of (6.3.4) 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 $k \geq 0$ all nonnegative [24], a property that enables SDA of [24] (with some minor but crucial implementation changes [42]) to compute $\varPhi$ with deserved entrywise relative accuracy as argued in [42].

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

$$\alpha \geq \alpha_{\rm opt} \overset{\rm def}{=} \max_i A_{(i,i)}, \quad \beta \geq \beta_{\rm opt} \overset{\rm def}{=} \max_j B_{(j,j)}, \tag{6.3.6}$$

but 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 in Theorem 6.3.3, 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.

That the rate of convergence of a doubling algorithm on a matrix Riccati-type equation is dependent on the product of some two spectral radii is not new. In fact, the convergence analysis in [22, 24, 29] all suggested that.

[3] This is the case studied in [24].

The assumption (7.0.1) implies that A and B are nonsingular M-matrices by Lemma 2.8.1(e). Therefore both $\alpha_{\text{opt}} > 0$ and $\beta_{\text{opt}} > 0$.

Lemma 6.3.1 (Wang, Wang and Li)**.** *Assume* (7.0.1)*. If* $\alpha > 0$ *and* $\beta > 0$*, then* A_β*,* B_α*,* $U_{\alpha\beta}$*, and* $V_{\alpha\beta}$ *defined in* (6.2.4) *and* (6.2.5) *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 2.8.1(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 2.8.1(e). □

Theorem 6.3.2 (Wang, Wang and Li)**.** *Assume* (7.0.1) *and* (6.3.6)*.*

(a) *We have*

$$E_0 \leq 0,\ F_0 \leq 0,\ \mathscr{C}(R; \beta, \alpha) \leq 0,\ \mathscr{C}(S; \alpha, \beta) \leq 0, \tag{6.3.7}$$

$$0 \leq X_0 \leq \varPhi,\ 0 \leq Y_0 \leq \varPsi. \tag{6.3.8}$$

If W *is also irreducible, then*

$$E_0 < 0,\ F_0 < 0,\ \mathscr{C}(R; \beta, \alpha) < 0,\ \mathscr{C}(S; \alpha, \beta) < 0, \tag{6.3.7$'$}$$

$$0 \leq X_0 < \varPhi,\ 0 \leq Y_0 < \varPsi. \tag{6.3.8$'$}$$

(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*

$$E_k \geq 0,\ F_k \geq 0,\ 0 \leq X_{k-1} \leq X_k \leq \varPhi,\ 0 \leq Y_{k-1} \leq Y_k \leq \varPsi \quad \textit{for } k \geq 1. \tag{6.3.9}$$

If W is also irreducible, then

$$E_k > 0,\ F_k > 0,\ 0 \leq X_{k-1} < X_k < \varPhi,\ 0 \leq Y_{k-1} < Y_k < \varPsi \quad \textit{for } k \geq 1. \tag{6.3.9$'$}$$

Proof. Our proof is largely the same as the proofs in [22, p.1088].

(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 (6.3.6). For E_0 and F_0, we note

$$E_0 = V_{\alpha\beta}^{-1}[V_{\alpha\beta} - (\beta+\alpha)I] \tag{6.3.10a}$$

$$= V_{\alpha\beta}^{-1}(B - \beta I - DA_{\beta}^{-1}C), \tag{6.3.10b}$$

$$F_0 = U_{\alpha\beta}^{-1}[U_{\alpha\beta} - (\beta+\alpha)I] \tag{6.3.10c}$$

$$= U_{\alpha\beta}^{-1}(A - \alpha I - CB_{\alpha}^{-1}D). \tag{6.3.10d}$$

Since A_β, B_α, $V_{\alpha\beta}$, and $U_{\alpha\beta}$ are nonsingular M-matrices by Lemma 6.3.1, 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 (6.3.2b) and (6.3.2c) for $k = 0$ yields $\varPhi - X_0 \geq 0$ and $\varPsi - Y_0 \geq 0$, respectively.

Now suppose W is irreducible. By Lemma 6.3.1, 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 $V_{\alpha\beta} - (\beta+\alpha)I$ and $U_{\alpha\beta} - (\beta+\alpha)I$ both of which are nonpositive are zeros. Therefore $E_0 < 0$ and $F_0 < 0$ by (6.3.10a) and (6.3.10c). Theorem 6.3.1(b) implies that $(S+\beta I)^{-1} > 0$, $(R+\alpha I)^{-1} > 0$, and no columns of $S - \alpha I$ and $R - \beta I$ both of which are nonpositive 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

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

because $\varPhi > 0$ and $\varPsi > 0$ by Theorem 6.3.1(b) and (6.3.7′).

(b) and **(c)** We have $I - X_0 Y_0 \geq I - \varPhi\varPsi$ and $I - Y_0 X_0 \geq I - \varPsi\varPhi$. Suppose for the moment that W is nonsingular. Then both $I - \varPhi\varPsi$ and $I - \varPsi\varPhi$ are nonsingular M-matrices by Theorem 6.3.1(e), and thus $I - X_0 Y_0$ and $I - Y_0 X_0$ are nonsingular M-matrices, too, by Lemma 2.8.1(b).

Now suppose W is an irreducible singular matrix. By Theorem 6.3.1(d), we have $\varPsi\varPhi v_1 \leq v_1$, where $v_1 > 0$ is defined in Theorem 6.3.1(d). So $\rho(\varPsi\varPhi) \leq 1$ by [9, Theorem 1.11, p.28]. By part **(a)** of this theorem, $0 \leq X_0 < \varPhi$ and $0 \leq Y_0 < \varPsi$. Therefore $0 \leq X_0 Y_0 < \varPhi\varPsi$. Since $\varPhi\varPsi$ is irreducible, we conclude by [9, Corollary 1.5, p.27]

$$\rho(Y_0 X_0) = \rho(X_0 Y_0) < \rho(\varPsi\varPhi) = \rho(\varPhi\varPsi) \leq 1,$$

and thus $I - Y_0 X_0$ and $I - X_0 Y_0$ are nonsingular M-matrices. This proves part **(b)** for $k = 0$.

Since $E_0 \leq 0$ and $F_0 \leq 0$, and $I - Y_0 X_0$ and $I - X_0 Y_0$ are nonsingular M-matrices, we deduce from (6.2.12) that

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

By (6.3.2b) and (6.3.2c),

$$\varPhi - X_1 = F_1 \varPhi \left[\mathscr{C}(R; \beta, \alpha)\right]^2, \quad \varPsi - Y_1 = E_1 \varPsi \left[\mathscr{C}(S; \alpha, \beta)\right]^2, \tag{6.3.11}$$

yielding $\varPhi - X_1 \geq 0$ and $\varPsi - Y_1 \geq 0$, respectively. Consider now W is also irreducible. We have, by (6.3.7′) and (6.3.8′) and (6.2.12),

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

and then $X_1 < \Phi$ and $Y_1 < \Psi$ by (6.3.11). 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 6.3.2 is that all formulas in section 6.2 for ADDA are well-defined under the assumptions (7.0.1) and (6.3.6).

Next we look into choosing α and β subject to (6.3.6) to optimize the convergence speed. We have (6.3.2) which yields

$$\begin{aligned} 0 \leq \Phi - X_k &= (I - X_k\Psi)\,[\mathscr{C}(S;\alpha,\beta)]^{2^k}\,\Phi\,[\mathscr{C}(R;\beta,\alpha)]^{2^k} && \text{(6.3.12a)} \\ &\leq \qquad\qquad [\mathscr{C}(S;\alpha,\beta)]^{2^k}\,\Phi\,[\mathscr{C}(R;\beta,\alpha)]^{2^k}\,, && \text{(6.3.12b)} \\ 0 \leq \Psi - Y_k &= (I - Y_k\Phi)\,[\mathscr{C}(R;\beta,\alpha)]^{2^k}\,\Psi\,[\mathscr{C}(S;\alpha,\beta)]^{2^k} && \text{(6.3.12c)} \\ &\leq \qquad\qquad [\mathscr{C}(R;\beta,\alpha)]^{2^k}\,\Psi\,[\mathscr{C}(S;\alpha,\beta)]^{2^k}\,. && \text{(6.3.12d)} \end{aligned}$$

Now if W is a nonsingular M-matrix, then both R and S are nonsingular M-matrices, too, by Theorem 6.3.1(c). Therefore

$$\rho(\mathscr{C}(R;\beta,\alpha)) < 1,\ \rho(\mathscr{C}(S;\alpha,\beta)) < 1 \text{ under (6.3.4)}, \tag{6.3.13}$$

implying $X_k \to \Phi$ and $Y_k \to \Psi$ as $k \to \infty$. This is what was proved in [24]. 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 [22] that one of the spectral radii in (6.3.13) 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, [22, Theorem 4.4] implies that the best choice is given by (6.3.5) in the sense that both spectral radii in $\rho(\mathscr{C}(R;\alpha,\alpha))$ and $\rho(\mathscr{C}(S;\alpha,\alpha))$ are minimized subject to $\alpha \geq \max_{i,j}\{A_{(i,i)}, B_{(j,j)}\}$.

We can do better by allowing α and β to be different, with the help of Theorem 6.1.1. The main result is summarized in the following theorem.

Theorem 6.3.3 (Wang, Wang and Li)**.** *Assume* (7.0.1) *and* (6.3.6). *We have*

$$\limsup_{k\to\infty} \|\Phi - X_k\|^{1/2^k} \le \rho(\mathscr{C}(S;\alpha,\beta)) \cdot \rho(\mathscr{C}(R;\beta,\alpha)), \tag{6.3.14a}$$

$$\limsup_{k\to\infty} \|\Psi - Y_k\|^{1/2^k} \le \rho(\mathscr{C}(R;\beta,\alpha)) \cdot \rho(\mathscr{C}(S;\alpha,\beta)). \tag{6.3.14b}$$

The optimal α and β that minimize the right-hand sides of (6.3.14) *are $\alpha = \alpha_{\text{opt}}$ and $\beta = \beta_{\text{opt}}$.*

Proof. Since all matrix norms are equivalent, we may assume that $\|\cdot\|$ is consistent. By (6.3.12b), 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 $k \to \infty$, unless $\Phi = 0$ in which case both sides are 0 for all k. Thus (6.3.14a) holds. Similarly we have (6.3.14b). 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 6.1.1, $\rho(\mathscr{C}(R;\beta,\alpha)) \cdot \rho(\mathscr{C}(S;\alpha,\beta))$ is either strictly increasing if at least one of R and S is nonsingular or identically 1, subject to (6.3.6). So in any case, $\alpha = \alpha_{\text{opt}}$ and $\beta = \beta_{\text{opt}}$ minimize the product $\rho(\mathscr{C}(S;\alpha,\beta)) \cdot \rho(\mathscr{C}(R;\beta,\alpha))$. □

6.4 Optimal ADDA

We are now ready to present our ADDA, basing on the framework in section 6.2 and analysis in section 6.3.

Algorithm 6.4.1.

ADDA for an MARE $XDX - AX - XB + C = 0$ and,

as a by-product, for the cMARE $YCY - YA - BY + D = 0$.

```
1   Pick $\alpha \geq \alpha_{\text{opt}}$ and $\beta \geq \beta_{\text{opt}}$;
2   $A_\beta \overset{\text{def}}{=} A + \beta I$, $B_\alpha \overset{\text{def}}{=} B + \alpha I$;
3   Compute $A_\beta^{-1}$ and $B_\alpha^{-1}$;
4   Compute $V_{\alpha\beta}$ and $U_{\alpha\beta}$ as in (6.2.5) and then their inverses;
5   Compute $E_0$ by (6.3.10b), $F_0$ by (6.3.10d), $X_0$ and $Y_0$ by (6.2.7);
6   Compute $(I - X_0 Y_0)^{-1}$ and $(I - Y_0 X_0)^{-1}$;
7   Compute $X_1$ and $Y_1$ by (6.2.12c) and (6.2.12d);
8   For $k = 1, 2, \ldots$, until convergence
9       Compute $E_k$ and $F_k$ by (6.2.12a) and (6.2.12b)
        (after substituting $k + 1$ for $k$);
10      Compute $(I - X_k Y_k)^{-1}$ and $(I - Y_k X_k)^{-1}$;
11      Compute $X_{k+1}$ and $Y_{k+1}$ by (6.2.12c) and (6.2.12d);
12  Enddo
```

Remark **6.4.1.** ADDA differs from SDA of [24] only in its initial setup – Lines 1 – 5 that build two parameters α and β into the algorithm. In [42], 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, 43] to invert all nonsingular M-matrices. Every comment in [42, 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 6.3.3 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 (6.3.10b) and F_0 by (6.3.10d) 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 $\varPhi$ and $\varPsi$ 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 6.4.1 ADDA can encounter overflown in E_k or F_k before X_k and Y_k converge with a desired accuracy. This happened in one of our test runs. To cure this, we notice that scaling 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., $\eta = \sqrt{\|F_k\|/\|E_k\|}$, and scale E_k and F_k to ηE_k and $\eta^{-1}F_k$. Which matrix norm $\|\cdot\|$ is not particularly important and in our tests, we used the ℓ_1-operator norm $\|\cdot\|_1$. $\diamond$

The *optimal ADDA* is the one with $\alpha = \alpha_{\text{opt}}$ and $\beta = \beta_{\text{opt}}$. Since there is little reason not to use the optimal ADDA, except for the situation we mentioned in Remark 6.4.1 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.

6.5 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 [24], and SDA combined with the shrink-and-shift technique (SDA-ss) of [11].

The right-hand sides in (6.3.14) 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

$$\alpha_{\text{opt}} \overset{\text{def}}{=} \max_i A_{(i,i)} \geq \beta_{\text{opt}} \overset{\text{def}}{=} \max_i B_{(i,i)}. \tag{6.5.1}$$

Let $\lambda_{\min}(S)$ be the eigenvalue of S in (6.2.2′) with the smallest real part among all its eigenvalues. We know $\lambda_{\min}(S) \geq 0$, and let $\lambda_{\min}(R)$ be the same for R also in (6.2.2′).

We have the convergence rate for the optimal ADDA

$$r_{\text{adda}} = \frac{\alpha_{\text{opt}} - \lambda_{\min}(S)}{\beta_{\text{opt}} + \lambda_{\min}(S)} \cdot \frac{\beta_{\text{opt}} - \lambda_{\min}(R)}{\alpha_{\text{opt}} + \lambda_{\min}(R)}. \tag{6.5.2}$$

Estimates in (6.3.14) with $\alpha = \beta$ hold for SDA. Apply [22, Theorem 4.4] to conclude that the convergence rate for the optimal SDA is

$$r_{\text{sda}} = \frac{\alpha_{\text{opt}} - \lambda_{\min}(S)}{\alpha_{\text{opt}} + \lambda_{\min}(S)} \cdot \frac{\alpha_{\text{opt}} - \lambda_{\min}(R)}{\alpha_{\text{opt}} + \lambda_{\min}(R)} \tag{6.5.3}$$

upon noticing (6.5.1).

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

$$\beta \geq \beta_{\text{opt}} \overset{\text{def}}{=} \max_j B_{(j,j)}, \tag{6.5.4}$$

set

$$\widehat{H} = I - \beta^{-1}H, \quad \widehat{A} = I + \beta^{-1}A, \quad \widehat{B} = I - \beta^{-1}B, \tag{6.5.5}$$

where H is defined as in (6.2.2). With S and R given by (6.2.2′), we have

$$\widehat{H}\begin{pmatrix} I \\ \Phi \end{pmatrix} = \begin{pmatrix} I \\ \Phi \end{pmatrix}\widehat{R}, \qquad \widehat{H}\begin{pmatrix} \Psi \\ I \end{pmatrix}\widehat{S} = \begin{pmatrix} \Psi \\ I \end{pmatrix}, \tag{6.5.6a}$$

$$\widehat{R} = I - \beta^{-1}R, \qquad\qquad \widehat{S} = (I + \beta^{-1}S)^{-1}. \tag{6.5.6b}$$

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

$$\widehat{M_0} = \begin{pmatrix} \widehat{E}_0 & 0 \\ -\widehat{X}_0 & I \end{pmatrix}, \quad \widehat{L}_0 = \begin{pmatrix} I & -\widehat{Y}_0 \\ 0 & \widehat{F}_0 \end{pmatrix}, \tag{6.5.7}$$

where

$$\widehat{E}_0 = \widehat{B} + \beta^{-2} D \widehat{A}^{-1} C, \quad \widehat{Y}_0 = \beta^{-1} D \widehat{A}^{-1}, \tag{6.5.8a}$$

$$\widehat{F}_0 = \widehat{A}^{-1}, \qquad \widehat{X}_0 = \beta^{-1} \widehat{A}^{-1} C. \tag{6.5.8b}$$

It can be verified that $\widehat{H} = \widehat{L}_0^{-1} \widehat{M_0}$, substituting which into the equations in (6.5.6) 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 [24] (and also in section 6.3). SDA-ss seeks to construct a sequence of pairs $\{\widehat{M_k}, \widehat{L}_k\}$, $k = 0, 1, 2, \ldots$ such that

$$\widehat{M_k} \begin{pmatrix} I \\ \Phi \end{pmatrix} = \widehat{L}_k \begin{pmatrix} I \\ \Phi \end{pmatrix} \widehat{R}^{2^k}, \quad \widehat{M_k} \begin{pmatrix} \Psi \\ I \end{pmatrix} \widehat{S}^{2^k} = \widehat{L}_k \begin{pmatrix} \Psi \\ I \end{pmatrix}, \tag{6.5.9}$$

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.,

$$\widehat{M_k} = \begin{pmatrix} \widehat{E}_k & 0 \\ -\widehat{X}_k & I \end{pmatrix}, \quad \widehat{L}_k = \begin{pmatrix} I & -\widehat{Y}_k \\ 0 & \widehat{F}_k \end{pmatrix}. \tag{6.5.10}$$

The formulas (6.2.12) 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 $\widehat{Y}_k$:

$$\Phi - \widehat{X}_k = (I - \widehat{X}_k \Psi) \widehat{S}^{2^k} \Phi \widehat{R}^{2^k} \le \widehat{S}^{2^k} \Phi \widehat{R}^{2^k}, \tag{6.5.11}$$

$$\Psi - \widehat{Y}_k = (I - \widehat{Y}_k \Phi) \widehat{R}^{2^k} \Psi \widehat{S}^{2^k} \le \widehat{R}^{2^k} \Psi \widehat{S}^{2^k}. \tag{6.5.12}$$

Consequently

$$\limsup_{k\to\infty} \|\mathit{\Phi} - \widehat{X}_k\|^{1/2^k},\ \limsup_{k\to\infty} \|\mathit{\Psi} - \widehat{Y}_k\|^{1/2^k} \le \rho(\widehat{R}) \cdot \rho(\widehat{S}). \tag{6.5.13}$$

In view of this inequality, (6.5.4) and Theorem 6.1.1, we conclude that the convergence rate of the optimal SDA-ss is

$$r_{\text{sda-ss}} = \frac{1 - \beta_{\text{opt}}^{-1}\lambda_{\min}(R)}{1 + \beta_{\text{opt}}^{-1}\lambda_{\min}(S)} = \frac{\beta_{\text{opt}} - \lambda_{\min}(R)}{\beta_{\text{opt}} + \lambda_{\min}(S)}. \tag{6.5.14}$$

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

$$\begin{aligned}
\frac{r_{\text{adda}}}{r_{\text{sda}}} &= \frac{\beta - \lambda_R}{\alpha - \lambda_R} \cdot \frac{\alpha + \lambda_S}{\beta + \lambda_S} \\
&= 1 - \frac{(\lambda_R + \lambda_S)(\alpha - \beta)}{(\alpha - \lambda_R)(\beta + \lambda_S)}, && (6.5.15)\\
\frac{r_{\text{adda}}}{r_{\text{sda-ss}}} &= \frac{\alpha - \lambda_S}{\alpha + \lambda_R} \\
&= 1 - \frac{\lambda_R + \lambda_S}{\alpha + \lambda_R}, && (6.5.16)\\
\frac{r_{\text{sda-ss}}}{r_{\text{sda}}} &= \frac{\beta - \lambda_R}{\beta + \lambda_S} \cdot \frac{\alpha + \lambda_S}{\alpha - \lambda_S} \cdot \frac{\alpha + \lambda_R}{\alpha - \lambda_R} \\
&= 1 - \frac{(\lambda_R + \lambda_S)[\alpha(\alpha - \beta) - \lambda_S(\alpha - \lambda_R) - \alpha(\beta - \lambda_R)]}{(\beta + \lambda_S)(\alpha - \lambda_S)(\alpha - \lambda_R)}. && (6.5.17)
\end{aligned}$$

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 [13]. Suppose, in what follows, that $\lambda_R + \lambda_S > 0$, and recall (6.5.1). 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

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

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 (6.5.5) is done. For example, instead of (6.5.1), if

$$\max_i A_{(i,i)} < \max_i B_{(i,i)}. \tag{6.5.18}$$

Then we still have (6.5.14), but, instead of (6.5.3),

$$r_{\text{sda}} = \frac{\beta-\lambda_S}{\beta+\lambda_S} \cdot \frac{\beta-\lambda_R}{\beta+\lambda_R}. \tag{6.5.19}$$

Then

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

always, indicating SDA-ss is slower than SDA. To overcome this, when (6.5.18) holds, SDA-ss should be applied to the cMARE (6.0.1), instead, and as a by-product, Φ is computed as the minimal nonnegative solution to the complementary MARE of the cMARE (6.0.1).

6.6 Doubling Algorithms by General Bilinear Transformations

The doubling algorithms SDA, SDA-ss, and ADDA are constructed, respectively, by

$$\begin{array}{rll} \textit{Cayley}\text{ transformation:} & t \to \mathscr{C}(t;\alpha,\alpha) = (t-\alpha)/(t+\alpha) & \text{for SDA,} \\ \textit{shrink-and-shift}\text{ transformation:} & t \to t/\beta - 1 & \text{for SDA-ss,} \\ \textit{generalized Cayley}\text{ transformation:} & t \to \mathscr{C}(t;\alpha,\beta) = (t-\alpha)/(t+\beta) & \text{for ADDA.} \end{array}$$

These transformations are three special cases of the following more general *bilinear* (also called *Möbius*) transformation:

$$t \to \mathscr{B}(t;\alpha,\alpha_1,\beta,\beta_1) \overset{\text{def}}{=} (\alpha_1 t-\alpha)/(\beta_1 t+\beta). \tag{6.6.1}$$

It is tempting to ask if some faster doubling algorithm than ADDA could be constructed with this bilinear transformation because of two additional parameters α_1 and β_1 to work with. In what follows we shall explain that optimal ADDA is still the best among all possible doubling algorithms coming out of (6.6.1).

The framework in section 6.2 can be modified to accommodate $\mathscr{B}(t;\alpha,\alpha_1,\beta,\beta_1)$ upon noticing that, similar to (6.2.3),

$$(\beta_1 H-\beta I)\begin{pmatrix} I \\ X \end{pmatrix} = (\alpha_1 H+\alpha I)\begin{pmatrix} I \\ X \end{pmatrix}\mathscr{B}(R;\beta,\beta_1,\alpha,\alpha_1), \tag{6.6.2a}$$

$$(\beta_1 H-\beta I)\begin{pmatrix} Y \\ I \end{pmatrix}\mathscr{B}(S;\alpha,\alpha_1,\beta,\beta_1) = (\alpha_1 H+\alpha I)\begin{pmatrix} Y \\ I \end{pmatrix}. \tag{6.6.2b}$$

Assuming no breakdown occurs, i.e., all involved inverses exist, in the end we will have error equations, similar to those in (6.3.2),

$$\varPhi - X_k = (I-X_k\varPsi)\left[\mathscr{B}(S;\alpha,\alpha_1,\beta,\beta_1)\right]^{2^k}\varPhi\left[\mathscr{B}(R;\beta,\beta_1,\alpha,\alpha_1)\right]^{2^k}, \tag{6.6.3a}$$

$$\varPsi - Y_k = (I-Y_k\varPhi)\left[\mathscr{B}(R;\beta,\beta_1,\alpha,\alpha_1)\right]^{2^k}\varPsi\left[\mathscr{B}(S;\alpha,\alpha_1,\beta,\beta_1)\right]^{2^k}. \tag{6.6.3b}$$

There are four cases to consider

1. $\alpha_1 \neq 0$ and $\beta_1 \neq 0$. Since $\mathscr{B}(t;\alpha,\alpha_1,\beta,\beta_1) = (\alpha_1/\beta_1)\cdot\mathscr{C}(t;\alpha/\alpha_1,\beta/\beta_1)$, both equations in (6.6.3) are the same as those for ADDA with the generalized Cayley transformation $\mathscr{C}(t;\alpha/\alpha_1,\beta/\beta_1)$. This implies that any resulting doubling algorithm is an ADDA.

2. $\alpha_1 \neq 0$, $\beta_1 = 0$ (and then $\beta \neq 0$ in order for $\mathscr{B}(t;\alpha,\alpha_1,\beta,\beta_1)$ to be well-defined):
   (a) If $\alpha = 0$, then $\mathscr{B}(t;\alpha,\alpha_1,\beta,\beta_1) = (\alpha_1/\beta)t$ and thus the equations in (6.6.3) become

   $$\varPhi - X_k = (I - X_k\varPsi)S^{2^k}\varPhi R^{-2^k}, \quad \varPsi - Y_k = (I - Y_k\varPhi)R^{-2^k}\varPsi S^{2^k}. \quad (6.6.4)$$

   Convergence of X_k and Y_k to $\varPhi$ and $\varPsi$, respectively, is no longer guaranteed.

   (b) If $\alpha \neq 0$, then $\mathscr{B}(t;\alpha,\alpha_1,\beta,\beta_1) = (\alpha/\beta)[t(\alpha/\alpha_1)^{-1} - 1]$ and thus the equations in (6.6.3) are the same as those for an SDA-ss. This implies that any resulting doubling algorithm is an SDA-ss.
3. $\alpha_1 = 0$ (and then $\alpha \neq 0$ in order for $\mathscr{B}(t;\beta,\beta_1,\alpha,\alpha_1)$ to be well-defined), $\beta_1 \neq 0$. This case is essentially the same as the previous one: $\alpha_1 \neq 0$, $\beta_1 = 0$.
4. $\alpha_1 = \beta_1 = 0$, i.e., $\mathscr{B}(t;\alpha,\alpha_1,\beta,\beta_1)$ is constant. This is the trivial case. Convergence of X_k and Y_k to $\varPhi$ and $\varPsi$, respectively, is not possible because no information on H is built into the algorithm.

In summary, possible doubling algorithms derivable from the general bilinear transformation are SDA, SDA-ss, ADDA, the trivial ones by $\mathscr{B}(t;\alpha,\alpha_1,\beta,\beta_1) \equiv 1$ or $\mathscr{B}(t;\alpha,\alpha_1,\beta,\beta_1) \equiv 0$, and the one by $\mathscr{B}(t;\alpha,\alpha_1,\beta,\beta_1) = t$. Among all, optimal ADDA is the best.

In principle, possible doubling algorithms can also be constructed by noticing that, similar to (6.2.3) and (6.6.2),

$$h(H)\begin{pmatrix} I \\ X \end{pmatrix} = \begin{pmatrix} I \\ X \end{pmatrix} h(R), \quad h(H)\begin{pmatrix} Y \\ I \end{pmatrix}[h(S)]^{-1} = \begin{pmatrix} Y \\ I \end{pmatrix},$$

where $h(\cdot)$ is a rational function (or any other more complicated function). But without knowing a particular effective $h(\cdot)$, such a generality has no practical value.

CHAPTER 7

d-ADDA: Deflating irreducible singular M-matrix Riccati equation

Doubling algorithms are linear convergence for critical case. So in this chapter we will propose a deflation technique to improve the rate of convergence of doubling algorithms. Since the necessary condition for being in the critical case is H being singular, to speed up the convergence, Guo, Iannazzo, and Meini [22] proposed to shift away its eigenvalue 0 to a properly chosen positive number η:

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

before SDA is applied, where $z = \begin{pmatrix} x \\ y \end{pmatrix}$, and $w \in \mathbb{R}^{m+n}$ is entrywise nonnegative such that $w^{\mathrm{T}} z = 1$. Dramatic improvements in reducing the number of iterative steps required for convergence were witnessed. In this chapter, 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.

In the rest of this chapter, A, B, C, and D, unless explicitly stated differently, are reserved for the coefficient matrices of an MARE (1.0.1) for which

$$\boxed{\begin{array}{l}W \text{ defined by (1.0.2) is an irreducible singular } M\text{-}\\ \text{matrix, and (6.0.2) holds, where } 0 < u, x \in \mathbb{R}^m \text{ and}\\ 0 < v, y \in \mathbb{R}^n.\end{array}} \tag{7.0.1}$$

Note that assuming (6.0.2) 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 (6.0.2).

This chapter is organized as follows. We begin by laying out our deflating framework and its convergent analysis in section 7.2, followed by the analysis of convergence in section 7.3 and then give out two efficient numerical realizations of the framework in sections 7.4 and 7.5. We outline the shifting approach of Guo, Iannazzo, and Meini [22] in section 7.6 for comparison purpose.

7.1 Deflating an Irreducible Singular MARE

Assume that (7.0.1) holds. We have three cases: $\mu = u^{\mathrm{T}}x - v^{\mathrm{T}}y > 0$, $\mu = 0$, and $\mu < 0$. The case $\mu < 0$ can be converted to the case $\mu > 0$ by transposing (1.0.1) to get

$$ZD^{\mathrm{T}}Z - ZA^{\mathrm{T}} - B^{\mathrm{T}}Z + C^{\mathrm{T}} = 0, \tag{7.1.1}$$

where $Z = X^{\mathrm{T}}$. This MARE has the unique minimal nonnegative solution $\varPhi^{\mathrm{T}}$, and

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

as the result of (6.0.2), and the new μ for (7.1.1) is positive. By Theorem 6.3.1, we have $\varPhi^{\mathrm{T}}v = u$.

If $m = 1$ and $\mu \geq 0$, then $B - D\varPhi = 0$ by Theorem 6.3.1(c). An MARE (1.0.1) after setting $X = \varPhi$ becomes $C - A\varPhi = 0$ to give $\varPhi = A^{-1}C$ because A is a nonsingular M-matrix.

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

$$\mu = u^{\mathrm{T}}x - v^{\mathrm{T}}y \geq 0, \quad m \geq 2. \tag{7.1.2}$$

By Theorem 6.3.1, $\varPhi x = y$. In what follows, we will first present a general framework for deflating an irreducible singular MARE with (7.1.2), and then its convergence analysis. Two numerical realizations of the framework will be discussed in detail in Remark 7.3.2.

7.2 General Framework

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

$$V^{-1}z = \delta e_1, \quad z = \begin{pmatrix} x \\ y \end{pmatrix}. \tag{7.2.1}$$

Any numerical realization of this framework in Remark 7.3.2 is simply a way of constructing such a matrix V.

$\varPhi$ satisfies the MARE (1.0.1), or equivalently,

$$H \begin{pmatrix} I \\ \varPhi \end{pmatrix} = \begin{pmatrix} I \\ \varPhi \end{pmatrix} R, \quad R = B - D\varPhi \tag{7.2.2}$$

which is equivalent to

$$V^{-1}HVV^{-1} \begin{pmatrix} I \\ \varPhi \end{pmatrix} = V^{-1} \begin{pmatrix} I \\ \varPhi \end{pmatrix} R. \tag{7.2.3}$$

Partition

$$V^{-1} = \begin{array}{c} \\ m \\ n \end{array} \begin{array}{c} \begin{array}{cc} m & n \end{array} \\ \begin{pmatrix} U_{11} & U_{12} \\ U_{21} & U_{22} \end{pmatrix} \end{array}. \tag{7.2.4}$$

Assuming that $(U_{11}+U_{12}\varPhi)^{-1}$ exists, we have from (7.2.3)

$$V^{-1}HV\,V^{-1}\begin{pmatrix} I \\ \varPhi \end{pmatrix}(U_{11}+U_{12}\varPhi)^{-1}$$

$$= V^{-1}\begin{pmatrix} I \\ \varPhi \end{pmatrix}(U_{11}+U_{12}\varPhi)^{-1}\left[(U_{11}+U_{12}\varPhi)\,R\,(U_{11}+U_{12}\varPhi)^{-1}\right]. \quad (7.2.5)$$

Since

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

we rewrite (7.2.5) as

$$V^{-1}HV\begin{pmatrix} I \\ \widetilde{\varPhi} \end{pmatrix} = \begin{pmatrix} I \\ \widetilde{\varPhi} \end{pmatrix}\widetilde{R}, \quad (7.2.6)$$

where

$$\widetilde{\varPhi} = (U_{21}+U_{22}\varPhi)\,(U_{11}+U_{12}\varPhi)^{-1}, \quad (7.2.7)$$

$$\widetilde{R} = (U_{11}+U_{12}\varPhi)\,R\,(U_{11}+U_{12}\varPhi)^{-1}. \quad (7.2.8)$$

Lemma 7.2.1 (Wang, Wang and Li)**.** *The first column of $V^{-1}HV$ is* 0*; so is that of $\widetilde{\varPhi}$.*

Proof. We have from (7.2.1) $Ve_1 = \delta^{-1}z$. Thus $V^{-1}HVe_1 = \delta^{-1}V^{-1}Hz = 0$, i.e., the first column of $V^{-1}HV$ is 0. To show $\widetilde{\varPhi}e_1 = 0$, we notice

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

which gives

$$x = \delta\,(U_{11}+U_{12}\varPhi)^{-1}e_1, \quad (U_{21}+U_{22}\varPhi)\,x = 0. \quad (7.2.9)$$

Therefore $\delta\widetilde{\varPhi}e_1 = (U_{21}+U_{22}\varPhi)\,x = 0$ yielding $\widetilde{\varPhi}e_1 = 0$, as claimed. □

Keeping in mind Lemma 7.2.1, we define matrices $\widetilde{A}$, $\widetilde{B}$, $\widetilde{C}$, $\widetilde{D}$, and $\widehat{A}$, $\widehat{B}$, $\widehat{C}$, $\widehat{D}$ by the following partitioning

$$V^{-1}HV = \begin{array}{c} \\ m \\ n \end{array}\!\!\begin{array}{c} \begin{array}{cc} m & n \end{array} \\ \begin{pmatrix} \widetilde{B} & -\widetilde{D} \\ \widetilde{C} & -\widetilde{A} \end{pmatrix} \end{array} = \begin{array}{c} \\ 1 \\ m-1 \\ n \end{array}\!\!\begin{array}{c} \begin{array}{ccc} 1 & m-1 & n \end{array} \\ \begin{pmatrix} 0 & b & -d \\ 0 & \widehat{B} & -\widehat{D} \\ 0 & \widehat{C} & -\widehat{A} \end{pmatrix} \end{array}. \tag{7.2.10}$$

In particular,

$$\widetilde{A} = \widehat{A}, \quad \widetilde{B} = \begin{array}{c} \\ 1 \\ m-1 \end{array}\!\!\begin{array}{c} \begin{array}{cc} 1 & m-1 \end{array} \\ \begin{pmatrix} 0 & b \\ 0 & \widehat{B} \end{pmatrix} \end{array}, \quad \widetilde{C} = \begin{array}{c} \begin{array}{cc} 1 & m-1 \end{array} \\ \begin{pmatrix} 0 & \widehat{C} \end{pmatrix} \end{array}, \quad \widetilde{D} = \begin{array}{c} \\ 1 \\ m-1 \end{array}\!\!\begin{array}{c} n \\ \begin{pmatrix} d \\ \widehat{D} \end{pmatrix} \end{array}. \tag{7.2.11}$$

Equation (7.2.6) says $\widetilde{X} = \widetilde{\varPhi}$ satisfies the following ARE

$$\widetilde{X}\widetilde{D}\widetilde{X} - \widetilde{A}\widetilde{X} - \widetilde{X}\widetilde{B} + \widetilde{C} = 0. \tag{7.2.12}$$

This ARE may have many solutions, and $\widetilde{X} = \widetilde{\varPhi}$ is just one of them. If this particular solution $\widetilde{X} = \widetilde{\varPhi}$ is known, then the minimal nonnegative solution $\varPhi$ of an MARE (1.0.1) can be recovered as follows:

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

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

$$\varPhi = (-U_{22} + \widetilde{\varPhi}U_{12})^{-1}(U_{21} - \widetilde{\varPhi}U_{11}). \tag{7.2.13}$$

While this formula suggests that it needs to do two matrix multiplications and to solve m linear systems of dimension n to recover $\varPhi$ from $\widetilde{\varPhi}$ in general, later we will see for the two realizations in Remark 7.3.2. It actually costs negligibly $O(m+n)$ and $O(mn)$ flops (in comparison to the cost that will be incurred by Algorithm 7.2.1 later for computing $\widetilde{\varPhi}$), respectively.

Lemma 7.2.1 allows us to write

$$\widetilde{\varPhi} = \begin{pmatrix} 0 & \widehat{\varPhi} \end{pmatrix}, \quad \widehat{\varPhi} = \widetilde{\varPhi}_{(:,2:m)}. \tag{7.2.14}$$

In what follows, we look for a determining ARE for $\widehat{\varPhi}$. To this end, we substitute $\widetilde{\varPhi} = \begin{pmatrix} 0 & \widehat{\varPhi} \end{pmatrix}$ and the expressions in (7.2.11) for $\widetilde{A}$, $\widetilde{B}$, $\widetilde{C}$, $\widetilde{D}$ into (7.2.12) to get

$$\begin{aligned} & \begin{pmatrix} 0 & \widehat{\varPhi} \end{pmatrix} \begin{pmatrix} d \\ \widehat{D} \end{pmatrix} \begin{pmatrix} 0 & \widehat{\varPhi} \end{pmatrix} - \widetilde{A} \begin{pmatrix} 0 & \widehat{\varPhi} \end{pmatrix} - \begin{pmatrix} 0 & \widehat{\varPhi} \end{pmatrix} \begin{pmatrix} 0 & b \\ 0 & \widehat{B} \end{pmatrix} + \begin{pmatrix} 0 & \widehat{C} \end{pmatrix} = 0 \\ \Leftrightarrow \quad & \begin{pmatrix} 0 & \widehat{\varPhi}\widehat{D}\widehat{\varPhi} \end{pmatrix} - \begin{pmatrix} 0 & \widetilde{A}\widehat{\varPhi} \end{pmatrix} - \begin{pmatrix} 0 & \widehat{\varPhi}\widehat{B} \end{pmatrix} + \begin{pmatrix} 0 & \widehat{C} \end{pmatrix} = 0 \\ \Leftrightarrow \quad & \widehat{\varPhi}\widehat{D}\widehat{\varPhi} - \widehat{A}\widehat{\varPhi} - \widehat{\varPhi}\widehat{B} + \widehat{C} = 0. \end{aligned}$$

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

$$\widehat{X}\widehat{D}\widehat{X} - \widehat{A}\widehat{X} - \widehat{X}\widehat{B} + \widehat{C} = 0 \tag{7.2.15}$$

which is equivalent to

$$\widehat{H} \begin{pmatrix} I_{m-1} \\ \widehat{X} \end{pmatrix} = \begin{pmatrix} I_{m-1} \\ \widehat{X} \end{pmatrix} (\widehat{B} - \widehat{D}\widehat{X}), \quad \widehat{H} = \begin{array}{c c} & \begin{array}{cc} m-1 & n \end{array} \\ \begin{array}{c} m-1 \\ n \end{array} & \begin{pmatrix} \widehat{B} & -\widehat{D} \\ \widehat{C} & -\widehat{A} \end{pmatrix} \end{array}. \tag{7.2.16}$$

The *complementary algebraic Riccati equation* (cARE) of (7.2.15) is

$$\widehat{Y}\widehat{C}\widehat{Y} - \widehat{Y}\widehat{A} - \widehat{B}\widehat{Y} + \widehat{D} = 0, \tag{7.2.17}$$

or equivalently

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

In the above deflation framework, we assume that both

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

are invertible. Later in Remark 7.3.2, this assumption will be verified for the two realizations of this framework there.

Theorem 7.2.1 (Wang, Wang and Li)**.** *Assume* (7.0.1) *and* (7.1.2)*. Suppose* $U_{11}+U_{12}\varPhi$ *is nonsingular, and define* $\widehat{\varPhi}$ *as in* (7.2.14)*. Then*

$$\operatorname{eig}(\widehat{H})=\{\lambda_1,\cdots,\lambda_{m-1},\lambda_{m+1},\ldots,\lambda_{m+n}\},\tag{7.2.18}$$

$$\operatorname{eig}(\widehat{B}-\widehat{D}\widehat{\varPhi})=\{\lambda_1,\ldots,\lambda_{m-1}\},\tag{7.2.19}$$

and cARE (7.2.17) *has a unique solution* $\widehat{\varPsi}$*, if exists, satisfying*

$$\operatorname{eig}(\widehat{A}-\widehat{C}\widehat{\varPsi})=\{-\lambda_{m+1},\ldots,-\lambda_{m+n}\},\tag{7.2.20}$$

where λ_i *(*$i=1,\ldots,m+n$*) are* H*'s eigenvalues as specified in* Theorem 6.3.1*.*

Proof. Equation (7.2.18) is a consequence of Theorem 6.3.1, the preceding reduction that leads to the definition of $\widehat{H}$ in (7.2.16), and (7.2.10).

We have (7.2.6) – (7.2.8). Since $Rx=(B-D\varPhi)x=0$ by Theorem 6.3.1, using (7.2.9) we find

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

and thus the partitioning

$$\widetilde{R}=\widetilde{B}-\widetilde{D}\widetilde{\varPhi}=\begin{array}{c|cc} & 1 & m-1\\ \hline 1 & 0 & \widetilde{R}_{12}\\ m-1 & 0 & \widetilde{R}_{22}\end{array}\tag{7.2.21}$$

which together with (7.2.11) and $\widetilde{\varPhi} = \begin{pmatrix} 0 & \widehat{\varPhi} \end{pmatrix}$ give $\widetilde{R}_{22} = \widehat{B} - \widehat{D}\widehat{\varPhi}$. Since

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

and $0 = \lambda_m < \mathrm{Re}\lambda_{m-1} \le \cdots \le \mathrm{Re}\lambda_1$ by Theorem 6.3.1, we have (7.2.19).

Let $Z \in \mathbb{R}^{(m+n-1)\times n}$ be a basis matrix of $\widehat{H}$'s invariant subspace associated with the eigenvalues $\lambda_{m+1}, \dots, \lambda_{m+n}$. If $Z_{(m:m+n-1,:)}$ is invertible, then $\widehat{\varPsi}$ exists and is unique, and moreover $\widehat{\varPsi} = Z_{(1:m-1,:)}[Z_{(m:m+n-1,:)}]^{-1}$ and (7.2.20) holds [28]. □

Theorem 7.2.2 (Wang, Wang and Li)**.** *Assume* (7.0.1) *and* (7.1.2). *Suppose both* $U_{11} + U_{12}\varPhi$ *and* $-U_{22} + \widetilde{\varPhi}U_{12}$ *are nonsingular. Then ARE* (7.2.15) *constructed as above has a particular solution* $\widehat{X} = \widehat{\varPhi}$ *characterized uniquely by* (7.2.19)*, and the minimal nonnegative solution* $\varPhi$ *can be recovered by* (7.2.13) *with* $\widetilde{\varPhi} = \begin{pmatrix} 0 & \widehat{\varPhi} \end{pmatrix}$.

Proof. The existence of $\widehat{\varPhi}$ is a consequence of the constructive deflation procedure above, and $\widehat{\varPhi}$ satisfies (7.2.19) by Theorem 7.2.1. That this particular solution $\widehat{X} = \widehat{\varPhi}$ is uniquely characterized by (7.2.19) follows from the relation between the solutions of ARE (7.2.15) and the invariant subspaces of $\widehat{H}$ [28]. □

Theorem 7.2.2 suggests a natural way to compute $\varPhi$ by first solving ARE (7.2.15) for $\widehat{\varPhi}$ by Algorithm 6.4.1 and then recovering $\varPhi$ by (7.2.13). This leads to the following *deflated Alternating-Directional Doubling Algorithm* (dADDA).

Algorithm 7.2.1.

dADDA for an MARE $XDX - AX - XB + C = 0$ **with** (7.0.1)**.**

```
Compute $\mu = u^{\mathrm{T}}x - v^{\mathrm{T}}y$;
If $\mu \ge 0$, then
    compute $\widehat{A}$, $\widehat{B}$, $\widehat{C}$, and $\widehat{D}$ as defined by (7.2.10) and (7.2.11);
    solve (7.2.15) by Algorithm 6.4.1 for $\widehat{\varPhi}$;
    recover $\varPhi$ by (7.2.13) with $\widetilde{\varPhi} = \begin{pmatrix} 0 & \widehat{\varPhi} \end{pmatrix}$;
```

```
6   else
7       compute $\varPhi^{\mathrm{T}}$ instead by working with (7.1.1);
8   Enddo
```

Remark **7.2.1.** There are a few practically important issues to resolve for this dADDA.

1. In building ARE (7.2.15), we need $U_{11}+U_{12}\varPhi$ to be nonsingular, and in recovering $\varPhi$ by (7.2.13), we need $-U_{22}+\widetilde{\varPhi}U_{12}$ to be nonsingular. These requirements are satisfied for each of the realizations in Remark 7.3.2, where we will also investigate the conditioning of both matrices.
2. Both (7.2.19) and (7.2.20) uniquely characterize the particular solution $\widehat{\varPhi}$ of (7.2.15) and the particular solution $\widehat{\varPsi}$, if exists, of (7.2.17), respectively. Specifically, *$\widehat{\varPhi}$ is the unique solution of* (7.2.15) *such that all eigenvalues of $\widehat{B}-\widehat{D}\widehat{\varPhi}$ have positive real parts and $\widehat{\varPsi}$ is the unique solution of* (7.2.17) *such that all eigenvalues of $\widehat{A}-\widehat{C}\widehat{\varPsi}$ have nonpositive real parts.* These characterizations in principle can be used to verify that the computed solution of (7.2.15) at Line 4 of Algorithm 7.2.1 is the right one. But such a verification can only be performed at the end of the iterative process. In the next subsection we will show that with a proper restriction on α and β, this kind of verification becomes unnecessary, i.e., Line 4 of Algorithm 7.2.1 will always produces the right $\widehat{\varPhi}$.
3. What should α and β be for fast convergence at Line 4 of Algorithm 7.2.1?

Remark **7.2.2.** So far, the existence of $\widehat{\varPsi}$ is assumed, not proven. If it exists, it is uniquely characterized by (7.2.20). One way to look into this existence issue, naturally, is to relate $\widehat{\Psi}$ to the minimal nonnegative solution Ψ of the original the cMARE (6.0.1). We shall do it now. Ψ satisfies the cMARE (6.0.1), or equivalently,

$$H\begin{pmatrix}\Psi\\ I\end{pmatrix}=\begin{pmatrix}\Psi\\ I\end{pmatrix}(-S),\quad S=A-C\Psi. \tag{7.2.22}$$

In the same way as we gotten (7.2.6), we can get

$$V^{-1}HV\begin{pmatrix}\widetilde{\Psi}\\ I\end{pmatrix}=\begin{pmatrix}\widetilde{\Psi}\\ I\end{pmatrix}(-\widetilde{S}),\quad \widetilde{S}=\widetilde{A}-\widetilde{C}\widetilde{\Psi}, \tag{7.2.23}$$

where

$$\widetilde{\Psi}=(U_{11}\Psi+U_{12})(U_{21}\Psi+U_{22})^{-1}, \tag{7.2.24}$$

$$\widetilde{S}=(U_{21}\Psi+U_{22})S(U_{21}\Psi+U_{22})^{-1}, \tag{7.2.25}$$

assuming $(U_{21}\Psi+U_{22})^{-1}$ exists. Equation (7.2.23) says $\widetilde{Y}=\widetilde{\Psi}$ satisfies the following ARE

$$\widetilde{Y}\widetilde{C}\widetilde{Y}-\widetilde{Y}\widetilde{A}-\widetilde{B}\widetilde{Y}+\widetilde{D}=0 \tag{7.2.26}$$

which is the complementary ARE of (7.2.12). Partition

$$\widetilde{\Psi}=\begin{array}{c}{\scriptstyle 1}\\ {\scriptstyle m-1}\end{array}\begin{pmatrix}\psi\\ \widehat{\Psi}\end{pmatrix} \tag{7.2.27}$$

and substitute this and (7.2.11) into (7.2.26) to get

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

This says that $\widehat{Y}=\widehat{\Psi}$ is a solution of the complementary ARE (7.2.17) and ψ satisfies $\psi(\widehat{A}-\widehat{C}\widehat{\Psi})=-b\widehat{\Psi}+d$. Thus $\widehat{\Psi}$ 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 Remark 7.3.2. 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., $\widehat{\Psi}$ may still exist even if $U_{21}\Psi+U_{22}$ is singular. For example, $\widehat{\Psi}$ still exists in all the critical case examples in section 8.2 and in [41]. ◇

Lemma 7.2.2 (Wang, Wang and Li)**.** *If $\mu=0$, then $(U_{21}\Psi+U_{22})y=0$ and thus $U_{21}\Psi+U_{22}$ is always singular in the critical case.*

Proof. In the critical case $\mu=0$, $\Psi y=x$ by Theorem 6.3.1. 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 $(U_{21}\Psi+U_{22})y=0$. □

7.3 Convergence Analysis

Assume, as in ADDA for the original MARE (1.0.1), that

$$\alpha \geq \alpha_{\text{opt}} \overset{\text{def}}{=} \max_i A_{(i,i)}, \quad \beta \geq \beta_{\text{opt}} \overset{\text{def}}{=} \max_j B_{(j,j)}. \tag{6.3.6}$$

By Theorem 7.2.1, $\widehat{X} = \widehat{\varPhi} = \widetilde{\varPhi}_{(:,2:m)}$ and $\widehat{\varPsi}$ are such that

$$\widehat{H}\begin{pmatrix} I \\ \widehat{\varPhi} \end{pmatrix} = \begin{pmatrix} I \\ \widehat{\varPhi} \end{pmatrix}\widehat{R}, \qquad \widehat{R} = \widehat{B} - \widehat{D}\widehat{\varPhi}, \quad \operatorname{eig}(\widehat{R}) = \{\lambda_1, \dots, \lambda_{m-1}\}, \tag{7.3.1a}$$

$$\widehat{H}\begin{pmatrix} \widehat{\varPsi} \\ I \end{pmatrix} = \begin{pmatrix} \widehat{\varPsi} \\ I \end{pmatrix}(-\widehat{S}), \quad \widehat{S} = \widehat{A} - \widehat{C}\widehat{\varPsi}, \quad \operatorname{eig}(\widehat{S}) = \{-\lambda_{m+1}, \dots, -\lambda_{m+n}\}. \tag{7.3.1b}$$

Lemma 7.3.1 (Wang, Wang and Li)**.** *Assume* (7.0.1) *and* (6.3.6)*. Let $R = B - D\varPhi$ and $S = A - C\varPsi$, and $\widehat{R}$ and $\widehat{S}$ as given by* (7.3.1)*. Then*

$$\rho(\mathscr{C}(\widehat{S};\alpha,\beta)) = \rho(\mathscr{C}(S;\alpha,\beta)), \quad \rho(\mathscr{C}(\widehat{R};\beta,\alpha)) < \rho(\mathscr{C}(R;\beta,\alpha)) \tag{7.3.2}$$

and in particular

$$\rho(\mathscr{C}(\widehat{S};\alpha,\beta)) \cdot \rho(\mathscr{C}(\widehat{R};\beta,\alpha)) < \rho(\mathscr{C}(S;\alpha,\beta)) \cdot \rho(\mathscr{C}(R;\beta,\alpha)) \leq 1. \tag{7.3.3}$$

Proof. By Theorem 6.3.1(b), both R and S are irreducible M-matrices. Since by (6.3.6)

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

we have $\rho(\mathscr{C}(S;\alpha,\beta)) \cdot \rho(\mathscr{C}(R;\beta,\alpha)) \leq 1$ by analysis in section 6.3. This is the second inequality in (7.3.3). The first inequality is a consequence of (7.3.2) which we now prove. It follows from Theorem 6.3.1(d) and Theorem 7.2.1 that

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

Thus $\rho(\mathscr{C}(\widehat{S};\alpha,\beta)) = \rho(\mathscr{C}(S;\alpha,\beta))$. The proof of [41, Theorem 2.1] implies that

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

where $\lambda_{\min}(R) = 0$ is the eigenvalue of R with the smallest absolute value among all eigenvalues of R. Since $-\mathscr{C}(R;\beta,\alpha) = -(\beta I - R)(\alpha I + R)^{-1} > 0$, by the Perron-Frobenius theorem 3.1.1, we know $\rho(\mathscr{C}(R;\beta,\alpha))$ is a simple eigenvalue with the greatest magnitude among all eigenvalues of $-\mathscr{C}(R;\beta,\alpha)$, i.e., $\rho(\mathscr{C}(R;\beta,\alpha))$ is strictly larger than the absolute value of any other eigenvalue of $-\mathscr{C}(R;\beta,\alpha)$. Since $\lambda_{\min}(R) = 0 \notin \operatorname{eig}(\widehat{R}) \subset \operatorname{eig}(R)$, the eigenvalues of $-\mathscr{C}(\widehat{R};\beta,\alpha)$ are precisely those of $-\mathscr{C}(R;\beta,\alpha)$, except $\rho(\mathscr{C}(R;\beta,\alpha))$. Thus $\rho(\mathscr{C}(R;\beta,\alpha))$ is bigger than the absolute value of any eigenvalue of $-\mathscr{C}(\widehat{R};\beta,\alpha)$. Therefore

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

as was to be shown. □

Theorem 7.3.1 (Wang, Wang and Li)**.** *Assume* (7.0.1) *and* (7.1.2)*. Suppose* $U_{11} + U_{12}\Phi$ *is nonsingular. Let* $\{\widehat{E}_k\}$, $\{\widehat{F}_k\}$, $\{\widehat{X}_k\}$, $\{\widehat{Y}_k\}$ *be the sequences generated by ADDA applied to* (7.2.15) *with no breakdowns, i.e., all involved inverses exist. If* (6.3.6) *holds, then* $\widehat{X}_k$ *and* $\widehat{Y}_k$ *converge quadratically to* $\widehat{\Phi}$ *and* $\widehat{\Psi}$*, respectively, and*

$$\limsup_{k\to\infty} \|\widehat{\Phi} - \widehat{X}_k\|^{1/2^k} \le \rho(\mathscr{C}(\widehat{S};\alpha,\beta)) \cdot \rho(\mathscr{C}(\widehat{R};\beta,\alpha)) < 1, \tag{7.3.4a}$$

$$\limsup_{k\to\infty} \|\widehat{\Psi} - \widehat{Y}_k\|^{1/2^k} \le \rho(\mathscr{C}(\widehat{R};\beta,\alpha)) \cdot \rho(\mathscr{C}(\widehat{S};\alpha,\beta)) < 1, \tag{7.3.4b}$$

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

Proof. Inequalities in (7.3.4) are the consequences of

$$\widehat{\Phi} - \widehat{X}_k = (I - \widehat{X}_k\widehat{\Psi}) \left[\mathscr{C}(\widehat{S};\alpha,\beta)\right]^{2^k} \widehat{\Phi} \left[\mathscr{C}(\widehat{R};\beta,\alpha)\right]^{2^k}, \tag{7.3.5a}$$

$$\widehat{\Psi} - \widehat{Y}_k = (I - \widehat{Y}_k\widehat{\Phi}) \left[\mathscr{C}(\widehat{R};\beta,\alpha)\right]^{2^k} \widehat{\Psi} \left[\mathscr{C}(\widehat{S};\alpha,\beta)\right]^{2^k}. \tag{7.3.5b}$$

Take (7.3.4a) for example. We have by (7.3.5a)

$$(\widehat{\varPhi} - \widehat{X}_k)\left(I - \widehat{\varPsi}\left[\mathscr{C}(\widehat{S};\alpha,\beta)\right]^{2^k}\widehat{\varPhi}\left[\mathscr{C}(\widehat{R};\beta,\alpha)\right]^{2^k}\right)$$
$$= (I - \widehat{\varPhi}\widehat{\varPsi})\left[\mathscr{C}(\widehat{S};\alpha,\beta)\right]^{2^k}\widehat{\varPhi}\left[\mathscr{C}(\widehat{R};\beta,\alpha)\right]^{2^k}. \quad (7.3.6)$$

Since by Lemma 7.3.1

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

$\varGamma \overset{\text{def}}{=} \widehat{\varPsi}\left[\mathscr{C}(\widehat{S};\alpha,\beta)\right]^{2^k}\widehat{\varPhi}\left[\mathscr{C}(\widehat{R};\beta,\alpha)\right]^{2^k} \to 0$ as $k \to \infty$. Therefore for sufficiently large k, $(I-\varGamma)^{-1}$ exists and[1]

$$\|\widehat{\varPhi} - \widehat{X}_k\|^{1/2^k} \le \|(I-\varGamma)^{-1}\|^{1/2^k}\,\|I - \widehat{\varPhi}\widehat{\varPsi}\|^{1/2^k}$$
$$\times\left\|\left[\mathscr{C}(\widehat{S};\alpha,\beta)\right]^{2^k}\right\|^{1/2^k}\|\widehat{\varPhi}\|^{1/2^k}\left\|\left[\mathscr{C}(\widehat{R};\beta,\alpha)\right]^{2^k}\right\|^{1/2^k}. \quad (7.3.7)$$

Letting $k \to \infty$ in both sides of (7.3.7) leads to (7.3.4a) because as $k \to \infty$,

$$\|(I-\varGamma)^{-1}\|^{1/2^k} \to 1, \quad \|I - \widehat{\varPhi}\widehat{\varPsi}\|^{1/2^k} \to 1, \quad \|\widehat{\varPhi}\|^{1/2^k} \to 1,$$
$$\left\|\left[\mathscr{C}(\widehat{S};\alpha,\beta)\right]^{2^k}\right\|^{1/2^k} \to \rho(\mathscr{C}(\widehat{S};\alpha,\beta)), \quad \left\|\left[\mathscr{C}(\widehat{R};\beta,\alpha)\right]^{2^k}\right\|^{1/2^k} \to \rho(\mathscr{C}(\widehat{R};\beta,\alpha)).$$

That $\widehat{X}_k$ and $\widehat{Y}_k$ converge quadratically to $\widehat{\varPhi}$ and $\widehat{\varPsi}$, respectively, is a consequence of the inequalities in (7.3.4). □

Remark **7.3.1.** A few comments are in order:

1. If $\mu \neq 0$, ADDA applied to the original MARE (1.0.1) is already quadratically convergent [41]. But it is only linearly convergent if $\mu = 0$ [13]. Theorem 7.3.1 says that ADDA applied to the deflated ARE (7.2.15) is still quadratically convergent.

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

2. ADDA applied to the original MARE (1.0.1) generates monotonic sequences, under (6.3.6). But this monotonicity property is generally lost in the sequences $\{\widehat{X}_k\}$ and $\{\widehat{Y}_k\}$ generated by ADDA applied to (7.2.15).
3. Theorem 6.3.2 says that under (6.3.6) $\rho(\mathscr{C}(S;\alpha,\beta))\cdot\rho(\mathscr{C}(R;\beta,\alpha))$ is minimized at $\alpha=\alpha_{\rm opt}$ and $\beta=\beta_{\rm opt}$, leading to the optimal ADDA in [41]. For the current case, for fast convergence we should pick α and β such that $\rho(\mathscr{C}(\widehat{S};\alpha,\beta))\cdot\rho(\mathscr{C}(\widehat{R};\beta,\alpha))$ is minimized subject to (6.3.6). While it is not clear whether $\rho(\mathscr{C}(\widehat{S};\alpha,\beta))\cdot\rho(\mathscr{C}(\widehat{R};\beta,\alpha))$ is also minimized at $\alpha=\alpha_{\rm opt}$ and $\beta=\beta_{\rm opt}$, intuitively selecting $\alpha=\alpha_{\rm opt}$ and $\beta=\beta_{\rm opt}$ should be good. This is what we will do in our numerical tests in section 8.2. ◇

Remark **7.3.2.** *Two numerical realizations of the deflating framework given in section 7.2 will be discussed in detail in next two sections. Assume, throughout these two sections,* (7.0.1) *and* (7.1.2).

7.4 By Elimination

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

$$P^{\rm T}=(e_{i_0},e_2,\dots,e_{i_0-1},e_1,e_{i_0+1},\dots,e_{m+n})\in\mathbb{R}^{(m+n)\times(m+n)}, \tag{7.4.1}$$

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 (7.2.1). Set

$$L^{-1}=\begin{pmatrix}1 & \\ -\hat{z} & I_{m+n-1}\end{pmatrix},\qquad L=\begin{pmatrix}1 & \\ \hat{z} & I_{m+n-1}\end{pmatrix}, \tag{7.4.2a}$$

$$V^{-1}=L^{-1}P,\qquad V=P^{\rm T}L, \tag{7.4.2b}$$

where

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

Then $V^{-1}z = z_{(i_0)}e_1$. We just mentioned that Pz serves as a pivoting strategy. We call it a *complete pivoting* if $i_0 = \operatorname{argmax}_i z_{(i)}$, and a *partial pivoting* if $i_0 = \operatorname{argmax}_{1\le i\le m} z_{(i)}$. Simply setting $i_0 = 1$ corresponds to no pivoting. For the complete pivoting, $\|V\|_1\|V^{-1}\|_1 \le (m+n)^2$; 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 $i_0 \le m$, especially so when $i_0 = 1$. In all of our examples in section 8.2 as well as those in the literature, $z = \mathbf{1}_{m+n}$ and thus it makes no difference with or without a pivoting strategy.

We can write

$$P^{\mathrm{T}} = P = I - ww^{\mathrm{T}}, \quad w = e_1 - e_{i_0}. \tag{7.4.3}$$

Partition

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

Use (7.4.2a) and (7.4.3) to see

$$L_{11} = \begin{pmatrix} 1 & \\ -\hat{z}_{(1:m-1)} & I_{m-1} \end{pmatrix}, \quad L_{21} = -\hat{z}_{(m:m+n-1)}e_1^{\mathrm{T}}, \tag{7.4.4a}$$

$$L = \begin{array}{cc} & \begin{array}{cc} m & n \end{array} \\ \begin{array}{c} m \\ n \end{array} & \begin{pmatrix} L_{11}^{-1} & 0 \\ -L_{21} & I \end{pmatrix} \end{array}, \qquad L_{11}^{-1} = \begin{pmatrix} 1 & \\ \hat{z}_{(1:m-1)} & I_{m-1} \end{pmatrix}, \tag{7.4.4b}$$

$$P_{ii} = I - w_iw_i^{\mathrm{T}}, \qquad P_{ij} = -w_iw_j^{\mathrm{T}} \text{ for } i \ne j. \tag{7.4.4c}$$

So the four submatrices U_{ij} of $V^{-1} = L^{-1}P$ partitioned as in (7.2.4) are

$$U_{11} = L_{11}(I - w_1w_1^{\mathrm{T}}), \qquad U_{12} = -L_{11}w_1w_2^{\mathrm{T}}, \tag{7.4.5a}$$

$$U_{21} = L_{21}(I - w_1w_1^{\mathrm{T}}) - w_2w_1^{\mathrm{T}}, \quad U_{22} = -L_{21}w_1w_2^{\mathrm{T}} + I - w_2w_2^{\mathrm{T}}. \tag{7.4.5b}$$

Equations (7.2.7) and (7.2.13) that relate Φ and $\widetilde{\Phi}$ remain valid, provided that $U_{11} + U_{12}\Phi$ and $-U_{22} + \widetilde{\Phi}U_{12}$ are invertible, as ensured by Theorem 7.4.1 below.

Lemma 7.4.1 (Sherman-Morrison-Woodbury). *Let E, $F \in \mathbb{R}^{p\times q}$. The matrix $I_p - EF^{\mathrm{T}}$ is invertible if and only if $I_q - F^{\mathrm{T}}E$ is nonsingular. Moreover*

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

Proof. Suppose matrix $I_p - EF^{\mathrm{T}}$ is invertible, then

$$\begin{array}{cc} & \begin{array}{cc} p & q \end{array} \\ \begin{array}{c} p \\ q \end{array} & \begin{pmatrix} I_p & 0 \\ -F^{\mathrm{T}} & I_q \end{pmatrix} \end{array} \begin{array}{c} \begin{array}{cc} p & q \end{array} \\ \begin{pmatrix} I_p - EF^{\mathrm{T}} & E \\ 0 & I_q \end{pmatrix} \end{array} \begin{array}{c} \begin{array}{cc} p & q \end{array} \\ \begin{pmatrix} I_p & 0 \\ F^{\mathrm{T}} & I \end{pmatrix} \end{array} = \begin{array}{cc} & \begin{array}{cc} p & q \end{array} \\ \begin{array}{c} p \\ q \end{array} & \begin{pmatrix} I_p & 0 \\ -F^{\mathrm{T}} & I_q \end{pmatrix} \end{array} \begin{array}{c} \begin{array}{cc} p & q \end{array} \\ \begin{pmatrix} I_p & E \\ F^{\mathrm{T}} & I_q \end{pmatrix} \end{array}$$

$$= \begin{array}{cc} & \begin{array}{cc} p & q \end{array} \\ \begin{array}{c} p \\ q \end{array} & \begin{pmatrix} I_p & E \\ 0 & I_q - F^{\mathrm{T}}E \end{pmatrix} \end{array}$$

It is easy to see that with non-singularity of $I_p - EF^{\mathrm{T}}$, we also have $I_q - F^{\mathrm{T}}E$ invertible. The proof of the other direction is almost the same. Moreover, we have

$$\begin{aligned} & (I_p + E(I_q - F^{\mathrm{T}}E)^{-1}F^{\mathrm{T}})(I_p - EF^{\mathrm{T}}) \\ =& I_p - EF^{\mathrm{T}} + E(I_q - F^{\mathrm{T}}E)^{-1}F^{\mathrm{T}}(I_p - EF^{\mathrm{T}}) \\ =& I_p - EF^{\mathrm{T}} + E(I_q - F^{\mathrm{T}}E)^{-1}(F^{\mathrm{T}} - F^{\mathrm{T}}EF^{\mathrm{T}}) \\ =& I_p - EF^{\mathrm{T}} + E(I_q - F^{\mathrm{T}}E)^{-1}(I_q - F^{\mathrm{T}}E)F^{\mathrm{T}} \\ =& I_p - EF^{\mathrm{T}} + EF^{\mathrm{T}} \\ =& I_p. \end{aligned}$$

Since $I_p - EF^{\mathrm{T}}$ is square, we have $(I_p - EF^{\mathrm{T}})^{-1} = I_p + E(I_q - F^{\mathrm{T}}E)^{-1}F^{\mathrm{T}}$. □

Theorem 7.4.1 (Wang, Wang and Li). *Let U_{ij} be defined by* (7.4.1) – (7.4.5). *Then both $U_{11} + U_{12}\Phi$ and $-U_{22} + \widetilde{\Phi}U_{12}$ are invertible, where $\widetilde{\Phi}$ relates to Φ by* (7.2.7).

Proof. We have by (7.4.5)

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

Since L_{11} is invertible, $U_{11} + U_{12}\Phi$ is invertible if and only if $I - w_1(w_1^{\mathrm{T}} + w_2^{\mathrm{T}}\Phi)$ is. By Lemma 7.4.1, $I - w_1(w_1^{\mathrm{T}} + w_2^{\mathrm{T}}\Phi)$ is invertible if and only if

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

There are three cases to consider:

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

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

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

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

   since $\Phi > 0$ by Theorem 6.3.1.

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

$$\begin{aligned} (U_{11} + U_{12}\Phi)^{-1} &= \left[I + \zeta^{-1} w_1(w_1^{\mathrm{T}} + w_2^{\mathrm{T}}\Phi)\right] L_{11}^{-1} \\ &= \begin{cases} L_{11}^{-1}, & \text{for } i_0 = 1, \\ \left[I - \ w_1 w_1^{\mathrm{T}}\right] L_{11}^{-1}, & \text{for } 1 < i_0 \leq m, \\ \left[I + \Phi_{(i_0-m,1)}^{-1} e_1(e_1^{\mathrm{T}} - \Phi_{(i_0-m,:)})\right] L_{11}^{-1}, & \text{for } m < i_0. \end{cases} \end{aligned} \tag{7.4.6}$$

Getting to $-U_{22} + \widetilde{\varPhi} U_{12}$, we have

$$-U_{22} + \widetilde{\varPhi} U_{12} = L_{21} w_1 w_2^{\mathrm{T}} - I + w_2 w_2^{\mathrm{T}} - \widetilde{\varPhi} L_{11} w_1 w_2^{\mathrm{T}}, \tag{7.4.7}$$

$$U_{21} + U_{22}\varPhi = L_{21}(I - w_1 w_1^{\mathrm{T}}) - w_2 w_1^{\mathrm{T}} + (-L_{21} w_1 w_2^{\mathrm{T}} + I - w_2 w_2^{\mathrm{T}})\varPhi, \tag{7.4.8}$$

$$\left(U_{11} + U_{12}\varPhi\right)^{-1} L_{11} w_1 w_2^{\mathrm{T}} = \left[I + \zeta^{-1}\, w_1(w_1^{\mathrm{T}} + w_2^{\mathrm{T}}\varPhi)\right] w_1 w_2^{\mathrm{T}}. \tag{7.4.9}$$

Again there are three cases to consider:

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

$$\left(U_{11} + U_{12}\varPhi\right)^{-1} L_{11} w_1 w_2^{\mathrm{T}} = \zeta^{-1} w_1 w_2^{\mathrm{T}}. \tag{7.4.10}$$

Therefore by (7.4.8) and (7.4.10)

$$\begin{aligned}
\widetilde{\varPhi} L_{11} w_1 w_2^{\mathrm{T}} &= \left(U_{21} + U_{22}\varPhi\right)\left(U_{11} + U_{12}\varPhi\right)^{-1} L_{11} w_1 w_2^{\mathrm{T}} \\
&= \left[L_{21}(I - w_1 w_1^{\mathrm{T}}) - w_2 w_1^{\mathrm{T}} + (-L_{21} w_1 w_2^{\mathrm{T}} + I - w_2 w_2^{\mathrm{T}})\varPhi\right] \zeta^{-1} w_1 w_2^{\mathrm{T}} \\
&= \zeta^{-1}\left[-w_2 w_2^{\mathrm{T}} + \zeta L_{21} w_1 w_2^{\mathrm{T}} + \varPhi w_1 w_2^{\mathrm{T}} + \zeta w_2 w_2^{\mathrm{T}}\right] \\
&= (1 - \zeta^{-1}) w_2 w_2^{\mathrm{T}} + L_{21} w_1 w_2^{\mathrm{T}} + \zeta^{-1}\varPhi w_1 w_2^{\mathrm{T}}.
\end{aligned}$$

Combine this with (7.4.7) to get

$$\begin{aligned}
-U_{22} + \widetilde{\varPhi} U_{12} &= -I + \zeta^{-1} w_2 w_2^{\mathrm{T}} - \zeta^{-1}\varPhi w_1 w_2^{\mathrm{T}} \\
&= -\left[I - \zeta^{-1}(w_2 - \varPhi w_1) w_2^{\mathrm{T}}\right] \qquad (7.4.11)
\end{aligned}$$

which, by Lemma 7.4.1, is invertible if

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

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

$$\left(-U_{22}+\widetilde{\varPhi}U_{12}\right)^{-1}=\begin{cases}-I, & \text{for } i_0\le m,\\ -\left[I-(e_{i_0-m}+\varPhi_{(:,1)})e_{i_0-m}^{\mathrm{T}}\right], & \text{for } i_0>m.\end{cases}\tag{7.4.12}$$

This completes the proof. □

The inversion formulas (7.4.6) and (7.4.12), together with (7.4.4) and (7.4.5), lead to fast algorithms via (7.2.7) and (7.2.13) to go from one of $\varPhi$ and $\widetilde{\varPhi}$ to the other at the cost of $O(m+n)$ flops. The numerical stability of going from $\widetilde{\varPhi}$ to $\varPhi$ this way depends on $\|U_{11}+U_{12}\varPhi\|_1\|(U_{11}+U_{12}\varPhi)^{-1}\|_1$ and $\|-U_{22}+\widetilde{\varPhi}U_{12}\|_1\|(-U_{22}+\widetilde{\varPhi}U_{12}^{-1}\|_1$ for which we have, provided $|\hat{z}_{(i)}|\le 1$ for $1\le i\le m+n-1$,

$$\|U_{11}+U_{12}\varPhi\|_1\le\begin{cases}(m+1), & \text{if } i_0\le m,\\ (m+1)\left(1+\max\limits_{1\le j\le m}\varPhi_{(i_0-m,j)}\right), & \text{if } i_0>m,\end{cases}\tag{7.4.13a}$$

$$\|(U_{11}+U_{12}\varPhi)^{-1}\|_1\le\begin{cases}(m+1), & \text{if } i_0\le m,\\ (m+1)\left(\varPhi_{(i_0-m,1)}^{-1}+\max\limits_{2\le j\le m}\frac{\varPhi_{(i_0-m,j)}}{\varPhi_{(i_0-m,1)}}\right), & \text{if } i_0>m,\end{cases}\tag{7.4.13b}$$

and

$$\|-U_{22}+\widetilde{\varPhi}U_{12}\|_1\begin{cases}=1, & \text{if } i_0\le m,\\ \le 1+\varPhi_{(i_0-m,1)}^{-1}(1+\|\varPhi_{(1,:)}\|_1), & \text{if } i_0>m,\end{cases}\tag{7.4.14a}$$

$$\|(-U_{22}+\widetilde{\varPhi}U_{12})^{-1}\|_1\begin{cases}=1, & \text{if } i_0\le m,\\ \le 1+\|\varPhi_{(1,:)}\|_1, & \text{if } i_0>m.\end{cases}\tag{7.4.14b}$$

In particular, if $i_0\le m$, all bounds by (7.4.13) and (7.4.14) are independent of $\varPhi$ and $\widetilde{\varPhi}$, and thus calculating $\varPhi$ or $\widetilde{\varPhi}$ via (7.2.7) or (7.2.13) is numerically stable.

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

$$\begin{aligned} V^{-1}HV &= (I - \tilde{z}e_1^{\mathrm{T}})PHP(I + \tilde{z}e_1^{\mathrm{T}}) \\ &= PHP - \tilde{z}(e_1^{\mathrm{T}}PHP) + (PHP\tilde{z})e_1^{\mathrm{T}} - (e_1^{\mathrm{T}}PHP\tilde{z})\tilde{z}e_1^{\mathrm{T}}. \qquad (7.4.15) \end{aligned}$$

where $\tilde{z} = (0, \hat{z}^{\mathrm{T}})^{\mathrm{T}}$. The right-hand side of (7.4.15) lends itself for a fast evaluation of $V^{-1}HV$. In the case $i_0 = 1$, we have[2]

$$\widetilde{\varPhi} = \begin{pmatrix} 0 & \varPhi_{(:,2:m)} \end{pmatrix}, \quad \widehat{\varPhi} = \varPhi_{(:,2:m)}, \quad \varPhi_{(:,1)} = x_{(1)}^{-1}\left[y - \varPhi_{(:,2:m)}x_{(2:m)}\right], \qquad (7.4.16)$$

and

$$\begin{aligned} \widehat{B} &= B_{(2:m,2:m)} - x_{(1)}^{-1}x_{(2:m)}B_{(1,2:m)}, & \widehat{D} &= D_{(2:m,:)} - x_{(1)}^{-1}x_{(2:m)}D_{(1,:)}, \qquad (7.4.17\text{a}) \\ \widehat{C} &= C_{(:,2:m)} - x_{(1)}^{-1}yB_{(1,2:m)}, & \widehat{A} &= A - x_{(1)}^{-1}yD_{(1,:)}. \qquad (7.4.17\text{b}) \end{aligned}$$

Note also in this case

$$\widehat{A} - \widehat{\varPhi}\widehat{D} = A - \varPhi D. \qquad (7.4.18)$$

This is because $\varPhi_{(:,1)} = x_{(1)}^{-1}\left[y - \widehat{\varPhi}x_{(2:m)}\right]$, and thus

$$\begin{aligned} \widehat{A} - \widehat{\varPhi}\widehat{D} &= A - x_{(1)}^{-1}yD_{(1,:)} - \widehat{\varPhi}(D_{(2:m,:)} - x_{(1)}^{-1}x_{(2:m)}D_{(1,:)}) \\ &= A - x_{(1)}^{-1}\left[y - \widehat{\varPhi}x_{(2:m)}\right]D_{(1,:)} - \widehat{\varPhi}D_{(2:m,:)} \\ &= A - \varPhi_{(:,1)}D_{(1,:)} - \widehat{\varPhi}D_{(2:m,:)} \\ &= A - \varPhi D. \end{aligned}$$

In Remark 7.2.2, we show $\widehat{\varPsi}$ exists if $U_{21}\varPsi + U_{22}$ is nonsingular, and in Lemma 7.2.2 we show $U_{21}\varPsi + U_{22}$ is always singular if $\mu = 0$. Theorem 7.4.2 asserts that $U_{21}\varPsi + U_{22}$ is guaranteed nonsingular if $\mu \neq 0$. Thus the existence of $\widehat{\varPsi}$ is unresolved for the case $\mu = 0$, but otherwise $\widehat{\varPsi}$ exists. We point out that $\widehat{\varPsi}$ does exist for all our critical case examples in section 8.2 though.

[2] This is not a misprint: the last $m-1$ columns of $\widetilde{\varPhi}$ are the same as those of $\varPhi$.

Theorem 7.4.2 (Wang, Wang and Li)**.** *Let U_{ij} be defined by* (7.4.1) – (7.4.5)*. Then $U_{21}\varPsi + U_{22}$ is singular when and only when $\mu = 0$.*

Proof. We already know that $U_{21}\varPsi + U_{22}$ is singular when $\mu = 0$ by Lemma 7.2.2. But the conclusion of the theorem is stronger than this. The proof below uses the explicit expressions for U_{ij} given in (7.4.5) which gives

$$U_{21}\varPsi + U_{22} = \left[L_{21}(I - w_1w_1^{\mathrm{T}}) - w_2w_1^{\mathrm{T}}\right]\varPsi - L_{21}w_1w_2^{\mathrm{T}} + I - w_2w_2^{\mathrm{T}}. \qquad (7.4.19)$$

There are three cases to consider.

1. If $i_0 = 1$, then $w_1 = 0$ and $w_2 = 0$ and thus (7.4.19) becomes

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

which is nonsingular if and only if $1 - x_{(1)}^{-1}\, e_1^{\mathrm{T}}\varPsi y \neq 0$. Now for $\mu > 0$, $\varPsi y < x$ by Theorem 6.3.1 and then $x_{(1)}^{-1}\, e_1^{\mathrm{T}}\varPsi y < x_{(1)}^{-1}\, e_1^{\mathrm{T}}x = 1$ implying $1 - x_{(1)}^{-1}\, e_1^{\mathrm{T}}\varPsi y > 0$. But for $\mu = 0$, $\varPsi y = x$ by Theorem 6.3.1 and then $x_{(1)}^{-1}\, e_1^{\mathrm{T}}\varPsi y = x_{(1)}^{-1}\, e_1^{\mathrm{T}}x = 1$ implying $1 - x_{(1)}^{-1}\, e_1^{\mathrm{T}}\varPsi y = 0$.

2. If $1 < i_0 \leq m$, then $w_1 = e_1 - e_{i_0}$ and $w_2 = 0$. Write $P_1 = I - w_1w_1^{\mathrm{T}}$ which is the permutation matrix that swaps the first entry and the i_0th entry of x. (7.4.19) becomes

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

which is nonsingular if and only if $1 - x_{(i_0)}^{-1}\, e_1^{\mathrm{T}}P_1\varPsi y \neq 0$. Now for $\mu > 0$, $\varPsi y < x$ by Theorem 6.3.1 and then $x_{(i_0)}^{-1}\, e_1^{\mathrm{T}}P_1\varPsi y < x_{(i_0)}^{-1}\, e_1^{\mathrm{T}}P_1x = 1$ implying $1 - x_{(i_0)}^{-1}\, e_1^{\mathrm{T}}P_1\varPsi y > 0$. But for $\mu = 0$, $\varPsi y = x$ by Theorem 6.3.1 and then $x_{(i_0)}^{-1}\, e_1^{\mathrm{T}}P_1\varPsi y = x_{(i_0)}^{-1}\, e_1^{\mathrm{T}}P_1x = 1$ implying $1 - x_{(i_0)}^{-1}\, e_1^{\mathrm{T}}P_1\varPsi y = 0$.

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

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

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

$$\begin{aligned} U_{21}\Psi + U_{22} &= -w_2 w_1^{\mathrm{T}}\Psi - L_{21} w_1 w_2^{\mathrm{T}} + I - w_2 w_2^{\mathrm{T}} \\ &= e_{j_0} e_1^{\mathrm{T}} \Psi - \hat{y} e_{j_0}^{\mathrm{T}} + I - e_{j_0} e_{j_0}^{\mathrm{T}} \\ &= I - (\hat{y} + e_{j_0}) e_{j_0}^{\mathrm{T}} + e_{j_0} e_1^{\mathrm{T}} \Psi \\ &= I - \begin{pmatrix} \hat{y} + e_{j_0} & -e_{j_0} \end{pmatrix} \begin{pmatrix} e_{j_0}^{\mathrm{T}} \\ e_1^{\mathrm{T}} \Psi \end{pmatrix} \end{aligned}$$

which, by Lemma 7.4.1, is invertible if and only if

$$I_2 - \begin{pmatrix} e_{j_0}^{\mathrm{T}} \\ e_1^{\mathrm{T}} \Psi \end{pmatrix} \begin{pmatrix} \hat{y} + e_{j_0} & -e_{j_0} \end{pmatrix} \tag{7.4.20}$$

is invertible. Use $\hat{y} + e_{j_0} = y_{(j_0)}^{-1}\, y + y_{(j_0)}^{-1} x_{(1)}\, e_{j_0}$ to simplify the matrix (7.4.20) to

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

whose determinant is $y_{(j_0)}^{-1} \left[e_1^{\mathrm{T}} \Psi y - x_{(1)} \right]$. Now if $\mu > 0$, then $\Psi y < x$ by Theorem 6.3.1 and thus $y_{(j_0)}^{-1} \left[e_1^{\mathrm{T}} \Psi y - x_{(1)} \right] < 0$. If $\mu = 0$, then $\Psi y = x$ by Theorem 6.3.1 and thus $y_{(j_0)}^{-1} \left[e_1^{\mathrm{T}} \Psi y - x_{(1)} \right] = 0$ implying $U_{21}\Psi + U_{22}$ is singular.

This completes the proof. □

7.5 By Orthogonal Transformation

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

$$Q = \begin{array}{c c} & \begin{array}{cc} m & n \end{array} \\ \begin{array}{c} m \\ n \end{array} & \begin{pmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{pmatrix} \end{array}. \tag{7.5.1}$$

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

$$\widetilde{\Phi} = \left(Q_{12}^{\mathrm{T}} + Q_{22}^{\mathrm{T}}\Phi\right)\left(Q_{11}^{\mathrm{T}} + Q_{21}^{\mathrm{T}}\Phi\right)^{-1}, \tag{7.5.2a}$$

$$\Phi = (-Q_{22}^{\mathrm{T}} + \widetilde{\Phi}Q_{21}^{\mathrm{T}})^{-1}(Q_{12}^{\mathrm{T}} - \widetilde{\Phi}Q_{11}^{\mathrm{T}}), \tag{7.5.2b}$$

assuming $Q_{11}^{\mathrm{T}} + Q_{21}^{\mathrm{T}}\Phi$ and $-Q_{22}^{\mathrm{T}} + \widetilde{\Phi}Q_{21}^{\mathrm{T}}$ are invertible. We know $\widetilde{\Phi}e_1 = 0$ by Lemma 7.2.1, and $\widehat{\Phi} = \widetilde{\Phi}_{(:,2:m)}$ satisfies ARE (7.2.15).

Possible candidates for Q include a product of $m + n - 1$ Givens rotations or a Householder transformation [16]. 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 $Q_{11}^{\mathrm{T}} + Q_{21}^{\mathrm{T}}\Phi$ and $-Q_{22}^{\mathrm{T}} + \widetilde{\Phi}Q_{21}^{\mathrm{T}}$ are guaranteed invertible[3] by Theorem 7.5.1 below.

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

$$Q = I - 2ww^{\mathrm{T}}, \quad w = \frac{z - \delta e_1}{\|z - \delta e_1\|_2} = \frac{z - \delta e_1}{\gamma}, \tag{7.5.3}$$

where

$$\delta = -\|z\|_2, \quad \gamma = \|z - \delta e_1\|_2 = \sqrt{2x_{(1)}\|z\|_2 + 2\|z\|_2^2}. \tag{7.5.4}$$

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

$$0 < w_1 = \gamma^{-1}(x - \delta e_1) \in \mathbb{R}^m, \quad 0 < w_2 = \gamma^{-1}y \in \mathbb{R}^n. \tag{7.5.5}$$

[3] This is not so for the Householder transformation such that $Qz = \|z\|_2\, e_1$. For example, $m = n = 2$, $B = \begin{pmatrix} 3 & -1 \\ -1 & 3 \end{pmatrix}$, $D = \mathbf{1}_{2,2}$, $A = B$, and $C = D$. For this example $W\mathbf{1}_4 = 0$, $\mathbf{1}_4^{\mathrm{T}}W = 0$, $\Phi = \frac{1}{2}\mathbf{1}_{2,2}$, $\Psi = \frac{1}{2}\mathbf{1}_{2,2}$, and thus $\mu = 0$. We have $Q_{11}^{\mathrm{T}} + Q_{21}^{\mathrm{T}}\Phi = \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix}$ for the Householder transformation Q such that $Q\mathbf{1}_4 = 2e_1$, but $Q_{11}^{\mathrm{T}} + Q_{21}^{\mathrm{T}}\Phi = \begin{pmatrix} -1 & -1 \\ -2/3 & 2/3 \end{pmatrix}$ for the Householder transformation Q such that $Q\mathbf{1}_4 = -2e_1$.

Then the four submatrices Q_{ij} as defined by (7.5.1) are

$$Q_{11} = I_m - 2w_1 w_1^{\mathrm{T}}, \qquad Q_{12} = -2w_1 w_2^{\mathrm{T}}, \tag{7.5.6a}$$

$$Q_{22} = I_n - 2w_2 w_2^{\mathrm{T}}, \qquad Q_{21} = -2w_2 w_1^{\mathrm{T}}. \tag{7.5.6b}$$

Theorem 7.5.1 (Wang, Wang and Li)**.** *Let* $Q \in \mathbb{R}^{(m+n)\times(m+n)}$ *be the Householder transformation as given by* (7.5.3) *and* (7.5.4)*. Then both* $Q_{11}^{\mathrm{T}}+Q_{21}^{\mathrm{T}}\varPhi$ *and* $-Q_{22}^{\mathrm{T}}+\widetilde{\varPhi}Q_{21}^{\mathrm{T}}$ *are invertible, where* $\widetilde{\varPhi}$ *relates to* $\varPhi$ *by* (7.5.2a)*.*

Proof. We have (7.5.3) – (7.5.6), and thus

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

By Lemma 7.4.1, $Q_{11}^{\mathrm{T}} + Q_{21}^{\mathrm{T}}\varPhi$ is invertible if and only if $1-2(w_1^{\mathrm{T}}+w_2^{\mathrm{T}}\varPhi)w_1 \neq 0$ which we will verify. We have

$$\begin{aligned}
\zeta &\overset{\text{def}}{=} 1 - 2(w_1^{\mathrm{T}} + w_2^{\mathrm{T}}\varPhi)w_1 && (7.5.7)\\
&= 1 - 2w_1^{\mathrm{T}}w_1 - 2w_2^{\mathrm{T}}\varPhi w_1\\
&= 1 - 2\,\frac{\|x-\delta e_1\|_2^2}{\gamma^2} - 2w_2^{\mathrm{T}}\varPhi w_1\\
&= -\frac{x_{(1)}\|z\|_2 + \|x\|_2^2}{x_{(1)}\|z\|_2 + \|z\|_2^2} - 2w_2^{\mathrm{T}}\varPhi w_1 < 0 && (7.5.8)
\end{aligned}$$

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

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

Next we have

$$-Q_{22}^{\mathrm{T}} + \widetilde{\varPhi}Q_{21}^{\mathrm{T}} = -I + 2w_2w_2^{\mathrm{T}} - 2\widetilde{\varPhi}w_1w_2^{\mathrm{T}} = -\left[I - 2(w_2 - \widetilde{\varPhi}w_1)w_2^{\mathrm{T}}\right]$$

which is invertible if and only if

$$1 - 2w_2^{\mathrm{T}}(w_2 - \widetilde{\varPhi}w_1) = 1 - 2(w_2^{\mathrm{T}}w_2 - w_2^{\mathrm{T}}\widetilde{\varPhi}w_1) \neq 0$$

which we will now verify. We have

$$\begin{aligned}
w_2^{\mathrm{T}}\left(Q_{12}^{\mathrm{T}}+Q_{22}^{\mathrm{T}}\mathit{\Phi}\right) &= w_2^{\mathrm{T}}\left[-2w_2w_1^{\mathrm{T}}+(I-2w_2w_2^{\mathrm{T}})\mathit{\Phi}\right] \\
&= (-2w_2^{\mathrm{T}}w_2)w_1^{\mathrm{T}}+(1-2w_2^{\mathrm{T}}w_2)w_2^{\mathrm{T}}\mathit{\Phi}, \\
\left(Q_{11}^{\mathrm{T}}+Q_{21}^{\mathrm{T}}\mathit{\Phi}\right)^{-1}w_1 &= \left[1+\frac{2w_1^{\mathrm{T}}w_1+2w_2^{\mathrm{T}}\mathit{\Phi}w_1}{1-2w_1^{\mathrm{T}}w_1-2w_2^{\mathrm{T}}\mathit{\Phi}w_1}\right]w_1 \\
&= \frac{1}{1-2w_1^{\mathrm{T}}w_1-2w_2^{\mathrm{T}}\mathit{\Phi}w_1}\,w_1, \\
w_2^{\mathrm{T}}\widetilde{\mathit{\Phi}}w_1 &= w_2^{\mathrm{T}}\left(Q_{12}^{\mathrm{T}}+Q_{22}^{\mathrm{T}}\mathit{\Phi}\right)\cdot\left(Q_{11}^{\mathrm{T}}+Q_{21}^{\mathrm{T}}\mathit{\Phi}\right)^{-1}w_1 \\
&= \frac{(-2w_2^{\mathrm{T}}w_2)w_1^{\mathrm{T}}w_1+(1-2w_2^{\mathrm{T}}w_2)w_2^{\mathrm{T}}\mathit{\Phi}w_1}{1-2w_1^{\mathrm{T}}w_1-2w_2^{\mathrm{T}}\mathit{\Phi}w_1}, \\
w_2^{\mathrm{T}}w_2-w_2^{\mathrm{T}}\widetilde{\mathit{\Phi}}w_1 &= \frac{w_2^{\mathrm{T}}w_2-w_2^{\mathrm{T}}\mathit{\Phi}w_1}{1-2w_1^{\mathrm{T}}w_1-2w_2^{\mathrm{T}}\mathit{\Phi}w_1}, \\
1-2(w_2^{\mathrm{T}}w_2-w_2^{\mathrm{T}}\widetilde{\mathit{\Phi}}w_1) &= \frac{1-2w_1^{\mathrm{T}}w_1-2w_2^{\mathrm{T}}w_2}{1-2w_1^{\mathrm{T}}w_1-2w_2^{\mathrm{T}}\mathit{\Phi}w_1} \\
&= -\frac{1}{1-2w_1^{\mathrm{T}}w_1-2w_2^{\mathrm{T}}\mathit{\Phi}w_1}>0,
\end{aligned}$$

as expected. □

Remark **7.5.1.** Theorem 7.5.1 is proved under the inherited conditions $x>0$, $y>0$, $\mathit{\Phi}>0$, and $\mathit{\Phi}x=y$. Carefully examining the proof, one finds that the condition of the theorem can be relaxed to

$$x\geq 0,\quad x\neq 0,\quad y\geq 0,\quad \mathit{\Phi}\geq 0,$$

and $\widetilde{\mathit{\Phi}}$ relates to $\mathit{\Phi}$ by (7.5.2a). Since $\mathit{\Phi}x=y$ is never referenced, it is not required. ◇

The above proof also yields

$$\left(Q_{11}^{\mathrm{T}}+Q_{21}^{\mathrm{T}}\mathit{\Phi}\right)^{-1} = I_m+2\zeta^{-1}w_1(w_1^{\mathrm{T}}+w_2^{\mathrm{T}}\mathit{\Phi}), \tag{7.5.9a}$$

$$\left(-Q_{22}^{\mathrm{T}}+\widetilde{\mathit{\Phi}}Q_{21}^{\mathrm{T}}\right)^{-1} = -\left[I_n-2\zeta(w_2-\widetilde{\mathit{\Phi}}w_1)w_2^{\mathrm{T}}\right], \tag{7.5.9b}$$

where ζ is defined by (7.5.7). With the help of (7.5.9), we can express any one of $\varPhi$ and $\widetilde{\varPhi}$ in terms of the other via a rank-one update. Details are as follows. By (7.5.2), we have

$$\begin{aligned}
\widetilde{\varPhi} &= \left[-2w_2w_1^{\mathrm T} + (I - 2w_2w_2^{\mathrm T})\varPhi\right]\left[I + 2\zeta^{-1}w_1(w_1^{\mathrm T} + w_2^{\mathrm T}\varPhi)\right] \\
&= \left[\varPhi - 2w_2(w_1^{\mathrm T} + w_2^{\mathrm T}\varPhi)\right]\left[I + 2\zeta^{-1}w_1(w_1^{\mathrm T} + w_2^{\mathrm T}\varPhi)\right] \\
&= \varPhi + 2\zeta^{-1}\varPhi w_1(w_1^{\mathrm T} + w_2^{\mathrm T}\varPhi) \\
&\quad - 2w_2(w_1^{\mathrm T} + w_2^{\mathrm T}\varPhi) - 4\zeta^{-1}w_2\underbrace{(w_1^{\mathrm T} + w_2^{\mathrm T}\varPhi)w_1}_{\text{scalar}}(w_1^{\mathrm T} + w_2^{\mathrm T}\varPhi) \\
&= \varPhi + 2\zeta^{-1}\varPhi w_1(w_1^{\mathrm T} + w_2^{\mathrm T}\varPhi) - 2\left[1 + 2\zeta^{-1}(w_1^{\mathrm T} + w_2^{\mathrm T}\varPhi)w_1\right]w_2(w_1^{\mathrm T} + w_2^{\mathrm T}\varPhi) \\
&= \varPhi + 2\left\{\zeta^{-1}\varPhi w_1 - \left[1 + 2\zeta^{-1}(w_1^{\mathrm T} + w_2^{\mathrm T}\varPhi)w_1\right]w_2\right\}(w_1^{\mathrm T} + w_2^{\mathrm T}\varPhi), \qquad \text{(7.5.10a)} \\
\varPhi &= \left[-I_n + 2\zeta(w_2 - \widetilde{\varPhi}w_1)w_2^{\mathrm T}\right]\left[-2w_2w_1^{\mathrm T} - \widetilde{\varPhi}(I - 2w_1w_1^{\mathrm T})\right] \\
&= \left[-I_n + 2\zeta(w_2 - \widetilde{\varPhi}w_1)w_2^{\mathrm T}\right]\left[-\widetilde{\varPhi} - 2(w_2 - \widetilde{\varPhi}w_1)w_1^{\mathrm T}\right] \\
&= \widetilde{\varPhi} + 2(w_2 - \widetilde{\varPhi}w_1)w_1^{\mathrm T} \\
&\quad - 2\zeta(w_2 - \widetilde{\varPhi}w_1)w_2^{\mathrm T}\widetilde{\varPhi} - 4\zeta(w_2 - \widetilde{\varPhi}w_1)\underbrace{w_2^{\mathrm T}(w_2 - \widetilde{\varPhi}w_1)}_{\text{scalar}}w_1^{\mathrm T} \\
&= \widetilde{\varPhi} + 2\left[1 - 2\zeta w_2^{\mathrm T}(w_2 - \widetilde{\varPhi}w_1)\right](w_2 - \widetilde{\varPhi}w_1)w_1^{\mathrm T} - 2\zeta(w_2 - \widetilde{\varPhi}w_1)w_2^{\mathrm T}\widetilde{\varPhi} \\
&= \widetilde{\varPhi} + 2(w_2 - \widetilde{\varPhi}w_1)\left\{\left[1 - 2\zeta w_2^{\mathrm T}(w_2 - \widetilde{\varPhi}w_1)\right]w_1^{\mathrm T} - \zeta w_2^{\mathrm T}\widetilde{\varPhi}\right\}. \qquad \text{(7.5.10b)}
\end{aligned}$$

Equation (7.5.10b) will become handy in coding up Algorithm 7.2.1, where recovering $\varPhi$ is needed from computed $\widehat{\varPhi}$ by (7.5.10b) with $\widetilde{\varPhi} = \begin{pmatrix} 0 & \widehat{\varPhi} \end{pmatrix}$. Equation (7.5.10a) expresses $\widetilde{\varPhi}$ in terms of $\varPhi$. The cost of getting one of $\varPhi$ and $\widetilde{\varPhi}$ from the other is only $O(mn)$ flops. The numerical stability of doing so depends on $\|Q_{11}^{\mathrm T} + Q_{21}^{\mathrm T}\varPhi\|_2\|(Q_{11}^{\mathrm T} +$

$Q_{21}^{\mathrm{T}}\Phi)^{-1}\|_2$ and $\| -Q_{22}^{\mathrm{T}}+\widetilde{\Phi}Q_{21}^{\mathrm{T}}\|_2\|(-Q_{22}^{\mathrm{T}}+\widetilde{\Phi}Q_{21}^{\mathrm{T}})^{-1}\|_2$ for which we have, upon using $\|w_i\|_2 \le 1$ for $i = 1, 2$,

$$\|Q_{11}^{\mathrm{T}}+Q_{21}^{\mathrm{T}}\Phi\|_2 \le 1+2(1+\|\Phi\|_2), \tag{7.5.11a}$$

$$\|(Q_{11}^{\mathrm{T}}+Q_{21}^{\mathrm{T}}\Phi)^{-1}\|_2 \le 1+2|\zeta^{-1}|(1+\|\Phi\|_2), \tag{7.5.11b}$$

$$\| -Q_{22}^{\mathrm{T}}+\widetilde{\Phi}Q_{21}^{\mathrm{T}}\|_2 \le 1+2(1+\|\widetilde{\Phi}\|_2), \tag{7.5.11c}$$

$$\|(-Q_{22}^{\mathrm{T}}+\widetilde{\Phi}Q_{21}^{\mathrm{T}})^{-1}\|_2 \le 1+2|\zeta|(1+\|\widetilde{\Phi}\|_2). \tag{7.5.11d}$$

Lower and upper bound on $|\zeta|$ can be easily gotten from (7.5.8), for example

$$\|x\|_2^2/\|z\|_2^2 \le |\zeta| \le 1+2\|\Phi\|_2.$$

Thus calculating Φ or $\widetilde{\Phi}$ via (7.5.2) is numerically stable unless $\|x\|_2^2 \ll \|z\|_2^2$.

Extractions of the coefficient matrices $\widehat{A}$, $\widehat{B}$, $\widehat{C}$, and $\widehat{D}$ for ARE (7.2.15) can be easily done from the partitioning (7.2.10) for

$$\begin{aligned} V^{-1}HV &= (I-2ww^{\mathrm{T}})H(I-2ww^{\mathrm{T}}) \\ &= H-2ww^{\mathrm{T}}H-2Hww^{\mathrm{T}}+4(w^{\mathrm{T}}Hw)ww^{\mathrm{T}}, \end{aligned} \tag{7.5.12}$$

where the expression in the right-hand side of (7.5.12) suggests an economical way to numerically compute $V^{-1}HV$.

In Remark 7.2.2, we show $\widehat{\Psi}$ exists if $U_{21}\Psi+U_{22}$ is nonsingular, and in Lemma 7.2.2 we show $U_{21}\Psi+U_{22}$ is always singular if $\mu = 0$. Theorem 7.5.2 asserts that $U_{21}\Psi+U_{22}$ is guaranteed nonsingular if $\mu \ne 0$. Thus the existence of $\widehat{\Psi}$ is unresolved for the case $\mu = 0$, but otherwise $\widehat{\Psi}$ exists. We point out that $\widehat{\Psi}$ does exist for all our critical case examples in section 8.2 though.

Theorem 7.5.2 (Wang, Wang and Li)**.** *Let $Q \in \mathbb{R}^{(m+n)\times(m+n)}$ be the Householder transformation as given by* (7.5.3) *and* (7.5.4)*. $U_{21}\Psi+U_{22}$ is singular when and only when $\mu = 0$.*

Proof. We have by (7.5.6) and $U_{ij} = Q_{ji}^{\mathrm{T}}$ that

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

which is invertible if and only if $1 - 2(w_2^{\mathrm{T}} + w_1^{\mathrm{T}}\varPsi)w_2 \neq 0$ which we now verify. Recall (7.5.4) and (7.5.5) and that $\varPsi y < x$ for $\mu > 0$ and $\varPsi y = x$ for $\mu = 0$. We have

$$\begin{aligned} 2(w_2^{\mathrm{T}} + w_1^{\mathrm{T}}\varPsi)w_2 &= \frac{2y^{\mathrm{T}}y + 2(x + \|z\|_2e_1)^{\mathrm{T}}\varPsi y}{\gamma^2} \\ &\leq \frac{y^{\mathrm{T}}y + (x + \|z\|_2e_1)^{\mathrm{T}}x}{x_{(1)}\|z\|_2 + \|z\|_2^2} \\ &= 1, \end{aligned}$$

where the equality occurs when and only when $\mu = 0$. Therefore $1-2(w_2^{\mathrm{T}}+w_1^{\mathrm{T}}\varPsi)w_2 \geq 0$ with equality when and only when $\mu = 0$. □

7.6 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 [22] 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 an MARE (1.0.1) with (7.0.1) and $\mu = u^{\mathrm{T}}x - v^{\mathrm{T}}y \geq 0$. Pick $\eta \in \mathbb{R}$ to be specified in a moment, and let

$$\widehat{H} = H + \eta z w^{\mathrm{T}} \equiv \begin{array}{c@{}c} & \begin{array}{cc} m & n \end{array} \\ \begin{array}{c} m \\ n \end{array} & \begin{pmatrix} \widehat{B} & -\widehat{D} \\ \widehat{C} & -\widehat{A} \end{pmatrix} \end{array}, \tag{7.6.1}$$

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

$$\widehat{X}\widehat{D}\widehat{X} - \widehat{A}\widehat{X} - \widehat{X}\widehat{B} + \widehat{C} = 0. \tag{7.6.2}$$

It is proved in [22] that $\widehat{X} = \varPhi$ is the solution of (7.6.2) uniquely characterized by

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

and at the same time the complementary ARE of (7.6.2) has the solution $\widehat{\varPsi}$ uniquely characterized by

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

where

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

In solving (7.6.2) by SDA [24], Guo, Iannazzo, and Meini [22] picked

$$w = \mathbf{1}_{m+n}/(\mathbf{1}_{m+n}^{\mathrm{T}} z) \tag{7.6.3}$$

for simplicity, and

$$\alpha = \beta = \eta = \max\{\alpha_{\mathrm{opt}}, \beta_{\mathrm{opt}}\} \tag{7.6.4}$$

to ensure[4] $\eta \in \operatorname{eig}(\widehat{R})$ contributes nothing to $\rho(\mathscr{C}(\widehat{R}; \eta, \eta))$, where α_{opt} and β_{opt} are as in (6.3.6).

It has been noted [41] that compared to ADDA, SDA will experience slow convergence if α_{opt} and β_{opt} differ substantially. Naturally applying ADDA to (7.6.2) would likely lead to a faster algorithm for the same reason. The rate of convergence of ADDA on (7.6.2) is determined by $\rho(\mathscr{C}(\widehat{S}; \alpha, \beta)) \cdot \rho(\mathscr{C}(\widehat{R}; \beta, \alpha))$, and we will pick

$$\alpha = \alpha_{\mathrm{opt}}, \quad \beta = \beta_{\mathrm{opt}}, \quad \eta = \beta_{\mathrm{opt}}, \tag{7.6.5}$$

as discussed in Remark 7.3.1 and to make sure $\eta \in \operatorname{eig}(\widehat{R})$ contributes nothing to $\rho(\mathscr{C}(\widehat{R}; \beta, \alpha))$.

[4] Recall that SDA is ADDA (Algorithm 6.4.1) after setting $\alpha = \beta$, and its rate of convergence is determined by $\rho(\mathscr{C}(\widehat{S}; \alpha, \alpha)) \cdot \rho(\mathscr{C}(\widehat{R}; \alpha, \alpha))$.

For their references in the section 8.2, we denote these two methods for solving the MARE (1.0.1) via ARE (7.6.2) by SDAs and ADDAs, respectively, with the suffix "s" standing for the shift in (7.6.1). We will use the parameters in (7.6.3) and (7.6.4) for SDAs and those in (7.6.3) and (7.6.5) for ADDAs.

CHAPTER 8

Numerical examples

In this chapter, we present numerical examples for ADDA in section 8.1 first and then for d-ADDA in section 8.2.

8.1 Numerical Examples of ADDA

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 [42]. We will use two error measures to gauge accuracy in a computed solution $\widehat{\Phi}$: the Normalized Residual (`NRes`)

$$\mathtt{NRes} = \frac{\|\widehat{\Phi}D\widehat{\Phi} - A\widehat{\Phi} - \widehat{\Phi}B + C\|}{\|\widehat{\Phi}\|(\|\widehat{\Phi}\|\|D\| + \|A\| + \|B\|) + \|C\|}, \tag{8.1.1}$$

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

$$\mathtt{ERErr} = \max_{i,j} |(\widehat{\Phi} - \Phi)_{(i,j)}|/\Phi_{(i,j)} \tag{8.1.2}$$

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 using (8.1.1) hereafter, we use ℓ_1-operator norm $\|\cdot\|_1$ as an example. For all practical purpose, any matrix norm should work just fine.

Both errors defined by (8.1.1) and (8.1.2) 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 [42] and then present them along with our numerical results.

Let[1] W be perturbed to $\widetilde{W}$ in such a way that

$$|\widetilde{A} - A| \le \epsilon |A|,\ |\widetilde{B} - B| \le \epsilon |B|,\ |\widetilde{C} - C| \le \epsilon C,\ |\widetilde{D} - D| \le \epsilon D, \tag{8.1.3}$$

where $0 \le \epsilon < 1$. It has been shown [42] that $\widetilde{\Phi}_{(i,j)} = 0$ if and only if $\Phi_{(i,j)} = 0$, under (8.1.3) 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

$$A = D_1 - N_1, \quad D_1 = \operatorname{diag}(A), \tag{8.1.4a}$$

$$B = D_2 - N_2, \quad D_2 = \operatorname{diag}(B). \tag{8.1.4b}$$

Correspondingly

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

and set

$$\lambda_1 = \rho(D_1^{-1}(N_1 + \Phi D)), \quad \lambda_2 = \rho(D_2^{-1}(N_2 + D\Phi)), \quad \lambda = \max\{\lambda_1, \lambda_2\}, \tag{8.1.5}$$

$$\tau_1 = \frac{\min_i A_{(i,i)}}{\max_j B_{(j,j)}}, \qquad \tau_2 = \frac{\min_j B_{(j,j)}}{\max_i A_{(i,i)}}. \tag{8.1.6}$$

If W is nonsingular, then $A - \Phi D$ and $B - D\Phi$ are nonsingular M-matrices by Theorem 6.3.1; so $\lambda_1 < 1$ and $\lambda_2 < 1$ [38, Theorem 3.15 on p.90] and thus $0 \le \lambda < 1$. If W is an irreducible singular M-matrix, then by Theorem 6.3.1(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 $\lambda_2 < 1$;

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

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 [21] 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 [42]. 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 (8.1.3) is sufficiently small, and $\widetilde{W}$ is an M-matrix. We have

1. $$|\Phi - \widetilde{\Phi}| \le \left[2\gamma\epsilon\, \mathbf{1}_{n,m} + O\left(\epsilon^2\right)\right] \Phi, \tag{8.1.7}$$

   where γ is given by

   $$(A - \Phi D)\Upsilon + \Upsilon(B - D\Phi) = D_1\Phi + \Phi D_2, \quad \gamma = \max_{i,j} \Upsilon_{(i,j)}/\Phi_{(i,j)}. \tag{8.1.8}$$

2. $$|\Phi - \widetilde{\Phi}| \le \left[2mn\, \kappa\chi\, \epsilon + O\left(\epsilon^2\right)\right] \Phi, \tag{8.1.9}$$

   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,

   $$\chi = \max\left\{\frac{1 + \lambda_1 + (1 + \lambda_2)\tau_1^{-1}}{1 - \lambda_1 + (1 - \lambda_2)\tau_1^{-1}}, \frac{1 + \lambda_2 + (1 + \lambda_1)\tau_2^{-1}}{1 - \lambda_2 + (1 - \lambda_1)\tau_2^{-1}}\right\} \le \frac{1 + \lambda}{1 - \lambda}. \tag{8.1.10}$$

Example	r_{adda}	$r_{\text{sda-ss}}$	r_{sda}	$\varrho(I - \Phi\Psi)$	$\varrho(I - \Psi\Phi)$
8.1.1 $(\xi = 1.5)$	0.58	0.75	0.64	0.5	0.5
8.1.1 $(\xi = 1 + 10^{-6})$	$1 - 10^{-6}$	$1 - 7 \cdot 10^{-7}$	$1 - 10^{-6}$	$1 - 2 \cdot 10^{-6}$	$1 - 2 \cdot 10^{-6}$
8.1.2	0.06	0.14	0.25	$6.3 \cdot 10^{-2}$	$6.3 \cdot 10^{-2}$
8.1.3	0.11	0.11	$1 - 2 \cdot 10^{-4}$	$5.9 \cdot 10^{-2}$	$1.1 \cdot 10^{-1}$

Table 8.1. Rates of convergence of ADDA, SDA-ss, and SDA

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

$$\chi = 2 \times \begin{cases} \dfrac{1 + \lambda_1 + 2\tau_1^{-1}}{1 - \lambda_1}, & \text{if } u_1^{\mathrm{T}} v_1 > u_2^{\mathrm{T}} v_2, \\ \dfrac{1 + \lambda_2 + 2\tau_2^{-1}}{1 - \lambda_2}, & \text{if } u_1^{\mathrm{T}} v_1 < u_2^{\mathrm{T}} v_2. \end{cases} \tag{8.1.11}$$

It is proved both γ and κ are finite [42]. Between (8.1.7) and (8.1.9), the linear term in the former is sharp while the one in the latter is not. But (8.1.9) 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 (8.1.3) 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.

All computations are performed in MATLAB with $\boldsymbol{u} = 1.11 \times 10^{-16}$. Optimal parameters as specified in section 6.5 are used for ADDA, SDA, and SDA-ss. Kahan's stopping criteria [43]:

$$\frac{(X_{k+1} - X_k)_{(i,j)}^2}{(X_k - X_{k-1})_{(i,j)} - (X_{k+1} - X_k)_{(i,j)}} \leq \epsilon \cdot (X_{k+1})_{(i,j)} \quad \text{for all } i \text{ and } j \tag{8.1.12}$$

is used to terminate iterations, 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.

Example	λ_1	λ_2	2γ	κ	$\kappa\chi$
8.1.1 ($\xi = 1.5$)	0.78	1.0	15.0	3.0	84.0
8.1.1 ($\xi = 1 + 10^{-6}$)	$1 - 6.7 \cdot 10^{-7}$	1.0	$6.0 \cdot 10^6$	$1.0 \cdot 10^6$	$1.2 \cdot 10^{13}$
8.1.2	1	0.4	$3.2 \cdot 10^2$	30.9	$1.6 \cdot 10^2$
8.1.3	0.11	1	$2.1 \cdot 10^4$	1.1	$4.8 \cdot 10^4$

Table 8.2. Parameters in the first order error bounds

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

$$\alpha_{\text{opt}} \overset{\text{def}}{=} \max_i A_{(i,i)} \neq \beta_{\text{opt}} \overset{\text{def}}{=} \max_i B_{(i,i)}.$$

We will present three examples here. More examples can be found in [41]. Table 8.1 summarizes rates of convergence for ADDA, SDA-ss, and SDA for the examples, computed according to (6.5.2), (6.5.3), and (6.5.14). 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 [43]. Table 8.2 summarizes various stability parameters in the first order error bounds at the beginning of this section. They can and will be used to explain the entrywise relative accuracy in computed $\widehat{\Phi}$.

Example 8.1.1. In this example, $m = n = 2$ and

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

Making $\xi = 1$ and scaling W by 10^{-3} recovers a null recurrent case example in [6] (see also [22, Test 7.2]). It can be verified that $\Phi = \frac{1}{2}\mathbf{1}_{2,2}$ and $\Psi = \frac{1}{2\xi}\mathbf{1}_{2,2}$. Note also W is an irreducible singular M-matrix:

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

Figure 8.1 shows plots for $\xi = 1.5$ and $\xi = 1 + 10^{-6}$: the *left* ones for `NRes` and the *right* ones for `ERErr`. The horizontal dotted line in the right plots are $(m+n)\gamma\boldsymbol{u}$. If `ERErr` falls below the dotted line, we regard the computed $\widehat{\Phi}$ as having the deserved entrywise relative accuracy. We will follow this way of presenting iteration histories in the rest of examples.

The case in which $\xi = 1$ is the critical case for which the doubling algorithms still converge but only linearly [13]. But for $0 < \xi \neq 1$ all three methods converge quadratically. In Figure 8.1 for $\xi = 1.5$, ADDA is the fastest, SDA comes in second, and SDA-ss is the slowest. Little differences between SDA and ADDA for $\xi = 1 + 10^{-6}$ as expected and both are faster than SDA-ss, but not by much, and all three algorithms take about 24 iteration steps, about 3 times as many as that for $\xi = 1.5$. $\diamond$

Example 8.1.2.

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

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[2] Φ and Ψ by the computerized algebra system *Maple* with 100 decimal digits. This "exact" solution Φ's

[2] Thanks to an anonymous referee, these exact solutions can also be constructed explicitly. However, evaluating such explicitly constructed solutions does not guarantee the smallest entries in magnitude to be fully accurate due to harmful cancelations, unless the evaluation is done in a floating point arithmetic environment with precision about twice as much as the IEEE double precision floating point arithmetics. We outline the construction as follows. Since A is the sum of I_n and a special circulant matrix, we have [10, p.356] $A = Q\Lambda Q^*$, where Q is unitary and Λ is diagonal and both are complex and known explicitly. Here Q^* is the complex conjugate transpose of Q. Let

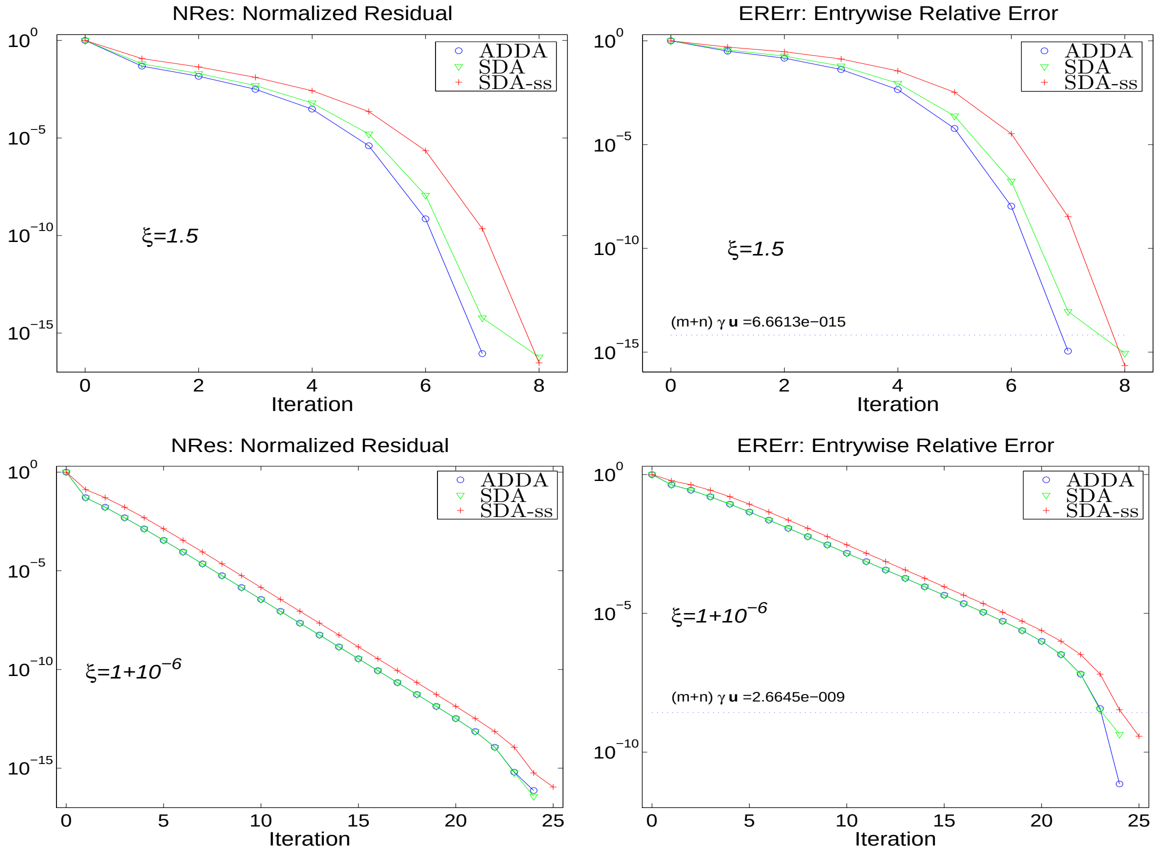

Figure 8.1. Example 8.1.1 for $\xi = 1.5$ and $\xi = 1 + 10^{-6}$. The case for $\xi = 1 + 10^{-6}$ is so much close to the critical case, convergence by the three algorithms looks like linear, except towards the very end. Note also much larger error bounds for the case $\xi = 1 + 10^{-6}$ than for the case $\xi = 1.5$. SDA-ss is actually slightly slower than SDA (and ADDA) for the two runs. .

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 of magnitudes in their entries, all three methods

$\Phi_Q = Q^* \Phi Q$. MARE $\Phi D \Phi - A\Phi - \Phi B + C = 0$ can be transformed to $20\Phi_Q^2 - \Lambda \Phi_Q - 10\Phi_Q \Lambda + 2I = 0$ whose interested solution can be constructed from a basis matrix of the invariant subspace of $\begin{pmatrix} 10\Lambda & -20I \\ 2I & -\Lambda \end{pmatrix}$ associated with those eigenvalues of positive real parts. It can be seen that one such a basis matrix takes the form $(X_1^{\mathrm{T}}, X_2^{\mathrm{T}})^{\mathrm{T}}$ with diagonal X_i, and consequently $\Phi_Q = X_2 X_1^{-1}$ is diagonal. The n diagonal entries of Φ_Q can then be computed by solving n scalar quadratic equations $20t^2 - 11\mu t + 2 = 0$ in t for each diagonal entry μ of Λ, and picking the root t such that $\mu > t$ (because $B - D\Phi = Q(20\Lambda - 20\Phi_Q)Q^*$. Similarly $\Psi C \Psi - \Psi A - B\Psi + D = 0$ can be transformed to

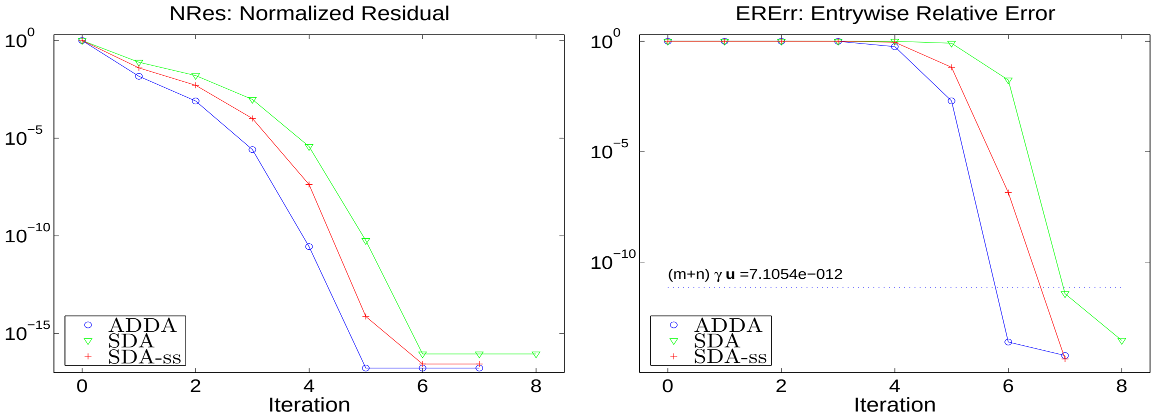

Figure 8.2. Example 8.1.2. Uneven convergence towards entries with widely different magnitudes. ERErr is still large even when NRes is already tiny before $\widehat{\Phi}$ is fully entrywise converged. .

are able to deliver computed $\widehat{\Phi}$ and $\widehat{\Psi}$ with entrywise relative errors at the level of $O(\boldsymbol{u})$. See Figure 8.2. 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 8.1.3 ([6, 22])**.** This is essentially the example of a positive recurrent Markov chain with nonsquare coefficients, originally from [6]. Here

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

$2\Psi_Q^2 - \Psi_Q \Lambda - 10\Lambda\Psi_Q + 20I = 0$ whose interested solution is also diagonal for the same reason, where $\Psi_Q = Q^*\Psi Q$. As by-product, one can argue that $\Psi_Q = 10\Phi_Q$ to conclude $\Psi = 10\Phi$.

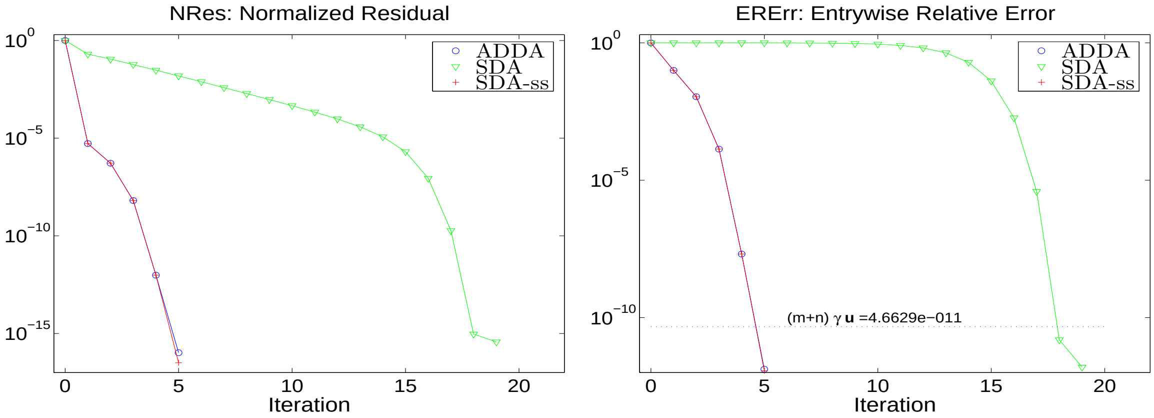

Figure 8.3. Example 8.1.3. ADDA and SDA-ss are barely distinguishable. Both are much faster than SDA. .

It is known $\Phi = \frac{1}{18} \cdot \mathbf{1}_{2,18} = \Psi^{\mathrm{T}}$. In this example, A and B differ a great deal in magnitude. Figure 8.3 shows the performance of the three methods. We see that ADDA and SDA-ss are about the same, and both are much faster than SDA. $\diamond$

Along with three 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 factor 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 6.5, all three methods are capable of delivering computed $\widehat{\Phi}$ with the deserved entrywise relative accuracy as warranted by the input data.

8.2 Numerical Examples of d-ADDA

In this section, we will also present three numerical examples to test numerical effectiveness of d-ADDA, in comparison with ADDA, SDA, and SDA-ss. As in the last section, we will use the *normalized residual* (`NRes`) error (8.1.1) to gauge accuracy in a computed solution Φ and the *entrywise relative error* (`ERErr`) (8.1.2). Moreover, we will use *normalized error* (NErr),

$$\text{NErr} = \frac{\|\boldsymbol{\Phi} - \Phi\|_1}{\|\Phi\|_1} \tag{8.2.1}$$

which is not available in actual computations but is made available here for testing purpose like `NRes` in section 8.1. These errors defined in (8.1.1), (8.1.2) and (8.2.1) are 0 if Φ is exact, but numerically they can only be made as small as $O(\boldsymbol{u})$, where $\boldsymbol{u}$ is the unit machine roundoff.

In [41, 42], it was argued that the doubling algorithms SDA [24, 22], SDA-ss [11], and ADDA [41] 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 (7.2.15) and the shifted ARE (7.6.2) are no longer MAREs and thus there is no guarantee that all harmful cancelations can be avoided when SDA or ADDA is applied to either one of them. This means that in general computed minimal nonnegative solutions Φ may not have deserved entrywise relative accuracy if some of the entries of Φ are very tiny relative to others, even though `NRes` is reduced to the level of $O(\boldsymbol{u})$. For this reason, we will use `NRes` $\leq 5 \times 10^{-14}$ as the stopping criteria in our tests here, instead of Kahan's criteria [43, 41] designed to stop the iterations only when Φ is computed to its deserved entrywise relative accuracy.

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

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

Table 8.3. NRes, NErr, and ERErr at convergence for all examples. Boldfaced entries are worth paying attention to. For the critical case (Example 8.2.1 with $\xi = 1$), ADDA on the original MAREs returns solutions with NErr about $O(\sqrt{\boldsymbol{u}})$, consistent with the error analysis in [21], even though the corresponding NRes is already $O(\boldsymbol{u})$. Examples 8.2.1 ($\xi = 10$) is special in that Φ's entries varies greatly in magnitude and consequently SDAs, ADDAs, dADDAe, and dADDAq have trouble getting tiny entries of Φ correct, even though all NErr are already $O(\boldsymbol{u})$. ADDA would have computed Φ to nearly full entrywise relative accuracy if it had continued for two more iterations as in last section.

1. ADDA introduced in Chapter 6. We use it as a representative for all doubling algorithms derivable from bilinear transformations, including SDA [24, 22] and SDA-ss [11], since ADDA is the fastest among all.
2. SDAs of [22] (as outlined in section 7.6). It is the first method ever proposed to improve SDA for irreducible singular MAREs.
3. ADDAs (as outlined in section 7.6). Since ADDA improves SDA, naturally we would expect ADDAs improves SDAs.
4. dADDAe which is Algorithm 7.2.1 combined with the elimination approach in subsection 7.4. For simplicity, all $i_0 = 1$. Actually in all examples, $z = \mathbf{1}_{m+n}$; so there is no need to do pivoting to control $\|V\|_1 \|V^{-1}\|_1$.

5. dADDAq which is Algorithm 7.2.1 combined with the Householder transformation approach in subsection 7.5.

Example 8.2.1 (Section 8.1 Example 8.1.2)**.**

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

W is an irreducible singular M-matrix:

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

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

$$7.4339\cdot 10^{-4} \le \Phi_{(i,j)} \le 3.8270\cdot 10^{-1}, \quad \text{for } \xi = 1, \tag{8.2.2}$$

$$5.7251\cdot 10^{-30} \le \Phi_{(i,j)} \le 6.3012\cdot 10^{-1}, \quad \text{for } \xi = 10. \tag{8.2.3}$$

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

Figure 8.4 plots the convergence histories of the five methods. For $\xi = 1$, ADDA converges linearly because the case falls into the critical case [13]. All methods are able to reduce `NRes` to about $O(\boldsymbol{u})$ as they should. Since Φ's entries vary in magnitude by a factor about 500, we would expect that `ERErr` for all be about $O(500\boldsymbol{u}) = O(10^{-12})$ which is true for all methods, except ADDA as shown in Table 8.3. 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 [21] that roughly speaking a perturbation of size ϵ to W will result in an error in Φ about $O(\sqrt{\epsilon})$. On the other hand, the shifting technique built

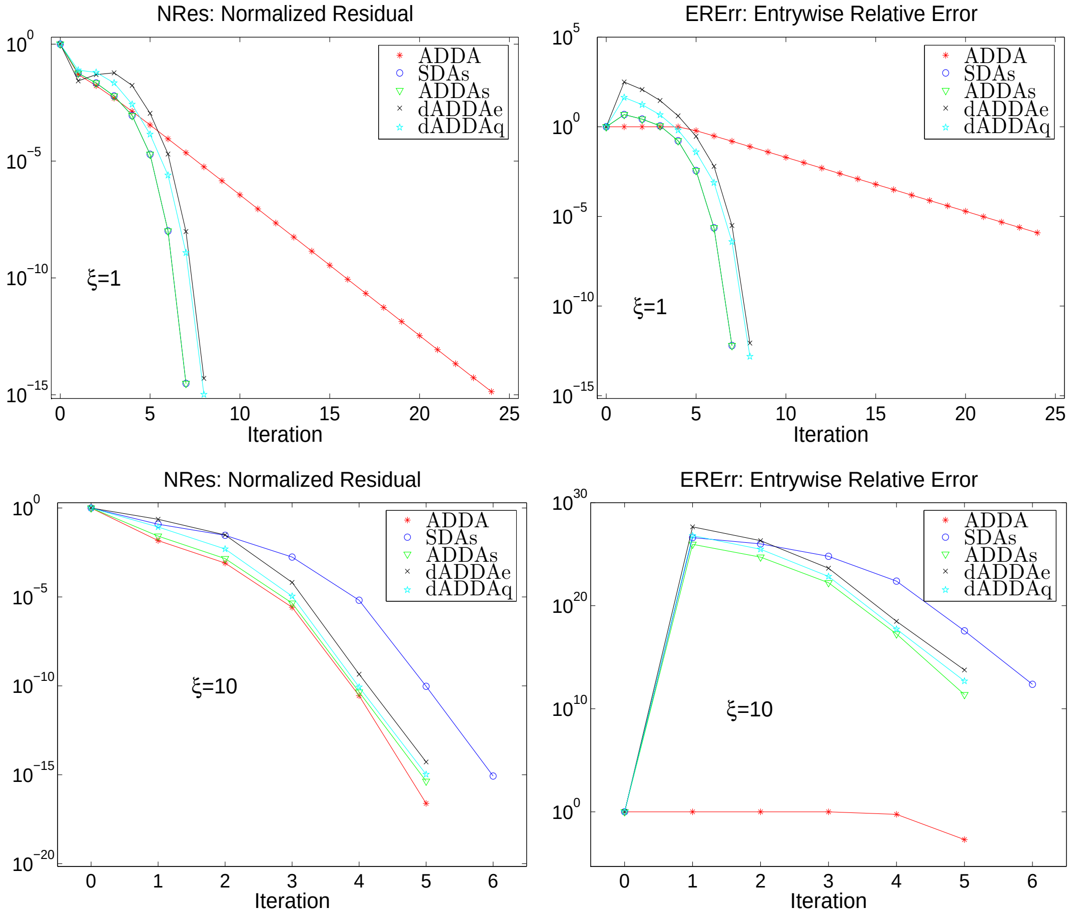

Figure 8.4. Example 8.2.1. For $\xi = 1$ ADDA converges linearly and for $\xi = 10$ ADDA performs the best. Also for $\xi = 10$, all methods, except ADDA (which took 7 iterations in last secion, two more than here, to deliver Φ with about 15 correct decimal digits entrywise), fail to compute accurately Φ's tiny entries..

into SDAs and ADDAs and the deflating technique built into dADDAe and dAADAq make the resulting ARE (7.2.15) and (7.6.2) sufficiently well-conditioned to be solved accurately. Guo, Iannazzo, and Meini [22] 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

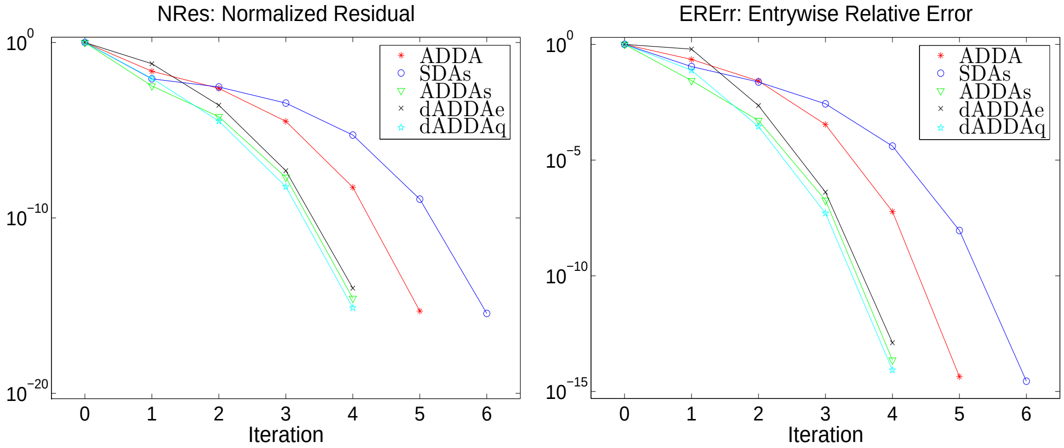

Figure 8.5. Example 8.2.2. ADDA is even faster than SDAs. ADDAs, dADDAe, and dADDAq work about equally well..

ADDA's superior performance. ADDA would have computed Φ to with almost full entrywise accuracy if it had not been stopped prematurely by one stopping criteria `NRes` $\le 5 \times 10^{-14}$ used for all. In fact, this example is the same as example 8.1.2 in Section 8.1, where ADDA delivered Φ to have almost 15 correct decimal digits entrywise in 7 iterations. The inability of the other methods to compute Φ's tiny entries accurately is evident from the right-bottom plot in Figure 8.4 and Table 8.3, even though at the same time all methods are able to reduce `NRes` to about $O(\boldsymbol{u})$. $\diamond$

Example 8.2.2. W is an irreducible singular M-matrix, randomly generated by the following piece of MATLAB code:

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

In the end, $W\mathbf{1}_{2n} = 0$, and with $m = n$, the coefficient matrices A, B, C, and D for an MARE (1.0.1) 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" Φ for testing purpose. For this saved W, we find that

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

So all entries of Φ have about the same magnitude which suggests that tiny `NRes` implies tiny `ERErr`. This is clearly the case as shown in Figure 8.5. 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 8.2.3. This is essentially the example of a positive recurrent Markov chain with nonsquare coefficients, originally from [6]. 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 known $\Phi = \frac{1}{18} \cdot \mathbf{1}_{2,18} = \Psi^{\mathrm{T}}$ 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[3], to say the least. In fact, our Maple code for ADDAs with arbitrary α and β but $\eta = \beta$ gives, in exact arithmetic,

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

We did not see this phenomenon in Examples 8.2.1 and 8.2.2 both of which are nontrivial, relatively speaking. So this kind of pleasant surprise shouldn't be expected in general. Figure 8.6 displays convergence histories for all tested methods. That both

[3] More examples like this can be found in last section.

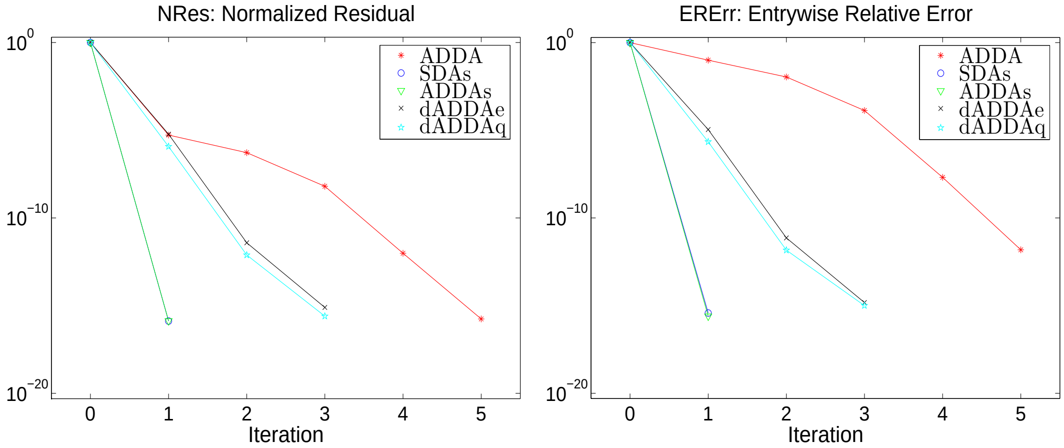

Figure 8.6. Example 8.2.3. Again SDAs and ADDAs get the solution in X_0 from their initial setup as they should because $X_0 \equiv \Phi$, independent of α and β. Both dADDAe and dADDAq take two iterations after the initial setup, while ADDA takes five iterations. .

NErr and **ERErr** for ADDA at convergence are about 10^{-12} can be explained by the relevant parameters in Table 8.2. ◇

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 Φ, even when some of the entries of Φ are extremely tiny relative to others. But entrywise accuracy in computed Φ is limited to about $O(\sqrt{\boldsymbol{u}})$.

2. The shifting technique of Guo, Iannazzo, and Meini [22] and the deflating technique in Chapter 7 can greatly improve the conditioning of an MARE in the critical case, enabling Φ to be computed much more accurately in the sense of making normalized error NErr to about $O(\boldsymbol{u})$. But when Φ's entries vary too much in magnitude, tiny entries may lose some or even all significant digits. When that happens, ADDA should be used directly to the original MARE.

3. The last example is accidental for both ADDAs and SDAs in that $X_0 \equiv \mathit{\Phi}$, independent of the parameters α and β. In general, ADDAs is faster than SDAs as one might expect from the conclusion in [41] that ADDA is at least as good as SDA and can be faster if A and B are very different in magnitude.

CHAPTER 9

Conclusion

Throughout this thesis, after some fundamental knowledge, we first mentioned the ADI method and the doubling algorithm before introducing our ADDA (Alternating-Directional Doubling Algorithm), which is the combination of the alternating-directional idea of ADI for Sylvester equation and the idea of SDA. We have proved theoretically and numerically that our ADDA method is always the best comparing with all the other doubling algorithms. Next, for the critical case, for which our ADDA method converges linearly (instead of quadratically), we established a deflation method called d-ADDA (deflated ADDA) to deflate an irreducible singular MARE. It is widely accepted that in computing eigen-decompositions, deflation approaches are often preferred to shifting mechanisms. There we also demonstrated the effectiveness of d-ADDA and compared it with existing methods by theoretical analysis and numerical examples.

APPENDIX A

Application to M-matrix Sylvester equation

When $D = 0$, MARE (1.0.1) degenerates to a Sylvester equation:

$$AX + XB = C. \tag{A.0.1}$$

The assumption (7.0.1) 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 (A.0.1) is an *M-Matrix Sylvester Equation* (MSE) as defined in [43]: both A and B have positive diagonal entries and nonpositive off-diagonal entries and $P = I_m \otimes A + B^{\mathrm{T}} \otimes I_n$ is a nonsingular M-matrix, and $C \geq 0$.

MSE (A.0.1) has the unique solution $\varPhi \geq 0$ and its cMARE has the solution $\varPsi = 0$. Apply ADDA to (A.0.1) to get

$$E_0 = \mathscr{C}(B; \beta, \alpha) \equiv (B + \alpha I)^{-1}(B - \beta I), \tag{A.0.2a}$$

$$F_0 = \mathscr{C}(A; \alpha, \beta) \equiv (A + \beta I)^{-1}(A - \alpha I), \tag{A.0.2b}$$

$$X_0 = (\beta + \alpha)(A + \beta I)^{-1} C (B + \alpha I)^{-1}, \tag{A.0.2c}$$

and for $k \geq 0$

$$E_{k+1} = E_k^2, \qquad F_{k+1} = F_k^2, \tag{A.0.2d}$$

$$X_{k+1} = X_k + F_k X_k E_k. \tag{A.0.2e}$$

The associated error equation is

$$0 \leq \varPhi - X_k = [\mathscr{C}(A; \alpha, \beta)]^{2^k} \varPhi \, [\mathscr{C}(B; \beta, \alpha)]^{2^k}. \tag{A.0.3}$$

Smith's method [36, 43] is obtained after setting $\alpha = \beta$ in (A.0.2) always.

Alternatively, we can derive (A.0.2) through a combination of an *Alternating-Directional-Implicit* (ADI) iteration and Smith's idea in [36]. Given an approximation $\boldsymbol{X} \approx \varPhi$, 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

$$\boldsymbol{Z} = X_0 + F_0 \boldsymbol{X} E_0, \tag{A.0.4}$$

where E_0, F_0, and X_0 are the same as in (A.0.2a) – (A.0.2c). With $\boldsymbol{X} = 0$, keep iterating (A.0.4) 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 $\varPhi = \boldsymbol{Z}_\infty = \sum_{i=0}^{\infty} F_0^i X_0 E_0^i$. 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 $\{\boldsymbol{Z}_i\}$, (A.0.2) computes only the 2^kth members. In view of its connection to ADI and Smith's method [36], we call (A.0.2) an *Alternating-Directional Smith Method* (ADSM) for MSE (A.0.1). This connection to ADI is also the reason for us to name our Algorithm 6.4.1 an *Alternating-Directional Doubling Algorithm* (ADDA).

Equation (A.0.3) gives

$$\limsup_{k\to\infty} \|\varPhi - X_k\|^{1/2^k} \le \rho(\mathscr{C}(A; \alpha, \beta)) \cdot \rho(\mathscr{C}(B; \beta, \alpha)), \tag{A.0.5}$$

suggesting us to pick α and β to minimize the right-hand side of (A.0.5) for fastest convergence. Subject to again

$$\alpha \ge \alpha_{\text{opt}} \overset{\text{def}}{=} \max_i A_{(i,i)}, \quad \beta \ge \beta_{\text{opt}} \overset{\text{def}}{=} \max_j B_{(j,j)} \tag{6.3.6}$$

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 6.1.1 that $\alpha = \alpha_{\text{opt}}$ and $\beta = \beta_{\text{opt}}$ minimize the right-hand side of (A.0.5).

APPENDIX B

Spectral Radius

Here we only give the basic definition of spectral radius and an important property of it.

For a square matrix A, the number

$$\rho(A) = \max_{\lambda \in \sigma(A)} |\lambda|$$

is called the spectral radius of A, where $\sigma(A)$ stands for the spectral of A. It is not uncommon for applications to require only a bound on the eigenvalues of A, i.e. precise knowledge of each eigenvalue may not be required, but only just an upper bound on $\rho(A)$ is all that is often needed.

A rather crude but cheap and useful property about spectral radius is that

$$\rho(A) \leq \|A\|.$$

The proof of this is easy. Take (λ, x) as an eigenpair of A, then we have $\lambda x = Ax$, which implies

$$|\lambda| \|x\| = \|\lambda x\| = \|Ax\| \leq \|A\| \|x\|.$$

So $|\lambda| \leq \|A\|$ for all $\lambda \in \sigma(A)$.

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] S. Barnett and C. Storey, *Matrix methods in stability theory*, Nelson, 1970.

[6] 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.

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

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

[9] 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.

[10] D. S. Bernstein, *Matrix Mathematics: Theory, Facts, and Formulas*, Princeton University Press, Princeton, NJ, 2009. 2nd ed.

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

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

[13] 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.

[14] 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.

[15] 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.

[16] J. Demmel, *Applied Numerical Linear Algebra*, SIAM, Philadelphia, PA, 1997.

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

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

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

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

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

[22] 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.

[23] 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.

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

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

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

[27] D. Kincaid and W. Cheney, *Numerical Analysis*, Brooks/Cole Publishing Company, Pacific Grove, CA, 1991.

[28] P. Lancaster and L. Rodman, *Algebraic Riccati Equations*, Oxford University Press, New York, USA, 1995.

[29] 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.

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

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

[32] C. D. Meyer, *Matrix Analysis and Applied Linear Algebra*, SIAM, Philadelphia, PA, 2000.

[33] F. Poloni, *Quadratic vector equations*, Linear Algebra Appl., 438 (2010), pp. 1627–1644.

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

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

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

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

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

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

[40] S.-J. Wang and X.-X. Guo, *On monotone convergence rate of alternately linearized implicit iteration method*, J. Numer. meth. Comput. Appl., 31 (2010), pp. 76–80.

[41] W.-G. Wang, W.-C. Wang, and R.-C. Li, *ADDA: Alternating-directional doubling algorithm for M-matrix algebraic Riccati equations*, SIAM J. Matrix Anal. Appl., 33 (2012), pp. 170–194.

[42] J. Xue, S. Xu, and R.-C. Li, *Accurate solutions of M-matrix algebraic Riccati equations*, Numer. Math., 120 (2012), pp. 671–700.

[43] ——, *Accurate solutions of M-matrix sylvester equations*, Numer. Math., 120 (2012), pp. 639–670.

BIOGRAPHICAL STATEMENT

Weichao Wang was born in Dalian, China, in 1984. He graduated from University of Science and Technology of China in 2008, in Hefei, China, with a Bachelor of Computational Mathematics. In 2008 he entered the Computational Mathematics program at The University of Texas at Arlington and completed his PhD in August 2013. In the same year he did an internship in Link-Quest Inc. and then joined the company as a senior algorithm engineer.