Journal of Scientific Computing (2020) 84:15
https://doi.org/10.1007/s10915-020-01263-4

Accurate Numerical Solution for Shifted *M*-Matrix Algebraic Riccati Equations

Changli Liu[1] · Jungong Xue[2] · Ren-Cang Li[3]

Received: 23 April 2020 / Revised: 5 June 2020 / Accepted: 12 June 2020 / Published online: 6 July 2020

Abstract
An algebraic Riccati equation (ARE) is called a shifted *M*-matrix algebraic Riccati equation (MARE) if it can be turned into an MARE after its matrix variable is partially shifted by a diagonal matrix. Such an ARE can arise from computing the invariant density of a Markov modulated Brownian motion. Sufficient and necessary conditions for an ARE to be a shifted MARE are obtained. Based on the conditions, a highly accurate implementation of the alternating directional doubling algorithm (ADDA) is established to compute the extremal solution of a shifted MARE, as well as a quantity needed for computing the invariant density in the application, with high entrywise relative accuracy. Numerical examples are presented to demonstrate the theory and algorithms.

Keywords M-matrix; algebraic Riccati equation · Markov modulated Brownian motion · Minimal nonnegative solution · Doubling algorithm

Mathematics Subject Classification 15A24 · 60J65 · 65F30 · 65H10

1 Introduction

In general an algebraic Riccati equation (ARE) takes the form [21]

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

✉ Ren-Cang Li
rcli@uta.edu

Changli Liu
chliliu@hotmail.com

Jungong Xue
xuej@fudan.edu.cn

1 College of Mathematics, Sichuan University, Chengdu 610065, People's Republic of China

2 School of Mathematical Science, Fudan University, Shanghai 200433, People's Republic of China

3 Department of Mathematics, University of Texas at Arlington, P.O. Box 19408, Arlington, TX 76019-0408, USA

where $A \in \mathbb{F}^{n\times n}$, $B \in \mathbb{F}^{m\times m}$, $C \in \mathbb{F}^{n\times m}$, and $D \in \mathbb{F}^{m\times n}$ are known coefficient matrices, and $X \in \mathbb{F}^{n\times m}$ is to be found. Here $\mathbb{F} = \mathbb{C}$ (the set of complex numbers) or $\mathbb{R}$ (the set of real numbers), but in what follows, we are interested in $\mathbb{F} = \mathbb{R}$ only. Let

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

an $(m+n)\times(m+n)$ matrix that defines the ARE (1.1). Following [30], we will call ARE (1.1) an *M-Matrix algebraic Riccati equation* (MARE) if W is a nonsingular or an irreducible singular M-matrix. This kind of AREs arise in applied probability and transportation theory (see [10,13–15,18,19,26] and the references therein). It is shown in [10,14] that when ARE (1.1) is an MARE, it has a unique minimal nonnegative solution Φ, i.e., entrywise

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

Thanks to several fundamental researches in the past 20 years or so, such an MARE is very well understood nowadays both theoretically and numerically. C. Guo and his collaborators completed much of the studies on the existence and basic properties of the unique minimal nonnegative solution Φ [10,12,13]. The first structure-preserving doubling algorithm (SDA) was proposed by Guo et al. [15] in 2006, and it was immediately clear at that moment that SDA is superior to Newton's method in solving MARE for the unique minimal nonnegative solution Φ. Soon after, SDA was improved by two other more efficient doubling algorithms SDA-ss [7] and ADDA [28]. A highly accurate implementation of ADDA was discovered first by Nguyen and Poloni [23] for a singular but irreducible W and then by Xue and Li [29] for nonsingular W as well. An entrywise relative perturbation theory for MARE was established earlier in [30,31]. More details can be found in the recent books [6, chapter 5], [16, chapter 6].[1]

In this paper we introduce a class of AREs that can be turned into MAREs after a partial shift to the diagonal of their matrix variables. Given[2] $1 \le p \le \min\{m,n\}$, we partition $X \in \mathbb{R}^{n\times m}$ as

$$X = \begin{array}{c} \\ p \\ n-p \end{array}\!\!\begin{array}{c} \begin{array}{cc} p & m-p \end{array} \\ \begin{bmatrix} X_{11} & X_{12} \\ X_{21} & X_{22} \end{bmatrix} \end{array}. \tag{1.3}$$

Perform a change of variable by

$$X_\Omega = X + \Omega \quad \text{with } \Omega = \begin{array}{c} \\ p \\ n-p \end{array}\!\!\begin{array}{c} \begin{array}{cc} p & m-p \end{array} \\ \begin{bmatrix} \Lambda & 0 \\ 0 & 0 \end{bmatrix} \end{array} \tag{1.4}$$

and then plug $X = X_\Omega - \Omega$ into ARE (1.1) to yield an ARE in X_Ω

$$X_\Omega D X_\Omega - A_\Omega X_\Omega - X_\Omega B_\Omega + C_\Omega = 0 \tag{1.5}$$

whose associated matrix W_Ω is

$$W_\Omega = \begin{bmatrix} B_\Omega & -D \\ -C_\Omega & A_\Omega \end{bmatrix}, \tag{1.6}$$

where

$$A_\Omega = A + \Omega D,\ \ B_\Omega = B + D\Omega,\ \ C_\Omega = \Omega D\Omega + A\Omega + \Omega B + C. \tag{1.7}$$

[1] In [16], an MARE is more broadly defined for the one with W being an M-matrix.

[2] The case $p = 0$ is the usual MARE case which was thoroughly studied; see, e.g., in [6,16].

Definition 1.1 ARE (1.1) is called a *shifted* MARE *of index* p if there exists a diagonal matrix $\Lambda_0 \in \mathbb{R}^{p\times p}$ such that ARE (1.5) is an MARE for any diagonal matrix Λ with

$$\Lambda_{(i,i)} > [\Lambda_0]_{(i,i)} \quad \text{for } i = 1, 2, \ldots, p,$$

i.e., the defining matrix W_Ω is a nonsingular M-matrix or an irreducible singular M-matrix.

Later in "Appendix 1", we will explain how a shifted MARE arises from computing the invariant density of a Markov modulated Brownian motion (MMBM). In that application, a portion of a particular solution to (1.1), among others, are sought, preferably with high entrywise relative accuracy. The main goals of this paper are three-fold: (1) to establish necessary and sufficient conditions for an ARE (1.1) to be a shifted MARE of index p, (2) to compute the shift Ω, and (3) to accurately solve shifted MARE (1.1) so that, when applied to MMBM, the associated density can be computed with high entrywise relative accuracy.

The rest of this paper is organized as follows. Section 2 collects some of the relevant results regarding M-matrices and MARE. In Sect. 3, we launch our detailed study on shifted MARE in terms of coefficient matrix structures, sufficient and necessary conditions for an ARE to be an MARE, and how to construct the shifting matrix Λ in (1.4). We explain how to adopt the idea of highly accurate ADDA [23,29] for solving shifted MARE with high entrywise relative accuracy in our general setting. In Sect. 5, we return to explain an accurate calculation of a quantity needed by MMBM. Numerical results are reported in Sect. 6 and finally we draw our conclusions in Sect. 7.

Notation $\mathbb{R}^{m\times m}$ is the set of all $m \times m$ 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. The superscript ".T" takes transpose. $\mathbf{1}_n \in \mathbb{R}^n$ is the n-dimensional vector of all ones and $\mathbf{1}_{m\times n} \in \mathbb{R}^{m\times n}$ is the $m \times n$ matrix of all ones. For $X \in \mathbb{R}^{m\times n}$, $X_{(i,j)}$ refers to its (i, j)th entry, $|X|$ is in $\mathbb{R}^{m\times n}$ with its (i, j)th entry $|X_{(i,j)}|$. Inequality $X \le Y$ means $X_{(i,j)} \le Y_{(i,j)}$ for all (i, j), and similarly for $X < Y$, $X \ge Y$, and $X > Y$. In particular, $X \ge 0$ means that X is entrywise nonnegative. For $X \in \mathbb{R}^{n\times n}$, denote by $\mathrm{eig}(X)$ the set of its eigenvalues counted algebraic multiplicities, by $\mathrm{diag}(X) \in \mathbb{R}^{n\times n}$ the diagonal matrix extracted out of X, and by $\mathrm{offdiag}(X) = X - \mathrm{diag}(X) \in \mathbb{R}^{n\times n}$.

2 *M*-Matrix Essentials

$A \in \mathbb{R}^{n\times n}$ is called a *Z-matrix* if $A_{(i,j)} \le 0$ for all $i \ne j$ [5, p. 284]. Any Z-matrix A can be written as $sI - B$ with $B \ge 0$, and it is called an *M-matrix* if $s \ge \rho(B)$, a *singular M-matrix* if $s = \rho(B)$, and a *nonsingular M-matrix* if $s > \rho(B)$, where $\rho(B)$ is the spectral radius of B.

The single most important property of a nonsingular M-matrix A is $A^{-1} \ge 0$. We refer the reader to, e.g., [5,8,22,27], for many other characterizations of M-matrices, but we will single out two results that we will frequently use during our developments later in this paper.

Proposition 2.1 (a) *Let $A \in \mathbb{R}^{n\times n}$ be a nonsingular M-matrix or an irreducible singular M-matrix, conformally partitioned as*

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

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 irreducible singular M-matrices if A is an irreducible singular M-matrix.

(b) *Suppose that A is a nonsingular M-matrix. A is irreducible if and only if $A^{-1} > 0$.*

The key ingredient in the recent work [23,29] to achieve high entrywise relative accuracy is the GTH-like algorithm for inverting a nonsingular M-matrix due to Alfa et al. [2]. The key idea is to represent a nonsingular M-matrix A by the so-called *triplet representation* which can determine A^{-1} entrywise to high relative accuracy. Specifically, a triplet representation $(\text{offdiag}(A), \boldsymbol{u}, \boldsymbol{v})$ of an M-matrix $A \in \mathbb{R}^{n\times n}$ consists of

$$\text{offdiag}(A) = A - \text{diag}(A),\ 0 < \boldsymbol{u} \in \mathbb{R}^n, \text{ and } \boldsymbol{v} = A\boldsymbol{u} \geq 0.$$

Often for convenience, we will not distinguish A from its triplet representation and write $A = (\text{offdiag}(A), \boldsymbol{u}, \boldsymbol{v})$. It is proved [3] that if all entries of $\text{offdiag}(A)$, $\boldsymbol{u}$, and $\boldsymbol{v}$ are known to high entrywise relative accuracy, then all entries of A^{-1} are determined to a comparable high relative accuracy, or equivalently the solution $\boldsymbol{x}$ to $A\boldsymbol{x} = \boldsymbol{b}$ for any $\boldsymbol{b} \geq 0$ is determined to a comparable high entrywise relative accuracy. Numerically, the GTH-like algorithm of Alfa et al. [2,3], using the idea in [9], computes the LU decomposition $A = LU$, via the Gaussian elimination without pivoting and without any cancellation[3] and, consequently, L and U are computed with high entrywise relative accuracy. Moreover, the diagonal entries of L are all 1 and its off-diagonal entries are non-positive, and U has positive diagonal entries and non-positive off-diagonal entries. These properties of L and U ensure that the solution $\boldsymbol{x}$ of $A\boldsymbol{x} = \boldsymbol{b} \geq 0$ can be computed to the claimed accuracy. For more detail, the reader is referred to [16].

But to serve our purpose later in Sect. 4, we shall introduce an alternative one, which we will call a *left* triplet representation, denoted by $A = (\text{offdiag}(A), \boldsymbol{u}, \boldsymbol{v})_{\text{L}}$, consisting of

$$\text{offdiag}(A) = A - \text{diag}(A),\ 0 < \boldsymbol{u} \in \mathbb{R}^n, \text{ and } \boldsymbol{v}^{\text{T}} = \boldsymbol{u}^{\text{T}}A \geq 0.$$

Essentially, there is no difference between the two types of triplet representations. In fact, if A is an M-matrix, then so is A^{T}. It can be seen that $A = (\text{offdiag}(A), \boldsymbol{u}, \boldsymbol{v})_{\text{L}}$ yields $A^{\text{T}} = (\text{offdiag}(A^{\text{T}}), \boldsymbol{u}, \boldsymbol{v})$ and $A^{-1} = [(A^{\text{T}})^{-1}]^{\text{T}}$. Therefore, the existing theory of Alfa et al. [2,3] implies that all entries of A^{-1} are also determined to a comparable high entrywise relative accuracy to what a left triplet representation $A = (\text{offdiag}(A), \boldsymbol{u}, \boldsymbol{v})_{\text{L}}$ has. Numerically, the GTH-like algorithm can be applied to $A^{\text{T}} = (\text{offdiag}(A^{\text{T}}), \boldsymbol{u}, \boldsymbol{v})$ to obtain $A^{\text{T}} = LU$ with L and U enjoying the same desirable properties as before. Finally, the solution $\boldsymbol{x}$ of $A\boldsymbol{x} = \boldsymbol{b} \geq 0$ can be computed entrywise accurately. For future reference, we summarize this alternative GTH-like algorithm as Algorithm 2.1, given a left triplet representation $A = (\text{offdiag}(A), \boldsymbol{u}, \boldsymbol{v})_{\text{L}}$. We still credit the algorithm to [2] because it is essentially the same one there.

An important characterization of the unique minimal nonnegative solution Φ of an MARE will be needed later. Consider ARE (1.1) in general with W in (1.2), and let

$$\mathscr{H} = \begin{bmatrix} I_m & \\ & -I_n \end{bmatrix} W.$$

[3] By cancellation, we mean any subtraction of one number from another of the same sign.

Algorithm 2.1 GTH-like Algorithm [2]

Require: A nonsingular M-matrix $A \in \mathbb{R}^{n\times n}$ in the form of a left triplet representation $A = (\text{offdiag}(A), \boldsymbol{u}, \boldsymbol{v})_{\text{L}}$, and $\mathbb{R}^n \ni \boldsymbol{b} \geq 0$

Ensure: Solution $\boldsymbol{x}$ to the linear system $A\boldsymbol{x} = \boldsymbol{b}$.

1: **for** $i = 1, 2, \ldots, n-1$ **do**
2: $A_{(i,i)} = \left[\boldsymbol{v}_{(i)}^{\text{T}} - \boldsymbol{u}_{(i+1:n)}^{\text{T}} A_{(i+1:n,i)}\right]/\boldsymbol{u}_{(i)}$;
3: $A_{(i,i+1:n)} = A_{i,(i+1:n)}/A_{(i,i)}$;
4: $A_{(i+1:n,i+1:n)} = A_{(i+1:n,i+1:n)} - A_{(i+1:n,i)} A_{(i,i+1:n)}$; % ignore the diagonal entries
5: $\boldsymbol{v}_{(i+1:n)}^{\text{T}} = \boldsymbol{v}_{(i+1:n)}^{\text{T}} - \boldsymbol{v}_{(i)}^{\text{T}} A_{(i,i+1:n)}$;
6: **end for**
7: extract the lower triangular part of A to give U^{T};
8: extract the strict upper triangular part of A to give L^{T} and set all of its diagonal entries to 1;
9: solve $U^{\text{T}}\boldsymbol{y} = \boldsymbol{b}$ for $\boldsymbol{y}$ by the forward substitution;
10: solve $L^{\text{T}}\boldsymbol{x} = \boldsymbol{y}$ for $\boldsymbol{x}$ by the backward substitution;
11: **return** $\boldsymbol{x}$.

The following connection between ARE (1.1) and the eigenvalue problem for $\mathscr{H}$ is well-known [16,21]. In fact, it can be verified that (1.1) is equivalent to

$$\mathscr{H}\begin{bmatrix} I_m \\ X \end{bmatrix} = \begin{bmatrix} I_m \\ X \end{bmatrix}(B - DX).$$

This means that any solution X to ARE (1.1) yields a basis matrix[4] $Z := \begin{bmatrix} I_m \\ X \end{bmatrix}$ of an m-dimensional eigenspace of $\mathscr{H}$ associated with its partial spectrum $\text{eig}(B - DX)$.

Proposition 2.2 which is extracted from [16, Theorem 6.2] is due to C. Guo and his collaborators. Item (c) uniquely characterizes the minimal nonnegative solution Φ.

Proposition 2.2 *Suppose that (1.1) is an* MARE*, i.e., W in (1.2) is a nonsingular or an irreducible singular M-matrix, and let* $\text{eig}(\mathscr{H}) = \{\lambda_1, \ldots, \lambda_{m+n}\}$*, where all λ_i are ordered by their nonincreasing real parts, i.e., $\Re(\lambda_j) \leq \Re(\lambda_i)$ for $i < j$.*

(a) *λ_m and λ_{m+1} are real, $\Re(\lambda_{m+2}) < 0 < \Re(\lambda_{m-1})$, and*

$$\begin{aligned} \Re(\lambda_{m+n}) \leq \cdots \leq \Re(\lambda_{m+2}) \leq \lambda_{m+1} \\ \leq 0 \leq \lambda_m \leq \Re(\lambda_{m-1}) \leq \cdots \leq \Re(\lambda_1). \end{aligned} \tag{2.1}$$

In particular, this implies $\lambda_{m+1} < 0 < \lambda_m$ if W is nonsingular.

(b) *When W is an irreducible singular M-matrix, there are three possibilities:*

$$\begin{cases} \lambda_m = 0, \ \lambda_{m+1} < 0, & \textit{case (i)}, \\ \lambda_m = \lambda_{m+1} = 0, & \textit{case (ii)}, \\ \lambda_m > 0, \ \lambda_{m+1} = 0, & \textit{case (iii)}. \end{cases} \tag{2.2}$$

In case (ii), 0 is a double eigenvalue coming from the 2×2 Jordan block $\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$.

(c) *$\mathscr{H}$ has a unique m-dimensional eigenspace associated with its eigenvalues in $\mathbb{C}_{0+} = \{z \in \mathbb{C} : \Re(z) \geq 0\}$, and $\begin{bmatrix} I_m \\ \Phi \end{bmatrix}$ is a basis matrix of the eigenspace.*

[4] By which we mean that the columns the matrix form a basis of the subspace.

3 Shifted MARE

The concept of shifted MARE we introduced in Sect. 1 was largely inspired by ARE (1.1) with (A.13) as explained in "Appendix 1". In this section, we will present sufficient and necessary conditions for a general ARE (1.1) to be a shifted MARE and define its extremal solution.

3.1 Sufficient and Necessary Conditions

We start by looking at the simple case where $m = n = p$ for which $\Omega = \Lambda \in \mathbb{R}^{n\times n}$ in (1.4) to get some idea. For W_Ω in (1.6) to be an M-matrix, necessarily

$$D \geq 0, \;\; \Omega D\Omega + A\Omega + \Omega B + C \geq 0, \;\; \text{and } B + D\Omega \text{ and } A + \Omega D \text{ are } M\text{-matrices.}$$

Hence D has to be diagonal; otherwise some off-diagonal entries of $B + D\Omega$ would become positive as the diagonal entries of Λ tends to infinity, contradicting that $B + D\Omega$ is an M-matrix. Since $A + \Omega D$ and $B + D\Omega$ are M-matrices and $\Omega D = D\Omega$ (both D and Ω are diagonal), A and B must be Z-matrices, i.e., their off-diagonal entries are non-positive. In order for $\Omega D\Omega + A\Omega + \Omega B + C$ to be nonnegative, A and B have to be diagonal and $-C$ a Z-matrix. We summarize our findings into the following theorem.

Theorem 3.1 *Suppose $m = n = p$. If* ARE (1.1) *is a shifted* MARE *of index p, then A, B, and D are diagonal, $D \geq 0$, and $-C$ is a Z-matrix.*

Remark 3.1 The conditions in Theorem 3.1 are not sufficient, however. For example, $A = -I$ and $D = 0$. Then for any Ω, $A_\Omega = A + \Omega D = A = -I$ and thus W_Ω can never become an M-matrix. It is also important to note that an MARE may not be a shifted MARE of any $p \geq 1$. As an example, if D has a positive off-diagonal entry in its leading p-by-p principal submatrix, say $D_{(i,j)} > 0$ for some $1 \leq i, j \leq p$ and $i \neq j$, then $[A_\Omega]_{(i,j)} = A_{(i,j)} + \Lambda_{(i,i)} D_{(i,j)} > 0$ for sufficiently large $\Lambda_{(i,i)}$ and thus cannot be an M-matrix.

Next we consider the case $1 \leq p \leq \min\{m, n\}$. Let[5] A, B, C, and D be partitioned consistently with X in (1.3):

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}, \quad C = \begin{bmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{bmatrix}, \quad D = \begin{bmatrix} D_{11} & D_{12} \\ D_{21} & D_{22} \end{bmatrix}, \quad B = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix},$$

i.e., all A_{11}, C_{11}, D_{11}, $B_{11} \in \mathbb{R}^{p\times p}$ and the sizes of all other submatrices are determined conformally according to the partitioning. Then

$$A_\Omega := A + \Omega D = \begin{bmatrix} A_{11} + \Lambda D_{11} & A_{12} + \Lambda D_{12} \\ A_{21} & A_{22} \end{bmatrix},$$

$$B_\Omega := B + D\Omega = \begin{bmatrix} B_{11} + D_{11}\Lambda & B_{12} \\ B_{21} + D_{21}\Lambda & B_{22} \end{bmatrix}.$$

When the diagonal entries of Λ are sufficiently large, A_Ω and B_Ω need to be M-matrices and hence D_{11} is diagonal and non-negative, $D_{12} = 0$, $D_{21} = 0$, and A_{11}, A_{22}, B_{11}, B_{22} are Z-matrices, and A_{12}, A_{21}, B_{12}, and B_{21} are non-positive matrices. Finally,

$$C_\Omega := \Omega D\Omega + A\Omega + \Omega B + C$$

[5] When $p = \min\{m, n\}$, some block submatrices are null, i.e., nonexistence, For example, $A = A_{11}$, $C = [C_{11}, C_{12}]$, and $D = \begin{bmatrix} D_{11} \\ D_{21} \end{bmatrix}$ in the case $p = n < m$.

$$= \begin{bmatrix} \Lambda D_{11}\Lambda + A_{11}\Lambda + \Lambda B_{11} + C_{11} & \Lambda B_{12} + C_{12} \\ C_{21} + A_{21}\Lambda & C_{22} \end{bmatrix} \geq 0 \tag{3.1}$$

for the diagonal matrix Λ with sufficiently large diagonal entries yields that $B_{12} = 0, A_{21} = 0$, A_{11} and B_{11} are diagonal, C_{12}, C_{21} and C_{22} are non-negative, and $-C_{11}$ is a Z-matrix. We therefore obtain the following necessary conditions for ARE (1.1) to be a shifted MARE of index p.

Theorem 3.2 *Suppose $1 \leq p \leq \min\{m, n\}$. If* ARE (1.1) *is a shifted* MARE *of index p, then A, B, C, and D admit the following structures:*

$$A = \begin{bmatrix} A_{11} & A_{12} \\ & A_{22} \end{bmatrix}, \quad B = \begin{bmatrix} B_{11} & \\ B_{21} & B_{22} \end{bmatrix}, \quad C = \begin{bmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{bmatrix}, \quad D = \begin{bmatrix} D_{11} & \\ & D_{22} \end{bmatrix}, \tag{3.2}$$

where A and B are Z-matrices with A_{11} and B_{11} being diagonal matrices, $-C_{11}$ is a Z-matrix, $D_{11} \geq 0$ is diagonal, and all $C_{12},\ C_{21},\ C_{22},\ D_{22} \geq 0$.

Suppose A, B, C, and D admit the block structures as described in Theorem 3.2, and let

$$A_\Omega := A + \Omega D = \begin{bmatrix} A_{11} + \Lambda D_{11} & A_{12} \\ & A_{22} \end{bmatrix}, \tag{3.3a}$$

$$B_\Omega := B + D\Omega = \begin{bmatrix} B_{11} + D_{11}\Lambda & \\ B_{21} & B_{22} \end{bmatrix}, \tag{3.3b}$$

$$C_\Omega := \Omega D\Omega + A\Omega + \Omega B + C = \begin{bmatrix} \Lambda D_{11}\Lambda + A_{11}\Lambda + \Lambda B_{11} + C_{11} & C_{12} \\ C_{21} & C_{22} \end{bmatrix}. \tag{3.3c}$$

Theorem 3.2 says that D_{11} is diagonal and nonnegative. Let

$$\mathbb{I}_0 = \{i \,:\, [D_{11}]_{(i,i)} = 0\},\ \mathbb{I}_+ = \{i \,:\, [D_{11}]_{(i,i)} > 0\}. \tag{3.4}$$

The next theorem states sufficient and necessary conditions for ARE (1.1) to be a shifted MARE.

Theorem 3.3 *Suppose that A, B, C, and D are as described in Theorem 3.2.* ARE (1.1) *is a shifted* MARE *of index p if and only if*

$$W_0 := \begin{array}{c} \\ m-p \\ p \\ n-p \end{array} \begin{array}{c} \begin{array}{ccc} m-p & p & n-p \end{array} \\ \begin{bmatrix} B_{22} & B_{21}D_{11} & -D_{22} \\ -C_{12} & A_{11}B_{11} - C_{11}D_{11} & A_{12} \\ -C_{22} & -C_{21}D_{11} & A_{22} \end{bmatrix} \end{array} \tag{3.5}$$

is either a nonsingular M-matrix with

$$[B_{11}]_{(i,i)} > 0 \quad for\ i \in \mathbb{I}_0, \tag{3.6}$$

or an irreducible singular M-matrix with $\mathbb{I}_0 = \emptyset$. Moreover, when ARE (1.1) *is a shifted* MARE *of index p, the corresponding Λ_0 in* Definition 1.1 *can be given by* (3.13) *below.*

Proof Suppose that ARE (1.1) is a shifted MARE. Then W_Ω defined as in (1.6) can be written as

$$W_\Omega = \begin{array}{c} \\ p \\ m-p \\ p \\ n-p \end{array} \begin{array}{c} \begin{array}{cccc} p & m-p & p & n-p \end{array} \\ \begin{bmatrix} B_{11} + D_{11}\Lambda & 0 & -D_{11} & 0 \\ B_{21} & B_{22} & 0 & -D_{22} \\ -\Lambda D_{11}\Lambda - A_{11}\Lambda - \Lambda B_{11} - C_{11} & -C_{12} & A_{11} + \Lambda D_{11} & A_{12} \\ -C_{21} & -C_{22} & 0 & A_{22} \end{bmatrix} \end{array}, \tag{3.7}$$

and it is a nonsingular M-matrix, or an irreducible singular M-matrix for all Λ such that $\Lambda_{(i,i)} > [\Lambda_0]_{(i,i)}$ for $1 \le i \le p$ for some Λ_0 (see Definition 1.1). By Proposition 2.1(a), $B_{11} + D_{11}\Lambda$ M-matrix and thus it must be diagonal with positive diagonal entries. Again by Proposition 2.1(a), the Schur complement $\widehat{W}_\Omega$ of $B_{11} + D_{11}\Lambda$ in W_Ω,

$$\widehat{W}_\Omega = \begin{bmatrix} B_{22} & B_{21}(B_{11} + D_{11}\Lambda)^{-1}D_{11} & -D_{22} \\ -C_{12} & A_{11} - (C_{11} + A_{11}\Lambda)(B_{11} + D_{11}\Lambda)^{-1}D_{11} & A_{12} \\ -C_{22} & -C_{21}(B_{11} + D_{11}\Lambda)^{-1}D_{11} & A_{22} \end{bmatrix}, \tag{3.8}$$

is a nonsingular M-matrix, or an irreducible singular M-matrix. Notice that $(B_{11} + D_{11}\Lambda)^{-1}D_{11} = D_{11}(B_{11} + D_{11}\Lambda)^{-1}$ because D_{11} is also diagonal. Thus

$$\widehat{W}_\Omega = W_0 \begin{bmatrix} I_{n-p} & 0 & 0 \\ 0 & (B_{11} + D_{11}\Lambda)^{-1} & 0 \\ 0 & 0 & I_{m-p} \end{bmatrix}. \tag{3.9}$$

Therefore W_0 is a nonsingular M-matrix when W_Ω is nonsingular, or an irreducible singular M-matrix when W_Ω is irreducible and singular. That $B_{11} + D_{11}\Lambda$ is diagonal with positive diagonal entries immediately leads to (3.6). When W_Ω is irreducible, each diagonal entry of D_{11} is positive, i.e., $\mathbb{I}_0 = \emptyset$; otherwise W_Ω would be reducible, a contradiction.

suppose that W_0 of (3.5) is either a nonsingular M-matrix with (3.6) or an irreducible singular M-matrix with $\mathbb{I}_0 = \emptyset$. By Proposition 2.1(a), $A_{11}B_{11} - C_{11}D_{11}$ is a nonsingular M-matrix, and thus we have $[A_{11}]_{(i,i)} > 0$ for $i \in \mathbb{I}_0$. As a consequence,

$$[A_{11}]_{(i,i)} + [B_{11}]_{(i,i)} > 0 \quad \text{for } i \in \mathbb{I}_0.$$

Now we look for diagonal Λ such that

$$[\Lambda D_{11}\Lambda + A_{11}\Lambda + \Lambda B_{11} + C_{11}]_{(i,i)} > 0, \quad [B_{11} + D_{11}\Lambda]_{(i,i)} > 0. \tag{3.10}$$

The inequalities in (3.10) are equivalent to

$$\Lambda_{(i,i)}\Big([A_{11}]_{(i,i)} + [B_{11}]_{(i,i)}\Big) + [C_{11}]_{(i,i)} > 0 \quad \text{for } i \in \mathbb{I}_0, \tag{3.11a}$$

and for $i \in \mathbb{I}_+$

$$\Lambda^2_{(i,i)}[D_{11}]_{(i,i)} + \Lambda_{(i,i)}\Big([A_{11}]_{(i,i)} + [B_{11}]_{(i,i)}\Big) + [C_{11}]_{(i,i)} > 0, \tag{3.11b}$$

$$[B_{11}]_{(i,i)} + [D_{11}]_{(i,i)}\Lambda_{(i,i)} > 0. \tag{3.11c}$$

Moments ago, we mentioned that $A_{11}B_{11} - C_{11}D_{11}$ is a nonsingular M-matrix and thus it has positive diagonal entries, i.e.,

$$[A_{11}]_{(i,i)}[B_{11}]_{(i,i)} - [C_{11}]_{(i,i)}[D_{11}]_{(i,i)} > 0,$$

noting that A_{11}, B_{11}, and D_{11} are diagonal. By

$$[A_{11}]^2_{(i,i)} + [B_{11}]^2_{(i,i)} \ge 2[A_{11}]_{(i,i)}[B_{11}]_{(i,i)},$$

we find for $i \in \mathbb{I}_+$

$$\begin{aligned} \Delta_i &:= \Big([A_{11}]_{(i,i)} + [B_{11}]_{(i,i)}\Big)^2 - 4[C_{11}]_{(i,i)}[D_{11}]_{(i,i)} \\ &= [A_{11}]^2_{(i,i)} + [B_{11}]^2_{(i,i)} + 2[A_{11}]_{(i,i)}[B_{11}]_{(i,i)} - 4[C_{11}]_{(i,i)}[D_{11}]_{(i,i)} \\ &\ge 4[A_{11}]_{(i,i)}[B_{11}]_{(i,i)} - 4[C_{11}]_{(i,i)}[D_{11}]_{(i,i)} \end{aligned} \tag{3.12}$$

$$
\begin{aligned}
&= 4\Big(\big[A_{11}\big]_{(i,i)}\big[B_{11}\big]_{(i,i)} - \big[C_{11}\big]_{(i,i)}\big[D_{11}\big]_{(i,i)}\Big)\\
&> 0.
\end{aligned}
$$

It can be verified that any Λ such that $\Lambda_{(i,i)} > [\Lambda_0]_{(i,i)}$ for $1 \le i \le p$ satisfies (3.10), where

$$
\big[\Lambda_0\big]_{(i,i)} = \begin{cases} \max\left\{ -\dfrac{\big[B_{11}\big]_{(i,i)}}{\big[D_{11}\big]_{(i,i)}},\ \dfrac{-\Big(\big[A_{11}\big]_{(i,i)}+\big[B_{11}\big]_{(i,i)}\Big)+\sqrt{\Delta_i}}{2\big[D_{11}\big]_{(i,i)}} \right\} & \text{for } i \in \mathbb{I}_+, \\ \dfrac{-\big[C_{11}\big]_{(i,i)}}{\big[A_{11}\big]_{(i,i)}+\big[B_{11}\big]_{(i,i)}} & \text{for } i \in \mathbb{I}_0. \end{cases} \tag{3.13}
$$

Note that we still have (3.9). When $\Lambda_{(i,i)} > [\Lambda_0]_{(i,i)}$ for $1 \le i \le p$, $\widehat{W}_\Omega$ is a Z-matrix. Now if W_0 is a nonsingular M-matrix, then so is $\widehat{W}_\Omega$ and thus W_Ω too. If W_0 is an irreducible singular M-matrix with $\mathbb{I}_0 = \emptyset$, then it can be verified that $\widehat{W}_\Omega$ and thus W_Ω are singular M-matrices. It remains to show that W_Ω is irreducible. To this end, define for $\epsilon > 0$,

$$
W_{\Omega,\epsilon} := W_\Omega + \begin{matrix} \\ p \\ m+n-p \end{matrix} \begin{matrix} \begin{matrix} p & m+n-p \end{matrix} \\ \begin{bmatrix} 0 & 0 \\ 0 & \epsilon I \end{bmatrix} \end{matrix}, \quad \widehat{W}_{\Omega,\epsilon} := \widehat{W}_\Omega + \epsilon I. \tag{3.14}
$$

It can be seen that $\widehat{W}_{\Omega,\epsilon}$ happens to be the Schur complement of $B_{11} + D_{11}\Lambda$ in $W_{\Omega,\epsilon}$. Thus $\widehat{W}_{\Omega,\epsilon}$ is an irreducible and nonsingular M-matrix, which yields $\widehat{W}^{-1}_{\Omega,\epsilon} > 0$ by Proposition 2.1(b). Noting that D_{11} has positive diagonal entries and the nonnegative matrix $\Lambda D_{11}\Lambda + A_{11}\Lambda + \Lambda B_{11} + C_{11}$ (see (3.10) above) also has a positive diagonal, we have

$$
\begin{aligned}
U_{\Omega,\epsilon} &:= (B_{11} + D_{11}\Lambda)^{-1}\big[0, D_{11}, 0\big]\widehat{W}^{-1}_{\Omega,\epsilon} > 0,\\
V_{\Omega,\epsilon} &:= \widehat{W}^{-1}_{\Omega,\epsilon}\begin{bmatrix} -B_{21} \\ \Lambda D_{11}\Lambda + A_{11}\Lambda + \Lambda B_{11} + C_{11} \\ C_{21} \end{bmatrix}(B_{11} + D_{11}\Lambda)^{-1} > 0,\\
Z_{\Omega,\epsilon} &:= (B_{11} + D_{11}\Lambda)^{-1} + (B_{11} + D_{11}\Lambda)^{-1}[0, D_{11}, 0]V_{\Omega,\epsilon} > 0,
\end{aligned}
$$

which yield

$$
W^{-1}_{\Omega,\epsilon} = \begin{bmatrix} Z_{\Omega,\epsilon} & U_{\Omega,\epsilon} \\ V_{\Omega,\epsilon} & \widehat{W}^{-1}_{\Omega,\epsilon} \end{bmatrix} > 0.
$$

By Proposition 2.1(b), $W_{\Omega,\epsilon}$ and thus W_Ω by (3.14) are irreducible. □

Corollary 3.1 *Suppose that A, B, C, and D are as described in Theorem 3.2, and* ARE (1.1) *is a shifted* MARE *of index p. If W_0 is an irreducible M-matrix, then D_{11} is diagonal and has positive diagonal entries, and $\mathbb{I}_0 = \emptyset$.*

Proof It is a consequence of (3.7) in the proof of Theorem 3.3. □

3.2 Extremal Solution

A *shifted* MARE may admit more than one solution. We now specify the solution that is of interest to us.

Definition 3.1 Suppose that ARE (1.1) is a shifted MARE of index p with A, B, C, and D as described in Theorem 3.2, and let Ω be as in (1.4) with diagonal Λ having sufficiently

Algorithm 3.1 Compute shift Λ_0 for shifted MARE (1.1)

Require: A, B, C, and D as described in Theorem 3.2 and assume that ARE (1.1) is a shifted MARE of index p (i.e., satisfying the conditions in Theorem 3.3);
Ensure: $\Lambda_0 \in \mathbb{R}^{p\times p}$ as described in Definition 1.1.
1: determine $\mathbb{I}_0$ and $\mathbb{I}_+$ by (3.4);
2: compute Δ_i for $i \in \mathbb{I}_+$ according to (3.12);
3: compute Λ_0 according to (3.13);
4: **return** Λ_0.

large diagonal entries such that ARE (1.5) is an MARE whose minimal nonnegative solution is denoted by Φ_Ω. We call $\Phi = \Phi_\Omega - \Omega$ the *extremal solution* to the shifted MARE (1.1).

As this definition stands, Φ is a function of the shifting matrix Ω, and thus potentially it may not be well-defined as stated. The next theorem settles the issue.

Theorem 3.4 *The extreme solution Φ defined in Definition 3.1 is independent of Ω.*

Proof Recall W defined as in (1.2) and W_Ω defined as in (1.6). Let

$$\mathscr{H} = \begin{bmatrix} I_m & \\ & -I_n \end{bmatrix} W, \quad \mathscr{H}_\Omega = \begin{bmatrix} I_m & \\ & -I_n \end{bmatrix} W_\Omega. \tag{3.15}$$

It can be verified that

$$\begin{bmatrix} I_m & \\ \Omega & I_n \end{bmatrix}^{-1} \mathscr{H} \begin{bmatrix} I_m & \\ -\Omega & I_n \end{bmatrix} = \mathscr{H}_\Omega. \tag{3.16}$$

Hence $\mathscr{H}$ and $\mathscr{H}_\Omega$ are similar and thus have the same eigenvalues, and there is one-one correspondence between the eigenspaces of $\mathscr{H}$ and $\mathscr{H}_\Omega$ as well. According to Proposition 2.2, $\begin{bmatrix} I_m \\ \Phi_\Omega \end{bmatrix}$ is a basis matrix of the unique m-dimensional eigenspace associated with the eigenvalues of $\mathscr{H}_\Omega$ in $\mathbb{C}_{0+}$, and, correspondingly, $\begin{bmatrix} I_m \\ \Phi \end{bmatrix}$ is a basis matrix of the unique m-dimensional eigenspace associated with the eigenvalues of $\mathscr{H}$ in $\mathbb{C}_{0+}$. Since $\mathscr{H}$ is independent of Ω, so must be Φ. □

4 Highly Accurate Doubling Algorithm for Φ_Ω

To set the stage, we will assume that, throughout the rest of this section, ARE (1.1) is a shifted MARE of index p, with A, B, C, and D as described in Theorem 3.2, Φ is its extremal solution defined by Definition 3.1, and Ω is a shifting matrix such that W_Ω given by (1.6) is a nonsingular or an irreducible singular M-matrix.

Partition Φ as

$$\Phi = \begin{array}{c} \\ p \\ n-p \end{array} \begin{array}{c} \begin{array}{cc} p & m-p \end{array} \\ \begin{bmatrix} \Phi_{11} & \Phi_{12} \\ \Phi_{21} & \Phi_{22} \end{bmatrix} \end{array},$$

similarly to (1.3). In the context of the applications detailed in "Appendix 1", it is $\Phi_{(:,p+1:m)} = \begin{bmatrix} \Phi_{12} \\ \Phi_{22} \end{bmatrix}$ and $A - \Phi D$ that are required to be computed accurately. For Λ with sufficiently large diagonal entries,

$$\Phi_\Omega = \Phi + \Omega = \begin{bmatrix} \Phi_{11} + \Lambda & \Phi_{12} \\ \Phi_{21} & \Phi_{22} \end{bmatrix}$$

is the minimal nonnegative solution to MARE (1.5), and thus it can be computed by the doubling algorithm ADDA [28] or preferably its highly accurate version accADDA developed in [23,29]. accADDA can compute Φ_Ω in a cancellation-free manner to high entrywise relative accuracy, provided a triplet representation of the M-matrix W_Ω is available with high entrywise relative accuracy.

We outline ADDA as follows [16,28,29]. Choose two parameters α and β satisfying

$$0 \le \alpha \le \alpha_{\rm opt} := \big(\max_i [A_\Omega]_{(i,i)}\big)^{-1},\ 0 \le \beta \le \beta_{\rm opt} := \big(\max_i [B_\Omega]_{(i,i)}\big)^{-1}, \tag{4.1a}$$

$$\max\{\alpha, \beta\} \ne 0. \tag{4.1b}$$

ADDA converges the fastest with $\alpha = \alpha_{\rm opt}$ and $\beta = \beta_{\rm opt}$ [28, Theorem 3.3]. Initially $E_0 \in \mathbb{R}^{m\times m}$, $F_0 \in \mathbb{R}^{n\times n}$, $X_0 \in \mathbb{R}^{n\times m}$ and $Y_0 \in \mathbb{R}^{m\times n}$ are determined by the linear matrix equation

$$\begin{bmatrix} \alpha B_\Omega + I & -\beta D \\ -\alpha C_\Omega & \beta A_\Omega + I \end{bmatrix} \begin{bmatrix} E_0 & Y_0 \\ X_0 & F_0 \end{bmatrix} = \begin{bmatrix} I - \beta B_\Omega & \alpha D \\ \beta C_\Omega & I - \alpha A_\Omega \end{bmatrix}, \tag{4.2}$$

and for $k \ge 0$, E_{k+1}, F_{k+1}, X_{k+1} and Y_{k+1} are then iteratively computed by the doubling iterative formulas:

$$E_{k+1} = E_k(I_m - Y_k X_k)^{-1} E_k, \tag{4.3a}$$

$$F_{k+1} = F_k(I_n - X_k Y_k)^{-1} F_k, \tag{4.3b}$$

$$X_{k+1} = X_k + F_k(I_n - X_k Y_k)^{-1} X_k E_k, \tag{4.3c}$$

$$Y_{k+1} = Y_k + E_k(I_m - Y_k X_k)^{-1} Y_i F_k, \tag{4.3d}$$

until convergence. A detailed derivation of these formulas can be found in [16, pp. 20–21].

With α and β satisfying (4.1), X_k is monotonically increasing and converges to Φ_Ω quadratically, except in the case when $\mathscr{H}_\Omega$ has a double eigenvalue 0 coming from a 2×2 Jordan block, the convergence is linear with the linear rate $1/2$. For more detailed statements of ADDA's convergence, the reader is referred to [16, Theorem 6.3].

In general, ADDA can solve any shifted MARE very efficiently, but if implemented as is written, cancellations can potentially cause some of the tiny entries of Φ_Ω to be much less accurate than its largest entries. To avoid those cancellations, we may have to implement it differently, i.e., use accADDA [23,29] instead, which needs a triplet representation of W_Ω to begin with. In what follows, we describe an alternative version that is essentially the same as accADDA there. In this alternative version, we rely on a left triplet representation of W_Ω. A major reason for the change is to also have $A - \Phi D$ computed with high entrywise relative accuracy at the end. For that purpose, we will need a couple of assumptions as stated in Assumption 4.1 below.

Assumption 4.1 (i) *The diagonal matrices A_{11}, B_{11} and D_{11} are accurately known with high entrywise relative accuracy;*

(ii) *A left triplet representation of W_0 in (3.5) is known. Specifically, $W_0 = (\text{offdiag}(W_0), \boldsymbol{u}, \boldsymbol{v})_{\rm L}$ is known with high entrywise relative accuracy.*

Assumption 4.1(b) requiring a accurate left triplet representation of W_0 reads a bit of unnatural because the second block column is quadratic in A, B, C, and D. However, for the ARE arising from the application in "Appendix 1" that motivates this study, it is rather natural, as explained in Remarks 8.2 and 8.3 there.

With the help of Assumption 4.1, in what follows, we will describe an alternative accADDA.

Consistently with the partition of W_0 in (3.5), we partition $\boldsymbol{u}$ and $\boldsymbol{v}$ as

$$\boldsymbol{u}^{\mathrm{T}} = \begin{bmatrix} \boldsymbol{u}_1^{\mathrm{T}} & \boldsymbol{u}_2^{\mathrm{T}} & \boldsymbol{u}_3^{\mathrm{T}} \end{bmatrix}, \quad \boldsymbol{v}^{\mathrm{T}} = \begin{bmatrix} \boldsymbol{v}_1^{\mathrm{T}} & \boldsymbol{v}_2^{\mathrm{T}} & \boldsymbol{v}_3^{\mathrm{T}} \end{bmatrix}$$

and expand $\boldsymbol{u}^{\mathrm{T}} W_0 = \boldsymbol{v}^{\mathrm{T}}$ as

$$\begin{bmatrix} \boldsymbol{u}_1^{\mathrm{T}} & \boldsymbol{u}_2^{\mathrm{T}} & \boldsymbol{u}_3^{\mathrm{T}} \end{bmatrix} W_0 = \begin{bmatrix} \boldsymbol{v}_1^{\mathrm{T}} & \boldsymbol{v}_2^{\mathrm{T}} & \boldsymbol{v}_3^{\mathrm{T}} \end{bmatrix}. \tag{4.4}$$

Let $\boldsymbol{u}_0 \in \mathbb{R}^p$ be given by

$$\big[\boldsymbol{u}_0\big]_{(i)} = \begin{cases} \dfrac{\big[\boldsymbol{u}_1^{\mathrm{T}}(-B_{21})\big]_{(i)} + \big[\boldsymbol{u}_3^{\mathrm{T}} C_{21}\big]_{(i)} + \big[\boldsymbol{u}_2^{\mathrm{T}}(A_{11}\Lambda + \Lambda B_{11} + C_{11})\big]_{(i)}}{\big[B_{11}\big]_{(i,i)}}, & \text{if } i \in \mathbb{I}_0, \\ \dfrac{\big[A_{11}\big]_{(i,i)} + \Lambda_{(i,i)}}{\big[D_{11}\big]_{(i,i)}} \big[\boldsymbol{u}_2\big]_{(i)}, & \text{if } i \in \mathbb{I}_+. \end{cases} \tag{4.5}$$

For Λ with sufficiently large diagonal entries, we can make $\boldsymbol{u}_0 > 0$ since $\big[D_{11}\big]_{(i,i)} > 0$ for $i \in \mathbb{I}_+$ and $\big[A_{11}\big]_{(i,i)} + \big[B_{11}\big]_{(i,i)} > 0$ for $i \in \mathbb{I}_0$ and $-B_{21} \geq 0$. Moreover, catastrophic cancellations can be completely avoided when calculating $A_{11} + \Lambda D_{11}$ and $A_{11}\Lambda + \Lambda B_{11} + C_{11}$ so that $\boldsymbol{u}_0$ can be obtained entrywise accurately. Next, we define $\boldsymbol{v}_0,\ \hat{\boldsymbol{v}}_0 \in \mathbb{R}^p$ by

$$\big[\boldsymbol{v}_0\big]_{(i)} = \begin{cases} 0, & \text{if } i \in \mathbb{I}_0, \\ \big[\boldsymbol{v}_2\big]_{(i)} \Big/ \big[D_{11}\big]_{(i,i)}, & \text{if } i \in \mathbb{I}_+, \end{cases} \quad \big[\hat{\boldsymbol{v}}_0\big]_{(i)} = \begin{cases} \big[\boldsymbol{u}_2^{\mathrm{T}} A_{11}\big]_{(i)}, & \text{if } i \in \mathbb{I}_0, \\ 0, & \text{if } i \in \mathbb{I}_+. \end{cases}$$

It is straightforward to check that

$$\begin{bmatrix} \boldsymbol{u}_0^{\mathrm{T}} & \boldsymbol{u}_1^{\mathrm{T}} & \boldsymbol{u}_2^{\mathrm{T}} & \boldsymbol{u}_3^{\mathrm{T}} \end{bmatrix} W_\Omega = \begin{bmatrix} \boldsymbol{v}_0^{\mathrm{T}} & \boldsymbol{v}_1^{\mathrm{T}} & \hat{\boldsymbol{v}}_0^{\mathrm{T}} & \boldsymbol{v}_3^{\mathrm{T}} \end{bmatrix} \geq 0 \tag{4.6}$$

which leads to an entrywise relatively accurate left triplet representation for W_Ω, as needed.

Remark 4.1 According to Theorem 3.3, when W_0 is an irreducible singular M-matrix, we have $\mathbb{I}_0 = \emptyset$ and $\boldsymbol{v} = \boldsymbol{0}$ and, thus, it follows from (4.5) that

$$\boldsymbol{u}_0 = D_{11}^{-1}(A_{11} + \Lambda D_{11})\boldsymbol{u}_2, \quad \boldsymbol{v}_0 = 0, \quad \hat{\boldsymbol{v}}_0 = 0. \tag{4.7}$$

Return to (4.2) that determines E_0, F_0, X_0, and Y_0. First, we notice that

$$\begin{bmatrix} \alpha B_\Omega + I & -\beta D \\ -\alpha C_\Omega & \beta A_\Omega + I \end{bmatrix} = I + W_\Omega \begin{bmatrix} \alpha I & \\ & \beta I \end{bmatrix} \tag{4.8}$$

is a nonsingular M-matrix for any $\alpha,\ \beta \geq 0$. Using (4.6), we conclude

$$\begin{bmatrix} \boldsymbol{u}_0 \\ \boldsymbol{u}_1 \\ \boldsymbol{u}_2 \\ \boldsymbol{u}_3 \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} \alpha B_\Omega + I & -\beta D \\ -\alpha C_\Omega & \beta A_\Omega + I \end{bmatrix} = \begin{bmatrix} \boldsymbol{u}_0 + \alpha \boldsymbol{v}_0 \\ \boldsymbol{u}_1 + \alpha \boldsymbol{v}_1 \\ \boldsymbol{u}_2 + \beta \hat{\boldsymbol{v}}_0 \\ \boldsymbol{u}_3 + \beta \boldsymbol{v}_3 \end{bmatrix}^{\mathrm{T}}, \tag{4.9}$$

yielding a left triplet representation for the matrix in (4.8).

The results in Lemma 4.1 below are essentially the ones in [29, Theorem 3.2], except for $k = 0$, the case for which we will provide a proof.

Lemma 4.1 *Let $E_0 \in \mathbb{R}^{m\times m}$, $F_0 \in \mathbb{R}^{n\times n}$, $X_0 \in \mathbb{R}^{n\times m}$ and $Y_0 \in \mathbb{R}^{m\times n}$ be as in (4.2) with α and β satisfying (4.1). Then*

$$\begin{bmatrix} \boldsymbol{u}_0^{\mathrm{T}} & \boldsymbol{u}_1^{\mathrm{T}} & \boldsymbol{u}_2^{\mathrm{T}} & \boldsymbol{u}_3^{\mathrm{T}} \end{bmatrix} \begin{bmatrix} E_k & Y_k \\ X_k & F_k \end{bmatrix} \leq \begin{bmatrix} \boldsymbol{u}_0^{\mathrm{T}} & \boldsymbol{u}_1^{\mathrm{T}} & \boldsymbol{u}_2^{\mathrm{T}} & \boldsymbol{u}_3^{\mathrm{T}} \end{bmatrix} \quad \textit{for all } k \geq 0. \tag{4.10}$$

In particular, if $\boldsymbol{v}^{\mathrm{T}} \equiv \begin{bmatrix}\boldsymbol{v}_1^{\mathrm{T}} & \boldsymbol{v}_2^{\mathrm{T}} & \boldsymbol{v}_3^{\mathrm{T}}\end{bmatrix} = 0$, *then*

$$\begin{bmatrix}\boldsymbol{u}_0^{\mathrm{T}} & \boldsymbol{u}_1^{\mathrm{T}} & \boldsymbol{u}_2^{\mathrm{T}} & \boldsymbol{u}_3^{\mathrm{T}}\end{bmatrix}\begin{bmatrix}E_k & Y_k \\ X_k & F_k\end{bmatrix} = \begin{bmatrix}\boldsymbol{u}_0^{\mathrm{T}} & \boldsymbol{u}_1^{\mathrm{T}} & \boldsymbol{u}_2^{\mathrm{T}} & \boldsymbol{u}_3^{\mathrm{T}}\end{bmatrix} \quad \textit{for all } k \geq 0. \tag{4.11}$$

Proof Only the case for $k = 0$ needs a proof. It can be verified that

$$I - \begin{bmatrix}\beta I & \\ & \alpha I\end{bmatrix} W_\Omega = \left(I + \begin{bmatrix}\alpha I & \\ & \beta I\end{bmatrix} W_\Omega\right) - (\alpha + \beta) W_\Omega.$$

Upon using (4.6), we find

$$\begin{aligned}&\begin{bmatrix}\boldsymbol{u}_0^{\mathrm{T}} & \boldsymbol{u}_1^{\mathrm{T}} & \boldsymbol{u}_2^{\mathrm{T}} & \boldsymbol{u}_3^{\mathrm{T}}\end{bmatrix}\left(I - \begin{bmatrix}\beta I & \\ & \alpha I\end{bmatrix} W_\Omega\right) \\ &= \begin{bmatrix}\boldsymbol{u}_0^{\mathrm{T}} & \boldsymbol{u}_1^{\mathrm{T}} & \boldsymbol{u}_2^{\mathrm{T}} & \boldsymbol{u}_3^{\mathrm{T}}\end{bmatrix}\left(I + \begin{bmatrix}\alpha I & \\ & \beta I\end{bmatrix} W_\Omega\right) - (\alpha + \beta)\begin{bmatrix}\boldsymbol{v}_0^{\mathrm{T}} & \boldsymbol{v}_1^{\mathrm{T}} & \hat{\boldsymbol{v}}_0^{\mathrm{T}} & \boldsymbol{v}_3^{\mathrm{T}}\end{bmatrix}.\end{aligned} \tag{4.12}$$

It follows from (4.2) that

$$\begin{bmatrix}E_0 & Y_0 \\ X_0 & F_0\end{bmatrix} = \left(I - \begin{bmatrix}\beta I & \\ & \alpha I\end{bmatrix} W_\Omega\right)\left(I + \begin{bmatrix}\alpha I & \\ & \beta I\end{bmatrix} W_\Omega\right)^{-1}.$$

Therefore making use of (4.12), we get

$$\begin{aligned}&\begin{bmatrix}\boldsymbol{u}_0^{\mathrm{T}} & \boldsymbol{u}_1^{\mathrm{T}} & \boldsymbol{u}_2^{\mathrm{T}} & \boldsymbol{u}_3^{\mathrm{T}}\end{bmatrix}\begin{bmatrix}E_0 & Y_0 \\ X_0 & F_0\end{bmatrix} \\ &= \begin{bmatrix}\boldsymbol{u}_0^{\mathrm{T}} & \boldsymbol{u}_1^{\mathrm{T}} & \boldsymbol{u}_2^{\mathrm{T}} & \boldsymbol{u}_3^{\mathrm{T}}\end{bmatrix} - (\alpha + \beta)\begin{bmatrix}\boldsymbol{v}_0^{\mathrm{T}} & \boldsymbol{v}_1^{\mathrm{T}} & \hat{\boldsymbol{v}}_0^{\mathrm{T}} & \boldsymbol{v}_3^{\mathrm{T}}\end{bmatrix}\left(I + \begin{bmatrix}\alpha I & \\ & \beta I\end{bmatrix} W_\Omega\right)^{-1} \\ &\leq \begin{bmatrix}\boldsymbol{u}_0^{\mathrm{T}} & \boldsymbol{u}_1^{\mathrm{T}} & \boldsymbol{u}_2^{\mathrm{T}} & \boldsymbol{u}_3^{\mathrm{T}}\end{bmatrix}\end{aligned} \tag{4.13}$$

because the matrix to be inverted in (4.13) is a nonsingular M-matrix. This gives (4.10) for $k = 0$. In particular, if $\boldsymbol{v} = \boldsymbol{0}$, plug (4.7) into (4.13) to get (4.11) for $k = 0$. □

To implement the doubling algorithm in a cancellation-free manner as inspired by [29], we introduce auxiliary vectors, for $k \geq 0$,

$$[\boldsymbol{w}^{(k)}]^{\mathrm{T}} \equiv \begin{bmatrix}\boldsymbol{w}_1^{(k)} \\ \boldsymbol{w}_2^{(k)} \\ \boldsymbol{w}_3^{(k)} \\ \boldsymbol{w}_4^{(k)}\end{bmatrix}^{\mathrm{T}} := \begin{bmatrix}\boldsymbol{u}_0 \\ \boldsymbol{u}_1 \\ \boldsymbol{u}_2 \\ \boldsymbol{u}_3\end{bmatrix}^{\mathrm{T}} - \begin{bmatrix}\boldsymbol{u}_0 \\ \boldsymbol{u}_1 \\ \boldsymbol{u}_2 \\ \boldsymbol{u}_3\end{bmatrix}^{\mathrm{T}}\begin{bmatrix}E_k & Y_k \\ X_k & F_k\end{bmatrix}. \tag{4.14}$$

They are nonnegative in general and exactly zero if $\boldsymbol{v} = \boldsymbol{0}$, as ensured by Lemma 4.1. It can be seen that

$$[\boldsymbol{w}^{(0)}]^{\mathrm{T}}\begin{bmatrix}I + \alpha B_\Omega & -\alpha D \\ -\beta C_\Omega & I + \beta A_\Omega\end{bmatrix} = (\alpha + \beta)\begin{bmatrix}\boldsymbol{v}_0^{\mathrm{T}} & \boldsymbol{v}_1^{\mathrm{T}} & \hat{\boldsymbol{v}}_0^{\mathrm{T}} & \boldsymbol{v}_3^{\mathrm{T}}\end{bmatrix}, \tag{4.15}$$

using (4.13). The big matrix in the left-hand side of (4.15) is a nonsingular M-matrix whose inverse can be computed with high entrywise relative accuracy by the GTH-like algorithm—Algorithm 2.1, made possible by Lemma 4.2 below.

Lemma 4.2 *Assume* (4.1) *holds.*

(a) *If* $\alpha = 0$ *but* $\beta > 0$, *then*

$$\begin{bmatrix} I & 0 \\ -\beta C_\Omega & I + \beta A_\Omega \end{bmatrix}^{-1} = \begin{bmatrix} I & 0 \\ \beta(I + \beta A_\Omega)^{-1} C_\Omega & (I + \beta A_\Omega)^{-1} \end{bmatrix}$$

and a left triplet representation for $I + \beta A_\Omega$ *can be read off from*

$$\begin{bmatrix} \boldsymbol{u}_2 \\ \boldsymbol{u}_3 \end{bmatrix}^{\mathrm{T}} (I + \beta A_\Omega) = \beta \begin{bmatrix} \hat{\boldsymbol{v}}_0^{\mathrm{T}} & \boldsymbol{v}_3^{\mathrm{T}} \end{bmatrix} + \begin{bmatrix} \boldsymbol{u}_2 \\ \boldsymbol{u}_3 \end{bmatrix}^{\mathrm{T}} + \beta \begin{bmatrix} \boldsymbol{u}_0 \\ \boldsymbol{u}_1 \end{bmatrix}^{\mathrm{T}} D.$$

(b) *If* $\beta = 0$ *but* $\alpha > 0$, *then*

$$\begin{bmatrix} I + \alpha B_\Omega & -\alpha D \\ 0 & I \end{bmatrix}^{-1} = \begin{bmatrix} (I + \alpha B_\Omega)^{-1} & \alpha(I + \alpha B_\Omega)^{-1} D \\ 0 & I \end{bmatrix}$$

and a left triplet representation for $I + \alpha B_\Omega$ *can be read off from*

$$\begin{bmatrix} \boldsymbol{u}_0 \\ \boldsymbol{u}_3 \end{bmatrix}^{\mathrm{T}} (I + \alpha B_\Omega) = \alpha \begin{bmatrix} \boldsymbol{v}_0^{\mathrm{T}} & \boldsymbol{v}_1^{\mathrm{T}} \end{bmatrix} + \alpha \begin{bmatrix} \boldsymbol{u}_2 \\ \boldsymbol{u}_3 \end{bmatrix}^{\mathrm{T}} C_\Omega + \begin{bmatrix} \boldsymbol{u}_0 \\ \boldsymbol{u}_1 \end{bmatrix}^{\mathrm{T}}.$$

(c) *If* $\alpha > 0$ *and* $\beta > 0$, *then*

$$\begin{bmatrix} (1/\alpha)\boldsymbol{u}_0 \\ (1/\alpha)\boldsymbol{u}_1 \\ (1/\beta)\boldsymbol{u}_2 \\ (1/\beta)\boldsymbol{u}_3 \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} I + \alpha B_\Omega & -\alpha D \\ -\beta C_\Omega & I + \beta A_\Omega \end{bmatrix} = \begin{bmatrix} (1/\alpha)\boldsymbol{u}_0 \\ (1/\alpha)\boldsymbol{u}_1 \\ (1/\beta)\boldsymbol{u}_2 \\ (1/\beta)\boldsymbol{u}_3 \end{bmatrix}^{\mathrm{T}} + \begin{bmatrix} \boldsymbol{v}_0^{\mathrm{T}} \\ \boldsymbol{v}_1 \\ \hat{\boldsymbol{v}}_0^{\mathrm{T}} \\ \boldsymbol{v}_3 \end{bmatrix}^{\mathrm{T}}.$$

Starting from $\boldsymbol{w}^{(0)}$, we can compute $\boldsymbol{w}^{(k)}$ recursively as follows. Notice that

$$\begin{bmatrix} E_{k+1} & Y_{k+1} \\ X_{k+1} & F_{k+1} \end{bmatrix} = \begin{bmatrix} E_k & Y_k \\ X_k & F_k \end{bmatrix} - \left(I - \begin{bmatrix} E_k & Y_k \\ X_k & F_k \end{bmatrix} \right) \begin{bmatrix} I_m & -Y_k \\ -X_k & I_n \end{bmatrix}^{-1} \begin{bmatrix} E_k & 0 \\ 0 & F_k \end{bmatrix}$$

to get

$$\begin{aligned} [\boldsymbol{w}^{(k+1)}]^{\mathrm{T}} &= [\boldsymbol{w}^{(k)}]^{\mathrm{T}} + [\boldsymbol{w}^{(k)}]^{\mathrm{T}} \begin{bmatrix} I & -Y_k \\ -X_k & I \end{bmatrix}^{-1} \begin{bmatrix} E_k & 0 \\ 0 & F_k \end{bmatrix} \\ &= [\boldsymbol{w}^{(k)}]^{\mathrm{T}} + [\boldsymbol{w}^{(k)}]^{\mathrm{T}} \begin{bmatrix} (I - Y_k X_k)^{-1} E_k & Y_k (I - X_k Y_k)^{-1} F_k \\ X_k (I - Y_k X_k)^{-1} E_k & (I - X_k Y_k)^{-1} F_k \end{bmatrix}^{-1}. \end{aligned} \tag{4.16}$$

As in [29], we can show that

$$\begin{aligned} \begin{bmatrix} \boldsymbol{u}_2 \\ \boldsymbol{u}_3 \end{bmatrix}^{\mathrm{T}} (I - X_k Y_k) &= \begin{bmatrix} \boldsymbol{w}_3^{(k)} \\ \boldsymbol{w}_4^{(k)} \end{bmatrix}^{\mathrm{T}} + \begin{bmatrix} \boldsymbol{u}_2 \\ \boldsymbol{u}_3 \end{bmatrix}^{\mathrm{T}} F_k + \left(\begin{bmatrix} \boldsymbol{u}_0 \\ \boldsymbol{u}_1 \end{bmatrix}^{\mathrm{T}} E_k + \begin{bmatrix} \boldsymbol{w}_1^{(k)} \\ \boldsymbol{w}_2^{(k)} \end{bmatrix}^{\mathrm{T}} \right) Y_k \\ &=: \begin{bmatrix} \boldsymbol{u}_1^{(k)} \\ \boldsymbol{u}_2^{(k)} \end{bmatrix}^{\mathrm{T}}, \end{aligned} \tag{4.17a}$$

Algorithm 4.1 Highly accurate doubling algorithm for shifted MARE (1.1)

Require: A, B, C, and D as described in Theorem 3.2 and a left triplet representation for W_0 in (3.5), $1 \le p \le \min\{m, n\}$;
Ensure: the extremal solution Φ to shifted MARE (1.1) of index p.
1: compute Λ_0 by Algorithm 3.1;
2: let $\Lambda = 1.01 \times \Lambda_0$, and formally $\Omega = \mathrm{diag}(\Lambda, 0)$ as in (1.4);
3: compute A_Ω, B_Ω, and C_Ω according to (1.7);
4: compute a left triplet representation of W_Ω according to (4.4) – (4.6);
5: compute E_0, F_0, X_0 and Y_0 according to (4.2) by the GTH-like algorithm – Algorithm 2.1 – using the left triplet representation yielded by (4.9) for the coefficient matrix;
6: compute $\boldsymbol{w}^{(0)}$ according to (4.15) by the GTH-like algorithm;
7: $k = -1$;
8: **repeat**
9: $k = k + 1$;
10: compute $\boldsymbol{u}_i^{(k)}$ for $1 \le i \le 4$ as defined in (4.17) and generate the left triplet representations for $I - Y_k X_k$ and $I - X_k Y_k$ as in (4.18);
11: compute E_{k+1}, F_{k+1}, X_{k+1} and Y_{k+1} according to (4.3) by the GTH-like algorithm using the left triplet representations (4.18) for $I - Y_k X_k$ and $I - X_k Y_k$;
12: compute $\boldsymbol{w}_j^{(k+1)}$ for $1 \le j \le 4$ according to (4.16) (reuse $E_k(I - Y_k X_k)^{-1}$ and $F_k(I - X_k Y_k)^{-1}$ that appear in implementing line 11 to reduce work);
13: **until** convergence;
14: **return** the last X_k as an approximation to Φ_Ω, and then $\Phi \approx X_k - \Omega$.

and similarly,

$$\begin{aligned}\begin{bmatrix}\boldsymbol{u}_0\\ \boldsymbol{u}_1\end{bmatrix}^{\mathrm{T}}(I - Y_k X_k) &= \begin{bmatrix}\boldsymbol{w}_1^{(k)}\\ \boldsymbol{w}_2^{(k)}\end{bmatrix}^{\mathrm{T}} + \begin{bmatrix}\boldsymbol{u}_0\\ \boldsymbol{u}_1\end{bmatrix}^{\mathrm{T}} E_k + \left(\begin{bmatrix}\boldsymbol{u}_2\\ \boldsymbol{u}_3\end{bmatrix}^{\mathrm{T}} F_k + \begin{bmatrix}\boldsymbol{w}_3^{(k)}\\ \boldsymbol{w}_4^{(k)}\end{bmatrix}^{\mathrm{T}}\right) X_k \\ &=: \begin{bmatrix}\boldsymbol{u}_3^{(k)}\\ \boldsymbol{u}_4^{(k)}\end{bmatrix}^{\mathrm{T}}. \end{aligned} \tag{4.17b}$$

Finally, we obtain the following left triplet representations for $I - X_k Y_k$ and $I - Y_k X_k$,

$$I - X_k Y_k = \left(\mathrm{offdiag}(I - X_k Y_k), \begin{bmatrix}\boldsymbol{u}_2\\ \boldsymbol{u}_3\end{bmatrix}, \begin{bmatrix}\boldsymbol{u}_1^{(k)}\\ \boldsymbol{u}_2^{(k)}\end{bmatrix}\right)_{\mathrm{L}}, \tag{4.18a}$$

$$I - Y_k X_k = \left(\mathrm{offdiag}(I - Y_k X_k), \begin{bmatrix}\boldsymbol{u}_0\\ \boldsymbol{u}_1\end{bmatrix}, \begin{bmatrix}\boldsymbol{u}_3^{(k)}\\ \boldsymbol{u}_4^{(k)}\end{bmatrix}\right)_{\mathrm{L}}. \tag{4.18b}$$

In the same way as in [29], we can implement the doubling algorithm in a cancellation-free manner which is outlined in Algorithm 4.1. It can compute $\Phi_\Omega = \Phi + \Omega$ and, consequently, the extremal solution Φ with high entrywise relative accuracy, except possibly the first p diagonal entries of Φ due to possibly cancellations in $\Phi_{(i,i)} = [\Phi_\Omega]_{(i,i)} - \Lambda_{(i,i)}$ for $1 \le i \le p$.

To stop the iteration: lines 8–13 of Algorithm 4.1, we can simply use the one proposed in [29, section 5] (see also [16, p. 90]):

$$\max_{i,j} \frac{\left([X_{k-1} - X_k]_{(i,j)}\right)^2}{\left([X_{k-2} - X_{k-1}]_{(i,j)} - [X_{k-1} - X_k]_{(i,j)}\right)[X_k]_{(i,j)}} \le \mathtt{rtol} \tag{4.19}$$

for the purpose to achieve $X_{k+1} \approx \Phi_\Omega$ with high entrywise relative accuracy, where `rtol` is a given relative tolerance. We also employ a safeguard by checking the entrywise relative residual [29, section 4] as promoted there, whenever (4.19) is satisfied.

In the next section, we will define an auxiliary set of vectors along with the doubling iteration for the purpose of calculating $A - \Phi D$ with high entrywise relative accuracy. To that end, we will employ an additional stopping criterion to go with (4.19) at the same time.

5 Accurate Computation of $A - \Phi D$

The highly accurate accADDA, Algorithm 4.1, can compute Φ_Ω with high entrywise relative accuracy. Since

$$\Phi_\Omega = \begin{bmatrix} \Phi_{11} + \Lambda & \Phi_{12} \\ \Phi_{21} & \Phi_{22} \end{bmatrix},$$

Φ is computed entrywise accurately, except possibly its first p diagonal entries. The other matrix of interest is $A - \Phi D$ as in the application described in "Appendix 1". In this section we will present a way to compute $A - \Phi D$ without any cancellation.

According to Theorem 3.2,

$$\begin{aligned} A - \Phi D &= \begin{bmatrix} A_{11} & A_{12} \\ & A_{22} \end{bmatrix} - \begin{bmatrix} \Phi_{11} & \Phi_{12} \\ \Phi_{21} & \Phi_{22} \end{bmatrix} \begin{bmatrix} D_{11} & \\ & D_{22} \end{bmatrix} \\ &= \begin{bmatrix} A_{11} - \Phi_{11} D_{11} & A_{12} - \Phi_{12} D_{22} \\ -\Phi_{21} D_{11} & A_{22} - \Phi_{22} D_{22} \end{bmatrix}. \end{aligned} \tag{5.1}$$

Note that Φ and Φ_Ω share the same off-diagonal entries. By examining the signs of the entries of A_{ij} and D_{ij} stated in Theorem 3.2, we find that the off-diagonal entries of $A - \Phi D$ can be computed without any cancellation. In fact, by (5.1),

$$\operatorname{offdiag}(A - \Phi D) = \begin{bmatrix} -\operatorname{offdiag}(\Phi_{11}) D_{11} & A_{12} - \Phi_{12} D_{22} \\ -\Phi_{21} D_{11} & \operatorname{offdiag}(A_{22} - \Phi_{22} D_{22}) \end{bmatrix}. \tag{5.2}$$

It remains to explain how to compute its diagonal entries also in a cancellation-free way. We will do so by finding left triplet representations for the diagonal blocks $A_{11} - \Phi_{11} D_{11}$ and $A_{22} - \Phi_{22} D_{22}$. To this end, we introduce vectors

$$\hat{\boldsymbol{u}}_0 := \boldsymbol{u}_0 - \Lambda \boldsymbol{u}_2, \quad \begin{bmatrix} \widetilde{\boldsymbol{w}}_1 \\ \widetilde{\boldsymbol{w}}_2 \end{bmatrix}^{\mathrm{T}} := \begin{bmatrix} \hat{\boldsymbol{u}}_0 \\ \boldsymbol{u}_1 \end{bmatrix}^{\mathrm{T}} - \begin{bmatrix} \boldsymbol{u}_2 \\ \boldsymbol{u}_3 \end{bmatrix}^{\mathrm{T}} \Phi. \tag{5.3}$$

The vectors $\widetilde{\boldsymbol{w}}_j$ for $j = 1, 2$ are, in fact, nonnegative as guaranteed by Lemma 5.1 below, and $\hat{\boldsymbol{u}}_0$ can be explicitly stated as, upon using (4.5),

$$[\hat{\boldsymbol{u}}_0]_{(i)} = \begin{cases} \dfrac{\left[\boldsymbol{u}_1^{\mathrm{T}}(-B_{21})\right]_{(i)} + \left[\boldsymbol{u}_3^{\mathrm{T}} C_{21}\right]_{(i)} + \left[\boldsymbol{u}_2^{\mathrm{T}}(A_{11}\Lambda + C_{11})\right]_{(i)}}{[B_{11}]_{(i,i)}}, & \text{if } i \in \mathbb{I}_0, \\ \dfrac{[A_{11}]_{(i,i)}}{[D_{11}]_{(i,i)}} [\boldsymbol{u}_2]_{(i)}, & \text{if } i \in \mathbb{I}_+. \end{cases}$$

But in our algorithm, there is no need to compute $\hat{\boldsymbol{u}}_0$, and also we cannot and won't compute $\widetilde{bw}_j$ as defined in (5.3), either, because it contains cancellations.

Lemma 5.1 *The sequence defined by*

$$\begin{bmatrix} \widetilde{\boldsymbol{w}}_1^{(k)} \\ \widetilde{\boldsymbol{w}}_2^{(k)} \end{bmatrix}^{\mathrm{T}} := \begin{bmatrix} \boldsymbol{w}_1^{(k)} \\ \boldsymbol{w}_2^{(k)} \end{bmatrix}^{\mathrm{T}} + \begin{bmatrix} \boldsymbol{u}_0 \\ \boldsymbol{u}_1 \end{bmatrix}^{\mathrm{T}} E_k \geq 0 \quad \text{for } k \geq 1 \tag{5.4}$$

Algorithm 5.1 Highly Accurate $A - \Phi D$ for shifted MARE (1.1)

Require: the same as Algorithm 4.1;
Ensure: $\Phi_{(:,p+1:m)}$ and a triplet representation for $A - \Phi D$.

1: call Algorithm 4.1 with a minor addition to include computing $\widetilde{\boldsymbol{w}}_1^{(k+1)}$ and $\widetilde{\boldsymbol{w}}_2^{(k+1)}$ as defined by (5.4), with the stopping criteria (4.19) and (5.8), to return $X_{k+1} \approx \Phi + \Omega$ and $\begin{bmatrix} \widetilde{\boldsymbol{w}}_1^{(k+1)} \\ \widetilde{\boldsymbol{w}}_2^{(k+1)} \end{bmatrix} \approx \begin{bmatrix} \widetilde{\boldsymbol{w}}_1 \\ \widetilde{\boldsymbol{w}}_2 \end{bmatrix}$;
2: evaluate $\widetilde{\boldsymbol{w}}_0$ as in (5.9);
3: **return** the left triplet representation (5.11) for $A - \Phi D$ and $[X_{k+1}]_{(:,p+1:m)} \approx \Phi_{(:,p+1:m)}$.

converges, and

$$\lim_{k\to\infty} \begin{bmatrix} \widetilde{\boldsymbol{w}}_1^{(k)} \\ \widetilde{\boldsymbol{w}}_2^{(k)} \end{bmatrix} = \begin{bmatrix} \widetilde{\boldsymbol{w}}_1 \\ \widetilde{\boldsymbol{w}}_2 \end{bmatrix} := \begin{bmatrix} \boldsymbol{u}_0 \\ \boldsymbol{u}_1 \end{bmatrix}^{\mathrm{T}} - \begin{bmatrix} \boldsymbol{u}_2 \\ \boldsymbol{u}_3 \end{bmatrix}^{\mathrm{T}} \Phi_\Omega \geq 0. \tag{5.5}$$

Proof The fact that all $\widetilde{\boldsymbol{w}}_j^{(k)} \geq 0$ is because all $\boldsymbol{w}_j^{(k)} \geq 0, \boldsymbol{u}_0 \geq 0, \boldsymbol{u}_1 \geq 0$, and all $E_k \geq 0$. By (4.14), we have

$$\begin{bmatrix} \boldsymbol{w}_1^{(k)} \\ \boldsymbol{w}_2^{(k)} \end{bmatrix}^{\mathrm{T}} = \begin{bmatrix} \boldsymbol{u}_0 \\ \boldsymbol{u}_1 \end{bmatrix}^{\mathrm{T}} - \begin{bmatrix} \boldsymbol{u}_0 \\ \boldsymbol{u}_1 \end{bmatrix}^{\mathrm{T}} E_k - \begin{bmatrix} \boldsymbol{u}_2 \\ \boldsymbol{u}_3 \end{bmatrix}^{\mathrm{T}} X_k. \tag{5.6}$$

Plug (5.6) into (5.4) to get

$$\begin{bmatrix} \widetilde{\boldsymbol{w}}_1^{(k)} \\ \widetilde{\boldsymbol{w}}_2^{(k)} \end{bmatrix}^{\mathrm{T}} = \begin{bmatrix} \boldsymbol{u}_0 \\ \boldsymbol{u}_1 \end{bmatrix}^{\mathrm{T}} - \begin{bmatrix} \boldsymbol{u}_2 \\ \boldsymbol{u}_3 \end{bmatrix}^{\mathrm{T}} X_k. \tag{5.7}$$

Letting $k \to \infty$ in (5.7) and noting $\lim_{k\to\infty} X_k = \Phi_\Omega \equiv \Phi + \Omega$, we obtain the second equality in (5.5). □

Remark 5.1 In the case when $\mathscr{H}_\Omega$ has no eigenvalues on the imaginary axis or when it has a double eigenvalue 0, E_k converges to 0 by [16, Theorems 3.26 and 6.3], and thus

$$\lim_{k\to\infty} \begin{bmatrix} \boldsymbol{w}_1^{(k)} \\ \boldsymbol{w}_2^{(k)} \end{bmatrix} = \lim_{k\to\infty} \begin{bmatrix} \widetilde{\boldsymbol{w}}_1^{(k)} \\ \widetilde{\boldsymbol{w}}_2^{(k)} \end{bmatrix} = \begin{bmatrix} \widetilde{\boldsymbol{w}}_1 \\ \widetilde{\boldsymbol{w}}_2 \end{bmatrix} \geq 0.$$

Consequently, there is no need to compute $\widetilde{\boldsymbol{w}}_j^{(k)}$ to get the limiting vectors $\widetilde{\boldsymbol{w}}_j$. However, when W_Ω has just a simple eigenvalue 0 on the imaginary axis, there is no guarantee that E_k converges to 0.

Equation (5.7) implies that as X_k increasingly converges to Φ_Ω, while $\widetilde{\boldsymbol{w}}_j^{(k)}$ for $j = 1, 2$ decreasingly converge to $\widetilde{\boldsymbol{w}}_j$ for $j = 1, 2$, respectively. Because $\widetilde{\boldsymbol{w}}_j^{(k)}$ for $j = 1, 2$ can be computed in a cancellation-free manner according to (5.4), adding

$$\max_{1\leq i\leq m} \frac{\left(\left[\widetilde{\boldsymbol{w}}^{(k)} - \widetilde{\boldsymbol{w}}^{(k+1)}\right]_{(i)}\right)^2}{\left(\left[\widetilde{\boldsymbol{w}}^{(k-1)} - \widetilde{\boldsymbol{w}}^{(k)}\right]_{(i)} - \left[\widetilde{\boldsymbol{w}}^{(k)} - \widetilde{\boldsymbol{w}}^{(k+1)}\right]_{(i)}\right)\left[\widetilde{\boldsymbol{w}}^{(k+1)}\right]_{(i)}} \leq \texttt{rtol} \tag{5.8}$$

to the stopping criteria of Algorithm 4.1, we will be able to obtain $\widetilde{\boldsymbol{w}}_j$ for $j = 1, 2$ with high entrywise relative accuracy.

We are ready to construct a left triplet representation for $A - \Phi D$ in terms of $\widetilde{\boldsymbol{w}}_j$ for $j = 1, 2$. In fact, we have by (5.3)

$$\begin{bmatrix} \boldsymbol{u}_2 \\ \boldsymbol{u}_3 \end{bmatrix}^{\mathrm{T}} (A - \Phi D) = \left[\widetilde{\boldsymbol{w}}_0^{\mathrm{T}}\ \boldsymbol{u}_2^{\mathrm{T}} A_{12} + \boldsymbol{u}_3^{\mathrm{T}} A_{22} - \boldsymbol{u}_1^{\mathrm{T}} D_{22}\right] + \left[\widetilde{\boldsymbol{w}}_1^{\mathrm{T}} D_{11}\ \widetilde{\boldsymbol{w}}_2^{\mathrm{T}} D_{22}\right],$$

where $\widetilde{\boldsymbol{w}}_0^{\mathrm{T}} = \boldsymbol{u}_2^{\mathrm{T}} A_{11} - \hat{\boldsymbol{u}}_0^{\mathrm{T}} D_{11}$, or explicitly,

$$\left[\widetilde{\boldsymbol{w}}_0\right]_{(i)} = \begin{cases} \left[\boldsymbol{u}_2^{\mathrm{T}} A_{11}\right]_{(i)}, & \text{if } i \in \mathbb{I}_0, \\ 0, & \text{if } i \in \mathbb{I}_+. \end{cases} \tag{5.9}$$

By the proof of Theorem 3.3, $[A_{11}]_{(i,i)} > 0$ for $i \in \mathbb{I}_0$, and thus $\left[\boldsymbol{u}_2^{\mathrm{T}} A_{11}\right]_{(i)} > 0$ for $i \in \mathbb{I}_0$.

Using (4.4), we get $\boldsymbol{u}_2^{\mathrm{T}} A_{12} + \boldsymbol{u}_3^{\mathrm{T}} A_{22} - \boldsymbol{u}_1^{\mathrm{T}} D_{22} = \boldsymbol{v}_3^{\mathrm{T}}$. Thus

$$\begin{bmatrix} \boldsymbol{u}_2 \\ \boldsymbol{u}_3 \end{bmatrix}^{\mathrm{T}} (A - \Phi D) = \left[\widetilde{\boldsymbol{w}}_1^{\mathrm{T}} D_{11} + \widetilde{\boldsymbol{w}}_0^{\mathrm{T}}\ \widetilde{\boldsymbol{w}}_2^{\mathrm{T}} D_{22} + \boldsymbol{v}_3^{\mathrm{T}}\right]. \tag{5.10}$$

A left triplet representation for $A - \Phi D$ immediately follows:

$$A - \Phi D = \left(\mathrm{offdiag}(A - \Phi D), \begin{bmatrix} \boldsymbol{u}_2 \\ \boldsymbol{u}_3 \end{bmatrix}, \begin{bmatrix} D_{11}\widetilde{\boldsymbol{w}}_1 + \widetilde{\boldsymbol{w}}_0 \\ D_{22}^{\mathrm{T}}\widetilde{\boldsymbol{w}}_2 + \boldsymbol{v}_3 \end{bmatrix}\right)_{\mathrm{L}}, \tag{5.11}$$

and, as a consequence, its diagonal entries can be computed to high entrywise relative accuracy.

6 Numerical Examples

In this section, we will experiment three examples to illustrate superior entrywise relative accuracy in Φ_Ω and $K = -(A - \Phi D)$ achieved by Algorithms 4.1 and 5.1 which are essentially based on accADDA [23,29] in comparison with that delivered by ADDA [28]. For testing purpose, we compute the "exact" Φ_Ω and K by MATLAB's variable-precision arithmetic so that we can calculate entrywise relative errors

$$\max_{i,j} \frac{|X_\Omega - \Phi_\Omega|_{(i,j)}}{[\Phi_\Omega]_{(i,j)}}, \quad \max_{i,j} \frac{|K_{\mathrm{comput}} - K|_{(i,j)}}{|K|_{(i,j)}} \tag{6.1}$$

in an approximation $X_\Omega \approx \Phi_\Omega$ and $K_{\mathrm{comput}} \approx K$, where $0/0$ is treated as 0. We also report the normalized residual

$$\frac{\|X_\Omega D X_\Omega - A_\Omega X_\Omega - X_\Omega B_\Omega + C_\Omega\|_{\mathrm{F}}}{\|X\|_{\mathrm{F}}(\|X_\Omega\|_2\|D\|_2 + \|A_\Omega\|_2 + \|B_\Omega\|_2) + \|C_\Omega\|_{\mathrm{F}}}, \tag{6.2}$$

commonly used in practice because of its availability, where $\|\cdot\|_2$ and $\|\cdot\|_{\mathrm{F}}$ are the matrix spectral and Frobenius norm, respectively. Except for the calculation of the "exact" Φ_Ω and K, all calculations are done in MATLAB with the default IEEE double precision whose unit roundoff is $2^{-53} \approx 1.11 \cdot 10^{-16}$.

Example 6.1 This example is the testing problem `NP15` in [24], a modification of [23, Example 5.1]. In the notation of "Appendix 1", $\mathbb{S} = \{1,\ 2,\ 3,\ 4,\ 5,\ 6\}$, $\mathbb{S}_b = \mathbb{S}$, $\mathbb{S}_u = \mathbb{S}_d = \emptyset$,

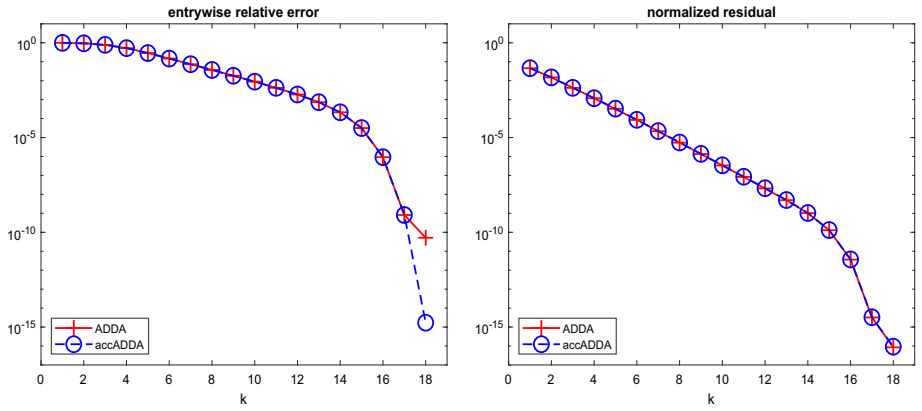

Fig. 1 Example 6.1. ADDA loses about 6 significant decimal digits in the last step, better than the worst case scenario, whereas Algorithm 4.1 computes all entries of Φ_Ω accurately to 16 decimal digits

and

$$Q = \begin{bmatrix} -4 & 0 & 0 & 0 & 0 & 4 \\ 0 & -15-10^{-8} & 5 & 5 & 5 & 10^{-8} \\ 0 & 5 & -15 & 5 & 5 & 0 \\ 0 & 5 & 5 & -15 & 5 & 0 \\ 0 & 5 & 5 & 5 & -15 & 0 \\ 4 & 1 & 0 & 0 & 0 & -5 \end{bmatrix}, \quad V_b = I_6,$$

and $\widehat{R}_b = \mathrm{diag}(1, 1, 1, -1.001, -1.001, -1.001)$. Recall the dimension parameters given by (A.3). The coefficient matrices A, B, C, and D of ARE (1.1) are given by (A.13):

$$A = 0_{n\times n}, \quad B = \widehat{R}_b V_b^{-1}, \quad C = Q_{bb} V_b^{-1}, \quad D = I_n.$$

Now W_0 is given by (A.17) and we construct its *left* triplet representation as described in Remark 8.2(c). For this example, we find

$$1.2952 \cdot 10^{-10} \le [\Phi_\Omega]_{(i,j)} \le 1.4585,$$

indicating the plain ADDA [28] could possibly lose up to about 10 significant decimal digits in the computed smallest entries of Φ_Ω. Figure 1 shows two history plots for ADDA and Algorithm 4.1: the entrywise relative error as defined by (6.1) (left plot) and the normalized residual as defined by (6.2) (right plot) against the iteration index. As can be seen from the left plot, ADDA loses about 6 significant decimal digits in the last step, better than the worst case scenario of 10 digits, whereas Algorithm 4.1 computes all entries of Φ_Ω accurately to 16 decimal digits, even though the normalized residuals for both methods match perfectly.

Example 6.2 This slightly differs from the one in Example 6.1: the same Q and $\mathbb{S}$ but with $\mathbb{S}_b = \{1,\ 2,\ 3,\ 4,\ 5\}$, $\mathbb{S}_u = \emptyset$, $\mathbb{S}_d = \{6\}$, and

$$V_b = I_5, \quad \widehat{R}_b = \mathrm{diag}(1, 1, 1, -1.001, -1.001, -1.001), \quad \widehat{R}_d = -1.001.$$

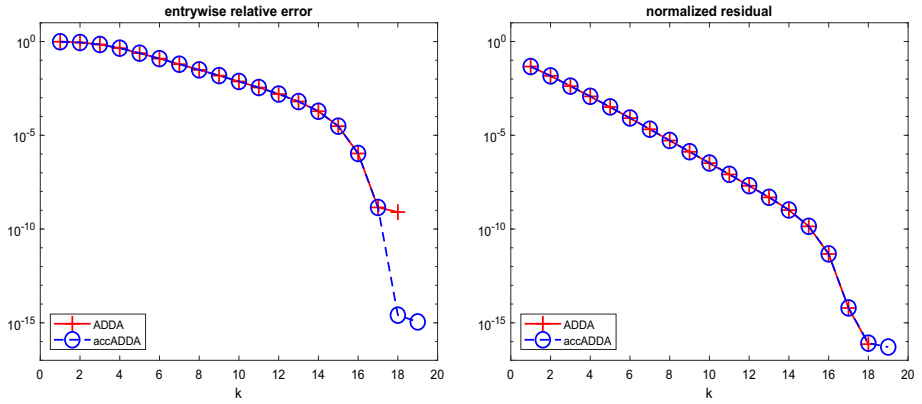

Fig. 2 Example 6.2. ADDA loses about 6 significant decimal digits in the last step, better than the worst case scenario, whereas Algorithm 4.1 computes all entries of Φ_Ω accurately to 16 decimal digits

Again the dimension parameters are given by (A.3). According to (A.13), we get

$$A = 0_{n\times n},\ \ C = \left[Q_{(1:n_b,1:n_b)} V_b^{-1}, -Q_{(1:n_b,n+1:N)}/\widehat{R}_d \right],$$
$$B = \begin{bmatrix} \widehat{R}_b V_b^{-1} & 0 \\ -Q_{(n+1:N,1:n_b)} V_b^{-1} & Q_{(n+1:N,n+1:N)}/\widehat{R}_d \end{bmatrix},\quad D = \begin{bmatrix} I_{n_b} \\ 0_{n_d\times n_b} \end{bmatrix}.$$

Again W_0 is given by (A.17) and we construct its *left* triplet representation as described there. For this example, we find

$$3.6946 \cdot 10^{-10} \le [\Phi_\Omega]_{(i,j)} \le 1.3957,$$

indicating the plain ADDA [28] could possibly lose up to about 10 significant decimal digits. We ran ADDA and Algorithm 4.1 and generated two history plots in Fig. 2. We observed the same phenomena as we saw in Example 6.1.

For Example 6.1, $A - \Phi D = -\Phi$, and for Example 6.2, $A - \Phi D = -\Phi_{(:,1:5)}$. Upon examining numerical results, we find that all diagonal entries of Φ_Ω and Φ are of $O(1)$ and there is no catastrophic cancellation happened in evaluating $\Phi_\Omega - \Omega$. Therefore, $A - \Phi D = -\Phi$ or $-\Phi_{(:,1:5)}$ calculated in the usual way will have high entrywise relative accuracy without the need of going into the complicated way as we described in Sect. 5.

Example 6.3 This is a random generated example (A.18) with $n_b = n_u = n_d = 10$ and $s = 0.006$, where Q is by MATLAB's `sprandn`, V_b, $\widehat{R}_u$ and $\widehat{R}_d$ by `randn`. We save the example after its generation for repeatability. Now W_0 is given by (A.20). We construct its left triplet representation as (A.21). For this example, we find

$$0.1236 \cdot 10^{-15} \le [\Phi_\Omega]_{(i,j)} \le 0.9650,$$

indicating the plain ADDA [28] could possible lose up to about 15 significant decimal digits. Numerically, it loses about 11 significant decimal digits in the smallest entries, as demonstrated in Fig. 3. We also find

$$4.633 \times 10^{-15} \le |K|_{(i,j)} \le 7.022.$$

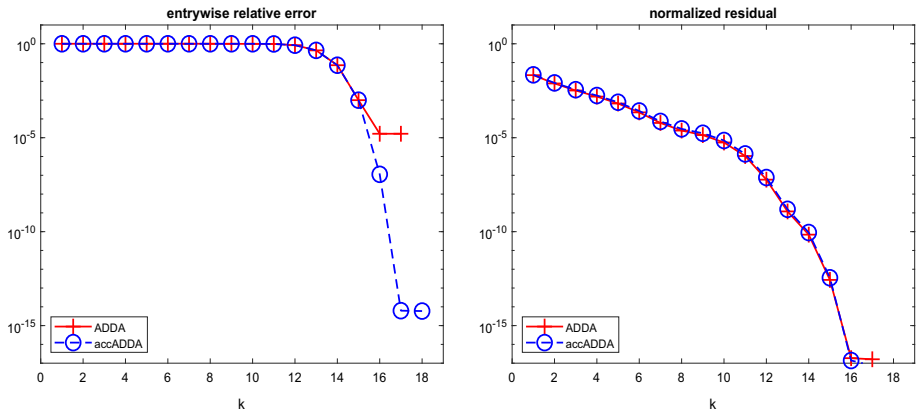

Fig. 3 Example 6.3. ADDA loses about 11 significant decimal digits in the last step whereas Algorithm 4.1 computes all entries of Φ_Ω accurately to 16 decimal digits

Denote by $K_{adda} = -(A_\Omega - \Phi_\Omega D)$ the one calculated in the usual way with Φ_Ω returned by ADDA, and $K_{\rm acc}$ by Algorithm 5.1. The entrywise relative errors in K_{adda} and $K_{\rm acc}$ are

$$\max_{i,j} \frac{|K_{adda} - K|_{(i,j)}}{|K|_{(i,j)}} = 1.2 \times 10^{-5}, \quad \max_{i,j} \frac{|K_{\rm acc} - K|_{(i,j)}}{|K|_{(i,j)}} = 5.4 \times 10^{-15},$$

demonstrating near full entrywise relative accuracy in $K_{\rm acc}$ by Algorithm 5.1, while K_{adda} by ADDA loses up to about 10 significant decimal digits in some of the entries.

7 Conclusions

An algebraic Riccati equation (ARE) $XDX - AX - XB + C = 0$ is called a shifted M-Matrix algebraic Riccati equation (MARE) of index p if the resulting ARE (1.5) in X_Ω upon substitution (1.4) is an MARE, i.e., the coefficient matrix W_Ω in (1.6) is a nonsingular or an irreducible singular M-matrix. Such an ARE arises from computing the invariant density of a Markov modulated Brownian motion (MMBM) [1,4,20], where the solution $X_\Omega = \Phi_\Omega$, the minimal nonnegative solution of the resulting MARE, is of interest for follow-up computations. In order for these follow-up computations to have high accuracy, it is imperative to compute Φ_Ω and $A - \Phi D \equiv A_\Omega - \Phi_\Omega D$ with high entrywise relative accuracy. Three major contributions of this paper are

1. a complete description of shifted MARE, and sufficient and necessary conditions for an ARE to be a shifted MARE,
2. an algorithm to compute shifting matrix Ω in (1.4), and
3. algorithms to accurately compute the minimal nonnegative solution Φ_Ω and $A - \Phi D$ with high entrywise relative accuracy they deserve, essentially based on accADDA of [23,29].

Numerical examples are presented to demonstrate our theory and confirm that our numerical methods can deliver numerical results with claimed accuracy.

Acknowledgements The authors are grateful to two anonymous referees for their helpful comments and suggestions that improve the presentation. Liu is supported in part by the National Natural Science Foundation of China: NSFC-11501388. Xue is supported in part by the National Natural Science Foundation of China: NSFC-11771100, the National Key R & D Program of China: 2018YFA0703900, and by Laboratory of Mathematics for Nonlinear Science, Fudan University. Li is supported in part by the US National Science Foundation: DMS-1719620 and DMS-2009689.

Appendix

Markov Modulated Brownian Motion

Consider a general Markov modulated Brownian motion (MMBM) $\{(F(t), J(t)) : t \geq 0\}$, where $F(t)$ is the level of fluid in a container at time t and $J(t)$ is the state at time t of a continuous time Markov chain modulating the level process. Let Q be the infinitesimal generator of the Markov chain which has a finite state space partitioned as

$$\mathbb{S} = \mathbb{S}_b \cup \mathbb{S}_u \cup \mathbb{S}_d \cup \mathbb{S}_0. \tag{A.1}$$

With respect to this state space partition, the dynamics of the fluid level can be described as follows:

1. when the chain is in a state $i \in \mathbb{S}_b$, the level evolves according to a Brownian motion process with drift r_i and diffusion coefficient $\frac{1}{2}v_i > 0$, where r_i can be positive, negative, or 0;
2. when the chain is in a state $i \in \mathbb{S}_u$, the level increases at rate $|r_i|$ with $r_i > 0$;
3. when the chain is in a state $i \in \mathbb{S}_d$, the level decrease at rate $|r_i|$ with $r_i < 0$, provided the container is nonempty;
4. when the chain is in a state $i \in \mathbb{S}_0$, the level remains unchanged.

When the Markov chain falls within the subset of states $\mathbb{S}_u \cup \mathbb{S}_d \cup \mathbb{S}_0$, the fluid level changes very much like a Markov modulated fluid flow model (MMFF).

In what follows, we adopt the convention that $v_i = 0$ for $i \in \mathbb{S}_u \cup \mathbb{S}_d \cup \mathbb{S}_0$, and, without loss of generality, we may assume that $\mathbb{S}_0 = \emptyset$ because it can be censored out [1]. Therefore, instead of (A.1), we will have

$$\mathbb{S} = \mathbb{S}_b \cup \mathbb{S}_u \cup \mathbb{S}_d. \tag{A.2}$$

Let $N = |\mathbb{S}|$, the cardinality of $\mathbb{S}$, i.e., the number of states in $\mathbb{S}$, and, for future references,

$$n_b = |\mathbb{S}_b|,\ n_u = |\mathbb{S}_u|,\ n_d = |\mathbb{S}_d|,\ N = |\mathbb{S}| = n_b + n_u + n_d, \tag{A.3a}$$

$$m = n_b + n_d,\ n = n_b + n_u. \tag{A.3b}$$

Besides Q, let

$$\hat{R} = \operatorname{diag}(r_i)_{i\in\mathbb{S}}, \quad V = \operatorname{diag}(v_i/2)_{i\in\mathbb{S}}.$$

We assume the infinitesimal generator Q is irreducible and denote by $\boldsymbol{\pi} \in \mathbb{R}^N$ its stationary distribution, i.e. $\boldsymbol{\pi} > 0, \boldsymbol{\pi}^{\mathrm{T}} Q = 0$ and $\boldsymbol{\pi}^{\mathrm{T}}\mathbf{1} = 1$. In the model, $\nu = \boldsymbol{\pi}^{\mathrm{T}}\hat{R}\mathbf{1} < 0$ which means the fluid level is positive recurrent and has a steady-state distribution.

We are interested in computing its invariant density vector $\boldsymbol{p} : \mathbb{R}_{0+} \to \mathbb{R}_{0+}^{1\times N}$. It satisfies the following second-order ordinary differential equation

$$\boldsymbol{p}''(x)V - \boldsymbol{p}'(x)\hat{R} + \boldsymbol{p}(x)Q = 0, \tag{A.4}$$

with suitable boundary conditions. It has been proved [17,20] that $\boldsymbol{p}(x)$ takes the matrix-exponential form

$$\boldsymbol{p}(x) = \boldsymbol{c}e^{Kx}\left[I_n\ \Gamma\right], \tag{A.5}$$

where $\boldsymbol{c} \in \mathbb{R}^{1\times n}$ can be determined by the boundary conditions, $K \in \mathbb{R}^{n\times n}$ and $\Gamma \in \mathbb{R}^{n\times n_d}$. The pair $(K,\ \Gamma)$ has a probabilistic meaning: $\Gamma \geq 0$ is the matrix recording the first-return probabilities of the time-reversed process, and K is the sub-generator matrix for the downward-record process with the property that $-K$ is a Z-matrix and $K\mathbf{1} \leq 0$, which implies that $-K$ is an M-matrix.

Substituting (A.5) into (A.4) gives

$$K^2\left[I\ \Gamma\right]V - K\left[I\ \Gamma\right]\hat{R} + \left[I\ \Gamma\right]Q = 0, \tag{A.6}$$

where unknowns are K and Γ. The computation of $\boldsymbol{p}$ is thus reduced to solving the matrix Eq. (A.6). Nguyen and Poloni [24] proposed to first transform (A.6) into a matrix equation of the form $X^2\widehat{A} - X\widehat{B} + \widehat{C} = 0$ with $\widehat{A},\ \widehat{C} \in \mathbb{R}_{0+}^{N\times N}$, $\widehat{B} \in \mathbb{R}^{N\times N}$, and $\widehat{B} - \widehat{A} - \widehat{C}$ being a regular M-matrix and then solve it by the cyclic reduction (CR) that can produce a solution with high entrywise relative accuracy. Later K and Γ are recovered from the computed solution X. It is an elegant and effective approach. The only drawback is perhaps the transformed equation has a larger scale than (A.6).

Remark 8.1 When $\mathbb{S} = \mathbb{S}_b$, V is nonsingular and (A.6) reduces to $K^2V - K\hat{R} + Q = 0$, or equivalently, $K^2 - K\hat{R}V^{-1} + QV^{-1} = 0$. Guo [11] investigated the quadratic matrix equation of the form $X^2 + XE + F = 0$ with a diagonal matrix E and an M-matrix F and showed such an equation has a unique M-matrix solution.

Inspired by the idea in Ahn and Ramaswami [1], in what follows, we seek to transform (A.6) into an $n \times m$ ARE which turns out to be a shifted MARE and then propose to solve the shifted MARE by the highly accurate ADDA detailed in Sects. 4 and 5.

Consistently with (A.2), we partition Q, V and $\hat{R}$ as

$$Q = \begin{bmatrix} Q_{bb} & Q_{bu} & Q_{bd} \\ Q_{ub} & Q_{uu} & Q_{ud} \\ Q_{db} & Q_{du} & Q_{dd} \end{bmatrix}, \quad V = \begin{bmatrix} V_b & & \\ & 0 & \\ & & 0 \end{bmatrix}, \quad \widehat{R} = \begin{bmatrix} \widehat{R}_b & & \\ & \widehat{R}_u & \\ & & \widehat{R}_d \end{bmatrix}, \tag{A.7}$$

where $V_b = \mathrm{diag}(v_i)_{i\in\mathbb{S}_b}$ with all $v_i > 0$, $\widehat{R}_b = \mathrm{diag}(r_i)_{i\in\mathbb{S}_b}$, $\widehat{R}_u = \mathrm{diag}(r_i)_{i\in\mathbb{S}_u}$ with all $r_i > 0$, and $\widehat{R}_d = \mathrm{diag}(r_i)_{i\in\mathbb{S}_d}$ with all $r_i < 0$. Similarly, corresponding to $\mathbb{S}_b \cup \mathbb{S}_u$, we partition K and Γ as

$$K = \begin{array}{c} \\ n_b \\ n_u \end{array}\!\!\begin{array}{c} \begin{array}{cc} n_b & n_u \end{array} \\ \begin{bmatrix} K_{11} & K_{12} \\ K_{21} & K_{22} \end{bmatrix} \end{array}, \quad \Gamma = \begin{array}{c} \\ n_b \\ n_u \end{array}\!\!\begin{array}{c} n_d \\ \begin{bmatrix} \Gamma_1 \\ \Gamma_2 \end{bmatrix} \end{array}.$$

Then (A.6) can be expressed as

$$\begin{bmatrix} K_{11} & K_{12} \\ K_{21} & K_{22} \end{bmatrix}^2 \begin{bmatrix} I & 0 & \Gamma_1 \\ 0 & I & \Gamma_2 \end{bmatrix} \begin{bmatrix} V_b & & \\ & 0 & \\ & & 0 \end{bmatrix} - \begin{bmatrix} K_{11} & K_{12} \\ K_{21} & K_{22} \end{bmatrix} \begin{bmatrix} I & 0 & \Gamma_1 \\ 0 & I & \Gamma_2 \end{bmatrix} \begin{bmatrix} \widehat{R}_b & & \\ & \widehat{R}_u & \\ & & \widehat{R}_d \end{bmatrix}$$
$$+ \begin{bmatrix} I & 0 & \Gamma_1 \\ 0 & I & \Gamma_2 \end{bmatrix} \begin{bmatrix} Q_{bb} & Q_{bu} & Q_{bd} \\ Q_{ub} & Q_{uu} & Q_{ud} \\ Q_{db} & Q_{du} & Q_{dd} \end{bmatrix} = 0, \tag{A.8}$$

where the unknowns are all K_{ij} and Γ_i. To reduce the number of unknowns, we claim that $\begin{bmatrix} K_{12} \\ K_{22} \end{bmatrix}$ can be obtained by some affine transformation on $\begin{bmatrix} \Gamma_1 \\ \Gamma_2 \end{bmatrix}$. In fact, equating the second block columns on both sides of (A.8), we arrive at

$$-\begin{bmatrix} K_{11} & K_{12} \\ K_{21} & K_{22} \end{bmatrix}\begin{bmatrix} I & 0 & \Gamma_1 \\ 0 & I & \Gamma_2 \end{bmatrix}\begin{bmatrix} 0 \\ \widehat{R}_u \\ 0 \end{bmatrix} + \begin{bmatrix} I & 0 & \Gamma_1 \\ 0 & I & \Gamma_2 \end{bmatrix}\begin{bmatrix} Q_{bu} \\ Q_{uu} \\ Q_{du} \end{bmatrix} = 0,$$

or equivalently,

$$-\begin{bmatrix} K_{12} \\ K_{22} \end{bmatrix}\widehat{R}_u + \begin{bmatrix} Q_{bu} \\ Q_{uu} \end{bmatrix} + \begin{bmatrix} \Gamma_1 \\ \Gamma_2 \end{bmatrix} Q_{du} = 0 \quad \Rightarrow \quad \begin{bmatrix} K_{12} \\ K_{22} \end{bmatrix} = \begin{bmatrix} Q_{bu} \\ Q_{uu} \end{bmatrix}\widehat{R}_u^{-1} + \begin{bmatrix} \Gamma_1 \\ \Gamma_2 \end{bmatrix} Q_{du}\widehat{R}_u^{-1}. \tag{A.9}$$

Thus $\begin{bmatrix} K_{12} \\ K_{22} \end{bmatrix}$ can be eliminated from (A.8), leaving $\begin{bmatrix} K_{11} \\ K_{21} \end{bmatrix}$ and $\begin{bmatrix} \Gamma_1 \\ \Gamma_2 \end{bmatrix}$ as the ones to be determined. Next, we group the unknowns into one[6]

$$X := \begin{bmatrix} K_{11} & \Gamma_1 \\ K_{21} & \Gamma_2 \end{bmatrix} \in \mathbb{R}^{n\times m} \tag{A.10}$$

and seek to establish an equation in X, based on (A.8). To this end, we extract the first and the third block columns in (A.8) to yield

$$\begin{bmatrix} K_{11} & K_{12} \\ K_{21} & K_{22} \end{bmatrix}^2\begin{bmatrix} I & 0 & \Gamma_1 \\ 0 & I & \Gamma_2 \end{bmatrix}\begin{bmatrix} V_b & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} - \begin{bmatrix} K_{11} & K_{12} \\ K_{21} & K_{22} \end{bmatrix}\begin{bmatrix} I & 0 & \Gamma_1 \\ 0 & I & \Gamma_2 \end{bmatrix}\begin{bmatrix} \widehat{R}_b & 0 \\ 0 & 0 \\ 0 & \widehat{R}_d \end{bmatrix}$$
$$+\begin{bmatrix} I & 0 & \Gamma_1 \\ 0 & I & \Gamma_2 \end{bmatrix}\begin{bmatrix} Q_{bb} & Q_{bd} \\ Q_{ub} & Q_{ud} \\ Q_{db} & Q_{dd} \end{bmatrix} = 0. \tag{A.11}$$

Using (A.9), we get

$$\begin{bmatrix} K_{11} & K_{12} \\ K_{21} & K_{22} \end{bmatrix} = \begin{bmatrix} 0_{n_b\times n_b} & Q_{bu}\widehat{R}_u^{-1} \\ 0 & Q_{uu}\widehat{R}_u^{-1} \end{bmatrix} + \begin{bmatrix} K_{11} & \Gamma_1 \\ K_{21} & \Gamma_2 \end{bmatrix}\begin{bmatrix} I & \\ & Q_{du}\widehat{R}_u^{-1} \end{bmatrix}, \tag{A.12}$$

plugging which into (A.11), we finally obtain an ARE in the form of (1.1) with

$$A = \begin{bmatrix} 0_{n_b\times n_b} & -Q_{bu}\widehat{R}_u^{-1} \\ 0 & -Q_{uu}\widehat{R}_u^{-1} \end{bmatrix}, \quad B = \begin{bmatrix} \widehat{R}_b V_b^{-1} & \\ -Q_{db}V_b^{-1} & -Q_{dd}(-\widehat{R}_d)^{-1} \end{bmatrix}, \tag{A.13a}$$

$$C = \begin{bmatrix} Q_{bb}V_b^{-1} & Q_{bd}(-\widehat{R}_d)^{-1} \\ Q_{ub}V_b^{-1} & Q_{ud}(-\widehat{R}_d)^{-1} \end{bmatrix}, \quad D = \begin{bmatrix} I & \\ & Q_{du}\widehat{R}_u^{-1} \end{bmatrix}. \tag{A.13b}$$

We now take a close look at X in (A.10). The matrices K_{21}, Γ_1 and Γ_2 are nonnegative, while K_{11} has nonnegative off-diagonal entries and negative diagonal entries. If we add a sufficiently large diagonal matrix, say Λ, to K_{11}, then the resulting matrix

$$X_\Omega := \begin{bmatrix} K_{11}+\Lambda & \Gamma_1 \\ K_{21} & \Gamma_2 \end{bmatrix} = \begin{bmatrix} K_{11} & \Gamma_1 \\ K_{21} & \Gamma_2 \end{bmatrix} + \begin{bmatrix} \Lambda & 0 \\ 0 & 0 \end{bmatrix} =: X + \Omega \tag{A.14}$$

can be made nonnegative. Plug $X = X_\Omega - \Omega$ into (1.1) to yield an ARE in X_Ω in the form of (1.5) with $p = n_b$. The associated matrix of ARE (1.5)

[6] This idea of conceiving this new matrix X is borrowed from [1, p. 74].

$$W_\Omega = \begin{bmatrix} B_\Omega & -D \\ -C_\Omega & A_\Omega \end{bmatrix}$$

$$= \begin{array}{c} \\ n_b \\ n_d \\ n_b \\ n_u \end{array} \begin{array}{c} \begin{array}{ccccc} n_b & & n_d & n_b & n_u \end{array} \\ \begin{bmatrix} \Lambda + \widehat{R}_b V_b^{-1} & 0 & -I_{n_b} & 0 \\ -Q_{db}V_b^{-1} & -Q_{dd}(-\widehat{R}_d)^{-1} & 0 & -Q_{du}\widehat{R}_u^{-1} \\ -Q_{bb}V_b^{-1} - \Lambda\widehat{R}_b V_b^{-1} - \Lambda^2 & -Q_{bd}(-\widehat{R}_d)^{-1} & \Lambda & -Q_{bu}\widehat{R}_u^{-1} \\ -Q_{ub}V_b^{-1} & -Q_{ud}(-\widehat{R}_d)^{-1} & 0 & -Q_{uu}\widehat{R}_u^{-1} \end{bmatrix} \end{array} \tag{A.15}$$

is a nonsingular M-matrix (for sufficiently large Λ), making (1.5) an MARE, and, as a result, ARE (1.1) with (A.13) is a shifted MARE. The interesting solution X_Ω corresponding to the interesting one of (1.1) for the underlying application turns out to be the unique minimal nonnegative solution of (1.5). This is because $-K$ is an M-matrix and, by (A.9) and (A.10),

$$-K = (A - XD) = A_\Omega - X_\Omega D, \tag{A.16}$$

following the solution properties of an MARE as summarized in Proposition 2.2.

Remark 8.2 A few comments are in order about ARE (1.1) with (A.13).

(a) It can be verified that A, B, C, and D given by (A.13) admit the forms of those in (3.2) with the properties specified in Theorem 3.2.
(b) Assumption 4.1(a), i.e., that the diagonal matrices A_{11}, B_{11} and D_{11} are accurately known with high entrywise relative accuracy, is satisfied by those in (A.13). In particular, $A_{11} = 0$ and $D_{11} = I$ here.
(c) To realize Assumption 4.1(b), we notice

$$W_0 = -\begin{bmatrix} Q_{dd} & Q_{db} & Q_{du} \\ Q_{bd} & Q_{bb} & Q_{bu} \\ Q_{ud} & Q_{ub} & Q_{uu} \end{bmatrix} \begin{bmatrix} -\hat{R}_d^{-1} & & \\ & V_b^{-1} & \\ & & \hat{R}_u^{-1} \end{bmatrix}. \tag{A.17}$$

Consistently with (A.2), we partition the stationary distribution $\boldsymbol{\pi}$ of the infinitesimal generator Q as $\boldsymbol{\pi} = \begin{bmatrix} \boldsymbol{\pi}_b & \boldsymbol{\pi}_u & \boldsymbol{\pi}_d \end{bmatrix} \in \mathbb{R}^{1\times N}$ which can be computed with high entrywise relative accuracy by the GTH algorithm [25, Theorem 2]. We have $\begin{bmatrix} \boldsymbol{\pi}_d & \boldsymbol{\pi}_b & \boldsymbol{\pi}_u \end{bmatrix} W_0 = 0$, a very natural thing to have. This leads to a *left* triplet representation of W_0, as needed for computing Φ_Ω and $K = -(A - \Phi D)$ entrywise accurately by the algorithms in Sects. 4 and 5.

What we have described so far is for a type of shifted MAREs as arising from computing the invariant density vector $\boldsymbol{p}$ of MMBM as determined by (A.4). It is about a steady-state analysis. In the time-dependent analysis of MMBM, Ahn and Ramaswami [1] established AREs of the first passage time distribution. These AREs provide ample examples of shifted MAREs, too. They are in the form of (1.1) with

$$A = \begin{bmatrix} -\widehat{R}_b V_b^{-1} & -\sqrt{2}V_b^{-\frac{1}{2}}Q_{bu} \\ 0 & -\widehat{R}_u^{-1}(Q_{uu} - sI) \end{bmatrix}, \quad B = \begin{bmatrix} 0 & 0 \\ \widehat{R}_d^{-1}Q_{db} & \widehat{R}_d^{-1}(Q_{dd} - sI) \end{bmatrix}, \tag{A.18a}$$

$$C = \begin{bmatrix} \sqrt{2}V_b^{-\frac{1}{2}}(Q_{bb} - sI) & \sqrt{2}V_b^{-\frac{1}{2}}Q_{bd} \\ \widehat{R}_u^{-1}Q_{ub} & \widehat{R}_u^{-1}Q_{ud} \end{bmatrix}, \quad D = \begin{bmatrix} (2V_b)^{-\frac{1}{2}} & \\ & (-\widehat{R}_d)^{-1}Q_{du} \end{bmatrix}, \tag{A.18b}$$

where $s \in \mathbb{C}_{0+}$ in general but we will consider $s \geq 0$ only. Now we shift the corresponding ARE (1.1) as in (1.4) – (1.7) with $\Lambda = \hat{R}_b(2V_b)^{-1/2} + \sqrt{\hat{R}_b^2(2V_b)^{-1} + 2\Lambda_{Q_{bb}} + 2sI}$ to get

$$A_\Omega = \begin{bmatrix} -\widehat{R}_b(2V_b)^{-1} + (2V_b)^{-\frac{1}{2}}\sqrt{(2V_b)^{-1}\widehat{R}_b^2 + 2\Lambda_{Q_{b,b}} + 2sI} & -\sqrt{2}V_b^{-\frac{1}{2}}Q_{bu} \\ 0 & -\widehat{R}_u^{-1}(Q_{uu} - sI) \end{bmatrix}, \tag{A.19a}$$

$$B_\Omega = \begin{bmatrix} (2V_b)^{-1}\widehat{R}_b + (2V_b)^{-\frac{1}{2}}\sqrt{(2V_b)^{-1}\widehat{R}_b^2 + 2\Lambda_{Q_{b,b}} + 2sI} & 0 \\ \widehat{R}_d^{-1}Q_{db} & \widehat{R}_d^{-1}(Q_{dd} - sI) \end{bmatrix}, \tag{A.19b}$$

$$C_\Omega = \begin{bmatrix} \sqrt{2}V_b^{-\frac{1}{2}}(Q_{bb} + \Lambda_{Q_{b,b}}) & \sqrt{2}V_b^{-\frac{1}{2}}Q_{bd} \\ \widehat{R}_u^{-1}Q_{ub} & \widehat{R}_u^{-1}Q_{ud} \end{bmatrix}, \tag{A.19c}$$

where $\Lambda_{Q_{bb}} = \mathrm{diag}(-[Q_{bb}]_{(i,i)})_{i\in\mathbb{S}_b}$. The resulting ARE (1.5) is the same as the one in [1, Theorem 5.1] which provably is an MARE. Ahn and Ramaswami [1] suggested to solve the MARE, as a nonlinear equation, by Newton's method.

Remark 8.3 The first two comments in Remark 8.2 still applies to ARE (1.1) with (A.18). To realize Assumption 4.1(b), we now have

$$W_0 = \begin{bmatrix} -\widehat{R}_d^{-1} & & \\ & \sqrt{2}V_b^{-1/2} & \\ & & \widehat{R}_u^{-1} \end{bmatrix} \begin{bmatrix} sI - Q_{dd} & -Q_{db} & -Q_{du} \\ -Q_{bd} & sI - Q_{bb} & -Q_{bu} \\ -Q_{ud} & -Q_{ub} & sI - Q_{uu} \end{bmatrix} \times \begin{bmatrix} I_{n_d} & & \\ & (2V_b)^{-1/2} & \\ & & I_{n_u} \end{bmatrix}. \tag{A.20}$$

Again let $\boldsymbol{\pi}$ be as in Remark 8.2(c). Then $\boldsymbol{u}^{\mathrm{T}}W_0 = \boldsymbol{v}^{\mathrm{T}}$, where

$$\boldsymbol{u}^{\mathrm{T}} = \begin{bmatrix}\boldsymbol{\pi}_d & \boldsymbol{\pi}_b & \boldsymbol{\pi}_u\end{bmatrix} \mathrm{diag}\left(-\widehat{R}_d, \frac{1}{\sqrt{2}}V_b^{1/2}, \widehat{R}_u\right), \quad \boldsymbol{v}^{\mathrm{T}} = s\begin{bmatrix}\boldsymbol{\pi}_d, \boldsymbol{\pi}_b(2V_b)^{-1/2}, \boldsymbol{\pi}_u\end{bmatrix}. \tag{A.21}$$

We commented that $\boldsymbol{\pi}$ can be computed with high entrywise relative accuracy. Since $\widehat{R}_d$, V_b, and $\widehat{R}_u$ are diagonal, $\boldsymbol{u}$ and $\boldsymbol{v}$ as in (A.21) can be computed with high entrywise relative accuracy, too.

References

1. Ahn, S., Ramaswami, V.: A quadratically convergent algorithm for first passage time distributions in the Markov-modulated Brownian motion. Stoch. Models **33**(1), 59–96 (2017)
2. Alfa, A.S., Xue, J., Ye, Q.: Accurate computation of the smallest eigenvalue of a diagonally dominant M-matrix. Math. Comput. **71**, 217–236 (2002)
3. Alfa, A.S., Xue, J., Ye, Q.: Entrywise perturbation theory for diagonally dominant M-matrices with applications. Numer. Math. **90**(3), 401–414 (2002)
4. Asmussen, S.: Stationary distributions for fluid flow models with or without Brownian noise. Commun. Stat. Stoch. Models **11**(1), 21–49 (1995)
5. Berman, A., Plemmons, R.J.: Nonnegative Matrices in the Mathematical Sciences. SIAM, Philadelphia: This SIAM edition is a corrected reproduction of the work first published in 1979 by Academic Press. San Diego, CA (1994)
6. Bini, D.A., Iannazzo, B., Meini, B.: Numerical Solution of Algebraic Riccati Equations. SIAM, Philadelphia (2012)

7. Bini, D.A., Meini, B., Poloni, F.: Transforming algebraic Riccati equations into unilateral quadratic matrix equations. Numer. Math. **116**, 553–578 (2010)
8. Fiedler, M.: Special Matrices and Their Applications in Numerical Mathematics, 2nd edn. Dover Publications Inc, Mineola (2008)
9. Grassmann, W., Taksar, M., Heyman, D.: Regenerative analysis and steady-state distributions for Markov chains. Oper. Res. **33**, 1107–1116 (1985)
10. Guo, C.H.: Nonsymmetric algebraic Riccati equations and Wiener–Hopf factorization for M-matrices. SIAM J. Matrix Anal. Appl. **23**, 225–242 (2001)
11. Guo, C.H.: On a quadratic matrix equation associated with an M-matrix. IMA J. Numer. Anal. **23**(1), 11–27 (2003)
12. Guo, C.H.: A new class of nonsymmetric algebraic Riccati equations. Linear Algebra Appl. **426**(2–3), 636–649 (2007)
13. Guo, C.H., Higham, N.: Iterative solution of a nonsymmetric algebraic Riccati equation. SIAM J. Matrix Anal. Appl. **29**, 396–412 (2007)
14. Guo, C.H., Laub, A.J.: On the iterative solution of a class of nonsymmetric algebraic Riccati equations. SIAM J. Matrix Anal. Appl. **22**, 376–391 (2000)
15. Guo, X., Lin, W., Xu, S.: A structure-preserving doubling algorithm for nonsymmetric algebraic Riccati equation. Numer. Math. **103**, 393–412 (2006)
16. Huang, T.M., Li, R.-C., Lin, W.-W.: Structure-Preserving Doubling Algorithms for Nonlinear Matrix Equations, Fundamentals of Algorithms, vol. 14. SIAM, Philadelphia (2018)
17. Ivanovs, J.: Markov-modulated Brownian motion with two reflectig barriers. J. Appl. Probab. **47**(4), 1034–1047 (2010)
18. Juang, J.: Existence of algebraic matrix Riccati equations arising in transport theory. Linear Algebra Appl. **230**, 89–100 (1995)
19. Juang, J., Lin, W.-W.: Nonsymmetric algebraic Riccati equations and Hamiltonian-like matrices. SIAM J. Matrix Anal. Appl. **20**(1), 228–243 (1998)
20. Karandikar, R.L., Kulkarni, V.G.: Second-order fluid flow models: reflected Brownian motion in a random environment. Oper. Res. **43**, 77–88 (1995)
21. Lancaster, P., Rodman, L.: Algebraic Riccati Equations. Oxford University Press, New York (1995)
22. Meyer, C.D.: Stochastic complementation, uncoupling Markov chains, and the theory of nearly reducible systems. SIAM Rev. **31**, 240–272 (1989)
23. Nguyen, G.T., Poloni, F.: Componentwise accurate fluid queue computations using doubling algorithms. Numer. Math. **130**(4), 763–792 (2015)
24. Nguyen, G.T., Poloni, F.: Componentwise accurate Brownian motion computations using Cyclic Reduction (2016). arXiv:1605.01482
25. O'Cinneide, C.A.: Entrywise perturbation theory and error analysis for Markov chains. Numer. Math. **65**, 109–120 (1993)
26. Rogers, L.: Fluid models in queueing theory and Wiener–Hopf factorization of Markov chains. Ann. Appl. Probab. **4**, 390–413 (1994)
27. Varga, R.: Matrix Iterative Analysis. Prentice Hall, Englewood Cliffs (1962)
28. Wang, W.G., Wang, W.C., Li, R.-C.: Alternating-directional doubling algorithm for M-matrix algebraic Riccati equations. SIAM J. Matrix Anal. Appl. **33**(1), 170–194 (2012)
29. Xue, J., Li, R.-C.: Highly accurate doubling algorithms for M-matrix algebraic Riccati equations. Numer. Math. **135**(3), 733–767 (2017)
30. Xue, J., Xu, S., Li, R.-C.: Accurate solutions of M-matrix algebraic Riccati equations. Numer. Math. **120**(4), 671–700 (2012)
31. Xue, J., Xu, S., Li, R.-C.: Accurate solutions of M-matrix Sylvester equations. Numer. Math. **120**(4), 639–670 (2012)