Linear Algebra and its Applications 583 (2019) 1–45

Contents lists available at ScienceDirect

Linear Algebra and its Applications

www.elsevier.com/locate/laa

Highly accurate doubling algorithm for quadratic matrix equation from quasi-birth-and-death process

Cairong Chen [a,1], Ren-Cang Li [b,*,2], Changfeng Ma [c,3]

[a] *School of Mathematics and Systems Science, Beihang University, Beijing, 100191, PR China*
[b] *Department of Mathematics, University of Texas at Arlington, P.O. Box 19408, Arlington, TX 76019-0408, USA*
[c] *College of Mathematics and Informatics, Fujian Normal University, Fuzhou 350007, PR China*

ARTICLE INFO

Article history:
Received 19 October 2018
Accepted 16 August 2019
Available online 23 August 2019
Submitted by B. Meini

MSC:
15A24
65F30
65H10

Keywords:
Doubling algorithm
Quasi-birth-and-death process
Nonnegative solution
Entrywise relative accuracy
SF1

ABSTRACT

A highly accurate doubling algorithm to solve the most fundamental quadratic matrix equation in the quasi-birth-and-death (QBD) process is developed. It follows from the general framework of the doubling algorithm for the first standard form (SF1) but can be implemented to compute the minimal nonnegative solution with high entrywise relative accuracy for all entries, large or tiny. The algorithm is globally and quadratically convergent, except for QBD equations in the critical case where convergence is linear with the linear rate 1/2. Numerical examples are presented to demonstrate and confirm our claims. The development here parallels the recent work of Xue and Li (2017) [3] on the M-matrix algebraic Riccati equation.

* Corresponding author.
E-mail addresses: chencr0611@163.com (C. Chen), rcli@uta.edu (R.-C. Li), macf@fjnu.edu.cn (C. Ma).

[1] Part of this work was done while this author was visiting the University of Texas at Arlington from February 2017 to January 2018.

[2] Supported in part by NSF grants CCF-1527104 and DMS-1719620.
[3] Supported in part by NNSFC grant 11071041 and Fujian NSF grants 2016J01005 and 2015J01578.

https://doi.org/10.1016/j.laa.2019.08.018
0024-3795/

1. Introduction

One of the most fundamental quadratic matrix equations in the quasi-birth-and-death (QBD) process is

$$A_0 + A_1X + A_2X^2 = X, \tag{1.1}$$

where A_0, A_1 and A_2 are blocks in an infinite block-tridiagonal transition matrix [1, Chapter 8] and are $n \times n$ nonnegative matrices. In the application, $I - A_0 - A_1 - A_2$ is also irreducible and singular, and, in particular,

$$(A_0 + A_1 + A_2)\mathbf{1}_n = \mathbf{1}_n, \tag{1.2}$$

where $\mathbf{1}_n$ (often simply $\mathbf{1}$ when its dimension is clear from the context) is the column n-vector of all ones. It is known that (1.1) has a minimal nonnegative solution Φ for which $\Phi\mathbf{1} \leq \mathbf{1}$ [1, pp. 168–172]. By a minimal nonnegative solution, we mean it is a solution to (1.1) and also satisfies entrywise

$$0 \leq \Phi \leq X \quad \text{for any nonnegative solution } X \text{ to (1.1)}.$$

This solution Φ is the one of interest. By definition such a minimal nonnegative solution Φ is necessarily unique. In the QBD process, the entries of Φ represent probabilities of events. Larger entries correspond to frequent events, while rare events result in tiny entries. Usually frequent events are more important than less frequent ones, but sometimes rare events can be practically significant, too, and hence getting tiny entries right can be critically useful. There are examples in which the entries of Φ vary from as tiny as $O(10^{-50})$ to $O(1)$. On the other hand, conventional wisdom suggests that any entries that are of $O(10^{-16})$ have little chance to be computed to even a single correct decimal digit in the IEEE double precision floating point environment. That is bad news. Fortunately, as we will demonstrate in this paper, the QBD equation (1.1) is special enough that we can compute Φ in a clever way by the doubling algorithm to high entrywise relative accuracy as warranted by the data.

Our work here is inspired by recent studies on M-matrix algebraic Riccati equations (MARE) [2–4] and by Ye's highly accurate implementation of the Latouche-Ramaswami algorithm [5,6].

Current methods for the QBD equation include the Latouche-Ramaswami algorithm [5,6], the method of cyclic reductions [7–9], Newton's method [10], and a few other fixed point iterative methods (see, e.g., [11,12] and references therein). The fixed point iterative methods are usually linearly convergent and sometimes can be very slow, and for this reason we will not discuss them hereafter. All other methods are quadratically convergent unless the involved QBD equation is in the critical case (see section 4 for definition). Because there is a generalized Sylvester equation to solve in each Newton's

iterative step, which, unfortunately, can be as expensive as solving the QBD equation itself by other methods, Newton's method is not competitive. As far as computational cost is concerned, the method of cyclic reduction is comparable, but a highly accurate implementation has not yet been developed. Ye's highly accurate implementation of Latouche-Ramaswami algorithm appears to be the only method known today to be able to compute Φ entrywise accurately.

Throughout the rest of this paper, we will broadly consider the quadratic matrix equation of form (1.1) that includes the original ones from the QBD process as special cases. Specifically, we assume, besides $A_i \geq 0$ for $0 \leq i \leq 2$ throughout the rest of this paper, that either

$$I - A_0 - A_1 - A_2 \text{ is a nonsingular } M\text{-matrix}, \tag{1.3a}$$

or

$$\boxed{\begin{array}{l} I - A_0 - A_1 - A_2 \text{ is an irreducible singular } M\text{-matrix} \\ \text{and } A_0 \text{ and } A_2 \text{ are nonzero matrices.} \end{array}} \tag{1.3b}$$

We will call any equation (1.1) that satisfies (1.3) *a* QBD *equation*. The case when one of A_0 and A_2 is zero can be considered trivial. In fact, if $A_0 = 0$, then $X = 0$ is clearly the minimal nonnegative solution; if $A_2 = 0$, then (1.1) becomes $(I - A_1)X = A_0$ which has a unique solution whenever $I - A_1$ is nonsingular.

Equation (1.1) originally from the QBD process falls into (1.3b), but not (1.3a). The Latouche-Ramaswami algorithm [13,5,6] with minor modifications will work for the case (1.3). More comments will come later in section 9.

It can be verified that (1.1) is equivalent to

$$\mathscr{A}\begin{bmatrix} I \\ X \end{bmatrix} := \begin{bmatrix} 0 & I \\ A_0 & A_1 - I \end{bmatrix}\begin{bmatrix} I \\ X \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & -A_2 \end{bmatrix}\begin{bmatrix} I \\ X \end{bmatrix} M =: \mathscr{B}\begin{bmatrix} I \\ X \end{bmatrix} M, \tag{1.4}$$

where $M \in \mathbb{R}^{n\times n}$. Necessarily, $M = X$. Now if also $I - A_1$ is nonsingular, then we can set

$$P_1 = \begin{bmatrix} I & 0 \\ 0 & -(I - A_1)^{-1} \end{bmatrix}, \quad P_2 = \begin{bmatrix} I & -I \\ 0 & I \end{bmatrix}.$$

We have

$$\mathscr{A}_0 := P_2P_1\mathscr{A} = \begin{bmatrix} (I - A_1)^{-1}A_0 & 0 \\ -(I - A_1)^{-1}A_0 & I \end{bmatrix} =: \begin{array}{c} \\ n \\ n \end{array}\begin{array}{c} \begin{array}{cc} n & n \end{array} \\ \begin{bmatrix} E_0 & 0 \\ -X_0 & I \end{bmatrix} \end{array}, \tag{1.5a}$$

$$\mathscr{B}_0 := P_2P_1\mathscr{B} = \begin{bmatrix} I & -(I-A_1)^{-1}A_2 \\ 0 & (I-A_1)^{-1}A_2 \end{bmatrix} =: \begin{matrix} n \\ n \end{matrix} \overset{\begin{matrix} n & n \end{matrix}}{\begin{bmatrix} I & -Y_0 \\ 0 & F_0 \end{bmatrix}} \tag{1.5b}$$

which is in the first standard form, (SF1) in short, as defined in [14] (see also [15]). Moreover, pre-multiply (1.4) by P_2P_1 to get

$$\mathscr{A}_0 \begin{bmatrix} I \\ X \end{bmatrix} = \mathscr{B}_0 \begin{bmatrix} I \\ X \end{bmatrix} X. \tag{1.6}$$

Now that the matrix pencil $\mathscr{A}_0 - \lambda\mathscr{B}_0$ is in (SF1), it is natural for us to apply the doubling algorithm for (SF1) [14] to solve (1.6).

The goals of this paper are twofold: to analyze the convergence of the doubling algorithm for (SF1) on solving (1.6) and to present a highly accurate implementation of it to compute the minimal nonnegative solution Φ to (1.1) to high entrywise relative accuracy in the sense that each entry of Φ, regardless of its magnitude, will be computed to almost full machine accuracy. This is important, because in the QBD process, Φ is used later to determine a matrix-geometric stationary distribution [16]. A highly entrywise accurate Φ will lead to a highly entrywise accurate stationary distribution.

Many results regarding the QBD equation (1.1) itself and its associated polynomial $\phi(\lambda) := \det(A_0 + \lambda(A_1 - I) + \lambda^2 A_2)$ are not altogether new, but we attempt to present them in a coherent way based on matrix analysis techniques for easy access by the numerical linear algebra community.

The rest of this paper is organized as follows. In section 2 we state a few basic results on nonnegative and M-matrices, relevant to our later developments. We present an elementary proof about the existence of the minimal nonnegative solution in section 3. Section 4 characterizes the spectral properties of the associated matrix pencil $\mathscr{A} - \lambda\mathscr{B}$ defined in (1.4). The plain doubling algorithm for the QBD equation and its convergence analysis are given in sections 5 and 6, respectively. Section 7 introduces a new entrywise relative residual that will be demonstrated more appropriate than the usual normalized residual, similar to the same concept in [3] for MARE. In section 8, we present our highly accurate doubling algorithm for the QBD equation (1.1), following the same line as accADDA of [3]. Numerical examples are given in section 9 to demonstrate our points made in the previous sections. Conclusions are made in section 10. Finally, in appendix Appendix A we investigate the sparsity pattern in approximations by the doubling algorithm.

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 $\mathbb{R} = \mathbb{R}^1$. When $\mathbb{R}$ is replaced by $\mathbb{C}$, these sets are understood as in the field of complex numbers. I_n (or simply I if its dimension is clear from the context) is the $n \times n$ identity matrix. The

superscript "$\cdot^{\mathrm{T}}$" takes transpose. 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 a matrix X, $\mathcal{R}(X)$ and $\mathcal{N}(X)$ are the column space and the null space of X, respectively. When X is square, we denote by $\rho(X)$ its spectral radius and by $\mathrm{eig}(X)$ its spectrum. Given $A, B \in \mathbb{C}^{n\times n}$ such that $A - \lambda B$ is a regular matrix pencil,[4] we denote by $\mathrm{eig}(A,B) := \{\mu \in \mathbb{C} : \det(A - \mu B) = 0\}$ the spectrum of the matrix pencil. $\mathbf{1}_n \in \mathbb{R}^n$ is the n-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. The symbol $\mathfrak{u}$ is the unit machine roundoff, and it is $2^{-53} \approx 1.1 \times 10^{-16}$ for the IEEE double precision.

2. Preliminaries

In this section, we collect a few important results on nonnegative matrices and M-matrices. These results are well-known and can be found in, e.g., [18,19]. They lay the foundation of our technical arguments.

A matrix $A \in \mathbb{R}^{m\times n}$ is *nonnegative*, denoted by $A \ge 0$, if all of its entries are nonnegative, and *positive*, denoted by $A > 0$, if all its entries are positive. The same understanding goes to vectors. In Theorem 2.1, we summarize a few relevant results on nonnegative matrices. They are parts of the Perron-Frobenius theory.

Theorem 2.1. *Let $A \in \mathbb{R}^{n\times n}$ be a nonnegative matrix.*

(a) *The spectral radius, $\rho(A)$, is an eigenvalue of A. If A is also irreducible, then $\rho(A)$ is a simple eigenvalue and positive.*
(b) *There exist a nonnegative right eigenvector $\boldsymbol{x}$ and a nonnegative left eigenvector $\boldsymbol{y}$ associated with the eigenvalue $\rho(A)$: $A\boldsymbol{x} = \rho(A)\boldsymbol{x}$ and $\boldsymbol{y}^{\mathrm{T}}A = \rho(A)\boldsymbol{y}^{\mathrm{T}}$. If A is also irreducible, then $\boldsymbol{x} > 0$ and $\boldsymbol{y} > 0$.*
(c) *Let $B \in \mathbb{R}^{n\times n}$. If $B \ge A$, then $\rho(B) \ge \rho(A)$. If B is also irreducible and $B \ne A$, then $\rho(B) > \rho(A)$.*
(d) *If $A\boldsymbol{x} \le \beta\boldsymbol{x}$ for some $\boldsymbol{x} > 0$, then $\rho(A) \le \beta$. If, in addition, also $A\boldsymbol{x} \ne \beta\boldsymbol{x}$, then $\rho(A) < \beta$.*
(e) *If $\boldsymbol{x}$ is a positive eigenvector of A, then $\boldsymbol{x}$ corresponds to $\rho(A)$.*

A matrix $A \in \mathbb{R}^{n\times n}$ is called a *Z-matrix* if $A_{(i,j)} \le 0$ for all $i \ne j$. Any Z-matrix A can be written as $sI - N$ with $N \ge 0$, and it is called an *M-matrix* if $s \ge \rho(N)$. Specifically, it is a *singular M-matrix* if $s = \rho(N)$, and a *nonsingular M-matrix* if $s > \rho(N)$. Theorem 2.2 lists four equivalent statements for a nonsingular M-matrix.

[4] I.e., $\det(A - \lambda B) \not\equiv 0$ for $\lambda \in \mathbb{C}$ [17].

Theorem 2.2. *Let $A \in \mathbb{R}^{n\times n}$ be a Z-matrix. Then the following statements are equivalent:*

(a) *A is a nonsingular M-matrix;*
(b) *$A^{-1} \geq 0$;*
(c) *$A\boldsymbol{u} > 0$ for some positive vector $\boldsymbol{u} \in \mathbb{R}^n$;*
(d) *All eigenvalues of A are in the open right half plane.*

Two other results on M-matrices that are useful to us are stated in the next theorem.

Theorem 2.3. *Let $A \in \mathbb{R}^{n\times n}$ be an M-matrix.*

(a) *If A is singular and irreducible, then 0 is a simple eigenvalue and its corresponding eigenvector $\boldsymbol{v}$ can be taken positive.*
(b) *Let $B \in \mathbb{R}^{n\times n}$ be a Z-matrix. If A is nonsingular and $B \geq A$, then B is also a nonsingular M-matrix.*

The key ingredient in recent work [2,3,6] to achieve high entrywise relative accuracy in solving certain nonlinear matrix equations is the GTH-like algorithm for inverting a nonsingular M-matrix due to Alfa, Xue, and Ye [20]. It is made possible by a brilliant idea of theirs: that is to represent a nonsingular M-matrix A in an alternative way, the so-called *triplet representation* of A. In particular, the representation determines A^{-1} entrywise to high relative accuracy. The reader is referred to [21, section 2] for a brief survey.

An M-matrix A can have infinite many triplet representations, but for the purpose of computation, any one is just as good as any other. A triplet representation $\{N_A, \boldsymbol{u}, \boldsymbol{v}\}$ of the M-matrix $A \in \mathbb{R}^{n\times n}$ consists of

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

where $\mathrm{diag}(A)$ is the diagonal matrix obtained from extracting the diagonal part of A. For convenience, we will not distinguish A from its triplet representation and write $A = \{N_A, \boldsymbol{u}, \boldsymbol{v}\}$ whenever it is more convenient to do so.

The main theoretical contribution in [22] is that if all entries of N_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, using the trick of [23], Alfa, Xue, and Ye [20] presented the GTH-like algorithm to compute the LU decomposition of $A = \{N_A, \boldsymbol{u}, \boldsymbol{v}\}$, via the Gaussian elimination without pivoting, without any cancellation and, consequently, to compute the solution $\boldsymbol{x}$ of $A\boldsymbol{x} = \boldsymbol{b} \geq 0$ to the claimed accuracy. For more detail, the reader is referred to [14].

3. Minimal nonnegative solution

Earlier, we mentioned that in the QBD application,[5] (1.1) has a unique minimal nonnegative solution. In this section, we will show that it is still true under, more generally, (1.3).

The following observation, as a result of the assumption (1.3), plays a critical role in the argument that follows. Since $I - A_0 - A_1 - A_2$ is either a nonsingular M-matrix or an irreducible and singular M-matrix, by Theorems 2.2 and 2.3(a), there exists a positive vector $\boldsymbol{u} > 0$ in $\mathbb{R}^n$ such that

$$\boldsymbol{v} = (I - A_0 - A_1 - A_2)\boldsymbol{u} \begin{cases} > 0, & \text{in the case of (1.3a)}, \\ = 0, & \text{in the case of (1.3b)}. \end{cases} \tag{3.1}$$

Throughout the rest of this paper, $\boldsymbol{v}$ and $\boldsymbol{u}$ are reserved for the ones here. For the QBD equation (1.1) originally from the QBD process, $\boldsymbol{u} = \mathbf{1}_n$ and $\boldsymbol{v} = 0$.

Lemma 3.1. *Suppose* (1.3). *Then* $I - A_1$ *is a nonsingular* M*-matrix. In particular,* $(I - A_1)^{-1} \geq 0$.

Proof. Since $I - A_1 \geq I - A_0 - A_1 - A_2$, $I - A_1$ is a nonsingular M-matrix under (1.3a) by Theorem 2.3(b). Under (1.3b), we have $(A_0 + A_1 + A_2)\boldsymbol{u} = \boldsymbol{u}$ by (3.1) and thus $\rho(A_0 + A_1 + A_2) = 1$ by Theorem 2.1. Since

$$A_1 \leq A_0 + A_1 + A_2 \text{ and } A_1 \neq A_0 + A_1 + A_2,$$

it follows from Theorem 2.1(c) that $\rho(A_1) < 1$, which implies that $I - A_1$ is also a nonsingular M-matrix under (1.3b). □

Part of next theorem for the case of (1.3b) was well known in the QBD application. The theorem as a whole is also implied by [24, Theorem 2.3]. However, it seems to be the first time that the inequality (3.2) is explicitly formulated.

Theorem 3.2. *Under the assumption* (1.3), *the quadratic equation* (1.1) *has a unique minimal nonnegative solution* Φ. *Moreover, it holds that* $\Phi \geq X_0$ *and*

$$\Phi\boldsymbol{u} \leq \boldsymbol{u} - (I - A_1)^{-1}\boldsymbol{v}, \tag{3.2}$$

where $X_0 = (I - A_1)^{-1}A_0$ *is as defined in* (1.5a).

[5] In the application, $A_i \geq 0$ for $i = 0, 1, 2$ and $I - A_0 - A_1 - A_2$ is irreducible and singular, and (1.2) holds.

Proof. We still present the following constructive proof because some part of the proof is needed later for concluding important solution properties.

Because of Lemma 3.1, the following matrix-valued function

$$G(X) = (I - A_1)^{-1}(A_0 + A_2X^2)$$

from $\mathbb{R}^{n\times n}$ to $\mathbb{R}^{n\times n}$ is well-defined. Now construct the sequence $\{Z_k\}_{k=0}^{\infty}$ by

$$Z_0 = 0 \quad \text{and} \quad Z_{k+1} = G(Z_k) \text{ for } k \geq 0. \tag{3.3}$$

This is essentially a special case of a more general traditional iterative scheme [16, (1.2.19) on p. 13] (see also [25, (6.7) on p. 144]) for solving a nonlinear matrix equation in a Markov chain. In particular, $Z_1 = X_0$. We claim that for $k \geq 0$

$$0 \leq Z_k \leq Z_{k+1} \text{ and } Z_k\boldsymbol{u} \leq \boldsymbol{u} - (I - A_1)^{-1}\boldsymbol{v}. \tag{3.4}$$

To see this, we note $Z_1 = (I - A_1)^{-1}A_0 \geq 0 = Z_0$ and

$$\begin{aligned}\boldsymbol{u} - (I - A_1)^{-1}\boldsymbol{v} &= (I - A_1)^{-1}\left[(I - A_1)\boldsymbol{u} - \boldsymbol{v}\right] \\ &= (I - A_1)^{-1}(A_0 + A_2)\boldsymbol{u} \\ &\geq 0 = Z_0\boldsymbol{u}.\end{aligned}$$

That is the inequalities in (3.4) are valid for $k = 0$. Now suppose that they hold for $k = \ell$. We have

$$Z_{\ell+2} - Z_{\ell+1} = (I - A_1)^{-1}A_2(Z_{\ell+1}^2 - Z_\ell^2) \geq 0,$$

because $Z_{\ell+1}^2 - Z_\ell^2 = Z_{\ell+1}(Z_{\ell+1} - Z_\ell) + (Z_{\ell+1} - Z_\ell)Z_\ell \geq 0$. Thus, $Z_{\ell+2} \geq Z_{\ell+1} \geq Z_\ell \geq 0$, which gives the first part of (3.4) for $k = \ell + 1$. Also, we have, upon using $Z_\ell\boldsymbol{u} \leq \boldsymbol{u} - (I - A_1)^{-1}\boldsymbol{v} \leq \boldsymbol{u}$,

$$\begin{aligned}Z_{\ell+1}\boldsymbol{u} &= (I - A_1)^{-1}(A_0 + A_2Z_\ell^2)\boldsymbol{u} \\ &\leq (I - A_1)^{-1}A_0\boldsymbol{u} + (I - A_1)^{-1}A_2\boldsymbol{u} \\ &= (I - A_1)^{-1}(A_0 + A_2)\boldsymbol{u} \\ &= (I - A_1)^{-1}\left[(I - A_1)\boldsymbol{u} - \boldsymbol{v}\right] \\ &= \boldsymbol{u} - (I - A_1)^{-1}\boldsymbol{v},\end{aligned}$$

which proves the second part of (3.4) for $k = \ell + 1$. This completes the proof of (3.4) for $k \geq 0$.

By (3.4), the sequence $\{Z_k\}_{k=0}^{\infty}$ is monotonically increasing and bounded from above. So it converges. Let Φ be the limit. Evidently, it is nonnegative and $X_0 = Z_1 \leq \Phi$.

Letting $k \to \infty$ in $Z_{k+1} = G(Z_k)$, we find that Φ is a nonnegative solution of (1.1). Letting $k \to \infty$ in (3.4), we find that Φ satisfies (3.2).

Lastly, we claim that it is minimal among all nonnegative solutions of (1.1). To this end, let X be any nonnegative solution of (1.1). First, $Z_0 = 0 \le X$. Suppose $Z_k \le X$ holds for $k = \ell$. Then

$$\begin{aligned} Z_{\ell+1} - X &= (I - A_1)^{-1}(A_0 + A_2 Z_\ell^2 - A_0 - A_2 X^2) \\ &= (I - A_1)^{-1} A_2 (Z_\ell^2 - X^2) \le 0. \end{aligned}$$

By the induction principle, $Z_k \le X$ for all $k \ge 0$. Let $k \to \infty$ in $Z_k \le X$ conclude $\Phi \le X$, i.e., Φ is the minimal nonnegative solution of (1.1). □

Lemma 3.3. *Suppose* (1.3a), *i.e.,* $I - A_0 - A_1 - A_2$ *is a nonsingular M-matrix. Then* $\rho(X) \ne 1$ *for any nonnegative solution* X *of* (1.1).

Proof. Suppose, to the contrary, that $\rho(X) = 1$, where X is a nonnegative solution of (1.1). Then according to Theorem 2.1(b), there exists a nonzero and nonnegative vector $\boldsymbol{z} \in \mathbb{R}^n$ such that $X\boldsymbol{z} = \boldsymbol{z}$ and thus

$$(A_0 + A_1 X + A_2 X^2)\boldsymbol{z} = X\boldsymbol{z} = \boldsymbol{z}$$

to give $(I - A_0 - A_1 - A_2)\boldsymbol{z} = 0$. This contradicts the assumption that $I - A_0 - A_1 - A_2$ is nonsingular. □

We end this section by introducing the *dual equation* of (1.1):

$$A_2 + A_1 Y + A_0 Y^2 = Y. \tag{3.5}$$

It differs from (1.1), which we will call the *primal equation*, slightly in that the roles of A_0 and A_2 are switched. Since our main assumption (1.3) is symmetrical with respect to the roles of A_0 and A_2. Theorem 3.2 is applicable to (3.5). In particular, this dual equation (3.5) also has a unique minimal nonnegative solution, denoted by Ψ hereafter. For convenience, we summarize in Theorem 3.4 below some of the important results that will be useful to us later, but point out that item (c) will be expanded with more detail later in Theorems 4.5 and 4.6.

Theorem 3.4. *Suppose* (1.3), *and let* Φ *and* Ψ *be the minimal nonnegative solutions to* (1.1) *and* (3.5), *respectively. The following statements hold.*

(a) *We have*

$$0 \le X_0 = (I - A_1)^{-1} A_0 \le \Phi, \quad \Phi\boldsymbol{u} \le \boldsymbol{u} - (I - A_1)^{-1}\boldsymbol{v}, \tag{3.6a}$$

$$0 \le Y_0 = (I - A_1)^{-1} A_2 \le \Psi, \quad \Psi\boldsymbol{u} \le \boldsymbol{u} - (I - A_1)^{-1}\boldsymbol{v}. \tag{3.6b}$$

(b) $\rho(\Phi) < 1$ *and* $\rho(\Psi) < 1$ *under* (1.3a).
(c) $\rho(\Phi) \leq 1$ *and* $\rho(\Psi) \leq 1$ *under* (1.3b).
(d) $I - \Phi\Psi$ *and* $I - \Psi\Phi$ *are* M*-matrices and they are nonsingular under* (1.3a).

Proof. Item (a) is a consequence of Theorem 3.2.

We have $\Phi \boldsymbol{u} \leq \boldsymbol{u}$ by (3.6a). Since $\boldsymbol{u} > 0$, we conclude $\rho(\Phi) \leq 1$ by Theorem 2.1(d), and thus item (c).

It can be verified that $\Phi\Psi \boldsymbol{u} \leq \boldsymbol{u}$ and $\Psi\Phi \boldsymbol{u} \leq \boldsymbol{u}$ and thus both $I - \Phi\Psi$ and $I - \Psi\Phi$ are M-matrices. In the case of (1.3a), we know $\rho(\Phi) \neq 1$ because of Lemma 3.3 and thus $\rho(\Phi) < 1$. Since $(I - A_1)^{-1} \geq 0$ and $\boldsymbol{v} > 0$, we deduce from (3.6) that $\Phi \boldsymbol{u} < \boldsymbol{u}$ and $\Psi \boldsymbol{u} < \boldsymbol{u}$. Therefore, $\Phi\Psi \boldsymbol{u} \leq \Phi \boldsymbol{u} < \boldsymbol{u}$ and $\Psi\Phi \boldsymbol{u} \leq \Psi \boldsymbol{u} < \boldsymbol{u}$. Consequently, $\rho(\Phi\Psi) = \rho(\Psi\Phi) < 1$, implying that $I - \Phi\Psi$ and $I - \Psi\Phi$ are nonsingular M-matrices because they are clearly Z-matrices. □

It can be verified that (3.5) is equivalent to

$$\mathscr{A} \begin{bmatrix} Y \\ I \end{bmatrix} N = \mathscr{B} \begin{bmatrix} Y \\ I \end{bmatrix}, \tag{3.7}$$

where $\mathscr{A} - \lambda\mathscr{B}$ is the same as the one defined in (1.4). Necessarily, $N = Y$ in (3.7). Because of the way $\mathscr{A}_0 - \lambda\mathscr{B}_0$ in (1.5) is constructed, we also have

$$\mathscr{A}_0 \begin{bmatrix} Y \\ I \end{bmatrix} N = \mathscr{B}_0 \begin{bmatrix} Y \\ I \end{bmatrix}. \tag{3.8}$$

In [14], the dual equation is introduced through the so-called *dual matrix pencil* of $\mathscr{A}_0 - \lambda\mathscr{B}_0$ defined as

$$\mathscr{A}_0^{(\mathrm{d})} - \lambda\mathscr{B}_0^{(\mathrm{d})} := \Pi^{\mathrm{T}}(\mathscr{B}_0 - \lambda\mathscr{A}_0)\Pi \quad \text{with} \quad \Pi = \begin{bmatrix} 0 & I_n \\ I_n & 0 \end{bmatrix}. \tag{3.9}$$

With it, (3.8) becomes

$$\mathscr{A}_0^{(\mathrm{d})} \begin{bmatrix} I \\ Y \end{bmatrix} = \mathscr{B}_0^{(\mathrm{d})} \begin{bmatrix} I \\ Y \end{bmatrix} N. \tag{3.8'}$$

One implication of (3.9) is that the eigenvalues of $\mathscr{A}_0^{(\mathrm{d})} - \lambda\mathscr{B}_0^{(\mathrm{d})}$ are exactly the reciprocals of those of $\mathscr{A}_0 - \lambda\mathscr{B}_0$.

Henceforward, notations Φ and Ψ are reserved for the minimal nonnegative solutions, if exist, to (1.1) and (3.5), respectively.

4. Spectrum of $\mathscr{A} - \lambda\mathscr{B}$

Let $\mathscr{A} - \lambda\mathscr{B}$ be defined by (1.4), which is a linearization of the quadratic eigenvalue problem (QEP) [26]

$$[A_0 + \lambda(A_1 - I) + \lambda^2 A_2]\boldsymbol{x} = 0. \tag{4.1}$$

We will assume that $\mathscr{A} - \lambda\mathscr{B}$ is regular,[6] i.e., $\det(\mathscr{A} - \lambda\mathscr{B}) \not\equiv 0$ for $\lambda \in \mathbb{C}$. We have

$$\det(\mathscr{A} - \lambda\mathscr{B}) = (-1)^n \det(A_0 + \lambda(A_1 - I) + \lambda^2 A_2) = (-1)^n \phi(\lambda), \tag{4.2}$$

where $\phi(\lambda) := \det(A_0 + \lambda(A_1 - I) + \lambda^2 A_2)$. To see this, we note

$$\begin{aligned}(\mathscr{A} - \lambda\mathscr{B}) \begin{bmatrix} I_n & I_n \\ 0 & \lambda I_n \end{bmatrix} &= \begin{bmatrix} -\lambda I_n & I_n \\ A_0 & A_1 - I_n + \lambda A_2 \end{bmatrix} \begin{bmatrix} I_n & I_n \\ 0 & \lambda I_n \end{bmatrix} \\ &= \begin{bmatrix} -\lambda I_n & 0 \\ A_0 & A_0 + \lambda(A_1 - I) + \lambda^2 A_2 \end{bmatrix}.\end{aligned}$$

Take their determinants to get

$$\lambda^n \det(\mathscr{A} - \lambda\mathscr{B}) = (-\lambda)^n \det(A_0 + \lambda(A_1 - I) + \lambda^2 A_2),$$

leading to (4.2). This means that the eigenvalues of the matrix pencil $\mathscr{A} - \lambda\mathscr{B}$ are the same as those of QEP (4.1), which are the zeros of $\phi(\lambda)$ defined in (4.2). In the case when A_2 is singular, by convention that is widely used to treat a regular matrix pencil such as $\mathscr{A} - \lambda\mathscr{B}$ here with a singular $\mathscr{B}$ [17], ∞ is counted as a zero with an appropriate algebraic multiplicity so that the total number of zeros of $\phi(\lambda)$ is always $2n$.

Theorem 4.1. *Suppose* (1.3a). *Then $\mathscr{A} - \lambda\mathscr{B}$ has exactly n eigenvalues in the open unit disk and n eigenvalues outside of the closed unit disk; so do $\mathscr{A}_0 - \lambda\mathscr{B}_0$ in* (1.5) *and $\mathscr{A}_0^{(\mathrm{d})} - \lambda\mathscr{B}_0^{(\mathrm{d})}$ in* (3.9).

Proof. Let Φ and Ψ be the minimal nonnegative solution to (1.1) and (3.5), respectively. By Theorem 3.4(d), both $I - \Phi\Psi$ and $I - \Psi\Phi$ are invertible and thus

$$\begin{bmatrix} I & \Psi \\ \Phi & I \end{bmatrix}^{-1} = \begin{bmatrix} (I - \Psi\Phi)^{-1} & -\Psi(I - \Phi\Psi)^{-1} \\ -(I - \Phi\Psi)^{-1}\Phi & (I - \Phi\Psi)^{-1} \end{bmatrix}.$$

On the other hand, we have (1.4) with $X = \Phi$ and (3.7) with $Y = \Psi$, yielding

$$\mathscr{A} \begin{bmatrix} I & \Psi \\ \Phi & I \end{bmatrix} \begin{bmatrix} I & 0 \\ 0 & \Psi \end{bmatrix} = \mathscr{B} \begin{bmatrix} I & \Psi \\ \Phi & I \end{bmatrix} \begin{bmatrix} \Phi & 0 \\ 0 & I \end{bmatrix}. \tag{4.3}$$

This implies that $\mathrm{eig}(\mathscr{A}, \mathscr{B})$ is the multiset union of $\mathrm{eig}(\Phi)$ and $\{1/\lambda : \lambda \in \mathrm{eig}(\Psi)\}$. The conclusion follows because $\rho(\Phi) < 1$ and $\rho(\Psi) < 1$ by Theorem 3.4(b). □

[6] This is the same as requiring $\phi(\lambda) \not\equiv 0$ for $\lambda \in \mathbb{C}$. This requirement does not follow from the assumptions in (1.3) because they do not exclude the situation, e.g., where A_0, A_1, and A_2 all have two parallel columns/rows, although such a situation cannot happen in any practical QBD process.

Complication for characterizing $\mathrm{eig}(\mathscr{A}, \mathscr{B})$ arises for the case (1.3b). We recall the positive vector $\boldsymbol{u}$ in (3.1) which satisfies

$$(A_0 + A_1 + A_2)\boldsymbol{u} = \boldsymbol{u}. \tag{4.4a}$$

Therefore $A_0 + (A_1 - I) + A_2$ is singular and thus $\phi(1) = 0$ and $1 \in \mathrm{eig}(\mathscr{A}, \mathscr{B})$. By Theorem 2.1, there is a positive vector $\boldsymbol{z}$ in $\mathbb{R}^n$ such that

$$\boldsymbol{z}^{\mathrm{T}}(A_0 + A_1 + A_2) = \boldsymbol{z}^{\mathrm{T}}. \tag{4.4b}$$

Define

$$\mu = \boldsymbol{z}^{\mathrm{T}}(A_2 - A_0)\boldsymbol{u}. \tag{4.5}$$

Since the QBD process can be viewed as an M/G/1-type Markov chain or a G/M/1-type Markov chain [25, p. 127], the spectral analysis in [27] can be applied to lead to a complete description on how the zeros of $\phi(\lambda)$ distribute relative to the unit circle, depending on the quantity [28, p. 541]

$$\max\left\{\text{integer } k \geq 0 \;\middle|\; \begin{array}{l} z^{-n/k}\det(A_0 + z^{1/k}(A_1 - I) + z^{2/k}A_2) \\ \text{is single-valued in } |z| \leq 1 \end{array}\right\}. \tag{4.6}$$

For the same purpose, [25, Theorem 5.20 on p. 128] presents a complete description under two conditions, **Conditions** 5.1 and 5.2 on [25, pp. 110–111]. However, the quantity in (4.6) is not intuitive and difficulty to find out numerically (if at all possible). At the same time, it will take quite some page space to clearly explain **Condition** 5.2 to someone who does not work in the area of Markov chains. For this reason and in order to well serve the numerical linear algebra community, in what follows, we will adopt a matrix analysis approach to investigate the zeros of $\phi(\lambda)$, under mild assumptions. Some results are not completely new, however.

Two results in Lemmas 4.2 and 4.4 from [16, section 1.3] form the foundation of the approach. Let

$$\chi(t) = \rho(A_0 + tA_1 + t^2A_2) \quad \text{for } 0 < t \leq 1. \tag{4.7}$$

We claim that $\ln \chi(e^{-s})$ is convex for $0 < s < \infty$ [16, p. 15]. This is because the (i, j)th entry of $A_0 + tA_1 + t^2A_2$ is either identically 0 or take the form $\alpha_0 + \alpha_1 t + \alpha_2 t^2$, where α_i $(0 \leq i \leq 1)$ are nonnegative and at least one of them is positive. We have

$$\frac{d^2}{ds^2}\ln(\alpha_0 + \alpha_1 e^{-s} + \alpha_2 e^{-2s}) = \frac{\alpha_0\alpha_1 e^{-s} + \alpha_1\alpha_2 e^{-3s} + 4\alpha_2\alpha_0 e^{-2s}}{[\alpha_0 + \alpha_1 e^{-s} + \alpha_2 e^{-2s}]^2} > 0.$$

By [29, Corollary 1], $\ln \chi(e^{-s})$ is convex for $0 < s < \infty$.

Lemma 4.2 *([16, p. 17]). Suppose* (1.3b) *and consider the equation*

$$t = \chi(t) \quad \textit{for } 0 < t \leq 1. \tag{4.8}$$

The equation has at least a root 1 *but at most two roots, and also* $1 \in \mathrm{eig}(\mathscr{A}, \mathscr{B})$. *Specifically,*

(a) *if* $\mu \leq 0$, *then* $t = 1$ *is the only root;*
(b) *if* $\mu > 0$ *and if* $\rho(A_0) > 0$, *then it has a second root* t_0, *where* $0 < t_0 < 1$.

Proof. We provide a proof in order to fill in some additional detail for being more rigorous. Without loss of generality, we may scale $\boldsymbol{z}$ or $\boldsymbol{u}$ such that $\boldsymbol{z}^{\mathrm{T}}\boldsymbol{u} = 1$. Then

$$\boldsymbol{z}^{\mathrm{T}}(A_0 + A_1 + A_2)\boldsymbol{u} = \boldsymbol{z}^{\mathrm{T}}\boldsymbol{u} = 1.$$

The derivative of $\chi(t)$ at $t = 1$ from the left is [30]

$$\chi'(1-) = \boldsymbol{z}^{\mathrm{T}}(A_1 + 2A_2)\boldsymbol{u} = \boldsymbol{z}^{\mathrm{T}}(A_1 + 2A_2)\boldsymbol{u} - [\boldsymbol{z}^{\mathrm{T}}(A_0 + A_1 + A_2)\boldsymbol{u} - 1] = 1 + \mu.$$

The rest of the proof is essentially the same as the proof of [16, Lemma 1.3.4] on p. 17 but with additional detail. In fact, the equation (4.8) is, upon substitution $t = e^{-s}$, equivalent to $s = -\ln \chi(e^{-s})$ for $s \geq 0$. It has a root $s = 0$. We know that $-\ln \chi(e^{-s})$ is concave (i.e., convex down) for $s > 0$ and its right derivative at $s = 0$ is $1 + \mu$.

If $\mu < 0$, then as s increases from $s = 0$, the graph of $-\ln \chi(e^{-s})$ moves below the bisectrix line s in the first quadrant. Thus $s = 0$ is the only solution to $s = -\ln \chi(e^{-s})$ for $s \geq 0$.

If $\mu = 0$, then the graphs of $-\ln \chi(e^{-s})$ and the bisectrix line s are tangent to each other at $s = 0$. We claim that the two graphs will break away from each other as soon as s becomes positive, i.e., the only intersection is at $s = 0$. Otherwise, suppose $s_0 = -\ln \chi(e^{-s_0})$ for some $0 < s_0 < \infty$. Then $s = -\ln \chi(e^{-s})$ for all $0 \leq s \leq s_0$ because $-\ln \chi(e^{-s})$ is concave, or equivalently, $t = \chi(t)$ for $t_0 := e^{-s_0} \leq t \leq 1$. Note $0 < t_0 = e^{-s_0} < 1$. This implies $\phi(t) = 0$ for all $t \in [t_0, 1]$ and thus $[t_0, 1] \subset \mathrm{eig}(\mathscr{A}, \mathscr{B})$, a contradiction because $\mathrm{eig}(\mathscr{A}, \mathscr{B})$ contains at most $2n$ distinct points on the extended complex plane (the complex plane with infinities).

If $\mu > 0$, then as s increases from 0, $-\ln \chi(e^{-s})$ increases and moves above the bisectrix line s for sufficiently tiny s. But as $s \to \infty$, $-\ln \chi(e^{-s})$ increases to its horizontal asymptote $-\ln \chi(0) = -\ln \rho(A_0)$ and thus the graph of $-\ln \chi(e^{-s})$ will intersect with the bisectrix line s at another point s_0 $(0 < s_0 < \infty)$, giving the second root $t_0 = e^{-s_0}$ $(0 < t_0 < 1)$. □

Remark 4.3. In [16, Lemma 1.3.4], $\rho(A_0) = 0$ is allowed in the case of $\mu > 0$, but then a condition that the derivative of $-\ln \chi(e^{-s})$ is less than 1 at some point $s = s_1 > 0$ is added. For simplicity, we go with $\rho(A_0) > 0$.

The next lemma is a corollary of Lemma 1.3.5 of [16, p. 18] which dealt with a more general Markov chain.

Lemma 4.4 *([16, Lemma 1.3.5]). Suppose* (1.3b). *If* $\rho(\Phi) > 0$, *then* $\rho(\Phi)$ *which is no bigger than* 1 *is the smallest positive root of the equation* $t = \chi(t)$.

Proof. We provide a proof for being self-contained. Let $0 < t_0 \leq 1$ be the smallest positive root. Then $\widehat{A} := A_0 + A_1 t_0 + A_2 t_0^2 \geq 0$ and it is irreducible. Let $\hat{\boldsymbol{u}} > 0$ be such that $\widehat{A}\hat{\boldsymbol{u}} = \rho(\widehat{A})\hat{\boldsymbol{u}} = t_0\hat{\boldsymbol{u}}$. Now instead of the iteration scheme (3.3), we use the following one [16]

$$Z_0 = 0 \quad \text{and} \quad Z_{k+1} = A_0 + A_1 Z_k + A_2 Z_k^2 \text{ for } k \geq 0.$$

We claim that $Z_k \leq \Phi$ and $Z_k\hat{\boldsymbol{u}} \leq t_0\hat{\boldsymbol{u}}$ for all k. This can be proved by induction. Evidently, $Z_0 = 0 \leq \Phi$ and $Z_0\hat{\boldsymbol{u}} = 0 \leq t_0\hat{\boldsymbol{u}}$. Suppose that $Z_k \leq \Phi$ and $Z_k\hat{\boldsymbol{u}} \leq t_0\hat{\boldsymbol{u}}$. We have $Z_k^2 \leq \Phi^2$, $Z_k^2\hat{\boldsymbol{u}} \leq t_0^2\hat{\boldsymbol{u}}$, and thus

$$\begin{aligned} Z_{k+1} &= A_0 + A_1 Z_k + A_2 Z_k^2 \leq A_0 + A_1\Phi + A_2\Phi^2 = \Phi, \\ Z_{k+1}\hat{\boldsymbol{u}} &\leq (A_0 + A_1 t_0 + A_2 t_0^2)\hat{\boldsymbol{u}} = \widehat{A}\hat{\boldsymbol{u}} = t_0\hat{\boldsymbol{u}}, \end{aligned}$$

completing the induction proof. The sequence $\{Z_k\}_{k=0}^{\infty}$ is monotonically increasing and thus convergent. Let the limit be Φ'. By letting k go to ∞, we find that Φ' is also a nonnegative solution to (1.1) and at the same time $\Phi' \leq \Phi$. Hence $\Phi' = \Phi$ because Φ is the minimal nonnegative solution. On the other hand, $\Phi\hat{\boldsymbol{u}} = \Phi'\hat{\boldsymbol{u}} \leq t_0\hat{\boldsymbol{u}}$, yielding $0 < \rho(\Phi) \leq t_0$. Since $\rho(\Phi)$ is also a root of the equation $t = \chi(t)$, it must hold that $\rho(\Phi) = t_0$, as was to be shown. □

Theorem 4.5. *Suppose* (1.3b). *The following statements hold.*

(a) *If* $\mu > 0$ *and if* $\rho(A_0) > 0$, *then* $\rho(\Phi) < 1$ *and* $\rho(\Psi) = 1$. *Moreover,* $\phi(\lambda)$ *has* n *zeros in the open unit disk, one simple zero equal to* 1, *and the other* $n - 1$ *zeros on or outside the unit circle (*1 *is not one of them).*

(b) *If* $\mu < 0$ *and if* $\rho(A_2) > 0$, *then* $\rho(\Phi) = 1$ *and* $\rho(\Psi) < 1$. *Moreover,* $\phi(\lambda)$ *has* n *zeros outside the unit disk, one simple zero equal to* 1, *and the other* $n - 1$ *zeros on or inside the unit circle (*1 *is not one of them).*

In both cases, $I - \Phi\Psi$ *and* $I - \Psi\Phi$ *are nonsingular M-matrices.*

Proof. We first prove item (a). Since $\Phi \geq A_0$, we have $\rho(\Phi) \geq \rho(A_0) > 0$ by Theorem 2.1(c). Now use Lemmas 4.2(b) and 4.4 to conclude $\rho(\Phi) < 1$. We claim that

$$\begin{bmatrix} I & \Psi \\ \Phi & I \end{bmatrix} \quad \text{is nonsingular.} \tag{4.9}$$

This is because the columns of $\begin{bmatrix} I \\ \Phi \end{bmatrix}$ form a basis of the eigenspace of $\mathscr{A} - \lambda\mathscr{B}$ associated with its eigenvalues in $\mathrm{eig}(\Phi)$, all inside the unit circle since $\rho(\Phi) < 1$, while those of $\begin{bmatrix} \Psi \\ I \end{bmatrix}$ form a basis of its eigenspace associated with its eigenvalues in $\{1/\lambda : \lambda \in \mathrm{eig}(\Psi)\}$, all on or outside the unit circle since $\rho(\Psi) \le 1$. Consequently, their column vectors together are linearly independent, implying (4.9). The Schur complement of I at the top-left corner of the matrix in (4.9) is $I - \Phi\Psi$ which has to be nonsingular; so is $I - \Psi\Phi$. Because $\Phi\Psi\boldsymbol{u} \le \Phi\boldsymbol{u} \le \boldsymbol{u}$ and $\Psi\Phi\boldsymbol{u} \le \boldsymbol{u}$ by Theorem 3.4(a), both $I - \Phi\Psi$ and $I - \Psi\Phi$ are also M-matrices and thus they are nonsingular M-matrices by definition.

On the other hand, by (4.3) and (4.9), we conclude that

$$1 \in \mathrm{eig}(\mathscr{A}, \mathscr{B}) = \mathrm{eig}(\Phi) \cup \{1/\lambda : \lambda \in \mathrm{eig}(\Psi)\} \quad \Rightarrow \quad 1 \in \{1/\lambda : \lambda \in \mathrm{eig}(\Psi)\}.$$

Hence $1 \in \mathrm{eig}(\Psi)$. We already know $\rho(\Psi) \le 1$ and thus $\rho(\Psi) = 1$.

It remains to show that 1 is a simple zero of $\phi(\lambda)$, or equivalently, a simple eigenvalue of $\mathscr{A} - \lambda\mathscr{B}$. To this end, we will prove $\dim\ker(\mathscr{A} - \mathscr{B}) = 1$, and that there is no Jordan block of size bigger than 1 for eigenvalue 1. The kernel $\ker(\mathscr{A} - \mathscr{B})$ consists of all vectors $\begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{bmatrix} \in \mathbb{R}^{2n}$ such that

$$\begin{bmatrix} -I & I \\ A_0 & A_1 - I + A_2 \end{bmatrix} \begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{bmatrix} = 0,$$

where $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$. Equivalently, $\boldsymbol{x} = \boldsymbol{y}$ and

$$(A_0 + A_1 + A_2)\boldsymbol{x} = \boldsymbol{x}.$$

Since $A_0 + A_1 + A_2$ is nonnegative and irreducible, we have $\boldsymbol{x} = \alpha\boldsymbol{u}$ for some scalar α. Therefore $\dim\ker(\mathscr{A} - \mathscr{B}) = 1$. Suppose, to the contrary, that there is a Jordan block of size bigger than 1 for eigenvalue 1. Then the second vector $\begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{bmatrix} \in \mathbb{R}^{2n}$ in its Jordan chain must satisfy the following equation

$$\mathscr{A}\begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{bmatrix} = \mathscr{B}\begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{bmatrix} + \mathscr{B}\begin{bmatrix} \boldsymbol{u} \\ \boldsymbol{u} \end{bmatrix}. \tag{4.10}$$

Equivalently, (4.10) gives

$$\boldsymbol{y} = \boldsymbol{x} + \boldsymbol{u}, \quad A_0\boldsymbol{x} + (A_1 - I + A_2)\boldsymbol{y} = -A_2\boldsymbol{u}.$$

Substituting the first equation into the second equation and noticing (4.4a), we obtain

$$\underbrace{(A_0 + A_1 + A_2 - I)}_{=:C}\boldsymbol{x} = (A_0 - A_2)\boldsymbol{u}. \tag{4.11}$$

We claim that this equation is not solvable, i.e., it has no solution. To see this, we need to prove $(A_0 - A_2)\boldsymbol{u} \notin \mathcal{R}(C) = \mathcal{N}(C^{\mathrm{T}})^{\perp}$. Recalling (4.4b), we have $C^{\mathrm{T}}\boldsymbol{z} = 0$, i.e., $\boldsymbol{z} \in \mathcal{N}(C^{\mathrm{T}})$. Since $\boldsymbol{z}^{\mathrm{T}}(A_0 - A_2)\boldsymbol{u} = -\mu < 0$, we conclude that $(A_0 - A_2)\boldsymbol{u} \notin \mathcal{N}(C^{\mathrm{T}})^{\perp}$, as expected.

By switching the roles of A_0 and A_2, we find that item (b) is a corollary of item (a). □

In general, the case $\mu = 0$ is more difficult to deal with, and the next theorem provides a partial description on the distribution of $\mathrm{eig}(\mathscr{A}, \mathscr{B})$. More comments come after the theorem.

Theorem 4.6. *Suppose* (1.3b) *and* $\mu = 0$. *Then* 1 *is a zero of* $\phi(\lambda)$, *and it is an eigenvalue of* $\mathscr{A} - \lambda\mathscr{B}$ *with geometric multiplicity* 1 *and algebraic multiplicity at least* 2. *Also* $\mathscr{A} - \lambda\mathscr{B}$ *has*[7] *at least* n *eigenvalues in the closed unit disk and at least* n *eigenvalues on or outside the unit circle. Moreover, the following statements hold.*

(a) *If* $\rho(A_0) > 0$, *then* $\rho(\Phi) = 1$, $1 \in \mathrm{eig}(\Phi)$, *and* $\Phi\boldsymbol{u} = \boldsymbol{u}$;
(b) *If* $\rho(A_2) > 0$, *then* $\rho(\Psi) = 1$, $1 \in \mathrm{eig}(\Psi)$, *and* $\Psi\boldsymbol{u} = \boldsymbol{u}$.

In particular, if both $\rho(A_0) > 0$ *and* $\rho(A_2) > 0$, *then both* $I - \Phi\Psi$ *and* $I - \Psi\Phi$ *are singular and, moreover,* $(I - \Phi\Psi)\boldsymbol{u} = (I - \Psi\Phi)\boldsymbol{u} = 0$.

Proof. Previously, we already commented that $\phi(1) = 0$ and $1 \in \mathrm{eig}(\mathscr{A}, \mathscr{B})$. Our argument in the proof of Theorem 4.5 for proving $\dim\ker(\mathscr{A} - \mathscr{B}) = 1$ remains valid here, i.e., the geometric multiplicity of the eigenvalue 1 is 1. In order to claim its algebraic multiplicity is at least 2, it suffices to show that the associated Jordan chain of vectors has length at least 2. It follows from the proof of Theorem 4.5 that we can take the eigenvector associated with the eigenvalue 1 to be $\begin{bmatrix}\boldsymbol{u}\\ \boldsymbol{u}\end{bmatrix}$. The second vector $\begin{bmatrix}\boldsymbol{x}\\ \boldsymbol{y}\end{bmatrix} \in \mathbb{R}^{2n}$ in its Jordan chain must satisfy (4.10), yielding $\boldsymbol{y} = \boldsymbol{x} + \boldsymbol{u}$ with $\boldsymbol{x}$ determined by (4.11). We claim that (4.11) now is solvable, i.e., it has a solution. To see this, we need to prove that $(A_0 - A_2)\boldsymbol{u}$ is in $\mathcal{R}(C)$. Recalling (4.4b), we have $C^{\mathrm{T}}\boldsymbol{z} = 0$, i.e., $\boldsymbol{z} \in \mathcal{N}(C^{\mathrm{T}})$. It can be argued that $\dim\mathcal{N}(C^{\mathrm{T}}) = 1$ since $-C$ is irreducible, and so $\mathcal{N}(C^{\mathrm{T}}) = \mathcal{R}(\boldsymbol{z})$. On the other hand, it follows from $\boldsymbol{z}^{\mathrm{T}}(A_0 - A_2)\boldsymbol{u} = 0$ that $(A_0 - A_2)\boldsymbol{u} \perp \boldsymbol{z}$. Therefore

$$(A_0 - A_2)\boldsymbol{u} \in \mathcal{R}(\boldsymbol{z})^{\perp} = \mathcal{N}(C^{\mathrm{T}})^{\perp} = \mathcal{R}(C).$$

Thus (4.11) has a solution $\boldsymbol{x}$. Finally, we have

$$\mathscr{A}\begin{bmatrix}\boldsymbol{u} & \boldsymbol{x}\\ \boldsymbol{u} & \boldsymbol{x}+\boldsymbol{u}\end{bmatrix} = \mathscr{B}\begin{bmatrix}\boldsymbol{u} & \boldsymbol{x}\\ \boldsymbol{u} & \boldsymbol{x}+\boldsymbol{u}\end{bmatrix}\begin{bmatrix}1 & 1\\ 0 & 1\end{bmatrix}.$$

[7] This is not the same as saying it has n eigenvalues in the closed unit disk and the other n eigenvalues on or outside the unit circle, as we would like to be able to show.

The vectors $\begin{bmatrix} \boldsymbol{u} \\ \boldsymbol{u} \end{bmatrix}$ and $\begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{x}+\boldsymbol{u} \end{bmatrix}$ are the first two vectors in a Jordan chain of $\mathscr{A}-\lambda\mathscr{B}$ associated with the eigenvalue 1. That $\mathscr{A}-\lambda\mathscr{B}$ has at least n eigenvalues in the closed unit disk and at least n eigenvalues on or outside the unit circle are consequences of the equations:

$$\mathscr{A}\begin{bmatrix} I \\ \Phi \end{bmatrix} = \mathscr{B}\begin{bmatrix} I \\ \Phi \end{bmatrix}\Phi \text{ and } \rho(\Phi)\le 1, \quad \mathscr{A}\begin{bmatrix} \Psi \\ I \end{bmatrix}\Psi = \mathscr{B}\begin{bmatrix} \Psi \\ I \end{bmatrix} \text{ and } \rho(\Psi)\le 1.$$

For item (a), we note $\rho(\Phi)\ge\rho(A_0)>0$. By Lemmas 4.2 and 4.4, we conclude $\rho(\Phi)=1$ and thus $1\in\mathrm{eig}(\Phi)$ because $\Phi\ge 0$. Let $\hat{\boldsymbol{u}}\ge 0$ be an eigenvector of Φ corresponding to its eigenvalue 1, i.e., $\Phi\hat{\boldsymbol{u}}=\hat{\boldsymbol{u}}$ and $\hat{\boldsymbol{u}}\ne 0$. Post-multiply $\Phi=A_0+A_1\Phi+A_2\Phi^2$ by $\hat{\boldsymbol{u}}$ to get

$$(A_0+A_1+A_2)\hat{\boldsymbol{u}}=\hat{\boldsymbol{u}},$$

i.e., $\hat{\boldsymbol{u}}$ is a nonnegative eigenvector of $A_0+A_1+A_2$, which is assumed irreducible, corresponding to its top eigenvalue 1. Therefore $\hat{\boldsymbol{u}}=\alpha\boldsymbol{u}$ for some $\alpha>0$, implying $\Phi\boldsymbol{u}=\boldsymbol{u}$, as expected.

Item (b) is a corollary of item (a) upon switching the roles of A_0 and A_2.

Finally if both $\rho(A_0)>0$ and $\rho(A_2)>0$, then $\Phi\boldsymbol{u}=\boldsymbol{u}$ and $\Psi\boldsymbol{u}=\boldsymbol{u}$, yielding $(I-\Phi\Psi)\boldsymbol{u}=(I-\Psi\Phi)\boldsymbol{u}=0$. □

A couple of comments are in order. When $\mu\ne 0$, we have

$$\mathrm{eig}(\mathscr{A},\mathscr{B})=\mathrm{eig}(\Phi)\cup\{1/\lambda:\lambda\in\mathrm{eig}(\Psi)\} \text{ and } \mathrm{eig}(\Phi)\cap\{1/\lambda:\lambda\in\mathrm{eig}(\Psi)\}=\emptyset, \tag{4.12}$$

as guaranteed by Theorem 4.5, where the union is in the sense of the multiset union which allows same value appears two or more times, representing different eigenvalues having the same value. Theorem 4.6 provides a partial description on the distribution of $\mathrm{eig}(\mathscr{A},\mathscr{B})$ for the case $\mu=0$. In particular, the second relation in (4.12) is no longer true because 1 is an eigenvalue of multiplicity at least 2 and it belongs to both $\mathrm{eig}(\Phi)$ and $\mathrm{eig}(\Psi)$. However, it is not clear whether the first equation in (4.12) remains valid or not.

Previously, we mentioned that in [25, p. 128] a complete description on the zeros of $\phi(\lambda)$ relative to the unit disk is given even for the case $\mu=0$ but with conditions[8] that would take a couple of pages to explain. Using the description, one would be able to conclude that, under the conditions, $\phi(\lambda)$ has $n-1$ zeros in the open unit disk, one zero of multiplicity two equal to 1, and $n-1$ zeros outside the closed unit disk. In general, however, without assuming the conditions, $\phi(\lambda)$ may have two or more different

[8] They are **Conditions** 5.1 and 5.2 on [25, pp. 110–111] we previously mentioned.

eigenvalues on the unit circle! The following example is provided by an anonymous referee.

$$A_0 = A_2 = \frac{1}{4}\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix},\ A_1 = \frac{1}{2}I_2,\ \text{and } \Phi = \Psi = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \tag{4.13}$$

for which it can be calculated that

$$\phi(\lambda) = -\frac{1}{4}(\lambda+1)^2(\lambda-1)^2, \quad \text{eig}(\Phi) = \text{eig}(\Psi) = \{\pm 1\},$$

and $\mathscr{A} - \lambda\mathscr{B}$ has two 2×2 Jordan blocks associated with its multiple eigenvalues 1 and -1:

$$\mathscr{A}\begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & 2 & -1 & 0 \\ 1 & 2 & 1 & 0 \end{bmatrix} = \mathscr{B}\begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & 2 & -1 & 0 \\ 1 & 2 & 1 & 0 \end{bmatrix}\begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 1 \\ 0 & 0 & 0 & -1 \end{bmatrix}.$$

Lemma 4.7. *We have* $\rho(A_1 + A_2\Phi + A_2) < 1$ *under* (1.3a)*, and* $\rho(A_1 + A_2\Phi + A_2) \le 1$ *under* (1.3b)*,* $\mu \neq 0$*,* $\rho(A_0) > 0$ *and* $\rho(A_2) > 0$*.*

Proof. Consider the case (1.3a). We have

$$(A_1 + A_2\Phi)\boldsymbol{u} \le (A_1 + A_2)\boldsymbol{u} \le (A_0 + A_1 + A_2)\boldsymbol{u} = \boldsymbol{u} - \boldsymbol{v} < \boldsymbol{u},$$

since $\boldsymbol{v} > 0$ and thus $\rho(A_1 + A_2\Phi) < 1$. Write $K = I - (A_1 + A_2\Phi)$, which is a nonsingular M-matrix. It can be verified that

$$\begin{aligned} A_0 + \lambda(A_1 - I) + \lambda^2 A_2 &= (\lambda A_2 + A_2\Phi + A_1 - I)(\lambda I - \Phi) \\ &= -(K - \lambda A_2)(\lambda I - \Phi) \\ &= -(I - \lambda A_2 K^{-1})K(\lambda I - \Phi). \end{aligned} \tag{4.14}$$

By Theorem 4.1, we conclude that $\text{eig}(\Phi)$ consists of exactly the n eigenvalues of $\mathscr{A} - \lambda\mathscr{B}$ in the open unit disk, while reciprocals of those in $\text{eig}(A_2K^{-1})$ give exactly the n eigenvalues of $\mathscr{A} - \lambda\mathscr{B}$ outside of the closed unit disk. In particular, $\rho(A_2K^{-1}) < 1$ and $I - A_2K^{-1}$ is a nonsingular M-matrix. On the other hand,

$$I - A_2K^{-1} = (K - A_2)K^{-1} = [I - (A_1 + A_2\Phi + A_2)]K^{-1},$$

yielding $[I-(A_1+A_2\Phi+A_2)]^{-1} = K^{-1}(I-A_2K^{-1})^{-1} \ge 0$. Evidently, $I-(A_1+A_2\Phi+A_2)$ is a Z-matrix. By Theorem 2.2, it is a nonsingular M-matrix and thus $\rho(A_1+A_2\Phi+A_2) < 1$.

Now turn to the case (1.3b). Define $A_0(t) = tA_0$ and $A_2(t) = tA_2$ for $t \le 1$. Consider the matrix equation

$$A_0(t) + A_1X + A_2(t)X^2 = X, \tag{4.15}$$

taking the same form as (1.1) but falling into the case (1.3a) when $t < 1$ because

$$0 \le (A_0(t) + A_1 + A_2(t))\boldsymbol{u} = \boldsymbol{u} + (t-1)(A_0 + A_2)\boldsymbol{u} \le \boldsymbol{u}$$

but the equality sign does not hold. Note that $A_0(t) + A_1 + A_2(t)$ is irreducible. Hence, by Theorem 2.1(d), $\rho(A_0(t)+A_1+A_2(t)) < 1$ for $t < 1$. The equation (4.15) has a unique nonnegative minimal solution, denoted by $\Phi(t)$ and satisfying

$$\rho(A_1 + A_2(t)\Phi(t) + A_2(t)) < 1. \tag{4.16}$$

Now we if we can prove that $\Phi(t)$ goes to $\Phi(1) = \Phi$ as $t \to 1^-$, then letting $t \to 1^-$ will complete the proof.

Similarly to what we had before, corresponding to (4.15), we have a matrix pencil $\mathscr{A}(t) - \lambda\mathscr{B}(t)$ defined in the same as in (1.4) but with $A_0(t)$, A_1, and $A_2(t)$. Let $\Phi(t)$ be the minimal nonnegative solution to (4.15), and $\Psi(t)$ be that to the dual equation of (4.15). We will also have as in (4.12), for $t < 1$,

$$\mathrm{eig}(\mathscr{A}(t), \mathscr{B}(t)) = \mathrm{eig}(\Phi(t)) \cup \mathbb{E}(t) \text{ and } \mathrm{eig}(\Phi(t)) \cap \mathbb{E}(t) = \emptyset,$$

where $\mathbb{E}(t) = \{1/\lambda : \lambda \in \mathrm{eig}(\Psi(t))\}$, and moreover, $\mathrm{eig}(\Phi(t))$ lies inside the unit circle while $\mathbb{E}(t)$ lies outside the unit circle. Since the eigenvalues of a regular matrix pencil are continuous functions of the matrix entries [31,17], in consideration of Theorem 4.5 and (4.12), we find that as $t \to 1^-$, the elements of $\mathrm{eig}(\Phi(t))$ goes to those of $\mathrm{eig}(\Phi)$ and the elements of $\mathbb{E}(t)$ goes to those of $\mathbb{E} := \{1/\lambda : \lambda \in \mathrm{eig}(\Psi)\}$. On the other hand,

$$\begin{bmatrix} I \\ \Phi(t) \end{bmatrix}, \quad \begin{bmatrix} I \\ \Phi \end{bmatrix}$$

are the basis matrices of the unique eigenspace of $\mathscr{A}(t) - \lambda\mathscr{B}(t)$ associated with its n eigenvalues in $\mathrm{eig}(\Phi(t))$ and that of $\mathscr{A} - \lambda\mathscr{B}$ associated with its n eigenvalues in $\mathrm{eig}(\Phi)$, respectively. Because the eigenspace associated with a cluster of eigenvalues is continuous in the metric of the canonical angles [32, Theorem 5.7], $\Phi(t)$ goes to Φ as $t \to 1^-$, as was to be shown. □

Define the linear operator

$$\mathscr{L}_\Phi : X \to X - (A_1 + A_2\Phi)X - A_2X\Phi \tag{4.17a}$$

whose matrix presentation is given by

$$P = I_{2n} - I_n \otimes (A_1 + A_2\Phi) - \Phi^{\mathrm{T}} \otimes A_2. \tag{4.17b}$$

Theorem 4.8. *P, which is defined as in (4.17b), is a nonsingular M-matrix under (1.3a) or under (1.3b) with $\mu \neq 0$, $\rho(A_0) > 0$ and $\rho(A_2) > 0$.*

Proof. By (4.2) and (4.14), the eigenvalues of $\mathscr{A} - \lambda\mathscr{B}$ is the multiset union of $\mathrm{eig}(\Phi)$ and $\mathrm{eig}(A_2\Phi + A_1 - I, -A_2)$. The two sets of eigenvalues have no intersection by Theorems 4.1 and 4.5. Therefore $\mathscr{L}_\Phi$ is an invertible linear operator. Using the Schur decompositions [17, p. 276] of Φ^T and $(A_1 + A_2\Phi) - \lambda A_2$, we find that

$$\mathrm{eig}(I_n \otimes (A_1 + A_2\Phi) + \Phi^{\mathrm{T}} \otimes A_2) = \bigcup_{\lambda \in \mathrm{eig}(\Phi)} \mathrm{eig}(A_1 + A_2\Phi + \lambda A_2),$$

which was also used in the proof of [8, Theorem 5.1]. Since $|\lambda| \leq 1$ for $\lambda \in \mathrm{eig}(\Phi)$, we also have $\rho(A_1 + A_2\Phi + \lambda A_2) \leq \rho(A_1 + A_2\Phi + A_2) \leq 1$, where the last inequality is a consequence of Lemma 4.7. Therefore

$$\rho(I_n \otimes (A_1 + A_2\Phi) + \Phi^{\mathrm{T}} \otimes A_2) \leq 1. \tag{4.18}$$

Combining this with the fact that P is a nonsingular Z-matrix, we conclude that (4.18) must be a strict inequality, i.e., P is a nonsingular M-matrix under the conditions of the theorem. □

5. Structure-preserving doubling algorithm

Geometrically, (1.6) says that the column space $\mathcal{R}(\begin{bmatrix} I \\ X \end{bmatrix})$ is an eigenspace of the matrix pencil $\mathscr{A}_0 - \lambda\mathscr{B}_0$ associated with its eigenvalues that are given by $\mathrm{eig}(X)$. The same statement can be made for (1.4). The basic idea of the structure-preserving doubling algorithm for (SF1) [33,14] for solving (1.6) is to recursively construct a sequence of matrix pencils $\mathscr{A}_k - \lambda\mathscr{B}_k$ for $k \geq 1$ that have the same block structure (thus structure-preserving) as $\mathscr{A}_0 - \lambda\mathscr{B}_0$:

$$\mathscr{A}_k = \begin{matrix} n \\ n \end{matrix} \overset{\begin{matrix} n & n \end{matrix}}{\begin{bmatrix} E_k & 0 \\ -X_k & I \end{bmatrix}}, \quad \mathscr{B}_k = \begin{matrix} n \\ n \end{matrix} \overset{\begin{matrix} n & n \end{matrix}}{\begin{bmatrix} I & -Y_k \\ 0 & F_k \end{bmatrix}} \quad \text{for } k = 1, 2, \ldots \tag{5.1}$$

and at the same time

$$\mathscr{A}_k \begin{bmatrix} I \\ X \end{bmatrix} = \mathscr{B}_k \begin{bmatrix} I \\ X \end{bmatrix} M^{2^k} \quad \text{for } k = 0, 1, \ldots,$$

where $M = X$ because of (1.4). Now if also $\rho(M) < 1$ (as we see later, $\rho(M) = 1$ is allowed, too), then we will have

$$\mathscr{A}_k \begin{bmatrix} I \\ X \end{bmatrix} \to 0 \quad \text{as } k \to \infty,$$

provided that $\{\|\mathscr{B}_k\|\}_{k=0}^{\infty}$ is bounded, and as a consequence $X_k \to X$, a solution of (1.6) and thus of (1.4), too. We outline the doubling algorithm as Algorithm 5.1. The interested reader is referred to [14] for how it is derived.

Algorithm 5.1 Doubling algorithm for (SF1) [14].

Input: X_0, Y_0, E_0, $F_0 \in \mathbb{R}^{n\times n}$.
Output: X_∞ as the limit of X_k if it converges.

1: **for** $k = 0, 1, \ldots,$ until convergence **do**
2: compute E_{k+1}, F_{k+1}, X_{k+1}, Y_{k+1} according to

$$E_{k+1} = E_k(I_n - Y_kX_k)^{-1}E_k, \tag{5.2a}$$

$$F_{k+1} = F_k(I_n - X_kY_k)^{-1}F_k, \tag{5.2b}$$

$$X_{k+1} = X_k + F_k(I_n - X_kY_k)^{-1}X_kE_k, \tag{5.2c}$$

$$Y_{k+1} = Y_k + E_k(I_n - Y_kX_k)^{-1}Y_kF_k. \tag{5.2d}$$

3: **end for**
4: **return** X_k at convergence as the computed solution.

Moments ago, we mentioned that X_k is intended to approach a solution to (1.6) (and thus (1.1)). As a by-product, the matrix Y_k approaches to something interesting, too. That is a solution to (3.8) (and thus (3.5)). Indeed, we also have

$$\mathscr{A}_k \begin{bmatrix} Y \\ I \end{bmatrix} N^{2^k} = \mathscr{B}_k \begin{bmatrix} Y \\ I \end{bmatrix} \quad \text{for } k = 0, 1, \ldots,$$

where $N = Y$ because of (3.7). Now if also $\rho(N) < 1$ (as we see later, $\rho(N) = 1$ is allowed, too), then we will have

$$\mathscr{B}_k \begin{bmatrix} Y \\ I \end{bmatrix} \to 0 \quad \text{as } k \to \infty,$$

provided that $\{\|\mathscr{A}_k\|\}_{k=0}^{\infty}$ is bounded, and as a consequence $Y_k \to Y$, a solution to (3.8).

The next section is devoted to the convergence analysis of Algorithm 5.1 for solving the QBD equation (1.1) under the assumption (1.3).

6. Convergence analysis

We start by noting that under (1.3), $(I - A_1)^{-1} \geq 0$ as guaranteed by Lemma 3.1. Immediately, we conclude that X_0, Y_0, E_0, and F_0 are well-defined as in (1.5) and they are all nonnegative. As we will soon see, a major inequality that governs the speed of convergence of X_k and Y_k to their respective targets Φ and Ψ is

$$\limsup_{k\to\infty} \|\Phi - X_k\|^{1/2^k}, \limsup_{k\to\infty} \|\Psi - Y_k\|^{1/2^k} \leq \rho(\Phi)\rho(\Psi), \tag{6.1}$$

where $\|\cdot\|$ is any matrix norm. This inequality is only useful when $\rho(\Phi)\rho(\Psi) < 1$ which indicates that the convergence is at least quadratic; otherwise it does not guarantee convergence, not to mention revealing anything about the convergence speed.

Previously in section 4, we show that

$\rho(\Phi)\rho(\Psi) < 1$ if either (1.3a) holds or $\{\mu \neq 0,\ \rho(A_0) > 0$ and $\rho(A_2) > 0\}$ in the case of (1.3b), where μ is defined by (4.5), but $\rho(\Phi)\rho(\Psi) = 1$ if $\mu = 0,\ \rho(A_0) > 0$ and $\rho(A_2) > 0$ in the case of (1.3b).

This, combined with (6.1), suggest that X_k and Y_k generated by Algorithm 5.1 converge quadratically to Φ and Ψ, respectively, under (1.3a) or $\{\mu \neq 0,\ \rho(A_0) > 0$ and $\rho(A_2) > 0\}$ in the case of (1.3b), provided all E_k, F_k, X_k, Y_k are well-defined. It turns out that for the case $\mu = 0$, convergence to the minimal nonnegative solution still happens but is linear at the rate of 1/2 under a condition on the Jordan eigen-structure (see (6.11) below).

We observe that as long as E_k, F_k, X_k, Y_k are well-defined (so are $\mathscr{A}_k$ and $\mathscr{B}_k$), we will have

$$\mathscr{A}_k \begin{bmatrix} I \\ \Phi \end{bmatrix} = \mathscr{B}_k \begin{bmatrix} I \\ \Phi \end{bmatrix} \Phi^{2^k}, \quad \mathscr{A}_k \begin{bmatrix} \Psi \\ I \end{bmatrix} \Psi^{2^k} = \mathscr{B}_k \begin{bmatrix} \Psi \\ I \end{bmatrix},$$

where $\mathscr{A}_k$ and $\mathscr{B}_k$ are defined as in (5.1). Or, equivalently,

$$\Phi - X_k = F_k \Phi^{2^k+1}, \quad E_k = (I - Y_k\Phi)\Phi^{2^k}, \tag{6.2a}$$

$$\Psi - Y_k = E_k \Psi^{2^k+1}, \quad F_k = (I - X_k\Psi)\Psi^{2^k}. \tag{6.2b}$$

The next theorem is essentially [33, Theorem 4.1] which was proved in the case of the M-matrix algebraic Riccati equation (MARE). The only difference lies in how initially (E_0, F_0, X_0, Y_0) is defined for MARE there and the QBD equation here. We still present a proof here because some part of the proof is needed in the proof of Theorem 6.2.

Theorem 6.1. *Under* (1.3a)*, the sequence* $\{(E_k, F_k, X_k, Y_k)\}_{k=0}^{\infty}$ *in* Algorithm 5.1 *is well-defined and, moreover, for* $k \geq 0$*,*

(a) $E_k = (I - Y_k\Phi)\Phi^{2^k} \geq 0$,
(b) $F_k = (I - X_k\Psi)\Psi^{2^k} \geq 0$,
(c) $I - X_kY_k$ *and* $I - Y_kX_k$ *are nonsingular* M*-matrices,*
(d) $0 \leq X_k \leq X_{k+1} \leq \Phi,\ 0 \leq Y_k \leq Y_{k+1} \leq \Psi$*, and*

$$0 \leq \Phi - X_k \leq \Psi^{2^k}\Phi\Phi^{2^k},\ 0 \leq \Psi - Y_k \leq \Phi^{2^k}\Psi\Psi^{2^k}. \tag{6.3}$$

As a consequence, the inequality (6.1) *holds.*

Proof. We start by proving item (c), which implies that $\{(E_k, F_k, X_k, Y_k)\}_{k=0}^{\infty}$ is well-defined, and for all $k \geq 0$

$$E_k \geq 0,\ F_k \geq 0,\ 0 \leq X_k \leq \Phi,\ 0 \leq Y_k \leq \Psi, \tag{6.4}$$

by mathematical induction. By (1.5) and Theorem 3.4, we see that (6.4) holds for $k = 0$. Therefore,

$$I - X_0 Y_0 \geq I - \Phi\Psi, \quad I - Y_0 X_0 \geq I - \Psi\Phi.$$

By Theorem 3.4(d), both $I - \Phi\Psi$ and $I - \Psi\Phi$ are nonsingular M-matrices; so are $I - X_0 Y_0$ and $I - Y_0 X_0$ according to Theorem 2.3(b). This completes the proof of (6.4) and item (c) for $k = 0$. Suppose that they hold for $k = \ell$. Hence $E_{\ell+1}$, $X_{\ell+1}$, $F_{\ell+1}$, $Y_{\ell+1}$ are well-defined by (5.2), which, together with the induction hypothesis, guarantee that

$$E_{\ell+1} \geq 0,\ F_{\ell+1} \geq 0,\ 0 \leq X_\ell \leq X_{\ell+1},\ 0 \leq Y_\ell \leq Y_{\ell+1}. \tag{6.5}$$

On the other hand, (6.2) for $k = \ell + 1$ says

$$\Phi - X_{\ell+1} = F_{\ell+1}\Phi^{2^{\ell+1}+1} \geq 0,\ \Psi - Y_{\ell+1} = E_{\ell+1}\Psi^{2^{\ell+1}+1} \geq 0.$$

So we have (6.4) for $k = \ell + 1$, and thus

$$I - X_{\ell+1}Y_{\ell+1} \geq I - \Phi\Psi, \quad I - Y_{\ell+1}X_{\ell+1} \geq I - \Psi\Phi.$$

By the same reasoning above, we conclude that $I - X_{\ell+1}Y_{\ell+1}$ and $I - Y_{\ell+1}X_{\ell+1}$ are nonsingular M-matrices. This is item (c) for $k = \ell + 1$. This completes the proof of item (c) and (6.4).

We deduce from (6.2) that

$$0 \leq \Phi - X_k = (I - X_k\Psi)\Psi^{2^k}\Phi\Phi^{2^k} \leq \Psi^{2^k}\Phi\Phi^{2^k},$$
$$0 \leq \Psi - Y_k = (I - Y_k\Phi)\Phi^{2^k}\Psi\Psi^{2^k} \leq \Phi^{2^k}\Psi\Psi^{2^k}.$$

That gives (6.3). The rest of the claims, except (6.1), are immediate consequences of (6.4), item (c), (6.2), and the recursive formulas in (5.2). It remains to show (6.1) which implies that X_k and Y_k converge quadratically to Φ and Ψ, respectively, as $k \to \infty$. Since all matrix norms on $\mathbb{R}^{n\times n}$ are equivalently, meaning each can be bounded by another modulo a constant factor depending only on n, without loss of generality, we may consider any consistent matrix norm that is monotonic on nonnegative matrices, e.g., the ℓ_1-operator norm. We get

$$\|\Phi - X_k\|^{1/2^k} \le \|\Psi^{2^k}\|^{1/2^k} \|\Phi\|^{1/2^k} \|\Phi^{2^k}\|^{1/2^k},$$
$$\|\Psi - Y_k\|^{1/2^k} \le \|\Phi^{2^k}\|^{1/2^k} \|\Psi\|^{1/2^k} \|\Psi^{2^k}\|^{1/2^k}.$$

Letting $k \to \infty$, we arrive at (6.1). □

Next, we consider the case of (1.3b). Recall $E_0 = (I - A_1)^{-1}A_0$ and $F_0 = (I - A_1)^{-1}A_2$ from (1.5).

Theorem 6.2. *Suppose* (1.3b) *and*[9]

$$\text{either } E_0\boldsymbol{u} > 0 \text{ or } F_0\boldsymbol{u} > 0. \tag{6.6}$$

Then all conclusions of Theorem 6.1 *are valid and, in addition, for* $k \ge 0$,

$$\text{either } E_k\boldsymbol{u} > 0 \text{ if } E_0\boldsymbol{u} > 0, \text{ or } F_k\boldsymbol{u} > 0 \text{ if } F_0\boldsymbol{u} > 0. \tag{6.7}$$

Proof. A proof can be given by simply modifying the induction argument in the proof of Theorem 6.1 to also include (6.7). Note that in (1.5) $E_0 = X_0$ and $F_0 = Y_0$. Hence $E_0\boldsymbol{u} > 0$ implies $\Phi\boldsymbol{u} \ge X_0\boldsymbol{u} = E_0\boldsymbol{u} > 0$, and similarly $F_0\boldsymbol{u} > 0$ implies $\Psi\boldsymbol{u} > 0$. In particular, we have $\Phi^m\boldsymbol{u} > 0$ and $\Psi^m\boldsymbol{u} > 0$ for any integer $m \ge 1$.

For $k = 0$, we have (6.4) and (6.7), but the argument there for claiming that $I - X_0Y_0$ and $I - Y_0X_0$ are nonsingular M-matrices no longer works because now $I - \Phi\Psi$ and $I - \Psi\Phi$ are singular (as stated in Theorem 6.3 below). We will have to do something different. Suppose that $E_0\boldsymbol{u} > 0$. Then it follows from the second equation in (6.2a) that

$$0 < E_0\boldsymbol{u} = (I - Y_0\Phi)\boldsymbol{u}. \tag{6.8}$$

Because $I - Y_0\Phi$ is also a Z-matrix, it is a nonsingular M-matrix by (6.8) and Theorem 2.2(c). Now $I - Y_0X_0$ is also a Z-matrix and $I - Y_0X_0 \ge I - Y_0\Phi$. By Theorem 2.3(b), $I - Y_0X_0$ is a nonsingular M-matrix; so is $I - X_0Y_0$ because $\rho(X_0Y_0) = \rho(Y_0X_0)$. Similarly, we can deal with the case $F_0v > 0$.

Suppose that item (c), (6.4), and (6.7) hold for $k = \ell$. Hence $E_{\ell+1}$, $X_{\ell+1}$, $F_{\ell+1}$, and $Y_{\ell+1}$ are well-defined, and for the same reasoning as in the proof of Theorem 6.1, we have (6.4) for $k = \ell+1$ and (6.5). As to (6.7) for $k = \ell+1$, we observe that $E_\ell\boldsymbol{u} > 0$ is equivalent to that no row of E_ℓ is zero. Thus if $E_\ell\boldsymbol{u} > 0$ then $E_{\ell+1}\boldsymbol{u} = E_\ell(I - Y_\ell X_\ell)^{-1}E_\ell\boldsymbol{u} > 0$, and if $F_\ell\boldsymbol{u} > 0$ then $F_{\ell+1}\boldsymbol{u} = F_\ell(I - X_\ell Y_\ell)^{-1}F_\ell\boldsymbol{u} > 0$. It remains to show item (c) for $k = \ell + 1$. For that purpose, we note that (6.2) for $k = \ell + 1$ says

$$\Phi - X_{\ell+1} = F_{\ell+1}\Phi^{2^{\ell+1}+1} \ge 0, \ \Psi - Y_{\ell+1} = E_{\ell+1}\Psi^{2^{\ell+1}+1} \ge 0. \tag{6.9}$$

[9] The assumption (6.6) is weaker than either $A_0\boldsymbol{u} > 0$ or $A_2\boldsymbol{u} > 0$.

Suppose now $E_\ell \boldsymbol{u} > 0$ and then $E_{\ell+1}\boldsymbol{u} > 0$ as we just argued. It follows from the second equation in (6.2a) that

$$0 < E_{\ell+1}\boldsymbol{u} = (I - Y_{\ell+1}\Phi)\Phi^{2^{\ell+1}}\boldsymbol{u}. \tag{6.10}$$

Because $I - Y_{\ell+1}\Phi$ is a Z-matrix and $\Phi^{2^{\ell+1}}\boldsymbol{u} > 0$, it is a nonsingular M-matrix by (6.10) and Theorem 2.2(c). Noting (6.9) and by the same reasoning above, we conclude that both $I - Y_{\ell+1}X_{\ell+1}$ and $I - X_{\ell+1}Y_{\ell+1}$ are nonsingular M-matrices. Similarly, we can deal with the case $F_\ell \boldsymbol{u} > 0$. □

Theorem 6.2 doesn't guarantee that the convergence of X_k to Φ and/or the convergence of Y_k to Ψ for the case $\mu = 0$ and $\rho(A_0) > 0$ and $\rho(A_2) > 0$ because then $\rho(\Phi) = \rho(\Psi) = 1$ by Theorem 4.6 and thus Ψ^{2^k} and Φ^{2^k} do not go to 0 as $k \to \infty$. It turns out that such a case falls into the so-called *critical case* for (1.6) studied in [14, section 3.8]. In what follows, we will cite the result for the current situation.

Suppose (1.3b), $\mu = 0$, $\rho(A_0) > 0$ and $\rho(A_2) > 0$, where μ is defined as in (4.5). By Theorem 4.6, $\mathscr{A} - \lambda\mathscr{B}$ (and thus $\mathscr{A}_0 - \lambda\mathscr{B}_0$, too) always have a multiple eigenvalue 1 and, as the example in (4.13) shows, it may have other multiple eigenvalues on the unit circle, too. Unfortunately, Theorem 4.6 fails to reveal a complete picture as how the eigenvalues of $\mathscr{A}_0 - \lambda\mathscr{B}_0$ distribute relative to the unit circle. In order to apply the convergence analysis of [14, Theorem 3.26 on p. 42], we will make the following assumption [14, (3.82) on p. 40].

> the partial multiplicities for all eigenvalues of $\mathscr{A}_0 - \lambda\mathscr{B}_0$ on the unit circle are even, i.e., the sizes of all Jordan blocks associated with its eigenvalues on the unit circle are even. (6.11)

Then Φ and Ψ can also be constructed from the Weierstrass canonical forms [34, p. 28] of $\mathscr{A}_0 - \lambda\mathscr{B}_0$ and $\mathscr{B}_0 - \lambda\mathscr{A}_0$, respectively, as in [14, pp. 40–41] because of the equivalent relation between the solution to the quadratic matrix equation (1.1) and the eigenvalue problem (1.6) and the equivalent relation between the solution to the dual quadratic matrix equation (3.5) and the eigenvalue problem (3.8'), combined with the existence of the nonnegative solutions Φ and Ψ with $\rho(\Phi) \le 1$ and $\rho(\Psi) \le 1$ by Theorem 3.2.

Finally, we state the following theorem, as a corollary of [14, Theorem 3.26 on p. 42].

Theorem 6.3. *To the conditions of* Theorem 6.2 *add these:* $\mu = 0$, $\rho(A_0) > 0$ *and* $\rho(A_2) > 0$. *Suppose* (6.11). *Then, besides all conclusions of* Theorem 6.2, *we have*

(i) $\max\{\|E_k\|, \|F_k\|\} \le O(2^{-k})$, *as* $k \to \infty$,
(ii) $\|\Phi - X_k\|, \|\Psi - Y_k\| \le O(2^{-k})$, *as* $k \to \infty$,
(iii) $I - X_kY_k$ *and* $I - Y_kX_k$ *approach singular matrices* $I - \Phi\Psi$ *and* $I - \Psi\Phi$. *Furthermore,*

$$(I - \Psi\Phi)\boldsymbol{u} = 0 \quad \textit{and} \quad (I - \Phi\Psi)\boldsymbol{u} = 0.$$

7. Entrywise relative residual

Given an approximation $\widetilde{\Phi} \approx \Phi$, the following normalized residual in norm (NRes):

$$\mathrm{NRes}(\widetilde{\Phi}) = \frac{\|A_0 + A_1\widetilde{\Phi} + A_2\widetilde{\Phi}^2 - \widetilde{\Phi}\|}{\|\widetilde{\Phi}\|(\|A_2\|\|\widetilde{\Phi}\| + \|A_1 - I\|) + \|A_0\|} \tag{7.1}$$

is the commonly used legacy measure to gauge how accurate $\widetilde{\Phi}$ may be, because of its computational availability, where $\|\cdot\|$ is some matrix norm, such as the ℓ_1 operator norm $\|\cdot\|_1$ which is the one we will use later for its computational convenience. NRes usually works well in the situation where only normwise accuracy is concerned. Although not guaranteed, often the relative error in norm (RErr), defined by

$$\mathrm{RErr}(\widetilde{\Phi}) = \frac{\|\widetilde{\Phi} - \Phi\|}{\|\widetilde{\Phi}\|}, \tag{7.2}$$

correlates well with $\mathrm{NRes}(\widetilde{\Phi})$ in the sense that both are tiny together. But NRes is not a good indicator on entrywise relative accuracy in general, unless all entries of Φ have comparable magnitudes. To overcome this shortcoming, we propose the so-called *entrywise relative residual* (ERRes) for the QBD equation (1.1)

$$\mathrm{ERRes}(\widetilde{\Phi}) = \max_{i,j} \frac{|(A_0 + A_1\widetilde{\Phi} + A_2\widetilde{\Phi}^2) - \widetilde{\Phi}|_{(i,j)}}{\widetilde{\Phi}_{(i,j)}}, \tag{7.3}$$

following the similar practice in [3]. Theorem 7.1 shows that ERRes is tiny, comparable to the entrywise accuracy in the entries of $\widetilde{\Phi}$. To state the theorem, we adopt the following floating point arithmetic model

$$\mathrm{fl}(\alpha \odot \beta) = (\alpha \odot \beta)(1 + \varepsilon), \quad |\varepsilon| \le \mathfrak{u} \text{ for } \odot \in \{+, -, \times, \div\}, \tag{7.4}$$

where $\mathrm{fl}(\cdot)$ is the computed result of an expression. All today's commercially significant machines run the IEEE floating point arithmetic [35,36] and thus conform to (7.4).

Theorem 7.1. *Suppose that all entries of the coefficient matrices A_i for $i = 0, 1, 2$ are floating point numbers and suppose that*

$$\widetilde{\Phi}_{(i,j)} = \Phi_{(i,j)}(1 + \varepsilon_{ij}) \quad \textit{with } |\varepsilon_{ij}| \le \delta < 1 \textit{ for all } i, j.$$

Then $\mathrm{fl}(A_0 + A_1\widetilde{\Phi} + A_2\widetilde{\Phi}^2) = \widetilde{\Phi} + \widetilde{E}$ *with*

$$|\widetilde{E}| \le [2n^2\mathfrak{u} + 3\delta + O(\mathfrak{u}^2 + \delta^2 + \delta\mathfrak{u})]\widetilde{\Phi}.$$

Proof. Keeping in mind that A_i for $0 \le i \le 2$ and $\widetilde{\Phi}$ are nonnegative, we have

$$\mathrm{fl}(A_0 + A_1\widetilde{\Phi} + A_2\widetilde{\Phi}^2) = A_0 + A_1\widetilde{\Phi} + A_2\widetilde{\Phi}^2 + E,$$

with $|E| \le [2n^2\mathfrak{u} + O(\mathfrak{u}^2)](A_0 + A_1\widetilde{\Phi} + A_2\widetilde{\Phi}^2)$. Also we notice

$$A_0 + A_1\widetilde{\Phi} + A_2\widetilde{\Phi}^2 = A_0 + A_1\Phi + A_2\Phi^2 + \widehat{E} = \Phi + \widehat{E}$$

with $|\widehat{E}| \le [2\delta + O(\delta^2)](A_0 + A_1\widetilde{\Phi} + A_2\widetilde{\Phi}^2) = [2\delta + O(\delta^2)]\Phi$. Write $\widetilde{\Phi} = \Phi + F$. We have

$$\mathrm{fl}(A_0 + A_1\widetilde{\Phi} + A_2\widetilde{\Phi}^2) = \Phi + \widehat{E} + E = \widetilde{\Phi} + \widehat{E} + E - F =: \widetilde{\Phi} + \widetilde{E},$$

where

$$\begin{aligned} |\widetilde{E}| = |\widehat{E} + E - F| &\le [2\delta + O(\delta^2)]\Phi + [2n^2\mathfrak{u} + O(\mathfrak{u}^2)]\Phi + \delta\,\Phi \\ &= [2n^2\mathfrak{u} + 3\delta + O(\mathfrak{u}^2 + \delta^2 + \delta\mathfrak{u})]\widetilde{\Phi}, \end{aligned}$$

as expected. □

The next theorem says that if $\mathrm{ERRes}(\widetilde{\Phi})$ is sufficiently tiny, then some multiple of it by a constant factor, also called the *condition number* but in the entrywise sense, can tell entrywise relative accuracy in $\widetilde{\Phi}$ as an approximation to Φ, much like the role played by NRes in telling the relative error (7.2) of $\widetilde{\Phi}$ in norm.

Theorem 7.2. *Let $\widetilde{\Phi} \approx \Phi$ such that $\widetilde{\Phi}$ and Φ share the same entrywise nonzero pattern. Suppose the* QBD *equation* (1.1) *is not in the critical case, i.e., either* (1.3a) *or $\{\mu \ne 0$, $\rho(A_0) > 0$ and $\rho(A_2) > 0\}$. If $\mathrm{ERRes}(\widetilde{\Phi}) \le \varepsilon$ and if ε is sufficiently tiny, then*

$$\begin{aligned} |(\Phi - \widetilde{\Phi}) \oslash \Phi| &\le \varepsilon\Upsilon \oslash \Phi + O\left(\varepsilon^2\right) \\ &\le \gamma\varepsilon\,\mathbf{1}_{n\times n} + O\left(\varepsilon^2\right), \end{aligned} \tag{7.5}$$

where $\oslash$ denotes the entrywise division, Υ and γ are defined by

$$(I - A_1 - A_2\Phi)\Upsilon - A_2\Upsilon\Phi = \Phi, \quad \gamma = \max_{i,j}(\Upsilon \oslash \Phi)_{(i,j)}.$$

Proof. Write $\Delta\Phi = \Phi - \widetilde{\Phi}$ and $A_0 + A_1\widetilde{\Phi} + A_2\widetilde{\Phi} = \widetilde{\Phi} + E$. By the definition (7.3),

$$|E| \le \varepsilon\,\widetilde{\Phi} = \varepsilon\,\Phi + \varepsilon\,(\Delta\Phi).$$

Then $\widetilde{\Phi} = \Phi - \Delta\Phi$ and

$$A_0 + A_1(\Phi - \Delta\Phi) + A_2(\Phi - \Delta\Phi)^2 = \Phi - \Delta\Phi + E,$$

which, after rearrangement, becomes

$$(I - A_1 - A_2\Phi)(\Delta\Phi) - A_2(\Delta\Phi)\Phi = E - A_2(\Delta\Phi)^2. \tag{7.6}$$

Define the linear operator $\mathscr{L}_\Phi$ as in (4.17a) which is invertible and $\mathscr{L}_\Phi^{-1}$ is nonnegative in the sense that it maps nonnegative matrices into nonnegative ones by Theorem 4.8. Following Stewart's argument [17, p. 242], we use the following iteration

$$Z_0 = 0,\ Z_{i+1} = \mathscr{L}_\Phi^{-1}(E - A_2 Z_i^2) \text{ for } i \geq 0$$

to conclude that for sufficiently tiny ε, (7.6) has a unique solution $\Delta\Phi = O(\varepsilon)$. Therefore

$$|\Delta\Phi| \leq \mathscr{L}_\Phi^{-1}\big(|E| + O(\varepsilon^2)\big) = \varepsilon\,\mathscr{L}_\Phi^{-1}\big(\Phi\big) + O(\varepsilon^2) = \varepsilon\,\Upsilon + O(\varepsilon^2)$$

which yields (7.5) since Φ and $\widetilde{\Phi}$ are assumed to have the same entrywise nonzero pattern. □

One of the conditions of Theorem 7.2 is that $\widetilde{\Phi}$ and Φ have the same entrywise nonzero pattern. In section Appendix A of the appendix, some sufficient conditions are given for the approximations by the doubling algorithms to have the same entrywise nonzero pattern as the exact Φ.

Example 7.1. Here we will use an example from [8] to numerically illustrate the superiority of ERRes over NRes in revealing the entrywise relative error

$$\mathrm{ERErr}(\widetilde{\Phi}) = \max_{i,j} \frac{|(\widetilde{\Phi} - \Phi)_{(i,j)}|}{\Phi_{(i,j)}}. \tag{7.7}$$

In this example, $n = 24$, and the matrices A_0, A_1, and A_2 are set up as follows. We first define 24×24 matrices A_0', A_1' and A_2' by

$$(A_0')_{ij} = \begin{cases} 192(1 - i/24), & i = j, \\ 0, & i \neq j, \end{cases} \qquad (A_2')_{ij} = \begin{cases} 192\rho_d, & i = j, \\ 0, & i \neq j, \end{cases}$$

$$(A_1')_{ij} = \begin{cases} \alpha r(\beta - i)/\beta, & i - j = -1, \\ ir, & i - j = 1, \\ \xi_i, & i - j = 0, \\ 0, & \text{elsewhere}, \end{cases}$$

where $0 \leq i, j \leq 23$, α, r, β and ρ_d are parameters, and all ξ_i are determined by $(A_0' + A_1' + A_2')\mathbf{1} = 0$. Then let

$$A_0 = -(A_1')^{-1}A_0', \quad A_1 = 0, \quad A_2 = -(A_1')^{-1}A_2'.$$

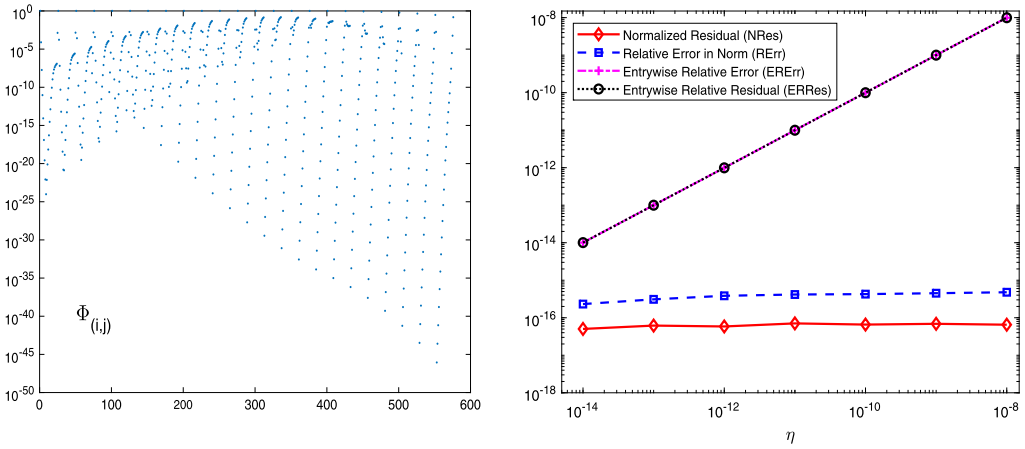

Fig. 7.1. Example 7.1. *Left:* Entries of Φ, many of which are much tinier than the machine roundoff $\mathfrak{u} \approx 10^{-16}$; *Right:* ERRes and ERErr move in sync, whereas NRes and RErr remain "constant". The curves for ERRes and ERErr indistinguishably collapse together.

Here we just consider one set of parameters: $\beta = 512$, $r = 1/300$, $\alpha = 18.244$ and $\rho_d = 0.280$. We compute the "exact" solution Φ by the computerized algebra system *Maple* with 100 decimal digits and we find that

$$8.6097 \cdot 10^{-47} \leq \Phi_{(i,j)} \leq 9.9868 \cdot 10^{-1}.$$

In fact, many entries of Φ are much tinier than the machine roundoff $\mathfrak{u} = 2^{-53} \approx 10^{-16}$, as is clear from the left plot in Fig. 7.1. We purposely perturb this Φ to get $\widetilde{\Phi}$ such that (1) $\|\widetilde{\Phi} - \Phi\|_1$ is always about $O(\mathfrak{u})$ and (2) ERErr varies from $O(\mathfrak{u})$ to about 10^{-8}. Specifically, we let in MATLAB

$$\widetilde{\Phi} = \Phi + \texttt{sign}(\texttt{randn}(\texttt{n})).*\texttt{min}(\texttt{Phimax}*\texttt{rand}(\texttt{n})*\texttt{eps}, (\texttt{Phi}*\texttt{eta}).*\texttt{rand}(\texttt{n})),$$

where $\texttt{Phimax} = 9.9868 \cdot 10^{-1}$ is the largest entry in Φ, `eta` varies from 10^{-14} to 10^{-8} so as to make ERErr vary from 10^{-15} to 10^{-8}, and `sign(randn(n))` is to make sure the entries of Φ are perturbed randomly up or down. Fig. 7.1 shows how NRes, RErr, ERRes, and ERErr change as `eta` varies. It clearly shows that ERRes and ERErr move in sync, whereas NRes and RErr remain "constant". This numerically demonstrates the capability of ERRes in revealing the entrywise relative accuracy in an approximation $\widetilde{\Phi}$, in addition to the theoretical justification we have in Theorem 7.2. In practice, since ERErr is not available because exact Φ is not known, ERRes is the perfect candidate to use because it is easily computable, just as we commonly use NRes (or some comparable quantities) in various numerical linear algebra problems. ◇

8. Highly accurate implementation

In this section, we will extend the idea in [3] to present a highly accurate implementation of Algorithm 5.1 for solving (1.1). It is made possible by the GTH-like algorithm of Alfa, Xue, and Ye [20] to invert nonsingular M-matrices

$$I - A_1, \quad I - X_k Y_k, \quad I - Y_k X_k \tag{8.1}$$

to almost full entrywise relative accuracy. For that purpose, we will need to generate entrywise accurate triplet representations for these matrices in (8.1). A key prerequisite for being able to do that is knowing A_0, A_1, A_2, and the vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ in (3.1) to almost full entrywise relative accuracy, which we will assume.

To begin with, we have for $I - A_1$

$$\hat{\boldsymbol{v}} := (I - A_1)\boldsymbol{u} = \boldsymbol{v} + (A_0 + A_2)\boldsymbol{u} \tag{8.2}$$

which can be computed without cancellation to give a triplet representation

$$I - A_1 = \{N_{I-A_1}, \boldsymbol{u}, \hat{\boldsymbol{v}}\}$$

to almost full entrywise relative accuracy, where N_{I-A_1} is the opposite in sign of off-diagonal part of $I - A_1$. Thus $(I - A_1)^{-1}$ can be computed to almost full entrywise relative accuracy by the GTH-like algorithm. Consequently, X_0, Y_0, E_0, and F_0 defined by (1.5) can be computed to almost full entrywise relative accuracy. Let

$$\begin{bmatrix} \boldsymbol{w}_1^{(0)} \\ \boldsymbol{w}_2^{(0)} \end{bmatrix} := \begin{bmatrix} \boldsymbol{u} \\ \boldsymbol{u} \end{bmatrix} - \begin{bmatrix} E_0 & Y_0 \\ X_0 & F_0 \end{bmatrix} \begin{bmatrix} \boldsymbol{u} \\ \boldsymbol{u} \end{bmatrix} \tag{8.3}$$

$$\begin{aligned} &= \begin{bmatrix} \boldsymbol{u} - (E_0 + Y_0)\boldsymbol{u} \\ \boldsymbol{u} - (X_0 + F_0)\boldsymbol{u} \end{bmatrix} \\ &= \begin{bmatrix} \boldsymbol{u} - (I - A_1)^{-1}(A_0 + A_2)\boldsymbol{u} \\ \boldsymbol{u} - (I - A_1)^{-1}(A_0 + A_2)\boldsymbol{u} \end{bmatrix} \\ &= \begin{bmatrix} (I - A_1)^{-1}(I - A_1 - A_0 - A_2)\boldsymbol{u} \\ (I - A_1)^{-1}(I - A_1 - A_0 - A_2)\boldsymbol{u} \end{bmatrix} \\ &= \begin{bmatrix} (I - A_1)^{-1}\boldsymbol{v} \\ (I - A_1)^{-1}\boldsymbol{v} \end{bmatrix} \geq 0. \end{aligned} \tag{8.4}$$

Here (8.3) defines $\boldsymbol{w}_i^{(0)}$ for $i = 1, 2$ but their actual computation is done according to (8.4) to almost full entrywise relative accuracy, for the same reason as we just argued.

The constructions of triplet representations for $I - X_k Y_k$ and $I - Y_k X_k$ are exactly the same as in [3]. The next theorem is essentially [3, Theorem 3.2], except for the case $k = 0$.

Theorem 8.1. *Suppose* (1.3a), *or suppose* (1.3b) *with* (6.6). *Let* $\{E_k, F_k, X_k, Y_k\}_{k=0}^{\infty}$ *be generated by* Algorithm 5.1. *Then*

$$\begin{bmatrix} E_k & Y_k \\ X_k & F_k \end{bmatrix} \begin{bmatrix} \boldsymbol{u} \\ \boldsymbol{u} \end{bmatrix} \le \begin{bmatrix} \boldsymbol{u} \\ \boldsymbol{u} \end{bmatrix} \text{ for } k \ge 0. \tag{8.5}$$

Moreover, they are equalities if $\boldsymbol{v} = 0$.

Proof. The inequality (8.5) holds for $k = 0$ and it is an equality if $\boldsymbol{v} = 0$, because of (8.4). The proof for $k \ge 1$ is exactly the same as that of [3, Theorem 3.2]. □

As in [3], we define

$$\begin{bmatrix} \boldsymbol{w}_1^{(k)} \\ \boldsymbol{w}_2^{(k)} \end{bmatrix} := \begin{bmatrix} \boldsymbol{u} \\ \boldsymbol{u} \end{bmatrix} - \begin{bmatrix} E_k & Y_k \\ X_k & F_k \end{bmatrix} \begin{bmatrix} \boldsymbol{u} \\ \boldsymbol{u} \end{bmatrix}. \tag{8.6}$$

It is nonnegative by Theorem 8.1. We emphasize that (8.6) is for definition only. A cancellation-free way to compute $\boldsymbol{w}_i^{(k)}$ to almost full entrywise relative accuracy is detailed in the next theorem.

Theorem 8.2 *([3]). Let* $\boldsymbol{w}_i^{(k)}$ $(i = 1, 2)$ *be defined by* (8.6). *Then*

$$\boldsymbol{w}_1^{(k+1)} = \boldsymbol{w}_1^{(k)} + E_k(I - Y_k X_k)^{-1} \left[\boldsymbol{w}_1^{(k)} + Y_k \boldsymbol{w}_2^{(k)} \right], \tag{8.7a}$$

$$\boldsymbol{w}_2^{(k+1)} = \boldsymbol{w}_2^{(k)} + F_k(I - X_k Y_k)^{-1} \left[X_k \boldsymbol{w}_1^{(k)} + \boldsymbol{w}_2^{(k)} \right]. \tag{8.7b}$$

Finally, we have [3],

$$I - Y_k X_k = \{N_{I - Y_k X_k}, \boldsymbol{u}, \boldsymbol{v}_1^{(k)}\}, \tag{8.8a}$$

$$I - X_k Y_k = \{N_{I - X_k Y_k}, \boldsymbol{u}, \boldsymbol{v}_2^{(k)}\}, \tag{8.8b}$$

where

$$\boldsymbol{v}_1^{(k)} \equiv (I - Y_k X_k)\boldsymbol{u} = \boldsymbol{w}_1^{(k)} + E_k \boldsymbol{u} + Y_k \left[F_k \boldsymbol{u} + \boldsymbol{w}_2^{(k)} \right] \ge 0, \tag{8.9a}$$

$$\boldsymbol{v}_2^{(k)} \equiv (I - X_k Y_k)\boldsymbol{u} = \boldsymbol{w}_2^{(k)} + F_k \boldsymbol{u} + X_k \left[E_k \boldsymbol{u} + \boldsymbol{w}_1^{(k)} \right] \ge 0. \tag{8.9b}$$

With all these, we outline our highly accurate doubling algorithm as in Algorithm 8.1 to solve (1.1). Except in its initialization phase at lines 1 – 3, Algorithm 8.1 is exactly the same as [3, Algorithm 5.1]

Our discussion below on when to stop at line 10 is essentially the same as the third comment in [3, p. 753] for MARE. There are three viable options for use as stopping criteria:

Algorithm 8.1 accDAQBD: highly accurate doubling algorithm for QBD equation (1.1).

Input: A_i for $i = 0, 1, 2$ and the vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ as in (3.1);
Output: minimal nonnegative solutions Φ and Ψ (if needed).
1: compute a triplet representation of $I - A_1 = \{N_{I-A_1}, \boldsymbol{u}, \hat{\boldsymbol{v}}\}$ according to (8.2);
2: compute E_0, F_0, X_0 and Y_0 according to (1.5) by the GTH-like algorithm using the triplet representation for $I - A_1$;
3: compute $\boldsymbol{w}_1^{(0)}$ and $\boldsymbol{w}_2^{(0)}$ according to (8.4) by the GTH-like algorithm using the triplet representation for $I - A_1$;
4: $k = -1$;
5: **repeat**
6: $k = k + 1$;
7: compute $\boldsymbol{v}_1^{(k)}$ and $\boldsymbol{v}_2^{(k)}$ as defined in (8.9) and generate the triplet representations for $I - Y_k X_k$ and $I - X_k Y_k$ as in (8.8);
8: compute $E_{k+1}, F_{k+1}, X_{k+1}$ and Y_{k+1} according to (5.2) by the GTH-like algorithm using the triplet representations for $I - Y_k X_k$ and $I - X_k Y_k$
9: compute $\boldsymbol{w}_1^{(k+1)}$ and $\boldsymbol{w}_2^{(k+1)}$ according to (8.7) (reuse $E_k(I - Y_k X_k)^{-1}$ and $F_k(I - X_k Y_k)^{-1}$ that appear in implementing line 8 to reduce work);
10: **until** convergence;
11: **return** the last X_k and Y_k as approximations to Φ and Ψ, respectively.

$$|X_{k+1} - X_k| \le \varepsilon \cdot X_{k+1}, \tag{8.10a}$$

$$\mathrm{ERRes}(X_{k+1}) \le \varepsilon, \tag{8.10b}$$

$$\frac{(X_{k+1} - X_k)^2_{(i,j)}}{(X_k - X_{k-1})_{(i,j)} - (X_{k+1} - X_k)_{(i,j)}} \le \varepsilon \cdot (X_{k+1})_{(i,j)} \quad \text{for all } i \text{ and } j, \tag{8.10c}$$

where ε is a pre-selected tolerance. The first one (8.10a) is the simplest and also cheapest one to use, the second one (8.10b) is based on our newly proposed entrywise relative residual (7.3), and the third one (8.10c) is Kahan's stopping criterion, previously in [37,21,4]. Both the simple (8.10a) and Kahan's stopping criterion (8.10c) can be too conservative in the case of a monotonically quadratically convergent sequence in the sense that they stop iterations unnecessarily late, wasting the last one or two iterations. With the same ε, (8.10a) is even more conservative than (8.10c) because, in the phase of quadratic convergence,

$$(X_k - X_{k-1})_{(i,j)} - (X_{k+1} - X_k)_{(i,j)} \approx (X_k - X_{k-1})_{(i,j)} \gg (X_{k+1} - X_k)_{(i,j)},$$

and when $X_{k+1} - X_k = O([X_k - X_{k-1}]^2)$, the left hand side of (8.10c) is $O([(X_{k+1} - X_k)_{(i,j)}]^{3/2})$ which is much tinier than $(X_{k+1} - X_k)_{(i,j)}$. Another shortcoming for both is a possible pitfall: false-convergence in the sense that the iteration may be stopped due to a period of very slow moving X_k. The second stopping criterion (8.10b) is most expensive to use among the three, especially $\mathrm{ERRes}(X_{k+1})$ is not needed in the doubling iteration kernel. But it does not have the pitfall mentioned above.

In view of this discussion, we propose to use Kahan's stopping criterion (8.10c) with a safeguard, in the sense that when Kahan's stopping criterion is satisfied we check if (8.10b) (probably with a different ε) is also satisfied to avoid possible false-convergence. After numerous numerical experiments, we find that in the non-critical case ε about 10^{-10} to 10^{-12} works the best for computed solution to achieve its deserved entrywise

relative accuracy about $O(10^{-15})$ without wasting the last iteration step (although not guaranteed). But in the critical case, ε should be set to about 10^{-14} to 10^{-16} because of linear convergence.

9. Numerical examples

In this section, we will present five numerical examples to illustrate the superior performance of Algorithm 8.1 in delivering entrywise accuracy in computed $\widetilde{\Phi}$ as an approximation to the exact solution Φ. Four algorithms will be tested.

1. DAQBD: the plain doubling algorithm, i.e., Algorithm 5.1 with inputs $X_0, Y_0, E_0, F_0 \in \mathbb{R}^{n\times n}$ determined by (1.5). It simply uses the usual Gaussian elimination with partial pivoting, such as MATLAB's operators "\" and "/", to carry out all the inversions in (1.5) and (5.2).
2. accDAQBD: Algorithm 8.1.
3. accDAQBD-lite: a lite version of Algorithm 8.1. It simply sets, in view of (8.9),

$$\boldsymbol{v}_1^{(k)} = (I - Y_k X_k)\boldsymbol{u}, \quad \boldsymbol{v}_2^{(k)} = (I - X_k Y_k)\boldsymbol{u} \tag{9.1}$$

   to replace line 7 there.
4. accLRQBD: Ye's highly accurate implementation of the Latouche-Ramaswami algorithm [6, Algorithm 3]. It is a highly accurate implementation of the logarithmic reduction algorithm of Latouche and Ramaswami [5]. Basically what it does is to carry out all inversions of the original Latouche-Ramaswami algorithm by the GTH-like algorithm. However, this algorithm only works for QBD equations with $\boldsymbol{u} = \mathbf{1}_n$ and $\boldsymbol{v} = 0$, and with minor modifications,[10] it can be made to work for those with $\boldsymbol{v} = 0$ but $\boldsymbol{u} \neq \mathbf{1}_n$. Latouche-Ramaswami algorithm without changes does work for $\boldsymbol{v} \neq 0$ and $\boldsymbol{u} = \mathbf{1}_n$ [13], but Ye's implementation remains to be extended. For this reason, we will not test the Latouche-Ramaswami algorithm for the case when $\boldsymbol{v} \neq 0$ and leave it for future study.

As stated in [3], the use of (9.1) may render some entries of computed $\boldsymbol{v}_i^{(k)}$ $(i = 1, 2)$ negative, albeit tiny, due to roundoff errors, especially in the critical case where $(I - Y_k X_k)\boldsymbol{u}$ and $(I - X_k Y_k)\boldsymbol{u}$ are known to converge to 0. Since in theory $\boldsymbol{v}_i^{(k)} \geq 0$, if there are negative entries in the computed $\boldsymbol{v}_i^{(k)}$ by (9.1) then the value of such an entry must be comparable to the roundoff error in evaluating it. For this reason, as a safe-guard, we reset all negative entries, if any, in the computed $\boldsymbol{v}_i^{(k)}$ to 0.

At convergence, DAQBD, accDAQBD-lite and accDAQBD output both Φ and Ψ (even though Ψ, the minimal nonnegative solution of the dual equation (3.5), may not

[10] Replace all A_i by $D^{-1}A_iD$, where $D = \mathrm{diag}(\boldsymbol{u})$, and finally at convergence recover the solution as $D^{-1}XD$, where X is the computed solution with the modified A_i.

be needed), while accLRQBD computes Φ only. Per iterative step, each of DAQBD, accDAQBD-lite and accDAQBD requires two matrix inversions and eight matrix multiplications, while accLRQBD uses one matrix inversion and six matrix multiplications and thus costs less. But the original accLRQBD as is does not work for the case $\boldsymbol{v} \neq 0$.

In reporting numerical results, we will plot iterative history curves for entrywise relative residual ERRes defined by (7.3), normalized residual NRes defined by (7.1), and entrywise relative error ERErr defined by (7.7) which is not available in practice but made available here for testing purposes. In fact, in what follows, Φ is known explicitly only for the first example and for all other examples, it is computed by the computerized algebra system Maple with 100 decimal digits for our testing purpose.

All computations are done in MATLAB. We will use Kahan's stopping criterion (8.10c) for the purpose to achieve $X_{k+1} \approx \Phi$ entrywise. We set $\varepsilon = 10^{-14}$ for Example 9.1 (because it is in the critical case), and $\varepsilon = 10^{-12}$ for all others.

Example 9.1 *([38]).* Consider the QBD equation (1.1) with

$$A_0 = \begin{bmatrix} 0.25 & 0 \\ 0.25 & 0 \end{bmatrix}, \quad A_1 = \begin{bmatrix} 0.25 & 0.25 \\ 0.25 & 0.25 \end{bmatrix}, \quad A_2 = \begin{bmatrix} 0 & 0.25 \\ 0 & 0.25 \end{bmatrix}.$$

Its exact minimal nonnegative solution is

$$\Phi = \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix}.$$

Fig. 9.1 plots the convergence history for Example 9.1 (data points appear in the plot only when they are positive). This figure clearly shows linear convergence for all methods, but there is a major difference. All ERRes and NRes curves reach about $O(10^{-15})$ at iterative step 24 or 25. But the ERErr curves for DAQBD and accDAQBD-lite behave differently from those for accDAQBD and accLRQBD in that the ones for DAQBD and accDAQBD-lite refuse to move below $O(10^{-9})$ while the ones for accDAQBD and accLRQBD move down all the way to $O(10^{-15})$. Also, the ERErr curve for DAQBD starts to wobble at iteration 26, while that for accDAQBD-lite starts to flatten out at iteration 27. Evidently, $\varepsilon = 10^{-14}$ is too tiny for the doubling iterations in both DAQBD and accDAQBD-lite to stop. In fact, their computed Φ at convergence are

$$\text{DAQBD}: \begin{bmatrix} 0.999999998798091 & 0 \\ 0.999999998798091 & 0 \end{bmatrix},$$

$$\text{accDAQBD-lite}: \begin{bmatrix} 0.999999993661981 & 0 \\ 0.999999993661981 & 0 \end{bmatrix},$$

$$\text{accDAQBD}: \begin{bmatrix} 0.999999999999993 & 0 \\ 0.999999999999993 & 0 \end{bmatrix},$$

$$\text{accLRQBD}: \begin{bmatrix} 0.999999999999993 & 0 \\ 0.999999999999993 & 0 \end{bmatrix},$$

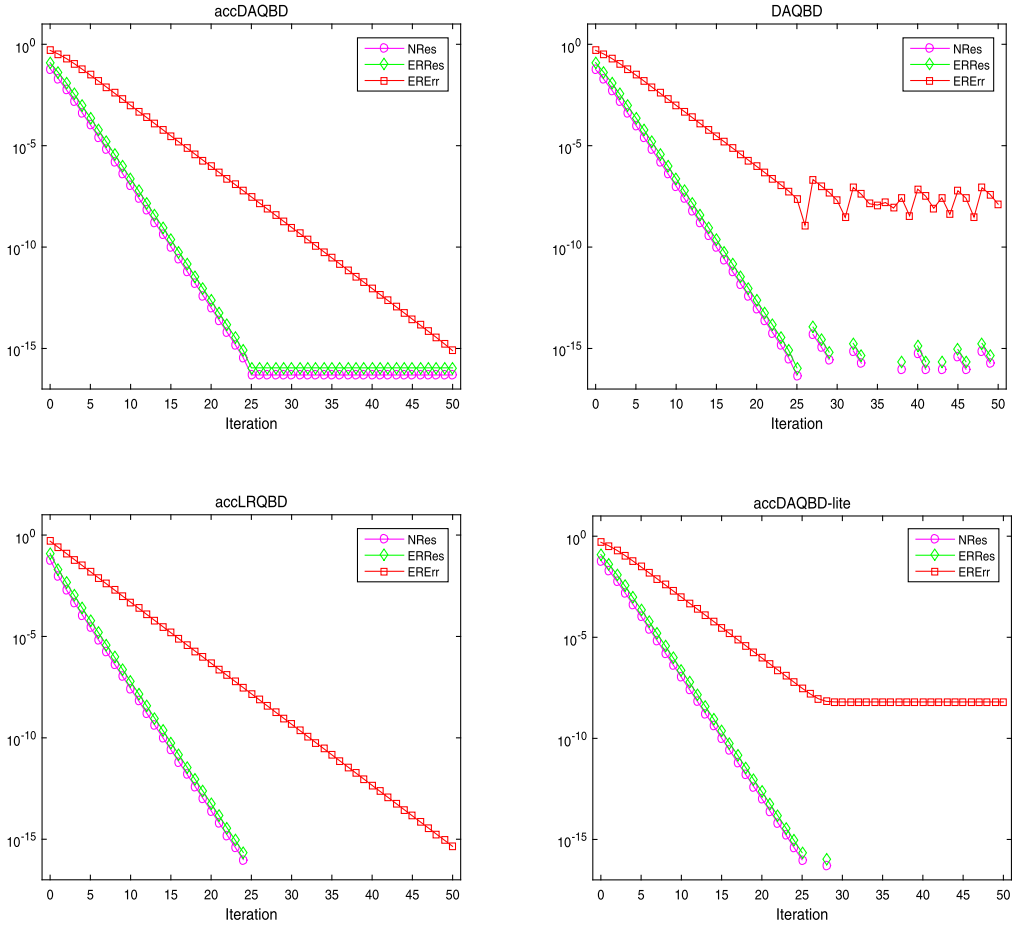

Fig. 9.1. Convergence history curves for Example 9.1. Both accDAQBD and accLRQBD can compute Φ to full entrywise relative accuracy, while DAQBD and accDAQBD-lite cannot. Some NRes and ERRes values are accidentally computed to 0.0 by roundoff and they are revealed by no markers shown.

respectively. We note all get[11] 0 exactly, but the accuracies in approximating the entry 1 are different, with DAQBD and accDAQBD-lite getting errors of just $O(\sqrt{u})$. This difference is caused by with or without using the GTH-like algorithm for inversions. The same phenomenon occurred to numerical solutions of MARE in [3]. ◇

As in Example 9.1, DAQBD and accDAQBD-lite always perform no better, if not significantly worse, than accDAQBD in all our tests, including many not reported here. For this reason and in order to save space, in what follows, we will omit reporting numerical results by DAQBD and accDAQBD-lite.

[11] These entries start with 0 and never get touched during the doubling iterative process.

Table 9.1
Ranges of Φ for Example 9.2 (all $\max_{i,j} \Phi_{(i,j)} \approx 1.0$).

$r = 1/300$, $\alpha = 18.244$, $\rho_d = 0.280$						
β	64	256	1024	4096	16384	65536
$\min_{i,j} \Phi_{(i,j)}$	$6.4 \cdot 10^{-59}$	$2.0 \cdot 10^{-57}$	$4.1 \cdot 10^{-57}$	$5.0 \cdot 10^{-57}$	$5.2 \cdot 10^{-57}$	$5.3 \cdot 10^{-57}$
$r = 1/100$, $\alpha = 18.244$, $\beta = 512$						
ρ_d	0.075	0.1	0.18	0.26	0.29	0.29568
$\min_{i,j} \Phi_{(i,j)}$	$2.5 \cdot 10^{-55}$	$7.7 \cdot 10^{-54}$	$4.5 \cdot 10^{-50}$	$2.4 \cdot 10^{-47}$	$1.6 \cdot 10^{-46}$	$2.2 \cdot 10^{-46}$

Table 9.2
Example 9.2 ($r = 1/300$, $\alpha = 18.244$, $\rho_d = 0.280$).

β	accDAQBD		accLRQBD	
	Iter	ERErr	Iter	ERErr
64	18	9.7875×10^{-15}	17	8.0522×10^{-15}
256	20	8.4767×10^{-15}	19	6.5757×10^{-15}
1024	22	8.9616×10^{-15}	21	8.3499×10^{-15}
4096	24	8.7352×10^{-15}	23	1.4428×10^{-15}
16384	27	7.6983×10^{-15}	26	7.9121×10^{-15}
65536	34	1.2750×10^{-14}	33	1.4831×10^{-14}

Table 9.3
Example 9.2 ($r = 1/100$, $\alpha = 18.244$, $\beta = 512$).

ρ_d	accDAQBD		accLRQBD	
	Iter	ERErr	Iter	ERErr
0.075	13	1.4158×10^{-14}	12	7.1819×10^{-15}
0.1	14	1.3486×10^{-14}	13	3.1512×10^{-14}
0.18	16	3.3683×10^{-15}	15	2.7801×10^{-15}
0.26	19	4.1455×10^{-15}	18	3.9467×10^{-15}
0.29	21	6.6874×10^{-15}	20	6.8350×10^{-15}
0.29568	30	7.6001×10^{-15}	29	4.2545×10^{-15}

Example 9.2 *([8]).* This is a revisit of Example 7.1. We will perform two groups of tests with varying parameters. In our first group of tests, we vary β while keeping $r = 1/300$, $\alpha = 18.244$ and $\rho_d = 0.280$, and for our second group, we vary ρ_d while keeping $r = 1/100$, $\alpha = 18.244$ and $\beta = 512$. The entries of the "exact" solutions Φ vary wildly in magnitude, as shown in Table 9.1.

In Tables 9.2 and 9.3, we recorded the numbers of iterations (Iter) needed for solving the associated QBD equations for different parameter choices, along with ERErr at convergence, by accDAQBD and accLRQBD. We find that both accDAQBD and accLRQBD are able to deliver X_k with ERErr of $O(10^{-15})$, and accLRQBD uses one fewer iteration than accDAQBD.

Fig. 9.2 plots the convergence history curves for the two most difficult cases: $\beta = 65536$ or $\rho_d = 0.29568$, respectively. We see very slow convergence at the beginning, especially for ERErr, but eventually quadratic convergence near the end. ◇

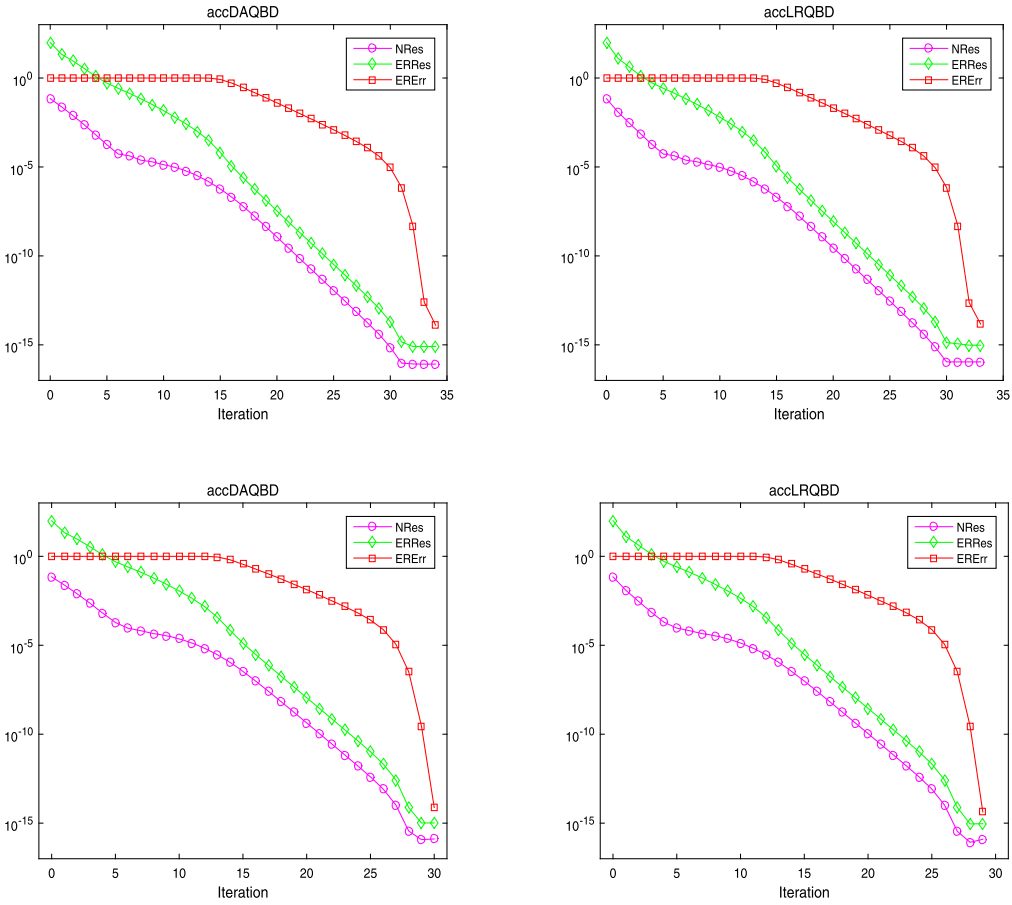

Fig. 9.2. Convergence history curves for Example 9.2. The top two are for $\beta = 65536$ while $r = 1/300$, $\alpha = 18.244$ and $\rho_d = 0.280$. The bottom two are for $\rho_d = 0.29568$ while $r = 1/100$, $\alpha = 18.244$ and $\beta = 512$.

The next example is about the case $\boldsymbol{v} = (I - A_0 - A_1 - A_2)\boldsymbol{u} = 0$ but $\boldsymbol{u} \neq \mathbf{1}_n$.

Example 9.3. We modify Example 9.2 to generate a QBD equation with $\boldsymbol{v} = 0$ but $\boldsymbol{u} \neq \mathbf{1}_n$. Specifically, we take $\boldsymbol{u} = [1, 2, \cdots, 24]^{\mathrm{T}}$, transform all A_i to DA_iD^{-1}, and again reassign the resulting matrices to A_i for notational simplicity, where $D = \mathrm{diag}(\boldsymbol{u})$. It is not hard to see that $(A_0 + A_1 + A_2)\boldsymbol{u} = \boldsymbol{u}$ for the updated A_i. For this example, accLRQBD is not directly applicable because it was designed for $\boldsymbol{u} = \mathbf{1}$, but, as we commented before, a minor modification that essentially undoes what we just did in constructing this example will solve the issue. In fact, the modification turns this example back into Example 9.2. Hence, numerically accLRQBD on this example is essentially the same as itself on Example 9.2, as we observed. We also observed that accDAQBD performs very much the same as itself on the previous example. For this reason, we omit the detail.

Table 9.4
Numbers of iterations (Iter) and ERErr for Example 9.4.

n	Method		δ			
			10^{-2}	10^{-4}	10^{-6}	10^{-8}
64	accDAQBD	Iter	11	17	23	29
		ERErr	2.2×10^{-15}	1.2×10^{-15}	8.8×10^{-16}	3.5×10^{-15}
	accLRQBD	Iter	10	16	22	28
		ERErr	1.8×10^{-15}	2.6×10^{-15}	8.8×10^{-16}	3.5×10^{-15}

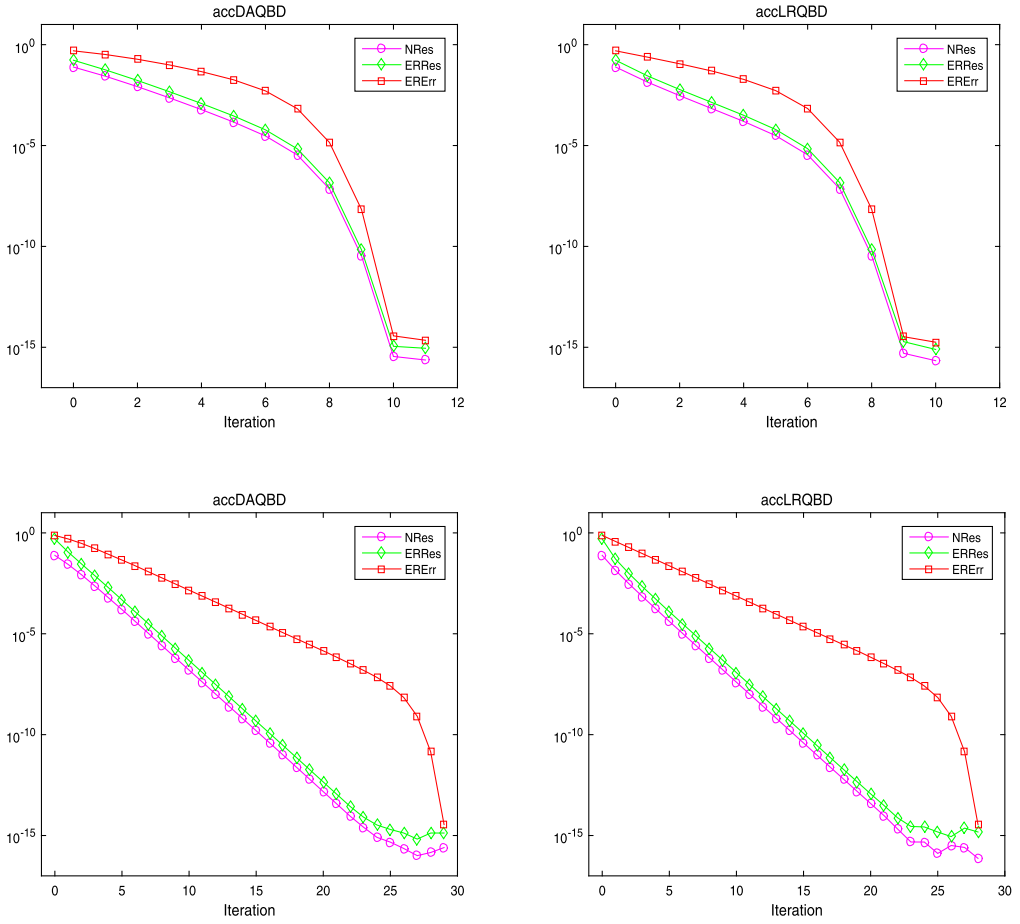

Fig. 9.3. Convergence history curves for Example 9.4 with $(\delta, n) = (64, 10^{-2})$ (top two plots) and $(\delta, n) = (64, 10^{-8})$ (bottom two plots).

Example 9.4 *([8,9])*. Let $A_0 = R + \delta I$, $A_1 = A_2 = R \in \mathbb{R}^{n\times n}$, where R has null diagonal entries but constant off-diagonal entries, and $0 < \delta < 1$. To ensure $(I - A_0 - A_1 - A_2)\mathbf{1} = 0$, we find the value of the off-diagonal entries of R to be $\frac{1-\delta}{3(n-1)}$. We will test accDAQBD and accLRQBD for four different values of δ and $n = 64$.

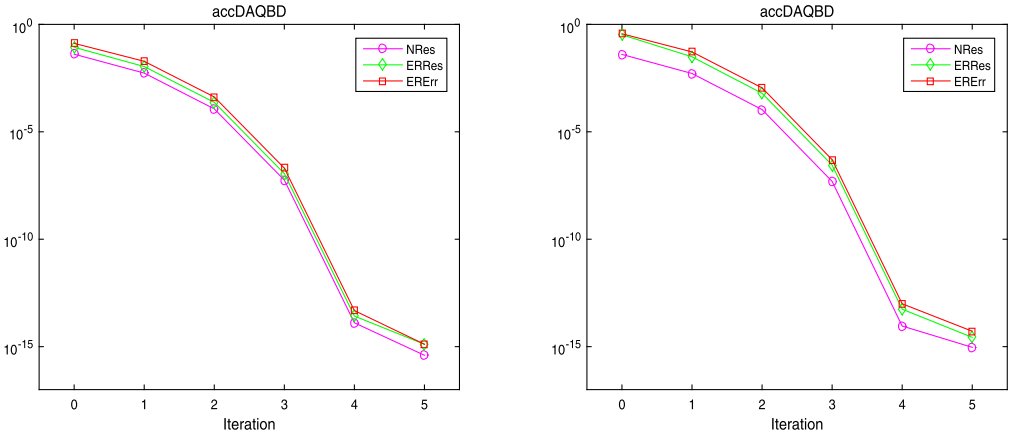

Fig. 9.4. Convergence history curves for Example 9.5 with $(\delta, n) = (64, 10^{-2})$ (left plot) and $(\delta, n) = (64, 10^{-8})$ (right plot).

Table 9.5
Numbers of iterations (Iter) and ERErr for Example 9.5.

n	Method		δ			
			10^{-2}	10^{-4}	10^{-6}	10^{-8}
64	accDAQBD	Iter	5	5	5	5
		ERErr	1.3×10^{-15}	1.2×10^{-15}	1.7×10^{-15}	5.2×10^{-15}

Table 9.4 displays the numbers of iterations (Iter) and ERErr by accDAQBD and accLRQBD. Again, we find that both accDAQBD and accLRQBD deliver X_k with ERErr of $O(10^{-15})$, almost full entrywise relative accuracy in the working precision. In addition, accLRQBD takes one fewer iteration than accDAQBD. Fig. 9.3 shows the convergence history curves for $(\delta, n) = (64, 10^{-2})$ and for $(\delta, n) = (64, 10^{-8})$. Quadratic convergence clearly shows. ◇

Finally, we present an example for the case $\boldsymbol{v} = (I - A_0 - A_1 - A_2)\boldsymbol{u} > 0$, in contrast to all examples so far. It is a simple modified version of Example 9.4.

Example 9.5. Let A_i for $0 \le i \le 3$ have the same form as in Example 9.4. To make sure $\boldsymbol{v} = (I - A_0 - A_1 - A_2)\mathbf{1} > 0$ (and thus $\boldsymbol{u} = \mathbf{1}$), we take the constant off-diagonal entries of R to be $\frac{1-\delta}{4(n-1)}$ which is less than $\frac{1-\delta}{3(n-1)}$, the value used in Example 9.4. Analogously, we will test accDAQBD for four different values of δ with $n = 64$. accLRQBD as is does not work for this example because $\boldsymbol{v} > 0$.

Table 9.5 reports the numbers of iterations (Iter) and ERErr by accDAQBD on this example. accDAQBD delivers X_k with ERErr of $O(10^{-15})$, almost full entrywise relative accuracy in the working precision. Fig. 9.4 shows convergence histories for $(\delta, n) = (64, 10^{-2})$ and for $(\delta, n) = (64, 10^{-8})$. Quadratic convergence again clearly shows. ◇

10. Conclusions

The structure-preserving doubling algorithm for the first standard form (SF1) [14] is extended to compute the minimal nonnegative solutions of a type of quadratic matrix equations that include the ones from the quasi-birth-and-death (QBD) process as a special case. It is shown that the approximations generated by the algorithm are globally and monotonically convergent, and the convergence is quadratical, except in the critical case where the convergence is only linear but with a respectable linear rate 1/2.

A highly accurate implementation of the algorithm, mostly along the lines in [3], is presented. The implementation resorts the GTH-like algorithm to invert all nonsingular M-matrices $I - X_kY_k$ and $I - Y_kX_k$ in the doubling iteration kernel without any substraction and, consequently, can compute the minimal nonnegative solution to high entrywise relative accuracy for entries, large and small, as warranted by the input data.

We also proposed a new entrywise relative residual (7.3) whose magnitude reflects the entrywise relative error (7.7), while the usually legacy normalized residual (7.1) can only reflect the relative error in norm (7.2) well.

Several numerical examples are presented to demonstrate the capability of the highly accurate implementation in delivering highly entrywise accurate solutions.

Declaration of competing interest

None.

Acknowledgements

The authors are grateful to both reviewers for their helpful comments and suggestions, and in particular for their pointing out an error in our earlier citation of the results in [25, chapter 5] on the zeroes of $\phi(\lambda)$, leading to our rewriting of section 4 based entirely on matrix analysis approach that is more accessible to the numerical linear algebra community than before. The authors also thank the editors for comments which have helped to improve the paper.

Appendix A. Sparsity of E_k, F_k, X_k, Y_k

In this section, we investigate the nonzero patterns in E_k, F_k, X_k, Y_k generated by the doubling iteration (5.2). The main result is stated in Theorem Appendix A.5, which presents sufficient conditions under which the iterative approximations X_k and Y_k by the doubling algorithm have the same zero-and-nonzero pattern as the minimal nonnegative solution Φ of the QBD equation (1.1) and as the minimal nonnegative solution Ψ of the dual QBD equation (3.5), respectively.

Xue and Li [3] introduced a partial ordering on nonnegative matrices with respect to their entrywise nonzero patterns. Let $P \geq 0$, $Q \geq 0$ be of the same size, we say

that Q *majorizes* P with respect to the entrywise nonzero pattern, written as $P \overset{0}{\preceq} Q$, if $Q_{(i,j)} = 0$ implies $P_{(i,j)} = 0$, and, Q and P are the same with respect to the entrywise nonzero pattern, written as $P \overset{0}{=} Q$ if $P \overset{0}{\preceq} Q$ and $Q \overset{0}{\preceq} P$.

Evidently, $0 \le P \le Q$ implies $P \overset{0}{\preceq} Q$, but not the other way around. Lemma Appendix A.1 is rather straightforward.

Lemma Appendix A.1 *([3]).*

(a) *If* $0 \le P_i \le Q_i$ *for* $i = 1, 2$, *all having the same size, then* $P_1 + P_2 \overset{0}{\preceq} Q_1 + Q_2$ *and* $P_1 P_2 \overset{0}{\preceq} Q_1 Q_2$.
(b) *If* $P \overset{0}{\preceq} Q$, *then* $P + Q \overset{0}{=} Q$.

The next three lemmas are also taken from [3].

Lemma Appendix A.2 *([3]). Let* P, Q *be nonsingular* M*-matrices, which are split as*

$$P = D_P - N_P, \quad D_P = \operatorname{diag}(P),$$
$$Q = D_Q - N_Q, \quad D_Q = \operatorname{diag}(Q).$$

If $N_P \overset{0}{\preceq} N_Q$, *then* $P^{-1} \overset{0}{\preceq} Q^{-1}$. *In particular, if* $N_P \overset{0}{=} N_Q$, *then* $P^{-1} \overset{0}{=} Q^{-1}$.

Lemma Appendix A.3 *([3]). For a nonsingular* M*-matrix* P, $P^{-1} \overset{0}{=} P^{-k}$ *for* $k \ge 1$, *and* $(\alpha I - P^{-1})^{-1} \overset{0}{=} P^{-1}$ *for* $\alpha > \rho(P^{-1})$.

Lemma Appendix A.4 *([3]). Let* P *be a nonsingular* M*-matrix and* $Q \ge 0$, *which are split* P, Q *as*

$$P = D_P - N_P, \quad D_P = \operatorname{diag}(P),$$
$$Q = D_Q + \widetilde{N}_Q, \quad D_Q = \operatorname{diag}(Q).$$

If $Q_{(i,i)} > 0$ *for all* i *and* $\widetilde{N}_Q \overset{0}{\preceq} N_P$, *then* $P^{-1}Q \overset{0}{=} P^{-1}$.

Our main result in this section is the next theorem.

Theorem Appendix A.5. *Let* E_0, F_0, X_0, Y_0 *be as in* (1.5) *and let* E_k, F_k, X_k, Y_k *be produced by the doubling iteration* (5.2). *Split* A_0, A_1, A_2 *as*

$$A_0 = D_{A_0} + \widetilde{N}_{A_0}, \quad D_{A_0} = \operatorname{diag}(A_0),$$
$$A_1 = D_{A_1} + \widetilde{N}_{A_1}, \quad D_{A_1} = \operatorname{diag}(A_1),$$
$$A_2 = D_{A_2} + \widetilde{N}_{A_2}, \quad D_{A_2} = \operatorname{diag}(A_2).$$

If $(A_0)_{(i,i)} > 0$ *and* $(A_2)_{(i,i)} > 0$ *for all* i *and if* $\widetilde{N}_{A_0} \overset{0}{\preceq} \widetilde{N}_{A_1}, \widetilde{N}_{A_2} \overset{0}{\preceq} \widetilde{N}_{A_1}$, *then for* $k \geq 0$

$$\begin{bmatrix} E_{k+1} & Y_{k+1} \\ X_{k+1} & F_{k+1} \end{bmatrix} \overset{0}{\preceq} \begin{bmatrix} E_k & Y_k \\ X_k & F_k \end{bmatrix}, \tag{A.1}$$

$X_{k+1} \overset{0}{=} X_k \overset{0}{=} \Phi$, *and* $Y_{k+1} \overset{0}{=} Y_k \overset{0}{=} \Psi$.

Proof. It is clear that $X_k \overset{0}{\preceq} X_{k+1}$ and $Y_k \overset{0}{\preceq} Y_{k+1}$ because of (5.2c) and (5.2d). So if (A.1) is proven true, then we will immediately have $X_{k+1} \overset{0}{=} X_k \overset{0}{=} \Phi$ and $Y_{k+1} \overset{0}{=} Y_k \overset{0}{=} \Psi$.

We will prove (A.1) by induction.

Consider first the base case $k = 0$. Note that $(A_0)_{(i,i)} > 0$ for all i and thus

$$E_0 = X_0 = (I - A_1)^{-1} A_0 \overset{0}{=} (I - A_1)^{-1}(I + \widetilde{N}_{A_0}) = (I - A_1)^{-1} + (I - A_1)^{-1}\widetilde{N}_{A_0}.$$

Therefore, immediately, $E_0 = X_0 \overset{0}{\succeq} (I - A_1)^{-1}$. On the other hand, we also have

$$E_0 = X_0 \overset{0}{\preceq} (I - A_1)^{-1} + (I - A_1)^{-1} A_1 \overset{0}{=} (I - A_1)^{-1}.$$

Together, we get

$$E_0 = X_0 = (I - A_1)^{-1} A_0 \overset{0}{=} (I - A_1)^{-1}. \tag{A.2a}$$

In exactly the same way, we can also get

$$F_0 = Y_0 = (I - A_1)^{-1} A_2 \overset{0}{=} (I - A_1)^{-1}. \tag{A.2b}$$

Let $Q = (I - A_1)^{-1} \geq 0$. It follows from (A.2a), (A.2b), and Lemma Appendix A.3 that

$$E_0 = X_0 \overset{0}{=} F_0 = Y_0 \overset{0}{=} (\alpha I - Q)^{-1}, \quad \alpha > \rho(Q),$$

from which we conclude that

$$\begin{bmatrix} E_0 & Y_0 \\ X_0 & F_0 \end{bmatrix} \overset{0}{=} \begin{bmatrix} (\alpha I - Q)^{-1} & (\alpha I - Q)^{-1} \\ (\alpha I - Q)^{-1} & (\alpha I - Q)^{-1} \end{bmatrix} =: P. \tag{A.3}$$

In particular,

$$\begin{bmatrix} 0 & Y_0 \\ X_0 & 0 \end{bmatrix} \overset{0}{\preceq} P, \quad \begin{bmatrix} E_0 & 0 \\ 0 & F_0 \end{bmatrix} \overset{0}{\preceq} P. \tag{A.4}$$

By Lemmas Appendix A.1 and Appendix A.3, we have

$$P^k \overset{0}{=} P \text{ for all } k \geq 1. \tag{A.5}$$

In addition, since $I \overset{0}{\preceq} (\alpha I - Q)^{-1} = \frac{1}{\alpha}(I + \frac{1}{\alpha}Q + \frac{1}{\alpha^2}Q^2 + \cdots)$, we have

$$\begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix} \overset{0}{\preceq} P. \tag{A.6}$$

By Lemma Appendix A.2, (A.3), (A.5) and (A.6), we have

$$\begin{aligned} \begin{bmatrix} I & -Y_0 \\ -X_0 & I \end{bmatrix}^{-1} &= \left(\begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix} - \begin{bmatrix} 0 & Y_0 \\ X_0 & 0 \end{bmatrix} \right)^{-1} \\ &\overset{0}{\preceq} \left(\begin{bmatrix} \gamma I & 0 \\ 0 & \gamma I \end{bmatrix} - \begin{bmatrix} E_0 & Y_0 \\ X_0 & F_0 \end{bmatrix} \right)^{-1} \\ &= \frac{1}{\gamma} \sum_{i=0}^{\infty} \left(\frac{1}{\gamma} \begin{bmatrix} E_0 & Y_0 \\ X_0 & F_0 \end{bmatrix} \right)^i \\ &\overset{0}{=} \sum_{i=0}^{\infty} P^i \overset{0}{=} P \end{aligned} \tag{A.7}$$

for γ large enough. Combine (A.7) and (A.4) to get

$$\begin{aligned} \begin{bmatrix} E_1 & Y_1 \\ X_1 & F_1 \end{bmatrix} &= \begin{bmatrix} 0 & Y_0 \\ X_0 & 0 \end{bmatrix} + \begin{bmatrix} E_0 & 0 \\ 0 & F_0 \end{bmatrix} \begin{bmatrix} I & -Y_0 \\ -X_0 & I \end{bmatrix}^{-1} \begin{bmatrix} E_0 & 0 \\ 0 & F_0 \end{bmatrix} \\ &\overset{0}{\preceq} P + P^3 \\ &\overset{0}{=} P + P \\ &\overset{0}{=} P \\ &\overset{0}{=} \begin{bmatrix} E_0 & Y_0 \\ X_0 & F_0 \end{bmatrix}, \end{aligned}$$

which implies (A.1) for $k = 0$.

Now suppose (A.1) holds for $k = \ell$. Notice

$$\begin{bmatrix} E_{\ell+1} & 0 \\ 0 & F_{\ell+1} \end{bmatrix} \overset{0}{\preceq} \begin{bmatrix} E_\ell & 0 \\ 0 & F_\ell \end{bmatrix}, \quad \begin{bmatrix} 0 & Y_{\ell+1} \\ X_{\ell+1} & 0 \end{bmatrix} \overset{0}{\preceq} \begin{bmatrix} 0 & Y_\ell \\ X_\ell & 0 \end{bmatrix},$$
$$\begin{bmatrix} I & -Y_{\ell+1} \\ -X_{\ell+1} & I \end{bmatrix}^{-1} \overset{0}{\preceq} \begin{bmatrix} I & -Y_\ell \\ -X_\ell & I \end{bmatrix}^{-1},$$

and in exactly the same way, we use

$$\begin{bmatrix} E_{\ell+2} & Y_{\ell+2} \\ X_{\ell+2} & F_{\ell+2} \end{bmatrix} = \begin{bmatrix} 0 & Y_{\ell+1} \\ X_{\ell+1} & 0 \end{bmatrix} + \begin{bmatrix} E_{\ell+1} & 0 \\ 0 & F_{\ell+1} \end{bmatrix} \begin{bmatrix} I & -Y_{\ell+1} \\ -X_{\ell+1} & I \end{bmatrix}^{-1} \begin{bmatrix} E_{\ell+1} & 0 \\ 0 & F_{\ell+1} \end{bmatrix},$$

and Lemma Appendix A.1 to prove (A.1) for $k = \ell + 1$. By mathematical induction, (A.1) holds for all $k \geq 0$. □

References

[1] G. Latouche, V. Ramaswami, Introduction to Matrix Analytic Methods in Stochastic Modeling, SIAM, Philadelphia, 1999.
[2] G.T. Nguyen, F. Poloni, Componentwise accurate fluid queue computations using doubling algorithms, Numer. Math. 130 (4) (2015) 763–792.
[3] J. Xue, R.-C. Li, Highly accurate doubling algorithms for M-matrix algebraic Riccati equations, Numer. Math. 135 (3) (2017) 733–767.
[4] J. Xue, S. Xu, R.-C. Li, Accurate solutions of M-matrix algebraic Riccati equations, Numer. Math. 120 (4) (2012) 671–700.
[5] G. Latouche, V. Ramaswami, A logarithmic reduction algorithm for quasi-birth-death processes, J. Appl. Probab. 30 (3) (1993) 650–674.
[6] Q. Ye, On Latouche-Ramaswami's logarithmic reduction algorithm for quasi-birth-and-death processes, Stoch. Models 18 (2002) 449–467.
[7] D.A. Bini, B. Meini, Improved cyclic reduction for solving queueing problems, Numer. Algorithms 15 (1997) 57–74.
[8] C.-Y. He, B. Meini, N.H. Rhee, A shifted cyclic reduction algorithm for quasi-birth-death problems, SIAM J. Matrix Anal. Appl. 23 (3) (2002) 679–691.
[9] B. Meini, Solving QBD problems: the cyclic reduction algorithm versus the invariant subspace method, Adv. Perform. Anal. 1 (1998) 215–225.
[10] G. Latouche, Newton's iteration for nonlinear equations in Markov chains, IMA J. Numer. Anal. 14 (1994) 583–598.
[11] P. Favati, B. Meini, Relaxed functional iteration techniques for the numerical solution of m/g/1 type Markov chains, BIT 38 (3) (1998) 510–526.
[12] C.-H. Guo, On the numerical solution of a nonlinear matrix equation in Markov chains, Linear Algebra Appl. 288 (1999) 175–186.
[13] S. Ahn, V. Ramaswami, Efficient algorithms for transient analysis of stochastic fluid flow models, J. Appl. Probab. 42 (2) (2005) 531–549.
[14] T.-M. Huang, R.-C. Li, W.-W. Lin, Structure-Preserving Doubling Algorithms for Nonlinear Matrix Equations, Fundamentals of Algorithms, vol. 14, SIAM, Philadelphia, 2018.
[15] C.-Y. Chiang, E.K.-W. Chu, C.-H. Guo, T.-M. Huang, W.-W. Lin, S.-F. Xu, Convergence analysis of the doubling algorithm for several nonlinear matrix equations in the critical case, SIAM J. Matrix Anal. Appl. 31 (2) (2009) 227–247.
[16] M.F. Neuts, Matrix Geometric Solutions in Stochastic Models: An Algorithmic Approach, The Johns Hopkins University Press, Baltimore, 1981.
[17] G.W. Stewart, J.-G. Sun, Matrix Perturbation Theory, Academic Press, Boston, 1990.
[18] A. Berman, 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.
[19] C.D. Meyer, Matrix Analysis and Applied Linear Algebra, SIAM, Philadelphia, 2000.
[20] A.S. Alfa, J. Xue, Q. Ye, Accurate computation of the smallest eigenvalue of a diagonally dominant M-matrix, Math. Comp. 71 (2002) 217–236.
[21] J. Xue, S. Xu, R.-C. Li, Accurate solutions of M-matrix Sylvester equations, Numer. Math. 120 (4) (2012) 639–670.
[22] A.S. Alfa, J. Xue, Q. Ye, Entrywise perturbation theory for diagonally dominant M-matrices with applications, Numer. Math. 90 (3) (2002) 401–414.
[23] W. Grassmann, M. Taksar, D. Heyman, Regenerative analysis and steady-state distributions for Markov chains, Oper. Res. 33 (1985) 1107–1116.
[24] J. Meng, S.-H. Seo, H.-M. Kim, Condition numbers and backward error of a matrix polynomial equation arising in stochastic models, J. Sci. Comput. 76 (2) (2018) 759–776.
[25] D.-A. Bini, G. Latouche, B. Meini, Numerical Methods for Structured Markov Chains, Oxford University Press, Oxford, 2005.
[26] I. Gohberg, P. Lancaster, L. Rodman, Matrix Polynomials, Academic Press, New York, 1982.
[27] H.R. Gail, S.L. Hantler, B.A. Taylor, Spectral analysis of M/G/1 and G/M/1 type Markov chains, Adv. Appl. Probab. 28 (1) (1996) 114–165.
[28] H.R. Gail, S.L. Hantler, B.A. Taylor, Matrix-geometric invariant measures for G/M/1 type Markov chains, Commun. Stat., Stoch. Models 14 (3) (1998) 537–569.
[29] J.F.C. Kingman, A convexity property of positive matrices, Q. J. Math. 12 (1) (1961) 283–284.

[30] A. Greenbaum, R.-C. Li, M.L. Overton, First-order perturbation theory for eigenvalues and eigenvectors, SIAM Rev. (2019), in press.
[31] R.-C. Li, On the variations of the spectra of matrix pencils, Linear Algebra Appl. 139 (1990) 147–164.
[32] G.W. Stewart, Error and perturbation bounds for subspaces associated with certain eigenvalue problems, SIAM Rev. 15 (1973) 727–764.
[33] X. Guo, W.-W. Lin, S. Xu, A structure-preserving doubling algorithm for nonsymmetric algebraic Riccati equation, Numer. Math. 103 (2006) 393–412.
[34] F.R. Gantmacher, The Theory of Matrices, vols. I, II, Chelsea Publishing Company, New York, 1959.
[35] American National Standards Institute and Institute of Electrical and Electronic Engineers, IEEE standard for binary floating-point arithmetic, ANSI/IEEE Standard, Std 754-1985, New York.
[36] D. Goldberg, What every computer scientist should know about floating-point arithmetic, ACM Comput. Surv. 23 (1) (1991) 5–47.
[37] W.-G. Wang, W.-C. Wang, R.-C. Li, Alternating-directional doubling algorithm for M-matrix algebraic Riccati equations, SIAM J. Matrix Anal. Appl. 33 (1) (2012) 170–194.
[38] C.-H. Guo, Convergence analysis of the Latouche-Ramaswami algorithm for null recurrent quasi-birth-death process, SIAM J. Matrix Anal. Appl. 23 (2002) 744–760.