IMA Journal of Numerical Analysis (2013) Page 1 of 19
doi:10.1093/imanum/drn000

Performance enhancement of doubling algorithms for a class of complex nonsymmetric algebraic Riccati equations

CHUN-HUA GUO†
Department of Mathematics and Statistics, University of Regina, Regina, SK S4S 0A2, Canada

CHANGLI LIU‡
School of Mathematical Science, Sichuan University, Chengdu 610065, P. R. China

AND

JUNGONG XUE§
School of Mathematical Science, Fudan University, Shanghai 200433, P. R. China

[Received on 9 December 2013]

A new class of complex nonsymmetric algebraic Riccati equations has been studied by Liu & Xue (2012, *SIAM J. Matrix Anal. Appl.*, **33**, 569–596), which is related to the M-matrix algebraic Riccati equations. Doubling algorithms, with properly chosen parameters, are used there for equations in this new class. It is pointed out that the number of iterations for the doubling algorithms may be relatively large in some situations. In this paper, we show that the performance of the doubling algorithms can often be improved significantly if a proper preprocessing procedure is used on the given Riccati equation. There are some difficult cases for which the preprocessing procedure does not help much by itself. We then propose new strategies for choosing parameters for doubling algorithms after using the preprocessing procedure. Numerical experiments show that our preprocessing procedure and the new parameter strategies are very effective.

Keywords: algebraic Riccati equation; M-matrix; doubling algorithm.

1. Introduction

We use $[A]_{ij}$ to denote the (i,j) entry of a matrix A. For $A,B\in\mathbb{R}^{m\times n}$, we write $A\geqslant B\,(A>B)$ if $[A]_{ij}\geqslant[B]_{ij}\,([A]_{ij}>[B]_{ij})$ for all i,j. A real square matrix A is called a nonsingular M-matrix if all its off-diagonal entries are nonpositive and $Av>0$ for some vector $v>0$. For a complex number z, its real part, imaginary part and modulus will be denoted by $\mathrm{Re}(z),\mathrm{Im}(z)$ and $|z|$, respectively. For $A\in\mathbb{C}^{n\times n}$, we use $\rho(A)$ to denote its spectral radius and use $\mathrm{diag}(A)$ to denote its diagonal part. The $n\times n$ identity matrix is denoted by I_n or simply I. For $A\in\mathbb{C}^{m\times n}$, its absolute value $|A|\in\mathbb{R}^{m\times n}$ is defined by $[|A|]_{ij}=|[A]_{ij}|$. For $A\in\mathbb{C}^{n\times n}$, the comparison matrix of A, denoted by $\widehat{A}$, is defined in (Liu & Xue, 2012) by

$$[\widehat{A}]_{ij}=\begin{cases}\mathrm{Re}([A]_{ii}), & i=j,\\ -|[A]_{ij}|, & i\neq j.\end{cases}$$

†Email: chun-hua.guo@uregina.ca
‡Email: chliliu@hotmail.com
§Email: xuej@fudan.edu.cn

We consider the nonsymmetric algebraic Riccati equation (NARE)

$$XCX - XD - AX + B = 0, \tag{1.1}$$

where A, B, C, D are complex matrices of sizes $m \times m, m \times n, n \times m, n \times n$, respectively. Associated with the NARE (1.1) is the matrix

$$Q = \begin{bmatrix} D & -C \\ -B & A \end{bmatrix}. \tag{1.2}$$

The NARE (1.1) is said to be in class H^* (see (Liu & Xue, 2012)) if the comparison matrix $\widehat{Q}$ is a nonsingular M-matrix. The class H^* is an extension of the class H^+ studied earlier in (Guo, 2007), where the diagonal entries of Q are required to be positive. The NARE in class H^* arises in the study of Markov modulated fluid flows; see (Liu & Xue, 2012) and the references therein.

The study of the NARE in class H^+ or H^* is through comparison with a NARE (1.1) for which the matrix Q in (1.2) is a nonsingular M-matrix. The later is said to be in class M or called an M-matrix algebraic Riccati equation, which has been studied extensively (see (Bini *et al.*, 2012), (Guo, 2001) and (Guo & Higham, 2007), for example). Any NARE in class M has a minimal nonnegative solution. In the study of the NARE (1.1) in class H^*, we may assume without loss of generality (Liu & Xue, 2012) that $\widehat{Q}\mathbf{1} > 0$, where $\mathbf{1}$ is the vector of ones. We have the following result of (Liu & Xue, 2012), which is a useful generalization of (Guo, 2007, Theorem 8).

THEOREM 1.1 Suppose the NARE (1.1) is in class H^* and $\widehat{Q}\mathbf{1} > 0$. Let $\widetilde{\Phi}$ be the minimal nonnegative solution of the NARE

$$X\widetilde{C}X - X\widetilde{D} - \widetilde{A}X + \widetilde{B} = 0, \tag{1.3}$$

where

$$\widetilde{Q} = \begin{bmatrix} \widetilde{D} & -\widetilde{C} \\ -\widetilde{B} & \widetilde{A} \end{bmatrix}, \tag{1.4}$$

partitioned as for Q in (1.2), is a nonsingular M-matrix satisfying $\widetilde{Q} \leqslant \widehat{Q}$ and $\widetilde{Q}\mathbf{1} > 0$. Then the NARE (1.1) has a unique solution Φ such that $|\Phi| \leqslant \widetilde{\Phi}$. Similarly, the dual equation of (1.1)

$$YBY - YA - DY + C = 0 \tag{1.5}$$

has a unique solution Ψ such that $|\Psi| \leqslant \widetilde{\Psi}$, where $\widetilde{\Psi}$ is the minimal nonnegative solution of the NARE

$$Y\widetilde{B}Y - Y\widetilde{A} - \widetilde{D}Y + \widetilde{C} = 0. \tag{1.6}$$

In (Liu & Xue, 2012) it is shown that the special solutions Φ and Ψ in Theorem 1.1 are the solutions required in applications and that these two solutions can be found simultaneously by existing doubling algorithms if the parameters in the doubling algorithms are chosen properly. In this paper, we show that the performance of the doubling algorithms can often be improved significantly if a proper preprocessing procedure is used on the given Riccati equation. We also propose new strategies for choosing parameters for doubling algorithms. For some difficult cases, these strategies can provide significant further improvement after using the preprocessing procedure.

2. Doubling algorithms

Three existing doubling algorithms are used in (Liu & Xue, 2012) to find the solutions Φ and Ψ. They are SDA of (Guo *et al.*, 2006), SDA-ss of (Bini *et al.*, 2010), and ADDA of (Wang *et al.*, 2012). SDA

and SDA-ss involve one parameter. ADDA involves two parameters and is reduced to SDA when the two parameters are equal. Experiments in (Liu & Xue, 2012) show that ADDA is consistently faster than SDA and SDA-ss for the NAREs in class H^*, just like for NAREs in class M (see the theory and experiments in (Wang *et al.*, 2012)), when good parameters are used for each of these algorithms. In this paper, we therefore consider ADDA, which includes SDA as a special case.

The details of ADDA can be found in (Wang *et al.*, 2012). The algorithm is also briefly reviewed in (Liu & Xue, 2012). Here we give a very brief presentation of ADDA, using the notation in (Liu & Xue, 2012).

Choose parameters α and β such that $A+\beta I$ and $D+\alpha I$ are nonsingular and set

$$A_\beta = A+\beta I, \quad D_\alpha = D+\alpha I, \tag{2.1}$$

$$W_{\alpha\beta} = A_\beta - BD_\alpha^{-1}C, \quad V_{\alpha\beta} = D_\alpha - CA_\beta^{-1}B. \tag{2.2}$$

Further suppose that $W_{\alpha\beta}$ and $V_{\alpha\beta}$ are also nonsingular, and let

$$E_0 = I-(\alpha+\beta)V_{\alpha\beta}^{-1}, \quad F_0 = I-(\alpha+\beta)W_{\alpha\beta}^{-1}, \tag{2.3}$$

$$G_0 = (\alpha+\beta)D_\alpha^{-1}CW_{\alpha\beta}^{-1}, \quad H_0 = (\alpha+\beta)W_{\alpha\beta}^{-1}BD_\alpha^{-1}. \tag{2.4}$$

The ADDA generates, as for SDA, the sequences $\{E_k\}$, $\{F_k\}$, $\{G_k\}$, $\{H_k\}$ using the iteration:

$$\begin{aligned} E_{k+1} &= E_k(I-G_kH_k)^{-1}E_k, \\ F_{k+1} &= F_k(I-H_kG_k)^{-1}F_k, \\ G_{k+1} &= G_k+E_k(I-G_kH_k)^{-1}G_kF_k, \\ H_{k+1} &= H_k+F_k(I-H_kG_k)^{-1}H_kE_k, \end{aligned} \tag{2.5}$$

assuming that the matrices $I-G_kH_k$ and $I-H_kG_k$ are all nonsingular. Provided that the sequences $\{H_k\}$ and $\{G_k\}$ are bounded, we have

$$\limsup_{k\to\infty} \|\Phi-H_k\|^{1/2^k} \leqslant \rho(\mathscr{R})\rho(\mathscr{S}), \quad \limsup_{k\to\infty} \|\Psi-G_k\|^{1/2^k} \leqslant \rho(\mathscr{R})\rho(\mathscr{S}), \tag{2.6}$$

for any matrix norm $\|\cdot\|$, where

$$\mathscr{R} = (R-\beta I)(R+\alpha I)^{-1}, \quad \mathscr{S} = (S-\alpha I)(S+\beta I)^{-1} \tag{2.7}$$

with

$$R = D-C\Phi, \quad S = A-B\Psi. \tag{2.8}$$

It follows that the sequences $\{H_k\}$ and $\{G_k\}$ converge quadratically to Φ and Ψ, respectively, when $\rho(\mathscr{R})\rho(\mathscr{S})<1$.

For the NARE (1.1) in class H^* with $\widehat{Q}\mathbf{1}>0$, the following two positive constants are introduced in (Liu & Xue, 2012):

$$\gamma_1 = \max_{1\leqslant i\leqslant n} p_i, \quad \gamma_2 = \max_{n+1\leqslant i\leqslant n+m} p_i, \tag{2.9}$$

where for $1\leqslant i\leqslant m+n$

$$p_i = \frac{\mathrm{Re}([Q]_{ii})+q_i}{2} + \frac{(\mathrm{Im}([Q]_{ii}))^2}{2(\mathrm{Re}([Q]_{ii})-q_i)} \tag{2.10}$$

with

$$q_i = \sum_{j \neq i} |[Q]_{ij}|. \tag{2.11}$$

The following result has been proved in (Liu & Xue, 2012).

THEOREM 2.1 Suppose the NARE (1.1) is in class H^* and $\widehat{Q}\mathbf{1} > 0$. Apply ADDA to the NARE (1.1) with parameters $\alpha > \gamma_2$ and $\beta > \gamma_1$. Then the sequences $\{E_k\}, \{F_k\}, \{G_k\}$, and $\{H_k\}$ are well-defined and $\{H_k\}$ and $\{G_k\}$ converge quadratically to Φ and Ψ, respectively.

Note that SDA is a special case of ADDA with $\alpha = \beta$. So we would require $\alpha > \max\{\gamma_1, \gamma_2\}$ to guarantee the convergence of the SDA. It is already mentioned in Remark 6.2 of (Liu & Xue, 2012) that the convergence of SDA could be slow when $(\mathrm{Im}([Q]_{ii}))^2$ is large compared to $\mathrm{Re}([Q]_{ii}) - q_i$ for some $i \in [1, m+n]$. Similar comment can be made about ADDA. More precisely, the convergence of ADDA could be slow if p_i is large for some $i \in [1, n]$ and for some $i \in [n+1, n+m]$.

In the next section, we show that the performance of the doubling algorithms can be improved significantly if a proper preprocessing procedure is used on the given Riccati equation.

3. A preprocessing procedure

We first rewrite p_i in (2.10) in a more compact form:

$$p_i = \frac{|[Q]_{ii}|^2 - q_i^2}{2(\mathrm{Re}([Q]_{ii}) - q_i)}.$$

For a given NARE, $\max_{1 \leqslant i \leqslant m+n} p_i$ may be very large. But it is quite possible that we can transform the given NARE to a new NARE for which $\max_{1 \leqslant i \leqslant m+n} p_i$ is much smaller and the solution sets of the two NAREs are related in a simple way.

The idea is very simple. Roughly speaking, we just need to transform a given NARE in class H^* to a new NARE in class H^* that resembles a NARE in class H^+ as much as possible.

For the given NARE (1.1) in class H^*, with $\widehat{Q}\mathbf{1} > 0$, we consider the new NARE

$$X(\omega C)X - X(\omega D) - (\omega A)X + \omega B = 0, \tag{3.1}$$

where ω is on the unit circle. Obviously, the new equation has the same solution set as the orginal one. The matrix corresponding to the new equation is

$$Q_\omega = \begin{bmatrix} \omega D & -\omega C \\ -\omega B & \omega A \end{bmatrix}. \tag{3.2}$$

Note that the comparison matrix of Q_ω differs from that of Q only on the main diagonal. We now have $[\widehat{Q}_\omega]_{ii} = \mathrm{Re}(\omega[Q]_{ii})$. Corresponding to the matrix Q_ω, we have the quantities

$$p_{i,\omega} = \frac{|\omega[Q]_{ii}|^2 - q_i^2}{2(\mathrm{Re}(\omega[Q]_{ii}) - q_i)} = \frac{|[Q]_{ii}|^2 - q_i^2}{2(\mathrm{Re}(\omega[Q]_{ii}) - q_i)}, \tag{3.3}$$

where ω is such that $\mathrm{Re}(\omega[Q]_{ii}) - q_i > 0$ (so the equation (3.1) is still in class H^*). Note that $|[Q]_{ii}|^2 - q_i^2 \geqslant (\mathrm{Re}([Q]_{ii}))^2 - q_i^2 > 0$ since $\widehat{Q}\mathbf{1} > 0$. Thus, for each fixed i, the positive quantity $p_{i,\omega}$ is minimized when $\omega = \frac{|[Q]_{ii}|}{[Q]_{ii}}$. If all complex numbers $[Q]_{ii}$ are on the same line passing throught the origin, then a

common value $\omega = e^{-\mathrm{i}\phi}$ will make all $\omega[Q]_{ii}$ on the positive real axis. In other words, the new NARE (3.1) is in class H^+ with this ω. In general, we cannot find a fixed $\omega = e^{-\mathrm{i}\phi}$ to minimize $p_{i,\omega}$ for all i. So we let

$$f_i(\phi) = \frac{|[Q]_{ii}|^2 - q_i^2}{\mathrm{Re}(e^{-\mathrm{i}\phi}[Q]_{ii}) - q_i}, \quad \phi \in \mathbb{R},$$

and try to find ϕ such that

$$f(\phi) = \max_{1 \leqslant i \leqslant m+n} f_i(\phi)$$

is minimized, subject to the condition that $\mathrm{Re}(e^{-\mathrm{i}\phi}[Q]_{ii}) - q_i > 0$ for all i.

THEOREM 3.1 Suppose the NARE (1.1) is class H^* and $\widehat{Q}\mathbf{1} > 0$. Then the function $f(\phi)$ has a unique minimizer $\phi_* \in (-\pi, \pi)$ and ϕ_* can be computed by a bisection procedure.

Proof. The existence of a minimizer is quite obvious. Let $d = \max_{1 \leqslant i \leqslant m+n} f_i(0) > 0$. For each i let Δ_i be the set of all values $\phi \in (-\pi, \pi)$ such that $0 < f_i(\phi) \leqslant d$. It is clear that Δ_i is a closed interval containing 0. Let $\Delta = \bigcap_{1 \leqslant i \leqslant m+n} \Delta_i$. Then Δ is also a closed interval containing 0. Now $\min f(\phi) = \min_{\phi \in \Delta} f(\phi)$ is attained at some $\phi_* \in \Delta$ since f is continuous on Δ.

Next we describe a (somewhat unusual) bisection procedure that determines a unique minimizer. The procedure is based on the following simple observation:

Let θ_i be the angles (arguments) of the complex numbers $[Q]_{ii}$. Note that $-\frac{\pi}{2} < \theta_i < \frac{\pi}{2}$ and that $f_i(\phi)$ is minimized at $\phi = \theta_i$. Moreover, $f_i(\phi)$ is strictly decreasing on the left of θ_i and is strictly increasing on the right of θ_i.

Based on this observation, the interval Δ above can be given explicitly as follows. Let $\underline{\delta}_i \leqslant \overline{\delta}_i$ be the two (usually different) solutions of $f_i(\phi) = d$. Namely, $\underline{\delta}_i = \theta_i - \psi_i$ and $\overline{\delta}_i = \theta_i + \psi_i$ with

$$\psi_i = \arccos\left(\frac{1}{|[Q]_{ii}|}\left(q_i + \frac{|[Q]_{ii}|^2 - q_i^2}{d}\right)\right), \quad 0 \leqslant \psi_i < \frac{\pi}{2}.$$

Now, $\Delta_i = [\underline{\delta}_i, \overline{\delta}_i]$ and $\Delta = [\max \underline{\delta}_i, \min \overline{\delta}_i]$. The interval Δ may be large even when all θ_i are equal to θ_*, in which case we know that θ_* is the unique minimizer of $f(\phi)$. To avoid using an unnecessarily large search interval in situations like this, we let $\theta_{min} = \min \theta_i$ and $\theta_{max} = \max \theta_i$ and we claim that any minimizer of $f(\phi)$ must be in $[\theta_{min}, \theta_{max}]$. In fact,

$$f_i(\phi) = \frac{|[Q]_{ii}|^2 - q_i^2}{\mathrm{Re}(e^{\mathrm{i}(\theta_i - \phi)}|[Q]_{ii}|) - q_i} = \frac{|[Q]_{ii}|^2 - q_i^2}{\cos(\theta_i - \phi)|[Q]_{ii}| - q_i}.$$

For any $\phi \in (-\pi, \theta_{min})$ such that $f_i(\phi) > 0$ for each i, we have $0 < \theta_i - \phi < \frac{\pi}{2}$ and $f_i(\phi) > f_i(\theta_{min}) > 0$ for each i. So any minimizer ϕ_* must satisfy $\phi_* \geqslant \theta_{min}$. Similarly we can show that $\phi_* \leqslant \theta_{max}$.

The initial search interval for a minimizer of $f(\phi)$ is then $[\phi_{min}, \phi_{max}]$, where

$$\phi_{min} = \max\{\max \underline{\delta}_i, \theta_{min}\}, \quad \phi_{max} = \min\{\min \overline{\delta}_i, \theta_{max}\}.$$

The first step of the bisection procedure is to take $\phi_1 = \frac{1}{2}(\phi_{min} + \phi_{max})$. Let

$$a_1 = \max_{\theta_i > \phi_1} f_i(\phi_1), \quad b_1 = \max_{\theta_i < \phi_1} f_i(\phi_1), \quad c_1 = \max_{\theta_i = \phi_1} f_i(\phi_1).$$

where the maximum over an empty set is defined to be 0. If $c_1 \geqslant \max\{a_1, b_1\}$, then ϕ_1 is a unique minimizer. Suppose $c_1 < \max\{a_1, b_1\}$. If $a_1 = b_1$ then ϕ_1 is still a unique minimizer. If $a_1 \neq b_1$, any

minimizer must be in $[\phi_1, \phi_{max}]$ if $a_1 > b_1$ and must be in $[\phi_{min}, \phi_1]$ if $a_1 < b_1$. So for the second step of bisection we take

$$\phi_2 = \begin{cases} \frac{1}{2}(\phi_1 + \phi_{max}), & \text{if } a_1 > b_1, \\ \frac{1}{2}(\phi_1 + \phi_{min}), & \text{if } a_1 < b_1. \end{cases}$$

Let

$$a_2 = \max_{\theta_i > \phi_2} f_i(\phi_2), \quad b_2 = \max_{\theta_i < \phi_2} f_i(\phi_2), \quad c_2 = \max_{\theta_i = \phi_2} f_i(\phi_2).$$

As before we can determine whether ϕ_2 is a unique minimizer. If not, for the case $a_1 > b_1$ any minimizer must be in $[\phi_2, \phi_{max}]$ if $a_2 > b_2$ and must be in $[\phi_1, \phi_2]$ if $a_2 < b_2$; for the case $a_1 < b_1$ any minimizer must be in $[\phi_2, \phi_1]$ if $a_2 > b_2$ and must be in $[\phi_{min}, \phi_2]$ if $a_2 < b_2$. We can then continue with the bisection procedure. Unless a unique minimizer is found in a finite number of steps, we get a sequence $\{\phi_i\}$ with $\lim_{i\to\infty} \phi_i = \phi_*$ and $|\phi_i - \phi_*| \leqslant (\phi_{max} - \phi_{min})/2^i < \pi/2^i$. By construction, this ϕ_* is the only candidate for the minimizer. So it is the unique minimizer since the existence is already known. □

In step k of the above bisection procedure, the θ_i are divided into three piles: one with $\theta_i > \phi_k$, one with $\theta_i < \phi_k$, and the other with $\theta_i = \phi_k$. We have assumed that this division is done in exact arithmetic. In practice this division is done by a computer and may be different from the division in exact arithmetic when some θ_i are extremely close to ϕ_k. But this will have very little effect on the accuracy of the computed ϕ_*. The situation here is similar to that for the usual bisection method for finding a root of a continuous function $g(x)$ on an interval $[a,b]$, where $g(a)g(b) < 0$. In the first step of the usual bisection method we compute $p = (a+b)/2$ and let the computer decide whether $g(p) > 0$, $g(p) < 0$, or $g(p) = 0$.

When $m = n$ for example, our bisection procedure requires $O(n)$ operations each step, while ADDA (or any other doubling algorithm) requires $O(n^3)$ operations each iteration. We have already seen in the proof of Theorem 3.1 that the ith approximation ϕ_i to the minimizer ϕ_* satisfies $|\phi_i - \phi_*| < \pi/2^i$. So when n is large, the computational work for using the bisection procedure to approximate ϕ_* to machine precision is negligible compared to the work for one ADDA iteration. In practice, there is no need to compute ϕ_* so accurately. If the ith approximation is obtained by $\phi_i = \frac{1}{2}(\phi_a + \phi_b)$, we will stop the bisection if we already have $|\phi_b - \phi_a| < \tau$. We will take $\tau = 10^{-6}$ in our numerical experiments, although $\tau = 10^{-2}$ is probably already small enough. The computational work in either case is negligible.

A simple preprocessing procedure for the NARE (1.1) is then as follows: Use the bisection method described in the proof of Theorem 3.1 to determine a good approximation $\tilde{\phi}$ to ϕ_*, let $\omega = e^{-\mathrm{i}\tilde{\phi}}$ and transform the NARE (1.1) to the NARE (3.1).

Let

$$\widetilde{\gamma}_1 = \max_{1\leqslant i\leqslant n} p_{i,\omega}, \quad \widetilde{\gamma}_2 = \max_{n+1\leqslant i\leqslant n+m} p_{i,\omega},$$

where $p_{i,\omega}$ are as in (3.3). Then we can apply ADDA to (3.1), with $\alpha > \widetilde{\gamma}_2$ and $\beta > \widetilde{\gamma}_1$. We often have the situation that the two largest values of $f_i(\tilde{\phi})$, say $f_{i_1}(\tilde{\phi})$ and $f_{i_2}(\tilde{\phi})$ with $i_1 < i_2$, are very close. If it happens that $i_1 \in [1,n]$ and $i_2 \in [n+1, n+m]$, then $\widetilde{\gamma}_1 \approx \widetilde{\gamma}_2$ and we may take $\alpha = \beta$ for ADDA. In this case, ADDA is reduced to SDA.

While the solution sets for the equations (1.1) and (3.1) are the same, there are many solutions in the set. We still need to make sure that the required solutions Φ and Ψ are obtained when ADDA is applied to the transformed equation (3.1).

THEOREM 3.2 Suppose the NARE (1.1) is in class H^* and $\widehat{Q}\mathbf{1} > 0$. Let ω be any unimodular number such that $\mathrm{Re}(\omega[Q]_{ii}) - q_i > 0$ for all $i \in [1, n+m]$ ($\omega = e^{-\mathrm{i}\tilde{\phi}}$ in particular). Let $\{E_k\}, \{F_k\}, \{G_k\}$, and $\{H_k\}$ be generated by ADDA applied to the NARE (3.1) with parameters

$$\alpha > \max_{n+1\leqslant i\leqslant n+m} p_{i,\omega}, \quad \beta > \max_{1\leqslant i\leqslant n} p_{i,\omega}.$$

Then $\{H_k\}$ and $\{G_k\}$ converge quadratically to Φ and Ψ, respectively, where Φ and Ψ are the same as in Theorem 1.1.

Proof. Note that the comparison matrix $\widehat{Q}_\omega$ of Q_ω is such that $\widehat{Q}_\omega \mathbf{1} > 0$. Let $\widetilde{Q} = \min\{\widehat{Q}, \widehat{Q}_\omega\}$, where the minimum is taken entrywise. Since $\widehat{Q}$ and $\widehat{Q}_\omega$ have the same offdiagonal entries, we have $\widetilde{Q}\mathbf{1} > 0$. By applying Theorem 1.1 to the NARE (1.1), we know that the NARE (1.1) has a unique solution Φ with $|\Phi| \leqslant \widetilde{\Phi}$. By applying Theorem 1.1 to the NARE (3.1), we know that the NARE (3.1) has a unique solution Φ_ω with $|\Phi_\omega| \leqslant \widetilde{\Phi}$. Since (1.1) and (3.1) have the same solution set, we have $\Phi_\omega = \Phi$. We also have similar conclusions for the dual equations of (1.1) and (3.1). The results in the theorem are then obtained by applying Theorem 2.1 to (3.1). □

Our preprocessing procedure in this section should also be used if one chooses to use SDA or SDA-ss.

4. Further improvement of doubling algorithms after preprocessing

We now assume that the equation (1.1) in class H^* (with $\widehat{Q}\mathbf{1} > 0$) has already gone through the preprocessing procedure in the previous section. So the equation is in some sense as close as possible to an equation in class H^+. At the moment we need to take $\alpha > \gamma_2$ and $\beta > \gamma_1$ for ADDA (see Theorem 2.1). In some adverse situations, these parameters may still be very large, leading to slower convergence of the doubling algorithm. To address this problem, we need to expand the convergence region of parameters α and β for ADDA.

The general idea for achieving this goal is to compare (as in (Guo, 2007) and (Liu & Xue, 2012)) the sequences $\{E_k\}$, $\{F_k\}$, $\{G_k\}$, $\{H_k\}$ from ADDA with their counterparts generated by ADDA applied to a NARE in class M. To determine a suitable NARE in class M, we construct $\widetilde{Q}$ with entries given by

$$[\widetilde{Q}]_{ij} = \begin{cases} q_i + \varepsilon, & i = j, \\ -|[Q]_{ij}|, & i \neq j, \end{cases} \tag{4.1}$$

where $\varepsilon > 0$ and q_i is defined as in (2.11). Then $\widetilde{Q}$ is a nonsingular M-matrix with $\widetilde{Q}\mathbf{1} = \varepsilon\mathbf{1} > 0$. We take ε to be sufficiently small so that $\widetilde{Q} \leqslant \widehat{Q}$. Let $\widetilde{Q}$ be partitioned in the form of (1.4). Let $\widetilde{\Phi}$ and $\widetilde{\Psi}$ be the minimal nonnegative solutions to the NARE (1.3) and its dual NARE (1.6), which are in class M, respectively. By Theorem 1.1,

$$|\Phi| \leqslant \widetilde{\Phi}, \quad |\Psi| \leqslant \widetilde{\Psi}.$$

Let $\{\widetilde{E}_k\}$, $\{\widetilde{F}_k\}$, $\{\widetilde{G}_k\}$, $\{\widetilde{H}_k\}$ be generated by ADDA with parameters $\widetilde{\alpha}$ and $\widetilde{\beta}$ applied to the NARE (1.3). The following result is shown in (Wang *et al.*, 2012).

THEOREM 4.1 If $\widetilde{\alpha} \geqslant \max_{n+1\leqslant i\leqslant m+n}[\widetilde{Q}]_{ii}$ and $\widetilde{\beta} \geqslant \max_{1\leqslant i\leqslant n}[\widetilde{Q}]_{ii}$, then

1. $\widetilde{E}_0, \widetilde{F}_0 \leqslant 0$ and $\widetilde{E}_k, \widetilde{F}_k \geqslant 0$ for $k \geqslant 1$;
2. $I - \widetilde{G}_k\widetilde{H}_k$ and $I - \widetilde{H}_k\widetilde{G}_k$ are nonsingular M-matrices for $k \geqslant 0$;
3. $0 \leqslant \widetilde{H}_k \leqslant \widetilde{H}_{k+1} \leqslant \widetilde{\Phi}$, $0 \leqslant \widetilde{G}_k \leqslant \widetilde{G}_{k+1} \leqslant \widetilde{\Psi}$ for $k \geqslant 0$ and

$$\limsup_{k\to\infty} \|\widetilde{\Phi} - \widetilde{H}_k\|^{1/2^k} \leqslant \rho(\widetilde{\mathscr{R}})\rho(\widetilde{\mathscr{S}}) < 1, \quad \limsup_{k\to\infty} \|\widetilde{\Psi} - \widetilde{G}_k\|^{1/2^k} \leqslant \rho(\widetilde{\mathscr{R}})\rho(\widetilde{\mathscr{S}}) < 1, \tag{4.2}$$

where

$$\widetilde{\mathscr{R}} = (\widetilde{R} - \widetilde{\beta}I)(\widetilde{R} + \widetilde{\alpha}I)^{-1}, \quad \widetilde{\mathscr{S}} = (\widetilde{S} - \widetilde{\alpha}I)(\widetilde{S} + \widetilde{\beta}I)^{-1}. \tag{4.3}$$

with

$$\widetilde{R} = \widetilde{D} - \widetilde{C}\widetilde{\Phi}, \quad \widetilde{S} = \widetilde{A} - \widetilde{B}\widetilde{\Psi}. \tag{4.4}$$

We will also need the following result.

LEMMA 4.1 Let $A \in \mathbb{R}^{n\times n}$ be a nonsingular M-matrix and be written as $A = D_1 - N_1$, where D_1 is diagonal with positive diagonal entries and $N_1 \geqslant 0$. Let $B \in \mathbb{C}^{n\times n}$ be written as $B = D_2 - N_2$, where D_2 is diagonal. If

$$D_1 \leqslant |D_2|, \quad |N_2| \leqslant N_1,$$

then B is nonsingular and

$$\left|B^{-1}\right| \leqslant A^{-1}C, \quad \left|B^{-1}\right| \leqslant CA^{-1},$$

where $C = \left|D_1 D_2^{-1}\right|$.

Proof. Since $\left|D_2^{-1}N_2\right| \leqslant D_1^{-1}N_1$ and A is a nonsingular M-matrix, $\rho\left(D_2^{-1}N_2\right) \leqslant \rho\left(D_1^{-1}N_1\right) < 1$ (see (Berman & Plemmons, 1994)). Then $B = D_2(I - D_2^{-1}N_2)$ is nonsingular and

$$\begin{aligned}
|B^{-1}| &= \left|\left(I - D_2^{-1}N_2\right)^{-1} D_2^{-1}\right| \\
&\leqslant \sum_{k=0}^{\infty} \left|D_2^{-1}N_2\right|^k \left|D_2^{-1}\right| \\
&\leqslant \sum_{k=0}^{\infty} \left(D_1^{-1}N_1\right)^k D_1^{-1} \left|D_1 D_2^{-1}\right| \\
&= A^{-1}C.
\end{aligned}$$

Similarly, $\left|B^{-1}\right| \leqslant CA^{-1}$. □

The next result gives a larger convergence region of parameters α and β for ADDA. Its proof is a refinement of the proof of Theorem 6.2 in (Liu & Xue, 2012). The above lemma plays an important role here.

THEOREM 4.2 Suppose the NARE (1.1) is in class H^* and $\widehat{Q}\mathbf{1} > 0$. Let Φ and Ψ be as in Theorem 1.1. Apply ADDA to the NARE (1.1) with parameters $\alpha, \beta > 0$. If α and β satisfy

$$\frac{\alpha + q_i}{|\alpha + [Q]_{ii}|}|\beta - [Q]_{ii}| < \beta - q_i, \quad 1 \leqslant i \leqslant n, \tag{4.5a}$$

$$\frac{\beta + q_j}{|\beta + [Q]_{jj}|}|\alpha - [Q]_{jj}| < \alpha - q_j, \quad n+1 \leqslant j \leqslant n+m. \tag{4.5b}$$

then the sequences $\{E_k\}$, $\{F_k\}$, $\{H_k\}$, $\{G_k\}$ are well-defined and $\{H_k\}$ and $\{G_k\}$ converge quadratically to Φ and Ψ, respectively.

Proof. Let $\widetilde{Q}$ be defined by (4.1), and let $\{\widetilde{E}_k\}$, $\{\widetilde{F}_k\}$, $\{\widetilde{G}_k\}$, $\{\widetilde{H}_k\}$ be generated by ADDA applied to the NARE (1.3) with parameters $\widetilde{\alpha} = \alpha$ and $\widetilde{\beta} = \beta$. Since $\widehat{Q}\mathbf{1} > 0$, we can take $\varepsilon > 0$ sufficiently small such that

$$\mathrm{Re}([Q]_{ii}) \geqslant q_i + \varepsilon, \quad 1 \leqslant i \leqslant m+n.$$

Furthermore, we assume that ε is small enough such that the inequalities in (4.5) strictly hold with q_i and q_j replaced by $q_i + \varepsilon$ and $q_j + \varepsilon$, respectively. Set

$$\widetilde{A}_{\widetilde{\beta}} = \widetilde{A} + \widetilde{\beta}I, \quad \widetilde{D}_{\widetilde{\alpha}} = \widetilde{D} + \widetilde{\alpha}I,$$

$$\widetilde{W}_{\widetilde{\alpha}\widetilde{\beta}} = \widetilde{A}_{\widetilde{\beta}} - \widetilde{B}\widetilde{D}_{\widetilde{\alpha}}^{-1}\widetilde{C}, \quad \widetilde{V}_{\widetilde{\alpha}\widetilde{\beta}} = \widetilde{D}_{\widetilde{\alpha}} - \widetilde{C}\widetilde{A}_{\widetilde{\beta}}^{-1}\widetilde{B}.$$

Then,

$$\widetilde{E}_0 = I - (\widetilde{\alpha}+\widetilde{\beta})\widetilde{V}_{\widetilde{\alpha}\widetilde{\beta}}^{-1}, \quad \widetilde{F}_0 = I - (\widetilde{\alpha}+\widetilde{\beta})\widetilde{W}_{\widetilde{\alpha}\widetilde{\beta}}^{-1},$$
$$\widetilde{G}_0 = (\widetilde{\alpha}+\widetilde{\beta})\widetilde{D}_{\widetilde{\alpha}}^{-1}\widetilde{C}\widetilde{W}_{\widetilde{\alpha}\widetilde{\beta}}^{-1}, \quad \widetilde{H}_0 = (\widetilde{\alpha}+\widetilde{\beta})\widetilde{W}_{\widetilde{\alpha}\widetilde{\beta}}^{-1}\widetilde{B}\widetilde{D}_{\widetilde{\alpha}}^{-1}.$$

As $\widetilde{\alpha} \geqslant \max_{n+1\leqslant j\leqslant n+m}[\widetilde{Q}]_{ii}$, $\widetilde{\beta} \geqslant \max_{1\leqslant i\leqslant n}[\widetilde{Q}]_{ii}$, Theorem 4.1 applies. We need to show that ADDA is well defined when applied to the NARE (1.1) and that the sequences $\{H_k\}$ and $\{G_k\}$ are bounded. It is already explained in (Liu & Xue, 2012) that it is enough to show

$$|E_0| \leqslant |\widetilde{E}_0|, \quad |F_0| \leqslant |\widetilde{F}_0|, \quad |G_0| \leqslant |\widetilde{G}_0|, \quad |H_0| \leqslant |\widetilde{H}_0|, \tag{4.6}$$

where $E_0,\ F_0,\ G_0,\ H_0$ are as in (2.3) and (2.4). Let

$$N = \operatorname{diag}(\widetilde{D}_{\widetilde{\alpha}})\,|\operatorname{diag}(D_\alpha)|^{-1}, \quad M = \operatorname{diag}(\widetilde{A}_{\widetilde{\beta}})\,\big|\operatorname{diag}(A_\beta)\big|^{-1},$$

where $A_\beta,\ D_\alpha$ are as in (2.1). Then $0 \leqslant N \leqslant I_n,\ 0 \leqslant M \leqslant I_m$, and the inequalities in (4.5) can be written as

$$N|\operatorname{diag}(\beta I - D)| \leqslant \operatorname{diag}(\widetilde{\beta}I - \widetilde{D}), \quad M|\operatorname{diag}(\alpha I - A)| \leqslant \operatorname{diag}(\widetilde{\alpha}I - \widetilde{A}). \tag{4.7}$$

Because $\widetilde{A}_{\widetilde{\beta}}$, $\widetilde{D}_{\widetilde{\alpha}}$, $\widetilde{W}_{\widetilde{\alpha}\widetilde{\beta}}$, and $\widetilde{V}_{\widetilde{\alpha}\widetilde{\beta}}$ are nonsingular M-matrices (see (Wang *et al.*, 2012) for example), it follows from Lemma 4.1 that

$$|A_\beta^{-1}| \leqslant \widetilde{A}_{\widetilde{\beta}}^{-1}M \leqslant \widetilde{A}_{\widetilde{\beta}}^{-1}, \quad |D_\alpha^{-1}| \leqslant \widetilde{D}_{\widetilde{\alpha}}^{-1}N \leqslant \widetilde{D}_{\widetilde{\alpha}}^{-1},$$

and thus

$$\begin{aligned}
|W_{\alpha\beta}^{-1}| &= |(A_\beta - BD_\alpha^{-1}C)^{-1}| \\
&\leqslant |(I - A_\beta^{-1}BD_\alpha^{-1}C)^{-1}||A_\beta^{-1}| \\
&\leqslant (I - \widetilde{A}_{\widetilde{\beta}}^{-1}\widetilde{B}\widetilde{D}_{\widetilde{\alpha}}^{-1}\widetilde{C})^{-1}\widetilde{A}_{\widetilde{\beta}}^{-1}M \\
&= \widetilde{W}_{\widetilde{\alpha}\widetilde{\beta}}^{-1}M \leqslant \widetilde{W}_{\widetilde{\alpha}\widetilde{\beta}}^{-1}.
\end{aligned}$$

Similarly,

$$|V_{\alpha\beta}^{-1}| \leqslant \widetilde{V}_{\widetilde{\alpha}\widetilde{\beta}}^{-1}N \leqslant \widetilde{V}_{\widetilde{\alpha}\widetilde{\beta}}^{-1}.$$

It follows from (2.4) that $|G_0| \leqslant \widetilde{G}_0,\ |H_0| \leqslant \widetilde{H}_0$. With the inequalities in (4.7), we have

$$\begin{aligned}
|E_0| &= |V_{\alpha\beta}^{-1}(D - \beta I - CA_\beta^{-1}B)| \\
&\leqslant \widetilde{V}_{\widetilde{\alpha}\widetilde{\beta}}^{-1}N(|\beta I - D| + |CA_\beta^{-1}B|) \\
&\leqslant \widetilde{V}_{\widetilde{\alpha}\widetilde{\beta}}^{-1}(N|\beta I - D| + \widetilde{C}\widetilde{A}_{\widetilde{\beta}}^{-1}\widetilde{B}) \\
&\leqslant \widetilde{V}_{\widetilde{\alpha}\widetilde{\beta}}^{-1}(\widetilde{\beta}I - \widetilde{D} + \widetilde{C}\widetilde{A}_{\widetilde{\beta}}^{-1}\widetilde{B}) \\
&= -\widetilde{E}_0.
\end{aligned}$$

Similarly, $|F_0| \leqslant -\widetilde{F}_0$. To complete the proof, we only need to show that $\rho(\mathscr{R})\rho(\mathscr{S}) < 1$, in view of (2.6). We have

$$|(R+\alpha I)^{-1}| = |(D-C\Phi+\alpha I)^{-1}| \leqslant |(I-D_\alpha^{-1}C\Phi)^{-1}||D_\alpha^{-1}| \leqslant (I-\widetilde{D}_{\widetilde{\alpha}}^{-1}\widetilde{C}\widetilde{\Phi})^{-1}\widetilde{D}_{\widetilde{\alpha}}^{-1}N = (\widetilde{R}+\widetilde{\alpha}I)^{-1}N$$

and

$$\begin{aligned} |R-\beta I| &= |D-\beta I-C\Phi| \\ &\leqslant |\beta I-D|+|C\Phi| \\ &\leqslant N^{-1}(\widetilde{\beta}I-\widetilde{D})+\widetilde{C}\widetilde{\Phi} \\ &\leqslant N^{-1}(\widetilde{\beta}I-\widetilde{D}+\widetilde{C}\widetilde{\Phi}) \\ &= -N^{-1}(\widetilde{R}-\widetilde{\beta}I). \end{aligned}$$

Then

$$|\mathscr{R}| \leqslant |(R-\beta I)||(R+\alpha I)^{-1}| \leqslant -N^{-1}(\widetilde{R}-\widetilde{\beta}I)(\widetilde{R}+\widetilde{\alpha}I)^{-1}N = -N^{-1}\widetilde{\mathscr{R}}N.$$

Similarly, $|\mathscr{S}| \leqslant -M^{-1}\widetilde{\mathscr{S}}M$. Therefore $\rho(\mathscr{R})\rho(\mathscr{S}) \leqslant \rho(\widetilde{\mathscr{R}})\rho(\widetilde{\mathscr{S}}) < 1$. □

It is straightforward to rewrite the inequalities in (4.5) as

$$(\alpha+p_i)(\beta-p_i) > -s_i^2,\ \beta > q_i, \quad 1 \leqslant i \leqslant n, \tag{4.8}$$

$$(\beta+p_j)(\alpha-p_j) > -s_j^2,\ \alpha > q_j, \quad n+1 \leqslant j \leqslant n+m, \tag{4.9}$$

where for $1 \leqslant i \leqslant n+m$,

$$\begin{aligned} p_i &= \frac{\mathrm{Re}([Q]_{ii})+q_i}{2} + \frac{(\mathrm{Im}([Q]_{ii}))^2}{2(\mathrm{Re}([Q]_{ii})-q_i)}, \\ s_i &= \frac{\mathrm{Re}([Q]_{ii})-q_i}{2} + \frac{(\mathrm{Im}([Q]_{ii}))^2}{2(\mathrm{Re}([Q]_{ii})-q_i)}. \end{aligned}$$

Note that the p_i's are the same as in (2.10). It is readily seen that the parameters $\alpha \geqslant \gamma_2$ and $\beta \geqslant \gamma_1$ (a slight extension of those in Theorem 2.1) always satisfy (4.8) and (4.9).

When $|\mathrm{Im}([Q]_{ii})|$ is large compared to $\mathrm{Re}([Q]_{ii})-q_i$ for some i, the numbers γ_1 and/or γ_2 will be large and the convergence of ADDA may be slow. To improve the performance of ADDA, we will find smaller parameters from the convergence region given by (4.8) and (4.9). The idea is to use the straight line $\beta = c\alpha$ to cut the convergence region, where the slope $c > 0$ is to be chosen properly.

THEOREM 4.3 Suppose the NARE (1.1) is in class H^* and $\widehat{Q}\mathbf{1} > 0$. Let Φ and Ψ be as in Theorem 1.1. Apply ADDA to the NARE (1.1) with parameters $\alpha, \beta > 0$. Suppose that $\beta = c\alpha$ with α satisfying

$$\alpha > \max\{\eta_1(c), \eta_2(c)\}, \tag{4.10}$$

where

$$\eta_1(c) = \max_{1\leqslant i\leqslant n} r_i(c), \quad \eta_2(c) = \max_{n+1\leqslant j\leqslant n+m} r_j(c)$$

with

$$r_i(c) = \frac{1}{2c}\left(-(c-1)p_i + \sqrt{(c-1)^2p_i^2 + 4c(p_i^2-s_i^2)}\right), \quad 1 \leqslant i \leqslant n, \tag{4.11}$$

and

$$r_j(c) = \frac{1}{2c}\left((c-1)p_j + \sqrt{(c-1)^2 p_j^2 + 4c(p_j^2 - s_j^2)}\right), \quad n+1 \leqslant j \leqslant n+m. \tag{4.12}$$

Then the sequences $\{E_k\}$, $\{F_k\}$, $\{H_k\}$, $\{G_k\}$ are well-defined and $\{H_k\}$ and $\{G_k\}$ converge quadratically to Φ and Ψ, respectively.

Proof. With $\beta = c\alpha$, (4.8) becomes

$$(\alpha + p_i)(c\alpha - p_i) > -s_i^2,\ \alpha > q_i/c, \quad 1 \leqslant i \leqslant n.$$

Let $r_i(c)$ be the larger root of $(\alpha + p_i)(c\alpha - p_i) + s_i^2 = 0$, given by (4.11). It is easy to show that $r_i(c) > q_i/c$ for $1 \leqslant i \leqslant n$. Thus (4.8) is satisfied when $\alpha > \eta_1(c)$. Similarly, (4.9) becomes

$$(c\alpha + p_j)(\alpha - p_j) > -s_j^2,\ \alpha > q_j, \quad n+1 \leqslant j \leqslant n+m.$$

Let $r_j(c)$ be the larger root of $(c\alpha + p_j)(\alpha - p_j) + s_j^2 = 0$, given by (4.12). It is easy to show that $r_j(c) > q_j$ for $n+1 \leqslant j \leqslant n+m$. Thus (4.9) is satisfied when $\alpha > \eta_2(c)$. When $\alpha > \max\{\eta_1(c), \eta_2(c)\}$, (4.8) and (4.9) are both satisfied and thus the sequences $\{E_k\}$, $\{F_k\}$, $\{H_k\}$, $\{G_k\}$ are well-defined and $\{H_k\}$ and $\{G_k\}$ converge quadratically to Φ and Ψ, respectively. □

We mentioned in the previous section that it is likely to have $\gamma_1 \approx \gamma_2$ after preprocessing. In that case, we can simply use SDA instead of ADDA. We therefore present the following special case of Theorem 4.3, with $c = 1$.

COROLLARY 4.1 Suppose the NARE (1.1) is in class H^* and $\widehat{Q}\mathbf{1} > 0$. Let Φ and Ψ be as in Theorem 1.1. Apply SDA to the NARE (1.1) with parameter α satisfying

$$\alpha > \max_{1 \leqslant i \leqslant m+n} \tau_i, \tag{4.13}$$

where

$$\tau_i = \sqrt{p_i^2 - s_i^2} = \sqrt{\left(\mathrm{Re}([Q]_{ii}) + \frac{(\mathrm{Im}([Q]_{ii}))^2}{\mathrm{Re}([Q]_{ii}) - q_i}\right) q_i}, \quad 1 \leqslant i \leqslant m+n.$$

Then, the sequences $\{E_k\}$, $\{F_k\}$, $\{G_k\}$ and $\{H_k\}$ are well-defined and $\{H_k\}$ and $\{G_k\}$ converge quadratically to Φ and Ψ, respectively.

Notice that $\max_{1 \leqslant i \leqslant m+n} \tau_i < \max_{1 \leqslant i \leqslant m+n} p_i = \max\{\gamma_1, \gamma_2\}$. So we now allow smaller values of the parameter α for SDA. By a more careful choice of c in Theorem 4.3, we can allow smaller values of the parameters α and β for ADDA.

PROPOSITION 4.4 Under the conditions in Theorem 4.3, there is a unique $c^* > 0$ such that $\eta_1(c^*) = \eta_2(c^*)$. Let $\alpha^* = \eta_1(c^*)$ and $\beta^* = c^*\alpha^*$. Then for any α and β satisfying (4.8) and (4.9), we have $\alpha > \alpha^*$ and $\beta > \beta^*$. In particular, $\gamma_2 > \alpha^*$ and $\gamma_1 > \beta^*$.

Proof. By computing derivatives it is easy to show that $r_i(c)(1 \leqslant i \leqslant n)$ is strictly decreasing on $(0,\infty)$, from ∞ to 0, and that $r_j(c)(n+1 \leqslant j \leqslant n+m)$ is strictly increasing on $(0,\infty)$, from $(p_j^2 - s_j^2)/p_j$ to p_j. It follows that $\eta_1(c)$ is strictly decreasing on $(0,\infty)$, from ∞ to 0, and that $\eta_2(c)$ is strictly increasing on $(0,\infty)$, from $\max_{n+1 \leqslant j \leqslant n+m}(p_j^2 - s_j^2)/p_j$ to γ_2. We only need to show that $\alpha > \alpha^*$ and $\beta > \beta^*$ for any α and β satisfying (4.8) and (4.9). In fact, $\beta = c\alpha$ with $c = \beta/\alpha$ and the conditions (4.8) and

(4.9) requires that $\alpha > \max\{\eta_1(c), \eta_2(c)\} \geqslant \max\{\eta_1(c^*), \eta_2(c^*)\} = \alpha^*$. For any $c > 0$ we replace α by β/c in (4.8) and (4.9), and find as before that (4.8) is equivalent to $\beta > c\eta_1(c)$ and (4.9) is equilvalent to $\beta > c\eta_2(c)$. So we need $\beta > \max\{c\eta_1(c), c\eta_2(c)\}$. As before we can show that $c\eta_1(c)$ is strictly decreasing on $(0,\infty)$, from γ_1 to $\max_{1\leqslant i\leqslant n}(p_i^2 - s_i^2)/p_i$, and that $c\eta_2(c)$ is strictly increasing on $(0,\infty)$, from 0 to ∞. It follows that $\beta > \max\{c\eta_1(c), c\eta_2(c)\} \geqslant \max\{c^*\eta_1(c^*), c^*\eta_2(c^*)\} = \beta^*$. □

From the proof, we can also see that

$$\alpha^* > \underline{\alpha} = \max_{n+1\leqslant j\leqslant n+m}(p_j^2 - s_j^2)/p_j, \quad \beta^* > \underline{\beta} = \max_{1\leqslant i\leqslant n}(p_i^2 - s_i^2)/p_i.$$

It follows that

$$\underline{\beta}/\gamma_2 < c^* < \gamma_1/\underline{\alpha}.$$

So c^* can be found by the usual bisection method applied to the function $\eta_1(c) - \eta_2(c)$ on the interval $[\underline{\beta}/\gamma_2, \gamma_1/\underline{\alpha}]$.

While Theorem 4.3 allows us to use smaller parameters α and β for ADDA, the smaller parameters will not always provide better convergence. The inequalities in (2.6) imply that, generally speaking, the smaller $\rho(\mathscr{R})\rho(\mathscr{S})$ is, the faster ADDA converges. In this regard, we are to choose parameters α and β to make $\rho(\mathscr{R})\rho(\mathscr{S})$ as small as possible. Once again we fix $c > 0$ and let $\beta = c\alpha$. We will try to find good values for α.

Let λ and μ be any eigenvalues of R and S, respectively, where R and S are given in (2.8). Note that (see (Liu & Xue, 2012)) λ is an eigenvalue of

$$H = \begin{bmatrix} D & -C \\ B & -A \end{bmatrix} \tag{4.14}$$

with positive real part and $-\mu$ is an eigenvalue of H with negative real part. It follows from Gershgorin's theorem and triangle inequality that

$$|\lambda| \leqslant \max_{1\leqslant i\leqslant n}(|[Q]_{ii}| + q_i), \quad |\mu| \leqslant \max_{n+1\leqslant i\leqslant n+m}(|[Q]_{ii}| + q_i).$$

The eigenvalues of $\mathscr{R}$ and $\mathscr{S}$ are $\frac{\lambda - c\alpha}{\lambda + \alpha}$ and $\frac{\mu - \alpha}{\mu + c\alpha}$, respectively.

PROPOSITION 4.5 Let $c > 0$ and $z = a + b\mathrm{i}$ with $a > 0$ and $b \in \mathbb{R}$. Then the function

$$f(\alpha) = \left|\frac{z - c\alpha}{z + \alpha}\right|$$

is increasing when

$$\alpha \geqslant \frac{-(c^2-1)|z|^2 + \sqrt{(c^2-1)^2|z|^4 + 4a^2c(c+1)^2|z|^2}}{2ac(c+1)}. \tag{4.15}$$

Proof. A simple computation shows that $f'(\alpha) \geqslant 0$ if and only if

$$-c(a - c\alpha)\left((a+\alpha)^2 + b^2\right) - (a+\alpha)\left((a - c\alpha)^2 + b^2\right) \geqslant 0,$$

which is equivalent to

$$ac(c+1)\alpha^2 + (c^2-1)|z|^2\alpha - a(c+1)|z|^2 \geqslant 0.$$

This inequality holds when α satisfies (4.15). □

If we use SDA, then $c = 1$ and (4.15) simplifies to $\alpha \geqslant |z|$. If follows that $\rho(\mathscr{R})\rho(\mathscr{S})$ is increasing in α if

$$\alpha \geqslant q^* = \max_{1\leqslant i\leqslant m+n}(|[Q]_{ii}| + q_i). \tag{4.16}$$

Note however that this is only a sufficient condition and $\rho(\mathscr{R})\rho(\mathscr{S})$ may be smaller for some smaller α values. In this regard, when $q^* \geqslant \gamma^* = \max\{\gamma_1, \gamma_2\}$ we will take $\alpha = \gamma^*$, i.e., we stick to the original strategy for choosing α for SDA. Now suppose $q^* < \gamma^*$. Since strict inequality is required in (4.13), we will require $\alpha \geqslant \tau^* = \delta \max_{1\leqslant i\leqslant m+n} \tau_i$, where δ is slightly bigger than 1 (we take $\delta = 1.01$ in our numerical experiments). Our strategy for choosing α in SDA is then to take

$$\alpha = \max\{\tau^*, \frac{1}{2}q^*\},$$

where the factor $\frac{1}{2}$ is introduced to account for the fact that (4.16) is only a sufficient condition. SDA with this new parameter strategy will be denoted by SDAn.

The situation for ADDA is more complicated since the inequality in (4.15) is complicated when $c \neq 1$. When $c \geqslant 1$ it is easy to see that (4.15) holds if $\alpha \geqslant |z|/\sqrt{c}$. But no such simplification is available when $c < 1$. When we apply Proposition 4.5 to $\frac{\lambda - c\alpha}{\lambda + \alpha}$ and $\frac{\mu - \alpha}{\mu + c\alpha} = \frac{\mu - \frac{1}{c}(c\alpha)}{\mu + c\alpha}$, we will run into difficulties when $c \neq 1$ since either c or $\frac{1}{c}$ will be smaller than 1. Since we have no useful monotonicity results to apply for ADDA, our parameter strategy is solely based on Proposition 4.4. We compute c^* by bisection method and take $\alpha = \delta\alpha^* = \delta\eta_1(c^*)$ and $\beta = c^*\alpha$, where δ is slightly bigger than 1 (we take $\delta = 1.01$ in our numerical experiments). ADDA with this new parameter strategy will be denoted by ADDAn.

Since there is more uncertainty about ADDAn, it may be appropriate to use SDAn when $0.1 < \gamma_1/\gamma_2 < 10$ and use ADDAn otherwise. This method will be denoted by DAn. Since the bounds 0.1 and 10 are somewhat arbitrary, one cannot expect DAn to be always better than SDAn and ADDAn.

5. Numerical results

In this section, we present some numerical results to illustrate the effectiveness of our preprocessing procedure and our new strategies for choosing parameters for SDA and ADDA. The experiments are performed in MATLAB (version 7.12) and the machine precision is 2.22×10^{-16}. An algorithm for computing the minimal nonnegative solution of (1.1) is terminated when the approximate solution Φ satisfies NRes $< 10^{-12}$, where

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

is the normalized residual.

We first give two examples to show the effectiveness of the preprocessing procedure. We apply SDA and ADDA to the NARE (1.1) directly and to the NARE (3.1) with $\omega \approx e^{-\mathrm{i}\phi_*}$. Note that the normalized residual for equation (3.1) is the same as that for equation (1.1). For SDA we take $\alpha = \max\{\gamma_1, \gamma_2\}$, and for ADDA we take $\alpha = \gamma_2$ and $\beta = \gamma_1$.

The first example is taken from (Liu & Xue, 2012).

EXAMPLE 5.1 Let $A, B, C, D \in \mathbb{C}^{n\times n}$ be given by

$$A = P + (\eta\mathrm{i})I_n, \quad D = P + (\eta\mathrm{i})I_n, \quad B = C = \xi I_n,$$

where $\xi \in (0,2), \eta \in \mathbb{R}$, and

$$P = \begin{bmatrix} 3 & -1 & & \\ & 3 & \ddots & \\ & & \ddots & -1 \\ -1 & & & 3 \end{bmatrix}.$$

So the NARE (1.1) is in class H^* and moreover $\widehat{Q}\mathbf{1} > 0$. For this example,

$$\gamma_1 = \gamma_2 = \frac{4+\xi}{2} + \frac{\eta^2}{2(2-\xi)}.$$

We can then take $\alpha = \beta$ in ADDA. So ADDA is reduced to SDA. Numerical results for SDA are already reported in (Liu & Xue, 2012) for $\xi = 1, 1.5, 1.9$ and $\eta = 0.1, 0.8, 1.5, 4$, where $\alpha = 1.01\gamma_1$. It was observed in (Liu & Xue, 2012) that the numer of iterations for SDA increases as ξ and η increase. We now take $\alpha = \gamma_1$ and have the same observation.

With the finding in this paper, we know that we should first preprocess the NARE (1.1) and then apply the SDA. For this special example, we can get the optimal ω without running the bisection procedure. We see directly that $\omega = \frac{|3+\eta i|}{3+\eta i}$ and transform the NARE (1.1) to the NARE (3.1). The NARE (3.1) is in class H^+ and the corresponding γ_1 and γ_2 are

$$\gamma_1 = \gamma_2 = \frac{1}{2}(\sqrt{9+\eta^2} + 1 + \xi).$$

The numbers of SDA iterations with preprocessing are given in Table 1, with the numbers of SDA iterations without preprocessing also given in parentheses. We have added $\eta = 10$ and $\xi = 1.999$. We can see that the simple preprocessing procedure is very effective. The improvement is more significant when ξ is closer to 2 or η is larger. We did not use very large η values in the table since for larger η values the less expensive fixed-point iterations in (Liu & Xue, 2012) also become more efficient and thus we cannot take too much credit for the more significant improvement brought to ADDA by preprocessing for large η values. On the other hand, the more significant improvement brought to ADDA by preprocessing for ξ values closer to 2 are meaningful since fixed-point iterations become less efficient as ξ gets closer to 2.

Table 1. Number of SDA iterations for Example 5.1.

	$\xi = 1$	$\xi = 1.5$	$\xi = 1.9$	$\xi = 1.999$
$\eta = 0.1$	4 (4)	4 (4)	5 (5)	7 (9)
$\eta = 0.8$	4 (4)	4 (4)	4 (6)	5 (12)
$\eta = 1.5$	4 (5)	4 (5)	4 (7)	4 (13)
$\eta = 4$	4 (6)	4 (7)	4 (9)	4 (15)
$\eta = 10$	4 (8)	4 (9)	4 (11)	4 (18)

The above example is somewhat too simple. But the preprocessing procedure can also be easily applied to practical situations. In fact, in (Liu & Xue, 2012) the NARE in class H^* is obtained in the

following way. Let

$$T = \begin{bmatrix} T_{11} & T_{12} & T_{13} \\ T_{21} & T_{22} & T_{23} \\ T_{31} & T_{32} & T_{33} \end{bmatrix}$$

be the infinitetesimal generator of a Markov chain, where the diagonal blocks are all square. Let s be a complex number in the open right half-plane. Then the comparison matrix of $sI - T$ is a nonsingular M-matrix with positive row sums. Let $T(s)$ be the Schur complement of $sI - T_{33}$ in $sI - T$. Then the comparison matrix $\widehat{T}(s)$ of $T(s)$ is a nonsingular M-matrix and $\widehat{T}(s)\mathbf{1} > 0$ by Lemma 2.4 of (Liu & Xue, 2012). The matrix associated with the NARE (1.1) is then given by

$$Q = \begin{bmatrix} C_2^{-1} & \\ & C_1^{-1} \end{bmatrix} \begin{bmatrix} & I \\ I & \end{bmatrix} T(s) \begin{bmatrix} & I \\ I & \end{bmatrix},$$

where C_1 and C_2 are diagonal matrices with positive diagonal entries. It is easily seen that $\widehat{Q}$ is a nonsingular M-matrix with $\widehat{Q}\mathbf{1} > 0$. However, we should not apply ADDA directly to the NARE (1.1). Instead we can determine ω by the bisection procedure, with very little additional cost, and then apply ADDA to the NARE (3.1). A specific example of this type is given below.

EXAMPLE 5.2 We take $C_1 = C_2 = I$, $T_{12} = 0.7I, T_{13} = 0.3I, T_{23} = T_{31} = 0.4I$, $T_{21} = T_{32} = 0.6I$, and

$$T_{11} = T_{22} = T_{33} = \begin{bmatrix} -3 & 2 & & & \\ 1 & -4 & 2 & & \\ & 1 & \ddots & \ddots & \\ & & \ddots & -4 & 2 \\ & & & 1 & -2 \end{bmatrix},$$

where all matrices are 100×100. We then take $s = 0.1 + \eta\mathrm{i}$ for $\eta = 1, 10, 20, 30, 40$. In each case, the matrix Q is a 200×200 dense matrix whose entries are not very regular.

In Table 2, we report the number k of bisection steps (we take $\tau = 10^{-6}$ in the stopping criterion in section 3, although a much larger τ, say 10^{-2}, usually works just as well) to get the approximation $\tilde{\phi}$ to ϕ_*, the number $\omega = e^{-\mathrm{i}\tilde{\phi}}$, the number of ADDA iterations Ite_ω with preprocessing, and the number of ADDA iterations Ite without preprocessing. Once again, we see that our simple preprocessing procedure is very effective.

Table 2. Number of ADDA iterations for Example 5.2.

s	$0.1+\mathrm{i}$	$0.1+10\mathrm{i}$	$0.1+20\mathrm{i}$	$0.1+30\mathrm{i}$	$0.1+40\mathrm{i}$
k	19	19	18	18	17
ω	$0.97-0.24\mathrm{i}$	$0.38-0.93\mathrm{i}$	$0.20-0.98\mathrm{i}$	$0.14-0.99\mathrm{i}$	$0.10-0.99\mathrm{i}$
Ite_ω	5	4	4	4	4
Ite	6	10	12	13	14

For Example 5.2 with $s = 0.1 + 20\mathrm{i}$, we plot in Fig. 1 the number of ADDA iterations for various ϕ. The minimal number of iterations is achieved on a large interval, with our $\phi_* = 1.3687$ near the midpoint of that interval. The figure also shows that there is no need to compute ϕ_* accurately.

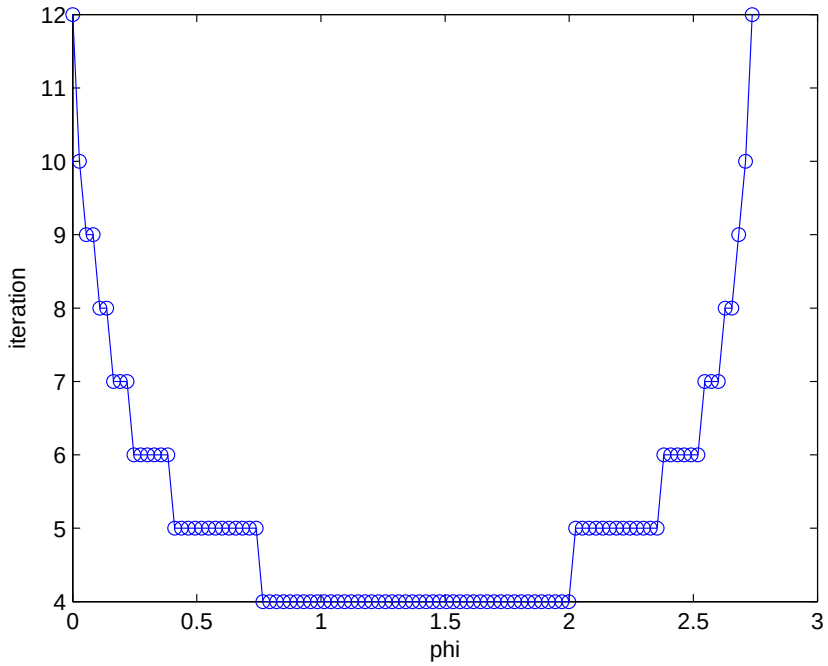

FIG. 1. Example 5.2 with $s = 0.1 + 20\mathrm{i}$

There are also some examples for which our preprocessing procedure is not helpful. In those situations, our new strategies for choosing parameters for SDA and ADDA can offer significant improvement.

EXAMPLE 5.3 Let

$$P = \begin{bmatrix} 2+\xi & -1 \\ -1 & 2+\xi \end{bmatrix}$$

and

$$A = D = P + \mathrm{i}\begin{bmatrix} \eta & \\ & -\eta \end{bmatrix}, \quad B = C = I,$$

where $\xi, \eta > 0$. For this example the preprocessing procedure chooses $\omega = 1$ and does not transform the original NARE (1.1). We test the new parameter strategy for SDA with $\xi = 1, 10^{-2}, 10^{-4}$ and $\eta = 1,\ 5$. The numbers of SDAn iterations are given in Table 3, with the numbers of SDA iterations also given in parentheses. The reduction achieved by the new parameter strategy is significant.

EXAMPLE 5.4 Let

$$A = \begin{bmatrix} 2+\mathrm{i} & -1 \\ -1 & 2-\mathrm{i} \end{bmatrix}, \quad D = \begin{bmatrix} \eta+\eta\mathrm{i} & -(\eta-1) \\ -(\eta-1) & \eta-\eta\mathrm{i} \end{bmatrix},$$

Table 3. Number of SDAn (SDA) iterations for Example 5.3.

	$\xi = 1$	$\xi = 10^{-2}$	$\xi = 10^{-4}$
$\eta = 1$	3 (3)	6 (8)	10 (15)
$\eta = 5$	5 (6)	8 (13)	12 (19)

Table 4. Number of iterations for Example 5.4.

Methods	SDA	ADDA	SDAn	ADDAn	DAn
$(\varepsilon,\eta) = (10^{-1},10)$	10	5	7	4	4
$(\varepsilon,\eta) = (10^{-2},10)$	13	7	9	6	6
$(\varepsilon,\eta) = (10^{-2},100)$	17	5	11	4	4

and $B = C = (1-\varepsilon)I$, where $0 < \varepsilon < 1,\ \eta > 1$. For this example the preprocessing procedure chooses $\omega = 1$ and does not transform the original NARE. We test the new parameter strategies for SDA and ADDA with $(\varepsilon,\eta) = (10^{-1},10),\ (10^{-2},10),\ (10^{-2},100)$. From Table 4 we can see that SDAn is significantly better than SDA and that ADDAn is better than ADDA, which is already very good. For this example, ADDAn is better than SDAn and DAn picks ADDAn each time.

EXAMPLE 5.5 Let

$$P = \text{tridiag}(-1,0,-1) \in \mathbb{R}^{2m\times 2m},$$

$$A = 0.1P + \xi I + \eta \mathrm{i} \begin{bmatrix} I_m & \\ & -I_m \end{bmatrix}, \quad D = 0.1P + 0.31I + \eta \mathrm{i} \begin{bmatrix} I_m & \\ & -I_m \end{bmatrix},$$

and $B = C = 0.1I_{2m}$. For this example (with $\xi > 0.3,\ \eta > 0$) the preprocessing procedure chooses $\omega = 1$ and does not transform the original NARE. We test the new parameter strategies for SDA and ADDA with $m = 100$, $\xi = 0.4,\ 0.5,\ 2,\ 4,\ 5,\ 20$, and $\eta = 10,\ 20$. In Table 5 we present the numbers of iterations for various methods for different pairs of (ξ,η). We can see that SDAn is significantly better than SDA and that ADDAn is better than ADDA, particularly when ξ is smaller. For this example, ADDAn is better than SDAn most of the time and DAn picks the better of SDAn and ADDAn most of the time.

EXAMPLE 5.6 For $m = 50$, let

$$A = 3I + \eta \mathrm{i} \begin{bmatrix} I_m & \\ & -I_m \end{bmatrix}, \quad B = I_{2m}.$$

We generate a random matrix $R = 0.1\text{rand}(2m,4m)$, and define

$$W = 1.01\ \text{diag}(R\mathbf{1}) - R(1:2m,1:2m).$$

We then take

$$D = W + \eta \mathrm{i} \begin{bmatrix} I_m & \\ & -I_m \end{bmatrix}, \quad C = R(1:2m,2m+1:4m).$$

For this example, we test the new parameter strategies for SDA and ADDA with $m = 50$ and $\eta =$ 5, 10, 20, 50, 100, after using the preprocessing procedure. We see from Table 6 that the new parameter strategies provide significant improvement.

Table 5. Number of iterations for Example 5.5 with $m = 100$.

Methods	SDA	ADDA	SDAn	ADDAn	DAn
$(0.4, 10)$	18	16	12	11	12
$(0.4, 20)$	20	18	13	12	13
$(0.5, 10)$	18	14	11	10	10
$(0.5, 20)$	20	16	12	11	11
$(2, 10)$	16	9	9	8	8
$(2, 20)$	18	11	10	9	9
$(4, 10)$	15	8	8	7	7
$(4, 20)$	17	9	9	8	8
$(5, 10)$	14	7	8	7	7
$(5, 20)$	16	9	9	8	8
$(20, 10)$	12	7	6	7	7
$(20, 20)$	14	7	7	7	7

Table 6. Number of iterations for Example 5.6 with $m = 50$.

Methods	SDA	ADDA	SDAn	ADDAn	DAn
$\eta = 5$	8	5	7	5	5
$\eta = 10$	10	7	8	6	6
$\eta = 20$	12	9	9	7	7
$\eta = 50$	15	11	10	8	8
$\eta = 100$	17	13	11	9	9

6. Conclusion

We have presented a simple preprocessing procedure for solving a class of complex nonsymmetric algebraic Riccati equations by doubling algorithms. It can significantly improve the performance of doubling algorithms, at a negligible cost. The preprocessing procedure is therefore highly recommended for solving these equations by doubling algorithms. We have also presented new strategies for choosing parameters for doubling algorithms. For some difficult cases, these strategies can offer significant further improvement after the preprocessing.

Acknowledgements

The work of C. G. was supported in part by a grant from the Natural Sciences and Engineering Research Council of Canada. The work of J. X. was supported in part by the National Science Foundation of China under Grant No. 11371105 and Laboratory of Mathematics for Nonlinear Science, Fudan University.

REFERENCES

BERMAN, A. & PLEMMONS, R. J. (1994) *Nonnegative Matrices in the Mathematical Sciences*. Philadelphia: SIAM.

BINI, D. A., IANNAZZO, B. & MEINI, B. (2012) *Numerical Solution of Algebraic Riccati Equations*. Philadelphia: SIAM.

BINI, D. A., MEINI, B. & POLONI, F. (2010) Transforming algebraic Riccati equations into unilateral quadratic matrix equations. *Numer. Math.*, **116**, 553–578.

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

GUO, C.-H.(2007) A new class of nonsymmetric algebraic Riccati equations. *Linear Algebra Appl.*, **426**, 636–649.

GUO, C.-H. & HIGHAM, N. J. (2007) Iterative solution of a nonsymmetric algebraic Riccati equation. *SIAM J. Matrix Anal. Appl.*, **29**, 396–412.

GUO, X.-X., LIN, W.-W. & XU, S.-F. (2006) A structure-preserving doubling algorithm for nonsymmetric algebraic Riccati equation. *Numer. Math.*, **103**, 393–412.

LIU, C. & XUE, J. (2012) Complex nonsymmetric algebraic Riccati equations arising in Markov modulated fluid flows. *SIAM J. Matrix Anal. Appl.*, **33**, 569–596.

WANG, W.-G., WANG, W.-C. & LI, R.-C. (2012) Alternating-directional doubling algorithm for M-matrix algebraic Riccati equations. *SIAM J. Matrix Anal. Appl.*, **33**, 170–194.