Solution of a nonsymmetric algebraic Riccati equation from a one-dimensional multistate transport model

Tiexiang Li

Department of Mathematics, Southeast University, Nanjing 211189,
People's Republic of China
txli@seu.edu.cn

Eric King-wah Chu*

School of Mathematical Sciences, Building 28, Monash University,
Victoria 3800, Australia
*Corresponding author: eric.chu@sci.monash.edu.au

Jong Juang

Department of Applied Mathematics, National Chiao Tung University, Hsinchu 300, Taiwan
jjuang@math.nctu.edu.tw

and

Wen-Wei Lin

Centre of Mathematical Modelling and Scientific Computing, Centre of Theoretical Sciences
and Department of Mathematics, National Taiwan University, No. 1, Sec. 4,
Roosevelt Road, Taipei 10617 Taiwan
wwlin@math.nctu.edu.tw

For the steady-state solution of a differential equation from a one-dimensional multistate model in transport theory, we shall derive and study a nonsymmetric algebraic Riccati equation $B^- - XF^- - F^+X + XB^+X = 0$, where $F^\pm \equiv (I - F)D^\pm$ and $B^\pm \equiv BD^\pm$ with positive diagonal matrices $D^\pm$ and possibly low-ranked matrices F and B. We prove the existence of the minimal positive solution X^* under a set of physically reasonable assumptions and study its numerical computation by fixed-point iteration, Newton's method and the doubling algorithm. We shall also study several special cases. For example when B and F are low ranked then $X^* = \Gamma \circ \left(\sum_{i=1}^4 U_i V_i^{\mathrm{T}} \right)$ with low-ranked U_i and V_i that can be computed using more efficient iterative processes. Numerical examples will be given to illustrate our theoretical results.

Keywords: algebraic Riccati equation; doubling algorithm; fixed-point iteration; Newton's method; reflection; transport theory.

1. Introduction

Transport theory has been an active area of research, associated with masters like R. E. Bellman and S. Chandrasekhar (see the references in Juang (1995)). A one-dimensional model was studied first in Juang (1995), starting a series of numerical studies, for example, in Lu (2005), Bai *et al.* (2008),
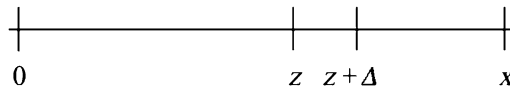
FIG. 1. The one-dimensional rod.

Bini *et al.* (2008), Lin *et al.* (2008), and Mehrmann & Xu (2008), in the past 15 years. We shall study a different one-dimensional multistate model from Bellman *et al.* (1973) and Bellman & Wing (1975, equation (1.37), p. 15) that is generalized slightly in this paper.

We start from a simple one-dimensional 'rod' or line segment that extends from 0 to x and denote a generic point in the rod by z, as in Fig. 1. Particles move to the right and left along this rod without colliding with one another while interacting with the rod itself without affecting it. We first assume that all particles are of the same type and have the same speed. The objective is to obtain information about the density of the beam of particles as a function of the position z.

We further assume that the probability of a particle at z (moving in either direction) interacting with the rod while moving a distance of Δ is given by the expression

$$\sigma(z)\Delta + \mathrm{o}(\Delta), \tag{1.1}$$

where $\sigma(\cdot) > 0$ is the macroscopic cross-section and $\mathrm{o}(\cdot)$ denotes higher-order terms. As a result of this interaction, an expected average of $f(z)$ and $b(z)$ new particles emerge at the point z in the same (forward) and opposite (backward) directions, respectively, as the original particle. Particles travelling to the left of $z = 0$ and the right of $z = x$ are lost to the system. Particles injected at the left and right ends, together with the new particles generated through the collision process, make up the total particle population of the system.

Initially, let us assume a time-independent state, where the expected particle population is stationary and independent of the time at which the system is observed. We define $u(z)$ and $v(z)$ to be the *expected* numbers of right- and left-moving particles, respectively, passing through the point z each second. (The adjective 'expected' is sometimes neglected but is necessary due to the stochastic nature of the problem.) From the definition in (1.1) the probability of particles passing through from z to $z + \Delta$ without interacting with the rod is $1 - \sigma(z)\Delta + \mathrm{o}(\Delta)$. The expected contribution to $u(z + \Delta)$ from this type of occurrence is

$$[1 - \sigma(z)\Delta + \mathrm{o}(\Delta)]u(z) = [(1 - \sigma(z)\Delta)]u(z) + \mathrm{o}(\Delta). \tag{1.2}$$

However, some right- (or left-) moving particles passing through z will interact with the rod before reaching $z + \Delta$. Each such event will produce an expected number $f(z)$ (or $b(z)$) of particles proceeding in the direction of interest. The expected contributions to $u(z + \Delta)$ from these types of occurrences are

$$\sigma(z)f(z)u(z)\Delta + \mathrm{o}(\Delta), \ \ \sigma(z)b(z)v(z)\Delta + \mathrm{o}(\Delta). \tag{1.3}$$

Other events can take place but the contributions are $\mathrm{o}(\Delta)$ and insignificant. Hence, summing the contributions in (1.2) and (1.3) yields

$$u(z + \Delta) = [(1 - \sigma(z)\Delta)]u(z) + \sigma(z)f(z)u(z)\Delta + \sigma(z)b(z)v(z)\Delta + \mathrm{o}(\Delta). \tag{1.4}$$

Taking the limit $\Delta \to 0$ in (1.4) leads to the following differential equation for u:

$$\frac{\mathrm{d}u}{\mathrm{d}z} = \sigma(z)\{[f(z) - 1]u(z) + b(z)v(z)\}, \quad u(0) = 0. \tag{1.5}$$

A similar particle counting process also produces the following differential equation for v:

$$-\frac{\mathrm{d}v}{\mathrm{d}z} = \sigma(z)\{b(z)u(z) + [f(z) - 1]v(z)\}, \quad v(x) = 1. \tag{1.6}$$

For a multistate model we allow n different states (e.g., speed, energy, type or any other features other than direction that distinguish between the particles). The macroscopic cross-section for the jth state is $\sigma_j(z) > 0$ and the probability in (1.1) resulting in the emission of particles in the jth state then reads

$$\sigma_j(z)\varDelta + \mathrm{o}(\varDelta).$$

Similarly, we have functions $u_j(z)$ and $v_j(z)$ ($j = 1, \ldots, n$) representing the expected number of particles in state j, moving to the right and left, respectively, past the point z each second. We define the matrix functions $F(z) = [f_{ij}(z)]$, $B(z) = [b_{ij}(z)]$, $\widetilde{F}(z) = [\widetilde{f}_{ij}(z)]$ and $\widetilde{B}(z) \equiv [\widetilde{b}_{ij}(z)]$, where

$$\widetilde{f}_{ij}(z) = \sigma_i(z)[f_{ij}(z) - \delta_{ij}], \quad \widetilde{b}_{ij}(z) = \sigma_i(z)b_{ij}(z), \qquad (i, j = 1, \ldots, n),$$

δ_{ij} is the Kronecker δ, and $f_{ij}(z)$ and $b_{ij}(z)$ are the expected numbers of particles travelling, respectively, in the forward and backward directions, respectively, after the collision of a particle of state j emitting particles of state i. A similar argument to that leading to (1.5) and (1.6) then produces

$$\frac{\mathrm{d}u}{\mathrm{d}z} = \widetilde{F}(z)u + \widetilde{B}(z)v, \quad -\frac{\mathrm{d}v}{\mathrm{d}z} = \widetilde{B}(z)u + \widetilde{F}(z)v,$$

with $u = [u_1, \ldots, u_n]^{\mathrm{T}}$, $v = [v_1, \ldots, v_n]^{\mathrm{T}}$ and the convention $M = [m_{ij}]$ (capital letters for matrices and the corresponding lower-case letters with indices for elements), and the initial conditions $u_i(0) = 0$ and $v_i(x) = \delta_{ij}$ (corresponding to the initial injection of a particle of state j from the right).

From the above discussion we expect $B, F > 0$ to satisfy

$$\sum_i (f_{ij} + b_{ij}) < 1 \quad \forall j. \tag{1.7}$$

We will allow the critical case of equality in (1.7) (the 'pure scattering' case in Bellman & Wing (1975, equation (4.1), p. 55)) later.

To carry out the invariant imbedding procedure the functions R and T are introduced, where $r_{ij}(x)$ is the expected number of particles emergent each second at $z = x$ in state i from a rod of length x when the only input is one particle per second in state j at the right end $z = x$, and $t_{ij}(x)$ is defined similarly except the emergence is at the other end $z = 0$. Consider a rod of length $x + \varDelta$ with the sub-rod of length x imbedded. Assuming that the reflecting response function $R(z) = [r_{ij}(z)]$ is known, the transmission response function $T(z) = [t_{ij}(z)]$ can be defined through a differential equation derived from a particle counting process.

Counting all the significant events as enumerated in Wing (1962), and allowing different macroscopic cross-sections $\sigma_j^{\pm}(z) > 0$ for sources from the left and the right, the following equation for $R(x) \equiv [r_{ij}(x)]$ can be derived:

$$\frac{\mathrm{d}R(x)}{\mathrm{d}x} = B^-(x) - R(x)F^-(x) - F^+(x)R(x) + R(x)B^+(x)R(x), \quad R(0) = 0, \tag{1.8}$$

where $B^{\pm} \equiv BD^{\pm}$, $F^{\pm} \equiv (I - F)D^{\pm}$, $D^{\pm} \equiv \mathrm{diag}\{\sigma_k^{\pm}\} > 0$, and B and F are possibly low ranked. (In Wing (1962) the signs of the linear terms on the right-hand side of (1.8) were positive. We change these

signs to make the resulting nonsymmetric algebraic Riccati equation (NARE) in (1.9) more consistent with notation in other papers on NAREs. Note that the $\{\sigma_k\}$ were independent of direction in Bellman *et al.* (1973) and Wing (1962).) After the determination of R, the function T can be derived from the simpler equation

$$\frac{\mathrm{d}T(x)}{\mathrm{d}x} = T(x)[\widetilde{F}(x) + \widetilde{B}(x)R(x)], \quad T(0) = I.$$

For the steady-state solution for a particular x, (1.8) leads to

$$B^- - XF^- - F^+X + XB^+X = 0, \tag{1.9}$$

with R replaced by the usual variable X for NAREs.

2. Existence of a solution

Some relevant definitions are as follows. For any matrices $\widehat{A}, \widehat{B} \in \mathbb{R}^{m \times n}$ we write $\widehat{A} \geqslant \widehat{B}$ or $\widehat{A} > \widehat{B}$ if their elements satisfy $\widehat{a}_{ij} \geqslant \widehat{b}_{ij}$ or $\widehat{a}_{ij} > \widehat{b}_{ij}$, respectively, for all i and j. A real square matrix $\widehat{A}$ is called a Z-matrix if all of its off-diagonal elements are nonpositive. It is clear that any Z-matrix $\widehat{A}$ can be written as $sI - \widehat{B}$ with $\widehat{B} \geqslant 0$. A Z-matrix $\widehat{A}$ is called an M-matrix if $s \geqslant \rho(\widehat{B})$, where $\rho(\cdot)$ is the spectral radius, and it is a singular M-matrix if $s = \rho(\widehat{B})$ and a nonsingular M-matrix if $s > \rho(\widehat{B})$. We have the following useful results from Berman & Plemmons (1994) and Guo & Higham (2007, Theorem 1.1).

LEMMA 2.1 For a Z-matrix $\widehat{A}$ the following statements are equivalent:

  (a) $\widehat{A}$ is a nonsingular M-matrix;
  (b) $\widehat{A}^{-1} \geqslant 0$;
  (c) $\widehat{A}v > 0$ for some vector $v > 0$.

THEOREM 2.2 Let us consider the NARE

$$X\widehat{C}X - X\widehat{D} - \widehat{A}X + \widehat{B} = 0, \tag{2.1}$$

where $\widehat{A}, \widehat{B}, \widehat{C}$ and $\widehat{D}$ are real matrices of sizes $m \times m$, $m \times n$, $n \times m$ and $n \times n$, respectively. Assume that

$$M = \begin{bmatrix} \widehat{D} & -\widehat{C} \\ -\widehat{B} & \widehat{A} \end{bmatrix} \tag{2.2}$$

is a nonsingular M-matrix or an irreducible singular M-matrix. Then the NARE has a minimal non-negative solution S. If M is irreducible, then $S > 0$, and $\widehat{A} - S\widehat{C}$ and $\widehat{D} - \widehat{C}S$ are irreducible M-matrices. If M is a nonsingular M-matrix, then $\widehat{A} - S\widehat{C}$ and $\widehat{D} - \widehat{C}S$ are nonsingular M-matrices. If M is an irreducible singular M-matrix with positive left and right null vectors $[u_1^{\mathrm{T}}, u_2^{\mathrm{T}}]^{\mathrm{T}}$ and $[v_1^{\mathrm{T}}, v_2^{\mathrm{T}}]^{\mathrm{T}}$ (where $u_1, v_1 \in \mathbb{R}^n$ and $u_2, v_2 \in \mathbb{R}^m$) satisfying

$$u_1^{\mathrm{T}}v_1 \neq u_2^{\mathrm{T}}v_2,$$

then

$$M_{\mathrm{S}} = I_n \otimes (\widehat{A} - S\widehat{C}) + (\widehat{D} - \widehat{C}S)^{\mathrm{T}} \otimes I_m$$

is a nonsingular M-matrix. If M is an irreducible singular M-matrix with $u_1^{\mathrm{T}}v_1 = u_2^{\mathrm{T}}v_2$, then M_{S} is an irreducible singular M-matrix.

We have the following existence result.

THEOREM 2.3 Under assumption (1.7), the unique minimal non-negative solution X^* of (1.9) exists.

Proof. From Theorem 2.2 we need to show that the Z-matrix

$$M = \begin{bmatrix} (I - F)D^- & -BD^+ \\ -BD^- & (I - F)D^+ \end{bmatrix} = \left\{ I - \begin{bmatrix} F & B \\ B & F \end{bmatrix} \right\} \begin{bmatrix} D^- & \\ & D^+ \end{bmatrix} \tag{2.3}$$

is a nonsingular M-matrix. Note that A is an M-matrix if and only if A^{T} is an M-matrix. Applying Lemma 2.1, we need to find a vector $v > 0$ such that $M^{\mathrm{T}}v > 0$, which is trivial from (1.7). $\square$

2.1 *NARE as an eigenvalue problem*

The NARE (1.9) can be reformulated as the following eigenvalue problem:

$$H \begin{bmatrix} I \\ X \end{bmatrix} = \begin{bmatrix} I \\ X \end{bmatrix} S, \quad H \equiv \begin{bmatrix} F^- & -B^+ \\ B^- & -F^+ \end{bmatrix} = \left\{ \begin{bmatrix} I & \\ & -I \end{bmatrix} + \begin{bmatrix} -F & -B \\ B & F \end{bmatrix} \right\} \begin{bmatrix} D^- & \\ & D^+ \end{bmatrix}. \tag{2.4}$$

From (1.7) it is easy to see that the eigenvalues of H are shifted slightly from $\pm\sigma_k^\mp$, splitting equally on opposite sides of the imaginary axis. Using the Gerschgorin theorem, with $\mathcal{D}(a, r) \equiv \{x \in \mathbb{C} : |x - a| \leqslant r\}$, the eigenvalues are in $[\bigcup_k \mathcal{D}(\sigma_k^-, \alpha\sigma_k^-)] \cup [\bigcup_k \mathcal{D}(-\sigma_k^+, \alpha\sigma_k^+)]$, divided equally on opposite sides of the imaginary axis, with $\alpha \equiv (\|F + B\|_1) < 1$ from (1.7).

REMARK 2.4 For the critical case with $\alpha = 1$ a simple application of the Gerschgorin theorem implies that the matrices H in (2.4) and M in (2.3) may be singular. However, the potential singularity may be detected or excluded by applying the extensions of the Gerschgorin theorem in Horn & Johnson (1985, Section 6.2). Consider all of the Gerschgorin disks of H^{T} containing the origin. At least one of the corresponding inequalities should not be satisfied with equality. In other words, we may have to exclude the ultra-critical case that all of the first or last n rows have their corresponding off-diagonal row sums equal to unity.

Note that, even if H or M are singular, the existence result in Theorem 2.2 still holds provided that M is irreducible. With the additional requirement for the null vectors as in Theorem 2.2, Newton's method in Section 4.1 will be quadratically convergent.

2.2 *NARE as a nonlinear equation*

To compute the minimal non-negative solution X^* for the NARE (1.9), consider it in component form as follows:

$$(\sigma_j^- + \sigma_i^+)x_{ij} = (BD^-)_{ij} + x_{i.}(FD^-)_{.j} + (FD^+)_{i.}x_{.j} + x_{i.}(BD^+)x_{.j}.$$

(Here $x_{i.}$ and $x_{.j}$ denote the ith row and jth column, respectively, of X, and $(\cdot)_{ij}$ denotes the element (i, j) of a matrix.) Equivalently, we have

$$X = \phi(X) \equiv \Gamma \circ (BD^- + XFD^- + FD^+X + XBD^+X), \quad \Gamma \equiv [(\sigma_j^- + \sigma_i^+)^{-1}], \tag{2.5}$$

with X being a Hadamard product. Note that Theorem 2.3, (2.5) and the assumption $B > 0$ imply that $X^* > 0$.

We have the following theorem for X^* and the fixed-point iteration.

THEOREM 2.5 Let $X^{(0)} = 0$ and $X^{(k+1)} = \phi(X^{(k)})$. Then, under assumption (1.7), we have the following

(i) the iterates satisfy $X^* \geqslant X^{(k+1)} \geqslant X^{(k)} \geqslant \Gamma \circ BD^- > 0$;

(ii) $X^{(k)} \to X^*$ as $k \to \infty$.

Proof. It is easy to show that $X^{(1)} = \Gamma \circ BD^- > 0$ from (2.5). For (i) consider the difference between $X^* = \phi(X^*)$ and $X^{(k)} = \phi(X^{(k-1)})$ as follows:

$$
\begin{aligned}
X^* - X^{(k)} = \Gamma \circ [&(X^* - X^{(k-1)})FD^- + FD^+(X^* - X^{(k-1)}) \\
&+ (X^* - X^{(k-1)})BD^+ X^* + X^{(k-1)}BD^+(X^* - X^{(k-1)})].
\end{aligned}
$$

Induction will then complete the argument for $X^* - X^{(k)} \geqslant 0$ in (i). Similarly, induction on the difference between $X^{(k+1)} = \phi(X^{(k)})$ and $X^{(k)} = \phi(X^{(k-1)})$ leads to $X^{(k+1)} - X^{(k)} \geqslant 0$ in (i).

For (ii) convergence is implied by (i) with the limit $\widetilde{X}^* \equiv \lim_{k\to\infty} X^{(k)} = X^*$ because (i) implies that $X^* \geqslant \widetilde{X}^* > 0$ and X^* is minimal. □

3. Low-ranked B and F

When the full-ranked decompositions $F = F_1 F_2^{\mathrm{T}}$ and $B = B_1 B_2^{\mathrm{T}}$ are of rank m and p, respectively, (2.5) implies that

$$
X = \Gamma \circ (B_1 B_2^{\mathrm{T}} D^- + Z_1 F_2^{\mathrm{T}} D^- + F_1 Z_2^{\mathrm{T}} + Z_3 Z_4^{\mathrm{T}}) \tag{3.1}
$$

with the auxiliary variables

$$
Z_1 \equiv X F_1, \quad Z_2 \equiv X^{\mathrm{T}} D^+ F_2, \quad Z_3 \equiv X B_1, \quad Z_4 \equiv X^{\mathrm{T}} D^+ B_2. \tag{3.2}
$$

Substituting X in (3.1) into (3.2), we have $2(m+p)n$ nonlinear equations for the $2(m+p)n$ unknowns in $Z_j (j = 1, \dots, 4)$ as follows:

$$
\begin{cases}
Z_1 = [\Gamma \circ (B_1 B_2^{\mathrm{T}} D^- + Z_1 F_2^{\mathrm{T}} D^- + F_1 Z_2^{\mathrm{T}} + Z_3 Z_4^{\mathrm{T}})]F_1, \\
Z_2 = [\Gamma^{\mathrm{T}} \circ (D^- B_2 B_1^{\mathrm{T}} + D^- F_2 Z_1^{\mathrm{T}} + Z_2 F_1^{\mathrm{T}} + Z_4 Z_3^{\mathrm{T}})]D^+ F_2, \\
Z_3 = [\Gamma \circ (B_1 B_2^{\mathrm{T}} D^- + Z_1 F_2^{\mathrm{T}} D^- + F_1 Z_2^{\mathrm{T}} + Z_3 Z_4^{\mathrm{T}})]B_1, \\
Z_4 = [\Gamma^{\mathrm{T}} \circ (D^- B_2 B_1^{\mathrm{T}} + D^- F_2 Z_1^{\mathrm{T}} + Z_2 F_1^{\mathrm{T}} + Z_4 Z_3^{\mathrm{T}})]D^+ B_2,
\end{cases} \tag{3.3}
$$

(cf. the $2n$ equations in $2n$ unknowns in Juang (1995) for a simpler NARE with $m = p = 1$). Similarly, X can be retrieved using (3.1) after the Z_j are obtained. It is obvious from Theorem 2.5 and (3.3) that $Z_j > 0 (j = 1, \dots, 4)$ and $X > 0$.

The convergence of various iterative schemes for the set of nonlinear equations (3.3) can be shown. First let $R_j (j = 1, \dots, 4)$ be the jth right-hand side in (3.3). Starting from $Z_j^{(0)} = 0 (j = 1, \dots, 4)$, we shall consider the following iterative methods.

(1) Simple iteration (SI): for $j, l = 1, \dots, 4$ we have

$$
Z_j^{(k+1)} = R_j(\dots, Z_l^{(k_{jl})}, \dots), \quad k_{jl} = k \qquad (\forall\, j, l). \tag{3.4}
$$

The right-hand sides R_j only involve the previous iterates $Z_l^{(k)}$.

(2) Modified simple iteration (MSI): for $j, l = 1, \ldots, 4$ we have

$$Z_j^{(k+1)} = R_j(\ldots, Z_l^{(\widetilde{k}_{jl})}, \ldots), \quad \widetilde{k}_{jl} = \begin{cases} k+1 & \text{if } l < j, \\ k & \text{otherwise.} \end{cases} \tag{3.5}$$

The right-hand sides R_j involve $Z_l^{(k)}$, as well as $Z_l^{(k+1)}$ if they have been computed.

(3) Nonlinear block Jacobi method (NBJ): for $j, l = 1, \ldots, 4$ we have

$$Z_j^{(k+1)} = R_j(\ldots, Z_l^{(\widehat{k}_{jl})}, \ldots), \quad \widehat{k}_{jl} = \begin{cases} k+1 & \text{if } l = j, \\ k & \text{otherwise.} \end{cases} \tag{3.6}$$

The right-hand sides R_j involve $Z_l^{(k)}$, as well as $Z_l^{(k+1)}$ with $l = j$ and the corresponding terms moved to the left-hand sides.

(4) Nonlinear block Gauss–Seidel method (NBGS): for $j, l = 1, \ldots, 4$ we have

$$Z_j^{(k+1)} = R_j(\ldots, Z_l^{(\check{k}_{jl})}, \ldots), \quad \check{k}_{jl} = \begin{cases} k+1 & \text{if } l \leqslant j, \\ k & \text{otherwise.} \end{cases} \tag{3.7}$$

The right-hand sides R_j involve $Z_l^{(k)}$ and $Z_l^{(k+1)}$ (when available), as well as $Z_l^{(k+1)}$ with $l = j$ and the corresponding terms moved to the left-hand sides.

From the above formulae we obviously have the following inequalities for the indices:

$$k = k_{jl} \leqslant \widetilde{k}_{jl} \leqslant \check{k}_{jl}, \quad k = k_{jl} \leqslant \widehat{k}_{jl} \leqslant \check{k}_{jl} \quad (j, l = 1, \ldots, 4). \tag{3.8}$$

The following simple lemma will be repeatedly applied in the proof of Theorem 3.2.

LEMMA 3.1 Let $U = [u_{ij}]$, $V = [v_{ij}]$ and $W = [w_{ij}] \in \mathbb{R}^{n \times q}$ be non-negative, and let $\widetilde{\Gamma} = [\widetilde{\tau}_{ij}] \in \mathbb{R}^{n \times n}$ be positive. Consider the linear system

$$U - [\widetilde{\Gamma} \circ (UV^{\mathrm{T}})]W = \widetilde{R}, \tag{3.9}$$

and its ith row has the form

$$u_i(I - P_i) = \widetilde{r}_i, \tag{3.10}$$

where

$$(P_i)_{sj} = \sum_{l=1}^{n} \widetilde{\tau}_{il} v_{ls} w_{lj}, \tag{3.11}$$

for $i = 1, \ldots, n$ and $s, j = 1, \ldots, q$. In addition, assume that

$$u_i^*(I - P_i^*) = \widetilde{r}_i^*, \tag{3.12}$$

where P_i^* is constructed as in (3.11) with v_{ls} replaced by $v_{ls}^* = (V^*)_{ls}$, $\widetilde{r}_i^*$ is the ith row of $\widetilde{R}^*$ and

$$\widetilde{R}^* \geqslant \widetilde{R} > 0, \quad U^* > 0, \quad V^* \geqslant V. \tag{3.13}$$

Then $I - P_i^*$ and $I - P_i$ are nonsingular M-matrices and

$$P_i^* \geqslant P_i, \quad (I - P_i^*)^{-1} \geqslant (I - P_i)^{-1} \geqslant 0. \tag{3.14}$$

Proof. From (3.11) and (3.13) we have that $u_i^*, \widetilde{r}_i^* > 0$ and $I - P_i^*$ is a Z-matrix. Consequently, the transpose of (3.12) and Lemma 2.1 imply that (the transpose of) $I - P_i^*$ is a nonsingular *M*-matrix with a non-negative inverse and $(I - P_i^*)v > 0$ for some vector $v > 0$. From (3.11) P_i is linear in V, with $V^* \geqslant V$ implying that $P_i^* \geqslant P_i$, $I - P_i \geqslant I - P_i^*$ and $(I - P_i)v \geqslant (I - P_i^*)v > 0$. Lemma 2.1 then implies that $I - P_i$ is a nonsingular *M*-matrix with a non-negative inverse and (3.14) follows. □

With the additional subscripts $\mathcal{I} = \mathcal{S}, \mathcal{M}, \mathcal{J}, \mathcal{G}$ for the four different methods (3.4)–(3.7), respectively (and ignoring them when the result holds for all the methods), we have the following results that are similar to those in Guo & Lin (2010).

THEOREM 3.2 We shall assume that (1.7) holds and we have the splitting $F = F_1 F_2^{\mathrm{T}}$ and $B = B_1 B_2^{\mathrm{T}}$, with the full-ranked $B_1, B_2, F_1, F_2 \geqslant 0$. We have for $j = 1, \ldots, 4$ and $k = 0, 1, \ldots$ that the following holds

(i) the iterates satisfy $Z_j^* \geqslant Z_j^{(k+1)} \geqslant Z_j^{(k)} > 0$, except $Z_j^{(0)} = 0$;

(ii) $Z_j^{(k)} \to Z_j^*$ as $k \to \infty$;

(iii) $0 \leqslant Z_{\mathcal{S},j}^{(k)} \leqslant Z_{\mathcal{M},j}^{(k)} \leqslant Z_{\mathcal{G},j}^{(k)}$;

(iv) $0 \leqslant Z_{\mathcal{S},j}^{(k)} \leqslant Z_{\mathcal{J},j}^{(k)} \leqslant Z_{\mathcal{G},j}^{(k)}$.

Proof. For NBJ and NBGS the formulae in (3.3) (ignoring the superscripts $(k+1)$ on the left-hand sides and (k) on the right for Z_j as in (3.6) and (3.7)) are equivalent to

$$Z_1 - [\Gamma \circ (Z_1 F_2^{\mathrm{T}})]D^- F_1 = \widetilde{R}_1 \equiv [\Gamma \circ (B_1 B_2^{\mathrm{T}} D^- + F_1 Z_2^{\mathrm{T}} + Z_3 Z_4^{\mathrm{T}})]F_1, \tag{3.15}$$

$$Z_2 - [\Gamma^{\mathrm{T}} \circ (Z_2 F_1^{\mathrm{T}})]D^+ F_2 = \widetilde{R}_2 \equiv [\Gamma^{\mathrm{T}} \circ (D^- B_2 B_1^{\mathrm{T}} + D^- F_2 Z_1^{\mathrm{T}} + Z_4 Z_3^{\mathrm{T}})]D^+ F_2, \tag{3.16}$$

$$Z_3 - [\Gamma \circ (Z_3 Z_4^{\mathrm{T}})]B_1 = \widetilde{R}_3 \equiv [\Gamma \circ (B_1 B_2^{\mathrm{T}} D^- + Z_1 F_2^{\mathrm{T}} D^- + F_1 Z_2^{\mathrm{T}})]B_1, \tag{3.17}$$

$$Z_4 - [\Gamma^{\mathrm{T}} \circ (Z_4 Z_3^{\mathrm{T}})]D^+ B_2 = \widetilde{R}_4 \equiv [\Gamma^{\mathrm{T}} \circ (D^- B_2 B_1^{\mathrm{T}} + D^- F_2 Z_1^{\mathrm{T}} + Z_2 F_1^{\mathrm{T}})]D^+ B_2. \tag{3.18}$$

The operators on the left-hand side of (3.15)–(3.18) are of similar form and we need to invert them with known right-hand sides $\widetilde{R}_j$. For the generic term $U - [\widetilde{\Gamma} \circ (UV^{\mathrm{T}})]W$ for $\widetilde{\Gamma} = [\widetilde{\tau}_{ij}] \in \mathbb{R}^{n \times n}$ and $U = [u_{ij}], V = [v_{ij}], W = [w_{ij}] \in \mathbb{R}^{n \times q}$ (with $q = m$ or p), the (i, j) component equals

$$\{U - [\widetilde{\Gamma} \circ (UV^{\mathrm{T}})]W\}_{ij} = u_{ij} - \sum_{s=1}^{q} u_{is} \left(\sum_{l=1}^{n} \widetilde{\tau}_{il} v_{ls} w_{lj} \right), \tag{3.19}$$

implying that the ith row in (3.15)–(3.18) has the generic form (3.10) with $P_i \in \mathbb{R}^{q \times q}$ ($i = 1, \ldots, n$) as in (3.11). Note that u_i and $\widetilde{r}_i$ in (3.10) are the ith rows of U and the right-hand side $\widetilde{R}$ in (3.15)–(3.18) or Table 1 below, respectively.

We shall prove (i) by induction.

For the $k = 0$ case in (i), except for Z_4 in NBGS, it is easy to see from (3.3) and (3.4)–(3.7) that $Z_j^{(1)}$ are well defined and

$$Z_j^* \geqslant Z_j^{(1)} > Z_j^{(0)} = 0 \quad (j = 1, \ldots, 4), \tag{3.20}$$

where the limits (indicated by $(\cdot)^*$) are guaranteed to exist by Theorems 2.3 or 4.1 together with (3.2). Note from Table 1 and $Z_j^{(0)} = 0$ that the P_i are constant in (3.10) for Z_1 and Z_2 in NBJ and NBGS, and $P_i = 0$ for Z_3 and Z_4 in NBJ as well as Z_3 in NBGS (because the corresponding Vs in Table 1 vanish).

TABLE 1 *Construction of P_i in (3.10) or (3.19) for NBJ and NBGS*

Method	U	$\widetilde{\Gamma}$	V	W	$\widetilde{R}$
Both	$Z_1^{(k+1)}$	ΓD^-	F_2	F_1	$\widetilde{R}_1$
	$Z_2^{(k+1)}$	$\Gamma^{\mathrm{T}} D^+$	F_1	F_2	$\widetilde{R}_2$
	$Z_3^{(k+1)}$	Γ	$Z_4^{(k)}$	B_1	$\widetilde{R}_3$
NBJ	$Z_4^{(k+1)}$	$\Gamma^{\mathrm{T}} D^+$	$Z_3^{(k)}$	B_2	$\widetilde{R}_4$
NBGS	$Z_4^{(k+1)}$	$\Gamma^{\mathrm{T}} D^+$	$Z_3^{(k+1)}$	B_2	$\widetilde{R}_4$

For Z_4 in NBGS the iteration has the following form that is similar to (3.10):

$$u_i^{(1)}[I - P_i^{(1)}] = \widetilde{r}_i^{(0)}, \tag{3.21}$$

where $P_i^{(1)}$ is linear in $V = Z_3^{(1)}$, which has been proved to satisfy

$$Z_3^* \geqslant Z_3^{(1)} > 0. \tag{3.22}$$

From (1.7) and (3.20) we have $u_i^* > 0$ (from methods other than NBGS) and $\widetilde{r}_i^* \geqslant \widetilde{r}_i^{(0)} > 0$. With (3.21) in place of (3.10), or $u_i^{(1)}$, $P_i^{(1)}$ and $\widetilde{r}_i^{(0)}$ in place of u_i, P_i and $\widetilde{r}_i$, respectively, Lemma 3.1 then implies that $I - P_i^{(1)}$ is nonsingular and $(I - P_i^*)^{-1} \geqslant [I - P_i^{(1)}]^{-1} \geqslant 0$. Consequently, $u_i^{(1)}$ and thus $Z_4^{(1)}$ are well defined, and we have

$$u_i^* = \widetilde{r}_i^*(I - P_i^*)^{-1} \geqslant \widetilde{r}_i^{(0)}[I - P_i^{(1)}]^{-1} = u_i^{(1)} \quad \Longleftrightarrow \quad Z_4^* \geqslant Z_4^{(1)} > Z_4^{(0)} = 0.$$

We have proved the $k = 0$ case in (i).

Assuming that (i) holds up to some value of k, we shall prove the $(k + 1)$ case. The conclusion can be easily drawn for SI and MSI by considering the differences between (3.4) and (3.5) for successive values of k as well as the limiting case when $k \to \infty$. For NBJ and NBGS in (3.6) and (3.7) for Z_1 and Z_2, we have that $P_i^{(s)} = P_i^* = P_i$ (for all s) are constant as $V = F_2, F_1$ from Table 1. The limiting case in (3.12) or a trivial application of Lemma 3.1 imply that $[I - P_i^{(s)}]^{-1} \geqslant 0$ (for all s). For NBJ and NBGS for Z_3 and Z_4, the iterations take the following generic form, as in (3.10):

$$u_i^{(s+1)}[I - P_i^{(s+1)}] = \widetilde{r}_i^{(s)} \quad (s = 0, 1, \ldots, k + 1), \tag{3.23}$$

with $\widetilde{r}_i^{(s)}$ and $P_i^{(s+1)}$ dependent on $V = Z_j^{(s)}$ or $Z_3^{(s+1)}$ (for NBGS for the iteration for Z_4, where $Z_3^{(k+2)} \geqslant Z_3^{(k+1)}$ when the iterations in (3.7) are executed in the intended order $j = 1, \ldots, 4$). From the induction hypothesis and the linearity of $\widetilde{R}$ with respect to Z_j (for all j) in (3.15)–(3.18), we have $U^* > 0$, $V^* \geqslant V > 0$ and $\widetilde{R}^* \geqslant \widetilde{R}^{(s)} > 0$. With (3.23) in place of (3.10), or $u_i^{(s+1)}$, $P_i^{(s+1)}$ and $\widetilde{r}_i^{(s)}$ in place of u_i, P_i and $\widetilde{r}_i$, respectively, Lemma 3.1 implies that $I - P_i^{(s+1)}$ ($s = 0, 1, \ldots, k + 1$) are nonsingular M-matrices with non-negative inverses, $P_i^* \geqslant P_i^{(k+2)} \geqslant P_i^{(k+1)}$ and $(I - P_i^*)^{-1} \geqslant$

$[I - P_i^{(k+2)}]^{-1} \geqslant [I - P_i^{(k+1)}]^{-1} \geqslant 0$. From successive values of $s = k, k+1$ and the limiting case $k \to \infty$, appropriate differences yield

$$
\begin{aligned}
u_i^* - u_i^{(k+2)} &= \widetilde{r}_i^*[I - P_i^*]^{-1} - \widetilde{r}_i^{(k+1)}[I - P_i^{(k+2)}]^{-1} \\
&= (\widetilde{r}_i^* - \widetilde{r}_i^{(k+1)})(I - P_i^*)^{-1} + \widetilde{r}_i^{(k+1)}[I - P_i^{(k+2)}]^{-1}[P_i^* - P_i^{(k+2)}](I - P_i^*)^{-1} \geqslant 0.
\end{aligned}
$$

Similarly, we have

$$
u_i^{(k+2)} - u_i^{(k+1)} = \widetilde{r}_i^{(k+1)}[I - P_i^{(k+2)}]^{-1} - \widetilde{r}_i^{(k)}[I - P_i^{(k+1)}t]^{-1} \geqslant 0.
$$

Thus a $Z_j^* \geqslant Z_j^{(k+2)} \geqslant Z_j^{(k+1)}$ ($j = 1, \ldots, 4$) and the induction for (i) is complete.

For (ii) a similar argument as in the proof of (ii) in Theorem 2.5 can be applied. For (iii) and (iv) we note that the iterates $Z_j^{(k)}$ are increasing towards their respective limits Z_j^*, and R_j and $\widetilde{R}_j$ preserve the order of positivity of their arguments. We shall prove the inequalities again by induction. The initial cases for $k = 0$ are obvious. Assume that the results hold for some value of k. From (3.8), for $j = 1, \ldots, 4$ and $k = 0, 1, \ldots$ we have

$$
\begin{aligned}
Z_{\mathcal{S},j}^{(k+1)} &= R_j(\ldots, Z_{\mathcal{S},l}^{(k)}, \ldots) \leqslant R_j(\ldots, Z_{\mathcal{M},l}^{(k)}, \ldots) \leqslant R_j(\ldots, Z_{\mathcal{M},l}^{(\widetilde{k}_{jl})}, \ldots) = Z_{\mathcal{M},j}^{(k+1)}, \\
Z_{\mathcal{S},j}^{(k+1)} &= R_j(\ldots, Z_{\mathcal{S},l}^{(k)}, \ldots) \leqslant R_j(\ldots, Z_{\mathcal{J},l}^{(k)}, \ldots) \leqslant R_j(\ldots, Z_{\mathcal{J},l}^{(\widehat{k}_{jl})}, \ldots) = Z_{\mathcal{J},j}^{(k+1)}, \\
Z_{\mathcal{M},j}^{(k+1)} &= R_j(\ldots, Z_{\mathcal{M},l}^{(\widetilde{k}_{jl})}, \ldots) \leqslant R_j(\ldots, Z_{\mathcal{G},l}^{(\widetilde{k}_{jl})}, \ldots) \leqslant R_j(\ldots, Z_{\mathcal{G},l}^{(\check{k}_{jl})}, \ldots) = Z_{\mathcal{G},j}^{(k+1)}.
\end{aligned}
$$

For the right-most inequalities in (iv) consider the iterations in the general form (3.23). We then have

$$
u_{\mathcal{J},i}^{(k+1)} = \widetilde{r}_{\mathcal{J},i}^{(k)}[I - P_{\mathcal{J},i}^{(k+1)}]^{-1} \leqslant \widetilde{r}_{\mathcal{G},i}^{(k)}[I - P_{\mathcal{J},i}^{(k+1)}]^{-1} \leqslant \widetilde{r}_{\mathcal{G},i}^{(k)}[I - P_{\mathcal{G},i}^{(k+1)}]^{-1} = u_{\mathcal{G},i}^{(k+1)}
$$

since $P_{\mathcal{J},i}^{(k+1)} \leqslant P_{\mathcal{G},i}^{(k+1)}$ and $\widetilde{r}_{\mathcal{J},i}^{(k)} \leqslant \widetilde{r}_{\mathcal{G},i}^{(k)}$. Thus $Z_{\mathcal{J},j}^{(k+1)} \leqslant Z_{\mathcal{G},j}^{(k+1)}$, and induction is complete. $\qquad\square$

REMARK 3.3 The assumption that $B_i, F_i \geqslant 0$ ($i = 1, 2$) is just a convenient sufficient condition for Theorem 3.2. There are many other weaker but more tedious sufficient conditions that we can write down. For example, by careful application of (3.15)–(3.18) in the proof, we can make the alternative assumption, with $W_1 \equiv (\Gamma D^-) \circ B$ and $W_2 \equiv (D^+ \Gamma D^-) \circ B$, that the following matrices are positive:

$$
F_2 F_1^{\mathrm{T}}; \; W_1 B_1, \; W_1 F_1, \; B_2^{\mathrm{T}} W_1 B_1; \; F_2^{\mathrm{T}} W_2, \; B_2^{\mathrm{T}} W_2 B_1.
$$

4. General case

For the general case with B and F being full ranked, the NARE (1.9), namely,

$$
B^- - XF^- - F^+X + XB^+X = 0,
$$

or the equivalent (2.5), can be solved by fixed-point iteration (as in Theorem 2.5), Newton's method (Lu, 2005; Guo & Higham, 2007; Lin *et al.*, 2008) or doubling (Guo *et al.*, 2006; Chiang *et al.*, 2009). The existence of the unique minimal positive solution X^* of (1.9) is guaranteed by Theorem 2.3.

4.1 *Newton's method*

Considering the NARE (1.9), let $R(X)$ denote the left-hand side of the equation. At the $(k+1)$th iteration with $X^{(k)}$ being an approximate solution and $X^{(k+1)} = X^{(k)} + \delta X^{(k+1)}$, Newton's method requires the solution of the Sylvester equation

$$(F^+ - X^{(k)}B^+)\delta X^{(k+1)} + \delta X^{(k+1)}(F^- - B^+ X^{(k)}) = R(X^{(k)}). \tag{4.1}$$

The convergence of Newton's method is guaranteed by the following theorem quoted from Guo & Higham (2007, Theorem 2.3).

THEOREM 4.1 Let S be the minimal positive solution of (1.9). Then, under assumption (1.7), for the Newton iteration (4.1) with $X^{(0)} = 0$, the sequence $\{X^{(k)}\}$ is well defined, $X^{(k)} \leqslant X^{(k+1)} \leqslant S$ for all $k \geqslant 0$, and $\lim_{k\to\infty} X^{(k)} = S$.

The proof makes use of selected results from Theorem 2.2. In particular, when vectorized the above Sylvester operator can be written as the matrix operator M_S (with $m = n$) as in Theorem 2.2.

4.2 *Doubling*

We shall quote the doubling algorithm for the general NARE (2.1), with the matrix M in (2.2) being a nonsingular M-matrix, from Guo *et al.* (2006). Note that, per iteration, the doubling algorithm is faster than Newton's method, as concluded in Guo *et al.* (2006), Guo (2007) and Table 2, and we refer the reader to the details in these references.

For the general NARE

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

with the corresponding matrix M in (2.2) being a nonsingular M-matrix, we first transform $\widehat{A}$, $\widehat{B}$, $\widehat{C}$ and $\widehat{D}$ to

$$E_\gamma = I - 2\gamma V_\gamma^{-1}, \quad G_\gamma = 2\gamma D_\gamma^{-1}\widehat{C}W_\gamma^{-1}, \quad F_\gamma = I - 2\gamma W_\gamma^{-1}, \quad H_\gamma = 2\gamma W_\gamma^{-1}\widehat{B}D_\gamma^{-1},$$

with the parameter $\gamma \geqslant \max\{\widehat{a}_{ii}, \ldots, \widehat{a}_{nn}; \widehat{d}_{11}, \ldots, \widehat{d}_{nn}\}$ and

$$A_\gamma = \widehat{A} + \gamma I, \quad D_\gamma = \widehat{D} + \gamma I, \quad W_\gamma = A_\gamma - \widehat{B}D_\gamma^{-1}\widehat{C}, \quad V_\gamma = D_\gamma - \widehat{C}A_\gamma^{-1}\widehat{B}.$$

The doubling algorithm can then be summarized as follows:

$$E_0 = E_\gamma, \quad F_0 = F_\gamma, \quad G_0 = G_\gamma, \quad H_0 = H_\gamma,$$
$$E_{k+1} = E_k(I - G_k H_k)^{-1}E_k, \quad F_{k+1} = F_k(I - H_k G_k)^{-1}F_k,$$
$$G_{k+1} = G_k + E_k(I - G_k H_k)^{-1}G_k F_k, \quad H_{k+1} = H_k + F_k(I - H_k G_k)^{-1}H_k E_k.$$

TABLE 2 *Operation counts per iteration*

B, F	Method	Flops
Low-ranked	NBGS	$8(m + p + 2)n^2$
General	Fixed-point iteration	$4n^3$
	Newton's method	$36n^3$
	Doubling	$10\frac{2}{3}n^3$

The iterates are well defined with $I - H_k G_k$ and $I - G_k H_k$ being nonsingular M-matrices for all k, and $H_k \to X$ and $G_k \to Y$ (the solution to the adjoint algebraic Riccati equation to (2.1)) from below quadratically as $k \to \infty$ (see Guo *et al.*, 2006, Theorem 5.1).

Note that, if $D^\pm = D$, $\widehat{B} = \widehat{C} = BD$ and $\widehat{A} = \widehat{D} = (I - F)D$, then we can halve the computation as $E_k = F_k$ and $G_k = H_k$ for all k. Some saving in computation can also be made in Newton's method as the Sylvester equations in the iteration become Lyapunov equations.

5. Numerical examples

For comparison, we shall summarize the operation counts per iteration of various iterative methods in Table 2. We shall show only the dominant terms, assuming that $n \gg m, p$. For low-ranked B and F, only the fastest method NBGS is considered. The slow fixed-point iteration method is also included for comparison.

REMARK 5.1 For the simpler NARE considered in Lu (2005), Bai *et al.* (2008), Lin *et al.* (2008) and Mehrmann & Xu (2008), the associated structure gave rise to iterative solution processes (analogous to our SI, MSI, NBJ and NBGS methods) of O(n) computational complexity. The 'fast' Newton method in Bini *et al.* (2008) is of O(n^2) complexity and is uncompetitive. However, our NARE in (1.9), for both the low-ranked and the general cases, has very different structures. It is likely that faster solution methods can be found but we do not anticipate methods of less complexity than the O(n^2) NBGS method (for the low-ranked case) and the O(n^3) doubling method (for the general case).

We shall consider two randomly generated examples for $n = 64, 128, 256, 512, 1024$ and 2048. Example 1 has B and F being full ranked, and Example 2 has B and F of rank 10. For the examples the assumptions in Theorems 2.3 and 3.2 are satisfied. The numerical computation has been carried out using MATLAB R2008b on a laptop with precision eps $= 2.2204 \times 10^{-16}$ (MathWorks, 2002).

For Example 1, fixed-point iteration, Newton's method and the doubling algorithm have been compared for various values of n. The iterations have been run until convergence with tolerance tol $= 10^{-15}$. The results are summarized in Table 3, with t_n denoting the CPU time, $r_n \equiv t_n / t_{n/2}$ and #It being the number of iterations required, for particular values of n. The iterates are also plotted in Fig. 2 for $n = 1024$. Note that the residuals in Figs 2 and 3 are plotted using a logarithmic scale.

Table 3 and Fig. 2 seem to indicate that the doubling algorithm performs better than Newton's method in CPU-time and the fixed-point iteration method is the slowest, as predicted in Table 2. The ratios r_n illustrate the O(n^3) complexity of the methods. The graphs in Fig. 2 illustrate the quadratic convergence of the doubling algorithm and Newton's method, with the fixed-point iteration method obviously converging linearly. Newton's method is two to three times faster than the doubling method in terms of number of iterations, but the latter has an advantage in operation count per iteration by a factor of 3.6, resulting in its better efficiency in terms of CPU time. Note that the cputime command in MATLAB (MathWorks, 2002) is not an exact reflection of CPU time consumed and should be used as a rough guide only. Also, users have no control over some parts of the algorithms, such as the inversion of the Sylvester operators by the MATLAB command lyap (MathWorks, 2002) in Newton's method.

For Example 2, only the fastest iteration method NBGS has been tested against the doubling method and the results are summarized in Table 4 (for $n = 64, 128, 256, 512, 1024, 2048$) and Fig. 3 (for $n = 1024$), with tol $= 10^{-15}$. The O(n^2) complexity of NBGS and the O(n^3) complexity of the doubling method are illustrated in the ratios r_n in Table 4. The linear convergence of NBGS and the quadratic convergence of the doubling method can be seen clearly in Fig. 3. NBGS usually requires less iterations than the doubling method and is also more efficient in terms of CPU time because of its superior
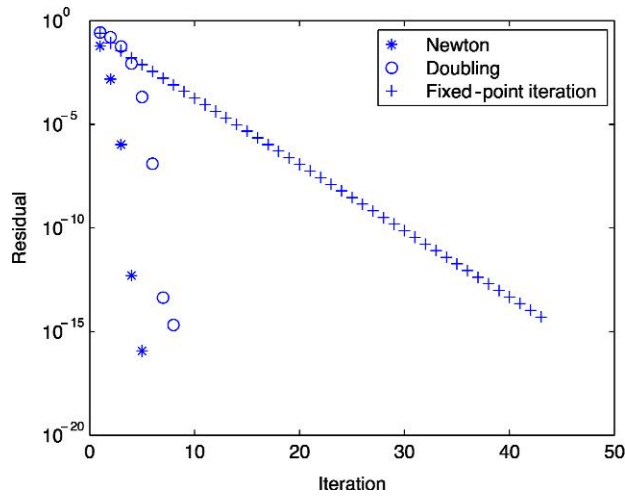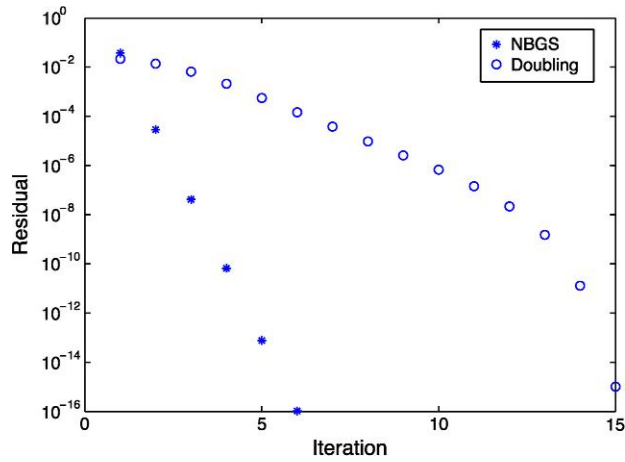
FIG. 2. Residuals for Example 1 ($n = 1024$).

FIG. 3. Residuals for Example 2 ($n = 1024$).

TABLE 3 *CPU times and iteration numbers for Example* 1

	Fixed-point iteration			Newton			Doubling		
n	t_n	r_n	#It	t_n	r_n	#It	t_n	r_n	#It
64	0.249	—	79	0.062	—	5	0.062	—	12
128	1.809	7.26	96	1.108	17.8	5	0.561	9.04	14
256	9.111	5.04	98	3.994	3.60	5	2.371	4.23	13
512	78.35	8.59	125	20.05	5.02	6	18.22	7.68	16
1024	560.2	7.15	132	186.0	9.27	6	124.1	6.81	15
2048	5147	9.18	150	1965	10.5	6	1112	8.96	18

TABLE 4 *CPU times and iteration numbers for Example* 2 *with NBGS and doubling and the methods*

	NBGS			Doubling		
n	t_n	r_n	#It	t_n	r_n	#It
64	0.468	—	12	0.125	—	13
128	1.203	2.57	10	0.625	5.00	13
256	3.718	3.09	8	4.094	6.55	14
512	13.67	3.67	7	25.66	6.27	15
1024	51.62	3.77	6	207.0	8.07	15
2048	170.1	3.29	5	1031	4.98	17

operation count per iteration. For fixed ranks of B and F the number of iterations required decreases with respect to n, as indicated in the fourth column of Table 4.

6. Concluding remarks

For the one-dimensional multistate model in transport theory we need to solve a differential equation to obtain the reflection function R. For the steady-state solution we have derived an NARE from the differential equation. We have proved the existence and uniqueness of the minimal positive solution of the NARE. When B and F are low ranked the NBGS method of $O(n^2)$ complexity solves the NARE efficiently. For the general case the doubling algorithm seems to be more efficient than Newton's method. The numerical results support our theoretical findings.

For future work we need to improve on the efficiency of the numerical algorithms for large values of n. Finally, there are other similar models and problems in transport theory (Bellman & Wing, 1975) worthy of investigation.

Funding

REFERENCES

BAI, Z.-Z., GAO, Y.-H. & LU, L.-Z. (2008) Fast iterative schemes for nonsymmetric algebraic Riccati equations arising from transport theory. *SIAM J. Sci. Comput.*, **30**, 804–818.

BELLMAN, R. E., UENO, S. & VASUDEVAN, R. (1973) On the matrix Riccati equation of transport processes. *J. Math. Anal. Appl.*, **44**, 472–481.

BELLMAN, R. E. & WING, G. M. (1975) *An Introduction to Invariant Imbedding*. New York: Wiley.

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

BINI, D. A., IANNAZZO, B. & POLONI, F. (2008) A fast Newton's method for a nonsymmetric algebraic Riccati equation. *SIAM J. Matrix Anal. Appl.*, **30**, 276–290.

CHIANG, C.-Y., CHU, E. K.-W., GUO, C.-H., HUANG, T.-M., LIN, W.-W. & XU, S.-F. (2009) Convergence analysis of the doubling algorithm for several nonlinear matrix equations in the critical case. *SIAM J. Matrix Anal. Appl.*, **31**, 227–247.

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, C.-H. & LANCASTER, P. (2006) Analysis and modification of Newton's method for algebraic Riccati equations. *Math. Comput.*, **67**, 1089–1105.

GUO, C.-H. & LIN, W.-W. (2010) Convergence rates of some iterative methods for nonsymmetric algebraic Riccati equations arising in transport theory. *Linear Algebra Appl.*, **432**, 283–291.

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

HORN, R. A. & JOHNSON, C. R. (1985) *Matrix Analysis*. Cambridge: Cambridge University Press.

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

LIN, Y., BAO, L. B. & WEI, Y. (2008) A modified Newton method for solving non-symmetric algebraic Riccati equations arising in transport theory. *IMA J. Numer. Anal.*, **29**, 215–224.

LU, L.-Z. (2005) Newton iterations for a non-symmetric algebraic Riccati equation. *Numer. Linear Algebra Appl.*, **12**, 191–200.

MATHWORKS. (2002) *MATLAB User's Guide*. Natick, MA: MathWorks.

MEHRMANN, V. & XU, H. (2008) Explicit solutions for a Riccati equation from transport theory. *SIAM J. Matrix Anal.*, **30**, 1339–1357.

WING, G. M. (1962) *An Introduction to Transport Theory*. New York: Wiley.