Explicit Solutions for a Riccati Equation from Transport Theory

Volker Mehrmann* Hongguo Xu†

November 19, 2007

In memoriam of Gene H. Golub.

Abstract

We derive formulas for the minimal positive solution of a particular non-symmetric Riccati equation arising in transport theory. The formulas are based on the eigenvalues of an associated matrix. We use the formulas to explore some new properties of the minimal positive solution and to derive fast and highly accurate numerical methods. Some numerical tests demonstrate the properties of the new methods.

Keywords. non-symmetric Riccati equation, secular equation, eigenvalues, minimal positive solution, Cauchy matrix, transport theory, quadrature formula
AMS subject classification. 15A24, 65F15, 82C70, 65H05.

1 Introduction

We consider non-symmetric matrix Riccati equations of the special form

$$XA + DX - XBX - C = 0, \tag{1}$$

with

$$A = \Gamma - pe^T, \quad D = \Delta - ep^T, \quad B = pp^T, \quad C = ee^T$$

*Institut für Mathematik, TU Berlin, Str. des 17. Juni 136, D-10623 Berlin, FRG. `mehrmann@math.tu-berlin.de`. Partially supported by *Deutsche Forschungsgemeinschaft*, through the DFG Research Center MATHEON *Mathematics for Key Technologies* in Berlin.

†Department of Mathematics, University of Kansas, Lawrence, KS 44045, USA. Partially supported by *the University of Kansas General Research Fund allocation # 2301717* and by *Deutsche Forschungsgemeinschaft* through the DFG Research Center MATHEON *Mathematics for key technologies* in Berlin. *Part of the work was done while this author was visiting TU Berlin whose hospitality is gratefully acknowledged.*

where

$$\Gamma := \mathrm{diag}(\gamma_1, \dots, \gamma_n), \quad \Delta := \mathrm{diag}(\delta_1, \dots, \delta_n),$$
$$p = [p_1, \dots, p_n]^T, \quad e = [1, \dots, 1]^T,$$

and $\gamma_n > \dots > \gamma_1 > 0$, $\delta_n > \dots, \delta_1 > 0$, and $p_1, \dots, p_n > 0$.

Such Riccati equations arise in Markov models [27] and in nuclear physics [6, 16, 20]. In the latter application, to study the transport of particles, one introduces integral equations of the form

$$\left[\frac{1}{x+\alpha} + \frac{1}{y-\alpha}\right] T(x,y) = \beta \left[1 + \frac{1}{2}\int_{-\alpha}^{1} \frac{T(t,y)}{t+\alpha} dt\right] \left[1 + \frac{1}{2}\int_{\alpha}^{1} \frac{T(x,t)}{t-\alpha} dt\right]. \tag{2}$$

where the unknown function $T(x,y) : [-\alpha, 1] \times [\alpha, 1] \mapsto \mathbb{R}^+$ is called the *scattering function*, $\alpha \in [0, 1)$ is an angular shift, and $\beta \in [0, 1]$ is the average of the total number of particles emerging from a collision. (Here $\mathbb{R}^+$ denotes the set of positive real numbers.)

To solve this integral equation numerically, one approximates the integrals via classical quadrature formulas [28]. For this the function $T(x, y)$ is approximated via a matrix $X = [x_{ij}]$, where x_{ij} is an approximation of $T(\mu_i, \nu_j)$ with μ_i, ν_j being the ith and jth nodes of the quadrature formula on $[-\alpha, 1]$ and $[\alpha, 1]$, respectively, e.g. [16].

In this discretization the matrix X has to satisfy the matrix Riccati equation (1) with coefficient matrices

$$\gamma_j = \frac{1}{\beta(1-\alpha)\omega_j}, \quad \delta_j = \frac{1}{\beta(1+\alpha)\omega_j}, \quad p_j = \frac{c_j}{2\omega_j}, \tag{3}$$

for $j = 1, 2, \dots, n$, where $\{c_j\}_{j=1}^n$, $\{w_j\}_{j=1}^n$ are the sets of weights and nodes of the specific quadrature rule that is used on the interval $[0, 1]$. These typically satisfy

$$c_1, \dots, c_n > 0, \quad \sum_{j=1}^{n} c_j = 1; \qquad 1 > \omega_1 > \dots > \omega_n > 0. \tag{4}$$

In [18] it is shown that the Riccati equation (1) has two entry-wise positive solutions $X = [x_{ij}], Y = [y_{ij}] \in \mathbb{R}^{n,n}$, which satisfy $X \leq Y$, where we use the notation that $X \leq Y$ if $x_{ij} \leq y_{ij}$ for all $i, j = 1, \dots, n$.

In the applications from transport theory only X, the smaller one of the two positive solutions is of interest. Therefore, in this paper we only consider the computation of the minimal positive solution X. The computation of this minimal solution has been investigated in several publications. Various direct and iterative methods [1, 10, 11, 12, 13, 14, 15, 17, 16, 24] have been proposed by either directly solving the Riccati equation or by computing specific invariant subspaces of the $2n \times 2n$ matrix

$$H = \begin{bmatrix} A & -B \\ C & -D \end{bmatrix} \tag{5}$$

that is formed from the coefficient matrices.

In [18] even an explicit solution formula has been derived that is based on the eigenvalues H. Motivated by this result, we derive different explicit formulas, one of which is mathematically equivalent to the one in [18], but of a much simpler form. We will use these formulas to derive both entry-wise and norm-wise bounds for the solution matrix and show that the entries of the solution have a graded entry property. We will also use the formulas to develop fast and highly accurate numerical algorithms for the minimal positive solution of (1).

The paper is organized in follows. In Section 2, we will reformulate the associated eigenvalue problem via an appropriate balancing strategy. We use the associated secular function to derive some properties of the eigenvalues of H. In Section 3 we then derive four formulas for the minimal positive solution based on the eigenvalues. Entry-wise and norm-wise bounds for the minimal positive solution are provided in Section 4. Numerical algorithms and an error analysis are presented in Section 5 and some numerical examples are shown in Section 6. A conclusion is given in Section 7.

Throughout the paper, $\lambda(A)$ denotes the spectrum of a square matrix A, I_n (or simply I) is the $n \times n$ identity matrix. The norm used in this paper is the spectral norm.

2 Spectral properties of the matrix H

In this section we analyze the spectral properties of the matrix H in (5) defined by the coefficient matrices of (1).

In order for all the eigenvalues of H to be real, we assume that the condition

$$1 - \sum_{j=1}^{n} p_j \left(\frac{1}{\gamma_j} + \frac{1}{\delta_j} \right) \geq 0 \tag{6}$$

holds, which follows directly from the definition of the coefficients in (3) and (4).

The first step in our analysis is a balancing of the coefficient matrices. Since the entries of the vector p are positive, we may define

$$\Phi := \mathrm{diag}(\sqrt{p_1}, \ldots, \sqrt{p_n}), \qquad \phi := [\sqrt{p_1}, \ldots, \sqrt{p_n}]^T.$$

Using Φ to scale the Riccati equation (1) via

$$\begin{aligned} \tilde{X} &= \Phi X \Phi \\ \tilde{A} &= \Phi^{-1} A \Phi = \Gamma - \phi\phi^T, \\ \tilde{D} &= \Phi D \Phi^{-1} = \Delta - \phi\phi^T, \\ \tilde{B} &= \Phi^{-1} B \Phi^{-1} = \phi\phi^T, \\ \tilde{C} &= \Phi C \Phi = \phi\phi^T = \tilde{B}, \end{aligned}$$

we obtain the equivalent Riccati equation

$$\tilde{X}\tilde{A} + \tilde{D}\tilde{X} - \tilde{X}\tilde{B}\tilde{X} - \tilde{B} = 0, \tag{7}$$

and obviously, X is a solution to (1) if and only if $\tilde{X} = \Phi X \Phi$ is a solution to (7). For the associated matrix formed from the coefficients we then have

$$\begin{aligned} \tilde{H} &= \begin{bmatrix} \Phi^{-1} & 0 \\ 0 & \Phi \end{bmatrix} H \begin{bmatrix} \Phi & 0 \\ 0 & \Phi^{-1} \end{bmatrix} \\ &= \begin{bmatrix} \tilde{A} & -\tilde{B} \\ \tilde{B} & -\tilde{D} \end{bmatrix} = \begin{bmatrix} \Gamma & 0 \\ 0 & -\Delta \end{bmatrix} - \begin{bmatrix} \phi \\ -\phi \end{bmatrix} \begin{bmatrix} \phi \\ \phi \end{bmatrix}^T, \end{aligned} \tag{8}$$

and we see that $\tilde{H}$ is similar to H and it is a rank one modification of a diagonal matrix, which is analogous to the real symmetric rank-one updating problem discussed in [7]. It follows that the eigenvalues of $\tilde{H}$ can be obtained cheaply and accurately via the solution of secular equations by using a method similar to the one discussed in [8, Sec. 8.5].

It is furthermore well-known, see e.g. [21], that $\tilde{X}$ is a solution to (7) if and only if $\tilde{X}$ satisfies the invariant subspace equation

$$\tilde{H} \begin{bmatrix} I \\ \tilde{X} \end{bmatrix} = \begin{bmatrix} I \\ \tilde{X} \end{bmatrix} (\tilde{A} - \tilde{B}\tilde{X}).$$

In [18] it was shown (for the original solution X) that $\tilde{X}$ is the minimal positive solution if and only if all the eigenvalues of $\tilde{A} - \tilde{B}\tilde{X}$ are nonnegative.

In order to analyze the properties of the matrix $\tilde{H}$ and thus also of the similar matrix H, we first derive some properties of the eigenvalues of $\tilde{H}$.

Consider the rational function

$$\chi(\lambda) = 1 + \sum_{j=1}^{n} \frac{p_j}{\lambda - \gamma_j} - \sum_{j=1}^{n} \frac{p_j}{\lambda + \delta_j}. \tag{9}$$

Then, since

$$\det(\lambda I - \tilde{H}) = \chi(\lambda) \left(\prod_{j=1}^{n} (\lambda - \gamma_j)(\lambda + \delta_j) \right), \tag{10}$$

it follows that the eigenvalues of $\tilde{H}$ are just the roots of the *secular equation* $\chi(\lambda) = 0$ and thus the computation of the spectrum of $\tilde{H}$ can be obtained very efficiently by solving the secular equation. Furthermore, we have the following interlacing properties.

Lemma 2.1 *Consider the matrix $\tilde{H}$ defined via the coefficients of the Riccati equation (7) and suppose that (6) holds. Then $\tilde{H}$ has $2n$ real eigenvalues, $-\nu_1 < \ldots < -\nu_n \leq 0$, $0 \leq \lambda_1 < \ldots < \lambda_n$ that satisfy the inequalities*

$$0 \leq \nu_1 < \delta_1 < \nu_2 < \delta_2 < \ldots < \nu_{n-1} < \delta_{n-1} < \nu_n < \delta_n,$$

and

$$0 \leq \lambda_1 < \gamma_1 < \lambda_2 < \gamma_2 < \ldots < \lambda_{n-1} < \gamma_{n-1} < \lambda_n < \gamma_n.$$

Moreover, the following cases can be considered.

1. *$\nu_1 = 0$ and $\lambda_1 > 0$ if and only if $\chi(0) = 0$ and $\chi'(0) > 0$.*
2. *$\nu_1 < 0$ and $\lambda_1 = 0$ if and only if $\chi(0) = 0$ and $\chi'(0) < 0$.*
3. *$\nu_1 = \lambda_1 = 0$ if and only if $\chi(0) = \chi'(0) = 0$. In this case, $\tilde{H}$ has a 2×2 Jordan block associated with the eigenvalue 0.*

Proof. The proof is basically given already in [18] based on the properties of the secular function $\chi(\lambda)$. Note that assumption (6) implies that $\chi(0) \geq 0$.

It remains to show that in the third case, if 0 is a double eigenvalue of $\tilde{H}$, then it has geometric multiplicity 1. Let $x \in \mathbb{R}^{2n,2n} \backslash \{0\}$ be in the kernel of $\tilde{H}$, i.e.

$$\tilde{H}x = 0,$$

then

$$\begin{bmatrix} \Gamma & 0 \\ 0 & -\Delta \end{bmatrix} x = \zeta \begin{bmatrix} -\phi \\ \phi \end{bmatrix}, \qquad \text{with } \zeta = [\phi^T, \phi^T]x.$$

Therefore, x has the form

$$x = -\zeta \begin{bmatrix} \Gamma^{-1}\phi \\ \Delta^{-1}\phi \end{bmatrix}.$$

This shows that the eigenspace corresponding to 0 is one-dimensional and hence the geometric multiplicity of 0 must be one. $\square$

Remark 2.2 Suppose the quadrature formula that is used to discretize the integral equation (2) is of order greater than or equal to 3, i.e.,

$$\sum_{j=1}^{n} c_j w_j^k = \frac{1}{k+1}, \qquad k = 0, 1, 2, 3.$$

With (3) it is easily verified that

$$\begin{aligned}
\chi(0) &= 1 - \sum_{j=1}^{n} \left(\frac{p_j}{\gamma_j} + \frac{p_j}{\delta_j} \right) = 1 - \beta \sum_{j=1}^{n} c_j = 1 - \beta, \\
\chi'(0) &= \sum_{j=1}^{n} \left(-\frac{p_j}{\gamma_j^2} + \frac{p_j}{\delta_j^2} \right) = 2\alpha\beta^2 \sum_{j=1}^{n} c_j w_j = \alpha\beta^2, \\
\chi''(0) &= -2 \sum_{j=1}^{n} \left(\frac{p_j}{\gamma_j^3} + \frac{p_j}{\delta_j^3} \right) = -2(1+3\alpha^2)\beta^3 \sum_{j=1}^{n} c_j w_j^2 = -\frac{2}{3}(1+3\alpha^2)\beta^3, \\
\chi'''(0) &= 6 \sum_{j=1}^{n} \left(-\frac{p_j}{\gamma_j^4} + \frac{p_j}{\delta_j^4} \right) = 24\alpha(1+\alpha^2)\beta^4 \sum_{j=1}^{n} c_j w_j^3 = 6\alpha(1+\alpha^2)\beta^4.
\end{aligned}$$

Since $\chi'(0) \geq 0$, we have that Case 1. in Lemma 2.1 happens when $\beta = 1$ and $\alpha > 0$ and Case 3. happens when $\beta = 1$ and $\alpha = 0$. Case 2. will never happen.

3 Formulas for the minimal positive solution

In this section we will derive explicit formulas for the minimal positive solution of (1) in terms of the eigenvalues $-\nu_1, \ldots, -\nu_n, \lambda_1, \ldots, \lambda_n$ of H (or $\tilde{H}$). For this we need the following lemma.

Lemma 3.1 *Suppose in the following that $\tilde{X} \in \mathbb{R}^{n,n}$. The following statements are equivalent.*

(a) $\tilde{X}$ is the minimal positive solution of (7).

(b) $\tilde{X}$ *satisfies*

$$\tilde{H}\begin{bmatrix} I_n \\ \tilde{X} \end{bmatrix} = \begin{bmatrix} I_n \\ \tilde{X} \end{bmatrix}\tilde{R}_1,$$

where $\tilde{R}_1 = \tilde{A} - \tilde{B}\tilde{X}$ *and* $\sigma(\tilde{R}_1) = \{\lambda_1, \ldots, \lambda_n\}$.

(c) $\tilde{X}^T$ *is the minimal positive solution to the dual Riccati equation*

$$\tilde{Y}\tilde{D} + \tilde{A}\tilde{Y} - \tilde{Y}\tilde{B}\tilde{Y} - \tilde{B} = 0. \tag{11}$$

(d) $\tilde{X}$ *satisfies*

$$\tilde{H}\begin{bmatrix} \tilde{X}^T \\ I_n \end{bmatrix} = \begin{bmatrix} \tilde{X}^T \\ I_n \end{bmatrix}\tilde{R}_2, \tag{12}$$

where $\tilde{R}_2 = -(\tilde{D} - \tilde{B}\tilde{X}^T)$ *and* $\sigma(\tilde{R}_2) = \{-\nu_1, \ldots, -\nu_n\}$.

Proof. The equivalence of (a) and (b) is given in [18].

The equivalence between (a) and (c) is obvious by taking the transpose on both sides of (7) or (11).

Just as the relation between (a) and (b), $\tilde{X}^T$ is the minimal positive solution of (11) if and only if

$$\begin{bmatrix} \tilde{D} & -\tilde{B} \\ \tilde{B} & -\tilde{A} \end{bmatrix}\begin{bmatrix} I \\ \tilde{X}^T \end{bmatrix} = \begin{bmatrix} I \\ \tilde{X}^T \end{bmatrix}(\tilde{D} - \tilde{B}\tilde{X}^T), \tag{13}$$

and the eigenvalues of $\tilde{D}-\tilde{B}\tilde{X}^T$ are the rightmost n eigenvalues of $\begin{bmatrix} \tilde{D} & -\tilde{B} \\ \tilde{B} & -\tilde{A} \end{bmatrix}$. Identity (13) can be written as

$$\begin{bmatrix} -\tilde{A} & \tilde{B} \\ -\tilde{B} & \tilde{D} \end{bmatrix}\begin{bmatrix} \tilde{X}^T \\ I \end{bmatrix} = \begin{bmatrix} \tilde{X}^T \\ I \end{bmatrix}(\tilde{D} - \tilde{B}\tilde{X}^T).$$

Since

$$\begin{bmatrix} -\tilde{A} & \tilde{B} \\ -\tilde{B} & \tilde{D} \end{bmatrix} = -\tilde{H},$$

we have

$$\tilde{H}\begin{bmatrix} \tilde{X}^T \\ I \end{bmatrix} = \begin{bmatrix} \tilde{X}^T \\ I \end{bmatrix}\tilde{R}_2, \quad \tilde{R}_2 = -(\tilde{D} - \tilde{B}\tilde{X}^T),$$

which is (12). Clearly, the eigenvalues of $\tilde{R}_2$ are the n leftmost eigenvalues of $\tilde{H}$, which are $-\nu_1, \ldots, -\nu_n$. This shows the equivalence between (c) and (d). $\square$

With formulas for $\tilde{R}_1, \tilde{R}_2$ as in Lemma 3.1 and the formulas for $\tilde{A}$, $\tilde{D}$ and $\tilde{B}$, it follows that the minimal positive solution $\tilde{X}$ of (7) satisfies the following relations.

$$\Gamma - \phi\tilde{\xi}^T = \tilde{R}_1, \qquad \sigma(\tilde{R}_1) = \{\lambda_1, \ldots, \lambda_n\}, \tag{14}$$

$$\Delta - \phi\tilde{\eta}^T = -\tilde{R}_2, \qquad \sigma(-\tilde{R}_2) = \{\nu_1, \ldots, \nu_n\}, \tag{15}$$

$$\tilde{X}\Gamma + \Delta\tilde{X} = \tilde{\eta}\tilde{\xi}^T, \tag{16}$$

where

$$\tilde{\xi} = (I + \tilde{X}^T)\phi, \qquad \tilde{\eta} = (I + \tilde{X})\phi.$$

The last equation is a reformulation of (7). It thus follows that if the vectors $\tilde{\xi}$ an $\tilde{\eta}$ can be determined, then $\tilde{X}$ can be easily formulated based on the simple Sylvester equation (16).

The following result shows that $\tilde{\xi}$ and $\tilde{\eta}$ can be determined based on the relations (14) and (15).

Proposition 3.2 ([25]) *Suppose that matrices A, B are given such that $A = \mathrm{diag}(a_1, \ldots, a_n)$ with distinct diagonal entries $a_1, \ldots, a_n \in \mathbb{R}$, and $B \in \mathbb{R}^{n,n}$ with $\lambda(B) = \{b_1, \ldots, b_n\}$ for distinct $b_1, \ldots, b_n \in \mathbb{R}$.*

Let $q_1, q_2, \ldots, q_n \in \mathbb{R} \setminus \{0\}$ and define

$$q = [q_1, q_2, \ldots, q_n]^T, \quad Q = \mathrm{diag}(q_1, q_2, \ldots, q_n)$$

as well as

$$f = \left[\frac{\prod_{j=1}^{n}(a_1 - b_j)}{\prod_{j\neq 1}(a_1 - a_j)}, \ldots, \frac{\prod_{j=1}^{n}(a_k - b_j)}{\prod_{j\neq k}(a_k - a_j)}, \ldots, \frac{\prod_{j=1}^{n}(a_n - b_j)}{\prod_{j\neq n}(a_n - a_j)} \right]^T.$$

If a vector $z \in \mathbb{R}^n$ satisfies $A - qz^T = B$, then

$$z = Q^{-1}f = \left[\frac{f_1}{q_1}, \ldots, \frac{f_n}{q_n}\right]^T. \tag{17}$$

Using (17), (14), (15), and (16), we obtain the following explicit formulas for X.

Theorem 3.3 *Consider the Riccati equation (1). Introduce for $k = 1, \dots, n$ the scalar quantities*

$$\xi_k = \frac{\prod_{j=1}^{n}(\gamma_k - \lambda_j)}{\prod_{j \neq k}(\gamma_k - \gamma_j)}, \quad \eta_k = \frac{\prod_{j=1}^{n}(\delta_k - \nu_j)}{\prod_{j \neq k}(\delta_k - \delta_j)}, \quad \kappa_k = \frac{\prod_{j=1}^{n}(\gamma_k + \delta_j)}{\prod_{j=1}^{n}(\gamma_k + \nu_j)}, \quad \epsilon_k = \frac{\prod_{j=1}^{n}(\delta_k + \gamma_j)}{\prod_{j=1}^{n}(\delta_k + \lambda_j)},$$

the associated vectors and matrices

$$\begin{aligned}
\xi &= [\xi_1, \dots, \xi_n]^T, & \Xi = \operatorname{diag}(\xi_1, \dots, \xi_n), \\
\eta &= [\eta_1, \dots, \eta_n]^T, & E = \operatorname{diag}(\eta_1, \dots, \eta_n), \\
\kappa &= [\kappa_1, \dots, \kappa_n]^T, & K = \operatorname{diag}(\kappa_1, \dots, \kappa_n), \\
\epsilon &= [\epsilon_1, \dots, \epsilon_n]^T, & \mathcal{E} = \operatorname{diag}(\epsilon_1, \dots, \epsilon_n),
\end{aligned} \tag{18}$$

and the Cauchy matrix

$$\Theta = \left[\frac{1}{\delta_i + \gamma_j}\right].$$

Let

$$P = \operatorname{diag}(p_1, \dots, p_n),$$

with the p_i defined in (1). Then we have the following solution formulas for (1).

$$X = P^{-1} E \Theta \Xi P^{-1}, \tag{19}$$
$$X = P^{-1} E \Theta K, \tag{20}$$
$$X = \mathcal{E} \Theta \Xi P^{-1}, \tag{21}$$
$$X = \mathcal{E} \Theta K. \tag{22}$$

Proof. To prove the formulas, we apply Proposition 3.2 to (14) and obtain

$$\tilde{\xi} = \Phi^{-1} \xi,$$

where ξ is defined in (18). Similarly, from (15) we obtain

$$\tilde{\eta} = \Phi^{-1} \eta,$$

where η is defined in (18). By solving the Sylvester equation (15) we obtain

$$\tilde{X} = \Phi^{-1} E \Theta \Xi \Phi^{-1},$$

with E, Ξ as in (18). Then, (19) follows by using $X = \Phi^{-1}\tilde{X}\Phi^{-1}$ and $P = \Phi^2$.

In order to get the other formulas we only need to show that $\Xi = PK$ and $E = P\mathcal{E}$.

Since $-\nu_1, \ldots, -\nu_n, \lambda_1, \ldots, \lambda_n$ are the eigenvalues of $\tilde{H}$, it follows from (10) that

$$\prod_{j=1}^{n}(\lambda-\lambda_j)\prod_{j=1}^{n}(\lambda+\nu_j) = \sum_{m=1}^{n} p_m \prod_{j\neq m}(\lambda-\gamma_j)\prod_{j=1}^{n}(\lambda+\delta_j)$$
$$-\sum_{m=1}^{n} p_m \prod_{j=1}^{n}(\lambda-\gamma_j)\prod_{j\neq m}(\lambda+\delta_j) + \prod_{j=1}^{n}(\lambda-\gamma_j)\prod_{j=1}^{n}(\lambda+\delta_j). \quad (23)$$

By inserting $\lambda = \gamma_k$, we obtain

$$\prod_{j=1}^{n}(\gamma_k-\lambda_j)\prod_{j=1}^{n}(\gamma_k+\nu_j) = p_k \prod_{j\neq k}(\gamma_k-\gamma_j)\prod_{j=1}^{n}(\gamma_k+\delta_j),$$

which implies that

$$\xi_k = p_k\kappa_k, \quad k = 1, 2, \ldots, n.$$

We then have $\Xi = PK$.

Similarly, by inserting $\lambda = -\delta_k$ in (23) we get

$$\eta_k = p_k\epsilon_k, \quad k = 1, \ldots, n,$$

and thus $E = P\mathcal{E}$. Then the other formulas follow. □

Note that formula (20) only needs the eigenvalues $\nu_1, \ldots, \nu_n$, while formula (21) only needs the eigenvalues $\lambda_1, \ldots, \lambda_n$. Numerically, these two formulas provide very cheap procedures to compute the minimal solution X of (1).

Remark 3.4 In [18] already an explicit formula for the minimal solution of (1) was given that is equivalent to (21). However, there a different expression for ϵ_k was introduced as

$$\epsilon_k = 1 + \sum_{m=1}^{n} \frac{1}{\delta_k+\lambda_m} \frac{\prod_{j=1}^{n}(\gamma_j-\lambda_m)}{\prod_{j\neq m}(\lambda_j-\lambda_m)}.$$

This expression is less compact and its evaluation has a higher complexity than the expression in Theorem 3.3.

In this section we have derived new explicit formulas for the minimal solution X of (1) and we will use them in the next section to derive some further properties of X.

4 Properties and bounds for the minimal positive solution

The simple expressions of the quantities $\xi_k, \kappa_k, \eta_k, \epsilon_k$ in the explicit formulas (19)–(22) and the eigenvalue interlacing property for the eigenvalues of $\tilde{H}$ allow to derive further properties of the minimal positive solution of (1). For this we first prove the following Lemma.

Lemma 4.1 *The coefficients γ_k, δ_k in (1), the eigenvalues ν_k, λ_k of $\tilde{H}$ in (8)and the quantities $\xi_k, \eta_k, \kappa_k, \epsilon_k$, $k = 1, \ldots, n$ in (18) satisfy the following inequalities.*

1.

$$
\begin{aligned}
&0 < a_k < \eta_k < \delta_k - \nu_1 \le \delta_k, \qquad &&0 < b_k < \xi_k < \gamma_k - \lambda_1 \le \gamma_k,\\
&1 < \epsilon_k < \frac{\delta_k + \gamma_n}{\delta_k + \lambda_1} \le \frac{\delta_k + \gamma_n}{\delta_k}, \qquad &&1 < \kappa_k < \frac{\gamma_k + \delta_n}{\gamma_k + \nu_1} \le \frac{\gamma_k + \delta_n}{\gamma_k},
\end{aligned}
$$

where

$$
\begin{aligned}
a_k &= \begin{cases} \frac{(\delta_k - \nu_k)(\nu_{k+1} - \delta_k)}{\delta_n - \delta_k} & 1 \le k < n, \\ \delta_n - \nu_n & k = n, \end{cases}\\
b_k &= \begin{cases} \frac{(\gamma_k - \lambda_k)(\lambda_{k+1} - \gamma_k)}{\gamma_n - \gamma_k} & 1 \le k < n, \\ \gamma_n - \lambda_n & k = n. \end{cases}
\end{aligned}
$$

2.

$$
1 < \epsilon_n < \epsilon_{n-1} < \ldots < \epsilon_1, \qquad 1 < \kappa_n < \kappa_{n-1} < \ldots < \kappa_1.
$$

Proof. To prove the first part, we use the interlacing property in Lemma 2.1, and obtain

$$
0 < \frac{\delta_k - \nu_j}{\delta_k - \delta_{j-1}} < 1, \quad 1 < j \le k; \qquad \frac{\delta_k - \nu_j}{\delta_k - \delta_j} > 1, \quad 1 \le j < k
$$

and

$$
0 < \frac{\delta_k - \nu_j}{\delta_k - \delta_j} < 1, \quad k < j \le n; \qquad \frac{\delta_k - \nu_{j+1}}{\delta_k - \delta_j} > 1, \quad k < j < n.
$$

For $1 \leq k < n$, then

$$\eta_k = \frac{(\delta_k - \nu_k)(\delta_k - \nu_{k+1})}{\delta_k - \delta_n} \prod_{j=1}^{k-1} \frac{\delta_k - \nu_j}{\delta_k - \delta_j} \prod_{j=k+1}^{n-1} \frac{\delta_k - \nu_{j+1}}{\delta_k - \delta_j} > a_k,$$

and

$$\eta_k = (\delta_k - \nu_1) \prod_{j=1}^{k-1} \frac{\delta_k - \nu_{j+1}}{\delta_k - \delta_j} \prod_{j=k+1}^{n} \frac{\delta_k - \nu_j}{\delta_k - \delta_j} < \delta_k - \nu_1 \leq \delta_k.$$

Finally, for $k = n$, we obtain

$$\eta_n = (\delta_n - \nu_n) \prod_{j=1}^{n-1} \frac{\delta_n - \nu_j}{\delta_n - \delta_j} > \delta_n - \nu_n =: a_n,$$

and

$$\eta_n = (\delta_n - \nu_1) \prod_{j=1}^{n-1} \frac{\delta_n - \nu_{j+1}}{\delta_n - \delta_j} < \delta_n - \nu_1 \leq \delta_n.$$

This proves the inequalities for the η_k and clearly we have $a_k > 0$ for $k = 1, \ldots, n$.

The inequalities for the ξ_k can be derived in the same way by using the interlacing property for the eigenvalues $\lambda_1, \ldots, \lambda_n$. This interlacing property also gives

$$\epsilon_k = \prod_{j=1}^{n} \frac{\delta_k + \gamma_j}{\delta_k + \lambda_j} > 1,$$

and

$$\epsilon_k = \frac{\delta_k + \gamma_n}{\delta_k + \lambda_1} \prod_{j=1}^{n-1} \frac{\delta_k + \gamma_j}{\delta_k + \lambda_{j+1}} < \frac{\delta_k + \gamma_n}{\delta_k + \lambda_1} \leq \frac{\delta_k + \gamma_n}{\delta_k}.$$

Similarly, one can prove the inequalities for κ_k.

To prove part 2. we consider the function

$$\psi(t) = \prod_{j=1}^{n} \frac{t + \gamma_j}{t + \lambda_j} = \prod_{j=1}^{n} \left(1 + \frac{\gamma_j - \lambda_j}{t + \lambda_j}\right).$$

Since $\gamma_j - \lambda_j \geq 0$ for $j = 1, \ldots, n$, it follows that $\psi(t)$ is decreasing as t increases. Since $\psi(\delta_k) = \epsilon_k$ for $k = 1, \ldots, n$, and $\delta_1 < \ldots < \delta_n$, we thus have

$$\epsilon_1 > \epsilon_2 > \ldots > \epsilon_n.$$

Obviously $\psi(t) > 1$ for any $t > 0$ and hence $\epsilon_n = \psi(\delta_n) > 1$.

The monotonicity $\kappa_1 > \ldots > \kappa_n > 1$ follows in the same way. □

With the help of Lemma 4.1 we can now prove the following entry-wise monotonicity property of the minimal positive solution X of (1).

Theorem 4.2 *Let $X = [x_{ij}] \in \mathbb{R}^{n,n}$ be the minimal positive solution of (1). Then for any $i \geq k$ and $j \geq l$ with $(i, j) \neq (k, l)$, the entries of X satisfy*

$$x_{ij} > x_{kl}$$

Proof. Since

$$0 < \gamma_1 < \ldots < \gamma_n, \quad 0 < \delta_1 < \ldots < \delta_n,$$

and by Lemma 4.1,

$$1 < \epsilon_n < \ldots \epsilon_1, \quad 1 < \kappa_n < \ldots < \kappa_1,$$

with (22), for $1 \leq i, j \leq n$, if $i < n$, it follows that

$$x_{ij} = \frac{\epsilon_i \kappa_j}{\delta_i + \gamma_j} > \frac{\epsilon_{i+1} \kappa_j}{\delta_{i+1} + \gamma_j} = x_{i+1,j}.$$

If $j < n$, then

$$x_{ij} = \frac{\epsilon_i \kappa_j}{\delta_i + \gamma_j} > \frac{\epsilon_i \kappa_{j+1}}{\delta_i + \gamma_{j+1}} = x_{i,j+1}.$$

□

The quantities in Lemma 4.1 also provide upper and lower bounds for the entries of the minimal positive solution X of (1).

Theorem 4.3 *Let $X = [x_{ij}] \in \mathbb{R}^{n,n}$ be the minimal positive solution of (1). Then*

$$\frac{w_{ij}}{\delta_i + \gamma_j} < x_{ij} < \frac{W_{ij}}{\delta_i + \gamma_j},$$

where

$$\begin{aligned} w_{ij} &= \max\left\{\frac{a_i b_j}{p_i p_j}, \frac{a_i}{p_i}, \frac{b_j}{p_j}, 1\right\}, \\ W_{ij} &= \min\left\{\frac{\delta_i \gamma_j}{p_i p_j}, \frac{\delta_i(\gamma_j + \delta_n)}{p_i \gamma_j}, \frac{(\delta_i + \gamma_n)\gamma_j}{\delta_i p_j} \frac{(\delta_i + \gamma_n)(\gamma_j + \delta_n)}{\delta_i \gamma_j}\right\}. \end{aligned}$$

Proof. The bounds follow from the formulas (19) - (22) and the inequalities given in the first part of Lemma 4.1. □

Corollary 4.4 *Let $X = [x_{ij}] \in \mathbb{R}^{n,n}$ be the minimal positive solution of (1) and let w_{ij}, W_{ij} be as in Theorem 4.3. Then*

$$\frac{w_{nn}}{\delta_n + \gamma_n} < x_{nn} \leq x_{ij} \leq x_{11} < \frac{W_{11}}{\delta_1 + \gamma_1}$$

for $i, j = 1, \ldots, n$.

Proof. The inequalities follow from Theorems 4.2 and 4.3. □

We also obtain a bound for the spectral norm of the minimal positive solution X of (1).

Theorem 4.5 *Let $\tilde{X} \in \mathbb{R}^{n,n}$ be the minimal positive solution of (7). Then*

$$\|\tilde{X}\| \leq 1,$$

and $\|\tilde{X}\| = 1$ if and only if $\chi(0) = 0$ and $\chi'(0) = 0$.

Moreover, the minimal positive solution X of (1) satisfies

$$\|X\| \leq \frac{1}{\min_j p_j}.$$

Proof. Define the matrix function

$$\tilde{H}(t) = \begin{bmatrix} \Gamma & 0 \\ 0 & -\Delta \end{bmatrix} - t \begin{bmatrix} \phi \\ -\phi \end{bmatrix} \begin{bmatrix} \phi \\ \phi \end{bmatrix}^T$$

with $0 \leq t \leq 1$. Let $\chi_t(\lambda)$ be the corresponding secular function as in (10). Using the assumption (6), it follows that $\chi_t(0) > 0$ for $0 \leq t < 1$. So $\tilde{H}(t)$ has $2n$ real eigenvalues $-\nu_1(t), \ldots, -\nu_n(t)$ and $\lambda_1(t), \ldots, \lambda_n(t)$, and the same interlacing properties as in Lemma 2.1 hold, i.e.,

$$0 < \nu_1(t) < \delta_1 < \nu_2(t) < \delta_2 < \ldots < \delta_{n-1} < \nu_n(t) < \delta_n$$

and

$$0 < \lambda_1(t) < \gamma_1 < \lambda_2(t) < \gamma_2 < \ldots < \gamma_{n-1} < \lambda_n(t) < \gamma_n,$$

for $0 \leq t < 1$.

Since

$$\tilde{H}(1) = \tilde{H}, \qquad \tilde{H}(0) = \begin{bmatrix} \Gamma & 0 \\ 0 & -\Delta \end{bmatrix}.$$

we have that

$$\lambda_j(1) = \lambda_j, \quad \nu_j(1) = \nu_j; \qquad \lambda_j(0) = \gamma_j, \quad \nu_j(0) = \delta_j$$

for $j = 1, \ldots, n$.

Let $v_1(t), \ldots, v_n(t) \in \mathbb{R}^{2n}$ be the eigenvectors associated with $\lambda_1(t), \ldots, \lambda_n(t)$, respectively, and let

$$\tilde{V}(t) = [v_1(t),\ \ldots,\ v_n(t)].$$

Then $\tilde{V}(t)$ satisfies

$$\tilde{H}(t)\tilde{V}(t) = \tilde{V}(t)\Lambda(t), \qquad \Lambda(t) = \mathrm{diag}(\lambda_1(t), \ldots, \lambda_n(t)). \tag{24}$$

Because $\lambda_1(t), \ldots, \lambda_n(t)$ are distinct for $0 \le t < 1$, such a matrix $\tilde{V}(t)$ always exists and has full rank. Since $\{\lambda_1(t), \ldots, \lambda_n(t)\} \cap \{-\nu_1(t), \ldots, -\nu_n(t)\} = \emptyset$, we may construct it in such a way that $\tilde{V}(t)$ is a continuous function of t and $\tilde{V}(0) = \left[\begin{smallmatrix} I \\ 0 \end{smallmatrix}\right]$, see e.g. [29].

With

$$\Sigma = \begin{bmatrix} I_n & 0 \\ 0 & -I_n \end{bmatrix}.$$

it is easily verified that $\Sigma\tilde{H}(t)$ is real symmetric. By taking the transpose on both sides of (24) we have

$$(\Sigma\tilde{V}(t))^T\tilde{H}(t) = \Lambda(t)(\Sigma\tilde{V}(t))^T,$$

i.e., the columns of $\Sigma\tilde{V}(t)$ form a basis of the left invariant subspace associated with the eigenvalues $\{\lambda_1(t), \ldots, \lambda_n(t)\}$ of $\tilde{H}$. Then

$$S(t) := \tilde{V}(t)^T\Sigma\tilde{V}(t)$$

is nonsingular for $0 \le t < 1$. Because $\tilde{V}(0) = \left[\begin{smallmatrix} I \\ 0 \end{smallmatrix}\right]$, we have $S(0) = I$, which is positive definite. Then, since $S(t)$ is a continuous function of t and $\det S(t) \neq 0$, it follows that $S(t)$ is positive definite for $0 \le t < 1$. Thus, with the partition

$$\tilde{V}(t) = \begin{bmatrix} \tilde{V}_1(t) \\ \tilde{V}_2(t) \end{bmatrix}, \qquad \tilde{V}_1(t), \tilde{V}_2(t) \in \mathbb{R}^{n,n}$$

and using the relation

$$S(t) = \tilde{V}_1(t)^T\tilde{V}_1(t) - \tilde{V}_2(t)^T\tilde{V}_2(t),$$

it follows that $\tilde{V}_1(t)$ must be nonsingular, and $\tilde{X}(t) = \tilde{V}_2(t)\tilde{V}_1(t)^{-1}$ is the minimal positive solution of the Riccati equation of the from (7) associated with $\tilde{H}(t)$. Because $I - \tilde{X}(t)^T\tilde{X}(t)$ is also positive definite, we have $\|\tilde{X}(t)\| < 1$ for $0 \le t < 1$. By taking the limit $t \to 1$ we have $\|\tilde{X}\| \le 1$.

If $\|\tilde{X}\| = 1$ then $S(1)$ is singular. This implies

$$\{\lambda_1, \ldots, \lambda_n\} \cap \{\nu_1, \ldots, \nu_n\} \neq \emptyset.$$

But due to the interlacing properties for the eigenvalues, this happens only when $\lambda_1 = \nu_1 = 0$, i.e., when $\chi(0) = 0$ and $\chi'(0) = 0$. On the other hand, if $\chi(0) = 0$ and $\chi'(0) = 0$, then by Lemma 2.1, $\lambda_1 = \nu_1 = 0$ and 0 is a defective eigenvalue of $\tilde{H}$. In this case $S(1)$ must be singular, or equivalently $\|\tilde{X}\| = 1$.

The upper bound for $\|X\|$ follows from the relation $X = \Phi^{-1}\tilde{X}\Phi^{-1}$. □

Various lower bounds for $\|X\|$ can also be derived by using the inequalities for the entries of X, but we will not pursue this topic here.

In the end of this section we also provide a formula for the inverse of X.

Theorem 4.6 *The minimal positive solution $X = [x_{ij}]$ of (1) is invertible and with P, Θ as in Theorem 3.3, its inverse is given by*

$$X^{-1} = PQ\Theta^T GP,$$

where

$$Q = \operatorname{diag}(q_1, \ldots, q_n), \qquad G = \operatorname{diag}(g_1, \ldots, g_n),$$

with

$$q_k = \prod_{j=1}^{n} \frac{\gamma_k + \delta_j}{\gamma_k - \lambda_j}, \qquad g_k = \prod_{j=1}^{n} \frac{\delta_k + \gamma_j}{\delta_k - \nu_j},$$

for $k = 1, \ldots, n$.

Proof. Since $\gamma_n > \ldots > \gamma_1 > 0$ and $\delta_n > \ldots > \delta_1 > 0$, it follows (see e.g. [5]) that the Cauchy matrix Θ is invertible and

$$\Theta^{-1} = \hat{Q}\Theta^T\hat{G},$$

where

$$\hat{Q} = \operatorname{diag}(\hat{q}_1, \ldots, \hat{q}_n), \quad \hat{G} = \operatorname{diag}(\hat{g}_1, \ldots, \hat{g}_n),$$

with

$$\hat{q}_k = \frac{\prod_{j=1}^{n}(\gamma_k + \delta_j)}{\prod_{j \neq k}(\gamma_k - \gamma_j)}, \qquad \hat{g}_k = \frac{\prod_{j=1}^{n}(\delta_k + \gamma_j)}{\prod_{j \neq k}(\delta_k - \delta_j)},$$

for $k = 1, \ldots, n$. Since all the diagonal matrices in (19) are invertible, it follows that X is also invertible and the formula for X^{-1} follows from (19) using Θ^{-1}. □

5 Numerical algorithms

The formulas given in Section 3 can be used to develop the following numerical algorithms for computing the minimal positive solution of (1).

Algorithm 5.1 For the Riccati equation (1) this algorithm computes the minimal positive solution.

1. Compute the eigenvalues $\nu_1, \ldots, \nu_n, \lambda_1, \ldots, \lambda_n$ of $\tilde{H}$ in (8) by applying a root finding solver to the secular equation $\chi(\lambda) = 0$ given by (9).
2. Use either of the formulas (19) or (22) to compute the minimal positive solution X of (1).

We can also use the formula (20) or (21).

Algorithm 5.2 For the Riccati equation (1) this algorithm computes the minimal positive solution.

1. Compute the eigenvalues $\nu_1, \ldots, \nu_n$ of $\tilde{H}$ in (8) by applying a root finding solver to the secular equation $\chi(\lambda) = 0$ given by (9).
2. Use Formula (20) to compute the minimal positive solution X of (1).

Algorithm 5.3 For the Riccati equation (1) this algorithm computes the minimal positive solution.

1. Compute the eigenvalues $\lambda_1, \ldots, \lambda_n$ of $\tilde{H}$ in (8) by applying a secular equation solver to $\chi(\lambda) = 0$.
2. Use Formula (21) to compute the minimal positive solution X of (1).

Note that Algorithms 5.2 and 5.3 only need to computed half of the eigenvalues.

The success of these three algorithms depends on how fast and accurately the eigenvalues can be computed and how sensitive the evaluation of the formulas (19)–(22) is. This requires an efficient and reliable secular equation solver. The osculatory interpolation methods of [2, 23] that were developed in the context of the divide-and-conquer eigenvalue methods ([8, Sec. 8.5], [3, 4, 7]) may not be applicable directly, since the secular function $\chi(\lambda)$ has quite different properties than the secular equation derived in the symmetric divide-and-conquer method. For this reason we propose the following hybrid method for the computation of roots of the secular function. We only consider the case for computing the eigenvalues λ_k, the method for computing the eigenvalues ν_k is analogous. Our approach treats λ_1 differently than the other eigenvalues $\lambda_2, \ldots, \lambda_n$, because of the different properties that λ_1 has.

5.1 Computation of λ_k with $k > 1$.

1. Initial guess. To compute an initial guess, we basically follow the procedure suggested in [23]. We first evaluate $\chi(m_k)$, where m_k is the mid-point of the interval (γ_k, γ_{k+1}). Because $\chi(\lambda)$ has only one root in (γ_k, γ_{k+1}), and since $\lim_{\lambda\to\gamma_k^+}\chi(\lambda) = \infty$, and $\lim_{\lambda\to\gamma_{k+1}^-}\chi(\lambda) = -\infty$, based on the sign of $\chi(m_k)$, we can easily determine in which half of the interval λ_k is located. Simple geometry shows that if $\chi(m_k) > 0$ then λ_k is closer to γ_{k+1}, and if $\chi(m_k) < 0$ then λ_k is closer to γ_k. We then consider the equation
$$\frac{p_k}{\lambda-\gamma_k} + \frac{p_{k+1}}{\lambda-\gamma_{k+1}} + r_k = 0,$$
with right hand side $r_k = \chi(m_k) - p_k/(m_k-\gamma_k) - p_{k+1}/(m_k-\gamma_{k+1})$, which can be obtained during the evaluation of $\chi(m_k)$ without any extra cost. We then take the root of this equation in (γ_k, γ_{k+1}) as our initial guess z_k^0. It is easily verified that z_k^0 and λ_k are in the same half interval. We also choose an initial interval so that the χ values on end-points have opposite signs, (which guarantees that λ_k is in this interval). If $\chi(m_k)\chi(z_k^0) < 0$, then we use m_k, z_k^0 for the interval. Otherwise, we use the asymptotic properties of χ to find another λ value to replace m_k. Let us denote the resulting interval by $[u_0, v_0]$.

2. Iteration step. For a current approximation z_k^j, we first evaluate $\chi'(z_k^j)$ and use one step of Newton's method to determine the next approximate z_k^{j+1}. If z_k^{j+1} is inside the current interval $[u_j, v_j]$ we evaluate $\chi(z_k^{j+1})$. We then replace one of u_j, v_j and its corresponding χ value with z_k^{j+1} and $\chi(z_k^{j+1})$ based on the sign of $\chi(z_k^{j+1})$ and move on to the next iteration. If z_k^{j+1} is outside $[u_j, v_j]$ (maybe even outside of (γ_k, γ_{k+1})), then we apply one step of the secant method with u_j, v_j and their corresponding χ values to get z_k^{j+1}. We then evaluate $\chi(z_k^{j+1})$, update $[u_j, v_j]$, and continue. If this z_k^{j+1} is still outside of $[u_j, v_j]$ we use one step of the bisection method with u_j, v_j to get z_k^{j+1}.

   When the iterates z_k^j get close to the root λ_j, then due to rounding errors it becomes more difficult to compute a reliable value of $\chi(z_k^j)$. (This happens typically for small roots.) This may cause the sign of χ to alternate between positive and negative values in the Newton iteration and the secant iteration and may have the effect that the sequence $\{z_k^j\}$ does not converge. If we observe such a behavior and the function values for χ are also small in absolute value, then we run a step of the bisection method. This procedure has turned out to be very successful during our numerical tests.

3. Stopping criterion. In order to compute the root λ_k accurately, we actually use the shift $s = \lambda - \gamma_k$ or $s = \lambda - \gamma_{k+1}$ initially, depending on whether λ_k is closer to γ_k or γ_{k+1}. The iteration step is then applied to the new variable s to generate a sequence of approximate values $s_0, s_1, \ldots, s_j, \ldots$. The iteration can be written as

$$s_{j+1} = s_j + \Delta s_j,$$

where Δs_j is the jth correction.

We use the stopping criterion

$$|\Delta s_j| < c\varepsilon_M |s_{j+1}|, \tag{25}$$

where ε_M is the machine precision and c is a modest constant (which is set to 48 in our tests).

The procedure for the computation of ν_k $k = 2, \ldots, n$ is analogous.

5.2 Computation of λ_1

1. Initial guess. The strategy for choosing starting values z_1^0 and starting intervals $[u_0,\, v_0]$ is slightly different than in the case of the other eigenvalues. Since we know that $\lambda_1 \in [0, \gamma_1)$, we first evaluate $\chi(m_1)$, where $m_1 = \gamma_1/2$. We use the sign of $\chi(m_1)$ to determine if λ_1 is closer to 0 or γ_1. We then use the root $z_1^0 \in [0, \gamma)$ of the equation

$$\frac{p_1}{\lambda - \gamma_1} + r_1 = 0,$$

with $r_1 = \chi(m_1) - p_1/(m_1 - \gamma_1)$, as the initial starting value.

If $\chi(m_1)\chi(z_1^0) < 0$, then we use m_1, z_1^0 to form the initial interval $[u_0,\, v_0]$. If $\chi(m_1), \chi(z_1^0) > 0$, then we replace m_1 by another value such that the corresponding χ value is negative, by using the fact $\lim_{\lambda \to \gamma_1^-}(\lambda) = -\infty$. In the case that $\chi(m_1), \chi(z_1^0) < 0$, if $\chi(0) > 0$, we replace m_1 with 0. If $\chi(0) = 0$ we still need to check the sign of $\chi'(0)$. If $\chi'(0) > 0$ we may use it to find a small positive number such that its corresponding χ is positive. We then replace m_1 with this number. If $\chi'(0) \leq 0$, we simply set $\lambda_1 = 0$, and no iteration is required.

Note that for transport theory problem $\chi(0)$ and $\chi'(0)$ can be easily determined by the formulas given in Remark 2.2.

2. Iteration step. We first use the same iteration steps as described for the eigenvalues λ_k, $k \geq 2$ to an approximation of λ_1. This usually works well for $\lambda_1 > c_1\sqrt{\varepsilon_M}$ with some positive constant c_1. If, however, λ_1 is too small, then it is difficult to get accurate function values for χ and χ', which then may cause convergence problems. In order to overcome this difficulty, once we observe that the jth approximate z_1^j satisfies $z_1^j < c_1\sqrt{\varepsilon_M}$ (we used $c_1 = 100$ in our tests), we evaluate $\chi(z_1^j)$ and $\chi'(z_1^j)$ by using their corresponding Taylor polynomials at 0, given by

$$\begin{aligned}\chi(z_1^j) &\approx \chi(0) + z_1^j\chi'(0) + \frac{(z_1^j)^2}{2}\chi''(0)\\ \chi'(z_1^j) &\approx \chi'(0) + z_1^j\chi''(0) + \frac{(z_1^j)^2}{2}\chi'''(0)\end{aligned}$$

and use these values in the next step of the Newton iteration. If $\chi'(z_1^j)$ is also very small in modulus, then we approximate $\chi''(z_1^j)$ by

$$\chi''(z_1^j) \approx \chi''(0) + z_1^j\chi'''(0).$$

We then use the approximations for $\chi(z_1^j)$, $\chi'(z_1^j)$, $\chi''(z_1^j)$ to construct the second degree Taylor polynomial for χ at z_1^j, and use one of the roots of this polynomial (if it exists) as our next iterate z_1^{j+1}.

For a general secular equation, the computation of $\chi(0)$, $\chi'(0)$, $\chi''(0)$, and $\chi'''(0)$ requires extra cost and it is not clear if the values can be really evaluated accurately. In the secular equation from the transport problem, however, this computation is essentially cost-free since we may use the formulas in Remark 2.2, and because of the simple formulations the values can be computed accurately.

3. Stopping criterion. We use again the stopping criterion (25).

The procedure for the computation of ν_1 is analogous.

5.3 Costs.

The main cost in Algorithms 5.1–5.3 is the evaluation of χ and χ' during each iteration step. In order to evaluate $\chi(\lambda)$ and $\chi'(\lambda)$, we first compute $\lambda - \gamma_j$, $\lambda + \delta_j$ for $j = 1, \ldots, n$. We then compute $p_j/(\lambda - \gamma_j)$ and $p_j/(\lambda + \delta_j)$. After this $\chi(\lambda)$ can be evaluated. We continue to compute $[p_j/(\lambda - \gamma_j)]/(\lambda - \gamma_j)$ and $[p_j/(\lambda + \delta_j)]/(\lambda + \delta_j)]$, which costs one extra flop for each term and then evaluate $\chi'(\lambda)$. So if the Newton iteration is used in the iteration step, then the cost per iteration step and per eigenvalue is about $10n$ flops. If the average number of iterations is M, then the cost for Algorithm 5.1 is about $(20M + 9)n^2$ flops, and the cost for Algorithms 5.2 and 5.3 is about $(10M + 9)n^2$ flops. Note that it requires $3n^2$ flops to compute each set of the values $\xi_k, \eta_k, \kappa_k, \epsilon_k$, and it requires another $3n^2$ flops to compute the components of X. Note also that in these complexity estimates we did not count the cost for the computation of the initial values.

5.4 Error analysis.

To analyze the computational errors in the described procedures, we first estimate the errors in the computed eigenvalues. We assume that the iteration for each eigenvalue stops when (25) holds, and the computed sequence satisfies the conditions in the following lemma observed by Kahan (see e.g. [23]).

Lemma 5.4 *Let $\{x_j\}_{j=1}^{\infty}$ be a sequence of real numbers, produced by some rapidly convergent iteration scheme, such that $\lim_{j\to\infty} x_j = x^*$. If the sequence of ratios $\frac{|x_{j+1}-x_j|}{|x_j-x_{j-1}|}$ is decreasing for $j \geq k$, and if $\frac{|x_{k+1}-x_k|}{|x_k-x_{k-1}|} < 1$, then*

$$|x_{k+1} - x^*| < \frac{|x_{k+1} - x_k|^2}{|x_k - x_{k-1}| - |x_{k+1} - x_k|}.$$

Let λ_j, ν_j be the exact eigenvalues of H and let $\hat{\lambda}_j$, $\hat{\nu}_j$ be the corresponding computed eigenvalues. With the discussed properties of the eigenvalues, the presented procedures and Lemma 5.4, it is reasonable to assume that the computed eigenvalues satisfy

$$\begin{aligned} |\lambda_j - \hat{\lambda}_j| &< C_{\lambda_j}\varepsilon_M \min\{\gamma_{j+1} - \lambda_j,\ \lambda_j - \gamma_j\}, &(26)\\ |\nu_j - \hat{\nu}_j| &< C_{\nu_j}\varepsilon_M \min\{\delta_{j+1} - \nu_j,\ \nu_j - \delta_j\}, &(27)\end{aligned}$$

for $j = 1, \ldots, n$, where C_{λ_j}, C_{ν_j} are some modest constants. We then have the following Lemma.

Lemma 5.5 *Suppose that the computed eigenvalues $\hat{\lambda}_j$, $\hat{\nu}_j$ of H as in (5) satisfy (26) and (27). Let $\hat{\xi}_k$, $\hat{\eta}_k$, $\hat{\epsilon}_k$, $\hat{\kappa}_k$ be the computed quantities determined via the formulas given in Theorem 3.3. Then*

$$\begin{aligned} \hat{\xi}_k &= \xi_k(1 + nC_{\xi_k}\varepsilon_M), \quad \hat{\eta}_k = \eta_k(1 + nC_{\eta_k}\varepsilon_M),\\ \hat{\kappa}_k &= \kappa_k(1 + nC_{\kappa_k}\varepsilon_M), \quad \hat{\epsilon}_k = \epsilon_k(1 + nC_{\epsilon_k}\varepsilon_M),\end{aligned}$$

for $k = 1, \ldots, n$, where C_{ξ_k}, C_{η_k}, C_{κ_k}, C_{ϵ_k} are constants.

Proof. For the proof we just consider the first order error.

Note that $\hat{\xi}_k$ is actually computed by the formula

$$\prod_{j=1}^{n}(\gamma_k - \hat{\lambda}_j) / \prod_{j\neq k}(\gamma_k - \gamma_j),$$

i.e., λ_j is replaced with $\hat{\lambda}_j$. By (26),

$$\begin{aligned} |\gamma_k - \hat{\lambda}_j| &= |\gamma_k - \lambda_j + \varepsilon_M C_{kj} \min\{\gamma_{j+1} - \lambda_j,\, \lambda_j - \gamma_j\}| \\ &= |\gamma_k - \lambda_j| \left| 1 + C_{kj}\varepsilon_M \frac{\min\{\gamma_{j+1} - \lambda_j,\, \lambda_j - \gamma_j\}}{|\gamma_k - \lambda_j|} \right|. \end{aligned}$$

By the interlacing property of the eigenvalues we have

$$\frac{\min\{\gamma_{j+1} - \lambda_j,\, \lambda_j - \gamma_j\}}{|\gamma_k - \lambda_j|} \leq 1$$

for $j = 1, \ldots, n$ and hence

$$|\gamma_k - \hat{\lambda}_j| = |\gamma_k - \lambda_j|(1 + \tilde{C}_{kj}\varepsilon_M),$$

for some constant $\tilde{C}_{kj}$. With this relation, it is not difficult to obtain that

$$\hat{\xi}_k = \xi_k(1 + nC_{\xi_k}\varepsilon_M),$$

where C_{ξ_k} is a constant. The corresponding relations for the other terms follow in the same way. □

Using this Lemma we obtain the following relative errors for the components of the minimal positive solution computed by the formulas given in Section 3.

Theorem 5.6 *Consider the problem of computing the minimal positive solution $X = [x_{ij}]$ of (1) using formulas (19)–(22) and suppose that the computed eigenvalues satisfy the relations (26) and (27). Then for the computed solution $\hat{X} = [\hat{x}_{ij}]$, the relative error estimate*

$$\frac{|\hat{x}_{ij} - x_{ij}|}{x_{ij}} = D_{ij} n \varepsilon_M, \qquad i, j = 1, \ldots, n$$

holds, where D_{ij}'s are positive constants.

Proof. The relative error estimates follow from Lemma 5.5. □

6 Numerical Examples

In this section we present some numerical test results for the problems from transport theory, see [18, 19]. The weights $c_1, \ldots, c_n$ and nodes $\omega_1, \ldots, \omega_n$ are generated from the composite four-node Gauß-Legendre quadrature formula on [0, 1] with $n/4$ equally spaced subintervals, see e.g. [28]. All the numerical examples were tested in MATLAB version 7.1.0 with machine precision $\varepsilon_M \approx 2.22e-16$. We solved the problem for various numbers of the parameters α and β and the size n. We used all four formulas to compute the minimal positive solution, with a secular equation solver as described in Section 5.

The computed minimal positive solution via formulas (19)–(22) are denoted by $X^{(1)}, X^{(2)}, X^{(3)}, X^{(4)}$, respectively. In the following we display the test results. We present one table for each pair (α, β) and various values of n. (The used norm is always the spectral norm). In each table, we list the following results:

- Maximum residual:

$$R = \max_{j \in \{1,2,3,4\}} \|X^{(j)}\Gamma + \Delta X^{(j)} - (e + X^{(j)}p)(e^T + p^T X^{(j)})\|$$

- Maximum and minimum entry-wise relative errors:

$$RE_{\max} = \max_{\substack{i,j \in \{1,2,3,4\} \\ i \neq j}} \max_{k,l \in \{1,\ldots,n\}} \frac{|x_{kl}^{(i)} - x_{kl}^{(j)}|}{\min\{x_{kl}^{(i)}, x_{kl}^{(j)}\}}$$

$$RE_{\min} = \min_{\substack{i,j \in \{1,2,3,4\} \\ i \neq j}} \max_{k,l \in \{1,\ldots,n\}} \frac{|x_{kl}^{(i)} - x_{kl}^{(j)}|}{\min\{x_{kl}^{(i)}, x_{kl}^{(j)}\}}$$

- Largest entry x_{11} (determined by one of the four solutions)

- Smallest entry x_{nn} (determined by one of the four solutions)

- Norm $\|X\|$ (X is one of the four solutions). Note that we have proved that $\|\tilde{X}\| \leq 1$ which translates to $\|X\| \leq 1/\min p_j$.

- Number of iterations for ν_1: N_-

- Number of iterations for λ_1: N_+

- Average of number of iterations for all $2n$ eigenvalues: N

We also give the eigenvalues $-\nu_1, \lambda_1$ in the caption.

We can summarize the numerical results as follows.

1. The values of R in the tables are usually the residual of $X^{(1)}$. The other residuals are basically the same but some can be one order smaller.

2. Since we do not know the exact solution, we use $RE_{\max}$ and $RE_{\min}$ to detect if high relative accuracy can be actually achieved. The values of $RE_{\max}$ and $RE_{\min}$ do support the high relative accuracy result. (Note that x_{nn} is small in all examples.)

3. The number of iterations for ν_1 and λ_1 increases as $\alpha \to 0$ and $\beta \to 1$. This shows the numerical difficulty when the eigenvalues $-\nu_1$ and λ_1 are getting close to each other. However, our computed values of ν_1, λ_1 are much more accurate than those obtained by running the MATLAB code *eig* on $\tilde{H}$.

4. Our MATLAB implementation of the root finder based on the secular equation is still not very robust. In general, about .5% of the eigenvalues need 100 iterations, the maximum iteration number used in our experimental code. Some further improvement could improve these convergence properties.

n	R	$RE_{\max}$	$RE_{\min}$	x_{11}	x_{nn}	$\Vert X\Vert$	N_-	N_+	N
64	2.70e-13	1.83e-14	6.80e-15	.263	8.23e-04	7.87e+00	8	7	5
128	1.27e-12	6.72e-14	3.33e-14	.263	4.09e-04	1.57e+01	9	8	5
256	5.35e-12	1.64e-13	7.73e-14	.264	2.04e-04	3.15e+01	9	9	5
512	1.97e-11	2.70e-13	1.34e-13	.264	1.02e-04	6.29e+01	10	8	5

Table 1: $\alpha = 0.5, \quad \beta = .5, \quad (-\nu_1, \lambda_1) \approx (-1.166, 3.996)$

n	R	$RE_{\max}$	$RE_{\min}$	x_{11}	x_{nn}	$\Vert X\Vert$	N_-	N_+	N
64	5.16e-13	2.65e-14	1.23e-14	2.70	2.19e-03	6.12e+01	8	6	5
128	2.43e-12	9.67e-14	4.06e-14	2.72	1.08e-03	1.22e+02	10	5	5
256	8.48e-12	1.46e-13	7.03e-14	2.72	5.37e-04	2.45e+02	9	5	5
512	3.48e-11	4.21e-13	2.04e-13	2.72	2.67e-04	4.89e+02	10	6	6

Table 2: $\alpha = 0.1, \quad \beta = 0.99, \quad (-\nu_1, \lambda_1) \approx (-7.98e-02, 3.83e-01)$

n	R	$RE_{\max}$	$RE_{\min}$	x_{11}	x_{nn}	$\|X\|$	N_-	N_+	N
64	2.46e-11	1.48e-12	7.35e-13	4.19	2.24e-03	8.59e+01	23	16	5
128	1.02e-10	5.16e-12	2.57e-12	4.21	1.10e-03	1.72e+02	26	25	5
256	4.66e-11	1.24e-12	5.60e-13	4.22	5.48e-04	3.43e+02	19	25	5
512	5.43e-10	7.02e-12	3.48e-12	4.22	2.73e-04	6.87e+02	34	25	6

Table 3: $\alpha = 10^{-4}, \quad \beta = 1 - 10^{-8}, \quad (-\nu_1, \lambda_1) \approx (-7.91e-05, 3.79e-04)$

n	R	$RE_{\max}$	$RE_{\min}$	x_{11}	x_{nn}	$\|X\|$	N_-	N_+	N
64	6.09e-13	2.52e-14	1.02e-14	4.19	2.24e-03	8.59e+01	28	26	6
128	2.72e-12	7.80e-14	3.15e-14	4.21	1.10e-03	1.72e+02	28	26	5
256	1.02e-11	1.85e-13	8.30e-14	4.22	5.48e-04	3.44e+02	28	26	5
512	4.28e-11	4.12e-13	1.60e-13	4.22	2.73e-04	6.87e+02	28	26	6

Table 4: $\alpha = 10^{-14}, \quad \beta = 1 - 10^{-14}, \quad (-\nu_1, \lambda_1) \approx (-1.73e-07, 1.73e-07)$

n	R	$RE_{\max}$	$RE_{\min}$	x_{11}	x_{nn}	$\|X\|$	N_-	N_+	N
64	7.74e-13	4.84e-14	1.94e-14	4.19	2.24e-03	8.59e+01	0	30	5
128	2.95e-12	8.97e-14	4.07e-14	4.21	1.10e-03	1.72e+02	0	30	5
256	1.21e-11	1.76e-13	7.39e-14	4.22	5.48e-04	3.44e+02	0	32	5
512	4.51e-11	4.14e-13	1.87e-13	4.22	2.73e-04	6.87e+02	0	30	6

Table 5: $\alpha = 10^{-8}, \quad \beta = 1, \quad (-\nu_1, \lambda_1) = (0, 3.00e-08)$

n	R	$RE_{\max}$	$RE_{\min}$	x_{11}	x_{nn}	$\|X\|$	N_-	N_+	N
64	6.97e-13	3.39e-14	1.42e-14	4.19	2.24e-03	8.59e+01	0	55	5
128	2.71e-12	7.83e-14	2.91e-14	4.21	1.10e-03	1.72e+02	0	55	5
256	1.02e-11	1.60e-13	7.47e-14	4.22	5.48e-04	3.44e+02	0	55	5
512	4.19e-11	3.71e-13	1.53e-13	4.22	2.73e-04	6.87e+02	0	55	5

Table 6: $\alpha = 10^{-15}, \quad \beta = 1, \quad (-\nu_1, \lambda_1) = (0, 3.00e-15)$

7 Conclusion

We have presented four formulas for the minimal positive solution of the non-symmetric Riccati equation (1) that depend on the eigenvalues of the associated matrix. With the help of the formulas we have given some properties and entry-wise bounds for the minimal positive solution. We also have derived a norm-wise upper bound by using the invariant subspace connection. We have used the formulas to develop fast numerical algorithms for computing the minimal positive solution. If the eigenvalues can be computed accurately, then the computed minimal positive solution has high relative accuracy.

References

[1] Z.-Z. Bai, X.-X. Guo, an S.-F.. Xu. Alternately linearized implicit iteration methods for the minimal nonnegative solutions of the nonsymmetric algebraic Riccati equations. *Numer. Linear Algebra Appl.*, 13:655 – 574, 2006.

[2] J.R. Bunch, Ch.P. Nielson, and D.C. Sorensen. Rank-one modification of the symmetric eigenproblem. *Numer. Math.*, 31:31 – 48, 1978.

[3] J.J.M. Cuppen. A divided and conquer method for the symmetric eigenproblem. *Numer. Math.*, 36:177 – 195, 1981.

[4] J.J. Dongarra and D.C. Sorensen. A fully parallel algorithm for the symmetric eigenvalue problem. *SIAM J. Sci. and Stat. Comp.*, 8:139 – 154, 1987.

[5] T. Finck, G. Heinig, and K. Rost. An inversion formula and fast algorithms for Cauchy-Vandermonde matrices. *Linear Algebra Appl.*, 183:179 – 197, 1993.

[6] B.D. Ganapol. An investigation of a simple transport model. *Transport Theory Stat. Physics*, 21:1 – 37, 1992.

[7] G.H. Golub. Some modified matrix eigenvalue problems. *SIAM Review*, 15:318 – 344, 1973.

[8] G.H. Golub and C.F. Van Loan. *Matrix Computations (3rd Edition).* Johns Hopkins University Press, Baltimore and London, 1996.

[9] M. Gu and S.C. Eisenstat. A divide-and conquer algorithm for the symmetric tridiagonal eigenproblem. *SIAM J. Matrix Anal. Appl.*, 16:172 – 191, 1995.

[10] C.-H. Guo. Nonsymmetric algebraic Riccati equations and Wiener-Hopf factorization for M-matrices. *SIAM J. Matrix Anal. Appl.*, 23:225 – 242, 2001.

[11] C.-H. Guo. A note on the minimal nonnegative solution of a nonsymmetric algebraic Riccati equation. *Linear Algebra Appl.*, 357:299 – 302, 2002.

[12] C.-H. Guo and N.J. Higham. Iterative solution of a nonsymmetric algebraic Riccati equation. *SIAM J. Matrix Anal. Appl.*, 29:396 – 412, 2007.

[13] C.-H. Guo, B. Iannazzo and B. Meini. On the doubling algorithm for a (shifted) nonsymmetric algebraic Riccati equations. Technical Report TR 1628, Dipartimento di Matematica, University of Pisa, May 2006.

[14] C.-H. Guo and A.J. Laub. On the iterative solution of a class of nonsymmetric algebraic Riccati equations. *SIAM J. Matrix Anal. Appl.*, 22:376 – 391, 2000.

[15] X.-X. Guo, W.-W. Lin, and S.-F. Xu. A structure-preserving doubling algorithm for nonsymmetric algebraic Riccati equation. *Numer. Math.*, 103:393 – 412, 2006.

[16] J. Juang. Existence of algebraic matrix Riccati equations arising in transport theory. *Linear Algebra Appl.* 230:89 – 100, 1995.

[17] J. Juang and I.-D. Chen. Iterative solution for a certain class of algebraic matrix Riccati equations arising in transport theory. *Transport Theory Statist. Phys.*, 22:65 – 80, 1993.

[18] J. Juang and W.-W. Lin. Nonsymmetric algebraic Riccati equations and Hamiltonian-like matrices. *SIAM J. Matrix Anal. Appl.*, 20:228 – 243, 1998.

[19] J. Juang and Z.-T. Lin. Convergence of an iterative technique for algebraic matrix Riccati equations and applications to transport theory. *Transport Theory Statist. Phys.*, 21:87 – 100, 1992.

[20] J. Juang and P. Nelson. Global existence, asymptotic and uniqueness for the reflection kernel of the angularly shifted transport equation. *Math. Models Methods Appl. Sci.*, 5:239 – 251, 1995.

[21] P. Lancaster and L. Rodman. *The Algebraic Riccati Equation*. Oxford University Press, Oxford, 1995.

[22] R.-C. Li. Solving secular equations stably and efficiently. Technical Report, *Department of Mathematics, University of California, Berkeley, CA*, LAPACK Working Note 89, 1993.

[23] R.-C. Li. Solving secular equations stably and efficiently. *Technical Report UCB//CSD-94-851*, Computer Science Division, Department of EECS, UC Berkeley. (LAPACK working note N. 93), 1994.

[24] L.-Z. Lu. Solution form and simple iteration of a nonsymmetric algebraic Riccati equation arising in transport theory. *SIAM J. Matrix Anal. Appl.*, 26:679 – 685, 2005.

[25] V. Mehrmann and H. Xu. Choosing poles so that the single-input pole placement problem is well-conditioned. *SIAM J. Matrix Anal. Appl.*, 19:664 – 681, 1998.

[26] A. Melman. Numerical solution of a secular equation. *Numer. Math.*, 69:483 – 493, 1995.

[27] L.C.G. Rogers. Fluid models in queueing theory and Wiener-Hopf factorization of Markov Chains. *Ann. Appl. Probab.*, 4:390 – 413, 1994.

[28] G.W. Stewart. *Afternotes on Numerical Analysis.* SIAM, Philadelphia, 1996.

[29] G.W. Stewart and J.-G. Sun. *Matrix Perturbation Theory.* Academic Press, Boston, 1990.