Existence of Algebraic Matrix Riccati Equations 
Arising in Transport Theory

Jonq Juang*
Department of Applied Mathematics
National Chiao Tung University
Hsinchu, Taiwan

Submitted by Richard A. Brualdi

ABSTRACT

We consider the existence of positive solutions of a certain class of algebraic 
matrix Riccati equations with two parameters, $c$ ($0 \leq c \leq 1$) and $\alpha$ ($0 \leq \alpha \leq 1$). 
Here $c$ denotes the fraction of scattering per collision, and $\alpha$ is an angular shift. 
Equations of this class are induced via invariant imbedding and the shifted Gauss-
Legendre quadrature formula from a simple transport model. By establishing the 
existence of positive solutions of such equations, the problem of the convergence of 
some iterative schemes for solving them can be completely solved.

1. INTRODUCTION

Consider the algebraic matrix Riccati equation of the form

$$B - AS - SD + SCS = 0.$$ (1)

* The work was partially supported by the National Science Council of R.O.C.
Here $A$, $B$, $C$, and $D$ are matrices having the following structure:

\[ A_{N \times N} = \text{diag} \left[ \frac{1}{c(w_1^- + \alpha)}, \frac{1}{c(w_2^- + \alpha)}, \ldots, \frac{1}{c(w_N^- + \alpha)} \right] \]

\[- \left[ \begin{array}{c} \vdots \\ 1 \end{array} \right] \left[ \begin{array}{c} c_1^- \\ 2(w_1^- + \alpha) \\ \vdots \\ 2(w_N^- + \alpha) \end{array} \right] \]

\[ := D_A - ia^T, \]

where $i = \left[ \begin{array}{c} 1 \\ \vdots \\ 1 \end{array} \right]$

\[ D_{N^+ \times N^+} = \text{diag} \left[ \frac{1}{c(w_1^+ - \alpha)}, \frac{1}{c(w_2^+ - \alpha)}, \ldots, \frac{1}{c(w_N^+ - \alpha)} \right] \]

\[- i \left[ \begin{array}{c} c_1^+ \\ 2(w_1^+ - \alpha) \\ \vdots \\ 2(w_N^+ - \alpha) \end{array} \right] \]

\[ := D_D - id^T; \]

\[ B = ii^T; \]

\[ C = da^T. \]

Equation (1) contains two parameters $c$ and $\alpha$. Here $c$ denotes the average total number of particles emerging from a collision, which is assumed to be conservative, (i.e., $0 \leq c \leq 1$), and $\alpha$ ($0 \leq \alpha \leq 1$) is an angular shift. The dimensionally dependent quantities $w_i^-$ and $w_i^+$ denote the Gauss-Legendre sets (see, e.g., [12]) on $[-\alpha, 1]$ and $[\alpha, 1]$, respectively; and $c_i^-$ and $c_i^+$ are, respectively, their corresponding weights. Without loss of generality, we shall assume that

\[-\alpha < w_1^- < w_2^- < \ldots < w_N^- < 1 \quad \text{and} \quad \alpha < w_1^+ < w_2^+ < \ldots < w_N^+, < 1. \]

Such an equation is induced via invariant imbedding (see, e.g., [1, 2, 15]) and the shifted Gauss-Legendre quadrature formula from a simple transport model [5, 6].
For $\alpha = 0$, two iterative procedures for finding the minimal positive solution (in the componentwise sense) of Equation (1), one corresponding to a nonlinear version of the Gauss-Jocobi (GJ) method and the other associated with a nonlinear version of the Gauss-Seidel (GS) method, were proposed, respectively, by Shimizu and Aoki [13] and by Juang and Lin [9]. While such iterative procedures have been proved quite effective in practice (see [10–11] and the work cited therein), their convergence has not yet been fully investigated. Sufficient conditions for convergence of the GJ and GS methods were given in [10] and [9], respectively. However, it was noted (see Table 2 of [8]) that those sufficient conditions will fail if $c$ is not far away from 1. And it was also observed (see Theorem 1 of [8]) that the existence of a positive solution of (1) implies the convergence of both iterations. This observation can be easily extended to the case that $\alpha \neq 0$. Therefore, to completely solve the convergence problem one needs to find a direct method for establishing the existence of positive solutions of Equation (1) for all $0 \leq c \leq 1$ and $0 \leq \alpha \leq 1$. This is what motivates our work here.

In this article, we first show that an a priori bound, which is independent of $c$ and $\alpha$, can be obtained by introducing a one-parameter ($k_1$, $0 < k_1 < 1$) family. Therefore, degree theory is applied to show the existence of positive solutions. Some applications and concluding remarks are given in Section 3.

2. MAIN RESULTS

To derive our main results, we first write Equation (1) in the component form

$$
\left( \frac{1}{w_i^- + \alpha} + \frac{1}{w_j^+ - \alpha} \right) S_{ij} = c \left( 1 + \frac{1}{2} \sum_{k=1}^{N^-} \frac{c_k^- S_{kj}}{w_k^- + \alpha} \right) \left( 1 + \frac{1}{2} \sum_{k=1}^{N^+} \frac{c_k^+ S_{ik}}{w_k^+ - \alpha} \right).
$$

(2)

The structure of (2) suggests that we seek a solution of the form

$$
S_{ij} = \frac{c(w_i^- + \alpha)(w_j^+ - \alpha)}{w_i^- + w_j^+} h_i h_j,
$$

(3a)
where
\[ h_i = 1 + \frac{1}{2} \sum_{k=1}^{N^+} \frac{c_k^+}{w_k^- - \alpha} s_{ik}, \quad 1 \leq i \leq N^-, \] (3b)
\[ l_j = 1 + \frac{1}{2} \sum_{k=1}^{N^-} \frac{c_k^-}{w_k^- + \alpha} s_{kj}, \quad 1 \leq j \leq N^+. \] (3c)

Substituting (3a) into (3b) and (3c), respectively, we obtain
\[ h_i = 1 + \frac{c}{2} \sum_{k=1}^{N^+} \frac{c_k^+ (w_i^- + \alpha)}{w_i^- + w_k^+} h_i l_k, \quad 1 \leq i \leq N^-, \] (4a)
\[ l_j = 1 + \frac{c}{2} \sum_{k=1}^{N^-} \frac{c_k^- (w_j^+ - \alpha)}{w_k^- + w_j^+} h_j l_k, \quad 1 \leq j \leq N^+. \] (4b)

Set \( h_i = 1/\tilde{h}_i, \) \( l_j = 1/\tilde{l}_j; \) then Equation (4) can be equivalently reduced to
\[ \tilde{h}_i = 1 - \frac{c}{2} \sum_{k=1}^{N^+} \frac{c_k^+}{\tilde{l}_k} + \frac{c}{2} \sum_{k=1}^{N^+} \frac{(w_k^+ - \alpha)c_k^+}{(w_i^- + w_k^+)\tilde{l}_k}, \quad 1 \leq i \leq N^-, \] (5a)
\[ \tilde{l}_j = 1 - \frac{c}{2} \sum_{k=1}^{N^-} \frac{c_k^-}{\tilde{h}_k} + \frac{c}{2} \sum_{k=1}^{N^-} \frac{(w_k^- + \alpha)c_k^-}{(w_k^- + w_j^+)\tilde{h}_k}, \quad 1 \leq j \leq N^+. \] (5b)

Define
\[ x = \frac{c}{2} \sum_{k=1}^{N^-} \frac{c_k^-}{\tilde{h}_k} \] (6a)
and
\[ y = \frac{c}{2} \sum_{k=1}^{N^+} \frac{c_k^+}{\tilde{l}_k}. \] (6b)
Multiplying (5a) by \( cc_i^{-}/2 \tilde{h}_i \), and summing the resulting equation over the index \( i \), we have that

\[
\frac{c}{2} \sum_{i=1}^{N^-} c_i^- = \frac{c(1 + \alpha)}{2} = x - xy + \frac{c}{2} \left[ \sum_{i=1}^{N^-} c_i^- \left( \frac{c}{2} \sum_{k=1}^{N^+} \frac{(w_k^+ - \alpha) e_k^+}{w_i^- + w_k^+} \tilde{l}_i \right) \right]
\]

\[= x - xy + a. \quad (7a)\]

A similar procedure is applied to (5b) to get

\[
\frac{c}{2} \sum_{i=1}^{N^+} c_i^+ = \frac{c(1 - \alpha)}{2} = y - xy + \frac{c}{2} \left[ \sum_{i=1}^{N^+} c_i^+ \left( \frac{c}{2} \sum_{k=1}^{N^-} \frac{(w_i^- + \alpha) e_i^-}{w_i^- + w_k^+} \tilde{l}_i \right) \right]
\]

\[= y - xy + b \quad (7b)\]

Adding (7a) and (7b), we obtain

\[(1 - x)(1 - y) = 1 - c. \quad (8)\]

**Remark.**

1. For \( \alpha = 0 \), the quantities \( h_i \) and \( l_j \) are the discrete version of Chandrasekhar’s well-known \( H \) functions [3, 4].

2. For \( \alpha = 0 \), (8) reduces to a discrete version of some expressions [3, 4] concerning the properties of \( H \) functions.

Since \( a + b = xy \), we see immediately, for \( \alpha \neq 1 \), that if \( h_i \) and \( l_j \) are positive solutions of (4), then there must exist two positive numbers \( k_1 \) and \( k_2 \), where \( 0 < k_1, k_2 < 1 \) and \( k_1 + k_2 = 1 \), such that

\[a = k_1 xy \quad \text{and} \quad b = k_2 xy. \quad (9a, b)\]
It then follows from (7), (8), and (9) that the following holds:

\[
x = \frac{1 - \frac{c}{2}(1 - \alpha) + k_1c \pm \sqrt{[1 - \frac{c}{2}(1 - \alpha) + k_1c]^2 - 2k_1(1 + \alpha)c}}{2k_1} =: a_1 \pm b_1,
\]

\[
y = \frac{1 - \frac{c}{2}(1 + \alpha) + k_2c \pm \sqrt{[1 - \frac{c}{2}(1 + \alpha) + k_2c]^2 - 2k_2(1 - \alpha)c}}{2k_2} =: a_2 \pm b_2.
\]

Since \(k_1\) and \(k_2\) are to be treated as real parameters, necessary conditions for (10) to be meaningful are that both \([1 - \frac{c}{2}(1 - \alpha) + k_1c]^2 - 2k_1c(1 + \alpha)\) and \([1 - \frac{c}{2}(1 + \alpha) + k_2c]^2 - 2k_2c(1 - \alpha)\) are nonnegative. However, these are so if \(0 \leq a \leq 1\) and \(0 \leq c \leq 1\). To see this, we note that, for \(c \neq 0\), \(f_1(k_1) := [1 - \frac{c}{2}(1 - \alpha) + k_1c]^2 - 2k_1c(1 + \alpha)\) has a minimum \((1 + \alpha)(1 - \alpha)(1 - c)\), which is nonnegative whenever \(0 \leq a \leq 1\) and \(0 \leq c \leq 1\).

We denote by \(F\) the feasible region \(\{(k, c, \alpha) : 0 < k < 1, 0 \leq c \leq 1, \text{ and } 0 \leq \alpha \leq 1\}\) for the solution of (1). The properties and signs of \(1 - x\) and \(1 - y\) will be examined in the next lemma.

**Lemma 1.**

(i) \(1 - a_1 + b_1 > 0\) and \(1 - a_1 - b_1 < 0\) for all \((k_1, c, \alpha) \in F\).

(ii) \(1 - a_2 + b_2 > 0\) and \(1 - a_2 - b_2 < 0\) for all \((k_2, c, \alpha) \in F\).

(iii) Let \(c\) be sufficiently small, say \(0 \leq c \leq \frac{1}{8}\). Then \(1 - a_1 + b_1 > \frac{1}{2}\) and \(1 - a_2 + b_2 > \frac{6}{7}\) for all \(k_1, k_2, 0 < k_1, k_2 < 1, \text{ and all } \alpha, 0 \leq \alpha \leq 1\).

**Proof.** Since the computation leading to (i) and (ii) is similar, we shall only prove (i). To see (i), it suffices to show that \(b_1^2 \geq (1 - a_1)^2\), or equivalently

\[
\left[1 - \frac{c}{2}(1 - \alpha) + k_1c\right]^2 - 2k_1c(1 + \alpha) - \left[(2k_1 - 1)\left(1 - \frac{c}{2}\right) - \frac{c\alpha}{2}\right]^2 \geq 0.
\]
Since the left-hand side of the inequality is equal to $4(1 - k_1)(k_1)(1 - c)$, the assertion of Lemma 2(i) thus follows.

To prove (iii), we see that if $0 \leq c \leq \frac{1}{8}$, then

$$a_i - b_1 = \frac{(1 + \alpha)c}{1 - (c/2)(1 - \alpha) + k_1 c + \sqrt{[1 - (c/2)(1 - \alpha) + k_1 c]^2 - 2k_1 c(1 + \alpha)}}$$

$$\leq 2c(1 + \alpha) \leq \frac{1}{2}$$

for all $k_1$ and $\alpha$. Thus, $1 - a_1 + b_1 \geq \frac{1}{2}$, as asserted. Similarly, we have

$$a_2 - b_2 = \frac{(1 - \alpha)c}{1 - (c/2)(1 + \alpha) + k_2 c + \sqrt{[1 - (c/2)(1 + \alpha) + k_2 c]^2 - 2k_2 c(1 - \alpha)}}$$

$$\leq \frac{c}{1 - (c/2)(1 + \alpha)} \leq \frac{c}{1 - c} \leq \frac{1}{7}.$$

Therefore, $1 - a_2 + b_2 \geq \frac{6}{7}$, as asserted. 

In view of (8), we see that if $h_i$ and $l_j$ are solutions of (4), then either

$$1 - x \geq 0 \quad \text{and} \quad 1 - y \geq 0 \quad (11a)$$

or

$$1 - x \leq 0 \quad \text{and} \quad 1 - y \leq 0. \quad (11b)$$

An a priori bound, which is independent of $k_1$, $c$, and $\alpha$, is obtained in the following lemma.

**Lemma 2.** Let $\tilde{h}_i$ and $\tilde{l}_j$ be any positive solutions of (5) satisfying (11a). Then there is an $m > 0$ such that $\min\{\tilde{h}_i, \tilde{l}_j\} \geq m$ for all $i, j$, all $0 \leq \alpha \leq 1$, and all $0 \leq c \leq 1$.

**Proof.** Using (4), we see clearly that $\tilde{h}_i \leq 1$ and $\tilde{l}_j \leq 1$ for all $i, j$. Therefore,

$$\tilde{h}_i \geq 1 - y + \frac{c}{2} \sum_{k=1}^{N^+} \frac{(w_k^+ - \alpha)c_k^+}{1 + w_k^+}$$
and

\[ \tilde{L}_j \geq 1 - x + \frac{c}{2} \sum_{k=1}^{N^-} \frac{w_k^- c_k^-}{w_k^- + 1}. \]

Since \(1 - y \geq 0\) and \(1 - x \geq 0\), there must exist positive constants \(k_1\) and \(k_2\), \(k_1 + k_2 = 1\), such that \(1 - x = 1 - a_1 + b_1\) and \(1 - y = 1 - a_2 + b_2\), where \(a_1 - b_1\) and \(a_2 - b_2\) are defined in (10). Now, via Lemma 1(iii), \(1 - x \geq \frac{1}{2}\) for \(0 \leq c \leq \frac{1}{8}\). Consequently,

\[ \tilde{L}_j \geq \min \left\{ \frac{1}{2}, \frac{1}{16} \sum_{k=1}^{N^-} \frac{w_k^- c_k^-}{(w_k^- + 1)} \right\} =: m_2 > 0 \]

for all \(j\), all \(0 \leq c \leq 1\), and all \(0 \leq \alpha \leq 1\). On the other hand,

\[ y = \frac{c}{2} \sum_{k=1}^{N^+} c_k^+ \leq \frac{c(1 - \alpha)}{2m_2}, \]

and so \(1 - y \geq 1 - c(1 - \alpha)/2m_2\). Hence, if \(0 \leq c \leq m_2\) or \(\alpha \geq 1 - m_2\), then \(1 - y \geq \frac{1}{2}\). However, if \(1 \geq c \geq m_2\) and \(0 \leq \alpha \leq 1 - m_2\), then

\[ \frac{c}{2} \sum_{k=1}^{N^+} \frac{(w_k^+ - \alpha)c_k^+}{1 + w_k^+} \geq \frac{m_2}{2} \sum_{k=1}^{N^+} \frac{(w_k^+ - 1 + m_2)c_k^+}{1 + w_k^+} =: \bar{m}_1 > 0. \]

Consequently, \(\bar{h}_i \geq \min \{\frac{1}{2}, \bar{m}_1\} =: m_1\). The assertion of the lemma now follows on choosing \(m = \min \{m_1, m_2\}\).

**Theorem 1.** Equation (1) has positive solutions satisfying (11a) for all \(0 \leq c \leq 1\) and \(0 \leq \alpha \leq 1\).

**Proof.** Using (3a), Lemma 2, and the fact that \(w_j^+ \geq \alpha\) and \(w_j^- \geq -\alpha\) for all \(i, j\), we conclude that for any positive solution \(S_{ij}\) of (2) satisfying (11a),

\[ S_{ij} \leq \frac{c(1 - \alpha^2)}{2} m^2 \leq \frac{m^2}{2} \]
for all $i,j$, all $0 \leq c \leq 1$, and all $0 \leq \alpha \leq 1$. Here $m$ is defined as in Lemma 2. Let $\tilde{S}$ be a column vector defined as

$$\tilde{S} = (S_{11}, \cdots, S_{1N+}, S_{21}, S_{22}, \cdots, S_{2N+}, \cdots, S_{N-1}, \cdots, S_{N-N+})^T.$$

Then (2) can be formulated as

$$\tilde{S} = F_c(\tilde{S}),$$

where $F_c$ is a continuous and nonlinear map from $R^{N-N+}$ to $R^{N-N+}$. Choose

$$r = \left(1 + \frac{m^2}{4} \sum_{k=1}^{N-} \frac{c_k^-}{w_k^+ + \alpha}\right) \left(1 + \frac{m^2}{4} \sum_{k=1}^{N+} \frac{c_k^+}{w_k^+ - \alpha}\right) > 0,$$

and let $D = \{ x \in R^{N-N+} : \|x\|_\infty < r \}$. Clearly, $D$ is a bounded open set in $R^{N-N+}$, and $F_c$ is continuous on $\overline{D}$. Consider the homotopy $H_c = I - F_c$, and suppose that $\tilde{S} - F_c(\tilde{S}) = 0$ for $\tilde{S} \in \overline{D}$; then $\|\tilde{S}\|_\infty = \|F_c(\tilde{S})\|_\infty < r/2 < r$. Thus $\tilde{S} \in D$. Hence, by homotopy invariance (see, e.g., Theorem 13.2.11(ii) of [7]),

$$d(H_c, 0, D) = d(H_0, 0, D) = d(I, 0, D) = 1.$$

The above argument is true for all $0 \leq \alpha \leq 1$. Therefore, we conclude that Equation (1) has positive solutions for all $0 \leq c \leq 1$ and $0 \leq \alpha \leq 1$. □

Remark. We are motivated by the work of Stuart [14] to use the homotopy argument to show the existence of positive solutions of (1).

3. APPLICATIONS AND CONCLUDING REMARKS

As in the case that $\alpha = 0$ (i.e., no angular shift), the iterative procedures for solving the minimal positive solution of the equation (1) can be classified into three types: first, the iteration of Aoki and Shimizu, which is essentially a nonlinear version of Gauss-Jocobi (GJ); second, the iteration of Juang and Lin, which is essentially a nonlinear version of Gauss-Seidel (GS); third, a nonlinear version of SOR, whose effectiveness has yet to be studied theoreti-
cally. We now define the GJ and GS methods, respectively, as follows:

\[
S^{(p+1)}_{ij} = \frac{c(w_i^- + \alpha)(w_j^+ - \alpha)}{w_i^- + w_j^-} \left( 1 + \frac{1}{2} \sum_{k=1}^{N^-} c_k^- S^{(p)}_{kj} \right) \left( 1 + \frac{1}{2} \sum_{k=1}^{N^+} c_k^+ S^{(p)}_{ik} \right),
\]

(12a)

\[
S^{(0)}_{ij} = 0 \quad \text{for all } i, j,
\]

(12b)

and

\[
\tilde{S}^{(p+1)}_{ij} = \frac{c(w_i^- + \alpha)(w_j^+ - \alpha)}{w_i^- + w_j^+} \left( 1 + \frac{1}{2} \sum_{k=1}^{i-1} c_k^- \tilde{S}^{(p+1)}_{kj} \right) \left( 1 + \frac{1}{2} \sum_{k=i}^{N^+} c_k^+ \tilde{S}^{(p)}_{ik} \right) \times \left( 1 + \frac{1}{2} \sum_{k=1}^{j-1} c_k^+ \tilde{S}^{(p+1)}_{ik} \right) \left( 1 + \frac{1}{2} \sum_{k=j}^{N^+} c_k^- \tilde{S}^{(p)}_{ik} \right),
\]

(13a)

\[
\tilde{S}^{(0)}_{ij} = 0 \quad \text{for all } i, j.
\]

(13b)

Let \( S = (S_{ij}) \) be a positive solution of Equation (1), whose existence is assured by Theorem 1.

An easy induction will give

\[
\max \left\{ S^{(p)}_{ij}, \tilde{S}^{(p)}_{ij} \right\} \leq S_{ij} \quad \text{for all } i, j \text{ and all } p.
\]

It is also clear that for each \( i, j \), the iterations \( \{S^{(p)}_{ij}\}^{\infty}_{p=0} \) and \( \{\tilde{S}^{(p)}_{ij}\}^{\infty}_{p=0} \) are monotonically increasing. Therefore, the limits of both iterations exist; they will be denoted by \( S^{(\infty)}_{ij} \) and \( \tilde{S}^{(\infty)}_{ij} \), respectively. Furthermore, \( (S^{(\infty)}_{ij}) = (\tilde{S}^{(\infty)}_{ij}) \).

To see this, we first note that \( (S^{(\infty)}_{ij}) \) and \( (\tilde{S}^{(\infty)}_{ij}) \) both are positive solutions of (1). Therefore, an induction will give \( S^{(p)}_{ij} \leq S^{(\infty)}_{ij} \) for all \( i, j, \text{ and } p \). Thus, \( S^{(\infty)}_{ij} \leq \tilde{S}^{(\infty)}_{ij} \). Similarly, \( S^{(\infty)}_{ij} \geq \tilde{S}^{(\infty)}_{ij} \). We summarize the above results as follows.

**Theorem 2.** For all \( 0 \leq c \leq 1 \) and \( 0 \leq \alpha \leq 1 \), the iterations \( \{S^{(p)}_{ij}\}^{\infty}_{p=0} \) and \( \{\tilde{S}^{(p)}_{ij}\}^{\infty}_{p=0} \) converge to the minimal positive solution \( S_{\min} \) of (1).

The minimal positive solution \( S_{\min} \) of (1) is defined in the following sense: if \( S \) is a positive solution of (1), then \( S \geq S_{\min} \), i.e., \( S_{ij} \geq (S_{\min})_{ij} \) for all \( i, j \).
THEOREM 3. Equation (1) has a unique positive solution satisfying (11a) for $0 \leq c \leq 1$ and $0 \leq \alpha \leq 1$.

Proof. Let $(h_{\text{min}})_i$ and $(l_{\text{min}})_j$ be the minimal positive solution pair of (4), whose existence assured by Theorem 2. Let $x_{\text{min}}$ and $y_{\text{min}}$ be defined as in (6) except that $h_i$ and $l_j$ are replaced by $(h_{\text{min}})_i$ and $(l_{\text{min}})_j$, respectively. Consequently, $x_{\text{min}} \leq x$ and $y_{\text{min}} \leq y$. On the other hand, for $c \neq 1$, $1 - x_{\text{min}} = (1 - c)/(1 - y_{\text{min}}) \leq (1 - c)/(1 - y) = 1 - x$. Hence, $x_{\text{min}} = x$, so that $h_i = (h_{\text{min}})_i$ for all $i$. Similarly, $l_j = (l_{\text{min}})_j$. The uniqueness of the positive solution of (1) satisfying (11a) now follows from (3a).

We conclude this paper by suggesting the following further related matters:

1. It would be interesting to study the bifurcation diagram of positive solutions of Equation (1) as $c$ and $\alpha$ vary from 0 to 1.
2. It is of interest to investigate the effectiveness of the NSOR.
3. Additional complexities of the reflections or scattering problem can also be considered, such as anisotropic scattering, spatially distributed sources, and time dependence.

REFERENCES


Received 21 June 1993; final manuscript accepted 15 December 1993