ELSEVIER

Journal of Computational and Applied Mathematics 54 (1994) 377–387

JOURNAL OF COMPUTATIONAL AND APPLIED MATHEMATICS

Quadratic operator equations and periodic operator continued fractions

Robert C. Busby [a,*], Wyman Fair [b]

[a] *Department of Mathematics and Computer Science, Drexel University, Philadelphia, PA 19104, United States*

[b] *Eckerd College, St. Petersburg, FL 33711, United States*

Received 15 January 1993

Abstract

Conditions are given that assure convergence of an operator-valued periodic continued fraction of period two. These results and techniques are applied to get a solution of the quadratic operator equation in a complex Hilbert space. Special attention is then given to the important case of the quadratic matrix equation connected with the steady-state solution of the matrix Riccati equation from control theory. It is shown that a modification of the traditional matrix power approximation technique leads to a new, efficient and highly simplified method of approximating the unique nonnegative definite solution that exists in many important special cases.

Keywords: Continued fraction; Matrix Riccati equation; Control theory

1. Introduction

Many of the standard theorems of continued fractions with entries from a field have been generalized to the case of noncommutative entries, see [1–9,11]. The applications given in [1,6,12] are evidence of the usefulness of such generalizations (see also the extensive references in [4] to applied mathematics and physics. In this paper we continue the work of [2,3] and give convergence criteria for periodic noncommutative continued fractions of period two.

These results are of particular interest because we can use them to construct a continued fraction solution to the quadratic matrix equation describing the steady state of the matrix Riccati equation, a topic of great interest throughout much of control theory (see also [1,10]). The resulting procedure shows that in many cases a simple modification of the classical power method of approximation will yield a solution of the matrix quadratic equation. It also yields precise numerical information, including the rate of convergence of the method to the solution.

* Corresponding author. e-mail: rbusby@mcs.drexel.edu.

SSDI 0377-0427(93)E0167-K

2. Notation and basic lemmas

Let H be a complex Hilbert space. We refer the reader to [2] for definitions and basic properties of spectral operators and, in particular, for the proof of the following lemma.

Lemma 2.1. *Let W_1 and W_2 be spectral operators on H such that*

$$\inf\{|\lambda| \mid \lambda \in \mathrm{sp}(W_1)\} > \sup\{|\lambda| \mid \lambda \in \mathrm{sp}(W_2)\}.$$

Then W_1 is invertible and $\lim_{n\to\infty} \|W_1^{-n}\|\|W_2^n\| = 0$.

Now let $\boldsymbol{H} = H \oplus H$, the Hilbert space direct sum of H with itself. We now recall some of the definitions and results from [2]. Let $U_i : H \to \boldsymbol{H}$, $i = 1, 2$, be linear isometries such that $\boldsymbol{H}_1 = U_1(H)$ and $\boldsymbol{H}_2 = U_2(H)$ intersect only in the zero vector of $\boldsymbol{H}$. If $\boldsymbol{H} = \boldsymbol{H}_1 \oplus \boldsymbol{H}_2$, then the set $\{U_1, U_2\}$ is called an *H-basis* for $\boldsymbol{H}$. The canonical H-basis for $\boldsymbol{H}$ is defined by $V_1(x) = (x, 0)$ and $V_2(y) = (0, y)$ for x and y in H.

If $\{U_1, U_2\}$ is an H-basis for $\boldsymbol{H}$, then if $x \in \boldsymbol{H}$, $x = x_1 + x_2$, $x_i \in U_i(H)$. Define $U_i^*(x) = U_i^{-1}(x_i)$, $i = 1, 2$. Then U_i^* is a continuous linear map from $\boldsymbol{H}$ onto H, and $U_i^* U_j = \delta_{ij} I$ where δ_{ij} is the Kronecker delta and I is the identity on H. Also $U_i U_i^* = P_i$, $i = 1, 2$, where P_i is the projection of $\boldsymbol{H}$ onto $\boldsymbol{H}_i$ along the complementary subspace.

Lemma 2.2. *For each $T \in L(\boldsymbol{H})$, the set of all bounded linear operators on $\boldsymbol{H}$, let $T_{ij} = U_i^* T U_j \in L(H)$, and let*

$$\boldsymbol{M}_T = \begin{bmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{bmatrix} \in M(2, H),$$

the algebra of two by two matrices with entries from $L(H)$ with matrix operator norm and matrix operations. Then the correspondence $T \to \boldsymbol{M}_T$ is an isomorphism from $L(\boldsymbol{H})$ onto $M(2, H)$.

Lemma 2.3. *"Change of basis" works in the above context: let $\{U_1, U_2\}$ and $\{W_1, W_2\}$ be H-bases for $\boldsymbol{H}$, let $T \in L(\boldsymbol{H})$ and let $\boldsymbol{M}_T$ and $\boldsymbol{N}_T$ be the respective matrix representations of T in the H-bases just defined. Let $\boldsymbol{S} = [S_{ij}]$ be the matrix in $M(2, H)$ with entries $S_{ij} = U_i^* W_j$, $i, j = 1, 2$. Then, $\boldsymbol{S}^{-1} \boldsymbol{M}_T \boldsymbol{S} = \boldsymbol{N}_T$.*

3. Convergence of a periodic continued fraction

Let A_i and B_i, $i = 1, 2, 3, \ldots$, be entries in $L(H)$, H a complex Hilbert space. The formal expression

$$\frac{A_1}{B_1+} \quad \frac{A_2}{B_2+} \quad \frac{A_3}{B_3+\cdots} \tag{3.1}$$

is a noncommutative continued fraction, see [11] for basic properties of these fractions. Provided that the appropriate inverses exist, the numerator P_n and the denominator Q_n of the nth approximate $P_n Q_n^{-1}$ satisfy the relations

$$\begin{aligned} P_{n+1} &= P_n B_{n+1} + P_{n-1} A_{n+1}, \quad P_{-1} = I, \quad P_0 = 0, \\ Q_{n+1} &= Q_n B_{n+1} + Q_{n-1} A_{n+1}, \quad Q_{-1} = 0, \quad Q_0 = I. \end{aligned} \tag{3.2}$$

We restrict ourselves to the two-periodic case

$$\frac{A_1}{B_1+} \quad \frac{A_2}{B_2+} \quad \frac{A_1}{B_1+\cdots}. \tag{3.3}$$

If we define

$$\boldsymbol{A} = \begin{bmatrix} B_1 & I \\ A_1 & 0 \end{bmatrix} \begin{bmatrix} B_2 & I \\ A_2 & 0 \end{bmatrix} = \begin{bmatrix} B_1 B_2 + A_2 & B_1 \\ A_1 B_2 & A_1 \end{bmatrix} = \begin{bmatrix} Q_2 & Q_1 \\ P_2 & P_1 \end{bmatrix}, \tag{3.4}$$

then

$$\boldsymbol{A}^n = \begin{bmatrix} Q_{2n} & Q_{2n-1} \\ P_{2n} & P_{2n-1} \end{bmatrix}. \tag{3.5}$$

Here $\boldsymbol{A}$ can be assumed (see Lemma 2.2) to be the matrix, relative to the canonical H-basis $\{V_1, V_2\}$, of an operator A in $L(\boldsymbol{H})$. We assume that A is spectral and satisfies the following additional conditions.

(α) If σ is the spectrum of A, then $\sigma = \sigma_1 \cup \sigma_2$ where

(i) $\sigma_1 \cap \sigma_2 = \emptyset$;

(ii) the ranges of the disjoint spectral projections $E(\sigma_1)$ and $E(\sigma_2)$ are isomorphic with H;

$$\text{(iii)} \qquad \inf\{|\lambda| \mid \lambda \in \sigma_1\} > \sup\{|\lambda| \mid \lambda \in \sigma_2\}. \tag{3.6}$$

In this case we will let $\gamma \equiv \gamma(A, \sigma_1, \sigma_2) = \sup\{|\lambda| \mid \lambda \in \sigma_2\} / \inf\{|\lambda| \mid \lambda \in \sigma_1\} < 1$.

Now $\boldsymbol{H}_1 = E(\sigma_1)\boldsymbol{H}$ and $\boldsymbol{H}_2 = E(\sigma_2)\boldsymbol{H}$ are invariant subspaces of $\boldsymbol{H}$, and by part (i) and the first line of (α), we see that $\boldsymbol{H} = \boldsymbol{H}_1 \oplus \boldsymbol{H}_2$. By (ii) of ($\alpha$) we may choose isometries $U_i : H \to \boldsymbol{H}_i$, $i = 1, 2$, so that $\{U_1, U_2\}$ is an H-basis for $\boldsymbol{H}$. Relative to this basis, A has the "Jordan canonical form"

$$\boldsymbol{\Lambda} = \begin{bmatrix} \Lambda_1 & 0 \\ 0 & \Lambda_2 \end{bmatrix}, \quad \Lambda_i \in L(H). \tag{3.7}$$

By property (iii) of (α) and [2, Lemma 3.1], we have

$$\lim_{n \to \infty} \|\Lambda_1^{-n}\| \, \|\Lambda_2^n\| = 0. \tag{3.8}$$

Lemma 2.3 tells us that if $S_{ij} = V_i^* U_j$ and

$$\boldsymbol{S} = \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix}, \tag{3.9}$$

then

$$\boldsymbol{S}^{-1} \boldsymbol{A} \boldsymbol{S} = \boldsymbol{\Lambda}. \tag{3.10}$$

Also if we define

$$\boldsymbol{T} = \boldsymbol{S}^{-1} = \begin{bmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{bmatrix}, \tag{3.11}$$

then $T_{ij} = U_i^* V_j$. We can now state and prove a convergence theorem for (3.3).

Theorem 3.1. *Let* (α) *hold, and let* S_{11}, S_{21}, T_{11} *and* T_{12} *be invertible. Then,*
(a) *all component matrices* S_{ij} *and* T_{ij}, $1 \leqslant i, j \leqslant 2$, *are invertible*;
(b) Q_r^{-1} *exists for sufficiently large* r *and the sequence* $\{P_r Q_r^{-1}\}$ *converges to* $S_{21}S_{11}^{-1}$.

Proof. (a) From the fact that $\boldsymbol{T} = \boldsymbol{S}^{-1}$, we have the relation $S_{21}T_{11} = -S_{22}T_{21}$. By assumption, $S_{21}T_{11}$ is invertible and therefore so is $S_{22}T_{21}$. This implies that both S_{22} and T_{21} are invertible. A similar argument works for S_{12} and T_{22}.

(b) From (3.5) and (3.10),

$$\begin{bmatrix} Q_{2n} & Q_{2n-1} \\ P_{2n} & P_{2n-1} \end{bmatrix} = A^n = S \begin{bmatrix} \Lambda_1^n & 0 \\ 0 & \Lambda_2^n \end{bmatrix} T, \tag{3.12}$$

yielding

$$\begin{aligned} P_{2n} &= [I + S_{22}\Lambda_2^n T_{21} T_{11}^{-1} \Lambda_1^{-n} S_{21}^{-1}] S_{21} \Lambda_1^n T_{11}, \\ Q_{2n} &= [I + S_{12}\Lambda_2^n T_{21} T_{11}^{-1} \Lambda_1^{-n} S_{11}^{-1}] S_{11} \Lambda_1^n T_{11}, \\ P_{2n-1} &= [I + S_{22}\Lambda_2^n T_{22} T_{12}^{-1} \Lambda_1^{-n} S_{21}^{-1}] S_{21} \Lambda_1^n T_{12}, \\ Q_{2n-1} &= [I + S_{12}\Lambda_2^n T_{22} T_{12}^{-1} \Lambda_1^{-n} S_{11}^{-1}] S_{11} \Lambda_1^n T_{12} \end{aligned} \tag{3.13}$$

or

$$\begin{aligned} P_{2n} &= [I + X_n^{(1)}] S_{21} \Lambda_1^n T_{11}, \qquad & Q_{2n} &= [I + X_n^{(2)}] S_{11} \Lambda_1^n T_{11}, \\ P_{2n-1} &= [I + X_n^{(3)}] S_{21} \Lambda_1^n T_{12}, \qquad & Q_{2n-1} &= [I + X_n^{(4)}] S_{11} \Lambda_1^n T_{12}. \end{aligned} \tag{3.13′}$$

We see that $\|X_n^{(i)}\| \leqslant C_i \|\Lambda_1^{-n}\| \, \|\Lambda_2^n\|$ for some constants C_i, and so $\lim_{n\to\infty} \|X_n^{(i)}\| = 0$, $i = 1, \ldots, 4$, by Lemma 2.1. Thus for large enough n, $I + X_n^{(i)}$ is invertible, and, by the assumptions of the theorem, so are Q_{2n} and Q_{2n-1}. Furthermore, $\lim_{n\to\infty} P_{2n}Q_{2n}^{-1} = \lim_{n\to\infty} P_{2n-1}Q_{2n-1}^{-1} = S_{21}S_{11}^{-1}$. Since the even and odd terms of the sequence have the same limit, we have $\lim_{r\to\infty} P_r Q_r^{-1} = S_{21}S_{11}^{-1}$. □

4. Solution of a quadratic operator equation

Let G_{ij}, $i, j = 1, 2$, be in $L(H)$. The operator equation under consideration is

$$XG_{12}X + XG_{11} - G_{22}X - G_{21} = 0, \tag{4.1}$$

which is to be solved for X in $L(H)$. Let the corresponding matrix $\boldsymbol{G}$ in $M(2, H)$ be given by

$$\boldsymbol{G} = \begin{bmatrix} G_{11} & G_{12} \\ G_{21} & G_{22} \end{bmatrix}.$$

We will need to consider the equivalent equation

$$XG_{12}X + X(G_{11} + \delta I) + (-G_{22} - \delta I)X - G_{21} = 0, \tag{4.2}$$

with corresponding matrix

$$\boldsymbol{G}_\delta = \begin{bmatrix} G_{11}+\delta & G_{12} \\ G_{21} & G_{22}+\delta \end{bmatrix} = \boldsymbol{G} + \delta \boldsymbol{I}, \tag{4.3}$$

in which $\delta > 0$ is arbitrary and will be assigned later.

If there exist operators

$$\boldsymbol{S} = \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix}, \qquad \boldsymbol{T} = \begin{bmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{bmatrix}, \qquad \boldsymbol{\Lambda} = \begin{bmatrix} \Lambda_{11} & 0 \\ 0 & \Lambda_{22} \end{bmatrix}$$

in $M(2, H)$ such that

$$\boldsymbol{T} = \boldsymbol{S}^{-1} \quad \text{and} \quad \boldsymbol{G}_\delta = \boldsymbol{S}\boldsymbol{\Lambda}\boldsymbol{T}, \tag{4.4}$$

then a direct computation and comparison of matrix elements shows that $X = S_{21}S_{11}^{-1}$ is a solution to (4.2) (and therefore to (4.1)). Note that there are no conditions placed on $\boldsymbol{S}$, $\boldsymbol{T}$ and $\boldsymbol{\Lambda}$ other than those implied in (4.4). The similarity between (4.4) and (3.10) suggests that one might be able to find a two-periodic continued fraction

$$\frac{A_1}{B_1+} \quad \frac{A_2}{B_2+} \quad \frac{A_1}{B_1+\cdots} \tag{4.5}$$

such that the corresponding matrix $\boldsymbol{A}$ is equal to $\boldsymbol{G}_\delta$. This may not be possible for $\boldsymbol{G}$, particularly if the elements G_{ij} are not invertible, but it can be done for $\boldsymbol{G}_\delta$, if δ is large enough. In fact, if $\delta > \|G_{22}\|$, then $G_{22} + \delta I = \delta[(1/\delta)G_{22} + I]$ is invertible, and a comparison of the entries of (4.5) with those of the middle matrix in (3.4) shows that $\boldsymbol{G}_\delta$ will equal $\boldsymbol{A}$ if

$$\begin{aligned} A_1 &= G_{22} + \delta I, & B_1 &= G_{12}, \\ A_2 &= (G_{11} + \delta I) - G_{12}(G_{22} + \delta I)^{-1}G_{21}, & B_2 &= (G_{22} + \delta I)^{-1}G_{21}. \end{aligned} \tag{4.6}$$

Whether or not $\boldsymbol{G}_\delta$ arises from a continued fraction, we can apply the same analysis to it that we used for $\boldsymbol{A}$ prior to Theorem 3.1 (and note that the $\boldsymbol{S}$ and $\boldsymbol{T}$ of that theorem will be the same $\boldsymbol{S}$ and $\boldsymbol{T}$ of (4.4)). That analysis does not depend on the nature of the entries of $\boldsymbol{A}$, but only on relation (3.10) (which in this case is given by (4.4)), the relative locations of the entries in the powers of $\boldsymbol{A}$, and the conditions on $\boldsymbol{S}$, $\boldsymbol{T}$ and $\boldsymbol{\Lambda}$. We get the following immediate result.

Theorem 4.1. *If δ in* (4.2) *can be chosen so that*

(i) *Condition* (α) *of Section* 3 *holds for the matrix* $\boldsymbol{G}_\delta$;

(ii) *The entries* S_{11}, S_{21}, T_{11}, T_{12} (*and therefore all entries*) *of the corresponding matrices* $\boldsymbol{S}$ *and* $\boldsymbol{T}$ *are invertible.*

Then,

(a) *if* $G_{22} + \delta I$ *is invertible,* (4.5) *converges to a solution of* (4.2) (*and therefore* (4.1)) *if* A_1, A_2, B_1, B_2 *are defined by* (4.6);

(b) *in any case if*

$$(\boldsymbol{G}_\delta)^n = \begin{bmatrix} \Gamma_{11}^{(n)} & \Gamma_{12}^{(n)} \\ \Gamma_{21}^{(n)} & \Gamma_{22}^{(n)} \end{bmatrix},$$

then for large n, $\Gamma_{11}^{(n)}$ and $\Gamma_{12}^{(n)}$ are invertible, and the sequences $\Gamma_{21}^{(n)}[\Gamma_{11}^{(n)}]^{-1}$ and $\Gamma_{22}^{(n)}[\Gamma_{12}^{(n)}]^{-1}$ converge to the solution $S_{21}S_{11}^{-1}$ of (4.1).

Remark 4.2. Condition (ii) above depends only on $\boldsymbol{G}$, and therefore holds for all δ if it holds for one. This is because if such an $\boldsymbol{S}$ and $\boldsymbol{T}$ produce a canonical form for a matrix, they will do the same when any multiple of the identity is added to that matrix.

5. Application to a quadratic matrix equation

We now apply the previous development to the important case of the matrix quadratic equation

$$XMX + XC + C^*X - D = 0, \tag{5.1}$$

associated with the steady-state solutions of the matrix Riccati equation. Here M and D are non-negative definite $r \times r$ matrices, and certain conditions (called *controllability* or *observability*) are often imposed on C. The situation arises in optimal control theory, and a host of other engineering applications (see [10]). The corresponding matrix is

$$\boldsymbol{G} = \begin{bmatrix} C & M \\ D & -C^* \end{bmatrix}, \tag{5.2}$$

which is a $2r \times 2r$ matrix.

Lemma 5.1. *The spectrum of $\boldsymbol{G}$ above is symmetric with respect to the imaginary axis.*

Proof. Let

$$\boldsymbol{T} = \begin{bmatrix} 0 & -I \\ I & 0 \end{bmatrix} \quad \text{and} \quad \boldsymbol{I} = \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix}.$$

Then it is easily seen that $\boldsymbol{T}^2 = -\boldsymbol{I}$ and

$$\boldsymbol{TGT} = \begin{bmatrix} C^* & D \\ M & -C \end{bmatrix} = \boldsymbol{G}^*.$$

For any complex number λ, $\boldsymbol{T}(\boldsymbol{G} - \lambda \boldsymbol{I})\boldsymbol{T} = \boldsymbol{TGT} - \lambda \boldsymbol{T}^2 = \boldsymbol{G}^* + \lambda \boldsymbol{I}$. Thus $\lambda \in \sigma(\boldsymbol{G})$ (where σ denotes spectrum) if and only if $-\lambda \in \sigma(\boldsymbol{G}^*)$ if and only if $-\overline{\lambda} \in \sigma(\boldsymbol{G})$. □

This adaptation of a proof in [10] applies equally well if the matrices are operators on a Hilbert space.

Theorem 5.2. *Let $\boldsymbol{G}$ be the $2r \times 2r$ matrix given above. Then the following holds.*

(1) *If $\boldsymbol{G}$ has no purely imaginary eigenvalues, and $\sigma_1 = \{\lambda_1, \lambda_2, \ldots, \lambda_k\}$ is the set of distinct eigenvalues of $\boldsymbol{G}$ having positive real part, then $\boldsymbol{G}_\delta$ will satisfy* (i) *of Theorem* 4.1 (*i.e., condition* (α) *of Section* 3) *if and only if*

$$\delta > \max_{1 \leqslant i,j \leqslant k} \frac{|\lambda_i|^2 - |\lambda_j|^2}{2\,\mathrm{Re}(\lambda_i + \lambda_j)}. \tag{5.3}$$

(2) *If both conditions of Theorem* 4.1 *hold for some* G_δ, *then the iterative process described in part* (b) *of that theorem will converge to a solution of* (5.1). *The rate of convergence of this process can be described as follows. If, as in Theorem* 4.1(b),

$$(G_\delta)^n = \begin{bmatrix} \Gamma_{11}^{(n)} & \Gamma_{12}^{(n)} \\ \Gamma_{21}^{(n)} & \Gamma_{22}^{(n)} \end{bmatrix} = SDT = SDS^{-1},$$

with D *blockdiagonal, so that* $\lim_{n\to\infty} \Gamma_{21}^{(n)}[\Gamma_{11}^{(n)}]^{-1} = S_{21}S_{11}^{-1}$ *is a solution of* (5.1), *then for large enough* n,

$$S_{21}S_{11}^{-1} - \Gamma_{21}^{(n)}[\Gamma_{11}^{(n)}]^{-1} = \frac{p(n)(C_\delta)^n}{1 - q(n)(C_\delta)^n}, \tag{5.4}$$

where $p(n)$ *and* $q(n)$ *are polynomials in* n *of degree* r, *and*

$$C_\delta = \max_{1\leqslant i,j\leqslant k} \frac{|\lambda_i - \delta|}{|\lambda_j + \delta|}. \tag{5.5}$$

A similar result holds for $S_{21}S_{11}^{-1} - \Gamma_{22}^{(n)}[\Gamma_{12}^{(n)}]^{-1}$ *with different polynomials* p *and* q.

(3) *If* G *has a diagonal Jordan canonical form, and the conditions of Theorem* 4.1 *hold for some* G_δ (*in particular* G *should have no purely imaginary eigenvalues*), *then the solution of* (5.1) *referred to in part* (2) *will be positive definite, and is the unique solution of* (5.1) *with that property. Moreover, in this case, the rate of convergence of the iterative process is eventually geometric of order* C_δ; *specifically, for large enough* n,

$$S_{21}S_{11}^{-1} - \Gamma_{21}^{(n)}[\Gamma_{11}^{(n)}]^{-1} = \frac{a(C_\delta)^n}{1 - b(C_\delta)^n}, \tag{5.6}$$

where a *and* b *are constants. Again a similar result holds for* $S_{21}S_{11}^{-1} - \Gamma_{22}^{(n)}[\Gamma_{12}^{(n)}]^{-1}$.

In particular, (5.4) *and* (5.6) *tell us that we should choose* δ *so as to minimize* (5.5), *subject to the constraint of* (5.3).

Proof. (1) Since G has no imaginary eigenvalues, the preceding lemma shows that $\sigma(G) = s_1 \cup -\bar{s}_1 \equiv s_1 \cup s_2$. Therefore G has a Jordan canonical form

$$\Lambda = \begin{bmatrix} \Lambda_1 & 0 \\ 0 & -\overline{\Lambda}_1 \end{bmatrix},$$

where Λ_1 is $r \times r$, and has diagonal elements only from s_1. Let S be the matrix that implements the similarity between G and Λ. Then,

$$S^{-1}G_\delta S = S^{-1}(G+\delta I)S = \begin{bmatrix} \Lambda_1 + \delta & 0 \\ 0 & \delta - \overline{\Lambda}_1 \end{bmatrix}. \tag{5.7}$$

Thus the eigenvalues of G_δ fall into two sets: $\sigma_1 = \{\lambda_1 + \delta, \lambda_2 + \delta, \ldots, \lambda_k + \delta\}$ corresponding to Λ_1 and $\sigma_2 = \{\delta - \bar{\lambda}_1, \delta - \bar{\lambda}_2, \ldots, \delta - \bar{\lambda}_k\}$ corresponding to Λ_2. The form of (5.5) shows that, with the above definition of σ_1 and σ_2, (i) and (ii) of condition (α) hold. Part (iii) of (α) will hold if and only if $\min_{1\leqslant i\leqslant k} |\lambda_i + \delta| > \max_{1\leqslant j\leqslant k} |\delta - \bar{\lambda}_j|$. Equivalently,

$$1 > C_\delta \equiv \frac{\max_{1\leqslant j\leqslant k}|\delta-\overline{\lambda}_j|}{\min_{1\leqslant i\leqslant k}|\lambda_i+\delta|} = \max_{1\leqslant i,j\leqslant k}\frac{|\delta-\overline{\lambda}_j|}{|\lambda_i+\delta|} = \max_{1\leqslant i,j\leqslant k}\frac{|\lambda_j-\delta|}{|\lambda_i+\delta|}. \tag{5.8}$$

Thus we need that for every $1 \leqslant i, j \leqslant k$,

$$0 < |\lambda_i+\delta|^2 - |\lambda_j-\delta|^2 = |\lambda_i|^2 - |\lambda_j|^2 + 2\delta\,\mathrm{Re}(\lambda_i) + 2\delta\,\mathrm{Re}(\lambda_j).$$

Since the real parts of the λ_i are positive, we get that condition (α) holds if and only if (5.3) holds.

(2) The proof of [2, Lemma 3.1] shows that if σ_1 and σ_2 are as above, then

$$\left(\|(\Lambda_1+\delta)^{-1}\|\,\|\delta-\overline{\Lambda}_1\|\right)^n \leqslant p'(n)(C_\delta)^n, \tag{5.9}$$

for sufficiently large n, where $p'(n)$ is a polynomial of degree n. This follows from the observation that if N is a nilpotent matrix of order r, then $N^s = 0$ when $s \geqslant r$. Now, borrowing the notation of the proof of Theorem 3.1, we have

$$\Gamma_{21}^{(n)}\left[\Gamma_{11}^{(n)}\right]^{-1} = \left(I + X_n^{(1)}\right)S_{21}S_{11}^{-1}\left(I + X_n^{(2)}\right)^{-1}. \tag{5.10}$$

Then, using (5.9), we have

$$\|X_n^{(1)}\| \leqslant \|S_{22}\|\,\|T_{21}\|\,\|T_{11}^{-1}\|\,\|S_{21}^{-1}\|\left(\|(\Lambda_1+\delta)^{-1}\|\,\|\delta-\overline{\Lambda}_1\|\right)^n \leqslant p_1(n)(C_\delta)^n, \tag{5.11}$$

and similar computations hold for $X_n^{(2)}$. For sufficiently large n, $X_n^{(2)} < 1$, and so $(I + X_n^{(2)})^{-1}$ is expandable in the usual geometric series, convergent in matrix norm. This is dominated by the corresponding norm series. Thus if we use (5.10) and (5.11), and sum the resulting scalar series, we have

$$\|S_{21}S_{11}^{-1} - \Gamma_{21}^{(n)}\left[\Gamma_{11}^{(n)}\right]^{-1}\| \leqslant \|S_{21}\|\,\|S_{11}^{-1}\|\left[\frac{p_1(n)(C_\delta)^n}{1-p_2(n)(C_\delta)^n}\right].$$

(3) In [10] it is shown that, under the conditions we have specified for (5.1), a unique positive definite solution will exist if $\boldsymbol{G}$ has diagonal Jordan canonical form and satisfies invertibility conditions on some of the matrices S_{ij}. Then the solution is constructed from these S_{ij} and has a similar, but not identical, form to our solution. Since the forms are not identical, and as a convenience to the reader, we briefly sketch a modification of some of his arguments to show that our solution $S_{21}[S_{11}]^{-1}$ is the desired positive definite solution to (5.1).

If $\boldsymbol{\Lambda}$ is diagonal with diagonal elements $\{\lambda_1, \lambda_2, \ldots, \lambda_{2n}\}$, and if the columns of $\boldsymbol{S}$ are $\{\overline{a}_1, \overline{a}_2, \ldots, \overline{a}_{2n}\}$, then $\boldsymbol{G}\overline{a}_k = \lambda_k\overline{a}_k$, $k = 1, \ldots, 2n$, since $\boldsymbol{S}^{-1}\boldsymbol{G}\boldsymbol{S} = \boldsymbol{\Lambda}$. Let P be the $n \times n$ matrix $S_{11}^*S_{21}$, and let $\boldsymbol{T}$ be as in Lemma 5.1. Then an easy computation shows that

$$\text{if } \boldsymbol{S}^*\boldsymbol{T}\boldsymbol{S} \equiv \begin{bmatrix} \Gamma & \cdot \\ \cdot & \cdot \end{bmatrix}, \text{ then } \Gamma = S_{12}^*S_{11} - S_{11}^*S_{21} = P^* - P.$$

By considering the columns of $\boldsymbol{S}$, we have that

$$[P^* - P]_{ij} = [\Gamma]_{ij} = \overline{a}_i^*\boldsymbol{T}\overline{a}_j, \quad \text{for } 1 \leqslant i, j \leqslant n.$$

We can see in a similar way that if

$$\boldsymbol{\Lambda}^*\boldsymbol{S}^*\boldsymbol{T}\boldsymbol{S} \equiv \begin{bmatrix} \Phi & \cdot \\ \cdot & \cdot \end{bmatrix} \quad \text{and} \quad \boldsymbol{S}^*\boldsymbol{T}\boldsymbol{S}\boldsymbol{\Lambda} \equiv \begin{bmatrix} \Theta & \cdot \\ \cdot & \cdot \end{bmatrix},$$

then

$$[\Phi]_{ij} = \overline{\lambda}_i \overline{a}_i^* T \overline{a}_j \quad \text{and} \quad [\Theta]_{ij} = \overline{a}_i^* T \overline{a}_j \lambda_j, \qquad \text{for } 1 \leqslant i, j \leqslant n.$$

Now since the first n columns of P correspond to eigenvalues in set s_1 and these have positive real parts, we can never have $\overline{\lambda}_i + \lambda_j = 0$ if $1 \leqslant i, j \leqslant n$. Thus,

$$[P^* - P]_{ij} = \overline{a}_i^* T \overline{a}_j = \frac{1}{(\overline{\lambda}_i + \lambda_j)} [\overline{\lambda}_i \overline{a}_i^* T \overline{a}_j + \overline{a}_i^* T \overline{a}_j \lambda_j] = [\Phi + \Theta]_{ij}, \quad \text{for } 1 \leqslant i, j \leqslant n.$$

On the other hand, we note that $\Lambda^* S^* T S = S^* G^* T S = S^* (TGT) T S$ (see the proof of Lemma 5.1) $= -S^* T G S = -S^* T S \Lambda$. Thus,

$$\mathbf{0} = \Lambda^* S^* T S + S^* T S \Lambda = \begin{bmatrix} \Phi + \Theta & \cdot \\ \cdot & \cdot \end{bmatrix}.$$

Thus $P^* - P = 0$ and P is self-adjoint. But then so is $S_{21} S_{11}^{-1} = [S_{11}^{-1}]^* S_{11}^* S_{21} S_{11}^{-1} = [S_{11}^{-1}]^* P S_{11}^{-1}$. The proof that $S_{21} S_{11}^{-1}$ is nonnegative definite is obtained by modifying the proof in [10] a similar way.

Finally, if the Jordan form of G (and therefore $\delta I \pm G$) is diagonal, then the polynomial $p'(n)$ of (5.9) reduces to a constant. Thus (5.4) reduces to (5.6). □

Remark 5.3. If a reasonably good estimate of the eigenvalues is available, one can often plot (5.4) as a function of δ, using easily available software products. This provides a graphical way to approximate the optimal δ and corresponding C_δ.

Remark 5.4. Instead of computing the powers $(G_\delta)^n$, it is just as easy to compute $(G_\delta)^{2^n}$, by squaring, then squaring that result, etc. In practice this is what we do, resulting in far better convergence properties.

Discussion. The usual way of solving (5.1) is to find precisely the eigenvalues and eigenvectors of the corresponding matrix G in (5.2), use these to construct the matrix S that diagonalizes G, and then construct the solution $S_{21} S_{11}^{-1}$ or a closely related one (see [10]). This procedure is numerically costly and it is difficult to get precise error estimates for the solution in terms of error analysis for the eigenvectors. By contrast, our method is a simple adaptation of the familiar power method, and requires only enough knowledge of the eigenvalues to verify (5.3) and the fact that none of them are imaginary. No knowledge of eigenvectors is needed to approximate the solution.

It is certainly true that verification of the invertibility conditions on the entries of the similarity matrix S and its inverse (Theorem 4.1(ii)) would require a knowledge of the eigenvectors of G, or at least of certain geometric properties of the eigenspaces. However, if one has a good estimate of the rate of convergence (5.4) or (5.6), one may certainly apply the method formally to the point where one should have a solution to the desired accuracy, and check the result in (5.1). Thus it is usually not necessary to verify the invertibility conditions directly.

The invertibility conditions do not always hold for a controllable system. Experimentation has shown that when this is true, the algorithm very quickly becomes singular, or at least violates convergence formula (5.4).

6. Numerical examples

We consider several low-order examples from standard treatises on control theory. The exact answers are given, so as to evaluate the convergence.

Example 6.1. Consider the matrix quadratic equation

$$X\begin{bmatrix}0 & 0\\ 0 & 1\end{bmatrix}X+X\begin{bmatrix}0 & -1\\ 0 & 1\end{bmatrix}+\begin{bmatrix}0 & 0\\ -1 & 1\end{bmatrix}X-\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}=\begin{bmatrix}0 & 0\\ 0 & 0\end{bmatrix}. \tag{6.1}$$

The corresponding matrix G is given by

$$G=\begin{bmatrix}0 & -1 & 0 & 0\\ 0 & 1 & 0 & 1\\ 1 & 0 & 0 & 0\\ 0 & 1 & 1 & -1\end{bmatrix},$$

and it is easily seen that the eigenvalues of G are 1 and -1, each of algebraic multiplicity two, and so $\sigma_1=\{1\}$ (Theorem 5.2(1)).

Clearly, $\delta=1$ produces a best C_δ of 0. We have two choices of approximating sequence: $\Gamma_{21}^{(n)}[\Gamma_{11}^{(n)}]^{-1}$ and $\Gamma_{22}^{(n)}[\Gamma_{12}^{(n)}]^{-1}$ and for $n=2$ we get the exact unique positive definite solution (easily verified) of $\left[\begin{smallmatrix}2 & 1\\ 1 & 1\end{smallmatrix}\right]$.

Example 6.2. Again consider the matrix quadratic equation

$$X\begin{bmatrix}1 & 0\\ 0 & 0\end{bmatrix}X+X\begin{bmatrix}0 & 0\\ -1 & \sqrt{2}\end{bmatrix}+\begin{bmatrix}0 & 1\\ 0 & -\sqrt{2}\end{bmatrix}X-\begin{bmatrix}0 & 0\\ 0 & 1\end{bmatrix}=\begin{bmatrix}0 & 0\\ 0 & 0\end{bmatrix}, \tag{6.2}$$

whose unique positive definite solution is known to be

$$\begin{bmatrix}2-\sqrt{2} & 3-2\sqrt{2}\\ 3-2\sqrt{2} & 6-4\sqrt{2}\end{bmatrix}.$$

In this case,

$$G=\begin{bmatrix}0 & 0 & 1 & 0\\ -1 & \sqrt{2} & 0 & 0\\ 0 & 0 & 0 & 1\\ 0 & 1 & 0 & -\sqrt{2}\end{bmatrix},$$

and the eigenvalues are again 1 and -1, of multiplicity two.

As in the previous example, $\delta=1$ is allowable and produces the minimal C_δ of 0. Eq. (5.6) then implies that for some n, $\Gamma_{22}^{(n)}[\Gamma_{12}^{(n)}]^{-1}$ will be the exact answer. Recall that we write

$$(G_\delta)^n=\begin{bmatrix}\Gamma_{11}^{(n)} & \Gamma_{12}^{(n)}\\ \Gamma_{21}^{(n)} & \Gamma_{22}^{(n)}\end{bmatrix}.$$

Computing symbolically, we have

$$(G_\delta)^2 = \begin{bmatrix} 1 & 0 & 2 & 1 \\ -2-\sqrt{2} & (1+\sqrt{2})^2 & -1 & 0 \\ 0 & 1 & 1 & 2-\sqrt{2} \\ -1 & 2 & 0 & (-1+\sqrt{2})^2 \end{bmatrix},$$

and so

$$\Gamma_{21}^{(2)}\left[\Gamma_{11}^{(2)}\right]^{-1} = \begin{bmatrix} 0 & 1 \\ -1 & 2 \end{bmatrix}\begin{bmatrix} 1 & 0 \\ -2-\sqrt{2} & (1+\sqrt{2})^2 \end{bmatrix}^{-1}$$

$$= \begin{bmatrix} 2+\sqrt{2}/(1+\sqrt{2})^2 & 1/(1+\sqrt{2})^2 \\ 1/(1+\sqrt{2})^2 & 2/(1+\sqrt{2})^2 \end{bmatrix} = \begin{bmatrix} 2-\sqrt{2} & 3-2\sqrt{2} \\ 3-2\sqrt{2} & 6-4\sqrt{2} \end{bmatrix}.$$

Thus $\Gamma_{21}^{(2)}\left[\Gamma_{11}^{(2)}\right]^{-1}$ is already exactly correct. We may also do the approximation with the sequence $\Gamma_{22}^{(n)}\left[\Gamma_{12}^{(n)}\right]^{-1}$, which again is exact for $n = 2$.

References

[1] C. Ahlbrandt, Continued fraction representations of maximal and minimal solutions of a discrete matrix Riccati equation, *SIAM J. Math. Anal.* **24** (1993) 1597–1621.

[2] R.C. Busby and W. Fair, Iterative solution of spectral operator polynomial equations and a related continued fraction, *SIAM J. Math. Anal. Appl.* **50** (1975) 113–134.

[3] R. Busby and W. Fair, Convergence of 'periodic in the limit' operator continued fractions, *SIAM J. Math. Anal.* **10** (1979) 512–522.

[4] H. Denk and M. Riederle, A generalization of a theorem of Pringshiem, *J. Approx. Theory* **35** (1982) 355–363.

[5] W. Fair, Non commutative continued fractions, *SIAM J. Math. Anal.* **2** (1971) 226–232

[6] D. Field, Convergence theorems for matrix continued fractions, *SIAM J. Math. Anal.* **10** (1979) 1220–1227.

[7] T. Hayden, Continued fractions in Banach spaces, *Rocky Mountain J. Math.* **4** (1974) 367–370.

[8] N. Negoescu, Convergence theorems on non commutative continued fractions, *Rev. Anal. Numér. Théor. Approx.* **5** (1976) 165–180.

[9] S. Peng and A. Hessel, Convergence of non commutative continued fractions, *SIAM J. Math. Anal.* **6** (1975) 724–727.

[10] J.E. Potter, Matrix quadratic solutions, *J. SIAM Appl. Math.* **14** (3) (1966) 496–501.

[11] P. Wynn, Continued fractions whose coefficients obey a non-commutative law of multiplication, *Arch. Rational Mech. Anal.* **12** (1963) 273–312.

[12] A. Zemanian, Non-uniform semi-infinite grounded grids, *SIAM J. Math. Anal.* **13** (1982) 770–788.