Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER 17636.17-MA	2. GOVT ACCESSION NO. AD A130 721	3. RECIPIENT'S CATALOG NUMBER N/A
4. TITLE (and Subtitle) Some Properties of Matrix Sign Functions Derived from Continued Fractions		5. TYPE OF REPORT & PERIOD COVERED Reprint 6. PERFORMING ORG. REPORT NUMBER N/A
7. AUTHOR(s) L. S. Shieh Y. T. Tsay R. E. Yates		8. CONTRACT OR GRANT NUMBER(s) DAAG29 80 K 0077
9. PERFORMING ORGANIZATION NAME AND ADDRESS University of Houston Houston, TX 77004		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS N/A
11. CONTROLLING OFFICE NAME AND ADDRESS U. S. Army Research Office P. O. Box 12211 Research Triangle Park, NC 27709		12. REPORT DATE May 83 13. NUMBER OF PAGES 8
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) Unclassified 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Submitted for announcement only.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

Some properties of matrix sign functions derived from continued fractions

Prof. L.S. Shieh, Ph.D., Y.T. Tsay, M.Sc., and R.E. Yates, Ph.D.

Indexing terms: Mathematical techniques, Approximation theory, Eigenfunctions, Linear systems.

Abstract: This paper proposes an alternate representation of a matrix sign function based on an irrational function described by a continued fraction. The properties of the continued fraction and the truncated continued fractions are investigated. Also, new algorithms for computing the matrix sign function are developed. The matrix sign function is then extended to a generalised matrix sign function for directly solving discrete-time system problems.

1 Introduction

Since Roberts [1] initially introduced the matrix sign function and its applications to linear systems, many applications for solving system problems have been developed [1, 2, 3]. A Newton-Raphson type algorithm proposed by Roberts [1] and an improved algorithm by Balzer [4] have been used as standard algorithms for computing the matrix sign function.

One main feature of the matrix sign function is that it preserves the eigenvectors of the original matrix. This property is useful both for studying the eigenstructures of matrices and for developing applications to engineering problems.

The matrix sign function S of a square matrix $A \in \mathbb{C}^{n\times n}$ with $\mathrm{Re}(\sigma(A)) \neq 0$ is defined by [1]

$$S \triangleq \mathrm{sign}(A) = 2\,\mathrm{sign}^{+}(A) - I_n \tag{1a}$$

where I_n is an $n \times n$ identity matrix and

$$\mathrm{sign}^{+}(A) = \frac{1}{2\pi i}\oint_{c_+} (\lambda I_n - A)^{-1}\,d\lambda \tag{1b}$$

c_+ is a simple closed contour in right halfplane of λ and encloses all the right-halfplane eigenvalues of A.

Following the definition of eqn. 1, if A has a Jordan form:

$$J = \text{block diag}\,[J_+, J_-] \triangleq J_+ \oplus J_- \tag{2a}$$

then

$$A = MJM^{-1} \tag{2b}$$

and

$$\begin{aligned} S = \mathrm{sign}(A) &= M[\mathrm{sign}(J_+) \oplus \mathrm{sign}(J_-)]M^{-1} \\ &= M[I_{n1} \oplus (-I_{n2})]M^{-1} \end{aligned} \tag{2c}$$

where $J_+ \in \mathbb{C}^{n1\times n1}$, $J_- \in \mathbb{C}^{n2\times n2}$, and $n = n1 + n2$ are the collection of Jordan blocks with $\mathrm{Re}(\sigma(A)) > 0$ and $\mathrm{Re}(\sigma(A)) < 0$, respectively. $M \in \mathbb{C}^{n\times n}$ is a modal matrix of A.

The extended matrix sign function $\hat{S}$ of A including $\mathrm{Re}(\sigma(A)) = 0$ is defined by [3]

$$\begin{aligned} \hat{S} \triangleq \mathrm{sign}(A) &= M[\mathrm{sign}(J_+) \oplus \mathrm{sign}(J_-) \oplus \mathrm{sign}(J_0)]M^{-1} \\ &= M[I_{n1} \oplus (-I_{n2}) \oplus 0_{n3}]M^{-1} \end{aligned} \tag{3}$$

where $J_0 \in \mathbb{C}^{n3\times n3}$ is the collection of Jordan blocks with $\mathrm{Re}(\sigma(A)) = 0$, 0_{n3} is an $n3 \times n3$ null matrix, and $n1 + n2 + n3 = n$.

Paper 2474 D, first received 10th March 1982 and in revised form 7th February 1983

Prof. Shieh and Mr. Tsay are with the Department of Electrical Engineering, University of Houston Central Campus, Houston, TX 77004, USA, and Dr. Yates is Director of the Guidance and Control Directorate, US Army Missile Command, Redstone Arsenal, AL 35809, USA

A recursive scheme for computing the matrix sign function, given by Roberts [1] and improved by Balzer [4], is as follows:

$$S_{k+1} = \alpha_k S_k + \beta_k S_k^{-1}; \qquad S_0 = A$$

$$\alpha_k + \beta_k = 1 \quad \text{and} \quad \lim_{k\to\infty} \alpha_k = \lim_{k\to\infty} \beta_k = \tfrac{1}{2},$$

$$\text{if} \quad \mathrm{Re}(\sigma(A)) \neq 0 \tag{4}$$

The algorithm for the extended matrix sign function is given by [3]

$$\hat{S}_{k+1} = \hat{S}^{+}_{k+1} + \hat{S}^{-}_{k+1} \tag{5a}$$

where $\hat{S}^{+}_{k+1}$ and $\hat{S}^{-}_{k+1}$ denote the $(k+1)$th iteration of the matrix sign algorithm in eqn. 4 using $\hat{S}^{+}_0 = A + \epsilon I_n$ and $\hat{S}^{-}_0 = A - \epsilon I_n$, respectively. ϵ is given by

$$\epsilon = \frac{\mu}{\|A^D\|} \tag{5b}$$

where $0 < \mu < 1$ and A^D is the Drazin inverse of A [3].

The algorithm in eqn. 4 is known to be a Newton-Raphson type and is often used as a standard method for computing the matrix sign function.

In this paper, we define an alternate representation of matrix sign functions based on an irrational function described by a continued fraction. The properties of the irrational function and the convergence of the truncated continued fractions are investigated. This leads to new algorithms for computing the matrix sign function. It is shown that the standard algorithm in Eqn. 4 is a special case of the new algorithms derived. The matrix sign function is then extended to a generalised matrix sign function for directly solving discrete-time system problems.

2 Scalar sign function

In order to develop a new algorithm for computing the matrix sign function, we define a new representation of a (scalar) sign function as follows.

Definition

The (scalar) sign function of a complex value $\lambda \in \sigma(A)$ is defined by

$$\mathrm{sign}(\lambda) = \begin{cases} +1 & \text{if} \quad \mathrm{Re}(\lambda) > 0 \\ -1 & \text{if} \quad \mathrm{Re}(\lambda) < 0 \end{cases} \tag{6a, 6b}$$

$$= \begin{cases} \dfrac{\lambda}{g(\lambda)} & \text{if} \quad \mathrm{Re}(\lambda) > 0 \\ \dfrac{\lambda}{g(\lambda)} & \text{if} \quad \mathrm{Re}(\lambda) < 0 \end{cases} \tag{6c, 6d}$$

$$= \begin{cases} \dfrac{g(\lambda)}{\lambda} & \text{if} \quad \mathrm{Re}(\lambda) > 0 \qquad (6e) \\ \dfrac{g(\lambda)}{\lambda} & \text{if} \quad \mathrm{Re}(\lambda) < 0 \qquad (6f) \end{cases}$$

where $g(\lambda)$ is defined by

$$g(\lambda) = \begin{cases} \lambda & \text{if} \quad \mathrm{Re}(\lambda) > 0 \\ -\lambda & \text{if} \quad \mathrm{Re}(\lambda) < 0 \end{cases}$$

$$= \begin{cases} |\lambda| e^{j\phi} & \text{if} \quad -\dfrac{\pi}{2} < \phi < \dfrac{\pi}{2} \qquad (6g) \\ |\lambda| e^{j(\phi - \pi)} & \text{if} \quad \dfrac{\pi}{2} < \phi < \dfrac{3\pi}{2} \qquad (6h) \end{cases}$$

Note that $\mathrm{Re}(\lambda) = 0$ is not included in the definition.

When λ is a real value, σ, then $g(\lambda)$ in eqns. 6*g* and 6*h* is obviously an absolute value function and

$$g(\sigma) = \sqrt{\sigma^2} \qquad (7)$$

In a similar fashion, when λ is a complex function, $g(\lambda)$ may be defined by

$$g(\lambda) = \sqrt{\lambda^2} \qquad (8)$$

with proper selection of branch cut to match the definitions of eqn. 7 and eqn. 8. To derive the function $g(\lambda)$ in eqn. 6 or eqn. 8, we consider a square-root function $f: z \to \sqrt{z}$ that has the branch cut on the negative real axis and the first Riemann sheet with $|\arg(z)| < \pi$ as the domain of z. The continued fraction expansion of $\sqrt{z}$ is given by [5]

$$f(z) = \sqrt{z} = 1 + \cfrac{z-1}{2 + \cfrac{z-1}{2 + \cfrac{z-1}{\ddots}}} \qquad (9)$$

In order to study the domain of z so that the irrational function $f(z)$ can fully be described by the continued fraction expansion in Eqn. 9, we investigate the properties of the continued fraction.

Define the kth truncation of $f(z)$ as $f_k(z)$, or

$$f_k(z) = 1 + 2f_{k-1}^*(z) \qquad (10a)$$

where $f_0^*(z) = 0$ and

$$f_k^*(z) \triangleq \cfrac{\frac{1}{4}(z-1)}{1 + \cfrac{\frac{1}{4}(z-1)}{1 + \cfrac{\frac{1}{4}(z-1)}{\ddots \; \frac{1}{4}(z-1)}}} \triangleq \frac{a_k}{b_k}, \quad k \geq 1 \qquad (10b)$$

The recursive forms for a_k and b_k are

$$a_k = a_{k-1} + \tfrac{1}{4}(z-1)a_{k-2} \qquad a_{-1} = 1, \quad a_0 = 0 \qquad (11a)$$

$$b_k = b_{k-1} + \tfrac{1}{4}(z-1)b_{k-2} \qquad b_{-1} = 0, \quad b_0 = 1 \qquad (11b)$$

The difference equation for both a_k and b_k satisfies

$$1 - q^{-1} - \tfrac{1}{4}(z-1)q^{-2} = 0 \qquad (11c)$$

where q^{-1} is a backward shift operator, or $q^{-1} a_k = a_{k-1}$.

The zeros of the equation in eqn. 11*c* become

$$q_1 = \frac{1 + \sqrt{z}}{2} \quad \text{and} \quad q_2 = \frac{1 - \sqrt{z}}{2} \qquad (11d)$$

From eqn. 11*d* we observe that

$$|q_1| \geq |q_2| \text{ if } -\pi < \arg(z) < \pi \qquad (11e)$$

The general form for eqn. 10*a* (see Reference 5) is as follows:

$$f_k(z) = \frac{\sqrt{z}\,[(1+\sqrt{z})^k + (1-\sqrt{z})^k]}{(1+\sqrt{z})^k - (1-\sqrt{z})^k}$$

$$= \frac{\sum_{j=0}^{p} {}_kC_{2j} z^j}{\sum_{j=0}^{r} {}_kC_{2j+1} z^j} \qquad (12a)$$

where $p = \lfloor k/2 \rfloor$ and $r = \lfloor (k-1)/2 \rfloor$ are integers for $k \geq 1$ and ${}_kC_{2j}$ are the coefficients of a binomial expansion.

Substituting eqn. 11*d* into eqn. 12*a* yields

$$f_k(z) = \frac{\sqrt{z}\,(q_1^k + q_2^k)}{q_1^k - q_2^k} = \frac{\sqrt{z}\left[1 + \left(\frac{q_2}{q_1}\right)^k\right]}{1 - \left(\frac{q_2}{q_1}\right)^k}$$

$$= \sqrt{z}\left[1 + \left(\frac{q_2}{q_1}\right)^k\right]\left[1 - \left(\frac{q_2}{q_1}\right)^k\right]^{-1}$$

$$k \geq 1 \quad \text{and} \quad q_1 \neq q_2 \qquad (12b)$$

$f_k(z)$ for $k = 1, \ldots, 4$ are as follows:

$$f_1(z) = 1 \qquad (13a)$$

$$f_2(z) = \frac{z+1}{2} \qquad (13b)$$

$$f_3(z) = \frac{3z+1}{z+3} \qquad (13c)$$

$$f_4(z) = \frac{z^2 + 6z + 1}{4z + 4} \qquad (13d)$$

Some properties of the continued fractions in eqns. 9 and 10 which will be used to derive the $g(\lambda)$ in eqns. 6 and 8 are as follows:

Property 1

If $|q_1| > |q_2|$, then we have

$$f(z) = \lim_{k \to \infty} f_k(z) = \sqrt{z} \qquad (14a)$$

The important result in eqn. 14*a* can be verified by a ratio test of the series, which can be obtained by expanding $[1 - (q_2/q_1)^k]^{-1}$ in eqn. 12*b*. Also, the convergent condition $|q_1| > |q_2|$ implies that the domain of z is in $\bar{\mathbb{C}}^{\mathbb{R}^-}$ where $\mathbb{C}^{\mathbb{R}^-} \triangleq \mathbb{C} - \mathbb{R}^-$ and $\mathbb{R}^-$ is the negative real axis $\mathbb{R}^- = (-\infty, 0]$, or that the complex variable z must be $-\pi < \arg(z) < \pi$. Thus, the function $f(z)$ uniformly converges to the desired

function $\sqrt{z}$, and $\sqrt{z}$ can fully be represented by the continued fraction if $z \in \bar{\mathbb{C}}^{\mathbb{R}^-}$.

Property 2
If $|q_1| = |q_2|$ and $z = 0$, then

$$f(z) = \lim_{k\to\infty} f_k(z) = 0 \tag{14b}$$

Property 3
If $|q_1| = |q_2|$, $|z| \neq 0$, and $|\arg(z)| = \pi$, then

$$\lim_{k\to\infty} f_k(z) \to \infty \tag{14c}$$

The results in eqn. 14*b* and *c* can be verified as follows:

When $|q_1| = |q_2|$, from eqn. 11*d* we have $|1 + \sqrt{z}| = |1 - \sqrt{z}|$. If $z = 0$, then we have $a_k = -(k/2)(1/2)^k$ and $b_k = (k + 1)(1/2)^k$ in eqn. 11. Thus, $f_k(z)$ in eqn. 10*a* becomes $1/(k + 1)$ and

$$f(z) = \lim_{k\to\infty} f_k(z) = \lim_{k\to\infty} \frac{1}{k+1} \to 0 \tag{14d}$$

The function $f(z)$ converges to zero if $z = 0$.

On the other hand, when z is a nonzero negative real, then $\sqrt{z} = j\omega$. Thus $1 + \sqrt{z} = \sqrt{1+\omega^2}\ e^{j\phi}$ and $1 - \sqrt{z} = \sqrt{1+\omega^2}\ e^{-j\phi}$ where $\phi = \tan^{-1}\omega$. Therefore, we have

$$f_k(z) = j\omega \frac{1 + e^{-j2k\phi}}{1 - e^{-j2k\phi}} = \frac{\omega}{\tan(k\phi)} \tag{14e}$$

The function $f_k(z)$ diverges as $z \in (-\infty, 0)$.

Property 4
The poles and zeros of $f_k(z)$ and $k \geqslant 2$ alternate on the negative real axis.

From eqn. 12 we have the poles of $f_k(z)$

$$P_m = -\left(\tan\frac{m\pi}{k}\right)^2,$$

$$m = 1, 2, \ldots, \begin{cases} (k-1)/2 \text{ for odd } k \\ (k-2)/2 \text{ for even } k \end{cases} \tag{14f}$$

and the zeros

$$z_m = -\left[\tan\frac{(m-\frac{1}{2})\pi}{k}\right]^2$$

$$m = 1, 2, \ldots, \begin{cases} (k-1)/2 \text{ for odd } k \\ k/2 \quad\quad \text{for even } k \end{cases} \tag{14g}$$

Since the tangent function $\tan\theta$ is monotonically increasing for $0 < \theta < \pi/2$, the poles and zeros of $f_k(z)$ for $k \geqslant 2$ alternate on the negative real axis.

Using the function $f(z)$ and the properties obtained in eqn. 14 we are now able to derive the desired function $g(\lambda)$ in eqn. 8 using the principal square-root function $f(z)$ of eqn. 9. Let $z = \lambda^2$, we have

$$f(z) \triangleq g(\lambda) = \sqrt{\lambda^2}, \qquad \lambda \in \bar{\mathbb{C}}^{\mathrm{I}}$$

$$= 1 + \cfrac{\lambda^2 - 1}{2 + \cfrac{\lambda^2 - 1}{2 + \cfrac{\lambda^2 - 1}{\ddots}}} \tag{15a}$$

where $\bar{\mathbb{C}}^{\mathrm{I}} \triangleq \mathbb{C} - \mathrm{I}$ and I is the entire imaginary axis. From the properties shown in eqn. 14, the domain of z is in $\bar{\mathbb{C}}^{\mathbb{R}^-}$, therefore the domain of λ must be in $\bar{\mathbb{C}}^{\mathrm{I}}$. In other words, the function $g(\lambda)$ converges if

$$\lambda = |\lambda| e^{j\phi} \text{ and } |\phi| \neq \frac{\pi}{2} \tag{15b}$$

Based on the convergent condition derived in eqn. 14*a*, we have the desired function $g(\lambda)$ in eqn. 6 as follows:

$$g(\lambda) = \lim_{k\to\infty} g_k(\lambda) = \sqrt{\lambda^2}$$

$$= \begin{cases} \sqrt{|\lambda|^2 e^{j2\phi}} = |\lambda| e^{j\phi} = \lambda & \text{if } -\dfrac{\pi}{2} < \phi < \dfrac{\pi}{2} \\ \sqrt{|\lambda|^2 e^{j2(\phi-\pi)}} = |\lambda| e^{j(\phi-\pi)} = -\lambda & \text{if } \dfrac{\pi}{2} < \phi < \dfrac{3\pi}{2} \end{cases} \tag{16a, 16b}$$

where

$$g_k(\lambda) = f_k(z) \quad \text{with} \quad z = \lambda^2 \tag{16c}$$

Thus, the scalar sign functions defined in eqn. 6 can be expressed by

$$\text{sign}(\lambda) \triangleq \frac{\lambda}{g(\lambda)} = \lim_{k\to\infty} \text{sign}_{(k)}^{\{1\}}(\lambda) \tag{17a}$$

or

$$\text{sign}(\lambda) \triangleq \frac{g(\lambda)}{\lambda} = \lim_{k\to\infty} \text{sign}_{(k)}^{\{2\}}(\lambda) \tag{17b}$$

where

$$\text{sign}_{(k)}^{\{1\}}(\lambda) \triangleq \frac{\lambda}{g_k(\lambda)} = \frac{(1+\lambda)^k - (1-\lambda)^k}{(1+\lambda)^k + (1-\lambda)^k} \tag{17c}$$

$$\text{sign}_{(k)}^{\{2\}}(\lambda) \triangleq \frac{g_k(\lambda)}{\lambda} = \frac{(1+\lambda)^k + (1-\lambda)^k}{(1+\lambda)^k - (1-\lambda)^k} \tag{17d}$$

and

$$g_k(\lambda) = \lambda \frac{(1+\lambda)^k + (1-\lambda)^k}{(1+\lambda)^k - (1-\lambda)^k} \qquad k = 1, 2, \ldots \tag{17e}$$

is the kth truncation of the continued fraction

$$g(\lambda) = \lim_{k\to\infty} g_k(\lambda)$$

$$= 1 + \cfrac{\lambda^2 - 1}{2 + \cfrac{\lambda^2 - 1}{2 + \cfrac{\lambda^2 - 1}{\ddots}}} \qquad \lambda \in \bar{\mathbb{C}}^{\mathrm{I}} \tag{17f}$$

3 Matrix sign function

The scalar sign function derived in Section 2 can be extended to a matrix sign function. For this extension we need to investigate a matrix function generated by a scalar analytic function.

Consider a matrix $A \in \mathbb{C}^{n\times n}$ with spectra $\sigma(A) = \{\lambda_1, \ldots, \lambda_l\}$ where $l \leqslant n$. If a scalar function $p(\lambda)$ is analytic at λ_j, $j = 1, \ldots, l$, then the matrix function $P(A)$ generated by

$p(\lambda)$ can be defined as [6]

$$P(A) = \frac{1}{2\pi i} \oint_c p(\lambda)(\lambda I_n - A)^{-1} d\lambda \tag{18}$$

where c is a simple closed contour which encloses $\lambda_j, j = 1, \ldots, l$. The matrix function described in eqn. 18 has the following properties [6]:

Lemma 1
Let A be defined as above and $p(\lambda)$, $q(\lambda)$ and $r(\lambda)$ are analytic at $\lambda_j \in \sigma(A)$, $\lambda_j = 1, \ldots, l \leqslant n$, then

(i) if $p(\lambda) = k$, then $P(A) = kI_n$
(ii) if $p(\lambda) = \lambda$, then $P(A) = A$
(iii) if $p(\lambda) = q(\lambda) + r(\lambda)$, then $P(A) = q(A) + r(A)$
(iv) if $p(\lambda) = q(\lambda)\, r(\lambda)$, then $P(A) = q(A)\, r(A) = r(A)\, q(A)$
(v) if $p(\lambda) = r(q(\lambda))$ and $r(\lambda)$ is analytic at $q(\lambda_i)$, and there exists $\lambda_j \in \sigma(A)$, then $P(A) = r(q(A))$.

When A has k_j Jordan chains of length t_{jkj} corresponding to $\lambda_j \in \sigma(A)$, $j = 1, \ldots, l$ and $\sum_{h=1}^{kj} t_{jh} = m_j$ where m_j is the multiplicity of λ_j, then A can be represented by

$$A = MJM^{-1} \tag{19a}$$

where

$$J = J_1 \oplus J_2 \oplus \ldots \oplus J_l \tag{19b}$$

and

$$J_j = J_{j1} \oplus J_{j2} \oplus \ldots \oplus J_{jk_j} \qquad j = 1, \ldots, l \tag{19c}$$

$$J_{jh} = \begin{bmatrix} \lambda_j & 1 & 0 & . & 0 \\ 0 & \lambda_j & 1 & . & 0 \\ 0 & 0 & . & . & . \\ . & . & . & . & 1 \\ 0 & 0 & 0 & . & \lambda_j \end{bmatrix} \in \mathbf{C}^{t_{jh} \times t_{jh}} \tag{19d}$$

$$h = 1, \ldots, k_j \text{ and } j = 1, \ldots, l$$

Using lemma 1, a matrix function $g(A)$ generated by $g(\lambda)$ becomes

$$g(A) = M\frac{1}{2\pi i}\oint_c g(\lambda)(\lambda I_n - J)^{-1} d\lambda M^{-1} = Mg(J)M^{-1}$$

$$= M[g(J_{j1}) \oplus g(J_{j2}) \oplus \ldots \oplus g(J_{jk_j})]M^{-1},$$

$$j = 1, \ldots, l \tag{20a}$$

where

$$g(J_{jh}) = \sum_{t=0}^{t_{jk_j}-1} \frac{g^{(t)}(\lambda_j)}{t!} H_{t_{jh}}^t, \qquad h = 1, \ldots, k_j,$$

$$j = 1, \ldots, l \tag{20b}$$

$g^{(t)}(\lambda)$ is the tth derivative of $g(\lambda)$ for $t \geqslant 1$, and $g^{(0)}(\lambda) \triangleq g(\lambda)$. $H_{t_{jh}}$, a shift operator, is the $t_{jh} \times t_{jh}$ matrix with null diagonal entries and all 1s on the super diagonal entries. Since $g(J)$ is an upper triangular matrix, we have the following result [6].

Lemma 2
Given $A \in \mathbf{C}^{n \times n}$, $\sigma(A) = \{\lambda_j, j = 1, \ldots, l \leqslant n\}$ and $g(\lambda)$ is analytic at each $\lambda_j \in \sigma(A)$, and $g(\lambda_j)$ is finite, then $\sigma(g(A)) = \{g(\lambda_j), j = 1, \ldots, l \leqslant n\}$.

Since the poles and zeros of $f(z)$ in eqn. 14 alternate on the negative real axis, the poles and zeros of $g(\lambda)$ in eqn. 17c or in eqn. 17d alternate on the imaginary axis. As a result, if $\sigma(A) \in \bar{\mathbf{C}}^{\mathrm{I}}$, or no $\lambda_j \in \sigma(A)$ exist on the imaginary axis, then $g(\lambda_j)$, $j = 1, \ldots, l$ are finite. Thus, from lemma 2 and the result shown in eqn. 16, we have

$$g(\lambda_k) = \begin{cases} \lambda_k & \text{if } \mathrm{Re}(\lambda_k) > 0 \\ -\lambda_k & \text{if } \mathrm{Re}(\lambda_k) < 0 \end{cases} \tag{21a, 21b}$$

Since

$$g'(\lambda_k) = \begin{cases} 1 & \text{if } \mathrm{Re}(\lambda_k) > 0 \\ -1 & \text{if } \mathrm{Re}(\lambda_k) < 0 \end{cases} \tag{21c, 21d}$$

we have

$$g(J_{kh}) = \begin{cases} J_{kh} & \text{if } \mathrm{Re}(\lambda_k) > 0 \\ -J_{kh} & \text{if } \mathrm{Re}(\lambda_k) < 0 \end{cases} \tag{21e, 21f}$$

Using the result of eqns. 21e and 21f, eqn. 20a becomes

$$g(A) = M[J_1 \oplus J_2 \oplus \ldots \oplus (-J_{m+1}) \oplus \ldots \oplus (-J_l)]M^{-1} \tag{21g}$$

where $\mathrm{Re}(\lambda_k) > 0$ for $1 \leqslant k \leqslant m$ and $\mathrm{Re}(\lambda_k) < 0$ for $m < k \leqslant l$. Finally, the desired matrix sign functions can be obtained from eqns. 21g and 17 as follows:

$$\mathrm{sign}^{(1)}(A) \triangleq \frac{1}{2\pi i}\oint_c \lambda g(\lambda)^{-1}(\lambda I_n - A)^{-1} d\lambda$$

$$= A[g(A)]^{-1} = M[I_{m_1} \oplus I_{m_2} \oplus \ldots \oplus (-I_{m_{m+1}}) \oplus \ldots \oplus (-I_{m_l})]M^{-1} \tag{22a}$$

or

$$\mathrm{sign}^{(2)}(A) \triangleq \frac{1}{2\pi i}\oint_c \lambda^{-1} g(\lambda)(\lambda I_n - A)^{-1} d\lambda$$

$$= A^{-1}[g(A)] \tag{22b}$$

and

$$\mathrm{sign}(A) = \mathrm{sign}^{(1)}(A) = \mathrm{sign}^{(2)}(A) \tag{22c}$$

Thus using lemma 1 and the results in eqn. 17, we have the following theorem.

Theorem 1
Given a matrix $A \in \mathbf{C}^{n \times n}$ and $\sigma(A) \in \bar{\mathbf{C}}^{\mathrm{II}}$, the matrix sign function of A can be described by a matrix continued fraction [7] as

$$\mathrm{sign}(A) = A[I_n + (A^2 - I_n)[2I_n + (A^2 - I_n) \times [2I_n + (A^2 - I_n)[\ldots]^{-1}]^{-1}]^{-1}]^{-1} \tag{23a}$$

$$= A^{-1}[I_n + (A^2 - I_n)[2I_n + (A^2 - I_n) \times [2I_n + (A^2 - I_n)[\ldots]^{-1}]^{-1}]^{-1}] \tag{23b}$$

Corollary 1
The matrix sign function of A as defined in Theorem 1 can be approximated by

$$\mathrm{sign}_{(k)}^{(1)}(A) = [(I_n + A)^k - (I_n - A)^k][(I_n + A)^k + (I_n - A)^k]^{-1}$$

$$= \left[\sum_{j=0}^{\left\lfloor\frac{k-1}{2}\right\rfloor} {}_kC_{2j+1}A^{2j+1}\right]\left[\sum_{j=0}^{\left\lfloor\frac{k}{2}\right\rfloor} {}_kC_{2j}A^{2j}\right]^{-1}$$

$$k = 1, 2, \ldots \tag{24a}$$

or

$$\text{sign}_{(k)}^{\{2\}}(A) = [(I_n + A)^k + (I_n - A)^k]\,[(I_n + A)^k - (I_n - A)^k]^{-1}$$

$$= \left[\sum_{j=0}^{\left\lfloor\frac{k}{2}\right\rfloor} {}_kC_{2j}A^{2j}\right]\left[\sum_{j=0}^{\left\lfloor\frac{k-1}{2}\right\rfloor} {}_kC_{2j+1}A^{2j+1}\right]^{-1}$$

$$k = 1, 2, \ldots \tag{24b}$$

For convenience of application we list some matrix sign functions, $\text{sign}_{(k)}^{\{2\}}(A)$ for $k = 1, \ldots, 7$, as follows:

$$\text{sign}_{(1)}^{\{2\}}(A) = A^{-1} \tag{25a}$$

$$\text{sign}_{(2)}^{\{2\}}(A) = (A^2 + I_n)(2A)^{-1} \tag{25b}$$

$$\text{sign}_{(3)}^{\{2\}}(A) = (3A^2 + I_n)(A^3 + 3A)^{-1} \tag{25c}$$

$$\text{sign}_{(4)}^{\{2\}}(A) = (A^4 + 6A^2 + I_n)(4A^3 + 4A)^{-1} \tag{25d}$$

$$\text{sign}_{(5)}^{\{2\}}(A) = (5A^4 + 10A^2 + I_n) \times (A^5 + 10A^3 + 5A)^{-1} \tag{25e}$$

$$\text{sign}_{(6)}^{\{2\}}(A) = (A^6 + 15A^4 + 15A^2 + I_n) \times (6A^5 + 20A^3 + 6A)^{-1} \tag{25f}$$

$$\text{sign}_{(7)}^{\{2\}}(A) = (7A^6 + 35A^4 + 21A^2 + I_n) \times (A^7 + 21A^5 + 35A^3 + 7A)^{-1} \tag{25g}$$

Observe that $\text{sign}_{(2)}^{\{2\}}(A)$ in eqn. 25*b* is the standard matrix sign algorithm in eqn. 4 for $k = 0$ and $\alpha_k = \beta_k = \frac{1}{2}$.

For a system which contains $\text{Re}(\sigma(A)) = 0$, the matrix A is modified as shown in eqn. 5. Then the proposed matrix sign algorithms in eqn. 24 can be applied to determine the extended matrix sign function [3].

4 Computational considerations

For online computation, the matrix sign function in eqn 24 or eqn. 25 can be used to approximate the matrix sign function in eqn. 23. However, if the matrix A contains both large and small eigenvalues in modules, for a large k of $\text{sign}_{(k)}^{\{1\}}(A)$ or $\text{sign}_{(k)}^{\{2\}}(A)$ in eqn. 24, numerical difficulty might occur. To resolve this difficulty, the following recursive algorithms are derived.

From eqn. 17*a* we have

$$\text{sign}_{(k)}^{\{1\}}(\lambda) = \frac{1 - \left(\frac{1-\lambda}{1+\lambda}\right)^k}{1 + \left(\frac{1-\lambda}{1+\lambda}\right)^k} \tag{26a}$$

Rearranging eqn. 26*a* gives

$$\left(\frac{1-\lambda}{1+\lambda}\right)^k = \frac{1 - \text{sign}_{(k)}^{\{1\}}(\lambda)}{1 + \text{sign}_{(k)}^{\{1\}}(\lambda)} \tag{26b}$$

Let $k = k_1 k_2$, we have

$$\text{sign}_{(k_1 k_2)}^{\{1\}}(\lambda) = \frac{1 - \left(\frac{1-\lambda}{1+\lambda}\right)^{k_1 k_2}}{1 + \left(\frac{1-\lambda}{1+\lambda}\right)^{k_1 k_2}}$$

$$= \frac{1 - \left[\frac{1 - \text{sign}_{(k_1)}^{\{1\}}(\lambda)}{1 + \text{sign}_{(k_1)}^{\{1\}}(\lambda)}\right]^{k_2}}{1 + \left[\frac{1 - \text{sign}_{(k_1)}^{\{1\}}(\lambda)}{1 + \text{sign}_{(k_1)}^{\{1\}}(\lambda)}\right]^{k_2}}$$

$$= \text{sign}_{(k_2)}^{\{1\}}[\text{sign}_{(k_1)}^{\{1\}}(\lambda)] \tag{27a}$$

$$= \text{sign}_{(k_1)}^{\{1\}}[\text{sign}_{(k_2)}^{\{1\}}(\lambda)] \tag{27b}$$

Letting $k = k_1 k_2 \ldots k_m$ and using eqn. 27*a* repeatedly yields

$$\text{sign}_{(k)}^{\{1\}}(\lambda) = \text{sign}_{(k_1)}^{\{1\}}[\text{sign}_{(k_2)}^{\{1\}}[\ldots[\text{sign}_{(k_m)}^{\{1\}}(\lambda)]]] \tag{27c}$$

Similar recursive algorithm can be derived for $\text{sign}_{(k)}^{\{2\}}(\lambda)$ in eqn. 17*b*.

Theorem 2

Recursive algorithms for computing the matrix sign functions of A for $\sigma(A) \in \bar{\mathbb{C}}^{\text{II}}$ are

$$\text{sign}_{(n_{k+1})}^{\{1\}}(A) = \text{sign}_{(f_k)}^{\{1\}}[\text{sign}_{(n_k)}^{\{1\}}(A)]; \quad \text{sign}_{(1)}^{\{1\}}(A) = A \tag{28a}$$

or

$$\text{sign}_{(n_{k+1})}^{\{2\}}(A) = \text{sign}_{(f_k)}^{\{2\}}[\text{sign}_{(n_k)}^{\{2\}}(A)]; \quad \text{sign}_{(1)}^{\{2\}}(A) = A^{-1} \tag{28b}$$

where $n_{k+1} = f_k n_k$ for $k = 1, 2, \ldots, f_k > 1$ and $n_1 \geq 1$

Remark 1

Using the following property of a matrix sign function $\text{sign}(A) = \text{sign}(A^{-1})$, we can set the initial condition of eqn. 28*b* to be

$$\text{sign}_{(1)}^{\{2\}}(A) = A \tag{28c}$$

Remark 2

The standard matrix sign algorithm in eqn. 4 is in fact a special case of the matrix sign algorithm derived herein by choosing $f_k = 2$ in eqns. 28*a* or *b*. For example, from eqn. 25*b* we have

$$\text{sign}_{(2)}^{\{2\}}(A) = (A^2 + I_n)(2A)^{-1} = \tfrac{1}{2}(A + A^{-1}) \tag{28d}$$

and

$$\text{sign}_{(4)}^{\{2\}}(A) = \text{sign}_{(2)}^{\{2\}}[\text{sign}_{(2)}^{\{2\}}(A)] = (A^4 + 6A^2 + I_n)(4A^3 + 4A)^{-1} \tag{28e}$$

The result in eqn. 28*e* is identical to that of the recursive algorithm in eqn. 4 using $k = 1$ and $\alpha_k = \beta_k = 1/2$. Other new recursive algorithms, $\text{sign}_{(k)}^{\{2\}}(A)$ for prime number $k \geq 3$ in eqn. 25 can be considered as basic recursive algorithms for computing the matrix sign function.

Remark 3

From eqn. 26*a* and lemma 2, if $\lambda_l = 1 + \rho e^{j\phi_l}$, $0 < |\rho| \ll 1$, is an eigenvalue of A, then λ_{lk} is the corresponding eigenvalue of $\text{sign}_{(k)}^{\{1\}}(A)$:

$$\lambda_{lk} = \frac{1 - \left(\frac{1-\lambda_l}{1+\lambda_l}\right)^k}{1 + \left(\frac{1-\lambda_l}{1+\lambda_l}\right)^k}$$

$$= 1 - \frac{2}{1 + \left(-1 - \frac{2}{\rho}e^{-j\phi_l}\right)^k}$$

$$\simeq 1 - 2\left(-\frac{\rho}{2}\right)^k e^{jk\phi_l} \tag{29a}$$

Similarly, if $\lambda_i = -1 + \rho e^{j\phi_i}$, $0 < |\rho| << 1$, then we have

$$\lambda_{ik} = \frac{2}{\left(\frac{1-\lambda_i}{1+\lambda_i}\right)^k + 1} - 1 \simeq -1 + 2\left(\frac{\rho}{2}\right)^k e^{jk\phi_i} \qquad (29b)$$

Therefore, the order of convergence in the neighbourhood of ± 1 is k. Furthermore, if $\lambda_i = \pm 1$, then the corresponding eigenvalue of $\text{sign}_{(k)}^{(1)}(A)$ stays at ± 1. The same remarks are true for $\text{sign}_{(k)}^{(2)}(A)$.

Remark 4
If $\text{sign}_{(N)}^{(1)}(A)$ or $\text{sign}_{(N)}^{(2)}(A)$ is a satisfactory approximation of sign (A), and if the recursive algorithm in eqn. 28 with constant order f_k and number of iterations m is used, then the number of iterations needed for large/small eigenvalues is $\log_{f_k} N$. This result can be verified as follows:

$$\text{Let} \quad N \leqslant (f_k)^m, \quad \text{then} \quad m \simeq \log_{f_k} N \qquad (29c)$$

Thus, the result obtained by using the approximate model in eqn. 24 with $k = (f_k)^m \simeq N$ is equivalent to the result using the recursive algorithm in eqn. 28 which has a constant order f_k and the number of iterations (m) shown in eqn. 29c.

5 Generalisation of matrix sign functions

The matrix sign function defined in Section 3 can be viewed as a nonlinear mapping which maps the eigenvalues at the right and left-hand side of the imaginary axis to $+1$ and -1, respectively. Also, the matrix sign function preserves the eigenvectors of the original matrix. The above properties are useful for examining the eigenstructure of a matrix and for solution of control system problems.

The matrix sign function can be generalised to map the eigenvalues of a matrix to $+1$ for those eigenvalues located at one side of a simple closed curve in $\mathbb{C}$ and to -1 for those at the other side of the simple closed curve.

Theorem 3
Let $L \subset \mathbb{C}$ be a simple closed curve which can be mapped onto the imaginary axis by a conformal mapping $h(\lambda)$. Assume that a matrix $A \in \mathbb{C}^{n \times n}$ with $\sigma(A) = \{\lambda_i, i = 1, \ldots, l\}$ exists such that $\sigma(A) \cap L = \phi$ and $h(\lambda)$ is analytic at λ_i. Define

$$\bar{S}(A) \triangleq \frac{1}{2\pi i} \oint_c \frac{g(h(\lambda))}{h(\lambda)} (\lambda I_n - A)^{-1} \, d\lambda \qquad (30a)$$

or

$$\bar{S}(A) \triangleq \frac{1}{2\pi i} \oint_c \frac{h(\lambda)}{g(h(\lambda))} (\lambda I_n - A)^{-1} \, d\lambda \qquad (30b)$$

where

$$g(\lambda) = 1 + \cfrac{\lambda^2 - 1}{2 + \cfrac{\lambda^2 - 1}{2 + \cfrac{\lambda^2 - 1}{\ddots}}} = \lim_{k \to \infty} \frac{\lambda[(1+\lambda)^k + (1-\lambda)^k]}{(1+\lambda)^k - (1-\lambda)^k} \qquad (30c)$$

Then we have

$$\bar{S}(A) = \text{sign}[h(A)] = M[I_{m_1} \oplus \ldots \oplus I_{m_m} \oplus (-I_{m_{m+1}}) \oplus \ldots \oplus (-I_{m_l})]M^{-1} \qquad (30d)$$

if the Jordan decompostion of A is given by

$$A = M[J_1 \oplus \ldots \oplus J_m \oplus J_{m+1} \oplus \ldots \oplus J_l]M^{-1} \qquad (30e)$$

where J_i is a generalised Jordan block, and

$$\sigma(J_i) = \{\lambda_i : \text{Re}(h(\lambda_i)) > 0 \text{ for } 1 \leqslant i \leqslant m \text{ and } \text{Re}(h(\lambda_i)) < 0 \text{ for } m < i \leqslant l\} \qquad (30f)$$

Proof
Since L is a simple closed curve in $\mathbb{C}$ and $h(\lambda)$ is conformal, which maps L onto $j\omega$-axis, the whole complex plane is separated into two regions $\mathbb{C}_1$ and $\mathbb{C}_2$ by L, such that $\text{Re}(h(\lambda)) > 0$ for $\lambda \in \mathbb{C}_1$ and $\text{Re}(h(\lambda)) < 0$ for $\lambda \in \mathbb{C}_2$. Since $h(\lambda)$ is defined to be analytic at $\lambda_i \in \sigma(A)$, $i = 1, \ldots, l$, from eqn. 18 we have

$$h(A) = \frac{1}{2\pi i} \oint_c h(\lambda)(\lambda I_n - A)^{-1} \, d\lambda \qquad (31a)$$

Assuming that $\sigma(A) \cap L = \phi$ yields, $h(\lambda_i) \in \sigma(h(A))$, $i = 1, \ldots, l$, which are not on the imaginary axis, or $h(\lambda_i) \neq j\omega$, $\omega \in \mathbf{R}$. Thus, we conclude that $h(\lambda_i)$ is in the domain of sign (λ).

$$\text{Define} \quad \bar{S}(\lambda) = \text{sign}(h(\lambda)) \qquad (31b)$$

where

$$\text{sign}(\lambda) = \frac{g(\lambda)}{\lambda} = \frac{\lambda}{g(\lambda)} \qquad (31c)$$

Thus

$$\bar{S}(A) = \frac{1}{2\pi i} \oint_c \bar{S}(\lambda)(\lambda I_n - A)^{-1} \, d\lambda = \frac{1}{2\pi i} \oint_c \frac{g(h(\lambda))}{h(\lambda)} (\lambda I_n - A)^{-1} \, d\lambda = \frac{1}{2\pi i} \oint_c \frac{h(\lambda)}{g(h(\lambda))} (\lambda I_n - A)^{-1} \, d\lambda \qquad (31d)$$

If A can be decomposed into the form of eqn. 30e, from eqn. 20 we have

$$h(A) = M[h(J_1) \oplus \ldots \oplus h(J_m) \oplus h(J_{m+1}) \oplus \ldots \oplus h(J_l)]M^{-1} \qquad (31e)$$

Furthermore, since $\text{Re}(h(\lambda_i)) > 0$, $1 \leqslant i \leqslant m$ and $\text{Re}(h(\lambda_i)) < 0$, $m < i \leqslant l$, from eqn. 22, we have

$$\bar{S}(A) = \text{sign}(h(A)) = M[I_{m_1} \oplus \ldots \oplus I_{m_m} \oplus (-I_{m_{m+1}}) \oplus \ldots \oplus (-I_{m_l})]M^{-1} \qquad (31f)$$

In a manner similar to that of theorem 2, we have the following computational algorithm for the generalised matrix sign function.

Corollary 2
Let $h(\lambda)$ and A be defined as in theorem 3. Then we have

$$\bar{S}(A) = \text{sign}(h(A)) = \lim_{k \to \infty} \bar{S}_{n_k}(A) \qquad (32a)$$

where

$$\bar{S}_{n_{k+1}}(A) = \bar{S}_{f_k}[\bar{S}_{n_k}(A)]; \quad \bar{S}_{n_1}(A) = h(A) \tag{32b}$$

$$n_{k+1} = f_k n_k \quad \text{for} \quad k = 1, 2, \ldots, f_k > 1, n_1 \geq 1$$

and

$$\bar{S}_{n_k}(A) = [(I_n + h(A))^{n_k} - (I_n - h(A))^{n_k}] \times [(I_n + h(A))^{n_k} + (I_n - h(A))^{n_k}]^{-1} \tag{32c}$$

$$\text{for} \quad k = 1, 2, \ldots$$

With theorem 3, we can develop appropriate matrix sign functions for several applications. For example, in discrete-time control system problems, we usually need to separate the eigenvalues of a matrix A by the unit circle, $|\lambda| = 1$, which can be mapped onto the imaginary axis by a class of conformal mappings [8], $h(\lambda) = (a\lambda + b)/(c\lambda + d)$ with $a = 1, b = -1, c = 1$ and $d = 1$, or

$$h(\lambda) = \frac{\lambda - 1}{\lambda + 1} \tag{33a}$$

Replacing λ in eqn. 17*a* by $h(\lambda)$ in eqn. 33*a* gives

$$\bar{S}^{(1)}_{(k)}(\lambda) = \text{sign}^{(1)}_{(k)}(h(\lambda)) = \frac{\lambda^k - 1}{\lambda^k + 1} \tag{33b}$$

Also, similarly, eqn. 17*b* yields

$$\bar{S}^{(2)}_{(k)}(\lambda) = \frac{\lambda^k + 1}{\lambda^k - 1} \tag{33c}$$

Thus, the corresponding generalised matrix sign functions for the mapping shown in eqn. 33*a* become

$$\bar{S}^{(1)}_{(k)}(A) = (A^k - I_n)(A^k + I_n)^{-1} \quad k = 1, 2, \ldots \tag{34a}$$

or

$$\bar{S}^{(2)}_{(k)}(A) = (A^k + I_n)(A^k - I_n)^{-1} \quad k = 1, 2, \ldots \tag{34b}$$

and

$$\bar{S}(A) = \lim_{k \to \infty} \bar{S}^{(1)}_{(k)}(A) = \lim_{k \to \infty} \bar{S}^{(2)}_{(k)}(A) \tag{34c}$$

The algorithms derived in eqn. 34 for computing generalised matrix sign functions can be directly applied to solve the algebriac discrete Riccati equation (see Appendix), discrete regulator problem [9], and discrete-time stability problem [10].

6 Conclusions

This paper has proposed an alternate representation of the matrix sign function based on an irrational function described by a continued fraction. It has been shown that an irrational function $\sqrt{z}$ of a complex variable z can be fully represented by a continued fraction if z is not a negative real value. Also, the poles and zeros of the truncated continued fraction alternate on the negative real axis. The principal square-root function, $f(z) = \sqrt{z}$, is then extended to generate a matrix sign function. It has been shown that, when the matrix of interest has no eigenvalues on the imaginary axis, the matrix sign function can be fully described by a matrix continued fraction.

Based on the structure properties of continued fractions, new recursive algorithms for computing the matrix sign function have been derived. It has been shown that the most commonly used matrix sign algorithm is a special case of the recursive algorithms derived in this paper. Finally, the matrix sign function is extended to a generalised matrix sign function such that the newly developed matrix sign algorithms can be directly applied for solution of discrete-time system problems.

For continuous-time systems, the proposed matrix sign functions can be applied to solve stability problems [3], Riccati-type and spectral factorisation problems [2].

7 Acknowledgment

This work was supported in part by the US Army Research Office, under research grant DAAG-29-80-K-0077, and by the US Army Missile R & D Command, under contract DAAH01-82-C-A139.

The authors wish to express their gratitude for the valuable remarks and suggestions from Dr. Jagdish Chandra, Director of Mathematics Division, US Army Research Office, and from the referees.

8 References

1 ROBERTS, J.D.: 'Linear model reduction and solution of the algebraic Riccati equation by use of the sign function', CUED/B-CONTROL/TR 13 Report, Cambridge University, 1971, also published in *Int. J. Control*, 1980, **32**, pp. 677–687
2 DENMAN, E.D., and BEAVERS, A.N.: 'The matrix sign function and computations in systems', *Appl. Math. & Comput.*, 1976, **2**, pp. 63–94
3 MATTHEYS, R.L.: 'Stability analysis via the extended matrix sign function', *Proc. IEE*, 1978, **125**, (3), pp. 241–243
4 BALZER, L.A.: 'Accelerated convergence of the matrix sign function method of solving Lyapunov, Riccati and other matrix equations', *Int. J. Control*, 1980, **32**, pp. 1057–1078
5 KHOVANSKII, A.N.: 'The application of continued fractions and their generalizations to problems in approximation theory' (P. Noordhoff, Netherlands, 1963)
6 RINEHART, R.F.: 'The equivalent of definition of a matrix function', *Am. Math. Monthly*, 1955, **62**, pp. 395–414
7 SHIEH, L.S., and GAUDIANO, F.F.: 'Some properties and applications of matrix-continued fraction', *IEEE Trans.*, 1975, **CAS-22**, pp. 721–728
8 KUO, B.C.: 'Digital control system' (Holt, Rinehart and Winston, New York, 1980)
9 HALBERSBERG, A., and BAR-NESS, Y.: 'Solution of the discrete regulator problem using the sign matrix function', *Electron. Lett.*, 1978, **14**, (9), pp. 286–287
10 ATTARZADEH, F.: 'Relative stability test for continuous and sampled-data control systems using the generalized sign matrix', *IEE Proc. D, Control Theory & Appl.*, 1982, **129**, (5), pp. 189–192

9 Appendix

To show the procedure of using the matrix sign functions for solving discrete-time Riccati equation and discrete regulator problems, we consider the following discrete-time system:

$$X(k+1) = AX(k) + Bu(k) \tag{35a}$$

$$y(k) = CX(k) \tag{35b}$$

where $A \in \mathbf{C}^{n \times n}$, $B \in \mathbf{C}^{n \times m}$, $C \in \mathbf{C}^{p \times n}$ and A is nonsingular. The infinite-time performance index to be minimised is given by

$$J = \sum_{i=0}^{\infty} [X^T(i)QX(i) + u^T(i)Ru(i)] \tag{36}$$

where $Q = \hat{Q}\hat{Q}^T$ is a symetric nonnegative matrix and R is a symmetric positive definite matrix. Assume that the pair $\{A, \hat{Q}\}$ is detectable and $\{A, B\}$ is stablisable, then the steady-state feedback optimal control law [8] becomes

$$u(k) = -(R + B^TPB)^{-1} B^TPAX(k) \tag{37a}$$

where the nonnegative definite matrix P is the solution of the following algebraic nonlinear discrete-time Riccati equation:

$$P = Q + A^T PA - A^T PB\,(R + B^T PB)^{-1} B^T PA \tag{37b}$$

The procedures to determine the optimal control law using the matrix sign functions are described as follows:

Define a $2n \times 2n$ Hamiltonian matrix [8]:

$$G = \begin{bmatrix} A^{-1} & A^{-1}BR^{-1}B^T \\ QA^{-1} & A^T + QA^{-1}BR^{-1}B^T \end{bmatrix} \tag{38a}$$

The modal matrix of G and its inversion are defined as

$$M \triangleq \begin{bmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{bmatrix} \quad M^{-1} \triangleq \begin{bmatrix} \bar{M}_{11} & \bar{M}_{12} \\ \bar{M}_{21} & \bar{M}_{22} \end{bmatrix}$$

$$M_{ij} \in \mathbb{C}^{n \times n};\ i,j = 1,2$$

$$\bar{M}_{ij} \in \mathbb{C}^{n \times n};\ i,j = 1,2 \tag{38b}$$

Thus,

$$M^{-1}GM = \text{block diag}[\Lambda, \Lambda^{-T}] \quad \Lambda \in \mathbb{C}^{n \times n} \tag{38c}$$

where $\Lambda(\Lambda^{-T})$ is the Jordan block corresponding to the eigenvalues of G outside (inside) the unit circle. The Riccati matrix gain P in eqn. 37b can be determined [8] by

$$P = M_{21}M_{11}^{-1} \tag{39}$$

P in eqn. 39 can be indirectly computed via the matrix sign functions as follows:

$$\begin{aligned} \text{sign}(G) &= M \cdot \text{block diag}[\text{sign}(\Lambda), \text{sign}(\Lambda^{-T})] \cdot M^{-1} \\ &= M \cdot \text{block diag}[I_n, -I_n] \cdot M^{-1} = M\bar{I}M^{-1} \end{aligned} \tag{40}$$

where $\text{sign}(\Lambda) = I_n$; $\text{sign}(\Lambda^{-T}) = -I_n$; $\bar{I} \triangleq \text{block diag}[I_n, -I_n]$ and I_n is an $n \times n$ identity matrix.

Define a new matrix W, or

$$\begin{aligned} W &\triangleq \text{sign}(G) + \bar{I} = M[\bar{I}M^{-1} + M^{-1}\bar{I}] \\ &= \text{M} \cdot \text{block diag}\,[2\bar{M}_{11}, -2\bar{M}_{22}] \\ &= 2 \begin{bmatrix} M_{11}\bar{M}_{11} & -M_{12}\bar{M}_{22} \\ M_{21}\bar{M}_{11} & -M_{22}\bar{M}_{22} \end{bmatrix} \triangleq \begin{bmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{bmatrix} \end{aligned}$$

$$W_{ij} \in \mathbb{C}^{n \times n};\ i,j = 1,2 \tag{41}$$

Thus P in eqn. 39 can be indirectly determined using the partitioning matrices W_{21} and W_{11} as

$$\begin{aligned} P &= W_{21}W_{11}^{-1} = (2M_{21}\bar{M}_{11})\,(2M_{11}\bar{M}_{11})^{-1} \\ &= M_{21}M_{11}^{-1} \end{aligned} \tag{42}$$

As a result, the optimal control law in eqn. 37a can be obtained.

In practice, an approximate sign(G) is often used to determine the P in eqn. 42. The matrix sign algorithms for computing the approximate sign (G) (defined as $\hat{G}_j$) are

$$\text{sign}(G) = \lim_{j \to \infty} (G^j - I_{2n})\,(G^j + I_{2n})^{-1} \tag{43a}$$

$$= \lim_{j \to \infty} (G^j + I_{2n})\,(G^j - I_{2n})^{-1} \tag{43b}$$

$$\simeq \hat{G}_j \text{ for a finite } j \tag{43c}$$

The index j of $\hat{G}_j$ in eqn. 43c can be determined when

$$|\text{trace}[(\hat{G}_j)^2] - 2n|/2n < \epsilon_j \tag{43d}$$

where ϵ_j is a desired error tolerance.

Note that, when j is a large value, the roundoff errors due to direct computations of G^j in eqn. 43 may occur. To reduce the errors, the recursive algorithm in eqn. 28a with $A = (G - I_{2n})\,(G + I_{2n})^{-1}$ can be applied to determine $\hat{G}_j$ in eqn. 43c where $\hat{G}_j = \text{sign}_j^{(1)}(A)$. Using the $\hat{G}_j$, the approximate W (defined as $\hat{W}_j$) yields

$$\hat{W}_j \triangleq \hat{G}_j + \bar{I} = \begin{bmatrix} (\hat{W}_{11})_j & (\hat{W}_{12})_j \\ (\hat{W}_{21})_j & (\hat{W}_{22})_j \end{bmatrix}$$

$$(\hat{W}_{ik})_j \in \mathbb{C}^{n \times n} \qquad i,k = 1,2 \tag{44}$$

Thus the approximate Riccati gain matrix P (defined as $\hat{P}_j$) becomes

$$\hat{P}_j = (\hat{W}_{21})_j\,(\hat{W}_{11})_j^{-1} \tag{45a}$$

and the approximate optimal control law and gain F (defined as $\hat{F}_j$) become

$$u(k) = -\hat{F}_j X(k) \tag{45b}$$

where

$$\hat{F}_j = (R + B^T\hat{P}_jB)^{-1}B^T\hat{P}_jA \tag{45c}$$

A 21