### Title
Some Properties of Matrix Sign Functions Derived from Continued Fractions

### Author(s)
L. S. Shieh  
Y. T. Tsay  
R. L. Yates

### Performing Organization Name and Address
University of Houston  
Houston, TX  77004

### Report Date
May 83

### Number of Pages
8

### Distribution Statement (of this report)
Submitted for announcement only.

### Key Words (Continue on reverse side if necessary and identify by block number)

### Abstract (Continue on reverse side if necessary and identify by block number)

---

**Note:** The image contains a form with various fields filled out, but the content is not fully legible. The fields include report number, title, author(s), performing organization, report date, number of pages, distribution statement, and abstract. The form is marked as unclassified with some fields containing specific information such as contract grant number and program element, project, task, and work unit numbers.
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 algorithms for engineering problems.

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

$$ S = \text{sign}(A) = \frac{1}{\pi} \int_{C} \frac{\lambda I_n - A}{\lambda - \sigma(A)} d\lambda $$

where $\lambda$ is a simple closed contour in right half-plane of $\lambda$ and encloses all the right-half-plane eigenvalues of $A$.

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

$$ J = \text{block diag} \{ J_1, J_2 \} \preceq J, \preceq J, $$

then

$$ A = MJM^{-1} $$

and

$$ S = \text{sign}(A) = M [\text{sign}(J_1) \ast \text{sign}(J_2)] M^{-1} $$

where $J_1 \in \mathbb{C}^{n_1 \times n_1}, J_2 \in \mathbb{C}^{n_2 \times n_2}$ and $n = n_1 + n_2$ are the collection of Jordan blocks with $\text{Re}(\sigma(A)) > 0$ and $\text{Re}(\sigma(A)) < 0$, respectively. $M \in \mathbb{C}^{n \times n}$ is a modal matrix of $A$.

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

$$ S \triangleq \text{sign}(A) = M [\text{sign}(J_1) \ast \text{sign}(J_2) \ast \text{sign}(J_0)] M^{-1} $$

where $J_0 \in \mathbb{C}^{n_3 \times n_3}$ is the collection of Jordan blocks with $\text{Re}(\sigma(A)) = 0$. $O_{n_3}$ is an $n_3 \times n_3$ null matrix, and $n_1 + n_2 + n_3 = n$.

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

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

$$ \alpha_k + \beta_k = 1 \quad \text{and} \quad \lim \alpha_k = \lim \beta_k = \frac{1}{2}. $$

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

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

$$ \hat{S}_{k+1} = \hat{S}_k + \hat{S}_k^{-1} $$

where $\hat{S}_k$ 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. $\epsilon$ is given by

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

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

$$ \text{sign}(\lambda) = \begin{cases} +1 & \text{if } \text{Re}(\lambda) > 0 \\ -1 & \text{if } \text{Re}(\lambda) < 0 \\ \frac{\lambda}{g(\lambda)} & \text{if } \text{Re}(\lambda) > 0 \\ \frac{\lambda}{g(\lambda)} & \text{if } \text{Re}(\lambda) < 0 \end{cases} $$

where

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

and

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

and
The recursive forms for continued fraction expansion in Eqn. 6e, where \( f(z) \) is defined by and \( g(z) \) is obviously an absolute value function and
\[
g(\phi) = \sqrt{\phi^2} \tag{7}
\]
In a similar fashion, when \( \lambda \) is a complex function, \( g(\lambda) \) may be defined by
\[
g(\lambda) = \sqrt{\lambda^2} \tag{8}
\]
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 \( f(z) = \sqrt{z} \) 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]
\[
f(z) = \sqrt{z} = 1 + \frac{z-1}{2 + \frac{z-1}{2 + \frac{z-1}{2 + \cdots}}} \tag{9}
\]
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 \( k \)th truncation of \( f(z) \) as \( f_k(z) \), or
\[
f_k(z) = 1 + 2f_{k-1}(z) \tag{10a}
\]
where \( f_0(z) = 0 \) and
\[
f_k(z) = \frac{1}{1 + \frac{1}{1 + \frac{1}{1 + \cdots}}} = \frac{a_k}{b_k}, \quad k \geq 1 \tag{10b}
\]
The recursive forms for \( a_k \) and \( b_k \) are
\[
a_k = a_{k-1} + \frac{1}{2} (z-1)a_{k-2}, \quad a_{-1} = 1, \quad a_0 = 0 \tag{11a}
\]
\[
b_k = b_{k-1} + \frac{1}{2} (z-1)b_{k-2}, \quad b_{-1} = 0, \quad b_0 = 1 \tag{11b}
\]
The difference equation for both \( a_k \) and \( b_k \) satisfies
\[
1 - q^{-1} - \frac{1}{2} (z-1)q^{-2} = 0 \tag{11c}
\]
where \( q^{-1} \) is a backward shift operator, or \( q^{-1} a_k = a_{k-1} \).

The zeros of the equation in Eqn. 11c become
\[
q_1 = \frac{1 + \sqrt{2}}{2} \quad \text{and} \quad q_2 = \frac{1 - \sqrt{2}}{2} \tag{11d}
\]
From Eqn. 11d we observe that
\[
|q_1| > |q_2| \quad \text{if} \quad -\pi < \arg(z) < \pi \tag{11e}
\]
The general form for Eqn. 10a (see Reference 5) is as follows:
\[
f_k(z) = \sqrt{\left(1 + \sqrt{2}\right)^k + \frac{1}{2}} - \frac{1}{\left(1 + \sqrt{2}\right)^k} = \sum_{j=0}^{p} \frac{C_{2j}z^j}{\sum_{j=0}^{r} C_{2j+1}z^j} \tag{12a}
\]
where \( p = \lfloor k/2 \rfloor \) and \( r = \lfloor (k-1)/2 \rfloor \) are integers for \( k \geq 1 \) and \( C_{2j} \) are the coefficients of a binomial expansion.

Substituting Eqn. 11d into Eqn. 12a yields
\[
f_k(z) = \frac{\sqrt{2}}{2} \left[ \left(1 + \sqrt{2}\right)^k + \frac{1}{2} \right] = \left(1 + \sqrt{2}\right)^k - \left(1 - \sqrt{2}\right)^k
\]
\[
= \left(1 + \left(\frac{q_2}{q_1}\right)^k\right) \left(1 - \left(\frac{q_2}{q_1}\right)^k\right) = \left(1 - \left(\frac{q_2}{q_1}\right)^k\right)
\]
\[
k \geq 1 \quad \text{and} \quad q_1 \neq q_2 \tag{12b}
\]

\( f_k(z) \) for \( k = 1, \ldots, 4 \) are as follows:
\[
f_1(z) = 1 \tag{13a}
\]
\[
f_2(z) = \frac{z + 1}{2} \tag{13b}
\]
\[
f_3(z) = \frac{z^2 + 1}{z + 3} \tag{13c}
\]
\[
f_4(z) = \frac{z^2 + 6z + 1}{4z + 4} \tag{13d}
\]
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
\[
f(z) = \lim_{k \to \infty} f_k(z) = \sqrt{z} \tag{14a}
\]
The important result in Eqn. 14a can be verified by a ratio test of the series, which can be obtained by expanding \( \left[1 - \left(\frac{q_2}{q_1}\right)^k\right]^{-1} \) in Eqn. 12b. Also, the convergent condition \( |q_1| > |q_2| \) implies that the domain of \( z \) is in \( \mathbb{C}^* \) where \( C^* = \mathbb{C} - \mathbb{R}^+ \) 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 \( \sqrt{\bar{z}} \) can fully be represented by the continued fraction if \( z \in \mathbb{C}^+ \).

**Property 2**

If \( |z_1| = |z_2| \) and \( z = 0 \), then

\[
f(z) = \lim_{k \to \infty} f_k(z) = 0
\]

**Property 3**

If \( |z_1| = |z_2| \), \( \Re \neq 0 \), and \( \arg(z) = \pi \), then

\[
\lim_{k \to \infty} f_k(z) = \infty
\]

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

When \( |z_1| = |z_2| \), from eqn. 11d 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. 10a becomes \( 1/(k + 1) \) and

\[
f(z) = \lim_{k \to \infty} f_k(z) = \lim_{k \to \infty} \frac{1}{k + 1} \to 0
\]

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 \( 1 + j\omega = \sqrt{1 + \omega^2} e^{j\phi} \) and \(-\sqrt{z} = \sqrt{1 + \omega^2} e^{-j\phi} \) where \( \phi = \tan^{-1}\omega \). Therefore, we have

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

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

**Property 4**

The poles and zeros of \( f_k(z) \) and \( k \geq 2 \) alternate on the negative real axis.

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

\[
P_m = \left( \tan \frac{m\pi}{k} \right)^2
\]

\[
m = 1, 2, \ldots, (k - 1)/2 \text{ for odd } k
\]

\[
m = 1, 2, \ldots, (k - 2)/2 \text{ for even } k
\]

and the zeros

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

\[
m = 1, 2, \ldots, (k - 1)/2 \text{ for odd } k
\]

\[
m = 1, 2, \ldots, k/2 \text{ for even } k
\]

Since the tangent function \( \tan \theta \) is monotonically increasing for \( 0 < \theta < \pi/2 \), the poles and zeros of \( f_k(z) \) for \( k \geq 2 \) 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

\[
f(z) \triangleq g(\lambda) = \sqrt{\lambda^2}, \quad \lambda \in \mathbb{C}^i
\]

\[
= 1 + \frac{\lambda^2 - 1}{\lambda^2 - 1}
\]

\[
= \frac{1}{2 + \frac{\lambda^2 - 1}{\lambda^2 - 1}}
\]

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

\[
\lambda = |\lambda| e^{j\phi} \quad \text{and } |\phi| \neq \frac{\pi}{2}
\]

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

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

\[
= \begin{cases} 
|\lambda| e^{j\phi} = \lambda & \text{if } -\frac{\pi}{2} < \phi < \frac{\pi}{2} \\
|\lambda| e^{-j(\phi - \pi)} = -\lambda & \text{if } -\frac{\pi}{2} < \phi < \frac{\pi}{2}
\end{cases}
\]

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

\[
\text{sign}(\lambda) \triangleq \frac{\lambda}{g(\lambda)} = \lim_{k \to \infty} \text{sign}^{(k)}(\lambda)
\]

where

\[
g_k(\lambda) = f_k(z) \quad \text{with } z = \lambda^2
\]

The 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 \( \sigma(A) = \{\lambda_1, \ldots, \lambda_n\} \) where \( f \leq n \). If a scalar function \( p(\lambda) \) is analytic at \( \lambda_j \), \( j = 1, \ldots, f \), then the matrix function \( P(A) \) generated by

\[
P(A) = \begin{pmatrix}
1 + \frac{\lambda^2 - 1}{\lambda^2 - 1} \\
2 + \frac{\lambda^2 - 1}{\lambda^2 - 1}
\end{pmatrix}
\]

\[
\lambda \in \mathbb{C}^i
\]
\( p(\lambda) \) can be defined as [6]

\[
P(A) = \frac{1}{2\pi i} \oint_{c} p(\lambda)(\lambda J_n - A)^{-1} \, d\lambda
\]  

(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 \( \lambda_j, j = 1, \ldots, l \) 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 \( p(\lambda) = q(\lambda) \), then \( P(A) = q(A) r(A) = q(A) q(A) \)

(v) if \( p(\lambda) = r(\lambda) \) and \( r(\lambda) \) is analytic at \( \lambda_j \), and there exists \( \lambda_j \in \sigma(A) \), then \( P(A) = r(q(A)) \).

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

\[
A = MJM^{-1}
\]

(19a)

where

\[
J = J_1 \oplus J_2 \oplus \cdots \oplus J_t
\]

(19b)

and

\[
J_j = J_{j_1} \oplus J_{j_2} \oplus \cdots \oplus J_{j_{k_j}} \quad j = 1, \ldots, l
\]

(19c)

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

\[
g(A) = M(1) (\lambda J_n - J)^{-1} M^{-1} = MJM^{-1}
\]

(20a)

where

\[
g(J_{jh}) = \sum_{t=0}^{t_{jh}} \frac{g^{(j)}(\lambda_j)}{t!} H^t_{jh}, \quad h = 1, \ldots, k_j, \quad j = 1, \ldots, l
\]

(20b)

\( g^{(j)}(\lambda) \) is the \( r \)th derivative of \( g(\lambda) \) for \( t > 1 \), and \( g^{(j)}(\lambda) \triangleq g(\lambda) \).

\( H^t_{jh} \), a shift operator, is the \( t_{jh} \times t_{jh} \) matrix with all 1s on the diagonal entries and all 0s on the super diagonal entries. Since \( g(J) \) is an upper triangular matrix, we have the following result [6]:

**Lemma 2**

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

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 \( \sigma(A) \in \mathbb{C} \), 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

\[
g(\lambda_j) = \begin{cases} 
\lambda_j & \text{if } \text{Re}(\lambda_j) > 0 \\
-\lambda_j & \text{if } \text{Re}(\lambda_j) < 0
\end{cases}
\]

(21a)

Since

\[
g'(\lambda_j) = \begin{cases} 
1 & \text{if } \text{Re}(\lambda_j) > 0 \\
-1 & \text{if } \text{Re}(\lambda_j) < 0
\end{cases}
\]

(21b)

we have

\[
g(J_{jh}) = \begin{cases} 
J_{jh} & \text{if } \text{Re}(\lambda_j) > 0 \\
-J_{jh} & \text{if } \text{Re}(\lambda_j) < 0
\end{cases}
\]

(21c)

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

\[
g(A) = M[1 \oplus J_2 \oplus \cdots \oplus (J_{m_1})] \oplus \cdots \oplus (-J_{m_1})] M^{-1}
\]

(21d)

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

\[
\text{sign}^{(1)}(A) = \frac{1}{2\pi i} \oint_{c} \lambda^{i \lambda_j} (\lambda J_n - A)^{-1} \, d\lambda
\]

(21e)

\[
= A [g(\lambda)]^{-1} = M[I_{m_1} \oplus I_{m_2} \cdots \oplus (-I_{m_1})] M^{-1}
\]

(22a)

or

\[
\text{sign}^{(2)}(A) = \frac{1}{2\pi i} \oint_{c} \lambda^{-i \lambda_j} (\lambda J_n - A)^{-1} \, d\lambda
\]

(22b)

and

\[
\text{sign}(A) = \text{sign}^{(1)}(A) = \text{sign}^{(2)}(A)
\]

(22c)

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

**Theorem 1**

Given a matrix \( A \in \mathbb{C}^{n \times n} \) and \( \sigma(A) \in \mathbb{C} \), the matrix sign function of \( A \) can be described by a matrix continued fraction [7] as

\[
\text{sign}(A) = A[I_n + (A^2 - I_n) \{ [2I_n + (A^2 - I_n) \} \cdots (2I_n + (A^2 - I_n)) \}]
\]

(23a)

\[
= A^{-1} [I_n + (A^2 - I_n) \{ [2I_n + (A^2 - I_n) \} \cdots (2I_n + (A^2 - I_n)) \}]
\]

(23b)

The matrix sign function of \( A \) as defined in Theorem 1 can be approximated by

\[
\text{sign}_L^{(2)}(A) = [(I_n + A) \cdots (I_n - A)^k] [(I_n + A)^k + (I_n - A)^k]^{-1}
\]

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

\[
\text{sign}_{[k]}(A) = \text{sign}_{[k_1]} \left[ \text{sign}_{[k_2]} \left[ \ldots \left[ \text{sign}_{[k_m]}(A) \right] \ldots \right] \right]
\]  

(27e)

Similar recursive algorithm can be derived for \( \text{sign}_{[k]}(A) \) in eqn. 17b.

**Theorem 2**

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

\[
\text{sign}_{[k]}(A) = A^{-1} \text{ sign}_{[k]}(A) \cdot A \text{ sign}_{[k]}(A)
\]  

(28a)

or

\[
\text{sign}_{[k]}(A) = A \text{ sign}_{[k]}(A) \cdot A^{-1} \text{ sign}_{[k]}(A)
\]  

(28b)

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

**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. 28a to be

\[
\text{sign}_{[k]}(A) = A
\]  

(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. 28a or b. For example, from eqn. 25b we have

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

(28d)

and

\[
\text{sign}_{[2]}(A) = \text{sign}(A) \left[ \text{sign}(A) \right]
\]  

(28e)

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

**Remark 3**

From eqn. 26a and lemma 2, if \( \lambda_k = 1 + \rho e^{i\phi} \), \( 0 < |\rho| << 1 \) is an eigenvalue of \( A \), then \( \lambda_k \) is the corresponding eigenvalue of \( \text{sign}_{[n]}(A) \):

\[
\lambda_k = \frac{1 - \left( 1 - \lambda_k \right)^{k_k}}{1 + \left( 1 - \lambda_k \right)^{k_k}}
\]  

(29a)

where

\[
1 - \left( 1 - \lambda_k \right)^{k_k} = \frac{1 - \text{sign}_{[k]}(\lambda)}{1 + \text{sign}_{[k]}(\lambda)}
\]  

(29b)

\[
\lambda_k = \frac{1 - \left( 1 - \lambda_k \right)^{k_k}}{1 + \left( 1 - \lambda_k \right)^{k_k}}
\]  

(29c)

\[
1 - \left( 1 - \lambda_k \right)^{k_k} \approx 1 - 2 \left( \frac{\rho}{2} \right)^{k_k} e^{i\phi}
\]  

(29d)
Similarly, if \( \lambda_0 = -1 + pe^{\phi t}, 0 < |p| < 1, \) then we have

\[
\lambda_{th} = \frac{2}{1 - \lambda^*} - 1 \simeq -1 + 2 \left( \frac{\phi}{2} \right) e^{\phi t} \tag{29b}
\]

Therefore, the order of convergence in the neighbourhood of \( \pm 1 \) is \( k \). Furthermore, if \( \lambda_0 = \pm 1, \) then the corresponding eigenvalue of \( \text{sign}(A) \) stays at \( \pm 1 \). The same remarks are true for \( \text{sign}^2(A) \).

**Remark 4**

If \( \text{sign}^2(A) \) or \( \text{sign}^3(A) \) is a satisfactory approximation of \( \text{sign}(A) \), 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 \( \log_h N \). This result can be verified as follows:

Let \( N \leq (f_k)^m \), then \( m \geq \log_h N \) \( \tag{29c} \)

Thus, the result obtained by using the approximate model in eqn. 24 with \( k = (f_k)^m \geq N \) is equivalent to the result using the recursive algorithm in eqn. 28 with 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 \( \sigma(A) \cap L = \emptyset \) and \( h(\lambda) \) is analytic at \( \lambda_i \). Define

\[
\tilde{S}(A) = \frac{1}{2\pi i} \int \frac{g(h(\lambda))}{h(\lambda)} \left( \lambda I_n - A \right)^{-1} d\lambda \tag{30a}
\]

or

\[
\tilde{S}(A) = \int \frac{1}{2\pi i} \frac{h(\lambda)}{g(h(\lambda))} \left( \lambda I_n - A \right)^{-1} d\lambda \tag{30b}
\]

where

\[
g(\lambda) = \frac{\lambda^2 - 1}{2 + \lambda^2 - 1} \Rightarrow \frac{\lambda((1 + \lambda)^k + (1 - \lambda)^k)}{(1 + \lambda)^k - (1 - \lambda)^k} \tag{30c}
\]

Then we have

\[
\tilde{S}(A) = \text{sign}(h(A)) = M \left[ I_{m_1} \otimes \ldots \otimes I_{m_m} \right.
\]

\[
\left. \otimes (-I_{m_{m+1}}) \otimes \ldots \otimes (-I_{m_l}) \right] M^{-1} \tag{30d}
\]

if the Jordan decomposition of \( A \) is given by

\[
A = M \left[ J_1 \otimes \ldots \otimes J_m \otimes J_{m+1} \otimes \ldots \otimes J_l \right] M^{-1} \tag{30e}
\]

where \( J_i \) is a generalised Jordan block, and

\[
\sigma(J_i) = \{ \lambda_i : \text{Re}(h(A)) > 0 \text{ for } 1 < i < m \text{ and } \text{Re}(h(A)) < 0 \text{ for } m < i < l \} \tag{30f}
\]

**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 \( C_1 \) and \( C_2 \) by \( L \), such that \( \text{Re}(h(\lambda)) > 0 \) for \( \lambda \in C_1 \), and \( \text{Re}(h(\lambda)) < 0 \) for \( \lambda \in C_2 \). Since \( h(\lambda) \) is defined to be analytic at \( \lambda_i \in \sigma(A), i = 1, \ldots, l \), from eqn. 18 we have

\[
h(\lambda) = \frac{1}{2\pi i} \int \frac{g(h(\lambda))}{h(\lambda)} (\lambda I_n - A)^{-1} d\lambda \tag{31a}
\]

Assuming that \( \sigma(A) \cap L = \emptyset \) yields, \( h(\lambda_i) \in \sigma(\text{sign}(A)), i = 1, \ldots, l \), which are not on the imaginary axis, or \( h(\lambda_i) \neq j\omega, \omega \in \mathbb{R} \).

Thus, we conclude that \( h(\lambda_i) \) is in the domain of \( \text{sign}(\lambda) \).

Define \( \tilde{S}(A) = \text{sign}(h(\lambda)) \) \( \tag{31b} \)

Thus

\[
\tilde{S}(A) = \frac{1}{2\pi i} \int \frac{h(\lambda)}{g(h(\lambda))} (\lambda I_n - A)^{-1} d\lambda \tag{31c}
\]

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

\[
h(\lambda) = M \left[ h(J_1) \otimes \ldots \otimes h(J_m) \right. \otimes (-I_{m_{m+1}}) \otimes \ldots \otimes (-I_{m_l}) \right] M^{-1} \tag{31d}
\]

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

\[
\tilde{S}(A) = \text{sign}(h(A)) = M \left[ I_{m_1} \otimes \ldots \otimes I_{m_m} \otimes (-I_{m_{m+1}}) \otimes \ldots \otimes (-I_{m_l}) \right] M^{-1} \tag{31f}
\]

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

\[
\tilde{S}(A) = \text{sign}(h(\lambda)) = \lim_{k \to \infty} \tilde{S}_k(A) \tag{32a}
\]

where
\[ S_{n+1}(A) = S_n[S_n(A)]; \quad S_1(A) = h(A) \] (32b)
and
\[ S_n(A) = [(U_n + h(A))^n - U_n - h(A)] \times \]
\[ [(U_n + h(A))^n + U_n - h(A)]^{-1} \] (32c)
for \( k = 1, 2, \ldots \).

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, \(|A| = 1\), which can be mapped onto the imaginary axis by a class of conformal mappings [8].

The infinite-time performance index to be minimised is
\[ J = \int_x^t \{X^T Q X + u^T R u\} dt \]
(37)
where \( Q = Q^T \) is a symmetric nonnegative matrix and \( R \) is a symmetric positive definite matrix. Assume that the pair \([A, Q]\) is detectable and \([A, B]\) is stabilisable, then the steady-state feedback optimal control law [8] becomes
\[ u(k) = - (R + B^T P B)^{-1} B^T P A x(k) \] (37a)
where the nonnegative definite matrix $P$ is the solution of the following algebraic nonlinear discrete-time Riccati equation:

$$P = Q + A^T PA - A^T PB (R + B^T PB)^{-1} B^T PA$$  \hspace{1cm} (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 $[8]$

$$G = \begin{bmatrix} A^{-1} & A^{-1}BR^{-1}B^T \\ QA^{-1} & AT + QA^{-1}BR^{-1}B^T \end{bmatrix}$$  \hspace{1cm} (38a)

The modal matrix of $G$ and its inversion are defined as

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

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

Thus,

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

$$P = M_{21}M_{11}^{-1}$$  \hspace{1cm} (39)

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

$$\text{sign}(G) = M \cdot \text{block diag}[\text{sign}(\Lambda), \text{sign}(\Lambda^{-T})] \cdot M^{-1}$$

$$= M \cdot \text{block diag}[I_n, -I_n] \cdot M^{-1} = M \tilde{M}^{-1}$$  \hspace{1cm} (40)

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

Define a new matrix $W$, or

$$W \triangleq \text{sign}(G) + \tilde{T} = M[\tilde{M}^{-1} + M^{-1}I]$$

$$= M \cdot \text{block diag}[2\tilde{M}_{11}, -2\tilde{M}_{22}]$$

$$= 2 \begin{bmatrix} M_{11} & \tilde{M}_{12} \\ M_{21} & \tilde{M}_{22} \end{bmatrix}$$

$$W_{ij} \in \mathbb{C}^{n \times n}; \quad i, j = 1, 2$$  \hspace{1cm} (41)

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

$$P = W_{21}W_{11}^{-1} = (2M_{21}\tilde{M}_{11})(2M_{11}\tilde{M}_{11})^{-1}$$

$$= M_{21}M_{11}^{-1}$$  \hspace{1cm} (42)

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

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

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

$$\simeq \tilde{G}_j$$ for a finite $j$$  \hspace{1cm} (43a)

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

$$\text{trace}[(\tilde{G}_j)^T] - 2n/2n < \varepsilon_j$$  \hspace{1cm} (43d)

where $\varepsilon_j$ 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 $\tilde{G}_j$ in eqn. 43c where $G_j = \text{sign}^{(1)}(A)$. Using the $\tilde{G}_j$, the approximate $W$ (defined as $\tilde{W}_j$) yields

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

$$(\tilde{W}_{ik})_j \in \mathbb{C}^{n \times n} \quad i, k = 1, 2$$  \hspace{1cm} (44)

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

$$\tilde{P}_j = (\tilde{W}_{21})_j (\tilde{W}_{11})_j^{-1}$$  \hspace{1cm} (45a)

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

$$u(k) = -\tilde{F}_j x(k)$$  \hspace{1cm} (45b)

where

$$\tilde{F}_j = (R + B^T PB)^{-1} B^T P_{21}$$  \hspace{1cm} (45c)