Research Article

On a Cubically Convergent Iterative Method for Matrix Sign

M. Sharifi, S. Karimi Vanani, F. Khaksar Haghani, M. Arab, and S. Shateyi

1Department of Mathematics, Islamic Azad University, Shahrekord Branch, Shahrekord, Iran
2Department of Mathematics and Applied Mathematics, University of Venda, Thohoyandou 0950, South Africa

Correspondence should be addressed to S. Shateyi; stanford.shateyi@univen.ac.za

Received 20 July 2014; Revised 29 August 2014; Accepted 30 August 2014

Academic Editor: Predrag S. Stanimirovic

Copyright © 2015 M. Sharifi et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

We propose an iterative method for finding matrix sign function. It is shown that the scheme has global behavior with cubical rate of convergence. Examples are included to show the applicability and efficiency of the proposed scheme and its reciprocal.

1. Introduction

It is known that the function of sign in the scalar case is defined for any \( z \in \mathbb{C} \) not on the imaginary axis by

\[
\text{sign}(z) = \begin{cases} 
1, & \text{Re}(z) > 0, \\
-1, & \text{Re}(z) < 0.
\end{cases}
\]

(1)

An extension of (1) for the matrix case was given firstly by Roberts in [1]. This extended matrix function is of clear importance in several applications (see, e.g., [2] and the references therein).

Assume that \( A \in \mathbb{C}^{n \times n} \) is a matrix with no eigenvalues on the imaginary axis. To define this matrix function formally, let

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

(2)

be a Jordan canonical form arranged so that \( J = \text{diag}(J_1, J_2) \), where the eigenvalues of \( J_1 \in \mathbb{C}^{p \times p} \) lie in the open left half-plane and those of \( J_2 \in \mathbb{C}^{q \times q} \) lie in the open right half-plane; then

\[
S = \text{sign}(A) = T \begin{pmatrix} -I_p & 0 \\ 0 & I_q \end{pmatrix} T^{-1},
\]

(3)

where \( p + q = n \). A simplified definition of the matrix sign function for Hermitian case (eigenvalues are all real) is

\[
S = U \text{diag}(\text{sign}(\lambda_1), \ldots, \text{sign}(\lambda_n)) U^*,
\]

(4)

where

\[
U^*AU = \text{diag}(\lambda_1, \ldots, \lambda_n)
\]

(5)

is a diagonalization of \( A \).

The importance of computing \( S \) is also due to the fact that the sign function plays a fundamental role in iterative methods for matrix roots and the polar decomposition [3].

Note that although \( \text{sign}(A) \) is a square root of the identity matrix, it is not equal to \( I \) or \( -I \) unless the spectrum of \( A \) lies entirely in the open right half-plane or open left half-plane, respectively. Hence, in general, \( \text{sign}(A) \) is a nonprimary square root of \( I \).

In this paper, we focus on iterative methods for finding \( S \). In fact, such methods are Newton-type schemes which are in essence fixed-point-type methods by producing a convergent sequence of matrices via applying a suitable initial matrix.

The most famous method of this class is the quadratic Newton method defined by

\[
X_{k+1} = \frac{1}{2} \left( X_k + X_k^{-1} \right).
\]

(6)

It should be remarked that iterative methods, such as (6), and the Newton-Schultz iteration

\[
X_{k+1} = \frac{1}{2} X_k \left( 3I - X_k^2 \right)
\]

(7)

or the cubically convergent Halley method

\[
X_{k+1} = \left[ I + 3X_k^2 \right] \left[ X_k \left( 3I + X_k^2 \right) \right]^{-1},
\]

(8)
are all special cases of the Padé family proposed originally in [4]. The Padé approximation belongs to a broader category of rational approximations. Coincidentally, the best uniform approximation of the sign function on a pair of symmetric but disjoint intervals can be expressed as a rational function.

Note that although (7) does not possess a global convergence behavior, on state-of-the-art parallel computer architectures, matrix inversions scale less satisfactorily than matrix multiplications do, and subsequently (7) is useful in some problems. However, due to local convergence behavior, it is excluded from our numerical examples in this work.

The rest of this paper is organized as follows. In Section 2, we discuss how to construct a new iterative method for finding (3). It is also shown that the constructed method is convergent with cubical rate. It is noted that its reciprocal iteration obtained from our main method is also convergent. Numerical examples are furnished to show the higher numerical accuracy for the constructed solvers in Section 3. The paper ends in Section 4 with some concluding comments.

2. A New Method

The connection of matrix iteration methods with the sign function is not immediately obvious, but in fact such methods can be derived by applying a suitable root-finding method to the nonlinear matrix equation

$$X^2 = I$$  

and when of course \(\text{sign}(A)\) is one solution of this equation (see for more [5]).

Here, we consider the following root-solver:

$$x_{k+1} = x_k - \frac{10 - 4L(x_k) f(x_k)}{10 - 9L(x_k) f'(x_k)},$$  

with \(L(x_k) = f''(x_k)f(x_k)/f'(x_k)^2\). In what follows, we observe that (10) possesses third order of convergence.

Theorem 1. Let \(\alpha \in D\) be a simple zero of a sufficiently differentiable function \(f : D \subseteq \mathbb{C} \rightarrow \mathbb{C}\), which contains \(x_0\) as an initial approximation. Then the iterative expression (10) satisfies

$$e_{k+1} = \left(\frac{e_2}{5} - c_3\right) e_3 + O(e_4),$$  

where \(c_j = f^{(j)}(\alpha)/j!f'(\alpha), e_k = x_k - \alpha\).

Proof. The proof would be similar to the proofs given in [6].

Applying (10) on the matrix equation (9) will result in the following new matrix fixed-point-type iteration for finding (3):

$$X_{k+1} = (2I + 15X^2 + 3X^4) \left[9X_k + 11X^3_k\right]^{-1},$$  

where \(X_0 = A\). This is named PM1 from now on.

The proposed scheme (12) is not a member of Padé family [4]. Furthermore, applying (10) on the scalar equation \(g(x) = x^2 - 1\) provides a global convergence in the complex plane (except the points lying on the imaginary axis). This global behavior, which is kept for matrix case, has been illustrated in Figure 1 by drawing the basins of attraction for (6) and (8). The attraction basins for (7) (local convergence) and (12) (global convergence) are also portrayed in Figure 2.

Theorem 2. Let \(A \in \mathbb{C}^{n \times n}\) have no pure imaginary eigenvalues. Then, the matrix sequence \(\{X_k\}_{k=0}^{k=\infty}\) defined by (12) converges to \(S\), choosing \(X_0 = A\).

Proof. We remark that all matrices, whether they are diagonalizable or not, have a Jordan normal form \(A = JTJ^{-1}\), where the matrix \(J\) consists of Jordan blocks. For this reason, let \(A\) have a Jordan canonical form arranged as

$$T^{-1} A T = \Lambda = \begin{bmatrix} C & 0 \\ 0 & N \end{bmatrix},$$  

where
where \( T \) is a nonsingular matrix and \( C, N \) are square Jordan blocks corresponding to eigenvalues lying in \( \mathbb{C}^- \) and \( \mathbb{C}^+ \), respectively. We have

\[
\text{sign}(\Lambda) = \text{sign}(T^{-1}AT) = T^{-1}\text{sign}(A)T
\]

\[
= \text{diag}\left(\text{sign}(\lambda_1), \ldots, \text{sign}(\lambda_p), \text{sign}(\lambda_{p+1}), \ldots, \text{sign}(\lambda_n)\right).
\]  

(14)

If we define \( D_k = T^{-1}X_kT \), then, from the method (12), we obtain

\[
D_{k+1} = \left(2I + 15D_k^2 + 3D_k^4\right)\left[9D_k + 11D_k^3\right]^{-1}.
\]  

(15)

Note that if \( D_0 \) is a diagonal matrix, then, based on an inductive proof, all successive \( D_k \) are diagonal too. From (15), it is enough to show that \( \{D_k\} \) converges to \( \text{sign}(\Lambda) \). We remark that the case at which \( D_0 \) is not diagonal will be discussed later in the proof.

In the meantime, we can write (15) as \( n \) uncoupled scalar iterations to solve \( g(x) = x^2 - 1 = 0 \), given by

\[
d_{k+1}^i = \left(2 + 15d_k^2 + 3d_k^4\right)\left[9d_k + 11d_k^3\right]^{-1},
\]  

(16)

where \( d_k^i = (D_k)_{ii} \) and \( 1 \leq i \leq n \). From (15) and (16), it is enough to study the convergence of \( \{d_k^i\} \) to \( \text{sign}(\lambda_i) \).

It is known that \( \text{sign}(\lambda_i) = s_i = \pm 1 \). Thus, we attain

\[
\frac{d_{k+1}^i - 1}{d_{k+1}^i + 1} = \frac{(-1 + d_k^i)^3 (2 + 3d_k^i)}{(1 + d_k^i)^3 (2 + 3d_k^i)}.
\]  

(17)

Since \( |d_0^i| = |\lambda_i| > 0 \), we have

\[
\lim_{k \to \infty} \left|\frac{d_{k+1}^i - 1}{d_{k+1}^i + 1}\right| = 0.
\]  

(18)

and \( \lim_{k \to \infty} |d_k^i| = |\text{sign}(\lambda_i)| \). This shows that \( \{d_k^i\} \) is convergent.

In the convergence proof, \( D_0 \) may not be diagonal. Since the Jordan canonical form of some matrices may not be diagonal, thus, one cannot write (15) as \( n \) uncoupled scalar iterations (16). We comment that in this case our method is also convergent. To this goal, we must pursue the scalar relationship among the eigenvalues of the iterates for the studied rational matrix iteration.

In this case, the eigenvalues of \( X_k \) are mapped from the iterate \( k \) to the iterate \( k+1 \) by the following relation:

\[
\lambda_{k+1}^i = \left(2 + 15\lambda_k^2 + 3\lambda_k^4\right)\left[9\lambda_k + 11\lambda_k^3\right]^{-1}.
\]  

(19)

So, (19) clearly shows that the eigenvalues in the general case are convergent to \( \pm 1 \); that is to say,

\[
\lim_{k \to \infty} \left|\frac{\lambda_{k+1}^i - 1}{\lambda_{k+1}^i + 1}\right| = 0.
\]  

(20)

Consequently, we have

\[
\lim_{k \to \infty} X_k = T\left(\lim_{k \to \infty} D_k\right)T^{-1} = T\text{sign}(A)T^{-1} = \text{sign}(A).
\]  

(21)

The proof is ended.

\[ \square \]

**Theorem 3.** Let \( A \in \mathbb{C}^{n\times n} \) have no pure imaginary eigenvalues. Then the proposed method (12) converges cubically to the sign matrix \( S \).

**Proof.** Clearly, \( X_k \) are rational functions of \( A \) and, hence, like \( A \), commute with \( S \). On the other hand, we know that \( S^2 = I \),
Table 1: Results of comparisons for Example 5 using \( X_0 = A \).

| Methods | NM     | HM     | PM1    | PM2    |
|---------|--------|--------|--------|--------|
| IT      | 14     | 9      | 8      | 8      |
| \( R_{k+1} \) | \( 1.41584 \times 10^{-249} \) | \( 1.0266 \times 10^{-299} \) | \( 2.5679 \times 10^{-298} \) | \( 1.45091 \times 10^{-337} \) |
| \( \rho \) | 1.99077 | 3      | 3      | 3      |

Table 2: Results of comparisons for Example 6 using \( X_0 = A \).

| Methods | NM     | HM     | PM1    | PM2    |
|---------|--------|--------|--------|--------|
| IT      | 10     | 7      | 6      | 6      |
| \( R_{k+1} \) | \( 5.7266 \times 10^{-155} \) | \( 5.80819 \times 10^{-203} \) | \( 8.38265 \times 10^{-153} \) | \( 1.55387 \times 10^{-143} \) |
| \( \rho \) | 2.00228 | 3.00001 | 3.00015 | 3      |

\( S^{-1} = S, S^{2j} = I, \text{ and } S^{2j+1} = S, j \geq 1 \). Using the replacement or
\[ B_k = 9X_k + 11X_k^3, \]
we have

\[ X_{k+1} - S = (2I + 15X_k^2 + 3X_k^4)B_k^{-1} - S \]
\[ = (2I + 15X_k^2 + 3X_k^4 - SB_k)B_k^{-1} \]
\[ = (2I + 15X_k^2 + 3X_k^4 - 9SX_k - 11SX_k^3)B_k^{-1} \]
\[ = -(2S - 15X_k^2 - 3SX_k^3 + 9X_k + 11X_k^3) \]
\[ \times S^{-1}B_k^{-1} \]
\[ = (X_k - S)^3 (2I - 3SX_k)S^{-1}B_k^{-1}. \]

Now, using any matrix norm from both sides of (22), we attain
\[ \|X_{k+1} - S\| \leq \left( \|B_k^{-1}\| \left\| S^{-1} \right\| \left\| 2I - 3SX_k \right\| \right) \|X_k - S\|^3. \] (23)

This reveals the cubical rate of convergence for the new method (12). The proof is complete.

It should be remarked that the reciprocal iteration obtained from (12) is also convergent to the sign matrix (3) as follows:

\[ X_{k+1} = (9X_k + 11X_k^3) \left[ 2I + 15X_k^2 + 3X_k^4 \right]^{-1}, \] (24)

where \( X_0 = A \). This is named PM2. Similar convergence results as the ones given in Theorems 2-3 hold for (24).

A scaling approach to accelerate the beginning phase of convergence is normally necessary since the convergence rate cannot be seen in the initial iterates. Such an idea was discussed fully in [7] for Newton's method. An effective way to enhance the initial speed of convergence is to scale the iterates prior to each iteration; that is, \( X_k \) is replaced by \( \mu_k X_k \). Subsequently, we can present the accelerated forms of our proposed methods as follows:

\[ X_0 = A, \]
\[ \mu_k = \text{is the scaling parameter computed by (27),} \]

\[ X_{k+1} = (2I + 15\mu_k^2X_k^2 + 3\mu_k^4X_k^4) \left[ 9\mu_kX_k + 11\mu_k^3X_k^3 \right]^{-1}, \] (25)

\[ X_{k+1} = (9\mu_kX_k + 11\mu_k^3X_k^3) \left[ 2I + 15\mu_k^2X_k^2 + 3\mu_k^4X_k^4 \right]^{-1}, \] (26)
We apply here double precision arithmetic with the stop termination $R_{k+1} = \|X_{k+1}^2 - I\|_{\infty} \leq 10^{-5}$. Results are given in Figure 3.

**Example 5 (academic test).** We compute the matrix sign for the following complex test problem:

$$A = \begin{pmatrix} 0 & 10 & i & 7 + i \\ 7 & -5 & 6 & -5 \\ 0 & 60 & -2 & 9 \\ 0 & 5 & 9 & i \end{pmatrix},$$

where

$$S = \begin{pmatrix} 0.882671 + 0.0118589i & 0.461061 - 0.0519363i \\ 0.219355 + 0.00464485i & 0.136809 - 0.00840032i \\ -0.566306 - 0.0184534i & 2.22878 + 0.0471091i \\ 0.145285 + 0.00157401i & -0.57165 + 0.000347003i \end{pmatrix}.$$
Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgment

The authors would like to thank the referees for their helpful corrections and suggestions.

References

[1] J. D. Roberts, “Linear model reduction and solution of the algebraic Riccati equation by use of the sign function,” International Journal of Control, vol. 32, no. 4, pp. 677–687, 1980.

[2] C. S. Kenney and A. J. Laub, “The matrix sign function,” IEEE Transactions on Automatic Control, vol. 40, no. 8, pp. 1330–1348, 1995.

[3] N. J. Higham, Functions of Matrices: Theory and Computation, Society for Industrial and Applied Mathematics, Philadelphia, Pa, USA, 2008.

[4] C. Kenney and A. J. Laub, “Rational iterative methods for the matrix sign function,” SIAM Journal on Matrix Analysis and Applications, vol. 12, no. 2, pp. 273–291, 1991.

[5] F. Soleymani, P. S. Stanimirović, S. Shateyi, and F. K. Haghani, “Approximating the matrix sign function using a novel iterative method,” Abstract and Applied Analysis, vol. 2014, Article ID 105301, 9 pages, 2014.

[6] F. Soleymani, “Some high-order iterative methods for finding all the real zeros,” Thai Journal of Mathematics, vol. 12, no. 2, pp. 313–327, 2014.

[7] C. Kenney and A. J. Laub, “On scaling Newton’s method for polar decomposition and the matrix sign function,” SIAM Journal on Matrix Analysis and Applications, vol. 13, no. 3, pp. 698–706, 1992.

[8] S. Wagon, Mathematica in Action, Springer, New York, NY, USA, 3rd edition, 2010.

[9] F. Soleymani, E. Tohidi, S. Shateyi, and F. Haghani, “Some matrix iterations for computing matrix sign function,” Journal of Applied Mathematics, vol. 2014, Article ID 425654, 9 pages, 2014.