Hindawi Publishing Corporation
Abstract and Applied Analysis
Volume 2014, Article ID 105301, 9 pages
http://dx.doi.org/10.1155/2014/105301

Research Article

Approximating the Matrix Sign Function Using a Novel Iterative Method

F. Soleymani,[1] P. S. Stanimirović,[2] S. Shateyi,[3] and F. Khaksar Haghani[1]

[1] *Department of Mathematics, Islamic Azad University, Shahrekord Branch, Shahrekord, Iran*
[2] *Faculty of Sciences and Mathematics, University of Niš, Višegradska 33, 18000 Niš, Serbia*
[3] *Department of Mathematics and Applied Mathematics, University of Venda, Private Bag X5050, Thohoyandou 0950, South Africa*

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

Received 19 April 2014; Accepted 6 July 2014; Published 17 July 2014

Academic Editor: Juan R. Torregrosa

This study presents a matrix iterative method for finding the sign of a square complex matrix. It is shown that the sequence of iterates converges to the sign and has asymptotical stability, provided that the initial matrix is appropriately chosen. Some illustrations are presented to support the theory.

1. Introduction

The generic matrix function $f(A)$ of a given matrix $A \in \mathbb{C}^{n\times n}$ is defined formally by the integral representation

$$f(A) = \frac{1}{2\pi i}\oint_{\gamma} f(\zeta)(\zeta I - A)^{-1} d\zeta, \tag{1}$$

where $f : \Omega \to \mathbb{C}$ is an analytic function, $\Omega \subseteq \mathbb{C}$, and γ is a closed curve which encircles all eigenvalues of A (it should be contained in the domain of analyticity of f). The integral representation (1) is known as the Cauchy integral formula [1]. The integral of a matrix M should be understood as the matrix whose entries are the integrals of the entries of M. However, this mathematically appealing formula for computing the matrix functions is complicated and needs complex analysis to be fully understandable. Hence, several important other strategies for computing the matrix functions have been proposed and investigated, such as the Jordan canonical form and iterative methods for applied numerical problems (see, e.g., [1–3]).

In 1971, Roberts in [4] introduced the matrix sign function as a tool for model reduction and for solving Lyapunov and algebraic Riccati equations. He defined the sign function as a Cauchy integral and obtained the following integral representation of $\operatorname{sign}(A)$:

$$\operatorname{sign}(A) = S = \frac{2}{\pi}\int_0^{\infty}\left(t^2 I + A^2\right)^{-1} dt. \tag{2}$$

The matrix sign function is widely exploited in numerical linear algebra, especially in the computation of invariant subspaces and solutions of Riccati equations [5–7]. Note that the application of this function enables a matrix to be decomposed into two components whose spectra lie on opposite sides of the imaginary axis. The matrix sign function is a valuable tool for the numerical solution of Sylvester and Lyapunov matrix equations (see, e.g., [8]). An application of a generalization of the Newton iteration for the matrix sign function to the solution of the generalized algebraic Bernoulli equations was considered in [9].

Another application of this matrix function as a simple and direct method to derive some fundamental results in the theory of surface waves in anisotropic materials was presented in [10]. The authors of paper [11] investigated some practical iterations for matrix sector function which is a generalization of the matrix sign function.

Due to the applicability of the matrix sign function along with the difficulty of representation (2), stable iterative methods have become some viable choices.

The most important general family of matrix iterations for finding the matrix sign function S was introduced in [12] using Padé approximants to $f(\xi) = (1-\xi)^{-1/2}$ and the following characterization:

$$\operatorname{sign}(z) = s = \frac{z}{(z^2)^{1/2}} = \frac{z}{(1-\xi)^{1/2}}, \tag{3}$$

where $\xi = 1 - z^2$ and ξ is less than 1 in magnitude. Let the (m,n)-Padé approximant to $f(\xi)$ be $P_{m,n}(\xi)/Q_{m,n}(\xi)$ and $m + n \geq 1$. The iteration

$$z_{k+1} = \frac{z_k P_{m,n}\left(1-z_k^2\right)}{Q_{m,n}\left(1-z_k^2\right)} := \varphi_{2m+1,2n} \tag{4}$$

has been proved to be convergent to 1 and −1 with the order of convergence $m+n+1$ for $m \geq n-1$. Generally speaking, the iterations of Kenny and Laub (4), generated by the $[m/m]$ and $[(m-1)/m]$ Padé approximants, are globally convergent. A list of different iterations (in the scalar form) is given in Table 1.

Note that Iannazzo in [13] pointed out that these iterations can be obtained from the general König family (which goes back to Schröder [14, 15]) applied to the equation $z^2 - 1 = 0$. For a recent method in this area, one may refer to [16].

A lot of known methods could be extracted from the Padé family (4). For example, the Newton's iteration can be deduced as the reciprocal of the iteration corresponding to the case $m = 0$ and $n = 1$ in Table 1:

$$X_{k+1} = \frac{1}{2}\left(X_k + X_k^{-1}\right). \tag{5}$$

Choosing $m = n = 2$ yields the following fifth order method:

$$X_{k+1} = \left[I + 5X_k^2\left(2I + X_k^2\right)\right]\left[X_k\left(5I + 10X_k^2 + X_k^4\right)\right]^{-1}. \tag{6}$$

Similarly, options $m = 1$ and $n = 3$, or $m = 3$ and $n = 1$, or $m = 0$ and $n = 4$ result in the following fifth order methods, respectively:

$$X_{k+1} = -\left(-5I - 45X_k^2 - 15X_k^4 + X_k^6\right)\left[8X_k\left(3I + 5X_k^2\right)\right]^{-1}, \tag{7}$$

$$X_{k+1} = \left(8I + 56X_k^2\right)\left[X_k\left(35I + 35X_k^2 - 7X_k^4 + X_k^6\right)\right]^{-1}, \tag{8}$$

$$X_{k+1} = \left(35I + 140X_k^2 - 70X_k^4 + 28X_k^6 - 5X_k^8\right)\left[128X_k\right]^{-1}. \tag{9}$$

The remaining sections of this work are organized as follows. Section 2 presents the construction and the derivation of a new matrix iteration for finding S using the following *new* nonlinear equation solver:

$$\begin{aligned} y_k &= x_k - \frac{2}{3}\frac{f(x_k)}{f'(x_k)}, \\ z_k &= x_k - \frac{1}{2}\frac{3f'(y_k) + f'(x_k)}{3f'(y_k) - f'(x_k)}\frac{f(x_k)}{f'(x_k)}, \\ x_{k+1} &= z_k - \frac{f(z_k)}{f[y_k, z_k]}, \end{aligned} \tag{10}$$

where $f[y_k, z_k] = (f(z_k) - f(y_k))/(z_k - y_k)$. (10) is a combination of Jarratt's method [17] and a secant approach [18].

Note that (10) is a *novel* three-step iterative method (for the scalar case). Section 2 also studies the stability of the method and verifies its asymptotical stability. In Section 3, we discuss some other aspects of the new method, applicable in the implementation. Therein, we derive a new inversion-free method as well as a scaled method for finding S. Section 4 is devoted to the numerical examples for illustrating the convergence behavior of the new method against the existing ones. We emphasize that our constructed solver possesses the same global convergence rate as (6), but numerical example will reveal that it has an equal or faster behavior for solving many randomly generated matrices. This would be a clear advantage of our solver in contrast to (6). Finally, concluding remarks will be drawn in Section 5.

2. A Novel Iterative Method

Throughout this work it is assumed that $A \in \mathbb{C}^{n\times n}$ has no eigenvalues on the imaginary axis, so that $\operatorname{sign}(A)$ is defined. Note that this assumption implies that A is nonsingular.

As discussed in the fundamental article of Kenny and Laub in 1995 [19], the construction of new matrix methods for the matrix sign S is related to the iterative methods for finding the solution of nonlinear scalar equations. Let us apply (10) to the following nonlinear matrix equation:

$$X^2 = I, \tag{11}$$

in which I is the identity matrix. In fact, the sign S is a solution of the matrix equation (11). After application and simplification of its reciprocal, we obtain

$$X_{k+1} = \left(7X_k + 30X_k^3 + 11X_k^5\right)\left[I + 20X_k^2 + 25X_k^4 + 2X_k^6\right]^{-1}, \tag{12}$$

where $X_0 = A$.

Iannazzo in [20] mentioned that a matrix convergence is governed by the corresponding scalar convergence. Since the method (10) reads the following error equation:

$$e_{k+1} = \frac{1}{3}\left(c_2^4 - c_2^2 c_3\right)e_k^5 + O\left(e_k^6\right), \tag{13}$$

wherein $c_j = f^{(j)}(\alpha)/(j!f'(\alpha))$ and $e_k = x_k - \alpha$, therefore its corresponding matrix method (12) converges with fifth order

Table 1: Iteration functions for finding the sign (in the scalar form) from the Padé family (4).

	$n = 0$	$n = 1$	$n = 2$	$n = 3$	$n = 4$
$m = 0$	x	$\frac{2x}{1+x^2}$	$\frac{8x}{3+6x^2-x^4}$	$\frac{16x}{5+15x^2-5x^4+x^6}$	$\frac{128x}{35+140x^2-70x^4+28x^6-5x^8}$
$m = 1$	$\frac{x}{2}(3-x^2)$	$\frac{x(3+x^2)}{1+3x^2}$	$\frac{4x(1+x^2)}{1+6x^2+x^4}$	$-\frac{8x\left(3+5x^2\right)}{-5-45x^2-15x^4+x^6}$	$\frac{64x\left(3+7x^2\right)}{35+420x^2+210x^4-28x^6+3x^8}$
$m = 2$	$\frac{x}{8}(15-10x^2+3x^4)$	$\frac{x}{4}\frac{15+10x^2-x^4}{1+5x^2}$	$\frac{x(5+10x^2+x^4)}{1+10x^2+5x^4}$	$\frac{6x+20x^3+6x^5}{1+15x^2+15x^4+x^6}$	$-\frac{16x\left(3+7x^2\left(2+x^2\right)\right)}{-7-140x^2-210x^4-28x^6+x^8}$
$m = 3$	$-\frac{x\left(-35+35x^2-21x^4+5x^6\right)}{16}$	$\frac{x\left(35+35x^2-7x^4+x^6\right)}{8+56x^2}$	$\frac{35x+105x^3+21x^5-x^7}{6+84x^2+70x^4}$	$\frac{x\left(7+35x^2+21x^4+x^6\right)}{1+7x^2\left(3+5x^2+x^4\right)}$	$\frac{8\left(x+7x^3+7x^5+x^7\right)}{1+28x^2+70x^4+28x^6+x^8}$
$m = 4$	$\frac{x\left(315-420x^2+378x^4-180x^6+35x^8\right)}{128}$	$\frac{315x+420x^3-126x^5+36x^7-5x^9}{64+576x^2}$	$\frac{105x+420x^3+126x^5-12x^7+x^9}{16+288x^2+336x^4}$	$\frac{63x+420x^3+378x^5+36x^7-x^9}{8\left(1+3x^2\left(9+7x^2\left(3+x^2\right)\right)\right)}$	$\frac{x\left(3+x^2\right)\left(3+27x^2+33x^4+x^6\right)}{1+36x^2+126x^4+84x^6+9x^8}$

of convergence. But the question is whether this convergence is local or global. To answer this question, we draw the basins of attraction for the new scheme along with the existing methods of various orders.

It is shown in Figures 1–3 that methods (5), (6), and (12) are globally convergent, while methods (7), (8), and (9) are locally convergent (if one chooses a matrix A (in Figure 2 (left)) with one eigenvalue with negative real part, but in a green petal, then the matrix iteration will not converge to S).

The higher order convergence of (12) made its basins larger and lighter in contrast to (5). Hence, the new method could be of interest due to its global fifth order of convergence for finding S.

Now, an important challenge, that must be proven, is to show the stability of the new method (12) for finding the matrix sign function. This will be done formally in what follows.

Definition 1 (stability, see [1]). Consider an iteration $X_{k+1} = g(X_k)$ with a fixed point X. Assume that g is Fréchet differentiable at X. The iteration is stable in a neighborhood of X if the Fréchet derivative $L_g(X)$ has bounded powers; that is, there exists a constant c such that $\|L_g^i(X)\| \le c$ for all $i > 0$.

We investigate the stability of (12) for finding S in a neighborhood of the solution of (11). In fact, we analyze how a small perturbation at the kth iterate is amplified or damped along the iterates. Stability concerns behavior close to convergence and so is an asymptotic property.

Lemma 2. *The sequence* $\{X_k\}_{k=0}^{k=\infty}$ *generated by* (12) *is asymptotically stable.*

Proof. If X_0 is a function of A, then the iterates from (12) are all functions of A and hence commute with A. Commutativity properties are frequently used when deriving a matrix iteration for finding S.

In this study, we restrict the analysis to asymptotically small perturbations; that is, we use the differential error analysis.

Let ΔX_k be the numerical perturbation introduced at the kth iterate of (12). Next, one has

$$\widetilde{X}_k = X_k + \Delta X_k. \tag{14}$$

Here, we perform a first order error analysis; that is, we formally use approximations $(\Delta X_k)^i \approx 0$, since $(\Delta X_k)^i$, $i \ge 2$, is close to the zero (matrix). This formal manipulation is meaningful if ΔX_k is sufficiently small. We have

$$\begin{aligned}\widetilde{X}_{k+1} &= \left(7\widetilde{X}_k + 30\widetilde{X}_k^3 + 11\widetilde{X}_k^5\right)\left[I + 20\widetilde{X}_k^2 + 25\widetilde{X}_k^4 + 2\widetilde{X}_k^6\right]^{-1} \\ &= \left(7\left(X_k + \Delta X_k\right) + 30\left(X_k + \Delta X_k\right)^3 + 11\left(X_k + \Delta X_k\right)^5\right) \\ &\quad \times \left[I + 20\left(X_k + \Delta X_k\right)^2 \right. \\ &\qquad \left. + 25\left(X_k + \Delta X_k\right)^4 + 2\left(X_k + \Delta X_k\right)^6\right]^{-1} \\ &\approx \left(S + \frac{100}{48}\Delta X_k + \frac{52}{100}S\Delta X_k S\right) \\ &\quad \times \left(I - \frac{19}{12}S\Delta X_k - \frac{19}{12}\Delta X_k S\right) \\ &\approx S - \frac{24}{48}\Delta X_k + \frac{24}{48}S\Delta X_k S,\end{aligned} \tag{15}$$

where the following identities are used (for any nonsingular matrix B and the matrix C):

$$(B + C)^{-1} \approx B^{-1} - B^{-1}CB^{-1}. \tag{16}$$

Note that after some algebraic manipulations and using $\Delta X_{k+1} = \widetilde{X}_{k+1} - X_{k+1}$, we have (assuming $X_k \approx \operatorname{sign}(A) = S$ for enough large k)

$$\Delta X_{k+1} \approx -\frac{1}{2}\Delta X_k + \frac{1}{2}S\Delta X_k S. \tag{17}$$

We used the following facts on the matrix sign function $S^2 = I$, and $S^{-1} = S$. We can now conclude that the perturbation at the iterate $k+1$ is bounded; that is,

$$\|\Delta X_{k+1}\| \le \frac{1}{2^{k+1}}\|S\Delta X_0 S - \Delta X_0\|. \tag{18}$$

Therefore, the sequence $\{X_k\}_{k=0}^{k=\infty}$ generated by (12) is asymptotically stable. This ends the proof. □

Theorem 3. *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 the matrix sign* S.

Proof. Let A have a Jordan canonical form arranged as

$$V^{-1}AV = \Lambda = \begin{bmatrix} C & 0 \\ 0 & N \end{bmatrix}, \tag{19}$$

where V is a nonsingular matrix and C, N are square Jordan blocks corresponding to eigenvalues lying in $\mathbb{C}^-$ (open left-half complex plane) and $\mathbb{C}^+$ (open right-half complex plane), respectively. Denote by $\lambda_1, \ldots \lambda_p$ and $\lambda_{p+1}, \ldots \lambda_n$ values lying on the main diagonals of blocks C and N, respectively.

Since V is invertible, we know that sign(A) is diagonalizable and [1]

$$\operatorname{sign}(A) = V\begin{bmatrix} -I_p & 0 \\ 0 & I_{n-p} \end{bmatrix}V^{-1}. \tag{20}$$

Therefore,

$$\begin{aligned}&\operatorname{sign}(\Lambda) \\ &\quad = \operatorname{sign}\left(V^{-1}AV\right) = V^{-1}\operatorname{sign}(A)V \\ &\quad = \operatorname{diag}\left(\operatorname{sign}(\lambda_1), \ldots, \operatorname{sign}(\lambda_p),\right. \\ &\qquad\qquad \left.\operatorname{sign}(\lambda_{p+1}), \ldots, \operatorname{sign}(\lambda_n)\right).\end{aligned} \tag{21}$$

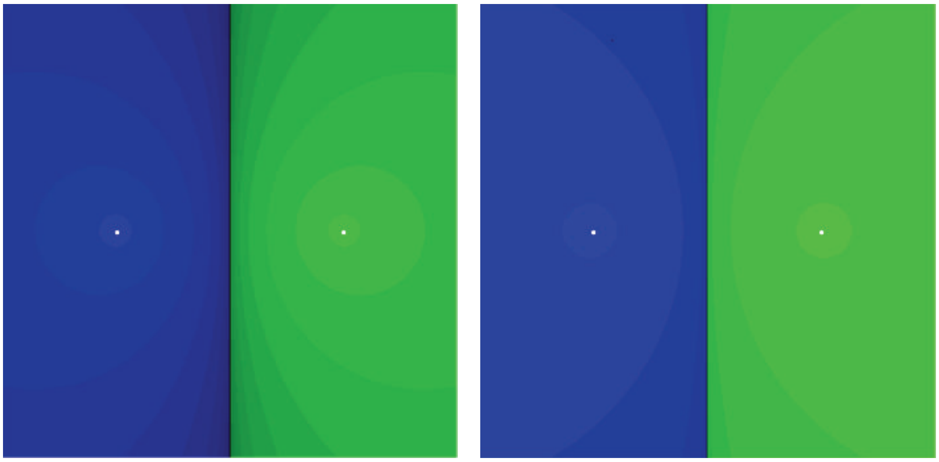

Figure 1: The basins of attraction for (5) (left) and (6) (right), for the polynomial $x^2 - 1 = 0$ (shaded by the number of iterations to obtain the solution).

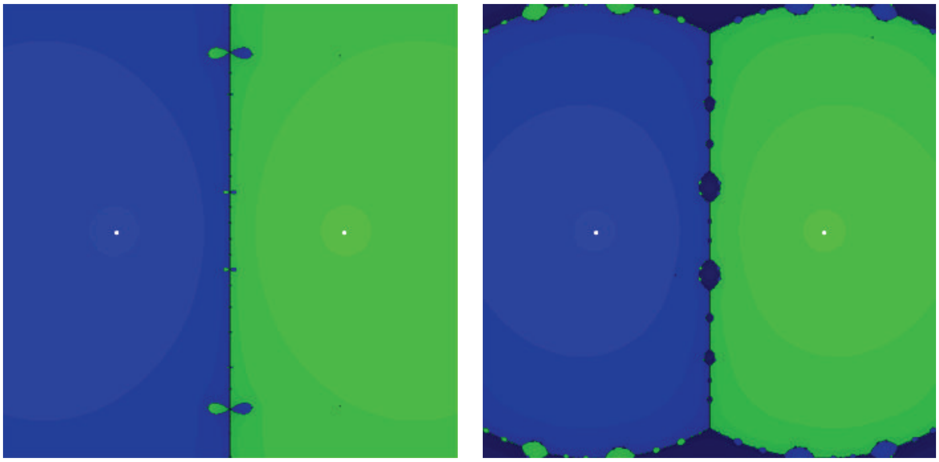

Figure 2: The basins of attraction for (7) (left) and (8) (right), for the polynomial $x^2 - 1 = 0$ (shaded by the number of iterations to obtain the solution).

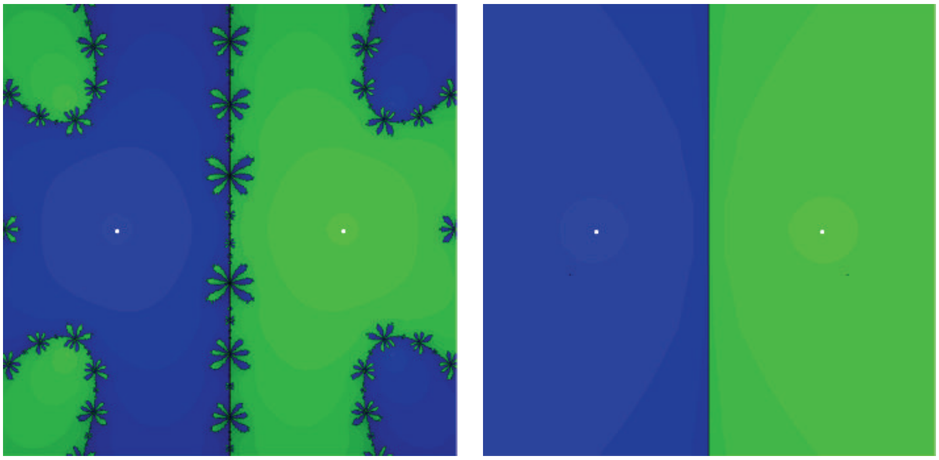

Figure 3: The basins of attraction for (9) (left) and (12) (right), for the polynomial $x^2 - 1 = 0$ (shaded by the number of iterations to obtain the solution).

On the other hand, if we define $D_k = V^{-1}X_kV$, then from the equation (12) we obtain

$$D_{k+1} = \left(7D_k + 30D_k^3 + 11D_k^5\right)\left[I + 20D_k^2 + 25D_k^4 + 2D_k^6\right]^{-1}. \tag{22}$$

Notice that if D_0 is a diagonal matrix, then all successive D_k are diagonal too. From (22), it is enough to prove that $\{D_k\}$ converges to $\operatorname{sign}(\Lambda)$, in order to ensure the convergence of the sequence generated by (12) to $\operatorname{sign}(A)$.

We can write (22) as n uncoupled scalar iterations to solve $g(x) = x^2 - 1 = 0$, given by

$$d_{k+1}^i = \frac{7d_k^i + 30{d_k^i}^3 + 11{d_k^i}^5}{1 + 20{d_k^i}^2 + 25{d_k^i}^4 + 2{d_k^i}^6}, \tag{23}$$

where $d_k^i = (D_k)_{i,i}$ and $1 \le i \le n$. From (22) and (23), it is enough to study the convergence of $\{d_k^i\}$ to $\operatorname{sign}(\lambda_i)$, for all $1 \le i \le n$.

From (23) and since the eigenvalues of A are not pure imaginary, we have that $\operatorname{sign}(\lambda_i) = s_i = \pm 1$. Thus, we attain

$$\frac{d_{k+1}^i - 1}{d_{k+1}^i + 1} = -\frac{\left(-1 + d_k^i\right)^5\left(-1 + 2d_k^i\right)}{\left(1 + d_k^i\right)^5\left(1 + 2d_k^i\right)}. \tag{24}$$

From the other point of view, since $|d_0^i| = |\lambda_i| > 0$, we obtain

$$\lim_{k\to\infty}\left|\frac{d_{k+1}^i - 1}{d_{k+1}^i + 1}\right| = 0, \tag{25}$$

and $\lim_{k\to\infty}|d_k^i| = 1 = |\operatorname{sign}(\lambda_i)|$. This shows that $\{d_k^i\}$ is convergent. Now, one concludes that $\lim_{k\to\infty}D_k = \operatorname{sign}(\Lambda)$.

Recalling $D_k = V^{-1}X_kV$, we have

$$\lim_{k\to\infty} X_k = V\left(\lim_{k\to\infty} D_k\right)V^{-1} = V\operatorname{sign}(\Lambda)\,V^{-1} = \operatorname{sign}(A), \tag{26}$$

and subsequently the convergence is established. The proof is complete. □

Theorem 4. *Consider the same conditions of Lemma 2 and Theorem 3. Then the proposed method* (12) *has fifth order of convergence 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 $S^2 = I$, $S^{-1} = S$, $S^{2j} = I$, and $S^{2j+1} = S$, $j \ge 1$. Using the replacement $B_k = I + 20X_k^2 + 25X_k^4 + 2X_k^6$ (for the sake of simplicity), we have

$$\begin{aligned}
X_{k+1} - S
&= \left(7X_k + 30X_k^3 + 11X_k^5\right)B_k^{-1} - S\\
&= \left(7X_k + 30X_k^3 + 11X_k^5 - SB_k\right)B_k^{-1}\\
&= \left(-S + 7X_k - 20SX_k^2 + 30X_k^3\right.\\
&\qquad \left.-25SX_k^4 + 11X_k^5 - 2SX_k^6\right)B_k^{-1}\\
&= \left(-S^5 + 5S^4X_k + 2X_k - 10S^3X_k^2 - 10SX_k^2\right.\\
&\qquad + 10S^2X_k^3 + 20X_k^3 - 5SX_k^4 - 20SX_k^4\\
&\qquad \left.+ X_k^5 + 10X_k^5 - 2SX_k\right)B_k^{-1}\\
&= \left(\left(X_k - S\right)^5 + 2X_k - 10SX_k^2 + 20X_k^3\right.\\
&\qquad \left.-20SX_k^4 + 10X_k^5 - 2SX_k\right)B_k^{-1}\\
&= \left(\left(X_k - S\right)^5 - 2SX_k\right.\\
&\qquad \left.\times\left(-S + 5X_k - 10SX_k^2 + 10X_k^3 - 5SX_k^4 + X_k^5\right)\right)B_k^{-1}\\
&= \left(\left(X_k - S\right)^5 - 2SX_k\left(X_k - S\right)^5\right)B_k^{-1}\\
&= \left(X_k - S\right)^5\left(I - 2SX_k\right)B_k^{-1}.
\end{aligned} \tag{27}$$

Now, using any matrix norm from both sides of (27), we attain

$$\left\|X_{k+1} - S\right\| \le \left(\left\|B_k^{-1}\right\|\left\|I - 2SX_k\right\|\right)\left\|X_k - S\right\|^5. \tag{28}$$

This reveals the fifth order of convergence for the new method (12). The proof is complete. □

3. Multiplication-Rich and Scaled Variants of the New Method

In general, reduction of a problem in numerical linear algebra to the matrix inversion problem is not an advisable technique. The iterations (5)–(9) as well as (12) require explicit computation of a matrix inverse in each iterative step. Since explicit usage of the inverse matrix is relatively rare in numerical analysis, there is a normal aspiration to approximate the inverse. Such discussions and variants for (12) will be given in the following subsections.

3.1. Solve a Matrix Equation Instead of the Matrix Inverse. One of the ways to avoid explicit usage of the inverse matrix is to solve corresponding system of linear matrix equations instead of the matrix inverse. This is done in Algorithm 1.

3.2. Use Approximation of the Matrix Inverse. Another tendency is to give matrix multiplication-rich iterations which retain the convergence rate of the method. For example, the inverse X_k^{-1} in (5) can be replaced by one step of Schulz method for the matrix inverse, which has the form $X_k(2I - X_k^2)$. This replacement produces the Newton-Schulz iteration

$$X_{k+1} = \frac{1}{2}X_k\left(3I - X_k^2\right), \qquad X_0 = A, \tag{29}$$

```
Given a suitable $X_0 \in \mathbb{C}^{n\times n}$
for $k = 0, 1, \ldots$ until convergence do
    $B_k = I + 20X_k^2 + 25X_k^4 + 2X_k^6$
    Solve the linear system $B_k X_{k+1} = 7X_k + 30X_k^3 + 11X_k^5$ for $X_{k+1}$
end for
```

Algorithm 1: The new method for computing the matrix sign.

```
Given a suitable $X_0 \in \mathbb{C}^{n\times n}$
use (12) until $\|X_k^2 - I\|_F < \varepsilon = 1$
set $X_0 = X_k$
for $k = 0, 1, \ldots$ until convergence do
    $C_k = (I - 40X_k^2 - 450X_k^4 - 1004X_k^6 - 705X_k^8 - 100X_k^{10} - 4X_k^{12})$
    $X_{k+1} = X_k (7I + 30X_k^2 + 11X_k^4)(I + 20X_k^2 + 25X_k^4 + 2X_k^6) C_k$
end for
```

Algorithm 2: The new (inversion-free) method for computing the matrix sign.

which is multiplication-rich and retains the quadratic convergence of Newton's method. However, it is only locally convergent, with convergence guaranteed for $\|I-A^2\| < 1$ (see [1]). We apply similar idea to (12). For the sake of simplicity, we use the notation $B_k = I + 20X_k^2 + 25X_k^4 + 2X_k^6$. Replacing B_k^{-1} with $B_k(2I-B_k^2)$ in (12), we get the following iterative rule for computing the matrix sign function:

$$\begin{aligned} X_{k+1} = X_k &\left(7I + 30X_k^2 + 11X_k^4\right) \\ &\times \left(I + 20X_k^2 + 25X_k^4 + 2X_k^6\right) \\ &\times \left(I - 40X_k^2 - 450X_k^4 - 1004X_k^6 \right. \\ &\qquad \left. -705X_k^8 - 100X_k^{10} - 4X_k^{12}\right). \end{aligned} \tag{30}$$

Formula (30) is multiplication-rich with convergence guaranteed for $\|I - A^2\| < 1$.

Note that inverse-free algorithms are suitable for the implementation on vector and parallel computers. The iterative scheme (30) includes 9 matrix multiplications, while the complexity of (12) contains 6 matrix multiplications and one matrix inversion to achieve fifth convergence order.

3.3. A Hybrid Method. An efficient algorithm for finding the sign by avoiding the computation of matrix inverse is to use (12) or (33) in initial iterations, until X_k^2 is close enough to I, and in a subsequent stage apply (30). Such a switching approach was proposed originally in [19]. A similar idea is exploited in Algorithm 2.

3.4. Scaling Method. For lower order methods, such as (5), the convergence is slow at the beginning. In fact, once the error is sufficiently small (in practice, less than, say, 0.5), successive errors decrease rapidly, each being approximately the square of the previous one. However, initial convergence can be slow if the iteration X_k has a large eigenvalue, that is, in the case $\|X_k\| \gg 1$. Hence, a scaling approach to accelerate the beginning of this phase is necessary and can be done in what follows [7] for the Newton's method:

$$\begin{gathered} X_0 = A, \\ \mu_k = \text{is the scaling parameter computed by } (32), \\ X_{k+1} = \frac{1}{2}\left(\mu_k X_k + \mu_k^{-1} X_k^{-1}\right), \end{gathered} \tag{31}$$

wherein

$$\mu_k = \begin{cases} \sqrt{\dfrac{\|X_k^{-1}\|}{\|X_k\|}}, & (\text{norm scaling}), \\ \sqrt{\dfrac{\rho\left(X_k^{-1}\right)}{\rho\left(X_k\right)}}, & (\text{spectral scaling}), \\ \sqrt{\left|\det\left(X_k\right)\right|^{-1/n}}, & (\text{determinantal scaling}). \end{cases} \tag{32}$$

Such an approach could be done to refine the initial matrix and to provide a much more robust initial matrix to arrive at the convergence phase rapidly.

The new iteration (12) is quite fast and reliable due to the discussions in Sections 2 and 3. However, a way is open for speeding up its initial phase of convergence via the concept of scaling.

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 $\mu_k X_k$. Such an idea can simply be done in what follows:

$$\begin{gathered} X_0 = A, \\ \mu_k = \text{is the scaling parameter computed by } (32), \\ X_{k+1} = \left(7\mu_k X_k + 30\mu_k^3 X_k^3 + 11\mu_k^5 X_k^5\right) \\ \qquad \times \left[I + 20\mu_k^2 X_k^2 + 25\mu_k^4 X_k^4 + 2\mu_k^6 X_k^6\right]^{-1}, \end{gathered} \tag{33}$$

```
Given a suitable $X_0 \in \mathbb{C}^{n\times n}$
use (33) until $\|X_k^2 - I\|_F < \varepsilon = 1$
set $X_0 = X_k$
for $k = 0, 1, \ldots$ until convergence do
    $C_k = (I - 40X_k^2 - 450X_k^4 - 1004X_k^6 - 705X_k^8 - 100X_k^{10} - 4X_k^{12})$
    $X_{k+1} = X_k (7I + 30X_k^2 + 11X_k^4)(I + 20X_k^2 + 25X_k^4 + 2X_k^6) C_k$
end for
```

Algorithm 3: The new scaling method for computing the matrix sign.

Table 2: Results of comparisons for Example 5.

Matrices	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10
Number of iterations for (5)	15	16	14	14	15	14	13	13	13	13
Number of iterations for (6)	7	7	6	6	7	6	6	6	6	6
Number of iterations for (12)	6	7	6	6	6	6	5	6	6	5

Table 3: Results of comparisons for Example 6 with norm scaling.

Matrices	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10
Number of iterations for the scaled (5)	12	11	11	12	13	12	11	10	14	10
Number of iterations for the scaled (6)	5	6	5	5	5	5	5	5	5	5
Number of iterations for the scaled (12)	5	5	5	5	5	5	4	5	5	4

where $\lim_{k\to\infty}\mu_k = 1$ and $\lim_{k\to\infty}X_k = S$. Algorithm 3 illustrates an efficient method based on (33) for finding S.

4. Experimental Results

There are basically two general ways to terminate a matrix iteration for finding S, that is, to stop when X_k has relative error below a tolerance or to stop when some suitable residual (such as $X_k^2 - I$) is below the tolerance. The relevant aim will in general be problem dependent. However, the stop termination is really important in matrix methods.

The considered stopping termination in this section would be the safe strategy introduced in [1] as follows:

$$\frac{\left\|X_k^2 - I\right\|_*}{\left\|X_k\right\|_*^2} \leq \epsilon. \tag{34}$$

For comparisons, we have used the matrix globally convergent methods (5), (6), and (12) using Mathematica 8 built-in precision, [21]. We used Mathematica function `Inverse[]` to compute the required matrix inverse. Implementation details of the function `Inverse[]` are based on efficient row reduction (Gaussian elimination) based on numerical approximation.

Example 5. In this test, we examine the behavior of different iterative methods for finding the matrix sign function of 10 randomly generated complex square 70 × 70 matrices as follows:

```
n = 70; number = 10; SeedRandom[12345];
Table[A[l] = RandomComplex[{−5 − I, 5 + I}, {n,
n}];, {l, number}].
```

The results of comparisons in terms of the number of iterations have been reported in Table 2. We remark that whatever the eigenvalues of a matrix are closer to the imaginary axis, the speed of convergence for different methods becomes slower and more risky to face with singular matrices X_k, whose inverses could not be computed.

In this example, we have used the stopping criterion (34) with $\epsilon = 10^{-10}$, matrix norm one, and $X_0 = A$ as the initial matrix.

Example 6. In order to confirm the acceleration via scaling, we have reran Example 5 with the norm scaling. The results are summarized in Table 3. The numerics reverify the effectiveness of the new method (12) in finding the matrix sign.

5. Summary

Interest in the sign function grew steadily starting from the 1970s and 1980s, initially among engineers and later among numerical analysts. Following such a trend, in this study we proposed a fifth order new iterative method for finding the matrix sign S.

Applying the basins of attraction in the complex plane, based on the results from [20], we concluded that the introduced method (12) has global convergence. We then theoretically found that it is asymptotically stable. Some numerical experiments have been studied in order to show the faster convergence on the basis of a smaller number of iterations.

Several modifications of the introduced method have been established, such as a new inversion-free method, a composite method, and a scaled version.

Conflict of Interests

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

Acknowledgments

The second author gratefully acknowledges support from the Research Project 174013 of the Serbian Ministry of Science. The authors are grateful to an anonymous reviewer for carefully reading their paper and helping to improve the presentation.

References

[1] N. J. Higham, *Functions of Matrices: Theory and Computation*, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, Pa, USA, 2008.

[2] J. L. Howland, "The sign matrix and the separation of matrix eigenvalues," *Linear Algebra and Its Applications*, vol. 49, pp. 221–232, 1983.

[3] J. Leyva-Ramos, "A note on mode decoupling of linear time-invariant systems using the generalized sign matrix," *Applied Mathematics and Computation*, vol. 219, no. 22, pp. 10817–10821, 2013.

[4] 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.

[5] Z. Bai and J. Demmel, "Using the matrix sign function to compute invariant subspaces," *SIAM Journal on Matrix Analysis and Applications*, vol. 19, no. 1, pp. 205–225, 1998.

[6] R. Byers, C. He, and V. Mehrmann, "The matrix sign function method and the computation of invariant subspaces," *SIAM Journal on Matrix Analysis and Applications*, vol. 18, no. 3, pp. 615–632, 1997.

[7] C. S. Kenney, A. J. Laub, and P. M. Papadopoulos, "Matrix-sign algorithms for Riccati equations," *IMA Journal of Mathematical Control and Information*, vol. 9, no. 4, pp. 331–344, 1992.

[8] P. Benner and E. S. Quintana-Ortí, "Solving stable generalized Lyapunov equations with the matrix sign function," *Numerical Algorithms*, vol. 20, no. 1, pp. 75–100, 1999.

[9] S. Barrachina, P. Benner, and E. S. Quintana-Ort, "Efficient algorithms for generalized algebraic Bernoulli equations based on the matrix sign function," *Numerical Algorithms*, vol. 46, no. 4, pp. 351–368, 2007.

[10] A. N. Norris, A. L. Shuvalov, and A. A. Kutsenko, "The matrix sign function for solving surface wave problems in homogeneous and laterally periodic elastic half-spaces," *Wave Motion*, vol. 50, no. 8, pp. 1239–1250, 2013.

[11] B. Laszkiewicz and K. Zietak, "Algorithms for the matrix sector function," *Electronic Transactions on Numerical Analysis*, vol. 31, pp. 358–383, 2008.

[12] 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.

[13] B. Iannazzo, *Numerical solution of certain nonlinear matrix equations [Ph.D. thesis]*, Universita degli Studi di Pisa, Facolta di Scienze Matematiche, Fisiche e Naturali, Pisa, Italy, 2007.

[14] E. Schröder, "Ueber unendlich viele Algorithmen zur Auflösung der Gleichungen," *Mathematische Annalen*, vol. 2, no. 2, pp. 317–365, 1870.

[15] E. Schröder, "On infinitely many algorithms for solving equations," Tech. Rep. TR-92-121, Department of Computer Science, University of Maryland, College Park, Md, USA, 1992.

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

[17] P. Jarratt, "Some fourth order multi-point iterative methods for solving equations," *Mathematics of Computation*, vol. 20, no. 95, pp. 434–437, 1966.

[18] A. Iliev and N. Kyurkchiev, *Nontrivial Methods in Numerical Analysis: Selected Topics in Numerical Analysis*, LAP Lambert, 2010.

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

[20] B. Iannazzo, "A family of rational iterations and its application to the computation of the matrix pth root," *SIAM Journal on Matrix Analysis and Applications*, vol. 30, no. 4, pp. 1445–1462, 2008.

[21] M. Trott, *The Mathematica Guide-Book for Numerics*, Springer, New York, NY, USA, 2006.