Computational and Applied Mathematics (2019) 38:53
https://doi.org/10.1007/s40314-019-0800-1

# New iterative methods for finding matrix sign function: derivation and application

**Mohammad Momenzadeh**[1] · **Taher Lotfi**[1]

Received: 22 September 2018 / Revised: 22 November 2018 / Accepted: 16 January 2019 /
Published online: 12 March 2019


**Abstract**
The objective of this research was to provide two new methods for the sign function of a matrix. It is discussed that the schemes are novel and present global convergence behaviors. Then, the high convergence speeds of these iterative methods are proved and confirmed for calculating the matrix sign of different types of nonsingular matrices to reveal their applicability over the existing iterative methods of the same type.

**Keywords** Sign function · Matrix iterations · Eigenvalues · High order · Attraction basin

**Mathematics Subject Classification** 65F30 · 41A25 · 65F60

## 1 Introductory notes

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 $\gamma$ 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 (Higham 2008). 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.

Communicated by Jinyun Yuan.

✉ Taher Lotfi
lotfi@iauh.ac.ir; lotfitaher@yahoo.com

Mohammad Momenzadeh
m.momenzadeh12@yahoo.com

[1] Department of Mathematics, Hamedan Branch, Islamic Azad University, Hamedan, Iran

Filbir 1994; Howland 1983). Among wide application of matrix functions in mathematics, we refer the readers to computing the inverse of ill-conditioned matrices that occur when solving financial PDEs using radial basis function (RBF) or finite difference (FD) methods (Company et al. 2016; Soleymani and Zaka Ullah 2018; Soleymani et al. 2018).

In 1971, Roberts in Roberts (1980) 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 sign$(A)$:

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

The matrix sign function has basic theoretical and algorithmic relations with the matrix square root, the polar decomposition (see e.g., Higham et al. 2004), and from time to time, with the matrix $p$th roots [see for more Higham (2008, chapter 5)]. For example, a large class of iterations for the matrix square root can be obtained from corresponding iterations for the matrix sign function, and due to this discussing and designing new iterative schemes for finding matrix sign function ($S$) is requisite.

Here we suppose that matrix $A \in \mathbb{C}^{n\times n}$ has no eigenvalues on the imaginary axis. This means that the matrix sign function $S$ can be uniquely defined ($A$ is a nonsingular square matrix). The most concise definition of the matrix sign decomposition is given in Higham (1994), Kenney and Laub (1991) as follows:

$$A = SN = A(A^2)^{-1/2}(A^2)^{1/2}, \tag{3}$$

whereas $S = A(A^2)^{-1/2}$ is the matrix sign function and $1/2$ denotes the principal matrix square root of a given matrix.

As a matter of fact, matrix disc function can be used to obtain invariant subspaces in an analogous way as for the matrix sign function. For other applications of matrix sign function, we refer the reader to Misrikhanov and Ryabchenko (2008), Shieh et al. (1984).

This matrix function has several properties. Some of them are given by Kenney and Laub (1991):

1. $S$ is involutory, viz, $S^2 = I$.
2. $S$ is diagonalizable with eigenvalues $\pm 1$.
3. $SA = AS$.
4. If $A$ is real, then $S$ is real.
5. $(I + S)/2$ and $(I - S)/2$ are projectors onto the invariant subspaces associated with the eigenvalues in the right half-plane and left half-plane, respectively.

For more information on the matrix sign function the reader is referred to Chapter 5 in Iannazzo (2007). Recall that a primary matrix function with a non-primary flavor is the matrix sign function, which for a matrix $A \in \mathbb{C}^{n\times n}$ is a (generally) non-primary square root of $I$ that depends on $A$ Higham (2008, p. 16).

A number of matrix functions $f(A)$ are amenable to computation by iteration functions of the following form Higham (2008, p. 91):

$$X_{k+1} = g(X_k), \tag{4}$$

where for the iterations used in practice, $X_0$ is not arbitrary but is a fixed function of $A$. Taking into account of computational burden makes it obvious that $g$ is a polynomial or rational function. A rational $g$ would require the solution of linear systems with multiple right-hand sides, or even explicit matrix inversion.

It is necessary to recall that outcomes and intuition from scalar nonlinear iterations do not necessarily generalize to the matrix case. As an illustration, standard convergence conditions expressed in terms of derivatives of $g$ at a fixed point in the scalar case do not directly translate into analogous conditions on the Frechét and higher order derivatives in the matrix case.

The main motivation of the present study is to complete the recent discussion given in Cordero et al. (2017, 2016) by proposing higher order method and a way to increase the convergence order for finding the matrix sign function via applying more steps in the main iterative method considered for solving nonlinear algebraic equations. We here provide two new methods of convergence speeds, five and seven, and it is discussed that they converge to the sign matrix if and only if the underlying matrix $A$ does not have any pure imaginary eigenvalues.

The rest of this work is unfolded as follows: In Sect. 2, some of the most important existing methods of the literature are reviewed. Then in Sect. 3 several new methods are derived. The derivation is based on an extension of one of the known matrix iterations. It is proved that the methods are convergence and possess global convergence behavior, which is an important feature for tackling matrix functions numerically. Thence, the application of the new methods for finding matrix sign functions of matrices with different structures is given in Sect. 4. Final remarks are pointed out in Sect. 5 along with some future work outlines.

## 2 The literature on iterative methods

The most common and well-known way for finding the sign of a square nonsingular matrix is the following numerical method:

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

which is also known as Newton's method (NM) and converges quadratically when

$$X_0 = A, \tag{6}$$

has been chosen as an initial matrix.

Noting that the accelerated Newton's method (ANM) is defined by

$$\begin{cases} X_0 = A, \\ \mu_k = \sqrt{\frac{\|X_k^{-1}\|}{\|X_k\|}}, \\ X_{k+1} = \frac{1}{2}\left(\mu_k X_k + \mu_k^{-1} X_k^{-1}\right), \end{cases} \tag{7}$$

which is accelerated via norm scaling to push the iterations towards convergence more rapidly at the cost of computing two matrix $l_2$-norms.

Although iteration (5) is quite efficient, several authors tried to improve it in terms of convergence acceleration and scaling. To this target, a general family of matrix iterative methods for finding $S$ was discussed in Kenney and Laub (1995), Gomilko et al. (2012) using the Padé approximants to

$$f(\xi) = (1-\xi)^{-1/2}. \tag{8}$$

Here consider that the $(m, n)$-Padé approximant to $f(\xi)$ is given by

$$\frac{P_{m,n}(\xi)}{Q_{m,n}(\xi)}, \tag{9}$$

where $m + n \geq 1$. Then, the following general iterative expression

$$x_{k+1} = \frac{x_k P_{m,n}(1 - x_k^2)}{Q_{m,n}(1 - x_k^2)} := \varphi_{2m+1,2n}, \tag{10}$$

has been proved to be convergent to 1 and $-1$ with convergence speed $m + n + 1$ for any $m \geq n - 1$.

The interesting point is that several known matrix schemes for computing $S$, such as Newton–Schultz iteration (NSM)

$$X_{k+1} = \frac{1}{2} X_k \left(3I - X_k^2\right), \tag{11}$$

and Halley's method (HM)

$$X_{k+1} = [I + 3X_k^2][X_k(3I + X_k^2)]^{-1}, \tag{12}$$

are all members of the Padé family or its reciprocal.

Similar to (11) and (12), 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:

$$X_{k+1} = -(-5I - 45X_k^2 - 15X_k^4 + X_k^6)[8X_k(3I + 5X_k^2)]^{-1}, \tag{13}$$
$$X_{k+1} = (8I + 56X_k^2)[X_k(35I + 35X_k^2 - 7X_k^4 + X_k^6)]^{-1}, \tag{14}$$
$$X_{k+1} = (35I + 140X_k^2 - 70X_k^4 + 28X_k^6 - 5X_k^8)[128X_k]^{-1}. \tag{15}$$

For obtaining more background about such matrix iterations, one may refer to Sharifi et al. (2015) and the references cited therein.

## 3 New iterative methods: derivation

The derivation of matrix iterations for finding $S$ has a direct link to the construction and application of new and useful iterative methods for solving nonlinear algebraic equations. Let us consider the following nonlinear equation:

$$f(x) = 0, \tag{16}$$

wherein $f : D \subseteq \mathbb{R} \to \mathbb{R}$ is a scalar function. Authors in Cordero et al. (2017) proved that the following member of the Chebyshev–Halley method has third order of convergence with global convergence behavior when it is extended for finding matrix sign functions

$$x_{k+1} = x_k - \left(1 + \frac{1}{2}\left(\frac{\mathfrak{L}(x_k)}{1 + 2\mathfrak{L}(x_k)}\right)\right)\frac{f(x_k)}{f'(x_k)}, \qquad k = 0, 1, 2, \ldots, \tag{17}$$

wherein

$$\mathfrak{L}(x_k) = \frac{f''(x_k) f(x_k)}{f'(x_k)^2}. \tag{18}$$

Noting that $f'(x_k)$ and $f''(x_k)$ stand for the first and second derivative of the function, to start contributing, we first should propose new iterative methods for solving (16). The new methods should not aim at having the *optimal convergence order* [in the sense of Kung–Traub, see, e.g., Soleymani et al. 2011] since such schemes lose the global convergence behavior after extension for finding $S$. Thus, the idea is to impose several steps, at which the denominator of each added step is approximated linear via a secant-like approximation,

(see, for more Traub 1964). In fact, we should design the base method so as to obtain a new method for matrix sign with global convergence behavior, because of this we applied divided difference operators in the next sub-steps.

Thus, we may write

$$\begin{cases} y_k = x_k - \left(1 + \frac{1}{2}\left(\frac{\mathfrak{L}(x_k)}{1+2\mathfrak{L}(x_k)}\right)\right)\frac{f(x_k)}{f'(x_k)}, \\ z_k = y_k - \frac{f(y_k)}{f[y_k,x_k]}, \\ x_{k+1} = z_k - \frac{f(z_k)}{f[z_k,x_k]}, \end{cases} \tag{19}$$

at which

$$f[\imath, \jmath] := \frac{f(\jmath) - f(\imath)}{\jmath - \imath}. \tag{20}$$

Similarly, we can propose the following iterative method for (16):

$$\begin{cases} y_k = x_k - \left(1 + \frac{1}{2}\left(\frac{\mathfrak{L}(x_k)}{1+2\mathfrak{L}(x_k)}\right)\right)\frac{f(x_k)}{f'(x_k)}, \\ z_k = y_k - \frac{f(y_k)}{f[y_k,x_k]}, \\ x_{k+1} = z_k - \frac{f(z_k)}{f[z_k,y_k]}. \end{cases} \tag{21}$$

**Theorem 3.1** *Assume that $f(x)$ is sufficiently smooth in a neighborhood of its simple zero $\alpha$. If an initial guess $x_0$ is close enough to $\alpha$, then, convergence order of (19) is at least five.*

***Proof*** We write the Taylor expansion of the function $f$ and its derivatives about simple zero $\alpha$ in the $k$-th iteration. Note that for simplicity we assume $c_n = \left(\frac{1}{n!}\right)\frac{f^{(n)}(\alpha)}{f'(\alpha)}, n \geq 2$. Also consider $e_k = x_k - \alpha$. Therefore, $f(x_k) = f'(\alpha)[e_k + c_2e_k^2 + c_3e_k^3 + c_4e_k^4 + c_5e_k^5 + O(e_k^6)]$. Furthermore, we obtain $f'(x_k) = f'(\alpha)[1 + 2c_2e_k + 3c_3e_k^2 + 4c_4e_k^3 + 5c_5e_k^4 + O(e_k^5)]$. Dividing these two on each other gives us $\frac{f(x_k)}{f'(x_k)} = e_k - c_2e_k^2 + 2(c_2^2 - c_3)e_k^3 + (7c_2c_3 - 4c_2^3 - 3c_4)e_k^4 + O(e_k^5)$. Now, we have the following error equation at the end of the first step:

$$y_k - \alpha = \left(6c_2^2 - c_3\right)e_k^3 + \left(-53c_2^3 + 36c_3c_2 - 3c_4\right)e_k^4 + O(e_k^5). \tag{22}$$

Using (22), it is possible to obtain

$$z_k - \alpha = (6c_2^3 - c_2c_3)e_k^4 + O(e_k^5). \tag{23}$$

And now by similar Taylor expanding, (23), the structure of the last step in (19) and extensive simplifications, we obtain the final error equation as follows:

$$e_{k+1} = \left(6c_2^4 - c_2^2c_3\right)e_k^5 + O\left(e_k^6\right). \tag{24}$$

This finishes the proof. □

**Theorem 3.2** *Assume that $f(x)$ is sufficiently smooth in a neighborhood of its simple zero $\alpha$. If an initial guess $x_0$ is close enough to $\alpha$, then convergence order of (21) is at least seven and satisfies the following error equation:*

$$e_{k+1} = \left(c_2c_3 - 6c_2^3\right)^2 e_k^7 + O\left(e_k^8\right). \tag{25}$$

***Proof*** The steps of proving the convergence order for this iterative method is similar to the proof in Theorem 3.1. It is hence omitted. □

Noting that by considering more than three steps and similar Secant-like approximations in the denominators of the new added steps, it is possible to propose general methods of various order for nonlinear equations and then for finding $S$.

Here, we consider solving the following nonlinear matrix equation in order to propose efficient matrix iterations for matrix sign functions

$$X^2 = I, \tag{26}$$

where $I$ is the identity matrix.

Applying (19) to (26), we obtain (in the reciprocal form):

$$X_{k+1} = \left(18X_k - 20X_k^3 - 30X_k^5\right)\left[5I + 15X_k^2 - 45X_k^4 - 7X_k^6\right]^{-1}. \tag{27}$$

Further simplifying results to

$$X_{k+1} = X_k\left(18I - 20X_k^2 - 30X_k^4\right)\left[5I + 15X_k^2 - 45X_k^4 - 7X_k^6\right]^{-1}, \tag{28}$$

which requires five matrix products and one matrix inverse to have a high convergence speed five. Moreover, by applying (21) to (26) and some simplifications, we obtain

$$\begin{aligned} X_{k+1} =& X_k\left(105I - 252X_k^2 - 210X_k^4 + 564X_k^6 + 49X_k^8\right) \\ &\times\left[25I + 21(4X_k^2 - 26X_k^4 + 20X_k^6 + 13X_k^8)\right]^{-1}. \end{aligned} \tag{29}$$

The novel method (29) has six matrix products and one matrix inverse to have a high convergence speed seven. Note that $X_k$ $(k \geq 0)$ are rational functions of $A$ and hence, like $A$, commute with $S$. These high order matrix iterations for computing $S$ are versatile ways of solving Riccati equations Lancaster and Rodman (1995, chapter 22).

It is quite obvious that the sign matrix may be used to determine the number of eigenvalues of a given matrix $A$ to the right or left of any straight line $x = \mathfrak{a}$, $(\mathfrak{a} \in \mathbb{R})$ in the complex $(x, y)$ plane (Howland 1983). To be more precise, the above iterations may be used to determine the number of eigenvalues inside a vertical strip bounded by the lines $x = \mathfrak{b}$ and $x = \mathfrak{c}$ with $\mathfrak{b}, \mathfrak{c} \in \mathbb{R}$ and $\mathfrak{b} < \mathfrak{c}$, provided that no eigenvalues of $A$ lie on these lines.

The eminence of the usefulness of the new methods (28) and (29) is that beside not being included as a member of the Padé family of matrix iteration (10), they should have global convergence behavior when applied to solve (26). Accordingly, it is needed to observe the dynamical behavior of these iterative methods applied to the scalar equation

$$\mathfrak{f}(x) := x^2 - 1 = 0. \tag{30}$$

To draw the basins of attractions, we consider a square $D = [-2, 2] \times [-2, 2] \in \mathbb{C}$ and assign a color to each point $x_0 \in \mathbb{D}$ according to the simple zero, at which the new methods (or the existing methods for comparisons) converge. Subsequently, we mark the point as black if the method does not converge. Here, we take into account the stopping criterion for convergence to be $|\mathfrak{f}(x_k)| \leq 10^{-2}$. The highest possible cycles that the schemes could reach is set to 120 and the programs were written in a computer algebra system Mathematica 11.0, (see for more Trott 2006).

In this way, we observe that if the convergence is not global when a point belonging to one side of the complex plane gets the color associated to the root which is located on the other side of the complex plane.

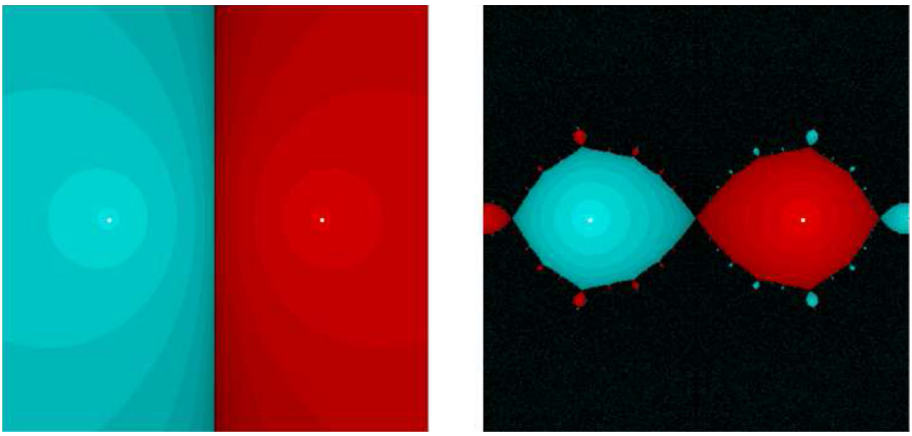

**Fig. 1** Attraction basins for (5) in left, and (11) in right, which are shaded according to the number of iterations

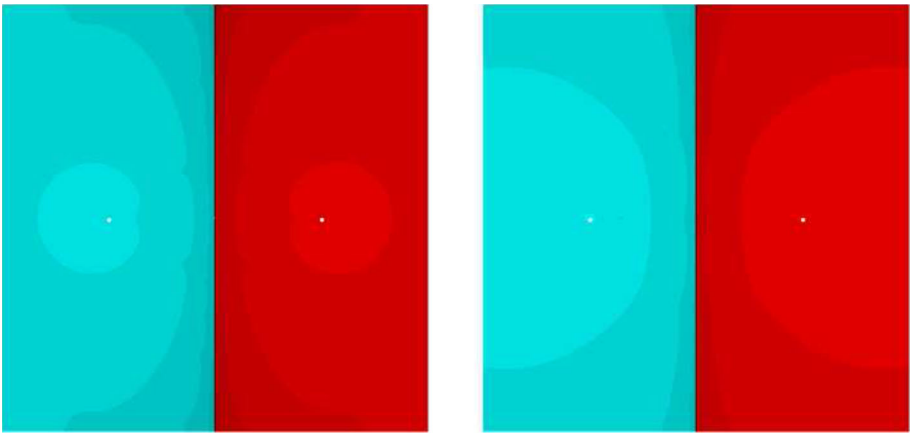

**Fig. 2** Attraction basins for (28) in left, and (29) in right, which are shaded according to the number of iterations

The results for different schemes are given in Figs. 1, 2, which indicate that the new schemes have global convergence behavior. As can be seen in Fig. 1-right, the Newton–Schulz method has local convergence behaviour and the initial matrix $X_0$ should be chosen so sharply to ensure the convergence. The black areas show that if an eigenvalue be in these domains, then the scheme would diverge. So, the Newton–Schulz method is fruitful whenever the computation matrix inverse are high and we rely on methods only based on matrix matrix multiplications.

To observe the dynamical behavior of the new methods in the larger domain $D = [-10, 10] \times [-10, 10] \in \mathbb{C}$, Fig. 3 has been provided which shows larger domain of convergence for the new methods, specially for (29), which has seventh-order convergence.

Now, it is shown that the proposed schemes are convergent, under standard conditions, viz, when there are no pure imaginary eigenvalues in the absence of the round-off errors.

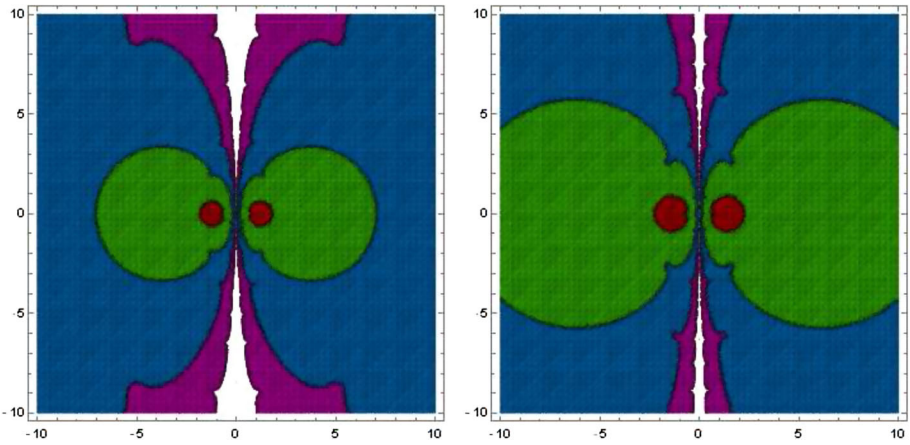


**Fig. 3** The 2D density plot of the basins of attractions (28) in left, and (29) in right, which are shaded according to the number of iterations

**Theorem 3.3** *Assume that the square matrix $A \in \mathbb{C}^{n\times n}$ have no pure imaginary eigenvalues. The matrix iterations produced as $\{X_k\}_{k=0}^{\infty}$ via (28) converges to the matrix sign S, choosing $X_0 = A$.*

***Proof*** Let $R$ be rational operator associated with (28). As any complex matrix $X \in \mathbb{C}^{n\times n}$ has a Jordan canonical form, there exists a matrix $Z$ such that $X = ZJZ^{-1}$. Then

$$R(X) = ZR(J)Z^{-1}. \tag{31}$$

An eigenvalue $\lambda$ of $X_k$ gets mapped into the eigenvalue of $R(\lambda)$ of $X_{k+1}$ by applying the iteration matrix schemes (28). This scalar relationship between eigenvalues means that we need to look at how $R(\lambda)$ maps the complex plane into itself. It is recalled that $R$ has the feature of sign preservation, viz,

$$\text{sign}(R(x)) = \text{sign}(x) \quad \text{for all } x \in \mathbb{C}. \tag{32}$$

Moreover, it should have the global convergence, i.e., the sequence defined by $x_{k+1} = R(x_k)$ with $x_0 = x$ converges to $\text{sign}(x)$ for any $x$ not on the imaginary axis. Now assume that the square matrix $A$ have a Jordan canonical form arranged as Higham (2008, p. 107):

$$Z^{-1}AZ = \Lambda = \begin{bmatrix} C & 0 \\ 0 & N \end{bmatrix}, \tag{33}$$

where $Z$ is a nonsingular matrix and $C$, $N$ are square Jordan blocks corresponding to eigenvalues lying in $\mathbb{C}^-$ and $\mathbb{C}^+$, respectively. We 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. Using (33), we have

$$\text{sign}(A) = Z \begin{bmatrix} -I_p & 0 \\ 0 & I_{n-p} \end{bmatrix} Z^{-1}. \tag{34}$$

Taking (34) into account, it is easy to deduce

$$\begin{aligned}\operatorname{sign}(\Lambda) &= \operatorname{sign}(Z^{-1}AZ)\\ &= Z^{-1}\operatorname{sign}(A)Z = \begin{pmatrix} \operatorname{sign}(\lambda_1) & & & & & \\ & \ddots & & & & \\ & & \operatorname{sign}(\lambda_p) & & & \\ & & & \operatorname{sign}(\lambda_{p+1}) & & \\ & & & & \ddots & \\ & & & & & \operatorname{sign}(\lambda_n) \end{pmatrix}. \end{aligned} \tag{35}$$

From $D_0 = Z^{-1}AZ$, we define $D_k = Z^{-1}X_kZ$ in order to obtain a convergent sequence to $\operatorname{sign}(\Lambda)$. Then from the method (28), we simply can write that

$$D_{k+1} = D_k\left(18I - 20D_k^2 - 30D_k^4\right)\left[5I + 15D_k^2 - 45D_k^4 - 7D_k^6\right]^{-1}. \tag{36}$$

If $D_0$ is a diagonal matrix then using mathematical induction, all successive $D_k$ are diagonal as well. The other case when $D_0$ is not diagonal will be discussed later in the proof.

By re-arranging (36) as $n$ un-coupled scalar iterations to tackle (30) as comes next:

$$d_{k+1}^i = \frac{18d_k^i - 20{d_k^i}^3 - 30{d_k^i}^5}{5 + 15{d_k^i}^2 - 45{d_k^i}^4 - 7{d_k^i}^6}, \tag{37}$$

where

$$d_k^i = (D_k)_{i,i}, \qquad 1 \le i \le n. \tag{38}$$

Using (36) and (37), we should study the convergence of $\{d_k^i\}$ to $\operatorname{sign}(\lambda_i)$, for all $1 \le i \le n$. From (37) and since the eigenvalues of $A$ are not pure imaginary, we have that

$$\operatorname{sign}(\lambda_i) = s_i = \pm 1. \tag{39}$$

Thus, we attain

$$\frac{d_{k+1}^i - s_i}{d_{k+1}^i + s_i} = \left(\frac{-s_i + d_k^i}{s_i + d_k^i}\right)^5 \frac{s_i + 5d_k^i}{-5s_i + 7d_k^i}. \tag{40}$$

Note that the factor $\frac{s_i+5d_k^i}{-5s_i+7d_k^i}$, is bounded and does not affect the convergence of the scheme (28). On the other hand, due to choosing an appropriate initial matrix $X_0 = A$, and the fact that

$$\left|\frac{d_0^i - s_i}{d_0^i + s_i}\right| < 1, \tag{41}$$

we attain

$$\lim_{k\to\infty}\left|\frac{d_{k+1}^i - s_i}{d_{k+1}^i + s_i}\right| = 0, \tag{42}$$

and, therefore, $\lim_{k\to\infty}(d_k^i) = s_i = \operatorname{sign}(\lambda_i)$. Now, it could be easy to conclude that $\lim_{k\to\infty} D_k = \operatorname{sign}(\Lambda)$.

Recall that if $D_0$ is not diagonal, we should pursue the scalar relationship among the eigenvalues of the iterates for (28). As described shortly at the beginning of the proof, the

eigenvalues of $X_k$ are mapped from the iterate $k$ to the iterate $k+1$, by the following relation:

$$\lambda_{k+1}^i = \left(18\lambda_k^i - 20{\lambda_k^i}^3 - 30{\lambda_k^i}^5\right) \times \left[5 + 15{\lambda_k^i}^2 - 45{\lambda_k^i}^4 - 7{\lambda_k^i}^6\right]^{-1}, \quad 1 \le i \le n. \tag{43}$$

The relation (43) shows that the eigenvalues in the general case are convergent to $s_i = \pm 1$, that is to say

$$\lim_{k\to\infty} \left| \frac{\lambda_{k+1}^i - s_i}{\lambda_{k+1}^i + s_i} \right| = 0. \tag{44}$$

In the final stage, it would be straightforward to conclude that

$$\lim_{k\to\infty} X_k = Z\left(\lim_{k\to\infty} D_k\right) Z^{-1} = Z\,\mathrm{sign}(\Lambda)\,Z^{-1} = \mathrm{sign}(A). \tag{45}$$

This finishes the proof of convergence for (28) to calculate $S$. □

Further analysis could reveal that if

$$\Theta_k = 5I + 15X_k^2 - 45X_k^4 - 7X_k^6, \tag{46}$$

then the new scheme reads the following error inequality:

$$\|X_{k+1} - S\| \le \left(\|\Theta_k^{-1}\|\|5I + 7SX_k\|\right)\|X_k - S\|^5. \tag{47}$$

The inequality (47) shows the fifth order of convergence, as $\|\Theta_k^{-1}\|\|5I + 7SX_k\|$ is bounded. This is because in the convergence phase, viz, by choosing a suitable initial approximation, we have

$$\lim_{k\to\infty} \Theta_k = \lim_{k\to\infty} \left(5I + 15X_k^2 - 45X_k^4 - 7X_k^6\right) \simeq \lim_{k\to\infty} (5I + 15I - 45I - 7I) = -32I. \tag{48}$$

And also, we have

$$\lim_{k\to\infty} (5I + 7SX_k) \simeq \lim_{k\to\infty} (5I + 7I) = 12I. \tag{49}$$

***Remark 3.1*** If we apply a first-order error analysis, then it is straightforward to obtain the new methods that are asymptotically stable, i.e., the propagation of the errors can be controlled as follows:

$$\|\Delta_{k+1}\| \le \frac{1}{2}\|\Delta_0 - S\Delta_0 S\|, \tag{50}$$

wherein $\Delta_k$ is a numerical perturbation introduced at the $k$th iterate, and $\widetilde{X}_k = X_k + \Delta_k$.

**Theorem 3.4** *Assume that the square matrix $A \in \mathbb{C}^{n\times n}$ has no pure imaginary eigenvalues. The matrix iterations produced as $\{X_k\}_{k=0}^{\infty}$ via (29) converges to the matrix sign $S$, choosing $X_0 = A$ and its convergence speed is at least seven.*

***Proof*** The proof of this theorem is similar to the proof of Theorem 3.3. It is hence omitted. □

It is requisite to notice that we can accelerate the performance of the new schemes by applying the same strategy as in (7). But since the computation of the scaling parameter $\mu_k$ is occasionally costly, we do not study it deeply for our matrix iterations.

**Table 1** Comparison of number of iterations for Experiment 4.1

| Matrix no. | NM | ANM | HM | PM1 | PM2 |
|---|---|---|---|---|---|
| $A_{50\times 50}$ | 13 | 10 | 8 | 7 | 6 |
| $A_{100\times 100}$ | 14 | 12 | 9 | 7 | 6 |
| $A_{150\times 150}$ | 14 | 13 | 9 | 7 | 6 |
| $A_{200\times 200}$ | 14 | 14 | 9 | 7 | 6 |
| $A_{250\times 250}$ | 25 | 23 | 16 | 12 | 8 |
| $A_{300\times 300}$ | 16 | 17 | 10 | 8 | 7 |
| $A_{350\times 350}$ | 18 | 20 | 12 | 9 | 8 |
| $A_{400\times 400}$ | 17 | 17 | 11 | 9 | 8 |
| $A_{450\times 450}$ | 18 | 18 | 12 | 9 | 8 |
| $A_{500\times 500}$ | 20 | 20 | 13 | 10 | 7 |
| $A_{550\times 550}$ | 18 | 16 | 11 | 9 | 8 |
| $A_{600\times 600}$ | 20 | 23 | 13 | 10 | 10 |
| $A_{650\times 650}$ | 19 | 20 | 12 | 10 | 8 |
| $A_{700\times 700}$ | 18 | 18 | 12 | 9 | 8 |
| $A_{750\times 750}$ | 20 | 22 | 13 | 10 | 8 |
| $A_{800\times 800}$ | 23 | 24 | 15 | 12 | 11 |
| $A_{850\times 850}$ | 18 | 19 | 11 | 9 | 9 |
| $A_{900\times 900}$ | 21 | 21 | 14 | 11 | 8 |
| $A_{950\times 950}$ | 22 | 24 | 14 | 11 | 9 |
| $A_{1000\times 1000}$ | 18 | 19 | 12 | 9 | 9 |
| Mean | 18.3 | 18.5 | 11.8 | 9.25 | 7.9 |

## 4 Application

Herein, several experiments are discussed for the computation of matrix sign function. The direct application of the new formulas for finding $S$ is given below, though the application for computing the polar decomposition, finding the Yang–Baxter matrix equation can be given similarly. The simulations are run on an office laptop with Windows 7 Ultimate equipped Intel(R) Core(TM) i5-2430M CPU 315 2.40 GHz processor and 16.00 GB of RAM on a 64-bit operating system. In addition, the simulations are done in 316 Mathematica 11.0 (Wellin et al. 2005).

Different methods are compared in terms of number of iterations and the computational CPU time. We only apply methods with global convergence behavior for comparison. The compared schemes are NM, HM, ANM, PM1 [i.e., (28)] and PM2 [i.e., (29)]. We do not include comparisons with methods having local convergence behavior such as the Newton–Schulz method (11) or the methods from different categories such as the ones based on the computation of the Cauchy integral (2).

The stopping criterion for our simulations is defined by

$$\|X_k^2 - I\|_2 \leq 10^{-4}. \tag{51}$$

**Experiment 4.1** *Here, we calculate the matrix sign function of the following 20 randomly generated* complex *matrices (with uniform distributions via the following piece of codes in the Mathematica environment):*

```
SeedRandom[789]; number = 20;
```

**Table 2** Comparison of the elapsed time for Experiment 4.1

| Matrix no. | NM | ANM | HM | PM1 | PM2 |
|---|---|---|---|---|---|
| $A_{50\times 50}$ | 0.013 | 0.019 | 0.008 | 0.012 | 0.012 |
| $A_{100\times 100}$ | 0.111 | 0.139 | 0.047 | 0.045 | 0.077 |
| $A_{150\times 150}$ | 0.178 | 0.377 | 0.146 | 0.133 | 0.129 |
| $A_{200\times 200}$ | 0.294 | 0.709 | 0.232 | 0.242 | 0.225 |
| $A_{250\times 250}$ | 0.867 | 1.794 | 0.690 | 0.684 | 0.533 |
| $A_{300\times 300}$ | 0.876 | 1.978 | 0.719 | 0.790 | 0.798 |
| $A_{350\times 350}$ | 1.429 | 3.324 | 1.201 | 1.362 | 1.305 |
| $A_{400\times 400}$ | 1.959 | 3.973 | 1.629 | 1.936 | 1.853 |
| $A_{450\times 450}$ | 2.836 | 5.801 | 2.415 | 2.848 | 2.799 |
| $A_{500\times 500}$ | 4.299 | 8.826 | 3.557 | 3.801 | 3.630 |
| $A_{550\times 550}$ | 5.271 | 9.236 | 3.959 | 4.993 | 4.534 |
| $A_{600\times 600}$ | 7.290 | 17.192 | 6.435 | 7.225 | 6.801 |
| $A_{650\times 650}$ | 8.579 | 19.139 | 7.310 | 8.519 | 7.056 |
| $A_{700\times 700}$ | 10.277 | 21.446 | 9.176 | 8.767 | 8.644 |
| $A_{750\times 750}$ | 13.214 | 32.084 | 12.204 | 11.898 | 10.452 |
| $A_{800\times 800}$ | 18.774 | 42.910 | 16.690 | 16.776 | 20.270 |
| $A_{850\times 850}$ | 17.347 | 40.220 | 14.931 | 15.345 | 17.780 |
| $A_{900\times 900}$ | 24.301 | 51.941 | 21.102 | 21.953 | 19.188 |
| $A_{950\times 950}$ | 29.330 | 70.299 | 24.985 | 25.427 | 23.819 |
| $A_{1000\times 1000}$ | 28.182 | 64.479 | 25.208 | 24.758 | 22.159 |
| Mean | 8.771 | 19.794 | 7.632 | 7.876 | 7.603 |

```
Table[A[l] = RandomComplex[{-5 - 5 I, 5 + 5 I},
                           {50 l, 50 l}];, {l,number}];
```

*noting that here* $I = \sqrt{-1}$.

The computational results are furnished in Tables 1, 2 for various sizes of the input matrices based on the required number od steps and the elapsed CPU times. The results uphold the theoretical aspects and discussions of Sects. 2, 3. They show that there is a reduction in the number of iterations and computational time using (28) and (29) are the best methods in terms of computational times. As a matter of fact, the mean of number of iterations and the CPU times listed in the last rows of each table indicate that the scheme (29) has the best performance in general.

It is pointed out that the calculation of $X_k^2$ per cycle for calculating the stopping condition adds one matrix product for NM, while the HM and the proposed methods form this matrix during the process of each step.

Technically speaking, the complexity analysis of different methods relies on the relations among their orders of convergence and their required matrix–matrix multiplications and the inverses per cycle. An important issue here, which makes the development of higher order methods such as PM1 and PM2 necessary, is because of incorporating the safe stopping termination (51), which is based on Higham (2008, chapter 5). In fact, the computation of one $l_2$ norm per cycle makes the process more cumbersome and since methods of higher orders, requiring fewer number of iterates, yield better performances for finding the sign function of a matrix.

**Table 3** Comparison of number of iterations for Experiment 4.1 using a different random matrices

| Matrix no. | NM | ANM | HM | PM1 | PM2 |
|---|---|---|---|---|---|
| $A_{50\times 50}$ | 14 | 13 | 9 | 7 | 6 |
| $A_{100\times 100}$ | 15 | 14 | 10 | 8 | 7 |
| $A_{150\times 150}$ | 16 | 15 | 10 | 8 | 7 |
| $A_{200\times 200}$ | 14 | 16 | 9 | 7 | 6 |
| $A_{250\times 250}$ | 15 | 15 | 10 | 8 | 6 |
| $A_{300\times 300}$ | 17 | 15 | 11 | 9 | 7 |
| $A_{350\times 350}$ | 16 | 15 | 10 | 8 | 7 |
| $A_{400\times 400}$ | 17 | 16 | 11 | 9 | 8 |
| $A_{450\times 450}$ | 18 | 16 | 11 | 9 | 7 |
| $A_{500\times 500}$ | 18 | 21 | 12 | 9 | 8 |
| Mean | 16.0 | 15.6 | 10.3 | 8.2 | 6.9 |

**Table 4** Time comparisons for Experiment 4.1 using a different random matrices

| Matrix no. | NM | ANM | HM | PM1 | PM2 |
|---|---|---|---|---|---|
| $A_{50\times 50}$ | 0.015 | 0.025 | 0.012 | 0.012 | 0.012 |
| $A_{100\times 100}$ | 0.068 | 0.126 | 0.066 | 0.059 | 0.058 |
| $A_{150\times 150}$ | 0.244 | 0.496 | 0.210 | 0.180 | 0.176 |
| $A_{200\times 200}$ | 0.370 | 0.939 | 0.361 | 0.312 | 0.287 |
| $A_{250\times 250}$ | 0.658 | 1.699 | 0.552 | 0.586 | 0.557 |
| $A_{300\times 300}$ | 1.510 | 2.405 | 0.997 | 1.151 | 1.167 |
| $A_{350\times 350}$ | 1.630 | 3.621 | 1.557 | 1.633 | 1.419 |
| $A_{400\times 400}$ | 2.773 | 5.326 | 2.163 | 2.636 | 2.613 |
| $A_{450\times 450}$ | 3.834 | 7.625 | 3.271 | 3.510 | 3.090 |
| $A_{500\times 500}$ | 5.607 | 13.587 | 4.975 | 4.552 | 4.543 |
| Mean | 1.671 | 3.585 | 1.416 | 1.463 | 1.392 |

We also repeat the process of finding matrix sign function for the first ten random matrices generated in Experiment 4.1, but with a different seed defined by `SeedRandom[345]`. The main aim is to check the results for different random matrices as shown in Tables 3, 4. Numerical evidences show a good agreement and efficacy for our proposed iteration to calculate the sign of a matrix function iteratively.

We also provide a Mathematica implementation of the new scheme in the Appendix so as to ease up its understanding and implementation.

## 5 Final remarks

In various fields of numerical linear algebra, and scientific computing, the theory and computation of matrix functions are very much useful. In the modern computational algebra, it underlies an efficient technology enabling one to resolve the topical problems of the control theory. The function of a matrix can be defined in several ways, of which the following three are generally the most useful: Jordan canonical form, polynomial interpolation, and finally Cauchy integral.

In this research work, we focus on iterative methods for this purpose. Hence, two high-order nonlinear equation solvers have been employed for constructing new methods for computing the sign of a matrix which does not have pure imaginary eigenvalues.

It was shown that the convergence is global via attraction basins in the complex plane and the rates of convergence are fifth and seventh. Finally, some applications of the new schemes via numerical experiments in double precision arithmetic were performed to manifest the superiority and applicability of the novel schemes. Outlines for future works can be forced to extend the discussed matrix iterations for calculating polar decompositions in future studies based on the application of the new schemes.

**Acknowledgements** The authors are thankful to two anonymous referees for several corrections and comments, which directly contribute to the readability of this paper.

## Appendix

Given a square nonsingular matrix `A[1]`, our proposed iteration could be coded as follows:

```
tolerance = 10^-4; k = 0; max = 50;
{n, n} = Dimensions[A[1]];
Id = SparseArray[{i_, i_} -> 1., {n, n}];

U = A[1]; U2 = U.U;
R[0] = Norm[U2 - Id, 2];
Quiet@While[k < max && R[k] >= tolerance,
    {
     U4 = U2.U2;
     U6 = U4.U2;
     U8 = U6.U2;
     UU = Inverse[25 Id + 21 (4 U2 - 26 U4 + 20 U6 + 13 U8)];
     U = ( U.(105 Id - 252 U2 - 210 U4 + 564 U6 + 49 U8)).UU;
     U2 = SparseArray@Chop[U.U, 10^-6];
     R[k + 1] = Norm[U2 - Id, 2];
     k++;
     }
    ]; // AbsoluteTiming
k
U2 // MatrixPlot
```

## References

Cordero A, Soleymani F, Torregrosa JR, Zaka Ullah M (2017) Numerically stable improved Chebyshev–Halley type schemes for matrix sign function. J Comput Appl Math 318:189–198

Cordero A, Soleymani F, Torregrosa JR, Zaka Ullah M (2016) Approximating the matrix sign function by means of Chebyshev–Halley type method. In: proceedings of the 16th international conference on computational and mathematical methods in science and engineering. pp 396–405

Company R, Egorova V, Jódar L, Soleymani F (2016) Computing prices for the American option problems with multi assets. Model Eng Human Behav XVIII:1–6

Filbir FD (1994) Computation of the structured stability radius via matrix sign function. Syst Control Lett 22:341–349

Gomilko O, Greco F, Ziętak K (2012) A Padé family of iterations for the matrix sign function and related problems. Numer Linear Algebra Appl 19:585–605
Higham NJ (1994) The matrix sign decomposition and its relation to the polar decomposition. Linear Algebra Appl 212(213):3–20
Higham NJ, Mackey DS, Mackey N, Tisseur F (2004) Computing the polar decomposition and the matrix sign decomposition in matrix groups. SIAM J Matrix Anal Appl 25:1178–1192
Higham NJ (2008) Functions of matrices: theory and computation. Society for Industrial and Applied Mathematics, Philadelphia
Howland JL (1983) The sign matrix and the separation of matrix eigenvalues. Linear Algebra Appl 49:221–232
Iannazzo B (2007) Numerical solution of certain nonlinear matrix equations, Ph.D. thesis, Dipartimento di Matematica. Università di Pisa, Pisa
Kenney C, Laub AJ (1991) Rational iterative methods for the matrix sign function. SIAM J Matrix Anal Appl 12:273–291
Kenney CS, Laub AJ (1995) The matrix sign function. IEEE Trans Autom Control 40:1330–1348
Lancaster P, Rodman L (1995) Algebraic Riccati equations. Oxford University Press, Oxford
Misrikhanov MSh, Ryabchenko VN (2008) Matrix sign function in the problems of analysis and design of the linear systems. Autom Remote Control 69:198–222
Roberts JD (1980) Linear model reduction and solution of the algebraic Riccati equation by use of the sign function. Int J Control 32:677–687
Sharifi M, Karimi Vanani S, Khaksar Haghani F, Arab M, Shateyi S (2015) On a cubically convergent iterative method for matrix sign. Sci World J 964257:6
Shieh LS, Tsay YT, Wang CT (1984) Matrix sector functions and their applications to system theory. IEEE Proc 131:171–181
Soleymani F, Karimi Vanani S, Afghani A (2011) A general three-step class of optimal iterations for nonlinear equations. Math Prob Eng 2011(469512):10
Soleymani F, Stanimirović PS, Shateyi S, Haghani FK (2014) Approximating the matrix sign function using a novel iterative method. Abstr Appl Anal 2014(105301):9
Soleymani F, Zaka Ullah M (2018) A multiquadric RBF–FD scheme for simulating the financial HHW equation utilizing exponential integrator. Calcolo 55(51):1–26
Soleymani F, Barfeie M, Khaksar Haghani F (2018) Inverse multi-quadric RBF for computing the weights of FD method: application to American options. Commun Nonlinear Sci Numer Simul 64:74–88
Traub JF (1964) Iterative methods for the solution of equations. Prentice Hall, New York
Trott M (2006) The mathematica guidebook for numerics. Springer, New York
Wellin PR, Gaylord RJ, Kamin SN (2005) An introduction to programming with mathematica. Cambridge University Press, Cambridge