Cordero Barbero, A.; Soleymani, F.; Torregrosa Sánchez, JR.; Ullah, MZ. (2017). Numerically stable improved Chebyshev-Halley type schemes for matrix sign function. Journal of Computational and Applied Mathematics. 318:189-198. doi:10.1016/j.cam.2016.10.025

Numerically stable improved Chebyshev-Halley type schemes for matrix sign function ☆

Alicia Cordero[a], F. Soleymani[a], Juan R. Torregrosa[a,*], M. Zaka Ullah[b]

[a]*Instituto de Matemática Multidisciplinar, Universitat Politècnica de València*
Camino de Vera s/n, 46022 València, Spain
[b]*Department of Mathematics, King Abdulaziz University, Jeddah 21589, Saudi Arabia*

Abstract

A general family of iterative methods including a free parameter is derived and proved to be convergent for computing matrix sign function, under some restrictions on the parameter. Several special cases including global convergence behavior are dealt with. It is analytically shown that they are asymptotically stable. A variety of numerical experiments for matrices with different sizes are considered to put on show the effectiveness of the proposed members of the family.

Keywords: Matrix sign function; stability; iterative methods; Chebyshev-Halley family; eigenvalues.

1. Motivation

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

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

An extension of this function for the matrix case was given firstly by Roberts in [19], who introduced the matrix sign function as a tool for model reduction and for solving Lyapunov and algebraic Riccati equations.

The problem of computing a function of a matrix, named by $f(A)$, is of growing significance, though as yet numerical methods are developed for this purpose. In between, matrix sign function is undoubtedly of crystal clear importance in the theory and application of matrix functions (e.g. one may refer to [3, 6, 21]). The matrix sign function has basic theoretical and algorithmic relations with the matrix square root, the polar decomposition and with the matrix pth roots (see for more [11, 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 is requisite.

The matrix sign function is a valuable tool for the numerical solution of Sylvester and Liapunov matrix equations [1]. A generalization of the Newton iteration for the matrix sign function to the solution of the generalized algebraic Bernoulli equations was presented in [2]. This matrix function is used in [18] as a simple and direct method to derive some fundamental results in the theory of surface waves in anisotropic materials. For other applications of matrix sign function, we refer the reader to [17]. Due to the applicability of the matrix sign function, stable iterative schemes have become some viable choices for approximating this matrix.

Here we suppose that matrix $A \in \mathbb{C}^{n\times n}$ has no eigenvalues on the imaginary axis. To define this matrix function formally, let $A = PJP^{-1}$ be the Jordan canonical form arranged so that $J = diag(J_1, J_2)$, where the eigenvalues of

☆This research was supported by Ministerio de Economía y Competitividad MTM2014-52016-C2-2-P.
*Corresponding author
Email addresses: acordero@mat.upv.es (Alicia Cordero), fazlollah.soleymani@gmail.com. (F. Soleymani), jrtorre@mat.upv.es (Juan R. Torregrosa), mzhussain@kau.edu.sa (M. Zaka Ullah)

$J_1 \in \mathbb{C}^{p\times p}$ lie in the open left half-plane and those of $J_2 \in \mathbb{C}^{n-p\times n-p}$ lie in the open right-plane, then

$$S = sign(A) = P \begin{pmatrix} -I_p & 0 \\ 0 & I_{n-p} \end{pmatrix} P^{-1}.$$

This matrix function can be uniquely defined (A is a nonsingular square matrix). The most concise definition of the matrix sign decomposition is given in [9, 14] as follows:

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

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 we have said, this matrix sign function was introduced by Roberts in 1971. It is important to note that the matrix disk function was introduced in the same paper by Roberts [19]. 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.

This matrix function has several properties. Some of them are given by [14]:

1. $S^2 = I$ (S is involutory).
2. S is diagonalizable with eigenvalues ± 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.

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 [11, p. 16].

A number of matrix functions $f(A)$ are amenable to computation by iteration functions of the following form [11, p. 91]:

$$X_{k+1} = g(X_k), \tag{2}$$

where for the iterations used in practice, X_0 is not arbitrary but is a fixed function of A. Taking into account the computational burden makes it obvious that g is a polynomial or rational function. Rational g 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 most common and well-known way for finding the sign of a square nonsingular matrix is the following numerical method

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

which is also known as Newton's method (NM) and converges quadratically when $X_0 = A$ has been chosen as an initial matrix.

Although iteration (3) 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 [15] using the Padé approximants to

$$f(\xi) = (1 - \xi)^{-1/2}. \tag{4}$$

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

$$\frac{P_{m,n}(\xi)}{Q_{m,n}(\xi)}, \tag{5}$$

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

$$x_{k+1} = \frac{x_k P_{m,n}(1-x_k^2)}{Q_{m,n}(1-x_k^2)} := \varphi_{2m+1,2n}, \tag{6}$$

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)

$$X_{k+1} = \frac{1}{2} X_k \left(3I - X_k^2\right), \tag{7}$$

and Halley's method (HM)

$$X_{k+1} = [I + 3X_k^2][X_k(3I + X_k^2)]^{-1}, \tag{8}$$

are all members of the Padé family or its reciprocal. Such high order schemes for computing S are versatile ways of solving Riccati equations [16, chapter 22].

It is requisite to recall that the iterative expressions in the Padé family or in the reciprocal Padé family have the minimum sum of the degrees of the numerator and the denominator among all rational iterations of a fixed order [5].

Motivated by the recent developments in this area [5, 10, 22], we here propose some variants of Chebyshev-Halley type scheme possessing a free parameter. An improvement of this family is given as our main contribution to possess high rate of convergence with global convergence behavior for some of its special members. The stability of the schemes are considered to show that the rounding errors remain under control.

The paper is divided into several sections and is organized as follows. In Section 2, a Chebyshev-Halley type family of schemes is proposed. Some discussions on several members of the family are given. Section 3 includes the analysis of convergence while Section 4 is dedicated to study the stability. In Section 5, various numerical examples are considered to confirm the theoretical results. A comparison with the existing methods is also presented therein. Concluding remarks are given in Section 6.

2. Construction of the family of schemes

Gutiérrez et al. in [7] developed a Chebyshev-Halley type family iterative methods (in Banach spaces) for finding simple zeros of the nonlinear (operator) equation $f(x) = 0$. This scheme can be written as follows:

$$x_{k+1} = x_k - \left(1 + \frac{1}{2}\frac{L(x_k)}{1 - aL(x_k)}\right)\frac{f(x_k)}{f'(x_k)}, \tag{9}$$

wherein $a \in \mathbb{R}$, $L(x_k) = \frac{f''(x_k)f(x_k)}{f'(x_k)^2}$ and the convergence order is cubic. The application of some special cases of this scheme has recently been discussed in [22]. Here, we consider the general expression (9) to solve the following nonlinear matrix equation

$$X^2 = I, \tag{10}$$

where I is the identity matrix and obtain the following iterative expression in the reciprocal form

$$X_{k+1} = \left(-4aX_k + 4(-2+a)X_k^3\right)\left[I - 3X_k^2(2I + X_k^2) + 2a(-I + X_k^4)\right]^{-1}. \tag{11}$$

The main aim and motivation in constructing iterative methods for matrix sign is to attain as fast as possible order of convergence with minimal computational costs.

The structure (11) is costly to attain matrix sign function since it requires *five* matrix matrix multiplications (mmm) and one inverse per computing cycle to reach the local order 3. To improve this family of schemes and construct an economic family of iterations, we propose the following nonlinear solver for solving (10)

$$\begin{cases} y_k = x_k - \left(1 + \frac{1}{2}\left(\frac{L(x_k)}{1 - aL(x_k)}\right)\right)\frac{f(x_k)}{f'(x_k)}, \\ x_{k+1} = y_k - \frac{f(y_k)}{f[y_k, x_k]}, \end{cases} \tag{12}$$

whereas $f[y_k, x_k] = \frac{f(y_k)-f(x_k)}{y_k-x_k}$ and attain uniquely the following new family of improved Chebyshev-Halley type iterative methods:

$$X_{k+1} = \left(X_k - 6aX_k + 2(-7+2a)X_k^3 + (-3+2a)X_k^5\right)\left[(1-2a)I - 2(3+2a)X_k^2 + (-11+6a)X_k^4\right]^{-1}. \quad (13)$$

Further simplifying results to

$$X_{k+1} = X_k\left((1-6a)I + 2(-7+2a)X_k^2 + (-3+2a)X_k^4\right)\left[(1-2a)I - 2(3+2a)X_k^2 + (-11+6a)X_k^4\right]^{-1}, \quad (14)$$

which requires *four* mmm and one matrix inverse to reach a higher rate of convergence *four* in contrast to (11). Note that, X_k ($k \geq 0$), are rational functions of A and hence, like A, commute with S.

Theorem 1. *Let $f(x)$ be a function at least three times differentiable in a neighborhood of its simple zero α. If an initial approximations x_0 is sufficiently close to α, then the convergence order of (12) is at least four, for any value of parameter a, being its error equation*

$$e_{k+1} = c_2(-2(a-1)c_2^2 - c_3)e_k^4 + O(e_k^5),$$

where $c_q = \frac{1}{q!}\frac{f^{(q)}(\alpha)}{f'(\alpha)}$, $q \geq 2$, and $e_k = x_k - \alpha$.

Proof. The proof of this theorem is based on Taylor expansion and it is similar to those given in [4]. This is hence skipped over. ∎

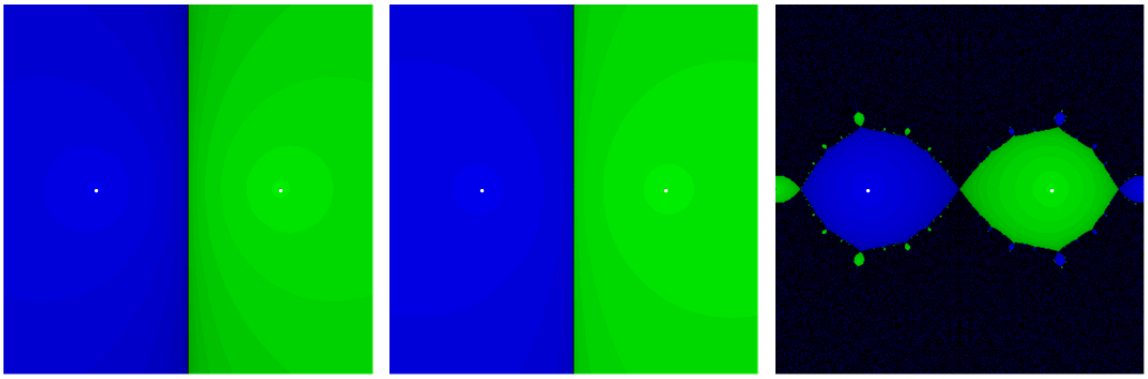
On modern computers with hierarchical memories, matrix multiplication is usually much faster than solving a matrix equation or inverting a matrix, so iterations such as (14) that are multiplication-rich, which means having a rational function g, are preferred.

In what follows, we list some special cases from the family (14).

- Choosing $a = 0$ results in (PM1): $X_{k+1} = X_k\left(-I + 14X_k^2 + 3X_k^4\right)\left[-I + 6X_k^2 + 11X_k^4\right]^{-1}$.
- Choosing $a = 1/2$ results in (PM2): $X_{k+1} = \left(I + 6X_k^2 + X_k^4\right)\left[4(X_k + X_k^3)\right]^{-1}$.
- Choosing $a = -1/2$ results in (PM3): $X_{k+1} = X_k\left(-2I + 8X_k^2 + 2X_k^4\right)\left[-I + 2X_k^2 + 7X_k^4\right]^{-1}$.
- Choosing $a = 1$ results in (PM4): $X_{k+1} = X_k\left(5I + 10X_k^2 + X_k^4\right)\left[I + 5X_k^2(2 + X_k^2)\right]^{-1}$.
- Choosing $a = -1$ results in (PM5): $X_{k+1} = X_k\left(-7I + 18X_k^2 + 5X_k^4\right)\left[-3I + 2X_k^2 + 17X_k^4\right]^{-1}$.
- Choosing $a = -2$ results in (PM6): $X_{k+1} = X_k\left(-13I + 22X_k^2 + 7X_k^4\right)\left[-5I - 2X_k^2 + 23X_k^4\right]^{-1}$.
- Choosing $a = 3/2$ results in (PM7): $X_{k+1} = X_k\left(4(I + X_k^2)\right)\left[I + 6X_k^2 + X_k^4\right]^{-1}$.
- Choosing $a = -3/2$ results in (PM8): $X_{k+1} = X_k\left(-5I + 10X_k^2 + 3X_k^4\right)\left[-2I + 10X_k^4\right]^{-1}$.
- Choosing $a = -4/5$ results in (PM9): $X_{k+1} = X_k\left(-29I + 86X_k^2 + 23X_k^4\right)\left[-13I + 14X_k^2 + 79X_k^4\right]^{-1}$.

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 $x = a$, ($a \in \mathbb{R}$) in the complex (x, y) plane [12]. To be more precise, the above iterations may be used to determine whether a matrix is stable. It is also apparent that we may easily determine the number of eigenvalues inside a vertical strip bounded by the lines $x = b$ and $x = c$ with $b, c \in \mathbb{R}$ and $b < c$, provided that no eigenvalues of A lie on these lines.

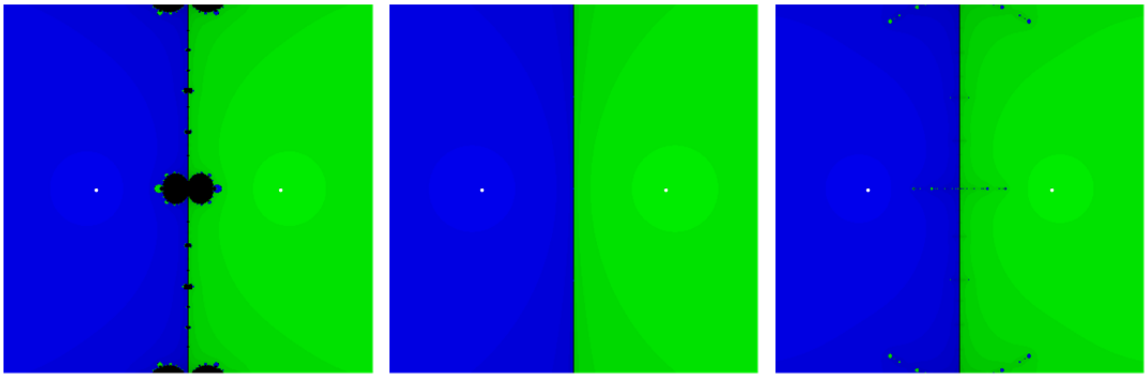
In the rest of this section it is discussed that for which values of the free parameter a, one may attain an efficient scheme for computing matrix sign function. We remark that a method for computing S must be globally convergent and it is of practical interest if it does not belong to the general Padé family of iterations (6).



(a) NM (b) HM (c) NSM

Figure 1: Basins of attraction for Newton, Halley and Newton-Schulz methods

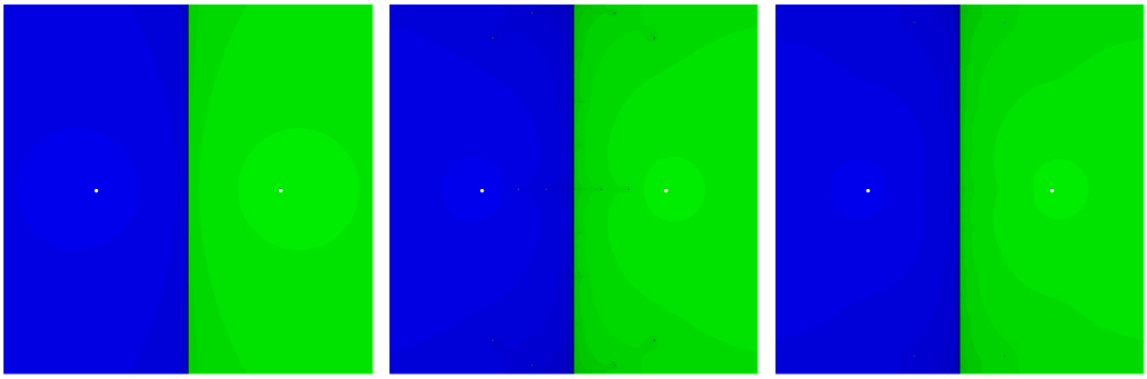
So, it must be checked that for which values of a the convergence is global. To pursue this aim, it is enough to draw the basins of attraction for the scheme (14) to solve the scalar equation $f(x) := x^2 - 1 = 0$ (for more information on pure matrix methods and their global convergence behavior one should consult the thesis [13]). We take a rectangle $\mathbb{D} = [-2, 2] \times [-2, 2] \in \mathbb{C}$ and assign a color to each point $z_0 \in \mathbb{D}$ according to the simple zero at which the scheme from (14) converges and we mark the point as black if the method does not converge. Here, we take into account the stopping criterion for convergence to be $|f(x_k)| \leq 10^{-2}$ wherein the maximum number of full cycles for each method is 200 in the written Mathematica codes [23]. Following such a procedure, we distinguish the attraction basins by their colors for different methods. To clearly illustrate the behavior of the proposed family in the complex plane, we made a short clip attached to this work as a supplementary material which shows the moving attractions basins of the proposed schemes when the free parameters changes from -2 to +2.



(a) $PM1$ (b) $PM2$ (c) $PM3$

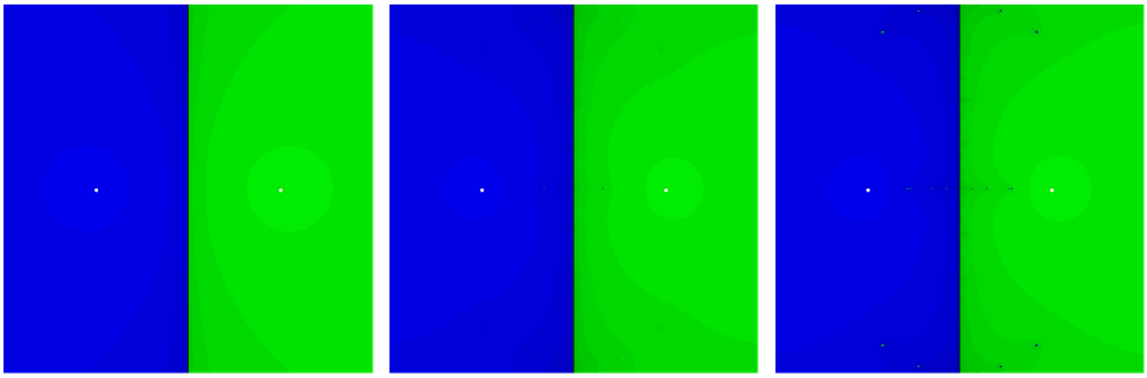
Figure 2: Basins of atraction for $a = 0$, $a = 0.5$ and $a = -0.5$,

Results of dynamical behaviors for different cases are brought forward in Figures 1-4. Checking the results and comparing by the schemes from Padé family, it is yielded that PM1, PM3, PM5, PM8 and PM9 are not of global convergence and they would not be of interest further. On the other hand, PM2, PM4 and PM7 are members from the Padé family. While this shows the generality of the proposed Chebyshev-Halley type method, we can deduce that PM6 is new with



(a) $PM4$ (b) $PM5$ (c) $PM6$

Figure 3: Basins of atraction for $a = 1$, $a = -1$ and $a = -2$,



(a) $PM7$ (b) $PM8$ (c) $PM9$

Figure 4: Basins of atraction for $a = 3/2$, $a = -3/2$ and $a = -4/5$,

global convergence.

3. Convergence analysis

In this section, it is showed that the proposed family of Chebyshev-Halley type schemes (14) is convergent, under standard conditions.

Theorem 2. *Let $A \in \mathbb{C}^{n\times n}$ have no pure imaginary eigenvalues. Then, the matrix sequence $\{X_k\}_{k=0}^{\infty}$ defined by* (14) *converges to the matrix sign S, choosing $X_0 = A$.*

Proof. Let R be the rational operator associated to (14). 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

$$R(X) = ZR(J)Z^{-1}. \tag{15}$$

An eigenvalue λ of X_k gets mapped into the eigenvalue of $R(\lambda)$ of X_{k+1} by applying the iteration matrix schemes (14). This scalar relationship between eigenvalues means that we need to look at how $R(\lambda)$ maps the complex plane into itself. Precisely speaking, rational operator R should have two properties:

(i) Sign preservation, that is, $\text{sign}(R(x)) = \text{sign}(x), \quad$ for all $x \in \mathbb{C}$, and
(ii) Global convergence, that is, 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.

To do this process formally, let A have a Jordan canonical form arranged as [11, p. 107]:

$$Z^{-1}AZ = \Lambda = \begin{bmatrix} C & 0 \\ 0 & N \end{bmatrix}, \tag{16}$$

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 (16), we have

$$\text{sign}(A) = Z \begin{bmatrix} -I_p & 0 \\ 0 & I_{n-p} \end{bmatrix} Z^{-1}. \tag{17}$$

Therefore, it is clear to write

$$\text{sign}(\Lambda) = \text{sign}(Z^{-1}AZ) = Z^{-1}\text{sign}(A)Z = \begin{pmatrix} \text{sign}(\lambda_1) & & & & & & \\ & \ddots & & & & & \\ & & \text{sign}(\lambda_p) & & & & \\ & & & \text{sign}(\lambda_{p+1}) & & & \\ & & & & \ddots & & \\ & & & & & \text{sign}(\lambda_n) & \end{pmatrix}. \tag{18}$$

From $D_0 = Z^{-1}AZ$, we define $D_k = Z^{-1}X_kZ$, $k = 1, 2, \ldots$, so as to have a sequence converging to $\text{sign}(\Lambda)$. Then, from method (14), we simply can write that

$$D_{k+1} = D_k \left((1-6a)I + 2(-7+2a)D_k^2 + (-3+2a)D_k^4\right) \left[(1-2a)I - 2(3+2a)D_k^2 + (-11+6a)D_k^4\right]^{-1}. \tag{19}$$

If D_0 is a diagonal matrix, then using mathematical induction, all successive D_k are diagonal as well. We note that the case when D_0 is not diagonal can be treated similarly and it is given later in the proof.

Let us re-write (19) in the form of n uncoupled scalar iterative methods to solve $f(x) = x^2 - 1 = 0$ as follows:

$$d_{k+1}^i = \frac{d_k^i - 6ad_k^i + 2(-7+2a){d_k^i}^3 + (-3+2a){d_k^i}^5}{1 - 2a - 2(3+2a){d_k^i}^2 + (-11+6a){d_k^i}^4}, \tag{20}$$

where $d_k^i = (D_k)_{i,i}$ and $1 \leq i \leq n$. From (19) and (20), we should study the convergence of $\{d_k^i\}$ to $\text{sign}(\lambda_i)$, for all $1 \leq i \leq n$.

From (20) and since the eigenvalues of A are not pure imaginary, we have that $\text{sign}(\lambda_i) = s_i = \pm 1$. Thus, we attain

$$\frac{d_{k+1}^i - s_i}{d_{k+1}^i + s_i} = \left(\frac{-s_i + d_k^i}{s_i + d_k^i}\right)^4 \frac{-s_i - 3d_k^i + 2a(s_i + d_k^i)}{s_i - 3d_k^i + 2a(-s_i + d_k^i)}. \tag{21}$$

It can be checked that the second factor of expression (21) is bounded for $i = 1, 2, \ldots, n$, as can be observed in Figure 5.

On the other hand, due to choosing an appropriate initial matrix $X_0 = A$, $\left|\dfrac{d_0^i - s_i}{d_0^i + s_i}\right| < 1$, we have

$$\lim_{k\to\infty} \left|\frac{d_{k+1}^i - s_i}{d_{k+1}^i + s_i}\right| = 0, \tag{22}$$

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

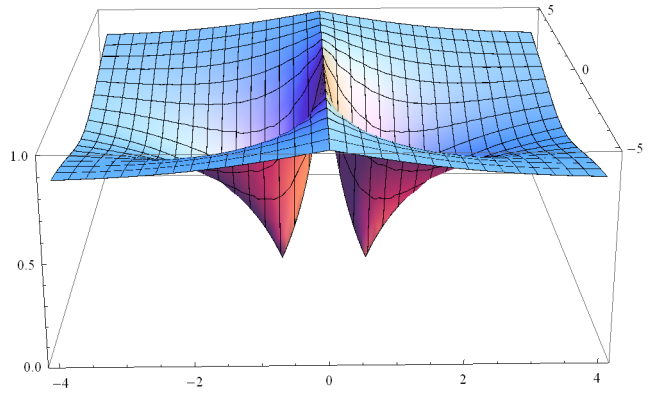


Figure 5: Bounded factor of expression (21)

Recall that if D_0 is not diagonal, we should pursue the scalar relationship among the eigenvalues of the iterates for the studied rational improved Chebyshev-Halley type matrix iteration. 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

$$\lambda_{k+1}^i = (\lambda_k^i - 6a\lambda_k^i + 2(-7+2a)\lambda_k^{i\,3} + (-3+2a)\lambda_k^{i\,5})[1-2a-2(3+2a)\lambda_k^{i\,2} + (-11+6a)\lambda_k^{i\,4}]^{-1}, \quad 1 \leq i \leq n. \tag{23}$$

Once again and in a similar methodology, (23) manifests that the eigenvalues in the general case are convergent to $s_i = \pm 1$, that is to say

$$\lim_{k\to\infty} \left| \frac{\lambda_{k+1}^i - s_i}{\lambda_{k+1}^i + s_i} \right| = 0. \tag{24}$$

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

$$\lim_{k\to\infty} X_k = Z \left(\lim_{k\to\infty} D_k \right) Z^{-1} = Z\, \mathrm{sign}(\Lambda)\, Z^{-1} = \mathrm{sign}(A). \tag{25}$$

This finishes the proof of convergence for our modification of Chebyshev-Halley type family of matrix schemes (14) for computing matrix sign function. □

Although the previous theorem discussed the convergence analysis, in what follows we study the convergence rate of (14).

Theorem 3. *Let $A \in \mathbb{C}^{n\times n}$ have no pure imaginary eigenvalues. Then, the matrix sequence $\{X_k\}_{k=0}^{\infty}$ defined by* (14) *has* at least *fourth rate of convergence to S, choosing $X_0 = A$.*

Proof. To show the convergence rate theoretically, let us use the five features of matrix sign function S stated in the motivation section and then consider

$$B_k = (1-2a)I - 2(3+2a)X_k^2 + (-11+6a)X_k^4. \tag{26}$$

We can write

$$\begin{aligned}
X_{k+1}-S &= \left(X_k-6aX_k+2(-7+2a)X_k^3+(-3+2a)X_k^5\right)B_k^{-1}-S\\
&= \left(X_k-6aX_k+2(-7+2a)X_k^3+(-3+2a)X_k^5-SB_k\right)B_k^{-1}\\
&= \Big(X_k-6aX_k+2(-7+2a)X_k^3+(-3+2a)X_k^5-(1-2a)S\\
&\qquad +2(3+2a)SX_k^2-(-11+6a)SX_k^4\Big)B_k^{-1}\\
&= S(X_k-S)^4-X_k\left((X_k-S)^4+(2-2a)^4X_k^4-4\left(\frac{6}{4}-\frac{6}{4}a\right)SX_k^3\right.\\
&\qquad \left.+6\left(\frac{4}{6}-\frac{4}{6}a\right)X_k^2-4(-1+a)SX_k+(-6+6a)I+S(2aS^2-2S)\right)B_k^{-1}\\
&= \left((X_k-S)^4(-I-3SX_k+2a(I+SX_k))\right)B_k^{-1}.
\end{aligned} \tag{27}$$

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

$$\|X_{k+1}-S\| \le \left(\|B_k^{-1}\|\|-I-3SX_k+2a(I+SX_k)\|\right)\|X_k-S\|^4. \tag{28}$$

The inequality (28) shows the fourth order of convergence, as $\|B_k^{-1}\|\|-I-3SX_k+2a(I+SX_k)\|$ is bounded. It is also evident that the choice $a=1$, makes the convergence rate to be five. The proof is now complete. □

Here it is remarked that for some special choices of the family (14) the convergence rate is higher than four. For instance, choosing $a=1$ which resulted in PM4 provides fifth order of convergence with global behavior.

4. Numerical stability issues

In this section, we study the stability of (14) for finding S in a vicinity of the solution of (10). To be more clear, we analyze how a small perturbation at kth iterate is amplified or damped along the iterates, which could be considered as asymptotical stability. In this way, it is tried to show that the perturbation errors are controllable by applying the methods with global convergence extracted from (14).

Theorem 4. *Using the same assumptions as in Theorem 3, matrix sequence* $\{X_k\}_{k=0}^{\infty}$ *generated by* (14) *is stable.*

Proof. If X_0 is a function of A, then the iterates from (14) are all functions of A and hence commute with A. To study the stability of the proposed scheme, let us assume that Δ_k is a numerical perturbation introduced at the kth iterate of (14). As a result, we may write

$$\widetilde{X}_k = X_k+\Delta_k. \tag{29}$$

To cut a long story short, we perform a first-order error analysis, say, we formally take advantage of approximations $(\Delta_k)^i \approx 0$, since $(\Delta_k)^i$, $i\ge 2$ is close to zero matrix. This consideration is meaningful when Δ_k is sufficiently small. We have

$$\begin{aligned}
\widetilde{X}_{k+1} &= \widetilde{X}_k\left((1-6a)I+2(-7+2a)\widetilde{X}_k^2+(-3+2a)\widetilde{X}_k^4\right)\left[(1-2a)I-2(3+2a)\widetilde{X}_k^2+(-11+6a)\widetilde{X}_k^4\right]^{-1}\\
&= \left((1-6a)(X_k+\Delta_k)+2(-7+2a)(X_k+\Delta_k)^3+(-3+2a)(X_k+\Delta_k)^5\right)\\
&\quad \left[(1-2a)I-2(3+2a)(X_k+\Delta_k)^2+(-11+6a)(X_k+\Delta_k)^4\right]^{-1}.
\end{aligned} \tag{30}$$

Using the following statements [8] for any nonsingular matrix B and matrix C:

$$(B+C)^{-1} \simeq B^{-1}-B^{-1}CB^{-1}, \tag{31}$$

and

$$S^2=I, \quad \text{and} \quad S^{-1}=S, \tag{32}$$

we have (assuming $X_k \simeq \text{sign}(A) = S$ for enough large k)

$$\begin{aligned}\widetilde{X}_{k+1} &\simeq (-16S - 36\Delta_k + 8a\Delta_k - 20S\Delta_k S + 8aS\Delta_k S)\\ &\quad (-16I - 28S\Delta_k - 28\Delta_k S + 8aS\Delta_k + 8a\Delta_k S)^{-1}\\ &\simeq (-16S - 36\Delta_k + 8a\Delta_k - 20S\Delta_k S + 8aS\Delta_k S)\\ &\quad \left(\frac{-1}{16}I + \frac{28}{16^2}S\Delta_k + \frac{28}{16^2}\Delta_k S - \frac{8}{16^2}aS\Delta_k - \frac{8}{16^2}a\Delta_k S\right)^{-1}\\ &\simeq \left(S + \frac{1}{2}\Delta_k - \frac{1}{2}S\Delta_k S\right).\end{aligned} \tag{33}$$

After some simplifications and using $\Delta_{k+1} = \widetilde{X}_{k+1} - X_{k+1} \simeq \widetilde{X}_{k+1} - S$, one can verify that:

$$\Delta_{k+1} \simeq \frac{1}{2}\Delta_k - \frac{1}{2}S\Delta_k S. \tag{34}$$

At this moment, we draw as a conclusion that the perturbation at the iterate $k+1$ is bounded, i.e.,

$$\|\Delta_{k+1}\| \leq \frac{1}{2}\|\Delta_0 - S\Delta_0 S\|. \tag{35}$$

Consequently, the sequence $\{X_k\}_{k=0}^{\infty}$ generated by (14) is stable. This ends the proof. □

5. Numerical experiments

This section addresses issues related to the numerical precision of the computation of matrix sign function, using Mathematica 8 built-in precision [25]. The value of machine precision that produced the results included here is 15.96 digits, which corresponds to a double precision number with a mantissa of 53 digit binary [24, chapter 8].

In this work, the computer specifications are Windows 7 Ultimate with Intel(R) Core(TM) i5-2430M CPU 2.40GHz processor and 8.00 GB of RAM on a 64-bit operating system.

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, PM4, PM7, PM6 and ANM (accelerated Newton's method) which is defined by

$$\begin{cases} X_0 = A,\\ \mu_k = \sqrt{\frac{\|X_k^{-1}\|}{\|X_k\|}},\\ X_{k+1} = \frac{1}{2}\left(\mu_k X_k + \mu_k^{-1}X_k^{-1}\right). \end{cases} \tag{36}$$

One can similarly accelerate the performance of the new schemes from the improved Chebyshev-Halley type family (14) using some strategy as in (36). But since the computation of the scaling parameter μ_k is occasionally costly, we do not study it deeply for our family of iterations. The stopping termination in this work is

$$\|X_k^2 - I\|_2 \leq 10^{-8}. \tag{37}$$

Example 1. *In this series of experiments, we compute the matrix sign function of the following 10 randomly generated matrices*

```
SeedRandom[1234]; number = 10;
Table[A[l] = RandomReal[{-100, 100}, {100 l, 100 l}];, {l, number}];
```

Matrix No.	NM	ANM	HM	PM7	PM4	PM6
$A_{100\times100}$	17	11	11	9	8	10
$A_{200\times200}$	19	14	12	10	8	12
$A_{300\times300}$	20	16	13	10	9	12
$A_{400\times400}$	24	18	15	12	11	14
$A_{500\times500}$	20	16	13	10	9	12
$A_{600\times600}$	23	21	14	12	10	14
$A_{700\times700}$	22	18	14	11	10	13
$A_{800\times800}$	23	21	15	12	10	14
$A_{900\times900}$	23	19	14	12	10	14
$A_{1000\times1000}$	23	21	15	12	10	14

Table 1: Comparison of number of iterations for Example 1.

Matrix No.	NM	ANM	HM	PM7	PM4	PM6
$A_{100\times100}$	0.0368623	0.0531857	0.0275516	0.0263713	0.0235674	0.0269054
$A_{200\times200}$	0.158083	0.273685	0.121207	0.113714	0.0977962	0.135305
$A_{300\times300}$	0.44521	0.843781	0.353374	0.315238	0.289896	0.367553
$A_{400\times400}$	1.09422	1.91654	0.88232	0.842601	0.829421	0.984137
$A_{500\times500}$	1.57968	2.92306	1.37999	1.19248	1.1876	1.45492
$A_{600\times600}$	2.97305	6.32247	2.37401	2.30263	2.16222	2.7574
$A_{700\times700}$	4.54777	8.51585	3.7262	3.3351	3.27217	3.94201
$A_{800\times800}$	7.25574	15.3073	6.00425	5.25829	4.87302	6.25886
$A_{900\times900}$	10.361	20.1839	8.11459	7.57771	6.7404	9.08985
$A_{1000\times1000}$	14.3216	31.2905	11.6792	10.3742	9.34322	12.3002

Table 2: Comparison of the elapsed time for Example 1.

The results are displayed in Tables 1-2 on random matrices of size $100i \times 100i$, $i = 1, 2, \ldots, 10$. The results are in good harmony with the theoretical aspects of Sections 2-4. They show that there is a reduction in the number of iterations and computational time using PM4, PM7 and PM6. PM4 and PM7 are the best methods in terms of computational time. Note that the computation of X_k^2 per cycle for calculating the stopping condition adds one matrix-matrix multiplication for NM, while the HM and the proposed methods form this matrix during the process of each step.

Similar numerical experiments have been carried out on variety of problems which confirm the above conclusions to a great extent. Finally, we can conclude from numerical experiments that new proposed schemes confirm the theoretical results and show consistent convergence behavior.

6. Summary

A matrix function can be defined and computed in several ways, such as Cauchy integral, polynomial interpolation, and Jordan canonical form. However, a practical way in most problems is to apply iterative methods for this purpose.

Under this motivation, in this paper we have introduced and demonstrated a general modification of Chebyshev-Halley type methods possessing *at least* fourth order of convergence for finding the matrix sign function. The proposed methods consist of one matrix inversion per cycle and are asymptotically stable. It is discussed that how several new methods with global convergence behavior can be deduced from the main proposed family.

Finally, the consistency and efficiency of the contributed methods have also been tested numerically for finding the matrix sign functions to support the theoretical parts. Now we draw the attention to the fact that matrix sector function,

introduced in [20], is a generalization of the matrix sign function, so extension of the discussions given in this work for computing matrix sector functions can be taken into consideration for future works in this active research line.

Acknowledgements

The authors thank Dr. Ben Nolting for useful discussions and several help in obtaining moving attraction basins in Section 2.

[1] P. Benner, E.S. Qintana-Ortí, Solving stable generalized Lyapunov equations with the matrix sign function, Numer. Algor. 20(1) (1999) 75–100.

[2] S. Barrachina, P. Benner, E.S. Qintana-Ortí, Efficient algorithms for generalized algebraic Bernoulli equations based on the matrix sign function, Numer. Algor. 46(4) (2007) 351–368.

[3] F.D. Filbir, Computation of the structured stability radius via matrix sign function, Sys. & Contr. Lett. 22 (1994), 341–349.

[4] Y.H. Geum, Y.I. Kim, A multi-parameter family of three-step eighth-order iterative methods locating a simple root, Appl. Math. Comput. 215 (2010) 3375–3382.

[5] F. Greco, B. Iannazzo, F. Poloni, The Padé iterations for the matrix sign function and their reciprocals are optimal, Lin. Alg. Appl. 436 (2012) 472-477.

[6] O. Gomilko, F. Greco, K. Ziętak, A Padé family of iterations for the matrix sign function and related problems, Numer. Lin. Alg. Appl. 19 (2012) 585-605.

[7] J.M. Gutiérrez, M.A. Hernández, A family of Chebyshev-Halley type methods in Banach spaces, Bull. Austral. Math. Soc. 55 (1997) 113–130.

[8] H.V. Henderson, S.R. Searle, On deriving the inverse of a sum of matrices, SIAM Rev. 23 (1981) 53–60.

[9] N.J. Higham, The matrix sign decomposition and its relation to the polar decomposition, Lin. Alg. Appl. 212/213 (1994) 3–20.

[10] N.J. Higham, D.S. Mackey, N. Mackey, F. Tisseur, Computing the polar decomposition and the matrix sign decomposition in matrix groups, SIAM Matrix Anal. Appl. 25 (2004) 1178–1192.

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

[12] J.L. Howland, The sign matrix and the separation of matrix eigenvalues, Lin. Alg. Appl. 49 (1983) 221–232.

[13] B. Iannazzo, Numerical solution of certain nonlinear matrix equations, Ph.D. thesis, Dipartimento di Matematica, Università di Pisa, 2007.

[14] C. Kenney, A.J. Laub, Rational iterative methods for the matrix sign function, SIAM Matrix Anal. Appl. 12 (1991) 273–291.

[15] C.S. Kenney, A.J. Laub, The matrix sign function, IEEE Trans. Automat. Cont., 40 (1995) 1330–1348.

[16] P. Lancaster, L. Rodman, Algebraic Riccati Equations, Oxford University Press, 1995.

[17] M.Sh. Misrikhanov, V.N. Ryabchenko, Matrix sign function in the problems of analysis and design of the linear systems, Auto. Remote Control 69 (2008) 198–222.

[18] A.N. Norris, A.L. Shuvalov, A.A. Kutsenko, The matrix sign function for solving surface wave problems in homogeneous and laterally periodic elastic half-spaces, Wave Motion 50(8) (2013) 1239–1250.

[19] J.D. Roberts, Linear model reduction and solution of the algebraic Riccati equation by use of the sign function, Int. J. Control. 32 (1980) 677–687.

[20] L.S. Shieh, Y.T. Tsay, C.T. Wang, Matrix sector functions and their applications to system theory, IEE Proc. 131 (1984) 171–181.

[21] A.R. Soheili, F. Toutounian, F. Soleymani, A fast convergent numerical method for matrix sign function with application in SDEs, Comput. Appl. Math. 282 (2015) 167–178.

[22] F. Soleymani, P.S. Stanimirović, I. Stojanović, A novel iterative method for polar decomposition and matrix sign function, Disc. Dyn. Nature Soc., Vol. 2015, Art. ID 649423, 11 pages.

[23] M. Trott, The Mathematica GuideBook for Numerics, Springer, NY, USA, 2006.

[24] P.R. Wellin, R.J. Gaylord, S.N. Kamin, An Introduction to Programming with Mathematica, Cambridge University Press, UK, 2005.

[25] Wolfram Research, Inc., Mathematica, Version 10.4, Champaign, IL 2016.