Mathematics

http://www.aimspress.com/journal/Math

AIMS Mathematics, 8(5): 12456–12471.
DOI:10.3934/math.2023625
Received: 07 January 2023
Revised: 23 February 2023
Accepted: 09 March 2023
Published: 27 March 2023

Research article

A numerically stable high-order Chebyshev-Halley type multipoint iterative method for calculating matrix sign function

Xiaofeng Wang* **and Ying Cao**

School of Mathematical Sciences, Bohai University, Jinzhou 121000, Liaoning, China

* **Correspondence:** Email: xiaofengwang@bhu.edu.cn.

Abstract: A new eighth-order Chebyshev-Halley type iteration is proposed for solving nonlinear equations and matrix sign function. Basins of attraction show that several special cases of the new method are globally convergent. It is analytically proven that the new method is asymptotically stable and the new method has the order of convergence eight as well. The effectiveness of the theoretical results are illustrated by numerical experiments. In numerical experiments, the new method is applied to a random matrix, Wilson matrix and continuous-time algebraic Riccati equation. Numerical results show that, compared with some well-known methods, the new method achieves the accuracy requirement in the minimum computing time and the minimum number of iterations.

Keywords: Chebyshev-Halley family; iterative method; basin of attraction; matrix sign function; stability
Mathematics Subject Classification: 65B99, 65H05

1. Introduction

The Chebyshev-Halley method is a popular iterative method for solving the simple roots of the nonlinear equation $f(x) = 0$. In fact, the Chebyshev-Halley method has been first provided by Hernández and Salanova [1]. Gutiérrez and Hernández [2] have provided a modification for the Chebyshev-Halley type iterative methods in Banach spaces. The base for constructing the family is the third-order method

$$x_{k+1} = x_k - (1 + \frac{1}{2}(\frac{L(x_k)}{1 - aL(x_k)}))\frac{f(x_k)}{f'(x_k)}, \tag{1.1}$$

where $a \in \mathbb{R}$, $L(x_k) = \frac{f''(x_k)f(x_k)}{f'(x_k)^2}$.

In 2008, Osada [3] gave two modifications of family of Chebyshev-Halley methods for analytic functions. With the developments of the theory of iteration processes, Kim et al. [4] proposed several new families of Chebyshev-Halley type methods based on weight function. Ivanov [5] established convergence theorems of Chebyshev-Halley iteration family for multiple polynomial zeros. In this paper, we proposed a new Chebyshev-Halley type method for solving the matrix sign functions.

In addition, the applications obtained for the matrix sign functions are significant. For example, matrix sign function can be used as a tool for solving of the algebraic Riccati equation [6–8], Lyapunov matrix equations [9], generalized algebraic Bernoulli equations [10] and separation problem of matrix eigenvalues [11]. It is used as a valuable method to compute the matrix square root, the matrix pth roots and the polar decomposition [12]. Due to the applicability of the matrix sign function, stable iterative schemes with global convergence have become viable choices for computing this matrix. Basin of attractions can help us to obtain the iterative schemes with global convergence, for example, Soleymani et al. [13, 14] imposed some high order iterative methods for solving matrix sign function with application in stochastic differential equations. Other high-order iterative methods with global convergent have been proposed for solving matrix sign function, see [15–19] and the references therein.

The sign of a matrix function generalizes the scalar sign, then the scalar sign function for any $z \in \mathbb{C}$ not lying on the imaginary axis is given by

$$sign(z) = \begin{cases} 1, & ifRe(z) > 0, \\ -1, & ifRe(z) < 0. \end{cases} \tag{1.2}$$

Indeed, this function for the matrix case was introduced by Roberts [19] for solving the problem of model reduction and the algebraic Riccati equations.

We suppose that the matrix $A \in \mathbb{C}^{n\times n}$ has no eigenvalues on the imaginary axis. If $A = TJ_AT^{-1}$ is a Jordan canonical form arranged so that

$$J_A = \begin{pmatrix} J_1 & 0 \\ 0 & J_2 \end{pmatrix}, \tag{1.3}$$

where the eigenvalues of $J_1 \in \mathbb{C}^{q\times q}$ and $J_2 \in \mathbb{C}^{(n-q)\times(n-q)}$ locate in the open left half-plane and the open right half-plane, respectively. Then, the matrix sign function is given by

$$S = sign(A) = T\begin{pmatrix} -I_q & 0 \\ 0 & I_{n-q} \end{pmatrix}T^{-1}. \tag{1.4}$$

Actually, for any positive integer p, the matrix p-sector function [20] can be defined by

$$sect_p(A) = A(A^p)^{-1/p}, \tag{1.5}$$

where the matrix sign function $S = A(A^2)^{-1/2}$ is taken in the case $p = 2$. In fact, S has various properties, as shown in [21].

The well known iteration for computation of S is Newton's method (NM)

$$X_{k+1} = \frac{1}{2}(X_k + X_k^{-1}). \tag{1.6}$$

We know that iteration (1.6) is by no means the only rational matrix method for computing S, as a lot of other iterations have been derived to get a higher order of convergence, and to improve the convergence acceleration. Hopefully, Laub and Kenney established the Padé family of iterations in [22]. For non-pure imaginary $z \in \mathbb{C}$, we have the following characterization:

$$sign(z) = \frac{z}{(1-(1-z^2))^{1/2}} = \frac{z}{(1-\xi)^{1/2}}, \tag{1.7}$$

where $\xi = 1 - z^2$, hence the task of approximating sign(z) leads to that of approximating

$$f(\xi) = (1-\xi)^{-1/2}, \tag{1.8}$$

where ξ is less than 1 in magnitude. Let the (l,n)- Padé approximant to $f(\xi)$ be

$$\frac{P_{l,n}(\xi)}{Q_{l,n}(\xi)}, \tag{1.9}$$

and $l + n \geq 1$. More precisely, Kenney and Laub set up the rational iterations of the form

$$x_{k+1} = g_r(x_k) = \frac{x_k P_{l,n}(1-x_k^2)}{Q_{l,n}(1-x_k^2)} := f_{ln}, \tag{1.10}$$

where $P_{l,n}/Q_{l,n}$ denotes the (l,n)- Padé approximant to the function $(1-\xi)^{-1/2}$. Obviously, for any suitable l and n, the matrix versions of the iterations are given by

$$X_{k+1} = X_k P_{l,n}(I - X_k^2) Q_{l,n}(I - X_k^2)^{-1} := g_r(X_k),\ X_0 = A, \tag{1.11}$$

where $r = l + n + 1$. It has been proved that the iterations with $l = n$ and $l = n - 1$ are globally convergent, while those with $l \geq n + 1$ have local convergence. The convergence rate being $l + n + 1$ in every case. Similarly, we can give the reciprocal Padé iterations as follows:

$$X_{k+1} = \frac{Q_{l,n}(I - X_k^2)}{X_k P_{l,n}(I - X_k^2)} := \frac{1}{g_r(X_k)}. \tag{1.12}$$

Table 1 gives the principal Padé iteration and its reciprocal for the order $4 \leq r \leq 10$. PM_r denotes the principal Padé iteration, while RPM_r denotes its reciprocal Padé iteration.

Table 1. Principal Padé iterations and their reciprocals for the order 4≤ r ≤10.

Method	(l,n)	Expressions
PM4	$(1,2)$	$4X_k(X_k^2 + I)(X_k^4 + 6X_k^2 + I)^{-1}$
PM5	$(2,2)$	$X_k(X_k^4 + 10X_k^2 + 5I)(5X_k^4 + 10X_k^2 + I)^{-1}$
PM6	$(2,3)$	$2X_k(3X_k^4 + 10X_k^2 + 3I)(X_k^6 + 15X_k^4 + 15X_k^2 + I)^{-1}$
PM7	$(3,3)$	$X_k(X_k^6 + 21X_k^4 + 35X_k^2 + 7I)(7X_k^6 + 35X_k^4 + 21X_k^2 + I)^{-1}$
PM8	$(3,4)$	$8X_k(X_k^6 + 7X_k^4 + 7X_k^2 + I)(X_k^8 + 28X_k^6 + 70X_k^4 + 28X_k^2 + I)^{-1}$
PM9	$(4,4)$	$X_k(X_k^8 + 36X_k^6 + 126X_k^4 + 84X_k^2 + 9I)(9X_k^8 + 84X_k^6 + 126X_k^4 + 36X_k^2 + I)^{-1}$
PM10	$(4,5)$	$2X_k(5X_k^8 + 60X_k^6 + 126X_k^4 + 60X_k^2 + 5I)(X_k^{10} + 45X_k^8 + 210X_k^6 + 210X_k^4 + 45X_k^2 + I)^{-1}$
RPM4	$(1,2)$	$(X_k^4 + 6X_k^2 + I)(4X_k(X_k^2 + I))^{-1}$
RPM5	$(2,2)$	$(5X_k^4 + 10X_k^2 + I)(X_k(X_k^4 + 10X_k^2 + I))^{-1}$
RPM6	$(2,3)$	$(X_k^6 + 15X_k^4 + 15X_k^2 + I)(2X_k(3X_k^4 + 10X_k^2 + 3I))^{-1}$
RPM7	$(3,3)$	$(7X_k^6 + 35X_k^4 + 21X_k^2 + I)(X_k(X_k^6 + 21X_k^4 + 35X_k^2 + 7I))^{-1}$
RPM8	$(3,4)$	$(X_k^8 + 28X_k^6 + 70X_k^4 + 28X_k^2 + I)(8X_k(X_k^6 + 7X_k^4 + 7X_k^2 + I))^{-1}$
RPM9	$(4,4)$	$(9X_k^8 + 84X_k^6 + 126X_k^4 + 36X_k^2 + I)(X_k(X_k^8 + 36X_k^6 + 126X_k^4 + 84X_k^2 + 9I))^{-1}$
RPM10	$(4,5)$	$(X_k^{10} + 45X_k^8 + 210X_k^6 + 210X_k^4 + 45X_k^2 + I)(2X_k(5X_k^8 + 60X_k^6 + 126X_k^4 + 60X_k^2 + 5I))^{-1}$

In particular, we study the cases $l = n - 1$ and $l = n$, which we call the principal Padé iterations [23]. They satisfy the equation

$$g_r(x) = \frac{xP_{l,n}(1 - x_k^2)}{Q_{l,n}(1 - x_k^2)} = \frac{(1 + x)^r - (1 - x)^r}{(1 + x)^r + (1 - x)^r}. \tag{1.13}$$

Note that $g_r(x) = p_r(x)/q_r(x)$, where $q_r(x)$ and $p_r(x)$ are the even and odd parts of $(1 + x)^r$, respectively. On the other hand, lots of known iterative methods are contained in the Padé family or its reciprocal. Newton-Schultz iteration (NSM) can be retrieved in the case of $l = 1$ and $n = 0$,

$$X_{k+1} = \frac{1}{2}X_k(3I - X_k^2). \tag{1.14}$$

Choosing $l = 1$ and $n = 1$ yields the Halley's method (HM):

$$X_{k+1} = [I + 3X_k^2][X_k(3I + X_k^2)]^{-1}. \tag{1.15}$$

After a brief introduction, we are interested in constructing more efficient methods for solving the matrix sign function. We also focus on some intriguing properties for finding the matrix sign function, including higher order convergence, global convergence, stability and the efficiency. In Section 2, we propose a variant of Chebyshev-Halley family with a free parameter. The convergence order of new Chebyshev-Halley type family is eight. Fractal results show that some special cases of the new family have global convergence. Some special cases of the new family are not in the Padé family. This means that some new iterative methods are obtained for solving the matrix sign function. We theoretically prove that the new family is convergent and stable in Sections 3 and 4, respectively. In Section 5, we compare our method with the existing methods by numerical experiments. The proposed method is applied to a random matrix,Wilson matrix and continuous-time algebraic Riccati equation. Numerical results show the effectiveness of the proposed methods. Finally, Section 6 concludes the findings.

2. A new Chebyshev-Halley type iterative method

Here, we give the nonlinear matrix equation below:

$$X^2 - I = 0, \tag{2.1}$$

in which I is a unit matrix. Moreover, the sign S is a solution of (2.1).

In fact, before we obtain a new iterative method to solve the matrix equation $X^2 - I = 0$, we should discuss two important problems about the new matrix iteration. The first is not in the Padé family or its reciprocal. Second, it must be globally convergent.

Now, we propose a new iterative method of Chebyshev-Halley type with a free parameter:

$$\begin{cases} y_k = x_k - (1 + \frac{1}{2}(\frac{L(x_k)}{1-aL(x_k)}))\frac{f(x_k)}{f'(x_k)}, \\ z_k = y_k - \frac{f(y_k)}{f[y_k,x_k]}, \\ x_{k+1} = z_k - \frac{f(z_k)}{2f[z_k,y_k]-f'(y)}, \end{cases} \tag{2.2}$$

where $a \in \mathbb{R}$, noting that

$$L(x_k) = \frac{f''(x_k)f(x_k)}{f'(x_k)^2}, f[y_k, x_k] = \frac{f(y_k) - f(x_k)}{y_k - x_k}, f[z_k, y_k] = \frac{f(z_k) - f(y_k)}{z_k - y_k}. \tag{2.3}$$

By inserting Eqs (2.2) and (2.3) into Eq (2.1), we attain iteration in the reciprocal form as follows:

$$\begin{aligned} X_{k+1} =& X_k((2-16a+24a^2)I+(-40+128a+32a^2)X_k^2+(140+224a-112a^2)X_k^4 \\ &+(344-256a+32a^2)X_k^6+(66-80a+24a^2)X_k^8)\times \\ &[(1-2a)^2I+(-11+4a+52a^2)X_k^2+(-14+280a-56a^2)X_k^4 \\ &+(322-56a-56a^2)X_k^6+(205-212a+52a^2)X_k^8+(9-12a+4a^2)X_k^{10}]^{-1}. \end{aligned} \tag{2.4}$$

We can achieve the method (2.4) for calculating the sign function. Note that, $X_k (k \geq 0)$ are rational functions of A and, hence, like A, commute with S.

Theorem 1. *Let $f(x) = 0$ be a function around the simple root α. If the initial point x_0 is sufficiently close to α, then the order of convergence for (2.2) is eight, for any value of parameter a, with the following error equation:*

$$\varepsilon_{k+1} = c_2{}^3(2(-1+a)c_2{}^2+c_3)^2\varepsilon_k^8+o(\varepsilon_k^9), \tag{2.5}$$

where $c_j = \frac{f^{(j)}(\alpha)}{j!f'(\alpha)}$, $j \geq 2$, and $\varepsilon_k = x_k - \alpha$.

Proof. The result is based on Taylor's series and symbolic computation in Mathematica [2], where this is skipped over.

Now, some different cases of the family (2.4) are given below.

If $a = -1$, we have (M1)

$$\begin{aligned} X_{k+1} =&(42X_k-136X_k^3-196X_k^5+632X_k^7+170X_k^9)[9I+37X_k^2-350X_k^4 \\ &+322X_k^6+469X_k^8+25X_k^{10}]^{-1}. \end{aligned} \tag{2.6}$$

If $a = -2$, we have (M2)

$$\begin{aligned} X_{k+1} =&(130X_k-168X_k^3-756X_k^5+984X_k^7+322X_k^9)[25I+189X_k^2-798X_k^4 \\ &+210X_k^6+837X_k^8+49X_k^{10}]^{-1}. \end{aligned} \tag{2.7}$$

If $a = 1$, we have (M3)

$$X_{k+1} = (10X_k+120X_k^3+252X_k^5+120X_k^7+10X_k^9)[I+45X_k^2+210X_k^4+210X_k^6+45X_k^8+X_k^{10}]^{-1}. \tag{2.8}$$

If $a = 2$, we have (M4)

$$X_{k+1} = (66X_k+344X_k^3+140X_k^5-40X_k^7+2X_k^9)[9I+205X_k^2+322X_k^4-14X_k^6-11X_k^8+X_k^{10}]^{-1}. \tag{2.9}$$

If $a = -1/2$, we have (M5)

$$X_{k+1} = (4X_k-24X_k^3+120X_k^7+28X_k^9)[I-42X_k^4+84X_k^6+81X_k^8+4X_k^{10}]^{-1}. \tag{2.10}$$

If $a = 1/2$, we have (M6)

$$X_{k+1} = (8X_k+56X_k^3+56X_k^5+8X_k^7)[I+28X_k^2+70X_k^4+28X_k^6+X_k^8]^{-1}. \tag{2.11}$$

If $a = 0$, we have (M7)

$$X_{k+1} = (2X_k-40X_k^3+140X_k^5+344X_k^7+66X_k^9)[I-11X_k^2-14X_k^4+322X_k^6+205X_k^8+9X_k^{10}]^{-1}. \tag{2.12}$$

If $a = 3/4$, we have (M8)

$$X_{k+1} = (14X_k+296X_k^3+980X_k^5+680X_k^7+78X_k^9)[I+85X_k^2+658X_k^4+994X_k^6+301X_k^8+9X_k^{10}]^{-1}. \tag{2.13}$$

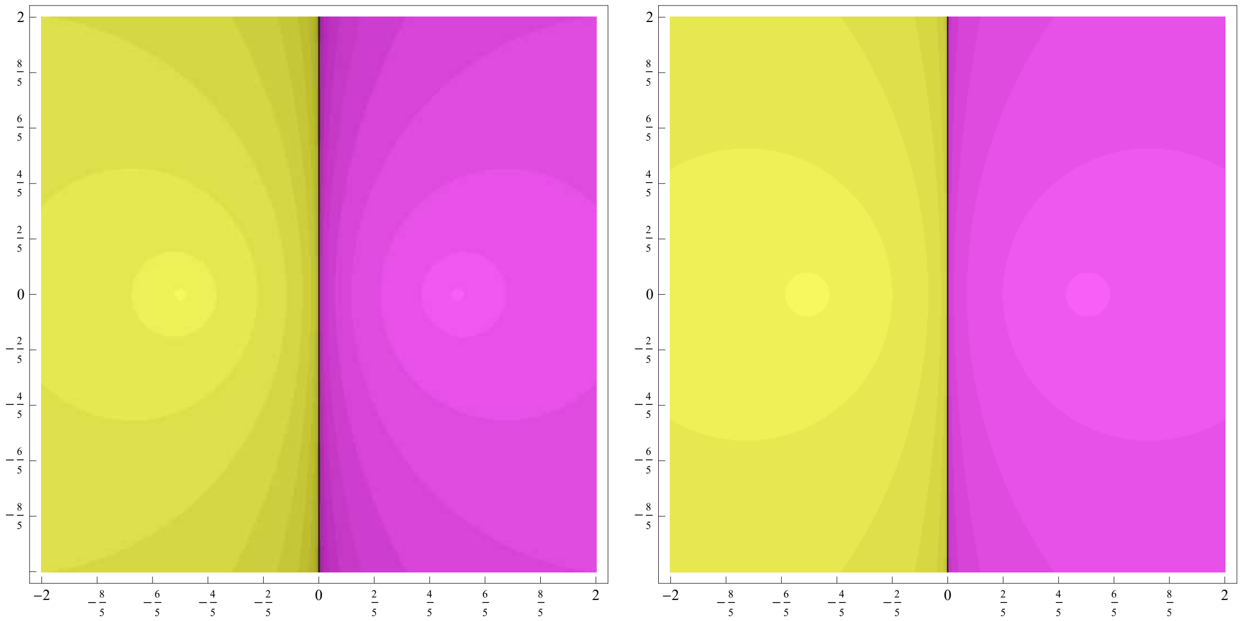
□

One might want to attain an efficient scheme for selecting values of the free parameter a to compute the matrix sign function. We remake that the proposed methods should satisfy two cases, that is, to possess global convergence and to not be in the general Padé family of iterations (1.10).

Now, we can observe the convergence behavior of the members of the family (2.4) by drawing the attraction basins (for more information see [15]). For such a case, the associated attraction basins in terms of some values of parameter a to solve the scalar equation $x^2 - 1 = 0$ are presented. We already know that the attraction basins for the Newton-Schultz iteration (1.14) has local convergence, so we neglect to draw it.

Here, we take the domain $\Gamma = [-2, 2]\times[-2, 2] \in \mathbb{C}$. Each starting point $z_0 \in \Gamma$ is allocated a color by the simple zero where the new scheme (2.4) converges. The point is painted in black when the method diverges. The stopping criterion is $|f(x_k)| \leq 10^{-3}$ in our programs. In addition, we set the maximum number of iterations to 30. Note that the roots are plotted in two white points.

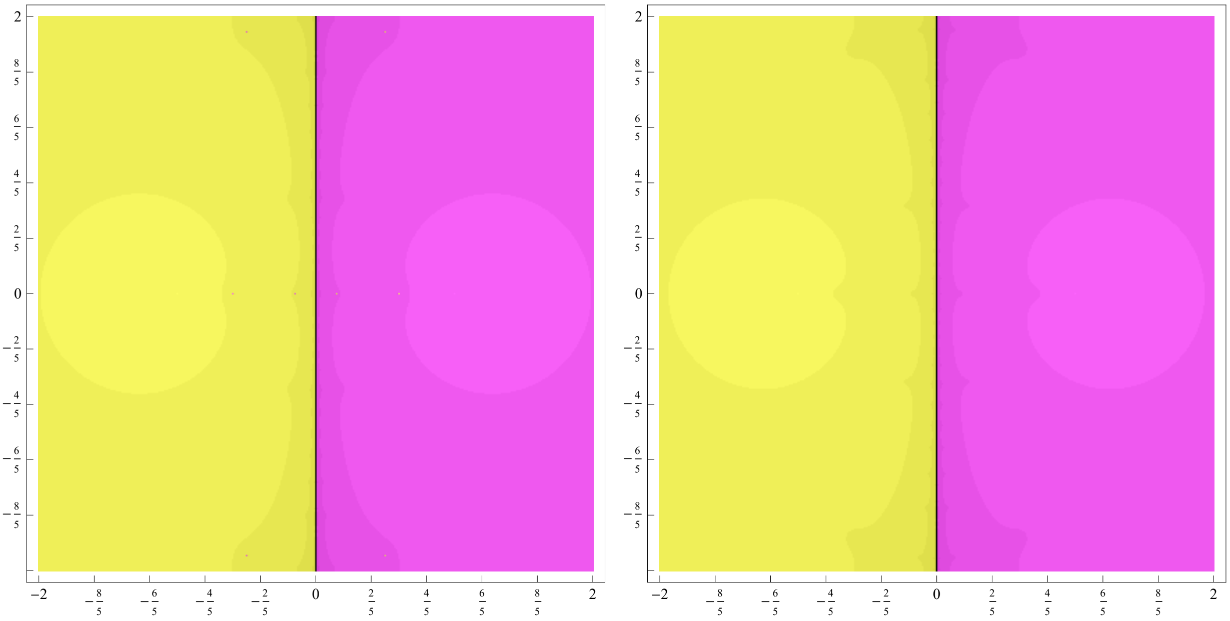
First of all, in Figure 1, the attraction basins of NM and HM are shown. We also provide different kinds of attraction basins for $a = \pm 1$, $a = \pm 1/2$, $a = \pm 2$, $a = 0$ and $a = 3/4$ in Figures 2–5, respectively. We can see that methods M1, M2, M4, M5 and M7 have local convergence. Methods M3, M6 and M8 corresponding to $a = 1$, $a = 1/2$, $a = 3/4$ perform the global convergence. But, M3 and M6 are the members from the Padé family, which are not new iterative schemes. M8 does not belong to the member of the Padé family. So, method M8 is a new iterative scheme with global convergence. Indeed, our main aim in constructing iterative schemes for the matrix sign is to reach the convergence order as fast as possible, and also to minimize computational cost. Taking into account the difficulty of finding a general family of matrix sign iterations, we should consider other essential properties of iterations for calculating the sign of a matrix. The following part is the convergence analysis.



(a) (NM)

(b) (HM)

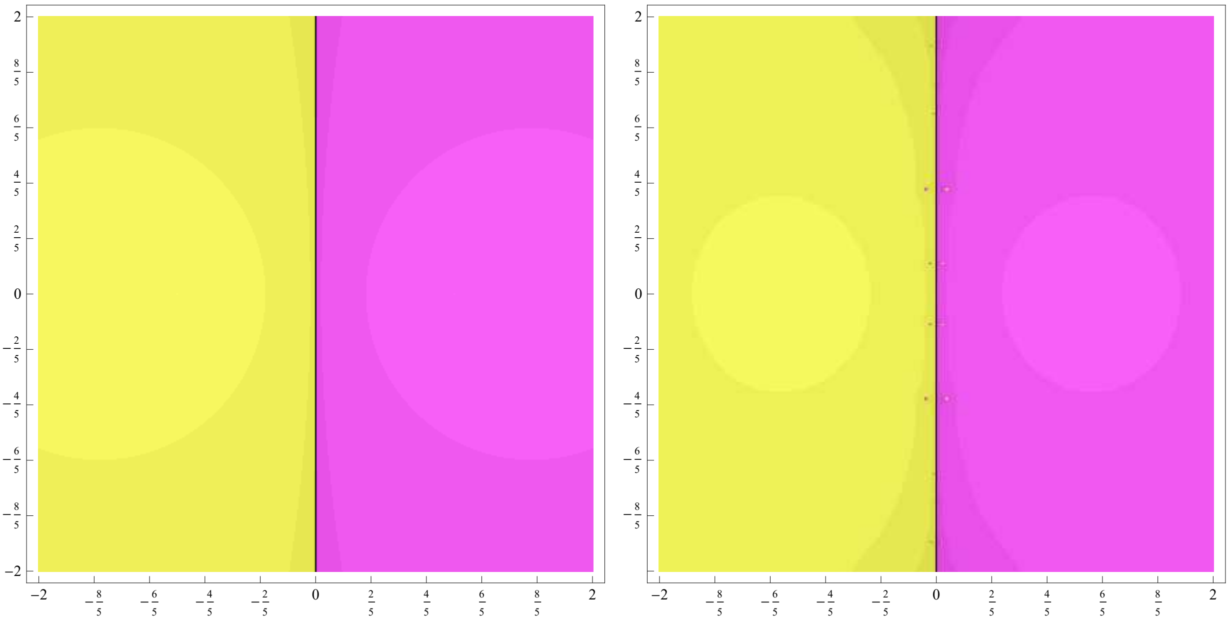
Figure 1. The basins of attraction for (1.6) (left) and (1.15) (right), for the polynomial (shaded by the number of iterations).



(a) (M1)

(b) (M2)

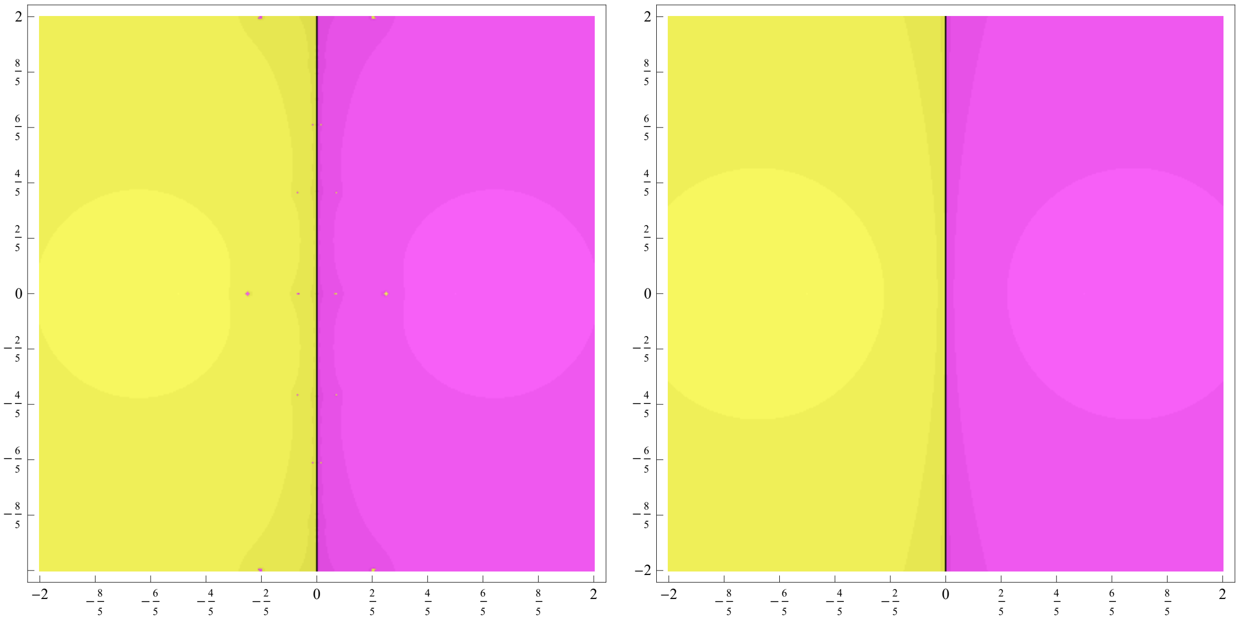
Figure 2. Basins of attraction for ($a = -1$) (left) and ($a = -2$) (right), for the polynomial $x^2 - 1 = 0$.



(a) (M3)

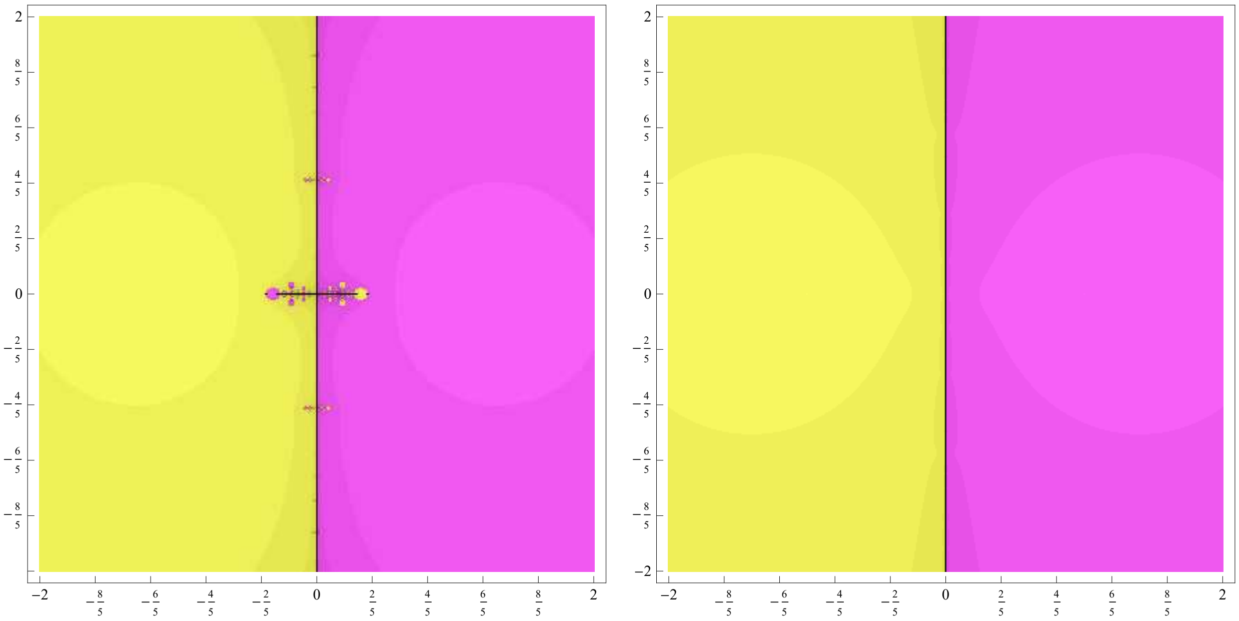
(b) (M4)

Figure 3. Basins of attraction for ($a = 1$) (left) and ($a = 2$) (right), for the polynomial $x^2 - 1 = 0$.



(a) (M5) (b) (M6)

Figure 4. Basins of attraction for ($a = -1/2$) (left) and ($a = 1/2$) (right), for the polynomial $x^2 - 1 = 0$.



(a) (M7) (b) (M8)

Figure 5. Basins of attraction for ($a = 0$) (left) and ($a = 3/4$) (right), for the polynomial $x^2 - 1 = 0$.

3. Convergence analysis

Here, we will show a precise convergence analysis of the method (2.4) under some assumptions. The following theorem is done to provide insight into the analysis.

Theorem 2. *Let $A \in \mathbb{C}^{n\times n}$ have no pure imaginary eigenvalues. Then, the proposed iterates $\{X_k\}_{k=0}^{k=\infty}$ of (2.4) converge to the matrix sign S, choosing X_0=A.*

Proof. Let A have a Jordan canonical form arranged as

$$T^{-1}AT = \Lambda = \begin{bmatrix} C & 0 \\ 0 & N \end{bmatrix}, \tag{3.1}$$

where T is a nonsingular matrix and C, N are the square Jordan blocks corresponding to eigenvalues lying in $\mathbb{C}^-$ and $\mathbb{C}^+$, respectively. Denote by λ_1 , ...,λ_q and λ_{q+1}, ...,λ_n values lying on the main diagonals of blocks C and N, respectively.

Of course, recall that

$$sign(A) = T\begin{pmatrix} -I_q & 0 \\ 0 & I_{n-q} \end{pmatrix}T^{-1}. \tag{3.2}$$

Thus,

$$sign(\Lambda) = sign(T^{-1}AT) = T^{-1}sign(A)T = diag(sign(\lambda_1), ..., sign(\lambda_q), sign(\lambda_{q+1}), ..., sign(\lambda_n)). \tag{3.3}$$

Furthermore, we give the definition of $D_k = T^{-1}X_kT$, and it follows from (2.4) that

$$\begin{aligned} D_{k+1} =& ((2-16a+24a^2)D_k + (-40+128a+32a^2)D_k^3 + (140+224a-112a^2)D_k^5 \\ &+ (344-256a+32a^2)D_k^7 + (66-80a+24a^2)D_k^9)\times \\ &[(1-4a+4a^2)I + (-11+4a+52a^2)D_k^2 + (-14+280a-56a^2)D_k^4 \\ &+ (322-56a-56a^2)D_k^6 + (205-212a+52a^2)D_k^8 + (9-12a+4a^2)D_k^{10}]^{-1}. \end{aligned} \tag{3.4}$$

Notice that if D_0 is a diagonal matrix, then using inductive proof, all successive D_k are diagonal matrices as well.

From the Eq (3.4), we can see that $\{D_k\}$ converges to sign(Λ). By re-arranging (3.4) as n uncoupled scalar iterations to solve $x^2 - 1 = 0$, we can derive the following equation

$$\begin{aligned} d_{k+1}^i =& ((2-16a+24a^2)d_k^i + (-40+128a+32a^2){d_k^i}^3 + (140+224a-112a^2){d_k^i}^5 \\ &+ (344-256a+32a^2){d_k^i}^7 + (66-80a+24a^2){d_k^i}^9)((1-4a+4a^2) + (-11+4a+52a^2){d_k^i}^2 \\ &+ (-14+280a-56a^2){d_k^i}^4 + (322-56a-56a^2){d_k^i}^6 + (205-212a+52a^2){d_k^i}^8 \\ &+ (9-12a+4a^2){d_k^i}^{10})^{-1}, \end{aligned} \tag{3.5}$$

where $d_k^i = (D_k)_{i,i}$ and $1 \le i \le n$. Similarly, for all $1 \le i \le n$, application of (3.4) and (3.5) lead to the convergence of $\{d_k^i\}$ to sign(λ_i).

Since the eigenvalues of A are not pure imaginary values and from (3.5), we have $sign(\lambda_i) = s_i = \pm 1$. Therefore, it follows that

$$\frac{d_{k+1}^i - 1}{d_{k+1}^i + 1} = -\frac{(d_k^i - 1)^8(1 + 3d_k^i - 2a(1 + d_k^i))^2}{(d_k^i + 1)^8(1 - 3d_k^i + 2a(-1 + d_k^i))^2}. \tag{3.6}$$

Since $|d_0^i| = |\lambda_i| > 0$ and $|\frac{d_0^i-1}{d_0^i+1}| < 1$, thus we have $\lim\limits_{k\to\infty} |\frac{d_{k+1}^i-1}{d_{k+1}^i+1}| = 0$, and $\lim\limits_{k\to\infty} |d_k^i| = 1 = |sign(\lambda_i)|$. This shows that $\{d_k^i\}$ is convergent. Then, we can obtain that $\lim\limits_{k\to\infty} D_k = sign(\Lambda)$. Recalling $D_k = T^{-1}X_kT$, we have

$$\lim_{k\to\infty} X_k = T(\lim_{k\to\infty} D_k)T^{-1} = T sign(\Lambda)T^{-1} = sign(A). \tag{3.7}$$

Finally, this finishes the proof. □

Clearly, X_k are rational functions of A and therefore, similar to A, commute with S. Note that we know that $S^{-1} = S$, $S^2 = I$, $S^{2j+1} = S$, and $S^{2j} = I$, $j \geq 1$. Hence, it is easy to show that the method (2.4) reads the following error inequality

$$||X_{k+1} - S|| \leq (||B_k^{-1}||||I + 3X_k - 2a(I + X_k)||^2)||X_k - S||^8, \tag{3.8}$$

where $B_k = (1-2a)^2I + (-11+4a+52a^2)X_k^2 + (-14+280a-56a^2)X_k^4 + (322-56a-56a^2)X_k^6$ $+(205-212a+52a^2)X_k^8+(9-12a+4a^2)X_k^{10}$. The inequality (3.8) reveals the eighth order of convergence.

4. Stability

This section begins with analysis of the stability of (2.4). We consider to prove the stability of (2.4) for finding S in a neighborhood of the solution. Take into account that how a small perturbation at the kth iterate is amplified or damped along the iterates.

Theorem 3. *Let $A \in \mathbb{C}^{n\times n}$ have no pure imaginary eigenvalues. The sequence $\{X_k\}_{k=0}^{k=\infty}$ generated by (2.4) is asymptotically stable.*

Proof. If X_0 is a function of A, then the iterations of (2.4) are all functions of A and, thus, commute with A. Let $\triangle X_k$ be the numerical perturbation introduced at the kth iterate of (2.4). Then, we have

$$\widetilde{X}_k = X_k + \triangle X_k. \tag{4.1}$$

Here, we formally use approximations $(\triangle X_k)^i \approx 0$, since $(\triangle X_k)^i \approx 0$, $i \geq$2, is small. For small $\triangle X_k$, we can neglect the value. In this case, we get

$$\begin{aligned}
\widetilde{X}_{k+1} &= ((2-16a+24a^2)\widetilde{X}_k + (-40+128a+32a^2)\widetilde{X}_k^3 + (140+224a-112a^2)\widetilde{X}_k^5 \\
&\quad + (344-256a+32a^2)\widetilde{X}_k^7 + (66-80a+24a^2)\widetilde{X}_k^9)\times \\
&\quad [(1-4a+4a^2)I + (-11+4a+52a^2)\widetilde{X}_k^2 + (-14+280a-56a^2)\widetilde{X}_k^4 \\
&\quad + (322-56a-56a^2)\widetilde{X}_k^6 + (205-212a+52a^2)\widetilde{X}_k^8 + (9-12a+4a^2)\widetilde{X}_k^{10}]^{-1} \\
&= ((2-16a+24a^2)(X_k+\triangle X_k) + (-40+128a+32a^2)(X_k+\triangle X_k)^3 \\
&\quad + (140+224a-112a^2)(X_k+\triangle X_k)^5 + (344-256a+32a^2)(X_k+\triangle X_k)^7 \\
&\quad + (66-80a+24a^2)(X_k+\triangle X_k)^9)\times[(1-4a+4a^2)I + (-11+4a+52a^2)(X_k+\triangle X_k)^2 \\
&\quad + (-14+280a-56a^2)(X_k+\triangle X_k)^4 + (322-56a-56a^2)(X_k+\triangle X_k)^6 \\
&\quad + (205-212a+52a^2)(X_k+\triangle X_k)^8 + (9-12a+4a^2)(X_k+\triangle X_k)^{10}]^{-1}.
\end{aligned} \tag{4.2}$$

By using the identity for any nonsingular matrix B and C [2]

$$(B+C)^{-1} \approx B^{-1} - B^{-1}CB^{-1}, \tag{4.3}$$

the Eq (4.2) can be written as

$$\begin{aligned}\widetilde{X}_{k+1} &\approx (512S + (2048 - 512a) \triangle X_k + (1536 - 512a)S \triangle X_k S) \\ &\quad \times (512I + (1792 - 512a)S \triangle X_k + (1792 - 512a) \triangle X_k S)^{-1} \\ &\approx (S + (4 - a) \triangle X_k + (3 - a)S \triangle X_k S) \times (I - (\frac{7}{2} - a)S \triangle X_k - (\frac{7}{2} - a) \triangle X_k S) \\ &\approx S + \frac{1}{2} \triangle X_k - \frac{1}{2} S \triangle X_k S.\end{aligned} \tag{4.4}$$

Further, we apply $\triangle X_{k+1} = \widetilde{X}_{k+1} - X_{k+1} = \widetilde{X}_{k+1} - S$, and it is easy to show that

$$\triangle X_{k+1} \approx \frac{1}{2} \triangle X_k - \frac{1}{2} S \triangle X_k S. \tag{4.5}$$

We can know that the perturbation at the $(k + 1)st$ iteration is bounded; that is, we have

$$\| \triangle X_{k+1} \| \leq \frac{1}{2} \| S \triangle X_0 S - \triangle X_0 \|. \tag{4.6}$$

To summarize, the sequence $\{X_k\}_{k=0}^{k=\infty}$ generated by (2.4) is stable. The proof is ended. □

5. Numerical experiments

This section describes some numerical experiments that demonstrate the effectiveness of the proposed method. On the other hand, we now only use iterative methods with global convergence. In the following tests, the compared schemes are NM, HM, PM4, M3, M6 and M8.

For a fairer comparison, all computations are performed on same laptop equipped with Core i5, 7th generation CPU. The stopping termination $\|X_k^2 - I\|_2 \leq 10^{-4}$ is used in Tables 2 and 3. We have examined each method of 9 randomly generated matrices. The numerical results for different methods on random matrices of sizes 5×5, 10×10, 20×20, 50×50, 100×100, $150 \times 150, 200 \times 200, 250 \times 250$ and 300×300 are given in Tables 2 and 3. Finally, we have recorded their average in the last line.

Table 2. Comparisons of number of iterations.

$Matrix No.$	NM	HM	$PM4$	$M3$	$M6$	$M8$
$A_{5\times5}$	11	8	6	4	5	3
$A_{10\times10}$	11	9	7	4	5	4
$A_{20\times20}$	14	8	7	5	5	4
$A_{50\times50}$	16	11	8	5	6	5
$A_{100\times100}$	18	12	9	6	6	5
$A_{150\times150}$	21	12	10	6	7	6
$A_{200\times200}$	23	13	10	7	6	6
$A_{250\times250}$	22	13	11	7	7	7
$A_{300\times300}$	23	15	12	8	7	7
Mean	17.7	11.2	8.9	5.8	6	5.2

Table 3. Comparisons of elapsed time (s).

Matrix No.	*NM*	*HM*	*PM4*	*M3*	*M6*	*M8*
$A_{5\times5}$	0.00448	0.00636	0.00432	0.00446	0.00435	0.00434
$A_{10\times10}$	0.00390	0.00406	0.00142	0.00496	0.00487	0.00492
$A_{20\times20}$	0.01830	0.02399	0.00457	0.01247	0.01757	0.01204
$A_{50\times50}$	0.04240	0.04327	0.06968	0.03576	0.08218	0.02653
$A_{100\times100}$	0.22346	0.19254	0.19842	0.09127	0.21662	0.09236
$A_{150\times150}$	1.31892	1.24589	0.35491	0.28424	0.40908	0.24713
$A_{200\times200}$	1.34231	1.53269	1.23860	0.54604	0.70944	0.52773
$A_{250\times250}$	1.46769	1.57901	1.59231	1.21704	1.27971	1.12810
$A_{300\times300}$	2.04023	1.91356	1.64894	1.82464	1.67087	1.66469
Mean	0.71797	0.72682	0.56813	0.44676	0.48830	0.41198

We study the behavior of different methods for finding the well-known Wilson matrix:

$$A = \begin{pmatrix} 10 & 7 & 8 & 7 \\ 7 & 5 & 6 & 5 \\ 8 & 6 & 10 & 9 \\ 7 & 5 & 9 & 10 \end{pmatrix}, \tag{5.1}$$

where we use the stopping termination $\|X_k^2 - I\|_2 \leq 10^{-5}$ in Figure 6. We can see that the results show a stable behavior of the proposed iterative method for finding S.

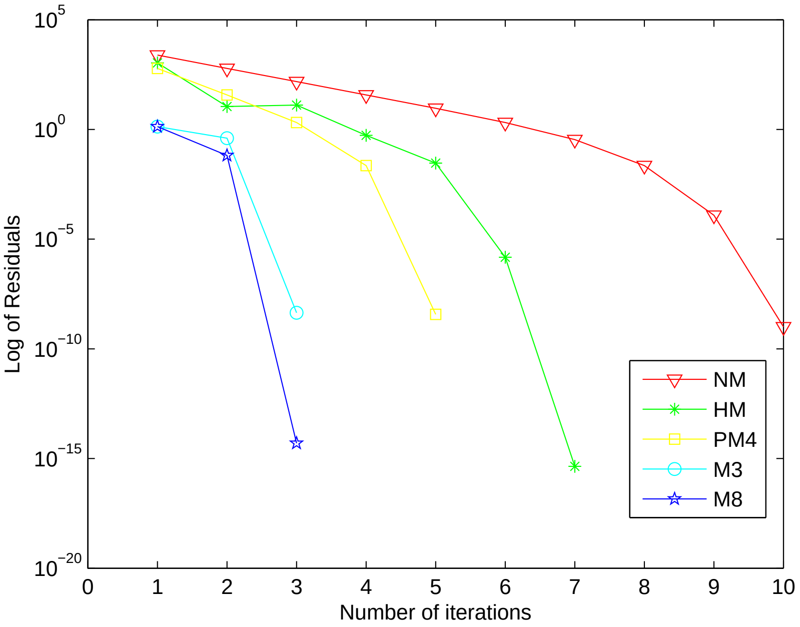


Figure 6. Convergence history of different methods in solving the Wilson matrix.

Now, we investigate the problem of solving the algebraic Riccati equation (ARE) with matrix sign functions. Consider the continuous-time and discrete-time algebraic Riccati equation of the forms

$$XA + A^T X + Q - XBR^{-1}B^T X = 0, \tag{5.2}$$

$$A^T XA + AXB(R + B^T XB)^{-1}B^T XA + Q - X = 0, \tag{5.3}$$

where the quantities $A \in \mathbb{R}^{n\times n}$, $Q = Q^T \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times n}$ are positive semi-definite, $X \in \mathbb{R}^{n\times n}$ is the unknown matrix, $R = R^T \in \mathbb{C}^{n\times n}$ is positive definite. The solutions of ARE are closely related to Symplectic matrices and Hamiltonian matrices [8].

First of all, we introduce the continuous-time algebraic Riccati equation. The stabilizing solution of the *ARE* (5.2) is a real matrix X for which all eigenvalues of $A - BR^{-1}B^T X$ have negative real part and, therefore,

$$\begin{pmatrix} A & BR^{-1}B^T \\ Q & -A^T \end{pmatrix}\begin{pmatrix} I & 0 \\ -X & I \end{pmatrix} = \begin{pmatrix} I & 0 \\ -X & 0 \end{pmatrix}\begin{pmatrix} A - BR^{-1}B^T X & BR^{-1}B^T \\ 0 & -A^T + XBR^{-1}B^T \end{pmatrix}. \tag{5.4}$$

Moreover,

$$H = \begin{pmatrix} A & BR^{-1}B^T \\ Q & -A^T \end{pmatrix}, \tag{5.5}$$

is Hamiltonian matrices. Thus, we have

$$L = \begin{pmatrix} L_{11} & L_{12} \\ L_{21} & L_{22} \end{pmatrix} = sign(H) = \begin{pmatrix} I & 0 \\ -X & I \end{pmatrix}\begin{pmatrix} -I & K \\ 0 & I \end{pmatrix}\begin{pmatrix} I & 0 \\ -X & I \end{pmatrix}^{-1}, \tag{5.6}$$

for a suitable matrix K.

Now, we have X in the following form

$$\begin{pmatrix} L_{11} & L_{12} \\ L_{21} & L_{22} \end{pmatrix}\begin{pmatrix} I \\ -X \end{pmatrix} = \begin{pmatrix} -I \\ X \end{pmatrix}, \tag{5.7}$$

and, thus,

$$-\begin{pmatrix} L_{12} \\ L_{22} \end{pmatrix}X + \begin{pmatrix} L_{11} \\ L_{21} \end{pmatrix} + \begin{pmatrix} -I \\ X \end{pmatrix} = 0, \tag{5.8}$$

which implies

$$\begin{pmatrix} L_{12} \\ L_{22} + I \end{pmatrix}X = \begin{pmatrix} L_{11} + I \\ L_{21} \end{pmatrix}. \tag{5.9}$$

Once the sign of H is calculated, we can solve the overdetermined system (5.9) by using the standard algorithm to get the required solution. Note that the matrix H can not have eigenvalues on the imaginary axis.

For the algebraic Riccati Eq (5.3), we have the Symplectic matrices M as follows:

$$M = \begin{pmatrix} A + BR^{-1}B^T A^{-T}Q & -BR^{-1}B^T A^{-T} \\ A^{-T}Q & A^{-T} \end{pmatrix}. \tag{5.10}$$

The details are not explained here, see more information in [8].

In this test, we use the stopping termination $\|X_k^2 - I\|_\infty \leq 10^{-8}$. We apply the proposed scheme M8 to solve the ARE with the following matrices

$$A = -\frac{1}{82}\begin{pmatrix} 7 & 12 \\ 30 & 28 \end{pmatrix}, B = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \tag{5.11}$$

$$Q = \frac{1}{41}\begin{pmatrix} 474 & -66 \\ -66 & 45 \end{pmatrix}, R = \begin{pmatrix} 10 & -\frac{7}{3} \\ -\frac{7}{3} & 1 \end{pmatrix}. \tag{5.12}$$

We get

$$M = \frac{1}{2}\begin{pmatrix} -5 & 3 & 0 & 3 \\ 138 & -20 & 12 & 0 \\ 372 & -78 & 28 & -30 \\ -150 & 27 & -12 & 7 \end{pmatrix}. \tag{5.13}$$

By using method M8 (leaving 4 decimal places), it can be seen that

$$sign(\widetilde{M}) = \begin{pmatrix} 1.0000 & 0 & 0 & 0 \\ -36.0569 & 18.9889 & -0.0049 & 19.9785 \\ -88.0674 & 35.9626 & -1.0088 & 35.9439 \\ 36.0541 & -17.9895 & 0.0044 & -18.9801 \end{pmatrix}. \tag{5.14}$$

The resulting matrix, which is the solution of (5.3), would be

$$X = \begin{pmatrix} 0.1251 & 0.2500 \\ 0.2504 & 1.6117 \end{pmatrix}. \tag{5.15}$$

From the approximate solution (5.15), we know that method M8 is valid for solving the algebraic Riccati equation. In Table 4, the speed of solving the Riccati equation is further improved by using the M8 method. In Tables 2 and 3, essentially, in terms of the number of iterations and the computational CPU time, they imply that M8 has the best performance in general. We can demonstrate that the proposed method affirms the theoretical parts from the results. From numerical experiments, the proposed method presents consistent convergence behavior.

Table 4. Results of comparisons for the algebraic Riccati equation.

$Method$	NM	HM	$PM4$	$M8$
$Iterations$	6	5	4	2
$Time(s)$	0.011512	0.011809	0.011409	0.010774
$Residual$	2.2446e-009	7.4238e-011	7.7876e-011	7.2760e-011

6. Conclusions

In this paper, we propose a new family of Chebyshev-Halley type iterative method (2.4) with eighth-order. We theoretically prove that the new method (2.4) is convergent and asymptotically stable. Method M8 is a special case of method (2.4) with parameter $a = \frac{3}{4}$, which does not belong to the member of the Padé family. So, method M8 is a new iterative scheme for solving the matrix sign function. Attraction basins in Figures 1–5 are performed to show the convergence behaviors of different methods. From Figure 5, we know that method M8 is globally convergent. Method M8 is applied to a random matrix, Wilson matrix and continuous-time algebraic Riccati equation. Numerical results show that method M8 costs less computing time and requires less number of iterations. This means that method M8 has good convergence behavior.

Acknowledgments

This research was supported by the National Natural Science Foundation of China (No. 61976027), the Natural Science Foundation of Liaoning Province (Nos. 2022-MS-371, 2019-ZD-0502), Educational Commission Foundation of Liaoning Province of China (Nos. LJKMZ20221492, LJKMZ20221498) and the Key Project of Bohai University (No. 0522xn078).

Conflict of interest

The authors declare no conflict of interest.

References

1. M. Hernández, M. Salanova, A family of Chebyshev-Halley type methods, *Int. J. Comput. Math.*, **47** (1993), 59–63. http://dx.doi.org/10.1080/00207169308804162
2. J. Gutiérrez, M. Hernández, A family of Chebyshev-Halley type methods in Banach spaces, *Bull. Austral. Math. Soc.*, **55** (1997), 113–130. http://dx.doi.org/10.1017/S0004972700030586
3. N. Osada, Chebyshev-Halley methods for analytic functions, *J. Comput. Appl. Math.*, **216** (2008), 585–599. http://dx.doi.org/10.1016/j.cam.2007.06.020
4. Y. Kim, R. Behl, S. Motsa, Higher-order efficient class of Chebyshev-Halley type methods, *Appl. Math. Comput.*, **273** (2016), 1148–1159. http://dx.doi.org/10.1016/j.amc.2015.09.013
5. S. Ivanov, Unified convergence analysis of Chebyshev-Halley methods for multiple polynomial zeros, *Mathematics*, **10** (2022), 135. http://dx.doi.org/10.3390/math10010135
6. Z. Bai, J. Demmel, Using the matrix sign function to compute invariant subspaces, *SIAM J. Matrix Anal. Appl.*, **19** (1998), 205–225. http://dx.doi.org/10.1137/S0895479896297719
7. R. Byers, C. He, V. Mehrmann, The matrix sign function method and the computation of invariant subspaces, *SIAM J. Matrix Anal. Appl.*, **18** (1997), 615–632. http://dx.doi.org/10.1137/S0895479894277454
8. C. Kenney, A. Laub, P. Papadopoulos, Matrix-sign algorithms for Riccati equations, *IMA J. Math. Control I.*, **9** (1992), 331–344. http://dx.doi.org/10.1093/imamci/9.4.331
9. N. Higham, *Functions of matrices: theory and computation*, Philadelphia: Society for Industrial and Applied Mathematics, 2008.
10. A. Norris, A. Shuvalov, A. Kutsenko, The matrix sign function for solving surface wave problems in homogeneous and laterally periodic elastic half-spaces, *Wave Motion*, **50** (2013), 1239–1250. http://dx.doi.org/10.1016/j.wavemoti.2013.03.010

11. J. van den Eshof, A. Frommer, T. Lippert, K. Schilling, H. van der Vorst, Numerical methods for the QCD overlap operator. I. Sign-function and error bounds, *Comput. Phys. Commun.*, **146** (2002), 203–224. http://dx.doi.org/10.1016/S0010-4655(02)00455-1

12. P. Benner, E. Quintana-Ortí, Solving stable generalized Lyapunov equations with the matrix sign function, *Numerical Algorithms*, **20** (1999), 75–100. http://dx.doi.org/10.1023/A:1019191431273

13. F. Soleymani, P. Stanimirović, S. Shateyi, F. Khaksar Haghani, Approximating the matrix sign function using a novel iterative method, *Abst. Appl. Anal.*, **2014** (2014), 105301. http://dx.doi.org/10.1155/2014/105301

14. A. Soheili, F. Toutounian, F. Soleymani, A fast convergent numerical method for matrix sign function with application in SDEs, *J. Comput. Appl. Math.*, **282** (2015), 167–178. http://dx.doi.org/10.1016/j.cam.2014.12.041

15. A. Cordero, F. Soleymani, J. Torregrosa, M. Zaka Ullah, Numerically stable improved Chebyshev-Halley type schemes for matrix sign function, *J. Comput. Appl. Math.*, **318** (2017), 189–198. http://dx.doi.org/10.1016/j.cam.2016.10.025

16. X. Wang, W. Li, Stability analysis of simple root seeker for nonlinear equation, *Axioms*, **12** (2023), 215. http://dx.doi.org/10.3390/axioms12020215

17. X. Wang, X. Chen, Derivative-free Kurchatov-type accelerating iterative method for solving nonlinear systems: dynamics and applications, *Fractal Fract.*, **6** (2022), 59. http://dx.doi.org/10.3390/fractalfract6020059

18. D. Jung, C. Chun, X. Wang, Construction of stable and globally convergent schemes for the matrix sign function, *Linear Algebra Appl.*, **580** (2019), 14–36. http://dx.doi.org/10.1016/j.laa.2019.06.019

19. J. Roberts, Linear model reduction and solution of the algebraic Riccati equation by use of the sign function, *Int. J. Control*, **32** (1980), 677–687. http://dx.doi.org/10.1080/00207178008922881

20. L. Shieh, Y. Tsay, C. Wang, Matrix sector functions and their applications to systems theory, *IEE Proceedings D*, **131** (1984), 171–181. http://dx.doi.org/10.1049/ip-d.1984.0029

21. B. Iannazzo, Numerical solution of certain nonlinear matrix equations, Ph. D. Thesis, Università di Pisa, 2007.

22. C. Kenney, A. Laub, Rational iterative methods for the matrix sign function, *SIAM Matrix Anal. Appl.*, **12** (1991), 273–291. http://dx.doi.org/10.1137/0612020

23. M. Misrikhanov, V. Ryabchenko, Matrix sign function in the problems of analysis and design of the linear systems, *Autom. Remote Control*, **69** (2008), 198–222. http://dx.doi.org/10.1134/S0005117908020033

AIMS Press

©2023 the Author(s), licensee AIMS Press. This is an open access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0)