Available online at http://ijim.srbiau.ac.ir/
Int. J. Industrial Mathematics (ISSN 2008-5621)
Vol. 14, No. 2, 2022 Article ID IJIM-1475, 11 pages
DOI: http://dx.doi.org/10.30495/ijim.2022.19413
Research Article

Developing a Stable Method for Computing the Matrix Sign Function with Applications to Algebraic Riccati and Sylvester Equations

P. Ataei Delshad *, T. Lotfi †‡

Received Date: 2020-11-28 Revised Date: 2021-08-22 Accepted Date: 2021-12-08

Abstract

This paper aims to propose a constructive methodology for determining the matrix sign function for a stable variant of the Kung-Traub method. It analytically shows that the new scheme is asymptotically stable. Different numerical experiments compare the new scheme's behavior with the existing matrix iteration of the same type. Finally, the given approach applies to solve the algebraic Riccati equation and the Sylvester equation.

Keywords : Matrix sign function; Kung-Traub method; Algebraic Riccati equation; Stable Sylvester equation.

1 Introduction

In 1858, Cayley [8] introduced the square root of a matrix, and it was not long before the definition of a matrix was proposed by Sylvester and others [37]. Recently, the problem of finding a function f of a matrix A, named by $f(A)$, becoming one of the most studied topics in the field of applied mathematics with widespread applications in science and engineering especially in control theory [7, 10, 13, 16, 17, 14, 29, 32].

One of the fundamental computational problems in control theory is to find the solution of the matrix algebraic Riccati equation. As mathematical models of physical systems get larger, it is vital to develop some reliable and efficient techniques for solving the matrix algebraic Riccati equation. Some of them are: Gardiner and Laub [18], Pandey, Kenney, and Laub [35], Charlier and Van Dooren [11], Gardiner [19].

The other problem in control theory is the Sylvester equation. This equation is applied widely in different fields such as control theory, image restoration, signal processing, model reduction, filtering, decoupling techniques for ordinary and partial differential equations see, e.g., [1, 9, 12, 16]. Bartels-Stewart method [3] and the Hessenberg-Schur method [13, 15] are standard methods for Sylvester equations of the form (7.26). Some iterative schemes for solving Sylvester equation have been proposed in [9, 22, 40]. In [4, 5] the authors investigated the

*Department of Mathematics, Hamedan Branch, Islamic Azad University, Hamedan, Iran.
†Corresponding author. lotfitaher@yahoo.com, Tel:+98(81)344163085.
‡Department of Mathematics, Hamedan Branch, Islamic Azad University, Hamedan, Iran.

numerical solution of stable Sylvester equation via iterative schemes, Newton iteration, Newton-Schulz iteration, Halley method, for computing the sign function of a matrix. Recently, we studied the local convergence analysis of the family of Kung-Traub's two-point method and obtained the convergence ball for this family. Moreover, we studeid the dynamical behavior on quadratic and cubic polynomials for this family [2].

The purpose of the present paper is two topics. One of our intentions is to expand the root-finding Kung-Traub two-point method for the matrix sign function S. Stability of the scheme will be shown analytically. The other aim is to solve the algebraic Ricatti equation and the stable Sylvester equation as an application of the contributed method.

The organization of the paper is as follows. In section 3, some fundamental definitions and properties for Kung-Traub two-point method are presented. Convergence of the method is analysed in section 4, while section 5 is devoted to investigating the stability. The numerical examples for illustrating the method's convergence behavior are devoted to Section 6. Section 7 is dedicated to solving the algebraic Ricatti equation and the Sylvester equation. Section 8 concludes this article with a summary.

2 Theoretical Background

In what follows, we briefly recall the basic definitions and properties of a matrix sign function. A primary matrix function is the matrix sign function. It was introduced by Robert in [34] as a tool for solving the algebraic Riccati equation and the Lyapunov equation. The function of sign for any non-imaginary number z is given as follows.

$$sign(z) = \begin{cases} 1, & Re(z) > 0; \\ -1, & Re(z) < 0. \end{cases} \tag{2.1}$$

It is supposed that $A \in \mathbb{C}^{n\times n}$ does not have any eigenvalues on the imaginary axis. Also, $A = PJP^{-1}$ is the Jordan canonical form arranged where $J = diag(J_1, J_2)$ and the eigenvalues of $J_1 \in \mathbb{C}^{P\times P}$ and the eigenvalues of $J_2 \in \mathbb{C}^{(n-p)\times(n-p)}$ lie in the open left half-plane and the open right half-plane, respectively. Therefore, the matrix sign function of A is defined as

$$S = sign(A) = P \begin{pmatrix} -I_p & 0 \\ 0 & I_{n-p} \end{pmatrix} P^{-1}. \tag{2.2}$$

We can define this matrix uniquely (A is a non-singular square matrix). Certain significant properties of the matrix sign function are outlined in Lemma 2.1.

Lemma 2.1. *(See [4, 10, 23]) Let $A \in \mathbb{C}^{n\times n}$ with no eigenvalues on the imaginary axis. Then the matrix sign function has the following properties*

1. $sign(A)^2 = I$.
2. $sign(A)$ *is diagonalizable with eigenvalues* ± 1.
3. $sign(A)\ A = A\ sign(A)$.
4. *If A is real, then $sign(A)$ is real.*
5. *If A is stable, then*

$$sign(A) = -I_n, \qquad sign(-A) = I_n. \tag{2.3}$$

According to property (2.1) of the previous lemma, solving the following nonlinear matrix equation

$$F(X) = X^2 - I,$$

where I is the identity matrix, by a appropriate root finding method could yield to $S = sign(A)$ if the starting point is chosen as A.

The matrix iteration of Newton, defined as below, is one the most useful and broadly applicable method for computing S.

$$X_{k+1} = \frac{1}{2}(X_k + X_k^{-1}), \tag{2.4}$$

that converges quadratically when $X_0 = A$ chosen as an initial matrix with an ultimately quadratic convergence [34]. Now consider $w_k = x_k + \beta F(x_k)$, $F[x_k, w_k]$ is the two point divided and $\beta \in \mathbb{R} \setminus \{0\}$. We apply the Steffensen family of [38]

$$x_{k+1} = x_k - \frac{F(x_k)}{F[x_k, w_k]}, \quad k = 0, 1, ...$$

in order to compute $sign(A)$. Therefore, we obtain the generalized Steffensen iteration for the matrix sign function ($|\beta| \leq 0.001$) [25]

$$X_{k+1} = \left(I + X_k^2 - \beta X_k + \beta X_k^3\right) \left(2X_k - \beta I + \beta X_k^2\right)^{-1}, \quad k = 0, 1, ... \quad (2.5)$$

The importance of the Steffensen method is in the fact that it has the same order and computational cost as the Newton method.

3 Kung-Traub two-point method

The problem of finding a simple zero of a nonlinear equation $f(x) = 0$, is an often discussed problem in many applications of science and technology [17, 30, 31]. In 1974 Kung and Traub proposed an optimal fourth-order method [28, 33] for finding a simple zero of a nonlinear equation $f(x) = 0$. Let $F : \mathcal{D} \subset \mathbb{X} \to \mathbb{Y}$ be a nonlinear Fréchet differentiable operator in open convex domain $\mathcal{D}$. Let $F'(x_0)^{-1} \in \mathcal{L}(\mathbb{Y}, \mathbb{X})$, where $\mathcal{L}(\mathbb{Y}, \mathbb{X})$ is the set of bounded linear operators from $\mathbb{Y}$ into $\mathbb{X}$. Assume that α is a simple real zero of a real function $F(x)$ and x_0 is an initial approximation to α. The Kung-Traub two-point method can be represented by

$$\begin{cases} y_n = x_n - \frac{F(x_n)}{F'(x_n)}, \\ x_{n+1} = y_n - \frac{F(x_n)^2 F(y_n)}{F'(x_n)\left(F(y_n) - F(x_n)\right)^2}. \end{cases} \quad (3.6)$$

According to property (2.1) of Lemma 2.1, solving the following nonlinear matrix equation

$$F(X) := X^2 - I, \quad (3.7)$$

by a appropriate root finding method could yield to $S = sign(A)$ if the starting point is chosen as A. Now, we consider the Eq.(3.6) to solve the Eq.(3.7) and derive an iterative formula in the reciprocal form as follows.

$$X_{k+1} = \left(I + 3X_k^2 + 23X_k^4 + 5X_k^6\right) \left(2X_k + 12X_k^3 + 18X_k^5\right)^{-1} \quad (3.8)$$

First, we show that the method (3.8) is convergence by using the basin of attraction. In order to indicate this, it is sufficient to plot the basin of attraction of the scheme (3.8) for solving the equation $g(x) = x^2 - 1 = 0$ (for more information see [23] or [24]).

We take square $[-2, 2] \times [-2, 2]$ of the complex plane with a mesh 500×500, while the maximum number of iterations are set to 50 in our written programs. The area of convergence to the roots is painted in sky blue and violet, while the divergence area (if it exists) painted in black (See Figure 1). The exact location of the simple roots of (3.7), i.e. ± 1 is marked with white color.

Figure 1 (b) shows the basins of attraction for (3.8). As you can see we do not have any black region, so the scheme (3.8) is convergence. The local convergence analysis of the family of Kung-Traub's two-point method and the convergence ball for this family are obtained in [2]. Moreover, we studied the dynamical behavior on quadratic and cubic polynomials for this family.

4 Convergence analysis

In this section, we present the theoretical behavior of iterative expression (3.8).

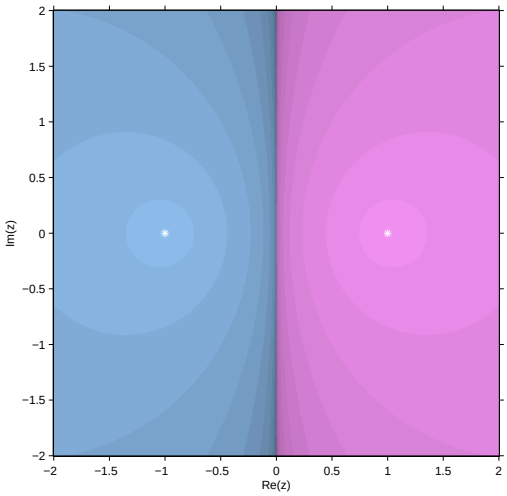
Theorem 4.1. *Assume that $A \in \mathbb{C}^{n\times n}$ possess no pure imaginary eigenvalues. Then, by choosing $X_0 = A$, the matrix sequence $\{X_k\}_{k=0}^{\infty}$ defined by (3.8) is convergent to the matrix sign S.*

Proof. Suppose that R is the rational operator associated to (3.8). If complex matrix $X \in \mathbb{C}^{n\times n}$ has a Jordan canonical form, i.e. there is a matrix Z so that $X = ZJZ^{-1}$, Then

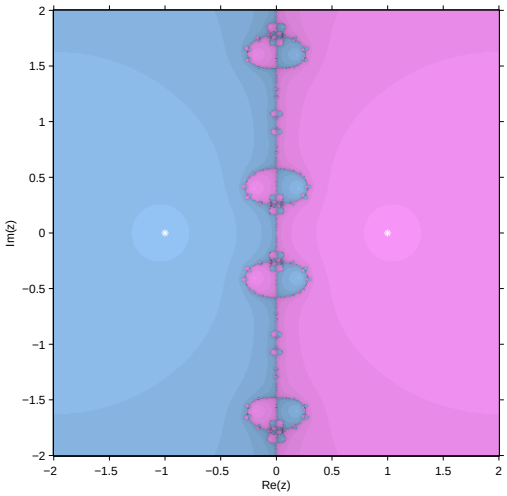
$$R(x) = ZR(J)Z^{-1}.$$

Thus, an eigenvalue λ of X_k gets mapped into the eigenvalue of $R(\lambda)$ of X_{k+1} by applying the matrix iteration (3.8). This scalar relationship between eigenvalues denotes that it is needed to regard how the complex plane is mapped into itself by $R(\lambda)$. The rational operator R must satisfy following properties.

i. Sign preservation:$sign(R(x)) = sign(x)$, $\forall x \in \mathbb{C}$.



(a). The Newton Method



(b). The Kung-Traub Method .

Figure 1: The basins of attractions for the equation $g(x) = x^2 - 1$.

ii. Global convergence: the sequence defined as $x_{k+1} = R(x_k)$ with $x_0 = x$, converges to $sign(x)$ for any x not on the imaginary axis.

Suppose that A has a Jordan canonical form as follows [21]

$$Z^{-1}AZ = \Lambda = \begin{bmatrix} C & 0 \\ 0 & N \end{bmatrix}. \tag{4.9}$$

where Z is a nonsingular matrix and C, N are square Jordan blocks regarding to eigenvalues which are loocated in $\mathbb{C}^-$ and $\mathbb{C}^+$, respectively. Let us consider $\lambda_1, ..., \lambda_p$ and $\lambda_{p+1}, ..., \lambda_n$ are the values locating on the main diagonals of blocks C and N, respectively. By utilizing (4.9), we have

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

Hence, it is clear that

$$sign(\Lambda) = sign(Z^{-1}AZ) = Z^{-1}sign(A)Z$$

Consider $D_0 = Z^{-1}AZ$, we define $D_k = Z^{-1}X_kZ$, $k = 1, 2, ...$, then from the method (3.8), we observe that

$$D_{k+1} = \left(I + 3D_k^2 + 23D_k^4 + 5D_k^6\right) \left[2D_k + 12D_k^3 + 18D_k^5\right]^{-1}. \tag{4.10}$$

It is noteworthy that if D_0 is a diagonal matrix then all successive D_k are diagonal as well. This can be shown by an inductive proof. The case when D_0 is not diagonal can be treated in similar fashion. This will be proved later.

It is sufficient show that $\{D_k\}$ converges to $sign(\Lambda)$. Now, (4.10) is rewritten in the form of n uncoupled scalar iterative methods to solve $f(x) = x^2 - 1 = 0$ as follows:

$$d_{k+1}^i = \frac{1 + 3{d_k^i}^2 + 23{d_k^i}^4 + 5{d_k^i}^6}{2d_k^i + 12{d_k^i}^3 + 18{d_k^i}^5}, \tag{4.11}$$

where $d_k^i = (D_k)_{i,i}$ and $i = 1, ..., n$. Using Eq.(4.10) and Eq.(4.11), we must investigate the convergence of $\{d_k^i\}$ to $sign(\lambda_i)$, for $i = 1, ..., n$.

Because the eigenvalues of A are not pure imaginary and using Eq.(4.11), we get $sign(\lambda_i) = s_i = \pm 1$. Hence, we obtain

$$\frac{d_{k+1}^i - 1}{d_{k+1}^i + 1} = \left(\frac{d_k^i - 1}{d_k^i + 1}\right)^4 \frac{1 + 2d_k^i + 5{d_k^i}^2}{1 - 2d_k^i + 5{d_k^i}^2}. \tag{4.12}$$

Since $|d_0^i| = |\lambda_i| > 0$ and $|\frac{d_0^i - 1}{d_0^i + 1}| < 1$, we have

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

and $\lim_{k\to\infty} |d_k^i| = 1 = |sign(\lambda_i)|$. So we can conclude that $\lim_{k\to\infty} D_k = sign(\Lambda)$.

Now, consider D_0 is not diagonal. Because the Jordan of some matrices may not be diagonal, it is not possible to write Eq.(4.10) as n uncoupled scalar iterations (4.11). In this case, the following relation maps the eigenvalues of X_k from the iteration k to the iteration $k+1$.

$$\lambda_{k+1}^i = \Big(-I - 3{\lambda_k^i}^2 + 23{\lambda_k^i}^4 + 5{\lambda_k^i}^6 \Big) \Big[2\lambda_k^i + 12{\lambda_k^i}^3 + 18{\lambda_k^i}^5 \Big]^{-1}. \tag{4.13}$$

According to the process described above, Eq.(4.13) shows that the eigenvalues are convergent to ± 1 generally, that is to say

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

Finally, we have

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

That is establishing the claim. □

Theorem 4.2. *Assume that $A \in \mathbb{C}^{n\times n}$ has no pure imaginary eigenvalues. Then, by choosing $X_0 = A$, the matrix sequence $\{X_k\}_{k=0}^{\infty}$ defined by (3.8) is convergent to S by fourth rate.*

Proof. The x_k are rational functions of A and hence, like A, commute with S. We know that

$$S^2 = I, \quad S^{-1} = S, \quad S^{2j} = I, \quad S^{2j+1} = S, \text{ for } j \geq 1.$$

Let us consider

$$B_k = 2X_k + 12X_k^3 + 18X_k^5,$$

we observe that

$$\begin{aligned} X_{k+1} - S &= (I + 3X_k^2 + 23X_k^4 + 5X_k^6)B_k^{-1} - S \\ &= \Big(I + 3X_k^2 + 23X_k^4 + 5X_k^6 - SB_k\Big)B_k^{-1} \\ &= \Big(I + 3X_k^2 + 23X_k^4 + 5X_k^6 - 2SX_k \\ &\quad - 12SX_k^3 - 18SX_k^5\Big)B_k^{-1} \\ &= \Big(S^6 - 2S^5X_k + 3S^4X_k^2 - 12S^3X_k^3 + 23S^2X_k^4 \\ &\quad - 18SX_k^5 + 5X_k^6\Big)B_k^{-1} \\ &= (X_k - S)^4\Big(I + X_k(2S + 5X_k)\Big)B_k^{-1}. \end{aligned} \tag{4.14}$$

Now, using any matrix norm from both side of (4.14), we have

$$\|X_{k+1} - S\| \leq \Big(\|B_k^{-1}\|\|I + X_k(2S + 5X_k)\|\Big)\|X_k - S\|^4.$$

The above inequality shows the fourth order of convergence. The proof of the theorem now is clear and completed. □ □

5 Stability

Theorem 5.1. *With identical hypothesis in Theorem 4.2, matrix sequence $\{X_k\}_{k=0}^{\infty}$ produced by (3.8) is stable.*

Proof. If X_0 is a function A, then the iterates form (3.8) are all functions of A and hence commute with A. Let Δ_k be the numerical perturbation presented at the k-th iteration of (3.8). therefore, It can be written as follows

$$\tilde{X}_k = X_k + \Delta_k. \tag{5.15}$$

Here, a first-order error analysis is carried out; that is, we formally neglect quadratic terms such as $(\Delta X_k)^2$, since $(\Delta_k)^i$, $i \geq 2$ is near to zero matrix. This discussion will be significant if Δ_k is small enough. We get

$$\begin{aligned} \tilde{X}_{k+1} &= \Big(I + 3\tilde{X}_k^2 + 23\tilde{X}_k^4 + 5\tilde{X}_k^6\Big) \\ &\quad \Big[2\tilde{X}_k + 12\tilde{X}_k^3 + 18\tilde{X}_k^5\Big]^{-1} \\ &= \Big(I + 3(X_k + \Delta_k)^2 + 23(X_k + \Delta_k)^4 \\ &\quad + 5(X_k + \Delta_k)^6\Big)\Big[2(X_k + \Delta_k) \\ &\quad + 12(X_k + \Delta_k)^3 + 18(X_k + \Delta_k)^5\Big]^{-1}. \end{aligned}$$

For any nonsingular matrix B and C we have the following statement [20]

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

and

$$S^2 = I, \quad and \quad S^{-1} = S.$$

By assuming $X_k \simeq sign(A) = S$ where k is large enough, we obtain

$$\begin{aligned} \tilde{X}_{k+1} &\approx (32I + 97S\Delta_k + 31\Delta_k S) \\ &\quad \left(32S + 32\Delta_k + 96S\Delta_k S\right)^{-1} \\ &\approx (32I + 97S\Delta_k + 31\Delta_k S) \\ &\quad \left(\frac{1}{32}S - \frac{1}{32}S\Delta_k S - \frac{3}{32}\Delta_k\right) \\ &\approx (S + \frac{1}{2}S\Delta_k S - \frac{1}{2}\Delta_k). \end{aligned}$$

Now, after some simplification and by $\Delta_{k+1} = \tilde{X}_{k+1} - X_{k+1}$, we observe that

$$\Delta_{k+1} = \frac{1}{2}(S\Delta_k S - \Delta_k). \tag{5.16}$$

Therefore, we can conclude that the perturbation is bounded at the iteration $k+1$, in other words

$$\|\Delta_{k+1}\| \leq \left(\frac{1}{2}\right)^{k+1} \|S\Delta_0 S - \Delta_0\|.$$

Hence, the sequence $\{X_k\}_{k=0}^{\infty}$ generated by (3.8) is asymptotically stable. The proof is ended. □

6 Numerical Experiments

Here, the result of comparisons in terms of number of iteration and the residual norms is presented for different matrix iterations.

The convergence my be slow if there is a large eigenvalue in iteration X_k , i.e. in the case $\|X_k\| \gg 1$. Therefore, we can speed up the convergence of the proposed iteration through scaling. For this purpose, the scaling parameter μ_k is introduced as follows [26]

$$\mu_k = \begin{cases} \sqrt{\frac{\|X_k^{-1}\|}{\|X_k\|}}, & \text{(norm scaling)}, \\ \sqrt{\frac{\rho(X_k^{-1})}{\rho(X_k)}}, & \text{(spectral scaling,)} \\ \sqrt{|\det(X_k)|^{\frac{-1}{n}}}, & \text{(determinantal scaling.)} \end{cases} \tag{6.17}$$

The new scheme can be expressed as follows

$$\begin{cases} X_0 = A, \\ \mu_k = is\ the\ scaling\ parameter\ computed\ by\ (6.17). \\ X_{k+1} = \left(I + 3\mu_k^2 X_k^2 + 23\mu_k^4 X_k^4 + 5\mu_k^6 X_k^6\right) \\ \left[2\mu_k X_k + 12\mu_k^3 X_k^3 + 18\mu_k^5 X_k^5\right]^{-1}, \end{cases}$$

where $\lim_{k\to\infty} \mu_k = 1$ and $\lim_{k\to\infty} X_k = S$. However, the computation of the scaling parameter μ_k is not studied in depth for the iteration method due to its high cost in some cases. In this work, the stopping termination is considered as follows.

$$\|X_k^2 - I\|_* \leq \epsilon \tag{6.18}$$

where ϵ is the tolerance and $\|.\|_*$ is an appropriate matrix norm. For complex and real input matrix, l_2 and l_∞ should be taken , respectively [36].

In order to comparison, we implement the compared methods, Kung-Traub method abbreviated as KTM, Newton method denoted by NM and Steffensen method denoted by SM1 with $\beta = 0.001$ and SM2 with $\beta = 0.0001$ in Mathematica [39]. The computer specifications are Microsoft Windows 7, 32-bit, Intel(R) Core(TM)i5 CPU 2.27GHz, with 4GB of RAM.

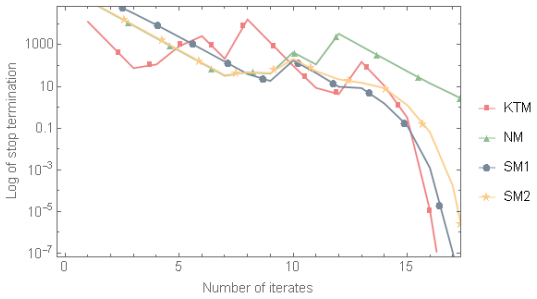


Figure 2: History of convergence of various methods for solving Example 6.1

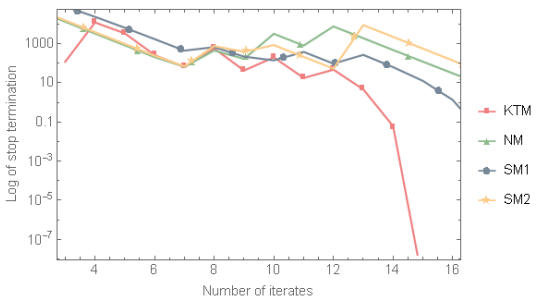


Figure 3: Convergence history of different methods in solving Example 6.2

Example 6.1. *In this example, the behavior of different methods for the following* 250×250 *randomly complex matrix* ($I = \sqrt{-1}$) *is investigated to find the matrix sign function*

$$n = 250;\ SeedRandom[123];$$
$$A = RandomComplex$$
$$[\{\ -100 - I,\ 100 + I\ \}, \{n,\ n\}];$$

The results of comparisons are displayed in Figure 2. In this example, the stopping criterion (6.18) *with* $\epsilon = 10^{-8}$, l_2 *(the input matrix is complex) has been and* $X_0 = A$ *is taken as the initial matrix.*

Example 6.2. *In this test, we run Example 6.1 for the dimension* $n = 400$. *The results in this case are shown in Figure 3*

From these numerical cases, we conclude that the Kung-Traub two-point method produced the best approximation which matches the theoretical fourth order of convergence.

7 Applications

In this section, the iterative method (3.8) is employed to solve the algebraic Riccati equation and the Sylvester equation.

7.1 Algebraic Riccati Equation

Let us consider the following algebraic Riccati equation

$$R(X) = XA + A^T X + Q - XBR^{-1}B^T X = 0, \quad (7.19)$$

where $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, $Q = Q^T \in \mathbb{R}^{n\times n}$ is positive semi-definite, $R = R^T \in \mathbb{R}^{m\times m}$ is positive definite and $X \in \mathbb{R}^{n\times n}$ is the unknown matrix [6, 27, 34]. Generally, the desirable solution is stabilizing because the eigenvalues of $A - BR^{-1}B^T X$ have negative real parts.

Theorem 7.1. *Eq.* (7.19) *can have a unique stabilizing solution* $X \in \mathbb{R}^{n\times n}$ *if* (A, B) *is stabilizable and* (A, Q) *is detectable. Furthermore* X *is symmetric and positive semidefinite.*

Equation (7.19) holds is and only if

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

Now consider

$$W = \begin{pmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{pmatrix} = sign(H) = \begin{pmatrix} I & 0 \\ -X & I \end{pmatrix} \begin{pmatrix} -I & K \\ 0 & I \end{pmatrix} \begin{pmatrix} I & 0 \\ -X & I \end{pmatrix}^{-1}, \quad (7.20)$$

where K is a suitable matrix and

$$H = \begin{pmatrix} A & BR^{-1}B^T \\ Q & -A^T \end{pmatrix} \quad (7.21)$$

Therefore we can find X as follows

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

so

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

Therefore, we have

$$\begin{pmatrix} W_{12} \\ W_{22} + I \end{pmatrix} X = \begin{pmatrix} W_{11} + I \\ W_{21} \end{pmatrix} \quad (7.22)$$

Thus we get the required solution by solving the overdetermined system (7.22). This solution can be computed with the QR decomposition or the method of least squares.

To verify the efficacy of the method we solve a simple example [36]. Consider

$$A = \begin{pmatrix} 2 & -1 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 \\ 0 & -1 & 2 & -1 & 0 \\ 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & -1 & 2 \end{pmatrix},$$

$$B = \begin{pmatrix} 0.8 & 0 & 0 & -1.6 & 0 \\ 0 & 0.8 & 0 & 0 & -1.6 \\ 0 & 0 & 0.8 & 0 & 0 \\ -1.6 & 0 & 0 & 0.8 & 0 \\ 0 & -1.6 & 0 & 0 & 0.8 \end{pmatrix}, \quad (7.23)$$

$$Q = \begin{pmatrix} 4.55719 & 0 & 0 & 0 & 0 \\ 0 & 9.77826 & 0 & 0 & 0 \\ 0 & 0 & 9.43215 & 0 & 0 \\ 0 & 0 & 0 & 9.62216 & 0 \\ 0 & 0 & 0 & 0 & 3.02348 \end{pmatrix},$$

$$R = \begin{pmatrix} 500 & 100 & -200 & 0 & 0 \\ 100 & 600 & -100 & 0 & -200 \\ -200 & -100 & 500 & 0 & -200 \\ 0 & 0 & 0 & 400 & 0 \\ 0 & -200 & -200 & 0 & 400 \end{pmatrix}. \quad (7.24)$$

By attention to Eq.(7.21) , we apply the iterative expression (3.8) to obtain $sign(H)$ by the stop termination (6.18) in the infinity norm and the tolerance 10^{-12}. We solve system (7.22) by the Mathematica function LeastSquares and get

$$X = \begin{pmatrix} 1265.8 & -587.5 & -483.8 & 1027.6 & -448.5 \\ -587.5 & 719.4 & 10.2 & -539.2 & 506.0 \\ -483.8 & 10.2 & 1252.8 & -598.0 & 57.2 \\ 1027.6 & -539.2 & -598.1 & 1349.1 & -672.0 \\ -448.5 & 506.0 & 57.2 & -672.0 & 1129.9 \end{pmatrix} \quad (7.25)$$

Using Eq.(7.25), we compute the residual norm of (7.19) in the infinity norm and we obtain $\|R(X)\|_\infty = 4.03814 \times 10^{-6}$, which confirms the accuracy of the approximation solution using the approach of matrix sign function based on the Kung-Traub two-point method.

7.2 Sylvester Equation

Consider the Sylvester equation

$$R(X) = AX + XB + C = 0, \quad (7.26)$$

where $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{m\times m}$, $C \in \mathbb{R}^{n\times m}$ and $X \in \mathbb{R}^{n\times m}$ is the proper solution. Equation (7.26) has a unique solution if and only if $\alpha + \beta \neq 0$ for all $\alpha \in \Lambda(A)$ and $\beta \in \Lambda(B)$, where $\Lambda(Z)$ symbolizes the spectrum of the matrix Z. This property is established for stable Sylvester equation, while both $\Lambda(A)$ and $\Lambda(B)$ are in the open left half plane. The antistable case can be turned into the stable case by multiplying (7.26) by -1 [4, 5].

In this section, by computation of the matrix sign function, we use the iterative schemes (3.8) for solving Sylvester equations in stable case.

Provided that X is a solution of (7.26), the similarity transformation defined as $\begin{pmatrix} I_n & X \\ 0 & I_m \end{pmatrix}$, can be used to block-diagonalize the block upper triangular matrix

$$H = \begin{pmatrix} A & C \\ 0 & -B \end{pmatrix}, \quad (7.27)$$

as follows

$$\begin{aligned} &\begin{pmatrix} I_n & X \\ 0 & I_m \end{pmatrix}^{-1} \begin{pmatrix} A & C \\ 0 & -B \end{pmatrix} \begin{pmatrix} I_n & X \\ 0 & I_m \end{pmatrix} \\ &= \begin{pmatrix} I_n & -X \\ 0 & I_m \end{pmatrix} \begin{pmatrix} A & C \\ 0 & -B \end{pmatrix} \begin{pmatrix} I_n & X \\ 0 & I_m \end{pmatrix} \\ &= \begin{pmatrix} A & 0 \\ 0 & -B \end{pmatrix}. \end{aligned} \quad (7.28)$$

By utilizing the matrix sign function of H, the relation given in (7.28) and Eq.(2.3), we can derive the following expression for the solution of the Sylvester equation (7.26)

$$sign(H) = \begin{pmatrix} -I_n & 2X \\ 0 & I_m \end{pmatrix}. \quad (7.29)$$

Therefore, in order to solve (7.26), we apply the Kung-Traub two-point schemes suggested for computing sign function.

Now, we solve a simple example to verify the efficacy of the method. First, we construct [4]

$$\begin{aligned} \hat{A} &= diag(-1, -a, -a^2, \cdots, -a^{n-1}), \quad a > 1, \\ \hat{B} &= diag(-1, -b, -b^2, \cdots, -b^{n-1}), \quad b > 1, \\ \hat{C} &= diag(1, 2, 3, \cdots, n), \end{aligned}$$

Here, the spectra of A and B are adjusted by the parameters a and b, respectively.

In the second step, a transformation matrix $K \in \mathbb{R}^{n\times n}$ defined as follows is employed.

$$K = H_2 S H_1, \quad (7.30)$$

where

$$\begin{aligned} H_1 &= I_n - \frac{2}{n} h_1 h_1^T, \qquad h_1 = [1, 1, ..., 1]^T, \\ H_2 &= I_n - \frac{2}{n} h_2 h_2^T, h_2 = [1, -1, ..., (-1)^{n-1}]^T, \\ S &= diag(1, s, \cdots, s^{n-1}), \end{aligned}$$

for transforming the equation matrices as

$$A = (K^{-1})^T \hat{A} K^T, B = K \hat{B} K^{-1}, C = (K^{-1})^T \hat{C} K^{-1}.$$

Here, the scalar s is applied for adjusting the conditioning of K. By setting the parameters $a = 1.03$, $b = 1.008$, $s = 1.001$ and $n = 5$, the solution of the Sylvester equation (7.26) is obtained

$$X = \begin{pmatrix} 1.77 & 0.08 & 0.63 & 0.23 & 0.00 \\ 0.083 & 1.92 & -0.07 & 0.00 & -0.23 \\ 0.61 & -0.07 & 1.44 & 0.07 & -0.61 \\ 0.22 & 0.00 & 0.07 & 0.98 & -0.07 \\ 0.00 & -0.22 & -0.59 & -0.07 & 1.13 \end{pmatrix}.$$

The residual norm of (7.26) in l_∞ is equaled $\|R(X)\|_\infty = 1.99862 \times 10^{-15}$, which confirm the accuracy of the approximation approach of matrix sign function and the method (3.8). Figure 4 shows the accuracy of the Kung-Traub iteration for different values of n.

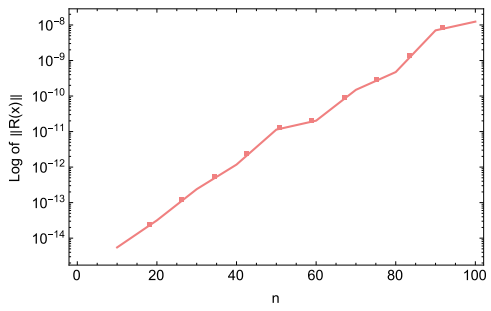


Figure 4: Relative errors of the Sylvester equation solvers for different n.

8 Conclusions

This paper devoted the Kung-Traub method to the computation of the matrix sign function. We showed that this method is convergence via attraction basin in the complex plane. Some numerical examples performed the contributed method's consistency and efficiency. Moreover, we discussed applying the sign function method for solving the algebraic Riccati equation and a class of the stable Sylvester equation. The numerical results are well in line with the theoretical aspects.

References

[1] F. A. Aliev, V. B. Larin, Optimization of Linear Control Systems: Analytical Methods and Computational Algorithms, volume 8 of Stability and Control: Theory, *Methods and Applications, Gordon and Breach*, 1998.

[2] P. Ataei Delshad, T. Lotfi, On the local convergence of Kung-Traubs two-point method and its dynamics, *Appl Math* 65 (2020) 379-406. http://dx.doi.org/10.21136/AM.2020.0322-18/

[3] R. H. Bartels, G. W. Stewart, Solution of the matrix equation $AX + XB = C$, *Algorithm 432. Comm. ACM.* 15 (1972) 820-826.

[4] P. Benner, E. S. Quintana-Ortí, G. Quintana-Ortí, Solving Stable Sylvester Equations via Rational Iterative Schemes, *Journal of Scientific Computing* 28 (2006) 51-83. http://dx.doi.org/10.1007/s10915-005-9007-2/

[5] P. Benner, Factorized Solution of Sylvester Equations with Applications in Control, *In Proc. of the 16th International Symposium on Mathematical Theory of Network and Systems*, 2004.

[6] A. Bunse-Gerstner, Computational solution of the algebraic Riccati equation, *Journal of the Society of Instrument and Control Engineers (SICE)* 38 (1996) 632-639.

[7] S. Barrachina, P. Benner, E. S. Quintana-Ortí, Efficient algorithms for generalized algebraic Bernoulli equations based on the matrix sign function,*Numerical Algorithms* 46 (2007) 351-368.

[8] A. Cayley, A memoir on the theory of matrices, *Philos. Trans. Roy. Soc. London* 148 (1858) 17-37.

[9] D. Calvetti, L. Reichel, Application of ADI iterative methods to the restoration of noisy images, *SIAM J. Matrix Anal. Appl.* 17 (1996) 165-186.

[10] A. Cordero, F. Soleymani, J. R.Torregrosa, M. ZakaUllah, Numerically stable improved Chebyshev-Halley type schemes for matrix sign function, *Journal of Computational and Applied Mathematics* 318 (2017) 189-198.

[11] J. P. Charlier, P. Van Dooren, A systolic algorithm for Riccati and Lyapunov equations, *Math. Control Signals Systems* 2 (1989) 109-136.

[12] L. Dieci, M. R. Osborne, R. D. Russell, A Riccati transformation method for solving linear bvps, I: Theoretical aspects, *SIAM J. Numer. Anal.* 25 (1988) 1055-1073.

[13] W. H. Enright, Improving the efficiency of matrix operations in the numerical solution of stiff ordinary differential equations, *ACM Trans. Math. Softw.* 4 (1987) 127-136.

[14] F. Filbir, Computation of the structured stability radius via matrix sign function, *Systems and Control Letters* 22 (1994) 341-349.

[15] G. H. Golub, S. Nash, C. F. Van Loan, A Hessenberg-Schur method for the problem $AX+XB = C$, *IEEE Trans. Automat. Control AC.* 24 (1979) 909-913.

[16] G. H. Golub, C. F. Van Loan, Matrix Computations, *Johns Hopkins University Press, Baltimore, third edition*, 1996.

[17] M. Ghorbanzadeh, K. Mahdiani, F. Soleymani, T. Lotfi, A Class of Kung-Traub-Type Iterative Algorithms for Matrix Inversion, *International Journal of Applied and Computational Mathematics* 2 (2016) 641-648.

[18] J. D. Gardiner, A. J. Laub, Parallel algorithms for algebraic Riccati equations, *Internat. J. Control* 54 (1991) 1317-1333.

[19] J. D. Gardiner, A Stabilized matrix sign function algorithm for solving algebraic Riccati equations, *SIAM J. SCI. COMPUT* 18 (1997) 1393-1411.

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

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

[22] D. Y. Hu, L. Reichel, Application of ADI iterative methods to the restoration of noisy images, *Linear Algebra Appl.* 172 (1992) 283-313.

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

[24] B. Iannazzo, A family of rational iterations and its application to the computation of the matrix pth root, *SIAM J. Matrix Anal. Appl.* 30 (2008) 1445-1462.

[25] F. Khaksar Haghani, A generalized Steffensen's method for matrix sign function, *Applied Mathematics and Computation* 260 (2015) 249-256.

[26] C. S. Kenney, A. J. Laub, On scaling Newton's method for polar decomposition and the matrix sign function, *SIAM J. Matrix Anal. Appl.* 13 (1992) 688-706.

[27] C. S. Kenney, A. J. Laub, P. M. Papadopoulos, Matrix-sign algorithms for Riccati equations, *IMA J. Math. Cont. Infor.* 9 (1992) 331-344.

[28] H. T. Kung, J. F. Traub, Optimal order of one-point and multi-point iteration, *J. ACM.* 21 (1974) 643-651.

[29] N. Kyurkchiev, A. Iliev, A refinement of some overre-laxation algorithms for solving a system of linear equations, *Serdica Journal of Computing* 7 (2013) 245-256.

[30] T. Lotfi, S. Sharifi, M. Salimi, S. Siegmund, A new class of three-point methods with optimal convergence order eight and its dynamics, *Numerical examples* 68 (2015) 261-288.

[31] T. Lotfi, P. Bakhtiari, A. Cordero, K. Mahdiani, J. R. Torregrosa, Some new efficient

multipoint iterative methods for solving nonlinear systems of equations, *International Journal of Computer Mathematics* 92 (2015) 1921-1934

[32] M. S. Misrikhanov, V. N. Ryabchenko, A matrix sign function in problems of the analysis and design of linear systems, *Automation and Remote Control* 69 (2008) 198-222.

[33] M. S. Petković, B. Neta, L. D. Petković, J. Džunić, Multipoint Methods for Solving Nonlinear Equations, *Elsevier, New York*, 2013.

[34] J. D. Roberts, Linear model reduction and solution of the algebraic Riccati equation by use of the sign function,*Reprint of Technical Report No. TR-13, CUED/B-Control, Cambridge University, Engineering Department*, 1971.

[35] P. Pandey, C. Kenney, A. J. Laub, parallel algorithm for the matrix sign function, *Internat. J. High Speed Computing* 2 (1990) 181-191.

[36] A. R. Soheili, F. Toutounian, F. Soleymani, A fast convergent numerical method for matrix sign function with application in SDEs, *Journal of Computational and Applied Mathematics* 282 (2015) 167-178.

[37] J. J. Sylvester, Additions to the articles, "On a New Class of Theorems," and "On Pascal's Theorem", *Philosophical Magazine* 37 (1850) 363-370.

[38] J. F. Steffensen, Remarks on iteration, *Skandinavisk Aktuarietidskrift* 16 (1933) 64-72.

[39] M. Trott, The Mathematica Guide book for Numerics, *Springer, NewYork, NY, USA*, 2006.

[40] E. L. Wachspress, Iterative solution of the Lyapunov matrix equation, *Appl. Math. Letters* 107 (1988) 87-90.

Parandoosh Ateai Delshad has got her B.S degree in Mathematics from Islamic Azad University, Hamedan, Iran in 2006 , and her M.S degree in Mathematics from Islamic Azad University, Science and Research Branch, Tehran Iran in 2011. She has got the PHD degree from Islamic Azad University, Hamedan, Iran, in 2020. Her research interest is local and semi-local convergence, dynamical stability, approximating numerical solutions and Fuzzy concepts. She has published some articles nationally and internationally in this field.

Taher Lotfi is an associate professor in applied mathematics and numerical analysis. His research interest includes iterative methods for approximating numerical solution of nonlinear systems of equations.