The Matrix Sign Function Method and the Computation of Invariant Subspaces

Ralph Byers * Chunyang He† Volker Mehrmann †

November 5. 1994

Abstract

A perturbation analysis shows that if a numerically stable procedure is used to compute the matrix sign function, then it is competitive with conventional methods for computing invariant subspaces. Stability analysis of the Newton iteration improves an earlier result of Byers and confirms that ill-conditioned iterates may cause numerical instability. Numerical examples demonstrate the theoretical results.

1 Introduction

If $A \in \mathbf{R}^{n \times n}$ has no eigenvalue on the imaginary axis, then the matrix sign function $\mathrm{sign}(A)$ may be defined as

$$\mathrm{sign}(A) = \frac{1}{\pi i} \int_{\gamma} (zI - A)^{-1} dz - I, \tag{1}$$

where γ is any simple closed curve in the complex plane enclosing all eigenvalues of A with positive real part. The sign function is used to compute eigenvalues and invariant subspaces [3, 5, 7, 10, 11] and to solve Riccati and Sylvester equations [9, 12, 13, 24]. The matrix sign function is attractive for machine computation, because it can be efficiently evaluated by relatively

*University of Kansas, Dept. of Mathematics, Lawrence, Kansas 66045, USA. Partial support received from National Science Foundation grant INT-8922444. Some support was also received from University of Kansas General Research Allocation 3514-20-0038.

†TU-Chemnitz-Zwickau, Fak. f. Mathematik, D-09107 Chemnitz, FRG. Partial support was received from Deutsche Forschungsgemeinschaft, Projekt La 767/3-2.

simple numerical methods. Some of these are surveyed in [24]. It is particularly attractive for large dense problems to be solved on computers with advanced architectures [3, 13, 28].

Beavers and Denman use the following equivalent definition [7, 10]. Let $A = XJX^{-1}$ be the Jordan canonical decomposition of a matrix A having no eigenvalues on the imaginary axis. Let the diagonal part of J be given by the matrix $D = \mathrm{diag}(d_1, \ldots, d_n)$. If $S = \mathrm{diag}(s_1, \ldots, s_n)$, where

$$s_i = \begin{cases} +1 & \text{if } \Re(d_i) > 0, \\ -1 & \text{if } \Re(d_i) < 0, \end{cases}$$

then $\mathrm{sign}(A)$ is given by $\mathrm{sign}(A) = XSX^{-1}$.

Let $\mathcal{V}^+ = \mathcal{V}^+(A)$ be the invariant subspace of A corresponding to eigenvalues with positive real part, let $\mathcal{V}^- = \mathcal{V}^-(A)$ be the invariant subspace of A corresponding to eigenvalues with negative real part, let $P^+(A) = P^+$ be the skew projection onto $\mathcal{V}^+$ parallel to $\mathcal{V}^-$, and let $P^- = P^-(A)$ be the skew projection onto $\mathcal{V}^-$ parallel to $\mathcal{V}^+$. Using the same contour γ as in (1), the projection P^+ has the resolvent integral representation [19, Page 67] [2]

$$P^+ = \frac{1}{2\pi i} \int_\gamma (zI - A)^{-1} dz. \tag{2}$$

It follows from (1) and (3) that $\mathrm{sign}(A) = P^+ - P^- = 2P^+ - I = I - 2P^-$.

The matrix sign function was introduced using definition (1) by Roberts in a 1971 technical report [29] which was not published until 1980 [30]. Kato [19, Page 67] reports that the resolvent integral (2) goes back to 1946 [35] and 1949 [17, 18].

There is some concern about the numerical stability of numerical methods based upon the matrix sign function [3, 8, 16]. In this paper, we demonstrate that evaluating the matrix sign function is a more ill-conditioned computational problem than the problem of finding bases of the invariant subspaces $\mathcal{V}^+$ and $\mathcal{V}^-$. (Sometimes it is tremendously more ill-conditioned. See Example 1 in Section 3.) Never-the-less, we also give perturbation and error analyses, which show that (at least for Newton's method for the computation of the matrix sign function [8, 9]) *in most circumstances* the accuracy is competitive with conventional methods for computing invariant subspaces. Our analysis improves some of the perturbation bounds in [4, 8, 15, 20].

In Section 2 we establish some notation and clarify the relationship between the matrix sign function and the Schur decomposition. The next two sections give a perturbation analysis of the matrix sign function and its invariant subspaces. Section 5 gives *a posteriori* bounds on the forward and backward error associated with a corrupted value of $\mathrm{sign}(S)$. Section 6 is a stability analysis of the Newton iteration. Section 7 demonstrates the results with some numerical examples.

Throughout the paper, $\|\cdot\|$ denotes the spectral norm, $\|\cdot\|_1$ the 1-norm (or column sum norm), and $\|\cdot\|_F$ the Frobenius norm $\|\cdot\|_F = \sqrt{\sum |a_{ij}|^2}$. The set of eigenvalues of a matrix A is denoted by $\lambda(A)$. The open left half plane is denoted by $\mathbf{C}^-$ and the open right half plane is denoted by $\mathbf{C}^+$. Borrowing some terminology from engineering, we refer to the invariant subspace $\mathcal{V}^- = \mathcal{V}^-(A)$ of a matrix $A \in \mathbf{R}^{n\times n}$ corresponding to eigenvalues in $\mathbf{C}^-$ as the *stable invariant subspace* and the subspace $\mathcal{V}^+ = \mathcal{V}^+(A)$ corresponding to eigenvalues in $\mathbf{C}^+$ as the *unstable invariant subspace*. We use $P^+ = P^+(A)$ for the the skew projection onto $\mathcal{V}^+$ parallel to $\mathcal{V}^-$ and $P^- = P^-(A)$ for the skew projection onto $\mathcal{V}^-$ parallel to $\mathcal{V}^+$.

2 Relationship with the Schur Decomposition

Suppose that A has the Schur form

$$Q^H A Q = \begin{array}{c c} & \begin{array}{cc} k & n-k \end{array} \\ \begin{array}{c} k \\ n-k \end{array} & \begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix} \end{array}, \tag{3}$$

where $\lambda(A_{11}) \subset \mathbf{C}^-$ and $\lambda(A_{22}) \subset \mathbf{C}^+$ [14]. If Y is a solution of the Sylvester equation

$$YA_{22} - A_{11}Y = 2A_{12}, \tag{4}$$

then

$$Q^H \mathrm{sign}(A) Q = \begin{array}{c c} & \begin{array}{cc} k & n-k \end{array} \\ \begin{array}{c} k \\ n-k \end{array} & \begin{bmatrix} -I & Y \\ 0 & I \end{bmatrix} \end{array}, \tag{5}$$

$$Q^H P^- Q = \begin{array}{c c} & \begin{array}{cc} k & n-k \end{array} \\ \begin{array}{c} k \\ n-k \end{array} & \begin{bmatrix} I & -\frac{1}{2}Y \\ 0 & 0 \end{bmatrix} \end{array},$$

and

$$Q^H P^+ Q = \begin{array}{c c} & \begin{array}{cc} k & n-k \end{array} \\ \begin{array}{c} k \\ n-k \end{array} & \begin{bmatrix} 0 & \frac{1}{2}Y \\ 0 & I \end{bmatrix} \end{array}.$$

The solution of (4) has the integral representation

$$Y = \frac{1}{\pi i} \int_\gamma (zI - A_{11})^{-1} A_{12} (zI - A_{22})^{-1} dz, \tag{6}$$

where γ is a closed contour containing all eigenvalues of A with positive real part [25, 31]).

The stable invariant subspace of A is the range (or column space) of $\text{sign}(A) - I = -2P^-$. If

$$(\text{sign}(A) - I)\Pi = QR = \begin{bmatrix} Q_1 & Q_2 \end{bmatrix} \begin{bmatrix} R_1 & R_2 \\ 0 & 0 \end{bmatrix} \tag{7}$$

is a QR factorization with column pivoting [1, 14], then the columns of Q_1 form an orthonormal basis of this subspace. Here Q is orthogonal, Π is a permutation matrix, R is upper triangular, and R_1 is nonsingular.

It is not difficult to use the singular value decomposition of Y to show that [4]

$$\|\text{sign}(A)\| = \frac{1}{2}\|Y\| + \sqrt{1 + \frac{1}{4}\|Y\|^2}. \tag{8}$$

It follows from (4) that

$$\|Y\| \leq \frac{2\|A_{12}\|}{\text{sep}(A_{11}, A_{22})}, \tag{9}$$

where sep is defined as in [14] by $\text{sep}(A_{11}, A_{22}) = \min_{Z \neq 0} \frac{\|A_{11}Z - ZA_{22}\|_F}{\|Z\|_F}$.

3 The Effect of Backward Errors

In this section we discuss the sensitivity of the matrix sign function subject to perturbations. Based on Fréchet derivatives, Kenney and Laub [20] presented a first order perturbation theory for the matrix sign function via the solution of a Sylvester equation. Mathias [26] derives an expression for the Fréchet derivative using the Schur decomposition. Kato's encyclopedic monograph [19] includes an extensive study of series representations and of perturbation bounds for eigenprojections. In this section we derive an expression for the Fréchet derivative using integral formulas.

For a perturbation matrix E, we give estimates for $\operatorname{sign}(A+E)$ in terms of powers of $\|E\|$. Partition E conformally with (3) as

$$Q^H E Q = \begin{array}{c} \\ k \\ n-k \end{array} \!\!\begin{array}{c} \begin{array}{cc} k & n-k \end{array} \\ \begin{bmatrix} E_{11} & E_{12} \\ E_{21} & E_{22} \end{bmatrix} \end{array}. \tag{10}$$

Consider first the relatively simple case in which A is block diagonal.

Lemma 1 *Suppose A is block diagonal,*

$$A = \begin{bmatrix} A_{11} & 0 \\ 0 & A_{22} \end{bmatrix}$$

where $\lambda(A_{11}) \subset \mathbf{C}^-$ and $\lambda(A_{22}) \subset \mathbf{C}^+$. Partition the perturbation $E \in \mathbf{R}^{n\times n}$ conformally with A as

$$E = \begin{bmatrix} E_{11} & E_{12} \\ E_{21} & E_{22} \end{bmatrix}. \tag{11}$$

If $\|E\|$ is sufficiently small, then

$$\operatorname{sign}(A+E) = \operatorname{sign}(A) + 2\left(\begin{bmatrix} 0 & F_{12} \\ F_{21} & 0 \end{bmatrix}\right) + O(\|E\|^2)$$

where F_{12} and F_{21} satisfy the Sylvester equations

$$\begin{aligned} A_{22}F_{21} - F_{21}A_{11} &= E_{21} && (12)\\ F_{12}A_{22} - A_{11}F_{12} &= E_{12}. && (13) \end{aligned}$$

Proof. Note that the eigenvalues of $A_{11} + E_{11}$ have negative real part and the eigenvalues of $A_{22} + E_{22}$ have positive real part. In the definition (1) choose the contour γ to enclose $\lambda(A_{22})$ and $\lambda(A_{22} + E_{22})$ but neither $\lambda(A_{11})$ nor $\lambda(A_{11} + E_{11})$. So,

$$\begin{aligned} \operatorname{sign}(A+E) &= \frac{1}{\pi i}\int_\gamma (zI - (A+E))^{-1}\,dz - I \\ &= \frac{1}{\pi i}\int_\gamma ((zI-A)^{-1} + (zI-A)^{-1}E(zI-A)^{-1})\,dz - I \\ &\quad + O(\|E\|^2) \\ &= \operatorname{sign}(A) + 2F + O(\|E\|^2), \end{aligned}$$

where

$$F = \frac{1}{2\pi i} \int_\gamma (zI - A)^{-1} E (zI - A)^{-1} \, dz.$$

Partitioning F conformally with E and A, then we have

$$\begin{aligned}
F_{11} &= \frac{1}{2\pi i} \int_\gamma (zI - A_{11})^{-1} E_{11} (zI - A_{11})^{-1} \, dz \\
F_{12} &= \frac{1}{2\pi i} \int_\gamma (zI - A_{11})^{-1} E_{12} (zI - A_{22})^{-1} \, dz \\
F_{21} &= \frac{1}{2\pi i} \int_\gamma (zI - A_{22})^{-1} E_{21} (zI - A_{11})^{-1} \, dz \\
F_{22} &= \frac{1}{2\pi i} \int_\gamma (zI - A_{22})^{-1} E_{22} (zI - A_{22})^{-1} \, dz.
\end{aligned}$$

As in (6), F_{12} and F_{21} are the solutions to the Sylvester equations (12) and (13) [25, 31]. The contour γ encloses no eigenvalues of A_{11}, so $(zI - A_{11})^{-1} E_{11} (zI - A_{11})^{-1}$ is analytic inside γ and $F_{11} = 0$.

We first prove that $F_{22} = 0$ in the case that A_{22} is diagonalizable, say $A_{22} = X \Lambda X^{-1}$ where $\Lambda = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_{n-k})$. Then

$$F_{22} = X \left(\frac{1}{2\pi i} \int_\gamma (zI - \Lambda)^{-1} (X^{-1} E_{22} X)(zI - \Lambda)^{-1} \, dz \right) X^{-1}.$$

Each component of the above integral is of the form $\int_\gamma c(z - \lambda_j)^{-1}(z - \lambda_k)^{-1} \, dz$ for some constant c. If $j = k$ then this is the integral of a residue free holomorphic function and hence it vanishes. If $j \neq k$, then

$$\int_\gamma \frac{c}{(z - \lambda_i)(z - \lambda_j)} \, dz = \int_\gamma \frac{c}{\lambda_i - \lambda_j} \left(\frac{1}{z - \lambda_i} - \frac{1}{z - \lambda_j} \right) dz = 0.$$

The general case follows by taking limits of the diagonalizable case and using the dominated convergence theorem [36]. $\Box$

Theorem 1 *Let the Schur form of A be given as in (3) and let E be as in (10). If $\|E\|$ is sufficiently small, then*

$$\mathrm{sign}(A + E) = \mathrm{sign}(A) + E_t - \mathrm{sign}(A) E_p \, \mathrm{sign}(A) + O(\|E\|^2),$$

where

$$\begin{aligned} E_t &= Q \begin{bmatrix} 0 & 2\tilde{E}_{12} + \frac{Y\tilde{E}_{21}Y}{2} \\ \tilde{E}_{21} & 0 \end{bmatrix} Q^H \\ E_p &= Q \begin{bmatrix} 0 & 0 \\ \tilde{E}_{21} & 0 \end{bmatrix} Q^H, \end{aligned}$$

$\tilde{E}_{21}$ satisfies the Sylvester equation

$$A_{22}\tilde{E}_{21} - \tilde{E}_{21}A_{11} = E_{21}, \tag{14}$$

Y satisfies (4), and $\tilde{E}_{12}$ satisfies

$$\tilde{E}_{12}A_{22} - A_{11}\tilde{E}_{12} = E_{12} - \frac{YE_{22}}{2} + \frac{E_{11}Y}{2} - \frac{YE_{21}Y}{4}. \tag{15}$$

Proof. If $S = \begin{bmatrix} I & -\frac{Y}{2} \\ 0 & I \end{bmatrix}$, then

$$S \begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix} S^{-1} = \begin{bmatrix} A_{11} & 0 \\ 0 & A_{22} \end{bmatrix}$$

and

$$S \begin{bmatrix} E_{11} & E_{12} \\ E_{21} & E_{22} \end{bmatrix} S^{-1} = \begin{bmatrix} E_{11} - \frac{YE_{21}}{2} & E_{12} - \frac{YE_{22}}{2} + \frac{E_{11}Y}{2} - \frac{YE_{21}Y}{4} \\ E_{21} & \frac{E_{21}Y}{2} + E_{22} \end{bmatrix}.$$

It follows from Lemma 1 that

$$\operatorname{sign}(SQ^H(A+E)QS^{-1}) = \begin{bmatrix} -I & 0 \\ 0 & I \end{bmatrix} + 2 \begin{bmatrix} 0 & \tilde{E}_{12} \\ \tilde{E}_{21} & 0 \end{bmatrix} + O(\|E\|^2).$$

Since $\operatorname{sign}(SAS^{-1}) = S \operatorname{sign}(A) S^{-1}$, multiplying QS^{-1} on the left side and SQ^H on the right side of the above equation, we have

$$\operatorname{sign}(A+E) = \operatorname{sign}(A) + Q \begin{bmatrix} Y\tilde{E}_{21} & 2\tilde{E}_{12} - \frac{Y\tilde{E}_{21}Y}{2} \\ 2\tilde{E}_{21} & -\tilde{E}_{21}Y \end{bmatrix} Q^H + O(\|E\|^2). \tag{16}$$

It is easy to verify that

$$\begin{bmatrix} Y\tilde{E}_{21} & 2\tilde{E}_{12} - \frac{Y\tilde{E}_{21}Y}{2} \\ 2\tilde{E}_{21} & -\tilde{E}_{21}Y \end{bmatrix} = \tag{17}$$

$$\begin{bmatrix} 0 & 2\tilde{E}_{12} + \frac{Y\tilde{E}_{21}Y}{2} \\ \tilde{E}_{21} & 0 \end{bmatrix} - \begin{bmatrix} -I & Y \\ 0 & I \end{bmatrix} \begin{bmatrix} 0 & 0 \\ \tilde{E}_{21} & 0 \end{bmatrix} \begin{bmatrix} -I & Y \\ 0 & I \end{bmatrix}.$$

The theorem follows from

$$Q^H \operatorname{sign}(A) Q = \operatorname{sign}(Q^H A Q) = \begin{bmatrix} -I & Y \\ 0 & I \end{bmatrix}.$$

□

Of course Theorem 1 also gives first order perturbations for the projections $P^+ = P^+(A)$ and $P^- = P^-(A)$.

Corollary 1 *Let the Schur form of A be given as in (3) and let E be as in (10). If $\|E\|$ is sufficiently small, then*

$$\begin{aligned} & P^{\pm}(A+E) \\ = \quad & P^{\pm}(A) + \frac{1}{2}\left(E_t - \operatorname{sign}(A) E_p \operatorname{sign}(A)\right) + O(\|E\|^2) \\ = \quad & P^{\pm}(A) + \frac{1}{2}\left(E_t - \left(P^{\pm}(A) - P^{\mp}(A)\right) E_p \left(P^{\pm}(A) - P^{\mp}(A)\right)\right) \\ & + O(\|E\|^2) \\ = \quad & P^{\pm}(A) + \frac{1}{2}\left(E_t - \left(2P^{\pm}(A) - I\right) E_p \left(2P^{\pm}(A) - I\right)\right) + O(\|E\|^2) \end{aligned}$$

where E_t and E_p are as in the statement of Theorem 1.

Taking norms in Theorem 1 gives first order perturbation bounds.

Corollary 2 *Let the Schur form of A be given as in (3), E as in (10) and let $0 < \delta = \operatorname{sep}(A_{11}, A_{22})$, then the first order perturbation of the matrix sign function stated in Theorem 1 is bounded by*

$$\|E_t - \operatorname{sign}(A) E_p \operatorname{sign}(A)\| \le \frac{4}{\delta}\left(1 + \frac{\|A_{12}\|}{\delta}\right)^2 \|E\|.$$

The corollary follows from the sum of the above bounds. □

On first examination, Corollary 2 is discouraging. It shows that calculating the matrix sign function may be more ill-conditioned than finding bases of the stable and unstable invariant subspace. If the matrix A whose Schur decomposition appears in (10) is perturbed to $A+E$, then the stable invariant subspace, $\operatorname{Im}(Q_1)$, is perturbed to $\operatorname{Im}(Q_1 + Q_2 E_q)$ where $\|E_q\| \le 2\|E\|/\delta$ [32, 34]. Corollary 2 and the following example show that $\|\operatorname{sign}(A+E)\|$ may differ from $\operatorname{sign}(A)$ by a factor of δ^{-3} which may be much larger than $\|E\|/\delta$.

Example 1 Let

$$\begin{aligned} A &= \begin{bmatrix} -\eta & 1 \\ 0 & \eta \end{bmatrix} \\ E &= \begin{bmatrix} 0 & 0 \\ \epsilon & 0 \end{bmatrix}. \end{aligned}$$

The matrix A is already in Schur form, so $\mathrm{sep}(A_{11}, A_{22}) = 2\eta$. If $\epsilon < \eta < 1$, then we have

$$\begin{aligned} \mathrm{sign}(A) &= \begin{bmatrix} -1 & \eta^{-1} \\ 0 & 1 \end{bmatrix} \\ \mathrm{sign}(A+E) &= \frac{1}{\sqrt{\eta^2+\epsilon}} \begin{bmatrix} -\eta & 1 \\ \epsilon & \eta \end{bmatrix}. \end{aligned}$$

The difference is

$$\mathrm{sign}(A+E) - \mathrm{sign}(A) = \epsilon \begin{bmatrix} \eta^{-2}/2 & -\eta^{-3}/2 \\ \eta^{-1} & -\eta^{-2}/2 \end{bmatrix} + O(\epsilon^2).$$

Perturbing A to $A+E$ does indeed perturb the matrix sign function by a factor of δ^{-3}.

Of course there is no rounding error in Example 1, so the stable invariant subspace of $A+E$ is also the stable invariant subspace of $\mathrm{sign}(A+E)$ and, in particular, evaluating the matrix sign function exactly has done no more damage than perturbing A. The stable invariant subspace of A is $\mathcal{V}^-(A) = \mathrm{Im}(\begin{bmatrix} 1 \\ 0 \end{bmatrix})$; the stable invariant subspace of $A+E$ and $\mathrm{sign}(A+E)$ is

$$\mathcal{V}^-(A+E) = \mathrm{Im}(\begin{bmatrix} 1 \\ \frac{-\epsilon}{\eta+\sqrt{\eta^2+\epsilon}} \end{bmatrix}) = \mathrm{Im}(\begin{bmatrix} 1 \\ \frac{-\epsilon}{2\eta} \end{bmatrix}) + O(\epsilon^2).$$

For a general small perturbation matrix E, the angle between $\mathcal{V}^-(A)$ and $\mathcal{V}^+(A+E)$ is of order no larger than $O(1/\eta)$ [14, 32, 34]. *The matrix sign function (and the projections P^- and P^+) may be significantly more ill-conditioned than the stable and unstable invariant subspaces.* Never-the-less, we argue in the next section that despite the possible poor conditioning of the matrix sign function, the invariant subspaces are usually preserved about as accurately as their native conditioning permits.

4 Perturbation Theory for Invariant Subspaces of the Matrix Sign Function

In this section we discuss the accuracy of the computation of the stable invariant subspace of A via the matrix sign function.

An easy first observation is that if a backward stable method was used to compute the matrix sign function, then the computed value of $\operatorname{sign}(A)$ is the exact value of $\operatorname{sign}(A+E)$ for some perturbation matrix E proportional to the precision of the arithmetic. The exact stable invariant subspace of $\operatorname{sign}(A+E)$ is also an invariant subspace of $A+E$.

However, in general, we can not guarantee that the computed value of $\operatorname{sign}(A)$ is exactly the value of $\operatorname{sign}(A+E)$ for a small perturbation E. We probably can not even represent such $\operatorname{sign}(A+E)$ within the limits of finite precision arithmetic. The best that can be hoped for is to compute $\operatorname{sign}(A+E)+F$ for some small perturbation matrices E and F. Consider now the effect of the hypothetical forward error F.

Let A have Schur form (3) and let E be a perturbation matrix partitioned conformally as in (10). Let Q_1 be the first k columns of Q and Q_2 be the remaining $n-k$ columns. If

$$\frac{\|E_{21}\|\left(\|A_{12}\|+\|E_{12}\|\right)}{\operatorname{sep}(A_{11},A_{22})-\|E_{11}\|-\|E_{22}\|}<\frac{1}{4},$$

then A has stable invariant subspace $\mathcal{V}^-(A)=\operatorname{Im}(Q_1)$ and $A+E$ has an invariant subspace $\operatorname{Im}(Q_1+Q_2W)$ where W satisfies

$$\|W\|\le\frac{2\|E_{21}\|}{\operatorname{sep}(A_{11},A_{22})-\|E_{11}\|-\|E_{22}\|} \tag{18}$$

[14, 32, 34]. The singular values of W are the tangents of the canonical angles between $\mathcal{V}^-=\operatorname{Im}(Q_1)$ and $\operatorname{Im}(Q_1+Q_2W)$. In particular, the canonical angles are at most of order $1/\operatorname{sep}(A_{11},A_{22})$.

For simplicity of notation, ignore for the moment the backward error matrix E and consider only the forward error. Let $B=\operatorname{sign}(A)+F$ where F represents the forward error in evaluating the matrix sign function and A has Schur form (3). Let $\operatorname{sign}(A)$ and F have the forms

$$Q^H\operatorname{sign}(A)Q=\begin{array}{c c} & \begin{array}{cc} k & n-k\end{array}\\ \begin{array}{c} k\\ n-k\end{array} & \left[\begin{array}{cc} -I & Y\\ 0 & I\end{array}\right]\end{array}$$

and

$$Q^H F Q = \begin{array}{c c} & \begin{array}{cc} k & n-k \end{array} \\ \begin{array}{c} k \\ n-k \end{array} & \left[\begin{array}{cc} F_{11} & F_{12} \\ F_{21} & F_{22} \end{array}\right] \end{array},$$

where Q is the unitary factor from the Schur decomposition of A (4). Now consider the stable invariant subspace $\mathcal{V}^-(A) = \mathcal{V}^-(\text{sign}(A)) = \text{Im}(Q_1)$. If $\|F_{21}\|(\|Y\| + \|F_{12}\|) < (\text{sep}(I, -I) - \|F_{11}\| - \|F_{22}\|)/4$, then perturbing $\text{sign}(A)$ to $\text{sign}(A) + F$ perturbs the invariant subspace $\text{Im}(Q_1)$ to $\text{Im}(Q_1 + Q_2 W_s)$ where $\|W_s\| \leq 2\|F_{21}\|/(2 - \|F_{11}\| - \|F_{22}\|)$ [14, 32, 34]. If $\|F_{21}\| \leq \epsilon\|\text{sign}(A)\|$, then by (8) and (9)

$$\begin{aligned} \|F_{21}\| &\leq \epsilon\|\text{sign}(A)\| \\ &\leq \epsilon\left(\frac{1}{2}\|Y\| + \sqrt{1 + \frac{\|Y\|^2}{4}}\right) \\ &\leq \epsilon\left(2\frac{\|A_{12}\|}{\text{sep}(A_{11}, A_{22})} + 1\right). \end{aligned}$$

Since $\text{sep}(A_{11}, A_{22}) \leq 2\|A\|_F$, W_s obeys the bounds

$$\|W_s\| \leq 2\epsilon \frac{2\frac{\|A_{12}\|}{\text{sep}(A_{11},A_{22})} + 1}{2 - \|F_{11}\| - \|F_{22}\|} \tag{19}$$

$$\leq 4\epsilon\left(\frac{\|A\|_F}{\text{sep}(A_{11}, A_{22})}\right) + O(\epsilon^2). \tag{20}$$

Comparing (18) with (20) we see that perturbing the computed value of $\text{sign}(A)$ by a relative error ϵ to a nearby sign matrix, disturbs the stable invariant subspace no more than twice as much as perturbing the original data A by a relative error of size ϵ might.

In order to illustrate the results, we give a comparison of our perturbation bounds and the bounds given by Bai and Demmel [4] for both the matrix sign function and the invariant subspaces in the case of Example 1. The distance to the ill-posed problem

$$d_A = \min_{\mu} \sigma_{min}(A - \mu i I),$$

where $\sigma_{min}(A - \mu i I)$ is the smallest singular value of $(A - \mu i I)$, in which μ is real and $i = \sqrt{-1}$, leads to overestimating bounds in [4]. Since $d_A \approx \eta^{-2}$, the bounds given in [4] are, respectively, $O(\eta^{-4})$ for the matrix sign function and $O(\eta^{-2})$ for the invariant subspaces.

5 A Posteriori Backward and Forward Error Bounds

A priori backward and forward error bounds for evaluation of the matrix sign function remain elusive even for the simplest algorithms. However, it isn't difficult to derive a posteriori error bounds for both backward and forward error.

We will need the following lemma to estimate the distance between a matrix S and $\operatorname{sign}(S)$.

Lemma 2 *If $S \in \mathbf{R}^{n\times n}$ has no eigenvalue with zero real part and $\|\operatorname{sign}(S)S^{-1} - I\| < 1$, then $\|\operatorname{sign}(S) - S\| \leq \|S^{-1} - S\|$.*

Proof. Let $F = \operatorname{sign}(S) - S$. The matrices F, S, and $\operatorname{sign}(S)$ commute, so

$$I = \operatorname{sign}(S)^2 = (S+F)^2 = S^2 + 2SF + F^2.$$

This implies that

$$\frac{S^{-1} - S}{2} - \frac{S^{-1}F^2}{2} = F.$$

Taking norms and using $\|FS^{-1}\| = \|\operatorname{sign}(S)S^{-1} - I\| < 1$ we get

$$\frac{1}{2}\|S^{-1} - S\| + \frac{1}{2}\|F\| \geq \|F\|$$

and the lemma follows. $\square$

It is clear from the proof of the Lemma 2 that $(\operatorname{sign}(S) - S) \approx (S^{-1} - S)/2$ is asymptotically correct as $\|\operatorname{sign}(S) - S\|$ tends to zero. The bound in the lemma tends to over estimate smaller values of $\|\operatorname{sign}(S) - S\|$ by a factor of two.

Suppose that a numerical procedure for evaluating $\operatorname{sign}(A)$ applied to a matrix $A \in \mathbf{R}^{n\times n}$ produces an approximation $S \in \mathbf{R}^{n\times n}$. Consider finding small norm solutions $E \in \mathbf{R}^{n\times n}$ and $F \in \mathbf{R}^{n\times n}$ to

$$\operatorname{sign}(A + E) = S + F. \tag{21}$$

Of course, E and F are not uniquely determined by (21). Common algorithms for evaluating $\operatorname{sign}(A)$ like Newton's method for the square root of I guarantee that S is very nearly a square root of I [16], i.e., S is a close approximation of $\operatorname{sign}(S)$. In the following theorem, we have arbitrarily taken $F = \operatorname{sign}(S) - S$.

Theorem 2 *If* $\|\operatorname{sign}(S)S^{-1} - I\| < 1$*, then (21) admits a solution with* $\|F\| \le \|S^{-1} - S\|$ *and*

$$\frac{\|E\|}{\|A\|} \le \frac{\|SA - AS\|}{\|A\|} + 2\|S^{-1} - S\|. \tag{22}$$

(The right-hand-side of (22) is easily computed or estimated from the known values of A and S, but it is subject to subtractive cancellation of significant digits.)

Proof. The matrices $S+F$ and $A+E$ commute, so an underdetermined, consistent system of equations for E in terms of S, A, and $F = \operatorname{sign}(S) - S$ is

$$E(S+F)-(S+F)E = \operatorname{sign}(S)A - A\operatorname{sign}(S) = (SA-AS)+(FA-AF). \tag{23}$$

Let

$$U^H \operatorname{sign}(S)U = \begin{bmatrix} -I & Y \\ 0 & I \end{bmatrix} \tag{24}$$

be a Schur decomposition of $\operatorname{sign}(S)$ whose unitary factor is U and whose triangular factor is on the left-hand-side of (24). Partition $U^H EU$ and $U^H AU$ conformally with the right-hand-side of (24) as

$$U^H EU = \begin{bmatrix} E_{11} & E_{12} \\ E_{21} & E_{22} \end{bmatrix}$$

and

$$U^H AU = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}.$$

Multiplying (23) on the left by U^H and on the right by U and partitioning gives

$$\begin{bmatrix} YE_{21} & E_{11}Y - YE_{22} + 2E_{12} \\ -2E_{21} & E_{21}Y \end{bmatrix} = \begin{bmatrix} YA_{21} & A_{11}Y - YA_{22} + 2A_{12} \\ -2A_{21} & A_{21}Y \end{bmatrix}.$$

A (hopefully) small norm solution for E is

$$U^H EU = \begin{bmatrix} E_{11} & E_{12} \\ E_{21} & E_{22} \end{bmatrix} = \begin{bmatrix} 0 & \frac{1}{2}(A_{11}Y - YA_{22} + 2A_{12}) \\ A_{21} & 0 \end{bmatrix}.$$

For this choice of E, we have

$$
\begin{aligned}
\|E\| &\le \|\operatorname{sign}(S)A - A\operatorname{sign}(S)\| \\
&\le \|SA - AS\| + \|FA - AF\| \\
&\le \|SA - AS\| + 2\|S^{-1} - S\|\,\|A\|
\end{aligned}
$$

from which the lemma follows. □

Lemma 2 and Theorem 2 agree well with intuition. In order to assure small forward error, S must be a good approximate square root of I and, in addition, to assure small backward error, S must nearly commute with the original data matrix A. Newton's method for a root of $X^2 - I$ tends to do a good job of both [16]. (Note that in general, Newton's method makes a poor algorithm to find a square root of a matrix. The square root of I is a special case. See [16] for details.)

6 The Newton Iteration for the Computation of the Matrix Sign Function

There are several numerical methods for computing the matrix sign function [21, 5]. Among the simplest and most commonly used is the Newton-Raphson method for a root of $X^2 - I$ starting with initial guess $X_0 = A$ [29, 30]. It is easily implemented using matrix inversion subroutines from widely available, high quality linear algebra packages like LAPACK [1, 3, 5]. It has been extensively studied and many variations have been suggested [3, 6, 9, 15, 21, 23, 22, 24].

Algorithm 1 Newton Iteration (without scaling)

$X_0 = A$

FOR $k = 0, 1, 2, \ldots$

$$X_{k+1} = (X_k + X_k^{-1})/2$$

If A has no eigenvalues on the imaginary axis, then Algorithm 1 converges globally and locally quadratically in a neighborhood of $\operatorname{sign}(A)$ [24]. Although the iteration ultimately converges rapidly, initially convergence may be slow. However, the initial convergence rate may be improved by scaling [3, 6, 9, 15, 21, 23, 22, 24].

Theorem 1 shows that the first order perturbation of $\operatorname{sign}(A)$ may be as large as $\|\operatorname{sign}(A)\|^2\epsilon$ where ϵ is the relative uncertainty in A. (If there is no other uncertainty, then ϵ is at least as large as the unit round of the finite precision arithmetic.) Thus, it is reasonable to stop the Newton iteration when

$$\|X_{k+1} - X_k\|_1 \leq c\epsilon \|X_{k+1}\|_1^2. \tag{25}$$

The *ad hoc* constant c is chosen in order to avoid extreme situations, e.g., $c = 1000n$. Experience shows furthermore that it is often advantageous to take an extra step of the iteration after the stopping criterion is satisfied.

In exact arithmetic, the stable and unstable invariant subspaces of the iterates X_k are the same as those of A. However, in finite precision arithmetic, rounding errors perturb these subspaces. The numerical stability of the Newton iteration for computing the stable invariant subspace has been analyzed in [8], we give an improved error bound here.

Let X and X^+ be, respectively, the computed k-th and $(k+1)-st$ iterate of the Newton iteration starting from

$$X_0 = A = Q \begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix} Q^H.$$

Suppose that X and X^+ have the form

$$X = Q \begin{bmatrix} X_{11} & X_{12} \\ E_{21} & X_{22} \end{bmatrix} Q^H, \qquad X^+ = Q \begin{bmatrix} X_{11}^+ & X_{12}^+ \\ E_{21}^+ & X_{22}^+ \end{bmatrix} Q^H. \tag{26}$$

A successful rounding error analysis must establish the relationship between E_{21}^+ and E_{21}. In order to do so we assume that some stable algorithm is applied to compute the inverse X^{-1} in the Newton iteration. More precisely we assume that X^+ satisfies

$$X^+ = \frac{(X + E_X) + (X + E_X)^{-1}}{2} + E_Z \tag{27}$$

where

$$\begin{aligned} \|E_X\| &\leq c\epsilon\|X\| && (28) \\ \|E_Z\| &\leq c\epsilon(\|X\| + \|X^{-1}\|), && (29) \end{aligned}$$

for some constant c. (Note that this is a nontrivial assumption. Ordinarily, if Gaussian elimination with partial pivoting is used to compute the inverse, the above error bound can be shown to hold only for each column separately [8, 33].) Write E_X and E_Z as

$$E_X = Q \begin{bmatrix} E'_{11} & E'_{12} \\ E'_{21} & E'_{22} \end{bmatrix} Q^H \tag{30}$$

$$E_Z = Q \begin{bmatrix} E''_{11} & E''_{12} \\ E''_{21} & E''_{22} \end{bmatrix} Q^H. \tag{31}$$

The following theorem bounds $\|E_{21}\|$ and indirectly the perturbation in the stable invariant subspace.

Theorem 3 *Let X, X^+, E_X, and E_Z be as in (26), (27), and (30). If $\frac{1}{2} < 1 - c\epsilon\|X\|\|X_{11}^{-1}\|$, $\frac{1}{2} < 1 - c\epsilon\|X\|\|X_{22}^{-1}\|$, and*

$$0 < \eta = 1 - 4(\|E_{21}\| + c\epsilon\|X\|)\|X_{22}^{-1}\|\|X_{11}^{-1}\|(\|X_{12}\| + c\epsilon\|X\|),$$

where c is as in (28), then

$$\|E_{21}^+\| \leq \frac{1}{2}(\|E_{21}\| + c\epsilon\|X\|)(1 + \frac{4\|X_{22}^{-1}\|\|X_{11}^{-1}\|}{\eta}) + c\epsilon(\|X\| + \|X^{-1}\|).$$

Proof. We start with (27). In fact the relationship between E_{21} and E_{21}^+ follows from applying the explicit formula for the inverse of $(X + E_X)$ in [27].

$$Q^H(X + E_X)^{-1}Q = \begin{bmatrix} \tilde{X}_{11}^{-1} + \tilde{X}_{11}^{-1}\tilde{X}_{12}\tilde{X}_c^{-1}(E_{21} + E'_{21})\tilde{X}_{11}^{-1} & -\tilde{X}_{11}^{-1}\tilde{X}_{12}\tilde{X}_c^{-1} \\ -\tilde{X}_c^{-1}(E_{21} + E'_{21})\tilde{X}_{11}^{-1} & \tilde{X}_c^{-1} \end{bmatrix}.$$

Here,

$$\begin{aligned} \tilde{X}_{11} &= X_{11} + E'_{11} \\ \tilde{X}_{12} &= X_{12} + E'_{12} \\ \tilde{X}_{22} &= X_{22} + E'_{22} \\ \tilde{X}_c &= \tilde{X}_{22} - (E_{21} + E'_{21})\tilde{X}_{11}^{-1}\tilde{X}_{12}. \end{aligned}$$

Then

$$
\begin{aligned}
X_{11}^{+} &= \frac{1}{2}(\tilde{X}_{11} + \tilde{X}_{11}^{-1} + \tilde{X}_{11}^{-1}\tilde{X}_{12}\tilde{X}_c^{-1}(E_{21} + E'_{21})\tilde{X}_{11}^{-1}) + E''_{11} \\
X_{12}^{+} &= \frac{1}{2}(\tilde{X}_{12} - \tilde{X}_{11}^{-1}\tilde{X}_{12}\tilde{X}_c^{-1}) + E''_{12} \\
E_{21}^{+} &= \frac{1}{2}((E_{21} + E'_{21}) - \tilde{X}_c^{-1}(E_{21} + E'_{21})\tilde{X}_{11}^{-1}) + E''_{21} \qquad (32) \\
X_{22}^{+} &= \frac{1}{2}(\tilde{X}_{22} + \tilde{X}_c^{-1}) + E''_{22}.
\end{aligned}
$$

Using the Neumann lemma that if $\|B\| < 1$, then $\|(I-B)^{-1}\| < (1-\|B\|)^{-1}$, [14], we have

$$
\|\tilde{X}_{11}^{-1}\| \le \frac{\|X_{11}^{-1}\|}{1 - \|X_{11}^{-1}\|\|E'_{11}\|} \le \frac{\|X_{11}^{-1}\|}{1 - c\epsilon\|X_{11}^{-1}\|\|X\|} \le 2\|X_{11}^{-1}\|.
$$

The following inequalities are established similarly.

$$
\begin{aligned}
\|\tilde{X}_{22}^{-1}\| &\le 2\|X_{22}^{-1}\| \\
\|\tilde{X}_{12}\| &\le \|X_{12}\| + c\epsilon\|X\| \\
\|\tilde{X}_c^{-1}\| &\le \frac{\|\tilde{X}_{22}^{-1}\|}{1 - \|\tilde{X}_{22}^{-1}\|(\|E_{21}\| + \|E'_{21}\|)\|\tilde{X}_{11}^{-1}\|\|\tilde{X}_{12}\|} \le \frac{2\|X_{22}^{-1}\|}{\eta}.
\end{aligned}
$$

Inserting these inequalities in (32) we obtain

$$
\|E_{21}^{+}\| \le \frac{1}{2}(\|E_{21}\| + c\epsilon\|X\|)(1 + \frac{4\|X_{22}^{-1}\|\|X_{11}^{-1}\|}{\eta}) + \|E''_{21}\|.
$$

□

The bound in Theorem 3 is stronger than the bound of Byers in [8]. A step of Newton iteration is backward stable if and only if

$$
\frac{\|E_{21}^{+}\|}{\mathrm{sep}(X_{11}^{+}, X_{22}^{+})} \le \frac{\|E_{21}\|}{\mathrm{sep}(X_{11}, X_{22})}.
$$

The term $\mathrm{sep}(X_{11}^{+}, X_{22}^{+})$ is dominated by

$$
\mathrm{sep}(\frac{X_{11} + X_{11}^{-1}}{2}, \frac{X_{22} + X_{22}^{-1}}{2}).
$$

In order to guarantee numerical stability, the factors in the bound of Theorem 3, $\|X_{11}^{-1}\|\|X_{22}^{-1}\|$ and $(\|X\|+\|X^{-1}\|)$, should be not so large as to violate the inequality

$$\|E_{12}^{+}\| \leq \frac{\operatorname{sep}(\frac{X_{11}+X_{11}^{-1}}{2}, \frac{X_{22}+X_{22}^{-1}}{2})}{\operatorname{sep}(X_{11}, X_{22})}\|E_{21}\|. \tag{33}$$

Roughly speaking, to have numerical stability throughout the algorithm, neither $\|X_{11}^{-1}\|\|X_{22}^{-1}\|$ nor $(\|X\| + \|X^{-1}\|)$ should be much larger than $1/\operatorname{sep}(A_{11}, A_{22})$.

The following example from [5] shows violation of inequality (33), which explains the numerical instability.

Example 2 Let

$$A_{11} = \begin{bmatrix} 1-\alpha & & & \alpha \\ \alpha & 1-\alpha & & \\ & \ddots & \ddots & \\ & & \alpha & 1-\alpha \end{bmatrix},$$

be a 10×10 real matrix, and let $A_{22} = -A_{11}^T$. Form $R = \begin{bmatrix} A_{11} & A_{12} \\ E_{21} & A_{22} \end{bmatrix}$ and $A = QRQ^T$, where the orthogonal matrix Q is chosen to be the unitary factor of the QR factorization of a matrix with entries chosen randomly uniformly distributed in the interval $[0, 1]$. The parameter α is taken as $\alpha = (1-10^{-5})/2$ so that there are two eigenvalues of A close to the imaginary axis from the left and right side. The entries of A_{12} are chosen randomly uniformly distributed in the interval $[0, 1]$, too. The entries of E_{21} are chosen randomly uniformly distributed in the interval $[0, \text{eps}]$, where $\text{eps} = 2.22 \times 10^{-16}$ is the machine precision.

In this example, $\operatorname{sep}(A_{11}, A_{22}) = 2.0000 \times 10^{-5}$ and $\sigma_{min}(A) = 3.3796 \times 10^{-10}$. The following table shows the evolution of $\|E_{21}\|_1/\operatorname{sep}(X_{11}, X_{22})$ during the Newton iteration starting with $X_0 = A$ and $X_0 = R$, respectively, where E_{21} is as in (26). The norm is taken to be the 1-norm.

k	$\|E_{21}\|_1/\operatorname{sep}(X_{11},X_{22})$		$\operatorname{sep}(X_{11},X_{22})$
	A	R	
0	8.7451e-11	7.0512e-11	2.0000e-05
1	7.7083e-07	1.5779e-07	1.0955e+00
2	5.0378e-07	1.0905e-07	7.9263e-01
3	1.2093e-07	2.5501e-08	1.6948e+00
4	8.3733e-08	1.2150e-08	1.7786e+00
5	7.3034e-08	5.4025e-09	2.0000
6	7.3164e-08	2.7012e-09	2.0000
7	7.2020e-08	1.3506e-09	2.0000
8	7.1731e-08	6.7532e-10	2.0000
9	7.1866e-08	3.3766e-10	2.0000
10	7.1888e-08	1.6883e-10	2.0000
11	7.1909e-08	8.4426e-11	2.0000
12	7.1926e-08	4.2231e-11	2.0000
13	7.1934e-08	2.1151e-11	2.0000
14	7.1938e-08	1.0646e-11	2.0000
15	7.1938e-08	5.4637e-12	2.0000
16	7.1937e-08	3.0055e-12	2.0000
17	7.1938e-08	2.0001e-12	2.0000
18	7.1937e-08	1.7474e-12	2.0000
19	7.1937e-08	1.7291e-12	2.0000
20	7.1937e-08	1.7290e-12	2.0000
21	7.1937e-08	1.7290e-12	2.0000

Because $\|A_{11}^{-1}\|_1\|A_{22}^{-1}\|_1 = 1.0000\times 10^{10}$, $\|A^{-1}\|_1 = 2.2516\times 10^9$, inequality (33) is violated in the first step of the Newton iteration for starting matrix A, which is shown in the first column of the table. Newton's method never recovers from this.

It is remarkable, however, that Newton's method applied to R directly seems to recover from the loss in accuracy in the first step. The second column shows that although $\|E_{21}\|_1/\operatorname{sep}(X_{11},X_{22}) = 1.5779\times 10^{-7}$ at the first step, it is reduced by the factor 1/2 every step until it reaches 1.7290×10^{-12} which is approximately $\|E_{21}\|_1/\operatorname{sep}(A_{11},A_{22})$. Observing that in this case the perturbation E''_{21} in E_Z as in (27) is zero and $\|E^{+}_{21}\|_1$ is dominated by $\frac{1}{2}(\|E_{21}\|_1+\|X_{22}^{-1}E_{21}X_{11}^{-1}\|_1)$. It is surprising to see that from the second step on $\|X_{11}^{-1}E_{21}X_{22}^{-1}\|_1$ is as small as eps, since A_{11}^{-1} and A_{22}^{-1} do not explicitly appear in the term $X_{11}^{-1}E_{21}X_{22}^{-1}$ after the first step.

By our analysis, the Newton iteration may be unstable when X_k is ill-conditioned. To overcome this difficulty the Newton iteration may be carried out with a shift along the imaginary line. In this case we have to use complex arithmetic.

Algorithm 2 Newton Iteration With Shift

$X_0 = A - \beta i I$
FOR $k = 0, 1, 2, \ldots$
$\quad X_{k+1} = (X_k + X_k^{-1})/2$
END

The real parameter β is chosen such that $\sigma_{min}(A - \beta i I)$ is not small. For Example 2, when β is taken to be 0.8, we have $\|E_{21}\|_1/\operatorname{sep}(X_{11}, X_{22}) = 7.3134 \times 10^{-12}$ for $k = 21$. Then by our analysis the computed invariant subspace is guaranteed to be accurate.

We can combine Algorithm 1 and Algorithm 2 in the following way.

Algorithm 3 Computing the Stable Invariant Subspace

1. Call Algorithm 1 with the stopping criterion (25) and get X_{k+1}.

2. Perform a QR factorization of $X_{k+1} - I$ *and partition* $Q = (Q_1, Q_2)$.

3. (Stability test) If $\|(Q_2^H A Q_1)\|_1/\|A\|_1 \leq n * \epsilon * \|X_{k+1}\|_1$, *where* ϵ *is the machine precision, then use* Q_1 *as the orthonormal basis of the computed stable invariant subspace. Otherwise call Algorithm 2 and start the step 1 again.*

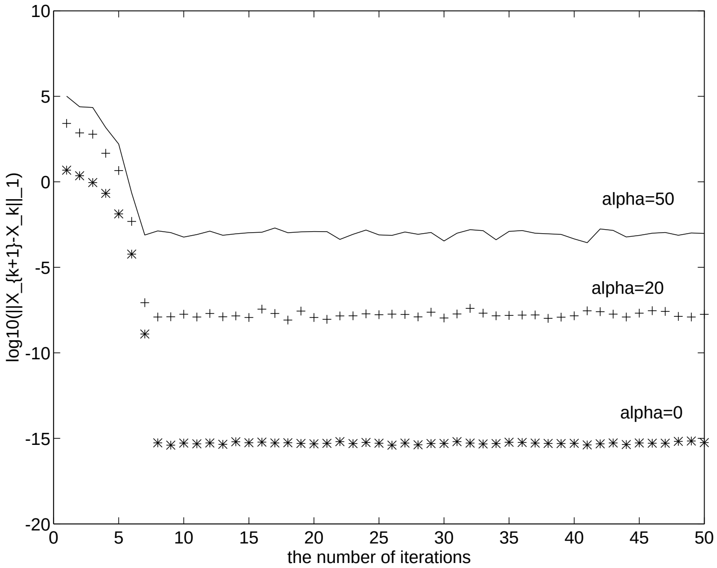
7 Numerical Experiments

In this section we demonstrate the theoretical results of this paper with some numerical experiments: The numerical algorithms were implemented in MATLAB 4.1 on a HP 715/33 workstation with $eps = 2.2204 * 10^{-16}$. The stopping criterion for the Newton iteration is as in (25) with $c = 1000$ and an extra step of the Newton iteration is performed after the stopping criterion is satisfied.

Example 3 This example is devoted to demonstrate the validity of our stopping criterion. We constructed a 10×10 matrix

$$A = QRQ^H,$$

where Q is a random unitary matrix and R an upper triangular matrix with diagonal elements $-1 \pm 0.2i, -2.0, -2.5, -3.0, -4.0, -4.5, 2 \pm 0.2i, 6.0$, a parameter α in the $(k, k+2)$ position and zero everywhere else. We chose α such that the norm $\|\operatorname{sign}(A)\|_1$ varies from small to large. The typical numerical behavior of $\log_{10}(\|X_{k+1} - X_k\|_1)$ is that it goes down and then becomes staionary. This behavior is shown in the following graph for the cases $\alpha = 0, \alpha = 20$ and $\alpha = 50$ in which $\|\operatorname{sign}(A)\|_1$ is 2.9132, $1.7418 * 10^3$ and $6.2279 * 10^4$ respectively. The Newton iteration with our stopping criterion stops at the 9-th step for $\alpha = 0$ and at the 8-th step for $\alpha = 20$ and $\alpha = 50$.



Example 4 In this test 100 random matrices of size 100×100 were considered. In all cases, the condition $\|Q_2^H * A * Q_1\|_1 / \|A\|_1 \leq n * eps * \|X_{k+1}\|_1$ is satisfied which indicates by the perturbation analysis in Section 4 that the computed stable invariant subspace is acceptable.

In Example 2, however, we have $\|Q_2^H * A * Q_1\|_1 / \|A\|_1 = 4.1616 \times 10^{-9}$, $n * eps * \|X_{k+1}\|_1 = 6.3351 \times 10^{-11}$ and hence the computed stable invariant subspace is not acceptable. However, when the Newton iteration is started with $X_0 = A - 0.8iI$, the stability condition is satisfied.

8 Conclusions

We have given a first order perturbation theory for the matrix sign function and an error analysis for Netwon's method to compute it. This analysis suggests that computing the stable (or unstable) invariant subspace of a matrix with the Newton iteration in most circumstances yields results as good as those obtained from the Schur form.

9 Acknowledgments

The authors would like to express their thanks to N. Higham for valuable comments on an earlier draft of the paper and Z. Bai and P. Benner for helpful discussions.

References

[1] E. Anderson, Z. Bai, C. Bischof, J. Demmel, J. Dongarra, J. D. Croz, A. Greenbaum, S. Hammarling, A. McKenney, S. Ostrouchov, and D. Sorensen. *LAPACK Users' Guide*. Society for Industrial and Applied Mathematics, Philadelphia, 1992.

[2] Z. Bai and J.W. Demmel. Design of a parallel nonsymmetric eigenroutine toolbox,Part I. Technical Report CDS-92-718, Dept. of Mathematics, University of California, Berkeley, Berkeley, Ca, 1992.

[3] Z. Bai and J.W. Demmel. Design of a parallel nonsymmetric eigenroutine toolbox. In *Proc. of the 6th SIAM Conf. on Parallel Processing for Scientific Computing*, Philadelphia, 1993. SIAM. Long version available from Dept. of Mathematics, University of California, Berkeley.

[4] Z. Bai and J.W. Demmel. Design of a parallel nonsymmetric eigenroutine toolbox,Part II. Technical report, Dept. of Mathematics, University of California, Berkeley, Berkeley, Ca, 1994.

[5] Z. Bai, J.W. Demmel, and M. Gu. Inverse free parallel spectral divide and conquer algorithms for nonsymmetric eigenproblems. Technical report, Dept. of Mathematics, University of California, Berkeley, Berkeley, Ca, 1994.

[6] L.A. Balzer. Accelerated convergence of the matrix sign function method of solving Lyapunov, Riccati and other matrix equations. *Int. J. Control*, 32:1057–1078, 1980.

[7] A.N. Beavers and E.D. Denman. A computational method for eigenvalues and eigenvectors of a matrix with real eigenvalues. *Numer. Math.*, 21:389–396, 1973.

[8] R. Byers. Numerical stability and instability in matrix sign function based algorithms. In C.I. Byrnes and A. Lindquist, editors, *Computational and Combinatorial Methods in Systems Theory*, pages 185–199, Elsevier, North Holland, 1986.

[9] R. Byers. Solving the algebraic Riccati equation with the matrix sign function. *Lin. Alg. Appl.*, 85:267–279, 1987.

[10] E.D. Denman and A.N. Beavers. The matrix sign function and computations in systems. *Appl. Math. and Comput.*, 2:63–94, 1976.

[11] E.D. Denman and J. Leyva-Ramos. Spectral decomposition of a matrix using the generalized sign matrix. *Appl. Math. and Comput.*, 8:237–250, 1981.

[12] J.D. Gardiner and A.J. Laub. A generalization of matrix-sign-functions solution for algebraic Riccati equations. *Int. J. Control*, 44:823–832, 1986.

[13] J.D. Gardiner and A.J. Laub. Parallel algorithms for algebraic Riccati equations. *Int. J. Control*, 54:1317–1333, 1991.

[14] G.H. Golub and C.F. Van Loan. *Matrix Computations*. The Johns Hopkins University Press, Baltimore, Maryland, 2nd edition, 1989.

[15] N.J. Higham. Computing the polar decomposition - with application. *SIAM J. Sci. Stat. Comput.*, 7:1160–1174, 1986.

[16] N.J. Higham. Newton's method for the matrix square root. *Math. Comp.*, 46(174):537–549, April 1986.

[17] T. Kato. On the convergence of the perturbation method, i. *Progr. Theor. Phys.*, 4:514–523, 1949.

[18] T. Kato. On the convergence of the perturbation method, ii. *Progr. Theor. Phys.*, 5:207–212, 1950.

[19] T. Kato. *Perturbation Theory for Linear Operators*. Springer-Verlag, Berlin, 1980.

[20] C. Kenney and A.J. Laub. Polar decompositions and matrix sign function condition estimates. *SIAM J. Sci. Stat. Comp.*, 12:488–504, 1991.

[21] C. Kenney and A.J. Laub. Rational iteration methods for the matrix sign function. *SIAM J. Math. Anal. Appl.*, 21:487–494, 1991.

[22] C. Kenney and A.J. Laub. Matrix-sign algorithms for Ricatti equations. *IMA J. Math. Contr. Info.*, 9:331–344, 1992.

[23] C. Kenney and A.J. Laub. On scaling Newton's method for polar decompositions and the matrix sign function. *SIAM Matrix Anal. Appl.*, 13:688–706, 1992.

[24] C. Kenney and A.J. Laub. The matrix sign runction. Technical report, Department of Electrical and Computing Engineering, University of California, Santa Barbara, 1994.

[25] P. Lancaster and M. Tismenetsky. *The Theory of Matrices*. Academic Press, Orlando, 1985.

[26] R. Mathias. Condition estimation for the matrix sign function via the schur decomposition. In John G. Lewis, editor, *Proceedings of the Fifth SIAM Conference on Applied Linear Algebra*, pages 85–89, 1994.

[27] D. V. Ouellette. Schur complements and statistics. *Lin. Alg. Appl.*, 36:187–295, 1981.

[28] C. Kenney P. Pandey and A.J. Laub. A parallel algorithm for the matrix sign function. *Int. J. High Speed Computing*, 2:181–191, 1990.

[29] J. Roberts. Linear model reduction and solution of algebraic Riccati equations by use of the sign function. Technical Report Engineering Report, CUED/B-Control, Tr-13, Cambridge University, Cambridge, England, 1971.

[30] J.D. Roberts. Linear model reduction and solution of the algebraic riccati equation. *Int. J. Control*, 32:677–687, 1980. Reprint of technical report, Cambridge Univ. 1971.

[31] M. Rosenblum. On the operator equation "$BX - XA = Q$". *Duke Math. J.*, 23:263–269, 1956.

[32] G. W. Stewart. Error and perturbation bounds for subspaces associated with certain eigenvalue problems. *SIAM Review*, pages 727–764, 1973.

[33] G. W. Stewart. *Introduction to Matrix Computations.* Academic Press, New York, 1973.

[34] G. W. Stewart and Ji-Guang Sun. *Matrix Perturbation Theory.* Academic Press, London, 1990.

[35] B. Sz.-Nagy. Perturbations des transformations autoadjointes dans l'espace de Hilbert. *Comment. Math. Helv.*, 19:347–366, 1946/47.

[36] J. H. Wilkinson. *The Algebraic Eigenvalue Problem.* Clarendon Press, Oxford University Press, Walton Street, Oxford OX2 6DP, 1965.