SIAM J. Matrix Anal. Appl.
Vol. 19, No. 1, pp. 205–225, January 1998

USING THE MATRIX SIGN FUNCTION TO COMPUTE INVARIANT SUBSPACES[*]

ZHAOJUN BAI[†] AND JAMES DEMMEL[‡]

Abstract. The matrix sign function has several applications in system theory and matrix computations. However, the numerical behavior of the matrix sign function, and its associated divide-and-conquer algorithm for computing invariant subspaces, are still not completely understood. In this paper, we present a new perturbation theory for the matrix sign function, the conditioning of its computation, the numerical stability of the divide-and-conquer algorithm, and iterative refinement schemes. Numerical examples are also presented. An extension of the matrix-sign-function-based algorithm to compute left and right deflating subspaces for a regular pair of matrices is also described.

Key words. matrix sign function, Newton's method, eigenvalue problem, invariant subspace, deflating subspaces

AMS subject classifications. 65F15, 65F35, 65F30, 15A18

PII. S0895479896297719

1. Introduction. Since the matrix sign function was introduced in the early 1970s, it has been the subject of numerous studies and used in many applications. For example, see [30, 31, 11, 26, 23] and references therein. Our main interest here is to use the matrix sign function to build parallel algorithms for computing invariant subspaces of nonsymmetric matrices, as well as their associated eigenvalues. It is a challenge to design a parallel algorithm for the nonsymmetric eigenproblem that uses coarse grain parallelism effectively, scales for larger problems on larger machines, does not waste time dealing with the parts of the spectrum in which the user is not interested, and deals with highly nonnormal matrices and strongly clustered spectra. In the work of [2], after reviewing the existing approaches, we proposed a design of a parallel nonsymmetric eigenroutine toolbox, which includes the basic building blocks (such as LU factorization, matrix inversion, and the matrix sign function), standard eigensolver routines (such as the QR algorithm), and new algorithms (such as spectral divide-and-conquer using the sign function). We discussed how these tools could be used in different combinations on different problems and architectures, for extracting all or some of the eigenvalues of a nonsymmetric matrix, and/or their corresponding invariant subspaces. Rather than using "black box" eigenroutines such as provided by EISPACK [32, 21] and LAPACK [1], we expect the toolbox approach to allow us more flexibility in developing efficient problem-oriented eigenproblem solvers on high-performance machines, especially on parallel distributed memory machines.

[*] Received by the editors January 25, 1996; accepted for publication (in revised form) by V. Mehrmann February 4, 1997.
http://www.siam.org/journals/simax/19-1/29771.html

[†] Department of Mathematics, University of Kentucky, Lexington, KY 40506 (bai@ms.uky.edu). The research of this author was supported in part by ARPA grant DM28E04120 and P-95006 via a subcontract from Argonne National Laboratory, by NSF grant ASC-9313958, and in part by DOE grant DE-FG03-94ER25219 via subcontracts from the University of California at Berkeley.

[‡] Computer Science Division and Mathematics Department, University of California, Berkeley, CA 94720 (demmel@cs.berkeley.edu). The research of this author was funded in part by ARPA contract DAAH04-95-1-0077 through University of Tennessee subcontract ORA7453.02, ARPA contract DAAL03-91-C-0047 through University of Tennessee subcontract ORA4466.02, NSF contracts ASC-9313958 and ASC-9404748, DOE contracts DE-FG03-94ER25219, DE-FG03-94ER25206, DOE contract W-31-109-Eng-38 through subcontract 951322401 with Argonne National Laboratory, and NSF Infrastructure Grants CDA-8722788 and CDA-9401156.

However, the numerical accuracy and stability of the matrix sign function and divide-and-conquer algorithms based on it are poorly understood. In this paper, we will address these issues. Much of this work also appears in [3].

Let us first restate some of basic definitions and ideas to establish notation. The matrix sign function of a matrix A is defined as follows [30]: let

$$A = X \operatorname{diag}(J_+, J_-) X^{-1}$$

be the Jordan canonical form of a matrix $A \in \mathbf{C}^{n\times n}$, where the eigenvalues of J_+ lie in the open right half-plane ($\mathbf{C}_+$) and those of J_- lie in the open left half-plane ($\mathbf{C}_-$). Then the matrix sign function of A is

$$\operatorname{sign}(A) = X \operatorname{diag}(I, -I) X^{-1}.$$

We assume that no eigenvalue of A lies on the imaginary axis; otherwise, $\operatorname{sign}(A)$ is not defined. It is easy to show that the spectral projection corresponding to the eigenvalues of A in the open right and left half-planes are $P_\pm = \frac{1}{2}(I \pm \operatorname{sign}(A))$, respectively. Let the leading columns of an orthogonal matrix Q span the range space of P_+ (for example, Q may be computed by the rank-revealing QR decomposition of P_+). Then Q yields the spectral decomposition

$$Q^T A Q = \begin{pmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{pmatrix}, \tag{1}$$

where $\lambda(A_{11})$ are the eigenvalues of A in $\mathbf{C}_+$, and $\lambda(A_{22})$ are the eigenvalues of A in $\mathbf{C}_-$. The algorithm proceeds in a divide-and-conquer fashion by computing the eigenvalues of A_{11} and A_{22}.

Rather than using the Jordan canonical form to compute $\operatorname{sign}(A)$, it can be shown that $\operatorname{sign}(A)$ is the limit of the following Newton iteration:

$$A_{k+1} = \frac{1}{2}(A_k + A_k^{-1}) \quad \text{for} \quad k = 0, 1, 2, \ldots, \quad \text{with} \quad A_0 = A. \tag{2}$$

The iteration is globally and ultimately quadratic convergent. There exist different scaling schemes to speed up the convergence of the iteration, and make it more suitable for parallel computation. By computing the matrix sign function of a Möbius transformation of A, the spectrum can be divided along arbitrary lines and circles, rather than just along the imaginary axis. See report [2] and references therein for more details.

Unfortunately, in finite precision arithmetic, the ill conditioning of a matrix A_k with respect to inversion and rounding errors may destroy the convergence of the Newton iteration (2) or cause convergence to the wrong answer. Consequently, the left bottom corner block of the matrix $Q^T AQ$ in (1) may be much larger than $\mathbf{u}\|A\|$, where $\mathbf{u}$ denotes machine precision. This means that it is not numerically stable to approximate the eigenvalues of A by the eigenvalues of A_{11} and A_{22}, as we would like.

In this paper, we will first study the perturbation theory of the matrix sign function, its conditioning, and the numerical stability of the overall divide-and-conquer algorithm based on the matrix sign function. We realize that it is very difficult to give a complete and clear analysis. We only have a partial understanding of when we can expect the Newton iteration to converge and how accurate it is. In a coarse analysis, we can also bound the condition numbers of intermediate matrices in the Newton iteration. Artificial and possibly very pathological test matrices are constructed to verify our theoretical analysis. Besides these artificial tests, we also test a large number of eigenvalue problems of random matrices, and a few eigenvalue problems from applications, such as electrical power system analysis, numerical simulation of chemical reactions, and aerodynamics stability analysis. Through these examples, we conclude that the most bounds for numerical sensitivity and stability of matrix sign function computation and its based algorithms are reachable for some very pathological cases, but they are often very pessimistic. The worst cases happen rarely.

In addition, we discuss iterative refinement of an approximate invariant subspace and outline an extension of the matrix-sign-function-based algorithms to compute both left and right deflating subspaces for a regular matrix pencil $A - \lambda B$.

The rest of this paper is organized as follows. Section 2 presents a new perturbation bound for the matrix sign function. Section 3 discusses the numerical conditioning of the matrix sign function. The backward error analysis of computed invariant subspace and remarks on the matrix-sign-function-based algorithm versus the QR algorithm are presented in section 4. Section 5 presents some numerical examples for the analysis of sections 2, 3, and 4. Section 6 describes the iteration refinement scheme to improve an approximate invariant subspace. Section 7 outlines an extension of the matrix-sign-function-based algorithms for the generalized eigenvalue problem. Concluding remarks are presented in section 8.

2. A perturbation bound for the matrix sign function.

When a matrix A has eigenvalues on the pure-imaginary axis, its matrix sign function is not defined. In other words, the set of *ill-posed problems* for the matrix sign function is the set of matrices with at least one pure-imaginary eigenvalue. Computationally, we have observed that when there are the eigenvalues of A close to the pure-imaginary axis, the Newton iteration and its variations are very slowly convergent and may be misconvergent. Moreover, even when the iteration converges, the error in the computed matrix sign function could be too large to use. It is desirable to have a perturbation analysis of the matrix sign function related to the distance from A to the nearest ill-posed problem.

Perturbation theory and condition number estimation of the matrix sign function are discussed in [25, 23, 29]. However, none of the existing error bounds explicitly reveals the relationship between the sensitivity of the matrix sign function and the distance to the nearest ill-posed problem. In this section, we will derive a new perturbation bound which explicitly reveals such relationship. We will denote all the eigenvalues of A with positive real part by $\lambda_+(A)$, i.e., $\lambda_+(A) = \{\lambda | \lambda \in \lambda(A), \Re(\lambda) > 0\}$. $\sigma_{\min}(A)$ denotes the smallest singular value of A. In addition, we recall the well-known inequality

$$\|(I - X)^{-1}\| \leq \frac{1}{1 - \|X\|} \quad \text{if } \|X\| < 1, \tag{3}$$

where $\| \cdot \|$ is the matrix 2-norm.

Theorem 2.1. *Suppose A has no pure-imaginary and zero eigenvalues, $A + \delta A$ is a perturbation of A, and $\epsilon \equiv \|\delta A\|$. Let*

$$\omega = \max_{\tau \in \mathbb{R}} \|(i\tau I - A)^{-1}\| = \frac{1}{\min\limits_{\tau \in \mathbb{R}} \sigma_{\min}(i\tau I - A)} \equiv \frac{1}{d_A}. \tag{4}$$

Then

$$\|\text{sign}(A)\| \leq \frac{4}{\pi}\omega\|A\| + 3. \tag{5}$$

Furthermore, if

$$\omega\epsilon < 1, \tag{6}$$

then

$$\|\mathrm{sign}(A+\delta A) - \mathrm{sign}(A)\| \le \frac{4}{\pi}\frac{\omega^2\epsilon}{1-\omega\epsilon}(\|A\|+\epsilon) + 2\frac{\epsilon}{\|A\|}. \tag{7}$$

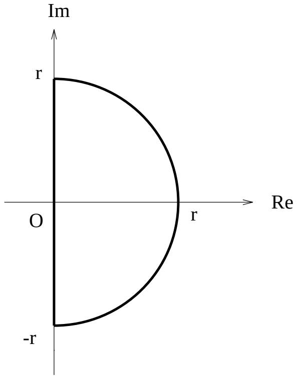

FIG. 1. *The semicircle* Γ.

Proof. We only prove the bound (7). The bound (5) can be proved by using a similar technique. Following Roberts [30] (or Kato [24]), the matrix sign function can also be defined using Cauchy integral representation:

$$\mathrm{sign}(A) = 2\,\mathrm{sign}^+(A) - I, \tag{8}$$

where

$$\mathrm{sign}^+(A) = \frac{1}{2\pi i}\int_\Gamma (\zeta I - A)^{-1} d\zeta,$$

Γ is any simple closed curve with positive direction enclosing $\lambda_+(A)$. $\mathrm{sign}^+(A)$ is the spectral projector for $\lambda_+(A)$. Here, without loss of generality, we take Γ to be a semicircle with radius $r = 2\max\{\|A\|, \|A+\delta A\|\}$ (see Figure 1). From the definition (8) of $\mathrm{sign}(A)$, it is seen that to study the stability of the matrix sign function of A to the perturbation δA, it is sufficient to just study the sensitivity of the projection $\mathrm{sign}^+(A)$.

Let $\mathrm{sign}^+(A+\delta A)$ be the projection corresponding to $\lambda_+(A+\delta A)$, from the condition (6), no eigenvalues of A are perturbed across or on the pure imaginary axis, and the semicircle Γ also encloses $\lambda_+(A+\delta A)$. Therefore, we have

$$\begin{aligned}\mathrm{sign}^+(A+\delta A) - \mathrm{sign}^+(A) &= \frac{1}{2\pi i}\int_\Gamma [(\zeta I - A - \delta A)^{-1} - (\zeta I - A)^{-1}]d\zeta \\ &= \frac{1}{2\pi i}\int_{-r}^{r} [(i\tau I - A - \delta A)^{-1} - (i\tau I - A)^{-1}]id\tau \\ &\quad + \frac{1}{2\pi i}\int_{-\pi/2}^{\pi/2} [(re^{i\theta} I - A - \delta A)^{-1} - (re^{i\theta} I - A)^{-1}]ire^{i\theta}d\theta \\ &\equiv \mathcal{I}_1 + \mathcal{I}_2,\end{aligned}$$

where the first integral, denoted $\mathcal{I}_1$, is the integral over the straight line of the semicircle Γ, the second integral, denoted $\mathcal{I}_2$, is the integral over the curved part of the semicircle Γ. Now, by taking the spectral norm of the first integral term, and noting the definition of ω, the condition (6), and the inequality (3), we have

$$\begin{aligned}
\|\mathcal{I}_1\| &\le \frac{1}{2\pi}\int_{-r}^{r} \|[(i\tau I - A - \delta A)^{-1} - (i\tau I - A)^{-1}]\|\ |d\tau| \\
&= \frac{1}{2\pi}\int_{-r}^{r} \|[(i\tau I - A - \delta A)^{-1}\delta A(i\tau I - A)^{-1}]\|\ |d\tau| \\
&= \frac{1}{2\pi}\int_{-r}^{r} \|(I - (i\tau I - A)^{-1}\delta A)^{-1}(i\tau I - A)^{-1}\delta A(i\tau I - A)^{-1}\|\ |d\tau| \\
&\le \frac{1}{2\pi}\int_{-r}^{r} \frac{\|(i\tau I - A)^{-1}\|^2\|\delta A\|}{1 - \|(i\tau I - A)^{-1}\delta A\|}|d\tau| \\
&\le \frac{1}{2\pi}\frac{\omega^2\|\delta A\|}{1-\omega\|\delta A\|}\ 2r.
\end{aligned}$$

By taking the spectral norm of the second integral term $\mathcal{I}_2$, we have

$$\begin{aligned}
\|\mathcal{I}_2\| &\le \frac{1}{2\pi}\int_{-\pi/2}^{\pi/2} \|(re^{i\theta}I - A - \delta A)^{-1}\delta A(re^{i\theta}I - A)^{-1}\|\ r\ |d\theta| \\
&\le \frac{1}{2\pi}\int_{-\pi/2}^{\pi/2} \left\|\left(I - \frac{A+\delta A}{re^{i\theta}}\right)^{-1}\right\|\ \|\delta A\|\ \left\|\left(I - \frac{A}{re^{i\theta}}\right)^{-1}\right\| \frac{1}{r}|d\theta| \\
&\le \frac{1}{2\pi}\left(\frac{1}{1-\|A+\delta A\|/r}\right)\ \|\delta A\|\ \left(\frac{1}{1-\|A\|/r}\right)\ \frac{1}{r}\ \pi \\
&\le \frac{2\|\delta A\|}{r} \le \frac{\|\delta A\|}{\|A\|},
\end{aligned}$$

where the third inequality follows from (3) and the fourth follows from the choice of the radius r of the semicircle Γ. The desired bound (7) follows from the bounds on $\|\mathcal{I}_1\|$ and $\|\mathcal{I}_2\|$ and the identity

$$\operatorname{sign}(A+\delta A) - \operatorname{sign}(A) = 2(\operatorname{sign}^+(A+\delta A) - \operatorname{sign}^+(A)). \qquad \square$$

A few remarks are in order:

1. In the language of pseudospectra [35], the condition (6) means that the $\|\delta A\|$-pseudospectra of A do not cross the pure-imaginary axis.
2. From the perturbation bound (7), we see that the stability of the matrix sign function to the perturbation requires not only the $\|\delta A\|$-pseudospectra of the A to be bounded away from the pure-imaginary axis but also $\omega^2 = 1/d_A^2$ to be small (recall that d_A is the distance from A to the nearest matrix with a pure-imaginary eigenvalue).
3. It is natural to take $\omega^2 = 1/d_A^2$ as the condition number of the matrix sign function. Algorithms for computing d_A and related problems can be found in [14, 9, 8, 12].
4. The bound (7) is similar to the bound of the norm of the Fréchet derivative of the matrix sign function of A at X given by Roberts [30]:

$$\|\mathcal{F}(\operatorname{sign}(A), X)\| \le \frac{l_\Gamma}{2\pi}\left(\max_{\zeta \text{ on } \Gamma}\|(\zeta I - A)^{-1}\|^2\right)\|X\|,$$

where l_Γ is the length of the closed contour Γ.

Recently, an asymptotic perturbation bound of sign(A) was given by Byers, He, and Mehrmann [13]. They show that to first order in δA

$$\|\mathrm{sign}(A+\delta A) - \mathrm{sign}(A)\| \le \frac{4}{\delta}\left(1 + \frac{\|A_{12}\|}{\delta}\right)^2 \|\delta A\|, \tag{9}$$

where A is assumed to have the form of (1), $\|\delta A\|$ is sufficiently small, and

$$\delta = \mathrm{sep}(A_{11}, A_{22}) = \sigma_{\min}(I \otimes A_{11} - A_{22}^T \otimes I), \tag{10}$$

the separation of the matrices A_{11} and A_{22} [33]. $\otimes$ is the Kronecker product. Comparing the bounds (7) and (9), we note that first the bound (7) is a global bound and (9) is an asymptotic bound. Second, the assumption (6) for the bound (7) has a simple geometric interpretation (see remark 2 above). It is unspecified how to interpret the assumption on sufficiently small $\|\delta A\|$ for the bound (9).

3. Conditioning of matrix sign function computation.

In [2], we point out that it may be much more efficient to compute $S = \mathrm{sign}(A)$ to half-machine precision only, i.e., to compute S with an absolute error bounded by $\mathbf{u}^{1/2}\|S\|$. To avoid ill conditioning in the Newton iteration and achieve the half-machine precision, we believe that the matrix A must have condition number less than $\mathbf{u}^{-1/2}$. If A is ill conditioned, say having singular values less than $\mathbf{u}^{1/2}\|A\|$, we need to use a preprocessing step to deflate small singular values by a unitary similarity transformation, and obtain a submatrix having condition number less than $\mathbf{u}^{-1/2}$, and then compute the matrix sign function of this submatrix. Such a deflation procedure may be also needed for the intermediate matrices in the Newton iteration in the worst case.

We now look more closely at the situation of near convergence of the Newton iteration and relate the error to the distance to the nearest ill-posed problem [18]. As before, the ill-posed problems are those matrices with pure-imaginary eigenvalues. Without loss of generality, let us assume A is of the form

$$A = \begin{pmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{pmatrix}, \tag{11}$$

where $\lambda(A_{11}) \in \mathbf{C}_+$ and $\lambda(A_{22}) \in \mathbf{C}_-$. Otherwise, for any matrix B, by the Schur decomposition, we can write $B = Q^H A Q$, where A has the above form, and then $\mathrm{sign}(B) = Q^H \mathrm{sign}(A) Q$. Let R be the solution of the Sylvester equation

$$A_{11}R - RA_{22} = -A_{12}, \tag{12}$$

which must exist and be unique since A_{11} and A_{22} have no common eigenvalues. Then it is known that the spectral projector P corresponding to the eigenvalues of A_{11} is

$$P = \begin{pmatrix} I & R \\ 0 & 0 \end{pmatrix},$$

and $\|P\| = \sqrt{1 + \|R\|^2}$. The following lemma relates R and the norm of the projection P to sign(A) and its condition number.

LEMMA 3.1. *Let A and R be as above. Let $\rho = \|R\| + \sqrt{1 + \|R\|^2}$. Then*

1. $S \equiv \mathrm{sign}(A) = \begin{pmatrix} I & -2R \\ 0 & -I \end{pmatrix}$.

2. *$\|S\| = \|S^{-1}\| = \rho$, and, therefore, $\kappa(S) = \rho^2$.*

Proof.

1. Let $X = \begin{pmatrix} I & R \\ 0 & I \end{pmatrix}$. It is easy to verify that if R satisfies (12), then $X^{-1}AX = \operatorname{diag}(A_{11}, A_{22})$. Therefore,

$$\begin{aligned}\operatorname{sign}(A) &= \operatorname{sign}(X\operatorname{diag}(A_{11}, A_{22})X^{-1}) = X\operatorname{sign}(\operatorname{diag}(A_{11}, A_{22}))X^{-1}\\ &= X\operatorname{diag}(I, -I)X^{-1} = \begin{pmatrix} I & -2R \\ 0 & -I \end{pmatrix}.\end{aligned}$$

2. Using the singular value decomposition (SVD) of R $URV^H = \Sigma = \operatorname{diag}(\sigma_i)$, one can reduce computing the SVD of S to computing the SVD of

$$\begin{pmatrix} U & 0 \\ 0 & V \end{pmatrix} S \begin{pmatrix} U^H & 0 \\ 0 & V^H \end{pmatrix} = \begin{pmatrix} I & -2\Sigma \\ 0 & -I \end{pmatrix}$$

which, by permutations, is equivalent to computing the SVDs of the 2×2 matrices $\begin{pmatrix} 1 & -2\sigma_j \\ 0 & -1 \end{pmatrix}$. This is, in turn, a simple calculation. □

We note that for the solution R of the Sylvester equation (12) we have

$$\|R\| \leq \frac{\|A_{12}\|}{\operatorname{sep}(A_{11}, A_{22})},$$

where the equality is attainable [33]. From Lemma 3.1, we see that the conditioning of the matrix sign function computation is closely related to the norm of the projection P, therefore the norm of R, which in turn is closely related to the quantity $\operatorname{sep}(A_{11}, A_{22})$. Specifically, when $\|R\|$ is large,

$$\|S\| = \|\operatorname{sign}(A)\| \leq \frac{2\|A_{12}\|}{\operatorname{sep}(A_{11}, A_{22})} \tag{13}$$

and

$$\kappa(S) \leq \frac{4\|A_{12}\|^2}{\operatorname{sep}^2(A_{11}, A_{22})}.$$

If $\|A_{12}\|$ is moderate, an ill-conditioned matrix sign function means large $\|R\|$, which in turn means small $\operatorname{sep}(A_{11}, A_{22})$. Following Stewart [33], it means that it is harder to separate the invariant subspaces corresponding to the matrices A_{11} and A_{22}.

The following theorem discusses the conditioning of the eigenvalues of $\operatorname{sign}(A)$ and the distance from $\operatorname{sign}(A)$ to the nearest ill-posed problem.

THEOREM 3.2. *Let A and R be as in Lemma 3.1. Then we have the following:*

1. *Let δS have the property that $S + \delta S$ has a pure-imaginary eigenvalue. Then δS may be chosen with $\|\delta S\| = 1/\|S\|$ but no smaller. In the language of [35], the ϵ-pseudospectrum of S excludes the imaginary axis for $\epsilon < 1/\|S\|$, and intersects it for $\epsilon \geq 1/\|S\|$.*
2. *The condition number of the eigenvalues of S is $\|P\|$. In other words, perturbing S by a small δS perturbs the eigenvalues by at most $\|P\|\,\|\delta S\| + \mathcal{O}(\|\delta S\|^2)$.*
3. *If A is close to S and $\kappa(S) < \mathbf{u}^{-1/2}$, then Newton iteration (2) in floating point arithmetic will compute S with an absolute error bounded by $\mathbf{u}^{1/2}\|S\|$.*

Proof.

1. The problem is to minimize $\sigma_{\min}(S - i\zeta I)$ over all real ζ, where $\sigma_{\min}$ is the smallest singular value of $S - i\zeta I$. Using the same unitary similarity transformation and permutation as in the part 1 of Lemma 3.1, we see that this is equivalent to minimizing
$$\sigma_{\min}\left(\begin{pmatrix} 1-i\zeta & -2\sigma_j \\ 0 & -1-i\zeta \end{pmatrix}\right)$$
over all σ_j and real ζ. This is a straightforward calculation, with the minimum being obtained for $\zeta = 0$ and $\sigma_j = \|R\|$.
2. The condition number of a semisimple eigenvalue is equal to the secant of the acute angle between its left and right eigenvectors [24, 17]. Using the above reduction to 2×2 subproblems (this unitary transformation of coordinates does not changes angles between vectors), this is again a straightforward calculation.
3. Since $\|S\| = \|S^{-1}\|$, the absolute error δS in computing $\frac{1}{2}(S+S^{-1})$ is bounded essentially by the error in computing S^{-1}:
$$\|\delta S\| \lesssim \mathbf{u}(\|S\| \cdot \|S^{-1}\|)\|S^{-1}\| = \mathbf{u}\|S\|^3 < \mathbf{u}^{1/2}\|S\|.$$
For the Newton iteration to converge, δS cannot be so large that $S + \delta S$ has pure-imaginary eigenvalues; from the result 1, this means $\|\delta S\| < \|S\|^{-1}$. Therefore, if $\mathbf{u}^{1/2}\|S\| < \|S\|^{-1}$, i.e., $\kappa(S) < \mathbf{u}^{-1/2}$, then Newton iteration (2) will compute S with an absolute error bounded by $\mathbf{u}^{1/2}\|S\|$. □

It is naturally desired to have an analysis from which we know the conditioning of the intermediate matrices A_k in the Newton iteration. It will help us in addressing the question of how to detect the possible appearance of pure-imaginary eigenvalues and to modify or terminate the iteration early if necessary. Unfortunately, it is difficult to make a clean analysis far from convergence because we are unable to relate the error in each step of the iteration to the conditioning of the problem. We can do a coarse analysis, however, in the case that the matrix is diagonalizable.

THEOREM 3.3. *Let A have eigenvalues λ_j (none pure imaginary or zero), right eigenvectors x_j, and left eigenvectors y_j normalized so $\|x_j\| = \|y_j\| = 1$. Let $s_j = \operatorname{sign}(\Re(\lambda_j))$, and*

$$\sigma = \min_j \frac{|y_j^H x_j|}{n} \cdot \frac{|\lambda_j + s_j| - |\lambda_j - s_j|}{|\lambda_j + s_j| + |\lambda_j - s_j|}. \tag{14}$$

Let A_k be the matrix obtained at the kth Newton iteration (2). Then for all k, $\sigma_{\max}(A_k) \le 1/\sigma$ and $\sigma_{\min}(A_k) \ge \sigma$, i.e.,

$$\kappa(A_k) = \frac{\sigma_{\max}(A_k)}{\sigma_{\min}(A_k)} \le \frac{1}{\sigma^2}. \tag{15}$$

Proof. We may express the eigendecomposition of A as $A = \sum_{j=1}^n \lambda_j x_j y_j^H / y_j^H x_j$. Then $A_k = \sum_{j=1}^n \lambda_{j,k} x_j y_j^H / y_j^H x_j$, where $\lambda_{j,k} = \frac{1}{2}(\lambda_{j,k-1}^{-1} + \lambda_{j,k-1})$ with $\lambda_{j,0} = \lambda_j$. We wish to bound $|\lambda_{j,k}|$ from above and below for all k. This is easily done by noting that

$$\frac{\lambda_{j,k+1} - s_j}{\lambda_{j,k+1} + s_j} = \left(\frac{\lambda_{j,k} - s_j}{\lambda_{j,k} + s_j}\right)^2$$

so that all $\lambda_{j,k}$ lie inside a disk defined by

$$\left|\frac{\lambda_{j,k}-s_j}{\lambda_{j,k}+s_j}\right| \le \left|\frac{\lambda_j-s_j}{\lambda_j+s_j}\right| \equiv c_j < 1.$$

This disk is symmetric about the real axis, so its points of minimum and maximum absolute value are both real. Solving for these extreme points yields

$$\frac{1-c_j}{1+c_j} \le |\lambda_{j,k}| \le \frac{1+c_j}{1-c_j}.$$

This means

$$\sigma_{\max}(A_k) = \|A_k\| = \left\|\sum_{j=1}^{n} \lambda_{j,k}\frac{x_j y_j^H}{y_j^H x_j}\right\| \le \sum_{j=1}^{n} \frac{|\lambda_{j,k}|}{|y_j^H x_j|} \le \max_j \frac{n}{|y_j^H x_j|}\cdot\frac{1+c_j}{1-c_j}.$$

Similarly

$$\sigma_{\min}^{-1}(A_k) = \|A_k^{-1}\| = \left\|\sum_{j=1}^{n} \lambda_{j,k}^{-1}\frac{x_j y_j^H}{y_j^H x_j}\right\| \le \sum_{j=1}^{n} \frac{|\lambda_{j,k}^{-1}|}{|y_j^H x_j|} \le \max_j \frac{n}{|y_j^H x_j|}\cdot\frac{1+c_j}{1-c_j},$$

which proves the bound (15). □

As we know, the error introduced at each step of the iteration is mainly caused by the computation of matrix inverse, which is approximately bounded in norm by

$$\mathbf{u}(\kappa(A_k)\|A_k^{-1}\| + \|A_k\|) \le \mathbf{u}(\sigma^{-3}+\sigma^{-1}) \approx \mathbf{u}\sigma^{-3}$$

when $\sigma \ll 1$. If $\mathbf{u}\sigma^{-3} < \sigma_{\min}(A_k)$, then this error cannot make an intermediate A_k become singular and cause the iteration to fail. Our analysis shows that if $\mathbf{u}\sigma^{-3} < \sigma$, or $\sigma > \mathbf{u}^{1/4}$, then the iteration will not fail. This very coarse bound generalizes result 3 of Theorem 2.

We note that if A is symmetric, by the orthonormal eigendecomposition of $A = \sum_{j=1}^{n} \lambda_j q_j q_j^T$, where $q_j^T q_j = 1$, $q_j^T q_k = 0$ if $j \ne k$, then from Theorem 3 we have

$$\sigma = \min_j \begin{cases} \frac{1}{|\lambda_j|} & \text{if } |\lambda_j| \ge 1, \\ |\lambda_j| & \text{if } |\lambda_j| < 1. \end{cases}$$

Therefore,

$$\kappa(A_k) \le \max_j \begin{cases} \lambda_j^2 & \text{if } |\lambda_j| \ge 1, \\ \frac{1}{\lambda_j^2} & \text{if } |\lambda_j| < 1. \end{cases} \tag{16}$$

It shows that when A is symmetric, the condition number of the intermediate matrices A_k, which affects the numerical stability of the Newton iteration, is essentially determined by the square of the distance of the eigenvalues to the imaginary axis.[1]

When A is nonsymmetric and diagonalizable, from Theorem 3.3, we also see that the condition number of the intermediate matrices A_k is related to the norms

[1] A referee predicted that in the symmetric case, the condition number of A_k might be determined only by the distance, not the square of the distance. We were not able to prove such prediction.

of the spectral projectors $P_j = x_j y_j^H/(y_j^H x_j)$ corresponding to the eigenvalues λ_j ($\|P_j\| = 1/|y_j^H x_j|$) and the quantities of the form

$$\tilde{\sigma}_j = \frac{|\lambda_j + s_j| - |\lambda_j - s_j|}{|\lambda_j + s_j| + |\lambda_j - s_j|},$$

where $s_j = \mathrm{sign}(\Re(\lambda_j))$. If we write $\lambda_j = \alpha_j + i\beta_j$, by a simple algebraic manipulation, we have

$$\tilde{\sigma}_j = \frac{1}{2|\alpha_j|}\left[1 + \alpha_j^2 + \beta_j^2 - \sqrt{(\alpha_j^2 - 1)^2 + 2(1 + \alpha_j^2)\beta_j^2 + \beta_j^4}\,\right].$$

From this expression, we see that if there is an eigenvalue λ_j of A very near to the pure imaginary axis, i.e., α_j is small, then by the first-order Taylor expansion of $\tilde{\sigma}_j$ in terms of α_j, we have

$$\tilde{\sigma}_j = \frac{|\alpha_j|}{1 + \beta_j^2} + \mathcal{O}(\alpha_j^2). \tag{17}$$

Therefore, to first order in α_j, the condition numbers of the intermediate matrices A_k satisfy

$$\kappa(A_k) \le \frac{1}{\sigma^2} = \max_j \left(\frac{|\alpha_j|}{n\|P_j\|(1 + \beta_j^2)} + \mathcal{O}\left(\frac{\alpha_j^2}{\|P_j\|}\right)\right)^{-2}. \tag{18}$$

This implies that even if the eigenvalues of A are well conditioned (i.e., the $\|P_j\|$ are not too large), if there are also eigenvalues of A closer to the imaginary axis than $\mathbf{u}^{1/2}$, then the condition number of A_k could be large, $\kappa(A_k) \ge \mathbf{u}^{-1}$, and so the Newton iteration could fail to converge.

4. Backward stability of computed invariant subspace. As discussed in the previous section, because of possible ill conditioning of a matrix with respect to inversion and rounding errors during the Newton iteration, we generally only expect to be able to compute the matrix sign function to the square root of the machine precision, provided that the initial matrix A has condition number smaller than $\mathbf{u}^{-1/2}$. This means that when Newton iteration converges, the computed matrix sign function $\widehat{S}$ satisfies

$$\widehat{S} = S + F \quad \text{with} \quad \|F\| \le \mathcal{O}(\sqrt{\mathbf{u}})\|S\|. \tag{19}$$

Under this assumption, $\widehat{P} = \frac{1}{2}(\widehat{S} + I)$ is an approximate spectral projection corresponding to $\lambda_+(A)$. Therefore, if $\ell = \mathrm{rank}(\widehat{P})$, the first ℓ columns $\widehat{Q}_1$ of $\widehat{Q} \equiv Q + \delta Q = (\widehat{Q}_1,\ \widehat{Q}_2)$ in the rank revealing QR decomposition of $\widehat{P}$ span an approximate invariant subspace. $\widehat{Q}^H A \widehat{Q}$ has the form

$$\widehat{Q}^H A \widehat{Q} = (Q + \delta Q)^H A (Q + \delta Q) = \begin{pmatrix} \widehat{A}_{11} & \widehat{A}_{12} \\ E_{21} & \widehat{A}_{22} \end{pmatrix}$$

with $\lambda(\widehat{A}_{11})$ being the approximate eigenvalues of A in $\mathbf{C}_+$, and $\lambda(\widehat{A}_{22})$ being the approximate eigenvalues of A in $\mathbf{C}_-$. Since we expect the computed matrix sign function to be of half-machine precision, it is reasonable to expect computing the invariant subspace to half-precision too. This in turn means that the backward error $\|E_{21}\|$ in the computed decomposition $\widehat{Q}^H A \widehat{Q}$ is bounded by $\mathcal{O}(\sqrt{\mathbf{u}})\|A\|$, provided that the problem is not very ill conditioned. In this section, we will try to justify such expectation.

To this end, we first need to bound the error in the space spanned by the leading $\ell = \operatorname{rank}(P)$ columns of the transformation matrix Q, i.e., we need to know how much a right singular subspace of the exact projection matrix $P = \frac{1}{2}(S + I)$ is perturbed when P is perturbed by a matrix of norm η. Since P is a projector, the subspace is spanned by the right singular vectors corresponding to all nonzero singular values of P (call the set of these singular values $\mathcal{S}$). In practice, of course, this is a question of rank determination. From the well-known perturbation theory of the singular value decomposition [34, page 260], the space spanned by the corresponding singular vectors is perturbed by at most $\mathcal{O}(\eta)/\mathrm{gap}_{\mathcal{S}}$, where $\mathrm{gap}_{\mathcal{S}}$ is defined by

$$\mathrm{gap}_{\mathcal{S}} \equiv \min_{\substack{\sigma \in \mathcal{S} \\ \bar{\sigma} \notin \mathcal{S}}} |\sigma - \bar{\sigma}| \ .$$

To compute $\mathrm{gap}_{\mathcal{S}}$, we note that there is always a unitary change of basis in which a projector is of the form $\begin{pmatrix} I & \Sigma \\ 0 & 0 \end{pmatrix}$, where $\Sigma = \operatorname{diag}(\sigma_1, \ldots, \sigma_l)$ is diagonal with $\sigma_1 \geq \cdots \geq \sigma_l \geq 0$. By straightforward calculation, we find that the singular values of the projector are $\{\sqrt{1+\sigma_1^2}, \ldots, \sqrt{1+\sigma_\ell^2}, 1, \ldots, 1, 0, \ldots, 0\}$, where the number of ones in the set of singular values is equal to $\max\{2\ell - n, 0\}$. Since $\mathcal{S} = \{\sqrt{1+\sigma_1^2}, \ldots, \sqrt{1+\sigma_\ell^2}, 1, \ldots, 1\}$, we have

$$\mathrm{gap}_{\mathcal{S}} = \begin{cases} \sqrt{1+\sigma_\ell^2} & \text{if } 2\ell \leq n, \\ 1 & \text{if } 2\ell > n. \end{cases}$$

Thus, the error δQ in Q is bounded by

$$\|\delta Q\| \leq \frac{\mathcal{O}(\|F\|)}{\mathrm{gap}_{\mathcal{S}}} \leq \frac{\mathcal{O}(\sqrt{\mathbf{u}})\|S\|}{\mathrm{gap}_{\mathcal{S}}}. \tag{20}$$

Hence, the backward error in the computed spectral decomposition is bounded by

$$\begin{aligned} \|E_{21}\| &\leq \|(Q + \delta Q)^H A (Q + \delta Q) - Q^H A Q\| \\ &= \|\delta Q^H A Q + Q^H A \delta Q + \delta Q^H A \delta Q\| \\ &\leq 2\|\delta Q\| \, \|A\| + \mathcal{O}(\epsilon^2), \end{aligned}$$

where $\mathcal{O}(\epsilon^2)$ is the second-order perturbation term of $\|\delta Q\|$. Therefore, if $2\ell \leq n$, we have the following first-order bound on the backward stability of computed invariant subspace:

$$\frac{\|E_{21}\|}{\|A\|} \leq \frac{\mathcal{O}(\sqrt{\mathbf{u}})\|S\|}{\mathrm{gap}_{\mathcal{S}}} = \frac{\mathcal{O}(\sqrt{\mathbf{u}})\|S\|}{\sqrt{1+\sigma_\ell^2}}. \tag{21}$$

If we use the bound (5) of the matrix sign function S, then from (21) we have

$$\frac{\|E_{21}\|}{\|A\|} \leq \frac{\mathcal{O}(\sqrt{\mathbf{u}})\|A\|}{d_A \sqrt{1+\sigma_\ell^2}}, \tag{22}$$

where d_A, defined in (4), is the distance to the ill-posed problem. On the other hand, if we use the bound (13) for the matrix sign function S, then from (21) again we have

$$\frac{\|E_{21}\|}{\|A\|} \le \frac{\mathcal{O}(\sqrt{\mathbf{u}})\|A\|}{\delta\sqrt{1+\sigma_\ell^2}}, \tag{23}$$

where $\delta = \text{sep}(A_{11}, A_{22})$ is the separation of the matrices A_{11} and A_{22}, if A is assumed to have the form (11). We note that the error bound (23) is essentially the same as the error bound given by Byers, He, and Mehrmann [13], although we use a different approach. In [13], it is assumed that $\|F_{21}\| \lesssim \mathcal{O}(\mathbf{u})\|S\|$ in (19), where F_{21} is the (2,1) block of the matrix F. Therefore, the $\mathcal{O}(\sqrt{\mathbf{u}})$ term in (23) is replaced by $\mathcal{O}(\mathbf{u})$.

The bounds (22) and (23) reveal two important features of the matrix-sign-function-based algorithm for computing the invariant subspace. First, they indicate that the backward error in the computed approximate invariant subspace appears no larger than the absolute error in the computed matrix sign function, provided that the spectral decomposition problem is not very ill conditioned (i.e., d_A or δ is not tiny). Second, if $2\ell \le n$, the backward error is a decreasing function of σ_l. If σ_ℓ is large, this means σ_1 and so $\|P\| = \sqrt{1+\sigma_1^2}$ are large, and this in turn means the eigenvalues close to the imaginary axis are ill conditioned. It is harder to divide these eigenvalues. Of course as they become ill conditioned, d_A decreases at the same time, which must counterbalance the increase in σ_ℓ in a certain range.

It is interesting to ask which error bound (22) and (23) is sharper, i.e., which one of the quantities d_A and $\delta = \text{sep}(A_{11}, A_{22})$ is larger. In [13], an example of a 2×2 matrix is given to show that the quantity δ is larger than the quantity d_A. However, we can also devise simple examples to show that d_A can be larger than $\delta = \text{sep}(A_{11}, A_{22})$. For example, let $A = \text{diag}(A_{11}, A_{22})$ with

$$A_{11} = \begin{pmatrix} \eta & 2 & 3 \\ 0 & \eta & 2 \\ 0 & 0 & \eta \end{pmatrix}, \qquad A_{22} = \begin{pmatrix} -\eta & 2 & 3 \\ 0 & -\eta & 2 \\ 0 & 0 & -\eta \end{pmatrix}.$$

When $\eta = 10^{-3}$, we have $d_A \approx 2.50 \times 10^{-10}$, and $\delta = \text{sep}(A_{11}, A_{22}) \approx 2.81 \times 10^{-16}$. More generally, by choosing A_{11} to be a large Jordan block with a tiny eigenvalue, and $A_{22} = -A_{11}$, d_A is close to the square root of δ. d_A is computed using "numerical brute force" to plot the function $d_A(\tau)$ on a wide range of $\tau \in \mathbb{R}$, and search for the minimal value.

Note that by modifying A to be $A - \sigma I$, where σ is a (sufficiently small) real number, d_A will change but δ will not. Thus, d_A and δ are not completely comparable quantities. We believe d_A to be a more natural quantity to use than δ, since δ does not always depend on the distance to the nearest ill-posed problem. This is reminiscent of the difference between the quantities $\delta = \text{sep}(A_{11}, A_{22})$ and $\text{sep}_\lambda(A_{11}, A_{22})$ [18].

In practice, we will use the a posteriori bound $\|E_{21}\|/\|A\|$ anyway, since if we block upper triangularize $\widehat{Q}^H A \widehat{Q}$ by setting the $(2,1)$ block to zero, $\|E_{21}\|/\|A\|$ is precisely the backward error we introduce.

Before ending this section, let us comment on the stability of the matrix-sign-function-based algorithm versus the QR algorithm. The QR algorithm is a numerical backward stable method for computing the Schur decomposition of a general nonsymmetric matrix A. The computed Schur form $\widehat{T}$ and Schur vectors $\widehat{Q}$ by the QR algorithm satisfy

$$\widehat{Q}^H(A+E)\widehat{Q} = \widehat{T},$$

where E is of the order of $\mathbf{u}\|A\|$. Numerical software for the QR algorithm is available in EISPACK [32] and LAPACK [1]. Although nonconvergent examples have been found, they are quite rare in practice [6, 16]. We note that the eigenvalues on the (block) diagonal of $\widehat{T}$ may appear in any order. Therefore, if an application requires an invariant subspace corresponding to the eigenvalues in a specific region in complex plane, a second step of reordering eigenvalues on the diagonal of $\widehat{T}$ is necessary. A guaranteed stable implementation of this reordering is described in [7].

The matrix-sign-function-based algorithm can be regarded as an algorithm to combine these two steps into one. If the matrix sign function can be computed within the order of $\mathbf{u}\|S\|$, then the analysis in this section shows that the matrix-sign-function-based algorithm could be as stable as the QR algorithm plus reordering. Unfortunately, if the matrix is ill conditioned with respect to matrix inversion (which does not affect the QR algorithm), numerical instability is anticipated in the computed matrix sign function. Therefore, in general, the matrix sign function is less stable than the QR algorithm plus reordering.

5. Numerical experiments.

In this section, we will present numerical examples to verify the above analysis. We will see the numerical stability of the Newton iteration (2) and the backward accuracy of computed spectral decomposition (1) under the influence of the conditioning of the matrix A with respect to inversion, the condition number $\kappa(S)$ of $S = \mathrm{sign}(A)$, and the distance $\Delta(A)$ of the eigenvalues of A to the pure-imaginary axis, where $\Delta(A) = \min_i |\Re(\lambda_i(A))|$. We use the easily computed quantity $\Delta(A)$ as a surrogate of the quantity d_A in (4).

Let us recall that the analysis of sections 3 and 4 essentially claims the following:

(1) If $\Delta(A) < \mathbf{u}^{1/2}$, then the Newton iteration may fail to converge or fail to compute the matrix sign function within the absolute error $\mathbf{u}^{1/2}\|S\|$, even when the matrix sign function is well conditioned. See (18).

(2) If $\kappa(S) > \mathbf{u}^{-1/2}$, then even the distance $\Delta(A)$ is not small, and the Newton iteration may still fail to compute the matrix sign function in the absolute error of $\mathcal{O}(\mathbf{u}^{1/2}\|S\|)$. See part 3 of Theorem 3.2.

(3) In general, the backward error in the computed spectral decomposition will be smaller than the absolute error in the computed matrix sign function. See (21).

The following numerical examples will illustrate these claims. Our numerical experiments were performed on a SUN workstation 10 with machine precision $\varepsilon_M = 2.2204 \times 10^{-16} \approx \mathbf{u}$. All the algorithms are implemented in MATLAB 4.0a. We use the simple Newtion iteration (2) to compute the matrix sign function with the stopping criterion

$$\|A_{k+1} - A_k\| \le 10n\varepsilon_M \|A_k\|.$$

The maximal number of iterations is set to be 70. At the convergence, we have $\lim_{k\to\infty} A_k = \widehat{S}$, the computed matrix sign function. We use the QR decomposition with column pivoting as the rank revealing scheme. $\frac{1}{2}(\widehat{S} + I) = \widehat{Q}\widehat{R}\Pi$, and finally compute

$$\widehat{Q}^H A \widehat{Q} = \begin{pmatrix} \widehat{A}_{11} & \widehat{A}_{12} \\ E_{21} & \widehat{A}_{22} \end{pmatrix},$$

where the first $\ell = \mathrm{rank}(\widehat{R})$ columns of $\widehat{Q}$ spans the invariant subspaces corresponding

Table 1

Numerical results for Example 1.

c	$\Delta(A)=s$	$\kappa(A)$	$\kappa(S)$	iter	$\frac{\Vert S-\bar{S}\Vert}{\Vert S\Vert}$	$\frac{\Vert E_{21}\Vert}{\Vert A\Vert}$
10	1.0e + 00	1.9e + 03	2.7e + 03	7	2.9e − 14	3.9e − 17
	1.0e − 02	8.6e + 02	1.5e + 02	13	8.4e − 14	8.4e − 16
	1.0e − 04	3.8e + 01	8.1e + 01	20	1.3e − 11	1.3e − 13
	1.0e − 06	4.7e + 02	9.0e + 02	30	4.1e − 09	4.1e − 12
	1.0e − 08	4.3e + 02	1.0e + 03	33	2.8e − 07	2.8e − 10
	1.0e − 09	2.8e + 02	1.8e + 03	36	8.0e − 06	8.0e − 09
	1.0e − 10	3.6e + 01	3.7e + 02	40	2.2e − 05	2.2e − 07
	1.0e − 12	5.5e + 01	1.0e + 03	46	4.0e − 03	4.0e − 06
10^3	1.0e − 06	7.8e + 06	1.7e + 07	$26(10^{-11})$	2.1e − 06	5.4e − 12
	1.0e − 08	1.7e + 06	1.0e + 07	$33(10^{-11})$	5.1e − 04	1.8e − 09

to $\lambda(\widehat{A}_{11})$, which are the approximate eigenvalues of A in $\mathbf{C}_+$. $\Vert E_{21}\Vert/\Vert A\Vert$ is the backward error committed by the algorithm.

All our test matrices are constructed of the form

$$A = U^T \begin{pmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{pmatrix} U, \tag{24}$$

where U is an orthogonal matrix generated from the QR decomposition of a random matrix with normal distribution having mean 0.0 and variance 1.0. We will choose different submatrices A_{11}, A_{22}, and A_{12} so that the generated matrices A have different specific features in order to observe our theoretical results in practice.

The exact matrix sign function $S = \text{sign}(A)$ of A and the condition number of S are computed as described in Lemma 3.1. The condition number of A is computed by Matlab function `cond`.

In the following tables, iter is the number of iterations of the Newton iteration. A number 10^α in parenthesis next to an iteration number iter indicates that the convergence of the Newton iteration was stationary about $\mathcal{O}(10^\alpha)$ from the iter^{th} iteration forward, and failed to satisfy the stopping criterion even after the allowed maximal number of iterations.

We have experimented with numerous matrices with different pathological ill conditioning in terms of the distance to the pure-imaginary axis, the condition numbers of $\kappa(A)$ and $\kappa(S)$, and the different values of $\text{sep}(A_{11}, A_{22})$ and so on. Two selected examples presented here are typical of behaviors we observed.

Example 1. In this example, the matrices A are of the form (24) with

$$A_{11} = \begin{pmatrix} s & 1 \\ -1 & s \end{pmatrix}, \quad A_{22} = \begin{pmatrix} -s & 1 \\ -1 & -s \end{pmatrix},$$

and $A_{12} = -(A_{11}R - RA_{22})$, where R is a random 2×2 matrix with normal distribution $(0, 1)$ multiplying by a parameter c. The generated matrix A has two complex conjugate eigenpairs $s \pm i$ and $-s \pm i$. As $s \to 0$, the distance $\Delta(A) = s \to 0$ too. The size of the parameter c will adjust the conditioning of the resulted matrix A and its matrix sign function.

Table 1 reports the computed results for different values of $\Delta(A) = s$. From the table, we see that when the matrices are well conditioned and the corresponding

TABLE 2
Numerical results of Example 2.

d	$\Delta(A)$	$\kappa(A)$	$\kappa(S)$	iter	$\frac{\Vert S-\bar{S}\Vert}{\Vert S\Vert}$	$\frac{\Vert E_{21}\Vert}{\Vert A\Vert}$
1.0	1.2e − 01	5.8e + 02	6.4e + 01	9	1.4e − 14	1.7e − 15
0.7	6.5e − 02	1.3e + 04	1.3e + 04	10	2.0e − 13	1.6e − 14
0.5	8.5e − 02	3.4e + 04	2.2e + 05	$10(10^{-13})$	9.5e − 12	1.7e − 13
0.3	2.0e − 02	3.9e + 06	5.9e + 08	$10(10^{-09})$	3.5e − 09	4.6e − 11
0.25	6.2e − 03	2.9e + 07	1.2e + 09	$13(10^{-10})$	2.8e − 08	1.5e − 10
0.09	1.1e − 02	4.9e + 09	5.5e + 13	$12(10^{-06})$	3.0e − 03	1.0e − 07

matrix sign function is also well conditioned, as stated in the claim (1), the convergence rate and accuracy of the Newton iteration is clearly determined by the distance $\Delta(A)$. When the distance becomes smaller, there is a steady increase in the number of Newton iterations required to convergence and the loss of the accuracy in the computed matrix sign function and, therefore, the desired invariant subspace. From the table, we also see that when both $\Delta(A)$ and $\kappa(S)$ are moderate, the Newton iteration fails to compute the matrix sign function in half-machine precision. Nevertheless, the computed invariant subspace seems to still have half-machine precision; see the claim (3).

Example 2. In this example, the test matrices A are of the form (24). A_{12} are 5×5 $(1,0)$ normally distributed random matrices. The submatrices A_{11} and A_{22} are first set by 5×5 $(1,0)$ normally distributed random upper tridiagonal matrices, and then the diagonal elements of A_{11} and A_{22} are replaced by $d|a_{ii}|$ and $-d|a_{ii}|$, respectively, where $a_{ii}(1 \leq i \leq n)$ are random numbers with normal distribution $(0,1)$, d is a positive parameter. A_{12} are 5×5 $(1,0)$ normally distributed random matrices.

The numerical results are reported in Table 2. For the given parameter d, the eigenvalues are well separated away from the pure-imaginary axis ($\Delta(A)$ is not small), however, as stated in the claim (2), we see the influence of the condition numbers $\kappa(S)$ to the convergence of the Newton iteration and, therefore, the accuracy of the computed matrix sign function and the invariant subspace.

6. Refining estimates of approximate invariant subspaces. When we use the matrix-sign-function-based algorithm to deflate an invariant subspace of matrix A, we end up with the form

$$\widehat{Q}^H A \widehat{Q} = (\widehat{Q}_1,\ \widehat{Q}_2)^H A (\widehat{Q}_1,\ \widehat{Q}_2) = \begin{pmatrix} \widehat{A}_{11} & \widehat{A}_{12} \\ E_{21} & \widehat{A}_{22} \end{pmatrix}, \tag{25}$$

where the size of $\Vert E_{21}\Vert/\Vert A\Vert$ reveals the accuracy and backward stability of computed invariant subspace spanning by $\widehat{Q}_1$ of A. If higher accuracy is desired, we may use iterative refinement techniques to improve the accuracy of computed invariant subspace. The methods are due to Stewart [33], Dongarra, Moler, and Wilkinson [20], and Chatelin [15]. Even though these methods all apparently solve different equations, as shown by Demmel [19], after changing variables, they all solve the same Riccati equation in the inner loop.

Let us follow Stewart's approach to present the first class of methods. From (25), we know that $\widehat{Q}_1$ spans an approximate invariant subspace and $\widehat{Q}_2$ spans an orthogonal complementary subspace. If we let the true invariant subspace be represented by

$\widehat{Q}_1 + \widehat{Q}_2 Y$ and, therefore, its orthogonal complementary subspace as $\widehat{Q}_2 - \widehat{Q}_1 Y^H$, then Y is derived as follows: $\widehat{Q}_1 + \widehat{Q}_2 Y$ will be an invariant subspace if and only if the lower left block of

$$(\widehat{Q}_1 + \widehat{Q}_2 Y,\ \widehat{Q}_2 - \widehat{Q}_1 Y^H)^{-1} A (\widehat{Q}_1 + \widehat{Q}_2 Y,\ \widehat{Q}_2 - \widehat{Q}_1 Y^H)$$

is zero, i.e., if the lower left corner of

$$\begin{pmatrix} I & -Y^H \\ Y & I \end{pmatrix} \begin{pmatrix} \widehat{A}_{11} & \widehat{A}_{12} \\ E_{21} & \widehat{A}_{22} \end{pmatrix} \begin{pmatrix} I & Y^H \\ -Y & I \end{pmatrix}$$

is zero. Thus, Y must satisfy the equation

$$\widehat{A}_{22} Y - Y \widehat{A}_{11} = E_{21} - Y \widehat{A}_{12} Y,$$

which is the well-known algebraic Riccati equation. We may use the following two iterative methods to solve it:

1. the simple Newton iteration

$$\widehat{A}_{22} Y_k - Y_k \widehat{A}_{11} = E_{21} - Y_{k-1} \widehat{A}_{12} Y_{k-1} \tag{26}$$

with $Y_0 = 0$, $k = 1, 2, \ldots$;

2. the modified Newton iteration

$$(\widehat{A}_{22} - Y_{k-1}\widehat{A}_{12}) Y_k - Y_k (\widehat{A}_{11} + \widehat{A}_{12} Y_{k-1}) = -E_{21} - Y_{k-1} \widehat{A}_{12} Y_{k-1} \tag{27}$$

with $Y_0 = 0$, $k = 1, 2, \ldots$.

Therefore, we only need to solve a Sylvester equation in the inner loop of the iterative refinement.

In the following numerical example, we only use the simple Newton iteration (26) to refine the approximate invariant subspace computed by the matrix-sign-function-based algorithm, with the following stopping criterion:

$$\|Y_k - Y_{k-1}\|_1 \le 10 n \varepsilon_M \|Y_{k-1}\|_1.$$

Example 3. We continue Example 2. Table 3 lists the $\mathrm{sep}(A_{11}, A_{22})$, the number of iterative refinement steps, and the backward accuracy of improved invariant subspace.

As shown in the convergence analysis for the iterative solvers (26) and (27) of the Riccati equation by Stewart [33] and Demmel [18], if we let

$$\kappa = (\|\widehat{A}_{12}\|_F \|E_{21}\|_F)/\mathrm{sep}^2(\widehat{A}_{11}, \widehat{A}_{22}),$$

then under the assumptions $k < 1/4$ and $k < 1/12$, the iterations (26) and (27) converge, respectively. Therefore, $\mathrm{sep}(\widehat{A}_{11}, \widehat{A}_{22})$ is a key factor to the convergence of the iterative refinement schemes. The above examples verify such analysis. From the analysis of section 3, we recall that $\mathrm{sep}(\widehat{A}_{11}, \widehat{A}_{22})$ also affects the backward stability of the computed invariant subspace by the matrix-sign-function-based algorithm in the first place (before iterative refinement).

TABLE 3
Iterative refinement results of Example 2.

d	$\mathrm{sep}(A_{11}, A_{22})$	iter	$\frac{\|E'_{21}\|}{\|A\|}$
1.0	2.4e − 2	2	6.6e − 31
0.7	2.4e − 3	3	6.3e − 30
0.5	2.3e − 3	3	1.1e − 28
0.3	2.0e − 5	4	2.0e − 25
0.25	3.8e − 5	4	2.5e − 25
0.09	5.1e − 7	$5(10^{-12})$	1.1e − 21

7. Extension to the generalized eigenproblem. In this section, we outline a scheme to extend the matrix-sign-function-based algorithm to solve the generalized eigenvalue problem of a regular matrix pencil $A - \lambda B$. A matrix pencil $A - \lambda B$ is regular if $A - \lambda B$ is square and $\det(A - \lambda B)$ is not identically zero. In [22], Gardiner and Laub have considered an extension of the Newton iteration for computing the matrix sign function to a matrix pencil for solving generalized algebraic Riccati equations. Here we discuss another possible approach, which includes the computation of both left and right deflating subspaces.

For the given matrix pencil $A - \lambda B$, the problem of the spectral decomposition is to seek a pair of left and right deflating subspaces $\mathcal{L}$ and $\mathcal{R}$ corresponding to the eigenvalues of the pencil in a specified region $\mathcal{D}$ in complex plane. In other words, we want to find a pair of unitary matrices Q_L and Q_R so that if $Q_L = (Q_{L1}, Q_{L2})$, $\mathrm{span}(Q_{L1}) = \mathcal{L}$ and $Q_R = (Q_{R1}, Q_{R2})$, $\mathrm{span}(Q_{R1}) = \mathcal{R}$, then

$$Q_L^H A Q_R = \begin{pmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{pmatrix}, \quad Q_L^H B Q_R = \begin{pmatrix} B_{11} & B_{12} \\ 0 & B_{22} \end{pmatrix}, \tag{28}$$

where the eigenvalues of $A_{11} - \lambda B_{11}$ are the eigenvalues of $A - \lambda B$ in a selected region $\mathcal{D}$ in complex plane. Here, we will only discuss the region $\mathcal{D}$ to be the open *right* half-complex plane. As the same treatment in the standard eigenproblem, by employing Möbius transformations $(\alpha A + \beta B)(\gamma A + \delta B)^{-1}$ and divide-and-conquer, $\mathcal{D}$ can be the union of intersections of arbitrary half-planes and (complemented) disks, and so a rather general region.

To this end, by directly applying the Newton iteration to AB^{-1}, we have

$$Y_{k+1} = \frac{1}{2}(Y_k + Y_k^{-1}), \quad k = 0, 1, 2, \ldots, \quad Y_0 = AB^{-1}.$$

At convergence, $Y_\infty = \mathrm{sign}(AB^{-1})$. In practice, we do not want to invert B if it is ill conditioned. Hence, by letting $Z_k = Y_k B$, then the above iteration becomes

$$Z_{k+1}B^{-1} = \frac{1}{2}(Z_k B^{-1} + BZ_k^{-1}) = \frac{1}{2}(Z_k + BZ_k^{-1}B)B^{-1}.$$

This leads to the following iteration:

$$Z_{k+1} = \frac{1}{2}(Z_k + BZ_k^{-1}B)$$

for $k = 0, 1, 2, \ldots$ with $Z_0 = A$. Z_j converges quadratically to a matrix Z_∞. Then $Z_\infty B^{-1} = Y_\infty = \mathrm{sign}(AB^{-1})$. Next, to find the desired deflating subspace, we use the rank revealing QR decomposition to calculate the range space of the projection $P = \frac{1}{2}(I + Z_\infty B^{-1})$ corresponding to the spectral in the open *right* half-plane, which has the same range space as $2PB = Z_\infty + B$. Thus, by computing the rank revealing QR decomposition of $Z_\infty + B = Q_L R_L \Pi_L$, we obtain the invariant subspace of AB^{-1} without inverting B, i.e.,

$$Q_L^H A B^{-1} Q_L = \begin{pmatrix} C_R & C_{12} \\ 0 & C_L \end{pmatrix}, \tag{29}$$

where $\lambda(C_R)$ are the eigenvalues of the pencil $A - \lambda B$ in the open right half-plane, $\lambda(C_L)$ are the ones of $A - \lambda B$ in the open left half-plane. Therefore, we have obtained the left deflating subspace of $A - \lambda B$.

To compute the right deflating subspace of $A - \lambda B$, we can apply the above idea to $A^H - \lambda B^H$, since transposing swaps right and left spaces. The Newton iteration implicitly applying to $A^H B^{-H}$ turns out to be

$$\tilde{Z}_{k+1} = \frac{1}{2}(\tilde{Z}_k + B^H \tilde{Z}_k^{-1} B^H)$$

for $k = 0, 1, 2, \ldots$ with $Z_0 = A^H$. $\tilde{Z}_j$ converges quadratically to a matrix $\tilde{Z}_\infty$. Using the same arguments as above, after computing the rank revealing QR decomposition of $\tilde{Z}_\infty - B = \tilde{Q}_R R_R \Pi_R$, we have

$$\tilde{Q}_R^H A^H B^{-H} \tilde{Q}_R = \begin{pmatrix} D_L & D_{12} \\ 0 & D_R \end{pmatrix},$$

where $\lambda(D_L)$ are the eigenvalues of the pencil $A - \lambda B$ in the open left half-plane, $\lambda(D_R)$ are the ones of $A - \lambda B$ in the open right half-plane. Note that for the desired spectral decomposition, after transposing, we need to first compute the deflating subspace corresponding to the eigenvalues in the open *left* half-plane. Let $Q_R = \tilde{Q}_R \tilde{\Pi}$, where $\tilde{\Pi}$ is an antidiagonal identity matrix[2]; then we have

$$Q_R^H A^H B^{-H} Q_R = \begin{pmatrix} D_R & 0 \\ D_{12} & D_L \end{pmatrix}. \tag{30}$$

From (29) and (30), we immediately have

$$Q_L^H A Q_R = \begin{pmatrix} C_R & C_{12} \\ 0 & C_L \end{pmatrix} Q_L^H B Q_R, \tag{31}$$

$$Q_L^H A Q_R = Q_L^H B Q_R \begin{pmatrix} D_R^H & D_{12}^H \\ 0 & D_L^H \end{pmatrix}. \tag{32}$$

Let $Q_L^H A Q_R$ and $Q_L^H B Q_R$ have the partitions

$$Q_L^H A Q_R = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}, \quad Q_L^H B Q_R = \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix};$$

we have

$$\begin{pmatrix} C_R & C_{12} \\ 0 & C_L \end{pmatrix} \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix} = \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix} \begin{pmatrix} D_R^H & D_{12}^H \\ 0 & D_L^H \end{pmatrix}.$$

[2] The permutation $\tilde{\Pi}$ can be avoided if we use the rank revealing QL decomposition.

Then B_{21} satisfies

$$C_L B_{21} - B_{21} D_R^H = 0.$$

Note that $\lambda(C_L)$ are the eigenvalues of the pencil $A - \lambda B$ in the open left half-plane, $\lambda(D_R)$ are the eigenvalues of the pencil $A-\lambda B$ in the open right half-plane. Therefore, the above homogeneous Sylvester equation has only the solution $B_{21} = 0$. From (31) or (32), we have $A_{21} = 0$. The computed unitary orthogonal matrices Q_L and Q_R give the desired spectral decomposition (28).

8. Closing remarks.

In this paper, we have presented a number of new results and approaches to further analyze the numerical behavior of the matrix sign function and algorithms using it to compute spectral decompositions of nonsymmetric matrices. From this analysis and numerical experiments, we conclude that if the spectral decomposition problem is not ill conditioned, the algorithm is a practical approach to solve the nonsymmetric eigenvalue problem. Performance evaluation of the matrix-sign-function-based algorithm on parallel distributed memory machines, such as the Intel Delta and CM-5, is reported in [4].

During the course of this work, we have discovered a new approach which essentially computes the same spectral projection matrix as the matrix sign function approach does, and also uses basic matrix operations, namely, matrix multiplication and the QR decomposition. However, it avoids the matrix inverse. From the point of view of accuracy, this is a more promising approach. The new approach is based on the work of Bulgakov and Godunov [10] and Malyshev [27, 28]. In [5], we have improved their results in several important ways, and made it a truly practical and *inverse-free* highly parallel algorithm for both the standard and generalized spectral decomposition problems. In brief, the difference between the matrix sign function and inverse-free methods is as follows. The matrix sign function method is significantly faster than the inverse-free method when it converges, but there are some very difficult problems where the inverse-free algorithm gives a more accurate answer than the matrix sign function algorithm. The interested reader may see paper [5] for details.

Acknowledgments. The authors would like to acknowledge Ralph Byers, Chunyang He, Nick Higham, and Volker Mehrmann for fruitful discussions on the subject. We would also like to thank the referees for their valuable comments on the manuscript.

The information presented here does not necessarily reflect the position or the policy of the U.S. Government and no official endorsement should be inferred.

REFERENCES

[1] E. Anderson, Z. Bai, C. Bischof, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenney, S. Ostrouchov, and D. Sorensen, *LAPACK Users' Guide*, 2nd ed., SIAM, Philadelphia, PA, 1995.

[2] Z. Bai and J. Demmel, *Design of a parallel nonsymmetric eigenroutine toolbox, Part* I, in Proc. Sixth SIAM Conference on Parallel Processing for Scientific Computing, R. F. Sincovec et al., eds., SIAM, Philadelphia, PA, 1993; also available as Computer Science Report CSD-92-718, University of California, Berkeley, CA, 1992.

[3] Z. Bai and J. Demmel, *Design of a Parallel Nonsymmetric Eigenroutine Toolbox, Part* II, Department of Mathematics Research Report 95-11, University of Kentucky, Lexington, KY, 1995.

[4] Z. Bai, J. Demmel, J. Dongarra, A. Petitet, H. Robinson, and K. Stanley, *The spectral decomposition of nonsymmetric matrices on distributed memory parallel computers*, SIAM J. Sci. Comput., 18 (1997), pp. 1446–1461.

[5] Z. Bai, J. Demmel, and M. Gu, *Inverse free parallel spectral divide and conquer algorithms for nonsymmetric eigenproblems*, Numer. Math., 76 (1997), pp. 279–308.

[6] S. Batterson, *Convergence of the shifted QR algorithm on* 3 *by* 3 *normal matrices*, Numer. Math., 58 (1990), pp. 341–352.

[7] A. W. Bojanczyk and P. Van Dooren, *Reordering diagonal blocks in real Schur form*, in Linear Algebra for Large Scale and Real-Time Applications, G. H. Golub, M. S. Moonen, and B. L. R. De Moor, eds., Kluwer Academic Publishers, Amsterdam, 1993.

[8] S. Boyd and V. Balakrishnan, *A regularity result for the singular values of a transfer matrix and a quadratically convergent algorithm for computing its* L_∞*-norm*, Systems Control Lett., 15 (1990), pp. 1–7.

[9] S. Boyd, V. Balakrishnan, and P. Kabamba, *A bisection method for computing the* H_∞ *norm of a transfer matrix and related problems*, Math. Control Signals Systems, 2 (1989), pp. 207–219.

[10] A. Y. Bulgakov and S. K. Godunov, *Circular dichotomy of the spectrum of a matrix*, Siberian Math. J., 29 (1988), pp. 734–744.

[11] R. Byers, *Solving the algebraic Riccati equation with the matrix sign function*, Linear Algebra Appl., 85 (1987), pp. 267–279.

[12] R. Byers, *A bisection method for measuring the distance of a stable matrix to the unstable matrices*, SIAM J. Sci. Statist. Comput., 9 (1988), pp. 875–881.

[13] R. Byers, C. He, and V. Mehrmann, *The Matrix Sign Function Method and the Computation of Invariant Subspaces*, Technical report preprint SPC 94-25, Fakultät für Mathematik, TU Chemnitz-Zwickau, Germany, 1994.

[14] R. Byers and N. K. Nichols, *On the stability radius of a generalized state-space system*, Linear Algebra Appl., 188–189 (1993), pp. 113–134.

[15] F. Chatelin, *Simultaneous Newton's iteration for the eigenproblem*, Comput. Suppl., 5 (1984), pp. 67–74.

[16] D. Day, *How the QR Algorithm Fails to Converge and How to Fix It*, Tech. Rep. SAND 96-09135, Sandia National Laboratories, Albuquerque, NM, 1996.

[17] J. Demmel, *The condition number of equivalence transformations that block diagonalize matrix pencils*, SIAM J. Numer. Anal., 20 (1983), pp. 599–610.

[18] J. Demmel, *On condition numbers and the distance to the nearest ill-posed problem*, Numer. Math., 51 (1987), pp. 251–289.

[19] J. Demmel, *Three methods for refining estimates of invariant subspaces*, Computing, 38 (1987), pp. 43–57.

[20] J. Dongarra, C. Moler, and J. H. Wilkinson, *Improving the accuracy of computed eigenvalues and eigenvectors*, SIAM J. Numer. Anal., 20 (1984), pp. 46–58.

[21] B. S. Garbow, J. M. Boyle, J. J. Dongarra, and C. B. Moler, *Matrix Eigensystem Routines – EISPACK Guide Extension*, Lecture Notes in Comput. Sci. 51, Springer-Verlag, Berlin, 1977.

[22] J. Gardiner and A. Laub, *A generalization of the matrix-sign function solution for algebraic Riccati equations*, Internat. J. Control, 44 (1986), pp. 823–832.

[23] N. J. Higham, *The matrix sign decomposition and its relation to the polar decomposition*, Linear Algebra Appl., 212/213 (1994), pp. 3–20.

[24] T. Kato, *Perturbation Theory for Linear Operators*, 2nd ed., Springer-Verlag, Berlin, 1980.

[25] C. Kenney and A. Laub, *Polar decomposition and matrix sign function condition estimates*, SIAM J. Sci. Statist. Comput., 12 (1991), pp. 488–504.

[26] C. Kenney, A. Laub, and P. Papadopoulos, *Matrix sign function algorithms for Riccati equations*, in Proc. of IMA Conference on Control: Modelling, Computation, Information, Southend-On-Sea, IEEE Press, Piscataway, NJ, 1992, pp. 1–10.

[27] A. N. Malyshev, *Guaranteed accuracy in spectral problems of linear algebra, Parts* I *and* II, Siberian Adv. Math., 2 (1992), pp. 144–197, 153–204.

[28] A. N. Malyshev, *Parallel algorithm for solving some spectral problems of linear algebra*, Linear Algebra Appl., 188/189 (1993), pp. 489–520.

[29] R. Mathias, *Condition estimation for the matrix function via the Schur decomposition*, SIAM J. Matrix Anal. Appl., 16 (1995), pp. 565–578.

[30] J. Roberts, *Linear model reduction and solution of the algebraic Riccati equation*, Internat. J. Control, 32 (1980), pp. 677–687.

[31] L. Shieh, H. Dib, and R. Yates, *Separation of matrix eigenvalues and structural decomposition of large-scale systems*, IEEE Proceedings: Control Theory Appl., 133 (1986), pp. 90–96.

[32] B. T. Smith, J. M. Boyle, J. J. Dongarra, B. S. Garbow, Y. Ikebe, V. C. Klema, and C. B. Moler, *Matrix Eigensystem Routines – EISPACK Guide*, Lecture Notes in Comput. Sci. 6, Springer-Verlag, Berlin, 1976.

[33] G. W. Stewart, *Error and perturbation bounds for subspaces associated with certain eigenvalue problems*, SIAM Rev., 15 (1973), pp. 727–764.

[34] G. W. Stewart and J.-G. Sun, *Matrix Perturbation Theory*, Academic Press, New York, 1990.

[35] L. N. Trefethen, *Pseudospectra of matrices*, in Numerical Analysis 1991, D. F. Griffiths and G. A. Watson, eds., Longman Sci. Tech. Publ., Harlow, UK, 1992, pp. 234–266.