# Design of a Parallel Nonsymmetric Eigenroutine Toolbox, Part II

Zhaojun Bai* and James Demmel†

January 18, 1996

**Abstract**

In Part I of this report [2], we proposed to build a toolbox for solving the dense nonsymmetric eigenvalue problem on distributed memory parallel computers. Part I described basic building blocks, and their uses in different combinations, for extracting the eigenvalues and/or their corresponding invariant subspaces of a nonsymmetric matrix. However, the numerical accuracy and stability of some of our proposed building blocks, such as the matrix sign function, are poorly understood. Part II of this report focusses on the numerical analysis of these building blocks. We discuss perturbation theory of the matrix sign function, the conditioning of its computation, the numerical stability of the overall 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 the generalized eigenvalue problem of a regular matrix pencil is also described.

*Department of Mathematics, University of Kentucky, Lexington, KY 40506. The author was supported in part by ARPA grant DM28E04120 and P-95006 via a subcontract from Argonne National Laboratory and by an NSF grant ASC-9313958 and in part by an DOE grant DE-FG03-94ER25219 via subcontracts from University of California at Berkeley.

†Computer Science Division and Mathematics Department, University of California, Berkeley, CA 94720. The author was funded in part by the Advanced Research Projects Agency (DOD) 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 No. W-31-109-Eng-38 through subcontract No. 951322401 with Argonne National Laboratory, and NSF Infrastructure Grant Nos. CDA-8722788 and CDA-9401156. The information presented here does not necessarily reflect the position or the policy of the Government and no official endorsement should be inferred.

*"Excellence is the product of EXTENSIVE correction."*
EMERY

# 1 Introduction

As stated at the beginning of Part I of this report [2]: "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." 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 matrix sign function), standard eigensolver routines (such as the QR algorithm) and newly developed algorithms (such as the matrix sign function based algorithm). We discussed how these tools can be used in different combinations on different problems and architectures for extracting all or part of eigenvalues of a nonsymmetric matrix and/or their corresponding invariant subspaces. Rather than using "black box" eigenroutines such as provided by EISPACK [28, 19] 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.

However, some of building blocks and algorithms proposed for the toolbox, such as the matrix sign function and divide-and-conquer algorithms based on it, are poorly understood in terms of their numerical accuracy and stability. In this report, we will address these issues. Let us first restate some of basic definitions and ideas in Part I to establish notation.

The matrix sign function of a matrix $A$ is defined as follows [27]: Let $A = X\text{diag}(J_+, J_-)X^{-1}$ 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 (denoted by $\text{sign}(A)$) of $A$ is: $\text{sign}(A) = X\text{diag}(I, -I)X^{-1}$. We assume that no eigenvalue of $A$ lies on the imaginary axis; otherwise, $\text{sign}(A)$ 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 $P_\pm = \frac{1}{2}(I \pm \text{sign}(A))$, 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

$$Q^T A Q = \begin{pmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{pmatrix}, \tag{1.1}$$

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 $\text{sign}(A)$, it can be shown that $\text{sign}(A)$ is the limit of the following Newton iteration

$$A_{k+1} = \frac{1}{2}(A_k + A_k^{-1}), \quad \text{for} \quad k = 0, 1, 2, \ldots,. \quad \text{with} \quad A_0 = A. \tag{1.2}$$

The iteration is globally and ultimately quadratic convergent. There exist different scaling schemes to speedup 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 Part I of this report [2] and the references therein for more details.

Unfortunately, in finite precision arithmetic, the ill-conditioning of a matrix $A_k$ with respect to inversion, together with rounding errors, may destroy the convergence of the Newton iteration (1.2), or cause convergence to the wrong answer. Consequently, the left bottom corner block of the matrix $Q^TAQ$ in (1.1) may be much larger than $\mathbf{u}\|A\|$ where $\mathbf{u}$ denotes machine epsilon. 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 report, 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 Newtion 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 the eigenvalue problem of random matrices and a few eigenvalue problems from applications, such as electrical power system analysis, numerical simulation of chemical reactions, and areodynamics 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, however, they are also often very pessimistic. The worst cases happen rarely.

In addition, in this report, we also discuss a post-processing phase for the 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 the generalized eigenvalue problem of a regular matrix pencil $A - \lambda B$.

The rest of this report is organized as the following: 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 approach, 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. Finally, concluding remarks are presented in section 8.

## 2 A perturbation bound for the matrix sign function

When a matrix $A$ has eigenvalues on 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 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. We will denote this distance by $d_A$.

Perturbation theory and condition number estimation of the matrix sign function are discussed in [23, 21, 26]. 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

$$\|(I - X)^{-1}\| \leq \frac{1}{1 - \|X\|}, \quad \text{if } \|X\| < 1. \tag{2.3}$$

**Theorem 1** *Suppose $A$ has no pure imaginary eigenvalues, $A + \delta A$ is a perturbation of $A$ and $\epsilon \equiv \|\delta A\|$. Let*

$$\omega = \max_{\tau \in \mathbb{R}} \|(i\tau I - A)^{-1}\| = \frac{1}{\min_{\tau \in \mathbb{R}} \sigma_{\min}(i\tau I - A)} \equiv \frac{1}{d_A}. \tag{2.4}$$

*Then*

*1.*

$$\|\mathrm{sign}(A)\| \leq \frac{4}{\pi}\omega\|A\| + 3 \tag{2.5}$$

*2. If*

$$\omega\epsilon < 1, \tag{2.6}$$

*then*

$$\|\mathrm{sign}(A + \delta A) - \mathrm{sign}(A)\| \leq \frac{4}{\pi}\frac{\omega^2\epsilon}{1 - \omega\epsilon}(\|A\| + \epsilon) + 2\frac{\epsilon}{\|A\|}. \tag{2.7}$$

**Proof.** We only prove the bound (2.7). The bound (2.5) can be proved by using a similar technique to that illustrated below.

Following Roberts [27] (or Kato [22]), the matrix sign function can also be defined using Cauchy integral representation:

$$\mathrm{sign}(A) = 2\,\mathrm{sign}^+(A) - I \tag{2.8}$$

where

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

$\Gamma$ 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 $\Gamma$ to be a semi-circle with radius $r = 2\max\{\|A\|, \|A + \delta A\|\}$ (see Figure 1). From the definition (2.8) of $\mathrm{sign}(A)$, it is seen that to study the stability of the matrix sign function of $A$ to the perturbation $\delta A$, it is sufficient to just study the sensitivity of the projection $\mathrm{sign}^+(A)$.

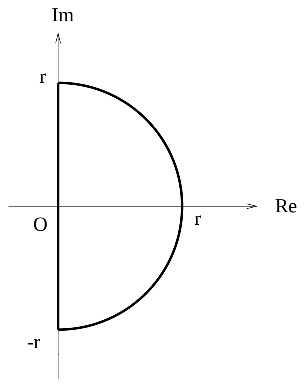


Figure 1: The semi-circle $\Gamma$

Let $\text{sign}^+(A+\delta A)$ be the projection corresponding to $\lambda_+(A+\delta A)$, from the condition (2.6), no eigenvalues of $A$ are perturbed across or on the pure imaginary axis, and the semi-circle $\Gamma$ also encloses $\lambda_+(A+\delta A)$. Therefore we have

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

where the first integral, denoted $\mathcal{I}_1$, is the integral over the straight line of the semi-circle $\Gamma$, the second integral, denoted $\mathcal{I}_2$, is the integral over the curved part of the semi-circle $\Gamma$. Now, by taking the spectral norm of the first integral term, and noting the definition of $\omega$, the condition (2.6) and the inequality (2.3), 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

$$\|\mathcal{I}_2\| \le \frac{1}{2\pi}\int_{-\pi/2}^{\pi/2} \|(re^{i\theta} I - A - \delta A)^{-1}\delta A(re^{i\theta} I - A)^{-1}\|\ r\ |d\theta|$$

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

where the third and fourth inequalities are obtained by the choice of the radius $r$ of the semi-circle $\Gamma$ and the inequality (2.3). The desired bound (2.7) follows from the bounds on $\|\mathcal{I}_1\|$ and $\|\mathcal{I}_2\|$ and the identity

$$\mathrm{sign}(A+\delta A) - \mathrm{sign}(A) = 2(\mathrm{sign}^+(A+\delta A) - \mathrm{sign}^+(A)).$$

∎

A few remarks are in order:

1. In the language of pseudospectra [32], the condition (2.6) means that the $\|\delta A\|$-pseudospectra of $A$ do not cross the pure imaginary axis.
2. From the perturbation bound (2.7), 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 [12, 8, 7, 10].
4. The bound (2.7) is similar to the bound of the norm of the Fréchet derivative of the matrix sign function of $A$ at $X$ given by Roberts [27]:

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

   where $l_\Gamma$ is the length of the closed contour $\Gamma$.

Recently, an asymptotic perturbation bound of $\mathrm{sign}(A)$ was given by Byers, He and Mehrmann [11]. They show that to first order in $\delta A$,

$$\|\mathrm{sign}(A+\delta A) - \mathrm{sign}(A)\| \leq \frac{4}{\delta}\left(1+\frac{\|A_{12}\|}{\delta}\right)^2 \|\delta A\|, \tag{2.9}$$

where $A$ is assumed to have the form of (1.1), $\|\delta A\|$ is sufficiently small, and

$$\delta = \mathrm{sep}(A_{11}, A_{22}) = \sigma_{\min}(I \otimes A_{11} - A_{22}^T \otimes I), \tag{2.10}$$

the separation of the matrices $A_{11}$ and $A_{22}$ [29]. $\otimes$ is the Kronecker product. Comparing the bounds (2.7) and (2.9), we note that first the bound (2.7) is a global bound and (2.9) is an asymptotic bound. Second, the assumption (2.6) for the bound (2.7) has a simple geometric interpretation (see Remark 2 above). It is unspecified how to interpret the assumption on sufficient small $\|\delta\|$ for the bound (2.9). Third, the bound (2.7) is essentially $\mathcal{O}(d_A^{-2})$ and the bound (2.9) is $\mathcal{O}(\delta^{-3})$. Which one is sharper? In section 4, we will compare these two quantities.

## 3 Conditioning of matrix sign function computation

In Part I, 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. In theory, in the worst case, such a deflation procedure may be also needed for the intermediate matrices in the Newton iteration.

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 [16]. 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

$$A = \begin{pmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{pmatrix}, \tag{3.11}$$

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 $B = Q^H AQ$, 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

$$A_{11}R - RA_{22} = -A_{12}, \tag{3.12}$$

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 $\mathrm{sign}(A)$ and its condition number.

**Lemma 1** *Let $A$ and $R$ be as above. Let $\rho = \|R\| + \sqrt{1+\|R\|^2}$. When $\|R\|$ is large, $\rho \approx 2\|P\| \approx 2\|R\|$. 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 (3.12), then $X^{-1}AX = \mathrm{diag}(A_{11}, A_{22})$. Therefore

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

2. Using the singular value decomposition (SVD) of $R$: $URV^H = \Sigma = \mathrm{diag}(\sigma_i)$, 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 by 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 (3.12), we have

$$\|R\| \le \frac{\|A_{12}\|}{\mathrm{sep}(A_{11}, A_{22})},$$

where the equality is attainable [29]. From Lemma 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 $\mathrm{sep}(A_{11}, A_{22})$. Specifically, when $\|R\|$ is large,

$$\|S\| = \|\mathrm{sign}(A)\| \le \frac{2\|A_{12}\|}{\mathrm{sep}(A_{11}, A_{22})} \tag{3.13}$$

and

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

If $\|A_{12}\|$ is moderate, an ill-conditioned matrix sign function means large $\|R\|$, which in turn means small $\mathrm{sep}(A_{11}, A_{22})$. Following Stewart [29], 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 $\mathrm{sign}(A)$, and the distance from $\mathrm{sign}(A)$ to the nearest ill-posed problem.

**Theorem 2** *Let $A$ and $R$ be as in Lemma 1. Then*

1. *Let $\delta S$ have the property that $S + \delta S$ has a pure imaginary eigenvalue. Then $\delta S$ may be chosen with $\|\delta S\| = 1/\|S\|$ but no smaller. In the language of [32], the $\epsilon$-pseudospectrum of $S$ excludes the imaginary axis for $\epsilon < 1/\|S\|$, and intersects it for $\epsilon \ge 1/\|S\|$.*

2. *The condition number of the eigenvalues of $S$ is $\|P\|$. In other words, perturbing $S$ by a small $\delta S$ 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 (1.2) in floating point arithmetic will compute $S$ with an absolute error bounded by $\mathbf{u}^{1/2}\|S\|$. We cannot expect the Newton iteration method to work if $\kappa(S) > \mathbf{u}^{-1/2}$.*

**Proof.**

1. The problem is to minimize $\sigma_{\min}(S - i\zeta I)$ over all real $\zeta$, where $\sigma_{\min}$ is the smallest singular value of $S - i\zeta I$. Using the same unitary similarity and permutation as in the part 1 of Lemma 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 $\sigma_j$ and real $\zeta$. This is a straightforward calculation, with the minimum being obtained for $\zeta = 0$ and $\sigma_j = \|R\|$.

2. The condition number of a semi-simple eigenvalue is equal to the secant of the acute angle between its left and right eigenvectors [22, 15]. Using the above reduction to 2 by 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 $\delta S$ 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, $\delta S$ 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 (1.2) 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 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** *Let $A$ have eigenvalues $\lambda_j$ (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*

$$\sigma = \min_j \frac{|y_j^H x_j|}{n} \cdot \frac{|\lambda_j + s_j| - |\lambda_j - s_j|}{|\lambda_j + s_j| + |\lambda_j - s_j|}. \tag{3.14}$$

*Let $A_k$ be the matrix obtained at the $k$th Newton iteration (1.2). Then for all $k$, $\sigma_{\max}(A_k) \leq 1/\sigma$ and $\sigma_{\min}(A_k) \geq \sigma$, i.e.,*

$$\kappa(A_k) = \frac{\sigma_{\max}(A_k)}{\sigma_{\min}(A_k)} \leq \frac{1}{\sigma^2}. \tag{3.15}$$

**Proof.** We may express the eigen-decomposition 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

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

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

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

This means

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

Similarly

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

which proves the bound (3.15). ∎

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

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

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 so 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 $j \neq k$, then from Theorem 3, we have

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

Therefore,

$$\kappa(A_k) \leq \max_j \begin{cases} \lambda_j^2 & \text{if } |\lambda_j| \geq 1, \\ \frac{1}{\lambda_j^2} & \text{if } |\lambda_j| < 1. \end{cases} \tag{3.16}$$

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.

When $A$ is nonsymmetric and diagonalizable, from Theorem 3, we also see that the condition number of the intermediate matrices $A_k$ is related to the norms of the spectral projectors $P_j = x_j y_j^H/(y_j^H x_j)$ corresponding to the eigenvalues $\lambda_j$ ($\|P_j\| = 1/|y_j^H x_j|$) 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 $\lambda_j$ of $A$ very near to the pure imaginary axis, i.e, $\alpha_j$ is small, then by the first order Taylor expansion of $\tilde{\sigma}_j$ in term of $\alpha_j$, we have

$$\tilde{\sigma}_j = \frac{|\alpha_j|}{1 + \beta_j^2} + \mathcal{O}(\alpha_j^2). \tag{3.17}$$

Therefore, to first order in $\alpha_j$, the condition numbers of the intermediate matrices $A_k$ satisfy

$$\kappa(A_k) \leq \frac{1}{\sigma^2} = \max_j \left( \frac{|\alpha_j|}{n\|P_j\|(1 + \beta_j^2)} + \mathcal{O}\left(\frac{\alpha_j^2}{\|P_j\|}\right)\right)^{-2}. \tag{3.18}$$

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 the intermediate matrices $A_k$ in the Newton iteration could be large, $\kappa(A_k) \leq \mathbf{u}^{-1}$, 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, together with rounding errors during the Newton iteration, we generally only expect to be able to compute the matrix sign function to half machine precision, provided that the initial matrix $A$ has condition number smaller than $\mathbf{u}^{-1/2}$. This means that when Newtion iteration converges, the computed matrix sign function $\bar{S}$ satisfies

$$\bar{S} = S + E \quad \text{with} \quad \|E\| \lesssim \mathcal{O}(\sqrt{\mathbf{u}})\|S\|, \tag{4.19}$$

i.e., we expect and assume that the forward error of the computed matrix sign function is about half machine precision. Under this assumption, $\bar{P} = \frac{1}{2}(\bar{S} + I)$ is an approximate spectral projection corresponding to $\lambda_+(A)$. Therefore, if $l = \mathrm{rank}(\bar{P})$, the first $l$ columns $\bar{Q}_1$ of $\bar{Q} \equiv Q + \delta Q = (\bar{Q}_1,\ \bar{Q}_2)$ in the rank revealing QR decomposition of $\bar{P}$ span an approximate invariant subspace. $A$ is then divided into the form

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

with $\lambda(\bar{A}_{11})$ being the approximate eigenvalues of $A$ in $\mathbf{C}_+$, and $\lambda(\bar{A}_{22})$ 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 $\bar{Q}^T A \bar{Q}$ 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 these expectations.

To this end, we first need to bound the error in the space spanned by the leading $l = \text{rank}(P)$ 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 $\eta$. 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 (see [31, page 260], for example), the space spanned by the corresponding singular vectors is perturbed by at most $\mathcal{O}(\eta)/\text{gap}_{\mathcal{S}}$, where $\text{gap}_{\mathcal{S}}$ is the shortest distance from any singular value in $\mathcal{S}$ to any singular value not in $\mathcal{S}$:

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

In order to compute $\text{gap}_{\mathcal{S}}$, 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 $\Sigma = \text{diag}(\sigma_1, \ldots, \sigma_l)$ 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 $\{\sqrt{1+\sigma_1^2}, \ldots, \sqrt{1+\sigma_l^2}, 1, \cdots, 1, 0, \ldots, 0\}$, where the number of ones in the set of singular values is equal to $\max\{2l-n, 0\}$. Since $\mathcal{S} = \{\sqrt{1+\sigma_1^2}, \ldots, \sqrt{1+\sigma_l^2}, 1, \cdots, 1\}$, we get

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

Thus, we get that the error $\delta Q$ in $Q$ is bounded by

$$\|\delta Q\| \leq \frac{\mathcal{O}(\|E\|)}{\text{gap}_{\mathcal{S}}} = \frac{\mathcal{O}(\sqrt{\mathbf{u}})\|S\|}{\text{gap}_{\mathcal{S}}}. \tag{4.20}$$

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 $2l \leq n$, we have the following first order bound on the backward stability of computed invariant subspace:

$$\frac{\|E_{21}\|}{\|A\|} \leq \frac{\mathcal{O}(\sqrt{\mathbf{u}})\|S\|}{\text{gap}_{\mathcal{S}}} = \frac{\mathcal{O}(\sqrt{\mathbf{u}})\|S\|}{\sqrt{1+\sigma_l^2}}. \tag{4.21}$$

If we use the bound (2.5) of the matrix sign function $S$, then from (4.21), we have

$$\frac{\|E_{21}\|}{\|A\|} \le \frac{\mathcal{O}(\sqrt{\mathbf{u}})\|A\|}{d_A\sqrt{1+\sigma_l^2}}, \tag{4.22}$$

where $d_A$, defined in (2.4), is the distance to the ill-posed problem. On the other hand, if we use the bound (3.13) for the matrix sign function $S$, then from (4.21) again, we have

$$\frac{\|E_{21}\|}{\|A\|} \le \frac{\mathcal{O}(\sqrt{\mathbf{u}})\|A\|}{\delta\sqrt{1+\sigma_l^2}}, \tag{4.23}$$

where $\delta = \mathrm{sep}(A_{11}, A_{22})$ is the separation of the matrices $A_{11}$ and $A_{22}$, if $A$ is assumed to have the form (3.11). We note that the error bound (4.23) is essentially the same as the error bound given by Byers, He and Mehrmann [11], although we use a different approach. In [11], it is assumed that $\|E\| \lesssim \mathcal{O}(\mathbf{u})\|S\|$ in (4.19). Therefore, $\mathcal{O}(\sqrt{\mathbf{u}})$ term in (4.23) is replaced by $\mathcal{O}(\mathbf{u})$. We consider the $\mathcal{O}(\mathbf{u})\|S\|$ accuracy for the computed matrix sign function to be an unrealistic assumption.

The bounds (4.22) and (4.23) 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 $\delta$ is not tiny). Second, if $2l \le n$, the backward error is a decreasing function of $\sigma_l$. If $\sigma_l$ is large, this means $\sigma_1$ and so $\|P\| = \sqrt{1+\sigma_1^2}$ are large, and this in turn means the eigenvalues close 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 $\sigma_l$ in a certain range.

It is interestingn to ask whether error bound (4.22) or (4.23) is sharper, i.e., which one of the quantities $d_A$ and $\delta = \mathrm{sep}(A_{11}, A_{22})$ is larger. In [11], an example of a 2 by 2 matrix is given to show that the quantity $\delta$ is larger than the quantity $d_A$. However, we can also devise simple examples to show that $d_A$ can be larger than $\delta = \mathrm{sep}(A_{11}, A_{22})$. For example, let $A = \mathrm{diag}(A_{11}, A_{22})$ and

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

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

Note that by modifying $A$ to be $A - \sigma I$, where $\sigma$ is a (sufficiently small) real number, $d_A$ will change, but $\delta$ will not. Thus $d_A$ and $\delta$ are not completely comparable quantities, and care is needed in comparing them. We believe $d_A$ to be a more natural quantity to use than $\delta$, since $\delta$ 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 $\text{sep}_\lambda(A_{11}, A_{22})$ [16].

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

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

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

where $E$ is of the order of $\mathbf{u}\|A\|$. Numerical software for the QR algorithm is available in EISPACK [28] and LAPACK [1]. Although nonconvergent examples have been found, they are quite rare in practice [5, 14]. We note that the eigenvalues on the (block)-diagonal of $\bar{T}$ 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 re-ordering eigenvalues on the diagonal of $\bar{T}$ is necessary. A guaranteed stable implementation of this reordering is described in [6].

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 basked 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 unstable is anticipated in the computed matrix sign function. Therefore, in general, the matrix sign function is less stable than the the QR algorithm plus reordering.

# 5 Numerical Experiments

In this section, we will present numerical examples to verify the our analysis above. We will see that the numerical stability of the Newton iteration (1.2) and the backward accuracy of computed spectral decomposition (1.1) under the influence of the conditioning of the matrix $A$ with respect to inversion, the condition number $\kappa(S)$ of the matrix sign function $S = \text{sign}(A)$ 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 an analog of the quantity $d_A$ in (2.4).

The analysis of sections 3 and 4 essentially claims that

(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 (3.18).

(2) If $\kappa(S) > \mathbf{u}^{-1/2}$, then even the distance $\Delta(A)$ is not small, 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 the part 3 of Theorem 2.

(3) Generally speaking, the backward error in the computed spectral decomposition will be smaller than the absolute error in the computed matrix sign function. See (4.21).

The following numerical examples will illustrate these claims.

Our numerical experiments were performed on a SUN workstation 1+ 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 (1.2) to compute the matrix sign function with the stopping criterion

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

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

$$\bar{Q}^T A \bar{Q} = \begin{pmatrix} \bar{A}_{11} & \bar{A}_{12} \\ E_{21} & \bar{A}_{22} \end{pmatrix}.$$

where the first $l = \operatorname{rank}(\bar{R})$ columns of $\bar{Q}$ spans the invariant subspaces corresponding to $\lambda(\bar{A}_{11})$, which are the approximate eigenvalues of $A$ in $\mathbf{C}_+$. $\|E_{21}\|/\|A\|$ is the backward error committed by the algorithm.

All our test matrices are constructed of the form

$$A = U^T \begin{pmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{pmatrix} U, \tag{5.24}$$

where $U$ is an orthogonal matrix generated from the QR decomposition 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 $S = \operatorname{sign}(A)$ of $A$ and the condition number of $S$ are computed as described in Lemma 1. The condition of $A$ is computed by Matlab function cond.

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

We have experimented numerous matrices with different pathologically 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 $\operatorname{sep}(A_{11}, A_{22})$ and so on. Two selected examples presented here are of typical behaviors we observed.

**Example 1**. In this example, the matrices $A$ are of the form (5.24) with

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

Table 1: Numerical Results for Example 1

| $c$ | $\Delta(A)=s$ | $\kappa(A)$ | $\kappa(S)$ | iter | $\frac{\|S-\tilde{S}\|}{\|S\|}$ | $\frac{\|E_{21}\|}{\|A\|}$ |
|---|---|---|---|---|---|---|
| 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 | $4.0e-03$ | $4.0e-06$ |
| $10^3$ | $1.0e-06$ | $7.8e+06$ | $1.7e+07$ | $26(10^{-11})$ | $2.1e-06$ | $5.4e-12$ |
| | $1.0e-08$ | $1.7e+06$ | $1.0e+07$ | $33(10^{-11})$ | $5.1e-04$ | $1.8e-09$ |

The solution $R$ of the Sylvester equation (3.12) is a random 2 by 2 matrix with normal distribution $(0,1)$ multiplying by a parameter $c$. Then let $A_{12} = -(A_{11}R - RA_{22})$.

The generated matrix $A$ have 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 Table 1, we see that when the generated test matrices are well conditioned ($\kappa(A)$ is small) and the corresponding matrix sign function is well-conditioned too ($\kappa(S)$ is small), 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 the Newton iteration required to convergence and the loss of the accuracy in the computed matrix sign function and therefore the desired invariant subspace.

¿From Table 1, 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 still have half machine precision, see the claim (3).

**Example 2**. In this example, the test matrices $A$ are of the form (5.24). $A_{12}$ are 5 by 5 $(1,0)$–normally distributed random matrices. The submatrices $A_{11}$ and $A_{22}$ are first set by 5 by 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 $a_{ii} (1 \le i \le n)$ are random numbers with normal distribution $(0,1)$, $d$ is a positive parameter. $A_{12}$ are 5 by 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

Table 2: Numerical Results of Example 2

| $d$ | $\Delta(A)$ | $\kappa(A)$ | $\kappa(S)$ | iter | $\frac{\|S-\bar{S}\|}{\|S\|}$ | $\frac{\|E_{21}\|}{\|A\|}$ |
|---|---|---|---|---|---|---|
| 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 | $1.1e-02$ | $4.9e+09$ | $5.5e+13$ | $12(10^{-06})$ | $3.0e-03$ | $1.0e-07$ |

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

$$\bar{Q}^T A \bar{Q} = (\bar{Q}_1,\ \bar{Q}_2)^T A (\bar{Q}_1,\ \bar{Q}_2) = \begin{pmatrix} A_{11} & A_{12} \\ E_{21} & A_{22} \end{pmatrix}, \tag{6.25}$$

where the size of $\|E_{21}\|/\|A\|$ reveals the accuracy and backward stability of computed invariant subspace spanning by $\bar{Q}_1$ of $A$. If higher accuracy is desired, we may use iterative refinement techniques to improve the accuracy of computed invariant subspace. There are two classes of methods, the first one is due to Stewart [29], Dongarra, Moler and Wilkinson [18], and Chatelin [13]. Even though these methods all apparently solve different equations, as shown by Demmel [17], after changing variables, they all solve the same Riccati equation in the inner loop. The second class of methods can be described as a Jacobi type technique to compute the Schur decomposition of a nonhermitian matrix [30], although originally, it is not motivated for refining estimates of an invariant subspace. Let us examine these two method in details.

## 6.1 Newton iterative refinement

Let us follow Stewart's approach to present the first class of methods. ¿From (6.25), we know that $\bar{Q}_1$ spans an approximate invariant subspaces and $\bar{Q}_2$ spans an orthogonal complementary subspace. If let the true invariant subspace is represented as $\bar{Q}_1 + \bar{Q}_2 Y$ and therefore its orthogonal complementary subspace as $\bar{Q}_2 - \bar{Q}_1 Y^H$. Then $Y$ is derived as follows: $\bar{Q}_1 + \bar{Q}_2 Y$ will be an invariant subspace if and only if the lower left block of

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

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

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

is zero. Thus, $Y$ must satisfy the equation

$$A_{22}Y - YA_{11} = E_{21} - YA_{12}Y,$$

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

1. The simple Newton iteration:
$$A_{22}Y_k - Y_k A_{11} = E_{21} - Y_{k-1}A_{12}Y_{k-1} \tag{6.26}$$
with $Y_0 = 0$, $k = 1, 2, \ldots,$.

2. The modified Newton iteration:
$$(A_{22} - Y_{k-1}A_{12})Y_k - Y_k(A_{11} + A_{12}Y_{k-1}) = -E_{21} - Y_{k-1}A_{12}Y_{k-1} \tag{6.27}$$
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 following numerical example, we only use the simple Newton iteration (6.26) to refine the approximate invariant subspace computed by matrix sign function based algorithm, with the following stopping criterion:

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

**Example 3** We continue the 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 (6.26) and (6.27) of the Riccati equation by Stewart [29] and Demmel [16], if we let $\kappa = (\|A_{12}\|_F \|E_{21}\|_F)/\mathrm{sep}^2(A_{11}, A_{22})$, then under the assumptions $k < 1/4$ and $k < 1/12$, the iterations (6.26) and (6.27) converge, respectively. Therefore, $\mathrm{sep}(A_{11}, A_{22})$ 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 $\mathrm{sep}(A_{11}, A_{22})$ also affects the backward stability of the computed invariant subspace by the matrix sign function based algorithm in the first place (before iterative refinement).

## 6.2 Jacobi iterative refinement

The second iterative refinement scheme is to use a Jacobi-like algorithm. The method we describe here is due to Stewart [30], which was proposed to compute the Schur decomposition of a nonhermitian matrix in parallel. In the framework of iterative refinement of an

Table 3: Iterative Refinement Reresults of Example 2

| $d$ | $\text{sep}(A_{11}, A_{22})$ | 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 | $5.1e-7$ | $5(10^{-12})$ | $1.1e-21$ |

approximate invariant subspace, we may further assume that the matrices $A_{11}$ and $A_{22}$ in (6.25) are in Schur forms, i.e., the matrix $\bar{Q}^T A \bar{Q}$ is of the form

$$\begin{pmatrix} \times & \times & \times & \times & \times & \times \\ \epsilon_1 & \times & \times & \times & \times & \times \\ \epsilon_1 & \epsilon_1 & \times & \times & \times & \times \\ \epsilon_2 & \epsilon_2 & \epsilon_2 & \times & \times & \times \\ \epsilon_2 & \epsilon_2 & \epsilon_2 & \epsilon_1 & \times & \times \\ \epsilon_2 & \epsilon_2 & \epsilon_2 & \epsilon_1 & \epsilon_1 & \times \end{pmatrix},$$

where $\epsilon_1$ is at the order of $\mathbf{u}\|A\|$, and $\epsilon_2$ is about the order of $\mathbf{u}^{1/2}\|A\|$. In this situation, Stewart's complex Jacobi-Schur method can be applied to eliminate the strictly lower triangular entries. The scheme is ultimately quadratic convergent. Therefore, in general, we can expect to reduce the $\epsilon_2$ entries to the order of $\epsilon_1$ entries in only about one to two sweeps. The method is fine grain and suitable for parallel implementation [30, 33]. However, it is not clear whether we only need real arithmetic (for the real Schur decomposition) if the matrix $A$ is real.

# 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 [20], 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 $\mathcal{R}$ (i.e., $A\mathcal{R} \subseteq \mathcal{L}$ and $B\mathcal{R} \subseteq \mathcal{L}$) 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 $Q_L =$

$(Q_{L1}, Q_{L2})$, $\text{span}(Q_{L1}) = \mathcal{L}$ and $Q_R = (Q_{R1}, Q_{R2})$, $\text{span}(Q_{R1}) = \mathcal{R}$, then

$$Q_L^H A Q_R = \begin{pmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{pmatrix}, \quad Q_L^H B Q_R = \begin{pmatrix} B_{11} & B_{12} \\ 0 & B_{22} \end{pmatrix}, \tag{7.28}$$

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, $Y_\infty = \text{sign}(AB^{-1})$. 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 $Z_\infty$. Then $Z_\infty B^{-1} = Y_\infty = \text{sign}(AB^{-1})$. 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.,

$$Q_L^H AB^{-1} Q_L = \begin{pmatrix} C_R & C_{12} \\ 0 & C_L \end{pmatrix}, \tag{7.29}$$

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 applying 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 anti-diagonal identity matrix[1], then we have

$$Q_R^H A^H B^{-H} Q_R = \begin{pmatrix} D_R & 0 \\ D_{12} & D_L \end{pmatrix}, \tag{7.30}$$

¿From (7.29) and (7.30), we immediately have

$$Q_L^H A Q_R = \begin{pmatrix} C_R & C_{12} \\ 0 & C_L \end{pmatrix} Q_L^H B Q_R \tag{7.31}$$

$$Q_L^H A Q_R = Q_L^H B Q_R \begin{pmatrix} D_R^H & D_{12}^H \\ 0 & D_L^H \end{pmatrix}. \tag{7.32}$$

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

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

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 solution $B_{21}=0$. From (7.31) or (7.32), we have $A_{21}=0$. The computed unitary orthogonal matrices $Q_L$ and $Q_R$ give the desired spectral decomposition (7.28).

# 8 Closing Remarks

In this report, 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 [3].

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,

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

and also uses basic matrix operations, namely, matrix multiplication and the QR decomposition. However, it avoids the matrix inverse. Numerically, this is a more promising approach. The new approach is based on the work of Bulgakov, Godunov and Malyshev [9, 24, 25]. In [4], 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 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 the report [4] for details.

## 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 (second edition)*. SIAM, Philadelphia, 1995.

[2] Z. Bai and J. Demmel. Design of a parallel nonsymmetric eigenroutine toolbox, Part I. In R. F. Sincovec *et al*, editor, *Proceedings of the Sixth SIAM Conference on Parallel Processing for Scientific Computing*. SIAM, 1993. Long version available as Computer Science Report CSD-92-718, University of California, Berkeley, 1992.

[3] Z. Bai, J. Demmel, J. Dongarra, A. Petitet, H. Robinson, and K. Stanley. The spectral decomposition of nonsymmetric matrices on distributed memory parallel computers. Computer Science Department Techical Report CS-95-273, University of Tennessee at Knoxville, 1995. To appear in SIAM J. Sci. Comp.

[4] Z. Bai, J. Demmel, and M. Gu. Inverse free parallel spectral divide and conquer algorithms for nonsymmetric eigenproblems. Department of Mathematics RR 94-01, University of Kentucky, 1994. To appear in Numer. Math.

[5] S. Batterson. Convergence of the shifted QR algorithm on 3 by 3 normal matrices. *Num. Math.*, 58:341–352, 1990.

[6] A. Bojanczyk and P. Van Dooren. personal communication, 1993.

[7] S. Boyd and V. Balakrishnan. A regularity result for the singular values of a transfer matrix and a quadratically convergent algorithm for computing its $\mathbf{L}_\infty$-norm. *Systems Control Letters*, 15:1–7, 1990.

[8] S. Boyd, V. Balakrishnan, and P. Kabamba. A bisection method for computing the $\mathbf{H}_\infty$ norm of a transfer matrix and related problems. *Mathematics of Control, Signals, and Systems*, 2(3):207–219, 1989.

[9] A. Ya. Bulgakov and S. K. Godunov. Circular dichotomy of the spectrum of a matrix. *Siberian Math. J.*, 29(5):734–744, 1988.

[10] R. Byers. A bisection method for measuring the distance of a stable matrix to the unstable matrices. *SIAM J. Sci. Stat. Comp.*, 9(5):875–881, 1988.

[11] 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.

[12] R. Byers and N. K. Nichols. On the stability radius of a generalized state-space system. *Lin. Alg. Appl.*, 188–189:113–134, 1993.

[13] F. Chatelin. Simultaneous Newton's iteration for the eigenproblem. *Comput. Suppl.*, 5:67–74, 1984.

[14] D. Day. How the QRalgorithm fails to converge and how to fix it. in preparation.

[15] J. Demmel. The condition number of equivalence transformations that block diagonalize matrix pencils. *SIAM J. Num. Anal.*, 20(3):599–610, 1983.

[16] J. Demmel. On condition numbers and the distance to the nearest ill-posed problem. *Num. Math.*, 51(3):251–289, 1987.

[17] J. Demmel. Three methods for refining estimates of invariant subspaces. *Computing*, 38:43–57, 1987.

[18] J. Dongarra, C. Moler, and J. H. Wilkinson. Improving the accuracy of computed eigenvalues and eigenvectors. *SIAM J. Num. Anal.*, 20:46–58, 1984.

[19] B. S. Garbow, J. M. Boyle, J. J. Dongarra, and C. B. Moler. *Matrix Eigensystem Routines – EISPACK Guide Extension*, volume 51 of *Lecture Notes in Computer Science*. Springer-Verlag, Berlin, 1977.

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

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

[22] T. Kato. *Perturbation Theory for Linear Operators.* Springer Verlag, Berlin, 2 edition, 1980.

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

[24] A. N. Malyshev. Guaranteed accuracy in spectral problems of linear algebra, I,II. *Siberian Adv. in Math.*, 2(1,2):144–197,153–204, 1992.

[25] A. N. Malyshev. Parallel algorithm for solving some spectral problems of linear algebra. *Lin. Alg. Appl.*, 188,189:489–520, 1993.

[26] R. Mathias. Condition estimation for the matrix sign function via the Schur decomposition. In J. G. Lewis, editor, *Proceedings of the Fifth SIAM Conference on Applied Linear Algebra.* SIAM, 1994.

[27] J. Roberts. Linear model reduction and solution of the algebraic Riccati equation. *Inter. J. Control*, 32:677–687, 1980.

[28] 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*, volume 6 of *Lecture Notes in Computer Science*. Springer-Verlag, Berlin, 1976.

[29] G. W. Stewart. Error and perturbation bounds for subspaces associated with certain eigenvalue problems. *SIAM Review*, 15(4):727–764, 1973.

[30] G. W. Stewart. A Jacobi-like algorithm for computing the Schur decomposition of a non-Hermitian matrix. *SIAM J. Sci. Stat. Comput.*, 6:853–864, 1985.

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

[32] L. N. Trefethen. Pseudospectra of matrices. In *1991 Dundee Numerical Analysis Conference Proceedings*, Dundee, Scotland, June 1991.

[33] R. van de Geijn. A novel storage scheme for parallel jacobi methods. Computer Science Dept. TR-88-26, University of Texas, Austin, TX, July 1988.