University of Maryland College Park

Institute for Advanced Computer Studies TR–2007–45
Department of Computer Science TR–4890

Analysis of the Residual Arnoldi Method*

Che-Rung Lee†
G. W. Stewart‡

October 2007

ABSTRACT

The Arnoldi method generates a nested sequences of orthonormal bases $U_1, U_2, \ldots$ by orthonormalizing Au_k against U_k. Frequently these bases contain increasingly accurate approximations of eigenparis from the periphery of the spectrum of A. However, the convergence of these approximations stagnates if u_k is contaminated by error. It has been observed that if one chooses a Rayleigh–Ritz approximation (μ_k, z_k) to a chosen target eigenpair (λ, x) and orthonormalizes the residual $Az_k - \mu_k z_k$, the approximations to x (but not the other eigenvectors) continue to converge, even when the residual is contaminated by error. The same is true of the shift-invert variant of Arnoldi's method. In this paper we give a mathematical analysis of these new methods.

*This report is available by anonymous ftp from `thales.cs.umd.edu` in the directory `pub/reports` or on the web at `http://www.cs.umd.edu/~stewart/`. This work was supported in part by the National Science Foundation under grant CCR0204084.

†Department of Computer Science, University of California, Davis, CA 95616 (`leech@cs.ucdavis.edu`).

‡Department of Computer Science and Institute for Advanced Computer Studies, University of Maryland, College Park, MD 20742 (`stewart@cs.umd.edu`).

Analysis of the Residual Arnoldi Method

Che-Rung Lee
G. W. Stewart

ABSTRACT

The Arnoldi method generates a nested squences of orthonormal bases $U_1, U_2, \ldots$ by orthonormalizing Au_k against U_k. Frequently these bases contain increasingly accurate approximations of eigenparis from the periphery of the spectrum of A. However, the convergence of these approximations stagnates if u_k is contaminated by error. It has been observed that if one chooses a Rayleigh–Ritz approximation (μ_k, z_k) to a chosen target eigenpair (λ, x) and orthonormalizes the residual $Az_k - \mu_k z_k$, the approximations to x (but not the other eigenvectors) continue to converge, even when the residual is contaminated by error. The same is true of the shift-invert variant of Arnoldi's method. In this paper we give a mathematical analysis of these new methods.

1. Introduction

This paper is concerned with modifications of the classic Arnoldi/Rayleigh–Ritz method for approximating eigenpairs of a large matrix A. To fix our notation, we will begin with a brief description of the two algorithms that underlie the new methods.

The Arnoldi algorithm proceeds by building orthonormal bases for a sequence of nested subspaces as follows (see [1] and [10, §5]). The basis U_1 consists of a normalized starting vector u_1. Given the orthonormal basis $U_k = (u_1 \ \cdots \ u_k)$, we generate U_{k+1} as follows.

1. Set $v_k = Au_k$.
2. Orthonormalize v_k against U_k to get u_{k+1}.[1] (1.1)
3. $U_{k+1} = (u_1 \ \cdots \ u_{k+1})$.

The results of the procedure can be summarized by the ARNOLDI RELATION

$$AU_k = U_k \hat{H}_k + \rho_k u_{k+1} \mathbf{e}_k^*, \tag{1.2}$$

in which $\hat{H}_k$ is an upper Hessenberg matrix whose elements along with ρ_k are the coefficients from the orthonormalizations and $\mathbf{e}_k$ is the kth unit vector of dimension k.

[1] This step can fail if the projection of v_k lies in the the column space of U_k, in which case the Arnoldi procedure terminates and U_k spans an invariant subspace of A. We will not consider this case in what follows.

Conversely, the existence of such a decomposition implies that U_k was obtained by the Arnoldi process (1.1).

Unless the starting vector u_1 is poorly chosen, the KRYLOV SUBSPACES $\mathcal{K}_k$ spanned by the U_k tend to contain increasingly accurate approximations to eigenvectors corresponding to eigenvalues on the periphery of of the spectrum of A. These eigenpairs can be extracted by the Rayleigh–Ritz algorithm (see [10, §4.4]). Here we will focus on obtaining an approximation (μ, z) to a TARGET EIGENEPAIR (λ, x).

1. Compute the RAYLEIGH QUOTIENT $H_k = U_k^* A U_k$.
2. Of the eigenpairs of (μ_i, w_i) of H_k choose a CANDIDATE pair, say (μ, w), such that μ approximates the target eigenvalue λ.
3. Let $(\mu, z) = (\mu, U_k w)$.

(1.3)

The pair (μ, z) is called a RITZ PAIR, and its components μ and z are its RITZ VALUE and RITZ VECTOR. The vector w is called a PRIMITIVE RITZ VECTOR. If U_k in (1.3) is the same as in (1.2), then it is easy to see that $\hat{H}_k = H_k$. The reason for making the distinction is that in our new methods we will always compute the Rayleigh quotient directly from U_k as in (1.3).

If all goes well, the Rayleigh-Ritz candidates converge to their respective targets. However, existing convergence theory is rather weak [7] and tends to understate the power of the Arnoldi/Rayleigh–Ritz method (which has served as a basis for widely used packages, such as ARPACK [6]).

The method is not good at finding eigenpairs from the interior of the spectrum of A. The cure is to choose a SHIFT σ near the desired eigenvalues and work with the matrix

$$S = (A - \sigma I)^{-1}. \tag{1.4}$$

This transformation moves the eigenvalues of A near the shift to the periphery of the spectrum of S, where good convergence can be expected. Thus the SHIFT-AND-INVERT Arnoldi procedure goes as follows.

1. Solve the system $(A - \sigma I)v_k = u_k$.
2. Orthonormalize v_k against U_k to get u_{k+1}.
3. $U_{k+1} = (u_1 \ \cdots \ u_{k+1})$.

(1.5)

The Rayleigh-Ritz procedure can be applied directly to S using the Rayleigh quotient formed from the Arnoldi process. Alternatively, one can use the Rayleigh quotient $U_k^* A U_k$, at the cost of having to compute and store the vectors Au_i. In our modified algorithm we will do the latter — indeed it is essential for the algorithm to work.

An important difficulty with the Arnoldi method, is that the process is sensitive to errors in the computation of v_{k+1}. Specifically, if v_{k+1} is computed to relative error ϵ, the Ritz pairs stagnate at approximately the same level. This is illustrated by the first plot in Figure 1.1. In these experiments, A is a matrix of order 100 with eigenvalues

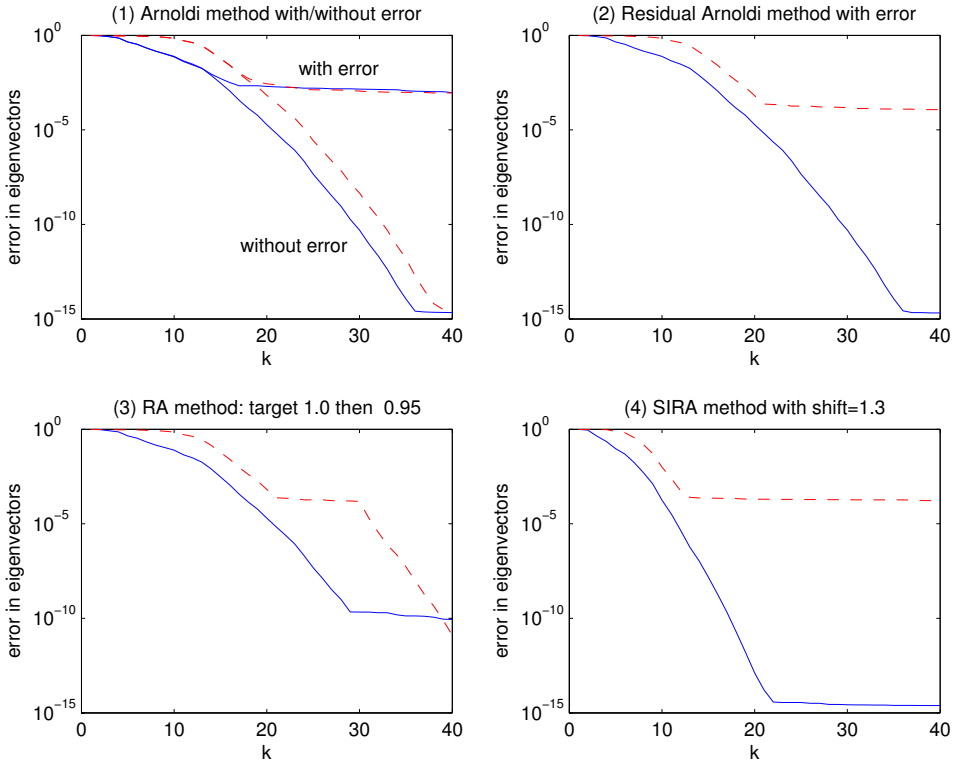

Figure 1.1: Arnoldi and residual Arnoldi methods

$1.00, 0.95, 0.95^2, \ldots, 0.95^{99}$. We consider the eigenvalue 1.00 and 0.95 as targets. The errors in the candidate eigenvectors are plotted against the iteration number k — solid lines representing $\lambda = 1.00$ and dashed lines representing $\lambda = 0.95$.

The lines labeled 'without error' show the ordinary course of the Arnoldi/Rayleigh–Ritz method. Both approximations converge to an accuracy of 10^{-15} — close to best that can be expected from computations in IEEE double-precision arithmetic. Note that the curves are slightly concave, indicating a mildly superlinear convergence, which is not predicted by the theory (however, see [2]). The pair of lines labeled 'with error' show what happens when we add a relative error of 10^{-3} into v_k. Both approximations converge to about 10^{-3} and then stagnate. Although we have introduced errors at every stage, a single error in u_2 will also cause the stagnation.[2]

[2]This is not entirely unexpected. The convergence theory depends critically on the fact that any vector in the column space of U_k can be written in the form $p_{k-1}(A)u_1$, where $p_{k-1}(A)$ is a polynomial of degree not greater than $k-1$. A single error in any u_j will contaminate this relation for all subsequent

As a partial solution to this problem, we propose that we choose a specific target eigenpair (λ, x) and replace v_k in (1.1) by

$$r_k = Az_k - \mu_k z_k,$$

where (μ_k, z_i) is the candidate Ritz pair for the target in question. We call the resulting method the **residual Arnoldi (RA)** method. Thus, given a target (λ, x), the Arnoldi step becomes:

1. Compute the candidate Ritz pair (μ_k, z_k) corresponding to the target (λ, x).
2. $r_k = Az_k - \mu_k z_k$.
3. Orthonormalize r_k against U_k to get u_{k+1}.
4. $U_{k+1} = (u_1 \ \cdots \ u_{k+1})$.

(1.6)

It is easy to see that in the absence of error the residual Arnoldi and the Arnoldi method produce the same bases U_k. However, when we introduce errors, the methods behave quite differently.

The second plot in Figure 1.1 shows the effects of errors on the residual Arnoldi method. The approximations to the target eigenvector converge as before, while the approximations to the second eigenvector stagnate. Although it is hard to see from the graphs, the convergence curves for the target eigenpair are essentially the same for both plots.

The third plot shows what happens when we change targets at the 30th iteration. Now the convergence to the first eigenvector stagnates, while the convergence to the second picks up. This suggests the possibility of computing several eigenpairs by switching the target after each converges. All this can be done with a residuals computed to a low relative error.

The shift-and-invert Arnoldi method (1.5) also stagnates when the system in line 1 is solved inaccurately. The cure here is to replace the right-hand side u_k with r_k. We will call this procedure the SHIFT-AND-INVERT RESIDUAL ARNOLDI (SIRA) method.

1. Using the matrix A, compute the candidate Ritz pair (μ_k, z_k) corresponding to the target (λ, x).
2. $r_k = Az_k - \mu_k z_k$.
3. Solve the system $(A - \sigma I)v_k = r_k$.
4. Orthonormalize v_k against U_k to get u_{k+1}.
5. $U_{k+1} = (u_1 \ \cdots \ u_{k+1})$.

(1.7)

The fourth plot in Figure 1.1 shows what happens when we use the SIRA method on our original matrix with $\sigma = 1.3$ and $\epsilon = 10^{-3}$. Convergence — somewhat accelerated — is achieved for the first eigenpair, the candidates for the second eigenvector stagnate.

subspaces.

The purpose of this paper is to give a mathematical analysis of the RA and SIRA methods (for the practical details of their implementation see [5]). In particular, we will be concerned to explain the Krylov-like convergence of the candidate pairs to the target pair. It would also be useful to show why the method continues to work when the target is changed (see Plot 3 in Figure 1.1); but that is beyond the scope of this paper.

The RA and SIRA methods have their origins in the Jacobi–Davidson method [9]. In [10, §6.2] one of us (Stewart) gave an analysis of that method which related it to Newton's method. It implied that the method should converge linearly when the residuals were computed with constant relative error. Later, in a visit to Utrecht in the spring of 2001, he noted that the convergence was Krylov-like and further noted that it remained so when the projections associated with the Jacobi-Davidson method were removed to give what we call here the SIRA method. During some discussions with H. A. van der Vorst and G. L. G. Sleijpen, the latter suggested that we look at the behavior of what was essentially the RA method. The analysis given here was begun about two years later.

The methods treated here should not be confused with inexact Krylov methods for solving linear systems. These methods require that the products used to form the Krylov sequence be initially calculated to full accuracy but allow increasingly reduced accuracy as the method converges. For more see the excellent survey by Simoncini and Szyld [8].

In Section 2, we give some basic results from perturbation theory and state the assumptions under which we will analyze the methods. In Section 3 we give the analysis of the RA method and in Section 4 we give the analysis of the SIRA method.

Throughout this paper $\|\cdot\|$ will denote the vector 2-norm and its subordinate matrix spectral norm. We assume that

> *All eigenvectors and Ritz vectors, as well as the matrix A, are normalized to have norm one.*

2. Preliminaries

In this section we establish the base for our analysis. To begin with we suppose that:

> *The target eigenpair (λ, x) is simple.*

We will need some results from the perturbation theory of eigenpairs. Let $(x\ X)$ be a unitary matrix whose first column is x. Then

$$\begin{pmatrix} x^* \\ X^* \end{pmatrix} A(x\ X) = \begin{pmatrix} \lambda & h^* \\ 0 & L \end{pmatrix}, \tag{2.1}$$

where

$$h^* = x^*AX \quad \text{and} \quad L = X^*AX$$

We will call the decomposition (2.1) a SCHUR REDUCTION OF A FOR λ.[3]

Lemma 2.1. *The matrix $\lambda I - L$ is nonsingular and hence*

$$\operatorname{sep}(\lambda, L) \stackrel{\text{def}}{=} \|(\lambda I - L)^{-1}\|^{-1} > 0. \tag{2.2}$$

If $\tilde{A} = A + E$, then for all sufficiently small E, there are constants C_λ, C_L and C_x, depending only on $\|A\|$ and $\operatorname{sep}(\lambda, L)$, and a Schur reduction

$$\begin{pmatrix} \tilde{x}^* \\ \tilde{X}^* \end{pmatrix} \tilde{A}(\tilde{x}\ \tilde{X}) = \begin{pmatrix} \tilde{\lambda} & \tilde{h}^* \\ 0 & \tilde{L} \end{pmatrix},$$

such that

$$|\tilde{\lambda} - \lambda| \le C_\lambda \|E\|, \quad \|\tilde{L} - L\| \le C_L \|E\|, \quad \text{and} \quad \|\tilde{x} - x\| \le C_x \|E\|.$$

Moreover,

$$\operatorname{sep}(\tilde{\lambda}, \tilde{L}) \ge \operatorname{sep}(\lambda, L) - |\tilde{\lambda} - \lambda| - \|\tilde{L} - L\|.$$

(This last inequality holds for any $\tilde{\lambda}$ and $\tilde{L}$.)

Proofs may be found in [10, §1.3].

A second result concerns the relation between the norm of the error in a normalized approximation z to x and the norm of its residual

$$r = Az - \mu z, \quad \mu = z^*Az. \tag{2.3}$$

It is well known that the above value of μ minimizes r.

Lemma 2.2. *Let z be a normalized approximation to x and let r be defined by* (2.3). *Then*

$$\|r\| \le 2\|z - x\|. \tag{2.4}$$

Furthermore, for any μ,

$$\frac{1}{\sqrt{2}}\|z - x\| \le |\sin \angle(x, z)| \le \frac{\|r\|}{\operatorname{sep}(\mu, L)}. \tag{2.5}$$

[3]So called because it is the first step in Schur's unitary reduction to triangular form.

Proof. For (2.4) we have

$$\begin{aligned}
\|r\| &= \|Az - \mu z\| \\
&\le \|Az - \lambda z\| && \text{by the optimality of } \mu \\
&= \|A(z-x) - \lambda(z-x)\| && \text{since } Ax = \lambda x \\
&\le 2\|z - x\| && \text{since } \|A\|, |\lambda| \le 1.
\end{aligned}$$

For a proof of (2.5) see [3, 4]. ■

We now turn to the technical assumptions that underlie our analysis. We will state them here for the RA method and modify them in §4 for the SIRA method.

In the course of the RA method, we compute the residual

$$r_k = Az_k - \mu_k z_k$$

with an error. More precisely, we will make the following RELATIVE ERROR ASSUMPTION.

Let

$$\tilde{r}_k = r_k + f_k,$$

be the computed residual. Then there is an ϵ (independent of k) such that

$$\frac{\|f_k\|}{\|r_k\|} \le \epsilon. \tag{2.6}$$

Note that ϵ is a free parameter to the extent that we can control the accuracy of the computations. To stress the dependency of the RA method on ϵ and A, we will write it $\mathrm{RA}_\epsilon(A)$. Note that $\mathrm{RA}_0(A)$ is the exact RA method, which is equivalent to the ordinary Arnoldi method.

The analysis is based on the following fact, established in Lemma 3.1. There is a matrix E_k such that U_k from $\mathrm{RA}_\epsilon(A)$ is the basis for the Krylov subspace $\mathcal{K}_k$ generated by $\mathrm{RA}_0(\tilde{A}_k)$, where

$$\tilde{A}_k = A + E_k.$$

We will further assume that

There is an integer constant C_E such that

$$\|E_k\| \le \epsilon C_\mathrm{E}. \tag{2.7}$$

Thus k steps of $\mathrm{RA}_\epsilon(A)$ produces the same subspace as k steps of $\mathrm{RA}_0(A + E_k)$. Moreover, the size of the perturbation E_k is proportional to ϵ uniformly in k. Although

there is considerable empirical evidence for the truth of this boundedness assumption — at least for reasonably restricted values of k — we have been unable to prove it. This matter is treated further in the appendix to this paper.

In describing the strategy of our analysis, we must be careful with our notation. We will denote by (μ_k, z_k) the Ritz pair generated at the kth stage $\mathrm{RA}_\epsilon(A)$. It comes from the Rayleigh quotient

$$H_k = U_k^* A U_k$$

which we will assume to have a Schur reduction of the form

$$\begin{pmatrix} w_k^* \\ \hat{W}_k^* \end{pmatrix} H_k (w_k \;\; \hat{W}_k) = \begin{pmatrix} \mu_k & g_k^* \\ 0 & P_k \end{pmatrix}$$

We assume that there is a constant $\gamma > 0$ such that

$$\mathrm{sep}(\mu_k, P_k),\ \mathrm{sep}(\lambda, L),\ \mathrm{sep}(\mu_k, L) \geq \gamma. \tag{2.8}$$

If E_k is sufficiently small, by Lemma 2.1 we can assume that $\tilde{A}_k$ has an eigenpair $(\tilde{\lambda}_k, \tilde{x}_k)$ and a Schur reduction

$$\begin{pmatrix} \tilde{x}_k^* \\ \tilde{X}_k^* \end{pmatrix} \tilde{A}_k (\tilde{x}_k \;\; \tilde{X}_k) = \begin{pmatrix} \tilde{\lambda}_k & \tilde{h}_k^* \\ 0 & \tilde{L}_k \end{pmatrix}. \tag{2.9}$$

Note that as $E_k \to 0$, the Schur reduction (2.9) approaches (2.1).

The Rayleigh quotient at the kth stage of $\mathrm{RA}_0(\tilde{A}_k)$ is

$$\tilde{H}_k = U_k^* \tilde{A}_k U_k,$$

which we assume has a Schur reduction of the form

$$\begin{pmatrix} \tilde{w}_k^* \\ \tilde{W}_k^* \end{pmatrix} \tilde{H}_k (\tilde{w}_k \;\; \tilde{W}_k) = \begin{pmatrix} \tilde{\mu}_k & \tilde{g}_k^* \\ 0 & \tilde{P}_k \end{pmatrix} \tag{2.10}$$

Since $\tilde{H}_k \to H_k$ as $E_k \to 0$, by Lemma 2.1 and (2.8) we may assume that for E_k sufficiently small, we have

$$\mathrm{sep}(\tilde{\mu}_k, \tilde{P}_k),\ \mathrm{sep}(\lambda, \tilde{L}_k),\ \mathrm{sep}(\tilde{\mu}_k, L) \geq \tilde{\gamma}, \tag{2.11}$$

where, say, $\tilde{\gamma} = 0.9\gamma$.

The reader should note that $\tilde{z}_{k-1}$ is not the $(k-1)$th Ritz vector of $\mathrm{RA}_0(\tilde{A}_k)$; rather it is the $(k-1)$th Ritz vector of $\mathrm{RA}_0(\tilde{A}_{k-1})$. To put it another way, if we think of an array whose jth row is the sequence of Ritz vectors of $\mathrm{RA}_0(\tilde{A}_j)$, then we have selected the diagonal entries in that array and named them $\tilde{z}_1, \tilde{z}_2, \ldots$. Similarly for the other values associated with the $\tilde{A}_k$.

Finally, we need a motor to drive the convergence of the RA method. One possibility is to mimic existing convergence theory. But as we have pointed out, This theory is rather weak. Instead we will assume that $\mathrm{RA}_0(A)$ converges and use that to prove the convergence of $\mathrm{RA}_\epsilon(A)$.

Specifically, denote the Ritz pairs of $\mathrm{RA}_0(A)$ by $(\mu_k^{(0)}, z_k^{(0)})$ and suppose there are constants $\kappa_k > 0$ such that

$$\|z_k^{(0)} - x\| \leq \kappa_k.$$

(Since $\|z_n^{(0)} - x\| = 0$, the κ_k converge to zero in n steps. But for our results to be interesting, the κ_k should behave like the the first plot in Figure 1.1). By the continuity of the RA algorithm,[4] if $\tilde{\kappa}_k = 1.1\kappa_k$, then for E_k sufficiently small, we have

$$\|\tilde{z}_k - \tilde{x}_k\| \leq \tilde{\kappa}_k. \tag{2.12}$$

By (2.7), we can control the sizes of the E_k by reducing ϵ. Hence we can assume

There is a constant C_{eps}, such that if $\epsilon \leq C_{\mathrm{eps}}$ then (2.8) and (2.12) are satisfied.

In what follows it will be tacitly understood that $\epsilon < C_{\mathrm{eps}}$. In addition, we will readjust C_{eps} from time to time to insure that certain conditions obtain.

Our assumptions imply the convergence of the $\tilde{z}_k - \tilde{x}_k$ [conditioned on the convergence of $\mathrm{RA}_0(A)$]. What we want is the convergence of $z_k - x$. To get from here to there we use the inequality

$$\|z_k - x\| \leq \|\tilde{z}_k - \tilde{x}_k\| + \|z_k - \tilde{z}_k\| + \|\tilde{x}_k - x\|. \tag{2.13}$$

The converge analysis of the next two sections amounts to computing suitable bounds on $\|z_k - \tilde{z}_k\|$ and $\|\tilde{x}_k - x\|$.

A final word on the assumptions made in this sections. Although they may appear arbitrary, they are all but one quite natural. The relative error assumption (2.6) is less an assumption than a statement that we intend to compute the residual to a specified accuracy. The assumptions in (2.8) is just what we would expect of a well-behaved Arnoldi iteration, and the bounds in (2.11) follow from them by a continuity argument. The assumption (2.12) says that the shape of the convergence curves is not affected by perturbing the matrix by a small quantity. In fact, you are invoking (2.12) if you believe that single and double precision computations exhibit the same behavior up to the point where the single precision computation stagnates. The curves in the first plot in Figure 1.1 confirm the assumption for $\epsilon = 10^{-3}$.

[4]There are two ways this continuity can fail. First, by the appearance of an invariant subspace, noted in footnote 1, which we have excluded. Second, by a failure of the Rayleigh–Ritz procedure, which is excluded by (2.11).

The only assumption that is not justified by general considerations is the condition (2.7) on the perturbations E_k. It is essential to our analysis, and we have not been able to prove it — although in the appendix we give some informal reasoning to why it is likely to be true.

3. Convergence of the residual Arnoldi method

At each stage of $\mathrm{RA}_\epsilon(A)$ we compute a candidate Ritz pair $(\mu_k, z_k) = (\mu_k, U_k w_k)$. We then compute its residual $Az_k - \mu_k z_k$ with error, so that what we actually compute is

$$\tilde{r}_k = AU_k w_k - \mu_k U_k w_k + f_k,$$

where f_k satisfies the relative error condition (2.6). This contaminated residual is then orthonormalized against U_k to give

$$u_{k+1} = \frac{(I - U_k U_k^*)\tilde{r}_k}{\|(I - U_k U_k^*)\tilde{r}_k\|} \equiv \frac{(I - U_k U_k^*)\tilde{r}_k}{\rho_k}. \tag{3.1}$$

The following lemma delivers the E_k promised in Section 2.

Lemma 3.1. *There is a matrix E_k, defined by* (3.5) *below, such that U_k is a basis for the kth Krylov subspace generated by* $\mathrm{RA}_0(A + E_k)$.

Proof. Write

$$\begin{aligned}(I - U_k U_k^*)\tilde{r}_k &= (I - U_k U_k^*)(r_k + f_k) \\ &= r_k - U_k g_k + f_k^\perp,\end{aligned}$$

where

$$g_k = U_k^* r_k \quad \text{and} \quad f_k^\perp = (I - U_k U_k^*) f_k.$$

Then from (3.1)

$$\rho_k u_{k+1} = r_k - U_k g_k + f_k^\perp.$$

Since $r_k = AU_k w_k - \mu_k U_k w_k$, we have $\rho_k u_{k+1} = AU_k w_k - \mu_k U_k w_k - U_k g_k + f_k^\perp$, or

$$AU_k w_k = U_k(\mu_k w_k + g_k) + \rho_k u_{k+1} - f_k^\perp. \tag{3.2}$$

Now let

$$\hat{g}_j = \begin{pmatrix} g_j \\ \rho_j \\ 0_{k-j-1} \end{pmatrix}, \quad j = 1, \ldots, k-1.$$

and let

$$G_k = (\hat{g}_1 \ \cdots \ \hat{g}_{k-1} \ g_k). \tag{3.3}$$

Also let W_k be the upper triangular matrix formed from the w_i. Then from (3.2)

$$AU_kW_k = U_k(W_kM_k + G_k) + \rho_k u_{k+1}\mathbf{e}_k^* - F_k^{\perp},$$

where

$$M_k = \operatorname{diag}(\mu_1, \ldots, \mu_k) \quad \text{and} \quad F_k^{\perp} = (f_1^{\perp} \ \cdots \ f_k^{\perp}).$$

Since we have explicitly excluded the case where the iteration terminates, the diagonals of W_k are all nonzero. Hence we can write

$$AU_k = U_k(W_kM_k + G_k)W_k^{-1} + \frac{\rho_k}{\omega_k}u_{k+1}\mathbf{e}_k^* - F_k^{\perp}W_k^{-1}, \tag{3.4}$$

where ω_k is the kth diagonal of W_k. If we set

$$E_k = F_k^{\perp}W_k^{-1}U_k^*, \tag{3.5}$$

then

$$(A + E_k)U_k = U_k(W_kM_k + G_k)W_k^{-1} + \frac{\rho_k}{\omega_k}u_{k+1}\mathbf{e}_k^*. \tag{3.6}$$

Now the matrix $(W_kM_k + G_k)W_k^{-1}$ is upper Hessenberg. Hence (3.6) is an Arnoldi relation, which establishes the lemma. ■

There are two comments to make about this lemma. First, there are many matrices E_k that will make $\mathcal{U}_k$ a Krylov subspace of $A + E_k$, and it is easy to calculate the one of minimal norm. However, the matrix defined by (3.5), though not optimal, has special structure, which we will use in establishing the convergence of the RA method.

Second since the residual r_k from a Ritz approximation is orthogonal to $\mathcal{U}_k$, the vectors g_i are zero. Equivalently, the matrix G_k is nonzero only on its subdiagonal. We have included these vectors in the definition of E_k because in the analysis of the SIRA method they are nonzero.

We now turn to bounding the terms in (2.13), starting with $\|z_k - \tilde{z}_k\|$. It turns out this quantity is zero.

Lemma 3.2. *We have*

$$(\mu_k, z_k) = (\tilde{\mu}_k, \tilde{z}_k).$$

Proof. From (3.6) is follows that the Rayleigh quotient for $\tilde{A}_k$ is

$$\tilde{H}_k = (W_kM_k + G_k)W_k^{-1}.$$

On the other hand from (3.4) it follows that that the Rayleigh quotient for A is

$$H_k = (W_kM_k + G_k)W_k^{-1} - U_k^*F_k^{\perp}W_k^{-1}.$$

Hence

$$\begin{aligned}
\mu_k w_k &= H_k w_k \\
&= (W_k M_k + G_k) W_k^{-1} w_k - U_k^* F_k^\perp W_k^{-1} w_k \\
&= (W_k M_k + G_k) W_k^{-1} w_k - U_k^* F_k^\perp \mathbf{e}_k \\
&= (W_k M_k + G_k) W_k^{-1} w_k \qquad \text{since } U_k^* f_k^\perp = 0 \\
&= \tilde{H}_k w_k.
\end{aligned}$$

Since μ_k is a simple eigenvalue [see (2.8)], this implies $w_k = \tilde{w}_k$ up to a scaling factor. The result now follows from the fact that $z_k = U_k w_k = U_k \tilde{w}_k = \tilde{z}_k$. ■

The proof of this lemma suggests why the Ritz vectors other than the candidate stagnate. Specifically, the proof depends on the fact that the primitive Ritz vector w_k satisfies $W_k^{-1} w_k = \mathbf{e}_k$, a relation which is not satisfied by the other primitive Ritz vectors.

If it were the case that $E_k \to 0$, then we would be finished. For then we would have $\tilde{x}_k \to x$; hence by (2.12) and Lemma 3.2, $z_k = \tilde{z}_k \to x$. Unfortunately, the E_k do not diminish. For from (3.5), we have $\|E_k\| = \|F_k^\perp W_k^{-1}\|$. Since W_k^{-1} is upper triangular and its nonzero elements remain unchanged as k increases, it follows that the columns of $F_k^\perp$ remain unchanged as k increases. Thus, the norms of all the $\|F_k^\perp\|$ are bounded below by the norm of the first column.

We now turn to bounding $\|\tilde{x}_k - x\|$ in (2.13). We do it in two steps.

Lemma 3.3.

$$\|\tilde{x}_k - x\| \le \sqrt{2}\,\tilde{\gamma}^{-1} \|E_k x\|.$$

Proof. We have

$$\begin{aligned}
\tfrac{1}{\sqrt{2}} \|\tilde{x}_k - x\| &\le |\sin \angle(x, \tilde{x}_k)| && \text{by (2.5)} \\
&\le \frac{\|\tilde{A}_k x - \lambda x\|}{\operatorname{sep}(\lambda, \tilde{L}_k)} && \text{by (2.5)} \\
&= \frac{\|(A + E_k) x - \lambda x\|}{\operatorname{sep}(\lambda, \tilde{L}_k)} \\
&= \frac{\|E_k x\|}{\operatorname{sep}(\lambda, \tilde{L}_k)} \\
&\le \frac{\|E_k x\|}{\tilde{\gamma}} && \text{by (2.11).} \ \blacksquare
\end{aligned}$$

The next lemma bounds $\|E_k x\|$.

Lemma 3.4.

$$\|E_k x\| \leq \epsilon \|r_k\| (1 + \tilde{\gamma}^{-1} C_{\mathrm{E}}).$$

Proof: Let $x = \alpha_k z_k + q_k$, where $\alpha_k = z_k^* x$ is the cosine of the $\angle(z_k, x)$. Hence, $\|q_k\|$ is the absolute value of $\sin \angle(z_k, x)$. From (2.5) and(2.11)

$$\|q_k\| \leq \tilde{\gamma}^{-1} \|r_k\|. \tag{3.7}$$

Now

$$\begin{aligned} E_k x &= \alpha_k E_k z_k + E_k q_k \\ &= \alpha_k F_k^{\perp} W_k^{-1} U_k^* z_k + E_k q_k \end{aligned}$$

Since $z_k = U_k w_k$, the first term becomes $\alpha_k F_k^{\perp} \mathbf{e}_k = \alpha_k f_k^{\perp}$. Hence

$$\begin{aligned} \|E_k x\| &\leq |\alpha_k| \|f_k^{\perp}\| + \|E_k\| \|q_k\| \\ &\leq \epsilon \|r_k\| + \tilde{\gamma}^{-1} \|E_k\| \|r_k\| \quad \text{by (3.7)} \\ &\leq \epsilon \|r_k\| + \epsilon \tilde{\gamma}^{-1} C_{\mathrm{E}} \|r_k\|. \quad \text{by (2.7).} \ \blacksquare \end{aligned}$$

Combining the results of these two lemmas we get

$$\|x - \tilde{x}_k\| \leq \epsilon C_{\mathrm{x}} \|r_k\|, \tag{3.8}$$

where

$$C_{\mathrm{x}} = \sqrt{2} \tilde{\gamma}^{-1} (1 + \tilde{\gamma}^{-1} C_{\mathrm{E}}).$$

Theorem 3.5. *Let* $\epsilon < C_{\mathrm{eps}}$. *If*

$$\tau_\epsilon \stackrel{\mathrm{def}}{=} 2\epsilon C_{\mathrm{x}} < 1, \tag{3.9}$$

then

$$\|z_k - x\| \leq \frac{\tilde{\kappa}_k}{1 - \tau_\epsilon}. \tag{3.10}$$

Proof. From Lemma 3.2 and (2.12), we have

$$\|z_k - \tilde{x}_k\| = \|\tilde{z}_k - \tilde{x}_k\| \leq \tilde{\kappa}_k.$$

Hence from (3.8)

$$\begin{aligned} \|z_k - x\| &\leq \|z_k - \tilde{x}_k\| + \|\tilde{x}_k - x\| \\ &\leq \tilde{\kappa}_k + \tfrac{1}{2} \tau_\epsilon \|r_k\|. \\ &\leq \tilde{\kappa}_k + \tau_\epsilon \|z_k - x\| \qquad \text{by (2.4).} \end{aligned}$$

If (3.9) is satisfied, we may solve this inequality to get for $\|z_k - x\|$ to get (3.10). ■

If $\epsilon < 1/(2C_{\mathrm{x}})$, the condition (3.9) will be satisfied. Moreover, when ϵ is sufficiently small, the bound (3.10) is effectively κ_k — that is, $\mathrm{RA}_\epsilon(A)$ has essentially the same properties (with respect to the target eigenpair) as $\mathrm{RA}_0(A)$, which agrees with the numerical experiments in Section 1. Of course, the constant C_{x} is not computable, and even if it were it would probably be too large. Thus a workable value of ϵ must be determined by trial.

4. Analysis of the inexact SIRA method

We now turn to the analysis of the SIRA method (1.7). The analysis parallels the analysis of the RA method — with some differences.

1. The error in $\mathrm{SIRA}_\epsilon(A)$ occurs in the computation of $v_k = Sr_k$, not in the computation of r_k.
2. The backward error is in S, and we must bound a corresponding error in A.
3. In addition to the separation hypotheses (2.11), we must postulate additional ones for S.
4. The vectors z_k and $\tilde{z}_k$ are no longer the same, and we must bound their difference.

In what follows we will use the notation and assumptions of §3.

Regarding item 1, we assume the following.

Let

$$v_k = Sr_k = (A - \sigma I)^{-1} r_k.$$

Then

$$\tilde{v}_k = v_k + f_k, \tag{4.1}$$

where

$$\|f_k\| \leq \epsilon \|v_k\|,$$

Since $\|v_k\| \leq \|S\|\|r_k\|$, we also have

$$\|f_k\| \leq \epsilon \|S\|\|r_k\|. \tag{4.2}$$

We now turn to the construction of the backward error matrix $\hat{E}_k$ in S. In order to do so, we impose a condition on the Ritz values μ_k of $\mathrm{SIRA}_\epsilon(A)$.

$$|\mu_k - \sigma| \geq \eta > 0. \tag{4.3}$$

Lemma 4.1. *There is a matrix $\hat{E}_k$, defined by (4.4) below, such that U_k spans the Krylov subspace generated by the kth step of* $\mathrm{SIRA}_0(\mathrm{S}+\hat{\mathrm{E}}_\mathrm{k})$.

Proof. From (4.1) we have

$$(I - UU^*)\tilde{v}_k = (I - U_kU_k^*)(v_k + f_k) = v_k - U_kg_k + f_k^\perp,$$

where $g_k = U_k^*v_k$ and $f_k^\perp = (I - U_kU_k^*)f_k$. If we write $\rho_k u_{k+1} = (I - UU^*)\tilde{v}_k$, we get

$$v_k = U_kg_k + \rho_k u_{k+1} - f_k^\perp.$$

Now

$$SA = (A - \sigma I)^{-1}A = (A - \sigma)^{-1}[(A - \sigma I) + \sigma I] = I + \sigma S.$$

Hence

$$v_k = Sr_k = S(Az_k - \mu_k z_k) = z_k + (\sigma - \mu_k)Sz_k.$$

It follows that

$$z_k + (\sigma - \mu_k)Sz_k = U_kw_k + (\sigma - \mu_k)SU_kw_k = U_kg_k + \rho_k u_{k+1} - f_k^\perp,$$

or equivalently

$$(\sigma - \mu_k)SU_kw_k = U_k(g_k - w_k) + \rho_k u_{k+1} - f_k^\perp.$$

Using the definition of G_k and W_k above [see (3.3)], we get

$$SU_kW_k(\sigma I - M_k) = U(G_k - W_k) + \rho_k u_{k+1}\mathbf{e}_k^* - F_k^\perp,$$

where

$$M_k = \mathrm{diag}(\mu_1, \ldots, \mu_k) \quad \text{and} \quad F_k^\perp = (f_1^\perp \ \cdots \ f_k^\perp).$$

Postmultiplying by $(\sigma I - M_k)^{-1}W_k^{-1}$, whose existence is ensured by (4.3), we get

$$SU_k = U(G_k - W_k)(\sigma I - M_k)^{-1}W_k^{-1} + \frac{\rho_k}{(\sigma - \mu_k)\omega_k}u_k\mathbf{e}_k^* - F_k^\perp(\sigma I - M_k)^{-1}W_k^{-1}$$

Hence if we define

$$\hat{E}_k = F_k^\perp(\sigma I - M_k)^{-1}W_k^{-1}U_k^*, \tag{4.4}$$

we have

$$(S + \hat{E}_k)U_k = U_k(G_k - W_k)(\sigma I - M_k)^{-1}W_k^{-1} + \frac{\rho_k}{(\sigma - \mu_k)\omega_k}u_{k+1}\mathbf{e}_k^*. \tag{4.5}$$

The matrix $(G_k - W_k)(\sigma I - M_k)^{-1}W_k^{-1}$ is upper Hessenberg. Hence (4.5) is a Krylov relation for

$$\tilde{S}_k = S_k + \hat{E}_k,$$

which establishes the lemma. ■

As with the case of the RA analysis, we will make the following assumption.

There is a constant $C_{\hat{\mathrm{E}}}$ such that

$$\|\hat{E}_k\| \leq \epsilon C_{\hat{\mathrm{E}}}.$$

This means that we can adjust C_{eps} so that

$$\epsilon < C_{\mathrm{eps}} \implies \tilde{S}_k \text{ is nonsingular.} \tag{4.6}$$

Since the Ritz values in the SIRA method are calculated from A rather than S, we need to derive a bound for the perturbation E_k in A corresponding to $\hat{E}_k$. Specifically, let

$$\tilde{A}_k = \tilde{S}_k^{-1} + \sigma I.$$

Equivalently

$$\tilde{S}_k = S + \hat{E}_k = (\tilde{A}_k - \sigma I)^{-1}.$$

Premultplying by $\tilde{S}_k^{-1}$ and postmultiplying by S^{-1}, we get

$$A - \sigma I = (\tilde{A}_k - \sigma I) + (\tilde{A}_k - \sigma I)\hat{E}_k(A - \sigma I),$$

or

$$\tilde{A}_k = A - (\tilde{A}_k - \sigma I)\hat{E}_k(A - \sigma I). \tag{4.7}$$

Hence

$$E_k = \tilde{A}_k - A = -(\tilde{A}_k - \sigma I)\hat{E}_k(A - \sigma I). \tag{4.8}$$

Now we cannot use (4.8) directly to bound $\|E_k\|$, since $\tilde{A}_k$ depends on E_k. However, if we write

$$E_k = -E_k\hat{E}_k(A - \sigma I) - (A - \sigma I)\hat{E}_k(A - \sigma I),$$

then

$$\|E_k\|(1 - \|\hat{E}_k\|\|A - \sigma I\|) \leq \|A - \sigma I\|^2\|\hat{E}_k\|.$$

If we adjust C_{eps} so that $\|\hat{E}_k\|\|A - \sigma I\|$ is less than $1/2$ whenever $\epsilon < C_{\mathrm{eps}}$, then with

$$C_{\mathrm{E}} = 2C_{\hat{\mathrm{E}}}\|A - \sigma I\|^2$$

we have

$$\|E_k\| \leq \epsilon C_{\mathrm{E}}.$$

It is worth noting that, since in practice σ will approximate an eigenvalue of A, the quantity $\|A - \sigma I\|$ is unlikely be much greater than $2\|A\|$. Consequently $C_{\hat{\mathrm{E}}}$ and C_{E} are approximately of the same order of magnitude.

At this point we must introduce two more separation assumptions. Let $\theta = (\lambda - \sigma)^{-1}$ be the target eigenvalue of S. By Lemma 2.1, if $\hat{E}_k$ is sufficiently small, $\tilde{S}_k = S + \hat{E}_k$ has a Schur reduction of the form

$$\begin{pmatrix} \tilde{x}^* \\ \tilde{X}_\perp^* \end{pmatrix} \tilde{S}_k (\tilde{x} \;\; \tilde{X}_\perp) = \begin{pmatrix} \tilde{\theta}_k & \tilde{h}_k^* \\ 0 & \tilde{N}_k \end{pmatrix}.$$

Then we assume that we adjust C_{eps} so that

$$\epsilon < C_{\text{eps}} \implies \text{sep}(\theta, \tilde{N}_k) \geq \tilde{\gamma}. \tag{4.9}$$

The second assumption concerns the Rayleigh quotient $\tilde{H}_k$, whose Schur reduction is given by (2.10).

Let

$$\check{\mu}_k = w_k^* \tilde{H}_k w_k. \tag{4.10}$$

Then we assume that

$$\text{sep}(\check{\mu}_k, \tilde{N}_k) \geq \tilde{\gamma}. \tag{4.11}$$

As in the RA analysis, we have two sets of candidate primitive Ritz vectors: w_k from $H_k = U_k^* A U_k$ and $\tilde{w}_k$ from $\tilde{H}_k = U_k^* \tilde{A}_k U_k$. Unlike the RA case, however, they are not the same. But we can bound the difference.

Lemma 4.2.

$$\|\tilde{w}_k - w_k\| \leq \epsilon 2\sqrt{2}\, \frac{\|A - \sigma I\| \|S_k\| \|r_k\|}{\tilde{\gamma}}.$$

Proof. From (4.7) we have

$$\tilde{H}_k = H_k - U_k^* (\tilde{A}_k - \sigma I) \hat{E}_k (A - \sigma I) U_k.$$

We will regard w_k as an approximation to $\tilde{w}_k$. Let $p_k = \tilde{H}_k w_k - \check{\mu}_k w_k$. Then by (2.5) and (4.11) we have

$$\|\tilde{w}_k - w_k\| \leq \sqrt{2}\, \frac{\|p_k\|}{\tilde{\gamma}} \tag{4.12}$$

Now

$$\begin{aligned} p_k &= \tilde{H}_k w_k - \mu_k w_k \\ &= H_k w_k - U_k^* (\tilde{A}_k - \sigma I) \hat{E}_k (A - \sigma I) U_k w_k - \mu_k w_k \\ &= U_k^* (\tilde{A}_k - \sigma I) \hat{E}_k (A - \sigma I) U_k w_k. \end{aligned}$$

Since $\hat{E}_k = F_k^{\perp}(\sigma I - M_k)^{-1} W_k^{-1} U_k^*$, we have

$$\begin{aligned}
\hat{E}_k(A - \sigma I)U_k w_k &= F_k^{\perp}(\sigma I - M_k)^{-1} W_k^{-1} U_k^*(A - \sigma I)U_k w_k \\
&= F_k^{\perp}(\sigma I - M_k)^{-1} W_k^{-1}(H_k - \sigma I)w_k \\
&= F_k^{\perp}(\sigma I - M_k)^{-1} W_k^{-1} w_k(\mu_k - \sigma) \\
&= F_k^{\perp}(\sigma I - M_k)^{-1} W_k^{-1} w_k(\mu_k - \sigma) \\
&= -F_k^{\perp}\mathbf{e}_k = -f_k^{\perp}
\end{aligned}$$

Hence from (4.12)

$$\begin{aligned}
\|w_k - \tilde{w}_k\| &\le \sqrt{2}\frac{\|p_k\|}{\tilde{\gamma}_k} \\
&\le \sqrt{2}\frac{\|U_k^*(\tilde{A}_k - \sigma I)f_k^{\perp}\|}{\tilde{\gamma}_k} \\
&\le \sqrt{2}\frac{\|\tilde{A}_k - \sigma I\|\|f_k^{\perp}\|}{\tilde{\gamma}} \\
&\le \epsilon\sqrt{2}\frac{\|\tilde{A}_k - \sigma I\|\|S\|\|r_k\|}{\tilde{\gamma}} \quad \text{by (4.2).}
\end{aligned}$$

Because $\tilde{A}_k = A + E_k$, we can adjust C_{eps} so that $\|\tilde{A}_k - \sigma I\| \le 2\|A - \sigma I\|$. Hence

$$\|w_k - \tilde{w}_k\| \le \epsilon 2\sqrt{2}\frac{\|A - \sigma I\|\|S\|\|r_k\|}{\tilde{\gamma}}. \blacksquare$$

If we set

$$C_{\text{z}} = 2\sqrt{2}\,\frac{\|A - \sigma I\|\|S\|}{\tilde{\gamma}}.$$

then we have

$$\|\tilde{z}_k - z_k\| \le \epsilon\|r_k\|C_{\text{z}}, \tag{4.13}$$

The next two lemmas parallel Lemmas 3.3 and 3.4 in the analysis of the RA method. As usual, let x be the target eigenvector, and let $\tilde{x}_k$ be the corresponding eigenvector of $\tilde{A}_k$. Note that x is an eigenvector of S and $\tilde{x}_k$ is an eigenvector of $\tilde{S}_k$.

Lemma 4.3.

$$\|\tilde{x}_k - x\| \le \sqrt{2}\,\tilde{\gamma}^{-1}\|\hat{E}_k x\|. \tag{4.14}$$

Proof. We have

$$\begin{aligned}
\frac{1}{\sqrt{2}}\|\tilde{x}_k - x\| &\leq |\sin\angle(x,\tilde{x}_k)| && \text{by (2.5)}\\
&\leq \frac{\|\tilde{S}_k x - \theta x\|}{\mathrm{sep}(\theta,\tilde{N}_k)} && \text{by (2.5)}\\
&\leq \frac{\|(S+\hat{E}_k)x - \theta x\|}{\tilde{\gamma}} && \text{by (4.9)}\\
&= \frac{\|\hat{E}_k x\|}{\tilde{\gamma}}. \blacksquare
\end{aligned}$$

The second lemma bounds $\|\hat{E}_k x\|$.

Lemma 4.4.

$$\|\hat{E}_k x\| \leq \epsilon\|r_k\|\left(\frac{\|S\|}{\eta} + \frac{C_{\hat{\mathrm{E}}}}{\tilde{\gamma}}\right), \tag{4.15}$$

where η is defined by (4.3)

Proof: Let $x = \alpha_k z_k + q_k$, where $\alpha_k = z_k^* x$ is the cosine of the $\angle(z_k, x)$. Hence, $\|q_k\|$ is the absolute value of $\sin\angle(z_k, x)$. From (2.5),

$$\|q_k\| \leq \frac{\|r_k\|}{\mathrm{sep}(\mu_k, L)} \leq \tilde{\gamma}^{-1}\|r_k\|.$$

Now

$$\begin{aligned}
\hat{E}_k x &= \alpha_k \hat{E}_k z_k + \hat{E}_k q_k\\
&= \alpha_k F_k^{\perp}(\sigma I - M_k)^{-1} W_k^{-1} U_k^* z_k + \hat{E}_k q_k\\
&= \frac{\alpha_k}{\sigma - \mu_k} f_k^{\perp} + \hat{E}_k q_k.
\end{aligned}$$

Hence

$$\begin{aligned}
\|\hat{E}_k x\| &\leq \frac{|\alpha_k|}{|\sigma - \mu_k|}\|f_k^{\perp}\| + \|\hat{E}_k\|\|q_k\|\\
&\leq \frac{\epsilon\|S\|\|r_k\|}{\eta} + \frac{\|\hat{E}_k\|\|r_k\|}{\tilde{\gamma}}\\
&\leq \frac{\epsilon\|S\|\|r_k\|}{\eta} + \frac{\epsilon C_{\hat{\mathrm{E}}}\|r_k\|}{\tilde{\gamma}}. \blacksquare
\end{aligned}$$

In follows from (4.14) and (4.15) that if we define

$$C_{\mathrm{x}} = \frac{\sqrt{2}}{\tilde{\gamma}}\left(\frac{\|S\|}{\eta} + \frac{C_{\hat{\mathrm{E}}}}{\tilde{\gamma}}\right),$$

then

$$\|\tilde{x}_k - x\| \leq \epsilon \|r_k\| C_{\mathrm{x}}. \tag{4.16}$$

We are now in a position to prove the convergence of $\mathrm{SIRA}_\epsilon(\mathrm{A})$.

Theorem 4.5. *Let*

$$\tau_\epsilon = 2\epsilon(C_{\mathrm{z}} + C_{\mathrm{x}}).$$

If $\tau_\epsilon < 1$, *the*

$$\|z_k - x\| \leq \frac{\tilde{\kappa}_k}{1 - \tau_\epsilon}. \tag{4.17}$$

Proof. We have

$$\begin{aligned}
\|z_k - x\| &\leq \|\tilde{z}_k - \tilde{x}_k\| + \|z_k - \tilde{z}_k\| + \|\tilde{x}_k - x\| \\
&\leq \tilde{\kappa}_k + \|z_k - \tilde{z}_k\| + \|\tilde{x}_k - x\| && \text{by (2.12)} \\
&\leq \tilde{\kappa}_k + \epsilon \|r_k\| (C_{\mathrm{z}} + C_{\mathrm{x}}) && \text{by (4.13) and (4.14)} \\
&= \tilde{\kappa}_k + \tfrac{1}{2}\tau_\epsilon \|r_k\| \\
&\leq \tilde{\kappa}_k + \tau_\epsilon \|z_k - x\| && \text{by (2.4).}
\end{aligned}$$

If $\tau_\epsilon < 1$ we can solve this inequality for $\|z_k - x\|$ to give (4.17).

5. Comments

We have established the convergence of the RA and SIRA algorithms under conditions that are likely to hold when the underlying problem is well behaved. Unfortunately, we have not been able to prove the boundedness of the backward errors E_k upon which the convergence proof is based. The following appendix gives some insight into this problem.

The practical use of the algorithm requires that we be able to switch targets once satisfactory convergence has been achieved for a given target. This situation is illustrated by the third plot in Figure 1.1. Unfortunately, this is a more difficult problem than proving convergence to a single target, and we have not treated it here.

The proofs are relative, in the sense that they assume that the the error free algorithm behaves well for the problem at hand. This means that we do not have to intermingle general convergence proofs for Krylov sequences with the particular convergence proofs for the RA and the SIRA methods. We feel that this approach may be useful in other contexts.

6. Appendix: The matrix E_k

Throughout this paper we have made the assumption that the matrix E_k is proportional to ϵ. Although we have not been able to prove the empirically observed fact, in this appendix we will analyze a specific example that will show why it might be expected to hold in general.

From (3.5), $E_k = F_k^{\perp} W_k^{-1} U_k^*$. Since U_k is orthonormal,

$$\|E_k\| = \|F_k^{\perp} W_k^{-1}\|.$$

The natural approach to bounding $\|E_k\|$ is to apply the submultiplicative norm inequality to get the bound $\|E_k\| \leq \|F_k^{\perp}\| \|W_k^{-1}\|$ and then determine bounds for $\|F_k^{\perp}\|$ and $\|W_k^{-1}\|$. Unfortunately this approach will not work because $\|W_k^{-1}\|$ grows with increasing k. Thus we must investigate the interaction between the matrices $F_k^{\perp}$ and W_k^{-1}. To do this we will make some simplifying assumptions.

We will consider the kth stage of the RA method, and suppose that n is much larger than k, so that for practical purposes the vector x can be considered infinite dimensional. We will work in the U_n coordinate system and assume that $U_n^* x$ has components that decrease geometrically. Thus we assume there is a $\beta \in (0,1)$ such that (in the U_n-coordinate system)

$$x = \gamma \begin{pmatrix} 1 \\ \beta \\ \beta^2 \\ \vdots \end{pmatrix}, \qquad \gamma = \sqrt{1-\beta^2}.$$

We will also assume that the primitive Ritz vector has the corresponding components of x; i.e.,

$$w_k = \gamma_k \begin{pmatrix} 1 \\ \beta \\ \vdots \\ \beta^{k-1} \end{pmatrix}, \qquad \gamma_k = \sqrt{(1-\beta^2)/(1-\beta^{2k})}.$$

This amounts to saying that the Rayleigh–Ritz procedure returns the best possible approximation from the column space of U_k.

It follows that the matrix W_k can be written

$$W_k = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ & \beta & \cdots & \beta \\ & & \ddots & \vdots \\ & & & \beta^{k-1} \end{pmatrix} \begin{pmatrix} \gamma_1 & & & \\ & \gamma_2 & & \\ & & \ddots & \\ & & & \gamma_k \end{pmatrix}.$$

The inverse of W_k is

$$W_k^{-1} = \begin{pmatrix} \gamma_1^{-1} & & & \\ & \gamma_2^{-1} & & \\ & & \ddots & \\ & & & \gamma_k^{-1} \end{pmatrix} \begin{pmatrix} 1 & -\beta^{-1} & & & \\ & \beta^{-1} & -\beta^{-2} & & \\ & & \ddots & \ddots & \\ & & & \beta^{k-2} & -\beta^{k-1} \\ & & & & \beta^{k-1} \end{pmatrix}.$$

The first column of $P_k = F_k^{\perp} W_k^{-1}$ is $p_1^{(k)} = f_1^{\perp}$. The ith column is

$$p_i^{(k)} = \frac{1}{\beta^{i-1}} \left(\frac{f_i^{\perp}}{\gamma_i} - \frac{f_{i-1}^{\perp}}{\gamma_{i-1}} \right).$$

Since $\gamma_i \geq \gamma$, we have

$$\|p_i^{(k)}\| \leq \frac{1}{\gamma \beta^{i-1}} (\|f_i^{\perp}\| + \|f_{i-1}^{\perp}\|). \tag{6.1}$$

Thus our next step is to bound $\|f_i^{\perp}\|$. By the relative error condition (2.6) and by (2.4), we have

$$\|f_i^{\perp}\| \leq \|f_i\| \leq \epsilon \|r_i\| \leq 2\epsilon \|x - z_i\|.$$

Now in the U_n-coordinate system

$$\|x - z_i\|^2 = (\gamma_i - \gamma)^2 (1 + \beta^2 + \cdots \beta^{2(i-1)}) + \gamma^2 (\beta^{2i} + \beta^{2(i+1)} + \cdots).$$

The first term in this expression can be written

$$\begin{aligned} \frac{(\gamma_i - \gamma)^2}{\gamma_i^2} &= \left(1 - \frac{\gamma}{\gamma_i}\right)^2 \\ &= \left(1 - \sqrt{1 - \beta^{2i}}\right)^2 \\ &\leq (\beta^{2i})^2 \\ &\leq \beta^{2i}. \end{aligned}$$

The second term is simply β^{2i}. Hence

$$\|f_i^{\perp}\| \leq 2\sqrt{2}\epsilon \beta^i.$$

It now follows from (6.1) that

$$\|p_i^{(k)}\| \leq 2\sqrt{2}\epsilon (\beta^i + \beta^{i-1}) \frac{\beta^{-(i-1)}}{\gamma} = \frac{2\sqrt{2}}{\gamma}(1 + \beta)\epsilon.$$

And finally

$$\|E_k\| \leq \frac{2\sqrt{2k}}{\gamma}(1+\beta)\epsilon. \tag{6.2}$$

Under the assumptions of this appendix, the columns of W_k^{-1} increase in size while the columns of $F_k^{\perp}$ decrease in size. The decrease is exactly enough to compensate for the increase in the columns of W_k^{-1}. Here the bidiagonality of W_k^{-1} is critical. It prevents large values of the kth of W_k^{-1} column from combining with the large columns at the beginning of $F_k^{\perp}$. In our numerical experiments we do not observe this strict locality. However, the larger elements of any column of W_k^{-1} are near the diagonal, and the elements decrease as one moves up the column enough to compensate for the increasing norms of the $f_i^{\perp}$ as one moves to the beginning of $F_k^{\perp}$.

If (6.2) is to be believed, the bound assumed in (2.7) must take the form

$$\|E_k\| \leq \frac{2\sqrt{2n}}{\gamma}(1+\beta)\epsilon, \tag{6.3}$$

in which the factor $2\sqrt{2n}$ can be quite large. Fortunately, our proofs are predicated on the convergence of $RA_0(A)$, which restricts the values of k to be less than the value k_{final} at which we declare convergence. Thus we can assume a bound of the form

$$\|E_k\| \leq \frac{2\sqrt{2k_{\text{final}}}}{\gamma}(1+\beta)\epsilon,$$

which for large n may be considerably smaller than (6.3).

Acknowledgements

One of the authors (Stewart) would like the Mathematical and Computational Sciences Division of the National Institute of Standards and Technology for the use of their facilities during the development of this project.

References

[1] W. E. Arnoldi. The principle of minimized iterations in the solution of the matrix eigenvalue problem. *Quarterly of Applied Mathematics*, 9:17–29, 1951.

[2] C. Beattie, M. Embree, and J. Rossi. Convergence of restarted Krylov subspaces to invariant subspaces. Report 01/21, Oxford University Computing Laboratory, Numerical Analysis Group, 2001.

[3] Z. Jia and G. W. Stewart. On the convergence of Ritz values, Ritz vectors, and refined Ritz vectors. Technical Report TR–3986, Department of Computer Science, University of Maryland, College Park, 1999.

[4] Z. Jia and G. W. Stewart. An analysis of the Rayleigh–Ritz method for approximating eigenspaces. *Mathematics of Computation*, 70:637–647, 2000.

[5] Che-Rung Lee. *Residual Arnoldi Methods: Theory, Package, and Experiments.* PhD thesis, Department of Computer Science, University of Maryland at College Park, 2007. Available at `https://drum.umd.edu/dspace/handle/1903/2`.

[6] R. B. Lehoucq, D. C. Sorensen, and C Yang. ARPACK Users' Guide: Solution of Large Scale Eigenvalue Problems by Implicitly Restarted Arnoldi Methods. Available at
`http//www.caam.rice.edu/software/ARPACK/index.html`, 1997.

[7] Y. Saad. On the rates of convergence of the Lanczos and the block Lanczos methods. *SIAM Journal on Numerical Analysis*, 17:687–706, 1980.

[8] V. Simoncini and D. B. Szyld. Recent computational developments in Krylov subspace methods for linear systems. Research Report 05-9-25, Department of Mathematics, Temple University, 2005. Revised May 2006. To Appear in *Numerical Linear Algebra with Applications.*.

[9] G. L. G. Sleijpen and H. A. van der Vorst. A generalized Jacobi–Davidson iteration method for linear eigenvalue problems. Preprint 856, Department of Mathematics, Utrecht University, 1994.

[10] G. W. Stewart. *Matrix Algorithms II: Eigensystems.* SIAM, Philadelphia, 2001.