Convergence Analysis of an Inexact Truncated RQ-Iteration

Chao Yang

CRPC-TR98747-S
April 1998

Center for Research on Parallel Computation
Rice University
6100 South Main Street
CRPC - MS 41
Houston, TX 77005

Submitted August 1998

CONVERGENCE ANALYSIS OF AN INEXACT TRUNCATED RQ-ITERATION [*]

CHAO YANG[†]

Abstract. The Truncated RQ-iteration (TRQ) can be used to calculate interior or clustered eigenvalues of a large sparse and/or structured matrix A. This method requires solving a sequence of linear equations. When these equations can be solved accurately by a direct solver, the convergence of each eigenvalue is quadratic in general and cubic if A is hermitian. An important question is whether the TRQ iteration will still converge if these equations are approximately solved by a preconditioned iterative solver. If it does converge, how fast is the convergence rate? In this paper, we analyze the convergence of an inexact TRQ iteration in which linear systems are solved iteratively with some error. We show that under some appropriate conditions, the convergence rate of the inexact TRQ is at least linear with a small convergence factor.

Key words. Arnoldi method, Lanczos method, eigenvalues, Truncated RQ-iteration

AMS subject classifications. Primary 65F15, Secondary 65G05

1. Introduction.

In this paper, we are concerned with solving

$$Ax = \lambda x,$$

where $A \in \mathrm{C}^{n\times n}$ is large sparse or structured. By that, we mean the matrix A has very few nonzero elements, or it has a special structure that allows $y \leftarrow Ax$ to be implemented efficiently in much less than n^2 floating point operations (FLOPS). (An example of this is the discrete Fourier transform matrix.) We are particularly interested in finding several interior or clustered eigenvalues of A. This problem is often solved by applying the Arnoldi [1] or Lanczos [3] method to $(A - \sigma I)^{-1}$, where σ is a target shift. This technique is usually referred to as *shift and invert*. In a shifted and inverted Arnoldi or Lanczos iteration, one must solve a sequence of linear equations accurately in order to capture all desired eigenvalues. Loss of accuracy in solving shift-invert equations of the form

$$(A - \sigma I)w = v$$

may result in the corruption of the Krylov subspace from which eigenvalue and eigenvector approximations are drawn.

The recently proposed Truncated RQ-iteration (TRQ) [8] provides an alternative for calculating the interior or clustered eigenvalues. The TRQ iteration also solves a sequence of linear systems of the form $(A - \mu I)w = v$. However, numerical examples shown in [8] have indicated that the solution accuracy of these linear equations can be relaxed. Rapid convergence has been observed when these equations are solved approximately by a preconditioned iterative solver. In the following, we will use the term *inexact TRQ* to refer to the TRQ iteration in which the linear equations are solved approximately.

In this paper, we analyze the convergence of the inexact TRQ iteration and show that under some appropriate conditions, the inexact TRQ iteration converges linearly

[*] This work was supported in part by NSF cooperative agreement CCR-9120008, and by ARPA contract number DAAL03-91-C-0047 (administered by the U.S. Army Research Office).

[†] Department of Computational and Applied Mathematics, Rice University, Houston, TX 77005-1892, (chao@caam.rice.edu).

```
Algorithm 1: Implicitly Shifted $RQ$-iteration

Input: $(A, V, H)$ with $AV = VH$, $V^H V = I$, and $H$ is upper
        Hessenberg.
Output: $(V, H)$ such that $AV = VH, V^H V = I$ and $H$ is upper triangular.

1. for $j = 1, 2, 3, \ldots$ until converged,
   1.1. Select a shift $\mu \leftarrow \mu_j$;
   1.2. Factor $H - \mu I = RQ$;
   1.3. $H \leftarrow QHQ^H$ ; $V \leftarrow VQ^H$;
2. end;
```

FIG. 2.1. *Implicitly Shifted RQ-iteration.*

with a small convergence factor. Moreover, the analysis recovers the quadratic (or cubic if A is Hermitian) convergence of the TRQ when the linear systems are solved exactly.

The organization of the paper is as follows. In Section 2, we review the TRQ iteration. The inexact TRQ iteration is introduced in Section 3. The linear convergence of the inexact TRQ is proved in Section 4. Some numerical examples are shown in Section 5 to confirm the convergence analysis.

2. The Truncated RQ-iteration.

Before introducing the TRQ iteration, let us examine the full RQ-iteration. The RQ-iteration is similar to the familiar QR algorithm. It begins with a Hessenberg reduction

$$AV = VH, \tag{2.1}$$

where $V^H V = I$ and H is upper Hessenberg. This reduction is followed by the actions described in Figure 2.1, which eventually drives H into an upper triangular form with eigenvalues exposed on the diagonal. If we let $V_+ = VQ^H$, $H_+ = QHQ^H$, $v_1^+ = V_+ e_1$ and $v_1 = Ve_1$, it is easy to verify that in a single RQ iterate

$$(A - \mu I)v_1^+ = v_1 \rho_{1,1},$$

where $\rho_{1,1} = e_1^T R e_1$. This implies that the first column V_+ is what one would have obtained by applying one step of inverse iteration to v_1 with the shift μ. This property is preserved in all subsequent RQ iterates. Thus, one would expect very rapid convergence of leading columns of V to an invariant subspace of A.

In the large scale setting it is generally impossible to carry out the full iteration involving $n \times n$ orthogonal similarity transformations. It would be desirable to truncate this update procedure after k steps to maintain and update only the leading portion of the factorizations occurring in this sequence. A truncated Hessenberg reduction can be produced by an Arnoldi iteration which yields

$$AV_k = V_k H_k + f_k e_k^T, \quad V_k^H V_k = I_k, \quad \text{and} \quad V_k^H f_k = 0. \tag{2.2}$$

The matrix $V_k \in \mathrm{C}^{n \times k}$ can be viewed as the leading k columns of the $V \in \mathrm{C}^{n \times n}$ that appears in the full Hessenberg reduction (2.1), and H_k can be viewed the $k \times k$ leading principle submatrix of H.

To carry out the last k steps of the RQ update within the truncated Hessenberg reduction, one must find out the consequence of the first $n-k$ steps of the full RQ factorization. In particular, one must determine the $(k+1)$-st column of both $V\tilde{Q}$ and $H\tilde{Q}$, where

$$\tilde{Q} = \begin{pmatrix} I_k & 0 \\ 0 & \hat{Q} \end{pmatrix}$$

is the product of Given's rotations used to drive the lower $(n-k) \times (n-k)$ portion of H to an upper triangular form. The determination of these intermediate vectors leads to solving a set of equations to be defined below.

To be precise, let us partition $V = (V_k, \hat{V})$ where V_k denotes the leading k columns of V (produced by an Arnoldi iteration,) and let

$$H = \begin{pmatrix} H_k & M \\ \beta_k e_1 e_k^T & \hat{H} \end{pmatrix}$$

be partitioned conformally so that

$$A(V_k, \hat{V}) = (V_k, \hat{V}) \begin{pmatrix} H_k & M \\ \beta_k e_1 e_k^T & \hat{H} \end{pmatrix}. \tag{2.3}$$

Let μ be some given shift. In a full RQ-iteration, we would begin factoring $H - \mu I$ from right to left using Givens method, say, to obtain

$$H - \mu I = \begin{pmatrix} H_k - \mu I_k & \hat{M} \\ \beta_k e_1 e_k^T & \hat{R} \end{pmatrix} \begin{pmatrix} I_k & 0 \\ 0 & \hat{Q} \end{pmatrix}$$

where $\hat{H} - \mu I = \hat{R}\hat{Q}$ and $\hat{M} = M\hat{Q}$. Postmultiplying (2.3) by $\begin{pmatrix} I_k & 0 \\ 0 & \hat{Q} \end{pmatrix}$ yields

$$(A - \mu I)(V_k, \hat{V}\hat{Q}^H) = (V_k, \hat{V}) \begin{pmatrix} H_k - \mu I_k & \hat{M} \\ \beta_k e_1 e_k^T & \hat{R} \end{pmatrix}. \tag{2.4}$$

Note that in a truncated RQ-iteration, we do not have $\hat{M}$, $\hat{H}$, or $\hat{V}$. However, at this point, all one needs to know in order to complete the RQ factorization are the first columns of $\hat{V}\hat{Q}$, $\hat{M}$ and $\hat{R}$.

If these vectors can be computed without forming and applying $\hat{Q}$, then a truncated version of the RQ-iteration is possible. To determine these vectors, let us examine the first $k+1$ columns of (2.4). Let $v = \hat{V}e_1$, $v_+ = \hat{V}\hat{Q}^H e_1$, $h = \hat{M}e_1$, and $\alpha = e_1^T \hat{R} e_1$. It follows from (2.4) that

$$(A - \mu I)(V_k, v_+) = (V_k, v) \begin{pmatrix} H_k - \mu I_k & h \\ \beta_k e_k^T & \alpha \end{pmatrix}. \tag{2.5}$$

Note that the vector v is the normalized f_k produced by the Arnoldi iteration. From (2.5), it follows that v_+ must satisfy

$$(A - \mu I)v_+ = V_k h + v\alpha, \tag{2.6}$$

with $V_k^H v_+ = 0$ and $\|v_+\| = 1$ since the columns of (V_k, v_+) must be orthonormal.

These conditions may be expressed succinctly through the *TRQ equation*

$$\begin{pmatrix} A-\mu I & V_k \\ V_k^H & 0 \end{pmatrix} \begin{pmatrix} v_+ \\ -h \end{pmatrix} = \begin{pmatrix} v\alpha \\ 0 \end{pmatrix}, \quad \|v_+\| = 1. \tag{2.7}$$

Note that the unknowns in (2.7) are v_+, h and α. The conditions $V_k^H v_+ = 0$ and $\|v_+\| = 1$ allow one to solve

$$\begin{pmatrix} A-\mu I & V_k \\ V_k^H & 0 \end{pmatrix} \begin{pmatrix} w \\ z \end{pmatrix} = \begin{pmatrix} v\gamma \\ 0 \end{pmatrix}$$

first for any $\gamma \neq 0$. The vector v_+ can be computed by simply normalizing w. Once we have v_+, premultiplying (2.6) with V_k and v yields

$$h = V_k^H(A-\mu I)v_+ \quad \text{and} \quad \alpha = v^H(A-\mu I)v_+.$$

The existence and uniqueness of the solution to the TRQ equation is carefully established in [8]. Using the additional property (2.2), we can further simplify the TRQ equation (2.7) to devise a solution scheme that avoids a block Gaussian elimination. It is shown in [8] that the TRQ equation can be solved as follows:

1. $w \leftarrow (I - V_k V_k^H)(A-\mu I)^{-1}(V_k s)$, for some appropriate s;
2. $v_+ \leftarrow w/\|w\|$;
3. $h \leftarrow V_k^H A v_+$; $\alpha \leftarrow v^H(A-\mu I)v_+$;

Once the TRQ equation is solved, the RQ update can be applied to (2.5) to finish a complete cycle. The TRQ algorithm is shown in Figure 2.2. We refer the interested reader to [8] for many implementation details.

```
Algorithm 2: (TRQ) Truncated RQ-iteration

Input: $(A, V_k, H_k, f_k)$ with $AV_k = V_k H_k + f_k e_k^T$, $V_k^H V_k = I$, $H_k$
       is upper Hessenberg.
Output: $(V_k, H_k)$ such that $AV_k = V_k H_k$, $V_k^H V_k = I$ and $H_k$ is upper
        triangular.

1. Put $\beta_k = \|f_k\|$ and put $v = f_k/\beta_k$;
2. for $j = 1, 2, 3, \ldots$ until converged,
   2.1. Select a shift $\mu \leftarrow \mu_j$;
   2.2. Solve $\begin{pmatrix} A-\mu I & V_k \\ V_k^H & 0 \end{pmatrix} \begin{pmatrix} v_+ \\ -h \end{pmatrix} = \begin{pmatrix} v\alpha \\ 0 \end{pmatrix}$ with $\|v_+\| = 1$;
   2.3. Factor $\begin{pmatrix} H_k-\mu I_k & h \\ \beta_k e_k^T & \alpha \end{pmatrix} = \begin{pmatrix} R_k & r \\ 0 & \rho \end{pmatrix} \begin{pmatrix} Q_k & q \\ \sigma e_k^T & \gamma \end{pmatrix}$;
   2.4. $V_k \leftarrow V_k Q_k^H + v_+ q^H$;
   2.5. $\beta_k \leftarrow \sigma e_k^T R_k e_k$; $v \leftarrow v_k\bar{\sigma} + v_+\bar{\gamma}$;
   2.6. $H_k \leftarrow Q_k R_k + \mu I_k$;
3. end;
```

FIG. 2.2. *The Truncated RQ-iteration.*

3. The Inexact TRQ Iteration.

If the cost of factoring $A-\mu I$ is moderate, the TRQ iteration provides a clean and efficient way of obtaining accurate approximations to interior or clustered eigenvalues. Otherwise, we must resort to other means to solve the TRQ equation (2.7). A preconditioned iterative solver is a natural candidate. In the following, we will present an algorithm based on the idea of incorporating an iterative solver in the TRQ iteration, and analyze the convergence of this method.

We shall ask the question of whether the TRQ iteration will still provide accurate eigenvalue approximations if the accuracy of the solution to the TRQ equation is relaxed. If convergence does occur in this inexact scheme, how fast can it be? To address these questions, let us first examine the consequence of replacing the exact solution, v_+, of (2.7) with some approximation $\tilde{v}_+$.

The vector $\tilde{v}_+$ can be computed by applying an iterative solver to

$$(A-\mu I)w = V_k s, \tag{3.1}$$

for some $s \neq 0$, followed by

$$\tilde{v}_+ \leftarrow (I - V_k V_k^H)w, \quad \tilde{v}_+ \leftarrow \frac{\tilde{v}_+}{||\tilde{v}_+||}. \tag{3.2}$$

The orthogonalization and normalization guarantee that

$$V_k^H \tilde{v}_+ = 0 \ \text{ and } \ ||\tilde{v}_+|| = 1$$

are satisfied. To continue the TRQ iteration, we shall compute h and α such that

$$(A-\mu I)\tilde{v}_+ = V_k h + v\alpha \tag{3.3}$$

holds. However, the following lemma indicates that it is generally difficult to find a perfect match for (3.3).

LEMMA 3.1. *Suppose we solve* (3.1) *by a Krylov subspace method with a zero starting vector and no preconditioner to obtain an approximation w. If $\tilde{v}_+ \equiv (I - V_k V_k^H)w \neq 0$, then*

$$(A-\mu I)\tilde{v}_+ \notin span\{V_k, v\}.$$

Proof. Recall that V_k and v are generated by an Arnoldi process. Thus, if we let v_1 be the first column of V_k, then

$$V_k = \text{span}\{v_1, Av_1, A^2v_1, ..., A^{k-1}v_1\} \ \text{ and } \ v \in \text{span}\{v_1, Av_1, A^2v_1, ..., A^kv_1\}.$$

It follows that

$$V_k s = p(A)v_1, \tag{3.4}$$

for some polynomial $p(\lambda)$ of degree at most $k-1$. Applying a Krylov subspace solver to (3.1) yields an approximate solution

$$w = q(A)V_k s,$$

where $q(\lambda)$ is another polynomial associated with the Krylov linear solver. It follows from (3.4) that $w = q(A)p(A)v_1$. If we put $\psi(\lambda) = q(\lambda)p(\lambda)$, it follows from our assumption that the degree of $\psi(\lambda)$ must be at least k for otherwise $\psi(A)v_1 \in \text{span}\{V_k\}$, and $\tilde{v}_+ = (I - V_k V_k^H)w = 0$. Now, let $z = V_k^H(A - \mu I)w$. Since V_k spans a k dimensional Krylov subspace associated with A and v_1, the vector $V_k z$ can be expressed as

$$V_k z = r(A)v_1,$$

for some polynomial $r(\lambda)$ of degree at most $k-1$. Hence, if we let $\beta = \|(I - V_k V_k^H)w\|$, then

$$\begin{aligned}(A - \mu I)\tilde{v}_+ &= (A - \mu I)(I - V_k V_k^H)w/\beta \\ &= (A - \mu I)\Big[w - V_k z\Big]/\beta \\ &= (A - \mu I)\Big[\psi(A)v_1 - r(A)v_1\Big]/\beta \\ &= \phi(A)v_1,\end{aligned}$$

where $\phi(\lambda) = (\lambda - \mu)\Big[\psi(\lambda) - r(\lambda)\Big]/\beta$, a polynomial of degree of at least $k+1$. Therefore, we conclude that

$$\tilde{v}_+ \notin \text{span}\{v_1, Av_1, ..., A^k v_1\} = \text{span}\{V_k, v\}.$$

□

Consequently, the best one can hope for from (3.3) is a weak solution $(\tilde{h}, \tilde{\alpha})$ derived from

$$V_k^H\Big((A - \mu I)\tilde{v}_+ - V_k\tilde{h} - v\tilde{\alpha}\Big) = 0 \tag{3.5}$$

$$v^H\Big((A - \mu I)\tilde{v}_+ - V_k\tilde{h} - v\tilde{\alpha}\Big) = 0 \tag{3.6}$$

or equivalently,

$$\tilde{h} = V_k^H A\tilde{v}_+, \quad \text{and} \quad \tilde{\alpha} = v^H(A - \mu I)\tilde{v}_+.$$

Due to the error remaining in (3.3), the truncated Hessenberg reduction (2.5) is now inexact. We can express this inexact reduction by

$$(A - \mu I)(V_k, \tilde{v}_+) = (V_k, v)\begin{pmatrix} H_k - \mu I_k & \tilde{h} \\ \beta_k e_k^T & \tilde{\alpha} \end{pmatrix} + z e_{k+1}^T, \tag{3.7}$$

where z is the residual error defined as

$$z \equiv (A - \mu I)\tilde{v}_+ - (V_k, v)\begin{pmatrix} \tilde{h} \\ \tilde{\alpha} \end{pmatrix}.$$

Recall from (3.5) and (3.6) that $V_k^H z = 0$ and $v^H z = 0$. If we now proceed by applying a sequence of Given's rotations from the right to (3.7) to annihilate the sub-diagonal elements of

$$\begin{pmatrix} H_k - \mu I_k & \tilde{h} \\ \beta_k e_k^T & \tilde{\alpha} \end{pmatrix},$$

the residual error will be mixed into all columns of V_k. Consequently, the updated basis vectors are no longer valid Arnoldi vectors. However, as we will show in the next section, the first column v_1^+ of this updated basis satisfies

$$(A - \mu I)v_1^+ = \rho_{11} v_1 + z\delta, \tag{3.8}$$

where δ is a product of sines associated with the aforementioned Given's rotations. This observation reveals that an approximate inverse iteration remains in this inexact TRQ update. The error associated with this inverse iteration is likely to be considerably smaller than $||z||$ due the factor δ. Therefore, a simple remedy for correcting the contaminated Arnoldi basis is to recompute an Arnoldi factorization from the very first column of the updated V_k. An algorithm based on the above discussion is given in Figure 3.1.

```
Algorithm 3: (ITRQ) Inexact TRQ-iteration

Input: $(A, V_k, H_k, f_k)$ with $AV_k = V_k H_k + f_k e_k^T$, $V_k^H V_k = I$,
        $H_k$ is upper Hessenberg.
Output: $(V_k, H_k)$ such that $AV_k = V_k H_k$, $V_k^H V_k = I$ and $H_k$ is
        upper triangular.

1. Put $\beta_k = ||f_k||$ and put $v = f_k/\beta_k$;
2. for $j = 1, 2, 3, \ldots$ until convergence,
   2.1. Select a shift $\mu \leftarrow \mu_j$;
   2.2. Solve $(I - V_k V_k^H)(A - \mu I)(I - V_k V_k^H)w = v$ approximately;
   2.3. $w \leftarrow (I - V_k V_k^H)w$, $v_+ \leftarrow w/||w||$;
   2.4. $h \leftarrow V_k^H A v_+$, $\alpha \leftarrow v^H (A - \mu I) v_+$ ;
   2.5. Factor $\begin{pmatrix} H_k - \mu I_k & h \\ \beta_k e_k^T & \alpha \end{pmatrix} = \begin{pmatrix} R_k & r \\ 0 & \rho \end{pmatrix} \begin{pmatrix} Q_k & q \\ \sigma e_k^T & \gamma \end{pmatrix}$;
   2.6. $v_1 \leftarrow V_k Q_k^H e_1 + v_+ q^H e_1$;
   2.7. $(H_k, V_k, v, \beta_k) \leftarrow$ Arnoldi$(A, v_1)$;
3. end;
```

FIG. 3.1. *Inexact TRQ iteration.*

Although this algorithm is similar to an explicitly restarted Arnoldi iteration with the starting vector generated from solving

$$(A - \mu I)w = v_1$$

iteratively, it offers the additional advantage of error damping which is absent in the latter approach. Therefore, it is likely to be more effective than a simple explicit restart. We will illustrate this point in Section 5.2 with a numerical example.

4. Convergence Analysis.

This section focuses on analyzing the convergence of this inexact TRQ scheme. In particular, we are interested in understanding the tradeoff between the accuracy of the solution to the TRQ equation and the rate of convergence of each eigenpair in TRQ. For a simple case in which Rayleigh quotient shifts are used throughout the TRQ iteration, we establish the local linear convergence of the first Arnoldi basis vector to an eigenvector of A. The convergence factor depends on $||\delta z||$, the magnitude of the damped residual error in (3.7), and the "gap" between two consecutive eigenvalues sought.

To begin the analysis, let us assume that an inexact Hessenberg reduction (3.7) has been obtained, and $k-1$ rotations Q_1^H, Q_2^H,...,Q_{k-1}^H, each of the form

$$Q_i^H = \begin{pmatrix} I_{k-i} & & & 0 \\ & \gamma_i & \sigma_i & \\ & -\sigma_i & \gamma_i & \\ 0 & & & I_{i-1} \end{pmatrix}, \quad \sigma_i^2 + \gamma_i^2 = 1 \tag{4.1}$$

have been applied to (3.7) from the right to annihilate the sub-diagonal elements of

$$\begin{pmatrix} H_k - \mu I_k & \tilde{h} \\ \beta_k e_k^T & \tilde{\alpha} \end{pmatrix}.$$

The first two columns of the new matrix equation satisfy

$$(A - \mu I)(v_1, \tilde{v}_2) = (v_1, v_2) \begin{pmatrix} 0 & \eta \\ \epsilon & \rho \end{pmatrix} + (0, z\hat{\sigma}), \tag{4.2}$$

where $\tilde{v}_2 = (V_k, \tilde{v}_+)Q_1^H Q_2^H \cdots Q_{k-1}^H e_2$, $\hat{\sigma} = \sigma_1\sigma_2\cdots\sigma_{k-1}$ and $\epsilon = ||(A-\mu I)v_1||$. Now, let

$$\tau = \sqrt{\epsilon^2 + \rho^2}, \quad \sigma_k = \frac{\rho}{\tau}, \quad \text{and} \quad \gamma_k = \frac{\epsilon}{\tau}.$$

Applying $\begin{pmatrix} \gamma_k & \sigma_k \\ -\sigma_k & \gamma_k \end{pmatrix}$ to (4.2) from the right yields

$$(A - \mu I)(v_1^+, \tilde{v}_2^+) = (v_1, v_2) \begin{pmatrix} -\sigma_k\eta & \gamma_k\eta \\ 0 & \tau \end{pmatrix} + (-\sigma_k\hat{\sigma}z, \gamma_k\hat{\sigma}z), \tag{4.3}$$

where $v_1^+ = \gamma_k v_1 - \sigma_k\tilde{v}_2$ and $\tilde{v}_2^+ = \sigma_k v_1 + \gamma_k\tilde{v}_2$.

We will analyze the convergence of the inexact TRQ iteration by examining the norm of $r_+ = (A - \mu_+ I)v_1^+$, where $\mu_+ = (v_1^+)^H A v_1^+$. We define the damped residual error ν as

$$\nu = ||\hat{\sigma}z||. \tag{4.4}$$

Note that

$$|\hat{\sigma}| = |\sigma_1\sigma_2\cdots\sigma_{k-1}| < 1.$$

This is because each σ_i is a sine used to construct the Given's rotation in (4.1).

The following theorem asserts that if the inexact TRQ is converging to an isolated eigenvalue of A, r_+ must satisfy $||r_+|| \le \psi(\mu,\nu)||r||$, where $\psi(\mu,\nu)$ is uniformly bounded if μ is sufficiently close to an isolated eigenvalue of A, and if ν is not too large.

THEOREM 4.1. *Let $r = (A - \mu I)v_1$ and $r_+ = (A - \mu_+)v_1^+$, where v_1 and v_1^+ are as defined in* (4.3)*, and μ, μ_+ are Rayleigh quotients of A with respect to v_1 and v_1^+ respectively. If $A - \mu I$ is nonsingular, and μ is convergent to a simple eigenvalue of A. Then*

$$||r_+|| \le \psi(\mu,\nu)||r||, \tag{4.5}$$

where the magnitude of the function ψ depends on μ and the size of the damped error ν defined in (4.4). Let $V \equiv (V_k, \hat{V}_{n-k})$ be unitary, where V_k consists of Arnoldi basis vectors generated by Step 2.7 in Algorithm 3. Repartition V as $V = (v_1, \hat{V}_{n-1})$, and let

$$C = \hat{V}_{n-1}^H A V_{n-1}, \quad \textit{and} \quad \zeta = \|(C - \mu I)^{-1}\|^{-1}.$$

If $\nu < \zeta$, then

$$|\psi(\mu, \nu)| \leq \frac{|\epsilon\eta|}{\zeta^2 - \nu^2} + \frac{|\epsilon|\nu}{\zeta^2 - \nu^2} + \frac{\nu}{\sqrt{\zeta^2 - \nu^2}}, \tag{4.6}$$

where $\epsilon = \|(A - \mu)v_1\|$ and $\eta = v_1^H A v_2$. Furthermore, if $\nu < \zeta/\sqrt{2}$, then

$$|\psi(\mu, \nu)| < 1$$

holds asymptotically.

Proof. For clarity, we drop the subscripts of σ_k and γ_k in the following. Note that

$$\begin{aligned} r_+ &= (A - \mu_+ I)v_1^+ \\ &= (A - \mu I)v_1^+ + (\mu - \mu_+)v_1^+ \\ &= (-\sigma\eta)v_1 + (\mu - \mu_+)v_1^+ + (-z\hat{\sigma})\sigma. \end{aligned} \tag{4.7}$$

The last step of the above derivation used the relation

$$(A - \mu I)v_1^+ = v_1(-\sigma_k\eta) - \hat{\sigma}\sigma_k z,$$

which appeared in the first column of (4.3). The distance between μ_+ and μ may be estimated as follows:

$$\begin{aligned} \mu_+ - \mu &= (v_1^+)^H A v_1^+ - \mu \\ &= (v_1^+)^H (A - \mu I) v_1^+ \\ &= (v_1^+)^H (-\sigma\eta v_1 - z\hat{\sigma}\sigma) \\ &= (\gamma v_1 + \sigma\tilde{v}_2)^H (-\sigma\eta v_1 - z\hat{\sigma}\sigma) \\ &= -\gamma\sigma\eta - \sigma^2\hat{\sigma}\tilde{v}_2^H z. \end{aligned} \tag{4.8}$$

The last equality follows from the fact that $v_1^H z = 0$ and $v_1^H \tilde{v}_2 = 0$.

We will transform r_+ to $V^H r_+$ before checking its norm. (Since $V^H V = I$, $\|r_+\| = \|V^H r_+\|$.) Recall that $V = (v_1, \hat{V}_{n-1})$. Put $p = \hat{V}_{n-1}^H \tilde{v}_2$ and $\hat{z} = \hat{V}_{n-1}^H z$. We will need the following formulae to simplify the expression for $V^H r_+$:

$$V^H v_1^+ = V^H(\gamma v_1 - \sigma\tilde{v}_2) = \begin{pmatrix} \gamma \\ -\sigma p \end{pmatrix}, \quad \text{and} \quad V^H z = \begin{pmatrix} 0 \\ \hat{z} \end{pmatrix}.$$

Since $V_k^H z = 0$ and $v^H z = 0$, the first k components of $\hat{z}$ are zeros. Clearly, $\|\hat{z}\| = \|z\|$. Now, it follows from (4.7) and (4.8) that

$$\begin{aligned} V^H r_+ &= V^H (A - \mu_+ I) v_1^+ \\ &= (-\sigma\eta) V^H v_1 + (\mu - \mu_+) V^H v_1^+ - (\hat{\sigma}\sigma) V^H z \\ &= (-\sigma\eta) e_1 + \left[\gamma\sigma\eta + \sigma^2(\tilde{v}_2^H z)\hat{\sigma}\right] \begin{pmatrix} \gamma \\ -\sigma p \end{pmatrix} - \sigma \begin{pmatrix} 0 \\ \hat{z}\hat{\sigma} \end{pmatrix} \\ &= (-\sigma\eta) \begin{pmatrix} 1 - \gamma^2 \\ \gamma\sigma p \end{pmatrix} + \sigma^2(\tilde{v}_2^H z)\hat{\sigma} \begin{pmatrix} \gamma \\ -\sigma p \end{pmatrix} - \sigma \begin{pmatrix} 0 \\ \hat{z}\hat{\sigma} \end{pmatrix} \\ &= (-\sigma\eta) \begin{pmatrix} \sigma^2 \\ \gamma\sigma p \end{pmatrix} + \sigma^2(\tilde{v}_2^H z)\hat{\sigma} \begin{pmatrix} \gamma \\ -\sigma p \end{pmatrix} - \sigma \begin{pmatrix} 0 \\ \hat{z}\hat{\sigma} \end{pmatrix}. \end{aligned}$$

It is easy to verify that $||p|| = 1$ since $p = \hat{V}_{n-1}^H \tilde{v}_2$, $\hat{V}_{n-1}^H \hat{V}_{n-1} = I_{n-1}$ and $\tilde{v}_2^H \tilde{v}_2 = 1$. Thus,

$$||r_+|| = ||V^H r_+|| = ||(-\sigma\eta)\begin{pmatrix} \sigma^2 \\ \gamma\sigma p \end{pmatrix} + \sigma^2(\tilde{v}_2^H z)\hat{\sigma}\begin{pmatrix} \gamma \\ -\sigma p \end{pmatrix} - \sigma\begin{pmatrix} 0 \\ \hat{z}\hat{\sigma} \end{pmatrix}||$$

$$\leq \sigma^2|\eta| + \sigma^2||z\hat{\sigma}|| + \sigma||z\hat{\sigma}|| \tag{4.9}$$

$$= \sigma^2|\eta| + \sigma^2\nu + \sigma\nu. \tag{4.10}$$

Recall that σ is generated to annihilate the sub-diagonal element of

$$\begin{pmatrix} 0 & \eta \\ \epsilon & \rho \end{pmatrix},$$

which appears in (4.2). Now, since

$$|\sigma| = \frac{|\epsilon|}{\sqrt{\epsilon^2 + \rho^2}} \leq |\frac{\epsilon}{\rho}| \quad \text{and} \quad |\epsilon| = ||r||,$$

we conclude that

$$||r_+|| \leq \psi(\mu, z)||r||,$$

where

$$\psi(\mu, z) = \frac{|\epsilon\eta|}{\rho^2} + \frac{|\epsilon|\nu}{\rho^2} + \frac{\nu}{|\rho|}.$$

Clearly, the factor $\psi(\mu, z)$ can be bounded uniformly if ν is not too large, and if $|\rho|$ can be bounded away from zero. Of course, one would not know the size of ρ until $k-1$ rotations $Q_1, Q_2, ..., Q_{k-1}$ have been applied. The following arguments provide an *a prior* lower bound for $|\rho|$. It asserts that $|\rho|$ can be bounded from below if ν is sufficiently small.

Recall from (4.3) that

$$(A - \mu I)v_1^+ = v_1(-\sigma\eta) + (-\sigma z\hat{\sigma}).$$

This is equivalent to

$$V^H(A - \mu I)VV^H v_1^+ = V^H v_1(-\sigma\eta) - \sigma\begin{pmatrix} 0 \\ \hat{z}\hat{\sigma} \end{pmatrix},$$

or

$$\begin{pmatrix} 0 & h^H \\ \epsilon e_1 & C - \mu I \end{pmatrix}\begin{pmatrix} \gamma \\ -\sigma p \end{pmatrix} = \begin{pmatrix} -\sigma\eta \\ -\sigma\hat{\sigma}\hat{z} \end{pmatrix},$$

where $h = v_1^H A\hat{V}_{n-1}$. It follows that

$$h^H p = \eta, \quad \text{and} \tag{4.11}$$

$$\rho e_1 - (C - \mu I)p = -\hat{\sigma}\hat{z}. \tag{4.12}$$

Since $e_1^T\hat{z} = 0$, it follows from (4.12) that

$$(C - \mu I)p = \begin{pmatrix} \rho \\ \hat{\sigma}\check{z} \end{pmatrix},$$

where $\breve{z}$ denotes the vector consisting of the last $n-2$ components of z. The assumption that μ is convergent to a simple eigenvalue of A ensures that $C - \mu I$ is nonsingular. Thus

$$p = (C - \mu I)^{-1} \begin{pmatrix} \rho \\ \hat{\sigma}\breve{z} \end{pmatrix}.$$

Recall that $p = \hat{V}_{n-1}\tilde{v}_2$ has unit length. Therefore,

$$1 = ||p|| \leq ||(C - \mu I)^{-1}|| \sqrt{\rho^2 + \hat{\sigma}^2 ||z||^2}.$$

or,

$$\frac{1}{||(C - \mu I)^{-1}||} \leq \sqrt{\rho^2 + \nu^2}. \tag{4.13}$$

Clearly, to establish a lower bound on ρ, one must prevent ν from getting too large.

Let $\zeta = 1/||(C - \mu)^{-1}||$. It follows from the assumption that $\nu < \zeta$ and equation (4.13) that

$$\zeta^2 - \nu^2 \leq \rho^2, \quad \text{or} \quad |\rho| \geq \sqrt{\zeta^2 - \nu^2}.$$

Consequently, we have

$$\psi(\mu, z) \leq \frac{|\epsilon\eta|}{\zeta^2 - \nu^2} + \frac{|\epsilon|\nu}{\zeta^2 - \nu^2} + \frac{\nu}{\sqrt{\zeta^2 - \nu^2}}. \tag{4.14}$$

As μ becomes sufficiently close to the desired eigenvalue, we may ignore the effect of the first two terms of (4.14), and focus on the dominating third term. It is easy to verify that if

$$\nu < \frac{\zeta}{\sqrt{2}},$$

$|\psi(\mu, z)|$ can be strictly bounded by 1. Monotonic convergence can be expected in this case. □

Remark 1 Note that the above convergence analysis is a local analysis. A global convergence analysis can be substantially more complicated, and is beyond the scope of this paper.

Remark 2 The above analysis is valid when the TRQ equation is solved exactly. One recovers the quadratic (or cubic if A is Hermitian) convergence rate of the Rayleigh quotient iteration. To see this, we replace equations (4.7) and (4.8) with

$$\begin{aligned} r_+ &= (-\sigma\eta)v_1, \\ \mu_+ - \mu &= -\gamma\sigma\eta. \end{aligned}$$

and conclude from (4.10) that

$$||r_+|| = ||V^H r_+|| = \sigma^2 |\eta|.$$

Since $\sigma = \epsilon/(\epsilon^2 + \rho^2)$ and $\epsilon = ||r||$,

$$||r_+|| = \frac{|\eta|}{\sqrt{\epsilon^2 + \rho^2}} ||r||^2 \leq \left|\frac{\eta}{\rho}\right| ||r||^2. \tag{4.15}$$

It follows from (4.12) that $\rho e_1 = (C - \mu I)p$. Thus $|\rho|$ is bounded below by $1/||(C - \mu I)^{-1}||$, and quadratic convergence follows from (4.15). When A is Hermitian, $|\eta| = |h^H p| \leq ||h||||p|| = |\epsilon| = ||r||$, and the cubic convergence rate follows.

Remark 3 We should point out that the bound given by (4.14) is not tight. This is a consequence of using the triangular inequality in (4.9). Thus in practice, the requirement $\nu \leq \zeta$ may be relaxed.

Remark 4 For Hermitian problems, ζ is approximately the gap between the eigenvalue to which the inexact TRQ is converging to and the eigenvalue nearest to it. This quantity can often be estimated by examining $|\mu - \hat{\mu}|$, where $\hat{\mu}$ is the eigenvalue of H_k that is nearest to μ.

5. Numerical Examples. In this section, we will demonstrate the convergence of the inexact TRQ by numerical examples. All computations shown in this section are performed in MATLAB 5.1 on a SUN-Ultra2. Two iterative solvers MINRES [4] and GMRES [7] are used in the following examples. Both solvers construct approximate solutions to a linear system from a Krylov subspace. The MINRES algorithm is mainly used for solving symmetric indefinite systems. Since it can be implemented by a short recurrence, only a few vectors need to be stored. A k-step GMRES algorithm requires an orthogonal basis of a k-dimensional Krylov subspace to be saved. To reduced the storage cost, the GMRES algorithm is often restarted using the approximate solution obtained in the previous run as a starting guess. We will use the notation `GMRES(k,m)` to denote a k-step GMRES with a maximum of m restarts. We set the convergence tolerance in both solvers to be 10^{-8}. We will also use `ITRQ(k,m)` to denote an inexact TRQ iteration in which k eigenpairs are to be computed and a m-step Arnoldi factorization is maintained.

5.1. Example 1 - Linear Convergence. The validity of the analysis presented in Section 4 is verified by a simple numerical example here. We choose the familiar 100×100 tridiagonal matrix with 2 on the diagonal and -1 on the sup- and sub-diagonal as the test matrix. The gap between the first two smallest eigenvalues is $\xi = 2.9 \times 10^{-3}$. To show the local convergence rate, we choose the starting vector v_0 of the initial Arnoldi factorization to be

$$v_0 = z_1 + 0.01 \cdot r,$$

where z_1 is the eigenvector corresponding to the smallest eigenvalue (λ_1) of A, and r is a normally distributed random vector. We apply `GMRES(10,5)` to (3.1) to obtain the vector $\tilde{v}_+$ used in (3.3). The residual error associated with the equation (3.3) and the first sub-diagonal element β_1 of the tridiagonal matrix are displayed in Table 5.1. The $(1,1)$-entry of the matrix, denoted by α_1, is also listed there. As ITRQ converges, we expect to see $\alpha_1 \rightarrow \lambda_1$, $v_1 \rightarrow z_1$, and $\beta_1 = ||Av_1 - \alpha_1 v_1|| \rightarrow 0$.

We observe from Column 4 that β_1 decreases monotonically in a linear fashion. This is in agreement with the theory developed in Section 4 since the damped residual error of (3.3) (Column 2 of Table 5.1 is less than the distance between the first two eigenvalues of A.)

iter.	α_1	$\|\|z\|\|$	β_1
1	3.0244×10^{-3}	-	7.8×10^{-3}
2	9.6750×10^{-4}	1.9×10^{-3}	7.3×10^{-5}
3	9.6742×10^{-4}	2.5×10^{-3}	3.2×10^{-6}
4	9.6744×10^{-4}	5.7×10^{-4}	1.5×10^{-7}
5	9.6744×10^{-4}	1.6×10^{-3}	6.7×10^{-9}
6	9.6744×10^{-4}	9.2×10^{-4}	3.2×10^{-10}
7	9.6744×10^{-4}	1.3×10^{-3}	1.6×10^{-11}

TABLE 5.1
The convergence of inexact TRQ.

5.2. Example 2 - Compute several eigenvalues. The following example illustrates that one can use the inexact TRQ iteration to compute more than one eigenpair. We also demonstrate that ITRQ is superior to the seemingly equivalent *inverse iteration with Wielandt deflation* [8] (INVWD.) The matrix used in the example corresponds to a discretized linear operator used in the stability analysis of the Brusselator wave model (BWM) [2]. Eigenvalues with the largest real parts are of interest. They help to determine the existence of stable periodic solutions to the Brusselator wave equation as a parameter varies. The dimension of the matrix is 200 × 200. The 32 rightmost eigenvalues are plotted in Figure 5.1. We place the target shift at $\sigma = 1.0$, and use `ITRQ(4,5)` to find 4 eigenvalues closest to σ. The equation (3.1) is solved by `GMRES(10,5)`. The residual norm of each approximate eigenpair is plotted against FLOPS in Figure 5.2. We marked residual norm at each iteration with a circle, and link these circles with solid (for ITRQ) or dotted lines (for INVWD) to show the convergence history of both methods.

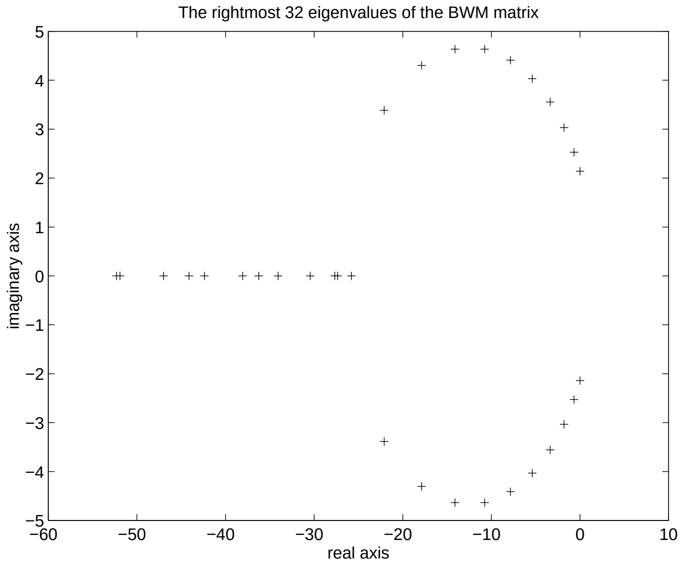

FIG. 5.1. *The 32 rightmost eigenvalues of a* 200 × 200 *BWM matrix.*

It appears that the convergence of four approximate eigenpairs occurs sequentially, i.e., the residual of the $j+1$st Ritz pair does not show significant decrease until the

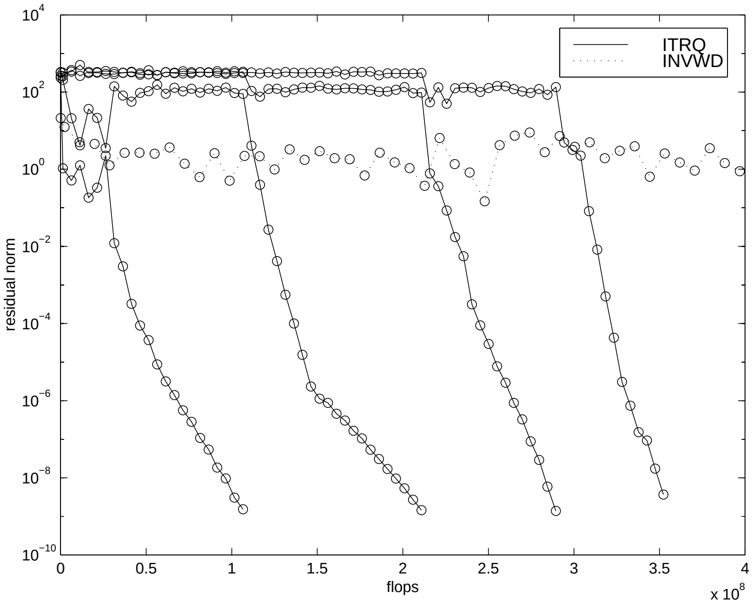

FIG. 5.2. *The convergence history of ITRQ and INVWD.*

j-th eigenpair has been found. Since we have shown in Section 4 that an approximate inverse iteration occurs in the inexact TRQ iteration, it will be interesting to compare the performance of ITRQ with an accelerated inverse iteration combined with Schur-Wielandt deflation [6, pp. 117]. The INVWD algorithm computes one eigenpair at a time by an approximate inverse iteration in which the linear system $(A - \mu I)w = v$ is solved by an iterative method. Instead of continuing the inverse iteration with

$$v \leftarrow \frac{w}{||w||}, \mu = \frac{v^H A v}{v^H v},$$

we compute an Arnoldi factorization using w as the starting vector, and choose (μ, v) from the Ritz pairs associated with this factorization. This approach can also be viewed as a sequence of restarted Arnoldi iterations in which the starting vector is repeatedly enhanced by an approximation inverse iteration. Once some eigenpairs have converged, we may apply Schur-Wielandt deflation to expose the subsequent eigenpairs. We will refer the interested reader to [8] for the detail of this algorithm. Unlike ITRQ, there is no error damping in INVWD. The equation $(A - \mu I)w = v$ must be solved rather accurately to guarantee the convergence of the inverse iteration. In this example, we use `GMRES(20,5)`. To make a fair comparison, we use a 5-step Arnoldi factorization to accelerate the inverse iteration. We observe from Figure 5.2 that, the residual curve corresponding to INVWD (dotted curve) zigzags around 1.0 and never shows significant decrease.

5.3. Example 3 - The Effect of Preconditioning. The previous two examples were presented merely to illustrate that it is possible to combine an iterative solver with the TRQ iteration. We should point out that in practice both problems can be solved with an exact TRQ or a shifted and inverted Arnoldi iteration because matrices involved in both examples can be efficiently factored.

As we mentioned before, the inexact TRQ is ideal for problems in which matrix factorization is prohibitively expensive. The following example carries this characteristic. We will show that with the help of a good preconditioner the speed of convergence of ITRQ can be drastically improved. The matrix $A \in \mathrm{R}^{256\times 256}$ used here arises from reactive scattering calculation [5]. Its sparsity pattern is shown in Figure 5.3. Although the matrix itself has only 6% non-zeros, a sparse factorization tends to fill up the entire matrix with non-zeros regardless of the reordering scheme used.

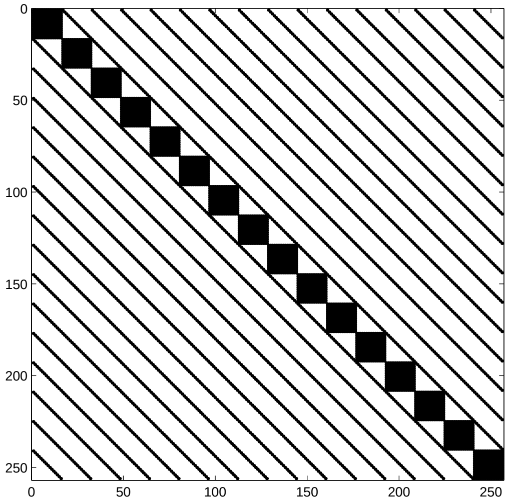

Fig. 5.3. *The sparsity pattern of the reactive scattering matrix.*

In this application, eigenvalues near zero are of interest. For clarity, we only show the convergence of the first eigenvalue. The convergence pattern for other eigenvalues is similar to this one. The solid curve in Figure 5.4 corresponds to the residual norm of the Ritz pair obtained from running a preconditioned *ITRQ(4,5)*.

The equation (3.1) is solved by MINRES with a convergence tolerence of 10^{-8}. A maximum of 100 steps are allowed in MINRES. If MINRES does not converge in 100 steps, the approximation generated at the 100th iteration is used to continue the inexact TRQ calculation. The preconditioner we used here is a matrix consisting of the dense diagonal blocks of the original matrix.

Without a preconditioner, `ITRQ(4,5)` (the dash-dotted curve) performs well at the beginning of the iteration when $A-\mu I$ is relatively well conditioned. The convergence slows down as the desired Ritz value gets close to the smallest eigenvalue. Even with the effect of damping, the residual error remained in (3.8) is not small enough to produce a qualitatively good starting vector for a subsequent Arnoldi factorization. With a good preconditioner, one can reduce the residual norm of (3.8) to a level which, combined with the TRQ damping, satisfies the conditions given in Theorem 4.1.

We also plotted, in Figure 5.4, the residual of the Ritz approximation obtained from an implicitly restarted Lanczos calculation, `IRL(1,29)` (the dotted curve) for comparison. (We use the notation `IRL(k,m)` to represent an IRL calculation in which the number of desired eigenvalue is k and the number of shifts applied during each restart is m.) We observe that IRL converges at a much slower rate in comparison with the preconditioned ITRQ.

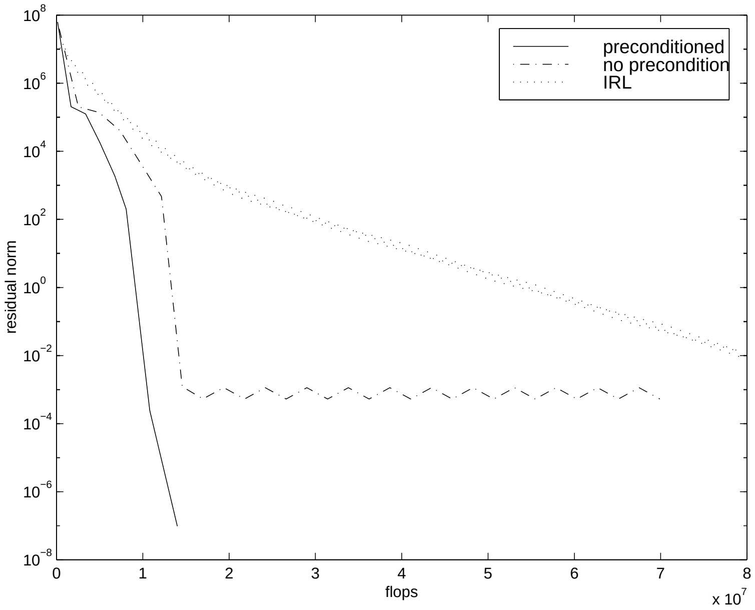

FIG. 5.4. *Comparison of (preconditioned) ITRQ with IRL.*

6. Conclusion.

We have analyzed the convergence of an inexact TRQ iteration, and showed that under some appropriate assumptions, the inexact TRQ iteration converges linearly with a small convergence factor. Our numerical examples confirmed our convergence analysis, and indicated the importance of constructing a good preconditioner for the iterative solver used in solving the TRQ equation.

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] Z. Bai, R. Barrett, D. Day, J. Demmel, and J. Dongarra. Test matrix collection (non-hermitian eigenvalue problems). Research report, Department of Mathematics, University of Kentucky, 1995.

[3] C. Lanzcos. An iteration method for the solution of the eigenvalue problem of linear differential and integral operators. *Journal of Research of the National Bureau of Standards*, 45(4):255–282, October 1950. Research Paper 2133.

[4] C. C. Paige and M. A. Saunders. Solutions of sparse indefinite systems of linear equations. *SIAM Journal on Numerical Analysis*, 12:617–629, 1975.

[5] P. Pendergast, Z. Darakjian, E. F. Hayes, and D. C. Sorensen. Scalable algorithms for three-dimensional reactive scattering: Evaluation of a new algorithm for obtaining surface functions. *Journal of Computational Physics*, 113(2):201–214, 1994.

[6] Y. Saad. *Numerical Methods for Large Eigenvalue Problems*. Halsted Press, 1992.

[7] Y. Saad and M. H. Schultz. GMRES: a generalized minimal residual algorithm for solving nonsymmetric linear systems. *SIAM Journal on Scientific and Statistical Computing*, 7:856–869, 1986.

[8] D.C. Sorensen and C. Yang. A truncated RQ-iteration for large scale eigenvalue calculations. Technical Report TR96-06, Department of Computational & Applied Mathematics, Rice Univeristy, Houston, TX 77005, 1996. To appear in SIAM Journal on Matrix Analysis and Applications.