BIT
2000, Vol. 40, No. 4, pp. 671–684
0006-3835/00/4004-0671 $15.00

INEXACT INVERSE ITERATION FOR GENERALIZED EIGENVALUE PROBLEMS *

GENE H. GOLUB[1†] and QIANG YE[2‡]

[1] *Department of Computer Science, Stanford University, Stanford, CA 94305, USA. email: golub@sccm.stanford.edu.*

[2] *Department of Mathematics, University of Kentucky, Lexington, KY 40506-0027, USA. email: qye@ms.uky.edu.*

Abstract.

In this paper, we study an inexact inverse iteration with inner-outer iterations for solving the generalized eigenvalue problem $Ax = \lambda Bx$, and analyze how the accuracy in the inner iterations affects the convergence of the outer iterations. By considering a special stopping criterion depending on a threshold parameter, we show that the outer iteration converges linearly with the inner threshold parameter as the convergence rate. We also discuss the total amount of work and asymptotic equivalence between this stopping criterion and a more standard one. Numerical examples are given to illustrate the theoretical results.

AMS subject classification: 65F15, 65F10.

Key words: Inverse iteration, shift-and-invert, inner-outer iterations.

1 Introduction.

The shift-and-invert transformation is a major spectral enhancement (preconditioning) technique used for solving large matrix eigenvalue problems. It is usually used in combination with an iterative method such as the power method (inverse iteration), subspace iteration [19], the Lanczos algorithm [7] and the Arnoldi algorithm [1] (see also [4, 13]). It also lies at the center of the recently developed Jacobi–Davidson method [15] and the rational Krylov subspace method [11]. Traditionally, it requires the inversion and thus factorization of a shifted matrix and is, therefore, limited to problems of moderate size. For many large scale problems, factorization of the matrix involved is either impractical or inefficient. In those cases, it has been proposed in recent years to use an iterative method (inner iterations) to solve the linear system, leading to an inexact shift-and-invert transformation with inner-outer iterations. Such an inexact technique has been studied for several methods, such as the Davidson and the Lanczos algorithm [2, 5, 10, 18], the rational Arnoldi algorithm and truncated

*Received March 1999. Communicated by Lars Eldén.

†Research supported in part by National Science Foundation Grant DMS-9403899.

‡Research supported by Natural Sciences and Engineering Research Council of Canada while this author was with University of Manitoba, Winnipeg, Canada.

RQ iterations [8, 9, 17] and the Jacobi–Davidson method [15]. Unfortunately, there is limited theoretical analysis of these methods under inexact solves and how the accuracy in the inner iterations affect the convergence. A challenging problem in implementations is how to choose the stopping threshold for the inner iterations.

In this paper, we are concerned with an inexact inverse iteration for solving the generalized eigenvalue problem $Ax = \lambda Bx$. In recent works [6, 16], it has been proved that, if the inner thresholds satisfy certain conditions, the inexact inverse iteration converges linearly at essentially the same rate as the exact case. Some practical ways of choosing the inner thresholds have also been suggested there. While it is probably not entirely surprising that the asymptotic convergence rate of the exact inverse iterations can be recovered with inexact solves, our interest here is to consider low accuracy solutions in the inner iterations and their effects on the outer iteration, rather than to recover the exact convergence rate of the inverse iterations. For that purpose, we shall consider a special inner stopping criterion depending on a threshold parameter, and show that the outer iteration converges linearly with the convergence rate directly given by that threshold parameter. We also discuss the total amount of work involved and equivalence between this stopping criterion and a more standard one. Numerical examples are given to illustrate the theoretical results.

We note that while the inverse iteration itself is rarely competitive compared with Krylov subspace projection methods, our interest in its behaviour under inexact solves is a first step towards understanding of more sophisticated methods, such as the inexact Jacobi–Davidson method. Indeed, the inexact shifted inverse iterations can be regarded as an extreme case of the inexact Jacobi–Davidson method and they both have the property that the inner accuracy mainly affects the convergence rate of the outer iterations but not the final accuracy achievable for the computed eigenvalues. We also note that the inverse iteration (or the subspace iteration) is still widely used in certain user communities because of its simplicity and ease in use. Indeed, our results will show that the inverse iteration is barely affected by the inexact solves, and thus may be more competitive in the context of inexact shift-and-invert transformations.

The paper is organized as follows. In Section 2, we present and analyze the inexact inverse iterations. In Section 3, we discuss the total amount of work and equivalence of two types of stopping criteria. Finally, we present some numerical examples illustrating the results in Section 4 and some concluding remarks in Section 5.

2 Inexact inverse iterations.

We consider solving the generalized eigenvalue problem

$$Ax = \lambda Bx,$$

with the following standard inverse iterations (A is typically a shifted matrix $A - \sigma B$ in the shift-and-invert strategy).

Inverse Iteration:
Given x_0;
For $k = 0, 1, 2, \ldots$ until convergence
$y_{k+1} = A^{-1}Bx_k$;
$x_{k+1} = y_{k+1}/\|y_{k+1}\|$.
End

In many applications, A^{-1} (i.e. factorizing A) is not explicitly available. Then at each iteration, an iterative method can be used to solve $Ay_{k+1} = Bx_k$, called inner iterations. In that case, the inverse iteration itself will be called the outer iteration. Thus, at each step of the outer iteration, we seek an approximate solution to $Ay_{k+1} = Bx_k$, such that the inner residual

$$q_k = Ay_{k+1} - Bx_k$$

satisfies a certain termination criterion. If we use y_k as an initial approximation to y_{k+1}, we aim at solving

$$Ad_k = Bx_k - Ay_k \quad \text{with } y_{k+1} = y_k + d_k. \tag{2.1}$$

Such a shifted system is also exploited in the context of Cayley transform in [8, 9]. The main purpose of this section is to analyze the convergence characteristic under inexact solves.

In light of solving (2.1) by an inner iteration, a natural stopping criterion for the inner iteration is

$$\|q_k\| < \epsilon\|Bx_k - Ay_k\|, \tag{2.2}$$

where ϵ is a threshold parameter. While this criterion is convenient for implementations, we shall also consider the following stopping criterion:

$$\|q_k\| < \epsilon_k\|y_{k+1}\|, \tag{2.3}$$

where ϵ_k is a stopping threshold. Here we allow ϵ_k to vary from step to step. Involving $\|y_{k+1}\|$, this termination criterion is less natural, but will turn out to be theoretically easier to analyze. Indeed, we shall give a theoretical analysis for (2.3) and show that, for a linearly convergent ϵ_k, the two stopping criteria are asymptotically equivalent, and thus indirectly provide an analysis of (2.2) which is practically more convenient. We also note that a stopping criterion of the form (2.3) has also been suggested in [12] for a different purpose.

Let $\max(y)$ denote the first entry of the vector y that has the maximal absolute value among all entries. We consider the following two versions of inexact inverse iterations, depending on which of (2.2) and (2.3) is used:

Inexact Inverse Iteration:
Given x_0; set $y_0 = 0$.
For $k = 0, 1, \ldots$ until convergence
$r_k = Bx_k - Ay_k$;
Solve $Ad_k = r_k$ by an inner iteration such that

$q_k = Ad_k - r_k$ satisfies (2.3) (or (2.2));
$y_{k+1} = y_k + d_k$;
$\sigma_{k+1} = \max(y_{k+1})$;
$x_{k+1} = y_{k+1}/\sigma_{k+1}$.

REMARK 2.1. The norm used in the inner iteration stopping criterion can be any of the 1-norm, 2-norm or ∞-norm. By choosing $\sigma_{k+1} = \max(y_{k+1})$, y_{k+1} is normalized with respect to the ∞-norm. Furthermore, at convergence, $\mu_k = 1/\sigma_k$ is an approximation to an eigenvalue and thus $r_k = Bx_k - \sigma_k Ax_k$ is a (outer) residual (scaled by σ_k). This is also true if we choose $\sigma_{k+1} = \ell^T y_{k+1}$ for a fixed vector ℓ, but not so if we simply normalize y_{k+1} with $\sigma_{k+1} = \|y_{k+1}\|$.

2.1 Convergence analysis.

We now discuss the convergence properties of the inexact algorithm. In this section we shall only consider the stopping criterion (2.3).

Let $\lambda_1, \lambda_2, \ldots, \lambda_n$ be the eigenvalues of $B^{-1}A$ ordered such that

$$|\lambda_1| \le |\lambda_2| \le \cdots \le |\lambda_n|.$$

Throughout this work, we assume that $B^{-1}A$ is diagonalizable and

$$|\lambda_1| < |\lambda_2|, \quad \text{or} \quad \rho = |\lambda_1/\lambda_2| < 1.$$

Let $u_1, u_2, \ldots, u_n$ (and $v_1, v_2, \ldots, v_n$) be the corresponding left (and right, resp.) eigenvectors normalized such that,

$$u_i^T A = \lambda_i u_i^T B, \quad Av_i = \lambda_i Bv_i, \quad \|v_i\| = 1, \quad u_i^T Av_j = \delta_{ij},$$

where δ_{ij} is the Kronecker symbol. Then $U^T AV = I$, where

$$U = [u_1, u_2, \ldots, u_n], \qquad V = [v_1, v_2, \ldots, v_n].$$

Define $x_k^{(i)} = u_i^T Ax_k$. Then $x_k^{(i)}$ is just the component of x_k on the ith eigenvector v_i, i.e.

$$x_k = \sum_{i=1}^{n} x_k^{(i)} v_i.$$

We further define, for $x_k^{(1)} \neq 0$,

$$t_k = \frac{\|(x_k^{(2)}, \ldots, x_k^{(n)})^T\|}{|x_k^{(1)}|},$$

where $\|\cdot\|$ is the 1-norm, 2-norm or ∞-norm (for example, $t_k = \sum_{i=2}^{n} |x_k^{(i)}|/|x_k^{(1)}|$, if $\|\cdot\|$ is the 1-norm). Clearly, t_k is a measure of the approximation of x_k to v_1. Indeed, the following proposition shows that it is an upper bound on the relative error of the approximation:

PROPOSITION 2.1. *Let t_k be defined as above. Then we have*

$$\frac{t_k}{\|V^{-1}\|} \le \frac{\|x_k - x_k^{(1)} v_1\|}{\|x_k^{(1)} v_1\|} \le \|V\| t_k.$$

In particular, for the 1-norm, we have $\|x_k - x_k^{(1)} v_1\|_1 / \|x_k^{(1)} v_1\|_1 \le t_k$. The same holds for the 2-norm if V is orthogonal.

PROOF. Let $\hat{V} = [v_2, \ldots, v_n]$. Clearly, $x_k = x_k^{(1)} v_1 + \hat{V}(x_k^{(2)}, \ldots, x_k^{(n)})^T$. Thus,

$$\frac{\|x_k - x_k^{(1)} v_1\|}{\|x_k^{(1)} v_1\|} = \frac{\|\hat{V}(x_k^{(2)}, \ldots, x_k^{(n)})^T\|}{|x_k^{(1)}|} \le \|\hat{V}\| t_k \le \|V\| t_k.$$

On the other hand,

$$\|\hat{V}(x_k^{(2)}, \ldots, x_k^{(n)})^T\| = \|V(0, x_k^{(2)}, \ldots, x_k^{(n)})^T\| \ge \frac{\|(x_k^{(2)}, \ldots, x_k^{(n)})^T\|}{\|V^{-1}\|},$$

which shows the lower bound. Finally, for the 1-norm, $\|V\|_1 = 1$ since $\|v_i\|_1 = 1$, and for the 2-norm, $\|V\|_2 = 1$ if V is orthogonal. □

We now discuss the convergence of t_k.

LEMMA 2.2. *Let $\rho = |\lambda_1/\lambda_2| < 1$ and $x_{k+1}^{(1)} \ne 0$. Then,*

$$\begin{aligned} t_{k+1} &\le \rho t_k + \epsilon_k \|U^T\| \frac{\|x_{k+1}\|}{|x_{k+1}^{(1)}|}(1 + \rho t_k) \\ &\le \rho t_k + \epsilon_k \|U^T\| \|V\| f(t_{k+1})(1 + \rho t_k) \end{aligned} \tag{2.4}$$

where $f(t) = \sqrt{t^2+1}$, $t+1$ or $\max\{1, t\}$ if $\|\cdot\|$ is the 2-norm, the 1-norm or the ∞-norm respectively.

PROOF. From the algorithm, $\sigma_{k+1} A x_{k+1} = B x_k + q_k$. Writing $\mu_{k+1} = 1/\sigma_{k+1}$, we have

$$A x_{k+1} = \mu_{k+1} B x_k + \mu_{k+1} q_k = \mu_{k+1} B x_k + e_k, \tag{2.5}$$

where $e_k = \mu_{k+1} q_k$ and, by the bound on q_k, we have $\|e_k\| < \epsilon_k \|\mu_{k+1} y_{k+1}\| = \epsilon_k \|x_{k+1}\|$. Thus, $u_i^T A x_{k+1} = \mu_{k+1} \lambda_i^{-1} u_i^T A x_k + u_i^T e_k$ or $x_{k+1}^{(i)} = \mu_{k+1} \lambda_i^{-1} x_k^{(i)} + \eta_k^{(i)}$, where $\eta_k^{(i)} = u_i^T e_k$. Hence, for $2 \le i \le n$,

$$\frac{x_{k+1}^{(i)}}{x_{k+1}^{(1)}} = \frac{\mu_{k+1} \lambda_i^{-1} x_k^{(i)} + \eta_k^{(i)}}{\mu_{k+1} \lambda_1^{-1} x_k^{(1)} + \eta_k^{(1)}} = \frac{\lambda_1}{\lambda_i} \frac{x_k^{(i)}}{x_k^{(1)}} + \frac{\eta_k^{(i)}}{x_{k+1}^{(1)}} - \frac{\eta_k^{(1)}}{x_{k+1}^{(1)}} \frac{\lambda_1}{\lambda_i} \frac{x_k^{(i)}}{x_k^{(1)}}.$$

Stacking the above in the vector form (for $2 \le i \le n$) and taking the norm, we obtain

$$t_{k+1} \le \left|\frac{\lambda_1}{\lambda_2}\right| t_k + \frac{\|(\eta_k^{(2)}, \ldots, \eta_k^{(n)})^T\|}{|x_{k+1}^{(1)}|} + \frac{|\eta_k^{(1)}|}{|x_{k+1}^{(1)}|} \left|\frac{\lambda_1}{\lambda_2}\right| t_k$$

$$\leq \rho t_k + \frac{\|(\eta_k^{(1)}, \eta_k^{(2)}, \ldots, \eta_k^{(n)})^T\|}{|x_{k+1}^{(1)}|} \left(1 + \left|\frac{\lambda_1}{\lambda_2}\right| t_k\right)$$

$$\leq \rho t_k + \frac{\|U^T\|\epsilon_k\|x_{k+1}\|}{|x_{k+1}^{(1)}|}(1 + \rho t_k) \tag{2.6}$$

where we have used $\|(\eta_k^{(1)}, \eta_k^{(2)}, \ldots, \eta_k^{(n)})^T\| = \|U^T e_k\| \leq \|U^T\|\epsilon_k\|x_{k+1}\|$. This proves the first part of the bound.

Furthermore, from the definition of t_k we have $\|[x_k^{(1)}, x_k^{(2)}, \ldots, x_k^{(n)}]^T\|/|x_k^{(1)}| = f(t_k)$. Then

$$|x_k^{(1)}| = \frac{\|U^T A x_k\|}{f(t_k)} \geq \frac{\|x_k\|}{f(t_k)\|(U^T A)^{-1}\|},$$

which proves the second part of the bound. □

We next give two results concerning bounds for the coefficient of ϵ_k.

LEMMA 2.3. *If*

$$\epsilon_k \leq \epsilon = \frac{(1-\rho)t_0}{\|U^T\|\|V\|(1+\rho t_0)(1+t_0)},$$

for all k, then $t_k \leq t_0$.

PROOF. We prove $t_k \leq t_0$ by induction. Supposing $t_k \leq t_0$ is true for some k, we show $t_{k+1} \leq t_0$. First, if $x_{k+1}^{(1)} \neq 0$, it follows from Lemma 2.2 and $f(t) \leq t+1$ that

$$\begin{aligned} t_{k+1} &\leq \rho t_k + \epsilon_k \|U^T\|\|V\|(t_{k+1}+1)(1+\rho t_k) \\ &\leq \rho t_0 + \epsilon \|U^T\|\|V\|(t_{k+1}+1)(1+\rho t_0), \end{aligned}$$

which implies

$$t_{k+1} \leq \frac{\rho t_0 + \|U^T\|\|V\|\epsilon(1+\rho t_0)}{1 - \|U^T\|\|V\|\epsilon(1+\rho t_0)} \leq t_0,$$

where we note that $1 - \|U^T\|\|V\|\epsilon(1+\rho t_0) > 0$ by the assumption on ϵ. Now, if $x_{k+1}^{(1)} = 0$, let $\tilde{y}_{k+1} = y_{k+1} + \delta v_1$ and $\tilde{x}_{k+1} = \tilde{y}_{k+1}/\tilde{\sigma}_{k+1}$ where $\tilde{\sigma}_{k+1} = \max(\tilde{y}_{k+1})$. Then

$$A\tilde{y}_{k+1} = Bx_k + q_k + \delta A v_1 = Bx_k + \tilde{q}_k$$

where $\tilde{q}_k = q_k + \delta A v_1$. Since $\|q_k\| < \epsilon_k\|y_{k+1}\|$, we have $\|\tilde{q}_k\| < \epsilon_k\|\tilde{y}_{k+1}\|$ for sufficiently small δ. Now, define $\tilde{t}_{k+1}$ from $\tilde{x}_{k+1}$ as in t_{k+1}. Since $\tilde{x}_{k+1}^{(1)} = \delta/\tilde{\sigma}_{k+1} \neq 0$, we have $\tilde{t}_{k+1} \leq t_0$. On the other hand, letting $\delta \to 0$, we obtain $\tilde{t}_{k+1} \to \infty$ from Proposition 2.1, a contradiction. Therefore $x_{k+1}^{(1)} \neq 0$ and hence $t_{k+1} \leq t_0$. This completes the proof. □

LEMMA 2.4. *If* $|x_k^{(1)}| > \alpha\|x_k\|$ *for all k, then* $t_k \leq T = \|U^T A\|/\alpha$.

PROOF. $[x_k^{(1)}, x_k^{(2)}, \ldots, x_k^{(n)}]^T = U^T A x_k$. Thus

$$\|[x_k^{(2)}, \ldots, x_k^{(n)}]^T\| \leq \|U^T A x_k\| \leq \|U^T A\|\|x_k\|.$$

It follows that $t_k \leq \|U^T A\|/\alpha$. □

Under the assumption of Lemma 2.3 (or Lemma 2.4), t_k is bounded above. Then from Lemma 2.2,

$$t_{k+1} \le \rho t_k + \epsilon_k C$$

where $C \le \|U^T\|\|V\|(1+t_0)(1+\rho t_0) \le \|U^T\|\|V\|(1+t_0)^2$. This suggests that t_k decreases linearly, though not necessarily converges to 0. It is therefore natural to consider linearly convergent $\epsilon_k = a\gamma^k$.

THEOREM 2.5. *Let*

$$C = \|U^T\| \max_{k\ge 0} \frac{\|x_{k+1}\|(1+\rho t_k)}{|x_{k+1}^{(1)}|}.$$

If $\epsilon_k = a\gamma^k$ *with* $\gamma \le 1$ *and*

$$a \le \frac{(1-\rho)t_0}{\|U^T\|\|V\|(1+\rho t_0)(1+t_0)},$$

then $C \le \|U^T\|\|V\|(1+t_0)^2$. *Furthermore, we have*

$$t_k \le \begin{cases} \rho^k t_0 + \frac{\gamma^k - \rho^k}{\gamma - \rho} aC, & \text{if } \gamma \ne \rho, \\ t_k \le \rho^k t_0 + k\rho^{k-1} aC, & \text{if } \gamma = \rho. \end{cases}$$

PROOF. From the proof of Lemma 2.2, $\|x_{k+1}\|/|x_{k+1}^{(1)}| \le \|V\| f(t_{k+1})$. Then the bound on C follows from Lemma 2.3. From Lemma 2.2, we have

$$t_{k+1} \le \rho t_k + \epsilon_k C. \tag{2.7}$$

Next, consider the sequence g_k defined by the difference equation

$$g_{k+1} = \rho g_k + \epsilon_k C, \quad g_0 = t_0, \tag{2.8}$$

which has the solution $g_k = \rho^k t_0 + \frac{\gamma^k - \rho^k}{\gamma - \rho} aC$ if $\gamma \ne \rho$, and $g_k = \rho^k t_0 + k\rho^{k-1} aC$, if $\gamma = \rho$. It is easy to show that $t_k \le g_k$ and the theorem is proved. □

REMARK 2.2. We can also consider the strategy $\epsilon_k = \epsilon$. Then it can be proved that

$$t_k \le \rho^k t_0 + \epsilon \frac{1-\rho^k}{1-\rho} C.$$

So t_k decreases to the level of ϵ at the rate ρ and need not converge to 0. However, this is clearly a less efficient method and we shall not discuss it further.

The conclusion of the above theorem is that t_k decreases linearly at the rate of $\max\{\rho, \gamma\}$. The condition on a is to ensure convergence and is clearly not a necessary condition.

Since choosing $\gamma < \rho$ leads to no gain in convergence rate and is almost certain to be inefficient, we shall concentrate now on the case $\gamma > \rho$. The following corollary gives a more precise bound for the constant C and hence for t_k at the convergence stage:

COROLLARY 2.6. *Let $1 > \gamma > \rho$ and $\epsilon_k = a\gamma^k$ with a satisfying the condition in Theorem* 2.5. *Then for sufficiently large k_0,*

$$C_0 = \|U^T\| \max_{k \geq k_0} \frac{\|x_{k+1}\|(1+\rho t_k)}{|x_{k+1}^{(1)}|} \sim \|U^T\|$$

and $\limsup \frac{t_k}{a\gamma^{k-1}} \leq (1-\rho/\gamma)^{-1}\|U^T\|$. *Furthermore, for $k \geq k_0$,*

$$t_k \leq \rho^{k-k_0} t_{k_0} + \frac{\gamma^k - \rho^k}{\gamma - \rho} a C_0 \sim a\gamma^{k-1} \min\{k-k_0, (1-\rho/\gamma)^{-1}\}\|U^T\|.$$

PROOF. By Prop. 2.1, we have $x_k/x_k^{(1)} \to v_1$ and then $\lim \frac{\|x_{k+1}\|(1+\rho t_k)}{|x_{k+1}^{(1)}|} = 1$. Thus $\lim_{k_0 \to \infty} C_0 = \|U^T\|$. Apply Theorem 2.5 to t_k starting from $k = k_0$, we obtain

$$\begin{aligned} t_k &\leq \rho^{k-k_0} t_{k_0} + \frac{\gamma^{k-k_0} - \rho^{k-k_0}}{\gamma - \rho} a\gamma^{k_0} C_0 \\ &\leq \rho^{k-k_0} t_{k_0} + \gamma^{k-k_0-1} \min\{k-k_0, (1-\rho/\gamma)^{-1}\} a\gamma^{k_0} C_0 \\ &\sim a\gamma^{k-1} \min\{k-k_0, (1-\rho/\gamma)^{-1}\}\|U^T\|. \end{aligned}$$

Take $k \to \infty$ first and then $k_0 \to \infty$ in the first inequality, we obtain the bound for $\limsup \frac{t_k}{a\gamma^{k-1}}$. □

3 Asymptotic analysis of computational work.

In this section, we discuss the effect of γ on the total amount of work and show that the two kinds of stopping criteria (2.2) and (2.3) are asymptotically equivalent. Again, we consider the case $\gamma > \rho$ only.

An examination of the algorithm for inexact inverse iteration shows that the computational work at each step of the outer iteration depends predominantly on the accuracy requirement (i.e. the inner residual reduction) in the inner iteration. With the termination criterion (2.3) $\|q_k\| \leq \epsilon_k \|y_{k+1}\|$, the number of inner iterations required at step k is determined by the (inner) residual reduction ratio

$$\delta_k = \frac{\epsilon_k \|y_{k+1}\|}{\|r_k\|}$$

while with (2.2), it is by the constant ϵ. We observe that, although $\epsilon_k\|y_{k+1}\| \sim a\gamma^k\lambda_1^{-1}$ is required to be smaller and smaller, as we will see, $\|r_k\| \sim t_k$ decreases as well. Indeed, asymptotically, they decrease at the same rate and thus the ratio δ_k is essentially a constant. Mathematically, we state it as the following theorem:

THEOREM 3.1. *Let $1 > \gamma > \rho$ and $\epsilon_k = a\gamma^k$ with a satisfying the condition in Theorem* 2.5. *Assume that the maximum entry of v_1 in absolute value is unique. Then*

$$E_1\gamma \leq \liminf \delta_k \leq \limsup \delta_k \leq E_2\gamma,$$

where

$$E_1^{-1} = \frac{\max_{i\neq 1}|\lambda_i - \lambda_1|}{1-\rho/\gamma}\|U^T\|\Big(1+\frac{1}{\|v_1\|_\infty}\Big)\|B\|\,\|V\| + \frac{\|B\|\,\|B^{-1}\|}{\|v_1\|_\infty},$$

and

$$E_2^{-1} = \frac{\min_{i\neq 1}|\lambda_i - \lambda_1|}{\|(BV)^{-1}\|}\liminf\frac{t_k}{a\gamma^{k-1}}.$$

PROOF. From $t_k \to 0$ and Proposition 2.1, we obtain $x_k/x_k^{(1)} \to v_1$. Then, $\lim\|x_k\| = \lim\|x_k\|/\|x_k\|_\infty = \lim\|x_k/x_k^{(1)}\|/\|x_k/x_k^{(1)}\|_\infty = 1/\|v_1\|_\infty$ and $\lim|x_k^{(1)}| = \lim\|x_k^{(1)}v_1\| = \lim\|x_k\|$. Now, write $z = [z_1, z_2, \ldots, z_n]^T$ and $\hat{z} = [0, z_2, \ldots, z_n]^T$ where $z_i = (\lambda_i - \mu_k)x_k^{(i)}$. Then

$$\begin{aligned}\frac{\|r_k\|}{|\sigma_k|} &= \|(A-\mu_k B)x_k\| = \Big\|\sum_{i=1}^n(\lambda_i-\mu_k)x_k^{(i)}Bv_i\Big\| \\ &= \|BVz\| \geq \|z\|/\|(BV)^{-1}\| \\ &\geq \|\hat{z}\|/\|(BV)^{-1}\| \geq t_k|x_k^{(1)}|\min_{i\neq 1}|\lambda_i-\mu_k|/\|(BV)^{-1}\|\end{aligned}$$

where we note that $\|\cdot\|$ is a 1, 2, or ∞-norm. Thus

$$\liminf\frac{\|r_k\|}{\epsilon_{k-1}|\sigma_k|\|x_{k+1}\|} \geq \frac{\min_{i\neq 1}|\lambda_i-\lambda_1|}{\|(BV)^{-1}\|}\liminf\frac{t_k}{\epsilon_{k-1}}. \tag{3.1}$$

On the other hand, let the unique maximum entry of v_1 in absolute value is the j-th entry. Then, for sufficiently large k, the j-th entry of x_k is 1, since $\max(x_k) = 1$. From (2.5), we have that $\mu_k x_{k-1} = B^{-1}Ax_k - B^{-1}e_{k-1} = \lambda_1 x_k + (B^{-1}A - \lambda_1)x_k - B^{-1}e_{k-1}$. It follows from comparing the j-th entry that

$$\begin{aligned}|\mu_k - \lambda_1| &\leq \|(B^{-1}A-\lambda_1)x_k\| + \|B^{-1}e_{k-1}\| \\ &\leq \Big\|\sum_{i=2}^n(\lambda_i-\lambda_1)x_k^{(i)}v_i\Big\| + \|B^{-1}\|\|e_{k-1}\| \\ &\leq |\lambda_p-\lambda_1|\,\|V\|t_k|x_k^{(1)}| + \epsilon_{k-1}\|B^{-1}\|\|x_{k-1}\|.\end{aligned}$$

where $|\lambda_p - \lambda_1| = \max_{i\neq 1}|\lambda_i - \lambda_1|$. Therefore, $\lim\sigma_k = \lim 1/\mu_k = 1/\lambda_1$. Furthermore,

$$\begin{aligned}\frac{\|r_k\|}{|\sigma_k|} &= \|(\lambda_1-\mu_k)x_k^{(1)}Bv_1 + BV\hat{z}\| \\ &\leq |\lambda_1-\mu_k|\,|x_k^{(1)}|\,\|B\| \;+ t_k|x_k^{(1)}|\,|\lambda_p-\lambda_1|\,\|BV\| \\ &\leq t_k|x_k^{(1)}|(1+|x_k^{(1)}|)|\lambda_p-\lambda_1|\,\|B\|\,\|V\| + \epsilon_{k-1}|x_k^{(1)}|\|B\|\,\|B^{-1}\|\|x_{k-1}\|.\end{aligned}$$

Thus, using Corollary 2.6,

$$\begin{aligned}\limsup\frac{\|r_k\|}{\epsilon_{k-1}|\sigma_k|\|x_{k+1}\|} \leq (1-\rho/\gamma)^{-1}\|U^T\|\Big(1+\frac{1}{\|v_1\|_\infty}\Big)|\lambda_p-\lambda_1|\,\|B\|\,\|V\| \\ + \frac{\|B\|\,\|B^{-1}\|}{\|v_1\|_\infty}.\end{aligned}$$

Finally, this and (3.1) lead to the theorem. □

This result shows that for sufficiently large k,

$$E_1\gamma \le \delta_k \le E_2\gamma.$$

Thus the accuracy requirement for the inner iterations is at most $\|q_k\|/\|r_k\| \le E_1\gamma$. In this way, (2.3) amounts to an asymptotically constant reduction in $\|q_k\|/\|r_k\| \sim E\gamma$ for some $E_1 \le E \le E_2$. Namely, at the convergence stage, (2.3) is effectively the same as (2.2) with $\epsilon \sim E\gamma$. This provides an explanation to the observed convergence of the inexact inverse iteration with the termination criterion (2.2). We note however that, in the case of (2.2), the threshold parameter ϵ is not explicitly related to the convergence rate of the outer iteration, which is $\gamma \sim \epsilon/E$. Namely, with (2.2), we only know that the outer iteration converges at the rate of ϵ/E for some unknown E. In contrast, with (2.3), the rate is explicitly γ (see Example 3 in Section 4).

In order to analyze the total work, we assume that the inner iterative method converges linearly with the rate ι. We note that such a model may not reflect the true convergence behaviour of most Krylov subspace methods; but an assumption is necessary in order to analyze the total work. Then the number of inner iterations required to achieve the reduction $\frac{\|q_k\|}{\|r_k\|} \le E\gamma$ is $p(\gamma, k) \sim \ln(E\gamma)/\ln(\iota)$. Since the outer iteration converges at the rate of γ, the number of outer steps required to reduce t_k by e^{-1} is $q(\gamma) = -1/\ln(\gamma)$. Thus, asymptotically, the total number of inner iteration steps required for the reduction of the outer residual by e^{-1} is

$$q(\gamma)p(\gamma, k) = -\frac{1}{\ln(\iota)}\left(1 + \frac{\ln(E)}{\ln(\gamma)}\right)$$

This formula shows that the total number of inner iterations has relatively insensitive dependence on γ, but it also suggests that it approaches ∞ as $\gamma \to 1$, which merely reflects the theoretical situation that when the inner solution has no accuracy at all, the outer iteration is not expected to converge. In practice, however, we still observed convergence even when $\gamma = 1$, and this is due to the fact that the inner solution (by say one iteration) has an accuracy much better than 1. Thus, an inexact inverse iteration with $\gamma = 1$ would in effect be the same as one carried out with a smaller γ. This shows the limitation of characterizing the accuracy of the inner solutions by the stopping threshold; namely, the solutions are usually more accurate than the threshold imposes (see Example 1 in Section 4).

We also note that, in our numerical tests, the total number of inner iterations increases only slightly as γ increases and, in any case, it is not significantly affected by γ in the range $\rho < \gamma < 1$. This suggests that $\ln(E)$ and hence the second term $\frac{\ln(E)}{\ln(\gamma)}$ are small compared with 1 when γ is not too close to 1. We note that the linear system solver takes $-1/\ln(\iota)$ iterations to reduce the (linear system) residual by e^{-1}. Hence, the numbers of iterations required to achieve a certain reduction for the outer eigenvector residual and for the (linear system) residual are comparable.

Finally, as a practical conclusion, we suggest that γ should be chosen to ensure $\gamma \geq \rho$ while as small as possible under that condition. One benefit of choosing smaller γ in the range $\gamma \geq \rho$ is that it would also reduce the number of outer iterations.

4 Numerical examples.

In this section, we present some numerical examples to illustrate the theoretical results. We shall consider $\epsilon_k = \gamma^k$ only. All inner iterations are solved by restarted GMRES(10) [14] unless otherwise specified.

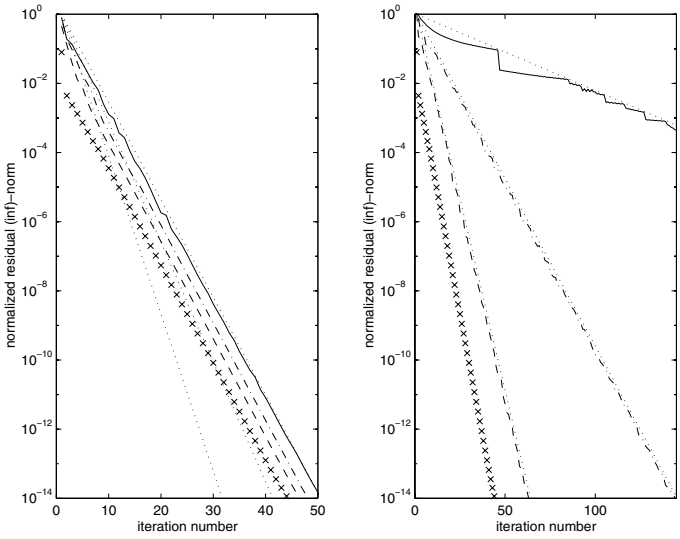

Figure 4.1: Example 4.1: Residual convergence history for various γ .

EXAMPLE 4.1. The matrix is the finite-difference discretization (center difference) on a 32×32 grid of the following eigenvalue problem of the convection diffusion operator:

$$-\triangle u + 5u_x + 5u_y = \lambda u \quad \text{on} \quad (0,1)^2,$$

with the homogeneous Dirichlet boundary condition. We consider finding the smallest eigenvalue (for which $\rho \approx 0.5225$) by Algorithm 2 with a random initial vector. We consider in this example the convergence behaviour of the outer iteration under different parametric values of γ. In Figure 4.1, we present the convergence history of the residual $\|r_k\|_\infty$ and the threshold $\epsilon_k = \gamma^k$ (in dotted lines) for various values of γ. The residuals are plotted in the solid, dash-dotted, and dashed lines while ϵ_k is plotted in the dotted lines from the top down for $\gamma = \rho$, 0.45, and 0.35 respectively in the left figure and for $\gamma = 0.95$, 0.8, and 0.6 respectively in the right figure. On both, the residual for the exact inverse iteration is plotted in the "$\times$" mark.

The results clearly demonstrates the linear convergence behaviour and the convergence rate is the same as γ for $\gamma \geq \rho$ and it is ρ for $\gamma < \rho$. For the larger γ (the case $\gamma = 0.95$), because the inner solutions are usually more accurate than the termination criterion imposes, the residual curve is not strictly linear but its linear trend is still clear.

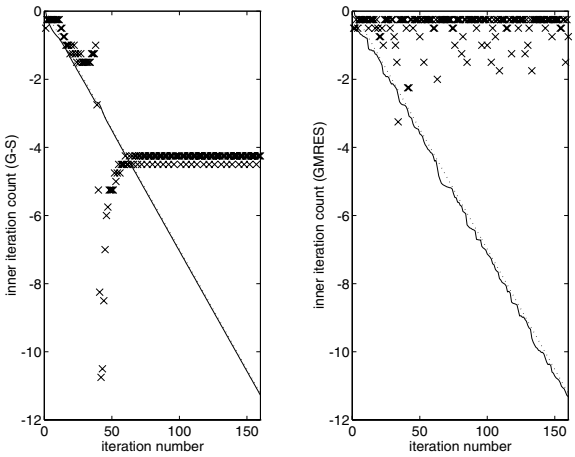

Figure 4.2: Example 4.2: Number of inner iterations at outer steps.

EXAMPLE 4.2. In this example, the matrix is the same as in Example 4.1 and we consider the number of inner iterations at each step of the outer iterations. As we have shown in section 3, the (inner) residual reduction as required is asymptotically constant and so is the number of the inner iterations if the inner iterative method used is linearly convergent. However, this need not be the case for Krylov subspace methods which often converge superlinearly. For this, we consider using the Gauss–Seidel method and the full GMRES in the inner iterations (with $\gamma = 0.85$) and the results are given in Figure 4.2 (left and right, respectively). In the figure, the number of the inner iterations (as scaled by $-1/4$) is plotted against the outer iteration steps in the "$\times$" mark. We also plot the corresponding (outer) residual (in solid line) and the threshold ϵ_k(in dotted line) (in the log scale).

We first observe that the convergence rate of the outer iteration is not affected by the use of different inner iterative methods. Furthermore, the number of inner iterations is clearly near a constant for the Gauss–Seidel method at the convergence stage as is expected. For GMRES, it is also near a constant at most steps with some large variations at others.

EXAMPLE 4.3. In this example, we study how the threshold parameter γ affects the total number of inner iterations (to reduce the residual to 10^{-12}). We consider both stopping criteria (2.2) and (2.3) and run them for a range of γ and ϵ. We test it for two matrices. One is the convection diffusion matrix in

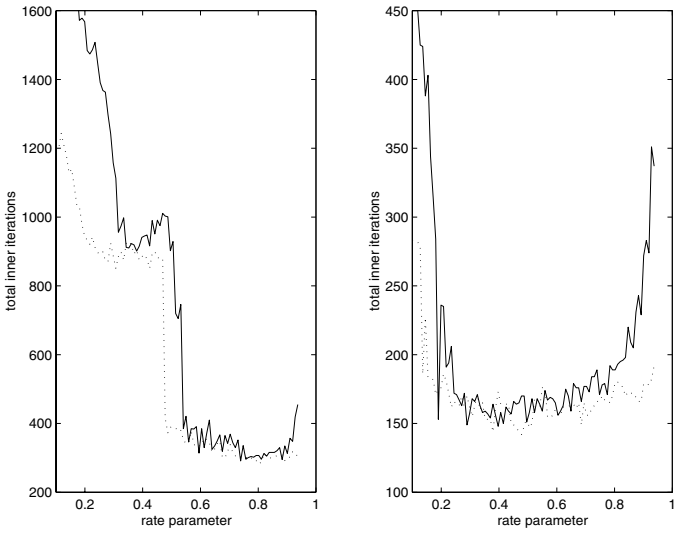

Figure 4.3: Example 4.3: Total inner iterations vs. rate parameter (γ or ϵ) .

Example 4.1 (for which $\rho \approx 0.5225$) and the other is the JPWH991 matrix (for which $\rho \approx 0.2799$) in the Harwell-Boeing collection [3] of sparse test matrices and we present the results in Figure 4.3 on the left and right respectively. The total number of inner iterations are plotted in Figure 4.3 in the solid line for (2.3) against γ and in the dotted line for (2.2) against ϵ.

For (2.3) (the solid line), the total number of inner iterations remains comparable for γ in the range $\rho < \gamma < 1$ on the left figure but increases slightly as γ increases on the right figure. The results show a similar behaviour for (2.2) (dotted line) but the parameter ϵ seems to have a wider range for which the total iteration count is near minimum. In this regard, (2.2) might be preferred in implementations. However, unlike (2.3), this range is not explicitly determined by ρ.

5 Concluding remarks.

We have developed a convergence analysis for the inexact inverse iterations and completely characterized the convergence rate of the outer iteration in terms of the inner threshold parameter. Specifically, for the termination criterion (2.3), we have proved that the inexact inverse iteration converges linearly at the rate of $\max\{\gamma, \rho\}$. We have also shown that the total number of inner iterations required has an insensitive dependence on the inner threshold parameter, which leads to the practical strategy of choosing a γ that is in the range of $\rho \leq \gamma < 1$ and as small as possible.

While it has been shown that the termination criteria (2.2) and (2.3) are asymptotically equivalent, no explicit relation between γ and ϵ has been obtained and thus the present result does not provide a definite range for choosing ϵ in

(2.2). For future work, it would also be interesting to consider generalizations and implications of the present analysis to other methods that employ inexact shift-and-inverse transformations.

REFERENCES

1. W. E. Arnoldi, *The principle of minimized iterations in the solutions of the matrix eigenvalue problem*, Quart. Appl. Math., 9 (1951), pp. 17–29.
2. M. Crouzeix, B. Philippe, and M. Sadkane, *The Davidson method*, SIAM J. Sci. Stat. Comput., 15 (1994), pp. 62–76.
3. I. S. Duff, R. G. Grimes, and J. G. Lewis, *Sparse matrix test problems*, ACM Trans. Math. Software, 15 (1989), pp. 1–14.
4. G. H. Golub and C. F. Van Loan, *Matrix Computations*, The Johns Hopkins University Press, Baltimore, MD, 1983.
5. G. H. Golub, Z. Zhang, and H. Zha, *Large sparse symmetric eigenvalue problems with homogeneous linear constraints: The Lanczos process with inner-outer iterations*, Linear Algebra Appl., 309 (2000), pp. 289–306.
6. Y. Lai, K. Lin, and W. Lin, *An inexact inverse iteration for large sparse eigenvalue problems,* Numer. Linear Algebra Appl., 4 (1997), pp. 425–437.
7. C. Lanczos, *An iteration method for the solutions of the eigenvalue problem of linear differential and integral operators*, J. Res. Nat. Bur. Stand., 45 (1950), pp. 255–282.
8. R. Lehoucq and K. Meerbergen, *Using generalized Cayley transformations within an inexact rational Krylov sequence method,* SIAM J. Matrix Anal. Appl., 18 (1997), pp. 131–148.
9. K. Meerbergen and D. Roose, *The restarted Arnoldi method applied to iterative linear solvers for the computation of rightmost eigenvalues*, SIAM J. Matrix Anal. Appl., 20 (1999), pp. 1–20.
10. R. Morgan and D. Scott, *Preconditioning the Lanczos algorithm for sparse symmetric eigenvalue problems*, SIAM J. Sci. Stat. Comput. 14 (1993), pp. 585–593.
11. A. Ruhe, *Rational Krylov: A practical algorithm for large sparse nonsymmetric matrix pencils*, SIAM J. Sci. Stat. Comput., 19 (1998), pp. 1535–1551.
12. A. Ruhe and T. Wiberg, *The method of conjugate gradients used in inverse iteration*, BIT, 12 (1972), pp. 543–554.
13. Y. Saad, *Numerical Methods for Large Eigenvalue Problems*, Manchester University Press, Manchester, UK, 1992.
14. Y. Saad, and M. H. Schultz, *GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems*, SIAM J. Sci. Stat. Comput., 7 (1986), pp. 856–869.
15. G. Sleijpen and H. Van der Vorst, *A Jacobi–Davidson iteration method for linear eigenvalue problems*, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 401–425.
16. P. Smit and M. Paardekooper, *The effects of inexact solvers in algorithms for symmetric eigenvalue problems*, Linear Algebra Appl., 287 (1999), pp. 337–357.
17. D. Sorensen and C. Yang, *A truncated RQ iteration for large scale eigenvalue calculations*, SIAM J. Matrix Anal. Appl., 19 (1998), pp. 1045–1073.
18. A. Stathopoulos, Y. Saad, and C. Fisher, *Robust preconditioning of large sparse symmetric eigenvalue problems*, J. Comp. Appl. Math., 64 (1995), pp. 197–215.
19. G. W. Stewart, *Simultaneous iterations for computing invariant subspaces of non-Hermitian matrices*, Numer. Math., 25 (1976), pp. 123–136.