

## CONVERGENCE ANALYSIS OF INEXACT RAYLEIGH QUOTIENT ITERATION\*

YVAN NOTAY†

**Abstract.** We consider the computation of the smallest eigenvalue and associated eigenvector of a Hermitian positive definite pencil. Rayleigh quotient iteration (RQI) is known to converge cubically, and we first analyze how this convergence is affected when the arising linear systems are solved only approximately. We introduce a special measure of the relative error made in the solution of these systems and derive a sharp bound on the convergence factor of the eigenpair in a function of this quantity. This analysis holds independently of the way the linear systems are solved and applies to any type of error. For instance, it applies to rounding errors as well.

We next consider the Jacobi–Davidson method. It acts as an inexact RQI method in which the use of iterative solvers is made easier because the arising linear systems involve a projected matrix that is better conditioned than the shifted matrix arising in classical RQI. We show that our general convergence result straightforwardly applies in this context and permits us to trace the convergence of the eigenpair in a function of the number of inner iterations performed at each step. On this basis, we also compare this method with some form of inexact inverse iteration, as recently analyzed by Neymeyr and Knyazev.

**Key words.** eigenvalue, Rayleigh quotient, Jacobi–Davidson, preconditioning

**AMS subject classifications.** 65F10, 65B99, 65N20

**PII.** S0895479801399596

**1. Introduction.** We consider the computation of the smallest eigenvalue and associated eigenvector of a Hermitian positive definite pencil  $A - \lambda B$ .

In this context, the Rayleigh quotient iteration (RQI) method is known to converge very quickly, and cubically in the asymptotic phase [1, 15]. However, it requires solving at each step a system with the shifted matrix  $A - \theta B$ , with shift  $\theta$  equal to the Rayleigh quotient, i.e., changing from step to step. For large sparse matrices, this makes the use of direct solvers impractical, and, therefore, several works focus on the use of iterative solvers either by a direct approach [2, 19, 24] or indirectly via the use of the Jacobi–Davidson (JD) method [3, 14, 20, 21, 22, 23]. However, how an inexact solution may affect the convergence seems up to now not very well understood, despite the various analyses developed in these papers. The answer is actually far from obvious because, on the one hand, the systems to solve are very ill conditioned, and hence reducing the error measured with respect to any standard norm may involve a lot of numerical effort. On the other hand, it has been known for a long time from the error analysis made in connection with direct solvers that large errors in the computed solution do not necessarily spoil the convergence [16, 26].

In this paper, we first bring some new light on the actual convergence of inexact RQI. We introduce a special measure of the relative error made in the solution of the linear systems and bound the convergence factor of the eigenpair in a function of this quantity. Moreover, we show that the bound is sharp, indicating that the analysis takes the errors into proper account. This is further demonstrated by showing that

---

\*Received by the editors December 12, 2001; accepted for publication (in revised form) by H. van der Vorst July 16, 2002; published electronically January 17, 2003. This research was supported by the Fonds National de la Recherche Scientifique, Maître de recherches.

<http://www.siam.org/journals/simax/24-3/39959.html>

†Service de Métrologie Nucléaire, Université Libre de Bruxelles (C.P. 165/84), 50, Av. F.D. Roosevelt, B-1050 Brussels, Belgium (ynotay@ulb.ac.be).

our bound allows a straightforward analysis of the rounding errors arising with a backward stable direct solver when  $\theta$  is numerically equal to  $\lambda_1$ .

We next consider the JD method. Although it may be motivated in a different way (see [21]), it acts as an inexact RQI method and may even be seen as one of the easiest ways to implement robustly iterative solvers within RQI, since the ill-conditioned systems are not attacked directly (see the above references or section 4 for details). Here we show that our special measure of the error is equal to some standard relative error for the linear systems arising in the JD method. Hence our general convergence result straightforwardly applies, allowing us to trace the convergence of the eigenpair in a function of the number of *inner* iterations. This also allows some comparison with the predicted convergence of schemes based on inexact inverse iteration, as analyzed by Neymeyr [11, 12] and Knyazev and Neymeyr [9] (see [5, 10] for alternative analyses of inexact inverse iteration that, however, do not allow us to directly bound the convergence rate).

The remainder of the paper is organized as follows. In section 2, we recall some needed results on the convergence of RQI with exact solution of the arising linear systems. Our convergence analysis of inexact RQI is developed in section 3, and the JD method is discussed in section 4.

**Notation.** Throughout this paper,  $A$  and  $B$  are Hermitian  $n \times n$  matrices. We further assume that  $B$  is positive definite and that the smallest eigenvalue of the pencil  $A - \lambda B$  is simple. The eigenpairs are denoted  $(\lambda_i, \mathbf{u}_i)$ ,  $i = 1, \dots, n$ , with the eigenvalues ordered increasingly (i.e.,  $\lambda_1 < \lambda_2 \leq \dots \leq \lambda_n$ ) and the eigenvectors orthonormal with respect to the  $(\cdot, B \cdot)$  inner product (i.e.,  $(\mathbf{u}_i, B \mathbf{u}_j) = \delta_{ij}$ ).

For any symmetric and positive definite matrix  $C$ , we denote  $\|\cdot\|_C$  as the  $C$ -norm, that is, the norm associated to the  $(\cdot, C \cdot)$  inner product:  $\|\mathbf{v}\|_C = \sqrt{(\mathbf{v}, C \mathbf{v})}$  for all  $\mathbf{v}$ .

**2. Convergence of standard RQI.** Let first recall the basic algorithm: if  $\mathbf{u}$  is some approximate eigenvector and

$$(2.1) \quad \theta = \frac{(\mathbf{u}, A \mathbf{u})}{(\mathbf{u}, B \mathbf{u})}$$

is the associated Rayleigh quotient, the RQI method computes the next approximate eigenpair  $(\hat{\mathbf{u}}, \hat{\theta})$  as

$$(2.2) \quad \hat{\mathbf{u}} = (A - \theta B)^{-1} B \mathbf{u},$$

$$(2.3) \quad \hat{\theta} = \frac{(\hat{\mathbf{u}}, A \hat{\mathbf{u}})}{(\hat{\mathbf{u}}, B \hat{\mathbf{u}})}.$$

(In practice, some form of normalization is performed on  $\hat{\mathbf{u}}$ , but this does not matter for the discussion here.) Note that the RQI method favors the convergence toward the eigenvalue closest to  $\theta$ . Here we analyze the convergence toward the smallest eigenvalue, and thus we assume that

$$(2.4) \quad \theta < \frac{\lambda_1 + \lambda_2}{2},$$

which implies (since  $\theta$  cannot be smaller than  $\lambda_1$ ) that

$$(2.5) \quad \frac{\theta - \lambda_1}{\lambda_2 - \theta} \in [0, 1).$$

To assess the convergence, we introduce the decompositions

$$\begin{aligned} \mathbf{u} &= \|\mathbf{u}\|_B (\cos \varphi \mathbf{u}_1 + \sin \varphi \mathbf{v}), \\ \widehat{\mathbf{u}} &= \|\widehat{\mathbf{u}}\|_B (\cos \widehat{\varphi} \mathbf{u}_1 + \sin \widehat{\varphi} \widehat{\mathbf{v}}), \end{aligned}$$

where  $(\mathbf{v}, B \mathbf{u}_1) = (\widehat{\mathbf{v}}, B \mathbf{u}_1) = 0$  and  $\|\mathbf{v}\|_B = \|\widehat{\mathbf{v}}\|_B = 1$ . Then (see [15, p. 73])

$$(2.6) \quad \begin{aligned} \tan \widehat{\varphi} &= (\theta - \lambda_1) \|(A - \theta B)^{-1} B \mathbf{v}\|_B \tan \varphi \\ &\leq \frac{\theta - \lambda_1}{\lambda_2 - \theta} \tan \varphi, \end{aligned}$$

and the cubic convergence follows from  $(\theta - \lambda_1) = \mathcal{O}(\sin^2 \varphi)$ .

Now, to prove our main theorem, we need a sharp bound on  $\widehat{\theta}$ . This is obtained with Knyazev’s analysis as developed in [6, 7]. Indeed, particularizing [6, Theorem 2.3.1] to our context (see also [7, Theorem 2.5]), one gets

$$(2.7) \quad \frac{\widehat{\theta} - \lambda_1}{\lambda_2 - \widehat{\theta}} \leq \left( \frac{\theta - \lambda_1}{\lambda_2 - \theta} \right)^3,$$

which is simpler than (2.6) to work with.

Knyazev’s proof is general and elegant. (It covers a family of methods and not only the RQI method; see [13, Theorem 4.4] for an English translation.) However, for our analysis, we need to know in which cases the above bound is sharp. This can be seen by deriving (2.7) directly from (2.6). To this purpose, let

$$\eta = (\mathbf{v}, A \mathbf{v}), \quad \widehat{\eta} = (\widehat{\mathbf{v}}, A \widehat{\mathbf{v}})$$

be the Rayleigh quotients associated to  $\mathbf{v}$ ,  $\widehat{\mathbf{v}}$ , respectively (remember that  $\|\mathbf{v}\|_B = \|\widehat{\mathbf{v}}\|_B = 1$ ). Note that  $\widehat{\eta} \leq \eta$  because  $\widehat{\mathbf{v}}$  is the vector resulting from one step of the shift and invert iteration applied to  $\mathbf{v}$  with shift  $\theta$  smaller than the smallest eigenvalue for which  $\mathbf{v}$  has a nonzero component in the direction of the corresponding eigenvector. Since

$$\theta = \cos^2 \varphi \lambda_1 + \sin^2 \varphi \eta,$$

one has

$$(2.8) \quad \begin{aligned} \theta - \lambda_1 &= \sin^2 \varphi (\eta - \lambda_1), \\ \eta - \theta &= \cos^2 \varphi (\eta - \lambda_1), \end{aligned}$$

whence

$$(2.9) \quad \tan^2 \varphi = \frac{\theta - \lambda_1}{\eta - \theta},$$

and, similarly,

$$\tan^2 \widehat{\varphi} = \frac{\widehat{\theta} - \lambda_1}{\widehat{\eta} - \widehat{\theta}} \geq \frac{\widehat{\theta} - \lambda_1}{\eta - \widehat{\theta}}.$$

Inequality (2.6) therefore implies that (squaring both sides)

$$\widehat{\theta} - \lambda_1 \leq \frac{(\theta - \lambda_1)^3}{(\lambda_2 - \theta)^2} \frac{\eta - \widehat{\theta}}{\eta - \theta},$$

whence (2.7) because the last term of the right-hand side is a decreasing function of  $\eta \geq \lambda_2$ .

From these developments, one sees that the bound (2.7) is sharp when  $\mathbf{v} = \mathbf{u}_2$ , i.e., when  $\mathbf{u} \in \text{span}\{\mathbf{u}_1, \mathbf{u}_2\}$ ; then, (2.6) becomes indeed an equality, whereas one has  $\eta = \hat{\eta} = \lambda_2$ , entailing that equality is attained in (2.7). Note that, since asymptotically  $\mathbf{v}$  converges toward  $\mathbf{u}_2$ , it also means that the bound (2.7) gives the correct value of the asymptotic convergence factor.

Finally, observe that it is relevant to characterize the convergence by the ratio  $(\theta - \lambda_1)/(\lambda_2 - \theta)$  even when one is primarily interested in the accuracy of the eigenvector. Indeed, the  $B^{-1}$ -norm of the residual

$$(2.10) \quad \mathbf{r} = (A - \theta B) \mathbf{u}$$

satisfies

$$(2.11) \quad \frac{\|\mathbf{r}\|_{B^{-1}}^2}{\|\mathbf{u}\|_B^2} \geq (\theta - \lambda_1)(\lambda_2 - \theta)$$

[17, Lemma 3.2], whence, with (2.9),

$$(2.12) \quad \tan \varphi \leq \sqrt{\frac{\theta - \lambda_1}{\lambda_2 - \theta}} \leq \frac{1}{\lambda_2 - \theta} \frac{\|\mathbf{r}\|_{B^{-1}}}{\|\mathbf{u}\|_B}.$$

The convergence factor for the eigenvector is, however, only the square root of the one for the ratio  $(\theta - \lambda_1)/(\lambda_2 - \theta)$ . Note also that this ratio actually has to be made very small to satisfy a stopping criterion based on the residual norm.

**3. Convergence of inexact RQI.** Assume that some errors are introduced in the computation of  $\hat{\mathbf{u}} = (A - \theta B)^{-1} B \mathbf{u}$ . Let  $\tilde{\mathbf{u}}$  be the resulting vector. To analyze the influence on the convergence factor, we need a proper measure of these errors. The error vector  $\mathbf{x} = \hat{\mathbf{u}} - \tilde{\mathbf{u}}$  is by itself meaningless because the scaling of  $\tilde{\mathbf{u}}$  is unimportant. Among other possibilities, one may consider

$$\hat{\mathbf{u}} - \frac{(\hat{\mathbf{u}}, \mathbf{v})}{(\tilde{\mathbf{u}}, \mathbf{v})} \tilde{\mathbf{u}}$$

for some vector  $\mathbf{v}$  not orthogonal to  $\hat{\mathbf{u}}, \tilde{\mathbf{u}}$ . Somewhat arbitrarily, we select  $\mathbf{v} = B \mathbf{u}$ . Note, nevertheless, that  $(\hat{\mathbf{u}}, B \mathbf{u}) = 0$  is not possible because this would imply  $\hat{\theta} - \theta = \|\hat{\mathbf{u}}\|_B^{-1} (\hat{\mathbf{u}}, (A - \theta B) \hat{\mathbf{u}}) = 0$ , which contradicts (2.7). Accordingly, since  $\tilde{\mathbf{u}}$  approximates  $\hat{\mathbf{u}}$  at least in direction, it is not very restrictive to assume that  $(\tilde{\mathbf{u}}, B \mathbf{u}) \neq 0$ .

This choice leads us to characterize the error with

$$(3.1) \quad \mathbf{y} = \hat{\mathbf{u}} - \frac{(\hat{\mathbf{u}}, B \mathbf{u})}{(\tilde{\mathbf{u}}, B \mathbf{u})} \tilde{\mathbf{u}}$$

for which we have still to choose an appropriate norm. Here we state the following lemma, which is a straightforward generalization of [14, Lemma 3.1].

**LEMMA 3.1.** *Let  $A, B$  be  $n \times n$  Hermitian matrices. Assume that  $B$  is positive definite, and let  $\lambda_1 < \lambda_2 \leq \dots \leq \lambda_n$  be the eigenvalues of the pencil  $A - \lambda B$ . For any nonzero vector  $\mathbf{u}$ , one has*

$$\min_{\substack{\mathbf{z} \perp B \mathbf{u} \\ \mathbf{z} \neq \mathbf{0}}} \frac{(\mathbf{z}, (A - \theta B) \mathbf{z})}{(\mathbf{z}, B \mathbf{z})} \geq \lambda_1 + \lambda_2 - 2\theta,$$

where  $\theta = \frac{(\mathbf{u}, A \mathbf{u})}{(\mathbf{u}, B \mathbf{u})}$ .

Moreover, the bound is sharp: if  $\mathbf{u} \in \text{span}\{\mathbf{u}_1, \mathbf{u}_2\}$ , where  $\mathbf{u}_1, \mathbf{u}_2$  are eigenvectors associated to  $\lambda_1, \lambda_2$ , then (3.1) becomes an equality, the lower bound being attained for the vectors  $\mathbf{z}$  in the one-dimensional subspace  $\text{span}\{\mathbf{u}_1, \mathbf{u}_2\} \cap B\mathbf{u}^\perp$ .

Hence, when the condition (2.4) holds,  $A - \theta B$  is positive definite on  $B\mathbf{u}^\perp$ , and

$$(3.2) \quad \|\cdot\|_{A-\theta B} = \sqrt{(\cdot, (A - \theta B)\cdot)}$$

defines a particular (energy) norm on that subspace. Since  $\mathbf{y}$  belongs to that subspace, we may therefore use that norm, and we find that this makes the theoretical analysis easier.

Now, results are often better expressed in a function of relative errors. In this view, we compare the actual norm of  $\mathbf{y}$  with the norm one would obtain with  $\tilde{\mathbf{u}} = \mathbf{u}$ , that is, if no progress at all were made in the computation of the eigenpair. We thus propose to measure the errors introduced in the RQI process with the number

$$(3.3) \quad \gamma = \frac{\left\| \hat{\mathbf{u}} - \frac{(\hat{\mathbf{u}}, B \mathbf{u})}{(\hat{\mathbf{u}}, B \mathbf{u})} \tilde{\mathbf{u}} \right\|_{A-\theta B}}{\left\| \hat{\mathbf{u}} - \frac{(\hat{\mathbf{u}}, B \mathbf{u})}{(\hat{\mathbf{u}}, B \mathbf{u})} \mathbf{u} \right\|_{A-\theta B}}.$$

This looks somewhat unusual, but, as recalled in the introduction, standard measures of the error are often meaningless as far as the convergence of the eigenvector is concerned. Moreover, we shall see in the next section that this measure allows a straightforward analysis of the JD method.

We now state our main result.

**THEOREM 3.2.** *Let  $A, B$  be  $n \times n$  Hermitian matrices. Assume that  $B$  is positive definite, and let  $\lambda_1 < \lambda_2 \leq \dots \leq \lambda_n$  be the eigenvalues of the pencil  $A - \lambda B$ . Let  $\mathbf{u}$  be any nonzero vector such that*

$$\theta = \frac{(\mathbf{u}, A \mathbf{u})}{(\mathbf{u}, B \mathbf{u})}$$

satisfies

$$\theta < \frac{\lambda_1 + \lambda_2}{2}.$$

Let

$$\hat{\mathbf{u}} = (A - \theta B)^{-1} B \mathbf{u},$$

and let  $\tilde{\mathbf{u}}$  be a vector such that  $(\tilde{\mathbf{u}}, B \mathbf{u}) \neq 0$  and

$$\gamma = \frac{\left\| \hat{\mathbf{u}} - \frac{(\hat{\mathbf{u}}, B \mathbf{u})}{(\tilde{\mathbf{u}}, B \mathbf{u})} \tilde{\mathbf{u}} \right\|_{A-\theta B}}{\left\| \hat{\mathbf{u}} - \frac{(\hat{\mathbf{u}}, B \mathbf{u})}{(\tilde{\mathbf{u}}, B \mathbf{u})} \mathbf{u} \right\|_{A-\theta B}} \leq 1.$$

Then

$$\tilde{\theta} = \frac{(\tilde{\mathbf{u}}, A \tilde{\mathbf{u}})}{(\tilde{\mathbf{u}}, B \tilde{\mathbf{u}})}$$

satisfies

$$(3.4) \quad \frac{\tilde{\theta} - \lambda_1}{\lambda_2 - \tilde{\theta}} \leq \sigma^2 \frac{\theta - \lambda_1}{\lambda_2 - \theta},$$

where

$$(3.5) \quad \sigma = \frac{(\theta - \lambda_1) + \gamma(\lambda_2 - \theta)}{(\lambda_2 - \theta) + \gamma(\theta - \lambda_1)}.$$

Moreover, the bound is sharp: let  $\mathbf{u}_1, \mathbf{u}_2$  be eigenvectors associated to  $\lambda_1, \lambda_2$ ; if  $\mathbf{u} \in \text{span}\{\mathbf{u}_1, \mathbf{u}_2\}$ , then, for all  $\mathbf{x} \in \text{span}\{\mathbf{u}_1, \mathbf{u}_2\}$  such that  $|(\mathbf{x}, B\mathbf{u})| < |(\hat{\mathbf{u}}, B\mathbf{u})|$ , one has that equality is attained in (3.4) for either  $\tilde{\mathbf{u}} = \hat{\mathbf{u}} + \mathbf{x}$  or  $\tilde{\mathbf{u}} = \hat{\mathbf{u}} - \mathbf{x}$ .

*Proof.* Assume (without loss of generality) that  $\|\mathbf{u}\|_B = 1$ , let  $\hat{\theta}$  be defined by (2.3) and  $\mathbf{y}$  by (3.1), and let

$$\hat{\delta} = \theta - \hat{\theta}, \quad \tilde{\delta} = \theta - \tilde{\theta}.$$

First observe that  $(\hat{\mathbf{u}}, (A - \theta B)\hat{\mathbf{u}}) = (\hat{\mathbf{u}}, B\mathbf{u})$ , that  $(\hat{\mathbf{u}}, (A - \theta B)\mathbf{u}) = \|\mathbf{u}\|_B = 1$ , and that  $(\mathbf{u}, (A - \theta B)\mathbf{u}) = 0$ . Hence

$$(3.6) \quad \|\hat{\mathbf{u}} - (\hat{\mathbf{u}}, B\mathbf{u})\mathbf{u}\|_{A-\theta B}^2 = -(\hat{\mathbf{u}}, B\mathbf{u}).$$

Further,

$$(3.7) \quad \hat{\delta} = -\frac{(\hat{\mathbf{u}}, (A - \theta B)\hat{\mathbf{u}})}{(\hat{\mathbf{u}}, B\hat{\mathbf{u}})} = -\frac{(\hat{\mathbf{u}}, B\mathbf{u})}{\|\hat{\mathbf{u}}\|_B^2},$$

and, since  $(\mathbf{y}, (A - \theta B)\hat{\mathbf{u}}) = (\mathbf{y}, B\mathbf{u}) = 0$ ,

$$(3.8) \quad \begin{aligned} \tilde{\delta} &= -\frac{(\tilde{\mathbf{u}}, (A - \theta B)\tilde{\mathbf{u}})}{(\tilde{\mathbf{u}}, B\tilde{\mathbf{u}})} \\ &= -\frac{((\hat{\mathbf{u}} + \mathbf{y}), (A - \theta B)(\hat{\mathbf{u}} + \mathbf{y}))}{\|\hat{\mathbf{u}} + \mathbf{y}\|_B^2} \\ &= -\frac{(\hat{\mathbf{u}}, (A - \theta B)\hat{\mathbf{u}}) + \|\mathbf{y}\|_{A-\theta B}^2}{\|\hat{\mathbf{u}} + \mathbf{y}\|_B^2} \\ &= -(1 - \gamma^2) \frac{(\hat{\mathbf{u}}, B\mathbf{u})}{\|\hat{\mathbf{u}} + \mathbf{y}\|_B^2}. \end{aligned}$$

On the other hand, consider the projector  $P = I - \mathbf{u}(B\mathbf{u})^*$ . Observe that  $(\mathbf{v}, B P \mathbf{w}) = (P \mathbf{v}, B \mathbf{w})$  for all  $\mathbf{v}, \mathbf{w}$ , i.e., that  $P$  is orthogonal with respect to the  $(\cdot, B \cdot)$  inner product. Since  $P \mathbf{y} = \mathbf{y}$ , one has

$$(3.9) \quad \begin{aligned} \|\hat{\mathbf{u}} + \mathbf{y}\|_B^2 &= \|(I - P)\hat{\mathbf{u}}\|_B^2 + \|P\hat{\mathbf{u}} + \mathbf{y}\|_B^2 \\ &\leq \|(I - P)\hat{\mathbf{u}}\|_B^2 + \|P\hat{\mathbf{u}}\|_B^2 + \|\mathbf{y}\|_B^2 + 2\|P\hat{\mathbf{u}}\|_B \|\mathbf{y}\|_B \\ &= \|\hat{\mathbf{u}}\|_B^2 + \|\mathbf{y}\|_B^2 + 2\|P\hat{\mathbf{u}}\|_B \|\mathbf{y}\|_B. \end{aligned}$$

Hence, using Lemma 3.1,

$$(3.10) \quad \begin{aligned} \|\hat{\mathbf{u}} + \mathbf{y}\|_B^2 &\leq \|\hat{\mathbf{u}}\|_B^2 + \frac{1}{\lambda_1 + \lambda_2 - 2\theta} (\|\mathbf{y}\|_{A-\theta B}^2 + 2\|P\hat{\mathbf{u}}\|_{A-\theta B} \|\mathbf{y}\|_{A-\theta B}) \\ &= \|\hat{\mathbf{u}}\|_B^2 - (\gamma^2 + 2\gamma) \frac{(\hat{\mathbf{u}}, B\mathbf{u})}{\lambda_1 + \lambda_2 - 2\theta}, \end{aligned}$$

and, therefore, with (3.7), (3.8),

$$\tilde{\delta} \geq \frac{1 - \gamma^2}{\widehat{\delta}^{-1} + \frac{\gamma^2 + 2\gamma}{\lambda_1 + \lambda_2 - 2\theta}}.$$

We now use (2.7) to bound  $\widehat{\delta}$ . With  $\beta = (\theta - \lambda_1)/(\lambda_2 - \theta)$ , the latter inequality may be rewritten as

$$\frac{\theta - \lambda_1 - \widehat{\delta}}{\beta^{-1}(\theta - \lambda_1) + \widehat{\delta}} \leq \beta^3,$$

i.e.,

$$\widehat{\delta} \geq (\theta - \lambda_1) \frac{1 - \beta^2}{1 + \beta^3} = (\theta - \lambda_1) \frac{1 - \beta}{1 - \beta + \beta^2}.$$

We thus have, since  $\lambda_1 + \lambda_2 - 2\theta = (\theta - \lambda_1)(\beta^{-1} - 1)$ ,

$$\tilde{\delta} \geq (\theta - \lambda_1) \frac{(1 - \gamma^2)(1 - \beta)}{1 - \beta + \beta^2 + \beta(\gamma^2 + 2\gamma)},$$

whence, letting  $D = 1 - \beta + \beta^2 + \beta(\gamma^2 + 2\gamma)$ ,

$$\begin{aligned} \tilde{\theta} - \lambda_1 &= \theta - \lambda_1 - \tilde{\delta} \leq \frac{\theta - \lambda_1}{D} (D - (1 - \gamma^2)(1 - \beta)) \\ &= \frac{\theta - \lambda_1}{D} (\beta + \gamma)^2 \end{aligned}$$

and

$$\begin{aligned} \lambda_2 - \tilde{\theta} &= \lambda_2 - \theta + \tilde{\delta} \geq \frac{\lambda_2 - \theta}{D} (D + \beta(1 - \gamma^2)(1 - \beta)) \\ &= \frac{\lambda_2 - \theta}{D} (1 + \beta\gamma)^2. \end{aligned}$$

Therefore, (3.4) holds with

$$\sigma = \frac{\beta + \gamma}{1 + \beta\gamma} = \frac{(\theta - \lambda_1) + \gamma(\lambda_2 - \theta)}{(\lambda_2 - \theta) + \gamma(\theta - \lambda_1)}.$$

To prove the sharpness, first observe that the only inequalities used in the proof are (2.7), (3.9), (3.10). Moreover, it has already been noted in section 2 that (2.7) becomes an equality when  $\mathbf{u} \in \text{span}\{\mathbf{u}_1, \mathbf{u}_2\}$ . On the other hand, under the given assumptions, both  $P\widehat{\mathbf{u}}$  and  $\mathbf{y}$  belong to the one-dimensional subspace  $\text{span}\{\mathbf{u}_1, \mathbf{u}_2\} \cap B\mathbf{u}^\perp$ . For this subspace, it is shown in Lemma 3.1 that the inequality used to obtain (3.10) from (3.9) is actually also an equality, whereas, since  $P\widehat{\mathbf{u}}$  and  $\mathbf{y}$  are aligned, one has necessarily

$$|(P\widehat{\mathbf{u}}, B\mathbf{y})| = \|P\widehat{\mathbf{u}}\|_B \|\mathbf{y}\|_B;$$

i.e., (3.9) becomes an equality too if and only if  $(P\widehat{\mathbf{u}}, B\mathbf{y})$  is positive. Let then  $\widetilde{\mathbf{u}} = \widehat{\mathbf{u}} + c\mathbf{x}$ , where  $c$  equals either 1 or  $-1$ . One finds

$$\mathbf{y} = \frac{c}{1 + c \frac{(\mathbf{x}, B\mathbf{u})}{(\widehat{\mathbf{u}}, B\mathbf{u})}} \left( \frac{(\mathbf{x}, B\mathbf{u})}{(\widehat{\mathbf{u}}, B\mathbf{u})} \widehat{\mathbf{u}} - \mathbf{x} \right),$$

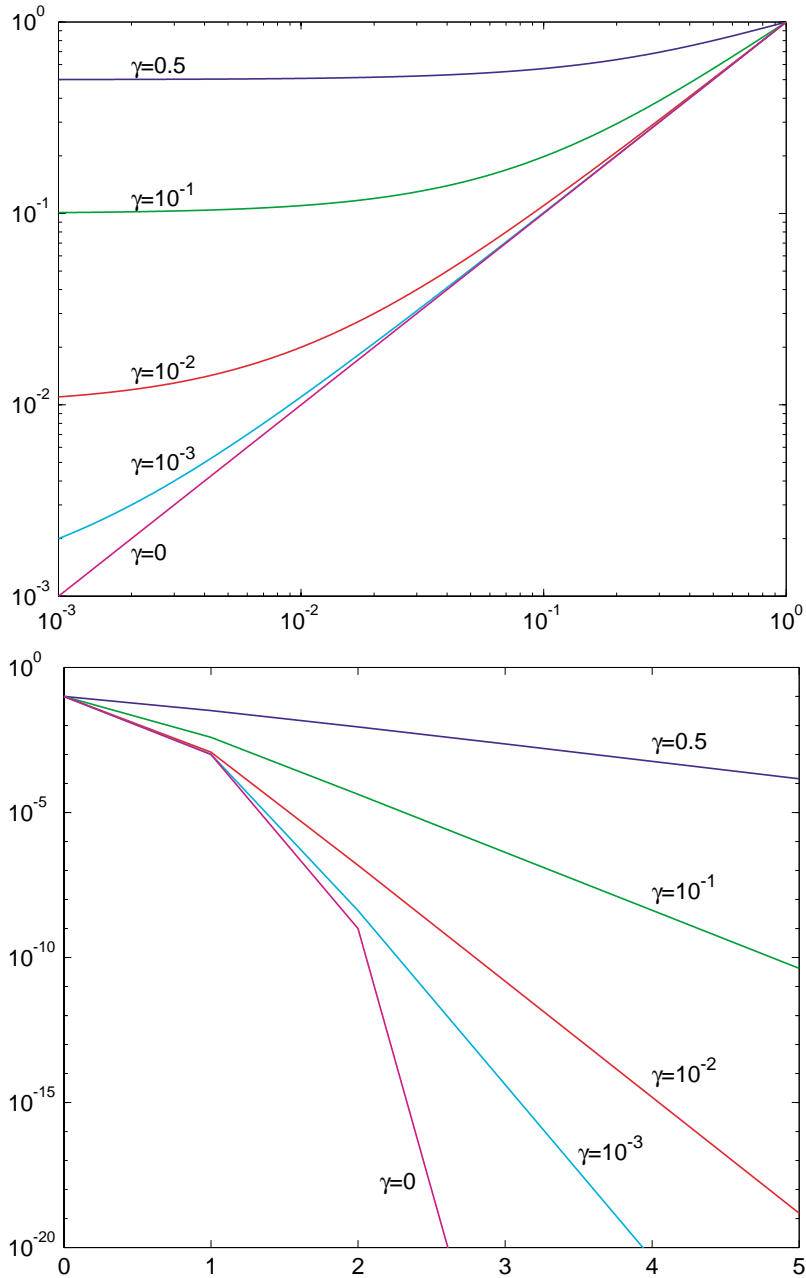


FIG. 1.  $\sigma$  versus  $(\theta - \lambda_1)/(\lambda_2 - \theta)$  (top) and evolution of  $(\theta - \lambda_1)/(\lambda_2 - \theta)$  in a function of the number of steps (bottom).

showing that, since  $|(\mathbf{x}, B\mathbf{u})| < |(\hat{\mathbf{u}}, B\mathbf{u})|$ , one can always, by choosing appropriately the sign of  $c$ , select the direction of  $\mathbf{y}$  in such a way that  $(P\hat{\mathbf{u}}, B\mathbf{y}) > 0$  holds.  $\square$

Observe that the sharpness of the bound is not proved only for one special orientation of the error vector but that it holds for a two-dimensional subspace that includes both vectors aligned with  $\mathbf{u}_1$  and vectors orthogonal to it.

To illustrate our result, we have plotted on Figure 1 (top) the convergence factor  $\sigma$

against  $(\theta - \lambda_1)/(\lambda_2 - \theta)$  for several values of  $\gamma$ . One sees that  $\sigma \rightarrow \gamma$  when  $\theta \rightarrow \lambda_1$ ; more precisely, one has

$$(3.11) \quad \sigma \approx \gamma \quad \text{when} \quad \frac{\theta - \lambda_1}{\lambda_2 - \theta} \ll \gamma.$$

On this figure (bottom), we also display the evolution of  $(\theta - \lambda_1)/(\lambda_2 - \theta)$  in a function of the number of RQI steps. (More precisely, we display it for the worst-case scenario, according to the bound (3.4).) One sees that it is not necessary to make the errors very small to essentially preserve the cubic convergence rate.

**Practical estimation of  $\gamma$ .** In practical situations, one generally does not have access to the exact value of  $\gamma$ . Nevertheless, some estimate can be obtained, based on the following reasoning. The situation is essentially similar to the one met in the context of the solution of Hermitian positive definite linear systems: theoretical results are expressed in a function of the energy norm of the error, which is not available in practical computations. However, one is generally satisfied with the computation of the residual norm, because it expresses the same error with respect to a different but equivalent norm, and in practice it most often happens that, on the whole, both measures of the error evolve similarly.

Here we want to follow the same approach, but we need to be careful because (3.2) defines a norm only on a particular subspace. Let then

$$(3.12) \quad P = I - \mathbf{u}(\mathbf{u}, B\mathbf{u})^{-1}(B\mathbf{u})^*$$

be the projector with range  $B\mathbf{u}^\perp$  and kernel  $\text{span}\{\mathbf{u}\}$ , and note that  $P^*(A - \theta B)P$  is Hermitian with range  $B\mathbf{u}^\perp$ . Hence, the pencil  $P^*(A - \theta B)P - \lambda B$  possesses  $n - 1$  eigenvectors forming a  $B$ -orthonormal basis of  $B\mathbf{u}^\perp$  and whose corresponding eigenvalues are, by Lemma 3.1, not smaller than  $\lambda_1 + \lambda_2 - 2\theta$  and not larger than  $\lambda_n - \theta$ . Therefore, by expanding  $\mathbf{v} \in B\mathbf{u}^\perp$  on this basis, one obtains, since  $P\mathbf{v} = \mathbf{v}$  and thus  $\|\mathbf{v}\|_{A-\theta B} = \|\mathbf{v}\|_{P^*(A-\theta B)P}$ ,

$$(3.13) \quad \alpha_1 \|\mathbf{v}\|_{A-\theta B} \leq \|P^*(A - \theta B)\mathbf{v}\|_{B^{-1}} \leq \alpha_2 \|\mathbf{v}\|_{A-\theta B},$$

where  $\alpha_1 = \sqrt{\lambda_1 + \lambda_2 - 2\theta}$  and  $\alpha_2 = \sqrt{\lambda_n - \theta}$ .

On the other hand,

$$\begin{aligned} P^*(A - \theta B) \left( \hat{\mathbf{u}} - \frac{(\hat{\mathbf{u}}, B\mathbf{u})}{(\tilde{\mathbf{u}}, B\mathbf{u})} \tilde{\mathbf{u}} \right) &= P^* \left( B\mathbf{u} - \frac{(\hat{\mathbf{u}}, B\mathbf{u})}{(\tilde{\mathbf{u}}, B\mathbf{u})} (A - \theta B) \tilde{\mathbf{u}} \right) \\ &= -\frac{(\hat{\mathbf{u}}, B\mathbf{u})}{(\tilde{\mathbf{u}}, B\mathbf{u})} P^*(A - \theta B) \tilde{\mathbf{u}} \\ &= \frac{(\hat{\mathbf{u}}, B\mathbf{u})}{(\tilde{\mathbf{u}}, B\mathbf{u})} P^* \mathbf{g}, \end{aligned}$$

where

$$(3.14) \quad \mathbf{g} = B\mathbf{u} - (A - \theta B) \tilde{\mathbf{u}}$$

is the residual of the linear system solved within the RQI process. Similarly, one finds

$$\begin{aligned} P^*(A - \theta B) \left( \hat{\mathbf{u}} - \frac{(\hat{\mathbf{u}}, B\mathbf{u})}{(\mathbf{u}, B\mathbf{u})} \mathbf{u} \right) &= -\frac{(\hat{\mathbf{u}}, B\mathbf{u})}{(\mathbf{u}, B\mathbf{u})} P^*(A - \theta B) \mathbf{u} \\ &= -\frac{(\hat{\mathbf{u}}, B\mathbf{u})}{(\mathbf{u}, B\mathbf{u})} \mathbf{r}, \end{aligned}$$

where  $\mathbf{r}$  is the current residual of the eigenproblem (2.10).

Hence, with (3.13),

$$(3.15) \quad \alpha^{-1} \tilde{\gamma} \leq \gamma \leq \alpha \tilde{\gamma},$$

where  $\alpha = \sqrt{\frac{\lambda_n - \theta}{\lambda_1 + \lambda_2 - 2\theta}}$  and where

$$(3.16) \quad \tilde{\gamma} = \frac{(\mathbf{u}, B \mathbf{u})}{|(\tilde{\mathbf{u}}, B \mathbf{u})|} \frac{\|P^*(A - \theta B) \tilde{\mathbf{u}}\|_{B^{-1}}}{\|(A - \theta B) \mathbf{u}\|_{B^{-1}}} = \frac{(\mathbf{u}, B \mathbf{u})}{|(\tilde{\mathbf{u}}, B \mathbf{u})|} \frac{\|P^* \mathbf{g}\|_{B^{-1}}}{\|\mathbf{r}\|_{B^{-1}}}.$$

**Error analysis for  $\theta \rightarrow \lambda_1$ .** For the sake of simplicity, here we confine ourselves to standard eigenproblems ( $B = I$ ).

In the final phase of the process, one reaches  $\theta = \lambda_1$  up to machine accuracy before the eigenvector has converged (see (2.12)). Some subtle reasoning is then needed to show that the ill-conditioning of  $A - \theta I$  does not prevent further progress despite that the computed solution to  $(A - \theta I) \mathbf{v} = \mathbf{u}$  cannot be accurate even with a backward stable direct solver [16, 26].

Here our results offer a straightforward way to prove that one more step is then enough to compute an accurate eigenvector. Indeed, with a backward stable direct solver, the computed solution  $\tilde{\mathbf{u}}$  is such that

$$\|\mathbf{g}\| = \|\mathbf{u} - (A - \theta I) \tilde{\mathbf{u}}\| \leq c \varepsilon_{\text{mach}} \|A\| \|\tilde{\mathbf{u}}\|.$$

Hence, since  $P$  is orthogonal (thus  $\|P \mathbf{g}\| \leq \|\mathbf{g}\|$ ) and using (2.11),

$$\begin{aligned} \tilde{\gamma} &\leq c \varepsilon_{\text{mach}} \frac{\|\tilde{\mathbf{u}}\| \|\mathbf{u}\|}{|(\tilde{\mathbf{u}}, \mathbf{u})|} \frac{\|A\| \|\mathbf{u}\|}{\|\mathbf{r}\|} \\ &\leq \frac{c \varepsilon_{\text{mach}}}{\cos(\tilde{\mathbf{u}}, \mathbf{u})} \frac{\|A\|}{\sqrt{(\theta - \lambda_1)(\lambda_2 - \theta)}}. \end{aligned}$$

Thus

$$\sigma \leq \frac{\theta - \lambda_1}{\lambda_2 - \theta} + \alpha \frac{c \varepsilon_{\text{mach}}}{\cos(\tilde{\mathbf{u}}, \mathbf{u})} \frac{\|A\|}{\sqrt{(\theta - \lambda_1)(\lambda_2 - \theta)}},$$

whence

$$\tan(\tilde{\mathbf{u}}, \mathbf{u}_1) \leq \sqrt{\frac{\tilde{\theta} - \lambda_1}{\lambda_2 - \tilde{\theta}}} \leq \left(\frac{\theta - \lambda_1}{\lambda_2 - \theta}\right)^{3/2} + \alpha \frac{c \varepsilon_{\text{mach}}}{\cos(\tilde{\mathbf{u}}, \mathbf{u})} \frac{\|A\|}{\lambda_2 - \theta};$$

that is, for  $\theta \rightarrow \lambda_1$ ,

$$(3.17) \quad \tan(\tilde{\mathbf{u}}, \mathbf{u}_1) \leq \alpha \frac{c \varepsilon_{\text{mach}}}{\cos(\tilde{\mathbf{u}}, \mathbf{u})} \frac{\|A\|}{\lambda_2 - \lambda_1},$$

which is not far from the best attainable accuracy; see [26, pp. 69–70] (note that  $\tan(\tilde{\mathbf{u}}, \mathbf{u}) = \mathcal{O}(\tan(\mathbf{u}, \mathbf{u}_1)) = \mathcal{O}(\sqrt{\theta - \lambda_1})$ ).

**4. Convergence of the JD method.** The JD method [3, 14, 20, 21, 22, 23] combines some form of inexact RQI with a Galerkin approach.

Let  $\mathbf{u}$  be the current approximate eigenvector which we assume is normalized with respect to the  $B$ -norm. With this method, one first computes a correction  $\mathbf{t}$  orthogonal to  $B\mathbf{u}^\perp$  by solving (approximately) the so-called *correction equation*

$$(4.1) \quad P^*(A - \tilde{\lambda}B)P\mathbf{t} = -\mathbf{r}; \quad (\mathbf{t}, B\mathbf{u}) = 0,$$

where  $\tilde{\lambda}$  is an approximation of the “target” eigenvalue, where  $\mathbf{r}$  is the residual (2.10), and where

$$P = I - \mathbf{u}(B\mathbf{u})^*$$

is the projector (3.12). Next, one applies the Galerkin principle: the initial approximation and the successive corrections are gathered to form the basis of a subspace from which one extracts the best approximation of the searched eigenpair by the Rayleigh–Ritz procedure (see, e.g., [22] for algorithmic details).

The exact solution to (4.1) is

$$(4.2) \quad \hat{\mathbf{t}} = \frac{1}{(B\mathbf{u}, (A - \tilde{\lambda}B)^{-1}B\mathbf{u})} (A - \tilde{\lambda}B)^{-1}B\mathbf{u} - \mathbf{u},$$

and hence the equivalence with RQI is recovered if one has used  $\tilde{\lambda} = \theta$  and if the next approximate eigenvector is  $\mathbf{u} + \hat{\mathbf{t}}$ .

In practice, one does not always select  $\tilde{\lambda} = \theta$ . One first sets  $\tilde{\lambda}$  equal to some fixed target, for instance,  $\tilde{\lambda} = 0$ , if one searches for the smallest eigenvalue of a Hermitian positive definite eigenproblem;  $\tilde{\lambda} = \theta$  is then used when, according to some heuristic criterion, one detects that  $\theta$  entered its final interval (see [14, 25] for examples of such criteria). Here we confine ourselves to this final phase.

On the other hand, the next approximate eigenvector resulting from the Rayleigh–Ritz procedure is generally not equal to  $\mathbf{u} + \mathbf{t}$ . However, in the context considered here, this procedure selects the vector from the subspace for which the associated Rayleigh quotient is minimal. Hence the convergence as measured through the evolution of the ratio  $(\theta - \lambda_1)/(\lambda_2 - \theta)$  can only be improved by this approach. Note, however, that little improvement is expected in the final phase, at least if the correction equation is solved sufficiently accurately: RQI converges then so quickly that one hardly accelerates it. Actually, the Galerkin approach is mainly useful in the first phase, to bring  $\theta$  into its final interval quickly and to avoid misconvergence if one has selected  $\tilde{\lambda} = \theta$  too early.

We thus continue assuming that one has selected  $\tilde{\lambda} = \theta < (\lambda_1 + \lambda_2)/2$ , and we bound the convergence factor by analyzing the Rayleigh quotient associated to

$$\tilde{\mathbf{u}} = \mathbf{u} + \tilde{\mathbf{t}},$$

where  $\tilde{\mathbf{t}}$  is the computed approximate solution to (4.1). Note that  $(\tilde{\mathbf{t}}, B\mathbf{u}) = 0$ , whence  $(\tilde{\mathbf{u}}, B\mathbf{u}) = (\mathbf{u}, B\mathbf{u}) = 1$ . Thus, since, by (4.2),

$$\hat{\mathbf{u}} = (\hat{\mathbf{u}}, B\mathbf{u}) (\hat{\mathbf{t}} + \mathbf{u}),$$

one has

$$\hat{\mathbf{u}} - \frac{(\hat{\mathbf{u}}, B\mathbf{u})}{(\tilde{\mathbf{u}}, B\mathbf{u})} \tilde{\mathbf{u}} = (\hat{\mathbf{u}}, B\mathbf{u}) (\hat{\mathbf{t}} - \tilde{\mathbf{t}}),$$

whereas

$$\widehat{\mathbf{u}} - \frac{(\widehat{\mathbf{u}}, B \mathbf{u})}{(\mathbf{u}, B \mathbf{u})} \mathbf{u} = (\widehat{\mathbf{u}}, B \mathbf{u}) \widehat{\mathbf{t}}.$$

Hence, Theorem 3.2 applies with

$$(4.3) \quad \gamma = \frac{\|\widehat{\mathbf{t}} - \widetilde{\mathbf{t}}\|_{A-\theta B}}{\|\widehat{\mathbf{t}}\|_{A-\theta B}};$$

i.e.,  $\gamma$  is here the relative error in the correction equation (4.1) measured with respect to the standard energy norm for that equation (remember that, for all  $\mathbf{v} \in B\mathbf{u}^\perp$ , one has  $P\mathbf{v} = \mathbf{v}$ , and therefore  $\|\mathbf{v}\|_{A-\theta B} = \|\mathbf{v}\|_{P^*(A-\theta B)P}$ ).

Now, since the correction equation is positive definite on  $B\mathbf{u}^\perp$ , one may solve it with the preconditioned conjugate gradient (PCG) method, as advised in [14]. Then  $\gamma$  may be directly bounded in a function of the number  $k$  of inner iterations [4, 18]: with the zero initial guess, one has

$$(4.4) \quad \gamma \leq 2 \left( \left( \frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} \right)^k + \left( \frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} \right)^{-k} \right)^{-1},$$

where  $\kappa$  is the spectral condition number.

Considering (3.11) again, one will achieve  $\sigma \approx \gamma$  if one is wise enough to stop inner iterations before  $\gamma$  becomes too small, so that further progress is useless. One then recovers our main conclusion from [14], where we analyze the evolution of the residual norm: with a proper stopping criterion, the convergence of the eigenvector goes along with that of the successive linear systems, and the main additional cost to compute the eigenpair compared with a mere linear system solution comes from the need to periodically restart the linear solver.

Now, one may wonder about the value of  $\kappa$  for such a projected system. To simplify the discussion, we assume here (and throughout the paper) that  $A$  is positive definite. (The general case is easily recovered with the shift transformation  $A - \lambda B \rightarrow (A + \tau B) - (\lambda + \tau) B$ .) We also recall that it is not advised to try to directly precondition the projected matrix (which is dense) nor even the shifted matrix  $A - \theta B$  (which is indefinite). Instead, set up your favorite (positive definite) preconditioner  $K$  for  $A$ , and precondition  $P^*(A - \theta B)P$  with

$$M = P^* K P.$$

Note that the singularity of  $M$  raises no practical difficulty; see [23]. Moreover, out of the four projection steps associated with the system matrix and the preconditioner, only one needs to be performed in practice; see [3, 14] for details.

Now, with such a preconditioner,

$$\kappa = \frac{\max_{\substack{\mathbf{z} \perp B\mathbf{u} \\ \mathbf{z} \neq 0}} \frac{(\mathbf{z}, (A - \theta B) \mathbf{z})}{(\mathbf{z}, K \mathbf{z})}}{\min_{\substack{\mathbf{z} \perp B\mathbf{u} \\ \mathbf{z} \neq 0}} \frac{(\mathbf{z}, (A - \theta B) \mathbf{z})}{(\mathbf{z}, K \mathbf{z})}}$$

may be bounded in a function of

$$\kappa(K^{-1}A) = \frac{\lambda_{\max}(K^{-1}A)}{\lambda_{\min}(K^{-1}A)}.$$

Indeed, one has (see also [14])

$$\begin{aligned} \max_{\substack{\mathbf{z} \perp_B \mathbf{u} \\ \mathbf{z} \neq 0}} \frac{(\mathbf{z}, (A - \theta B) \mathbf{z})}{(\mathbf{z}, K \mathbf{z})} &\leq \max_{\mathbf{z} \neq 0} \frac{(\mathbf{z}, A \mathbf{z})}{(\mathbf{z}, K \mathbf{z})} \\ &= \lambda_{\max}(K^{-1}A), \end{aligned}$$

whereas, with Lemma 3.1,

$$\begin{aligned} \min_{\substack{\mathbf{z} \perp_B \mathbf{u} \\ \mathbf{z} \neq 0}} \frac{(\mathbf{z}, (A - \theta B) \mathbf{z})}{(\mathbf{z}, K \mathbf{z})} &= \min_{\substack{\mathbf{z} \perp_B \mathbf{u} \\ \mathbf{z} \neq 0}} \left( \frac{(\mathbf{z}, A \mathbf{z})}{(\mathbf{z}, K \mathbf{z})} \left( 1 + \frac{\theta (\mathbf{z}, B \mathbf{z})}{(\mathbf{z}, (A - \theta B) \mathbf{z})} \right)^{-1} \right) \\ &\geq \lambda_{\min}(K^{-1}A) \left( 1 + \frac{\theta}{\lambda_1 + \lambda_2 - 2\theta} \right)^{-1}. \end{aligned}$$

Therefore,

$$(4.5) \quad \kappa \leq \kappa(K^{-1}A) \left( 1 + \frac{\theta}{\lambda_1 + \lambda_2 - 2\theta} \right),$$

which, together with (4.4), allows us to bound  $\gamma$  and thus  $\sigma$  in a function of  $k$ ,  $\kappa(K^{-1}A)$ ,  $\theta$ ,  $\lambda_1$ , and  $\lambda_2$ .

Concerning the estimation of  $\gamma$  through  $\tilde{\gamma}$  (3.16), note that

$$\begin{aligned} P^* \mathbf{g} &= P^* (B \mathbf{u} - (A - \theta B)(\mathbf{u} + \tilde{\mathbf{t}})) \\ &= -\mathbf{r} - P^*(A - \theta B) \tilde{\mathbf{t}} \\ &= \mathbf{g}_{\text{ce}}, \end{aligned}$$

where  $\mathbf{g}_{\text{ce}}$  is the residual in the correction equation (4.1). Hence, since  $(\tilde{\mathbf{u}}, B \mathbf{u}) = (\mathbf{u}, B \mathbf{u})$ ,

$$(4.6) \quad \tilde{\gamma} = \frac{\|\mathbf{g}_{\text{ce}}\|_{B^{-1}}}{\|\mathbf{r}\|_{B^{-1}}},$$

which confirms that  $\tilde{\gamma}$  expresses the same error as  $\gamma$  but with respect to the residual norm instead of the energy norm.

**Convergence of inexact inverse iteration.** It is interesting to compare the above result with the convergence analysis of inexact inverse iteration as developed in [9, 11, 12]. Indeed, if schemes based on the RQI method are expected to converge faster in general, it does not mean that they are always more cost effective. On the other hand, the results in these papers also offer the best bounds to date for schemes based on nonlinear conjugate gradients as developed in, e.g., [8]. Thus the comparison may also give some insight into how the JD method compares with such methods.

Let first recall that inexact (or “preconditioned”) inverse iteration also sets

$$\tilde{\mathbf{u}} = \mathbf{u} + \tilde{\mathbf{t}},$$

but here  $\tilde{\mathbf{t}}$  is obtained by solving approximately

$$(4.7) \quad A \mathbf{t} = -\mathbf{r}.$$

Stricto sensu, the analysis in [9, 11, 12] covers only the case in which  $\tilde{\mathbf{t}} = -K^{-1}\mathbf{r}$  for some positive definite preconditioner  $K$ . However, looking closely at Lemma 2.1

in [11] (which is the root of everything else), it clearly turns out that the main results also apply when several inner iterations are performed, with parameter  $\gamma$  equal to the relative error in (4.7) measured with respect to the energy norm, i.e.,

$$\gamma = \frac{\|\tilde{\mathbf{t}} + A^{-1}\mathbf{r}\|_A}{\|A^{-1}\mathbf{r}\|_A}.$$

The main result is precisely a bound similar to (3.4) with convergence factor

$$(4.8) \quad \sigma \leq \bar{\sigma}(\gamma, \theta) \leq \bar{\sigma}(\gamma, \lambda_1) = 1 - (1 - \gamma) \left(1 - \frac{\lambda_1}{\lambda_2}\right).$$

The first inequality is sharp for any  $\theta$  and is due to Neymeyr [11, 12]; the resulting expression for  $\bar{\sigma}(\gamma, \theta)$ , however, is so complicated that it is not interesting to reproduce it here. The second inequality is due to Knyazev and Neymeyr [9] and actually gives a good approximation of the first one (see the figures).

**Illustration and comparison.** For both the JD method and inexact inverse iteration, we are able to bound the convergence factor  $\sigma$  in a function of the number of inner iterations  $k$ , given  $\kappa(K^{-1}A)$ ,  $\theta$ ,  $\lambda_1$ , and  $\lambda_2$ . Let  $\bar{\sigma}(k, \theta)$  be the resulting bound. The most interesting quantity is  $\bar{\sigma}^{1/k}(k, \theta)$ , which represents the convergence in the outer process *per inner iteration*. We have plotted it on Figure 2 against  $(\theta - \lambda_1)/(\lambda_2 - \theta)$ . We consider different values of  $k$  for the JD method, but, to keep the figures readable, for inverse iteration we display  $\bar{\sigma}^{1/k}(k, \theta)$  only for the value of  $k$  that minimizes  $\bar{\sigma}^{1/k}(k, \lambda_1)$ .

One sees that, for the JD method, the optimal value of  $k$  depends on  $\theta$ , that is, on how far we are in the convergence process. On the other hand, with a proper choice of  $k$ , the JD method clearly outmatches inverse iteration, despite that we use for the latter the exact value of the condition number, whereas, for the JD method, the bound (4.4) is based on the worst-case estimate (4.5).

This is confirmed in Figure 3, where we have plotted the evolution of  $(\theta - \lambda_1)/(\lambda_2 - \theta)$  (worst-case scenario) against the cumulated number of inner iterations. One sees that the JD method converges faster when  $k$  is increased from step to step. In practice, this requires an adaptive stopping criterion such as the one proposed in [14].

Finally, we compared these results with the actual convergence on the following model example:  $n = 10000$ ;  $A = \text{diag}(\lambda_i)$  with  $\lambda_1 = 2$  and  $\lambda_i = 3 + i$ ,  $i = 2, \dots, n$ ;  $K = \text{diag}(\lambda_i(1 + \eta_i))$ ,  $i = 1, \dots, n$ , where the  $\eta_i$  are at random in  $(0, 1)$ ; the initial approximate eigenvector is given by  $(\mathbf{u}_0)_i = \lambda_i^{-2}$ .

Thus, we have  $\lambda_2/\lambda_1 = 5/2$ ,  $\kappa(K^{-1}A) \approx 2$ , and  $(\theta_0 - \lambda_1)/(\lambda_2 - \theta) = 0.953$ ; i.e., this situation is very similar to the one simulated in Figures 2 and 3 (top part). We therefore performed 2, 4, and 8 inner iterations (with the zero initial guess) in the successive JD steps and plotted on Figure 4 the corresponding evolution of  $(\theta - \lambda_1)/(\lambda_2 - \theta)$  against the cumulated number of inner iterations. Note that we do not consider the further improvement that could be obtained with the Galerkin approach mentioned at the beginning of this section and that the quantity given within inner steps corresponds to the quantity one would get if inner iterations were stopped at that moment. For illustration purposes, this actual convergence is compared with the theoretical bound and with the convergence of locally optimal block preconditioned conjugate gradient (LOBPCG) [8], which may be seen as an optimized version of preconditioned inverse iteration. The latter method is not of inner-outer type, and thus we here plot  $(\theta - \lambda_1)/(\lambda_2 - \theta)$  against the number of iterations.

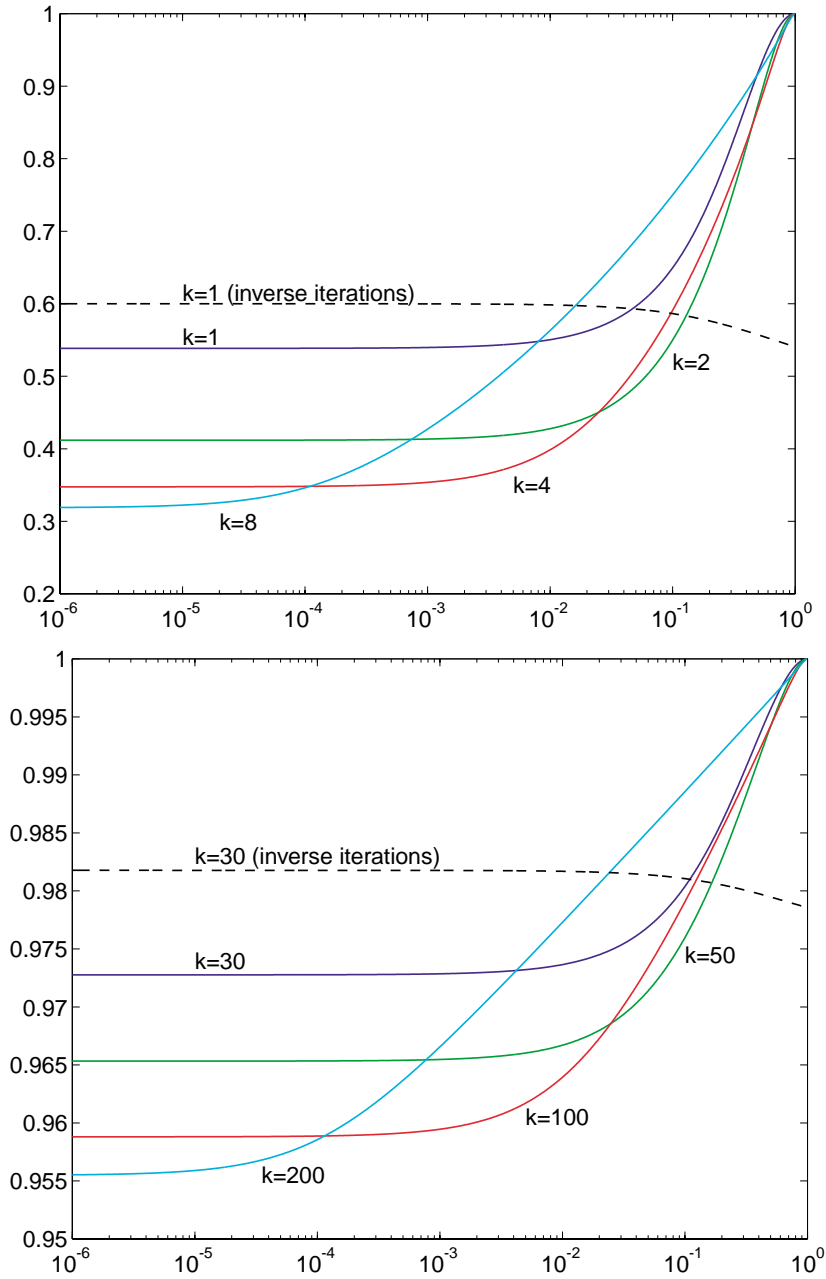


FIG. 2.  $\bar{\sigma}^{1/k}(k, \theta)$  versus  $(\theta - \lambda_1)/(\lambda_2 - \theta)$  for  $\kappa(K^{-1}A) = 2$  (top) and  $\kappa(K^{-1}A) = 1000$  (bottom);  $\lambda_2/\lambda_1 = 5/2$  in both cases.

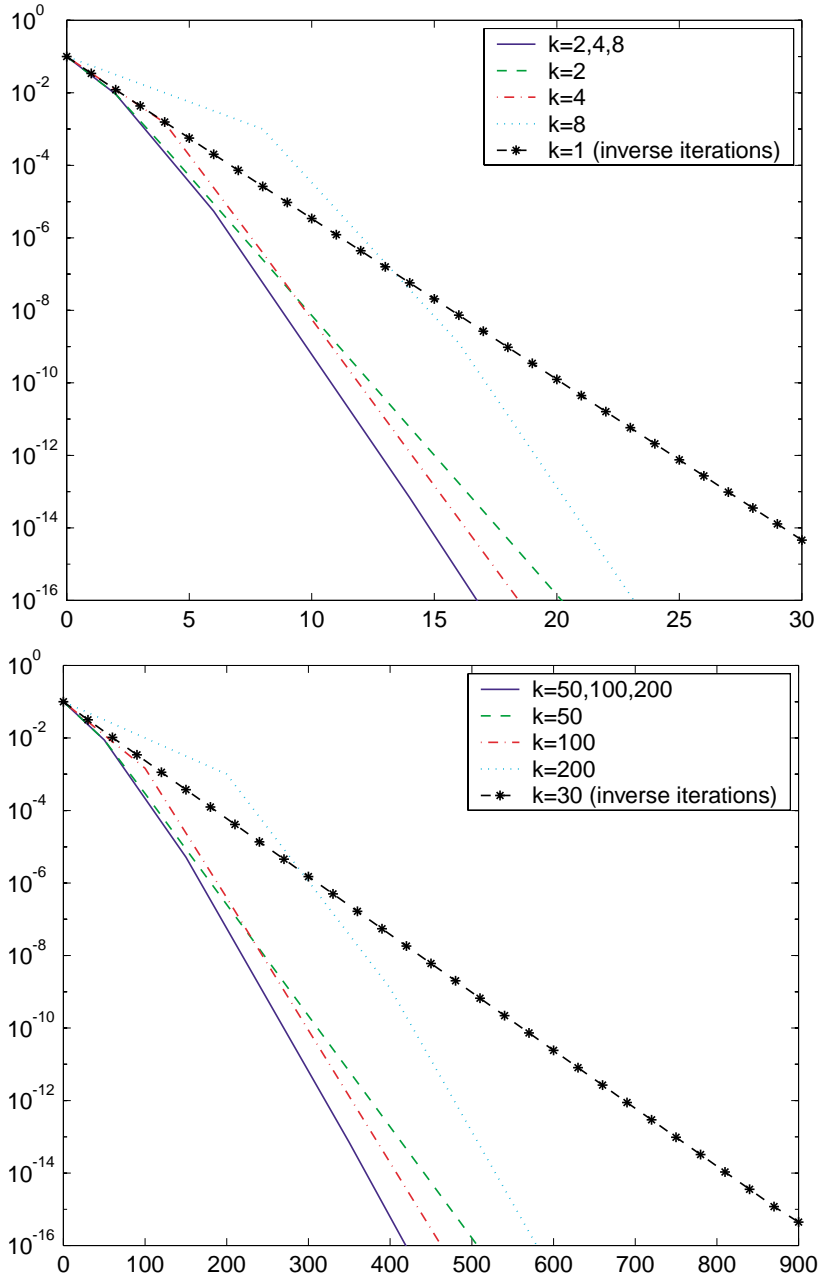


FIG. 3. Evolution of  $(\theta - \lambda_1)/(\lambda_2 - \theta)$  in function of the cumulated number of inner iterations for  $\kappa(K^{-1}A) = 2$  (top) and  $\kappa(K^{-1}A) = 1000$  (bottom);  $\lambda_2/\lambda_1 = 5/2$  in both cases.

**Acknowledgment.** I thank Prof. A. Knyazev for having drawn my attention to (2.7) and for further stimulating discussions.

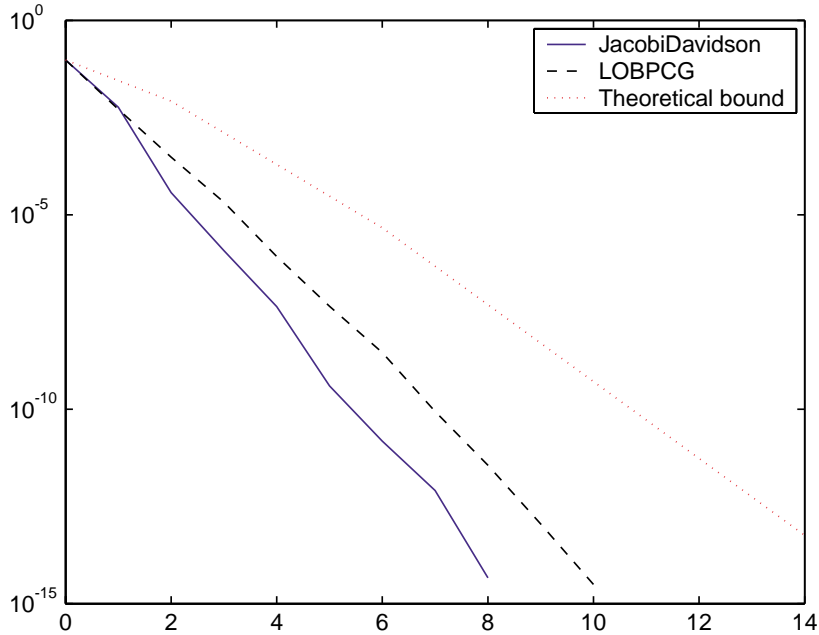


FIG. 4. Evolution of  $(\theta - \lambda_1)/(\lambda_2 - \theta)$  in function of the cumulated number of inner iterations (JD, theoretical bound) or in function of the number of iterations (LOBPCG) for the model example.

## REFERENCES

- [1] Z. BAI, J. DEMMEL, J. DONGARRA, A. RUHE, AND H. A. VAN DER VORST, EDs., *Templates for the Solution of Algebraic Eigenvalue Problems: A Practical Guide*, Software Environ. Tools 11, SIAM, Philadelphia, 2000.
- [2] F. A. DUL, *MINRES and MINERR are better than SYMMLQ in eigenpair computations*, SIAM J. Sci. Comput., 19 (1998), pp. 1767–1782.
- [3] D. R. FOKKEMA, G. L. G. SLEIJPEN, AND H. A. VAN DER VORST, *Jacobi–Davidson style QR and QZ algorithms for the reduction of matrix pencils*, SIAM J. Sci. Comput., 20 (1998), pp. 94–125.
- [4] G. H. GOLUB AND C. F. VAN LOAN, *Matrix Computations*, 3rd ed., The Johns Hopkins University Press, Baltimore, 1996.
- [5] G. H. GOLUB AND Q. YE, *Inexact inverse iterations for the generalized eigenvalue problems*, BIT, 40 (2000), pp. 672–684.
- [6] A. V. KNYAZEV, *Computation of Eigenvalues and Eigenvectors for Mesh Problems: Algorithms and Error Estimates*, Dept. Numerical Math., USSR Academy of Sciences, Moscow, 1986 (in Russian).
- [7] A. V. KNYAZEV, *Convergence rate estimates for iterative methods for a mesh symmetric eigenvalue problem*, Soviet J. Numer. Anal. Math. Modelling, 2 (1987), pp. 371–396.
- [8] A. V. KNYAZEV, *Toward the optimal preconditioned eigensolver: Locally optimal block preconditioned conjugate gradient method*, SIAM J. Sci. Comput., 23 (2001), pp. 517–541.
- [9] A. V. KNYAZEV AND K. NEYMEYR, *A geometric theory for preconditioned inverse iteration III: A short and sharp convergence estimate for generalized eigenvalue problems*, Linear Algebra Appl., 358 (2003), pp. 95–114; also available online from <http://www-math.cudenver.edu/ccmreports/repl73.pdf>, CU-Denver, 2001.
- [10] Y.-L. LAI, K.-Y. LIN, AND W.-W. LIN, *An inexact inverse iteration for large sparse eigenvalue problems*, Numer. Linear Algebra Appl., 4 (1997), pp. 425–437.
- [11] K. NEYMEYR, *A geometric theory for preconditioned inverse iteration I: Extrema of the Rayleigh quotient*, Linear Algebra Appl., 322 (2001), pp. 61–85.
- [12] K. NEYMEYR, *A geometric theory for preconditioned inverse iteration II: Convergence estimates*, Linear Algebra Appl., 322 (2001), pp. 87–104.

- [13] K. NEYMEYR, *A Hierarchy of Preconditioned Eigensolvers for Elliptic Differential Operators*, habilitationsschrift an der mathematischen fakultät, Universität Tübingen, Tübingen, Germany, 2001.
- [14] Y. NOTAY, *Combination of Jacobi-Davidson and conjugate gradients for the partial symmetric eigenproblem*, Numer. Linear Algebra Appl., 9 (2002), pp. 21–44.
- [15] B. N. PARLETT, *The Symmetric Eigenvalue Problem*, Prentice–Hall, Englewood Cliffs, NJ, 1980.
- [16] G. PETERS AND J. H. WILKINSON, *Inverse iteration, ill-conditioned equations and Newton’s method*, SIAM Rev., 21 (1979), pp. 339–360.
- [17] Y. SAAD, *Numerical Methods for Large Eigenvalue Problems*, Halstead Press, New York, 1992.
- [18] Y. SAAD, *Iterative Methods for Sparse Linear Systems*, PWS, New York, 1996.
- [19] V. SIMONCINI AND L. ELDÉN, *Inexact Rayleigh quotient-type methods for eigenvalue computations*, BIT, 42 (2002), pp. 159–182.
- [20] G. L. G. SLEIJPEN, A. BOOTEN, D. R. FOKKEMA, AND H. A. VAN DER VORST, *Jacobi-Davidson type methods for generalized eigenproblems and polynomial eigenproblems*, BIT, 36 (1996), pp. 595–633.
- [21] G. L. G. SLEIJPEN AND H. A. VAN DER VORST, *A Jacobi–Davidson iteration method for linear eigenvalue problems*, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 401–425.
- [22] G. L. G. SLEIJPEN AND H. A. VAN DER VORST, *Hermitian eigenvalue problems*, *Generalized Hermitian eigenvalue problems*, in Templates for the Solution of Algebraic Eigenvalue Problems: A Practical Guide, Software Environ. Tools 11, SIAM, Philadelphia, 2000, Chapters 4.7, 5.6.
- [23] G. L. G. SLEIJPEN, H. A. VAN DER VORST, AND E. MEIJERINK, *Efficient expansion of subspaces in the Jacobi-Davidson method for standard and generalized eigenproblems*, Electron. Trans. Numer. Anal., 7 (1998), pp. 75–89.
- [24] P. SMIT AND M. H. C. PAARDEKOOPER, *The effects of inexact solvers in algorithms for symmetric eigenvalue problems*, Linear Algebra Appl., 287 (1999), pp. 337–357.
- [25] D. B. SZYLD, *Criteria for combining inverse and Rayleigh quotient iteration*, SIAM J. Numer. Anal., 25 (1988), pp. 1369–1375.
- [26] J. H. WILKINSON, *The Algebraic Eigenvalue Problem*, Clarendon, Oxford, UK, 1965.