CONTROLLING INNER ITERATIONS IN THE JACOBI–DAVIDSON METHOD

MICHIEL E. HOCHSTENBACH* AND YVAN NOTAY†

Abstract. The Jacobi–Davidson method is an eigenvalue solver which uses an inner-outer scheme. In the outer iteration one tries to approximate an eigenpair while in the inner iteration a linear system has to be solved, often iteratively, with the ultimate goal to make progress for the outer loop. In this paper we prove a relation between the residual norm of the inner linear system and the residual norm of the eigenvalue problem. We show that the latter may be estimated inexpensively during the inner iterations. On this basis, we propose a stopping strategy for the inner iterations to maximally exploit the strengths of the method. These results extend previous results obtained for the special case of Hermitian eigenproblems with the conjugate gradient or the symmetric QMR method as inner solver. The present analysis applies to both standard and generalized eigenproblems, does not require symmetry, and is compatible with most iterative methods for the inner systems. It can also be extended to other type of inner–outer eigenvalue solvers, such as inexact inverse iteration or inexact Rayleigh quotient iteration. The effectiveness of our approach is illustrated by a few numerical experiments, including the comparison of a standard Jacobi–Davidson code with the same code enhanced by our stopping strategy.

Key words. eigenvalue, Jacobi–Davidson, inverse iteration, Rayleigh quotient iteration, correction equation, subspace expansion, inner iterations

AMS subject classifications. 65F15, 65F10

1. Introduction. The Jacobi–Davidson (JD) method was introduced around a decade ago by Sleijpen and Van der Vorst ([22], see also [24]). It is considered to be one of the best eigenvalue solvers, especially for eigenvalues in the interior of the spectrum. It may be applied to both standard eigenproblems

$$A\,\mathbf{x} \;=\; \lambda\,\mathbf{x} \tag{1.1}$$

and the generalized eigenproblem (GEP) [6]

$$A\,\mathbf{x} \;=\; \lambda\,B\,\mathbf{x} \tag{1.2}$$

or in homogeneous form if B may be singular [6, 11]

$$(\beta\,A - \alpha\,B)\,\mathbf{x} \;=\; 0\;. \tag{1.3}$$

In these equations, A and B are $n \times n$ real or complex matrices, typically large and sparse, so that solving a linear system, for instance of the form $(A-\zeta\,I)\mathbf{v} = \mathbf{y}$, where I is the identity, by a direct method is unattractive or even infeasible. The JD method computes a few selected eigenpairs $(\lambda\,,\,\mathbf{x})$ or $(\alpha/\beta\,,\,\mathbf{x})$, for instance those for which the eigenvalue λ, respectively α/β, is closest to a given target τ.

JD belongs to the class of subspace methods, which means that approximate eigenvectors are sought in a subspace. Each iteration of these methods has two important phases: the subspace extraction, in which a sensible approximate eigenpair

*Université Libre de Bruxelles, Service de Métrologie Nucléaire (C.P. 165-84), 50 Av. F.D. Roosevelt, B-1050 Brussels, Belgium. Research supported by the Belgian FNRS and NSF grant DMS-0405387. Current address: Department of Mathematics and Computing Science, Eindhoven University of Technology, PO Box 513, 5600 MB, The Netherlands, `m.e.hochstenbach@tue.nl`, `www.win.tue.nl/~hochsten`.

†Université Libre de Bruxelles, Service de Métrologie Nucléaire (C.P. 165-84), 50 Av. F.D. Roosevelt, B-1050 Brussels, Belgium. `ynotay@ulb.ac.be`, `homepages.ulb.ac.be/~ynotay`. Research supported by the Belgian FRNS ("Directeur de recherches").

is sought, of which the approximate vector is in the search space; and the subspace expansion, in which the search space is enlarged by adding a new basis vector to it, hopefully leading to better approximate eigenpairs in the next extraction phase.

In this paper we analyze the expansion phase. In JD, the expansion vector is computed by solving an inner linear system, often referred to as the *correction equation*, iteratively (and in general inaccurately). We show how the progress in the outer iteration (extraction phase) is connected to the residual norm in the correction equation, and how this progress may be estimated inexpensively during inner iterations. As will be discussed in §2.5, these results also apply to inexact inverse iteration [8, 12, 26] and inexact Rayleigh quotient iteration (RQI) [20, 2]. Note, however, that our analysis measures the progress in the outer iteration with respect to a particular vector in the next search space; that is, we do not consider here issues related to subspace acceleration. Nevertheless, as discussed in §2.2 below, the outer convergence estimate considered in our results is often a worst case estimate for the method supplemented with a proper subspace extraction algorithm.

JD may be seen as an inexact Newton method [5, 23, 32]. In such methods, it is often challenging to find a proper stopping criterion for inner iterations. Indeed, after some time, the solution of the inner system is accurate enough and any extra work is superfluous. On the other hand, with the exact solution of the linear system, the convergence of the outer iteration is quadratic or even cubic in the Hermitian case [22, 24, 16]. To benefit from this, one ought to compute more accurate solutions of inner systems as the outer convergence proceeds. In this paper, we show how to build efficient stopping criteria for JD, using an inexpensive assessment of the outer convergence during inner iterations. In fact, we extend the results in [15, 27], where similar stopping criteria were proposed but with a much more restricted scope. More precisely, the results in [15, 27] apply to the specific case of Hermitian standard eigenproblems with the conjugate gradient (CG) method [10] or the symmetric QMR method [7] as inner solver, whereas the results in this paper apply to generalized and standard eigenproblems, Hermitian or not, and are compatible with most iterative inner solvers.

The remainder of this paper is organized as follows. In §2 we develop the necessary mathematical results; more precisely, we prove a general theorem in §2.1, and consider its application to several contexts: JD for standard eigenproblems (§2.2), JD for the GEP with skew projection (as often used in the Hermitian case) (§2.3), JD for the GEP with orthogonal projection (as in [6]) (§2.4), inexact inverse iteration and inexact RQI (§2.5). Stopping criteria are discussed in §3. The results of a few numerical experiments and some conclusions are presented in Sections 4 and 5.

2. Theoretical results.

2.1. General theorem. The following theorem contains our main mathematical result, stated in its general form. How it applies to different forms of inner-outer eigensolvers is made clear in the corollaries given in Sections 2.2 through 2.5.

THEOREM 2.1. *Let Q, S be $n \times n$ matrices and $\mathbf{p}$ a vector in $\mathbb{C}^n$ with two-norm $\|\mathbf{p}\| = 1$. For all $\mathbf{z} \in \mathbb{C}^n$ such that $\mathbf{p}^* S\mathbf{z} \neq 0$, letting*

$$r \;=\; \min_{\xi} \|Q\mathbf{z} - \xi\, S\mathbf{z}\| \;, \tag{2.1}$$

there holds

$$\frac{|\, g - \beta\, s\,|}{\sqrt{1+s^2}} \;\le\; r \;\le\; \begin{cases} \sqrt{g^2+\beta^2} & \text{if } \beta < g\, s \\ \dfrac{g+\beta\, s}{\sqrt{1+s^2}} & \text{otherwise}\ , \end{cases} \tag{2.2}$$

where

$$g = \big\|(I - \mathbf{p}\,\mathbf{p}^*)Q\mathbf{z}\big\|\ , \tag{2.3}$$

$$s = \frac{\|(I - \mathbf{p}\,\mathbf{p}^*)S\mathbf{z}\|}{|\mathbf{p}^* S\mathbf{z}|}\ , \tag{2.4}$$

$$\beta = |\mathbf{p}^* Q\mathbf{z}|\ . \tag{2.5}$$

Moreover, if $\big(I - \mathbf{p}\,\mathbf{p}^*\big)Q\mathbf{z} \perp S\mathbf{z}$ *,*

$$r \;=\; \sqrt{g^2 + \frac{\beta^2\, s^2}{1+s^2}}\ . \tag{2.6}$$

Proof. One has

$$r^2 = \|Q\mathbf{z} - \theta\, S\mathbf{z}\|^2 = \|Q\mathbf{z}\|^2 \;-\; |\theta|^2\, \|S\mathbf{z}\|^2 \tag{2.7}$$

where

$$\theta \;=\; \frac{(S\mathbf{z})^* Q\mathbf{z}}{\|S\mathbf{z}\|^2}\ .$$

Further, letting

$$\mathbf{g} \;=\; \big(I - \mathbf{p}\,\mathbf{p}^*\big)Q\mathbf{z}$$

(thus $g = \|\mathbf{g}\|$), since $\mathbf{g} \perp \mathbf{p}$, there holds

$$\|Q\mathbf{z}\|^2 = \|\mathbf{g} + \mathbf{p}\ (\mathbf{p}^* Q\mathbf{z})\,\|^2 \;=\; g^2 \;+\; \beta^2\ . \tag{2.8}$$

On the other hand, let

$$\mathbf{s} \;=\; \big(\mathbf{p}^* S\mathbf{z}\big)^{-1} S\mathbf{z} - \mathbf{p}$$

(thus $s = \|\mathbf{s}\|$). Since

$$S\mathbf{z} = \big(\mathbf{p}^* S\mathbf{z}\big)(\mathbf{p} + \mathbf{s})\ , \tag{2.9}$$

letting

$$\widetilde{\beta} = \mathbf{p}^* Q\mathbf{z}\ ,$$

(thus $\beta = |\widetilde{\beta}|$), one has

$$\begin{aligned} \|S\mathbf{z}\|^2\, \theta &= \big(\mathbf{p}^* S\mathbf{z}\big)^*\ \big((\mathbf{p}+\mathbf{s})^* Q\mathbf{z}\big) \\ &= \big(\mathbf{p}^* S\mathbf{z}\big)^*\ \Big(\widetilde{\beta} + \mathbf{s}^* Q\mathbf{z}\Big) \\ &= \big(\mathbf{p}^* S\mathbf{z}\big)^*\ \Big(\widetilde{\beta} + \mathbf{s}^* \mathbf{g}\Big)\ , \end{aligned} \tag{2.10}$$

the last equality holding because $\mathbf{s} \perp \mathbf{p}$, hence $\mathbf{s}^* = \mathbf{s}^*\big(I - \mathbf{p}\,\mathbf{p}^*\big)$. Observe also that, since $\mathbf{s} \perp \mathbf{p}$, (2.9) implies

$$|\mathbf{p}^* S\mathbf{z}|^2 \;=\; \frac{\|S\mathbf{z}\|^2}{\|\mathbf{p}+\mathbf{s}\|^2} \;=\; \frac{\|S\mathbf{z}\|^2}{1+s^2}\;. \tag{2.11}$$

Letting

$$\delta \;=\; \frac{\mathbf{s}^*\mathbf{g}}{s\,g}$$

if $s\,g \neq 0$, and $\delta = 0$ otherwise. one then finds, with (2.7), (2.8), (2.10), and (2.11),

$$\begin{aligned} r^2 &= g^2 \;+\; \beta^2 \;-\; (1+s^2)^{-1}\;\left|\widetilde{\beta} + \delta\, s\, g\right|^2 \\ &= (1+s^2)^{-1}\;\Big(g^2\;\big(1+s^2\,(1-|\delta|^2)\big) + \beta^2\,s^2 - 2\;\mathrm{Re}\left(\widetilde{\beta}^*\delta\right)\;s\,g\Big)\;. \end{aligned}$$

When $\mathbf{g} \perp S\mathbf{z}$, one has also $\mathbf{g} \perp \mathbf{s}$ since $S\mathbf{z}$ is a multiple of $\mathbf{p}+\mathbf{s}$ and $\mathbf{g} \perp \mathbf{p}$. Then $\delta = 0$ and the last equality gives (2.6). In the general case $0 \leq |\delta| \leq 1$ and one obtains (because also $\mathrm{Re}\left(\widetilde{\beta}^*\delta\right) \geq -\beta\,|\delta|$)

$$\begin{aligned} r^2 &\leq (1+s^2)^{-1}\;\max_{0\leq|\delta|\leq 1}\big(g^2\,(1+s^2\,) + \beta^2\,s^2 - g^2\,s^2\,|\delta|^2 + 2\,\beta\,s\,g\,|\delta|\big)\;, \\ &= g^2 + \beta^2 - (1+s^2)^{-1}\;\min_{0\leq|\delta|\leq 1}(|\delta|\,g\,s-\beta)^2 \\ &= \begin{cases} g^2+\beta^2 & \text{if } \beta < g\,s \\ g^2+\beta^2-(1+s^2)^{-1}(g\,s-\beta)^2 & \text{otherwise} \end{cases} \end{aligned}$$

hence the upper bound in (2.2). On the other hand,

$$r^2 \;\geq\; (1+s^2)^{-1}\;\min_{0\leq|\delta|\leq 1}\big(g^2\,(1+s^2\,) + \beta^2\,s^2 - g^2\,s^2\,|\delta|^2 - 2\,\beta\,s\,g\,|\delta|\big)\;,$$

which gives the lower bound (obtained for $|\delta| = 1$). □

The following lemma helps to understand the relation between the upper bound in (2.2) and the right hand side of (2.6).

LEMMA 2.2. *Let g, β, s, be nonnegative numbers, and define*

$$\overline{r} = \begin{cases} \sqrt{g^2+\beta^2} & \textit{if } \beta < g\,s \\ \dfrac{g+\beta\,s}{\sqrt{1+s^2}} & \textit{otherwise}\;, \end{cases}$$

$$\widetilde{r} = \sqrt{g^2 + \frac{\beta^2\,s^2}{1+s^2}}\;.$$

Then:

$$\widetilde{r} \;\leq\; \overline{r} \;\leq\; \widetilde{r}\,\sqrt{1+\tfrac{1}{1+s^2}}. \tag{2.12}$$

Proof. For $\beta < g\,s$,

$$\left(\frac{\overline{r}}{\widetilde{r}}\right)^2 \;=\; \frac{(g^2+\beta^2)(1+s^2)}{g^2(1+s^2)+\beta^2\,s^2} \;=\; 1+\frac{\beta^2}{g^2(1+s^2)+\beta^2\,s^2}$$

hence $\overline{r} \geq \widetilde{r}$, whereas the upper bound holds if and only if $\frac{\beta^2}{g^2(1+s^2)+\beta^2\,s^2} \leq \frac{1}{1+s^2}$, that is, if and only if $g^2(1+s^2)+\beta^2\,s^2-\beta^2(1+s^2) = g^2(1+s^2)-\beta^2$ is nonnegative, which is always true for $\beta < g\,s$.

On the other hand, for $\beta \geq g\,s$,

$$\left(\frac{\overline{r}}{\widetilde{r}}\right)^2 \;=\; \frac{(g+\beta\,s)^2}{g^2(1+s^2)+\beta^2\,s^2} \;=\; 1+\frac{g\,s\,(2\,\beta-g\,s)}{g^2(1+s^2)+\beta^2\,s^2}$$

hence $\overline{r} \geq \widetilde{r}$, whereas the upper bound holds if and only if $\frac{g\,s\,(2\,\beta-g\,s)}{g^2(1+s^2)+\beta^2\,s^2} \leq \frac{1}{1+s^2}$, that is, if and only if $g^2(1+s^2)+\beta^2\,s^2-g\,s\,(2\,\beta-g\,s)(1+s^2) = (g(1+s^2)-\beta\,s)^2$ is nonnegative. □

2.2. JD for standard eigenproblems. Consider the standard eigenproblem (1.1), let $\mathbf{u}$ be the current approximation to the sought eigenvector, and assume for convenience $\|\mathbf{u}\| = 1$. JD computes an orthogonal correction $\mathbf{t}$ to $\mathbf{u}$ by (approximately) solving the correction equation

$$(I-\mathbf{u}\,\mathbf{u}^*)(A-\zeta\,I)\,\mathbf{t} \;=\; -\,\mathbf{r} \qquad \text{with } \mathbf{t} \perp \mathbf{u}\,, \tag{2.13}$$

where

$$\mathbf{r} \;=\; (A-\theta\,I)\;\mathbf{u}$$

is the eigenvalue residual with

$$\theta \;=\; \mathbf{u}^*A\mathbf{u}.$$

If we solve this equation exactly, then we find [22]

$$\mathbf{t} \;=\; -\,\mathbf{u}+\gamma\,(A-\zeta\,I)^{-1}\mathbf{u}\,, \tag{2.14}$$

where $\gamma = (\mathbf{u}^*(A-\zeta I)^{-1}\mathbf{u})^{-1}$ is such that $\mathbf{t} \perp \mathbf{u}$. The solution is used to expand the search space. Since $\mathbf{u}$ is already in this space, the expansion vector is effectively $(A-\zeta\,I)^{-1}\mathbf{u}$, which is the same as for inverse iteration if $\zeta = \tau$, and the same as for RQI if $\zeta = \theta$. Our analysis applies independently of the choice made for ζ. For the sake of completeness, let us mention that, in general, one selects $\zeta = \tau$ in the initial phase (to enforce convergence towards eigenvalue closest to the target τ), and $\zeta = \theta$ in the final phase (to benefit from the fast convergence of RQI), see [15, 29] for a discussion.

Since $\mathbf{t}$ aims at being a correction to $\mathbf{u}$, the efficiency of the expansion phase may be measured by assessing the eigenvalue residual norm associated to $\mathbf{u}+\mathbf{t}$. This can be done with Theorem 2.1, as made clear in Corollary 2.3 below. Note that, therefore, this analysis measures the progress in the outer iteration with respect to a particular vector in the next search space; that is, it is is based on a simplified extraction phase. Issues related to more sophisticated (subspace) extraction algorithms are outside of the scope of the present work. Note, however, that the aim is to *improve* the convergence, compared with the "simple" vector $\mathbf{u}+\mathbf{t}$. If this goal is reached, then the estimate

for the outer iteration given in Corollary 2.3 is also a worst case estimate for the real process. Moreover, this worst case estimate property is always true if the extraction is implemented in such a way that $\mathbf{u}+\mathbf{t}$ is selected when the subspace algorithm fails to provide a better eigenvector approximation (that is, if the eigenvalue residual associated with this approximation is larger than the eigenvalue residual associated to $\mathbf{u}+\mathbf{t}$).

COROLLARY 2.3 (JD for standard eigenproblems). *Let A be an $n \times n$ matrix, ζ a complex number, and $\mathbf{u}$ a vector in $\mathbb{C}^n$ such that $\|\mathbf{u}\| = 1$. Let $\mathbf{t}$ be a vector in $\mathbb{C}^n$ such that $\mathbf{t} \perp \mathbf{u}$, and let*

$$\mathbf{r}_{\rm in} \;=\; -\,\mathbf{r} - (I - \mathbf{u}\,\mathbf{u}^*)\,\big(A - \zeta\, I\big)\mathbf{t} \tag{2.15}$$

where $\mathbf{r} = (A - \theta\, I)\ \mathbf{u}$ with $\theta = \mathbf{u}^* A \mathbf{u}$. Then

$$r_{\rm eig} \;=\; \min_{\xi} \frac{\|(A - \xi\, I)(\mathbf{u}+\mathbf{t})\|}{\|\mathbf{u}+\mathbf{t}\|} \tag{2.16}$$

satisfies

$$\frac{|\, g - \beta\, s\,|}{1+s^2} \;\le\; r_{\rm eig} \;\le\; \begin{cases} \frac{\sqrt{g^2+\beta^2}}{\sqrt{1+s^2}} & \text{if } \beta < g\, s \\ \frac{g+\beta\, s}{1+s^2} & \text{otherwise}\,, \end{cases} \tag{2.17}$$

where

$$g = \|\mathbf{r}_{\rm in}\|\,, \tag{2.18}$$

$$s = \|\mathbf{t}\|\,, \tag{2.19}$$

$$\beta = |\,\theta - \zeta + \mathbf{u}^*\big(A - \zeta\, I\big)\mathbf{t}\,|\,. \tag{2.20}$$

Moreover, if $\mathbf{r}_{\rm in} \perp \mathbf{t}$,

$$r_{\rm eig} \;=\; \sqrt{\frac{g^2}{1+s^2} + \left(\frac{\beta\, s}{1+s^2}\right)^2}\,. \tag{2.21}$$

Proof. Apply Theorem 2.1 with $Q = A - \zeta\, I$, $S = I$, $\mathbf{p} = \mathbf{u}$, $\mathbf{z} = \mathbf{u}+\mathbf{t}$, and take into account that $\mathbf{t} \perp \mathbf{u}$ implies $\|\mathbf{u}+\mathbf{t}\| = \sqrt{1+s^2}$, whereas $\big(I - \mathbf{p}\,\mathbf{p}^*\big)Q\mathbf{z} \perp S\mathbf{z}$ amounts to $\mathbf{r}_{\rm in} \perp (\mathbf{u}+\mathbf{t})$, hence the stated condition since $\mathbf{r}_{\rm in} \perp \mathbf{u}$. ∎

The condition $\mathbf{r}_{\rm in} \perp \mathbf{t}$ holds in the Hermitian case when solving (2.13) by the (preconditioned) CG method with the zero vector as initial approximation ([10]; see, e.g., [15] for a formal proof in the preconditioned case). This is used in [15] to derive a theoretical result similar to Corollary 2.3 but restricted to that particular context of Hermitian eigenproblems with inner CG iterations. In [27], the close relation between symmetric QMR and CG is exploited to extend the approach, which nevertheless remains limited to the Hermitian case with specific inner solvers.

With CG, $\mathbf{r}_{\rm in} \perp \mathbf{t}$ holds because this method makes the (inner) residual orthogonal to the space in which the approximate solution is sought. In the non-Hermitian case, this property is also satisfied by the full orthogonal method (FOM) [18]. However, this method is seldom used in practice: GMRES has about the same cost per iteration and is generally preferred. With this method, the (inner) residual is minimized; this is still equivalent to an orthogonality condition on $\mathbf{r}_{\rm in}$, but with respect to an

another subspace, which, in general, does not contain $\mathbf{t}$. Similar remarks apply to the symmetric (or Hermitian) variant MINRES, and also to BiCG and related methods (BiCGSTAB, QMR), which are built from an orthogonality condition with respect to an auxiliary subspace, which, again, in general does not contain the computed approximate solution.

Hence, as a general rule, $\mathbf{r}_{\text{in}}$ will not be orthogonal to $\mathbf{t}$ in the non-Hermitian case. However, letting $\overline{r}_{\text{eig}}$ denote the upper bound in (2.17), one sees that it corresponds to $\overline{r}$ in Lemma 2.2 divided by $\sqrt{1+s^2}$. Thus,

$$\sqrt{\frac{g^2}{1+s^2}+\left(\frac{\beta\, s}{1+s^2}\right)^2} \;\le\; \overline{r}_{\text{eig}} \;\le\; \sqrt{1+\tfrac{1}{1+s^2}}\,\sqrt{\frac{g^2}{1+s^2}+\left(\frac{\beta\, s}{1+s^2}\right)^2}\;; \qquad (2.22)$$

that is, the upper bound in (2.17) is, in any case, within a factor of at most $\sqrt{2}$ of the value (2.21) obtained when $\mathbf{r}_{\text{in}} \perp \mathbf{t}$.

We now illustrate our results with two numerical examples. In Example 1, the matrix is *SHERMAN4* from the Harwell-Boeing Collection available on matrix market [13]. The matrix is real, unsymmetric, and of order $n = 1104$. Its eigenvalues are real, the smallest ones being (in 3 decimal places) 0.031, 0.085, 0.278, 0.399, 0.432, 0.590. The target eigenvalue is the one closest to $\tau = 0.5$, i.e., the fifth eigenvalue. Letting $\mathbf{z}$ be the corresponding eigenvector (with unit norm) and $\mathbf{v}$ a vector with random entries uniformly distributed in the interval $(-1, 1)$, we set $\mathbf{u} = \|\mathbf{z}+\xi\mathbf{v}\|^{-1}(\mathbf{z}+\xi\mathbf{v})$ with $\xi = 0.25/\|\mathbf{v}\|$. We solve the correction equation (2.13) with $\zeta = \tau = 0.5$, using full GMRES [19] with the zero vector as initial approximation.

The second matrix (Example 2), which we call *BandRand* for further reference, is a band lower triangular matrix A of order $n = 1000$ with prescribed diagonal $a_{ii} = \sqrt{i}$ and five subdiagonals with random entries uniformly distributed in the interval $(-1, 1)$. Because the matrix is lower triangular, the diagonal entries give the eigenvalues, and we target the sixth one: $\sqrt{6}$. Letting $\mathbf{z}$ be the corresponding eigenvector (with unit norm) and $\mathbf{v}$ a vector with random entries uniformly distributed in the interval $(-1, 1)$, we set $\mathbf{u} = \|\mathbf{z}+\xi\mathbf{v}\|^{-1}(\mathbf{z}+\xi\mathbf{v})$ with $\xi = 0.025/\|\mathbf{v}\|$. We solve the correction equation (2.13) with $\zeta = \theta = \mathbf{u}^* A\mathbf{u}$, again using full GMRES with zero initial approximation.

In Figure 2.1, we show the different quantities referred to in Corollary 2.3 against the number of inner iterations. After a few iterations, both s and β converge smoothly towards their asymptotic values (which correspond to (2.19) and (2.20) with $\mathbf{t}$ equal to the exact solution (2.14)). On the other hand, $g = \|\mathbf{r}_{\text{in}}\|$ decreases as the convergence proceeds; then, as expected from (2.22), r_{eig} follows closely $g/\sqrt{1+s^2}$ until $g/\sqrt{1+s^2} \approx \beta\, s/(1+s^2)$ which means $g \approx \beta\, s/\sqrt{1+s^2}$. From there, r_{eig} stagnates to a value close to $\beta\, s/(1+s^2)$. The latter value is itself close to the eigenvalue residual norm obtained with the exact solution of the correction equation, that is, with an exact inverse iteration step (Example 1), or an exact RQI step (Example 2). We can see this by setting $g = 0$ in (2.17).

We now discuss the practical computation of β and s during the inner iterations. If A is Hermitian, $\mathbf{u}^*(A-\zeta\, I)\mathbf{t} = \mathbf{r}^*\mathbf{t}$. With CG or symmetric QMR, β may therefore be computed from the inner products needed by the method, that is, without any extra work, see [15, 27] for details. In the general case, one may use the following observation. In most iterative methods, $\mathbf{t}$ is computed as a linear combination of basis vectors $\mathbf{v}_i$ (sometimes called "search directions") [18, 31]. Since iterative methods require the multiplication of these basis vectors by the system matrix, which is here $(I - \mathbf{u}\,\mathbf{u}^*)(A - \zeta\, I)$, the inner products $\mathbf{u}^*(A - \zeta\, I)\mathbf{v}_i$ are needed by the method.

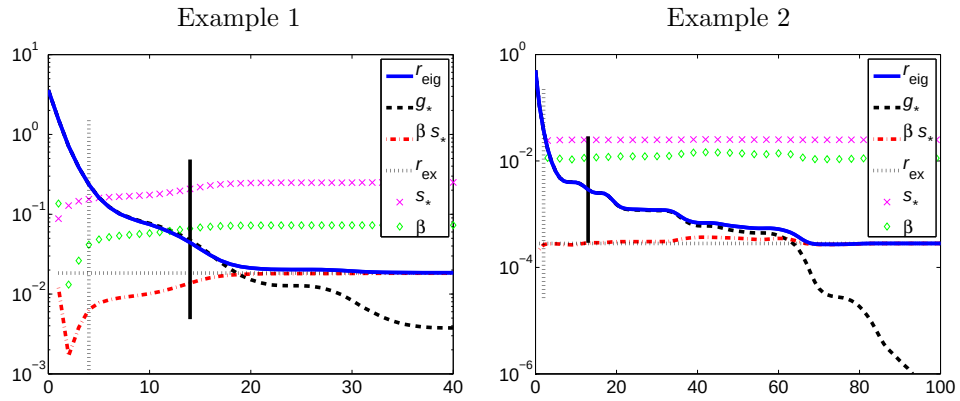

FIG. 2.1. *Quantities referred to in Corollary 2.3 when solving the correction equation* (2.13) *for Examples 1 and 2 with full GMRES;* $g_* = g/\sqrt{1+s^2}$, $s_* = s/(1+s^2)$, *and* $r_{\rm ex}$ *is the eigenvalue residual norm* (2.16) *corresponding to the exact solution of the correction equation* (2.13)*; the dotted vertical line corresponds to the first iteration for which the relative residual error is below* 0.1 *(i.e.,* $g < 10^{-1}\|\mathbf{r}\|$*), and the solid vertical line corresponds to the exit point according to criterion* (B) *proposed in §3.*

Storing them allows us to compute β at no cost from the coefficients of the linear combination (or the recursion) that expresses $\mathbf{t}$ in the basis vectors.

In most cases, the computation of s is straightforward, but some iterative methods like GMRES do not form the approximate solution at intermediate steps, and computing it would require significant extra work. However, in practice, s and $\beta\, s/(1+s^2)$ often do not vary much after the relative residual error in the linear system has been decreased by a modest factor, say 10. This suggests that relevant estimates can be obtained computing these quantities only a few times. A practical implementation of these principles is suggested in §3.

2.3. JD for the GEP with skew projection. There are two main variants of JD for the GEP. Considering the problem (1.2) with approximate eigenvector $\mathbf{u}$, the first variant uses the correction equation with skew projection

$$\left(I - \frac{B\mathbf{u}\,\mathbf{u}^*}{\mathbf{u}^*B\mathbf{u}}\right)(A - \zeta\, B)\,\mathbf{t} \;=\; -\,\mathbf{r} \qquad \text{with } \mathbf{t} \perp \mathbf{v}\;, \tag{2.23}$$

where

$$\mathbf{r} \;=\; (A - \theta\, B)\;\mathbf{u}$$

with

$$\theta \;=\; \frac{\mathbf{u}^*A\mathbf{u}}{\mathbf{u}^*B\mathbf{u}}$$

is the eigenvalue residual; $\mathbf{v}$ is an auxiliary vector that determines the subspace in which the solution is sought, for instance $\mathbf{v} = B\mathbf{u}$. For us, the choice made for $\mathbf{v}$ is unimportant.

This variant of JD is often preferred for generalized Hermitian eigenvalue problems, that is, when A, B are Hermitian and B is positive definite [25]. Indeed, in this case the method is equivalent to standard JD applied to the transformed problem

$B^{-1/2}A\,B^{-1/2}\mathbf{z} = \lambda\,\mathbf{z}$. Our results, however, will be for general A and B, assuming only $\mathbf{u}^*B\mathbf{u} \neq 0$, as already implicitly done when writing (2.23) [1] .

The exact solution to (2.23) is

$$\mathbf{t} \;=\; -\,\mathbf{u} + \gamma\,(A-\zeta\,B)^{-1}B\,\mathbf{u}\;, \tag{2.24}$$

where γ is such that $\mathbf{t} \perp \mathbf{v}$. Hence, with an approximate solution to (2.23), the expansion vector $\mathbf{u}+\mathbf{t}$ approximates an inverse iteration step or an RQI step, according to the choice of ζ. The following corollary shows how the corresponding eigenvalue residual norm may be estimated.

As for any eigenproblem, the eigenvalue residual norm is meaningful only if the approximate eigenvector is normalized in a certain way. Our results are stated for the case where one searches the solution to (1.2) such that $\|B\mathbf{z}\| = 1$. That is, we assume that $\|B\mathbf{u}\| = 1$ and we estimate the eigenvalue residual norm for $\frac{\mathbf{u}+\mathbf{t}}{\|B(\mathbf{u}+\mathbf{t})\|}$. The reason for this is that the statements become somewhat simpler, more transparent, and sometimes computationally slightly more efficient. However, this scaling is no real restriction (Note that the assumption $\mathbf{u}^*B\mathbf{u} \neq 0$ implies $B\mathbf{u} \neq 0$).

Corollary 2.4 (JD for the GEP, skew projection). *Let A, B be $n\times n$ matrices, ζ a complex number, $\mathbf{u}$ a nonzero vector in $\mathbb{C}^n$ such that $\|B\mathbf{u}\| = 1$ and $\mathbf{u}^*B\mathbf{u} \neq 0$. Set*

$$\mathbf{p} \;=\; B\mathbf{u}\;.$$

Let $\mathbf{t}$ be a vector in $\mathbb{C}^n$, and let

$$\mathbf{r}_{\rm in} \;=\; -\,\mathbf{r} - \Big(I - \tfrac{B\mathbf{u}\,\mathbf{u}^*}{\mathbf{u}^*B\mathbf{u}}\Big)\,(A-\zeta\,B)\mathbf{t} \tag{2.25}$$

where $\mathbf{r} = (A-\theta\,B)\;\mathbf{u}$ with $\theta = \frac{\mathbf{u}^*A\mathbf{u}}{\mathbf{u}^*B\mathbf{u}}$. Assume $\mathbf{p}^*B\mathbf{t} \neq -1$. Then

$$r_{\rm eig} \;=\; \min_{\xi} \frac{\|(A-\xi\,B)(\mathbf{u}+\mathbf{t})\|}{\|B(\mathbf{u}+\mathbf{t})\|} \tag{2.26}$$

satisfies

$$\frac{|\,g-\beta\,s\,|}{\alpha\,(1+s^2)} \;\le\; r_{\rm eig} \;\le\; \begin{cases} \frac{\sqrt{g^2+\beta^2}}{\alpha\,\sqrt{1+s^2}} & \text{if } \beta < g\,s \\ \frac{g+\beta\,s}{\alpha\,(1+s^2)} & \text{otherwise}\;, \end{cases} \tag{2.27}$$

where

$$g = \|\,(I-\mathbf{p}\,\mathbf{p}^*)\mathbf{r}_{\rm in}\| \;=\; \sqrt{\|\mathbf{r}_{\rm in}\|^2 - |\mathbf{p}^*\mathbf{r}_{\rm in}|^2}\;, \tag{2.28}$$

$$\alpha = |\,1+\mathbf{p}^*B\mathbf{t}|\;, \tag{2.29}$$

$$s = \alpha^{-1}\sqrt{\|B\mathbf{t}\|^2 - |\mathbf{p}^*B\mathbf{t}|^2}\;, \tag{2.30}$$

$$\beta = |\,\theta-\zeta+\mathbf{p}^*\mathbf{r}+\mathbf{p}^*(A-\zeta\,B)\mathbf{t}\,|\;. \tag{2.31}$$

[1] $\mathbf{u}^*B\mathbf{u}$ may become small when searching for infinite eigenvalues, but then it is better to use the approach based on (1.3), which considered in §2.4

Proof. Apply Theorem 2.1 with $Q = A - \zeta B$, $S = B$, $\mathbf{z} = \mathbf{u} + \mathbf{t}$; take into account that $(I - \mathbf{p}\,\mathbf{p}^*)\left(I - \frac{B\mathbf{u}\,\mathbf{u}^*}{\mathbf{u}^*B\mathbf{u}}\right) = I - \mathbf{p}\,\mathbf{p}^*$, and that (2.11) implies $\|B(\mathbf{u}+\mathbf{t})\| = \sqrt{1+s^2}\,|\mathbf{p}^*B(\mathbf{u}+\mathbf{t})| = \sqrt{1+s^2}\,|1 + \mathbf{p}^*B\mathbf{t}|$. ∎

The practical use of this result raises the same comments as in §2.2. Two additional difficulties need to be mentioned. Firstly, computing g requires an additional inner product. However, in several tests we experienced that the stopping strategy that will be proposed in §3 was equally effective using the upper bound $\|\mathbf{r}_{\text{in}}\|$ instead of the exact value of g. Secondly, $\mathbf{p}^*B\mathbf{t}$ is needed to compute α, and the observation mentioned in §2.2 to compute β does not apply because the projector in the correction equation (2.23) is not $I - \mathbf{p}\,\mathbf{p}^*$. However, the comment made about the computation of s may be extended here, that is, the extra work may be kept modest by computing these quantities only a few times.

2.4. JD for the GEP with orthogonal projection. The second main variant of JD for the GEP has been proposed in [6]. Considering the problem (1.3) with approximate eigenvector $\mathbf{u}$, it uses the correction equation

$$(I - \mathbf{p}\,\mathbf{p}^*)\,(\eta A - \zeta B)\,\mathbf{t} \;=\; -\,\mathbf{r} \qquad \text{with } \mathbf{t} \perp \mathbf{v}\,, \tag{2.32}$$

where

$$\begin{aligned}
\eta &= \frac{\mathbf{p}^*B\mathbf{u}}{\sqrt{|\mathbf{p}^*A\mathbf{u}|^2 + |\mathbf{p}^*B\mathbf{u}|^2}}\,,\\
\zeta &= \frac{\mathbf{p}^*A\mathbf{u}}{\sqrt{|\mathbf{p}^*A\mathbf{u}|^2 + |\mathbf{p}^*B\mathbf{u}|^2}}\,,\\
\mathbf{r} &= (\eta\,A - \zeta\,B)\,\mathbf{u}\,,\\
\mathbf{p} &= \frac{(\nu A + \mu B)\mathbf{u}}{\|(\nu A + \mu B)\mathbf{u}\|}
\end{aligned}$$

for some given numbers ν, μ (see [6] for a discussion of two possible choices); $\mathbf{v}$ is an auxiliary vector that determines the subspace in which the solution is sought, but is not of importance here. Note that $\mathbf{r} \perp \mathbf{p}$; the exact solution to (2.32) is

$$\mathbf{t} \;=\; -\,\mathbf{u} + \gamma\,(\eta A - \zeta B)^{-1}(\nu A + \mu B)\,\mathbf{u}\,, \tag{2.33}$$

where γ is such that $\mathbf{t} \perp \mathbf{v}$. Therefore, with an approximate solution to (2.23), the expansion vector $\mathbf{u} + \mathbf{t}$ approximates an RQI step. The following corollary shows how the corresponding eigenvalue residual norm may be estimated. Again, we use a form of normalization: we scale $\mathbf{u}$ such that $\|(\nu A + \mu B)\mathbf{u}\| = 1$, and we estimate the eigenvalue residual norm for $\frac{\mathbf{u}+\mathbf{t}}{\|(\nu A+\mu B)(\mathbf{u}+\mathbf{t})\|}$.

Corollary 2.5 (JD for the GEP, orthogonal projection). *Let A, B be $n \times n$ matrices, η, ζ as above, ν, μ be complex numbers, $\mathbf{u} \in \mathbb{C}^n$ with $\|(\nu A + \mu B)\mathbf{u}\| = 1$, and set*

$$\mathbf{p} = (\nu A + \mu B)\mathbf{u}\,.$$

Let $\mathbf{t}$ be a vector in $\mathbb{C}^n$, and let

$$\mathbf{r}_{\text{in}} \;=\; -\,\mathbf{r} - (I - \mathbf{p}\,\mathbf{p}^*)\,(\eta A - \zeta B)\mathbf{t} \tag{2.34}$$

where $\mathbf{r} = (\eta\, A - \zeta\, B)\,\mathbf{u}$. *Assume that* $\mathbf{p}^*(\nu A + \mu B)\mathbf{t} \neq -1$. *Then*

$$r_{\text{eig}} \;=\; \min_{\xi} \frac{\big\|\big((\eta A - \zeta B) - \xi\,(\nu A + \mu B)\big)(\mathbf{u}+\mathbf{t})\big\|}{\|(\nu A + \mu B)(\mathbf{u}+\mathbf{t})\|} \tag{2.35}$$

satisfies

$$\frac{|\,g - \beta\, s\,|}{\alpha\,(1+s^2)} \;\leq\; r_{\text{eig}} \;\leq\; \begin{cases} \frac{\sqrt{g^2+\beta^2}}{\alpha\,\sqrt{1+s^2}} & \text{if } \beta < g\, s \\ \frac{g+\beta\, s}{\alpha\,(1+s^2)} & \text{otherwise}\ , \end{cases} \tag{2.36}$$

where

$$g = \|\mathbf{r}_{\text{in}}\|\ , \tag{2.37}$$

$$\alpha = |1 + \mathbf{p}^*(\nu A + \mu B)\mathbf{t}|\ , \tag{2.38}$$

$$s = \alpha^{-1}\sqrt{\|(\nu A + \mu B)\mathbf{t}\|^2 - |\mathbf{p}^*(\nu A + \mu B)\mathbf{t}|^2}\ , \tag{2.39}$$

$$\beta = |\,\mathbf{p}^*\big(\eta A - \zeta B\big)\mathbf{t}\,|\ . \tag{2.40}$$

Proof. Theorem 2.1 applies with $Q = \eta A - \zeta B$, $S = \nu A + \mu B$, $\mathbf{z} = \mathbf{u} + \mathbf{t}$ and $g = \|\mathbf{r}_{\text{in}}\|$. The stated result then follows noting that (2.11) implies $\|(\nu A + \mu B)(\mathbf{u}+\mathbf{t})\| = \sqrt{1+s^2}\ |\mathbf{p}^*(\nu A + \mu B)(\mathbf{u}+\mathbf{t})| = \sqrt{1+s^2}\ |1 + \mathbf{p}^*(\nu A + \mu B)\mathbf{t}|$. □

Here again, the practical use of this result raises the same comments as in §2.2; the observation proposed there to compute $\mathbf{u}^*(A - \zeta\, I)\mathbf{t}$ may be used here to obtain $\mathbf{p}^*\big(\eta A - \zeta B\big)\mathbf{t}$ and hence β ; $\mathbf{p}^*(\nu A + \mu B)\mathbf{t}$ can also be computed efficiently in a similar manner.

2.5. Inexact inverse iteration and RQI. Consider, for the sake of generality, the GEP (1.2) (the standard case is obtained with $B = I$). Let $\mathbf{u}$ be the current eigenvector approximation. Inverse iteration and RQI compute a new eigenvector approximation by solving

$$\big(A - \zeta\, B\big)\,\mathbf{z} \;=\; B\,\mathbf{u}\ . \tag{2.41}$$

RQI uses ζ equal to the Rayleigh quotient $\frac{\mathbf{u}^* A \mathbf{u}}{\mathbf{u}^* B \mathbf{u}}$, whereas inverse iteration uses some fixed shift. The "inexact" variants are obtained when (2.41) is solved approximately by an iterative method. Theorem 2.1 applies also here, as shown in the following corollary.

COROLLARY 2.6 (Inexact inverse iteration and RQI). *Let* A , B *be* $n \times n$ *matrices,* ζ *a complex number,* $\mathbf{u} \in \mathbb{C}^n$ *such that* $\|B\mathbf{u}\| = 1$, *and set*

$$\mathbf{p} \;=\; B\mathbf{u}\ .$$

Let $\mathbf{z}$ *be a vector in* $\mathbb{C}^n$, *and let*

$$\mathbf{r}_{\text{in}} \;=\; B\mathbf{u} - \big(A - \zeta\, B\big)\mathbf{z}\ . \tag{2.42}$$

Assume $\mathbf{p}^* B\mathbf{z} \neq 0$. *Then*

$$r_{\text{eig}} \;=\; \min_{\xi} \frac{\|(A - \xi\, B)\mathbf{z}\|}{\|B\mathbf{z}\|} \tag{2.43}$$

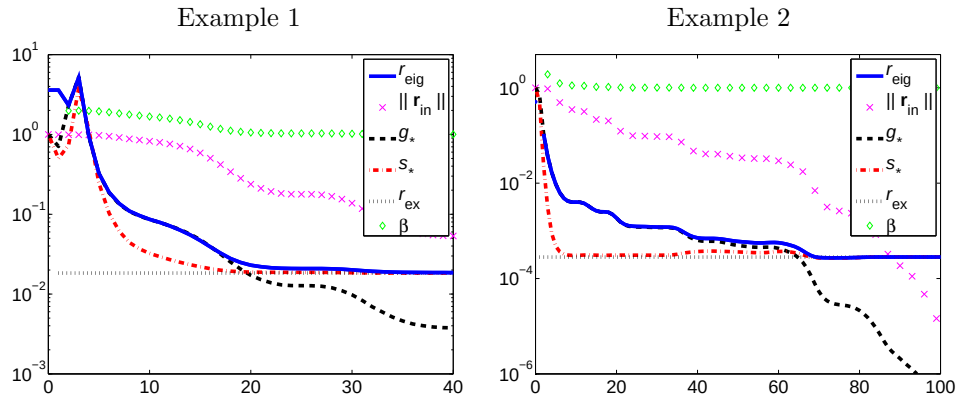

FIG. 2.2. *Quantities referred to in Corollary 2.6 when solving* (2.41) *for Examples 1 and 2 with full GMRES* ($B = I$, $\mathbf{p} = \mathbf{u}$); $g_* = g/(|\mathbf{u}^*\mathbf{z}|\sqrt{1+s^2})$, $s_* = s/(|\mathbf{u}^*\mathbf{z}|(1+s^2))$ *and* $r_{\rm ex}$ *is the eigenvalue residual norm* (2.43) *corresponding to the exact solution of* (2.41).

satisfies

$$\frac{|\,g-\beta\,s\,|}{|\mathbf{p}^*B\mathbf{z}|\,(1+s^2)} \;\le\; r_{\rm eig} \;\le\; \begin{cases} \frac{\sqrt{g^2+\beta^2}}{|\mathbf{p}^*B\mathbf{z}|\,\sqrt{1+s^2}} & \text{if } \beta < g\,s \\ \frac{g+\beta\,s}{|\mathbf{p}^*B\mathbf{z}|\,(1+s^2)} & \text{otherwise}\,, \end{cases} \tag{2.44}$$

where

$$g = \|(I-\mathbf{p}\,\mathbf{p}^*)\mathbf{r}_{\rm in}\| \;=\; \sqrt{\|\mathbf{r}_{\rm in}\|^2 - |\mathbf{p}^*\mathbf{r}_{\rm in}|^2}\,, \tag{2.45}$$

$$s = \frac{\sqrt{\|B\mathbf{z}\|^2 - |\mathbf{p}^*B\mathbf{z}|^2}}{|\mathbf{p}^*B\mathbf{z}|}\,, \tag{2.46}$$

$$\beta = |\,1-\mathbf{p}^*\mathbf{r}_{\rm in}\,|\,. \tag{2.47}$$

Proof. Apply Theorem 2.1 with $Q = A-\zeta\,B$, $S = B$, and take into account that $\big(A-\zeta\,B\big)\mathbf{z} = B\mathbf{u}-\mathbf{r}_{\rm in}$ to obtain the given expression for β. ∎

Note that, letting $\overline{r}_{\rm eig}$ be the upper bound in (2.44), one has, with the help of Lemma 2.2,

$$\frac{1}{|\mathbf{p}^*B\mathbf{z}|}\sqrt{\frac{g^2}{1+s^2}+\left(\frac{\beta\,s}{1+s^2}\right)^2} \;\le\; \overline{r}_{\rm eig} \;\le\; \frac{\sqrt{1+\frac{1}{1+s^2}}}{|\mathbf{p}^*B\mathbf{z}|}\sqrt{\frac{g^2}{1+s^2}+\left(\frac{\beta\,s}{1+s^2}\right)^2}\,. \tag{2.48}$$

To illustrate this result, we consider the same example matrices as in §2.2 (and $B = I$), with the same starting approximate eigenvectors and values of ζ. Here we solve (2.41) with full GMRES, and, in Figure 2.2, we show the different quantities referred to in Corollary 2.6 against the number of inner iterations. One sees that β converges smoothly towards 1, and that the eigenvalue residual norm $r_{\rm eig}$ follows $g/(|\mathbf{u}^*\mathbf{z}|\sqrt{1+s^2})$ until $g \approx s/\sqrt{1+s^2}$. Note that $g/(|\mathbf{u}^*\mathbf{z}|\sqrt{1+s^2}) = \|(I-\mathbf{u}\,\mathbf{u}^*)\mathbf{r}_{\rm in}\|/\|\mathbf{z}\|$. The importance of $\|\mathbf{z}\|$ in the convergence process is stressed in [20], where a stopping criterion based on monitoring its value is proposed. Here we see that the norm g of the projected inner residual may also play a role.

3. Stopping criteria for JD. Next, we propose a stopping strategy and give its motivation afterwards. We remark that this strategy is based on the mathematical results in the preceding section supplemented with heuristic choices. These choices, as well as the suggested default values for the thresholds τ_1 , τ_2 , τ_3 , have been validated using a set of test problems. However, we do not claim that they represent the definitive truth. The criteria below may be seen as an illustration on how to go from theory to practice, with some freedom left. For instance, practitioners may experiment with threshold values, and our last criterion (C) below is optional.

The proposed stopping strategy can be described as follows. Let $\varepsilon^{(\mathrm{out})}$ be the (outer) tolerance on the norm of the eigenvalue residual $\mathbf{r}$, or some slightly smaller value as discussed below. Let $\alpha^{(\mathrm{est})}$, $s^{(\mathrm{est})}$, $\beta^{(\mathrm{est})}$ be the values computed according to:

- (2.19), (2.20) with $\alpha^{(\mathrm{est})} = 1$ when solving a correction equation of the form (2.13) (with inner residual given by (2.15));
- (2.29), (2.30), (2.31) when solving a correction equation of the form (2.23) (with inner residual given by (2.25));
- (2.38), (2.39), (2.40) when solving a correction equation of the form (2.32) (with inner residual given by (2.34)).

For efficiency reasons, we propose to compute these values only twice in the solution of the inner system: for the first time when the relative inner residual has been reduced by a factor of τ_1 (i.e., $\|\mathbf{r}_{\mathrm{in}}\| < \tau_1\|\mathbf{r}\|$: we assume that the linear solver is used with the zero vector as initial approximation), and next when the relative inner residual has been reduced by a factor of τ_2 (i.e., $\|\mathbf{r}_{\mathrm{in}}\| < \tau_2\|\mathbf{r}\|$). Here, $\tau_1 > \tau_2$ are suitably chosen thresholds; we recommend $\tau_1 = 10^{-1/2}$ and $\tau_2 = 10^{-1}$. Let

$$g_k \;=\; \|\mathbf{r}_{\mathrm{in}}\|$$

be the current inner residual norm (at inner iteration k). Set

$$r_{\mathrm{eig}}^{(k)} \;=\; \begin{cases} \sqrt{\frac{g_k^2}{1+ s^{(\mathrm{est})\,2}} + \left(\frac{\beta^{(\mathrm{est})}\, s^{(\mathrm{est})}}{1+ s^{(\mathrm{est})\,2}}\right)^2} & \text{if (2.13), (2.15) apply and } \mathbf{r}_{\mathrm{in}} \perp \mathbf{t} \\ \frac{\sqrt{g_k^2+ \beta^{(\mathrm{est})\,2}}}{\alpha^{(\mathrm{est})}\sqrt{1+ s^{(\mathrm{est})\,2}}} & \text{otherwise, if } \beta^{(\mathrm{est})} < g_k\, s^{(\mathrm{est})} \\ \frac{g_k+\beta^{(\mathrm{est})}\, s^{(\mathrm{est})}}{\alpha^{(\mathrm{est})}\left(1+ s^{(\mathrm{est})\,2}\right)} & \text{otherwise .} \end{cases}$$

The g_k is computed every inner step, but the $\alpha^{(\mathrm{est})}$, $s^{(\mathrm{est})}$, $\beta^{(\mathrm{est})}$ are only computed twice during the inner iterations.

We propose the three following stopping criteria, where τ_3 is another threshold value (from experiments $\tau_3 = 15$ seems a reasonable value); we exit the inner iterations

if at least one of (A), (B) or (C) holds.

$$\textbf{Exit if} \quad g_k < \tau_1 \|\mathbf{r}\| \quad \textbf{and} \quad r_{\text{eig}}^{(k)} < \varepsilon^{(\text{out})} \tag{A}$$

$$\textbf{Exit if} \quad g_k < \tau_1 \|\mathbf{r}\| \quad \textbf{and} \quad \frac{\beta^{(\text{est})}\, s^{(\text{est})}}{\alpha^{(\text{est})}\big(1 + {s^{(\text{est})}}^2\big)} > \frac{\varepsilon^{(\text{out})}}{2} \quad \textbf{and} \quad g_k < \tau_3 \frac{\beta^{(\text{est})}\, s^{(\text{est})}}{\sqrt{1 + {s^{(\text{est})}}^2}} \tag{B}$$

$$\textbf{Exit if} \quad g_k < \tau_1 \|\mathbf{r}\| \quad \textbf{and} \quad \frac{\beta^{(\text{est})}\, s^{(\text{est})}}{\alpha^{(\text{est})}\big(1 + {s^{(\text{est})}}^2\big)} > \frac{\varepsilon^{(\text{out})}}{2} \quad \textbf{and} \quad k > 1 \quad \textbf{and}$$
$$\begin{cases} \left(\frac{g_k}{g_{k-1}}\right)^2 > \left(2 - \left(\frac{g_{k-1}}{g_{k-2}}\right)^2\right)^{-1} & \text{if a norm minimizing method (GMRES, QMR) is used} \\ g_k > g_{k-1} & \text{otherwise .} \end{cases} \tag{C}$$

We first comment on the condition $g_k < \tau_1 \|\mathbf{r}\|$, which is present in all three stopping criteria. To start with, as explained above, $\alpha^{(\text{est})}$, $s^{(\text{est})}$, $\beta^{(\text{est})}$ are computed when this condition is first met. Before this, we are therefore not able to check the other clauses in (A), (B), and (C). But this condition means also that we ask for a minimal reduction of the relative linear system residual error. From many test examples we concluded that it seems beneficial to avoid exiting the inner iterations too early. In particular, this prevents stagnation in the outer loop, which otherwise may take place [2]. The default threshold $\tau_1 = 10^{-1/2}$ seems to work well in practice; it may be dynamically decreased if stagnation is detected in the outer loop, as may for instance occur for difficult-to-find interior eigenvalues.

Criterion (A) allows us to exit when the outer tolerance has been reached. Depending on the implementation, it may be wise to set $\varepsilon^{(\text{out})}$ slightly smaller than the outer tolerance, for instance taking half of it as also suggested in [27]. Indeed, the residual norm of the approximate eigenpair after the outer loop may sometimes be slightly larger than $r_{\text{eig}}^{(k)}$. This depends on several factors. First, on the extraction process, which works with the entire search space instead of just the span of $\mathbf{u}$ and $\mathbf{t}$; but of course, one may decide to choose $\mathbf{u} + \mathbf{t}$ as approximate eigenvector if (A) is met. Second, on whether or not one has the exact residual norm of the correction equation: with some linear solvers a closely related pseudo-norm of the residual may be easier to compute. Third, one has to take into account possible inaccuracies in α, s, and β; these are avoided if one recomputes them for safety when (A) holds with the values estimated so far.

The condition $\beta^{(\text{est})}\, s^{(\text{est})} / \big(\alpha^{(\text{est})}\big(1 + {s^{(\text{est})}}^2\big)\big) > \varepsilon^{(\text{out})}/2$ present in criteria (B) and (C) ensures that we do not stop if we can solve the correction equation accurately enough to meet the main criterion (A). Consider for instance the standard eigenvalue problem where $\alpha = 1$. If $\beta^{(\text{est})}\, s^{(\text{est})} / \big(\alpha^{(\text{est})}\big(1 + {s^{(\text{est})}}^2\big)\big) < \varepsilon^{(\text{out})}/2$ then from (2.22) it follows that $r_{\text{eig}}^{(k)} < \varepsilon^{(\text{out})}$ if g_k is small enough. In practice, however, this condition has a limited impact, allowing only to avoid an extra outer iteration in a limited number of cases.

[2] For Hermitian matrices, stagnation of the outer loop may also be prevented with the GD+1 variant from [27], at least when inner iterations are stopped early and amount to one application of the preconditioner.

Therefore, the essential clause in criterion (B) is the last one, in which we compare g_k with $\beta^{(\mathrm{est})}\, s^{(\mathrm{est})}/\sqrt{1+{s^{(\mathrm{est})}}^2}$. This is motivated by the observations in §2.2, in particular by (2.22). This is also in line with the strategy in [27] (there, g_k is compared to $r_{\mathrm{eig}}^{(k)}$, but this amounts to the same comparison since (2.21) holds). Alternatively, one could use a comparison of the ratios g_k/g_{k-1} and $r_{\mathrm{eig}}^{(k)}/r_{\mathrm{eig}}^{(k-1)}$ as in [15]. But, following [27], we found the present approach more effective on average. The default threshold $\tau_3 = 15$ is not far from the one proposed in [27] (when translated to our notation) and is based on many numerical experiments. One explanation of this relatively large value is that the convergence of Krylov subspace methods is often irregular.

The threshold also "absorbs" possible inaccuracies in α, s, and β. Consider for instance Figure 2.1 where criterion (B) is illustrated with a solid vertical lines indicating the exit point. The criterion was checked with $\alpha^{(\mathrm{est})}$, $s^{(\mathrm{est})}$, $\beta^{(\mathrm{est})}$ equal to the values at the iteration marked by the dotted vertical line; that is, according the strategy suggested above, at the first iteration for which $g_k < 10^{-1}\|\mathbf{r}\|$. One sees that, for Example 1, $\beta s/(1+s^2)$ has not yet stabilized at that iteration. The asymptotic value is thus underestimated. Thanks to the relatively large threshold, however, this is harmless: the inner iterations are effectively exited before the outer residual starts to stagnate; that is, the main goal of criterion (B) has been achieved despite the inaccurate estimation of $\beta s/(1+s^2)$.

Criterion (C) is motivated by the possibly irregular convergence of the inner iterative method. There are two main classes of Krylov subspace linear solvers; Galerkin methods may converge irregularly with many peaks, whereas norm minimizing methods present smooth convergence curves, but may stagnate [9]. When a Galerkin type method is used, for instance CG, it may be beneficial to exit when one enters a peak, that is when $g_k > g_{k-1}$ since it is not known in advance how many iterations will be needed before the residual decreases again. It is slightly more difficult to properly detect a stagnation in a norm minimizing method. Therefore, we use the following idea. Each such method is closely related to a Galerkin method, and the residual norm $\hat{g}_k$ in the latter is related to the residual norm in the norm minimizing method g_k by

$$\hat{g}_k \;=\; \frac{g_k}{\sqrt{1-\left(\frac{g_k}{g_{k-1}}\right)^2}}$$

[9, 18, 30]. From this relation, one may check that $\hat{g}_k > \hat{g}_{k-1}$ if and only if $\left(\frac{g_k}{g_{k-1}}\right)^2 > \left(2-\left(\frac{g_{k-1}}{g_{k-2}}\right)^2\right)^{-1}$. This motivates the first alternative in criterion (C).

4. Numerical results. All experiments have been done within the MATLAB environment (version 7.1.0.183 (R14) Service Pack 3), running under Linux on a PC with Intel Pentium 4 CPU at 3.0 GHz.

In the first experiment, we consider again the matrices *SHERMAN4* and *BandRand* described in §2.2. We solve the corresponding eigenvalue problem (1.1) for the 10 smallest eigenvalues, using the standard JDQR code by Sleijpen [21], and the same code in which our stopping strategy has been implemented. All options are set to their default, except the preconditioning type: the default is left preconditioned GMRES, which does not allow to monitor the true residual norm as needed by our stopping

Stopping test	Type prec.	1 eigenvalue #MV	1 eigenvalue #Prec.Sol.	1 eigenvalue Time	10 eigenvalues #MV	10 eigenvalues #Prec.Sol.	10 eigenvalues Time
			Matrix *SHERMAN4*				
(A),(B),(C)	Right (F)	34	27	0.18	251	235	1.77
standard	Right (F)	34	27	0.36	225	209	3.58
standard	Right	34	36	0.37	225	286	3.58
standard	Left	36	37	0.34	247	304	3.83
(A),(B),(C)	No	157	-	0.45	868	-	3.21
standard	No	180	-	1.10	825	-	6.00
			Matrix *BandRand*				
(A),(B),(C)	No	125	-	0.43	757	-	3.21
standard	No	103	-	0.68	596	-	5.03

TABLE 4.1

Numerical results for the JDQR code with and without our stopping strategy; whenever used, the preconditioning is an incomplete LU preconditioning as obtained with the MATLAB function luinc *with drop tolerance set to* 10^{-2} *(such a LU preconditioning makes no sense for the lower triangular matrix* BandRand*); #MV is the number of needed matrix vector multiplications, and #Prec.Sol. is the number of times the preconditioner is applied; Time is elapsed CPU time in seconds; RIGHT (LEFT) means right (left) preconditioned GMRES; (F) means FGMRES implementation.*

strategy. For our modified version, we therefore select right preconditioned GMRES, and further chose the FGMRES implementation [17] (referred to as "implicit preconditioning" in [21]), which is more convenient for the computation of $\alpha^{(\mathrm{est})}$, $s^{(\mathrm{est})}$, $\beta^{(\mathrm{est})}$. For the sake of completeness we include the original JDQR method with all possible preconditioning types in the comparison. Note that because GMRES is a method with a long recurrence relation, it needs either to be restarted, or stopped when a maximal number of iteration is reached, regardless of our criteria (A), (B), and (C). Our implementation stops the iterations after 15 GMRES steps.

The results are given in Table 4.1. One sees that our stopping strategy does not reduce the total number of matrix vector products. The default stopping criterion in JDQR exits quite quickly (in practice the method nearly always uses 5 inner iterations), thus it tends to do fewer inner iterations and more outer ones, trusting on the extraction process in the outer loop. However, such subspace iterations are much more costly than inner GMRES iterations, despite the fact that GMRES is a relatively costly method. Therefore, the code is much faster with our stopping strategy.

In the second experiment, we consider generalized Hermitian eigenproblems from the matrix market [13]. Here we compare JDQZ [21], the version of JDQR for the GEP, and JDRPCG_GEP [14]. This code is an extension of the JDCG code [14] to generalized Hermitian eigenproblems. Note that JDCG itself was proposed as a simplified version of JDQR for the particular case of Hermitian eigenproblems [15]. JDRPCG_GEP implements our stopping strategy, uses CG for inner iterations, and, unlike standard implementations, it uses a regular preconditioning (whence "RP" in the acronym; it means that the preconditioner used for the correction equation (2.23) is a mere approximation to $A - \zeta B$, without any projection involved; see also [28] for a related discussion). The matrices have moderate size ($n = 1074$ for *BCSSTK08* and $n = 2003$ for *BCSSTK13*) but exhibit bad conditioning (the ratio between the largest diagonal element and the smallest eigenvalue is close to 10^9).

Code	Prec.	1 eigenvalue #MV	1 eigenvalue #Prec.Sol.	1 eigenvalue Time	10 eigenvalues #MV	10 eigenvalues #Prec.Sol.	10 eigenvalues Time
			Matrices *BCSSTK08*, *BCSSTM08*				
JDRPCG_GEP	CG like	155	148	0.33	1514	1445	3.31
JDQZ	Right (F)	148	147	0.85	1535	1543	12.3
JDQZ	Right	148	174	0.83	1313	1541	10.9
JDQZ	Left	113	132	0.69	1457	1707	11.5
JDRPCG_GEP	No	12424	-	5.27	(5 eigs found in 10^5 iter.)		
JDQZ	No	(no convergence)			(no convergence)		
			Matrices *BCSSTK13*, *BCSSTM13*				
JDRPCG_GEP	CG like	242	229	1.34	1616	1524	9.44
JDQZ	Right (F)	(no convergence)			(no convergence)		
JDQZ	Right	(no convergence)			(no convergence)		
JDQZ	Left	555	649	8.50	3290	3847	57.5

TABLE 4.2

Numerical results for the JDRPCG_GEP and JDQZ codes; whenever used, the preconditioning is an incomplete LU preconditioning as obtained with the MATLAB function luinc *with drop tolerance set to* 10^{-2} *for the first problem and* 10^{-3} *for the second problem; #MV is the number of needed matrix vector multiplications, and #Prec. Sol. is the number of times the preconditioner is applied; Time is elapsed CPU time in seconds; RIGHT (LEFT) means right (left) preconditioned GMRES; (F) means FGMRES implementation.*

For this reason, we do not use JDQZ with the default tolerance (10^{-8}), but use 10^{-6} instead. Otherwise, we use all default options in JDQZ, except that we set the "TestSpace" parameter to "SearchSpace", because this is favorable for generalized Hermitian eigenproblems (JDQZ does not converge with the default choice). This makes also the outer loop similar to the one in JDRPCG_GEP: both methods use standard Rayleigh–Ritz, enabling a fair comparison. The JDRPCG_GEP code is also used with all options to their default, except the tolerance which is adapted to guarantee the same accuracy as obtained with JDQZ (both codes use different criteria to check the outer convergence).

The results are given in Table 4.2. For the second problem, we selected a smaller drop tolerance for the LU preconditioner because otherwise JDQZ does not converge with any of the tested options. Note that here and in the table, "no convergence" means stagnation of the outer loop. JDRPCG_GEP is more robust and always converges if one allows enough matrix vector products. It is also much faster, especially on the second test problem which seems more difficult than the first one.

5. Conclusions. We considered the JD method for several types of eigenvalue problems. In each case, we showed that the (outer) eigenvalue residual norm may be estimated as

$$r_{\text{eig}} \;\approx\; \alpha^{-1}\sqrt{\left(\frac{\|\mathbf{r}_{\text{in}}\|}{\alpha\sqrt{1+s^2}}\right)^2+\left(\frac{\beta\, s}{1+s^2}\right)^2}\,,$$

where $\mathbf{r}_{\text{in}}$ is the residual of the correction equation, and where α, β, s are scalar parameters easy to compute or estimate while solving this equation iteratively. Here $\frac{\beta\, s}{\alpha\,(1+s^2)}$ converges quickly to the eigenvalue residual norm obtained with the exact

solution of the correction equation, that is, with an exact RQI or inverse iteration step, according to the choice of the shift. Moreover, $\alpha = 1$ for standard JD, and, more generally, $\alpha \approx 1$ when the outer method is close to the solution.

Hence, when solving the correction equation iteratively, one has $r_{\rm eig} \approx \frac{\|\mathbf{r}_{\rm in}\|}{\alpha\sqrt{1+s^2}}$ until a certain point where progress is no longer possible since $r_{\rm eig}$ reached its asymptotic value $\frac{\beta\, s}{\alpha\,(1+s^2)}$, so that a further reduction of $\|\mathbf{r}_{\rm in}\|$ is useless. From a theoretical point of view, this shows how and why the JD method converges with an inexact solution of the correction equation.

From a practical point of view, this enables us to propose efficient stopping criteria for the inner iterations that take into account the progress in the outer convergence. We proposed such a strategy which, for symmetric standard eigenproblems, is along the lines of those proposed in [15, 27]. In fact, it has already been implemented verbatim in the JADAMILU software [3], which was found efficient on challenging real life applications, for both exterior and interior eigenproblems [4]. JADAMILU is based on the JD algorithm as implemented in the Matlab code JDRPCG, combined with a proper preconditioning. The results obtained in this paper with JDRPCG_GEP (the extension of JDRPCG to GEP) indicate that a similar extension of JADAMILU to the GEP, which has just been released, should prove similarly efficient. Finally, the proposed stopping strategy can also be applied to nonsymmetric eigenproblems. In this context, our results suggest that it can also improve existing software like JDQR.

Acknowledgments. We thank the referees for useful comments.

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*, SIAM, Philadelphia, PA, 2000.

[2] J. Berns-Müller, I. G. Graham, and A. Spence, *Inexact inverse iteration for symmetric matrices*, Linear Algebra Appl., 416 (2006), pp. 389–413.

[3] M. Bollhöfer and Y. Notay, *JADAMILU software and documentation*. Available online at `http://homepages.ulb.ac.be/~jadamilu/`.

[4] M. Bollhöfer and Y. Notay, *JADAMILU: a software code for computing selected eigenvalues of large sparse symmetric matrices*, Computer Physics Communications, 177 (2007), pp. 951–964.

[5] A. Edelman, T. A. Arias, and S. T. Smith, *The geometry of algorithms with orthogonality constraints*, SIAM J. Matrix Anal. Appl., 20 (1999), pp. 303–353.

[6] D. Fokkema, 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 (1999), pp. 94–125.

[7] R. W. Freund and N. M. Nachtigal, *Software for simplified Lanczos and QMR algorithms*, Appl. Num. Math., 19 (1995), pp. 319–341.

[8] G. H. Golub and Q. Ye, *Inexact inverse iterations for generalized eigenvalue problems*, BIT, 40 (2000), pp. 672–684.

[9] A. Greenbaum, *Iterative Methods for Solving Linear Systems*, vol. 17 of Frontiers in Applied Mathematics, SIAM, Philadelphia, PA, 1997.

[10] M. R. Hestenes and E. Stiefel, *Methods of conjugate gradients for solving linear systems*, J. Res. Nat. Bur. Standards, 49 (1952), pp. 409–436.

[11] M. E. Hochstenbach and Y. Notay, *Homogeneous Jacobi–Davidson*, Electronic Trans. Numer. Anal., 29 (2007), pp. 19–30.

[12] Y.-L. Lai, K.-Y. Lin, and W.-W. Lin, *An inexact inverse iteration for large sparse eigenvalue problems*, Numer. Lin. Alg. Appl., 4 (1997), pp. 425–437.

[13] *The Matrix Market.* `http://math.nist.gov/MatrixMarket`, a repository for test matrices.

[14] Y. Notay, *JDCG, JDRPCG and JDRPCG_GEP codes*. Available online at `http://homepages.ulb.ac.be/~ynotay/`.

[15] ———, *Combination of Jacobi-Davidson and conjugate gradients for the partial symmetric eigenproblem*, Numer. Lin. Alg. Appl., 9 (2002), pp. 21–44.
[16] B. N. Parlett, *The Symmetric Eigenvalue Problem*, SIAM, Philadelphia, PA, 1998. Corrected reprint of the 1980 original.
[17] Y. Saad, *A flexible inner-outer preconditioned GMRES algorithm*, SIAM J. Sci. Comput., 14 (1993), pp. 461–469.
[18] ———, *Iterative Methods for Sparse Linear Systems*, SIAM, Philadelphia, PA, 2003. Second ed.
[19] Y. Saad and M. H. Schultz, *GMRES: a generalized minimal residual algorithm for solving nonsymmetric linear systems*, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 856–869.
[20] V. Simoncini and L. Eldén, *Inexact Rayleigh quotient-type methods for eigenvalue computations*, BIT, 42 (2002), pp. 159–182.
[21] G. Sleijpen, *JDQR and JDQZ codes*.
Available online at `http://www.math.uu.nl/people/sleijpen`.
[22] 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.
[23] ———, *The Jacobi–Davidson method for eigenvalue problems and its relation with accelerated inexact Newton schemes*, in Iterative Methods in Linear Algebra II, S. Margenov and P. S. Vassilevski, eds., Series in Computational and Applied Mathematics Vol.3, IMACS, 1996, pp. 377–389.
[24] ———, *A Jacobi–Davidson iteration method for linear eigenvalue problems*, SIAM Review, 42 (2000), pp. 267–293.
[25] ———, *Jacobi–Davidson Methods*, in Bai et al. [1], 2000, ch. 5.6.
[26] 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.
[27] A. Stathopoulos, *Nearly optimal preconditioned methods for Hermitian eigenproblems under limited memory. Part I: Seeking one eigenvalue*, SIAM J. Sci. Comput., 29 (2007), pp. 481–514.
[28] A. Stathopoulos and J. R. McCombs, *Nearly optimal preconditioned methods for hermitian eigenproblems under limited memory. Part II: Seeking many eigenvalues*, SIAM J. Sci. Comput., 29 (2007), pp. 2162–2188.
[29] D. B. Szyld, *Criteria for combining inverse and Rayleigh quotient iteration*, SIAM J. Numer. Anal., (1988), pp. 1369–1375.
[30] H. A. van der Vorst, *Computational methods for large eigenvalue problems*, in Handbook of Numerical Analysis, Vol. VIII, North-Holland, Amsterdam, 2002, pp. 3–179.
[31] ———, *Iterative Krylov Methods for Large Linear systems*, Cambridge University Press, Cambridge, 2003.
[32] K. Wu, Y. Saad, and A. Stathopoulos, *Inexact Newton preconditioning techniques for large symmetric eigenvalue problems*, Electronic Trans. Numer. Anal., 7 (1998), pp. 202–214.