Electronic Transactions on Numerical Analysis.
Volume 7, 1998, pp. 56-74.

ISSN 1068-9613.

ETNA
Kent State University
etna@mcs.kent.edu

A BLOCK RAYLEIGH QUOTIENT ITERATION WITH LOCAL QUADRATIC CONVERGENCE *

JEAN-LUC FATTEBERT †

Abstract. We present an iterative method, based on a block generalization of the Rayleigh Quotient Iteration method, to search for the p lowest eigenpairs of the generalized matrix eigenvalue problem $Au = \lambda Bu$. We prove its local quadratic convergence when $B^{-1}A$ is symmetric. The benefits of this method are the well-conditioned linear systems produced and the ability to treat multiple or nearly degenerate eigenvalues.

Key words. Subspace iteration, Rayleigh Quotient Iteration, Rayleigh-Ritz procedure.

AMS subject classifications. 65F15.

1. Introduction.

Many scientific applications require the solution of a generalized eigenvalue problem

$$Au = \lambda Bu,$$

where A and B are real $N \times N$ sparse matrices, and B is positive definite. A well-known example is the electronic structure calculation of molecules or solids. In the context of density functional theory, some recent developments in the numerical schemes for *ab initio* electronic structure calculation methods have been obtained by describing the electronic wave functions in finite dimensional vector spaces of larger and larger dimension, or more recently by the use of finite difference schemes on tridimensional grids. In this field, a discretized stationary Schrödinger-like eigenvalue problem (the Kohn-Sham equations) has to be solved. Typically, we are interested in the lowest one hundred eigenpairs from matrices of order larger than 10^5. Due to the diagonal dominance of the matrices, Davidson's method and the preconditioned Lanczos method [2, 3, 15] are very popular in this field. Other methods based on the simultaneous Rayleigh-Quotient minimization methods [12] or subspace preconditioning algorithms [1] are also very common, sometimes in combination with conjugate gradient techniques [5].

These approaches require only a very approximate resolution of linear systems (by conjugate gradient for instance) or the solution of very simple linear systems (typically diagonal). But the resolution of large linear systems is improving because of ever more powerful computers and sophisticated algorithms such as the multigrid method. As a result, iterative eigensolvers requiring an accurate resolution of numerous linear systems have to be considered from a new point of view. New preconditioners can be investigated for classical methods, or direct implementation of methods based on the inverse iteration algorithm can be used. If sufficiently accurate linear solvers are available, subspace iterative methods based on inverse iteration can be implemented without expanding the dimension of the search subspace at each step.

In this article, we present an iterative eigensolver based on a block generalization of the Rayleigh Quotient Iteration (RQI) method [13] and prove its local quadratic convergence. We consider matrices A and B such that $B^{-1}A$ is symmetric (for instance A symmetric, B the identity matrix) with eigenvalues $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_p < \lambda_{p+1} \leq \ldots \leq \lambda_N \in R$. We look

*Received January 29, 1998. Accepted for publication September 2, 1998. Recommended by R. Lehoucq. This research was performed when the author was at the Mathematics Department of the Ecole Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland

† Department of Physics, North Carolina State University, Raleigh, NC 27695-8202. (fatteber@nemo.physics.ncsu.edu).

for the subspace

$$\mathcal{E}_0 = \sum_{j=1}^{p} \mathrm{Ker}(B^{-1}A - \lambda_j I),$$

that is, the subspace spanned by the eigenvectors associated to the lowest eigenvalues of $B^{-1}A$. To prove the local convergence of the algorithm, the main assumptions will be on the starting trial subspace.

The algorithm concerned here is described in Section 2 and a small numerical example is provided to illustrate its convergence rate. In Section 3, technical results are derived concerning the subspaces spanned by the Ritz elements obtained by the Rayleigh-Ritz procedure. These results will be used in Section 4 where a precise description of the algorithm and a proof of its local quadratic convergence are presented. Concluding remarks are presented in Section 5. Some technical lemmas and proofs are given in the Appendix. A variant of the method was first applied to the electronic structure calculations in [7]. More details on its application in this field can be found in [6, 8].

Notations and general assumptions. Throughout this paper we consider the space R^N with the usual scalar product $(x, y) = \sum_{i=1}^{N} x_i y_i$ and the induced norm $\|x\| = (x, x)^{1/2}$, $x, y \in R^N$. To the set $\mathcal{M}_N$ of the real matrices $N \times N$ we associate the spectral norm

$$\|M\| = \max_{x \in R^N, \|x\|=1} \|Mx\|$$

for $M \in \mathcal{M}_N$. We denote by I the identity matrix.

The orthogonal complement of a subspace $\mathcal{V} \subset R^N$ is denoted by $\mathcal{V}^{\perp}$. If $\mathcal{V} =$ Span$\{v_1, \dots, v_m\}$, we denote $M\mathcal{V} =$ Span$\{Mv_1, \dots, Mv_m\}$ for $M \in \mathcal{M}_N$.

Let $z \in R^N$, $\mathcal{V}$ and $\mathcal{W}$ be two subspaces of R^N. According to Kato [10], we define the distance from z to $\mathcal{V}$ by

$$\delta(z, \mathcal{V}) = \min_{v \in \mathcal{V}} \|z - v\|, \tag{1.1}$$

and the distance from $\mathcal{V}$ to $\mathcal{W}$ by

$$\delta(\mathcal{V}, \mathcal{W}) = \max_{v \in \mathcal{V}, \|v\|=1} \delta(v, \mathcal{W}). \tag{1.2}$$

For an interval $\mathcal{I}$ of R and a matrix $M \in \mathcal{M}_N$, symmetric, with eigenvalues $\lambda_1 \le \lambda_2 \le \dots \le \lambda_N$, we define the subspace of R^N

$$\mathcal{E}_M(\mathcal{I}) = \sum_{\lambda_i \in \mathcal{I}} \mathrm{Ker}(M - \lambda_i I). \tag{1.3}$$

2. The Block Rayleigh Quotient Iteration method.

2.1. The algorithm. The algorithm we address here contains two main parts. In the first part, for a given subspace whose dimension is the number of searched eigenpairs, we compute approximate eigenvectors applying the Rayleigh-Ritz procedure (Step 2 of Algorithm 2.1). In the second part, this subspace is updated by computing corrections for each of these trial eigenvectors using a generalization of the RQI method (Steps 3-5 of Algorithm 2.1). A general outline of the Block Rayleigh Quotient Iteration method (BRQI) algorithm is as follows:

ALGORITHM 2.1. *BRQI*

1. *Let the tolerance* ϵ *and an initial* $N \times p$ *matrix* $W^0 = (w_1^0, \ldots, w_p^0)$ *be given. Let* $k = 0$.

2. *Let* X *be a real* $p \times p$ *matrix and* $U^k = (u_1^k, \ldots, u_p^k) = W^k X$, *be such that* $U^{k\,T} U^k = I$ *and* $U^{k\,T} B^{-1} A U^k = \Lambda$ *where* Λ *is a real diagonal* $p \times p$ *matrix whose diagonal elements are ordered by* ($\Lambda_{11} \leq \ldots \leq \Lambda_{pp}$). *Check for convergence by testing the condition*

$$\|AU^k - BU^k \Lambda\| \leq \epsilon.$$

3. *For* $j = 1, \ldots, p$, *let* m_j *and* n_j *be given integers such that* $0 \leq m_j < j$ *and* $0 \leq n_j \leq p - j$. *Define the subspace*

$$U_j^k = (u_{j-m_j}^k, \ldots, u_{j+n_j}^k)$$

(of dimension $1 + m_j + n_j \geq 1$*) and let* Q_j^k *be the orthogonal projector onto the subspace* $\left(BU_j^k\right)^{\perp}$.

4. *For* $j = 1, \ldots, p$, *compute the correction* z_j *such that*

$$Q_j(A - \Lambda_{jj} B)(u_j^k + z_j) = 0 \tag{2.1}$$

and $z_j^T B U_j^k = 0$.

5. *Set* $W^{k+1} = (u_1 + z_1, \ldots, u_p + z_p)$. *Increment* k *by* 1 *and go to step 2).*

At Step 3, the parameters m_j and n_j are integers chosen such that $Q_j^k(A - \Lambda_{jj}B)\Big|_{(BU_j^k)^{\perp}}$ is well-conditioned. For instance, select

$$m_j = \max_{\{i | \Lambda_{jj} - \Lambda_{ii} \leq \alpha\}} j - i, \qquad n_j = \max_{\{i | \Lambda_{ii} - \Lambda_{jj} \leq \alpha\}} i - j$$

for a given real constant $\alpha > 0$. It is easy to see that, in the case $B = Identity$, this would ensure that

$$\|Q_j^k(A - \Lambda_{jj})x\| \geq \min(\alpha, \lambda_{p+1} - \lambda_j)\|x\|, \qquad \forall x \in (U_j^k)^{\perp}$$

at convergence of the algorithm. We will be more precise on this point in §4.1.

It is easy to see that in the particular case $m_j = n_j = 0$, the vector $u_j^k + z_j$ (Step 2.1) is equal (in exact arithmetic), once properly normalized, to the vector u_j^{k+1} updated by a classical RQI iteration [7] (we have in fact $u_j^{k+1} = \gamma(u_j^k + z_j)$, $\gamma \in R$). In this particular case, the equation defining the correction z_j is the same as the one used in the Jacobi-Davidson method [16]. If $max(m_j, n_j) > 0$, z_j is restricted to be in a smaller subspace—meaning that we do not to correct u_j^k in some directions considered before—and the arguments used to prove the convergence of the Jacobi-Davidson method or classical RQI are no longer valid. Nevertheless, because those directions are included in the subspace BU^k (provided $m_j < j$ and $n_j \leq p - j$), convergence can still be attained (as it will be shown in Section 4) by the "mixing" of the updated trial eigenvectors in the Rayleigh-Ritz procedure. Moreover, an appropriate choice of the coefficients m_j and n_j leads to well-conditioned linear problems at Step 4 of the algorithm, even for multiple or nearly degenerate eigenvalues. A related algorithm can be found in [9] where the purpose is to build a multigrid eigensolver. Also, the method is designed to get well-conditioned linear systems adapted to a multigrid resolution in the inverse iteration steps.

For small systems, the pseudo-inverse of the matrix $Q_j(A - \Lambda_{jj}B)Q_j$ can be computed and applied to $-Q_j(A - \Lambda_{jj}B)u_j^k$ to find z_j at Step 4. However when A and B are large sparse matrices, the application of the composed operator $Q_j(A - \Lambda_{jj}B)$ on a vector is not expensive and iterative linear solvers are more appropriate to solve (2.1). In addition to the well conditioned linear systems, the iterative resolution is also made easier by the fact that we just look for a small correction z_j that we can approximate by zero at the first iteration.

As presented here, the algorithm BRQI require building and diagonalizing the matrix ${W^k}^T B^{-1} A W^k$ which is assumed to be symmetric. Nevertheless, in practical applications, the inversion of B can be avoided, replacing Step 2 of Algorithm 2.1 with the resolution of the generalized eigenvalue problem

$$ {W^k}^T A W^k X = {W^k}^T B W^k X \Lambda $$

(see §5).

The algorithm BRQI requires a relatively good starting trial subspace (as in the RQI method). If this subspace is not accurate enough, it may converge to another eigenspace corresponding to larger eigenvalues. But BRQI has proved to be efficient for electronic structure calculations. In this case, due to the nonlinearity of the operator, a series of eigenvalue problems has to be solved (one at each step of a fixed point algorithm for an operator that is slightly modified between two steps). Here the solutions of the eigenvalue problem at a given step provide good approximations to start the calculation at the next step.

Compared to the Jacobi-Davidson algorithm [16], where the same kind of projected inverse iteration equations are used, the method described here requires a more precise resolution of better conditioned linear systems, but does not require generating search subspaces of larger dimension than the number of eigenpairs we look for.

2.2. Example. Let us consider the symmetric eigenvalue problem

$$ Au = \lambda u, $$

where

$$ A = \begin{pmatrix} Y_1 & Y_2 & 0 \\ Y_2 & Y_1 & Y_2 \\ 0 & Y_2 & Y_1 \end{pmatrix} \in \mathcal{M}_{27} $$

is defined by

$$ Y_1 = \begin{pmatrix} X_1 & X_2 & 0 \\ X_2 & X_1 & X_2 \\ 0 & X_2 & X_1 \end{pmatrix} \in \mathcal{M}_9,\; Y_2 = \begin{pmatrix} X_2 & 0 & 0 \\ 0 & X_2 & 0 \\ 0 & 0 & X_2 \end{pmatrix} \in \mathcal{M}_9 $$

and

$$ X_1 = \begin{pmatrix} 6 & -1 & 0 \\ -1 & 6 & -1 \\ 0 & -1 & 6 \end{pmatrix},\; X_2 = \begin{pmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{pmatrix}. $$

(The matrix A is obtained for a finite difference discretization of the Laplacian with Dirichlet boundary conditions in 3D.) The first eigenvalues of A are $\lambda_1 = 6 - 3\sqrt{2}, \lambda_2 = \lambda_3 = \lambda_4 = 6 - 2\sqrt{2}, \lambda_5 = \lambda_6 = \lambda_7 = \lambda_8 = \lambda_9 = \lambda_{10} = 6 - \sqrt{2}$. We apply the algorithm BRQI (with $B = I$) to find the four smallest eigenvalues of A. At Step 3 of the algorithm BRQI, we use a subspace U_j^k of dimension 1 for $j = 1$ (containing the vector of index 1 only) and 3 for

$j = 2, 3, 4$ (containing the vectors of indices $2, 3, 4$). The trial eigenvectors at Step $k = 0$ are chosen to be the exact ones plus a random error of small amplitude. The numerical results in Figure 2.1 show the distance from the trial subspace to the exact one (as defined in (1.2)), and the errors on the eigenvalues as a function of the number of iterations. The method's quadratic convergence rate can be observed— note that the errors on the eigenvalues are already within the 15 decimals working precision after the third step.

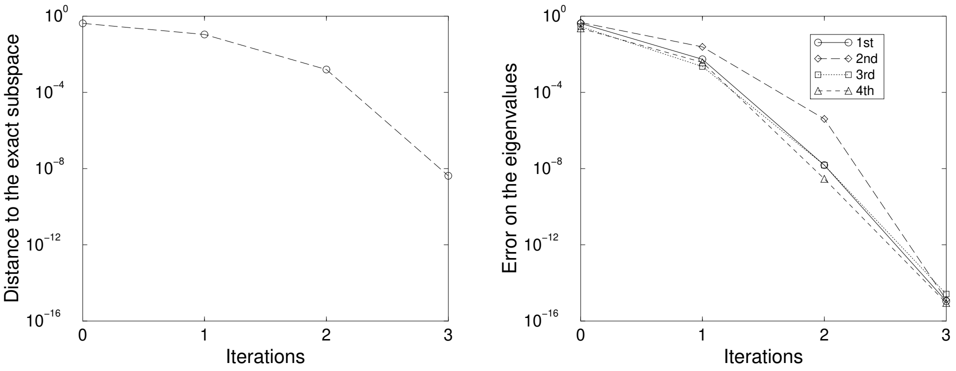

FIG. 2.1. *Distance between the trial subspace and the exact one and errors on the eigenvalues as a function of the iteration number for the example of* §2.2.

3. Some properties of the Ritz elements. In this section, we review the Rayleigh-Ritz algorithm. Then we derive some results on the eigenvectors approximations obtained by this procedure, and on the invariant subspaces approximations spanned by these vectors. These results will be useful in Section 4.

Let:

- $S \in \mathcal{M}_N$ be a symmetric matrix,
- $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_p < \lambda_{p+1} \leq \ldots \leq \lambda_N \in R$ be the eigenvalues of S,
- $\mathcal{E}_0 = \sum_{i=1}^{p} \mathrm{Ker}(S - \lambda_i I)$,
- Π, be the orthogonal projector onto $\mathcal{E}_0$,
- $\mathcal{W} \subset R^N$ be a subspace of dimension p, approximation of $\mathcal{E}_0$, given by $\mathcal{W} = \mathrm{Span}\{w_1, \ldots, w_p\}$, where the vectors w_j, $j = 1, \ldots, p$ are orthonormalized,
- P be the orthogonal projector onto $\mathcal{W}$,
- $\bar{W} = (w_1, \ldots, w_p) \in \mathcal{M}_{N\times p}$.

3.1. Rayleigh-Ritz procedure. We define the Rayleigh-Ritz procedure (see [13] for example) by the following algorithm:

ALGORITHM 3.1 (Rayleigh-Ritz).

(i) *Compute the $p \times p$ symmetric matrix $H = \bar{W}^T S \bar{W}$.*

(ii) *Compute the p orthonormalized eigenvectors $g_j \in R^p$ and the p eigenvalues $\theta_j \in R$, solutions of $Hg_j = g_j\theta_j, j = 1, \ldots, p$.*

(iii) *Compute the p Ritz vectors $y_j = \bar{W} g_j, j = 1 \ldots, p$.*

REMARK 3.1. *The Ritz elements $(\theta_j, y_j), j = 1, \ldots, p$, constructed in Algorithm 3.1, are independent of the chosen orthonormalized basis $\{w_j\}_{j=1}^{p}$ of $\mathcal{W}$. The vectors y_j are orthonormalized and give a basis for $\mathcal{W}$. Moreover, they satisfy the property*

$$P(S - \theta_j)y_j = 0, \quad j = 1, \ldots, p. \tag{3.1}$$

Using this remark with Lemmas A.1 and A.2, we easily prove the following result:

LEMMA 3.2. *Let* $r_j = (S - \theta_j)y_j$ *for* $j = 1, \ldots, p$. *Then, we have*

$$\|r_j\| \le 4\delta(\mathcal{E}_0, \mathcal{W})\|S\|\|y_j\|.$$

With respect to the Ritz values, we have the following lemma (see [13]):

LEMMA 3.3. *There exists an injective application*

$$\beta : \{1, \ldots, p\} \to \{1, \ldots, N\}$$

such that

$$|\theta_j - \lambda_{\beta(j)}| \le \|(I - P)SP\|, \quad j = 1, \ldots, p. \tag{3.2}$$

Moreover, the right hand-side of (3.2) satisfies the following lemma:

LEMMA 3.4.

$$\|(I - P)SP\| \le 4\sqrt{p}\delta(\mathcal{E}_0, \mathcal{W})\|S\|.$$

Proof. We have

$$\begin{aligned} \|(I - P)SP\| &= \sup_{v \in R^N, \|v\|=1} \|(I - P)SPv\| \\ &= \sup_{\eta \in R^p, \|\eta\|=1} \|(I - P)S \sum_{j=1}^{p} \eta_j y_j\| \\ &\le \sup_{\eta \in R^p, \|\eta\|=1} \sum_{j=1}^{p} |\eta_j| \|(I - P)Sy_j\|, \end{aligned}$$

where η_j denotes the component j of η. By property (3.1), we obtain

$$\|(I - P)SP\| \le \sup_{\eta \in R^p, \|\eta\|=1} \sum_{j=1}^{p} |\eta_j| \|Sy_j - \theta_j y_j\|.$$

Applying Lemma 3.2 gives

$$\|(I - P)SP\| \le \left(\sup_{\eta \in R^p, \|\eta\|=1} \sum_{j=1}^{p} |\eta_j| \right) 4\delta(\mathcal{E}_0, \mathcal{W})\|S\|. \tag{3.3}$$

Using the Cauchy-Schwarz inequality, we also have

$$\sum_{j=1}^{p} |\eta_j| \le \sqrt{p} \left(\sum_{j=1}^{p} \eta_j^2 \right)^{1/2} = \sqrt{p}. \tag{3.4}$$

The desired result follows by (3.3) and (3.4). □

ETNA
Kent State University
etna@mcs.kent.edu

The next results will require $\delta(\mathcal{E}_0, \mathcal{W})$ to be small enough. In the following we will assume that:

ASSUMPTION 3.5.

$$\delta(\mathcal{E}_0, \mathcal{W}) < \frac{\lambda_{p+1} - \lambda_p}{8\sqrt{p}\|S\|}.$$

PROPOSITION 3.6. *If Assumption 3.5 holds, then*

$$\lambda_j \leq \theta_j \leq \lambda_j + 4\sqrt{p}\delta(\mathcal{E}_0, \mathcal{W})\|S\|.$$

Proof. The first inequality follows by the Courant-Fischer Theorem. To prove the second one, we use Lemma 3.3 applied to

$$\mathcal{W}(t) = \text{Span}\{w_1(t), \ldots, w_p(t)\},$$

where $w_j(t) = \Pi w_j + t(I - \Pi)w_j$, $j = 1, \ldots, p, 0 \leq t \leq 1$.

Suppose $\bar{W}(t) = (w_1(t), \ldots, w_p(t)) \in \mathcal{M}_{N\times p}$, $H(t) = \bar{W}(t)^T S\bar{W}(t)$, and $\theta_1(t) \leq \theta_2(t) \leq \ldots \leq \theta_p(t)$, be the eigenvalues of $H(t)$. Let

$$\nu = 4\sqrt{p}\delta(\mathcal{E}_0, \mathcal{W})\|S\|.$$

Lemma A.7 gives $\delta(\mathcal{W}(t), \mathcal{E}_0) \leq \delta(\mathcal{W}, \mathcal{E}_0)$, $0 \leq t \leq 1$. From Lemmas 3.3 and 3.4, it follows that there exists p distinct indices j' such that

$$|\theta_j(t) - \lambda_{j'}| \leq \nu, \quad j = 1, \ldots, p, \qquad 0 \leq t \leq 1. \tag{3.5}$$

By Assumption 3.5, we have $\lambda_p + \nu < \lambda_{p+1} - \nu$. From (3.5) we thus obtain

$$\theta_j(t) \notin (\lambda_p + \nu, \lambda_{p+1} - \nu), \quad j = 1, \ldots, p, \qquad 0 \leq t \leq 1. \tag{3.6}$$

For $t = 0$, we are allowed to choose $j' = j$ (because $\mathcal{W}(0) = \mathcal{E}_0$). By the continuity of $\theta_j(t)$ as a function of t, (3.6) gives, using a proof by contradiction,

$$\theta_j(t) \leq \lambda_p + \nu, \quad j = 1, \ldots, p, \qquad 0 \leq t \leq 1.$$

For $0 \leq t \leq 1$, the p indices j' have to be chosen in $\{1, \ldots, p\}$. Using again a proof by contradiction, we clearly see that we can choose $j' = j$ and we thus have

$$\theta_j - \lambda_j \leq \nu$$

in $t = 1$. $\square$

3.2. Distances between invariant subspaces. Let $0 < \delta < \lambda_{p+1} - \lambda_p$ be a given real constant. For $j = 1, \ldots, p$, using (1.3), we define the subspace

$$\mathcal{E}_j = \mathcal{E}_S([\lambda_j - \delta, \lambda_j + \delta]). \tag{3.7}$$

Clearly, we have $\mathcal{E}_j \subset \mathcal{E}_0$. Let Π_j denote the orthogonal projector onto $\mathcal{E}_j$ and let $\sigma(S)$ be the spectrum of S. Let $\triangle > 0$ be such that, for $j = 1, \ldots, p$,

$$(\lambda_j - \delta - \triangle, \lambda_j - \delta) \cap \sigma(S) = \emptyset, \tag{3.8}$$

$$(\lambda_j + \delta, \lambda_j + \delta + \triangle) \cap \sigma(S) = \emptyset. \tag{3.9}$$

Let

$$\tilde{\delta} = \delta + \frac{\triangle}{2} \tag{3.10}$$

and define

$$\mathcal{W}_j = \mathcal{E}_{PS|_{\mathcal{W}}}([\theta_j - \tilde{\delta}, \theta_j + \tilde{\delta}]). \tag{3.11}$$

Let P_j denote the orthogonal projector onto $\mathcal{W}_j$.

To present the main results of this section, let us first make the following sufficient assumption:

ASSUMPTION 3.7.

$$\delta(\mathcal{E}_0, \mathcal{W}) \leq \frac{\triangle}{16\sqrt{p}\|S\|}.$$

REMARK 3.2. *Equation (3.9) for $j = p$ implies in particular that $\triangle < \lambda_{p+1} - \lambda_p$. It follows that if the assumption above is true, Assumption 3.5 will also be true.*

PROPOSITION 3.8. *If Assumption 3.7 is true, we have:*

$$dim(\mathcal{W}_j) = dim(\mathcal{E}_j), \quad j = 1, \ldots, p.$$

The proof of this proposition, relating the dimensions of the exact and approximate invariant subspaces, $\mathcal{E}_j$ and $\mathcal{W}_j$, is given in Appendix B.1.

In [14], an upper bound is given for the angle $\phi(\tilde{u}, u)$ between an eigenvector u of the matrix S, associated with a simple eigenvalue λ, and its approximation $\tilde{u}$. We have the following inequality:

$$\sin \phi(\tilde{u}, u) \leq \frac{\|(S - \tilde{\lambda})\tilde{u}\|}{\epsilon\|\tilde{u}\|},$$

where ϵ denotes the distance between $\tilde{\lambda} = (\tilde{u}, S\tilde{u})/(\tilde{u}, \tilde{u})$ and the remaining part of the spectrum of S, that is, $\epsilon = \min_i\{|\lambda_i - \tilde{\lambda}|, \lambda_i \neq \lambda\}$. This result has been generalized by Knyazev[11] to invariant subspaces of dimension larger than 1. According to [11] (theorem 4.3) and using the notations introduced in this section, we have:

PROPOSITION 3.9. *Let j be a given integer, $1 \leq j \leq p$. If*

$$\tilde{d} = \inf_{\tilde{\nu} \in \sigma(RSR|_{\mathrm{Im}R})} |\tilde{\nu} - \lambda_j| > \delta, \quad R = P - P_j, \tag{3.12}$$

then

$$\|(I - P_j)\Pi_j\|^2 \leq \left[1 + \frac{\|(I-P)SP\|^2}{(\tilde{d} - \delta)^2}\right] \|(I - P)\Pi_j\|^2. \tag{3.13}$$

Applying this proposition gives the following theorem:

THEOREM 3.10. *If Assumption 3.7 holds, then*

$$\delta(\mathcal{E}_j, \mathcal{W}_j) \leq 2\delta(\mathcal{E}_j, \mathcal{W}), j = 1, \ldots, p. \tag{3.14}$$

This is the main result of the section. Theorem 3.10 means that if the trial subspace $\mathcal{W}$ is good enough, the subspaces $\mathcal{W}_j \subset \mathcal{W}, j = 1, \ldots, p$, spanned by the Ritz vectors, will be good approximations of the invariant subspaces $\mathcal{E}_j$. Its proof is given in Appendix B.2.

4. Properties of the algorithm. In this section we present the BRQI algorithm and prove its local quadratic convergence rate. We first show a property of coercivity for the operator which appears in the generalized inverse iteration equations (§4.1). Then we derive some properties of the eigenvectors' corrections obtained in solving these equations (§4.2). Finally we detail the algorithm and give a convergence theorem (§4.3).

Assume that we have a p-dimensional subspace $\mathcal{W} \subset R^N$, that is, a good approximation of $\mathcal{E}_0$. Let P denote the orthogonal projector onto $\mathcal{W}$.

4.1. Coercivity of the inverse iteration operator. Let j be given, $1 \leq j \leq p$, $\tilde{\delta} > 0$ be a given real constant and $\mu_j \in R$, $\lambda_1 \leq \mu_j < \lambda_{p+1}$. Let:

- $\mathcal{W}_j = \mathcal{E}_{PB^{-1}A\big|_{\mathcal{W}}}([\mu_j - \tilde{\delta}, \mu_j + \tilde{\delta}])$,
- P_j, the orthogonal projector onto $\mathcal{W}_j$,
- Q_j, the orthogonal projector onto $(B\mathcal{W}_j)^{\perp}$.

ASSUMPTION 4.1. *We assume that there is a constant $\alpha > 0$ and an invariant subspace of $B^{-1}A$, $\mathcal{E}_j \subset \mathcal{E}_0$, such that:*

$$\mathcal{E}_{B^{-1}A}([\mu_j - \alpha, \mu_j + \alpha]) \subset \mathcal{E}_j, \tag{4.1}$$

$$dim(\mathcal{E}_j) = dim(\mathcal{W}_j). \tag{4.2}$$

Let

$$\gamma = \inf_{\eta \in R} \|I - \eta B\|, \tag{4.3}$$

$$\kappa = \inf_{\eta \in R} \|I - \eta B^2\|, \tag{4.4}$$

and Π_j denote the orthogonal projector onto $\mathcal{E}_j$.

REMARK 4.1. *If the matrix B is symmetric positive definite, we easily see that* $0 \leq \gamma < 1$ *and* $0 \leq \kappa < 1$.

We begin with a Lemma that will be useful.

LEMMA 4.2. *Let $\lambda \in R, c > 0, S \in \mathcal{M}_N$ symmetric, $\mathcal{I} \subset R$ an interval, $x \in \mathcal{E}_S(\mathcal{I})$. Then:*

(i) *If $\mathcal{I} \subset [\lambda - c, \lambda + c]$, then $\|(S - \lambda I)x\| \leq c\|x\|$.*
(ii) *If $\mathcal{I} \cap (\lambda - c, \lambda + c) = \emptyset$, then $\|(S - \lambda I)x\| \geq c\|x\|$.*

In Step 4 of the algorithm 2.1, we have to solve the linear system

$$G_j = Q_j(A - \mu_j B)\Big|_{(B\mathcal{W}_j)^{\perp}} \tag{4.5}$$

where $G_j : (B\mathcal{W}_j)^{\perp} \to (B\mathcal{W}_j)^{\perp}$. This operator has the following property:

PROPOSITION 4.3. *If Assumption 4.1 holds, then there exists strictly positive constants ϵ and C, depending only on B, α and $\lambda_{p+1} - \lambda_1$, such that if $\delta(\mathcal{E}_j, \mathcal{W}_j) < \epsilon$, then*

$$\|G_j x\| \geq C\|x\|, \ \forall x \in (B\mathcal{W}_j)^{\perp}.$$

This proposition is easy to prove when B is the identity matrix. In Appendix B.3, we give the (rather technical) proof in the general case.

4.2. The generalized inverse iteration. Let $(\theta_j, u_j) \in R \times R^N$, $j = 1, \ldots, p$, denote the Ritz elements for the symmetric matrix $B^{-1}A$ in the subspace $\mathcal{W}$ (see §3.1). As in §3.2, we choose δ and $\triangle \in R$, $0 < \delta < \lambda_{p+1} - \lambda_p$, $0 < \triangle < \lambda_{p+1} - \lambda_p$ such that

$$(\lambda_j - \delta - \triangle, \lambda_j - \delta) \cap \sigma(B^{-1}A) = \emptyset, \tag{4.6}$$

$$(\lambda_j + \delta, \lambda_j + \delta + \triangle) \cap \sigma(B^{-1}A) = \emptyset, \tag{4.7}$$

for $j = 1, \ldots, p$, where $\sigma(B^{-1}A)$ denotes the spectrum of $B^{-1}A$. Moreover, we impose here

$$0 < \triangle \leq 2\delta. \tag{4.8}$$

In the following, we define:

$$\mathcal{E}_j = \mathcal{E}_{B^{-1}A}([\lambda_j - \delta, \lambda_j + \delta]), \quad j = 1, \ldots, p, \tag{4.9}$$

$$\tilde{\delta} = \delta + \frac{\triangle}{2}, \tag{4.10}$$

$$\mathcal{W}_j = \mathcal{E}_{PB^{-1}A\big|_{\mathcal{W}}}([\theta_j - \tilde{\delta}, \theta_j + \tilde{\delta}]), \quad j = 1, \ldots, p. \tag{4.11}$$

Let P_j denote the orthogonal projector onto $\mathcal{W}_j$, and Q_j denote the orthogonal projector onto $(B\mathcal{W}_j)^{\perp}$.

In order to apply the results of Section 3.2, we will assume that $\delta(\mathcal{E}_0, \mathcal{W})$ satisfies:

ASSUMPTION 4.4.

$$\delta(\mathcal{E}_0, \mathcal{W}) \leq \frac{\triangle}{16\sqrt{p}\|B^{-1}A\|}.$$

Under Assumption 4.4, and using Remark 3.2, Proposition 3.6 gives

$$|\theta_j - \lambda_j| \leq 4\sqrt{p}\delta(\mathcal{E}_0, \mathcal{W})\|B^{-1}A\| \leq \frac{\triangle}{4}, \quad j = 1, \ldots, p.$$

By (4.8), we thus have

$$|\theta_j - \lambda_j| \leq \frac{\delta}{2}, \quad j = 1, \ldots, p,$$

that implies

$$\mathcal{E}_{B^{-1}A}([\theta_j - \delta/2, \theta_j + \delta/2]) \subset \mathcal{E}_{B^{-1}A}([\lambda_j - \delta, \lambda_j + \delta]) = \mathcal{E}_j \tag{4.12}$$

for $j = 1, \ldots, p$.

By (4.12) and Proposition 3.8, Assumption 4.1 holds if $\mu_j = \theta_j$ and $\alpha = \delta/2$. In particular it gives

$$\dim(\mathcal{E}_j) = \dim(\mathcal{W}_j). \tag{4.13}$$

In this context, using Theorem 3.10 gives

$$\delta(\mathcal{E}_j, \mathcal{W}_j) \leq 2\delta(\mathcal{E}_j, \mathcal{W}) \leq 2\delta(\mathcal{E}_0, \mathcal{W}).$$

We then rewrite Proposition 4.3 as follows.

PROPOSITION 4.5. *Suppose that Assumption 4.4 holds. Then there exist strictly positive constants ϵ_c and C_c depending only on B, δ and $\lambda_{p+1} - \lambda_1$, such that, if $\delta(\mathcal{E}_0, \mathcal{W}) < \epsilon_c$,*

$$\|Q_j(A - \theta_j B)x\| \geq C_c\|x\|, \; \forall x \in (B\mathcal{W}_j)^{\perp}.$$

This proposition ensures that we can define $z_j \in (B\mathcal{W}_j)^{\perp}$, $j = 1, \ldots, p$, as the only solution of the linear problem

$$Q_j(A - \theta_j B)(u_j + z_j) = 0, \tag{4.14}$$

ETNA
Kent State University
etna@mcs.kent.edu

provided $\delta(\mathcal{E}_0, \mathcal{W})$ is sufficiently small. Equation (4.14) is a generalization of a RQI iteration (see §2.1) whose purpose is to improve the approximation u_j of the j^{th} eigenvector by the correction z_j. The following proposition gives a few properties of z_j and $u_j + z_j$.

PROPOSITION 4.6. *Let $1 \leq j \leq p$. Suppose that Assumption 4.4 holds and that $\delta(\mathcal{E}_0, \mathcal{W}) < \epsilon_c$ for the constant ϵ_c of Proposition 4.5. Then, for z_j solution of (4.14),*

$$\|z_j\| \leq 2\tau C_c^{-1} \beta \delta(\mathcal{E}_0, \mathcal{W}), \tag{4.15}$$

$$\|(I - \Pi_j)(u_j + z_j)\| \leq \sigma \left(\delta(\mathcal{E}_0, \mathcal{W})\right)^2, \tag{4.16}$$

where

$$\tau = 2\|B\|\|B^{-1}A\|, \tag{4.17}$$

$$\beta = \|B^{-1}\|\|B\| + 1, \tag{4.18}$$

$$\sigma = \frac{4\tau\|B^{-1}\|}{\delta}\left(1 + C_c^{-1}\beta\tau\right), \tag{4.19}$$

and C_c is the same constant as in Proposition 4.5.

This theorem is a key element in proving the local quadratic convergence of the algorithm BRQI. It shows that the updated approximations $u_j + z_j, j = 1, \ldots, p$ have only a second order component orthogonal to $\mathcal{E}_0$, after having been corrected by a first order component z_j.

Proof. In this proof, we will regularly use Lemma A.1 and the fact that $\|u_j\| = 1$.

We decompose $u_j = \Pi_j u_j + (I - \Pi_j)u_j$. Then

$$\begin{aligned} \|Q_j(A - \theta_j B)u_j\| \leq\; & \|Q_j(A - \theta_j B)\Pi_j u_j\| \\ & + \|Q_j(A - \theta_j B)(I - \Pi_j)u_j\|. \end{aligned} \tag{4.20}$$

Note that $(A - \theta_j B)\Pi_j u_j \in B\mathcal{E}_j$, and so we obtain by Lemma A.5,

$$\begin{aligned} \|Q_j(A - \theta_j B)\Pi_j u_j\| &\leq \left(\|A\| + |\theta_j|\|B\|\right)\delta(B\mathcal{E}_j, B\mathcal{W}_j) \\ &\leq \left(\|A\| + |\theta_j|\|B\|\right)\|B^{-1}\|\|B\|\delta(\mathcal{E}_j, \mathcal{W}_j). \end{aligned} \tag{4.21}$$

On the other hand, because

$$\|(I - \Pi_j)u_j\| = \|(I - \Pi_j)P_j u_j\| \leq \delta(\mathcal{W}_j, \mathcal{E}_j),$$

it follows by Lemma A.2 and (4.13) that

$$\begin{aligned} \|Q_j(A - \theta_j B)(I - \Pi_j)u_j\| &\leq \left(\|A\| + |\theta_j|\|B\|\right)\delta(\mathcal{W}_j, \mathcal{E}_j) \\ &= \left(\|A\| + |\theta_j|\|B\|\right)\delta(\mathcal{E}_j, \mathcal{W}_j). \end{aligned} \tag{4.22}$$

By (4.20)–(4.22), we have

$$\begin{aligned} \|Q_j(A - \theta_j B)u_j\| &\leq \left(\|A\| + |\theta_j|\|B\|\right)\left(\|B^{-1}\|\|B\| + 1\right)\delta(\mathcal{E}_j, \mathcal{W}_j) \\ &\leq 2\|B\|\|B^{-1}A\|\left(\|B^{-1}\|\|B\| + 1\right)\delta(\mathcal{E}_j, \mathcal{W}_j). \end{aligned} \tag{4.23}$$

Applying Theorem 3.10 gives

$$\begin{aligned} \|Q_j(A - \theta_j B)u_j\| &\leq 4\|B\|\|B^{-1}A\|\left(\|B^{-1}\|\|B\| + 1\right)\delta(\mathcal{E}_j, \mathcal{W}) \\ &\leq 4\|B\|\|B^{-1}A\|\left(\|B^{-1}\|\|B\| + 1\right)\delta(\mathcal{E}_0, \mathcal{W}). \end{aligned}$$

Since z_j is solution of the equation

$$Q_j(A - \theta_j B)z_j = -Q_j(A - \theta_j B)u_j,$$

we obtain the inequality (4.15) by using Proposition 4.5.

To prove the second inequality, we first define $\left(B^{-1}A-\theta_j\right)_*$ as the restriction of the operator $\left(B^{-1}A-\theta_j\right)$ to $\mathcal{E}_j^{\perp}$, invariant subspace of $B^{-1}A$. There exists an inverse of $\left(B^{-1}A-\theta_j\right)_*$, denoted $\left(B^{-1}A-\theta_j\right)_*^{-1}$, whose norm is bounded by $2/\delta$ (see Eq. (4.12) and Lemma 4.2). Using the fact that z_j is solution of (4.14), we obtain

$$\begin{aligned}
&(I-\Pi_j)(u_j+z_j)\\
&\quad=\left(B^{-1}A-\theta_j\right)_*^{-1}\left(B^{-1}A-\theta_j\right)_*(I-\Pi_j)(u_j+z_j)\\
&\quad=\left(B^{-1}A-\theta_j\right)_*^{-1}(I-\Pi_j)\left(B^{-1}A-\theta_j\right)(u_j+z_j)\\
&\quad=\left(B^{-1}A-\theta_j\right)_*^{-1}(I-\Pi_j)B^{-1}(Q_j+(I-Q_j))\,(A-\theta_jB)\,(u_j+z_j)\\
&\quad=\left(B^{-1}A-\theta_j\right)_*^{-1}(I-\Pi_j)B^{-1}(I-Q_j)\,(A-\theta_jB)\,(u_j+z_j).
\end{aligned}$$

Consequently

$$\begin{aligned}
&\|(I-\Pi_j)(u_j+z_j)\|\\
&\quad=\|\left(B^{-1}A-\theta_j\right)_*^{-1}(I-\Pi_j)B^{-1}(I-Q_j)\,(A-\theta_jB)\,(u_j+z_j)\|\\
&\quad\leq\frac{2}{\delta}\|(I-\Pi_j)B^{-1}(I-Q_j)\|\left(\|\,(A-\theta_jB)\,u_j\|+\|\,(A-\theta_jB)\,z_j\|\right).
\end{aligned}\tag{4.24}$$

From the definition of $(I-Q_j)$, $B^{-1}(I-Q_j)x\in\mathcal{W}_j\quad\forall x\in R^N$, it follows that

$$\|(I-\Pi_j)B^{-1}(I-Q_j)\|\leq\delta(\mathcal{W}_j,\mathcal{E}_j)\|B^{-1}\|=\delta(\mathcal{E}_j,\mathcal{W}_j)\|B^{-1}\|.\tag{4.25}$$

Moreover, given the Ritz pair (θ_j,u_j), Lemma 3.2 implies

$$\begin{aligned}
\|\,(A-\theta_jB)\,u_j\|&\leq\|B\|\|\left(B^{-1}A-\theta_j\right)u_j\|\\
&\leq\|B\|4\delta(\mathcal{E}_0,\mathcal{W})\|B^{-1}A\|\\
&=2\tau\delta(\mathcal{E}_0,\mathcal{W}).
\end{aligned}\tag{4.26}$$

Using (4.15), we also have

$$\|\,(A-\theta_jB)\,z_j\|\leq\left(2\|B\|\|B^{-1}A\|\right)\|z_j\|\leq 2C_c^{-1}\beta\tau^2\delta(\mathcal{E}_0,\mathcal{W}).\tag{4.27}$$

Now it follows by (4.24)–(4.27), and Theorem 3.10, that inequality (4.16) holds. □

PROPOSITION 4.7. *Supposing that Assumption 4.4 holds and that $\delta(\mathcal{E}_0,\mathcal{W})<\epsilon_c$ for the constant ϵ_c given in Proposition 4.5, let*

$$\mathcal{W}^{new}=\mathit{Span}\{u_1+z_1,\ldots,u_p+z_p\}$$

for z_j solution of Equation (4.14), $j=1,\ldots,p$. Then there exist constants $\epsilon_q>0$ and $\chi<1$, independent of $\mathcal{W}$, such that if $\delta(\mathcal{E}_0,\mathcal{W})<\epsilon_q$, then

$$\mathit{dim}(\mathcal{W}^{new})=\mathit{dim}(\mathcal{W}),\tag{4.28}$$

$$\delta(\mathcal{E}_0,\mathcal{W}^{new})\leq\varrho\left(\delta(\mathcal{E}_0,\mathcal{W})\right)^2,\tag{4.29}$$

$$\delta(\mathcal{E}_0,\mathcal{W}^{new})\leq\chi\delta(\mathcal{E}_0,\mathcal{W}),\tag{4.30}$$

where $\varrho=2\sqrt{p}\sigma$, for σ defined by (4.19).

The proof of this proposition, based on Proposition 4.6, is given in Appendix B.4.

4.3. A convergence theorem. Using the subspace notations and the mathematical tools developed in the previous sections, the algorithm BRQI described in §2.1 can be written as:

ALGORITHM 4.8. *BRQI*

1. *Let $\mathcal{W}^0 \subset R^N$ be a given subspace of dimension p. Let $k = 0$.*
2. *Build the pairs $(\theta_j, u_j), j = 1, \dots, p$ by the Rayleigh-Ritz procedure (Algorithm 3.1) for the matrix $B^{-1}A$ in $\mathcal{W} = \mathcal{W}^k$.*
3. *For $j = 1, \dots, p$, define the subspaces $\mathcal{W}_j$ according to (4.11) for $\mathcal{W} = \mathcal{W}^k$.*
4. *For $j = 1, \dots, p$, compute z_j solution of Eq. (4.14).*
5. *Let $\mathcal{W}^{k+1} = Span\{u_1 + z_1, \dots, u_p + z_p\}$. Increment k by 1 and go to step 2).*

For this algorithm, we have the following local convergence result:

THEOREM 4.9. *There exist constants $\epsilon_0 > 0$, ϱ, $\chi < 1$, $C_\theta > 0$ such that, if $\delta(\mathcal{E}_0, \mathcal{W}^0) < \epsilon_0$, Algorithm 4.8 is well defined and the following properties hold for $k = 0, 1, 2, \dots$,*

$$dim(\mathcal{W}^k) = p, \tag{4.31}$$

$$\delta(\mathcal{E}_0, \mathcal{W}^{k+1}) \leq \varrho \left(\delta(\mathcal{E}_0, \mathcal{W}^k)\right)^2, \tag{4.32}$$

$$\delta(\mathcal{E}_0, \mathcal{W}^{k+1}) \leq \chi \delta(\mathcal{E}_0, \mathcal{W}^k), \tag{4.33}$$

$$|\theta_j - \lambda_j| \leq C_\theta \delta(\mathcal{E}_0, \mathcal{W}^k), j = 1, \dots, p. \tag{4.34}$$

Moreover, the algorithm converges, that is:

$$\lim_{k \to \infty} \delta(\mathcal{E}_0, \mathcal{W}^k) = 0.$$

Proof. Relations (4.31), (4.32), (4.33) follow by Proposition 4.7. The inequality (4.34) follows by Proposition 3.6 for $C_\theta = 4\sqrt{p}\|B^{-1}A\|$. □

5. Concluding remarks. A straightforward implementation of Algorithm 4.8 is not always obvious or even adequate for large-scale eigenvalue problems. First, Step 2 implies the use of the matrix $B^{-1}A$, hence the need to solve numerous linear systems with the matrix B. This can be done efficiently if the matrix B is very well conditioned (common in practical applications) or can be efficiently factored. But replacing the Rayleigh-Ritz procedure at Step 2 by a Petrov-Galerkin approach using $B\mathcal{W}$ as test (or left) subspace instead of $\mathcal{W}$, is often easier and less expensive, leading to a generalized eigenvalue problem of dimension p (see §2.1)

$$Hg_j = \theta_j G g_j.$$

This approach (that we call Block Galerkin Inverse Iteration - BGII [4]) can work quite well in practice. But since here $G^{-1}H$ is not symmetric in general and can generate complex eigenvalues θ_j, the proof of convergence presented in this paper is no longer valid—and probably much more difficult to establish in this case.

On the other hand, Algorithm 4.8 depends, by (4.11), on the quantity $\tilde{\delta}$. Moreover, the assumptions on $\delta(\mathcal{E}_0, \mathcal{W})$ can be difficult to satisfy in practice, depending on the value of $\triangle$ that has to satisfy (4.8), (4.6) and (4.7). In practice $\triangle$ can be quite small depending on the spectrum of $B^{-1}A$ and the choice of δ. This remains an issue, even if we replace δ by coefficients δ_j^- and δ_j^+ depending on j and define

$$\mathcal{E}_j = \mathcal{E}_{B^{-1}A}([\lambda_j - \delta_j^-, \lambda_j + \delta_j^+]), j = 1, \dots, p.$$

An appropriate choice of δ_j^- and δ_j^+, $j = 1, \dots, p$ would then allow to replace Assumption 4.4 by a less restricting one.

In practice, we choose a coefficient $\tilde{\delta}$ so that the linear systems are well-defined but also so that projectors Q_j are not expensive to apply. This approach, in combination with the Petrov-Galerkin version of the algorithm, has given good results for quantum physics problems [4, 6, 7, 8]. This approach should also work well on other eigenvalue problems.

Appendix A. Some relations between subspaces of R^N.

In this appendix, we give some technical lemmas concerning the properties of $\delta(\mathcal{V}, \mathcal{W})$, the distance from one subspace $\mathcal{V}$ to another subspace $\mathcal{W}$ (see (1.2)). Their proofs are not difficult to establish or are given in references.

LEMMA A.1. *Let $\mathcal{V}$ and $\mathcal{W}$ be two subspaces of R^N, P and Q being the associated orthogonal projectors. Then, we have*

$$\delta(\mathcal{V}, \mathcal{W}) = \|(I - Q)P\|. \tag{A.1}$$

LEMMA A.2. *Let $\mathcal{V}$ and $\mathcal{W}$ be two subspaces of R^N of the same dimension, P and Q the associated orthogonal projectors. Then, we have*

$$\delta(\mathcal{V}, \mathcal{W}) = \delta(\mathcal{W}, \mathcal{V}) = \|P - Q\|. \tag{A.2}$$

Moreover, if $\|P - Q\| < 1$, then $P\big|_{\mathcal{W}}$ defines a bijection between $\mathcal{W}$ and $\mathcal{V}$.

Proof. See [10]. □

LEMMA A.3. *Let $\mathcal{V}$ and $\mathcal{W}$ be two subspaces of R^N, $\mathcal{V}^\perp$ and $\mathcal{W}^\perp$ be their orthogonal complement in R^N. Then*

$$\delta(\mathcal{V}^\perp, \mathcal{W}^\perp) = \delta(\mathcal{W}, \mathcal{V}).$$

Proof. See [10]. □

LEMMA A.4. *Let $A \in \mathcal{M}_N$ be a regular matrix, $\mathcal{V}$ be a subspace of R^N. Then, we have*

$$\delta(\mathcal{V}, A\mathcal{V}) \leq \min_{\eta \in R} \|I - \eta A\|. \tag{A.3}$$

LEMMA A.5. *Let $A \in \mathcal{M}_N$ be a regular matrix, $\mathcal{V}$ and $\mathcal{W}$ be two subspaces of R^N of the same dimension. Then, we have*

$$\delta(A\mathcal{V}, A\mathcal{W}) \leq \delta(\mathcal{V}, \mathcal{W})\|A\|\|A^{-1}\|. \tag{A.4}$$

LEMMA A.6. *Let $A \in \mathcal{M}_N$ be a symmetric, positive definite matrix, $\mathcal{V}$ be a subspace of R^N. Then, we have*

$$A\mathcal{V}^\perp = (A^{-1}\mathcal{V})^\perp. \tag{A.5}$$

LEMMA A.7. *Let $\mathcal{V}$ and $\mathcal{W} = Span\{w_1, \ldots, w_p\}$ be two subspaces of R^N of dimension p, P be the orthogonal projector onto $\mathcal{V}$. We assume $\delta(\mathcal{W}, \mathcal{V}) < 1$. Let*

$$\mathcal{W}(t) = Span\{Pw_1 + t(I - P)w_1, \ldots, Pw_p + t(I - P)w_p\}.$$

Then the dimension of $\mathcal{W}(t)$ is p and $\delta(\mathcal{W}(t), \mathcal{V}) \leq \delta(\mathcal{W}, \mathcal{V})$, $0 \leq t \leq 1$.

Appendix B. Some proofs.

B.1. Proof of Proposition 3.8. *Proof*. Let j be a given integer, $1 \leq j \leq p$. By Proposition 3.6, we have

$$\lambda_i \leq \theta_i \leq \lambda_i + 4\sqrt{p}\delta(\mathcal{E}_0, \mathcal{W})\|S\|, \quad i = 1, \dots, p.$$

Using Assumption 3.7 yields

$$\lambda_i \leq \theta_i \leq \lambda_i + \frac{\triangle}{4}, \quad i = 1, \dots, p.$$

Let k be such that $\lambda_k \in [\lambda_j - \delta, \lambda_j + \delta]$. We then have

$$\theta_j - \tilde{\delta} = \theta_j - \delta - \frac{\triangle}{2} < \lambda_j - \delta - \frac{\triangle}{4} < \lambda_k \leq \theta_k \leq \lambda_k + \frac{\triangle}{4}$$
$$\leq \lambda_j + \delta + \frac{\triangle}{4} \leq \theta_j + \delta + \frac{\triangle}{4} < \theta_j + \delta + \frac{\triangle}{2} = \theta_j + \tilde{\delta}.$$

We thus have

$$\theta_k \in (\theta_j - \tilde{\delta}, \theta_j + \tilde{\delta}). \tag{B.1}$$

Let k be such that $\lambda_k \in \quad (-\infty, \lambda_j - \delta - \triangle]$. It follows that

$$\theta_k \leq \lambda_k + \frac{\triangle}{4} \leq \lambda_j - \delta - \triangle + \frac{\triangle}{4}$$
$$= \lambda_j - \delta - \frac{3\triangle}{4} < \lambda_j - \delta - \frac{\triangle}{2} \leq \theta_j - \delta - \frac{\triangle}{2} = \theta_j - \tilde{\delta}.$$

We thus have

$$\theta_k \notin [\theta_j - \tilde{\delta}, \theta_j + \tilde{\delta}]. \tag{B.2}$$

Let $k \leq p$ be such that $\lambda_k \in [\lambda_j + \delta + \triangle, \infty)$. It follows that

$$\theta_j + \tilde{\delta} \leq \lambda_j + \frac{\triangle}{4} + \delta + \frac{\triangle}{2} < \lambda_j + \delta + \triangle \leq \lambda_k \leq \theta_k.$$

We thus have

$$\theta_k \notin [\theta_j - \tilde{\delta}, \theta_j + \tilde{\delta}]. \tag{B.3}$$

Relations (3.8), (3.9), (B.1), (B.2) and (B.3) imply then

$$\theta_i \in [\theta_j - \tilde{\delta}, \theta_j + \tilde{\delta}] \Leftrightarrow \lambda_i \in [\lambda_j - \delta, \lambda_j + \delta], \quad i = 1, \dots, p.$$

We conclude by noting that dim$(\mathcal{W}_j)$ (respectively dim$(\mathcal{E}_j)$) is equal to the number of eigenvalues (according to their multiplicities) of $PS\big|_{\mathcal{W}}$ (respectively S) in $[\theta_j - \tilde{\delta}, \theta_j + \tilde{\delta}]$ (respectively $[\lambda_j - \delta, \lambda_j + \delta]$). ☐

B.2. Proof of Theorem 3.10. *Proof*. Let us first compute the term $(\tilde{d}-\delta)^2$ in equation (3.13). By using (3.11), we obtain

$$(\tilde{d}-\delta)^2 \geq min((|\theta_j-\delta-\triangle/2-\lambda_j|-\delta)^2,(|\theta_j+\delta+\triangle/2-\lambda_j|-\delta)^2). \tag{B.4}$$

Proposition 3.6 and Assumption 3.7 yield

$$0 \leq \theta_j-\lambda_j \leq \triangle/4.$$

We thus have

$$\theta_j-\delta-\triangle/2-\lambda_j < 0$$

and

$$\theta_j+\delta+\triangle/2-\lambda_j > 0.$$

The inequality (B.4) thus yields

$$\begin{aligned}(\tilde{d}-\delta)^2 &\geq min((-\theta_j+\triangle/2+\lambda_j)^2,(\theta_j-\lambda_j+\triangle/2)^2)\\ &\geq min((\triangle/4)^2,(\triangle/2)^2)=(\triangle/4)^2>0.\end{aligned}$$

Since (3.12) holds, we can apply Proposition 3.9. Since by Lemma 3.4 and Assumption 3.7,

$$\|(I-P)SP\| \leq \triangle/4,$$

we obtain

$$\|(I-P_j)\Pi_j\|^2 \leq 2\|(I-P)\Pi_j\|^2. \tag{B.5}$$

By Lemma A.1, we have

$$\|(I-P_j)\Pi_j\| = \delta(\mathcal{E}_j,\mathcal{W}_j) \tag{B.6}$$

and

$$\|(I-P)\Pi_j\| = \delta(\mathcal{E}_j,\mathcal{W}). \tag{B.7}$$

The relations (B.6) and (B.7), used with (B.5), and Proposition 3.8, complete the proof. □

B.3. Proof of Proposition 4.3. *Proof*. Let $x \in (B\mathcal{W}_j)^\perp$, $z=G_jx$. Since $x=Q_jx$, we have, using Lemma A.1,

$$\|\Pi_jx\| = \|\Pi_jQ_jx\| \leq \delta((B\mathcal{W}_j)^\perp,\mathcal{E}_j^\perp)\|x\|.$$

By Lemmas A.3, A.4, using (4.2),(4.3), and the triangle inequality for the distance $\delta(.,.)$ between subspaces of the same dimension, we get

$$\begin{aligned}\|\Pi_jx\| &\leq \delta(\mathcal{E}_j,B\mathcal{W}_j)\|x\| \leq (\delta(\mathcal{E}_j,\mathcal{W}_j)+\delta(\mathcal{W}_j,B\mathcal{W}_j))\|x\|\\ &\leq (\delta(\mathcal{E}_j,\mathcal{W}_j)+\inf_{\eta\in R}\|I-\eta B\|)\|x\| = (\delta(\mathcal{E}_j,\mathcal{W}_j)+\gamma)\|x\|,\end{aligned} \tag{B.8}$$

while the reverse triangle inequality yields

$$\|(I-\Pi_j)x\| \geq \|x\|-\|\Pi_jx\| \geq (1-\delta(\mathcal{E}_j,\mathcal{W}_j)-\gamma)\|x\|. \tag{B.9}$$

On the other hand, we have

$$z = Q_j(A - \mu_j B)x = Q_j(A - \mu_j B)(I - \Pi_j + \Pi_j)x = Q_j y_1 + Q_j y_2,$$

with $y_1 = (A - \mu_j B)(I - \Pi_j)x$ and $y_2 = (A - \mu_j B)\Pi_j x$. Still using the reverse triangle inequality, we thus obtain

$$\|z\| \geq \|Q_j y_1\| - \|Q_j y_2\|. \tag{B.10}$$

We note that $y_2 \in B\mathcal{E}_j$, which leads, by Lemmas A.1 and A.5, to

$$\|Q_j y_2\| \leq \delta(B\mathcal{E}_j, B\mathcal{W}_j)\|y_2\| \leq \delta(\mathcal{E}_j, \mathcal{W}_j)\|B\|\|B^{-1}\|\|y_2\|. \tag{B.11}$$

Moreover, we have,

$$\begin{aligned}\|y_2\| &= \|(A - \mu_j B)\Pi_j x\| = \|B(B^{-1}A - \mu_j)\Pi_j x\| \\ &\leq \|B\|\|(B^{-1}A - \mu_j)\Pi_j x\|.\end{aligned}$$

Since $\Pi_j x \in \mathcal{E}_0$ and $\lambda_1 \leq \mu_j < \lambda_{p+1}$, we obtain

$$\begin{aligned}\|y_2\| &\leq \|B\| \max(\lambda_{p+1} - \mu_j, \mu_j - \lambda_1)\|\Pi_j x\| \\ &\leq \|B\|(\lambda_{p+1} - \lambda_1)\|\Pi_j x\|.\end{aligned} \tag{B.12}$$

From (B.8), (B.11), (B.12), it follows

$$\|Q_j y_2\| \leq \delta(\mathcal{E}_j, \mathcal{W}_j)\|B^{-1}\|\|B\|^2(\lambda_{p+1} - \lambda_1)(\delta(\mathcal{E}_j, \mathcal{W}_j) + \gamma)\|x\|. \tag{B.13}$$

Concerning y_1, we have

$$\|Q_j y_1\| \geq \|y_1\| - \|(I - Q_j)y_1\|. \tag{B.14}$$

We also have

$$y_1 = B(B^{-1}A - \mu_j)(I - \Pi_j)x \in B\mathcal{E}_j^{\perp},$$

which implies

$$\|(I - Q_j)y_1\| \leq \delta(B\mathcal{E}_j^{\perp}, (B\mathcal{W}_j)^{\perp})\|y_1\|. \tag{B.15}$$

By Lemmas A.3 and A.6,

$$\begin{aligned}\delta(B\mathcal{E}_j^{\perp}, (B\mathcal{W}_j)^{\perp}) &= \delta((B^{-1}\mathcal{E}_j)^{\perp}, (B\mathcal{W}_j)^{\perp}) = \delta(B\mathcal{W}_j, B^{-1}\mathcal{E}_j) \\ &\leq \left(\delta(B\mathcal{W}_j, B\mathcal{E}_j) + \delta(B\mathcal{E}_j, B^{-1}\mathcal{E}_j)\right) \\ &\leq \left(\delta(\mathcal{W}_j, \mathcal{E}_j)\|B\|\|B^{-1}\|) + \inf_{\eta \in R}\|I - \eta B^2\|\right) \\ &= \left(\delta(\mathcal{W}_j, \mathcal{E}_j)\|B\|\|B^{-1}\| + \kappa\right);\end{aligned}$$

thus,

$$\|(I - Q_j)y_1\| \leq (\kappa + \delta(\mathcal{W}_j, \mathcal{E}_j)\|B\|\|B^{-1}\|)\|y_1\|. \tag{B.16}$$

Moreover, by (4.1) and Lemma 4.2, we have

$$\begin{aligned}\|y_1\| &= \|B(B^{-1}A - \mu_j)(I - \Pi_j)x\| \\ &\geq \|B^{-1}\|^{-1}\|(B^{-1}A - \mu_j)(I - \Pi_j)x\| \\ &\geq \|B^{-1}\|^{-1}\alpha\|(I - \Pi_j)x\|.\end{aligned}$$

Hence, by (B.9), we get

$$\|y_1\| \geq \|B^{-1}\|^{-1}\alpha(1-\delta(\mathcal{E}_j,\mathcal{W}_j)-\gamma)\|x\|. \tag{B.17}$$

From (B.14), (B.16), and (B.17), we obtain, for $\delta(\mathcal{W}_j,\mathcal{E}_j)$ sufficiently small,

$$\begin{aligned}\|Q_j y_1\| &\geq (1-\kappa-\delta(\mathcal{W}_j,\mathcal{E}_j)\|B\|\|B^{-1}\|)\|y_1\| \\ &\geq (1-\kappa-\delta(\mathcal{W}_j,\mathcal{E}_j)\|B\|\|B^{-1}\|) \\ &\quad \cdot\|B^{-1}\|^{-1}\alpha(1-\delta(\mathcal{E}_j,\mathcal{W}_j)-\gamma)\|x\|.\end{aligned} \tag{B.18}$$

Finally, from (B.10), (B.13) and (B.18), we get, for $\delta(\mathcal{W}_j,\mathcal{E}_j)$ sufficiently small,

$$\begin{aligned}\|z\| \geq \big(\ &(1-\kappa-\delta(\mathcal{W}_j,\mathcal{E}_j)\|B\|\|B^{-1}\|)\|B^{-1}\|^{-1}\alpha(1-\delta(\mathcal{E}_j,\mathcal{W}_j)-\gamma) \\ &-\delta(\mathcal{E}_j,\mathcal{W}_j)\|B^{-1}\|\|B\|^2(\lambda_{p+1}-\lambda_1)(\delta(\mathcal{E}_j,\mathcal{W}_j)+\gamma)\big)\|x\|.\end{aligned} \tag{B.19}$$

Now (4.2) implies that $\mathcal{E}_j$ and $\mathcal{W}_j$ have same dimension and $\delta(\mathcal{E}_j,\mathcal{W}_j)=\delta(\mathcal{W}_j,\mathcal{E}_j)$. Proposition 4.3 is thus a direct consequence of (B.19). □

B.4. Proof of Proposition 4.7. *Proof.* Let $\eta \in R^p$, be a vector of components $\eta_j, j = 1,\dots,p$, such that $\|\eta\|=1$. Assuming that the vectors $u_j, j=1,\dots,p$, are orthonormal, we have

$$\begin{aligned}\|\sum_{j=1}^p \eta_j(u_j+z_j)\| &\geq \|\sum_{j=1}^p \eta_j u_j\| - \|\sum_{j=1}^p \eta_j z_j\| \\ &\geq 1-\max_{j=1,\dots,p}\|z_j\|\sum_{j=1}^p|\eta_j|.\end{aligned} \tag{B.20}$$

Using the Cauchy-Schwarz inequality, we have

$$\sum_{j=1}^p |\eta_j| \leq \sqrt{p}\left(\sum_{j=1}^p \eta_j^2\right)^{1/2} = \sqrt{p}. \tag{B.21}$$

By Proposition 4.6, for $\delta(\mathcal{E}_0,\mathcal{W})$ sufficiently small, we can assume $\|z_j\| \leq (2\sqrt{p})^{-1}$. By (B.20), we thus have

$$\|\sum_{j=1}^p \eta_j(u_j+z_j)\| \geq 1-\frac{1}{2\sqrt{p}}\sqrt{p} = \frac{1}{2}. \tag{B.22}$$

Since (B.22) is true for all normalized $\eta \in R^p$, we obtain that the vectors $u_j+z_j, j=1,\dots,p$ are linearly independent and (4.28) holds.

On the other hand, we have

$$\begin{aligned}\delta(\mathcal{W}^{new},\mathcal{E}_0) &= \max_{v\in\mathcal{W}^{new}}\frac{\|(I-\Pi)v\|}{\|v\|} \\ &= \max_{\eta\in R^p,\|\eta\|=1}\frac{\|(I-\Pi)\sum_{j=1}^p\eta_j(u_j+z_j)\|}{\|\sum_{j=1}^p\eta_j(u_j+z_j)\|}.\end{aligned} \tag{B.23}$$

By (B.21), we have

$$\begin{aligned}\|(I-\Pi)\sum_{j=1}^p\eta_j(u_j+z_j)\| &\leq \max_{j=1,\dots,p}\|(I-\Pi)(u_j+z_j)\|\sum_{j=1}^p|\eta_j| \\ &\leq \max_{j=1,\dots,p}\|(I-\Pi)(u_j+z_j)\|\sqrt{p}.\end{aligned} \tag{B.24}$$

From (B.22), (B.23), and (B.24), it follows

$$\delta(\mathcal{W}^{new}, \mathcal{E}_0) \leq 2\sqrt{p} \max_{j=1,\ldots,p} \|(I - \Pi)(u_j + z_j)\|. \tag{B.25}$$

By Proposition 4.6, we have

$$\|(I - \Pi)(u_j + z_j)\| \leq \|(I - \Pi_j)(u_j + z_j)\| \leq \sigma\left(\delta(\mathcal{E}_0, \mathcal{W})\right)^2. \tag{B.26}$$

Relation (4.29) comes from (B.25), (B.26), and, using (4.28), comes from $\delta(\mathcal{W}^{new}, \mathcal{E}_0) = \delta(\mathcal{E}_0, \mathcal{W}^{new})$.

Finally, (4.30) is a consequence of (4.29). □

Acknowledgments. I am indebted to Professor Jean Descloux for numerous valuable discussions about this work. I also thank Dr. R. Lehoucq for useful suggestions that have helped to improve the manuscript.

REFERENCES

[1] J. BRAMBLE, A. KNYAZEV AND J. PASCIAK, *A subspace preconditioning algorithm for eigenvector/eigenvalue computation*, Adv. Comput. Math., 6 (1997), pp. 159–189.
[2] M. CROUZEIX, B. PHILIPPE AND M. SADKANE, *The Davidson method*, SIAM J. Sci. Comput., 15 (1994), pp. 62–76.
[3] E. R. DAVIDSON, *The iterative calculation of a few of the lowest eigenvalues and corresponding eigenvectors of large real symmetric matrices*, J. Comput. Phys., 17 (1975), pp. 87–94.
[4] J. DESCLOUX, J.-L. FATTEBERT, AND F. GYGI, *RQI (Rayleigh Quotient Iteration), an old recipe for solving modern large scale eigenvalue problems*, Comput. Phys., 12 (1998), pp. 22–27.
[5] A. EDELMAN AND S. SMITH, *On conjugate gradient-like methods for eigen-like problems*, BIT, 36 (1996), pp. 494–508.
[6] J.-L. FATTEBERT, *Finite difference schemes and block Rayleigh Quotient Iteration for electronic structure calculations on composite grids*, J. Comput. Phys. (to appear).
[7] ———, *An inverse iteration method using multigrid for quantum chemistry*, BIT, 36 (1996), pp. 509–522.
[8] ———, *Une méthode numérique pour la résolution des problèmes aux valeurs propres liés au calcul de structure électronique moléculaire*, Ph.D. thesis, Thèse No 1640, Ecole Polytechnique Fédérale de Lausanne, 1997.
[9] W. HACKBUSCH, *Multi-grid Methods and Applications*, Springer, Berlin, 1985.
[10] T. KATO, *Perturbation Theory for Linear Operators*, 2nd Ed., Springer, Berlin Heidelberg, 1976.
[11] A. KNYAZEV, *New estimates for Ritz vectors*, Math. Comp., 66 (1997), pp. 985–995.
[12] D. LONGSINE AND S. MCCORMICK, *Simultaneous Raleigh-Quotient minimization methods for $Ax = \lambda Bx$*, Linear Algebra Appl., 34 (1980), pp. 195–234.
[13] B. N. PARLETT, *The Symmetric Eigenvalue Problem*, Prentice-Hall, Englewood Cliffs, NJ, 1980.
[14] Y. SAAD, *Numerical Methods for Large Eigenvalue Problems*, Manchester University Press, Manchester, 1992.
[15] Y. SAAD, A. STATHOPOULOS, J. CHELIKOWSKY, K. WU AND S. OGUT, *Solution of large eigenvalue problems in electronic structure calculations*, BIT, 36 (1996), pp. 563–578.
[16] G. SLEIJPEN AND H. V. D. VORST, *A generalized Jacobi-Davidson iteration method for linear eigenvalue problem*, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 401–425.