Universiteit Utrecht

Department of Mathematics

Convergence Estimates for Preconditioned Gradient Subspace Iteration Eigensolvers

by

E. Ovtchinnikov

Preprint

nr. 1244

June, 2002

CONVERGENCE ESTIMATES FOR PRECONDITIONED GRADIENT SUBSPACE ITERATION EIGENSOLVERS*

E. OVTCHINNIKOV†

Abstract

Subspace iteration for computing several eigenpairs (i.e. eigenvalues and eigenvectors) of an eigenvalue problem is an alternative to the deflation technique whereby the eigenpairs are computed successively by projecting the problem onto the subspace orthogonal to the already found eigenvectors. The main advantage of the subspace iteration over the deflation is its 'cluster robustness': even if some of the computed eigenvalues form a cluster (i.e. are very close to each other), the convergence does not deteriorate. For standard subspace iteration eigensolvers the above fact is well-known, and it is supported by convergence estimates. This paper tackles the so-called preconditioned gradient subspace iteration eigensolvers – a relatively new class of methods designed to efficiently compute several extreme eigenpairs of large-scale eigenvalue problems. Using a new approach to the convergence analysis of subspace iterations, based on dealing with eigenvalue sums rather than individual eigenvalues, the paper presents new convergence results for a class of preconditioned gradient subspace iteration eigensolvers which are fully cluster robust, i.e. involve the distances between the eigenvalues in a cluster neither in the assumptions nor in the estimates themselves.

1 Introduction

This paper is concerned with the convergence of numerical methods for computing several smallest eigenvalues and corresponding eigenvectors of the problem

$$Lu = \lambda M u \tag{1}$$

where L and M are symmetric positive definite operators in a Euclidean space $\mathcal{E}$. The paper's focus is on methods which can be applied efficiently to large-scale problems resulting from the discretization (e.g. finite element one) of partial differential ones, i.e. problems where L is a discretization of a partial differential (unbounded) operator and M of a non-differential (bounded) one.

The large scale of targeted problems rules out the use of matrix transformation methods (QZ-algorithm and the like), which work with full matrix representations of L and M and, hence, require N^2 storage, where N is the dimension of $\mathcal{E}$. Hence, such problems are tackled by vector iterations, i.e. iterative methods in which L and M are only used to compute Lv and Mv (or a linear combination thereof) for a given vector v. This paper studies the convergence of a class of such methods known as 'preconditioned gradient methods' [14, 13].

Preconditioned gradient methods are the oldest among iterative methods for eigenvalue problems which use *preconditioning* – the first (asymptotic) convergence result for such methods was obtained as early as in 1958 [25].[1] They are also among the best

* Presented at the mini-simpozium "Iterative Methods", University of Utrecht, 12 June 2002

† Harrow School of Computer Science, University of Westminster, London UK.

[1] It is quite remarkable that the asymptotic convergence factor given in [25] (see (5)) still remains the smallest proven for the method studied there, and that it was not until recent paper [20] that the corresponding non-asymptotic estimate had been obtained.

studied, especially as far as computing the smallest eigenvalue of (1) is concerned, as demonstrated by the extensive list of references in [13]. The idea behind these methods is quite simple and natural: the minimization of (or just reducing) the Rayleigh quotient in the direction of its gradient taken in some auxiliary scalar product. Their implementation is also fairly simple compared to various two-level methods built around the powerful shift-and-invert technique, such as Rayleigh quotient iterations, Jacobi-Davidson method etc.[2] (see e.g. [1] and the references therein), which invariably involve complicated stopping criteria for the inner iterations (see e.g. [18]). One of the most promising methods of the class, the so-called *locally optimal block preconditioned conjugate gradient* (LOBPCG) method [9] is shown numerically to outperform many of the preconditioned eigensolvers used nowadays (see [11, 12]).

The application of the preconditioned gradient technique to computing several smallest eigenvalues of (1) brings about quite a few theoretical and practical difficulties. The straightforward *deflation* technique whereby the orthogonalization to already computed eigenvectors is used (see e.g. [1]) may lead to slow convergence when eigenvalues form a cluster, i.e. a close group. This situation is adequately described by the theoretical convergence estimates (see e.g. (9) below) which show that, assuming that $\lambda_j < \lambda_{j+1}$ (throughout the paper the eigenvalues λ_j of (1) are enumerated in increasing order with each one counted as many times as its multiplicity) the asymptotic convergence factor q_j for λ_j is controlled by the ratio λ_j/λ_{j+1} with $q_j \to 1$ as $\lambda_{j+1} \to \lambda_j$.

An efficient remedy in dealing with clusters is to use *subspace* iterations rather than vector ones. Indeed, this technique coupled e.g. with the inverse iterations leads to the reduction of the error in computing λ_j by a factor of $\mathcal{O}\left((\lambda_j/\lambda_n)^{2i}\right)$ after i iterations, where n is the dimension of the iterated subspace (see e.g. [24]). This estimate is an obvious improvement over $\mathcal{O}\left((\lambda_j/\lambda_{j+1})^{2i}\right)$ estimate for the inverse iterations with deflation, and it shows that subspace iterations are 'cluster robust', i.e. even in the presence of a cluster (a one which does not include λ_n) the convergence to clustered eigenvalues does not depend on the distances between them[3].

A similar improved convergence estimate for a particular preconditioned gradient subspace iteration method was obtained in [3]. However, the assumptions on the problem and the initial subspace in [3] are quite strong: the eigenvalues are assumed to be simple and the distance between the initial subspace and the corresponding invariant subspace of (1) (measured in a certain metric) is assumed to be not greater than a rather small quantity depending strongly on the distances between the computed eigenvalues. Furthermore, this estimate is not recursive – the errors on the i-th iteration are estimated in terms of the initial errors rather than previous ones, and the reduction in the errors on each iteration is not guaranteed, which prevents applying this estimate to other practically important preconditioned gradient methods such as LOBPCG. The most recent estimate for preconditioned gradient subspace iteration methods given in [12] is recursive and, moreover, sharp but, according to this estimate, the reduction in the error for λ_j is determined by λ_j/λ_{j+1} – just like for the deflation technique. Apart from (seemingly) being in contradiction with the results of [3] such an estimate does

[2]Actually, some of these two-level methods can also be interpreted as preconditioned gradient ones with a different (more complicated) kind of preconditioning (see [20]) – the true distinction between the two approaches compared lies in whether or not the preconditioning used involves some inner iterations, i.e. is the method in hand one-level or two-level one.

[3]In more precise terms, there is an upper bound for the asymptotic convergence factor (cf. left-hand side of (5)) which is less than 1 and does not depend on the distances between the eigenvalues in the cluster.

not explain the numerical results in [12] itself (see also [13]), which clearly demonstrate cluster robustness of preconditioned gradient subspace iterations.

The main goal of this paper is to establish convergence estimates for (a class of) preconditioned gradient subspace iteration methods which would be truly cluster robust, i.e. would not involve the distances between the eigenvalues in a cluster at all – neither in the assumptions nor in the estimates themselves. The convergence results in [22] (for a special kind of preconditioned gradient subspace iterations) suggest that cluster robustness can be achieved by working with eigenvalue sums rather than individual eigenvalues, and the present paper follows this approach. The convergence analysis for eigenvalue sums is, however, rather complicated and involves quite a lot of technicalities, which are presented in the appendices.

2 Preconditioned gradient methods

Let us denote by $\lambda(u)$ the Rayleigh quotient for (1) on the vector u and by $r(u)$ the corresponding residual vector, i.e.

$$\lambda(u) = \frac{(Lu, u)}{(Mu, u)}, \quad r(u) = (L - \lambda(u)M)u \tag{2}$$

The minimal eigenvalue λ_0 of (1) is the minimum of $\lambda(u)$, therefore it can be found by applying to $\lambda(u)$ one of the methods for the minimization of a functional, e.g. the steepest descent method

$$u^{i+1} = u^i - \tau_i \nabla\lambda(u^i) \tag{3}$$

where τ_i are parameters which have to be chosen in such way that $\lambda^i \equiv \lambda(u^i)$ converges to λ_0. One possible choice for τ_i is the value which minimizes u^{i+1}, and it is this choice that has become associated with the term 'steepest descent method'.[4]

The gradient direction $\nabla\lambda(u)$ depends on the scalar (inner) product used. In the standard steepest descent method the usual scalar product $(\cdot,\cdot)$ in $\mathcal{E}$ is used, which leads to a very slow convergence of the iterations (3) when the ratio $(\lambda_1 - \lambda_0)/\lambda_{N-1}$ is small. In the preconditioned steepest descent the scalar product $(\cdot,\cdot)_{K^{-1}} \equiv (K^{-1}\cdot,\cdot)$ is used where K is a symmetric positive definite operator which will be specified later on. The gradient of $\lambda(u)$ in this scalar product becomes

$$\nabla_K \lambda(u) = \frac{2}{(Mu, u)} Kr(u)$$

and (3) becomes

$$u^{i+1} = u^i - \tau_i Kr(u^i) \tag{4}$$

The above iterative scheme was first studied in [25] where the following estimate for the asymptotic convergence factor has been obtained:[5]

$$q_\infty \equiv \overline{\lim}_{i\to\infty} \left(\frac{\lambda^i - \lambda_0}{\lambda^0 - \lambda_0} \right)^{\frac{1}{i}} \le \left(\frac{b_0 - a_0}{b_0 + a_0} \right)^2 \tag{5}$$

[4] In fact, this association is rather confusing: 'steepest' actually refers to the direction of descent rather than to the particular point in that direction where $\lambda(u)$ reaches its (locally) minimal value. It would be more proper to call iterations (3) with the (locally) optimal choice of τ_i 'locally optimal steepest descent'.

[5] The very same asymptotic result was later reproduced (apparently independently and in a somewhat different context) in [19]

where a_0 and b_0 are resp. the minimal positive and the maximal eigenvalue of $K(L - \lambda_0 M)$. Subsequently, various non-asymptotic convergence estimates for (4) have been obtained (see e.g. [8] or Chapter 9 in the monograph [6] and the references therein) – however, all of them, apart from that in the recent paper [20], lead to asymptotic constants which are greater than q_∞ above in the general case, i.e. for K that does not commute with L^{-1}.

In [16] the iterative scheme (4) is interpreted in a different manner. In the simplest version of the well-known inverse iteration method for (1) (see e.g. [5]) u^{i+1} is found by solving $Lu^{i+1} = \lambda^i M u^i$. The last equation can be solved using the iterative scheme[6]

$$v^{k+1} = v^k - \tau K(Lv^k - \lambda^i M u^i) \tag{6}$$

where K is a *preconditioner* for L, i.e. an operator 'similar' to L^{-1} in the following sense:

$$aL^{-1} \leq K \leq bL^{-1} \tag{7}$$

If only one iteration (6) is performed, we obtain the iterative scheme which is called in [16] *preconditioned inverse iteration* (PINVIT):

$$u^{i+1} = u^i - \tau K r(u^i) \tag{8}$$

We observe that PINVIT is a particular case of (4) with $\tau_i = \tau$. Further, we observe that K in (4) can be interpreted as a preconditioner. Hence, it is natural to call (4) a *preconditioned gradient* method.

In [17] a sharp convergence estimate for PINVIT is given. In its simplified form presented in [12] it reads:

$$\frac{\lambda^{i+1} - \lambda_j}{\lambda_{j+1} - \lambda^{i+1}} \leq q^2 \frac{\lambda^i - \lambda_j}{\lambda_{j+1} - \lambda^i}, \quad q = \gamma + (1-\gamma)\frac{\lambda_j}{\lambda_{j+1}} \tag{9}$$

where $\gamma = (b-a)/(b+a)$ and j is such that $\lambda_j \leq \lambda^i < \lambda_{j+1}$. According to (9) the upper estimate for the asymptotic convergence factor is q^2 and it is somewhat greater than that in [25],[19] and [20] (for $j = 0$, which is the case considered in the latter papers). This is quite expectable, as the value of τ in PINVIT is not optimal neither locally nor globally (cf. [12, 13]).

Just like in the case of solving linear systems, the convergence of the preconditioned steepest descent can be accelerated by using data from previous iterations. The simplest approach, underpinning a group of methods known as *preconditioned conjugate gradient* methods, is to use an iterative scheme of the form

$$u^{i+1} = u^i - \tau_i K r(u^i) - \upsilon_i u^{i-1} \tag{10}$$

Particular methods of this group differ by the choice of the parameters τ_i and υ_i. In the so-called locally optimal preconditioned conjugate gradient method (LOPCG) [9] τ_i and υ_i are such that $\lambda(u^{i+1})$ is minimal possible (hence, the above convergence estimates apply to LOPCG as well). Some other choices are considered e.g. in [7]: respective convergence estimates for the general case $a < b$ are not yet available[7], but the numerical experiments demonstrate essentially the same convergence as with LOPCG [21].

[6] In [16] K is scaled so that $\tau = 1$.

[7] In the case $a = b = 1$ (i.e. $K = L^{-1}$) asymptotical convergence results are given in [2].

A straightforward generalization of (10) is to introduce more previous approximate eigenvectors (i.e. u^{i-2}, u^{i-3} etc.) into the right-hand side. However, as the numerical experiments in [13] show, this approach does not lead to any noticeable further acceleration of the convergence. Furthermore, numerical experiments in [21] show that the same is true for the generalized Davidson method [23, 24, 1, 20] where *all* previous u^i are used. Some insights on why this might be so are provided in [9, 11], where locally optimal preconditioned conjugate gradient iterations for the eigenvalue problem (1) are compared with those for the linear system $(L - \lambda_0 M)u = 0$. In the latter case the preconditioned conjugate gradient methods mentioned above are equivalent, since they all produce the best (in the semi-norm induced by $L - \lambda_0 M$) approximation to u in the Krylov subspace for $K(L - \lambda_0 M)$. Consequently, invoking u^{i-2} etc. does not affect the convergence at all (in exact arithmetic), as the approximate solution remains in the same subspace. While the same does not hold for the eigenvalue problem (1), the two cases at hand are closely related because the functionals minimized on each preconditioned conjugate gradient iteration for (1) and for the above linear system are asymptotically close to each other, as indicated in [9].

3 Preconditioned gradient subspace iterations

Preconditioned gradient subspace iteration eigensolvers combine the preconditioned gradient descent technique (4) with the Rayleigh-Ritz method. A straightforward subspace generalization of (4), known as the block preconditioned steepest descent (see e.g. [12]), defines new approximations $\{\lambda_j^{i+1}, u_j^{i+1}\}$ to the eigenpairs $\{\lambda_j, u_j\}$ of (1) as the Rayleigh-Ritz eigenpairs in the subspace

$$\mathcal{I}^{i+1} \equiv \operatorname{span}\{u_j^{i+\frac{1}{2}}\}_{j=0,n-1}, \quad u_j^{i+\frac{1}{2}} = u_j^i - \tau_{ij} K r(u_j^i) \tag{11}$$

The use of the above scheme raises the issue of the choice of parameters τ_{ij}. In [4] locally optimal ones for each vector are chosen, i.e. τ_{ij} minimizes $\lambda(u_j^{i+\frac{1}{2}})$. No convergence proof for this scheme is available at present, but the numerical results in [4] demonstrate that it is working. The paper [3] assumes that

$$(1-\gamma)L^{-1} \le K \le (1+\gamma)L^{-1} \tag{12}$$

where $\gamma < 1$, and suggests $\tau_{ij} = 1$, which leads to the following convergence result: assuming that

$$\sum_{j=0}^{n-1} \|(1-P_0)u_j\|_L^2 \le \frac{(1-\gamma)^2}{1999\Delta^2}\left(\frac{\lambda_0}{\lambda_{n-1}}\left(1 - \frac{\lambda_{n-1}}{\lambda_n}\right)\right)^4 \tag{13}$$

where

$$\Delta = \max_{0 \le j < n} \frac{\lambda_{j+1} + \lambda_j}{\lambda_{j+1} - \lambda_j}$$

and P_0 is the $(\cdot,\cdot)_L$-orthogonal projection onto $\mathcal{I}^0$, one has:

$$1 - \frac{\lambda_j}{\lambda_j^i} \le 1.03 \frac{\lambda_n}{\lambda_n - \lambda_j} \bar{\delta}_j^{2i} \|(1-P_0)u_j\|_L^2 \tag{14}$$

where

$$\bar{\delta}_j = \delta_j + \frac{1-\delta_{n-1}}{2}\frac{\lambda_j}{\lambda_{n-1}}\sqrt{\frac{\lambda_n-\lambda_{n-1}}{\lambda_n-\lambda_j}}, \quad \delta_j = \gamma + (1-\gamma)\frac{\lambda_j}{\lambda_n}$$

We note that although the estimates (14) are asymptotically cluster robust, the same cannot be said about the above convergence result as a whole because of the condition (13) on the initial subspace $\mathcal{I}^0$, where the distances between eigenvalues feature prominently.

The same choice of τ_{ij} under the same condition on K is also considered in [15], where the following sharp recursive convergence estimates are given:

$$\frac{\lambda_j^{i+1}-\lambda_{k_j}}{\lambda_{k_j+1}-\lambda_j^{i+1}} \le q\left(\gamma, \frac{\lambda_{k_j}}{\lambda_{k_j+1}}\right)^2 \frac{\lambda_j^{i}-\lambda_{k_j}}{\lambda_{k_j+1}-\lambda_j^{i}}, \quad q(u,v) = u + (1-u)v \tag{15}$$

where k_j is such that $\lambda_{k_j} \le \lambda_j^i < \lambda_{k_j+1}$.

A radical solution to the problem of optimal choice of τ_{ij} is the use of the subspace [8]

$$\mathcal{I}^{i+\frac{1}{2}} = \mathrm{span}\{u_j^i, Kr(u_j^i)\}_{j=0,n-1} \tag{16}$$

instead of $\mathcal{I}^{i+1}$ given by (11). The new subspace $\mathcal{I}^{i+1}$ is then defined as the one spanning the Ritz eigenvectors in $\mathcal{I}^{i+\frac{1}{2}}$ corresponding to n smallest Ritz eigenvalues. It turns out, that, just like in computing a single eigenvalue, adding the previous approximations u_j^{i-1} to $\mathcal{I}^{i+\frac{1}{2}}$ dramatically improves the convergence. A group of methods built around this idea is known under the collective name of locally optimal block preconditioned conjugate gradient (LOBPCG) method [11]. Again, a surprising fact is that increasing $\mathcal{I}^{i+\frac{1}{2}}$ any further does not have any tangible effect [11, 13].

Due to their recursiveness and in view of the minimax principle, estimates (15) apply to any iterative scheme which uses the Rayleigh-Ritz method in a subspace containing $\mathcal{I}^{i+\frac{1}{2}}$ given by (16), in particular, to LOBPCG (although they, of course, do not explain the remarkable convergence features of the latter). As admitted in [12], a serious disadvantage of estimates (15) is that they are not 'cluster robust' in the sense specified in Introduction, whereas there is at least numerical evidence that methods using the subspace given by (16) are (see [12, 13]). Furthermore, assuming that $k_j = j$, the (upper estimate of the) asymptotic convergence factor according to (15) is the square of $\gamma+(1-\gamma)\lambda_j/\lambda_{j+1}$, whereas according to [3] it should be the square of $\gamma + (1-\gamma)\lambda_j/\lambda_n$. Given that (15) are claimed to be sharp, there may seem to be a contradiction. However, there is none, because estimates (15) are sharp in a very specific sense: (15) for a given j covers *any* subspace $\mathcal{I}^i$ with the same Ritz eigenvalue λ_j^i and *any* preconditioner K satisfying 7 with a given ratio b/a. But if $\mathcal{I}^i$ happened to be a very 'bad' subspace as far as the convergence to some eigenvalue is concerned, the subsequently calculated subspaces $\mathcal{I}^{i+k}$ may improve in this respect, as can be seen e.g. from [3] or from the new estimates below. Finally, the preconditioner K remains the same throughout the iterations, and if $\mathcal{I}^i$ and K happened to be a 'bad' pair, this does not necessarily apply to $\mathcal{I}^{i+k}$ and K. In short, estimates (15), despite being 'sharp and short', are somewhat 'too pessimistic' and certainly not 'cluster robust'.

The new estimates presented in the next section achieve 'cluster robustness' by dealing with eigenvalue sums rather than individual eigenvalues. It should be noted

[8] This version of the block preconditioned steepest descent might also be called 'locally optimal'

that the same idea was used in [22]; however, the method suggested there proved to be less efficient than LOBPCG. The convergence results of the present paper produce smaller estimates for the asymptotic convergence factor than that in [22] and, moreover, they apply to a class of preconditioned gradient methods which includes LOBPCG.

4 New convergence estimates

The new estimates below apply to a class of preconditioned gradient subspace iteration methods based on iterative schemes of the following form:

$$\mathcal{I}^{i+1} = \operatorname{span}\{u_j(\mathcal{X}^i)\}_{j=0,n-1}, \quad \mathcal{X}^i \supset \mathcal{I}^i + \operatorname{span}\{Kr_j(\mathcal{I}^i)\}_{j=0,n-1} \tag{17}$$

where $\mathcal{I}^i$ is the iterated subspace, $r_j(\mathcal{X}) \equiv r(u_j(\mathcal{X}))$ and $u_j(\mathcal{X})$ are the Ritz eigenvectors in the subspace $\mathcal{X}$, i.e.

$$(r(u_j(\mathcal{X})), v) = 0 \quad \forall v \in \mathcal{X}$$

enumerated in the increasing order of $\lambda(u_j(\mathcal{X}))$ starting from 0. We observe that the above class of methods includes e.g. (locally optimal) block preconditioned steepest descent (cf. (16)), the original version of LOBPCG and the block version of the generalized Davidson method in [11]. It is important to note, however, that with regard to the last two methods the estimates of this paper can only be considered as preliminary ones, as the methods themselves demonstrate much better convergence in practical calculations (see [11, 12, 13]).

The estimates below are given in terms of the inverses $\mu_j = \lambda_j^{-1}$ of the eigenvalues of (1) and those of the Ritz eigenvalues λ_j^i, i.e. $\mu_j^i = (\lambda_j^i)^{-1}$, which proved to be more convenient for the convergence analysis. This is equivalent to rewriting (1) as $Mu = \mu Lu$. Accordingly, we denote the Rayleigh quotient for the latter problem by $\mu(u)$, i.e. $\mu(u) = \lambda(u)^{-1}$, and use the notation $s(u) = (\mu(u)L - M)u$ and $s_j^i = s(u_j^i)$ for the residuals. The enumeration of μ_j and μ_j^i starts from 0 and is in decreasing order.

The gap between subspaces $\mathcal{X}$ and $\mathcal{Y}$ measured in the scalar product $\langle \cdot, \cdot \rangle$ is denoted by $\theta\langle \cdot, \cdot \rangle$. Further, we denote

$$\mathcal{I}_m \equiv \operatorname{span}\{u_j\}_{j=0,m-1}, \quad \mathcal{I}_m^i \equiv \operatorname{span}\{u_j^i\}_{j=0,m-1}$$

$$\theta_{i,m} \equiv \theta(\mathcal{I}_m^i, \mathcal{I}_m)_L, \quad t_{i,m} \equiv \tan(\mathcal{I}_m^i, \mathcal{I}_m)_L \equiv \frac{\theta_{i,m}}{\sqrt{1 - \theta_{i,m}^2}}$$

$$\rho_{i,m} \equiv \sum_{j=0}^{m-1} \frac{\|s_j^i\|_{L^{-1}}^2}{\mu_j^i}$$

$$\alpha_{j,k}^i \equiv \frac{\mu_{j-1}^i}{\mu_{j-1}^i - \mu_k}, \quad \alpha_j^i \equiv \alpha_{j,j}^i, \quad \beta_j^i \equiv \frac{\mu_j}{\mu_{j-1}^i - \mu_j}$$

We start with two preliminary results. The first one relates to the convergence of a group of approximate eigenvalues μ_j^i, $j = 0, \ldots, m-1$, where $m \leq n$ is such that $\mu_{m-1}^0 > \mu_m$.

Theorem 1 *If K satisfies* (7) *and $\mu^0_{m-1} > \mu_m$ for some $0 < m \le n$ then μ^i_j converges to μ_j, $j = 0, \ldots, m-1$, and, furthermore, the following convergence estimate holds for the iterative scheme* (17)*:*

$$\sum_{j=0}^{m-1} (\mu_j - \mu_j^{i+1}) \le \frac{q_{i,0}^2 + \epsilon_i^0}{1 + \epsilon_i^0} \sum_{j=0}^{m-1} (\mu_j - \mu_j^i) \tag{18}$$

where

$$q_{i,0} = \frac{\kappa_{i,0} - 1}{\kappa_{i,0} + 1}, \quad \kappa_{i,0} = (1 + t_{i,m}^2)^2 \alpha_m^i \frac{b}{a}$$

$$\epsilon_i^0 = \frac{b}{a}(1 + \rho_{i,m})(a_i^0 \rho_{i,m} + b_i^0 t_{i,m}^2)$$

and a_i^0 and b_i^0 depend on α_m^i, $\frac{\mu_0}{\mu_m}$, m and $t_{i,m}$ only, in a continuous and monotonically increasing way.

Proof. First, we transform the problem (1) into the one with $L = I$ as described in Appendix A.1 (note that the reverse transformation turns $\theta(\mathcal{I}_m^i, \mathcal{I}_m)$ into $\theta(\mathcal{I}_m^i, \mathcal{I}_m)_L$, $\tan(\mathcal{I}_m^i, \mathcal{I}_m)$ into $\tan(\mathcal{I}_m^i, \mathcal{I}_m)_L$ and $\|s_j^i\|$ into $\|s_j^i\|_{L^{-1}}$). The convergence of $\mathcal{I}_m^i$ to $\mathcal{I}_m$ (and, hence, μ_j^i to μ_j) follows from the minimax principle, lemma 23 and lemma 10. Applying lemma 33 with $k = m$ and $l = 0$, we obtain the estimate of this theorem with $a_i^0 = c_4$ and $b_i^0 = c_5/(1 + t_{i,m}^2) + c_6$. □

The second preliminary result improves the asymptotic convergence factor in theorem 1 for rightmost eigenvalues μ_j in the above group.

Theorem 2 *Assuming that K satisfies* (7) *and that $\mu^0_{m-1} > \mu_m$ and $\mu^0_{k-1} > \mu_k$ for some $0 < k < m \le n$, the following convergence estimate holds for the iterative scheme* (17)*:*

$$\sum_{j=0}^{k-1} (\mu_j - \mu_j^{i+1}) \le \frac{q_i^2 + \epsilon_i^*}{1 + \epsilon_i^*} \sum_{j=0}^{k-1} (\mu_j - \mu_j^i) \tag{19}$$

where

$$q_i = \frac{\kappa_i - 1}{\kappa_i + 1}, \quad \kappa_i = (1 + t_{i,m}^2)^2 \alpha_{k,m}^i \frac{b}{a}$$

$$\epsilon_i^* = \frac{b}{a}(1 + \rho_{i,k})(a_i^* \rho_{i,k} + b_i^* t_{i,m}^2 + c_i^* t_{i,k}^2)$$

and a_i^*, b_i^* and c_i^* depend on $\alpha_{k,m}^i$, α_k^i, $\frac{\mu_0}{\mu_k}$, k, $t_{i,m}$ and $t_{i,k}$ only, in a continuous and monotonically increasing way.

Proof. Just like with theorem 1 we apply lemma 33 (with $l = 0$) to obtain the estimate of this theorem with $a_i^* = c_4$, $b_i^* = c_5/(1 + t_{i,m}^2)$ and $c_i^* = c_6$. □

For the main result below some more notation is needed:

$$\mathcal{I}_{l,k} \equiv \operatorname{span}\{u_j\}_{j=l,k-1}, \quad \mathcal{I}_{l,k}^i \equiv \operatorname{span}\{u_j^i\}_{j=l,k-1}$$

$$t_{i,l,k} = \frac{\theta(\mathcal{I}_{l,k}, \mathcal{I}_{l,k}^i)_L}{\sqrt{1 - \theta(\mathcal{I}_{l,k}, \mathcal{I}_{l,k}^i)_L^2}}$$

Theorem 3 *Assume that K satisfies* (7) *and that $\mu_{m-1}^{i_0} > \mu_m$, $\mu_{k-1}^{i_0} > \mu_k$ and $\mu_{l-1}^{i_0} > \mu_l$ for some i_0 and some $0 < l < k < m \le n$. If*

$$\frac{1}{\mu_0}\sum_{j=0}^{l-1}(\mu_j - \mu_j^{i_0}) \le \frac{1}{2(k-l)^2}\frac{a}{b}\left(\left(2+\frac{\mu_0}{\mu_k}\right)\beta_l^{i_0}\right)^{-2}\left(\frac{\mu_k}{\mu_0}\right)^2 \tag{20}$$

and

$$\frac{1}{\mu_0}\sum_{j=0}^{l-1}(\mu_j - \mu_j^{i_0}) \le \frac{1}{8}\left(\left(2+\frac{\mu_m+\mu_0}{\mu_l}\right)\left(1+m\beta_l^{i_0}\frac{\mu_0}{\mu_l}\right)\beta_l^{i_0}\right)^{-1}\frac{\mu_l}{\mu_0} \tag{21}$$

then the following convergence estimate for (17) *holds for $i \ge i_0$:*

$$\sum_{j=l}^{k-1}(\mu_j - \mu_j^{i+1}) \le \frac{q_i^2+\epsilon_i}{1+\epsilon_i}\sum_{j=l}^{k-1}(\mu_j - \mu_j^{i}) \tag{22}$$

where q_i is defined in theorem 2,

$$\epsilon_i = 2\frac{b}{a}(1+\rho_{i,k})(a_i\rho_{i,k} + b_i t_{i,m}^2 + c_i t_{i,l,k}^2)$$

and a_i, b_i and c_i depend on $\alpha_{k,m}^i$, α_k^i, α_l^i, β_l^i, $\frac{\mu_0}{\mu_l}$, $\frac{\mu_l}{\mu_k}$, $k-l$, $t_{i,m}$ and $t_{i,l,k}$ only, in a continuous and monotonically increasing way.

Proof. Again, it is enough to apply lemma 33 (with $\xi = \zeta = 0.5$) and to notice that if (20) and (21) are valid for a given i_0 then they remain so for $i > i_0$. □

Remark 1 *The (fairly technical) results of § A.5 can be used to derive explicit formulas for the coefficients in theorems 1-3. Those, however, are bound to be extremely cumbersome, which is the only reason why they are not presented in the paper.*

Remark 2 *The above results remain valid if the left-hand side inequality in* (7) *is replaced with*

$$K \ge a(I-\pi)L^{-1}$$

where π is the spectral projection onto $\mathcal{I}_m$ (cf. remarks 3 and 4).

Taking a closer look at the above convergence result one observes that it involves four groups of parameters:

1. The (spectral) condition number b/a of KL and the 'cluster depth' parameter $\alpha_{k,m}^i$. These two parameters might be called 'essential' as they determine the asymptotic convergence factor (cf. (23) below).

2. Parameters related to 'eigenvalue macrostructure' or 'cluster parameters' (here and below 'cluster' refers to μ_j, $j = l, \ldots, k-1$): $k-l$, μ_0/μ_l, μ_0/μ_k and μ_l/μ_m.

3. Those related to the 'cluster resolution': α_k^i, α_l^i, β_l^i. We observe that these parameters are monotonically decreasing in i, and their upper bounds (at $i = 0$) depend on the 'quality' of the initial guess $\mathcal{I}^0$ measured by μ_j^0, whereas their lower bounds, which they approach as one proceeds with the subspace iterations (17), are relative distances between the cluster limits and the rest of the spectrum.

4. 'Asymptotically insignificant' parameters $\rho_{i,k}$, $t_{i,m}$ and $t_{i,l,k}$. We observe that $\rho_{i,k}$, $t_{i,m}$ and $t_{i,l,k}$ can be estimated, using corollaries 4, 3 and lemma 19, in terms of the parameters of the previous two groups. Furthermore, from lemmas 10, 23 and 24 one can obtain the upper bounds for $t_{i,m}^2$ proportional to the total error of μ_j^i, $j = 0, \ldots, m-1$, for $\rho_{i,k}$ to that of μ_j^i, $j = 0, \ldots, k-1$, and for $t_{i,l,k}^2$ to that of μ_j^i, $j = l, \ldots, k-1$, all the constants involved again depending only on the parameters of the previous two groups.

It is important to emphasize that the 'eigenvalue microstructure' parameters (i.e. the distances between the eigenvalues in a cluster) never appear in the convergence estimates of theorems 1-3. Furthermore, even the 'macrostructure' parameters and 'cluster resolution' parameters only appear in asymptotically insignificant terms.

We observe further that theorem 1 implies that conditions (20) and (21) are satisfied after a number of iterations which depend only on the parameters of the above first three groups. Finally, we observe that the asymptotic convergence factor for μ_j^i, $j = l, \ldots, k-1$, is estimated by

$$q_\infty = \left(\frac{1-\xi}{1+\xi}\right)^2, \quad \xi = \frac{a}{b}\left(1 - \frac{\mu_m}{\mu_{k-1}}\right) \tag{23}$$

Thus, by taking $a = b = 1$, $k = l+1$ and $m = n$ one can obtain from (23) the well-known convergence result for inverse iterations (see e.g. [5]).

The above considerations show how the convergence estimates of theorems 1-3 can be used in order to ascertain the robustness of a particular application of (17) with respect to a certain parameter t on which the problem (1) depends (e.g. a discretization parameter). If a preconditioner is used for which b/a can be estimated from above uniformly with respect to t, and if similar estimates can be obtained for the parameters describing 'eigenvalue macrostructure' in the above sense, then starting with an initial guess of a sufficient quality (i.e. with $\mu_{m-1}^0 > \mu_m$ and with α_m^0 not exceeding a given arbitrary value independent of t) one has convergence rate estimates (18), (19) and, eventually, (22) which are independent of t.

Let us now turn to the comparison between the above new convergence results and (15).

Estimates (15) are certainly impressive: they are explicit, sharp, take minimal assumptions on the preconditioner K and no assumptions at all on the current subspace $\mathcal{I}^i$. However, they do not adequately reflect some convergence aspects of the preconditioned gradient subspace iterations. Firstly, they are not cluster robust. A closely related issue is that of the asymptotic convergence factor: the esimate for this factor that can be derived from (15) is larger than that given by (23); furthermore, it is determined by μ_{k_j+1}/μ_{k_j}, and, hence, (15) does not demonstrate any advantage of subspace iterations compared to the deflation technique. Yet another limitation of (15) is even more serious.

From the minimax principle and from lemma 23 it follows that the Ritz eigenvalues μ_j^i converge to eigenvalues of (1) for any initial subspace $\mathcal{I}^0$. However, unless some assumptions on $\mathcal{I}^0$ are taken, it is impossible to identify to which eigenvalues they will have converged. An assumption which is often taken is $\mu_{m-1}^0 > \mu_m$ for some $m \le n$ (see e.g. [6, 8]): it appears to be the weakest possible in terms of eigenvalues only which guarantees the convergence of μ_j^i to μ_j, $j = 0, \ldots, m-1$, and it is used in the above new results. By estimating $q_{i,0}$ and ϵ_i^0 from above it is not difficult to obtain from (18) an estimate of the number of iterations needed to compute μ_j,

$j = 0, \dots, m-1$, to a given accuracy (such estimates are important in ascertaining the robustness of the convergence for parameter-dependent problems – cf. the discussion above), and theorems 2 and 3 allow one to improve this estimate for the respective groups of eigenvalues. Based on (15), however, such an estimate can only be obtained for those μ_j for which the interval $[\mu_j, \mu_{j+1}]$ contains at least one μ_k^i. Hence, to estimate the number of iterations needed to compute all μ_j , $j = 0, \dots, m-1$, to a given accuracy one has to assume that each interval $[\mu_j, \mu_{j+1}]$ contains *precisely* as many Ritz eigenvalues as the multiplicity of μ_j. This assumption is obviously much stronger than the assumption $\mu_{m-1}^0 > \mu_m$, especially in the presence of closely situated eigenvalues.

A Proofs

A.1 Notation

All the vectors and matrices featuring below are real-valued.

Below $\mathcal{E}$ denotes a Euclidean space. As usual, $(\cdot, \cdot)$ denotes the scalar (inner) product and $\|\cdot\|$ the associated norm in $\mathcal{E}$, and $\mathcal{L}(\mathcal{E})$ denotes the space of linear operators acting from $\mathcal{E}$ into $\mathcal{E}$. The norm of $A \in \mathcal{L}(\mathcal{E})$ subordinated to $\|\cdot\|$ (the square root of the maximal eigenvalue of A^TA) is denoted by $\|A\|$. The unit operator in $\mathcal{L}(\mathcal{E})$ is denoted by I (or, where the dimension N of $\mathcal{E}$ matters, by I_N), and the null operator by 0. The trace of $A \in \mathcal{L}(\mathcal{E})$ (i.e. the sum of the eigenvalues of A) is denoted by $\mathrm{Tr}(A)$, and $\|A\|_F$ denotes the Frobenius norm of A, i.e. $\|A\|_F^2 = \mathrm{Tr}(A^TA)$.

If $P \in \mathcal{L}(\mathcal{E})$ is a projection (i.e. $P^2 = P$) then $\bar{P} \equiv I - P$. The orthogonal projection onto $\mathcal{X} \subset \mathcal{E}$ is denoted by $P_{\mathcal{X}}$. Since $P_{\mathcal{X}}\bar{P}_{\mathcal{Y}} = P_{\mathcal{X}}(P_{\mathcal{X}} - P_{\mathcal{Y}})$ and $\bar{P}_{\mathcal{X}}P_{\mathcal{Y}} = (P_{\mathcal{X}} - P_{\mathcal{Y}})P_{\mathcal{X}}$ we have

$$\|P_{\mathcal{X}}\bar{P}_{\mathcal{Y}}\| \le \theta(\mathcal{X}, \mathcal{Y}), \quad \|\bar{P}_{\mathcal{X}}P_{\mathcal{Y}}\| \le \theta(\mathcal{X}, \mathcal{Y}), \quad \|\bar{P}_{\mathcal{X}}P_{\mathcal{Y}}\bar{P}_{\mathcal{X}}\| \le \theta(\mathcal{X}, \mathcal{Y})^2 \tag{24}$$

If $\theta(\mathcal{X}, \mathcal{Y}) < 1$ then

$$\tan(\mathcal{X}, \mathcal{Y}) \equiv \frac{\theta(\mathcal{X}, \mathcal{Y})}{\sqrt{1 - \theta(\mathcal{X}, \mathcal{Y})^2}}$$

For $0 \le A = A^T \in \mathcal{L}(\mathcal{E})$ the notation $(u, v)_A \equiv (Au, v)$ and $\|u\|_A \equiv \sqrt{(u, u)_A}$ is used. If $0 < A = A^T \in \mathcal{L}(\mathcal{E})$ then $(\cdot, \cdot)_A$ is a scalar product in $\mathcal{E}$, and $P_{\mathcal{H},A}$ denotes the projection onto $\mathcal{H} \subset \mathcal{E}$ which is orthogonal in this scalar product, i.e. $((1 - P_{\mathcal{H},A})u, v)_A = 0$ for any $u \in \mathcal{E}$ and $v \in \mathcal{H}$.

In order to simplify notation, we rewrite the problem (1) and the iterative scheme (17) as follows. Let us denote (in this paragraph only) $\tilde{M} = L^{-1/2}ML^{-1/2}$, $\tilde{K} = L^{-1/2}KL^{-1/2}$ and $\tilde{u} = L^{1/2}u$ for any $u \in \mathcal{E}$. Then (1) becomes

$$\tilde{u} = \lambda\tilde{M}\tilde{u} \tag{25}$$

or else $\tilde{M}\tilde{u} = \mu\tilde{u}$, and the subspace $\mathcal{X}^i$ in (17) becomes

$$\mathcal{X}^i \supset \mathcal{I}^i + \mathrm{span}\{\tilde{K}\tilde{r}_j^i\}_{j=0,n-1} \tag{26}$$

where $\tilde{r}_j^i = (I - \lambda_j^i\tilde{M})\tilde{u}_j^i$ and $\{\lambda_j^i, \tilde{u}_j^i\}$ are the Ritz eigenpairs for (25) in $\mathcal{X}^i$. Comparing (25) with (1) and (26) with (17) we observe that in (1) and (17) one can assume

$$L = I \tag{27}$$

without any loss of generality. Below we take this assumption and, accordingly, assume that

$$aI \leq K \leq bI \tag{28}$$

If $\mathcal{I}$ is an invariant subspace of M then $\pi_{\mathcal{I}} \equiv P_{\mathcal{I}} = P_{\mathcal{I},M}$ and $\bar{I} = \bar{\pi}\mathcal{E}$. The invariant subspace of M corresponding to a part σ of its spectrum is denoted by $\mathcal{I}_\sigma$, and $\pi_\sigma \equiv \pi_{\mathcal{I}_\sigma}$. Conversely, the part of the spectrum of M corresponding to an invariant subspace $\mathcal{I}$ is denoted by $\sigma_{\mathcal{I}}$. For any $\mathcal{X}, \mathcal{Y} \subset \mathcal{E}$ and any non-zero $u \in \mathcal{E}$

$$\delta_\mu(u, \mathcal{X}) \equiv \min_{v \in \mathcal{X}} |\mu(u) - \mu(v)|$$

Finally, we denote

$$\rho_{\mathcal{X}} = \sum_{i=0}^{m-1} \frac{\|s_i(\mathcal{X})\|^2}{\mu_i(\mathcal{X})}$$

where m is the dimension of $\mathcal{X} \subset \mathcal{E}$, $\mu_i(\mathcal{X}) \equiv \mu(u_i(\mathcal{X}))$, $s_i(\mathcal{X}) \equiv s(u_i(\mathcal{X}))$, and $u_i(\mathcal{X})$ are the Ritz eigenvectors of M in $\mathcal{X}$. Ritz eigenvalues for M are normalized in $\|\cdot\|_M$, and the enumeration is in decreasing order of $\mu_i(\mathcal{X})$.

A fundamental result frequently used below is the so-called *minimax (Courant-Fisher) principle*:

$$\lambda_i = \min_{\substack{\mathcal{X} \subset \mathcal{E} \\ \dim \mathcal{X} = i+1}} \max_{0 \neq u \in \mathcal{X}} \lambda(u)$$

from which, in particular, it follows that $\lambda_i \leq \lambda_i(\mathcal{X}) \leq \lambda_i(\mathcal{Y})$ for any $\mathcal{Y} \subset \mathcal{X} \subset \mathcal{E}$.

A.2 General auxiliary results

Lemmas 1–3 below are elementary, hence, the proofs are omitted.

Lemma 1 *For any $A = A^T \in \mathcal{L}(\mathcal{E})$ and any $0 \leq B = B^T \in \mathcal{L}(\mathcal{E})$ one has*

$$\mathrm{Tr} AB \leq \|A\| \, \mathrm{Tr} B \tag{29}$$

Lemma 2 *Let $A = A^T \in \mathcal{L}(\mathcal{E})$. If $A < I$ then $(I - A)^{-1} \geq I + A$.*

Corollary 1 *Let $0 < A = A^T \in \mathcal{L}(\mathcal{E})$ and $B = B^T \in \mathcal{L}(\mathcal{E})$. If $B < A$ then $(A-B)^{-1} \geq A^{-1} + A^{-1}BA^{-1}$.*

Lemma 3 *Let $A = A^T \in \mathcal{L}(\mathcal{E})$ and $0 < B = B^T \in \mathcal{L}(\mathcal{E})$. If $-\alpha B \leq A \leq \alpha B$ for some $\alpha > 0$ then $|(Au, v)| \leq \alpha \|u\|_B \|v\|_B$ for any $u, v \in \mathcal{E}$.*

Lemma 4 *Let A_0 and B_0 be symmetric positive definite operators in $\mathcal{L}(\mathcal{E})$ and let $A = A_0 + \delta A$ and $B = B_0 + \delta B$, where δA and δB are symmetric. Let λ_i^0 and λ_i be the eigenvalues of $B_0^{-1}A_0$ and $B^{-1}A$ resp. enumerated in the same order. If*

$$-\alpha A_0 \leq \delta A \leq \alpha A_0, \quad -\beta B_0 \leq \delta B \leq \beta B_0, \quad B \geq \gamma B_0 \tag{30}$$

then

$$|\lambda_i - \lambda_i^0| \leq \frac{\alpha + \beta}{\gamma} \lambda_i^0 \tag{31}$$

Proof. For any non-zero u we have

$$\frac{(Au,u)}{(Bu,u)} = \frac{(A_0u,u)}{(B_0u,u)} + \frac{1}{(Bu,u)}\left((\delta Au,u) - \frac{(A_0u,u)}{(B_0u,u)}(\delta Bu,u)\right)$$

Hence, using (30) and the minimax principle we obtain (31). □

Lemma 5 *Let $A = \Lambda - \delta A$ be a symmetric and $B = I - \delta B$ a symmetric positive definite matrix, where Λ is a diagonal matrix. Then*

$$\mathrm{Tr}(B^{-1}A) = \mathrm{Tr}({D_B}^{-1}D_A) - \mathrm{Tr}(A_1) + \mathrm{Tr}(A_2) + \mathrm{Tr}(A_3) \tag{32}$$

where

$$A_1 = {D_B}^{-1}\delta B {D_B}^{-1}(\delta A + D), \quad A_2 = {D_B}^{-1}D_{\delta B}{D_B}^{-1}(D_{\delta A} + D),$$
$$A_3 = {D_B}^{-1}\bar{D}_B B^{-1}\bar{D}_B {D_B}^{-1}A,$$
$$D_X \equiv \mathrm{Diag}X, \quad \bar{D}_X = D_X - X$$

and D is any diagonal matrix.

Proof. We have

$$\begin{aligned} B^{-1} = (D_B - \bar{D}_B)^{-1} &= {D_B}^{-1}(I - \bar{D}_B{D_B}^{-1})^{-1} \\ &= {D_B}^{-1}(I + \bar{D}_B{D_B}^{-1} + \bar{D}_B B^{-1}\bar{D}_B{D_B}^{-1}) \\ &= {D_B}^{-1} + {D_B}^{-1}\bar{D}_B{D_B}^{-1} + {D_B}^{-1}\bar{D}_B B^{-1}\bar{D}_B{D_B}^{-1} \end{aligned}$$

$$\begin{aligned} B^{-1}A &= ({D_B}^{-1} + {D_B}^{-1}\bar{D}_B{D_B}^{-1} + {D_B}^{-1}\bar{D}_B B^{-1}\bar{D}_B{D_B}^{-1})(D_A - \bar{D}_A) \\ &= {D_B}^{-1}D_A + {D_B}^{-1}\bar{D}_B{D_B}^{-1}D_A - {D_B}^{-1}\bar{D}_A - {D_B}^{-1}\bar{D}_B{D_B}^{-1}\bar{D}_A \\ &\quad + {D_B}^{-1}\bar{D}_B B^{-1}\bar{D}_B{D_B}^{-1}A \end{aligned}$$

Since $\mathrm{Tr}(D_X\bar{D}_Y D_Z) = 0$ for any X, Y and Z we have

$$\mathrm{Tr}({D_B}^{-1}\bar{D}_B{D_B}^{-1}D_A) = 0, \quad \mathrm{Tr}({D_B}^{-1}\bar{D}_A) = 0$$

and since $\delta X = D_{\delta X} - \bar{D}_{\delta X} = D_{\delta X} + \bar{D}_X$, $X = A$ or B, we obtain

$$\begin{aligned} \mathrm{Tr}({D_B}^{-1}\bar{D}_B{D_B}^{-1}\bar{D}_A) &= \mathrm{Tr}({D_B}^{-1}\delta B{D_B}^{-1}\bar{D}_A) \\ = \mathrm{Tr}({D_B}^{-1}\delta B{D_B}^{-1}(\delta A + D)) &- \mathrm{Tr}({D_B}^{-1}\delta B{D_B}^{-1}(D_{\delta A} + D)) \\ = \mathrm{Tr}({D_B}^{-1}\delta B{D_B}^{-1}(\delta A + D)) &- \mathrm{Tr}({D_B}^{-1}D_{\delta B}{D_B}^{-1}(D_{\delta A} + D)) \end{aligned}$$

which leads to (32). □

Lemma 6 *In the notation of lemma 5, let $\delta A = \delta A_- + \delta A_+$, where $\mathrm{Tr}(\delta A_- - D_-) \leq 0$ and $\mathrm{Tr}(\delta A_+ - D_+) \geq 0$ for some diagonal matrices D_- and D_+, and let $\delta B \geq 0$. Then*

$$-(\alpha_- + \alpha_+) \leq \mathrm{Tr}(B^{-1}A) - \mathrm{Tr}({D_B}^{-1}D_A) \leq \alpha_- + \alpha_+ + \beta \tag{33}$$

where

$$\alpha_- \equiv \frac{\|\delta B\|\,\mathrm{Tr}(D_- - \delta A_-)}{(1 - \|\delta B\|)^2}, \quad \alpha_+ \equiv \frac{\|\delta B\|\,\mathrm{Tr}(\delta A_+ - D_+)}{(1 - \|\delta B\|)^2}, \quad \beta \equiv \frac{\|A\|\|\delta B\|_F^2}{(1 - \|\delta B\|)^3}$$

Proof. By lemma 5 we have

$$\begin{aligned}\mathrm{Tr}(B^{-1}A) - \mathrm{Tr}({D_B}^{-1}D_A) &= -\mathrm{Tr}(A_1) + \mathrm{Tr}(A_2) + \mathrm{Tr}(A_3)\\ &= \mathrm{Tr}(A_{1-}) - \mathrm{Tr}(A_{2-}) - \mathrm{Tr}(A_{1+}) + \mathrm{Tr}(A_{2+}) + \mathrm{Tr}(A_3)\end{aligned}$$

where

$$\begin{aligned}A_{1-} &= {D_B}^{-1}\delta B {D_B}^{-1}(D_- - \delta A_-),\ A_{1+} = {D_B}^{-1}\delta B {D_B}^{-1}(\delta A_+ - D_+),\\ A_{2-} &= {D_B}^{-1}D_{\delta B}{D_B}^{-1}(D_- - D_{\delta A}),\ A_{2+} = {D_B}^{-1}D_{\delta B}{D_B}^{-1}(D_{\delta A} - D_+)\end{aligned}$$

Since $\delta B \geq 0$, $D_- - \delta A_- \geq 0$ and $\delta A_+ - D_+ \geq 0$, the matrices A_{1-}, A_{1+}, A_{2-}, A_{2+}, and A_3 are positive semi-definite. Using the first inequality in (29) and the fact that $\|D_B\| = \|I - D_{\delta B}\| \geq 1 - \|D_{\delta B}\| \geq 1 - \|\delta B\|$ we obtain for any $Z = Z^T \geq 0$

$$\begin{aligned}\mathrm{Tr}({D_B}^{-1}\delta B {D_B}^{-1} Z) &\leq \|{D_B}^{-1}\delta B {D_B}^{-1}\|\,\mathrm{Tr} Z\\ &\leq \|D_B\|^{-2}\|\delta B\|\,\mathrm{Tr} Z \leq \frac{\|\delta B\|}{(1-\|\delta B\|)^2}\,\mathrm{Tr} Z\end{aligned}$$

and, thus, $\mathrm{Tr}(A_{1-}) \leq \alpha_-$ and $\mathrm{Tr}(A_{1+}) \leq \alpha_+$. Since $\|D_{\delta B}\| \leq \|\delta B\|$, in a similar way we obtain $\mathrm{Tr}(A_{2-}) \leq \alpha_-$ and $\mathrm{Tr}(A_{2+}) \leq \alpha_+$. Finally, for A_3 we use the second inequality in (29) to obtain

$$\begin{aligned}\mathrm{Tr}(A_3) \leq \mathrm{Tr}({D_B}^{-1}\bar{D}_B B^{-1}\bar{D}_B {D_B}^{-1})\|A\| &= \|{D_B}^{-1}\bar{D}_B B^{-1/2}\|_F^2\|A\|\\ &\leq \|D_B\|^{-2}\|B^{-1}\|\|A\|\|\bar{D}_B\|_F^2 \leq \frac{\|A\|\|\delta B\|_F^2}{(1-\|\delta B\|)^3}\end{aligned}$$

□

Lemma 7 *If, in the notation of lemma 5, $A \geq 0$, $D_{\delta A} \leq 0$ and $\delta B \leq 0$ then*

$$Tr(B^{-1}A) \geq Tr({D_B}^{-1}D_A) - Tr(\delta B)\|\delta A\| \tag{34}$$

Proof. By lemma 5 we have (32), where$A_2 \geq 0$ and $A_3 \geq 0$. Hence, using lemma 1, we obtain

$$Tr(B^{-1}A) \geq Tr({D_B}^{-1}D_A) - Tr({D_B}^{-1}\delta B {D_B}^{-1})\|\delta A\|$$

Since $B = I - \delta B \geq I$ we have $D_B \geq I$, and, thus, the above inequality leads to (34).□

Lemma 8 *Let $A = A^T > 0$ and $B = B^T \geq I_n$ be matrices of the size n represented as 2-by-2 symmetric block matrices with the blocks A_{ij} and B_{ij} resp. Denote by ν_i the eigenvalues of $B^{-1}A$ and by ν_i^j the eigenvalues of $B_{jj}^{-1}A_{jj}$ enumerated in ascending order starting from 1. If $\nu_m^0 < \nu_{m+1}$, where m is the size of the block A_{00}, then*

$$0 \leq 1 - \frac{\nu_k^1}{\nu_{m+k}} \leq \sum_{i=m+1}^{n} \frac{\|A_{01} - \nu_i B_{01}\|^2}{(\nu_i - \nu_m^0)^2}, \quad k = 1, \dots, n-m \tag{35}$$

If $\nu_m < \nu_1^1$ then

$$0 \leq \frac{\nu_k^0}{\nu_k} - 1 \leq \sum_{i=1}^{m} \frac{\|A_{10} - \nu_i B_{10}\|^2}{(\nu_i - \nu_1^1)^2}, \quad k = 1, \dots, m \tag{36}$$

Proof. Denote by x_k the eigenvector of $B^{-1}A$ corresponding to ν_k and normalized by $\|x_k\|_B = 1$. Let u_k and v_k be vectors of dimension m and $n-m$ incorporating the first m and the last $n-m$ components of x_k resp. For $k > m$ we have

$$L_k u_k = R_k v_k \tag{37}$$

where $L_k \equiv -A_{00} + \nu_k B_{00}$ and $R_k \equiv A_{01} - \nu_k B_{01}$. Since $\nu_k > \nu_m^0$ we have $L_k \geq (\nu_k - \nu_m^0) B_{00} \geq \nu_k - \nu_m^0$, and, hence, from (37) we obtain

$$\begin{aligned} \|u_k\|_{L_k}^2 = \|R_k v_k\|_{L_k^{-1}}^2 \leq \frac{\|R_k v_k\|^2}{\nu_k - \nu_m^0} \leq \frac{\|R_k\|^2 \|v_k\|^2}{\nu_k - \nu_m^0} \leq \frac{\|R_k\|^2 \|x_k\|^2}{\nu_k - \nu_m^0} \\ \leq \frac{\|R_k\|^2 \|x_k\|_B^2}{\nu_k - \nu_m^0} = \frac{\|R_k\|^2}{\nu_k - \nu_m^0} \end{aligned}$$

and

$$\|u_k\|_{B_{00}}^2 \leq \frac{\|R_k\|^2}{(\nu_k - \nu_m^0)^2} \tag{38}$$

Let

$$x = \sum_{i=m+1}^{n} a_i x_i$$

and let u be the vector of dimension m incorporating first m components of x. We have

$$\|u\|_{B_{00}}^2 \leq \sum_{i=m+1}^{n} a_i^2 \sum_{i=m+1}^{n} \|u_i\|_{B_{00}}^2 \leq \|x\|_B^2 \sum_{i=m+1}^{n} \frac{\|R_i\|^2}{(\nu_i - \nu_m^0)^2}$$

from which it follows that

$$\theta_B^2 \leq \sum_{i=m+1}^{n} \frac{\|R_i\|^2}{(\nu_i - \nu_m^0)^2}$$

where θ_B is the gap in the norm $\|\cdot\|_B$ between the invariant subspace of $B^{-1}A$ corresponding to $n-m$ largest eigenvalues and the subspace of vectors of dimension n with the first m components equal to 0. Since ν_i^1 are the Ritz eigenvalues of the problem $A_{11} v = \nu^1 B_{11} v$ in the subspace $\tilde{\mathcal{I}}$, applying lemma 3.1 from [3] to the matrix $B_{11}^{-1/2} A_{11} B_{11}^{-1/2}$ we arrive at (35). Similar calculations lead to (36). □

A.3 Auxiliary results for the Rayleigh-Ritz approximation

A.3.1 General results

Lemma 9 *Let $\pi \equiv \pi_{\mathcal{I}}$, where $\mathcal{I}$ is an invariant subspace of M. For any non-zero $u \in \mathcal{E}$*

$$((M - \mu(u)I)\bar{\pi}u, \bar{\pi}u) = (\mu(u) - \mu(\pi u))\|\pi u\|^2$$

Proof. We have

$$\begin{aligned} 0 = ((\mu(u)I - M)u, u) &= ((\mu(u)I - M)\pi u, \pi u) + ((\mu(u)I - M)\bar{\pi}u, \bar{\pi}u) \\ &= (\mu(u) - \mu(\pi u))\|\pi u\|^2 - ((M - \mu(u)I)\bar{\pi}u, \bar{\pi}u) \end{aligned}$$

□

Corollary 2

$$\frac{\|\bar{\pi} u\|^2}{\|u\|^2} = \frac{\mu(\pi u) - \mu(u)}{\mu(\pi u) - \mu(\bar{\pi} u)}$$

Lemma 10 *Let $\mathcal{I}$ be an invariant subspace of M. For any non-zero $u \in \mathcal{E}$*

$$\delta_\mu(u, \mathcal{I}) \|\pi_{\mathcal{I}} u\| \le \|s(u)\|$$

Proof. Denote $M_u = (M - \mu(u)I)^2$. The eigenvalues of $M_u v$ are $\nu_i = (\mu_i - \mu(u))^2$, the eigenvectors being the same as for M. Therefore, $\|s(u)\|^2 = \|u\|^2_{M_u} \ge \|\pi_{\mathcal{I}} u\|^2_{M_u} \ge \delta_\mu(u, \mathcal{I})^2 \|\pi_{\mathcal{I}} u\|^2$. □

Lemma 11 *Let $\mathcal{I}$ be an invariant subspace of M and let $u \in \mathcal{E}$ be a non-zero vector. If $\delta_\mu(u, \mathcal{I}) > 0$ then*

$$-\frac{\|s(u)\|^2}{\delta_\mu(u, \mathcal{I}^-)} \le ((M - \mu(u)I)\pi_{\mathcal{I}} u, \pi_{\mathcal{I}} u) \le \frac{\|s(u)\|^2}{\delta_\mu(u, \mathcal{I}^+)} \tag{39}$$

where $\mathcal{I}^\pm = \mathcal{I}_{\sigma^\pm}$ and $\sigma^- = \{\mu \in \sigma_{\mathcal{I}} : \mu < \mu(u)\}$, $\sigma^+ = \{\mu \in \sigma_{\mathcal{I}} : \mu > \mu(u)\}$.

Proof. Consider the eigenvalue problem

$$\pi_{\mathcal{I}}(M - \mu(u)I)\pi_{\mathcal{I}} v = \nu M_u v$$

where M_u is defined in the proof of lemma 10. Non-zero eigenvalues of this problem are $\nu_i = (\mu(u) - \mu_i)^{-1}$, and, thus, $\delta_\mu(u, \mathcal{I}^-)^{-1} \le \nu_i \le \delta_\mu(u, \mathcal{I}^+)^{-1}$ which leads to (39) since $\|s(u)\|^2 = \|u\|^2_{M_u}$. □

A.3.2 Results related to extreme eigenpairs

In what follows $\mathcal{I}$ is the invariant subspace of M corresponding to m largest eigenvalues $\mu_0 \ge \mu_1 \ge \ldots \ge \mu_{m-1} > \mu_m$, $\pi \equiv \pi_{\mathcal{I}}$ and $\tilde{\mathcal{I}} \subset \mathcal{E}$ is a subspace of dimension m. For the gaps between $\tilde{\mathcal{I}}$ and $\mathcal{I}$ the notation $\theta \equiv \theta(\tilde{\mathcal{I}}, \mathcal{I})$ and $\theta_M \equiv \theta(\tilde{\mathcal{I}}, \mathcal{I})_M$ is used, and we denote $t \equiv \tan(\tilde{\mathcal{I}}, \mathcal{I})$. We also denote

$$\tilde{\mu}_i \equiv \mu_i(\tilde{\mathcal{I}}), \quad \tilde{u}_i \equiv u_i(\tilde{\mathcal{I}}), \quad r_i \equiv r(\tilde{u}_i), \quad s_i \equiv s(\tilde{u}_i), \quad \delta \equiv \delta_\mu(\tilde{\mathcal{I}}, \bar{\mathcal{I}})$$

Lemma 12 *If $\tilde{\mu}_{m-1} > \mu_m$ then*

$$\theta^2 \le \frac{\mu_0 - \tilde{\mu}_{m-1}}{\mu_0 - \mu_m} < 1$$

Proof. See corollary 2. □

Corollary 3 *If $\tilde{\mu}_{m-1} > \mu_m$ then*

$$t^2 \le \frac{\mu_0}{\tilde{\mu}_{m-1} - \mu_m} \theta_L^2$$

Lemma 13 *Let*

$$M_\mu \equiv \mu\pi + (\mu I - M)\bar{\pi} \tag{40}$$

If $\mu > \mu_m$ then

$$(\mu - \mu_m)I \le M_\mu \le \mu I \tag{41}$$

Proof. The eigenvalues ν_i of $M_\mu v$ are, obviously, $\nu_i = \mu$, $i = 0, \dots, m-1$, and $\nu_i = \mu - \mu_i$, $i \geq m$. Since $\nu_m \leq \nu_i \leq \mu$, we arrive at (41). $\square$

Lemma 14 *Let $\mathcal{I}' \subset \mathcal{I}$ be an invariant subspace of M, and denote $\pi' \equiv \pi_{\mathcal{I}'}$. For any non-zero $u \in \tilde{\mathcal{I}}$*

$$\delta_\mu(u, \mathcal{I}')\|\pi' u\| \leq \theta \|s(u)\| \tag{42}$$

Proof.[9] We have

$$\begin{aligned}(\mu(\pi' u) - \mu(u))\|\pi' u\|^2 &= ((M - \mu(u)I)\pi' u, \pi' u) \\ &= (s(u), \pi' u) = (s(u), (\pi - P_{\tilde{\mathcal{I}}})\pi' u)\end{aligned}$$

Hence, $|\mu(u) - \mu(\pi' u)|\|\pi' u\|^2 \leq \theta \|s(u)\|\|\pi' u\|$, leading to (42) $\square$

Lemma 15 *In the notation of lemma 14, let $u \in \tilde{\mathcal{I}}$ be a non-zero vector for which $\delta_\mu(u, I') > 0$. Then*

$$-\frac{\theta^2 \|s(u)\|^2}{\delta_\mu(u, \mathcal{I}^-)} \leq ((M - \mu(u)I)\pi' u, \pi' u) \leq \frac{\theta^2 \|s(u)\|^2}{\delta_\mu(u, \mathcal{I}^+)} \tag{43}$$

where $\mathcal{I}^\pm = \mathcal{I}_{\sigma^\pm}$ and $\sigma^- = \{\mu \in \sigma_{\mathcal{I}'} : \mu < \mu(u)\}$, $\sigma^+ = \{\mu \in \sigma_{\mathcal{I}'} : \mu > \mu(u)\}$.

Proof. Denoting $\pi^+ \equiv \pi_{\mathcal{I}^+}$ and using (42) we have

$$\begin{aligned}((M - \mu(u)I)\pi^+ u, \pi^+ u) &= (s(u), \pi^+ u) = (s(u), (\pi - P_{\tilde{\mathcal{I}}})\pi^+ u) \\ &\leq \theta \|s(u)\|\|\pi^+ u\| \leq \frac{\theta^2 \|s(u)\|^2}{\delta_\mu(u, \mathcal{I}^+)}\end{aligned}$$

which leads to the right-hand side inequality in (43). The left-hand one is obtained in a similar way. $\square$

Lemma 16

$$\|(I - P_{\tilde{\mathcal{I}},M})u\|^2 \leq (1 + \rho_{\tilde{\mathcal{I}}})\|u\|^2 \quad \forall u \in \mathcal{E}$$

Proof. We have

$$\begin{aligned}\|(I - P_{\tilde{\mathcal{I}},M})u\|^2 &= \|u - \sum_{i=0}^{m-1} (Mu, \tilde{u}_i)\tilde{u}_i\|^2 \\ &= \|u\|^2 - 2\sum_{i=0}^{m-1} (Mu, \tilde{u}_i)(u, \tilde{u}_i) + \sum_{i=0}^{m-1} \frac{1}{\tilde{\mu}_i}(Mu, \tilde{u}_i)^2 \\ &= \|u\|^2 - 2\sum_{i=0}^{m-1} \frac{1}{\tilde{\mu}_i}(Mu, \tilde{u}_i)(u, s_i) - \sum_{i=0}^{m-1} \frac{1}{\tilde{\mu}_i}(Mu, \tilde{u}_i)^2 \\ &\leq \|u\|^2 + \sum_{i=0}^{m-1} \frac{1}{\tilde{\mu}_i}\left((Mu, \tilde{u}_i)^2 + (u, s_i)^2\right) - \sum_{i=0}^{m-1} \frac{1}{\tilde{\mu}_i}(Mu, \tilde{u}_i)^2 \\ &= \|u\|^2 + \sum_{i=0}^{m-1} \frac{1}{\tilde{\mu}_i}(u, s_i)^2 \leq \|u\|^2 + \|u\|^2 \sum_{i=0}^{m-1} \frac{1}{\tilde{\mu}_i}\|s_i\|^2\end{aligned}$$

$\square$

[9] Cf. also [10].

Lemma 17 *If $\theta < 1$ then $\|s_i\|^2 \le \|\bar{\pi} s_i\|^2/(1-\theta^2)$.*

Proof. Let $P = P_{\tilde{\mathcal{I}}}$. Since $Ps_i = 0$, we have

$$\|s_i\|^2 = \|\pi s_i\|^2 + \|\bar{\pi} s_i\|^2 = \|(\pi - P)s_i\|^2 + \|\bar{\pi} s_i\|^2 \le \theta^2 \|s_i\|^2 + \|\bar{\pi} s_i\|^2$$

□

Lemma 18 *If $\theta < 1$ then $\|s_i\|^2 \le \tilde{\mu}_i t^2$.*

Proof. Denoting $M_i \equiv M_{\tilde{\mu}_i}$, where M_μ is given by (40), we have $\bar{\pi} s_i = M_i \bar{\pi} \tilde{u}_i$, and, using (41) and lemma 17, we obtain

$$\|s_i\|^2 \le \frac{\|M_i \bar{\pi} \tilde{u}_i\|^2}{1-\theta^2} \le \tilde{\mu}_i^2 \frac{\|\bar{\pi} \tilde{u}_i\|^2}{1-\theta^2} \le \frac{\theta^2}{1-\theta^2} \tilde{\mu}_i^2 \|\tilde{u}_i\|^2 = t^2 \tilde{\mu}_i$$

□

Corollary 4 *If $\theta < 1$ then*

$$\rho_{\tilde{\mathcal{I}}} \le m t^2 \le \frac{m \mu_0}{\tilde{\mu}_{m-1} - \mu_m}$$

A.3.3 Results related to internal eigenpairs

Let $0 \le l < k \le m$. In the sequel we use the following notation

$$\mathcal{I}_- \equiv \operatorname{span}\{u_i\}_{i=0,l-1}, \quad \mathcal{I}_0 \equiv \operatorname{span}\{u_i\}_{i=l,k-1},$$

$$\mathcal{I}_+ \equiv \operatorname{span}\{u_i\}_{i=k,m-1}, \quad \mathcal{I}_* \equiv \mathcal{I}_- + \mathcal{I}_0$$

$$m_- \equiv l, \quad m_0 \equiv k - l, \quad m_+ \equiv m - k,$$

$$\pi_s \equiv \pi_{\mathcal{I}_s}, \quad s \in \{-, 0, +, *\}$$

$$\chi_s = \begin{cases} 0, & m_s = 0 \\ 1, & m_s > 0 \end{cases} \qquad s \in \{-, +\}, \quad \chi_\pm = \max\{\chi_-, \chi_+\}$$

Accordingly, we denote $\tilde{\mathcal{I}}_- = \operatorname{span}\{\tilde{u}_i\}_{i=0,l-1}$ etc., and

$$\theta_s \equiv \theta(\tilde{\mathcal{I}}_s, \mathcal{I}_s), \quad t_s \equiv \tan(\tilde{\mathcal{I}}_s, \mathcal{I}_s), \quad \rho_s \equiv \rho_{\tilde{\mathcal{I}}_s}, \quad s \in \{-, 0, +, *\}$$

Further,

$$\eta_0 \equiv \frac{\tilde{\mu}_{k-1}}{\tilde{\mu}_{k-1} - \mu_m}$$

For $l > 0$ we denote

$$\eta_- \equiv \frac{\tilde{\mu}_l}{\mu_{l-1} - \tilde{\mu}_l}, \quad \tilde{\eta}_- = \frac{\mu_l}{\tilde{\mu}_{l-1} - \mu_l}$$

and for $l = 0$ we set $\eta_- = 0$ and $\tilde{\eta}_- = 0$. For $k < m$ we denote

$$\eta_+ \equiv \frac{\tilde{\mu}_{k-1}}{\tilde{\mu}_{k-1} - \mu_k}$$

and for $k = m$ we set $\eta_+ = 0$. Finally, $\eta_\pm = \max\{\eta_-, \eta_+\}$.

Lemma 19 *If $\tilde\mu_{l-1} \ge \mu_l$ and $\tilde\mu_{k-1} \ge \mu_k$, $0 < l < k < m$, then*

$$\theta_0^2 \le \frac{\mu_l - \tilde\mu_{k-1}}{\tilde\mu_{l-1} - \tilde\mu_{k-1}} \frac{\tilde\mu_{l-1} - \mu_k}{\mu_l - \mu_k} \tag{44}$$

Proof. See [8]; cf. also [26]. □

Lemma 20 *If $\tilde\mu_{l-1} > \mu_l$ and $\tilde\mu_{k-1} > \mu_k$ then*

$$-\alpha_0 t_0^2 \sum_{i=l}^{k-1} \|s_i\|^2 \le \sum_{i=l}^{k-1} (\mu_i - \mu(\pi_0 \tilde u_i)) \le (\alpha_0 + \beta_0) t_0^2 \sum_{i=l}^{k-1} \|s_i\|^2 \tag{45}$$

where

$$\alpha_0 = (1 + t_0^2)(\eta_0 + (\eta_- + \eta_+)\theta^2)$$

$$\beta_0 = (k - l)(1 + t_0^2)^2 \left(\frac{\mu_l}{\mu_k} \eta_0^2 + (\eta_-^2 + \frac{\mu_l}{\mu_k} \eta_+^2)\theta^2 \right)$$

Proof. Denote $\tilde v_i = \sqrt{\tilde\mu_i}\tilde u_i$. By lemma 19 we have $\theta_0 < 1$ and hence the vectors $\tilde v_i^0 \equiv \pi_0 \tilde v_i$, $i = l, \ldots, k-1$, form a basis of $\mathcal{I}_0$. Using these vectors as the basis for the Rayleigh-Ritz method for M we obtain the generalized eigenvalue problem

$$Ax = \mu B x \tag{46}$$

where A and B are m_0-by-m_0 matrices with the entries $a_{ij} = (\tilde v_i^0, \tilde v_j^0)_M$ and $b_{ij} = (\tilde v_i^0, \tilde v_j^0)$ resp. Obviously, μ_i, $i = l, \ldots, k-1$, are the eigenvalues of (46), and $\mu(\pi \tilde v_i) = a_{ii}/b_{ii}$. This suggests using lemma 6 for obtaining the estimates (45). Since

$$\delta_{ij} = (\tilde v_i, \tilde v_j) = (\tilde v_i^0, \tilde v_j^0) + (\bar\pi_0 \tilde v_i, \bar\pi_0 \tilde v_j)$$

and

$$\delta_{ij}\tilde\mu_i = (\tilde v_i, \tilde v_j)_M = (\tilde v_i^0, \tilde v_j^0)_M + (\pi_- \tilde v_i, \pi_- \tilde v_j)_M + (\pi' \tilde v_i, \pi' \tilde v_j)_M$$

where $\pi' \equiv \pi_+ + \bar\pi$, we have: $A = \Lambda - \delta A_- - \delta A_+$ and $B = I - \delta B$ where $\Lambda = \mathrm{Diag}\{\tilde\mu_i\}_{i=l,k-1}$ and the matrices δA_-, δA_+ and δB have the entries $(\pi_- \tilde v_i, \pi_- \tilde v_j)_M$, $(\pi' \tilde v_i, \pi' \tilde v_j)_M$ and $(\bar\pi_0 \tilde v_i, \bar\pi_0 \tilde v_j)_M$ resp. Obviously, $\delta A_- + \delta A_+ \ge 0$, and, hence, $\|A\| \le \tilde\mu_l$. Further,

$$\|\delta B\| = \max_x \frac{(\delta B x, x)}{\|x\|^2} = \max_{v \in \tilde{\mathcal{I}}} \frac{\|\bar\pi_0 v\|^2}{\|v\|^2} = \theta_0^2$$

Let D_- and D_+ be diagonal m_0-by-m_0 matrices with the diagonal entries $\tilde\mu_i(\pi_- \tilde v_i, \pi_- \tilde v_i)$ and $\tilde\mu_i(\pi' \tilde v_i, \pi' \tilde v_i)$ resp. We have: $\mathrm{Tr}(\delta A_- - D_-) \ge 0$ and $\mathrm{Tr}(\delta A_+ - D_+) \le 0$. Furthermore, using lemmas 15 and 11 we obtain

$$\begin{aligned}
\mathrm{Tr}(\delta A_- - D_-) &= \sum_{i=l}^{k-1} ((M - \tilde\mu_i I)\pi_- \tilde v_i, \pi_- \tilde v_i) \\
&\le \frac{\chi_- \theta^2}{\delta_\mu(\tilde{\mathcal{I}}_0, \mathcal{I}_-)} \sum_{i=l}^{k-1} \tilde\mu_i \|s_i\|^2 \le \eta_- \theta^2 \sum_{i=l}^{k-1} \|s_i\|^2
\end{aligned}$$

and

$$\mathrm{Tr}(D_+ - \delta A_+) = \sum_{i=l}^{k-1} ((\tilde{\mu}_i I - M)\pi' \tilde{v}_i, \pi' \tilde{v}_i)$$
$$= \sum_{i=l}^{k-1} ((\tilde{\mu}_i I - M)\pi_+ \tilde{v}_i, \pi_+ \tilde{v}_i) + \sum_{i=l}^{k-1} ((\tilde{\mu}_i I - M)\bar{\pi} \tilde{v}_i, \bar{\pi} \tilde{v}_i)$$
$$\le \frac{\chi_+ \theta^2}{\delta_\mu(\tilde{\mathcal{I}}_0, \mathcal{I}_+)} \sum_{i=l}^{k-1} \tilde{\mu}_i \|s_i\|^2 + \frac{1}{\delta_\mu(\tilde{\mathcal{I}}_0, \bar{\mathcal{I}})} \sum_{i=l}^{k-1} \tilde{\mu}_i \|s_i\|^2$$
$$\le \eta_+ \theta^2 \sum_{i=l}^{k-1} \|s_i\|^2 + \eta_0 \sum_{i=l}^{k-1} \|s_i\|^2$$

Finally, using lemmas 10 and 14 we obtain

$$\|\delta B\|_F \le \sum_{i=l}^{k-1} \|\bar{\pi}_0 \tilde{v}_i\|^2 = \sum_{i=l}^{k-1} \|\pi_- \tilde{v}_i\|^2 + \sum_{i=l}^{k-1} \|\pi_+ \tilde{v}_i\|^2 + \sum_{i=l}^{k-1} \|\bar{\pi} \tilde{v}_i\|^2$$
$$\le \left(\frac{\chi_- \theta^2}{\delta_\mu(\tilde{\mathcal{I}}_0, \mathcal{I}_-)^2} + \frac{\chi_+ \theta^2}{\delta_\mu(\tilde{\mathcal{I}}_0, \mathcal{I}_+)^2} + \frac{1}{\delta_\mu(\tilde{\mathcal{I}}_0, \bar{\mathcal{I}})^2} \right) \sum_{i=l}^{k-1} \tilde{\mu}_i \|s_i\|^2$$
$$\le \left(\frac{\chi_- \theta^2}{\delta_\mu(\tilde{\mathcal{I}}_0, \mathcal{I}_-)^2} + \frac{\chi_+ \theta^2}{\delta_\mu(\tilde{\mathcal{I}}_0, \mathcal{I}_+)^2} + \frac{1}{\delta_\mu(\tilde{\mathcal{I}}_0, \bar{\mathcal{I}})^2} \right) \tilde{\mu}_l \sum_{i=l}^{k-1} \|s_i\|^2$$

which leads to (45). □

Lemma 21 *If $\tilde{\mu}_{l-1} > \mu_l$ and $\tilde{\mu}_{k-1} > \mu_k$ then for $l \le i < k$*

$$|\mu(\pi_0 \tilde{u}_i) - \tilde{\mu}_i| \le (1 + \eta_\pm t)(1 + t_0^2)\sqrt{\tilde{\mu}_i}\theta \|s_i\| \tag{47}$$

$$\mu(\pi_0 \tilde{u}_i) - \tilde{\mu}_i \le (1 + \eta_+ t^2)(1 + t_0^2)\tilde{\mu}_i \theta^2 \tag{48}$$

$$\mu(\pi_0 \tilde{u}_i) - \tilde{\mu}_i \le (\eta_0 + \eta_+ \theta^2)(1 + t_0^2)\|s_i\|^2 \tag{49}$$

Proof. Using lemmas 14 and 18 we obtain

$$|((\tilde{\mu}_i I - M)\bar{\pi}_0 \tilde{u}_i, \bar{\pi}_0 \tilde{u}_i)| \le |((\tilde{\mu}_i I - M)\bar{\pi} \tilde{u}_i, \bar{\pi} \tilde{u}_i)|$$
$$+ |((\tilde{\mu}_i I - M)\pi_\pm \tilde{u}_i, \pi_\pm \tilde{u}_i)|$$
$$= |(s_i, \bar{\pi} \tilde{u}_i)| + |(s_i, \pi_\pm \tilde{u}_i)| \le (\|\bar{\pi} \tilde{u}_i\| + \|\pi_\pm \tilde{u}_i\|)\|s_i\|$$
$$\le \left(\frac{\theta}{\sqrt{\tilde{\mu}_i}} + \frac{\chi_\pm \theta \|s_i\|}{\delta_\mu(\tilde{u}_i, \bar{\mathcal{I}}_0)} \right) \|s_i\| \le \left(\frac{1}{\sqrt{\tilde{\mu}_i}} + \frac{\chi_\pm t \sqrt{\tilde{\mu}_i}}{\delta_\mu(\tilde{u}_i, \bar{\mathcal{I}}_0)} \right) \theta \|s_i\|$$
$$\le \left(1 + \frac{\chi_\pm \tilde{\mu}_i t}{\delta_\mu(\tilde{u}_i, \bar{\mathcal{I}}_0)} \right) \frac{\theta \|s_i\|}{\sqrt{\tilde{\mu}_i}} \le (1 + \eta_\pm t) \frac{\theta \|s_i\|}{\sqrt{\tilde{\mu}_i}}$$

Further,

$$((\tilde{\mu}_i I - M)\bar{\pi}_0 \tilde{u}_i, \bar{\pi}_0 \tilde{u}_i) \le ((\tilde{\mu}_i I - M)\bar{\pi} \tilde{u}_i, \bar{\pi} \tilde{u}_i)$$
$$+ ((\tilde{\mu}_i I - M)\pi_+ \tilde{u}_i, \pi_+ \tilde{u}_i)$$

Therefore, using lemmas 10 and 15 we obtain

$$
\begin{aligned}
((\tilde{\mu}_i I - M)\bar{\pi}_0 \tilde{u}_i, \bar{\pi}_0 \tilde{u}_i) &\leq \frac{\|s_i\|^2}{\delta_\mu(\tilde{u}_i, \bar{\mathcal{I}})} + \frac{\chi_+ \theta^2 \|s_i\|^2}{\delta_\mu(\tilde{u}_i, \mathcal{I}_+)} \\
\leq \left(\frac{\tilde{\mu}_i}{\delta_\mu(\tilde{u}_i, \bar{\mathcal{I}})} + \frac{\chi_+ \tilde{\mu}_i}{\delta_\mu(\tilde{u}_i, \mathcal{I}_+)} \theta^2 \right) \frac{\|s_i\|^2}{\tilde{\mu}_i} &\leq (\eta_0 + \eta_+ \theta^2) \frac{\|s_i\|^2}{\tilde{\mu}_i}
\end{aligned}
\tag{50}
$$

and

$$
\begin{aligned}
((\tilde{\mu}_i I - M)\bar{\pi}_0 \tilde{u}_i, \bar{\pi}_0 \tilde{u}_i) &\leq \theta^2 + \frac{\chi_+ \theta^2 \|s_i\|^2}{\delta_\mu(\tilde{u}_i, \mathcal{I}_+)} \\
\leq \left(1 + \frac{\chi_+ \tilde{\mu}_i}{\delta_\mu(\tilde{u}_i, \mathcal{I}_+)} t^2 \right) \theta^2 &\leq (1 + \eta_+ t^2) \theta^2
\end{aligned}
$$

Using lemma 9 and the inequality

$$
\|\pi_0 \tilde{u}_i\|^2 \geq (1 - \theta^2) \|\tilde{u}_i\|^2 = \frac{1}{\tilde{\mu}_i (1 + t_0^2)}
$$

we arrive at (47) – (49). □

A.4 Auxiliary results for gradient type eigensolvers

Lemma 22 *Let A be an k-by-k matrix with the entries $a_{ij} = (\hat{u}_i, \hat{u}_j)$ where $0 < k \leq m$ and*

$$
\hat{u}_i = \tilde{u}_i - \tau (I - P) K r_i, \quad P \equiv P_{\tilde{\mathcal{I}}_*, M}, \quad \tau = \frac{1}{(1 + \rho_*) b}
$$

If $K = K^T \in \mathcal{L}(\mathcal{E})$ satisfies (28) then

$$
A \leq A_0 - \tau A_1 \tag{51}
$$

where $A_0 = \mathrm{Diag}\{\tilde{\lambda}_i\}_{i=0,k-1}$ and A_1 is an k-by-k matrix with the entries $(r_i, r_j)_K$.

Proof. Denoting $g_i = \bar{P} K r_i$, we have:

$$
\begin{aligned}
a_{ij} = (\hat{u}_i, \hat{u}_j) &= \tilde{\lambda}_i \delta_{ij} - 2\tau (\tilde{u}_i, g_j) + \tau^2 (g_i, g_j) \\
&= \tilde{\lambda}_i \delta_{ij} - 2\tau (\tilde{u}_i - \tilde{\lambda}_i M \tilde{u}_i, g_j) + \tau^2 (g_i, g_j) \\
&= \tilde{\lambda}_i \delta_{ij} - 2\tau (r_i, K r_j) + \tau^2 (\bar{P} K r_i, \bar{P} K r_j)
\end{aligned}
$$

that is, $A = A_0 - 2\tau A_1 + \tau^2 A_2$, where A_2 has the entries $(\bar{P} K r_i, \bar{P} K r_j)$. Let

$$
u = \sum_{i=0}^{k-1} x_i r_i
$$

Using (28) and lemma 16 we have

$$
(A_2 x, x) = \|\bar{P} K u\|^2 \leq (1 + \rho_*) \|K u\|^2 \leq (1 + \rho_*) b \|u\|_K^2 = \tau^{-1} (A_1 x, x)
$$

and, hence, $A \leq A_0 - \tau A_1$. □

Lemma 23 *Let* $\tilde{\mathcal{I}}' = \operatorname{span}\{\hat{u}_i\}_{i=0,k-1}$*, where* $\hat{u}_i$ *are given in lemma 22 and* $0 < k \le m$*. If* $K = K^T$ *satisfies* (28) *then* $\tilde{\mu}'_i \equiv \mu_i(\tilde{\mathcal{I}}') \ge \tilde{\mu}_i$*,* $i = 0, \dots, k-1$*, and*

$$\sum_{i=0}^{k-1}(\tilde{\mu}'_i - \tilde{\mu}_i) \ge \frac{1}{1+\rho_*}\frac{a}{b}\sum_{i=0}^{k-1}\|s_i\|^2 \tag{52}$$

Proof. Denoting $g_i = \bar{P}Kr_i$, and using linear independent vectors $\hat{u}_i = \tilde{u}_i - \tau g_i$ as the basis for the Rayleigh-Ritz method in $\tilde{\mathcal{I}}'$ we see that $\tilde{\mu}'_i$ are the eigenvalues of the problem

$$Bx = \tilde{\mu}' Ax$$

where A and B are k-by-k matrices with entries $a_{ij} = (\hat{u}_i, \hat{u}_j)$ and $b_{ij} = (\hat{u}_i, \hat{u}_j)_M$ resp. We have:

$$b_{ij} = \delta_{ij} + \tau^2(g_i, g_j)_M$$

that is, $B \ge I$, which, together with (51), implies that $\tilde{\mu}'_i \ge \tilde{\mu}_i$. Further, $A_0 - \tau A_1 \ge A \ge \lambda_0 B > 0$, and, hence, by corollary 1 we have $A^{-1} \ge A_0^{-1} + \tau A_0^{-1}A_1A_0^{-1}$, which yields

$$\begin{aligned}\sum_{i=0}^{k-1}\tilde{\mu}'_i \ge \operatorname{Tr}(A^{-1}) \ge \operatorname{Tr}(A_0^{-1}) + \tau\operatorname{Tr}(A_0^{-1}A_1A_0^{-1}) &= \sum_{i=0}^{k-1}(\tilde{\mu}_i + \tau\tilde{\mu}_i^2\|r_i\|_K^2)\\ &= \sum_{i=0}^{k-1}\tilde{\mu}_i + \tau\sum_{i=0}^{k-1}\|s_i\|_K^2 \ge \sum_{i=0}^{k-1}\tilde{\mu}_i + \frac{1}{1+\rho_*}\frac{a}{b}\sum_{i=0}^{k-1}\|s_i\|^2\end{aligned}$$

□

Corollary 5

$$\sum_{i=0}^{k-1}\|s_i\|^2 \le (1+\rho_*)\sum_{i=0}^{k-1}(\mu_i - \tilde{\mu}_i)$$

Lemma 24 *Let* $\tilde{\mu}'_i$ *be as in lemma 23. If* $\tilde{\mu}_{l-1} > \mu_l$ *and*

$$\sum_{j=0}^{l-1}(\mu_j - \tilde{\mu}_j) \le \frac{a}{b}\left(3 + \left(\frac{\mu_0}{\mu_k} - 1\right)\theta_{M,*}^2\right)^{-2}\frac{\tilde{\mu}_{k-1}^2}{\mu_0\tilde{\eta}_-^2}\frac{\xi}{(k-l)^2} \tag{53}$$

where $\xi < 1$ *and* $\theta_{M,*} = \theta(\tilde{\mathcal{I}}_*, \mathcal{I}_*)_M$*, then*

$$\sum_{i=l}^{k-1}(\tilde{\mu}'_i - \tilde{\mu}_i) \ge (1-\xi)\frac{1}{1+\rho_*}\frac{a}{b}\sum_{i=l}^{k-1}\|s_i\|^2$$

Proof. In the notation of the proof of lemma 23, let us split both A and B into the blocks A_{ij} and B_{ij}, $i = 0, 1$, $j = 0, 1$, resp., where A_{00} and B_{00} incorporate the entries a_{ij} and b_{ij} resp. with $0 \le i < l$ and $0 \le j < l$ etc. For the entries r_{ij} of the matrix $R_n = A_{01} - \tilde{\lambda}'_n B_{01}$, where $l \le n < k$ and $\tilde{\lambda}'_n = (\tilde{\mu}'_n)^{-1}$, we have

$$\begin{aligned}r_{ij} = ((I - \tilde{\lambda}'_n M)\hat{u}_i, \hat{u}_j) &= -((I - \tilde{\lambda}'_n M)\tilde{u}_i, g_j) - ((I - \tilde{\lambda}'_n M)\tilde{u}_j, g_i)\\ &+ ((I - \tilde{\lambda}'_n M)g_i, g_j) = -((I - \tilde{\lambda}_i M)\tilde{u}_i, g_j) - ((I - \tilde{\lambda}_i M)\tilde{u}_j, g_i)\\ &+ ((I - \tilde{\lambda}'_n M)g_i, g_j) = -2\tau(r_i, Kr_j) + \tau^2((I - \tilde{\lambda}'_n M)g_i, g_j)\end{aligned}$$

$$
\begin{aligned}
((I-\tilde{\lambda}'_n M)g_i, g_j) &= ((I-\tilde{\lambda}'_n M)\pi_* g_i, \pi_* g_j) + ((I-\tilde{\lambda}'_n M)\bar{\pi}_* g_i, \bar{\pi}_* g_j) \\
&\le \tau^2(\lambda_k-\lambda_0)\|\pi_*\bar{P}Kr_i\|_M\|\pi_*\bar{P}Kr_j\|_M + \tau^2\|\bar{\pi}_*\bar{P}Kr_i\|\|\bar{\pi}_*\bar{P}Kr_j\| \\
&\le \tau^2(\lambda_k-\lambda_0)\theta_{M,*}^2\|Kr_i\|_M\|Kr_j\|_M + \tau^2(1+\rho_*)\|Kr_i\|\|Kr_j\| \\
&\le \tau^2\left(\frac{\lambda_k-\lambda_0}{\lambda_0}\theta_{M,*}^2 + 1 + \rho_*\right)\|Kr_i\|\|Kr_j\| \\
&\le b^2\tau^2\left(\frac{\lambda_k-\lambda_0}{\lambda_0}\theta_{M,*}^2 + 1 + \rho_*\right)\|r_i\|\|r_j\|
\end{aligned}
$$

Thus

$$
|r_{ij}| \le (2 + (1+\rho_* + \frac{\lambda_k-\lambda_0}{\lambda_0}\theta_M^2)b\tau)b\tau\|r_i\|\|r_j\| \le \frac{\psi}{1+\rho_*}\|r_i\|\|r_j\|
$$

where

$$
\psi = 3 + \frac{\lambda_k-\lambda_0}{\lambda_0}\theta_{M,*}^2
$$

and, hence,

$$
\begin{aligned}
\|R_n\|^2 \le \|R_n\|_F^2 &\le \frac{\psi^2}{(1+\rho_*)^2}\sum_{i=l}^{k-1}\|r_i\|^2\sum_{j=0}^{l-1}\|r_j\|^2 \\
&\le \frac{\psi^2\tilde{\lambda}_{k-1}^2}{(1+\rho_*)^2}\sum_{i=l}^{k-1}\|s_i\|^2\sum_{j=0}^{l-1}\|r_j\|^2
\end{aligned}
$$

Denote by $\hat{\mu}_i$, $i=0,\dots,l-1$, the eigenvalues of $A_{00}^{-1}B_{00}$ and by $\hat{\mu}_i$, $i=l,\dots,k-1$, the eigenvalues of $A_{11}^{-1}B_{11}$ enumerated in ascending order. From $A \le A_0 - \tau A_1$ and from $B_{ii} \ge I$ it follows that $\hat{\mu}_i \ge \tilde{\mu}_i$ and, furthermore,

$$
\sum_{i=0}^{k-1}(\hat{\mu}_i - \tilde{\mu}_i) \ge \frac{1}{1+\rho_*}\frac{a}{b}\sum_{i=l}^{k-1}\|s_i\|^2
$$

(cf. the proof of lemma 23). Since $\tilde{\lambda}'_l \ge \lambda_l > \tilde{\lambda}_{l-1} \ge \hat{\lambda}_{l-1} \equiv (\hat{\mu}_{l-1})^{-1}$, we can apply lemma 8 (with $n=k$, $m=l$, $\nu_i = \tilde{\lambda}'_{i+1}$, $\nu_i^0 = \hat{\lambda}_{i+1}$ and $\nu_i^1 = \hat{\lambda}_{i+l-1}$) to obtain for $l \le p < k$

$$
\begin{aligned}
1 - \frac{\tilde{\mu}'_p}{\hat{\mu}_p} = 1 - \frac{\hat{\lambda}_p}{\tilde{\lambda}'_p} &\le \sum_{j=l}^{k-1}\frac{\|R_j\|^2}{(\tilde{\lambda}'_j - \tilde{\lambda}_{l-1})^2} \le \frac{1}{(\lambda_l - \tilde{\lambda}_{l-1})^2}\sum_{j=l}^{k-1}\|R_j\|^2 \\
&\le (k-l)\frac{\tilde{\lambda}_{k-1}^2\tilde{\lambda}_{l-1}^2}{(\lambda_l - \tilde{\lambda}_{l-1})^2}\frac{\psi^2}{(1+\rho_*)^2}\sum_{i=l}^{k-1}\|s_i\|^2\sum_{j=0}^{l-1}\|s_j\|^2 \\
&= (k-l)\frac{\tilde{\eta}_-^2}{\tilde{\mu}_{k-1}^2}\frac{\psi^2}{(1+\rho_*)^2}\sum_{i=l}^{k-1}\|s_i\|^2\sum_{j=0}^{l-1}\|s_j\|^2
\end{aligned}
$$

Thus,

$$
\begin{aligned}
\sum_{i=l}^{k-1}(\tilde{\mu}'_i - \tilde{\mu}_i) = \sum_{i=l}^{k-1}(\hat{\mu}_i - \tilde{\mu}_i) - \sum_{i=l}^{k-1}(\hat{\mu}_i - \tilde{\mu}'_i) &\ge \frac{1}{1+\rho_*}\frac{a}{b}\sum_{i=l}^{k-1}\|s_i\|^2 \\
&- (k-l)^2\frac{\mu_0\tilde{\eta}_-^2}{\tilde{\mu}_{k-1}^2}\frac{\psi^2}{(1+\rho_*)^2}\sum_{i=l}^{k-1}\|s_i\|^2\sum_{j=0}^{l-1}\|s_j\|^2
\end{aligned}
$$

$$\geq \left(1 - \psi^2 \frac{b}{a}(k-l)^2 \frac{\mu_0 \tilde{\eta}_-^2}{\tilde{\mu}_{k-1}^2} \sum_{j=0}^{l-1}(\mu_j - \tilde{\mu}_j)\right) \frac{1}{1+\rho_*} \frac{a}{b} \sum_{i=l}^{k-1} \|s_i\|^2$$

$$= (1-\xi) \frac{1}{1+\rho_*} \frac{a}{b} \sum_{i=l}^{k-1} \|s_i\|^2$$

□

Remark 3 *From lemma 17 it follows that if we replace the left-hand side inequality in* (28) *with* $K \geq a\bar{\pi}$ *then* (52) *becomes (cf. the last step in the proof of lemma 23)*

$$\sum_{i=0}^{k-1}(\tilde{\mu}_i' - \tilde{\mu}_i) \geq \frac{1-\theta^2}{1+\rho_*} \frac{a}{b} \sum_{i=0}^{k-1} \|s_i\|^2$$

and, consequently, the right-hand side of (53) *becomes multiplied by* $1-\theta^2$.

A.5 Auxiliary results related to theorems 1–3

Lemma 25 *Let* $P \equiv P_{\tilde{\mathcal{I}}}$, $M_i \equiv M_{\tilde{\mu}_i}$, *where* M_μ *is given by* (40), *and let*

$$T_i = \bar{\pi}\bar{P}K\bar{P}\bar{\pi}(\tilde{\mu}_i I - M)$$

If $\tilde{\mu}_{m-1} > \mu_m$ *and* $K = K^T$ *satisfies* (28) *then for any* $u \in \mathcal{E}$

$$a_i \|\bar{\pi}u\|_{M_i}^2 \leq (T_i u, u)_{M_i} \leq b_i \|\bar{\pi}u\|_{M_i}^2 \tag{54}$$

where $a_i = a(\tilde{\mu}_i - \mu_m)(1-\theta^2)^2$ *and* $b_i = b\tilde{\mu}_i$.

Proof. Since $T_i = \bar{\pi}\bar{P}K\bar{P}\bar{\pi}M_i$, we have

$$(T_i u, u)_{M_i} = (\bar{\pi}\bar{P}K\bar{P}\bar{\pi}M_i u, M_i u) = \|\bar{P}M_i\bar{\pi}u\|_K^2$$

and, in view of (28),

$$a\|\bar{P}M_i\bar{\pi}u\|^2 \leq (T_i u, u)_{M_i} \leq b\|\bar{P}M_i\bar{\pi}u\|^2$$

For $\|\bar{P}M_i\bar{\pi}u\|$ we easily obtain the estimates (cf. (24) for $\|\bar{\pi}P\bar{\pi}\|$)

$$\begin{aligned}\|M_i\bar{\pi}u\| \geq \|\bar{P}M_i\bar{\pi}u\| \geq \|\bar{\pi}\bar{P}M_i\bar{\pi}u\| &\geq \|M_i\bar{\pi}u\| - \|\bar{\pi}P\bar{\pi}M_i\bar{\pi}u\| \\ &\geq \|M_i\bar{\pi}u\| - \theta^2\|M_i\bar{\pi}u\| = (1-\theta^2)\|M_i\bar{\pi}u\|\end{aligned}$$

and from (41) we have

$$(\tilde{\mu}_i - \mu_m)\|\bar{\pi}u\|_{M_i}^2 \leq \|M_i\bar{\pi}u\|^2 \leq \tilde{\mu}_i\|\bar{\pi}u\|_{M_i}^2$$

which leads to (54). □

Corollary 6 *In the notation and under the assumptions of lemma 25, for*

$$\tau_i = \frac{2}{a_i + b_i}$$

we have

$$\|I - \tau_i T_i\|_{M_i} \leq q_i \equiv \frac{b_i - a_i}{b_i + a_i}$$

Remark 4 *From the proof of lemma 25 we observe that the above results for T_i remain valid if the left-hand side inequality in* (28) *is replaced with $K \ge a\bar{\pi}$.*

Lemma 26 *In the notation and under the assumptions of corollary 6, for*

$$\hat{u}_i = \tilde{u}_i - \tau_i \bar{P} K s_i$$

we have

$$\|\bar{\pi}\hat{u}_i\|_{M_i} \le q_i \|\bar{\pi}\tilde{u}_i\|_{M_i} + 2\theta^2 \frac{\|s_i\|}{\sqrt{\tilde{\mu}_i}} \tag{55}$$

Proof. We have

$$\begin{aligned} \bar{\pi}\hat{u}_i = \bar{\pi}\tilde{u}_i - \tau_i \bar{\pi}\bar{P}K s_i &= \bar{\pi}\tilde{u}_i - \tau_i \bar{\pi}\bar{P}K\bar{P} s_i = \bar{\pi}\tilde{u}_i - \tau_i \bar{\pi}\bar{P}K\bar{P}(\bar{\pi} + \pi) s_i \\ &= \bar{\pi}\tilde{u}_i - \tau_i \bar{\pi}\bar{P}K\bar{P}\bar{\pi} M_i \tilde{u}_i - \tau_i \bar{\pi}\bar{P}K\bar{P}\pi s_i = (I - \tau_i T_i)\bar{\pi}\tilde{u}_i - v_i \end{aligned}$$

where $v_i = \tau_i \bar{\pi}\bar{P}K\bar{P}\pi s_i = \tau_i \bar{\pi}\bar{P}K\bar{P}\pi\bar{P} s_i$. Using (41), (28) and (24) we obtain

$$\|v_i\|_{M_i} \le \sqrt{\tilde{\mu}_i}\|v_i\| \le \tau_i b \sqrt{\tilde{\mu}_i}\|\bar{P}\pi\bar{P}s_i\| \le \frac{2}{\sqrt{\tilde{\mu}_i}}\theta^2 \|s_i\|$$

which leads to (55). □

Lemma 27 *Let $\tilde{\mu}_{l-1} > \mu_l$, $\tilde{\mu}_{k-1} > \mu_k$ and $\tilde{\mu}_{m-1} > \mu_m$. In the notation and under the assumptions of lemma 26 the following inequality is valid for $l \le i < k$*

$$((\tilde{\mu}_i I - M)\bar{\pi}_0 \hat{u}_i, \bar{\pi}_0 \hat{u}_i) \le q_i^2 ((\tilde{\mu}_i I - M)\bar{\pi}_0 \tilde{u}_i, \bar{\pi}_0 \tilde{u}_i) + c_0 \frac{\theta^2 \|s_i\|^2}{\tilde{\mu}_i} \tag{56}$$

where $c_0 = 4\sqrt{\eta_0} + \eta_- + \eta_+ + 8\chi_+ + 4\theta^2$.

Proof. Using lemma 11, we have

$$\|\bar{\pi}\tilde{u}_i\|_{M_i}^2 = ((\tilde{\mu}_i I - M)\bar{\pi}\tilde{u}_i, \bar{\pi}\tilde{u}_i) \le \frac{\|s_i\|^2}{\delta_\mu(\tilde{u}_i, \bar{\mathcal{I}})} \le \eta_0 \frac{\|s_i\|^2}{\tilde{\mu}_i}$$

and, using lemma 15, we have

$$((\tilde{\mu}_i I - M)\pi_- \tilde{u}_i, \pi_- \tilde{u}_i) \ge -\frac{\chi_- \theta^2 \|s_i\|^2}{\delta_\mu(\tilde{u}_i, \mathcal{I}_-)} \ge -\eta_- \frac{\theta^2 \|s_i\|^2}{\tilde{\mu}_i} \tag{57}$$

and, hence,

$$\begin{aligned} \|\bar{\pi}\tilde{u}_i\|_{M_i}^2 = ((\tilde{\mu}_i I - M)\bar{\pi}\tilde{u}_i, \bar{\pi}\tilde{u}_i) &= ((\tilde{\mu}_i I - M)\bar{\pi}_0 \tilde{u}_i, \bar{\pi}_0 \tilde{u}_i) \\ -((\tilde{\mu}_i I - M)\pi_- \tilde{u}_i, \pi_- \tilde{u}_i) &- ((\tilde{\mu}_i I - M)\pi_+ \tilde{u}_i, \pi_+ \tilde{u}_i) \\ &\le ((\tilde{\mu}_i I - M)\bar{\pi}_0 \tilde{u}_i, \bar{\pi}_0 \tilde{u}_i) + \eta_- \frac{\theta^2 \|s_i\|^2}{\tilde{\mu}_i} \end{aligned}$$

Therefore, using lemma 26, we obtain

$$\begin{aligned} \|\bar{\pi}\hat{u}_i\|_{M_i}^2 &\le \left(q_i \|\bar{\pi}\tilde{u}_i\|_{M_i} + 2\frac{\theta^2 \|s_i\|}{\sqrt{\tilde{\mu}_i}} \right)^2 \\ = q_i^2 \|\bar{\pi}\tilde{u}_i\|_{M_i}^2 &+ 4 q_i \|\bar{\pi}\tilde{u}_i\|_{M_i} \frac{\theta^2 \|s_i\|}{\sqrt{\tilde{\mu}_i}} + 4\frac{\theta^4 \|s_i\|^2}{\tilde{\mu}_i} \\ &\le q_i^2 ((\tilde{\mu}_i I - M)\bar{\pi}_0 \tilde{u}_i, \bar{\pi}_0 \tilde{u}_i) + \eta_- \frac{\theta^2 \|s_i\|^2}{\tilde{\mu}_i} \\ &\qquad + 4\sqrt{\eta_0} \frac{\theta^2 \|s_i\|^2}{\tilde{\mu}_i} + 4\frac{\theta^4 \|s_i\|^2}{\tilde{\mu}_i} \end{aligned} \tag{58}$$

Further,

$$\begin{aligned}\|\bar\pi\hat u_i\|_{M_i}^2 = ((\tilde\mu_i I - M)\bar\pi\hat u_i, \bar\pi\hat u_i) &= ((\tilde\mu_i I - M)\bar\pi_0\hat u_i, \bar\pi_0\hat u_i)\\ -((\tilde\mu_i I - M)\pi_-\hat u_i, \pi_-\hat u_i) &- ((\tilde\mu_i I - M)\pi_+\hat u_i, \pi_+\hat u_i)\\ \geq ((\tilde\mu_i I - M)\bar\pi_0\hat u_i, \bar\pi_0\hat u_i) &- ((\tilde\mu_i I - M)\pi_+\hat u_i, \pi_+\hat u_i)\end{aligned}$$

that is,

$$((\tilde\mu_i I - M)\bar\pi_0\hat u_i, \bar\pi_0\hat u_i) \leq \|\bar\pi\hat u_i\|_{M_i}^2 + ((\tilde\mu_i I - M)\pi_+\hat u_i, \pi_+\hat u_i)$$

Since $\pi_+\hat u_i = \pi_+\tilde u_i - \tau_i\pi_+\bar P K s_i \equiv \pi_+\tilde u_i - w_i$, we have:

$$\begin{aligned}((\tilde\mu_i I - M)\pi_+\hat u_i, \pi_+\hat u_i) &\leq ((\tilde\mu_i I - M)\pi_+\tilde u_i, \pi_+\tilde u_i)\\ &+2|((\tilde\mu_i I - M)\pi_+\tilde u_i, w_i)| + ((\tilde\mu_i I - M)w_i, w_i)\end{aligned}$$

From lemma 15 it follows that (cf. (57))

$$((\tilde\mu_i I - M)\pi_+\tilde u_i, \pi_+\tilde u_i) \leq \eta_+ \frac{\theta^2\|s_i\|^2}{\tilde\mu_i}$$

Further,

$$\begin{aligned}|((\tilde\mu_i I - M)\pi_+\tilde u_i, w_i)| &= \tau_i|((\tilde\mu_i I - M)\pi_+\tilde u_i, \pi_+\bar P K s_i)|\\ &= \tau_i|(\pi_+ r_i, \pi_+\bar P K s_i)| = \tau_i|(\pi_+\bar P s_i, \pi_+\bar P K s_i)|\\ &\leq \tau_i\|\pi_+\bar P s_i\|\|\pi_+\bar P K s_i\| \leq \tau_i\|\pi\bar P s_i\|\|\pi\bar P K s_i\|\\ &\leq \tau_i\theta^2\|s_i\|\|K s_i\|\end{aligned}$$

and, using (28), we obtain

$$|((\tilde\mu_i I - M)\pi_+\tilde u_i, w_i)| \leq \tau_i b\theta^2\|s_i\|^2 \leq 2\frac{\theta^2\|s_i\|^2}{\tilde\mu_i}$$

Finally,

$$\begin{aligned}((\tilde\mu_i I - M)w_i, w_i) &= \tau_i^2((\tilde\mu_i I - M)\pi_+\bar P K s_i, \pi_+\bar P K s_i)\\ \leq \tau_i^2\tilde\mu_i\|\pi_+\bar P K s_i\|^2 &\leq \tau_i^2\tilde\mu_i\|\pi\bar P K s_i\|^2 \leq \tau_i^2\tilde\mu_i\theta^2\|K s_i\|^2\\ &\leq b^2\tau_i^2\tilde\mu_i\theta^2\|s_i\|^2 \leq 4\frac{\theta^2\|s_i\|^2}{\tilde\mu_i}\end{aligned}$$

that is

$$((\tilde\mu_i I - M)\bar\pi_0\hat u_i, \bar\pi_0\hat u_i) \leq \|\bar\pi\tilde u_i\|_{M_i}^2 + \chi_+(\eta_+ + 8)\frac{\theta^2\|s_i\|^2}{\tilde\mu_i}$$

which, together with (58) leads to (56). □

Lemma 28 *In the notation and under the assumptions of lemma 27 the following inequality is valid*

$$(\mu(\pi_0\tilde u_i) - \mu(\pi_0\hat u_i))\|\pi_0\hat u_i\|^2 \leq c_1\frac{\theta^2\|s_i\|^2}{\tilde\mu_i} \tag{59}$$

where $c_1 = 4(2 + (1 + \eta_\pm t + (1 + \eta_+ t^2)\theta^2)(1 + t_0^2))$.

Proof. We have

$$(\mu(\pi_0\tilde{u}_i) - \mu(\pi_0\hat{u}_i))\|\pi_0\hat{u}_i\|^2 = \mu(\pi_0\tilde{u}_i)\|\pi_0\hat{u}_i\|^2 - \|\pi_0\hat{u}_i\|_M^2$$

$$\pi_0\hat{u}_i = \pi_0\tilde{u}_i - \tau_i\pi_0\bar{P}Ks_i$$

$$\|\pi_0\hat{u}_i\|_M^2 = \|\pi_0\tilde{u}_i\|_M^2 - 2\tau_i(\tilde{u}_i, \pi_0\bar{P}Ks_i)_M + \tau_i^2\|\pi_0\bar{P}Ks_i\|_M^2$$

$$\|\pi_0\hat{u}_i\|^2 = \|\pi_0\tilde{u}_i\|^2 - 2\tau_i(\tilde{u}_i, \pi_0\bar{P}Ks_i) + \tau_i^2\|\pi_0\bar{P}Ks_i\|^2$$

Hence, $(\mu(\pi_0\tilde{u}_i) - \mu(\pi_0\hat{u}_i))\|\pi_0\hat{u}_i\|^2 = -2\tau_i\alpha + \tau_i^2\beta$, where

$$\alpha = ((\mu(\pi_0\tilde{u}_i)I - M)\tilde{u}_i, \pi_0\bar{P}Ks_i) = \alpha_1 + \alpha_2$$

$$\alpha_1 = ((\tilde{\mu}_iI - M)\tilde{u}_i, \pi_0\bar{P}Ks_i), \quad \alpha_2 = (\mu(\pi_0\tilde{u}_i) - \tilde{\mu}_i)(\tilde{u}_i, \pi_0\bar{P}Ks_i)$$

$$\beta = \mu(\pi_0\tilde{u}_i)\|\pi_0\bar{P}Ks_i\|^2 - \|\pi_0\bar{P}Ks_i\|_M^2$$

For α_1 we have

$$\begin{aligned}|\alpha_1| = |(s_i, \pi_0\bar{P}Ks_i)| = |(\pi_0\bar{P}s_i, \pi_0\bar{P}Ks_i)| &\le \|\pi_0\bar{P}s_i\|\|\pi_0\bar{P}Ks_i\| \\ &\le \|\pi\bar{P}s_i\|\|\pi\bar{P}Ks_i\| \le \theta^2\|s_i\|\|Ks_i\| \le b\theta^2\|s_i\|^2\end{aligned}$$

The second factor in α_2 can be estimated as follows

$$|(\tilde{u}_i, \pi_0\bar{P}Ks_i)| \le \frac{\|\pi_0\bar{P}Ks_i\|}{\sqrt{\tilde{\mu}_i}} \le \frac{\|\pi\bar{P}Ks_i\|}{\sqrt{\tilde{\mu}_i}} \le \frac{\theta\|Ks_i\|}{\sqrt{\tilde{\mu}_i}} \le b\theta\frac{\|s_i\|}{\sqrt{\tilde{\mu}_i}}$$

and for the first factor we can apply (47) to obtain

$$|\alpha_2| \le b(1 + \eta_\pm t)(1 + t_0^2)\theta^2\|s_i\|^2$$

Finally, β is easily estimated, with the help of (48), as

$$\begin{aligned}\beta &\le \mu(\pi_0\tilde{u}_i)\|\pi_0\bar{P}Ks_i\|^2 \le \mu(\pi_0\tilde{u}_i)\|\pi\bar{P}Ks_i\|^2 \le b^2\mu(\pi_0\tilde{u}_i)\theta^2\|s_i\|^2 \\ &= b^2(\tilde{\mu}_i + \mu(\pi_0\tilde{u}_i) - \tilde{\mu}_i)\theta^2\|s_i\|^2 \le b^2(1 + (1 + \eta_+t^2)(1 + t_0^2)\theta^2)\tilde{\mu}_i\theta^2\|s_i\|^2\end{aligned}$$

which, together with the above estimates for α_1 and α_2, leads to (59). □

Lemma 29 *In the notation and under the assumptions of lemma 27 the following inequality is valid*

$$(\mu(\pi_0\tilde{u}_i) - \tilde{\mu}_i)\|\bar{\pi}_0\hat{u}_i\|^2 \le (c_2t_0^2 + c_3\theta^2)\frac{\|s_i\|^2}{\tilde{\mu}_i} \qquad (60)$$

where $c_2 = 2(\eta_0 + \eta_+\theta^2)$ *and* $c_3 = 8(1 + \eta_+t^2)(1 + t_0^2)(1 + \theta^2)$.

Proof. We have

$$\|\bar{\pi}_0 \hat{u}_i\|^2 \le 2\|\bar{\pi}_0 \tilde{u}_i\|^2 + 2\tau_i^2 \|\bar{\pi}_0 \bar{P} K s_i\|^2$$

and

$$\begin{aligned} \|\bar{\pi}_0 \bar{P} K s_i\|^2 = \|\bar{\pi} \bar{P} K s_i\|^2 + \|\bar{\pi}_\pm \bar{P} K s_i\|^2 \le \|K s_i\|^2 + \|\pi \bar{P} K s_i\|^2 \\ \le (1+\theta^2) b^2 \|s_i\|^2 \end{aligned}$$

Thus,

$$\begin{aligned} (\mu(\pi_0 \tilde{u}_i) - \tilde{\mu}_i)\|\bar{\pi}_0 \hat{u}_i\|^2 &\le 2(\mu(\pi_0 \tilde{u}_i) - \tilde{\mu}_i)(\|\bar{\pi}_0 \tilde{u}_i\|^2 + \tau_i^2 b^2 (1+\theta^2)\|s_i\|^2) \\ &= 2(\mu(\pi_0 \tilde{u}_i) - \tilde{\mu}_i)(t_0^2 \|\pi_0 \tilde{u}_i\|^2 + \tau_i^2 b^2 (1+\theta^2)\|s_i\|^2) \end{aligned}$$

and, using (48) and (50), we arrive at (60). □

Lemma 30 *Let* $\hat{\mathcal{I}} = \operatorname{span}\{\hat{u}_i\}_{i=p,q}$*, where* $\hat{u}_i$ *are given in lemma 26 and* $0 \le p < q < m$*, and denote* $\hat{\mu}_i = \mu_{i-p}(\hat{\mathcal{I}})$*. If* K *satisfies* (28) *then*

$$\sum_{i=p}^{q} (\mu(\hat{u}_i) - \hat{\mu}_i) \le 16 \frac{\mu_p + \mu_m + \mu_0 \theta^2}{\tilde{\mu}_q} \rho_{\hat{\mathcal{I}}} \sum_{i=p}^{q} \|s_i\|^2 \tag{61}$$

Proof. Let A and B be $(q-p+1)$-by-$(q-p+1)$ matrices with the entries $a_{ij} = (v_i, v_j)_M$ and $b_{ij} = (v_i, v_j)$, $i, j = p, \ldots, q$, resp., where $v_i = \sqrt{\tilde{\mu}_i}\hat{u}_i$. Using the fact that $\bar{P} M \tilde{u}_i = \bar{P}(M - \tilde{\mu}_i I)\tilde{u}_i = -\bar{P} s_i = -s_i$ we have

$$\begin{aligned} (\hat{u}_i, \hat{v}_j) &= \delta_{ij} - \tau_i (\bar{P} K s_i, M \tilde{u}_j) - \tau_j (\bar{P} K s_j, M \tilde{u}_i) + \tau_i \tau_j (\bar{P} K s_i, \bar{P} K s_j)_M \\ &= \delta_{ij} + \tau_i (K s_i, s_j) + \tau_j (K s_j, s_i) + \tau_i \tau_j (\bar{P} K s_i, \bar{P} K s_j)_M \end{aligned}$$

Hence, $a_{ij} = \tilde{\mu}_i \delta_{ij} + a_{ij}^1 + a_{ij}^2 + a_{ij}^3$, where

$$a_{ij}^1 = a_{ji}^2 = \tau_i \sqrt{\tilde{\mu}_i \tilde{\mu}_j} (K s_i, s_j), \quad a_{ij}^3 = \tau_i \tau_j \sqrt{\tilde{\mu}_i \tilde{\mu}_j} (\bar{P} K s_i, \bar{P} K s_j)_M$$

Using (28) and the fact that $b \tau_i \tilde{\mu}_i < 2$ we easily obtain the following estimate

$$|a_{ij}^1| \le \tau_i \sqrt{\tilde{\mu}_i \tilde{\mu}_j} \|K s_i\| \|s_j\| \le 2 \sqrt{\frac{\tilde{\mu}_j}{\tilde{\mu}_i}} \|s_i\| \|s_j\|$$

Further,

$$\begin{aligned} |(\bar{P} K s_i, \bar{P} K s_j)_M| &\le |(\pi \bar{P} K s_i, \pi \bar{P} K s_j)_M| + |(\bar{\pi} \bar{P} K s_i, \bar{\pi} \bar{P} K s_j)_M| \\ &\le \|\pi \bar{P} K s_i\|_M \|\pi \bar{P} K s_j\|_M + \|\bar{\pi} \bar{P} K s_i\|_M \|\bar{\pi} \bar{P} K s_j\|_M \\ &\le \mu_0 \|\pi \bar{P} K s_i\| \|\pi \bar{P} K s_j\| + \mu_m \|\bar{\pi} \bar{P} K s_i\| \|\bar{\pi} \bar{P} K s_j\| \\ &\le \mu_0 \theta^2 \|K s_i\| \|K s_j\| + \mu_m \|K s_i\| \|K s_j\| \le (\mu_0 \theta^2 + \mu_m) b^2 \|s_i\| \|s_j\| \end{aligned}$$

and, hence, $|a_{ij}^3| \le 4(\mu_0 \theta^2 + \mu_m)(\tilde{\mu}_i \tilde{\mu}_j)^{-1} \|s_i\| \|s_j\|$. For b_{ij} we have

$$b_{ij} = \delta_{ij} + \tau_i \tau_j \sqrt{\tilde{\mu}_i \tilde{\mu}_j} (\bar{P} K s_i, \bar{P} K s_j) = \delta_{ij} + b_{ij}^1$$

Let A_n and δB be $(q-p+1)$-by-$(q-p+1)$ matrices with the entries a_{ij}^n, $n=1,2,3$, and b_{ij}^1 resp. and let $A_0 = \mathrm{Diag}\{\tilde{\mu}_i\}_{i=p,q}$. For $\delta A = A_1 + A_2 + A_3 = A_1 + A_1^T + A_3$ we easily obtain the following estimate

$$\begin{aligned}
\|\delta A\| \le \|\delta A\|_F \le \|\delta A_1\|_F + \|\delta A_2\|_F + \|\delta A_3\|_F = 2\|\delta A_1\|_F + \|\delta A_3\|_F \\
\le 4\left(\sum_{i=p}^{q} \tilde{\mu}_i \|s_i\|^2\right)^{\frac{1}{2}} \left(\sum_{i=p}^{q} \frac{\|s_i\|^2}{\tilde{\mu}_i}\right)^{\frac{1}{2}} + 4(\mu_m + \mu_0\theta^2) \sum_{i=p}^{q} \frac{\|s_i\|^2}{\tilde{\mu}_i} \\
\le 4(\mu_p + \mu_m + \mu_0\theta^2) \sum_{i=p}^{q} \frac{\|s_i\|^2}{\tilde{\mu}_i}
\end{aligned}$$

and for δB we have: $\delta B \ge 0$ and

$$\mathrm{Tr}(\delta B) = \sum_{i=p}^{q} b_{ii}^1 = \sum_{i=p}^{q} \tilde{\mu}_i \tau_i^2 \|\bar{P}Ks_i\|^2 \le 4\sum_{i=p}^{q} \frac{\|s_i\|^2}{\tilde{\mu}_i} \le \frac{4}{\tilde{\mu}_q} \sum_{i=p}^{q} \|s_i\|^2$$

It remains to apply lemma 7. □

Lemma 31 *Let $\tilde{\mathcal{I}}'' = \mathrm{span}\{\hat{u}_i\}_{i=0,k-1}$, where $\hat{u}_i$ are given in lemma 26, and denote $\tilde{\mu}_i'' = \mu_i(\tilde{\mathcal{I}}'')$. If K satisfies* (28) *and $\tilde{\mu}_{k-1} > \mu_k$ and either $l = 0$ or*

$$\sum_{i=0}^{l-1} (\mu_i - \tilde{\mu}_i) \le \frac{\zeta}{4}\left(2 + \frac{\mu_m + \mu_0\theta^2}{\mu_l}\right)^{-1} \frac{\tilde{\mu}_{l-1} - \mu_l}{1 + \rho_-} \tag{62}$$

where $\zeta < 1$, then

$$\sum_{i=l}^{k-1} (\mu(\hat{u}_i) - \tilde{\mu}_i'') \le c_4 \rho_* \sum_{i=l}^{k-1} \|s_i\|^2 \tag{63}$$

where

$$c_4 = 4\frac{\mu_l}{\mu_k}\left(8 + 4\frac{\mu_0}{\mu_l}\theta^2 + \left(\frac{(k-l)\tilde{\eta}_-}{1-\zeta}\left(1 + 3\frac{\mu_0}{\mu_l}\right)\right)^2\right)$$

Proof. For $l = 0$ we have $\tilde{\eta}_- = 0$ and (63) follows directly from (61). Thus, it remains to consider the case $l > 0$.

Let A and B be k-by-k matrices with the entries $a_{ij} = (\hat{v}_i, \hat{v}_j)_M$ and $b_{ij} = (\hat{v}_i, \hat{v}_j)$ resp., $i, j = 0, \ldots, k-1$, where $\hat{v}_i = \sqrt{\tilde{\mu}_i} u_i$, and let A and B be split into blocks A_{ij} and B_{ij}, $i, j = 0, 1$, where the size of A_{00} and B_{00} is l-by-l. Denote by $\hat{\mu}_i$, $i = 0, \ldots, l-1$, the eigenvalues of $B_{00}^{-1}A_{00}$ and by $\hat{\mu}_i$, $i = l, \ldots, k-1$, the eigenvalues of $B_{11}^{-1}A_{11}$ enumerated in descending order. We have: $a_{ij} = \tilde{\mu}_{ij}\delta_{ij} + a_{ij}^1 + a_{ij}^2 + a_{ij}^3$ and $b_{ij} = \delta_{ij} + b_{ij}^1$ where a_{ij}^n and b_{ij}^1 are defined in the proof of lemma 30. Hence, $A_{00} = D_{00} + \delta A_{00}$, where $D_{00} = \mathrm{Diag}\{\tilde{\mu}_i\}_{i=0,l-1}$, and $B_{00} = I + \delta B_{00}$. Using the estimates for a_{ij}^n in the proof of lemma 30 and the fact that $A_{00} \ge \tilde{\mu}_{l-1}$, we easily obtain

$$-\frac{\alpha_{00}}{\tilde{\mu}_{l-1}} A_{00} \le \delta A_{00} \le \frac{\alpha_{00}}{\tilde{\mu}_{l-1}} A_{00}$$

where

$$\alpha_{00} = 4(\mu_l + \mu_m + \mu_0\theta^2)\rho_- \le 4\left(1 + \frac{\mu_m + \mu_0\theta^2}{\mu_l}\right) \sum_{i=0}^{l-1} \|s_i\|^2$$

Further,

$$|b_{ij}^1| \le \tau_i \tau_j b^2 \sqrt{\tilde{\mu}_i \tilde{\mu}_j} \|s_i\| \|s_j\| \le 4 \frac{\|s_i\| \|s_j\|}{\sqrt{\tilde{\mu}_i \tilde{\mu}_j}}$$

and, hence, $0 \le \delta B_{00} \le \beta_{00} \le \beta_{00} B_{00}$, where

$$\beta_{00} = 4 \sum_{i=0}^{l-1} \frac{\|s_i\|^2}{\tilde{\mu}_i} \le \frac{4}{\tilde{\mu}_{l-1}} \sum_{i=0}^{l-1} \|s_i\|^2$$

Using lemma 4, we obtain

$$\begin{aligned} \hat{\mu}_{l-1} &\ge \tilde{\mu}_{l-1} \left(1 - \frac{4}{\tilde{\mu}_{l-1}} \left(2 + \frac{\mu_m + \mu_0 \theta^2}{\mu_l} \right) \sum_{i=0}^{l-1} \|s_i\|^2 \right) \\ &= \tilde{\mu}_{l-1} - 4 \left(2 + \frac{\mu_m + \mu_0 \theta^2}{\mu_l} \right) \sum_{i=0}^{l-1} \|s_i\|^2 \end{aligned}$$

and, using lemma 23 and (62),

$$\begin{aligned} \hat{\mu}_{l-1} \ge \tilde{\mu}_{l-1} - 4 \left(2 + \frac{\mu_m + \mu_0 \theta^2}{\mu_l} \right) (1 + \rho_-) \sum_{i=0}^{l-1} (\mu_i - \tilde{\mu}_i) \\ \ge \tilde{\mu}_{l-1} - \zeta (\tilde{\mu}_{l-1} - \mu_l) \end{aligned}$$

that is, $\hat{\mu}_{l-1} - \mu_l \ge (1 - \zeta)(\tilde{\mu}_{l-1} - \mu_l)$. Now, let $\mu = \tilde{\mu}''_n$, where $n \ge l$, and denote $r_{ij} = a_{ij} - \mu b_{ij}$. We have

$$\begin{aligned} r_{ij} = ((M - \mu I) \hat{v}_i, \hat{v}_j) = \sqrt{\tilde{\mu}_i \tilde{\mu}_j} ((M - \mu I) \hat{u}_i, \hat{u}_j) \\ = \sqrt{\tilde{\mu}_i \tilde{\mu}_j} (-\tau_i (\bar{P} K s_i, (M - \mu I) \tilde{u}_j) - \tau_j ((M - \mu I) \tilde{u}_i, \bar{P} K s_j) \\ + \tau_i \tau_j ((M - \mu I) \bar{P} K s_i, \bar{P} K s_j)) \end{aligned}$$

and, since $\bar{P} \tilde{u} = 0$ for any $\tilde{u} \in \tilde{\mathcal{I}}$,

$$\begin{aligned} r_{ij} = \sqrt{\tilde{\mu}_i \tilde{\mu}_j} (\tau_i (K s_i, s_j) + \tau_j (s_i, K s_j) + \tau_i \tau_j ((M - \mu L) \bar{P} K s_i, \bar{P} K s_j)) \\ = a_{ij}^1 + a_{ij}^2 + r_{ij}^3 \end{aligned}$$

We have

$$|a_{ij}^1| = |a_{ji}^2| \le \sqrt{\tilde{\mu}_i \tilde{\mu}_j} \tau_i \|K s_i\| \|s_j\| \le \sqrt{\tilde{\mu}_i \tilde{\mu}_j} \tau_i b \|s_i\| \|s_j\| \le 2 \sqrt{\frac{\tilde{\mu}_j}{\tilde{\mu}_i}} \|s_i\| \|s_j\|$$

Since $-\mu_0 I \le M - \mu I \le \mu_0 I$, applying lemma 3, we obtain

$$\begin{aligned} |((M - \mu I) \bar{P} K s_i, \bar{P} K s_j)| \le \mu_0 \|\bar{P} K s_i\| \|\bar{P} K s_j\| \\ \le \mu_0 \|K s_i\| \|P K s_j\| \le \mu_0 \|K s_i\| \|K s_j\| \le \mu_0 b^2 \|s_i\| \|s_j\| \end{aligned}$$

and, thus,

$$|r_{ij}^3| = \tau_i \tau_j \sqrt{\tilde{\mu}_i \tilde{\mu}_j} |((M - \mu I) \bar{P} K s_i, \bar{P} K s_j)| \le 4 \mu_0 \frac{\|s_i\| \|s_j\|}{\sqrt{\tilde{\mu}_i \tilde{\mu}_j}}$$

For $l \le i < k$ and $0 \le j < l$ we have

$$|r_{ij}| \le 2(\tilde{\mu}_i + \tilde{\mu}_j + 2\mu_0)\frac{\|s_i\|\|s_j\|}{\sqrt{\tilde{\mu}_i\tilde{\mu}_j}} \le \frac{2}{\sqrt{\tilde{\mu}_{k-1}}}(\mu_l + 3\mu_0)\|s_i\|\frac{\|s_j\|}{\sqrt{\tilde{\mu}_j}}$$

that is

$$\|A_{10} - \mu R_{10}\|_F^2 \le \frac{4\mu_l^2}{\tilde{\mu}_{k-1}}\left(1 + 3\frac{\mu_0}{\mu_l}\right)^2 \rho_- \sum_{i=l}^{k-1}\|s_i\|^2$$

By the minimax principle $\tilde{\mu}_i'' \le \mu_i$ and, hence, $\hat{\mu}_{l-1} - \tilde{\mu}_l'' \ge \hat{\mu}_{l-1} - \mu_l \ge (1-\zeta)(\tilde{\mu}_{l-1} - \mu_l)$. Applying lemma 8, we obtain

$$\begin{aligned}\hat{\mu}_i - \tilde{\mu}_i'' &\le \frac{\tilde{\mu}_i''(k-l)}{(1-\zeta)^2(\tilde{\mu}_{l-1} - \mu_l)^2}\frac{4\mu_l^2}{\tilde{\mu}_{k-1}}\left(1 + 3\frac{\mu_0}{\mu_l}\right)^2 \rho_- \sum_{i=l}^{k-1}\|s_i\|^2\\ &= 4(k-l)\frac{\tilde{\mu}_i''}{\tilde{\mu}_{k-1}}\frac{\tilde{\eta}_-^2}{(1-\zeta)^2}\left(1 + 3\frac{\mu_0}{\mu_l}\right)^2 \rho_- \sum_{i=l}^{k-1}\|s_i\|^2\end{aligned}$$

and, thus,

$$\sum_{i=l}^{k-1}(\hat{\mu}_i - \tilde{\mu}_i'') \le 4(k-l)^2\frac{\mu_l}{\mu_k}\frac{\tilde{\eta}_-^2}{(1-\zeta)^2}\left(1 + 3\frac{\mu_0}{\mu_l}\right)^2 \rho_- \sum_{i=l}^{k-1}\|s_i\|^2$$

Finally, using lemma 30, we obtain

$$\sum_{i=l}^{k-1}(\mu(\hat{u}_i) - \hat{\mu}_i) \le 16\frac{2\mu_l + \mu_0\theta^2}{\mu_k}\rho_0 \sum_{i=l}^{k-1}\|s_i\|^2$$

and we arrive at (63). □

Lemma 32 *In the notation and under the assumptions of lemma 31*

$$\sum_{i=l}^{k-1}(\mu_i - \tilde{\mu}_i'') \le q_{k-1}^2 \sum_{i=l}^{k-1}(\mu_i - \tilde{\mu}_i) + (c_4\rho_* + c_5\theta^2 + c_6 t_0^2)\sum_{i=l}^{k-1}\|s_i\|^2 \qquad (64)$$

where $c_5 = c_0 + c_1 + c_3$, $c_6 = c_2 + (1 - q_{k-1}^2)(\alpha_0 + \beta_0)$, α_0 and β_0 are given in lemma 20, and $c_0, \dots, c_4$ are given in lemmas 27, 28, 29 and 31.

Proof. It is easy to verify that

$$\begin{aligned}((\tilde{\mu}_i I - M)\bar{\pi}_0\hat{u}_i, \bar{\pi}_0\hat{u}_i) &= (\mu(\pi_0\tilde{u}_i) - \mu(\hat{u}_i))\|\hat{u}_i\|^2\\ &+ (\mu(\pi_0\hat{u}_i) - \mu(\pi_0\tilde{u}_i)\|\pi_0\hat{u}_i\|^2 + (\tilde{\mu}_i - \mu(\pi_0\tilde{u}_i)\|\bar{\pi}_0\hat{u}_i\|^2\\ &\ge (\mu(\pi_0\tilde{u}_i) - \mu(\hat{u}_i))\|\tilde{u}_i\|^2 + (\mu(\pi_0\hat{u}_i) - \mu(\pi_0\tilde{u}_i)\|\pi_0\hat{u}_i\|^2\\ &+ (\tilde{\mu}_i - \mu(\pi_0\tilde{u}_i)\|\bar{\pi}_0\hat{u}_i\|^2\end{aligned}$$

Thus,

$$\begin{aligned}\mu(\pi_0\tilde{u}_i) - \mu(\hat{u}_i) &\le \tilde{\mu}_i((\tilde{\mu}_i I - M)\bar{\pi}_0\hat{u}_i, \bar{\pi}_0\hat{u}_i)\\ + \tilde{\mu}_i(\mu(\pi_0\tilde{u}_i) - \mu(\pi_0\hat{u}_i))\|\pi_0\hat{u}_i\|^2 &+ \tilde{\mu}_i(\mu(\pi_0\tilde{u}_i) - \tilde{\mu}_i)\|\bar{\pi}_0\hat{u}_i\|^2\end{aligned}$$

and, using lemmas 27, 28 and 29, we obtain

$$\begin{aligned}\mu(\pi_0\tilde{u}_i) - \mu(\hat{u}_i) \le q_i^2 \tilde{\mu}_i((\tilde{\mu}_i I - M)\bar{\pi}_0\tilde{u}_i, \bar{\pi}_0\tilde{u}_i) + c_0\theta^2\|s_i\|^2 \\ + c_1\theta_L^2\|s_i\|^2 + (c_2 t_0^2 + c_3\theta^2)\|s_i\|^2 \\ = q_i^2(\mu(\pi_0\tilde{u}_i) - \tilde{\mu}_i) + (c_5\theta^2 + c_2 t_0^2)\|s_i\|^2\end{aligned}$$

Using the above inequality together with lemmas 31 and 20, we finally obtain

$$\begin{aligned}\sum_{i=l}^{k-1}(\mu_i - \tilde{\mu}_i'') &= \sum_{i=l}^{k-1}(\mu_i - \mu(\pi_0\tilde{u}_i)) + \sum_{i=l}^{k-1}(\mu(\pi_0\tilde{u}_i) - \mu(\hat{u}_i)) + \sum_{i=l}^{k-1}(\mu(\hat{u}_i) - \tilde{\mu}_i'') \\ &\le \sum_{i=l}^{k-1}(\mu_i - \mu(\pi_0\tilde{u}_i)) + q_{k-1}^2\left(\sum_{i=l}^{k-1}(\mu_i - \tilde{\mu}_i) + \sum_{i=l}^{k-1}(\mu(\pi_0\tilde{u}_i) - \mu_i)\right) \\ &\qquad + (c_5\theta^2 + c_2 t_0^2)\|s_i\|^2 + c_4\rho_*\|s_i\|^2 \\ &\le q_{k-1}^2\sum_{i=l}^{k-1}(\mu_i - \tilde{\mu}_i) + (((1 - q_{k-1}^2)(\alpha_0 + \beta_0) + c_2)t_0^2 + c_5\theta^2 + c_4\rho_*)\|s_i\|^2\end{aligned}$$

□

Lemma 33 *Assuming that K satisfies (28) and $\tilde{\mu}_{k-1} > \mu_k$ and either $l = 0$ or the conditions (53) and (62) are satisfied, for any $\mathcal{X} \supset \tilde{\mathcal{I}}_* + \operatorname{span}\{Ks_i\}_{i=0,k-1}$*

$$\sum_{i=l}^{k-1}(\mu_i - \mu_i(\mathcal{X})) \le \gamma \sum_{i=l}^{k-1}(\mu_i - \tilde{\mu}_i) \tag{65}$$

where

$$\gamma = \frac{q_{k-1}^2 + \epsilon}{1 + \epsilon}, \quad \epsilon = \frac{b}{a}\frac{1 + \rho_*}{1 - \chi_- \xi}(c_4\rho_* + c_5\theta^2 + c_6 t_0^2) \tag{66}$$

and ξ is given in lemma 24, c_4 in lemma 31 and c_5 and c_6 in lemma 32.

Proof. Denote $\mu_i^{\mathcal{X}} \equiv \mu_i(\mathcal{X})$. Since $\mathcal{X} \supset \tilde{\mathcal{I}}'$ and $\mathcal{X} \supset \tilde{\mathcal{I}}''$, where $\tilde{\mathcal{I}}'$ and $\tilde{\mathcal{I}}''$ are defined in lemmas 23 and 31 resp., by the minimax principle $\mu_i^{\mathcal{X}} \ge \tilde{\mu}_i'$ and $\mu_i^{\mathcal{X}} \ge \tilde{\mu}_i''$. Hence, using lemma 32 and either lemma 23 (for $l = 0$) or lemma 24 (for $l > 0$), we obtain

$$\begin{aligned}\sum_{i=l}^{k-1}(\mu_i - \mu_i^{\mathcal{X}}) &\le q_{k-1}^2\sum_{i=l}^{k-1}(\mu_i - \tilde{\mu}_i) + (c_4\rho_* + c_5\theta^2 + c_6 t_0^2)\sum_{i=l}^{k-1}\|s_i\|^2 \\ &\le q_{k-1}^2\sum_{i=l}^{k-1}(\mu_i - \tilde{\mu}_i) + \epsilon\sum_{i=l}^{k-1}(\mu_i^{\mathcal{X}} - \tilde{\mu}_i) \\ &= (q_{k-1}^2 + \epsilon)\sum_{i=l}^{k-1}(\mu_i - \tilde{\mu}_i) - \epsilon\sum_{i=l}^{k-1}(\mu_i - \mu_i^{\mathcal{X}})\end{aligned}$$

which leads to (65). □

References

[1] Z. Bai *et al.* (eds). *Templates for the Solution of Algebraic Eigenvalue Problems. A Practical Guide.* SIAM, 2000.

[2] L. Bergamaschi, G. Gambolati and G. Pini. Asymptotic convergence of conjugate gradient methods for the partial symmetric eigenproblems. *Numer. Linear Algebra Appl.*, **4**:69-84, 1997.

[3] J. H. Bramble, J. E. Pasciak and A. V. Knyazev. A subspace preconditioning algorithm for eigenvector/eigenvalue computation. *Adv. Comput. Math.*, **6**: 159-189, 1996.

[4] V. E. Bulgakov, M. V. Belyi and K. M. Mathisen. Multilevel aggregation method for solving large-scale generalized eigenvalue problems in structural dynamics. *Internat. J. Numer. Methods Engrg.*, **40**:453-471, 1997.

[5] F. Chatelin. *Eigenvalues of Matrices*. Wiley, 1993.

[6] E. G. D'yakonov. *Optimization in Solving Elliptic Problems*. CRC Press, 1996.

[7] Y. T. Feng and D. R. Owen. Conjugate gradient methods for solving the smallest eigenpair of large symmetric eigenvalue problems. *Int. J. Numer. Meth. Engrg.*, **39**:2209-2229, 1996.

[8] A. V. Knyazev. Convergence rate estimates for iterative methods for mesh symmetric eigenvalue problem. *Soviet J. Numer. Anal. Math. Modelling*, **2**:371-396, 1987.

[9] A. V. Knyazev. A preconditioned conjugate gradient method for eigenvalue problems and its implementation in a subspace. *International Series of Numerical Mathematics*, v. 96, pp. 143-154, 1991.

[10] A. V. Knyazev. New estimates for Ritz vectors. *Math. Comput.*, **66**:985-995, 1997.

[11] A. V. Knyazev, Toward the optimal preconditioned eigensolver: Locally optimal block preconditioned conjugate gradient method. *SIAM J. Sci. Comp.*, **2**:517-541, 2001.

[12] A. V. Knyazev and K. Neymeyr, A geometric theory for preconditioned inverse iteration. III: A short and sharp convergence estimate for generalized eigenvalue problems. To appear in *Linear Algebra and Applications*, 2001.

[13] A. V. Knyazev and K, Neymeyr. Efficient solution of symmetric eigenvalue problems using multigrid preconditioners in the locally optimal block conjugate gradient method. To appear in *ETNA*, 2001.

[14] A. V. Knyazev and A. L. Skorokhodov. The preconditioned gradient-type iterative methods in a subspace for partial generalized symmetric eigenvalue problem. *SIAM J. Numer. Anal.*, **31**:1226-1239, 1994.

[15] K. Neymeyr. A geometric theory for preconditioned inverse iteration applied to a subspace. Accepted for publication in *Math. Comput.*, 2000.

[16] K. Neymeyr. A geometric theory for preconditioned inverse iteration. I: Extrema of the Rayleigh quotient. *Linear Algebra Appl.*, **322**:61-85, 2001.

[17] K. Neymeyr. A geometric theory for preconditioned inverse iteration. II: Convergence estimates. *Linear Algebra Appl.*, **322**:87-104, 2001.

[18] Y. Notay. Combination of Jacobi-Davidson and conjugate gradients for the partial symmetric eigenproblem. To appear in *Num. Lin. Alg. Appl.*, 2001.

[19] S. Oliveira. On the convergence rate of a preconditioned subspace eigensolver. *Computing*, **63**(3): 219-231, 1999.

[20] E. Ovtchinnikov. Convergence estimates for the generalized Davidson method for symmetric eigenvalue problems I. The preconditioning aspect. Submitted to *SIAM J. Numer. Anal.*, 2002.

[21] E. Ovtchinnikov. Generalized Davidson versus Conjugate Gradient Methods: A Numerical Study. In preparation, 2002.

[22] E. E. Ovtchinnikov and L. S. Xanthis. Successive eigenvalue relaxation: a new method for generalized eigenvalue problems and convergence estimates. *Proc. R. Soc. Lond. A*, **457**:441-451, 2001.

[23] B. N. Parlett. The Rayleigh quotient iteration and some generalizations. *Math. Comp.*, **28**:679-693, 1974.

[24] Y. Saad. *Numerical methods for large eigenvalue problems*, Algorithms and Architectures for Advanced Scientific Computing Series, Manchester University Press and Halsted Press, 1992.

[25] B. Samokish. The steepest descent method for an eigenvalue problem with semi-bounded operators. *Izv. Vyssh. Uchebn. Zaved. Mat.*, **5**:105-114, 1958.

[26] H. F. Weinberger, *Variational Methods for Eigenvalue Approximation*, SIAM, Philadelphia, PA, 1974.