

# Adaptive Eigenvalue Computation - Complexity Estimates\*

W. Dahmen, T. Rohwedder, R. Schneider, A. Zeiser

November 7, 2007

## Abstract

This paper is concerned with the design and analysis of a fully adaptive eigenvalue solver for linear symmetric operators. After transforming the original problem into an equivalent one formulated on  $\ell_2$ , the space of square summable sequences, the problem becomes sufficiently well conditioned so that a gradient type iteration can be shown to reduce the error by some fixed factor per step. It then remains to realize these (ideal) iterations within suitable dynamically updated error tolerances. It is shown under which circumstances the adaptive scheme exhibits in some sense asymptotically optimal complexity.

## 1 Introduction

### 1.1 Background

In a Gelfand triple  $\mathcal{H} \xrightarrow{d} \mathcal{X} \xrightarrow{d} \mathcal{H}'$  of Hilbert spaces, where  $\mathcal{H}$  is densely embedded in  $\mathcal{X}$ , and the dual pairing induced by the inner product  $\langle \cdot, \cdot \rangle$  of  $\mathcal{X}$ , let

$$\mathcal{L} : \mathcal{H} \rightarrow \mathcal{H}' \tag{1.1}$$

be a linear operator that takes  $\mathcal{H}$  onto its normed dual  $\mathcal{H}'$ . We wish to find, under certain assumptions on  $\mathcal{L}$ , an eigenpair  $(\lambda, u)$  of the problem

$$\mathcal{L}u = \lambda \mathcal{E}u \tag{1.2}$$

corresponding to the smallest eigenvalue  $\lambda$ . Here  $\mathcal{E} : u \mapsto \langle \cdot, u \rangle$  canonically embeds  $\mathcal{H}$  in  $\mathcal{H}'$ . Of course, (1.2) is to be understood in the weak sense, i.e.

$$\langle v, \mathcal{L}u \rangle = \lambda \langle v, \mathcal{E}u \rangle, \quad v \in \mathcal{H}, \tag{1.3}$$

---

\*This research was supported in part by the Leibniz Programme of the DFG, by the SFB 401 funded by DFG, the DFG Priority Program SPP1145 and by the EU NEST project BigDFT.

In view of the above choice of the dual pairing  $\langle \cdot, \cdot \rangle$  we shall from now on simply write  $\langle v, w \rangle$  instead of  $\langle v, \mathcal{E}w \rangle$  for  $v, w \in \mathcal{H}$ .

For typical applications, one can think of  $\mathcal{X}$  as an  $L_2$ -space over some domain  $\Omega$  and  $\mathcal{H}$  as a Sobolev space  $H^t$  of positive real order  $t$  (or as a closed subspace of a Sobolev space, determined e.g. by homogeneous boundary conditions).

We shall be concerned with the case that the operator  $\mathcal{L}$  is symmetric, i.e.

$$\langle \mathcal{L}v, w \rangle = \langle v, \mathcal{L}w \rangle, \quad \text{for all } v, w \in \mathcal{H} \quad (1.4)$$

and bounded and strongly positive, i.e. there exist positive constants  $c_L, C_L$  such that

$$c_L \|v\|_{\mathcal{H}}^2 \leq \langle \mathcal{L}v, v \rangle =: \|v\|_{\mathcal{L}}^2 \leq C_L \|v\|_{\mathcal{H}}^2, \quad v \in \mathcal{H}. \quad (1.5)$$

Note that if the (real) spectrum of  $\mathcal{L}$  is bounded from below by a negative number,  $\mathcal{L}$  can be shifted by a suitable parameter  $\mu$  such that  $\mathcal{L} - \mu\mathcal{E}$  satisfies (1.5). With this slight modification, this assumption is met by e.g. the electronic Schrödinger operator [ReSi], one particle Schrödinger operators in  $\mathbb{R}^3$  with certain potentials [Weid, HiSi] and eigenvalue problems for strongly elliptic (satisfying a Garding inequality) differential operators on bounded domains as well as by eigenvalue problems for Fredholm integral operators.

Note that (1.5) also implies that  $\mathcal{L}$  is a norm-isomorphism from  $\mathcal{H}$  onto  $\mathcal{H}'$  whose condition is bounded by  $C_L/c_L$ .

We shall exclusively deal with eigenvalue problems for which the infimum of the spectrum is an isolated eigenvalue  $\underline{\lambda}$ . For simplicity we shall actually assume that

$$\underline{\lambda} := \inf \frac{\langle \mathcal{L}u, u \rangle}{\langle u, u \rangle}$$

is a simple eigenvalue with corresponding eigenvector  $u$ , and that that the rest of the spectrum is bounded from below by  $\Lambda > \underline{\lambda}$ , which means that

$$\frac{\langle \mathcal{L}v, v \rangle}{\langle v, v \rangle} \geq \Lambda \quad \text{holds for all } v \in \mathcal{H} \quad \text{with} \quad \langle \mathcal{L}u, v \rangle = 0. \quad (1.6)$$

The conventional approach is to discretize (1.2), e.g. by finite elements or finite differences, which gives rise to a finite dimensional discrete problem

$$\mathbf{A}_h \mathbf{u}_h = \mu \mathbf{C}_h \mathbf{u}_h \quad (1.7)$$

where, for a given basis of the trial space,  $\mathbf{A}_h, \mathbf{C}_h$  are the corresponding stiffness matrix of  $\mathcal{L}$  and the mass matrix, respectively, and  $\mathbf{u}_h$  denotes the coefficient vector of the approximate eigenvector. Now the issue becomes to solve (1.7) efficiently within a suitable accuracy tolerance associated with the discretization. There is a vast literature on this issue, see e.g. Parlett [Parl], Golub/Van Loan [GvL], which are standard text books in

numerical linear algebra, or Chatelin [Chat], Babuska-Osborn [BabO] concerning Galerkin discretizations and Dyaknov [Dyak], Knyazev/Neymair [BPK] for preconditioned iteration schemes. See also [RSZ] for further references and comments.

In this paper we follow a different line. While the subsequent analysis will apply to finite dimensional problems of the form (1.7) as well, our primary interest is to avoid the separation of the discretization and solution process. Rather, we will design an abstract iteration scheme that solves the original *infinite dimensional* problem within a given accuracy tolerance, where the iterates of this scheme are taken, in principle, from all of  $\mathcal{H}$ . We will then show how to realize these updates (up to a perturbation) in a finite dimensional setting, particularly directing our attention to carrying out this task at possibly low computational cost. After completion of this work, we became aware of related work in [GG] where an adaptive finite element scheme for elliptic eigenvalue problems is shown to converge without giving complexity estimates though.

In principle, such infinite dimensional iterations can be formulated in different ways, focussing either on the convergence of Rayleigh quotients or of eigendirections. While the first option - although with the same motivation as in the present work - has been adopted in [RSZ], we address here the algorithmic realization of the second option. In fact, in [RSZ], from a somewhat different perspective, we focus on convergence of preconditioned iteration schemes per se, whereas in the present paper we develop and analyze a convergent adaptive algorithm. Our analysis given here provides (a posteriori) criteria for adaptive updates and yields complexity estimates that prove the (asymptotic) optimality of the scheme.

We note that problems of the kind (1.3) may be treated by means of inverse iteration or better by Davidson or Jacobi Davidson type methods as well. These approaches require the solution of a linear system in each iteration step. For this purpose one may apply the adaptive solution strategies proposed in [CDD1, CDD2, CDD3] based again on various types of (infinite dimensional) iteration schemes. In contrast to this strategy, in the present approach we will intertwine the outer loop, e.g. inverse iteration or Jacobi-Davidson and the iterative solver in the inner loop by updating the Rayleigh quotient in each inner iteration step. In fact, the present approach can be viewed as a preconditioned steepest descent method for minimizing the Rayleigh quotient. At any rate, one should stress that for large systems of linear equations the use of iterative methods would be inevitable, anyway.

As in Jacobi Davidson or in Krylov space methods, one can also apply subspace acceleration techniques to improve the present scheme by computing Ritz values of a rather small system. This requires the computation of scalar products like those appearing in the Rayleigh quotient. Since we deal with infinite matrices in the present paper we can only compute these values approximatively, see also Section 4.3 for further details. Regarding

these possible improvements, we content ourselves here with brief indications and further studies are required.

We proceed now with showing how the problem (1.2) is transformed into an equivalent one formulated over a sequence space which is, however, in some sense better conditioned. To this end, we begin with collecting a few facts that are relevant for carrying out this program. Moreover, this will motivate the assumptions made in subsequent sections.

## 1.2 Transformed Problems

Instead of projecting the problem (1.2) to a (fixed) finite dimensional subspace  $\mathcal{H}_h \subset \mathcal{H}$  of  $\mathcal{H}$ , we shall transform (1.2) into an *equivalent* problem defined on the sequence space  $\ell_2(\mathcal{I})$  of square summable sequences of some possibly infinite index set  $\mathcal{I}$  endowed with the inner product

$$\langle \mathbf{v}, \mathbf{w} \rangle = \sum_{i \in \mathcal{I}} v_i w_i, \quad \text{for all } \mathbf{v}, \mathbf{w} \in \ell_2(\mathcal{I})$$

and induced norm  $\|\cdot\|$ . Note that we use the same symbols for the inner product on  $\ell_2(\mathcal{I})$  as for the dual pairing on  $\mathcal{H}' \times \mathcal{H}$ .

The key ingredient is a suitable Riesz basis  $\Psi = \{\psi_i : i \in \mathcal{I}\}$  of  $\mathcal{H}$ , i.e. there exist positive constants  $c_\Psi, C_\Psi$  such that

$$c_\Psi \|\mathbf{v}\| \leq \left\| \sum_{i \in \mathcal{I}} v_i \psi_i \right\|_{\mathcal{H}} \leq C_\Psi \|\mathbf{v}\|, \quad \text{for all } \mathbf{v} \in \ell_2(\mathcal{I}). \quad (1.8)$$

Then (1.2) is equivalent to

$$\mathbf{A}\mathbf{u} = \lambda \mathbf{C}\mathbf{u}, \quad (1.9)$$

where

$$\mathbf{A} := (\langle \mathcal{L}\psi_j, \psi_i \rangle)_{i,j \in \mathcal{I}}, \quad \mathbf{C} := (\langle \psi_j, \psi_i \rangle)_{i,j \in \mathcal{I}}. \quad (1.10)$$

For the operator  $\mathbf{A}$  the properties (1.5) of  $\mathcal{L}$  together with (1.8) imply (see e.g. [D]) that

$$\|\mathbf{A}\|_{\ell_2 \rightarrow \ell_2} \leq C_L C_\Psi^2, \quad \|\mathbf{A}^{-1}\|_{\ell_2 \rightarrow \ell_2} \leq c_L^{-1} c_\Psi^{-2}, \quad (1.11)$$

for the problem on  $\ell_2(\mathcal{I})$ , which in turn means that

$$c_L c_\Psi^2 \|\mathbf{v}\|^2 \leq \|\mathbf{v}\|_{\mathbf{A}}^2 \leq C_L C_\Psi^2 \|\mathbf{v}\|^2, \quad (1.12)$$

where  $\|\cdot\|_{\mathbf{A}}^2 := \langle \mathbf{A}\cdot, \cdot \rangle$ . Thus,  $\mathbf{A}$ -ellipticity is equivalent to bounded invertibility of  $\mathbf{A}$  on  $\ell_2(\mathcal{I})$ . One could say that the transformation has a built-in preconditioning effect: The original  $\mathcal{L}$  has been transformed to an operator  $\mathbf{A}$  which is well conditioned on  $\ell_2(\mathcal{I})$  according to (1.12), a fact that will play a crucial role in solving (1.9) numerically.

As for the properties of  $\mathbf{C}$ , in the above setting the operator  $\mathbf{C}$  stems from the inner product on  $\mathcal{X}$  and is therefore symmetric, bounded and positive definite. The boundedness

of  $\mathbf{C}$  as a linear operator on  $\ell_2(\mathcal{I})$  results from the continuous embedding of  $\mathcal{H}$  in  $\mathcal{X}$ . However  $\mathbf{C}$  is typically not coercive on  $\ell_2(\mathcal{I})$ .

The role of  $\mathbf{C}$  is further illuminated when specifying the setting slightly to the following situation which may actually serve as a guiding example. In fact, consider again the example  $\mathcal{H} = H^t$ ,  $\mathcal{X} = L_2$ . Here, a suitable choice of the basis  $\Psi$  is a wavelet-type basis. In fact, such wavelet bases are available by now for a wide range of domains and exhibit an important property, namely that a scaled version of  $\Psi$  is also a Riesz basis for the pivot space  $\mathcal{X}$ . That means, setting

$$\mathbf{D} := \text{diag}(d_i := \langle \psi_i, \psi_i \rangle^{1/2} = \|\psi_i\|_{\mathcal{X}} : i \in \mathcal{I}),$$

the collection

$$\Psi^\circ := \{d_i^{-1}\psi_i : i \in \mathcal{I}\}$$

satisfies

$$c'_\Psi \|\mathbf{D}\mathbf{v}\| \leq \left\| \sum_{i \in \mathcal{I}} v_i \psi_i \right\|_{\mathcal{X}} \leq C'_\Psi \|\mathbf{D}\mathbf{v}\|, \quad \mathbf{D}\mathbf{v} \in \ell_2(\mathcal{I}), \quad (1.13)$$

where  $c'_\Psi, C'_\Psi$  are again fixed positive constants. It is well-known that for  $\mathcal{X} = L_2$  and  $\mathcal{H} = H^t$  one has  $d_i = \langle \psi_i, \psi_i \rangle^{1/2} \sim 2^{-t|i|}$ , where  $|i|$  denotes the dyadic level of the wavelet  $\psi_i$ . Moreover, the matrices  $\mathbf{A}$  and  $\mathbf{D}^{-1}\mathbf{C}\mathbf{D}^{-1}$  are spectrally equivalent, i.e. there are constants  $c^*, C^*$  such that

$$c^* \|\mathbf{v}\|_{\mathbf{D}^{-1}\mathbf{C}\mathbf{D}^{-1}} \leq \|\mathbf{v}\|_{\mathbf{A}} \leq C^* \|\mathbf{v}\|_{\mathbf{D}^{-1}\mathbf{C}\mathbf{D}^{-1}}, \quad \mathbf{v} \in \ell_2(\mathcal{I}). \quad (1.14)$$

**Remark 1.** *In both of the above settings,  $\mathbf{C}$  is symmetric positive definite and bounded but typically not coercive on  $\ell_2(\mathcal{I})$ . However,  $\mathbf{C}$  is coercive on the space  $\{\mathbf{x} : \mathbf{D}\mathbf{x} \in \ell_2(\mathcal{I})\}$  equipped with the proper norm in the second example.*

*Note also that if the basis functions  $\psi_i$  are pairwise orthogonal with respect to the pivot inner product  $\langle \cdot, \cdot \rangle$ , then  $\mathbf{C}$  is actually a diagonal matrix.*

The spectrum of the generalized eigenvalue problem  $\mathbf{A}\mathbf{u} = \lambda\mathbf{C}\mathbf{u}$  coincides with the original spectrum of (1.2). In particular, since  $\mathbf{C}$  is positive definite (but not necessarily coercive) on  $\ell_2(\mathcal{I})$ , 2.1 is equivalent to  $\mathbf{C}^{-1/2}\mathbf{A}\mathbf{C}^{-1/2}\mathbf{y} = \lambda\mathbf{y}$ . The minimal eigenvalue is then given by

$$0 < \underline{\lambda} = \min_v \frac{\langle \mathcal{L}v, v \rangle}{\langle v, v \rangle} = \min_{\mathbf{y}} \frac{\langle \mathbf{C}^{-1/2}\mathbf{A}\mathbf{C}^{-1/2}\mathbf{y}, \mathbf{y} \rangle}{\langle \mathbf{y}, \mathbf{y} \rangle} = \min_{\mathbf{x}} \frac{\langle \mathbf{A}\mathbf{x}, \mathbf{x} \rangle}{\langle \mathbf{C}\mathbf{x}, \mathbf{x} \rangle}, \quad (1.15)$$

so that the minimal eigenvalue of the transformed problem equals the one of the original problem and  $\underline{\lambda}$  is simple. Denoting by  $\mathbf{u} \in \ell_2(\mathcal{I})$  the corresponding eigenvector, one has

$$\frac{\langle \mathbf{A}\mathbf{v}, \mathbf{v} \rangle}{\langle \mathbf{C}\mathbf{v}, \mathbf{v} \rangle} \geq \underline{\lambda}, \quad \text{for all } \mathbf{v} \in \ell_2(\mathcal{I}) \text{ with } \langle \mathbf{A}\mathbf{v}, \mathbf{u} \rangle = 0. \quad (1.16)$$

The properties collected above will guide us later when formulating the precise conditions all subsequent developments will be based upon.

### 1.3 Objectives and Layout

The above considerations indicate how to arrive at transformed equivalent eigenproblems on sequence spaces (1.9) that are better conditioned in the sense that the spectrum of the resulting matrix  $\mathbf{A}$  is enclosed in a bounded interval of the positive real semi-axis. In particular, the resulting ellipticity on certain orthogonal complements will be seen to give rise to iteration schemes that reduce the error of the approximate smallest eigenvalue and also the deviation of the corresponding approximate eigenspace from the exact one by at least some fixed factor less than one. This is the case for the full infinite dimensional problem as well as for any finite dimensional discretization resulting from projecting onto spans of subsets of the Riesz basis  $\Psi$ .

Once the fixed error reduction has been established for the full infinite dimensional problem, the principal idea is to mimic this ideal iteration on the infinite dimensional problem (that still contains all information) by ultimately carrying out these iterations only approximately. One then faces the following two tasks:

- (i) Find appropriate stage dependent tolerances within which each iteration is to be solved so as to still guarantee convergence to the exact solution;
- (ii) Devise numerical schemes that realize the approximate application in the involved operators within the given target accuracy at possibly low computational cost.

In this paper we explore the potential of such an adaptive solution strategy on a primarily theoretical level. Our central objective is to analyze the intrinsic complexity of accuracy oriented eigenvalue computations. In particular, we shall show under which circumstances such a scheme has asymptotically optimal complexity in the following sense: If the solution  $u$  belongs to some approximation space  $\mathcal{A}^s$ , which means that  $N$  terms in the expansion of  $u$  suffice to approximate  $u$  in  $\mathcal{H}$  within accuracy of order  $N^{-s}$ , then the computational work for realizing a target  $\varepsilon$  remains bounded by a fixed multiple of  $\varepsilon^{-1/s}$ , when  $\varepsilon$  tends to zero.

While the present work is certainly inspired by prior work on adaptive solution schemes for operator equations (see e.g. [CDD2, CDD3]), there are some noteworthy differences. On one hand the iteration is nonlinear, on the other hand, one approximates coefficient arrays that are only determined up to normalization.

The layout of the paper is as follows. After summarizing the fundamental properties of the problem in Section 2, we formulate in Section 3 an iterative scheme for the computation of the smallest eigenvalue of the infinite dimensional problem (1.2) (or equivalently (1.9)) and a corresponding eigenspace. Moreover, we analyze the convergence of this idealized iteration making essential use of the ellipticity of  $\mathbf{A} - \lambda\mathbf{C}$  on certain orthogonal complements. Section 4 is devoted to the numerical realization of the idealized scheme

based on the approximate realization of residuals. We first show under which circumstances and for which accuracy tolerances such perturbed schemes converge to the exact solution of (1.9). Then, in a second step, we analyze the complexity of such schemes under the assumption that the involved operators  $\mathbf{A}, \mathbf{C}$  are in a certain sense *quasi-sparse*. We note that this property is known to hold for a wide range of operators. The main result is that the proposed adaptive scheme exhibits in some sense asymptotically optimal complexity. While the underlying adaptive scheme is to be viewed as one possible prototype realization with certain asymptotic properties, we conclude the section with indicating ways of quantitative improvements. Finally, in Section 5 we briefly indicate a typical application background.

## 2 Problem formulation

Motivated by the above considerations, we will in the remainder of this paper restrict our treatment to determining the smallest eigenvalue  $\underline{\lambda}$  and the associated eigenvector  $\mathbf{u}$  of the generalized eigenvalue problem

$$\mathbf{A}\mathbf{u} = \underline{\lambda}\mathbf{C}\mathbf{u}, \quad (2.1)$$

formulated on the infinite dimensional space  $\ell_2(\mathcal{I})$ . As in the previous examples we will only consider the case where  $\mathbf{A}, \mathbf{C}$  are symmetric, positive definite matrices defined on  $\ell_2(\mathcal{I})$  endowed as above with the norm  $\|\mathbf{x}\|^2 := \sum_{i \in \mathcal{I}} |x_i|^2 =: \langle \mathbf{x}, \mathbf{x} \rangle$ . Let us again stress that the index set  $\mathcal{I}$  could (and actually will) be infinite.

Further, we will impose the following conditions on our problem (2.1) reflecting the properties we have identified in the framework discussed before.

**Property 1.** *There exist positive constants  $\gamma, \Gamma$  such that*

$$\gamma \|\mathbf{x}\|^2 \leq \|\mathbf{x}\|_{\mathbf{A}}^2 \leq \Gamma \|\mathbf{x}\|^2, \quad (2.2)$$

where  $\|\cdot\|_{\mathbf{A}} := \langle \mathbf{A}\cdot, \cdot \rangle^{\frac{1}{2}}$  denotes the  $\mathbf{A}$ -(energy)-norm, see (1.12).

**Property 2.** *For the minimal generalized simple eigenvalue  $\underline{\lambda}$  of 2.1, there holds*

$$0 < \underline{\lambda} = \min_{\mathbf{x}} \frac{\langle \mathbf{A}\mathbf{x}, \mathbf{x} \rangle}{\langle \mathbf{C}\mathbf{x}, \mathbf{x} \rangle}, \quad (2.3)$$

while there exists a  $\Lambda > \underline{\lambda}$  one has for all  $\mathbf{x}$  with  $\langle \mathbf{A}\mathbf{u}, \mathbf{x} \rangle = 0$

$$\frac{\langle \mathbf{A}\mathbf{x}, \mathbf{x} \rangle}{\langle \mathbf{C}\mathbf{x}, \mathbf{x} \rangle} \geq \Lambda > \underline{\lambda}. \quad (2.4)$$

**Remark 2.** In the framework discussed in the previous section it was already observed that  $\mathbf{C}$  is bounded. Note that this actually follows from Properties 1, 2, namely one has for  $C_{\mathbf{C}} := \Gamma/\underline{\lambda}$  that

$$\|\mathbf{x}\|_{\mathbf{C}} \leq C_{\mathbf{C}} \|\mathbf{x}\|. \quad (2.5)$$

On the other hand, it follows from (2.2) together with (1.15) that  $\mathbf{C}$  is coercive on all of  $\ell_2$  if and only if  $\mathcal{L}$  is bounded with respect to the  $\mathcal{X}$ -inner product, a condition that is usually not fulfilled, for instance when  $\mathcal{L}$  is a differential operator.

### 3 Basic Algorithm - A Richardson-style method for calculating the smallest eigenvalue

In this section, we shall now formulate an “ideal iteration scheme” for (2.1). This scheme and the tools used to prove convergence will be utilized later in the analysis for our adaptive algorithm to be introduced in Section 4; therefore, the convergence analysis will be given in detail. Let  $\tilde{\underline{\lambda}}$  and  $\tilde{\Lambda}$  be lower and upper bounds for  $\underline{\lambda}$  and  $\Lambda$ , respectively. Then the iteration reads as follows.

#### 3.1 Basic Algorithm:

---

##### ***MINIT***

---

**Require:** initial value  $\mathbf{x}_0 \in \ell_2(\mathcal{I})$ ,  $\|\mathbf{x}_0\| = 1$ ,

$$\alpha = 2((1 - \tilde{\underline{\lambda}}/\tilde{\Lambda})\gamma + \Gamma)^{-1}$$

**Iteration:**

For  $n = 0, 1, \dots$  do

Calculate the Rayleigh quotient  $\lambda^{(n)} = \frac{\langle \mathbf{A}\mathbf{x}_n, \mathbf{x}_n \rangle}{\langle \mathbf{C}\mathbf{x}_n, \mathbf{x}_n \rangle}$ ;

Calculate the residual  $\mathbf{r}_n = \mathbf{A}\mathbf{x}_n - \lambda^{(n)}\mathbf{C}\mathbf{x}_n$ .

Let  $\hat{\mathbf{x}}_{n+1} = \mathbf{x}_n - \alpha\mathbf{r}_n$ ;

Normalize  $\mathbf{x}_{n+1} = \langle \hat{\mathbf{x}}_{n+1}, \hat{\mathbf{x}}_{n+1} \rangle^{-\frac{1}{2}} \hat{\mathbf{x}}_{n+1}$ .

endfor

---

This algorithm resembles a preconditioned Richardson iteration  $\hat{\mathbf{x}}_{n+1} = \Phi^{(n)}\mathbf{x}_n$  with a stage dependent iteration matrix

$$\Phi^{(n)} := \mathbf{I} - \alpha(\mathbf{A} - \lambda^{(n)}\mathbf{C})$$

depending on the  $n$ -th iterate  $\mathbf{x}_n$  and a relaxation parameter  $\alpha$ . On the other hand, defining the (generalized) Rayleigh quotient  $\lambda(\mathbf{x}) = \frac{\langle \mathbf{A}\mathbf{x}, \mathbf{x} \rangle}{\langle \mathbf{C}\mathbf{x}, \mathbf{x} \rangle}$  one has  $\nabla\lambda(\mathbf{x}_n) = \frac{2}{\langle \mathbf{C}\mathbf{x}_n, \mathbf{x}_n \rangle}\mathbf{r}_n$ , so that a steepest descent step with stepsize  $\beta_n$  takes the form

$$\mathbf{x}_n - \frac{2\beta_n}{\langle \mathbf{C}\mathbf{x}_n, \mathbf{x}_n \rangle}\mathbf{r}_n.$$

Thus, the above iteration could be viewed as a steepest descent method for the generalized Rayleigh quotient. This iteration scheme is also known as a preconditioned inverse iteration scheme PINVIT, see e.g. [KN, RSZ]. Note that the preconditioning is already provided through the formulation as a well-conditioned problem in  $\ell_2(\mathcal{I})$ .

This algorithm can be improved by subspace acceleration techniques. For instance, the Rayleigh quotient might be minimized according to

$$\mathbf{x}_{n+1} := \operatorname{argmin} \{ \lambda(\mathbf{v}) : \mathbf{v} \in \operatorname{span} \{ \mathbf{x}_n, \mathbf{r}_n \}, \|\mathbf{v}\| = 1 \}. \quad (3.1)$$

This corresponds to optimal line search in a steepest descent step. However, in order to have a technically simpler exposition of the principal mechanisms of our approach we confine the subsequent discussion first to the above simple Richardson type iteration and will take up possible improvements along with computational consequences later in Section 4.3.

In the remainder of this section, we will analyze the properties of Algorithm *MINIT* and in particular prove convergence to the smallest eigenvector, provided the starting vector lies sufficiently close to the target  $\mathbf{u}$ . The main results will then be compiled in Theorem 1 at the end of this section.

## 3.2 Error reduction

We shall show that for suitable damping parameters  $\alpha$  the above (ideal) iteration (on the full infinite dimensional space  $\ell_2(\mathcal{I})$ ) provides approximations to the searched eigenpair with a guaranteed fixed error reduction per step.

### 3.2.1 Preliminary considerations

Setting  $\boldsymbol{\delta}_n := \boldsymbol{\delta}(\mathbf{x}_n) := \mathbf{x}_n - \mathbf{u}$  for the normalized iterates  $\mathbf{x}_n$ , we will show that the error components that are orthogonal to  $\mathbf{u}$  tend to zero in the Euclidean  $\ell_2(\mathcal{I})$ -norm. Denoting by  $\mathbf{P}\mathbf{x} = \frac{\langle \mathbf{x}, \mathbf{u} \rangle}{\langle \mathbf{u}, \mathbf{u} \rangle} \mathbf{u}$  the  $\ell_2$ -orthogonal projection of  $\mathbf{x}$  into the eigenspace  $U_0$  containing the eigenvector  $\mathbf{u}$ , the orthogonal error component is given by  $\boldsymbol{\delta}_n^\perp = (\mathbf{I} - \mathbf{P})\boldsymbol{\delta}_n$ , we introduce

$$\|\boldsymbol{\delta}_n^\perp\| = \|(\mathbf{I} - \mathbf{P})\boldsymbol{\delta}_n\| = \|(\mathbf{I} - \mathbf{P})\mathbf{x}_n\| =: \sin \angle(\mathbf{x}_n, \mathbf{u}) . \quad (3.2)$$

Note that  $\|\boldsymbol{\delta}_n^\perp\|$  does not depend on the normalization of  $\mathbf{u}$ . We shall now show that we have an at least linear reduction of the orthogonal error component of the iterates  $\mathbf{x}_n$  of (3.1), i.e.

$$\|\boldsymbol{\delta}_{n+1}^\perp\| \leq \xi \|\boldsymbol{\delta}_n^\perp\|, \quad (3.3)$$

with  $\xi < 1$ , provided the initial vector has a sufficiently small angle with the solution  $\mathbf{u}$  corresponding to the smallest eigenvalue  $\underline{\lambda}$ .

Towards this end, we first note that  $\mathbf{x}_n \perp \mathbf{r}_n$ , so that we have for the non-normalized iterates

$$\|\widehat{\mathbf{x}}_{n+1}\|^2 = \|\mathbf{x}_n - \alpha \mathbf{r}_n\|^2 = \|\mathbf{x}_n\|^2 + \alpha^2 \|\mathbf{r}_n\|^2 \geq \|\mathbf{x}_n\|^2 = 1. \quad (3.4)$$

Hence, the error of the  $\ell_2$ -normalized iterates  $\mathbf{x}_{n+1}$  can be estimated by

$$\|\delta_{n+1}^\perp\| = \frac{1}{\|\widehat{\mathbf{x}}_{n+1}\|} \|(\mathbf{I} - \mathbf{P})\widehat{\mathbf{x}}_{n+1}\| \leq \|(\mathbf{I} - \mathbf{P})\widehat{\mathbf{x}}_{n+1}\| =: \|\widehat{\delta}_{n+1}^\perp\|. \quad (3.5)$$

To go further, we decompose  $\widehat{\delta}_{n+1}^\perp$  as follows

$$\begin{aligned} \|\widehat{\delta}_{n+1}^\perp\| &= \|(\mathbf{I} - \mathbf{P})\Phi^{(n)}\mathbf{x}_n\| = \|(\mathbf{I} - \mathbf{P}) \left( (\mathbf{I} - \alpha(\mathbf{A} - \lambda^{(n)}\mathbf{C})) \mathbf{x}_n \right)\| \\ &\leq \|(\mathbf{I} - \mathbf{P}) (\mathbf{I} - \alpha(\mathbf{A} - \underline{\lambda}\mathbf{C}))\delta_n\| + \|\alpha(\lambda^{(n)} - \underline{\lambda})(\mathbf{I} - \mathbf{P})\mathbf{C}\mathbf{x}_n\|, \end{aligned} \quad (3.6)$$

where we have used that

$$\mathbf{M}\mathbf{u} := (\mathbf{A} - \underline{\lambda}\mathbf{C})\mathbf{u} = \mathbf{0}.$$

Moreover, using the fact that  $\text{range}\mathbf{P} \subseteq \ker\mathbf{M}$  so that

$$\mathbf{M} = \mathbf{M}(\mathbf{I} - \mathbf{P}), \quad (3.7)$$

we conclude, upon using  $(\mathbf{I} - \mathbf{P})^2 = \mathbf{I} - \mathbf{P}$  and

$$(\mathbf{A} - \underline{\lambda}\mathbf{C})\delta_n = \mathbf{M}\delta_n = \mathbf{M}(\mathbf{I} - \mathbf{P})\delta_n = \mathbf{M}\delta_n^\perp,$$

that

$$\|\widehat{\delta}_{n+1}^\perp\| \leq \|(\mathbf{I} - \mathbf{P}) (\mathbf{I} - \alpha\mathbf{M}) \delta_n^\perp\| + C_{\mathbf{C}} \alpha |\lambda^{(n)} - \underline{\lambda}|. \quad (3.8)$$

We now estimate the two parts in the right hand side of (3.8) separately in the next two subsections.

### 3.2.2 Linear part of the iteration scheme

The main goal of this subsection will be to show the following property of the iteration matrix, from which in Lemma 4, an estimate for the left part of (3.8) will easily be derived.

**Lemma 1.** *Under the assumptions stated above, the matrix*

$$\mathbf{M} := \mathbf{A} - \underline{\lambda}\mathbf{C} \quad (3.9)$$

*is bounded and  $\ell_2$ -elliptic on the set*

$$V_0 = \{\mathbf{x} \in \ell_2(\mathcal{I}) : \langle \mathbf{x}, \mathbf{u} \rangle = 0\}, \quad (3.10)$$

*that is, there exist  $\theta, \Theta > 0$  such that*

$$\theta \|\mathbf{x}\|^2 \leq \langle \mathbf{M}\mathbf{x}, \mathbf{x} \rangle \leq \Theta \|\mathbf{x}\|^2 \quad \text{for all } \mathbf{x} \in V_0. \quad (3.11)$$

For the proof of Lemma 1, we need the following general facts included for the convenience of the reader in form of the next two lemmata.

**Lemma 2.** *Let  $(\mathcal{X}, \langle \cdot, \cdot \rangle)$  be any inner product space with norm  $\|\cdot\|^2 := \langle \cdot, \cdot \rangle$ . Moreover, assume that  $x, v \in X$  fulfil the conditions  $\|v\| = \|x\|$  and  $\langle x, v \rangle \geq 0$ . Then, for  $P = \frac{\langle x, v \rangle}{\langle v, v \rangle} v$ , we have*

$$\frac{1}{\sqrt{2}} \|x - v\| \leq \|(I - P)x\| \leq \|x - v\|.$$

**Proof:** First, we note that it suffices to show the equivalence for  $\|v\| = \|x\| = 1$  and  $v \neq x$ . The second inequality is clear due to the fact that  $P$  is an  $\ell_2$ -orthogonal projector. For the inverse estimate, let  $\alpha_0 = \langle x, v \rangle \geq 0$ . We then have  $\|x - v\|^2 = \|x\|^2 + \|v\|^2 - 2\alpha_0 = 2 - 2\alpha_0$  and  $\|(I - P)x\|^2 = 1 - \alpha_0^2$ , which together gives

$$\frac{\|(I - P)x\|^2}{\|v - x\|^2} = \frac{1 + \alpha_0}{2} \geq \frac{1}{2}.$$

□

**Lemma 3.** *Let  $\langle \cdot, \cdot \rangle_1, \langle \cdot, \cdot \rangle_2$  be two equivalent inner products on a Hilbert space  $\mathcal{H}$ , i.e. they induce equivalent norms in  $\mathcal{H}$ ,  $G$  a symmetric operator with respect to  $\langle \cdot, \cdot \rangle_2$  (i.e.  $\langle Gx, y \rangle_2 = \langle x, Gy \rangle_2$  for all  $x, y \in \mathcal{H}$ ) and  $u \in \ker G$ . Then, if  $G$  is  $\langle \cdot, \cdot \rangle_2$ -elliptic on the  $\langle \cdot, \cdot \rangle_1$ -orthogonal complement of  $u$ , that is*

$$\langle Gx, x \rangle_2 \geq c \langle x, x \rangle_2 \tag{3.12}$$

*holds for all  $x$  with  $\langle x, u \rangle_1 = 0$ , then  $G$  is also  $\langle \cdot, \cdot \rangle_2$ -elliptic on the  $\langle \cdot, \cdot \rangle_2$ -orthogonal complement of  $u$ , i.e. (3.12) holds for all  $x$  with  $\langle x, u \rangle_2 = 0$  with a possibly different constant  $c$ .*

**Proof:** Let  $x \in \mathcal{H}$  fulfil  $\langle x, u \rangle_2 = 0$ . Decomposing  $x = P_1 x + x^{\perp 1}$ , where  $P_i x = \frac{\langle x, u \rangle_i}{\langle u, u \rangle_i} u$  denotes the  $i$ -orthogonal projector, we obtain  $Gx = Gx^{\perp 1}$  and therefore

$$\langle Gx, x \rangle_2 = \langle x^{\perp 1}, Gx \rangle_2 = \langle Gx^{\perp 1}, x^{\perp 1} \rangle_2 \gtrsim \|x^{\perp 1}\|_2^2 \sim \|x^{\perp 1}\|_1^2.$$

To prove the assertion, it remains to see that  $\|x^{\perp 1}\|_1 \sim \|x^{\perp 2}\|_2 = \|x\|_2$ , where  $x^{\perp 2} = x - \frac{\langle x, u \rangle_2}{\langle u, u \rangle_2} u = x$ . This latter fact follows from Lemma 2 above, where we can choose  $v$  to be a multiple of  $u$  such that  $\|v\|_1 = \|x\|_1$  and  $\langle x, v \rangle_1 \geq 0$ , giving

$$\|x^{\perp 1}\|_1 \sim \|x - v\|_1 \sim \|x - v\|_2 \geq \|(I - P_2)(x - v)\|_2 = \|x^{\perp 2}\|_2.$$

This confirms the assertion. □

**Proof of lemma 1.** For the ellipticity, we show the claim for all  $\mathbf{x}$  with  $\langle \mathbf{x}, \mathbf{u} \rangle_{\mathbf{A}} = 0$ ; setting  $\langle \cdot, \cdot \rangle_1 = \langle \cdot, \cdot \rangle_{\mathbf{A}}$  in Lemma 3 proves then the coercivity on  $V_0$ . Indeed, for all such  $\mathbf{x}$ , we have  $\frac{\|\mathbf{x}\|_{\mathbf{A}}^2}{\|\mathbf{x}\|_{\mathbf{C}}^2} \geq \Lambda$ , by assumption 2, and therefore

$$\langle (\mathbf{A} - \underline{\lambda}\mathbf{C})\mathbf{x}, \mathbf{x} \rangle = \|\mathbf{x}\|_{\mathbf{A}}^2 - \underline{\lambda}\|\mathbf{x}\|_{\mathbf{C}}^2 \quad (3.13)$$

$$= \frac{\Lambda - \underline{\lambda}}{\Lambda} \|\mathbf{x}\|_{\mathbf{A}}^2 + \frac{\underline{\lambda}}{\Lambda} \|\mathbf{x}\|_{\mathbf{A}}^2 - \underline{\lambda}\|\mathbf{x}\|_{\mathbf{C}}^2 \quad (3.14)$$

$$\geq \frac{\Lambda - \underline{\lambda}}{\Lambda} \gamma \|\mathbf{x}\|^2 + \frac{\underline{\lambda}}{\Lambda} \Lambda \|\mathbf{x}\|_{\mathbf{C}}^2 - \underline{\lambda}\|\mathbf{x}\|_{\mathbf{C}}^2 \quad (3.15)$$

$$= \frac{\Lambda - \underline{\lambda}}{\Lambda} \gamma \|\mathbf{x}\|^2. \quad (3.16)$$

By

$$0 < \langle (\mathbf{A} - \lambda\mathbf{C})\mathbf{w}, \mathbf{w} \rangle \leq \langle \mathbf{A}\mathbf{w}, \mathbf{w} \rangle \leq \Gamma \|\mathbf{w}\|^2, \quad (3.17)$$

the boundedness of  $\mathbf{M}$  on  $V_0$  follows. □

**Lemma 4.** Let  $\Phi := (\mathbf{I} - \mathbf{P})(\mathbf{I} - \alpha\mathbf{M}) : V_0 \rightarrow V_0$  denote the matrix in the left part of (3.8). Then setting the parameter  $\alpha := 2((1 - \underline{\lambda}/\Lambda)\gamma + \Gamma)^{-1}$  in the Richardson method gives

$$\|\Phi\|_{V_0 \rightarrow V_0} \leq \beta < 1.$$

**Proof:** Due to the continuity and ellipticity of  $\mathbf{M}$  on  $V_0$ , we have

$$(1 - \alpha\Theta)\|\mathbf{x}\|^2 \leq \langle (\mathbf{I} - \alpha\mathbf{M})\mathbf{x}, \mathbf{x} \rangle \leq (1 - \alpha\theta)\|\mathbf{x}\|^2, \quad \forall \mathbf{x} \in V_0. \quad (3.18)$$

The same inequality applies to  $\Phi$  owing to the fact that  $\mathbf{P}$  is symmetric and therefore

$$\langle (\mathbf{I} - \mathbf{P})(\mathbf{I} - \alpha\mathbf{M})\mathbf{x}, \mathbf{x} \rangle = \langle (\mathbf{I} - \alpha\mathbf{M})\mathbf{x}, (\mathbf{I} - \mathbf{P})\mathbf{x} \rangle = \langle (\mathbf{I} - \alpha\mathbf{M})\mathbf{x}, \mathbf{x} \rangle, \quad \forall \mathbf{x} \in V_0.$$

We now let  $\beta = \frac{\Theta - \theta}{\Theta + \theta} < 1$ . Then, choosing  $\alpha := \frac{2}{\Theta + \theta} = 2((1 - \underline{\lambda}/\Lambda)\gamma + \Gamma)^{-1}$  (recalling the choice of  $\Theta$  and  $\theta$  from the proof of Lemma 1), equation (3.18) gives  $-\beta \leq \langle \Phi \mathbf{x}, \mathbf{x} \rangle \leq \beta$  and therefore  $\|\Phi\|_{V_0 \rightarrow V_0} \leq \beta < 1$ . □

### 3.2.3 Estimates for Rayleigh quotients and residuals

It remains to estimate the Rayleigh quotients in the right hand side of equation (3.8).

**Lemma 5.** Let  $\mathbf{x} \in \ell_2(\mathcal{I})$  with  $\|\mathbf{x}\| = 1$  and as before let  $\lambda(\mathbf{x})$  denote the corresponding Rayleigh quotient. Then we have

$$\lambda(\mathbf{x}) - \underline{\lambda} \leq \frac{2\Gamma}{\gamma} \lambda(\mathbf{x}) \|\delta(\mathbf{x})^\perp\|^2, \quad (3.19)$$

where the constants in inequality (3.19) are those from the norm estimate (2.2) and  $\boldsymbol{\delta}^\perp(\mathbf{x}) = (\mathbf{I} - \mathbf{P})(\mathbf{x} - \mathbf{u})$  as above.

**Proof:** Suppose that  $\mathbf{u}_x$  is an eigenvector belonging to the smallest eigenvalue  $\underline{\lambda}$  of the eigenvalue problem (2.1) which is normalized in such a way that  $\|\mathbf{u}_x\| = \|\mathbf{x}\|$  and  $\langle \mathbf{x}, \mathbf{u}_x \rangle \geq 0$ . Straightforward computation yields

$$\lambda(\mathbf{x}) - \underline{\lambda} = \frac{1}{\|\mathbf{x}\|_{\mathbf{C}}^2} (\|\mathbf{x} - \mathbf{u}_x\|_{\mathbf{A}}^2 - \underline{\lambda} \|\mathbf{x} - \mathbf{u}_x\|_{\mathbf{C}}^2). \quad (3.20)$$

Keeping the normalization of  $\mathbf{x}$  and (2.2) in mind and since there are no absolute values involved there, we infer from (3.20) that

$$\begin{aligned} \lambda(\mathbf{x}) - \underline{\lambda} &\leq \frac{1}{\|\mathbf{x}\|_{\mathbf{C}}^2} \|\mathbf{x} - \mathbf{u}_x\|_{\mathbf{A}}^2 = \frac{\|\mathbf{x}\|^2}{\|\mathbf{x}\|_{\mathbf{C}}^2} \|\mathbf{x} - \mathbf{u}_x\|_{\mathbf{A}}^2 \leq \frac{\|\mathbf{x}\|_{\mathbf{A}}^2}{\gamma \|\mathbf{x}\|_{\mathbf{C}}^2} \|\mathbf{x} - \mathbf{u}_x\|_{\mathbf{A}}^2 \\ &= \frac{\lambda(\mathbf{x})}{\gamma} \|\mathbf{x} - \mathbf{u}_x\|_{\mathbf{A}}^2 \leq \frac{\Gamma}{\gamma} \lambda(\mathbf{x}) \|\mathbf{x} - \mathbf{u}_x\|^2. \end{aligned}$$

Since Lemma 2 says that

$$\|\mathbf{x} - \mathbf{u}_x\| \leq \sqrt{2} \|(\mathbf{I} - \mathbf{P})\mathbf{x}\| = \sqrt{2} \|(\mathbf{I} - \mathbf{P})(\mathbf{x} - \mathbf{u}_x)\|,$$

the assertion (3.19) follows as well.  $\square$

**Remark 3.** As before let  $\boldsymbol{\delta} = \boldsymbol{\delta}(\mathbf{x}) := \mathbf{x} - \mathbf{u}$  and fix  $0 < a < \gamma/2\Gamma$ . If  $\|\boldsymbol{\delta}(\mathbf{x})^\perp\| \leq \sqrt{a}$ , one has for the corresponding Rayleigh quotient

$$\lambda(\mathbf{x}) \leq \underline{\lambda} \left(1 - \frac{2\Gamma a}{\gamma}\right)^{-1} =: K. \quad (3.21)$$

Taking e.g.  $a = \gamma/4\Gamma$  one has  $K = 2\underline{\lambda}$ .

**Proof:** From Lemma 5 we infer that  $\lambda(\mathbf{x}) \leq \underline{\lambda} \left(1 - \frac{2\Gamma}{\gamma} \|\boldsymbol{\delta}(\mathbf{x})^\perp\|^2\right)^{-1} \leq \underline{\lambda} \left(1 - \frac{2\Gamma a}{\gamma}\right)^{-1}$ .  $\square$

An immediate consequence of (3.21) can be stated as follows.

**Remark 4.** Under the property of Section 2 the operator  $\mathbf{C}$  is coercive on any sufficiently small fixed angular neighborhood of the direction  $\mathbf{u}$ , i.e. for  $K$  from (3.21) one has

$$\langle \mathbf{C}\mathbf{x}, \mathbf{x} \rangle \geq \gamma/K \quad \text{whenever } \|\mathbf{x}\| = 1, \ \|(\mathbf{u} - \mathbf{x})^\perp\| \leq \sqrt{a}. \quad (3.22)$$

The next observation is that the orthogonal error components are controlled by the residuals. To this end, note that

$$\mathbf{M}_\perp : V_0 \rightarrow V_0, \quad \mathbf{x} \mapsto (\mathbf{I} - \mathbf{P})\mathbf{M}\mathbf{x}$$

is bounded and elliptic so that  $\|\mathbf{M}_\perp^{-1}\|$  is a bounded positive number.

**Lemma 6.** For any  $\mathbf{x} \in \ell_2(\mathcal{I})$  with  $\|\mathbf{x}\| = 1$ , let  $\mathbf{r}(\mathbf{x}) := (\mathbf{A} - \lambda(\mathbf{x})\mathbf{C})\mathbf{x}$  be the corresponding residual with respect to the Rayleigh quotient  $\lambda(\mathbf{x}) := \frac{\langle \mathbf{A}\mathbf{x}, \mathbf{x} \rangle}{\langle \mathbf{C}\mathbf{x}, \mathbf{x} \rangle}$  and, as before, let  $\mathbf{x}^\perp = (\mathbf{I} - \mathbf{P})\mathbf{x}$  so that in particular  $\boldsymbol{\delta}(\mathbf{x})^\perp = (\mathbf{I} - \mathbf{P})(\mathbf{x} - \mathbf{u}) = (\mathbf{I} - \mathbf{P})\mathbf{x}$ . Assume that for the constant  $a$  from Remark 3

$$\|\boldsymbol{\delta}(\mathbf{x})^\perp\| \leq \min \left\{ \sqrt{a}, \frac{1}{2} \left\{ \left( 1 + \frac{\gamma}{\Gamma K C_{\mathbf{C}} \|\mathbf{M}_\perp^{-1}\|} \right)^{1/2} - 1 \right\} \right\}. \quad (3.23)$$

Then for  $M := 2\|\mathbf{M}_\perp^{-1}\|$  one has

$$\|\boldsymbol{\delta}(\mathbf{x})^\perp\| \leq M\|\mathbf{r}(\mathbf{x})^\perp\| \leq M\|\mathbf{r}(\mathbf{x})\| \leq M \left( \|\mathbf{M}\| + \frac{2\Gamma K C_{\mathbf{C}}}{\gamma} \|\boldsymbol{\delta}(\mathbf{x})^\perp\| \right) \|\boldsymbol{\delta}(\mathbf{x})^\perp\|. \quad (3.24)$$

**Proof:** Since

$$\mathbf{r}(\mathbf{x}) = \mathbf{M}(\mathbf{x} - \mathbf{u}) + (\underline{\lambda} - \lambda(\mathbf{x}))\mathbf{C}\mathbf{x} = \mathbf{M}(\mathbf{I} - \mathbf{P})(\mathbf{x} - \mathbf{u}) + (\underline{\lambda} - \lambda(\mathbf{x}))\mathbf{C}\mathbf{x},$$

we conclude, on account of Lemma 5 and Remark 3, that

$$\|\mathbf{r}(\mathbf{x})\| \leq \|\mathbf{M}\| \|\boldsymbol{\delta}^\perp(\mathbf{x})\| + \frac{2\Gamma K C_{\mathbf{C}}}{\gamma} \|\boldsymbol{\delta}^\perp(\mathbf{x})\|^2, \quad (3.25)$$

which confirms the upper estimate in (3.24).

As for the lower estimate, recall that by the definition of  $\mathbf{M}_\perp$ ,

$$\mathbf{M}_\perp \boldsymbol{\delta}(\mathbf{x})^\perp = (\mathbf{I} - \mathbf{P})(\mathbf{A} - \lambda(\mathbf{x})\mathbf{C})\boldsymbol{\delta}(\mathbf{x})^\perp + (\mathbf{I} - \mathbf{P})(\lambda(\mathbf{x}) - \underline{\lambda})\mathbf{C}\boldsymbol{\delta}(\mathbf{x})^\perp,$$

giving

$$\|\mathbf{M}_\perp^{-1}\|^{-1} \|\boldsymbol{\delta}(\mathbf{x})^\perp\| \leq \|\mathbf{M}_\perp \boldsymbol{\delta}(\mathbf{x})^\perp\| \leq \|(\mathbf{I} - \mathbf{P})(\mathbf{A} - \lambda(\mathbf{x})\mathbf{C})\boldsymbol{\delta}(\mathbf{x})^\perp\| + \|(\lambda(\mathbf{x}) - \underline{\lambda})\mathbf{C}\boldsymbol{\delta}(\mathbf{x})^\perp\|. \quad (3.26)$$

Now straightforward calculations yield

$$(\mathbf{A} - \lambda(\mathbf{x})\mathbf{C})\boldsymbol{\delta}(\mathbf{x})^\perp = \mathbf{r}(\mathbf{x}) - (\mathbf{A} - \lambda(\mathbf{x})\mathbf{C})\mathbf{P}\mathbf{x} = \mathbf{r}(\mathbf{x}) - \frac{\langle \mathbf{x}, \mathbf{u} \rangle}{\langle \mathbf{u}, \mathbf{u} \rangle} (\underline{\lambda} - \lambda(\mathbf{x}))\mathbf{C}\mathbf{u},$$

which gives

$$\|(\mathbf{I} - \mathbf{P})(\mathbf{A} - \lambda(\mathbf{x})\mathbf{C})\boldsymbol{\delta}(\mathbf{x})^\perp\| \leq \|\mathbf{r}(\mathbf{x})^\perp\| + C_{\mathbf{C}} |\underline{\lambda} - \lambda(\mathbf{x})|. \quad (3.27)$$

Invoking now Lemma 5, (3.19) together with Remark 3 and (3.26), yields

$$\|\mathbf{M}_\perp^{-1}\|^{-1} \|\boldsymbol{\delta}(\mathbf{x})^\perp\| \leq \|\mathbf{r}(\mathbf{x})^\perp\| + \frac{2\Gamma K C_{\mathbf{C}}}{\gamma} (1 + \|\boldsymbol{\delta}(\mathbf{x})^\perp\|) \|\boldsymbol{\delta}(\mathbf{x})^\perp\|^2. \quad (3.28)$$

Hence, whenever

$$\frac{2\Gamma K C_{\mathbf{C}}}{\gamma} (1 + \|\boldsymbol{\delta}(\mathbf{x})^\perp\|) \|\boldsymbol{\delta}(\mathbf{x})^\perp\| \leq \frac{1}{2} \|\mathbf{M}_\perp^{-1}\|^{-1}, \quad (3.29)$$

which is the case when (3.23) holds, we obtain

$$\|\boldsymbol{\delta}(\mathbf{x})^\perp\| \leq 2\|\mathbf{M}_\perp^{-1}\| \|\mathbf{r}(\mathbf{x})^\perp\|, \quad (3.30)$$

which is lower estimate in (3.24) with  $M := 2\|\mathbf{M}_\perp^{-1}\|$ .  $\square$

### 3.2.4 Convergence of the scheme

In summary, the above observations allow us to establish the following local convergence properties of the Richardson eigenvalue iteration.

**Theorem 1.** *Let the damping parameter  $\alpha$  in the basic algorithm **MINIT** from Section 3.1 be chosen according to Lemma 4, so that the bound  $\beta$  for  $\|\Phi\|_{V_0 \rightarrow V_0}$  in Lemma 4 satisfies  $\beta < 1$ . Furthermore assume that the initial error satisfies*

$$\|\delta_0^\perp\| \leq \min \left\{ a^{1/2}, \frac{1-\beta}{2\tilde{c}K} \right\} =: d, \quad (3.31)$$

where  $\tilde{c} := 2\alpha\Gamma C_C/\gamma$ . Then the following statements hold:

a) For  $\xi := (\beta + 1)/2 < 1$  one has

$$\|\delta_{n+1}^\perp\| \leq \xi \|\delta_n^\perp\|, \quad (3.32)$$

i.e. one has monotone linear error reduction with respect to the  $\ell_2$ -norm. Moreover, for any  $\epsilon > 0$  there is an  $n_\epsilon$  such that for  $n \geq n_\epsilon$  one can choose  $\xi < \beta + \epsilon$  in (3.32), i.e. one has an asymptotic error reduction of a rate  $\leq \beta$ . The orthogonal error components  $\delta_n^\perp$  also converge to zero in the  $\mathbf{A}$ -norm.

b) There exists a uniform constant  $C$  such that the error of the Rayleigh quotients  $\lambda(\mathbf{x}_n)$  is bounded by

$$\lambda(\mathbf{x}_n) - \underline{\lambda} \leq C \|\delta_n^\perp\|^2 \rightarrow 0. \quad (3.33)$$

c) There exist constants  $M, M' > 0$  and  $n_0 \in \mathbb{N}$  such that

$$\|\delta_n^\perp\| \leq M \|\mathbf{r}_n^\perp\| \leq M \|\mathbf{r}_n\| \leq M' \|\delta_n^\perp\|, \quad n \geq n_0. \quad (3.34)$$

Moreover, there exists  $\zeta < 1$  and  $n_1 \in \mathbb{N}$  such that

$$\|\mathbf{r}_{n+1}\| \leq \zeta \|\mathbf{r}_n\|, \quad n \geq n_1. \quad (3.35)$$

**Proof:** Recalling (3.5), (3.6) and employing Lemmata 4 and 5 provides

$$\begin{aligned} \|\delta_{n+1}^\perp\| &\leq \|\Phi\|_{V_0 \rightarrow V_0} \|\delta_n^\perp\| + \alpha \frac{2\Gamma C_C}{\gamma} \lambda(\mathbf{x}_n) \|\delta_n^\perp\|^2 \\ &\leq (\beta + \tilde{c} \lambda(\mathbf{x}_n) \|\delta_n^\perp\|) \|\delta_n^\perp\| =: \xi \|\delta_n^\perp\|, \end{aligned} \quad (3.36)$$

with  $\tilde{c} := 2\alpha\Gamma C_C/\gamma$ . It remains to show that  $\xi < 1$  for  $\|\delta_0^\perp\|$  sufficiently small. In fact, we know from Remark 3 that

$$\lambda(\mathbf{x}_0) \leq K, \quad (3.37)$$

provided that  $\delta_0 = \delta(\mathbf{x}_0)$  satisfies  $\|\delta_0^\perp\| \leq \sqrt{a}$  (cf. (3.31)). Moreover, by (3.36) for  $n = 0$ , we have

$$\|\delta_1^\perp\| \leq (\beta + \tilde{c}K \|\delta_0^\perp\|) \|\delta_0^\perp\|.$$

Now note that (3.31) implies that  $\tilde{c}K\|\boldsymbol{\delta}_0^\perp\| \leq (1 - \beta)/2$  so that

$$\beta + \tilde{c}K\|\boldsymbol{\delta}_0^\perp\| \leq \frac{1 + \beta}{2} =: \xi < 1, \quad (3.38)$$

which yields

$$\|\boldsymbol{\delta}_1^\perp\| < \|\boldsymbol{\delta}_0^\perp\|, \quad \lambda(\mathbf{x}_1) \leq K.$$

One easily shows now with the aid of Lemma 5 and (3.36) inductively that

$$\|\boldsymbol{\delta}_{n+1}^\perp\| \leq \xi_n \|\boldsymbol{\delta}_n^\perp\|, \quad \xi_n := \beta + \tilde{c}K\|\boldsymbol{\delta}_n^\perp\| \leq \xi < 1, \quad \lambda(\mathbf{x}_n) \leq K, \quad n \in \mathbb{N}. \quad (3.39)$$

This guarantees monotone convergence of the iteration. It is obvious from the above arguments that with  $\|\boldsymbol{\delta}_n^\perp\| \rightarrow 0$  the values of  $\xi_n$  drop down to  $\beta$ , showing the first assertion.

Convergence in the  $\mathbf{A}$ -norm is now an immediate consequence of (2.2) and (3.32) provided that  $\boldsymbol{\delta}_0$  satisfies (3.31). This proves a).

The convergence of the Rayleigh quotients  $\lambda(\mathbf{x}_n)$  now follows immediately from (3.19) and (3.39), giving (3.33), confirming b).

Concerning c), we have seen above that under the assumption (3.31) the estimates (3.39) hold. Thus, using the rough estimate  $a \leq 1$  one obtains

$$\frac{2\Gamma K C_{\mathbf{C}}}{\gamma} (1 + \|\boldsymbol{\delta}_n^\perp\|) \|\boldsymbol{\delta}_n^\perp\| \leq \frac{1}{2} \|\mathbf{M}_\perp^{-1}\|^{-1}, \quad n \geq n_0, \quad (3.40)$$

where

$$n_0 := \left\lceil \frac{\log(\gamma/(8\Gamma K C_{\mathbf{C}} \|\mathbf{M}_\perp^{-1}\|))}{\log \xi} \right\rceil.$$

Hence, for  $n \geq n_0$ , the hypotheses of Lemma 6 are satisfied for  $\mathbf{x} = \mathbf{x}_n$ , which provides the lower estimate of (3.34) again with  $M = 2\|\mathbf{M}_\perp^{-1}\|$  as in Lemma 6. The upper estimate is also an immediate consequence of Lemma 6 and (3.39).

As for the remaining claim, we find that

$$\begin{aligned} \|\mathbf{r}_{n+1}\| &= \|(\mathbf{A} - \lambda_{n+1}\mathbf{C})\mathbf{x}_{n+1}\| \\ &= \|\widehat{\mathbf{x}}_{n+1}\|^{-1} \|(\mathbf{A} - \lambda^{(n+1)}\mathbf{C})(\mathbf{x}_n - \alpha\mathbf{r}_n)\| \\ &= \|\widehat{\mathbf{x}}_{n+1}\|^{-1} \|(\mathbf{A} - \lambda^{(n)}\mathbf{C})\mathbf{x}_n \\ &\quad - \alpha(\mathbf{A} - \underline{\lambda}\mathbf{C})\mathbf{r}_n + (\lambda^{(n)} - \lambda^{(n+1)})\mathbf{C}\mathbf{x}_n - \alpha(\underline{\lambda} - \lambda^{(n+1)})\mathbf{C}\mathbf{r}_n\|. \end{aligned}$$

On account of Lemma 5, (3.19), (3.21), and the fact that  $\|\widehat{\mathbf{x}}_{n+1}\| \geq 1$  for the unnormalized iterates, we obtain further

$$\begin{aligned} \|\mathbf{r}_{n+1}\| &\leq \|(\mathbf{I} - \alpha(\mathbf{A} - \underline{\lambda}\mathbf{C}))\| \|\mathbf{r}_n\| + \frac{2C_{\mathbf{C}}\Gamma K}{\gamma} \|\boldsymbol{\delta}_n^\perp\|^2 (2\|\mathbf{x}_n\| + \alpha\|\mathbf{r}_n\|) \\ &\leq \left( \beta + \frac{M^2 2C_{\mathbf{C}}\Gamma K}{\gamma} (2 + \alpha\|\mathbf{r}_n\|) \|\mathbf{r}_n\| \right) \|\mathbf{r}_n\|, \end{aligned}$$

where  $M$  is from (3.24) and where we have also used Lemma 4, the continuity of  $\mathbf{C}$  and the estimate (3.24) just proven before. Recall that we have  $\beta < 1$ , thus showing the assertion for  $\|\mathbf{r}_n\|$  small enough. The fact that  $\|\mathbf{r}_n\|$  indeed becomes small follows from the upper estimate in (3.24) and (3.32) as well.  $\square$

## 4 Adaptive strategies for the eigenvalue problem

There are two points that we would like to stress from the start. First, all the above considerations are independent of the index set  $\mathcal{I}$  underlying the space  $\ell_2(\mathcal{I})$  being finite or infinite as long as our assumptions on  $\mathbf{A}$ ,  $\mathbf{C}$  are satisfied. If the basic Richardson iteration is considered on a fixed finite dimensional space, i.e. if we consider a fixed discretization of the original problem (1.2) like (1.7), the proposed scheme may not be the most favorable one since it offers at best linear error reduction. Nevertheless, the property (2.2) reflects the fact that one has been able to precondition the problem sufficiently well in the sense that a fixed error reduction is achieved *independently* of the (fixed) size of  $\mathcal{I}$ .

On the other hand, if one is interested in solving the original problem (1.2) within a desired target accuracy  $\varepsilon$  at possibly low cost (which means to understand how the cost depends on  $\varepsilon$  when  $\varepsilon$  tends to zero), the game changes completely. Following [CDD2] one may think of performing the iteration on the infinite eigenvalue problem (2.1) which is still *equivalent* to (1.2). In fact, solving (2.1) within target accuracy  $\varepsilon$  provides a solution to (1.2), due to the norm equivalence (2.2) and the mapping property (1.5), that has up to a constant factor the same accuracy in the (continuous) energy norm. Now, of course, the matrices  $\mathbf{A}$ ,  $\mathbf{C}$  are infinite so that even when the current iterate  $\mathbf{x}_n$  has finite support, the next iterate  $\mathbf{x}_{n+1}$  cannot be computed exactly. The idea is therefore to compute each update only approximately within a suitable dynamically varying accuracy tolerance in such a way that, on one hand, the convergence to the exact solution is preserved, while on the other hand, the computational cost of each perturbed iteration is possibly low.

In the following sections we shall carry out this program based on the above Richardson type iteration. It will be seen that, regardless of the order of the iteration, when applied to a fixed finite dimensional trial space, such an adaptive scheme may perform at an asymptotically optimal complexity.

We will proceed in two steps. First, each iteration step requires the approximate calculation of the residual  $(\mathbf{A} - \lambda(\mathbf{x}_n)\mathbf{C})\mathbf{x}_n$  which can be broken down to two types of tasks, namely:

- (I) to approximate matrix/vector products  $\mathbf{A}\mathbf{x}$ ,  $\mathbf{C}\mathbf{x}$  (where now  $\mathbf{A}$ ,  $\mathbf{C}$  are in principle infinite),

(II) to approximate the Rayleigh quotient  $\lambda(\mathbf{x}) = \frac{\langle \mathbf{A}\mathbf{x}, \mathbf{x} \rangle}{\langle \mathbf{C}\mathbf{x}, \mathbf{x} \rangle}$ .

Of course, (II) can be reduced to (I) but there are interesting alternatives. We distinguish the two cases at this point because, depending on the situation, the accuracy requirements may be somewhat different due to the fact that relative accuracy tolerances are needed in (II) unless  $\mathbf{C}$  has particularly favorable properties that essentially permit the exact calculation of  $\langle \mathbf{C}\mathbf{x}, \mathbf{x} \rangle$ . We shall return to this issue later in more detail.

Let us therefore suppose at this point that we have for  $\mathbf{B} \in \{\mathbf{A}, \mathbf{C}\}$  a routine with the following property at hand:

$APPLY(\mathbf{B}, \mathbf{x}, \eta) \rightarrow \mathbf{w}$  such that

$$\|\mathbf{B}\mathbf{x} - \mathbf{w}\| \leq \eta.$$

We shall postpone the actual description of  $APPLY$  to a later section. Given such a routine we shall first analyze which tolerances  $\eta$  are needed to ensure convergence of a correspondingly perturbed iteration. As a second step we shall then discuss the complexity of a corresponding numerical realization.

## 4.1 A perturbed iteration

We shall now collect the main ingredients of a perturbed version of *MINIT* from Section 3.1, assuming that the routine  $APPLY$  from above is available. The first one concerns the approximation of a given  $\mathbf{x} \in \ell_2(\mathcal{I})$  by one with possibly short support.

$APPROX(\mathbf{x}, \eta) \rightarrow \mathbf{z}$ : produces for a given (finitely supported)  $\mathbf{x}$  and any  $\eta > 0$  a finitely supported  $\mathbf{z}$  such that

$$\|\mathbf{x} - \mathbf{z}\| \leq \eta, \quad \#\text{supp } \mathbf{z} \leq \operatorname{argmin}_{\|\mathbf{w} - \mathbf{x}\| \leq \eta/2} \#\text{supp } \mathbf{w}. \quad (4.1)$$

This routine can be realized by replacing as many entries of the input as possible by zero, as long as the sum of their squares does not exceed  $\eta^2$ . One could of course use the tolerance  $\eta$  in the argmin which would mean to compute a best  $N$ -term approximation of  $\mathbf{x}$ . This in turn would require exact sorting of the coefficients introducing an additional log-factor of the support size of  $\mathbf{x}$ , see [CDD1] for details. Being content with quasi-sorting based on binary binning one can avoid the log-factor at the expense of a slightly larger support, see e.g. [B].

In addition to approximate matrix/vector products we need approximate Rayleigh quotients and hence scalar products. Recall that we have to perform such routines for normalized inputs  $\|\mathbf{x}\| = 1$ , which will be henceforth assumed. A straightforward way to

approximate the scalar products would be as follows:

$SCAL(\mathbf{B}, \mathbf{x}, \eta) \rightarrow s$ : Given any  $\eta > 0$ , the routine outputs a scalar  $s$  such that

$$|\langle \mathbf{B}\mathbf{x}, \mathbf{x} \rangle - s| \leq \eta \quad (4.2)$$

as follows:

- $APPROX(\mathbf{x}, \eta/(2\|\mathbf{B}\|)) \rightarrow \mathbf{z}$ ;
- $APPLY(\mathbf{B}, \mathbf{x}, \eta/2)|_{\text{supp } \mathbf{z}} \rightarrow \mathbf{w}$ ;
- $s = \langle \mathbf{z}, \mathbf{w} \rangle$ .

Here  $APPLY(\mathbf{B}, \mathbf{x}, \eta/2)|_{\text{supp } \mathbf{z}}$  means that the output of  $APPLY$  can be restricted to  $\text{supp } \mathbf{z}$ . To which extent this can be used to economize on the computational effort depends on the realization of  $APPLY$  and will therefore be postponed to Section 4.3.

It is easily seen that  $s$  given as above does satisfy (4.2). Note that  $\|\mathbf{A}\| := \sup_{\|\mathbf{x}\|=1} \|\mathbf{A}\mathbf{x}\| \leq \Gamma$ .

We also remark that one can think of different ways of approximating  $\langle \mathbf{B}\mathbf{x}, \mathbf{x} \rangle$ . For instance, one could decompose  $\mathbf{x}$  by determining  $\mathbf{x}_j$  of smallest support so that  $\|\mathbf{x} - \mathbf{x}_j\| \leq 2^j \eta^{1/2}$ , say, noting that  $\mathbf{x}_J = 0$  for  $J := \lceil \log_2 \eta^{1/2} \rceil$ . Then, setting  $\mathbf{z}_j := \mathbf{x}_j - \mathbf{x}_{j+1}$

$$\sum_{j=0}^J \langle \mathbf{z}_j, \mathbf{w}_j \rangle, \quad \mathbf{w}_j := APPLY(\mathbf{B}, \mathbf{x}, \epsilon_j)|_{\text{supp } \mathbf{z}_j}, \quad j = 0 \dots, J, \quad (4.3)$$

is an approximation to  $\langle \mathbf{x}, \mathbf{B}\mathbf{x} \rangle$  of the order  $\eta$ , provided that  $\sum_{j=0}^J 2^j \epsilon_j \sim \eta^{1/2}$ . The apparent advantage is that highly accurate matrix applications need to be computed only on typically small supports of coarse  $\mathbf{z}_j$ 's. It will be shown in Section 4.3 that such a strategy does offer asymptotic savings. Of course, as long as the matrix/vector products  $\mathbf{A}\mathbf{x}$ ,  $\mathbf{C}\mathbf{x}$  have to be approximated within a similar accuracy tolerance anyway in the iteration, one might as well stick with the simpler version of  $SCAL$  shown above which we will do for the time being.

Given the routine  $SCAL$  we can proceed to approximating the Rayleigh quotients  $\lambda(\mathbf{x})$ . We shall devise a routine:

$RAYL(\mathbf{x}, \eta) \rightarrow \bar{\lambda}$ : Given  $\eta > 0$  and any finitely supported  $\mathbf{x}$  with  $\|\mathbf{x}\| = 1$ ,  $RAYL$  outputs  $\bar{\lambda}$  such that

$$|\lambda(\mathbf{x}) - \bar{\lambda}| \leq \eta. \quad (4.4)$$

In many cases it is even possible to apply  $\mathbf{C}$  exactly at acceptable cost, e.g. when  $\mathbf{C}$  is diagonal, see e.g. Remark 1. In this case we need no further assumption on  $\mathbf{C}$  beyond boundedness in  $\ell_2(\mathcal{I})$  and can simply take

$$RAYL(\mathbf{x}, \eta) = \frac{SCAL(\mathbf{A}, \mathbf{x}, \eta \langle \mathbf{C}\mathbf{x}, \mathbf{x} \rangle)}{\langle \mathbf{C}\mathbf{x}, \mathbf{x} \rangle}. \quad (4.5)$$

If, on the other hand,  $\mathbf{C}$  can only be applied approximately with the aid of some routine *APPLY*, the routine *SCAL* is slightly more involved. Denoting by  $c_C := \gamma/K$  the local coercivity constant for  $\mathbf{C}$  from Remark 3, the following realization works.

**Lemma 7.** *Assume that  $\|(\mathbf{u} - \mathbf{x})^\perp\| \leq \sqrt{a}$  holds. Carrying out the following steps:*

- $SCAL(\mathbf{A}, \mathbf{x}, \eta c_C/2) \rightarrow \tilde{s}_A$ ,  $SCAL(\mathbf{C}, \mathbf{x}, c_C^2 \eta/(6\Gamma)) \rightarrow \tilde{s}_C$ ;
- $\bar{\lambda} := \tilde{s}_A/\tilde{s}_C$ ,

verifies (4.4), provided that

$$\eta \leq \min \{ \gamma/c_C, 3\Gamma/c_C \}. \quad (4.6)$$

**Proof:** To see the validity of (4.4), let  $s_A := \langle \mathbf{A}\mathbf{x}, \mathbf{x} \rangle$ ,  $s_C := \langle \mathbf{C}\mathbf{x}, \mathbf{x} \rangle$  and note first that

$$\frac{s_A}{s_C} - \frac{\tilde{s}_A}{\tilde{s}_C} = \frac{1}{s_C}(s_A - \tilde{s}_A) - \frac{\tilde{s}_A}{s_C \tilde{s}_C}(s_C - \tilde{s}_C),$$

so that

$$|\lambda(\mathbf{x}) - \bar{\lambda}| \leq \frac{\eta c_C}{2s_C} + \frac{\tilde{s}_A}{s_C \tilde{s}_C} \frac{c_C^2 \eta}{6\Gamma}. \quad (4.7)$$

Now note that under the assumption (4.6) one has (recalling that  $\|\mathbf{x}\| = 1$ )

$$|s_A - \tilde{s}_A| \leq \frac{\eta c_C}{2} = \frac{\eta c_C \|\mathbf{x}\|^2}{2} \leq \frac{\eta c_C s_A}{2\gamma} \leq \frac{1}{2} s_A,$$

so that  $\tilde{s}_A \leq \frac{3s_A}{2} \leq \frac{3\Gamma}{2}$ . Likewise, upon using (3.22), one obtains for  $\eta$  satisfying (4.6) that  $|s_C - \tilde{s}_C| \leq s_C/2$  so that  $\tilde{s}_C \geq s_C/2 \geq c_C/2$ . Combining these estimates yields

$$\frac{\tilde{s}_A}{s_C \tilde{s}_C} \leq \frac{6\Gamma}{2c_C^2},$$

which, on account of (4.7), confirms the claim (4.4).  $\square$

**Remark 5.** *On account of Remark 3, the tolerance in the evaluation of the scalar product remains proportional to the target accuracy  $\eta$  in the routine *RAYL* for either version provided that  $\|(\mathbf{u} - \mathbf{x})^\perp\| \leq \sqrt{a}$ .*

We are now able to describe a routine for approximating the scaled residual  $\alpha(\mathbf{A} - \lambda(\mathbf{x})\mathbf{C})\mathbf{x}$ :

$RES(\mathbf{x}, \eta) \rightarrow \mathbf{r}_\eta$ : Given any  $\eta > 0$  and any  $\mathbf{x}$  with  $\|\mathbf{x}\| = 1$ , the routine  $RES$  outputs a finitely supported vector  $\mathbf{r}_\eta$  such that

$$\|\mathbf{r}_\eta - \alpha(\mathbf{A} - \lambda(\mathbf{x})\mathbf{C})\mathbf{x}\| \leq \eta. \quad (4.8)$$

The routine  $RES$  can be realized as follows:

- 1)  $RAYL(\mathbf{x}, \eta/(4C_C\alpha)) \rightarrow \bar{\lambda}$ ;
- 2)  $RES(\mathbf{x}, \eta) = \alpha(APPLY(\mathbf{A}, \mathbf{x}, \eta/(2\alpha)) - \bar{\lambda}APPLY(\mathbf{C}, \mathbf{x}, \eta/(4\bar{\lambda}\alpha)))$

Of course, when the approximate evaluation of the scalar product involving  $\mathbf{A}$  simply relies on the application of  $APPLY$  one only needs to carry out this latter routine once in  $RES$  with respect to the minimal tolerance required in  $RAYL$  and in step 2) of the above scheme.

**Remark 6.** From now on we shall always assume that the parameter  $\alpha$  in the basic Richardson iteration is chosen according to Lemma 4, ensuring that  $\beta < 1$ .

The following lemma will help to understand the perturbed Richardson iteration.

**Lemma 8.** Assume that the approximation  $\bar{\mathbf{u}}$  to  $\mathbf{u}$  satisfies  $\|\delta(\bar{\mathbf{u}})^\perp\| \leq \bar{\varepsilon} \leq d$ , where  $d$  is the constant from (3.31), and let  $\xi$  be given by (3.38). Then, setting  $\eta_j := (1 - \xi)\bar{\varepsilon}2^{-j}$ ,  $\mathbf{v}_0 := \bar{\mathbf{u}}$ , and doing for  $j = 0, 1, 2, 3, \dots$

$$\mathbf{v}_j - RES(\mathbf{v}_j, \eta_j) \rightarrow \hat{\mathbf{v}}_{j+1}, \quad (4.9)$$

one has

$$\|\delta(\mathbf{v}_j)^\perp\| \leq \bar{\varepsilon}, \quad j = 0, 1, 2, \dots \quad (4.10)$$

and

$$\|\delta(\mathbf{v}_j)^\perp\| \leq \xi^j \bar{\varepsilon} / \beta, \quad j = 0, 1, 2, \dots \quad (4.11)$$

**Proof:** Setting  $\mathbf{r}_\eta := RES(\mathbf{v}_j, \eta)$ , we have, by definition,

$$\begin{aligned} \hat{\mathbf{v}}_{j+1} - \mathbf{u} &= \mathbf{v}_j - \mathbf{u} - \alpha\mathbf{r}(\mathbf{v}_j) + (\alpha\mathbf{r}(\mathbf{v}_j) - \mathbf{r}_{\eta_j}) \\ &= \mathbf{v}_j - \mathbf{u} - \alpha(\mathbf{A}(\mathbf{v}_j - \mathbf{u}) - \underline{\lambda}\mathbf{C}(\mathbf{v}_j - \mathbf{u})) + \alpha(\lambda(\mathbf{v}_j) - \underline{\lambda})\mathbf{C}\mathbf{v}_j \\ &\quad + (\alpha\mathbf{r}(\mathbf{v}_j) - \mathbf{r}_{\eta_j}) \\ &= (\mathbf{I} - \alpha\mathbf{M})(\mathbf{v}_j - \mathbf{u}) + \alpha(\lambda(\mathbf{v}_j) - \underline{\lambda})\mathbf{C}\mathbf{v}_j + (\alpha\mathbf{r}(\mathbf{v}_j) - \mathbf{r}_{\eta_j}). \end{aligned} \quad (4.12)$$

Therefore we obtain, on account of Lemma 4, Lemma 5, and the accuracy of  $\mathbf{r}_{\eta_j}$

$$\begin{aligned} \|\boldsymbol{\delta}(\mathbf{v}_{j+1})^\perp\| &\leq \beta\|\boldsymbol{\delta}(\mathbf{v}_j)^\perp\| + \frac{2\alpha\|\mathbf{C}\|\Gamma}{\gamma}\lambda(\mathbf{v}_j)\|\boldsymbol{\delta}(\mathbf{v}_j)^\perp\|^2 + \eta_j \\ &= \left(\beta + \frac{2\alpha\|\mathbf{C}\|\Gamma}{\gamma}\lambda(\mathbf{v}_j)\|\boldsymbol{\delta}(\mathbf{v}_j)^\perp\|\right)\|\boldsymbol{\delta}(\mathbf{v}_j)^\perp\| + \eta_j. \end{aligned} \quad (4.13)$$

Now recall from Remark 3 that  $\lambda(\mathbf{v}_j) \leq K$  as long as  $\|\boldsymbol{\delta}(\mathbf{v}_j)^\perp\| \leq \sqrt{a} \leq d$  (which is, in particular, valid by assumption at the initialization step  $\mathbf{v}_0 = \bar{\mathbf{u}}$ ).

From (3.39) we then know that

$$\|\boldsymbol{\delta}(\mathbf{v}_1)^\perp\| \leq \left(\beta + \frac{\alpha\|\mathbf{C}\|\Gamma}{\gamma}K\|\boldsymbol{\delta}(\bar{\mathbf{u}})^\perp\|\right)\|\boldsymbol{\delta}(\bar{\mathbf{u}})^\perp\| + (1 - \xi)\bar{\varepsilon} \leq \xi\|\boldsymbol{\delta}(\mathbf{v}_0)^\perp\| + (1 - \xi)\bar{\varepsilon} \leq \bar{\varepsilon},$$

where  $\xi < 1$  is given by (3.38). Hence, we conclude that  $\|\boldsymbol{\delta}(\mathbf{v}_1)^\perp\| \leq \bar{\varepsilon}$ . One easily concludes inductively that (4.10) holds (actually with strict inequality for  $j \geq 1$ ).

More precisely, repeating the reasoning in (4.13), we obtain upon elementary calculations and using that  $\xi = (1 + \beta)/2$ ,

$$\|\boldsymbol{\delta}(\mathbf{v}_{j+1})^\perp\| \leq \xi^{j+1}\bar{\varepsilon} + \sum_{i=0}^j \xi^{j-i}\eta_i \leq \bar{\varepsilon}\xi^{j+1}/\beta, \quad (4.14)$$

which was to be shown.  $\square$

Now recall from Lemma 6 that, once the orthogonal error component drops below another possibly smaller threshold

$$\bar{\delta} := \frac{1}{2}\left\{\left(1 + \frac{\gamma}{\Gamma K\|\mathbf{C}\|\|\mathbf{M}_\perp^{-1}\|}\right)^{1/2} - 1\right\}, \quad (4.15)$$

the orthogonal error component of the approximate eigenvectors behaves essentially as the residual.

**Remark 7.** *If  $\bar{\delta} < d$  at most*

$$m_0 = \left\lceil \frac{\log\left(\frac{\beta\bar{\delta}}{\bar{\varepsilon}}\right)}{\log \xi} \right\rceil$$

*iterations of the form (4.9) suffice to provide an approximation  $\mathbf{x}_{m_0}$  to  $\mathbf{u}$  satisfying*

$$\|\boldsymbol{\delta}(\mathbf{x}_{m_0})^\perp\| \leq \min\{d, \bar{\delta}\}. \quad (4.16)$$

We are now ready to formulate the main adaptive eigensolver. On account of Remark 7 we shall assume for simplicity without loss of generality that the initial guess already satisfies the somewhat more stringent accuracy tolerance (4.16).

(i) **Initialization:** Choose  $\varepsilon_0 \leq \min\{d, \bar{\delta}\}$  (see (3.31), (4.15)) and  $\|\mathbf{x}_0\| = 1$  s.t.  $\|\boldsymbol{\delta}(\mathbf{x}_0)^\perp\| \leq \varepsilon_0$ ;

set  $\bar{\varepsilon} := \varepsilon_0$ ,  $\bar{\mathbf{u}} := \mathbf{x}_0$ ;

(ii) **Iteration block:** set  $\bar{\mathbf{u}} \rightarrow \mathbf{v}$ ,  $(1 - \xi)\bar{\varepsilon} \rightarrow \eta$ ,  $j = 0$ ;

do

$RES(\mathbf{v}, \eta) \rightarrow \mathbf{r}_\eta$ ;  $\mathbf{v} - \mathbf{r}_\eta \rightarrow \hat{\mathbf{v}}$ ;  $\hat{\mathbf{v}}/\|\hat{\mathbf{v}}\| \rightarrow \mathbf{v}$ ;  $\eta/2 \rightarrow \eta$ ,  $j + 1 \rightarrow j$ ;

until  $j \geq \log(c_1\beta)/\log\xi$  or  $\eta + \|\mathbf{r}_\eta\| \leq c_1\bar{\varepsilon}/M$  for a fixed  $c_1 \leq \min\{2/(5\varepsilon_0), (\sqrt{3} + 5/2)^{-1}2^{-3/2}\}$ ;

(iii) **Coarsening:** do  $\mathbf{w} = APPROX(\mathbf{v}, 3c_1\bar{\varepsilon}/\sqrt{2})$ ;

normalize  $\bar{\mathbf{u}} = \frac{\mathbf{w}}{\|\mathbf{w}\|}$ ; if  $\bar{\varepsilon}/2 \leq \varepsilon$ , stop and set  $\lambda(\varepsilon) = RES(\bar{\mathbf{u}}, \varepsilon)$ ,  $\mathbf{u}(\varepsilon) := \bar{\mathbf{u}}$ ;

else  $\bar{\varepsilon}/2 \rightarrow \bar{\varepsilon}$ ; go to (ii).

---

Note that a block of perturbed iterations (ii) is interrupted by a coarsening step (iii) as soon as a threshold criterion is met. This criterion involves actually two alternatives that will both be seen to be met after a uniformly bounded finite number of steps in (ii). One of the stopping tests is an *a-posteriori* test based on the numerical residual and the rationale is that it may actually be met earlier than the other test which is based on the bounds in (4.11) which might be too pessimistic, in particular at later stages when the reduction constants decrease. The role of the coarsening step is to control the complexity of the overall scheme in a similar way as in adaptive schemes for operator equations (see e.g. [CDD2, CDD3]).

**Theorem 2.** *The scheme **MINIEIG**( $\mathbf{A}, \mathbf{C}, \varepsilon$ ) terminates for any given  $\varepsilon > 0$  after finitely many steps and outputs an approximate eigenpair  $(\lambda(\varepsilon), \mathbf{u}(\varepsilon))$  with  $\mathbf{u}$  normalized satisfying*

$$\|(\mathbf{u} - \mathbf{u}(\varepsilon))^\perp\| \leq \varepsilon, \quad |\lambda(\varepsilon) - \underline{\lambda}| \leq \varepsilon, \quad (4.17)$$

where  $(\underline{\lambda}, \mathbf{u})$  is an exact ground state eigenpair.

**Remark 8.** *Given  $\mathbf{u}(\varepsilon)$  satisfying the first estimate in (4.17) we can approximate  $\underline{\lambda}$  within a tolerance proportional to  $\varepsilon^2$  instead of (4.17) by applying  $RAYL(\mathbf{u}(\varepsilon), \varepsilon^2)$  which of course requires a correspondingly higher cost due to the the required more accurate applications of  $SCAL$  and ultimately of  $APPLY$ , see Section 4.3.*

**Proof of Theorem 2:** To analyze the effect of the various perturbations of the exact iteration suppose that  $\mathbf{u}_k := \bar{\mathbf{u}}$  is the output of the coarsening step after the  $k$ -th cycle through the coarsening step (iii), i.e. at this stage we have  $\bar{\varepsilon} = 2^{-k}\varepsilon_0$ . Moreover, let  $\mathbf{v}_j$  be the result after  $j$  perturbed iterations in the iteration block (ii) starting with  $\mathbf{v}_0 = \bar{\mathbf{u}} = \mathbf{u}_k$ . The corresponding tolerance  $\eta = \eta_j$  is then given by  $\eta = (1 - \xi)\bar{\varepsilon}2^{-j} = (1 - \xi)\varepsilon_02^{-k-j}$ . Invoking Lemma 8, ensures that  $\|\boldsymbol{\delta}(\mathbf{v}_j)^\perp\| \leq \bar{\varepsilon}$  and  $\|\boldsymbol{\delta}(\mathbf{v}_j)^\perp\| \leq \bar{\varepsilon}\xi^j/\beta$  and hence that the error reduces at least by a fixed rate. Hence, the stopping criterion in (ii) is met after at most  $J := \lceil \log(c_1\beta)/\log\xi \rceil$  steps giving

$$\|\boldsymbol{\delta}(\mathbf{v}_J)^\perp\| \leq c_1\bar{\varepsilon}. \quad (4.18)$$

Moreover, the initialization ensures that Lemma 6 applies to the iterates  $\mathbf{v}_j$ , which means that  $\|\boldsymbol{\delta}(\mathbf{v}_j)^\perp\|$  is controlled by the residuals. Thus

$$\|\boldsymbol{\delta}(\mathbf{v}_j)^\perp\| \leq M \|\mathbf{r}(\mathbf{v}_j)\| \leq M(\|\mathbf{r}_{\eta_j}\| + \eta_j). \quad (4.19)$$

On the other hand, also by Lemma 6,

$$\|\mathbf{r}_{\eta_j}\| \leq \|\mathbf{r}(\mathbf{v}_j)\| + \eta_j \leq M' \|\boldsymbol{\delta}(\mathbf{v}_j)^\perp\| + \eta_j \leq M' \bar{\varepsilon} \xi^j / \beta + \eta_j.$$

Thus  $\|\mathbf{r}_{\eta_j}\| + \eta_j$  will also drop below the threshold  $c_1 \bar{\varepsilon} / M$  after finitely many steps, and in fact, possibly earlier than the alternative step bound  $J$ , if the bounds in (4.11) are overly pessimistic.

In summary, the input  $\mathbf{v}$  in (iii) at this stage satisfies  $\|(\mathbf{u} - \mathbf{v})^\perp\| \leq c_1 \bar{\varepsilon}$ . Assuming without loss of generality that  $\langle \mathbf{u}, \mathbf{x} \rangle > 0$ , we can invoke Lemma 2 to conclude that

$$\|\mathbf{u}^\circ - \mathbf{v}\| \leq \sqrt{2} c_1 \bar{\varepsilon} =: \sigma, \quad (4.20)$$

where  $\mathbf{u}^\circ := \mathbf{u} / \|\mathbf{u}\|$ . Since by (iii), we have  $\mathbf{w} = \text{APPROX}(\mathbf{v}, 3\sigma/2)$ , we obtain by triangle inequality  $\|\mathbf{u}^\circ - \mathbf{w}\| \leq 5\sigma/2$ . Moreover, by definition of the routine *APPROX*, we have  $\langle \mathbf{w}, \mathbf{v} - \mathbf{w} \rangle = 0$  so that  $\|\mathbf{w}\| \geq (1 - (3\sigma/2)^2)^{1/2}$ . Thus, the normalized vector  $\bar{\mathbf{u}} := \mathbf{w} / \|\mathbf{w}\|$  satisfies with  $(1 - (3\sigma/2)^2)^{-1/2} =: 1 + g$

$$\begin{aligned} \|(\mathbf{u} - \bar{\mathbf{u}})^\perp\| &\leq \|\mathbf{u}^\circ - \bar{\mathbf{u}}\| \leq \|\mathbf{u}^\circ - \mathbf{w}\| + \|\mathbf{w} - \bar{\mathbf{u}}\| \\ &\leq \frac{5\sigma}{2} + \|\mathbf{w}\| \left( \frac{1}{\|\mathbf{w}\|} - 1 \right) \leq \frac{5\sigma}{2} + \|\mathbf{w}\| \left( \frac{1}{(1 - (3\sigma/2)^2)^{1/2}} - 1 \right) \\ &\leq \frac{5\sigma}{2} + g. \end{aligned}$$

Noting that for  $b \leq 1/2$  one has  $(1 - b^2)^{-1/2} \leq 1 + 2b/\sqrt{3}$ , we conclude that  $g \leq \sqrt{3}\sigma$ , so that

$$\|(\mathbf{u} - \bar{\mathbf{u}})^\perp\| \leq (\sqrt{3} + 5/2)\sigma = \sqrt{2} c_1 (\sqrt{3} + 5/2) \bar{\varepsilon} \leq \bar{\varepsilon} / 2, \quad (4.21)$$

by our choice of the constant  $c_1$ . Thus, in summary we have shown that after a uniformly bounded finite number of perturbed iterations in (ii) with initial accuracy  $\bar{\varepsilon}$ , one branches into (iii) whose output is either sufficiently accurate or serves as input for (ii) with improved accuracy  $\bar{\varepsilon}/2$ . Hence the algorithm terminates after finitely many cycles through (ii), (iii), namely as soon as  $2^{-k} \varepsilon_0 \leq \varepsilon$ .  $\square$

## 4.2 Complexity estimates

It remains to analyze the computational complexity of the above scheme *MINIEIG*. The subsequent analysis follows similar lines as used before in connection with adaptive solvers

for operator equations. We formulate an ideal benchmark which describes the minimal cost needed to achieve a desired accuracy tolerance  $\varepsilon$  for an approximate normalized eigenvector. This lower bound is simply the number of entries needed in any finitely supported sequence to approximate  $\mathbf{u}$  within accuracy  $\varepsilon$  (or equivalently the normalized  $\mathbf{u}^\circ$  within a fixed factor). This naturally leads to the notion of best  $N$ -term approximation that we briefly recall first.

Let  $\Sigma_k$  denote the set of all  $\mathbf{x} \in \ell_2(\mathcal{I})$  which have at most  $k$  nonvanishing entries. Then

$$\sigma_N(\mathbf{x}) := \inf_{\mathbf{z} \in \Sigma_N} \|\mathbf{x} - \mathbf{z}\|$$

is the error of best  $N$ -term approximation in  $\ell_2(\mathcal{I})$ . Then

$$|\mathbf{x}|_{\mathcal{A}^s} := \sup_{N \in \mathbb{N}} N^s \sigma_N(\mathbf{x})$$

is a (quasi-)seminorm and

$$\mathcal{A}^s := \{\mathbf{x} \in \ell_2(\mathcal{I}) : \|\mathbf{x}\|_{\mathcal{A}^s} := \|\mathbf{x}\| + |\mathbf{x}|_{\mathcal{A}^s} < \infty\}$$

is a (quasi-)Banach space. Thus, for  $s > 0$ , the unit ball of  $\mathcal{A}^s$  is the set of all those sequences in  $\ell_2(\mathcal{I})$  whose error of best  $N$ -term approximation decays at least as  $N^{-s}$  and hence a compact set in  $\ell_2(\mathcal{I})$ . Another way to view this is the following: In order to approximate a given  $\mathbf{x} \in \mathcal{A}^s$  with accuracy  $\varepsilon$  it takes at most  $N_\varepsilon = \varepsilon^{-1/s} |\mathbf{x}|_{\mathcal{A}^s}^{1/s}$  entries to do so, and in the worst case over the whole unit ball of  $\mathcal{A}^s$  the necessary order of entries is exactly  $\varepsilon^{-1/s}$ . In what follows this relation

$$\text{accuracy } \varepsilon \quad \leftrightarrow \quad \text{degrees of freedom } \varepsilon^{-1/s} \cdot |\mathbf{x}|_{\mathcal{A}^s}$$

reflecting  $s$ -sparsity of elements in  $\ell_2(\mathcal{I})$  will be a central orientation in the subsequent developments.

Let us pause to mention that  $\mathcal{A}^s$  can also be characterized as a *weak*  $\ell_p$  space. In fact, denote by  $\mathbf{x}^*$  the nonincreasing rearrangement of  $\mathbf{x} \in \ell_2(\mathcal{I})$ , i.e.  $x_{i+1}^* = |x_{j_{i+1}}| \leq x_i^* = |x_{j_i}|$ ,  $j = 1, 2, \dots$ . Then  $\ell_p^w(\mathcal{I})$  is comprized of all those  $\mathbf{x} \in \ell_2(\mathcal{I})$  for which

$$|\mathbf{x}|_{\ell_p^w(\mathcal{I})} := \sup_{n \geq 1} n^{1/p} x_n^* < \infty.$$

It is not hard to show that (see [DeVore])

$$\mathcal{A}^s = \ell_p^w, \quad \frac{1}{p} = s + \frac{1}{2}.$$

Moreover, it is easy to see that  $\ell_p \subset \ell_p^w$  but for any  $q < p$  one has  $\ell_p^w \subset \ell_q$ , so that  $s$ -sparse sequences are almost just  $\ell_p$  summable sequences with  $s$  and  $p$  related as above.

The first key ingredient of the analysis is the following *coarsening lemma* that explains, in particular, the role of step (iii) in **MINIEIG**, see [C, CDD3].

**Lemma 9.** Assume that  $\mathbf{v} \in \mathcal{A}^s$  for some  $s > 0$  and suppose that  $\mathbf{x} \in \ell_2(\mathcal{I})$  is any finitely supported sequence in  $\ell_2(\mathcal{I})$  such that  $\|\mathbf{v} - \mathbf{x}\| \leq \varepsilon$ . Moreover fix  $b > 0$  and set

$$\mathbf{w} := \operatorname{argmin}_{\|\mathbf{x}-\mathbf{z}\| \leq (1+b)\varepsilon} \#\operatorname{supp} \mathbf{z}. \quad (4.22)$$

Then there exists a constant  $C$  depending only on  $b$  and  $s$  when  $s \rightarrow 0$  such that

$$\|\mathbf{w}\|_{\mathcal{A}^s} \leq C\|\mathbf{v}\|_{\mathcal{A}^s}, \quad (4.23)$$

and

$$\#\operatorname{supp} \mathbf{w} \leq C\|\mathbf{v}\|_{\mathcal{A}^s}^{\frac{1}{s}} \varepsilon^{-\frac{1}{s}}, \quad \|\mathbf{v} - \mathbf{w}\| \leq (2+b)\varepsilon \leq \|\mathbf{v}\|_{\mathcal{A}^s} (\#\operatorname{supp} \mathbf{w})^{-s}, \quad (4.24)$$

hold uniformly in  $\varepsilon > 0$ .

In other words, thresholding a given finitely supported approximation at a slightly higher tolerance than the accuracy of approximation provides essentially a best  $N$ -term approximation to the (possibly unknown) approximand.

As mentioned before, strictly speaking the cost of determining  $\mathbf{w}$  is essentially  $(\#\operatorname{supp} \mathbf{x}) \log(\#\operatorname{supp} \mathbf{x})$ . However, at the expense of a slightly worse target accuracy than  $(1+b)\varepsilon$  one can get away with quasi-sorting based on binary binning at a computational cost that stays proportional to  $(\#\operatorname{supp} \mathbf{x})$ . For simplicity, we shall therefore ignore the log-factor in what follows and use that *APPROX* can be realized in linear complexity of the input size.

**Remark 9.** Let us denote again by  $\bar{\mathbf{u}}_k$  the result of the  $k$ -th application of step (iii) in *MINIEIG*. Then, if the ground state  $\mathbf{u}$  belongs to  $\mathcal{A}^s$  for some  $s > 0$ , we have

$$\|\bar{\mathbf{u}}_k\|_{\mathcal{A}^s} \lesssim \|\mathbf{u}\|_{\mathcal{A}^s}, \quad \#\operatorname{supp} \bar{\mathbf{u}}_k \lesssim (\varepsilon_0 2^{-k})^{-1/s} \|\mathbf{u}\|_{\mathcal{A}^s}^{1/s}, \quad \|\mathbf{u}^\circ - \bar{\mathbf{u}}_k\| \leq \varepsilon_0 2^{-k}, \quad (4.25)$$

uniformly in  $k \in \mathbb{N}$ .

**Proof:** The last relation in (4.25) has been already established in the proof of Theorem 2, see (4.21). The rest is then an immediate consequence of Lemma 9.  $\square$

Hence to estimate the computational complexity of *MINIEIG*( $\mathbf{A}, \mathbf{C}, \varepsilon$ ) depending on  $\varepsilon$  as  $\varepsilon \rightarrow 0$ , it remains to bound the computational work in each block (ii).

**Proposition 1.** Assume that for some  $s > 0$  one has  $\mathbf{u} \in \mathcal{A}^s$ , and that  $\mathbf{r}_\eta := \operatorname{RES}(\mathbf{v}, \eta)$  satisfies

$$\begin{aligned} \|\mathbf{r}_\eta\|_{\mathcal{A}^s} &\leq C(\|\mathbf{u}\|_{\mathcal{A}^s} + \|\mathbf{v}\|_{\mathcal{A}^s}), \\ \#\operatorname{supp} \mathbf{r}_\eta, \#\operatorname{flops}(\mathbf{r}_\eta) &\leq C\eta^{-1/s} (\|\mathbf{u}\|_{\mathcal{A}^s}^{1/s} + \|\mathbf{v}\|_{\mathcal{A}^s}^{1/s}), \end{aligned} \quad (4.26)$$

holds for some constant  $C$  independent of  $\eta$ . Here  $\#\text{flops}(\mathbf{r}_\eta)$  denotes the number of arithmetic operations needed to compute  $\mathbf{r}_\eta$ . Then the output  $(\lambda(\varepsilon), \mathbf{u}(\varepsilon))$  of  $\text{MINIEIG}(\mathbf{A}, \mathbf{C}, \varepsilon)$  satisfies in addition to (4.17)

$$\#\text{supp } \mathbf{u}(\varepsilon), \#\text{flops}(\mathbf{u}(\varepsilon)) \leq C\varepsilon^{-1/s} \|\mathbf{u}\|_{\mathcal{A}^s}^{1/s}, \quad \|\mathbf{u}(\varepsilon)\|_{\mathcal{A}^s} \leq C\|\mathbf{u}\|_{\mathcal{A}^s}, \quad (4.27)$$

for some constant  $C$  independent of  $\mathbf{u}$  and  $\varepsilon$ .

**Proof:** Applying if necessary the coarsening lemma to the initial guess (starting from a correspondingly slightly higher initial accuracy) we have  $|\mathbf{x}_0|_{\mathcal{A}^s} \lesssim \|\mathbf{u}\|_{\mathcal{A}^s}$ ,  $\#\text{supp } \mathbf{x}_0 \lesssim \varepsilon_0^{-1/s}$ . Suppose now that at the  $k$ -th stage of (ii) the input  $\bar{\mathbf{u}} = \bar{\mathbf{u}}_{k-1}$  satisfies for some constant  $C$  independent of  $k$

$$\#\text{supp } \bar{\mathbf{u}}, \#\text{flops}(\bar{\mathbf{u}}) \leq C\varepsilon^{-1/s} (\|\mathbf{u}\|_{\mathcal{A}^s}^{1/s} + \|\bar{\mathbf{u}}\|_{\mathcal{A}^s}^{1/s}). \quad (4.28)$$

By assumption (4.26), the same estimates hold for the intermediate iterates  $\mathbf{v}_j$  in (ii) (with  $\mathbf{v}_0 = \bar{\mathbf{u}}_{k-1}$ ) with constants however now depending on  $j$ . As shown in the proof of Theorem 2 each block (ii) has at most a uniformly bounded number  $J$  of iterations, so that (4.28) still holds for the input to step (iii) with some constant  $C = C(J)$ ,  $J$  being the upper bound of the number of iterations in (ii). As pointed out in Remark 9, the application of *APPROX* produces then  $\bar{\mathbf{u}} = \bar{\mathbf{u}}_k$  satisfying again (4.28) for  $\bar{\varepsilon}_k = \bar{\varepsilon}_{k-1}/2$  with a constant coming from the coarsening lemma that is independent of  $k$ . Thus summing the computational cost of each block (ii) and using a straightforward geometric series argument confirms the claim.  $\square$

Thus, it remains to verify (4.26) for the approximate residuals. According to the ingredients of the approximate residuals provided by the routine *RES*, given in the previous section, the key issue here is the efficient approximate application of the matrices  $\mathbf{A}$  and  $\mathbf{C}$  through the routine *APPLY*. It is well known by now that for a wide range of (local and global) operators and suitable wavelet bases the corresponding operator representations in wavelet coordinates are *nearly sparse*. A precise formulation of this property that is fulfilled in many concrete cases, reads as follows, see e.g. [CDD4, DHS].

**$s^*$ -compressibility:** Let  $s^*$  be a positive real.  $\mathcal{B}^{s^*}$  denotes the set of matrices (over  $\mathcal{I} \times \mathcal{I}$ ) with the following properties:  $\mathbf{B} \in \mathcal{B}^{s^*}$  if for every  $k \in \mathbb{N}$ , there is a matrix  $\mathbf{B}_k$  with at most  $2^k$  entries in each row and column satisfying  $\|\mathbf{B} - \mathbf{B}_k\| \leq C2^{-ks^*} \alpha_k$ , where  $(\alpha_k)_{k=0}^\infty$  is an  $\ell_1$ -sequence of positive numbers. Elements in  $\mathcal{B}^{s^*}$  are called  $s^*$ -compressible.

One can show that any  $\mathbf{B} \in \mathcal{B}^{s^*}$  maps the  $\mathcal{A}^s$ -spaces boundedly into themselves for  $s < s^*$ , cf. [CDD1], Section 3. For elements in  $\mathbf{B} \in \mathcal{B}^{s^*}$  a concrete realization of *APPLY*( $\mathbf{B}, \cdot, \cdot$ ) has been developed in [CDD1] whose properties are given in the following lemma.

**Lemma 10.** *Assume that  $\mathbf{B} \in \mathcal{B}^{s^*}$ . Given a tolerance  $\delta > 0$  and a vector  $\mathbf{x}$  with finite support, the algorithm  $APPLY(\mathbf{B}, \mathbf{x}, \delta)$  produces a vector  $\mathbf{w}$  which satisfies*

$$\|\mathbf{B}\mathbf{x} - \mathbf{w}\| \leq \delta.$$

Moreover, for  $s < s^*$  one has:

(i) *The output vector  $\mathbf{w}$  satisfies*

$$\|\mathbf{w}\|_{\mathcal{A}^s} \lesssim \|\mathbf{x}\|_{\mathcal{A}^s}; \quad \text{supp } \mathbf{w} \lesssim \|\mathbf{x}\|_{\mathcal{A}^s}^{\frac{1}{s}} \delta^{-\frac{1}{s}}.$$

(ii) *The number of entries of  $\mathbf{B}$  to be computed to obtain  $\mathbf{w}$  is  $\lesssim \|\mathbf{x}\|_{\mathcal{A}^s}^{\frac{1}{s}} \delta^{-\frac{1}{s}}$ .*

For the *proof*, again see [CDD1].

If for  $s < s^*$  the number of arithmetic operations needed to compute  $\mathbf{w}$  does not exceed  $C\|\mathbf{x}\|_{\mathcal{A}^s}^{\frac{1}{s}} \delta^{-\frac{1}{s}} + 2 \text{supp } \mathbf{x}$  (using quasi-sorting instead of exact sorting which would entail an additional log-factor), the matrix  $\mathbf{B}$  is called  *$s^*$ -computable*, [S, GS]. For the verification of  $s^*$ -computability for a wide class of operators, see [S, GS].

**$s^*$ -sparsity:** *The matrix  $\mathbf{B}$  is called  $s^*$ -sparse if there exists a scheme  $APPLY$  satisfying the properties listed in Lemma 10 whose computational complexity remains proportional to the output size.*

Clearly  $s^*$ -computable matrices are examples of  $s^*$ -sparse matrices. It is important to note though that there are further important examples. Recalling the form of  $\mathcal{L}$  in remark 1, the matrix  $\mathbf{A}$  may actually be the product of several matrices. An  $APPLY$ -scheme for such products can easily be obtained by composing individual  $APPLY$ -schemes designed e.g. for compressible matrices, see [CDD2] for the treatment of least squares formulations.

Another important case concerns the application of  $\mathbf{C}$  in a fairly general setting.

Actually  $\mathbf{C}$  represent the *inverse*  $\mathbf{C} = \mathbf{S}^{-1}$  of an (elliptic) operator  $\mathcal{S}$  in the sense of (1.5). Thus the  $APPLY$ -scheme for  $\mathbf{C}$  may just mean the adaptive approximate solution of an operator equation for an  $\mathcal{H}$ -elliptic operator  $\mathcal{S} : \mathcal{H} \rightarrow \mathcal{H}'$ . In many cases its output has been shown to satisfy the properties in Lemma 10 and thus gives rise to an  $s^*$ -sparse  $APPLY$ -scheme.

In summary, it is important to keep in mind that the actual realizations of  $APPLY$  for  $\mathbf{A}$  and  $\mathbf{C}$  may be completely different but should satisfy the properties of Lemma 10.

**Property 3.** *The matrices  $\mathbf{A}$  and  $\mathbf{C}$  are  $s^*$ -sparse for some  $s^* > 0$  (For  $\mathbf{C}$  this is trivially the case when  $\mathbf{C}$  can be applied exactly in linear time).*

The main result of this paper can now be formulated as follows.

**Theorem 3.** *Assume that  $\mathbf{A}$  and  $\mathbf{C}$  are  $s^*$ -sparse for some  $s^* > 0$  and that the parameters  $\alpha, \beta$  are chosen according to Lemma 4. Then for any  $\varepsilon > 0$ , the scheme **MINIEIG**( $\mathbf{A}, \mathbf{C}, \varepsilon$ ) after finitely many steps produces an approximate eigenpair  $(\lambda(\varepsilon), \mathbf{u}(\varepsilon))$  with normalized  $\mathbf{u}$  satisfying*

$$\|(\mathbf{u} - \mathbf{u}(\varepsilon))^\perp\| \leq \varepsilon, \quad |\underline{\lambda} - \lambda(\varepsilon)| \leq \varepsilon, \quad (4.29)$$

where  $(\underline{\lambda}, \mathbf{u})$  is the exact ground state solution of (2.1).

Moreover, if  $\mathbf{u} \in \mathcal{A}^s$  for some  $s < s^*$ , then one has

$$\#\text{flops } \mathbf{u}(\varepsilon), \#\text{supp } \mathbf{u}(\varepsilon) \lesssim \varepsilon^{-1/s} \|\mathbf{u}\|_{\mathcal{A}^s}^{1/s}, \quad \|\mathbf{u}(\varepsilon)\|_{\mathcal{A}^s} \lesssim \|\mathbf{u}\|_{\mathcal{A}^s}, \quad (4.30)$$

where the constants are independent of  $\varepsilon$  and  $\mathbf{u}$  but depend only on  $s$  when  $s$  approaches  $s^*$ .

**Proof:** (4.29) has been already shown in Theorem 2. To prove the rest of the claim we employ Proposition 1 which requires confirming the property (4.26). Towards this end, recall from Section 4.1 that the realization of  $RES(\mathbf{x}, \eta)$  requires the approximate application of  $\mathbf{A}, \mathbf{C}$  within tolerances that are uniformly bounded from below and above by fixed multiples of  $\eta$  (see (4.8)) as well as the computation of  $RAYL(\mathbf{x}, \eta/(4\alpha))$ . Now, this latter routine requires the evaluation of the routine  $SCAL$  for  $\mathbf{A}$  and, unless  $\mathbf{C}$  can be applied exactly, for  $\mathbf{C}$ . By Remark 5 and Lemma 7, the tolerances needed in the  $SCAL$  routines remain uniformly proportional to the target accuracy in  $RAYL(\mathbf{x}, \eta/(4\alpha))$  which is proportional to  $\eta$ . Since all these routines rely on the scheme  $APPLY$  with respect to tolerances proportional to  $\eta$ , the relations (4.26) follow from Lemma 10. This finishes the proof.  $\square$

### 4.3 Possible quantitative improvements

The scheme analyzed above should be viewed as one possible realization of an adaptive strategy. To achieve quantitative improvements of schemes of the above type one might try to exploit the fact that the (exact) Rayleigh quotient exhibits essentially the square accuracy of the corresponding approximate eigendirection. We shall only sketch such strategies whose details would essentially follow analogous lines as discussed above.

Recall that the evaluation of the scalar products requires the tightest accuracy tolerances (although still proportional to the target accuracy of  $RES$ ). Thus there are two angles that might help to reduce computational complexity, namely:

- a) Trying to speed up the calculation of scalar products;
- b) Reducing the number of calls of  $RAYL$ .

### 4.3.1 Fast evaluation of scalar products

As for a), the naive approach outlined above rests on the accurate calculation of a complete matrix/vector product although one computes at the end only a single number. We have already mentioned a possible approach in (4.3) that might help reducing the computational complexity.

Recalling that an approximate eigendirection of accuracy  $\varepsilon$  would give rise to a (precise) Rayleigh quotient that approximates  $\underline{\lambda}$  within accuracy of the order of  $\varepsilon^2$  (see Lemma 5), a first natural idea is to postprocess the result of  $\mathbf{MINIEIG}(\mathbf{A}, \mathbf{C}, \varepsilon)$ , satisfying (4.29), so as to obtain an approximation to  $\lambda$  of order  $\varepsilon^2$ . In fact, in view of Lemma 5, it would make sense to compute

$$\lambda^*(\varepsilon) = \text{RAYL}(\mathbf{u}(\varepsilon), \varepsilon^2), \quad \text{so that } |\lambda^*(\varepsilon) - \underline{\lambda}| \leq (1 + C)\varepsilon^2, \quad \text{with } C := 2\Gamma K/\gamma,$$

see (3.19) and (3.21). However, using the simple version of the routine *SCAL* described below (4.2), the computational complexity could be of the order of  $\varepsilon^{-2/s}$  when  $\mathbf{u} \in \mathcal{A}^s$ . Let us now point out that this cost can be reduced significantly by employing more refined versions of *SCAL* along the lines of (4.3). To this end, suppose that  $\mathbf{B}$  is an  $s^*$ -compressible matrix and recall e.g. from [CDD1, DHS] the following key idea of constructing an *APPLY* scheme for an approximate matrix/vector computation. Fix any  $\bar{s} < s^*$ . Given  $\zeta > 0$  and  $\mathbf{v}$  of finite support, let  $\mathbf{v}_\zeta := \text{APPROX}(\mathbf{v}, \zeta)$  and set

$$\mathbf{v}_{-1}(\zeta) := \mathbf{v} - \mathbf{v}_\zeta, \quad \mathbf{v}_j(\zeta) := \mathbf{v}_{2^{j\bar{s}}\zeta} - \mathbf{v}_{2^{(j+1)\bar{s}}\zeta}, \quad j = 0, \dots, J(\zeta) := \lceil \log_2(\|\mathbf{v}\|/\zeta)/\bar{s} \rceil.$$

Then, one has

$$\sum_{j=-1}^{J(\zeta)} \mathbf{v}_j(\zeta) = \mathbf{v}, \quad \|\mathbf{v}_{-1}(\zeta)\| \leq \zeta, \quad \|\mathbf{v}_j(\zeta)\| \leq (1 + 2^{\bar{s}})2^{j\bar{s}}\zeta, \quad j = 1, \dots, J(\zeta). \quad (4.31)$$

Moreover, it is shown in [DHS] that (when  $\mathbf{B}_j$  are the compressed versions of  $\mathbf{B} \in \mathcal{B}^{s^*}$  from the definition of  $s^*$ -compressibility)

$$\mathbf{w}_\zeta := \sum_{j=0}^{J(\zeta)} \mathbf{B}_j \mathbf{v}_j(\zeta) \quad \text{satisfies} \quad \|\mathbf{B}\mathbf{v} - \mathbf{w}_\zeta\| \leq C'\zeta, \quad (4.32)$$

for some uniform constant  $C'$  independent of  $\zeta$ . For simplicity we shall work with  $C' = 1$  which can always be arranged through the definition of the  $\mathbf{B}_j$  or by replacing  $\zeta$  by  $c\zeta$  in the decomposition of  $\mathbf{v}$ .

The announced improved version of *SCAL* is based on the following observation.

**Proposition 2.** *Assume that  $\mathbf{B} \in \mathcal{B}^{s^*}$  and  $\bar{s} < s^*$  is fixed. Setting*

$$\epsilon_j := \alpha_j(1 + 2^{\bar{s}})^{-1}2^{-\bar{s}j}\delta^{1/2}, \quad j = 0, \dots, J(\sqrt{\delta}), \quad \epsilon_{-1} := \alpha_{-1}\sqrt{\delta}, \quad \sum_{j=-1}^{\infty} \alpha_j = 1, \quad (4.33)$$

where the summable (to one) coefficients  $\alpha_j$  are again from the definition of  $s^*$ -compressibility and have algebraic decay, and defining

$$s(\delta, \mathbf{v}) := \sum_{j=-1}^{J(\sqrt{\delta})} \langle \mathbf{v}_j(\sqrt{\delta}), \mathbf{w}_{\epsilon_j} \rangle, \quad (4.34)$$

where the  $\mathbf{w}_{\epsilon_j}$  are given by (4.32), one has

$$|s(\delta, \mathbf{v}) - \langle \mathbf{B}\mathbf{v}, \mathbf{v} \rangle| \leq \delta. \quad (4.35)$$

Moreover, whenever  $\text{supp } \mathbf{v} \leq C\delta^{-1/2s} |\mathbf{v}|_{\mathcal{A}^s}^{1/s}$  for some  $s \leq \bar{s} < s^*$ , then the number of operations  $\text{ops}(\mathbf{v}, \delta)$  needed to compute  $s(\delta, \mathbf{v})$  is bounded by a constant multiple of  $\left( |\mathbf{v}|_{\mathcal{A}^s} \delta^{-\frac{1}{2s}} \right)^{\frac{\bar{s}+2s}{s+\bar{s}}}$ , where the constant depends only on  $C$ .

**Proof:** By (4.31) we can write

$$|s(\delta, \mathbf{v}) - \langle \mathbf{B}\mathbf{v}, \mathbf{v} \rangle| = \left| \sum_{j=-1}^{J(\sqrt{\delta})} \langle \mathbf{v}_j(\sqrt{\delta}), \mathbf{w}_{\epsilon_j} - \mathbf{B}\mathbf{v} \rangle \right| \leq \sum_{j=-1}^{J(\sqrt{\delta})} \alpha_j \delta \leq \delta,$$

where we have used Cauchy-Schwarz and (4.33) in the last step, confirming (4.35).

As for the work count, we shall estimate now first the number of operations needed to compute a single entry of  $\mathbf{w}_\zeta$ . Now note that (since  $\mathbf{v} \in \mathcal{A}^s$  for any  $s > 0$ ) for  $\Delta_j := \text{supp } \mathbf{v}_j(\zeta)$  one has

$$\#\Delta_j \lesssim (2^{j\bar{s}}\zeta)^{-1/s} |\mathbf{v}|_{\mathcal{A}^s}^{1/s}. \quad (4.36)$$

Further, recall that each row of  $\mathbf{B}_j$  has at most  $2^j$  entries, so that the computation of one entry of a contribution  $\mathbf{B}_j \mathbf{v}_j(\zeta)$  takes, in view of (4.36), at most  $2^j$  operations as long as  $j \leq j^*$  when  $j^* = j^*(\zeta)$  is the largest integer for which

$$2^{j^*} \leq |\mathbf{v}|_{\mathcal{A}^s}^{1/s} 2^{-j^*\bar{s}/s} \zeta^{-1/s} \iff j^* = \left\lfloor (s + \bar{s})^{-1} \log_2 \left( \frac{|\mathbf{v}|_{\mathcal{A}^s}}{\zeta} \right) \right\rfloor. \quad (4.37)$$

Thus the computation of a single entry of the partial sum  $\sum_{j=0}^{j^*} \mathbf{B}_j \mathbf{v}_j(\zeta)$  takes the order of

$$2^{j^*} \lesssim \zeta^{-1/(s+\bar{s})} |\mathbf{v}|_{\mathcal{A}^s}^{1/(s+\bar{s})}. \quad (4.38)$$

Likewise the number of computations required for a single entry of the remaining sum  $\sum_{j=j^*(\zeta)}^{J(\zeta)} \mathbf{B}_j \mathbf{v}_j(\zeta)$  is, by (4.36), of the order of

$$\zeta^{-1/s} |\mathbf{v}|_{\mathcal{A}^s}^{1/s} \sum_{j=j^*}^J 2^{-j\bar{s}/s} \lesssim \zeta^{-1/s} |\mathbf{v}|_{\mathcal{A}^s}^{1/s} 2^{-j^*\bar{s}/s} \lesssim |\mathbf{v}|_{\mathcal{A}^s}^{1/(s+\bar{s})} \zeta^{-1/(s+\bar{s})}. \quad (4.39)$$

Therefore

$$\#\text{ops}(\text{for computing one entry of } \mathbf{w}_\zeta) \lesssim \zeta^{-1/(s+\bar{s})} |\mathbf{v}|_{\mathcal{A}^s}^{1/(s+\bar{s})}. \quad (4.40)$$

We shall now estimate the work required by the computation of  $\langle \mathbf{v}_{(j)}, \mathbf{w}_{\epsilon_j} \rangle$ . Note first that

$$\#\text{supp } \mathbf{v}_{-1}(\sqrt{\delta}) \lesssim \delta^{-1/2s} |\mathbf{v}|_{\mathcal{A}^s}^{1/s}, \quad \#\text{supp } \mathbf{v}_j(\sqrt{\delta}) \lesssim 2^{-\bar{s}j/s} \delta^{-1/2s} |\mathbf{v}|_{\mathcal{A}^s}^{1/s}.$$

Hence, using (4.40) with  $\zeta = \epsilon_j$ , the computation of  $\langle \mathbf{v}_j(\sqrt{\delta}), \mathbf{w}_{\epsilon_j} \rangle$  in (4.34) takes, in view of (4.34), the order of

$$\#\text{supp } \mathbf{v}_j(\sqrt{\delta}) |\mathbf{v}|_{\mathcal{A}^s}^{1/(s+\bar{s})} \epsilon_j^{-1/(s+\bar{s})} \leq \alpha_j^{-\frac{1}{s+\bar{s}}} (1+2^{\bar{s}})^{\frac{1}{s+\bar{s}}} |\mathbf{v}|_{\mathcal{A}^s}^{\frac{2s+\bar{s}}{s(s+\bar{s})}} 2^{-j\bar{s}^2/(s\bar{s}+\bar{s}^2)} \delta^{-\frac{1}{2s} \left( \frac{\bar{s}+2s}{s+\bar{s}} \right)} \quad (4.41)$$

operations. Summing over  $j$  and recalling that the  $\alpha_j$  decay polynomially completes the proof.  $\square$

**Corollary 1.** *Assume that the hypotheses of Theorem 3 are valid. Given  $\epsilon > 0$ , let  $\mathbf{u}(\epsilon)$  be the output of  $\mathbf{MINIEIG}(\mathbf{A}, \mathbf{C}, \epsilon)$  and let  $\lambda^*(\epsilon) = \text{RAYL}(\mathbf{u}(\epsilon), \epsilon^2)$  where  $\text{RAYL}$  is based on a version of  $\text{SCAL}$  derived in an obvious manner from (4.34). Then one has*

$$|\underline{\lambda} - \lambda^*(\epsilon)| \leq (1 + 2\Gamma K/\gamma)\epsilon^2, \quad (4.42)$$

and, whenever  $\mathbf{u} \in \mathcal{A}^s$  for some  $s \leq \bar{s}$ , the computational complexity  $\#\text{ops}(\lambda^*(\epsilon))$  of  $\lambda^*(\epsilon)$  remains bounded by

$$\#\text{ops}(\lambda^*(\epsilon)) \lesssim |\mathbf{u}|_{\mathcal{A}^s}^{\frac{2s+\bar{s}}{s(s+\bar{s})}} \epsilon^{-\frac{1}{s} \left( \frac{\bar{s}+2s}{s+\bar{s}} \right)}, \quad (4.43)$$

where the constant is independent of  $\mathbf{u}$  and  $s$ .

**Proof:** The estimate (4.42) is an immediate consequence of (3.19) and the first relation in (4.29). The complexity estimate, in turn, follows from Proposition 2 applied to  $\mathbf{v} = \mathbf{u}(\epsilon)$  with  $\delta := \epsilon^2$  together with (4.30).  $\square$

Since  $g(s) := (\bar{s} + 2s)/(\bar{s} + s)$  increases in  $s$  and  $g(\bar{s}) = 3/2$ , the computational complexity of computing  $\lambda^*(\epsilon)$ , and hence the smallest eigenvalue  $\underline{\lambda}$  within a tolerance of order  $\epsilon^2$ , grows at most like  $\epsilon^{-3/2s}$ , which is of course much better than the cost  $\epsilon^{-2/s}$  that would result from applying the original simple version of  $\text{SCAL}$ . In fact, when  $s$  is very small one almost recovers cost of  $\epsilon^{-1/s}$  needed to approximate  $\mathbf{u} \in \mathcal{A}^s$  within tolerance  $\epsilon$ .

Note also that the coarsened versions of  $\mathbf{v}$  needed in the computation of  $\mathbf{w}_{\epsilon_j}$  are essentially the same as those in (4.34), so that they can be reused. Nevertheless, this version of  $\text{RAYL}$  is quantitatively still more involved as the original simple one based on approximating  $\mathbf{Bv}$  at the respective target accuracy. Since these vectors are needed anyway in the course of  $\mathbf{MINIEIG}$  it seems preferable to use the latter more elaborate version only at the end once  $\mathbf{u}(\epsilon)$  has been obtained.

### 4.3.2 Modified iterations

As for b), another option is to modify the ideal iteration itself, again trying to exploit the fact that the approximation rate of the (exact) Rayleigh quotients is faster than that of the eigendirections. We only give a very rough sketch of the idea. Instead of the final accuracy one could run *MINIEIG* first with some intermediate accuracy  $\varepsilon$  outputting again  $\mathbf{u}(\varepsilon)$  satisfying (4.29). One could then fix  $\bar{\lambda} = \lambda^*(\varepsilon)$ , set  $\mathbf{x}_0 := \mathbf{u}(\varepsilon)$  and iterate

$$\mathbf{x}_{i+1} = \mathbf{x}_i - \alpha(\mathbf{A} - \bar{\lambda}\mathbf{C})\mathbf{x}_i,$$

so that, by straightforward calculations, one obtains

$$(\mathbf{x}_{i+1} - \mathbf{u})^\perp = (\mathbf{x}_i - \mathbf{u})^\perp - \Phi(\mathbf{x}_i - \mathbf{u})^\perp + \alpha(\bar{\lambda} - \underline{\lambda})(\mathbf{I} - \mathbf{P})\mathbf{C}\mathbf{x}_i,$$

and hence by recursion

$$\|(\mathbf{x}_i - \mathbf{u})^\perp\| \leq \beta^i \varepsilon + \frac{(1 + 2K\Gamma\alpha/\gamma)\varepsilon^2}{1 - \beta} \|\mathbf{C}\|. \quad (4.44)$$

Thus, in principle, after  $|\log \varepsilon|$  steps one has quadratic accuracy of  $(\mathbf{x}_i - \mathbf{u})^\perp$  without further applications of *RAYL*. However, the iterates  $\mathbf{x}_i$  are no longer orthogonal to the residuals  $\bar{\mathbf{r}}_i := (\mathbf{A} - \bar{\lambda}\mathbf{C})\mathbf{x}_i$ . Observing that  $\langle \bar{\mathbf{r}}_i, \mathbf{x}_i \rangle = (\lambda(\mathbf{x}_i) - \bar{\lambda})\langle \mathbf{C}\mathbf{x}_i, \mathbf{x}_i \rangle$ , we have

$$\|\mathbf{x}_{i+1}\|^2 = \|\mathbf{x}_i\|^2 + \alpha^2 \|\bar{\mathbf{r}}_i\|^2 - 2\alpha(\lambda(\mathbf{x}_i) - \bar{\lambda})\|\mathbf{x}_i\|_{\mathbf{C}}^2 \geq \|\mathbf{x}_i\|^2 + \alpha^2 \|\bar{\mathbf{r}}_i\|^2 - 2\alpha(\lambda(\mathbf{x}_i) - \bar{\lambda})\|\mathbf{C}\|\|\mathbf{x}_i\|^2,$$

so that we cannot guarantee any more that the iterates have increasing norms. Nevertheless, since  $|\lambda(\mathbf{x}_j) - \bar{\lambda}|$  is expected to be of the order  $\varepsilon^2$  and  $\|\mathbf{x}_0\| = 1$ , there is still a fixed positive constant  $b \geq 1/2$  say, so that after  $J \sim |\log \varepsilon|$  steps, one still has  $\|\mathbf{x}_J\| \geq b$ , so that renormalizing the  $J$ th iterate preserves quadratic accuracy. This way, one expects to catch up quadratic accuracy of the approximate eigendirections without any intermediate computation of Rayleigh quotients. One can then continue such a block of iterations with initial guess  $\mathbf{x}_J/\|\mathbf{x}_J\|$  that approximates  $\mathbf{u}$  now within a tolerance of the order of  $\varepsilon^2$  so that a corresponding approximate eigenvalue can be computed within tolerance  $\varepsilon^4$  with the aid of the above fast computation of scalar products, etc. We leave the details to the reader.

Concerning optimal line search (3.1) and subspace acceleration techniques, we mention that this requires the solution of a two-dimensional (or more generally small) generalized eigenvalue problem. However, setting up the corresponding matrices requires the computation of scalar products. Since, generally, we cannot compute these quantities exactly, a careful assessment of the perturbation effects would be necessary. Also, the possibilities for realizing enhanced accuracy at reduced cost mentioned above might be relevant in this context.

## 5 Remarks concerning an application - The Schrödinger equation

We conclude with some brief comments concerning a scenario that has actually motivated part of this work.

We consider the time-independent electronic Schrödinger equation  $H\psi = E_0\psi$ , where  $H : H^1((\mathbb{R}^3 \times \{\pm\frac{1}{2}\})^N) \rightarrow H^{-1}((\mathbb{R}^3 \times \{\pm\frac{1}{2}\})^N)$  is the Hamilton operator of the molecular  $N$ -electron system under consideration, and  $E_0$  denotes the lowest eigenvalue of  $H$  corresponding to the ground state of the system. The spectrum of  $H$  is real and bounded from below by a certain  $\mu \in \mathbb{R}$  (cf. e.g. [Yser] and references therein), so that the shifted Hamiltonian  $H' := H - \mu\mathcal{E}$  satisfies the conditions (1.4) and (1.5) of Section 1. Letting  $\Phi = \{\phi_\nu | \nu \in \mathcal{I}\}$  be an orthonormal basis of  $L_2((\mathbb{R}^3 \times \{\pm\frac{1}{2}\})^N)$ , we have to solve the eigenvalue equation

$$\mathbf{H}\mathbf{c} = E_0\mathbf{c} \tag{5.1}$$

where the entries  $\mathbf{H}_{\nu',\nu}$  of the “matrix Hamiltonian” in this formulation are given by

$$\mathbf{H}_{\nu',\nu} = \langle H\psi_{\nu'}, \psi_\nu \rangle.$$

In view of our discussion in Section 1, the problem is posed in  $\ell_2(\mathcal{I})$ , which gives the so-called *complete CI* formulation for the Schrödinger equation. To obtain a Riesz basis for  $H^1$ , we can utilize a reference operator  $\mathcal{F} = \sum_{i=1}^N F_i$ , where the  $F_i : H^1(\mathbb{R}^3 \times \{\pm\frac{1}{2}\}) \rightarrow H^{-1}(\mathbb{R}^3 \times \{\pm\frac{1}{2}\})$  are single particle operators with a complete eigenvalue basis. The self consistent Fock operator from the Hartree Fock equations, see e.g. [Helg], can be modified to fit the present purpose. This operator is also bounded as an operator  $H^1((\mathbb{R}^3 \times \{\pm\frac{1}{2}\})^N) \rightarrow H^{-1}((\mathbb{R}^3 \times \{\pm\frac{1}{2}\})^N)$  and its spectrum is bounded from below, so that  $\mathcal{F}' := \mathcal{F} - \mu\mathcal{E}$  satisfies (1.5). In this case the eigenfunctions  $\phi_\nu$  and the eigenvalues  $\sigma_\nu$  of  $\mathcal{F}, \mathcal{F}'$  can easily be computed from those of the single particle operator  $F_i$ ,

$$\mathcal{F}'\phi_\nu = \sigma_\nu\phi_\nu, \quad \nu \in \mathcal{J}.$$

If we choose

$$\Psi = \{\psi_\nu := \sigma_\nu^{-\frac{1}{2}}\phi_\nu | \nu \in \mathcal{I}\}$$

this basis satisfies (1.8). The generalized eigenvalue problem (2.1) then resembles the symmetry-transformed variant of the Schrödinger equation with

$$\mathbf{C} := (\langle \sigma_\nu^{-\frac{1}{2}}\phi_\nu, \sigma_{\nu'}^{-\frac{1}{2}}\phi_{\nu'} \rangle)_{\nu,\nu' \in \mathcal{I}} \quad \text{and} \quad \mathbf{A} := \mathbf{C}^{\frac{1}{2}}\mathbf{H}'\mathbf{C}^{\frac{1}{2}} = \mathbf{C}^{\frac{1}{2}}(\mathbf{H} - \mu\mathbf{I})\mathbf{C}^{\frac{1}{2}}, \tag{5.2}$$

where the conditions (2.2) and (2.3) are valid.

With the above notation,  $\|\mathbf{x}\|_{\mathbf{C}}$  then gives the  $L_2((\mathbb{R}^3 \times \{\pm\frac{1}{2}\})^N)$ -error of the wave function  $\varphi := \sum_{\nu \in \mathcal{I}} x_\nu \psi_\nu = \sum_{\nu \in \mathcal{I}} c_\nu \phi_\nu$ . For estimating the  $H^1((\mathbb{R}^3 \times \{\pm\frac{1}{2}\})^N)$ -error of the wave function  $\varphi$ , note that the canonical  $H^1((\mathbb{R}^3 \times \{\pm\frac{1}{2}\})^N)$ -norm is equivalent to the norms induced by the inner products  $\langle\langle \cdot, \cdot \rangle\rangle := \langle H' \cdot, \cdot \rangle_{L_2}$  and  $\langle \mathcal{F}' \cdot, \cdot \rangle_{L_2}$ , defined by the Hamiltonian and the reference operator respectively, which gives

$$\begin{aligned} \|\mathbf{x}\|_{\mathbf{A}}^2 &= \langle \mathbf{C}^{\frac{1}{2}} \mathbf{H}' \mathbf{C}^{\frac{1}{2}} \mathbf{x}, \mathbf{x} \rangle &= \langle \mathbf{H}' \mathbf{C}^{\frac{1}{2}} \mathbf{x}, \mathbf{C}^{\frac{1}{2}} \mathbf{x} \rangle \\ &= \sum_{\nu \in \mathcal{I}} \sum_{\nu' \in \mathcal{I}} c_\nu c_{\nu'} \langle H' \phi_\nu, \phi_{\nu'} \rangle &= \langle H' \varphi, \varphi \rangle \simeq \|\varphi\|_{H^1}^2, \end{aligned}$$

and

$$\|\mathbf{x}\|_2 = \|\mathbf{C}^{\frac{1}{2}} \mathbf{x}\|_{\mathbf{C}^{-1}} = \|\mathbf{c}\|_{\mathbf{C}^{-1}} = \|\varphi\|_{\mathcal{F}'} \simeq \|\varphi\|_{H^1}.$$

This serves also as a typical example where  $\mathbf{C}$  is not coercive on  $\ell_2(\mathcal{I})$  (because  $\|\cdot\|_{\mathbf{C}}$  is equivalent to the  $L_2$ -norm), cf. Remark 1. The norms  $\|\cdot\|_{\mathbf{A}}$  and  $\|\cdot\|_2 \simeq \|\cdot\|_{\mathbf{A}}$  on  $\ell_2(\mathcal{I})$  can now be used to estimate the convergence of the  $CI$  solution with respect to the  $H^1$ -norm.

## References

- [BabO] I. Babuska, J. Osbourne, Eigenvalue problems, in: Handbook of numerical analysis, Vol. II, P.G. Ciarlet, J.L. Lions (editors), Elsevier Science Publisher, North-Holland, Amsterdam, 1991.
- [B] A. Barinka, Fast Evaluation Tools for Adaptive Wavelet Methods, Doctoral Dissertation, RWTH Aachen, Dec. 2004.
- [BPK] J.H. Bramble, J.E. Pasciak, A. V. Knyazev, A subspace preconditioning algorithm for eigenvector/eigenvalue computation, Advances in Computational Mathematics, Volume 6, Number 1, December 1996, pp. 159-189.
- [C] A. Cohen, Numerical analysis of wavelet methods, Studies in Mathematics and its Applications, Vol. 32, Elsevier, Amsterdam, 2003.
- [Chat] F. Chatelin, Eigenvalues of Matrices, Wiley, Chichester, 1993.
- [Ciarl] P.G. Ciarlet, J.L. Lions, Handbook of numerical analysis, Vol. II, North-Holland, Amsterdam, New York, 1991.
- [CDD1] A. Cohen, W. Dahmen, R. DeVore, Adaptive wavelet methods for elliptic operator equations: Convergence estimates, Math. Comp. 70 (2001), 27–75.
- [CDD2] A. Cohen, W. Dahmen, R. DeVore, Adaptive wavelet methods II - Beyond the elliptic case, Foundations of Computational Mathematics, 2 (2002), 203–245.

- [CDD3] A. Cohen, W. Dahmen, R. DeVore, Adaptive Wavelet Schemes for Nonlinear Variational Problems, *SIAM J. Numer. Anal.*, (5)41(2003), 1785–1823.
- [CDD4] A. Cohen, W. Dahmen, R. DeVore, Adaptive wavelet techniques in Numerical Simulation, in: *Encyclopedia of Computational Mechanics*, (R. De Borste, T. Hughes, E. Stein, eds.), Wiley-Interscience, 2004, 157–197.
- [D] W. Dahmen, Wavelet and Multiscale Methods for Operator Equations, *Acta Numerica*, Cambridge University Press, 6(1997), 55–228.
- [DHS] W. Dahmen, H. Harbrecht, R. Schneider, Adaptive methods for boundary integral equations - complexity estimates, *Math. Comp.* 76 (2007), 1243-1274.
- [DeVore] R.A. DeVore, Nonlinear approximation, *Acta Numerica* 7, 51–150 (1998)
- [Dyak] E.G. D’yakonov, Optimization in Solving Elliptic Problems, CRC, Boca Raton, 1995.
- [GS] T. Gantumur, R. Stevenson, Computation of differential operators in wavelet coordinates, *Math. Comp.*, 75(2006), 697–709.
- [GG] S. Giani, I. Graham, A convergent adaptive method for elliptic eigenvalue problems, BICS Preprint, June 2007, Bath Institute for Complex Systems.
- [GvL] G.H. Golub, C.F. van Loan, Matrix Computations, John Hopkins University Press, Baltimore, 1996.
- [Helg] T. Helgaker, P. Jørgensen, J. Olsen, Molecular electronic-structure theory, Wiley, New York, 2000.
- [HiSi] P.D. Hislop, I.M. Sigal, Introduction to spectral theory, Springer, New York, 1996.
- [Kato] T. Kato, Perturbation theory for linear operators, New York, 1966.
- [KN] A. V. Knyazev, K. Neymeyr, A geometric theory for preconditioned inverse iteration. III: A short and sharp convergence estimate for generalized eigenvalue problems, *Linear Algebra Appl.*, 358 (2003), 95–114.
- [Parl] B. N. Parlett, The symmetric eigenvalue problem, Society for Industrial and Applied Mathematics, Philadelphia, 1998.
- [ReSi] M. Reed, B. Simon, Methods of modern mathematical physics, Vol. IV: Analysis of operators, Academic Press, New York, 1978.
- [RSZ] T. Rohwedder, R. Schneider, A. Zeiser, Perturbed preconditioned inverse iteration for operator eigenvalue problems with applications to adaptive wavelet discretization, Preprint arXiv:0708.0517v1 [math.NA], submitted to Adv. Comp. Math.

- [S] R. Stevenson, On the compressibility of operators in wavelet coordinates, *SIAM J. Math. Anal.*, 35(5)(2004), 1110–1132.
- [Weid] J. Weidmann, *Lineare Operatoren in Hilberträumen, Teil 1: Grundlagen*, Springer, Stuttgart, 2000.
- [Yser] H. Yserentant, *On the electronic Schrödinger equation*, Lecture Notes, Universität Tübingen, 2003