Guaranteed a posteriori bounds for eigenvalues and eigenvectors: multiplicities and clusters

Eric Cancès, Geneviève Dusson, Yvon Maday, Benjamin Stamm, Martin Vohralík

Guaranteed *a posteriori* bounds for eigenvalues and eigenvectors: multiplicities and clusters *

Eric Cancès[1,6], Geneviève Dusson[2], Yvon Maday[3,4], Benjamin Stamm[5], and Martin Vohralík[6,1]

[1] *CERMICS, Ecole des Ponts ParisTech, 6 & 8 Av. Pascal, 77455 Marne-la-Vallée, France*
[2] *Université Bourgogne Franche-Comté, Laboratoire de Mathématiques de Besançon, UMR CNRS 6623, 16 route de Gray, 25030 Besançon, France*
[3] *Sorbonne Université and Université de Paris, CNRS, Laboratoire Jacques-Louis Lions (LJLL), 75005 Paris, France*
[4] *Institut Universitaire de France, 75005 Paris, France*
[5] *Center for Computational Engineering Science, RWTH Aachen University, Aachen, Germany*
[6] *Inria, 2 Rue Simone Iff, 75589 Paris, France*

August 10, 2020

Abstract

This paper presents *a posteriori* error estimates for conforming numerical approximations of eigenvalue clusters of second-order self-adjoint elliptic linear operators with compact resolvent. Given a cluster of eigenvalues, we estimate the error in the sum of the eigenvalues, as well as the error in the eigenvectors represented through the density matrix, i.e., the orthogonal projector on the associated eigenspace. This allows us to deal with degenerate (multiple) eigenvalues within the framework. All the bounds are valid under the only assumption that the cluster is separated from the surrounding smaller and larger eigenvalues; we show how this assumption can be numerically checked. Our bounds are guaranteed and converge with the same speed as the exact errors. They can be turned into fully computable bounds as soon as an estimate on the dual norm of the residual is available, which is presented in two particular cases: the Laplace eigenvalue problem discretized with conforming finite elements, and a Schrödinger operator with periodic boundary conditions of the form $-\Delta + V$ discretized with planewaves. For these two cases, numerical illustrations are provided on a set of test problems.

Contents

*Part of this work has been supported from French state funds managed by the CalSimLab LABEX and the ANR within the Investissements d'Avenir program (reference ANR-11-LABX-0037-01). The last author has also received funding from the European Research Council (ERC) under the European Unions Horizon 2020 research and innovation program (Grant Agreement No. 647134 GATIPOR). Part of this work was supported by the French "Investissements d'Avenir" program, project ISITE-BFC (contract ANR-15-IDEX-0003). YM and EC acknowledge funding from PICS- CNRS, PHC PROCOPE 2017 (Project No. 37855ZK), and European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 810367).

1 Introduction

Elliptic eigenvalue problems arise in many mathematical models used in science and engineering; often, precise approximations of eigenvalues and eigenvectors are crucial. To guarantee the quality of the approximations at stake, one needs to estimate the size of the errors for the computed quantities, namely the eigenvalues and eigenvectors. *A posteriori* error bounds aim at providing such estimates.

Already very good *a posteriori* estimates have been proposed for elliptic source problems, based for example on the theory of equilibrated fluxes for the Laplace source problem following Prager and Synge [39], see [33, 18, 6, 21] and the references therein. Nonetheless, the error estimation for eigenvalue problems seems more complex in comparison. Following Kato [32], Forsythe [23], Weinberger [44], and Bazley and Fox [2], recent works have been presented for the estimation of simple eigenvalues, possibly only the lowest one, see e.g. [38, 30, 31, 13, 34], see also the references therein. A thorough *a posteriori* analysis of errors in both simple eigenvectors and eigenvalues for the Laplace eigenvalue problem can be found in [9] (conforming discretization methods) and [10] (a unified framework including nonconforming discretization methods).

The above results however only hold for simple eigenvalues, whereas degenerate or near degenerate eigenvalues often appear in practice. This can dramatically deteriorate the estimates, especially when the latter depend on the gap between the estimated eigenvalue and the surrounding ones, as it is the case in [9, 10]. Only few results have been presented so far for the *a posteriori* estimation of multiple eigenvalues or clusters of eigenvalues. In [34, 43], guaranteed error estimates are presented for the eigenvalue error, but not the eigenvector error. The derivation of optimal eigenvalue convergence rates for adaptive finite element methods can be found in [25, 17] for conforming finite elements, in [24] and [4] for nonconforming and mixed finite elements, and in [5] for conforming and nonconforming finite elements of higher order. *A posteriori* error estimation for clusters of eigenvalues have been presented in the case of the discontinuous Galerkin method in [26] and for Crouzeix–Raviart nonconforming finite elements in [3]. Also, *a posteriori* bounds have been established in [28] and [1] for $\mathbb{P}_1$ finite elements with triangular meshes, where the bounds are directly derived on the eigenspace, and are therefore independent of the choice of the eigenvectors, as is the case in this work. Finally, a recent contribution deriving upper bounds on eigenvectors associated with multiple eigenvalues is [37]. Though the methodology also works on eigenspaces and the bounds are guaranteed, it does not seem to extend to a general cluster (the precision is limited by the difference of the largest and smallest eigenvalue in the cluster).

In this article, we extend the *a posteriori* error estimates presented in [9] to clusters of eigenvalues, which includes the possibility of degenerate eigenvalues. The estimators are derived for a generic second-order elliptic self-adjoint operator with compact resolvent denoted by A. More precisely, let $(\lambda_i, \varphi_i^0)_{i\geq 1}$ be the eigenvalues and associated eigenvectors of the operator A. We are interested in the cluster of eigenvalues $\lambda_m, \ldots, \lambda_M$, with $m, M \in \mathbb{N}\backslash\{0\}, m \leq M$. We first derive guaranteed bounds for the error in the sum of the eigenvalues. To derive these bounds, the only necessary assumption is that the cluster is separated from the surrounding lower and higher eigenvalues, as stated in Assumption 2.1, and a continuous–discrete gap condition summarized in Assumption 4.2. The problem is described in Section 2.1, and we consider a conforming discretization presented in Section 2.3.

In order to account for all the exact (respectively approximate) eigenvalues and eigenvectors of the cluster as a whole, the estimates rely on the use of density matrices, which are the orthogonal projectors on the exact (resp. approximate) eigenspaces spanned by the eigenvectors of the cluster. This allows to handle the nonuniqueness of the eigenvectors. Indeed, under the gap Assumption 2.1, the exact density matrix γ^0 is uniquely defined. The error estimates are therefore presented on the density matrix error in Hilbert–Schmidt and energy norms. All the definitions necessary to introduce the density matrix framework are presented in Section 2.2.

As we are aware that error estimates are usually provided in terms of eigenvectors instead of density matrix, we show in Section 3.1 that for a well-chosen norm, the error on the density matrix is in fact equivalent to the error on the eigenvectors for a particular choice of approximate eigenvectors, which essentially guaranties that they are aligned with the reference exact eigenvectors. We then in Section 3.2 introduce the residual, defined in this framework as an operator and not as a functional as in usual eigenvalue problems.

Generic error equivalences are presented in Section 4. More precisely, we provide estimates on density matrix errors and on the sum of eigenvalues error in terms of dual norms of the residual. The bounds are guaranteed, containing no unknown constant, but are not directly computable. Indeed, they depend on the dual norm of the residual, which is not always computable and can be difficult to estimate.

In Section 5, we transform these equivalences into fully computable error bounds in two cases. We first in Section 5.1 treat the case of the Laplace eigenvalue problem on an open Lipschitz polygon or polyhedron $\Omega \subset \mathbb{R}^d$ with Dirichlet boundary conditions discretized on simplicial meshes by conforming finite elements of degree p, based on [9] for the estimate of the dual norm of the residual. This estimate relies on the construction of an equilibrated flux requiring to solve mixed finite element local residual problems. The error bound for the sum of the eigenvalues in the considered cluster is given in Theorem 5.3 and reads

$$0 \leq \sum_{i=m}^{M} (\lambda_{ih} - \lambda_i) \leq \eta^2, \tag{1.1}$$

where λ_{ih} is the i^{th} approximate eigenvalue (counting multiplicities). Further, error bounds on the density matrix error are provided. In particular, Theorem 5.6 shows that

$$\| |\nabla| (\gamma^0 - \gamma_h) \|_{\mathfrak{S}_2(\mathcal{H})} \leq \eta, \tag{1.2}$$

γ_h being the approximate density matrix and $\|\cdot\|_{\mathfrak{S}_2(\mathcal{H})}$ the Hilbert–Schmidt norm associated with the $L^2(\Omega)$ Hilbert space. Moreover, Theorem 5.6 also shows that these bounds are efficient in the sense that

$$\eta \leq C \| |\nabla| (\gamma^0 - \gamma_h) \|_{\mathfrak{S}_2(\mathcal{H})}, \tag{1.3}$$

where C is a generic constant independent of the mesh size h and the polynomial degree p.

We distinguish two cases. In Case I, no assumption other than the gap Assumptions 2.1 and 4.2 are needed, and we give sufficient conditions to check them in practice, cf. Remark 4.3. There, the bound η^2 only depends on the flux reconstruction and on a lower bound for the relative gaps between the cluster of approximate eigenvalues and the surrounding exact eigenvalues, that is lower bounds of the quantities $\left(\dfrac{\lambda_{mh}}{\lambda_{m-1}} - 1\right)$ and $\left(1 - \dfrac{\lambda_{Mh}}{\lambda_{M+1}}\right)$. In Case II, which holds under an additional elliptic regularity assumption on the corresponding source problem as described in (5.11), the pre-factor in η can be brought to the optimal value of 1.

In Section 5.2, we then provide bounds for Schrödinger-type operators of the form $-\Delta + V$ on a cubic box with periodic boundary conditions, discretized with a planewave basis, in which case the dual norm

Assumption	Name	Implications
Assumption 2.1	Continuous gap conditions	
Assumption 2.6	Non-orthogonality of the exact and approximate eigenspaces	
Assumption 4.2	Continuous–discrete gap conditions	
Assumption 4.4	Availability of accurate enough lower bounds for λ_m and λ_{M+1}	implies Assumptions 2.1 and 4.2

Table 1: Assumptions and implications between assumptions

Assumptions	Reference of the results
2.1	Lemma 3.1 [Link between eigenvalue and eigenvector errors]
	Theorem 4.1 [Eigenvalue bounds]
	Theorem 4.5, (4.4) [Upper bounds for the density matrix error]
and 2.6	Lemma 3.3 [Link between density matrix and eigenvector errors]
	Theorem 4.6 [Lower bound for the density matrix error]
2.1 and 4.2	Lemma 4.7 [Bounds on the density matrix error]
	Theorem 5.3, Case II [Guaranteed bounds for the sum of eigenvalues]
	Theorem 5.6, Case II [Guaranteed bound for the density matrix error]
	Theorem 5.9 [Guaranteed bounds for the sum of eigenvalues]
	Theorem 5.10 [Guaranteed bound for the density matrix errors]
and 2.6	Theorem 4.5, (4.6) [Upper bounds for the density matrix error]
	Theorem 5.3, Case I [Guaranteed bounds for the sum of eigenvalues]
	Theorem 5.6, Case I [Guaranteed bound for the density matrix error]

Table 2: Summary of the main results with the employed assumptions.

of the residual is explicitly computable as the Laplace operator is diagonal in this basis. This allows to straightforwardly apply the bounds obtained in Section 4 and derive error estimates both for the sum of the eigenvalues error and the error on the density matrix built from the eigenvectors in the form similar to (1.1)–(1.3). For clarity, the different assumptions used throughout this article are collected in Table 1, and the main results and corresponding assumptions are listed in Table 2.

We present in Section 6 numerical results for (i) the Laplace operator discretized with conforming finite elements in a 2D setting, and (ii) a Schrödinger operator $-\Delta + V$ on a cubic box with periodic boundary conditions, discretized in a planewave basis, in a 1D and 2D setting. The error bounds fulfill the expectations, and in particular, the necessary assumptions already hold for coarse bases. Finally, some conclusions are drawn in Section 7, and Appendix A details the proof of a technical result.

2 Setting

We introduce here the considered eigenvalue problem, its generic conforming discretization, and the functional analysis setting that we adopt.

2.1 The eigenvalue problem

Let $\mathcal{H}$ be a real *separable Hilbert space* endowed with an inner product denoted by $(\cdot, \cdot)$, and a corresponding norm denoted by $\|\cdot\|$. We consider a *self-adjoint operator* A on $\mathcal{H}$ with domain $D(A)$, *bounded-below*, and with *compact resolvent*. For such an operator, there exists a non-decreasing sequence of real numbers $(\lambda_k)_{k\geq 1}$ such that $\lambda_k \to +\infty$ and an orthonormal basis $(\varphi_k^0)_{k\geq 1}$ of $\mathcal{H}$ consisting of vectors of $D(A)$ such that

$$\forall k \geq 1, \quad A\varphi_k^0 = \lambda_k \varphi_k^0. \tag{2.1}$$

In the following, we will often employ the Parseval identity, which states that for any $v \in \mathcal{H}$,

$$\|v\|^2 = \sum_{k\geq 1} |(v, \varphi_k^0)|^2. \tag{2.2}$$

Up to shifting the operator A by a constant $c \in \mathbb{R}_+$, we can assume without loss of generality that A is a positive definite operator, in which case $(\lambda_k)_{k\geq 1}$ is a sequence of positive numbers. This enables to define the operators A^s, $s \in \mathbb{R}$, by their domains

$$D(A^s) := \left\{ v \in \mathcal{H}; \quad \|A^s v\|^2 := \sum_{k\geq 1} \lambda_k^{2s} |(v, \varphi_k^0)|^2 < +\infty \right\} \tag{2.3a}$$

and expressions

$$A^s : v \in D(A^s) \mapsto \sum_{k\geq 1} \lambda_k^s (v, \varphi_k^0)\varphi_k^0 \in \mathcal{H}. \tag{2.3b}$$

In particular, the norm $\|A^{1/2} \bullet \|$ is referred to as the *energy norm*. We also remark that A^0 is the identity operator on $\mathcal{H}$, that $A^s A^t = A^{s+t}$ for all $s, t \in \mathbb{R}$, and that $D(A^s) = \mathcal{H}$ for all $s \leq 0$. Also, $D(A^s) \subset D(A^t)$ for $s \geq t$, so that for all $k \geq 1$, φ_k^0 from (2.1) belongs to $D(A^s)$ for all $s \in \mathbb{R}$ and

$$(A^{1/2}\varphi_k^0, A^{1/2}v) = \lambda_k(\varphi_k^0, v) \qquad \forall v \in D(A^{1/2}),\ \forall k \geq 1. \tag{2.4}$$

This in particular implies, as $\|\varphi_k^0\| = 1$,

$$\|A^{1/2}\varphi_k^0\|^2 = \lambda_k \qquad \forall k \geq 1. \tag{2.5}$$

2.2 Functional analysis setting

In this article, we focus on the error estimation of clusters of eigenvalues and their corresponding eigenvectors. More precisely, given $m, M \in \mathbb{N} \setminus \{0\}$, $m \leq M$, we consider the *eigenvalue cluster* composed of the $J := M - m + 1$ eigenvalues $(\lambda_m, \ldots, \lambda_M)$ from (2.1), counted with their multiplicities. For our *a posteriori* analysis, we will need to assume that the considered cluster is separated from the rest of the spectrum:

Assumption 2.1 (Continuous gap conditions)**.** *There holds $\lambda_{m-1} < \lambda_m$ if $m > 1$ and $\lambda_M < \lambda_{M+1}$.*

We show in Remark 4.3 below how this condition can be verified practically.

We denote an orthonormal set of corresponding *eigenvectors* by

$$\mathbf{\Phi}^0 := (\varphi_m^0, \ldots, \varphi_M^0). \tag{2.6}$$

Note that estimating the error between $\mathbf{\Phi}^0$ and given approximate eigenvectors $\mathbf{\Phi}_h = (\varphi_{mh}, \ldots, \varphi_{Mh})$ cannot in general be done without further assumptions on the choice of the eigenvectors. Indeed, in particular for multiple eigenvalues $\lambda_m = \ldots = \lambda_M$, for any matrix $\mathbf{U} \in O(J) = \left\{ \mathbf{U} \in \mathbb{R}^{J\times J} \middle| \mathbf{U}^T\mathbf{U} = \mathbf{1}_J \right\}$, the group of orthogonal matrices of order J, $\mathbf{\Phi}^0\mathbf{U}$ also form an orthonormal set of eigenvectors associated with $(\lambda_m, \ldots, \lambda_M)$.

To get rid of the above problematic nonuniqueness, we will measure and estimate the errors not on the eigenvectors directly, but in the spaces spanned by these eigenvectors, which are uniquely determined, even in the case of degenerate (multiple) eigenvalues, as long as the gap Assumption 2.1 is satisfied. In this case, the orthogonal projector for the inner product $(\cdot, \cdot)$ onto $\mathrm{Span}\{\varphi_m^0, \ldots, \varphi_M^0\}$, denoted by γ^0 and called *density matrix*, is also unique. It is the rank-J operator on $\mathcal{H}$ defined by

$$\forall v \in \mathcal{H}, \quad \gamma^0 v := \sum_{i=m}^{M} (v, \varphi_i^0)\varphi_i^0. \tag{2.7}$$

The exact and approximate eigenspaces can therefore be compared through their density matrices. In fact, we will introduce below a norm to measure the error on the density matrices which is equivalent to the energy norm of the error on the eigenvectors for the particular choice of eigenvectors for which the approximate eigenvectors are as much aligned as possible with the corresponding exact eigenvectors. Note that in particular $\gamma^0 v = v$ if $v \in \operatorname{Span}\{\varphi_m^0, \ldots, \varphi_M^0\}$ and $\gamma^0 v = 0$ if v is in the orthogonal complement of $\operatorname{Span}\{\varphi_m^0, \ldots, \varphi_M^0\}$, which will often be used below.

The functional setting of trace-class and Hilbert–Schmidt operators used to define this norm is presented in detail in [40, Chapter VI] and can also be found in [11]. We only briefly recall here the properties used in this article. We denote by $\mathcal{L}(\mathcal{H})$ the space of bounded linear operators on $\mathcal{H}$. If $B \in \mathcal{L}(\mathcal{H})$ is a *positive* operator, (i.e., $(v, Bv) \geq 0$ for any $v \in \mathcal{H}$), then the value of the sum

$$\operatorname{Tr}(B) := \sum_{k \geq 1} (e_k, Be_k) \in \mathbb{R}_+ \cup \{+\infty\} \tag{2.8}$$

is independent of the choice of the orthonormal basis $(e_k)_{k\geq 1}$ of $\mathcal{H}$. Let now $B \in \mathcal{L}(\mathcal{H})$ be arbitrary and let $B^\dagger$ denote the adjoint of B, i.e., $B^\dagger \in \mathcal{L}(\mathcal{H})$ such that $(B^\dagger v, w) = (v, Bw)$ for all $v, w \in \mathcal{H}$. We define $|B| := \sqrt{B^\dagger B}$,

$$\|B\|_{\mathfrak{S}_1(\mathcal{H})} := \operatorname{Tr}(|B|) = \sum_{k\geq 1} (e_k, |B| e_k), \tag{2.9}$$

and

$$\|B\|_{\mathfrak{S}_2(\mathcal{H})} := \operatorname{Tr}(B^\dagger B)^{1/2} = \left(\sum_{k\geq 1} \|Be_k\|^2 \right)^{1/2}. \tag{2.10}$$

Then the Banach space $\mathfrak{S}_1(\mathcal{H})$ of *trace-class* operators on $\mathcal{H}$ is the space of all $B \in \mathcal{L}(\mathcal{H})$ with $\|B\|_{\mathfrak{S}_1(\mathcal{H})} < +\infty$. In particular, if $B \in \mathcal{L}(\mathcal{H})$ is positive and *self-adjoint*, then $B \in \mathfrak{S}_1(\mathcal{H})$ if and only if $\operatorname{Tr}(B) < +\infty$. The Hilbert space $\mathfrak{S}_2(\mathcal{H})$ of *Hilbert–Schmidt* operators on $\mathcal{H}$ is the space of all $B \in \mathcal{L}(\mathcal{H})$ with $\|B\|_{\mathfrak{S}_2(\mathcal{H})} < +\infty$ endowed with the scalar product

$$(B, C)_{\mathfrak{S}_2(\mathcal{H})} := \operatorname{Tr}(B^\dagger C). \tag{2.11}$$

Recall that $\mathfrak{S}_1(\mathcal{H}) \subset \mathfrak{S}_2(\mathcal{H}) \subset \mathfrak{S}_\infty(\mathcal{H}) \subset \mathcal{L}(\mathcal{H})$, where $\mathfrak{S}_\infty(\mathcal{H})$ is the vector space of compact operators on $\mathcal{H}$, and that for any compact self-adjoint operator B on $\mathcal{H}$, we have

$$\|B\|_{\mathcal{L}(\mathcal{H})} = \max_{i\geq 1} |\mu_i|, \quad \|B\|_{\mathfrak{S}_2(\mathcal{H})} = \left(\sum_{i\geq 1} |\mu_i|^2 \right)^{1/2}, \quad \|B\|_{\mathfrak{S}_1(\mathcal{H})} = \sum_{i\geq 1} |\mu_i|,$$

where the μ_i's are the eigenvalues of B counting multiplicities, so that

$$\|B\|_{\mathcal{L}(\mathcal{H})} \leq \|B\|_{\mathfrak{S}_2(\mathcal{H})} \leq \|B\|_{\mathfrak{S}_1(\mathcal{H})} \quad \text{and} \quad \|B\|_{\mathfrak{S}_2(\mathcal{H})} \leq \|B\|_{\mathcal{L}(\mathcal{H})}^{1/2} \|B\|_{\mathfrak{S}_1(\mathcal{H})}^{1/2}.$$

For all $B \in \mathfrak{S}_2(\mathcal{H})$ and $C \in \mathfrak{S}_2(\mathcal{H})$, it follows that $BC \in \mathfrak{S}_1(\mathcal{H})$, $CB \in \mathfrak{S}_1(\mathcal{H})$, and

$$\operatorname{Tr}(BC) = \operatorname{Tr}(CB) \leq \|B\|_{\mathfrak{S}_2(\mathcal{H})} \|C\|_{\mathfrak{S}_2(\mathcal{H})}. \tag{2.12}$$

Note that $\gamma^0 \in \mathcal{L}(\mathcal{H})$ and γ^0 is positive, self-adjoint, and an (orthogonal) projector since $(\gamma^0)^2 = \gamma^0$. Furthermore, $\gamma^0 \in \mathfrak{S}_1(\mathcal{H})$ and its trace is equal to $J = M - m + 1$, the dimension of $\operatorname{Span}\{\varphi_m^0, \ldots, \varphi_M^0\}$. Indeed,

$$\operatorname{Tr}(\gamma^0) = \sum_{k\geq 1} (\varphi_k^0, \gamma^0 \varphi_k^0) = \sum_{k=m}^{M} |(\varphi_k^0, \varphi_k^0)|^2 = J. \tag{2.13}$$

Moreover, we have

$$\|\gamma^0\|_{\mathfrak{S}_2(\mathcal{H})}^2 = \operatorname{Tr}\Big((\gamma^0)^\dagger \gamma^0 \Big) = \operatorname{Tr}((\gamma^0)^2) = \operatorname{Tr}(\gamma^0) = J. \tag{2.14}$$

For the following, we set, for all $s \in \mathbb{R}$,

$$\forall \Psi = (\psi_m, \ldots, \psi_M) \in [D(A^s)]^J,\ \|A^s \Psi\| := \left(\sum_{i=m}^{M} \|A^s \psi_i\|^2 \right)^{1/2}. \tag{2.15}$$

A particular consequence of the definitions presented above is:

Lemma 2.2 (Difference of orthogonal projectors)**.** *Let γ_I and γ_L be two finite-rank orthogonal projectors of the same rank. There holds*

$$\|\gamma_L - \gamma_I\|^2_{\mathfrak{S}_2(\mathcal{H})} = 2\mathrm{Tr}(\gamma_L(1-\gamma_I)).$$

Proof. Note that the traces of γ_I and γ_L are equal (to their rank). Therefore,

$$\begin{aligned}
\|\gamma_L - \gamma_I\|^2_{\mathfrak{S}_2(\mathcal{H})} &= \mathrm{Tr}\left((\gamma_L - \gamma_I)(\gamma_L - \gamma_I)\right) \\
&= \mathrm{Tr}(\gamma_L^2) + \mathrm{Tr}(\gamma_I^2) - 2\mathrm{Tr}(\gamma_L\gamma_I) \\
&= \mathrm{Tr}(\gamma_L) + \mathrm{Tr}(\gamma_I) - 2\mathrm{Tr}(\gamma_L\gamma_I) \\
&= 2\left(\mathrm{Tr}(\gamma_L) - \mathrm{Tr}(\gamma_L\gamma_I)\right) \\
&= 2\mathrm{Tr}(\gamma_L(1-\gamma_I)).
\end{aligned}$$

□

2.3 Conforming discretizations

We consider conforming approximations of problem (2.1) in a space $V_h \subset D(A^{1/2})$. The approximate k-th eigenpair $(\varphi_{kh}, \lambda_{kh}) \in V_h \times \mathbb{R}_+$ is such that $(\varphi_{kh}, \varphi_{jh}) = \delta_{kj}$, $1 \le k, j \le \dim V_h$, and satisfies

$$(A^{1/2}\varphi_{kh}, A^{1/2}v_h) = \lambda_{kh}(\varphi_{kh}, v_h) \qquad \forall v_h \in V_h. \tag{2.16}$$

We number the approximate eigenvalues in increasing order, that is $0 < \lambda_{1h} \le \lambda_{2h} \le \ldots \le \lambda_{\dim V_h h}$, while counting multiplicities. Again, an immediate consequence of (2.16) is

$$\|A^{1/2}\varphi_{kh}\|^2 = \lambda_{kh}, \qquad \forall 1 \le k \le \dim V_h, \tag{2.17}$$

and, since the approximation is conforming, there holds

$$\lambda_k \le \lambda_{kh} \qquad \forall 1 \le k \le \dim V_h. \tag{2.18}$$

The *approximate eigenvalues* we are interested in are denoted by $(\lambda_{mh}, \ldots, \lambda_{Mh})$, where of course we suppose $\dim V_h \ge M$, and a corresponding set of orthonormal *approximate eigenvectors* by

$$\Phi_h := (\varphi_{mh}, \ldots, \varphi_{Mh}). \tag{2.19}$$

The *approximate density matrix* is then defined by

$$\forall v \in \mathcal{H}, \quad \gamma_h v := \sum_{i=m}^{M} (v, \varphi_{ih})\varphi_{ih}. \tag{2.20}$$

Remark 2.3 (Discrete gap condition)**.** *If the discrete gap condition $\lambda_{(m-1)h} < \lambda_{mh}$ if $m > 1$ and $\lambda_{Mh} < \lambda_{(M+1)h}$ is fulfilled, then the approximate density matrix γ_h is uniquely defined. Note that this uniqueness is not needed at this point, but is a consequence of Assumption 4.2 that is required starting from the upcoming Theorem 4.5 onward.*

Like the density matrix γ^0, the approximate density matrix γ_h is an orthogonal projector, hence $\gamma_h^2 = \gamma_h$, and there also holds $\mathrm{Tr}(\gamma_h) = \mathrm{Tr}(\gamma_h^2) = J$. We will measure the error between the exact and approximate density matrices using the quantities

$$\|\gamma^0 - \gamma_h\|_{\mathfrak{S}_2(\mathcal{H})} \quad \text{and} \quad \|A^{1/2}(\gamma^0 - \gamma_h)\|_{\mathfrak{S}_2(\mathcal{H})}.$$

The second quantity is indeed justified as we have:

Lemma 2.4 (Operators $A^{1/2}\gamma^0$ and $A^{1/2}\gamma_h$)**.** *Let the density matrices γ^0 and γ_h be respectively defined by (2.7) and (2.20). Then there holds $A^{1/2}\gamma^0$, $A^{1/2}\gamma_h \in \mathfrak{S}_2(\mathcal{H})$ and*

$$\|A^{1/2}\gamma^0\|^2_{\mathfrak{S}_2(\mathcal{H})} = \sum_{i=m}^{M} \lambda_i, \quad \|A^{1/2}\gamma_h\|^2_{\mathfrak{S}_2(\mathcal{H})} = \sum_{i=m}^{M} \lambda_{ih}. \tag{2.21}$$

Proof. Taking in (2.10) $e_k = \varphi_k^0$, since $\gamma^0 \varphi_k^0 = \varphi_k^0$ if $m \le k \le M$ and 0 otherwise, and employing (2.5),

$$\|A^{1/2}\gamma^0\|^2_{\mathfrak{S}_2(\mathcal{H})} = \sum_{k\ge 1} \|A^{1/2}\gamma^0\varphi_k^0\|^2 = \sum_{k=m}^{M} \|A^{1/2}\varphi_k^0\|^2 = \sum_{k=m}^{M} \lambda_k.$$

The result for $\|A^{1/2}\gamma_h\|_{\mathfrak{S}_2(\mathcal{H})}$ is obtained similarly upon completing φ_{kh} to an orthonormal basis of $\mathcal{H}$ used in (2.10) and employing (2.17). □

The following equalities will be useful in the upcoming analysis:

Lemma 2.5 (Orthogonal projector and Hilbert–Schmidt norm)**.** *Let γ_J be the orthogonal projector of rank $J = M - m + 1$ onto* $\mathrm{Span}\{\varphi_m^J, \ldots, \varphi_M^J\}$*, where $\varphi_m^J, \ldots, \varphi_M^J \in \mathcal{H}$ are orthonormal. Completing $\{\varphi_i^J\}_{i=m,\ldots,M}$ to an orthonormal basis of $\mathcal{H}$ denoted by $\{\varphi_i^J\}_{i\ge1}$, there holds*

$$\forall v \in \mathcal{H}, \quad \|\gamma_J v\|^2 = \sum_{i\ge1} \left|\left(\gamma_J v, \varphi_i^J\right)\right|^2 = \sum_{i=m}^{M} \left|\left(v, \varphi_i^J\right)\right|^2. \tag{2.22}$$

If in addition $\varphi_m^J, \ldots, \varphi_M^J \in D(A^{1/2})$, then there holds

$$\begin{aligned}\|A^{1/2}\gamma_J\|^2_{\mathfrak{S}_2(\mathcal{H})} &= \sum_{i\ge1} \|A^{1/2}\gamma_J\varphi_i^J\|^2 = \sum_{i=m}^{M} \|A^{1/2}\varphi_i^J\|^2 = \sum_{k\ge1}\sum_{i=m}^{M} \left|\left(A^{1/2}\varphi_i^J, \varphi_k^0\right)\right|^2 \\ &= \sum_{k\ge1}\sum_{i=m}^{M} \lambda_k \left|\left(\varphi_i^J, \varphi_k^0\right)\right|^2.\end{aligned} \tag{2.23}$$

Proof. Result (2.22) follows as in (2.2) and since γ_J is a projector. For the other claim, let us first note that $A^{1/2}\gamma_J \in \mathcal{L}(\mathcal{H})$, since $\mathrm{Ran}(\gamma_J) \subset D(A^{1/2})$. Then, (2.23) is a consequence of (2.10), (2.2), and (2.3a). □

For some of the results presented in the following, we will need to assume that the approximate eigenvectors are not orthogonal to the exact ones:

Assumption 2.6 (Non-orthogonality of the exact and approximate eigenspaces)**.** *There holds*

$$\forall v \in \mathrm{Span}\{\varphi_m^0, \ldots, \varphi_M^0\} \backslash \{0\}, \quad \|\gamma_h v\| \neq 0.$$

This assumption guarantees that every exact eigenvector is not orthogonal to the whole space spanned by the approximate eigenvectors. Note that this assumption, which in practice cannot be easily checked, is not needed for the first upper bound (4.4) below, which is used in Section 5.1 for finite element discretizations in Case II and in the planewave discretization in Section 5.2. However, in Case I in the finite element discretization, we prefer to use the improved bound based on (4.6), which requires this assumption. Assumption 2.6 is also needed to show an equivalence between eigenvectors and density matrix errors (Lemma 3.3), as well as to derive a lower bound for the density matrix error (Theorem 4.6).

3 Density matrix error and residuals

We develop in this section the links between the eigenvector errors and the density matrix errors. We will also define the residual and its dual norm, both for single and for cluster eigenpairs.

3.1 Eigenvector error and density matrix error equivalence

Since there is a choice in the (exact and approximate) eigenvectors of A, in particular for multiple eigenvalues, the approximate eigenvectors Φ_h might be far from the exact ones Φ^0 individually, which is measured in the energy norm $\|A^{1/2}(\Phi^0 - \Phi_h)\|$, while the density matrices and the eigenvalues are very close, or even equal. Traditionally, though, the error estimates are presented for the eigenvectors. Therefore, we first show that, given the exact eigenvectors Φ^0, there exists a choice of approximate eigenvectors $\Phi_h^0 =$

$(\varphi_{mh}^0, \dots, \varphi_{Mh}^0)$ constructed from Φ_h for which the error in the energy norm $\|A^{1/2}(\Phi^0 - \Phi_h^0)\|$ is equivalent to $\|A^{1/2}(\gamma^0 - \gamma_h)\|_{\mathfrak{S}_2(\mathcal{H})}$. This is valid under the sole assumption that the two eigenspaces are not orthogonal with respect to the $\mathcal{H}$ scalar product, as presented in Assumption 2.6. We also show that the density matrix error $\|A^{1/2}(\gamma^0 - \gamma_h)\|_{\mathfrak{S}_2(\mathcal{H})}$ can be easily expressed in terms of the eigenvectors.

Let us define the following unitary-transformed approximate eigenvectors by

$$\Phi_h^0 := (\varphi_{mh}^0, \dots, \varphi_{Mh}^0) := \mathrm{argmin}_{\mathbf{U} \in O(J)} \|\mathbf{U}\Phi_h - \Phi^0\|, \tag{3.1}$$

where we recall that $O(J)$ denotes the group of orthogonal matrices of order J. From [8, Lemma 4.3], the minimization problem (3.1) has a unique solution and therefore Φ_h^0 is well defined as soon as Assumption 2.6 is satisfied. Note that from this definition and the fact that the approximate eigenvectors are orthonormal, the rotated approximate eigenvectors are also orthonormal. Also, the approximate density matrix γ_h given by (2.20) can be equivalently written in terms of the rotated eigenvectors Φ_h^0 as

$$\forall v \in \mathcal{H}, \quad \gamma_h v = \sum_{i=m}^{M} (v, \varphi_{ih}^0)\varphi_{ih}^0. \tag{3.2}$$

To relate $\|A^{1/2}(\gamma^0 - \gamma_h)\|_{\mathfrak{S}_2(\mathcal{H})}$ to the energy norm of the eigenvector errors $\|A^{1/2}(\Phi^0 - \Phi_h^0)\|$, we first need a preliminary lemma, expressing these two quantities in terms of the exact and approximate eigenvalues and the eigenvectors of the operator A.

Lemma 3.1 (Link between eigenvalue and eigenvector errors). *Let Assumption 2.1 hold and let the density matrices γ^0 and γ_h be respectively defined by (2.7) and (2.20) and the eigenvectors Φ^0 and Φ_h^0 by (2.6) and (3.1). Then there holds*

$$\|A^{1/2}(\gamma^0 - \gamma_h)\|_{\mathfrak{S}_2(\mathcal{H})}^2 = \sum_{i=m}^{M} (\lambda_{ih} - \lambda_i) + 2\sum_{i=m}^{M} \lambda_i \|(1 - \gamma_h)\varphi_i^0\|^2, \tag{3.3}$$

and

$$\|A^{1/2}(\Phi^0 - \Phi_h^0)\|^2 = \sum_{i=m}^{M} (\lambda_{ih} - \lambda_i) + \sum_{i=m}^{M} \lambda_i \|\varphi_i^0 - \varphi_{ih}^0\|^2. \tag{3.4}$$

Remark 3.2. *Note that Assumption 2.1 is needed for γ^0 and Φ^0 to be well-defined, but is not needed stricto sensu in the proof of Lemma 3.1, which is still valid in the case where there is no discrete or continuous gap between the eigenvalue cluster and the rest of the spectrum. More precisely, this Lemma would hold for any orthogonal projectors γ^0, γ_h of rank $M - m + 1$ and associated orthogonal vectors Φ^0 and Φ_h^0 such that $\mathrm{Ran}(\gamma^0) = \mathrm{Span}(\Phi^0)$, $\mathrm{Ran}(\gamma_h) = \mathrm{Span}(\Phi_h)$, and satisfying for $i = m, \dots, M$, $(A^{1/2}\varphi_i^0, A^{1/2}v) = \lambda_i(\varphi_i^0, v)$ for all $v \in D(A^{1/2})$, and $(A^{1/2}\varphi_{ih}, A^{1/2}v_h) = \lambda_{ih}(\varphi_{ih}, v_h)$ for all $v_h \in V_h$. For simplicity of presentation, Lemma 3.1 is kept without these generalizations.*

Proof. To show (3.3), we first expand $\|A^{1/2}(\gamma^0 - \gamma_h)\|_{\mathfrak{S}_2(\mathcal{H})}^2$ and then use (2.21) to obtain

$$\begin{aligned}
\|A^{1/2}(\gamma^0 - \gamma_h)\|_{\mathfrak{S}_2(\mathcal{H})}^2 &= \|A^{1/2}\gamma^0\|_{\mathfrak{S}_2(\mathcal{H})}^2 - 2(A^{1/2}\gamma^0, A^{1/2}\gamma_h)_{\mathfrak{S}_2(\mathcal{H})} + \|A^{1/2}\gamma_h\|_{\mathfrak{S}_2(\mathcal{H})}^2 \\
&= \sum_{i=m}^{M} \lambda_i - 2\sum_{i \geq 1} (A^{1/2}\gamma^0\varphi_i^0, A^{1/2}\gamma_h\varphi_i^0) + \sum_{i=m}^{M} \lambda_{ih} \\
&= \sum_{i=m}^{M} \lambda_i - 2\sum_{i=m}^{M} (A^{1/2}\varphi_i^0, A^{1/2}\gamma_h\varphi_i^0) + \sum_{i=m}^{M} \lambda_{ih},
\end{aligned} \tag{3.5}$$

where we have also used that $\gamma^0\varphi_i^0 = \varphi_i^0$ if $m \leq i \leq M$ and 0 otherwise. Employing (2.4) leads to

$$\begin{aligned}
\|A^{1/2}(\gamma^0 - \gamma_h)\|_{\mathfrak{S}_2(\mathcal{H})}^2 &= \sum_{i=m}^{M} (\lambda_i + \lambda_{ih}) - 2\sum_{i=m}^{M} \lambda_i \left(\varphi_i^0, \gamma_h\varphi_i^0\right) \\
&= \sum_{i=m}^{M} (\lambda_{ih} - \lambda_i) + 2\sum_{i=m}^{M} \lambda_i \left(1 - (\varphi_i^0, \gamma_h\varphi_i^0)\right).
\end{aligned} \tag{3.6}$$

Writing $1 = (\varphi_i^0, \varphi_i^0)$ and observing that $(1-\gamma_h) = (1-\gamma_h)^2$, we obtain

$$\|A^{1/2}(\gamma^0 - \gamma_h)\|^2_{\mathfrak{S}_2(\mathcal{H})} = \sum_{i=m}^{M}(\lambda_{ih} - \lambda_i) + 2\sum_{i=m}^{M} \lambda_i \left(\varphi_i^0, (1-\gamma_h)^2 \varphi_i^0\right),$$

which leads to (3.3), using that $(1-\gamma_h)$ is self-adjoint.

To show (3.4), we complete Φ_h^0 to an orthonormal basis $(\varphi_{ih}^0)_{i\geq 1}$ of $\mathcal{H}$, employ this basis in (2.10), and use (2.21),

$$\sum_{i=m}^{M} \|A^{1/2}\varphi_{ih}^0\|^2 = \sum_{i\geq 1} \|A^{1/2}\gamma_h \varphi_{ih}^0\|^2 = \|A^{1/2}\gamma_h\|^2_{\mathfrak{S}_2(\mathcal{H})} = \sum_{i=m}^{M} \lambda_{ih}.$$

We next use definition (2.15) together with (2.5) and (2.4) to see that

$$\begin{aligned}\|A^{1/2}(\Phi^0 - \Phi_h^0)\|^2 &= \sum_{i=m}^{M} \left[\|A^{1/2}\varphi_i^0\|^2 - 2(A^{1/2}\varphi_i^0, A^{1/2}\varphi_{ih}^0) + \|A^{1/2}\varphi_{ih}^0\|^2\right] \\ &= \sum_{i=m}^{M} \left[\lambda_{ih} + \lambda_i - 2\lambda_i(\varphi_i^0, \varphi_{ih}^0)\right].\end{aligned}$$

Using that for all $i = m, \ldots, M$, φ_{ih}^0 as well as φ_i^0 are of norm 1 leads to

$$\begin{aligned}\|A^{1/2}(\Phi^0 - \Phi_h^0)\|^2 &= \sum_{i=m}^{M}(\lambda_{ih} - \lambda_i) + 2\sum_{i=m}^{M} \lambda_i \left(1 - (\varphi_i^0, \varphi_{ih}^0)\right) \\ &= \sum_{i=m}^{M}(\lambda_{ih} - \lambda_i) + \sum_{i=m}^{M} \lambda_i \|\varphi_i^0 - \varphi_{ih}^0\|^2,\end{aligned}$$

which concludes the proof. □

The following lemma relates the errors on the density matrix to the errors on the eigenvectors.

Lemma 3.3 (Link between density matrix and eigenvector errors)**.** *Let the assumptions of Lemma 3.1 hold, together with Assumption 2.6. Then*

$$\frac{1}{\sqrt{2}}\|\gamma^0 - \gamma_h\|_{\mathfrak{S}_2(\mathcal{H})} \leq \|\Phi^0 - \Phi_h^0\| \leq \|\gamma^0 - \gamma_h\|_{\mathfrak{S}_2(\mathcal{H})}. \tag{3.7}$$

Moreover,

$$\begin{aligned}\frac{1}{\sqrt{2}}\|A^{1/2}(\gamma^0 - \gamma_h)\|_{\mathfrak{S}_2(\mathcal{H})} &\leq \|A^{1/2}(\Phi^0 - \Phi_h^0)\| \\ &\leq \left(1 + \frac{\lambda_M}{4\lambda_m}\|\gamma^0 - \gamma_h\|^2_{\mathfrak{S}_2(\mathcal{H})}\right)^{1/2} \|A^{1/2}(\gamma^0 - \gamma_h)\|_{\mathfrak{S}_2(\mathcal{H})}, \end{aligned}\tag{3.8}$$

and in particular

$$\|A^{1/2}(\Phi^0 - \Phi_h^0)\| \leq \left(1 + \frac{J\lambda_M}{\lambda_m}\right)^{1/2} \|A^{1/2}(\gamma^0 - \gamma_h)\|_{\mathfrak{S}_2(\mathcal{H})}. \tag{3.9}$$

A proof for (3.7) in the case $m = 1$ can be found in Lemma 2.3 of [11], whereas (3.8) is proved in [20, Lemma 3.1] in a similar setting. For the sake of completeness, we present the proof of (3.8) in Appendix A in our specific setting. Using the (very) crude bound $\|\gamma^0 - \gamma_h\|^2_{\mathfrak{S}_2(\mathcal{H})} \leq 4J$, cf. (2.14), we immediately deduce (3.9) from (3.8). Using Lemma 3.3, it is possible to easily translate bounds expressed in terms of density matrices on bounds on the eigenvectors, as long as these eigenvectors are rotated correctly. In the rest of this paper, we will therefore focus on the estimation of density-matrix-based quantities only.

In terms of implementation, the natural outputs of an eigenvalue solver are often eigenvectors and not density matrices (note, however, that some algorithms directly compute density matrices using Cauchy's formula $\gamma^0 = \frac{1}{2i\pi}\oint_{\mathcal{C}}(z-A)^{-1}\,dz$, where $\mathcal{C}$ is a contour in the complex plane enclosing the eigenvalues $(\lambda_m,\dots,\lambda_M)$). The practical computation of $\|A^{1/2}(\gamma_h-\gamma^0)\|_{\mathfrak{S}_2(\mathcal{H})}$ can easily be done in terms of the eigenvectors since, using (3.5) and (2.20),

$$\|A^{1/2}(\gamma^0-\gamma_h)\|^2_{\mathfrak{S}_2(\mathcal{H})} = \sum_{i=m}^{M}\left[\|A^{1/2}\varphi_i^0\|^2 - 2\sum_{j=m}^{M}(A^{1/2}\varphi_i^0, A^{1/2}\varphi_{jh})(\varphi_{jh},\varphi_i^0) + \|A^{1/2}\varphi_{ih}\|^2\right].$$

To conclude this section, Table 3 presents a summary of the principal mathematical objects dealt with in the analysis, in the general case of a given self-adjoint operator A, as well as for the Laplace operator $-\Delta$ with Dirichlet boundary conditions and a Schrödinger operator $-\Delta+V$ on a cubic box with periodic boundary conditions, for which we present numerical simulations below in Section 6.

Table 3: Mathematical objects used in this analysis. For the Laplace operator case, Ω is a bounded Lipschitz domain of $\mathbb{R}^d$, $d\ge 1$. For the periodic Schrödinger operator case, $\Omega=[0,L)^d$, $d\ge1$, $L^2_{\#}(\Omega) := \{v\in L^2_{\rm loc}(\mathbb{R}^d)|\, v\, L\mathbb{Z}^d\text{-periodic}\}$, $H^s_{\#}(\Omega)=\{v\in H^s_{\rm loc}(\mathbb{R}^d)|\, v\, L\mathbb{Z}^d\text{-periodic}\}$, $s=1,2$, and $V\in L^\infty_{\#}(\Omega)$, $V\ge1$, $\gamma\in\mathcal{L}(\mathcal{H})$ is a rank-J orthogonal projector such that $\mathrm{Ran}(\gamma)=\mathrm{Span}\{\varphi_m^J,\dots,\varphi_M^J\}$, where $\varphi_m^J,\dots,\varphi_M^J\in D(A^{1/2})$ are orthonormal in $\mathcal{H}$.

	General framework	Laplace operator	Schrödinger operator
Hilbert space	$\mathcal{H}$	$L^2(\Omega)$	$L^2_{\#}(\Omega)$
Operator	A	$-\Delta$	$-\Delta+V$
Domain	$D(A)$	$\{v\in H^1_0(\Omega)\mid \Delta v\in L^2(\Omega)\}$	$H^2_{\#}(\Omega)$
Form domain	$D(A^{1/2})$	$H^1_0(\Omega)$	$H^1_{\#}(\Omega)$
Norm of v	$\|v\|$	$\left(\int_\Omega \|v\|^2\right)^{1/2}$	$\left(\int_\Omega \|v\|^2\right)^{1/2}$
Energy norm of v	$\|A^{1/2}v\|$	$\left(\int_\Omega \|\nabla v\|^2\right)^{1/2}$	$\left(\int_\Omega(\|\nabla v\|^2+V\|v\|^2)\right)^{1/2}$
Energy norm of γ	$\|A^{1/2}\gamma\|_{\mathfrak{S}_2(\mathcal{H})} = \left\{\sum_{i=m}^M \|A^{1/2}\varphi_i^J\|^2\right\}^{1/2}$	$\left\{\sum_{i=m}^M\int_\Omega\|\nabla\varphi_i^J\|^2\right\}^{1/2}$	$\left\{\sum_{i=m}^M\int_\Omega(\|\nabla\varphi_i^J\|^2 + V\|\varphi_i^J\|^2)\right\}^{1/2}$

3.2 Residuals and their dual norms

Classically, the derivation of *a posteriori* error estimates is based on the notion of the residual and its dual norm. In our setting, we can define the residual for a single eigenpair as follows, where $D(A^{1/2})'$ stands for the dual of $D(A^{1/2})$.

Definition 3.4 (Single eigenpair residual and its dual norm). *For any eigenpair $(\varphi_{ih},\lambda_{ih})\in V_h\times\mathbb{R}_+$ of (2.16), $m\le i\le M$, define the* residual $\mathrm{Res}(\varphi_{ih},\lambda_{ih})\in D(A^{1/2})'$ *by*

$$\langle \mathrm{Res}(\varphi_{ih},\lambda_{ih}),v\rangle_{D(A^{1/2})',D(A^{1/2})} := \lambda_{ih}(\varphi_{ih},v) - \left(A^{1/2}\varphi_{ih},A^{1/2}v\right)\qquad \forall v\in D(A^{1/2}). \tag{3.10a}$$

Its dual norm *is then*

$$\|\mathrm{Res}(\varphi_{ih},\lambda_{ih})\|_{D(A^{1/2})'} := \sup_{\substack{v\in D(A^{1/2})\\ \|A^{1/2}v\|=1}} \langle \mathrm{Res}(\varphi_{ih},\lambda_{ih}),v\rangle_{D(A^{1/2})',D(A^{1/2})}. \tag{3.10b}$$

To consider the error on the eigenvalue cluster in its globality, we now define a cluster residual, which is an operator measuring the error with respect to the equation for the whole targeted eigenspace. Note that this operator depends on the approximate density matrix γ_h only, and not on the exact density matrix γ^0, exactly as the single eigenpair residuals depend on the approximate eigenpairs only.

Definition 3.5 (Cluster residual). *For γ_h defined in (2.20), define the* cluster residual $\mathrm{Res}(\gamma_h)\in\mathcal{L}(\mathcal{H})$ *by*

$$\mathrm{Res}(\gamma_h) := A^{1/2}\gamma_h - A^{-1/2}\left(A^{1/2}\gamma_h\right)^\dagger A^{1/2}\gamma_h. \tag{3.11}$$

Note that $\mathrm{Res}(\gamma_h)$ is a finite-rank operator of $\mathcal{L}(\mathcal{H})$, as γ_h is finite-rank, $A^{1/2}\gamma_h$ is bounded by Lemma 2.4, and $A^{-1/2} \in \mathcal{L}(\mathcal{H})$. The choice of this definition is motivated by the following remark.

Remark 3.6 (Strong form of the cluster residual). *When the approximation space in (2.16) satisfies $V_h \subset D(A)$, which is the case for planewave discretizations of periodic Schrödinger operators, but not for Lagrange finite element discretizations of the Laplace operator, one could first define a cluster residual in $\mathcal{L}(\mathcal{H})$ by*

$$\mathrm{Res}^h_{m,M} := (1-\gamma_h)A\gamma_h.$$

The corresponding operator for the exact density matrix $(1-\gamma^0)A\gamma^0$ is indeed zero, as $\mathrm{Ran}(A\gamma^0) = \mathrm{Ran}(\gamma^0) \subset \mathrm{Ker}(1-\gamma^0)$.

Then $A^{-1/2}\mathrm{Res}^h_{m,M} = \mathrm{Res}(\gamma_h)$ and

$$\|\mathrm{Res}(\gamma_h)\|^2_{\mathfrak{S}_2(\mathcal{H})} = \|A^{-1/2}\mathrm{Res}^h_{m,M}\|^2_{\mathfrak{S}_2(\mathcal{H})} = \mathrm{Tr}\big(\mathrm{Res}^h_{m,M}{}^\dagger A^{-1}\mathrm{Res}^h_{m,M}\big).$$

In the case where $V_h \subsetneq D(A)$, the analog of $\mathrm{Res}^h_{m,M}$ is $(A^{1/2}(1-\gamma_h))^\dagger A^{1/2}\gamma_h = (A^{1/2})^\dagger A^{1/2}\gamma_h - (A^{1/2}\gamma_h)^\dagger A^{1/2}\gamma_h$, which multiplied on the left by $A^{-1/2}$ is well-defined.

We now show that the definitions of the single eigenpair and cluster residuals match in the sense that the sum of the dual norms of the single eigenpair residuals is equal to the Hilbert–Schmidt norm of the cluster residual. Therefore, we will be able to estimate the individual dual norms (3.10b) by existing tools in Section 5 below.

The following preliminary lemma relates the residual to the exact and approximate eigenpairs.

Lemma 3.7 (Residual expansion). *There holds*

$$\forall s \le 0, \quad \|A^s\mathrm{Res}(\gamma_h)\|^2_{\mathfrak{S}_2(\mathcal{H})} = \sum_{k\ge 1}\sum_{i=m}^{M} \lambda_k^{2s-1}(\lambda_k - \lambda_{ih})^2 \left|\big(\varphi_{ih}, \varphi_k^0\big)\right|^2. \tag{3.12}$$

Proof. First note that $A^s\mathrm{Res}(\gamma_h) \in \mathcal{L}(\mathcal{H})$ for $s \le 0$. Using (2.20) and since A^s is self-adjoint, (2.10) yields

$$\begin{aligned}
\|A^s\mathrm{Res}(\gamma_h)\|^2_{\mathfrak{S}_2(\mathcal{H})} &= \sum_{i=m}^{M} \|A^s\mathrm{Res}(\gamma_h)\varphi_{ih}\|^2 = \sum_{i=m}^{M} \big(\mathrm{Res}(\gamma_h)\varphi_{ih}, A^{2s}\mathrm{Res}(\gamma_h)\varphi_{ih}\big) \\
&= \sum_{i=m}^{M} \Big[\big(A^{1/2}\varphi_{ih}, A^{2s}A^{1/2}\varphi_{ih}\big) - 2\big(A^{1/2}\varphi_{ih}, A^{2s}A^{-1/2}\big(A^{1/2}\gamma_h\big)^\dagger A^{1/2}\varphi_{ih}\big) \\
&\quad + \big(A^{-1/2}\big(A^{1/2}\gamma_h\big)^\dagger A^{1/2}\varphi_{ih}, A^{2s}A^{-1/2}\big(A^{1/2}\gamma_h\big)^\dagger A^{1/2}\varphi_{ih}\big)\Big] \\
&= \sum_{i=m}^{M} \Big[\big(A^{1/2}\varphi_{ih}, A^{2s}A^{1/2}\varphi_{ih}\big) - 2\big(A^{1/2}\gamma_h A^{2s}\varphi_{ih}, A^{1/2}\varphi_{ih}\big) \\
&\quad + \big(\big(A^{1/2}\gamma_h\big)^\dagger A^{1/2}\varphi_{ih}, A^{2s-1}\big(A^{1/2}\gamma_h\big)^\dagger A^{1/2}\varphi_{ih}\big)\Big] \\
&=: \sum_{i=m}^{M} [T_{1i} + T_{2i} + T_{3i}].
\end{aligned}$$

We now treat the three terms separately while expanding the operators A^{2s}, $A^{1/2}$, and A^{2s-1} on the eigenvectors using (2.3b). This gives, noting that $A^{1/2}$ is self-adjoint,

$$T_{1i} = \left(A^{1/2}\varphi_{ih}, \sum_{k\ge 1}\lambda_k^{2s}\big(A^{1/2}\varphi_{ih}, \varphi_k^0\big)\varphi_k^0\right) = \sum_{k\ge 1}\lambda_k^{2s}\left|\big(A^{1/2}\varphi_{ih}, \varphi_k^0\big)\right|^2 = \sum_{k\ge 1}\lambda_k^{2s+1}\left|\big(\varphi_{ih}, \varphi_k^0\big)\right|^2$$

and similarly, also using (2.16),

$$
\begin{aligned}
T_{2i} &= -2\left(A^{1/2}\gamma_h \sum_{k\geq 1} \lambda_k^{2s}\left(\varphi_{ih}, \varphi_k^0\right)\varphi_k^0, A^{1/2}\varphi_{ih}\right) \\
&= -2\sum_{k\geq 1} \lambda_k^{2s}\left(\varphi_{ih}, \varphi_k^0\right)\left(A^{1/2}\gamma_h\varphi_k^0, A^{1/2}\varphi_{ih}\right) \\
&= -2\sum_{k\geq 1} \lambda_k^{2s}\left(\varphi_{ih}, \varphi_k^0\right)\lambda_{ih}\left(\varphi_{ih}, \gamma_h\varphi_k^0\right) = -2\sum_{k\geq 1}\lambda_k^{2s}\lambda_{ih}\left|\left(\varphi_{ih}, \varphi_k^0\right)\right|^2.
\end{aligned}
$$

Finally, relying again on (2.16),

$$
\begin{aligned}
T_{3i} &= \left(\left(A^{1/2}\gamma_h\right)^\dagger A^{1/2}\varphi_{ih}, \sum_{k\geq 1}\lambda_k^{2s-1}\left(\left(A^{1/2}\gamma_h\right)^\dagger A^{1/2}\varphi_{ih}, \varphi_k^0\right)\varphi_k^0\right) \\
&= \sum_{k\geq 1}\lambda_k^{2s-1}\left|\left(A^{1/2}\varphi_{ih}, A^{1/2}\gamma_h\varphi_k^0\right)\right|^2 = \sum_{k\geq 1}\lambda_k^{2s-1}\lambda_{ih}^2\left|\left(\varphi_{ih}, \varphi_k^0\right)\right|^2.
\end{aligned}
$$

Developing the square in (3.12) finishes the proof. □

We can now state the correspondence between the cluster residual and the single eigenpair residuals.

Lemma 3.8 (Relation between cluster and single eigenpair residuals)**.** *There holds*

$$
\|\mathrm{Res}(\gamma_h)\|^2_{\mathfrak{S}_2(\mathcal{H})} = \sum_{i=m}^{M}\|\mathrm{Res}(\varphi_{ih}, \lambda_{ih})\|^2_{D(A^{1/2})'}.
$$

Proof. For each $m \leq i \leq M$, define the Riesz representation of the residual $\mathscr{r}_{(ih)} \in D(A^{1/2})$ such that

$$
\left(A^{1/2}\mathscr{r}_{(ih)}, A^{1/2}v\right) = \langle \mathrm{Res}(\varphi_{ih}, \lambda_{ih}), v\rangle_{D(A^{1/2})', D(A^{1/2})} \qquad \forall v \in D(A^{1/2}). \tag{3.13}
$$

Consequently,

$$
\|A^{1/2}\mathscr{r}_{(ih)}\| = \|\mathrm{Res}(\varphi_{ih}, \lambda_{ih})\|_{D(A^{1/2})'}.
$$

Moreover, using that $A^{-1/2}$ is self-adjoint and $A^{-1/2}A^{1/2} = 1$, we see from (3.10a) that

$$
\left(A^{1/2}\mathscr{r}_{(ih)}, A^{1/2}v\right) = \lambda_{ih}\left(A^{-1/2}\varphi_{ih}, A^{1/2}v\right) - \left(A^{1/2}\varphi_{ih}, A^{1/2}v\right) \qquad \forall v \in D(A^{1/2}),
$$

so that

$$
A^{1/2}\mathscr{r}_{(ih)} = \lambda_{ih}A^{-1/2}\varphi_{ih} - A^{1/2}\varphi_{ih}.
$$

Expressing the norms related to $A^{-1/2}$, A^0, and $A^{1/2}$ via (2.3a), we conclude therefrom that

$$
\begin{aligned}
\|A^{1/2}\mathscr{r}_{(ih)}\|^2 &= \lambda_{ih}^2\|A^{-1/2}\varphi_{ih}\|^2 - 2\lambda_{ih}\left(\varphi_{ih}, \varphi_{ih}\right) + \|A^{1/2}\varphi_{ih}\|^2 \\
&= \lambda_{ih}^2\sum_{k\geq 1}\lambda_k^{-1}|(\varphi_{ih}, \varphi_k^0)|^2 - 2\lambda_{ih}\sum_{k\geq 1}|(\varphi_{ih}, \varphi_k^0)|^2 + \sum_{k\geq 1}\lambda_k|(\varphi_{ih}, \varphi_k^0)|^2,
\end{aligned} \tag{3.14}
$$

and the assertion follows using (3.12) with $s = 0$. □

4 Error equivalences

The framework is now ready to prove *a posteriori* estimates for $\|A^{1/2}(\gamma^0 - \gamma_h)\|_{\mathfrak{S}_2(\mathcal{H})}$ and the sum of the eigenvalues errors in terms of the cluster residual $\mathrm{Res}(\gamma_h)$. These results extend [9, Theorems 3.4, 3.5 and Lemmas 3.1, 3.2] to the case of eigenvalue clusters, and especially cover the case of degenerate eigenvalues.

4.1 Eigenvalue error equivalence

We first show how to estimate the sum of the eigenvalues errors in terms of errors on the density matrix.

Theorem 4.1 (Eigenvalue bounds)**.** *Let Assumption 2.1 hold and let the density matrices γ^0 and γ_h be respectively defined by* (2.7) *and* (2.20)*. Then*

$$\|A^{1/2}(\gamma^0 - \gamma_h)\|^2_{\mathfrak{S}_2(\mathcal{H})} - \lambda_M \|\gamma^0 - \gamma_h\|^2_{\mathfrak{S}_2(\mathcal{H})} \leq \sum_{i=m}^{M} (\lambda_{ih} - \lambda_i) \leq \|A^{1/2}(\gamma^0 - \gamma_h)\|^2_{\mathfrak{S}_2(\mathcal{H})}. \tag{4.1}$$

Proof. We start from (3.3), i.e.,

$$\|A^{1/2}(\gamma^0 - \gamma_h)\|^2_{\mathfrak{S}_2(\mathcal{H})} = \sum_{i=m}^{M} (\lambda_{ih} - \lambda_i) + 2 \sum_{i=m}^{M} \lambda_i \|(1 - \gamma_h)\varphi_i^0\|^2.$$

Noting that $2\sum_{i=m}^{M} \lambda_i \|(1-\gamma_h)\varphi_i^0\|^2 \geq 0$ easily proves the right-hand side of (4.1). Moreover, bounding the eigenvalues by the largest in the sum, expressing the sum of the projected eigenvectors as a trace, and using $(1 - \gamma_h)^2 = 1 - \gamma_h$ yields

$$2\sum_{i=m}^{M} \lambda_i \|(1 - \gamma_h)\varphi_i^0\|^2 \leq 2\lambda_M \sum_{i=m}^{M} \|(1 - \gamma_h)\varphi_i^0\|^2 = 2\lambda_M \mathrm{Tr}(\gamma^0(1 - \gamma_h)) = \lambda_M \|\gamma^0 - \gamma_h\|^2_{\mathfrak{S}_2(\mathcal{H})},$$

where we have used Lemma 2.2 with γ^0 and γ_h for the last equality. The left-hand side of (4.1) follows. $\square$

4.2 Eigenvector error equivalence

We next estimate the energy norm of the density matrix error in terms of the Hilbert–Schmidt norm of the cluster residual $\mathrm{Res}(\gamma_h)$. We henceforth often need the following assumption, in addition to Assumption 2.1:

Assumption 4.2 (Continuous–discrete gap conditions)**.** *There exist $\underline{\lambda}_{M+1}$, and $\overline{\lambda}_{m-1}$ if $m > 1$, which we take as $\overline{\lambda}_{m-1} = \lambda_{(m-1)h}$, such that there holds*

$$\lambda_{m-1} \leq \overline{\lambda}_{m-1} < \lambda_{mh} \text{ when } m > 1, \qquad \lambda_{Mh} < \underline{\lambda}_{M+1} \leq \lambda_{M+1}.$$

For practical use, we usually proceed as follows.

Remark 4.3 (Verification of Assumptions 2.1 and 4.2, choice of $\underline{\lambda}_{M+1}$, and uniqueness of the discrete projector γ_h)**.** *Let $\underline{\lambda}_m$ be a guaranteed lower bound for λ_m and $\underline{\lambda}_{M+1}$ a guaranteed lower bound for λ_{M+1} (obtained by, e.g.in a finite element discretization, employing the nonconforming finite element method on a coarse mesh and using the technique presented [13, Theorem 3.2] or [34, formula (6)]). If these bounds are accurate enough, it follows from* (2.18) *that we can request*

$$\lambda_{m-1} \leq \lambda_{(m-1)h} < \underline{\lambda}_m \leq \lambda_m \leq \lambda_{mh}, \quad \text{when } m > 1, \tag{4.2}$$

$$\lambda_M \leq \lambda_{Mh} < \underline{\lambda}_{M+1} \leq \lambda_{M+1} \leq \lambda_{(M+1)h}, \tag{4.3}$$

so that: 1) Assumption 2.1 is satisfied; 2) Assumption 4.2 is satisfied with $\overline{\lambda}_{m-1} = \lambda_{(m-1)h}$ and hence the constants c_h and $\tilde{c}_h$ in (4.5) *and* (4.15) *will be well-defined; 3) the discrete gap condition of Remark 2.3 is satisfied and hence the discrete projector γ_h is uniquely defined.*

In view of Remark 4.3, we introduce the following assumption, which implies in particular Assumptions 2.1 and 4.2 as well as Remark 2.3.

Assumption 4.4 (Availability of accurate enough lower bounds for λ_m and λ_{M+1})**.** *We know two real numbers $\underline{\lambda}_m$ and $\underline{\lambda}_{M+1}$ such that the inequalities* (4.2)*–*(4.3) *are satisfied.*

Theorem 4.5 (Upper bounds for the density matrix error)**.** *Let Assumption 2.1 hold, let the density matrices γ^0 and γ_h be respectively defined by (2.7) and (2.20), and let the cluster residual* $\mathrm{Res}(\gamma_h)$ *be defined by (3.11). Then, there holds*

$$\|A^{1/2}(\gamma^0 - \gamma_h)\|^2_{\mathfrak{S}_2(\mathcal{H})} \leq \|\mathrm{Res}(\gamma_h)\|^2_{\mathfrak{S}_2(\mathcal{H})} + (\lambda_M + \lambda_{Mh})\|\gamma^0 - \gamma_h\|^2_{\mathfrak{S}_2(\mathcal{H})}. \tag{4.4}$$

Let in addition Assumptions 2.6 and 4.2 hold and set

$$c_h := \max\left[\left(\frac{\lambda_{mh}}{\overline{\lambda}_{m-1}} - 1\right)^{-1}, \left(1 - \frac{\lambda_{Mh}}{\underline{\lambda}_{M+1}}\right)^{-1}\right], \tag{4.5}$$

the first term in the max *being discarded for $m = 1$. Then there also holds that*

$$\|A^{1/2}(\gamma^0 - \gamma_h)\|^2_{\mathfrak{S}_2(\mathcal{H})} \leq 2{c_h}^2\|\mathrm{Res}(\gamma_h)\|^2_{\mathfrak{S}_2(\mathcal{H})} + \frac{\lambda_M}{2}\|\gamma^0 - \gamma_h\|^4_{\mathfrak{S}_2(\mathcal{H})}. \tag{4.6}$$

Proof. To show (4.4), let us decompose $\|A^{1/2}(\gamma^0 - \gamma_h)\|^2_{\mathfrak{S}_2(\mathcal{H})}$ using (2.10) as

$$\begin{aligned}
\|A^{1/2}(\gamma^0 - \gamma_h)\|^2_{\mathfrak{S}_2(\mathcal{H})} &= \mathrm{Tr}\big(\big(A^{1/2}(\gamma^0 - \gamma_h)\big)^\dagger A^{1/2}(\gamma^0 - \gamma_h)\big) \\
&= \mathrm{Tr}\big(\big(A^{1/2}(\gamma^0 - \gamma_h)\big)^\dagger A^{1/2}\gamma^0\big) + \mathrm{Tr}\big(\big(A^{1/2}(\gamma_h - \gamma^0)\big)^\dagger A^{1/2}\gamma_h\big) \\
&=: T_1 + T_2.
\end{aligned}$$

On the one hand, using definition (2.7) of γ^0 and (2.4),

$$T_1 = \sum_{i=m}^{M}(A^{1/2}\varphi_i^0, A^{1/2}(\gamma^0 - \gamma_h)\varphi_i^0) = \sum_{i=m}^{M}\lambda_i(\varphi_i^0, (\gamma^0 - \gamma_h)\varphi_i^0).$$

Since for $i = m, \ldots, M$, $(\varphi_i^0, (\gamma^0 - \gamma_h)\varphi_i^0) = \big(1 - \|\gamma_h\varphi_i^0\|^2\big) \geq 0$, we can bound the above expression via Lemma 2.2 as

$$T_1 \leq \lambda_M \sum_{i=m}^{M}(\varphi_i^0, (\gamma^0 - \gamma_h)\varphi_i^0) = \lambda_M \mathrm{Tr}(\gamma^0(1 - \gamma_h)) = \frac{\lambda_M}{2}\|\gamma^0 - \gamma_h\|^2_{\mathfrak{S}_2(\mathcal{H})}.$$

On the other hand, writing $1 = \big(A^{1/2}\gamma_h A^{-1/2}\big)^\dagger + (1 - \big(A^{1/2}\gamma_h A^{-1/2}\big)^\dagger)$, using the definition of the cluster residual (3.11), and employing (2.16), we obtain

$$\begin{aligned}
T_2 &= \mathrm{Tr}\big(\big(A^{1/2}(\gamma_h - \gamma^0)\big)^\dagger \big(A^{1/2}\gamma_h A^{-1/2}\big)^\dagger A^{1/2}\gamma_h\big) \\
&\quad + \mathrm{Tr}\big(\big(A^{1/2}(\gamma_h - \gamma^0)\big)^\dagger (1 - \big(A^{1/2}\gamma_h A^{-1/2}\big)^\dagger) A^{1/2}\gamma_h\big) \\
&= \sum_{i=m}^{M}\big(A^{1/2}\gamma_h(\gamma_h - \gamma^0)\varphi_{ih}, A^{1/2}\varphi_{ih}\big) + \mathrm{Tr}\big(\big(A^{1/2}(\gamma_h - \gamma^0)\big)^\dagger \mathrm{Res}(\gamma_h)\big) \\
&= \sum_{i=m}^{M}\lambda_{ih}(\varphi_{ih}, (\gamma_h - \gamma^0)\varphi_{ih}) + \mathrm{Tr}\big(\big(A^{1/2}(\gamma_h - \gamma^0)\big)^\dagger \mathrm{Res}(\gamma_h)\big).
\end{aligned}$$

Using this time that, for all $i = m, \ldots, M$, $(\varphi_{ih}, (\gamma_h - \gamma^0)\varphi_{ih}) = \big(1 - \|\gamma^0\varphi_{ih}\|^2\big) \geq 0$, Lemma 2.2, and (2.12), Young's inequality leads to

$$\begin{aligned}
T_2 &\leq \lambda_{Mh}\mathrm{Tr}(\gamma_h(\gamma_h - \gamma^0)) + \|A^{1/2}(\gamma^0 - \gamma_h)\|_{\mathfrak{S}_2(\mathcal{H})}\|\mathrm{Res}(\gamma_h)\|_{\mathfrak{S}_2(\mathcal{H})} \\
&\leq \frac{\lambda_{Mh}}{2}\|\gamma^0 - \gamma_h\|^2_{\mathfrak{S}_2(\mathcal{H})} + \frac{1}{2}\|A^{1/2}(\gamma^0 - \gamma_h)\|^2_{\mathfrak{S}_2(\mathcal{H})} + \frac{1}{2}\|\mathrm{Res}(\gamma_h)\|^2_{\mathfrak{S}_2(\mathcal{H})}.
\end{aligned}$$

Putting these contributions together, we get

$$\begin{aligned}
\|A^{1/2}(\gamma^0 - \gamma_h)\|^2_{\mathfrak{S}_2(\mathcal{H})} &\leq \frac{\lambda_M + \lambda_{Mh}}{2}\|\gamma^0 - \gamma_h\|^2_{\mathfrak{S}_2(\mathcal{H})} \\
&\quad + \frac{1}{2}\|\mathrm{Res}(\gamma_h)\|^2_{\mathfrak{S}_2(\mathcal{H})} + \frac{1}{2}\|A^{1/2}(\gamma^0 - \gamma_h)\|^2_{\mathfrak{S}_2(\mathcal{H})},
\end{aligned}$$

from which we deduce (4.4).

To show (4.6), we start from (3.12) with $s=0$ which reads

$$\|\mathrm{Res}(\gamma_h)\|^2_{\mathfrak{S}_2(\mathcal{H})} = \sum_{k\geq 1}\sum_{i=m}^{M} \lambda_k \left(1-\frac{\lambda_{ih}}{\lambda_k}\right)^2 \left|\left(\varphi_{ih},\varphi_k^0\right)\right|^2.$$

Remark now that for $k\geq 1$, $k\notin\{m,\ldots,M\}$, $(1-\frac{\lambda_{ih}}{\lambda_k})^2 \geq {c_h}^{-2}$ under Assumption 4.2, similarly as in [9, proof of Lemma 3.1]. Thus, dropping some non-negative terms and introducing the density matrix γ_h, we obtain

$$\begin{aligned}
\|\mathrm{Res}(\gamma_h)\|^2_{\mathfrak{S}_2(\mathcal{H})} &\geq \sum_{\substack{k\geq 1\\ k\notin\{m,\ldots,M\}}} \sum_{i=m}^{M} \lambda_k \left(1-\frac{\lambda_{ih}}{\lambda_k}\right)^2 \left|\left(\varphi_{ih},\varphi_k^0\right)\right|^2 \\
&\geq {c_h}^{-2} \sum_{\substack{k\geq 1\\ k\notin\{m,\ldots,M\}}} \sum_{i=m}^{M} \lambda_k \left|\left(\varphi_{ih},\varphi_k^0\right)\right|^2 \\
&= {c_h}^{-2} \sum_{\substack{k\geq 1\\ k\notin\{m,\ldots,M\}}} \lambda_k \left(\varphi_k^0,\gamma_h\varphi_k^0\right).
\end{aligned}$$

Now, introducing the rotated discrete eigenvectors $(\varphi^0_{ih})_{i=1,\ldots,M}$ defined by (3.1) through the expression of the density matrix (3.2), using the orthonormality of the eigenvectors (φ_k^0), and definition (2.3a) of $\|A^{1/2}v\|$, there holds

$$\begin{aligned}
\|\mathrm{Res}(\gamma_h)\|^2_{\mathfrak{S}_2(\mathcal{H})} &\geq {c_h}^{-2} \sum_{\substack{k\geq 1\\ k\notin\{m,\ldots,M\}}} \sum_{i=m}^{M} \lambda_k \left|\left(\varphi^0_{ih}-\varphi^0_i,\varphi_k^0\right)\right|^2 \\
&= {c_h}^{-2}\left(\sum_{k\geq 1}\sum_{i=m}^{M} \lambda_k \left|\left(\varphi^0_{ih}-\varphi^0_i,\varphi_k^0\right)\right|^2 - \sum_{k=m}^{M}\sum_{i=m}^{M} \lambda_k \left|\left(\varphi^0_{ih}-\varphi^0_i,\varphi_k^0\right)\right|^2\right) \\
&\geq {c_h}^{-2}\left(\sum_{i=m}^{M} \|A^{1/2}(\varphi^0_{ih}-\varphi^0_i)\|^2 - \lambda_M \sum_{k=m}^{M}\sum_{i=m}^{M} \left|\left(\varphi^0_k,\varphi^0_{ih}-\varphi^0_i\right)\right|^2\right).
\end{aligned}$$

Since the eigenvectors composing Φ_h^0 are orthonormal and using Assumption 2.6, the $J\times J$ overlap matrix $\mathbf{M}_{\Phi^0,\Phi_h^0}$ with entries $\left(\mathbf{M}_{\Phi^0,\Phi_h^0}\right)_{i,k} = (\varphi^0_{ih},\varphi^0_k)$ is symmetric (see [8, Lemma 4.3]). Hence using once again that the eigenvectors are orthonormal, we obtain that for any $i,k=m,\ldots,M$, $i\neq k$,

$$\left(\varphi^0_k,\varphi^0_{ih}-\varphi^0_i\right) = \left(\varphi^0_k,\varphi^0_{ih}\right) = \left(\varphi^0_i,\varphi^0_{kh}\right) = \frac{1}{2}\left(\left(\varphi^0_{ih},\varphi^0_k\right)+\left(\varphi^0_{kh},\varphi^0_i\right)\right) = \frac{1}{2}\left(\varphi^0_{kh}-\varphi^0_k,\varphi^0_i-\varphi^0_{ih}\right).$$

Since for $i=m,\ldots,M$, $\left(\varphi^0_i,\varphi^0_{ih}-\varphi^0_i\right) = -\frac{1}{2}\|\varphi^0_i-\varphi^0_{ih}\|^2$, we obtain that for any $i,k=m,\ldots,M$,

$$\left(\varphi^0_k,\varphi^0_{ih}-\varphi^0_i\right) = \frac{1}{2}\left(\varphi^0_{kh}-\varphi^0_k,\varphi^0_i-\varphi^0_{ih}\right). \tag{4.7}$$

From (4.7), definition (2.15), and the Cauchy–Schwarz inequality, we deduce

$$\begin{aligned}
\|\mathrm{Res}(\gamma_h)\|^2_{\mathfrak{S}_2(\mathcal{H})} &\geq {c_h}^{-2}\left(\|A^{1/2}(\Phi^0-\Phi^0_h)\|^2 - \frac{\lambda_M}{4}\sum_{k=m}^{M}\sum_{i=m}^{M}\left|\left(\varphi^0_k-\varphi^0_{kh},\varphi^0_i-\varphi^0_{ih}\right)\right|^2\right) \\
&\geq {c_h}^{-2}\left(\|A^{1/2}(\Phi^0-\Phi^0_h)\|^2 - \frac{\lambda_M}{4}\left[\sum_{i=m}^{M}\|\varphi^0_i-\varphi^0_{ih}\|^2\right]^2\right) \\
&= {c_h}^{-2}\left(\|A^{1/2}(\Phi^0-\Phi^0_h)\|^2 - \frac{\lambda_M}{4}\|\Phi^0-\Phi^0_h\|^4\right).
\end{aligned}$$

Finally, using (3.7) and (3.8) finishes the proof of (4.6). $\square$

Theorem 4.6 (Lower bound for the density matrix error)**.** *Let Assumption 2.1 and 2.6 hold, let the density matrices γ^0 and γ_h be respectively defined by (2.7) and (2.20), and let the cluster residual* $\mathrm{Res}(\gamma_h)$ *be defined by (3.11). Set*

$$\bar{c}_h := \max\left\{\left(\frac{\lambda_{Mh}}{\lambda_1} - 1\right)^2, 1\right\}. \tag{4.8}$$

Then, there holds

$$\begin{aligned} &\|\mathrm{Res}(\gamma_h)\|^2_{\mathfrak{S}_2(\mathcal{H})} \\ &\leq \bar{c}_h \|A^{1/2}(\gamma^0 - \gamma_h)\|^2_{\mathfrak{S}_2(\mathcal{H})} + \frac{3(\lambda_M - \lambda_m)^2}{4\lambda_m}\|\gamma^0 - \gamma_h\|^4_{\mathfrak{S}_2(\mathcal{H})} + \frac{3}{\lambda_m}\left(1 + \frac{1}{4}\|\gamma^0 - \gamma_h\|^4_{\mathfrak{S}_2(\mathcal{H})}\right) \times \\ &\left[2\left(1 + \frac{\lambda_M}{4\lambda_m}\|\gamma^0 - \gamma_h\|^2_{\mathfrak{S}_2(\mathcal{H})}\right)^2 \|A^{1/2}(\gamma^0 - \gamma_h)\|^4_{\mathfrak{S}_2(\mathcal{H})} + 2(\lambda_M)^2\|\gamma^0 - \gamma_h\|^4_{\mathfrak{S}_2(\mathcal{H})}\right]. \end{aligned} \tag{4.9}$$

Proof. First, let us define the Lagrange multiplier matrix of the orthonormality constraints for Φ_h^0 defined in (3.1) by

$$\mathbf{\Lambda}^h = (\mathbf{\Lambda}^h_{ij})_{m\leq i,j\leq M} := ((A^{1/2}\varphi^0_{ih}, A^{1/2}\varphi^0_{jh}))_{m\leq i,j\leq M} \in \mathbb{R}^{J\times J}. \tag{4.10}$$

Note that the matrix $\mathbf{\Lambda}^h$ is not diagonal in general. However, the matrix of the Lagrange multipliers of the orthonormality constraints for Φ^0 is diagonal, from (2.4). It is denoted by

$$\mathbf{\Lambda} := (\delta_{ij}\lambda_i)_{m\leq i,j\leq M} \in \mathbb{R}^{J\times J}. \tag{4.11}$$

Using $1 = \gamma^0 + (1-\gamma^0)$ and $(1-\gamma^0)\gamma^0 = 0$, there holds

$$\|\mathrm{Res}(\gamma_h)\|^2_{\mathfrak{S}_2(\mathcal{H})} = \|(1-\gamma^0)\mathrm{Res}(\gamma_h)\|^2_{\mathfrak{S}_2(\mathcal{H})} + \|\gamma^0\mathrm{Res}(\gamma_h)\|^2_{\mathfrak{S}_2(\mathcal{H})}. \tag{4.12}$$

To estimate the first term in (4.12), we note that the development performed in Lemma 3.7 can be done similarly in this case, leading to

$$\|(1-\gamma^0)\mathrm{Res}(\gamma_h)\|^2_{\mathfrak{S}_2(\mathcal{H})} = \sum_{\substack{k\geq 1\\ k\notin\{m,\dots,M\}}} \sum_{i=m}^M \lambda_k\left(1 - \frac{\lambda_{ih}}{\lambda_k}\right)^2 \left|\left(\varphi_{ih}, \varphi^0_k\right)\right|^2.$$

Bounding the eigenvalue term by $\bar{c}_h$, using the self-adjointness of $A^{1/2}$, and employing the expansion (2.3b), the Parseval equality (2.2), the Hilbert–Schmidt norm definition (2.10), and the definitions of the projectors (2.7) and (2.20), we obtain

$$\begin{aligned} \|(1-\gamma^0)\mathrm{Res}(\gamma_h)\|^2_{\mathfrak{S}_2(\mathcal{H})} &\leq \max_{\substack{i\in\{m,\dots,M\}\\ k\notin\{m,\dots,M\}}}\left(1 - \frac{\lambda_{ih}}{\lambda_k}\right)^2 \sum_{\substack{k\geq 1\\ k\notin\{m,\dots,M\}}} \sum_{i=m}^M \lambda_k\left|\left(\varphi_{ih}, \varphi^0_k\right)\right|^2 \\ &\leq \bar{c}_h \sum_{\substack{k\geq 1\\ k\notin\{m,\dots,M\}}} \sum_{i=m}^M \left|\left(\varphi_{ih}, A^{1/2}\varphi^0_k\right)\right|^2 \\ &= \bar{c}_h\|(1-\gamma^0)A^{1/2}\gamma_h\|^2_{\mathfrak{S}_2(\mathcal{H})} \\ &\leq \bar{c}_h\|A^{1/2}(\gamma^0 - \gamma_h)\|^2_{\mathfrak{S}_2(\mathcal{H})}, \end{aligned}$$

where the last estimate follows by a Pythagorean equality as (4.12) and the fact that $(1-\gamma^0)A^{1/2}\gamma^0 = 0$.

To deal with the second term in (4.12), first note that from the definition of the residual (3.11)

$$\|\gamma^0\mathrm{Res}(\gamma_h)\|^2_{\mathfrak{S}_2(\mathcal{H})} = \left\|\gamma^0\left(A^{1/2}\gamma_h - A^{-1/2}\left(A^{1/2}\gamma_h\right)^\dagger A^{1/2}\gamma_h\right)\right\|^2_{\mathfrak{S}_2(\mathcal{H})}.$$

Expanding γ_h on the rotated eigenvector basis $(\varphi^0_{mh}, \dots, \varphi^0_{Mh})$ defined in (3.1) and using that γ^0 is self-adjoint leads to, as in Lemma 3.7,

$$\begin{aligned}\|\gamma^0 \mathrm{Res}(\gamma_h)\|^2_{\mathfrak{S}_2(\mathcal{H})} &= \sum_{i=m}^{M} \Big[\big(A^{1/2}\varphi^0_{ih}, \gamma^0 A^{1/2}\varphi^0_{ih}\big) - 2\big(A^{1/2}\varphi^0_{ih}, \gamma^0 A^{-1/2}\big(A^{1/2}\gamma_h\big)^\dagger A^{1/2}\varphi^0_{ih}\big) \\ &\quad + \big(\big(A^{1/2}\gamma_h\big)^\dagger A^{1/2}\varphi^0_{ih}, A^{-1/2}\gamma^0 A^{-1/2}\big(A^{1/2}\gamma_h\big)^\dagger A^{1/2}\varphi^0_{ih}\big)\Big] \\ &=: \sum_{i=m}^{M} [T_{1i} + T_{2i} + T_{3i}].\end{aligned}$$

First, expanding γ^0 and using the self-adjointness of $A^{1/2}$ leads to

$$T_{1i} = \sum_{k=m}^{M} \left|\big(\varphi^0_k, A^{1/2}\varphi^0_{ih}\big)\right|^2 = \sum_{k=m}^{M} \lambda_k \left|(\varphi^0_k, \varphi^0_{ih})\right|^2.$$

Second, we expand $A^{-1/2}$ on the eigenvectors φ^0_k following (2.3b) and we use the self-adjointness of $A^{1/2}$ and the definition of the Lagrange multipliers (4.10) to obtain

$$\begin{aligned}T_{2i} &= -2 \sum_{k=m}^{M} \frac{1}{\sqrt{\lambda_k}} \big(A^{1/2}\varphi^0_{ih}, \varphi^0_k\big)\big(\varphi^0_k, \big(A^{1/2}\gamma_h\big)^\dagger A^{1/2}\varphi^0_{ih}\big) \\ &= -2 \sum_{k=m}^{M} \big(\varphi^0_{ih}, \varphi^0_k\big)\big(\big(A^{1/2}\gamma_h\big)\varphi^0_k, A^{1/2}\varphi^0_{ih}\big) \\ &= -2 \sum_{k=m}^{M}\sum_{j=m}^{M} \big(\varphi^0_{ih}, \varphi^0_k\big)\big(A^{1/2}\varphi^0_{jh}, A^{1/2}\varphi^0_{ih}\big)\big(\varphi^0_{jh}, \varphi^0_k\big) \\ &= -2 \sum_{k=m}^{M}\sum_{j=m}^{M} \mathbf{\Lambda}^h_{ij}\big(\varphi^0_{ih}, \varphi^0_k\big)\big(\varphi^0_{jh}, \varphi^0_k\big).\end{aligned}$$

Third, using the definition of γ^0 and expanding $A^{-1/2}$ two-times on the eigenvectors φ^0_k as well as γ_h on the rotated eigenvector basis $(\varphi^0_{mh}, \dots, \varphi^0_{Mh})$ leads to

$$\begin{aligned}T_{3i} &= \sum_{k=m}^{M} \frac{1}{\lambda_k}\big(\big(A^{1/2}\gamma_h\big)^\dagger A^{1/2}\varphi^0_{ih}, \varphi^0_k\big)\big(\varphi^0_k, \big(A^{1/2}\gamma_h\big)^\dagger A^{1/2}\varphi^0_{ih}\big) \\ &= \sum_{k=m}^{M} \frac{1}{\lambda_k}\big(A^{1/2}\varphi^0_{ih}, A^{1/2}\gamma_h\varphi^0_k\big)\big(A^{1/2}\gamma_h\varphi^0_k, A^{1/2}\varphi^0_{ih}\big) \\ &= \sum_{k=m}^{M} \frac{1}{\lambda_k}\sum_{j=m}^{M}\sum_{p=m}^{M} \big(A^{1/2}\varphi^0_{ih}, A^{1/2}\varphi^0_{jh}\big)\big(\varphi^0_{jh}, \varphi^0_k\big)\big(A^{1/2}\varphi^0_{ph}, A^{1/2}\varphi^0_{ih}\big)\big(\varphi^0_{ph}, \varphi^0_k\big) \\ &= \sum_{k=m}^{M} \frac{1}{\lambda_k}\sum_{j=m}^{M}\sum_{p=m}^{M} \mathbf{\Lambda}^h_{ij}\mathbf{\Lambda}^h_{ip}\big(\varphi^0_{jh}, \varphi^0_k\big)\big(\varphi^0_{ph}, \varphi^0_k\big).\end{aligned}$$

Putting T_{1i}, T_{2i}, T_{3i} together, we can write

$$\begin{aligned}\|\gamma^0 \mathrm{Res}(\gamma_h)\|^2_{\mathfrak{S}_2(\mathcal{H})} &= \sum_{k=m}^{M}\sum_{i=m}^{M} \frac{1}{\lambda_k}\left(\lambda_k(\varphi^0_{ih}, \varphi^0_k) - \sum_{j=m}^{M} \mathbf{\Lambda}^h_{ij}(\varphi^0_{jh}, \varphi^0_k)\right)^2 \\ &= \sum_{k=m}^{M}\sum_{i=m}^{M} \frac{1}{\lambda_k}\Bigg((\delta_{ik}\lambda_k - \mathbf{\Lambda}^h_{ik})(\varphi^0_{kh}, \varphi^0_k) + (1-\delta_{ik})(\lambda_k - \mathbf{\Lambda}_{ii})(\varphi^0_{ih}, \varphi^0_k)\end{aligned}$$

$$
- \sum_{\substack{j=m\\ j\neq k}}^{M} (\mathbf{\Lambda}^h_{ij} - \delta_{ij}\mathbf{\Lambda}_{ij})(\varphi^0_{jh}, \varphi^0_k)\Bigg)^2.
$$

Using definition (4.11), the inequality $(a+b+c)^2 \le 3(a^2+b^2+c^2)$, and using that the eigenvectors are orthonormal, we obtain

$$
\begin{aligned}
\|\gamma^0 \mathrm{Res}(\gamma_h)\|^2_{\mathfrak{S}_2(\mathcal{H})} &= \sum_{k=m}^{M}\sum_{i=m}^{M} \frac{1}{\lambda_k}\Bigg((\mathbf{\Lambda}_{ik} - \mathbf{\Lambda}^h_{ik})(\varphi^0_{kh}, \varphi^0_k) + (1-\delta_{ik})(\mathbf{\Lambda}_{kk} - \mathbf{\Lambda}_{ii})(\varphi^0_{ih}, \varphi^0_k) \\
&\quad - \sum_{\substack{j=m\\ j\neq k}}^{M} (\mathbf{\Lambda}^h_{ij} - \mathbf{\Lambda}_{ij})(\varphi^0_{jh}, \varphi^0_k)\Bigg)^2 \\
&\le 3\sum_{k=m}^{M}\sum_{i=m}^{M} \frac{1}{\lambda_k}(\mathbf{\Lambda}_{ik} - \mathbf{\Lambda}^h_{ik})^2(\varphi^0_{kh}, \varphi^0_k)^2 \\
&\quad + 3\sum_{k=m}^{M}\sum_{i=m}^{M} \frac{1}{\lambda_k}(1-\delta_{ik})((\mathbf{\Lambda}_{kk} - \mathbf{\Lambda}_{ii}))^2(\varphi^0_{ih} - \varphi^0_i, \varphi^0_k)^2 \\
&\quad + 3\sum_{k=m}^{M}\sum_{i=m}^{M} \frac{1}{\lambda_k}\left(\sum_{\substack{j=m\\ j\neq k}}^{M} (\mathbf{\Lambda}^h_{ij} - \mathbf{\Lambda}_{ij})(\varphi^0_{jh} - \varphi^0_j, \varphi^0_k)\right)^2.
\end{aligned}
$$

Noting that $|(\varphi^0_{kh}, \varphi^0_k)| \le 1$ for $k = m, \dots, M$, using (4.7), the Cauchy–Schwarz inequality, and (3.7), we get

$$
\begin{aligned}
\|\gamma^0 \mathrm{Res}(\gamma_h)\|^2_{\mathfrak{S}_2(\mathcal{H})} &\le \frac{3}{\lambda_m}\|\mathbf{\Lambda} - \mathbf{\Lambda}^h\|^2_{\mathrm{F}} + \frac{3(\lambda_M - \lambda_m)^2}{4\lambda_m}\sum_{k=m}^{M}\sum_{i=m}^{M}(\varphi^0_{ih} - \varphi^0_i, \varphi^0_k - \varphi^0_{kh})^2 \\
&\quad + \frac{3}{4\lambda_m}\sum_{k=m}^{M}\sum_{i=m}^{M}\left(\sum_{\substack{j=m\\ j\neq k}}^{M} (\mathbf{\Lambda}^h_{ij} - \mathbf{\Lambda}_{ij})(\varphi^0_{jh} - \varphi^0_j, \varphi^0_k - \varphi^0_{kh})\right)^2 \\
&\le \frac{3}{\lambda_m}\|\mathbf{\Lambda} - \mathbf{\Lambda}^h\|^2_{\mathrm{F}} + \frac{3(\lambda_M - \lambda_m)^2}{4\lambda_m}\|\Phi^0_h - \Phi^0\|^4 + \frac{3}{4\lambda_m}\|\mathbf{\Lambda} - \mathbf{\Lambda}^h\|^2_{\mathrm{F}}\|\Phi^0_h - \Phi^0\|^4 \\
&\le \frac{3}{\lambda_m}\|\mathbf{\Lambda} - \mathbf{\Lambda}^h\|^2_{\mathrm{F}}\left(1 + \frac{1}{4}\|\gamma^0 - \gamma_h\|^4_{\mathfrak{S}_2(\mathcal{H})}\right) + \frac{3(\lambda_M - \lambda_m)^2}{4\lambda_m}\|\gamma^0 - \gamma_h\|^4_{\mathfrak{S}_2(\mathcal{H})},
\end{aligned}
$$

where $\|\cdot\|_{\mathrm{F}}$ is the matrix Frobenius (or Hilbert–Schmidt) norm. Combining the estimates for the two summands in (4.12), the dual norm of the residual can be bounded by

$$
\begin{aligned}
\|\mathrm{Res}(\gamma_h)\|^2_{\mathfrak{S}_2(\mathcal{H})} &\le \bar{c}_h\|A^{1/2}(\gamma^0 - \gamma_h)\|^2_{\mathfrak{S}_2(\mathcal{H})} + \frac{3(\lambda_M - \lambda_m)^2}{4\lambda_m}\|\gamma^0 - \gamma_h\|^4_{\mathfrak{S}_2(\mathcal{H})} \\
&\quad + \frac{3}{\lambda_m}\|\mathbf{\Lambda} - \mathbf{\Lambda}^h\|^2_{\mathrm{F}}\left(1 + \frac{1}{4}\|\gamma^0 - \gamma_h\|^4_{\mathfrak{S}_2(\mathcal{H})}\right).
\end{aligned} \tag{4.13}
$$

We are left with estimating the Lagrange multipliers error in the Frobenius norm. For $m \le i, j \le M$, and using (4.7), there holds

$$
\begin{aligned}
\mathbf{\Lambda}^h_{ij} - \mathbf{\Lambda}_{ij} &= (A^{1/2}\varphi^0_{ih}, A^{1/2}\varphi^0_{jh}) - (A^{1/2}\varphi^0_i, A^{1/2}\varphi^0_j) \\
&= (A^{1/2}(\varphi^0_{ih} - \varphi^0_i), A^{1/2}(\varphi^0_{jh} - \varphi^0_j)) \\
&\quad + (A^{1/2}\varphi^0_i, A^{1/2}(\varphi^0_{jh} - \varphi^0_j)) + (A^{1/2}(\varphi^0_{ih} - \varphi^0_i), A^{1/2}\varphi^0_j) \\
&= (A^{1/2}(\varphi^0_{ih} - \varphi^0_i), A^{1/2}(\varphi^0_{jh} - \varphi^0_j))
\end{aligned}
$$

$$
\begin{aligned}
&+ \lambda_i(\varphi_i^0, \varphi_{jh}^0 - \varphi_j^0) + \lambda_j(\varphi_j^0, \varphi_{ih}^0 - \varphi_i^0) \\
&= (A^{1/2}(\varphi_{ih}^0 - \varphi_i^0), A^{1/2}(\varphi_{jh}^0 - \varphi_j^0)) \\
&\quad + \frac{\lambda_i}{2}(\varphi_i^0 - \varphi_{ih}^0, \varphi_{jh}^0 - \varphi_j^0) + \frac{\lambda_j}{2}(\varphi_j^0 - \varphi_{jh}^0, \varphi_{ih}^0 - \varphi_i^0).
\end{aligned}
$$

Using the Cauchy–Schwarz inequality,

$$
\begin{aligned}
|\mathbf{\Lambda}_{ij}^h - \mathbf{\Lambda}_{ij}| \leq &\|A^{1/2}(\varphi_{jh}^0 - \varphi_j^0)\| \, \|A^{1/2}(\varphi_{ih}^0 - \varphi_i^0)\| \\
&+ \frac{\lambda_i + \lambda_j}{2} \|\varphi_{jh}^0 - \varphi_j^0\| \|\varphi_{ih}^0 - \varphi_i^0\|,
\end{aligned}
$$

from which we deduce that

$$
\begin{aligned}
|\mathbf{\Lambda}_{ij}^h - \mathbf{\Lambda}_{ij}|^2 \leq 2&\|A^{1/2}(\varphi_{jh}^0 - \varphi_j^0)\|^2 \|A^{1/2}(\varphi_{ih}^0 - \varphi_i^0)\|^2 \\
&+ 2\left(\frac{\lambda_i + \lambda_j}{2}\right)^2 \|\varphi_{jh}^0 - \varphi_j^0\|^2 \|\varphi_{ih}^0 - \varphi_i^0\|^2.
\end{aligned}
$$

Finally, using (3.8) and (3.7), the estimate for the Frobenius norm goes as

$$
\begin{aligned}
\|\mathbf{\Lambda}^h - \mathbf{\Lambda}\|_{\mathrm{F}}^2 &= \sum_{i,j=m}^{M} |\mathbf{\Lambda}_{ij}^h - \mathbf{\Lambda}_{ij}|^2 \\
&\leq 2\left(\sum_{i=m}^{M} \|A^{1/2}(\varphi_{ih}^0 - \varphi_i^0)\|^2\right)^2 + 2(\lambda_M)^2 \left(\sum_{i=m}^{M} \|\varphi_{ih}^0 - \varphi_i^0\|^2\right)^2 \\
&= 2\|A^{1/2}(\Phi_h^0 - \Phi^0)\|^4 + 2(\lambda_M)^2 \|\Phi_h^0 - \Phi^0\|^4 \\
&\leq 2\left(1 + \frac{\lambda_M}{4\lambda_m}\|\gamma^0 - \gamma_h\|_{\mathfrak{S}_2(\mathcal{H})}^2\right)^2 \|A^{1/2}(\gamma^0 - \gamma_h)\|_{\mathfrak{S}_2(\mathcal{H})}^4 + 2(\lambda_M)^2 \|\gamma^0 - \gamma_h\|_{\mathfrak{S}_2(\mathcal{H})}^4. \qquad (4.14)
\end{aligned}
$$

The result (4.9) follows from inserting (4.14) into (4.13). ☐

4.3 Bound on the $\mathcal{H}$-norm of the density matrix error

Finally, we provide two estimates for the Hilbert–Schmidt norm of the density matrix error. The second bound makes appear the Hilbert–Schmidt norm of the cluster residual $\mathrm{Res}(\gamma_h)$, already present in the bounds above. The first bound measures the residual further scaled by $A^{-1/2}$; it is typically sharper but can be less straightforward to estimate further.

Lemma 4.7 (Bounds on the density matrix error)**.** *Let Assumptions 2.1 and 4.2 hold, let the density matrices γ^0 and γ_h be respectively defined by (2.7) and (2.20), and let the cluster residual $\mathrm{Res}(\gamma_h)$ be defined by (3.11). Set*

$$
\tilde{c}_h := \max\left[(\overline{\lambda}_{m-1})^{-1/2}\left(\frac{\lambda_{mh}}{\overline{\lambda}_{m-1}} - 1\right)^{-1}, (\underline{\lambda}_{M+1})^{-1/2}\left(1 - \frac{\lambda_{Mh}}{\underline{\lambda}_{M+1}}\right)^{-1}\right], \qquad (4.15)
$$

the first term in the max *being discarded for $m = 1$, and recall c_h is defined in (4.5). Then there holds*

$$
\|\gamma^0 - \gamma_h\|_{\mathfrak{S}_2(\mathcal{H})} \leq \sqrt{2} c_h \|A^{-1/2}\mathrm{Res}(\gamma_h)\|_{\mathfrak{S}_2(\mathcal{H})} \qquad (4.16)
$$

and

$$
\|\gamma^0 - \gamma_h\|_{\mathfrak{S}_2(\mathcal{H})} \leq \sqrt{2} \tilde{c}_h \|\mathrm{Res}(\gamma_h)\|_{\mathfrak{S}_2(\mathcal{H})}. \qquad (4.17)
$$

Proof. First, starting from (3.12) with $s = 0$, neglecting again some positive terms in the sum, and bounding below the eigenvalue part with the help of $\tilde{c}_h$, we obtain

$$
\begin{aligned}
\|\mathrm{Res}(\gamma_h)\|^2_{\mathfrak{S}_2(\mathcal{H})} &= \sum_{k\geq 1}\sum_{i=m}^{M} \frac{(\lambda_k-\lambda_{ih})^2}{\lambda_k} |\left(\varphi_{ih},\varphi_k^0\right)|^2 \\
&\geq \sum_{\substack{k\geq 1\\ k\notin\{m,\dots,M\}}} \sum_{i=m}^{M} \frac{(\lambda_k-\lambda_{ih})^2}{\lambda_k} |\left(\varphi_{ih},\varphi_k^0\right)|^2 \\
&\geq \tilde{c}_h{}^{-2} \sum_{\substack{k\geq 1\\ k\notin\{m,\dots,M\}}} \sum_{i=m}^{M} |\left(\varphi_{ih},\varphi_k^0\right)|^2.
\end{aligned}
$$

Similarly, for the $A^{-1/2}$-scaled residual, there holds

$$
\begin{aligned}
\|A^{-1/2}\mathrm{Res}(\gamma_h)\|^2_{\mathfrak{S}_2(\mathcal{H})} &= \sum_{k\geq 1}\sum_{i=m}^{M} \frac{(\lambda_k-\lambda_{ih})^2}{(\lambda_k)^2} |\left(\varphi_{ih},\varphi_k^0\right)|^2 \\
&\geq \sum_{\substack{k\geq 1\\ k\notin\{m,\dots,M\}}} \sum_{i=m}^{M} \frac{(\lambda_k-\lambda_{ih})^2}{(\lambda_k)^2} |\left(\varphi_{ih},\varphi_k^0\right)|^2 \\
&\geq c_h{}^{-2} \sum_{\substack{k\geq 1\\ k\notin\{m,\dots,M\}}} \sum_{i=m}^{M} |\left(\varphi_{ih},\varphi_k^0\right)|^2.
\end{aligned}
$$

Moreover, from Lemma 2.2, expanding the expression in terms of the eigenvectors,

$$
\begin{aligned}
\|\gamma^0-\gamma_h\|^2_{\mathfrak{S}_2(\mathcal{H})} = 2\mathrm{Tr}(\gamma_h(1-\gamma^0)) &= 2\sum_{i=m}^{M}\left(\varphi_{ih},(1-\gamma^0)\varphi_{ih}\right) \\
&= 2\sum_{\substack{k\geq 1\\ k\notin\{m,\dots,M\}}} \sum_{i=m}^{M} |\left(\varphi_{ih},\varphi_k^0\right)|^2,
\end{aligned}
$$

from which we easily deduce (4.16) and (4.17). ☐

By combining equations (4.16) or (4.17) with (4.1) and (4.4) or (4.6), it is possible to obtain estimates for the errors on the density matrix as well as on the sum of the eigenvalues which only depend on the dual norm of the cluster residual together with the exact eigenvalues λ_{m-1}, λ_M, and λ_{M+1} (or their corresponding lower and upper bound according to Assumption 4.2); the converse estimate (4.9), not necessary in practice to guarantee upper bounds of the error and only used to theoretically assess the efficiency of the estimates, also employs λ_1 and λ_m. Note that practically computable bounds on these eigenvalues are obtained following Remark 4.3. Using such bounds, computable estimates are obtained provided the dual norm of the residual $\|\mathrm{Res}(\gamma_h)\|_{\mathfrak{S}_2(\mathcal{H})}$ can be evaluated or estimated. This is possible for specific operators and numerical methods, as illustrated in the next section.

5 Guaranteed and computable *a posteriori* error estimates

In this section, we transform the estimates presented in Section 4 into fully guaranteed and computable estimators in two particular cases. First, we focus on the Laplace operator $-\Delta$ with homogeneous Dirichlet conditions discretized with conforming finite elements, for which the dual norm of the residual was estimated in [9], based on [39, 18, 6, 21]. We then present estimates for a Schrödinger operator of the form $-\Delta+V$ on a cubic box with periodic boundary conditions, where V is a bounded-below periodic multiplicative potential, discretized with planewaves, in which case the dual norm of the residual can be easily computed.

5.1 Finite element discretization of the Laplace operator

In this section, we consider the Laplace eigenvalue problem with Dirichlet boundary conditions. Let $\Omega \subset \mathbb{R}^d$, $d = 2, 3$, be a polygonal/polyhedral domain with a Lipschitz boundary. In this setting, $A = -\Delta$, $\mathcal{H} = L^2(\Omega)$, and $D(A^{1/2}) = \mathcal{V} := H_0^1(\Omega)$. Let $\mathbf{H}(\mathrm{div}, \Omega)$ stand for the space of $[L^2(\Omega)]^d$ functions with weak divergences in $L^2(\Omega)$, and let $H^{-1}(\Omega)$ be the dual of $H_0^1(\Omega)$. The problem reads: find eigenvector and eigenvalue pairs (φ_k^0, λ_k) such that $-\Delta\varphi_k^0 = \lambda_k \varphi_k^0$ in Ω, subject to the orthonormality constraints $(\varphi_k^0, \varphi_j^0) = \delta_{kj}$, $k, j \geq 1$. In weak form, this reads: find $(\varphi_k^0, \lambda_k) \in \mathcal{V} \times \mathbb{R}_+$ with $(\varphi_k^0, \varphi_j^0) = \delta_{kj}$ such that

$$(\nabla\varphi_k^0, \nabla v) = \lambda_k(\varphi_k^0, v) \qquad \forall v \in \mathcal{V}. \tag{5.1}$$

Here, for $\omega \subset \Omega$, $(\nabla u, \nabla v)_\omega$ stands for $\int_\omega \nabla u{\cdot}\nabla v$ and $(u, v)_\omega$ for $\int_\omega uv$; we also denote $\|\nabla v\|_\omega^2 := \int_\omega |\nabla v|^2$ and $\|v\|_\omega^2 := \int_\omega v^2$ and drop the index whenever $\omega = \Omega$.

We consider a conforming finite element discretization of this problem. Let $\{\mathcal{T}_h\}_h$ be a family of meshes, matching simplicial partitions of the domain Ω. We suppose that it is shape regular in the sense that there exists a constant $\kappa_\mathcal{T} > 0$ such that the ratio of the element diameter and of the diameter of its largest inscribed ball is uniformly bounded by $\kappa_\mathcal{T}$, cf. Ciarlet [16]. A generic element of $\mathcal{T}_h$ is denoted by K. The set of vertices of $\mathcal{T}_h$ is denoted by $\mathcal{V}_h$, the set of interior vertices by $\mathcal{V}_h^{\mathrm{int}}$, the set of vertices located on the boundary by $\mathcal{V}_h^{\mathrm{ext}}$, and a generic vertex by $\mathbf{a}$. We denote by $\mathcal{T}_\mathbf{a}$ the patch of elements of $\mathcal{T}_h$ which share the vertex $\mathbf{a} \in \mathcal{V}_h$, by $\omega_\mathbf{a}$ the corresponding open subdomain, and by $\mathbf{n}_{\omega_\mathbf{a}}$ its outward unit normal. We will often tacitly extend functions defined on $\omega_\mathbf{a}$ by zero outside of $\omega_\mathbf{a}$, whereas $V_h(\omega_\mathbf{a})$ stands for the restriction of the space V_h to $\omega_\mathbf{a}$. Next, $\psi_\mathbf{a}$ for $\mathbf{a} \in \mathcal{V}_h$ stands for the piecewise affine "hat" function taking value 1 at the vertex $\mathbf{a}$ and zero at the other vertices. Note that $(\psi_\mathbf{a})_{\mathbf{a}\in\mathcal{V}_h}$ form a partition of unity since $\sum_{\mathbf{a}\in\mathcal{V}_h} \psi_\mathbf{a} = 1|_\Omega$.

Let $\mathbb{P}_s(K)$, $s \geq 0$, stand for the space polynomials on K of total degree at most s, and $\mathbb{P}_s(\mathcal{T}_h)$ for the space of piecewise polynomials on $\mathcal{T}_h$, without any continuity requirement at the element interfaces. The approximation space is $V_h := \mathbb{P}_p(\mathcal{T}_h) \cap \mathcal{V}$ for a given polynomial degree $p \geq 1$. Let also $\mathbf{V}_h \times Q_h \subset \mathbf{H}(\mathrm{div}, \Omega) \times L^2(\Omega)$ stand for the Raviart–Thomas–Nédélec (RTN) mixed finite element spaces of order $p+1$, i.e., $\mathbf{V}_h := \{\mathbf{v}_h \in \mathbf{H}(\mathrm{div}, \Omega); \mathbf{v}_h|_K \in [\mathbb{P}_{p+1}(K)]^d + \mathbb{P}_{p+1}(K)\mathbf{x}\}$ and $Q_h := \mathbb{P}_{p+1}(\mathcal{T}_h)$, see Brezzi and Fortin [7] or Roberts and Thomas [41]. We also denote by Π_{Q_h} the $L^2(\Omega)$-orthogonal projection onto Q_h.

The discretized eigenvalue problem then reads in this case: find $(\varphi_{kh}, \lambda_{kh}) \in V_h \times \mathbb{R}_+$ with $(\varphi_{kh}, \varphi_{jh}) = \delta_{kj}$, $1 \leq k, j \leq \dim V_h$, such that

$$(\nabla\varphi_{kh}, \nabla v_h) = \lambda_{kh}(\varphi_{kh}, v_h) \qquad \forall v_h \in V_h. \tag{5.2}$$

5.1.1 Residual norm estimate

In order to turn the error estimates obtained in Section 4 into practical ones, we need to estimate the dual norm of the residual $\|\mathrm{Res}(\gamma_h)\|_{\mathfrak{S}_2(\mathcal{H})}$, which in turn requires an estimate of $\|\mathrm{Res}(\varphi_{ih}, \lambda_{ih})\|_{H^{-1}(\Omega)}$ for $i = m, \ldots, M$. The latter estimate relies on previous works [39, 18, 6, 21] and has been presented for the Laplace eigenvalue problem in [9]. We recall the key points here for the sake of completeness.

From (5.1), it is easy to see that for all $i \geq 1$, there holds $-\nabla\varphi_i^0 \in \mathbf{H}(\mathrm{div}, \Omega)$, with the weak divergence equal to $\lambda_i\varphi_i^0$. However, this does not hold at the discrete level, i.e., in general, $-\nabla\varphi_{ih} \notin \mathbf{H}(\mathrm{div}, \Omega)$, and a fortiori $\nabla{\cdot}(-\nabla\varphi_{ih}) \neq \lambda_{ih}\varphi_{ih}$. We therefore introduce an *equilibrated flux reconstruction*, a vector field $\boldsymbol{\sigma}_{ih}$ constructed from $(\varphi_{ih}, \lambda_{ih})$, satisfying

$$\boldsymbol{\sigma}_{ih} \in \mathbf{H}(\mathrm{div}, \Omega), \tag{5.3a}$$

$$\nabla{\cdot}\boldsymbol{\sigma}_{ih} = \lambda_{ih}\varphi_{ih}. \tag{5.3b}$$

In the context of conforming finite elements, the flux reconstruction $\boldsymbol{\sigma}_{ih}$ for $i = m, \ldots, M$ can be constructed from the following *local constrained minimizations*:

Definition 5.1 (Equilibrated flux reconstruction)**.** *For a mesh vertex $\mathbf{a} \in \mathcal{V}_h$, set*

$$\begin{aligned}
&\left.\begin{aligned}\mathbf{V}_h^\mathbf{a} &:= \{\mathbf{v}_h \in \mathbf{V}_h(\omega_\mathbf{a}); \mathbf{v}_h{\cdot}\mathbf{n}_{\omega_\mathbf{a}} = 0 \textit{ on } \partial\omega_\mathbf{a}\},\\ Q_h^\mathbf{a} &:= \{q_h \in Q_h(\omega_\mathbf{a}); (q_h, 1)_{\omega_\mathbf{a}} = 0\},\end{aligned}\right\} \quad \mathbf{a} \in \mathcal{V}_h^{\mathrm{int}},\\
&\left.\begin{aligned}\mathbf{V}_h^\mathbf{a} &:= \{\mathbf{v}_h \in \mathbf{V}_h(\omega_\mathbf{a}); \mathbf{v}_h{\cdot}\mathbf{n}_{\omega_\mathbf{a}} = 0 \textit{ on } \partial\omega_\mathbf{a} \setminus \partial\Omega\},\\ Q_h^\mathbf{a} &:= Q_h(\omega_\mathbf{a}),\end{aligned}\right\} \quad \mathbf{a} \in \mathcal{V}_h^{\mathrm{ext}}.
\end{aligned}$$

Then define $\boldsymbol{\sigma}_{ih} := \sum_{\mathbf{a}\in\mathcal{V}_h} \boldsymbol{\sigma}_{ih}^{\mathbf{a}} \in \mathbf{V}_h$*, where* $\boldsymbol{\sigma}_{ih}^{\mathbf{a}} \in \mathbf{V}_h^{\mathbf{a}}$ *solve*

$$\boldsymbol{\sigma}_{ih}^{\mathbf{a}} := \arg \min_{\substack{\mathbf{v}_h\in\mathbf{V}_h^{\mathbf{a}},\\ \nabla\cdot\mathbf{v}_h=\Pi_{Q_h}(\lambda_{ih}\varphi_{ih}\psi_{\mathbf{a}}-\nabla\varphi_{ih}\cdot\nabla\psi_{\mathbf{a}})}} \|\psi_{\mathbf{a}}\nabla\varphi_{ih} + \mathbf{v}_h\|_{\omega_{\mathbf{a}}} \qquad \forall \mathbf{a} \in \mathcal{V}_h. \tag{5.4}$$

The Euler–Lagrange equations for (5.4) give the standard *mixed finite element formulation*, cf. [21, Remark 3.7]: find $\boldsymbol{\sigma}_{ih}^{\mathbf{a}} \in \mathbf{V}_h^{\mathbf{a}}$ and $p_h^{\mathbf{a}} \in Q_h^{\mathbf{a}}$ such that

$$(\boldsymbol{\sigma}_{ih}^{\mathbf{a}}, \mathbf{v}_h)_{\omega_{\mathbf{a}}} - (p_h^{\mathbf{a}}, \nabla\cdot\mathbf{v}_h)_{\omega_{\mathbf{a}}} = -(\psi_{\mathbf{a}}\nabla\varphi_{ih}, \mathbf{v}_h)_{\omega_{\mathbf{a}}} \qquad \forall \mathbf{v}_h \in \mathbf{V}_h^{\mathbf{a}}, \tag{5.5a}$$

$$(\nabla\cdot\boldsymbol{\sigma}_{ih}^{\mathbf{a}}, q_h)_{\omega_{\mathbf{a}}} = (\lambda_{ih}\varphi_{ih}\psi_{\mathbf{a}} - \nabla\varphi_{ih}\cdot\nabla\psi_{\mathbf{a}}, q_h)_{\omega_{\mathbf{a}}} \qquad \forall q_h \in Q_h^{\mathbf{a}}. \tag{5.5b}$$

Consequently, $\nabla\cdot\boldsymbol{\sigma}_{ih} = \lambda_{ih}\varphi_{ih}$, cf., e.g., [21, Lemma 3.5].

On each patch $\omega_{\mathbf{a}}$ around the vertex $\mathbf{a} \in \mathcal{V}_h$, define

$$H^1_*(\omega_{\mathbf{a}}) := \{v \in H^1(\omega_{\mathbf{a}});\ (v,1)_{\omega_{\mathbf{a}}} = 0\}, \qquad \mathbf{a} \in \mathcal{V}_h^{\mathrm{int}}, \tag{5.6a}$$

$$H^1_*(\omega_{\mathbf{a}}) := \{v \in H^1(\omega_{\mathbf{a}});\ v = 0 \text{ on } \partial\omega_{\mathbf{a}} \cap \partial\Omega\}, \quad \mathbf{a} \in \mathcal{V}_h^{\mathrm{ext}}. \tag{5.6b}$$

Following Carstensen and Funken [12, Theorem 3.1], Braess *et al.* [6, Section 3], or [21, Lemma 3.12], there exists a constant $C_{\mathrm{cont,PF}}$ only depending on the mesh regularity parameter $\kappa_{\mathcal{T}}$ such that

$$\|\nabla(\psi_{\mathbf{a}}v)\|_{\omega_{\mathbf{a}}} \le C_{\mathrm{cont,PF}}\|\nabla v\|_{\omega_{\mathbf{a}}} \qquad \forall v \in H^1_*(\omega_{\mathbf{a}}),\ \forall \mathbf{a} \in \mathcal{V}_h. \tag{5.7}$$

Moreover, the key result of Braess *et al.* [6, Theorem 7], see [22, Corollaries 3.3 and 3.6] for $d = 3$, states that the reconstructions of Definition 5.1 satisfy the following *stability* property,

$$\|\psi_{\mathbf{a}}\nabla\varphi_{ih} + \boldsymbol{\sigma}_{ih}^{\mathbf{a}}\|_{\omega_{\mathbf{a}}} \le C_{\mathrm{st}} \sup_{\substack{v\in H^1_*(\omega_{\mathbf{a}})\\ \|\nabla v\|_{\omega_{\mathbf{a}}}=1}} \{\langle \mathrm{Res}(\varphi_{ih},\lambda_{ih}), \psi_{\mathbf{a}}v\rangle_{H^{-1}(\Omega),H^1_0(\Omega)}\}. \tag{5.8}$$

The constant $C_{\mathrm{st}} > 0$ again only depends on $\kappa_{\mathcal{T}}$, and a computable upper bound on C_{st} is given in [21, Lemma 3.23].

In this setting, the dual norm of the residual can be bounded as follows.

Theorem 5.2 (Residual equivalences)**.** *For* $i = m, \ldots, M$*, let* $(\varphi_{ih}, \lambda_{ih}) \in V_h \times \mathbb{R}$ *be defined in* (5.2)*. Then, for the reconstruction* $\boldsymbol{\sigma}_{ih}$ *from Definition 5.1,*

$$\|\mathrm{Res}(\varphi_{ih},\lambda_{ih})\|_{H^{-1}(\Omega)} \le \|\nabla\varphi_{ih} + \boldsymbol{\sigma}_{ih}\|, \tag{5.9a}$$

$$\|\nabla\varphi_{ih} + \boldsymbol{\sigma}_{ih}\| \le (d+1)C_{\mathrm{st}}C_{\mathrm{cont,PF}}\|\mathrm{Res}(\varphi_{ih},\lambda_{ih})\|_{H^{-1}(\Omega)}. \tag{5.9b}$$

Therefore, there holds

$$\|\mathrm{Res}(\gamma_h)\|^2_{\mathfrak{S}_2(\mathcal{H})} \le \sum_{i=m}^{M}\|\nabla\varphi_{ih} + \boldsymbol{\sigma}_{ih}\|^2, \tag{5.10a}$$

$$\sum_{i=m}^{M}\|\nabla\varphi_{ih} + \boldsymbol{\sigma}_{ih}\|^2 \le (d+1)^2C_{\mathrm{st}}^2C_{\mathrm{cont,PF}}^2\|\mathrm{Res}(\gamma_h)\|^2_{\mathfrak{S}_2(\mathcal{H})}. \tag{5.10b}$$

Proof. Fix $v \in \mathcal{V}$ with $\|\nabla v\| = 1$. Starting from (3.10a), adding and subtracting $(\boldsymbol{\sigma}_{ih}, \nabla v)$, applying Green's theorem and using (5.3b) yields

$$\langle \mathrm{Res}(\varphi_{ih},\lambda_{ih}), v\rangle_{H^{-1}(\Omega),H^1_0(\Omega)} = \lambda_{ih}(\varphi_{ih}, v) - (\nabla\varphi_{ih}, \nabla v) = -(\nabla\varphi_{ih} + \boldsymbol{\sigma}_{ih}, \nabla v).$$

Then, definition (3.10b) of the dual norm of the residual and the Cauchy–Schwarz inequality yield (5.9a). This actually also holds when choosing for $\mathbf{V}_h$ the cheaper RTN space of order p (instead of $p+1$), as (5.3b) still holds for Definition 5.1 with this choice. As in [9], the proof of (5.9b) relies on [21, Lemma 3.22], where the weak norm $\|\mathrm{Res}(\varphi_{ih},\lambda_{ih})\|_{H^{-1}(\Omega)}$ is treated as in [15, Theorems 3.3 and 4.8]. Finally, the bounds (5.10a) and (5.10b) directly follow from (5.9a), and (5.9b) combined with Lemma 3.8. $\square$

5.1.2 Final estimates

We combine here the results of the previous sections to derive the actual guaranteed and fully computable bounds. We will denote by $\zeta_{(ih)}$ the solution of the Laplace *source problem* $-\Delta\zeta_{(ih)} = \mathscr{r}_{(ih)}$ in Ω, $\zeta_{(ih)} = 0$ on $\partial\Omega$, i.e., $\zeta_{(ih)} \in \mathcal{V}$ such that

$$(\nabla\zeta_{(ih)}, \nabla v) = (\mathscr{r}_{(ih)}, v) \qquad \forall v \in \mathcal{V}, \tag{5.11}$$

where $\mathscr{r}_{(ih)} \in \mathcal{V}$ is the *Riesz representation* of the residual defined by

$$(\nabla \mathscr{r}_{(ih)}, \nabla v) = \langle \mathrm{Res}(\varphi_{ih}, \lambda_{ih}), v\rangle_{H^{-1}(\Omega), H_0^1(\Omega)} \qquad \forall v \in \mathcal{V}, \tag{5.12a}$$

$$\|\nabla \mathscr{r}_{(ih)}\| = \|\mathrm{Res}(\varphi_{ih}, \lambda_{ih})\|_{H^{-1}(\Omega)}, \tag{5.12b}$$

cf. (3.13).

Theorem 5.3 (Guaranteed bounds for the sum of eigenvalues)**.** *Let $m, M \in \mathbb{N}\backslash\{0\}, M \geq m$, and let Assumptions 2.1 and 4.2 hold. For $i = m, \dots, M$, let $(\varphi_{ih}, \lambda_{ih}) \in V_h \times \mathbb{R}_+$ be given by (5.2). For $i = m, \dots, M$, let next $\boldsymbol{\sigma}_{ih}$ be constructed following Definition 5.1 and define*

$$\eta_{\mathrm{res}}^2 := \sum_{i=m}^{M} \|\nabla\varphi_{ih} + \boldsymbol{\sigma}_{ih}\|^2. \tag{5.13}$$

Recall the notations (4.5) and (4.15). Then

$$0 \leq \sum_{i=m}^{M} (\lambda_{ih} - \lambda_i) \leq \eta^2, \tag{5.14}$$

where we distinguish the following two cases:
Case I (General case) *Let Assumption 2.6 hold. Then (5.14) holds with*

$$\eta^2 := (2c_h^2 + 2\lambda_{Mh}\tilde{c}_h^4\eta_{\mathrm{res}}^2)\eta_{\mathrm{res}}^2. \tag{5.15}$$

Case II (Optimal estimates under elliptic regularity assumption) *Assume that for $i = m, \dots, M$, the solutions $\zeta_{(ih)}$ of problems (5.11) belong to the space $H^{1+\delta}(\Omega)$, $0 < \delta \leq 1$, so that the approximation and stability estimates*

$$\min_{v_h \in V_h} \|\nabla(\zeta_{(ih)} - v_h)\| \leq C_{\mathrm{I}} h^\delta |\zeta_{(ih)}|_{H^{1+\delta}(\Omega)}, \tag{5.16a}$$

$$|\zeta_{(ih)}|_{H^{1+\delta}(\Omega)} \leq C_{\mathrm{S}} \|\mathscr{r}_{(ih)}\| \tag{5.16b}$$

are satisfied. Then (5.14) holds with

$$\eta^2 := (1 + 4\lambda_{Mh} c_h^2 C_I^2 C_S^2 h^{2\delta})\eta_{\mathrm{res}}^2. \tag{5.17}$$

Proof. **(Case I)** Combining the estimates (4.1), (4.6), (4.17) together with (2.18) and (5.10a) yield the result.
(Case II) The proof is as in Case I, relying (4.4) instead of (4.6) and on (4.16) instead of (4.17). Using the characterization (3.12) and similarly as in (3.14) in Lemma 3.8, one can show that

$$\|A^{-1/2}\mathrm{Res}(\gamma_h)\|_{\mathfrak{S}_2(\mathcal{H})}^2 = \sum_{i=m}^{M} \|\mathscr{r}_{(ih)}\|^2,$$

where $\mathscr{r}_{(ih)}$ is defined in (5.12a). Now, using an Aubin–Nitsche trick, (5.11), (5.12a), (3.10a) , and the discrete problem equation (5.2), we get

$$\|\mathscr{r}_{(ih)}\|^2 = (\nabla\zeta_{(ih)}, \nabla \mathscr{r}_{(ih)}) = (\nabla(\zeta_{(ih)} - \zeta_{ih}), \nabla \mathscr{r}_{(ih)}),$$

where $\zeta_{ih} \in V_h$ is the minimizer in (5.16a). Using the Cauchy–Schwarz inequality, estimates (5.16), and the characterization (5.12b) altogether give

$$\|\mathscr{r}_{(ih)}\| \leq C_{\mathrm{I}} C_{\mathrm{S}} h^\delta \|\mathrm{Res}(\varphi_{ih}, \lambda_{ih})\|_{H^{-1}(\Omega)}.$$

Therefore, also using Lemma 3.8,

$$\|A^{-1/2}\mathrm{Res}(\gamma_h)\|^2_{\mathfrak{S}_2(\mathcal{H})} \le (C_{\mathrm{I}}C_{\mathrm{S}}h^\delta)^2 \sum_{i=m}^{M} \|\mathrm{Res}(\varphi_{ih},\lambda_{ih})\|^2_{H^{-1}(\Omega)} = (C_{\mathrm{I}}C_{\mathrm{S}}h^\delta)^2 \|\mathrm{Res}(\gamma_h)\|^2_{\mathfrak{S}_2(\mathcal{H})}. \tag{5.18}$$

Thus, estimates (4.1), (4.4), (4.16) together with (2.18) and (5.10a) yields the result.

□

Remark 5.4 (Constants C_{I} and C_{S}). *As discussed in [9], it is possible to obtain explicit bounds for the constants C_{I} and C_{S} in particular cases, e.g., when Ω is a convex polygon in $\mathbb{R}^2$. In this case, the solution of the source problem $\zeta_{(ih)}$ of (5.11) belongs to $H^2(\Omega)$ and $|\zeta_{(ih)}|_{H^2(\Omega)} = \|\Delta\zeta_{(ih)}\| = \|\mathscr{r}_{(ih)}\|$, so it is possible to take $\delta = 1$ and $C_{\mathrm{S}} = 1$, see [27, Theorem 4.3.1.4]. Computable bounds for C_{I} can be found in Liu and Kikuchi [35], Carstensen* et al. *[14], and Liu and Oishi [36, Section 2]. Note that in the particular case of a mesh formed by isosceles rectangular triangles, there holds $C_{\mathrm{I}} \le \frac{0.493}{\sqrt{2}}$.*

Remark 5.5 (Improved guaranteed upper bounds for the eigenvalues). *Similarly as in [9, Theorem 5.2], it is possible to estimate $\|\mathrm{Res}(\varphi_{ih},\lambda_{ih})\|_{H^{-1}(\Omega)}$ from below and combine this lower bound with (4.1) and (4.9) to obtain guaranteed improved upper bounds for the eigenvalues. For brevity, we do not state such results here.*

Theorem 5.6 (Guaranteed and polynomial-degree robust bound for the density matrix error). *Let the assumptions of Theorem 5.3 be verified. Then the energy density matrix error can be bounded via*

$$\||\nabla|(\gamma^0 - \gamma_h)\|_{\mathfrak{S}_2(\mathcal{H})} \le \eta, \tag{5.19}$$

where η is defined in the Case I by (5.15) and in Case II by (5.17). Moreover, the density matrix error can be bounded by

$$\|\gamma^0 - \gamma_h\|_{\mathfrak{S}_2(\mathcal{H})} \le \eta_{L^2}, \tag{5.20}$$

where

$$\eta_{L^2} := \begin{cases} \sqrt{2}\tilde{c}_h\eta_{\mathrm{res}} & \text{(Case I)}, \quad (5.21a) \\ \sqrt{2}c_h C_{\mathrm{I}}C_{\mathrm{S}}h^\delta\eta_{\mathrm{res}} & \text{(Case II)}. \quad (5.21b) \end{cases}$$

Recall finally the definition of $\bar{c}_h$ by (4.8). Under Assumption 2.6, the estimator η is efficient as

$$\begin{aligned} \eta^2_{\mathrm{res}} \le (d+1)^2 C^2_{\mathrm{st}} C^2_{\mathrm{cont,PF}} \Bigg(& \bar{c}_h \||\nabla|(\gamma^0-\gamma_h)\|^2_{\mathfrak{S}_2(\mathcal{H})} + \frac{3(\lambda_M-\lambda_m)^2}{4\lambda_m}\|\gamma^0-\gamma_h\|^4_{\mathfrak{S}_2(\mathcal{H})} \\ & + \frac{3}{\lambda_m}\left(1+\frac{1}{4}\|\gamma^0-\gamma_h\|^4_{\mathfrak{S}_2(\mathcal{H})}\right)\times \\ & \left[2\left(1+\frac{\lambda_M}{4\lambda_m}\|\gamma^0-\gamma_h\|^2_{\mathfrak{S}_2(\mathcal{H})}\right)^2 \||\nabla|(\gamma^0-\gamma_h)\|^4_{\mathfrak{S}_2(\mathcal{H})} + 2(\lambda_M)^2\|\gamma^0-\gamma_h\|^4_{\mathfrak{S}_2(\mathcal{H})}\right]\Bigg). \end{aligned} \tag{5.22}$$

This in particular implies that the bound (5.19) is efficient in the sense that

$$\eta \le C\||\nabla|(\gamma^0-\gamma_h)\|_{\mathfrak{S}_2(\mathcal{H})}, \tag{5.23}$$

where C is a constant independent of the mesh size h and the polynomial degree p.

Proof. The proof of (5.19) is actually contained in the proof of (5.14) which relies on (4.1). In Case I, the estimate (5.20) follows from (4.17) and (5.10a), whereas in Case II, the bound (5.20) can be derived from (4.16) and (5.18) combined with (5.10a). The bound (5.22) is a consequence of (4.9) and (5.10b). Finally, (5.23) follows from (5.15) or (5.17) in combination with (5.22), the (crude) bound $\|\gamma^0-\gamma_h\|^2_{\mathfrak{S}_2(\mathcal{H})} \le 4J$, cf. (2.14), the equivalence (3.7), the Poincaré inequality $\|\Phi^0 - \Phi^0_h\|^2 \le \|\nabla(\Phi^0 - \Phi^0_h)\|^2/\lambda_1$, and the equivalence (3.8). □

5.2 Planewave discretization of a Schrödinger operator

In this section, we consider a Schrödinger-type operator of the form $-\Delta + V$, with periodic boundary conditions. We denote by $\Omega \subset \mathbb{R}^d$, $d = 1, 2, 3$, the periodic cell, by $\mathcal{R}$ the periodic lattice, and by $\mathcal{R}^*$ the corresponding dual lattice. For simplicity, we assume that $\Omega = [0, L)^d$, $(L > 0)$, in which case $\mathcal{R}$ is the cubic lattice $L\mathbb{Z}^d$, and $\mathcal{R}^* = 2\pi\mathbb{Z}^d$. Our arguments can be easily extended to the general case. The potential V is multiplicative and satisfies $V \in L^\infty_\#(\Omega)$, where, for $s \geq 1$,

$$L^s_\#(\Omega) = \left\{ v \in L^s_{\rm loc}(\mathbb{R}^d), \quad v \quad \mathcal{R}\text{-periodic} \right\}.$$

Up to shifting the operator $-\Delta + V$ by a positive constant, we can assume that $V \geq 1$.

For $\mathbf{k} \in \mathcal{R}^*$, we denote by $e_\mathbf{k}(x) = |\Omega|^{-1/2} e^{i\mathbf{k}\cdot x}$ the planewave with wavevector $\mathbf{k}$. The family $(e_\mathbf{k})_{\mathbf{k}\in\mathcal{R}^*}$ forms an orthonormal basis of $L^2_\#(\Omega)$. Moreover, for all $v \in L^2_\#(\Omega)$,

$$v(x) = \sum_{\mathbf{k}\in\mathcal{R}^*} \hat{v}_\mathbf{k} e_\mathbf{k}(x), \quad \text{where} \quad \hat{v}_\mathbf{k} = (e_\mathbf{k}, v)_{L^2_\#(\Omega)} = |\Omega|^{-1/2} \int_\Omega v(x) e^{-i\mathbf{k}\cdot x} dx.$$

Let us take in this case $\mathcal{H} = L^2_\#(\Omega)$ and $D(A^{1/2}) = \mathcal{V} := H^1_\#(\Omega)$ endowed with the norm

$$\|v\|^2_{D(A^{1/2})} := \|\nabla v\|^2 + (v, Vv) \geq \|v\|^2_{H^1_\#(\Omega)},$$

where we endow the Sobolev spaces of real-valued $\mathcal{R}$-periodic functions

$$H^s_\#(\Omega) := \Bigg\{ v(x) = \sum_{\mathbf{k}\in\mathcal{R}^*} \hat{v}_\mathbf{k} e_\mathbf{k}(x), \quad \text{where} \\ \|v\|^2_{H^1_\#(\Omega)} := \sum_{\mathbf{k}\in\mathcal{R}^*} (1 + |\mathbf{k}|^2)^s |\hat{v}_\mathbf{k}|^2 < \infty, \text{ and } \forall \mathbf{k},\, \hat{v}_{-\mathbf{k}} = \hat{v}^*_\mathbf{k} \Bigg\},$$

with the inner products

$$(v, w)_{H^s_\#(\Omega)} := \sum_{\mathbf{k}\in\mathcal{R}^*} (1 + |\mathbf{k}|^2)^s \overline{\hat{v}_\mathbf{k}} \hat{w}_\mathbf{k}.$$

Note that the constraints $\hat{v}_{-\mathbf{k}} = \hat{v}^*_\mathbf{k}$ imply that the functions are real-valued.

The eigenvalue problem reads in this case: find eigenvector and eigenvalue pairs (φ^0_i, λ_i) subject to the orthonormality constraints $(\varphi^0_i, \varphi^0_j) = \delta_{ij}$, $i, j \geq 1$, such that $(-\Delta + V)\varphi^0_i = \lambda_i \varphi^0_i$ in Ω. In weak form, this reads: find $(\varphi^0_i, \lambda_i) \in \mathcal{V} \times \mathbb{R}_+$ with $(\varphi^0_i, \varphi^0_j) = \delta_{ij}$ such that

$$(\nabla\varphi^0_i, \nabla v) + (\varphi^0_i, Vv) = \lambda_i(\varphi^0_i, v) \qquad \forall v \in \mathcal{V}. \tag{5.24}$$

For $N \in \mathbb{N}\backslash\{0\}$, we consider the approximation space

$$\mathcal{V}_N := \left\{ \sum_{\substack{\mathbf{k}\in\mathcal{R}^* \\ |\mathbf{k}|\leq \frac{2\pi}{L}N}} \hat{v}_\mathbf{k} e_\mathbf{k}(x), \quad \forall \mathbf{k},\, \hat{v}_{-\mathbf{k}} = \hat{v}^*_\mathbf{k} \right\}.$$

The discrete problem then reads: find eigenpairs $(\varphi_{iN}, \lambda_{iN}) \in \mathcal{V}_N \times \mathbb{R}_+$ with $(\varphi_{iN}, \varphi_{jN}) = \delta_{ij}$, $1 \leq i, j \leq N$, such that

$$(\nabla\varphi_{iN}, \nabla v_N) + (\varphi_{iN}, Vv_N) = \lambda_{iN}(\varphi_{iN}, v_N) \qquad \forall v_N \in \mathcal{V}_N. \tag{5.25}$$

Given $m, M \in \mathbb{N}\backslash\{0\}$, $M \geq m$, we focus on the eigenvalue cluster $(\lambda_{mN}, \ldots, \lambda_{MN})$ and a set of the associated eigenvectors $(\varphi_{mN}, \ldots, \varphi_{MN})$. Note that in this case, there actually holds $\mathcal{V}_N \subset D(A)$ and not merely $\mathcal{V}_N \subset D(A^{1/2})$ as supposed generally in (2.16).

5.2.1 Estimation of the dual norm of the residual

In order to use the error estimates defined in Section 4, we need to estimate the Hilbert–Schmidt norm of the residual $\mathrm{Res}(\gamma_h)$ defined in (3.11). As

$$A \geq -\Delta + 1 \geq 0, \quad A^{-1/2} \leq (-\Delta + 1)^{-1/2}, \tag{5.26}$$

the Hilbert–Schmidt norm of the residual can be estimated as follows, using the framework of Remark 3.6.

Corollary 5.7 (Hilbert–Schmidt norm of the residual estimate). *There holds*

$$\begin{aligned}
\|\mathrm{Res}(\gamma_h)\|^2_{\mathfrak{S}_2(\mathcal{H})} = \|A^{-1/2}\mathrm{Res}^h_{m,M}\|^2_{\mathfrak{S}_2(\mathcal{H})} &\leq \|(-\Delta+1)^{-1/2}\mathrm{Res}^h_{m,M}\|^2_{\mathfrak{S}_2(\mathcal{H})} \\
&= \sum_{i=m}^{M} \|\mathrm{Res}(\varphi_{iN}, \lambda_{iN})\|^2_{H^{-1}_{\#}(\Omega)} \\
&= \sum_{i=m}^{M} \sup_{\substack{v \in H^1_{\#}(\Omega) \\ \|v\|_{H^1_{\#}(\Omega)}=1}} \langle \mathrm{Res}(\varphi_{ih}, \lambda_{ih}), v\rangle_{H^{-1}_{\#}(\Omega), H^1_{\#}(\Omega)}.
\end{aligned} \tag{5.27}$$

Note that in the planewave setting, the Laplace operator is diagonal, so that in this case the quantity $\sum_{i=m}^{M} \|\mathrm{Res}(\varphi_{iN}, \lambda_{iN})\|^2_{H^{-1}_{\#}(\Omega)}$ can actually be computed exactly at a negligible cost.

Remark 5.8 (Estimate (5.27)). *Remark that inequality (5.26) is in fact independent of the choice of discretization. Therefore, (5.27) can also be used in the finite element setting, generalizing the estimates of Section 5.1 for a Schrödinger operator on a torus.*

Actually, using the same argument, there holds

$$\begin{aligned}
\|A^{-1/2}\mathrm{Res}(\gamma_h)\|^2_{\mathfrak{S}_2(\mathcal{H})} = \|A^{-1}\mathrm{Res}^h_{m,M}\|^2_{\mathfrak{S}_2(\mathcal{H})} &\leq \|(-\Delta+1)^{-1}\mathrm{Res}^h_{m,M}\|^2_{\mathfrak{S}_2(\mathcal{H})} \\
&= \sum_{i=m}^{M} \|\mathrm{Res}(\varphi_{iN}, \lambda_{iN})\|^2_{H^{-2}_{\#}(\Omega)}.
\end{aligned} \tag{5.28}$$

Since for $i = m, \dots, M$, $\mathrm{Res}(\varphi_{iN}, \lambda_{iN}) \in \mathcal{V}_N{}^{\perp}$, the orthogonal space of $\mathcal{V}_N$ with respect to any $H^s_{\#}$ scalar product, there holds

$$\|\mathrm{Res}(\varphi_{iN}, \lambda_{iN})\|^2_{H^{-s}_{\#}(\Omega)} = \sum_{\substack{\mathbf{k}\in\mathcal{R}^* \\ |\mathbf{k}| \geq \frac{2\pi}{L}N}} (1 + |\mathbf{k}|^2)^{-s} |\hat{r}^i_{\mathbf{k}}|^2,$$

where for $i = m, \dots, M$, $(\hat{r}^i_{\mathbf{k}})_{\mathbf{k}\in\mathcal{R}^*}$ are the planewave coefficients of $\mathrm{Res}(\varphi_{iN}, \lambda_{iN})$. Hence,

$$\|\mathrm{Res}(\varphi_{iN}, \lambda_{iN})\|_{H^{-2}_{\#}(\Omega)} \leq \frac{L}{2\pi}\frac{1}{N}\|\mathrm{Res}(\varphi_{iN}, \lambda_{iN})\|_{H^{-1}_{\#}(\Omega)}. \tag{5.29}$$

5.2.2 Final estimates

We now state the guaranteed and fully computable error bounds for eigenvalues and density matrices of the operator $-\Delta + V$ discretized with planewaves.

Theorem 5.9 (Guaranteed bounds for the sum of eigenvalues). *Let $m, M \in \mathbb{N}\backslash\{0\}, M \geq m$, and let Assumption 2.1 hold. For $i = m, \dots, M$, let $(\varphi_{iN}, \lambda_{iN}) \in \mathcal{V}_N \times \mathbb{R}_+$ be defined in (5.25). Let $\underline{\lambda}_{M+1}$, and $\overline{\lambda}_{m-1}$ if $m > 1$ satisfying Assumption 4.2 respectively with λ_{mN} and λ_{MN} in place of λ_{mh} and λ_{Mh}, i.e.*

$$\lambda_{m-1} \leq \overline{\lambda}_{m-1} < \lambda_{mN} \text{ when } m > 1, \qquad \lambda_{MN} < \underline{\lambda}_{M+1} \leq \lambda_{M+1}, \tag{5.30}$$

where we take $\overline{\lambda}_{m-1} = \lambda_{m-1N}$. For $i = m, \dots, M$, define

$$\eta^2_{\mathrm{res}} := \sum_{i=m}^{M} \|\mathrm{Res}(\varphi_{iN}, \lambda_{iN})\|^2_{H^{-1}_{\#}(\Omega)}.$$

Set

$$c_N := \max\left[\left(\frac{\lambda_{mN}}{\underline{\lambda}_{m-1}} - 1\right)^{-1}, \left(1 - \frac{\lambda_{MN}}{\underline{\lambda}_{M+1}}\right)^{-1}\right], \tag{5.31}$$

with the first term in the max *discarded if* $m = 1$. *Then*

$$0 \le \sum_{i=m}^{M} (\lambda_{ih} - \lambda_i) \le \eta^2, \tag{5.32}$$

where

$$\eta^2 := \left(1 + \frac{1}{N^2}\frac{L^2\lambda_{MN}}{\pi^2}c_N^2\right)\eta_{\mathrm{res}}^2. \tag{5.33}$$

Proof. Combining the estimates (4.1), (4.4), (4.16), (5.27), (5.28), and (5.29) yields the result. □

Please note that in practice, condition (5.30) can be verified as in Remark 4.3.

Theorem 5.10 (Guaranteed and robust bound for the density matrix errors)**.** *Let the assumptions of Theorem 5.9 be verified. Then the energy density matrix error can be bounded via*

$$\|(-\Delta + V)^{1/2}(\gamma^0 - \gamma_h)\|_{\mathfrak{S}_2(\mathcal{H})} \le \eta, \tag{5.34}$$

where η *is defined by* (5.33). *Moreover, the density matrix error can be bounded by*

$$\|\gamma^0 - \gamma_h\|_{\mathfrak{S}_2(\mathcal{H})} \le \eta_{L^2} := \sqrt{2}c_N\frac{L}{2\pi N}\eta_{\mathrm{res}}. \tag{5.35}$$

Recall finally the definition of $\bar{c}_h$ *by* (4.8), *with* λ_{MN} *in place of* λ_{Mh}. *Under Assumption 2.6, the estimator* η *is efficient as*

$$\begin{aligned}\eta_{\mathrm{res}}^2 \le (\sup_\Omega V)\Bigg(&\bar{c}_h\|(-\Delta+V)^{1/2}(\gamma^0-\gamma_h)\|^2_{\mathfrak{S}_2(\mathcal{H})} + \frac{3(\lambda_M-\lambda_m)^2}{4\lambda_m}\|\gamma^0-\gamma_h\|^4_{\mathfrak{S}_2(\mathcal{H})}\\ &+\frac{3}{\lambda_m}\left(1+\frac{1}{4}\|\gamma^0-\gamma_h\|^4_{\mathfrak{S}_2(\mathcal{H})}\right)\times\\ &\left[2\left(1+\frac{\lambda_M}{4\lambda_m}\|\gamma^0-\gamma_h\|^2_{\mathfrak{S}_2(\mathcal{H})}\right)^2\|(-\Delta+V)^{1/2}(\gamma^0-\gamma_h)\|^4_{\mathfrak{S}_2(\mathcal{H})}\right.\\ &\left.+2(\lambda_M)^2\|\gamma^0-\gamma_h\|^4_{\mathfrak{S}_2(\mathcal{H})}\right]\Bigg).\end{aligned} \tag{5.36}$$

Proof. The proof of (5.34) follows from the proof of (5.32). The estimate (5.35) follows from (4.16), (5.28), and (5.29). Finally, the bound (5.36) is a consequence of (4.9) and the inequality

$$\forall v \in H^1_\#(\Omega), \quad \|(-\Delta+V)^{1/2}v\|^2 = \|\nabla v\|^2_{L^2_\#} + \int_\Omega Vv^2 \le (\sup_\Omega V)\|v\|^2_{H^1_\#(\Omega)},$$

which yields

$$\forall v \in H^{-1}_\#(\Omega), \quad \|(-\Delta+V)^{-1/2}v\|^2 \ge \frac{1}{(\sup_\Omega V)}\|v\|^2_{H^{-1}_\#}.$$

□

6 Numerical experiments

We now present some numerical results for two different examples. First, we perform simulations for the Laplace eigenvalue problem discretized with finite elements. Second, we show the estimates obtained for a Schrödinger operator on the torus discretized with planewaves.

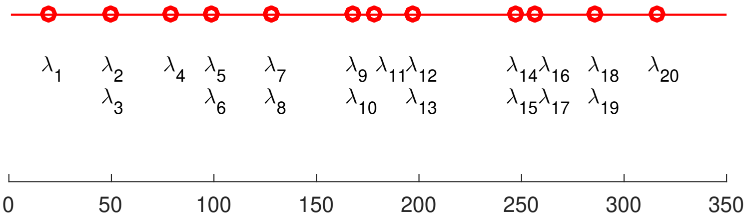


Figure 1: Plot of the first 20 eigenvalues of the Laplace operator with homogeneous Dirichlet boundary conditions on the unit square.

6.1 Laplace operator discretized with finite elements

We start with a series of numerical examples using the conforming finite element method with piecewise linear polynomials, i.e., $p = 1$, as presented in Section 5.1 for the Laplace eigenvalue problem. We consider either the square domain $\Omega = (0, 1)^2$ or an L-shaped domain with homogeneous Dirichlet conditions. For the flux equilibration, we use the cheap Raviart–Thomas–Nédélec space of degree $p = 1$. This still provides guaranteed upper bounds, see the proof of Theorem 5.2, and we do not observe any asymptotic loss of the effectivity. The numerical tests are performed with the FreeFem++ code [29].

Theorem 5.3 requires a lower bound $\underline{\lambda}_{M+1}$, (recall that for the upper bound $\overline{\lambda}_{m-1}$ if $m > 1$, we simple use the numerically computed eigenvalue, i.e., $\overline{\lambda}_{m-1} = \lambda_{(m-1)h} \geq \lambda_{m-1}$, relying on the variational principle (2.18)). A guaranteed lower bound $\underline{\lambda}_{M+1}$ is obtained by employing the nonconforming finite element method on a coarse mesh $\mathcal{T}_H$ and using the technique presented in formula (6) of [34].

In the presentation of the results, we use the following notation:

$$\mathtt{Err}_\lambda := \sum_{i=m}^{M} (\lambda_{ih} - \lambda_i), \quad \mathtt{Err}_{H^1} := \| |\nabla|(\gamma^0 - \gamma_h)\|_{\mathfrak{S}_2(\mathcal{H})}, \quad \mathtt{Err}_{L^2} := \|\gamma^0 - \gamma_h\|_{\mathfrak{S}_2(\mathcal{H})}. \tag{6.1}$$

The effectivity indices are then defined by

$$I_\lambda^{\mathtt{eff}} := \frac{\eta^2}{\mathtt{Err}_\lambda}, \qquad I_{H^1}^{\mathtt{eff}} := \frac{\eta}{\mathtt{Err}_{H^1}}, \qquad I_{L^2}^{\mathtt{eff}} := \frac{\eta_{L^2}}{\mathtt{Err}_{L^2}},$$

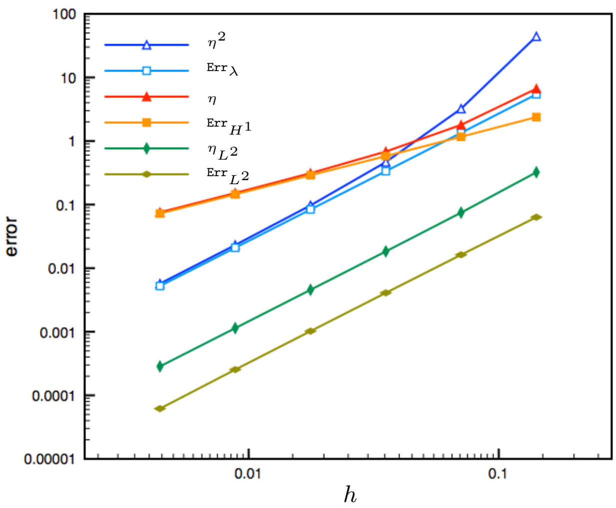
where η and η_{L^2} are respectively defined in (5.15) and (5.21a) for Case I and (5.17) and (5.21b) for Case II.

6.1.1 Unit square

We first consider the unit square $\Omega = (0, 1)^2$ where explicit eigenpairs are known. Indeed, the sequence of eigenvalues is given by $\pi^2(k^2 + l^2)$, $k, l \in \mathbb{N}$, and the corresponding eigenvectors are $u_{k,l} = \sin(k\pi x)\sin(l\pi y)$. The first few eigenvalues are therefore given by $\pi^2, 5\pi^2, 5\pi^2, 8\pi^2, \ldots$ yielding a gap between the first and second, and the third and forth eigenvalues for example. Figure 1 illustrates the first 20 eigenvalues and indicates the multiplicities. For small eigenvalues, we use a coarse mesh $\mathcal{T}_{H,1}$ consisting of 121 triangles and 320 degrees of freedom and for larger eigenvalues, we use a second coarse mesh $\mathcal{T}_{H,2}$ consisting of 441 triangles and 1,240 degrees of freedom. Since the domain is a convex polygon, we can apply **Case II** in Theorems 5.3 and 5.6 which exploits elliptic regularity results. We will here consider sequences of structured and uniformly refined meshes and use constants $C_{\mathrm{I}} = \frac{0.493}{\sqrt{2}}$, $C_{\mathrm{S}} = 1$, and $\delta = 1$ following Remark 5.4.

We first analyze the quality of the estimators for $m = 2, M = 3$. The guaranteed lower bound is computed on the coarse mesh $\mathcal{T}_{H,1}$ yielding $\underline{\lambda}_4 \approx 73.9444$. Figure 2 (top) illustrates the convergence of the error quantities $\mathtt{Err}_\lambda$, $\mathtt{Err}_{H^1}$, and $\mathtt{Err}_{L^2}$ as well as the corresponding upper bounds η^2, η, η_{L^2}, whereas Table 4 (top) reports the effectivity indices.

We next analyze the effectivity indices of the estimator as we increase the index of the eigenvalues, still considering clusters of size 2. Table 4 (bottom) compares the results for the clusters corresponding to



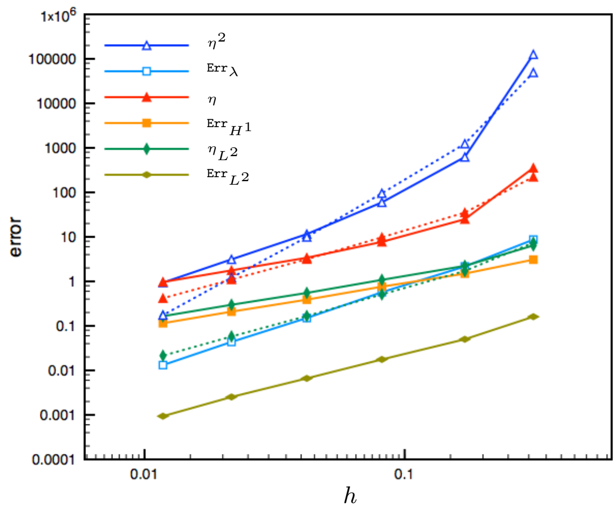


Figure 2: Convergence of various measures of the error and their upper bounds for finite elements and the unit square with $m = 2$, $M = 3$ (top) and for the L-shaped domain with $m = 3$, $M = 5$ (bottom). The dotted lines provide the estimators using Case II with $\delta = \frac{2}{3}$ and the empirical choice $C_{\mathrm{S}} = C_{\mathrm{I}} = 1$.

$m = 9, M = 10$ and $m = 18, M = 19$. Since higher eigenvalues are sought for, we considered the second coarse mesh $\mathcal{T}_{H,2}$ for computing $\underline{\lambda}_{M+1}$. This yields

$$\underline{\lambda}_{11} \approx 171.135, \quad \underline{\lambda}_{20} \approx 295.777.$$

The results confirm the theoretical findings of Section 5.1.2. In particular, all the bounds are guaranteed, with effectivity indices taking values above one. Moreover, numerically, we observe asymptotic exactness of the estimators η^2 and η of $\mathtt{Err}_\lambda$, respectively $\mathtt{Err}_{H^1}$, meaning that the corresponding effectivity indices $I_\lambda^{\mathtt{eff}}$ and $I_{H^1}^{\mathtt{eff}}$ tend to the optimal value of one. Additionally, we also numerically observe that the effectivity indices are robust with respect to increasing indices of the cluster of fixed size, which we could not cover in our theory. Indeed, for the efficiency bound (5.22) of Theorem 5.6, the (exploding) factor $\bar{c}_h$ appears.

	N	h	ndof	$\mathtt{Err}_\lambda$	η^2	$I_\lambda^{\mathtt{eff}}$	$\mathtt{Err}_{H^1}$	η	$I_{H^1}^{\mathtt{eff}}$	$\mathtt{Err}_{L^2}$	η_{L^2}	$I_{L^2}^{\mathtt{eff}}$
$m=2$	40	0.0354	1681	0.3351	0.4661	1.39	0.5788	0.6827	1.18	0.0041	0.0183	4.49
$M=3$	80	0.0177	6561	0.0837	0.0972	1.16	0.2890	0.3118	1.08	0.0010	0.0046	4.47
$\mathcal{T}_{H,1}$	160	0.0088	25921	0.0209	0.0231	1.10	0.1445	0.1521	1.05	0.0003	0.0011	4.49
	320	0.0044	103041	0.0052	0.0057	1.09	0.0722	0.0755	1.05	0.0001	0.0003	4.62
$m=9$	40	0.0354	1681	3.2698	3714.3421	1135.96	1.8235	60.9454	33.42	0.0194	0.3295	17.01
$M=10$	80	0.0177	6561	0.8151	76.6523	94.04	0.9037	8.7551	9.69	0.0049	0.0622	12.81
$\mathcal{T}_{H,2}$	160	0.0088	25921	0.2036	4.0755	20.02	0.4508	2.0188	4.48	0.0012	0.0148	12.17
	320	0.0044	103041	0.0509	0.2842	5.58	0.2253	0.5331	2.37	0.0003	0.0036	12.03
$m=18$	40	0.0354	1681	10.6565	10777.4005	1011.34	3.4872	103.8143	29.77	0.0729	0.5069	6.95
$M=19$	80	0.0177	6561	2.6465	166.0018	62.73	1.6537	12.8842	7.79	0.0183	0.0887	4.86
$\mathcal{T}_{H,2}$	160	0.0088	25921	0.6605	8.7166	13.20	0.8152	2.9524	3.62	0.0046	0.0209	4.57
	320	0.0044	103041	0.1651	0.6511	3.94	0.4061	0.8069	1.99	0.0011	0.0051	4.50

Table 4: [Finite elements, unit square, Case II] Errors, estimates, and effectivity indices for clusters of size 2 and increasing index of the eigenvalues. The values of m and M are indicated on the far left as well as the type of coarse mesh $\mathcal{T}_{H,i}$ used to obtain the auxiliary guaranteed lower bounds $\underline{\lambda}_{M+1}$.

	N	h	ndof	$\mathtt{Err}_\lambda$	η^2	$I_\lambda^{\mathtt{eff}}$	$\mathtt{Err}_{H^1}$	η	$I_{H^1}^{\mathtt{eff}}$	$\mathtt{Err}_{L^2}$	η_{L^2}	$I_{L^2}^{\mathtt{eff}}$
$m=1$	10	0.1414	121	13.5049	21673.5051	1604.86	4.1325	147.2192	35.63	0.2141	1.7415	8.13
$M=4$	20	0.0707	441	3.4018	98.8430	29.06	1.9076	9.9420	5.21	0.0554	0.2274	4.10
$\mathcal{T}_{H,1}$	40	0.0354	1681	0.8519	5.0687	5.95	0.9297	2.2514	2.42	0.0139	0.0521	3.75
	80	0.0177	6561	0.2131	0.4708	2.21	0.4619	0.6862	1.49	0.0035	0.0128	3.67
	160	0.0088	25921	0.0533	0.0728	1.37	0.2306	0.2698	1.17	0.0009	0.0032	3.67
	320	0.0044	103041	0.0133	0.0155	1.16	0.1152	0.1243	1.08	0.0002	0.0008	3.71
$m=1$	10	0.1414	121	72.9222	82403.2050	1130.02	9.3347	287.0596	30.75	0.3359	3.2521	9.68
$M=8$	20	0.0707	441	18.0492	281.4040	15.59	4.3588	16.7751	3.85	0.0874	0.3923	4.49
$\mathcal{T}_{H,2}$	40	0.0354	1681	4.4994	15.9735	3.55	2.1323	3.9967	1.87	0.0221	0.0893	4.04
	80	0.0177	6561	1.1240	1.8566	1.65	1.0603	1.3626	1.29	0.0055	0.0219	3.94
	160	0.0088	25921	0.2810	0.3445	1.23	0.5294	0.5869	1.11	0.0014	0.0054	3.94
	320	0.0044	103041	0.0702	0.0788	1.12	0.2646	0.2808	1.06	0.0003	0.0014	4.00

Table 5: [Finite elements, unit square, Case II] Errors, estimates, and effectivity indices for clusters of increasing size. The values of m and M are indicated on the far left as well as the type of coarse mesh $\mathcal{T}_{H,i}$ used to obtain the auxiliary guaranteed lower bounds $\underline{\lambda}_{M+1}$.

Next, we consider clusters of increasing size. We consider the choices $m = 1, M = 4$ resp. $m = 1, M = 8$ and present the results in Table 5. We observe that the effectivity indices are also numerically robust when doubling the size of the cluster.

6.1.2 L-shaped domain

We now address the case of an L-shaped domain $\Omega := (-1, 1)^2 \setminus ([0, 1] \times [-1, 0])$. Note that in this setting, only **Case I**, to our knowledge, is currently applicable in Theorems 5.3 and 5.6 to obtain guaranteed bounds (cf. Remark 5.4). The first few eigenvalues are known to high accuracy [42]

$$\lambda_1 \approx 9.6397238, \qquad \lambda_2 \approx 15.197252, \qquad \lambda_3 \approx 19.739209,$$
$$\lambda_4 \approx 29.521481, \qquad \lambda_5 \approx 31.912636, \qquad \lambda_6 \approx 41.474510.$$

We focus on the cluster from the third ($m = 3$) to the fifth ($M = 5$) eigenvalues. A sequence of non-structured quasi-uniform meshes is considered first. We test the estimator for the lower bounds of λ_6 that are computed on two different coarse meshes $\mathcal{T}_{H,1}$ and $\mathcal{T}_{H,1}$, with 105 triangles resulting in 272 degrees of freedom resp. with 372 triangles resulting in 1033 degrees of freedom. This yields a lower bound $\underline{\lambda}_6 \approx 34.0774$ resp. $\underline{\lambda}_6 \approx 39.1209$. The convergence plots are reported in Figure 2 (bottom) for the latter case, and Table 6 presents the effectivity indices in both cases. We remark that the bound η_{L^2} for $\|\gamma^0 - \gamma_h\|_{\mathfrak{S}_2(\mathcal{H})}$ is guaranteed but of a much worse quality in this case, as the effectivity index $I_{L^2}^{\mathtt{eff}}$ increases with the number of degrees of freedom.

While Case I is always applicable and thus justified, Case II also applies theoretically with $\delta = \frac{2}{3}$ for the L-shaped domain. The limiting issue is that the constants C_{S} and C_I are unknown so that any empirical

	N	h	ndof	$\mathtt{Err}_\lambda$	η^2	I_λ^{eff}	$\mathtt{Err}_{H^1}$	η	$I_{H^1}^{\mathrm{eff}}$	$\mathtt{Err}_{L^2}$	η_{L^2}	$I_{L^2}^{\mathrm{eff}}$
$m=3$	20	0.1703	372	2.1603	320733.4214	148468.40	1.4948	566.3333	378.87	0.0500	5.1000	101.92
$M=5$	40	0.0817	1426	0.5710	3020.5208	5289.65	0.7607	54.9593	72.25	0.0176	2.0122	114.26
$\mathcal{T}_{H,1}$	80	0.0421	5734	0.1503	211.0547	1403.82	0.3886	14.5277	37.39	0.0066	0.9843	148.78
	160	0.0216	22001	0.0436	35.1498	806.13	0.2089	5.9287	28.38	0.0025	0.5277	208.68
	320	0.0118	86787	0.0132	8.7007	661.24	0.1149	2.9497	25.68	0.0009	0.2917	311.83
$m=3$	10	0.3124	105	8.6772	126111.0898	14533.55	3.0801	355.1212	115.30	0.1608	6.4197	39.93
$M=5$	20	0.1703	372	2.1603	622.3367	288.08	1.4948	24.9467	16.69	0.0500	2.2311	44.59
$\mathcal{T}_{H,2}$	40	0.0817	1426	0.5710	59.5714	104.32	0.7607	7.7182	10.15	0.0176	1.0820	61.44
	80	0.0421	5734	0.1503	11.5424	76.77	0.3886	3.3974	8.74	0.0066	0.5505	83.21
	160	0.0216	22001	0.0436	3.1223	71.61	0.2089	1.7670	8.46	0.0025	0.2980	117.86
	320	0.0118	86787	0.0132	0.9370	71.21	0.1149	0.9680	8.43	0.0009	0.1652	176.63

Table 6: [Finite elements, L-shaped domain, Case I] Errors, estimates, and effectivity indices for the cluster with $m = 3$, $M = 5$. The type of the coarse mesh $\mathcal{T}_{H,i}$ used to obtain the auxiliary guaranteed lower bounds $\underline{\lambda}_{M+1}$ is indicated on the far left.

	N	h	ndof	$\mathtt{Err}_\lambda$	η^2	I_λ^{eff}	$\mathtt{Err}_{H^1}$	η	$I_{H^1}^{\mathrm{eff}}$	$\mathtt{Err}_{L^2}$	η_{L^2}	$I_{L^2}^{\mathrm{eff}}$
$m=3$	20	0.1703	372	2.1603	29382.3983	13601.19	1.4948	171.4129	114.67	0.0500	3.7863	75.67
$M=5$	40	0.0817	1426	0.5710	987.9690	1730.17	0.7607	31.4320	41.32	0.0176	0.9157	52.00
$\mathcal{T}_{H,1}$	80	0.0421	5734	0.1503	86.9332	578.23	0.3886	9.3238	24.00	0.0066	0.2876	43.47
	160	0.0216	22001	0.0436	10.0040	229.43	0.2089	3.1629	15.14	0.0025	0.0989	39.10
	320	0.0118	86787	0.0132	1.3585	103.25	0.1149	1.1656	10.15	0.0009	0.0365	38.99
$m=3$	10	0.3124	105	8.6772	49345.4041	5686.76	3.0801	222.1383	72.12	0.1608	7.3922	45.98
$M=5$	20	0.1703	372	2.1603	1237.8702	573.01	1.4948	35.1834	23.54	0.0500	1.7146	34.27
$\mathcal{T}_{H,2}$	40	0.0817	1426	0.5710	95.4292	167.12	0.7607	9.7688	12.84	0.0176	0.5097	28.94
	80	0.0421	5734	0.1503	9.9167	65.96	0.3886	3.1491	8.10	0.0066	0.1665	25.17
	160	0.0216	22001	0.0436	1.2135	27.83	0.2089	1.1016	5.27	0.0025	0.0578	22.86
	320	0.0118	86787	0.0132	0.1744	13.25	0.1149	0.4176	3.64	0.0009	0.0214	22.86

Table 7: [Finite elements, L-shaped domain, Case II (with the empirical choice $C_{\mathrm{S}} = C_{\mathrm{I}} = 1$)] Errors, estimates, and effectivity indices for the cluster with $m = 3$, $M = 5$. The type of the coarse mesh $\mathcal{T}_{H,i}$ used to obtain the auxiliary guaranteed lower bounds $\underline{\lambda}_{M+1}$ is indicated on the far left.

choice of these constants yields an error indicator but not a guaranteed estimator. We have tested this indicator in Case II with $\delta = \frac{2}{3}$ and $C_{\mathrm{S}} = C_{\mathrm{I}} = 1$ in order to obtain indicators that are no longer guaranteed but, on the other hand, have asymptotically the multiplicative pre-factor equal to 1, see also Remark 5.4. Table 7 presents the effectivity indices that are now all decreasing (including the one of η_{L^2}) and the dotted lines in Figure 2 (bottom) present this indicator in Case II.

We finally test an adaptive refinement strategy using the local character of the density matrix estimator (5.13) (Case I)

$$\eta^2 = \sum_{K \in \mathcal{T}_h} \eta_K^2$$

with

$$\eta_K^2 = (2c_h^2 + 2\lambda_{Mh}\tilde{c}_h^4\eta_{\mathrm{res}}^2) \sum_{i=m}^{M} \|\nabla\varphi_{ih} + \boldsymbol{\sigma}_{ih}\|_K^2.$$

We employ the Dörfler marking strategy [19] with $\theta = 0.6$ and the newest vertex bisection mesh refinement. The initial mesh is unstructured with 103 degrees of freedom. The same lower bounds (using the mesh $\mathcal{T}_{H,2}$) as for the uniform refinement have been used and we note that Assumption 4.2 is satisfied for the initial mesh.

Figure 3 illustrates the error in the eigenvalues and density matrix as well as their upper bounds (left) and the mesh at the 10th iteration of the adaptive mesh refinement procedure (right). The optimal convergence rates are indicated by dashed lines. Table 8 then presents more details including effectivity indices. We observe quasi-optimal convergence with respect to the number of unknowns on the generated sequence of meshes and a loss of effectivity of η_{L^2} as in the case of uniform refinement.

Level	h	ndof	$\mathtt{Err}_\lambda$	η^2	$I_\lambda^{\mathtt{eff}}$	$\mathtt{Err}_{H^1}$	η	$I_{H^1}^{\mathtt{eff}}$	$\mathtt{Err}_{L^2}$	η_{L^2}	$I_{L^2}^{\mathtt{eff}}$
4	0.2687	311	2.6872	971.5901	361.57	1.6678	31.1703	18.69	0.0569	2.5161	44.20
8	0.2000	1503	0.5247	50.8697	96.95	0.7264	7.1323	9.82	0.0114	1.0228	89.74
12	0.1000	6944	0.1086	7.6657	70.62	0.3299	2.7687	8.39	0.0024	0.4568	187.50
16	0.0500	31608	0.0233	1.5273	65.69	0.1548	1.2358	7.98	0.0005	0.2103	394.88

Table 8: [Adaptive mesh refinement, L-shaped domain, Case I]

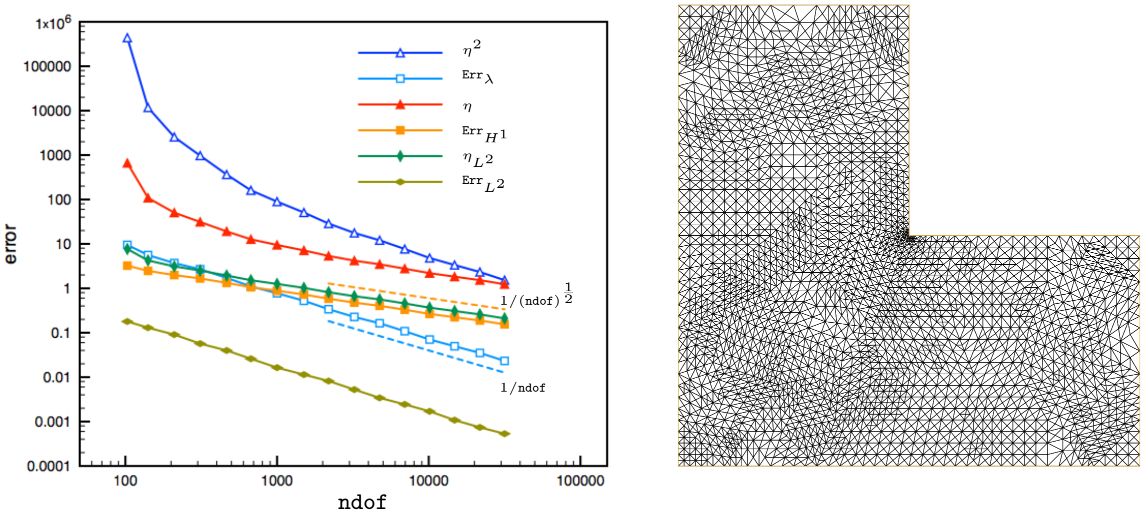


Figure 3: Convergence of various measures of the error and their upper bounds for finite elements on the L-shaped domain with $m = 3$, $M = 5$ using an adaptive refinement strategy (left). The mesh at the 10th iteration (right).

6.2 A Schrödinger operator discretized with planewaves

We consider in this section a Schrödinger operator of the form $-\Delta + V$ on $L^2_\#((0, 2\pi)^d)$, where $V \in L^\infty_\#((0, 2\pi)^d)$, $V \geq 1$, $d = 1, 2$. The problem is discretized with planewaves, and falls into the setting presented in Section 5.2. Using the notation of Section 5.2, $L = 2\pi$. The potential V is defined by its Fourier coefficients $\hat{V}_{\mathbf{k}}$, $\mathbf{k} \in \mathbb{Z}^d$, which are of the form

$$\forall \mathbf{k} \in \mathbb{Z}^d \backslash \{0\}, \quad \hat{V}_{\mathbf{k}} = \frac{\alpha}{|\mathbf{k}|^2}, \tag{6.2}$$

with $\alpha > 0$ given and $\hat{V}_0$ such that $\min_{\mathrm{x} \in (0,2\pi)^d} V(\mathrm{x}) = 1$.

For the implementation of the bounds, note that the eigenvalues of the operator $-\Delta + 1$, which are explicitly known, are lower bounds for the eigenvalues of $-\Delta + V$, since $V \geq 1$. Moreover, the eigenvalues computed in the basis with planewave cutoff N are upper bounds of the exact eigenvalues, due to the variational principle. The constant c_N defined in (5.31) can therefore be computed with these bounds of the eigenvalues.

The notation used here is similar to the notation of Section 6.1, see (6.1). In particular, the effectivity indices are defined by

$$I_\lambda^{\mathtt{eff}} := \frac{\eta^2}{\mathtt{Err}_\lambda}, \qquad I_{H^1}^{\mathtt{eff}} := \frac{\eta}{\mathtt{Err}_{H^1}}, \qquad I_{L^2}^{\mathtt{eff}} := \frac{\eta_{L^2}}{\mathtt{Err}_{L^2}},$$

where η and η_{L^2} are resp. defined in (5.33) and (5.35).

	N	ndof	$\mathtt{Err}_\lambda$	η^2	I_λ^{eff}	$\mathtt{Err}_{H^1}$	η	$I_{H^1}^{\mathrm{eff}}$	$\mathtt{Err}_{L^2}$	η_{L^2}	$I_{L^2}^{\mathrm{eff}}$
$m=2$	10	21	2.81e-06	2.71e-05	9.65	1.72e-03	5.21e-03	3.03	1.89e-04	1.68e-03	8.89
$m=3$	50	101	1.18e-09	1.59e-09	1.35	3.44e-05	3.99e-05	1.16	8.12e-07	6.91e-06	8.50
	90	181	6.39e-11	7.08e-11	1.11	8.00e-06	8.41e-06	1.05	1.06e-07	8.94e-07	8.46
	130	261	1.02e-11	1.08e-11	1.05	3.20e-06	3.28e-06	1.03	2.93e-08	2.48e-07	8.44
$m=10$	10	21	2.74e-05	4.20e-04	15.3	6.35e-03	2.05e-02	3.23	6.83e-04	2.68e-03	3.93
$M=11$	50	101	2.78e-09	4.70e-09	1.69	5.32e-05	6.85e-05	1.29	1.26e-06	5.91e-06	4.70
	90	181	1.44e-10	1.75e-10	1.21	1.20e-05	1.32e-05	1.10	1.59e-07	7.48e-07	4.71
	130	261	2.29e-11	2.51e-11	1.10	4.78e-06	5.01e-06	1.05	4.38e-08	2.06e-07	4.71
$m=16$	10	21	4.41e-04	1.74e-02	39.5	3.67e-02	1.32e-01	3.59	3.69e-03	1.14e-02	3.08
$M=17$	50	101	3.00e-09	1.19e-08	3.97	5.59e-05	1.09e-04	1.96	1.32e-06	8.21e-06	6.22
	90	181	1.43e-10	2.76e-10	1.93	1.20e-05	1.66e-05	1.38	1.59e-07	1.00e-06	6.31
	130	261	2.24e-11	3.22e-11	1.44	4.73e-06	5.68e-06	1.20	4.34e-08	2.74e-07	6.32
$m=1$	10	21	3.88e-05	3.82e-04	9.83	6.80e-03	1.95e-02	2.88	7.43e-04	3.04e-03	4.09
$M=9$	50	101	1.02e-08	1.41e-08	1.38	1.01e-04	1.19e-04	1.17	2.40e-06	1.02e-05	4.25
	90	181	5.46e-10	6.10e-10	1.12	2.34e-05	2.47e-05	1.06	3.09e-07	1.31e-06	4.24
	130	261	8.72e-11	9.22e-11	1.06	9.35e-06	9.60e-06	1.03	8.57e-08	3.62e-07	4.23
$m=1$	10	21	3.18e-03	1.49e-01	46.9	1.18e-01	3.86e-01	3.28	1.15e-02	1.53e-02	3.53
$M=17$	50	101	2.49e-08	1.36e-07	5.49	1.59e-04	3.69e-04	2.32	3.76e-06	2.22e-05	6.32
	90	181	1.26e-09	3.02e-09	2.39	3.57e-05	5.50e-05	1.54	4.71e-07	2.81e-06	6.34
	130	261	2.00e-10	3.33e-10	1.67	1.41e-05	1.82e-05	1.29	1.30e-07	7.74e-07	6.34

Table 9: [Planewaves, one-dimensional case, Schrödinger operator with $\alpha = 1$ in (6.2)] Errors, estimates, and effectivity indices for different clusters of eigenvalues. The values of m and M are indicated on the far left.

6.2.1 One-dimensional simulations

In the following simulations, we take $d = 1$ and $\alpha = 1$. For this potential, we compute reference eigenvectors and eigenvalues taking $N = 600$. We then compute approximate eigenvectors for different values of the discretization parameter N varying from 10 to 130. For all the chosen eigenvalue clusters, the assumptions required for Theorems 5.9 and 5.10 are already satisfied for $N = 10$.

We first assess the quality of the estimators for $m = 2, M = 3$. Figure 4 illustrates the convergence of the error quantities $\mathtt{Err}_\lambda$, $\mathtt{Err}_{H^1}$, and $\mathtt{Err}_{L^2}$ as well as the corresponding upper bounds η^2, η, η_{L^2} and Table 9 (top) reports the corresponding effectivity indices. We observe that the estimators η^2 and η are numerically asymptotically exact.

We then consider clusters of increasing indices and increasing size. Namely, we take $m = 10$, $M = 11$ and $m = 16$, $M = 17$, as well as $m = 1$, $M = 9$ and $m = 1$, $M = 17$. The results presented in Table 9 confirm excellent efficiency and robustness of the bounds in all the considered situations.

6.2.2 Two-dimensional simulations

We now take $d = 2$ and first use $\alpha = 0.1$. For this potential, we compute reference eigenvectors and eigenvalues taking $N = 50$, the number of degrees of freedom being $(2N+1)^2$. We then compute approximate eigenvectors and eigenvalues for different N varying from 5 to 25. We compute the error bounds as well as the effectivity indices for different clusters of eigenvalues, namely $m = 1$, $M = 5$, then $m = 6$, $M = 9$, and finally $m = 10$, $M = 13$. The eigenvalue clusters are chosen such that the gaps between the cluster and the surrounding eigenvalues are rather large, in practice > 0.87. The results, presented in Table 10, confirm excellent accuracy of the bounds in this case as well.

We, however, note that the parameter α, which determines the amplitude of the potential, has a large influence on the efficiency of the bounds. In table 11, we present the error bounds and the effectivity indices in the setting $\alpha = 0.5$ for two clusters $m = 6$, $M = 9$ and $m = 10$, $M = 13$. The efficiency is here reduced by one order of magnitude, though the assumptions required for the bounds to be valid are still satisfied from $N = 5$ onwards.

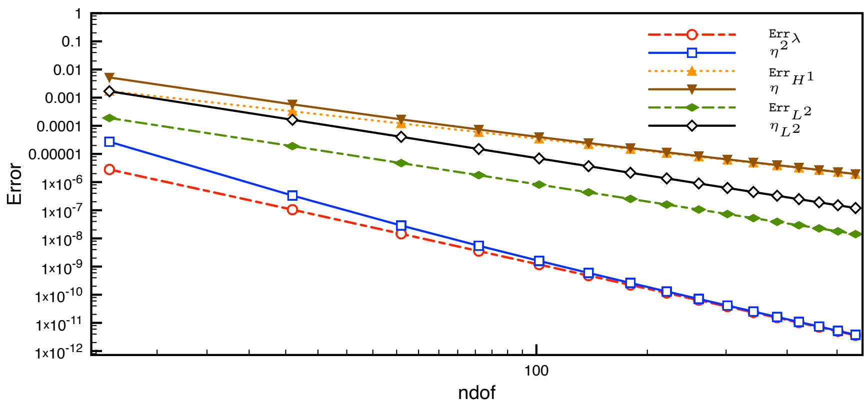


Figure 4: Convergence of the errors and their upper bounds for a 1D Schrödinger operator with periodic boundary conditions with $m = 2$, $M = 3$.

	N	ndof	$\mathtt{Err}_\lambda$	η^2	I_λ^{eff}	$\mathtt{Err}_{H^1}$	η	$I_{H^1}^{\mathrm{eff}}$	$\mathtt{Err}_{L^2}$	η_{L^2}	$I_{L^2}^{\mathrm{eff}}$
$m = 1$	5	121	2.62e-05	2.02e-04	7.70	5.32e-03	1.42e-02	2.67	9.94e-04	6.18e-03	6.22
$M = 5$	15	961	4.12e-07	7.31e-07	1.77	6.45e-04	8.55e-04	1.32	4.47e-05	2.62e-04	5.85
	25	2601	5.32e-08	7.22e-08	1.36	2.31e-04	2.69e-04	1.16	9.99e-06	5.80e-05	5.81
$m = 6$	5	121	5.12e-05	2.80e-04	5.47	7.60e-03	1.67e-02	2.20	1.41e-03	5.90e-03	4.17
$M = 9$	15	961	7.51e-07	1.15e-06	1.53	8.73e-04	1.07e-03	1.23	6.05e-05	2.43e-04	4.02
	25	2601	9.63e-08	1.22e-07	1.26	3.11e-04	3.49e-04	1.12	1.35e-05	5.38e-05	4.00
$m = 10$	5	121	3.81e-05	1.79e-03	46.9	6.83e-03	4.23e-02	6.19	1.28e-03	1.30e-02	10.1
$M = 13$	15	961	4.47e-07	2.93e-06	6.55	6.77e-04	1.71e-03	2.53	4.69e-05	4.87e-04	10.4
	25	2601	5.64e-08	1.80e-07	3.18	2.39e-04	4.24e-04	1.78	1.03e-05	1.07e-04	10.4

Table 10: [Planewaves, two-dimensional case, Schrödinger operator with $\alpha = 0.1$ in (6.2)] Errors, estimates, and effectivity indices for different clusters of eigenvalues. The values of m and M are indicated on the far left.

7 Conclusion

In this paper, we have introduced a new framework for error estimation in eigenvalue problems based on the density matrix formalism. This framework allows to deal with clusters of eigenvalues with possible degeneracies or near-degeneracies, as long as there is a gap between the considered eigenvalues and the rest of the spectrum. We propose *a posteriori* error estimates that are valid for conforming finite element and planewaves discretizations where in the first case, equilibrated flux reconstruction is used to bound the dual residual norms. The numerical results witness a very good quality of the derived methodology in a large set of test scenarios.

	N	ndof	$\mathtt{Err}_\lambda$	η^2	I_λ^{eff}	$\mathtt{Err}_{H^1}$	η	$I_{H^1}^{\mathrm{eff}}$	$\mathtt{Err}_{L^2}$	η_{L^2}	$I_{L^2}^{\mathrm{eff}}$
$m=6$	5	121	2.33e-04	9.87e-02	424	1.65e-02	3.14e-01	19.0	2.90e-03	1.01e-01	35.0
$M=9$	15	961	4.23e-06	2.03e-04	47.9	2.08e-03	1.42e-02	6.84	1.42e-04	4.55e-03	32.0
	25	2601	5.57e-07	1.05e-05	18.8	7.50e-04	3.24e-03	4.32	3.22e-05	1.02e-03	31.6
$m=10$	5	121	1.02e-04	1.44e-01	1410	1.12e-02	3.79e-01	33.9	2.00e-03	1.78e-02	8.89
$M=13$	15	961	1.61e-06	2.58e-04	161	1.29e-03	1.61e-02	12.5	8.80e-05	7.51e-04	8.53
	25	2601	2.10e-07	1.30e-05	61.6	4.61e-04	3.60e-03	7.81	1.98e-05	1.67e-04	8.44

Table 11: [Planewaves, two-dimensional case, Schrödinger operator with $\alpha = 0.5$ in (6.2)] Errors, estimates, and effectivity indices for different clusters of eigenvalues. The values of m and M are indicated on the far left.

Appendix

We present the proof of (3.8) from Lemma 3.3 in this appendix.

A Proof of (3.8) from Lemma 3.3

Proof. To show (3.8), let us first express $\|A^{1/2}(\Phi^0-\Phi_h^0)\|^2$ and $\|A^{1/2}(\gamma^0-\gamma_h)\|^2_{\mathfrak{S}_2(\mathcal{H})}$. From (3.3) and (3.4), there holds

$$\|A^{1/2}(\gamma^0-\gamma_h)\|^2_{\mathfrak{S}_2(\mathcal{H})} = \sum_{i=m}^{M}(\lambda_{ih}-\lambda_i) + 2\sum_{i=m}^{M}\lambda_i\|(1-\gamma_h)\varphi_i^0\|^2 \tag{A.1a}$$

$$\|A^{1/2}(\Phi^0-\Phi_h^0)\|^2 = \sum_{i=m}^{M}(\lambda_{ih}-\lambda_i) + \sum_{i=m}^{M}\lambda_i\|\varphi_i^0-\varphi_{ih}^0\|^2. \tag{A.1b}$$

Since the projector $(1-\gamma_h)$ applied to any approximate eigenvector in the cluster is equal to zero, we obtain

$$\|A^{1/2}(\gamma^0-\gamma_h)\|^2_{\mathfrak{S}_2(\mathcal{H})} = \sum_{i=m}^{M}(\lambda_{ih}-\lambda_i) + 2\sum_{i=m}^{M}\lambda_i\|(1-\gamma_h)(\varphi_i^0-\varphi_{ih}^0)\|^2. \tag{A.2}$$

To show the left inequality in (3.8), we use the fact that the operator norm of the projector $(1-\gamma_h)$ in $\mathcal{L}(\mathcal{H})$ is equal to 1 and (2.18) which, together with (A.2), yield

$$\|A^{1/2}(\gamma^0-\gamma_h)\|^2_{\mathfrak{S}_2(\mathcal{H})} \le 2\sum_{i=m}^{M}(\lambda_{ih}-\lambda_i) + 2\sum_{i=m}^{M}\lambda_i\|\varphi_i^0-\varphi_{ih}^0\|^2 = 2\|A^{1/2}(\Phi^0-\Phi_h^0)\|^2.$$

To show the right inequality in (3.8), we compute the difference $\|A^{1/2}(\Phi^0-\Phi_h^0)\|^2-\|A^{1/2}(\gamma^0-\gamma_h)\|^2_{\mathfrak{S}_2(\mathcal{H})}$. Starting from (A.1b) and (A.2), decomposing the identity as the sum of two orthogonal projectors $1 = \gamma_h + (1-\gamma_h)$, using (2.22) from Lemma 2.5, we obtain

$$\begin{aligned}\|A^{1/2}(\Phi^0-\Phi_h^0)\|^2-\|A^{1/2}(\gamma^0-\gamma_h)\|^2_{\mathfrak{S}_2(\mathcal{H})} &= \sum_{i=m}^{M}\lambda_i\|\varphi_i^0-\varphi_{ih}^0\|^2 \\ &\quad - 2\sum_{i=m}^{M}\lambda_i\|(1-\gamma_h)(\varphi_i^0-\varphi_{ih}^0)\|^2 \\ &\le \sum_{i=m}^{M}\lambda_i\|\varphi_i^0-\varphi_{ih}^0\|^2 \\ &\quad - \sum_{i=m}^{M}\lambda_i\|(1-\gamma_h)(\varphi_i^0-\varphi_{ih}^0)\|^2\end{aligned}$$

$$
\begin{aligned}
&= \sum_{i=m}^{M} \lambda_i \|\gamma_h(\varphi_i^0 - \varphi_{ih}^0)\|^2 \\
&= \sum_{i,j=m}^{M} \lambda_i \left| \left(\varphi_i^0 - \varphi_{ih}^0, \varphi_{jh}^0\right) \right|^2.
\end{aligned}
$$

Further, applying $(\varphi_{jh}^0, \varphi_i^0 - \varphi_{ih}^0) = \frac{1}{2}(\varphi_{jh}^0 - \varphi_j^0, \varphi_i^0 - \varphi_{ih}^0)$ which follows as (4.7) and using the Cauchy–Schwarz inequality together with (2.15), we get

$$
\begin{aligned}
\|A^{1/2}(\Phi^0 - \Phi_h^0)\|^2 - \|A^{1/2}(\gamma^0 - \gamma_h)\|^2_{\mathfrak{S}_2(\mathcal{H})} &\leq \sum_{i,j=m}^{M} \frac{\lambda_i}{4} \|\varphi_i^0 - \varphi_{ih}^0\|^2 \|\varphi_j^0 - \varphi_{jh}^0\|^2 \\
&\leq \frac{\lambda_M}{4} \|\Phi^0 - \Phi_h^0\|^4.
\end{aligned}
$$

Combining with (3.7), we obtain

$$
\|A^{1/2}(\Phi^0 - \Phi_h^0)\|^2 - \|A^{1/2}(\gamma^0 - \gamma_h)\|^2_{\mathfrak{S}_2(\mathcal{H})} \leq \frac{\lambda_M}{4} \|\gamma^0 - \gamma_h\|^4_{\mathfrak{S}_2(\mathcal{H})}.
$$

Also, using (3.6) from the proof of Lemma 3.1 together with $\|\varphi_i^0\| = \|\varphi_{ih}\| = 1$ and using $(\gamma^0)^2 = \gamma^0$, $(\gamma_h)^2 = \gamma_h$, together with (2.18) and (2.11), (2.14) yields

$$
\begin{aligned}
\|A^{1/2}(\gamma^0 - \gamma_h)\|^2_{\mathfrak{S}_2(\mathcal{H})} &= \sum_{i=m}^{M} \lambda_i \left(\varphi_i^0, \varphi_i^0\right) - 2 \sum_{i=m}^{M} \lambda_i \left(\varphi_i^0, \gamma_h \varphi_i^0\right) + \sum_{i=m}^{M} \lambda_{ih} \left(\varphi_{ih}, \varphi_{ih}\right) \\
&\geq \sum_{i=m}^{M} \lambda_i \left[\left(\varphi_i^0, \varphi_i^0\right) - 2 \left(\varphi_i^0, \gamma_h \varphi_i^0\right) + \left(\varphi_{ih}, \varphi_{ih}\right) \right] \\
&\geq \lambda_m \sum_{i=m}^{M} \left[\left(\varphi_i^0, \varphi_i^0\right) - 2 \left(\varphi_i^0, \gamma_h \varphi_i^0\right) + \left(\varphi_{ih}, \varphi_{ih}\right) \right] \\
&= \lambda_m \left[\mathrm{Tr}((\gamma^0)^2) - 2\mathrm{Tr}(\gamma^0 \gamma_h) + \mathrm{Tr}((\gamma_h)^2) \right] \\
&= \lambda_m \left[\|\gamma^0\|^2_{\mathfrak{S}_2(\mathcal{H})} - 2(\gamma^0, \gamma_h)_{\mathfrak{S}_2(\mathcal{H})} + \|\gamma_h\|^2_{\mathfrak{S}_2(\mathcal{H})} \right] \\
&= \lambda_m \|\gamma^0 - \gamma_h\|^2_{\mathfrak{S}_2(\mathcal{H})},
\end{aligned}
$$

from which we deduce that

$$
\|A^{1/2}(\Phi^0 - \Phi_h^0)\|^2 - \|A^{1/2}(\gamma^0 - \gamma_h)\|^2_{\mathfrak{S}_2(\mathcal{H})} \leq \frac{\lambda_M}{4\lambda_m} \|\gamma^0 - \gamma_h\|^2_{\mathfrak{S}_2(\mathcal{H})} \|A^{1/2}(\gamma^0 - \gamma_h)\|^2_{\mathfrak{S}_2(\mathcal{H})},
$$

which gives the right inequality of (3.8). □

References

[1] R. E. Bank, L. Grubišić, and J. S. Ovall, *A framework for robust eigenvalue and eigenvector error estimation and Ritz value convergence enhancement*, Appl. Numer. Math., 66 (2013), pp. 1–29.

[2] N. W. Bazley and D. W. Fox, *Lower bounds for eigenvalues of Schrödinger's equation*, Phys. Rev. (2), 124 (1961), pp. 483–492.

[3] D. Boffi, R. G. Durán, F. Gardini, and L. Gastaldi, *A posteriori error analysis for nonconforming approximation of multiple eigenvalues*, Math. Methods Appl. Sci., 40 (2017), pp. 350–369.

[4] D. Boffi, D. Gallistl, F. Gardini, and L. Gastaldi, *Optimal convergence of adaptive FEM for eigenvalue clusters in mixed form*, Math. Comp., 86 (2017), pp. 2213–2237.

[5] A. Bonito and A. Demlow, *Convergence and optimality of higher-order adaptive finite element methods for eigenvalue clusters*, SIAM J. Numer. Anal., 54 (2016), pp. 2379–2388.

[6] D. Braess, V. Pillwein, and J. Schöberl, *Equilibrated residual error estimates are p-robust*, Comput. Methods Appl. Mech. Engrg., 198 (2009), pp. 1189–1197.

[7] F. Brezzi and M. Fortin, *Mixed and hybrid finite element methods*, vol. 15 of Springer Series in Computational Mathematics, Springer-Verlag, New York, 1991.

[8] E. Cancès, R. Chakir, and Y. Maday, *Numerical analysis of the planewave discretization of some orbital-free and Kohn-Sham models*, Esaim Math. Model. Numer. Anal., 46 (2012), pp. 341–388.

[9] E. Cancès, G. Dusson, Y. Maday, B. Stamm, and M. Vohralík, *Guaranteed and robust a posteriori bounds for Laplace eigenvalues and eigenvectors: conforming approximations*, SIAM J. Numer. Anal., 55 (2017), pp. 2228–2254.

[10] E. Cancès, G. Dusson, Y. Maday, B. Stamm, and M. Vohralík, *Guaranteed and robust a posteriori bounds for Laplace eigenvalues and eigenvectors: a unified framework*, Numer. Math., 140 (2018), pp. 1033–1079.

[11] E. Cancès, G. Dusson, Y. Maday, B. Stamm, and M. Vohralík, *Post-processing of the planewave approximation of Schrödinger equations. Part I: linear operators*, IMA Journal of Numerical Analysis, (2020).

[12] C. Carstensen and S. A. Funken, *Fully reliable localized error control in the FEM*, SIAM J. Sci. Comput., 21 (1999/00), pp. 1465–1484.

[13] C. Carstensen and J. Gedicke, *Guaranteed lower bounds for eigenvalues*, Math. Comp., 83 (2014), pp. 2605–2629.

[14] C. Carstensen, J. Gedicke, and D. Rim, *Explicit error estimates for Courant, Crouzeix-Raviart and Raviart-Thomas finite element methods*, J. Comput. Math., 30 (2012), pp. 337–353.

[15] P. Ciarlet, Jr. and M. Vohralík, *Localization of global norms and robust a posteriori error control for transmission problems with sign-changing coefficients*, M2AN Math. Model. Numer. Anal., 52 (2018), pp. 2037–2064.

[16] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, vol. 4 of Studies in Mathematics and its Applications, North-Holland, Amsterdam, 1978.

[17] X. Dai, L. He, and A. Zhou, *Convergence and quasi-optimal complexity of adaptive finite element computations for multiple eigenvalues*, IMA J. Numer. Anal., 35 (2015), pp. 1934–1977.

[18] P. Destuynder and B. Métivet, *Explicit error bounds in a conforming finite element method*, Math. Comp., 68 (1999), pp. 1379–1396.

[19] W. Dörfler, *A convergent adaptive algorithm for Poisson's equation*, SIAM J. Numer. Anal., 33 (1996), pp. 1106–1124.

[20] G. Dusson, *Post-processing of the planewave approximation of Schrödinger equations. Part II: Kohn–sham models*, IMA Journal of Numerical Analysis, (2020).

[21] A. Ern and M. Vohralík, *Polynomial-degree-robust a posteriori estimates in a unified setting for conforming, nonconforming, discontinuous Galerkin, and mixed discretizations*, SIAM J. Numer. Anal., 53 (2015), pp. 1058–1081.

[22] ——, *Stable broken H^1 and $\boldsymbol{H}(\mathrm{div})$ polynomial extensions for polynomial-degree-robust potential and flux reconstruction in three space dimensions*. HAL Preprint 01422204, submitted for publication, 2016.

[23] G. E. Forsythe, *Asymptotic lower bounds for the fundamental frequency of convex membranes*, Pacific J. Math., 5 (1955), pp. 691–702.

[24] D. Gallistl, *Adaptive nonconforming finite element approximation of eigenvalue clusters*, Comput. Methods Appl. Math., 14 (2014).

[25] ——, *An optimal adaptive FEM for eigenvalue clusters*, Numer. Math., 130 (2015), pp. 467–496.

[26] S. Giani and E. J. C. Hall, *An a posteriori error estimator for hp-adaptive discontinuous Galerkin methods for elliptic eigenvalue problems*, Math. Models Methods Appl. Sci., 22 (2012), pp. 1250030, 35.

[27] P. Grisvard, *Elliptic problems in nonsmooth domains*, vol. 24 of Monographs and Studies in Mathematics, Pitman (Advanced Publishing Program), Boston, MA, 1985.

[28] L. Grubišić and J. S. Ovall, *On estimators for eigenvalue/eigenvector approximations*, Math. Comp., 78 (2009), pp. 739–770.

[29] F. Hecht, O. Pironneau, J. Morice, A. Le Hyaric, and K. Ohtsuka, *FreeFem++*, tech. rep., Laboratoire Jacques-Louis Lions, Université Pierre et Marie Curie, Paris, http://www.freefem.org/ff++, 2012.

[30] J. Hu, Y. Huang, and Q. Lin, *Lower bounds for eigenvalues of elliptic operators: by nonconforming finite element methods*, J. Sci. Comput., 61 (2014), pp. 196–221.

[31] J. Hu, Y. Huang, and Q. Shen, *The lower/upper bound property of approximate eigenvalues by nonconforming finite element methods for elliptic operators*, J. Sci. Comput., 58 (2014), pp. 574–591.

[32] T. Kato, *On the upper and lower bounds of eigenvalues*, J. Phys. Soc. Japan, 4 (1949), pp. 334–339.

[33] P. Ladevèze and D. Leguillon, *Error estimate procedure in the finite element method and applications*, SIAM J. Numer. Anal., 20 (1983), pp. 485–509.

[34] X. Liu, *A framework of verified eigenvalue bounds for self-adjoint differential operators*, Appl. Math. Comput., 267 (2015), pp. 341–355.

[35] X. Liu and F. Kikuchi, *Analysis and estimation of error constants for P_0 and P_1 interpolations over triangular finite elements*, J. Math. Sci. Univ. Tokyo, 17 (2010), pp. 27–78.

[36] X. Liu and S. Oishi, *Verified eigenvalue evaluation for the Laplacian over polygonal domains of arbitrary shape*, SIAM J. Numer. Anal., 51 (2013), pp. 1634–1654.

[37] X. Liu and T. Vejchodský, *Rigorous and fully computable a posteriori error bounds for eigenfunctions.* Preprint arXiv:1904.07903, submitted for publication, 2019.

[38] F. Luo, Q. Lin, and H. Xie, *Computing the lower and upper bounds of Laplace eigenvalue problem: by combining conforming and nonconforming finite element methods*, Sci. China Math., 55 (2012), pp. 1069–1082.

[39] W. Prager and J. L. Synge, *Approximations in elasticity based on the concept of function space*, Quart. Appl. Math., 5 (1947), pp. 241–269.

[40] M. Reed and B. Simon, *Methods of modern mathematical physics IV: Analysis of operators*, Academic Press, New York, 1978.

[41] J. E. Roberts and J.-M. Thomas, *Mixed and hybrid methods*, in Handbook of Numerical Analysis, Vol. II, North-Holland, Amsterdam, 1991, pp. 523–639.

[42] L. N. Trefethen and T. Betcke, *Computed eigenmodes of planar regions*, in Recent advances in differential equations and mathematical physics, vol. 412 of Contemp. Math., Amer. Math. Soc., Providence, RI, 2006, pp. 297–314.

[43] L. Wang, L. Chamoin, P. Ladevèze, and H. Zhong, *Computable upper and lower bounds on eigenfrequencies*, Comput. Methods Appl. Mech. Engrg., 302 (2016), pp. 27–43.

[44] H. F. Weinberger, *Upper and lower bounds for eigenvalues by finite difference methods*, Comm. Pure Appl. Math., 9 (1956), pp. 613–623.