EIGENVALUE ENCLOSURES AND APPLICATIONS TO THE MAXWELL OPERATOR

GABRIEL R. BARRENECHEA, LYONELL BOULTON,
AND NABILE BOUSSAÏD

ABSTRACT. This paper is concerned with methods for computing certified bounds for the isolated eigenvalues of self-adjoint operators. We examine in close detail the connections between an extension of the Temple-Lehmann-Goerisch method developed a few years ago by Zimmermann and Mertins, and a general framework considered by Davies and Plum. We propose employing the former as a highly effective tool for the pollution-free numerical estimation of the eigenfrequencies and field phasors of the resonant cavity problem on a bounded region filled with a generally anisotropic medium, by means of finite elements.

CONTENTS

Date: September 24, 2013.

Key words and phrases. eigenvalue enclosures, Maxwell equation, spectral pollution, finite element method.

1. Introduction

The goal of this paper is two-folded. On the one hand, we examine in close detail the equivalence between an extension of the Temple-Lehmann-Goerisch method [21] developed a few years ago by Zimmermann and Mertins [25], and a pollution-free eigenvalue bound calculation method considered by Davies and Plum [18]. On the other hand, we show that the former can be effectively applied for finite element computations in the context of the resonant cavity problem.

Let $\Omega \subset \mathbb{R}^3$ be a domain which is bounded, open and simply connected. Let $\partial\Omega$, the boundary of Ω, be sufficiently regular (see Section 5) and denote by $\mathbf{n}$ its outer normal vector. The numerical estimation of the angular frequencies $\omega \in \mathbb{R}$ and electromagnetic field phasors $(\boldsymbol{E}, \boldsymbol{H}) \neq 0$ of the Maxwell eigenvalue problem,

$$\begin{cases} \operatorname{curl} \boldsymbol{E} = i\omega\mu\boldsymbol{H} & \text{in } \Omega \\ \operatorname{curl} \boldsymbol{H} = -i\omega\epsilon\boldsymbol{E} & \text{in } \Omega \\ \boldsymbol{E} \times \mathbf{n} = 0 & \text{on } \partial\Omega, \end{cases} \tag{1}$$

is known to be extremely challenging for general data. Here and elsewhere the electric permittivity and the magnetic permeability, ϵ and μ respectively, are positive and such that

$$\epsilon,\ \frac{1}{\epsilon},\ \mu,\ \frac{1}{\mu} \in L^\infty(\Omega). \tag{2}$$

The physical phenomenon of electromagnetic oscillations in a resonator is described via equation (1), restricted to the solenoidal subspace which is characterized by the Gauss law:

$$\operatorname{div}(\epsilon\boldsymbol{E}) = 0 = \operatorname{div}(\mu\boldsymbol{H}). \tag{3}$$

The orthogonal complement of this subspace in a suitable inner product (see [7] and references therein) is the gradient space, which has infinite dimension and it lies in the kernel of the self-adjoint operator $\mathcal{M}$ associated to (1). In turns, this means that (1)-(3) and the unrestricted problem (1), have the same non-zero spectrum and the same corresponding eigenspace. In this paper we propose computing the non-zero angular frequencies and field phasors of the resonator, by means of the latter.

The operator $\mathcal{M}$ does not have a compact resolvent and it is strongly indefinite. The self-adjoint operator associated to (1)-(3) has a compact resolvent but it is still strongly indefinite. By considering the square of

$\mathcal{M}$ on the solenoidal subspace, one obtains a positive definite eigenvalue problem (involving the bi-curl, for example, if the medium is isotropic) which can in principle be discretized via the Galerkin method. One serious drawback of this idea for practical computations is the fact that the standard finite element spaces are not solenoidal. Usually, spurious modes associated to the infinite-dimensional kernel appear and give rise to spectral pollution. This has been well documented and it is known to be a manifested problem whenever the underlying mesh is unstructured, [2] and references therein.

In order to overcome the difficulties involved in the finite element treatment of (1), various ingenious methods have been considered in the past. One possible approach [10], is to enhance the divergence of the electric field in a fractional order negative Sobolev norm. Another possibility [12], is to combine nodal elements with a least squares formulation of (1)-(3) re-written in weak form. Moreover, the latter ansatz can also be incorporated into (1) by means of a Lagrange multiplier and then one can use continuous finite element spaces of a Taylor-Hood type, [13]. Perhaps the most effective among these methods, [9, 8], consists in re-writing the spectral problem associated to $\mathcal{M}^2$ in a mixed form and then employ edge finite elements. This approach turns out to be linked to deep results on the rigourous treatment of finite elements and it is at the core of elegant geometrical ideas, [2].

Unfortunately, and to the best of our knowledge, all these computational techniques exhibit two main limitations.

a) They are not certified. To be precise, computed eigenvalues are not necessarily guaranteed one-sided bounds of the exact eigenvalues in general, despite of the possible convergence of the method.

b) Detecting the multiplicity of an eigenvalue or the presence of a spectral cluster is extremely difficult.

Below we propose an alternative approach for computing the eigenvalues of (1) which addresses these limitations. The strategy is completely general in character and it is based on the method developed by Zimmermann and Mertins in [25]. The procedure is robust in the sense that any standard class of finite elements, including the ones based on nodal degrees of freedom, can be implemented to perform computations which are certified up to machine precision. In recent years, this method has been successfully used in the context of the radially reduced magnetohydrodynamics operator [25, 11], the calculation of complementary bounds for the Helmholtz equation [5] and the calculation of sloshing frequencies in the left definite case [4].

The method of Zimmermann and Mertins is closely linked to another pollution-free technique for eigenvalue computation which is based on a notion of approximated spectral distance. This other method was formulated by Davies in [16, 17] and was later developed by Davies and Plum in [18], but it is yet to be tested properly on models of dimension other than one. Below we develop further the arguments presented in [18, Section 6], in order to determine the precise nature of the equivalence between these two techniques.

In Section 2 we extend various canonical results from [18]. Notably, we include multiplicity counting (Proposition 1 and also the Appendix A) and a description of how eigenfunctions are approximated (Proposition 3). The method of Zimmermann and Mertins, on the other hand, is introduced in Section 3. We derive the latter in a self-contained manner independently from the work [25]. See Theorem 6 and Corollary 7.

Section 4 addresses the questions of convergence and upper bounds for residuals in both methods. The main statements in this respect are Theorem 10, Corollary 11 and Theorem 12, where we determine general convergence estimates with explicit residuals for a finite group of contiguous eigenvalues. The results of Section 3 and Theorem 12 are then employed in Section 5, in order to establish concrete approximation rates for the pollution-free numerical solution of (1) by means of the Zimmermann-Mertins method on nodal finite elements. Theorem 14 collects the main contribution in this respect. We show that the rate of convergence found is optimal for this class of trial spaces. Note that we have chosen the most widely available class of finite elements here, with the purpose of illustrating our findings in a concrete accessible manner. However the technique described in sections 2 and 3 is completely general in character. It can, for instance, be also implemented on other classes of basis functions for problem (1).

The final part of the paper is devoted to concrete computational applications. A certified numerical strategy is specified in the Procedure 1 of Section 6. According to Lemma 16, this strategy is convergent in a suitable regime for the finite element approximation of the solutions of (1). Section 7, on the other hand, contains various benchmark numerical experiments. A companion Comsol Multiphysics v4.3a Livelink code, which was employed to produce some of the results presented in Section 7, is included in the Appendix B.

2. Approximated local counting functions

This section is devoted to notions of approximated spectral distance and approximated local counting function for self-adjoint operators. We follow closely the framework established in [16, 17, 18]. These notions and their properties will lead in the next section to the formulation of a method for eigenvalue computation which has been examined in [25] and subsequent works [5, 4]. Various results in all these references can be recovered from the unified approach presented below.

Let $A : \mathrm{D}(A) \longrightarrow \mathcal{H}$ be a self-adjoint operator acting on a Hilbert space $\mathcal{H}$. Decompose the spectrum of A in the usual fashion, as the union of discrete and essential spectrum, $\sigma(A) = \sigma_{\mathrm{disc}}(A) \cup \sigma_{\mathrm{ess}}(A)$. Let J be any Borel subset of $\mathbb{R}$. The spectral projector associated to A is denoted by $\mathbb{1}_J(A) = \int_J \mathrm{d}E_\lambda$. Hence $\mathrm{Tr}\,\mathbb{1}_J(A) = \dim \mathbb{1}_J(A)\mathcal{H}$. We write $\mathcal{E}_J(A) = \oplus_{\lambda\in J} \ker(A-\lambda)$ with the convention $\mathcal{E}_\lambda(A) = \mathcal{E}_{\{\lambda\}}(A)$. Generally $\mathcal{E}_J(A) \subseteq \mathbb{1}_J(A)\mathcal{H}$, however there is no reason for these two subspaces to be equal.

Let $t \in \mathbb{R}$. Let $q_t : \mathrm{D}(A) \times \mathrm{D}(A) \longrightarrow \mathbb{C}$ be the closed bilinear form

$$q_t(u, w) = \langle (A-t)u, (A-t)w\rangle \qquad \forall u, w \in \mathrm{D}(A). \tag{4}$$

For any $u \in \mathrm{D}(A)$ we will constantly make use of the following t-dependant semi-norm, which is a norm if t is not an eigenvalue,

$$|u|_t = q_t(u, u)^{1/2} = \|(A-t)u\|. \tag{5}$$

By virtue of the min-max principle, q_t characterizes the spectrum which lies near the origin of the positive operator $(A-t)^2$. In turn, this gives rise to a notion of local counting function at t for the spectrum of A.

Let

$$\mathfrak{d}_j(t) = \inf_{\substack{\dim V = j \\ V \subset \mathrm{D}(A)}} \sup_{u\in V} \frac{|u|_t}{\|u\|}$$

so that $0 \le \mathfrak{d}_j(t) \le \mathfrak{d}_k(t)$ for $j < k$. Then $\mathfrak{d}_1(t)$ is the Hausdorff distance from t to $\sigma(A)$,

$$\mathfrak{d}_1(t) = \min\{\lambda \in \sigma(A) : |\lambda - t|\} = \inf_{u\in \mathrm{D}(A)} \frac{|u|_t}{\|u\|}. \tag{6}$$

Similarly $\mathfrak{d}_j(t)$ are the distances from t to the jth nearest point in $\sigma(A)$ counting multiplicity in a generalized sense. That is, stopping when the essential spectrum is reached. Moreover

$$\mathfrak{d}_j(t) = \mathfrak{d}_{j-1}(t) \iff \begin{cases} \text{either} & \dim \mathcal{E}_{[t-\mathfrak{d}_{j-1}(t), t+\mathfrak{d}_{j-1}(t)]}(A) > j-1 \\ \quad\text{or} & t + \mathfrak{d}_{j-1}(t) \in \sigma_{\mathrm{ess}}(A) \\ \quad\text{or} & t - \mathfrak{d}_{j-1}(t) \in \sigma_{\mathrm{ess}}(A). \end{cases}$$

Without further mention, below we will always count spectral points of A relative to t, regarding multiplicities in this generalized sense.

We now show how to extract certified information about $\sigma(A)$ in the vicinity of t from the action of A onto finite-dimensional trial subspaces $\mathcal{L} \subset \mathrm{D}(A)$, see [16, Section 3]. For $j \leq n = \dim \mathcal{L}$, let

$$F^j_{\mathcal{L}}(t) = \min_{\substack{\dim V = j \\ V \subset \mathcal{L}}} \max_{u \in V} \frac{|u|_t}{\|u\|}. \tag{7}$$

Then $0 \leq F^1_{\mathcal{L}}(t) \leq \ldots \leq F^n_{\mathcal{L}}(t)$ and $F^j_{\mathcal{L}}(t) \geq \mathfrak{d}_j(t)$ for all $j = 1, 2, \ldots, n$. Since $[t - \mathfrak{d}_j(t), t + \mathfrak{d}_j(t)] \subseteq [t - F^j_{\mathcal{L}}(t), t + F^j_{\mathcal{L}}(t)]$, there are at least j spectral points of A in the segment $\big[t - F^j_{\mathcal{L}}(t), t + F^j_{\mathcal{L}}(t)\big]$ including, possibly, the essential spectrum. That is

$$\operatorname{Tr} \mathbb{1}_{[t - F^j_{\mathcal{L}}(t), t + F^j_{\mathcal{L}}(t)]}(A) \geq j \qquad \forall j = 1, \ldots, n. \tag{8}$$

Hence $F^j_{\mathcal{L}}(t)$ is an approximated local counting function for $\sigma(A)$.

As a consequence of the triangle inequality, $F^j_{\mathcal{L}}$ is a Lipschitz continuous function such that

$$|F^j_{\mathcal{L}}(t) - F^j_{\mathcal{L}}(s)| \leq |t - s| \qquad \forall s, t \in \mathbb{R} \quad \text{and} \quad j = 1, \ldots, n. \tag{9}$$

Moreover, $F^j_{\mathcal{L}}(t)$ is the jth smallest eigenvalue μ of the non-negative weak problem:

$$\text{find } (\mu, u) \in [0, \infty) \times \mathcal{L} \setminus \{0\} \quad \text{such that} \quad q_t(u, v) = \mu^2 \langle u, v \rangle \qquad \forall v \in \mathcal{L}. \tag{10}$$

Hence

$$F^j_{\mathcal{L}}(t) = \max_{\substack{\dim V = j-1 \\ V \subset \mathcal{L}}} \min_{u \in \mathcal{L} \ominus V} \frac{|u|_t}{\|u\|} = \max_{\substack{\dim V = j-1 \\ V \subset \mathcal{H}}} \min_{u \in \mathcal{L} \ominus V} \frac{|u|_t}{\|u\|}. \tag{11}$$

We now show how to detect the spectrum of A to the left/right of t by means of $F^j_{\mathcal{L}}$ in an optimal setting. This turns out to be a crucial ingredient in the formulation of the strategy proposed in [16, 17, 18]. The following notation simplifies various statements below. Let

$$\mathfrak{n}^-_j(t) = \sup\{s < t : \operatorname{Tr} \mathbb{1}_{(s,t]}(A) \geq j\} \quad \text{and}$$
$$\mathfrak{n}^+_j(t) = \inf\{s > t : \operatorname{Tr} \mathbb{1}_{[t,s)}(A) \geq j\}.$$

Then $\mathfrak{n}^\mp_j(t)$ is the jth point in $\sigma(A)$ to the left$(-)$/right$(+)$ of t counting multiplicities. Here $t \in \sigma(A)$ is allowed and neither t nor $\mathfrak{n}^\mp_1(t)$ have to be isolated from the rest of $\sigma(A)$. Note that $\mathfrak{n}^-_j(t) = -\infty$ for $\operatorname{Tr} \mathbb{1}_{(-\infty,t]}(A) < j$ and $\mathfrak{n}^+_j(t) = +\infty$ for $\operatorname{Tr} \mathbb{1}_{[t,+\infty)}(A) < j$. Without further mention, all statements below regarding bounds on $\mathfrak{n}^\mp_j(t)$ will be void (hence redundant) in either of these two cases.

Proposition 1. *Let $t^- < t < t^+$. Then*

$$F_{\mathcal{L}}^j(t^-) \le t - t^- \quad \Rightarrow \quad t^- - F_{\mathcal{L}}^j(t^-) \le \mathfrak{n}_j^-(t)$$
$$F_{\mathcal{L}}^j(t^+) \le t^+ - t \quad \Rightarrow \quad t^+ + F_{\mathcal{L}}^j(t^+) \ge \mathfrak{n}_j^+(t). \tag{12}$$

Moreover, let $t_1^- < t_2^- < t < t_2^+ < t_1^+$. Then

(13)

$$F_{\mathcal{L}}^j(t_i^-) \le t - t_i^- \text{ for } i = 1,2 \quad \Rightarrow \quad t_1^- - F_{\mathcal{L}}^j(t_1^-) \le t_2^- - F_{\mathcal{L}}^j(t_2^-) \le \mathfrak{n}_j^-(t)$$
$$F_{\mathcal{L}}^j(t_i^+) \le t_i^+ - t \text{ for } i = 1,2 \quad \Rightarrow \quad t_1^+ + F_{\mathcal{L}}^j(t_1^+) \ge t_2^+ + F_{\mathcal{L}}^j(t_2^+) \ge \mathfrak{n}_j^+(t).$$

Proof. We firstly show (12). Suppose that $t \ge F_{\mathcal{L}}^j(t^-) + t^-$. Then

$$\operatorname{Tr} \mathbb{1}_{[t^- - F_{\mathcal{L}}^j(t^-), t]}(A) \ge j.$$

Since $\mathfrak{n}_j^-(t) \le \ldots \le \mathfrak{n}_1^-(t)$ are the only spectral points in the segment $[\mathfrak{n}_j^-(t), t]$, then necessarily

$$\mathfrak{n}_j^-(t) \in [t^- - F_{\mathcal{L}}^j(t^-), t].$$

The bottom of (12) is shown in a similar fashion.

The second statement follows by observing that the maps $t \mapsto t \pm F_{\mathcal{L}}^j(t)$ are monotonically increasing as a consequence of (9). ☐

The structure of the trial subspace $\mathcal{L}$ determines the existence of $t^\pm$ satisfying the hypothesis in (12). If we expect to detect $\sigma(A)$ at both sides of t, a necessary requirement on $\mathcal{L}$ should certainly be the condition

$$\min_{u\in\mathcal{L}} \frac{\langle Au, u\rangle}{\langle u, u\rangle} < t < \max_{u\in\mathcal{L}} \frac{\langle Au, u\rangle}{\langle u, u\rangle}. \tag{14}$$

By virtue of lemmas 4 and 5 below, for $j = 1$, the left hand side inequality of (14) implies the existence of t^- and the right hand side inequality implies the existence of t^+, respectively.

Remark 1. From Proposition 1 it follows that optimal lower bounds for $\mathfrak{n}_j^-(t)$ are achieved by finding $\hat{t}_j^- \le t$, the closer point to t, such that $F_{\mathcal{L}}^j(\hat{t}_j^-) = t - \hat{t}_j^-$. Indeed, (13) gives, $t^- - F_{\mathcal{L}}^j(t^-) \le \hat{t}_j^- - F_{\mathcal{L}}^j(\hat{t}_j^-) \le \mathfrak{n}_j^-(t)$ for any other t^- as in (12). Similarly, optimal upper bounds for $\mathfrak{n}_j^+(t)$ are found by analogous means. This observation will play a crucial role in Section 3.

We now determine further geometrical properties of $F_{\mathcal{L}}^1$ and its connection to the spectral distance. Let the Hausdorff distances from $t \in \mathbb{R}$ to $\sigma(A) \setminus (-\infty, t]$ and $\sigma(A) \setminus [t, \infty)$, respectively, be given by

$$\delta^+(t) = \inf\{\mu - t : \mu \in \sigma(A),\ \mu > t\} \qquad \text{and}$$
$$\delta^-(t) = \inf\{t - \mu : \mu \in \sigma(A),\ \mu < t\}. \tag{15}$$

In general, $t-\mathfrak{n}_1^-(t)\leq\delta^-(t)$ and $\mathfrak{n}_1^+(t)-t\leq\delta^+(t)$. In fact, we know that $|\mathfrak{n}_1^\pm(t)-t|=\delta^\pm(t)$ for $t\notin\sigma(A)$. However, these relations can be strict whenever $t\in\sigma(A)$. Indeed, $\mathfrak{n}_1^+(t)-t=\delta^+(t)$ iff there exists a decreasing sequence $t_n^+\in\sigma(A)$ such that $t_n^+\downarrow t$, whereas $t-\mathfrak{n}_1^-(t)=\delta^-(t)$ iff there exists an increasing sequence $t_n^-\in\sigma(A)$ such that $t_n^-\uparrow t$.

An emphasis in distinguishing $|\mathfrak{n}_1^\pm(t)-t|$ from $\delta^\pm(t)$ seems unnecessary at this stage. However, this distinction in the notation will be justified later on. Without further mention below we write $\delta^\pm(t)=\pm\infty$ to indicate that either of the sets on the right side of (15) is empty.

Let $\lambda\in\sigma(A)$ be an isolated point. If there exists a non-vanishing $u\in\mathcal{L}\cap\mathcal{E}_\lambda(A)$, then

$$\frac{|u|_s}{\|u\|}=|\lambda-s|=\mathfrak{d}_1(s)\qquad\forall s\in\left[\lambda-\frac{\delta^-(\lambda)}{2},\lambda+\frac{\delta^+(\lambda)}{2}\right].$$

According to the convergence analysis carried out in Section 4, the smaller the angle between $\mathcal{L}$ and the spectral subspace $\mathcal{E}_\lambda(A)$, the closer the $F_\mathcal{L}^1(t)$ is to $\mathfrak{d}_1(t)$ for $t\in\left(\lambda-\frac{\delta^-(\lambda)}{2},\lambda+\frac{\delta^+(\lambda)}{2}\right)$. The special case of this angle being zero is described by the following lemma.

Lemma 2. *For $\lambda\in\sigma(A)$ isolated from the rest of the spectrum, the following statements are equivalent.*

- *a) There exists a minimizer $u\in\mathcal{L}$ of the right side of (7) for $j=1$, such that $|u|_t=\mathfrak{d}_1(t)$ for a single $t\in\left(\lambda-\frac{\delta^-(\lambda)}{2},\lambda+\frac{\delta^+(\lambda)}{2}\right)$,*
- *b) $F_\mathcal{L}^1(t)=\mathfrak{d}_1(t)$ for a single $t\in\left(\lambda-\frac{\delta^-(\lambda)}{2},\lambda+\frac{\delta^+(\lambda)}{2}\right)$,*
- *c) $F_\mathcal{L}^1(s)=\mathfrak{d}_1(s)$ for all $s\in[\lambda-\frac{\delta^-(\lambda)}{2},\lambda+\frac{\delta^+(\lambda)}{2}]$,*
- *d) $\mathcal{L}\cap\mathcal{E}_\lambda(A)\neq\{0\}$.*

Proof. Since $\mathcal{L}$ is finite-dimensional, *a)* and *b)* are equivalent by the definitions of $\mathfrak{d}_1(t)$, $F_\mathcal{L}^1(t)$ and q_t. From the paragraph above the statement of the lemma it is clear that *d) ⇒ c) ⇒ b)*. Since $|u|_t/\|u\|$ is the square root of the Rayleigh quotient associated to the operator $(A-t)^2$, the fact that λ is isolated combined with the Rayleigh-Ritz principle, gives the implication *a)⇒d)*. □

As there can be a mixing of eigenspaces, it is not possible to replace *b)* in this lemma by an analogous statement including $t=\lambda\pm\frac{\delta^\pm(\lambda)}{2}$. If $\lambda'=\lambda+\delta^+(\lambda)$ is an eigenvalue, for example, then $F_\mathcal{L}^1\left(\frac{\lambda+\lambda'}{2}\right)=\mathfrak{d}_1\left(\frac{\lambda+\lambda'}{2}\right)$ ensures that $\mathcal{L}$ contains elements of $\mathcal{E}_\lambda(A)\oplus\mathcal{E}_{\lambda'}(A)$. However it is not guaranteed to be orthogonal to either of these two subspaces. See the Appendix A for similar results in the case $j>1$.

We conclude this section by examining extensions of the implications *b)* $\Rightarrow$ *d)* of Lemma 2 into a more general context. In combination with the results of Section 3, the next proposition shows how to obtain certified information about spectral subspaces. Some of its practical implications will be discussed later on.

Here and below $\{u_j^t\}_{j=1}^n \subset \mathcal{L}$ will denote an orthonormal family of eigenfunctions associated to the eigenvalues $\mu = F_{\mathcal{L}}^j(t)$ of the weak problem (10). In a suitable asymptotic regime for $\mathcal{L}$, the angle between these eigenfunctions and the spectral subspaces of $|A-t|$ in the vicinity of the origin is controlled by a residual which is as small as $\mathcal{O}\left(\sqrt{F_{\mathcal{L}}^j(t)-\mathfrak{d}_j(t)}\right)$ for $F_{\mathcal{L}}^j(t)-\mathfrak{d}_j(t)\to 0$.

Assumption 1. *Unless otherwise specified, from now on we will always fix the parameter $m \le n = \dim \mathcal{L}$ and suppose that*

$$[t-\mathfrak{d}_m(t), t+\mathfrak{d}_m(t)] \cap \sigma(A) \subseteq \sigma_{\mathrm{disc}}(A). \tag{16}$$

Set

$$\delta_j(t) = \operatorname{dist}\left[t, \sigma(A) \setminus \{t \pm \mathfrak{d}_k(t)\}_{k=1}^j\right].$$

By virtue of (16), $\delta_j(t) > \mathfrak{d}_j(t)$ for all $j \le m$.

Remark 2. If $t = \frac{\mathfrak{n}_j^-(t)+\mathfrak{n}_j^+(t)}{2}$ for a given j, the vectors ϕ_j^t introduced in Proposition 3 and invoked subsequently, might not be eigenvectors of A despite of the fact that $|A-t|\phi_j^t = \mathfrak{d}_j(t)\phi_j^t$. However, in any other circumstance ϕ_j^t are eigenvectors of A.

Proposition 3. *Let $t \in \mathbb{R}$ and $j \in \{1,\dots,m\}$. Assume that the difference $F_{\mathcal{L}}^j(t) - \mathfrak{d}_j(t)$ is small enough so that $0 < \varepsilon_j < 1$ holds true for the residuals constructed inductively as follows,*

$$\varepsilon_1 = \sqrt{\frac{F_{\mathcal{L}}^1(t)^2 - \mathfrak{d}_1(t)^2}{\delta_1(t)^2 - \mathfrak{d}_1(t)^2}}$$

$$\varepsilon_j = \sqrt{\frac{F_{\mathcal{L}}^j(t)^2 - \mathfrak{d}_j(t)^2}{\delta_j(t)^2 - \mathfrak{d}_j(t)^2} + \sum_{k=1}^{j-1} \frac{\varepsilon_k^2}{1-\varepsilon_k^2}\left(1 + \frac{\mathfrak{d}_j(t)^2 - \mathfrak{d}_k(t)^2}{\delta_j(t)^2 - \mathfrak{d}_j(t)^2}\right)}.$$

Then, there exists an orthonormal basis $\{\phi_j^t\}_{j=1}^m$ of $\mathcal{E}_{[t-\mathfrak{d}_m(t), t+\mathfrak{d}_m(t)]}(A)$ such that $\phi_j^t \in \mathcal{E}_{\{t-\mathfrak{d}_j(t), t+\mathfrak{d}_j(t)\}}(A)$,

$$\|u_j^t - \langle u_j^t, \phi_j^t\rangle \phi_j^t\| \le \varepsilon_j \qquad \textit{and} \tag{17}$$

$$|u_j^t - \langle u_j^t, \phi_j^t\rangle \phi_j^t|_t \le \sqrt{F_{\mathcal{L}}^j(t)^2 - \mathfrak{d}_j(t)^2 + \mathfrak{d}_j(t)^2\varepsilon_j^2}. \tag{18}$$

Proof. As it is clear from the context, in this proof we suppress the index t on top of any vector. We write $\Pi_{\mathcal{S}}$ to denote the orthogonal projection onto the subspace $\mathcal{S}$ with respect to the inner product $\langle \cdot, \cdot \rangle$.

Let us first consider the case $j = 1$. Let $\mathcal{S}_1 = \mathcal{E}_{\{t-\mathfrak{d}_1(t), t+\mathfrak{d}_1(t)\}}(A)$, and decompose $u_1 = \Pi_{\mathcal{S}_1} u_1 + u_1^\perp$ where $u_1^\perp \perp \mathcal{S}_1$. Since A is self-adjoint,

$$F_{\mathcal{L}}^1(t)^2 = \|(A-t)u_1\|^2 = \mathfrak{d}_1(t)^2 \|\Pi_{\mathcal{S}_1} u_1\|^2 + \|(A-t)u_1^\perp\|^2. \tag{19}$$

Hence

$$F_{\mathcal{L}}^1(t)^2 \geq \mathfrak{d}_1(t)^2(1 - \|u_1^\perp\|^2) + \delta_1(t)^2 \|u_1^\perp\|^2.$$

Since $\delta_1(t) > \mathfrak{d}_1(t)$, clearing from this identity $\|u_1^\perp\|^2$ yields $\|u_1^\perp\| \leq \varepsilon_1$. Hence $\|\Pi_{\mathcal{S}_1} u_1\|^2 \geq 1 - \varepsilon_1^2 > 0$. Let

$$\phi_1 = \frac{1}{\|\Pi_{\mathcal{S}_1} u_1\|} \Pi_{\mathcal{S}_1} u_1$$

so that $\|\Pi_{\mathcal{S}_1} u_1\| = |\langle u_1, \phi_1 \rangle|$. Then (17) holds immediately and (18) is achieved by clearing $\|(A-t)u_1^\perp\|^2$ from (19).

We define the needed basis, and show (17) and (18), for j up to m inductively as follows. Set

$$\phi_j = \frac{1}{\|\Pi_{\mathcal{S}_j} u_j\|} \Pi_{\mathcal{S}_j} u_j$$

where $\mathcal{S}_j = \mathcal{E}_{\{t-\mathfrak{d}_j(t), t+\mathfrak{d}_j(t)\}}(A) \ominus \operatorname{Span}\{\phi_l\}_1^{j-1}$ and $\Pi_{\mathcal{S}_j} u_j \neq 0$, all this for $1 \leq j \leq k-1$. Assume that (17) and (18) hold true for j up to $k-1$. Define $\mathcal{S}_k = \mathcal{E}_{\{t-\mathfrak{d}_k(t), t+\mathfrak{d}_k(t)\}}(A) \ominus \operatorname{Span}\{\phi_l\}_1^{k-1}$. We first show that $\Pi_{\mathcal{S}_k} u_k \neq 0$, and so we can define

$$\phi_k = \frac{1}{\|\Pi_{\mathcal{S}_k} u_k\|} \Pi_{\mathcal{S}_k} u_k \tag{20}$$

ensuring $\phi_k \perp \operatorname{Span}\{\phi_l\}_{l=1}^{k-1}$. After that we verify the validity of (17) and (18) for $j = k$.

Decompose

$$u_k = \Pi_{\mathcal{S}_k} u_k + \sum_{l=k-1}^{1} \langle u_k, \phi_l \rangle \phi_l + u_k^\perp$$

where $u_k^\perp \perp \mathrm{Span}\{\phi_l\}_{l=1}^{k-1} \oplus \mathcal{S}_k$. Then

$$
\begin{aligned}
{F_{\mathcal{L}}}^k(t)^2 &= \mathfrak{d}_k(t)^2 \|\Pi_{\mathcal{S}_k} u_k\|^2 + \sum_{l=k-1}^{1} \mathfrak{d}_l(t)^2 |\langle u_k, \phi_l\rangle|^2 + \|(A-t)u_k^\perp\|^2 \\
&\geq \mathfrak{d}_k(t)^2 \|\Pi_{\mathcal{S}_k} u_k\|^2 + \sum_{l=k-1}^{1} \mathfrak{d}_l(t)^2 |\langle u_k, \phi_l\rangle|^2 + \delta_k(t)^2 \|u_k^\perp\|^2 \\
&= \mathfrak{d}_k(t)^2 (1 - \|u_k^\perp\|^2) + \sum_{l=k-1}^{1} (\mathfrak{d}_l(t)^2 - \mathfrak{d}_k(t)^2) |\langle u_k, \phi_l\rangle|^2 + \delta_k(t)^2 \|u_k^\perp\|^2.
\end{aligned}
$$

The conclusion (17) up to $k-1$, implies $|\langle u_l, \phi_l\rangle|^2 \geq 1 - \varepsilon_l^2$ for index $l = 1, \ldots, k-1$. Since $\langle u_k, u_l\rangle = 0$ for $l \neq k$,

$$|\langle u_l, \phi_l\rangle||\langle u_k, \phi_l\rangle| = |\langle u_k, u_l - \langle u_l, \phi_l\rangle \phi_l\rangle|.$$

Then, the Cauchy-Schwarz inequality alongside with (17) yield

$$|\langle u_k, \phi_l\rangle|^2 \leq \frac{\varepsilon_l^2}{1 - \varepsilon_l^2}. \tag{21}$$

Hence, since $\mathfrak{d}_l(t) \leq \mathfrak{d}_k(t)$,

$$F_{\mathcal{L}}^k(t)^2 \geq \mathfrak{d}_k(t)^2 + \sum_{l=k-1}^{1} (\mathfrak{d}_l(t)^2 - \mathfrak{d}_k(t)^2) \frac{\varepsilon_l^2}{1 - \varepsilon_l^2} + (\delta_k(t)^2 - \mathfrak{d}_k(t)^2) \|u_k^\perp\|^2.$$

Clearing $\|u_k^\perp\|^2$ from this inequality and combining with the validity of (21) and (17) up to $k-1$, yields $\Pi_{\mathcal{S}_k} u_k \neq 0$.

Let ϕ_k be as in (20). Then (17) is guaranteed for $j = k$. On the other hand, (17) up to $j = k$, (21) and the identity

$$F_{\mathcal{L}}^k(t)^2 = \mathfrak{d}_k(t)^2 |\langle u_k, \phi_k\rangle|^2 + \|(A-t)(u_k - \langle u_k, \phi_k\rangle \phi_k)\|^2,$$

yield (18) up to $j = k$. □

The main result of this section is Proposition 1, which is central to the hierarchical method for finding eigenvalue inclusions examined a few years ago in [16, 17]. For fixed $\mathcal{L}$ this method leads to bounds for eigenvalues which are far sharper than those obtained from the obvious idea of estimating local minima of $F_{\mathcal{L}}^1(t)$. It was later shown [18] that this hierarchical method is equivalent to the method considered in [25], which extends to the indefinite case the classical Temple-Lehmann-Goerisch inequality. From an abstract perspective, Proposition 1 provides an intuitive insight on the mechanism for determining complementary bounds for eigenvalues (in the left definite case, for example). Even though the method proposed in [16, 17, 18] is yet to be explored more systematically in the practical setting, in most circumstances the technique described in [25] appears to be easier to implement.

3. The method of Zimmermann and Mertins

Let $t \in \mathbb{R}$ and $\mathcal{L} \subset \mathrm{D}(A)$ be a specified trial subspace as above. Recall that q_t is given by (4). Let $l_t : \mathrm{D}(A) \times \mathrm{D}(A) \longrightarrow \mathbb{C}$ be the (generally not closed) bilinear form associated to $(A - t)$,

$$l_t(u, w) = \langle (A - t)u, w\rangle \qquad \forall u, w \in \mathrm{D}(A).$$

Our next purpose is to characterize the optimal parameters $t^{\pm}$ in Proposition 1 as described in Remark 1 by means of the following weak eigenvalue problem,

$$(\mathrm{Z}_t^{\mathcal{L}}) \qquad \begin{array}{l} \text{find } u \in \mathcal{L} \setminus \{0\} \text{ and } \tau \in \mathbb{R} \text{ such that} \\ \tau q_t(u, v) = l_t(u, v) \qquad \forall v \in \mathcal{L}. \end{array}$$

This problem is central to the method of eigenvalue bounds calculation examined in [25] and it will be at the core of the numerical strategy presented in Section 6.

Let

$$\tau_1^-(t) \leq \ldots \leq \tau_{n^-}^-(t) < 0 \qquad \text{and} \qquad 0 < \tau_{n^+}^+(t) \leq \ldots \leq \tau_1^+(t),$$

be the negative and positive eigenvalues of $(\mathrm{Z}_t^{\mathcal{L}})$ respectively. Here and below $n^{\mp}(t)$ is the number of these negative and positive eigenvalues, which are both locally constant in t. Below we will denote eigenfunctions associated with $\tau_j^{\mp}(t)$ by $u_j^{\mp}(t)$.

The hypotheses (14) ensure the existence of $\tau_1^{\mp}(t)$. A more concrete connection with the framework of Section 2 is made precise in the following lemma. Its proof is straightforward, hence omitted.

Lemma 4. *In the following lists, the conditions stated are equivalent.*

- a^-) $F_{\mathcal{L}}^1(s) > t - s$ *for all* $s < t$
- b^-) $\frac{\langle Au,u\rangle}{\langle u,u\rangle} > t$ *for all* $u \in \mathcal{L}$
- c^-) *all the eigenvalues of* $(\mathrm{Z}_t^{\mathcal{L}})$ *are positive,*

- a^+) $F_{\mathcal{L}}^1(s) < s - t$ *for all* $s > t$
- b^+) $\frac{\langle Au,u\rangle}{\langle u,u\rangle} < t$ *for all* $u \in \mathcal{L}$
- c^+) *all the eigenvalues of* $(\mathrm{Z}_t^{\mathcal{L}})$ *are negative.*

Remark 3. Let $\mathcal{L} = \mathrm{Span}\{b_j\}_{j=1}^n$. The matrix $[q_t(b_j, b_k)]_{jk=1}^n$ is singular if and only if $\mathcal{E}_t(A) \cap \mathcal{L} \neq \{0\}$. On the other hand, the kernel of $(\mathrm{Z}_t^{\mathcal{L}})$ might be non-empty. If $n_0(t)$ is the dimension of this kernel and $n_\infty(t) = \dim(\mathcal{E}_t(A) \cap \mathcal{L})$, then $n = n_\infty(t) + n_0(t) + n^-(t) + n^+(t)$.

Assumption 2. *Note that $n_\infty(t) \geq 1$ if and only if $F^j_{\mathcal{L}}(t) = 0$ for $j = 1, \dots, n_\infty(t)$. In this case the conclusions of Lemma 5 and Theorem 6 below become void. In order to write our statements in a more transparent fashion, without further mention from now on we will suppose that*

$$\mathcal{L} \cap \mathcal{E}_t(A) = \{0\}. \tag{22}$$

By virtue of the next three results, finding the eigenvalues of $(\mathrm{Z}^{\mathcal{L}}_t)$ is equivalent to finding $s = \hat{t}^{\pm}_j \in \mathbb{R}$ such that

$$t - s = \mp F^j_{\mathcal{L}}(s), \tag{23}$$

and in this case $\hat{t}^{\pm}_j = t + \frac{1}{2\tau^{\pm}_j(t)}$. It then follows from Remark 1 that $(\mathrm{Z}^{\mathcal{L}}_t)$ encodes information about the optimal bounds for the spectrum around t, achievable by (13) in Proposition 1.

We begin with the case $j = 1$, see [18, Theorem 11].

Lemma 5. *Let $t \in \mathbb{R}$.*

(−) *The smallest eigenvalue $\tau = \tau^-_1(t)$ of $(\mathrm{Z}^{\mathcal{L}}_t)$ is negative if and only if there exists $s < t$ such that $F^1_{\mathcal{L}}(s) = t - s$. In this case $s = t + \frac{1}{2\tau^-_1(t)}$ and*

$$F^1_{\mathcal{L}}(s) = -\frac{1}{2\tau^-_1(t)} = \frac{|u^-_1(t)|_s}{\|u^-_1(t)\|}$$

for $u = u^-_1(t) \in \mathcal{L}$ the corresponding eigenvector.

(+) *The largest eigenvalue $\tau = \tau^+_1(t)$ of $(\mathrm{Z}^{\mathcal{L}}_t)$ is positive if and only if there exists $s > t$ such that $F^1_{\mathcal{L}}(s) = s - t$. In this case $s = t + \frac{1}{2\tau^+_1(t)}$ and*

$$F^1_{\mathcal{L}}(s) = \frac{1}{2\tau^+_1(t)} = \frac{|u^+_1(t)|_s}{\|u^+_1(t)\|}$$

for $u = u^+_1(t) \in \mathcal{L}$ the corresponding eigenvector.

Proof. We only show (−), as the proof of (+) is similar. For all $u \in \mathcal{L}$ and $s \in \mathbb{R}$,

$$\begin{aligned} &q_s(u,u) - F^1_{\mathcal{L}}(s)^2 \langle u, u\rangle \\ &\qquad = q_t(u,u) + 2(t-s) l_t(u,u) + \left((t-s)^2 - F^1_{\mathcal{L}}(s)^2\right) \langle u, u\rangle. \end{aligned}$$

Suppose that $F^1_{\mathcal{L}}(s) = t - s$. Then

$$q_s(u,u) - F^1_{\mathcal{L}}(s)^2 \langle u, u\rangle = q_t(u,u) + 2F^1_{\mathcal{L}}(s) l_t(u,u).$$

As the left side of this expression is non-negative,

$$\frac{l_t(u,u)}{q_t(u,u)} \geq -\frac{1}{2F^1_{\mathcal{L}}(s)}$$

for all $u \in \mathcal{L}\backslash\{0\}$ and the equality holds for some $u \in \mathcal{L}$. Hence $-\frac{1}{2F^1_{\mathcal{L}}(s)}$ is the smallest eigenvalue of $(\mathrm{Z}^{\mathcal{L}}_t)$, and thus necessarily equal to $\tau_1^-(t)$. In this case $s - F^1_{\mathcal{L}}(s) = t - 2F^1_{\mathcal{L}}(s) = t + \frac{1}{\tau_1^-(t)}$. Here the vector u for which equality is achieved is exactly $u = u_1^-(t)$.

Conversely, let $\tau_1^-(t)$ and $u_1^-(t)$ be as stated. Then

$$\tau_1^-(t) \leq \frac{l_t(u,u)}{q_t(u,u)}$$

for all $u \in \mathcal{L}$ with equality for $u = u_1^-(t)$. Re-arranging this expression yields

$$q_t(u,u) - \frac{1}{\tau_1^-(t)} l_t(u,u) \geq 0$$

for all $u \in \mathcal{L}$ with equality for $u = u_1^-(t)$. The substitution $t = s - \frac{1}{2\tau_1^-(t)}$ then yields

$$q_t(u,u) - \frac{1}{(2\tau_1^-(t))^2} \langle u,u \rangle \geq 0$$

for all $u \in \mathcal{L}$. The equality holds for $u = u_1^-(t)$. This expression further re-arranges as

$$\frac{|u|_s^2}{\|u\|^2} \geq \frac{1}{(2\tau_1^-(t))^2}.$$

Hence $F^1_{\mathcal{L}}(s)^2 = \frac{1}{(2\tau_1^-(t))^2}$, as needed. □

An extension to $j \neq 1$ is now found by induction.

Theorem 6. *Let $t \in \mathbb{R}$ and $1 \leq j \leq n$ be fixed.*

(−) *The number of negative eigenvalues $n^-(t)$ in $(\mathrm{Z}^{\mathcal{L}}_t)$ is greater than or equal to j if and only if*

$$\frac{\langle Au,u \rangle}{\langle u,u \rangle} < t \qquad \text{for some} \quad u \in \mathcal{L} \ominus \mathrm{Span}\{u_1^-(t), \ldots, u_{j-1}^-(t)\}.$$

Assuming this holds true, then $\tau = \tau_j^-(t)$ and $u = u_j^-(t)$ are solutions of $(\mathrm{Z}^{\mathcal{L}}_t)$ if and only if

$$F^j_{\mathcal{L}}\left(t + \frac{1}{2\tau_j^-(t)}\right) = -\frac{1}{2\tau_j^-(t)} = \frac{\left|u_j^-(t)\right|_{t+\frac{1}{2\tau_j^-(t)}}}{\|u_j^-(t)\|}.$$

(+) *The number of positive eigenvalues $n^+(t)$ in $(\mathrm{Z}^{\mathcal{L}}_t)$ is greater than or equal to j if and only if*

$$\frac{\langle Au,u \rangle}{\langle u,u \rangle} > t \qquad \text{for some} \quad u \in \mathcal{L} \ominus \mathrm{Span}\{u_1^+(t), \ldots, u_{j-1}^+(t)\}.$$

Assuming this holds true, then $\tau = \tau_j^+(t)$ *and* $u = u_j^+(t)$ *are solutions of* $(\mathrm{Z}_t^{\mathcal{L}})$ *if and only if*

$$F_{\mathcal{L}}^j\left(t+\frac{1}{2\tau_j^+(t)}\right)=\frac{1}{2\tau_j^+(t)}=\frac{\left|u_j^+(t)\right|_{t+\frac{1}{2\tau_j^+(t)}}}{\|u_j^+(t)\|}.$$

Proof. For $j=1$ the statements are Lemma 5 taking into consideration (14). For $j>1$, due to the symmetry of the eigenproblem $(\mathrm{Z}_t^{\mathcal{L}})$, it is enough to apply again Lemma 5 by fixing $\tilde{\mathcal{L}}=\mathcal{L}\ominus\mathrm{Span}\{u_1^{\mp}(t),\ldots,u_{j-1}^{\mp}(t)\}$ as trial spaces. Note that the eigenvalues of $(\mathrm{Z}_t^{\tilde{\mathcal{L}}})$ are those of $(\mathrm{Z}_t^{\mathcal{L}})$ except for $\tau_1^{\mp}(t),\ldots,\tau_{j-1}^{\mp}(t)$. □

A neat procedure for finding certified spectral bounds for A, as described in [25], can now be deduced from Theorem 6. By virtue of Proposition 1 and Remark 1, this procedure turns out to be optimal in the context of the approximated counting functions discussed in Section 2, see [18, Section 6]. We summarize the core statement as follows.

Corollary 7. *For all* $t\in\mathbb{R}$ *and* $j\in\{1,\ldots,n^{\pm}(t)\}$,

$$t+\frac{1}{\tau_j^-(t)}\le\mathfrak{n}_j^-(t)\qquad\text{and}\qquad\mathfrak{n}_j^+(t)\le t+\frac{1}{\tau_j^+(t)}. \tag{24}$$

In recent years, numerical techniques based on this statement have been designed to successfully compute eigenvalues for the radially reduced magnetohydrodynamics operator [25, 11], the Helmholtz equation [5] and the calculation of sloshing frequencies in the left definite case [4]. We will determine one such a numerical scheme for the case of the Maxwell operator in sections 5 and 6.

Remark 4. Since $\pm\frac{1}{\tau_j^{\pm}(t)}\ge\pm(\mathfrak{n}_j^{\pm}(t)-t)$ in the above,

$$\begin{aligned}\hat{t}_j^-&=t+\frac{1}{2\tau_j^-(t)}\le\frac{t+\mathfrak{n}_j^-(t)}{2}\le\frac{\mathfrak{n}_j^+(t)+\mathfrak{n}_j^-(t)}{2}\\&\le\frac{\mathfrak{n}_j^+(t)+t}{2}\le t+\frac{1}{2\tau_j^+(t)}=\hat{t}_j^+.\end{aligned}$$

Hence $\hat{t}_j^{\pm}$ is not further from $\mathfrak{n}_j^{\pm}(t)$ than it is to $\mathfrak{n}_j^{\mp}(t)$. Moreover

$$\hat{t}_j^{\pm}=\frac{\mathfrak{n}_j^+(t)+\mathfrak{n}_j^-(t)}{2}$$

renders $t\in\sigma(A)$ and

$$\frac{1}{\tau_j^{\pm}(t)}=\mathfrak{n}_j^{\pm}(t)-t.$$

This geometrical property for the solution of (23) will be relevant below, when examining the convergence properties of the estimates (24).

4. Convergence and error estimates

Our first goal in this section will be to show that, if $\mathcal{L}$ captures an eigenspace of A within a certain order of precision $\mathcal{O}(\varepsilon)$ as specified below, then the bounds consequence of Proposition 1 are

a) at least within $\mathcal{O}(\varepsilon)$ from the true spectral data for any $t \in \mathbb{R}$,
b) within $\mathcal{O}(\varepsilon^2)$ for $t \notin \sigma(A)$.

This will be the content of theorems 9 and 10, and Corollary 11. We will then show that, in turns, the estimates (24) have always residual of size $\mathcal{O}(\varepsilon^2)$ for any $t \in \mathbb{R}$. See Theorem 12. In the spectral approximation literature this property is known as optimal order of convergence/exactness, see [14, Chapter 6] or [24].

Recall Remark 2, and the assumptions 1 and 2. Below $\{\phi_j^t\}_{j=1}^m$ denotes an orthonormal set of eigenvectors of $\mathcal{E}_{[t-\mathfrak{d}_m(t),t+\mathfrak{d}_m(t)]}(A)$ which is ordered so that

$$|A-t|\phi_j^t = \mathfrak{d}_j(t)\phi_j^t \qquad \text{for} \qquad j=1,\ldots,m.$$

Whenever $0<\varepsilon_j<1$ is small, as specified below, the trial subspace $\mathcal{L} \subset \mathrm{D}(A)$ will be assumed to be close to $\mathrm{Span}\{\phi_j^t\}_{j=1}^m$ in the sense that there exist $w_j^t \in \mathcal{L}$ such that

$$(\mathrm{A}_0) \qquad \|w_j^t - \phi_j^t\| \le \varepsilon_j \qquad \text{and}$$

$$(\mathrm{A}_1) \qquad |w_j^t - \phi_j^t|_t \le \varepsilon_j.$$

We have split this condition into two, in order to highlight the fact that some times only (A_1) is required. Unless otherwise specified, the index j runs from 1 to m.

From (16) it follows that the family $\{\phi_j^s\}_{j=1}^m \subset \mathcal{E}_{[t-\mathfrak{d}_m(t),t+\mathfrak{d}_m(t)]}(A)$ and the family $\{w_j^s\}_{j=1}^m \subset \mathcal{L}$ above can always be chosen piecewise constant for s in a neighbourhood of t. Moreover, they can be chosen so that jumps only occur at $s \in \sigma(A)$.

Assumption 3. *Without further mention all t-dependant vectors below will be assumed to be locally constant in t with jumps only at the spectrum of A.*

A set $\{w_j^t\}_{j=1}^m$ subject to (A_0)-(A_1) is not generally orthonormal. However, according to the next lemma, it can always be substituted by an orthonormal set, provided ε_j is small enough.

Lemma 8. *There exists a constant $C > 0$ independent of $\mathcal{L}$ ensuring the following. If $\{w_j^t\}_{j=1}^m \subset \mathcal{L}$ is such that* (A$_0$)-(A$_1$) *hold for all ε_j such that*

$$\varepsilon = \sqrt{\sum_{j=1}^m \varepsilon_j^2} < \frac{1}{\sqrt{m}},$$

then there is a set $\{v_j^t\}_{j=1}^m \subset \mathcal{L}$ orthonormal in the inner product $\langle\cdot,\cdot\rangle$ such that

$$|v_j^t - \phi_j^t|_t + \|v_j^t - \phi_j^t\| < C\varepsilon.$$

Proof. As it is clear from the context, in this proof we suppress the index t on top of any vector. The desired conclusion is achieved by applying the Gram-Schmidt procedure. Let $G = [\langle w_k, w_l\rangle]_{kl=1}^m \in \mathbb{C}^{m\times m}$ be the Gram matrix associated to $\{w_j\}$. Set

$$v_j = \sum_{k=1}^m (G^{-1/2})_{kj}\, w_k.$$

Then

$$\begin{aligned}\|G - I\| &\le \sqrt{\sum_{kl=1}^m |\langle w_k, w_l\rangle - \langle \phi_k, \phi_l\rangle|^2} \\ &\le \sqrt{2\sum_{kl=1}^m \|w_k - \phi_k\|^2(\|w_l\| + \|\phi_l\|)^2} \\ &\le \sqrt{2}(2+\varepsilon)\varepsilon.\end{aligned}$$

Since

$$\begin{aligned}\|v_j - w_j\|^2 &= \left\|\sum_{k=1}^m (G^{-1/2} - I)_{kj}\, w_k\right\|^2 \\ &= \sum_{kl=1}^m (G^{-1/2} - I)_{kj}\overline{(G^{-1/2} - I)_{lj}}\langle w_k, w_l\rangle \\ &= \sum_{k=1}^m (G^{-1/2} - I)_{kj}\overline{\left(\sum_{l=1}^m G_{kl}(G^{-1/2} - I)_{lj}\right)} \\ &= \sum_{k=1}^m (G^{-1/2} - I)_{kj}(G^{1/2} - G)_{jk} \\ &= \left((I - G^{1/2})^2\right)_{jj}\end{aligned}$$

then

$$\|v_j - w_j\| \leq \|I - G^{1/2}\|.$$

As $G^{1/2}$ is a positive-definite matrix, for every $\underline{v} \in \mathbb{C}^m$ we have

$$\|(G^{1/2} + I)\underline{v}\|^2 = \|G^{1/2}\underline{v}\|^2 + 2\langle G^{1/2}\underline{v}, \underline{v}\rangle + \|\underline{v}\|^2 \geq \|\underline{v}\|^2.$$

Then $\det(I + G^{1/2}) \neq 0$ and $\|(I + G^{1/2})^{-1}\| \leq 1$. Hence

$$\|v_j - w_j\| \leq \|(I - G)(I + G^{1/2})^{-1}\| \leq \|I - G\| \, \|(I + G^{1/2})^{-1}\| \leq (2+\varepsilon)\varepsilon \,. \tag{25}$$

Now, identify $\underline{v} = (v_1, \ldots, v_m) \in \mathbb{C}^m$ with $v = \sum_{k=1}^m v_k \phi_k$. As

$$\|G^{1/2}\underline{v}\| = \left\| \sum_{j=1}^m \langle v, \phi_j\rangle w_j \right\| \geq \|v\| - \left\| \sum_{j=1}^m \langle v, \phi_j\rangle (w_j - \phi_j) \right\| \geq (1-\varepsilon)\|\underline{v}\|$$

then

$$\|G^{-1/2}\| \leq \frac{1}{1-\varepsilon}.$$

Hence

$$\begin{aligned} |v_j - w_j|_t &\leq \sum_{k=1}^m |(G^{-1/2} - I)_{jk}||w_k|_t \\ &\leq \sum_{k=1}^m |(G^{-1/2} - I)_{jk}|(\varepsilon_k + \mathfrak{d}_k(t)) \\ &\leq \sum_{kl=1}^m |(G^{-1/2})_{kl}||(G^{1/2} - I)_{lj}|(\varepsilon_k + \mathfrak{d}_k(t)) \\ &\leq \frac{\sqrt{m}(\varepsilon + \mathfrak{d}_m(t))(2+\varepsilon)}{1-\varepsilon}\varepsilon. \end{aligned} \tag{26}$$

The desired conclusion follows from (25) and (26). □

The next theorem addresses the claim *a)* made at the beginning of this section. According to Lemma 8, in order to examine the asymptotic behaviour of $F_{\mathcal{L}}^j(t)$ as $\varepsilon_j \to 0$ under the constraints (A_0)-(A_1), we can assume without loss of generality that the trial vectors w_j^t form an orthonormal set in the inner product $\langle\cdot,\cdot\rangle$.

Theorem 9. *Let $\{w_j^t\}_{j=1}^m \subset \mathcal{L}$ be a family of vectors which is orthonormal in the inner product $\langle\cdot,\cdot\rangle$ and satisfies* (A_1). *Then*

$$F_{\mathcal{L}}^j(t) - \mathfrak{d}_j(t) \leq \left(\sum_{k=1}^j \varepsilon_k^2\right)^{1/2} \qquad \forall j = 1, \ldots, m.$$

Proof. From the min-max principle we obtain

$$\begin{aligned} F^j_{\mathcal{L}}(t) &\leq \max_{\sum|c_k|^2=1} \left|\sum_{k=1}^j c_k w_k\right|_t \\ &\leq \max_{\sum|c_k|^2=1} \left|\sum_{k=1}^j c_k (w_k-\phi_k)\right|_t + \max_{\sum|c_k|^2=1} \left|\sum_{k=1}^j c_k \phi_k\right|_t \\ &= \max_{\sum|c_k|^2=1} \left|\sum_{k=1}^j c_k (w_k-\phi_k)\right|_t + \mathfrak{d}_j(t). \end{aligned}$$

This gives

$$\begin{aligned} F^j_{\mathcal{L}}(t) - \mathfrak{d}_j(t) &\leq \max_{\sum|c_k|^2=1} \sum_{k=1}^j |c_k||w_k-\phi_k|_t \\ &\leq \max_{\sum|c_k|^2=1} \left(\sum_{k=1}^j |c_k|^2\right)^{1/2} \left(\sum_{k=1}^j |w_k-\phi_k|_t^2\right)^{1/2} \leq \left(\sum_{k=1}^j \varepsilon_k^2\right)^{1/2} \end{aligned}$$

as needed. □

In terms of order of approximation, Theorem 9 will be superseded by Theorem 10 for $t \notin \sigma(A)$. However, if $t \in \sigma(A)$, the trial space $\mathcal{L}$ can be chosen so that $F^1_{\mathcal{L}}(t) - \mathfrak{d}_1(t)$ is linear in ε_1. Indeed, fixing any non-zero $u \in \mathrm{D}(A)$ and $\mathcal{L} = \mathrm{Span}\{u\}$, yields $F^1_{\mathcal{L}}(t) - \mathfrak{d}_1(t) = F^1_{\mathcal{L}}(t) = \varepsilon_1$. This shows that Theorem 9 is optimal, upon the presumption that t is arbitrary.

The next theorem addresses the claim *b)* made at the beginning of this section. Its proof is reminiscent of that of [23, Theorem 6.1].

Theorem 10. *Let $t \notin \sigma(A)$. Suppose that the ε_j in* (A_1) *are such that*

$$\sum_{j=1}^m \varepsilon_j^2 < \frac{\mathfrak{d}_1(t)^2}{6}. \tag{27}$$

Then,

$$F^j_{\mathcal{L}}(t) - \mathfrak{d}_j(t) \leq 3\frac{\mathfrak{d}_j(t)}{\mathfrak{d}_1(t)^2} \sum_{k=1}^j \varepsilon_k^2 \qquad \forall j = 1, \ldots, m. \tag{28}$$

Proof. Since $t \notin \sigma(A)$, then $(\mathrm{D}(A), q_t(\cdot,\cdot))$ is a Hilbert space. Let $P_{\mathcal{L}} : \mathrm{D}(A) \longrightarrow \mathcal{L}$ be the orthogonal projection onto $\mathcal{L}$ with respect to the inner product $q_t(\cdot,\cdot)$, so that

$$q_t(u - P_{\mathcal{L}}u, v) = 0 \qquad \forall v \in \mathcal{L}.$$

Then $|u|_t^2 = |P_{\mathcal{L}}u|_t^2 + |u - P_{\mathcal{L}}u|_t^2$ for all $u \in \mathrm{D}(A)$ and $|u - P_{\mathcal{L}}u|_t \leq |u - v|_t$ for all $v \in \mathcal{L}$. Hence

$$|\phi_k - P_{\mathcal{L}}\phi_k|_t \leq \varepsilon_k \qquad \forall k = 1, \dots, m. \tag{29}$$

Let $\mathcal{E}_j = \mathrm{Span}\{\phi_k\}_{k=1}^j$. Define

$$\mathcal{F}_j = \{\phi \in \mathcal{E}_j : \|\phi\| = 1\} \qquad \text{and}$$
$$\mu_{\mathcal{L}}^j(t) = \max_{\phi \in \mathcal{F}_j} \left|2\,\mathrm{Re}\langle \phi, \phi - P_{\mathcal{L}}\phi\rangle - \|\phi - P_{\mathcal{L}}\phi\|^2\right|.$$

Here $\mu_{\mathcal{L}}^j$ depends on t, as $P_{\mathcal{L}}$ does. We first show that, under hypothesis (27), $\mu_{\mathcal{L}}^j(t) < \frac{1}{2}$. Indeed, given $\phi \in \mathcal{F}_j$ we decompose it as $\phi = \sum_{k=1}^j c_k \phi_k$. Then

$$\begin{aligned}
|\langle \phi, \phi - P_{\mathcal{L}}\phi\rangle| &= \left|\sum_{k=1}^j c_k \langle \phi_k, \phi - P_{\mathcal{L}}\phi\rangle\right| = \left|\sum_{k=1}^j \frac{c_k}{\mathfrak{d}_k(t)^2} q_t(\phi_k, \phi - P_{\mathcal{L}}\phi)\right| \\
&= \left|q_t\left(\sum_{k=1}^j \frac{c_k}{\mathfrak{d}_k(t)^2}\phi_k, \phi - P_{\mathcal{L}}\phi\right)\right| \\
&= \left|q_t\left(\sum_{k=1}^j \frac{c_k}{\mathfrak{d}_k(t)^2}(\phi_k - P_{\mathcal{L}}\phi_k), \phi - P_{\mathcal{L}}\phi\right)\right| \\
&\leq \left|\sum_{k=1}^j \frac{c_k}{\mathfrak{d}_k(t)^2}(\phi_k - P_{\mathcal{L}}\phi_k)\right|_t \left|\sum_{k=1}^j c_k(\phi_k - P_{\mathcal{L}}\phi_k)\right|_t. \qquad (30)
\end{aligned}$$

For each multiplying term in the latter expression, the triangle and Cauchy-Schwarz's inequalities yield (take $\alpha_k = c_k$ or $\alpha_k = \frac{c_k}{\mathfrak{d}_k(t)^2}$)

$$\begin{aligned}
\left|\sum_{k=1}^j \alpha_k(\phi_k - P_{\mathcal{L}}\phi_k)\right|_t &\leq \sum_{k=1}^j |\alpha_k|\,|\phi_k - P_{\mathcal{L}}\phi_k|_t \\
&\leq \left(\sum_{k=1}^j |\alpha_k|^2\right)^{1/2} \left(\sum_{k=1}^j |\phi_k - P_{\mathcal{L}}\phi_k|_t^2\right)^{1/2}. \qquad (31)
\end{aligned}$$

Then

$$\begin{aligned}
|2\,\mathrm{Re}\langle \phi, \phi - P_{\mathcal{L}}\phi\rangle| &\leq 2\left(\sum_{k=1}^j \frac{|c_k|^2}{\mathfrak{d}_k(t)^4}\right)^{1/2} \left(\sum_{k=1}^j |c_k|^2\right)^{1/2} \sum_{k=1}^j \varepsilon_k^2 \\
&\leq \frac{2}{\mathfrak{d}_1(t)^2} \sum_{k=1}^j \varepsilon_k^2 \qquad (32)
\end{aligned}$$

for all $\phi \in \mathcal{F}_j$.

The other term in the expression for $\mu_{\mathcal{L}}^j(t)$ has an upper bound found as follows. According to the min-max principle

$$\|\phi - P_{\mathcal{L}}\phi\|^2 \le \frac{1}{\mathfrak{d}_1(t)^2} q_t(\phi - P_{\mathcal{L}}\phi, \phi - P_{\mathcal{L}}\phi). \tag{33}$$

Therefore, by repeating analogous steps as in (30) and (31), we get

$$\begin{aligned}
\|\phi - P_{\mathcal{L}}\phi\|^2 &\le \frac{1}{\mathfrak{d}_1(t)^2} \sum_{k=1}^{j} c_k q_t(\phi_k - P_{\mathcal{L}}\phi_k, \phi - P_{\mathcal{L}}\phi) \\
&= q_t\left(\sum_{k=1}^{j} \frac{c_k}{\mathfrak{d}_1(t)^2}(\phi_k - P_{\mathcal{L}}\phi_k), \phi - P_{\mathcal{L}}\phi\right) \\
&= q_t\left(\sum_{k=1}^{j} \frac{c_k}{\mathfrak{d}_1(t)^2}(\phi_k - P_{\mathcal{L}}\phi_k), \sum_{l=1}^{j} c_l(\phi_l - P_{\mathcal{L}}\phi_l)\right) \\
&\le \frac{1}{\mathfrak{d}_1(t)^2} \sum_{k=1}^{j} \varepsilon_k^2 . \qquad (34)
\end{aligned}$$

Hence, from (32) and (34),

$$\mu_{\mathcal{L}}^j(t) \le \frac{3}{\mathfrak{d}_1(t)^2} \sum_{k=1}^{j} \varepsilon_k^2 < \frac{1}{2} \tag{35}$$

as a consequence of (27).

Next, observe that $\dim(P_{\mathcal{L}}\mathcal{E}_j) = j$. Indeed $P_{\mathcal{L}}\psi = 0$ for $\|\psi\| = 1$ would imply

$$\mu_{\mathcal{L}}^j(t) \ge \left|2\operatorname{Re}\langle\psi, \psi - P_{\mathcal{L}}\psi\rangle - \|\psi - P_{\mathcal{L}}\psi\|^2\right| = \|\psi\|^2 = 1,$$

which would contradict the fact that $\mu_{\mathcal{L}}^j(t) < 1$. Then,

$$F_{\mathcal{L}}^j(t)^2 \le \max_{u \in P_{\mathcal{L}}\mathcal{E}_j} \frac{|u|_t^2}{\|u\|^2} = \max_{\phi \in \mathcal{E}_j} \frac{|P_{\mathcal{L}}\phi|_t^2}{\|P_{\mathcal{L}}\phi\|^2} = \max_{\phi \in \mathcal{F}_j} \frac{|P_{\mathcal{L}}\phi|_t^2}{\|P_{\mathcal{L}}\phi\|^2}.$$

As

$$\|P_{\mathcal{L}}\phi\|^2 = \|\phi\|^2 - 2\operatorname{Re}\langle\phi, \phi - P_{\mathcal{L}}\phi\rangle + \|\phi - P_{\mathcal{L}}\phi\|^2 \ge 1 - \mu_{\mathcal{L}}^j(t),$$

we get

$$F_{\mathcal{L}}^j(t)^2 \le \max_{\phi \in \mathcal{F}_j} \frac{|\phi|_t^2}{1 - \mu_{\mathcal{L}}^j(t)} = \max_{\sum |c_k|^2 = 1} \frac{\sum_{k=1}^{j} |c_k|^2 \mathfrak{d}_k(t)^2}{1 - \mu_{\mathcal{L}}^j(t)} = \frac{\mathfrak{d}_j(t)^2}{1 - \mu_{\mathcal{L}}^j(t)}. \tag{36}$$

Finally, (36) and (35) yield

$$\begin{aligned} F_{\mathcal{L}}^j(t)^2 - \mathfrak{d}_j(t)^2 &\leq \frac{\mu_{\mathcal{L}}^j(t)}{1-\mu_{\mathcal{L}}^j(t)}\mathfrak{d}_j(t)^2 \\ &\leq 2\mu_{\mathcal{L}}^j(t)\mathfrak{d}_j(t)^2 \\ &\leq 2\frac{3}{\mathfrak{d}_1(t)^2}\mathfrak{d}_j(t)^2\sum_{k=1}^{j}\varepsilon_k^2. \end{aligned} \tag{37}$$

The proof is completed by observing that $F_{\mathcal{L}}^j(t) + \mathfrak{d}_j(t) \geq 2\mathfrak{d}_j(t)$. □

As the next corollary shows, a quadratic order of decrease for the difference $F_{\mathcal{L}}^j(t) - \mathfrak{d}_j(t)$ is prevented for $t \in \sigma(A)$ in the context of theorems 9 and 10, only for j up to $\dim \mathcal{E}_t(A)$.

Corollary 11. *Let $t \in \sigma_{\rm disc}(A)$, $\ell = 1 + \dim \mathcal{E}_t(A)$ and $k \in \{\ell, \dots, m\}$. Let*

$$\alpha_k(t) = \frac{1}{4}\min\left\{|\mathfrak{d}_l(t) - \mathfrak{d}_{l-1}(t)| : \mathfrak{d}_l(t) \neq \mathfrak{d}_{l-1}(t),\, l = \ell, ..., k\right\} > 0.$$

There exists $\varepsilon > 0$ independent of k ensuring the following. If (A$_1$) *holds true for $\sqrt{\sum_{j=1}^m \varepsilon_j^2} < \varepsilon$, then*

$$F_{\mathcal{L}}^k(t) - \mathfrak{d}_k(t) \leq 3\frac{\mathfrak{d}_k(t)}{\alpha_k(t)^2}\sum_{j=1}^{k}\varepsilon_j^2.$$

Proof. Without loss of generality we assume that $t + \mathfrak{d}_k(t) \in \sigma(A)$. Otherwise $t - \mathfrak{d}_k(t) \in \sigma(A)$ and the proof is analogous to the one presented below.

Let $\tilde{t} = t + \alpha_k(t)$. Then $\tilde{t} \notin \sigma(A)$ and $t + \mathfrak{d}_k(t) = \tilde{t} + \mathfrak{d}_k(\tilde{t})$. Since the map $s \mapsto s + F_{\mathcal{L}}^j(s)$ is non-decreasing as a consequence of Proposition 1, Theorem 10 applied at $\tilde{t}$ yields

$$\begin{aligned} F_{\mathcal{L}}^k(t) - \mathfrak{d}_k(t) = t + F_{\mathcal{L}}^k(t) - (t + \mathfrak{d}_k(t)) &\leq \tilde{t} + F_{\mathcal{L}}^k(\tilde{t}) - (\tilde{t} + \mathfrak{d}_k(\tilde{t})) \\ = F_{\mathcal{L}}^k(\tilde{t}) - \mathfrak{d}_k(\tilde{t}) &\leq 3\frac{\mathfrak{d}_k(\tilde{t})}{\mathfrak{d}_1(\tilde{t})^2}\sum_{j=1}^{k}\varepsilon_k^2 \leq 3\frac{\mathfrak{d}_k(t)}{\alpha_k(t)^2}\sum_{j=1}^{k}\varepsilon_j^2 \end{aligned}$$

as needed. □

For the final part of this section, we formulate a precise statements on the convergence of the method of Zimmermann and Mertins. Theorem 12 below improves upon two crucial aspects of a similar result established in [11, Lemma 2]. It allows $j > 1$ and it allows $t \in \sigma(A)$. These two improvements are essential in order to obtain sharp bounds for those eigenvalues of the Maxwell operator which are either degenerate or form a tight cluster.

Remark 5. The constants $\tilde{\varepsilon}_t$ and $C_t^{\pm}$ below do have a dependance on t that may be determined explicitly from Theorem 10, Corollary 11 and the proof of Theorem 12. Despite of the fact that they can deteriorate as t approaches the isolated eigenvalues of A and they can have jumps precisely at these points, they may be chosen locally independent of t in compacts outside the spectrum. This has an impact on practical implementations of the computational method to be described in Section 6 which we do not fully understand at present. Our numerical tests in Section 7 indicate that the best results are achieved by choosing t relatively far form the spectral point being approximated.

Set

$$\nu_j^-(t) = \sup\{s < t : \operatorname{Tr} \mathbb{1}_{(s,t)}(A) \geq j\}$$
$$\nu_j^+(t) = \inf\{s > t : \operatorname{Tr} \mathbb{1}_{(t,s)}(A) \geq j\}.$$

Note that these are the spectral points of A which are strictly to the left and strictly to the right of t respectively. The inequality $\nu_j^{\pm}(t) \neq \mathfrak{n}_j^{\pm}(t)$ only occurs when t is an eigenvalue. In view of (15), $\delta^{\pm}(t) = |t - \nu_1^{\pm}(t)|$.

Theorem 12. *Let $J \subset \mathbb{R}$ be a bounded open segment such that the intersection $J \cap \sigma(A) \subseteq \sigma_{\mathrm{disc}}(A)$. Let $\{\phi_k\}_{k=1}^{\tilde{m}}$ be a family of eigenvectors of A such that $\operatorname{Span}\{\phi_k\}_{k=1}^{\tilde{m}} = \mathcal{E}_J(A)$. For fixed $t \in J$, there exist constants $\tilde{\varepsilon}_t > 0$ and $C_t^{\pm} > 0$ independent of the trial space $\mathcal{L}$, ensuring the following. If there are $\{w_j\}_{j=1}^{\tilde{m}} \subset \mathcal{L}$ such that*

$$\left(\sum_{j=1}^{\tilde{m}} \|w_j - \phi_j\|^2 + |w_j - \phi_j|_t^2 \right)^{1/2} \leq \varepsilon < \tilde{\varepsilon}_t, \tag{38}$$

then

$$\left| \nu_j^{\pm}(t) - \left(t + \frac{1}{\tau_j^{\pm}(t)} \right) \right| \leq C_t^{\pm} \varepsilon^2$$

for all $j \leq n^{\pm}(t)$ such that $\nu_j^{\pm}(t) \in J$.

Proof. We focus on the case of the plus sign, as the one with the minus sign is completely analogous. The hypotheses ensure that the number of indices $j \leq n^{\pm}(t)$ such that $\nu_j^{\pm}(t) \in J$ never exceeds $\tilde{m}$. Therefore this condition in the conclusion of the theorem is consistent.

Let

$$m(t) = \max\{m \in \mathbb{N} : [t - \mathfrak{d}_m(t), t + \mathfrak{d}_m(t)] \subset J\}.$$

Recall the Assumption 1 and the Remark 3. The hypothesis on $\mathcal{L}$ guarantees that (A_0)-(A_1) hold true for $m = m(t)$ and

$$\left(\sum_{j=1}^{m(t)} \varepsilon_j^2\right)^{1/2} < \varepsilon.$$

By combining Lemma 8, Theorem 9 and the fact that we can pick $\{w_j^t\}_{j=1}^{m(t)} \subseteq \{w_k\}_{k=1}^{\tilde{m}}$, there exists $\tilde{\varepsilon}_t > 0$ small enough, such that (38) yields

$$F_{\mathcal{L}}^j(s) - \mathfrak{d}_j(s) \leq \frac{\nu_1^+(t) - t}{2} \qquad \forall j = 1, \ldots, \tilde{m} \quad \text{and} \quad s \in J. \tag{39}$$

Let j be such that $\nu_j^{\pm}(t) \in J$. Since $t + \alpha - \nu_1^+(t) \leq \nu_j^+(t) - (\alpha + t)$ for all $0 \leq \alpha \leq \frac{\nu_j^+(t)+\nu_1^+(t)}{2} - t$, then

$$\mathfrak{d}_j(s) = \nu_j^+(t) - s \qquad \forall s \in \left[\frac{t + \nu_j^+(t)}{2}, \frac{\nu_1^+(t) + \nu_j^+(t)}{2}\right].$$

Let

$$g(\alpha) = \alpha - F_{\mathcal{L}}^j(t + \alpha).$$

Then g is an increasing function of α and $g(0) = -F_{\mathcal{L}}^j(t) < 0$. For the strict inequality in the latter, recall Assumption 2. Moreover, according to (39)

$$\begin{aligned} & g\left(\frac{\nu_j^+(t) + \nu_1^+(t)}{2} - t\right) \\ & = \nu_1^+(t) - t + \frac{\nu_j^+(t) - \nu_1^+(t)}{2} - F_{\mathcal{L}}^j\left(\frac{\nu_j^+(t) + \nu_1^+(t)}{2}\right) \\ & = \nu_1^+(t) - t + \mathfrak{d}_j\left(\frac{\nu_j^+(t) + \nu_1^+(t)}{2}\right) - F_{\mathcal{L}}^j\left(\frac{\nu_j^+(t) + \nu_1^+(t)}{2}\right) \\ & \geq \nu_1^+(t) - t - \frac{\nu_1^+(t) - t}{2} > 0\,. \end{aligned}$$

Hence, the Mean Value Theorem ensures the existence of

$$\tilde{\alpha} \in \left(0, \frac{\nu_1^+(t) + \nu_j^+(t)}{2} - t\right)$$

such that $\tilde{\alpha} = F_{\mathcal{L}}^j(t + \tilde{\alpha})$. According to Theorem 6 (+), $\tilde{\alpha}$ is unique and $\tilde{\alpha} = \frac{1}{2\tau_j^+(t)}$.

The proof is now completed as follows. By virtue of Remark 4,

$$\hat{t}_j^+(t) = t + \frac{1}{2\tau_j^+(t)} \in \left(\frac{t + \nu_j^+(t)}{2}, \frac{\nu_1^+(t) + \nu_j^+(t)}{2} \right)$$
$$\text{and} \quad F_{\mathcal{L}}^j(\hat{t}_j^+(t)) = \frac{1}{2\tau_j^+(t)}.$$

Then, Theorem 10 or Corollary 11, as appropriate, ensure the existence of $C_t^+ > 0$ yielding

$$\nu_j^+(t) - \left(t + \frac{1}{\tau_j^+(t)} \right) = F_{\mathcal{L}}^j(\hat{t}_j^+) - \mathfrak{d}_j(\hat{t}_j^+) \le C_t^+ \sum_{k=1}^{j} \varepsilon_k^2 < \varepsilon^2 ,$$

as needed. □

We conclude this section with a result on convergence of eigenfunctions.

Corollary 13. *Let $J \subset \mathbb{R}$ be a bounded open segment such that $J \cap \sigma(A) \subseteq \sigma_{\mathrm{disc}}(A)$. Let $\{\phi_k\}_{k=1}^{\tilde{m}}$ be a family of eigenvectors of A such that $\mathrm{Span}\{\phi_k\}_{k=1}^{\tilde{m}} = \mathcal{E}_J(A)$. For fixed $t \in J$, there exist constants $\tilde{\varepsilon}_t > 0$ and $C_t^{\pm} > 0$ independent of the trial space $\mathcal{L}$, ensuring the following. If there are $\{w_j\}_{j=1}^{\tilde{m}} \subset \mathcal{L}$ guaranteeing the validity of (38), for all $j \le n^{\pm}(t)$ such that $\nu_j^{\pm}(t) \in J$ we can find $\psi_j^{\varepsilon\pm} \in \mathcal{E}_{\{\nu_j^-(t),\nu_j^+(t)\}}(A)$ satisfying*

$$|u_j^{\pm}(t) - \psi_j^{\varepsilon\pm}|_t + \|u_j^{\pm}(t) - \psi_j^{\varepsilon\pm}\| \le C_t^{\pm}\varepsilon.$$

Proof. Fix $t \in J$. By virtue of Theorem 6, $u_j^{\pm}(t) = u_j^{\hat{t}_j^{\pm}}$ in the notation for eigenvectors employed in Proposition 3. The claimed conclusion is a consequence of the latter combined with Theorem 10 or Corollary 11, as appropriate. □

5. The finite element method for the Maxwell eigenvalue problem

Recall that Ω is an open, bounded, simply connected domain of $\mathbb{R}^3$. Below $\mathcal{D}(\Omega)$ denotes the infinitely differentiable test functions with compact support in Ω. The inner product of $L^2(\Omega)$ is $\langle \cdot, \cdot \rangle_\Omega$ and its norm $\|\cdot\|_{0,\Omega}$. The Sobolev space of order m is $\mathcal{H}^m(\Omega)$ and its norm is $\|\cdot\|_{m,\Omega}$. We do not distinguish in the notation between products and norms of scalar functions or vector fields with components in these linear spaces.

We define rigorously the domain of the operator $\mathcal{M}$ associated to the eigenvalue problem (1) by following closely the ideas of the work [7]. Let

$$\mathcal{H}(\mathrm{curl};\Omega) = \left\{\boldsymbol{u} \in [L^2(\Omega)]^3 : \mathrm{curl}\,\boldsymbol{u} \in [L^2(\Omega)]^3\right\}$$

equipped with the norm

$$\|\boldsymbol{u}\|^2_{\mathrm{curl},\Omega} = \|\boldsymbol{u}\|^2_{0,\Omega} + \|\,\mathrm{curl}\,\boldsymbol{u}\|^2_{0,\Omega}. \tag{40}$$

Let $\mathcal{R}_{\max}$ denote the operator defined by the expression "curl" acting on the domain $\mathrm{D}(\mathcal{R}_{\max}) = \mathcal{H}(\mathrm{curl};\Omega)$, the maximal domain. Let

$$\mathcal{R}_{\min} = \mathcal{R}^*_{\max} = \overline{\mathcal{R}_{\max} \restriction [\mathcal{D}(\Omega)]^3}.$$

The domain of $\mathcal{R}_{\min}$ is

$$\begin{aligned}\mathrm{D}(\mathcal{R}_{\min}) &= \mathcal{H}_0(\mathrm{curl};\Omega)\\ &= \{\boldsymbol{u} \in \mathcal{H}(\mathrm{curl};\Omega) : \langle \mathrm{curl}\,\boldsymbol{u}, \boldsymbol{v}\rangle_\Omega = \langle \boldsymbol{u}, \mathrm{curl}\,\boldsymbol{v}\rangle_\Omega \quad \forall \boldsymbol{v} \in \mathcal{H}(\mathrm{curl};\Omega)\}.\end{aligned}$$

By virtue of Green's identity for the rotational (see e.g. [20, Theorem I.2.11]), if Ω is Lipschitz in the sense of [1, Notation 2.1], then $\boldsymbol{u} \in \mathcal{H}_0(\mathrm{curl};\Omega)$ if and only if $\boldsymbol{u} \in \mathcal{H}(\mathrm{curl};\Omega)$ and $\boldsymbol{u} \times \mathbf{n} = \mathbf{0}$ on $\partial\Omega$.

Let

$$\mathcal{M}_1 = \begin{pmatrix} 0 & i\mathcal{R}_{\max} \\ -i\mathcal{R}_{\min} & 0 \end{pmatrix}$$

on the domain

$$\mathrm{D}(\mathcal{M}_1) = \mathrm{D}(\mathcal{R}_{\min}) \times \mathrm{D}(\mathcal{R}_{\max}) \subset [L^2(\Omega)]^6. \tag{41}$$

As $\mathcal{R}_{\max}$ and $\mathcal{R}_{\min}$ are mutually adjoints, $\mathcal{M}_1 : \mathrm{D}(\mathcal{M}_1) \longrightarrow [L^2(\Omega)]^6$ is a self-adjoint operator, [7, Lemma 1.2]. Now, write the system (1) as

$$\begin{pmatrix} \epsilon^{-1/2} & 0 \\ 0 & \mu^{-1/2} \end{pmatrix} \begin{pmatrix} 0 & i\,\mathrm{curl} \\ -i\,\mathrm{curl} & 0 \end{pmatrix} \begin{pmatrix} \epsilon^{-1/2} & 0 \\ 0 & \mu^{-1/2} \end{pmatrix} \begin{pmatrix} \tilde{\boldsymbol{E}} \\ \tilde{\boldsymbol{H}} \end{pmatrix} = \omega \begin{pmatrix} \tilde{\boldsymbol{E}} \\ \tilde{\boldsymbol{H}} \end{pmatrix}$$

with unknowns $(\tilde{\boldsymbol{E}}, \tilde{\boldsymbol{H}}) = (\epsilon^{1/2}\boldsymbol{E}, \mu^{1/2}\boldsymbol{H})$. Let

$$\mathcal{P} = \mathrm{diag}[\epsilon^{1/2} I_{3\times3}, \mu^{1/2} I_{3\times3}]$$

be the self-adjoint operator acting on $[L^2(\Omega)]^6$ given by the permittivity and permeability. The constraint (2) ensures that $\mathcal{P}$ is bounded and invertible with

$$\mathcal{P}^{-1} = \mathrm{diag}[\epsilon^{-1/2} I_{3\times3}, \mu^{-1/2} I_{3\times3}].$$

Define $\mathcal{M} = \mathcal{P}^{-1}\mathcal{M}_1\mathcal{P}^{-1}$ on the dense domain $\mathrm{D}(\mathcal{M}) = \mathcal{P}\,(\mathrm{D}(\mathcal{M}_1))$. Then $\mathcal{M}$ is a self-adjoint operator and its eigenvalues correspond exactly with the angular frequencies in (1). Every $(\tilde{\boldsymbol{E}}, \tilde{\boldsymbol{H}})^t \neq 0$ eigenfunction of $\mathcal{M}$ will produce a corresponding field phasor

$$(\boldsymbol{E}, \boldsymbol{H})^t = \mathcal{P}^{-1}(\tilde{\boldsymbol{E}}, \tilde{\boldsymbol{H}})^t \neq 0$$

satisfying (1) and vice-versa.

Assumption 4. *Here and everywhere below we assume that the non-zero spectrum of* $\mathcal{M}_1$ *is purely discrete and it does not accumulate at* $\omega = 0$. *This hypothesis can be verified whenever* Ω *is a polyhedron with Lipschitz boundary for example, see* [22, Corollary 3.49] *and* [7, Lemma 1.3]. *A more systematic analysis of the properties of* $\mathcal{M}$ *on more general regions* Ω *will be carried out elsewhere* [3].

Suppose that Ω is a polyhedron. We may consider applying the framework of Section 3 for $A = \mathcal{M}$ as follows. Fix $\{\mathcal{T}_h\}_{h>0}$ a family of shape-regular triangulations of $\overline{\Omega}$ [19], where the elements $K \in \mathcal{T}_h$ are simplexes with diameter h_K such that $h = \max_{K\in\mathcal{T}_h} h_K$. For $r \geq 1$, let

$$\begin{aligned}\mathbf{V}_h^r &= \{\boldsymbol{v}_h \in [C^0(\overline{\Omega})]^3 : \boldsymbol{v}_h|_K \in [\mathbb{P}_r(K)]^3 \ \forall K \in \mathcal{T}_h\}\\ \mathbf{V}_{h,0}^r &= \{\boldsymbol{v}_h \in \mathbf{V}_h^r : \boldsymbol{v}_h \times \mathbf{n} = \mathbf{0} \text{ on } \partial\Omega\}.\end{aligned}$$

Then

$$\mathcal{L}_h = \mathbf{V}_{h,0}^r \times \mathbf{V}_h^r \subset \mathrm{D}(\mathcal{M}_1) \tag{42}$$

and

$$\tilde{\mathcal{L}}_h = \mathcal{P}\mathcal{L}_h \subset \mathrm{D}(\mathcal{M}) \tag{43}$$

are finite element spaces of isotropic and anisotropic media, respectively. Recall that $\mathcal{P}$ is bounded and invertible as a consequence of (2).

By virtue of [22, Theorem 3.26] and the fact that $\mathcal{H}_0(\mathrm{curl};\Omega)$ is the closure in the curl norm of $C_0^\infty(\Omega)$, the family $\mathcal{L}_h$ is dense in $\mathrm{D}(\mathcal{M}_1)$. That is, for any $(\boldsymbol{F},\boldsymbol{G})^t \in \mathrm{D}(\mathcal{M}_1)$ there exists a sequence $\{(\boldsymbol{F}_h,\boldsymbol{G}_h)^t\}_{h>0}$ such that $(\boldsymbol{F}_h,\boldsymbol{G}_h)^t \in \mathcal{L}_h$ and

$$\lim_{h\to 0}\Big(\|\boldsymbol{F}-\boldsymbol{F}_h\|_{\mathrm{curl},\Omega} + \|\boldsymbol{G}-\boldsymbol{G}_h\|_{\mathrm{curl},\Omega}\Big) = 0\,. \tag{44}$$

In turns, this implies that for all $(\tilde{\boldsymbol{F}},\tilde{\boldsymbol{G}})^t = \mathcal{P}(\boldsymbol{F},\boldsymbol{G})^t \in \mathrm{D}(\mathcal{M})$, there exists a family $\{(\tilde{\boldsymbol{F}}_h,\tilde{\boldsymbol{G}}_h)^t\}_{h>0} \subset \tilde{\mathcal{L}}_h$ such that

$$\lim_{h\to 0}\left(\left\|\mathcal{M}\begin{pmatrix}\tilde{\boldsymbol{F}}-\tilde{\boldsymbol{F}}_h\\ \tilde{\boldsymbol{G}}-\tilde{\boldsymbol{G}}_h\end{pmatrix}\right\|_{0,\Omega} + \left\|\begin{pmatrix}\tilde{\boldsymbol{F}}-\tilde{\boldsymbol{F}}_h\\ \tilde{\boldsymbol{G}}-\tilde{\boldsymbol{G}}_h\end{pmatrix}\right\|_{0,\Omega}\right) = 0\,. \tag{45}$$

Let $\mathcal{I}_h$ denote the Lagrange interpolator on $\mathcal{L}_h$, [19]. Under the condition of regularity $(\boldsymbol{F},\boldsymbol{G})^t \in \mathcal{H}^{r+1}(\Omega)^6$,

$$\begin{aligned}\|\boldsymbol{F}-\mathcal{I}_h(\boldsymbol{F})\|_{\mathrm{curl},\Omega} &+ \|\boldsymbol{G}-\mathcal{I}_h(\boldsymbol{G})\|_{\mathrm{curl},\Omega}\\ &\leq C_r h^r\left(\|\boldsymbol{F}\|_{r+1,\Omega} + \|\boldsymbol{G}\|_{r+1,\Omega}\right)\end{aligned} \tag{46}$$

for a suitable constant $C_r > 0$. Hence, there also exists a constant $\tilde{C}_r(\epsilon,\mu) > 0$, such that

$$\left\| \mathcal{M} \begin{pmatrix} \tilde{\boldsymbol{F}} - \mathcal{I}_h(\tilde{\boldsymbol{F}}) \\ \tilde{\boldsymbol{G}} - \mathcal{I}_h(\tilde{\boldsymbol{G}}) \end{pmatrix} \right\|_{0,\Omega} + \left\| \begin{pmatrix} \tilde{\boldsymbol{F}} - \tilde{\boldsymbol{F}}_h \\ \tilde{\boldsymbol{G}} - \tilde{\boldsymbol{G}}_h \end{pmatrix} \right\|_{0,\Omega} \leq \tilde{C}_r(\epsilon,\mu) h^r. \tag{47}$$

As a consequence of Theorem 12 and Corollary 13, the estimates (45) and (47) lead to precise convergence and error estimates for the method of Section 3 in the case $A = \mathcal{M}$ and $\mathcal{L} = \tilde{\mathcal{L}}_h$. We summarize the corresponding statements in next two theorems.

Theorem 14. *Let $J \subset \mathbb{R}$ be a bounded open segment such that $0 \notin J$. Let $t \in J$. Let $\tau_{j,h}^+(t)$ and $\tau_{j,h}^-(t)$ be the corresponding positive and negative eigenvalues of $(\mathrm{Z}_t^{\mathcal{L}})$ for $\mathcal{L} = \tilde{\mathcal{L}}_h$. Then, for every j such that $\nu_j^{\pm}(t) \in J$,*

$$\lim_{h\to 0} \left| \left(t + \frac{1}{\tau_{j,h}^{\pm}(t)} \right) - \nu_j^{\pm}(t) \right| = 0.$$

Moreover, if in addition $\mathcal{P}^{-1}\mathcal{E}_J(\mathcal{M}) \subseteq \mathcal{H}^{r+1}(\Omega)^6$, then there exist constants $C_t^{\pm} > 0$ such that

$$\left| \left(t + \frac{1}{\tau_{j,h}^{\pm}(t)} \right) - \nu_j^{\pm}(t) \right| \leq C_t^{\pm} h^{2r} \tag{48}$$

for h sufficiently small and j such that $\nu_j^{\pm}(t) \in J$.

For $(\tilde{\boldsymbol{F}}, \tilde{\boldsymbol{G}})^t \in \mathrm{D}(\mathcal{M})$ and a subspace $\mathcal{E} \subseteq \mathrm{D}(\mathcal{M})$, let

$$\mathrm{dist}_{\mathcal{M}}[(\tilde{\boldsymbol{F}}, \tilde{\boldsymbol{G}}), \mathcal{E}] = \inf_{(\boldsymbol{X},\boldsymbol{Y})^t \in \mathcal{E}} \left[\left\| \mathcal{M} \begin{pmatrix} \tilde{\boldsymbol{F}} - \boldsymbol{X} \\ \tilde{\boldsymbol{G}} - \boldsymbol{Y} \end{pmatrix} \right\|_{0,\Omega} + \left\| \begin{pmatrix} \tilde{\boldsymbol{F}} - \boldsymbol{X} \\ \tilde{\boldsymbol{G}} - \boldsymbol{Y} \end{pmatrix} \right\|_{0,\Omega} \right].$$

Theorem 15. *Assume the same hypotheses as in Theorem 14. Let*

$$(\tilde{\boldsymbol{E}}_{j,h}^{\pm}(t), \tilde{\boldsymbol{H}}_{j,h}^{\pm}(t))^t \in \tilde{\mathcal{L}}_h$$

be the corresponding normalized eigenvectors of the eigenvalue problem $(\mathrm{Z}_t^{\tilde{\mathcal{L}}_h})$. Then, for every j such that $\nu_j^{\pm}(t) \in J$,

$$\lim_{h\to 0} \mathrm{dist}_{\mathcal{M}}[(\tilde{\boldsymbol{E}}_{j,h}^{\pm}(t), \tilde{\boldsymbol{H}}_{j,h}^{\pm}(t)), \mathcal{E}_{\{\nu_j^-(t),\nu_j^+(t)\}}(\mathcal{M})] = 0.$$

Moreover, if in addition $\mathcal{P}^{-1}\mathcal{E}_J(\mathcal{M}) \subseteq \mathcal{H}^{r+1}(\Omega)^6$, then there exist constants $C_t^{\pm} > 0$ such that

$$\mathrm{dist}_{\mathcal{M}}[(\tilde{\boldsymbol{E}}_{j,h}^{\pm}(t), \tilde{\boldsymbol{H}}_{j,h}^{\pm}(t)), \mathcal{E}_{\{\nu_j^-(t),\nu_j^+(t)\}}(\mathcal{M})] \leq C_t^{\pm} h^r$$

for h sufficiently small and j such that $\nu_j^{\pm}(t) \in J$.

Theorems 14 and 15 have various consequences for the numerical calculation of the eigenfrequencies associated to the resonant cavity problem which are worth highlighting. Note that convergence and absence of spectral pollution are guaranteed, despite of the fact that $\mathcal{L}_h$ are spaces of nodal finite elements with no particular mesh structure. These convergence properties are constrained to extremely mild assumptions on the coefficients ϵ and μ. Moreover, the order of approximation achieved is optimal in the context of the finite elements chosen.

Our analysis above relies on the regularity of the eigenspaces associated to the interval J only. This opens the possibility of approximating eigenvalues associated to regular eigenfunctions with high accuracy, if a priori information about their location is at hand. Refer to the numerical results below for concrete examples on this matter.

The discussion above was restricted finite elements of Lagrange type with the sole purpose of illustrating a concrete implementation. Analogous approximation results hold true for other choices of trial subspaces (made out of standard finite elements or otherwise) as long as they form a dense family in $\mathrm{D}(\mathcal{M})$. A control on the order of convergence will be achieved in a similar fashion, as long as interpolation estimates are available.

6. A certified numerical strategy

We now describe a certified numerical scheme for computing the eigenvalues of $\mathcal{M}$ which is based on Corollary 7. In an asymptotic regime, as specified below, this scheme provides small intervals which are guaranteed to contain spectral points. Its convergence will be deduced from Theorem 14.

Let $t > 0$. Let $\mathcal{L} = \tilde{\mathcal{L}}_h$ as in (42)-(43) satisfy (14). Bounds for the eigenvalues of $\mathcal{M}$ in a vicinity of t, can be found from (24). The inverse residuals $\tau_j^{\mp}(t)$ in (24) can be computed by solving $(\mathrm{Z}_t^{\tilde{\mathcal{L}}_h})$ as follows. Let $\{b_1, \ldots, b_{n(h)}\}$ be a basis of $\mathcal{L}_h$. Let $B_t,\, K_t \in \mathbb{C}^{n(h)\times n(h)}$ be determined by

$$[B_t]_{jk} = \big\langle(\mathcal{P}^{-1}\mathcal{M}_1 - t\mathcal{P})b_j, (\mathcal{P}^{-1}\mathcal{M}_1 - t\mathcal{P})b_k\big\rangle \qquad \text{and}$$
$$[K_t]_{jk} = \big\langle(\mathcal{P}^{-1}\mathcal{M}_1 - t\mathcal{P})b_j, \mathcal{P}b_k\big\rangle.$$

Then $\tau_j^{\mp}(t) = \eta_{\mp}^{-1}$ where $\eta_{\mp}$ is the negative$(-)$/positive$(+)$ eigenvalue of the pencil $B_t - \eta K_t$ which is in the jth place among those closer to 0.

Denote by $0 < t_{\mathrm{up}} < t_{\mathrm{low}}$ the corresponding position t set for computing upper and lower bounds by means of $\tau_j^-(t_{\mathrm{low}})$ and $\tau_j^+(t_{\mathrm{up}})$, respectively. Since $\mathcal{M}$ is strongly indefinite and $\tilde{\mathcal{L}}_h$ are dense in the graph norm of $\mathrm{D}(\mathcal{M})$ for suitable sub-families of mesh, we can always assume that the trial spaces are chosen such that

$$\min_{u\in\tilde{\mathcal{L}}_h} \frac{\langle \mathcal{M}u,u\rangle}{\langle u,u\rangle} < t_{\mathrm{up}} \qquad \text{and} \qquad t_{\mathrm{low}} < \max_{u\in\tilde{\mathcal{L}}_h} \frac{\langle \mathcal{M}u,u\rangle}{\langle u,u\rangle}. \tag{49}$$

Recall the condition (14).

The following procedure aims at computing intervals of enclosure for the eigenvalues of $\mathcal{M}$ which lie in the segment $(t_{\mathrm{up}}, t_{\mathrm{low}})$ for a prescribed tolerance set by the parameter $\delta > 0$. According to Lemma 16 below, these intervals will be certified in the regime $\delta \to 0$.

Procedure 1.

Input.
- *Initial* $t_{\mathrm{up}} > 0$.
- *Initial* $t_{\mathrm{low}} > t_{\mathrm{up}}$ *such that* $t_{\mathrm{low}} - t_{\mathrm{up}}$ *is fairly large.*
- *A sub-family* $\mathcal{F}$ *of finite element spaces* $\tilde{\mathcal{L}}_h$ *as in* (42)-(43), *dense in the graph norm of* $\mathrm{D}(\mathcal{M})$ *as* $h \to 0$.
- *A tolerance* $\delta > 0$ *fairly small compared with* $t_{\mathrm{low}} - t_{\mathrm{up}}$.

Output.
- *A prediction* $\tilde{m}(\delta) \in \mathbb{N}$ *of* $\operatorname{Tr} \mathbb{1}_{(t_{\mathrm{up}},t_{\mathrm{low}})}(\mathcal{M})$.
- *Predictions* $\omega_j^{\pm}(\delta)$ *of the endpoints of enclosures for the eigenvalues in* $\sigma(\mathcal{M})\cap(t_{\mathrm{up}}, t_{\mathrm{low}})$, *such that* $0 < \omega_j^+(\delta)-\omega_j^-(\delta) < \delta$ *for* $j = 1, \dots, \tilde{m}(\delta)$.

Steps.

a) Set initial $\mathcal{L} = \tilde{\mathcal{L}}_h \in \mathcal{F}$.

b) While

$$\omega^+_{j,h} - \omega^-_{j,h} \geq \delta \text{ or } \omega^-_{j,h} > \omega^+_{j,h} \text{ for some } j = 1, \dots, \tilde{m},$$

do c) - e).

c) Compute

$$\omega^+_{j,h} = t_{\mathrm{up}} + \frac{1}{\tau_j^+(t_{\mathrm{up}})} \qquad \text{for} \qquad j = 1, \dots, \tilde{m}_{\mathrm{up}}$$

where $\tilde{m}_{\mathrm{up}}$ *is such that all* $\omega^+_{j,h} < t_{\mathrm{low}}$ *and*

$$t_{\mathrm{up}} + \frac{1}{\tau^+_{\tilde{m}_{\mathrm{up}}+1}(t_{\mathrm{up}})} \geq t_{\mathrm{low}}.$$

d) Compute

$$\omega^-_{\tilde{m}_{\mathrm{low}}-k+1,h} = t_{\mathrm{low}} + \frac{1}{\tau_k^-(t_{\mathrm{low}})} \qquad \text{for} \qquad k = 1, \dots, \tilde{m}_{\mathrm{low}}$$

where $\tilde{m}_{\rm low}$ *is such that all* $\omega^-_{\tilde{m}_{\rm low}-k+1,h} > t_{\rm up}$ *and*

$$t_{\rm low} + \frac{1}{\tau^-_{\tilde{m}_{\rm low}+1}(t_{\rm low})} \leq t_{\rm up}.$$

e) If $\tilde{m}_{\rm low} \neq \tilde{m}_{\rm up}$*, decrease* h*, set new* $\mathcal{L} = \tilde{\mathcal{L}}_h \in \mathcal{F}$ *and go back to c). Otherwise set* $\tilde{m} = \tilde{m}_{\rm low} = \tilde{m}_{\rm up}$*, decrease* h*, set new* $\mathcal{L} = \tilde{\mathcal{L}}_h \in \mathcal{F}$ *and continue from b).*

f) Exit with $\tilde{m}(\delta) = \tilde{m}$ *and* $\omega^\pm_j(\delta) = \omega^\pm_{j,h}$ *for* $j = 1, \ldots, \tilde{m}$*.*

Let

$$(t_{\rm up}, t_{\rm low}) \cap \sigma(\mathcal{M}) = \{\omega_{k+1}, \ldots, \omega_{k+m}\}$$

where

$$m = \operatorname{Tr} \mathbb{1}_{(t_{\rm up}, t_{\rm low})}(\mathcal{M}) \qquad \text{and} \qquad k \geq 0.$$

Observe that, a priori, an interval (ω^-_j, ω^+_j) obtained as the output of Procedure 1 is not guaranteed to have a non-empty intersection with the spectrum of $\mathcal{M}$ or in fact include precisely the eigenvalue ω_{k+j}. However, as it is established by the following lemma, the latter is certainly true for δ small enough.

Lemma 16. *There exist* $t^0 > 0$ *and* $\delta_0 > 0$*, ensuring all the next items for all* $t_{\rm low} \geq t^0$ *and* $\delta < \delta_0$*.*

a) The conditional loop in Procedure 1 always exits in the regime $h \to 0$*.*

b) $m(\delta) = m$*.*

c) $\omega^-_j(\delta) \leq \omega_{k+j} \leq \omega^+_j(\delta)$ *for all* $j = 1, \ldots, n$*.*

Proof. Since $\nu^+_j(t_{\rm up}) = \omega_{k+j} = \nu^-_{n-j+1}(t_{\rm low})$ for all $j = 1, \ldots, n$, Theorem 14 alongside with the assumption on $\mathcal{F}$, confirms the existence of $\omega^\pm_{j,h}$ in Procedure 1-*c)* and *d)*, for all $j = 1, \ldots, n$ whenever h is small enough. Moreover

$$\omega^+_{j,h} \downarrow \omega_{k+j} \qquad \text{and} \qquad \omega^-_{j,h} \uparrow \omega_{k+j} \qquad \text{as } h \to 0.$$

This ensures the validity of the lemma. □

If the eigenfunctions of $\mathcal{M}$ lie in $\mathcal{H}^{r+1}(\Omega)^6$, where r is the degree of the polynomials in (42), then

$$\omega^+_{j,h} - \omega^-_{j,h} = O(h^{2r}).$$

This means that the exit rate of the conditional loop in Procedure 1 is also $O(h^{2r})$ as $h \to 0$.

A close examination of the constants involved in the proof of Theorem 14, indicates that they are of order $|t - \nu^\pm_1(t)|^{-1}$. See Theorem 10 and Corollary 11. Table 1 and other various numerical experiments not included in Section 7, strongly suggest that the accuracy improves significantly, as $t_{\rm up} \downarrow \nu_1^-(t_{\rm up})$ and $t_{\rm low} \uparrow \nu_1^+(t_{\rm low})$.

7. Benchmark examples

We now illustrate the practical applicability of the ideas discussed above by means of several examples. Two canonical references for benchmarks on the Maxwell eigenvalue problem are [15] and [9]. We validate some of the numerical bounds shown below against these benchmarks. All the experiments presented are performed for $\epsilon = \mu = 1$ and some of them consider the so-called two-dimensional Maxwell problem.

If the domain Ω has a cylindrical symmetry, say $\Omega = \tilde{\Omega} \times (0, \pi)$ for $\tilde{\Omega} \subset \mathbb{R}^2$ open and sufficiently regular, then (1) decouples. Indeed, by separation of variables, a non-zero ω is an eigenvalue of $\mathcal{M}_1$ if and only if either $\omega^2 = \lambda^2$ or $\omega^2 = \nu^2 + \rho^2$, where λ^2 is a Dirichlet eigenvalue of the Laplacian in $\tilde{\Omega}$, ν^2 is a non-zero Neumann eigenvalue of the Laplacian in $\tilde{\Omega}$ and $\rho \in \mathbb{N}$. In turns the Neumann problem can be re-written as

$$\begin{cases} \operatorname{curl} \boldsymbol{E} = i\mu H & \text{in } \tilde{\Omega} \\ \operatorname{curl} H = -i\omega \boldsymbol{E} & \text{in } \tilde{\Omega} \\ \boldsymbol{E} \cdot \mathbf{t} = \mathbf{0} & \text{on } \partial\tilde{\Omega}\,, \end{cases} \tag{50}$$

for non-zero $(\boldsymbol{E}, H)^t \in L^2(\tilde{\Omega})^3$ and $\nu = \omega \in \mathbb{R}$, where

$$\boldsymbol{E} = \begin{pmatrix} E_1 \\ E_2 \end{pmatrix}, \qquad \operatorname{curl} \boldsymbol{E} = \partial_x E_2 - \partial_y E_1, \qquad \operatorname{curl} H = \begin{pmatrix} \partial_y H \\ -\partial_x H \end{pmatrix}$$

and $\mathbf{t}$ is the unit tangent to $\partial\tilde{\Omega}$. This two-dimensional Maxwell problem suffers from all the complications concerning spectral pollution, as its three-dimensional counterpart.

We denote by $\tilde{\mathcal{M}}$ the self-adjoint operator associated to (50). This operator can be employed for numerical tests which can then be validated against numerical calculations for the original Neumann Laplacian via the Galerkin method, [15]. Indeed, note that the latter is a semi-definite operator with a compact resolvent, so it does not exhibit spectral pollution.

The ideas developed in Section 5 for the operator $\mathcal{M}$ have analogues for $\tilde{\mathcal{M}}$. In the lower-dimensional examples presented below, we have chosen the finite element spaces on a corresponding triangulation $\mathcal{T}_h$ of

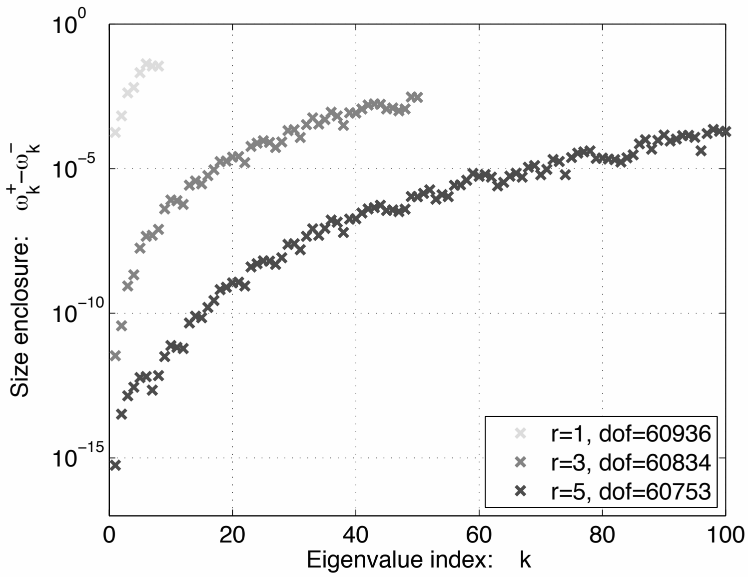


FIGURE 1. Semi-log graph associated to $\Omega_{\rm sqr}$. Vertical axis: $\omega_k^+ - \omega_k^-$. Horizontal axis: eigenvalue index k (not counting multiplicity). Here we use elements of order $r = 1, 3, 5$ on unstructured uniform meshes rendering roughly the same number of degrees of freedom.

$\tilde{\Omega}$ as

$$\begin{aligned} \mathbf{V}_h^{r,k} &= \{\boldsymbol{v}_h \in [C^0(\overline{\tilde{\Omega}})]^k : \boldsymbol{v}_h|_K \in [\mathbb{P}_r(K)]^k \ \forall K \in \mathcal{T}_h\} \qquad (k=1,2) \\ \mathbf{V}_{h,0}^{r,2} &= \{\boldsymbol{v}_h \in \mathbf{V}_h^{r,2} : \boldsymbol{v}_h \times \mathbf{n} = 0 \text{ on } \partial\tilde{\Omega}\} \qquad \text{and} \\ \mathcal{L}_h &= \mathbf{V}_{h,0}^{r,2} \times \mathbf{V}_h^{r,1}. \end{aligned}$$

This ensures that $\mathcal{L}_h \subset \mathrm{D}(\tilde{\mathcal{M}})$.

7.1. **Convex domains.** The eigenfunctions of (1) or (50) are regular in the interior of a convex domain. This leads to an improvement in accuracy as a consequence of (48). In this, the best possible case scenario, the Zimmermann-Mertins method for the resonant cavity problem achieves an optimal order of convergence in the context of the finite element method.

Enclosures on a square. Let $\tilde{\Omega} = \Omega_{\rm sqr} = (0,\pi)^2 \subset \mathbb{R}^2$. The eigenvalues of $\tilde{\mathcal{M}}$ are $\omega = \pm\sqrt{l^2+m^2}$ for $l, m \in \mathbb{N} \cup \{0\}$. In order to estimate $\omega_k^{\pm}$ numerically, we have picked

$$t_{\rm up} = \frac{1}{4}\omega_{k-1} + \frac{3}{4}\omega_k \qquad \text{and} \qquad t_{\rm low} = \frac{3}{4}\omega_k + \frac{1}{4}\omega_{k+1}$$

to machine precision. Here and below we substitute from the notation in previous sections the index j for eigenvalues by an index k, in order to highlight the fact that we do not always count multiplicities.

In our first experiment we have computed the width of the enclosure, $\omega_k^+ - \omega_k^-$, for $k = 1, \ldots, 100$ and $r = 1, 3, 5$. We have chosen $h = h(r)$ such that the subspaces $\mathcal{L}_h$ have roughly the same dimension ≈ 61K. Figure 1 shows the outcomes of this experiment. We have excluded enclosures with size above 10^{-1}. As it is natural to expect, for a fixed subspace $\mathcal{L}_h$, the accuracy deteriorates as the eigenvalue counting number j increases: high energy eigenfunctions have more oscillations so their approximation is more challenging. The accuracy increases with the polynomial order. The first 100 eigenvalues are approximated fairly accurately (note that $\omega_{100} = \sqrt{261}$) with polynomial order $r = 5$.

Convergence on a cube. We now consider $\Omega = \Omega_{\rm cbe} = (0,\pi)^3 \subset \mathbb{R}^3$. The non-zero eigenvalues are now $\omega = \pm\sqrt{l^2+m^2+n^2}$. The corresponding eigenfunctions are

$$\boldsymbol{E}(x,y,z) = \begin{pmatrix} \alpha_1 \cos(lx)\sin(my)\sin(nz) \\ \alpha_2 \sin(lx)\cos(my)\sin(nz) \\ \alpha_3 \sin(lx)\sin(my)\cos(nz) \end{pmatrix} \qquad \forall \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{pmatrix} \cdot \begin{pmatrix} l \\ m \\ n \end{pmatrix} = 0.$$

Here $\{l, m, n\} \subset \mathbb{N} \cup \{0\}$ and not two indices are allowed to vanish simultaneously. The vector $\underline{\alpha}$ determines the multiplicity of the eigenvalue for a given triplet (l, m, n). That is, for example, $\omega = \sqrt{2}$ (the first positive eigenvalue) has multiplicity 3 corresponding to indices $\{(1,1,0), (0,1,1), (1,0,1)\}$ each one of them contributing to one of the dimensions of the eigenspace. However, $\omega = \sqrt{3}$ (the second positive eigenvalue) corresponding to index $\{(1,1,1)\}$ has multiplicity 2 determined by $\underline{\alpha}$ on a plane.

In Figure 2 we have depicted the decrease in the enclosure width for the computation of the eigenvalue $\omega_2 = \sqrt{3}$ for Lagrange elements of order $r = 1, 2, 3$. We have chosen a sequence of unstructured tetrahedral mesh. The computed values for the slopes of the straight lines indicate that the enclosures obey the estimate

$$|\omega_j^{\pm} - \omega_j| \leq c h^{2r}. \tag{51}$$

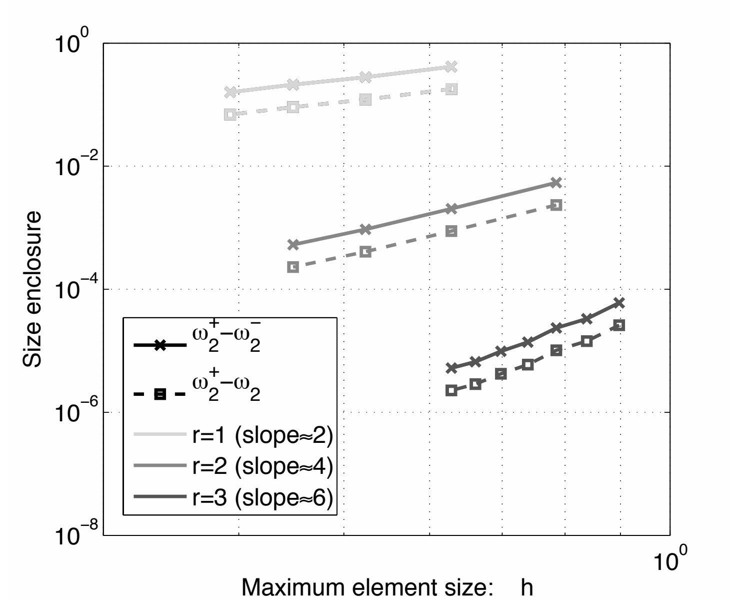


FIGURE 2. Log-log graph associated to $\Omega_{\rm cbe}$ and $\omega_2 = \sqrt{3}$. Vertical axis: enclosure width. Horizontal axis: maximum element size h. Here we have chosen Lagrange elements of order $r = 1, 2, 3$ on a sequence of unstructured meshes.

Therefore the conclusion (48) of Theorem 14 will be sharp. Note that in the picture, we have considered both the exact residual and the length of the enclosure.

The slashed cube. We now assume that

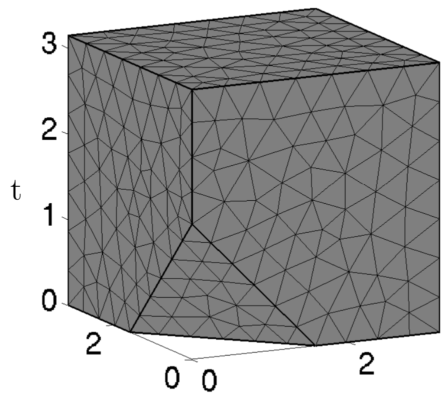
$$\Omega = \Omega_{\rm sla} = (0, \pi)^3 \setminus T \subset \mathbb{R}^3.$$

Here T is the closed tetrahedron with vertices given by

$$(0, 0, 0),\ (\pi/2, 0, 0),\ (0, \pi/2, 0) \quad \text{and} \quad (0, 0, \pi/2).$$

This domain does not have symmetries allowing a reduction into two-dimensions. However, as $\Omega_{\rm sla}$ is fairly close to $\Omega_{\rm cbe}$, we should expect that the structure of the spectrum in the two cases is reminiscent of one another.

In our first experiment on this region, we compute benchmark eigenvalue enclosures for (1). The table to the right of Figure 3 shows the outcomes of implementing the Procedure 1. We have run an algorithm based on this procedure for three fixed choices of $t_{\rm up}$ and $t_{\rm low}$ (third and fourth columns) with $\delta = 10^{-2}$. We have picked the family of mesh so that no more than five iterations were required to achieve the



k	$(\omega_k)^+_-$	t_{up} (l)	t_{low} (l)
1	1.412^{236}_{000}	0.5 (1)	1.6 (3)
2	1.430^{672}_{560}	0.5 (2)	1.6 (2)
3	1.430^{673}_{577}	0.5 (3)	1.6 (1)
4	1.755^{308}_{043}	1.5 (1)	2.1 (2)
5	1.755^{329}_{063}	1.5 (2)	2.1 (1)
6	2.22^{200}_{053}	1.8 (1)	2.6 (5)
7	2.237^{667}_{434}	1.8 (2)	2.6 (4)
8	2.237^{684}_{459}	1.8 (3)	2.6 (3)
9	2.239^{533}_{387}	1.8 (4)	2.6 (2)
10	2.270^{778}_{558}	1.8 (5)	2.6 (1)

FIGURE 3. Benchmark spectral approximation for Ω_{sla}. In the table we compute interval of enclosure for the first 10 eigenvalues of (1). In order to obtain this calculation we have employed Procedure 1. The trial spaces are made of Lagrange elements of order $r = 3$. The final mesh is the one shown on the right side. Total number of DOF=117102.

needed accuracy. The parameter l in this table counts the number of eigenvalues to the right of t_{up} or to the left of t_{low}, respectively.

In this experiment we have chosen trial spaces made out of Lagrange elements of order $r = 3$. All the final eigenvalue enclosures have a length of at most 2×10^{-3}. The mesh used in the last iteration is depicted on the left of Figure 3.

From the table it seems clear that there is a cluster of eigenvalues at the bottom of the positive spectrum near $\sqrt{2}$. The latter is the first positive eigenvalue of Ω_{cbe} which is of multiplicity 3. It appears that this eigenvalue splits into a single eigenvalue at the bottom of the spectrum and a seemingly double eigenvalue slightly above it. Another cluster occurs at ω_4 and ω_5 with strong indication that this is a double eigenvalue. This pair is near $\sqrt{3}$, the second eigenvalue of Ω_{cbe} which is indeed double. The next eigenvalues for the cube are 2 and $\sqrt{5}$ with total multiplicity 5. It is natural to conjecture that ω_j for $j = 5, \ldots, 10$ are perturbations of these eigenvalues, but the data shown in the table is inconclusive.

For our next experiment on this region, we have estimated numerically the electromagnetic fields corresponding to index up to 6 from the table in Figure 3. The purpose of the experiment is to set benchmarks for the eigenfunctions on Ω_{sla} and simultaneously illustrate Theorem 15. In Figure 4 we depict the density of electric and magnetic fields, $|\boldsymbol{E}|$

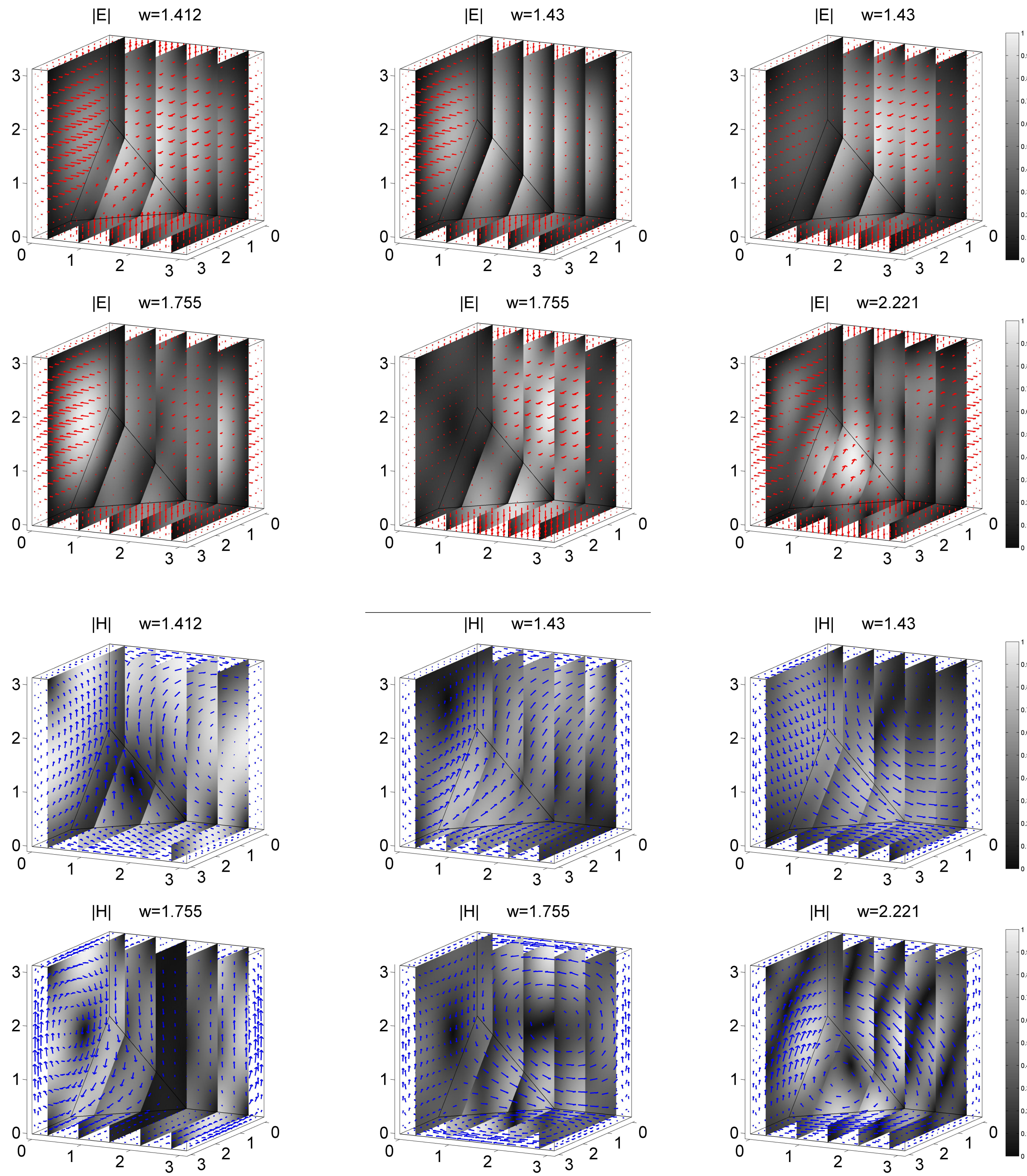


FIGURE 4. The first six eigenfunctions on Ω_{sla} for the first six positive eigenvalues. Densities $|\boldsymbol{E}|$ (top) and $|\boldsymbol{H}|$ (bottom). Corresponding arrow fields $\boldsymbol{E}$ (red) and $\boldsymbol{H}$ (blue) on $\partial\Omega_{\text{sla}}$.

and $|\boldsymbol{H}|$ both normalized to having maximum equal to 1. We also show arrows pointing towards the direction of these fields on $\partial\Omega_{\rm sla}$.

The mesh employed for these calculations is the one shown in Figure 3. According to Theorem 15 and the data presented in the table, the shown eigenfunctions should be close to the exact eigenfunctions in the curl norm. We remark that for both experiments on $\Omega_{\rm sla}$ a reasonable accuracy has been achieved even for the fairly coarse mesh depicted in Figure 3.

7.2. **Non-convex domains.** The numerical approximation of the eigenfrequencies and electromagnetic fields in the resonant cavity is known to be challenging when the domain is not convex. The main reason for this is the fact that the electromagnetic field might have a singularity and a low degree of regularity at re-entrant corners. See for example the discussion after [22, Lemma 3.56] and references therein.

In some of the examples of this section we consider a mesh adapted to the geometry of Ω. However, we do not pursue any specialized mesh refinement strategy. We show below that, even in the case where there is poor approximation due to low regularity of the eigenspace, the Procedure 1 provides a stable approxiation to the eigenvalues of (1).

The L-shaped domain. The region $\tilde{\Omega} = \Omega_{\rm L} = (0,\pi)^2 \setminus [0,\pi/2]^2$ is a classical benchmark domain both for the Maxwell and the Helmholtz problems, and it has been extensively examined in the past. Numerical computations for the eigenvalues of $\mathcal{M}$ were reported in [9, Table 5] via an implementation based on a mixed formulation of (50) and the help of edge finite elements. See also [15]. We now show how to achieve accurate enclosures for these eigenvalues with the help of nodal finite elements.

For this next set of experiments we consider unstructured triangulations of the domain, refined around the re-entrant corner. The polynomial order is set to $r = 3$. Figures 5 - 7 summarize our findings.

We produced the table in Figures 5 by implementing Procedure 1 in the same fashion as for the case of $\Omega_{\rm sla}$ discussed previously. For comparison in the second column of this table we have included the benchmark eigenvalue estimations found in [9] and [15]. Note that some of the computed eigenvalues associated to the mixed formulation are lower bounds of the true eigenvalues, and some, like the 9th, are upper bounds. This confirms that the latter approach is in general un-hierarchical as previously suggested in the literature.

j	ω_j from [9] (from [15])	$(\omega_j)^+_-$	$t_{\rm up}$ (l)	$t_{\rm low}$ (l)
1	0.768192684 (0.773334985176)	0.773334^{991}_{694}	0.1 (1)	2.1 (4)
2	1.196779010 (1.19678275574)	1.1967827557^{026}_{761}	0.1 (2)	2.1 (3)
3	1.999784988 (2.00000000000)	$^{2.00000000064}_{1.99999999933}$	1.5 (1)	2.5 (4)
4	1.999784988 (2.00000000000)	$^{2.00000000067}_{1.99999999936}$	1.5 (2)	2.5 (3)
5	2.148306309 (2.14848368266)	2.14848368^{365}_{199}	3.1 (5)	1.5 (3)
6	2.252760528	2.25729^{896}_{776}	1.5 (4)	3.1 (4)
7	2.828075317	2.8284271^{354}_{186}	1.5 (5)	3.7 (4)
8	2.938491109	2.94671^{343}_{112}	1.5 (6)	3.7 (3)
9	3.075901493	3.0758929^{738}_{571}	1.5 (7)	3.7 (2)
10	3.390427701	3.3980^{724}_{676}	1.5 (8)	3.7 (1)

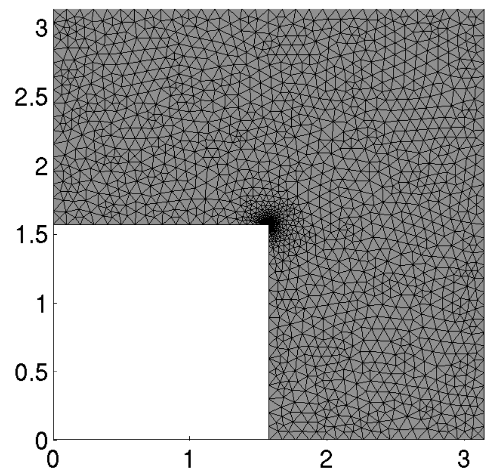


FIGURE 5. Enclosures for the first 10 positive eigenvalues of $\mathcal{M}$ on the region $\Omega_{\rm L}$. The next eigenvalue is above 3.7. Here Procedure 1 has been implemented on Lagrange elements of order 3. The final mesh shown on the right has a number of DOF=56055. The mesh has a maximum element size $h = 0.1$ and has been refined at $(\pi/2, \pi/2)$. For comparison on the second column we include the eigenvalue estimations found in [9] and [15].

From the third column of the table, it is clear that the accuracy depends on the regularity of the corresponding eigenspaces. The eigenfunctions associated to $\omega = 2$ and $\omega = \sqrt{8}$ are found by gluing together corresponding eigenfunctions of (1) on squares of side $\pi/2$. These eigenfunctions are smooth in the interior of $\Omega_{\rm L}$, while those associated to ω_1 and ω_2 are singular at the re-entrant corner. The electric field component of the former is known to be outside $\mathcal{H}^1(\Omega_{\rm L})^2$ while that of the latter is in $\mathcal{H}^1(\Omega_{\rm L})^2$. This explains the significant gain in accuracy in the calculation of ω_2 with respect to the one of ω_1.

Figure 6 depicts in log-log scale residuals versus maximum element size. We have considered here Lagrange elements of order $r = 3$ and $r = 5$. The hierarchy of mesh (not shown) was chosen unstructured, but with an uniform distribution of nodes. Since the eigenfunctions associated to ω_1 and ω_2 have a limited regularity, then there is no

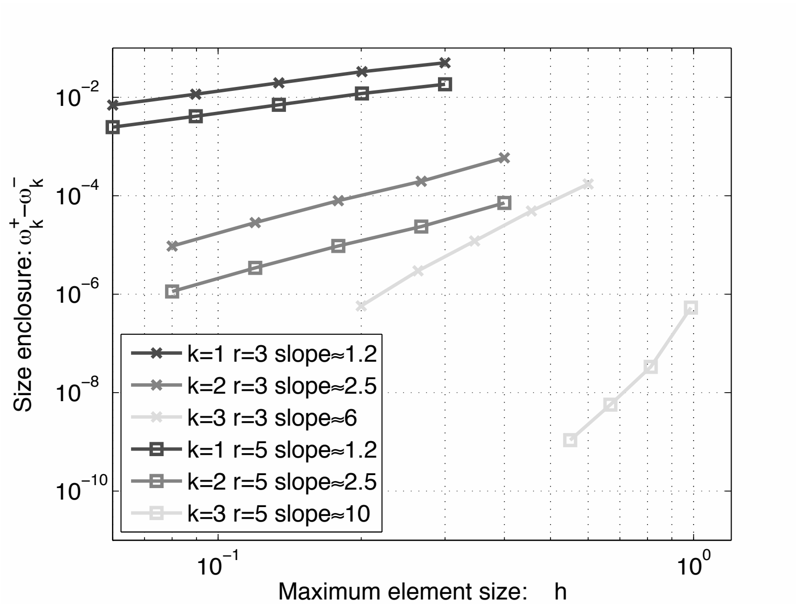


FIGURE 6. Compared order of approximation for different eigenvalues in the region Ω_{L}. The log-log plot shows residual versus maximum element size h for the calculation of enclosures for ω_k where $k = 1, 2, 3$ and $\mathcal{L}$ is generated by Lagrange elements of order $r = 3$ and $r = 5$. Note that $(\boldsymbol{E}, H) \notin \mathcal{H}^s(\Omega_{\mathrm{L}})^3$ for $k = 1$ and $s = 1$, and for $k = 2$ and $s = 1.5$. On the other hand, for $k = 3$ we have $(\boldsymbol{E}, H)$ smooth, as the eigenfunction is also solution of (1) on a square of side $\pi/2$.

noticeable improvement of convergence order as r increases. As the third eigenfunction is smooth, it does obey the estimate (51).

Benchmark approximated eigenfunctions are depicted in Figure 7. The mesh employed to produce these graphs is the one shown on the right of Figure 5. As some of the electric fields have a singularity at $(\pi/2, \pi/2)$ we have normalized each individual plot to a range in the interval $[0, 1]$.

The Fichera domain. In this next experiment we approximate numerically the eigenpairs of (1) associated to the region

$$\Omega = \Omega_{\mathrm{F}} = (0, \pi)^3 \setminus [0, \pi/2]^3 \subset \mathbb{R}^3.$$

Some of the eigenvalues can be obtained by domain decomposition and the corresponding eigenfunctions are regular. For example, eigenfunctions on the cube of side $\pi/2$ can be assembled in the obvious

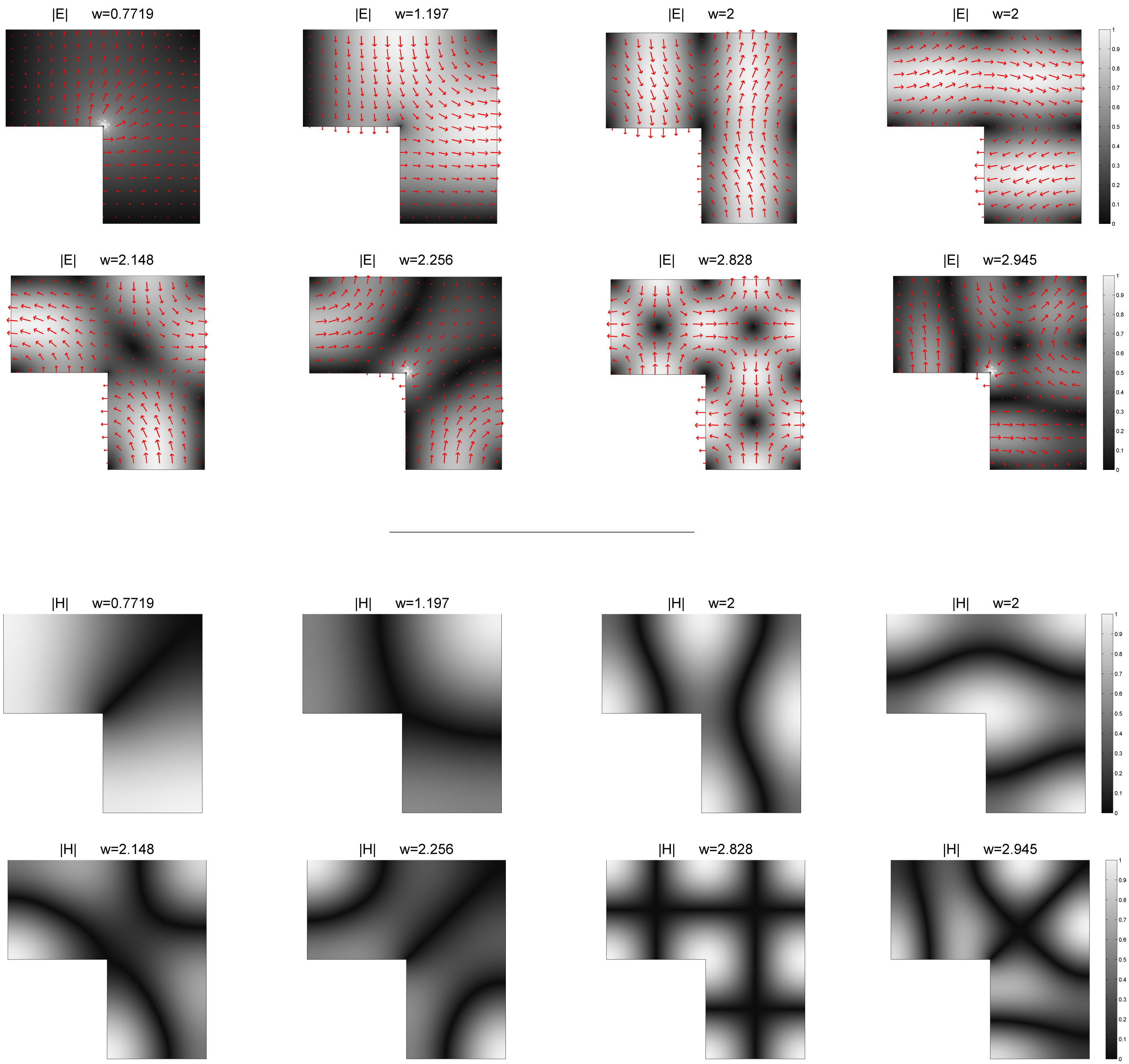
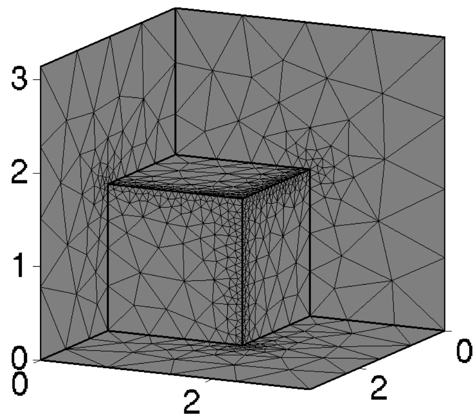


FIGURE 7. Eigenfunctions on Ω_{L} associated to the first eight positive eigenvalues. Densities $|\boldsymbol{E}|$ (top) and $|H|$ (bottom). Corresponding arrow fields $\boldsymbol{E}$. We have normalized each individual density to have as maximum the value 1.



k	$(\omega_k)^+_-$
1	1.1^{46}_{25}
2	1.54^{41}_{16}
3	1.54^{41}_{18}
4	2.0^{82}_{64}
5	2.0^{82}_{78}
6	2.0^{82}_{78}
7	2.2^{35}_{13}
8	2.2^{35}_{14}
9	2.32^{67}_{58}
10	2.33^{23}_{09}
11	2.33^{24}_{10}
12	$2.^{40}_{36}$
13	$2.^{60}_{59}$
14	$2.^{60}_{59}$
15	2.60^{56}_{09}

FIGURE 8. Spectral enclosures for the spectrum lying on the interval $(0, 2\sqrt{2})$ for the Fichera domain Ω_{F}. Here we have fixed $t_{\mathrm{up}} = 0.2$ and $t_{\mathrm{low}} = 2.8$. We considered mesh refined at the re-entrant edges as shown on the left. The final number of DOF=208680.

fashion, in order to build eigenfunctions on Ω_{F}. Therefore the set $\{\pm 2\sqrt{l^2 + m^2 + n^2}\}$ where not two indices vanish simultaneously certainly lies inside $\sigma(\mathcal{M})$. The first eigenvalue in this set is $2\sqrt{2}$. We conjecture that there are exactly 15 eigenvalues in the interval $(0, 2\sqrt{2})$. Furthermore, we conjecture that the multiplicity counting of the spectrum in this interval is

$$1,\ 2,\ 3,\ 2,\ 1,\ 2,\ 1,\ 3.$$

The table on the right of Figure 8 shows a numerical estimation of these eigenvalues. Here we have fixed $t_{\mathrm{up}} = 0.2$ and $t_{\mathrm{low}} = 2.8$. We have considered a family of mesh refined along the re-entrant edges. The final mesh is shown on the left side of Figure 8. We have stopped the algorithm when the tolerance $\delta = 0.05$ has been achieved. However, note that the accuracy is much higher for the indices $k = 2, 3, 9, 10, 11, 15$.

The slight numerical discrepancy shown in the table for the seemingly multiple eigenvalues appears to be a consequence of the fact that the meshes employed are not entirely symmetric with respect to permutation of the spacial coordinates. Figure 9 includes the corresponding

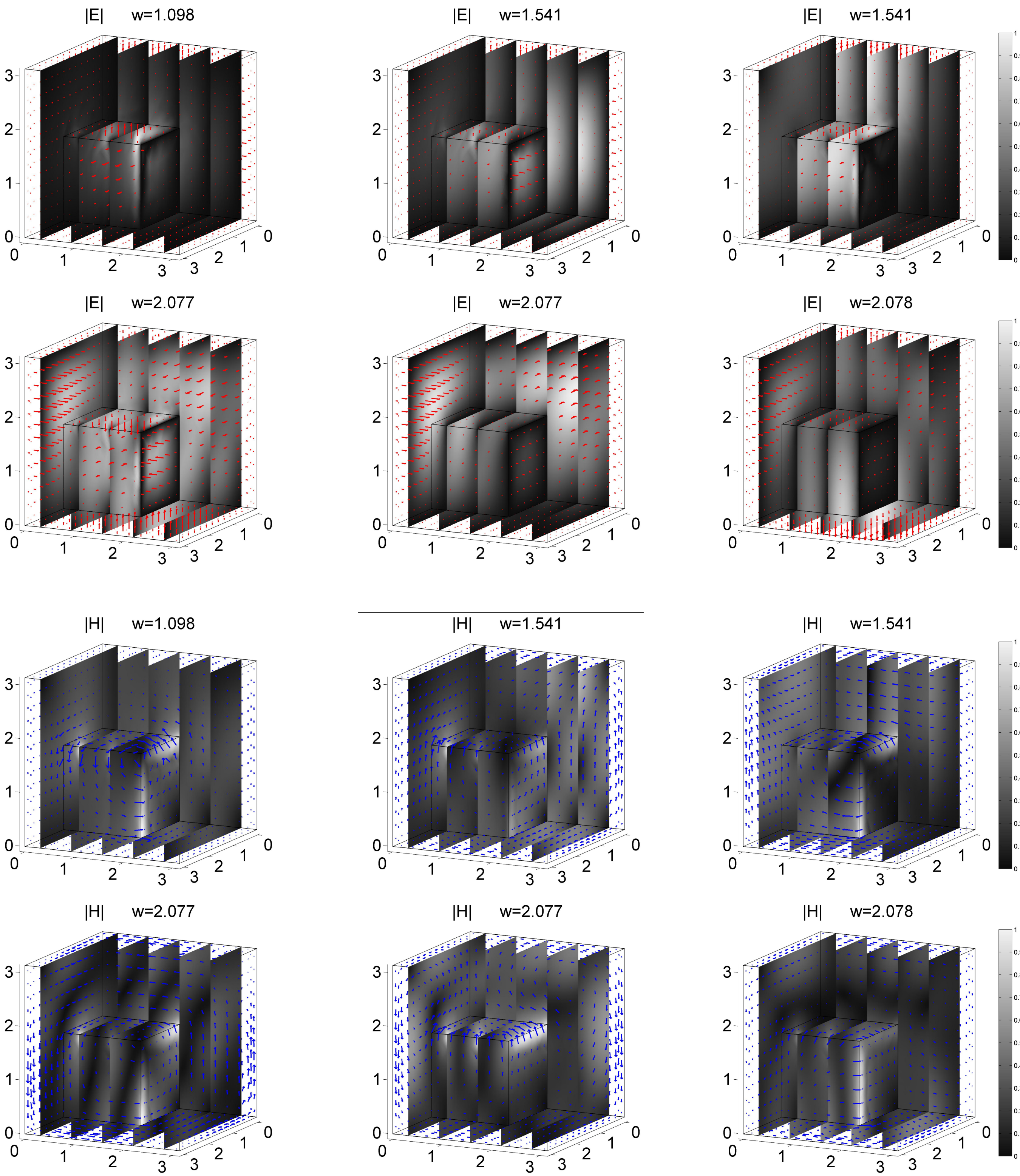


FIGURE 9. The first six eigenfunctions on Ω_{F} for the first six positive eigenvalues. Densities $|\boldsymbol{E}|$ (top) and $|\boldsymbol{H}|$ (bottom). Corresponding arrow fields $\boldsymbol{E}$ (red) and $\boldsymbol{H}$ (blue) on $\partial\Omega_{\mathrm{F}}$.

RF	DOF	$t_{\text{low}} = 1.95$ $(j=1)$ ω_3^-	$t_{\text{low}} = 2.05$ $(j=3)$ ω_3^-	$t_{\text{up}} = 1.05$ $(j=1)$ ω_3^+	$t_{\text{up}} = 0.7$ $(j=3)$ ω_3^+
1	4143	1.24764	1.26640	1.50395	1.3436
0.1	9648	1.25029	1.26830	1.49282	1.3336
0.01	74226	1.25063	1.26846	1.48899	1.3274

TABLE 1. Numerical experiment showing the dependence of the accuracy in the Zimmermann-Mertins method on the choice of t_{up} and t_{low}. It is preferable to pick t_{up} and t_{low} as far as possible from ω than to increase the dimension of $\mathcal{L}$.

approximated eigenfunctions. The mesh employed for this calculation is the same as that of Figure 8.

The slit square. As mentioned in Section 6, for a single trial space $\mathcal{L}$, the accuracy of the eigenvalue bounds produced by the Zimmermann-Mertins method depends on the position of t relative to the spectrum of $\mathcal{M}$. In this final experiment we demonstrate that this dependence might vary significantly with t. The numerical evidence suggests that a good choice of t_{up} and t_{low} plays a role in the design of efficient algorithms for eigenvalue calculation based on this method.

Let $\tilde{\Omega} = (0,\pi)^2 \setminus S$ for $S = [\pi/2,\pi] \times \{\pi/2\}$. Benchmarks on the eigenvalues of (50) are known by means of solving numerically the corresponding Neumann Laplacian problems, [15]. The first seven positive eigenvalues are

$$\omega_1 \approx 0.647375015,\ \omega_2 = 1,\ \omega_3 \approx 1.280686161,$$
$$\omega_4 = \omega_5 = 2,\ \omega_6 \approx 2.096486081 \quad \text{and} \quad \omega_7 \approx 2.229523505.$$

The eigenfunctions associated to ω_2, ω_4 and ω_5 are smooth, as they are also eigenfunctions on Ω_{sqr}. On the other hand, ω_1 and ω_3, correspond to singular eigenfunctions. Standard nodal elements are completely unsuitable for the computation of these eigenvalues, even with a significant refinement of the mesh on S.

Table 1 shows computation of $\omega_3^{\pm}$ on a mesh that is increasingly refined at S with a factor RF for two pairs of choices of t_{up} and t_{low}. Here $h = 0.1$ and we consider Lagrange elements of order $r = 1$. The choice of t_{up} and t_{low} further from ω_3 even with the very coarse mesh, provide qualitatively sharper $\omega_3^{\pm}$ than the other choices even with the finer mesh.

Appendix A. Further geometrical properties of $F^j_{\mathcal{L}}(t)$

Various extensions of Lemma 2 to the case $j>1$ are possible, however it is difficult to write these results in a neat fashion. The proposition below is one such an extension.

The following generalization of Danskin's Theorem is a direct consequence of [6, Theorem D1]. Let $J\subset\mathbb{R}$ be an open segment. Denote by

$$\partial_t^{\pm} f(t)=\lim_{\tau\to 0^+}\pm\frac{f(t\pm\tau)-f(t)}{\tau},$$

the one-side derivatives of a function $f:J\longrightarrow\mathbb{R}$. Let $\mathcal{V}$ be a compact topological space. For given $\mathcal{J}:J\times\mathcal{V}\longrightarrow\mathbb{R}$ we write

$$\tilde{\mathcal{J}}(t)=\max_{v\in\mathcal{V}}\mathcal{J}(t,v)\quad\text{and}\quad\tilde{\mathcal{V}}(t)=\left\{\tilde{v}\in\mathcal{V}:\tilde{\mathcal{J}}(t)=\mathcal{J}(t,\tilde{v})\right\}.$$

Lemma 17. *If the map $\mathcal{J}$ is upper semi-continuous and $\partial_t^{\pm}\mathcal{J}(t,v)$ exist for all $(t,v)\in J\times\mathcal{V}$, then also $\partial_t^{\pm}\tilde{\mathcal{J}}(t)$ exist for all $t\in J$ and*

$$\partial_t^{\pm}\tilde{\mathcal{J}}(t)=\max_{\tilde{v}\in\tilde{\mathcal{V}}(t)}\partial_t^{\pm}\tilde{\mathcal{J}}(t,\tilde{v}). \tag{52}$$

In the statement of this lemma, note that the left and right derivatives of both $\mathcal{J}$ and $\tilde{\mathcal{J}}$ might possibly be different.

Proposition 18. *Let $j=1,\dots,n$ and $t\in\mathbb{R}$ be fixed. The following assertions are equivalent.*

a) $|F^j_{\mathcal{L}}(t)-F^j_{\mathcal{L}}(s)|=|t-s|$ *for some* $s\neq t$.

b) There exists an open segment $J\subset\mathbb{R}$ containing t in its closure, such that

$$|F^j_{\mathcal{L}}(t)-F^j_{\mathcal{L}}(s)|=|t-s|\qquad\forall s\in\overline{J}.$$

c) There exists an open segment $J\subset\mathbb{R}$ containing t in its closure, such that

$$\forall s\in J,\ \textit{either}\quad \mathcal{L}\cap\mathcal{E}_{s+F^j_{\mathcal{L}}(s)}\neq\{0\}\quad\textit{or}\quad\mathcal{L}\cap\mathcal{E}_{s-F^j_{\mathcal{L}}(s)}(A)\neq\{0\}.$$

Proof.

<u>*a)* $\Rightarrow$ *b)*</u>. Assume *a)*. Since $r\mapsto r\pm F^j_{\mathcal{L}}(r)$ are continuous and monotonically increasing, then they have to be constant in the closure of

$$J=\{\tau t+(1-\tau)s:0<\tau<1\}.$$

This is precisely *b)*.

<u>*b)* $\Rightarrow$ *c)*</u>. Assume *b)*. Then $s\mapsto F^j_{\mathcal{L}}(s)$ is differentiable in J and its one-side derivatives are equal to 1 or -1 in the whole of this interval.

For this part of the proof, we aim at applying (52), in order to get another expression for these derivatives.

Let $\mathcal{F}_j$ be the family of $(j-1)-$dimensional linear subspaces of $\mathcal{L}$. Identify an orthonormal basis of $\mathcal{L}$ with the canonical basis of $\mathbb{C}^n$. Then any other orthonormal basis of $\mathcal{L}$ is represented by a matrix in $\mathrm{O}(n)$, the orthonormal group. By picking the first $(j-1)$ columns of these matrices, we cover all possible subspaces $V \in \mathcal{F}_j$. Indeed we just have to identify $(\underline{v}_1|\ldots|\underline{v}_{j-1})$ for $[\underline{v}_{kl}]_{kl=1}^n \in \mathrm{O}(n)$ with $V = \mathrm{Span}\{\underline{v}_k\}_{k=1}^{j-1}$.

Let

$$\mathcal{K}_j = \Big\{(\underline{v}_1,\ldots,\underline{v}_{j-1}) : [\underline{v}_{kl}]_{kl=1}^n \in \mathrm{O}(n)\Big\} \subset \underbrace{\mathbb{C}^n \times \ldots \times \mathbb{C}^n}_{j-1}.$$

Then $\mathcal{K}_j$ is a compact subset in the product topology of the right hand side. According to (11),

$$F_{\mathcal{L}}^j(s) = \max_{(\underline{v}_1,\ldots,\underline{v}_{j-1})\in\mathcal{K}_j} g(s;\underline{v}_1,\ldots,\underline{v}_{j-1})$$

where

$$g(s;\underline{v}_1,\ldots,\underline{v}_{j-1}) = \min_{\substack{(a_1,\ldots,a_{j-1})\in\mathbb{C}^{j-1}\\ \sum|a_k|^2=1}} \left|\sum a_k\tilde{v}_k\right|_s .$$

Here we have used the correspondence between $\underline{v}_k \in \mathbb{C}^n$ and $\tilde{v}_k \in \mathcal{L}$ in the orthonormal basis set above. We write

$$g(r,V) = g(r;\underline{v}_1,\ldots,\underline{v}_{j-1}) \quad \text{for} \quad V = \mathrm{Span}\{\underline{\tilde{v}}_k\}_{k=1}^{j-1} \in \mathcal{F}_j.$$

The map $g : J \times \mathcal{K}_j \longrightarrow \mathbb{R}^+$ is the minimum of a differentiable function, so the hypotheses of Lemma 17 are satisfied by $\mathcal{J} = -g$. Hence, by virtue of (52),

$$\partial_s^{\pm} g(s,V) = \min_{\substack{u\in\mathcal{L}\ominus V,\|u\|=1\\ |u|_s=g(s,V)}} \left(\frac{\mathrm{Re}\, l_s(u,u)}{|u|_s}\right) .$$

As minima of continuous functions, $g(s,V)$ and $\partial_s^{\pm} g(s,V)$ are upper semi-continuous. Therefore, a further application of Lemma 17 yields

$$\begin{aligned}\partial_s^{\pm} F_{\mathcal{L}}^j(s) &= \max_{\substack{(\underline{v}_1,\ldots,\underline{v}_{j-1})\in\mathcal{K}_j\\ g(s;\underline{v}_1,\ldots,\underline{v}_{j-1})=F_{\mathcal{L}}^j(s)}} \partial_s^{\pm} g(s,\underline{v}_1,\ldots,\underline{v}_{j-1})\\ &= \max_{\substack{V\in\mathcal{F}_j\\ g(s,V)=F_{\mathcal{L}}^j(s)}} \min_{\substack{u\in\mathcal{L}\ominus V,\|u\|=1\\ |u|_s=g(s,V)}} \left(\frac{\mathrm{Re}\, l_s(u,u)}{|u|_s}\right) .\end{aligned}$$

Now, this shows that

$$\left| \max_{\substack{V\in\mathcal{F}_j \\ g(s,V)=F_{\mathcal{L}}^j(s)}} \min_{\substack{u\in\mathcal{L}\ominus V, \|u\|=1 \\ |u|_s=g(s,V)}} \left(\frac{\operatorname{Re} l_s(u,u)}{|u|_s} \right) \right| = 1.$$

As $\mathcal{L}$ is finite dimensional, there exists a vector $u \in \mathcal{L}$ satisfying the identity $|u|_s = F_{\mathcal{L}}^j(s)$ such that

$$\frac{|\operatorname{Re} l_s(u,u)|}{|u|_s} = 1.$$

Thus $|\operatorname{Re}\langle (A-s)u,u\rangle| = \langle (A-s)u,(A-s)u\rangle = F_{\mathcal{L}}^j(s)$. Hence, according to the "equality" case in the Cauchy-Schwarz inequality, u must be an eigenvector of A associated with either $s+F_{\mathcal{L}}^j(s)$ or $s-F_{\mathcal{L}}^j(s)$. This is precisely *c)*.

c) $\Rightarrow$ *a)*. Under the condition *c)*, we know there exists an open segment $\tilde{J} \subseteq J$, possibly smaller, such that $t \in \overline{\tilde{J}}$ and $F_{\mathcal{L}}^j(s) = \mathfrak{d}_j(s)$ for all $s \in \tilde{J}$. As $|\mathfrak{d}_j(s) - \mathfrak{d}_j(r)| = |s-r|$, then either *a)* is immediate, or it follows by taking $r \to t$. □

Appendix B. A Comsol v4.3 LiveLink code

```
%     Comsol V4.3 LiveLink code for computing
%   fundamental frequencies on a resonant cavity
%      with perfect conductivity conditions
% the test geometry below is the Fichera domain.
%
%      Gabriel Barrenechea, Lyonell Boulton
%              and Nabile Boussaid
%                                       November 2012

% INITIALIZATION OF THE MODEL FROM SCRATCHES

model = ModelUtil.create('Model');
geom1=model.geom.create('geom1', 3);
mesh1=model.mesh.create('mesh1', 'geom1');
w=model.physics.create('w', 'WeakFormPDE', 'geom1',
              {'E1','E2', 'E3', 'H1', 'H2', 'H3'});

% CREATING THE GEOMETRY - IN THIS CASE THE FICHERA DOMAIN

hex1=geom1.feature.create('hex1', 'Hexahedron');
hex1.set('p',{'0' '0' '0' '0' 'pi' 'pi' 'pi' 'pi';
```

```
              '0' '0' 'pi' 'pi' '0' '0' 'pi' 'pi';
              '0' 'pi' 'pi' '0' '0' 'pi' 'pi' '0'});
hex2=geom1.feature.create('hex2', 'Hexahedron');
hex2.set('p',{'0' '0' '0' '0' 'pi/2' 'pi/2' 'pi/2' 'pi/2';
              '0' '0' 'pi/2' 'pi/2' '0' '0' 'pi/2' 'pi/2';
              '0' 'pi/2' 'pi/2' '0' '0' 'pi/2' 'pi/2' '0'});
dif1 = geom1.feature.create('dif1', 'Difference');
dif1.selection('input').set({'hex1'});
dif1.selection('input2').set({'hex2'});
geom1.run;

%CREATING THE GEOMETRY
model.mesh('mesh1').automatic(false);
model.mesh('mesh1').feature('size').set('custom', 'on');
model.mesh('mesh1').feature('size').set('hmax', '.8');
mesh1.run;

% PARAMETER t WHERE TO LOOK FOR EIGENVALUES
parat=2.2;

% WHETHER TO LOOK FOR THE EIGENVALUES TO THE LEFT (-) OR
% RIGHT (+) AND WHERE ABOUT
shi=-.3;
model.param.set('tt', num2str(parat));
searchtau=shi;

% FINITE ELEMENTS TO USE AND ORDER
w.prop('ShapeProperty').set('shapeFunctionType', 'shlag');
w.prop('ShapeProperty').set('order', 3);

% PHYSICS
w.feature('wfeq1').set('weak',1 ,'(H3y-H2z)*(H3y_test-H2z_test)-
i*2*tt*(H3y-H2z)*E1_test+tt^2*E1*E1_test+(i*(H3y-H2z)-tt*E1)*E1t_test');
w.feature('wfeq1').set('weak',2 ,'(H1z-H3x)*(H1z_test-H3x_test)-
i*2*tt*(H1z-H3x)*E2_test+tt^2*E2*E2_test+(i*(H1z-H3x)-tt*E2)*E2t_test');
w.feature('wfeq1').set('weak',3 ,'(H2x-H1y)*(H2x_test-H1y_test)-
i*2*tt*(H2x-H1y)*E3_test+tt^2*E3*E3_test+(i*(H2x-H1y)-tt*E3)*E3t_test');
w.feature('wfeq1').set('weak',4 ,'(E3y-E2z)*(E3y_test-E2z_test)+
i*2*tt*(E3y-E2z)*H1_test+tt^2*H1*H1_test+((-i)*(E3y-E2z)-tt*H1)*H1t_test');
w.feature('wfeq1').set('weak',5 ,'(E1z-E3x)*(E1z_test-E3x_test)+
i*2*tt*(E1z-E3x)*H2_test+tt^2*H2*H2_test+((-i)*(E1z-E3x)-tt*H2)*H2t_test');
w.feature('wfeq1').set('weak',6 ,'(E2x-E1y)*(E2x_test-E1y_test)+
```

```
i*2*tt*(E2x-E1y)*H3_test+tt^2*H3*H3_test+((-i)*(E2x-E1y)-tt*H3)*H3t_test');

% BOUNDARY CONDITIONS
cons1=model.physics('w').feature.create('cons1', 'Constraint');
cons1.set('R', 2, 'E2');
cons1.set('R', 3, 'E3');
cons1.selection.set([1 8 9]);
cons2=model.physics('w').feature.create('cons2', 'Constraint');
cons2.set('R', 1, 'E1');
cons2.set('R', 3, 'E3');
cons2.selection.set([2 5 7]);
cons3=model.physics('w').feature.create('cons3', 'Constraint');
cons3.set('R', 1, 'E1');
cons3.set('R', 2, 'E2');
cons3.selection.set([3 4 6]);

% HOW MANY EIGENVALUES TO LOOK FOR AROUND t
neval=3;

% SOLVING THE MODEL
std1=model.study.create('std1');
model.study('std1').feature.create('eigv', 'Eigenvalue');
model.study('std1').feature('eigv').set('shift', num2str(searchtau));
model.study('std1').feature('eigv').set('neigs', neval);
std1.run;

% STORING SOLUTION FOR POST PROCESSING
[SZ,NDOFS,DATA,NAME,TYPE]= mphgetp(model,'solname','sol1');



% DISPLAYING SOLUTION
for inde=1:neval,
tauinv=(real(DATA(inde)));
bd=parat+tauinv;
if tauinv<0, disp(['lower= ',num2str(bd,10)]);
else disp(['upper= ',num2str(bd,10)]);
end
disp(['DOF= ',num2str(NDOFS)])
end
```

Acknowledgements

We kindly thank Michael Levitin and Stefan Neuwirth for their suggestions during the preparation of this manuscript. We kindly thank Université de Franche-Comté, University College London and the Isaac Newton Institute for Mathematical Sciences (programme Spectral Theory of Relativistic Operators), for their hospitality. Funding was provided by MOPNET, the British-French project PHC Alliance (22817YA), the British Engineering and Physical Sciences Research Council (EP/I00761X/1) and the French Ministry of Research (ANR-10-BLAN-0101).

References

[1] C. Amrouche, C. Bernardi, M. Dauge, and V. Girault, Vector potentials in three-dimensional non-smooth domains, Math. Methods Appl. Sci., 21 (1998), pp. 823–864.

[2] D. Arnold, R. Falk, and R. Winther, Finite element exterior calculus: from hodge theory to numerical stability, Bulletin of the American Mathematical Society, 47 (2010), pp. 281–354.

[3] G. Barrenechea, L. Boulton, and N. Boussaïd, Some remarks on the spectral properties of the maxwell operator on rough domains and domains with symmetries, in preparation.

[4] H. Behnke, Lower and upper bounds for sloshing frequencies, Inequalities and Applications, (2009), pp. 13–22.

[5] H. Behnke and U. Mertins, Bounds for eigenvalues with the use of finite elements, Perspectives on Enclosure Methods, (2001), p. 119.

[6] P. Bernhard and A. Rapaport, On a theorem of Danskin with an application to a theorem of von Neumann-Sion, Nonlinear Anal., 24 (1995), pp. 1163–1181.

[7] M. Birman and M. Solomyak, The self-adjoint Maxwell operator in arbitrary domains, Leningrad Math. J, 1 (1990), pp. 99–115.

[8] D. Boffi, Finite element approximation of eigenvalue problems, Acta Numer., 19 (2010), pp. 1–120.

[9] D. Boffi, P. Fernandes, L. Gastaldi, and I. Perugia, Computational models of electromagnetic resonators: analysis of edge element approximation, SIAM J. Numer. Anal., 36 (1999), pp. 1264–1290 (electronic).

[10] A. Bonito and J.-L. Guermond, Approximation of the eigenvalue problem for the time harmonic Maxwell system by continuous Lagrange finite elements, Math. Comp., 80 (2011), pp. 1887–1910.

[11] L. Boulton and M. Strauss, Eigenvalue enclosures for the MHD operator, BIT Numerical Mathematics, (2012).

[12] J. H. Bramble, T. V. Kolev, and J. E. Pasciak, The approximation of the Maxwell eigenvalue problem using a least-squares method, Math. Comp., 74 (2005), pp. 1575–1598 (electronic).

[13] A. Buffa, P. Ciarlet, and E. Jamelot, Solving electromagnetic eigenvalue problems in polyhedral domains, Numer. Math., 113 (2009), pp. 497–518.

[14] F. Chatelin, Spectral Approximation of Linear Operators, Academic Press, New York, 1983.
[15] M. Dauge, Computations for Maxwell equations for the approximation of highly singular solutions, 2004, http://perso.univ-rennes1.fr/monique.dauge/benchmax.html.
[16] E. B. Davies, Spectral enclosures and complex resonances for general self-adjoint operators, LMS J. Comput. Math, 1 (1998), pp. 42–74.
[17] ——, A hierarchical method for obtaining eigenvalue enclosures, Math. Comp., 69 (2000), pp. 1435–1455.
[18] E. B. Davies and M. Plum, Spectral pollution, IMA J. Numer. Anal., 24 (2004), pp. 417–438.
[19] A. Ern and J.-L. Guermond, Theory and practice of finite elements, vol. 159 of Applied Mathematical Sciences, Springer-Verlag, New York, 2004.
[20] V. Girault and P.-A. Raviart, Finite element methods for Navier-Stokes equations, vol. 5 of Springer Series in Computational Mathematics, Springer-Verlag, Berlin, 1986. Theory and algorithms.
[21] F. Goerisch and J. Albrecht, The convergence of a new method for calculating lower bounds to eigenvalues, in Equadiff 6 (Brno, 1985), vol. 1192 of Lecture Notes in Math., Springer, Berlin, 1986, pp. 303–308.
[22] P. Monk, Finite element methods for Maxwell's equations, Clarendon Press, Cambridge, 2003.
[23] G. Strang and G. Fix, An Analysis of the Finite Element Method, Prentice Hall, London, 1973.
[24] H. F. Weinberger, Variational Methos for Eigenvalue Approximation, Society for Industrial and Applied Mathematics, Philadelphia, 1974.
[25] S. Zimmermann and U. Mertins, Variational bounds to eigenvalues of self-adjoint eigenvalue problems with arbitrary spectrum, Z. Anal. Anwendungen, 14 (1995), pp. 327–345.

Department of Mathematics and Statistics, University of Strathclyde, 26 Richmond Street, Glasgow G1 1XH, Scotland
E-mail address: `gabriel.barrenechea@strath.ac.uk`

Department of Mathematics and Maxwell Institute for Mathematical Sciences, Heriot-Watt University, Edinburgh, EH14 4AS, UK
E-mail address: `L.Boulton@hw.ac.uk`

Département de Mathématiques, Université de Franche-Comté, Besançon, France
E-mail address: `nboussai@univ-fcomte.fr`