ON SPECTRAL ANALYSIS AND A NOVEL ALGORITHM FOR TRANSMISSION EIGENVALUE PROBLEMS

TIEXIANG LI*, WEI-QIANG HUANG†, WEN-WEI LIN†, AND JIJUN LIU*

Abstract. The transmission eigenvalue problem, except for its critical role in inverse scattering problems, is of its own special interest due to the fact that the corresponding differential operator is neither elliptic nor self-adjoint. In this paper, we provide the spectral analysis and propose a novel iterative algorithm for the computation of a few positive real eigenvalues and the corresponding eigenfunctions of the transmission eigenvalue problem. Based on the continuous finite element method, we first derive an associated symmetric quadratic eigenvalue problem (QEP) for the transmission eigenvalue problem to eliminate the nonphysical zero eigenvalue while preserve all nonzero ones. In addition, the derived QEP enables us to consider more refined discretization to overcome the limitation on the number of degree of freedoms. We then transform the QEP to a parameterized symmetric definite generalized eigenvalue problem (GEP) and develop a secant-type iteration for solving the resulting GEPs. Moreover, we examine the spectral analysis for various existence intervals of desired positive real eigenvalues, since a few lowest positive real transmission eigenvalues are of practical interest in the estimation and the reconstruction of the index of refraction. Numerical experiments show that the proposed method can find those desired smallest positive real transmission eigenvalues accurately, efficiently, and robustly.

Key words. transmission eigenvalues, quadratic eigenvalue problems, symmetric positive definite, spectral analysis, secant-type iteration method

AMS subject classifications. 78A46, 65N30, 65N25, 65F15

1. Introduction.

The transmission eigenvalue problem has recently attracted much attention in the inverse scattering community [11, 20, 7, 17, 5, 6, 10, 8, 9]. This is due to the reason that one can determine the transmission eigenvalues from the far-field pattern of the scattered wave and then use them to estimate the material properties of the scattering object [4, 2, 3, 5, 6, 23]. On the other hand, transmission eigenvalues are also related to the validity of some recently developed reconstruction method for scattering problems such as linear sampling method and factorization method [9]. For the recent progress in the theories and applications of transmission eigenvalue problems, we refer to [8] and the references therein.

In this paper, we consider the case of scattering of acoustic waves by a bounded and simply connected inhomogeneous medium $\Omega \subset \mathbb{R}^2$. This scattering model for the 2D Helmholtz equation can be considered as a special case of the Maxwell equations for the electromagnetic wave scattering by a 3D cylinder model under some physical assumptions on incident waves and cylinder parameters. In this simplified 2D case, the transmission eigenvalue problem is to find $k \in \mathbb{C}$ and $u, v \in L^2(\Omega)$ with $u - v \in H^2(\Omega)$

*Department of Mathematics, Southeast University, Nanjing, 210096, People's Republic of China (txli@seu.edu.cn, jjliu@seu.edu.cn). The research of first and fourth authors were supported by the NSFC (No. 91330109), China.

†Department of Applied Mathematics, National Chiao Tung University, Hsinchu 300, Taiwan (wqhuang@math.nctu.edu.tw, wwlin@math.nctu.edu.tw). The second and third authors would like to acknowledge the support from the Ministry of Science and Technology (NSC 102-2628-M-009-002-), the National Center for Theoretical Sciences, and the ST Yau Center at the National Chiao Tung University in Taiwan.

such that

$$\Delta u + k^2 n(x) u = 0, \quad \text{in} \quad \Omega, \tag{1.1a}$$

$$\Delta v + k^2 v = 0, \quad \text{in} \quad \Omega, \tag{1.1b}$$

$$u - v = 0, \quad \text{on} \quad \partial\Omega, \tag{1.1c}$$

$$\frac{\partial u}{\partial \mathbf{n}} - \frac{\partial v}{\partial \mathbf{n}} = 0, \quad \text{on} \quad \partial\Omega, \tag{1.1d}$$

where $\mathbf{n}$ is the unit outer normal to the smooth boundary $\partial\Omega$ and the index of refraction $n(x)$ is positive. Any nonzero value k such that there are nontrivial solutions u and v of (1.1) is called a *transmission eigenvalue*.

Efficient numerical methods to determine transmission eigenvalues are required in estimating the index of refraction [5, 23] and, in addition, numerical evidence obtained for the discrete system might contribute to the progress of further theoretical developments such as the distribution of real eigenvalues for the original infinite dimensional system. Nonetheless, numerical techniques for solving the transmission eigenvalues are limited, and only a few papers have addressed the numerical computation on this topic in the past few years. It is challenging because, firstly, the problem (1.1) is neither elliptic nor self-adjoint so that it cannot be covered by the standard theory of elliptic partial differential equations. Secondly, owing to the non-self-adjointness, the resulting eigenvalue problem derived from the standard finite element method is non-Hermitian. Moreover, the nonphysical transmission eigenvalue $k = 0$ has an infinite-dimensional eigenspace as can be seen from the fact that any harmonic function on Ω is an eigenfunction of (1.1) with $k = 0$.

Among the numerical investigations on the computation of transmission eigenvalues, Colton, Monk and Sun [10] first proposed three finite element methods to compute the Helmholtz transmission eigenvalues. However, the mesh has to be kept rather coarse so that the QZ algorithm [18] can be used to find all approximate eigenpairs of the induced non-Hermitian eigenvalue problems. Then, Sun [24] proposed two iterative methods together with the convergence analysis based on the existence theory of the fourth order reformulation for the transmission eigenvalues [7, 20]. Ji, Sun and Turner [14] proposed a mixed finite element method and provided a MATLAB implementation on an adaptive algorithm to solve the corresponding generalized eigenvalue problem (GEP). Later, Monk and Sun [19] applied this method to the Maxwell's transmission eigenvalue problem. In [15], Ji, Sun and Xie used the multilevel correction method to transform the solution of the transmission problem to a series of solutions corresponding to linear boundary value problems and solved them by the multigrid method.

1.1. Contributions. The main contribution of this paper is to study the spectral analysis for a discrete model of (1.1) and to propose an efficient scheme for computing *positive real* transmission eigenvalues and associated eigenvectors. Based on the continuous finite element method in [10], we first show that the associated discretization model of (1.1) can induce a symmetric quadratic eigenvalue problem (QEP) which can completely exclude the inference of zero eigenvalues as presented in [10]. According to the spectral decompositions on the coefficient matrices of the QEP, which have particular structures, we then provide the spectral analysis for estimating various existing intervals of desired positive real eigenvalues of (1.1). Note that only a few smallest positive real transmission eigenvalues are of practical interest in the estimation and the reconstruction of the index of refraction [2, 23] while

a transmission eigenvalue with the smallest norm may not be the desire one owing to the existence of complex transmission eigenvalues. These theoretical analyses and practical applications further motivate us to develop an efficient and robust numerical method for computing desired *positive real* transmission eigenvalues but avoiding any possible complex eigenvalues.

1.2. Notations and overview. The following notations are frequently used throughout this paper. Other notations will be clearly defined whenever they are used. For convenience, we use λ to denote the square of the transmission eigenvalue k, i.e., $\lambda := k^2$. For a given mesh, ν and ρ are used to indicate the number of interior nodes and the number of boundary nodes respectively. I_N denotes the $N \times N$ identity matrix with the given size N, and $\mathbf{0}$ is the zero vectors or matrices with appropriate sizes. Given a real square matrix A, we write $A \succ 0$ if A is symmetric positive definite. The notation $\cdot^\top$ is used to represent the transpose of vectors or matrices.

This paper is organized as follows. In section 2, we review the continuous finite element method in [10] for the discretization of the transmission eigenvalue problem (1.1) and derive the associated symmetric QEP. In section 3, we study the spectral analysis of the QEP and estimates various existing intervals of positive real eigenvalues of (1.1). Based on the spectral analysis, in section 4, we develop an efficient and robust numerical method for the computation of positive real transmission eigenvalues. Numerical experiments with different $n(x)$ on various domains are presented in section 5 and concluding remarks are given in section 6.

2. Discretization of transmission eigenvalue problems. To treat the transmission eigenvalue problem (1.1), Colton, Monk and Sun [10] proposed three finite element methods: *the Argyris method*, *the continuous finite element approximation* and *the mixed finite element method*. The continuous finite element method ends up with much sparse matrices than the other two methods and the implementation is also easy because only a linear finite element is used. The Argyris method takes more works because it needs to calculate the affine transformation for each triangle of the mesh. However, as numerical results presented in [10], these three methods did not converge for computing a few real eigenvalues via the MATLAB built-in eigensolver `eigs` for sparse matrices. In all cases, it must compute all eigenvalues using the MATLAB function `eig`. This limits the maximum number of degree of freedoms that can be used. Furthermore, the GEPs for (1.1) derived from continuous finite element method and mixed finite element method contain a large number of spurious zeros, which considerably influences convergence so that the `eigs` fails.

In this section, we will transform the GEP constructed by the continuous finite element method into a QEP with symmetric and positive definite coefficient matrices so that it can exclude the nonphysical transmission eigenvalue, $k = 0$, while preserves all nontrivial ones of the original problem.

2.1. Continuous finite element approximation. We briefly review the discretization of the transmission eigenvalue problem (1.1) based on the standard piecewise linear finite element method (see [10] for detailed discussion). Let

$$\begin{aligned} S_h &= \text{The space of continuous piecewise linear functions on } \Omega, \\ S_h^0 &= \text{The subspace of functions in } S_h \text{ that have vanishing DoF on } \partial\Omega, \\ S_h^B &= \text{The subspace of functions in } S_h \text{ that have vanishing DoF in } \Omega, \end{aligned}$$

where DoF is the degrees of freedom. Taking test functions $\eta_h, \zeta_h \in S_h^0$ and $\gamma_h \in S_h^B$, and applying the integration by parts, one can show that the discretization of (1.1)

is to seek $u_{0,h}, v_{0,h} \in S_h^0$ and $u_{B,h} \in S_h^B$ satisfying

$$\begin{aligned}
&\int_\Omega \nabla(u_{0,h}+u_{B,h})\cdot\nabla\eta_h dx - \lambda\int_\Omega n(u_{0,h}+u_{B,h})\eta_h dx = 0,\\
&\int_\Omega \nabla(v_{0,h}+u_{B,h})\cdot\nabla\zeta_h dx - \lambda\int_\Omega (v_{0,h}+u_{B,h})\zeta_h dx = 0,\\
&\int_\Omega \nabla(u_{0,h}-v_{0,h})\cdot\nabla\gamma_h dx - \lambda\int_\Omega (n(u_{0,h}+u_{B,h})-(v_{0,h}+u_{B,h}))\gamma_h dx = 0,
\end{aligned}$$

for all $\eta_h, \zeta_h \in S_h^0$ and $\gamma_h \in S_h^B$. If $\{\phi_j\}_{j=1}^{\nu}$ and $\{\psi_j\}_{j=1}^{\rho}$ denote standard nodal bases for the finite element spaces of S_h^0 and S_h^B, respectively, then $u_{0,h}, v_{0,h}$ and $u_{B,h}$ can be written as $u_{0,h} = \sum_{j=1}^{\nu} u_j\phi_j$, $v_{0,h} = \sum_{j=1}^{\nu} v_j\phi_j$ and $u_{B,h} = \sum_{j=1}^{\rho} w_j\psi_j$.

Matrix	Dimension	Definition
$K \succ 0$	$\nu \times \nu$	Interior space stiffness matrix. $K_{ij} = \int_\Omega \nabla\phi_i\cdot\nabla\phi_j dx$
E	$\nu \times \rho$	Interior/Boundary stiffness matrix. $E_{ij} = \int_\Omega \nabla\phi_i\cdot\nabla\psi_j dx$
$M_n, M_1 \succ 0$	$\nu \times \nu$	Interior space mass matrices. $(M_n)_{ij} = \int_\Omega n\phi_i\phi_j dx,\ \ (M_1)_{ij} = \int_\Omega \phi_i\phi_j dx$
F_n, F_1	$\nu \times \rho$	Interior/Boundary mass matrices. $(F_n)_{ij} = \int_\Omega n\phi_i\psi_j dx,\ \ (F_1)_{ij} = \int_\Omega \phi_i\psi_j dx$
$G_n, G_1 \succ 0$	$\rho \times \rho$	Boundary mass matrices. $(G_n)_{ij} = \int_\Omega n\psi_i\psi_j dx,\ \ (G_1)_{ij} = \int_\Omega \psi_i\psi_j dx$

Table 1: Stiffness and mass matrices with $n(x) > 1$, $x \in \overline{\Omega}$.

We assume, without loss of generality, that $n(x) > 1$, $x \in \overline{\Omega}$, the analysis of the case for $0 < n(x) < 1$ is similar. Using the stiffness and mass matrices indicated in Table 1, we can represent the weak form of the continuous finite element method for problem (1.1) as a GEP

$$\mathcal{K}\mathbf{z} = \lambda\mathcal{M}\mathbf{z} \tag{2.1a}$$

with

$$\mathcal{K} := \begin{bmatrix} K & 0 & E \\ 0 & K & E \\ E^\top & -E^\top & 0 \end{bmatrix}, \quad \mathcal{M} := \begin{bmatrix} M_n & 0 & F_n \\ 0 & M_1 & F_1 \\ F_n^\top & -F_1^\top & G_n - G_1 \end{bmatrix}, \quad \mathbf{z} := \begin{bmatrix} \mathbf{u} \\ \mathbf{v} \\ \mathbf{w} \end{bmatrix}, \tag{2.1b}$$

where $\mathbf{u} = (u_1, \ldots, u_\nu)^\top$, $\mathbf{v} = (v_1, \ldots, v_\nu)^\top$, and $\mathbf{w} = (w_1, \ldots, w_\rho)^\top$ are the associated vectors of degree of freedoms. Note that the matrices K, M_n and M_1, all corresponding to the interior nodes, are symmetric positive definite.

2.2. Symmetric quadratic eigenvalue problems. We now show that the GEP (2.1) can induce a symmetric QEP. In order to make the following discussion

more concise, we first introduce some convenient notations. Let

$$M := M_n - M_1, \qquad F := F_n - F_1, \qquad G := G_n - G_1; \tag{2.2}$$

$$\widehat{K} = K - EG^{-1}F^\top, \quad \widehat{M}_1 := M_1 - F_1G^{-1}F^\top, \quad \widehat{M} := M - FG^{-1}F^\top; \tag{2.3}$$

$$\widehat{E} = E - KM^{-1}F, \quad \widehat{G} = G - F^\top M^{-1}F, \tag{2.4}$$

and, furthermore, we assume that

$$\begin{bmatrix} M & F \\ F^\top & G \end{bmatrix} \succ 0, \quad \text{or equivalently,} \quad \widehat{M} \succ 0 \text{ and } G \succ 0. \tag{2.5}$$

Let

$$W_\ell = \begin{bmatrix} I_\nu & \mathbf{0} & \mathbf{0} \\ I_\nu & -I_\nu & -FG^{-1} \\ \mathbf{0} & \mathbf{0} & I_\rho \end{bmatrix} \quad \text{and} \quad W_r = \begin{bmatrix} \mathbf{0} & I_\nu & \mathbf{0} \\ -I_\nu & I_\nu & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & I_\rho \end{bmatrix}.$$

Then (2.1) is equivalent to the equation $(W_\ell \mathcal{K} W_r)(W_r^{-1}\mathbf{z}) = \lambda (W_\ell \mathcal{M} W_r)(W_r^{-1}\mathbf{z})$ whose matrix-vector representation is given by

$$\begin{bmatrix} 0 & K & E \\ \widehat{K}^\top & 0 & 0 \\ E^\top & 0 & 0 \end{bmatrix} \begin{bmatrix} \mathbf{p} \\ \mathbf{u} \\ \mathbf{w} \end{bmatrix} = \lambda \begin{bmatrix} 0 & M_n & F_n \\ \widehat{M}_1^\top & \widehat{M} & 0 \\ F_1^\top & F^\top & G \end{bmatrix} \begin{bmatrix} \mathbf{p} \\ \mathbf{u} \\ \mathbf{w} \end{bmatrix}, \quad \mathbf{p} := \mathbf{u} - \mathbf{v}. \tag{2.6}$$

From the second row of equations in (2.6), we have

$$\lambda \mathbf{u} = \widehat{M}^{-1}(\widehat{K}^\top - \lambda \widehat{M}_1^\top)\mathbf{p}. \tag{2.7}$$

According to the third equations in (2.6), with the replacement of $\mathbf{u}$ by (2.7), $\mathbf{w}$ can be also expressed as a function of $\mathbf{p}$:

$$\lambda \mathbf{w} = G^{-1}[E^\top - F^\top \widehat{M}^{-1}\widehat{K}^\top - \lambda(F_1^\top - F^\top \widehat{M}^{-1}\widehat{M}_1^\top)]\mathbf{p}. \tag{2.8}$$

Plugging (2.7) and (2.8) into the first row of equations in (2.6), and further replacing M_n and F_n by $M + M_1$ and $F + F_1$, respectively (using (2.2)), one can show that if we collect the terms of $\widehat{M}^{-1}\widehat{M}_1^\top$ and $\widehat{M}^{-1}\widehat{K}^\top$ according to the degree of λ, then we can derive a quadratic equation in λ. Namely,

$$\begin{aligned}
& K\,\mathbf{u} + E\mathbf{w} = \lambda M_n \mathbf{u} + \lambda F_n \mathbf{w} \\
\Leftrightarrow\ & K\,(\lambda\mathbf{u}) + E(\lambda\mathbf{w}) - \lambda(M + M_1)(\lambda\mathbf{u}) - \lambda(F + F_1)(\lambda\mathbf{w}) = \mathbf{0} \\
\Leftrightarrow\ & \lambda^2\Big[\Big((M - FG^{-1}F^\top) + (M_1 - F_1G^{-1}F^\top)\Big)\widehat{M}^{-1}\widehat{M}_1^\top + FG^{-1}F_1^\top + F_1G^{-1}F_1^\top\Big]\mathbf{p} \\
& +\lambda\ \Big[-\Big((M - FG^{-1}G^\top) + (M_1 - F_1G^{-1}F^\top)\Big)\widehat{M}^{-1}\widehat{K}^\top - FG^{-1}E^\top \\
& \qquad -(K - EG^{-1}F^\top)\widehat{M}^{-1}\widehat{M}_1^\top - EG^{-1}F_1^\top - F_1G^{-1}E^\top\Big]\mathbf{p} \\
& + \quad \Big[(K - EG^{-1}F^\top)\widehat{M}^{-1}\widehat{K}^\top + EG^{-1}E^\top\Big]\mathbf{p} = \mathbf{0} \\
\Leftrightarrow\ & \lambda^2\Big(M_1 + \widehat{M}_1\widehat{M}^{-1}\widehat{M}_1^\top + F_1G^{-1}F_1^\top\Big)\mathbf{p} \\
& +\lambda\ \Big(-K - EG^{-1}F_1^\top - F_1G^{-1}E^\top - \widehat{K}\widehat{M}^{-1}\widehat{M}_1^\top - \widehat{M}_1\widehat{M}^{-1}\widehat{K}^\top\Big)\mathbf{p} \\
& + \quad \Big(\widehat{K}\widehat{M}^{-1}\widehat{K}^\top + EG^{-1}E^\top\Big)\mathbf{p} = \mathbf{0}.
\end{aligned}$$

Consequently, we show that the GEP (2.1) can be transformed to a QEP in $(\lambda, \mathbf{p})$:

$$Q(\lambda)\mathbf{p} := (\lambda^2 A_2 + \lambda A_1 + A_0)\mathbf{p} = \mathbf{0}, \tag{2.9}$$

where the coefficient matrices A_2, A_1 and A_0 are given by

$$A_2 = M_1 + \widehat{M}_1 \widehat{M}^{-1} \widehat{M}_1^\top + F_1 G^{-1} F_1^\top, \tag{2.10a}$$

$$A_1 = -K - \widehat{K}\widehat{M}^{-1}\widehat{M}_1^\top - \widehat{M}_1\widehat{M}^{-1}\widehat{K}^\top - EG^{-1}F_1^\top - F_1 G^{-1} E^\top, \tag{2.10b}$$

$$A_0 = \widehat{K}\widehat{M}^{-1}\widehat{K}^\top + EG^{-1}E^\top = KM^{-1}K + \widehat{E}\widehat{G}^{-1}\widehat{E}^\top. \tag{2.10c}$$

Note that the last equality in (2.10c) is not obvious and requires further explanation. The detail derivation can be found in the Appendix.

We then point out the particular matrix structures of the coefficient matrices in (2.10). They play critical roles not only in our theoretical analysis, but also in numerical computation of transmission eigenvalue problems.

THEOREM 2.1. *The coefficient matrices of the QEP* (2.9) *are symmetric and, in particular, A_2 and A_0 are symmetric positive definite. As a result, its eigenvalues are either real, but nonzero, or come in complex conjugate pairs $(\lambda, \overline{\lambda})$.*

Proof. The symmetry properties of the matrices A_2, A_1 and A_0 are obvious from (2.10). Moreover, we can further conclude that A_2 and A_0 are symmetric positive definite thanks to the fact that M_1, $\widehat{M}$ and G (see (2.5)) are symmetric positive definite. Finally, 0 is not an eigenvalue of $Q(\lambda)$ since A_0 is nonsingular. ∎

The following theorem explains the spectral relation between the QEP (2.9) as well as the GEP (2.1). More importantly, it points out that even though the kernel of the GEP (2.1) is as large as the number of boundary nodes, the derived QEP (2.9) can completely avoid the disturbance of these zero eigenvalues.

THEOREM 2.2. *It holds that*

$$\sigma(\mathcal{K} - \lambda\mathcal{M}) = \underbrace{\{0, \ldots, 0\}}_{\rho} \cup \sigma(Q(\lambda)). \tag{2.11}$$

Here, $\sigma(\cdot)$ denotes the spectrum of the associated matrix pencil.

Proof. By Theorem 2.1, the QEP (2.9) has 2ν nonzero eigenvalues of $\mathcal{K} - \lambda\mathcal{M}$. In contrast, by inspecting (2.1), it is easy to verify that

$$\begin{bmatrix} K & \mathbf{0} & E \\ \mathbf{0} & K & E \\ E^\top & -E^\top & \mathbf{0} \end{bmatrix} \begin{bmatrix} -K^{-1}E \\ -K^{-1}E \\ I_\rho \end{bmatrix} = \mathbf{0}.$$

Therefore, we conclude that the assertion of (2.11) holds. ∎

When the transmission eigenvalue problem (1.1) is associated with a constant index of refraction $n(x) = n > 0$, the corresponding QEP (2.9) can be expressed in a more concise formulation. As the index of refraction is a constant, according to the definitions in Table 1, M_n, F_n and G_n are equal to nM_1, nF_1 and nG_1, respectively. In this case, M, F, G in (2.2), $\widehat{K}$, $\widehat{M}_1$, $\widehat{M}$ in (2.3) and $\widehat{E}$, $\widehat{G}$ in (2.4) are reduced to

$$M = (n-1)M_1, \qquad F = (n-1)F_1, \qquad G = (n-1)G_1; \tag{2.12a}$$

$$\widehat{K} = K - EG_1^{-1}F_1^\top, \quad \widehat{M}_1 = M_1 - F_1 G_1^{-1} F_1^\top, \quad \widehat{M} = (n-1)\widehat{M}_1; \tag{2.12b}$$

$$\widehat{E} = E - KM_1^{-1}F_1, \quad \widehat{G} = (n-1)(G_1 - F_1^\top M_1^{-1} F_1). \tag{2.12c}$$

With the substitutions of equations (2.12) into equations (2.10), we see that

$$A_2 = M_1 + \frac{1}{n-1}(M_1 - F_1G_1^{-1}F^\top) + \frac{1}{n-1}F_1G_1^{-1}F_1^\top = \frac{n}{n-1}M_1, \tag{2.13a}$$

$$A_1 = -K - \frac{1}{n-1}(K - EG_1^{-1}F_1^\top) - \frac{1}{n-1}(K - F_1G_1^{-1}E^\top) - \frac{1}{n-1}EG_1^{-1}F_1^\top - \frac{1}{n-1}F_1G_1^{-1}E^\top = -\frac{n+1}{n-1}K, \tag{2.13b}$$

$$A_0 = \frac{1}{n-1}[KM_1^{-1}K + (E - KM_1^{-1}F_1)(G_1 - F_1^\top M_1^{-1}F_1)^{-1}(E - KM_1^{-1}F_1)^\top]. \tag{2.13c}$$

3. Spectral analysis of transmission eigenvalue problems. Based on the spirit of the proof for the existence of real eigenvalues for the continuous type transmission eigenvalue problem (1.1) in [7], in this section, we first prove the existence of positive real eigenvalues of the QEP (2.9) in some suitable interval. Moreover, in the case that the index of refraction is a constant, we will show that the associated QEP with coefficient matrices in (2.13) has at least $2(\nu - 2\rho)$ positive real eigenvalues.

3.1. Non-constant refractive index. Let $\overline{\kappa}$, $\underline{\kappa}$, $\overline{\mu}_1$, $\underline{\mu}_1$, d_M and d_m be scalars defined by

$$\overline{\kappa} := \lambda_{\max}(K), \quad \underline{\kappa} := \lambda_{\min}(K), \quad \overline{\mu}_1 := \lambda_{\max}(M_1), \quad \underline{\mu}_1 := \lambda_{\min}(M_1), \tag{3.1}$$

$$d_M := \max\left\{ \left\| \begin{bmatrix} \widehat{M}^{-\frac{1}{2}}\widehat{M}_1^\top \\ G^{-\frac{1}{2}}F_1^\top \end{bmatrix} \right\|^2, \|\widehat{M}^{-1}\|, \|G^{-1}\| \right\}, \tag{3.2}$$

$$d_m := \max\left\{ \sigma_{\min}^2 \left(\begin{bmatrix} \widehat{M}^{-\frac{1}{2}}\widehat{M}_1^\top \\ G^{-\frac{1}{2}}F_1^\top \end{bmatrix} \right), \sigma_{\min}\big(\widehat{M}^{-1}\big), \sigma_{\min}\big(G^{-1}\big) \right\}, \tag{3.3}$$

where $\lambda_{\max}(\cdot)$ and $\lambda_{\min}(\cdot)$ (respectively, $\sigma_{\max}(\cdot)$ and $\sigma_{\min}(\cdot)$) are the maximum and minimum eigenvalues (respectively, singular values) of a given matrix, respectively, and the notation $\|\cdot\|$ denotes the matrix 2-norm. Furthermore, we let

$$U_\sigma^\top \begin{bmatrix} \widehat{K}^\top \\ E^\top \end{bmatrix} V_\sigma = \begin{bmatrix} \operatorname{diag}\{\sigma_1, \ldots, \sigma_\nu\} \\ \mathbf{0}^\top \end{bmatrix} \tag{3.4}$$

be the singular value decomposition of $\begin{bmatrix} \widehat{K}^\top \\ E^\top \end{bmatrix}$ with $0 < \sigma_1 \le \cdots \le \sigma_\nu$.

THEOREM 3.1. *Assume that*

$$d_m(\underline{\mu}_1 + d_m) > d_M^2 \quad \textit{and} \quad d_M < \frac{\underline{\kappa}^2}{4(\underline{\kappa}\sigma_p + \overline{\mu}_1\sigma_p^2)} \tag{3.5}$$

for some p. Then there are at least p real eigenvalues of the QEP (2.9) in the interval $[\tau^0, \tau^*]$ and at least p real eigenvalues in $[\tau^*, \infty)$, where

$$0 < \tau^* = \frac{\underline{\kappa} - 2d_M\sigma_p}{2(\overline{\mu}_1 + d_M)}, \quad \textit{and} \quad 0 \le \tau^0 < \min\left\{ \tau^*, \left(d_m - \frac{d_M^2}{\underline{\mu}_1 + d_m} \right) \frac{\sigma_1^2}{\overline{\kappa}} \right\}. \tag{3.6}$$

Proof. Let τ be a positive parameter for representing $\lambda > 0$ in (2.9) and (2.10). From (2.10) and (2.3), we have

$$A_2 = M_1 + \begin{bmatrix}\widehat{M}_1 & F_1\end{bmatrix}\begin{bmatrix}\widehat{M}^{-1} & \mathbf{0}\\ \mathbf{0} & G^{-1}\end{bmatrix}\begin{bmatrix}\widehat{M}_1^\top\\ F_1^\top\end{bmatrix}, \tag{3.7a}$$

$$A_1 = -K - \begin{bmatrix}\widehat{K} & E\end{bmatrix}\begin{bmatrix}\widehat{M}^{-1} & \mathbf{0}\\ \mathbf{0} & G^{-1}\end{bmatrix}\begin{bmatrix}\widehat{M}_1^\top\\ F_1^\top\end{bmatrix} - \begin{bmatrix}\widehat{M}_1 & F_1\end{bmatrix}\begin{bmatrix}\widehat{M}^{-1} & \mathbf{0}\\ \mathbf{0} & G^{-1}\end{bmatrix}\begin{bmatrix}\widehat{K}^\top\\ E^\top\end{bmatrix}, \tag{3.7b}$$

$$A_0 = \begin{bmatrix}\widehat{K} & E\end{bmatrix}\begin{bmatrix}\widehat{M}^{-1} & \mathbf{0}\\ \mathbf{0} & G^{-1}\end{bmatrix}\begin{bmatrix}\widehat{K}^\top\\ E^\top\end{bmatrix}. \tag{3.7c}$$

Then, from (3.7) and (3.1)–(3.4), the Rayleigh quotient of $Q(\tau)$ in (2.9) with respect to a unit vector $\mathbf{v} \in \mathbb{R}^{\nu}$ has the following relation:

$$\begin{aligned}
&\mathbf{v}^\top Q(\tau)\mathbf{v} \\
&= \left\| \begin{bmatrix}\widehat{M}^{-\frac{1}{2}} & \mathbf{0}\\ \mathbf{0} & G^{-\frac{1}{2}}\end{bmatrix}\begin{bmatrix}\widehat{K}^\top\\ E^\top\end{bmatrix}\mathbf{v} - \tau\begin{bmatrix}\widehat{M}^{-\frac{1}{2}}\widehat{M}_1^\top\\ G^{-\frac{1}{2}}F_1^\top\end{bmatrix}\mathbf{v}\right\|^2 + \tau^2\mathbf{v}^\top M_1\mathbf{v} + \tau\mathbf{v}^\top K\mathbf{v} \\
&\geq d_m\left\|\begin{bmatrix}\widehat{K}^\top\\ E^\top\end{bmatrix}\mathbf{v}\right\|^2 - 2\tau d_M\left\|\begin{bmatrix}\widehat{K}^\top\\ E^\top\end{bmatrix}\mathbf{v}\right\| + d_m\tau^2 + \underline{\mu}_1\tau^2 - \overline{\kappa}\tau \\
&= (\underline{\mu}_1 + d_m)\left(\tau - \frac{d_M}{\underline{\mu}_1 + d_m}\left\|\begin{bmatrix}\widehat{K}^\top\\ E^\top\end{bmatrix}\mathbf{v}\right\|\right)^2 + \left(d_m - \frac{d_M^2}{\underline{\mu}_1 + d_m}\right)\left\|\begin{bmatrix}\widehat{K}^\top\\ E^\top\end{bmatrix}\mathbf{v}\right\|^2 - \overline{\kappa}\tau \\
&\geq \sigma_1^2\left(d_m - \frac{d_M^2}{\underline{\mu}_1 + d_m}\right) - \overline{\kappa}\tau
\end{aligned} \tag{3.8}$$

and we can deduce that $\mathbf{v}^\top Q(\tau)\mathbf{v} > 0$ provided $0 < \tau < \left(d_m - \frac{d_M^2}{\underline{\mu}_1 + d_m}\right)\frac{\sigma_1^2}{\overline{\kappa}}$. Since the eigenvalues of $Q(\tau)$ are continuous, from (3.8), the eigenvalue curves $\{\lambda_j(\tau)\}_{j=1}^n$ of $Q(\tau)$ with $\lambda_1(\tau) \leq \cdots \leq \lambda_n(\tau)$ are all larger than zero for $0 < \tau < \tau^0$, where τ^0 is given by (3.6).

On the other hand, from (3.1)–(3.4) and (3.7), the Rayleigh quotient of $Q(\tau)$ with respect to a unit vector $\mathbf{v} \in \operatorname{span}\{V_{\sigma,p}\}$ satisfies

$$\begin{aligned}
\mathbf{v}^\top Q(\tau)\mathbf{v} = \tau^2&\left(\mathbf{v}^\top M_1\mathbf{v} + \mathbf{v}^\top\begin{bmatrix}\widehat{M}_1 & F_1\end{bmatrix}\begin{bmatrix}\widehat{M}^{-1}\widehat{M}_1^\top\\ G^{-1}F_1^\top\end{bmatrix}\mathbf{v}\right) - \tau(\mathbf{v}^\top K\mathbf{v}) \\
&- \tau\mathbf{v}^\top\left(\begin{bmatrix}\widehat{K} & E\end{bmatrix}\begin{bmatrix}\widehat{M}^{-1}\widehat{M}_1^\top\\ G^{-1}F_1^\top\end{bmatrix} + \begin{bmatrix}\widehat{M}_1\widehat{M}^{-1} & F_1G^{-1}\end{bmatrix}\begin{bmatrix}\widehat{K}^\top\\ E^\top\end{bmatrix}\right)\mathbf{v} \\
&+ \mathbf{v}^\top\begin{bmatrix}\widehat{K} & E\end{bmatrix}\begin{bmatrix}\widehat{M}^{-1} & 0\\ 0 & G^{-1}\end{bmatrix}\begin{bmatrix}\widehat{K}^\top\\ E^\top\end{bmatrix}\mathbf{v} \\
&\leq \tau^2(\overline{\mu}_1 + d_M) - \tau\underline{\kappa} + 2\tau d_M\sigma_p + d_M\sigma_p^2
\end{aligned} \tag{3.9}$$

for $\tau > 0$. By assumption, it is clear that τ^* in (3.6) is positive and minimizes the right hand side of (3.9).

Substituting τ^* into the right-hand side of the last inequality in (3.9), we get

$$\mathbf{v}^\top Q(\tau^*)\mathbf{v} \leq -\frac{(\underline{\kappa} - 2d_M\sigma_p)^2}{4(\overline{\mu}_1 + d_M)} + d_M\sigma_p^2 < 0,$$

provided that the assumption (3.5) holds. Since σ_j is of the increasing order in j, we know that $d_M \leq \frac{\underline{\kappa}^2}{4(\underline{\kappa}\sigma_j+\overline{\mu}_1\sigma_j^2)}$ for $1 \leq j \leq p$, and hence we can deduce that $\mathbf{v}Q(\tau^*)\mathbf{v} < 0$ for $\mathbf{v} \in \operatorname{span}\{V_{\sigma,p}\}$.

For a given τ^0 satisfies (3.6) we see that the eigenvalue curves of $Q(\tau)$ in (2.9) satisfy $\lambda_j(\tau^*) < 0$ for $j = 1, \ldots, p$. It is clear that $\lambda_j(\infty) > 0$, $j = 1, \ldots, p$. By Intermediate Value Theorem, we conclude that $[\tau^0, \tau^*]$ and $[\tau^*, \infty)$, respectively, contain at least p real eigenvalues of $Q(\tau)$.
□

3.2. Constant refractive index. We now consider the case of constant index of refraction, i.e., $n(x) = n > 1$. In this situation, the coefficient matrices of the QEP (2.9) are as shown in (2.13).

For the theoretical derivation, we first note that since $G_1 - F_1^\top M_1^{-1} F_1$ is symmetric positive definite (cf. (2.5)), A_0 in (2.13c) can be further written as

$$A_0 = KM_1^{-1}K + K_0K_0^\top, \tag{3.10}$$

where $K_0 := (E - KM_1^{-1}F_1)(G_1 - F_1^\top M_1^{-1}F_1)^{-\frac{1}{2}}$ with $\operatorname{rank}(K_0) = \rho$. So, by (2.13) and (3.10), the QEP (2.9) with constant index of refraction $n > 1$ can be simplified to

$$Q_n(\lambda)\mathbf{p} := \Big[\lambda^2 nM_1 + \lambda(-n-1)K + (KM_1^{-1}K + K_0K_0^\top)\Big]\mathbf{p} = \mathbf{0}. \tag{3.11}$$

LEMMA 3.2. *Let P_2, P_1, P_0 be $\nu \times \nu$ so that $P_2^\top = P_2 \succ 0$, $P_1^\top = P_1$ and $P_0^\top = P_0 \succ 0$. Consider the linear symmetric GEP in $\tau \in \mathbb{R}$.*

$$P(\tau)\mathbf{x}(\tau) = \beta(\tau)P_0\mathbf{x}(\tau), \quad P(\tau) := -P_1 - \tau P_2. \tag{3.12}$$

Let $\beta_1(\tau) \leq \cdots \leq \beta_\nu(\tau)$ be the eigenvalue curves of the matrix pair $(P(\tau), P_0)$, and $\mathbf{x}_j(\tau)$ be the associated eigenvector satisfying $\mathbf{x}_i^\top P_0 \mathbf{x}_j = \delta_{ij}$, $i, j = 1, \ldots, \nu$, where δ_{ij} denoted the Kronecker delta. Then

(i) $\beta_j(\tau) \in \mathbb{R}$ is strictly decreasing in τ for each $j = 1, \ldots, \nu$ [13].

(ii) $(\lambda, \mathbf{x})$ is an eigenpair of the QEP

$$(\lambda^2 P_2 + \lambda P_1 + P_0)\mathbf{x} = \mathbf{0} \tag{3.13}$$

with $\mathbf{x}^\top P_0 \mathbf{x} = 1$ if and only if $(\beta(\lambda), \mathbf{x})$ is an eigenpair of (3.12) and

$$\beta(\lambda) = \frac{1}{\lambda}. \tag{3.14}$$

Proof.

(i) Given $\tau \in \mathbb{R}$. Since $P(\tau) = -P_1 - \tau P_2$ is symmetric and P_0 is symmetric positive definite, we know that all eigenvalues of (3.12) are real. From [16], the positive definiteness of P_0 implies that $\beta(\tau)$ is differentiable for all but a finite number of τ. Moreover, since, for $j = 1, \ldots, \nu$,

$$\beta_j(\tau) = \mathbf{x}_j^\top P(\tau)\mathbf{x}_j(\tau), \tag{3.15}$$

and $2\mathbf{p}_j^\top(\tau)' A_0 \mathbf{p}_j(\tau) = \frac{d}{d\tau}(\mathbf{p}_j^\top(\tau)A_0\mathbf{p}_j(\tau)) = \frac{d}{d\tau}1 = 0$, one can see that if we take the derivative of $\beta_j(\tau)$ in (3.15) with respect to τ and note that $A_2 \succ 0$, then

$$\begin{aligned}\beta_j'(\tau) &= 2\mathbf{p}_j^\top(\tau)' A(\tau)\mathbf{p}_j(\tau) + \mathbf{p}_j^\top(\tau)A'(\tau)\mathbf{p}_j(\tau) \\ &= 2\beta_j(\tau)\mathbf{p}_j^\top(\tau)' A_0\mathbf{p}_j(\tau) - \mathbf{p}_j^\top(\tau)A_2\mathbf{p}_j(\tau) = -\mathbf{p}_j^\top(\tau)A_2\mathbf{p}_j(\tau) < 0,\end{aligned}$$

which indicates that the eigenvalue curves $\beta_j(\tau)$, $j = 1, \ldots, \nu$, are strictly decreasing.

(ii) First, we note that $\lambda \neq 0$ since P_0 is invertible (cf. Theorem 2.1). The equation (3.14) follows directly from the fact that $(\lambda, \mathbf{x})$ with $\lambda \neq 0$ and $\mathbf{x}^\top P_0 \mathbf{x} = 1$ is an eigenpair of the QEP (3.13) if and only if

$$(-P_1 - \lambda P_2) = \frac{1}{\lambda} P_0 \mathbf{x}.$$

□

THEOREM 3.3. *A real eigenvalue of the QEP* (3.11) *(if it exists) must be positive.*

Proof. Consider the linear GEP

$$\Big((n+1)K - \tau n M_1\Big)\mathbf{p}(\tau) = \beta(\tau) A_0 \mathbf{p}(\tau), \quad \tau \in \mathbb{R},$$

where A_0 is defined as in (3.10). By Lemma 3.2 (ii), λ is a real eigenvalue of the QEP (3.11) if and only if $\beta_j(\lambda)$ intersects the hyperbola $y = 1/\tau$ for some j, i.e., $\beta_j(\lambda) = 1/\lambda$. When $\tau = 0$, since K and A_0 both are symmetric positive definite, and $n > 0$, we have $\beta_j(0) > 0$, $j = 1, \ldots, \nu$. Nevertheless, as shown is Lemma 3.2 (i), $\beta_j(\tau)$ is strictly decreasing in τ, for $j = 1, \ldots, \nu$, it follows that $\beta_j(\lambda) = 1/\lambda$ holds only for $\lambda > 0$. □

Consider the spectrum decomposition of (K, M_1),

$$K\mathbf{x}_i = \xi_i M_1 \mathbf{x}_i \tag{3.16a}$$

with eigenpairs $(\xi_i, \mathbf{x}_i)$ satisfying

$$0 < \xi_1 \leq \xi_2 \leq \cdots \leq \xi_\nu \quad \text{and} \quad \mathbf{x}_i^\top M_1 \mathbf{x}_j = \delta_{ij},\ i, j = 1, \ldots, \nu. \tag{3.16b}$$

THEOREM 3.4. *The QEP $Q_n(\lambda)$ in* (3.11) *has at least $2(\nu - 2\rho)$ real eigenvalues.*

To prove this theorem, we first introduce the following lemma.

LEMMA 3.5. *Consider $\nu \times \nu$ matrices $A^\top = A$, $\widehat{B} = \widehat{B}^\top \succ 0$ and $B = \widehat{B} + bb^\top$ with rank$(b) = \rho \ll \nu$. Let $\beta_1 \leq \cdots \leq \beta_\nu$ and $\widehat{\beta}_1 \leq \cdots \leq \widehat{\beta}_\nu$ be ordered eigenvalues of matrix pairs (A, B) and $(A, \widehat{B})$, respectively. Then, we have the inequalities*

$$\widehat{\beta}_j \geq \beta_j \geq \widehat{\beta}_{j-2\rho}, \quad j = 1, \ldots, \nu,$$

with $\widehat{\beta}_{j-2\rho} = -\infty$ if $j - 2\rho \leq 0$.

Proof. Let $\widehat{B} = LL^\top$ be the Cholesky decomposition of $\widehat{B}$. Then

$$(A, \widehat{B}) \overset{\text{eq.}}{\sim} (A_\ell, I_\nu), \qquad A_\ell := L^{-1} A L^{-\top}, \tag{3.17}$$

$$(A, B) \overset{\text{eq.}}{\sim} (A_\ell, (I_\nu + b_\ell b_\ell^\top)), \quad b_\ell := L^{-1} b. \tag{3.18}$$

Here, $\overset{\text{eq.}}{\sim}$ denotes the equivalence transformation between matrix pairs.

Since $I_\nu + b_\ell b_\ell^\top$ is symmetric positive definite, there exists an orthogonal Q_ℓ such that $I_\nu + b_\ell b_\ell^\top = Q_\ell \Big(I_\nu + \begin{bmatrix} D_\rho^2 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}\Big) Q_\ell^\top$ with $D_\rho^2 = \operatorname{diag}\{d_1^2, \ldots, d_\rho^2\} \succ 0$. Therefore, performing congruence transformations consecutively on the matrix pair $(A_\ell, (I_\nu + b_\ell b_\ell^\top))$ in (3.18) by matrices Q_ℓ and $J_\nu = I_\nu - \begin{bmatrix} \Delta_\rho \\ \mathbf{0} \end{bmatrix} \begin{bmatrix} I_\rho & \mathbf{0} \end{bmatrix}$ with $\Delta_\rho = \operatorname{diag}\Big\{1 - \frac{1}{\sqrt{1+d_i^2}}\Big\}_{i=1}^{\rho}$, we can obtain the following equivalence relation:

$$\begin{aligned} (A_\ell, (I_\nu + b_\ell b_\ell^\top)) &\overset{\text{eq.}}{\sim} \left(Q_\ell^\top A_\ell Q_\ell, \begin{bmatrix} I_\nu + D_\rho^2 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}\right) \\ &\overset{\text{eq.}}{\sim} \big(J_\nu Q_\ell^\top A_\ell Q_\ell J_\nu, I_\nu\big) = (Q_\ell^\top A_\ell Q_\ell - F_\ell F_\ell^\top, I_\nu), \end{aligned} \tag{3.19}$$

where $F_\ell = \left[\begin{array}{c|c} \Delta_\rho & \\ \mathbf{0} & A_\rho \end{array}\right]\left[\begin{array}{cc} \widehat{A}_\rho & I_\rho \\ I_\rho & \mathbf{0} \end{array}\right]\left[\begin{array}{cc} \Delta_\rho & \mathbf{0} \\ & A_\rho^\top \end{array}\right] \in \mathbb{R}^{\nu\times 2\rho}$ with $A_\rho = Q_\ell^\top A_\ell Q_\ell \left[\begin{array}{c} I_\rho \\ \mathbf{0} \end{array}\right] \in \mathbb{R}^{\nu\times\rho}$ and $\widehat{A}_\rho = \big[I_\rho\ \mathbf{0}\big]A_\rho \in \mathbb{R}^{\rho\times\rho}$.

Applying the interlace theorem [21, Chapter 10] to (3.17)–(3.19) we have that $\widehat{\beta}_j \geq \beta_j \geq \widehat{\beta}_{j-2\rho}$, for $j = 1, \ldots, \nu$. ∎

Proof of Theorem 3.4. Throwing away the low-rank term of $Q_n(\lambda)$ in (3.11), we define

$$\widehat{Q}_n(\lambda) := \lambda^2 n M_1 - \lambda(n+1)K + KM_1^{-1}K. \tag{3.20}$$

For $Q_n(\lambda)$ in (3.11), let $(\beta_j(\tau), \mathbf{p}_j(\tau))$, $j = 1, \ldots, \nu$, be eigenpairs of the GEP

$$\Big((n+1)K - \tau n M_1\Big)\mathbf{p}_j(\tau) = \beta_j(\tau)\Big(KM_1^{-1}K + K_0K_0^\top\Big)\mathbf{p}_j(\tau). \tag{3.21}$$

For $\widehat{Q}_n(\lambda)$ in (3.20), let $(\widehat{\beta}_j(\tau), \widehat{\mathbf{p}}_j(\tau))$, $j = 1, \ldots, \nu$, be eigenpairs of the GEP

$$\Big((n+1)K - \tau n M_1\Big)\widehat{\mathbf{p}}_j(\tau) = \widehat{\beta}_j(\tau)KM_1^{-1}K\widehat{\mathbf{p}}_j(\tau); \tag{3.22}$$

Applying Lemma 3.5 to (3.22) and (3.21) we get

$$\widehat{\beta}_j(\tau) \geq \beta_j(\tau) \geq \widehat{\beta}_{j-2\rho}(\tau). \tag{3.23}$$

According the eigendecomposition of (K, M_1) in (3.16), we can compute the 2ν's eigenvalues of $\widehat{Q}_n(\lambda)$ by solving the ν's quadratic equations $\mathbf{x}_i^\top \widehat{Q}_n(\lambda)\mathbf{x}_i = 0$, $i = 1, \ldots, \nu$. Actually, one can see that the 2ν's eigenvalues of $\widehat{Q}_n(\lambda)$ are real with values given by

$$\Big(\widehat{\lambda}_i^-, \widehat{\lambda}_i^+\Big) = \frac{1}{2n}\Big((n+1)\xi_i - \sqrt{\Delta_i}, (n+1)\xi_i + \sqrt{\Delta_i}\Big) = \Big(\frac{1}{n}\xi_i, \xi_i\Big),$$

where $\Delta_i = (n+1)^2\xi_i^2 - 4n\xi_i^2 = (n-1)^2\xi_i$ for $i = 1, \ldots, \nu$.

Now, we can connect the relation between the eigenvalues $\{\widehat{\lambda}_i^-, \widehat{\lambda}_i^+\}$ of $\widehat{Q}_n(\lambda)$ as well as the eigenvalue cures $\widehat{\beta}_j(\tau)$ in (3.22). By Lemma 3.2, $\widehat{\beta}_j(\tau)$ is strictly decreasing in τ so that $\widehat{\beta}_j(\tau)$ intersects the hyperbola $y = 1/\tau > 0$ at two points $\widehat{\lambda}_{\nu-j+1}^-$ and $\widehat{\lambda}_{\nu-j+1}^+$ with $\widehat{\beta}_j(\widehat{\lambda}_{\nu-j+1}^-) = n/\xi_{\nu-j+1}$ and $\widehat{\beta}_j(\widehat{\lambda}_{\nu-j+1}^+) = 1/\xi_{\nu-j+1}$, for $j = 1, \ldots, \nu$. From (3.23), we have

$$\beta_\nu(\widehat{\lambda}_\nu^+) \geq \beta_{\nu-1}(\widehat{\lambda}_\nu^+) \geq \cdots \geq \beta_{2\rho+1}(\widehat{\lambda}_\nu^+) \geq \widehat{\beta}_1(\widehat{\lambda}_\nu^+) = \frac{1}{\widehat{\lambda}_\nu^+} > 0.$$

Since $\beta_j(\tau)$ is also strictly decreasing, it follows that $\beta_j(\tau)$ must intersects $y = \frac{1}{\tau} > 0$ at two points, for $j = 2\rho+1, \ldots, \nu$. Therefore, we deduce that $Q_n(\lambda)$ has no less than $2(\nu - 2\rho)$ real eigenvalues. ∎

THEOREM 3.6. *The QEP* (3.11) *only has real eigenvalues when the index of refraction n is sufficiently large so that $\sqrt{n} + 1/\sqrt{n} \geq 2\theta/\xi_1$, where ξ_1 is the smallest positive eigenvalue of (K, M_1) and $\theta := \left\|\left[\begin{array}{c} M_1^{-\frac{1}{2}}K \\ K_0^\top \end{array}\right]\right\|$.*

Proof. Given $\mathbf{p} \in \mathbb{R}^\nu$ with $\mathbf{p}^\top M_1\mathbf{p} = 1$, we see that

$$\mathbf{p}^\top Q_n(\lambda)\mathbf{p} = n\lambda^2 - (n+1)(\mathbf{p}^\top K\mathbf{p})\lambda + \left\|\left[\begin{array}{c} M_1^{-\frac{1}{2}}K \\ K_0^\top \end{array}\right]\mathbf{p}\right\|^2. \tag{3.24}$$

According to the arrangement of ξ_i in (3.16b), the discriminant of (3.24) satisfies

$$(n+1)^2(\mathbf{p}^\top K\mathbf{p})^2 - 4n\left\|\begin{bmatrix} M_1^{-\frac{1}{2}}K \\ K_0^\top \end{bmatrix}\mathbf{p}\right\|^2 \geq (n+1)^2\xi_1^2 - 4n\theta^2 \geq 0,$$

provided that $\sqrt{n}+1/\sqrt{n} \geq 2\theta/\xi_1$. Thus, the QEP (3.24) has only real eigenvalues if n is sufficiently large. □

The foregoing theorems indicate the appropriate requirements of the index of refraction $n > 1$ for the QEP (3.11) to guarantee the existence of real eigenvalues. In contrast, the following theorem states that, under what conditions of n together with a specified eigenpair, the corresponding eigenvalue is complex.

THEOREM 3.7. *Let $(\lambda, \mathbf{p})$ be an eigenpair of the QEP (3.11). If $n > 1$ in $Q_n(\lambda)$ is sufficiently small with $\mathcal{O}(n-1) < \sigma_{\min}(K_0)$[1] and if the associated eigenvector $\mathbf{p}$ satisfies $\mathbf{p} \cap \operatorname{span}\{K_0\} \neq \emptyset$. Then λ is a complex eigenvalue.*

Proof. Write $\mathbf{p} = \sum_{i=1}^{\nu} \alpha_i \mathbf{x}_i$ with $\mathbf{p} \cap \operatorname{span}\{K_0\} \neq \emptyset$ and $\mathbf{p}^\top M_1 \mathbf{p} = 1$. Then, from (3.16) follows that

$$\begin{aligned}\mathbf{p}^\top Q_n(\lambda)\mathbf{p} &= n\lambda^2 - (n+1)\Big(\sum_{i=1}^{\nu}\xi_i\alpha_i^2\Big)\lambda + \sum_{i=1}^{\nu}\xi_i^2\alpha_i^2 + \mathbf{p}^\top K_0K_0^\top\mathbf{p} \\ &= [1+(n-1)]\lambda^2 - [2+(n-1)]\Big(\sum_{i=1}^{\nu}\xi_i\alpha_i^2\Big)\lambda + \sum_{i=1}^{\nu}\xi_i^2\alpha_i^2 + \mathbf{p}^\top K_0K_0^\top\mathbf{p}.\end{aligned}$$

Thus, when $\mathcal{O}(n-1) < \sigma_{\min}(K_0)$, the discriminant of $\mathbf{p}^\top Q_n(\lambda)\mathbf{p} = 0$ is negative as

$$4\Bigg[\underbrace{\Big(\sum_{i=1}^{\nu}\xi_i\alpha_i^2\Big)^2 - \sum_{i=1}^{\nu}\xi_i^2\alpha_i^2}_{=-\sum_{i<j}(\xi_i-\xi_j)^2\alpha_i^2\alpha_j^2} - \mathbf{p}^\top K_0K_0^\top\mathbf{p}\Bigg] + \mathcal{O}(n-1) < 0.$$

In other words, the eigenvalue λ of the QEP (3.11) is complex. □

4. The Secant-Type Iteration. The existence and location of positive real transmission eigenvalues are important in practice due to the fact that only a few lowest positive real eigenvalues of (1.1) are required in order to estimate the index of refraction in inverse scattering theory [2]. To this end, we develop an efficient and robust eigensolver for solving the derived QEP (2.9) to detect some *positive real* transmission eigenvalues of (1.1).

Linearization is a classical strategy to treat a QEP [25]. For example, in order to find the smallest positive eigenvalue of the QEP (2.9), one can, through the linearizing process, transform the QEP (2.9) into an equivalent GEP:

$$\begin{bmatrix} -A_1 & -A_2 \\ I_\nu & \mathbf{0} \end{bmatrix}\begin{bmatrix} \mathbf{p} \\ \lambda\mathbf{p} \end{bmatrix} = \frac{1}{\lambda}\begin{bmatrix} A_0 & \mathbf{0} \\ \mathbf{0} & I_\nu \end{bmatrix}\begin{bmatrix} \mathbf{p} \\ \lambda\mathbf{p} \end{bmatrix}. \tag{4.1}$$

On the one hand, the size of the enlarged problem (4.1) is smaller in contrast to (2.1); on the other hand, (4.1) does not suffer from the influence of the nonphysical transmission eigenvalue $\lambda = k^2 = 0$ as A_0 is symmetric positive (cf. Theorem 2.1 and Theorem 2.2).

[1] Here $\mathcal{O}$ denotes the "big O".

Then, we can compute the extreme eigenvalues and associated eigenvectors of (4.1) by applying iteration schemes, such as the Arnoldi method [1, 12]. However, the desired positive real eigenvalues may not be the extreme values owing to the possibility of the existence of complex transmission eigenvalues. Even though one can further consider the shifted-and-invert technique to detect the desired eigenvalues in some suitable region, we may still lose some positive real eigenvalues under inappropriate selections of the shift value. Moreover, in practice, an explicit factorization of the shifted operator, $\sigma^2 A_2 + \sigma A_1 + A_0$ with a shift σ, is hard to compute because the matrix $\widehat{K}\widehat{M}^{-1}\widehat{K}$ in (2.10c) is dense so that it is impossible to formally construct A_0 when ν is large.

Fortunately, based on the symmetry and positive definiteness of the matrices in (2.10), in what follows, we will propose a secant-type iteration to find a few, if any, smallest positive real eigenvalues of the QEP (2.9) via the idea in [13].

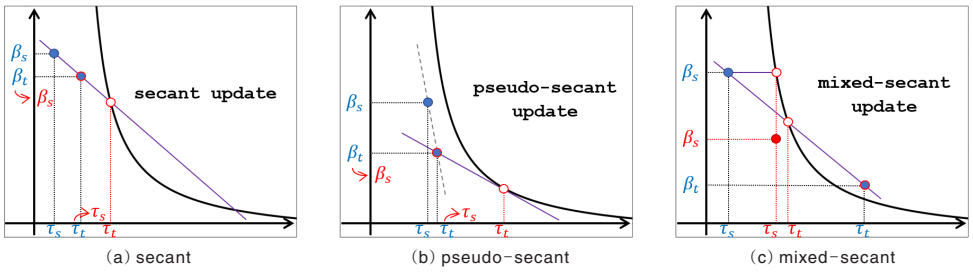
Given a $\tau \geq 0$[2], we now consider the linear symmetric eigenvalue problem in τ

$$A(\tau)\mathbf{p}(\tau) = \beta(\tau)A_0\mathbf{p}(\tau), \quad A(\tau) := -A_1 - \tau A_2. \tag{4.2}$$

By Lemma 3.2, we know that all eigenvalue curves $\beta_q(\tau)$ of (4.2) are real and are decreasing in τ for $q = \nu, \ldots, 1$. Furthermore, λ is a positive real eigenvalue of the QEP (2.9) if and only if it is a fixed point of one of the curves $1/\beta_q(\lambda)$ for some $q = \nu, \ldots, 1$, i.e.,

$$\beta_q(\lambda) = \frac{1}{\lambda}.$$

This motivates us to develop a secant-type iteration to compute some smallest positive eigenvalues of (2.9). Algorithm 4.1 summarizes the practical details.



(a) secant (b) pseudo-secant (c) mixed-secant

Fig. 1: Secant-type update

4.1. Geometrical explanation of the Secant-Type Iteration. To seek "the dth smallest" positive real eigenvalue λ_d of the QEP (2.9), the Secant-Type Iteration (Algorithm 4.1) intends to find the intersection point of "the qth largest", $q = \nu - d + 1$, eigenvalue curve $\beta_q(\tau)$ and the hyperbola $y = 1/\tau$. The iterative procedure starts with the initial guess (τ_s, τ_t) as in lines 2 and 4 of Algorithm 4.1, and successively updates the τ-pair via three appropriate strategies – *secant*, *pseudo-secant* and *mixed-secant* (see Figure 1). In all three cases, the point $(\tau_s, \beta_q(\tau_s))$ is always kept below the hyperbola (cf. Figure 1). For convenience, we omit the index q and denote β_s and

[2]Since we are only interested in finding the positive real eigenvalues of the QEP (2.9), we restrict our discussion on the case $\tau \geq 0$.

Algorithm 4.1: The Secant-Type Iteration

Input: A_2, A_1, A_0, s: number of desired eigenpairs, tol: convergent tolerance.
Output: The s smallest eigenpairs $(\lambda_d, \mathbf{p}_d)$, $d = 1, \ldots, s$, of the QEP (2.9).

for $d = 1, \ldots, s$ **do**
  set $\tau_s = \lambda_{d-1}$ with $\lambda_0 := 0$ and set $q = \nu - d + 1$;
  find the qth largest eigenpair $(\beta_q(\tau_s), \mathbf{z}_q(\tau_s))$ of $(A(\tau_s), A_0)$, where $A(\tau)$ is defined as in (4.2);
  set $\tau_t = \frac{1}{\beta_q(\tau_s)}$;
  repeat
    if $|\tau_s - \tau_t| < tol$ **then**
      return $(\lambda_d, \mathbf{p}_d) = (\tau_t, \mathbf{z}_q(\tau_t))$;
    else
      find the qth largest eigenpair $(\beta_q(\tau_t), \mathbf{z}_q(\tau_t))$ of $(A(\tau_t), A_0)$;
      set $a = \beta_q(\tau_t) - \beta_q(\tau_s)$, $b = \tau_t\beta_q(\tau_s) - \tau_s\beta_q(\tau_t)$ and $c = \tau_s - \tau_t$;
      if $\tau_t\beta_q(\tau_t) < 1$ **then**
        $\tau_s \leftarrow \tau_t$, $\beta_q(\tau_s) \leftarrow \beta_q(\tau_t)$ and $\mathbf{z}_q(\tau_s) \leftarrow \mathbf{z}_q(\tau_t)$;
        if $b^2 - 4ac > 0$ **then** % secant update
          $\tau_t \leftarrow \frac{-b+\mathtt{sign}(b)\sqrt{b^2-4ac}}{2a}$;
        else % pseudo-secant update
          $\tau_t \leftarrow \frac{1+\sqrt{1-\tau_t\beta_q(\tau_t)}}{\beta_q(\tau_t)}$;
        end
      else if $\tau_t\beta_q(\tau_t) > 1$ **then** % mixed-secant update
        find the qth largest eigenpair $(\beta_q^s, \mathbf{z}_q^s)$ of $(A(\frac{1}{\beta_q(\tau_s)}), A_0)$;
        $\tau_s \leftarrow \frac{1}{\beta_q(\tau_s)}$, $\beta_q(\tau_s) \leftarrow \beta_q^s$ and $\mathbf{z}_q(\tau_s) \leftarrow \mathbf{z}_q^s$;
        $\tau_t \leftarrow \frac{-b+\mathtt{sign}(b)\sqrt{b^2-4ac}}{2a}$;
      end
    end
  until *the dth largest positive real eigenpair* $(\lambda_d, \mathbf{p}_d)$ *is convergent*
end

β_t the values of the qth eigenvalue curve $\beta_q(\tau)$ evaluated at the points τ_s and τ_t, respectively.

• **Secant update:** The update criterion is primarily divided into two cases according to the location of the point (τ_t, β_t). When (τ_s, β_s) and (τ_t, β_t) both lie below the hyperbola, we first inspect whether the secant line through (τ_s, β_s) and (τ_t, β_t) intersects with the hyperbola curve $y = 1/\tau$. If so, we separately update the point (τ_s, β_s) by (τ_t, β_t) and update τ_t by the intersection point of which is closer to the vertical axis (see Figure 1(a)). If the difference of τ_s and τ_t is small enough, we have caught the desired eigenvalue; otherwise, we solve the GEP $(A(\tau_t), A_0)$ to obtain β_t and continue the step of the next iteration. This strategy is named by the *secant update*.

• **Pseudo-secant update:** The hyperbola and the secant line through the points (τ_s, β_s) and (τ_t, β_t), however, may not always intersect each other. In this case, the classical secant iteration, such as in [13], may update τ_t by the fixed-point iteration (or basic iteration), i.e., $\tau_t \leftarrow \frac{1}{\beta_t}$, for the next iterative process. Here, to accelerate

the convergence behavior, we modify this procedure by a *pseudo-secant update*. From Figure 1(b), one can see that we "create" a secant line from the point (τ_t, β_t) to an unknown point $(\tau, \frac{1}{\tau})$ with $\tau > \tau_t$ so that this secant line is tangent to the hyperbola at the point $(\tau, \frac{1}{\tau})$. Once, the unknown τ on the hyperbola is solved with $\tau > \tau_t$, we then update $(\tau_s, \beta_s) \leftarrow (\tau_t, \beta_t)$ and $\tau_t \leftarrow \tau$, respectively.

• **Mixed-secant update:** It is surely possible that we may encounter the case $\tau_t\beta_t > 1$ (as in Figure 1(c)). In this case, we know that the line from (τ_s, β_s) to (τ_t, β_t) must intersect the hyperbola and, moreover, the desired eigenvalue must be located between τ_s and τ_t. Therefore, we update τ_t using the strategy of the secant update. In addition, we also use the fixed-point iteration to update $\tau_s \leftarrow \frac{1}{\beta_s}$. In order to further update β_s, we need to solve the qth largest eigenpair of GEP $(A(\frac{1}{\beta_s}), A_0)$. On the one hand, such an update can maintain the updated (τ_s, β_s) is still below the hyperbola (since the eigenvalue curves are strictly decreasing). On the other hand, the fixed point iteration can be viewed as a pseudo-secant iteration by creating a secant line from (τ_s, β_s) to the point $(\frac{1}{\beta_s}, \beta_s)$ on the hyperbola. From this viewpoint, we call this update procedure the *mixed-secant update*.

4.2. Practical implementations. At the first glance, in order for the Secant-Type Iteration to solve the QEP (2.9), we have to generate the coefficient matrices A_2, A_1 and A_0 in (2.10), respectively. However, it is not possible to construct these matrices beforehand as the resulting matrices may be dense. Instead, we consider how to perform the matrix-vector multiplications and how to solve some linear systems through these matrices in (2.2)–(2.4) based on suitable eigensolvers to the symmetric definite GEP (4.2) in lines 3, 9 and 19 of Algorithm 4.1.

The generalized Lanczos scheme [26] is an efficient algorithm to solve the symmetric definite GEP (4.2) as A_0 is symmetric positive definite and, for a given $\tau \geq 0$, $A(\tau)$ is symmetric. Thus, in order to generate the (generalized) Lanczos vectors, we are required to compute the matrix-vector multiplications $A_2\mathbf{q}$, $A_1\mathbf{q}$ and $A_0\mathbf{q}$, and to solve the linear system $A_0\mathbf{x} = \mathbf{b}$ for given ν-vectors $\mathbf{q}$ and $\mathbf{b}$.

In the implementation, we will generate the matrices in (2.2)–(2.4) in advance and use the formulas (2.10) to compute the matrix-vector multiplications. For example, the multiplication of A_0 and $\mathbf{q}$, via the MATLAB notation, is to compute

$$\widehat{\mathtt{K}}*(\widehat{\mathtt{M}}\backslash(\widehat{\mathtt{K}}'*\mathbf{q})) + \widehat{\mathtt{E}} * (\mathtt{G}\backslash(\widehat{\mathtt{E}}'*\mathbf{q})).$$

On the other hand, to solve the linear system $A_0\mathbf{x} = \mathbf{b}$, we adopt the representation of A_0 in the second equality of (2.10c) and apply the Sherman-Morrison-Woodbury formula [12] to write the inverse of A_0 as follows

$$\begin{aligned} A_0^{-1} &= (KM^{-1}K + \widehat{E}\widehat{G}^{-1}\widehat{E}^\top)^{-1} \\ &= K^{-1}MK^{-1} - K^{-1}MK^{-1}\widehat{E}(\widehat{G} + \widehat{E}^\top K^{-1}MK^{-1}\widehat{E})^{-1}\widehat{E}^\top K^{-1}MK^{-1} \\ &= K^{-1}(M - Z_0C_0^{-1}Z_0^\top)K^{-1}, \end{aligned} \tag{4.3}$$

where $Z_0 = MK^{-1}\widehat{E}$ and $C_0 = \widehat{G} + (K^{-1}\widehat{E})^\top Z_0$. Finally, we remark that since $M \succ 0$ and $G \succ 0$, we have $C_0 \succ 0$. Therefore, to further improve the performance of solving the linear system $A_0\mathbf{x} = \mathbf{b}$ by (4.3), we can compute the Cholesky of K, M and $\widehat{G}$, respectively, before calling the Secant-Type Iteration.

5. Numerical experiments. In what follows, we will demonstrate that the derived QEPs (2.9) and (3.11) enable us to consider more finer mesh discretization to

overcome the limitation on the number of degree of freedoms in [10]. Furthermore, we will also show the efficiency and robustness of Algorithm 4.1 for computing some desired positive real transmission eigenvalues.

The numerical experiments in this paper are carried out using MATLAB R2013a on a MacBook Pro Retina with 2.6GHz Intel Core i5 processor and 8GB of RAM. The computed transmission eigenvalues by means of Algorithm 4.1 are consistent with values obtained by running the MATLAB codes in [14] for the attached coarse mesh data.

In all examples, we compute the first four positive real eigenvalues of the QEP (2.9) through Algorithm 4.1. The triangulation of a given domain was generated by a MATLAB toolbox called DistMesh [22]. The stopping criterion *tol* in Algorithm 4.1 and the tolerance by calling the generalized Lanczos scheme [26] are set by 10^{-6} and 10^{-14}, respectively. For a given mesh size, we use λ_j^h to denote the jth smallest positive real eigenvalue of the QEP (2.9) computed by Algorithm 4.1 and use $k_j^h := \sqrt{\lambda_j^h}$ to indicate the corresponding transmission eigenvalue, $j = 1, 2, 3, 4$.

5.1. Model problems with the constant index of refraction $n(x) = 16$. Setting $n = 16$, we first compute the four positive real transmission eigenvalues on various domains: (i) a disk centered at $(0, 0)$ with radius $1/2$; (ii) a unit square centered at the origin; (iii) a triangle with vertices $(-\frac{\sqrt{3}}{2}, -\frac{1}{2})$, $(\frac{\sqrt{3}}{2}, -\frac{1}{2})$ and $(0, 1)$; (iv) a dumbbell consisting of two disks, both with radius $1/2$ and centered at (-1,0) and (1,0) respectively, connected by a rectangular channel centered at $(0, 0)$ with the width 2 and the height 1; and (v) a peanut-like region enclosed by the equation $x_1^2 + x_2^2 = \frac{1}{4} + \frac{x_1^2}{x_1^2+x_2^2}$ (see Figure 2). The triangulation of each domain is of the regular mesh size $h \approx 0.004$.

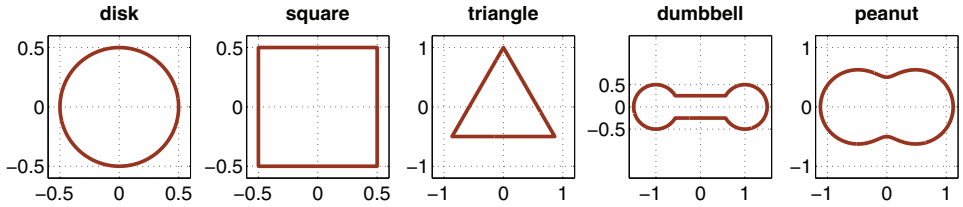


Fig. 2: Model domains

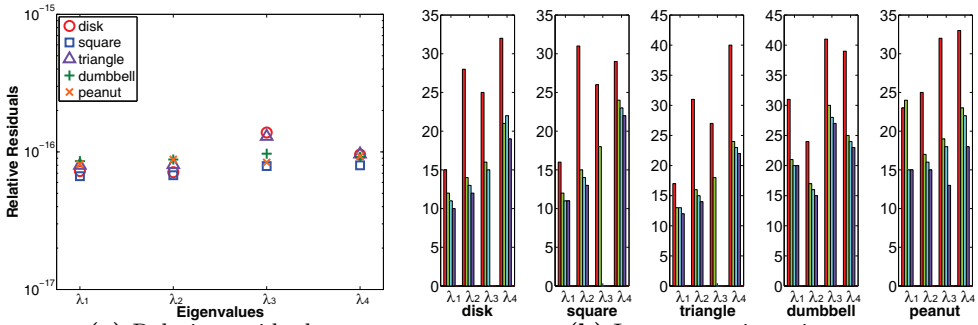
The first four desired transmission eigenvalues for each domain are shown in Table 2 and the contours plots of the associated **u**-vectors as well as **v**-vectors, which have the connections (2.7) and (2.8) together with each eigenpair $(\lambda, \mathbf{p})$, are given in Figure 4. Note that A_2, A_1, A_0 in the QEP (3.11) are $\nu \times \nu$ matrices and K_0 in (3.10) is a $\nu \times \rho$ matrix. Figure 3 presents the relative residuals

$$\frac{\|(\lambda^2 A_2 + \lambda A_1 + A_0)\frac{\mathbf{p}}{\|\mathbf{p}\|_2}\|_2}{|\lambda|^2|n(x)|\|M_1\|_2 + |\lambda||n(x)+1|\|K\|_2 + \|A_0\|_2}, \tag{5.1}$$

with $n(x) = 16$, and the inner-outer iterations for computing the desired eigenvalues of each domain. Here, the so-called "outer iteration" is the iteration numbers of the Secant-Type Iteration; and the "inner iteration" records the number of steps (i.e., the dimension of the associated Krylov subspace) by calling the generalized Lanczos scheme [26] to solve the GEPs in lines 3, 9 and 19 of Algorithm 4.1.

Domain	Degree of Freedoms		Transmission Eigenvalues			
	ν	ρ	k_1^h	k_2^h	k_3^h	k_4^h
disk	55,901	780	1.988092	2.613109	2.613123	3.226967
square	71,321	1,076	1.879649	2.444358	2.444358	2.866634
triangle	93,114	1,302	1.818525	2.287172	2.287173	2.837825
dumbbell	149,051	1,871	1.961928	1.961985	2.517941	2.518188
peanut	168,548	1,492	1.452506	1.503795	1.703846	1.987087

Table 2: The first four transmission eigenvalues with $n(x) = 16$ and $h \approx 0.004$.

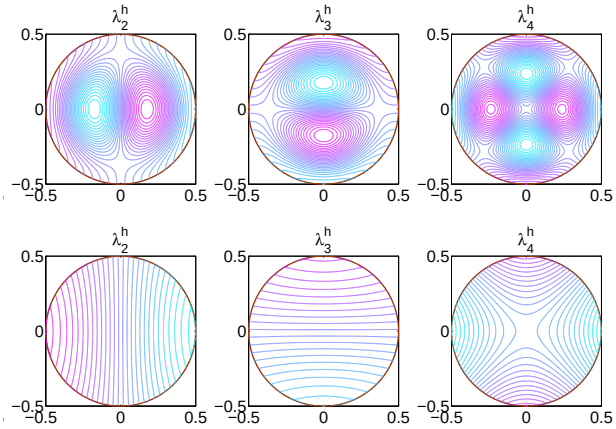


(a) Relative residuals **(b)** Inner-outer iterations

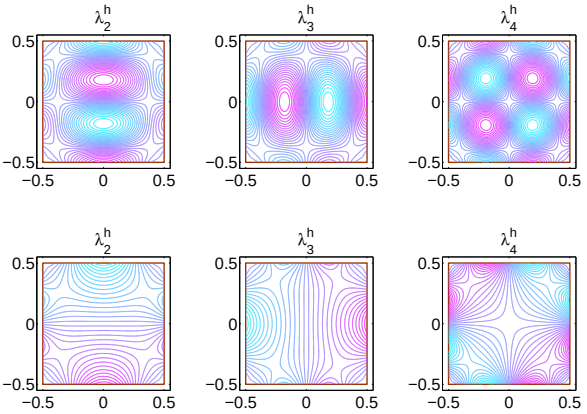
Fig. 3: (a) Relative residuals (5.1) of the QEP (3.11) with the desired approximate eigenpairs $(\lambda, \mathbf{p})$ for the case $n(x) = 16$. (b) The number of bars represents iteration numbers of the loop 5–24 in Algorithm 4.1; while the height of each bar is the number of steps by calling the generalized Lanczos scheme [26].

The numerical experiments reveal that, even we refine the mesh size so that $h \approx 0.004$ is less than one-tenth of the mesh sizes in [10, 24], the novel Secant-Type Iteration can *accurately* and *efficiently* compute some lowest positive real transmission eigenvalues of the problem (1.1). The relative residuals are roughly of the order `eps` and the Secant-Type Iteration can fast capture the desired eigenvalues with iteration numbers no more than four steps. From the observation of the numbers of outer iterations for computing λ_3^h on the domains *square* and *triangle*, we see that the Secant-Type Iteration can quickly capture λ_3^h with 2 iterations. This is because that the difference between λ_2^h and λ_3^h are very small (see Table 2) so that when λ_2^h is convergent, it is indeed a good initial guess τ_s instead of setting $\tau_s = 0$ for finding λ_3^h. Therefore, when we need to compute a few transmission eigenvalues, the experiment indicates that the latest computed approximate eigenvalue would be a good initial value for finding the next desired one (see line 2 of Algorithm 4.1), especially when the adjoint eigenvalues are very close to each other.

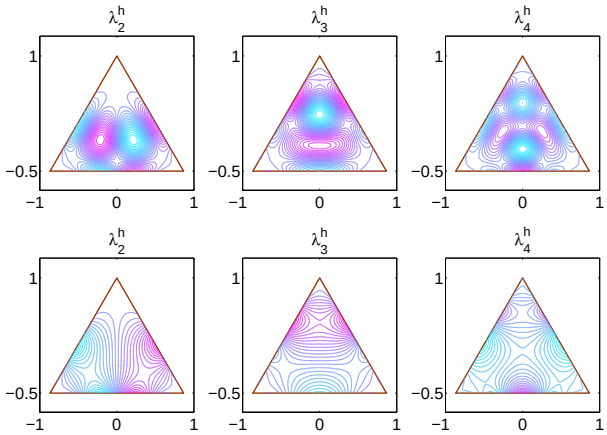
5.2. Model problems with non-constant $n(x)$. Next, we consider the case when the index of refraction $n(x)$ is not a constant. To compare with the known results presented in [23, 24, 14], we consider the domains *disk* and *square* as described



(a) Disk domain.

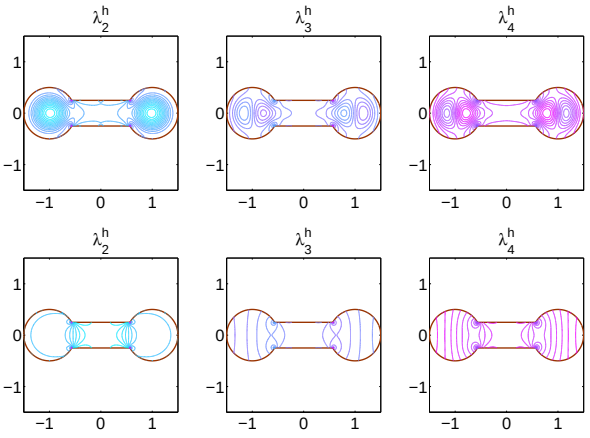


(b) Square domain.

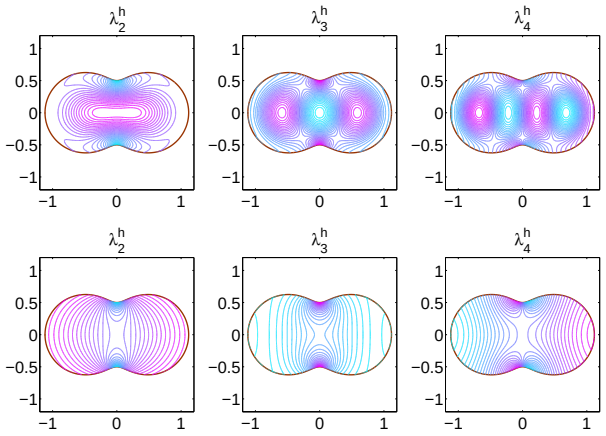


(c) Triangle domain.

Fig. 4: The contour plots of the **u**-vectors (the first row) and the **v**-vectors (the second row) corresponding to the transmission eigenvalues with $n(x) = 16$ and $h \approx 0.004$.



(d) Dumbbell domain.



(e) Peanut domain.

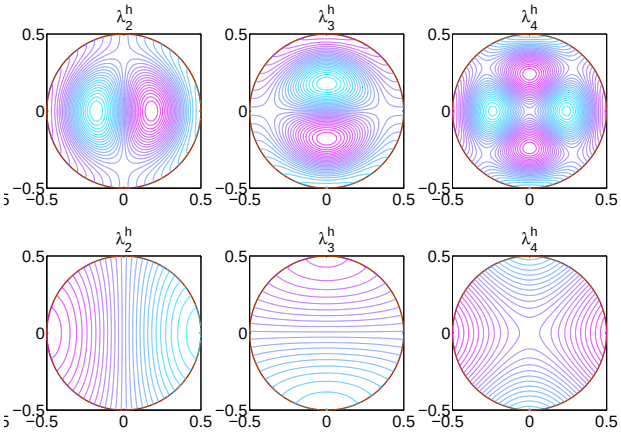
Fig. 4: The contour plots of the **u**-vectors (the first row) and the **v**-vectors (the second row) corresponding to the transmission eigenvalues with $n(x) = 16$ and $h \approx 0.004$.

in subsection 5.1, and choose the indices of refraction as $8 + 4|x|$ and $8 + x_1 - x_2$, respectively. The mesh sizes and triangulations are kept the same as in Table 2. The numerical results are shown in Table 3 and the contour plots of the **u**-vectors and **v**-vectors corresponding to each transmission eigenvalue are presented in Figure 5.

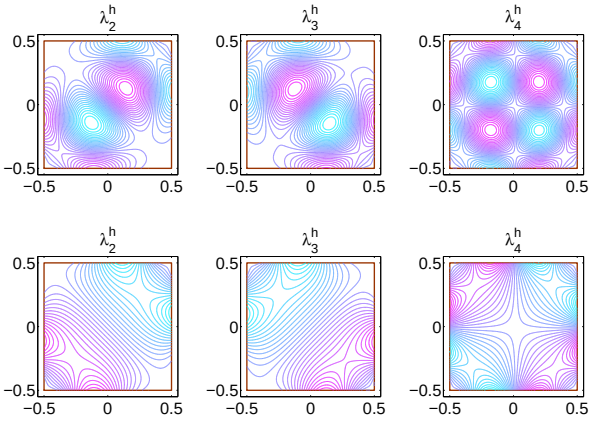
5.3. Efficiency and robustness of the Secant-Type Iteration. For the transmission eigenvalue problem (1.1), there are two major advantages on considering the QEP (2.9) and further on solving (2.9) by the Secant-Type Iteration method for the parameterized GEPs (4.2): (i) The influence of nonphysical zero eigenvalues of (1.1) can be eliminated; (ii) Positive real transmission eigenvalues can be computed, avoiding any complex ones. To demonstrate these superiorities, we compare

Domain	$n(x)$	Transmission Eigenvalues			
		k_1^h	k_2^h	k_3^h	k_4^h
disk	$8+4\|x\|$	2.759592	3.527535	3.527555	4.308419
square	$8+x_1-x_2$	2.822306	3.538893	3.539185	4.118040

Table 3: Transmission eigenvalues for non-constant index of refraction $n(x)$ with $h \approx 0.004$.



(a) Disk domain with $n(x) = 8 + 4|x|$.



(b) Square domain with $n(x) = 8 + x_1 - x_2$.

Fig. 5: The contour plots of the **u**-vectors (the first row) and the **v**-vectors (the second row) corresponding to the transmission eigenvalues with $h \approx 0.004$.

the numerical results of the QEP (2.9) with those of its linearized GEP (4.1).

In this example, we consider the discretization on the disk domain centered at $(0,0)$ with radius 0.5. The mesh size h is set to be 0.002 which produces a discrete triangular mesh with interior nodes $\nu = 225,134$ and boundary points $\rho = 1,571$. Moreover, the index of refraction is taken by $n(x) = 1.2$.

As mentioned in section 4, it is difficult to apply the shifted-and-invert technique for the QEP (2.9) or the corresponding linearized GEP (4.1) since the resulting matrices may be dense. Thus, we first rewrite the linearized GEP (4.1) to the following standard eigenvalue problem

$$\begin{bmatrix} -A_0^{-1}A_1 & -A_0^{-1}A_2 \\ I_\nu & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{p} \\ \lambda\mathbf{p} \end{bmatrix} = \frac{1}{\lambda} \begin{bmatrix} \mathbf{p} \\ \lambda\mathbf{p} \end{bmatrix}. \tag{5.2}$$

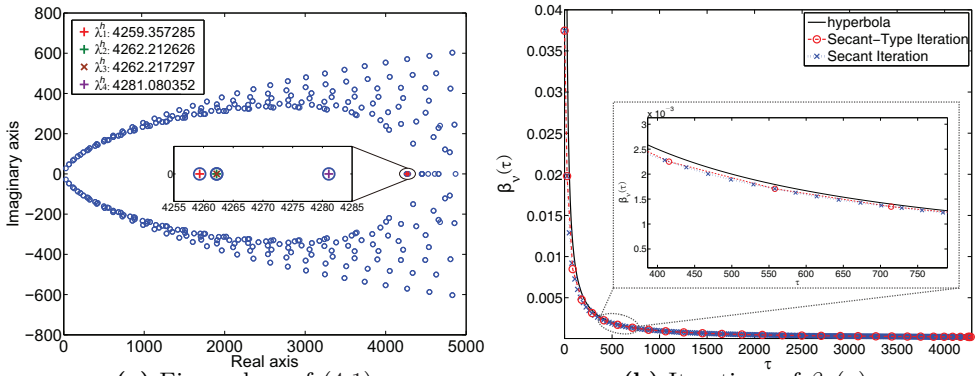
and then call the MATLAB built-in function `eigs` to compute the first 600 largest magnitude eigenvalues of (5.2) with a function handle computing matrix-vector products and solving linear systems without explicitly forming coefficient matrices (cf. subsection 4.2). The distribution of these eigenvalues are shown in Figure 6 (a). As a comparison, we implement the Secant-Type Iteration on the parameterized GEP (4.2) induced by (2.9) to compute the first four lowest positive real transmission eigenvalues λ_j^h, $j = 1,2,3,4$. The values of the approximate eigenvalues are also marked in Figure 6 (a) and they are exactly the first four positive real eigenvalues computed by (5.2) (with difference less than 10^{-7}). Moreover, we also give the contour plots of the $\mathbf{u}$-vectors and $\mathbf{v}$-vectors corresponding to each transmission eigenvalue in Figure 7.

To demonstrate the efficiency of Algorithm 4.1, we compare the iteration numbers of the Secant-Type Iteration with those of the *classical secant iteration*. The classical secant iteration will use the secant update (line 14 of Algorithm 4.1) when the hyperbola and the secant line pass through $(\tau_s, \beta_q(\tau_s))$ and $(\tau_t, \beta_q(\tau_t))$ have intersection points; otherwise, $(\tau_s, \beta_q(\tau_s)) \leftarrow (\tau_t, \beta_q(\tau_t))$ and $\tau_t \leftarrow \frac{1}{\beta_t}$. Table 4 records the iteration numbers by performing these two methods to compute the first four positive real transmission eigenvalues, and Figure 6 (b) presents the iteration processes of the eigenvalue curve $\beta_\nu(\tau)$ for finding λ_1.

• **Robustness:** From Figure 6 (a), we see that, when $n(x) = 1.2$, the transmission eigenvalue problem (1.1) has numerous complex eigenvalues, and the lowest positive real transmission eigenvalues are very far away from the origin. On the one hand, since the shifted-and-invert technique cannot be used to improve the efficiency, to solve the enlarged eigenvalue problem (5.2) will additionally compute a lot of unwanted complex eigenvalues. On the other hand, even if we can apply the shifted-and-invert technique, inappropriate choices of shift values may lose some desired real eigenvalues. In contrast, using the novel Secant-Type Iteration to solve the parameterized GEP (4.2) can accurately and robustly capture the wanted eigenvalues without losing the actual information.

• **Efficiency:** Table 4 shows that the classical secant iteration costs much more iteration processes, compared with the Secant-Type Iteration, to compute the desired eigenvalues. This is due to that, from Figure 6 (b), the eigenvalue curves are very close to the hyperbola which increases the difficulty for finding the intersection points. To our experiment, the Secant-Type Iteration takes 35 iteration numbers to find the first positive real eigenvalue of the QEP (2.9), but, the classical secant iteration requires 575 iteration steps, which is more than 16 times compared with our method, to capture this value. This indicates that the strategies of the pseudo-secant update and

the mixed-secant update in Algorithm 4.1 can indeed accelerate the convergence rate and reveals the efficiency of the novel Secant-Type Iteration proposed in this paper.



(a) Eigenvalues of (4.1) **(b)** Iterations of $\beta_\nu(\tau)$

Fig. 6: (a) The eigenvalue distribution of the GEP (4.1) with $n(x) = 1.2$ and the first four smallest positive real eigenvalues. (b) The iteration processes of the eigenvalue curve $\beta_\nu(\tau)$ computed by Secant-Type Iteration and the classical secant iteration.

Method	Iteration Numbers			
	λ_1^h	λ_2^h	λ_3^h	λ_4^h
Secant-Type Iteration	35	8	6	14
Classical Secant Iteration	575	11	4	63

Table 4: Iteration numbers by running the Secant-Type Iteration and the classical secant iteration to compute the first four positive transmission eigenvalues on the disk domain with $n(x) = 1.2$ and $h \approx 0.002$.

6. Conclusions. The study of efficient eigensolvers for the transmission eigenvalue problem is a challenging and important issue. To this end, based on the continuous finite element discretization method in [10], we derive a symmetric QEP induced from the GEP (2.1) to exclude the influence of nonphysical zero eigenvalues so as to detect a few desired transmission eigenvalues by existing iterative methods instead of finding the whole spectra of (2.1). According the derived QEP, we analyze various existence intervals of positive real eigenvalues and indicate some sufficient conditions for the possibility of complex eigenvalues.

In order to capture the positive real transmission eigenvalues which are of practical interest in the inverse scattering theory, we further transform the QEP to the symmetric definite GEP (4.2) with a parameter $\tau \geq 0$ so that we can exactly avoid any complex transmission eigenvalues and find the *positive real* eigenvalues by solving the intersection points of the hyperbola as well as the eigenvalue curves of (4.2). To achieve this goal, we develop a novel Secant-Type Iteration method (Algorithm 4.1) which is a modification of the classical secant iteration by introducing the pseudo-secant update in order to accelerate the convergence rate.

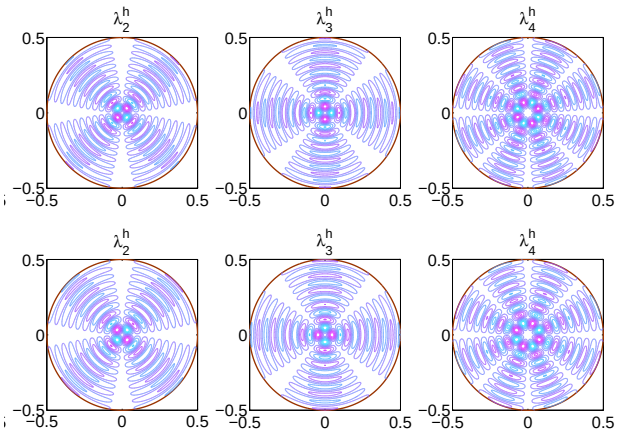


Fig. 7: The contour plots of the **u**-vectors (the first row) and the **v**-vectors (the second row) corresponding to the transmission eigenvalues of the disk domain with $n(x) = 1.2$ and $h \approx 0.002$.

Numerical examples demonstrate that we can consider more finer mesh based on the new derived QEP to remedy the limits of the number of degree of freedoms as appeared in [10]. Furthermore, via several experiments on various domains with different indices of refraction and additional comparisons with other numerical solvers, we can conclude that the novel Secant-Type Iteration method can accurately and efficiently compute the desired transmission eigenvalues. More importantly, it is a robust method for finding the positive real transmission eigenvalues even though the original problem (1.1) has numerous cluster of complex eigenvalues.

Appendix. In what follows, we show that the equality appearing in (2.10c) holds, that is, we prove that

$$A_0 := \widehat{K}\widehat{M}^{-1}\widehat{K} + EG^{-1}E^\top = KM^{-1}K + \widehat{E}\widehat{G}^{-1}\widehat{E}^\top, \tag{A.1}$$

where $\widehat{K}$, $\widehat{M}$ and $\widehat{E}$, $\widehat{G}$ are defined as in (2.3) and (2.4), respectively.

Proof of (A.1). First, according to the definitions of $\widehat{M}$ and $\widehat{G}$ in (2.3) and (2.4) respectively, the Sherman-Morrison-Woodbury formula [12] implies that

$$\widehat{M}^{-1} = (M - FG^{-1}F^\top)^{-1} = M^{-1} + M^{-1}F\widehat{G}^{-1}F^\top M^{-1}, \tag{A.2a}$$

$$\widehat{G}^{-1} = (G - F^\top M^{-1}F)^{-1} = G^{-1} + G^{-1}F^\top\widehat{M}^{-1}FG^{-1}. \tag{A.2b}$$

Furthermore, we have the equality

$$\widehat{G}G^{-1}F^\top = (G - F^\top M^{-1}F)G^{-1}F^\top = F^\top M^{-1}(M - FG^{-1}F^\top) = F^\top M^{-1}\widehat{M},$$

or equivalently,

$$G^{-1}F^\top\widehat{M}^{-1} = \widehat{G}^{-1}F^\top M^{-1}. \tag{A.3}$$

From (2.13c) and (2.3), we can rewrite A_0 as follows:

$$
\begin{aligned}
A_0 &= \widehat{K}\widehat{M}^{-1}\widehat{K}^\top + EG^{-1}E^\top \\
&= (K - EG^{-1}F^\top)\widehat{M}^{-1}(K - FG^{-1}E^\top) + EG^{-1}E^\top \\
&= \begin{bmatrix} K - EG^{-1}F^\top & E \end{bmatrix} \begin{bmatrix} \widehat{M}^{-1} & \mathbf{0} \\ \mathbf{0} & G^{-1} \end{bmatrix} \begin{bmatrix} K - FG^{-1}E^\top \\ E^\top \end{bmatrix} \\
&= \begin{bmatrix} K & E \end{bmatrix} \begin{bmatrix} I & \mathbf{0} \\ -G^{-1}F^\top & I \end{bmatrix} \begin{bmatrix} \widehat{M}^{-1} & \mathbf{0} \\ \mathbf{0} & G^{-1} \end{bmatrix} \begin{bmatrix} I & -FG^{-1} \\ \mathbf{0} & I \end{bmatrix} \begin{bmatrix} K \\ E^\top \end{bmatrix} \\
&= \begin{bmatrix} K & E \end{bmatrix} \begin{bmatrix} \widehat{M}^{-1} & -(G^{-1}F^\top\widehat{M}^{-1})^\top \\ -G^{-1}F^\top\widehat{M}^{-1} & \widehat{G}^{-1} \end{bmatrix} \begin{bmatrix} K \\ E^\top \end{bmatrix},
\end{aligned} \tag{A.4}
$$

Substituting (A.3) into (A.4) and replacing $\widehat{M}^{-1}$ by (A.2a), we then perform the congruence transformation on the middle matrix in the last equality of (A.4) with the matrix $\left[\begin{smallmatrix} I & \mathbf{0} \\ F^\top M^{-1} & I \end{smallmatrix}\right]$, to get

$$
\begin{aligned}
&\begin{bmatrix} I & M^{-1}F \\ \mathbf{0} & I \end{bmatrix} \begin{bmatrix} \widehat{M}^{-1} & -(G^{-1}F^\top\widehat{M}^{-1})^\top \\ -G^{-1}F^\top\widehat{M}^{-1} & \widehat{G}^{-1} \end{bmatrix} \begin{bmatrix} I & \mathbf{0} \\ F^\top M^{-1} & I \end{bmatrix} \\
= &\begin{bmatrix} I & M^{-1}F \\ \mathbf{0} & I \end{bmatrix} \begin{bmatrix} M^{-1} + M^{-1}F\widehat{G}^{-1}F^\top M^{-1} & -M^{-1}F\widehat{G}^{-1} \\ -\widehat{G}^{-1}F^\top M^{-1} & \widehat{G}^{-1} \end{bmatrix} \begin{bmatrix} I & \mathbf{0} \\ F^\top M^{-1} & I \end{bmatrix} \\
= &\begin{bmatrix} M^{-1} & \mathbf{0} \\ \mathbf{0} & \widehat{G}^{-1} \end{bmatrix}.
\end{aligned}
$$

As a result, (A.4) can be transformed by

$$
\begin{aligned}
A_0 &= \begin{bmatrix} K & E \end{bmatrix} \begin{bmatrix} \widehat{M}^{-1} & -(G^{-1}F^\top\widehat{M}^{-1})^\top \\ -G^{-1}F^\top\widehat{M}^{-1} & \widehat{G}^{-1} \end{bmatrix} \begin{bmatrix} K \\ E^\top \end{bmatrix} \\
&= \begin{bmatrix} K & E \end{bmatrix} \begin{bmatrix} I & -M^{-1}F \\ \mathbf{0} & I \end{bmatrix} \begin{bmatrix} M^{-1} & \mathbf{0} \\ \mathbf{0} & \widehat{G}^{-1} \end{bmatrix} \begin{bmatrix} I & \mathbf{0} \\ -F^\top M^{-1} & I \end{bmatrix} \begin{bmatrix} K \\ E^\top \end{bmatrix} \\
&= \begin{bmatrix} K & E - KM^{-1}F \end{bmatrix} \begin{bmatrix} M^{-1} & \mathbf{0} \\ \mathbf{0} & \widehat{G}_1 \end{bmatrix} \begin{bmatrix} K \\ E^\top - F^\top M^{-1}K \end{bmatrix} \\
&= KM^{-1}K + (E - KM^{-1}F)(G - F^\top M^{-1}F)^{-1}(E^\top - F^\top M^{-1}K) \\
&= KM^{-1}K + \widehat{E}\widehat{G}^{-1}\widehat{E}^\top,
\end{aligned}
$$

which completes the proof of (A.1). ∎

REFERENCES

[1] Z. Bai, J. Demmel, J. Dongarra, A. Ruhe, and H. van der Vorst, *Templates for the Solution of Algebraic Eigenvalue Problems: A Practical Guide*, SIAM, Philadelphia, PA, 2000.

[2] F. Cakoni, M. Çayören, and D. Colton, *Transmission eigenvalues and the nondestructive testing of dielectrics*, Inverse Probl., 24 (2008), p. 065016.

[3] F. Cakoni, D. Colton, and H. Haddar, *On the determination of Dirichlet or transmission eigenvalues from far field data*, C. R. Math., 348 (2010), pp. 379–383.

[4] F. Cakoni, D. Colton, and P. Monk, *On the use of transmission eigenvalues to estimate the index of refraction from far field data*, Inverse Probl., 23 (2007), pp. 507–522.
[5] F. Cakoni, D. Colton, P. Monk, and J. Sun, *The inverse electromagnetic scattering problem for anisotropic media*, Inverse Probl., 26 (2010), p. 074004.
[6] F. Cakoni, D. Gintides, and H. Haddar, *The existence of an infinite discrete set of transmission eigenvalues*, SIAM J. Math. Anal., 42 (2010), pp. 237–255.
[7] F. Cakoni and H. Haddar, *On the existence of transmission eigenvalues in an inhomogeneous medium*, Appl. Anal., 88 (2009), pp. 475–493.
[8] F. Cakoni and H. Haddar, *Transmission eigenvalues in inverse scattering theory*, in Inverse Problems and Applications: Inside Out II, G. Uhlmann, ed., vol. 60 of MSRI Publications, Cambridge University Press, 2012, pp. 527–578.
[9] D. Colton and R. Kress, *Inverse Acoustic and Electromagnetic Scattering Theory*, vol. 93 of Applied Mathematical Sciences, Springer, New York, 3rd ed., 2013.
[10] D. Colton, P. Monk, and J. Sun, *Analytical and computational methods for transmission eigenvalues*, Inverse Probl., 26 (2010), p. 045011.
[11] D. Colton, L. Päivärinta, and J. Sylvester, *The interior transmission problem*, Inverse Probl. Imaging, 1 (2007), pp. 13–28.
[12] G. H. Golub and C. F. Van Loan, *Matrix Computations*, The Johns Hopkins University Press, Baltimore, MD, 4th ed., 2012.
[13] J.-S. Guo, W.-W. Lin, and C.-S. Wang, *Numerical solutions for large sparse quadratic eigenvalue problems*, Linear Alg. Appl., 225 (1995), pp. 57–89.
[14] X. Ji, J. Sun, and T. Turner, *Algorithm 922: A mixed finite element method for Helmholtz transmission eigenvalues*, ACM Trans. Math. Softw., 38 (2012), pp. 29:1–29:8.
[15] X. Ji, J. Sun, and H. Xie, *A multigrid method for Helmholtz transmission eigenvalue problems*, J. Sci. Comput., (2013), pp. 1–19.
[16] T. Kato, *Perturbation Theory for Linear Operators*, vol. 132, Springer-Verlag, Berlin, 2nd ed., 1980.
[17] A. Kirsch, *On the existence of transmission eigenvalues*, Inverse Probl. Imaging, 3 (2009), pp. 155–172.
[18] C. B. Moler and G. W. Stewart, *An algorithm for generalized matrix eigenvalue problems*, SIAM J. Numer. Anal., 10 (1973), pp. 241–256.
[19] P. Monk and J. Sun, *Finite element methods for Maxwell's transmission eigenvalues*, SIAM J. Sci. Comput., 34 (2012), pp. B247–B264.
[20] L. Päivärinta and J. Sylvester, *Transmission eigenvalues*, SIAM J. Math. Anal., 40 (2008), pp. 738–753.
[21] B. N. Parlett, *The Symmetric Eigenvalue Problem*, SIAM, Philadelphia, PA, 1998.
[22] P.-O. Persson and G. Strang, *A simple mesh generator in MATLAB*, SIAM Rev., 46 (2004), pp. 329–345.
[23] J. Sun, *Estimation of transmission eigenvalues and the index of refraction from Cauchy data*, Inverse Probl., 27 (2011), p. 015009.
[24] J. Sun, *Iterative methods for transmission eigenvalues*, SIAM J. Numer. Anal., 49 (2011), pp. 1860–1874.
[25] F. Tisseur and K. Meerbergen, *The quadratic eigenvalue problem*, SIAM Rev., 43 (2001), pp. 235–286.
[26] H. A. van der Vorst, *A generalized Lanczos scheme*, Math. Comp., 39 (1982), pp. 559–561.