Journal of Scientific Computing (2026) 106:61
https://doi.org/10.1007/s10915-026-03192-0

A New Numerical Method for Scalar Eigenvalue Problems in Heterogeneous, Dispersive, Sign-Changing Materials

Martin Halla[1] · Thorsten Hohage[2] · Florian Oberender[2]

Received: 12 November 2024 / Revised: 27 November 2025 / Accepted: 8 January 2026

Abstract

We consider time-harmonic scalar transmission problems between dielectric and dispersive materials with generalized Lorentz frequency laws. For certain frequency ranges such equations involve a sign-change in their principle part. Due to the resulting loss of coercivity properties, the numerical simulation of such problems is demanding. Furthermore, the related eigenvalue problems are nonlinear and give rise to additional challenges. We present a new finite element method for both of these types of problems, which is based on a weakly coercive reformulation of the PDE. The new scheme can handle $C^{1,1}$-interfaces consisting piecewise of elementary geometries. Neglecting quadrature errors, the method allows for a straightforward convergence analysis. In our implementation we apply a simple, but non-standard quadrature rule to achieve negligible quadrature errors. We present computational experiments in two and three dimensions both for source and for eigenvalue problems. They confirm the stability and convergence of the new scheme.

Keywords Sign-changing coefficients · Dispersive materials · Plasmonics · Meta materials · Nonlinear eigenvalue problem · Finite element method

Mathematics Subject Classification 65N25 · 78M10

Support from DFG, CRC 1456 project 432680300 is gratefully acknowledged. Martin Halla acknowledges funding from DFG, project 541433971.

✉ Florian Oberender
j.oberender@math.uni-goettingen.de

Martin Halla
martin.halla@kit.edu

Thorsten Hohage
hohage@math.uni-goettingen.de

[1] Institute for Applied and Numerical Mathematics, Karlsruhe Institute of Technology, Englerstraße 2, 76131 Karlsruhe, Germany

[2] Institute of Numerical and Applied Mathematics, University of Göttingen, Lotzestraße 16-18, 37083 Göttingen, Germany

Published online: 24 January 2026

1 Introduction

The starting point of this work are time-harmonic electromagnetic transmission problems involving dispersive materials, modeled by Maxwell's equations. To simplify the setting we assume that the domain is invariant in one direction and bounded by a perfect conductor in the other two. Thus the equations are reduced to two uncoupled systems called the transverse magnetic (TM) and transverse electric (TE) problem. They can be further transformed into two scalar equations for the electromagnetic field (E,H) in the invariant direction with appropriate boundary conditions [9]. Both equations have the form

$$-\operatorname{div}(\sigma\nabla u) - \omega^2 \tau u = \tilde{f}$$

with the temporal frequency ω, the dispersive permeability $\mu = \mu(\omega, x)$ and permittivity $\epsilon = \epsilon(\omega, x)$, and $(\sigma, \tau) = (\mu^{-1}, \epsilon)$ for the TE problem and $(\sigma, \tau) = (\epsilon^{-1}, \mu)$ for the TM problem [8]. The case where σ is real valued and the sign of σ changes is of particular interest since then the arising bilinear forms are no longer (weakly) coercive, and classical theory fails [9]. On the other hand, the term associated to $-\omega^2\tau(\omega, \cdot)u$ constitutes a compact perturbation for each $\tau(\omega, \cdot) \in L^\infty$. Hence, this part is omitted for the Fredholmness and discretization analysis as its inclusion requires only standard arguments.

An important field where such sign-changing equations appear is the study of surface plasmons, which are electromagnetic waves that can form along the surface between a conductor and a dielectric material. They are the result of a resonance of light and free electrons on the surface of the conductor. This resonance of electrons essentially traps the light along the surface and is only possible if the frequency-dependent permittivities of the two materials have different signs [4]. These plasmons provide a unique way to concentrate and channel light and because of their properties there are many potential applications including light harvesting [3], the construction of miniaturized photonic circuits, and the detection of molecules [4].

To obtain numerical solutions to this problem different strategies have been developed. In the case of piecewise constant coefficients, the problem can be solved using the boundary element method [29]. Another approach suggested in [1, 2, 17] reduces the problem to a quadratic optimization problem for functions on the interface. If the optimization problem is solved iteratively, e.g., by the conjugate gradient method, PDEs with coefficients of constant signs have to be solved in each iteration step. A further issue is the proper choice of a stabilization parameter. Standard finite element methods in general only converge if the contrast (see (10) for a precise definition) is large enough as shown in [7] using T-coercivity techniques. However, the necessary bounds for the contrast are not known explicitly and simulations computed this way can be treacherous as shown in Fig. 5a and Fig. 5b. Sharp convergence results, with respect to the contrast, of finite element discretizations have been shown for polygonal interfaces in [6] if angles are rational multiples of 2π, meshes are chosen symmetric in a neighborhood of the flat parts, and special meshes at the corners are used. While this approach is promising in two dimensions, the construction of respective corner meshes in three dimensions is only possible for special cases (the Fichera corner) or leads to unsharp results. A very recent, new approach is to apply a primal-dual stabilization [10]. The eigenvalue problems associated to dispersive transmission problems are nonlinear and have received significantly less attention so far; see [12] for results on frozen coefficient simplifications, [29] for boundary element methods and [21] for a generalization of the finite element methods for polygonal interfaces in [6] to two-dimensional Maxwell eigenvalue problems. Concerning additional aspects involving sign-changing materials, we refer to [18, 26] for error estimators and to [14] for localized orthogonal decomposition techniques.

In summary, the only method for elliptic differential equations with variable sign-changing coefficients for three-dimensional domains and non-polygonal two-dimensional domains is the optimization-based approach introduced in [1]. However, the optimization process in [1] requires a large number of PDE solutions, making this approach computationally significantly more costly than a finite element discretization, especially when applied to eigenvalue problems. The main aim of this work is to propose a finite element type discretization of such problems for smooth interfaces with only standard requirements on the mesh.

The principle approach of our new scheme is to apply a suitable T-operator to the PDE yielding a weakly coercive equation and allowing for discretization with standard finite element spaces. Thus instead of using a T-operator as a theoretical tool and seeking compatible finite element spaces, our T-operator enters into the implementation of the numerical method. This idea already was mentioned briefly in [15, Section 2.3.2], but has not been studied in detail until now. In addition, our method can naturally be applied to the related eigenvalue problems. A similar approach for Stokes equations can be found in [16].

The remainder of this article is structured as follows. First we specify the considered problem. In Sect. 3 we introduce the applied reflection operator and the weakly coercive reformulation of the PDE. In Sect. 4 we discuss the implementation of the FEM and the used quadrature rules. In Sect. 5 we present several computational experiments which confirm the stability and convergence of the new scheme. In Appendix A we include some technical analysis on the bounds of the used reflection operators.

2 Notations and Problem Setting

For a domain $U \subset \mathbb{R}^d, d = 2, 3$ we denote by $(\cdot, \cdot)_U$ the scalar product of $L^2(U)$ with associated norm $\|\cdot\|_U$. Furthermore we denote by $\|\cdot\|_{H_0^1(U)} := \|\nabla\cdot\|_U$ the norm of $H_0^1(U)$ and by $\|\cdot\|_{H^{-1}(U)}$ the norm on $H^{-1}(U)$ that is given by the dual norm of $H_0^1(U)$ with respect to the norm $\|\nabla\cdot\|_U$. Unless specified otherwise, all spaces and scalar products are over $\mathbb{R}$.

We consider a bounded Lipschitz domain $\Omega \subset \mathbb{R}^d$ which is decomposed into two disjoint, nonempty Lipschitz subdomains $\Omega_\pm \subset \Omega$ with a $C^{1,1}$-interface $\Gamma := \partial\Omega_+ \cap \partial\Omega_-$ such that $\overline{\Omega_-} \subset \Omega$. We introduce the notation $u_+ := u|_{\Omega_+}$ and $u_- := u|_{\Omega_-}$ for the restrictions of a function u defined on Ω. Similarly, for a subdomain $\Sigma \subset \Omega$ we write $\Sigma_+ := \Sigma \cap \Omega_+$ and $\Sigma_- := \Sigma \cap \Omega_-$. We define the spaces of restrictions of H_0^1 on these subdomains by

$$H^1(\Sigma_\pm) := \{u|_{\Sigma_\pm} : u \in H_0^1(\Omega)\}.$$

and on these subspaces we consider the H^1-seminorms $|\cdot|_{H^1(\Sigma_\pm)} := \|\nabla\cdot\|_{\Sigma_\pm}$.

Source problems: Here we assume that the coefficient function $\sigma \in L^\infty(\Omega)$, $|\sigma|$ is essentially bounded from below by a positive constant and the restrictions of σ satisfy $\sigma_- < 0$ and $\sigma_+ > 0$. Lastly, we restrict ourselves to homogeneous Dirichlet boundary conditions and consider a source term $f \in H^{-1}(\Omega)$. This leads to the following problem:

$$\text{Find } u \in H_0^1(\Omega) \text{ such that } -\operatorname{div}(\sigma\nabla u) = f \text{ in } \Omega. \tag{1}$$

The former equation can alternatively be formulated in operator form using the operator $B \in \mathcal{L}(H_0^1(\Omega))$ defined by $(Bu, v)_{H_0^1(\Omega)} := (\sigma\nabla u, \nabla v)_{L^2(\Omega)}$:

$$\text{Find } u \in H_0^1(\Omega) \text{ such that } Bu = \tilde{f} \tag{2}$$

where $\tilde{f} \in H_0^1(\Omega)$ is the image of f under the canonical identification of $H^{-1}(\Omega)$ and $H_0^1(\Omega)$.

Eigenvalue problems: We also consider holomorphic eigenvalue problems related to the dispersive transmission problems. We restrict ourselves to local lossless passive materials which are nondispersive in Ω_+. Such materials are described by generalized Lorentz laws [13, Theorem 3.22], and convenient reconstructions from measurement data also take this form [20], cf. [13] for further discussions on the physical and mathematical requirements of dispersive material laws. Generalized Lorentz laws are described by coefficients $\sigma_0, \tau_0 \in L^\infty(\Omega)$ such that $\sigma_0, \tau_0 > 0$ and $\sigma_0^{-1}, \tau_0^{-1} \in L^\infty(\Omega)$, and take the form

$$\sigma(\omega, x) = \left(\sigma_0(x)\left(1 + \mathbb{1}_{\Omega_-}(x)\sum_{l=1}^{N_\sigma} \frac{c_{\sigma,l}^2}{\omega_{\sigma,l}^2 - \omega^2}\right)\right)^{-1} \qquad N_\sigma \in \mathbb{N}, \quad \omega_{\sigma,l}, c_{\sigma,l} \geq 0, \tag{3a}$$

$$\tau(\omega, x) = \tau_0(x)\left(1 + \mathbb{1}_{\Omega_-}(x)\sum_{l=1}^{N_\tau} \frac{c_{\tau,l}^2}{\omega_{\tau,l}^2 - \omega^2}\right) \qquad N_\tau \in \mathbb{N}, \quad \omega_{\tau,l}, c_{\tau,l} \geq 0. \tag{3b}$$

Setting

$$\Lambda := \mathbb{C} \setminus \left(\bigcup_{l=1}^{N_\sigma}\{\pm\omega_{\sigma,l}\} \cup \bigcup_{l=1}^{N_\tau}\{\pm\omega_{\tau,l}\} \cup \{\omega \in \mathbb{C} \colon \sigma(\omega) = 0\}\right),$$

we consider the following problem:

$$\begin{gathered}\text{Find } (\omega, u) \in \Lambda \times H_0^1(\Omega, \mathbb{C}) \setminus \{0\} \\ \text{such that } -\operatorname{div}(\sigma(\omega)\nabla u) - \omega^2\tau(\omega)u = 0 \text{ in } \Omega.\end{gathered} \tag{4}$$

We rewrite the former as the eigenvalue problem for a holomorphic operator function:

$$\text{Find } (\omega, u) \in \Lambda \times H_0^1(\Omega, \mathbb{C}) \setminus \{0\} \text{ such that } B(\omega)u = 0, \tag{5}$$

with $B(\cdot)\colon \Lambda \to \mathcal{L}(H_0^1(\Omega, \mathbb{C}))$ defined by

$$(B(\omega)u, v)_{H_0^1(\Omega,\mathbb{C})} := (\sigma(\omega, \cdot)\nabla u, \nabla v)_{L^2(\Omega,\mathbb{C}^d)} - (\omega^2\tau(\omega, \cdot)u, v)_{L^2(\Omega,\mathbb{C})}.$$

Note that for $\mp\operatorname{Im}(\omega^2) > 0$ we have

$$\pm\operatorname{Im}\left(\frac{-\omega^2}{\omega_{\tau,l}^2 - \omega^2}\right) \geq 0 \qquad \text{and} \qquad \pm\operatorname{Im}(\sigma(\omega, \cdot)) \geq 0.$$

This shows that for a solution (ω, u) to (5) with $\operatorname{Im}(\omega^2) \neq 0$ it follows from $\operatorname{Im}(B(\omega)u, u)_{H_0^1(\Omega,\mathbb{C})} = 0$ that $(\tau_0 u, u)_{L^2(\Omega)} = 0$ and hence $u = 0$. Since for $\operatorname{Im}(\omega^2) \neq 0$ the operator $B(\omega)$ is weakly coercive, it follows from the Fredholm alternative that the spectrum of $B(\cdot)$ is real. The challenging part is then to compute the part of the spectrum contained in $\Lambda_- := \{\omega \in \mathbb{R} \cap \Lambda \colon \sigma(\omega, \cdot) \not> 0\}$. Here $\sigma(\omega, \cdot) \not> 0$ means that there exists a set $\Omega' \subset \Omega$ of positive measure such that $\sigma(\omega, x) \leq 0$ for all $x \in \Omega'$.

3 The Weakly Coercive Reformulation

As it is an important concept in our analysis, we start with a definition of (weak) coercivity of an operator.

Definition 1 ((weak) coercivity) An operator $B \in \mathcal{L}(X)$ defined on a Hilbert space X is called coercive if there exists a constant $\alpha > 0$ such that

$$(Bv, v)_X \geq \alpha \|v\|_X^2 \quad \text{for all } v \in X.$$

An operator is called weakly coercive if it can be written as the sum of a compact and a coercive operator.

It is well known that for weakly coercive operators or operator functions Galerkin schemes yield asymptotically reliable solutions for source problems (see, e.g., [24, (13.7b)]) or for eigenvalue problems ([22, 23]), respectively. For the problem at hand, the operator B is not weakly coercive. One technique to deal with problems lacking weak coercivity is the so called T-coercivity approach that we will briefly review now. The idea is to construct a bijective operator $T \in \mathcal{L}(H_0^1(\Omega))$ such that $\tilde{B} := T^*B$ is weakly coercive. Then a solution to the source problem

$$\text{Find } u \in H_0^1(\Omega) \text{ such that } \tilde{B}u = T^*\tilde{f} \tag{6}$$

is a solution to the original problem (1) and vice-versa [15, Sect. 2.3.2]. For the eigenvalue problem we can proceed similarly if we are able to construct an operator T for which $T^*B(\omega)$ is weakly coercive for all $\omega \in \Lambda_-$. In this case we consider the eigenvalue problem:

$$\text{Find } (\omega, u) \in \Lambda' \times H_0^1(\Omega) \setminus \{0\} \text{ such that } T^*B(\omega)u = 0, \tag{7}$$

where $\Lambda' \subset \mathbb{C}$ is an open neighborhood of Λ_- for which $T^*B(\omega)$ is weakly coercive. Since weak coercivity is a continuous property and $T^*B(\omega)$ is weakly coercive on Λ_- such a neighborhood always exists, although to determine its exact shape an inspection of the frequency law is necessary. Due to our assumptions (3) the operator function $B(\cdot)$ has only real eigenvalues and since T is bijective the problem above leads to the same eigenvalues and eigenfunctions as the original problem.

Since T^*B or $T^*B(\omega)$ for $\omega \in \Lambda'$, respectively, are now weakly coercive, each Galerkin scheme yields an approximation that exhibits the usual convergence properties for all approximations which are fine enough specified in [24, (13.7b)] and [22, Thm. 2,Thm. 3], [23, Thm. 2,Thm. 3] respectively, and any convenient finite element spaces can be used.

3.1 Construction of the Operator *T*

Several approaches have been suggested to construct T-operators yielding existence and convergence results, see [6–8]. However, the operators T constructed in these references are not well suited for numerical implementations. Here we will work with a global reflection operator similar to [6] for polygons in contrast to the patch-wise approach used in [8]. We define $T \in \mathcal{L}(H_0^1(\Omega))$ as either

$$T_-u := \begin{cases} u_+ - 2\chi R_- u|_{\Sigma_-} & \text{in } \Omega_+ \\ -u_- & \text{in } \Omega_- \end{cases}$$
$$\text{or} \tag{8}$$
$$T_+u := \begin{cases} u_+ & \text{in } \Omega_+ \\ -u_- + 2\chi R_+ u|_{\Sigma_+} & \text{in } \Omega_- \end{cases},$$

where Σ is a neighbourhood of Γ,

$$R_\pm \in \mathcal{L}(H^1(\Sigma_\pm), H^1(\Sigma_\mp)) \tag{9}$$

are reflection operators which fulfill the so called matching condition $\left(R_\pm u|_{\Sigma_\pm}\right)|_\Gamma = u_\mp|_\Gamma$ and $\chi \in \mathcal{C}^1(\Omega, [0,1])$ is a cut-off function with support in Σ which equals 1 in an open neighborhood of Γ. The weak coercivity of T^*B then depends on the operator seminorms of $R_\pm$ with respect to the seminorms on $H^1(\Sigma_\pm)$ which are defined by

$$|R_\pm|_* := \sup_{|u|_{H^1(\Sigma_\pm)}=1} |u|_{H^1(\Sigma_\mp)}$$

and the so called contrasts of σ near the interface, that are given by

$$k_{+,\Sigma} := \frac{\inf_{x\in\Sigma_+} \sigma(x)}{\sup_{x\in\Sigma_-} |\sigma(x)|} \quad \text{and} \quad k_{-,\Sigma} := \frac{\inf_{x\in\Sigma_-} |\sigma(x)|}{\sup_{x\in\Sigma_+} \sigma(x)}. \tag{10}$$

Furthermore, we define $k_\pm := \inf_{\Sigma\supset\Gamma} k_{\pm,\Sigma}$ where the infimum is taken over all open neighborhoods of Γ. The precise relationship of the seminorms of the operators and the contrast is given by the following lemma, the technique of which is well known (see, e.g., [6–8]).

Lemma 1 *For $T_\pm$ be defined as above the following implication hold true:*

$$|R_\pm|_*^2 < k_{\pm,\Sigma} \quad \Rightarrow \quad T_\pm^* B \textit{ is weakly coercive.}$$

Proof We will only prove the statement for T_-, and we will write T instead of T_- for better readability. The statements for T_+ can be shown in the same way. To show that T^*B is weakly coercive under the given assumption on the contrast, we define the operators $A, K \in \mathcal{L}(H_0^1(\Omega))$ via bilinear forms as

$$\begin{aligned}
(Au, v)_{H_0^1(\Omega)} &:= (|\sigma|\nabla u, \nabla v)_\Omega - 2(\sigma\nabla u, \chi\nabla R_- v|_{\Sigma_-})_\Omega \qquad &&\forall u, v \in H_0^1(\Omega),\\
(Ku, v)_{H_0^1(\Omega)} &:= -2(\sigma\nabla u, \nabla\chi R_- v|_{\Sigma_-})_\Omega &&\forall u, v \in H_0^1(\Omega).
\end{aligned}$$

Note that due to the boundedness of R_- both bilinear forms are bounded and hence the operators are well-defined. Then $A + K = T^*B$, and we will show that A is coercive and K is compact. From the bound on the seminorm of $R_\pm$ we can derive the coercivity of A with standard arguments based on Young's inequality which can be found in the proof of Lemma 2 in [6].

For K we define the following operators to express K as a product of them. We write $\iota\colon H^1(\Sigma_-) \to L^2(\Sigma_-)$ for the compact embedding operator and define $M_{-2\sigma}\colon (L^2(\Omega))^d \to (L^2(\Omega))^d$ and $M_{\nabla\chi}\colon L^2(\Sigma_-) \to (L^2(\Omega))^d$ as the multiplication operators with symbols -2σ and $\nabla\chi$ respectively. Additionally, we write $P_{\Sigma_-}\colon H_0^1(\Omega) \to H^1(\Sigma_-)$ for the corresponding restriction operator. Because of the definitions of σ, χ and $H_{0,\Gamma}^1(\Sigma_-)$ all these operators are bounded, and we can now use the definition of K to obtain

$$\begin{aligned}
(Ku, v)_{H_0^1(\Omega)} &= (-2\sigma\nabla u, \nabla\chi R_- v|_{\Sigma_-})_\Omega\\
&= (M_{-2\sigma}\nabla u, M_{\nabla\chi}\iota R_- P_{\Sigma_-} v)_\Omega\\
&= (P_{\Sigma_-}^* R_-^* \iota^* M_{\nabla\chi}^* M_{-2\sigma}\nabla u, v)_{H_0^1(\Omega)}.
\end{aligned}$$

This implies

$$K = P_{\Sigma_-}^* R_-^* \iota^* M_{\nabla\chi}^* M_{-2\sigma}\nabla,$$

and because ι is compact, so is ι^*. This means that K is compact because it is the product of compact and bounded operators. It follows that T^*B is weakly coercive. □

Now we will provide an explicit construction of $R_\pm$ based on the geometry of the interface and provide upper bounds for their seminorm to clarify how they have to be constructed to achieve weak coercivity.

3.2 Global Reflection Operators

We construct $R_\pm$ via a $C^{0,1}$-homomorphism $\varphi\colon \Sigma \to \Sigma$ with $\varphi(\Sigma_\pm) = \Sigma_\mp$ for some neighborhood Σ of Γ such that $\varphi(x) = x$ for all $x \in \Gamma$ by

$$R_- w := w \circ \varphi, \qquad R_+ w := w \circ \varphi.$$

Recall that if $\varphi \in C^{0,1}(\Sigma)$, then by the Rademacher theorem the Jacobian $D_x\varphi$ exists for almost all $x \in \Sigma$, and $\operatorname{ess\,sup}_{x\in\Sigma} |D_x\varphi|_{\mathcal{L}(\mathbb{R}^2)} < \infty$. Moreover, if $w \in H^1(\Sigma_\pm)$, then $w \circ \varphi \in H^1(\Sigma_\pm)$ (see [30, Theorem 4.1]), and together with the invariance of Γ under φ this implies the mapping property (9). To explicitly construct φ under our assumptions, we can define the unit normal vector $n\colon \Gamma \to S^{d-1}$ pointing towards Ω_- everywhere on the interface and consider the functions

$$\begin{aligned} \Phi\colon \Gamma \times (-\delta, \delta) &\to \Omega, \quad (x, t) \mapsto x + tn(x), \\ M\colon \Gamma \times \mathbb{R} &\to \Gamma \times \mathbb{R}, \quad (x, t) \mapsto (x, -t). \end{aligned} \tag{11}$$

A similar construction also using reflections along normals of curved interfaces was used in [25]. Since Γ is $C^{1,1}$-smooth by assumption, it follows that $n \in C^{0,1}(\Gamma, \mathbb{R}^d)$, and hence $\Phi \in C^{0,1}(\Sigma, \mathbb{R}^d)$. If we can choose $\delta > 0$ small enough such that Φ is a $C^{0,1}$-homomorphism, then we can define φ by

$$\varphi := \Phi \circ M \circ \Phi^{-1}.$$

Subsequently, we define $\Sigma := \Phi(\Gamma \times (-\delta, \delta))$ and note that $\varphi(\Sigma_\pm) = \Sigma_\mp$. As $M = M^{-1}$, we have $\varphi = \varphi^{-1}$. For many practically relevant surfaces such as arcs, lines, planes and parts of spheres and cylinders an explicit computation of Φ^{-1} is feasible, but for general surfaces, one has to resort to numerical inversions. It is now possible to calculate upper bounds for the seminorms of R_+ and R_- which only depend on the geometry of the interface and δ.

Theorem 1 *In two dimensions the seminorms of the reflection operators are bounded by*

$$|R_-|_* \le \max\left(1, \left|\frac{1 - \delta \inf_{x\in\Gamma} \kappa(x)}{1 + \delta \inf_{x\in\Gamma} \kappa(x)}\right|\right), \; |R_+|_* \le \max\left(1, \left|\frac{1 + \delta \sup_{x\in\Gamma} \kappa(x)}{1 - \delta \sup_{x\in\Gamma} \kappa(x)}\right|\right),$$

where κ is the curvature of the interface. In three dimensions the bounds are

$$\begin{aligned} |R_-|_* &\le \max\left(1, \left|\frac{1 - \delta \inf_{x\in\Gamma} \kappa_1(x)}{1 + \delta \inf_{x\in\Gamma} \kappa_1(x)}\right|, \left|\frac{1 - \delta \inf_{x\in\Gamma} \kappa_2(x)}{1 + \delta \inf_{x\in\Gamma} \kappa_2(x)}\right|\right), \\ |R_+|_* &\le \max\left(1, \left|\frac{1 + \delta \sup_{x\in\Gamma} \kappa_1(x)}{1 - \delta \sup_{x\in\Gamma} \kappa_1(x)}\right|, \left|\frac{1 + \delta \sup_{x\in\Gamma} \kappa_2(x)}{1 - \delta \sup_{x\in\Gamma} \kappa_2(x)}\right|\right) \end{aligned}$$

where κ_1, κ_2 are the principal curvatures of the interface.

Remark 1 In practical applications we want these seminorms to be smaller than the square root of a given contrast k_+. Then these inequalities can be rearranged to obtain upper bounds on δ. E.g., for the seminorm $|R_+|_*$ and $\sqrt{k_+}$ we get $\delta < \frac{1}{\sup_{x\in\Gamma} \kappa(x)} \frac{\sqrt{k_+}-1}{\sqrt{k_+}+1}$ for $\sup_{x\in\Gamma} \kappa(x) > 0$.

The proof of these bounds relies on an application of the transformation formula and basic differential geometry. It is given in Appendix A. Using these explicit bounds, we can formulate conditions on the contrast of σ under which the construction of an operator T based on global reflection operators is possible such that T^*B is weakly coercive. For this we just combine our previous results with the fact that the bounds for the reflection operators decay to 1 when δ gets smaller.

Theorem 2 *(*Conditions for weak coercivity*) For an interface that is $C^{1,1}$ and piecewise C^2 there exists an operator T which can be constructed via a global reflection operator as in* (8) *such that T^*B is weakly coercive if the following conditions are satisfied:*

1. *There exists $\delta_0 > 0$ such that the map $\Phi\colon \Gamma \times (-\delta_0, \delta_0) \to \Omega$ defined by* (11) *is a $C^{0,1}$-homomorphism.*
2. *The curvature of Γ for the two dimensional case or the two principal curvatures of Γ for the three dimensional case are bounded.*
3. *One of the contrasts k_+ or k_- of σ is strictly greater than* 1*.*

Proof We only prove the case where $k_+ > 1$. The other case can be proven in the same way. From the second condition we know that the principal curvatures or the curvature is bounded in absolute value by a constant which we call κ_{max}. Now, since $k_+ > 1$ we can choose $\delta \in (0, \delta_0)$ small enough such that

$$1 \leq \frac{1 + \kappa_{max}\delta}{1 - \kappa_{max}\delta} < \sqrt{k_+}.$$

Next we consider the map $\Phi_\delta\colon \Gamma \times (-\delta, \delta) \to \Omega, (x, t) \mapsto x + tn(x)$ which is a $C^{0,1}$-homomorphism due to the first assumption. Therefore, as in Sect. 3.2, we can construct the global reflection operator R_+. Then we can use Theorem 1 and obtain

$$|R_+|_* \leq \max\left(1, \left|\frac{1 + \delta \sup_{x\in\Gamma} \kappa(x)}{1 - \delta \sup_{x\in\Gamma} \kappa(x)}\right|\right) \leq \frac{1 + \kappa_{max}\delta}{1 - \kappa_{max}\delta} < \sqrt{k_+}$$

for the two dimensional case and

$$\begin{aligned} |R_+|_* &\leq \max\left(1, \left|\frac{1 + \delta \sup_{x\in\Gamma} \kappa_1(x)}{1 - \delta \sup_{x\in\Gamma} \kappa_1(x)}\right|, \left|\frac{1 + \delta \sup_{x\in\Gamma} \kappa_2(x)}{1 - \delta \sup_{x\in\Gamma} \kappa_2(x)}\right|\right) \\ &\leq \frac{1 + \kappa_{max}\delta}{1 - \kappa_{max}\delta} < \sqrt{k_+} \end{aligned}$$

for the three dimensional case. Now Lemma 1 implies that the operator T_+^*B is weakly coercive with T_+ constructed via R_+. □

Remark 2 From the proof we can see that the last condition can be slightly weakened. In general it is enough to require one of the contrasts $k_{+,\Sigma}$ or $k_{-,\Sigma}$ to be bounded from below by 1 where Σ contains all points that are closer to Γ than a fixed distance $\delta > 0$ which can be arbitrary small. Furthermore, it may seem to be advantageous to choose δ as small as possible, but this comes with the price of a large gradient of the cut-off function.

Remark 3 We also note that the required bounds on the contrast in the two dimensional case coincide with the ones in [8] where the optimality of these bounds has been shown.

For simple geometries where the interface consists of circular arches and straight lines in two dimensions or planes, parts of spheres and cylinders in three dimensions the operator T can be implemented and used for finite element methods. Precise bounds for the necessary size of δ for a given contrast are presented in Appendix A. With such a suitable δ the convergence is then established by the weak coercivity using standard theory [24, (13.7b)] for source problems and [22, 23] for eigenvalue problems respectively as mentioned at the beginning of this section.

However there is one further challenge. For the full discretization the entries of the system matrix have to be computed numerically where integrals are approximated by quadrature rules. Usually this does not pose a major problem as long as the finite element functions and coefficient functions are smooth enough, because the Bramble-Hilbert lemma can be used to show that the quadrature error converges to zero for decreasing mesh sizes. Unfortunately, this is not the case for this method, because here we also have to consider integrals of finite element basis functions which have non-intersecting supports and numerically approximate integrals of the form

$$\int_{\mathcal{D}} \sigma(\nabla u)^\top \nabla(v \circ \varphi)\,\mathrm{d}x = \int_{\mathcal{D}} \sigma(\nabla u)^\top \nabla(v \circ \varphi)\mathbb{1}_{\varphi(\operatorname{supp} v)}\,\mathrm{d}x,$$

where we have left out the cut-off function χ for simplicity. The problem with the numerical approximation of such integrals is that even in the simplest case where φ is an affine transformation and ∇u and ∇v are constant, the function $\mathbb{1}_{\varphi(\operatorname{supp} v)}$ is only in $L^\infty(\mathcal{D})$ and therefore the classical methods fail. Additionally even if the jump of this function only occurs along a polygonal line, the quadrature approximation still does not get better with decreasing h because the function gets scaled as well. Finally the exact computation of the intersection of $\mathcal{D}$ and $\varphi(\operatorname{supp} v)$ is to costly especially if the boundary is curved and the mapping distorts the mesh geometry. We therefore cannot hope to achieve convergence of the quadrature error to zero for decreasing h. We can however see that the quadrature error decreases if a larger number of quadrature nodes is taken. Note that this generally is not true for L^∞ functions. For this we consider the simple example of a grid of quadrature nodes $(x_j)_{j\in J}$ such that a function f is approximated by a piece-wise constant function $\sum_{j\in J} f(x_j)\mathbb{1}_{C_j}$ where $(C_j)_j$ are cells around the x_j. For our method, we need to numerically approximate integrals of the form $\int_{\mathcal{D}_{\mathrm{ref}}} f\mathbb{1}_{\tilde{\varphi}(\mathcal{D}_{\mathrm{ref}})}\,\mathrm{d}x$ where all possible f are equicontinuous and bounded and $\tilde{\varphi}$ is a combination of ϕ and the affine transformations mapping the reference element to the element in the mesh and vice-versa. Therefore all $\tilde{\varphi}$ we need to consider are equicontinuous and bounded as well. Due to these equicontinuity properties, we can choose a quadrature with cells C_j small enough, such that the total area of cells which intersect the boundary of any $\tilde{\varphi}(\mathcal{D}_{\mathrm{ref}})$ is arbitrarily small, and such that any possible f is arbitrarily well approximated on cells which do not intersect the boundary of $\tilde{\varphi}(\mathcal{D}_{\mathrm{ref}})$. Such an approximation then produces an arbitrarily small quadrature error. To still get the usual convergence behaviour, we have to increase precision of our quadrature, when decreasing our mesh size h. The practical implications of this are explained further in the forthcoming sections.

4 Implementation

We implemented our method using the finite element library NGSolve [28]. The main effort lies in the implementation of the custom assemble procedure for the calculation of the stiffness matrix. To explain its details we recall in the following the convenient framework of a finite element implementation.

Let $\mathcal{T}_h = \bigcup_m \mathcal{D}_m$ be a mesh consisting of elements $\mathcal{D}_m$. Even though it is not necessary for our implementation, we assume that all elements are images of a single reference element $\mathcal{D}$ to simplify the presentation. Let us denote the corresponding transformations by $\Psi_m : \mathcal{D} \to \mathcal{D}_m$. The finite element space V_h is then implemented by providing a collection of shape functions $s_\alpha : \mathcal{D} \to \mathbb{R}$ for $\alpha = 1, \ldots, N_s$ on the reference element. For H^1-finite elements we can additionally use the gradients of the shape functions to calculate the gradient of a finite element function via the chain rule.

With these tools we are able to outline the calculation of the stiffness matrix $\mathbf{B} := (\mathbf{b}_{i,j})_{i,j=1,\dots,N}$ and the right-hand side vector $\mathbf{f} = (\mathbf{f}_i)_{i=1,\dots,N}$ defined by

$$\begin{aligned}\mathbf{b}_{i,j} &:= (Bv_j, Tv_i)_{H_0^1(\Omega)} = \int_\Omega \sigma(\nabla v_j)^\top \nabla(Tv_i)\,\mathrm{d}x,\\ \mathbf{f}_i &:= \int_\Omega \tilde{f}\,Tv_i\,\mathrm{d}x\end{aligned}$$

for finite element functions $v_i, v_j \in V_h \subset H_0^1(\Omega)$. In the following part we will only consider the case where T is defined by

$$T_- u := \begin{cases} u_+ - 2\chi R_- u|_{\Sigma_-} & \text{on } \Omega_+ \\ -u_- & \text{on } \Omega_- \end{cases},$$

because the implementation of the other case is essentially the same. We will also write R instead of R_-. With this definition of T we have

$$\begin{aligned}\mathbf{b}_{i,j} &= \int_\Omega \sigma(\nabla v_j)^\top \nabla(Tv_i)\,\mathrm{d}x\\ &= \int_{\Omega_+} \sigma(\nabla v_j)^\top \nabla(v_i - 2\chi R v_i|_{\Sigma_-})\,\mathrm{d}x + \int_{\Omega_-} \sigma(\nabla v_j)^\top \nabla(-v_i)\,\mathrm{d}x\\ &= \int_\Omega |\sigma|(\nabla v_j)^\top \nabla v_i\,\mathrm{d}x - 2\int_{\Sigma_+} \sigma(\nabla v_j)^\top \nabla(\chi R v_i|_{\Sigma_-})\,\mathrm{d}x\\ &=: \mathbf{b}_{i,j}^{(1)} - 2\mathbf{b}_{i,j}^{(2)}\end{aligned}$$

so we get $\mathbf{B} = \mathbf{B}^{(1)} - 2\mathbf{B}^{(2)}$. In the same way we can define $\mathbf{f}^{(1)}, \mathbf{f}^{(2)}$ with $\mathbf{f} = \mathbf{f}^{(1)} - 2\mathbf{f}^{(2)}$. We note that $\mathbf{B}^{(1)}$ is the stiffness matrix of a bilinear form without any special operators, so it can be calculated the usual way. The same is true for $\mathbf{f}^{(1)}$. Additionally, we see that the domain of integration required for $\mathbf{B}^{(2)}$ is just Σ_+, so for its calculation we only have to consider finite element functions with support in Σ. To make use of this and to generally simplify the implementation and further calculations, we subdivide $\Sigma =: \bigcup_{l=1}^L \Sigma^{(l)}$ according to the type of the interface by lines or planes which are perpendicular to the interface. An example of such a setup is depicted in Fig. 1. We then generate the mesh such that each element lies either in $\Omega \setminus \Sigma$, in $\Sigma^{(l)} \cap \Sigma_-$ or in $\Sigma^{(l)} \cap \Sigma_+$ for some $l = 1, \dots, L$. This is easy to achieve by standard mesh generators and simplifies the bookkeeping of the transformations. The additional subdivision based on the interface geometry $\Sigma = \bigcup_{l=1}^L \Sigma^{(l)}$ could be avoided by implementing one global map φ for the whole interface instead.

This subdivision allows us to write R as

$$(Rv)(x) = (v \circ \varphi^{(l)})(x) \text{ for } x \in \Sigma_+^{(l)}$$

where $\varphi^{(l)} \colon \Sigma^{(l)} \to \Sigma^{(l)}$ is a predefined transformation based on the interface geometry. Because the mesh respects this subdivision, on each finite element only one transformation has to be considered and transformations only have to be considered for elements in Σ.

Due to the presence of the transformation φ the assembly of $\mathbf{B}^{(2)}$ is non-standard and poses the main challenge for the application of the method. This is mainly caused by the fact that usual assembly procedures assume that only finite element functions which are supported on a common element contribute to the entries of the matrix. In our case functions v_j and $v_i \circ \varphi$

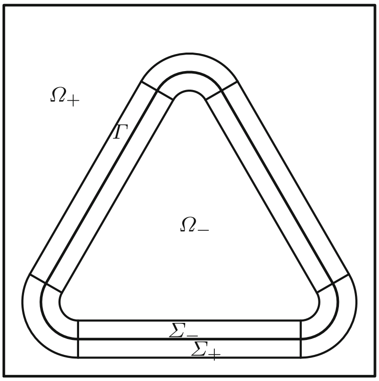


Fig. 1 Example for subdivided domain based on interface geometry

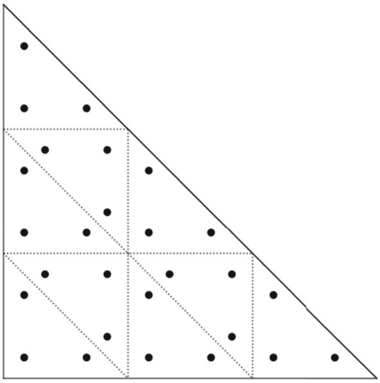


Fig. 2 Example for subdivided element with quadrature points

can have intersecting supports even though the corresponding finite element functions have disjoint supports. Additionally the intersection of the supports is usually not a finite element. To tackle these problems, the main idea is to reduce the problem to contributions from single quadrature points which can then be calculated explicitly.

For this let $\mathcal{D}_m$ be an element of $\mathcal{T}_h$ in Σ_+ and let $v_i, v_j \in V_h$ be such that $\operatorname{supp} v_j \cap \operatorname{supp} v_i \circ \varphi \cap \mathcal{D}_m \neq \emptyset$. For the sake of simplicity we confine our presentation to the case that on any element $\mathcal{D}_m$ a basis function v_j is given by a single shape function and can therefore be written as

$$v_j|_{\mathcal{D}_m} = s_{\alpha_{j,m}} \circ \Psi_m^{-1}. \tag{12}$$

(We may formally introduce a shape function $s_0 \equiv 0$ for the case the $\operatorname{supp} v_j$ and $\mathcal{D}_m$ are disjoint.)

The contribution of the element $\mathcal{D}_m$ to the entry $\mathbf{b}_{i,j}^{(2)}$ corresponding to v_i and v_j is then given by

$$\begin{aligned}\mathbf{b}_{i,j,m}^{(2)} &:= \int_{\mathcal{D}_m} \sigma(x) \nabla (s_{\alpha_{j,m}} \circ \Psi_m^{-1})(x)^\top \nabla(\chi \cdot R v_i)(x) \, \mathrm{d}x \\ &= \int_{\mathcal{D}} \sigma(\Psi_m(y)) \nabla (s_{\alpha_{j,m}} \circ \Psi_m^{-1})(\Psi_m(y))^\top \nabla(\chi \cdot R v_i)(\Psi_m(y)) \left|\det D_y \Psi_m\right| \, \mathrm{d}y,\end{aligned}$$

with the Jacobian $D_y \Psi_m$ of Ψ_m at y. Approximating the integral over the reference element $\mathcal{D}$ by a quadrature rule

$$\int_{\mathcal{D}} g(y)\,\mathrm{d}y \approx \sum_k w_k g(y_k),$$

the contribution of the quadrature point y_k is given by

$$\mathbf{b}^{(2)}_{i,jm,k} := w_k \sigma(\Psi_m(y_k)) \nabla(s_{\alpha_{j,m}} \circ \Psi_m^{-1})(\Psi_m(y_k))^\top \nabla(\chi \cdot R v_i)(\Psi_m(y_k)) \left|\det D_{y_k} \Psi_m\right|.$$

Overall, we have

$$\mathbf{b}^{(2)}_{i,j} = \sum_m \mathbf{b}^{(2)}_{i,j,m} \approx \sum_m \sum_k \mathbf{b}^{(2)}_{i,j,m,k}.$$

To compute $\mathbf{b}^{(2)}_{i,jm,k}$, we have to find the element which $\varphi(\Psi_m(y_k))$ belongs to: Suppose that $\varphi(\Psi_m(y_k)) \in \mathcal{D}_{n(m,k)}$. Then (12) implies $v_i|_{\mathcal{D}_{n(m,k)}} = s_{\alpha_{i,n(m,k)}} \circ \Psi^{-1}_{n(m,k)}$, and we obtain

$$(R v_i)(\Psi_m(y_k)) = (v_i \circ \varphi)\,(\Psi_m(y_k)) = \left(s_{\alpha_{i,n(m,k)}} \circ \Psi^{-1}_{n(m,k)} \circ \varphi\right)(\Psi_m(y_k))$$

and hence

$$\begin{aligned}\mathbf{b}^{(2)}_{i,j,m,k} = w_k \left|\det D_{y_k} \Psi_m\right| \sigma(\Psi_m(y_k)) \left(\nabla(s_{\alpha_{j,m}} \circ \Psi_m^{-1})(\Psi_m(y_k))\right)^\top \\ \nabla\left(\chi \cdot (s_{\alpha_{i,n(m,k)}} \circ \Psi^{-1}_{n(m,k)} \circ \varphi)\right)(\Psi_m(y_k)).\end{aligned}$$

This can be explicitly calculated via the chain rule as long as the values and derivatives of φ and χ can be computed.

To optimize the assembly procedure and make it more flexible, in practice all shape functions on an element can be considered at the same time, and the implementation is done in a way that the calculation of $\mathbf{b}^{(2)}_{i,j,m,k}$ can be easily replaced to allow for the computation of different integrands. For example,

$$\begin{aligned}\tilde{\mathbf{b}}^{(2)}_{i,j,m,k} =& w_k \left|\det D_{y_k} \Psi_m\right| \left((s_{\alpha_j} \circ \Psi_m^{-1})(\Psi_m(y_k))\right) \\ &\cdot \left(\chi \cdot (s_{\alpha_{i,n(m,k)}} \circ \Psi^{-1}_{n(m,k)} \circ \varphi)\right)(\Psi_m(y_k))\end{aligned}$$

can be used for L^2-terms.

To cope with the errors related to the quadrature one needs to use a high number of quadrature points. This could be done by just using higher quadrature orders but they are adapted to high order polynomials and do not work particularly well for piece-wise continuous functions. We therefore use a composite quadrature rule which is based on dividing an element into many smaller similar elements and then using a standard Gauss-Legendre quadrature rule on each smaller element [19]. An example for this is shown in Fig. 2.

Note that this specialized quadrature is only necessary in the small part $\Sigma_\pm \subset \Omega$. The required additional effort only mildly increases the asymptotic complexity of the total assemble procedure. The most expensive step for each element consists in finding the elements where the individual quadrature points get mapped to which can be done in $\mathcal{O}(\log N)$ using search trees. This leads to a total complexity of $\mathcal{O}(N \log N)$.

For the linear forms we can use a simple transformation to significantly simplify the calculations:

$$\begin{aligned}\mathbf{f}_i^{(2)} &= \int_{\Sigma_+} \tilde{f}(x)\chi(x)(v_i \circ \varphi)(x)\,\mathrm{d}x \\ &= \int_{\Sigma_-} \tilde{f}(\varphi(y))\chi(\varphi(y))v_i(y)|\det D_y\varphi|\,\mathrm{d}y.\end{aligned}$$

In this new form the transformation φ is now no longer composed with a finite element function so the calculation of $\mathbf{f}^{(2)}$ can then be performed by implementing a coefficient function $h(y) := \tilde{f}(\varphi(y))\chi(\varphi(y))|\det D_y\varphi|$ and then defining and assembling the linear form as usual.

As a last step we have to solve a linear system. Because of the incorporation of the operator T the matrix $\mathbf{B}$ is not symmetric and it is also has an unusual sparsity pattern as it is more dense for elements near the interface. Direct solvers which are not used to this setting therefore sometimes are not able to handle this problem and we use iterative methods instead. The fastest and most reliable solver which we have found so far and which is used in all numerical experiments is the generalized minimal residual method (GMRES) paired with a simple diagonal preconditioner. The loss of the symmetry and the reduction of sparsity is one of the main disadvantages of our approach. In particular, it limits the use of direct solvers to small problem instances, and it reduces the efficiency of standard direct solvers. For larger systems solving the linear system with diagonally preconditioned GMRES takes much longer than using a direct solver for the symmetric system. For example, for a circular domain and about 500,000 degrees of freedom diagonally preconditioned GMRES takes 50 times as long as a direct solver for a symmetric system, and this ratio becomes worse the larger the system gets. In general, the time needed for the solution is hard to estimate and varies depending on the geometry and the mesh size as this greatly impacts the sparsity of the matrix.

5 Numerical Experiments

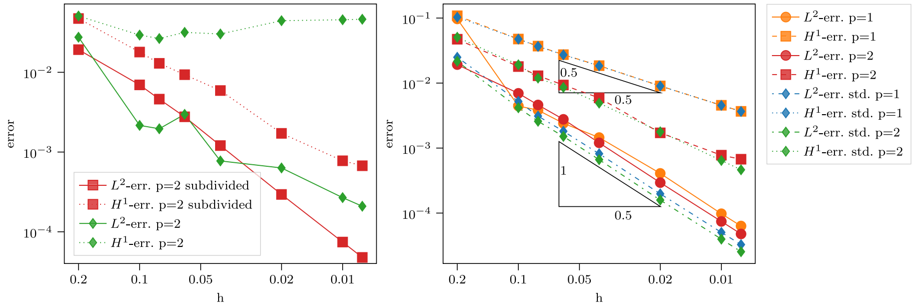
In this section we present different examples that illustrate our method for different domains. We consider the convergence rates, analyze the errors, and for settings which allow the application of classical finite elements we compare our results with those.

5.1 Examples in Two Dimensions

Our first example is a two dimensional domain, that consists of a ring and a disc by defining $\sigma : B_2 \to \mathbb{R}$ in polar coordinates by

$$\sigma(r) := \begin{cases} -1 & \text{for } r \leq 1 \\ 3 & \text{else} \end{cases},$$

where $B_r := \{x \in \mathbb{R}^2 \colon |x| < r\}$. This leads to the subdomains $\Omega_- = B_1 \subset \mathbb{R}^2$ and $\Omega_+ = B_2 \setminus \overline{B_1}$.



(a) (b)

Fig. 3 Numerical results for solution of $-\operatorname{div}(\sigma\nabla u) = f$ for a disc-shaped inclusion. (**a**) Influence of the custom quadrature rule for $p = 2$. The usual quadrature rule leads to a stagnation of the H^1-errors and erratic convergence of the L^2 errors. (**b**) Expected linear (for H^1) and quadratic (for L^2) decay of convergence errors for $p = 1$

We define the right hand-side f also in polar coordinates as

$$f(r) := \begin{cases} 4 & \text{for} \quad r \leq 1 \\ 4\frac{1-r}{r} & \text{for } 1 < r \leq 2 \end{cases},$$

which corresponds to the solution $u_{\text{ref}} \in H_0^1(\Omega)$ given by

$$u_{\text{ref}}(r) = \begin{cases} r^2 - \frac{2}{3} & \text{for} \quad r \leq 1 \\ \frac{1}{3}(r-2)^2 & \text{for } 1 < r \leq 2 \end{cases}, \tag{13}$$

According to Corollary 2 we can choose $\delta = 0.2$ in (11) because the squared norm of R is then bounded by 2.25, which is smaller than the contrast $k_- = 3$.

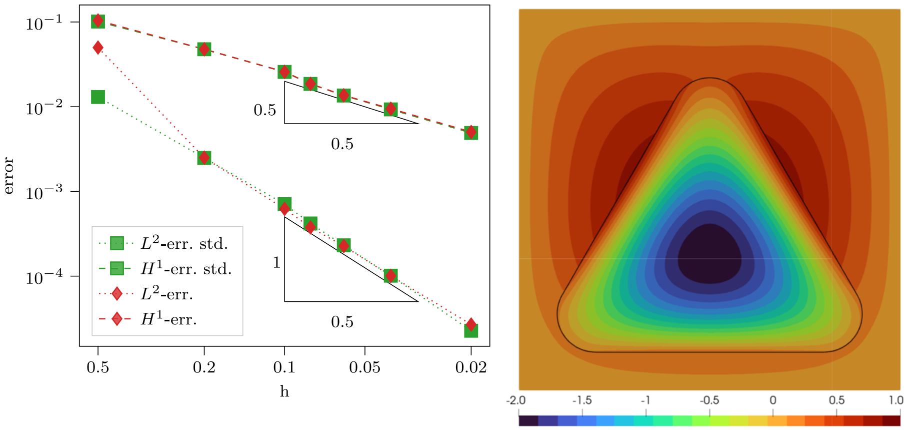
With this setup we now compute the solutions using H^1-conforming finite element spaces of order $p = 1$ and $p = 2$ for meshes with varying mesh sizes. At the boundary and the interface we use curved elements with a quadratic geometry approximation. In Fig. 3a we compare for our new method the achieved errors with standard quadrature rules vs. quadrature rules with subdivision, where each element is divided into 5 smaller similar elements for $h \geq 0.04$ and 625 smaller elements for $h < 0.04$. We observe in Fig. 3a that the adapted quadrature rule becomes necessary to achieve small H^1-errors. Henceforth, a quadrature rule with subdivision into 9 smaller elements is used for all following calculations using the new method, as this is enough in the following examples. We also do this for $p = 1$ even though in our experiments the finite element approximation error still dominates the quadrature error. Fig. 3b shows a log-log-plot of the relative errors in H^1- and in L^2-norm with respect to the reference solution. We compare our new method with a classical finite element method using the same meshes. We observe that the errors are of comparable size and converge with the same rate.

For the smaller examples, we also computed the condition numbers of the resulting linear systems. The results in Tab. 1 show that the unusual structure of the matrices in our method does not significantly impact their condition numbers. This underlines the stability of our method.

After we have investigated the convergence for a simple domain, we will now move on to a more complicated domain to show and inspect the applicability of the method for more realistic configurations. To this end we consider an equilateral triangle with rounded

Table 1 Comparison of condition numbers for the linear systems occurring in our method and the standard finite element method

	$p = 1$		$p = 2$	
h	std.	new	std.	new
0.2	64.78	65.35	717.68	757.98
0.1	284.77	292.13	804.39	804.42
0.08	469.12	475.63	789.56	884.68
0.06	894.35	910.12	903.76	1140.88
0.04	1920.35	1948.11	–	–



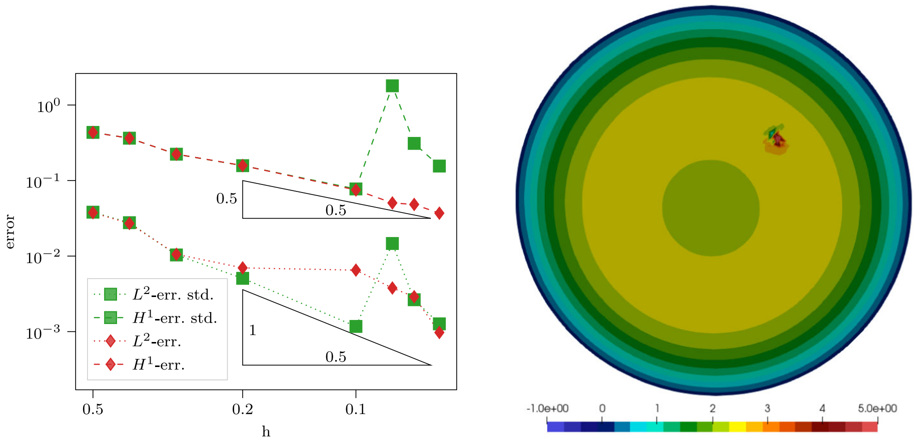
(a)

(b)

Fig. 4 Numerical results for solution of $-\operatorname{div}(\sigma \nabla u) = 1$ for a rounded triangle-shaped inclusion. (**a**) Expected linear (for H^1) and quadratic (for L^2) decay of errors, (**b**) Example solution u for $h = 0.04$ and $k_- = 1.1$

corners inside a square. Here Ω is the square with corners $(0, 0)$, $(10, 0)$, $(10, 10)$ and $(0, 10)$. Now we consider the equilateral triangle $\mathcal{D}$ inside this square with corners $(2, 2)$, $(8, 2)$ and $(5, 2 + 3\sqrt{3})$ and define $\Omega_- := \{x : \operatorname{dist}(x, \mathcal{D}) < 1\}$. This leads to a shape with a boundary that is comprised of three circular arcs with radius 1 connected by three straight lines. As usual we then set $\Omega_+ := \Omega \setminus \overline{\Omega_-}$ and we define σ to be piecewise constant such that $\sigma|_{\Omega_-} = -1$ and $\sigma|_{\Omega_+} = 10$. This enables us to choose $\delta = 0.5$. Finally, we choose $f = 1$ as right hand-side and compare our solutions to a reference solution that was computed using the usual finite element method on a finer mesh ($h = 0.008$).

The resulting errors for finite elements of order 1 are depicted in Fig. 4a. There we observe that the usual method and our method converge with approximately the same rates, which shows that the new method also performs well for more complicated interface geometries. In Fig. 4b we plot a computed solution for a contrast that is much closer to 1, which is obtained by using $\delta = h = 0.04$. While we have no analytical reference solution to compare with, we can still see the absence of singularities and the expected symmetry.



(a)

(b)

Fig. 5 Numerical results for solution of $-\operatorname{div}(\sigma\nabla u) = \sigma\frac{6|x|-9}{4}$ for a sphere-shaped inclusion. (**a**) Decay of errors for $p = 1$ with notable instabilities for the standard method, (**b**) Cross-section of solution computed using the standard method ($h = 0.08$) exhibiting artifacts near the interface

5.2 Example in Three Dimensions

We also present an example for a three dimensional domain. Again, to have an explicit reference solution we choose a domain that consists of a smaller ball inside a bigger one. This leads to the following domains

$$\Omega := \{x : |x| < 4\}, \qquad \Omega_- := \{x : |x| < 2\}, \qquad \Omega_+ := \Omega \setminus \overline{\Omega_-}.$$

Thence we use the right hand-side

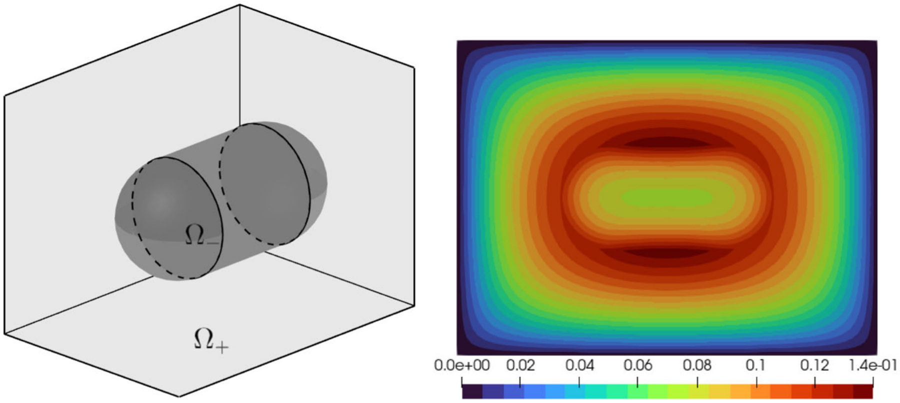
$$f(r) := \sigma(r)\frac{6r-9}{4}$$

corresponding to the solution

$$u_{\mathrm{ref}}(r) = -\frac{1}{8}(r^3 - 3r^2 - 16)$$

in spherical coordinates, where $\sigma|_{\Omega_-} := -1$ and $\sigma|_{\Omega_+} := 2$. We then choose $\delta = 0.2$ and compute the relative errors which are depicted in Fig. 5a.

We observe that the L^2- and H^1-errors for our method mostly decay with the expected rates apart from a small plateau in the L^2-error, which is most likely caused by the use of anisotropic meshes for $h > \delta$. In contrast, we notice that the errors for the standard method do not converge. This phenomenon is caused by the appearance of local singularities near the interface depicted in Fig. 5b, which may occur regardless of the mesh size and have been observed previously, see, e.g., [6]. Finally, we also present an example for the more complicated pill shaped inclusion depicted in Fig. 6a. A slice of the solution computed by our method is shown in Fig. 6b. Note that it has the expected symmetries and does not exhibit any singularities.



(a)

(b)

Fig. 6 Numerical results for the solution of $-\operatorname{div}(\sigma\nabla u) = 1$ for pill-shaped inclusion. (**a**) Sketch of domain geometry, (**b**) Cross-section of the solution $h = 0.06$ exhibiting expected symmetry and no artifacts

5.3 Example of a Dispersive Eigenvalue Problem

We use the same disc shaped geometry, which we considered in the very first example. Its symmetry allows us to compute the eigenvalues semi-analytically using Bessel functions and this enables us to obtain accurate reference solutions. We then consider the following eigenvalue problem:

$$\text{Find } (\omega, u) \in \mathbb{C} \times H_0^1(\Omega) \setminus \{0\} \text{ such that } -\operatorname{div}(\sigma(\omega)\nabla u) - \omega^2 u = 0$$

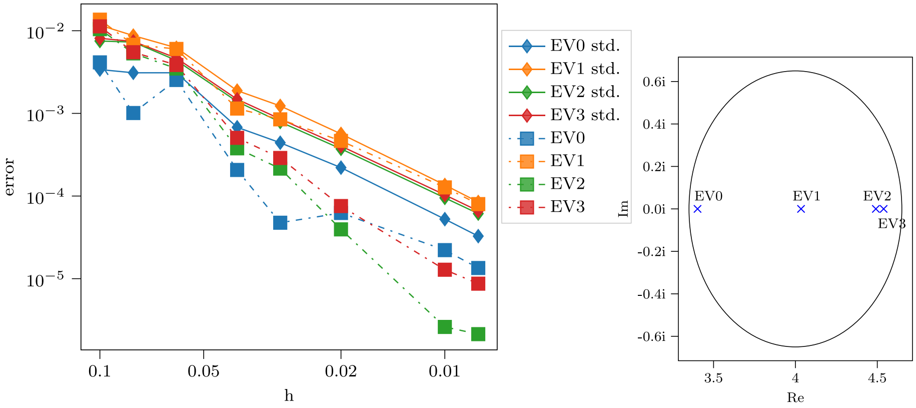
where

$$\sigma(\omega) := \begin{cases} \frac{\omega^2}{\omega^2 - 200} & x \in \Omega_- \\ 1 & \text{else} \end{cases}.$$

Our finite element method discretizes this problem into a holomorphic eigenvalue problem for a matrix, which is subsequently solved using the contour integral method proposed by Beyn [5]. As the contour we choose a circle with radius 0.65 centered at 4.0. For comparison, the same method is also used for the discrete system obtained via a standard FEM. In Fig. 7a we observe that both methods find the same 4 eigenvalues which are depicted in Fig. 7b and coincide with eigenvalues obtained by the semi-analytic method. We see that both methods reliably compute the eigenvalues with similar convergence rates.

6 Conclusions

We have presented a new numerical approach for solving both source problems and holomorphic eigenvalue problems with sign-changing coefficients in the leading order term and $C^{1,1}$-smooth interfaces, which is based on a finite element-type discretization. Our method does not impose any restrictions on the finite element mesh, except that it has to respect the interface and the neighborhood Σ of the interface. We have demonstrated the practicality and that the quadrature rule can be chosen such that the expected convergence rates (i.e., the



(a)

(b)

Fig. 7 Numerical results for the solution of a dispersive eigenvalue problem for a disc-shaped inclusion. (**a**) Decay of errors for $p = 1$ with similar or smaller errors compared to standard method, (**b**) Sketch of eigenvalues and contour used for their computation

order of the finite element interpolation error) for examples in two and three dimensions are achieved (Figs. 3b, 4a.)

The proposed method requires numerical quadrature on unfitted reflected meshes near the interface. Whereas for low order methods this can efficiently be achieved by standard quadratures on uniformly refined meshes, it poses a major challenge for high order methods. The assembly procedure, although non-standard, is of log-linear complexity and can be implemented using tools available in many finite element packages. Compared to standard finite element discretizations, the numerical solution of the linear system is complicated by the loss of symmetry and a denser and non-standard matrix structure. But we did not observe a significant increase of condition numbers and could reliably solve the linear systems using iterative methods. However, the proposed method is very flexible, and in contrast to some competing methods, the overall numerical effort is still of the same order of magnitude as standard finite element discretizations.

Appendix A Reflection Operator Bounds

In this appendix we present explicit bounds for the seminorms of the global reflection operators $R_\pm$, which are solely based on the curvature of the interface and δ. The analysis consists of two main steps. First we prove the following lemma:

Lemma 2 *Consider the operators $R_\pm : H^1(\Sigma_\pm) \to H^1(\Sigma_\mp)$ defined via φ as in Sect. 3.2. Then the seminorm of R defined by*

$$|R_\pm|_* := \sup_{|u|_{H^1(\Sigma_\pm)}=1} |u|_{H^1(\Sigma_\mp)}$$

is bounded by

$$|R_\pm|_* \le \sup_{y \in \Phi^{-1}(\Sigma_\mp)} \left\| (D_{M(y)}\Phi)(D_y M)(D_y \Phi)^{-1} \right\|_{\mathcal{L}(\mathbb{R}^d)} . \tag{14}$$

Proof For $w \in H^1(\Sigma_\pm)$ the chain rule yields

$$\begin{aligned}|R_\pm w|^2_{H^1(\Sigma_\mp)} &= \int_{\Sigma_\mp} \left|\nabla(w \circ \Phi \circ M \circ \Phi^{-1})(x)\right|^2 \,\mathrm{d}x\\ &= \int_{\Sigma_\mp} \Big|[(\nabla w)(\Phi \circ M \circ \Phi^{-1}(x))]^\top \, (D_{M\circ\Phi^{-1}(x)}\Phi)\\ &\qquad (D_{\Phi^{-1}(x)}M)(D_x\Phi^{-1})\Big|^2 \,\mathrm{d}x\end{aligned}$$

Now we use the transformation formula for $y := \Phi^{-1}(x)$ and get

$$\begin{aligned}|R_\pm w|^2_{H^1(\Sigma_\mp)} &= \int_{\Phi^{-1}(\Sigma_\mp)} \left|[(\nabla w)(\Phi \circ M(y))]^\top (D_{M(y)}\Phi)(D_yM)(D_{\Phi(y)}\Phi^{-1})\right|^2\\ &\qquad \left|\det D_y\Phi\right| \,\mathrm{d}y\\ &= \int_{\Phi^{-1}(\Sigma_\mp)} \left|[(\nabla w)(\Phi \circ M(y))]^\top (D_{M(y)}\Phi)(D_yM)(D_y\Phi)^{-1}\right|^2\\ &\qquad \left|\det D_y\Phi\right| \,\mathrm{d}y\\ &\le \int_{\Phi^{-1}(\Sigma_\mp)} |(\nabla w)(\Phi \circ M(y))|^2 \, \|(D_{M(y)}\Phi)(D_yM)(D_y\Phi)^{-1}\|^2_{\mathcal{L}(\mathbb{R}^d)}\\ &\qquad \left|\det D_y\Phi\right| \,\mathrm{d}y\end{aligned}$$

Denoting the right hand side of (14) by $C_\pm$ and applying the transformation formula two more times yields

$$\begin{aligned}|R_\pm w|^2_{H^1(\Sigma_\mp)} &\le C_\pm^2 \int_{\Phi^{-1}(\Sigma_\mp)} |(\nabla w)(\Phi \circ M(y))|^2 \left|\det D_y\Phi\right| \mathrm{d}y\\ &= C_\pm^2 \int_{M\circ\Phi^{-1}(\Sigma_\mp)} \left|(\nabla w)(\Phi(y'))\right|^2 \left|\det D_{y'}\Phi\right| \mathrm{d}y'\\ &= C_\pm^2 \int_{\Sigma_\pm} \left|(\nabla w)(x')\right|^2 \mathrm{d}x' = C_\pm^2 |w|^2_{H^1(\Sigma_\pm)}.\end{aligned}$$

□

As a second step we parameterize the interface and calculate the spectral norm of the matrix given by the lemma to get an explicit bound. This part is split into different sections based on the dimension.

A.1 Bounds in Two Dimensions

In case of two dimensions, the interface is a curve that is $C^{1,1}$. Hence, by the Rademacher theorem the second derivative of parameterizations exists almost everywhere and is uniformly bounded, so we can define the signed curvature $\kappa\colon \Gamma \to \mathbb{R}$ corresponding to the normal vector. We then obtain the following bounds for $R_\pm$.

Theorem 3 *In two dimensions the reflection operators are bounded by*

$$|R_-|_* \le \max\left(1, \left|\frac{1-\delta \inf_{x\in\Gamma}\kappa(x)}{1+\delta \inf_{x\in\Gamma}\kappa(x)}\right|\right), \qquad |R_+|_* \le \max\left(1, \left|\frac{1+\delta \sup_{x\in\Gamma}\kappa(x)}{1-\delta \sup_{x\in\Gamma}\kappa(x)}\right|\right).$$

Proof We take $x \in \Gamma$ fixed and choose a coordinate system such that $n(x) = (0, 1)^\top$. Now we can parameterize Γ around x by $\gamma : [-\alpha, \alpha] \to \Gamma$ such that

$$\gamma(0) = x, \quad \gamma'(0) = (1, 0)^\top.$$

Next we use the identity $\frac{\mathrm{d}}{\mathrm{d}s}[n(\gamma(s))] = \kappa(\gamma(s))\gamma'(s)$ [11, Chapter 1.5] to calculate $D\Phi$ at (x, t_0):

$$\begin{aligned} D_{(x,t_0)}\Phi &= \left(\tfrac{\mathrm{d}}{\mathrm{d}s}[\gamma(s) + tn(\gamma(s))]_{s=0,t=t_0} \middle| \tfrac{\mathrm{d}}{\mathrm{d}t}[\gamma(s) + tn(\gamma(s))]_{s=0,t=t_0} \right) \\ &= \left(\gamma'(0) + t_0 \tfrac{\mathrm{d}}{\mathrm{d}s}[n(\gamma(s))]_{s=0} \middle| n(\gamma(0)) \right) \\ &= \begin{pmatrix} 1 + t_0\kappa(x) & 0 \\ 0 & 1 \end{pmatrix}. \end{aligned}$$

Because this formula is independent of the parametrization γ, we can now calculate

$$\begin{aligned} (D_{M((x,t))}\Phi)(D_{(x,t)}M)(D_{(x,t)}\Phi)^{-1} &= (D_{(x,-t)}\Phi)(D_{(x,t)}M)(D_{(x,t)}\Phi)^{-1} \\ &= \begin{pmatrix} 1 - t\kappa(x) & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} 1 + t\kappa(x) & 0 \\ 0 & 1 \end{pmatrix}^{-1} \\ &= \begin{pmatrix} \frac{1-t\kappa(x)}{1+t\kappa(x)} & 0 \\ 0 & -1 \end{pmatrix} \end{aligned}$$

and get

$$\|(D_{M((x,t))}\Phi)(D_{(x,t)}M)(D_{(x,t)}\Phi)^{-1}\|_{\mathcal{L}(\mathbb{R}^2)} = \max\left(\left| \frac{1 - t\kappa(x)}{1 + t\kappa(x)} \right|, 1 \right).$$

Using Lemma 2, this leads to the following bound:

$$\begin{aligned} |R_-|_* &\leq \sup_{(x,t)\in\Phi^{-1}(\Sigma_+)} \|(D_{M((x,t))}\Phi)(D_{(x,t)}M)(D_{(x,t)}\Phi)^{-1}\|_{\mathcal{L}(\mathbb{R}^d)} \\ &= \sup_{(x,t)\in\Phi^{-1}(\Sigma_+)} \left| \frac{1 - t\kappa(x)}{1 + t\kappa(x)} \right| = \sup_{x\in\Gamma, t\in[0,\delta)} \left| \frac{1 - t\kappa(x)}{1 + t\kappa(x)} \right|. \end{aligned}$$

If $\kappa(x) \geq 0$, we always have that $\left|\frac{1-t\kappa(x)}{1+t\kappa(x)}\right|$ attains its maximum at $t = 0$ and the maximum is 1. If $\kappa(x) \leq 0$, the expression is monotonically increasing in t and its maximum is attained at $t = \delta$. With this, we can eliminate the dependence of the supremum on t and obtain that

$$|R_-|_* \leq \max\left(1, \left| \frac{1 - \delta \inf_{x\in\Gamma} \kappa(x)}{1 + \delta \inf_{x\in\Gamma} \kappa(x)} \right| \right).$$

The result for R_+ can be derived in the same way. □

A.2 Bounds in Three Dimensions

Similar to the two dimensional case, we will again derive a bound for the operator based on the curvature of the surface. For this, we consider the two principal curvatures $\kappa_1, \kappa_2 \colon \Gamma \mapsto \mathbb{R}$ corresponding to the sign convention in the previous definition of the normal vector. In the three dimensional case the map n is known as the Gauss map and we will use its properties to prove the following theorem.

Theorem 4 *In three dimensions the reflection operators are bounded by*

$$|R_-|_* \leq \max\left(1, \left|\frac{1-\delta\inf_{x\in\Gamma}\kappa_1(x)}{1+\delta\inf_{x\in\Gamma}\kappa_1(x)}\right|, \left|\frac{1-\delta\inf_{x\in\Gamma}\kappa_2(x)}{1+\delta\inf_{x\in\Gamma}\kappa_2(x)}\right|\right),$$
$$|R_+|_* \leq \max\left(1, \left|\frac{1+\delta\sup_{x\in\Gamma}\kappa_1(x)}{1-\delta\sup_{x\in\Gamma}\kappa_1(x)}\right|, \left|\frac{1+\delta\sup_{x\in\Gamma}\kappa_2(x)}{1-\delta\sup_{x\in\Gamma}\kappa_2(x)}\right|\right).$$

Proof The proof is very similar to the two dimensional case. We again consider a fixed point $p \in \Gamma$ and choose a coordinate system and a parametrization $\gamma : (-\alpha, \alpha) \times (-\beta, \beta) \to \Gamma$ such that

$$\gamma(0,0) = p, \quad \frac{\mathrm{d}}{\mathrm{d}x}\gamma(0,0) = \begin{pmatrix}1\\0\\0\end{pmatrix}, \quad \frac{\mathrm{d}}{\mathrm{d}y}\gamma(0,0) = \begin{pmatrix}0\\1\\0\end{pmatrix}, \quad n(p) = \begin{pmatrix}0\\0\\1\end{pmatrix}.$$

As in the two dimensional case, we then calculate

$$\begin{aligned} D_{(p,t_0)}\Phi &= \Big(\tfrac{\mathrm{d}}{\mathrm{d}x}[\gamma(x,y) + tn(\gamma(x,y))] \Big| \tfrac{\mathrm{d}}{\mathrm{d}y}[\gamma(x,y) + tn(\gamma(x,y))] \Big| \\ &\quad \tfrac{\mathrm{d}}{\mathrm{d}t}[\gamma(x,y) + tn(\gamma(x,y))] \Big) \\ &= \Big(\tfrac{\mathrm{d}}{\mathrm{d}x}\gamma(x,y) \Big| \tfrac{\mathrm{d}}{\mathrm{d}y}\gamma(x,y) \Big| n(\gamma(x,y)) \Big) \\ &\quad + t_0 \Big(\tfrac{\mathrm{d}}{\mathrm{d}x}n(\gamma(x,y)) \Big| \tfrac{\mathrm{d}}{\mathrm{d}y}n(\gamma(x,y)) \Big| 0 \Big) \\ &= \begin{pmatrix}1&0&0\\0&1&0\\0&0&1\end{pmatrix} + t_0 \Big(\tfrac{\mathrm{d}}{\mathrm{d}x}n(\gamma(x,y)) \Big| \tfrac{\mathrm{d}}{\mathrm{d}y}n(\gamma(x,y)) \Big| 0 \Big). \end{aligned}$$

Now, we use that the derivative of the Gauss map n at p is given by the Weingarten map $w\colon T_p\Gamma \mapsto T_p\Gamma$ where $T_p\Gamma$ is the tangent space of Γ at p. Next we use that in our chosen coordinate system w can be represented by a matrix W. Then we can write the second summand in the previous equation using the Weingarten map as

$$D_{(p,t_0)}\Phi = \begin{pmatrix}1&0&0\\0&1&0\\0&0&1\end{pmatrix} + t_0 \left(\begin{array}{c|c} W & 0 \\ \hline 0 & 0 \end{array}\right) = \left(\begin{array}{c|c} I + t_0 W & 0 \\ \hline 0 & 1 \end{array}\right).$$

We can now calculate

$$\begin{aligned} (D_{M((p,t))}\Phi)(D_{(p,t)}M)(D_{(p,t)}\Phi)^{-1} &= \left(\begin{array}{c|c} I - tW & 0 \\ \hline 0 & 1 \end{array}\right) \left(\begin{array}{c|c} I & 0 \\ \hline 0 & -1 \end{array}\right) \\ &\quad \left(\begin{array}{c|c} I + tW & 0 \\ \hline 0 & 1 \end{array}\right)^{-1} \\ &= \left(\begin{array}{c|c} (I - tW)(I + tW)^{-1} & 0 \\ \hline 0 & -1 \end{array}\right). \end{aligned}$$

The Weingarten map is diagonalizable and its eigenvalues are the two principal curvatures of Γ at p [11, Chapter 3.2]. We can therefore write it as $W =: SD_wS^{-1}$ with

$$D_w := \begin{pmatrix} \kappa_1(p) & 0 \\ 0 & \kappa_2(p) \end{pmatrix}$$

and S being orthonormal. Inserting this into the equation above leads to

$$\begin{aligned}
(D_{M((p,t))}\Phi)(D_{(p,t)}M)(D_{(p,t)}\Phi)^{-1} &= \left(\begin{array}{c|c} (I - tSD_wS^{-1})(I + tSD_wS^{-1})^{-1} & 0 \\ \hline 0 & -1 \end{array}\right) \\
&= \left(\begin{array}{c|c} S(I - tD_w)S^{-1}S(I + tD_w)^{-1}S^{-1} & 0 \\ \hline 0 & -1 \end{array}\right) \\
&= \left(\begin{array}{c|c} S(I - tD_w)(I + tD_w)^{-1}S^{-1} & 0 \\ \hline 0 & -1 \end{array}\right) \\
&= \tilde{S}\left(\begin{array}{c|c} (I - tD_w)(I + tD_w)^{-1} & 0 \\ \hline 0 & -1 \end{array}\right)\tilde{S}^{-1} \\
&= \tilde{S}\begin{pmatrix} \frac{1-t\kappa_1(p)}{1+t\kappa_1(p)} & 0 & 0 \\ 0 & \frac{1-t\kappa_2(p)}{1+t\kappa_2(p)} & 0 \\ 0 & 0 & -1 \end{pmatrix}\tilde{S}^{-1}
\end{aligned}$$

with

$$\tilde{S} := \left(\begin{array}{c|c} S & 0 \\ \hline 0 & 1 \end{array}\right).$$

This implies that

$$\|(D_{M((p,t))}\Phi)(D_{(p,t)}M)(D_{(p,t)}\Phi)^{-1}\|_{\mathcal{L}(\mathbb{R}^3)} = \max\left(\left|\frac{1-t\kappa_1(p)}{1+t\kappa_1(p)}\right|, \left|\frac{1-t\kappa_2(p)}{1+t\kappa_2(p)}\right|, 1\right).$$

Now the same calculations as in the two dimensional case lead to the claimed bounds. □

A.3 Bounds for Special Geometries

Finally, we apply the previous results to specific geometries which are used in our numerical experiments.

Corollary 1 *(Line or Plane) If the interface is a plane, the bounds are given by* $|R_\pm|_* \leq 1$.

Corollary 2 *(Part of a circle, sphere or cylinder) If the interface is a part of a sphere or cylinder with radius* r *and* $\delta < r$, *we have to distinguish if the vector pointing inwards points towards* Ω_- *or* Ω_+. *Then the following bounds hold:*

$$\begin{aligned}
&\textit{Towards } \Omega_-: && |R_-|_* \leq 1 && \textit{and} && |R_+|_* \leq \frac{r+\delta}{r-\delta}, \\
&\textit{Towards } \Omega_+: && |R_-|_* \leq \frac{r+\delta}{r-\delta} && \textit{and} && |R_+|_* \leq 1.
\end{aligned}$$

Proof For the simple geometries considered in the corollaries above, the principal curvatures are constant and either $\pm\frac{1}{r}$ or 0. By inserting these values into the bound from Theorem 4 the given bounds then follow immediately. □

Note that it is possible to combine these different parts to achieve interfaces which are rounded polygons or polyhedra.

Funding Open Access funding enabled and organized by Projekt DEAL.

Data Availability The code needed for reproducing all the numerical examples is available at https://doi.org/10.25625/BP0DPY (ref. [27]).

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Abdulle, A., Huber, M.E., Lemaire, S.: An optimization-based numerical method for diffusion problems with sign-changing coefficients. C.R. Math. **355**(4), 472–478 (2017). https://doi.org/10.1016/j.crma.2017.02.010
2. Abdulle, A., Lemaire, S.: An optimization-based method for sign-changing elliptic PDEs. ESAIM Math. Model. Numer. Anal. (2024). https://doi.org/10.1051/m2an/2024013
3. Aubry, A., Lei, D.Y., Fernández-Domínguez, A.I., Sonnefraud, Y., Maier, S.A., Pendry, J.B.: Plasmonic light-harvesting devices over the whole visible spectrum. Nano Lett. **10**(7), 2574–2579 (2010)
4. Barnes, W.L., Dereux, A., Ebbesen, T.W.: Surface plasmon subwavelength optics. Nature **424**(6950), 824–830 (2003)
5. Beyn, W.J.: An integral method for solving nonlinear eigenvalue problems. Linear Algebra Appl. **436**(10), 3839–3863 (2012)
6. Bonnet-Ben Dhia, A.S., Carvalho, C., Ciarlet, P.: Mesh requirements for the finite element approximation of problems with sign-changing coefficients. Numer. Math. **138**(4), 801–838 (2018). https://doi.org/10.1007/s00211-017-0923-5
7. Bonnet-Ben Dhia, A.S., Ciarlet, P., Jr., Zwölf, C.M.: Time harmonic wave diffraction problems in materials with sign-shifting coefficients. J. Comput. Appl. Math. **234**(6), 1912–1919 (2010)
8. Bonnet-BenDhia, A.S., Chesnel, L., Ciarlet, P.: T-coercivity for scalar interface problems between dielectrics and metamaterials. Math. Mod. Num. Anal. **46**, 363–1387 (2012). https://doi.org/10.1051/m2an/2012006
9. Bonnet-BenDhia, A.S., Chesnel, L., Ciarlet, P.: T-coercivity for the Maxwell problem with sign-changing coefficients. Comm. Partial Differential Equations **39**, 1007–1031 (2014). https://doi.org/10.1080/03605302.2014.892128
10. Burman, E., Ern, A., Preuss, J.: A stabilized hybridized Nitsche method for sign- changing elliptic PDEs. Math. Models Methods Appl. Sci. **35**(14), 2977–3009 (2025). https://doi.org/10.1142/S021820252550054X
11. do Carmo, M.P.: Differential geometry of curves and surfaces. Prentice-Hall, Inc. (1976)
12. Carvalho, C., Chesnel, L., Ciarlet, P., Jr.: Eigenvalue problems with sign-changing coefficients. C. R. Math. Acad. Sci. Paris **355**(6), 671–675 (2017). https://doi.org/10.1016/j.crma.2017.05.002
13. Cassier, M., Joly, P., Kachanovska, M.: Mathematical models for dispersive electromagnetic waves: an overview. Comput. Math. Appl. **74**(11), 2792–2830 (2017). https://doi.org/10.1016/j.camwa.2017.07.025
14. Chaumont-Frelet, T., Verfürth, B.: A generalized finite element method for problems with sign-changing coefficients. ESAIM Math. Model. Numer. Anal. **55**(3), 939–967 (2021). https://doi.org/10.1051/m2an/2021007
15. Chesnel, L., Ciarlet, P.j.: T-coercivity and continuous Galerkin methods: application to transmission problems with sign changing coefficients. Numer. Math. **124**(1), 1–29 (2013). https://doi.org/10.1007/s00211-012-0510-8
16. Ciarlet, P., Jamelot, E.: The T-coercivity approach for solving Stokes problem: stabilization of finite element pairs (2024). https://inria.hal.science/hal-04414789. Hal-report

17. Ciarlet, P., Jr., Lassounon, D., Rihani, M.: An optimal control-based numerical method for scalar transmission problems with sign-changing coefficients. SIAM J. Numer. Anal. **61**(3), 1316–1339 (2023). https://doi.org/10.1137/22M1495998
18. Ciarlet, P., Jr., Vohralík, M.: Localization of global norms and robust a posteriori error control for transmission problems with sign-changing coefficients. ESAIM Math. Model. Numer. Anal. **52**(5), 2037–2064 (2018). https://doi.org/10.1051/m2an/2018034
19. Davis, P.J., Rabinowitz, P.: Methods of numerical integration. Courier Corporation (2007)
20. Grabovsky, Y.: Reconstructing Stieltjes functions from their approximate values: a search for a needle in a haystack. SIAM J. Appl. Math. **82**(4), 1135–1166 (2022). https://doi.org/10.1137/21M1392279
21. Halla, M.: On the approximation of dispersive electromagnetic eigenvalue problems in two dimensions. IMA J. Numer. Anal. **43**(1), 535–559 (2023). https://doi.org/10.1093/imanum/drab100
22. Karma, O.: Approximation in eigenvalue problems for holomorphic Fredholm operator functions. I. Numer. Funct. Anal. Optim. **17**(3–4), 365–387 (1996). https://doi.org/10.1080/01630569608816699
23. Karma, O.: Approximation in eigenvalue problems for holomorphic Fredholm operator functions. II. (Convergence rate). Numer. Funct. Anal. Optim. **17**(3-4), 389–408 (1996). https://doi.org/10.1080/01630569608816700
24. Kress, R.: Linear integral equations, *Appl. Math. Sci.*, vol. 82, 3rd ed. edn. New York, NY: Springer (2014). https://doi.org/10.1007/978-1-4614-9593-2
25. Nguyen, H.M.: Limiting absorption principle and well-posedness for the helmholtz equation with sign changing coefficients. Journal de Mathématiques Pures et Appliquées **106**(2), 342–374 (2016)
26. Nicaise, S., Venel, J.: A posteriori error estimates for a finite element approximation of transmission problems with sign changing coefficients. J. Comput. Appl. Math. **235**(14), 4272–4282 (2011). https://doi.org/10.1016/j.cam.2011.03.028
27. Oberender, F.: Replication Data for: A new numerical method for scalar eigenvalue problems in heterogeneous, dispersive, sign-changing materials (2024). https://doi.org/10.25625/BP0DPY
28. Schöberl, J.: C++11 implementation of finite elements in NGSolve. Institut für Analysis und Scientific Computing, TU Wien, Technical Report 30/2014 (2014)
29. Unger, G.: Convergence analysis of a Galerkin boundary element method for electromagnetic resonance problems. Partial Differ. Equ. Appl. **2**(3), Paper No. 39 (2021). https://doi.org/10.1007/s42985-020-00049-5
30. Wloka, J.: Partial Differential Equations. Cambridge University Press (1987)