An optimization-based numerical method for diffusion problems with sign-changing coefficients

Assyr Abdulle, Martin E. Huber, Simon Lemaire

An optimization-based numerical method for diffusion problems with sign-changing coefficients

Assyr Abdulle [a], Martin E. Huber [a], Simon Lemaire [a]
[a] *École Polytechnique Fédérale de Lausanne (EPFL), SB-MATHICSE-ANMC, Station 8, 1015 Lausanne, Switzerland*

Abstract

A new optimization-based numerical method is proposed for the solution of diffusion problems with sign-changing conductivity coefficients. In contrast to existing approaches, our method does not rely on the discretization of a stabilized equation and the convergence of the scheme can be proved without any symmetry assumption on the mesh near the interface where the conductivity sign changes.

Résumé

Une méthode d'optimisation pour des problèmes de diffusion avec changement de signe. Nous proposons une nouvelle méthode, basée sur la résolution d'un problème de minimisation, pour l'approximation de problèmes de diffusion avec changement de signe. Cette approche, qui tire profit d'une reformulation du modèle initial sous la forme d'un problème de transmission, ne repose pas sur la discrétisation d'une équation stabilisée, et la convergence de la méthode est obtenue sans hypothèse de symétrie du maillage dans un voisinage de l'interface où la conductivité change de signe.

Version française abrégée

Dans cette note, nous introduisons une méthode d'optimisation pour l'approximation numérique de problèmes de diffusion dont la conductivité change de signe dans le domaine. La résolution numérique efficace de ce genre de problèmes est importante pour de nombreuses applications (e.g., super-lentilles, invisibilité), mais les méthodes existantes ne sont pour l'instant pas satisfaisantes. Dans [6], les deux approches envisagées reposent (i) sur la discrétisation d'une équation stabilisée, pour laquelle les taux de convergence obtenus sont sous-optimaux, ou (ii) sur des hypothèses de symétrie du maillage autour de l'interface où la conductivité change de signe, exigences pouvant s'avérer très contraignantes pour des interfaces générales (voir [3]) ou en 3D. La méthode numérique introduite ici, qui utilise une reformulation du modèle initial en un problème de transmission, ne repose pas sur l'ajout de dissipation à l'équation, et nous montrons sa convergence pour des problèmes elliptiques présentant un changement de signe sans aucune hypothèse de symétrie sur le maillage. Nous notons que cette méthode numérique a pour la première fois été introduite par Gunzburger et al. [9] (voir aussi [8]), dans un contexte de décomposition de domaine pour des équations elliptiques classiques, sans preuve de convergence. L'application de cet algorithme à des problèmes elliptiques présentant un changement de signe est introduite dans cette note, et la convergence de la méthode est démontrée.

Email addresses: assyr.abdulle@epfl.ch (Assyr Abdulle), martin.huber@epfl.ch (Martin E. Huber), simon.lemaire@epfl.ch (Simon Lemaire).

1. Introduction

Partial differential equations with sign-changing coefficients play an increasingly important role in the modeling of metamaterials, with applications ranging from superlensing to cloaking. In this paper, we consider in an open bounded polytopal domain $\Omega \subset \mathbb{R}^d$, $d \in \{2,3\}$ such that $\overline{\Omega} = \overline{\Omega_{\rm e}} \cup \overline{\Omega_{\rm i}}$, with $\Omega_{\rm e}, \Omega_{\rm i}$ disjoint open polytopal subsets of Ω with nonzero measure, the following sign-changing diffusion problem

$$-\mathrm{div}(s\,\mathbb{K}\boldsymbol{\nabla}\tilde{u}) = f \ \text{ in } \Omega, \qquad \tilde{u} = 0 \ \text{ on } \partial\Omega, \tag{1}$$

where $\mathbb{K}$ is a symmetric, uniformly elliptic and bounded matrix-valued field, and $s : \Omega \to \mathbb{R}$ is a sign function such that $s_{\rm e} := s_{|\Omega_{\rm e}} = 1$ and $s_{\rm i} := s_{|\Omega_{\rm i}} = -1$. The subscripts 'e' and 'i' respectively stand for the exterior (for which we assume that $\partial\Omega_{\rm e} \cap \partial\Omega$ has nonzero measure) and the interior subdomains, and we denote their interface $\Gamma = \partial\Omega_{\rm e} \cap \partial\Omega_{\rm i}$. We assume that $f \in L^2(\Omega)$ and study the following weak formulation of Problem (1): Find $\tilde{u} \in H^1_0(\Omega)$ such that

$$a(\tilde{u},\varphi) := (s\,\mathbb{K}\boldsymbol{\nabla}\tilde{u}, \boldsymbol{\nabla}\varphi)_\Omega = (f,\varphi)_\Omega \qquad \forall \varphi \in H^1_0(\Omega). \tag{2}$$

Several approaches have been developed to study the well-posedness of Problem (2). We first mention the T-coercivity theory of Bonnet-Ben Dhia et al. [4]. In this context, well-posedness (in the classical Hadamard sense or in the Fredholm sense) of Problem (2) is equivalent to finding a linear, bounded and bijective operator $\mathtt{T} : H^1_0(\Omega) \to H^1_0(\Omega)$ such that the bilinear form $a(\cdot, \mathtt{T}\cdot)$ is coercive (for classical well-posedness) or weakly coercive (for the Fredholm case). The operator $\mathtt{T}$ plays the role of an explicit inf − sup operator. However, proving T-coercivity can be difficult for complex geometries, especially in 3D. More recently, another viewpoint has come out that consists in studying the "limit" as $\delta \to 0^+$ of the well-posed stabilized problem: Find $\tilde{\mathrm{u}}_\delta \in \mathrm{H}^1_0(\Omega)$ such that

$$\mathrm{a}_\delta(\tilde{\mathrm{u}}_\delta,\varphi) := a(\tilde{\mathrm{u}}_\delta,\varphi) - \mathrm{i}\delta(\mathbb{K}\boldsymbol{\nabla}\tilde{\mathrm{u}}_\delta, \boldsymbol{\nabla}\varphi)_{\Omega_{\rm i}} = (f,\varphi)_\Omega \qquad \forall \varphi \in \mathrm{H}^1_0(\Omega), \tag{3}$$

where $\mathrm{H}^1_0(\Omega)$ is complex-valued. This approach has been developed in the context of Helmholtz equations by Nguyen [11,12], and makes use of transport operators called reflections.

Different strategies have been studied by Chesnel and Ciarlet Jr. in [6] for the numerical approximation of Problem (2) (assuming that it is well-posed in the classical Hadamard sense). The first approach is based on a simplicial discretization $\mathcal{T}_h$ of Ω, that respects the interface Γ and the construction of a conforming finite element space $V_0(\mathcal{T}_h)$ that is stable by the operator $\mathtt{T}$. Well-posedness and optimal convergence rates can be shown for this approach. In practice, T-stability is achieved by means of symmetric meshes near the interface, whose construction is a nontrivial task for complicated interfaces (see [3]) or 3D problems. On general meshes, two main approaches have been investigated by Chesnel and Ciarlet Jr. The first one consists in building a (mesh-dependent) operator $\mathtt{T}_h$, such that the bilinear form a is $\mathtt{T}_h$-coercive on $V_0(\mathcal{T}_h)$ (see, e.g., [13]). This kind of approach is limited by the fact that, in general, well-posedness cannot be proved for the whole range of admissible coefficients. The second idea consists in discretizing the stabilized equation (3), and in scaling the dissipation δ as a function of h. However, and as studied in [6], this approach leads to suboptimal convergence rates.

In this note, we aim at proposing a new discretization approach, that is based on a reformulation of Problem (1) as a transmission problem, whose solution is obtained numerically from an optimization procedure. The approach we propose is proved to converge, and does not rely on any symmetry assumption on the mesh. As opposed to T-coercivity theory, we do not need to assume well-posedness of Problem (2) in the Hadamard sense (i.e. for any $f \in H^{-1}(\Omega)$) to prove convergence of our numerical method. We only need to assume that, for the given $f \in L^2(\Omega)$ we consider, the solution to Problem (2) exists and is unique. This allows us to treat cases that cannot be analyzed using the T-coercivity theory (cf. Remark 3). The numerical method we consider here has been introduced by Gunzburger et al. [9] (see also [8]), in the context of domain decomposition for classical elliptic equations. However, up to now, the convergence

of this method has never been proved. We provide the first proof of convergence of such a numerical method, with application to sign-changing elliptic equations, for which this approach seems particularly promising.

2. Transmission problem and numerical method

We provide in this section, under suitable regularity assumptions, an alternative characterization of the solution to Problem (2), that will be our starting point for the design of the numerical method.

2.1. *Transmission problem*

Recall that $f \in L^2(\Omega)$. Hence, $s\,\mathbb{K}\boldsymbol{\nabla}\tilde{u} \in \boldsymbol{H}(\mathrm{div};\Omega)$. Denoting, for $\alpha \in \{\mathrm{e},\mathrm{i}\}$, by $\boldsymbol{n}_\alpha$ the unit normal vector to Γ pointing out of Ω_α, and introducing $\tilde{g}$ such that

$$\tilde{g} := \Big[(\mathbb{K}\boldsymbol{\nabla}\tilde{u})_{|\Omega_\mathrm{e}}\cdot\boldsymbol{n}_\mathrm{e}\Big]_{|\Gamma}, \tag{4}$$

we have that $\tilde{g} \in H^{-1/2}(\Gamma)$. Here, $H^{-1/2}(\Gamma)$ denotes the dual of $H^{1/2}(\partial\Omega_\mathrm{i})$ when $\Gamma = \partial\Omega_\mathrm{i}$ or of $H^{1/2}_{00}(\Gamma)$ otherwise. For $\alpha \in \{\mathrm{e},\mathrm{i}\}$, and for any $g \in H^{-1/2}(\Gamma)$, we consider in the subdomain Ω_α the problem

$$s_\alpha(\mathbb{K}\boldsymbol{\nabla}u_\alpha(g),\boldsymbol{\nabla}\varphi_\alpha)_{\Omega_\alpha} = (f,\varphi_\alpha)_{\Omega_\alpha} + s_\alpha\langle g,\varphi_\alpha\rangle_\Gamma \qquad \forall\varphi_\alpha \in H^1_{0\backslash\Gamma}(\Omega_\alpha), \tag{5}$$

where $H^1_{0\backslash\Gamma}(\Omega_\alpha)$ is the space of functions in $H^1(\Omega_\alpha)$ that vanish on $\partial\Omega_\alpha\backslash\Gamma$. In Ω_e, Problem (5) always admits a unique solution $u_\mathrm{e}(g) \in H^1_{0\backslash\Gamma}(\Omega_\mathrm{e})$, as the measure of $\partial\Omega_\mathrm{e}\cap\partial\Omega$ is nonzero. In Ω_i, Problem (5) also admits a unique solution $u_\mathrm{i}(g) \in H^1_{0\backslash\Gamma}(\Omega_\mathrm{i})$ if the measure of $\partial\Omega_\mathrm{i}\cap\partial\Omega$ is nonzero. Otherwise, the problem in Ω_i is purely Neumann and we assume that the flux $g \in H^{-1/2}(\Gamma)$ satisfies $(f,1)_{\Omega_\mathrm{i}} - \langle g,1\rangle_\Gamma = 0$ to ensure that Problem (5) admits a solution, that is unique up to an additive constant. We fix the constant by imposing $(u_\mathrm{i}(g),1)_\Gamma = (u_\mathrm{e}(g),1)_\Gamma$. Finally, for $g \in H^{-1/2}(\Gamma)$, we denote by $u(g)$ the function such that $u(g)_{|\Omega_\alpha} := u_\alpha(g)$, $\alpha \in \{\mathrm{e},\mathrm{i}\}$.

Proposition 2.1 (Characterization of the solution to Problem (2)) *We assume that (2) admits a unique solution $\tilde{u} \in H^1_0(\Omega)$, and that there exists $\tilde{s} > \frac{1}{2}$ such that $\tilde{u}_{|\Omega_\alpha} \in H^{1+\tilde{s}}(\Omega_\alpha)$ for $\alpha \in \{\mathrm{e},\mathrm{i}\}$. Then,*

- *the flux $\tilde{g}$ defined in (4) belongs to $L^2(\Gamma)$, and satisfies $\tilde{g} = -\Big[(-\mathbb{K}\boldsymbol{\nabla}\tilde{u})_{|\Omega_\mathrm{i}}\cdot\boldsymbol{n}_\mathrm{i}\Big]_{|\Gamma}$;*
- *$u(\tilde{g}) = \tilde{u} \in H^1_0(\Omega)$;*
- *almost everywhere on the interface Γ, $u_\mathrm{e}(\tilde{g}) = u_\mathrm{i}(\tilde{g})$;*
- *the problem $\inf\limits_{g\in L^2(\Gamma)} \|u_\mathrm{e}(g) - u_\mathrm{i}(g)\|^2_{0,\Gamma}$ admits $\tilde{g}$ as its unique solution.*

Remark 1 (Assumptions on $\tilde{u}$) The existence (and uniqueness) of a solution $\tilde{u} \in H^1_0(\Omega)$ to Problem (2) is satisfied in practice in a large variety of situations (cf., e.g., [12]). The assumption $\tilde{u}_{|\Omega_\alpha} \in H^{1+\tilde{s}}(\Omega_\alpha)$ for $\alpha \in \{\mathrm{e},\mathrm{i}\}$, $\tilde{s} > \frac{1}{2}$, made in Proposition 2.1 is convenient for the analysis, as it enables to work in $L^2(\Gamma)$ instead of working in $H^{-1/2}(\Gamma)$. This theoretical assumption is quite strict, but not mandatory in practice for the method to be applicable. In a forthcoming work [1], we will consider a test-case for which this regularity assumption is violated, and show that numerical convergence can still be observed.

2.2. *Minimization problem and numerical method*

We consider a family of simplicial conformal discretizations $\{\mathcal{T}_h\}_h$ of Ω, that respects the interface Γ and is shape-regular in the sense of Ciarlet [7]. The subscript h stands for the meshsize, i.e. the maximum diameter of all the simplices in $\mathcal{T}_h$. We denote by Γ_h the set of faces of the mesh $\mathcal{T}_h$ belonging to Γ. For $k \in \mathbb{N}^\star$ and $\alpha \in \{\mathrm{e},\mathrm{i}\}$, we introduce the space $V^k(\mathcal{T}_{h,\alpha}) := \left\{v \in H^1(\Omega_\alpha) \mid v_{|T} \in \mathbb{P}^k_d(T),\ \forall T \in \mathcal{T}_{h,\alpha}\right\}$, where $\mathbb{P}^k_d(T)$ is the space of d-variate polynomial functions of total degree less or equal to k in $T \in \mathcal{T}_h$. We also introduce the subspace $V^k_{0\backslash\Gamma}(\mathcal{T}_{h,\alpha}) := V^k(\mathcal{T}_{h,\alpha}) \cap H^1_{0\backslash\Gamma}(\Omega_\alpha)$.

For $\alpha \in \{\mathrm{e},\mathrm{i}\}$, and for any $g \in H^{-1/2}(\Gamma)$, we consider, in the subdomain Ω_α, the following conforming approximation of Problem (5):

$$s_\alpha(\mathbb{K}\boldsymbol{\nabla} u_{h,\alpha}(g), \boldsymbol{\nabla}\varphi_{h,\alpha})_{\Omega_\alpha} = (f, \varphi_{h,\alpha})_{\Omega_\alpha} + s_\alpha\langle g, \varphi_{h,\alpha}\rangle_\Gamma \qquad \forall \varphi_{h,\alpha} \in V^k_{0\backslash\Gamma}(\mathcal{T}_{h,\alpha}). \tag{6}$$

As in the continuous case, Problem (6) always admits a unique solution $u_{h,\mathrm{e}}(g) \in V^k_{0\backslash\Gamma}(\mathcal{T}_{h,\mathrm{e}})$ in Ω_e, and $u_{h,\mathrm{i}}(g) \in V^k_{0\backslash\Gamma}(\mathcal{T}_{h,\mathrm{i}})$ in Ω_i when the measure of $\partial\Omega_\mathrm{i} \cap \partial\Omega$ is nonzero. For the case of a pure Neumann problem in Ω_i, we assume that $(f,1)_{\Omega_\mathrm{i}} - \langle g, 1\rangle_\Gamma = 0$, ensuring existence of a solution and uniqueness up to an additive constant, that we fix imposing $(u_{h,\mathrm{i}}(g),1)_\Gamma = (u_{h,\mathrm{e}}(g),1)_\Gamma$. For $g \in H^{-1/2}(\Gamma)$, we introduce $u_h(g)$ such that $u_h(g)_{|\Omega_\alpha} := u_{h,\alpha}(g)$, $\alpha \in \{\mathrm{e},\mathrm{i}\}$.

To define the minimization problem, we introduce $F^k(\Gamma_h) := \left\{q \in L^2(\Gamma) \mid q_{|F} \in \mathbb{P}^k_{d-1}(F),\ \forall F \in \Gamma_h\right\}$ when the measure of $\partial\Omega_\mathrm{i} \cap \partial\Omega$ is nonzero, and we add the constraint $(q,1)_\Gamma = (f,1)_{\Omega_\mathrm{i}}$ otherwise. We then define the functional $J_h : F^k(\Gamma_h) \to \mathbb{R}_+$ such that $J_h(g_h) := \|u_{h,\mathrm{e}}(g_h) - u_{h,\mathrm{i}}(g_h)\|^2_{0,\Gamma} + \lambda(h)\|g_h\|^2_{0,\Gamma}$, where $\lambda : \mathbb{R}^\star_+ \to \mathbb{R}^\star_+$ is a function such that $\lim\limits_{h\to 0}\lambda(h) = 0$, and we consider the minimization problem

$$\inf_{g_h \in F^k(\Gamma_h)} J_h(g_h). \tag{7}$$

The continuous function J_h is coercive on $F^k(\Gamma_h)$, and hence admits at least one minimizer: there exists $\tilde{g}_h \in F^k(\Gamma_h)$ such that

$$J_h(\tilde{g}_h) \leqslant J_h(g_h) \quad \text{for any } g_h \in F^k(\Gamma_h). \tag{8}$$

The approximation of $\tilde{u}$ we then consider is $u_h(\tilde{g}_h)$. The following result is a direct consequence of (8) and of the properties of the L^2-orthogonal projector from $L^2(\Gamma)$ onto $F^k(\Gamma_h)$.

Lemma 2.1 (Estimate on $J_h(\tilde{g}_h)$) *Assume that $\tilde{g}$ defined in (4) is in $L^2(\Gamma)$. Then, the following holds:*

$$J_h(\tilde{g}_h) \leqslant \|u_{h,\mathrm{e}}(\tilde{g}) - u_{h,\mathrm{i}}(\tilde{g})\|^2_{0,\Gamma} + \lambda(h)\|\tilde{g}\|^2_{0,\Gamma}. \tag{9}$$

3. Convergence

From now on, we write $A \lesssim B$ when there exists a constant $c > 0$, possibly depending on $\mathbb{K}$, k, and on the geometry, but independent of h, and of $\tilde{u}$, $\tilde{g}$, such that $A \leqslant cB$.

Lemma 3.1 (Approximation properties) *Suppose that the assumptions of Proposition 2.1 are fulfilled. Then, letting $p = \min(k, \tilde{s})$, there exists $\delta \in \left(0, \frac{1}{2}\right]$ such that*

$$\|\boldsymbol{\nabla}_h(u_h(\tilde{g}) - \tilde{u})\|_{0,\Omega} \lesssim h^p\left(|\tilde{u}|_{1+p,\Omega_\mathrm{e}} + |\tilde{u}|_{1+p,\Omega_\mathrm{i}}\right),$$
$$\|u_{h,\mathrm{e}}(\tilde{g}) - u_{h,\mathrm{i}}(\tilde{g})\|_{0,\Gamma} \leqslant \sum_{\alpha\in\{\mathrm{e},\mathrm{i}\}} \|u_{h,\alpha}(\tilde{g}) - \tilde{u}\|_{0,\Gamma} \lesssim h^{p+\delta}\left(|\tilde{u}|_{1+p,\Omega_\mathrm{e}} + |\tilde{u}|_{1+p,\Omega_\mathrm{i}}\right).$$

Proof We first assume that we have, both in $\Omega_{\rm e}$ and $\Omega_{\rm i}$, mixed Dirichlet–Neumann boundary conditions. Then, classical arguments (particularly, Aubin–Nitsche duality argument, and a multiplicative continuous trace inequality) yield

$$\|u_{h,\alpha}(\tilde{g}) - \tilde{u}\|_{0,\Omega_\alpha} + h^{\frac{\delta_\alpha^{\rm m}}{2}} \|u_{h,\alpha}(\tilde{g}) - \tilde{u}\|_{0,\Gamma} + h^{\delta_\alpha^{\rm m}} \|\boldsymbol{\nabla}(u_{h,\alpha}(\tilde{g}) - \tilde{u})\|_{0,\Omega_\alpha} \lesssim h^{p+\delta_\alpha^{\rm m}} |\tilde{u}|_{1+p,\Omega_\alpha}, \tag{10}$$

for $\alpha \in \{{\rm e}, {\rm i}\}$ and with $\delta_\alpha^{\rm m} \in (0,1]$ only depending on the geometry of Ω_α. By setting $\delta = \min\left(\frac{\delta_{\rm e}^{\rm m}}{2}, \frac{\delta_{\rm i}^{\rm m}}{2}\right)$, we obtain the stated result. We next treat the case of a pure Neumann problem in $\Omega_{\rm i}$. Due to the choice $(u_{h,{\rm i}}(\tilde{g}), 1)_\Gamma = (u_{h,{\rm e}}(\tilde{g}), 1)_\Gamma$ to fix the constant, a straightforward Aubin–Nitsche duality argument fails. We therefore consider an auxiliary function $\overline{u}_{h,{\rm i}}(\tilde{g})$, defined as the unique discrete solution of the same Neumann problem in $\Omega_{\rm i}$, but for which we fix the constant as in the continuous problem, namely $(\overline{u}_{h,{\rm i}}(\tilde{g}), 1)_{\Omega_{\rm i}} = (\tilde{u}, 1)_{\Omega_{\rm i}}$. Then, there exists $\delta_{\rm i}^{\rm n} \in (0,1]$, only depending on the geometry of $\Omega_{\rm i}$, such that

$$\|\overline{u}_{h,{\rm i}}(\tilde{g}) - \tilde{u}\|_{0,\Omega_{\rm i}} + h^{\frac{\delta_{\rm i}^{\rm n}}{2}} \|\overline{u}_{h,{\rm i}}(\tilde{g}) - \tilde{u}\|_{0,\Gamma} + h^{\delta_{\rm i}^{\rm n}} \|\boldsymbol{\nabla}(\overline{u}_{h,{\rm i}}(\tilde{g}) - \tilde{u})\|_{0,\Omega_{\rm i}} \lesssim h^{p+\delta_{\rm i}^{\rm n}} |\tilde{u}|_{1+p,\Omega_{\rm i}}. \tag{11}$$

Note that, a priori, $\delta_{\rm i}^{\rm n} \neq \delta_{\rm i}^{\rm m}$ since regularity results for pure Neumann or mixed problems are different in general. Next, we observe that $u_{h,{\rm i}}(\tilde{g}) - \overline{u}_{h,{\rm i}}(\tilde{g}) = c_h$, for a constant $c_h \in \mathbb{R}$. We then write

$$|\Gamma|\, c_h = (u_{h,{\rm i}}(\tilde{g}) - \overline{u}_{h,{\rm i}}(\tilde{g}), 1)_\Gamma = (u_{h,{\rm e}}(\tilde{g}) - \overline{u}_{h,{\rm i}}(\tilde{g}), 1)_\Gamma = (u_{h,{\rm e}}(\tilde{g}) - \tilde{u}, 1)_\Gamma + (\tilde{u} - \overline{u}_{h,{\rm i}}(\tilde{g}), 1)_\Gamma,$$

which yields, using (10) and (11),

$$|c_h| \leqslant |\Gamma|^{-1/2} \left(\|u_{h,{\rm e}}(\tilde{g}) - \tilde{u}\|_{0,\Gamma} + \|\overline{u}_{h,{\rm i}}(\tilde{g}) - \tilde{u}\|_{0,\Gamma} \right) \lesssim h^{p+\frac{\delta_{\rm e}^{\rm m}}{2}} |\tilde{u}|_{1+p,\Omega_{\rm e}} + h^{p+\frac{\delta_{\rm i}^{\rm n}}{2}} |\tilde{u}|_{1+p,\Omega_{\rm i}}. \tag{12}$$

We finally obtain from (12)

$$\|u_{h,{\rm i}}(\tilde{g}) - \tilde{u}\|_{0,\Gamma} \leqslant \|u_{h,{\rm e}}(\tilde{g}) - \tilde{u}\|_{0,\Gamma} + 2\|\overline{u}_{h,{\rm i}}(\tilde{g}) - \tilde{u}\|_{0,\Gamma} \lesssim h^{p+\frac{\delta_{\rm e}^{\rm m}}{2}} |\tilde{u}|_{1+p,\Omega_{\rm e}} + h^{p+\frac{\delta_{\rm i}^{\rm n}}{2}} |\tilde{u}|_{1+p,\Omega_{\rm i}},$$

which enables to conclude the proof, setting in that case $\delta = \min\left(\frac{\delta_{\rm e}^{\rm m}}{2}, \frac{\delta_{\rm i}^{\rm n}}{2}\right)$. □

We note that, as a consequence of Lemma 3.1, it is always possible to choose $\lambda(h)$ in (7) such that $\|u_{h,{\rm e}}(\tilde{g}) - u_{h,{\rm i}}(\tilde{g})\|_{0,\Gamma}^2 = O(\lambda(h))$ or $\|u_{h,{\rm e}}(\tilde{g}) - u_{h,{\rm i}}(\tilde{g})\|_{0,\Gamma}^2 = o(\lambda(h))$. We then have the

Theorem 3.2 (Convergence of the method) *Suppose that the assumptions of Proposition 2.1 are fulfilled. Denote $u_h(\tilde{g}_h)$ by $\tilde{u}_h$. Then, up to a choice of $\lambda(h)$ in (7) such that $\|u_{h,{\rm e}}(\tilde{g}) - u_{h,{\rm i}}(\tilde{g})\|_{0,\Gamma}^2 = O(\lambda(h))$, there holds as $h \to 0$:*

$$\tilde{g}_h \rightharpoonup \tilde{g} \quad \text{in } L^2(\Gamma), \qquad \boldsymbol{\nabla}_h \tilde{u}_h \rightharpoonup \boldsymbol{\nabla} \tilde{u} \quad \text{in } L^2(\Omega)^d, \qquad \tilde{u}_h \to \tilde{u} \quad \text{in } L^2(\Omega). \tag{13}$$

If we further choose $\lambda(h)$ such that $\|u_{h,{\rm e}}(\tilde{g}) - u_{h,{\rm i}}(\tilde{g})\|_{0,\Gamma}^2 = o(\lambda(h))$, then we have

$$\tilde{g}_h \to \tilde{g} \quad \text{in } L^2(\Gamma), \qquad \boldsymbol{\nabla}_h \tilde{u}_h \to \boldsymbol{\nabla} \tilde{u} \quad \text{in } L^2(\Omega)^d, \qquad \tilde{u}_h \to \tilde{u} \quad \text{in } L^2(\Omega). \tag{14}$$

Proof We begin by remarking that, for $\alpha \in \{{\rm e}, {\rm i}\}$,

$$u_{h,\alpha}(\tilde{g}_h) = u_{h,\alpha}(\tilde{g}) + v_{h,\alpha}, \tag{15}$$

where $v_{h,\alpha} \in V_{0\backslash\Gamma}^k(\mathcal{T}_{h,\alpha})$ satisfies

$$(\mathbb{K}\boldsymbol{\nabla} v_{h,\alpha}, \boldsymbol{\nabla}\varphi_{h,\alpha})_{\Omega_\alpha} = ((\tilde{g}_h - \tilde{g}), \varphi_{h,\alpha})_\Gamma \qquad \forall \varphi_{h,\alpha} \in V_{0\backslash\Gamma}^k(\mathcal{T}_{h,\alpha}). \tag{16}$$

From (16), we readily infer

$$\|\boldsymbol{\nabla} v_{h,\alpha}\|_{0,\Omega_\alpha}^2 \lesssim \|\tilde{g}_h - \tilde{g}\|_{0,\Gamma} \|v_{h,\alpha}\|_{0,\Gamma}. \tag{17}$$

Besides, the following estimate can be obtained from (9):

$$\|\tilde{g}_h\|_{0,\Gamma}^2 \leqslant \frac{\|u_{h,\mathrm{e}}(\tilde{g}) - u_{h,\mathrm{i}}(\tilde{g})\|_{0,\Gamma}^2}{\lambda(h)} + \|\tilde{g}\|_{0,\Gamma}^2. \tag{18}$$

As soon as $\|u_{h,\mathrm{e}}(\tilde{g}) - u_{h,\mathrm{i}}(\tilde{g})\|_{0,\Gamma}^2/\lambda(h) \leqslant c$, the combination of (18), (17), and of a continuous trace inequality (as well as of an appropriate Poincaré inequality [5, Equation (1.1)] when the measure of $\partial\Omega_\mathrm{i} \cap \partial\Omega$ is zero) enables to infer that there exist $\tilde{g}_0 \in L^2(\Gamma)$, and $v_\alpha \in H^1_{0\backslash\Gamma}(\Omega_\alpha)$, $\alpha \in \{\mathrm{e},\mathrm{i}\}$, so that the following convergences hold as $h \to 0$, up to a subsequence (retaining the same notation):

$$\begin{aligned} \tilde{g}_h &\rightharpoonup \tilde{g}_0 \quad && \text{in } L^2(\Gamma), & \boldsymbol{\nabla} v_{h,\alpha} &\rightharpoonup \boldsymbol{\nabla} v_\alpha \quad && \text{in } L^2(\Omega_\alpha)^d, \\ v_{h,\alpha} &\to v_\alpha \quad && \text{in } L^2(\Omega_\alpha), & v_{h,\alpha} &\rightharpoonup v_\alpha \quad && \text{in } L^2(\Gamma). \end{aligned} \tag{19}$$

Now, from (9) we infer

$$\|u_{h,\mathrm{e}}(\tilde{g}_h) - u_{h,\mathrm{i}}(\tilde{g}_h)\|_{0,\Gamma}^2 \leqslant \|u_{h,\mathrm{e}}(\tilde{g}) - u_{h,\mathrm{i}}(\tilde{g})\|_{0,\Gamma}^2 + \lambda(h)\|\tilde{g}\|_{0,\Gamma}^2 \leqslant \left(c + \|\tilde{g}\|_{0,\Gamma}^2\right)\lambda(h). \tag{20}$$

Owing to the fact that $\lambda(h)$ tends to zero as h vanishes, we deduce from (20) that $\|u_{h,\mathrm{e}}(\tilde{g}_h) - u_{h,\mathrm{i}}(\tilde{g}_h)\|_{0,\Gamma} \to 0$ as $h \to 0$. Combining this last result with (15) and Lemma 3.1, we get that $\|v_{h,\mathrm{e}} - v_{h,\mathrm{i}}\|_{0,\Gamma} \to 0$ as $h \to 0$ and hence, owing to (19), that $v_\mathrm{e} = v_\mathrm{i}$ almost everywhere on Γ. Passing to the limit in (16) using (19) and a strongly converging interpolant for test functions, it can be shown that $v \in H^1_0(\Omega)$ such that $v_{|\Omega_\alpha} := v_\alpha$ for $\alpha \in \{\mathrm{e},\mathrm{i}\}$ satisfies $(s\,\mathbb{K}\boldsymbol{\nabla} v, \boldsymbol{\nabla}\varphi)_\Omega = 0$ for all $\varphi \in H^1_0(\Omega)$, which implies, from the well-posedness of Problem (2), that $v \equiv 0$ and $\tilde{g}_0 \equiv \tilde{g}$ (this last result is inferred considering the limit equation in a subdomain, and using a density argument along with the fact that $\tilde{g}_0, \tilde{g} \in L^2(\Gamma)$). The uniqueness of the limits implies that the whole sequences converge in (19). Collecting these last results, (15), and Lemma 3.1, we prove (13). If we further choose $\lambda(h)$ such that $\|u_{h,\mathrm{e}}(\tilde{g}) - u_{h,\mathrm{i}}(\tilde{g})\|_{0,\Gamma}^2/\lambda(h) \to 0$ as $h \to 0$, we get from (18) and from the weak convergence of $\tilde{g}_h$ towards $\tilde{g}$ that

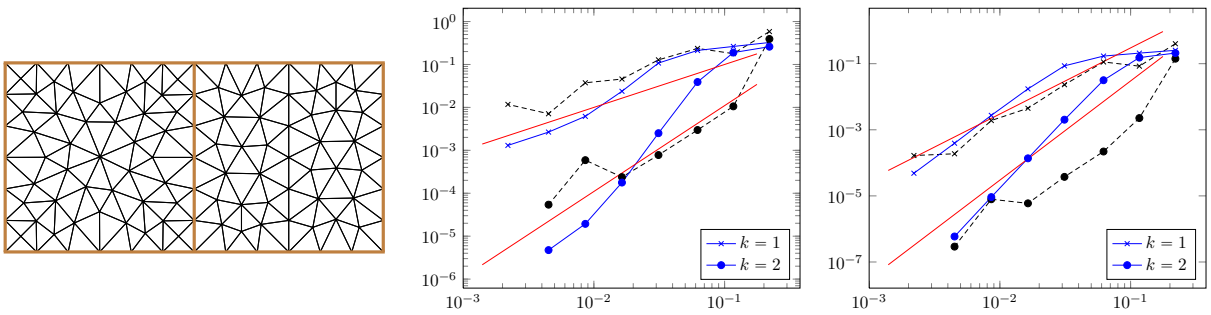
$$\|\tilde{g}\|_{0,\Gamma} \leqslant \liminf_{h\to 0} \|\tilde{g}_h\|_{0,\Gamma} = \limsup_{h\to 0} \|\tilde{g}_h\|_{0,\Gamma} \leqslant \|\tilde{g}\|_{0,\Gamma}, \tag{21}$$

which states the strong convergence of $\tilde{g}_h$ towards $\tilde{g}$ in $L^2(\Gamma)$. Testing (16) with $v_{h,\alpha}$, and using (21) combined with the weak convergence result for $v_{h,\alpha}$ in $L^2(\Gamma)$ of (19), we finally get the strong convergence of $\boldsymbol{\nabla} v_{h,\alpha}$ towards 0 in $L^2(\Omega_\alpha)^d$ for $\alpha \in \{\mathrm{e},\mathrm{i}\}$, proving (14) and concluding the proof. □

4. Numerical validation

We consider the 2D symmetric cavity test-case of [6], for which $\Omega := (-1,1) \times (0,1)$ and $\Omega_\mathrm{e} := (-1,0) \times (0,1)$ (it follows that $\Omega_\mathrm{i} = (0,1) \times (0,1)$ and $\Gamma = \{0\} \times [0,1]$). For this test-case, both the exterior and the interior problems feature mixed Dirichlet–Neumann boundary conditions. For the particular geometry considered here, elliptic regularity holds in both subdomains (see, e.g., [2, Remark I.3.6]), so that $\delta = \frac{1}{2}$ in Lemma 3.1. The tensor $\mathbb{K}$ is taken isotropic, with constant value in each subdomain ($\mathrm{k}_{|\Omega_\mathrm{e}} = \mathrm{k}_\mathrm{e} > 0$ and $\mathrm{k}_{|\Omega_\mathrm{i}} = \mathrm{k}_\mathrm{i} > 0$). For this particular setting, it is known [6] that Problem (2) is well-posed in the classical Hadamard sense if and only if the contrast $\nu := -\frac{\mathrm{k}_\mathrm{i}}{\mathrm{k}_\mathrm{e}}$ is different from -1. The exact solution $\tilde{u} \in H^1_0(\Omega)$ we consider is

$$\tilde{u}(x,y) := \begin{cases} \left((x+1)^2 - \dfrac{2\mathrm{k}_\mathrm{e} - \mathrm{k}_\mathrm{i}}{\mathrm{k}_\mathrm{e} - \mathrm{k}_\mathrm{i}}(x+1)\right)\sin(\pi y) & \text{in } \Omega_\mathrm{e}, \\ \dfrac{\mathrm{k}_\mathrm{e}}{\mathrm{k}_\mathrm{e} - \mathrm{k}_\mathrm{i}}(x-1)\sin(\pi y) & \text{in } \Omega_\mathrm{i}, \end{cases}$$



(a) Member of the mesh sequence (b) Relative H^1-norm error vs. h (c) Relative L^2-norm error vs. h

Figure 1. Relative errors on Ω vs. meshsize for the symmetric cavity test-case with $\nu = -1.001$.

and is associated to the right-hand side $f = -\mathrm{div}(s\,\mathrm{k}\boldsymbol{\nabla}\tilde{u}) \in L^2(\Omega)$. It can be easily seen that, for $\alpha \in \{\mathrm{e}, \mathrm{i}\}$, $\tilde{u}_{|\Omega_\alpha} \in H^{1+l}(\Omega_\alpha)$ for any $l > 0$, meaning that $p = k$ in Lemma 3.1.

As in [6], we choose $\mathrm{k_e} = 1$ and $\mathrm{k_i} = 1.001$, so that $\nu = -1.001$. We run the computations on a sequence of non-symmetric (with respect to the interface Γ), unstructured meshes. Results are depicted on Figure 1 for linear ($k = 1$) and second-order ($k = 2$) FEM, for our approach (solid blue), and for classical FEM (dashed black) applied to (2). For our approach, the parameter $\lambda(h)$ is chosen as $O(h^\beta)$ with $\beta \leqslant (2k+1)$, as required by Theorem 3.2 and Lemma 3.1. Here we choose $\lambda(h) = 0.002\,h^3$ for $k = 1$ and $\lambda(h) = 0.002\,h^{4.2}$ for $k = 2$. We first observe that all convergence plots for our approach are strictly monotone, as opposed to classical FEM. In the L^2-norm, for h sufficiently small, we observe a slightly super-convergent behavior for our approach for both $k = 1$ and $k = 2$. In the H^1-norm, for both orders, our method seems to reach the expected convergence rates for h sufficiently small, and clearly outperforms classical FEM. The choice of a small multiplicative coefficient 0.002 in the function λ is guided by the fact that the norm of $\tilde{g}$ is big for such a contrast. Hence, for coarse meshes for which convergence is far from being reached, one has to give less weight to the second term of the functional J_h in the minimization process, to have a chance to recover a correct approximation of $\tilde{g}$. This weight is useless when h is sufficiently small.

Remark 2 (Stabilization parameter) For coarse meshes, one has to tune the parameter $\lambda(h)$ in a nontrivial, contrast-depending way. This has to do with the regularization we use. If the L^2-orthogonal projection $\Pi_h^k\tilde{g}$ of $\tilde{g}$ onto $F^k(\Gamma_h)$ was known, then we would stabilize the functional J_h with $\mu(h)\|g_h - \Pi_h^k\tilde{g}\|_{0,\Gamma}^2$ instead of $\lambda(h)\|g_h\|_{0,\Gamma}^2$, yielding, in view of (18), an optimal estimate on $\|\tilde{g}_h - \Pi_h^k\tilde{g}\|_{0,\Gamma}$.

Remark 3 It has to be noted that the symmetric cavity test-case can be analyzed by means of T-coercivity. However, it cannot be on general meshes, which is the main outcome of our approach. Furthermore, there are also many cases (even in 2D) that cannot be analyzed, at the continuous level, using the T-coercivity approach. This is for example the case of the cloaking device studied in [10, Section 4]. In that case, the operator associated to the problem, viewed as an operator from $H_0^1(\Omega)$ to $H^{-1}(\Omega)$ is not Fredholm, as its range is not closed. T-coercivity is hence inapplicable. However, it is known that, for compatible right-hand sides $f \in L^2(\Omega)$, the solution exists and is unique in $H_0^1(\Omega)$ [10]. The convergence of our numerical method for compatible loadings as well as numerical experiments for such problems will be investigated in a forthcoming work [1].

References

[1] A. Abdulle and S. Lemaire. An optimization-based method for sign-changing PDEs. In preparation.

[2] C. Bernardi, Y. Maday, and F. Rapetti. *Discrétisations variationnelles de problèmes aux limites elliptiques*, volume 45 of *Mathématiques & Applications*. Springer-Verlag, Berlin, 2004.

[3] A.-S. Bonnet-Ben Dhia, C. Carvalho, and P. Ciarlet Jr. Mesh requirements for the finite element approximation of problems with sign-changing coefficients. Submitted (2016). Preprint hal-01335153.

[4] A.-S. Bonnet-Ben Dhia, L. Chesnel, and P. Ciarlet Jr. T-coercivity for scalar interface problems between dielectrics and metamaterials. *ESAIM: Math. Model. Numer. Anal. (M2AN)*, 46:1363–1387, 2012.

[5] S. C. Brenner. Poincaré–Friedrichs inequalities for piecewise H^1 functions. *SIAM J. Numer. Anal.*, 41(1):306–324, 2003.

[6] L. Chesnel and P. Ciarlet Jr. T-coercivity and continuous Galerkin methods: application to transmission problems with sign-changing coefficients. *Numer. Math.*, 124:1–29, 2013.

[7] P. G. Ciarlet. *The finite element method for elliptic problems*, volume 40 of *Classics in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2002. Reprint of the 1978 original [North-Holland, Amsterdam].

[8] M. D. Gunzburger, M. Heinkenschloss, and H. K. Lee. Solution of elliptic partial differential equations by an optimization-based domain decomposition method. *Appl. Math. Comput.*, 113(2-3):111–139, 2000.

[9] M. D. Gunzburger, J. S. Peterson, and H. K. Lee. An optimization-based domain decomposition method for partial differential equations. *Comput. Math. Appl.*, 37(10):77–93, 1999.

[10] H.-M. Nguyen. Negative index materials and their applications: recent mathematics progress. *Chinese Annals of Mathematics*. To appear.

[11] H.-M. Nguyen. Asymptotic behavior of solutions to the Helmholtz equations with sign-changing coefficients. *Transactions of the American Mathematical Society*, 367:6581–6595, 2015.

[12] H.-M. Nguyen. 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.

[13] S. Nicaise and J. Venel. A posteriori error estimates for a finite element approximation of transmission problems with sign-changing coefficients. *J. Comput. Appl. Math.*, 235:4272–4282, 2011.