An optimization-based method for sign-changing elliptic PDEs

Assyr Abdulle, Simon Lemaire

An optimization-based method for sign-changing elliptic PDEs

Assyr Abdulle (†)[1] and Simon Lemaire*[1,2]

[1]ANMC, Institute of Mathematics, École Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland
[2]Inria, Univ. Lille, CNRS, UMR 8524 – Laboratoire Paul Painlevé, 59000 Lille, France

June 25, 2023

This article is dedicated to the memory of Assyr Abdulle (1971–2021).

Abstract

We study the numerical approximation of sign-shifting problems of elliptic type. We fully analyze and assess the method briefly introduced in [1]. Our method, which is based on domain decomposition and optimization, is proved to be convergent as soon as, for a given loading, the continuous problem admits a unique solution of finite energy. Departing from the T-coercivity approach, which relies on the use of geometrically fitted mesh families, our method works for arbitrary (interface-compliant) meshes and anisotropic coefficients. Moreover, it is shown convergent for a class of problems for which T-coercivity is not applicable. A comprehensive set of test-cases complements our analysis.

1 Introduction

We are interested in this work in the numerical approximation of elliptic interface problems that present a sign shift. Our main motivation is the modeling of the interface between a classical material and a metamaterial.

Optical metamaterials are artificial micro-structured materials exhibiting effective electromagnetic properties that cannot be found in Nature, like an electric permittivity or/and a magnetic permeability with negative real part(s). Optical metamaterials are genuinely dispersive. Among them, the so-called *negative-index metamaterials (NIMs)* are of particular interest: they present over some frequency range a negative refractive index, i.e. simultaneously negative permittivity/permeability (we always refer to the real parts of these coefficients). The existence of such materials has been postulated in 1968 in the seminal work of Veselago [55]. The first effective design of a device exhibiting simultaneously negative permittivity/permeability was realized by Smith et al. in 2000 [53, 52]. NIMs have a tremendous amount of potential applications, among which superlensing [48, 50, 43] or cloaking (either using complementary media [35, 44], or via anomalous localized resonance [40, 10, 42]).

Several models exist in the literature to describe the effective properties of dispersive optical metamaterials. One can cite for instance the Drude–Lorentz class of materials. These

*Corresponding author: simon.lemaire@inria.fr

effective models can be mathematically justified by (high-contrast) homogenization, starting from the corresponding micro-structures. Typically, optical metamaterials are made up of small, highly conductive inclusions, which are periodically arranged within a dielectric matrix. We mention [9, 11, 36] and the references therein for examples of such settings. For non-lossy materials, the modeling of the interface between a classical material and a metamaterial raises new questions concerning the well-posedness and the approximability of the resulting models, owing to the possible (spatial) sign shift of the coefficients. For Maxwell's equations in the time domain, existence and uniqueness hold irrespectively of the problem data. However, the limiting amplitude principle is not always valid. We refer the reader to [18, 19] for an analysis in the case of a plane interface, and to [17] for a numerical study including corners. In the frequency domain, existence and uniqueness may depend on various parameters, including the frequency, the geometry, the coefficients, or the loading [25, 49, 29, 24, 34], which can be interpreted as a signature of the limiting amplitude principle conditional validity.

Among the different mathematical frameworks for studying the well-posedness of sign-shifting elliptic PDEs, two are especially worth discussing in details (we refer the reader to [37, 46] for comprehensive surveys). The first one is the T-coercivity theory, introduced by Bonnet-Ben Dhia et al. [6]. The T-coercivity theory aims at proving the well-posedness (in H^1) of the problem in the Fredholm sense. So far, the focus of T-coercivity has been on isotropic coefficients, and it has been shown that the contrast of the coefficients at the sign-changing interface plays a crucial role in the well-posedness of the model, with a super-critical value of the contrast equal to -1. In 2D, optimal conditions have been derived which provide a bounded closed interval of $(-\infty, 0)$ for the contrast (the so-called *critical interval*, which contains the super-critical value -1) outside which the problem is well-posed in the Fredholm sense. The well-posedness of critical (but non super-critical) situations has been tackled for some given configurations in [7], where Fredholmness is recovered in an augmented functional framework. Another interesting approach to study the well-posedness of sign-changing elliptic PDEs has been proposed by Nguyen [41, 45] in the context of Helmholtz equations. The idea is to introduce some loss in the negative material (i.e. a nonzero imaginary part) and to study the behavior, as the loss tends to zero, of the solution to the well-posed lossy system (limiting absorption principle). By means of reflection operators, it is possible to derive conditions on the coefficients under which the limit problem is well-posed. The advantage of such an approach is twofold: first it is designed for anisotropic coefficients, and second it can deal with configurations for which the corresponding operator is not Fredholm (including super-critical cases). On the other hand, it is currently limited to smooth sign-shifting interfaces.

As far as numerical approximation is concerned, when dealing with sign-shifting problems, one needs to deploy dedicated techniques in order to handle the indefiniteness of the model at hand. Interesting yet sub-optimal first attempts include [8, 47] (cf. also [21, Section 5.1]) and [21, Section 5.2] (based on limiting absorption). The most fruitful approach so far is based on the T-coercivity theory. T-coercivity based approximation [21, 5, 16, 33] takes advantage of the knowledge of the bijective operator T to infer meshing rules, under which conforming finite elements can be proved (optimally) convergent. Evidently, T-coercivity based approximation suffers from the same limitations as T-coercivity does; in particular, it can only apply to configurations for which the problem is well-posed in the classical sense. In addition, it is also bound to the explicit knowledge of the operator T, as well as to the use of geometrically fitted meshes locally to the sign-changing interface. The design of such meshes can become very intricate for interfaces with general shapes. Therefore, there is room for improvement in designing a numerical method that is applicable as soon as the model possesses, for a given

loading, a unique solution of finite energy, and which does not require the use of geometrically constrained meshes. This last criterion is particularly crucial in applications, for instance to simulate the micro-structures of [13], for which sign-shifting cell problems with potentially fairly general interfaces must be solved.

In this work, we fully analyze and validate the method summarily introduced in [1]. This method constitutes an alternative path for the numerical approximation of sign-shifting PDEs. It is based on a decomposition of the domain into signed subdomains (i.e. subdomains in which the coefficient is sign-definite), and on a recasting of the model into a transmission problem. The numerical method then consists in finding the discrete flux at the interface for which some criterion, quantifying the trace jump at the interface between the discrete solutions in both subdomains, is minimal. A key feature of the method is to relax at the discrete level the continuity of the solution at the interface while keeping a control on its jump through the minimization of an augmented functional. This method is proved convergent as soon as the problem admits, for a given loading, a unique solution of finite energy (in H^1). In particular, the problem is not required to be Fredholm. Furthermore, the convergence proof does not rely on any kind of geometrical constraints on the mesh family (the only needed assumption is that the mesh cells do not cut the interface). As standard with domain decomposition, the implementation of our method can benefit from parallel/distributed architectures. As already mentioned in [1], the type of cost functional we consider has first been deployed in [32, 31] in the context of optimization-based domain decomposition for classical elliptic equations. With respect to [32, 31], the novelty in our approach essentially lies in the numerical algorithm, in its analysis, and in its application to sign-changing PDEs. Note that, at the time this manuscript is finalized, another related approach (based on optimal control) has been introduced in [23], which remedies one limitation of our method (cf. Remark 6.9). Finally, note that our approach shares some common goals with [14, 15], in that it aims at approximating problems which are not necessarily well-posed in the classical sense.

The article is organized as follows. In Section 2 we introduce some useful functional analysis tools. In Section 3 we introduce the problem under study. In Section 4 we biefly motivate our approach, in particular we review the limitations of T-coercivity as an approximation method. In Section 5 we recast the continuous problem as an interface problem, and we provide a characterization of its solution on which we base our numerical algorithm. In Section 6 we introduce the numerical method, and we prove its convergence. In Section 7 we provide a comprehensive set of numerical experiments and demonstrate the efficiency of our approach. We also detail a possible algebraic realization of the method. Finally, in Appendix A we provide some basic background on Fredholm theory, whereas in Appendix B we collect (sharp) error estimates for the finite element solutions to nonhomogeneous mixed and purely Neumann variable diffusion problems, which are instrumental to finely tune our method.

2 Functional analysis tools

Let $\mathcal{D}$ be a *domain* in $\mathbb{R}^d$, $d \in \{2,3\}$, that is a bounded and connected Lipschitz open set of $\mathbb{R}^d$. We let $\Upsilon := \partial\mathcal{D}$ denote the boundary of $\mathcal{D}$. Since $\mathcal{D}$ is Lipschitz, a unit normal vector field $\boldsymbol{n}$ can be defined almost everywhere on Υ, which is assumed to point outward $\mathcal{D}$. The set Υ is further partitioned into two disjoint, relatively open Lipschitz subsets Υ_{t} and Υ_{f} with $\Upsilon_{\mathrm{f}} \neq \varnothing$ such that $\Upsilon = \overline{\Upsilon_{\mathrm{t}}} \cup \overline{\Upsilon_{\mathrm{f}}}$.

For $q \in \{1, d\}$, we classically let $L^2(\mathcal{D};\mathbb{R}^q)$ be the Hilbert space of distributions $\boldsymbol{v} :=$

$(v_1,\dots,v_q):\mathcal{D}\to\mathbb{R}^q$ (whenever $q=1$, we simply write v) such that $\int_{\mathcal{D}}|\boldsymbol{v}(\boldsymbol{x})|^2\mathrm{d}\boldsymbol{x}<\infty$. Irrespectively of q, the standard inner product and norm in $L^2(\mathcal{D};\mathbb{R}^q)$ are denoted by $(\boldsymbol{v},\boldsymbol{w})_{\mathcal{D}}:=\int_{\mathcal{D}}\boldsymbol{v}(\boldsymbol{x})\cdot\boldsymbol{w}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$ and $\|\boldsymbol{v}\|_{0,\mathcal{D}}:=\sqrt{(\boldsymbol{v},\boldsymbol{v})_{\mathcal{D}}}$. For $m\in\mathbb{N}^\star$, $\boldsymbol{\alpha}:=(\alpha_1,\dots,\alpha_d)\in\mathbb{N}^d$ a multi-index, and $\partial^{\boldsymbol{\alpha}}\boldsymbol{v}:=(\partial_1^{\alpha_1}..\partial_d^{\alpha_d}v_1,\dots,\partial_1^{\alpha_1}..\partial_d^{\alpha_d}v_q):\mathcal{D}\to\mathbb{R}^q$, we classically let $H^m(\mathcal{D};\mathbb{R}^q)$ be the Hilbert space of distributions $\boldsymbol{v}\in L^2(\mathcal{D};\mathbb{R}^q)$ such that $\|\partial^{\boldsymbol{\alpha}}\boldsymbol{v}\|_{0,\mathcal{D}}<\infty$ for all $\boldsymbol{\alpha}\in\mathbb{N}^d$ with $(1\leqslant)\,\alpha_1+\ldots+\alpha_d\leqslant m$. We equip $H^m(\mathcal{D};\mathbb{R}^q)$ with the following norm:

$$\|\boldsymbol{v}\|_{m,\mathcal{D}}^2:=\|\boldsymbol{v}\|_{m-1,\mathcal{D}}^2+|\boldsymbol{v}|_{m,\mathcal{D}}^2,\qquad |\boldsymbol{v}|_{m,\mathcal{D}}^2:=\sum_{\alpha_1+\ldots+\alpha_d=m}\|\partial^{\boldsymbol{\alpha}}\boldsymbol{v}\|_{0,\mathcal{D}}^2,$$

with the convention that $H^0\equiv L^2$. Next, for $\sigma\in(0,1)$, letting for $\boldsymbol{w}:\mathcal{D}\to\mathbb{R}^q$,

$$|\boldsymbol{w}|_{\sigma,\mathcal{D}}^2:=\int_{\mathcal{D}}\int_{\mathcal{D}}\frac{|\boldsymbol{w}(\boldsymbol{x})-\boldsymbol{w}(\boldsymbol{y})|^2}{|\boldsymbol{x}-\boldsymbol{y}|^{2\sigma+d}}\mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{y},$$

we classically let $H^\sigma(\mathcal{D};\mathbb{R}^q)$ be the fractional Hilbert space of distributions $\boldsymbol{v}\in L^2(\mathcal{D};\mathbb{R}^q)$ such that $|\boldsymbol{v}|_{\sigma,\mathcal{D}}<\infty$. In the same vein, for $s=m+\sigma$ with $m:=\lfloor s\rfloor\in\mathbb{N}$ and $\sigma:=s-m\in(0,1)$, we denote by $H^s(\mathcal{D};\mathbb{R}^q)$ the fractional Hilbert space of distributions $\boldsymbol{v}\in H^m(\mathcal{D};\mathbb{R}^q)$ such that $|\partial^{\boldsymbol{\alpha}}\boldsymbol{v}|_{\sigma,\mathcal{D}}<\infty$ for all $\boldsymbol{\alpha}\in\mathbb{N}^d$ with $\alpha_1+\ldots+\alpha_d=m$. Remark that this definition coincides with the above definition of $H^\sigma(\mathcal{D};\mathbb{R}^q)$ whenever $m=0$. We equip $H^s(\mathcal{D};\mathbb{R}^q)$ with the following Sobolev–Slobodeckij norm:

$$\|\boldsymbol{v}\|_{s,\mathcal{D}}^2:=\|\boldsymbol{v}\|_{m,\mathcal{D}}^2+|\boldsymbol{v}|_{s,\mathcal{D}}^2,\qquad |\boldsymbol{v}|_{s,\mathcal{D}}^2:=\sum_{\alpha_1+\ldots+\alpha_d=m}|\partial^{\boldsymbol{\alpha}}\boldsymbol{v}|_{\sigma,\mathcal{D}}^2.$$

Henceforth, for convenience, we simply write $L^2(\mathcal{D})$ or $H^s(\mathcal{D})$ in place of $L^2(\mathcal{D};\mathbb{R})$ or $H^s(\mathcal{D};\mathbb{R})$, and $\boldsymbol{L}^2(\mathcal{D})$ or $\boldsymbol{H}^s(\mathcal{D})$ in place of $L^2(\mathcal{D};\mathbb{R}^d)$ or $H^s(\mathcal{D};\mathbb{R}^d)$. As standard, we let $H_0^1(\mathcal{D})$ be the Hilbert space, closed subset of $H^1(\mathcal{D})$, obtained as the closure for the $\|\cdot\|_{1,\mathcal{D}}$-norm of $C_0^\infty(\mathcal{D})$. We further let $H^{-1}(\mathcal{D})$ denote the topological dual of $H_0^1(\mathcal{D})$, with duality pairing $\langle\cdot,\cdot\rangle_{\mathcal{D}}$. Endowed with the norm

$$\|t\|_{-1,\mathcal{D}}:=\sup_{v\in H_0^1(\mathcal{D})\setminus\{0\}}\frac{\langle t,v\rangle_{\mathcal{D}}}{\|v\|_{1,\mathcal{D}}},$$

$H^{-1}(\mathcal{D})$ is a Banach space.

Let us now turn to the definition of trace spaces, first on the whole domain boundary. We classically let $L^2(\Upsilon;\mathbb{R}^q)$ be the Hilbert space of distributions $\boldsymbol{\varphi}:=(\varphi_1,\dots,\varphi_q):\Upsilon\to\mathbb{R}^q$ (whenever $q=1$, we simply write φ) such that $\int_\Upsilon|\boldsymbol{\varphi}(\boldsymbol{x})|^2\mathrm{d}\boldsymbol{\sigma}(\boldsymbol{x})<\infty$. Irrespectively of q, the standard inner product and norm in $L^2(\Upsilon;\mathbb{R}^q)$ are denoted by $(\boldsymbol{\varphi},\boldsymbol{\psi})_\Upsilon:=\int_\Upsilon\boldsymbol{\varphi}(\boldsymbol{x})\cdot\boldsymbol{\psi}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{\sigma}(\boldsymbol{x})$ and $\|\boldsymbol{\varphi}\|_{0,\Upsilon}:=\sqrt{(\boldsymbol{\varphi},\boldsymbol{\varphi})_\Upsilon}$. For $\sigma\in(0,1)$, letting

$$|\boldsymbol{\varphi}|_{\sigma,\Upsilon}^2:=\int_\Upsilon\int_\Upsilon\frac{|\boldsymbol{\varphi}(\boldsymbol{x})-\boldsymbol{\varphi}(\boldsymbol{y})|^2}{|\boldsymbol{x}-\boldsymbol{y}|^{2\sigma+d-1}}\mathrm{d}\boldsymbol{\sigma}(\boldsymbol{x})\mathrm{d}\boldsymbol{\sigma}(\boldsymbol{y}),$$

we classically denote by $H^\sigma(\Upsilon;\mathbb{R}^q)$ the fractional Hilbert space of distributions $\boldsymbol{\varphi}\in L^2(\Upsilon;\mathbb{R}^q)$ such that $|\boldsymbol{\varphi}|_{\sigma,\Upsilon}<\infty$. We equip $H^\sigma(\Upsilon;\mathbb{R}^q)$ with the following Sobolev–Slobodeckij norm:

$$\|\boldsymbol{\varphi}\|_{\sigma,\Upsilon}^2:=\|\boldsymbol{\varphi}\|_{0,\Upsilon}^2+|\boldsymbol{\varphi}|_{\sigma,\Upsilon}^2.$$

Next, for $s\in(\frac{1}{2},1]$, we let $\boldsymbol{\gamma}:H^s(\mathcal{D};\mathbb{R}^q)\to H^{s-\frac{1}{2}}(\Upsilon;\mathbb{R}^q)$ (whenever $q=1$, we write γ) denote the (linear, bounded) trace operator. By definition, $\boldsymbol{\gamma}(\boldsymbol{v})$ coincides with $\boldsymbol{v}_{|\Upsilon}$ when $\boldsymbol{v}$ is sufficiently regular. There is $c_\gamma>0$ such that, for all $\boldsymbol{v}\in H^s(\mathcal{D};\mathbb{R}^q)$, there holds

$$\|\boldsymbol{\gamma}(\boldsymbol{v})\|_{s-\frac{1}{2},\Upsilon}\leqslant c_\gamma\|\boldsymbol{v}\|_{s,\mathcal{D}}.\tag{1}$$

The operator $\boldsymbol{\gamma}$ is also surjective, with bounded right inverse (cf. e.g. [30, Theorem 1.5.1.2]). Henceforth, for convenience, we simply write $H^\sigma(\Upsilon)$ in place of $H^\sigma(\Upsilon;\mathbb{R})$, and $\boldsymbol{H}^\sigma(\Upsilon)$ in place of $H^\sigma(\Upsilon;\mathbb{R}^d)$. We classically let $H^{-\frac12}(\Upsilon)$ denote the topological dual of $H^{\frac12}(\Upsilon)$, with duality pairing $\langle\cdot,\cdot\rangle_\Upsilon$. Endowed with the norm

$$\|\theta\|_{-\frac12,\Upsilon} := \sup_{\varphi\in H^{\frac12}(\Upsilon)\setminus\{0\}} \frac{\langle\theta,\varphi\rangle_\Upsilon}{\|\varphi\|_{\frac12,\Upsilon}},$$

$H^{-\frac12}(\Upsilon)$ is a Banach space. Let

$$\boldsymbol{H}(\mathrm{div};\mathcal{D}) := \big\{\boldsymbol{\theta}\in\boldsymbol{L}^2(\mathcal{D}) \mid \mathrm{div}\,\boldsymbol{\theta}\in L^2(\mathcal{D})\big\}.$$

For any $\boldsymbol{\theta}\in\boldsymbol{H}(\mathrm{div};\mathcal{D})$, by surjectivity of the trace operator γ, one can give a sense to the normal trace of $\boldsymbol{\theta}$ on Υ (denoted $\gamma_{\boldsymbol{n}}(\boldsymbol{\theta})$) in $H^{-\frac12}(\Upsilon)$ via the following divergence formula: for all $v\in H^1(\mathcal{D})$,

$$\langle\gamma_{\boldsymbol{n}}(\boldsymbol{\theta}),\gamma(v)\rangle_\Upsilon := (\boldsymbol{\theta},\boldsymbol{\nabla}v)_{\mathcal{D}} + (\mathrm{div}\,\boldsymbol{\theta},v)_{\mathcal{D}}.$$

By definition, $\gamma_{\boldsymbol{n}}(\boldsymbol{\theta})$ coincides with $\boldsymbol{\theta}_{|\Upsilon}\cdot\boldsymbol{n}$ when $\boldsymbol{\theta}$ is sufficiently regular.

Let us finally introduce the so-called Lions–Magenes trace space. We assume that $\Upsilon_{\mathrm{t}}\neq\varnothing$, so that both Υ_{t} and Υ_{f} are nonempty. The Lions–Magenes space on Υ_{f}, usually denoted $H^{1/2}_{00}(\Upsilon_{\mathrm{f}})$, is formally the space of distributions in $H^{\frac12}(\Upsilon_{\mathrm{f}})$ which can be extended by zero to distributions in $H^{\frac12}(\Upsilon)$. More rigorously, letting

$$|\varphi|^2_{\frac12,\Upsilon_{\mathrm{f}},00} := \int_{\Upsilon_{\mathrm{f}}}\frac{(\varphi(\boldsymbol{x}))^2}{\rho(\boldsymbol{x})}\mathrm{d}\boldsymbol{\sigma}(\boldsymbol{x}),$$

where $\rho(\boldsymbol{x}) := \min_{\boldsymbol{y}\in\partial\Upsilon_{\mathrm{f}}}|\boldsymbol{x}-\boldsymbol{y}|$ is the distance to $\partial\Upsilon_{\mathrm{f}}$, we define $H^{1/2}_{00}(\Upsilon_{\mathrm{f}})$ as the fractional Hilbert space of distributions $\varphi\in H^{\frac12}(\Upsilon_{\mathrm{f}})$ such that $|\varphi|_{\frac12,\Upsilon_{\mathrm{f}},00}<\infty$. We equip $H^{1/2}_{00}(\Upsilon_{\mathrm{f}})$ with the following Sobolev–Slobodeckij norm:

$$\|\varphi\|^2_{\frac12,\Upsilon_{\mathrm{f}},00} := \|\varphi\|^2_{\frac12,\Upsilon_{\mathrm{f}}} + |\varphi|^2_{\frac12,\Upsilon_{\mathrm{f}},00}.$$

There holds

$$H^{1/2}_{00}(\Upsilon_{\mathrm{f}}) = \big\{\varphi\in H^{\frac12}(\Upsilon_{\mathrm{f}}) \mid \exists\hat\varphi\in H^{\frac12}(\Upsilon) \text{ s.t. } \hat\varphi_{|\Upsilon_{\mathrm{f}}}=\varphi,\ \hat\varphi_{|\Upsilon_{\mathrm{t}}}=0\big\}\big(\subsetneq H^{\frac12}(\Upsilon_{\mathrm{f}})\big),$$

in such a way that, letting

$$H^1_{0\setminus\Upsilon_{\mathrm{f}}}(\mathcal{D}) := \big\{v\in H^1(\mathcal{D}) \mid \gamma(v)_{|\Upsilon_{\mathrm{t}}}=0\big\}, \tag{2}$$

we have $\gamma\big(H^1_{0\setminus\Upsilon_{\mathrm{f}}}(\mathcal{D})\big)_{|\Upsilon_{\mathrm{f}}} = H^{1/2}_{00}(\Upsilon_{\mathrm{f}})$ by surjectivity of the trace operator γ. We let $H^{-\frac12}(\Upsilon_{\mathrm{f}})$ denote the topological dual of $H^{1/2}_{00}(\Upsilon_{\mathrm{f}})$, with duality pairing $\langle\cdot,\cdot\rangle_{\Upsilon_{\mathrm{f}}}$. Endowed with the norm

$$\|\theta\|_{-\frac12,\Upsilon_{\mathrm{f}}} := \sup_{\varphi\in H^{1/2}_{00}(\Upsilon_{\mathrm{f}})\setminus\{0\}}\frac{\langle\theta,\varphi\rangle_{\Upsilon_{\mathrm{f}}}}{\|\varphi\|_{\frac12,\Upsilon_{\mathrm{f}},00}},$$

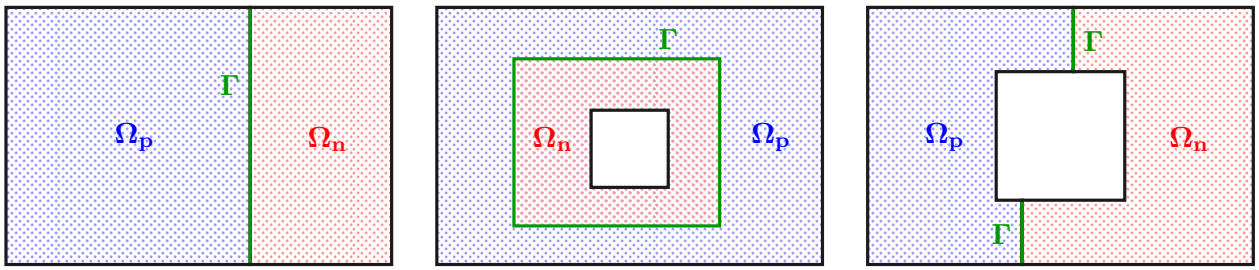
$H^{-\frac12}(\Upsilon_{\mathrm{f}})$ is a Banach space. For any $\boldsymbol{\theta}\in\boldsymbol{H}(\mathrm{div};\mathcal{D})$, it is possible to give a sense to the normal trace of $\boldsymbol{\theta}$ on Υ_{f} (denoted $\gamma_{\boldsymbol{n},\mathrm{f}}(\boldsymbol{\theta})$) in $H^{-\frac12}(\Upsilon_{\mathrm{f}})$ via the following divergence formula: for all $v\in H^1_{0\setminus\Upsilon_{\mathrm{f}}}(\mathcal{D})$,

$$\langle\gamma_{\boldsymbol{n},\mathrm{f}}(\boldsymbol{\theta}),\gamma(v)\rangle_{\Upsilon_{\mathrm{f}}} := (\boldsymbol{\theta},\boldsymbol{\nabla}v)_{\mathcal{D}} + (\mathrm{div}\,\boldsymbol{\theta},v)_{\mathcal{D}}.$$

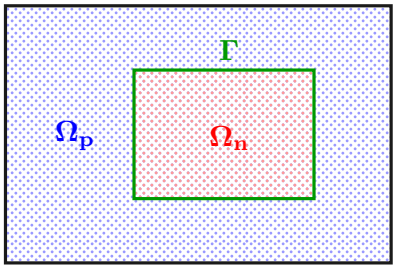
Above, we abuse the notation by writing $\gamma(v)$ in place of $\gamma(v)_{|\Upsilon_{\mathrm{f}}}$. By definition, $\gamma_{\boldsymbol{n},\mathrm{f}}(\boldsymbol{\theta})$ coincides with $\boldsymbol{\theta}_{|\Upsilon_{\mathrm{f}}}\cdot\boldsymbol{n}$ when $\boldsymbol{\theta}$ is sufficiently regular.

3 Setting of the problem

Let Ω be a domain in $\mathbb{R}^d$ (i.e. a bounded and connected Lipschitz open set of $\mathbb{R}^d$), $d \in \{2,3\}$. We assume that Ω is partitioned into two disjoint (nonempty) Lipschitz open subsets Ω_p and Ω_n so that $\overline{\Omega} = \overline{\Omega_\mathrm{p}} \cup \overline{\Omega_\mathrm{n}}$. As it will become clear in what follows, the subscripts 'p' and 'n' refer, respectively, to the positive and negative subdomains. The two subdomains Ω_p and Ω_n are assumed to be connected. We further suppose that Ω_p is such that $\partial\Omega_\mathrm{p} \cap \partial\Omega$ has nonzero $(d-1)$-dimensional measure. We let $\Gamma := \mathrm{int}(\partial\Omega_\mathrm{p} \cap \partial\Omega_\mathrm{n})$ denote the (open) interface between Ω_p and Ω_n, which is a Lipschitz $(d-1)$-dimensional manifold. Since Γ is Lipschitz, one can define a (unit) normal vector field almost everywhere on Γ. There holds $|\partial\Omega \cap \Gamma|_{d-1} = 0$. On Figure 1 are depicted various admissible configurations Ω in 2D. The meaning of the classification `2M` (for mixed-mixed coupling) and `MN` (for mixed-Neumann coupling) will be made completely precise in Section 5.3. The top configurations 1a are such that both $\partial\Omega_\mathrm{p} \cap \partial\Omega$ and $\partial\Omega_\mathrm{n} \cap \partial\Omega$ have nonzero lineic measures. The bottom configuration 1b is, at the opposite, such that $\partial\Omega_\mathrm{n} \cap \partial\Omega = \varnothing$. For the left and center configurations 1a, and for the configuration 1b, the interface Γ is connected, whereas it is not the case for the right configuration 1a. The left configuration 1a is referred to in the literature as the (symmetric or nonsymmetric, depending on the position of Γ) cavity; see e.g. [21, Section 3.3]. In the following, we will refer to the configuration 1b as the inclusion case.



(a) Configurations `2M`



(b) Configuration `MN`

Figure 1: Examples of configurations Ω in 2D

Let $\boldsymbol{\sigma} : \Omega \to \mathbb{R}^{d\times d}$ be a symmetric tensor field such that

$$0 < \sigma_\flat |\boldsymbol{\xi}|^2 \leqslant \boldsymbol{\sigma}(\boldsymbol{x})\boldsymbol{\xi}\cdot\boldsymbol{\xi} \leqslant \sigma_\sharp |\boldsymbol{\xi}|^2 < \infty \qquad \text{for a.e. } \boldsymbol{x} \in \Omega \text{ and all } \boldsymbol{\xi} \in \mathbb{R}^d\backslash\{\boldsymbol{0}\},$$

and let $\rho := \sigma_\sharp/\sigma_\flat \geqslant 1$ denote its heterogeneity/anisotropy ratio in Ω. We further assume that $\boldsymbol{\sigma}_\alpha := \boldsymbol{\sigma}_{|\Omega_\alpha} \in \mathbb{W}^{1,\infty}(\Omega_\alpha)$ (with obvious notation) for $\alpha \in \{\mathrm{p},\mathrm{n}\}$. When $\boldsymbol{\sigma}$ is isotropic, i.e. when there is $\sigma : \Omega \to \mathbb{R}$ satisfying $0 < \sigma_\flat \leqslant \sigma \leqslant \sigma_\sharp < \infty$ such that $\boldsymbol{\sigma} = \sigma\mathbb{1}_d$ where $\mathbb{1}_d$ is the $d \times d$

identity tensor, we let

$$\nu := -\frac{\sigma_{\mathrm{n}|\Gamma}}{\sigma_{\mathrm{p}|\Gamma}} \tag{3}$$

denote the coefficient *contrast* at the interface. Let $s : \Omega \to \{-1, +1\}$ be the sign function such that $s_{\mathrm{p}} := s_{|\Omega_{\mathrm{p}}} = +1$ and $s_{\mathrm{n}} := s_{|\Omega_{\mathrm{n}}} = -1$. For $f \in H^{-1}(\Omega)$, we study the following (anisotropic) sign-shifting problem: find $\tilde{u} \in H^1(\Omega)$ such that

$$\begin{cases} -\mathrm{div}\,(s\,\varpi\boldsymbol{\nabla}\tilde{u}) = f & \text{in } \Omega, \\ \tilde{u} = 0 & \text{on } \partial\Omega. \end{cases} \tag{4}$$

For further use, we also let $\tilde{u}_\alpha := \tilde{u}_{|\Omega_\alpha}$ for $\alpha \in \{\mathrm{p}, \mathrm{n}\}$. Let us insist on the fact that ϖ is a real-valued coefficient. We are thus looking for real-valued solutions to Problem (4). Note that we could also consider, up to slight adaptations of the method described in Section 6, more general boundary conditions for Problem (4), or/and more complex geometries for the subdomains Ω_{p} and Ω_{n}. We refer to Remark 6.10 for further insight on this question.

The weak form of Problem (4) writes: find $\tilde{u} \in H^1_0(\Omega)$ such that

$$a(\tilde{u}, v) := (s\,\varpi\boldsymbol{\nabla}\tilde{u}, \boldsymbol{\nabla}v)_\Omega = \langle f, v\rangle_\Omega \qquad \forall v \in H^1_0(\Omega). \tag{5}$$

4 A look into T-coercivity

4.1 Background on T-coercivity theory

For the reader not familiar with Fredholm theory on which we are going to rely below, we refer to Appendix A, where some fundamental definitions and results are recalled.

Let U be a real-valued Hilbert space, with norm $\|\cdot\|$. Let $U^\star$ denote the topological dual of U, with duality pairing $\langle\cdot,\cdot\rangle$. Let $b : U \times U \to \mathbb{R}$ be a continuous bilinear form, which is further assumed symmetric, i.e. $b(v,u) = b(u,v)$ for all $u, v \in U$. Under these assumptions, there exists a self-adjoint operator $\mathtt{B} \in \mathcal{L}(U, U^\star)$ such that, for all $u, v \in U$, $\langle \mathtt{B}(u), v\rangle = b(u,v)$. For a given $f \in U^\star$, we are interested in the following problem: find $\tilde{u} \in U$ such that

$$\mathtt{B}(\tilde{u}) = f \text{ in } U^\star. \tag{6}$$

The target application we have in mind is Problem (5), for which (i) $U := H^1_0(\Omega)$ with norm $\|\cdot\| := |\cdot|_{1,\Omega}$ and duality pairing $\langle\cdot,\cdot\rangle := \langle\cdot,\cdot\rangle_\Omega$, and (ii) $\mathtt{B} := \mathtt{A}$ where the self-adjoint operator $\mathtt{A} \in \mathcal{L}(H^1_0(\Omega), H^{-1}(\Omega))$ is such that $\langle \mathtt{A}(u), v\rangle_\Omega := a(u,v)$ for all $u, v \in H^1_0(\Omega)$ with (symmetric) bilinear form a defined in (5).

The T-coercivity theory [6] is a variational Hilbertian theory which aims at proving the well-posedness of Problem (6) in the Fredholm sense. Let us thus define this notion.

Definition 4.1 (Well-posedness in the Fredholm sense)**.** *Problem* (6) *is said to be well-posed in the* Fredholm sense *if the operator* $\mathtt{B}$ *is Fredholm of index* 0*.*

In the particular case of Problem (6), for which $\mathtt{B}$ is a self-adjoint operator, Proposition A.4 ensures that, as soon as $\mathtt{B}$ is Fredholm, its index is necessarily equal to zero. Proposition A.4 also provides a detailed characterization of the structure of the solutions to Problem (6). For an equivalent characterization in the non-necessarily self-adjoint case, we refer to [39, Theorem 2.27]. A subcase of well-posedness in the Fredholm sense is the well-posedness in the

Hadamard (or classical) sense. Problem (6) is well-posed in the Hadamard sense when it is well-posed in the Fredholm sense and the operator $\mathtt{B}$ is injective, i.e. when $\mathtt{B}$ is an isomorphism.

Let us now give the definition of $\mathtt{T}$-coercivity; cf. e.g. [21, Definition 3].

Definition 4.2 ($\mathtt{T}$-coercivity)**.** *The bilinear form b is said* $\mathtt{T}$-coercive *if there exists $\mathcal{T} \in \mathcal{L}(U)$ bijective so that there is $c > 0$ such that*

$$b(u, \mathcal{T}(u)) \geqslant c\|u\|^2 \qquad \forall u \in U. \tag{7}$$

In other words, the bilinear form b is $\mathtt{T}$-coercive as soon as the (continuous) bilinear form $b(\cdot, \mathtt{T}(\cdot))$ is coercive. The link between $\mathtt{T}$-coercivity and well-posedness is made explicit in the following proposition; see e.g. [21, Theorem 1].

Proposition 4.3. *Problem (6) is well-posed in the Hadamard sense if and only if the form b is $\mathcal{T}$-coercive.*

$\mathtt{T}$-coercivity is hence a necessary (and sufficient) condition for well-posedness in the classical sense. Note that $\mathtt{T}$-coercivity is, however, less general than the Banach–Nečas–Babuška (inf-sup) theory, as it is restricted to the Hilbertian setting (cf. [28, Remark 25.14]).

In practice, proving $\mathtt{T}$-coercivity may be difficult. This is for instance the case for Problem (5) when considering a general, non-smooth interface between the positive and negative subdomains. In this situation, what one can usually prove is a weaker result, namely weak $\mathtt{T}$-coercivity; see [5, Definition 2].

Definition 4.4 (Weak $\mathtt{T}$-coercivity)**.** *The bilinear form b is said* weakly $\mathtt{T}$-coercive *if there exist $\mathcal{T} \in \mathcal{L}(U)$ bijective and $\mathcal{C} \in \mathcal{L}(U)$ compact so that there are $c_1 > 0$ and $c_2 \in \mathbb{R}$ such that*

$$b(u, \mathcal{T}(u)) \geqslant c_1\|u\|^2 - c_2\|\mathcal{C}(u)\|^2 \qquad \forall u \in U. \tag{8}$$

The bilinear form b is hence weakly $\mathtt{T}$-coercive as soon as the (continuous) bilinear form $b(\cdot, \mathtt{T}(\cdot))$ fulfills a Gårding's inequality [51]. When $c_2 \leqslant 0$, one recovers (plain) $\mathtt{T}$-coercivity for the form b. In the present symmetric case, the link between weak $\mathtt{T}$-coercivity and well-posedness is given in the following proposition; cf. [5, Lemma 1].

Proposition 4.5. *Problem (6) is well-posed in the Fredholm sense if and only if the form b is weakly $\mathcal{T}$-coercive.*

For a symmetric bilinear form, weak $\mathtt{T}$-coercivity is thus a necessary (and sufficient) condition for well-posedness in the Fredholm sense.

4.2 An intrinsic limitation for sign-shifting problems

Let us consider Problem (5) for the 2D nonsymmetric cavity setting analyzed in [21, Section 3.3]. Let $\Omega := (-\underline{\zeta}, \overline{\zeta}) \times (0, 1)$ for $\underline{\zeta}, \overline{\zeta} > 0$ such that $\underline{\zeta} \neq \overline{\zeta}$, and let $\Omega_{\mathrm{p}} := (-\underline{\zeta}, 0) \times (0, 1)$ and $\Omega_{\mathrm{n}} := (0, \overline{\zeta}) \times (0, 1)$, in such a way that $\Gamma = \{0\} \times (0, 1)$; cf. the left panel of Figure 1a. The coefficient $\mathfrak{o}$ is chosen isotropic and homogeneous in Ω, i.e. $\mathfrak{o} := \sigma \mathbb{1}_2$ with σ any positive real number. This corresponds to the so-called super-critical case of a (constant) contrast at the interface of $\nu = -1$. Then, it can be proved that the self-adjoint operator $\mathtt{A}$ is injective but not surjective. More precisely, the range of $\mathtt{A}$ is not closed in $H^{-1}(\Omega)$ (cf. Proposition A.5). As a consequence, either $f \in \mathrm{Im}\mathtt{A}$ and Problem (5) admits a unique solution (of finite energy, i.e. in $H^1_0(\Omega)$), or $f \in H^{-1}(\Omega) \backslash \mathrm{Im}\mathtt{A}$ and Problem (5) does not have a solution.

It is clear that the latter self-adjoint operator $\mathtt{A}$ is not Fredholm. If it was, its index would be zero and injectivity would necessarily imply surjectivity. Since, for self-adjoint operators, weak $\mathtt{T}$-coercivity and Fredholmness (of index 0) are equivalent (cf. Proposition 4.5), we conclude that we cannot find $\mathtt{T} \in \mathcal{L}(U)$ such that a is weakly $\mathtt{T}$-coercive in this case. Consequently, there exist settings, for which the problem admits a unique solution of finite energy, but only for admissible loadings, which are not covered by $\mathtt{T}$-coercivity theory. Another example of such settings, this time with disconnected subdomain Ω_{p}, is given by the cloaking device of [44].

Remark 4.6. *We have focused so far on settings for which the (self-adjoint) operator is injective, but not surjective. Let us point out that there exist other non-Fredholm configurations, and thus other settings not covered by the $\mathtt{T}$-coercivity theory. For example, consider again Problem (5) for the 2D cavity setting with contrast -1 of [21, Section 3.3], but this time with $\underline{\zeta} = \overline{\zeta}$ (symmetric cavity). In this case, the operator $\mathtt{A}$ is such that* $\dim(\mathrm{Ker}\mathtt{A}) = \infty$*, this is hence a non-Fredholm configuration. Note that, for some configurations, it is possible to adapt the functional framework in order to recover Fredholmness of the problem; this is the approach pursued in [7] for critical (but non super-critical) contrasts.*

This intrinsic limitation of $\mathtt{T}$-coercivity for sign-shifting problems has obvious repercussions on the scope of application of (conforming) $\mathtt{T}$-coercivity based approximation for Problem (5).

4.3 $\mathtt{T}$-coercivity based approximation

We make the assumption that Problem (6) is well-posed in the Fredholm sense, and that it has a unique solution. Hence, Problem (6) is well-posed in the Hadamard sense and, according to Proposition 4.3, there exists $\mathtt{T} \in \mathcal{L}(U)$ bijective such that the form b is $\mathtt{T}$-coercive (with constant $c > 0$).

Let $(U_h)_{h>0}$ be a countable family of finite-dimensional vector spaces satisfying $U_h \subset U$ for all $h > 0$ in the family. The dimension of the discrete space U_h is meant to increase as h tends to zero. In practice, U_h is a space of piecewise polynomial functions on a partition $\mathcal{T}_h$ (of size h) of the domain. Let us define the notion of $\mathtt{T}$-conformity.

Definition 4.7 ($\mathtt{T}$-conformity)**.** *The family of discrete spaces $(U_h)_{h>0}$ is said* $\mathtt{T}$-conforming *if it is stable by $\mathtt{T}$, i.e. if $\mathtt{T}(U_h) \subseteq U_h$ for all $h > 0$ in the family.*

We consider the following conforming approximation of Problem (6): find $\tilde{u}_h \in U_h$ such that

$$b(\tilde{u}_h, v_h) = \langle f, v_h \rangle \qquad \forall v_h \in U_h. \tag{9}$$

The following result is adapted from [21, Corollary 1].

Proposition 4.8. *Assume that $(U_h)_{h>0}$ is $\mathtt{T}$-conforming. Then, for all $h > 0$ in the family, Problem (9) admits a unique solution $\tilde{u}_h \in U_h$, and the following estimate holds true:*

$$\|\tilde{u} - \tilde{u}_h\| \leqslant \frac{|||b||| \, |||\mathtt{T}|||}{c} \inf_{v_h \in U_h} \|\tilde{u} - v_h\|. \tag{10}$$

In the case of Problem (5), the operator $\mathtt{T}$ is derived from elementary geometrical transforms (symmetries and rotations) with respect to the sign-shifting interface. These transforms do preserve polynomials. However, since functions in U_h are defined piecewise on the partition $\mathcal{T}_h$, one has to make sure the overall transform maps a cell in Ω_{p} to another cell in Ω_{n}, or reciprocally. Enforcing $\mathtt{T}$-conformity thus boils down to the design of geometrically fitted meshes.

Their practical construction requires the operator $\mathtt{T}$ to be known explicitly. As already pointed out, for a general, non-smooth interface, proving $\mathtt{T}$-coercivity is usually difficult. What is often feasible, however, is to prove weak $\mathtt{T}'$-coercivity, for some (other) bijective operator $\mathtt{T}'$ built as a superposition of localized elementary geometrical transforms. The relevant notion of conformity then becomes $\mathtt{T}'$-conformity, and is a local one. In other words, the mesh constraints need only be imposed in this case in a neighborhood of the interface (see [5, Definition 3]). In 2D, such weak operators $\mathtt{T}'$ can be built for general polygonal interfaces; cf. [5, Theorem 1]. In 3D, only partial results exist; in particular, the case of general polyhedral interfaces is still open. When such a weak operator $\mathtt{T}'$ is available, a result like Proposition 4.8 is valid under a smallness assumption on h; cf. [5, Theorem 2] (in turn based on [21, Proposition 3]).

Proposition 4.9. *Assume that the form b is weakly $\mathtt{T}'$-coercive for some bijective operator $\mathtt{T}' \in \mathcal{L}(U)$. Assume that $(U_h)_{h>0}$ is $\mathtt{T}'$-conforming. Then, for all $h > 0$ small enough in the family, Problem (9) admits a unique solution $\tilde{u}_h \in U_h$, and the following estimate holds true for some $\xi > 0$:*

$$\|\tilde{u} - \tilde{u}_h\| \leqslant \xi \inf_{v_h \in U_h} \|\tilde{u} - v_h\|. \tag{11}$$

In practice, for sign-shifting problems, the dicrete space U_h is usually not stable by the operator $\mathtt{T}'$. The problem is not of a geometrical nature, but comes from the use of cut-off functions to localize the different transforms in $\mathtt{T}'$. As a by-product, functions in $\mathtt{T}'(U_h)$ are usually non-polynomial on each cell of $\mathcal{T}_h$. One has to introduce a new, uniformly (in h) bounded family of operators $(\mathtt{T}'_h)_{h>0}$ such that $\mathtt{T}'_h(U_h) \subseteq U_h$ for all $h > 0$ in the family. This family is constructed so that, for all $h > 0$ small enough in the family, and for all $u_h \in U_h$, $\|(\mathtt{T}' - \mathtt{T}'_h)(u_h)\| \leqslant \vartheta\, h \|u_h\|$ for some $\vartheta > 0$; cf. [5, Lemma 3]. With such an operator at hand, a result equivalent to that of Proposition 4.9 can be proved.

4.4 Towards an alternative approach

For sign-shifting problems of the form (5), $\mathtt{T}$-coercivity based approximation suffers from three important shortcomings:

- non-Fredholm situations are not covered by $\mathtt{T}$-coercivity theory, yet they may correspond to interesting practical configurations (often super-critical), for which Problem (5) admits a unique solution of finite energy for admissible loadings;
- the operator $\mathtt{T}$ must be known explicitly in order to design geometrically fitted mesh families, however it has not been made explicit yet for all 3D Fredholm configurations;
- $\mathtt{T}$-conform meshing may be delicate in practice for general interfaces, especially in 3D.

One can also add to the above list the fact that $\mathtt{T}$-coercivity theory is (at least presently) restricted to the case of isotropic coefficients.

In this work, we aim at developing an alternative approach to the numerical approximation of (the anisotropic) Problem (5), enjoying the following features:

a) to be applicable, without any a priori restriction, as soon as Problem (5) admits, for a given loading, a unique solution (of finite energy);

b) to be applicable without any particular geometrical constraints on the mesh family (except that the mesh cells do not cut the interface).

We will see in the next sections that the new method introduced in this paper fulfills the requirements a) and b) above.

5 Recasting of Problem (4) as a transmission problem

We henceforth assume that $f \in L^2(\Omega)$. The weak formulation (5) becomes: find $\tilde{u} \in H^1_0(\Omega)$ such that

$$a(\tilde{u}, v) = (f, v)_\Omega \qquad \forall v \in H^1_0(\Omega). \tag{12}$$

In this section, we recast Problem (12) into a transmission problem between the positive and negative subdomains.

5.1 Notation

Based on the functional analysis tools from Section 2, we begin by introducing some notation. For $\alpha \in \{\mathrm{p}, \mathrm{n}\}$, we let

$$\gamma_\alpha : H^1(\Omega_\alpha) \to H^{\frac{1}{2}}(\partial\Omega_\alpha)$$

denote the usual trace operator in Ω_α. We now define the space

$$H^1_{0\setminus\Gamma}(\Omega_\alpha) := \left\{ v_\alpha \in H^1(\Omega_\alpha) \mid \gamma_\alpha(v_\alpha)_{|\partial\Omega_\alpha\setminus\overline{\Gamma}} = 0 \right\}.$$

Letting

$$H^{1/2}_{00,\,\alpha}(\Gamma) := \left\{ \varphi_\alpha \in H^{\frac{1}{2}}(\Gamma) \mid \exists\, \hat{\varphi}_\alpha \in H^{\frac{1}{2}}(\partial\Omega_\alpha) \text{ s.t. } \hat{\varphi}_{\alpha\,|\Gamma} = \varphi_\alpha,\ \hat{\varphi}_{\alpha\,|\partial\Omega_\alpha\setminus\overline{\Gamma}} = 0 \right\},$$

there holds $\gamma_\alpha\big(H^1_{0\setminus\Gamma}(\Omega_\alpha)\big)_{|\Gamma} = H^{1/2}_{00,\,\alpha}(\Gamma)$.

When $\partial\Omega_\mathrm{n} = \Gamma$ (inclusion case), then $H^{1/2}_{00,\,\mathrm{n}}(\Gamma) = H^{\frac{1}{2}}(\partial\Omega_\mathrm{n})$. We assume in what follows that $H^{1/2}_{00,\,\mathrm{p}}(\Gamma) = H^{1/2}_{00,\,\mathrm{n}}(\Gamma)$, which holds true for Lipschitz interfaces Γ. We then denote this common Lions–Magenes trace space $H^{1/2}_{00}(\Gamma)$. One can easily remark that

$$H^{1/2}_{00}(\Gamma) = \left\{ v_{|\Gamma} := \gamma_\mathrm{p}(v_{|\Omega_\mathrm{p}})_{|\Gamma} = \gamma_\mathrm{n}(v_{|\Omega_\mathrm{n}})_{|\Gamma}, v \in H^1_0(\Omega) \right\}. \tag{13}$$

We denote by $H^{-\frac{1}{2}}(\Gamma)$ the topological dual of $H^{1/2}_{00}(\Gamma)$, and by $\langle\cdot,\cdot\rangle_\Gamma$ the duality pairing between $H^{-\frac{1}{2}}(\Gamma)$ and $H^{1/2}_{00}(\Gamma)$.

5.2 Weak continuity of the flux

We prove here a (somewhat classical) weak continuity property for the flux at the interface. Recall the notation for the normal flux trace introduced in Section 2. For $\alpha \in \{\mathrm{p}, \mathrm{n}\}$, let $\boldsymbol{n}_\alpha$ be the unit normal vector field to $\partial\Omega_\alpha$ pointing outward Ω_α, and define, for $\tilde{u} \in H^1_0(\Omega)$ solution to Problem (12),

$$\tilde{g}_{\alpha,\Gamma} := \gamma_{\boldsymbol{n}_\alpha,\Gamma}\big(s_\alpha \varpi_\alpha \boldsymbol{\nabla} \tilde{u}_\alpha\big). \tag{14}$$

Since $f \in L^2(\Omega)$, for $\alpha \in \{\mathrm{p}, \mathrm{n}\}$, $\tilde{\boldsymbol{g}}_\alpha := s_\alpha \varpi_\alpha \boldsymbol{\nabla} \tilde{u}_\alpha \in \boldsymbol{H}(\mathrm{div}; \Omega_\alpha)$. As a consequence, $\gamma_{\boldsymbol{n}_\alpha}(\tilde{\boldsymbol{g}}_\alpha) \in H^{-\frac{1}{2}}(\partial\Omega_\alpha)$ and $\tilde{g}_{\alpha,\Gamma} = \gamma_{\boldsymbol{n}_\alpha,\Gamma}(\tilde{\boldsymbol{g}}_\alpha)$ belongs to $H^{-\frac{1}{2}}(\Gamma)$.

Lemma 5.1 (Weak continuity of the flux)**.** *There holds* $\tilde{g}_{\mathrm{p},\Gamma} = -\tilde{g}_{\mathrm{n},\Gamma}$ *in* $H^{-\frac{1}{2}}(\Gamma)$.

Proof. The divergence formula in Ω_α first yields

$$(\operatorname{div} \tilde{\boldsymbol{g}}_\alpha, v_\alpha)_{\Omega_\alpha} + (\tilde{\boldsymbol{g}}_\alpha, \boldsymbol{\nabla} v_\alpha)_{\Omega_\alpha} = \langle \tilde{g}_{\alpha,\Gamma}, \gamma_\alpha(v_\alpha) \rangle_\Gamma \qquad \forall v_\alpha \in H^1_{0\backslash\Gamma}(\Omega_\alpha).$$

Since $\operatorname{div} \tilde{\boldsymbol{g}}_\alpha = -f$ almost everywhere in Ω_α, we then infer

$$-(f, v_\alpha)_{\Omega_\alpha} + (\tilde{\boldsymbol{g}}_\alpha, \boldsymbol{\nabla} v_\alpha)_{\Omega_\alpha} = \langle \tilde{g}_{\alpha,\Gamma}, \gamma_\alpha(v_\alpha) \rangle_\Gamma \qquad \forall v_\alpha \in H^1_{0\backslash\Gamma}(\Omega_\alpha).$$

Let now $v \in H^1_0(\Omega)$. Since $v_{|\Omega_\alpha} \in H^1_{0\backslash\Gamma}(\Omega_\alpha)$, setting $v_\alpha = v_{|\Omega_\alpha}$, there holds

$$-(f, v)_{\Omega_\alpha} + (\tilde{\boldsymbol{g}}_\alpha, \boldsymbol{\nabla} v)_{\Omega_\alpha} = \langle \tilde{g}_{\alpha,\Gamma}, v_{|\Gamma} \rangle_\Gamma,$$

where we recall that the notation $v_{|\Gamma}$ stands for $\gamma_{\mathrm{p}}(v_{|\Omega_{\mathrm{p}}})_{|\Gamma} = \gamma_{\mathrm{n}}(v_{|\Omega_{\mathrm{n}}})_{|\Gamma}$. Summing over $\alpha \in \{\mathrm{p}, \mathrm{n}\}$, and using Problem (12), then yields

$$\langle \tilde{g}_{\mathrm{p},\Gamma} + \tilde{g}_{\mathrm{n},\Gamma}, v_{|\Gamma} \rangle_\Gamma = 0 \qquad \forall v \in H^1_0(\Omega),$$

which, by (13), is finally equivalent to

$$\langle \tilde{g}_{\mathrm{p},\Gamma} + \tilde{g}_{\mathrm{n},\Gamma}, \varphi \rangle_\Gamma = 0 \qquad \forall \varphi \in H^{1/2}_{00}(\Gamma),$$

i.e. $\tilde{g}_{\mathrm{p},\Gamma} = -\tilde{g}_{\mathrm{n},\Gamma}$ in $H^{-\frac{1}{2}}(\Gamma)$. □

As a consequence of Lemma 5.1, for $\tilde{u} \in H^1_0(\Omega)$ solution to Problem (12), one can define

$$H^{-\frac{1}{2}}(\Gamma) \ni \tilde{g}_\Gamma := \tilde{g}_{\mathrm{p},\Gamma} = -\tilde{g}_{\mathrm{n},\Gamma}, \tag{15}$$

so that $\tilde{g}_{\alpha,\Gamma} = s_\alpha \tilde{g}_\Gamma$ for $\alpha \in \{\mathrm{p}, \mathrm{n}\}$.

5.3 Characterization of the solution

For $\alpha \in \{\mathrm{p}, \mathrm{n}\}$, and for any $g_\Gamma \in H^{-\frac{1}{2}}(\Gamma)$, we introduce in the subdomain Ω_α the problem: find $u_\alpha(g_\Gamma) \in H^1_{0\backslash\Gamma}(\Omega_\alpha)$ such that

$$\begin{aligned} a_\alpha(u_\alpha(g_\Gamma), v_\alpha) := s_\alpha \left(\varpi_\alpha \boldsymbol{\nabla} u_\alpha(g_\Gamma), \boldsymbol{\nabla} v_\alpha\right)_{\Omega_\alpha} \\ = (f, v_\alpha)_{\Omega_\alpha} + s_\alpha \langle g_\Gamma, \gamma_\alpha(v_\alpha) \rangle_\Gamma \qquad \forall v_\alpha \in H^1_{0\backslash\Gamma}(\Omega_\alpha). \end{aligned} \tag{16}$$

Recall that Ω_{p} and Ω_{n} are supposed connected. Problem (16) in Ω_{p} always admits a unique solution, since we have assumed that $\partial\Omega_{\mathrm{p}} \cap \partial\Omega$ has nonzero $(d-1)$-dimensional measure. The same holds true in Ω_{n} as soon as $|\partial\Omega_{\mathrm{n}} \cap \partial\Omega|_{d-1} \neq 0$. In the opposite (inclusion) case, $\partial\Omega_{\mathrm{n}} = \Gamma$ and we then assume that g_Γ satisfies $\langle g_\Gamma, 1 \rangle_\Gamma = (f, 1)_{\Omega_{\mathrm{n}}}$ to ensure that the problem admits a solution, which is unique up to an additive constant. We fix this constant by imposing $\left(\gamma_{\mathrm{n}}\left(u_{\mathrm{n}}(g_\Gamma)\right), 1\right)_\Gamma = \left(\gamma_{\mathrm{p}}\left(u_{\mathrm{p}}(g_\Gamma)\right), 1\right)_\Gamma$.

At this point, we can give a sense to the classification `2M, MN` used in Figure 1. Remark that the boundary conditions for Problem (16) in Ω_{p} are always mixed, whereas in Ω_{n} they can be mixed or purely Neumann. Configurations for which both $\partial\Omega_{\mathrm{p}} \cap \partial\Omega$ and $\partial\Omega_{\mathrm{n}} \cap \partial\Omega$ have nonzero $(d-1)$-dimensional measures feature two subproblems of mixed (`M`) type; they are hence denoted `2M`. Configurations for which $|\partial\Omega_{\mathrm{n}} \cap \partial\Omega|_{d-1} = 0$ (inclusion) feature one subproblem in Ω_{p} of mixed (`M`) type, and one subproblem in Ω_{n} of purely Neumann (`N`) type; they are hence denoted `MN`.

Definition 5.2 (Transmission solution)**.** *For $g_\Gamma \in H^{-\frac{1}{2}}(\Gamma)$, we denote by $u(g_\Gamma)$ the function defined on Ω (and not necessarily belonging to $H^1_0(\Omega)$) such that $u(g_\Gamma)_{|\Omega_\alpha} := u_\alpha(g_\Gamma)$ with $u_\alpha(g_\Gamma)$ unique solution to Problem (16) in Ω_α, $\alpha \in \{\mathrm{p},\mathrm{n}\}$.*

The following result establishes an equivalent characterization of the solution to (12).

Proposition 5.3 (Characterization of the solution to (12))**.** *Assume that Problem (12) admits a unique solution $\tilde{u} \in H^1_0(\Omega)$. Then, this solution satisfies $\tilde{u} = u(\tilde{g}_\Gamma)$, where $\tilde{g}_\Gamma \in H^{-\frac{1}{2}}(\Gamma)$ is defined by (15)–(14). Furthermore, $\tilde{g}_\Gamma$ is the unique solution to the minimization problem*

$$\inf_{g_\Gamma \in H^{-\frac{1}{2}}(\Gamma)} \left\| \gamma_\mathrm{p}\left(u_\mathrm{p}(g_\Gamma)\right) - \gamma_\mathrm{n}\left(u_\mathrm{n}(g_\Gamma)\right) \right\|^2_{\frac{1}{2},\Gamma,00}. \tag{17}$$

Proof. (i) Let us begin by proving that $\tilde{u} = u(\tilde{g}_\Gamma)$. We first check, using Problem (4), that in the (inclusion) case when $\partial\Omega_\mathrm{n} = \Gamma$, the compatibility condition $\langle \tilde{g}_\Gamma, 1\rangle_\Gamma = (f,1)_{\Omega_\mathrm{n}}$ does hold. From (16) and the fact that $\tilde{g}_{\alpha,\Gamma} = s_\alpha \tilde{g}_\Gamma$, we infer, for $\alpha \in \{\mathrm{p},\mathrm{n}\}$,

$$a_\alpha(u_\alpha(\tilde{g}_\Gamma), v_\alpha) = (f, v_\alpha)_{\Omega_\alpha} + \langle \tilde{g}_{\alpha,\Gamma}, \gamma_\alpha(v_\alpha)\rangle_\Gamma \qquad \forall v_\alpha \in H^1_{0\backslash\Gamma}(\Omega_\alpha).$$

Using the definition (14) of $\tilde{g}_{\alpha,\Gamma}$, we then get

$$a_\alpha(u_\alpha(\tilde{g}_\Gamma), v_\alpha) = (f, v_\alpha)_{\Omega_\alpha} + \langle \gamma_{\boldsymbol{n}_\alpha,\Gamma}\big(s_\alpha \varpi_\alpha \boldsymbol{\nabla} \tilde{u}_\alpha\big), \gamma_\alpha(v_\alpha)\rangle_\Gamma \qquad \forall v_\alpha \in H^1_{0\backslash\Gamma}(\Omega_\alpha),$$

which yields, by the divergence formula in Ω_α, since $f = -\mathrm{div}(s_\alpha \varpi_\alpha \boldsymbol{\nabla}\tilde{u}_\alpha)$ almost everywhere in Ω_α,

$$a_\alpha(u_\alpha(\tilde{g}_\Gamma) - \tilde{u}_\alpha, v_\alpha) = 0 \qquad \forall v_\alpha \in H^1_{0\backslash\Gamma}(\Omega_\alpha).$$

Testing with $v_\alpha = (u_\alpha(\tilde{g}_\Gamma) - \tilde{u}_\alpha) \in H^1_{0\backslash\Gamma}(\Omega_\alpha)$, using the uniform ellipticity of ϖ_α in Ω_α, and the fact that Ω_α is connected, we infer

$$\tilde{u}_\alpha - u_\alpha(\tilde{g}_\Gamma) = c_\alpha \in \mathbb{R} \qquad \text{in } \Omega_\alpha.$$

In Ω_p, since $|\partial\Omega_\mathrm{p} \cap \partial\Omega|_{d-1} \neq 0$, we always have $c_\mathrm{p} = 0$, hence $\tilde{u}_\mathrm{p} = u_\mathrm{p}(\tilde{g}_\Gamma)$. In Ω_n, when $|\partial\Omega_\mathrm{n} \cap \partial\Omega|_{d-1} \neq 0$, then $c_\mathrm{n} = 0$ and $\tilde{u}_\mathrm{n} = u_\mathrm{n}(\tilde{g}_\Gamma)$. In the opposite (inclusion) case, we fix the constant by imposing $\left(\gamma_\mathrm{n}\left(u_\mathrm{n}(\tilde{g}_\Gamma)\right), 1\right)_\Gamma = \left(\gamma_\mathrm{p}\left(u_\mathrm{p}(\tilde{g}_\Gamma)\right), 1\right)_\Gamma$, i.e.

$$\left(\gamma_\mathrm{n}\left(c_\mathrm{n} - \tilde{u}_\mathrm{n}\right), 1\right)_\Gamma = \left(\gamma_\mathrm{p}\left(\tilde{u}_\mathrm{p}\right), 1\right)_\Gamma.$$

Since $\tilde{u} \in H^1_0(\Omega)$, this yields $c_\mathrm{n} = 0$, and hence $\tilde{u}_\mathrm{n} = u_\mathrm{n}(\tilde{g}_\Gamma)$ as in the mixed case.

(ii) Let us now prove that $\tilde{g}_\Gamma \in H^{-\frac{1}{2}}(\Gamma)$ is the unique solution to the minimization problem (17). We first remark that $\tilde{g}_\Gamma$ is indeed a solution to the problem, owing to the non-negativity of the cost functional and the fact that $\gamma_\mathrm{p}\left(u_\mathrm{p}(\tilde{g}_\Gamma)\right) = \gamma_\mathrm{n}\left(u_\mathrm{n}(\tilde{g}_\Gamma)\right)$ in $H^{1/2}_{00}(\Gamma)$ since $u(\tilde{g}_\Gamma) = \tilde{u} \in H^1_0(\Omega)$. To show uniqueness now, we assume that there exists another minimizer $\check{g}_\Gamma \in H^{-\frac{1}{2}}(\Gamma)$. Then, one must have $\gamma_\mathrm{p}\left(u_\mathrm{p}(\check{g}_\Gamma)\right) = \gamma_\mathrm{n}\left(u_\mathrm{n}(\check{g}_\Gamma)\right)$ in $H^{1/2}_{00}(\Gamma)$, which means that $u(\check{g}_\Gamma) \in H^1_0(\Omega)$. In addition, for $\alpha \in \{\mathrm{p},\mathrm{n}\}$, $u_\alpha(\check{g}_\Gamma) \in H^1_{0\backslash\Gamma}(\Omega_\alpha)$ solves (16). Considering test functions $v_\alpha \in H^1_{0\backslash\Gamma}(\Omega_\alpha)$ in (16) such that $v_\alpha := v_{|\Omega_\alpha}$ for $v \in H^1_0(\Omega)$, we infer by summing over $\alpha \in \{\mathrm{p},\mathrm{n}\}$ that $u(\check{g}_\Gamma) \in H^1_0(\Omega)$ is solution to Problem (12). Since Problem (12) admits a unique solution $\tilde{u} \in H^1_0(\Omega)$, we infer that $u(\check{g}_\Gamma) = \tilde{u} = u(\tilde{g}_\Gamma)$. Then, from (16) we obtain, for $\alpha \in \{\mathrm{p},\mathrm{n}\}$, that $\langle \check{g}_\Gamma - \tilde{g}_\Gamma, \gamma_\alpha(v_\alpha)\rangle_\Gamma = 0$ for all $v_\alpha \in H^1_{0\backslash\Gamma}(\Omega_\alpha)$. Taking $\alpha = \mathrm{p}$ or $\alpha = \mathrm{n}$, we deduce that $\check{g}_\Gamma = \tilde{g}_\Gamma$ in $H^{-\frac{1}{2}}(\Gamma)$, which concludes the proof. □

Remark 5.4. *Note that the uniqueness of the solution to Problem (12) is not needed for the characterization $\tilde{u} = u(\tilde{g}_\Gamma)$ of Proposition 5.3. The uniqueness assumption is only needed to ensure the unique solution of the minimization problem (17).*

6 The numerical method

Henceforth, we assume that the domain Ω and the subdomains Ω_p, Ω_n are (Lipschitz) polytopes.

6.1 Discrete setting and discrete subproblems

Let us first precise our definition of an admissible mesh family.

Definition 6.1 (Admissible mesh family)**.** *A mesh family $(\mathcal{T}_h)_{h>0}$ is* admissible *if (i) for all $h > 0$ in the family, $\mathcal{T}_h$ is a matching simplicial discretization of Ω that is geometrically compliant with the interface Γ (in the sense that there is Γ_h, subset of inner faces of the mesh $\mathcal{T}_h$, such that $\overline{\Gamma} = \bigcup_{F\in\Gamma_h} \overline{F}$), and if (ii) $(\mathcal{T}_h)_{h>0}$ is shape-regular in the sense of Ciarlet [22].*

Our definition of admissibility ensures that no mesh cell can cut the interface Γ.

Let $\mathcal{T}_h$ be a member of an admissible mesh family. The subscript $h > 0$ stands for the meshsize, i.e. the maximum diameter of all the simplices in $\mathcal{T}_h$. For an integer $k \geqslant 1$, we introduce the discrete space

$$U_h^k := \left\{ v_h \in C_0^0(\Omega) \mid v_{h|T} \in \mathbb{P}_d^k(T)\ \forall T \in \mathcal{T}_h \right\} \subset H_0^1(\Omega),$$

where $\mathbb{P}_d^k(T)$ is the vector space of d-variate polynomials of total degree at most k in T. For $\alpha \in \{\mathrm{p}, \mathrm{n}\}$, we let $\mathcal{T}_{\alpha,h}$ denote the restriction of $\mathcal{T}_h$ to Ω_α, and we define

$$U_{\alpha,h}^k := \left\{ v_{\alpha,h} \in C^0(\overline{\Omega_\alpha}) \mid v_{\alpha,h|T} \in \mathbb{P}_d^k(T)\ \forall T \in \mathcal{T}_{\alpha,h},\ \gamma_\alpha(v_{\alpha,h})_{|\partial\Omega_\alpha\setminus\overline{\Gamma}} = 0 \right\} \subset H^1_{0\setminus\Gamma}(\Omega_\alpha).$$

We also introduce the discrete space of normal flux traces at the interface

$$G_{\Gamma,h}^k := \left\{ g_{\Gamma,h} \in L^2(\Gamma) \mid g_{\Gamma,h|F} \in \mathbb{P}_{d-1}^k(F)\ \forall F \in \Gamma_h \right\}, \tag{18}$$

and its affine subspace $G_{\Gamma,h}^{k,\mathtt{N}} := \left\{ g_{\Gamma,h} \in G_{\Gamma,h}^k \mid (g_{\Gamma,h}, 1)_\Gamma = (f,1)_{\Omega_\mathrm{n}} \right\}$, where $\mathbb{P}_{d-1}^k(F)$ denotes the space of $(d-1)$-variate polynomials of total degree at most k on F.

For $\alpha \in \{\mathrm{p}, \mathrm{n}\}$, and for any $g_\Gamma \in H^{-\frac{1}{2}}(\Gamma)$, we introduce the following conforming finite element approximation of Problem (16) in the subdomain Ω_α: find $u_{\alpha,h}(g_\Gamma) \in U_{\alpha,h}^k$ such that

$$a_\alpha(u_{\alpha,h}(g_\Gamma), v_{\alpha,h}) = (f, v_{\alpha,h})_{\Omega_\alpha} + s_\alpha \langle g_\Gamma, \gamma_\alpha(v_{\alpha,h})\rangle_\Gamma \qquad \forall v_{\alpha,h} \in U_{\alpha,h}^k. \tag{19}$$

Problem (19) always admits a unique solution in Ω_p, and the same holds true in Ω_n in the case of mixed boundary conditions. In the purely Neumann case of an inclusion, in which we assume that the flux g_Γ satisfies $\langle g_\Gamma, 1\rangle_\Gamma = (f,1)_{\Omega_\mathrm{n}}$, the solution to Problem (19) in Ω_n is unique up to an additive constant. We fix the constant imposing $\left(\gamma_\mathrm{n}\left(u_{\mathrm{n},h}(g_\Gamma)\right), 1\right)_\Gamma = \left(\gamma_\mathrm{p}\left(u_{\mathrm{p},h}(g_\Gamma)\right), 1\right)_\Gamma$.

Definition 6.2 (Discrete transmission solution)**.** *For $g_\Gamma \in H^{-\frac{1}{2}}(\Gamma)$, we denote by $u_h(g_\Gamma)$ the function defined on Ω (and not necessarily belonging to U_h^k) such that $u_h(g_\Gamma)_{|\Omega_\alpha} := u_{\alpha,h}(g_\Gamma)$ with $u_{\alpha,h}(g_\Gamma)$ unique solution to Problem (19) in Ω_α, $\alpha \in \{\mathrm{p}, \mathrm{n}\}$.*

6.2 Minimization procedure

We define the cost functional $J_h : G^k_{\Gamma,h} \to [0,\infty)$ such that, for any $g_{\Gamma,h} \in G^k_{\Gamma,h}$,

$$J_h(g_{\Gamma,h}) := \|\gamma_{\mathrm{p}}\left(u_{\mathrm{p},h}(g_{\Gamma,h})\right) - \gamma_{\mathrm{n}}\left(u_{\mathrm{n},h}(g_{\Gamma,h})\right)\|^2_{0,\Gamma} + \lambda(h)\,\sigma_\flat^{-2}\|g_{\Gamma,h}\|^2_{0,\Gamma}, \tag{20}$$

where $\lambda : (0,\infty) \to (0,\infty)$ is a function such that $\lim_{h\to 0}\lambda(h) = 0$. When $|\partial\Omega_{\mathrm{n}} \cap \partial\Omega|_{d-1} \neq 0$, we consider the minimization problem

$$\inf_{g_{\Gamma,h}\in G^k_{\Gamma,h}} J_h(g_{\Gamma,h}), \tag{21}$$

otherwise ($\partial\Omega_{\mathrm{n}} = \Gamma$) we consider the following variant:

$$\inf_{g_{\Gamma,h}\in G^{k,\mathtt{N}}_{\Gamma,h}} J_h(g_{\Gamma,h}). \tag{22}$$

Lemma 6.3 (Well-posedness of the minimization problems)**.** *Both minimization problems* (21) *and* (22) *admit a unique solution.*

Proof. We focus on Problem (21); Problem (22) can be treated similarly invoking that $G^{k,\mathtt{N}}_{\Gamma,h}$ is a closed convex subspace of $G^k_{\Gamma,h}$. The functional J_h is continuous, and $\lim_{\|g_{\Gamma,h}\|_{0,\Gamma}\to\infty} J_h(g_{\Gamma,h}) = +\infty$, hence Problem (21) admits at least one solution. Let $g_{\Gamma,h} \in G^k_{\Gamma,h}$. A straightforward computation yields, for all $i_h, j_h \in G^k_{\Gamma,h}$,

$$\begin{aligned}\mathrm{d}^2 J_h(g_{\Gamma,h})(i_h,j_h) = 2\left(\left[\gamma_{\mathrm{p}}\left(u'_{\mathrm{p},h}(i_h)\right) - \gamma_{\mathrm{n}}\left(u'_{\mathrm{n},h}(i_h)\right)\right],\left[\gamma_{\mathrm{p}}\left(u'_{\mathrm{p},h}(j_h)\right) - \gamma_{\mathrm{n}}\left(u'_{\mathrm{n},h}(j_h)\right)\right]\right)_\Gamma \\ + 2\,\lambda(h)\,\sigma_\flat^{-2}\,(i_h,j_h)_\Gamma,\end{aligned}$$

where, for $\alpha \in \{\mathrm{p},\mathrm{n}\}$, and $\iota_\Gamma \in L^2(\Gamma)$, $u'_{\alpha,h}(\iota_\Gamma) \in U^k_{\alpha,h}$ solves

$$a_\alpha(u'_{\alpha,h}(\iota_\Gamma), v_{\alpha,h}) = s_\alpha(\iota_\Gamma, \gamma_\alpha(v_{\alpha,h}))_\Gamma \qquad \forall v_{\alpha,h} \in U^k_{\alpha,h}. \tag{23}$$

Hence, for all $i_h \in G^k_{\Gamma,h}$,

$$\mathrm{d}^2 J_h(g_{\Gamma,h})(i_h,i_h) = 2\,\|\gamma_{\mathrm{p}}\left(u'_{\mathrm{p},h}(i_h)\right) - \gamma_{\mathrm{n}}\left(u'_{\mathrm{n},h}(i_h)\right)\|^2_{0,\Gamma} + 2\,\lambda(h)\,\sigma_\flat^{-2}\|i_h\|^2_{0,\Gamma}.$$

Thus, the cost functional J_h is strictly convex on $G^k_{\Gamma,h}$, meaning that the minimizer to Problem (21) is unique. □

Remark 6.4 (Tikhonov regularization)**.** *The addition of the term $\lambda(h)\,\sigma_\flat^{-2}\|\cdot\|^2_{0,\Gamma}$ in the cost functional J_h, which plays the role of a Tikhonov regularization [54] (see also [20]), ensures the uniqueness of the minimizer to Problems* (21) *and* (22)*. Without this term, the sole existence can be proved, as a consequence of the linear least-squares nature of Problems* (21) *and* (22)*.*

Mimicking, at the discrete level, the characterization of the continuous solution from Proposition 5.3, we let $\tilde{g}_{\Gamma,h} \in G^k_{\Gamma,h}$ (resp. $\tilde{g}_{\Gamma,h} \in G^{k,\mathtt{N}}_{\Gamma,h}$) denote the unique minimizer to Problem (21) (resp. (22)), and we define $\tilde{u}_h := u_h(\tilde{g}_{\Gamma,h})$ (cf. Definition 6.2) as our approximation of the solution $\tilde{u} \in H^1_0(\Omega)$ to Problem (12). Remark that $\tilde{u}_h$ is always well-defined, as a consequence of the well-posedness of the subproblems (19), and of that of the optimization problems (21) and (22). Note that $\tilde{u}_h$ does not a priori belong to U^k_h, and thus to $H^1_0(\Omega)$. For an algebraic realization of our method, we refer to Section 7.1 below.

Remark 6.5 (Link with T-coercivity based approximation)**.** *Assume that the form a from Problem (12) is $\mathcal{T}$-coercive for some operator $\mathcal{T}$, and denote by $\tilde{u}_h^{\mathrm{c}} \in U_h^k \subset H_0^1(\Omega)$ the conforming finite element approximation of $\tilde{u}$ on a $\mathcal{T}$-conforming mesh $\mathcal{T}_h$ ($\tilde{u}_h^{\mathrm{c}}$ is then known to be well-defined; cf. Proposition 4.8). Nothing guarantees, as in Proposition 5.3, that it may exist $\tilde{g}_{\Gamma,h} \in G_{\Gamma,h}^k$ (or even in $H^{-\frac{1}{2}}(\Gamma)$) such that $\tilde{u}_h^{\mathrm{c}} = u_h(\tilde{g}_{\Gamma,h})$. We hence do not know if the solution $\tilde{u}_h$ given by our approach on $\mathcal{T}_h$ degenerates towards $\tilde{u}_h^{\mathrm{c}}$ when removing the Tikhonov regularization from J_h. However, in the case the minimum value of J_h without regularization is zero, and is attained for some $\tilde{g}_{\Gamma,h}$, then $\tilde{u}_h = u_h(\tilde{g}_{\Gamma,h})$ is equal to $\tilde{u}_h^{\mathrm{c}} \in H_0^1(\Omega)$.*

6.3 Convergence of the method

Before proving our convergence result, we need to quantify the jump of $u_h(\tilde{g}_\Gamma)$ along the interface. Recall that $u_h(\tilde{g}_\Gamma)$ is the discrete transmission solution (cf. Definition 6.2) corresponding to the exact normal flux trace $\tilde{g}_\Gamma \in H^{-\frac{1}{2}}(\Gamma)$ defined by (14)–(15). The following lemma relies on several classical error estimates for the discrete solutions to variable elliptic problems featuring either mixed or purely Neumann boundary conditions, which are recalled in Appendix B, as well as on the notion of dual regularity exponent (cf. Assumption B.1).

Lemma 6.6 (Bound on the interface jump of $u_h(\tilde{g}_\Gamma)$)**.** *Let $\tilde{u} \in H_0^1(\Omega)$ be a solution to Problem (12). Let $m \geqslant 0$ be some exponent such that $\tilde{u}_{|\Omega_\alpha} \in H^{1+m}(\Omega_\alpha)$ for $\alpha \in \{\mathrm{p},\mathrm{n}\}$, and let $\tau := \min(m,k)$. Denote by $\varepsilon_{\mathrm{p}}, \varepsilon_{\mathrm{n}} \in (\frac{1}{2}, 1]$ the dual regularity exponents of the subproblems (16) in Ω_{p} and Ω_{n}, respectively. Then, letting $\tilde{\delta} := 2\tau + \min(\varepsilon_{\mathrm{p}}, \varepsilon_{\mathrm{n}}) > 0$, the following estimate holds true, for some constant $\tilde{c} > 0$:*

$$\|\gamma_{\mathrm{p}}\left(u_{\mathrm{p},h}(\tilde{g}_\Gamma)\right) - \gamma_{\mathrm{n}}\left(u_{\mathrm{n},h}(\tilde{g}_\Gamma)\right)\|_{0,\Gamma} \leqslant \tilde{c}\,\rho\,h^{\frac{\tilde{\delta}}{2}}\left(|\tilde{u}|_{1+\tau,\Omega_{\mathrm{p}}} + |\tilde{u}|_{1+\tau,\Omega_{\mathrm{n}}}\right). \tag{24}$$

Proof. According to Proposition 5.3 and Remark 5.4, $u(\tilde{g}_\Gamma) = \tilde{u} \in H_0^1(\Omega)$, hence $\gamma_{\mathrm{p}}(u_{\mathrm{p}}(\tilde{g}_\Gamma)) = \gamma_{\mathrm{n}}(u_{\mathrm{n}}(\tilde{g}_\Gamma))$ on Γ. This allows us to infer

$$\|\gamma_{\mathrm{p}}\left(u_{\mathrm{p},h}(\tilde{g}_\Gamma)\right) - \gamma_{\mathrm{n}}\left(u_{\mathrm{n},h}(\tilde{g}_\Gamma)\right)\|_{0,\Gamma} \leqslant \sum_{\alpha\in\{\mathrm{p},\mathrm{n}\}} \|\gamma_\alpha\left(u_\alpha(\tilde{g}_\Gamma) - u_{\alpha,h}(\tilde{g}_\Gamma)\right)\|_{0,\Gamma}.$$

To estimate the right-hand side of this inequality, we use the approximation results derived in Appendix B. Note that if $(\mathcal{T}_h)_{h>0}$ is admissible in the sense of Definition 6.1, then the mesh families $(\mathcal{T}_{\alpha,h})_{h>0}$, $\alpha \in \{\mathrm{p},\mathrm{n}\}$, are admissible in the sense of Definition B.3.

In Ω_{p}, the subproblem (16) is always endowed with mixed boundary conditions. We thus apply the results of Appendix B.2, with $\mathcal{D} := \Omega_{\mathrm{p}}$, $\Upsilon_{\mathrm{f}} := \Gamma$ (hence, $\Upsilon_{\mathrm{t}} = \partial\Omega_{\mathrm{p}}\backslash\overline{\Gamma}$), $\mathfrak{o} := \varpi_{\mathrm{p}}$, $r := f_{|\Omega_{\mathrm{p}}}$, $\phi := 0$, and $\theta := \tilde{g}_\Gamma$. By Remark B.10, we get

$$\|\gamma_{\mathrm{p}}\left(u_{\mathrm{p}}(\tilde{g}_\Gamma) - u_{\mathrm{p},h}(\tilde{g}_\Gamma)\right)\|_{0,\Gamma} \leqslant c_{\mathrm{p}}\,\rho\,h^{\frac{\delta_{\mathrm{p}}}{2}}|\tilde{u}|_{1+\tau,\Omega_{\mathrm{p}}}, \tag{25}$$

with $\delta_{\mathrm{p}} := 2\tau + \varepsilon_{\mathrm{p}}$.

Assume that the subproblem (16) in Ω_{n} is also endowed with mixed boundary conditions (case $|\partial\Omega_{\mathrm{n}} \cap \partial\Omega|_{d-1} \neq 0$). Applying again the results of Appendix B.2, this time with $\mathcal{D} := \Omega_{\mathrm{n}}$, $\Upsilon_{\mathrm{f}} := \Gamma$ (hence, $\Upsilon_{\mathrm{t}} = \partial\Omega_{\mathrm{n}}\backslash\overline{\Gamma}$), $\mathfrak{o} := \varpi_{\mathrm{n}}$, $r := -f_{|\Omega_{\mathrm{n}}}$, $\phi := 0$, and $\theta := \tilde{g}_\Gamma$, we get

$$\|\gamma_{\mathrm{n}}\left(u_{\mathrm{n}}(\tilde{g}_\Gamma) - u_{\mathrm{n},h}(\tilde{g}_\Gamma)\right)\|_{0,\Gamma} \leqslant c_{\mathrm{n}}\,\rho\,h^{\frac{\delta_{\mathrm{n}}}{2}}|\tilde{u}|_{1+\tau,\Omega_{\mathrm{n}}},$$

with $\delta_{\mathrm{n}} := 2\tau + \varepsilon_{\mathrm{n}}$. Recalling (25), and remarking that $\tilde{\delta} = \min(\delta_{\mathrm{p}}, \delta_{\mathrm{n}})$, (24) follows easily.

Now, assume that the subproblem (16) in Ω_{n} is of pure Neumann type (case $\partial\Omega_{\mathrm{n}} = \Gamma$). We apply the results of Appendix B.3, with $\mathcal{D} := \Omega_{\mathrm{n}}$, $\Lambda := \Gamma$, $\mathfrak{o} := \varpi_{\mathrm{n}}$, $r := -f_{|\Omega_{\mathrm{n}}}$, and $\theta := \tilde{g}_\Gamma$. Setting $\hat{\mathcal{D}} := \Omega$, $\hat{\mathfrak{o}} := \varpi$, and $\hat{u} := \tilde{u}$, there holds $\hat{\varrho} = \rho$ and $\kappa = |\Gamma|_{d-1}^{-1}(\gamma_{\mathrm{p}}(u_{\mathrm{p}}(\tilde{g}_\Gamma)), 1)_\Gamma$. We let $\kappa_h := |\Gamma|_{d-1}^{-1}(\gamma_{\mathrm{p}}(u_{\mathrm{p},h}(\tilde{g}_\Gamma)), 1)_\Gamma$. The Cauchy–Schwarz inequality then yields

$$|\Gamma|_{d-1}^{1/2}|\kappa - \kappa_h| \leqslant \|\gamma_{\mathrm{p}}(u_{\mathrm{p}}(\tilde{g}_\Gamma) - u_{\mathrm{p},h}(\tilde{g}_\Gamma))\|_{0,\Gamma},$$

hence, as a consequence of (25), the estimate (62) holds true with $\mathcal{D}' = \Omega_{\mathrm{p}}$, $u' = \tilde{u}_{\mathrm{p}}$, $\varrho' \leqslant \hat{\varrho} = \rho$, and $\delta' = \delta_{\mathrm{p}}$. Using the notation of Lemma B.13, $\hat{\delta} = \min(\delta, \delta')$, with $\delta = \delta_{\mathrm{n}}$ and $\delta' = \delta_{\mathrm{p}}$. Thus, $\hat{\delta} = \tilde{\delta}$, and we infer

$$\|\gamma_{\mathrm{n}}\left(u_{\mathrm{n}}(\tilde{g}_\Gamma) - u_{\mathrm{n},h}(\tilde{g}_\Gamma)\right)\|_{0,\Gamma} \leqslant c_{\mathrm{n}}\,\rho\,h^{\frac{\tilde{\delta}}{2}}\left(|\tilde{u}|_{1+\tau,\Omega_{\mathrm{p}}} + |\tilde{u}|_{1+\tau,\Omega_{\mathrm{n}}}\right),$$

which, combined to (25), yields (24). □

We are now ready to prove convergence of our optimization-based method. Let us just recall the principle of our approach. Our approach consists in defining an approximation of the solution $\tilde{u} \in H_0^1(\Omega)$ to Problem (12) as $\tilde{u}_h := u_h(\tilde{g}_{\Gamma,h})$, where $\tilde{g}_{\Gamma,h} \in G_{\Gamma,h}^k$ (resp. $\tilde{g}_{\Gamma,h} \in G_{\Gamma,h}^{k,\mathtt{N}}$) solves the well-posed minimization problem (21) (resp. (22)), and $u_h(\tilde{g}_{\Gamma,h})$ is the discrete transmission solution corresponding to $\tilde{g}_{\Gamma,h}$.

Theorem 6.7 (Convergence of the method)**.** *Suppose that Problem (12) admits a unique solution $\tilde{u} \in H_0^1(\Omega)$. Assume that $\tilde{g}_\Gamma \in L^2(\Gamma)$, with $\tilde{g}_\Gamma$ as defined in (14)–(15). Then, choosing $\lambda(h)$ in (20) such that $\lambda(h) = c\,h^\delta$ for some $c > 0$ and $0 < \delta < \tilde{\delta}$, where $\tilde{\delta}$ is the positive number introduced in Lemma 6.6, there holds, strongly as $h \to 0$:*

$$\tilde{g}_{\Gamma,h} \to \tilde{g}_\Gamma \quad \textit{in } L^2(\Gamma), \qquad \boldsymbol{\nabla}_h \tilde{u}_h \to \boldsymbol{\nabla}\tilde{u} \quad \textit{in } \boldsymbol{L}^2(\Omega), \qquad \tilde{u}_h \to \tilde{u} \quad \textit{in } L^2(\Omega), \tag{26}$$

where $(\boldsymbol{\nabla}_h)_{h>0}$ is the broken gradient operator on $(\mathcal{T}_h)_{h>0}$.

Proof. The proof proceeds in three steps.

(i) Weak convergence: By linearity, for $\alpha \in \{\mathrm{p}, \mathrm{n}\}$, we first write

$$u_{\alpha,h}(\tilde{g}_{\Gamma,h}) = u_{\alpha,h}(\tilde{g}_\Gamma) + \tilde{u}'_{\alpha,h}, \tag{27}$$

with $U_{\alpha,h}^k \ni \tilde{u}'_{\alpha,h} := u'_{\alpha,h}(\tilde{g}_{\Gamma,h} - \tilde{g}_\Gamma)$ as defined in (23), i.e. solution to

$$(\varpi_\alpha \boldsymbol{\nabla}\tilde{u}'_{\alpha,h}, \boldsymbol{\nabla}v_{\alpha,h})_{\Omega_\alpha} = (\tilde{g}_{\Gamma,h} - \tilde{g}_\Gamma, \gamma_\alpha(v_{\alpha,h}))_\Gamma \qquad \forall v_{\alpha,h} \in U_{\alpha,h}^k. \tag{28}$$

Testing (28) with $\tilde{u}'_{\alpha,h} \in U_{\alpha,h}^k$, and using the Cauchy–Schwarz inequality, yields

$$\|\boldsymbol{\nabla}\tilde{u}'_{\alpha,h}\|_{0,\Omega_\alpha}^2 \leqslant \sigma_\flat^{-1}\|\tilde{g}_{\Gamma,h} - \tilde{g}_\Gamma\|_{0,\Gamma}\|\gamma_\alpha(\tilde{u}'_{\alpha,h})\|_{0,\Gamma}. \tag{29}$$

In Ω_{p}, starting from (29), using the trace inequality (1) (with $\mathcal{D} \leftarrow \Omega_{\mathrm{p}}$ and $s = 1$), and applying a classical Poincaré–Steklov inequality in $H_{0\setminus\Gamma}^1(\Omega_{\mathrm{p}})$, we infer

$$\|\boldsymbol{\nabla}\tilde{u}'_{\mathrm{p},h}\|_{0,\Omega_{\mathrm{p}}} \leqslant c_{\mathrm{p}}\,\sigma_\flat^{-1}\|\tilde{g}_{\Gamma,h} - \tilde{g}_\Gamma\|_{0,\Gamma}. \tag{30}$$

An equivalent inequality can also be inferred in $\Omega_{\rm n}$. When $|\partial\Omega_{\rm n}\cap\partial\Omega|_{d-1}\neq 0$, the proof is identical to (30). When $\partial\Omega_{\rm n}=\Gamma$, the derivation is a bit less straightforward (the details are given in Remark 6.8 below), but leads to

$$\|\boldsymbol{\nabla}\tilde{u}'_{{\rm n},h}\|_{0,\Omega_{\rm n}}+\|\tilde{u}'_{{\rm n},h}\|_{0,\Omega_{\rm n}}\leqslant c_{\rm n}\,\sigma_\flat^{-1}\|\tilde{g}_{\Gamma,h}-\tilde{g}_\Gamma\|_{0,\Gamma}. \tag{31}$$

Now, we leverage the fact that $J_h(\tilde{g}_{\Gamma,h})\leqslant J_h(g_{\Gamma,h})$ for all $g_{\Gamma,h}\in G^k_{\Gamma,h}$ (resp. for all $g_{\Gamma,h}\in G^{k,\mathtt{N}}_{\Gamma,h}$ in the inclusion case), and we choose $g_{\Gamma,h}=\pi^k_h(\tilde{g}_\Gamma)$ the $L^2(\Gamma)$-orthogonal projection of $\tilde{g}_\Gamma$ onto $G^k_{\Gamma,h}$ (remark that, in the inclusion case, there holds $g_{\Gamma,h}\in G^{k,\mathtt{N}}_{\Gamma,h}$). Using the boundedness and orthogonality properties of the projector, we infer

$$\sigma_\flat^{-2}\|\tilde{g}_{\Gamma,h}\|^2_{0,\Gamma}\leqslant\frac{\|\gamma_{\rm p}(u_{{\rm p},h}(\tilde{g}_\Gamma))-\gamma_{\rm n}(u_{{\rm n},h}(\tilde{g}_\Gamma))\|^2_{0,\Gamma}}{\lambda(h)}+\sigma_\flat^{-2}\|\tilde{g}_\Gamma\|^2_{0,\Gamma}. \tag{32}$$

Owing to (24), and to the fact that $\lambda(h)=c\,h^\delta$ with $\delta<\tilde{\delta}$, we deduce from (32) that $\|\tilde{g}_{\Gamma,h}\|_{0,\Gamma}$ is uniformly bounded with respect to h. We can thus infer the existence of $\tilde{g}_{\Gamma,0}\in L^2(\Gamma)$ such that, up to a subsequence (retaining the same notation), $\tilde{g}_{\Gamma,h}\rightharpoonup\tilde{g}_{\Gamma,0}$ weakly in $L^2(\Gamma)$. From (30) and (31), together with the uniform boundedness of $(\tilde{g}_{\Gamma,h})_{h>0}$ in $L^2(\Gamma)$, we also infer the uniform boundedness of $(\tilde{u}'_{\alpha,h})_{h>0}$ in $H^1(\Omega_\alpha)$ for $\alpha\in\{{\rm p},{\rm n}\}$. Thus, by Rellich's theorem (and a standard limit regularity argument), there exist $\tilde{u}'_\alpha\in H^1_{0\setminus\Gamma}(\Omega_\alpha)$ such that, up to a subsequence (retaining the same notation),

$$\boldsymbol{\nabla}\tilde{u}'_{\alpha,h}\rightharpoonup\boldsymbol{\nabla}\tilde{u}'_\alpha\text{ weakly in }\boldsymbol{L}^2(\Omega_\alpha),\qquad\tilde{u}'_{\alpha,h}\to\tilde{u}'_\alpha\text{ strongly in }L^2(\Omega_\alpha), \tag{33}$$

$$\gamma_\alpha(\tilde{u}'_{\alpha,h})\rightharpoonup\gamma_\alpha(\tilde{u}'_\alpha)\text{ weakly in }L^2(\Gamma). \tag{34}$$

(ii) Identification of the limits: From the relation $J_h(\tilde{g}_{\Gamma,h})\leqslant J_h(g_{\Gamma,h})$ for all $g_{\Gamma,h}\in G^k_{\Gamma,h}$ applied to $g_{\Gamma,h}=\pi^k_h(\tilde{g}_\Gamma)$ (with adaptation as above in the inclusion case), from (24), and from the fact that $0<\delta<\tilde{\delta}$, we infer that

$$\begin{aligned}\|\gamma_{\rm p}(u_{{\rm p},h}(\tilde{g}_{\Gamma,h}))-\gamma_{\rm n}(u_{{\rm n},h}(\tilde{g}_{\Gamma,h}))\|^2_{0,\Gamma}&\leqslant\|\gamma_{\rm p}(u_{{\rm p},h}(\tilde{g}_\Gamma))-\gamma_{\rm n}(u_{{\rm n},h}(\tilde{g}_\Gamma))\|^2_{0,\Gamma}\\&\quad+\lambda(h)\,\sigma_\flat^{-2}\|\tilde{g}_\Gamma\|^2_{0,\Gamma}\leqslant C\rho^2h^\delta\mathcal{N}_\Omega(\tilde{u}),\end{aligned} \tag{35}$$

where $\mathcal{N}_\Omega(\tilde{u}):=|\tilde{u}|^2_{1+\tau,\Omega_{\rm p}}+|\tilde{u}|^2_{1+\tau,\Omega_{\rm n}}+\sigma_\sharp^{-2}\|\tilde{g}_\Gamma\|^2_{0,\Gamma}$. We then deduce from (35) that, as $h\to 0$,

$$\|\gamma_{\rm p}(u_{{\rm p},h}(\tilde{g}_{\Gamma,h}))-\gamma_{\rm n}(u_{{\rm n},h}(\tilde{g}_{\Gamma,h}))\|_{0,\Gamma}\to 0.$$

Combining this result with (24) and (27), we readily get that, as $h\to 0$,

$$\|\gamma_{\rm p}(\tilde{u}'_{{\rm p},h})-\gamma_{\rm n}(\tilde{u}'_{{\rm n},h})\|_{0,\Gamma}\to 0,$$

and hence, from (34), that $\gamma_{\rm p}(\tilde{u}'_{\rm p})=\gamma_{\rm n}(\tilde{u}'_{\rm n})$ almost everywhere on Γ. Passing to the limit $h\to 0$ in (28) (where both sides are multiplied by s_α), using (33), the weak convergence result $\tilde{g}_{\Gamma,h}\rightharpoonup\tilde{g}_{\Gamma,0}$, and a strongly convergent interpolant for test functions, one can show by sum over $\alpha\in\{{\rm p},{\rm n}\}$ that $\tilde{u}'\in H^1_0(\Omega)$, defined by $\tilde{u}'_{|\Omega_\alpha}:=\tilde{u}'_\alpha$ for $\alpha\in\{{\rm p},{\rm n}\}$, satisfies

$$a(\tilde{u}',v)=0\qquad\forall v\in H^1_0(\Omega).$$

This implies, from the injectivity of Problem (12), that $\tilde{u}' \equiv 0$. Also, the uniqueness of the limit implies that the whole sequences converge in (33)–(34). By (27) and the strong convergences of $(\boldsymbol{\nabla} u_{\alpha,h}(\tilde{g}_\Gamma))_{h>0}$ and $(u_{\alpha,h}(\tilde{g}_\Gamma))_{h>0}$ towards $\boldsymbol{\nabla}\tilde{u}_\alpha$ and $\tilde{u}_\alpha$, respectively in $\boldsymbol{L}^2(\Omega_\alpha)$ and $L^2(\Omega_\alpha)$ for $\alpha \in \{\mathrm{p},\mathrm{n}\}$, we have thus proved at this point that

$$\boldsymbol{\nabla}_h \tilde{u}_h \rightharpoonup \boldsymbol{\nabla}\tilde{u} \quad \text{weakly in } \boldsymbol{L}^2(\Omega), \qquad \tilde{u}_h \to \tilde{u} \quad \text{strongly in } L^2(\Omega).$$

Passing again to the limit in (28), using a strongly convergent interpolant for test functions, and the fact that $\tilde{u}' \equiv 0$, one obtains

$$(\tilde{g}_{\Gamma,0} - \tilde{g}_\Gamma, \gamma_\alpha(v_\alpha))_\Gamma = 0 \qquad \forall v_\alpha \in H^1_{0\backslash\Gamma}(\Omega_\alpha).$$

From there, fixing $\alpha \in \{\mathrm{p},\mathrm{n}\}$, since $\gamma_\alpha(H^1_{0\backslash\Gamma}(\Omega_\alpha))_{|\Gamma} = H^{1/2}_{00}(\Gamma)$ and $H^{1/2}_{00}(\Gamma)$ is dense in $L^2(\Gamma)$, we infer that $\tilde{g}_{\Gamma,0} = \tilde{g}_\Gamma$. The uniqueness of the limit implies that the whole sequence $(\tilde{g}_{\Gamma,h})_{h>0}$ converges towards $\tilde{g}_\Gamma$. We have thus proved that $\tilde{g}_{\Gamma,h} \rightharpoonup \tilde{g}_\Gamma$ weakly in $L^2(\Gamma)$.

(iii) Strong convergence: Passing to the limit in (32) (recall that $\lambda(h) = c\,h^\delta$ with $\delta < \tilde{\delta}$) yields, owing to the weak convergence of $(\tilde{g}_{\Gamma,h})_{h>0}$ towards $\tilde{g}_\Gamma$,

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

which readily implies the strong convergence of $(\tilde{g}_{\Gamma,h})_{h>0}$ towards $\tilde{g}_\Gamma$ in $L^2(\Gamma)$. Now, testing (28) with $v_{\alpha,h} = \tilde{u}'_{\alpha,h}$ and passing to the limit, owing to the strong convergence of $(\tilde{g}_{\Gamma,h})_{h>0}$ towards $\tilde{g}_\Gamma$ and to the weak convergence (34) of $(\gamma_\alpha(\tilde{u}'_{\alpha,h}))_{h>0}$, we infer the strong convergence of $(\boldsymbol{\nabla}\tilde{u}'_{\alpha,h})_{h>0}$ towards $\mathbf{0}$ in $\boldsymbol{L}^2(\Omega_\alpha)$, for $\alpha \in \{\mathrm{p},\mathrm{n}\}$. By (27), combined with the strong convergence of $(\boldsymbol{\nabla} u_{\alpha,h}(\tilde{g}_\Gamma))_{h>0}$, this finally proves (26), and concludes the proof. □

Remark 6.8 (Proof of (31), inclusion case)**.** *Combining the trace inequality (1) (with $\mathcal{D} \leftarrow \Omega_\mathrm{n}$ and $s = 1$) with a generalized Poincaré–Steklov inequality (cf. [27, Lemma 3.30]) and the fact that $(\gamma_\mathrm{n}(\tilde{u}'_{\mathrm{n},h}),1)_\Gamma = (\gamma_\mathrm{p}(\tilde{u}'_{\mathrm{p},h}),1)_\Gamma$ (owing to (27) along with the definition of the discrete transmission solution in the inclusion case), we first infer that*

$$\|\gamma_\mathrm{n}(\tilde{u}'_{\mathrm{n},h})\|_{0,\Gamma} \leqslant c_1 \left(\|\boldsymbol{\nabla}\tilde{u}'_{\mathrm{n},h}\|_{0,\Omega_\mathrm{n}} + |\Gamma|_{d-1}^{-\frac{1}{2}} |(\gamma_\mathrm{n}(\tilde{u}'_{\mathrm{n},h}),1)_\Gamma| \right) \leqslant c_1 \left(\|\boldsymbol{\nabla}\tilde{u}'_{\mathrm{n},h}\|_{0,\Omega_\mathrm{n}} + \|\gamma_\mathrm{p}(\tilde{u}'_{\mathrm{p},h})\|_{0,\Gamma} \right).$$

Then, by the trace inequality (1) (with $\mathcal{D} \leftarrow \Omega_\mathrm{p}$ and $s = 1$), a classical Poincaré–Steklov inequality in $H^1_{0\backslash\Gamma}(\Omega_\mathrm{p})$, and (30), we obtain

$$\|\gamma_\mathrm{n}(\tilde{u}'_{\mathrm{n},h})\|_{0,\Gamma} \leqslant c_2 \left(\|\boldsymbol{\nabla}\tilde{u}'_{\mathrm{n},h}\|_{0,\Omega_\mathrm{n}} + \|\boldsymbol{\nabla}\tilde{u}'_{\mathrm{p},h}\|_{0,\Omega_\mathrm{p}} \right) \leqslant c_3 \left(\|\boldsymbol{\nabla}\tilde{u}'_{\mathrm{n},h}\|_{0,\Omega_\mathrm{n}} + \sigma_\flat^{-1} \|\tilde{g}_{\Gamma,h} - \tilde{g}_\Gamma\|_{0,\Gamma} \right). \tag{36}$$

In Ω_n, starting from (29), and using (36), we thus get

$$\|\boldsymbol{\nabla}\tilde{u}'_{\mathrm{n},h}\|^2_{0,\Omega_\mathrm{n}} \leqslant c_3\, \sigma_\flat^{-1} \|\tilde{g}_{\Gamma,h} - \tilde{g}_\Gamma\|_{0,\Gamma} \left(\|\boldsymbol{\nabla}\tilde{u}'_{\mathrm{n},h}\|_{0,\Omega_\mathrm{n}} + \sigma_\flat^{-1} \|\tilde{g}_{\Gamma,h} - \tilde{g}_\Gamma\|_{0,\Gamma} \right),$$

which eventually yields, by Young's inequality, the estimate (31) on $\|\boldsymbol{\nabla}\tilde{u}'_{\mathrm{n},h}\|_{0,\Omega_\mathrm{n}}$. The estimate on $\|\tilde{u}'_{\mathrm{n},h}\|_{0,\Omega_\mathrm{n}}$ can be obtained leveraging the same arguments.

It is crucial to note that the convergence result of Theorem 6.7 is valid as soon as Problem (12) admits a unique solution for the loading $f \in L^2(\Omega)$ at hand. In particular, nothing needs to be assumed on the invertibility of the operator $\mathtt{A}$ associated with the problem, as it is the case for $\mathtt{T}$-coercivity based approximation. In this respect, our approach checks the requirement a) from Section 4.4. Furthermore, the convergence result does not rely on any particular geometrical constraint (with respect to the interface) to be imposed to mesh families, as it is the case for $\mathtt{T}$-coercivity based approximation. Our approach hence also gives a positive answer to the requirement b) from Section 4.4.

Remark 6.9 (Regularity assumption on $\tilde{g}_\Gamma$)**.** *We make the hypothesis in Theorem 6.7 that $\tilde{g}_\Gamma \in L^2(\Gamma)$, which is a rather strong assumption, that is fulfilled, e.g., when $\tilde{u}_\alpha \in H^{1+m}(\Omega_\alpha)$ for $m > \frac{1}{2}$, $\alpha \in \{\mathrm{p},\mathrm{n}\}$. This assumption enables us to manipulate $L^2(\Gamma)$-norms instead of fractional-order ones, which is particularly convenient from both the implementation and theoretical viewpoints. Let us point out that, in practice, this assumption does not seem necessary for our approach to be applicable. We will see in Section 7.3 that numerical convergence can still be observed, up to a slight adaptation of our method, in cases for which the assumption $\tilde{g}_\Gamma \in L^2(\Gamma)$ is violated. We finally point out that, at the time this manuscript is finalized, another related approach (based on optimal control) has been introduced in [23], which remedies this limitation. The key idea therein is the use of a bulk-supported control instead of a boundary-supported one.*

Remark 6.10 (Extension of the approach)**.** *Our approach is not restricted to the configurations or the boundary conditions for Problem (4) considered in Section 3. Under the only assumption on Problem (4) that the Dirichlet part of $\partial\Omega$ has nonzero $(d-1)$-dimensional measure, one can actually consider arbitrary (nonhomogeneous) boundary conditions, and relax the connectedness assumption on the two subdomains. Then, each connected part of a subdomain sharing a Dirichlet boundary with $\partial\Omega$ is treated as an $\mathtt{M}$ domain (cf. Figure 1), whereas every other connected part is treated as an $\mathtt{N}$ domain, for which the constant is fixed on a part of its boundary that is shared with an $\mathtt{M}$ domain (or which can be linked to one). For general geometries and boundary conditions, the method can be adapted and the analysis extended using the general approximation properties derived in Appendix B.*

7 Numerical results

For all the test-cases collected in this section, Problem (4) will be set in a polygonal domain $\Omega \subset \mathbb{R}^2$, and we will consider an isotropic coefficient $\mathfrak{o} := \sigma \mathbb{1}_2$, with $\sigma_\alpha = \sigma_{|\Omega_\alpha}$ a positive (real) constant for $\alpha \in \{\mathrm{p},\mathrm{n}\}$. In this case, the contrast (3) at the interface is simply $\nu = -\frac{\sigma_\mathrm{n}}{\sigma_\mathrm{p}}$.

7.1 Algebraic realization

We start by describing a possible algebraic realization for Problems (21) and (22). We let N_α, $\alpha \in \{\mathrm{p},\mathrm{n}\}$, be the dimension of the discrete space $U^k_{\alpha,h}$, and N_Γ be the dimension of $G^k_{\Gamma,h}$. For $\alpha \in \{\mathrm{p},\mathrm{n}\}$, we denote by $\mathbb{K}^{\alpha,\alpha}_h$ (size $N_\alpha \times N_\alpha$) the stiffness matrix in $U^k_{\alpha,h}$, written in the basis $(\psi^i_{\alpha,h})_{1\leqslant i\leqslant N_\alpha}$ of $U^k_{\alpha,h}$, and by $\mathbb{M}^{\Gamma,\Gamma}_h$ (size $N_\Gamma \times N_\Gamma$) the mass matrix in $G^k_{\Gamma,h}$, expressed in the basis $(\phi^j_{\Gamma,h})_{1\leqslant j\leqslant N_\Gamma}$ of $G^k_{\Gamma,h}$. We also let $\mathbb{T}^{\Gamma,\alpha}_h$ (size $N_\Gamma \times N_\alpha$) be the matrix representation of $\gamma_\alpha(U^k_{\alpha,h})_{|\Gamma}$, expressed in the basis $(\phi^j_{\Gamma,h})_{1\leqslant j\leqslant N_\Gamma}$ (remark that $\gamma_\alpha(U^k_{\alpha,h})_{|\Gamma} \subset G^k_{\Gamma,h}$).

To solve Problem (21), the first step is to compute, for $\alpha \in \{\mathrm{p},\mathrm{n}\}$, the solutions to Problem (19) for all the basis functions of $G^k_{\Gamma,h}$. In practice, we solve the $(N_\Gamma+1)$ following symmetric positive-definite (SPD) linear systems, of size $N_\alpha \times N_\alpha$:

$$\mathbb{K}_h^{\alpha,\alpha}\left(\mathfrak{u}_h^{\alpha,\Gamma} \;\middle|\; \mathbf{u}_h^\alpha\right) = \left([\mathbb{T}_h^{\Gamma,\alpha}]^\top \mathbb{M}_h^{\Gamma,\Gamma} \;\middle|\; s_\alpha \mathbf{F}_h^\alpha\right), \tag{37}$$

where $\mathbf{F}_h^\alpha \in \mathbb{R}^{N_\alpha}$ has i-th coordinate $(f,\psi^i_{\alpha,h})_{\Omega_\alpha}$. Then, for $\mathbf{g}_h^\Gamma \in \mathbb{R}^{N_\Gamma}$, the vector $\mathfrak{u}_h^\alpha(\mathbf{g}_h^\Gamma) \in \mathbb{R}^{N_\alpha}$ solution to Problem (19) in Ω_α is given by

$$\mathfrak{u}_h^\alpha(\mathbf{g}_h^\Gamma) = \mathbf{u}_h^\alpha + \mathfrak{u}_h^{\alpha,\Gamma}\mathbf{g}_h^\Gamma. \tag{38}$$

Solving Problem (21) is equivalent to solving $\inf_{\mathbf{g}_h^\Gamma \in \mathbb{R}^{N_\Gamma}} \mathcal{J}_h(\mathbf{g}_h^\Gamma)$, where the quadratic functional $\mathcal{J}_h : \mathbb{R}^{N_\Gamma} \to [0,\infty)$ is given by

$$\begin{aligned}\mathcal{J}_h(\mathbf{g}_h^\Gamma) := \left(\mathbb{T}_h^{\Gamma,\mathrm{p}}\mathfrak{u}_h^\mathrm{p}(\mathbf{g}_h^\Gamma) - \mathbb{T}_h^{\Gamma,\mathrm{n}}\mathfrak{u}_h^\mathrm{n}(\mathbf{g}_h^\Gamma)\right)^\top \mathbb{M}_h^{\Gamma,\Gamma}\left(\mathbb{T}_h^{\Gamma,\mathrm{p}}\mathfrak{u}_h^\mathrm{p}(\mathbf{g}_h^\Gamma) - \mathbb{T}_h^{\Gamma,\mathrm{n}}\mathfrak{u}_h^\mathrm{n}(\mathbf{g}_h^\Gamma)\right)\\ + \lambda(h)\sigma_\flat^{-2}\,(\mathbf{g}_h^\Gamma)^\top \mathbb{M}_h^{\Gamma,\Gamma}\mathbf{g}_h^\Gamma.\end{aligned} \tag{39}$$

One can easily compute

$$\boldsymbol{\nabla}^2\mathcal{J}_h = 2\left(\mathbb{T}_h^{\Gamma,\mathrm{p}}\mathfrak{u}_h^{\mathrm{p},\Gamma} - \mathbb{T}_h^{\Gamma,\mathrm{n}}\mathfrak{u}_h^{\mathrm{n},\Gamma}\right)^\top \mathbb{M}_h^{\Gamma,\Gamma}\left(\mathbb{T}_h^{\Gamma,\mathrm{p}}\mathfrak{u}_h^{\mathrm{p},\Gamma} - \mathbb{T}_h^{\Gamma,\mathrm{n}}\mathfrak{u}_h^{\mathrm{n},\Gamma}\right) + 2\lambda(h)\sigma_\flat^{-2}\,\mathbb{M}_h^{\Gamma,\Gamma},$$

so that, since $\mathcal{J}_h$ is quadratic, $\mathcal{J}_h(\mathbf{g}_h^\Gamma) = \frac{1}{2}(\mathbf{g}_h^\Gamma)^\top\left[\boldsymbol{\nabla}^2\mathcal{J}_h\right]\mathbf{g}_h^\Gamma + (\mathbf{g}_h^\Gamma)^\top\mathbf{V}_h^\Gamma + C$, where $\mathbf{V}_h^\Gamma \in \mathbb{R}^{N_\Gamma}$ and $C \in \mathbb{R}$ are inferred from (39). Writing the first-order necessary condition of optimality, solving Problem (21) finally consists in solving the SPD linear system, of size $N_\Gamma \times N_\Gamma$:

$$\left[\boldsymbol{\nabla}^2\mathcal{J}_h\right]\tilde{\mathbf{g}}_h^\Gamma = \mathbf{V}_h^\Gamma. \tag{40}$$

The seeked solution is finally $\mathfrak{u}_h^\alpha(\tilde{\mathbf{g}}_h^\Gamma)$, $\alpha \in \{\mathrm{p},\mathrm{n}\}$, as given by (38).

To solve Problem (22) (inclusion case), the first step is also to compute, for $\alpha \in \{\mathrm{p},\mathrm{n}\}$, the solutions to Problem (19) for all the basis functions of $G^k_{\Gamma,h}$. In Ω_p, one solves the $(N_\Gamma+1)$ SPD linear systems (37), of size $N_\mathrm{p}\times N_\mathrm{p}$. In Ω_n, the problems are of pure Neumann type. Let us first introduce some notations. Let $\mathbf{1}_h^\mathrm{n} \in \mathbb{R}^{N_\mathrm{n}}$ be the vector such that $\sum_{i=1}^{N_\mathrm{n}}[\mathbf{1}_h^\mathrm{n}]_i\psi^i_{\mathrm{n},h} \equiv 1$ in Ω_n. In turn, let $\mathbf{1}_h^\Gamma \in \mathbb{R}^{N_\Gamma}$ be the vector such that $\sum_{j=1}^{N_\Gamma}[\mathbf{1}_h^\Gamma]_j\phi^j_{\Gamma,h} \equiv 1$ on Γ. There holds $\mathbb{T}_h^{\Gamma,\mathrm{n}}\mathbf{1}_h^\mathrm{n} = \mathbf{1}_h^\Gamma$. Define also

$$\hat{\mathbb{M}}_h^{\Gamma,\Gamma} := |\Gamma|_{d-1}^{-1}\left([\mathbf{1}_h^\Gamma]^\top\mathbb{M}_h^{\Gamma,\Gamma}\right)^\top\left([\mathbf{1}_h^\Gamma]^\top\mathbb{M}_h^{\Gamma,\Gamma}\right).$$

In Ω_n, one solves the following $(N_\Gamma+1)$ SPD linear systems, of size $N_\mathrm{n}\times N_\mathrm{n}$:

$$\left(\mathbb{K}_h^{\mathrm{n},\mathrm{n}} + [\mathbb{T}_h^{\Gamma,\mathrm{n}}]^\top\hat{\mathbb{M}}_h^{\Gamma,\Gamma}\mathbb{T}_h^{\Gamma,\mathrm{n}}\right)\left(\mathfrak{u}_h^{\mathrm{n},\Gamma} \;\middle|\; \mathbf{u}_h^\mathrm{n}\right) = \left([\mathbb{T}_h^{\Gamma,\mathrm{n}}]^\top\left(\mathbb{M}_h^{\Gamma,\Gamma} - \hat{\mathbb{M}}_h^{\Gamma,\Gamma}\right) \;\middle|\; -\hat{\mathbf{F}}_h^\mathrm{n}\right), \tag{41}$$

where $\hat{\mathbf{F}}_h^\mathrm{n} \in \mathbb{R}^{N_\mathrm{n}}$ has i-th coordinate $(f,\psi^i_{\mathrm{n},h})_{\Omega_\mathrm{n}} - |\Gamma|_{d-1}^{-1}(f,1)_{\Omega_\mathrm{n}}\big(1,\gamma_\mathrm{n}(\psi^i_{\mathrm{n},h})\big)_\Gamma$. Remark that

$$[\mathbf{1}_h^\mathrm{n}]^\top[\mathbb{T}_h^{\Gamma,\mathrm{n}}]^\top\left(\mathbb{M}_h^{\Gamma,\Gamma} - \hat{\mathbb{M}}_h^{\Gamma,\Gamma}\right) = [\mathbf{0}_h^\Gamma]^\top \qquad \text{and} \qquad [\mathbf{1}_h^\mathrm{n}]^\top\hat{\mathbf{F}}_h^\mathrm{n} = 0.$$

Hence, the discrete solutions corresponding to $\mathfrak{v}_h^{\mathrm{n},\Gamma}$ and $\mathbf{u}_h^{\mathrm{n}}$ from (41) have zero mean over Γ. We amend a posteriori their expressions in the following way:

$$\mathfrak{v}_h^{\mathrm{n},\Gamma} \leftarrow \mathfrak{v}_h^{\mathrm{n},\Gamma} + |\Gamma|_{d-1}^{-1}\mathbf{1}_h^{\mathrm{n}}\left([\mathbf{1}_h^\Gamma]^\top \mathbb{M}_h^{\Gamma,\Gamma}\mathbb{T}_h^{\Gamma,\mathrm{p}}\mathfrak{v}_h^{\mathrm{p},\Gamma}\right), \quad \mathbf{u}_h^{\mathrm{n}} \leftarrow \mathbf{u}_h^{\mathrm{n}} + |\Gamma|_{d-1}^{-1}\mathbf{1}_h^{\mathrm{n}}\left([\mathbf{1}_h^\Gamma]^\top \mathbb{M}_h^{\Gamma,\Gamma}\mathbb{T}_h^{\Gamma,\mathrm{p}}\mathbf{u}_h^{\mathrm{p}}\right).$$

For $\mathbf{g}_h^\Gamma \in \mathbb{R}^{N_\Gamma}_{\mathtt{N}} := \left\{\mathbf{g}_h^\Gamma \in \mathbb{R}^{N_\Gamma} \mid [\mathbf{1}_h^\Gamma]^\top \mathbb{M}_h^{\Gamma,\Gamma}\mathbf{g}_h^\Gamma = (f,1)_{\Omega_\mathrm{n}}\right\}$, and $\alpha \in \{\mathrm{p},\mathrm{n}\}$, the vector $\mathfrak{u}_h^\alpha(\mathbf{g}_h^\Gamma) \in \mathbb{R}^{N_\alpha}$ solution to Problem (19) in Ω_α is finally given by

$$\mathfrak{u}_h^\alpha(\mathbf{g}_h^\Gamma) = \mathbf{u}_h^\alpha + \mathfrak{v}_h^{\alpha,\Gamma}\mathbf{g}_h^\Gamma. \tag{42}$$

Solving Problem (22) is equivalent to solving $\inf_{\mathbf{g}_h^\Gamma \in \mathbb{R}^{N_\Gamma}_{\mathtt{N}}} \mathcal{J}_h(\mathbf{g}_h^\Gamma)$, i.e. it consists in solving the well-posed saddle-point problem of size $(N_\Gamma+1)\times(N_\Gamma+1)$: find $(\tilde{\mathbf{g}}_h^\Gamma, \ell_\Gamma) \in \mathbb{R}^{N_\Gamma}\times\mathbb{R}$ such that

$$\begin{pmatrix} \boldsymbol{\nabla}^2\mathcal{J}_h & \mathbb{M}_h^{\Gamma,\Gamma}\mathbf{1}_h^\Gamma \\ [\mathbf{1}_h^\Gamma]^\top\mathbb{M}_h^{\Gamma,\Gamma} & 0 \end{pmatrix}\begin{pmatrix} \tilde{\mathbf{g}}_h^\Gamma \\ \ell_\Gamma \end{pmatrix} = \begin{pmatrix} \mathbf{V}_h^{\Gamma,\mathtt{N}} \\ (f,1)_{\Omega_\mathrm{n}} \end{pmatrix}, \tag{43}$$

where $\mathbf{V}_h^{\Gamma,\mathtt{N}} \in \mathbb{R}^{N_\Gamma}$. The seeked solution is finally $\mathfrak{u}_h^\alpha(\tilde{\mathbf{g}}_h^\Gamma)$, $\alpha \in \{\mathrm{p},\mathrm{n}\}$, as given by (42).

Remark 7.1 (Efficient implementation)**.** *Let us focus on Problem (21); similar considerations apply to Problem (22). The bottleneck in the solution to (21) is actually the solution to (37) for $\alpha \in \{\mathrm{p},\mathrm{n}\}$. As a matter of fact, once the hessian of $\mathcal{J}_h$ is computed (based on the solutions to (37)), solving the minimization problem then amounts to solving the small linear system (40). As standard in domain decomposition, Problem (37) can be solved in parallel in the two subdomains Ω_p and Ω_n. Each subproblem consists in solving a multi-rhs linear system, for which efficient solution techniques exist (Cholesky factorization for a direct solution, or Krylov subspace recycling for an iterative one).*

7.2 Test-case 1: nonsymmetric cavity with contrast -1

We consider the nonsymmetric cavity (cf. [21, Section 3.3]) with $\Omega := (-1,3)\times(0,1)$ and $\Gamma := \{0\}\times(0,1)$, so that $\Omega_\mathrm{p} = (-1,0)\times(0,1)$ and $\Omega_\mathrm{n} = (0,3)\times(0,1)$. This configuration is of type `2M`. We let $\sigma_\mathrm{p} = \sigma_\mathrm{n} = 1$, so that $\nu = -1$ (super-critical case). With such choices, the operator $\mathtt{A} \in \mathcal{L}(H_0^1(\Omega), H^{-1}(\Omega))$ from Problem (5) is injective, but not Fredholm (cf. Section 4.2). Since the operator is not Fredholm, the problem cannot be studied with `T`-coercivity theory, nor approximated using meshing rules inferred from the latter. However, our approach is applicable, and so as soon as the solution exists (it is then unique) for a given loading.

Let us consider the exact solution $\tilde{u} \in H_0^1(\Omega)$ defined by

$$\tilde{u}(x,y) := \begin{cases} \Big(2(x+1)^2 - 5(x+1)\Big)\sin(\pi y) & \text{in } \Omega_\mathrm{p}, \\ (x-3)\sin(\pi y) & \text{in } \Omega_\mathrm{n}, \end{cases}$$

which is associated to the loading $f \in L^2(\Omega)$ such that

$$f(x,y) = \begin{cases} \Big(2\pi^2(x+1)^2 - 5\pi^2(x+1) - 4\Big)\sin(\pi y) & \text{in } \Omega_\mathrm{p}, \\ -\pi^2(x-3)\sin(\pi y) & \text{in } \Omega_\mathrm{n}. \end{cases}$$

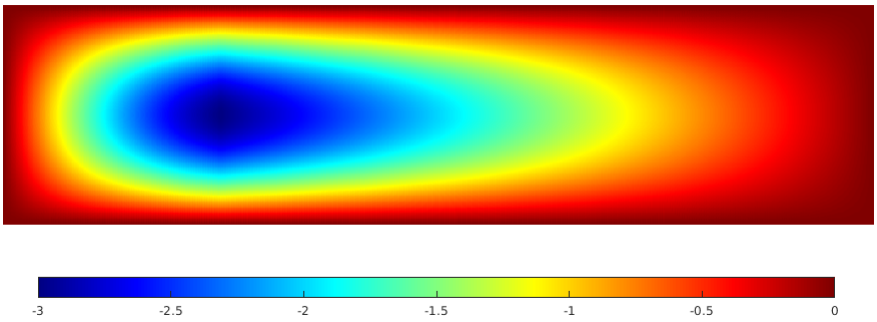
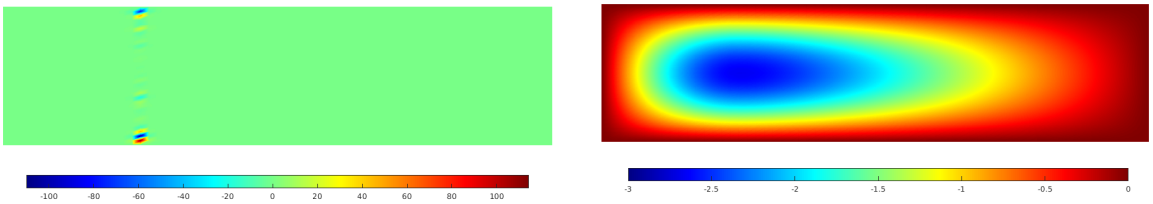


Figure 2: Test-case 1: exact solution for the nonsymmetric cavity with contrast -1

The exact solution is depicted on Figure 2. We have $\tilde{g}_\Gamma(y) = -\sin(\pi y)$ on Γ, where $\tilde{g}_\Gamma$ is defined by (14)–(15). For $\alpha \in \{\mathrm{p}, \mathrm{n}\}$, $\tilde{u}_\alpha \in H^{1+m}(\Omega_\alpha)$ for any $m \geqslant 0$. For the geometry considered here, full elliptic regularity holds true in both subdomains (cf. Remark B.2), meaning that the dual regularity exponents ε_p and ε_n are both equal to 1. Hence, the value of the parameter $\tilde{\delta}$ in Theorem 6.7 is $\tilde{\delta} = 2k + 1$.

We consider a structured triangulation of the domain Ω, with meshsize $h = 0.07$, which is admissible in the sense of Definition 6.1 (it is compliant with Γ), and we compare, for $k = 1$ (hence $\tilde{\delta} = 3$), the discrete solutions obtained with our approach (for $\lambda(h) = 0.01\, h^{2.9}$), and with a direct (non-stabilized) conforming finite element (cFE) approximation of the problem. Snapshots of the solutions are depicted on Figure 3. Whereas our approach provides a somewhat accurate solution (the relative error in L^2-norm is of 5.33×10^{-2}), which converges monotonically to the exact one as the mesh is refined (not shown here), the cFE solution exhibits very large spurious oscillations near the interface. This is a striking example of how unstable can be a non-stabilized method for such an ill-posed problem.



(a) cFE solution (rescaled by a factor 10^{13})

(b) Optimization-based solution

Figure 3: Test-case 1: discrete solutions for the nonsymmetric cavity with contrast -1

7.3 Test-case 2: low-regularity solution

We consider (a slight variant of) the test-case studied in [5, Section 3]. We let Ω be the hexagonal domain of Figure 4 (left), for which $\Gamma = \left\{(r, \theta) \in \Omega \mid \theta = 0 \text{ or } \theta = \frac{4\pi}{3}\right\}$, and Ω_p, Ω_n are respectively the top and bottom subdomains. The corresponding configuration is of type `2M`. We let $\sigma_\mathrm{p} = 1$, and we tune σ_n so as to change the value of the contrast ν. For such a configuration, Problem (5) is well-posed in the Hadamard sense if and only if $\nu \notin \left[-2, -\frac{1}{2}\right]$.

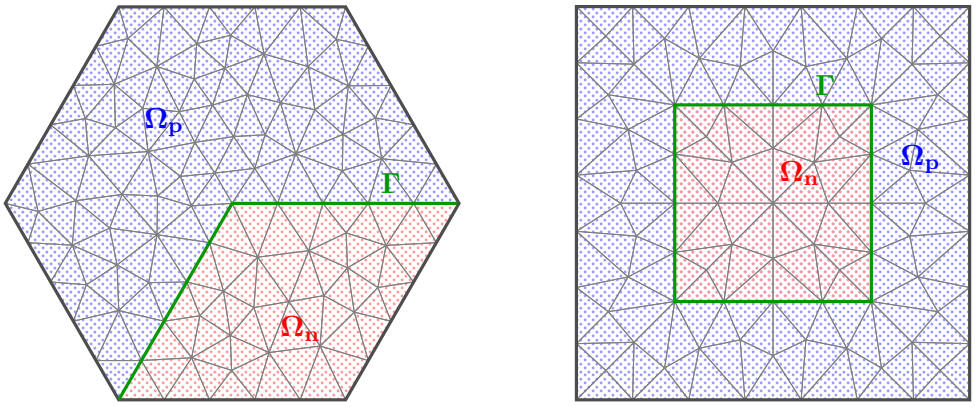


Figure 4: Test-cases 2 (left) and 3 (right): geometry and mesh family member

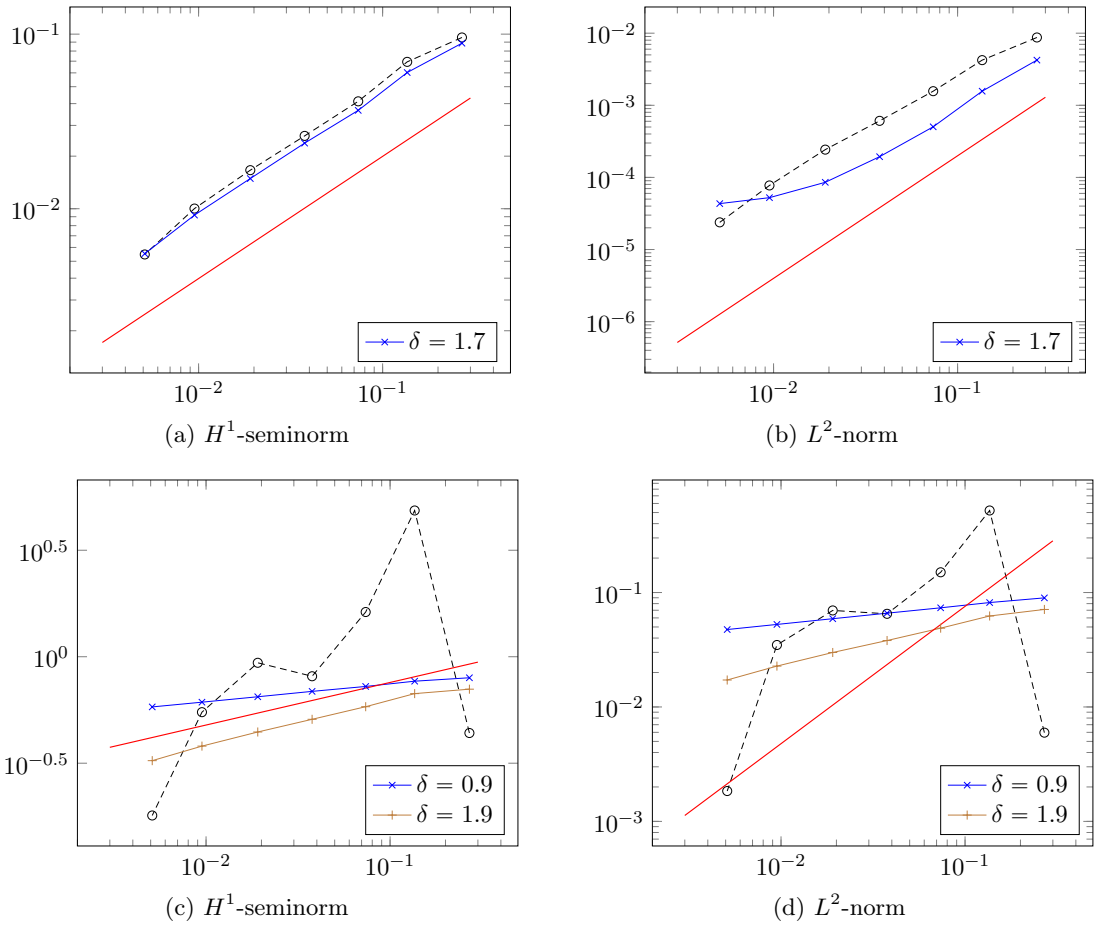
We consider the exact solution $\tilde{u} \in H^1(\Omega)$ defined by $\tilde{u}(r,\theta) := r^\kappa \Phi(\theta)$, where

$$\Phi(\theta) := \begin{cases} \dfrac{\cos\left(\kappa(\theta - \frac{2\pi}{3})\right)}{\cos\left(\kappa\frac{2\pi}{3}\right)} & \text{in } \Omega_{\rm p}, \\ \dfrac{\cos\left(\kappa(\theta - \frac{5\pi}{3})\right)}{\cos\left(\kappa\frac{\pi}{3}\right)} & \text{in } \Omega_{\rm n}, \end{cases}$$

and where $\kappa > 0$ depends on ν in the following way: κ is the smallest positive (real) solution to $\tan\left(\kappa\frac{2\pi}{3}\right) = -\nu \tan\left(\kappa\frac{\pi}{3}\right)$. The solution $\tilde{u}$ is associated to the loading $f \equiv 0$, and to the nonhomogeneous Dirichlet boundary data $\gamma(\tilde{u})$ on $\partial\Omega$. Besides, we have $\tilde{g}_\Gamma(r) = -\kappa\, r^{\kappa-1} \tan\left(\kappa\frac{2\pi}{3}\right)$ on Γ. The following regularity result holds true: for $\alpha \in \{\mathrm{p}, \mathrm{n}\}$, $\tilde{u}_\alpha \in H^{1+m}(\Omega_\alpha)$ for all $m < \kappa$. We are going to consider two values of ν outside of the critical interval, namely $\nu \in \{-10.57, -2.1\}$, for which the parameter κ is respectively such that $\kappa \gtrapprox \{0.7, 0.2\}$. When $\kappa \approx 0.7$, then we have $\tilde{g}_\Gamma \in L^2(\Gamma)$, and the assumptions of Theorem 6.7 are fulfilled. At the opposite, when $\kappa \approx 0.2$, the result of Theorem 6.7 is not valid ($\tilde{g}_\Gamma \notin L^2(\Gamma)$). For the geometry considered here, full elliptic regularity holds true in $\Omega_{\rm n}$, whereas the reentrant corner in $\Omega_{\rm p}$ induces a loss of regularity of 1/4 (cf. Remark B.2). We thus have $\varepsilon_{\rm n} = 1$ and $\frac{3}{4} - \epsilon < \varepsilon_{\rm p} < \frac{3}{4}$ for all $\epsilon > 0$. Since the subproblems in $\Omega_{\rm p}$ and $\Omega_{\rm n}$ feature nonhomogeneous (Dirichlet) boundary conditions on the part of their boundary which is shared with $\partial\Omega$, one has to use the result of Lemma B.9 in both $\Omega_{\rm p}$ and $\Omega_{\rm n}$ (and then take the min) to infer the value of the parameter $\tilde{\delta}$ from Theorem 6.7: a straightforward computation yields $2\kappa + \frac{1}{2} - \epsilon < \tilde{\delta} < 2\kappa + \frac{1}{2}$ for all $\epsilon > 0$, which is valid for any integer $k \geqslant 1$ as soon as $\frac{1}{2} < \kappa \leqslant 1$.

We consider a family of unstructured triangulations (see Figure 4 (left)) of the domain Ω, that is admissible in the sense of Definition 6.1, but which is not T-conforming, as opposed to the mesh family from [5, Figure 5 (right)]. For the latter mesh family, T-coercivity theory enables to prove the optimal convergence of cFE. At the opposite, in our T-nonconforming case, no theory applies. We compare, for $k = 1$, the results obtained with our approach and with cFE. We compute, for meshsizes between $h = 0.27$ and $h = 0.0051$, the relative errors over Ω in H^1-seminorm and in L^2-norm, for $\nu = -10.57$ and $\nu = -2.1$. In the following, the convergence rate of the error that one can expect with a T-conforming approximation will be referred to as the expected convergence rate. According to Proposition 4.8, the expected convergence rate in H^1-seminorm is h^κ. In L^2-norm, the expected convergence rate depends on the dual

regularity exponent of the sign-changing problem (see [21, Section 3.4] for some insight on the question), as well as on the regularity of the boundary data. We do not have a theoretical value in the present case, but we conjecture (based on an application of Lemma B.8 in each subdomain) $h^{\kappa+\frac{1}{2}}$ convergence. For $\nu = -10.57$ ($\kappa \approx 0.7$), we choose $\lambda(h) = 0.01\, h^{1.7}$ (remark that $\delta = 1.7 < 1.9 < 2\kappa + \frac{1}{2}$), whereas for $\nu = -2.1$ ($\kappa \approx 0.2$) we test two different possibilities. First, we choose the stabilization following the rationale of Theorem 6.7, even though the latter is not applicable in this case; we let $\lambda(h) = 0.01\, h^{0.9}$ (remark that $\delta = 0.9 < 2\kappa + \frac{1}{2}$). Second, we choose $\lambda(h) = 0.01\, h^{1.9}$, i.e. we decrease the magnitude of the stabilization. The heuristics behind this choice is elementary: since $\tilde{g}_\Gamma \notin L^2(\Gamma)$ in this case, we rescale the stabilization so as to formally embed an $H^{-\frac{1}{2}}(\Gamma)$-norm of $\tilde{g}_\Gamma$, i.e. we multiply the original stabilization by the square of $h^{\frac{1}{2}}$, yielding $\delta = 1.9$. The results are collected in Figure 5. For $\nu = -10.57$ (top), for both approaches, we observe the expected convergence rate in H^1-seminorm. For $\nu = -2.1$ (bottom), cFE presents a completely erratic behavior. At the opposite, our approach provides monotonic convergence. When the regularization exponent is fixed to $\delta = 0.9$, the method sub-converges, whereas for $\delta = 1.9$, the expected convergence rate is reached in H^1-seminorm.



(a) H^1-seminorm (b) L^2-norm

(c) H^1-seminorm (d) L^2-norm

Figure 5: Test-case 2: relative errors vs. h for $\nu = -10.57$ (top) and $\nu = -2.1$ (bottom). *Dotted black: cFE, Solid blue/brown: optimization-based, Solid red: h^κ for H^1, $h^{\kappa+1}$ for L^2.*

7.4 Test-case 3: inclusion

We consider an inclusion test-case with $\Omega := (-2,2) \times (-2,2)$ and $\Omega_{\mathrm{n}} := (-1,1) \times (-1,1)$; cf. Figure 4 (right). This configuration is of type MN. We let $\sigma_{\mathrm{p}} = 1$, and we tune σ_{n} so as to change the value of the contrast ν. For such a setting, Problem (5) is well-posed in the Fredholm sense if and only if $\nu \notin \left[-3, -\frac{1}{3}\right]$ (see [5, Theorem 1]).

Let us consider the exact solution $\tilde{u} \in H^1_0(\Omega)$ defined by

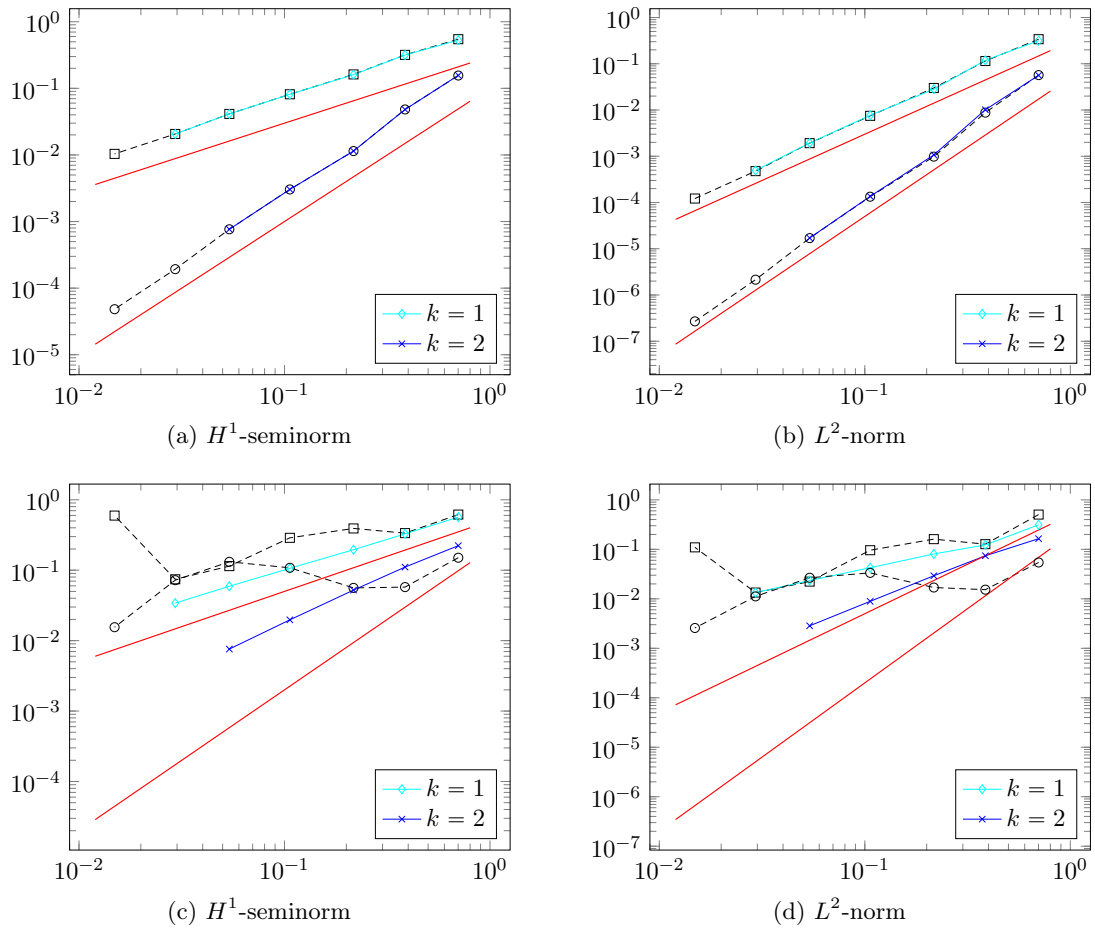
$$\tilde{u}(x,y) := \begin{cases} \sin(\pi x)\sin(\pi y) & \text{in } \Omega_{\mathrm{p}}, \\ \nu^{-1}\sin(\pi x)\sin(\pi y) & \text{in } \Omega_{\mathrm{n}}, \end{cases}$$

which is associated to the loading $f \in L^2(\Omega)$ such that $f(x,y) = 2\pi^2 \sin(\pi x)\sin(\pi y)$ in Ω. We have $\tilde{g}_\Gamma(x,y) = \pi\left(\sin(\pi x)\left(1_{y=1} - 1_{y=-1}\right) + \sin(\pi y)\left(1_{x=1} - 1_{x=-1}\right)\right)$ on Γ. For $\alpha \in \{\mathrm{p}, \mathrm{n}\}$, $\tilde{u}_\alpha \in H^{1+m}(\Omega_\alpha)$ for any $m \geqslant 0$. For the geometry considered here, full elliptic regularity holds true in Ω_{n}, but the reentrant corners in Ω_{p} induce a loss of regularity of 1/3 (cf. Remark B.2). The dual regularity exponents are thus such that $\varepsilon_{\mathrm{n}} = 1$ and $\frac{2}{3} - \epsilon < \varepsilon_{\mathrm{p}} < \frac{2}{3}$ for all $\epsilon > 0$. The parameter $\tilde{\delta}$ from Theorem 6.7 is, in turn, such that $2k + \frac{2}{3} - \epsilon < \tilde{\delta} < 2k + \frac{2}{3}$ for all $\epsilon > 0$. We consider two values of ν, one outside of the critical interval ($\nu = -4$), and the super-critical value ($\nu = -1$). For the first value of ν, we know that the operator $\mathtt{A} \in \mathcal{L}(H^1_0(\Omega), H^{-1}(\Omega))$ from Problem (5) is Fredholm (of index 0), whereas for the second we know that it is not (and hence T-coercivity is not applicable). For both values of ν, we assume in the following that the operator A is injective. Numerically, we have not found any evidence of non-uniqueness.

We consider the family of unstructured triangulations of Ω depicted on Figure 4 (right). This family is admissible in the sense of Definition 6.1, but is not (locally) T-conforming. For $\nu = -4$, we compare, for $k = 1$ and $k = 2$, the results obtained with our approach and with cFE. We compute the relative errors over Ω in H^1-seminorm and L^2-norm, for meshsizes between $h = 0.70$ and $h = 0.015$. In H^1-seminorm, according to Proposition 4.9, the expected convergence rate (i.e. relative to a locally T-conforming approximation, and for h small enough) is h^k. In L^2-norm, we do not have a theoretical value, but we conjecture (based on an application of Lemma B.8 in each subdomain) $h^{k+\frac{2}{3}}$ convergence. We choose $\lambda(h) = 0.01\, h^{2k+\frac{1}{2}}$, and we check that $\delta = 2k + \frac{1}{2} < 2k + \frac{2}{3}$. For $\nu = -1$, we perform the same comparisons. However, in this case, no theoretical convergence rate is available, even in H^1-seminorm. All the results are collected in Figure 6. For $\nu = -4$ (top), we remark that cFE and our approach give very similar results; the expected convergence rates are reached in H^1-seminorm. In L^2-norm, both approaches seem to converge with order $k+1$ (higher than expected). For $\nu = -1$ (bottom), we remark that cFE suffers, whereas our approach provides monotonic convergence in both H^1-seminorm and L^2-norm. The convergence orders are difficult to analyze. On Figure 7, we have depicted the discrete solutions obtained for $k = 2$ and $h = 0.054$. We observe spurious oscillations at the interface between the two subdomains for cFE, whereas our approach provides an oscillation-robust solution (the relative error in H^1-seminorm is more than 10 times smaller).

A Background on Fredholm theory

We collect in this appendix some classical definitions and results. We provide short proofs for the most important of them.



(a) H^1-seminorm (b) L^2-norm

(c) H^1-seminorm (d) L^2-norm

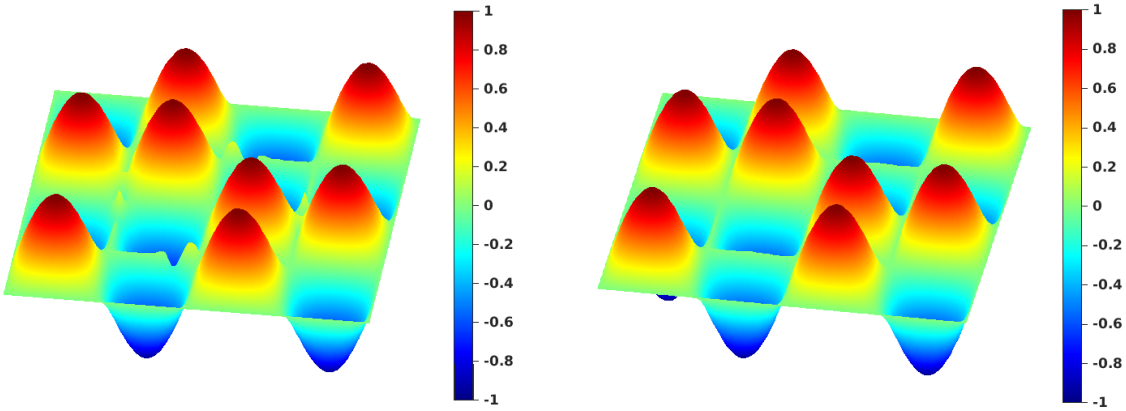
Figure 6: Test-case 3: relative errors vs. h for $\nu = -4$ (top) and $\nu = -1$ (bottom). *Dotted black: cFE (squares for $k = 1$, circles for $k = 2$), Solid cyan/blue: optimization-based, Solid red: h^k for H^1, h^{k+1} for L^2.*

For V, W real-valued Banach spaces, we let $\mathcal{L}(V, W)$ be the space of bounded linear operators from V to W. When $W = V$, we simply write $\mathcal{L}(V)$. Let U be a real-valued reflexive Banach space (e.g. a real-valued Hilbert space), with topological dual $U^\star := \mathcal{L}(U, \mathbb{R})$, and duality pairing $\langle \cdot, \cdot \rangle$. Since U is reflexive, there exists a natural (isometric) isomorphism between U and $U^{\star\star}$, and one can identify U with its double dual. Let us recall some definitions.

Definition A.1 (Adjoint operator)**.** *Let $B \in \mathcal{L}(U, U^\star)$. The* adjoint $B^\star$ *of the operator B is the unique operator in $\mathcal{L}(U, U^\star)$ such that, for all $u, v \in U$, $\langle B^\star(u), v \rangle = \langle B(v), u \rangle$. When $B^\star = B$, the operator B is said to be* self-adjoint*.*

In what follows, for $V \subset W$, we denote by W/V the quotient of the vector space W by the subspace V.

Definition A.2 (Fredholm operator [2, Definition 4.37])**.** *The operator $B \in \mathcal{L}(U, U^\star)$ is said to be* Fredholm *if its nullity* $\dim(\mathrm{Ker} B)$ *and defect* $\dim(U^\star/\mathrm{Im} B)$ *are both finite. Its* index *is then defined as* $\mathrm{ind}(B) := \dim(\mathrm{Ker} B) - \dim(U^\star/\mathrm{Im} B)$*.*



(a) cFE solution (b) Optimization-based solution

Figure 7: Test-case 3: discrete solutions for $\nu = -1$

As a by-product of Definition A.2, any Fredholm operator $\mathtt{B} \in \mathcal{L}(U, U^\star)$ of index 0 that is injective is also surjective and is an isomorphism from U to $U^\star$.

The following lemma holds true.

Lemma A.3 ([2, Lemma 4.38])**.** *Let $\mathtt{B} \in \mathcal{L}(U, U^\star)$ be such that its defect* $\dim(U^\star/\mathrm{Im}\mathtt{B})$ *is finite. Then,* $\mathrm{Im}\mathtt{B}$ *is closed in $U^\star$ and one has* $\dim(U^\star/\mathrm{Im}\mathtt{B}) = \dim(\mathrm{Ker}\mathtt{B}^\star)$.

We can now state the main results.

Proposition A.4. *Let $\mathtt{B} \in \mathcal{L}(U, U^\star)$ be a self-adjoint Fredholm operator. Then,* $\mathrm{ind}(\mathtt{B}) = 0$ *and the following alternative holds true:*

- *either $\mathtt{B}$ is injective, then $\mathtt{B}$ is an isomorphism from U to $U^\star$;*
- *or, letting* $1 \leqslant n := \dim(\mathrm{Ker}\mathtt{B}) < \infty$, *and* $\mathrm{Ker}\mathtt{B} := \mathrm{Span}\{v_1, \dots, v_n\}$ *for functions $v_1, \dots, v_n \in U$, one has* $\mathrm{Im}\mathtt{B} = \{f \in U^\star \mid \langle f, v_k \rangle = 0 \ \forall k = 1, \dots, n\}$.

Proof. $\mathtt{B}$ being self-adjoint, $\mathtt{B}^\star = \mathtt{B}$. Since $\mathtt{B}$ is Fredholm, its defect is finite and, by Lemma A.3, $\dim(U^\star/\mathrm{Im}\mathtt{B}) = \dim(\mathrm{Ker}\mathtt{B}^\star) = \dim(\mathrm{Ker}\mathtt{B})$, which yields $\mathrm{ind}(\mathtt{B}) = 0$. Then, if $\mathtt{B}$ is injective, $\dim(U^\star/\mathrm{Im}\mathtt{B}) = \dim(\mathrm{Ker}\mathtt{B}) = 0$ and $\mathtt{B}$ is also surjective. In the opposite case, using the relation $(\mathrm{Im}\mathtt{B})^\perp = \mathrm{Ker}\mathtt{B}^\star$ (cf. e.g. [2, Theorem 2.13]), one has $\overline{\mathrm{Im}\mathtt{B}} = ((\mathrm{Im}\mathtt{B})^\perp)^\perp = (\mathrm{Ker}\mathtt{B}^\star)^\perp = (\mathrm{Ker}\mathtt{B})^\perp$ (recall that U is reflexive), thus since $\mathrm{Im}\mathtt{B}$ is closed (by Lemma A.3), there holds

$$\mathrm{Im}\mathtt{B} = \{f \in U^\star \mid \langle f, v \rangle = 0 \ \forall v \in \mathrm{Ker}\mathtt{B}\}\,.$$

The conclusion follows from the finiteness of the nullity of $\mathtt{B}$. □

Proposition A.5. *Let $\mathtt{B} \in \mathcal{L}(U, U^\star)$ be a self-adjoint injective operator. Then, the following equivalence holds true: (i) $\mathtt{B}$ is surjective $\Leftrightarrow$ (ii) $\mathrm{Im}\mathtt{B}$ is closed.*

Proof. $(i) \Rightarrow (ii)$ is trivially true by Lemma A.3 (remark that $\dim(U^\star/\mathrm{Im}\mathtt{B}) = 0$ since $\mathtt{B}$ is surjective). To prove $(ii) \Rightarrow (i)$, let us assume that $\mathrm{Im}\mathtt{B}$ is closed. Then, one can show that $\mathrm{Im}\mathtt{B} = \overline{\mathrm{Im}\mathtt{B}} = ((\mathrm{Im}\mathtt{B})^\perp)^\perp = (\mathrm{Ker}\mathtt{B}^\star)^\perp = (\mathrm{Ker}\mathtt{B})^\perp$, where we have used that $\mathtt{B}$ is self-adjoint in the last identity. Since $\mathtt{B}$ is injective, $(\mathrm{Ker}\mathtt{B})^\perp = U^\star$, hence $\mathrm{Im}\mathtt{B} = U^\star$ and $\mathtt{B}$ is surjective. □

B Error estimates

We collect in this appendix some error estimates for the finite element solution to variable diffusion problems, endowed with either mixed or purely Neumann nonhomogeneous boundary conditions. These results are meant to be applied to problems of type (19), set in the positive or negative subdomain of configurations Ω of type 2M or MN (cf. Figure 1). Such estimates are instrumental to study the convergence of our method, and to finely tune the parameter $\lambda(h)$ in (20); cf. Sections 6.3 and 7. Should some of these results be quite classical we detail their proofs, both to keep the exposition as self-contained as possible, and because we aim at deriving the sharpest possible bounds.

B.1 Continuous and discrete settings

Following Section 2, let $\mathcal{D}$ be a (Lipschitz) domain in $\mathbb{R}^d$, $d \in \{2,3\}$, with boundary $\Upsilon := \partial\mathcal{D}$ such that $\Upsilon = \overline{\Upsilon_{\mathrm{t}}} \cup \overline{\Upsilon_{\mathrm{f}}}$ with $\Upsilon_{\mathrm{f}} \neq \varnothing$, and unit outer normal $\boldsymbol{n}$. We make the additional assumption that $\mathcal{D}$ is a polytope. Note that the boundary of $\mathcal{D}$ is not necessarily connected. Each (Lipschitz) subset Υ_{t} and Υ_{f} of the boundary Υ is assumed to be the finite union of $(d-1)$-dimensional polytopes. Note that Υ_{t} and Υ_{f} are not necessarily connected. In what follows, the set Υ_{t} is meant to be the trace/Dirichlet part of the boundary, whereas the set Υ_{f} is meant to be the flux/Neumann part.

Let $\mathbb{o} : \mathcal{D} \to \mathbb{R}^{d\times d}$ be a symmetric tensor field such that

$$0 < a_\flat|\boldsymbol{\xi}|^2 \leqslant \mathbb{o}(\boldsymbol{x})\boldsymbol{\xi}\cdot\boldsymbol{\xi} \leqslant a_\sharp|\boldsymbol{\xi}|^2 < \infty \qquad \text{for a.e. } \boldsymbol{x} \in \mathcal{D} \text{ and all } \boldsymbol{\xi} \in \mathbb{R}^d\backslash\{\boldsymbol{0}\},$$

and let $\varrho := a_\sharp/a_\flat \geqslant 1$ denote its heterogeneity/anisotropy ratio in $\mathcal{D}$. We further assume that $\mathbb{o} \in \mathbb{W}^{1,\infty}(\mathcal{D})$. Since $\mathcal{D}$ is a (Lipschitz) domain, thus $\mathcal{D}$ is quasiconvex and bounded, this is equivalent to assume that $\mathbb{o}$ is Lipschitz continuous in $\overline{\mathcal{D}}$. Also, since $\mathbb{o}$ is symmetric and uniformly elliptic, there is a unique symmetric and uniformly elliptic tensor field $\mathbb{o}^{1/2}$ such that $\mathbb{o} = \mathbb{o}^{1/2}\mathbb{o}^{1/2}$.

Let us consider, for $t \in L^2(\mathcal{D})$, the following problem (referred to as *dual* in the sequel), endowed with homogeneous boundary conditions:

$$\begin{cases} -\mathrm{div}(\mathbb{o}\boldsymbol{\nabla} z) = t & \text{in } \mathcal{D}, \\ z = 0 & \text{on } \Upsilon_{\mathrm{t}}, \\ \mathbb{o}\boldsymbol{\nabla} z\cdot\boldsymbol{n} = 0 & \text{on } \Upsilon_{\mathrm{f}}. \end{cases} \tag{44}$$

In the purely Neumann case $\Upsilon_{\mathrm{t}} = \varnothing$, we further assume that $(t,1)_{\mathcal{D}} = 0$, and we replace the condition $z = 0$ on Υ_{t} by $(z,1)_{\mathcal{D}} = 0$. Henceforth, we make the following assumption.

Assumption B.1 (Regularity of the dual solution)**.** *There is $\varepsilon \in (\frac{1}{2}, 1]$ (called* regularity exponent*), whose value may depend on the geometries of $\mathcal{D}$, Υ_{t} and Υ_{f}, and on $\mathbb{o}$, so that the solution z to Problem (44) belongs to $H^{1+\varepsilon}(\mathcal{D})$, and satisfies the following regularity estimate: there exists a constant $c_{\mathrm{r}} > 0$ such that*

$$\|z\|_{1+\varepsilon,\mathcal{D}} \leqslant c_{\mathrm{r}} a_\flat^{-1} \|t\|_{0,\mathcal{D}}. \tag{45}$$

Remark B.2 (Elliptic regularity)**.** *Let us discuss configurations for which the regularity assumption B.1 holds true. First, recall that $\mathbb{o}$ is Lipschitz continuous in $\overline{\mathcal{D}}$. Second, recall that Problem (44) is endowed with homogeneous boundary conditions.*

Let us begin by considering the purely Neumann case $\Upsilon_{\rm t} = \varnothing$. In that case, when the domain $\mathcal{D}$ is convex, Assumption B.1 holds true with regularity exponent $\varepsilon = 1$ (cf. [30, Theorem 3.2.1.3]). When $d = 2$ (so that $\mathcal{D}$ is a polygon), $\mathfrak{a} = a\,\mathbb{1}_2$ for $a > 0$, and the maximum angle ω in $\mathcal{D}$ is such that $\pi < \omega < 2\pi$ ($\mathcal{D}$ is not convex), Assumption B.1 holds true for all $\varepsilon < \varepsilon_0$ with $\varepsilon_0 = \frac{\pi}{\omega}$ (cf. [30, Theorem 4.4.3.7] and [4, Remark I.3.4]).

In the case of mixed Dirichlet-Neumann boundary conditions $\Upsilon_{\rm t} \neq \varnothing$, the situation is more complex. Here, we only state results in the case $\mathfrak{a} = a\,\mathbb{1}_d$ for $a > 0$. When $d = 3$, $\mathcal{D}$ is a rectangular cuboid, and $\Upsilon_{\rm f}$ is the union of (entire) faces of $\mathcal{D}$, Assumption B.1 holds true with regularity exponent $\varepsilon = 1$. When $d = 2$ (so that $\mathcal{D}$ is a polygon), (i) if $\mathcal{D}$ is convex, and the maximum angle $\omega_{\rm dn}$ between $\Upsilon_{\rm t}$ and $\Upsilon_{\rm f}$ is such that $\omega_{\rm dn} \leqslant \frac{\pi}{2}$, Assumption B.1 holds true with regularity exponent $\varepsilon = 1$; (ii) if one or both of the previous two assumptions is not satisfied, and if $\omega_{\rm dn} < \pi$, Assumption B.1 holds true for all $\varepsilon < \varepsilon_0$ with $\varepsilon_0 = \min\left(\frac{\pi}{\omega_{\rm d}}, \frac{\pi}{\omega_{\rm n}}, \frac{\pi}{2\omega_{\rm dn}}\right)$, where $\omega_{\rm d}$ and $\omega_{\rm n}$ are, respectively, the maximum angles in $\mathcal{D}$ internal to $\Upsilon_{\rm t}$ and to $\Upsilon_{\rm f}$ (cf. [30, Theorem 4.4.3.7] and [4, Remark I.3.6]).

Since we are going to consider finite element approximations, let us precise our definition of an admissible mesh family.

Definition B.3 (Admissible mesh family)**.** *A mesh family $(\mathfrak{T}_h)_{h>0}$ is* admissible *if (i) for all $h > 0$ in the family, $\mathfrak{T}_h$ is a matching simplicial discretization of $\mathcal{D}$ that is geometrically compliant with the partition of the boundary (in the sense that $\overline{\Upsilon_{\rm t}} = \bigcup \overline{F}$ and $\overline{\Upsilon_{\rm f}} = \bigcup \overline{F}$ with $\{F\}$ boundary faces of $\mathfrak{T}_h$), and if (ii) $(\mathfrak{T}_h)_{h>0}$ is shape-regular in the sense of Ciarlet [22].*

Let $\mathfrak{T}_h$ be a member of an admissible mesh family. For an integer $k \geqslant 1$, we introduce the discrete space

$$V_h^k := \left\{ v_h \in C^0(\overline{\mathcal{D}}) \mid v_{h|T} \in \mathbb{P}_d^k(T)\ \forall T \in \mathfrak{T}_h \right\} \subset H^1(\mathcal{D}).$$

The usual Lagrange interpolator from $C^0(\overline{\mathcal{D}})$ onto V_h^k is denoted $\mathcal{I}_h^{k,d}$, whereas $\mathcal{I}_h^{k,d-1}$ stands for the Lagrange interpolator (piecewise defined on each face of $\mathcal{D}$) from $C^0(\Upsilon)$ onto the space

$$\left\{ \varphi_h \in C^0(\Upsilon) \mid \varphi_{h|F} \in \mathbb{P}_{d-1}^k(F)\ \forall F \in \mathcal{F}_h^{\rm b} \right\},$$

where the set $\mathcal{F}_h^{\rm b}$ collects the boundary faces of the mesh $\mathfrak{T}_h$. It is an easy matter to verify that $\gamma \circ \mathcal{I}_h^{k,d} = \mathcal{I}_h^{k,d-1} \circ \gamma$ on $C^0(\overline{\mathcal{D}})$. In order to deal with mixed Dirichlet-Neumann boundary conditions ($\Upsilon_{\rm t} \neq \varnothing$), we will need the space $H^1_{0\backslash\Upsilon_{\rm f}}(\mathcal{D})$ defined in (2). In the purely Neumann case, we will instead consider the space

$$H^{1,0}(\mathcal{D}) := \left\{ v \in H^1(\mathcal{D}) \mid (v,1)_{\mathcal{D}} = 0 \right\}.$$

From a discrete viewpoint, we define $V_{0,h}^k := V_h^k \cap H^1_{0\backslash\Upsilon_{\rm f}}(\mathcal{D})$ and $V_h^{k,0} := V_h^k \cap H^{1,0}(\mathcal{D})$. We then let $\Pi_{0,h}^k : H^1_{0\backslash\Upsilon_{\rm f}}(\mathcal{D}) \to V_{0,h}^k$ and $\Pi_h^{k,0} : H^{1,0}(\mathcal{D}) \to V_h^{k,0}$ denote the respective $\mathfrak{a}$-weighted elliptic projections onto the previous discrete spaces, i.e. the orthogonal projectors for the inner product $(v,w) \mapsto (\mathfrak{a}\boldsymbol{\nabla} v, \boldsymbol{\nabla} w)_{\mathcal{D}}$. The following approximation result holds true.

Proposition B.4 (Approximation)**.** *Let V be either $H^1_{0\backslash\Upsilon_{\rm f}}(\mathcal{D})$ or $H^{1,0}(\mathcal{D})$ and, correspondingly, let V_h be either $V_{0,h}^k$ or $V_h^{k,0}$ and $\Pi_h : V \to V_h$ be either $\Pi_{0,h}^k$ or $\Pi_h^{k,0}$. Let $s \in [0,k]$. Then, there is $c_{\rm a} > 0$ such that, for all $v \in V$ satisfying $v \in H^{1+s}(\mathcal{D})$,*

$$\|\mathfrak{a}^{1/2}\boldsymbol{\nabla}(v - \Pi_h(v))\|_{0,\mathcal{D}} \leqslant c_{\rm a} a_\sharp^{1/2} h^s |v|_{1+s,\mathcal{D}}. \tag{46}$$

Proof. By definition of the $\mathfrak{d}$-weighted elliptic projection, there holds

$$\|\mathfrak{d}^{1/2}\boldsymbol{\nabla}(v-\Pi_h(v))\|_{0,\mathcal{D}} = \min_{v_h\in V_h}\|\mathfrak{d}^{1/2}\boldsymbol{\nabla}(v-v_h)\|_{0,\mathcal{D}}.$$

If $s=0$, choosing $v_h=0$ directly yields $\|\mathfrak{d}^{1/2}\boldsymbol{\nabla}(v-\Pi_h(v))\|_{0,\mathcal{D}} \leqslant a_\sharp^{1/2}|v|_{1,\mathcal{D}}$. Now, assume that $s\in(\frac{d}{2}-1,k]$. In that case, owing to the (continuous) embedding of $H^{1+s}(\mathcal{D})$ into $C^0(\overline{\mathcal{D}})$, one can resort to the Lagrange interpolate $\mathcal{I}_h^{k,d}(v)\in V_h^k$ of v. Indeed, if $v\in H^1_{0\backslash\Upsilon_\mathrm{f}}(\mathcal{D})$, $\mathcal{I}_h^{k,d}(v)\in V_{0,h}^k$. If $v\in H^{1,0}(\mathcal{D})$, $\mathcal{I}_h^{k,d}(v)\notin V_h^{k,0}$ a priori but

$$\min_{v_h\in V_h^{k,0}}\|\mathfrak{d}^{1/2}\boldsymbol{\nabla}(v-v_h)\|_{0,\mathcal{D}} = \min_{v_h\in V_h^{k}}\|\mathfrak{d}^{1/2}\boldsymbol{\nabla}(v-v_h)\|_{0,\mathcal{D}}. \tag{47}$$

In any case, one can thus write $\|\mathfrak{d}^{1/2}\boldsymbol{\nabla}(v-\Pi_h(v))\|_{0,\mathcal{D}}\leqslant a_\sharp^{1/2}\|\boldsymbol{\nabla}(v-\mathcal{I}_h^{k,d}(v))\|_{0,\mathcal{D}}$, and conclude invoking standard approximation results for $\mathcal{I}_h^{k,d}$ (see e.g. [27, Corollary 19.8]). When $d=2$, the proof is complete. When $d=3$, one still has to treat the case $s\in(0,\frac{1}{2}]$. Our proof makes use of the quasi-interpolation operator introduced in [26] (among other candidates). When $V=H^{1,0}(\mathcal{D})$, the conclusion follows from the trick (47) and from [27, Theorem 22.6] (together with the shape-regularity of the mesh family). When $V=H^1_{0\backslash\Upsilon_\mathrm{f}}(\mathcal{D})$, one has to use a quasi-interpolation operator which preserves the Dirichlet boundary condition. Such a construction is performed in [26] for purely Dirichlet boundary conditions. In such a configuration, the conclusion follows from [27, Theorem 22.14] (together with the shape-regularity of the mesh family). In our partially Dirichlet case, the arguments need to be slightly adapted. We will admit that the result from [27, Theorem 22.14] extends, and we refer the reader to [38]. □

We now treat separately the mixed and the purely Neumann cases.

B.2 Mixed boundary conditions

We here assume that $\Upsilon_\mathrm{t}\neq\varnothing$. We study the following problem:

$$\begin{cases} -\mathrm{div}(\mathfrak{d}\boldsymbol{\nabla}u) = r & \text{in } \mathcal{D},\\ u=\phi & \text{on } \Upsilon_\mathrm{t},\\ \mathfrak{d}\boldsymbol{\nabla}u\cdot\boldsymbol{n}=\theta & \text{on } \Upsilon_\mathrm{f}. \end{cases} \tag{48}$$

We assume that $r\in L^2(\mathcal{D})$, that θ belongs to $H^{-\frac{1}{2}}(\Upsilon_\mathrm{f})$ (as defined in Section 2), and that $\phi\in H^{\frac{1}{2}}(\Upsilon_\mathrm{t})$. Recall that Υ_t is Lipschitz in Υ. By Calderón's extension theorem (guaranteeing the existence of a bounded extension operator from $H^{\frac{1}{2}}(\Upsilon_\mathrm{t})$ to $H^{\frac{1}{2}}(\Upsilon)$) and the surjectivity of the trace operator (ensuring the existence of a bounded lifting operator from $H^{\frac{1}{2}}(\Upsilon)$ to $H^1(\mathcal{D})$), we infer the existence of $\overline{\phi}\in H^1(\mathcal{D})$ such that $\gamma(\overline{\phi})_{|\Upsilon_\mathrm{t}}=\phi$ and $\|\overline{\phi}\|_{1,\mathcal{D}}\leqslant c_\mathrm{s}\|\phi\|_{\frac{1}{2},\Upsilon_\mathrm{t}}$.

The weak formulation of Problem (48) writes: find $u\in H^1(\mathcal{D})$, $u=u_0+\overline{\phi}$, with $u_0\in H^1_{0\backslash\Upsilon_\mathrm{f}}(\mathcal{D})$ such that

$$(\mathfrak{d}\boldsymbol{\nabla}u_0,\boldsymbol{\nabla}v)_\mathcal{D} = (r,v)_\mathcal{D} + \langle\theta,\gamma(v)\rangle_{\Upsilon_\mathrm{f}} - (\mathfrak{d}\boldsymbol{\nabla}\overline{\phi},\boldsymbol{\nabla}v)_\mathcal{D} \qquad \forall v\in H^1_{0\backslash\Upsilon_\mathrm{f}}(\mathcal{D}). \tag{49}$$

We henceforth assume that the lifting $\overline{\phi}$ belongs to $H^{1+s}(\mathcal{D})$ for some $s>\frac{d}{2}-1$. Note that, since $s>\frac{d}{2}-1$, we have $\overline{\phi}\in C^0(\overline{\mathcal{D}})$.

Remark B.5 (Characterization of $H^{\frac{1}{2}+s}(\Upsilon_{\mathrm{t}})$)**.** *Formally, a necessary and sufficient condition for the existence of a regular lifting $\overline{\phi} \in H^{1+s}(\mathcal{D})$ is that "$\phi \in H^{\frac{1}{2}+s}(\Upsilon_{\mathrm{t}})$". The space $H^{\frac{1}{2}+s}(\Upsilon)$ has standard meaning for $s < \frac{1}{2}$, however its definition is unclear for $s \geqslant \frac{1}{2}$ without further regularity on Υ. Denoting by Υ_j, $1 \leqslant j \leqslant N$, the open faces of the polytopal domain $\mathcal{D}$, a necessary condition so as to ensure that $\phi = \gamma(\overline{\phi})_{|\Upsilon_{\mathrm{t}}}$ for some $\overline{\phi} \in H^{1+s}(\mathcal{D})$ is that $\phi_{|\Upsilon_j \cap \Upsilon_{\mathrm{t}}} \in H^{\frac{1}{2}+s}(\Upsilon_j \cap \Upsilon_{\mathrm{t}})$ for all $1 \leqslant j \leqslant N$. Of course this condition cannot be sufficient, and must be supplemented by some "jump" control between faces. To obtain necessary and sufficient conditions, one needs to finely characterize the range of the trace operator on $H^{1+s}(\mathcal{D})$. For Lipschitz polytopes, the range of the trace operator of order $n \in \mathbb{N}$ on $H^{\zeta}(\mathcal{D})$ such that $\zeta > n + \frac{1}{2}$ has been fully characterized in [30, Theorem 1.5.2.8] ($d = 2$) and in [3] ($d = 3$).*

We consider the following conforming finite element approximation of Problem (49): find $u_h \in V_h^k$, $u_h = u_{0,h} + \mathcal{I}_h^{k,d}(\overline{\phi})$, with $u_{0,h} \in V_{0,h}^k$ such that

$$(\mathfrak{a}\boldsymbol{\nabla} u_{0,h}, \boldsymbol{\nabla} v_h)_{\mathcal{D}} = (r, v_h)_{\mathcal{D}} + \langle \theta, \gamma(v_h) \rangle_{\Upsilon_{\mathrm{f}}} - (\mathfrak{a}\boldsymbol{\nabla}\mathcal{I}_h^{k,d}(\overline{\phi}), \boldsymbol{\nabla} v_h)_{\mathcal{D}} \qquad \forall v_h \in V_{0,h}^k. \tag{50}$$

Remark that there holds $\gamma(u_h) = \mathcal{I}_h^{k,d-1}(\phi)$ on Υ_{t}.

Lemma B.6 ($H^1(\mathcal{D})$-seminorm estimate)**.** *Assume that $u \in H^{1+m}(\mathcal{D})$, with $0 \leqslant m \leqslant s$. Let $\tau := \min(m, k)$. Then, the following estimate holds true, for some constant $c > 0$:*

$$\|\mathfrak{a}^{1/2}\boldsymbol{\nabla}(u - u_h)\|_{0,\mathcal{D}} \leqslant c\, a_\sharp^{1/2} h^\tau \left(|u|_{1+\tau,\mathcal{D}} + |\overline{\phi}|_{1+\tau,\mathcal{D}} \right). \tag{51}$$

Proof. Since $V_{0,h}^k \subset H^1_{0\setminus\Upsilon_{\mathrm{f}}}(\mathcal{D})$, the following orthogonality property holds true as a consequence of (49) and (50):

$$(\mathfrak{a}\boldsymbol{\nabla}(u - u_h), \boldsymbol{\nabla} v_h)_{\mathcal{D}} = 0 \qquad \forall v_h \in V_{0,h}^k. \tag{52}$$

From this, we obtain, for any $w_h \in V_h^k$ such that $\gamma(w_h) = \mathcal{I}_h^{k,d-1}(\phi)$ on Υ_{t},

$$\|\mathfrak{a}^{1/2}\boldsymbol{\nabla}(u - u_h)\|_{0,\mathcal{D}}^2 = (\mathfrak{a}\boldsymbol{\nabla}(u - u_h), \boldsymbol{\nabla}(u - w_h))_{\mathcal{D}} \leqslant \|\mathfrak{a}^{1/2}\boldsymbol{\nabla}(u - u_h)\|_{0,\mathcal{D}} \|\mathfrak{a}^{1/2}\boldsymbol{\nabla}(u - w_h)\|_{0,\mathcal{D}}.$$

Now, choosing $w_h = \Pi_{0,h}^k(u_0) + \mathcal{I}_h^{k,d}(\overline{\phi})$, we get

$$\|\mathfrak{a}^{1/2}\boldsymbol{\nabla}(u - u_h)\|_{0,\mathcal{D}} \leqslant \|\mathfrak{a}^{1/2}\boldsymbol{\nabla}(u_0 - \Pi_{0,h}^k(u_0))\|_{0,\mathcal{D}} + \|\mathfrak{a}^{1/2}\boldsymbol{\nabla}(\overline{\phi} - \mathcal{I}_h^{k,d}(\overline{\phi}))\|_{0,\mathcal{D}}.$$

We have $u_0 \in H^{1+m}(\mathcal{D})$ and $\overline{\phi} \in H^{1+s}(\mathcal{D}) \subset H^{1+m}(\mathcal{D})$. Hence, by the approximation properties of $\Pi_{0,h}^k$ (see Proposition B.4) and $\mathcal{I}_h^{k,d}$ (cf. e.g. [27, Corollary 19.8]), we infer

$$\|\mathfrak{a}^{1/2}\boldsymbol{\nabla}(u - u_h)\|_{0,\mathcal{D}} \leqslant c\, a_\sharp^{1/2} h^\tau \left(|u_0|_{1+\tau,\mathcal{D}} + |\overline{\phi}|_{1+\tau,\mathcal{D}} \right).$$

Since $|u_0|_{1+\tau,\mathcal{D}} \leqslant |u|_{1+\tau,\mathcal{D}} + |\overline{\phi}|_{1+\tau,\mathcal{D}}$, we obtain (51). □

Remark B.7 (Case $m > \frac{d}{2} - 1$)**.** *If $m > \frac{d}{2} - 1$ (then $u \in C^0(\overline{\mathcal{D}})$), one can choose $w_h = \mathcal{I}_h^{k,d}(u)$ in the proof of Lemma B.6. Doing so, one can prove in this case that (51) holds true with right-hand side simply proportional to $|u|_{1+\tau,\mathcal{D}}$.*

Lemma B.8 ($L^2(\mathcal{D})$-norm estimate)**.** *Assume that $u \in H^{1+m}(\mathcal{D})$, with $0 \leqslant m \leqslant s$. Let $\tau := \min(m,k)$, $\chi := \min\left(\frac{1}{2}+s, k+1\right)$, and $\eta := \min(\tau+\varepsilon, \chi)$, where $\varepsilon \in (\frac{1}{2}, 1]$ is the regularity exponent of the dual problem. Then, there is some constant $c > 0$ such that*

$$\|u - u_h\|_{0,\mathcal{D}} \leqslant c\, \varrho\, h^{\eta} \left(|u|_{1+\tau,\mathcal{D}} + |\overline{\phi}|_{1+\tau,\mathcal{D}} + \left(\sum_{j=1}^{N} |\phi|^2_{\chi,\Upsilon_j \cap \Upsilon_{\mathrm{t}}} \right)^{1/2} \right), \tag{53}$$

and there holds $\eta \in [\tau + \frac{1}{2}, k+1]$.

Proof. We resort to the Aubin–Nitsche duality argument. Recall that $\mathbb{d}$ is Lipschitz continuous in $\overline{\mathcal{D}}$, and that the dual solution z to (44) belongs to $H^{1+\varepsilon}(\mathcal{D})$ for $\varepsilon \in (\frac{1}{2}, 1]$ by Assumption B.1. As a consequence, $\mathbb{d}\boldsymbol{\nabla} z \in \boldsymbol{H}^{\varepsilon}(\mathcal{D})$, and there is a constant $c_{\mathrm{l}} > 0$, which depends linearly on the Lipschitz constant of $a_\sharp^{-1}\mathbb{d}$, such that

$$\|\mathbb{d}\boldsymbol{\nabla} z\|_{\varepsilon,\mathcal{D}} \leqslant c_{\mathrm{l}} a_\sharp \|\boldsymbol{\nabla} z\|_{\varepsilon,\mathcal{D}}. \tag{54}$$

Furthermore, one can give a sense to $\boldsymbol{\gamma}(\mathbb{d}\boldsymbol{\nabla} z)\cdot\boldsymbol{n}$ in $L^2(\Upsilon)$. We consider the following weak formulation of the dual Problem (44): find $z \in H^1_{0\setminus\Upsilon_{\mathrm{f}}}(\mathcal{D})$ such that

$$(\mathbb{d}\boldsymbol{\nabla} z, \boldsymbol{\nabla} w)_{\mathcal{D}} - (\boldsymbol{\gamma}(\mathbb{d}\boldsymbol{\nabla} z)\cdot\boldsymbol{n}, \gamma(w))_{\Upsilon_{\mathrm{t}}} = (t, w)_{\mathcal{D}} \qquad \forall w \in H^1(\mathcal{D}),$$

where we have leveraged the fact that $\mathbb{d}\boldsymbol{\nabla} z\cdot\boldsymbol{n} = 0$ on Υ_{f} to cancel out the Neumann boundary contribution. Testing with $w = (u - u_h) \in H^1(\mathcal{D})$, remarking that $\gamma(u - u_h) = \phi - \mathcal{I}_h^{k,d-1}(\phi)$ on Υ_{t}, and using the symmetry of $\mathbb{d}$, yields

$$(t, (u-u_h))_{\mathcal{D}} = (\boldsymbol{\nabla} z, \mathbb{d}\boldsymbol{\nabla}(u-u_h))_{\mathcal{D}} - (\boldsymbol{\gamma}(\mathbb{d}\boldsymbol{\nabla} z)\cdot\boldsymbol{n}, \phi - \mathcal{I}_h^{k,d-1}(\phi))_{\Upsilon_{\mathrm{t}}}.$$

Since $z \in H^1_{0\setminus\Upsilon_{\mathrm{f}}}(\mathcal{D})$, using the orthogonality property (52), we infer

$$(t, (u-u_h))_{\mathcal{D}} = (\boldsymbol{\nabla}(z - \Pi^k_{0,h}(z)), \mathbb{d}\boldsymbol{\nabla}(u-u_h))_{\mathcal{D}} - (\boldsymbol{\gamma}(\mathbb{d}\boldsymbol{\nabla} z)\cdot\boldsymbol{n}, \phi - \mathcal{I}_h^{k,d-1}(\phi))_{\Upsilon_{\mathrm{t}}},$$

hence, choosing $t = a_\sharp(u - u_h) \in L^2(\mathcal{D})$, there holds

$$\begin{aligned} a_\sharp \|u - u_h\|^2_{0,\mathcal{D}} \leqslant{}& \|\mathbb{d}^{1/2}\boldsymbol{\nabla}(z - \Pi^k_{0,h}(z))\|_{0,\mathcal{D}} \|\mathbb{d}^{1/2}\boldsymbol{\nabla}(u-u_h)\|_{0,\mathcal{D}} \\ &+ \|\boldsymbol{\gamma}(\mathbb{d}\boldsymbol{\nabla} z)\cdot\boldsymbol{n}\|_{0,\Upsilon_{\mathrm{t}}} \|\phi - \mathcal{I}_h^{k,d-1}(\phi)\|_{0,\Upsilon_{\mathrm{t}}} =: \mathfrak{T}_1 + \mathfrak{T}_2. \end{aligned} \tag{55}$$

Let us first estimate $\mathfrak{T}_1$ in (55). By the approximation properties of $\Pi^k_{0,h}$ (see Proposition B.4), the fact that $|z|_{1+\varepsilon,\mathcal{D}} \leqslant \|z\|_{1+\varepsilon,\mathcal{D}}$, and the regularity result (45), we infer

$$\mathfrak{T}_1 \leqslant c_{\mathrm{a}} c_{\mathrm{r}} \varrho^{\frac{1}{2}} a_\flat^{-\frac{1}{2}} h^{\varepsilon} \|t\|_{0,\mathcal{D}} \|\mathbb{d}^{1/2}\boldsymbol{\nabla}(u-u_h)\|_{0,\mathcal{D}}. \tag{56}$$

Let us now estimate $\mathfrak{T}_2$ in (55). Since $\varepsilon \in (\frac{1}{2}, 1]$, the trace theorem (1) followed by (54) yields

$$\|\boldsymbol{\gamma}(\mathbb{d}\boldsymbol{\nabla} z)\cdot\boldsymbol{n}\|_{0,\Upsilon_{\mathrm{t}}} \leqslant \|\boldsymbol{\gamma}(\mathbb{d}\boldsymbol{\nabla} z)\|_{0,\Upsilon} \leqslant \|\boldsymbol{\gamma}(\mathbb{d}\boldsymbol{\nabla} z)\|_{\varepsilon-\frac{1}{2},\Upsilon} \leqslant c_\gamma \|\mathbb{d}\boldsymbol{\nabla} z\|_{\varepsilon,\mathcal{D}} \leqslant c_\gamma c_{\mathrm{l}} a_\sharp \|\boldsymbol{\nabla} z\|_{\varepsilon,\mathcal{D}}.$$

Since $\|\boldsymbol{\nabla} z\|_{\varepsilon,\mathcal{D}} \leqslant \|z\|_{1+\varepsilon,\mathcal{D}}$, leveraging the regularity result (45), we infer

$$\mathfrak{T}_2 \leqslant c_\gamma c_{\mathrm{l}} c_{\mathrm{r}} \varrho \|t\|_{0,\mathcal{D}} \|\phi - \mathcal{I}_h^{k,d-1}(\phi)\|_{0,\Upsilon_{\mathrm{t}}}. \tag{57}$$

Plugging the estimates (56) and (57) into (55), recalling the definition of the function t, and using (i) (51) for $\mathfrak{T}_1$, and (ii) standard approximation properties for $\mathcal{I}_h^{k,d-1}$ (cf. e.g. [27, Corollary 19.8]) together with the admissibility of the mesh and the fact that $\phi_{|\Upsilon_j\cap\Upsilon_\mathrm{t}} \in H^{\frac12+s}(\Upsilon_j\cap\Upsilon_\mathrm{t})$ for all $1\leqslant j\leqslant N$ (see Remark B.5) for $\mathfrak{T}_2$, we infer

$$\|u-u_h\|_{0,\mathcal{D}} \leqslant c_1\varrho\, h^{\tau+\varepsilon}\left(|u|_{1+\tau,\mathcal{D}}+|\overline{\phi}|_{1+\tau,\mathcal{D}}\right)+c_2\varrho\, h^{\chi}\left(\sum_{j=1}^N |\phi|^2_{\chi,\Upsilon_j\cap\Upsilon_\mathrm{t}}\right)^{1/2},$$

which yields (53). To prove the upper bound on η, we just remark that $\tau+\varepsilon\leqslant k+1$ and $\chi\leqslant k+1$. For the lower bound, since $\chi-\frac12=\min(s,k+\frac12)$ and $\min(m,k)\leqslant\min(s,k+\frac12)$, one has $\chi\geqslant\tau+\frac12$. Together with $\varepsilon>\frac12$, this yields $\eta\geqslant\tau+\frac12$. □

Lemma B.9 ($L^2(\Upsilon)$-norm estimate)**.** *Assume that $u\in H^{1+m}(\mathcal{D})$, with $0\leqslant m\leqslant s$. Let $\tau:=\min(m,k)$, $\chi:=\min\left(\frac12+s,k+1\right)$, $\eta:=\min(\tau+\varepsilon,\chi)$, and $\delta:=\tau+\eta$, where $\varepsilon\in(\frac12,1]$ is the regularity exponent of the dual problem. Then, there is $c>0$ such that*

$$\|\gamma(u-u_h)\|_{0,\Upsilon}\leqslant c\,\varrho\, h^{\frac{\delta}{2}}\left(|u|_{1+\tau,\mathcal{D}}+|\overline{\phi}|_{1+\tau,\mathcal{D}}+\left(\sum_{j=1}^N|\phi|^2_{\chi,\Upsilon_j\cap\Upsilon_\mathrm{t}}\right)^{1/2}\right),\tag{58}$$

and there holds $\frac{\delta}{2}\in[\tau+\frac14,k+\frac12]$.

Proof. The estimate (58) is a consequence of the multiplicative trace inequality in $H^1(\mathcal{D})$ (cf. e.g. [12, (1.6.6)]):

$$\|\gamma(u-u_h)\|^2_{0,\Upsilon}\leqslant c_{\mathrm{mt}}\|u-u_h\|_{0,\mathcal{D}}\|u-u_h\|_{1,\mathcal{D}}.$$

Lemmas B.6 and B.8, and the fact that $\varrho\geqslant1$ and $\eta>\tau$ yield the conclusion. □

Remark B.10 (Case $\phi=0$)**.** *If $\phi=0$ on Υ_t, one can choose $\overline{\phi}=0$ in $\mathcal{D}$, hence $\chi=k+1$, $\eta=\tau+\varepsilon$, and the estimate (58) holds with $\frac{\delta}{2}=\tau+\frac{\varepsilon}{2}$ and right-hand side simply proportional to $|u|_{1+\tau,\mathcal{D}}$.*

B.3 Purely Neumann boundary conditions

We here assume that $\Upsilon_\mathrm{t}=\varnothing$. Let Λ be a nonempty relatively open Lipschitz subset of Υ, which is the finite union of $(d-1)$-dimensional polytopes, and satisfies $|\Lambda|_{d-1}=c_\Lambda|\Upsilon|_{d-1}$ for some $c_\Lambda\in(0,1]$ (note that $\Lambda=\Upsilon$ is allowed). For $\kappa\in\mathbb{R}$, we study the following problem:

$$\begin{cases}-\mathrm{div}(\mathfrak{a}\boldsymbol{\nabla}u)=r & \text{in } \mathcal{D},\\ \mathfrak{a}\boldsymbol{\nabla}u\cdot\boldsymbol{n}=\theta & \text{on } \Upsilon,\\ (\gamma(u),1)_\Lambda=|\Lambda|_{d-1}\kappa.\end{cases}\tag{59}$$

We assume that $r\in L^2(\mathcal{D})$, that θ belongs to $H^{-\frac12}(\Upsilon)$ (as defined in Section 2), and that $(r,1)_\mathcal{D}+\langle\theta,1\rangle_\Upsilon=0$, so that Problem (59) admits a unique solution.

The weak formulation of Problem (59) writes as follows: find $u\in H^1(\mathcal{D})$, $u=u^0+\iota^0$, $\iota^0:=\kappa-|\Lambda|_{d-1}^{-1}(\gamma(u^0),1)_\Lambda$, with $u^0\in H^{1,0}(\mathcal{D})$ such that

$$(\mathfrak{a}\boldsymbol{\nabla}u^0,\boldsymbol{\nabla}v)_\mathcal{D}=(r,v)_\mathcal{D}+\langle\theta,\gamma(v)\rangle_\Upsilon\qquad\forall v\in H^{1,0}(\mathcal{D}).\tag{60}$$

Let $\mathcal{D}'$ be another (Lipschitz) polytopal domain in $\mathbb{R}^d$ such that $\mathcal{D}$ and $\mathcal{D}'$ are disjoint and $\Lambda = \mathrm{int}(\partial\mathcal{D}' \cap \Upsilon)$. We let $\hat{\mathcal{D}}$ be the polytopal set of $\mathbb{R}^d$ such that $\overline{\hat{\mathcal{D}}} := \overline{\mathcal{D}} \cup \overline{\mathcal{D}'}$, and we assume that $\hat{\mathcal{D}}$ is Lipschitz. We henceforth assume that the solution $u \in H^1(\mathcal{D})$ to Problem (60) is such that $u = \hat{u}_{|\mathcal{D}}$ for some function $\hat{u} \in H^1(\hat{\mathcal{D}})$, and we also let $u' := \hat{u}_{|\mathcal{D}'}$. In the same vein, we suppose that $\mathfrak{a} = \hat{\mathfrak{a}}_{|\mathcal{D}}$ with $\hat{\mathfrak{a}} \in \mathbb{L}^\infty(\hat{\mathcal{D}})$ (in practice, $\hat{\mathfrak{a}}$ is a symmetric field such that $\mathfrak{a}' := \hat{\mathfrak{a}}_{|\mathcal{D}'}$ satisfies analogous properties to $\mathfrak{a}$, i.e. $\mathfrak{a}'$ is uniformly elliptic and belongs to $\mathbb{W}^{1,\infty}(\mathcal{D}')$). We denote by $\varrho' \geqslant 1$ (resp. $\hat{\varrho} \geqslant 1$) the heterogeneity/anisotropy ratio of $\mathfrak{a}'$ (resp. $\hat{\mathfrak{a}}$) in $\mathcal{D}'$ (resp. $\hat{\mathcal{D}}$), in such a way that $\max(\varrho, \varrho') \leqslant \hat{\varrho}$.

When Λ does not coincide with Υ (like it does when $\mathcal{D}$ is included inside $\mathcal{D}'$), we further assume that $\mathfrak{T}_h$ is geometrically compliant with Λ. We consider the following conforming approximation of Problem (60): find $u_h \in V_h^k$, $u_h = u_h^0 + \iota_h^0$, $\iota_h^0 := \kappa_h - |\Lambda|_{d-1}^{-1}(\gamma(u_h^0), 1)_\Lambda$, with $u_h^0 \in V_h^{k,0}$ such that

$$(\mathfrak{a}\boldsymbol{\nabla} u_h^0, \boldsymbol{\nabla} v_h)_\mathcal{D} = (r, v_h)_\mathcal{D} + \langle \theta, \gamma(v_h)\rangle_\Upsilon \qquad \forall v_h \in V_h^{k,0}. \tag{61}$$

The real number $\kappa_h = |\Lambda|_{d-1}^{-1}(\gamma(u_h), 1)_\Lambda$ is a given exterior approximation of κ, in the sense that it is inferred from $\mathcal{D}'$ in practice. Suppose that $\mathcal{D}'$ is also meshed, with same meshsize h, in such a way that the resulting global mesh on $\hat{\mathcal{D}}$ is admissible in the sense of Definition 6.1 (with $\Omega \leftarrow \hat{\mathcal{D}}$ and $\Gamma \leftarrow \Lambda$). Let $m \geqslant 0$ be some exponent such that both $u \in H^{1+m}(\mathcal{D})$ and $u' \in H^{1+m}(\mathcal{D}')$. Then, letting $\tau := \min(m, k)$, κ_h is assumed to satisfy

$$|\Lambda|_{d-1}^{1/2}|\kappa - \kappa_h| \leqslant c_\kappa \varrho' h^{\frac{\delta'}{2}} |u'|_{1+\tau,\mathcal{D}'}, \tag{62}$$

with $\delta' := 2\tau + \varepsilon'$ for some $\varepsilon' \in (\frac{1}{2}, 1]$. Note that, since $\hat{u} \in H^1(\hat{\mathcal{D}})$ by assumption, $\gamma(u) = \gamma'(u')$ on Λ so that $\kappa = |\Lambda|_{d-1}^{-1}(\gamma'(u'), 1)_\Lambda$. Assume that u' is solution to a mixed-type $\mathfrak{a}'$-weighted diffusion problem in $\mathcal{D}'$, with homogeneous Dirichlet boundary condition on some subset of $\partial\mathcal{D}'\backslash\overline{\Lambda}$, and dual regularity exponent ε'. Then, letting $\kappa_h := |\Lambda|_{d-1}^{-1}(\gamma'(u_h'), 1)_\Lambda$ with $u_h' \in (V_h^k)'$ finite element approximation of u' in $\mathcal{D}'$, the estimate (62) follows as a simple by-product of Remark B.10.

Lemma B.11 ($H^1(\mathcal{D})$-seminorm estimate)**.** *Assume that $u \in H^{1+m}(\mathcal{D})$, with $m \geqslant 0$. Let $\tau := \min(m, k)$. Then, the following estimate holds true, for some constant $c > 0$:*

$$\|\mathfrak{a}^{1/2}\boldsymbol{\nabla}(u - u_h)\|_{0,\mathcal{D}} \leqslant c\, a_\sharp^{1/2} h^\tau |u|_{1+\tau,\mathcal{D}}. \tag{63}$$

Proof. Since $V_h^{k,0} \subset H^{1,0}(\mathcal{D})$, the following orthogonality property holds true as a consequence of (60) and (61):

$$(\mathfrak{a}\boldsymbol{\nabla}(u^0 - u_h^0), \boldsymbol{\nabla} v_h)_\mathcal{D} = 0 \qquad \forall v_h \in V_h^{k,0}. \tag{64}$$

Therefore, $u_h^0 = \Pi_h^{k,0}(u^0)$. Now, since $\boldsymbol{\nabla}(u - u_h) = \boldsymbol{\nabla}(u^0 - u_h^0)$ and $|u_0|_{1+\tau,\mathcal{D}} = |u|_{1+\tau,\mathcal{D}}$, Proposition B.4 directly yields the result. □

Lemma B.12 ($L^2(\mathcal{D})$-norm estimate)**.** *Assume that $u \in H^{1+m}(\mathcal{D})$ and $u' \in H^{1+m}(\mathcal{D}')$, with $m \geqslant 0$. Let $\tau := \min(m, k)$, $\delta := 2\tau + \varepsilon$ where $\varepsilon \in (\frac{1}{2}, 1]$ is the regularity exponent of the dual problem in $\mathcal{D}$, and $\delta' := 2\tau + \varepsilon'$ for some $\varepsilon' \in (\frac{1}{2}, 1]$. Define $\hat{\delta} := \min(\delta, \delta')$. Then, there is some constant $c > 0$ such that*

$$\|u - u_h\|_{0,\mathcal{D}} \leqslant c\, \hat{\varrho}\, h^{\frac{\hat{\delta}}{2}} \left(|u|_{1+\tau,\mathcal{D}} + |u'|_{1+\tau,\mathcal{D}'}\right), \tag{65}$$

and there holds $\frac{\hat{\delta}}{2} \in (\tau + \frac{1}{4}, k + \frac{1}{2}]$.

Proof. Writing, for $\ell_{\mathcal{D}}$ the diameter of $\mathcal{D}$,

$$\|u-u_h\|_{0,\mathcal{D}} \leqslant \|u^0-u_h^0\|_{0,\mathcal{D}} + \ell_{\mathcal{D}}^{\frac12}|\Upsilon|_{d-1}^{1/2}|\iota^0-\iota_h^0|, \tag{66}$$

we first estimate $|\iota^0-\iota_h^0|$. We have

$$c_\Lambda^{\frac12}|\Upsilon|_{d-1}^{1/2}|\iota^0-\iota_h^0| = |\Lambda|_{d-1}^{1/2}|\iota^0-\iota_h^0| \leqslant |\Lambda|_{d-1}^{1/2}|\kappa-\kappa_h| + \|\gamma(u^0-u_h^0)\|_{0,\Lambda}.$$

Using the exterior estimate (62) on $|\Lambda|_{d-1}^{1/2}|\kappa-\kappa_h|$, the multiplicative trace inequality in $H^1(\mathcal{D})$ (cf. e.g. [12, (1.6.6)]), and the fact that $\boldsymbol{\nabla}(u^0-u_h^0)=\boldsymbol{\nabla}(u-u_h)$, we infer

$$\begin{aligned}c_\Lambda^{\frac12}|\Upsilon|_{d-1}^{1/2}|\iota^0-\iota_h^0| \leqslant{}& c_\kappa\varrho' h^{\frac{\delta'}{2}}|u'|_{1+\tau,\mathcal{D}'}\\ &+ c_{\mathrm{mt}}^{1/2}\left(\|u^0-u_h^0\|_{0,\mathcal{D}} + \|\boldsymbol{\nabla}(u-u_h)\|_{0,\mathcal{D}}^{1/2}\|u^0-u_h^0\|_{0,\mathcal{D}}^{1/2}\right).\end{aligned}$$

Plugging this estimate into (66), and using the $H^1(\mathcal{D})$-seminorm estimate (63), yields

$$\|u-u_h\|_{0,\mathcal{D}} \leqslant c_1\left(\|u^0-u_h^0\|_{0,\mathcal{D}} + \varrho^{\frac14}h^{\frac{\tau}{2}}|u|_{1+\tau,\mathcal{D}}^{1/2}\|u^0-u_h^0\|_{0,\mathcal{D}}^{1/2} + \varrho' h^{\frac{\delta'}{2}}|u'|_{1+\tau,\mathcal{D}'}\right). \tag{67}$$

Now, we invoke the Aubin–Nitsche duality argument to estimate $\|u^0-u_h^0\|_{0,\mathcal{D}}$. We consider the following weak formulation of Problem (44): find $z\in H^{1,0}(\mathcal{D})$ such that

$$(\mathfrak{a}\boldsymbol{\nabla}z,\boldsymbol{\nabla}w)_{\mathcal{D}} = (t,w)_{\mathcal{D}} \qquad \forall w\in H^{1,0}(\mathcal{D}).$$

Choosing $w=(u^0-u_h^0)\in H^{1,0}(\mathcal{D})$, we infer, by symmetry of $\mathfrak{a}$ and orthogonality (64),

$$(t,(u^0-u_h^0))_{\mathcal{D}} = (\boldsymbol{\nabla}z,\mathfrak{a}\boldsymbol{\nabla}(u^0-u_h^0))_{\mathcal{D}} = (\boldsymbol{\nabla}(z-\Pi_h^{k,0}(z)),\mathfrak{a}\boldsymbol{\nabla}(u^0-u_h^0))_{\mathcal{D}}.$$

Choosing $t=a_\sharp(u^0-u_h^0)\in L^2(\mathcal{D})$ (notice that $(t,1)_{\mathcal{D}}=0$), and leveraging the approximation result of Proposition B.4, combined with the regularity result (45) and the fact that $|z|_{1+\varepsilon,\mathcal{D}}\leqslant\|z\|_{1+\varepsilon,\mathcal{D}}$, as well as the $H^1(\mathcal{D})$-seminorm estimate (63), we obtain

$$\|u^0-u_h^0\|_{0,\mathcal{D}} \leqslant c_{\mathrm{a}}c_{\mathrm{r}}\varrho^{\frac12}a_\flat^{-\frac12}h^\varepsilon\|\mathfrak{a}^{1/2}\boldsymbol{\nabla}(u-u_h)\|_{0,\mathcal{D}} \leqslant c_2\varrho\, h^{\tau+\varepsilon}|u|_{1+\tau,\mathcal{D}}. \tag{68}$$

The conclusion follows from (67), together with $\varrho\geqslant 1$ and $\max(\varrho,\varrho')\leqslant\hat\varrho$. □

Lemma B.13 ($L^2(\Upsilon)$-norm estimate)**.** *Assume that $u\in H^{1+m}(\mathcal{D})$ and $u'\in H^{1+m}(\mathcal{D}')$, with $m\geqslant 0$. Let $\tau:=\min(m,k)$, $\delta:=2\tau+\varepsilon$ where $\varepsilon\in(\frac12,1]$ is the regularity exponent of the dual problem in $\mathcal{D}$, and $\delta':=2\tau+\varepsilon'$ for some $\varepsilon'\in(\frac12,1]$. Define $\hat\delta:=\min(\delta,\delta')$. Then, there is $c>0$ such that*

$$\|\gamma(u-u_h)\|_{0,\Upsilon} \leqslant c\,\hat\varrho\, h^{\frac{\hat\delta}{2}}\left(|u|_{1+\tau,\mathcal{D}}+|u'|_{1+\tau,\mathcal{D}'}\right), \tag{69}$$

and there holds $\frac{\hat\delta}{2}\in(\tau+\frac14,k+\frac12]$.

Proof. Starting from

$$\begin{aligned}\|\gamma(u-u_h)\|_{0,\Upsilon} &\leqslant \|\gamma(u^0-u_h^0)\|_{0,\Upsilon} + c_\Lambda^{-\frac12}|\Lambda|_{d-1}^{1/2}|\iota^0-\iota_h^0|\\ &\leqslant (1+c_\Lambda^{-\frac12})\|\gamma(u^0-u_h^0)\|_{0,\Upsilon} + c_\Lambda^{-\frac12}|\Lambda|_{d-1}^{1/2}|\kappa-\kappa_h|,\end{aligned}$$

we get, using the multiplicative trace inequality in $H^1(\mathcal{D})$ (cf. e.g. [12, (1.6.6)]) combined with the fact that $\boldsymbol{\nabla}(u^0 - u_h^0) = \boldsymbol{\nabla}(u - u_h)$, along with the exterior estimate (62) on $|\Lambda|_{d-1}^{1/2}|\kappa - \kappa_h|$,

$$\|\gamma(u-u_h)\|_{0,\Upsilon} \leqslant (1 + c_\Lambda^{-\frac{1}{2}}) c_{\mathrm{mt}}^{1/2} \left(\|u^0 - u_h^0\|_{0,\mathcal{D}} + \|\boldsymbol{\nabla}(u-u_h)\|_{0,\mathcal{D}}^{1/2} \|u^0 - u_h^0\|_{0,\mathcal{D}}^{1/2} \right) + c_\Lambda^{-\frac{1}{2}} c_\kappa \varrho' \, h^{\frac{\delta'}{2}} |u'|_{1+\tau,\mathcal{D}'}.$$

The conclusion follows from (68) and (63), as in the proof of Lemma B.12. □

References

[1] A. Abdulle, M. E. Huber, and S. Lemaire. An optimization-based numerical method for diffusion problems with sign-changing coefficients. *Comptes Rendus. Mathématique*, 355(4):472–478, 2017.

[2] Y. A. Abramovich and C. D. Aliprantis. *An invitation to operator theory*, volume 50 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2002.

[3] C. Bernardi, M. Dauge, and Y. Maday. Compatibilité de traces aux arêtes et coins d'un polyèdre. *C. R. Acad. Sci. Paris, Ser. I*, 331(9):679–684, 2000.

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

[5] A.-S. Bonnet-Ben Dhia, C. Carvalho, and P. Ciarlet Jr. Mesh requirements for the finite element approximation of problems with sign-changing coefficients. *Numer. Math.*, 138(4):801–838, 2018.

[6] 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.*, 46:1363–1387, 2012.

[7] A.-S. Bonnet-Ben Dhia, L. Chesnel, and X. Claeys. Radiation condition for a non-smooth interface between a dielectric and a metamaterial. *Math. Models Methods Appl. Sci.*, 23(9):1629–1662, 2013.

[8] A.-S. Bonnet-Ben Dhia, P. Ciarlet Jr., and C.-M. Zwölf. Time harmonic wave diffraction problems in materials with sign-shifting coefficients. *J. Comput. Appl. Math.*, 234(6):1912–1919, 2010. Corrigendum 2616.

[9] G. Bouchitté and D. Felbacq. Homogenization near resonances and artificial magnetism from dielectrics. *C. R. Acad. Sci. Paris, Ser. I*, 339(5):377–382, 2004.

[10] G. Bouchitté and B. Schweizer. Cloaking of small objects by anomalous localized resonance. *Quart. J. Mech. Appl. Math.*, 63(4):437–463, 2010.

[11] G. Bouchitté and B. Schweizer. Homogenization of Maxwell's equations in a split ring geometry. *SIAM Multiscale Model. Simul.*, 8(3):717–750, 2010.

[12] S. C. Brenner and L. R. Scott. *The mathematical theory of finite element methods*, volume 15 of *Texts in Applied Mathematics*. Springer, New York, third edition, 2008.

[13] R. Bunoiu and K. Ramdani. Homogenization of materials with sign changing coefficients. *Communications in Mathematical Sciences*, 14(4):1137–1154, 2016.

[14] E. Burman. Stabilized finite element methods for nonsymmetric, noncoercive, and ill-posed problems. Part I: Elliptic equations. *SIAM J. Sci. Comput.*, 35(6):A2752–A2780, 2013.

[15] E. Burman. Stabilised finite element methods for ill-posed problems with conditional stability. In G. R. Barrenechea, F. Brezzi, A. Cangiani, and E. H. Georgoulis, editors, *Building Bridges: Connections and Challenges in Modern Approaches to Numerical Partial Differential Equations*, volume 114 of *Lect. Notes Comput. Sci. Eng.*, pages 93–127. Springer, 2016.

[16] C. Carvalho, L. Chesnel, and P. Ciarlet Jr. Eigenvalue problems with sign-changing coefficients. *Comptes Rendus. Mathématique*, 355(6):671–675, 2017.

[17] C. Carvalho, P. Ciarlet Jr., and C. Scheid. Limiting amplitude principle and resonances in plasmonic structures with corners: numerical investigation. *Comput. Methods Appl. Mech. Engrg.*, 388:Paper No. 114207, 23, 2022.

[18] M. Cassier, C. Hazard, and P. Joly. Spectral theory for Maxwell's equations at the interface of a metamaterial. Part I: Generalized Fourier transform. *Comm. Partial Differential Equations*, 42(11):1707–1748, 2017.

[19] M. Cassier, C. Hazard, and P. Joly. Spectral theory for Maxwell's equations at the interface of a metamaterial. Part II: Limiting absorption, limiting amplitude principles and interface resonance. *Comm. Partial Differential Equations*, 47(6):1217–1295, 2022.

[20] J. Cheng and M. Yamamoto. One new strategy for a priori choice of regularizing parameters in Tikhonov's regularization. *Inverse Problems*, 16(4):L31–L38, 2000.

[21] 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.

[22] 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].

[23] P. Ciarlet Jr., D. Lassounon, and M. Rihani. An optimal control-based numerical method for scalar transmission problems with sign-changing coefficients. *SIAM J. Numer. Anal.*, 61(3):1316–1339, 2023.

[24] P.-H. Cocquet, P.-A. Mazet, and V. Mouysset. On the existence and uniqueness of a solution for some frequency-dependent partial differential equations coming from the modeling of metamaterials. *SIAM J. Math. Anal.*, 44(6):3806–3833, 2012.

[25] M. Costabel and E. Stephan. A direct boundary integral equation method for transmission problems. *J. Math. Anal. Appl.*, 106(2):367–413, 1985.

[26] A. Ern and J.-L. Guermond. Finite element quasi-interpolation and best approximation. *ESAIM: Math. Model. Numer. Anal.*, 51(4):1367–1385, 2017.

[27] A. Ern and J.-L. Guermond. *Finite Elements I: Approximation and interpolation*, volume 72 of *Texts in Applied Mathematics*. Springer, Cham, 2021.

[28] A. Ern and J.-L. Guermond. *Finite Elements II: Galerkin approximation, elliptic and mixed PDEs*, volume 73 of *Texts in Applied Mathematics*. Springer, Cham, 2021.

[29] P. Fernandes and M. Raffetto. Well-posedness and finite element approximability of time-harmonic electromagnetic boundary value problems involving bianisotropic materials and metamaterials. *Math. Models Methods Appl. Sci.*, 19(12):2299–2335, 2009.

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

[31] 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.

[32] 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.

[33] M. Halla. On the approximation of dispersive electromagnetic eigenvalue problems in two dimensions. *IMA J. Numer. Anal.*, 43(1):535–559, 2023.

[34] C. Hazard and S. Paolantoni. Spectral analysis of polygonal cavities containing a negative-index material. *Ann. H. Lebesgue*, 3:1161–1193, 2020.

[35] Y. Lai, H. Chen, Z.-Q. Zhang, and C. T. Chan. Complementary media invisibility cloak that cloaks objects at a distance outside the cloaking shell. *Phys. Rev. Lett.*, 102:093901, 2009.

[36] A. Lamacz and B. Schweizer. A negative-index meta-material for Maxwell's equations. *SIAM J. Math. Anal.*, 48(6):4155–4174, 2016.

[37] J. Li. A literature survey of mathematical study of metamaterials. *Int. J. Numer. Anal. Model.*, 13(2):230–243, 2016.

[38] M. W. Licht. Smoothed projections and mixed boundary conditions. *Math. Comp.*, 88(316):607–635, 2019.

[39] W. McLean. *Strongly elliptic systems and boundary integral equations*. Cambridge University Press, Cambridge, 2000.

[40] G. W. Milton and N.-A. Nicorovici. On the cloaking effects associated with anomalous localized resonance. *Proc. R. Soc. Lond. Ser. A*, 462(2074):3027–3059, 2006.

[41] H.-M. Nguyen. Asymptotic behavior of solutions to the Helmholtz equations with sign-changing coefficients. *Trans. Amer. Math. Soc.*, 367:6581–6595, 2015.

[42] H.-M. Nguyen. Cloaking via anomalous localized resonance for doubly complementary media in the quasistatic regime. *J. Eur. Math. Soc. (JEMS)*, 17(6):1327–1365, 2015.

[43] H.-M. Nguyen. Superlensing using complementary media. *Ann. Inst. H. Poincaré Anal. Non Linéaire*, 32(2):471–484, 2015.

[44] H.-M. Nguyen. Cloaking using complementary media in the quasistatic regime. *Ann. Inst. H. Poincaré Anal. Non Linéaire*, 33(6):1509–1518, 2016.

[45] H.-M. Nguyen. Limiting absorption principle and well-posedness for the Helmholtz equation with sign-changing coefficients. *J. Math. Pures Appl.*, 106(2):342–374, 2016.

[46] H.-M. Nguyen. Negative index materials: some mathematical perspectives. *Acta Math. Vietnam.*, 44(2):325–349, 2019.

[47] 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.

[48] N.-A. Nicorovici, R. C. McPhedran, and G. W. Milton. Optical and dielectric properties of partially resonant composites. *Phys. Rev. B*, 49:8479–8482, 1994.

[49] P. Ola. Remarks on a transmission problem. *J. Math. Anal. Appl.*, 196(2):639–658, 1995.

[50] J. B. Pendry. Negative refraction makes a perfect lens. *Phys. Rev. Lett.*, 85:3966–3969, 2000.

[51] A. H. Schatz. An observation concerning Ritz-Galerkin methods with indefinite bilinear forms. *Math. Comp.*, 28:959–962, 1974.

[52] R. A. Shelby, D. R. Smith, and S. Schultz. Experimental verification of a negative index of refraction. *Science*, 292:77–79, 2001.

[53] D. R. Smith, W. J. Padilla, D. C. Vier, S. C. Nemat-Nasser, and S. Schultz. Composite medium with simultaneously negative permeability and permittivity. *Phys. Rev. Lett.*, 84:4184–4187, 2000.

[54] A. N. Tikhonov. On the solution of ill-posed problems and the method of regularization. *Dokl. Akad. Nauk SSSR*, 151:501–504, 1963.

[55] V. G. Veselago. The electrodynamics of substances with simultaneously negative values of ε and μ. *Soviet Physics Uspekhi*, 10(4):509–514, 1968.