A stabilized hybridized Nitsche method for sign-changing elliptic PDEs

Erik Burman, Alexandre Ern, Janosch Preuss

Mathematical Models and Methods in Applied Sciences

A stabilized hybridized Nitsche method for sign-changing elliptic PDEs

Erik Burman

Department of Mathematics, University College London,
25 Gordon Street, WC1H 0AY London, United Kingdom,
e.burman@ucl.ac.uk

Alexandre Ern

CERMICS, ENPC, Institut Polytechnique de Paris, 77455 Marne-la-Vallee cedex, France and
Inria Paris, 48 rue Barrault, 75647 Paris, France
alexandre.ern@enpc.fr

Janosch Preuss

Department of Mathematics, University College London,
25 Gordon Street, WC1H 0AY London, United Kingdom,
j.preuss@ucl.ac.uk

Received (Day Month Year)
Revised (Day Month Year)
Communicated by (xxxxxxxxxx)

We present and analyze a stabilized hybridized Nitsche method for elliptic problems with sign-changing coefficients without imposing symmetry assumptions on the mesh around the material interfaces. The use of a stabilized primal-dual formulation allows us to cope with the sign-changing nature of the problem and to prove optimal error estimates under two assumptions on the continuous problem, namely that it admits a unique solution and that the contrast at the sign-changing interface lies outside a certain critical interval. The method can be used on arbitrary shape-regular meshes (fitted to material interfaces) and yields optimal convergence rates for smooth solutions. As an illustration, the method is applied to simulate a realistic acoustic cloaking device.

Keywords: sign-changing PDEs; Finite Elements; Stabilized methods; Hybridized Methods; Metamaterials.

AMS Subject Classification: 65N20, 65N30, 78A48

1. Introduction

In the last decades, extensive research was motivated by the aim to manufacture so-called metamaterials leading to propagation of electromagnetic and acoustic waves in ways that are unknown from naturally occuring materials. These synthetic materials have many interesting applications. Acoustic metamaterials, for example, can be used to cloak objects from incoming sound waves. We refer the reader to Ref. 31 for a review and to Ref. 36 for an overview of the mathematical theory

behind cloaking. Often, metamaterials are constructed by placing natural materials of size smaller than the wavelength of interest in repeating patterns to produce the desired effects. In the paper, we consider the fully homogenized version of such materials which already poses enough challenges for their mathematical analysis and numerical resolution. Indeed, when it comes to simulating wave propagation inside metamaterials, one is faced with the challenge to deal with the sign-changing nature of the material coefficients. This leads to variational problems which are not coercive, even when the wavenumber vanishes, thereby precluding the application of many established numerical methods relying on this property.

At the continuous level, the problem falls within the framework of the Fredholm alternative, and well-posedness holds provided the contrast between the material parameters at the sign-changing interface lies outside some critical interval of the negative real axis. In the numerical analysis literature, well-posedness at the continuous level is generally addressed by means of the notion of T-coercivity, introduced in Ref. 8. We refer to Ref. 5 in which scalar problems are studied using this approach and to Refs. 6 and 7 for applications to time-harmonic Maxwell problems. T-coercivity is a reformulation in the setting of Hilbert spaces of an inf-sup condition (see Remark 25.14 of Ref. 34). The T-operator allows one to keep track of the maximizer considered in the proof of the inf-sup condition. For well-posed problems in the sense of Hadamard, T-coercivity is equivalent to the Banach–Nečas–Babuška theorem in a symmetric setting (see, e.g., Chap. 25 of Ref. 34).

At the discrete level, even when the contrast lies outside the critical interval, standard Galerkin methods can struggle with sign-changing problems. Proving convergence on arbitrary regular meshes typically requires the additional assumption that the contrast is sufficiently large, see Refs. 8 and 25. Alternatively, certain conditions on the mesh must be respected for the Galerkin discretization to work properly. These conditions allow one to apply the T-coercivity framework, or, equivalently, to prove the relevant inf-sup condition at the discrete level, see, e.g., Refs. 27 and 4. Furthermore, we refer to Refs. 24 and 37 for an application to eigenvalue problems and to Ref. 25 for an application to multi-scale problems using the framework of localized orthogonal decomposition. Unfortunately, the above conditions on the mesh can be very challenging to realize if the interface at which the sign change occurs is geometrically complicated. They even become impossible to realize for certain advanced discretization techniques, e.g., geometrically unfitted methods. In the recent preprint 38, an approach was suggested to avoid these stringent assumptions on the mesh. However, its implementation involves an intricate assembly procedure and a delicate construction of adapted quadrature rules, which have so far been neglected in the corresponding error analysis.

Let us continue to discuss alternative approaches. In the case of piecewise constant coefficients, the problem can be treated using boundary element methods, see Section 5 in Ref. 43. For more general settings, we are aware of two other finite-element based approaches that can be applied on general meshes. An optimization-based method was proposed in Ref. 1, see also Ref. 2, and very recently also an

optimal control-based method was devised in Ref. 28 which overcomes a potentially restrictive regularity condition required in the former reference. These methods can be proven to converge. The method proposed in Ref. 1 even converges in natural norms when the contrast lies inside the critical interval. However, it seems that convergence rates have not been proven for either of these methods. Moreover, the solution of the associated optimization problems requires additional computational effort. Hence, there is the need for more research on this subject.

In the paper, we propose a stabilized finite element method for the numerical simulation of acoustic metamaterials. We proceed in the spirit of Ref. 10 in which a stabilized primal-dual framework is introduced to discretize non-coercive or ill-posed problems using the finite element method. This methodology has for example been applied to various unique continuation problems[11,16,17,19,20] for which it leads to optimal error estimates when combined with appropriate conditional stability estimates for the continuous problem[21]. In the paper, we show how to apply this framework to treat problems with sign-changing coefficients under the assumption that the (possibly curved) meshes are fitted to the interface. In particular, we derive optimal error estimates in the H^1-norm under the assumption that the problem admits a unique solution and the contrast lies outside the critical interval mentioned above. A hybridized Nitsche method is considered in which an interface variable is introduced to enforce the appropriate zero-jump conditions across the interface. Moreover, the discretization hinges, for both the primal and the dual variable, on continuous finite elements on both subdomains, and discontinuous finite elements for the interface variable. The hybridized Nitsche method has been introduced in Ref. 32 for an elliptic interface problem without sign-changes. We also notice that it is possible to employ a hybridized discontinuous Galerkin method in the subdomains, as done, e.g., in Refs. 12 and 13 for unique continuation problems.

Let us briefly distinguish our method from the ones already proposed in the literature. In contrast to the plain Galerkin discretization, our approach is applicable on arbitrary shape-regular meshes without any symmetry requirement. The price to pay is to solve for a larger problem since a dual variable is introduced and needs to be approximated as well. Furthermore, even though there already exist other methods in the literature[1,28,2] which can be applied on arbitrary shape-regular meshes, it seems that convergence rates for these methods have not been shown so far. As we are able to do so under the assumption of well-posedness, it thus seems that our method closes a gap in the literature. Let us recall that to guarantee the mentioned convergence rates in the H^1-norm, we require in particular that the contrast lies outside the critical interval.

Clearly, several open questions remain for future research. Firstly, if we lower the assumption on the continuous problem to uniqueness, we still obtain convergence, but in a fairly weak norm which does not provide much information of practical interest about the quality of the approximate solution. Secondly, our method requires that the solution of the continuous problem is subdomain-wise in H^s for $s > 3/2$, which is the same assumption stated in Proposition 2.1 of Ref. 1. This assumption

may fail to hold if the interface has very low regularity. We mention that some methods proposed in the literature operate under lower regularity assumptions as those in Refs. 28 and 38. Thus, the extension of our method to allow for weaker assumptions on the continuous problem and the study of its convergence (rates) remains a topic that deserves additional research.

The remainder of the paper is structured as follows. We define the hybridized Nitsche method in Sec. 2 and prove its convergence in Sec. 3. In Sec. 4, we conduct numerical experiments to investigate the performance of the method for academic test cases and an actual metamaterial proposed in the physics literature. We finish in Sec. 5 with a conclusion and give some perspectives on further research.

2. Continuous and discrete settings

In this section, we present the continuous and discrete settings. In Sec. 2.1, we introduce the continuous model problem and state certain assumptions related to its solvability. We then move on to the discrete setting by introducing the mesh and the finite element space in Sec. 2.2. The standard hybridized Nitsche method and its limitations when applied to sign-changing problems are discussed in Sec. 2.3. To resolve these issues, we combine this method in Sec. 2.4 with a stabilized primal-dual approach. Sec. 2.5 is devoted to the discussion of the concrete choice of stabilization terms for the sign-changing problem.

2.1. *Model problem*

We consider a bounded domain $\Omega \subset \mathbb{R}^d$ for $d \in \{2,3\}$ split by an interface Γ into two subdomains $\Omega_\pm$ in such a way that $\overline{\Omega} = \overline{\Omega}_+ \cup \overline{\Omega}_-$ and $\partial\Omega_+ \cap \partial\Omega_- = \Gamma$. For a pair of constants $\sigma_+ > 0$, $\sigma_- < 0$, a pair of functions $\mu_\pm \in L^\infty(\Omega_\pm)$ representing reaction coefficients, and a pair of functions $f_\pm \in L^2(\Omega_\pm)$ representing source terms, we consider the following model problem: Find $u := (u_+, u_-) \in H^1(\Omega_+) \times H^1(\Omega_-)$ such that

$$\mathcal{L}_\pm(u_\pm) := -\nabla\cdot(\sigma_\pm\nabla u_\pm) + \mu_\pm u_\pm = f_\pm \quad \text{in } \Omega_\pm, \tag{2.1a}$$

$$u_\pm = 0 \quad \text{on } \partial\Omega_\pm \setminus \Gamma, \tag{2.1b}$$

together with the jump interface conditions

$$[\![u]\!]_\Gamma = 0 \quad \text{on } \Gamma, \tag{2.2a}$$

$$[\![\sigma\nabla u]\!]_\Gamma\cdot\mathbf{n}_\Gamma = 0 \quad \text{on } \Gamma, \tag{2.2b}$$

where

$$[\![u]\!]_\Gamma := u_+|_\Gamma - u_-|_\Gamma, \qquad [\![\sigma\nabla u]\!]_\Gamma := \sigma_+\nabla u_+|_\Gamma - \sigma_-\nabla u_-|_\Gamma. \tag{2.3}$$

Here, we denote the outward unit normal vector of the subdomain $\Omega_\pm$ by $\mathbf{n}_\pm$ and conventionally set $\mathbf{n}_\Gamma := \mathbf{n}_+$ pointing from Ω_+ into Ω_-. Note that there is no assumption on the sign of $\mu_\pm$, so that (2.1a) covers, in particular, the case of the

Helmholtz equation. Owing to the jump condition (2.2a), it is meaningful to consider the function $\tilde{u} \in H^1(\Omega)$ such that $\tilde{u}|_{\Omega_\pm} = u_\pm$, and we notice that $\tilde{u} \in H^1_0(\Omega)$ owing to (2.1b). We slightly abuse the notation and write u instead of $\tilde{u}$. We also define the spaces

$$V_\pm^{\rm reg} := \{v \in H^1(\Omega_\pm) \mid -\nabla \cdot (\sigma_\pm \nabla v_\pm) \in L^2(\Omega_\pm)\}, \tag{2.4}$$

and notice that $u_\pm \in V_\pm^{\rm reg}$ owing to (2.1a) and our assumption $f_\pm \in L^2(\Omega_\pm)$. For later purpose, we define $\sigma_\flat := \min(\sigma_+, -\sigma_-)$ and $\sigma_\sharp := \max(\sigma_+, -\sigma_-)$.

To perform the error analysis, some assumptions on the model problem (2.1)-(2.2) are required. The weakest assumption under which our method can operate is that of uniqueness.

Assumption 1 (Uniqueness). The model problem (2.1)-(2.2) with zero right-hand side admits only the trivial solution in $H^1_0(\Omega)$.

Under this assumption we merely show convergence in a very weak norm. To obtain convergence in a stronger norm, we require a well-posedness result, or at least a conditional stability result. From Ref. 5, see also Refs. 30, 9, 27, it is known that the model problem (2.1)-(2.2) is Fredholm provided that the contrast σ_-/σ_+ lies outside of some critical interval $I^{\rm crit}$ of the negative real axis. This leads to the following assumption, under which we show optimal convergence in the H^1-norm. We remark that the analysis in the recently published Ref. 28 is performed under the same assumption.

Assumption 2 (Uniqueness and $\sigma_-/\sigma_+ \notin I^{\rm crit}$). There is $C^{\rm stab}$ such that, for all $f \in H^{-1}(\Omega)$, there exists a unique $u \in H^1_0(\Omega)$ so that $\sum_\pm (\sigma_\pm \nabla u_\pm, \nabla v_\pm)_{\Omega_\pm} + (\mu_\pm u_\pm, v_\pm)_{\Omega_\pm} = \langle f, v\rangle$ for all $v \in H^1_0(\Omega)$, and u fulfills the stability estimate

$$\Big\{ \sum_\pm |\sigma_\pm| \, \|\nabla u_\pm\|^2_{\Omega_\pm} \Big\}^{\frac{1}{2}} \le \sigma_\flat^{-\frac{1}{2}} C^{\rm stab} \, \|f\|_{H^{-1}(\Omega)}\,, \tag{2.5}$$

with $\|f\|_{H^{-1}(\Omega)} := \sup_{0\neq y\in H^1_0(\Omega)} \frac{\langle f,y\rangle}{\|\nabla y\|_\Omega}$ and where the factor $\sigma_\flat^{-\frac{1}{2}}$ is a natural scaling to make the constant $C^{\rm stab}$ nondimensional.

Generic constants that are independent of the problem parameters (including the size of Ω) are simply called C in what follows. The value of C can change at each occurrence. We characterize the dependence of the other named constants on the problem parameters explicitly.

We define the functional spaces $V_\pm := \{v_\pm \in H^1(\Omega_\pm) \mid v_\pm|_{\partial\Omega_\pm\setminus\Gamma} = 0\}$ and

$$V := V_+ \times V_-, \quad V_\Gamma := L^2(\Gamma), \quad \hat{V} := V \times V_\Gamma. \tag{2.6}$$

Note that for a function $v := (v_+, v_-) \in V$, we have in general $[\![v]\!]_\Gamma \neq 0$. For later use, we record here the following result.

Lemma 2.1 (Poincaré inequality). *There is a (nondimensional) constant $C^{\mathrm{P}} > 0$ (independent of the problem parameters) such that, for all $(z, z_\Gamma) \in \hat{V}$, we have*

$$\|z\|_{L^2(\Omega)}^2 = \sum_{\pm} \|z_\pm\|_{\Omega_\pm}^2 \le C^{\mathrm{P}} \ell_\Omega^2 \sum_{\pm} \Big\{ \|\nabla z_\pm\|_{\Omega_\pm}^2 + \ell_\Omega^{-1} \|z_\pm - z_\Gamma\|_\Gamma^2 \Big\}, \tag{2.7}$$

where ℓ_Ω is a length scale associated with Ω, e.g., its diameter.

Proof. Owing to the Peetre–Tartar Lemma (see, e.g., Theorem 2.1.3 in Ref. 35 or Lemma A.20 in Ref. 33), we infer that there is a constant C such that, for all $z := (z_+, z_-) \in V$,

$$\|z\|_{L^2(\Omega)}^2 \le C \ell_\Omega^2 \left(\sum_{\pm} \|\nabla z_\pm\|_{\Omega_\pm}^2 + \ell_\Omega^{-1} \|[\![z]\!]_\Gamma\|_\Gamma^2 \right),$$

since the right-hand side is positive definite on V and the embedding $V \hookrightarrow L^2(\Omega)$ is compact. The claim follows by adding and subtracting z_Γ in the jump term and using the triangle inequality. □

2.2. *Discrete setting*

We assume that Ω and $\Omega_\pm$ are all (Lipschitz) polygons/polyhedra so that they can be meshed exactly by a matching affine triangulation. Let $\mathcal{T}_h$ be such a triangulation, so that the subtriangulations $\mathcal{T}_h^\pm := \{T \in \mathcal{T}_h \mid T \subset \Omega_\pm\}$ fit the subdomains $\Omega_\pm$, respectively. The mesh elements are conventionally taken to be closed sets, and the interiors of any two distinct elements are disjoint. Moreover, setting

$$\Omega_h := \bigcup_{T \in \mathcal{T}_h} T, \qquad \Omega_{h,\pm} := \bigcup_{T \in \mathcal{T}_h^\pm} T, \qquad \Gamma_h := \partial\overline{\Omega}_{h,+} \cap \partial\overline{\Omega}_{h,-}, \tag{2.8}$$

we have

$$\overline{\Omega} = \Omega_h, \qquad \overline{\Omega}_\pm = \Omega_{h,\pm}, \qquad \Gamma = \Gamma_h. \tag{2.9}$$

We discuss in Sec. 3.4 appropriate extensions of our method in the presence of curved boundaries $\partial\Omega$ and Γ for which the equalities in (2.9) no longer hold.

Continuing now with the case of polygonal/polyhedral domains, let the set of all facets $\mathcal{F}_h$ of $\mathcal{T}_h$ be partitioned into

$$\mathcal{F}_h = \mathcal{F}_h^{\partial\Omega} \cup \mathcal{F}_h^\Gamma \cup \mathcal{F}_h^+ \cup \mathcal{F}_h^-, \tag{2.10}$$

where $\mathcal{F}_h^{\partial\Omega}$ and $\mathcal{F}_h^\Gamma$ denote the facets on $\partial\Omega$ and Γ, respectively, and $\mathcal{F}_h^\pm$ are the interior facets of $\Omega_\pm$, i.e., those facets that neither belong to $\partial\Omega$ nor to Γ.

For all $F \in \mathcal{F}_h^\pm$, $\mathbf{n}_F$ denotes the unit normal to $F = T_1 \cap T_2$ with an arbitrary but fixed orientation and the jump operator across F is defined as

$$[\![\nabla u_\pm]\!]_F := \nabla u_\pm|_{T_1}|_F(\mathbf{n}_F \cdot \mathbf{n}_{T_1}) + \nabla u_\pm|_{T_2}|_F(\mathbf{n}_F \cdot \mathbf{n}_{T_2}), \tag{2.11}$$

where $\mathbf{n}_{T_1}$ and $\mathbf{n}_{T_2}$ are the outward pointing normal vectors of T_1 and T_2, respectively. This definition leaves an arbitrariness in the sign of the jump operator which is, however, irrelevant in what follows.

To alleviate technicalities, we assume furthermore that $\mathcal{T}_h$ is quasi-uniform and use a single mesh size h. We use the notation $(v,w)_M := \int_M vw \,\mathrm{d}x$ and $\|v\|_M^2 := (v,v)_M$ to denote the L^2-scalar product and norm (with appropriate Lebesgue measure) over a subset $M \subset \overline{\Omega}$ which can be a collection of either mesh cells or mesh facets.

Let $l \geq 0$ be a polynomial degree, let $\mathbb{P}_l(T)$ be the space of d-variate polynomials of degree at most l on $T \in \mathcal{T}_h$, and let $\mathbb{P}_l(F)$ be the space of $(d-1)$-variate polynomials of degree at most l on $F \in \mathcal{F}_h^\Gamma$. We define the usual continuous finite element spaces on the subdomains: For $l \geq 1$,

$$V_{h,\pm}^l := \left\{ v_{h,\pm} \in H^1(\Omega_\pm) \mid v_{h,\pm}|_T \in \mathbb{P}_l(T),\ \forall T \in \mathcal{T}_h^\pm \mid v_{h,\pm}|_{\partial\Omega_\pm \setminus \Gamma} = 0 \right\} \subset V_\pm, \tag{2.12}$$

and the discontinuous finite element space on the interface: For $l \geq 0$,

$$V_{h,\Gamma}^l := \left\{ v_h \in L^2(\Gamma) \mid v_h|_F \in \mathbb{P}_l(F),\ \forall F \in \mathcal{F}_h^\Gamma \right\} \subset V_\Gamma. \tag{2.13}$$

We will invoke the following trace inequality:

$$h \left\| \nabla z_\pm \right\|_\Gamma^2 \leq C_\pm^{\mathrm{tr}} \left\| \nabla z_\pm \right\|_{\Omega_\pm}^2, \qquad \forall z_\pm \in V_{h,\pm}^l, \tag{2.14}$$

where $C_\pm^{\mathrm{tr}}$ depends on the polynomial degree l (see, e.g., Sec. 12.2 in Ref. 33).

For a real number $s \geq 1$, we consider the broken Sobolev spaces $H^s(\mathcal{T}_h^\pm) := \{v_\pm \in L^2(\Omega_\pm) \mid v_\pm|_T \in H^s(T),\ \forall T \in \mathcal{T}_h^\pm\}$, and set $|v_\pm|_{H^s(\mathcal{T}_h^\pm)}^2 := \sum_{T \in \mathcal{T}_h^\pm} |v_\pm|_{H^s(T)}^2$. Let $\Pi_\pm^{h,l}$ and $\Pi_\Gamma^{h,l}$ denote (quasi-)interpolation operators into the spaces $V_{h,\pm}^l$, respectively V_Γ^l, with the expected approximation properties:

$$|v_\pm - \Pi_\pm^{h,l}(v_\pm)|_{H^m(\mathcal{T}_h^\pm)} \leq C h^{s-m} |v_\pm|_{H^s(\mathcal{T}_h^\pm)}, \tag{2.15}$$

for all $s \in \{1, \dots, l+1\}$ and all $m \in \{0, \dots, s\}$, and

$$\|v_\pm - \Pi_\Gamma^{h,l}(v_\pm)\|_F \leq C h^{s-\frac{1}{2}} |v_\pm|_{H^s(T^\pm)}, \quad \forall F \in \mathcal{F}_h^\Gamma \cup \mathcal{F}_h^\pm, \tag{2.16}$$

for all $s \in \{1, \dots, l+1\}$, where $T^\pm \in \mathcal{T}_h^\pm$ are the two mesh cells of which F is a facet. We can take $\Pi_\pm^{h,l}$ to be the Scott–Zhang operator[42] or the L^1-stable quasi-interpolation operators from Chap. 22 of Ref. 33, and $\Pi_\Gamma^{h,l}$ to be the local L^2-projection.

2.3. *Hybridized Nitsche method*

The presence of the interface can lead to contrasting physical phenomena in the respective subdomains. For this reason, it is useful to employ a flexible discretization that allows to decouple the subdomain problems as much as possible. We opt here for a hybridized Nitsche method[32] in which the coupling occurs only at the interface via a hybrid variable. The method offers the possibilities of using independent meshes

in the subdomains and to apply static condensation to obtain linear systems for the hybrid variable only.

We recall that the hybridized Nitsche method is defined via the following bilinear form on $(V_{h,\pm}^k \times V_{h,\Gamma}^k) \times (V_{h,\pm}^k \times V_{h,\Gamma}^k)$:

$$\begin{aligned} a_\pm[(u_{h,\pm},u_{h,\Gamma});(v_{h,\pm},v_{h,\Gamma})] &:= (\sigma_\pm \nabla u_{h,\pm}, \nabla v_{h,\pm})_{\Omega_\pm} + (\mu_\pm u_{h,\pm}, v_{h,\pm})_{\Omega_\pm} \\ &\quad - (\sigma_\pm \nabla u_{h,\pm}\cdot\mathbf{n}_\pm, v_{h,\pm} - v_{h,\Gamma})_\Gamma - (\sigma_\pm \nabla v_{h,\pm}\cdot\mathbf{n}_\pm, u_{h,\pm} - u_{h,\Gamma})_\Gamma \\ &\quad + \frac{\lambda_\pm|\sigma_\pm|}{h}(u_{h,\pm} - u_{h,\Gamma}, v_{h,\pm} - v_{h,\Gamma})_\Gamma, \end{aligned} \tag{2.17}$$

for user-dependent (nondimensional) parameters $\lambda_\pm > 0$ to be chosen sufficiently large. Notice that, in order to enhance stability and symmetrize the linear system, the bilinear form $a_\pm$ contains additional terms that vanish when $(u_{h,\pm}, u_{h,\Gamma})$ is replaced by the exact solution $(u|_{\Omega_\pm}, u|_\Gamma)$ of (2.1)-(2.2). In particular,

$$0 = (\llbracket \sigma \nabla u \rrbracket \cdot \mathbf{n}_\Gamma, v_{h,\Gamma})_\Gamma = \sum_\pm (\sigma_\pm \nabla u_\pm \cdot \mathbf{n}_\pm, v_{h,\Gamma})_\Gamma. \tag{2.18}$$

Owing to the addition of these terms, the stability for coercive problems readily follows via a standard application of Young's inequality and the trace inequality (2.14). However, the story is different for sign-changing coefficients since carrying out this argument yields

$$\begin{aligned} &a_\pm[(v_{h,\pm},v_{h,\Gamma});(v_{h,\pm},v_{h,\Gamma})] \\ &\quad \geq \sigma'_\pm \|\nabla v_{h,\pm}\|_{\Omega_\pm}^2 + (\mu_\pm v_{h,\pm}, v_{h,\pm})_{\Omega_\pm} + (\lambda_\pm - 2C_\pm^{\mathrm{tr}})\frac{|\sigma_\pm|}{h}\|v_{h,\pm} - v_{h,\Gamma}\|_\Gamma^2, \end{aligned} \tag{2.19}$$

with $\sigma'_+ := \frac{1}{2}\sigma_+$ and $\sigma'_- := \frac{3}{2}\sigma_-$. Since $\sigma'_- < 0$, the hybridized Nitsche method on its own does not lead to stable discretizations for sign-changing problems. A preliminary idea to fix this deficit is add an additional term that stabilizes $\|\nabla v_{h,-}\|_{\Omega_-}^2$. However, this is inconsistent as $\nabla u_- \not\equiv 0$ in general and thus spoils the convergence of the method. The way forward, which we detail in the next section, is to impose the PDE via a Lagrange multiplier which can be stabilized (almost) as much as desired as it approximates zero.

2.4. *Primal-dual stabilized FEM*

To resolve the issue encountered in the previous section, we utilize the methodology that was proposed in Ref. 10 to solve non-coercive or ill-posed problems via a stabilized primal-dual formulation. We introduce an additional dual variable that may live in an approximation space different from the primal variable. To allow for a compact notation, we define the discrete spaces

$$\hat{V}_h := (V_{h,+}^k \times V_{h,-}^k) \times V_{h,\Gamma}^k, \qquad \hat{V}_h^* := (V_{h,+}^{k^*} \times V_{h,-}^{k^*}) \times V_{h,\Gamma}^{k_\Gamma^*}, \tag{2.20}$$

and we use the notation $\hat{u}_h := (u_h, u_{h,\Gamma})$, $u_h := (u_{h,+}, u_{h,-})$ for a generic element of the space $\hat{V}_h$ associated with the primal variable and $\hat{z}_h := (z_h, z_{h,\Gamma})$, $z_h :=$

$(z_{h,+}, z_{h,-})$ for a generic element of $\hat{V}_h^*$ associated with the dual variable. We assume that

$$k \geq \max(k^*, k_\Gamma^*), \qquad k_\Gamma^* \geq k-1. \tag{2.21}$$

We define a Lagrangian for the discrete problem as

$$L(\hat{u}_h, \hat{z}_h) = a[\hat{u}_h, \hat{z}_h] - \tilde{f}(z_h) + \frac{1}{2}s[\hat{u}_h - \hat{u}, \hat{u}_h - \hat{u}] - \frac{1}{2}\tilde{s}(z_h, z_h), \tag{2.22}$$

which represents a residual energy to be minimized. Its constituents are defined and motivated as follows: (i) The first two terms in (2.22) enforce the PDE constraint, where we defined

$$a[\hat{u}_h, \hat{z}_h] := \sum_{\pm} a_\pm[(u_{h,\pm}, u_{h,\Gamma}); (z_{h,\pm}, z_{h,\Gamma})], \tag{2.23}$$

with $a_\pm[\cdot;\cdot]$ from (2.17) and

$$\tilde{f}(z_h) := \sum_{\pm} (f_\pm, z_{h,\pm})_{\Omega_\pm}. \tag{2.24}$$

(ii) The third term $s[\cdot,\cdot]$ minimizes the distance between the exact solution u of (2.1)-(2.2) and the discrete approximation in an appropriate (semi-)norm. Here, $\hat{u} := ((u_+, u_-), u_\Gamma)$ with $u_\pm := u|_{\Omega_\pm}$ and $u_\Gamma := u|_\Gamma$. This third term serves to stabilize the primal variable and can be used to incorporate a priori knowledge about the continuous solution into the minimization problem. (iii) The fourth term $\tilde{s}(\cdot,\cdot)$ stabilizes the dual variable. This term is crucial to recover an inf-sup stable formulation for the sign-changing problem. Here we use the notation $\tilde{s}(z_h, z_h)$ instead of $\tilde{s}[\hat{z}_h, \hat{z}_h]$ to underline that the dual stabilization will not depend on $z_{h,\Gamma}$.

We search for a critical point of the Lagrangian L by solving the first-order optimality conditions:

$$\partial_{\hat{z}_h} L(\hat{y}_h) = a[\hat{u}_h, \hat{y}_h] - \tilde{f}(y_h) - \tilde{s}(z_h, y_h) = 0, \quad \forall \hat{y}_h \in \hat{V}_h^*, \tag{2.25a}$$

$$\partial_{\hat{u}_h} L(\hat{w}_h) = a^*[\hat{z}_h, \hat{w}_h] + s[\hat{u}_h - \hat{u}, \hat{w}_h] = 0, \qquad \forall \hat{w}_h \in \hat{V}_h. \tag{2.25b}$$

The second equation (2.25b) for the dual variable $\hat{z}_h$ is associated with the formal adjoint bilinear form $a^*[\hat{z}_h, \hat{w}_h] := a[\hat{w}_h, \hat{z}_h]$. Note that setting $\hat{u}_h = \hat{u}$ and $\hat{z}_h = 0$ solves the system (see Lemma 2.3 below). That is, $\hat{z}_h$ approximates zero which leaves significant freedom in choosing the dual stabilization $\tilde{s}(\cdot,\cdot)$. In practice, it is however desirable to opt for the minimal choice for which stability holds since an excessive dual stabilization perturbs the PDE constraint for the primal variable realized by (2.25a). We define the bilinear form

$$B[(\hat{v}_h, \hat{z}_h); (\hat{w}_h, \hat{y}_h)] := a^*[\hat{z}_h, \hat{w}_h] + a[\hat{v}_h, \hat{y}_h] + s[\hat{v}_h, \hat{w}_h] - \tilde{s}(z_h, y_h). \tag{2.26}$$

This allows us to rewrite the equations in a more compact notation: Find $(\hat{u}_h, \hat{z}_h) \in \hat{V}_h \times \hat{V}_h^*$ such that, for all $(\hat{w}_h, \hat{y}_h) \in \hat{V}_h \times \hat{V}_h^*$,

$$B[(\hat{u}_h, \hat{z}_h); (\hat{w}_h, \hat{y}_h)] = \tilde{f}(y_h) + s[\hat{u}, \hat{w}_h]. \tag{2.27}$$

Let us now sketch how this formulation can be used to overcome the stability issue observed for the original hybridized Nitsche method in Sec. 2.3. We have

$$B[(\hat{v}_h,\hat{z}_h);(\hat{z}_h,0)] = a[\hat{z}_h,\hat{z}_h] + s[\hat{v}_h,\hat{z}_h], \tag{2.28}$$

and we recognize the term $a[\hat{z}_h,\hat{z}_h]$ from (2.19) which is indefinite in Ω_-. However, now the problem has shifted to the dual variable, and we are allowed to compensate by means of the dual stabilization. Indeed, for a real number $\alpha > 0$, we have

$$B[(\hat{v}_h,\hat{z}_h);(\alpha\hat{v}_h,-\alpha\hat{z}_h)] = \alpha s[\hat{v}_h,\hat{v}_h] + \alpha\tilde{s}(z_h,z_h),$$

and the dual stabilization $\alpha\tilde{s}(z_h,z_h)$ can be chosen so as to absorb terms that prevent $a[\hat{z}_h,\hat{z}_h]$ from being coercive, e.g., $-\|\nabla z_{h,-}\|^2_{\Omega_-}$. This is legitimate since $\hat{z}_h$ approximates zero. We refer to Lemma 3.1 below for the full proof of discrete inf-sup stability.

2.5. *Choice of appropriate stabilization*

It remains to define and motivate the stabilization terms

$$s[\hat{v}_h,\hat{w}_h] := \sum_{\pm} s_\pm[(v_{h,\pm},v_{h,\Gamma});(w_{h,\pm},w_{h,\Gamma})], \qquad \tilde{s}(z_h,y_h) := \sum_{\pm}\tilde{s}_\pm(z_{h,\pm};y_{h,\pm}), \tag{2.29}$$

appearing in the variational formulation (2.27). We define the dual stabilization as

$$\tilde{s}_\pm(z_{h,\pm},y_{h,\pm}) := \gamma^*_\pm|\sigma_\pm|(\nabla z_{h,\pm},\nabla y_{h,\pm})_{\Omega_\pm} + \tilde{\mu}_\pm(z_{h,\pm},y_{h,\pm})_{\Omega_\pm},$$

for (nondimensional) stabilization parameters $\gamma^*_\pm$ and $\tilde{\mu}_\pm := \|\mu^\ominus_\pm\|_{L^\infty(\Omega_\pm)}$ with the negative-part operator $\mu^\ominus_\pm := \frac{1}{2}(|\mu_\pm| - \mu_\pm)$. For later purpose, we also define $\mu^\oplus_\pm := \frac{1}{2}(|\mu_\pm| + \mu_\pm)$ and notice that $\mu_\pm = \mu^\oplus_\pm - \mu^\ominus_\pm$, and we set $\mu_{\infty,\pm} := \|\mu_\pm\|_{L^\infty(\Omega_\pm)}$. We notice that $\tilde{\mu}_\pm = 0$ if $\mu_\pm \geq 0$ and already mention that we can choose $\gamma^*_- := 1$ and $\gamma^*_+ := 0$ in this case. This shows that the dual stabilization is only required where $a[\cdot,\cdot]$ fails to be coercive. However, it is possible and potentially useful to take $\gamma^*_+ := 1$, e.g., to facilitate the solution of the linear systems or to retain some symmetry of the original problem.

Let us proceed to the definition of the primal stabilization. We set for all $(v_\pm,v_\Gamma),(w_\pm,w_\Gamma) \in \hat{V}$ with $v_\pm, w_\pm \in V^l_{h,\pm} + V^{\mathrm{reg}}_\pm \cap H^s(\mathcal{T}^\pm_h)$, $s > \frac{3}{2}$ (this last requirement allows us to consider jumps of normal derivatives across the mesh interfaces in $\Omega_\pm$),

$$\begin{aligned} s_\pm[(v_\pm,v_\Gamma);(w_\pm,w_\Gamma)] := & \sum_{T\in\mathcal{T}^\pm_h}\gamma^{\mathrm{LS}}_\pm\frac{h^2}{|\sigma_\pm|}\left(\mathcal{L}_\pm(v_\pm),\mathcal{L}_\pm(w_\pm)\right)_T \\ & + \sum_{F\in\mathcal{F}^\pm_h}|\sigma_\pm|\,h\left([\![\nabla v_\pm]\!]_F\cdot\mathbf{n}_F,[\![\nabla w_\pm]\!]_F\cdot\mathbf{n}_F\right)_F \\ & + \frac{|\sigma_\pm|}{h}(v_\pm - v_\Gamma, w_\pm - w_\Gamma)_\Gamma, \end{aligned} \tag{2.30}$$

for a (nondimensional) stabilization parameter $\gamma_\pm^{\mathrm{LS}} > 0$. The first term is a Galerkin least-squares stabilization that minimizes the PDE residual at the element-level. The second term promotes smoothness of the discrete solution by penalizing the jump of the normal derivative across facets. The use of this stabilization is motivated by the following identity which follows from integration by parts.

Lemma 2.2 (Basic identity). *The following holds for all* $v_\pm \in V_{h,\pm}^l + V_\pm^{\mathrm{reg}} \cap H^s(\mathcal{T}_h^\pm)$, $s > \frac{3}{2}$, *and all* $w_\pm \in V_\pm$*:*

$$\sum_{T\in\mathcal{T}_h^\pm} (\mathcal{L}_\pm(v_\pm), w_\pm)_T = (\sigma_\pm \nabla v_\pm, \nabla w_\pm)_{\Omega_\pm} + (\mu_\pm v_\pm, w_\pm)_{\Omega_\pm} - \sum_{F\in\mathcal{F}_h^\pm} (\sigma_\pm [\![\nabla v_\pm]\!]_F \cdot \mathbf{n}_F, w_\pm)_F - (\sigma_\pm \nabla v_\pm \cdot \mathbf{n}_\pm, w_\pm)_\Gamma. \tag{2.31}$$

Using this identity, we control below the PDE residual in the H^{-1}-norm by a weak norm which is essentially carried by the stabilization and in which convergence of the discrete to the exact solution is shown. We refer the reader to Lemmas 3.5 and 3.2 below for the full arguments. Combining these results with Assumption 2 then allows us to prove convergence in the H^1-norm.

Before embarking on the error analysis, let us convince ourselves that the proposed stabilization is consistent.

Lemma 2.3 (Consistency). *Let* $u \in H_0^1(\Omega)$ *solve (2.1)-(2.2) and denote* $\hat{u} = ((u_+, u_-), u_\Gamma)$ *with* $u_\pm := u|_{\Omega_\pm}$ *and* $u_\Gamma := u|_\Gamma$*. Assume that* $u_\pm \in H^s(\Omega_\pm)$, $s > \frac{3}{2}$*. The following holds: For all* $(\hat{w}_h, \hat{y}_h) \in \hat{V}_h \times \hat{V}_h^*$*,*

$$B[(\hat{u}, 0), (\hat{w}_h, \hat{y}_h)] = \sum_\pm \Big\{ (f_\pm, y_{h,\pm})_{\Omega_\pm} + \sum_{T\in\mathcal{T}_h^\pm} \gamma_\pm^{\mathrm{LS}} \frac{h^2}{|\sigma_\pm|} (f_\pm, \mathcal{L}_\pm(w_\pm))_T \Big\}. \tag{2.32}$$

Proof. We have $B[(\hat{u}, 0), (\hat{w}_h, \hat{y}_h)] = a[\hat{u}, \hat{y}_h] + s[\hat{u}, \hat{w}_h]$. Summing (2.17) over $\pm$, using that $[\![\sigma\nabla u]\!]\cdot\mathbf{n}_\Gamma = 0$, and invoking the basic identity from Lemma 2.2 (recall from Sec. 2.1 that $u_\pm \in V_\pm^{\mathrm{reg}}$) gives $a[\hat{u}, \hat{y}_h] = \sum_\pm (f_\pm, y_{h,\pm})_{\Omega_\pm}$. Moreover, we have

$$s[\hat{u}, \hat{w}_h] = \sum_\pm \sum_{T\in\mathcal{T}_h^\pm} \gamma_\pm^{\mathrm{LS}} \frac{h^2}{|\sigma_\pm|} (f_\pm, \mathcal{L}_\pm(w_\pm))_T \tag{2.33}$$

since $\mathcal{L}_\pm(u_\pm) = f_\pm$ in $\Omega_\pm$. This completes the proof. □

Notice that (2.33) also shows that the variational formulation (2.27) can be implemented based solely on the given data.

Remark 2.1 (Penalty on jumps of higher-order derivatives). So far, we added the minimal stabilization that allows us to prove optimal convergence rates in the H^1-norm if Assumption 2 holds, i.e., the solution is unique and the contrast lies outside the critical interval. In this case, the solution depends continuously on

the data, so that we have strong stability. If the contrast lies instead inside the critical interval, we expect the problem to be far less stable so that adding more stabilization (or regularization) at the discrete level may turn out to be beneficial. Here, we mention the possibility of strengthening the primal stabilization $s_\pm$ defined in (2.30) by adding a penalty on the jumps of second-order derivatives across facets:

$$\sum_{F\in\mathcal{F}_h^\pm} h^3 \left(\llbracket D^2 v_\pm \rrbracket_F, \llbracket D^2 w_\pm \rrbracket_F\right)_F,$$

where D^2 denotes the Hessian. A similar stabilization is already analyzed in the application of a stabilized primal-dual FEM to the unique continuation problem for the Lamé system, see Ref. 22. Here, we will numerically investigate the potential of this stabilization for the sign-changing problem in Sec. 4.3, in which an experiment for a problem with contrast lying inside the critical interval is presented.

3. Error analysis

The error analysis proceeds in three steps. In Sec. 3.1, we prove the (discrete) inf-sup stability of the bilinear form B in a weak triple norm $|||\cdot|||$ defined in (3.2) below. A convergence result in this norm is derived in Sec. 3.2. The first two steps only require uniqueness of the solution to the continuous problem (2.1)-(2.2), i.e., Assumption 1. An error estimate in the H^1-norm is derived in Sec. 3.3 under the stronger Assumption 2. Finally, an extension of the method to allow for curved interfaces is presented in Sec. 3.4.

3.1. *Discrete stability*

Here, we deal with the stability of the discrete problem, and, for ease of notation, we drop the index h on the discrete variables. We define the stabilization (semi-)norm on $\hat{V}_h$ by $|\hat{v}|_s^2 := s[\hat{v},\hat{v}]$, i.e.,

$$|\hat{v}|_s^2 := \sum_\pm \Bigg\{ \sum_{T\in\mathcal{T}_h^\pm} \gamma_\pm^{\mathrm{LS}} \frac{h^2}{|\sigma_\pm|} \|\mathcal{L}_\pm(v_\pm)\|_T^2 + \sum_{F\in\mathcal{F}_h^\pm} |\sigma_\pm|\, h \,\|\llbracket \nabla v_\pm \rrbracket_F \cdot \mathbf{n}_F\|_F^2 + \frac{|\sigma_\pm|}{h} \|v_\pm - v_\Gamma\|_\Gamma^2 \Bigg\}. \tag{3.1}$$

Furthermore, we define the following triple norm on $\hat{V}_h \times \hat{V}_h^*$:

$$|||(\hat{v},\hat{y})|||^2 := |\hat{v}|_s^2 + \sigma_\sharp^{-1} h \,\|\llbracket \sigma \nabla v \rrbracket_\Gamma \cdot \mathbf{n}_\Gamma\|_\Gamma^2 + \sum_\pm \Big\{ |\sigma_\pm| \,\|\nabla y_\pm\|_{\Omega_\pm}^2 + \tilde{\mu}_\pm \|y_\pm\|_{\Omega_\pm}^2 + \frac{|\sigma_\pm|}{h} \|y_\pm - y_\Gamma\|_\Gamma^2 \Big\}. \tag{3.2}$$

Lemma 3.1 (Triple norm). *Let Assumption 1 hold true. Then $|||\cdot|||$ defines a norm on $\hat{V}_h \times \hat{V}_h^*$.*

Proof. Let $(\hat{v},\hat{z}) \in \hat{V}_h \times \hat{V}_h^*$ be such that $|||(\hat{v},\hat{y})||| = 0$. We need to prove that $v_\pm = 0$, $v_\Gamma = 0$, $y_\pm = 0$, and $y_\Gamma = 0$. Owing to the definition of the triple norm, we infer that $\mathcal{L}_\pm(v_\pm)|_T = 0$ for all $T \in \mathcal{T}_h^\pm$, $[\![\nabla v_\pm]\!]_F \cdot \mathbf{n}_F = 0$ for all $F \in \mathcal{F}_h^\pm$, $[\![\sigma\nabla v]\!]_\Gamma \cdot \mathbf{n}_\Gamma = 0$, $v_+ = v_\Gamma = v_-$ on Γ, as well as $\nabla y_\pm = 0$ in $\Omega_\pm$ and $y_+ = y_\Gamma = y_-$ on Γ. Since $\hat{y} \in \hat{V}_h^* \subset \hat{V}$, the Poincaré inequality from Lemma 2.1 readily gives $y_\pm = 0$, and thus $y_\Gamma = 0$. Let us now deal with $\hat{v}$. Since $v_+ = v_\Gamma = v_-$, we infer that $[\![v]\!]_\Gamma = 0$ on Γ, and we also have $[\![\sigma\nabla v]\!]_\Gamma \cdot \mathbf{n}_\Gamma = 0$. Moreover, $\mathcal{L}_\pm(v_\pm)|_T = 0$, for all $T \in \mathcal{T}_h^\pm$, and $[\![\nabla v_\pm]\!]_F \cdot \mathbf{n}_F = 0$, for all $F \in \mathcal{F}_h^\pm$, imply that $\mathcal{L}_\pm(v_\pm)|_{\Omega_\pm} = 0$. Hence, $v = (v_+, v_-)$ solves (2.1)-(2.2) with right-hand side $f_\pm = 0$, so that v is zero by Assumption 1. Then, also $v_\Gamma = v_\pm|_\Gamma = 0$, and this completes the proof. □

Lemma 3.2 (Bound on stabilization). *Set the least-squares stabilization parameters so that*

$$\gamma_\pm^{\mathrm{LS}} \le \gamma_{\sharp,\pm}^{\mathrm{LS}} := \Big\{ \max(1, (\sigma_\flat|\sigma_\pm|)^{-1} h^2 \ell_\Omega^2 \mu_{\infty,\pm}^2) \Big\}^{-1}. \tag{3.3}$$

There is C^{s}, independent of h and the problem parameters, so that, for all $\hat{y} \in \hat{V}_h^*$,

$$|\hat{y}|_s^2 \le C^{\mathrm{s}} \sum_\pm |\sigma_\pm| \Big\{ \|\nabla y_\pm\|_{\Omega_\pm}^2 + \frac{1}{h} \|y_\pm - y_\Gamma\|_\Gamma^2 \Big\}. \tag{3.4}$$

Proof. Invoking inverse inequalities, the assumption (3.3) and $h \le \ell_\Omega$, we infer that

$$\begin{aligned}
&\sum_\pm \sum_{T\in\mathcal{T}_h^\pm} \gamma_\pm^{\mathrm{LS}} \frac{h^2}{|\sigma_\pm|} \|\mathcal{L}_\pm(y_\pm)\|_T^2 \le 2 \sum_\pm \sum_{T\in\mathcal{T}_h^\pm} \gamma_\pm^{\mathrm{LS}} \Big\{ h^2 |\sigma_\pm| \, \|\Delta y_\pm\|_T^2 + h^2 \tfrac{\mu_{\infty,\pm}^2}{|\sigma_\pm|} \|y_\pm\|_T^2 \Big\} \\
&\le C \sum_\pm \gamma_\pm^{\mathrm{LS}} \Big\{ |\sigma_\pm| \, \|\nabla y_\pm\|_{\Omega_\pm}^2 + h^2 \tfrac{\mu_{\infty,\pm}^2}{|\sigma_\pm|} \|y_\pm\|_{\Omega_\pm}^2 \Big\} \\
&\le C \sum_\pm \gamma_\pm^{\mathrm{LS}} \max\Big(1, h^2 \ell_\Omega^2 \tfrac{\mu_{\infty,\pm}^2}{\sigma_\flat |\sigma_\pm|}\Big) \Big\{ |\sigma_\pm| \, \|\nabla y_\pm\|_{\Omega_\pm}^2 + \sigma_\flat \ell_\Omega^{-2} \|y_\pm\|_{\Omega_\pm}^2 \Big\} \\
&\le C \sum_\pm \Big\{ |\sigma_\pm| \, \|\nabla y_\pm\|_{\Omega_\pm}^2 + \sigma_\flat \ell_\Omega^{-2} \|y_\pm\|_{\Omega_\pm}^2 \Big\}.
\end{aligned}$$

We conclude by observing that

$$\sigma_\flat \ell_\Omega^{-2} \sum_\pm \|y_\pm\|_{\Omega_\pm}^2 \le C^{\mathrm{P}} \sum_\pm |\sigma_\pm| \Big\{ \|\nabla y_\pm\|_{\Omega_\pm}^2 + \frac{1}{h} \|y_\pm - y_\Gamma\|_\Gamma^2 \Big\},$$

owing to the Poincaré inequality from Lemma 2.1, $\sigma_\flat \le |\sigma_\pm|$ and $h \le \ell_\Omega$. □

We can now state the main stability result of this section.

Theorem 3.1 (Inf-sup stability). *Assume that the polynomial degrees satisfy* (2.21). *Assume that $\lambda_\pm \ge 2C_\pm^{\mathrm{tr}} + \frac{1}{2}$ and that $\gamma_\pm^{\mathrm{LS}}$ are prescribed by* (3.3). *The following holds:*

$$\inf_{0\neq(\hat{v},\hat{z})\in\hat{V}_h\times\hat{V}_h^*} \sup_{0\neq(\hat{w},\hat{y})\in\hat{V}_h\times\hat{V}_h^*} \frac{B[(\hat{v},\hat{z});(\hat{w},\hat{y})]}{|||(\hat{v},\hat{z})|||\,|||(\hat{w},\hat{y})|||} \ge \beta > 0, \tag{3.5}$$

where, setting $\lambda_\sharp := \max(\lambda_\pm)$,

$$\beta := \frac{1}{4}\big(2(\alpha^2 + C^{\mathrm{s}} + 2\max(C_\pm^{\mathrm{tr}}) + 2)\big)^{-\frac{1}{2}}, \quad \alpha := \max(2, C^{\mathrm{s}} + \tfrac{2}{3}\lambda_\sharp^2 + \tfrac{1}{4}). \tag{3.6}$$

Consequently, under Assumption 1, the discrete problem (2.27) *is well-posed.*

Proof. (1) Let $\alpha > 0$ be chosen as in (3.6). We have

$$B[(\hat v,\hat z);(\alpha\hat v,-\alpha\hat z)] = \alpha s[\hat v,\hat v] + \alpha\tilde s(z,z) = \alpha|\hat v|_s^2 + \sum_\pm \alpha\Big\{\gamma_\pm^*|\sigma_\pm|\|\nabla z_\pm\|_{\Omega_\pm}^2 + \tilde\mu_\pm\|z_\pm\|_{\Omega_\pm}^2\Big\}. \tag{3.7}$$

(2) Since $k \ge \max\{k^*, k_\Gamma^*\}$ owing to (2.21), it is legitimate to test with $\hat w := \hat z$. We observe that

$$B[(\hat v,\hat z);(\hat z,0)] = a[\hat z,\hat z] + s[\hat v,\hat z].$$

On the one hand, proceeding as in (2.19), we obtain

$$a_\pm[(z_\pm,z_\Gamma);(z_\pm,z_\Gamma)] \ge \sigma'_\pm\|\nabla z_\pm\|_{\Omega_\pm}^2 + (\mu_\pm z_\pm, z_\pm)_{\Omega_\pm} + (\lambda_\pm - 2C_\pm^{\mathrm{tr}})\frac{|\sigma_\pm|}{h}\|z_\pm - z_\Gamma\|_\Gamma^2,$$

with $\sigma'_+ := \frac{1}{2}\sigma_+$ and $\sigma'_- := \frac{3}{2}\sigma_-$. Summing over both subdomains and using the assumption on $\lambda_\pm$, this gives

$$a[\hat z,\hat z] \ge \sum_\pm \Big\{\sigma'_\pm\|\nabla z_\pm\|_{\Omega_\pm}^2 + (\mu_\pm z_\pm, z_\pm)_{\Omega_\pm} + \frac{1}{2}\frac{|\sigma_\pm|}{h}\|z_\pm - z_\Gamma\|_\Gamma^2\Big\}. \tag{3.8}$$

On the other hand, owing to Young's inequality and the estimate (3.4), we infer that

$$\begin{aligned} s[\hat v,\hat z] &\ge -C^{\mathrm{s}}|\hat v|_s^2 - \frac{1}{4C^{\mathrm{s}}}|\hat z|_s^2 \\ &\ge -C^{\mathrm{s}}|\hat v|_s^2 - \frac{1}{4}\sum_\pm |\sigma_\pm|\Big\{\|\nabla z_\pm\|_{\Omega_\pm}^2 + \frac{1}{h}\|z_\pm - z_\Gamma\|_\Gamma^2\Big\}. \end{aligned}$$

Taking into account (3.8), this gives

$$B[(\hat v,\hat z);(\hat z,0)] \ge -C^{\mathrm{s}}|\hat v|_s^2 + \sum_\pm \Big\{\sigma''_\pm\|\nabla z_\pm\|_{\Omega_\pm}^2 + (\mu_\pm z_\pm, z_\pm)_{\Omega_\pm} + \frac{1}{4}\frac{|\sigma_\pm|}{h}\|z_\pm - z_\Gamma\|_\Gamma^2\Big\}, \tag{3.9}$$

with $\sigma''_+ := \frac{1}{4}\sigma_+$ and $\sigma''_- := \frac{7}{4}\sigma_-$.

(3) Since $k_\Gamma^* \ge k-1$ owing to (2.21), it is legitimate to test with $\hat\zeta := (0,\zeta_\Gamma)$ with $\zeta_\Gamma := \frac{h}{\sigma_\sharp}[\![\sigma\nabla v]\!]_\Gamma\cdot\mathbf{n}_\Gamma$. Using the Cauchy–Schwarz inequality followed by Young's

inequality and $|\sigma_\pm| \le \sigma_\sharp$, and observing that $s^*(\hat z, \hat\zeta) = 0$, this gives

$$\begin{aligned} B[(\hat v,\hat z);(0,\hat\zeta)] = a[\hat v,\hat\zeta] &= \sum_\pm \Big\{ (\sigma_\pm \nabla v_\pm\cdot\mathbf{n}_\pm, \zeta_\Gamma)_\Gamma - \frac{\lambda_\pm|\sigma_\pm|}{h}(v_\pm - v_\Gamma, \zeta_\Gamma)_\Gamma \Big\} \\ &= \frac{h}{\sigma_\sharp}\|[\![\sigma\nabla v]\!]_\Gamma\cdot\mathbf{n}_\Gamma\|_\Gamma^2 - \sum_\pm \frac{\lambda_\pm|\sigma_\pm|}{\sigma_\sharp}(v_\pm - v_\Gamma, [\![\sigma\nabla v]\!]_\Gamma\cdot\mathbf{n}_\Gamma)_\Gamma \\ &\ge \frac14\frac{h}{\sigma_\sharp}\|[\![\sigma\nabla v]\!]_\Gamma\cdot\mathbf{n}_\Gamma\|_\Gamma^2 - \sum_\pm \frac{2\lambda_\pm^2}{3}\frac{|\sigma_\pm|}{h}\|v_\pm - v_\Gamma\|_\Gamma^2. \end{aligned} \tag{3.10}$$

(4) Combining (3.7), (3.9), and (3.10), we infer that

$$\begin{aligned} B[(\hat v,\hat z);(\hat w,\hat y)] &\ge (\alpha - C^{\mathrm{s}} - \tfrac23\lambda_\sharp^2)|\hat v|_s^2 + \frac14\frac{h}{\sigma_\sharp}\|[\![\sigma\nabla v]\!]_\Gamma\cdot\mathbf{n}_\Gamma\|_\Gamma^2 \\ &+\sum_\pm \Big\{ (\alpha\gamma_\pm^*|\sigma_\pm| + \sigma_\pm'')\|\nabla z_\pm\|_{\Omega_\pm}^2 + \alpha\tilde\mu_\pm\|z_\pm\|_{\Omega_\pm}^2 + (\mu_\pm z_\pm, z_\pm)_{\Omega_\pm} + \frac14\frac{|\sigma_\pm|}{h}\|z_\pm - z_\Gamma\|_\Gamma^2 \Big\}, \end{aligned}$$

with $\hat w := \alpha\hat v + \hat z$, $\hat y := -\alpha\hat z + \hat\zeta$. The condition $\alpha \ge 2 \ge \frac54$ ensures that $\alpha\tilde\mu_\pm\|z_\pm\|_{\Omega_\pm}^2 + (\mu_\pm z_\pm, z_\pm)_{\Omega_\pm} \ge \frac14\tilde\mu_\pm\|z_\pm\|_{\Omega_\pm}^2$. Since we also have $\alpha \ge C^{\mathrm{s}} + \frac23\lambda_\sharp^2 + \frac14$, we obtain

$$\begin{aligned} B[(\hat v,\hat z);(\hat w,\hat y)] &\ge \frac14|\hat v|_s^2 + \frac14\frac{h}{\sigma_\sharp}\|[\![\sigma\nabla v]\!]_\Gamma\cdot\mathbf{n}_\Gamma\|_\Gamma^2 \\ &\quad + \sum_\pm \Big\{ (\alpha\gamma_\pm^*|\sigma_\pm| + \sigma_\pm'')\|\nabla z_\pm\|_{\Omega_\pm}^2 + \frac14\tilde\mu_\pm\|z_\pm\|_{\Omega_\pm}^2 + \frac14\frac{|\sigma_\pm|}{h}\|z_\pm - z_\Gamma\|_\Gamma^2 \Big\}. \end{aligned}$$

Finally, since $\alpha \ge 2$, we have $\alpha\gamma_\pm^*|\sigma_\pm| + \sigma_\pm'' \ge \frac14|\sigma_\pm|$. We infer that

$$B[(\hat v,\hat z);(\hat w,\hat y)] \ge \frac14|||(\hat v,\hat z)|||^2.$$

(5) To conclude the proof of the inf-sup condition, we bound $|||(\hat w,\hat y)|||$. We have

$$|||(\hat w,\hat y)|||^2 \le 2\alpha^2|||(\hat v,\hat z)|||^2 + 2|||(\hat z,\hat\zeta)|||^2.$$

Moreover, since $\zeta_\pm = 0$, we have

$$\begin{aligned} |||(\hat z,\hat\zeta)|||^2 &= |\hat z|_s^2 + \sigma_\sharp^{-1}h\,\|[\![\sigma\nabla z]\!]_\Gamma\cdot\mathbf{n}_\Gamma\|_\Gamma^2 + \sum_\pm \frac{|\sigma_\pm|}{h}\,\|\zeta_\Gamma\|_\Gamma^2 \\ &\le |\hat z|_s^2 + \sigma_\sharp^{-1}h\,\|[\![\sigma\nabla z]\!]_\Gamma\cdot\mathbf{n}_\Gamma\|_\Gamma^2 + 2\frac{h}{\sigma_\sharp}\|[\![\sigma\nabla v]\!]_\Gamma\cdot\mathbf{n}_\Gamma\|_\Gamma^2, \end{aligned}$$

where we used the definition of ζ_Γ and $|\sigma_\pm| \le \sigma_\sharp$. Invoking (3.4) to bound the first term on the right-hand side and the trace inequality (2.14) to bound the second one, we infer that

$$|||(\hat z,\hat\zeta)|||^2 \le (C^{\mathrm{s}} + 2\max(C_\pm^{\mathrm{tr}}) + 2)|||(\hat v,\hat z)|||^2.$$

(6) Since the discrete problem (2.27) amounts to a square linear system, its well-posedness is a direct consequence of the inf-sup condition (3.5) together with Lemma 3.1. □

3.2. *Convergence in the triple norm*

For all $\hat{v} := (v, v_\Gamma)$ with $v_\pm \in V_{h,\pm}^l + V_\pm^{\mathrm{reg}} \cap H^s(\mathcal{T}_h^\pm)$, $s > \frac{3}{2}$, and $v_\Gamma \in L^2(\Gamma)$, we define the norm

$$\begin{aligned}
|||\hat{v}|||_\sharp^2 &:= |||(\hat{v},0)|||^2 \\
&\quad + \sum_\pm |\sigma_\pm| \Big\{ \|\nabla v_\pm\|_{\Omega_\pm}^2 + h \|\nabla v_\pm \cdot \mathbf{n}_\Gamma\|_\Gamma^2 + (\sigma_\flat |\sigma_\pm|)^{-1} \ell_\Omega^2 \mu_{\infty,\pm}^2 \|v_\pm\|_{\Omega_\pm}^2 \Big\} \\
&= |\hat{v}|_s^2 + \sigma_\sharp^{-1} h \|[\![\sigma \nabla v]\!]_\Gamma \cdot \mathbf{n}_\Gamma\|_\Gamma^2 \\
&\quad + \sum_\pm |\sigma_\pm| \Big\{ \|\nabla v_\pm\|_{\Omega_\pm}^2 + h \|\nabla v_\pm \cdot \mathbf{n}_\Gamma\|_\Gamma^2 + (\sigma_\flat |\sigma_\pm|)^{-1} \ell_\Omega^2 \mu_{\infty,\pm}^2 \|v_\pm\|_{\Omega_\pm}^2 \Big\}.
\end{aligned} \tag{3.11}$$

Lemma 3.3 (Continuity). *For all $\hat{v} = (v, v_\Gamma)$ with $v_\pm \in V_{h,\pm}^l + V_\pm^{\mathrm{reg}} \cap H^s(\mathcal{T}_h^\pm)$, $s > \frac{3}{2}$, and $v_\Gamma \in L^2(\Gamma)$, and for all $(\hat{w}_h, \hat{y}_h) \in \hat{V}_h \times \hat{V}_h^*$, the following holds:*

$$\big|B[(\hat{v},0),(\hat{w}_h,\hat{y}_h)]\big| \le C^{\mathrm{bnd}} |||\hat{v}|||_\sharp |||(\hat{w}_h,\hat{y}_h)|||, \tag{3.12}$$

with a constant C^{bnd} independent of h and the problem parameters.

Proof. We observe that $\big|B[(\hat{v},0),(\hat{w}_h,\hat{y}_h)]\big| \le \big|a[\hat{v},\hat{y}_h]\big| + \big|s[\hat{v},\hat{w}_h]\big|$ and bound the two terms on the right-hand side.

(1) Bound on $a[\hat{v},\hat{y}_h]$. Since $\hat{y}_h \in \hat{V}_h^*$, we can use the trace inequality (2.14) to obtain

$$\begin{aligned}
\sum_\pm (\sigma_\pm \nabla y_{h,\pm} \cdot \mathbf{n}_\pm, v_\pm - v_\Gamma)_\Gamma &\le (C_\pm^{\mathrm{tr}})^{\frac12} \sum_\pm \|\nabla y_{h,\pm}\|_{\Omega_\pm} |\sigma_\pm| \, h^{-\frac12} \|v_\pm - v_\Gamma\|_\Gamma \\
&\le (C_\pm^{\mathrm{tr}})^{\frac12} |||\hat{v}|||_\sharp |||(0,\hat{y}_h)|||.
\end{aligned}$$

Moreover, we estimate the reaction term as follows:

$$\begin{aligned}
\sum_\pm (\mu_\pm v_\pm, y_{h,\pm})_{\Omega_\pm} &\le \left(\sum_\pm \sigma_\flat^{-1} \ell_\Omega^2 \mu_{\infty,\pm}^2 \|v_\pm\|_{\Omega_\pm}^2 \right)^{\frac12} \left(\sigma_\flat \ell_\Omega^{-2} \sum_\pm \|y_{h,\pm}\|_{\Omega_\pm}^2 \right)^{\frac12} \\
&\le \left(\sum_\pm \sigma_\flat^{-1} \ell_\Omega^2 \mu_{\infty,\pm}^2 \|v_\pm\|_{\Omega_\pm}^2 \right)^{\frac12} \\
&\quad \times (C^{\mathrm{P}})^{\frac12} \left(\sum_\pm |\sigma_\pm| \Big\{ \|\nabla y_{h,\pm}\|_{\Omega_\pm}^2 + h^{-1} \|y_{h,\pm} - y_{h,\Gamma}\|_\Gamma^2 \Big\} \right)^{\frac12} \\
&\le \left(\sum_\pm \sigma_\flat^{-1} \ell_\Omega^2 \mu_{\infty,\pm}^2 \|v_\pm\|_{\Omega_\pm}^2 \right)^{\frac12} (C^{\mathrm{P}})^{\frac12} |||(0,\hat{y}_h)|||,
\end{aligned}$$

where we used the Poincaré inequality from Lemma 2.1 and $\sigma_\flat \le |\sigma_\pm|$. The remaining terms are easily bounded using the Cauchy-Schwarz inequality, leading to

$$\big|a[\hat{v},\hat{y}_h]\big| \le C |||\hat{v}|||_\sharp |||(0,\hat{y}_h)|||.$$

(2) Bound on $s[\hat{v},\hat{w}_h]$. The Cauchy–Schwarz inequality readily gives

$$\big|s[\hat{v},\hat{w}_h]\big| \le |\hat{v}|_s|\hat{w}_h|_s \le |||\hat{v}|||_\sharp|||(\hat{w}_h,0)|||.$$

(3) Combining the bounds from Steps 1 and 2 yields the assertion since $|||(\hat{w}_h,\hat{y}_h)|||^2 = |||(\hat{w}_h,0)|||^2 + |||(0,\hat{y}_h)|||^2$. □

The next result demonstrates that the error in the triple norm $|||(\cdot,\cdot)|||$ is bounded by the best-approximation error in the augmented triple norm $|||\cdot|||_\sharp$.

Theorem 3.2 (Convergence in triple norm). *Let $u \in H^1_0(\Omega)$ solve (2.1)-(2.2) and denote $\hat{u} := ((u_+,u_-),u_\Gamma)$ with $u_\pm := u|_{\Omega_\pm}$ and $u_\Gamma := u|_\Gamma$. Assume that $u_\pm \in H^s(\mathcal{T}_h^\pm)$, $s > \frac{3}{2}$. Let $(\hat{u}_h,\hat{z}_h) \in \hat{V}_h \times \hat{V}_h^*$ solve the discrete problem (2.27). Under Assumption 1, we have*

$$|||(\hat{u}-\hat{u}_h,\hat{z}_h)||| \le \left(1+\frac{C^{\rm bnd}}{\beta}\right) \inf_{\hat{v}\in\hat{V}_h} |||\hat{u}-\hat{v}_h|||_\sharp. \tag{3.13}$$

Proof. Let $\hat{v}_h \in \hat{V}_h$ be arbitrary. For all $(\hat{w}_h,\hat{y}_h) \in \hat{V}_h \times \hat{V}_h^*$, the consistency result from Lemma 2.3 gives

$$\begin{aligned} B[(\hat{u}_h-\hat{v}_h,\hat{z}_h);(\hat{w}_h,\hat{y}_h)] &= B[(\hat{u}_h,\hat{z}_h);(\hat{w}_h,\hat{y}_h)] - B[(\hat{v}_h,0);(\hat{w}_h,\hat{y}_h)] \\ &= B[(\hat{u},0);(\hat{w}_h,\hat{y}_h)] - B[(\hat{v}_h,0);(\hat{w}_h,\hat{y}_h)] \\ &= B[(\hat{u}-\hat{v}_h,0);(\hat{w}_h,\hat{y}_h)]. \end{aligned}$$

Using Lemma 3.3 then yields

$$B[(\hat{u}_h-\hat{v}_h,\hat{z}_h);(\hat{w}_h,\hat{y}_h)] \le C^{\rm bnd}|||\hat{u}-\hat{v}_h|||_\sharp|||(\hat{w}_h,\hat{y}_h)|||.$$

In view of the inf-sup condition from (3.5), this implies

$$|||(\hat{u}_h-\hat{v}_h,\hat{z}_h)||| \le \frac{C^{\rm bnd}}{\beta}|||\hat{u}-\hat{v}_h|||_\sharp.$$

The claim then follows from the triangle inequality $|||(\hat{u}-\hat{u}_h,\hat{z}_h)||| \le |||(\hat{u}-\hat{v}_h,0)|||+|||(\hat{v}_h-\hat{u}_h,\hat{z}_h)|||$, observing that $|||(\hat{u}-\hat{v}_h,0)||| \le |||\hat{u}-\hat{v}_h|||_\sharp$, and recalling that $\hat{v}_h \in \hat{V}_h$ is arbitrary. □

Corollary 3.1 (Convergence rates for smooth solutions). *Under the assumptions of Theorem 3.2 and assuming that $u_\pm \in H^{k+1}(\mathcal{T}_h^\pm)$, we have*

$$|||(\hat{u}-\hat{u}_h,\hat{z}_h)||| \le CC^{\rm app}h^k\sum_\pm |u_\pm|_{H^{k+1}(\mathcal{T}_h^\pm)}, \tag{3.14}$$

with $C^{\rm app} := \sigma_\sharp^{\frac{1}{2}}\max(1,(\sigma_\flat|\sigma_\pm|)^{-\frac{1}{2}}h\ell_\Omega\mu_{\infty,\pm})$.

Proof. Combine the estimate (3.13) with the approximation properties (2.15)-(2.16) (we used $\gamma_\pm^{\rm LS} \le 1$ owing to (3.3) to simplify some expressions). □

3.3. *Convergence in H^1*

To derive convergence rates in the H^1-norm, we assume well-posedness of the continuous problem (see Assumption 2). In particular, we would like to apply the stability estimate (2.5) to the error $u-u_h$. However, this is not possible since the discrete solution is not continuous across the interface. To overcome this issue, we interpolate the discrete solution into an $H^1(\Omega)$-conforming space. The following lemma ensures that the corresponding interpolation error can be bounded by the jump terms over the interface which are controlled by the triple norm.

Lemma 3.4 (Discontinuous to continuous interpolation). *There exists an interpolation operator Π_h^{c} from $\hat{V}_h$ into a subspace of $H^1(\Omega)$ such that, for all $\hat{w}_h \in \hat{V}_h$,*

$$\sum_{\pm} |\sigma_\pm| \Big\{ h^{-2} \left\| \Pi_h^{\mathrm{c}}(\hat{w}_h) - w_{h,\pm} \right\|_{\Omega_\pm}^2 + \left\| \nabla \left(\Pi_h^{\mathrm{c}}(\hat{w}_h) - w_{h,\pm} \right) \right\|_{\Omega_\pm}^2 \Big\} \leq C^{\Pi} \sum_{\pm} \frac{|\sigma_\pm|}{h} \left\| w_{h,\pm} - w_{h,\Gamma} \right\|_{\Gamma}^2, \quad (3.15)$$

with a constant C^{Π} independent of h and the problem parameters.

Proof. The claim follows from Lemma 3.2 & 5.3 and Remark 3.2 of Ref. 15 with the following minor modification in the construction: the value of $\Pi_h^{\mathrm{c}}(\hat{w}_h)$ at the mesh nodes located on Γ is prescribed by using $w_{h,\Gamma}$. □

We can now state our main error estimate.

Theorem 3.3. *Let $u \in H_0^1(\Omega)$ solve (2.1)-(2.2) and denote $\hat{u} := ((u_+, u_-), u_\Gamma)$ with $u_\pm := u|_{\Omega_\pm}$ and $u_\Gamma := u|_\Gamma$. Assume that $u_\pm \in H^s(\mathcal{T}_h^\pm)$, $s > \frac{3}{2}$. Let $(\hat{u}_h, \hat{z}_h) \in \hat{V}_h \times \hat{V}_h^*$ solve the discrete problem (2.27). Assume that $\gamma_\pm^{\mathrm{LS}} \geq \frac{1}{2}\gamma_{\sharp,\pm}^{\mathrm{LS}}$ with $\gamma_{\sharp,\pm}^{\mathrm{LS}}$ defined in* (3.3)*. Under Assumption 2, we have*

$$\Big\{ \sum_{\pm} |\sigma_\pm| \left\| \nabla(u - u_{h,\pm}) \right\|_{\Omega_\pm}^2 \Big\}^{\frac{1}{2}} \leq C C^{\mathrm{E}} \inf_{\hat{v} \in \hat{V}_h} |||\hat{u} - \hat{v}_h|||_\sharp, \quad (3.16)$$

with $C^{\mathrm{E}} := \max\left\{1, C^{\mathrm{stab}}\right\} \left(\frac{\sigma_\sharp}{\sigma_\flat}\right)^{\frac{1}{2}} \max\left\{1, |\sigma_\pm|^{-\frac{1}{2}} \ell_\Omega \mu_{\infty,\pm}^{\frac{1}{2}}, (\sigma_\flat |\sigma_\pm|)^{-\frac{1}{2}} h \ell_\Omega \mu_{\infty,\pm}\right\}$ and C^{stab} defined in Assumption 2.

Proof. We define the linear form $r_h \in H^{-1}(\Omega)$ so that, for all $y \in H_0^1(\Omega)$,

$$\langle r_h, y \rangle := \sum_{\pm} \Big\{ \left(\sigma_\pm \nabla \left(\Pi_h^{\mathrm{c}}(\hat{u}_h) - u_\pm \right), \nabla y \right)_{\Omega_\pm} + \left(\mu_\pm (\Pi_h^{\mathrm{c}}(\hat{u}_h) - u_\pm), y \right)_{\Omega_\pm} \Big\}. \quad (3.17)$$

It is shown in Lemma 3.5 below that

$$\|r_h\|_{H^{-1}(\Omega)} \leq C C^{\mathrm{R}} |||(\hat{u} - \hat{u}_h, \hat{z}_h)|||,$$

with $C^{\mathrm{R}} := \sigma_\sharp^{\frac{1}{2}} \max(1, |\sigma_\pm|^{-\frac{1}{2}} \ell_\Omega \mu_{\infty,\pm}^{\frac{1}{2}}, (\sigma_\flat|\sigma_\pm|)^{-\frac{1}{2}} h \ell_\Omega \mu_{\infty,\pm})$. Invoking the stability estimate (2.5) from Assumption 2, we infer that

$$\begin{aligned}\Big\{ \sum_{\pm} |\sigma_\pm| \, \|\nabla(\Pi_h^{\mathrm{c}}(\hat{u}_h) - u)\|_{\Omega_\pm}^2 \Big\}^{\frac{1}{2}} &\leq \sigma_\flat^{-\frac{1}{2}} C^{\mathrm{stab}} \, \|r_h\|_{H^{-1}(\Omega)} \\ &\leq C \sigma_\flat^{-\frac{1}{2}} C^{\mathrm{stab}} C^{\mathrm{R}} |||(\hat{u} - \hat{u}_h, \hat{z}_h)|||.\end{aligned}$$

By applying the triangle inequality

$$\begin{aligned}\Big\{ \sum_{\pm} |\sigma_\pm| \, \|\nabla(u - u_{h,\pm})\|_{\Omega_\pm}^2 \Big\}^{\frac{1}{2}} \leq \Big\{ 2 \sum_{\pm} |\sigma_\pm| \, \|\nabla(\Pi_h^{\mathrm{c}}(\hat{u}_h) - u_{h,\pm})\|_{\Omega_\pm}^2 \Big\}^{\frac{1}{2}} \\ + \Big\{ 2 \sum_{\pm} |\sigma_\pm| \, \|\nabla(\Pi_h^{\mathrm{c}}(\hat{u}_h) - u)\|_{\Omega_\pm}^2 \Big\}^{\frac{1}{2}},\end{aligned}$$

and using the estimate (3.15) for the first term on the right-hand side, we obtain (we absorb the constant C^{Π} in the generic constant C and use that $\sigma_\flat^{-\frac{1}{2}} C^{\mathrm{R}} \geq 1$)

$$\Big\{ \sum_{\pm} |\sigma_\pm| \, \|\nabla(u - u_{h,\pm})\|_{\Omega_\pm}^2 \Big\}^{\frac{1}{2}} \leq C \max(1, C^{\mathrm{stab}}) \sigma_\flat^{-\frac{1}{2}} C^{\mathrm{R}} |||(\hat{u} - \hat{u}_h, \hat{z}_h)|||.$$

The claim follows by invoking the error estimate in the triple norm from Theorem 3.2. □

Remark 3.1. We have $C^{\mathrm{E}} \leq \max\big\{1, C^{\mathrm{stab}}\big\} \big(\frac{\sigma_\sharp}{\sigma_\flat}\big)^{\frac{1}{2}} \max\big\{1, (\sigma_\flat|\sigma_\pm|)^{-\frac{1}{2}} \ell_\Omega^2 \mu_{\infty,\pm}\big\}$.

Lemma 3.5 (Bound on r_h). *Let $r_h \in H^{-1}(\Omega)$ be defined in* (3.17)*. The following holds:*

$$\|r_h\|_{H^{-1}(\Omega)} \leq C C^{\mathrm{R}} |||(\hat{u} - \hat{u}_h, \hat{z}_h)|||, \tag{3.18}$$

with $C^{\mathrm{R}} := \sigma_\sharp^{\frac{1}{2}} \max\big\{1, |\sigma_\pm|^{-\frac{1}{2}} \ell_\Omega \mu_{\infty,\pm}^{\frac{1}{2}}, (\sigma_\flat|\sigma_\pm|)^{-\frac{1}{2}} h \ell_\Omega \mu_{\infty,\pm}\big\}$.

Proof. Let $y \in H_0^1(\Omega)$. Set $y_\pm := y|_{\Omega_\pm}$ and $y_\Gamma := y|_\Gamma$. We have

$$\begin{aligned}\langle r_h, y \rangle = &\sum_{\pm} \Big\{ \big(\sigma_\pm \nabla \left(\Pi_h^{\mathrm{c}}(\hat{u}_h) - u_{h,\pm}\right), \nabla y\big)_{\Omega_\pm} + \big(\mu_\pm \left(\Pi_h^{\mathrm{c}}(\hat{u}_h) - u_{h,\pm}\right), y\big)_{\Omega_\pm} \Big\} \\ &+ \sum_{\pm} \Big\{ \big(\sigma_\pm \nabla \left(u_{h,\pm} - u_\pm\right), \nabla y\big)_{\Omega_\pm} + \big(\mu_\pm \left(u_{h,\pm} - u_\pm\right), y\big)_{\Omega_\pm} \Big\} =: I_1 + \tilde{I}.\end{aligned}$$

We need to decompose $\tilde{I}$ further. Invoking the basic identity (2.31) from Lemma 2.2 gives

$$\begin{aligned}\tilde{I} = &\sum_{\pm} \Big\{ \sum_{T \in \mathcal{T}_h^\pm} (\mathcal{L}_\pm(u_{h,\pm} - u_\pm), y_\pm)_T + \sum_{F \in \mathcal{F}_h^\pm} (\sigma_\pm [\![\nabla u_h]\!]_F \cdot \mathbf{n}_F, y_\pm)_F \Big\} \\ &+ ([\![\sigma \nabla u_h]\!]_\Gamma \cdot \mathbf{n}_\Gamma, y_\Gamma)_\Gamma.\end{aligned}$$

Moreover, using the variational formulation (2.27) with $\hat{w}_h = 0$ and since $\mathcal{L}_\pm(u_\pm) = f_\pm$, we obtain, for all $\hat{y}_h \in \hat{V}_h$,

$$\sum_\pm \sum_{T\in\mathcal{T}_h^\pm} (\mathcal{L}_\pm(u_\pm), y_{h,\pm})_T = B[(\hat{u}_h, \hat{z}_h), (0, \hat{y}_h)] = a[\hat{u}_h, \hat{y}_h] - \tilde{s}(z_h, y_h).$$

Invoking again the basic identity (2.31) to transform the expression of $a[\hat{u}_h, \hat{y}_h]$, we infer that

$$\begin{aligned} 0 = \sum_\pm \Big\{ \sum_{T\in\mathcal{T}_h^\pm} (\mathcal{L}_\pm(u_{h,\pm} - u_\pm), y_{h,\pm})_T + \sum_{F\in\mathcal{F}_h^\pm} (\sigma_\pm [\![\nabla u_h]\!]_F \cdot \mathbf{n}_F, y_{h,\pm})_F \Big\} \\ + ([\![\sigma\nabla u_h]\!]_\Gamma \cdot \mathbf{n}_\Gamma, y_{h,\Gamma})_\Gamma + \sum_\pm \Big\{ \frac{\lambda_\pm|\sigma_\pm|}{h}(u_{h,\pm} - u_{h,\Gamma}, y_{h,\pm} - y_{h,\Gamma})_\Gamma \\ - (\sigma_\pm \nabla y_{h,\pm}\cdot\mathbf{n}_\pm, u_{h,\pm} - u_{h,\Gamma})_\Gamma \Big\} - \tilde{s}(z_h, y_h). \end{aligned}$$

Subtracting the above identity from the above expression for $\tilde{I}$ and adding to I_1 gives $\langle r_h, y\rangle = \sum_{j\in\{1:7\}} I_j$, with I_1 defined above and

$$\begin{aligned} I_2 &:= \sum_\pm \sum_{T\in\mathcal{T}_h^\pm} (\mathcal{L}_\pm(u_{h,\pm} - u_\pm), y_\pm - y_{h,\pm})_T\,, \\ I_3 &:= \sum_\pm \sum_{F\in\mathcal{F}_h^\pm} (\sigma_\pm [\![\nabla(u_h - u)]\!]_F \cdot \mathbf{n}_F, y_\pm - y_{h,\pm})_F\,, \\ I_4 &:= ([\![\sigma\nabla(u_h - u)]\!]_\Gamma \cdot \mathbf{n}_\Gamma, y_\Gamma - y_{h,\Gamma})_\Gamma\,, \\ I_5 &:= \sum_\pm \frac{\lambda_\pm\,|\sigma_\pm|}{h}(u_{h,\pm} - u_{h,\Gamma}, y_{h,\Gamma} - y_{h,\pm})_\Gamma \\ I_6 &:= \sum_\pm (\sigma_\pm \nabla y_{h,\pm}\cdot\mathbf{n}_\pm, u_{h,\pm} - u_{h,\Gamma})_\Gamma, \quad I_7 := \tilde{s}(z_h, y_h), \end{aligned}$$

where we used that $[\![\nabla u]\!]_F \cdot \mathbf{n}_F = 0$ in the expression of I_3 and that $[\![\sigma\nabla u]\!]_\Gamma \cdot \mathbf{n}_\Gamma = 0$ in the expression of I_4. We now choose

$$y_{h,\pm} := \Pi_\pm^{h,k^*}(y_\pm), \qquad y_{h,\Gamma} := \Pi_\pm^{h,k_\Gamma^*}(y_\Gamma)$$

and bound the seven terms I_j in terms of $\|\nabla y\|_\Omega$ and $|||(\hat{u} - \hat{u}_h, \hat{z}_h)|||$. Invoking the Cauchy–Schwarz inequality, the Poincaré inequality for y, and the approximation property (3.15) gives (we absorb the constant C^Π in the generic constant C)

$$\begin{aligned} |I_1| &\le C \left(\sum_\pm |\sigma_\pm|\, \|\nabla(\Pi_h^{\mathrm{c}}(\hat{u}_h) - u_{h,\pm})\|_{\Omega_\pm} + \ell_\Omega \mu_{\infty,\pm}\, \|\Pi_h^{\mathrm{c}}(\hat{u}_h) - u_{h,\pm}\|_{\Omega_\pm} \right) \|\nabla y\|_\Omega \\ &\le C\sigma_\sharp^{\frac12} \max\left\{1, |\sigma_\pm|^{-1} h\ell_\Omega\mu_{\infty,\pm}\right\} |\hat{u} - \hat{u}_h|_s\, \|\nabla y\|_\Omega\,, \end{aligned}$$

where the second bound uses that $u_\pm|_\Gamma = u_\Gamma$. Using the Cauchy–Schwarz inequality, the (low-order) approximation properties of $\Pi_\pm^{h,k^*}$ and $\Pi_\pm^{h,k_\Gamma^*}$, and observing that

$1 \le (\gamma_\pm^{\mathrm{LS}})^{-1} \le 2\max\{1, \frac{h^2\ell_\Omega^2\mu_{\infty,\pm}^2}{\sigma_\flat|\sigma_\pm|}\}$ (see (3.3)), we infer that

$$|I_2 + I_3 + I_4| \le C\sigma_\sharp^{\frac12} \max\{1, (\sigma_\flat|\sigma_\pm|)^{-\frac12} h\ell_\Omega\mu_{\infty,\pm}\} |||(\hat u - \hat u_h, 0)||| \, \|\nabla y\|_\Omega .$$

To bound I_5, we write

$$I_5 = \sum_\pm \frac{\lambda_\pm |\sigma_\pm|}{h}(u_{h,\pm} - u_{h,\Gamma} - u_\pm + u_\Gamma, y_\pm - y_\Gamma - y_{h,\pm} + y_{h,\Gamma})_\Gamma,$$

so that, reasoning as above gives (we absorb the factor $\lambda_\sharp^{\frac12}$ in the generic constant C)

$$|I_5| \le C\sigma_\sharp^{\frac12} |||(\hat u - \hat u_h, 0)||| \, \|\nabla y\|_\Omega .$$

From the discrete trace inequality (2.14) (applied now on functions in $V_{h,\pm}^{k^*}$), we have $h\|\nabla y_{h,\pm}\|_\Gamma \le C_\pm^{\mathrm{tr}} \|\nabla y_{h,\pm}\|_{\Omega_\pm}$. Since $I_6 = \sum_\pm (\sigma_\pm \nabla y_{h,\pm}\cdot\mathbf{n}_\pm, u_{h,\pm} - u_{h,\Gamma} - (u_\pm - u_\Gamma))_\Gamma$, we infer that (we absorb the factor $C_\pm^{\mathrm{tr}}$ in the generic constant C)

$$|I_6| \le C\sigma_\sharp^{\frac12} |||(\hat u - \hat u_h, 0)||| \, \|\nabla y\|_\Omega ,$$

where we used the H^1-stability of $\Pi_\pm^{h,k^*}$. Finally, the Cauchy–Schwarz inequality, the Poincaré inequality, and the H^1-stability of $\Pi_\pm^{h,k^*}$ give

$$|I_7| \le C\sigma_\sharp^{\frac12} \max\{1, |\sigma_\pm|^{-\frac12}\ell_\Omega\mu_{\infty,\pm}^{\frac12}\} |||(0, \hat z_h)||| \, \|\nabla y\|_\Omega .$$

Combining the above bounds yields the assertion since $\sigma_\flat \le |\sigma_\pm|$. □

Finally, convergence rates are inferred by proceeding as in the proof of Corollary 3.1.

Corollary 3.2 (Convergence rates for smooth solutions). *Under the assumptions of Theorem 3.3 and assuming that $u_\pm \in H^{k+1}(\mathcal{T}_h^\pm)$, we have*

$$\Big\{ \sum_\pm |\sigma_\pm| \, \|\nabla(u - u_{h,\pm})\|_{\Omega_\pm}^2 \Big\}^{\frac12} \le CC^{\mathrm{app}}C^{\mathrm{E}}h^k \sum_\pm |u_\pm|_{H^{k+1}(\mathcal{T}_h^\pm)}, \tag{3.19}$$

with C^{app} defined in Corollary 3.1.

3.4. *Extension to curved interfaces*

We discuss in this section an extension of our method to cover the case where Γ is curved so that only $\mathrm{dist}(\Gamma, \Gamma_h) \le Ch^2$ can be guaranteed based on the affine triangulation $\mathcal{T}_h$. To improve the geometric accuracy, we assume that we can curve the elements of $\mathcal{T}_h$ by means of a piecewise diffeomorphism $\Theta : \Omega_h \to \tilde\Omega_h$ for $\tilde\Omega_h := \Theta(\Omega_h)$. For the mapped interface $\tilde\Gamma_h := \Theta(\Gamma_h)$, an accuracy of $\mathrm{dist}(\Gamma, \tilde\Gamma_h) \le Ch^{q+1}$, for $q \in \mathbb{N}$, can be expected, provided that the interface Γ is sufficiently smooth and $\Theta = \Theta_h$ is constructed based on a vector-valued finite element space of polynomial

order q. We refer the reader to Ref. 39, Ref. 3 or chapter 13 of Ref. 33 for detailed explanations about the construction and analysis of curved elements.

We remark in particular that this technique has already been analyzed in the context of a unique continuation problem in Ref. 23 using a discretization without hybridization. The interesting new aspect here is that the proof of discrete stability from Ref. 23 fails if a hybrid variable is added. We explain in the remainder of this section the reason for this defect and suggest a possible remedy to achieve stability. The remainder of the error analysis on curved meshes can then be performed along the lines of Ref. 23.

The discretization is based on the deformed mesh $\tilde{\mathcal{T}}_h := \Theta(\mathcal{T}_h)$ for which the spaces $V_{h,\pm}^l$ and $V_{h,\Gamma}^l$ from (2.12)-(2.13) (on which the PDE discretization is based) are replaced by their curved versions

$$\tilde{V}_{h,\pm}^l := \left\{ \tilde{v}_{h,\pm} = v_{h,\pm} \circ \Theta^{-1} \mid v_{h,\pm} \in V_{h,\pm}^l \right\}, \tag{3.20a}$$

$$\tilde{V}_{h,\Gamma}^l := \left\{ \tilde{v}_{h,\Gamma} = v_{h,\Gamma} \circ \Theta^{-1} \mid v_{h,\Gamma} \in V_{h,\Gamma}^l \right\}. \tag{3.20b}$$

The variational formulation is posed on the deformed elements $\tilde{K} := \Theta(K)$ for all $K \in \mathcal{T}_h$ and facets $\tilde{F} := \Theta(F)$ for all $F \in \mathcal{F}_h$ and pulled back to the piecewise affine configuration by using the transformation formula for integrals,

$$\int_{\tilde{K}} \tilde{v}_{h,\pm} \, \mathrm{d}\tilde{x} = \int_{\Theta(K)} v_{h,\pm} \circ \Theta^{-1} \, \mathrm{d}\tilde{x} = \int_K v_{h,\pm} \, |\det(D\Theta_K)| \, \mathrm{d}x, \tag{3.21}$$

where $D\Theta_K$ denotes the Jacobian of $\Theta_K := \Theta|_K$. For a hybridized method, it is of particular interest how the fluxes transform. Let us denote the outer normal vectors of an element $K_\pm \in \mathcal{T}_h^\pm$ by $\mathbf{n}_{h,K_\pm}$ and those of the transformed elements $\tilde{K}_\pm := \Theta(K_\pm)$ by $\tilde{\mathbf{n}}_{h,K_\pm}$. We define the global normal vectors $\tilde{\mathbf{n}}_{h,\pm}$ on $\tilde{\Gamma}_h$ elementwise as $\tilde{\mathbf{n}}_{h,\pm}|_{\tilde{F}} := \tilde{\mathbf{n}}_{h,K_\pm}|_{\tilde{F}}$ for all $\tilde{F} = \tilde{K}_+ \cap \tilde{K}_-$. We have the relations (see, e.g., Lemma 9.11 in Ref. 33): For all $\tilde{x} = \Theta_{K_\pm}(x) \in \partial\tilde{K}_\pm$,

$$\tilde{\mathbf{n}}_{h,K_\pm}(\tilde{x}) = \frac{1}{\left\| ((D\Theta_{K_\pm})^{-\mathrm{T}} \mathbf{n}_{h,K_\pm})(x) \right\|_2} ((D\Theta_{K_\pm})^{-\mathrm{T}} \mathbf{n}_{h,K_\pm})(x), \tag{3.22}$$

and for all $\tilde{x} = \Theta_{K_\pm}(x) \in \tilde{K}_\pm$,

$$\nabla \tilde{v}_{h,\pm}(\tilde{x}) = (D\Theta_{K_\pm})^{-\mathrm{T}}(x) \nabla v_{h,\pm}(x). \tag{3.23}$$

Owing to the presence of the inverse transpose of the Jacobians $D\Theta_{K_\pm}$ in (3.22)-(3.23), we conclude that in general

$$\nabla \tilde{v}_{h,\pm} \cdot \tilde{\mathbf{n}}_{h,\pm} \notin \tilde{V}_{h,\Gamma}^l. \tag{3.24}$$

It is well-known that this poses an issue for hybridized methods. Indeed, point (3) of the proof of Theorem 3.1 crucially hinges on the fact that we can choose the facet variable to control the flux.

To obtain stability on curved meshes, we require an additional stabilization term to compensate for the fact that the facet space does not contain the flux. This

stabilization is given by

$$\sum_{\pm} \frac{h}{\sigma_\sharp} \left(R_\pm(\tilde{v}_{h,\pm}), R_\pm(\tilde{w}_{h,\pm})\right)_{\tilde{\Gamma}_h}, \tag{3.25}$$

where

$$R_\pm(\tilde{v}_{h,\pm}) := \tilde{\Pi}_\Gamma^{h,l}(\sigma_\pm \nabla \tilde{v}_{h,\pm} \cdot \tilde{\mathbf{n}}_{h,\pm}) - \sigma_\pm \nabla \tilde{v}_{h,\pm} \cdot \tilde{\mathbf{n}}_{h,\pm}, \tag{3.26}$$

with $\tilde{\Pi}_\Gamma^{h,l}$ being a suitable (quasi-)interpolation operator into $\tilde{V}_{h,\Gamma}^l$. In practice, it is convenient to choose an L^2-orthogonal projection and incorporate it into the scheme by introducing an additional variable. Indeed, in step (3) of the proof of Theorem 3.1, we now test with $\zeta_\Gamma := \frac{h}{\sigma_\sharp} \tilde{\Pi}_\Gamma^{h,l}(\sum_{\pm} \sigma_\pm \nabla \tilde{v}_{h,\pm} \cdot \tilde{\mathbf{n}}_{h,\pm})$ so that

$$\sum_{\pm} (\sigma_\pm \nabla \tilde{v}_{h,\pm} \cdot \tilde{\mathbf{n}}_{h,\pm}, \zeta_\Gamma)_{\tilde{\Gamma}_h} \geq \frac{h}{2\sigma_\sharp} \left\| \sum_{\pm} \sigma_\pm \nabla \tilde{v}_{h,\pm} \cdot \tilde{\mathbf{n}}_{h,\pm} \right\|_{\tilde{\Gamma}_h}^2 - \frac{h}{\sigma_\sharp} \sum_{\pm} \|R_\pm(\tilde{v}_{h,\pm})\|_{\tilde{\Gamma}_h}^2 . \tag{3.27}$$

The term with the negative sign is then controlled by means of the additional stabilization introduced in (3.25).

4. Numerical experiments

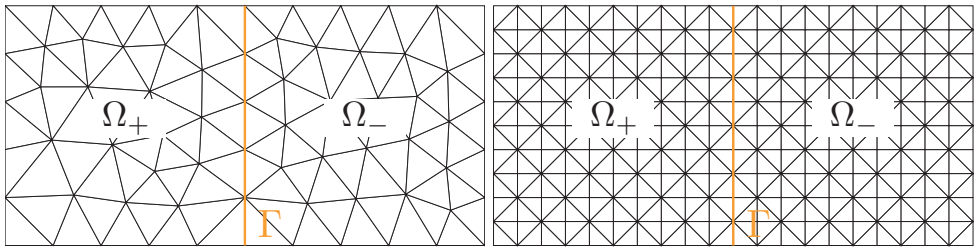
We present numerical experiments to investigate the performance of the proposed method. We start in Sec. 4.1 with an academic test case and proceed in Sec. 4.2 to an acoustic cloaking device proposed in Ref. 44. Finally, in Sec. 4.3, we explore the setting in which the stability Assumption 2 is violated.

The numerical experiments have been implemented in the finite element library `NGSolve`[40,41]. Reproduction material for the presented experiments is available at zenodo in the form of a docker image `https://doi.org/10.5281/zenodo.11067990` (including a document describing the choice of stabilization parameters).

4.1. *Symmetric cavity*

The symmetric cavity problem is one of the main benchmark tests in the numerical analysis literature on problems with sign-changing coefficients (see, e.g., Refs. 27, 1, 28). This is a pure diffusion problem, i.e. $\mu_\pm := 0$, in which the subdomains are given by $\Omega_+ := (-1,0)\times(0,1)$ and $\Omega_- := (0,1)\times(0,1)$. It is known from Section 3.3 of Ref. 27 that Assumption 2 holds true for $\sigma_+/\sigma_- \neq -1$. In this case, the solution is given by

$$u(x,y) := \begin{cases} ((x+1)^2 - (\sigma_+ + \sigma_-)^{-1}(2\sigma_+ + \sigma_-)(x+1))\sin(\pi y) & \text{in } \Omega_+, \\ (\sigma_+ + \sigma_-)^{-1}\sigma_+(x-1)\sin(\pi y) & \text{in } \Omega_-. \end{cases} \tag{4.1}$$



(a) Unstructured mesh (b) Symmetric mesh

Fig. 1: Meshes for the cavity problem.

The corresponding right-hand side $f_\pm := -\nabla \cdot (\sigma_\pm \nabla u_\pm)$ is

$$f(x,y) = \begin{cases} \sigma_+ \left[-2 + \pi^2((x+1)^2 - \frac{2\sigma_+ + \sigma_-}{\sigma_+ + \sigma_-}(x+1))\right] \sin(\pi y) & \text{in } \Omega_+, \\ \pi^2 \frac{\sigma_+ \sigma_-}{\sigma_+ + \sigma_-}(x-1)\sin(\pi y) & \text{in } \Omega_-. \end{cases} \tag{4.2}$$

The problem becomes numerically more difficult to handle if the critical contrast $\sigma_+/\sigma_- = -1$ is approached. Let us first test the hybridized Nitsche method for some contrasts away from the critical value. The relative H^1-errors on a sequence of unstructured meshes, see Fig. 1a for an example, are displayed in Fig. 2. We have taken here the minimal choice for the dual stabilization, $k^* = 1$ and $k_\Gamma^* = k - 1$. Clearly, the convergence rates predicted by Theorem 3.3 are achieved.

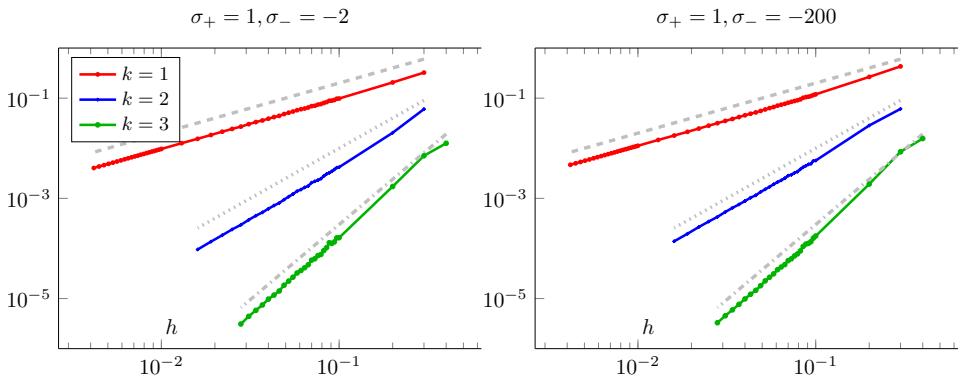


Fig. 2: Relative error $\|u - u_h\|_{H^1(\Omega_+ \cup \Omega_-)} / \|u\|_{H^1(\Omega_+ \cup \Omega_-)}$ for the symmetric cavity in the well-posed case on unstructured meshes.

Now let us approach the critical contrast by setting $\sigma_+ = 1$ and $\sigma_- = -1.001$. It is well-known (see, e.g., Refs. 27 and 1) that a naive Galerkin discretization obtained from the bilinear form $(u,v) \mapsto \int_\Omega \sigma \nabla u \cdot \nabla v$ with $\sigma|_{\Omega_\pm} = \sigma_\pm$, suffers from

instabilities on unstructured meshes, yet yields optimal convergence rates on symmetric meshes of the form shown in Fig. 1b. We compare the performance of our stabilized method with the plain Galerkin discretization in Fig. 3. We observe that the Galerkin method is unstable on unstructured meshes, whereas the stabilized method shows a fairly robust performance and optimal convergence rates. We have taken the full dual order $k^* = k$ and $k^*_\Gamma = k$ for this example since it was observed that this allows to reduce the size of the stabilization parameters without affecting the numerical stability, thereby leading to reduced errors on coarse meshes. Instead, on symmetric meshes, the error for the stabilized method and coarse mesh sizes is slightly higher than that produced by the Galerkin method. As our method contains a built-in Galerkin discretization, it is indeed possible to deactivate the stabilization on symmetric meshes and achieve the same errors as the plain Galerkin method. However, to keep the comparison fair, we did not resort to this device here.

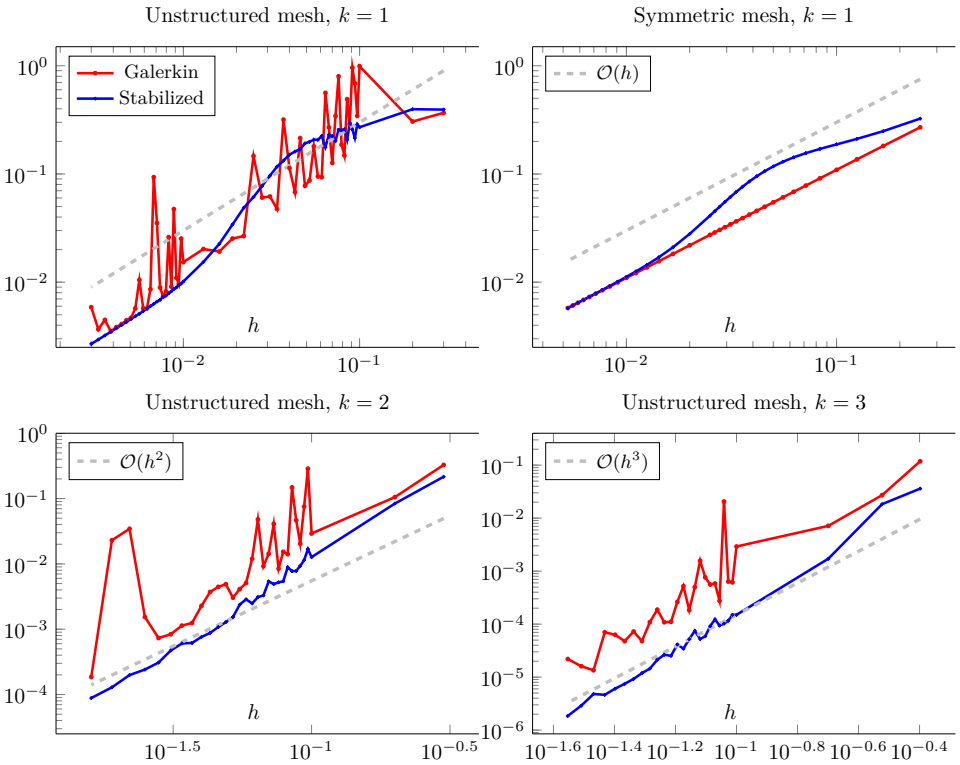


Fig. 3: Relative error $\|u - u_h\|_{H^1(\Omega_+\cup\Omega_-)} / \|u\|_{H^1(\Omega_+\cup\Omega_-)}$ for the symmetric cavity at contrast $\sigma_+/\sigma_- = -1.001$.

4.2. *Metamaterial*

Let us proceed to a more realistic test case. To this end, we consider the acoustic cloaking device from Ref. 44. The equation for a point source at $x_0 \in \mathbb{R}^2$ takes the form

$$-\nabla\cdot(\sigma(r)\nabla u) - \mu(r)u = \delta_{x_0},$$

for a piecewise constant σ and a radially varying μ given by

$$\sigma(r) := \begin{cases} 1/\rho_0 & \text{for } 0 < r < a, \\ 1/\rho_1 & \text{for } a < r < b, \\ -1/\rho_1 & \text{for } b < r < c, \\ 1/\rho_0 & \text{for } c < r, \end{cases} \qquad \mu(r) := \begin{cases} (\omega^2/\kappa_0)(b/a)^4 & \text{for } 0 < r < a, \\ -(\omega^2/\kappa_1)(b/r)^4 & \text{for } a < r < b, \\ \omega^2/\kappa_1 & \text{for } b < r < c, \\ \omega^2/\kappa_0 & \text{for } c < r, \end{cases}$$

where $r := \sqrt{x^2+y^2}$. The parameters are given by:

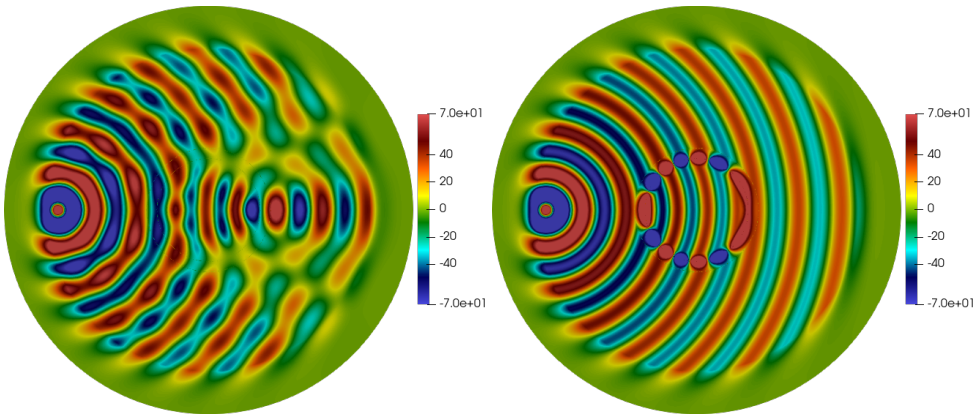
$$\kappa_0 := 2.19 \text{ GPa},\ \kappa_1 := 0.48\kappa_0,\ \rho_0 := 998 \text{ kg/m}^3,\ \rho_1 := \rho_0,$$
$$a := 1.0 \text{ m},\ b := 1.2 \text{ m},\ c := 1.44 \text{ m}.$$

We have the following contrast values at the sign-changing interfaces:

$$\frac{\lim_{r\uparrow a}\sigma(r)}{\lim_{r\downarrow a}\sigma(r)} = 1, \qquad \frac{\lim_{r\uparrow a}\mu(r)}{\lim_{r\downarrow a}\mu(r)} = -\frac{\kappa_1}{\kappa_0} = -0.48,$$
$$\frac{\lim_{r\uparrow b}\sigma(r)}{\lim_{r\downarrow b}\sigma(r)} = -1, \qquad \frac{\lim_{r\uparrow b}\mu(r)}{\lim_{r\downarrow b}\mu(r)} = -1,$$
$$\frac{\lim_{r\uparrow c}\sigma(r)}{\lim_{r\downarrow c}\sigma(r)} = -1, \qquad \frac{\lim_{r\uparrow c}\mu(r)}{\lim_{r\downarrow c}\mu(r)} = \frac{\kappa_0}{\kappa_1} = \frac{1}{0.48}.$$

The source is positioned at $x_0 := (-3.5, 0)$ m, and we truncate the computational domain by a perfectly matched layer (PML) (see, e.g., Refs. 29 and 26) active in the region $\Omega_{\text{PML}} := \{r \in (\tau, \eta)\}$, with $\tau = 3.75$ m and $\eta = 4.75$ m. A sketch of the computational setup is displayed in the upper left panel of Fig. 5. We choose equal polynomial order for the primal and dual variables in this example, as we observed that the constants in the error estimate (3.16) are significantly larger if the polynomial orders are chosen differently. Hence, the additional computational cost incurred by choosing equal polynomial order for the dual variable is compensated for by a significant gain in accuracy.

Before we embark on convergence studies, let us first study the effectiveness of the metamaterial, which is located in the layer $a < r < c$. The idea of this layer is to cloak the object contained in the region $r < a$. Since $\mu(r)$ differs for $r < a$ and $r > c$, we expect to see traces of the object in the propagating waves if no metamaterial is present, i.e., if $\sigma(r) = 1/\rho_0$ and $\mu(r) = \omega^2/\kappa_0$ uniformly for $r > a$. This is confirmed in Fig. 4a which shows the numerical solution computed with the stabilized method using $\omega = 2\pi\cdot 1481.5$ Hz. The waves are indeed strongly disturbed by the inhomogeneity. However, if the cloak is activated, as shown in Fig. 4b, the object becomes invisible.



(a) without cloak (b) with cloak

Fig. 4: Cloaking using a metamaterial. The function values in the cloaking domain are actually very high, but have been truncated here to ± 70 to aid presentation.

If the cloaking worked perfectly, we would expect that the solution in the exterior of the cloaked region $r > c$ be given by a spherical wave emanating from the point source at x_0 given by

$$u = \frac{i\rho_0}{4} H_0^{(1)} \left(\omega \sqrt{\frac{\rho_0}{\kappa_0}} \, \|x - x_0\| \right), \tag{4.3}$$

where $H_0^{(1)}$ denotes the Hankel function of the first kind of order zero. We will measure the convergence of the numerical solution against this reference solution in the two circular regions

$$\Omega_i := \{c < r < 1.7\}, \quad \Omega_e := \{1.7 < r < 3.25\},$$

as sketched in the upper left panel of Fig. 5. Note that Ω_i represents a buffer layer between Ω_e and the interface. The relative H^1-errors for the stabilized and the plain Galerkin methods in the various regions are displayed Fig. 5. We observe that the Galerkin method shows severe instabilities in the region Ω_i.

In contrast, the stabilized method always converges at the optimal rate in both regions. We notice that the material interfaces in this example are curved. For this reason, we have used curved elements and implemented the additional stabilization discussed in Sec. 3.4. We mention that this stabilization was not observed to be necessary though, as optimal convergence rates were obtained even when it was omitted.

Finally, we would like to mention that the convergence of the Galerkin method can even suffer in the far field Ω_e if the PML is not chosen carefully. We define the

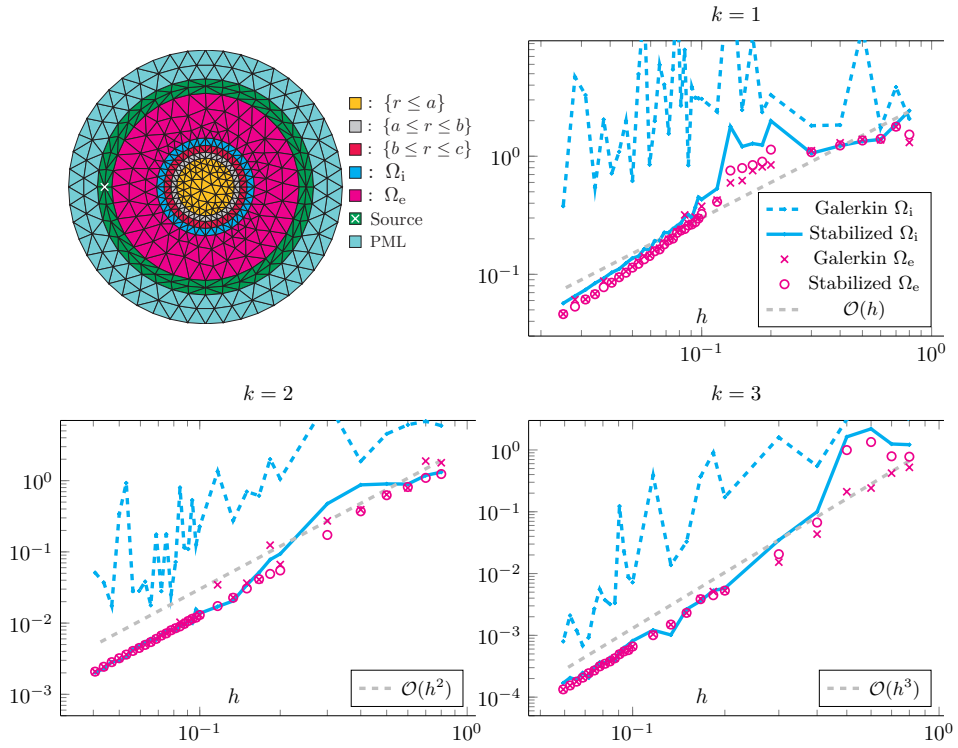


Fig. 5: Upper left panel: computational setup. Other panels: Relative H^1-error in Ω_i and Ω_e with respect to the reference solution (4.3) as a function of the mesh size. Here, h refers to an upper bound on the mesh size in $\{r < a\} \cup \{r > c\}$. The mesh size within the metamaterial $\{a \le r \le c\}$ is bounded from above by $h/3$.

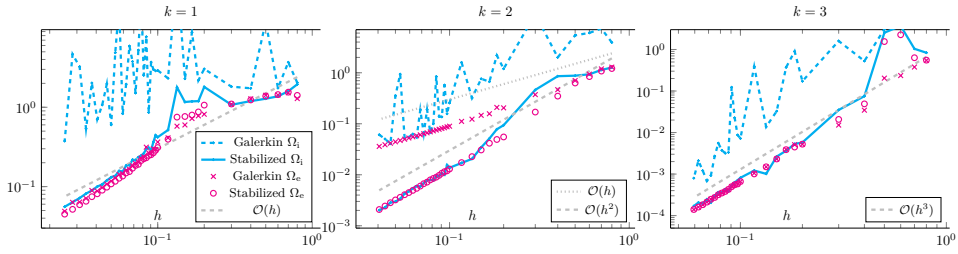


Fig. 6: Relative H^1-error in Ω_i and Ω_e with respect to the reference solution (4.3) when the parameter m in the PML profile (4.4) is chosen as $m = 0$.

complex stretching as

$$r \mapsto r + i\frac{\alpha_m}{m+1}\frac{(r-\tau)^{m+1}}{(\eta-\tau)^m}, \qquad \tau \le r \le \eta, \tag{4.4}$$

for a non-negative integer m which we consider as a free parameter and an amplitude $\alpha_m > 0$. In the experiment shown in Fig. 5, we have chosen $m = 1$ and $\alpha_1 = 4.5$, which leads to optimal convergence of the Galerkin method in Ω_e. However, when choosing $m = 0$ with $\alpha_0 = 1$ (i.e. a linear profile for $\hat{\alpha}$), we observe in Fig. 6 that the Galerkin method for $k = 2$ converges at a reduced rate in Ω_e. Notice that the stabilized method, which uses exactly the same PML, converges optimally. Thus, this behavior seems not to be an issue of the PML by itself, but rather a manifestation of the instability of the Galerkin method for sign-changing problems (on general meshes). These results highlight again the importance of using a reliable method for simulating wave propagation inside metamaterials.

4.3. *Non-symmetric cavity with contrast inside critical interval*

Let us finally consider a test case for which it is known, see Sec. 3.3 of Ref. 27 and Sec. 4.2 of Ref 2, that Assumption 2 fails. As in Sec. 4.1, we set $\Omega_+ := (-1, 0)\times(0, 1)$, but now $\Omega_- := (0, 3)\times(0, 1)$ which breaks the symmetry of the cavity. We set $\mu_\pm = 0$ and consider $\sigma_+ = 1, \sigma_- = -1$. Here, the contrast is inside the critical interval {-1}, so that the problem is not Fredholm, yet the weaker Assumption 1 of uniqueness holds true. We consider the exact solution from Sec. 7.2 of Ref. 2 defined as

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

The corresponding right hand side $f_\pm := -\nabla \cdot (\sigma_\pm \nabla u_\pm)$ is

$$f(x, y) := \begin{cases} \left[-4 + \pi^2(2(x + 1)^2 - 5(x + 1))\right] \sin(\pi y) & \text{in } \Omega_+, \\ -\pi^2(x - 3) \sin(\pi y) & \text{in } \Omega_-. \end{cases}$$

Notice that the theoretical convergence result in the H^1-norm derived in Theorem 3.3 cannot be applied here as it relies on the well-posedness of the continuous problem, but we do have the convergence result in the weaker triple norm as established in Theorem 3.2. The numerical results displayed in Fig. 7 prove the importance of making this distinction. The error in the triple norm $|||(\hat{u}-\hat{u}_h, 0)|||$ converges to zero at the optimal rate, while oscillations and stagnation can be observed in both the H^1- and L^2-norms. The stability improves if we modify the method by adding a penalty on the jump of the second-order derivatives as described in Remark 2.1.

However, despite this modification, we were only able to obtain second-order convergence up to a mesh-size of about $h \approx 0.05$. The right plot in Fig. 7 shows that, on finer meshes, the convergence deteriorates to a logarithmic rate. The plot of the absolute error in Fig. 7 indicates that this behavior seems to stem from a poor approximation close to the interface.

Overall, this example shows that our method can still be applied when the contrast lies inside the critical interval. However, only optimal convergence in the rather weak triple norm is achieved. Obtaining convergence in a stronger norm would apparently require modification of the stabilization. It is, however, neither

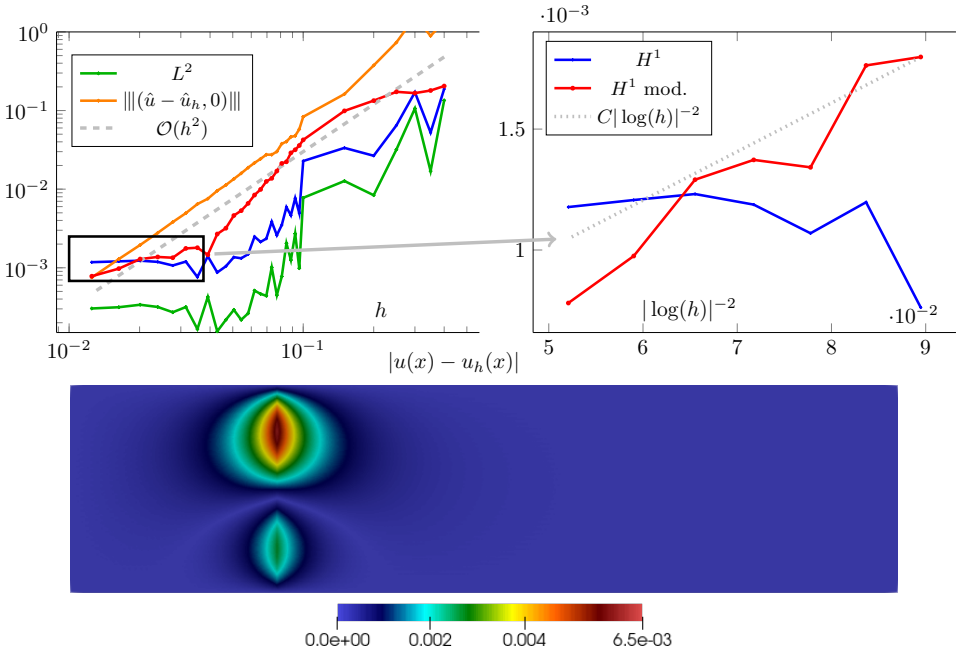


Fig. 7: Results for the non-symmetric cavity with contrast inside the critical interval obtained with the stabilized method using $k = 2$. The red line displays the error in the H^1-norm obtained with a modified method for which we added the additional stabilization discussed in Remark 2.1. All other graphs display the results obtained with the original method. Notice that the x-axis in the right plot is given in terms of $|\log(h)|^{-2}$ and that the error for the modified method looks almost like a straight line in this plot. The figure on the bottom displays the absolute error for the modified method on a mesh of size $h \approx 0.028$.

clear which stabilization to choose nor what type of convergence can be expected, since these factors are determined by the stability properties of the continuous problem which are not known explicitly when the contrast lies inside the critical interval.

5. Conclusion

We presented a stabilized primal-dual finite element method for the numerical approximation of acoustic metamaterials and proved optimal error estimates under a well-posedness assumption on the continuous problem. The method can be applied on general shape-regular meshes and has shown to be reliable and accurate in numerical experiments featuring physically relevant metamaterials. These results motivate to conduct further research on the proposed method. The following extensions seem interesting:

- At the discrete level, the solutions in the subdomains are coupled only via a trace variable defined on the interface Γ. This suggest to solve the linear system efficiently via static condensation, which seems particularly interesting for metamaterials composed of several layers.
- In the analysis and implementation of the method, we assumed that the mesh fits the interface. However, we expect that the method can be extended to unfitted discretizations by combining the techniques from Refs. 14 and 23.
- We also assumed that the solution u of (2.1)-(2.2) fulfills $u_\pm \in H^s(\Omega_\pm)$, $s > \frac{3}{2}$. Recently, an analysis of the primal dual stabilized FEM applied to a unique continuation problem with low regularity solutions ($1 \leq s < \frac{3}{2}$) has been given in Ref. 18. It would be interesting and practically relevant to apply these techniques to the sign-changing problem.
- We restricted our attention to acoustic metamaterials. An extension of the method to Maxwell's equations, which would be required to capture the electromagnetic characteristics of general metamaterials, is certainly of practical relevance.

Acknowledgment

E.B. and J.P. acknowledge funding by EPSRC grant EP/V050400/1.

References

1. A. Abdulle, M. E. Huber and S. Lemaire, An optimization-based numerical method for diffusion problems with sign-changing coefficients, *Comptes Rendus Math.* **355** (2017) 472–478.
2. A. Abdulle and S. Lemaire, An optimization-based method for sign-changing elliptic PDEs, *ESAIM: Math. Model. Numer. Anal.* **58** (2024) 2187–2223.
3. C. Bernardi, Optimal Finite-Element Interpolation on Curved Domains, *SIAM J. Numer. Anal.* **26** (1989) 1212–1240.
4. 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** (2018) 801–838.
5. 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** (2012) 1363–1387.
6. A.-S. Bonnet-Ben Dhia, L. Chesnel and P. Ciarlet, Jr., T-coercivity for the Maxwell problem with sign-changing coefficients, *Comm. Partial Diff. Equ.* **39** (2014) 1007–1031.
7. A.-S. Bonnet-Ben Dhia, L. Chesnel and P. Ciarlet, Jr., Two-dimensional Maxwell's equations with sign-changing coefficients, *Appl. Numer. Math.* **79** (2014) 29–41, workshop on Numerical Electromagnetics and Industrial Applications (NELIA 2011).
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. Comp. Appl. Math.* **234** (2010) 1912–1919, eighth International Conference on Mathematical and Numerical Aspects of Waves (Waves 2007).

9. A.-S. Bonnet-Ben Dhia, M. Dauge and K. Ramdani, Analyse spectrale et singularités d'un problème de transmission non coercif, *C. R. Math. Acad. Sci. Paris* **328** (1999) 717–720.
10. E. Burman, Stabilized finite element methods for nonsymmetric, noncoercive, and ill-posed problems. Part I: elliptic equations, *SIAM J. Sci. Comput.* **35** (2013) A2752–A2780.
11. E. Burman, Error estimates for stabilized finite element methods applied to ill-posed problems, *Comptes Rendus Math.* **352** (2014) 655–659.
12. E. Burman, G. Delay and A. Ern, A hybridized high-order method for unique continuation subject to the Helmholtz equation, *SIAM J. Numer. Anal.* **59** (2021) 2368–2392.
13. E. Burman, G. Delay and A. Ern, The unique continuation problem for the heat equation discretized with a high-order space-time nonconforming method, *SIAM J. Numer. Anal.* **61** (2023) 2534–2557.
14. E. Burman, D. Elfverson, P. Hansbo, M. G. Larson and K. Larsson, Hybridized CutFEM for elliptic interface problems, *SIAM J. Sci. Comput.* **41** (2019) A3354–A3380.
15. E. Burman and A. Ern, Continuous interior penalty *hp*-finite element methods for advection and advection-diffusion equations, *Math. Comp.* **76** (2007) 1119–1140.
16. E. Burman, P. Hansbo and M. G. Larson, Solving ill-posed control problems by stabilized finite element methods: an alternative to Tikhonov regularization, *Inverse Problems* **34** (2018) 035004.
17. E. Burman, M. G. Larson and L. Oksanen, Primal-dual mixed finite element methods for the elliptic Cauchy problem, *SIAM J. Numer. Anal.* **56** (2018) 3480–3509.
18. E. Burman, M. Lu and L. Oksanen, Solving the unique continuation problem for Schrödinger equations with low regularity solutions using a stabilized finite element method, *ESAIM: Math. Model. Numer. Anal.* To appear. URL: https://doi.org/10.1051/m2an/2025060.
19. E. Burman, M. Nechita and L. Oksanen, Unique continuation for the Helmholtz equation using stabilized finite element methods, *J. Math. Pures Appl.* **129** (2019) 1–22.
20. E. Burman, M. Nechita and L. Oksanen, A stabilized finite element method for inverse problems subject to the convection–diffusion equation. I: diffusion-dominated regime, *Numer. Math.* **144** (2020) 451–477.
21. E. Burman, M. Nechita and L. Oksanen, Optimal approximation of unique continuation, *Found. Comput. Math.* **25** (2023) 1025–1045.
22. E. Burman and J. Preuss, Unique continuation for the Lamé system using stabilized finite element methods, *GEM - Int. J. Geomath.* **14** (2023) 9.
23. E. Burman and J. Preuss, Unique continuation for an elliptic interface problem using unfitted isoparametric finite elements, *SMAI J. Comput. Math.* **11** (2025) 165–202.
24. C. Carvalho, L. Chesnel and P. Ciarlet, Jr., Eigenvalue problems with sign-changing coefficients, *Comptes Rendus Math.* **355** (2017) 671–675.
25. T. Chaumont-Frelet and B. Verfürth, A generalized finite element method for problems with sign-changing coefficients, *ESAIM: Math. Model. Numer. Anal.* **55** (2021) 939–967.
26. Z. Chen and X. Liu, An Adaptive Perfectly Matched Layer Technique for Time-harmonic Scattering Problems, *SIAM J. Numer. Anal.* **43** (2005) 645–671.
27. L. Chesnel and P. Ciarlet, Jr., T-coercivity and continuous Galerkin methods: application to transmission problems with sign changing coefficients, *Numer. Math.* **124** (2013) 1–29.
28. 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** (2023) 1316–1339.

29. F. Collino and P. Monk, The Perfectly Matched Layer in curvilinear coordinates, *SIAM J. Sci. Comput.* **19** (1998) 2061–2090.
30. M. Costabel and E. Stephan, A direct boundary integral equation method for transmission problems, *J. Math. Anal. Appl.* **106** (1985) 367–413.
31. S. A. Cummer, J. Christensen and A. Alù, Controlling sound with acoustic metamaterials, *Nature Reviews Materials* **1** (2016) 1–13.
32. H. Egger, A class of hybrid mortar finite element methods for interface problems with non-matching meshes, *preprint AICES-2009-2, Jan* URL: https://www.math.tugraz.at/%7Eherbert/pubs/hybridmortar.pdf.
33. A. Ern and J.-L. Guermond, *Finite Elements I: Approximation and Interpolation*, volume 72 of *Texts in Applied Mathematics* (Springer Nature, Cham, Switzerland, 2021).
34. 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).
35. V. Girault and P.-A. Raviart, *Finite element methods for Navier-Stokes equations: Theory and algorithms*, volume 5 of *Springer Series in Computational Mathematics* (Springer Berlin, Heidelberg, 1986).
36. A. Greenleaf, Y. Kurylev, M. Lassas and G. Uhlmann, Cloaking devices, electromagnetic wormholes, and transformation optics, *SIAM Rev.* **51** (2009) 3–33.
37. M. Halla, On the approximation of dispersive electromagnetic eigenvalue problems in two dimensions, *IMA J. Numer. Anal.* **43** (2021) 535–559.
38. M. Halla, T. Hohage and F. Oberender, A new numerical method for scalar eigenvalue problems in heterogeneous, dispersive, sign-changing materials, 2024, URL: https://arxiv.org/abs/2401.16368.
39. M. Lenoir, Optimal Isoparametric Finite Elements and Error Estimates for Domains Involving Curved Boundaries, *SIAM J. Numer. Anal.* **23** (1986) 562–580.
40. J. Schöberl, NETGEN An advancing front 2D/3D-mesh generator based on abstract rules, *Comput. Vis. Sci.* **1** (1997) 41–52.
41. J. Schöberl, C++11 implementation of finite elements in NGSolve, Technical report, ASC-2014-30, Institute for Analysis and Scientific Computing, September 2014, URL: http://hdl.handle.net/20.500.12708/28346.
42. R. L. Scott and S. Zhang, Finite element interpolation of nonsmooth functions satisfying boundary conditions, *Math. Comp.* **54** (1990) 483–493.
43. G. Unger, Convergence analysis of a Galerkin boundary element method for electromagnetic resonance problems, *Partial Differ. Equ. Appl.* **2** (2021) 39.
44. X. Zhu, B. Liang, W. Kan, X. Zou and J. Cheng, Acoustic cloaking by a superlens with single-negative materials, *Phys. Rev. Lett.* **106** (2011) 014301.