A HYBRIDIZED NITSCHE METHOD FOR SIGN-CHANGING ELLIPTIC PDES

ERIK BURMAN, ALEXANDRE ERN AND JANOSCH PREUSS

Abstract. We present a hybridized Nitsche method for acoustic metamaterials. The use of adapted stabilization terms allows us to cope with the sign-changing nature of the problem and to prove optimal error estimates under a well-posedness assumption. The method can be used on arbitrary shape-regular meshes (fitted to material interfaces) and yields optimal convergence rates when applied to simulate a realistic acoustic cloaking device.

1. Introduction

Acoustic metamaterials have many interesting applications, for example they can be used to cloak objects from incoming sound waves. We refer to the review paper [CCA16] and references therin for a thorough discussion of this subject. When it comes to simulating wave propagation inside metamaterials, one is faced with the challenge to deal with the sign-changing nature of the corresponding 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.

To start with, let us briefly outline the current approaches in the numerical analysis literature which have been proposed to cope with sign-changing coefficient problems. To investigate well-posedness of these problems at the continuous level, the approach of T-coercivity, introduced in [BBDCJZ10], has proved to be fruitful. We refer to [BBDCCJ12] in which scalar problems are studied using this approach and to references [BBDCCJ14a, BBDCCJ14b] for applications to time-harmonic Maxwell problems. It turns out that certain meshing rules must be respected in order to apply the T-coercivity framework. In this case, a standard Galerkin discretization on these meshes is suitable for the numerical solution of sign-changing coefficient problems, see e.g., [CCJ13, BBDCCJ18]. Furthermore, we refer to [CCC17, Hal21] for an application to eigenvalue problems and to [CFV21] for application to multi-scale problems using the framework of localized orthogonal decomposition. Note however, that these meshing rules can be very challenging to implement if the interface at which the sign change occurs is geometrically complicated and are by construction impossible to realize for certain advanced discretization techniques, e.g., geometrically unfitted methods. In the recent preprint [HHO24] an approach has been suggested to avoid these stringent assumption on the mesh. However,

Date: February 8, 2024.

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.

We are aware of two other approaches which can be applied on general meshes and are applicable to some more general settings that cannot be covered using the framework of T-coercivity. An optimization-based method was proposed in [AHL17], see also [AL23], and very recently also an optimal control-based method [CLR23] was devised which overcomes a potentially restrictive regularity condition required in the former reference. These methods can be proven to convergence, yet, it seems that convergence rates have not been proven for either of these methods. Hence, there is clearly the need for more research on this subject.

In this article we propose a stabilized finite element method for the numerical simulation of acoustic metamaterials. We proceed in the spirit of [Bur13] in which a primal-dual stabilized framework is introduced to render the numerical solution of non-coercive or ill-posed problems using the finite element method feasible. This methodology has for example been applied to various unique continuation problems [Bur14, BHL18, BLO18, BNO19, BNO20] for which it leads to optimal error estimates [BNO23] when combined with appropriate conditional stability estimates for the continuous problem. In this article we show how to apply this framework to treat sign-changing coefficient problems and derive optimal error estimates in the H^1-norm under the assumptions that the continuous problem is well-posed and that the solution is sufficiently regular.

This method has first been introduced in reference [Egg09] for a Poisson problem without sign-changes.

Let us briefly distinguish our method from the ones already proposed in the literature. In contrast to some of the methods derived from T-coercivity mentioned above, our approach is applicable on arbitrary shape-regular meshes (which are for simplicity assumed to be fitted to the interface). Even though there already exist methods in the literature [AHL17, CLR23, AL23] which can be applied on such meshes, it seems that convergence rates for these methods have so far not been shown. 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. Note, however, that the methods in references [AHL17, CLR23, AL23] appear to be applicable under weaker conditions than well-posedness. If we lower the assumption on the continuous problem to uniqueness, we merely obtain convergence in a fairly weak norm which does not provide much information of practical interest about the quality of the approximate solution. Thus, the development of numerical methods under weaker assumptions on the continuous problem and the study of their optimal convergence rates is certainly a topic that deserves additional research.

The remainder of this article is structured as follows. We define the hybridized Nitsche method in Section 2 and prove its convergence in Section 3. In Section 4 we conduct numerical experiments to investigate the performance of the method for academic toy problems and an actual metamaterial

proposed in the physics literature. We finish in Section 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.

2.1. **Model problem.** We consider a 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 outer 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 $\tilde{u}$.

To perform the error analysis, we will make one of the following two assumptions on the model problem (2.1)-(2.2). Clearly, Assumption 2 is stronger than Assumption 1.

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

Assumption 2 (Strong stability)**.** *There is* C^{stab} *such that, for all* $f := (f_+, f_-) \in L^2(\Omega)$*, the model problem* (2.1)-(2.2) *admits a unique solution* $u \in H^1_0(\Omega)$ *that fulfills the stability estimate*

$$\|u\|_{H^1(\Omega)} \leqslant C^{\mathrm{stab}} \|f\|_{L^2(\Omega)}. \tag{2.4}$$

⇝ **JP:** *To prove Theorem 3.6, we require Assumption 2 for $f \in H^{-1}(\Omega)$. If the principle part of the operator does not change sign, then there are standard tricks to derive bounds for $f \in H^{-1}(\Omega)$ from the bounds for $f \in L^2(\Omega)$. But I think that for sign-changing coefficients this technique no longer works and we really have to assume that*

$$\|u\|_{H^1(\Omega)} \leqslant C^{\mathrm{stab}} \|f\|_{H^{-1}(\Omega)}.$$

This is also the bound that has been shown in [CCJ13, Section 3.3] *for the symmetric cavity example we treat in Section 4.1.*

Notice that the constant C^{stab} may depend on the coefficients of $\mathcal{L}_\pm$. Generic constants that are independent of the problem parameters will simply be called C in what follows. We will characterize the dependence of the other named constans on the problem parameters explicitly. We define the following functional spaces

$$V_\pm := \{v \in H^1(\Omega_\pm) \mid v|_{\partial\Omega_\pm\setminus\Gamma} = 0\}, \quad V := V_+ \times V_-, \quad V_\Gamma := L^2(\Gamma), \quad \hat{V} := V \times V_\Gamma. \tag{2.5}$$

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 constant $C^{\mathrm{P}} > 0$ (independent of $\mathcal{L}_\pm$) such that, for all $(z, z_\Gamma) \in \hat{V}$, we have*

$$\|z\|_{L^2(\Omega)}^2 = \sum_\pm \|z_\pm\|_{\Omega_\pm}^2 \leqslant C^{\mathrm{P}} \sum_\pm \Big\{ \|\nabla z_\pm\|_{\Omega_\pm}^2 + \ell_\Gamma^{-1} \|z_\pm - z_\Gamma\|_\Gamma^2 \Big\}, \tag{2.6}$$

where ℓ_Γ is a length scale associated with Γ, e.g., its diameter.

Proof. Owing to the Peetre–Tartar theorem (see, e.g., [GR86, Theorem 2.1.3] or [EG21, Lemma A.20]), we infer that there is a constant C such that, for all $z := (z_+, z_-) \in V$,

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

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 that 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. Moreover, the set of all the facets $\mathcal{F}_h$ of $\mathcal{T}_h$ can be partitioned into

$$\mathcal{F}_h = \mathcal{F}_h^{\partial\Omega} \cup \mathcal{F}_h^\Gamma \cup \mathcal{F}_h^+ \cup \mathcal{F}_h^-, \tag{2.7}$$

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

$F = T_1 \cap T_2$ conventionally pointing from T_1 to T_2, and the jump operator across F is to be interpreted as

$$[\![\nabla u_\pm]\!]_F = \nabla u_\pm|_{T_1} - \nabla u_\pm|_{T_2}.$$

To alleviate technicalities, we assume in what follows that $\mathcal{T}_h$ is quasi-uniform and use a single mesh size h. All what is said henceforth extends to shape-regular triangulations by localizing the mesh size in the error analysis.

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:

$$V_{h,\pm}^l := \left\{ v_{h,\pm} \in H^1(\Omega_\pm) \mid v_{h,\pm}|_T \in \mathbb{P}_l(T),\ \forall T \in \mathcal{T}_h^\pm \mid v_{h,\pm}|_{\partial\Omega_\pm \setminus \Gamma} = 0 \right\} \subset V_\pm, \quad l \geq 1, \tag{2.8}$$

and the discontinuous finite element space on the interface:

$$V_{h,\Gamma}^l := \left\{ v_h \in L^2(\Gamma) \mid v_h|_F \in \mathbb{P}_l(F),\ \forall F \in \mathcal{F}_h^\Gamma \right\} \subset V_\Gamma, \quad l \geq 0. \tag{2.9}$$

We will invoke the following trace inequality:

$$h \left\| \nabla z_\pm \right\|_\Gamma^2 \leqslant C_{\mathrm{tr}} \left\| \nabla z_\pm \right\|_{\Omega_\pm}^2, \qquad \forall z_\pm \in V_{h,\Gamma}^l. \tag{2.10}$$

We will also need a version valid for continuous functions:

$$\left\| v \right\|_{\partial T} \leqslant C \left(h^{-1/2} \left\| v \right\|_T + h^{1/2} \left\| \nabla v \right\|_T \right), \quad \forall v \in H^1(T), \quad T \in \mathcal{T}_h. \tag{2.11}$$

For a real number $s \geq 1$, we consider the broken Sobolev spaces $H^s(\mathcal{T}_h^\pm) := \{ v_\pm \in L^2(\Omega_\pm) \mid v_\pm|_T \in H^s(T),\ \forall T \in \mathcal{T}_h^\pm \}$. For later use, we record here the following identity which results from elementwise integration by parts: For all $v_\pm \in V_{h,\pm}^l + H^s(\mathcal{T}_h^\pm)$, $s > \frac{3}{2}$, and all $w_\pm \in V_\pm$,

$$\begin{aligned} (\mathcal{L}_\pm(v_\pm), w_\pm)_{\mathcal{T}_h^\pm} &= (\sigma_\pm \nabla v_\pm, \nabla w_\pm)_{\Omega_\pm} + (\mu_\pm v_\pm, w_\pm)_{\Omega_\pm} \\ &\quad - \sum_{F \in \mathcal{F}_h^\pm} (\sigma_\pm [\![\nabla v_\pm]\!]_F \cdot \mathbf{n}_F, w_\pm)_F - (\sigma_\pm \nabla v_\pm \cdot \mathbf{n}_\pm, w_\pm)_\Gamma. \end{aligned} \tag{2.12}$$

2.3. **Hybridized Nitsche method.** 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. We use a hybridized Nitsche method to discretize (2.1)–(2.2). We consider polynomial degrees $k, k^* \geq 1$ and $k_\Gamma^* \geq 0$ can we take $k_\Gamma = k-1$?.

⇝ **JP:** *Do you mean here projecting the jumps? (Lehrenfeld-Schöberl stabilization)*

We assume that

$$k \geq \max(k^*, k_\Gamma^*), \qquad k_\Gamma^* \geq k-1. \tag{2.13}$$

We define the following bilinear forms on $(V_{h,\pm}^k \times V_{h,\Gamma}^k) \times (V_{h,\pm}^{k^*} \times V_{h,\Gamma}^{k_\Gamma^*})$ (to alleviate the notation, we omit the subscript h for all the arguments): do we need $\tilde{a}$ separately?

⇝ **JP:** *We use $\tilde{a}$ later in the proof of Theorem 3.6 to work around some regularity issues. Alternatively, we could just define it there.*

$$\tilde{a}_\pm[(v_\pm, v_\Gamma);(z_\pm, z_\Gamma)] := (\sigma_\pm \nabla v_\pm, \nabla z_\pm)_{\Omega_\pm} + (\mu_\pm v_\pm, z_\pm)_{\Omega_\pm} - (\sigma_\pm \nabla v_\pm \cdot \mathbf{n}_\pm, z_\pm - z_\Gamma)_\Gamma + \frac{\lambda_\pm |\sigma_\pm|}{h}(v_\pm - v_\Gamma, z_\pm - z_\Gamma)_\Gamma, \tag{2.14a}$$

$$a_\pm[(v_\pm, v_\Gamma);(z_\pm, z_\Gamma)] := \tilde{a}_\pm(v_\pm, v_\Gamma; z_\pm, z_\Gamma) - (\sigma_\pm \nabla z_\pm \cdot \mathbf{n}_\pm, v_\pm - v_\Gamma)_\Gamma, \tag{2.14b}$$

for user-dependent parameters $\lambda_\pm > 0$ to be chosen sufficiently large (see Proposition 3.2), as well as the stabilization bilinear forms

$$s_\pm[(v_\pm, v_\Gamma);(w_\pm, w_\Gamma)] := \sum_{T \in \mathcal{T}_h^\pm} \gamma_{\mathrm{GLS}} h^2 \left(\mathcal{L}_\pm(v_\pm), \mathcal{L}_\pm(w_\pm)\right)_T + \sum_{F \in \mathcal{F}_h^\pm} h\,|\sigma_\pm| \left([\![\nabla v_\pm]\!]_F \cdot \mathbf{n}_F, [\![\nabla w_\pm]\!]_F \cdot \mathbf{n}_F\right)_F + \frac{|\sigma_\pm|}{h}(v_\pm - v_\Gamma, w_\pm - w_\Gamma)_\Gamma, \tag{2.14c}$$

$$s_\pm^*(z_\pm, y_\pm) := \gamma_\pm^* |\sigma_\pm| (\nabla z_\pm, \nabla y_\pm)_{\Omega_\pm} + \tilde{\mu}_\pm (z_\pm, y_\pm)_{\Omega_\pm}, \tag{2.14d}$$

for user-dependent parameters $\gamma_{\mathrm{GLS}} > 0, \gamma_-^* > 0$ and $\gamma_+^* \geqslant 0$ (see again Proposition 3.2), and where $\tilde{\mu}_\pm := \|\mu_\pm^\ominus\|_{L^\infty(\Omega_\pm)}$ with the negative part operator $\mu_\pm^\ominus := \frac{1}{2}(|\mu_\pm| - \mu_\pm)$. Notice that $\tilde{\mu}_\pm = 0$ if $\mu_\pm \geq 0$. For later purposes let us also define $\mu_\pm^\oplus := \frac{1}{2}(|\mu_\pm| + \mu_\pm)$ and note that $\mu_\pm = \mu_\pm^\oplus - \mu_\pm^\ominus$ holds.

To allow for a more compact notation, we define the discrete spaces

$$\hat{V}_h := (V_{h,+}^k \times V_{h,-}^k) \times V_{h,\Gamma}^k, \qquad \hat{V}_h^* := (V_{h,+}^{k^*} \times V_{h,-}^{k^*}) \times V_{h,\Gamma}^{k^*_\Gamma}, \tag{2.15}$$

and we use the notation $\hat{v}_h := (v_h, v_{h,\Gamma})$, $v_h := (v_{h,+}, v_{h,-})$ for a generic element of $\hat{V}_h$ and $\hat{z}_h := (z_h, z_{h,\Gamma})$, $z_h := (z_{h,+}, z_{h,-})$ for a generic element of $\hat{V}_h^*$. We define the following bilinear forms:

$$\tilde{a}[\hat{v}, \hat{z}] := \sum_\pm \tilde{a}_\pm[(v_\pm, v_\Gamma);(z_\pm, z_\Gamma)], \qquad a[\hat{v}, \hat{z}] := \sum_\pm a_\pm[(v_\pm, v_\Gamma);(z_\pm, z_\Gamma)], \tag{2.16a}$$

as well as

$$s[\hat{v}, \hat{w}] := \sum_\pm s_\pm[(v_\pm, v_\Gamma);(w_\pm, w_\Gamma)], \qquad s^*[z, y] := \sum_\pm s_\pm^*(z_\pm; y_\pm), \tag{2.16b}$$

for all $\hat{v} \in \hat{V}_h$, all $\hat{w} \in \hat{V}_h$, all $\hat{z} \in \hat{V}_h^*$, and all $\hat{y} \in \hat{V}_h^*$. Putting everything together, we define the bilinear form

$$B[(\hat{v}, \hat{z});(\hat{w}, \hat{y})] := a[\hat{w}, \hat{z}] + a[\hat{v}, \hat{y}] + s[\hat{v}, \hat{w}] - s^*(z, y). \tag{2.17}$$

The discrete problem consists of finding $(\hat{u}_h, \hat{z}_h) \in \hat{V}_h \times \hat{V}_h^*$ such that, for all $(\hat{w}_h, \hat{y}_h) \in \hat{V}_h \times \hat{V}_h^*$,

$$B[(\hat{u}_h, \hat{z}_h);(\hat{w}_h, \hat{y}_h)] = \sum_\pm \Big\{ (f_\pm, y_{h,\pm})_{\Omega_\pm} + \sum_{T \in \mathcal{T}_h^\pm} \gamma_{\mathrm{GLS}} h^2 (f_\pm, \mathcal{L}_\pm(w_\pm))_T \Big\}. \tag{2.18}$$

Remark 1 (Stabilization)**.** *It can be useful to scale some of the stabilization terms by some mesh-independent constants to enhance the preasymptotic convergence rate in the numerical experiments. However, for the numerical analysis, these scalings are of no importance and therefore set to unity.*

3. Error analysis

The error analysis proceeds in three steps. In Section 3.1, we prove the (discrete) inf-sup stability of the bilinear form B in a weak triple norm $|||\cdot|||$. A convergence result in this norm is derived in Section 3.2. The first two steps only require uniqueness of the solution to the continuous problem (2.1)-(2.2), i.e., the weak stability Assumption 1. To derive an error estimate in the H^1-norm, we make the stronger stability Assumption 2. This is done in Section 3.3, where, in particular, we derive optimal convergence rates on the H^1-error for sufficiently smooth solutions.

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.,

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

Furthermore, letting $\sigma_\sharp := \max(\sigma_+,-\sigma_-)$ and $\sigma_\flat := \min(\sigma_+,-\sigma_-)$ we define the following triple norm on $\hat{V}_h\times\hat{V}_h^*$:

$$|||(\hat{v},\hat{z})|||^2 := |\hat{v}|_s^2+\sigma_\sharp^{-1}h\,\|[\![\sigma\nabla v]\!]_\Gamma\cdot\mathbf{n}_\Gamma\|_\Gamma^2 \tag{3.2}$$

$$+\sum_{\pm}\Big\{|\sigma_\pm|\,\|\nabla z_\pm\|_{\Omega_\pm}^2+\tilde{\mu}_\pm\|z_\pm\|_{\Omega_\pm}^2+\frac{|\sigma_\pm|}{h}\,\|z_\pm-z_\Gamma\|_\Gamma^2\Big\}. \tag{3.3}$$

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 $|||(\hat{v},\hat{z})|||=0$. We need to prove that $v_\pm=0$, $v_\Gamma=0$, $z_\pm=0$, and $z_\Gamma=0$. 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 $F\in\mathcal{F}_h^\pm$, $v_+=v_\Gamma=v_-$ on Γ, as well as $\nabla z_\pm=0$ in $\Omega_\pm$ and $z_+=z_\Gamma=z_-$ on Γ. Since $\hat{z}\in\hat{V}_h^*\subset\hat{V}$, the Poincaré inequality from Lemma 2.1 readily gives $z_\pm=0$, and thus $z_\Gamma=0$. Let us now deal with $\hat{v}$. Since $v_+=v_\Gamma=v_-$, we infer that $[\![v]\!]_\Gamma=0$. Moreover, we also have $[\![\sigma\nabla v]\!]_\Gamma\cdot\mathbf{n}_\Gamma=0$ by definition of the triple norm. An application of the integration by parts formula (2.12) using (3.1) shows that for any $w:=(w_+,w_-)\in V$, we have

$$(\sigma_\pm\nabla v_\pm,\nabla w)_{\Omega_\pm}+(\mu_\pm v_\pm,w)_{\Omega_\pm}=0.$$

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. □

For the sake of simplicity, we assume that the mesh is fine enough so that (to discuss if we keep or not)

$$|\mu_\pm| h^2 \le |\sigma_\pm|. \tag{3.4}$$

Invoking inverse inequalities and the Poincaré inequality from Lemma 2.1 together with (3.4) (and since $h \le \ell_\Gamma$) to bound the reaction term, we infer that there is C^s so that, for all $\hat z \in \hat V_h^*$,

$$|\hat z|_s^2 \leqslant C^s \sum_\pm |\sigma_\pm| \Big\{ \|\nabla z_\pm\|_{\Omega_\pm}^2 + \frac{1}{h} \|z_\pm - z_\Gamma\|_\Gamma^2 \Big\}. \tag{3.5}$$

Indeed, we have using assumption (3.4)

$$\begin{aligned}
\sum_\pm \sum_{T\in\mathcal{T}_h^\pm} \gamma_{\mathrm{GLS}} h^2 \|\mathcal{L}_\pm(z_\pm)\|_T^2 &\leqslant 2 \sum_\pm \sum_{T\in\mathcal{T}_h^\pm} \gamma_{\mathrm{GLS}} \left(h^2 |\sigma_\pm|^2 \|\nabla\cdot(\nabla u_\pm)\|_T^2 + \|h\mu_\pm z_\pm\|_T^2 \right) \\
&\leqslant C \sum_\pm \gamma_{\mathrm{GLS}} \left(|\sigma_\pm|^2 \|\nabla z_\pm\|_{\Omega_\pm}^2 + |\sigma_\pm| \|\mu_\pm\|_{L^\infty(\Omega_\pm)} \|z_\pm\|_{\Omega_\pm}^2 \right) \\
&\leqslant C \sum_\pm \gamma_{\mathrm{GLS}} \sigma_\sharp \left[1 + \sigma_\flat^{-1} \max_\pm \|\mu_\pm\|_{L^\infty(\Omega_\pm)} \right] \left(|\sigma_\pm| \|\nabla z_\pm\|_{\Omega_\pm}^2 + \sigma_\flat \|z_\pm\|_{\Omega_\pm}^2 \right),
\end{aligned}$$

which is bounded by the right hand side of (3.5) upon choosing

$$\gamma_{\mathrm{GLS}} = \sigma_\sharp^{-1} \left[1 + \sigma_\flat^{-1} \max_\pm \|\mu_\pm\|_{L^\infty(\Omega_\pm)} \right]^{-1} \tag{3.6}$$

and using that

$$\sigma_\flat \sum_\pm \|z_\pm\|_{\Omega_\pm}^2 \leqslant \sum_\pm |\sigma_\pm| \Big\{ \|\nabla z_\pm\|_{\Omega_\pm}^2 + \frac{1}{h} \|z_\pm - z_\Gamma\|_\Gamma^2 \Big\}.$$

holds thanks to Lemma 2.1 and $\sigma_\flat \leqslant |\sigma_\pm|$.

⇝ **JP:** *Since the coefficients appear quadratically in the GLS term, assumption (3.4) is not sufficient to completely get rid of the reaction term. To ensure that C^s and thus the inf-sup constant C^B are independent of the coefficients, I scaled the GLS-term with γ_{GLS}. However, this parameter will later appear in the constants C^E of Theorem 3.6 and C^{app} of Lemma 3.7.*

Proposition 3.2 (inf-sup stability)**.** *Assume that the polynomial degrees satisfy* (2.13) *and that the mesh size satisfies* (3.4)*. Assume that $\lambda_\pm \ge 2C_{\mathrm{tr}} + \frac12$, $\gamma_+^* = 0$, and $\gamma_-^* = 1$. The following holds:*

$$\inf_{(\hat v,\hat z)\in \hat V_h\times\hat V_h^*} \sup_{(\hat w,\hat y)\in \hat V_h\times\hat V_h^*} \frac{B[(\hat v,\hat z);(\hat w,\hat y)]}{|||(\hat v,\hat z)|||\,|||(\hat w,\hat y)|||} \geqslant C^B > 0, \tag{3.7}$$

where

$$C^B = \frac{1}{4\sqrt{2}\sqrt{\alpha^2 + C^s + 2C_{\mathrm{tr}} + 1}}, \quad \alpha \ge \max(2, C^s + \tfrac{2\lambda_\sharp^2}{3} + \frac14), \quad \lambda_\sharp := \max(\lambda_+, \lambda_-). \tag{3.8}$$

Proof. (1) Let $\alpha > 0$ to be chosen large enough later on. We have

$$B[(\hat v,\hat z);(\alpha\hat v,-\alpha\hat z)] = \alpha s[\hat v,\hat v] + \alpha s^*[z,z] = \alpha|\hat v|_s^2 + \sum_\pm \alpha \Big\{ \gamma_\pm^* |\sigma_\pm| \|\nabla z_\pm\|_{\Omega_\pm}^2 + \tilde\mu_\pm \|z_\pm\|_{\Omega_\pm}^2 \Big\}. \tag{3.9}$$

(2) Since $k \geqslant \max\{k^*, k_\Gamma^*\}$ owing to (2.13), 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, classical manipulations invoking the trace inequality (2.10) and Young's inequality lead to

$$\begin{aligned} a_\pm[(z_\pm,z_\Gamma);(z_\pm,z_\Gamma)] &\geq \sigma_\pm\|\nabla z_\pm\|_{\Omega_\pm}^2 + (\mu_\pm z_\pm, z_\pm)_{\Omega_\pm} + \frac{\lambda_\pm|\sigma_\pm|}{h}\|z_\pm - z_\Gamma\|_\Gamma^2 \\ &\quad - 2|\sigma_\pm|C_{\rm tr}^{\frac12}h^{-\frac12}\|\nabla z_\pm\|_{\Omega_\pm}\|z_\pm - z_\Gamma\|_\Gamma \\ &\geq \sigma_\pm'\|\nabla z_\pm\|_{\Omega_\pm}^2 + (\mu_\pm z_\pm, z_\pm)_{\Omega_\pm} + (\lambda_\pm - 2C_{\rm tr})\frac{|\sigma_\pm|}{h}\|z_\pm - z_\Gamma\|_\Gamma^2, \end{aligned}$$

with $\sigma_+' := \frac12\sigma_+$ and $\sigma_-' := \frac32\sigma_-$. Summing over both subdomains, this gives

$$a[\hat{z},\hat{z}] \geq \sum_\pm \Big\{\sigma_\pm'\|\nabla z_\pm\|_{\Omega_\pm}^2 + (\mu_\pm z_\pm, z_\pm)_{\Omega_\pm} + (\lambda_\pm - 2C_{\rm tr})\frac{|\sigma_\pm|}{h}\|z_\pm - z_\Gamma\|_\Gamma^2\Big\}. \tag{3.10}$$

On the other hand, owing to Young's inequality and the estimate (3.5), we infer that

$$\begin{aligned} s[\hat{v},\hat{z}] &\geq -C^s|\hat{v}|_s^2 - \frac{1}{4C^s}|\hat{z}|_s^2 \\ &\geq -C^s|\hat{v}|_s^2 - \frac14\sum_\pm|\sigma_\pm|\Big\{\|\nabla z_\pm\|_{\Omega_\pm}^2 + \frac1h\|z_\pm - z_\Gamma\|_\Gamma^2\Big\}. \end{aligned}$$

Taking into account (3.10), this gives

$$B[(\hat{v},\hat{z});(\hat{z},0)] \geq -C^s|\hat{v}|_s^2 + \sum_\pm\Big\{\sigma_\pm''\|\nabla z_\pm\|_{\Omega_\pm}^2 + (\mu_\pm z_\pm, z_\pm)_{\Omega_\pm} + (\lambda_\pm - 2C_{\rm tr} - \tfrac14)\frac{|\sigma_\pm|}{h}\|z_\pm - z_\Gamma\|_\Gamma^2\Big\}, \tag{3.11}$$

with $\sigma_+'' := \frac14\sigma_+$ and $\sigma_-'' := \frac74\sigma_-$.

(3) Since $k_\Gamma^* \geq k-1$ owing to (2.13), it is legitimate to test with $\hat{\zeta} := (0,\zeta_\Gamma)$ with $\zeta_\Gamma := \sigma_\sharp^{-1}h[\![\sigma\nabla v]\!]_\Gamma \cdot \mathbf{n}_\Gamma$. Using the Cauchy–Schwarz inequality followed by Young's inequality and recalling that $\sigma_\sharp := \max(\sigma_+,-\sigma_-)$, this gives

$$\begin{aligned} B[(\hat{v},\hat{z});(0,\hat{\zeta})] = a[\hat{v},\hat{\zeta}] &= \sum_\pm\Big\{(\sigma_\pm\nabla v_\pm\cdot\mathbf{n}_\pm,\zeta_\Gamma)_\Gamma - \frac{\lambda_\pm|\sigma_\pm|}{h}(v_\pm - v_\Gamma,\zeta_\Gamma)_\Gamma\Big\} \\ &\geq \frac14\sigma_\sharp^{-1}h\|[\![\sigma\nabla v]\!]_\Gamma\cdot\mathbf{n}_\Gamma\|_\Gamma^2 - \sum_\pm\frac{2\lambda_\pm^2}{3}\frac{|\sigma_\pm|}{h}\|v_\pm - v_\Gamma\|_\Gamma^2. \end{aligned} \tag{3.12}$$

(4) Combining (3.9), (3.11), and (3.12), we infer that

$$\begin{aligned} B[(\hat{v},\hat{z});(\hat{w},\hat{y})] &\geq (\alpha - C^s - \tfrac{2\lambda_\sharp^2}{3})|\hat{v}|_s^2 + \frac14\sigma_\sharp^{-1}h\|[\![\sigma\nabla v]\!]_\Gamma\cdot\mathbf{n}_\Gamma\|_\Gamma^2 \\ &\quad + \sum_\pm\Big\{(\alpha\gamma_\pm^*|\sigma_\pm| + \sigma_\pm'')\|\nabla z_\pm\|_{\Omega_\pm}^2 + \alpha\tilde{\mu}_\pm\|z_\pm\|_{\Omega_\pm}^2 + (\mu_\pm z_\pm, z_\pm)_{\Omega_\pm}\Big\} \\ &\quad + \sum_\pm\Big\{(\lambda_\pm - 2C_{\rm tr} - \tfrac14)\frac{|\sigma_\pm|}{h}\|z_\pm - z_\Gamma\|_\Gamma^2\Big\}, \end{aligned}$$

with $\hat{w} := \alpha\hat{v} + \hat{z}$, $\hat{y} := -\alpha\hat{z} + \hat{\zeta}$, and $\lambda_\sharp := \max(\lambda_+, \lambda_-)$. Taking $\alpha \geq \frac{5}{4}$ ensures that $\alpha\tilde{\mu}_\pm\|z_\pm\|^2_{\Omega_\pm} + (\mu_\pm z_\pm, z_\pm)_{\Omega_\pm} \geq \frac{1}{4}\tilde{\mu}_\pm\|z_\pm\|^2_{\Omega_\pm} \geq 0$. Therefore, under this condition, we obtain

$$B[(\hat{v},\hat{z});(\hat{w},\hat{y})] \geq (\alpha - C^s - \tfrac{2\lambda_\sharp^2}{3})|\hat{v}|_s^2 + \frac{1}{4}\sigma_\sharp^{-1}h\|[\![\sigma\nabla v]\!]_\Gamma \cdot \mathbf{n}_\Gamma\|^2_\Gamma \\ + \sum_\pm \Big\{(\alpha\gamma_\pm^*|\sigma_\pm| + \sigma_\pm'')\|\nabla z_\pm\|^2_{\Omega_\pm} + \tfrac{1}{4}\tilde{\mu}_\pm\|z_\pm\|^2_{\Omega_\pm} + (\lambda_\pm - 2C_{\mathrm{tr}} - \tfrac{1}{4})\frac{|\sigma_\pm|}{h}\|z_\pm - z_\Gamma\|^2_\Gamma\Big\}.$$

We choose $\lambda_\pm = \lambda_\sharp := 2C_{\mathrm{tr}} + \frac{1}{2}$ and require that $\alpha \geq C^s + \frac{2\lambda_\sharp^2}{3} + \frac{1}{4}$. This gives

$$B[(\hat{v},\hat{z});(\hat{w},\hat{y})] \geq \tfrac{1}{4}|\hat{v}|_s^2 + \frac{1}{4}\sigma_\sharp^{-1}h\|[\![\sigma\nabla v]\!]_\Gamma \cdot \mathbf{n}_\Gamma\|^2_\Gamma \\ + \sum_\pm \Big\{(\alpha\gamma_\pm^*|\sigma_\pm| + \sigma_\pm'')\|\nabla z_\pm\|^2_{\Omega_\pm} + \tfrac{1}{4}\tilde{\mu}_\pm\|z_\pm\|^2_{\Omega_\pm} + \tfrac{1}{4}\frac{|\sigma_\pm|}{h}\|z_\pm - z_\Gamma\|^2_\Gamma\Big\}.$$

Finally, we take $\gamma_+^* = 0$, $\gamma_-^* = 1$ and require that $\alpha \geq 2$. This gives $\alpha\gamma_\pm^*|\sigma_\pm| + \sigma_\pm'' \geq \frac{1}{4}|\sigma_\pm|$. In conclusion, whenever $\alpha \geq \max(2, C^s + \frac{2\lambda_\sharp^2}{3} + \frac{1}{4})$, we conclude that

$$B[(\hat{v},\hat{z});(\hat{w},\hat{y})] \geq \tfrac{1}{4}|||(\hat{v},\hat{z})|||^2.$$

Since (see Appendix A for the calculation)

$$|||(\hat{w},\hat{y})|||^2 \leqslant 2(\alpha^2 + C^s + 2C_{\mathrm{tr}} + 1)|||(\hat{v},\hat{z})|||^2,$$

this completes the proof and shows that C^B is given by (3.8). □

I STOPPED HERE!

3.2. **Convergence in the triple norm.** We define a strengthened triple norm

$$|||(\hat{v},\hat{z})|||^2_\square = |||(\hat{v},\hat{z})|||^2 + \sum_\pm \Big\{|\sigma_\pm|\,\|\nabla v_\pm\|^2_{\Omega_\pm} + h|\sigma_\pm|\,\|\nabla v_\pm \cdot \mathbf{n}_\Gamma\|^2_\Gamma + \big[\tilde{\mu}_\pm + \sigma_\flat^{-1}\,\big\|\mu_\pm^\oplus\big\|^2_{L^\infty(\Omega_\pm)}\big]\,\|v_\pm\|^2_{\Omega_\pm}\Big\} \tag{3.13}$$

to show continuity of the bilinear form $a[\cdot,\cdot]$ representing the PDE constraint.

Lemma 3.3. *For $\hat{v} = (v, v_\Gamma)$ with $v_\pm \in V^l_{h,\pm} + H^2(\mathcal{T}_h^\pm)$ and $v_\Gamma \in L^2(\Gamma)$ we have for any $\hat{z}_h \in \hat{V}_h^*$ that*

$$a[\hat{v},\hat{z}_h] \leqslant C|||(\hat{v},0)|||_\square|||(0,\hat{z}_h)|||, \tag{3.14}$$

for a constant C independent of h and the coefficients of $\mathcal{L}$.

Proof. Since $\hat{z}_h \in \hat{V}_h^*$ we can use the trace inequality (2.10) to obtain

$$\sum_\pm(\sigma_\pm\nabla z_{h,\pm}\cdot\mathbf{n}_\pm, v_\pm - v_\Gamma)_\Gamma \leqslant C_{\mathrm{tr}}^{1/2}\sum_\pm \|\nabla z_{h,\pm}\|_{\Omega_\pm}\,|\sigma_\pm|\,h^{-1/2}\,\|v_\pm - v_\Gamma\|_\Gamma \leqslant C_{\mathrm{tr}}^{1/2}|||(\hat{v},0)|||_\square|||(0,\hat{z}_h)|||.$$

Using the splitting $\mu_\pm = \mu_\pm^\oplus - \mu_\pm^\ominus$, we estimate the reaction term as follows:

$$\sum_\pm(\mu_\pm v_\pm, z_{h,\pm})_{\Omega_\pm} \leqslant I_1 + I_2,$$

where

$$I_1 = \sum_{\pm} \tilde{\mu}_{\pm} \|v_{\pm}\|_{\Omega_{\pm}} \|z_{h,\pm}\|_{\Omega_{\pm}} \leqslant \left(\sum_{\pm} \tilde{\mu}_{\pm} \|v_{\pm}\|^2_{\Omega_{\pm}} \right)^{1/2} |||(0, \hat{z}_h)|||.$$

Using the Poincaré inequality from Lemma 2.1 and $\sigma_\flat \leqslant |\sigma_\pm|$ we can estimate the second term:

$$\begin{aligned} I_2 &= \sum_{\pm} \left\|\mu^{\oplus}_{\pm}\right\|_{L^\infty(\Omega_\pm)} \|v_{\pm}\|_{\Omega_{\pm}} \|z_{h,\pm}\|_{\Omega_{\pm}} \leqslant \left(\sum_{\pm} \sigma_\flat^{-1} \left\|\mu^{\oplus}_{\pm}\right\|^2_{L^\infty(\Omega_\pm)} \|v_{\pm}\|^2_{\Omega_{\pm}} \right)^{1/2} \left(\sigma_\flat \sum_{\pm} \|z_{h,\pm}\|^2_{\Omega_{\pm}} \right)^{1/2} \\ &\leqslant \left(\sum_{\pm} \sigma_\flat^{-1} \left\|\mu^{\oplus}_{\pm}\right\|^2_{L^\infty(\Omega_\pm)} \|v_{\pm}\|^2_{\Omega_{\pm}} \right)^{1/2} (C^{\mathrm{P}})^{1/2} \left(\sum_{\pm} \sigma_\flat \Big\{ \|\nabla z_{h,\pm}\|^2_{\Omega_{\pm}} + h^{-1} \|z_{h,\pm} - z_{h,\Gamma}\|^2_{\Gamma} \Big\} \right)^{1/2} \\ &\leqslant \left(\sum_{\pm} \sigma_\flat^{-1} \left\|\mu^{\oplus}_{\pm}\right\|^2_{L^\infty(\Omega_\pm)} \|v_{\pm}\|^2_{\Omega_{\pm}} \right)^{1/2} (C^{\mathrm{P}})^{1/2} |||(0, \hat{z}_h)|||, \end{aligned}$$

The remaining terms are easily bounded using the Cauchy-Schwarz inequality. □

Recall that for $\hat{u} = (u, u_\Gamma)$ with u solving (2.1)-(2.2) and u_Γ the trace of u on Γ we have consistency

$$a[\hat{u}, \hat{y}_h] = \sum_{\pm} (\mathcal{L} u_{\pm}, y_{h,\pm})_{\Omega_\pm} + (\llbracket \sigma \nabla u \rrbracket_\Gamma \cdot \mathbf{n}_\Gamma, y_{h,\Gamma})_\Gamma = \sum_{\pm} (f_{\pm}, y_{h,\pm})_{\Omega_\pm}, \quad \hat{y}_h \in V_h^*. \tag{3.15}$$

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 $|||(\cdot,\cdot)|||_\square$.

Proposition 3.4. *Let $u \in H^1_0(\Omega) \cap [H^2(\Omega_+) \times H^2(\Omega_-)]$ solve (2.1)-(2.2) and denote $\hat{u} = (u, u_\Gamma)$ for $u_\Gamma = u|_\Gamma$. Let $(\hat{u}_h, \hat{z}_h) \in \hat{V}_h \times \hat{V}_h^*$ be the solution of the variational problem (2.18). Then we have*

$$|||(\hat{u} - \hat{u}_h, \hat{z}_h)||| \leqslant \left(1 + \frac{C}{C^B}\right) \inf_{\hat{v} \in \hat{V}_h} |||(\hat{u} - \hat{v}_h, 0)|||_\square. \tag{3.16}$$

Proof. Let $\hat{v}_h \in \hat{V}_h$ and $(\hat{w}_h, \hat{y}_h) \in \hat{V}_h \times \hat{V}_h^*$ be arbitrary. Then using (2.18), consistency and that u is sufficiently smooth to fulfill

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

we obtain

$$\begin{aligned} &B[(\hat{u}_h - \hat{v}_h, \hat{z}_h); (\hat{w}_h, \hat{y}_h)] = B[(\hat{u}_h, \hat{z}_h); (\hat{w}_h, \hat{y}_h)] - a[\hat{v}_h, \hat{y}_h] - s[\hat{v}_h, \hat{w}_h] \\ &= \sum_{\pm} \Big\{ (f_\pm, y_{h,\pm})_{\Omega_\pm} + \sum_{T \in \mathcal{T}_h^{\pm}} \gamma_{\mathrm{GLS}} h^2 (f_\pm, \mathcal{L}_\pm(w_{h,\pm}))_T \Big\} - a[\hat{v}_h, \hat{y}_h] - s[\hat{v}_h, \hat{w}_h] \\ &= a[\hat{u} - \hat{v}_h, \hat{y}_h] + s[\hat{u} - \hat{v}_h, \hat{w}_h]. \end{aligned}$$

Using Lemma 3.3 to control the first term yields

$$\begin{aligned} a[\hat{u} - \hat{v}_h, \hat{y}_h] + s[\hat{u} - \hat{v}_h, \hat{w}_h] &\leqslant C \left(|||(\hat{u} - \hat{v}_h, 0)|||_\square |||(0, \hat{y}_h)||| + |||(\hat{u} - \hat{v}_h, 0)||| |||(\hat{w}_h, 0)||| \right) \\ &\leqslant C |||(\hat{u} - \hat{v}_h, 0)|||_\square |||(\hat{w}_h, \hat{y}_h)|||. \end{aligned}$$

In view of the inf-sup condition from (3.7), this implies

$$|||(\hat{u}_h - \hat{v}_h, \hat{z}_h)||| \leqslant \frac{C}{C^B} |||(\hat{u} - \hat{v}_h, 0)|||_\square.$$

The claim then follows from the triangle inequality

$$|||(\hat{u} - \hat{u}_h, \hat{z}_h)||| \leqslant |||(\hat{u} - \hat{v}_h, 0)||| + |||(\hat{v}_h - \hat{u}_h, \hat{z}_h)|||,$$

combined with the fact that $\hat{v}_h \in \hat{V}_h$ was arbitrary. □

3.3. **Convergence in H^1.** To derive convergence rates in the H^1-norm, we will assume well-posedness of the continuous problem (see Assumption 2). In particular, we would like to apply the stability estimate (2.4) to the error $u - u_h$. However, this is not immediately possible since the discrete approximation is not exactly continuous across the interface. To overcome this issue, we will need to interpolate 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 3.5 (Discontinuous to continuous interpolation)**.** *There exists an interpolation operator Π_h^c from $V_{h,+}^k \times V_{h,-}^k$ into a subspace of $H^1(\Omega)$ such that for all $\hat{w}_h \in \hat{V}_h$ it holds that*

$$\sum_{\pm} h^{-1} \left\| \Pi_h^c w_h - w_{h,\pm} \right\|_{\Omega_\pm} + \left\| \nabla \left(\Pi_h^c w_h - w_{h,\pm} \right) \right\|_{\Omega_\pm} \leqslant C h^{-1/2} \sum_{\pm} \left\| w_{h,\pm} - w_{h,\Gamma} \right\|_\Gamma . \tag{3.17}$$

Proof. The claim follows from [BE07, Lemma 3.2 & 5.3 and Remark 3.2] upon making the following minor observation. In the reference, the bound on the right hand side is in terms of

$$\frac{1}{h^{1/2}} \sum_{F \in \mathcal{F}_h \setminus \mathcal{F}_h^{\partial\Omega}} \left\| [\![w_h]\!]_F \right\|_F ,$$

i.e. the sum runs over all interior facets of the triangulation. However, as $w_{h,\pm}$ are in $H^1(\Omega_\pm)$ only jumps over Γ remain. This concludes the argument upon noting that

$$\left\| [\![w_h]\!]_\Gamma \right\|_\Gamma \leqslant \sum_{\pm} \left\| w_{h,\pm} - w_{h,\Gamma} \right\|_\Gamma .$$

□

We will also need some more standard interpolation operators. Let $\Pi_\pm^{h,l}$ and $\Pi_\Gamma^{h,l}$ denote interpolation operators into the spaces $V_{h,\pm}^l$, respectively V_Γ^l, with the expected approximation properties:

$$\left\| v_\pm - \Pi_\pm^{h,l} v \right\|_{H^l(\Omega_\pm)} \leqslant C h^{s-l} \left\| v \right\|_{H^s(\Omega)}$$

for $1 \leqslant s \leqslant l+1,\ 0 \leqslant l \leqslant s$ and

$$\left\| v - \Pi_\Gamma^{h,l} v \right\|_F \leqslant h^{s+1/2} \left\| v \right\|_{H^{s+1}(T)}$$

for $0 \leqslant s \leqslant l$. Here $F \in \mathcal{F}_h^\Gamma$ is a facet on Γ and T is a corresponding volume element sharing this facet. We may take $\Pi_\pm^{h,l}$ as the Scott-Zhang operator and $\Pi_\Gamma^{h,l}$ as the local L^2-projection. Finally, we

can deduce an error estimate in the H^1-norm. Its proof is somewhat similar to the proof of [BP23, Theorem 21] in which unique continuation over an interface using an unfitted discretization has been treated.

Theorem 3.6. *Let $u \in H^1_0(\Omega) \cap [H^2(\Omega_+) \times H^2(\Omega_-)]$ solve (2.1)-(2.2) and denote $\hat{u} = (u, u_\Gamma)$ for $u_\Gamma = u|_\Gamma$. Let $(\hat{u}_h, \hat{z}_h) \in \hat{V}_h \times \hat{V}_h^*$ be the solution of the variational problem (2.18). Then under Assumption 2 the following quasi-optimal error bound holds true:*

$$\sum_{\pm} \|u - u_{h,\pm}\|_{H^1(\Omega_\pm)} \leqslant C^E \inf_{\hat{v} \in \hat{V}_h} |||(\hat{u} - \hat{v}_h, 0)|||_\square, \tag{3.18}$$

where the constant C^E is given by

$$C^E = C\left(1 + \frac{1}{C^B}\right)\left\{\sigma_\flat^{-1/2} + C^{\mathrm{stab}}\left(\sigma_\flat^{-1/2}\left[\sigma_\sharp + \sigma_\sharp^{1/2} \max_{\pm} \|\mu_\pm\|^{1/2}_{L^\infty(\Omega_\pm)}\right] + \sigma_\sharp^{1/2} + (\gamma_{\mathrm{GLS}})^{-1/2} + \max_{\pm}(\tilde{\mu}_\pm)^{1/2}\right)\right\}.$$

Proof. We would like to apply the continuous stability estimate to the error $u - u_h$. Since $u_h \notin H^1(\Omega)$, this is not directly possible. So we consider a projection $\Pi_h^{\mathrm{c}}(u_h)$ into continuous functions (see Lemma 3.5) instead. We have for any $y \in H^1_0(\Omega)$ that

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

We have to treat $\tilde{I}$ further. In line with the previously introduced notation we set $\hat{y} = (y, y_\Gamma)$, where $y = (y_+, y_-) := (y|_{\Omega_+}, y|_{\Omega_-})$ and $y_\Gamma := y|_\Gamma$. Using that $y_\pm = y_\Gamma$ on Γ we can write

$$\tilde{I} = \tilde{a}[\hat{u}_h, \hat{y}] - \sum_{\pm}(f_\pm, y_\pm)_{\Omega_\pm},$$

where we also used that u solves (2.1)-(2.2). Notice that the symmetrizing Nitsche term is omitted in $\tilde{a}[\cdot,\cdot]$ as it would not be well-defined for $y \in H^1_0(\Omega)$. On the other hand, using the discrete variational formulation (2.18) with $(\hat{w}_h, \hat{y}_h)$ defined by

$$\hat{w}_h = 0, \qquad y_{h,\pm} = \Pi_\pm^{h,k^*} y_\pm, \qquad y_{h,\Gamma} = \Pi_\pm^{h,k_\Gamma^*} y_\Gamma$$

yields:

$$\sum_{\pm}(f_\pm, y_{h,\pm})_{\Omega_\pm} - a[\hat{u}_h, \hat{y}_h] + s^*[\hat{z}_h, \hat{y}_h] = 0.$$

Adding this equation to $\tilde{I}$ and using that

$$\sum_{\pm}(f_\pm, y_\pm - y_{h,\pm})_{\Omega_\pm} = \tilde{a}[\hat{u}, \hat{y} - \hat{y}_h]$$

because u solves (2.1)-(2.2), we obtain

$$\tilde{I} = \tilde{a}[\hat{u}_h - \hat{u}, \hat{y} - \hat{y}_h] + s^*[\hat{z}_h, \hat{y}_h] + \sum_{\pm} (\sigma_\pm \nabla y_{h,\pm} \cdot \mathbf{n}_\pm, u_{h,\pm} - u_{h,\Gamma})_\Gamma.$$

From integration by parts, see (2.12), we obtain $\tilde{I} = \sum_{j=2}^{7} I_j$ where

$$I_2 = \sum_{\pm} (\mathcal{L}_\pm(u_{h,\pm} - u_\pm), y_\pm - y_{h,\pm})_{\mathcal{T}_h^\pm}, \quad I_3 = \sum_{\pm} \sum_{F \in \mathcal{F}_h^\pm} (\sigma_\pm [\![\nabla(u_{h,\pm} - u_\pm)]\!]_F \cdot \mathbf{n}_F, y_\pm - y_{h,\pm})_F,$$

$$I_4 = ([\![\sigma \nabla(u_{h,\pm} - u_\pm)]\!]_\Gamma \cdot \mathbf{n}_\Gamma, y_\Gamma - y_{h,\Gamma})_\Gamma, \; I_5 = \sum_{\pm} \frac{|\sigma_\pm| \, \lambda_\pm}{h} (u_{h,\pm} - u_{h,\Gamma} - u_\pm + u_\Gamma, y_\pm - y_\Gamma - y_{h,\pm} + y_{h,\Gamma})_\Gamma$$

$$I_6 = \sum_{\pm} (\sigma_\pm \nabla y_{h,\pm} \cdot \mathbf{n}_\pm, u_{h,\pm} - u_{h,\Gamma})_\Gamma, \quad I_7 = s^*[\hat{z}_h, \hat{y}_h].$$

We are going to control the I_j in terms of $\|y\|_{H^1(\Omega)}$ and $|||(\hat{u} - \hat{u}_h, \hat{z}_h)|||$. Using the assumption in (3.4), Lemma 3.5 and $\sigma_\flat \leqslant |\sigma_\pm|$ yields

$$\begin{aligned}
I_1 &\leqslant \|y\|_{H^1(\Omega)} \left(\sum_{\pm} \sigma_\sharp \|\nabla(\Pi_h^{\mathrm{c}} u_h - u_{h,\pm})\|_{\Omega_\pm} + \frac{\sigma_\sharp^{1/2} \max_\pm \|\mu_\pm\|_{L^\infty(\Omega_\pm)}^{1/2}}{h} \|\Pi_h^{\mathrm{c}} u_h - u_{h,\pm}\|_{\Omega_\pm} \right) \\
&\leqslant \|y\|_{H^1(\Omega)} \, \sigma_\flat^{-1/2} \left(\sigma_\sharp + \sigma_\sharp^{1/2} \max_\pm \|\mu_\pm\|_{L^\infty(\Omega_\pm)}^{1/2} \right) \sigma_\flat^{1/2} \left(\sum_{\pm} \|\nabla(\Pi_h^{\mathrm{c}} u_h - u_{h,\pm})\|_{\Omega_\pm} + \frac{1}{h} \|\Pi_h^{\mathrm{c}} u_h - u_{h,\pm}\|_{\Omega_\pm} \right) \\
&\leqslant \sigma_\flat^{-1/2} \left(\sigma_\sharp + \sigma_\sharp^{1/2} \max_\pm \|\mu_\pm\|_{L^\infty(\Omega_\pm)}^{1/2} \right) |||(\hat{u} - \hat{u}_h, 0)||| \, \|y\|_{H^1(\Omega)},
\end{aligned}$$

where the last inequality uses also that $u_\pm = u_\Gamma$ on Γ. Making use of the continuous trace inequality (2.11) and (low-order) interpolation estimates leads to

$$\begin{aligned}
\sum_{j=2}^{5} I_j &\leqslant C |||(\hat{u} - \hat{u}_h, 0)||| \sum_{\pm} \Big(h^{-1} \left((\gamma_{\mathrm{GLS}})^{-1/2} + |\sigma_\pm|^{1/2} \right) \|y_\pm - y_{h,\pm}\|_{\Omega_\pm} \\
&\quad + |\sigma_\pm|^{1/2} \|\nabla(y_\pm - y_{h,\pm})\|_{\Omega_\pm} + h^{-1/2} \sigma_\sharp^{1/2} \|y_\Gamma - y_{h,\Gamma}\|_\Gamma \Big) \\
&\leqslant C \left(\sigma_\sharp^{1/2} + (\gamma_{\mathrm{GLS}})^{-1/2} \right) |||(\hat{u} - \hat{u}_h, 0)||| \, \|y\|_{H^1(\Omega)}.
\end{aligned}$$

From the (discrete) trace inequality (2.10) we have $h^{1/2} \|\nabla y_{h,\pm}\|_\Gamma \leqslant C \|\nabla y_{h,\pm}\|_{\Omega_\pm}$ and hence

$$I_6 \leqslant C \|y\|_{H^1(\Omega)} \, \sigma_\sharp^{1/2} \sum_{\pm} |\sigma_\pm|^{1/2} \, h^{-1/2} \|u_{h,\pm} - u_{h,\Gamma} - (u_\pm - u_\Gamma)\|_\Gamma \leqslant C \sigma_\sharp^{1/2} |||(\hat{u} - \hat{u}_h, 0)||| \, \|y\|_{H^1(\Omega)}.$$

Finally,

$$I_7 \leqslant C |||(0, \hat{z}_h)||| \sum_{\pm} \{ |\sigma_\pm|^{1/2} \|\nabla y_h\|_{\Omega_\pm} + (\tilde{\mu}_\pm)^{1/2} \|y_h\|_{\Omega_\pm} \} \leqslant C \left(\sigma_\sharp^{1/2} + \max_\pm (\tilde{\mu}_\pm)^{1/2} \right) |||(0, \hat{z}_h)||| \, \|y\|_{H^1(\Omega)}.$$

Thus, we obtain by combining the stability estimate (2.4) and Proposition 3.4:

$$\|\Pi_h^{\mathrm{c}} u_h - u\|_{H^1(\Omega)} \leqslant C^{\mathrm{stab}} \|r_h\|_{H^{-1}(\Omega)} \leqslant C C^{\mathrm{stab}} C_0 \left(1 + \frac{1}{C^B} \right) \inf_{\hat{v} \in \hat{V}_h} |||(\hat{u} - \hat{v}_h, 0)|||_\square,$$

where

$$C_0 = \sigma_\flat^{-1/2}\left(\sigma_\sharp + \sigma_\sharp^{1/2}\max_\pm \|\mu_\pm\|^{1/2}_{L^\infty(\Omega_\pm)}\right) + \sigma_\sharp^{1/2} + (\gamma_{\mathrm{GLS}})^{-1/2} + \max_\pm(\tilde{\mu}_\pm)^{1/2}.$$

By applying the triangle inequality

$$\sum_\pm \|u - u_{h,\pm}\|_{H^1(\Omega_\pm)} \leqslant \sum_\pm \left\{\|\Pi_h^{\mathrm{c}} u_h - u_{h,\pm}\|_{H^1(\Omega_\pm)} + \|\Pi_h^{\mathrm{c}} u_h - u\|_{H^1(\Omega)}\right\}$$

and estimating the first term similarly to I_1 above, we then obtain the claim. □

Assuming smoothness of the solution we then obtain optimal decay of the error in the augmented triple norm from interpolation estimates.

Lemma 3.7. *Let $\hat{u} \in \hat{V}$ with $u_\pm \in H^{k+1}(\Omega_\pm)$. Then*

$$\inf_{\hat{v}\in\hat{V}_h} |||(\hat{u} - \hat{v}_h, 0)|||_\square \leqslant C^{\mathrm{app}} h^k \sum_\pm \|u_\pm\|_{H^{k+1}(\Omega_\pm)},$$

holds, where

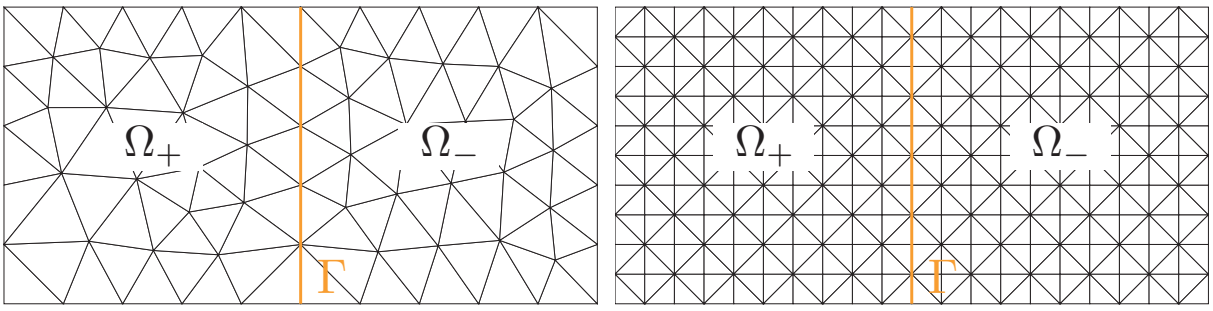
$$C^{\mathrm{app}} = C\left\{\sigma_\sharp^{1/2} + (\gamma_{\mathrm{GLS}})^{1/2}\left(\sigma_\sharp + \max_\pm \|\mu_\pm\|_{L^\infty(\Omega_\pm)}\right) + \max_\pm\left[\tilde{\mu}_\pm + \sigma_\flat^{-1}\left\|\mu_\pm^{\oplus}\right\|^2_{L^\infty(\Omega_\pm)}\right]^{1/2}\right\}.$$

Proof. Given in Appendix A. □

4. Numerical experiments

We present a series of numerical experiments to investigate the performance of the proposed method. We start in Section 4.1 with an academic toy problem and proceed in Section 4.2 to an acoustic cloaking device proposed in [ZLK+11]. Finally, in Section 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` [Sch97, Sch14]. Reproduction material is available at ... (Todo: upload once paper is finalized).



(A) Unstructured mesh (B) Symmetric mesh

Figure 1. Meshes for the cavity problem.

4.1. **Symmetric cavity.** The symmetric cavity problem seems to be one of the main benchmark tests in the numerical analysis literature on problems with sign-changing coefficients, see e.g. [CCJ13, AHL17, CLR23]. 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, see [CCJ13, Section 3.3], that Assumption 2 holds true for $\sigma_+ + \sigma_- \neq -1$. In this case the solution is given by

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

The problem becomes numerically more difficult to handle if the critical contrast $\sigma_+/\sigma_- = -1$ is

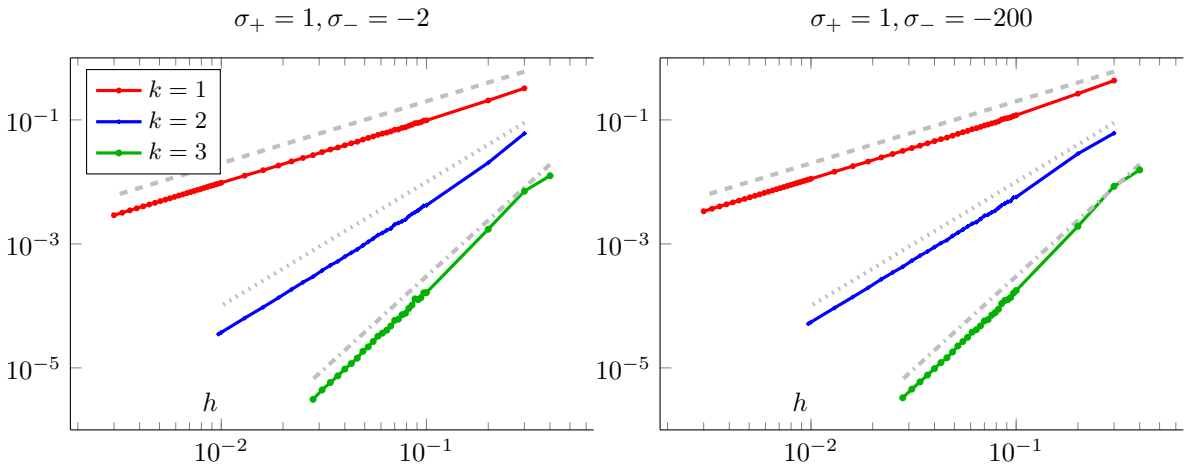


FIGURE 2. The relative error $\|u - u_h\|_{H^1(\Omega_+\cup\Omega_-)} / \|u\|_{H^1(\Omega_+\cup\Omega_-)}$ using $u_h|_{\Omega_\pm} = u_{h,\pm}$ for the symmetric cavity in the well-posed case on unstructured meshes.

approached. Let us therefore 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 3.6 are achieved.

Now let us approach the critical contrast setting $\sigma_+ = 1$ and $\sigma_- = -1.001$. It is well-known, see e.g. [CCJ13, AHL17], that a naive Galerkin discretization obtained from the bilinear form

$$(u,v) \mapsto \int_\Omega \sigma \nabla u \nabla v, \quad \sigma|_{\Omega_\pm} = \sigma_\pm,$$

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 clearly 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

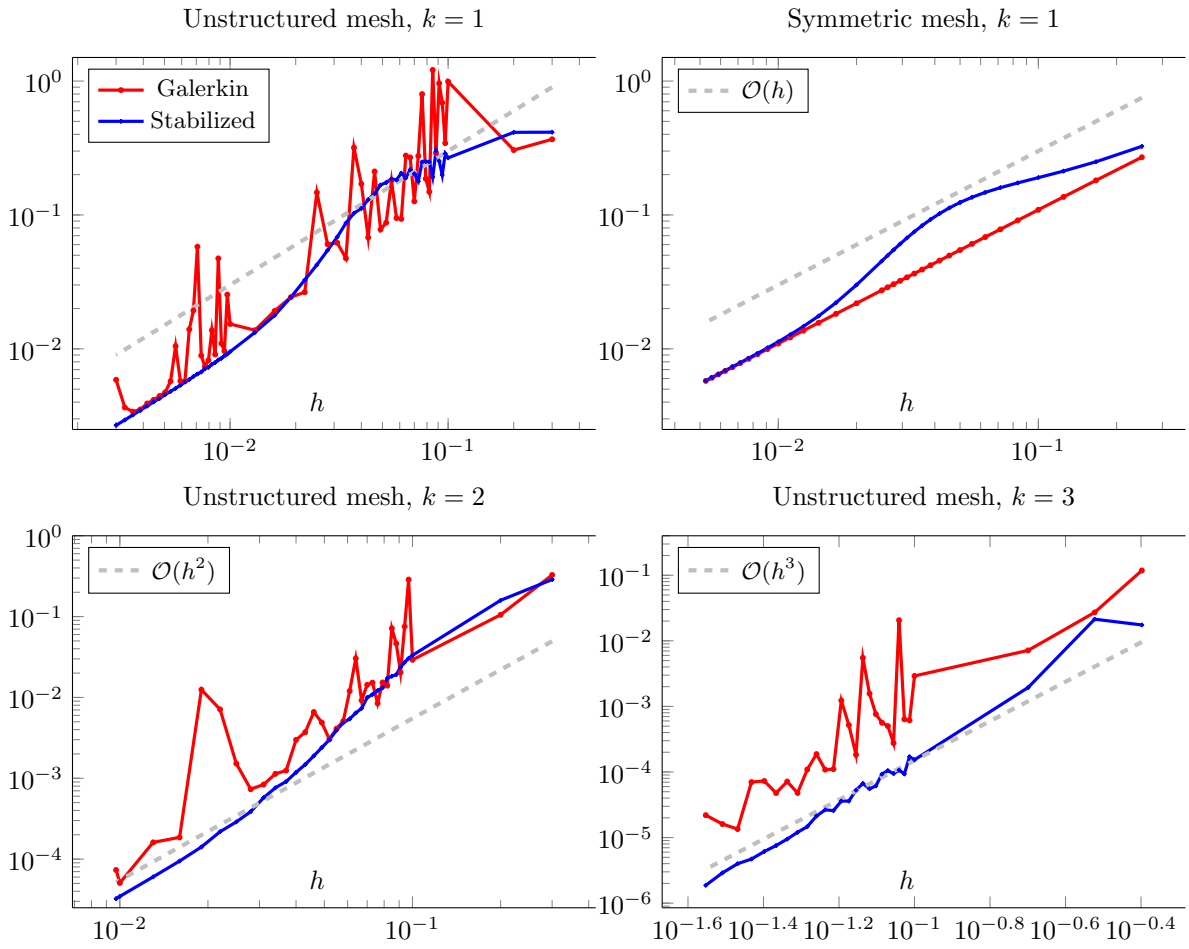


FIGURE 3. The relative error $\|u - u_h\|_{H^1(\Omega_+ \cup \Omega_-)} / \|u\|_{H^1(\Omega_+ \cup \Omega_-)}$ for the symmetric cavity at contrast $\sigma_+/\sigma_- = -1.001$.

was observed that this allows to reduce the size of the stabilization parameters without affecting the numerical stability which leads to a reduced error on coarse meshes. Still, on symmetric meshes the error for the stabilized method for coarse mesh sizes is slightly higher than for the Galerkin method. As our method contains an in-built 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 measure here.

4.2. **Metamaterial.** Let us proceed to a more realistic test case. To this end, we consider the acoustic cloaking device from [ZLK+11]. 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 radially varying μ given by

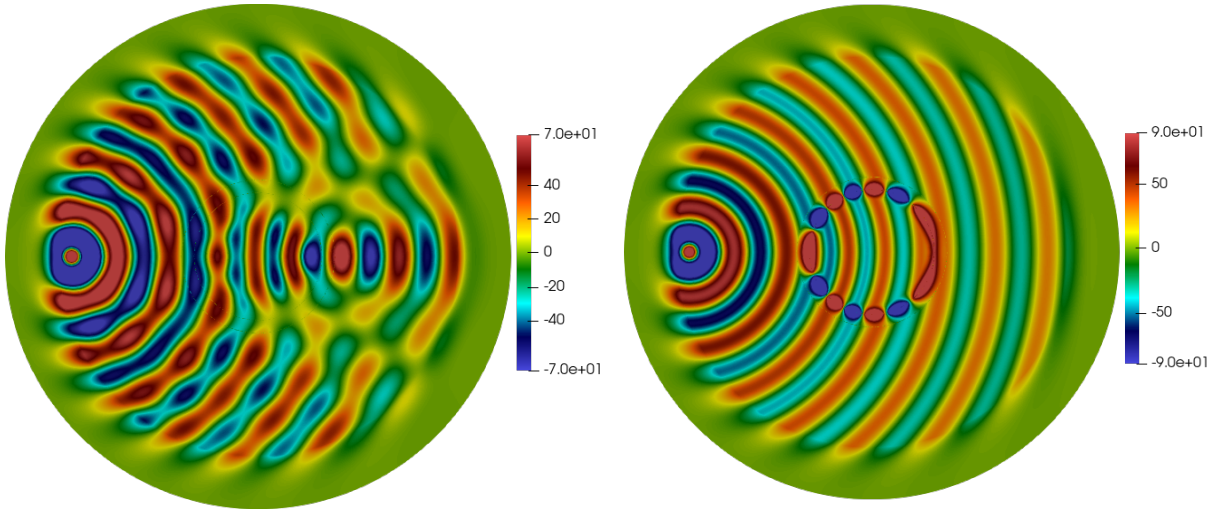
$$\sigma(r) = \begin{cases} 1/\rho_0 & \text{for } 0 < r < a, \\ 1/\rho_1 & \text{for } a < r < b, \\ -1/\rho_1 & \text{for } b < r < c, \\ 1/\rho_0 & \text{for } c < r, \end{cases} \qquad \mu(r) = \begin{cases} (\omega^2/\kappa_0)(b/a)^4 & \text{for } 0 < r < a, \\ -(\omega^2/\kappa_1)(b/r)^4 & \text{for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

where $r = \sqrt{x^2 + y^2}$ and according to [ZLK+11] the parameters are given as follows:

$$\kappa_0 = 2.19 \text{ GPa}, \ \kappa_1 = 0.48\kappa_0, \ \rho_0 = 998 \text{ kg/m}^3, \ \rho_1 = \rho_0$$

$$a = 1.0 \text{ m } b = 1.2 \text{ m}, \ c = 1.44 \text{ m}.$$

The source is positioned at $x_0 = (-3.5 \text{ m}, 0)$ and we truncate the computational domain by a perfectly matched layer (PML), see e.g. [CM98], extending from $r = 3.75$ m to $r = 4.75$ m. Recall that our method contains an in-built Galerkin discretization. So if we take the full dual order and omit the stabilization terms in the PML layer, the PML will work as usual to attenuate the waves. A sketch of the computational setup is displayed in Fig. 5.



(A) without cloak (B) with cloak

FIGURE 4. Cloaking using a metamaterial. The function values in the cloaking domain are actually very high, but have been truncated here to ± 90 to aid presentation.

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 $r < a$. Note that as $\mu(r)$ differs for $r < a$ and $r > c$ we would expect to see traces of the object in the propagating waves if no metamaterial was 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.

If the cloak works perfectly, we would expect that the solution in the exterior of the cloaked region $r > c$ is given by a spherical wave emanating from the point source at x_0 defined as

$$u = \frac{i\rho_0}{4} H_0^{(1)} \left(\omega \sqrt{\frac{\rho_0}{\kappa_0}} \, \|x - x_0\| \right), \tag{4.1}$$

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

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

as sketched in Fig. 5. Note that Ω_i represents a buffer layer between Ω_e and the interface.

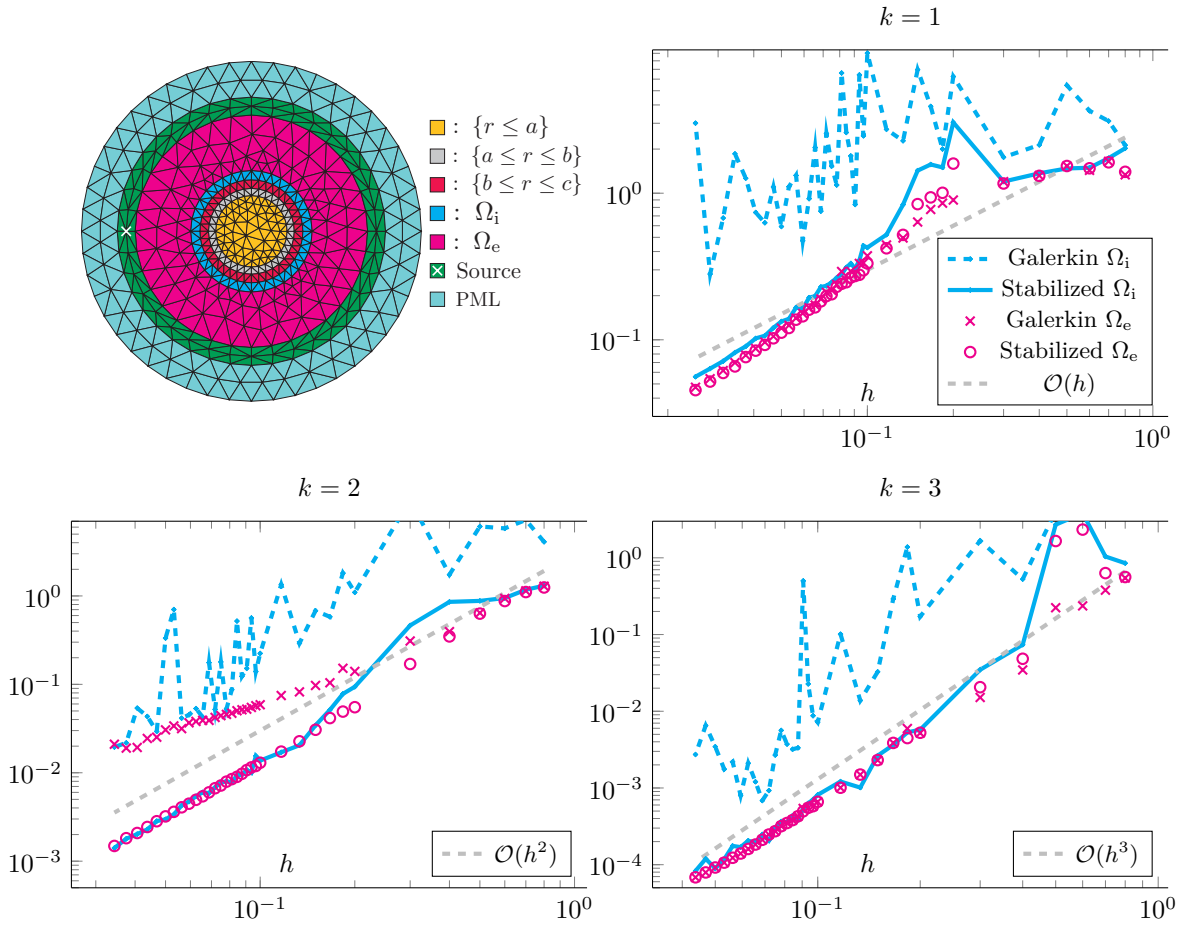


FIGURE 5. Relative H^1-error in Ω_i and Ω_e with respect to the reference solution (4.1). Here, h refers to an upper bound on the mesh size in $\{r < a\} \cup \{r > c\}$. The mesh size within the metamaterial $\{a \leqslant r \leqslant c\}$ is bounded from above by $h/3$.

The relative H^1-erors for the stabilized and the plain Galerkin method in these regions are displayed Fig. 5. We observe that the Galerkin method shows severe instabilities in the region Ω_i. Interestingly, for $k = 2$ these errors even pollute the approximation in the exterior region Ω_e so that the expected convergence rate is not reached there. For $k = 1, 3$ and $k = 4$ (not shown here), however, the Galerkin solution converges at the optimal rate in Ω_e. In contrast, the stabilized method converges at the optimal rate in both regions. These results clearly demonstrate the suitability and relevance of the proposed method for the reliable simulation of physically interesting metamaterials.

4.3. **Non-symmetric cavity at super-critical contrast.** Let us finally consider a test case for which it is known, see [CCJ13, Section 3.3] and [AL23, Section 4.2], that the well-posedness Assumption 2 is violated. Like in Section 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 have $\mu_\pm = 0$ and consider the super-critical contrast $\sigma_+ = 1, \sigma_- = -1$. In this case, the problem is not Fredholm, yet the weaker Assumption 1 of injectivity holds true. We consider then the exact solution from [AL23, Section 7.2] defined as

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

Note that the theoretical convergence result derived in Theorem 3.6 cannot be applied here as it relies on well-posedness of the problem. We would like to investigate whether convergence in H^1 can still be obtained with our method. The relative H^1-errors using polynomial degree $k = 2$ are displayed in Fig. 6. To enhance the accuracy of the method, we add the following second-order jump terms:

$$\sum_{F \in \mathcal{F}_h^\pm} h^3 \left([\![D^2 u_\pm]\!], [\![D^2 v_\pm]\!] \right)_F ,$$

where $D^2 u_\pm$ denotes the Hessian, to the primal stabilization $s_\pm(u_\pm, v_\pm)$. This is reasonable as the solutions in the subdomains are smooth. Despite these modifications, 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. 6 shows that on finer meshes the convergence deteriorates to a logarithmic rate. The plot of the absolute error in Fig. 6 shows that this seems to stem from a poor approximation close to the interface. To investigate this more qualitatively, we measured additionally the quantity (see blue line in Fig. 6)

$$\left(\frac{\|u_\Gamma - u_{h,\Gamma}\|_\Gamma \, \|\nabla_\Gamma (u_\Gamma - u_{h,\Gamma})\|_\Gamma}{\|u_\Gamma\|_\Gamma \, \|\nabla_\Gamma u_\Gamma\|_\Gamma} \right)^{1/2}$$

which, according to the Gagliardo-Nirenberg inequality, gives an upper bound for $\|u_\Gamma - u_{h,\Gamma}\|_{H^{1/2}(\Gamma)}$. We observe the same logarithmic convergence behavior in the asymptotic regime as for the error in the bulk and a significantly higher error constant. This supports the conclusion that the difficulty of capturing the behavior of the solution at the interface is responsible for the observed degeneration of the convergence rate. Overall, this example shows that our method can still be applied when the

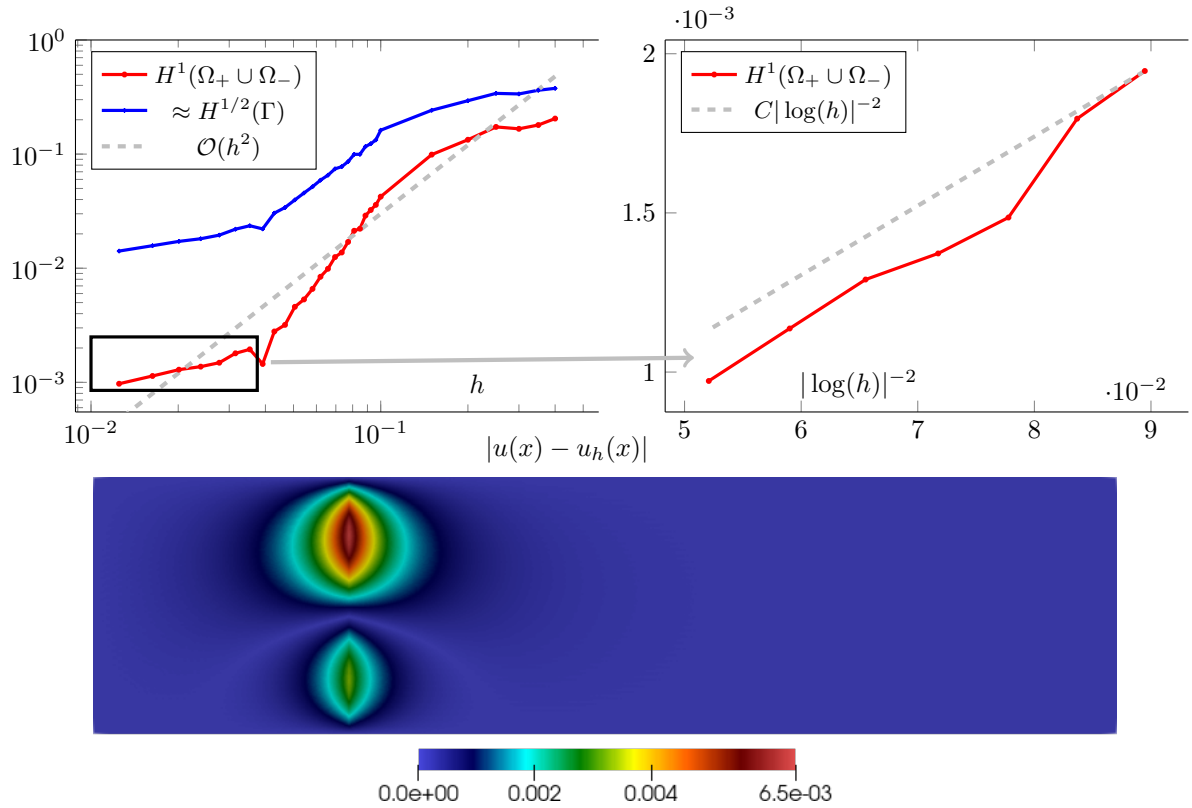


FIGURE 6. Results for the non-symmetric cavity at super-critical contrast obtained with the stabilized method using $k = 2$. The red line displays the relative error in the bulk while the blue line shows the relative errors for the quantity $\|u_\Gamma - u_{h,\Gamma}\|_\Gamma^{1/2} \|\nabla_\Gamma(u_\Gamma - u_{h,\Gamma})\|_\Gamma^{1/2}$. Note that the x-axis in the right plot is given in terms of $|\log(h)|^{-2}$ and that the error looks almost like a diagonal line in this plot. The figure on the bottom displays the absolute error on a mesh of size $h \approx 0.028$.

well-posedness assumption is violated, yet, the convergence rates proven in Theorem 3.6 are no longer valid. On the other hand, due to the lack of a continuum stability estimate, the optimal convergence rate one can expect from a numerical method applied to this problem is not known. This would be an interesting question to investigate in future research.

5. Conclusion

We presented a stabilized finite element method for the numerical approximation of acoustic metamaterials and proved its optimal convergence under a well-posedness assumption. The method can be applied on general shape-regular meshes and has shown reliable and accurate performance 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 as they often occur in applications.
- In the analysis and implementation of the method we have assumed that the mesh fits the interface. However, we expect that the method could be easily extended to unfitted discretizations by combining the techniques from references [BEH$^+$19] and [BP23].
- To prove convergence rates in H^1, we have assumed well-posedness of the continuous problem. In the numerical experiment of Section 4.3 where this assumption is violated we have seen that the method appears to convergence asymptotically at a logarithmic rate. It would be interesting to investigate theoretically whether this rate is optimal.
- In this article we have restricted 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, would certainly be of practical relevance.

References

[AHL17] Assyr Abdulle, Martin E. Huber, and Simon Lemaire. An optimization-based numerical method for diffusion problems with sign-changing coefficients. *Comptes Rendus Mathematique*, 355(4):472–478, 2017.

[AL23] Assyr Abdulle and Simon Lemaire. An optimization-based method for sign-changing elliptic PDEs. working paper or preprint, June 2023.

[BBDCCJ12] Anne-Sophie Bonnet-Ben Dhia, Lucas Chesnel, and Patrick Ciarlet Jr. T-coercivity for scalar interface problems between dielectrics and metamaterials. *ESAIM: M2AN*, 46(6):1363–1387, 2012.

[BBDCCJ14a] Anne-Sophie Bonnet-Ben Dhia, Lucas Chesnel, and Patrick Ciarlet Jr. T-coercivity for the Maxwell problem with sign-changing coefficients. *Communications in Partial Differential Equations*, 39(6):1007–1031, 2014.

[BBDCCJ14b] Anne-Sophie Bonnet-Ben Dhia, Lucas Chesnel, and Patrick Ciarlet Jr. Two-dimensional Maxwell's equations with sign-changing coefficients. *Applied Numerical Mathematics*, 79:29–41, 2014. Workshop on Numerical Electromagnetics and Industrial Applications (NELIA 2011).

[BBDCCJ18] Anne-Sophie Bonnet-Ben Dhia, Camille Carvalho, and Patrick Ciarlet Jr. Mesh requirements for the finite element approximation of problems with sign-changing coefficients. *Numerische Mathematik*, 138(4):801–838, 2018.

[BBDCJZ10] Anne-Sophie Bonnet-Ben Dhia, Patrick Ciarlet Jr, and Carlo Maria Zwölf. Time harmonic wave diffraction problems in materials with sign-shifting coefficients. *Journal of Computational and Applied Mathematics*, 234(6):1912–1919, 2010. Eighth International Conference on Mathematical and Numerical Aspects of Waves (Waves 2007).

[BE07] Erik Burman and Alexandre Ern. Continuous interior penalty *hp*-finite element methods for advection and advection-diffusion equations. *Mathematics of computation*, 76(259):1119–1140, 2007.

[BEH$^+$19] Erik Burman, Daniel Elfverson, Peter Hansbo, Mats G. Larson, and Karl Larsson. Hybridized CutFEM for Elliptic Interface Problems. *SIAM Journal on Scientific Computing*, 41(5):A3354–A3380, 2019.

[BHL18] Erik Burman, Peter Hansbo, and Mats G. Larson. Solving ill-posed control problems by stabilized finite element methods: an alternative to Tikhonov regularization. *Inverse Problems*, 34(3):035004, jan 2018.

[BLO18] Erik Burman, Mats G. Larson, and Lauri Oksanen. Primal-dual mixed finite element methods for the elliptic Cauchy problem. *SIAM Journal on Numerical Analysis*, 56(6):3480–3509, 2018.

[BNO19] Erik Burman, Mihai Nechita, and Lauri Oksanen. Unique continuation for the Helmholtz equation using stabilized finite element methods. *Journal de Mathématiques Pures et Appliquées*, 129:1–22, 2019.

[BNO20] Erik Burman, Mihai Nechita, and Lauri Oksanen. A stabilized finite element method for inverse problems subject to the convection–diffusion equation. I: diffusion-dominated regime. *Numerische Mathematik*, 144(3):451–477, 2020.

[BNO23] Erik Burman, Mihai Nechita, and Lauri Oksanen. Optimal finite element approximation of unique continuation, 2023.

[BP23] Erik Burman and Janosch Preuss. Unique continuation for an elliptic interface problem using unfitted isoparametric finite elements, 2023.

[Bur13] Erik Burman. Stabilized Finite Element Methods for Nonsymmetric, Noncoercive, and Ill-Posed Problems. Part I: Elliptic Equations. *SIAM Journal on Scientific Computing*, 35(6):A2752–A2780, 2013.

[Bur14] Erik Burman. Error estimates for stabilized finite element methods applied to ill-posed problems. *Comptes Rendus Mathematique*, 352(7):655–659, 2014.

[CCA16] Steven A Cummer, Johan Christensen, and Andrea Alù. Controlling sound with acoustic metamaterials. *Nature Reviews Materials*, 1(3):1–13, 2016.

[CCC17] Camille Carvalho, Lucas Chesnel, and Patrick Ciarlet. Eigenvalue problems with sign-changing coefficients. *Comptes Rendus. Mathématique*, 355(6):671–675, 2017.

[CCJ13] Lucas Chesnel and Patrick Ciarlet Jr. T-coercivity and continuous Galerkin methods: application to transmission problems with sign changing coefficients. *Numerische Mathematik*, 124(1):1–29, 2013.

[CFV21] Théophile Chaumont-Frelet and Barbara Verfürth. A generalized finite element method for problems with sign-changing coefficients. *ESAIM: M2AN*, 55(3):939–967, 2021.

[CLR23] Patrick Ciarlet, David Lassounon, and Mahran Rihani. An optimal control-based numerical method for scalar transmission problems with sign-changing coefficients. *SIAM Journal on Numerical Analysis*, 61(3):1316–1339, 2023.

[CM98] Francis Collino and Peter Monk. The Perfectly Matched Layer in Curvilinear Coordinates. *SIAM J. Sci. Comput.*, 19(6):2061–2090, November 1998.

[EG21] Alexandre Ern and Jean-Luc Guermond. *Finite Elements I: Approximation and Interpolation*, volume 72 of *Texts in Applied Mathematics*. Springer Nature, Cham, Switzerland, 2021.

[Egg09] Herbert Egger. A class of hybrid mortar finite element methods for interface problems with non-matching meshes. *preprint AICES-2009-2, Jan*, 2009.

[GR86] Vivette Girault and Pierre-Arnaud Raviart. *Finite element methods for Navier-Stokes equations: Theory and algorithms*, volume 5. 1986.

[Hal21] Martin Halla. On the approximation of dispersive electromagnetic eigenvalue problems in two dimensions. *IMA Journal of Numerical Analysis*, 43(1):535–559, 12 2021.

[HHO24] Martin Halla, Thorsten Hohage, and Florian Oberender. A new numerical method for scalar eigenvalue problems in heterogeneous, dispersive, sign-changing materials, 2024.

[Sch97] Joachim Schöberl. NETGEN An advancing front 2D/3D-mesh generator based on abstract rules. *Comput. Vis. Sci.*, 1(1):41–52, 1997.

[Sch14] Joachim Schöberl. C++11 implementation of finite elements in NGSolve. Technical report, ASC-2014-30, Institute for Analysis and Scientific Computing, September 2014.

[ZLK+11] Xuefeng Zhu, Bin Liang, Weiwei Kan, Xinye Zou, and Jianchun Cheng. Acoustic Cloaking by a Superlens with Single-Negative Materials. *Physical review letters*, 106(1):014301, 2011.

Appendix A. Selected proofs

The arguments given here are standard and can probably be omitted in the submitted version.

Additional calculations for Proof of Proposition 3.2

Proof. We give some details on the lower bound for

$$B[(\hat{v},\hat{z});(0,\hat{\zeta})] = a[\hat{v},\hat{\zeta}] = (\llbracket\sigma\nabla v\rrbracket_\Gamma\cdot\mathbf{n}_\Gamma,\zeta_\Gamma)_\Gamma - \sum_{\pm}\frac{\lambda_\pm|\sigma_\pm|}{h}(v_\pm - v_\Gamma,\zeta_\Gamma)_\Gamma\Big\}.$$

By Young's inequality

$$\begin{aligned}\sum_{\pm}\frac{\lambda_\pm|\sigma_\pm|}{h}(v_\pm - v_\Gamma,\zeta_\Gamma)_\Gamma &\leqslant \sum_{\pm}\frac{\lambda_\pm|\sigma_\pm|^{1/2}}{h^{1/2}}\left\|v_\pm - v_\Gamma\right\|_\Gamma\left\|\zeta_\Gamma\right\|_\Gamma\frac{\sigma_\sharp^{1/2}}{h^{1/2}}\\ &\leqslant\left(\sum_{\pm}\frac{\lambda_\pm^2}{2\varepsilon}\frac{|\sigma_\pm|}{h}\left\|v_\pm - v_\Gamma\right\|_\Gamma^2\right) + \varepsilon\frac{\sigma_\sharp}{h}\left\|\zeta_\Gamma\right\|_\Gamma^2.\end{aligned}$$

Choosing $\varepsilon = 3/4$ and recalling that $\zeta_\Gamma := \sigma_\sharp^{-1}h\llbracket\sigma\nabla v\rrbracket_\Gamma\cdot\mathbf{n}_\Gamma$ we obtain

$$B[(\hat{v},\hat{z});(0,\hat{\zeta})] \geq \frac{1}{4}\sigma_\sharp^{-1}h\|\llbracket\sigma\nabla v\rrbracket_\Gamma\cdot\mathbf{n}_\Gamma\|_\Gamma^2 - \sum_{\pm}\frac{2\lambda_\pm^2}{3}\frac{|\sigma_\pm|}{h}\|v_\pm - v_\Gamma\|_\Gamma^2.$$

We show the bound on the squared norm by appealing to (3.5) and (2.10):

$$\begin{aligned}&|||(\hat{w},\hat{y})|||^2 = |||(\alpha\hat{v}+\hat{z},-\alpha\hat{z}+\hat{\zeta})|||^2\\ &= |\alpha\hat{v}+\hat{z}|_s^2 + \sigma_\sharp^{-1}h\left\|\llbracket\alpha\sigma\nabla v + \sigma\nabla z\rrbracket_\Gamma\cdot\mathbf{n}_\Gamma\right\|_\Gamma^2 + \sum_{\pm}\Big\{\left|\sigma_\pm\right|\left\|\alpha\nabla z_\pm\right\|_{\Omega_\pm}^2 + \tilde{\mu}_\pm\|\alpha z_\pm\|_{\Omega_\pm}^2 + \frac{|\sigma_\pm|}{h}\left\|-\alpha z_\pm + \alpha z_\Gamma - \hat{\zeta}\right\|_\Gamma^2\Big\}\\ &\leqslant 2\alpha^2\left|\hat{v}\right|_s^2 + 2(\alpha^2+1)\sigma_\sharp^{-1}h\left\|\llbracket\sigma\nabla v\rrbracket_\Gamma\cdot\mathbf{n}_\Gamma\right\|_\Gamma^2 + 2\left|\hat{z}\right|_s^2 + 2\sigma_\sharp^{-1}h\left\|\llbracket\sigma\nabla z\rrbracket_\Gamma\cdot\mathbf{n}_\Gamma\right\|_\Gamma^2\\ &\quad + 2\alpha^2\sum_{\pm}\Big\{\left|\sigma_\pm\right|\left\|\nabla z_\pm\right\|_{\Omega_\pm}^2 + \tilde{\mu}_\pm\|z_\pm\|_{\Omega_\pm}^2 + \frac{|\sigma_\pm|}{h}\left\|z_\pm - z_\Gamma\right\|_\Gamma^2\Big\}\\ &\leqslant 2\alpha^2\left|\hat{v}\right|_s^2 + 2(\alpha^2+1)\sigma_\sharp^{-1}h\left\|\llbracket\sigma\nabla v\rrbracket_\Gamma\cdot\mathbf{n}_\Gamma\right\|_\Gamma^2\\ &\quad + 2\left(\alpha^2 + C^s + 2C_{\mathrm{tr}}\right)\sum_{\pm}\Big\{\left|\sigma_\pm\right|\left\|\nabla z_\pm\right\|_{\Omega_\pm}^2 + \tilde{\mu}_\pm\|z_\pm\|_{\Omega_\pm}^2 + \frac{|\sigma_\pm|}{h}\left\|z_\pm - z_\Gamma\right\|_\Gamma^2\Big\}\\ &\leqslant 2(\alpha^2 + C^s + 2C_{\mathrm{tr}} + 1)|||(\hat{v},\hat{z})|||^2.\end{aligned}$$

□

Proof of Lemma 3.7

Proof. We take $v_{h,\pm} = \Pi_\pm^{h,k}u_\pm$ and $v_{h,\Gamma} = \Pi_\Gamma^{h,k}u_\Gamma$. Using the continuous trace inequality (2.11) and the interpolation we can estimate the augmentation term in $|||(\cdot,\cdot)|||_\square$ as follows:

$$\sum_{\pm}\Big\{|\sigma_\pm|^{1/2}\left\|\nabla(u_\pm - \Pi_\pm^{h,k}u_\pm)\right\|_{\Omega_\pm} + h^{1/2}|\sigma_\pm|^{1/2}\left\|\nabla(u_\pm - \Pi_\pm^{h,k}u_\pm)\cdot\mathbf{n}_\Gamma\right\|_\Gamma$$

$$
\begin{aligned}
&+ \left[\tilde{\mu}_\pm + \sigma_\flat^{-1} \left\|\mu_\pm^\oplus\right\|^2_{L^\infty(\Omega_\pm)}\right]^{1/2} \left\|u_\pm - \Pi_\pm^{h,k} u_\pm\right\|_{\Omega_\pm} \Big\} \\
&\leqslant C \left(\sigma_\sharp^{1/2} + \max_\pm \left[\tilde{\mu}_\pm + \sigma_\flat^{-1} \left\|\mu_\pm^\oplus\right\|^2_{L^\infty(\Omega_\pm)}\right]^{1/2}\right) h^k \sum_\pm \|u_\pm\|_{H^{k+1}(\Omega_\pm)} .
\end{aligned}
$$

We treat the remaining terms one after another. For the Galerkin-least-squares term we have

$$
\begin{aligned}
&\sum_\pm \sum_{T \in \mathcal{T}_h^\pm} (\gamma_{\mathrm{GLS}})^{1/2} h \left\|\mathcal{L}_\pm (u_\pm - \Pi_\pm^{h,k} u_\pm)\right\|_T \\
&\leqslant C(\gamma_{\mathrm{GLS}})^{1/2} \left(\sigma_\sharp + \max_\pm \|\mu_\pm\|_{L^\infty(\Omega_\pm)}\right) \sum_\pm \sum_{T \in \mathcal{T}_h^\pm} h \left\|u_\pm - \Pi_\pm^{h,k} u_\pm\right\|_{H^2(T)} \\
&\leqslant C(\gamma_{\mathrm{GLS}})^{1/2} \left(\sigma_\sharp + \max_\pm \|\mu_\pm\|_{L^\infty(\Omega_\pm)}\right) h^k \sum_\pm \|u_\pm\|_{H^{k+1}(\Omega_\pm)} .
\end{aligned}
$$

We estimate the interior penalty term by another application of the continuous trace inequality (2.11):

$$
\begin{aligned}
&\sum_\pm \sum_{F \in \mathcal{F}_h^\pm} h^{1/2} |\sigma_\pm|^{1/2} \left\|[\![\nabla (u_\pm - \Pi_\pm^{h,k} u_\pm)]\!]_F \cdot \mathbf{n}_F\right\|_F \\
&\leqslant C \sigma_\sharp^{1/2} \sum_\pm \sum_{T \in \mathcal{T}_h^\pm} \left(\left\|\nabla \left(u_\pm - \Pi_\pm^h u_\pm\right)\right\|_T + h \left\|u_\pm - \Pi_\pm^h u_\pm\right\|_{H^2(T)}\right) \leqslant C \sigma_\sharp^{1/2} h^k \sum_\pm \|u_\pm\|_{H^{k+1}(\Omega_\pm)} .
\end{aligned}
$$

Similarly, we obtain

$$
\begin{aligned}
&\sum_\pm \frac{|\sigma_\pm|^{1/2}}{h^{1/2}} \left\|u_\pm - u_\Gamma - \Pi_\pm^{h,k} u_\pm + \Pi_\Gamma^{h,k} u_\Gamma\right\|_\Gamma \\
&\leqslant C \sigma_\sharp^{1/2} \sum_\pm \left(h^{-1} \left\|u_\pm - \Pi_\pm^{h,k} u_\pm\right\|_{\Omega_\pm} + \left\|\nabla \left(u_\pm - \Pi_\pm^{h,k} u_\pm\right)\right\|_{\Omega_\pm} + \left\|u_\Gamma - \Pi_\Gamma^{h,k} u_\Gamma\right\|_\Gamma\right) \\
&\leqslant C \sigma_\sharp^{1/2} h^k \sum_\pm \|u_\pm\|_{H^{k+1}(\Omega_\pm)} .
\end{aligned}
$$

A final application of inequality (2.11) gives

$$
\begin{aligned}
&\sigma_\sharp^{-1/2} h^{1/2} \left\|[\![\sigma \nabla \left(u - \Pi^{h,k} u\right)]\!] \cdot \mathbf{n}_\Gamma\right\|_\Gamma \\
&\leqslant C \sigma_\sharp^{1/2} \sum_\pm \sum_{T \in \mathcal{T}_h^\pm} \left(\left\|\nabla \left(u_\pm - \Pi_\pm^{h,k} u_\pm\right)\right\|_T + h \left\|u_\pm - \Pi_\pm^{h,k} u_\pm\right\|_{H^2(T)}\right) \leqslant C \sigma_\sharp^{1/2} h^k \sum_\pm \|u_\pm\|_{H^{k+1}(\Omega_\pm)} .
\end{aligned}
$$

□