A generalized finite element method for problems with sign-changing coefficients

Théophile Chaumont-Frelet, Barbara Verfürth

▶ **To cite this version:**

Théophile Chaumont-Frelet, Barbara Verfürth. A generalized finite element method for problems with sign-changing coefficients. 2020. hal-02496832v1

HAL Id: hal-02496832

https://inria.hal.science/hal-02496832v1

Preprint submitted on 3 Mar 2020 (v1), last revised 27 Aug 2020 (v2)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

A generalized finite element method for problems with sign-changing coefficients

Théophile Chaumont-Frelet*† Barbara Verfürth‡

Abstract. Problems with sign-changing coefficients occur, for instance, in the study of transmission problems with metamaterials. In this work, we present and analyze a generalized finite element method in the spirit of the Localized Orthogonal Decomposition, that is especially efficient when the negative and positive materials exhibit multiscale features. We derive optimal linear convergence in the energy norm independently of the potentially low regularity of the exact solution. Numerical experiments illustrate the theoretical convergence rates and show the applicability of the method for a large class of sign-changing diffusion problems.

Key words. generalized finite element method, multiscale method, sign-changing coefficients, T-coercivity

AMS subject classifications. 65N30, 65N12, 65N15, 78A48, 35J20

1. Introduction

Metamaterials with, for instance, negative refractive index have a attracted a lot of interest over the last years due to many applications [25, 31]. The related mathematical problems are characterized by so-called sign-changing coefficients. At the simplest example of a diffusion problem, it means that the diffusion coefficient σ takes strictly negative values, i.e., $\sigma \leq -|\sigma_-| < 0$ in some part Ω_- of the domain, while it takes strictly positive values, i.e., $\sigma \geq |\sigma_+| > 0$ on the complement Ω_+. Such a behavior of the coefficient in the PDE does not only appear for metamaterials with negative effective properties [31], but also for electric permittivities, which can have a negative real part for certain metals.

The sign-change of the PDE coefficient has tremendous effects on the analysis and numerics. The standard assumption of coercive bilinear forms is no longer valid, so that existence and uniqueness of solutions have to be studied anew. Employing the approach of T-coercivity [6], a large progress has been made in this area in the last years considering the diffusion problem [3, 6] as well as time-harmonic wave propagation [4, 5] and eigenvalue problems [8]. In particular, two-dimensional settings with a polygonal interface between the two regions (positive vs. negative coefficient) are very well understood now [2]. Essentially, the problem is well-posed if the contrast $|\sigma_+|/|\sigma_-|$ lies outside a so-called "critical interval" $I = [1/r, r]$, where $r \geq 1$ depends on the geometry of Γ.

When discretizing these problems with the standard finite element method, the questions of existence and uniqueness of the discrete solution as well as convergence rates for the error immediately arise. Simply speaking, they have been answered positively in two different scenarios, namely a) if the mesh satisfies certain symmetry properties around the interface Γ between Ω_-

*Inria Sophia Antipolis Méditerranée, 2004 Route des Lucioles, 06902 Valbonne, France
†Laboratoire J.A. Dieudonné UMR CNRS 7351, Parc Valrose, 06108 Nice, France
‡Institut für Mathematik, Universität Augsburg, Universitätsstr. 14, D-86159 Augsburg

and Ω_+, which is denoted as T-conformity [2], or b) if the contrast $|\sigma_+|/|\sigma_-|$ is outside an enlarged critical interval $\widetilde{I} = [1/\widetilde{r}, \widetilde{r}]$, where $\widetilde{r} > r$ [9, Section 5.1].

Besides the a priori stability and error analysis, a posteriori error indicators and their reliability and efficiency have been studied for the standard finite element method as well [12, 24]. Furthermore, an optimization-based scheme which does not require symmetric meshes is introduced in [1]. Apart from continuous Galerkin methods, we also mention that schemes in the discontinuous Galerkin framework have been presented and analyzed in [10] and [21].

The main contribution of the present work is the introduction and numerical analysis of a generalized finite element method in the spirit and framework of the Localized Orthogonal Decomposition (LOD) [18, 23, 26]. The latter has been successfully applied in various situations, where we mention in particular the reduction of the pollution effect for high-frequency Helmholtz problems [14, 27, 30]. The efficient implementation of the method is outlined in [13]. The LOD can also be interpreted in the context of homogenization [15] and domain decomposition methods [19, 20, 29], approaches that we however do not follow here.

We analyze the stability and convergence of the proposed method under the assumption that the interface is resolved by the mesh and that the contrast is "sufficiently large". While this restriction means that the interface Γ is essentially "macroscale", σ is allowed to exhibit a rough and multiscale behavior in Ω_- and Ω_+. Under these assumptions, the present method allows for optimal convergence orders on uniform meshes, even in the presence of corner singularities, which is already known for positive discontinuous diffusion coefficients. In contrast with standard FEM [9], considerable complications arise in the analysis of the LOD method in the presence of sign-changing coefficients. Indeed, while the LOD method has been analyzed for a rather large class of inf-sup stable problems ([22, Chap. 2]), these general arguments cannot be directly applied here, because of the inherently non-local procedure involved by the T-coercivity approach.

While our numerical analysis assumes an interface-resolving mesh as well as an hypothesis on the contrast, we present numerical experiments with general meshes, that do not necessarily resolve the sign-changing interface(s), as well as contrasts close to the critical interval. These results are very promising, and indicate the efficiency of the method in highly heterogeneous media. Finally, we mention that we consider the diffusion problem here, but the arguments and techniques might also be generalized to other settings such as the Helmholtz equation.

The paper is organized as follows. In Section 2, we introduce our model problem as well as necessary finite element notation. Our generalized finite element method is presented and analyzed in Section 3. The dedicated arguments required to take into account T-coercivity in the context of LOD are discussed in Section 4. In Section 5, we present several numerical experiments illustrating our theory and showing the applicability of the method even for meshes that do not resolve the interface, and contrasts close to the critical interval. Some technical finite element estimates related to quasi-interpolation are collected in Appendix A.

2. Setting

In this section we introduce the model problem, discuss the notion of T-coercivity and introduce the necessary finite element preliminaries. Throughout the whole article, we use standard notation on Sobolev spaces. $\|\cdot\|_0$ denotes the usual L^2-norm and $\|\cdot\|_1$ and $|\cdot|_1$ denote the H^1-norm and semi norm, respectively. We add an additional subscript for the domain over which the norm is taken if necessary.

2.1. Model problem and T-coercivity

We consider a polytopal domain $\Omega \subset \mathbb{R}^d$, with $d \in \{1, 2, 3\}$. We assume that $\overline{\Omega} = \overline{\Omega_+} \cup \overline{\Omega_-}$, where $\Omega_\pm \subset \Omega$ are two non-overlapping subdomains. We denote by

$$\Gamma := \partial\Omega_+ \cap \partial\Omega_-$$

the boundary shared by the two subdomains.

We consider a diffusion coefficient $\sigma \in L^\infty(\Omega)$ such that $\sigma|_{\Omega_-} \leq -\sigma_-$ and $\sigma|_{\Omega_+} \geq \sigma_+$, where $0 < \sigma_+ \leq \sigma_- < +\infty$ are fixed real numbers. Given $f \in L^2(\Omega)$, we seek $u \in H^1_0(\Omega)$ such that

$$a(u, v) = (f, v), \tag{2.1}$$

where

$$a(u, v) := (\sigma\boldsymbol{\nabla} u, \boldsymbol{\nabla} v)$$

For the sake of simplicity, we introduce the norm

$$|v|^2_{1,\sigma,\Omega} := \int_\Omega |\sigma||v|^2 \quad \forall v \in H^1_0(\Omega),$$

that is equivalent to the usual $\|\cdot\|_{1,\Omega}$-norm.

While the bilinear form $a(\cdot, \cdot)$ is not coercive, well-posedness in the sense of Hadamard can be guaranteed through a weaker property usually called T-coercivity, which is a particular case of "inf-sup stability". For convenience, we give a short illustration how T-coercivity is obtained for our model problem.

Proposition 2.1. *Let $V \subset H^1_0(\Omega)$ be a closed subspace. Assume that there exists an operator $\mathrm{T} \in \mathcal{L}(V)$ such that*

$$(\mathrm{T}v)|_{\Omega_-} = -v|_{\Omega_-} \tag{2.2a}$$

and

$$|v - \mathrm{T}v|_{1,\Omega_+} \leq C_\pm(\mathrm{T})|v|_{1,\Omega_-}, \tag{2.2b}$$

for all $v \in V$. Then, we have

$$a(v, \mathrm{T}v) \geq \left(1 - \frac{C_\pm(\mathrm{T})}{2}\left(\frac{\sup_{\Omega_+}\sigma}{\inf_{\Omega_+}\sigma}\right)\sqrt{\frac{\sigma_+}{\sigma_-}}\right)|v|^2_{1,\sigma,\Omega} \tag{2.3}$$

for all $v \in V$.

Proof. Pick an arbitrary element $v \in V$. Taking advantage of (2.2a), we may write

$$\begin{aligned} a(v, \mathrm{T}v) &= (\sigma\boldsymbol{\nabla} v, \boldsymbol{\nabla}(\mathrm{T}v))_{\Omega_+} - (\sigma\boldsymbol{\nabla} v, \boldsymbol{\nabla}(\mathrm{T}v))_{\Omega_-} \\ &= (\sigma\boldsymbol{\nabla} v, \boldsymbol{\nabla}(\mathrm{T}v))_{\Omega_+} + (\sigma\boldsymbol{\nabla} v, \boldsymbol{\nabla} v)_{\Omega_-} \\ &= |v|^2_{1,\sigma,\Omega} - (\sigma\boldsymbol{\nabla} v, \boldsymbol{\nabla}(v - \mathrm{T}v))_{\Omega_+} \end{aligned} \tag{2.4}$$

Then, we derive that

$$|(\sigma\boldsymbol{\nabla} v, \boldsymbol{\nabla}(v - \mathrm{T}v))_{\Omega_+}| \leq (\sup_{\Omega_+}\sigma)|v|_{1,\Omega_+}|v - \mathrm{T}v|_{1,\Omega_+} \tag{2.5}$$

$$\leq \left(\frac{\sup_{\Omega_+}\sigma}{\inf_{\Omega_+}\sigma}\right)\sigma_+|v|_{1,\Omega_+}|v - \mathrm{T}v|_{1,\Omega_+} \tag{2.6}$$

$$\leq C_{\pm}(\mathrm{T}) \left(\frac{\sup_{\Omega_+} \sigma}{\inf_{\Omega_+} \sigma} \right) \sigma_+ |v|_{1,\Omega_+} |v|_{1,\Omega_-}$$
$$\leq \frac{C_{\pm}(\mathrm{T})}{2} \left(\frac{\sup_{\Omega_+} \sigma}{\inf_{\Omega_+} \sigma} \right) \sqrt{\frac{\sigma_+}{\sigma_-}} |v|^2_{1,\sigma,\Omega},$$

where we have employed Young's inequality

$$\sigma_+ |v|_{1,\Omega_+} |v|_{1,\Omega_-} = \sqrt{\frac{\sigma_+}{\sigma_-}} \sqrt{\sigma_+} |v|_{1,\Omega_+} \sqrt{\sigma_-} |v|_{1,\Omega_-} \leq \frac{1}{2} \sqrt{\frac{\sigma_+}{\sigma_-}} |v|^2_{1,\sigma,\Omega}.$$

Estimate (2.3) then follows from (2.4) and (2.5). □

In the following we will assume that there exists an operator $\mathrm{T} \in \mathcal{L}(H^1_0(\Omega))$ satisfying (2.2). We will assume that

$$\sqrt{\frac{\sigma_-}{\sigma_+}} > \frac{C_{\pm}(\mathrm{T})}{2} \left(\frac{\sup_{\Omega_+} \sigma}{\inf_{\Omega_+} \sigma} \right) \tag{2.7}$$

and employ the notation

$$\alpha := 1 - \frac{C_{\pm}(\mathrm{T})}{2} \left(\frac{\sup_{\Omega_+} \sigma}{\inf_{\Omega_+} \sigma} \right) \sqrt{\frac{\sigma_+}{\sigma_-}}$$

Since $\alpha > 0$, estimate (2.3) shows that $a(\cdot,\cdot)$ is inf-sup stable, and (2.1) is well-posed.

We will further assume that T takes the particular form

$$\mathrm{T}u = \begin{cases} -u & \text{in } \Omega_- \\ u - 2\mathrm{S}u & \text{on } \Omega_+, \end{cases}$$

where $\mathrm{S} : H^1_{\partial\Omega}(\Omega_-) \to H^1_{\partial\Omega}(\Omega_+)$ is a "symmetrization" operator satisfying $\mathrm{S}v|_\Gamma = v|_\Gamma$ for all $v \in H^1_{\partial\Omega}(\Omega_-)$, and there exists a constant $C^0_{\pm}(\mathrm{T})$ such that

$$\|v - \mathrm{T}v\|_{0,\Omega_+} \leq C^0_{\pm}(\mathrm{T}) \|v\|_{0,\Omega_-} \qquad \forall v \in L^2(\Omega). \tag{2.8}$$

We point out that the above assumptions are not very restrictive. In particular, we refer the reader to [2] for an explicit construction of S employing local geometrical transformations (rotations and symmetry) for general polygonal interfaces. We also note that (2.2b) holds with

$$C_{\pm}(\mathrm{T}) = 2\|\mathrm{S}\|_{\mathcal{L}(H^1_{\partial\Omega}(\Omega_-);H^1_{\partial\Omega}(\Omega_+))}.$$

Remark 2.2. In the above, we (arbitrarily) assumed that $\sigma_+ \leq \sigma_-$. This is not a restrictive assumption, since in the case where $\sigma_- \leq \sigma_+$, we can always get back to this situation by applying a minus sign on both sides of (2.1). In particular, when we write that the contrast is "sufficiently large", it actually means it is "sufficiently far away from the critical interval".

2.2. Quasi-interpolation and T_H-coercivity

2.2.1. Notations related to the mesh

We consider a shape-regular quasi-uniform triangulation $\mathcal{T}_H$ of Ω. We assume that $\mathcal{T}_H$ that resolves the interface, i.e., that Γ is covered by faces. The standard conforming finite element space of lowest order Lagrange elements is denoted by $V_H \subset H^1_0(\Omega)$.

Given an element $K \in \mathcal{T}_H$ the notations

$$H_K := \sup_{x,y \in K} |x - y| \qquad \rho_K := \sup\{r > 0 \mid \exists x \in K;\ B(x,r) \subset K\}.$$

respectively denote the diameter of K and the radius of the largest balled contained in K. The assumptions of shape-regularity and quasi-uniformity imply the existence of a constant $\kappa > 1$ such that

$$\frac{H}{\rho} \leq \kappa,$$

where $H := \max_{K\in\mathcal{T}_H} H_K$ and $\rho := \min_{K\in\mathcal{T}_H} \rho_K$.

$\mathcal{V}_H$ is the set of vertices of $\mathcal{T}_H$, and $\mathcal{V}_H^{\text{int}}$ is the set of "interior" vertices that do not lie on $\partial\Omega$. If $\mathbf{a} \in \mathcal{V}_H$, we denote by $\psi^{\mathbf{a}}$ the associated hat function and set $\omega^{\mathbf{a}} := \operatorname{supp}\psi^{\mathbf{a}}$. We further split $\mathcal{V}_H^{\text{int}}$ into three categories of vertices:

$$\mathcal{V}_H^- := \left\{\mathbf{a} \in \mathcal{V}_H^{\text{int}} \mid \mathbf{a} \in \Omega_-\right\}, \quad \mathcal{V}_H^+ := \left\{\mathbf{a} \in \mathcal{V}_H^{\text{int}} \mid \mathbf{a} \in \Omega_+\right\}, \quad \mathcal{V}_H^0 := \left\{\mathbf{a} \in \mathcal{V}_H^{\text{int}} \mid \mathbf{a} \in \Gamma\right\}.$$

If $\mathbf{a} \in \mathcal{V}_H$, then

$$\mathcal{T}_H^{\mathbf{a}} := \{K \in \mathcal{T}_H \quad | \quad \mathbf{a} \in \mathcal{V}(K)\},$$

is the associated local mesh, and $\sharp\mathbf{a} := \operatorname{card}\mathcal{T}_H^{\mathbf{a}}$, is the number of elements touching $\mathbf{a}$. Finally, If $K \in \mathcal{T}_H$, then $\mathcal{V}(K) \subset \mathcal{V}_H$ is the set of vertices of K.

2.2.2. Oswald-type quasi-interpolation

Following [26], we consider a standard Oswald-type quasi-interpolation operator $I_H : H_0^1(\Omega) \to V_H$. For $v \in H_0^1(\Omega)$, it is defined as

$$I_H v := \sum_{\mathbf{a}\in\mathcal{V}_H^{\text{int}}} m^{\mathbf{a}}(v)\psi^{\mathbf{a}}, \tag{2.9}$$

with

$$m^{\mathbf{a}}(v) := \frac{1}{\sharp\mathbf{a}} \sum_{K\in\mathcal{T}_H^{\mathbf{a}}} (P_K v)(\mathbf{a}),$$

where $P_K v$ denotes the $L^2(K)$ projection onto $\mathcal{P}_1(K)$. Obviously, I_H is a projection onto V_H ($I_H \circ I_H = I_H$) and we furthermore have

$$\|v - I_H v\|_K + H\|\boldsymbol{\nabla}(v - I_H v)\|_K \lesssim H\|\boldsymbol{\nabla} v\|_{\mathrm{N}(K)} \qquad \forall K \in \mathcal{T}_H \tag{2.10}$$

for all $v \in H_0^1(\Omega)$, see [26]. While for the sake of simplicity, we work with the above mentioned operator I_H, we emphasize that other quasi-interpolation operators could be considered, and we refer the reader to [13] for the required properties.

2.2.3. T_H-coercivity

Our generalized finite element method hinges on variational problems defined on $W := \ker I_H$. While $a(\cdot,\cdot)$ is T-coercive on $H_0^1(\Omega)$, it does not automatically mean that it is T-coercive on W. Indeed, we do not have $\mathrm{T}w \in W$ for $w \in W$ in general. This motivates the introduction of the following modified version of the T operator:

$$\mathrm{T}_H v = \mathrm{T}v - \sum_{\mathbf{a}\in\mathcal{V}_H^0\cup\mathcal{V}_H^+} m^{\mathbf{a}}(\mathrm{T}v)\eta^{\mathbf{a}}, \tag{2.11}$$

for each $v \in H_0^1(\Omega)$, where $\eta^{\mathbf{a}}$ is the function defined in Lemma A.4 in the appendix. We postpone the detailed analysis of T_H, and state its essential properties.

$\mathrm{T}_H \in \mathcal{L}(W)$ satisfies (2.2) with

$$C_{\pm}(\mathrm{T}_H) := C_{\pm}(\mathrm{T}) + \widehat{C}\sqrt{2 + C_{\pm}^0(\mathrm{T})^2}$$

where the constant $\widehat{C}$ only depends on κ and is described in details in Section 4. In addition, as shown in the appendix, $\operatorname{supp}\eta_{\mathbf{a}} \subset \omega^{\mathbf{a}} \cap \Omega_+$ for all $\mathbf{a} \in \mathcal{V}_H$. As a result, we have

$$\operatorname{supp}(\mathrm{T}_H v) \cap \Omega_- = \operatorname{supp}(\mathrm{T} v) \cap \Omega_- \tag{2.12a}$$

and

$$\operatorname{supp}(\mathrm{T}_H v) \cap \Omega_+ = \left\{K \in \mathcal{T}_H \mid \operatorname{supp}(\mathrm{T} v) \cap \overline{K} \neq \emptyset\right\} \cap \Omega_+ \tag{2.12b}$$

for all $v \in H_0^1(\Omega)$.

As advertised in the introduction, we will derive the well-posedness and convergence of the proposed generalized finite element method under a "large contrast" assumption. Specifically, we assume from now on that

$$\sqrt{\frac{\sigma_-}{\sigma_+}} > \frac{C_{\pm}(\mathrm{T}_H)}{2}\left(\frac{\sup_{\Omega_+}\sigma}{\inf_{\Omega_+}\sigma}\right). \tag{2.13}$$

In view of (2.3), we have

$$a(w, \mathrm{T}_H w) \geq \alpha_\kappa |w|^2_{1,\sigma,\Omega} \qquad \forall w \in W \tag{2.14}$$

with

$$\alpha_\kappa := 1 - \frac{C_{\pm}(\mathrm{T}_H)}{2}\left(\frac{\sup_{\Omega_+}\sigma}{\inf_{\Omega_+}\sigma}\right)\sqrt{\frac{\sigma_-}{\sigma_+}},$$

which means that $a(\cdot,\cdot)$ is T_H-coercive on W, and in particular, satisfies an inf-sup condition.

3. The generalized finite element method

In this section, we derive our generalized finite element method and show its well-posedness as well as a priori error estimates. We use the notation introduced in Section 2 and in particular, we assume throughout that (2.1) is well-posed and that (2.13) is satisfied. To avoid the proliferation of constants, we use the notation $a \lesssim b$ (resp. $a \gtrsim b$) if $a \leq Cb$ (resp. $a \geq Cb$) with a constant C that only depends on κ, α_κ, σ_+, σ_-, and $\|\sigma\|_{L^\infty(\Omega)}$. We also write $a \approx b$ when $a \lesssim b$ and $a \gtrsim b$.

3.1. An ideal method

The method is based on the splitting $H_0^1(\Omega) = V_H \oplus W$ which we orthogonalize with respect to $a(\cdot,\cdot)$ in the following sense: The correction operator $\mathcal{Q} : V_H \to W$ is defined via

$$a(\mathcal{Q}v_H, w) = a(v_H, w) \qquad \text{for all} \quad w \in W. \tag{3.1}$$

We note that these corrector problems are automatically well-posed because of (2.14). Furthermore, $\mathcal{Q}$ can be written as $\mathcal{Q} = \sum_{K\in\mathcal{T}_H} \mathcal{Q}_K$, where $\mathcal{Q}_K$ is defined via

$$a(\mathcal{Q}_K v_H, w) = a_K(v_H, w) \qquad \text{for all} \quad w \in W.$$

In the ideal generalized finite element method we use $(\mathrm{id} - \mathcal{Q})V_H$ as new ansatz and test space, i.e., we seek $u_H \in V_H$ such that

$$a((\mathrm{id}-\mathcal{Q})u_H, (\mathrm{id}-\mathcal{Q})v_H) = (f, (\mathrm{id}-\mathcal{Q})v_H) \qquad \text{for all} \quad v_H \in V_H. \tag{3.2}$$

Proposition 3.1. *$a(\cdot,\cdot)$ satisfies the following inf-sup condition: There exists $\tilde{\alpha}_\kappa \approx \alpha_\kappa > 0$, independent of H, such that*

$$\inf_{v_H \in V_H \setminus \{0\}} \sup_{\psi_H \in V_H \setminus \{0\}} \frac{a((\mathrm{id} - \mathcal{Q})v_H, (\mathrm{id} - \mathcal{Q})\psi_H)}{\|v_H\|_1 \, \|\psi_H\|_1} \geq \tilde{\alpha}_\kappa. \tag{3.3}$$

Moreover, the unique solution u_H of (3.2) fulfills the following error estimate

$$\|u - (\mathrm{id} - \mathcal{Q})u_H\|_1 \lesssim H \|f\|_0.$$

Note that the inf-sup condition automatically implies the well-posedness of (3.2). A direct calculation then shows that u_H coincides with the interpolation of the exact solution $I_H u$. Since I_H is a stable quasi-interpolation onto V_H, $u_H = I_H u$ already contains many characteristic coarse features of the exact solution and hence, may be a sufficiently good approximation in many cases. Note that linear convergence of the error in Proposition 3.1 is optimal for lowest-order elements and moreover, that this result is independent of the regularity of the exact solution (which may be arbitrarily low, since $\sigma \in L^\infty(\Omega)$). Proposition 3.1 is classical for the LOD applied to inf-sup stable problems and we refer to [22, Chapter 2] for a proof.

3.2. Localization of the basis

The corrector problems (3.1) are global finescale problems and therefore as expensive to solve as the original problem on a fine mesh. In this section, we will show how to localize the computation of the correctors yielding a practical method. This localization step is motivated by a decay of the correctors which is exponential in units of H.

We emphasize that the present localization analysis requires a dedicated treatment, due to the underlying usage of T-coercivity. Indeed, the arguments for general inf-sup stable problems presented in [22, Chapter 2] requires a "locality assumption" in the inf-sup condition. This locality assumption essentially requires that for $w \in W$, there exists a function $w^\star \in W$ that realizes the inf-sup condition such that $\|w^\star\|_{1,D} \lesssim \|w\|_{\widetilde{D}}$ for $D \subset \Omega$, where $\widetilde{D}$ is a slightly "oversampled" version of D. In view of the nature of the operator T, that involves a symmetrization around Γ, this assumption is fundamentally violated here.

We denote by $\mathrm{N}(K)$ denote the union of all elements sharing a vertex with $K \in \mathcal{T}_H$, i.e.,

$$\mathrm{N}(K) := \{K' \in \mathcal{T}_H \quad | \quad \overline{K'} \cap \overline{K} \neq \emptyset\}.$$

This can inductively be generalized to m-layer patches: Define $\mathrm{N}^m(K) = \mathrm{N}^{m-1}(\mathrm{N}(K))$ for $m \in \mathbb{N}$, $m \geq 2$, and set for simplicity $\mathrm{N}^0(K) = K$. The shape regularity implies that there is a bound $C_{\mathrm{ol},m}$ (depending only on m) of the number of the elements in the m-layer patch, i.e.,

$$\max_{T \in \mathcal{T}_H} \mathrm{card}\{K \in \mathcal{T}_H \quad | \quad K \subset \mathrm{N}^m(T)\} \leq C_{\mathrm{ol},m}. \tag{3.4}$$

We note that since $\mathcal{T}_H$ is quasi-uniform, $C_{\mathrm{ol},m}$ grows at most polynomially with m.

As stated above, we need to modify usual proof because T_H involves a symmetrization operator and thus, is inherently non-local. This is why we introduce the following "symmetric" patches $\mathrm{P}^m(K)$

$$\mathrm{P}^m(K) \cap \Omega_- := \mathrm{N}^m(K) \cap \Omega_-,$$

and

$$\mathrm{P}^m(K) \cap \Omega_+ := \{K' \in \mathcal{T}_H \quad | \quad K' \cap \mathrm{supp}(\mathrm{T}v) \neq \emptyset \quad \text{for all} \quad v \in H^1_0(\mathrm{N}^m(K))\} \cap \Omega_+.$$

We emphasize that this does not require the mesh $\mathcal{T}_H$ to be symmetric. In view of (2.12), the idea of $\mathrm{P}^m(K)$ is that, for any function $v \in H_0^1(\Omega)$ with $\operatorname{supp} v \subset \mathrm{P}^m(K)$ we now have $\operatorname{supp} \mathrm{T}_H v \subset \mathrm{P}^m(K)$ as well. We now have an exponential decay of $\mathcal{Q}_K$ outside those symmetric patches, as stated in the following proposition, whose proof is postponed to Section 3.3.

Proposition 3.2. *There is $0 < \tilde{\gamma} < 1$, independent of H, such that for any $K \in \mathcal{T}_H$*

$$\|\mathcal{Q}_K v_H\|_{1,\Omega\setminus \mathrm{P}^m(K)} \lesssim \tilde{\gamma}^m \|v_H\|_{1,K}.$$

In order to localize the corrector problems, we introduce the space

$$W(\mathrm{P}^m(K)) := \{w \in W \quad | \quad w = 0 \quad \text{in} \quad \Omega \setminus \mathrm{P}^m(K)\}$$

and define for any $v_H \in V_H$ the localized element corrector $\mathcal{Q}_{K,m} v_H \in W(\mathrm{P}^m(K))$ as the solution of

$$a_{\mathrm{P}^m(K)}(\mathcal{Q}_{K,m} v_H, w) = a_K(v_H, w) \qquad \text{for all} \quad w \in W(\mathrm{P}^m(K)), \tag{3.5}$$

where $a_D(\cdot,\cdot)$ denotes the restriction of $a(\cdot,\cdot)$ to a subdomain $D \subset \Omega$. Due to $\mathrm{T}_H \in \mathcal{L}(W)$ and the definition of $\mathrm{P}^m(K)$, these localized corrector problems are well-posed because the T_H-coercivity of $a(\cdot,\cdot)$ thereby carries over from W to $W(\mathrm{P}^m(K))$.

We emphasize that, if $\mathrm{N}^m(K) \cap \Gamma = \emptyset$, $\mathrm{P}^m(K)$ consists of two disconnected domains and $\mathcal{Q}_{K,m} v_H$ is even zero outside the standard patch $\mathrm{N}^m(K)$ because of the localized right-hand side in (3.5). Hence, we can solve (3.5) on $\mathrm{N}^m(K)$ (as in the usual LOD) in the case $\mathrm{N}^m(K) \cap \Gamma = \emptyset$, resulting in the standard localized element corrector problems. In other words, we only need to define new and larger patches for $Q_{K,m}$ for element K close to the interface Γ. The truncated correction operator $\mathcal{Q}_m$ is now defined as the sum of these element correctors, i.e., $\mathcal{Q}_m := \sum_{K\in\mathcal{T}_H} \mathcal{Q}_{K,m}$.

Due to the exponential decay of the idealized correctors, we have the following estimate of the truncation or localization error, which again is proved in Section 3.3.

Theorem 3.3. *There exists $0 < \gamma < 1$, independent of H, such that for any $v_H \in V_H$*

$$\|(\mathcal{Q} - \mathcal{Q}_m) v_H\|_1 \lesssim C_{\mathrm{ol},m}^{1/2}\, \gamma^m \|v_H\|_1.$$

In our generalized finite element method, we now replace $\mathcal{Q}$ in (3.2) by $\mathcal{Q}_m$, exactly in the spirit of LOD. Hence, we seek $u_{H,m} \in V_H$ such that

$$a((\mathrm{id} - \mathcal{Q}_m) u_{H,m}, (\mathrm{id} - \mathcal{Q}_m) v_H) = (f, (\mathrm{id} - \mathcal{Q}_m) v_H) \qquad \text{for all} \quad v_H \in V_H. \tag{3.6}$$

The numerical analysis relies on the error estimate for the ideal method in Proposition 3.1 and the fact that the localization is a small perturbation thereof.

Theorem 3.4. *Let $m \gtrsim |\log(C_{\mathrm{ol},m}^{1/2} \tilde{\alpha}_\kappa)|$ with the inf-sup constant of Proposition 3.1. Then* (3.6) *is well-defined and the unique solution $u_{H,m}$ satisfies the error estimates*

$$\|u - (\mathrm{id} - \mathcal{Q}_m) u_{H,m}\|_1 \lesssim (H + C_{\mathrm{ol},m}^{1/2}\, \gamma^m) \|f\|_0, \tag{3.7}$$

$$\|u - u_{H,m}\|_0 \lesssim H \inf_{v_H \in V_H} |u - v_H|_1 + C_{\mathrm{ol},m}^{1/2}\, \gamma^m (H + C_{\mathrm{ol},m}^{1/2} \gamma^m) \|f\|_0. \tag{3.8}$$

Note that the oversampling condition $m \gtrsim |\log(C_{\mathrm{ol},m}^{1/2} \tilde{\alpha}_\kappa)|$ is independent of H. Since $C_{\mathrm{ol},m}$ grows only polynomially in m, it is fulfillable. With respect to the error estimates, we need to couple $m \approx |\log(C_{\mathrm{ol},m}^{1/2} H)|$ anyway to balance the terms in H and m, which in general is the dominating condition. We summarize that under this (standard) oversampling condition,

the method is well-posed, we have linear convergence in the $H^1(\Omega)$-norm (see (3.7)) and up to quadratic convergence of the FE part in the $L^2(\Omega)$-norm (see (3.8)). Note that the second term in (3.8) is of order H^2 for $m \approx |\log(C_{\mathrm{ol},m}^{1/2} H)|$. The exact convergence rate for the FE part depends on the (higher) regularity of the model problem (encoded in the best approximation of V_H), but we have at least linear convergence. To be more precise, (3.8) gives a convergence order of H^{1+s} if the exact solution is in $H^{1+s}(\Omega)$. This should be contrasted with the convergence order H^{2s} in $L^2(\Omega)$ for the standard FEM.

Proof. The well-posedness of (3.6) follows from an inf-sup condition on $V_{H,m}$ (see [30] for instance). This directly yields quasi-optimality and the error estimate (3.7), where we refer to [22, Chapter 2] for details.

Moreover, a standard duality argument can be employed to show

$$\|u - (\mathrm{id} - \mathcal{Q}_m) u_{H,m}\|_0 \lesssim (H + C_{\mathrm{ol},m}^{1/2}\, \gamma^m) \|u - (\mathrm{id} - \mathcal{Q}_m) u_{H,m}\|_1,$$

i.e., quadratic convergence in the $L^2(\Omega)$-norm. We refer to, e.g., [30] for details.

Finally, we have that

$$\|u - u_{H,m}\|_0 \leq \|u - I_H u\|_0 + \|I_H u - u_{H,m}\|_0 \lesssim H |u - I_H u|_1 + \|I_H u - u_{H,m}\|_1.$$

Due to the stability and projection property of I_H, we have $|u - I_H u|_1 \lesssim \inf_{v_H \in V_H} |u - v_H|_1$ so that it remains to estimate $\|I_H u - u_{H,m}\|_1$. We note that by the definition of $\mathcal{Q}$ and the stability of I_H it holds that

$$\|I_H u - u_{H,m}\|_1 = \|I_H (\mathrm{id} - \mathcal{Q})(I_H u - u_{H,m})\|_1 \lesssim \|(\mathrm{id} - \mathcal{Q})(I_H u - u_{H,m})\|_1.$$

Due to Proposition 3.1, there exists $\psi_H \in V_H$ with $\|\psi_H\| = 1$ such that

$$\|(\mathrm{id} - \mathcal{Q})(I_H u - u_{H,m})\|_1 \leq \tilde{\alpha}_\kappa^{-1} a((\mathrm{id} - \mathcal{Q})(I_H u - u_{H,m}), (\mathrm{id} - \mathcal{Q})\psi_H).$$

The definition of $\mathcal{Q}$, Galerkin orthogonality and Theorem 3.3 give that

$$\begin{aligned}
\|(\mathrm{id} - \mathcal{Q})(I_H u - u_{H,m})\|_1 &\leq \tilde{\alpha}_\kappa^{-1} a((\mathrm{id} - \mathcal{Q}) I_H u - (\mathrm{id} - \mathcal{Q}) u_{H,m}, (\mathrm{id} - \mathcal{Q})\psi_H) \\
&= \tilde{\alpha}_\kappa^{-1}\, a(u - (\mathrm{id} - \mathcal{Q}_m) u_{H,m}, (\mathrm{id} - \mathcal{Q})\psi_H) \\
&= \tilde{\alpha}_\kappa^{-1}\, a(u - (\mathrm{id} - \mathcal{Q}_m) u_{H,m}, (\mathcal{Q}_m - \mathcal{Q})\psi_H) \\
&\lesssim \tilde{\alpha}_\kappa^{-1}\, C_{\mathrm{ol},m}^{1/2}\, \gamma^m \|u - (\mathrm{id} - \mathcal{Q}_m) u_{H,m}\|_1.
\end{aligned}$$

Combination with the estimate for $\|u - (\mathrm{id} - \mathcal{Q}_m) u_{H,m}\|_1$ finishes the proof. □

3.3. Proof of the localization error

This section is devoted to the proofs of Proposition 3.2 and Theorem 3.3. In the proofs we will frequently make use of cut-off functions. We collect some properties for them in the following. Let $\eta \in H^1(\Omega)$ be a function with values in the interval $[0, 1]$ satisfying the bound $\|\boldsymbol{\nabla}\eta\|_{L^\infty(\Omega)} \lesssim H^{-1}$ and let $\mathcal{R} := \mathrm{supp}(\boldsymbol{\nabla}\eta)$. Given any subset $D \subset \Omega$ as the union of elements in $\mathcal{T}_H$, any $w \in W$ satisfies that

$$\|w\|_{L^2(D)} \lesssim H \|\boldsymbol{\nabla} w\|_{L^2(\mathrm{N}(D))}, \tag{3.9}$$

$$\|(\mathrm{id} - I_H)(\eta w)\|_{L^2(D)} \lesssim H \|\boldsymbol{\nabla}(\eta w)\|_{L^2(\mathrm{N}(D))}, \tag{3.10}$$

$$\|\boldsymbol{\nabla}(\eta w)\|_{L^2(D)} \lesssim \|\boldsymbol{\nabla} w\|_{L^2(D \cap \mathrm{supp}\,\eta)} + \|\boldsymbol{\nabla} w\|_{L^2(\mathrm{N}(D \cap \mathcal{R}))}. \tag{3.11}$$

These properties are proved in [14, Lemma 2].

Proof of Proposition 3.2. Fix $K \in \mathcal{T}_H$ and $v_H \in V_H$. Set $\phi := \mathcal{Q}_K v_H \in W$ and $\widetilde{\phi} = (\mathrm{id} - I_H)(\eta\phi)$ with the piecewise linear and globally continuous cut-off function η defined via

$$\eta = 0 \quad \text{in} \quad \mathrm{P}^{m-4}(K), \qquad \eta = 1 \quad \text{in} \quad \Omega \setminus \mathrm{P}^{m-3}(K).$$

We write $\mathcal{R} = \mathrm{supp}(\boldsymbol{\nabla}\eta)$ and use in the following $\mathrm{N}^k(\mathcal{R}) = \mathrm{P}^{m-3+k}(K) \setminus \mathrm{P}^{m-4-k}(K)$. Note that $\|\boldsymbol{\nabla}\eta\|_{L^\infty(\mathcal{R})} \lesssim H^{-1}$. Then

$$\|\phi\|_{1,\Omega\setminus\mathrm{P}^m(K)} = \|\phi - I_H\phi\|_{1,\Omega\setminus\mathrm{P}^m(K)} \leq \|\widetilde{\phi}\|_{1,\Omega}.$$

We have $\mathrm{T}_H\widetilde{\phi} \in W$ with support outside K due to the definition of $\mathrm{P}^m(K)$. Hence,

$$\|\phi\|^2_{1,\Omega\setminus\mathrm{P}^m(K)} \leq \|\widetilde{\phi}\|^2_{1,\Omega} \leq \alpha_\kappa^{-1} a(\widetilde{\phi}, \mathrm{T}_H\widetilde{\phi}) = \alpha_\kappa^{-1} a(\widetilde{\phi} - \phi, \mathrm{T}_H\widetilde{\phi}).$$

Note that $\mathrm{supp}(\widetilde{\phi}-\phi) \cap \mathrm{supp}(\mathrm{T}_H\widetilde{\phi}) \subset \mathrm{N}^1(\mathcal{R})$ and $\|\mathrm{T}_H\tilde{\phi}\|_{\mathrm{N}(\mathcal{R})} \lesssim \|\tilde{\phi}\|_{1,\mathrm{N}^2(\mathcal{R})}$ due to the definitions of $\mathrm{P}^m(K)$ and $\mathcal{R}$. Hence, we obtain with the continuity of $a(\cdot,\cdot)$

$$\begin{aligned}\alpha_\kappa\|\phi\|^2_{1,\Omega\setminus\mathrm{P}^m(K)} &\lesssim \|\widetilde{\phi} - \phi\|_{1,\mathrm{N}^1(\mathcal{R})}\, \|\mathrm{T}_H\widetilde{\phi}\|_{1,\mathrm{N}^1(\mathcal{R})} \\ &\lesssim \|\widetilde{\phi} - \phi\|_{1,\mathrm{N}^1(\mathcal{R})}\, (\|\widetilde{\phi} - \phi\|_{1,\mathrm{N}^2(\mathcal{R})} + \|\phi\|_{1,\mathrm{N}^2(\mathcal{R})}).\end{aligned}$$

Employing that $I_H\phi = 0$ and the properties (3.10) as well as (3.11), we deduce

$$\|\widetilde{\phi} - \phi\|_{1,\mathrm{N}^2(\mathcal{R})} = \|(\mathrm{id} - I_H)((1-\eta)\phi)\|_{1,\mathrm{N}^2(\mathcal{R})} \lesssim \|\phi\|_{1,\mathrm{N}^3(\mathcal{R})}$$

and analogously $\|\widetilde{\phi} - \phi\|_{1,\mathrm{N}^1(\mathcal{R})} \lesssim \|\phi\|_{1,\mathrm{N}^2(\mathcal{R})}$. All in all, this gives

$$\|\phi\|^2_{1,\Omega\setminus\mathrm{P}^m(K)} \leq \tilde{C}\|\phi\|^2_{1,\mathrm{P}^m(K)\setminus\mathrm{P}^{m-7}(K)} = \tilde{C}\|\phi\|^2_{1,\Omega\setminus\mathrm{P}^{m-7}(K)} - \tilde{C}\|\phi\|^2_{1,\Omega\setminus\mathrm{P}^m(K)}$$

for some constant $\tilde{C}$. This yields

$$\|\phi\|^2_{1,\Omega\setminus\mathrm{P}^m(K)} \leq \frac{\tilde{C}}{1+\tilde{C}}\, \|\phi\|^2_{1,\Omega\setminus\mathrm{P}^{m-7}(K)}$$

The repeated application of this argument finishes the proof with $\tilde{\gamma} = \frac{\tilde{C}}{1+\tilde{C}} < 1$. □

Note that the constant hidden in $\lesssim$ in Proposition 3.2 depends on the interpolation constant, the norm of T_H, the continuity constant of $a(\cdot,\cdot)$ and on α_κ^{-1}. In particular the latter may become very large depending on the contrast, see [12] and Section 5.

Proof of Theorem 3.3. We start by proving the following local estimate

$$\|(\mathcal{Q}_K - \mathcal{Q}_{K,m})v_H\|_1 \lesssim \tilde{\gamma}^m \|v_H\|_{1,K} \tag{3.12}$$

for some $0 < \tilde{\gamma} < 1$ and for any $v_H \in V_H$ and $K \in \mathcal{T}_H$. Note that $\mathcal{Q}_{K,m}v_H$ is the Galerkin approximation of $\mathcal{Q}_K v_H$ on the subspace $W(\mathrm{P}^m(K)) \subset W$. Due to the T_H-coercivity of $a(\cdot,\cdot)$ over $W(\mathrm{P}^m(K))$, we have the following standard quasi-optimality

$$\|(\mathcal{Q}_K - \mathcal{Q}_{K,m})v_H\|_1 \lesssim \inf_{w_{K,m} \in W(\mathrm{P}^m(K))} \|\mathcal{Q}_K v_H - w_{K,m}\|_1. \tag{3.13}$$

We choose now $w_{K,m} := (\mathrm{id} - I_H)(\eta\mathcal{Q}_K v_H)$ with a piecewise linear, globally continuous cut-off function η defined via

$$\eta = 0 \quad \text{in} \quad \Omega \setminus \mathrm{P}^m(K), \qquad \eta = 1 \quad \text{in} \quad \mathrm{P}^{m-2}(K).$$

Inserting this choice of $w_{K,m}$ into (3.13) and noting that $I_H(\mathcal{Q}_K v_H) = 0$, we obtain

$$\|(\mathcal{Q}_K - \mathcal{Q}_{K,m})v_H\|_1 \lesssim \|(\mathrm{id} - I_H)((1-\eta)\mathcal{Q}_K v_H)\|_1 \lesssim \|\mathcal{Q}_K v_H\|_{1,\Omega\setminus \mathrm{P}^m(K)},$$

where the last inequality follows from the properties (3.10) and (3.11) similar to the arguments in the proof of Proposition 3.2. Combination with Proposition 3.2 gives (3.12).

To prove Theorem 3.3, we define, for a given simplex $K \in \mathcal{T}_H$, the piecewise linear, globally continuous cut-off function η_K via

$$\eta_K = 0 \quad \text{in} \quad \mathrm{P}^{m+1}(K), \qquad \eta_K = 1 \quad \text{in} \quad \Omega \setminus \mathrm{P}^{m+2}(K).$$

For a given $v_H \in V_H$, denote $w := (\mathcal{Q} - \mathcal{Q}_m)v_H = \sum_{K\in\mathcal{T}_H} w_K$ with $w_K := (\mathcal{Q}_K - \mathcal{Q}_{K,m})v_H$. By the T_H-coercivity of $a(\cdot,\cdot)$ over W, we have

$$\alpha_\kappa \|w\|_{1,\Omega} \lesssim \alpha_\kappa |w|^2_{1,\sigma,\Omega} \le \sum_{K\in\mathcal{T}_H} a(w_K, \mathrm{T}_H w) \le \sum_{K\in\mathcal{T}_H} (A_K + B_{K,1} + B_{K,2}),$$

where, for any $K \in \mathcal{T}_H$, we abbreviate

$$A_K := |a(w_K, (1-\eta_K)\mathrm{T}_H w)|, \quad B_{K,1} := |a(w_K, (\mathrm{id} - I_H)(\eta_K \mathrm{T}_H w)|, \quad B_{K,2} := |a(w_K, I_H(\eta_K \mathrm{T}_H w))|.$$

Because $(\mathrm{id} - I_H)(\eta \mathrm{T}_H w) \in W$ with support outside $\mathrm{P}^m(K)$, we have $B_{K,1} = 0$. Using the property (3.11), the stability of I_H (2.10) and $\|\mathrm{T}_H w\|_{\mathrm{N}(\{\eta\neq 1\})} \lesssim \|w\|_{\mathrm{N}^2(\{\eta\neq 1\})}$, we deduce

$$A_K \lesssim \|w_K\|_1 \, \|w\|_{1,\mathrm{N}^1(\{\eta\neq 1\})}, \qquad B_{K,2} \lesssim \|w_K\|_1 \, \|w\|_{1,\mathrm{N}^2(\{\eta\neq 1\})}.$$

Combining these estimates and observing that $\{\eta \neq 1\} = \mathrm{P}^{m+2}(K)$, we obtain

$$\alpha_\kappa \|w\|_1^2 \lesssim \sum_{K\in\mathcal{T}_H} \|w_K\|_1 \|w\|_{1,\mathrm{P}^{m+4}(K)} \lesssim C_{\mathrm{ol},m}^{1/2} \|w\|_1 \Big(\sum_{K\in\mathcal{T}_H} \|w_K\|_1^2 \Big)^{1/2},$$

which in combination with (3.12) finishes the proof. □

3.4. Further remarks on the method

3.4.1. Weak T-coercivity

In this paragraph, we briefly discuss how our results transfer to the case that $a(\cdot,\cdot)$ is weakly T-coercive, which means that instead of (2.3), $a(v, \mathrm{T}v)$ only satisfies a Gårding-type inequality [6], namely

$$a(v, \mathrm{T}v) \ge \alpha |v|^2_{1,\sigma,\Omega} - \mu \|v\|^2_{0,\Omega},$$

where $\alpha, \mu > 0$ are positive constants, which yields a setting similar to the Helmholtz equation.

Assuming in addition that $a(\cdot,\cdot)$ satisfies an inf-sup condition, the problem can be approximated with the proposed generalized finite element method, but the described theory does not immediately apply. In particular, the study of the well-posedness of the corrector problems and their exponential decay requires additional arguments.

However, a similar situation, namely the Helmholtz equation (with positive coefficients), was analyzed in [14, 27, 30]. In particular, it is shown that the corrector problems are well-posed under a resolution condition on H because the L^2-perturbation in the Gårding inequality can be absorbed for functions in the kernel W due to the property (2.10) of I_H. The authors believe that this argument carries over to the weakly T-coercive setting for problems with sign-changing coefficients, so that we can establish strong T_H-coercivity of $a(\cdot,\cdot)$ over W under a resolution condition (smallness assumption) on H.

3.4.2. Fully discrete method

Although the corrector problems (3.5) are localized, they are not yet ready to use since the space W is still infinite-dimensional. In practice we therefore introduce a second, fine triangulation $\mathcal{T}_h$ of Ω as well as the corresponding Lagrange finite element space V_h. The corrector problems (3.5) are then defined on the discrete space $W(\mathrm{P}^m(K)) \cap V_h$.

The corresponding solution $u_{H,h,m}$ of our generalized finite element method (3.6) then approximates the FEM solution $u_h \in V_h$ on the fine mesh. In particular, u in Theorem 3.4 is replaced by u_h. By the triangle inequality, this gives error estimates for $u - u_{H,h,m}$ if $u - u_h$ is small. In other words, we assume that u_h is a good approximation of the exact solution u.

This requires the mesh $\mathcal{T}_h$ to be sufficiently fine, and in particular it needs to be T-conforming. We point out than then, $\mathrm{T}(V_h) \subset V_h$, and one easily checks that $\mathrm{T}_H(W \cap V_h) \subset (W \cap V_h)$. As a result, the authors strong believe the above analysis will still hold true with minor modifications due to the additional discretization. We refer the reader to [14] for details on the proof of the exponential decay in this case.

4. T-coercivity in the kernel of I_H

In this section, we analyze the operator T_H introduced at the end of Section 2.2. Specifically, we show that the kernel W of the Oswald-type interpolation operator I_H specified in (2.9) is stable under application of T_H, and that T_H satisfies (2.2). In view of Proposition 2.1, this ensures that for a sufficiently large contrast, the corrector problems (3.1) are well-posed. We first establish that T_H satisfies (2.2a).

Lemma 4.1. *Let $w \in W$, it holds that*

$$(\mathrm{T}_H w)|_{\Omega_-} = -w|_{\Omega_-}. \tag{4.1}$$

Proof. This is direct consequence from the fact that $\operatorname{supp} \eta^{\mathbf{a}} \subset \Omega_+$ for all $\mathbf{a} \in \mathcal{V}_H^0 \cup \mathcal{V}_H^+$. □

We now show that $\mathrm{T}_H \in \mathcal{L}(W)$.

Lemma 4.2. *We have $\mathrm{T}_H \in \mathcal{L}(W)$ with the operator norm bounded independently of H.*

Proof. We need to show that $\mathrm{T}_H w \in W$ for every $w \in W$. Let us thus pick an arbitrary $w \in W$, so that

$$m^{\mathbf{a}}(w) = 0 \quad \forall \mathbf{a} \in \mathcal{V}_H^{\mathrm{int}}. \tag{4.2}$$

Then, let $\mathbf{a} \in \mathcal{V}_H^{\mathrm{int}}$, we have

$$\begin{aligned} m^{\mathbf{a}}(\mathrm{T}_H w) &= m^{\mathbf{a}}(\mathrm{T}w) - \sum_{\mathbf{a}' \in \mathcal{V}_H^0 \cup \mathcal{V}_H^+} m^{\mathbf{a}'}(\mathrm{T}w) m^{\mathbf{a}}(\eta^{\mathbf{a}'}) \\ &= m^{\mathbf{a}}(\mathrm{T}w) - \sum_{\mathbf{a}' \in \mathcal{V}_H^0 \cup \mathcal{V}_H^+} m^{\mathbf{a}'}(\mathrm{T}w) \delta_{\mathbf{a}',\mathbf{a}}, \end{aligned}$$

and it follows that $m^{\mathbf{a}}(\mathrm{T}_H v) = 0$ whenever $\mathbf{a} \in \mathcal{V}_H^0 \cup \mathcal{V}_H^+$. If on the other hand $\mathbf{a} \in \mathcal{V}_H^-$, recalling (4.1) and observing that $\omega^{\mathbf{a}} \subset \Omega_-$, we have

$$m^{\mathbf{a}}(\mathrm{T}_H w) = m^{\mathbf{a}}(\mathrm{T}w) = -m^{\mathbf{a}}(w) = 0$$

since $w \in W$. This shows that $I_H(\mathrm{T}_H v) = 0$. The H-independent bound on the operator norm of T_H follows by the scalings of $m^{\mathbf{a}}$ and $\eta^{\mathbf{a}}$ in the appendix, see Lemmas A.1 and A.4. □

We now establish that stability property (2.2b) holds true for T_H, under the additional stability assumption (2.8) on T.

Lemma 4.3. *Assume that* T *satisfies* (2.8)*. Then, we have*

$$|w - \mathrm{T}_H w|_{1,\Omega_+} \leq C_\pm(\mathrm{T}_H)|w|_{1,\Omega_-}$$

for all $w \in W$*, with*

$$C_\pm(\mathrm{T}_H) = C_\pm(\mathrm{T}) + 2(d+1)\widehat{C}_{\mathrm{P}}\widehat{C}_m\widehat{C}_{\mathrm{dual}}\beta\kappa\sqrt{2 + C_\pm^0(\mathrm{T})^2},$$

where κ*,* $C_\pm(\mathrm{T})$*, and* $C_\pm^0(\mathrm{T})$ *are introduced in Section 2.2 and the other constants are explained in Appendix A.*

Proof. Let $w \in W$. We have

$$\begin{aligned} w - \mathrm{T}_H w &= w - \Big(\mathrm{T}w - \sum_{\mathbf{a}\in\mathcal{V}_H^0\cup\mathcal{V}_H^+} m^{\mathbf{a}}(\mathrm{T}w)\eta^{\mathbf{a}}\Big) \\ &= \Big(w - \mathrm{T}w\Big) - \sum_{\mathbf{a}\in\mathcal{V}_H^0\cup\mathcal{V}_H^+} m^{\mathbf{a}}(w - \mathrm{T}w)\eta^{\mathbf{a}}, \end{aligned}$$

so that

$$|w - \mathrm{T}_H w|_{1,\Omega_+} \leq |w - \mathrm{T}w|_{1,\Omega_+} + \Big|\sum_{\mathbf{a}\in\mathcal{V}_H^0\cup\mathcal{V}_H^+} m^{\mathbf{a}}(w - \mathrm{T}w)\eta^{\mathbf{a}}\Big|_{1,\Omega_+}.$$

We have

$$\Big|\sum_{\mathbf{a}\in\mathcal{V}_H^0\cup\mathcal{V}_H^+} (w - \mathrm{T}w, m^{\mathbf{a}})\eta^{\mathbf{a}}\Big|_{1,\Omega_+}^2 \leq (d+1)\sum_{\mathbf{a}\in\mathcal{V}_H^0\cup\mathcal{V}_H^+} |m^{\mathbf{a}}(w - \mathrm{T}w)|^2|\eta^{\mathbf{a}}|_{1,\Omega_+}^2.$$

Then, for each $\mathbf{a} \in \mathcal{V}_H^0 \cup \mathcal{V}_H^+$, it holds with Lemmas A.1 and A.4 that

$$\begin{aligned} |m^{\mathbf{a}}(w - \mathrm{T}w)||\eta^{\mathbf{a}}|_{1,\Omega_+} &\leq \widehat{C}_m\widehat{C}_{\mathrm{dual}}\left(\frac{1}{\min_{K\in\mathcal{T}_H^{\mathbf{a}}}|K|}\right)^{1/2}\left(\max_{K\in\mathcal{T}_H^{\mathbf{a}}}\frac{|K|^{1/2}}{\rho_K}\right)\|w - \mathrm{T}w\|_{0,\omega^{\mathbf{a}}} \\ &\leq \widehat{C}_m\widehat{C}_{\mathrm{dual}}\left(\frac{\max_{K\in\mathcal{T}_H^{\mathbf{a}}}|K|}{\min_{K\in\mathcal{T}_H^{\mathbf{a}}}|K|}\right)^{1/2}\frac{1}{\rho}\|w - \mathrm{T}w\|_{0,\omega^{\mathbf{a}}} \\ &\leq \frac{\widehat{C}_m\widehat{C}_{\mathrm{dual}}\beta}{\rho}\|w - \mathrm{T}w\|_{0,\omega^{\mathbf{a}}}. \end{aligned}$$

Furthermore, we have

$$\begin{aligned} \|w - \mathrm{T}w\|_{0,\omega^{\mathbf{a}}}^2 &= \|w - \mathrm{T}w\|_{0,\omega^{\mathbf{a}}\cap\Omega_-}^2 + \|w - \mathrm{T}w\|_{0,\omega^{\mathbf{a}}\cap\Omega_+}^2 \\ &= 2\|w\|_{0,\omega^{\mathbf{a}}\cap\Omega_-}^2 + \|w - \mathrm{T}w\|_{0,\omega^{\mathbf{a}}\cap\Omega_+}^2. \end{aligned}$$

Therefore, we obtain by combining these two estimates

$$\begin{aligned} \Big|\sum_{\mathbf{a}\in\mathcal{V}_H^0\cup\mathcal{V}_H^+} (w - \mathrm{T}v, m^{\mathbf{a}})\eta^{\mathbf{a}}\Big|^2 &\leq \Big(\frac{\widehat{C}_m\widehat{C}_{\mathrm{dual}}\beta}{\rho}\Big)^2(d+1)^2\big(2\|w\|_{0,\Omega_-}^2 + \|w - \mathrm{T}w\|_{0,\Omega_+}^2\big) \\ &\leq \Big(\frac{\widehat{C}_m\widehat{C}_{\mathrm{dual}}\beta}{\rho}\Big)^2(d+1)^2\left(2 + C_\pm^0(\mathrm{T})^2\right)\|w\|_{0,\Omega_-}^2. \end{aligned}$$

Moreover, we have by Lemma A.2

$$\|w\|_{0,\Omega_-}^2 \leq \frac{1}{d+1}\sum_{\mathbf{a}\in\mathcal{V}_H^-}\|w\|_{0,\omega^{\mathbf{a}}}^2 \leq \frac{(\widehat{C}_{\mathrm{P}}H)^2}{d+1}\sum_{\mathbf{a}\in\mathcal{V}_H^-}|w|_{1,\omega^{\mathbf{a}}}^2 \leq \widehat{C}_{\mathrm{P}}^2H^2|w|_{1,\Omega_-}^2.$$

Hence, combining all the foregoing estimates, we finally deduce

$$\left|\sum_{\mathbf{a}\in\mathcal{V}_H^0\cup\mathcal{V}_H^+}(w-\mathrm{T}w,m^{\mathbf{a}})\eta^{\mathbf{a}}\right| \leq (d+1)\widehat{C}_{\mathrm{P}}\widehat{C}_m\widehat{C}_{\mathrm{dual}}\beta\frac{H}{\rho}\sqrt{2+C_{\pm}^0(\mathrm{T})^2}\,|w|_{1,\Omega_-},$$

and the result follows. □

5. Numerical experiments

In the following numerical examples, we always consider $\Omega = [0,1]^2$. The diffusion coefficient σ is chosen piecewise constant. Specifically, Ω is partitioned into two subsets Ω_- and Ω_+, and $\sigma_\pm$ denotes the value taken by σ over $\Omega_\pm$. The corrector computations are discretized on a fine mesh of criss-cross type with $h = 2^{-8}$ which is T-conform in all settings described below except from the circular inclusion in Section 5.3. In Sections 5.1 and 5.3 the exact solution is known, whereas we compute a reference solution u_h with a standard FEM on the fine mesh in Sections 5.2 and 5.4. The LOD solution is computed on a series of meshes with $H = 2^{-1},\dots,2^{-6}$ and oversampling parameters $m \in \{1,2,3\}$. We refer to $(\mathrm{id}-\mathcal{Q}_m)u_{H,m}$ from (3.6) as the LOD solution and to $u_{H,m}$ as the macroscopic part of the LOD solution. Note that $u_{H,m}$ lies in the standard FE space. For comparison, we also compute the standard FE solution on the coarse grids $\mathcal{T}_H$ as well as the $L^2(\Omega)$-projection of the exact or reference solution onto V_H. The latter is referred to as the L^2-best approximation in V_H. We compute the absolute error of the LOD solution in the $H^1(\Omega)$-semi norm and compare it to the the absolute error of the standard FEM. From (3.7), we expect linear convergence of this LOD error. Moreover, we also consider the absolute error of the macroscopic part of the LOD solution in the $L^2(\Omega)$-norm and compare it to the absolute errors of the FEM solution and the L^2-best approximation in V_H. We expect that the macroscopic error of the LOD behaves like the L^2-best approximation error (cf. (3.8)).

Finally, we note that, although our theory guarantees a well-posedness of the corrector problems only if the contrast is outside a sufficiently large interval, which is larger than the analytical one, we never experienced any well-posedness issues in practice.

5.1. Flat interface with known exact solution

We define $\Omega_+ = \{x\in\Omega \mid x_2 < 0.5-2^{-7}\}$ and Ω_- accordingly as $\Omega_- = \{x\in\Omega \mid x_2 > 0.5-2^{-7}\}$. We set $\sigma_+ = 1$ and consider two different cases where $\sigma_- = 2$ or 1.1. Note that the model problem is well-posed for both choices of σ_- since the critical interval in this case only consists of the value -1. The meshes $\mathcal{T}_H$ do not resolve the interface and are not symmetric for any H and hence we expect a poor performance of the standard FEM. We consider the following piecewise smooth function fulfilling homogeneous Dirichlet boundary conditions

$$u(x,y) = \begin{cases} -\sigma_- x(x-1)y(y-1)(y-l), & (x,y)\in\Omega_+,\\ x(x-1)y(y-1)(y-l), & (x,y)\in\Omega_-,\end{cases}$$

where $l = 0.5-2^{-7}$ stands for the interface location. The right-hand side f is computed so that u is the exact solution. Precisely, $f(x,y) = \sigma_-(2y(y-1)(y-l)+x(x-1)(6y-2(l+1))$ and we note that f is globally smooth.

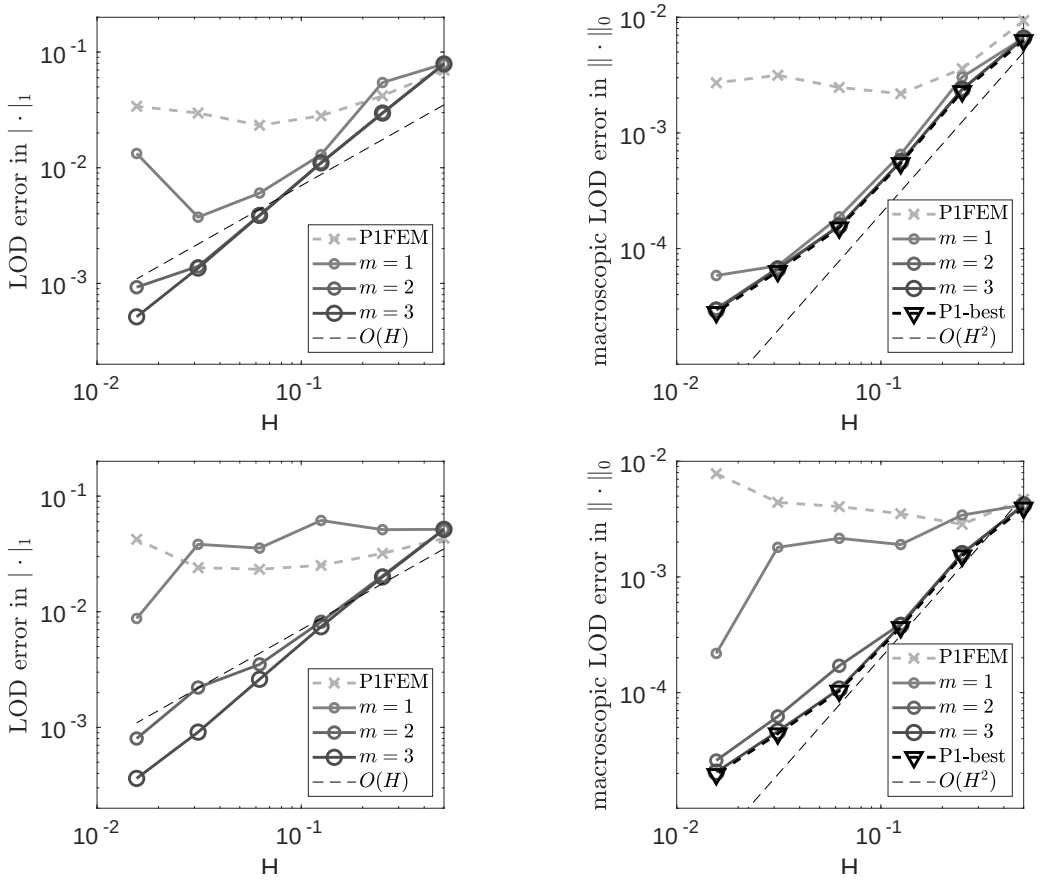

Figure 5.1: Convergence histories for the flat interface with $\sigma_- = 2$ (top) and $\sigma_- = 1.1$ (bottom) in Section 5.1.

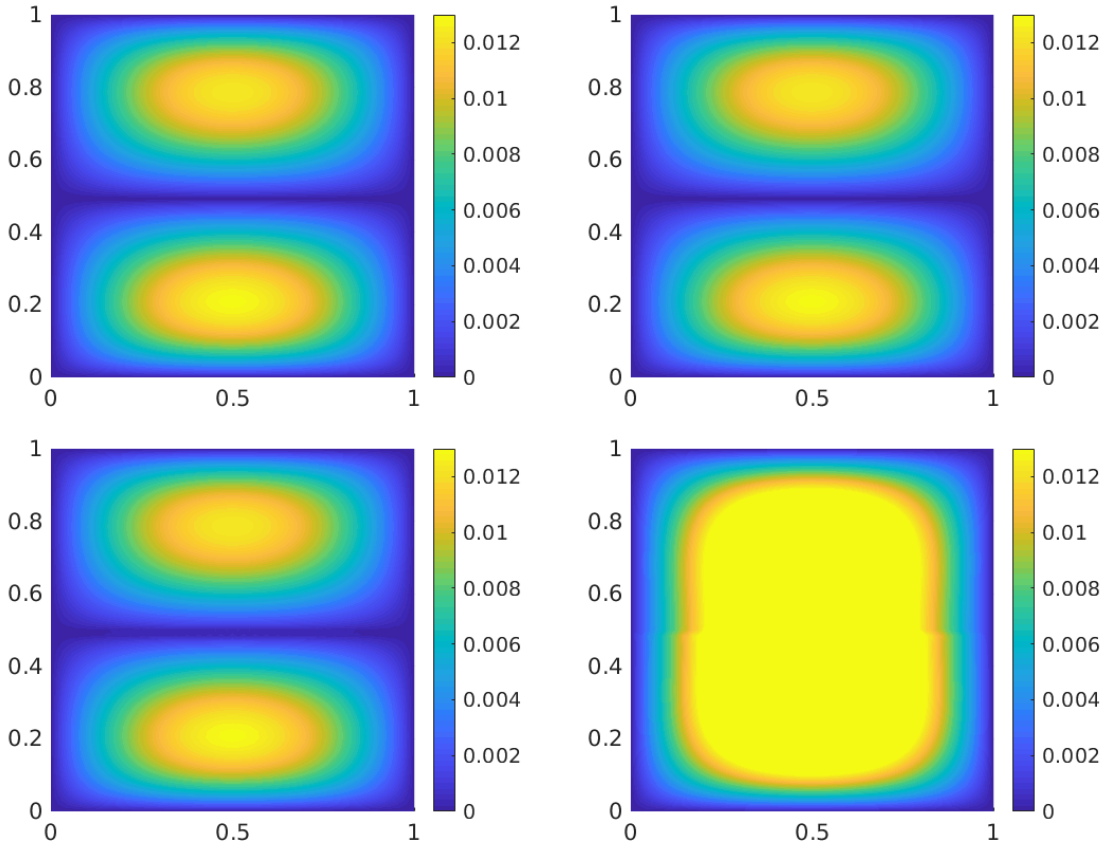

Figure 5.2: Various solutions for the flat interface with $\sigma_- = 1.1$ (Section 5.1): exact solution u (top left), LOD solution (top right), macroscopic part of the LOD solution ($u_{H,m}$, bottom left) and FE solution (bottom right).

The LOD error (in the $H^1(\Omega)$-semi norm) and the macroscopic LOD error (in the $L^2(\Omega)$-norm) for both choices of σ_- are depicted in Figure 5.1. We observe that an oversampling parameter $m = 3$ is sufficient to produce faithful LOD approximations. The LOD error in both cases converges linearly as expected and the macroscopic LOD error follows the L^2-best approximation. Note that the latter converges quadratically due to the piecewise smoothness of u. This nicely illustrates the findings of Theorem 3.4. In contrast to the good performance of the LOD, we see the failure of the standard FEM in Figure 5.1. Moreover, we observe that for $\sigma_- = 1.1$ we should select $m = 3$ as oversampling parameter in the LOD, whereas for $\sigma_- = 2$, $m = 2$ already yields good results, see Figure 5.1 top and bottom left. This effect is connected to the $\tilde{\alpha}_\kappa$-dependency of the exponential decay: Since $\sigma_- = 1.1$ is close to the critical interval, this constant in the T_H-coercivity is small so that the decay of the corrector is slow, which results in a larger oversampling region.

We now compare for $H = 2^{-6}$ and $m = 3$ the LOD solution, its macroscopic part, and the FE solution to the exact solution in the case $\sigma_- = 1.1$, see Figure 5.2. Strikingly, the FE solution has almost no resemblance with the exact solution, but the macroscopic part of the LOD (which lies in the same space V_H) is very close to the exact solution. For this example, one can hardly make out any differences between the exact solution, the LOD solution and its macroscopic part,

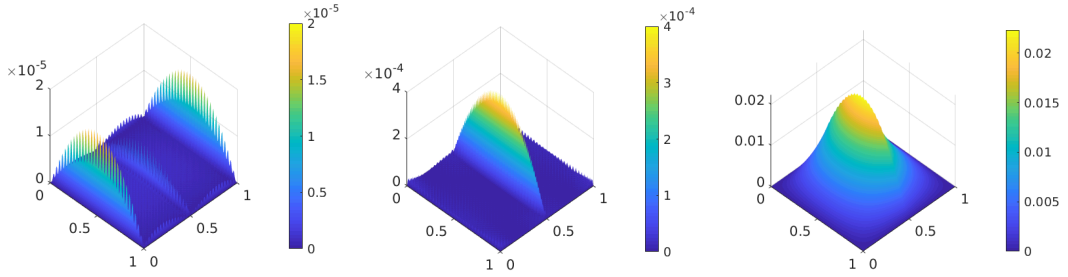

Figure 5.3: Errors of the different solutions to u for the interface with $\sigma_2 = -1.1$ (Section 5.1): error of LOD solution (left), error of macroscopic part of LOD solution (middle), and error of FE solution (right).

which clearly underlines the potential of our method. In particular, we emphasize once more that good approximations (in an $L^2(\Omega)$-sense) exist in the coarse FE space V_H, which are found by our approach but not by the standard finite element method.

To see more details, we visualize the absolute errors of the three solutions (i.e., $(\mathrm{id} - \mathcal{Q}_m)u_{H,m}$, $u_{H,m}$ and u_H) to u in Figure 5.3. Here, we clearly see a difference in the error distribution. The FE error (right) is very large close to the interface and this errors spreads out over a large part of the domain. In contrast, the macroscopic part of the LOD solution (middle) has a much smaller error which is furthermore very confined to the interface. A localization of the error close to the interface is expected because on the one hand, this jump in the coefficient is not resolved by the mesh and because on the other hand, the interesting effects happen there. In the full LOD solution (left), the error at the interface is largely reduced by the upscaling procedure so that interface and boundary errors are now of the same order.

5.2. Square inclusion

We consider $\Omega_- = [0.25+2^{-7}, 0.75-2^{-7}]$ and Ω_+ as the complement. For such a square inclusion, the critical interval for the model problem equals $[-3, -1/3]$. Hence, we choose $\sigma_+ = 1$ and $\sigma_- = 4$ and as right-hand side the function $f = 0.1\chi_{\{x_2<0.1\}} + \chi_{\{x_2>0.1\}}$. Since no exact solution is known, we compute a reference solution u_h on the fine mesh $\mathcal{T}_h$ using a standard FEM. Note that the fine mesh resolves the interface and is T-conform.

As in the previous section, we depict the convergence histories for the LOD error in the $H^1(\Omega)$-semi norm and the macroscopic LOD error in the $L^2(\Omega)$-norm in Figure 5.4. We again observe the expected linear convergence of the LOD solution in the $H^1(\Omega)$-semi norm. Moreover, the macroscopic LOD error follows the L^2-best approximation in the FE space, the best one can hope for. Note that for this experiment, the L^2-best approximation no longer converges quadratically. More precisely, both the macroscopic LOD error and the L^2-best approximation converge at an average approximate rate of 1.3 as we calculated by taking the average of the experimental orders of convergence. This corresponds very well to the analytically expected regularity of solutions due to the presence of corners at the interface: According to [7, 24], we expect $u \in H^{1+\lambda}(\Omega)$ with $\lambda \approx 0.37286$ for the square inclusion with our choice of σ_-.

When we compare again the different solutions (LOD solution, its macroscopic part, and the FE solution) to the reference solution in Figure 5.5, we see that the full LOD solution leads to clearly better results at the corners in comparison to its macroscopic part. This is a well-known feature of the LOD, observed also for classical elliptic diffusion problems with singularities at corners. Hence, even in a setting where the standard FEM is converging but at a reduced rate,

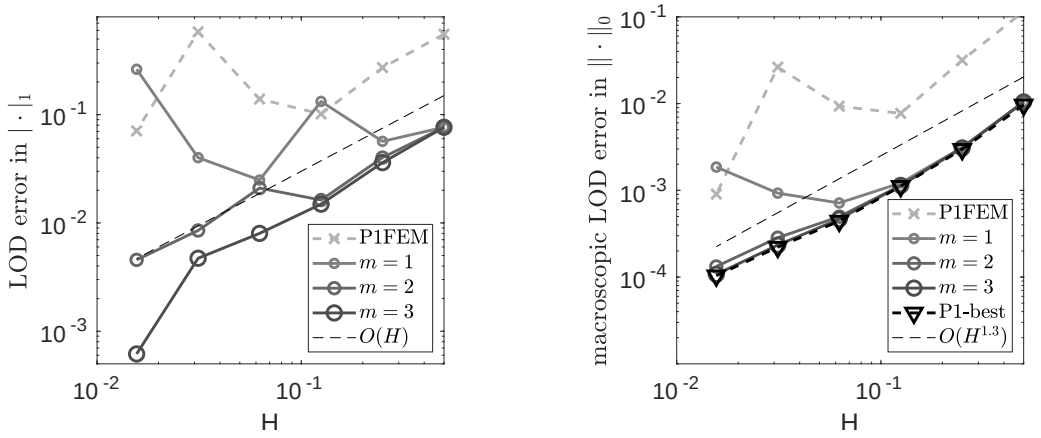

Figure 5.4: Convergence histories for the square inclusion in Section 5.2.

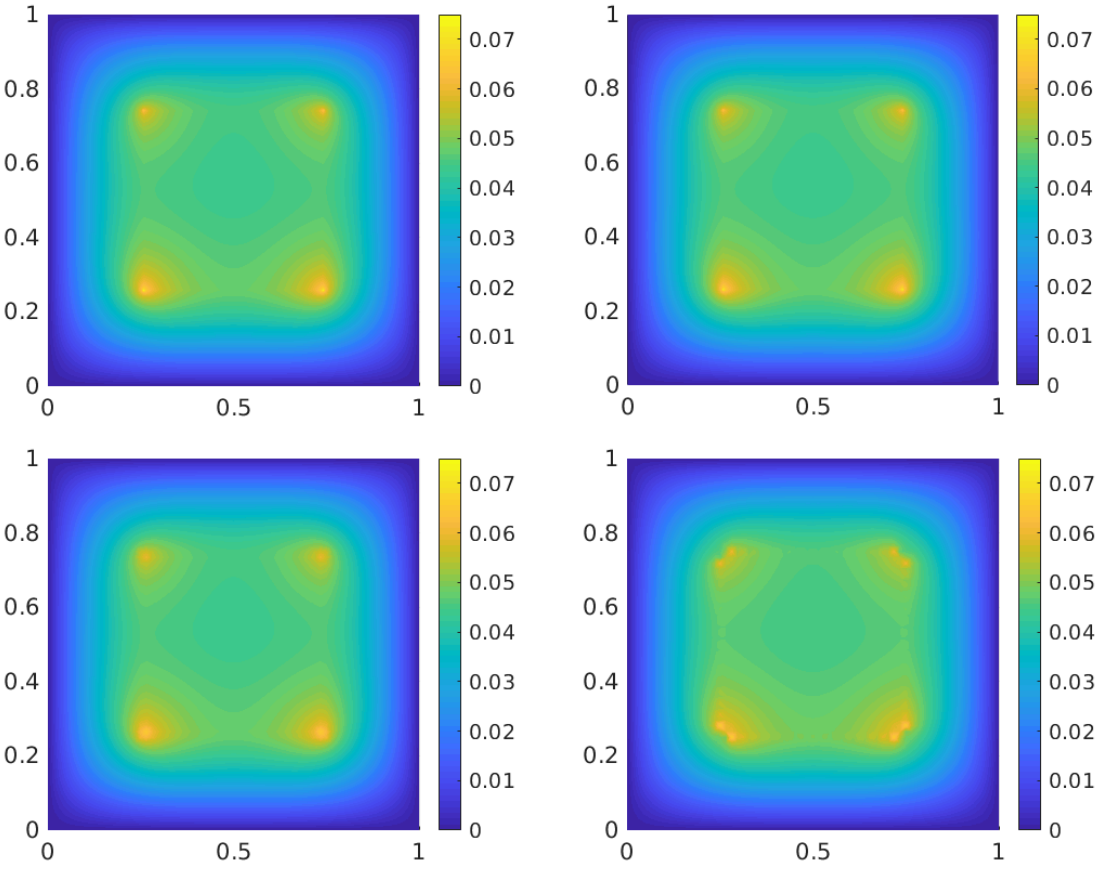

Figure 5.5: Various solutions for the square inclusion in Section 5.2: reference solution u_h (top left), LOD solution (top right), macroscopic part of the LOD solution ($u_{H,m}$, bottom left) and FE solution (bottom right).

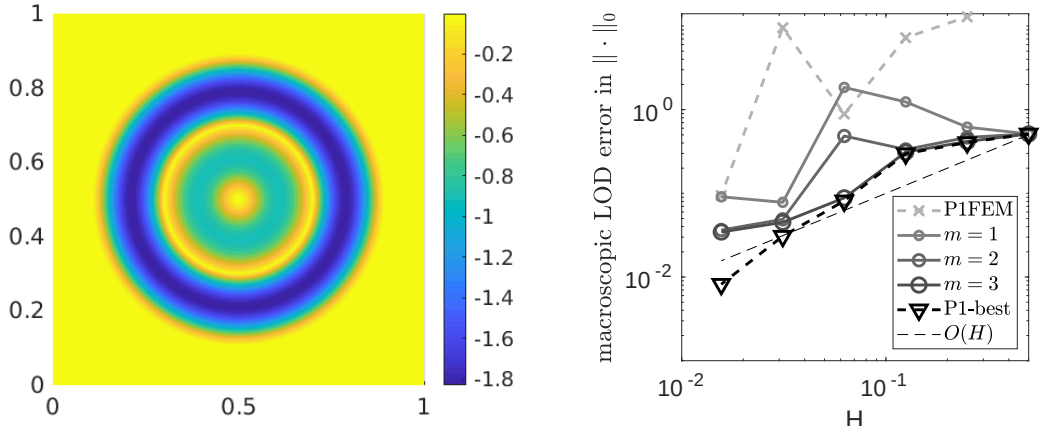

Figure 5.6: Exact solution (left) and convergence history for the macroscopic LOD error in the L^2-norm (right) for the circular inclusion in Section 5.3.

the full LOD might still be beneficial since the solution converges linearly in the $H^1(\Omega)$-semi norm independent of the regularity of the exact solution. We emphasize that the macroscopic part of the LOD solution qualitatively shows the correct behavior at the corners in contrast to the FE solution in Figure 5.5.

5.3. Circular inclusion with known exact solution

We consider $\Omega_- = B_{0.2}((0.5, 0.5))$, i.e., a circle with radius 0.2 around the point $(0.5, 0.5)$, and Ω_+ the complement. Since the boundary of Ω_- is smooth, the critical interval consists only of the value -1. Hence, we choose $\sigma_+ = 1$ and $\sigma_- = 2$ as in Section 5.1. We select a radially symmetric exact solution with homogeneous Dirichlet boundary conditions as follows. Let (r, φ) denote the standard polar coordinates and set $\tilde{r} = r - 0.5$. Then u is given by

$$u(\tilde{r}) = \begin{cases} A\tilde{r}^2(\tilde{r} - 0.2)(\tilde{r} - 0.4)^2, & \tilde{r} < 0.2, \\ -A\sigma_-\tilde{r}^2(\tilde{r} - 0.2)(\tilde{r} - 0.4)^2, & 0.2 < \tilde{r} < 0.4, \\ 0 & \text{else} \end{cases}$$

and f is calculated accordingly. The scalar factor A is used to scale the solution u to an $L^\infty(\Omega)$-norm of order 1, we pick here $A = 10000$. Note that the right-hand side f is piecewise smooth and does *not* possess a singularity at $(0.5, 0.5)$. The exact solution u is depicted in Figure 5.6, left.

The curved interface is never resolved, neither by the coarse meshes $\mathcal{T}_H$ nor by the fine reference mesh $\mathcal{T}_h$. In particular, the standard FEM solution on $\mathcal{T}_h$ may be not very reliable, which implies that the fine discretization in the LOD method might not be a faithful approximation either. In the present example, the absolute $L^2(\Omega)$-error between the exact solution u and the FEM solution on the fine grid $\mathcal{T}_h$ is of order 10^{-2}. Nevertheless, the convergence plot of the macroscopic LOD solution in the $L^2(\Omega)$-norm in Figure 5.6 shows rather promising results. At least for $m = 2, 3$, the macroscopic LOD error still follows the best approximation error – at least for coarse mesh sizes H. We observe a deviation from this desired best-approximation error for finer meshes because the discretization error on the underlying fine mesh $\mathcal{T}_h$ starts to dominate. Given these considerations and emphasizing once more that neither $\mathcal{T}_h$ nor the coarse meshes resolve the interface, the convergence results of Figure 5.6 are very satisfying.

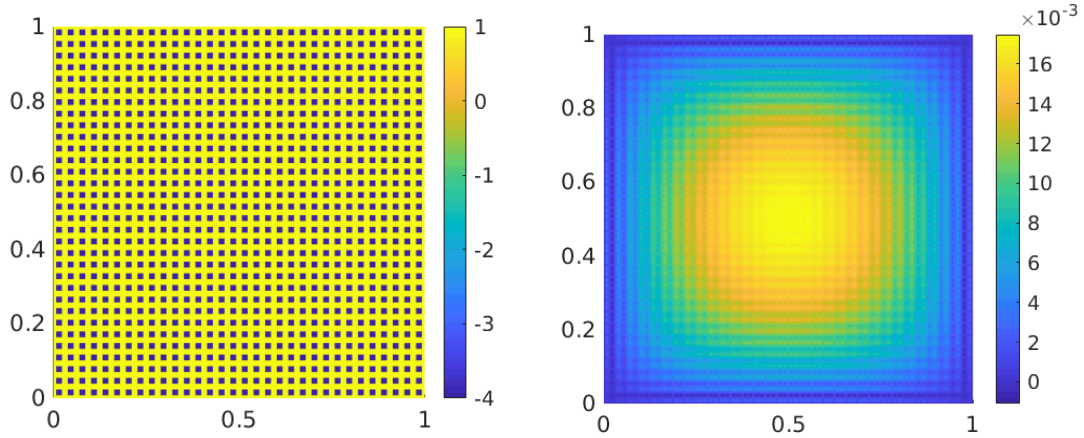

Figure 5.7: Coefficient (left) and fine FE solution (right) for the experiment in Section 5.4.

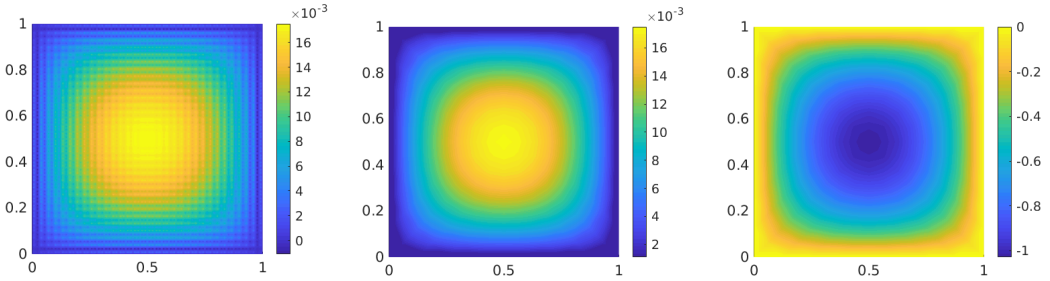

Figure 5.8: LOD solution (left), macroscopic part of LOD solution (middle), and FE solution (right) for $H = 2^{-4}$ and $m = 3$ in the experiment of Section 5.4.

5.4. Multiscale sign-changing coefficient

We consider a multiscale, sign-changing coefficient as depicted in Figure 5.7, left. It is periodic on a scale $\varepsilon = 2^{-5}$ and takes the values -4 (blue) and 1 (yellow). We set $f \equiv 1$ and compute a standard FE solution u_h on the mesh $\mathcal{T}_h$ as reference, see Figure 5.7, right. Note that $\mathcal{T}_h$ resolves all the jumps of the coefficient so that we can hope that u_h is a good approximation of the unknown exact solution u. In this example, we illustrate the homogenization feature of the LOD and its attractive performance even in the pre-asymptotic region, i.e., for meshes that do not resolve the discontinuities of the coefficient. For the coarse mesh $\mathcal{T}_H$ with $H = 2^{-4}$ and $m = 3$, we depict the LOD solution, its macroscopic part, and the FE solution in Figure 5.8. First of all, we observe that the standard FEM fails on this coarse mesh because the multiscale features of the coefficient are not resolved. This observation is already expected and well understood for the classical elliptic diffusion problem, see [26] for an excellent review. In contrast, the LOD produces faithful approximations. Its macroscopic part can be seen as a homogenized solution and already contains the main characteristic features of the solution. The full LOD solution also takes finescale features into account and thereby is even closer to the reference solution. This of course comes at the cost of higher computational complexity.

Conclusion

We presented and analyzed a generalized finite element method in the spirit of the Localized Orthogonal Decomposition for diffusion problems with sign-changing coefficients. Standard finite element basis functions are modified by including local corrections. The stability and the convergence of the method were analyzed under the assumption that the contrast is "sufficiently large". Our analysis involves a discrete T-coercivity argument, as well as "symmetrized" patches to compute the correctors associated with the elements close to the sign-changing interface. Numerical experiments illustrated the theoretically predicted optimal convergence rates. Furthermore, they showed the applicability of the method for general meshes, which do not resolve the interface, and highly heterogeneous coefficients.

The numerical experiments also outlined some possible future research questions. If the contrast is close to the critical interval, the patches for the corrector computations need to be rather large. This contrast-dependency might be reduced with the norm considered in [12], where we mention the connection with the LOD approach in weighted norms [16, 28]. Finally, it might be sufficient to only enrich finite element basis functions close to the interface.

References

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

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

[3] 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(6):1363–1387, 2012.

[4] A.-S. Bonnet-Ben Dhia, L. Chesnel, and P. Ciarlet Jr. T-coercivity for the Maxwell problem with sign-changing coefficients. *Comm. Partial Differential Equations*, 39(6):1007–1031, 2014.

[5] A.-S. Bonnet-Ben Dhia, L. Chesnel, and P. Ciarlet Jr. Two-dimensional Maxwell's equations with sign-changing coefficients. *Appl. Numer. Math.*, 79:29–41, 2014.

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

[7] A.-S. Bonnet-Ben Dhia, M. Dauge, and K. Ramdani. Analyse spectrale et singularités d'un problème de transmission non coercif. *C. R. Acad. Sci. Paris Sér. I Math.*, 328(8):717–720, 1999.

[8] C. Carvalho, L. Chesnel, and P. Ciarlet Jr. Eigenvalue problems with sign-changing coefficients. *C. R. Math. Acad. Sci. Paris*, 355(6):671–675, 2017.

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

[10] E. T. Chung and P. Ciarlet Jr. A staggered discontinuous Galerkin method for wave propagation in media with dielectrics and meta-materials. *J. Comput. Appl. Math.*, 239:189–207, 2013.

[11] P. G. Ciarlet. *The finite element method for elliptic problems*, volume 40 of *Classics in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2002.

[12] P. Ciarlet Jr. and M. Vohralík. Localization of global norms and robust a posteriori error control for transmission problems with sign-changing coefficients. *ESAIM Math. Model. Numer. Anal.*, 52(5):2037–2064, 2018.

[13] C. Engwer, P. Henning, A. Målqvist, and D. Peterseim. Efficient implementation of the localized orthogonal decomposition method. *Comput. Methods Appl. Mech. Engrg.*, 350:123–153, 2019.

[14] D. Gallistl and D. Peterseim. Stable multiscale Petrov-Galerkin finite element method for high frequency acoustic scattering. *Comput. Methods Appl. Mech. Engrg.*, 295:1–17, 2015.

[15] D. Gallistl and D. Peterseim. Computation of quasi-local effective diffusion tensors and connections to the mathematical theory of homogenization. *Multiscale Model. Simul.*, 15(4):1530–1552, 2017.

[16] F. Hellman and A. Målqvist. Contrast independent localization of multiscale problems. *Multiscale Model. Simul.*, 15(4):1325–1355, 2017.

[17] F. Hellman, A. Målqvist, and S. Wang. Numerical upscaling for heterogeneous materials in fractured domains. *arXiv pre-print*, 1908.03822, 2019.

[18] P. Henning and D. Peterseim. Oversampling for the multiscale finite element method. *Multiscale Model. Simul.*, 11(4):1149–1175, 2013.

[19] R. Kornhuber, D. Peterseim, and H. Yserentant. An analysis of a class of variational multiscale methods based on subspace decomposition. *Math. Comp.*, 87(314):2765–2774, 2018.

[20] R. Kornhuber and H. Yserentant. Numerical homogenization of elliptic multiscale problems by subspace decomposition. *Multiscale Model. Simul.*, 14(3):1017–1036, 2016.

[21] J. J. Lee and S. Rhebergen. A hybridized discontinuous Galerkin method for Poisson-type problems with sign-changing coefficients. *arXiv pre-print*, 1911.01984, 2019.

[22] R. Maier. *Computational multiscale methods in unstructured heterogeneous media*. PhD thesis, Universität Augsburg, 2020.

[23] A. Målqvist and D. Peterseim. Localization of elliptic multiscale problems. *Math. Comp.*, 83(290):2583–2603, 2014.

[24] S. Nicaise and J. Venel. A posteriori error estimates for a finite element approximation of transmission problems with sign changing coefficients. *J. Comput. Appl. Math.*, 235(14):4272–4282, 2011.

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

[26] D. Peterseim. Variational multiscale stabilization and the exponential decay of fine-scale correctors. In *Building bridges: connections and challenges in modern approaches to numerical partial differential equations*, volume 114 of *Lect. Notes Comput. Sci. Eng.*, pages 341–367. Springer, Cham, 2016.

[27] D. Peterseim. Eliminating the pollution effect in Helmholtz problems by local subscale correction. *Math. Comp.*, 86(305):1005–1036, 2017.

[28] D. Peterseim and R. Scheichl. Robust numerical upscaling of elliptic multiscale problems at high contrast. *Comput. Methods Appl. Math.*, 16(4):579–603, 2016.

[29] D. Peterseim, D. Varga, and B. Verfürth. From domain decomposition to homogenization theory. In *to appear in DD25 proceedings*, 2019.

[30] D. Peterseim and B. Verfürth. Computational high frequency scattering from high contrast media. *accepted for publication in Math. Comp.*, 2020.

[31] D. R. Smith, J. B. Pendry, and M. C. K. Wiltshire. Metamaterials and negative refractive index. *Science*, 305(5685):788–792, 2004.

A. Technical results used in Section 4

In this section, we prove a few technical results used in Section 4 combining standard scaling arguments for classical FE functions. Throughout the appendix, we use the notation introduced in Sections 2.2 and 4. Classical finite element scaling arguments use the mapping of elements in the mesh $\mathcal{T}_H$ onto the reference element. We use the standard notation of $\widehat{\cdot}$ for quantities (functions, constants, etc.) on the reference element. In particular, functions $\widehat{v}$ and v are connected with each other via the standard reference element mapping.

Lemma A.1. *Let $\mathbf{a} \in \mathcal{V}_H$. The estimate*

$$|m^{\mathbf{a}}(v)| \leq \widehat{C}_m \left(\frac{1}{\min_{K \in \mathcal{T}_H^{\mathbf{a}}} |K|} \right)^{1/2} \|v\|_{0,\omega^{\mathbf{a}}} \tag{A.1}$$

holds true for all $v \in H_0^1(\Omega)$, with

$$\widehat{C}_m := |\widehat{K}|^{1/2} \sup_{\substack{\widehat{v} \in \mathcal{P}_1(\widehat{K}) \\ \|\widehat{v}\|_{0,\widehat{K}}=1}} \|\widehat{v}\|_{0,\infty,\widehat{K}}.$$

Proof. Fix $\mathbf{a} \in \mathcal{V}_H$, $v \in H_0^1(\Omega)$, and recall the definition

$$m^{\mathbf{a}}(v) := \frac{1}{\sharp \mathbf{a}} \sum_{K \in \mathcal{T}_H^{\mathbf{a}}} (P_K v)(\mathbf{a}).$$

It is clear that

$$|m^{\mathbf{a}}(v)| \leq \frac{1}{\sharp \mathbf{a}} \sum_{K \in \mathcal{T}_H^{\mathbf{a}}} \|P_K v\|_{0,\infty,K} \leq \max_{K \in \mathcal{T}_H^{\mathbf{a}}} \|P_K v\|_{0,\infty,K} = \|P_{K_\star} v\|_{0,\infty,K_\star}$$

for some $K_\star \in \mathcal{T}_H^{\mathbf{a}}$. Then, since $w := P_{K_\star} v \in \mathcal{P}_1(K_\star)$, a standard scaling argument shows that

$$\|w\|_{0,\infty,K_\star}^2 = \|\widehat{w}\|_{0,\infty,\widehat{K}}^2 \leq \frac{\widehat{C}_m^2}{|\widehat{K}|} \|\widehat{w}\|_{0,\widehat{K}}^2 = \widehat{C}_m^2 \frac{1}{|K_\star|^2} \|w\|_{0,K_\star}^2,$$

from which (A.1) follows. □

Lemma A.2. *Let $\mathbf{a} \in \mathcal{V}_H$ and assume that $w \in H^1(\omega^{\mathbf{a}})$ satisfies $m^{\mathbf{a}}(w) = 0$. Then, it holds*

$$\|w\|_{0,\omega^{\mathbf{a}}} \leq \widehat{C}_{\mathrm{P}} H |w|_{1,\omega^{\mathbf{a}}}.$$

Proof. We assume that there exists a domain $\widehat{\omega}$ and a bilipschitz invertible mapping $\mathcal{F} : \widehat{\omega} \to \omega^{\mathbf{a}}$, whose restriction on each element $\widehat{K} := \mathcal{F}^{-1}(K)$ with $K \in \mathcal{T}_H^{\mathbf{a}}$ is affine. Since the mesh $\mathcal{T}_H$ is quasi uniform, we may further assume that there exists two constants $c_\star, c^\star$ such that $c_\star H \leq |D\mathcal{F}(\widehat{\mathbf{x}})| \leq c^\star H$ for all $\widehat{\mathbf{x}} \in \widehat{\omega}$. For the sake of simplicity, we also assume without loss of generality that $\mathcal{F}^{-1}(\mathbf{a}) = \mathbf{0}$.

If $v \in H^1(\omega^{\mathbf{a}})$, then $\widehat{v} := v \circ \mathcal{F}$ belongs to $H^1(\widehat{\omega})$, and we have $m(v) = \widehat{m}(\widehat{v})$, where

$$\widehat{m}(\widehat{v}) := \frac{1}{\sharp \mathbf{a}} \sum_{\widehat{K} \in \mathcal{F}^{-1}(\mathcal{T}_H^{\mathbf{a}})} (\mathcal{P}_{\widehat{K}} \widehat{v})(\mathbf{0}).$$

Now, we observe that for $\widehat{q} \in \mathcal{P}_0(\widehat{\omega})$, $\widehat{m}(\widehat{q}) = 0$ implies that $\widehat{q} = 0$. Then, a standard contradiction argument (see for instance [11, proof of Theorem 3.1.1]) shows that there exists a constant $\widehat{C}$ such that

$$\inf_{\widehat{q} \in \mathcal{P}_0(\widehat{\omega})} \|\widehat{v} - \widehat{q}\|_{1,\widehat{\omega}} \leq \widehat{C} \left(|\widehat{m}(\widehat{v})| + |\widehat{v}|_{1,\widehat{\omega}} \right),$$

and therefore,

$$\|\widehat{v}\|_{0,\widehat{\omega}} \leq \widehat{C} \widehat{C}_{\mathrm{P}}(\widehat{\omega}) |\widehat{v}|_{1,\widehat{\omega}}$$

for all $\widehat{v} \in H^1(\widehat{\omega})$ with $\widehat{m}(\widehat{v}) = 0$, where $\widehat{C}_{\mathrm{P}}(\widehat{\omega})$ is the Poincaré constant of $\widehat{\omega}$.

Then, the desired result is obtained by standard mapping arguments, with a constant $\widehat{C}_{\mathrm{P}}$ depending on $\widehat{C}$, $\widehat{C}_{\mathrm{P}}(\widehat{\omega})$, $c_\star$ and $c^\star$. □

The main aim of this appendix is to construct the function $\eta^{\mathbf{a}}$ used for the definition of T_H and study its scaling. For the ensuing construction to hold, we need to assume that $\mathcal{T}_H$ resolves the interface Γ. We emphasize, however, that no symmetry of the mesh is required. Moreover, we believe that a similar result holds if the interface does not cut the elements "too badly". We refer to [17] for a similar discussion in a different context.

Lemma A.3. *For all $\widehat{\lambda} \in L^2(\widehat{K})$, there exists a unique $\widehat{\eta} \in H^1_0(\widehat{K}) \cap \mathcal{P}_{d+2}(\widehat{K})$ such that*

$$(\widehat{\eta}, \widehat{v})_{\widehat{K}} = (\widehat{\lambda}, \widehat{v})_{\widehat{K}} \quad \forall \widehat{v} \in \mathcal{P}_1(\widehat{K}), \tag{A.2}$$

and we have

$$|\widehat{\eta}|_{1,\widehat{K}} \leq \widehat{C}_{\mathrm{dual}} \|\widehat{\lambda}\|_{0,\widehat{K}} \tag{A.3}$$

for some constant $\widehat{C}_{\mathrm{dual}}$ only depending on $\widehat{K}$. In addition, the equality

$$(\eta, v)_K = (\lambda, v)_K \quad \forall v \in \mathcal{P}_1(K) \tag{A.4}$$

and the estimate

$$|\eta|_{1,K} \leq \frac{\widehat{C}_{\mathrm{dual}}}{\rho_K} \|\lambda\|_{0,K} \tag{A.5}$$

hold true.

Proof. We introduce the "bubble function"

$$\widehat{b} := \prod_{\mathbf{a} \in \mathcal{V}(\widehat{K})} \widehat{\lambda}^{\mathbf{a}} \in \mathcal{P}_{d+1}(\widehat{K}),$$

defined using the barycentric coordinates $\widehat{\lambda}^{\mathbf{a}}$ of the reference element $\widehat{K}$. We observe that since $\widehat{b} \geq 1/2$ on an open set contained in $\widehat{K}$, the application $\widehat{v} \to \|b^{1/2}\widehat{v}\|_{0,\widehat{K}}$ is a norm on $\mathcal{P}_1(\widehat{K})$. Furthermore, since $\mathcal{P}_1(\widehat{K})$ is finite-dimensional, there exists a constant $\widehat{C}_{\mathrm{norm}}$ such that

$$\|\widehat{v}\|_{0,\widehat{K}}^2 \leq \widehat{C}_{\mathrm{norm}} \|\widehat{b}^{1/2}\widehat{v}\|_{0,\widehat{K}}^2 \tag{A.6}$$

for all $\widehat{v} \in \mathcal{P}_1(\widehat{K})$.

Now, let $\widehat{\lambda} \in L^2(\widehat{K})$. In virtue of the above discussion, there exists a unique $\widehat{w} \in \mathcal{P}_1(\widehat{K})$ such that

$$(\widehat{b}\widehat{w}, \widehat{v})_{\widehat{K}} = (\widehat{\lambda}, \widehat{v})_{\widehat{K}} \quad \forall \widehat{v} \in \mathcal{P}_1(\widehat{K}).$$

Then, one easily observes that $\widehat{\eta} := \widehat{b}\widehat{w} \in H_0^1(\widehat{K}) \cap \mathcal{P}_{d+2}(\widehat{K})$, satisfies (A.2). Furthermore, picking the test function $\widehat{v} = \widehat{w}$ in the definition of $\widehat{w}$ and employing (A.6), we have

$$\|\widehat{b}^{1/2}\widehat{w}\|_{0,\widehat{K}}^2 = (\widehat{\lambda}, \widehat{w})_{\widehat{K}} \leq \|\widehat{\lambda}\|_{0,\widehat{K}} \|\widehat{w}\|_{0,\widehat{K}} \leq \widehat{C}_{\mathrm{norm}} \|\widehat{\lambda}\|_{0,\widehat{K}} \|\widehat{b}^{1/2}\widehat{w}\|_{0,\widehat{K}}$$

and (A.3) follows since, recalling that $\widehat{b} \leq 1$, we have

$$\|\widehat{\eta}\|_{0,\widehat{K}} = \|\widehat{b}\widehat{w}\|_{0,\widehat{K}} \leq \|\widehat{b}^{1/2}\widehat{w}\|_{0,\widehat{K}} \leq \widehat{C}_{\mathrm{norm}} \|\widehat{\lambda}\|_{0,\widehat{K}},$$

and

$$|\widehat{\eta}|_{1,\widehat{K}} \leq C_{\mathrm{inv}} \|\widehat{\eta}\|_{0,\widehat{K}},$$

as $\widehat{\eta} \in \mathcal{P}_1(\widehat{K})$.

At this point (A.4) and (A.5) follows from usual scaling arguments. □

Lemma A.4. *For all $\mathbf{a} \in \mathcal{V}_H^+ \cup \mathcal{V}_H^0$, there exists $\eta^{\mathbf{a}} \in H_0^1(\Omega)$ with $\operatorname{supp} \eta^{\mathbf{a}} \subset \Omega_+$ such that*

$$m^{\mathbf{a}'}(\eta^{\mathbf{a}}) = \delta_{\mathbf{a}',\mathbf{a}} \qquad \forall \mathbf{a}' \in \mathcal{V}_H$$

and

$$|\eta^{\mathbf{a}}|_{1,\Omega_+} \leq \widehat{C}_{\mathrm{dual}} \max_{K \in \mathcal{T}_H^{\mathbf{a}}} \frac{|K|^{1/2}}{\rho_K},$$

where the constant $\widehat{C}_{\mathrm{dual}}$ only depends on the shape-regularity of the mesh.

Proof. Let $\mathbf{a} \in \mathcal{V}_H^+ \cup \mathcal{V}_H^0$ be arbitrary but fixed. There exists an element $K_\star \in \mathcal{T}_H$ such that $K_\star \subset \omega^{\mathbf{a}} \cap \Omega_+$. Following Lemma A.3 we consider a function $\eta^{\mathbf{a}} \in H_0^1(K_\star)$ such that $P_{K_\star}\eta^{\mathbf{a}} = \psi^{\mathbf{a}}|_{K_\star}$. Then, we obtain for any $\mathbf{a}' \in \mathcal{V}_H$ that

$$m^{\mathbf{a}'}(\eta^{\mathbf{a}}) = \frac{1}{\sharp \mathbf{a}'} \sum_{K \in \mathcal{T}_H^{\mathbf{a}'}} (P_K \eta^{\mathbf{a}})(\mathbf{a}') = \frac{1}{\sharp \mathbf{a}'} \psi^{\mathbf{a}}(\mathbf{a}') = \delta_{\mathbf{a},\mathbf{a}'}.$$

On the other hand, using (A.5), we have

$$|\eta^{\mathbf{a}}|_{1,\Omega_+} = |\eta^{\mathbf{a}}|_{1,K_\star} \leq \frac{\widehat{C}_{\mathrm{dual}}}{\rho_{K_\star}} \|\psi^{\mathbf{a}}\|_{0,K_\star} \leq \widehat{C}_{\mathrm{dual}} \frac{|K_\star|^{1/2}}{\rho_{K_\star}}.$$

□