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

Théophile Chaumont-Frelet, Barbara Verfürth

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 attracted a lot of interest over the last years due to many applications [27, 33]. The related mathematical problems are characterized by so-called sign-changing coefficients. At the simplest example of a diffusion problem in a domain $\Omega \subset \mathbb{R}^d$, $d \in \{2, 3\}$, 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., $\sigma \geq |\sigma_+| > 0$ in the complement Ω_+. The interface Γ between Ω_+ and Ω_- is then called the "sign-changing" interface. Such a behavior of the coefficient in the PDE does not only appear for metamaterials with negative effective properties [33], but also for electric permittivities, which can have a negative real part for certain metals.

The change of sign of σ 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 The standard assumption of coercive bilinear forms is no longer valid, so that existence and uniqueness of solutions have to be studied anew.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 [10]. 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,

[*] 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 Angewandte und Numerische Mathematik, Karlsruher Institut für Technologie (KIT), Englerstr. 2, D-76131 Karlsruhe

namely a) if the mesh satisfies certain symmetry properties around the interface Γ, 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 $\widetilde{r} > r$ [11, 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 [14, 26]. 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 [12] and [23].

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) [20, 25, 28]. The LOD is especially targeted at so-called multiscale problems, where the coefficient is subject to rapid spatial variations. Standard discretization schemes need to resolve all these features with their computational grid leading to an enormous and often infeasible computational effort. The basic idea of the LOD is to construct a low-dimensional solution space with very good H^1-approximation properties with respect to the exact solution. As standard finite element functions on a coarse mesh alone do not yield a faithful approximation space, problem-dependent multiscale functions are added. The latter are defined as solutions of local fine-scale problems. Since its introduction in [20, 25], the LOD has been successfully applied in various situations, where we mention in particular the reduction of the pollution effect for high-frequency Helmholtz problems [16, 29, 32]. The efficient implementation of the method is outlined in [15]. Note that the LOD is closely connected to domain decomposition methods [21, 22, 31]. Further, it can also be interpreted in the context of homogenization [17]. If σ is (locally) periodic one can thereby recover traditional (analytical) homogenization results, see [8, 9] for such results in the case of periodic sign-changing coefficients.

We analyze the stability and convergence of the proposed method when $d = 2$ or 3, 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 [11], 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 in [24, 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. Although they are limited to two-dimensional settings, 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. Our generalized finite element method is motivated in an ideal form in Section 3. There, we also discuss three main challenges of the ideal method, namely the well-posedness of the construction in the case of sign-changing coefficients, as well as the localization and discretization of the multiscale basis. The dedicated arguments required to take into account T-coercivity in the context of LOD are discussed in Section 4. The fully practical LOD is finally presented and analyzed in Section 5. In Section 6, 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. Settings

In this section we introduce the required functional spaces, the model problem under consideration and discuss the notion of T-coercivity.

2.1. Domain and coefficient

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

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

the boundary shared by the two subsets. For the sake of simplicity, we assume that Γ is polytopal. However, we do not require specific assumption about the topology of $\Omega_\pm$. In particular, Ω_+ and/or Ω_- can be multi-connected.

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. Note that by symmetry, one could alternatively consider $\sigma|_{\Omega_-} \geq \sigma|_{\Omega_+}$, but we only analyze the other case for the sake of simplicity. In the remaining of this work, we will call the positive real number $\mathscr{C} := \sigma_-/\sigma_+$ the "contrast".

2.2. Functional spaces

Throughout this work, if $D \subset \Omega$, $L^2(D)$ is the usual Lebesgue space of square integrable functions. We denote by $(\cdot,\cdot)_D$ and $\|\cdot\|_{0,D}$ the usual inner product and norm of $L^2(D)$. We employ the same notations for the inner product and norm of $\left(L^2(D)\right)^d$. Classically, $H^1(D) := \left\{v \in L^2(D) \mid \boldsymbol{\nabla} v \in \left(L^2(D)\right)^d\right\}$ denotes the usual Sobolev space, and if $\gamma \subset \partial D$, we employ the notation

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

that we equip with its usual semi-norm $|\cdot|_{1,D}$. Note that, thanks to Poincaré inequality, $|\cdot|_{1,D}$ is actually a norm on $H^1_\gamma(D)$ as long as γ has a strictly positive surface measure. Unless stated otherwise, we will always employ the notation $H^1_\gamma(D)$ when the surface measure of γ is strictly positive, and equip the space with $|\cdot|_{1,D}$ as norm. In particular, this norm is considered when defining the norm of linear operators. We will also use the usual notation $H^1_0(\Omega) := H^1_{\partial\Omega}(\Omega)$, and consider the equivalent norm

$$|v|^2_{1,\sigma,\Omega} := \int_\Omega |\sigma||\boldsymbol{\nabla} v|^2$$

for $v \in H^1_0(\Omega)$.

2.3. Model problem

Given $f \in L^2(\Omega)$, we seek $u \in H^1_0(\Omega)$ such that

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

where

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

When σ is positive, the bilinear form $a(\cdot,\cdot)$ actually corresponds to the inner product associated with the norm $|\cdot|_{1,\sigma,\Omega}$. In particular, $a(\cdot,\cdot)$ is coercive, and the well-posedness of (2.1) follows from Lax-Milgram Lemma. Here, the sign-change of σ prevents the coercivity of $a(\cdot,\cdot)$. Following an approach known as T-coercivity, we will show that instead, assuming that the contrast is sufficiently large, $a(\cdot,\cdot)$ satisfies an inf-sup condition, ensuring the well-posedness of (2.1).

Remark 2.1. Throughout the whole work, we assume $f \in L^2(\Omega)$. While the model problem and the generalized finite element method can be defined for $f \in H^{-1}(\Omega)$ as well, $f \in L^2(\Omega)$ is required to obtain convergence of the method, see Proposition 3.2.

2.4. Symmetrization and T-coercivity

In the broadest sense, we say that the bilinear form $a(\cdot,\cdot)$ is Y-coercive if there exists an operator $\mathrm{Y} \in \mathcal{L}(H^1_0(\Omega))$ such that

$$a(v, \mathrm{Y}v) \geq a_\star |v|^2_{1,\sigma,\Omega} \quad \forall v \in H^1_0(\Omega),$$

for some fixed $a_\star > 0$. This definition is equivalent to the usual "inf-sup" condition. It guarantees the well-posedness of Problem (2.1), and the stability constant then depends on $a_\star$ and $\|\mathrm{Y}\|_{\mathcal{L}(H^1_0(\Omega))}$.

In the context of problems with a sign-changing coefficient, a particular form of T-coercivity based on symmetrization has been developed [3, 2]. The key idea is to design an operator Y that "flips" the sign of its argument in Ω_-. This construction relies on a symmetrization operator S, that maps $H^1_{\partial\Omega}(\Omega_-)$ into $H^1_{\partial\Omega}(\Omega_+)$.

In the following, we assume that Ω_- and Ω_+ are such that a "symmetrization" operator S is available. By symmetrization operator, we mean that that $\mathrm{S} \in \mathcal{L}(H^1_{\partial\Omega}(\Omega_-); H^1_{\partial\Omega}(\Omega_+)) \cap \mathcal{L}(L^2(\Omega_-); L^2(\Omega_+))$, is a linear mapping that preserves the trace on Γ, i.e., $\mathrm{S}v|_\Gamma = v|_\Gamma$ for all $v \in H^1_{\partial\Omega}(\Omega_-)$. We refer the reader to [3] for examples of such symmetrization operators. Further, we point out that rather general polypotal interfaces can be treated [2] if one uses the concept of so-called *weak* T-*coercivity*. We discuss the extension of our work to this setting in Section 3.4.

Using S, we may now define an operator $\mathrm{T} \in \mathcal{L}(H^1_0(\Omega))$ for which $a(\cdot,\cdot)$ is T-coercive. Specifically, if $u \in H^1_0(\Omega)$, we set

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

One easily sees that we have

$$|v - \mathrm{T}v|_{1,\Omega_+} \leq C_\pm(\mathrm{T})|v|_{1,\Omega_-} \text{ and } \|v - \mathrm{T}v\|_{0,\Omega_+} \leq C^0_\pm(\mathrm{T})\|v\|_{0,\Omega_-} \tag{2.2}$$

for all $v \in H^1_0(\Omega)$, with

$$C_\pm(T) := 2\|\mathrm{S}\|_{\mathcal{L}(H^1_{\partial\Omega}(\Omega_-);H^1_{\partial\Omega}(\Omega_+))}, \qquad C^0_\pm(T) := 2\|\mathrm{S}\|_{\mathcal{L}(L^2(\Omega_-);L^2(\Omega_+))}.$$

We can readily employ T to establish an inf-sup condition for $a(\cdot,\cdot)$ on $H^1_0(\Omega)$, thus showing that Problem (2.1) is well-posed [3]. However, later in the analysis of the LOD method, we will need to show a similar inf-sup condition, but on the kernel of some quasi-interpolation operator, instead of the whole $H^1_0(\Omega)$ space. For this reason, we first give a general result linking the existence of an operator $\mathrm{Y} \in \mathcal{L}(V)$ for some $V \subset H^1_0(\Omega)$ and the inf-sup stability of $a(\cdot,\cdot)$ on V.

Theorem 2.2. *Let $V \subset H^1_0(\Omega)$ be a closed subspace. Assume that there exists an operator $\mathrm{Y} \in \mathcal{L}(V)$ such that*

$$(\mathrm{Y}v)|_{\Omega_-} = -v|_{\Omega_-} \tag{2.3a}$$

and

$$|v - \mathrm{Y}v|_{1,\Omega_+} \leq C_\pm(\mathrm{Y})|v|_{1,\Omega_-}, \tag{2.3b}$$

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

$$a(v, \mathrm{Y}v) \geq \left(1 - \frac{C_\pm(\mathrm{Y})}{2}\left(\frac{\sup_{\Omega_+}\sigma}{\inf_{\Omega_+}\sigma}\right)\sqrt{\frac{\sigma_+}{\sigma_-}}\right)|v|^2_{1,\sigma,\Omega} \tag{2.4}$$

for all $v \in V$.

Proof. Pick an arbitrary element $v \in V$. Taking advantage of (2.3a), we may write

$$\begin{aligned} a(v, \mathrm{Y}v) &= (\sigma \boldsymbol{\nabla} v, \boldsymbol{\nabla}(\mathrm{Y}v))_{\Omega_+} + (\sigma \boldsymbol{\nabla} v, \boldsymbol{\nabla}(\mathrm{Y}v))_{\Omega_-} \\ &= (\sigma \boldsymbol{\nabla} v, \boldsymbol{\nabla}(\mathrm{Y}v))_{\Omega_+} - (\sigma \boldsymbol{\nabla} v, \boldsymbol{\nabla} v)_{\Omega_-} \\ &= (|\sigma| \boldsymbol{\nabla} v, \boldsymbol{\nabla}(\mathrm{Y}v))_{\Omega_+} + (|\sigma| \boldsymbol{\nabla} v, \boldsymbol{\nabla} v)_{\Omega_-} \\ &= |v|^2_{1,\sigma,\Omega} - (|\sigma| \boldsymbol{\nabla} v, \boldsymbol{\nabla}(v - \mathrm{Y}v))_{\Omega_+} \end{aligned} \tag{2.5}$$

Then, we derive that

$$\begin{aligned} |(\sigma \boldsymbol{\nabla} v, \boldsymbol{\nabla}(v - \mathrm{Y}v))_{\Omega_+}| &\leq (\sup_{\Omega_+}\sigma)|v|_{1,\Omega_+}|v - \mathrm{Y}v|_{1,\Omega_+} \\ &\leq \left(\frac{\sup_{\Omega_+}\sigma}{\inf_{\Omega_+}\sigma}\right)\sigma_+|v|_{1,\Omega_+}|v - \mathrm{Y}v|_{1,\Omega_+} \\ &\leq C_\pm(\mathrm{Y})\left(\frac{\sup_{\Omega_+}\sigma}{\inf_{\Omega_+}\sigma}\right)\sigma_+|v|_{1,\Omega_+}|v|_{1,\Omega_-} \\ &\leq \frac{C_\pm(\mathrm{Y})}{2}\left(\frac{\sup_{\Omega_+}\sigma}{\inf_{\Omega_+}\sigma}\right)\sqrt{\frac{\sigma_+}{\sigma_-}}|v|^2_{1,\sigma,\Omega}, \end{aligned} \tag{2.6}$$

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 (2.4) then follows from (2.5) and (2.6). □

Recalling (2.2), an immediate consequence of Theorem 2.2 is that if the contrast is sufficiently large, $a(\cdot,\cdot)$ is T-coercive, and Problem (2.1) is well-posed.

Corollary 2.3. *Under the assumption that*

$$\sqrt{\mathscr{C}} > \left(\frac{\sup_{\Omega_+}\sigma}{\inf_{\Omega_+}\sigma}\right)\|\mathrm{S}\|_{\mathcal{L}(H^1_{\partial\Omega}(\Omega_-);H^1_{\partial\Omega}(\Omega_+))} \tag{2.7}$$

we have $a(v, \mathrm{T}v) \geq \alpha|v|^2_{1,\sigma,\Omega}$ for all $v \in H^1_0(\Omega)$, with

$$\alpha := 1 - \frac{C_\pm(\mathrm{T})}{2}\left(\frac{\sup_{\Omega_+}\sigma}{\inf_{\Omega_+}\sigma}\right)\sqrt{\frac{\sigma_+}{\sigma_-}} > 0.$$

In particular, Problem (2.1) is well-posed.

Remark 2.4. 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". Similarly, one could choose to define a symmetrization operator S mapping from Ω_+ to Ω_-. We only consider one direction for the sake of simplicity. We refer the reader to [3] for a detailed discussion.

3. An ideal generalized finite element method

In this section, we are concerned with the discretization of our model problem (2.1). We first introduce some finite element notation in Section 3.1. The generalized finite element method is built upon a quasi-interpolation operator, which we briefly introduce in Section 3.2, and then present the idea of the generalized finite element method in Section 3.3. Finally, in Section 3.4, we discuss the extension of the method to problems satisfying a "weak" T-coercivity condition.

3.1. Preliminaries and notation

We consider a shape-regular quasi-uniform triangulation $\mathcal{T}_H$ of Ω. $\mathcal{T}_H$ is supposed to be a coarse mesh in the sense that it does not necessarily resolve the variations and oscillations of σ nor the sign-changing interface Γ.

Remark 3.1. The definition of the method does not require the triangulation $\mathcal{T}_H$ to fit the sign-changing interface Γ, and numerical experiments seem to indicate that the method is efficient without this requirements. In our theoretical analysis however, we require that Γ (but not the oscillation of σ) is resolved by $\mathcal{T}_H$ as a technical assumption.

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:

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

For $K \in \mathcal{T}_H$, let $\mathcal{V}(K) \subset \mathcal{V}_H$ denote the set of vertices of K. 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 $\mathbf{a}$.

The standard conforming finite element space of lowest order Lagrange elements is denoted by $V_H \subset H_0^1(\Omega)$, i.e.,

$$V_H = \{v \in H_0^1(\Omega) \quad | \quad v|_K \in \mathcal{P}_1(K)\ \forall K \in \mathcal{T}_H\},$$

where $\mathcal{P}_1$ denotes the polynomials of total degree at most one. Note that a standard finite element discretization of (2.1) using V_H will fail to produce faithful approximations if $\mathcal{T}_H$ does not resolve the variations of σ.

For any $D \subset \Omega$, $\mathrm{N}(D)$ denotes the element patch around D defined as

$$\mathrm{N}(D) = \bigcup \{K \in \mathcal{T}_H, \overline{K} \cap \overline{D} \neq \emptyset\}.$$

For later use, we also define inductively the m-layer patch $\mathrm{N}^m(D)$ around D via $\mathrm{N}^m(D) = \mathrm{N}(\mathrm{N}^{m-1}(D))$ for $m \geq 2$. The idea of patches is illustrated in Figure 5.1 in Section 5.1 below.

3.1.1. Reference element and associated constants

In the remaining of this work, $\widehat{K} \subset \mathbb{R}^d$ is an arbitary but fixed "reference" simplex with diameter $\widehat{H} = 1$. We denote by $\widehat{b} \in \mathcal{P}_{d+1}(\widehat{K})$ the "bubble" function of $\widehat{K}$ that we define as the product of its $(d+1)$ barycentric coordinate functions. Since $0 \leq \widehat{b} \leq 1$ on $\widehat{K}$, $\|\cdot\|_{0,\widehat{K}}$ and $\|\widehat{b}^{1/2}\cdot\|_{0,\widehat{K}}$ are equivalent norms on $\mathcal{P}_1(\widehat{K})$ and we denote by

$$\widehat{C}_{norm} := \sup_{q \in \mathcal{P}_1(\widehat{K})\setminus\{0\}} \frac{\|q\|_{0,\widehat{K}}}{\|\widehat{b}^{1/2} q\|_{0,\widehat{K}}}, \tag{3.1}$$

the upper constant (note that since $\widehat{b} \leq 1$, the constant is 1 in the other direction). The constants

$$\widehat{C}_{inf} := |\widehat{K}|^{1/2} \sup_{q \in \mathcal{P}_1(\widehat{K})\setminus\{0\}} \frac{\|q\|_{0,\infty,\widehat{K}}}{\|q\|_{0,\widehat{K}}} \qquad \widehat{C}_{inv} := \sup_{q \in \mathcal{P}_1(\widehat{K})\setminus\{0\}} \frac{|q|_{1,\widehat{K}}}{\|q\|_{0,\widehat{K}}} \tag{3.2}$$

will also be useful in the sequel.

3.1.2. Reference patches and associated constants

We assume that there exists a finite set of "reference patches" $\widehat{\mathcal{R}}$ such that for all $\mathbf{a} \in \mathcal{V}_H$, there exist $\widehat{\mathcal{T}} \in \widehat{\mathcal{R}}$ and a bilipschitz invertible mapping $\mathcal{F} : \widehat{\omega} \to \omega^{\mathbf{a}}$, $\widehat{\omega}$ being the domain associated with $\widehat{\mathcal{T}}$, such that for each $\widehat{K} \in \widehat{\mathcal{T}}$ the restriction of $\mathcal{F}$ to $\widehat{K}$ is affine, and $\mathcal{F}(\widehat{K}) = K$ for some $K \in \mathcal{T}_H^{\mathbf{a}}$. 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 $\mathcal{F}^{-1}(\mathbf{a}) = \mathbf{0}$.

We assume that all elements $\widehat{K}$ in the reference patches satisfy $H_{\widehat{K}} \leq 1$ and $\rho_{\widehat{K}} \geq \widehat{\rho}$.

We employ the notation

$$\widehat{C}_{\mathrm{P}} := \max_{\mathcal{T} \in \widehat{\mathcal{R}}} \sup_{\widehat{v} \in H^1(\widehat{\omega})} \frac{\|\widehat{v}\|_{0,\widehat{\omega}} + |\widehat{v}|_{1,\widehat{\omega}}}{|\ell(\widehat{v})| + |\widehat{v}|_{1,\widehat{\omega}}}, \tag{3.3}$$

where

$$\ell(\widehat{v}) := \frac{1}{\sharp\widehat{\mathcal{T}}} \sum_{\widehat{K} \in \widehat{\mathcal{T}}} (\mathcal{P}_{\widehat{K}} \widehat{v})(\mathbf{0}).$$

As we detail in Appendix A, the above definition makes sense and we have $\widehat{C}_{\mathrm{P}} < +\infty$.

We will also need the constants

$$\widehat{C}_{\mathrm{u}} := \max_{\widehat{\mathcal{T}} \in \widehat{\mathcal{R}}} \left(\frac{\max_{K \in \widehat{\mathcal{T}}} |K|}{\min_{K \in \widehat{\mathcal{T}}} |K|} \right)$$

as well as

$$C_{\mathrm{u}}^{\mathbf{a}} := \frac{\max_{K \in \mathcal{T}_H^{\mathbf{a}}} |K|}{\min_{K \in \mathcal{T}_H^{\mathbf{a}}} |K|} \qquad C_{\mathrm{u}} := \max_{\mathbf{a} \in \mathcal{V}_H} C_{\mathrm{u}}^{\mathbf{a}}.$$

3.2. The quasi-interpolation operator

The LOD method hinges on a stable quasi-interpolation operator $I_H : H_0^1(\Omega) \to V_H$. Here, following [28], 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

$$I_H v := \sum_{\mathbf{a} \in \mathcal{V}_H^{\mathrm{int}}} m^{\mathbf{a}}(v) \psi^{\mathbf{a}}, \tag{3.4}$$

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 $\mathcal{P}_1(K)$.

I_H is a projection onto V_H ($I_H \circ I_H = I_H$) and we furthermore have

$$\|v - I_H v\|_{0,K} + H \|\boldsymbol{\nabla}(v - I_H v)\|_{0,K} \leq C_I H \|\boldsymbol{\nabla} v\|_{0,\mathrm{N}(K)} \qquad \forall K \in \mathcal{T}_H \tag{3.5}$$

for all $v \in H_0^1(\Omega)$, see [28, Sec. 4, eq. (16)] and the references therein. 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 [15] for the required properties.

3.3. Motivation and presentation of the ideal method

The aim of this section is to construct a generalized finite element method in order to approximate the solution u of (2.1) on the coarse mesh $\mathcal{T}_H$ even if σ is a multiscale coefficient and the standard finite element method on $\mathcal{T}_H$ therefore fails to produce a faithful approximation. The idea is to construct a generalized finite element space $\widetilde{V}_H$ of the same dimension as V_H, but with better approximation properties. We will explain and introduce this idea in detail following the lines of thought for an elliptic diffusion problem (see, e.g., [25, 24, 28] for more details) and discuss the occurring challenges.

We note first that the projection property of I_H implies the decomposition of $H_0^1(\Omega)$ into the finite element space V_H and the finescale space $W := \ker(I_H)$, i.e. $H_0^1(\Omega) = V_H \oplus W$. We stress that W represents the space of functions with potential finescale oscillations and is infinite-dimensional. Since I_H is a stable quasi-interpolation onto V_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, however, that $I_H u$ is typically not found by the finite element method.

A simple calculation shows that it holds for any $v \in H_0^1(\Omega)$ that

$$a(I_H u, v) = (f, v)_\Omega - a((\mathrm{id} - I_H) u, v).$$

The last term on the right-hand side vanishes if we restrict the test functions v to the space

$$\widetilde{V}_H := \{v \in H_0^1(\Omega) \quad | \quad a(w, v) = 0 \ \ \forall w \in W\}.$$

This means that $I_H u$ can be characterized as a Petrov-Galerkin solution with ansatz space V_H and test space $\widetilde{V}_H$. This ideal (test) space comes with the decomposition $H_0^1(\Omega) = \widetilde{V}_H \oplus W$ which additionally is orthogonal with respect to $a(\cdot, \cdot)$. We now provide an alternative characterization of $\widetilde{V}_H$ that in particular will show that $\dim V_H = \dim \widetilde{V}_H$. For this, we introduce a so-called correction operator $\mathcal{Q} : H_0^1(\Omega) \to W$ by

$$a(w, \mathcal{Q} v) = a(w, v) \qquad \text{for all} \quad w \in W. \tag{3.6}$$

As a direct consequence, we obtain $a(w, (\mathrm{id} - \mathcal{Q})v) = 0$ for all $w \in W$ and hence, the characterization of $\widetilde{V}_H$ from above. This implies

$$\widetilde{V}_H = (\mathrm{id} - \mathcal{Q})H^1_0(\Omega) = (\mathrm{id} - \mathcal{Q})\big(I_H(H^1_0(\Omega)) + (\mathrm{id} - I_H)H^1_0(\Omega)\big) = (\mathrm{id} - \mathcal{Q})V_H$$

because $I_H(H^1_0(\Omega)) = V_H$, $(\mathrm{id} - I_H)H^1_0(\Omega) = W$ and $(\mathrm{id} - \mathcal{Q})W = \{0\}$.

To sum up, we have $\widetilde{V}_H = (\mathrm{id} - \mathcal{Q})V_H$ and, hence, the desired property $\dim \widetilde{V}_H = \dim V_H$. We will use the space $\widetilde{V}_H$ not only as test space, but also as ansatz space in our generalized finite element (Galerkin) method. This means that we seek $u_H \in V_H$ such that

$$a((\mathrm{id} - \mathcal{Q})u_H, (\mathrm{id} - \mathcal{Q})v_H) = (f, (\mathrm{id} - \mathcal{Q})v_H)_\Omega \qquad \text{for all} \quad v_H \in V_H. \tag{3.7}$$

A direct consequence of this construction is that $I_H(\mathrm{id} - \mathcal{Q})u_H = u_H = I_H u$.

Before providing an a priori error estimate for this (ideal) generalized finite element method, let us discuss the challenges and open problems with the approach presented so far. These challenges will be addressed in the ensuing sections.

1. **Well-posedness of the corrector problems** (3.6)**:** Since $a(\cdot, \cdot)$ is not coercive and only satisfies an inf-sup condition over $H^1_0(\Omega)$, we need to show such an inf-sup condition over the space W as well (note that in contrast, coercivity is automatically inherited on W for coercive problem). More precisely, we will construct in Section 4 below an operator $\mathrm{T}_H \in \mathcal{L}(W)$ and show that there is $\alpha_\kappa > 0$ (independent of H) such that for a sufficiently large contrast, we have

$$a(w, \mathrm{T}_H w) \geq \alpha_\kappa |w|^2_{1,\sigma,\Omega} \qquad \text{for all} \quad w \in W. \tag{3.8}$$

2. **Non-locality of the correctors:** The corrector problems (3.6) are global finescale problems and therefore as expensive to solve as the original problem on a fine mesh. In Section 5.1, we will show how to localize the computation of the correctors to patches of elements. This localization step is motivated by a decay of the correctors which is exponential in units of H. Due to the T-coercivity of our problem, (technical) modifications in the construction of the patches for the localized correctors need to be introduced in comparison to standard elliptic problems.

3. **Infinite-dimensionality of the fine-scale space W:** Although the corrector problems are localized in Section 5.1 as discussed above, 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 and discretize the corrector problems using this mesh. This final step towards a practical method is discussed in Section 5.2.

Because of the second and third challenge we call the generalized finite element method in this section "ideal". We close its presentation with illustrating its good approximation properties, which will be preserved even through the localization and discretization of the corrector problems.

Proposition 3.2. *Assume that the corrector problems* (3.6) *are well-posed, i.e.,* (3.8) *holds. Then, we have*

$$\inf_{v_H \in V_H \setminus \{0\}} \sup_{\psi_H \in V_H \setminus \{0\}} \frac{a((\mathrm{id} - \mathcal{Q})v_H, (\mathrm{id} - \mathcal{Q})\psi_H)}{|v_H|_{1,\sigma,\Omega}\, |\psi_H|_{1,\sigma,\Omega}} \geq \tilde{\alpha}. \tag{3.9}$$

where $\widetilde{\alpha} = \alpha C_I^{-2} \, \|\mathrm{T}\|_{\mathcal{L}(H^1_0(\Omega))}$ *and* C_I *is the interpolation constant from* (3.5). *Moreover, the unique solution* u_H *of* (3.7) *fulfills the following error estimate*

$$|u - (\mathrm{id} - \mathcal{Q})u_H|_{1,\sigma,\Omega} \leq \alpha_\kappa^{-1} C_I \, \|\mathrm{T}_H\|_{\mathcal{L}(W)} \, H \|f\|_{0,\Omega}.$$

Note that the inf-sup condition automatically implies the well-posedness of (3.7). Further, we stress that $\|\mathrm{T}_H\|_{\mathcal{L}(W)}$ is independent of H. The linear convergence of the error in Proposition 3.2 is optimal for lowest-order elements and moreover, this result is independent of the regularity of the exact solution (which may be arbitrarily low, since $\sigma \in L^\infty(\Omega)$). Proposition 3.2 is classical for the LOD applied to inf-sup stable problems and we refer to [24, Chapter 2] for a proof. We emphasize that the assumption $f \in L^2(\Omega)$ is essential to obtain the linear rate, cf. [28] for a general discussion.

3.4. 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 (2.4), $a(v, \mathrm{T}v)$ only satisfies a Gårding-type inequality [6], namely

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

where $\alpha, \mu > 0$ are positive constants. As mentioned, this concept allows to treat rather general sign-changing problems with polytopal interfaces, see [3, 2]. We stress that in the case of the Helmholtz equation, one also considers a sesquilinear form satisfying a similar Gårding inequality (without an operator T, though).

Assuming in addition that the solution u to (2.1) is unique, i.e., (2.1) is well-posed, 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, the Helmholtz equation (with positive coefficients), was analyzed in [16, 29, 32]. 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 (3.5) 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.

4. T-coercivity in the kernel of I_H

As described above, the LOD method relies on "corrector" problems set in the kernel W of I_H. The purpose of this section is to show that the bilinear form $a(\cdot,\cdot)$ is inf-sup stable over W. To do so, we build a discrete counterpart T_H of the the operator T that maps the kernel W into itself.

4.1. Preliminary results

We start by recording two preliminary results in Lemma 4.1 and Lemma 4.2. The first is concerned with the scaling of the weights $m^{\mathbf{a}}$ appearing in the definition of I_H, while the second is a Poincaré-type inequality for functions in W. As the proofs are rather technical, we postpone them to Appendix A.1 to ease the reading.

Lemma 4.1. *Let $\mathbf{a} \in \mathcal{V}_H$. The estimate*

$$|m^{\mathbf{a}}(v)| \leq \widehat{C}_{inf} \left(\frac{1}{\min_{K \in \mathcal{T}_H^{\mathbf{a}}} |K|} \right)^{1/2} \|v\|_{0,\omega^{\mathbf{a}}} \tag{4.1}$$

holds true for all $v \in H^1_0(\Omega)$.

Lemma 4.2. *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*

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

where

$$C_{\mathrm{P}} := \frac{\sqrt{\widehat{C}_{\mathrm{u}} C_{\mathrm{u}}} \widehat{C}_{\mathrm{P}}}{\widehat{\rho}} |w|_{1,\omega^{\mathbf{a}}}.$$

4.2. A discrete operator T_H

A key ingredient in the construction of the operator T_H is the introduction of a "dual weight function" $\eta^{\mathbf{a}}$ associated with each vertex $\mathbf{a} \in \mathcal{V}_H^+ \cup \mathcal{V}_H^0$. The purpose of such functions, is to "rectify" the original T operator so that T_H maps into the kernel W of I_H. Importantly, these functions need to be supported in Ω_+, so that T_H has the same "symmetrization" property as T (see (2.3a)).

The actual construction of the dual functions is technical, so that the proof of the following Lemma is delayed until Appendix A.2.

Lemma 4.3. *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

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

We are now ready to introduce our "discrete" T-operator. For $v \in H_0^1(\Omega)$, it is defined as a modified version of T by:

$$\mathrm{T}_H v = \mathrm{T} v - \sum_{\mathbf{a} \in \mathcal{V}_H^0 \cup \mathcal{V}_H^+} m^{\mathbf{a}}(\mathrm{T} v) \eta^{\mathbf{a}}. \tag{4.2}$$

We will establish in the next Section that $a(\cdot,\cdot)$ is indeed T_H-coercive over W. In addition, let us remark that, 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

$$\operatorname{supp}(\mathrm{T}_H v) \cap \Omega_- = \operatorname{supp}(\mathrm{T} v) \cap \Omega_- \tag{4.3a}$$

and

$$\operatorname{supp}(\mathrm{T}_H v) \cap \Omega_+ = \left\{K \in \mathcal{T}_H \mid \operatorname{supp}(\mathrm{T} v) \cap \overline{K} \neq \emptyset\right\} \cap \Omega_+ \tag{4.3b}$$

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

Remark 4.4. An instinctive choice for the discrete operator is $\mathrm{T}_H := (\mathrm{id} - I_H) \circ \mathrm{T}$, or equivalently

$$\mathrm{T}_H v = \mathrm{T} v - \sum_{\mathbf{a} \in \mathcal{V}_H^0 \cup \mathcal{V}_H^+} m^{\mathbf{a}}(\mathrm{T} v) \psi^{\mathbf{a}}, \quad v \in H_0^1(\Omega).$$

While this definition automatically maps into W, it is not satisfactory, since this operator does not flip the sign of the argument in Ω_- (requirement (2.3a)). Indeed, the corrections at the vertices lying on the interface would "leak" in Ω_- as the support of $\psi^{\mathbf{a}}$ intersects Ω_- in this case.

Remark 4.5. In the definition of T_H, the correction functions $\eta^{\mathbf{a}}$ are supported in Ω_+, since we chose to symmetrize "from Ω_- to Ω_+". If the other direction, is considered (i.e. $\mathrm{S} \in \mathcal{L}(H^1_{\partial\Omega}(\Omega_+); H^1_{\partial\Omega}(\Omega_-)))$, then these functions have to be supported in Ω_-. As can be seen from the proof of Lemma 4.3 in the appendix, it is easy to design $\eta^{\mathbf{a}}$ with support in Ω_- instead of Ω_+, so that every "direction" can be considered for symmetrization purposes.

4.3. T-coercivity in the kernel of I_H

We are now ready to establish the inf-sup stability of $a(\cdot,\cdot)$ over W, under the assumption that the contrast $\mathscr{C}$ is sufficiently large. The proof is based on Theorem 2.2 with the operator $\mathrm{Y} := \mathrm{T}_H$. Hence, we verify that T_H satisfies the requirements of Theorem 2.2. We first show that $\mathrm{T}_H \in \mathcal{L}(W)$.

Lemma 4.6. *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

$$m^{\mathbf{a}}(w) = 0 \quad \forall \mathbf{a} \in \mathcal{V}_H^{\mathrm{int}}. \tag{4.4}$$

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 (4.5) 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 $\eta^{\mathbf{a}}$ (see Lemmas 4.1 and 4.3). □

We now verify that T_H satisfies requirement (2.3) of Theorem 2.2.

Lemma 4.7. *Let $w \in W$, it holds that*

$$(\mathrm{T}_H w)|_{\Omega_-} = -w|_{\Omega_-}. \tag{4.5}$$

In addition, we have

$$|w - \mathrm{T}_H w|_{1,\Omega_+} \leq C_{\pm}(\mathrm{T}_H) |w|_{1,\Omega_-}, \tag{4.6}$$

with

$$C_{\pm}(\mathrm{T}_H) := C_{\pm}(\mathrm{T}) + 2(d+1)\widehat{C}_{\mathrm{P}}\widehat{C}_{norm}\widehat{C}_{inf}\widehat{C}_{inv}C_{\mathrm{u}}\kappa\sqrt{2 + C_{\pm}^0(\mathrm{T})^2},$$

where κ, $C_{\pm}(\mathrm{T})$, and $C_{\pm}^0(\mathrm{T})$ are introduced in Section 2.4 and the other constants are explained in Sections 3.1.1 and 3.1.2.

Remark 4.8. We emphasize that the constant $C_{\pm}(\mathrm{T}_H)$ is bounded independently of the mesh size H. Actually, it only depends on the original operator T, and the mesh shape-regularity parameter κ.

Proof. Identity (4.5) is a 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 thus focus on (4.6). Let $w \in W$. We have

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

so that

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

We have

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

Then, for each $\mathbf{a} \in \mathcal{V}_H^0 \cup \mathcal{V}_H^+$, it holds with Lemmas 4.1 and 4.3 that

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

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

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

Moreover, we have by Lemma 4.2

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

Hence, combining all the foregoing estimates, we finally deduce

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

and the result follows. □

We can now conclude this section with Theorem 4.9 establishing T_H-coercivity of $a(\cdot,\cdot)$ in the kernel W. The proof is direct consequence of Theorem 2.2 and Lemma 4.7.

Theorem 4.9. *Under the assumption that*

$$\sqrt{\mathscr{C}} > \frac{C_{\pm}(\mathrm{T}_H)}{2} \left(\frac{\sup_{\Omega_+} \sigma}{\inf_{\Omega_+} \sigma} \right), \tag{4.7}$$

we have

$$a(w, \mathrm{T}_H w) \geq \alpha_\kappa |w|^2_{1,\sigma,\Omega} \qquad \forall w \in W \tag{4.8}$$

where

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

5. Towards a practical method

In this section, we address the second and third challenge discussed in Section 3.3, namely the localization of the corrector computations and their discretization. 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 $a \gtrsim b$.

5.1. Localized correctors

In this section, we will show how to localize the computation of the correctors defined in (3.6). Note that due to linearity, $\mathcal{Q}$ can be written as $\mathcal{Q} = \sum_{K \in \mathcal{T}_H} \mathcal{Q}_K$, where $\mathcal{Q}_K$ is defined via

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

Here and in the following, $a_D(\cdot,\cdot)$ denotes the restriction of $a(\cdot,\cdot)$ to a subdomain $D \subset \Omega$.

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 [24, 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 $\|w^\star\|_{1,D} \lesssim \|w\|_{1,\widetilde{D}}$ 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.

Recall the definition of the mth layer patch $\mathrm{N}^m(D)$ around $D \subset \Omega$ from Section 3.1. 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.,

$$\max_{T \in \mathcal{T}_H} \operatorname{card}\{K \in \mathcal{T}_H \quad | \quad K \subset \mathrm{N}^m(T)\} \leq C_{\mathrm{ol},m}. \tag{5.1}$$

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 the usual proof because T_H involves a symmetrization operator and thus, is inherently non-local. This is why we introduce the following "symmetric" patches $\mathrm{P}^m(K) := (\mathrm{P}^m(K) \cap \Omega_-) \cup (\mathrm{P}^m(K) \cap \Omega_+)$ by

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

and

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

We emphasize that this does not require the mesh $\mathcal{T}_H$ to be symmetric. In view of (4.3), the idea of $\mathrm{P}^m(K)$ is that, for any function $v \in H^1_0(\Omega)$ 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. Some examples of $\mathrm{P}^1(K)$ for an interface-resolving, but non-symmetric mesh are illustrated in Figure 5.1.

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 5.3.

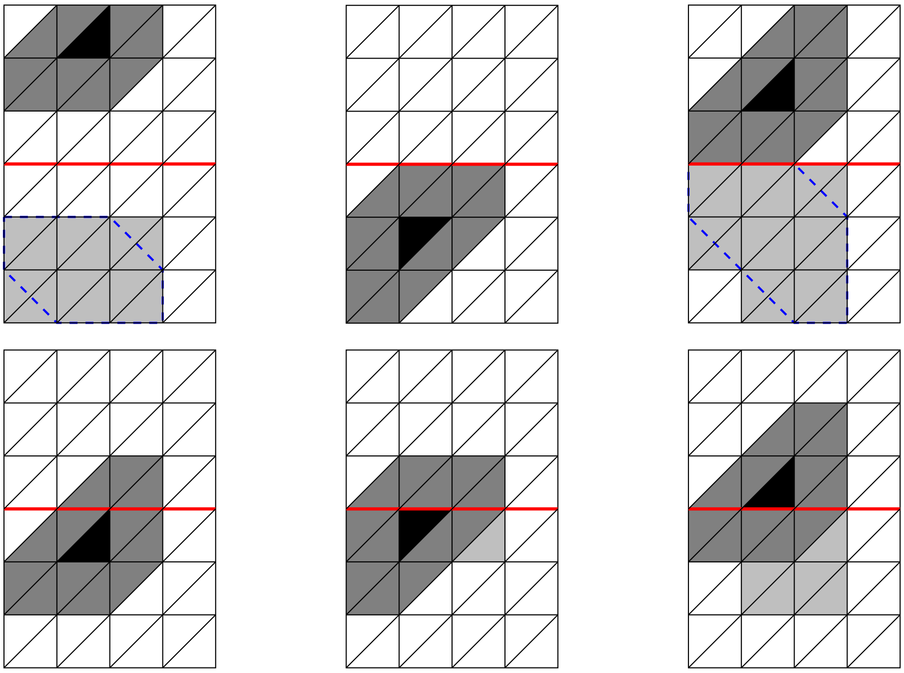

Figure 5.1: Illustration of $\mathrm{P}^1(K)$ for different triangles K. The red line is the interface Γ, Ω_- is the upper half and Ω_+ the lower half. Triangle K in black, $\mathrm{N}^1(K)$ consists of K and additional elements in gray, $\mathrm{P}^1(K)$ consists of $\mathrm{N}^1(K)$ and additional elements in light gray. In the top line, dashed blue lines indicate the area of $\mathrm{N}^1(K)$ under symmetrization.

Proposition 5.1. *There is $0 < \tilde{\gamma} < 1$, independent of H, such that for any $K \in \mathcal{T}_H$ and all $v_H \in V_H$*

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

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

$$a_{\mathrm{P}^m(K)}(\mathcal{Q}_{K,m} v_H, w) = a_K(v_H, w) \qquad \text{for all} \quad w \in W(\mathrm{P}^m(K)). \tag{5.2}$$

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 (5.2). Hence, we can solve (5.2) 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 elements 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 5.3.

Theorem 5.2. *There exists $0 < \gamma < 1$, independent of H, such that for any $v_H \in V_H$*

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

In our generalized finite element method, we now replace $\mathcal{Q}$ in (3.7) by $\mathcal{Q}_m$, exactly in the spirit of LOD. Hence, we seek $u_{H,m} \in V_H$ such that

$$a((\mathrm{id} - \mathcal{Q}_m)u_{H,m}, (\mathrm{id} - \mathcal{Q}_m)v_H) = (f, (\mathrm{id} - \mathcal{Q}_m)v_H)_\Omega \qquad \text{for all} \quad v_H \in V_H. \tag{5.3}$$

The numerical analysis relies on the error estimate for the ideal method in Proposition 3.2 and the fact that the localization is a small perturbation thereof.

Theorem 5.3. *Let $m \gtrsim |\log(C_{\mathrm{ol},m}^{1/2}\tilde{\alpha})|$ with the inf-sup constant $\tilde{\alpha}$ of Proposition 3.2. Then* (5.3) *is well-defined and the unique solution $u_{H,m}$ satisfies the error estimates*

$$\|u - (\mathrm{id} - \mathcal{Q}_m)u_{H,m}\|_{1,\Omega} \lesssim (H + C_{\mathrm{ol},m}^{1/2}\,\gamma^m)\|f\|_{0,\Omega}, \tag{5.4}$$

$$\|u - u_{H,m}\|_{0,\Omega} \lesssim H \inf_{v_H \in V_H} |u - v_H|_{1,\Omega} + C_{\mathrm{ol},m}^{1/2}\,\gamma^m(H + C_{\mathrm{ol},m}^{1/2}\gamma^m)\|f\|_{0,\Omega}. \tag{5.5}$$

Note that the oversampling condition $m \gtrsim |\log(C_{\mathrm{ol},m}^{1/2}\tilde{\alpha})|$ is independent of H. Since $C_{\mathrm{ol},m}$ grows only polynomially in m, it is fulfillable. We emphasize that, in order to balance the terms H and γ^m in the error estimates, the stronger, but standard, oversampling condition $m \approx |\log(C_{\mathrm{ol},m}^{1/2}H)|$ is required. We summarize that under this (standard) oversampling condition, the method is well-posed, we have linear convergence in the $H^1(\Omega)$-norm (see (5.4)) and up to quadratic convergence of the FE part in the $L^2(\Omega)$-norm (see (5.5)). Note that the second term in (5.5) 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, (5.5) 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 (5.3) follows from an inf-sup condition on $V_{H,m}$ (see [32] for instance). This directly yields quasi-optimality and the error estimate (5.4), where we refer to [24, Chapter 2] for details.

Moreover, a standard duality argument can be employed to show

$$\|u - (\mathrm{id} - \mathcal{Q}_m)u_{H,m}\|_{0,\Omega} \lesssim (H + C_{\mathrm{ol},m}^{1/2}\,\gamma^m)\|u - (\mathrm{id} - \mathcal{Q}_m)u_{H,m}\|_{1,\Omega},$$

i.e., quadratic convergence in the $L^2(\Omega)$-norm. We refer to, e.g., [32] for details.

Finally, we have that

$$\|u - u_{H,m}\|_{0,\Omega} \leq \|u - I_H u\|_{0,\Omega} + \|I_H u - u_{H,m}\|_{0,\Omega} \lesssim H|u - I_H u|_{1,\Omega} + \|I_H u - u_{H,m}\|_{1,\Omega}.$$

Due to the stability and projection property of I_H, we have $|u - I_H u|_{1,\Omega} \lesssim \inf_{v_H \in V_H} |u - v_H|_{1,\Omega}$ so that it remains to estimate $\|I_H u - u_{H,m}\|_{1,\Omega}$. We note that by the definition of $\mathcal{Q}$ and the stability of I_H it holds that

$$\|I_H u - u_{H,m}\|_{1,\Omega} = \|I_H(\mathrm{id} - \mathcal{Q})(I_H u - u_{H,m})\|_{1,\Omega} \lesssim \|(\mathrm{id} - \mathcal{Q})(I_H u - u_{H,m})\|_{1,\Omega}.$$

Due to Proposition 3.2, there exists $\psi_H \in V_H$ with $\|\psi_H\|_{1,\Omega} = 1$ such that

$$\|(\mathrm{id} - \mathcal{Q})(I_H u - u_{H,m})\|_{1,\Omega} \lesssim a((\mathrm{id} - \mathcal{Q})(I_H u - u_{H,m}), (\mathrm{id} - \mathcal{Q})\psi_H).$$

The definition of $\mathcal{Q}$, Galerkin orthogonality and Theorem 5.2 give that

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

Combination with the estimate for $\|u-(\mathrm{id}-\mathcal{Q}_m)u_{H,m}\|_{1,\Omega}$ finishes the proof. □

5.2. Fully discrete method

We now address the final challenge to obtain a fully practical generalized finite element method: the fact that the spaces W and $W(\mathrm{P}^m(K))$ are 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. $\mathcal{T}_h$ should be a shape-regular refinement of $\mathcal{T}_H$, but note that $\mathcal{T}_h$ is not required to be quasi-uniform. The corrector problems (5.2) are then defined on the discrete space $W(\mathrm{P}^m(K))\cap V_h$ and yield discrete localized correctors $\mathcal{Q}_{m,h}$.

This requires the mesh $\mathcal{T}_h$ to be sufficiently fine in the sense that all multiscale features and jumps of σ are resolved, and in particular it needs to be T-conforming. We point out that 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 strongly believe the above analysis will still hold true with minor modifications due to the additional discretization. We refer the reader to [16] for details on the proof of the exponential decay in this case.

The corresponding solution $u_{H,h,m}$ of our generalized finite element method (5.3) then approximates the FEM solution $u_h\in V_h$ on the fine mesh. In particular, we have by the triangle inequality that

$$\|u-(\mathrm{id}-\mathcal{Q}_{m,h})u_{H,h,m}\|_{1,\Omega}\le\|u-u_h\|_{1,\Omega}+\|u_h-(\mathrm{id}-\mathcal{Q}_{m,h})u_{H,h,m}\|_{1,\Omega}.$$

With the above mentioned modifications, an estimate for $\|u_h-u_{H,h,m}\|_{1,\Omega}$ similar to Theorem 5.3 should hold, namely

$$\|u_h-(\mathrm{id}-\mathcal{Q}_{m,h})u_{H,h,m}\|_{1,\Omega}\lesssim(H+C_{\mathrm{ol},m}^{1/2}\gamma^m)\|f\|_{0,\Omega}$$

with a constant hidden in $\lesssim$ that is independent of H, h, and m. Since $\mathcal{T}_h$ is a fine, not necessarily quasi-uniform, and T-conforming triangulation, it is reasonable to assume that the finescale discretization error $\|u-u_h\|_{1,\Omega}$ is sufficiently small in comparison to the LOD error $\|u_h-u_{H,h,m}\|_{1,\Omega}$. Finally, we note that u_h is not needed for computing $u_{H,h,m}$. However, in numerical experiments where often u is not available, we use u_h as reference solution and evaluate the error $\|u_h-(\mathrm{id}-\mathcal{Q}_{m,h})u_{H,h,m}\|_1$ only.

Concerning the practical implementation of the LOD, we refer the interested reader to [15], where, for instance, the (parallel) computation of the correctors is addressed in detail. In comparison with a standard finite element method on a fine (adaptive) mesh, our method has the advantage of a much smaller linear system to be solved at the cost of a slightly more dense matrix and additional computations (in form of the local correctors) during the assembly of the stiffness matrix. Therefore, the method is particularly attractive if a standard finite element method on a fine grid is not feasible due to the size of the system or if the same multiscale problem has to be solved for many different right-hand sides.

5.3. Proof of the localization error

This section is devoted to the proofs of Proposition 5.1 and Theorem 5.2. 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 $\mathcal{R} := \operatorname{supp}(\boldsymbol{\nabla}\eta)$. Given any subset $D \subset \Omega$ as the union of elements in $\mathcal{T}_H$, any $w \in W$ satisfies that

$$\|w\|_{0,D} \lesssim H\|\boldsymbol{\nabla} w\|_{0,\mathrm{N}(D)}, \tag{5.6}$$

$$\|(\mathrm{id} - I_H)(\eta w)\|_{0,D} \lesssim H\|\boldsymbol{\nabla}(\eta w)\|_{0,\mathrm{N}(D)}, \tag{5.7}$$

$$\|\boldsymbol{\nabla}(\eta w)\|_{0,D} \lesssim \|\boldsymbol{\nabla} w\|_{0,D\cap\operatorname{supp}\eta} + \|\boldsymbol{\nabla} w\|_{0,\mathrm{N}(D\cap\mathcal{R})}. \tag{5.8}$$

These properties are proved in [16, Lemma 2].

Proof of Proposition 5.1. Fix $K \in \mathcal{T}_H$, $v_H \in V_H$ and m. 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 $\mathcal{R} = \operatorname{supp}(\boldsymbol{\nabla}\eta)$ 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,

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

Note that $\operatorname{supp}(\widetilde{\phi}-\phi)\cap\operatorname{supp}(\mathrm{T}_H\widetilde{\phi}) \subset \mathrm{N}^1(\mathcal{R})$ and $\|\mathrm{T}_H\widetilde{\phi}\|_{\mathrm{N}(\mathcal{R})} \lesssim \|\widetilde{\phi}\|_{1,\mathrm{N}^2(\mathcal{R})}$ 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 (5.7) as well as (5.8), 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 5.1 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 [14] and Section 6.

Proof of Theorem 5.2. We start by proving the following local estimate

$$\|(\mathcal{Q}_K - \mathcal{Q}_{K,m})v_H\|_{1,\Omega} \lesssim \tilde{\gamma}^m \|v_H\|_{1,K} \tag{5.9}$$

for some $0 < \tilde{\gamma} < 1$ and for any $v_H \in V_H$ and $K \in \mathcal{T}_H$ as well as any (fixed) m. 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

$$\|(\mathcal{Q}_K - \mathcal{Q}_{K,m})v_H\|_{1,\Omega} \lesssim \inf_{w_{K,m} \in W(\mathrm{P}^m(K))} \|\mathcal{Q}_K v_H - w_{K,m}\|_{1,\Omega}. \tag{5.10}$$

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 (5.10) and noting that $I_H(\mathcal{Q}_K v_H) = 0$, we obtain

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

where the last inequality follows from the properties (5.7) and (5.8) similar to the arguments in the proof of Proposition 5.1. Combination with Proposition 5.1 gives (5.9).

To prove Theorem 5.2, we define, for a given simplex $K \in \mathcal{T}_H$ and a given number of layers m, 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

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

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 (5.8), the stability of I_H (3.5) and $\|\mathrm{T}_H w\|_{\mathrm{N}(\{\eta \neq 1\})} \lesssim \|w\|_{\mathrm{N}^2(\{\eta \neq 1\})}$, we deduce

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

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

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

which in combination with (5.9) finishes the proof. □

6. Numerical experiments

The numerical examples were carried out in MATLAB based upon preliminary code developed at the Chair for Computational Mathematics at University of Augsburg. We always consider

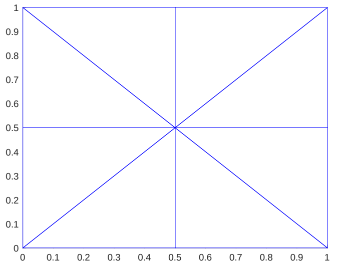

Figure 6.1: Building block for the meshes in the numerical experiments

$\Omega = [0,1]^2$. Our meshes are constructed out of blocks as depicted in Figure 6.1: A mesh size of $H = 2^{-N}$ with $N \geq 1$ means that the mesh consists of $N \times N$ blocks of Figure 6.1. The fine mesh has $h = 2^{-8}$ and is T-conform in all settings described below except from the circular inclusion in Section 6.3. This fine mesh is used for the corrector computations and, additionally, for the computation of a reference solution u_h using standard FEM in Sections 6.2 and 6.4. 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 (5.3) 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 absolute error of the standard FEM. From (5.4), 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. (5.5)).

Finally, we note that, although our theory guarantees 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.

6.1. 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. We shifted the interface Γ from the middle line in order to have meshes $\mathcal{T}_H$ that do not resolve the interface and that are not symmetric for any H. Hence, we expect a poor performance of the standard FEM. In this case, $C_\pm(T)$ can be analytically computed: We obtain $C_\pm(T) = 2\sqrt{\frac{0.5+2^{-7}}{0.5-2^{-7}}}$, such that $a(\cdot,\cdot)$ is T-coercive if $\frac{\sigma_-}{\sigma_+} > \frac{0.5+2^{-7}}{0.5-2^{-7}} \approx 1.0317$, see also [11]. Hence, the model problem is well-posed for our choices of σ_-, but note that the condition for T_H-coercivity is most probably violated.

We consider the following piecewise smooth function fulfilling homogeneous Dirichlet boundary conditions

$$u(x_1,x_2) = \begin{cases} -\sigma_- x_1(x_1-1)x_2(x_2-1)(x_2-l), & (x_1,x_2) \in \Omega_+, \\ x_1(x_1-1)x_2(x_2-1)(x_2-l), & (x_1,x_2) \in \Omega_-, \end{cases}$$

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, $f(x_1,x_2) = \sigma_-(2x_2(x_2-1)(x_2-l) + x_1(x_1-1)(6x_2 - 2(l+1))$ and we note that f is globally smooth.

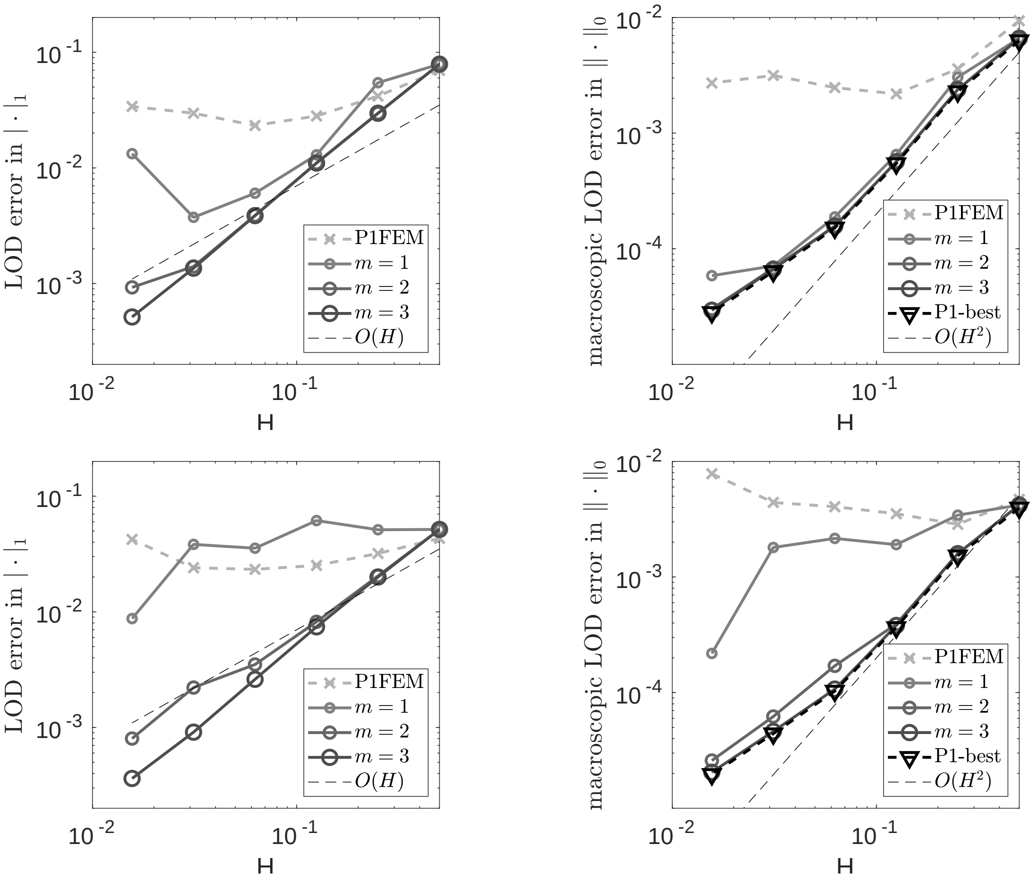

Figure 6.2: Convergence histories for the flat interface with $\sigma_- = 2$ (top) and $\sigma_- = 1.1$ (bottom) in Section 6.1.

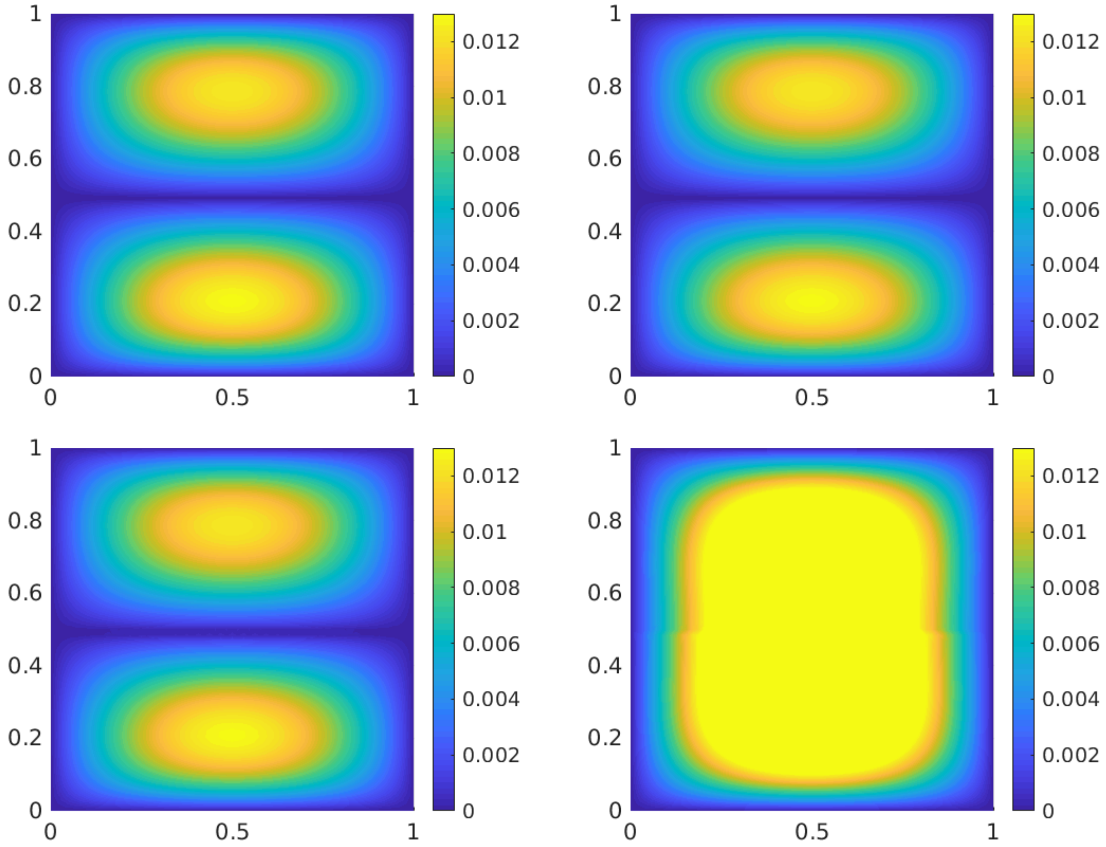

Figure 6.3: Various solutions for the flat interface with $\sigma_- = 1.1$ (Section 6.1): exact solution u (top left), LOD solution (top right), macroscopic part of the LOD solution ($u_{H,m}$, bottom left) and FE solution (bottom right).

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 6.2. 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 5.3. In contrast to the good performance of the LOD, we see the failure of the standard FEM in Figure 6.2. This is of course expected from the fact that $\mathcal{T}_H$ does not resolve the interface. 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 6.2 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 6.3. 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

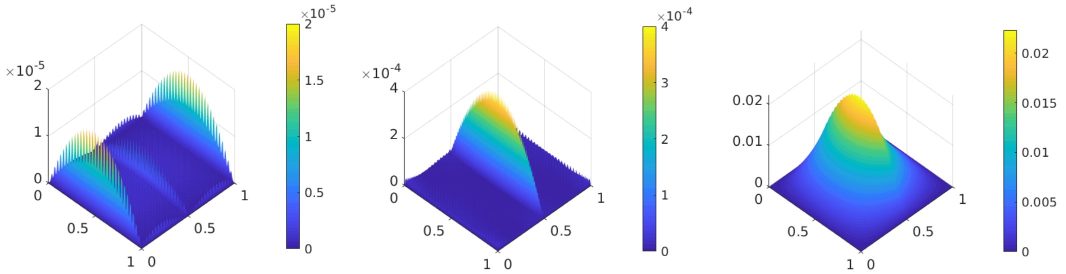

Figure 6.4: Errors of the different solutions to u for the interface with $\sigma_- = 1.1$ (Section 6.1): error of LOD solution (left), error of macroscopic part of LOD solution (middle), and error of FE solution (right).

make out any differences between the exact solution, the LOD solution and its macroscopic part, 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. Though expected, it is interesting to see how drastically the (slight) unfit of the meshes to interface influences the performance of the standard FEM.

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 6.4. Here, we clearly see a difference in the error distribution. The FE error (right) is very large close to the interface and this error 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.

6.2. Square inclusion

We consider $\Omega_- = [0.25, 0.75]^2$ and Ω_+ as the complement. The coefficient σ is spatially varying, more precisely $\sigma|_{\Omega_+}(x) = 0.75 + 0.125\cos(2\pi\frac{x_1}{\varepsilon}) + 0.125\sin(2\pi\frac{x_2}{\varepsilon})$ and $\sigma|_{\Omega_-}(x) = -5 + 0.5\sin(2\pi\frac{x_1}{\varepsilon}) + 0.5\cos(2\pi\frac{x_2}{\varepsilon})$ with $\varepsilon = 2^{-7}$. According to [3], the model problem is T-coercive for this geometry and choice of σ. We set $f = 0.1\chi_{\{x_2<0.1\}} + \chi_{\{x_2>0.1\}}$ to have a right-hand side only in $L^2(\Omega)$ and, at the same time, to keep the sign-change in σ and the jump in f apart from each other. The coefficient σ and the reference solution u_h, computed by a standard FEM on the fine mesh $\mathcal{T}_h$, are depicted in Figure 6.5. Note that the fine mesh resolves the oscillations of σ. All coarse meshes $\mathcal{T}_H$ resolve the interface and are T-conform, but note that they do not resolve the multiscale variations of σ.

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 6.6. We again observe the expected overall 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. The error for the standard finite element method is mostly decaying as well, but at a higher level in comparison to the LOD solution with $m = 3$. Moreover, the rate of convergence is definitely lower and we even have a stagnation of the error in the L^2-norm at around $H = 2^{-5}$, where the coefficient variations are not yet resolved. Note that for this

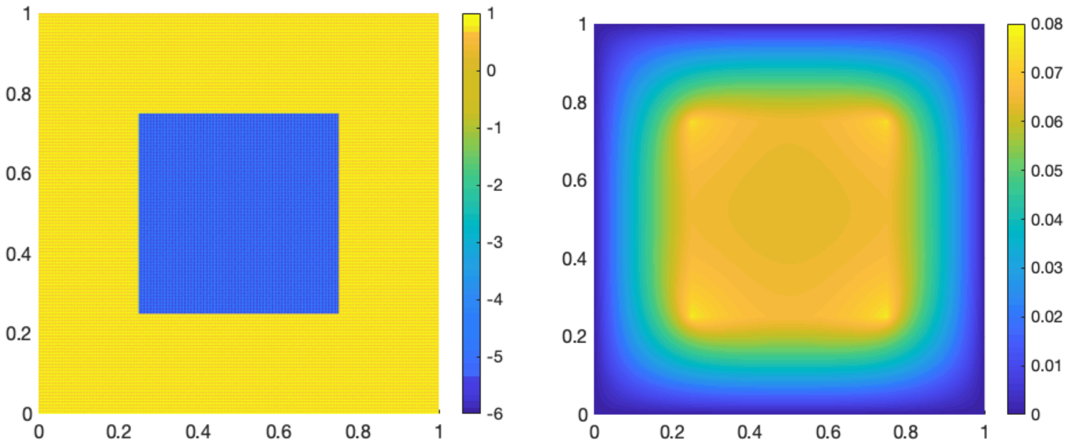

Figure 6.5: Coefficient (left) and fine FE solution (right) for the experiment in Section 6.2.

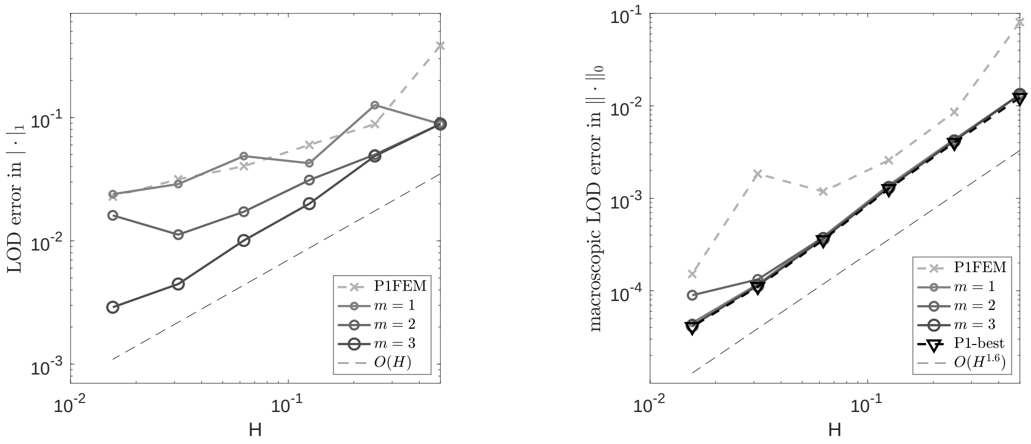

Figure 6.6: Convergence histories for the square inclusion in Section 6.2.

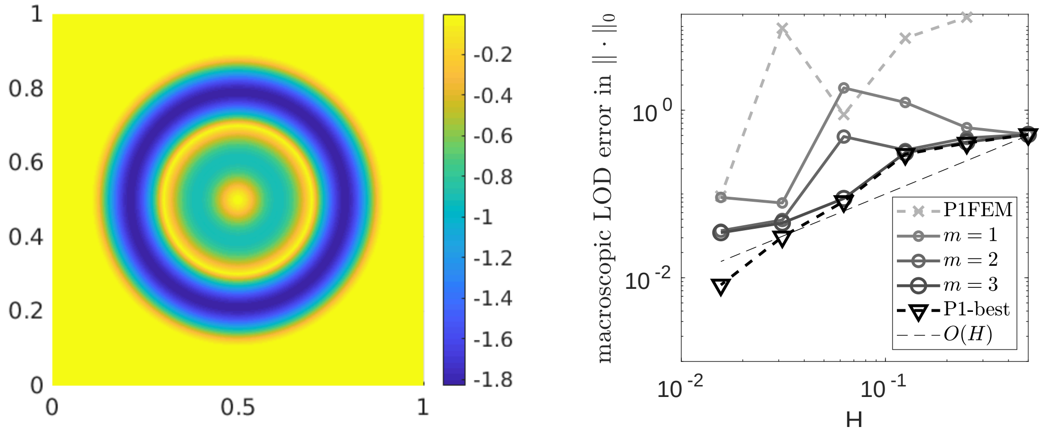

Figure 6.7: Exact solution (left) and convergence history for the macroscopic LOD error in the L^2-norm (right) for the circular inclusion in Section 6.3.

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 1.66 as we calculated by taking the average of the experimental orders of convergence. The regularity $u \in H^{1+\lambda}(\Omega)$ of the exact solution was studied for the present configuration with piecewise constant σ in [7, 26]. Inserting into these results the average values of $\sigma|_{\Omega_+}$ and $\sigma|_{\Omega_-}$, one obtains $\lambda \approx 0.52$. The L^2-best approximation is thus converging slightly faster than the simple ad-hoc regularity calculation for constant coefficients predicts.

This experiment underlines the applicability and advantages of the method for oscillating coefficients. Further, we emphasize that we have linear convergence of the LOD error in the H^1-semi-norm although the exact solution is definitely not in $H^2(\Omega)$ due to the corners at the interface.

6.3. 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 6.1. 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 6.7, 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. We note that in this example, a simple use of a isoparametric elements will most probably yield

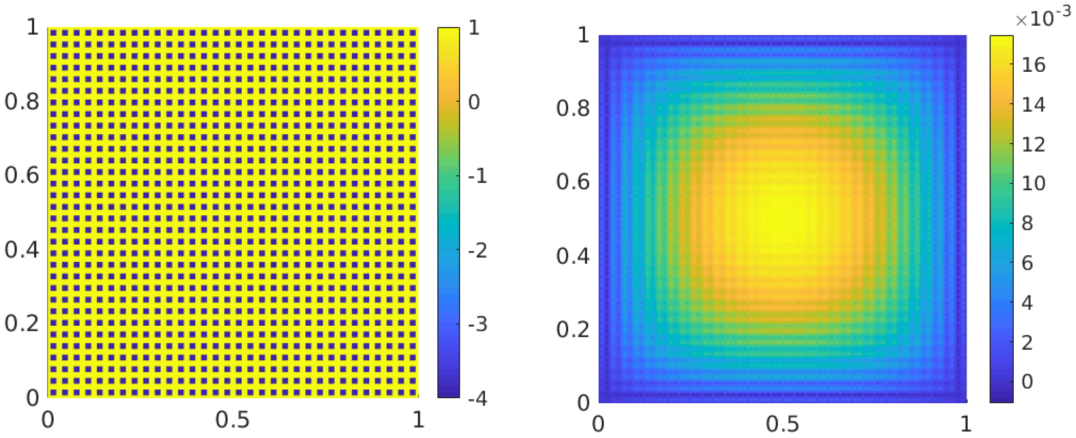

Figure 6.8: Coefficient (left) and fine FE solution (right) for the experiment in Section 6.4.

a good approximation with less computational effort than the LOD, but we nevertheless check the convergence rates of our method. 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 6.7 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 6.7 are very satisfying.

6.4. Multiscale sign-changing coefficient

We consider a multiscale, sign-changing coefficient as depicted in Figure 6.8, 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 6.8, 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 6.9. First of all, we observe that the standard FEM fails on this coarse mesh because the multiscale features of the coefficient are not resolved. To be more precise, FEM on the coarse mesh essentially calculates the solution to a diffusion problem with the coefficient $\tilde{\sigma}$ as the element-wise (arithmetic) mean of σ, i.e., $\tilde{\sigma}|_T = |T|^{-1} \int_T \sigma \, dx$ for all mesh elements T. Since for all coarse mesh elements T, we have $|T \cap \Omega_-| = \frac{1}{4}|T|$ and $|T \cap \Omega_+| = \frac{3}{4}|T|$ this average of σ equals $-\frac{1}{4}$ in this example, which nicely explains the “bump” pointing in the negative direction in Figure 6.9 (right). This observation is already expected and well understood for the classical elliptic diffusion problem, see [28] 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

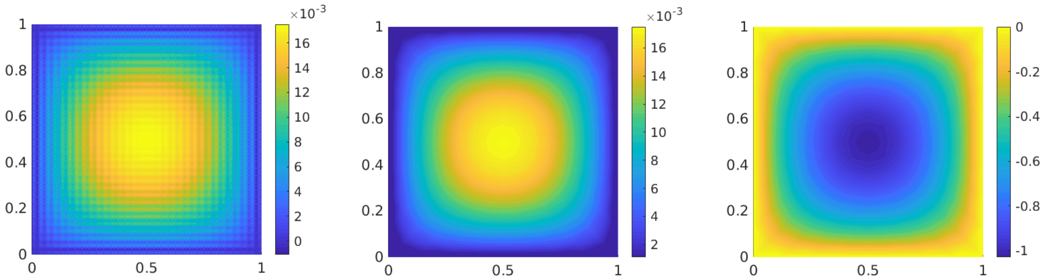

Figure 6.9: LOD solution (left), macroscopic part of LOD solution (middle), and FE solution (right) for $H = 2^{-4}$ and $m = 3$ in the experiment of Section 6.4.

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 coarse 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 [14], where we mention the connection with the LOD approach in weighted norms [18, 30].

Acknowledgments

Major parts of this work were carried out while BV was affiliated with University of Augsburg. BV’s work at KIT is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 258734477 – SFB 1173. We are very grateful to the referees for their valuable comments that helped to improve the paper.

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] E. Bonnetier, C. Dapogny, and F. Triki. Homogenization of the eigenvalues of the Neumann-Poincaré operator. *Arch. Ration. Mech. Anal.*, 234(2):777–855, 2019.

[9] R. Bunoiu and K. Ramdani. Homogenization of materials with sign changing coefficients. *Commun. Math. Sci.*, 14(4):1137–1154, 2016.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

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

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

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

[21] 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.

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

[23] 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.

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

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

[26] 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.

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

[28] 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.

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

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

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

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

[33] 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 3.1 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.

A.1. Key properties of the Oswald operator I_H

Proof of Lemma 4.1. 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 $w := P_{K_\star} v \in \mathcal{P}_1(K_\star)$, the first estimate of (3.2) shows that

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

from which (4.1) follows. □

Proof of Lemma 4.2. Recall the notation for the reference patches introduced at Section 3.1.2. 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 [13, proof of Theorem 3.1.1]) shows that there exists a constant $\widehat{C}$ such that

$$\|\widehat{w}\|_{0,\widehat{\omega}} + |\widehat{w}|_{1,\widehat{\omega}} \leq \widehat{C} \left(|\widehat{m}(\widehat{w})| + |\widehat{w}|_{1,\widehat{\omega}} \right), \quad \forall \widehat{w} \in H^1(\widehat{\omega})$$

justifying estimate (3.3).

Hence, applying (3.3), we have

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

At this point, employing (element-wise) usual scaling arguments, we easily see that

$$\|v\|_{0,\omega^{\mathbf{a}}} \leq \max_{K \in \mathcal{T}_H^{\mathbf{a}}} \max_{\widehat{K} \in \widehat{\mathcal{T}}} \sqrt{\frac{|K|}{|\widehat{K}|}} \|\widehat{v}\|_{0,\widehat{\omega}}$$

and

$$|v|_{1,\omega^{\mathbf{a}}} \leq \max_{K \in \mathcal{T}_H^{\mathbf{a}}} \max_{\widehat{K} \in \widehat{\mathcal{T}}} \sqrt{\frac{|\widehat{K}|}{|K|}} \frac{h_{\widehat{K}}}{\rho_K} \|\widehat{v}\|_{0,\widehat{\omega}}.$$

All in all, recalling that $h_{\widehat{K}} \leq 1$, we obtain that

$$\|v\|_{0,\omega^{\mathbf{a}}} \leq \widehat{C}_{\mathrm{P}} \max_{\widehat{K},\widehat{K}' \in \widehat{\mathcal{T}}} \sqrt{\frac{|\widehat{K}|}{|\widehat{K}'|}} \max_{K,K' \in \mathcal{T}_H^{\mathbf{a}}} \sqrt{\frac{|K|}{|K'|}} \frac{1}{\rho_K} |v|_{1,\omega^{\mathbf{a}}},$$

from which the result follows.

□

A.2. Construction of dual functions

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 [19] for a similar discussion in a different context.

Lemma A.1. *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*

$$(\widehat{\eta}, \widehat{v})_{\widehat{K}} = (\widehat{\lambda}, \widehat{v})_{\widehat{K}} \quad \forall \widehat{v} \in \mathcal{P}_1(\widehat{K}), \tag{A.1}$$

and we have

$$|\widehat{\eta}|_{1,\widehat{K}} \leq \widehat{C}_{norm} \widehat{C}_{inv} \|\widehat{\lambda}\|_{0,\widehat{K}} \tag{A.2}$$

In addition, the equality

$$(\eta, v)_K = (\lambda, v)_K \quad \forall v \in \mathcal{P}_1(K) \tag{A.3}$$

and the estimate

$$|\eta|_{1,K} \leq \frac{\widehat{C}_{norm} \widehat{C}_{inv}}{\rho_K} \|\lambda\|_{0,K} \tag{A.4}$$

hold true. Moreover, whenever $\lambda \in \mathcal{P}_1(K)$, we have

$$\lambda = P_K \eta. \tag{A.5}$$

Proof. Our proof rely on the bubble function $\widehat{b}$ defined in Section 3.1.1. Let $\widehat{\lambda} \in L^2(\widehat{K})$. 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 $\widehat{\eta} := \widehat{b}\widehat{w} \in H^1_0(\widehat{K}) \cap \mathcal{P}_{d+2}(\widehat{K})$ satisfies (A.1). Furthermore, picking the test function $\widehat{v} = \widehat{w}$ in the definition of $\widehat{w}$ and employing (3.1), we have

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

and (A.2) follows recalling (3.2) since

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

and

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

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

At this point (A.3) and (A.4) follows from usual scaling arguments, since $\widehat{H} = 1$, and (A.5) is a direct consequence of (A.3). □

Proof of Lemma 4.3. 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.1 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.4), we have

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

□