Computational Multiscale Methods in Unstructured Heterogeneous Media

Dissertation

zur Erlangung des akademischen Grades

Dr. rer. nat.

eingereicht an der

Mathematisch-Naturwissenschaftlich-Technischen Fakultät

der Universität Augsburg

von

Roland Maier

Augsburg, Januar 2020

Universität
Augsburg
University

Erstgutachter: Prof. Dr. Daniel Peterseim
Zweitgutachterin: Prof. Dr. Tatjana Stykel
Drittgutachter: Prof. Dr. Patrick Henning

Tag der mündlichen Prüfung: 09.03.2020

To my wife Katharina

Abstract

In this thesis, we consider the numerical approximation of solutions of partial differential equations that exhibit some kind of multiscale features. Such equations describe, for instance, the deformation of porous media, diffusion processes, or wave propagation and the multiscale behavior of corresponding solutions is typically the result of material coefficients that include variations on some fine scale. To avoid global computations on scales that resolve the microscopic quantities, the aim is to provide suitable approximations on some coarse discretization level while taking into account these fine-scale characteristics of underlying coefficients. To this end, we employ the framework of *Localized Orthogonal Decomposition* that is able to cope with general heterogeneous coefficients without the requirement for structural assumptions such as periodicity or an explicit characterization of a fine scale. The approach provides adapted finite element functions with improved approximation properties based on localized corrections of classical finite element functions. We introduce the method in an abstract stationary setting and rigorously analyze its convergence behavior in terms of theoretical and numerical investigations. We also present a higher-order generalization of the approach based on non-conforming spaces and study the interplay between the mesh parameter, the polynomial degree, and the localization parameter. We provide convergence results with explicit dependencies on the above-mentioned parameters and present numerical experiments. Further, we consider an inverse problem of recovering information about an underlying diffusion coefficient from given coarse-scale measurements. Instead of reconstructing the actual coefficient, we follow the idea of finding a coarse model in the spirit of general numerical homogenization methods that is able to satisfactorily reproduce the given data. Although this is a seemingly very different setting, the results of the inverse procedure provide a justification of general (forward) numerical homogenization methods (as, e.g., the Localized Orthogonal Decomposition) and therefore solidify the approach from a different point of view. Beyond these stationary problems, we apply the Localized Orthogonal Decomposition method to the wave equation and the multiphysics problem of linear poroelasticity. We provide rigorous convergence studies and numerical examples. The approach displays its full potential in these time-dependent settings in the sense of an overall complexity reduction. In the context of the wave equation, we focus on an explicit time stepping scheme and the effect of the method on the time step restriction. For the poroelastic problem, we use an implicit scheme and introduce an alternative approach that exploits the saddle point structure which arises if the system is first discretized in time.

Acknowledgments

First and foremost, I would like to express my gratitude to my supervisor Prof. Dr. Daniel Peterseim who offered me a position in his work group and provided me insight into the field of multiscale problems and corresponding methods. I cordially thank him for his great support and advice and his mentoring throughout my time in Augsburg and already earlier in Bonn. I am also very grateful to Prof. Dr. Tatjana Stykel and Prof. Dr. Patrick Henning for agreeing to be the co-referees of this thesis.

Besides, I would like to thank my co-authors Robert Altmann, Alfonso Caiazzo, Eric Chung, Stéphane Cotin, Shubin Fu, Paul Hennig, Markus Kästner, Christoph Paulus, Sai-Mang Pun, Dominik Schillinger, Benjamin Unger, and Barbara Verfürth. The content of this thesis is, directly or indirectly, the result of these fruitful collaborations during the last three years.

Further, I am very grateful to all current and former colleagues at the *Chair of Computational Mathematics* and the *Research Unit Applied Analysis* at the University of Augsburg for the nice working atmosphere, fruitful discussions, and their help. Particularly, I want to thank Robert Altmann, Fabian Kröpfl, and Barbara Verfürth for their proofreading and valuable comments.

I gratefully acknowledge financial support by the German Research Foundation (DFG) in the Priority Program 1748 *Reliable simulation techniques in solid mechanics. Development of non-standard discretization methods, mechanical and mathematical analysis* (PE2143/2-2). Besides, I acknowledge support from the Germany/Hong Kong Joint Research Scheme sponsored by the German Academic Exchange Service (DAAD) under the project 57334719.

Last but not least, I heartily thank my family and friends for being there and helping me wherever possible. Especially, I want to thank my parents, my sisters, and my wife Katharina for their never-ending support, their love, and their belief in me.

Augsburg, January 2020 *Roland Maier*

Contents

List of Figures

List of Tables

1 Introduction

1.1 Motivation

Many physical processes in nature, as for example fluid flows in porous media or general wave propagation through multi-layered soil, involve multiple scales. Typically, one distinguishes between the microscopic scales which describe the highly heterogeneous material properties including the possibly very complex textures of the materials on the one hand, and a macroscopic or effective scale on which the resulting physical phenomena can be observed on the other hand. The presence of multiple scales is also key in the manufacturing of modern composites where fiber-reinforced materials are produced to enhance the overall strength of the originally homogeneous workpiece. In this context, the artificial microstructure changes the macroscopic behavior of the material and has to be taken into account when modeling, for instance, the deformation under loading.

From a mathematical point of view, the physical processes such as flows in porous media or the deformation of a composite material are described by partial differential equations (PDEs) with one or more material coefficients that encode the physical properties. With multiple scales involved, this means that the coefficients and corresponding solutions of the PDE may vary on a microscopic scale. Nevertheless, in general only effective information, i.e., the behavior of the solutions on the macroscopic scale, is of interest for the understanding and simulation of the respective processes. Although it might seem natural, the straightforward approach of discarding micro-scale features of the coefficients in numerical simulations typically fails to provoke the desired effective solution on the macroscopic scale. Then again, resolving the microscopic coefficients would generally be too costly and thus unfeasible for computer simulation, which calls for an alternative strategy to overcome these problems. In particular, it is important to diligently treat the mismatch between microscopic material properties and the macroscopic observation scale. Corresponding techniques are commonly referred to as *homogenization*. The key idea of *classical homogenization* is to replace the original PDE by a homogenized or effective PDE whose solution describes the behavior of the original solution up to variations on a microscopic scale. Solutions of effective PDEs can then be simulated using standard numerical methods on the macroscopic scale because the microscopic scale has basically been removed. The main drawback of classical homogenization models, however, are structural assumptions such as a clear distinction of scales and periodicity which are required in analytical homogenization theory on which these methods are based. Although many manufactured composites like fiber-reinforced materials generally provide a clear separation of the involved scales, i.e., the size of the fibers and the size of the workpiece, and even a periodic structure, these assumptions are not fulfilled anymore in the presence of material imperfections or perturbations. In the above-mentioned geophysical processes such as flows in porous media, structural assumptions such as periodicity or scale separation are usually not fulfilled either.

In the general setting with only minimal assumptions on the microscopic structure of involved coefficients, so-called *numerical homogenization* methods provide an alternative to classical homogenization. The main idea of these approaches is to enhance standard finite element (FE) methods by modifying FE basis functions in a coefficient-adapted way to obtain optimal approximation spaces on the macroscopic scale of interest. These methods generally have in common an increased computational complexity in the sense that there is a moderate overhead in the support of the basis functions or the number of basis functions per mesh entity. Since this overhead can typically be controlled by the macroscopic scale of interest and retains locality in a reduced sense, these methods are called *quasi-local* in the following. Some more details on such methods and particular examples are given in the next section.

1.2 Overview of the literature

As already mentioned in the previous section, homogenization techniques can basically be divided into two groups: numerical homogenization methods and classical homogenization methods. The latter are based on the mathematical theory of homogenization, i.e., various types of convergence results for sequences of problems indexed by a fine-scale parameter ϵ which tends to zero.

A first convergence type is *G-convergence* introduced by Spagnolo [Spa68] for elliptic second-order symmetric operators. The main result is the existence of a so-called *G-limit* for any sequence of bounded and uniformly elliptic operators. This limit corresponds to the homogenized coefficient and the associated solution captures the effective behavior of the solutions of the ϵ-dependent problems.

To overcome the necessity of symmetric operators, Murat and Tartar [MT97a, MT97b, Tar78] generalized the concept of G-convergence to the non-symmetric case. This type of convergence is known as *H-convergence* and requires some additional assumptions on the sequence to compensate for the lack of symmetry. Note that there are constructive proofs to the existence result of the *H-limit*, also known as *method of oscillating test functions* or *energy method* [Tar78, MT97b] from which corresponding numerical methods can be derived.

Another kind of convergence is the so-called *Γ-convergence* and was introduced by De Giorgi [DG75, DG84]. It is characterized by the convergence of minimizers of ϵ-dependent functionals and thus valid in relatively general settings. The relevance of Γ-convergence to homogenization theory is mainly based on the fact that solving a linear, symmetric PDE is often closely connected to finding the minimizer of an appropriate functional.

A less general type of convergence is the so-called *two-scale convergence* which traces back to Nguetseng [Ngu89] and Allaire [All92]. It is based on the ansatz of a two-scale periodic expansion with a slow variable (macroscopic) and a fast one (microscopic) which is justified by a rigorous two-scale homogenization result. For more details on the different convergence types, see also, for instance, the overview provided in [All97].

The most popular numerical approaches to homogenization which are based on the above-mentioned theoretical results are the Multiscale Finite Element Method (MsFEM) by Hou and Wu [HW97], the Two-Scale Finite Element Method introduced by Matache and Schwab in [MS02], and the approach of E and Engquist [EE03,EE05] known as Heterogeneous Multiscale Method (HMM). The MsFEM uses a set of multiscale basis functions which are constructed by solving operator-adapted problems in each element of a coarse mesh. Further, these functions coincide with classical FE basis functions on the boundary of the elements. The approach of [MS02] builds a two-scale FE space based on a coarse mesh and a local fine-scale space consisting of ϵ-periodic functions for each coarse degree of freedom (DOF). The very general idea of the HMM is to approximate the homogenized coefficients from classical homogenization by solving discrete local cell problems on small patches around quadrature points. All these methods are powerful tools to deal with the discrepancy between microstructural quantities in PDEs and the desired effective behavior of respective solutions. The main drawback, however, are the restrictive assumptions that underlie the analysis of these approaches.

The aim to overcome the aforementioned structural restrictions gave rise to many numerical homogenization methods which are designed to work in very general settings. These methods have in common that they approximate the effective behavior of the solution of a PDE on some coarse scale H, which is typically the mesh size of the underlying FE grid. The involved coefficient and the corresponding solution are assumed to have some kind of fine-scale variations but an explicit characterization of a microscopic scale in terms of a parameter ϵ is generally not required. In particular, these methods aim for error estimates which do not depend on fine-scale variations of the coefficient and especially not on ϵ (if available), in contrast to classical FE methods where such variations severely impact the error estimates.

As far as elliptic problems are concerned, there are several methods which fall into this category. Note that the ones presented here do not at all represent a complete list. One of these methods is the Generalized Finite Element Method (GFEM) that is analyzed in [BL11] and traces back to earlier works [BO83, BCO94]. The main idea is to decompose the domain of interest into local (possibly overlapping) subsets and thus divide the global approximation space into local contributions in the spirit of a *partition of unity* approach.

On each of the subdomains, so-called *local particular solutions* are computed as well as a number of local eigenfunctions to approximate the space of harmonic functions with respect to the given diffusion coefficient. It can be shown that the accuracy of the local approximations depends nearly exponentially on the number of spectral problems, which means that the number of local functions should depend logarithmically on the mesh size H to obtain an overall first-order accurate method.

The general question on the necessary computational overhead was discussed in [GGS12] in the context of so-called Adaptive Local Bases (ALB). The theoretical results state that the overhead should be logarithmically dependent on the mesh size H. In particular, it was shown that, in d dimensions, choosing $\mathcal{O}\left(|\log H|^{d+1}\right)$ non-polynomial local basis functions per mesh entity is enough to retain an H^1-error of order H, independently of the actual fine-scale variations of the coefficient. However, the approach is not constructive in the sense that global fine-scale problems need to be solved in order to derive the method. This issue was later overcome with a fully practical approach in connection with the ALB, see [Wey16].

Målqvist and Peterseim [MP14] were the first who proved that the solution of quasi-local problems is sufficient to obtain a quasi-optimal approximation space. Their approach is known as Localized Orthogonal Decomposition (LOD) and was further refined by Henning and Peterseim in [HP13]. The construction is based on the decomposition of the solution space into a finite-dimensional coarse approximation space and a *fine-scale space* in the spirit of the Variational Multiscale Method (VMM) introduced in [HFMQ98]. The main concept of the LOD is to choose the approximation space as the orthogonal complement of the fine-scale space with respect to a coefficient-dependent bilinear form. The resulting space has improved approximation properties compared to classical finite elements with the same number of DOFs. There is even an explicit bijective transformation between the classical FE space on some prescribed coarse scale H and the new space. This allows one to write the basis functions of the improved space in terms of the classical FE basis functions by subtracting the solutions of auxiliary *corrector problems*. These problems may be localized to local patches of size $H|\log H|$ without an impact on the overall convergence rate, since the solutions of the corrector problems decay exponentially fast. As this thesis is substantially based on the LOD, this method is explained in more detail in Chapter 2.

Another ansatz is based on Rough Polyharmonic Splines (RPS) and is described in [OZB14]. There, a set of generalized splines which include fine-scale information is used to approximate the original problem. These generalized functions, however, require the solutions of more demanding bi-harmonic corrector problems.

The approach known as Generalized Multiscale Finite Element Method (GMsFEM) [EGH13] is based on the ideas of the MsFEM described above and is divided into an *offline* and an *online stage*. In the offline stage, local *snapshot* *spaces* consisting of local solutions are computed on coarse elements using a fine discretization. Then, a spectral decomposition is used to reduce the dimension of these spaces by only taking the eigenfunctions with large energy. In the online stage, when specific model parameters are given, the precomputed spaces are used to define a global multiscale space in order to solve the global problem on the coarse scale. One extension of this approach is the so-called Constraint Energy Minimizing GMsFEM (CEM-GMsFEM) [CEL18], where the spectral decomposition is used to compute new multiscale basis functions that minimize the problem-dependent energy and, additionally, fulfill an orthogonality property in the spirit of the LOD. The aim of this approach is to achieve decay of the basis functions even for problems with high contrast.

In contrast to the aforementioned approaches, Owhadi [Owh15,Owh17] studied the view on numerical homogenization from a game theoretical approach and introduced so-called *gamblets* which are also based on a decomposition of the solution space into orthogonal spaces similar to the LOD. Gamblets extend the classical LOD not only to a multilevel setting but also allow one to go beyond its conforming nature by writing the orthogonalization approach as constrained minimization problem, which enables a wide range of possible constraint conditions.

1.3 Goal and main contribution of this work

The overall purpose of this thesis is to show the potential of the LOD method introduced by Målqvist and Peterseim [MP14] and consolidate the approach from multiple perspectives. To this end, the LOD is first presented in a relatively general framework in Chapter 2 including a systematic derivation with the aim to obtain a first-order method that is able to cope with microscopic dependencies without resolving the underlying scale. The method is rigorously analyzed in the general setting, especially in terms of localization, and numerical examples that show the potential of the method are given.

Another contribution is the extension of the original method to a higher-order variant which enables convergence rates beyond first-order. These rates are generally only limited by the regularity of the right-hand side of the variational problem at hand. In this context, a rigorous analysis of the method in the elliptic setting is presented in Chapter 3 with special focus on how the method depends on the polynomial degree. Further, the interplay between the choice of the polynomial degree and the oversampling parameter is studied. Besides the theoretical investigation of this approach, also numerical studies are presented that indicate an even better behavior of the higher-order LOD method.

The structure and ideas of the LOD are then used in Chapter 4 to justify the general approach of using quasi-local effective models, i.e., models with a controlled variation from locality, to overcome the issues that arise in the presence of multiple scales as it is done for the LOD approach and in numerical homogenization in general. This is achieved in connection with inverse problems where a general coarse-scale model is reconstructed from a given set of coarse data using an iterative optimization technique. The key feature is to let the algorithm decide whether to deviate from locality or not. It can be observed numerically that a slightly non-local model is able to better capture macroscopic effects. In that context, a variant of the LOD is used to motivate the inversion algorithm in the sense that the multiscale model that is obtained with LOD is provably a possible solution of the optimization problem. Moreover, these findings also justify other numerical homogenization methods which are based on some computational overhead per mesh entity, such as the ones mentioned in Section 1.2.

Apart from the extension of the LOD to higher-order schemes and its general justification, this thesis shall also present advantageous side benefits that occur in connection with time-dependent problems where multiscale aspects in the PDE are independent of time and only depend on the spatial variables. As a model example, the acoustic wave equation is considered in Chapter 5 and the common procedure of applying the LOD to the stationary part of the PDE is used to derive a semi-discrete multiscale method which is then combined with an explicit time stepping scheme. The method is theoretically examined with particular focus on the errors introduced by discretization and localization. Further, the time step restriction, also known as Courant-Friedrichs-Lewy (CFL) condition, which is crucial for explicit time stepping, is investigated. The side benefit that comes along with the method is a relaxation of this condition in the sense that the time step only needs to be bounded in terms of the coarse mesh parameter and is independent of any fine discretization or microstructural quantity. This leads to computational savings not only in the spatial discretization but also in the temporal one and shows the true potential of the LOD.

Another time-dependent multiscale model discussed in this work is the problem of linear heterogeneous poroelasticity, which is described by two coupled PDEs, an elliptic and a parabolic one. Besides proving the applicability of the LOD to more involved multiphysics problems with multiple varying microscopic parameters, the main contribution in this part is a variation of the classical approach of applying the LOD to the stationary equation as proposed in [MP17] for the mathematically equivalent problem of linear thermoelasticity. Instead, the method is motivated by the time step dependent problem that arises when first discretizing with respect to the temporal variable. Since the resulting equations have a favorable saddle point structure, the coupling terms in the PDE can be discarded with the side benefit of decoupled corrector problems for the two equations of the poroelastic system. These corrections are still independent of the actual time point of the temporal discretization which leads to a simple multiscale method based on a modification of the classical LOD approach. This method is investigated in terms of a theoretical error analysis and numerical studies in Chapter 6.

Parts of this thesis have already been published or submitted to scientific journals. The work on the reconstruction of an effective model in connection with numerical homogenization was submitted for publication and is available as preprint [CMP19]. The findings on the LOD for the acoustic wave equation in combination with an explicit time discretization scheme were published in *BIT Numerical Mathematics* [MP19]. Finally, the content on the LOD with respect to the problem of linear poroelasticity was published in *Journal of Computational Mathematics* [ACM$^+$20]. The presentation of these findings partially follows the one in the corresponding journal or preprint versions. However, some parts are rephrased or extended and the notation might differ in order to be in line with the other content and the overall reading flow of this thesis.

During the work on this thesis, further research articles were written in the larger context of this thesis [AMU19, FAC$^+$19, HMP$^+$19]. These articles, however, are not directly taken into account in this work.

The numerical experiments presented throughout this thesis were generated either with Python using an adaption of the software of Hellman [Hel17] or with MATLAB based on preliminary code developed at the Chair of Computational Mathematics at the University of Augsburg. A detailed description on the implementation of the LOD method is provided in [EHMP19]. All computations were performed on an HPC Infiniband cluster.

Notation. Throughout this work, we use the following notation. We write C for any positive constant that is independent of the mesh sizes h or H, the polynomial degree p, the time step τ, the oversampling parameter ℓ, and the microscopic scale ϵ. Such constants are allowed to depend on the dimension d and the domain D. Note that C might change from line to line in the estimates. To indicate an explicit dependence on a parameter ξ, we may write C_ξ. We further abbreviate $a \leq C\,b$ and $a \leq C_\xi\,b$ by $a \lesssim b$ and $a \lesssim_\xi b$, respectively, and use $a \sim b$ if $a \lesssim b$ and $a \gtrsim b$.

2 The Classical Localized Orthogonal Decomposition Method

This chapter is devoted to a review of the classical Localized Orthogonal Decomposition (LOD) method introduced in [MP14] and further elaborated in [HP13] for an elliptic model problem. As already mentioned in the introduction, the objective of this technique is to provide suitable approximations of solutions of PDEs on a coarse scale of interest. While classical FE approaches are generally well suited for the approximation of such problems, these methods fail to satisfactorily describe the behavior of PDE solutions if the respective problems involve one or more heterogeneous coefficients which may vary on some microscopic scale. We show throughout this chapter that the LOD approach is able to overcome the discrepancy between microscopic information and a coarse approximation scale and works under minimal structural assumptions. This is achieved by the idea of decomposing a given solution space into a fine-scale space and its coarse complement in a problem-adapted fashion. Since the complementary space is well suited for computations on the coarse scale, the idea of the method is a continuous Galerkin (cG) approach using a localized version of this space. It is computed based on quasi-local auxiliary problems which explains the name *Localized Orthogonal Decomposition*. The method is designed to work for relatively general settings and presents, to some extend, a natural generalization of classical homogenization approaches. That is, in certain periodic regimes with an explicit characterization of microscopic coefficients, the (ideal) method recovers the classical homogenization limit in the elliptic setting; see [GP17]. Note that there exist also alternative formulations of the method as an iterative approach based on an overlapping domain decomposition. This more abstract way of interpreting the LOD in terms of an additive Schwarz method is, for example, investigated in [KY16, KPY18].

In the following, we formulate the classical LOD developed in [MP14,HP13] in a relatively general framework that includes the cases of the subsequent chapters, e.g., the stationary problem in connection with linear poroelasticity (Chapter 6). A general setting has already been considered in [Pet16] but a complete generalized error analysis was only indicated. In this chapter, we fill this gap and provide a rigorous derivation and analysis of the LOD approach in the general case and identify sufficient conditions for the applicability of the method.

2.1 Model problem

Let $d \in \mathbb{N}$ and $D \subseteq \mathbb{R}^d$ be a bounded, convex, and polytopal Lipschitz domain. Further, let $\Gamma \subseteq \partial D$ be the Dirichlet boundary with non-zero $(d-1)$-dimensional Hausdorff measure, i.e., $|\Gamma| > 0$, and denote with $H^1_\Gamma(D)$ the space of H^1 functions with values in $\mathbb{R}$ and vanishing traces on Γ. If $\Gamma = \partial D$, we write $H^1_0(D) := H^1_{\partial D}(D)$. Due to the *Friedrichs inequality*, also known as *Poincaré-Friedrichs inequality* (see, e.g., [Bre03]), we equip the space $H^1_\Gamma(D)$ with the H^1-seminorm $|\cdot|_{H^1(D)} := \|\nabla \cdot\|_{L^2(D)}$, which is a full norm in $H^1_\Gamma(D)$. For some $n \in \mathbb{N}$, let

$$\mathcal{H} := \left[L^2(D)\right]^n \quad \text{and} \quad \mathcal{V} := H^1_{\Gamma_1}(D) \times \ldots \times H^1_{\Gamma_n}(D),$$

where $\Gamma_i \subseteq \partial D$ (with $|\Gamma_i| > 0$) denotes the Dirichlet boundary of the ith component. Let $\mathcal{V}^*$ and $\mathcal{H}^*$ be the dual spaces of $\mathcal{V}$ and $\mathcal{H}$, respectively, and observe that

$$\mathcal{V} \hookrightarrow \mathcal{H} \cong \mathcal{H}^* \hookrightarrow \mathcal{V}^*,$$

where $\hookrightarrow$ denotes a continuous embedding. For completeness, we also introduce the space

$$\bar{\mathcal{V}} := \left[H^1(D)\right]^n \hookleftarrow \mathcal{V}$$

without boundary conditions. Further, define $\mathcal{V}(S)$ and $\mathcal{H}(S)$ as the restrictions of functions in $\mathcal{V}$ and $\mathcal{H}$, respectively, to a subdomain $S \subseteq D$, i.e.,

$$\mathcal{V}(S) = \{v|_S : \ v \in \mathcal{V}\} \quad \text{and} \quad \mathcal{H}(S) = \{v|_S : \ v \in \mathcal{H}\}.$$

In this chapter, we consider the general variational model problem of finding the solution $u \in \mathcal{V}$ of

$$\mathfrak{a}(u, v) = \mathcal{F}(v) \tag{2.1}$$

for all $v \in \mathcal{V}$, where $\mathcal{F} \in \mathcal{V}^*$ is a bounded linear functional, and $\mathfrak{a}\colon \mathcal{V} \times \mathcal{V} \to \mathbb{R}$ is a bilinear form which is bounded from above by

$$|\mathfrak{a}(v, w)| \leq \beta \, \|v\|_{\mathcal{V}} \, \|w\|_{\mathcal{V}} \tag{2.2}$$

for all v, $w \in \mathcal{V}$ and that fulfills the inf-sup condition

$$0 < \alpha := \inf_{v \in \mathcal{V}} \sup_{w \in \mathcal{V}} \frac{\mathfrak{a}(v, w)}{\|v\|_{\mathcal{V}} \, \|w\|_{\mathcal{V}}} = \inf_{w \in \mathcal{V}} \sup_{v \in \mathcal{V}} \frac{\mathfrak{a}(v, w)}{\|v\|_{\mathcal{V}} \, \|w\|_{\mathcal{V}}}. \tag{2.3}$$

Here and in the following, zero is implicitly excluded in the infima and suprema. With regard to the possible choices of the space $\mathcal{V}$, we can think of (2.1) as the variational problem corresponding to a general linear second-order PDE. Note that we do not require symmetry of the bilinear form $\mathfrak{a}$. Under the above assumptions, it follows that (2.1) has a unique solution $u \in \mathcal{V}$ which is bounded by

$$\|u\|_{\mathcal{V}} \leq \alpha^{-1} \|\mathcal{F}\|_{\mathcal{V}^*}, \tag{2.4}$$

see, e.g., [Bab71] for the details.

2.2 Finite-dimensional approximation

In this section, we are concerned with finite-dimensional approximations of problem (2.1). To this end, let $\{\mathcal{T}_H\}_{H>0}$ be a family of regular decompositions (also referred to as *meshes*) of the domain D into *d-rectangles* as described in [Cia78, Ch. 2 & 3]. That is, any $(d-1)$-dimensional face of a d-rectangle (or *element*) $K \in \mathcal{T}_H$ is either a subset of the boundary ∂D or a face of another element. In particular, we pose the assumption that the domain D is such that a decomposition into elements as described above is possible. However, we remark that this condition is not necessarily required since, e.g., curved elements (see [CR72, Zla73]) or non-matching decompositions could be used. Further, we assume *quasi-uniformity* of the family $\{\mathcal{T}_H\}_{H>0}$ in the sense that there are constants c_{qu}, $C_{\mathrm{qu}} > 0$ such that for any mesh $\mathcal{T}_H$ with characteristic mesh parameter H, all elements $K \in \mathcal{T}_H$ satisfy

$$c_{\mathrm{qu}}\, H_K \leq H \leq C_{\mathrm{qu}}\, H_K,$$

where H_K is the diameter of K. The quasi-uniformity allows us to only use the mesh parameter $H > 0$ in the following, instead of the specific diameters H_K of elements $K \in \mathcal{T}_H$.

Let now $H > 0$ be fixed and denote with $V_H \subseteq \mathcal{V}$ the corresponding conforming Q_1 FE space, i.e.,

$$V_H := \left\{ \begin{array}{r} v \in \mathcal{V} :\ \forall K \in \mathcal{T}_H :\ v|_K \text{ is a polynomial of coordinate} \\ \text{degree} \leq 1 \text{ in every component} \end{array} \right\}.$$

Alternatively, we could as well consider decompositions of D into simplices. In this case, V_H denotes the P_1 FE space of piecewise affine and continuous functions. We note that the following construction works analogously if P_1 finite elements are considered instead of Q_1 elements and restrict ourselves to decompositions $\mathcal{T}_H$ into d-rectangles and the corresponding spaces. Further, we emphasize that in view of the inclusion $V_H \subseteq \mathcal{V}$ we also pose the assumption that the Dirichlet boundaries Γ_i, $i \in \{1, \ldots, n\}$, are unions of faces of elements in $\mathcal{T}_H$.

2.2.1 Classical finite element method

A straightforward approach of discretizing problem (2.1) with the classical cG FE method reads as follows: find $u_H \in V_H$ that solves

$$\mathfrak{a}(u_H, v_H) = \mathcal{F}(v_H) \tag{2.5}$$

for all $v_H \in V_H$. In this general setting, the well-posedness of (2.5) requires a discrete inf-sup condition, similar to the one in (2.3), i.e.,

$$0 < \alpha_H := \inf_{v_H \in V_H} \sup_{w_H \in V_H} \frac{\mathfrak{a}(v_H, w_H)}{\|v_H\|_{\mathcal{V}}\, \|w_H\|_{\mathcal{V}}}. \tag{2.6}$$

Let us also assume that there exists a constant $\alpha_0 > 0$ with

$$\alpha_0 \leq \inf_{H>0} \alpha_H. \tag{2.7}$$

With these additional assumptions on the discrete spaces, we can show the following quasi-optimality result (cf. [XZ03, Thm. 2]), also known as *Céa's Lemma.*

Lemma 2.2.1 (Céa's Lemma)**.** *Suppose that the assumptions* (2.2)*,* (2.3)*,* (2.6)*, and* (2.7) *hold. Then the discrete solution* u_H *of* (2.5) *is quasi-optimal in the sense that*

$$\|u - u_H\|_{\mathcal{V}} \leq \frac{\beta}{\alpha_0} \inf_{v_H \in V_H} \|u - v_H\|_{\mathcal{V}},$$

where $u \in \mathcal{V}$ *is the solution of* (2.1)*.*

Proof. Let $\mathcal{G}\colon \mathcal{V} \to V_H$ be the *Galerkin projection* defined, for any $v \in \mathcal{V}$, as the solution of

$$\mathfrak{a}(\mathcal{G}v, w_H) = \mathfrak{a}(v, w_H)$$

for all $w_H \in V_H$, which is well-posed with (2.6). Note that $\mathcal{G}u = u_H$ by the *Galerkin orthogonality*

$$\mathfrak{a}(u - u_H, w_H) = \mathcal{F}(w_H) - \mathcal{F}(w_H) = 0$$

which holds for all $w_H \in V_H$. Since $\mathcal{G}$ is a projection, we obtain for any $v_H \in V_H$

$$\begin{aligned} \|u - u_H\|_{\mathcal{V}} &= \|(\mathtt{id} - \mathcal{G})u\|_{\mathcal{V}} = \|(\mathtt{id} - \mathcal{G})(u - v_H)\|_{\mathcal{V}} \\ &\leq \|\mathtt{id} - \mathcal{G}\|_{\mathcal{L}(\mathcal{V},\mathcal{V})} \, \|u - v_H\|_{\mathcal{V}} = \|\mathcal{G}\|_{\mathcal{L}(\mathcal{V},\mathcal{V})} \, \|u - v_H\|_{\mathcal{V}} \end{aligned} \tag{2.8}$$

employing that $\|\mathtt{id} - \mathcal{G}\|_{\mathcal{L}(\mathcal{V},\mathcal{V})} = \|\mathcal{G}\|_{\mathcal{L}(\mathcal{V},\mathcal{V})}$ (see, e.g., [Szy06]). Here, $\mathtt{id}$ denotes the identity operator. By (2.6), (2.7), and (2.2), we get that

$$\|\mathcal{G}v\|_{\mathcal{V}} \leq \alpha_0^{-1} \sup_{w_H \in V_H} \frac{\mathfrak{a}(\mathcal{G}v, w_H)}{\|w_H\|_{\mathcal{V}}} = \alpha_0^{-1} \sup_{w_H \in V_H} \frac{\mathfrak{a}(v, w_H)}{\|w_H\|_{\mathcal{V}}} \leq \frac{\beta}{\alpha_0} \|v\|_{\mathcal{V}}. \tag{2.9}$$

Combining (2.8) and (2.9), we obtain

$$\|u - u_H\|_{\mathcal{V}} \leq \frac{\beta}{\alpha_0} \|u - v_H\|_{\mathcal{V}}.$$

Taking the infimum over all $v_H \in V_H$ yields the assertion. □

Céa's Lemma allows us to reduce the problem of writing down an error estimate for the cG solution to the problem of finding any discrete function in V_H that is able to suitably approximate the function u. Thus, from classical interpolation results (see, e.g., [BS08, Thm. 4.6.14]) we may obtain an error estimate of order H if the solution u fulfills additional regularity assumptions, which typically requires additional regularity of the right-hand side $\mathcal{F}$. These estimates are optimal in cases where the bilinear form $\mathfrak{a}$ does not include any multiscale behavior in the sense of, e.g., a dependence of $\mathfrak{a}$, and thus u, on a fine-scale parameter ϵ.

In the setting where the bilinear form $\mathfrak{a}$ depends on such a parameter ϵ and the function u has microscopic features on the scale ϵ, the standard FE space V_H is not able to provide a convenient discrete function that approximates the solution u for $\epsilon < H$ satisfactorily. In terms of explicit error estimates, this means that error estimates of the form

$$\|u - u_H\|_{\mathcal{V}} \leq C_{\epsilon,\mathcal{F}} H^s \tag{2.10}$$

for some $s > 0$ involve a multiplicative constant $C_{\epsilon,\mathcal{F}}$ that blows up when ϵ tends to zero. This especially means that H needs to resolve the microscopic scale, i.e., $H \lesssim \epsilon$, in order to obtain a viable estimate. In practical computations, one observes a stagnation of the error curve in the regime $H \gtrsim \epsilon$, and only in the case $H \lesssim \epsilon$ the expected convergence rate is obtained; see also Figure 2.1 in Section 2.5.2 for an illustration of this behavior. This observation is known as *pre-asymptotic effect* and calls for a thorough treatment of the fine-scale features of the bilinear form $\mathfrak{a}$. In the following section, we present the construction of a multiscale method that is able to achieve ϵ-independent error estimates. In particular, the results are also valid if a characterization of the fine scale in terms of an explicit parameter ϵ is not available.

2.2.2 General construction by orthogonal decomposition

As already mentioned in the previous subsection, the classical cG solution $u_H \in V_H$ of (2.5) fails to produce an acceptable approximation of the solution $u \in \mathcal{V}$ of (2.1) in the $\mathcal{V}$-norm if the discretization parameter H does not resolve the microscopic scale. While a similar statement is still true if the error is measured in the weaker $\mathcal{H}$-norm, there actually exist functions in V_H that are able to satisfactorily approximate u with respect to the $\mathcal{H}$-norm. One may think of such a function as one that approximates u in a macroscopic sense, since the effect of microscopic oscillations mainly appears in the stronger $\mathcal{V}$-norm. The first goal of the following construction is to find such a macroscopic representation in the space $\mathcal{H}$, which is then further adjusted to also obtain optimal error rates in the $\mathcal{V}$-norm.

The construction is built upon a linear, local, and projective *quasi-interpolation operator* $\mathcal{I}_H$, i.e., a linear projection $\mathcal{I}_H \colon \mathcal{H} \to V_H$ which fulfills suitable stability and approximation properties. To be more precise, we assume that for $v \in \mathcal{H}$, it holds that

$$\|\mathcal{I}_H v\|_{\mathcal{H}} \leq C_{\mathcal{I}_H} \|v\|_{\mathcal{H}}, \tag{2.11}$$

and, for $v \in \mathcal{V}$ and any element $K \in \mathcal{T}_H$,

$$\|H^{-1}(v - \mathcal{I}_H v)\|_{\mathcal{H}(K)} + \|\mathcal{I}_H v\|_{\mathcal{V}(K)} \leq C_{\mathcal{I}_H} \|v\|_{\mathcal{V}(\mathsf{N}(K))}, \tag{2.12}$$

where $\mathsf{N}(S)$, for any $S \subseteq D$, denotes the *element patch* around S defined by

$$\mathsf{N}(S) := \bigcup \big\{ K \in \mathcal{T}_H : \overline{K} \cap \overline{S} \neq \emptyset \big\}.$$

For later use, we also define for $\ell \in \mathbb{N}_0$ the *element patch of order* ℓ (or ℓ-*neighborhood*) around S by

$$\begin{aligned} \mathsf{N}^\ell(S) &:= \mathsf{N}(\mathsf{N}^{\ell-1}(S)), \quad \ell \geq 1, \\ \mathsf{N}^0(S) &:= \bigcup \big\{ K \in \mathcal{T}_H : S \cap \overline{K} \subseteq \overline{K} \big\}. \end{aligned} \tag{2.13}$$

Due to the locality in (2.12), a global result of the form

$$\|H^{-1}(v - \mathcal{I}_H v)\|_{\mathcal{H}} + \|\mathcal{I}_H v\|_{\mathcal{V}} \leq C_{\mathcal{I}_H} \|v\|_{\mathcal{V}} \tag{2.14}$$

for any $v \in \mathcal{V}$ directly follows from summation over all elements in $\mathcal{T}_H$. Note that the constants in (2.11), (2.12), and (2.14) are not necessarily identical. For simplicity, we use the constant $C_{\mathcal{I}_H}$ whenever one of the three estimates is employed.

The projection property of the operator $\mathcal{I}_H$ leads to a unique decomposition of a function $v \in \mathcal{V}$ into its *finite element part* $\mathcal{I}_H v \in V_H$ and its *fine-scale part* $v - \mathcal{I}_H v$, i.e., the space $\mathcal{V}$ can be decomposed as

$$\mathcal{V} = V_H \oplus \mathcal{W}$$

with the so-called *fine-scale space* $\mathcal{W}$ defined by

$$\mathcal{W} := (\mathtt{id} - \mathcal{I}_H)\mathcal{V} = \ker \mathcal{I}_H|_{\mathcal{V}}.$$

Regarding a suitable approximation of the solution $u \in \mathcal{V}$ of (2.1) in the space V_H with respect to the $\mathcal{H}$-norm, the finite element part $\mathcal{I}_H u \in V_H$ seems to be a good candidate. Indeed, with (2.11) and the projection property, it directly follows that $\mathcal{I}_H u$ is quasi-optimal with respect to the $\mathcal{H}$-norm. To be more precise, as in Lemma 2.2.1, it holds for $v_H \in V_H$ that

$$\|u - \mathcal{I}_H u\|_{\mathcal{H}} = \|(\mathtt{id} - \mathcal{I}_H)(u - v_H)\|_{\mathcal{H}} \leq C_{\mathcal{I}_H} \|u - v_H\|_{\mathcal{H}}$$

and thus

$$\|u - \mathcal{I}_H u\|_{\mathcal{H}} \leq C_{\mathcal{I}_H} \inf_{v_H \in V_H} \|u - v_H\|_{\mathcal{H}}. \tag{2.15}$$

Further, from (2.14) and (2.4) we get the error bound

$$\|u - \mathcal{I}_H u\|_{\mathcal{H}} \leq C_{\mathcal{I}_H} H \, \|u\|_{\mathcal{V}} \leq \alpha^{-1} C_{\mathcal{I}_H} H \, \|\mathcal{F}\|_{\mathcal{V}^*}. \tag{2.16}$$

Although the existence of an appropriate macroscopic approximation in V_H becomes evident from the above inequality, it remains unclear how to obtain

$\mathcal{I}_H u$ if the solution u is not known a priori. To overcome this issue, we first note that for any $v \in \mathcal{V}$, it holds that

$$\mathfrak{a}(\mathcal{I}_H u, v) = \mathfrak{a}(u, v) - \mathfrak{a}((\mathtt{id} - \mathcal{I}_H)u, v) = \mathcal{F}(v) - \mathfrak{a}((\mathtt{id} - \mathcal{I}_H)u, v).$$

This especially means that $\mathcal{I}_H u$ is a solution of the continuous Petrov-Galerkin (cPG) formulation which seeks $\bar{u}_H \in V_H$ that solves

$$\mathfrak{a}(\bar{u}_H, \tilde{v}_H) = \mathcal{F}(\tilde{v}_H) \tag{2.17}$$

for all $\tilde{v}_H \in \tilde{V}_H$, where the test space is defined by

$$\tilde{V}_H := \{v \in \mathcal{V} : \ \forall w \in \mathcal{W} : \ \mathfrak{a}(w, v) = 0\}. \tag{2.18}$$

This result is the basis of the original LOD and known from the Variational Multiscale Method, see e.g. [HS07]. It even holds for more general bounded linear projection operators $\mathcal{I}_H$. The *ideal test space* $\tilde{V}_H$ comes along with the alternative decomposition

$$\mathcal{V} = \tilde{V}_H \oplus \mathcal{W}$$

which satisfies the orthogonality property

$$\mathfrak{a}(\mathcal{W}, \tilde{V}_H) = 0. \tag{2.19}$$

Note that, with the aforementioned assumptions, this construction does not automatically provide the uniqueness of the cPG solution $\bar{u}_H$ in (2.17). In order to obtain uniqueness, a condition of the form

$$\dim V_H = \dim \tilde{V}_H \tag{2.20}$$

must hold. The next subsection is concerned with an explicit construction of the test space $\tilde{V}_H$ which guarantees that condition (2.20) is fulfilled.

2.2.3 Characterization of the ideal test space

In order to show that the dimensions of V_H and $\tilde{V}_H$ are equal, we derive a characterization of the space $\tilde{V}_H$ in terms of V_H. In this subsection, we explicitly construct a bijective operator $\mathcal{R}^* \colon V_H \to \tilde{V}_H$ that quantifies the connection between the two spaces. In the Petrov-Galerkin setting, such an operator is usually referred to as the *trial-to-test operator.*

We start the construction by introducing a *correction operator* $\mathcal{C}^* \colon \mathcal{V} \to \mathcal{W}$ defined for any $v \in \mathcal{V}$ by

$$\mathfrak{a}(w, \mathcal{C}^* v) = \mathfrak{a}(w, v) \tag{2.21}$$

for all $w \in \mathcal{W}$. Note that the well-posedness of (2.21) does not follow automatically and requires the inf-sup condition

$$\alpha_{\mathcal{W}} \leq \inf_{v \in \mathcal{W}} \sup_{w \in \mathcal{W}} \frac{\mathfrak{a}(v, w)}{\|v\|_{\mathcal{V}} \, \|w\|_{\mathcal{V}}} = \inf_{w \in \mathcal{W}} \sup_{v \in \mathcal{W}} \frac{\mathfrak{a}(v, w)}{\|v\|_{\mathcal{V}} \, \|w\|_{\mathcal{V}}} \tag{2.22}$$

with some constant $\alpha_{\mathcal{W}} > 0$.

The correction operator $\mathcal{C}^*$ provides an alternative characterization of the space $\tilde{V}_H$ since the direct consequence that

$$\mathfrak{a}(w, (\mathtt{id} - \mathcal{C}^*)v) = 0$$

for all $w \in \mathcal{W}$ is exactly the condition in the definition of the space $\tilde{V}_H$ in (2.18). This yields

$$\tilde{V}_H = (\mathtt{id} - \mathcal{C}^*)\mathcal{V} = (\mathtt{id} - \mathcal{C}^*)(\mathcal{I}_H\mathcal{V} + (\mathtt{id} - \mathcal{I}_H)\mathcal{V}) = (\mathtt{id} - \mathcal{C}^*)V_H$$

since $(\mathtt{id} - \mathcal{I}_H)\mathcal{V} = \mathcal{W}$ and $(\mathtt{id} - \mathcal{C}^*)\mathcal{W} = \{0\}$. Thus,

$$\mathcal{R}^* := (\mathtt{id} - \mathcal{C}^*)|_{V_H} \colon V_H \to \tilde{V}_H \tag{2.23}$$

defines a bijective operator from V_H to $\tilde{V}_H$ with inverse $\mathcal{I}_H|_{\tilde{V}_H} \colon \tilde{V}_H \to V_H$. Due to this explicit characterization, we also use the alternative notation

$$\mathcal{R}^* V_H = \tilde{V}_H$$

in the following. This also means that given a basis $\mathfrak{B}$ of V_H, we directly get a basis of $\tilde{V}_H$ by $\tilde{\mathfrak{B}} = \mathcal{R}^*\mathfrak{B}$.

Finally, we remark that condition (2.20) and thus the well-posedness of the cPG problem (2.17) follow from the inf-sup condition (2.22), which is required for the above construction. That is, we ultimately need an additional inf-sup condition to obtain existence and uniqueness of the finite-dimensional problem (2.17), as for the classical FE approach (2.5).

2.3 Fine-scale correction of the discrete trial space

In this section, we extend the method of the previous section in order to obtain a good approximation of the solution u of (2.1) not only with respect to the $\mathcal{H}$-norm but also with respect to the stronger $\mathcal{V}$-norm. This is achieved by adding a specific function from the fine-scale space $\mathcal{W}$ to the solution $\bar{u}_H \in V_H$ of the cPG problem (2.17), which is discussed in the next subsection.

2.3.1 Ideal trial space

The starting point of the approach is the observation that, due to the orthogonality property (2.19), it holds for any $w \in \mathcal{W}$ that

$$\mathfrak{a}(\bar{u}_H + w, \tilde{v}_H) = \mathfrak{a}(\bar{u}_H, \tilde{v}_H) = \mathcal{F}(\tilde{v}_H) \tag{2.24}$$

for all $\tilde{v}_H \in \mathcal{R}^*V_H$. As for the test space, the idea is thus to connect the trial space V_H to an appropriate subspace of $\mathcal{V}$ with the same dimension by subtracting suitable fine-scale corrections. The goal is to replace the trial space V_H in (2.17) by a new space without losing the well-posedness of the discrete problem. This can be achieved exploiting the property (2.24).

Similarly to (2.21), we define the *correction operator* $\mathcal{C}\colon \mathcal{V} \to \mathcal{W}$ by

$$\mathfrak{a}(\mathcal{C}v, w) = \mathfrak{a}(v, w) \tag{2.25}$$

for all $w \in \mathcal{W}$ and define

$$\mathcal{R} := (\mathtt{id} - \mathcal{C})|_{V_H} \colon V_H \to \mathcal{R}V_H \tag{2.26}$$

with inverse $\mathcal{I}_H|_{\mathcal{R}V_H} : \mathcal{R}V_H \to V_H$. Note that the well-posedness of (2.25) follows from the inf-sup condition (2.22). Further, observe that by (2.25) and (2.26) we also have

$$\mathcal{R}V_H = \{v \in \mathcal{V} :\ \forall w \in \mathcal{W} :\ \mathfrak{a}(v, w) = 0\}. \tag{2.27}$$

The introduction of the operator $\mathcal{R}$ now provides an equivalent formulation of problem (2.17) in terms of the *ideal trial space* $\mathcal{R}V_H$ and the test space $\mathcal{R}^*V_H$: find $\tilde{u}_H \in \mathcal{R}V_H$ such that

$$\mathfrak{a}(\tilde{u}_H, \tilde{v}_H) = \mathcal{F}(\tilde{v}_H) \tag{2.28}$$

for all $\tilde{v}_H \in \mathcal{R}^*V_H$. We call (2.28) the *ideal method* and refer to $\tilde{u}_H$ as the *ideal approximation.* A direct consequence of (2.24) is that $\mathcal{I}_H\tilde{u}_H = \bar{u}_H = \mathcal{I}_H u$. The subsequent theorem shows that the solution $\tilde{u}_H \in \mathcal{R}V_H$ provides quasi-optimal error estimates in the $\mathcal{V}$-norm and the $\mathcal{H}$-norm under additional regularity assumptions on the functional $\mathcal{F}$. To be more precise, we assume that

$$\mathcal{F}(v) = (f, v)_{\mathcal{H}} \tag{2.29}$$

for some function $f \in \mathcal{H}$, where $(\cdot,\cdot)_{\mathcal{H}}$ denotes the scalar product in $\mathcal{H}$.

Theorem 2.3.1 (Error of the ideal method)**.** *Suppose that the inf-sup conditions* (2.3) *and* (2.22) *hold, $\mathcal{F}$ fulfills the regularity condition* (2.29)*, and $\mathfrak{a}$ is bounded according to* (2.2)*. Then the solution $u \in \mathcal{V}$ of* (2.1) *and the ideal approximation $\tilde{u}_H \in \mathcal{R}V_H$ of* (2.28) *satisfy the error estimates*

$$\|u - \tilde{u}_H\|_{\mathcal{V}} \le \alpha_{\mathcal{W}}^{-1} C_{\mathcal{I}_H} H\, \|f\|_{\mathcal{H}} \tag{2.30}$$

and

$$\|u - \tilde{u}_H\|_{\mathcal{H}} \le \alpha_{\mathcal{W}}^{-1} C_{\mathcal{I}_H}^2 H^2\, \|f\|_{\mathcal{H}} \tag{2.31}$$

independently of possible oscillations of coefficients encoded in $\mathfrak{a}$ on some microscopic scale ϵ.

Proof. By construction, we have that $\mathcal{I}_H(u - \tilde{u}_H) = 0$ and thus $u - \tilde{u}_H \in \mathcal{W}$. We can even show that the error between the two functions is exactly the correction of u, i.e.,

$$u - \tilde{u}_H = u - (\mathtt{id} - \mathcal{C})\mathcal{I}_H u = u - (\mathtt{id} - \mathcal{C})u + (\mathtt{id} - \mathcal{C})(\mathtt{id} - \mathcal{I}_H)u = \mathcal{C}u,$$

using the fact that $\mathcal{C}$ defines a projection onto $\mathcal{W}$. From the inf-sup condition (2.22), we further get the existence of a function $w \in \mathcal{W}$ with $\|w\|_{\mathcal{V}} = 1$ such that

$$\begin{aligned}\|u - \tilde{u}_H\|_{\mathcal{V}} = \|\mathcal{C}u\|_{\mathcal{V}} &\leq \alpha_{\mathcal{W}}^{-1}\, \mathfrak{a}(\mathcal{C}u, w) = \alpha_{\mathcal{W}}^{-1}\, \mathfrak{a}(u, w) = \alpha_{\mathcal{W}}^{-1}\, (f, w)_{\mathcal{H}}\\ &\leq \alpha_{\mathcal{W}}^{-1}\, \|f\|_{\mathcal{H}}\, \|w\|_{\mathcal{H}} = \alpha_{\mathcal{W}}^{-1}\, \|f\|_{\mathcal{H}}\, \|(\mathtt{id} - \mathcal{I}_H)w\|_{\mathcal{H}}\\ &\leq \alpha_{\mathcal{W}}^{-1} C_{\mathcal{I}_H} H\, \|f\|_{\mathcal{H}}\end{aligned}$$

with the constant $C_{\mathcal{I}_H}$ from (2.14). This proves (2.30). Again exploiting the fact that $\mathcal{C}u \in \mathcal{W}$, we directly get

$$\|u - \tilde{u}_H\|_{\mathcal{H}} = \|\mathcal{C}u\|_{\mathcal{H}} = \|(\mathtt{id} - \mathcal{I}_H)\mathcal{C}u\|_{\mathcal{H}} \leq C_{\mathcal{I}_H} H\, \|\mathcal{C}u\|_{\mathcal{V}}$$

and thus

$$\|u - \tilde{u}_H\|_{\mathcal{H}} \leq \alpha_{\mathcal{W}}^{-1} C_{\mathcal{I}_H}^2 H^2\, \|f\|_{\mathcal{H}}.$$

This completes the proof. □

Theorem 2.3.1 shows that the function $\tilde{u}_H$, which is obtained as the solution of the finite-dimensional problem (2.28), is a suitable approximation of the solution u of (2.1). However, (2.28) does not provide a practicable method because the spaces $\mathcal{R}V_H$ and $\mathcal{R}^*V_H$ are constructed by solving the infinite-dimensional corrector problems (2.25) and (2.21). Before we address this issue in Section 2.4, we first show in the subsequent subsection how problem (2.28) can be reinterpreted as a variational problem in the full space $\mathcal{V}$ subject to a finite number of constraints.

2.3.2 Reformulation as saddle point problem

The following results provide a useful alternative characterization of the multiscale spaces $\mathcal{R}V_H$ and $\mathcal{R}^*V_H$, which allows us to circumvent an explicit computation of the fine-scale space $\mathcal{W}$. Moreover, the alternative representation creates a basis for an extension of the method to a higher-order method as it is introduced in Chapter 3.

Theorem 2.3.2 (Alternative characterization of $\mathcal{R}$ and $\mathcal{R}^*$)**.** *Assume that the inf-sup condition* (2.22) *holds and let $\mathcal{R}$ and $\mathcal{R}^*$ be the operators defined in* (2.26) *and* (2.23)*, respectively. Then, for any $v_H \in V_H$, the function $\mathcal{R}v_H \in \mathcal{V}$ solves the saddle point problem*

$$\begin{aligned}\mathfrak{a}(\mathcal{R}v_H, w) \quad + \quad (\lambda_{v_H}, \mathcal{I}_H w)_{\mathcal{H}} \quad &= \quad 0,\\ (\mathcal{I}_H \mathcal{R}v_H, \mu_H)_{\mathcal{H}} \quad &= \quad (v_H, \mu_H)_{\mathcal{H}}\end{aligned} \tag{2.32}$$

for all $w \in \mathcal{V}$ and all $\mu_H \in V_H$, where $\lambda_{v_H} \in V_H$ is the associated Lagrange multiplier. Likewise, $\mathcal{R}^*v_H \in \mathcal{V}$ solves

$$\begin{aligned}\mathfrak{a}(w, \mathcal{R}^*v_H) \quad + \quad (\mathcal{I}_H w, \lambda_{v_H}^*)_{\mathcal{H}} \quad &= \quad 0,\\ (\mu_H, \mathcal{I}_H \mathcal{R}^*v_H)_{\mathcal{H}} \quad &= \quad (\mu_H, v_H)_{\mathcal{H}}\end{aligned} \tag{2.33}$$

for all $w \in \mathcal{V}$ and all $\mu_H \in V_H$, where $\lambda^*_{v_H} \in V_H$ is the corresponding Lagrange multiplier. Further, the solutions $(\mathcal{R}v_H, \lambda_{v_H})$ of (2.32) *and $(\mathcal{R}^*v_H, \lambda^*_{v_H})$ of* (2.33) *are unique.*

Proof. Let $v_H \in V_H$ and define $\tilde{v}_H = \mathcal{R}v_H = (\mathtt{id} - \mathcal{C})v_H \in \mathcal{R}V_H$. Further, let $\lambda_{v_H} \in V_H$ be the solution of the auxiliary problem

$$(\lambda_{v_H}, w_H)_\mathcal{H} = -\mathfrak{a}(v_H, \mathcal{R}^* w_H) \tag{2.34}$$

for all $w_H \in V_H$. Note that (2.34) has a unique solution by the *Lax-Milgram Theorem* (see, e.g., [BS08, Thm. 2.7.7]) and the *inverse inequality*

$$\|w_H\|_\mathcal{V} \leq C_{\text{inv}} H^{-1} \|w_H\|_\mathcal{H} \tag{2.35}$$

for $w_H \in V_H$ (see, e.g., [Sch98, GHS05, Geo08]). Thus, using (2.21), (2.25), and the auxiliary problem (2.34), we get

$$\begin{aligned}\mathfrak{a}(\tilde{v}_H, w) &= \mathfrak{a}(\tilde{v}_H, \mathcal{I}_H w) + \mathfrak{a}(\tilde{v}_H, (\mathtt{id} - \mathcal{I}_H)w)\\ &= \mathfrak{a}(\tilde{v}_H, \mathcal{I}_H w) = \mathfrak{a}(v_H, \mathcal{R}^*\mathcal{I}_H w)\\ &= -(\lambda_{v_H}, \mathcal{I}_H w)_\mathcal{H}\end{aligned}$$

for any $w \in \mathcal{V}$. Since

$$\mathcal{I}_H \tilde{v}_H = \mathcal{I}_H(\mathtt{id} - \mathcal{C})v_H = v_H,$$

the pair $(\mathcal{R}v_H, \lambda_{v_H})$ solves (2.32). From classical saddle point theory and with the inf-sup condition (2.22), it follows that the solution of (2.32) is also unique (see, e.g., [BBF13, Thm. 4.2.3]). Introducing $\lambda^*_{v_H} \in V_H$ as the unique solution of

$$(w_H, \lambda^*_{v_H})_\mathcal{H} = -\mathfrak{a}(\mathcal{R}w_H, v_H)$$

for all $w_H \in V_H$, it follows with the same arguments as above that $(\mathcal{R}^*v_H, \lambda^*_{v_H})$ is the unique solution of (2.33). □

As a direct consequence of Theorem 2.3.2, the spaces $\mathcal{R}V_H$ and $\mathcal{R}^*V_H$ may be obtained without explicitly defining the fine-scale space $\mathcal{W}$. Besides, the reformulation also provides an alternative characterization of the solution $\tilde{u}_H$ of (2.28). This result is stated as a corollary.

Corollary 2.3.3 (Equivalent saddle point formulation)**.** *Assume that* (2.3)*,* (2.22)*, and* (2.2) *hold. Then the solution $\tilde{u}_H \in \mathcal{V}$ of* (2.28) *can be equivalently described as the solution of the saddle point formulation*

$$\begin{array}{rcrcl}\mathfrak{a}(\tilde{u}_H, w) & + & (\tilde{\lambda}_H, \mathcal{I}_H w)_\mathcal{H} & = & 0,\\ (\mathcal{I}_H \tilde{u}_H, \mu_H)_\mathcal{H} & & & = & (\mathcal{I}_H u, \mu_H)_\mathcal{H}\end{array} \tag{2.36}$$

for all $w \in \mathcal{V}$ and all $\mu_H \in V_H$, where $\tilde{\lambda}_H \in V_H$ is a uniquely defined Lagrange multiplier and $u \in \mathcal{V}$ is the solution of (2.1)*.*

Remark 2.3.4. If the bilinear form $\mathfrak{a}$ is symmetric, the saddle point problems (2.32) and (2.33) are equivalent, i.e., it holds that $\mathcal{R} = \mathcal{R}^*$. Further, one may reformulate these problems as a constrained energy minimization problem. That is,

$$\mathcal{R}v_H := \operatorname*{arg\,min}_{v \in \mathcal{V}} \mathfrak{a}(v,v) \quad \text{subject to} \quad \mathcal{I}_H v = v_H \tag{2.37}$$

for any $v_H \in V_H$. The fact that energy-minimizing functions that fulfill a finite number of constraints present suitable trial and test spaces in the context of multiscale problems is also the basis of the technique described in [Owh15, Owh17] based on gamblets.

2.4 Fully discrete approximation

As already mentioned in the previous sections, the ideal approximation $\tilde{u}_H$ in (2.28) is not a practicable discrete approximation in the sense that its computation involves the solution of infinite-dimensional global problems. In this section, we address this issue and present a strategy to derive a fully discrete multiscale approach. The procedure consists of three main steps that are treated in the next subsections: the splitting, localization, and discretization of the correction operators $\mathcal{C} \colon \mathcal{V} \to \mathcal{W}$ and $\mathcal{C}^* \colon \mathcal{V} \to \mathcal{W}$ introduced in (2.25) and (2.21), respectively. Since the strategies for $\mathcal{C}$ and $\mathcal{C}^*$ follow the same arguments, we only consider the operator $\mathcal{C}$.

2.4.1 Splitting of the correction operator

The first step towards a fully discrete method consists in splitting the restricted correction operator $\mathcal{C}|_{V_H}$ into its contributions of a (local) basis. Since a splitting in terms of conforming FE basis functions leads to a pollution of the error estimate in terms of a negative power of H (see [MP14]), we follow the approach of [HP13] and further decompose these basis functions into its discontinuous element-wise contributions. This alternative strategy turns out to enable much better decay estimates.

To this end, we define for $K \in \mathcal{T}_H$ the nodal basis of $\mathcal{V}(K)$ by $\{\Lambda_{K,j}\}_{j=1}^{m_K}$, where m_K is the number of vertices of the element K. We remark that any function $v_H \in V_H$ can be written as

$$v_H = \sum_{K \in \mathcal{T}_H} \sum_{j=1}^{m_K} v_H(x_{K,j})\, \Lambda_{K,j},$$

where $\{x_{K,j}\}_{j=1}^{m_K}$ are the vertices of $K \in \mathcal{T}_H$. Based on the above characterization, we define for $K \in \mathcal{T}_H$ and $j \in \{1, \dots, m_K\}$ the jth *basis corrector* $q_{K,j} \in \mathcal{W}$ by

$$\mathfrak{a}(q_{K,j}, w) = \mathfrak{a}(\Lambda_{K,j}, w) \tag{2.38}$$

for all $w \in \mathcal{W}$. At this point, we have to assume that $\mathfrak{a}$ is even well-defined on the restricted spaces $\mathcal{V}(S_1) \times \mathcal{V}(S_2)$ for subdomains S_1, $S_2 \subseteq D$ to justify the right-hand side of (2.38). In that context, we also suppose that the boundedness of $\mathfrak{a}$ in (2.2) holds in a more local sense, i.e., we suppose that for $v \in \mathcal{V}(S_1)$ and $w \in \mathcal{V}(S_2)$

$$|\mathfrak{a}(v|_{S_1}, w|_{S_2})| \leq \beta_{\mathrm{loc}} \, \|v\|_{\mathcal{V}(S)} \, \|w\|_{\mathcal{V}(S)} \tag{2.39}$$

with $S = S_1 \cap S_2$. Although this additional assumption seems restrictive at first glance, such estimates are natural in the context of variational formulations of linear second-order PDEs, which are typically defined by integrals.

From (2.38) and the linearity of $\mathfrak{a}$ with respect to the first argument, we now get that

$$\mathcal{C} v_H = \sum_{K \in \mathcal{T}_H} \sum_{j=1}^{m_K} v_H(x_{K,j}) \, q_{K,j} \tag{2.40}$$

for any $v_H \in V_H$. Note that the functions $q_{K,j}$ in general have global support, even though the right-hand side of (2.38) is restricted to the element K. Thus, the splitting of the operator $\mathcal{C}|_{V_H}$ in (2.40) does not lead to localized contributions. However, it is very valuable for the localization procedure, which is discussed in the subsequent subsections.

2.4.2 Decay of the basis correctors

This subsection is devoted to proving that the basis correctors $q_{K,j}$, defined in (2.38), decay exponentially fast away from the support of the associated element K. This observation is the key property to deriving a fully discrete method and allows us to localize the computation of all the correctors (see Section 2.4.3).

In the general setting of this chapter, we need to assume that the inf-sup condition (2.22) holds in a more generalized form, i.e., we assume that there exists a constant $\alpha_{\mathcal{W},\mathrm{dec}} > 0$ such that

$$\alpha_{\mathcal{W},\mathrm{dec}} \leq \inf_{v \in \mathcal{W}^{\mathrm{c}}_{\ell,K}} \sup_{w \in \mathcal{W}^{\mathrm{c}}_{\ell,K}} \frac{\mathfrak{a}(v,w)}{\|v\|_{\mathcal{V}} \, \|w\|_{\mathcal{V}}} = \inf_{w \in \mathcal{W}^{\mathrm{c}}_{\ell,K}} \sup_{v \in \mathcal{W}^{\mathrm{c}}_{\ell,K}} \frac{\mathfrak{a}(v,w)}{\|v\|_{\mathcal{V}} \, \|w\|_{\mathcal{V}}} \tag{2.41}$$

for any $K \in \mathcal{T}_H$ and $\ell \in \mathbb{N}$, where

$$\mathcal{W}^{\mathrm{c}}_{\ell,K} := \big\{ w \in \mathcal{W} : \, \mathrm{supp}(w) \subseteq D \setminus \mathsf{N}^{\ell}(K) \big\}.$$

Note that here and in the following we implicitly assume that ℓ is small enough such that the space $\mathcal{W}^{\mathrm{c}}_{\ell,K}$ is non-empty. We emphasize, however, that the subsequent results trivially hold if ℓ is such that $\mathcal{W}^{\mathrm{c}}_{\ell,K} = \emptyset$.

Theorem 2.4.1 (Decay of the basis correctors). *Assume that the inf-sup condition* (2.41) *and the local boundedness condition* (2.39) *are fulfilled. Further,* *let $K \in \mathcal{T}_H$, $j \in \{1,\dots,m_K\}$, $\ell \in \mathbb{N}$, and $q_{K,j} \in \mathcal{W}$ be the solution of* (2.38). *Then it holds that*

$$\|q_{K,j}\|_{\mathcal{V}(D\setminus\mathsf{N}^\ell(K))} \lesssim \exp(-C_{\mathrm{dec}}\,\ell)\,\|q_{K,j}\|_{\mathcal{V}} \tag{2.42}$$

with a constant C_{dec} that depends on $C_{\mathcal{I}_H}$, $\alpha_{\mathcal{W},\mathrm{dec}}$, and β_{loc}.

Proof. We abbreviate $q := q_{K,j} \in \mathcal{W}$ and $\Lambda := \Lambda_{K,j} \in \mathcal{V}(K)$. For fixed $\ell \in \mathbb{N}$, we choose a cutoff function $\eta \in W^{1,\infty}(D)$ with the following properties:

$$\begin{aligned}
&0 \le \eta \le 1,\\
&\eta \;=\; 0 \quad \text{in } \mathsf{N}^{\ell+2}(K),\\
&\eta \;=\; 1 \quad \text{in } D\setminus\mathsf{N}^{\ell+3}(K),\\
&\|\nabla\eta\|_{L^\infty(D)} \;\le\; C_\eta\,H^{-1}.
\end{aligned}$$

Then, by (2.41) there exists a function $w \in \mathcal{W}^{\mathrm{c}}_{\ell+1,K}$ with $\|w\|_{\mathcal{V}} = 1$ such that

$$\begin{aligned}
\|q\|_{\mathcal{V}(D\setminus\mathsf{N}^{\ell+4}(K))} &= \|(\mathtt{id}-\mathcal{I}_H)q\|_{\mathcal{V}(D\setminus\mathsf{N}^{\ell+4}(K))} \le \|(\mathtt{id}-\mathcal{I}_H)(\eta q)\|_{\mathcal{V}}\\
&\le \alpha_{\mathcal{W},\mathrm{dec}}^{-1}\,\mathfrak{a}((\mathtt{id}-\mathcal{I}_H)(\eta q), w)\\
&= \alpha_{\mathcal{W},\mathrm{dec}}^{-1}\,\big(\mathfrak{a}((\mathtt{id}-\mathcal{I}_H)q, w) - \mathfrak{a}((\mathtt{id}-\mathcal{I}_H)((1-\eta)q), w)\big)\\
&= \alpha_{\mathcal{W},\mathrm{dec}}^{-1}\,\big(\mathfrak{a}(\Lambda, w) - \mathfrak{a}((\mathtt{id}-\mathcal{I}_H)((1-\eta)q), w)\big),
\end{aligned}$$

where we use the fact that $\mathcal{I}_H$ increases the support of a function by at most one layer of elements due to (2.12). Since $\operatorname{supp}(\Lambda)\cap\operatorname{supp}(w) = \emptyset$ and

$$\operatorname{supp}((\mathtt{id}-\mathcal{I}_H)((1-\eta)q)) \cap \operatorname{supp}(w) = \mathsf{N}^{\ell+4}(K)\setminus\mathsf{N}^{\ell+1}(K),$$

we further get

$$\begin{aligned}
\|q\|_{\mathcal{V}(D\setminus\mathsf{N}^{\ell+4}(K))} &\le \alpha_{\mathcal{W},\mathrm{dec}}^{-1}\beta_{\mathrm{loc}}\,\|(\mathtt{id}-\mathcal{I}_H)((1-\eta)q)\|_{\mathcal{V}(\mathsf{N}^{\ell+4}(K)\setminus\mathsf{N}^{\ell+1}(K))}\\
&\lesssim \alpha_{\mathcal{W},\mathrm{dec}}^{-1}\beta_{\mathrm{loc}}C_{\mathcal{I}_H}\,\|(1-\eta)q\|_{\mathcal{V}(\mathsf{N}^{\ell+4}(K)\setminus\mathsf{N}^{\ell}(K))}\\
&\lesssim \alpha_{\mathcal{W},\mathrm{dec}}^{-1}\beta_{\mathrm{loc}}C_{\mathcal{I}_H}^2C_\eta\,\|q\|_{\mathcal{V}(\mathsf{N}^{\ell+4}(K)\setminus\mathsf{N}^{\ell}(K))}
\end{aligned}$$

employing (2.12) and the product rule. From the above computations and with the identity

$$\mathsf{N}^{\ell+4}(K)\setminus\mathsf{N}^{\ell}(K) = \big(D\setminus\mathsf{N}^{\ell}(K)\big)\setminus\big(D\setminus\mathsf{N}^{\ell+4}(K)\big),$$

we obtain

$$\|q\|^2_{\mathcal{V}(D\setminus\mathsf{N}^{\ell+4}(K))} \le C\,\|q\|^2_{\mathcal{V}(D\setminus\mathsf{N}^{\ell}(K))} - C\,\|q\|^2_{\mathcal{V}(D\setminus\mathsf{N}^{\ell+4}(K))}$$

and thus

$$\|q\|^2_{\mathcal{V}(D\setminus\mathsf{N}^{\ell+4}(K))} \le \frac{C}{C+1}\,\|q\|^2_{\mathcal{V}(D\setminus\mathsf{N}^{\ell}(K))} \le \Big(\frac{C}{C+1}\Big)^{\lfloor\ell/4\rfloor}\|q\|^2_{\mathcal{V}}.$$

With the estimate

$$\Big(\frac{C}{C+1}\Big)^{\lfloor\ell/4\rfloor} \lesssim \exp\big(-\tfrac14\,\big|\log\big(\tfrac{C}{C+1}\big)\big|\,(\ell+4)\big)$$

and a shift in ℓ, this yields (2.42) with the constant $C_{\mathrm{dec}} := \frac18\,|\log(\frac{C}{C+1})|$. □

2.4.3 Localization of the correction operator

The exponential decay of the basis correctors allows us to localize (2.38) to patches around an element $K \in \mathcal{T}_H$. For $j \in \{1, \dots, m_K\}$ and $\ell \in \mathbb{N}$, we define the *localized basis corrector* $q^\ell_{K,j} \in \mathcal{W}_{\ell,K}$ by

$$\mathfrak{a}(q^\ell_{K,j}, w) = \mathfrak{a}(\Lambda_{K,j}, w) \tag{2.43}$$

for all $w \in \mathcal{W}_{\ell,K}$, where the local fine-scale space $\mathcal{W}_{\ell,K}$ is given by

$$\mathcal{W}_{\ell,K} := \big\{ w \in \mathcal{W} : \operatorname{supp}(w) \subseteq \mathsf{N}^\ell(K) \big\}\,.$$

Within the general setting of this chapter, we need to assume well-posedness of (2.43), i.e., we suppose that there exists a constant $\alpha_{\mathcal{W},\mathrm{loc}} > 0$ such that

$$\alpha_{\mathcal{W},\mathrm{loc}} \leq \inf_{v \in \mathcal{W}_{\ell,K}} \sup_{w \in \mathcal{W}_{\ell,K}} \frac{\mathfrak{a}(v,w)}{\|v\|_{\mathcal{V}}\,\|w\|_{\mathcal{V}}} = \inf_{w \in \mathcal{W}_{\ell,K}} \sup_{v \in \mathcal{W}_{\ell,K}} \frac{\mathfrak{a}(v,w)}{\|v\|_{\mathcal{V}}\,\|w\|_{\mathcal{V}}} \tag{2.44}$$

for any $K \in \mathcal{T}_H$ and $\ell \in \mathbb{N}$. In this subsection, we prove that the *localized correction operator* $\mathcal{C}^\ell \colon V_H \to \mathcal{W}$, defined for any $v_H \in V_H$ by

$$\mathcal{C}^\ell v_H := \sum_{K \in \mathcal{T}_H} \sum_{j=1}^{m_K} v_H(x_{K,j})\, q^\ell_{K,j}, \tag{2.45}$$

only introduces a moderate error if the so-called *localization* (or *oversampling*) *parameter* ℓ is chosen appropriately. As a first step, we quantify the error introduced by replacing one basis corrector by its localized counterpart.

Lemma 2.4.2. *Let $K \in \mathcal{T}_H$, $j \in \{1, \dots, m_K\}$, and $\ell \in \mathbb{N}$. Assume that the inf-sup conditions* (2.22)*,* (2.41)*, and* (2.44) *as well as the local boundedness condition* (2.39) *hold. Then the solutions $q_{K,j} \in \mathcal{W}$ of* (2.38) *and $q^\ell_{K,j} \in \mathcal{W}_{\ell,K}$ of* (2.43) *satisfy*

$$\|q_{K,j} - q^\ell_{K,j}\|_{\mathcal{V}} \lesssim \exp(-C_{\mathrm{dec}}\,\ell)\, \|q_{K,j}\|_{\mathcal{V}} \tag{2.46}$$

with the constant C_{dec} from Theorem 2.4.1.

Proof. We use the short-hand notation $q := q_{K,j} \in \mathcal{W}$, $q^\ell := q^\ell_{K,j} \in \mathcal{W}_{\ell,K}$, and $\Lambda := \Lambda_{K,j} \in \mathcal{V}(K)$. As in the proof of Theorem 2.4.1, we choose a cutoff function $\eta \in W^{1,\infty}(D)$ that fulfills

$$\begin{aligned} &0 \leq \eta \leq 1, \\ &\eta = 0 \quad \text{in } \mathsf{N}^{\ell+1}(K), \\ &\eta = 1 \quad \text{in } D \setminus \mathsf{N}^{\ell+2}(K), \\ &\|\nabla \eta\|_{L^\infty(D)} \leq C_\eta\, H^{-1}. \end{aligned}$$

By (2.44), we know that there exists a function $w \in \mathcal{W}_{\ell+3,K}$ such that

$$\begin{aligned}\|q - q^{\ell+3}\|_{\mathcal{V}} &\leq \|(\mathtt{id} - \mathcal{I}_H)(\eta q)\|_{\mathcal{V}} + \|(\mathtt{id} - \mathcal{I}_H)((1-\eta)q) - q^{\ell+3}\|_{\mathcal{V}} \\ &\lesssim C_{\mathcal{I}_H}^2 C_\eta \, \|q\|_{\mathcal{V}(D\setminus \mathsf{N}^\ell(K))} + \alpha_{\mathcal{W},\mathrm{loc}}^{-1}\, \mathfrak{a}((\mathtt{id} - \mathcal{I}_H)((1-\eta)q) - q^{\ell+3}, w) \\ &\lesssim C_{\mathcal{I}_H}^2 C_\eta \, \|q\|_{\mathcal{V}(D\setminus \mathsf{N}^\ell(K))} + \alpha_{\mathcal{W},\mathrm{loc}}^{-1}\, \mathfrak{a}((q - q^{\ell+3}) - (\mathtt{id} - \mathcal{I}_H)(\eta q), w) \\ &\lesssim C_{\mathcal{I}_H}^2 C_\eta \, \|q\|_{\mathcal{V}(D\setminus \mathsf{N}^\ell(K))} - \alpha_{\mathcal{W},\mathrm{loc}}^{-1}\, \mathfrak{a}((\mathtt{id} - \mathcal{I}_H)(\eta q), w) \\ &\lesssim C_{\mathcal{I}_H}^2 C_\eta \, \|q\|_{\mathcal{V}(D\setminus \mathsf{N}^\ell(K))} + \alpha_{\mathcal{W},\mathrm{loc}}^{-1} \beta C_{\mathcal{I}_H}^2 C_\eta \, \|q\|_{\mathcal{V}(D\setminus \mathsf{N}^\ell(K))} \\ &\lesssim (1 + \alpha_{\mathcal{W},\mathrm{loc}}^{-1}\beta) C_{\mathcal{I}_H}^2 C_\eta \exp(-C_{\mathrm{dec}}\, \ell)\, \|q\|_{\mathcal{V}},\end{aligned}$$

where we employ Theorem 2.4.1 in the last step. This proves the assertion. □

Remark 2.4.3. Although Theorem 2.4.1 and Lemma 2.4.2 only quantify the exponential decay and the localization error, respectively, of the basis correctors $q_{K,j}$, the results hold analogously for any function $v_K \in \mathcal{V}(K)$ and its corresponding correction q_K given by

$$q_K := \sum_{j=1}^{m_K} v_K(x_{K,j})\, q_{K,j}.$$

That is, we actually have

$$\begin{aligned}\|q_K\|_{\mathcal{V}(D\setminus \mathsf{N}^\ell(K))} &\lesssim \exp(-C_{\mathrm{dec}}\, \ell)\, \|q_K\|_{\mathcal{V}} \\ &\lesssim \exp(-C_{\mathrm{dec}}\, \ell)\, \alpha_{\mathcal{W}}^{-1} \beta_{\mathrm{loc}}\, \|v_K\|_{\mathcal{V}(K)}\end{aligned} \tag{2.47}$$

and

$$\begin{aligned}\Big\| \sum_{j=1}^{m_K} v_K(x_{K,j})\, (q_{K,j} - q_{K,j}^\ell) \Big\|_{\mathcal{V}} &\lesssim \exp(-C_{\mathrm{dec}}\, \ell)\, \|q_K\|_{\mathcal{V}} \\ &\lesssim \exp(-C_{\mathrm{dec}}\, \ell)\, \alpha_{\mathcal{W}}^{-1} \beta_{\mathrm{loc}}\, \|v_K\|_{\mathcal{V}(K)}\end{aligned} \tag{2.48}$$

using the upper bound on $\|q_K\|_{\mathcal{V}}$ which can be shown with (2.22) and (2.39).

With the above localization results, we are prepared to prove the main theorem of this subsection, which quantifies the error between the restricted operator $\mathcal{C}|_{V_H}$ and its localized version $\mathcal{C}^\ell$.

Theorem 2.4.4 (Localization error). *Let $\ell \in \mathbb{N}$. Suppose that the inf-sup conditions* (2.22), (2.41), *and* (2.44) *are satisfied and that $\mathfrak{a}$ fulfills the boundedness condition* (2.39). *Then, for any $v_H \in V_H$, the global localization error is bounded by*

$$\|(\mathcal{C} - \mathcal{C}^\ell) v_H\|_{\mathcal{V}} \lesssim \ell^{(d-1)/2} \exp(-C_{\mathrm{dec}}\, \ell)\, \|v_H\|_{\mathcal{V}}. \tag{2.49}$$

Proof. Let $v_H \in V_H$. As before, we can write

$$v_H = \sum_{K \in \mathcal{T}_H} \sum_{j=1}^{m_K} v_H(x_{K,j})\, \Lambda_{K,j}.$$

Further, we define for any $K \in \mathcal{T}_H$ a cutoff function $\eta_K \in W^{1,\infty}(D)$ which fulfills

$$\begin{aligned} &0 \leq \eta_K \leq 1, \\ &\eta_K \;=\; 0 \quad \text{in } \mathsf{N}^{\ell+1}(K), \\ &\eta_K \;=\; 1 \quad \text{in } D \setminus \mathsf{N}^{\ell+2}(K), \\ &\|\nabla \eta_K\|_{L^\infty(D)} \;\leq\; C_\eta \, H^{-1}. \end{aligned}$$

Since $(\mathcal{C} - \mathcal{C}^{\ell+3})v_H \in \mathcal{W}$, there exists a function $w \in \mathcal{W}$ with $\|w\|_{\mathcal{V}} = 1$ such that

$$\begin{aligned} \|(\mathcal{C} - \mathcal{C}^{\ell+3})v_H\|_{\mathcal{V}} &\leq \alpha_{\mathcal{W}}^{-1} \, \mathfrak{a}((\mathcal{C} - \mathcal{C}^{\ell+3})v_H, w) \\ &= \alpha_{\mathcal{W}}^{-1} \sum_{K \in \mathcal{T}_H} \mathfrak{a}\Big(\sum_{j=1}^{m_K} v_H(x_{K,j}) \, (q_{K,j} - q_{K,j}^{\ell+3}), w \Big). \end{aligned} \tag{2.50}$$

For any $K \in \mathcal{T}_H$, it holds that

$$\begin{aligned} &\mathfrak{a}\Big(\sum_{j=1}^{m_K} v_H(x_{K,j}) \, (q_{K,j} - q_{K,j}^{\ell+3}), w \Big) \\ &\quad = \mathfrak{a}\Big(\sum_{j=1}^{m_K} v_H(x_{K,j}) \, (q_{K,j} - q_{K,j}^{\ell+3}), (\mathtt{id} - \mathcal{I}_H)((1-\eta_K)w) + (\mathtt{id} - \mathcal{I}_H)(\eta_K w) \Big) \\ &\quad = \mathfrak{a}\Big(\sum_{j=1}^{m_K} v_H(x_{K,j}) \, \Lambda_{K,j}, (\mathtt{id} - \mathcal{I}_H)(\eta_K w) \Big) \\ &\qquad - \mathfrak{a}\Big(\sum_{j=1}^{m_K} v_H(x_{K,j}) \, q_{K,j}^{\ell+3}, (\mathtt{id} - \mathcal{I}_H)(\eta_K w) \Big), \end{aligned}$$

where we use (2.38), (2.43), and the fact that $(\mathtt{id} - \mathcal{I}_H)((1-\eta_K)w) \in \mathcal{W}_{\ell+3,K}$. Since $\operatorname{supp}((\mathtt{id} - \mathcal{I}_H)(\eta_K w)) \cap K = \emptyset$, we further get

$$\begin{aligned} &\mathfrak{a}\Big(\sum_{j=1}^{m_K} v_H(x_{K,j}) \, (q_{K,j} - q_{K,j}^{\ell+3}), w \Big) \\ &\quad = -\mathfrak{a}\Big(\sum_{j=1}^{m_K} v_H(x_{K,j}) \, q_{K,j}^{\ell+3}, (\mathtt{id} - \mathcal{I}_H)(\eta_K w) \Big) \\ &\quad = \mathfrak{a}\Big(\sum_{j=1}^{m_K} v_H(x_{K,j}) \, (q_{K,j} - q_{K,j}^{\ell+3})|_{\mathsf{N}^{\ell+3}(K) \setminus \mathsf{N}^{\ell}(K)}, (\mathtt{id} - \mathcal{I}_H)(\eta_K w) \Big) \\ &\qquad - \mathfrak{a}\Big(\sum_{j=1}^{m_K} v_H(x_{K,j}) \, q_{K,j}|_{\mathsf{N}^{\ell+3}(K) \setminus \mathsf{N}^{\ell}(K)}, (\mathtt{id} - \mathcal{I}_H)(\eta_K w) \Big) \\ &\quad \lesssim \beta_{\mathrm{loc}} C_{\mathcal{I}_H}^2 C_\eta \, \Big\| \sum_{j=1}^{m_K} v_H(x_{K,j}) \, (q_{K,j} - q_{K,j}^{\ell+3}) \Big\|_{\mathcal{V}} \, \|w\|_{\mathcal{V}(\mathsf{N}^{\ell+4}(K) \setminus \mathsf{N}^{\ell}(K))} \\ &\qquad + \beta_{\mathrm{loc}} C_{\mathcal{I}_H}^2 C_\eta \, \Big\| \sum_{j=1}^{m_K} v_H(x_{K,j}) \, q_{K,j} \Big\|_{\mathcal{V}(\mathsf{N}^{\ell+3}(K) \setminus \mathsf{N}^{\ell}(K))} \, \|w\|_{\mathcal{V}(\mathsf{N}^{\ell+4}(K) \setminus \mathsf{N}^{\ell}(K))}. \end{aligned}$$

We now employ the estimates (2.47) and (2.48). Altogether, this yields

$$\begin{aligned}\mathfrak{a}\Big(\sum_{j=1}^{m_K} v_H(x_{K,j})\,(q_{K,j}-q_{K,j}^{\ell+3}),w\Big)&\\ &\lesssim \exp(-C_{\mathrm{dec}}\,\ell)\,\Big\|\sum_{j=1}^{m_K} v_H(x_{K,j})\,\Lambda_{K,j}\Big\|_{\mathcal{V}(K)}\,\|w\|_{\mathcal{V}(\mathsf{N}^{\ell+4}(K)\setminus\mathsf{N}^{\ell}(K))}.\end{aligned}$$

Going back to the estimate (2.50) and using the discrete Cauchy-Schwarz inequality, we obtain

$$\begin{aligned}&\|(\mathcal{C}-\mathcal{C}^{\ell+3})v_H\|_{\mathcal{V}}\\ &\quad\lesssim \exp(-C_{\mathrm{dec}}\,\ell)\sum_{K\in\mathcal{T}_H}\Big(\Big\|\sum_{j=1}^{m_K} v_H(x_{K,j})\,\Lambda_{K,j}\Big\|_{\mathcal{V}(K)}\,\|w\|_{\mathcal{V}(\mathsf{N}^{\ell+4}(K)\setminus\mathsf{N}^{\ell}(K))}\Big)\\ &\quad\lesssim \exp(-C_{\mathrm{dec}}\,\ell)\Big(\sum_{K\in\mathcal{T}_H}\|v_H|_K\|^2_{\mathcal{V}(K)}\Big)^{1/2}\Big(\sum_{K\in\mathcal{T}_H}\|w\|^2_{\mathcal{V}(\mathsf{N}^{\ell+4}(K)\setminus\mathsf{N}^{\ell}(K))}\Big)^{1/2}\\ &\quad\lesssim \ell^{(d-1)/2}\exp(-C_{\mathrm{dec}}\,\ell)\,\|v_H\|_{\mathcal{V}}\,\|w\|_{\mathcal{V}}.\end{aligned}$$

With $\|w\|_{\mathcal{V}}=1$ and a shift in ℓ, this proves the assertion. □

Theorem 2.4.4 allows us to replace the operators $\mathcal{R}$ and $\mathcal{R}^*$ by their localized counterparts $\mathcal{R}^\ell\colon V_H\to\mathcal{R}^\ell V_H$ and $\mathcal{R}^{*,\ell}\colon V_H\to\mathcal{R}^{*,\ell}V_H$ defined by

$$\mathcal{R}^\ell := \mathtt{id}-\mathcal{C}^\ell \quad\text{ and }\quad \mathcal{R}^{*,\ell} := \mathtt{id}-\mathcal{C}^{*,\ell}.$$

With the localized spaces $\mathcal{R}^\ell V_H$ and $\mathcal{R}^{*,\ell}V_H$ and under the assumptions of Theorem 2.4.4, we can formulate the *classical LOD method* that seeks $\tilde{u}_H^\ell\in\mathcal{R}^\ell V_H$ that solves

$$\mathfrak{a}(\tilde{u}_H^\ell,\tilde{v}_H)=\mathcal{F}(\tilde{v}_H) \tag{2.51}$$

for all $\tilde{v}_H\in\mathcal{R}^{*,\ell}V_H$. With Theorem 2.4.4, we directly get the following result.

Theorem 2.4.5 (Error of the classical LOD method)**.** *Let $\ell\in\mathbb{N}$. Suppose that the inf-sup conditions* (2.3), (2.22), (2.41), *and* (2.44) *hold. Further, assume that $\mathcal{F}$ fulfills the regularity condition* (2.29) *and $\mathfrak{a}$ the boundedness condition* (2.39)*. Then the solution $u\in\mathcal{V}$ of* (2.1) *and the solution $\tilde{u}_H^\ell\in\mathcal{R}^\ell V_H$ of* (2.51) *satisfy the error estimate*

$$\|u-\tilde{u}_H^\ell\|_{\mathcal{V}}\lesssim H\,\|f\|_{\mathcal{H}}+\ell^{(d-1)/2}\exp(-C_{\mathrm{dec}}\,\ell)\,\|f\|_{\mathcal{H}}. \tag{2.52}$$

Moreover, if $\ell\gtrsim|\log H|$, we get

$$\|u-\tilde{u}_H^\ell\|_{\mathcal{V}}\lesssim H\,\|f\|_{\mathcal{H}}. \tag{2.53}$$

Proof. First, we remark that $\tilde{u}_H^\ell$ fulfills a quasi-optimality result in the space $\mathcal{R}^\ell V_H$ similar to the one presented in Lemma 2.2.1, i.e.,

$$\|u - \tilde{u}_H^\ell\|_{\mathcal{V}} \lesssim \inf_{\tilde{v}_H \in \mathcal{R}^\ell V_H} \|u - \tilde{v}_H\|_{\mathcal{V}}.$$

Thus, we obtain

$$\begin{aligned}\|u - \tilde{u}_H^\ell\|_{\mathcal{V}} &\lesssim \inf_{\tilde{v}_H \in \mathcal{R}^\ell V_H} \|u - \tilde{v}_H\|_{\mathcal{V}} \lesssim \|u - (\mathtt{id} - \mathcal{C}^\ell)\mathcal{I}_H u\|_{\mathcal{V}} \\ &\lesssim \|u - (\mathtt{id} - \mathcal{C})\mathcal{I}_H u\|_{\mathcal{V}} + \|(\mathcal{C}^\ell - \mathcal{C})\mathcal{I}_H u\|_{\mathcal{V}} \\ &\lesssim H\,\|f\|_{\mathcal{H}} + \ell^{(d-1)/2}\exp(-C_{\mathrm{dec}}\,\ell)\,\|f\|_{\mathcal{H}}\end{aligned}$$

using Theorem 2.3.1, Theorem 2.4.4, (2.14), (2.4), and (2.29). This proves (2.52). The estimate (2.53) follows directly with the choice $\ell \gtrsim |\log H|$. □

Note that the LOD method in (2.51) is still not fully computable since the operators $\mathcal{C}^\ell$ and $\mathcal{C}^{*,\ell}$ are defined by the solutions of (2.43) which are infinite-dimensional problems. This issue is resolved in the following.

2.4.4 Discretization at the microscopic scale

In this subsection, we introduce an additional discretization at the microscopic scale in order to obtain a computable method. There are essentially two possibilities to approach this last step. The idea of the first one is to discretize the localized basis correctors by approximating (2.43) in the discrete space $V_h \cap \mathcal{W}_{\ell,K}$ based on a standard finite element space $V_h \subseteq \mathcal{V}$ with suitable mesh parameter $h < H$. This strategy is, for instance, used in Chapter 5 in the context of the wave equation. On the other hand, the whole construction of this chapter can also be done when replacing the infinite-dimensional space $\mathcal{V}$ with the discrete space V_h. That is, instead of computing $\mathcal{I}_H u$ in (2.17), where $u \in \mathcal{V}$ is the solution of (2.1), we compute $\mathcal{I}_H u_h$, where $u_h \in V_h$ is the classical cG FE solution that solves

$$\mathfrak{a}(u_h, v_h) = \mathcal{F}(v_h) \tag{2.54}$$

for all $v_h \in V_h$. Note that if the conditions (2.6) and (2.7) hold, the problem (2.54) is well-posed.

Replacing $\mathcal{V}$ by V_h (and also $\mathcal{R}$ by $\mathcal{R}_h$, $\mathcal{R}^\ell$ by $\mathcal{R}_h^\ell$, etc.) has the direct consequence that the fine-scale space W_h in the decomposition

$$V_h = \mathcal{R}_h V_H \oplus W_h$$

is also finite-dimensional such that the correctors in (2.43) become computable. After localization as described above, the *fully discrete LOD method* reads: find $\tilde{u}_{H,h}^\ell \in \mathcal{R}_h^\ell V_H$ that solves

$$\mathfrak{a}(\tilde{u}_{H,h}^\ell, \tilde{v}_{H,h}) = \mathcal{F}(\tilde{v}_{H,h}) \tag{2.55}$$

for all $\tilde{v}_{H,h} \in \mathcal{R}_h^{*,\ell} V_H$.

The bases of the spaces $\mathcal{R}_h^{\ell} V_H$ and $\mathcal{R}_h^{*,\ell} V_H$ are obtained by solving quasi-local corrector problems in the form of discrete versions of (2.43) on the scale $h < H$. To avoid an explicit characterization, these problems are usually computed using the saddle point structure presented in Section 2.3.2. We remark that the proofs of Theorem 2.4.1, Lemma 2.4.2, and Theorem 2.4.4 need to be slightly adjusted following the proofs presented in [GP15], but the overall results remain valid in the fully discrete setting if the corresponding inf-sup conditions are satisfied.

Before we quantify the total error of the fully discrete LOD approach, we need to estimate the error between the fine-scale cG solution $u_h \in V_h$ of (2.54) and the solution $u \in \mathcal{V}$ of (2.1). Similarly as in (2.10), we assume that this error can be bounded by

$$\|u - u_h\|_{\mathcal{V}} \leq C_\epsilon \, h \, \|f\|_{\mathcal{H}} \tag{2.56}$$

with a constant C_ϵ that depends on the scale of microscopic oscillations. Therefore, choosing h appropriately allows us to retain the convergence rate of order H. The final result is given in the following theorem.

Theorem 2.4.6 (Error of the fully discrete LOD method)**.** *Let $\ell \gtrsim |\log H|$ and suppose that the assumptions of Theorem 2.4.5 hold in the case where $\mathcal{V}$ is replaced by V_h. Further assume that* (2.56) *is fulfilled and h is small enough to resolve the microscopic scale in the sense that*

$$C_\epsilon \, h \lesssim H. \tag{2.57}$$

Then the fully discrete LOD approximation $\tilde{u}_{H,h}^{\ell} \in \mathcal{R}_h^{\ell} V_H$ in (2.55) *and the solution $u \in \mathcal{V}$ of* (2.1) *satisfy the error estimate*

$$\|u - \tilde{u}_{H,h}^{\ell}\|_{\mathcal{V}} \lesssim H \, \|f\|_{\mathcal{H}}.$$

Proof. If $\ell \gtrsim |\log H|$, we get from Theorem 2.4.5 in the case where V_h replaces $\mathcal{V}$ that

$$\|u_h - \tilde{u}_{H,h}^{\ell}\|_{\mathcal{V}} \lesssim H \, \|f\|_{\mathcal{H}},$$

where $u_h \in V_h$ is the solution of (2.54). With the classical FE estimate (2.56) and the resolution condition (2.57), we further get that

$$\|u - \tilde{u}_{H,h}^{\ell}\| \leq \|u - u_h\|_{\mathcal{V}} + \|u_h - \tilde{u}_{H,h}^{\ell}\|_{\mathcal{V}} \lesssim H \, \|f\|_{\mathcal{H}},$$

which completes the proof. □

We emphasize that the purpose of the resolution condition (2.57) in Theorem 2.4.6 on the fine mesh size h is mainly to retain the convergence rate of order H. Alternatively, one could take a step back from the idea of rigorously tracing convergence rates. That is, one may prescribe some fixed tolerance and balance H and h with the aim to obtain an overall error below the given threshold.

2.5 Numerical experiments

In this section, we present some illustrative examples that show the practical performance of the LOD method. We remark that although the aim of the previous sections was to quantify the full error between the exact solution $u \in \mathcal{V}$ of (2.1) and the fully discrete LOD approximation $\tilde{u}^{\ell}_{H,h} \in \mathcal{R}^{\ell}_{h} V_H$ of (2.55) (see Theorem 2.4.6), in practical computations generally only the error between the solution $u_h \in V_h$ of (2.54) and $\tilde{u}^{\ell}_{H,h} \in \mathcal{R}^{\ell}_{h} V_H$ can be measured. Therefore, it is always implicitly assumed that u_h is a sufficiently good approximation of u, e.g., in the sense of a given tolerance, as discussed in Section 2.4.4.

For our numerical experiments, we consider the following model problem: let $D = (0,1)^2$ and seek $u \in H^1_0(D)$ that solves

$$\int_D A_\epsilon \nabla u \cdot \nabla v \,\mathrm{d}x = \int_D f v \,\mathrm{d}x \tag{2.58}$$

for all $v \in H^1_0(D)$ and given right-hand side f, which is the variational problem corresponding to an elliptic PDE with scalar diffusion coefficient A_ϵ that is bounded from above and below by positive constants and varies on the scale ϵ. We study this problem in more detail in Chapter 3. However, we remark that (2.58) is well-posed and, due to the coercivity of the involved bilinear form, the inf-sup conditions in the above derivations are all satisfied automatically.

As mentioned above, for the error estimates below we compare the coarse-scale solutions to a fine FE solution that resolves the fine-scale oscillations of A_ϵ. In our experiments, this reference solution is computed on a mesh with mesh parameter $h = 2^{-9}$. Further, the errors are computed in the *energy norm* $\|\cdot\|_a := \|A^{1/2}\nabla \cdot\|_{L^2(D)}$, which is equivalent to the classical norm on $H^1_0(D)$.

Before we present numerical examples, we first introduce an explicit quasi-interpolation operator with the properties quantified in Section 2.2.2.

2.5.1 Choice of the quasi-interpolation operator

In this subsection, we briefly present the quasi-interpolation operator that is used for all the experiments in this thesis. We remark that this choice is not unique and any other operator that fulfills the required properties (2.11) and (2.12) could be considered as well.

We set $\mathcal{I}_H := \pi_H \circ \Pi^1_H$, where Π^1_H is the piecewise L^2-projection onto $Q_1(\mathcal{T}_H)$, the space of possibly discontinuous functions which are polynomials of coordinate degree at most one in every component when restricted to an element. Moreover, π_H denotes the averaging operator that maps $Q_1(\mathcal{T}_H)$ to V_H by assigning to each free vertex and each component the arithmetic mean of the corresponding function values of the neighboring elements. Rigorously, for any $v_H \in Q_1(\mathcal{T}_H)$ and $i \in \{1, \dots, n\}$, the ith component of $\pi_H(v_H)$ is characterized

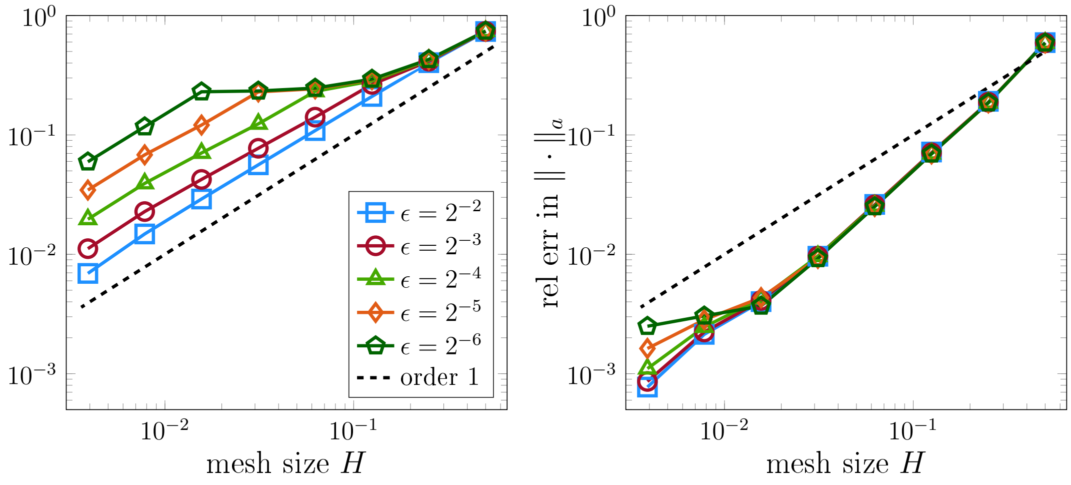


Figure 2.1: Illustration of the relative energy errors of the FEM (left) and the LOD for $\ell = 2$ (right) with respect to the mesh size H and for multiple checkerboard coefficients on the scale ϵ.

by

$$\big(\pi_H(v_H)\big)_i(z) := \sum_{\substack{K\in\mathcal{T}_H:\\ z\in K}} \big(v_H|_K\big)_i(z)\cdot\frac{1}{\operatorname{card}\{T\in\mathcal{T}_H\,:\,z\in T\}}$$

for all vertices z of $\mathcal{T}_H$ with $z\notin\Gamma_i$.

We emphasize that this choice of $\mathcal{I}_H$ satisfies the stability property (2.11) as well as (2.12) and refer to, e.g., [Osw93, Bre94, EG17] for a proof of these conditions.

2.5.2 Comparison between finite elements and LOD

In a first experiment, we study the behavior of classical finite elements in the presence of oscillating coefficients. For a given scale ϵ, let $A_\epsilon\colon (0,1)^2 \to \{1,2\}$ be the periodic and piecewise constant checkerboard coefficient that oscillates between 1 and 2 on the mesh $\mathcal{T}_\epsilon$. Besides, we choose the right-hand side $f(x) = \mathbb{1}_{\{x_1>0.5\}}$, where $\mathbb{1}_S$ denotes the indicator function for the set $S\subseteq D$.

For different oscillation scales ϵ, the relative energy errors of the FE method are depicted in Figure 2.1 (left). One observes the expected first-order convergence rate in terms of the mesh parameter H if the scale ϵ is resolved. However, in the regime $H\gtrsim\epsilon$ the FE solution is not able to provide an appropriate approximation of the exact solution and the error curve stagnates although the mesh size is decreased. This is the pre-asymptotic effect mentioned in Section 2.2.1. Our experiment indicates that a resolution condition of the form $H\lesssim\epsilon$ indeed should hold to observe the expected convergence rate. The experiment also shows that such a bound is sharp.

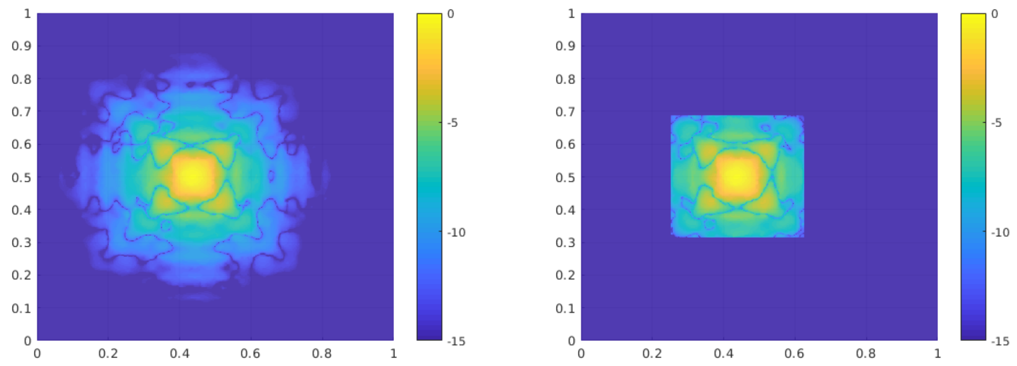


Figure 2.2: Multiscale basis functions ($H = 2^{-4}$) in logarithmic scale: ideal basis function (left) and localized basis function for $\ell = 2$ (right).

For the same model, the energy errors of the LOD approximations with fixed localization parameter $\ell = 2$ are given in Figure 2.1 (right). Due to the fine-scale corrections, the LOD does not suffer from a pre-asymptotic effect and shows the expected convergence behavior which is actually slightly better than first-order. The experiment also shows that a condition of the form $\ell \gtrsim |\log H|$ might even be too pessimistic in certain regimes where the coefficient fulfills additional properties such as periodicity.

We emphasize that the comparison between the FE method and the LOD in Figure 2.1 is only in terms of the convergence behavior. Of course, in terms of computation time a FE method is always faster than a multiscale construction as described above. The main goal of the LOD, however, is to avoid global computations on a fine scale, which could as well be seen as a distribution of complexity. That is, since the computations of the correctors are independent of each other, they can be parallelized and the parallelization procedure is only limited by the specifications of the available computer system. Nevertheless, the method shows its full potential if multiple right-hand sides to the same diffusion coefficient are given or if the PDE at hand is time-dependent, see also Chapters 5 and 6.

2.5.3 Convergence studies in an unstructured setting

As a second example, we consider (2.58) with right-hand side

$$f(x) = (1 + \sin(\pi\, x_1))(1 + 2\,\cos(\tfrac{\pi}{3}\, x_2))$$

and a scalar heterogeneous coefficient that is piecewise constant on the mesh $\mathcal{T}_\epsilon$ with mesh size $\epsilon = 2^{-7}$. In each element $K \in \mathcal{T}_\epsilon$, the value of the coefficient is obtained from a uniform distribution with values in $[0.5, 10]$, i.e., $A|_K \sim U(0.5, 10)$. Further, we choose a nodal basis function $\Lambda_1 \in V_H$, $H = 2^{-4}$,

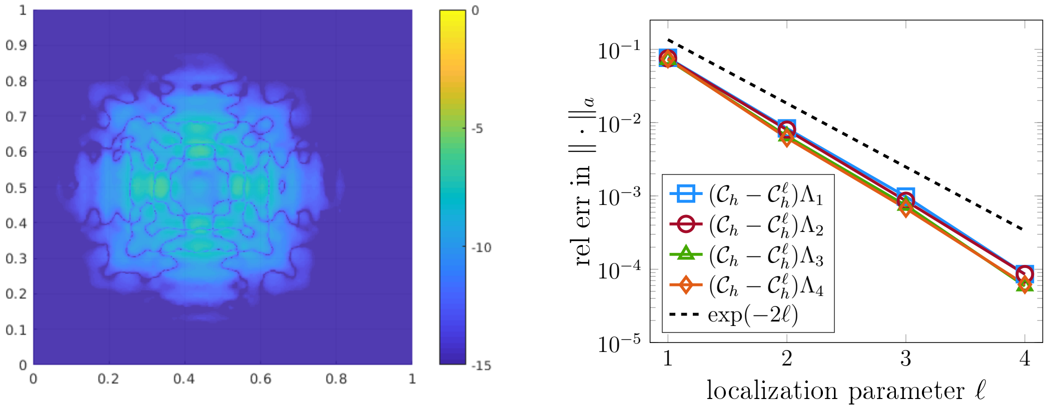


Figure 2.3: Illustration of the localization error $|(\mathcal{C}_h - \mathcal{C}_h^\ell)\Lambda_1|$ in logarithmic scale (left) and localization error in the relative energy norm for different basis functions (right) on the scale $H = 2^{-4}$.

and compute its multiscale counterpart $\tilde{\Lambda}_1 = \mathcal{R}_h\Lambda_1$ corresponding to the vertex $z_1 = (0.4375, 0.5)$. The absolute value of $\tilde{\Lambda}_1$ in logarithmic scale is depicted in Figure 2.2 (left) and illustrates the decay property of such functions; cf. also Theorem 2.4.1. Its localized counterpart $\mathcal{R}_h^\ell\Lambda_1$, $\ell = 2$, is shown in Figure 2.2 (right) and the error between these two basis functions is depicted in Figure 2.3 (left) in logarithmic scale. These illustrations show that the localized function captures the essential characteristics of the global function provided that ℓ is chosen appropriately. This can also be observed in Figure 2.3 (right), where we present the localization errors $\|(\mathcal{C}_h - \mathcal{C}_h^\ell)\Lambda_i\|_a / \|\mathcal{C}_h\Lambda_i\|_a$ for different values of ℓ and the nodal basis functions $\Lambda_i \in V_H$ associated with the nodes z_i, $i \in \{1, \dots, 4\}$, given by

$$z_1 = (0.4375, 0.5), \quad z_2 = (0.5625, 0.5), \quad z_3 = (0.0625, 0.5625), \quad z_4 = (0.125, 0.5).$$

As a reference, we include the behavior of the function $\exp(-2\ell)$ which confirms the theoretical findings of an exponential decay in ℓ as quantified in Theorem 2.4.4.

Finally, the total errors of LOD approximations in the relative energy norm on different discretization scales and for different localization parameters ℓ are depicted in Figure 2.4 (left). One can observe a convergence rate that is even slightly better than the expected first-order rate provided that ℓ is chosen large enough as predicted by the theory. If ℓ is not increased for smaller values of H, the error curve stagnates since the effect of the localization dominates the overall error. This is in line with the assertion of Theorem 2.4.5. Additionally, we also provide L^2-errors of the finite element parts of LOD solutions for different H and ℓ in Figure 2.4 (right). Already for small ℓ, we observe at least first-order convergence which can be expected from the above theory; see, e.g., the

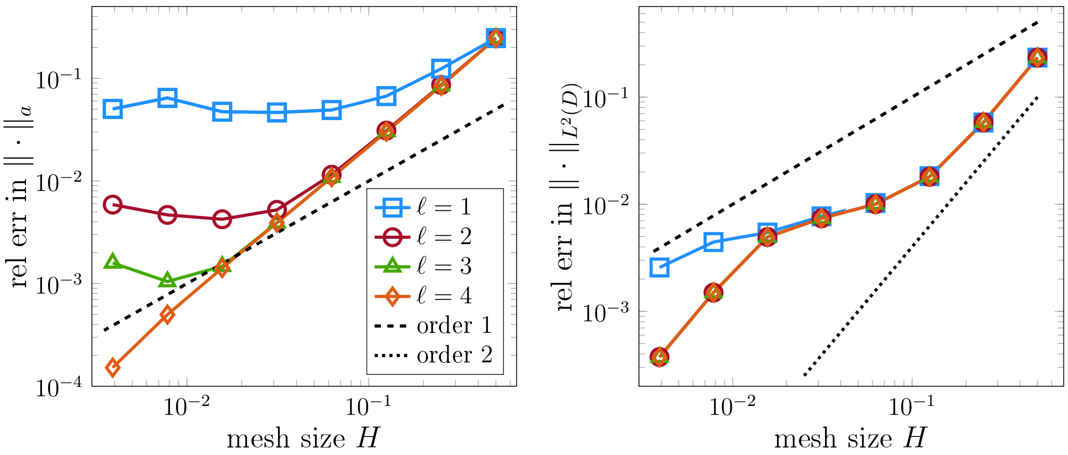


Figure 2.4: Errors of the LOD approximations for different localization parameters ℓ in the relative energy norm (left) and relative L^2-errors of their finite element parts (right) with respect to the mesh size H.

ideal error estimate (2.16). The error curve even partially indicates second-order convergence in the pre-asymptotic regime and if the scale ϵ is resolved. This behavior is for instance discussed in [GP17]; see also Section 4.1.3 and Theorem 4.1.1.

Overall, the numerical experiments verify the theoretical results presented in this chapter when applied to the elliptic setting. In particular, the examples show that the localization procedure described above is justified and the considered localized multiscale method is first-order accurate already for moderate choices of the localization parameter ℓ. Moreover, the approach does not suffer from pre-asymptotic effects in the presence of microscopic coefficients and provides reasonable approximations beyond structural assumptions such as periodicity.

3 A Higher-Order Extension of the Localized Orthogonal Decomposition Method

In the previous chapter, we have presented an approach based on a first-order FE space that constructs a multiscale space that is able to cope with heterogeneous and possibly microscopic properties of, e.g., an underlying material coefficient. In general, one could generalize the idea and consider higher-order conforming discrete spaces as used in the context of *hp* methods; we refer to, e.g., [BG96, Sch98] for further details on *hp* finite elements. Although there exist quasi-interpolation operators for such spaces that fulfill properties similar to (2.11) and (2.12) without restrictive regularity assumptions [Mel05], a construction as in Chapter 2 does not provide higher-order convergence rates with respect to H for general non-smooth coefficients. Therefore, the derivation of a higher-order multiscale method calls for an appropriate adjustment of the construction presented in Chapter 2.

In this chapter, we consider the use of discontinuous FE spaces for a higher-order multiscale construction. This idea traces back to [EGMP13] and [HP13]. In [HP13], local corrections of element-wise discontinuous functions were considered as described in Section 2.4.1. It turned out that a splitting of conforming FE functions into element-wise discontinuous contributions has a favorable effect on the localization procedure in connection with the classical (conforming) LOD as presented in Chapter 2. Then again, a truly discontinuous approach was used in [EGMP13] to construct a first-order discontinuous Galerkin (dG) multiscale method for an elliptic model problem. The approach is based on the decomposition of a fine discontinuous FE space into a coarse discontinuous multiscale FE space and the remaining (discontinuous) fine-scale space which are orthogonal with respect to the mesh-dependent bilinear form that arises in connection with a *symmetric interior penalty approach* (see, e.g., [DD76, Arn82, HSW07]).

Here, we base the method on an orthogonal decomposition of the infinite-dimensional space $\mathcal{V}$, as in Chapter 2, and build the higher-order ansatz on the saddle point formulation described in Section 2.3.2, which allows for a generalization using discontinuous spaces. This approach is, for instance, also employed in connection with gamblets [Owh15, Owh17], usually with spaces consisting of piecewise constant functions. The aim of this chapter is to extend these ideas to construct a higher-order variant of the LOD based on piecewise polynomials which allows for a thorough treatment of not only the convergence behavior with respect to the mesh size H but also the polynomial degree p.

Since the abstract theory of Chapter 2 is not directly applicable to the higher-order setting of this chapter, as discussed above, and the extension of the method requires more refined arguments to successfully trace the involved parameters, we restrict ourselves to an elliptic setting as introduced in the following section. We emphasize that the overall construction also works for a more general setting but the results presented below do not immediately follow and need to be adjusted to the respective framework.

3.1 Elliptic model problem

In this section, we present the model problem used throughout this chapter. We consider the variational formulation corresponding to the prototypical second-order diffusion problem

$$\begin{aligned} -\operatorname{div}(A\nabla u) &= f \quad \text{in } D,\\ u &= 0 \quad \text{on } \partial D, \end{aligned} \tag{3.1}$$

where $D \subseteq \mathbb{R}^d$, $d \in \{1,2,3\}$, is a bounded, convex, and polytopal Lipschitz domain and $f \in L^2(D)$. We assume the coefficient A to encode microscopic features of the medium on some scale ϵ and to be *admissible*, i.e., it belongs to the set

$$\mathfrak{A} := \left\{ \begin{aligned} &A \in L^\infty(D;\mathbb{R}^{d\times d}_{\mathrm{sym}}) \,:\, \exists\, 0 < \alpha \leq \beta < \infty \,: \\ &\forall \xi \in \mathbb{R}^d,\ \text{a.a. } x \in D \,:\, \alpha|\xi|^2 \leq A(x)\xi\cdot\xi \leq \beta|\xi|^2 \end{aligned} \right\} \tag{3.2}$$

with minimal assumptions. For a given coefficient $A \in \mathfrak{A}$, we write α for the largest possible choice of α in the definition (3.2) and β for the L^∞-norm of A, i.e., $\beta = \|A\|_{L^\infty(D;\mathbb{R}^{d\times d}_{\mathrm{sym}})}$, although this choice of β might not be the minimal constant with respect to the estimate in (3.2). We emphasize that also positive and bounded scalar coefficients are admissible, since these coefficients may simply be multiplied by the identity matrix.

With regard to the spaces in Chapter 2, we have $\mathcal{V} = H^1_0(D)$ as well as $\mathcal{H} = L^2(D)$. To derive the variational formulation of (3.1), we multiply its first line with a test function $v \in H^1_0(D)$, integrate by parts, and obtain

$$\int_D A\nabla u \cdot \nabla v \,\mathrm{d}x = \int_D f v \,\mathrm{d}x \tag{3.3}$$

using the boundary condition of u. The left-hand side of (3.3) motivates the definition of the symmetric bilinear form $a\colon H^1_0(D) \times H^1_0(D) \to \mathbb{R}$,

$$a(v,w) := \int_D A\nabla v \cdot \nabla w \,\mathrm{d}x \tag{3.4}$$

for any v, $w \in H_0^1(D)$. As in Chapter 2, we deduce from the Friedrichs inequality that the H^1-seminorm $|\cdot|_{H^1(D)} = \|\nabla \cdot\|_{L^2(D)}$ is actually a norm on $H_0^1(D)$ which is equivalent to the standard H^1-norm. Using this and the fact that $A \in \mathfrak{A}$, we directly get boundedness and coercivity of a, i.e., for any v, $w \in H_0^1(D)$, it holds that

$$a(v, w) \leq \beta \, \|\nabla v\|_{L^2(D)} \, \|\nabla w\|_{L^2(D)}, \tag{3.5}$$

making use of the Hölder inequality, and

$$a(v, v) \geq \alpha \, \|\nabla v\|_{L^2(D)}^2. \tag{3.6}$$

Note that from the definition of a in terms of an integral, we directly get that the bilinear form a fulfills the local boundedness condition (2.39) that had to be explicitly assumed in Chapter 2.

With the bounds (3.5) and (3.6), we get from Chapter 2 or directly with the Lax-Milgram Theorem that there exists a unique solution $u \in H_0^1(D)$ that solves

$$a(u, v) = (f, v)_{L^2(D)} \tag{3.7}$$

for all $v \in H_0^1(D)$. Further, it holds that

$$\|\nabla u\|_{L^2(D)} \leq \alpha^{-1} \, \|f\|_{L^2(D)}, \tag{3.8}$$

see also (2.4).

3.2 Construction of higher-order multiscale spaces

Inspired by the findings presented in Chapter 2, the multiscale approach of this chapter, which aims at finding a discrete approximation of u in (3.7), is also based on the idea of decomposing the space $H_0^1(D)$ into a coarse FE-type space V_H on some scale H and an infinite-dimensional fine-scale space $\mathcal{W}$. While this decomposition was chosen in a conforming fashion in Chapter 2, i.e.,

$$V_H \subseteq H_0^1(D) \quad \text{and} \quad \mathcal{W} \subseteq H_0^1(D),$$

the construction of our higher-order variant is explicitly based on non-conforming spaces. However, the multiscale space $\tilde{V}_H$ constructed from V_H and $\mathcal{W}$ should again be a conforming space such that the final multiscale decomposition

$$H_0^1(D) = \tilde{V}_H \oplus (\mathcal{W} \cap H_0^1(D))$$

consists of two conforming spaces in contrast to the decomposition

$$H_0^1(D) \subseteq V_H \oplus \mathcal{W}$$

with two non-conforming spaces. The main problem with this generalization to non-conforming spaces is the fact that many of the arguments used in Chapter 2 explicitly rely on the fact that V_H and $\mathcal{W}$ are subspaces of $H^1_0(D)$. Nevertheless, the saddle point formulation in Section 2.3.2 presents an ideal basis for the non-conforming construction. Before we get into the details, we introduce the discrete framework of this chapter.

3.2.1 Discontinuous discrete spaces

Let, as in Section 2.2, $\{\mathcal{T}_H\}_{H>0}$ be a family of regular decompositions of the domain D into quasi-uniform d-rectangles on the scale H and denote with V_H^p the space of piecewise polynomial functions with prescribed maximal coordinate degree, i.e.,

$$V_H^p := \left\{ \begin{aligned} v \in L^2(D) : \forall K \in \mathcal{T}_H :\ v|_K \text{ is a polynomial} \\ \text{of coordinate degree } \leq p \end{aligned} \right\}.$$

Note that we explicitly indicate the dependence on the polynomial degree $p \in \mathbb{N}$ because in this chapter the convergence not only with respect to the mesh parameter H is investigated but also with respect to the polynomial degree. For any $S \subseteq D$, we further write $V_H^p(S)$ for the restriction of V_H^p to the subdomain S. In particular, for any $K \in \mathcal{T}_H$, the restricted space $V_H^p(K)$ is exactly the space of polynomials up to degree p in each coordinate direction on the element K. For later use, we also define for $k \in \mathbb{N}$ the *broken Sobolev space* $H^k(\mathcal{T}_H)$ by

$$H^k(\mathcal{T}_H) := \{v \in L^2(D) : \forall K \in \mathcal{T}_H :\ v|_K \in H^k(K)\}.$$

with the seminorm

$$|\cdot|^2_{H^k(\mathcal{T}_H)} := \sum_{K \in \mathcal{T}_H} |\cdot|^2_{H^k(K)},$$

where $|\cdot|_{H^k(S)} := \|\nabla^k \cdot\|_{L^2(S)}$ denotes the H^k-seminorm on $S \subseteq D$.

As before, the next step of the construction consists in defining a projection operator onto the space V_H^p that fulfills local stability and approximation properties in the sense of (2.11) and (2.12). The non-conforming nature of the space V_H^p, however, allows us to use a truly local projection. Here, we choose the L^2-projection $\Pi_H^p\colon L^2(D) \to V_H^p$ defined for any $v \in L^2(D)$ by

$$\big(\Pi_H^p v, w_H\big)_{L^2(D)} = \big(v, w_H\big)_{L^2(D)} \tag{3.9}$$

for all $w_H \in V_H^p$. The above-mentioned locality of Π_H^p comes from the elementwise definition of the space V_H^p and the possible discontinuities across element boundaries. That is, the definition of Π_H^p in (3.9) is equivalent to the elementwise characterization

$$\big((\Pi_H^p v)|_K, q\big)_{L^2(K)} = \big(v, q\big)_{L^2(K)} \tag{3.10}$$

for all $q \in V_H^p(K)$ and $K \in \mathcal{T}_H$. For the sake of readability, in the following we abbreviate $\Pi := \Pi_H^p$ if p and H are explicitly given and there is no possibility of confusion.

For any $K \in \mathcal{T}_H$, the L^2-stability of Π follows directly from equation (3.10) with the choice $q = (\Pi v)|_K$ and reads

$$\|\Pi v\|_{L^2(K)} \leq \|v\|_{L^2(K)} \tag{3.11}$$

for all $v \in L^2(K)$. Further, it holds that

$$\|(\mathtt{id} - \Pi)v\|_{L^2(K)} \leq C_\Pi \frac{H}{p} \, \|\nabla v\|_{L^2(K)} \tag{3.12}$$

for all $v \in H^1(K)$; see, e.g., [Sch98, HSS02, Geo03]. If $v \in H^k(K)$ for $k \in \mathbb{N}$ and $k \leq p+1$, we even have

$$\|(\mathtt{id} - \Pi)v\|_{L^2(K)} \leq C_\Pi \, \Phi(p,k) \, H^k \, |v|_{H^k(K)} \tag{3.13}$$

with a constant C_Π that does not depend on H or p and

$$\Phi(p,k) := \left(\frac{(p+1-k)!}{(p+1+k)!}\right)^{1/2}. \tag{3.14}$$

We emphasize that due to the true locality of the inequalities (3.11) and (3.12), the results immediately generalize to unions of elements and, in particular, to a global result on the domain D in the sense of an element-wise gradient on the right-hand side. Based on the projection Π, we define, as before, the fine-scale space $\mathcal{W}$ as the kernel of Π with respect to the space $H_0^1(D)$, i.e.,

$$\mathcal{W} := (\mathtt{id} - \Pi) H_0^1(D) = \ker \Pi|_{H_0^1(D)}.$$

At this point, we also introduce the inverse inequality for polynomials which states that

$$\|\nabla q\|_{L^2(K)} \leq C_{\mathrm{inv}} H^{-1} p^2 \, \|q\|_{L^2(K)} \tag{3.15}$$

for $K \in \mathcal{T}_H$ and for all polynomials $q \in V_H^p(K)$; see, e.g., [Sch98, GHS05, Geo08]. As above, this result also holds globally, i.e.,

$$|v_H|_{H^1(\mathcal{T}_H)} \leq C_{\mathrm{inv}} H^{-1} p^2 \, \|v_H\|_{L^2(D)}$$

for all $v_H \in V_H^p$. We emphasize that $\Pi \colon L^2(D) \to V_H^p$ is obviously surjective as an operator from $L^2(D)$ to the non-conforming space V_H^p. Next, we show that the projection operator Π is also surjective when restricted to functions in $H_0^1(D)$. To prove this assertion, we need the following lemma.

Lemma 3.2.1 (Local inf-sup condition)**.** *Let $K \in \mathcal{T}_H$. Then the inf-sup condition*

$$\inf_{q \in V_H^p(K)} \sup_{v \in H_0^1(K)} \frac{(q,v)_{L^2(K)}}{\|q\|_{L^2(K)} \, \|\nabla v\|_{L^2(K)}} \geq \gamma(H,p) > 0 \tag{3.16}$$

holds with $\gamma(H,p) \sim Hp^{-2}$.

Proof. Let $\kappa \subseteq K$ be such that the edges, faces, etc. of κ are parallel to the ones of K. According to [Geo08, Lem. 3.7], there exists a choice of κ such that $\operatorname{dist}(\kappa, \partial K) = C_{\text{dist}}\, H p^{-2}$ and

$$\|q\|^2_{L^2(\kappa)} \geq \frac{1}{4}\|q\|^2_{L^2(K)} \tag{3.17}$$

for all $q \in V^p_H(K)$, where $\operatorname{dist}(\cdot,\cdot)$ denotes the Hausdorff distance. Now, let $\rho \in W^{1,\infty}(K) \cap H^1_0(K)$ be a bubble function with

$$\begin{aligned} 0 \;&\leq\; \rho \;\leq\; 1, \\ \rho \;&\equiv\; 1 \quad \text{in } \kappa, \\ \|\nabla\rho\|_{L^\infty(K)} \;&\leq\; C_\rho\, H^{-1}p^2, \end{aligned}$$

where C_ρ depends on C_{dist}. Using (3.17) and

$$\begin{aligned} \|\nabla(\rho q)\|_{L^2(K)} &\leq \|\nabla\rho\|_{L^\infty(K)}\, \|q\|_{L^2(K)} + \|\rho\|_{L^\infty(K)}\, \|\nabla q\|_{L^2(K)} \\ &\leq H^{-1}p^2(C_\rho + C_{\text{inv}})\, \|q\|_{L^2(K)}, \end{aligned} \tag{3.18}$$

we get for any $q \in V^p_H(K)$ that

$$\begin{aligned} \sup_{v \in H^1_0(K)} \frac{(q,v)_{L^2(K)}}{\|q\|_{L^2(K)}\, \|\nabla v\|_{L^2(K)}} &\geq \frac{(q,\rho q)_{L^2(K)}}{\|q\|_{L^2(K)}\, \|\nabla(\rho q)\|_{L^2(K)}} \\ &\geq \frac{1}{4}\frac{\|q\|^2_{L^2(K)}}{\|q\|_{L^2(K)}\, \|\nabla(\rho q)\|_{L^2(K)}} \\ &= \frac{H}{4p^2\,(C_\rho + C_{\text{inv}})} =: \gamma(H,p) > 0. \end{aligned}$$

Taking the infimum over $q \in V^p_H(K)$, we obtain the assertion. □

Theorem 3.2.2 (Surjectivity)**.** *The restricted operator $\Pi|_{H^1_0(D)}$ is surjective, i.e., for any $w_H \in V^p_H$, there exists a function $w \in H^1_0(D)$ such that $\Pi w = w_H$. Further, among all possible candidates exists a choice of w such that*

$$\|\nabla w\|_{L^2(D)} \lesssim \frac{p^2}{H}\, \|w_H\|_{L^2(D)}. \tag{3.19}$$

Proof. Let $w_H \in V^p_H$. We define $w \in H^1_0(D)$ as the solution of

$$\begin{aligned} a(w,v) \quad &+ \quad (\lambda_{w_H}, v)_{L^2(D)} &= \quad 0, \\ (w,\mu_H)_{L^2(D)} & &= \quad (w_H,\mu_H)_{L^2(D)} \end{aligned} \tag{3.20}$$

for all $v \in H^1_0(D)$ and all $\mu_H \in V^p_H$. From classical saddle point theory (see, e.g., [BBF13, Cor. 4.2.1]), we know that (3.20) has a unique solution if the inf-sup condition

$$\inf_{v_H \in V^p_H} \sup_{v \in H^1_0(D)} \frac{(v_H, v)_{L^2(D)}}{\|v_H\|_{L^2(D)}\, \|\nabla v\|_{L^2(D)}} \geq \tilde{\gamma}(H,p) > 0 \tag{3.21}$$

holds and a is coercive. To show the inf-sup condition (3.21), let $v_H \in V_H^p$. From the construction in the proof of Lemma 3.2.1, we get for any $K \in \mathcal{T}_H$ the existence of a function $v_K \in H_0^1(K)$ which fulfills

$$(v_H, v_K)_{L^2(K)} \gtrsim \|v_H\|_{L^2(K)}^2 \tag{3.22}$$

and similarly to (3.18) also

$$\|\nabla v_K\|_{L^2(K)} \lesssim H^{-1}p^2 \, \|v_H\|_{L^2(K)}. \tag{3.23}$$

Using these local contributions, the inclusion

$$\bigcup_{K\in\mathcal{T}_H} H_0^1(K) \subseteq H_0^1(D),$$

and the estimates (3.22) and (3.23), we compute

$$\begin{aligned}\sup_{v\in H_0^1(D)} \frac{(v_H, v)_{L^2(D)}}{\|v_H\|_{L^2(D)}\,\|\nabla v\|_{L^2(D)}} &\geq \frac{\sum_{K\in\mathcal{T}_H}(v_H, v_K)_{L^2(K)}}{\|v_H\|_{L^2(D)}\big(\sum_{K\in\mathcal{T}_H}\|\nabla v_K\|_{L^2(K)}^2\big)^{1/2}}\\ &\geq C\,Hp^{-2}\frac{\sum_{K\in\mathcal{T}_H}\|v_H\|_{L^2(K)}^2}{\|v_H\|_{L^2(D)}\big(\sum_{K\in\mathcal{T}_H}\|v_H\|_{L^2(K)}^2\big)^{1/2}} = C\,Hp^{-2} > 0.\end{aligned}$$

That is, the inf-sup condition (3.21) holds with $\tilde{\gamma}(H,p) \sim Hp^{-2}$. Thus, (3.20) is well-posed and the stability estimates

$$\|\lambda_{w_H}\|_{L^2(D)} \leq \frac{\beta}{\tilde{\gamma}(H,p)^2}\,\|w_H\|_{L^2(D)}$$

and

$$\|\nabla w\|_{L^2(D)} \leq \frac{2\beta^{1/2}}{\alpha^{1/2}\tilde{\gamma}(H,p)}\,\|w_H\|_{L^2(D)}$$

hold (cf. [BBF13, Cor. 4.2.1]). Finally, we remark that the equality $\Pi w = w_H$ follows by construction. □

The construction in the proof of Theorem 3.2.2 is based on local subspaces of $H_0^1(D)$ and, thus, allows us to even find a conforming preimage $w \in H_0^1(D)$ under Π of a function $w_H \in V_H^p$ which is supported only in the elements where w_H is non-zero. This straightforward consequence is given in the following corollary.

Corollary 3.2.3 (Local bubble function)**.** *Let $\{K_j\}_{j=1}^{n_R} \subseteq \mathcal{T}_H$ be a set of elements and $w_H \in V_H^p$ such that*

$$w_H|_{D\setminus R} = 0, \qquad \textit{where} \qquad R = \bigcup_{j=1}^{n_R} K_j.$$

Then there exists a function $w \in H_0^1(R)$ with $w|_{D\setminus R} = 0$ such that $\Pi w = w_H$ and

$$\|\nabla w\|_{L^2(R)} \lesssim \frac{p^2}{H}\,\|w_H\|_{L^2(R)}. \tag{3.24}$$

3.2.2 Ideal trial and test space

In the spirit of Chapter 2, we can now construct an operator $\mathcal{R}\colon V_H^p \to H_0^1(D)$ that assigns to each $v_H \in V_H^p$ a continuous function whose L^2-projection is exactly v_H. From Theorem 3.2.2 or Corollary 3.2.3, we know that such functions exist but, as before, we particularly want the space $\mathcal{R}V_H^p$ to have improved approximation properties compared to a classical FE space for which error estimates typically depend on the scale of microscopic oscillations.

To this end, we start our construction by adopting the definition of $\mathcal{R}$ presented in Section 2.3.2. We remark that in the setting of this chapter, it holds that $\mathcal{R}^* = \mathcal{R}$ since a is symmetric. Thus, we do not distinguish between the two operators in the following. We define $\mathcal{R}\colon V_H^p \to H_0^1(D)$ for any $v_H \in V_H^p$ as the solution of the saddle point problem

$$\begin{array}{lcllcl} a(\mathcal{R}v_H, v) & + & (\lambda_{v_H}, v)_{L^2(D)} & = & 0, \\ (\mathcal{R}v_H, \mu_H)_{L^2(D)} & & & = & (v_H, \mu_H)_{L^2(D)} \end{array} \tag{3.25}$$

for all $v \in H_0^1(D)$ and all $\mu_H \in V_H^p$, where $\lambda_{v_H} \in V_H^p$ is the associated Lagrange multiplier. From the construction in the proof of Theorem 3.2.2, we know that there exists a unique solution $(\mathcal{R}v_H, \lambda_{v_H}) \in H_0^1(D) \times V_H^p$ of (3.25) and that

$$\|\nabla \mathcal{R}v_H\|_{L^2(D)} \lesssim \frac{p^2}{H}\, \|v_H\|_{L^2(D)} \tag{3.26}$$

for any $v_H \in V_H^p$. Note that due to the symmetry of a, the operator $\mathcal{R}$ is equivalently defined by

$$\mathcal{R}v_H := \operatorname*{arg\,min}_{v \in H_0^1(D)} a(v,v) \quad \text{subject to} \quad \Pi v = v_H. \tag{3.27}$$

We now set $\tilde{V}_H^p := \mathcal{R}V_H^p \subseteq H_0^1(D)$ and observe that $\dim \tilde{V}_H^p = \dim V_H^p$ because $\mathcal{R}\colon V_H^p \to \tilde{V}_H^p$ is a bijection with inverse $\Pi|_{\tilde{V}_H^p}$. We use $\tilde{V}_H^p$ as test and trial space to obtain a finite-dimensional approximation of (3.7) in the next subsection.

Remark 3.2.4. In the one-dimensional setting with a constant coefficient, the above definition of $\mathcal{R}$ produces the classical spline space of order $p+2$. This smoothing property is, for instance, employed in [HMP+19] in connection with a diffuse approximation of jumping coefficients to avoid spurious oscillations.

3.2.3 The ideal method

In this subsection, we introduce and analyze an *ideal method* to discretize problem (3.7) with a cG FE approach based on the space $\tilde{V}_H^p$ introduced in the previous subsection: find $\tilde{u}_H \in \tilde{V}_H^p$ such that

$$a(\tilde{u}_H, \tilde{v}_H) = (f, \tilde{v}_H)_{L^2(D)} \tag{3.28}$$

for all $\tilde{v}_H \in \tilde{V}_H^p$. As for the variational problem (3.7), we directly get the well-posedness of (3.28) from the Lax-Milgram Theorem using the coercivity of the bilinear form a and the conformity of $\tilde{V}_H^p$.

Before we further analyze the method, we state the following useful result.

Lemma 3.2.5 (Equivalent formulation)**.** *Let $u \in H_0^1(D)$ be the solution of* (3.7)*. Then the solution $\tilde{u}_H \in \tilde{V}_H^p$ of* (3.28) *is equivalently defined as the function $\tilde{u}_H \in H_0^1(D)$ that solves*

$$\begin{array}{rcrcl} a(\tilde{u}_H, v) & + & (\lambda_{\Pi u}, \Pi v)_{L^2(D)} & = & 0, \\ (\Pi \tilde{u}_H, \mu_H)_{L^2(D)} & & & = & (\Pi u, \mu_H)_{L^2(D)} \end{array} \tag{3.29}$$

for all $v \in H_0^1(D)$ and $\mu_H \in V_H^p$, where $\lambda_{\Pi u} \in V_H^p$ is the associated Lagrange multiplier.

Proof. The assertion follows with similar arguments as in the proof of Theorem 2.3.2. For $\tilde{v}_H = \mathcal{R} v_H \in \tilde{V}_H^p$, we compute

$$\begin{aligned} a(\mathcal{R}\Pi u, \tilde{v}_H) &= a(u, \tilde{v}_H) - a((\mathtt{id} - \mathcal{R}\Pi)u, \tilde{v}_H) \\ &= (f, \tilde{v}_H)_{L^2(D)} - a((\mathtt{id} - \mathcal{R}\Pi)u, \tilde{v}_H). \end{aligned}$$

Since $\Pi(\mathtt{id} - \mathcal{R}\Pi)u = 0$, we get with (3.25) that

$$a((\mathtt{id} - \mathcal{R}\Pi)u, \tilde{v}_H) = 0.$$

Therefore, $\mathcal{R}\Pi u$ is the (unique) solution of problem (3.28). □

The next theorem states that under additional (piecewise) regularity assumptions on the right-hand side f, the error between the solutions of (3.7) and (3.28) scales optimally with respect to H and p and does not depend on the oscillations of the coefficient.

Theorem 3.2.6 (Error of the ideal method)**.** *Assume that $f \in H^k(\mathcal{T}_H)$, $k \in \mathbb{N}_0$, and define $s := \min\{k, p+1\}$. Further, let $u \in H_0^1(D)$ and $\tilde{u}_H \in \tilde{V}_H^p$ be the solutions of* (3.7) *and* (3.28)*, respectively. Then*

$$\|\nabla(u - \tilde{u}_H)\|_{L^2(D)} \lesssim \frac{\Phi(p,s)}{p} H^{s+1} |f|_{H^s(\mathcal{T}_H)} \tag{3.30}$$

and

$$\|u - \tilde{u}_H\|_{L^2(D)} \lesssim \frac{\Phi(p,s)}{p^2} H^{s+2} |f|_{H^s(\mathcal{T}_H)}, \tag{3.31}$$

with the notation $H^0(\mathcal{T}_H) := L^2(D)$ and $|\cdot|_{H^0(\mathcal{T}_H)} := \|\cdot\|_{L^2(D)}$.

Proof. Using the Galerkin orthogonality, (3.12), and (3.13), for $k \geq 1$ we obtain

$$\begin{aligned}\alpha\,\|\nabla(u-\tilde{u}_H)\|^2_{L^2(D)} &\leq a(u-\tilde{u}_H, u-\tilde{u}_H) = a(u, u-\tilde{u}_H)\\ &= (f, u-\tilde{u}_H)_{L^2(D)} = (f-\Pi f, u-\tilde{u}_H)_{L^2(D)}\\ &\leq \|f-\Pi f\|_{L^2(D)}\, C_\Pi\,\frac{H}{p}\,\|\nabla(u-\tilde{u}_H)\|_{L^2(D)}\\ &\leq C_\Pi\,\Phi(p,s)\,H^s\,|f|_{H^s(\mathcal{T}_H)}\,C_\Pi\,\frac{H}{p}\,\|\nabla(u-\tilde{u}_H)\|_{L^2(D)}\end{aligned}$$

employing that $\Pi(u-\tilde{u}_H)=0$ by Lemma 3.2.5. Thus,

$$\|\nabla(u-\tilde{u}_H)\|_{L^2(D)} \leq \alpha^{-1} C_\Pi^2\,\frac{\Phi(p,s)}{p}\,H^{s+1}\,|f|_{H^s(\mathcal{T}_H)}.$$

With the same arguments but without inserting Πf, we get in the case $k=0$ that

$$\|\nabla(u-\tilde{u}_H)\|_{L^2(D)} \leq \alpha^{-1} C_\Pi\,\frac{H}{p}\,\|f\|_{L^2(D)}.$$

This proves (3.30). To show the L^2-error estimate, we use once again that $\Pi(u-\tilde{u}_H)=0$. Therefore, we get with (3.12) that

$$\|u-\tilde{u}_H\|_{L^2(D)} \leq C_\Pi\,\frac{H}{p}\,\|\nabla(u-\tilde{u}_H)\|_{L^2(D)}.$$

Combining the last estimate with (3.30), we deduce (3.31). □

Remark 3.2.7. If $p=1$ in the above construction, the error estimate in Theorem 3.2.6 is comparable to the one presented in Chapter 2 in connection with the classical conforming approach; see Theorem 2.3.1.

3.3 Derivation of a practical method

As already addressed in the previous chapter for the classical LOD, the ideal method given in (3.28) is a finite-dimensional approximation of the solution u of (3.7) but the construction of the space $\tilde{V}_H^p$ involves the solution of infinite-dimensional problems. Thus, we follow the strategy presented in Section 2.4 and adapt it to the setting of this chapter with the non-conforming spaces V_H^p and $\mathcal{W}$ in order to derive a fully practical method. First, we investigate the decay properties of functions in $\tilde{V}_H^p$ and especially focus on the dependence on the polynomial degree p.

3.3.1 Decay of the basis functions

As a first step, we identify a suitable choice of a basis of $\tilde{V}_H^p$ which is constructed from a basis of V_H^p. For any $K \in \mathcal{T}_H$, let

$$\mathfrak{B}_K := \{\Lambda_{K,j}\}_{j=1}^{m_K} \quad \text{with} \quad m_K = (p+1)^d$$

be a basis of $V_H^p(K)$ and

$$\mathfrak{B} := \bigcup_{K \in \mathcal{T}_H} \mathfrak{B}_K$$

the corresponding (local) basis of V_H^p. In our numerical computations, we choose shifted Legendre polynomials on each element K, which are orthogonal with respect to the L^2-scalar product $(\cdot,\cdot)_{L^2(K)}$.

Using the isomorphism $\mathcal{R}$ between V_H^p and $\tilde{V}_H^p$, we directly get that $\tilde{\mathfrak{B}} := \mathcal{R}\mathfrak{B}$ is a basis of $\tilde{V}_H^p$. In the following, we show that for any basis function $\Lambda \in \mathfrak{B}$, the corresponding basis function $\mathcal{R}\Lambda \in \tilde{\mathfrak{B}}$ decays exponentially fast away from the support of the function Λ, which is exactly one element of $\mathcal{T}_H$.

Theorem 3.3.1 (Decay of the basis functions)**.** *Let $\ell \in \mathbb{N}$, $K \in \mathcal{T}_H$, and $\Lambda \in \mathfrak{B}_K$. Further, define $\tilde{\Lambda} = \mathcal{R}\Lambda \in \tilde{\mathfrak{B}}$. Then it holds that*

$$\|\nabla\tilde{\Lambda}\|_{L^2(D\setminus \mathsf{N}^\ell(K))} \lesssim \exp(-C_{\mathrm{dec}}\,\ell/p)\,\|\nabla\tilde{\Lambda}\|_{L^2(D)} \tag{3.32}$$

with a constant C_{dec} that depends on C_Π, α, and β.

Proof. We choose a cutoff function $\eta \in W^{1,\infty}(D)$ with the following properties:

$$\begin{aligned} &0 \le \eta \le 1, \\ &\eta \;=\; 0 \quad \text{in } \mathsf{N}^\ell(K), \\ &\eta \;=\; 1 \quad \text{in } D \setminus \mathsf{N}^{\ell+1}(K), \\ &\|\nabla\eta\|_{L^\infty(D)} \;\le\; C_\eta\, H^{-1}. \end{aligned} \tag{3.33}$$

Define $R := \mathsf{N}^{\ell+1}(K) \setminus \mathsf{N}^\ell(K)$. Since R is a union of elements of $\mathcal{T}_H$ and $\Pi(\tilde{\Lambda}\eta)|_{D\setminus R} = 0$, we know from Corollary 3.2.3 that there exists a bubble function $b \in H_0^1(R)$ which fulfills $\Pi b = \Pi(\tilde{\Lambda}\eta)$ and

$$\|\nabla b\|_{L^2(R)} \lesssim \frac{p^2}{H}\|\tilde{\Lambda}\eta\|_{L^2(R)}. \tag{3.34}$$

We compute

$$\begin{aligned} \alpha\,\|\nabla\tilde{\Lambda}\|^2_{L^2(D\setminus\mathsf{N}^{\ell+1}(K))} &\le \Big|\int_D A\nabla\tilde{\Lambda}\cdot\nabla(\tilde{\Lambda}\eta)\,\mathrm{d}x\Big| + \Big|\int_D A\nabla\tilde{\Lambda}\cdot\nabla\eta\,\tilde{\Lambda}\,\mathrm{d}x\Big| \\ &= \Big|\int_D A\nabla\tilde{\Lambda}\cdot\nabla(\tilde{\Lambda}\eta - b)\,\mathrm{d}x\Big| \\ &\qquad + \Big|\int_D A\nabla\tilde{\Lambda}\cdot\nabla b\,\mathrm{d}x\Big| + \Big|\int_D A\nabla\tilde{\Lambda}\cdot\nabla\eta\,\tilde{\Lambda}\,\mathrm{d}x\Big| \\ &= \Big|\int_R A\nabla\tilde{\Lambda}\cdot\nabla b\,\mathrm{d}x\Big| + \Big|\int_R A\nabla\tilde{\Lambda}\cdot\nabla\eta\,\tilde{\Lambda}\,\mathrm{d}x\Big|, \end{aligned}$$

where we use the fact that, by definition (3.25), $a(\tilde{\Lambda}, v) = 0$ for $v \in H_0^1(D)$ with $\Pi v = 0$. Therefore, we get with (3.12), $\Pi\tilde{\Lambda}|_R = 0$, (3.33), and (3.34) that

$$\|\nabla\tilde{\Lambda}\|^2_{L^2(D\setminus\mathsf{N}^{\ell+1}(K))} \le Cp\,\|\nabla\tilde{\Lambda}\|^2_{L^2(R)},$$

which leads to

$$\|\nabla\tilde{\Lambda}\|^2_{L^2(D\setminus\mathsf{N}^{\ell+1}(K))} \leq \frac{Cp}{Cp+1}\,\|\nabla\tilde{\Lambda}\|^2_{L^2(D\setminus\mathsf{N}^{\ell}(K))} \leq \left(\frac{Cp}{Cp+1}\right)^{\ell+1}\|\nabla\tilde{\Lambda}\|^2_{L^2(D)}$$

as in the proof of Theorem 2.4.1. We further obtain

$$\Big(\frac{Cp}{Cp+1}\Big)^{\ell} = \exp\big(-\,|\log\big(\tfrac{Cp}{Cp+1}\big)|\,\ell\big) \leq \exp\big(-\tfrac{1}{2C}\,\ell/p\big).$$

Taking the square root, we deduce (3.32) with $C_{\mathrm{dec}} := \frac{1}{4C}$ after a shift in ℓ. □

Remark 3.3.2. Although Theorem 3.3.1 only quantifies the decay of basis functions $\tilde{\Lambda} \in \tilde{\mathfrak{B}}$, with the same arguments the result also holds for any function $\mathcal{R}q$, where $q \in V_H^p(K)$ and $K \in \mathcal{T}_H$. That is, we have

$$\|\nabla\mathcal{R}q\|_{L^2(D\setminus\mathsf{N}^{\ell}(K))} \lesssim \exp(-C_{\mathrm{dec}}\,\ell/p)\,\|\nabla\mathcal{R}q\|_{L^2(D)}. \tag{3.35}$$

Remark 3.3.3. The p-dependence in Theorem 3.3.1 seems pessimistic and could possibly be improved. If, for instance, $\nabla\eta$ is only supported on a portion of the ring R in the proof of Theorem 3.3.1, one could expect some additional (fractional) powers of p in the estimate (3.34) in the sense of

$$\|\nabla b\|_{L^2(R)} \lesssim \frac{p^2}{H}\|\tilde{\Lambda}\eta\|_{L^2(R)} \lesssim \frac{p^2}{H}\,p^{-\delta}\,\|\tilde{\Lambda}\|_{L^2(R)}$$

for some $\delta > 0$. However, decreasing the support of $\nabla\eta$ has an influence on its L^∞-bound. For that matter, one may relax the restriction in (3.33) to

$$\|\nabla\eta\|_{L^\infty(D)} \leq C_\eta\,H^{-1}p$$

without an impact on the final estimates in the proof of Theorem 3.3.1 and possibly even further dependent on δ.

The decay property of the basis functions in $\mathfrak{B}$ that is proven in Theorem 3.3.1 is the key ingredient to define a localized version of the operator $\mathcal{R}$. This localization procedure is explained and investigated in the following subsection.

3.3.2 Localized computation of the approximation space

As in Chapter 2, we base the definition of a localized operator $\mathcal{R}^\ell$ on truncated versions of the basis functions in $\tilde{\mathfrak{B}}$. Thus, for a given oversampling parameter $\ell \in \mathbb{N}$ and any $\Lambda \in \mathfrak{B}$ with $\mathrm{supp}(\Lambda) = K \in \mathcal{T}_H$, we define $\tilde{\Lambda}^\ell \in H^1_0(\mathsf{N}^\ell(K))$ as the unique solution of the saddle point problem

$$\begin{array}{lclcl} a(\tilde{\Lambda}^\ell, v) & + & (\lambda^\ell_\Lambda, v)_{L^2(D)} & = & 0, \\ (\tilde{\Lambda}^\ell, \mu_H)_{L^2(D)} & & & = & (\Lambda, \mu_H)_{L^2(D)} \end{array} \tag{3.36}$$

for all $v \in H_0^1(\mathsf{N}^\ell(K))$ and $\mu_H \in V_H^p(\mathsf{N}^\ell(K))$ with associated Lagrange multiplier $\lambda_\Lambda^\ell \in V_H^p(\mathsf{N}^\ell(K))$. Then for any function $v_H \in V_H^p$ which can be expanded as

$$v_H = \sum_{K \in \mathcal{T}_H} \sum_{j=1}^{m_K} c_{K,j}\, \Lambda_{K,j},$$

we define the corresponding function $\mathcal{R}^\ell v_H \in H_0^1(D)$ by

$$\mathcal{R}^\ell v_H := \sum_{K \in \mathcal{T}_H} \sum_{j=1}^{m_K} c_{K,j}\, \tilde{\Lambda}_{K,j}^\ell. \tag{3.37}$$

We set $\tilde{V}_H^{p,\ell} := \mathcal{R}^\ell V_H^p$ and remark that $\tilde{\mathfrak{B}}^\ell := \mathcal{R}^\ell \mathfrak{B}$ is a basis of $\tilde{V}_H^{p,\ell}$ by construction. We use this space to compute an approximation of the ideal finite-dimensional solution $\tilde{u}_H \in \tilde{V}_H^p$ of (3.28), i.e., we want to find $\tilde{u}_H^\ell \in \tilde{V}_H^{p,\ell}$ that solves

$$a(\tilde{u}_H^\ell, \tilde{v}_H) = (f, \tilde{v}_H)_{L^2(D)} \tag{3.38}$$

for all $\tilde{v}_H \in \tilde{V}_H^{p,\ell}$. With regard to Chapter 2, we refer to $\tilde{u}_H^\ell$ as the *LOD solution.* As a next step, we show an error estimate for the error $u - \tilde{u}_H^\ell$.

Theorem 3.3.4 (Error of the LOD method). *Let $\ell \in \mathbb{N}$, $f \in H^k(\mathcal{T}_H)$, $k \in \mathbb{N}_0$, and define $s := \min\{k, p+1\}$. Further, let $u \in H_0^1(D)$ be the solution of* (3.7) *and $\tilde{u}_H^\ell \in \tilde{V}_H^{p,\ell}$ the solution of* (3.38). *Then it holds that*

$$\begin{aligned}\|\nabla(u - \tilde{u}_H^\ell)\|_{L^2(D)} & \\ &\lesssim \frac{\Phi(p,s)}{p}\, H^{s+1}\, |f|_{H^s(\mathcal{T}_H)} + \frac{p^3}{H}\, \ell^{(d-1)/2} \exp(-C_{\mathrm{dec}}\, \ell/p)\, \|f\|_{L^2(D)}\end{aligned} \tag{3.39}$$

with the constant C_{dec} from Theorem 3.3.1.

Proof. First, we observe that $\tilde{u}_H^\ell$ is quasi-optimal by the Galerkin orthogonality; see also Lemma 2.2.1. Therefore, we obtain

$$\|\nabla(u - \tilde{u}_H^\ell)\|_{L^2(D)} \le \frac{\beta}{\alpha} \inf_{\tilde{v}_H \in \tilde{V}_H^{p,\ell}} \|\nabla(u - \tilde{v}_H)\|_{L^2(D)} \le \frac{\beta}{\alpha}\, \|\nabla(u - \bar{u}_H^\ell)\|_{L^2(D)},$$

where $\bar{u}_H^\ell := \mathcal{R}^\ell \Pi u \in \tilde{V}_H^{p,\ell}$. With the triangle inequality and the solution $\tilde{u}_H \in \tilde{V}_H^p$ of (3.28), we get that

$$\|\nabla(u - \bar{u}_H^\ell)\|_{L^2(D)} \le \|\nabla(u - \tilde{u}_H)\|_{L^2(D)} + \|\nabla(\tilde{u}_H - \bar{u}_H^\ell)\|_{L^2(D)}. \tag{3.40}$$

The first term can be estimated with Theorem 3.2.6, i.e.,

$$\|\nabla(u - \tilde{u}_H)\|_{L^2(D)} \lesssim \frac{\Phi(p,s)}{p}\, H^{s+1}\, |f|_{H^s(\mathcal{T}_H)}.$$

For the second term, we set $w := \tilde{u}_H - \bar{u}_H^\ell$. Further, for $K \in \mathcal{T}_H$ we define a cutoff function $\eta_K \in W^{1,\infty}(D)$ with

$$\begin{aligned} &0 \leq \eta_K \leq 1, \\ &\eta_K \,= 0 \quad \text{in } \mathsf{N}^{\ell-1}(K), \\ &\eta_K \,= 1 \quad \text{in } D \setminus \mathsf{N}^{\ell}(K), \\ &\|\nabla \eta_K\|_{L^\infty(D)} \,\leq C_\eta \, H^{-1}. \end{aligned}$$

We set $R_K := \mathsf{N}^\ell(K) \setminus \mathsf{N}^{\ell-1}(K)$. By (3.36) and (3.37), for each $K \in \mathcal{T}_H$ there exists a Lagrange multiplier $\lambda_K^\ell \in V_H^p(\mathsf{N}^\ell(K))$ such that

$$\begin{aligned} a(\mathcal{R}^\ell(\Pi u|_K), v) \quad &+ \quad (\lambda_K^\ell, v)_{L^2(D)} &&= \quad 0, \\ (\mathcal{R}^\ell(\Pi u|_K), \mu_H)_{L^2(D)} & &&= \quad (\Pi u|_K, \mu_H)_{L^2(D)} \end{aligned} \tag{3.41}$$

for all $v \in H_0^1(\mathsf{N}^\ell(K))$ and $\mu_H \in V_H^p(\mathsf{N}^\ell(K))$. Noting that

$$(1 - \eta_K) w \in H_0^1(\mathsf{N}^\ell(K)) \quad \text{and} \quad \Pi w = 0,$$

we obtain with (3.25) and (3.41)

$$\begin{aligned} \alpha \, \|\nabla w\|_{L^2(D)}^2 &\leq \sum_{K \in \mathcal{T}_H} a(\mathcal{R}(\Pi u|_K) - \mathcal{R}^\ell(\Pi u|_K), w) \\ &= \sum_{K \in \mathcal{T}_H} -a(\mathcal{R}^\ell(\Pi u|_K), (1 - \eta_K) w + \eta_K w) \\ &= \sum_{K \in \mathcal{T}_H} \Big((\lambda_K^\ell, (1 - \eta_K) w)_{L^2(R_K)} - a(\mathcal{R}^\ell(\Pi u|_K), \eta_K w) \Big) \\ &\lesssim \sum_{K \in \mathcal{T}_H} \Big(\|\lambda_K^\ell\|_{L^2(R_K)} \, \|w\|_{L^2(R_K)} \\ &\qquad\qquad + \|\nabla \mathcal{R}^\ell(\Pi u|_K)\|_{L^2(R_K)} \, \|\nabla(\eta_K w)\|_{L^2(R_K)} \Big) \\ &\lesssim \sum_{K \in \mathcal{T}_H} (p+1) \, \|\nabla \mathcal{R}^\ell(\Pi u|_K)\|_{L^2(R_K)} \, \|\nabla w\|_{L^2(R_K)}. \end{aligned} \tag{3.42}$$

In the last step, we use that

$$\|w\|_{L^2(R_K)} \leq C_\Pi \frac{H}{p} \, \|\nabla w\|_{L^2(R_K)}$$

by the approximation result (3.12),

$$\|\nabla(\eta_K w)\|_{L^2(R_K)} \leq C_\eta C_\Pi \, p^{-1} \, \|\nabla w\|_{L^2(R_K)} + \|\nabla w\|_{L^2(R_K)},$$

and

$$\|\lambda_K^\ell\|_{L^2(R_K)} \lesssim H^{-1} p^2 \, \|\nabla \mathcal{R}^\ell(\Pi u|_K)\|_{L^2(R_K)}.$$

The last estimate follows from the arguments in the proof of Lemma 3.2.1. More precisely, for any $T \in \mathcal{T}_H$, there exists a bubble function $\rho_T \in W^{1,\infty}(T) \cap H_0^1(T)$ as in the proof of Lemma 3.2.1 such that

$$\begin{aligned}\|\lambda_K^\ell\|_{L^2(T)}^2 &\leq 4\,(\lambda_K^\ell, \rho_T \lambda_K^\ell)_{L^2(T)} \\ &= -4\,a(\mathcal{R}^\ell(\Pi u|_K), \rho_T \lambda_K^\ell) \\ &\lesssim H^{-1} p^2\, \|\nabla \mathcal{R}^\ell(\Pi u|_K)\|_{L^2(T)}\, \|\lambda_K^\ell\|_{L^2(T)},\end{aligned} \tag{3.43}$$

where we employ the estimates (3.17) and (3.18).

Using Theorem 3.3.1 and Remark 3.3.2, which both equivalently hold with $\mathcal{R}$ replaced by $\mathcal{R}^\ell$, we get with (3.42) and (3.35) that

$$\begin{aligned}\|\nabla w\|_{L^2(D)}^2 &\lesssim \sum_{K\in\mathcal{T}_H} (p+1)\, \|\nabla \mathcal{R}^\ell(\Pi u|_K)\|_{L^2(D\setminus \mathsf{N}^{\ell-1}(K))}\, \|\nabla w\|_{L^2(R_K)} \\ &\lesssim \frac{p^3}{H} \exp(-C_{\mathrm{dec}}\,\ell/p) \Big(\sum_{K\in\mathcal{T}_H} \|\Pi u|_K\|_{L^2(K)}^2 \Big)^{1/2} \Big(\sum_{K\in\mathcal{T}_H} \|\nabla w\|_{L^2(R_K)}^2 \Big)^{1/2} \\ &\lesssim \frac{p^3}{H}\, \ell^{(d-1)/2} \exp(-C_{\mathrm{dec}}\,\ell/p)\, \|\Pi u\|_{L^2(D)}\, \|\nabla w\|_{L^2(D)}.\end{aligned}$$

Here, we employ the discrete Cauchy-Schwarz inequality and the stability of (3.41), i.e.,

$$\|\nabla \mathcal{R}^\ell(\Pi u|_K)\|_{L^2(D)} \lesssim \frac{p^2}{H}\, \|\Pi u|_K\|_{L^2(K)}$$

for any $K \in \mathcal{T}_H$. We now go back to (3.40) and obtain

$$\begin{aligned}&\|\nabla(u - \tilde{u}_H^\ell)\|_{L^2(D)} \\ &\quad\lesssim \frac{\Phi(p,s)}{p} H^{s+1}\, |f|_{H^s(\mathcal{T}_H)} + \frac{p^3}{H}\, \ell^{(d-1)/2} \exp(-C_{\mathrm{dec}}\,\ell/p)\, \|\Pi u\|_{L^2(D)} \\ &\quad\lesssim \frac{\Phi(p,s)}{p} H^{s+1}\, |f|_{H^s(\mathcal{T}_H)} + \frac{p^3}{H}\, \ell^{(d-1)/2} \exp(-C_{\mathrm{dec}}\,\ell/p)\, \|f\|_{L^2(D)},\end{aligned}$$

where we use the stability of Π and (3.8). This completes the proof. □

Remark 3.3.5. The additional H in the denominator of the estimate in Theorem 3.3.4 may be explained by the fact that the localization error $\tilde{u}_H - \bar{u}_H^\ell$ is measured in the H^1-norm while Πu is measured in the L^2-norm. Although this seems suboptimal, the pollution in terms of H in the second term of (3.39) is also observed in our numerical experiments; see Section 3.4.

We can now use Theorem 3.3.4 to quantify the choice of the oversampling parameter ℓ with respect to the polynomial degree p and the mesh size H dependent on the regularity of the right-hand side f.

Corollary 3.3.6. *Let $f \in H^k(\mathcal{T}_H)$, $k \in \mathbb{N}_0$, and define $s := \min\{k, p+1\}$. Further, let $u \in H^1_0(D)$ be the solution of* (3.7)*, and $\tilde{u}^\ell_H \in \tilde{V}^{p,\ell}_H$ the solution of* (3.38)*. Then, for*

$$\ell \gtrsim |\log H|\, p\,(s+1) + (\log p)\, p\,(s+1), \tag{3.44}$$

it holds that

$$\|\nabla(u - \tilde{u}^\ell_H)\|_{L^2(D)} \lesssim \frac{\Phi(p,s)}{p}\, H^{s+1}\, |f|_{H^s(\mathcal{T}_H)} + \Big(\frac{H}{p}\Big)^{s+1} \|f\|_{L^2(D)}.$$

Note that if $k = 0$ and $p = 1$, Corollary 3.3.6 provides a similar error estimate as in the conforming case of Chapter 2 with the same scaling of ℓ. Of course, if we increase p, the oversampling parameter ℓ in Theorem 3.3.4 needs to grow as well in order to maintain the high convergence rate of Theorem 3.2.6 with respect to H and p. Nevertheless, the experiments in Section 3.4 indicate that the p-dependence of ℓ in (3.44) might be too pessimistic and the decay property of Theorem 3.3.1 even slightly improves for larger values of p. Before we turn our attention to these numerical investigations of the higher-order method, we first need to discuss the last step towards a fully practical method, i.e., the discretization at the microscopic scale.

3.3.3 Microscopic discretization

As discussed in Section 2.4.4, the localized operator $\mathcal{R}^\ell$ does not provide a fully discrete method since the localized basis functions (3.36) are obtained by solving infinite-dimensional auxiliary problems. The easiest approach to resolve this issue is to introduce a (conforming) fine FE space $V_{h,p'} \subseteq H^1_0(D)$ based on a decomposition $\mathcal{T}_h$ with mesh parameter h and polynomial degree p' that replaces the space $H^1_0(D)$ in the above construction. Ideally, the classical cG solution in $V_{h,p'}$ should fulfill an estimate similar to the one in Theorem 3.2.6. Motivated by error estimates of the hp FE method (see, e.g., [BG96, Sch98]), for $f \in H^k(D)$, $k \in \mathbb{N}_0$, we assume that

$$\|\nabla(u - u_h)\|_{L^2(D)} \lesssim \frac{\Phi(p',s')}{p'}\, (C_\epsilon\, h)^{s'+1}\, |f|_{H^{s'}(D)}, \tag{3.45}$$

where $u \in H^1_0(D)$ is the solution of (3.7), $s' := \min\{k, p'+1\}$, and $u_h \in V_{h,p'}$ is the solution of

$$a(u_h, v_h) = (f, v_h)_{L^2(D)} \tag{3.46}$$

for all $v_h \in V_{h,p'}$. Note that the right-hand side of (3.45) depends on the fine-scale parameter ϵ through the constant C_ϵ. This is typical for classical FE spaces which do not take into account microscopic information.

We emphasize that on the one hand, the ideal approximation $\tilde{u}_H \in \tilde{V}^p_H$ characterized by (3.28) fulfills the higher-order estimate quantified in Theorem 3.2.6

by the (piecewise) regularity of the right-hand side f only. On the other hand, in order to obtain a higher-order estimate of the form (3.45) for the classical FE space $V_{h,p'}$, the regularity of f needs to hold globally. Further, one requires additional smoothness assumptions on the domain D as well as on the coefficient A (see, e.g., [Eva10, Thm. 5 in Sec. 6.3]) and, in particular, the microscopic scale ϵ needs to be resolved. Another problem that occurs when discretizing the fine scales is the fact that the proof of the inf-sup condition in Lemma 3.2.1 is explicitly based on the space $H_0^1(D)$. The result does not directly follow for subspaces of $H_0^1(D)$ and a similar inf-sup condition needs to be proven for the respective discrete space $V_{h,p'}$ at hand.

With these problems in mind, the following lemma provides a condition on the fine mesh parameter h for which the inf-sup condition (3.16) and thus the surjectivity results in Theorem 3.2.2 and Corollary 3.2.3 remain valid if $H_0^1(D)$ is replaced by the first-order space $V_h \subseteq H_0^1(D)$, for which we omit the subscript 1. The explicit choice of the polynomial degree $p' = 1$ is motivated by the fact that higher-order estimates for the classical conforming FE space $V_{h,p'}$ would require additional smoothness assumptions as mentioned above.

Lemma 3.3.7 (Discrete local inf-sup condition)**.** *Let $K \in \mathcal{T}_H$. Then there exists a constant $C > 0$ independent of h, H, and p such that for*

$$h \leq C\,Hp^{-2}$$

the inf-sup condition

$$\inf_{q\in V_H^p(K)} \sup_{v_h\in V_h\cap H_0^1(K)} \frac{(q,v_h)_{L^2(K)}}{\|q\|_{L^2(K)}\,\|\nabla v_h\|_{L^2(K)}} \geq \gamma_h > 0 \tag{3.47}$$

holds with $\gamma_h \sim Hp^{-2}$.

Proof. As in the proof of Lemma 3.2.1, let $\kappa \subseteq K$ be such that its edges, faces, etc. are parallel to the ones of K, $\operatorname{dist}(\kappa, \partial K) = C_{\text{dist}}\,Hp^{-2}$, and

$$\|q\|_{L^2(\kappa)}^2 \geq \frac{1}{4}\|q\|_{L^2(K)}^2 \tag{3.48}$$

for all $q \in V_H^p(K)$. Now, let $\rho \in W^{1,\infty}(K) \cap H_0^1(K)$ with

$$\begin{aligned} 0 \,&\leq\, \rho \,\leq\, 1,\\ \rho \,&=\, 1 \quad \text{in } \kappa,\\ \|\nabla\rho\|_{L^\infty(K)} \,&\leq\, C_\rho\, H^{-1}p^2, \end{aligned}$$

where C_ρ depends on C_{dist}. Next, we define for any $q \in V_H^p(K)$ the function $w_q \in V_h \cap H_0^1(K)$ as the solution of

$$(w_q, v_h)_{L^2(K)} = (q, v_h)_{L^2(K)}$$

for all $v_h \in V_h \cap H^1_0(K)$. Note that w_q is unique by the inverse inequality

$$\|\nabla v_h\|_{L^2(K)} \leq C_{\mathrm{inv},h}\, h^{-1}\, \|v_h\|_{L^2(K)}$$

and the Lax-Milgram Theorem. The last auxiliary ingredient is an estimate of the form

$$\|q\|_{L^2(K)} \lesssim \|w_q\|_{L^2(K)}$$

which can be obtained using a projection operator $\mathcal{I}_h^K \colon L^2(K) \to V_h \cap H^1_0(K)$ which fulfills stability and approximation properties as in (2.11) and (2.14). That is, for all $v \in L^2(K)$, it holds that

$$\|\mathcal{I}_h^K v\|_{L^2(K)} \leq C_{\mathcal{I}_h^K} \|v\|_{L^2(K)}$$

and, for any $v \in H^1_0(K)$, we have

$$\|h^{-1}(v - \mathcal{I}_h^K v)\|_{L^2(K)} + \|\nabla \mathcal{I}_h^K v\|_{L^2(K)} \leq C_{\mathcal{I}_h^K}\, \|\nabla v\|_{L^2(K)}.$$

For an explicit choice of $\mathcal{I}_h^K$, see Section 2.5.1. With the above inequalities, we can show that

$$\begin{aligned}
\tfrac{1}{4}\|q\|^2_{L^2(K)} \leq \|q\|^2_{L^2(\kappa)} \leq (q, \rho q)_{L^2(K)} &= (q, \mathcal{I}_h^K(\rho q))_{L^2(K)} + (q, (\mathtt{id} - \mathcal{I}_h^K)(\rho q))_{L^2(K)}\\
&= (w_q, \mathcal{I}_h^K(\rho q))_{L^2(K)} + (q, (\mathtt{id} - \mathcal{I}_h^K)(\rho q))_{L^2(K)}\\
&\leq \|w_q\|_{L^2(K)}\, C_{\mathcal{I}_h^K}\, \|\rho q\|_{L^2(K)} + \|q\|_{L^2(K)}\, C_{\mathcal{I}_h^K} h\, \|\nabla(\rho q)\|_{L^2(K)}\\
&\leq C_{\mathcal{I}_h^K}\, \|w_q\|_{L^2(K)}\, \|q\|_{L^2(K)} + C_{\mathcal{I}_h^K}(C_\rho + C_{\mathrm{inv}})\, hH^{-1}p^2\, \|q\|^2_{L^2(K)}
\end{aligned}$$

and thus

$$\|q\|_{L^2(K)} \leq 8\, C_{\mathcal{I}_h^K}\, \|w_q\|_{L^2(K)}$$

provided that

$$C_{\mathcal{I}_h^K}(C_\rho + C_{\mathrm{inv}})\, hH^{-1}p^2 \leq \frac{1}{8}.$$

With all the above estimates, it holds that

$$\begin{aligned}
\inf_{q \in V_H^p(K)} \sup_{v_h \in V_h \cap H^1_0(K)} \frac{(q, v_h)_{L^2(K)}}{\|q\|_{L^2(K)}\, \|\nabla v_h\|_{L^2(K)}} &\geq \inf_{q \in V_H^p(K)} \frac{(w_q, w_q)_{L^2(K)}}{\|q\|_{L^2(K)}\, \|\nabla w_q\|_{L^2(K)}}\\
&\geq \inf_{q \in V_H^p(K)} \frac{1}{8\, C_{\mathcal{I}_h^K}} \frac{\|w_q\|^2_{L^2(K)}}{\|w_q\|_{L^2(K)}\, \|\nabla w_q\|_{L^2(K)}}\\
&\geq \frac{h}{8\, C_{\mathcal{I}_h^K} C_{\mathrm{inv},h}} =: \gamma_h > 0.
\end{aligned}$$

For $h \sim Hp^{-2}$, this is the assertion. For $h \lesssim Hp^{-2}$, there exists an auxiliary $h' \sim Hp^{-2}$ such that $V_{h'} \subseteq V_h$ and thus

$$\begin{aligned}
&\inf_{q \in V_H^p(K)} \sup_{v_h \in V_h \cap H^1_0(K)} \frac{(q, v_h)_{L^2(K)}}{\|q\|_{L^2(K)}\, \|\nabla v_h\|_{L^2(K)}}\\
&\geq \inf_{q \in V_H^p(K)} \sup_{v_{h'} \in V_{h'} \cap H^1_0(K)} \frac{(q, v_{h'})_{L^2(K)}}{\|q\|_{L^2(K)}\, \|\nabla v_{h'}\|_{L^2(K)}}\\
&\geq \gamma_{h'} \sim Hp^{-2}.
\end{aligned}$$

This completes the proof. □

With Lemma 3.3.7, we can replace $H_0^1(D)$ (and the solution $u \in H_0^1(D)$ of (3.7)) in the construction of this chapter by a conforming Q_1 FE space V_h (and the classical cG approximation $u_h \in V_h$) provided that h is sufficiently small with respect to H and p and, additionally, resolves the microscopic information on the scale ϵ. This is quantified with the resolution conditions

$$C_\epsilon\, h \lesssim \frac{\Phi(p,s)}{p}\, H^{s+1} \quad \text{and} \quad h \lesssim Hp^{-2}, \tag{3.49}$$

where the constant C_ϵ indicates the dependence on the microscopic scale ϵ as in (3.45). While a resolution condition on h with respect to H and p of the form $h \leq Hp^{-s}$ for some $s \geq 1$ seems natural to resolve higher-order functions, the left condition in (3.49) is mainly motivated by the aim to retain the convergence properties with respect to H and p as derived in the previous subsections. In a more practical manner, one could alternatively prescribe some certain tolerance and balance h, p', H, and p such that the given tolerance is reached with the respective approximation. We remark that a discrete inf-sup condition as in Lemma 3.3.7 may also be obtained for a higher-order conforming FE space and relaxes the resolution condition $h \lesssim Hp^{-2}$ dependent on the choice of p'. If additional smoothness conditions hold, the use of a higher-order space can further provide a relaxation of the left resolution condition in (3.49) on h if p' is suitably coupled to h and ϵ in the spirit of [PS12, Cor. 5.3].

Although such higher-order constructions may generally be considered for the fine discretization, we restrict ourselves to the first-order setting with $p' = 1$ which only requires minimal regularity assumptions. Similar to the notation in Section 2.4.4, we introduce the additional parameter h in the above construction if $H_0^1(D)$ is replaced by V_h, i.e., we write

$$\mathcal{R}_h,\, \mathcal{R}_h^\ell,\, \tilde{V}_{H,h}^p,\, \tilde{V}_{H,h}^{p,\ell} \quad \text{instead of} \quad \mathcal{R},\, \mathcal{R}^\ell,\, \tilde{V}_H^p,\, \tilde{V}_H^{p,\ell}.$$

Further, the solution $\tilde{u}_{H,h} \in \tilde{V}_{H,h}^{p,\ell}$ of the *fully discrete LOD method* is determined by

$$a(\tilde{u}_{H,h}^\ell, \tilde{v}_{H,h}) = (f, \tilde{v}_{H,h})_{L^2(D)} \tag{3.50}$$

for all $\tilde{v}_{H,h} \in \tilde{V}_{H,h}^{p,\ell}$. The error of the fully discrete approach is quantified in the next theorem.

Theorem 3.3.8 (Error of the fully discrete LOD method)**.** *Assume $f \in H^k(\mathcal{T}_H)$, $k \in \mathbb{N}_0$, and let $s := \min\{k, p+1\}$. Further, suppose that the resolution conditions* (3.49) *hold and let $u \in H_0^1(D)$ be the solution of* (3.7) *and $\tilde{u}_{H,h}^\ell \in \tilde{V}_{H,h}^{p,\ell}$ the solution of* (3.50)*. Then, with the choice*

$$\ell \gtrsim |\log H|\, p\,(s+1) + (\log p)\, p\,(s+1),$$

it holds that

$$\|\nabla(u - \tilde{u}_{H,h}^\ell)\|_{L^2(D)} \lesssim \frac{\Phi(p,s)}{p}\, H^{s+1}\left(\|f\|_{L^2(D)} + |f|_{H^s(\mathcal{T}_H)}\right) + \left(\frac{H}{p}\right)^{s+1} \|f\|_{L^2(D)}.$$

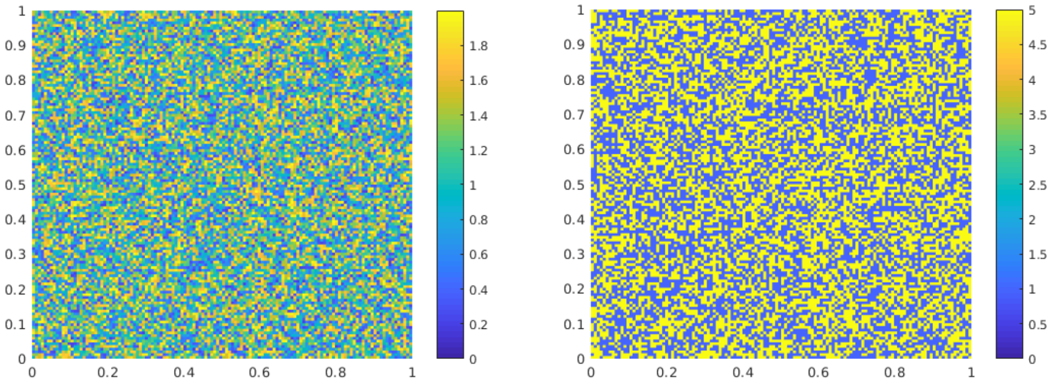


Figure 3.1: Multiscale coefficients A_1 (left) and A_2 (right) on the scale $\epsilon = 2^{-7}$.

Proof. The assertion follows from a simple triangle inequality, the estimate (3.45) with $s' = 0$, the resolution conditions (3.49), and Corollary 3.3.6 in the case where $H^1_0(D)$ is replaced by V_h. To be more precise, with the solution $u_h \in V_h$ of (3.46), we obtain

$$\begin{aligned}\|\nabla(u-\tilde{u}^\ell_{H,h})\|_{L^2(D)} &\leq \|\nabla(u-u_h)\|_{L^2(D)} + \|\nabla(u_h-\tilde{u}^\ell_{H,h})\|_{L^2(D)} \\ &\lesssim C_\epsilon\, h\,\|f\|_{L^2(D)} + \frac{\Phi(p,s)}{p}\,H^{s+1}\,|f|_{H^s(\mathcal{T}_H)} + \Big(\frac{H}{p}\Big)^{s+1}\|f\|_{L^2(D)} \\ &\lesssim \frac{\Phi(p,s)}{p}\,H^{s+1}\left(\|f\|_{L^2(D)} + |f|_{H^s(\mathcal{T}_H)}\right) + \Big(\frac{H}{p}\Big)^{s+1}\|f\|_{L^2(D)}\end{aligned}$$

employing the estimates mentioned above. □

3.4 Numerical experiments

In this section, we present some examples to verify the results of the previous sections. As in Section 2.5, we remark that if the exact solution $u \in H^1_0(D)$ of (3.7) is not explicitly given, only the errors between the discrete solutions $u_h \in V_h$ of (3.46) and $\tilde{u}^\ell_{H,h} \in \tilde{V}^{p,\ell}_{H,h}$ of (3.50) can be measured. Thus, we need to pose the assumption that the chosen mesh parameter h is indeed small enough as quantified in Section 3.3.3, and use u_h as the reference solution. As before, we measure the errors in the energy norm $\|\cdot\|_a := \|A^{1/2}\nabla\cdot\|_{L^2(D)}$.

3.4.1 Two-dimensional examples

For the experiments of this subsection, we consider the domain $D = (0,1)^2$ as well as the two scalar diffusion coefficients A_1 and A_2 as depicted in Figure 3.1. These coefficients are piecewise constant on a mesh $\mathcal{T}_\epsilon$ with mesh parameter $\epsilon = 2^{-7}$. In each element $K \in \mathcal{T}_\epsilon$, the value of A_1 is obtained as a uniformly

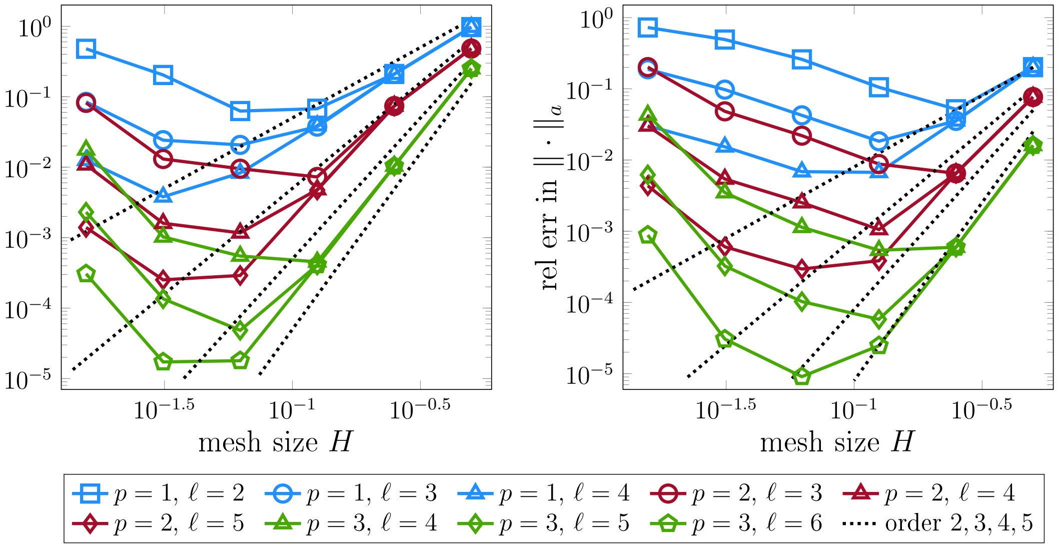


Figure 3.2: Errors of the higher-order LOD in the relative energy norm for the first (left) and the second model (right) with respect to H for different values of ℓ and p.

distributed random number in $[0.2, 2]$. Similarly, A_2 takes values in $\{1, 5\}$. Further, we take the right-hand sides

$$f_1(x) = \sin(5\pi\, x_1)\cos(3\pi\, x_2)$$

and

$$f_2(x) = (x_1 + \sin(3\pi\, x_1))\, x_2 \cos(\pi\, x_2).$$

For the first model, we choose the coefficient $A = A_1$ and the right-hand side $f = f_1$ in (3.7) and compute the solution $\tilde{u}^{\ell}_{H,h} \in \tilde{V}^{p,\ell}_{H,h}$ of (3.50) for multiple choices of the polynomial degree p and the localization parameter ℓ. The relative energy errors of these approximations with respect to the reference solution on the scale $h = 2^{-9}$ are depicted in Figure 3.2 (left). Similarly, we present the energy errors for the second model with the coefficient A_2 and the right-hand side f_2 in Figure 3.2 (right), where again $h = 2^{-9}$. The error curves in both examples show a convergence rate between $p + 1$ and $p + 2$ with respect to H for different polynomial degrees p if ℓ is chosen large enough. These results are in line with the findings in Theorem 3.3.4 which predicts a convergence rate of up to order $p + 2$ in H dependent on the regularity of f and provided that the second term in the estimate (3.39) is small enough. For the first model, we also provide the relative errors in Table 3.1 as well as the respective experimental orders of convergence (EOCs). For two mesh sizes $H_1 > H_2$ with corresponding errors e_1 and e_2, the EOC is defined by EOC $:= \log\left(\frac{e_1}{e_2}\right) / \log\left(\frac{H_1}{H_2}\right)$.

Apart from the observed higher-order rates for appropriate parameter regimes, the two examples also indicate that there might be a pollution in terms of some

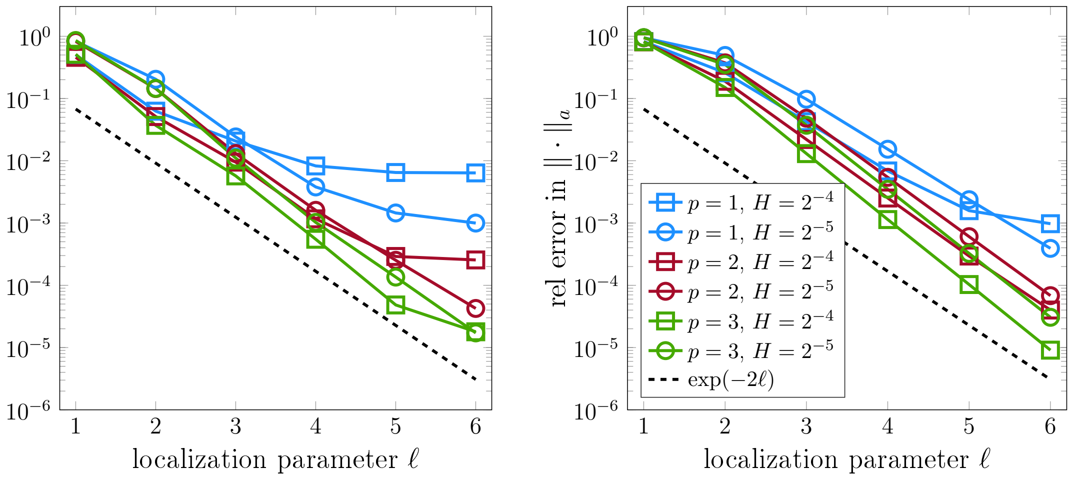


Figure 3.3: Errors of the higher-order LOD in the relative energy norm for the first (left) and second model (right) with respect to ℓ for different values of H and p.

negative power of H as obtained from the theory. That is, instead of a stagnation of the error curve for smaller H, the overall error grows again if ℓ is not chosen appropriately. We further study this effect in Section 3.4.2.

For completeness, we present the errors of the LOD method also with respect to the localization parameter ℓ in Figure 3.3. The plots show the exponential convergence rate in ℓ as in the theory. The curves stagnate for larger values of ℓ where the localization error is small enough and the first term in the estimate (3.39) dominates the overall error.

Since the previous experiments indicate that the exponential convergence in ℓ even slightly improves when p is increased, we further investigate the sharpness of the decay estimate quantified in Theorem 3.3.1. To this end, for $H = 2^{-4}$ we choose an element $K \in \mathcal{T}_H$ in the middle of the domain and compute the relative energy error between the ideal multiscale basis functions $\tilde{\Lambda}_{K,j} := \mathcal{R}_h \Lambda_{K,j}$ and its localized versions $\tilde{\Lambda}^{\ell}_{K,j} := \mathcal{R}^{\ell}_h \Lambda_{K,j}$ for different values of ℓ and $j \in \{1, \dots, m_K\}$. For the first model, Figure 3.4 (left) shows the decay of the localization error for different basis functions with respect to ℓ. To be more precise, for each p, we show the localization error corresponding to the highest-order basis function $\Lambda_{K,j}$ (with maximal polynomial degree p in both components). The results seem to contradict the scaling in p as predicted by Theorem 3.3.1. Instead, the rate even slightly improves when the polynomial degree p is increased, which is possibly due to the fact that the decay estimates are not sharp as explained in Remark 3.3.3. In Figure 3.4 (right), we show the localization error for different ℓ and p corresponding to the respective lowest-order basis function, i.e., the one whose L^2-projection onto $V^p_H(K)$ is constant. Again, the curves show an error reduction when p is increased which is slightly amplified by ℓ. That is, these

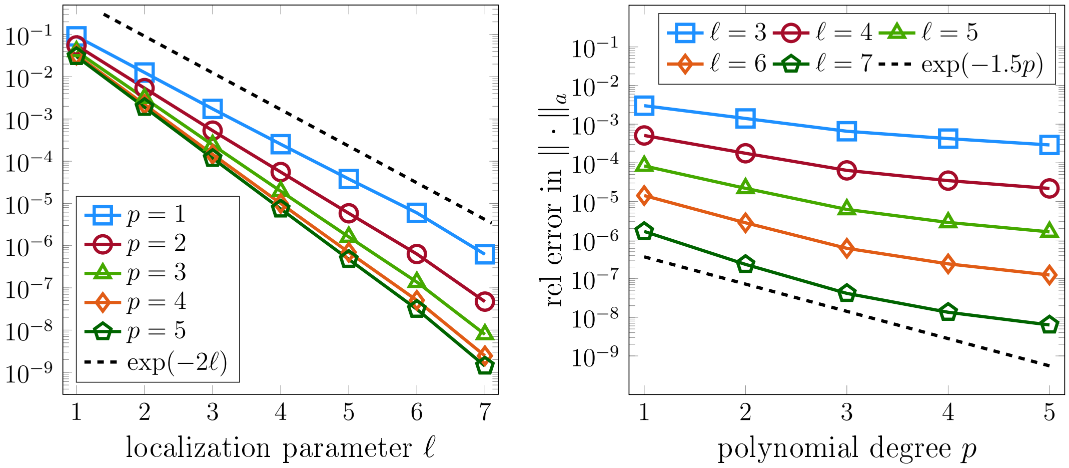


Figure 3.4: Localization errors of the higher-order LOD basis functions on the scale $H = 2^{-4}$ for the first model with respect to ℓ (left) and p (right) in the relative energy norm.

results also indicate a better scaling in p than quantified in Theorem 3.3.1. The commencing stagnation of the errors in Figure 3.4 (right) for larger p is probably related to the fact that $h = 2^{-9}$ is not fine enough to handle higher polynomial degrees. This issue is addressed in the following.

3.4.2 One-dimensional considerations

The aim of this subsection is to provide a study of the higher-order method in one spatial dimension. The motivation of this is the fact that the resolution conditions derived above require h to be much smaller than H subject to the choice of p. In this regard, the setting of this subsection allows us to compute LOD solutions with higher polynomial degree p and smaller H.

We consider a coefficient A which is piecewise constant on the scale $\epsilon = 2^{-12}$ with element-wise randomly chosen values in $[0.5, 10]$. Further, we set

$$f(x) = \sin(5\pi\, x).$$

With respect to the mesh size H, the errors of the higher-order LOD compared to a fine-scale solution on the scale $h = 2^{-16}$ are depicted in Figure 3.5. The plot seems to confirm the higher-order decay as quantified in Theorem 3.3.4 as well as the presence of a polluting term proportional to H^{-1}. The errors presented in Figure 3.6 with respect to p indicate that the dependence on p of the second term in (3.39) is probably too pessimistic and that there might even be some positive scaling with respect to p which is amplified by increasing values of ℓ. Lastly, we mention that the exponential convergence rate with respect to p as quantified in Corollary 3.3.6 is observed in Figure 3.6 provided that ℓ is

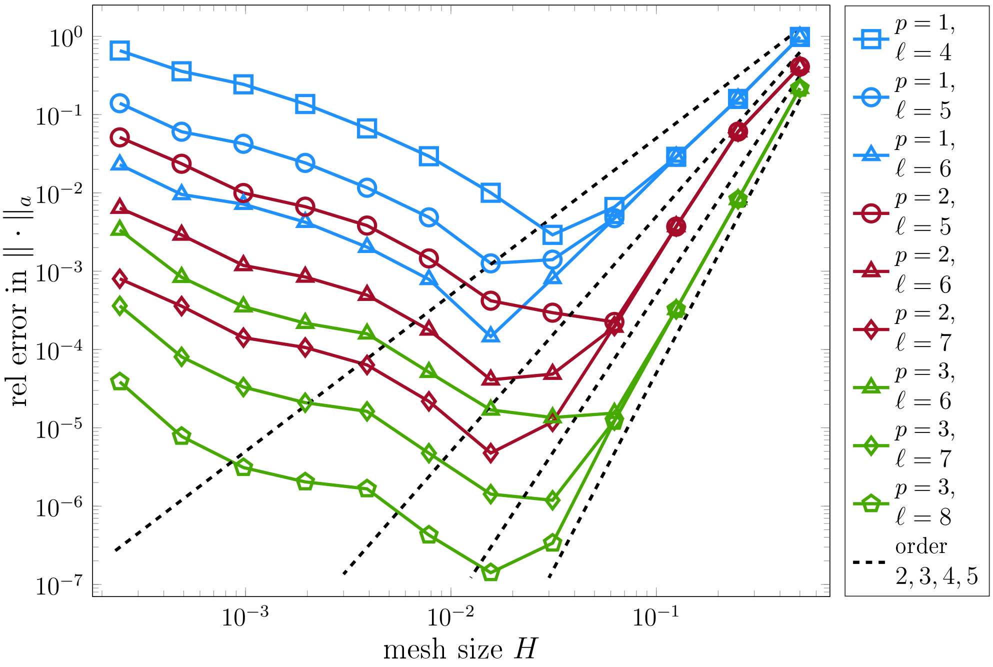


Figure 3.5: Relative errors of the higher-order LOD in one dimension in the energy norm with respect to H for different values of p and ℓ.

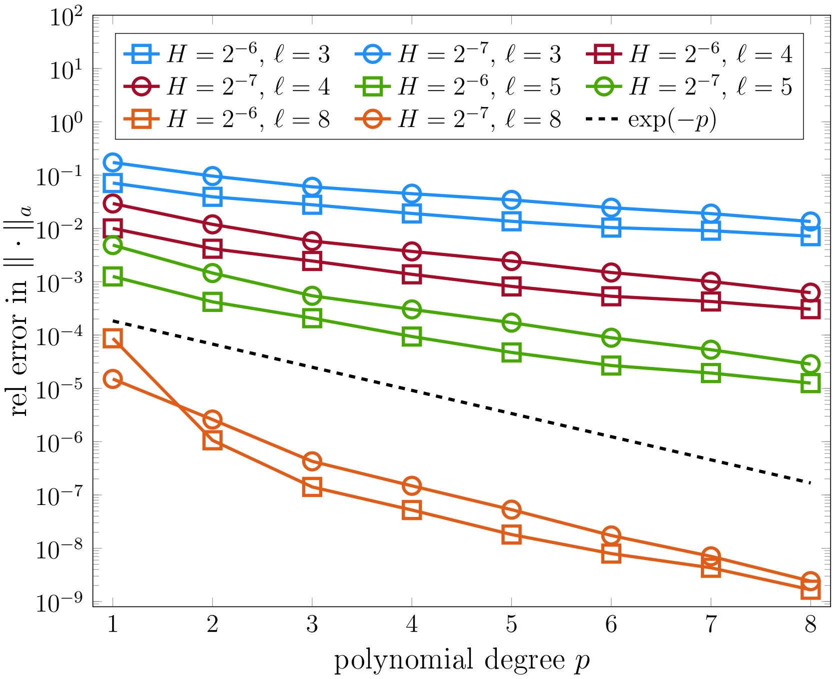


Figure 3.6: Relative errors of the higher-order LOD in one dimension in the energy norm with respect to p for different values of H and ℓ.

chosen large enough. If the localization parameter is too small, the error curve stagnates.

The numerical experiments of this section overall confirm the theoretical results for the higher-order construction considered in this chapter. The only deviation is in the scaling with respect to the polynomial degree p which seems to be better than predicted by the theory. That is, the result presented in Theorem 3.3.1 is most likely not sharp with respect to p and can possibly be improved. An enhanced estimate would also directly relax the condition on ℓ which is quantified in (3.44).

Note that although the approach numerically and theoretically shows a pollution of the total error for small mesh sizes H, this issue can be compensated for by a correct scaling of ℓ. Nevertheless, the method shows its best potential for relatively coarse mesh sizes which, combined with higher-order polynomials, already provide very good approximations. Moreover, the locality of the higher-order construction in principle allows us to even choose different polynomial degrees on the respective coarse elements.

Table 3.1: Relative errors and EOCs of the higher-order LOD for the first two-dimensional model in the energy norm for different values of the mesh size H, the polynomial degree p, and the localization parameter ℓ.

ℓ	H	$p=1$	$p=2$	$p=3$	$\mathrm{EOC}_{p=1}$	$\mathrm{EOC}_{p=2}$	$\mathrm{EOC}_{p=3}$
1	2^{-1}	0.97207	0.48187	0.25088	–	–	–
1	2^{-2}	0.21859	0.07310	0.09682	2.15	2.72	1.37
1	2^{-3}	0.20948	0.20747	0.19511	0.06	-1.51	-1.01
1	2^{-4}	0.50299	0.45793	0.51348	-1.26	-1.14	-1.40
1	2^{-5}	0.82963	0.82471	0.85400	-0.72	-0.85	-0.73
1	2^{-6}	0.95424	0.96866	0.98334	-0.20	-0.23	-0.20
2	2^{-1}	0.97207	0.48187	0.25088	–	–	–
2	2^{-2}	0.21018	0.07184	0.01204	2.21	2.75	4.38
2	2^{-3}	0.06752	0.03169	0.02515	1.64	1.18	-1.06
2	2^{-4}	0.06258	0.05128	0.03704	0.11	-0.69	-0.56
2	2^{-5}	0.20347	0.14435	0.14524	-1.70	-1.49	-1.97
2	2^{-6}	0.47862	0.50940	0.66576	-1.23	-1.82	-2.20
3	2^{-1}	0.97207	0.48187	0.25088	–	–	–
3	2^{-2}	0.20940	0.07370	0.01037	2.21	2.71	4.60
3	2^{-3}	0.03759	0.00726	0.00260	2.48	3.34	2.00
3	2^{-4}	0.02069	0.00950	0.00571	0.86	-0.39	-1.14
3	2^{-5}	0.02416	0.01306	0.01103	-0.22	-0.46	-0.95
3	2^{-6}	0.08455	0.08169	0.12815	-1.81	-2.65	-3.54
4	2^{-1}	0.97207	0.48187	0.25088	–	–	–
4	2^{-2}	0.20940	0.07370	0.01037	2.21	2.71	4.60
4	2^{-3}	0.03891	0.00480	0.00045	2.43	3.94	4.52
4	2^{-4}	0.00826	0.00117	0.00055	2.24	2.04	-0.27
4	2^{-5}	0.00380	0.00160	0.00102	1.12	-0.45	-0.90
4	2^{-6}	0.01251	0.01081	0.01765	-1.72	-2.76	-4.11
5	2^{-1}	0.97207	0.48187	0.25088	–	–	–
5	2^{-2}	0.20940	0.07370	0.01037	2.21	2.71	4.60
5	2^{-3}	0.03888	0.00473	0.00041	2.43	3.96	4.67
5	2^{-4}	0.00648	0.00029	0.00005	2.59	4.03	3.08
5	2^{-5}	0.00145	0.00025	0.00014	2.16	0.20	-1.49
5	2^{-6}	0.00174	0.00139	0.00233	-0.26	-2.47	-4.11
6	2^{-1}	0.97207	0.48187	0.25088	–	–	–
6	2^{-2}	0.20940	0.07370	0.01037	2.21	2.71	4.60
6	2^{-3}	0.03884	0.00474	0.00041	2.43	3.96	4.65
6	2^{-4}	0.00638	0.00026	0.00002	2.61	4.21	4.52
6	2^{-5}	0.00100	0.00004	0.00002	2.68	2.59	0.06
6	2^{-6}	0.00024	0.00018	0.00031	2.08	-2.09	-4.16

4 Justification of Quasi-Local Numerical Homogenization Methods by Inversion

In this chapter, we consider the numerical homogenization technique described in the previous chapters from a different perspective. So far, we have considered an LOD approach on some coarse scale H that is able to cope with general microscopic quantities encoded in, e.g., a diffusion coefficient corresponding to a second-order elliptic PDE. This includes the case of variations on some known fine scale $\epsilon \ll H$. The approach is quasi-local in the sense that communication among the degrees of freedom (DOFs), which can be associated with the vertices in the underlying mesh, includes not only communication between neighboring DOFs but also between those that are within ℓ layers of elements for some oversampling parameter ℓ. This is due to the fact that the basis functions of the constructed multiscale space are supported on some subdomain consisting of ℓ layers of elements. The previous chapters have shown that the choice of ℓ with respect to H (and possibly the polynomial degree p) is crucial to avoid reduced orders of convergence. Further, the fact that the deviation from true locality, i.e., only neighbor-to-neighbor communication, is to some extent controlled by the parameter H (and conceivably p) marks the difference between the described quasi-local approach and a fully non-local one.

This chapter aims at illustrating the advantage of quasi-local approaches (such as the LOD) compared to truly local ones to deal with general microscopic coefficients which only fulfill minimal assumptions. To this end, we consider the inverse problem of reconstructing a coarse model that satisfactorily reproduces given coarse data corresponding to measurements of solutions of an elliptic PDE for different boundary conditions. The idea is to allow, but not enforce, increased communication between the DOFs depending on multiple choices of ℓ and compare the respective results. Note that this approach was first presented in [CMP19].

Besides the main intention to show the potential of quasi-local models from a different point of view, this chapter also provides an actual strategy to handle inverse problems, where the microscopic coefficient to be reconstructed may vary on a very fine scale. If only coarse data are given and no a priori knowledge on a parametrization or the structure of the coefficient is available, a straightforward approach to recover the coefficient on the fine scale is computationally unfeasible and possibly fails to provide any meaningful information about the underlying coefficient.

Before we consider the actual inverse problem, it is useful to first understand the forward model and introduce forward operators that can be interpreted as the available data for the inversion procedure. This is treated in the following section.

4.1 Microscopic forward problem and effective approximation

In this section, we use the ideal setting as presented in Section 2.3.1 to derive a discrete forward operator which is then used to formulate the inverse problem. Therefore, the theory from the previous chapters is reused and adapted to the inhomogeneous setting.

4.1.1 Problem setting

We reconsider the model problem (3.1) from Chapter 3 but with inhomogeneous boundary conditions, i.e., the prototypical second-order linear elliptic diffusion problem

$$\begin{aligned} -\operatorname{div}(A\nabla u) &= f \quad && \text{in } D,\\ u &= u_0 && \text{on } \partial D, \end{aligned} \tag{4.1}$$

where $D \subseteq \mathbb{R}^d$, $d \in \{1, 2, 3\}$, is a bounded, convex, and polytopal Lipschitz domain and the admissible diffusion coefficient $A \in \mathfrak{A}$ encodes the microstructure of the medium with minimal structural assumptions; see (3.2) for the definition of $\mathfrak{A}$.

Since solutions of problem (4.1) do not necessarily exist in the classical sense, we are interested in the weak solution of (4.1) in the Sobolev space $\bar{\mathcal{V}} := H^1(D)$, which is characterized by the following variational formulation. Given $A \in \mathfrak{A}$, $u_0 \in \mathcal{X} := H^{1/2}(\partial D)$, and $f \in L^2(D)$, we seek $u \in \bar{\mathcal{V}}$ such that

$$\begin{aligned} a(u, v) &= (f, v)_{L^2(D)} \quad && \text{for all } v \in \mathcal{V} := H^1_0(D),\\ \operatorname{tr} u &= u_0 && \text{on } \partial D, \end{aligned} \tag{4.2}$$

where for v, $w \in \bar{\mathcal{V}}$, the bilinear form a is given by

$$a(v, w) = \int_D A\nabla v \cdot \nabla w \,\mathrm{d}x$$

and $\operatorname{tr}\colon \mathcal{V} \to \mathcal{X}$ is the trace operator. Note that instead of (4.1), we could as well consider a general second-order linear PDE in divergence form with additional lower-order terms. Such a generalization is straightforward.

In practice, it is favorable to rewrite problem (4.2) as a problem with homogeneous Dirichlet boundary conditions in $\mathcal{V}$. Let $E^b\colon \mathcal{X} \to \bar{\mathcal{V}}$ be a linear extension operator, which also defines the restriction operator $R\colon \bar{\mathcal{V}} \to \mathcal{V}$ by $R := \mathtt{id} - E^b\,\mathrm{tr}$. Then, we can decompose $u = Ru + (\mathtt{id} - R)u = Ru + E^b u_0$ and problem (4.2) reduces to finding $Ru \in \mathcal{V}$ such that

$$a(Ru, v) = (f, v)_{L^2(D)} - a(E^b u_0, v) \tag{4.3}$$

for all $v \in \mathcal{V}$. The next subsection deals with a discretization of this variational problem in terms of an LOD approach as introduced in Chapter 2. Again, we use the coarse parameter H for the scale on which we want to obtain a reliable coarse approximation of (4.3). With respect to the inverse problem, H is the scale on which the data are available.

4.1.2 Effective model via LOD

As in Section 2.2, let $\mathcal{T}_H$ be a mesh of quasi-uniform d-rectangles with characteristic mesh size H and denote again with $Q_1(\mathcal{T}_H)$ the corresponding space of piecewise polynomials with coordinate degree at most one in each element. In the present setting, we define the discrete spaces $\bar{V}_H := Q_1(\mathcal{T}_H) \cap \bar{\mathcal{V}}$, $V_H := \bar{V}_H \cap \mathcal{V}$, and $X_H := \mathrm{tr}\,\bar{V}_H$ of dimensions $\bar{m} = \dim \bar{V}_H$, $m = \dim V_H$, and $n = \dim X_H$, respectively. The choice of these FE spaces is not unique and other standard FE spaces could be used (see, e.g., Chapter 3). As before, we require a linear and projective quasi-interpolation operator $\mathcal{I}_H\colon L^2(D) \to V_H$ which fulfills the approximation and stability properties (2.12) and (2.11). Further, we define as in Chapter 2 the fine-scale space $\mathcal{W} := \ker \mathcal{I}_H|_{\mathcal{V}}$ and the correction operators

$$\mathcal{C}\colon \mathcal{V} \to \mathcal{W} \quad \text{and} \quad \mathcal{C}^\ell\colon V_H \to \mathcal{W}$$

by (2.25) and (2.45), respectively, and recall that

$$a((\mathtt{id} - \mathcal{C})v_H, w) = 0 \tag{4.4}$$

for $v_H \in V_H$ and $w \in \mathcal{W}$. As shown for a more general setting in Theorem 2.4.4 (see also [HP13]), we have for any $v_H \in V_H$ and $\ell \in \mathbb{N}$ that

$$\|\nabla(\mathcal{C} - \mathcal{C}^\ell)v_H\|_{L^2(D)} \lesssim \ell^{(d-1)/2} \exp(-C_{\mathrm{dec}}\,\ell)\, \|\nabla v_H\|_{L^2(D)}. \tag{4.5}$$

Since the forward problem merely serves as motivation, a fine-scale discretization of the correction operator is not considered in the present setting.

With these preliminary considerations, we can formulate an LOD method with inhomogeneous boundary conditions and with an adapted right-hand side. Given a discretized extension operator

$$E_H^b\colon X_H \to \bar{V}_H \quad \text{which fulfills} \quad E^b|_{X_H} = E_H^b$$

and the corresponding restriction operator

$$R_H\colon \bar{V}_H \to V_H, \qquad R_H := \mathtt{id} - E_H^b\, \mathrm{tr},$$

a possible discretized version of (4.3) reads: find $u_H = R_H u_H + E_H^b u_{H,0} \in \bar{V}_H$ such that

$$a(\mathcal{R}^\ell R_H u_H, \mathcal{R}^\ell v_H) = (f_H, v_H)_{L^2(D)} - a(E_H^b u_{H,0}, \mathcal{R}^\ell v_H) \tag{4.6}$$

for all $v_H \in V_H$, where $\mathcal{R}^\ell = \mathtt{id} - \mathcal{C}^\ell$ as in Chapter 2. Further, $f_H := \Pi_H f$ is the L^2-projection of f onto $\bar{V}_H$ and $u_{H,0}$ a FE approximation of u_0. In the context of inverse problems, it is reasonable to consider that u_0 is defined as the first-order FE approximation of coarse experimental boundary data which approximate the real data up to order H in the $H^{1/2}$-norm. Thus, in the following we assume that $u_0 = u_{H,0}$. For completeness, we now also define the correction operators for functions $v \in \bar{\mathcal{V}}$ and $v_H \in \bar{V}_H$, i.e., we set $\mathcal{C}v := \mathcal{C}Rv$ and $\mathcal{C}^\ell v_H := \mathcal{C}^\ell R_H v_H$, respectively.

Note that, in contrast to the previous chapters, we are only interested in the FE part u_H in (4.6) and not in the corrected variant $(\mathtt{id} - \mathcal{C}^\ell)u_H$. This is motivated by the fact that only coarse data without additional information on fine-scale corrections are available for the inverse problem.

4.1.3 Error estimates

In this subsection, we investigate the L^2-error between the solutions $u \in \bar{\mathcal{V}}$ of (4.3) and $u_H \in \bar{V}_H$ of (4.6), which is important to quantify the error between the solution operator

$$\begin{aligned} \mathfrak{L}_A\colon X_H \times L^2(D) \to\ & \bar{\mathcal{V}}, \\ (u_0, f) \mapsto\ & u, \text{ where } u \text{ solves (4.3)}, \end{aligned} \tag{4.7}$$

and its discretized version

$$\begin{aligned} \mathfrak{L}_{A,\ell}^{\mathrm{eff}}\colon X_H \times L^2(D) \to\ & \bar{V}_H, \\ (u_0, f) \mapsto\ & u_H, \text{ where } u_H \text{ solves (4.6)}. \end{aligned} \tag{4.8}$$

The following theorem shows that the error between $u \in \bar{\mathcal{V}}$ and the FE part $u_H \in \bar{V}_H$ scales optimally with H and that it is independent of the variations of the diffusion coefficient. The theorem adapts ideas from Chapter 2.

Theorem 4.1.1 (Error of the forward effective model)**.** *Let $u \in \bar{\mathcal{V}}$ be the solution of* (4.3) *and $u_H \in \bar{V}_H$ the solution of* (4.6) *for given boundary data $u_0 \in X_H$, a right-hand side $f \in L^2(D)$, as well as an oversampling parameter $\ell \in \mathbb{N}$. For $g \in L^2(D)$, denote with $\hat{u}(g) \in \bar{\mathcal{V}}$ the solution of* (4.3) *with right-hand* *side g and boundary condition $u_0 = 0$. Further, we introduce the worst-case best-approximation error*

$$\mathbf{wcba}(A, \mathcal{T}_H) := \sup_{g \in L^2(D)} \inf_{v_H \in V_H} \frac{\|R\hat{u}(g) - v_H\|_{L^2(D)}}{\|g\|_{L^2(D)}}.$$

Then it holds that

$$\|u - u_H\|_{L^2(D)} \lesssim \big(H^2 + \exp(-C_{\mathrm{dec}}\,\ell) + \mathbf{wcba}(A, \mathcal{T}_H)\big)\,\big(\|f\|_{L^2(D)} + \|u_0\|_{\mathcal{X}}\big).$$

Proof. We split the error $u - u_H = (u - \bar{u}_H) + (\bar{u}_H - \tilde{u}_H) + (\tilde{u}_H - u_H)$ with the solutions $\bar{u}_H$ and $\tilde{u}_H$ of the auxiliary problems

$$a(R_H\bar{u}_H, (\mathtt{id} - \mathcal{C})v_H) = (f, v_H)_{L^2(D)} - a(E_H^b u_0, (\mathtt{id} - \mathcal{C})v_H)$$

and

$$a(R_H\tilde{u}_H, (\mathtt{id} - \mathcal{C}^\ell)v_H) = (f_H, v_H)_{L^2(D)} - a(E_H^b u_0, (\mathtt{id} - \mathcal{C}^\ell)v_H)$$

for all $v_H \in V_H$. To bound $e_H := u_H - \tilde{u}_H$, we observe using the orthogonality property (4.4) that

$$\begin{aligned} a((\mathtt{id} - \mathcal{C}^\ell)e_H,&(\mathtt{id} - \mathcal{C}^\ell)v_H) \\ &= a(\mathcal{C}^\ell R_H\tilde{u}_H, (\mathtt{id} - \mathcal{C}^\ell)v_H) = a(\mathcal{C}^\ell R_H\tilde{u}_H, (\mathcal{C} - \mathcal{C}^\ell)v_H). \end{aligned} \tag{4.9}$$

Testing with $v_H = e_H$ in (4.9) and using (4.5), (2.14), and the fact that $e_H = \mathcal{I}_H(\mathtt{id} - \mathcal{C}^\ell)e_H$, it follows that

$$\begin{aligned} \alpha\,\|\nabla(\mathtt{id} - \mathcal{C}^\ell)e_H\|_{L^2(D)}^2 &\leq a((\mathtt{id} - \mathcal{C}^\ell)e_H, (\mathtt{id} - \mathcal{C}^\ell)e_H) \\ &= a(\mathcal{C}^\ell R_H\tilde{u}_H, (\mathcal{C} - \mathcal{C}^\ell)e_H) \\ &\lesssim \exp(-C_{\mathrm{dec}}\,\ell)\,\|\nabla\mathcal{C}^\ell R_H\tilde{u}_H\|_{L^2(D)}\,\|\nabla(\mathtt{id} - \mathcal{C}^\ell)e_H\|_{L^2(D)} \end{aligned}$$

and thus

$$\|e_H\|_{L^2(D)} \lesssim \|\nabla(\mathtt{id} - \mathcal{C}^\ell)e_H\|_{L^2(D)} \lesssim \exp(-C_{\mathrm{dec}}\,\ell)\big(\|f\|_{L^2(D)} + \|u_0\|_{\mathcal{X}}\big), \tag{4.10}$$

where we use (2.11) and the Friedrichs inequality. As a next step, we bound the error $\bar{e}_H := \tilde{u}_H - \bar{u}_H$. We note that

$$a(\bar{e}_H, (\mathtt{id} - \mathcal{C})v_H) = a(R_H\tilde{u}_H + E_H^b u_0, (\mathcal{C}^\ell - \mathcal{C})v_H)$$

for any $v_H \in V_H$. With $v_H = \bar{e}_H$ and similar arguments as above, we obtain

$$\|\bar{e}_H\|_{L^2(D)} \lesssim \|\nabla(\mathtt{id} - \mathcal{C})\bar{e}_H\|_{L^2(D)} \lesssim \exp(-C_{\mathrm{dec}}\,\ell)\big(\|f\|_{L^2(D)} + \|u_0\|_{\mathcal{X}}\big). \tag{4.11}$$

The error $u - \bar{u}_H$ can be estimated using [GP17, Prop. 1], which also holds for inhomogeneous Dirichlet boundary conditions, i.e.,

$$\|u - \bar{u}_H\|_{L^2(D)} \lesssim \big(H^2 + \mathbf{wcba}(A, \mathcal{T}_H)\big)\big(\|f\|_{L^2(D)} + \|u_0\|_{\mathcal{X}}\big). \tag{4.12}$$

The triangle inequality, (4.10), (4.11), and (4.12) yield the desired estimate. □

We emphasize that, choosing ℓ large enough (i.e., $\ell \gtrsim |\log H|$), we have $\exp(-C_{\mathrm{dec}}\,\ell) \lesssim H$ or even $\exp(-C_{\mathrm{dec}}\,\ell) \lesssim H^2$. As discussed in [GP17], the worst-case best-approximation error is at least $\mathcal{O}(H)$, and it scales possibly even better with H in certain regimes (cf. also Figure 2.4 (right)).

To prepare the setting of the inverse problem, we go back to the operators defined in (4.7) and (4.8) and observe that $\mathfrak{L}_A$ (and similarly also $\mathfrak{L}^{\mathrm{eff}}_{A,\ell}$) can be written as

$$\mathfrak{L}_A(u_0, f) = \mathfrak{L}_A(u_0, 0) + \mathfrak{L}_A(0, f) \tag{4.13}$$

with the linear operators $\mathfrak{L}_A(\cdot, 0)\colon X_H \to \bar{\mathcal{V}}$ and $\mathfrak{L}_A(0, \cdot)\colon L^2(D) \to \bar{\mathcal{V}}$. For simplicity, we assume in the following that f is a fixed function. The generalization to the case where f is also part of the input data is conceptually straightforward but slightly more involved. The decomposition (4.13) motivates the distance function between operators defined by

$$\operatorname{dist}_f(\mathfrak{C}, \mathfrak{D}) := \Big(\|\mathfrak{C}(\cdot, 0) - \mathfrak{D}(\cdot, 0)\|^2_{\mathcal{L}(X_H; L^2(D))} + \|\mathfrak{C}(0, f) - \mathfrak{D}(0, f)\|^2_{L^2(D)}\Big)^{1/2}$$

for all $\mathfrak{C}$, $\mathfrak{D}\colon X_H \times L^2(D) \to L^2(D)$.

Remark 4.1.2. If we consider the case $f = 0$, coefficients that only differ by a multiplicative constant produce the same solution operator. In view of the inverse problem in the next section, in this case one should fix an additional parameter, e.g., the mean value of A.

Using Theorem 4.1.1, we obtain the following result which quantifies the error between the two solution operators $\mathfrak{L}_A$ and $\mathfrak{L}^{\mathrm{eff}}_{A,\ell}$.

Corollary 4.1.3 (Error of the effective forward operator)**.** *Let $\ell \gtrsim |\log H|$. Then it holds that*

$$\operatorname{dist}_f(\mathfrak{L}_A, \mathfrak{L}^{\mathrm{eff}}_{A,\ell}) \lesssim H.$$

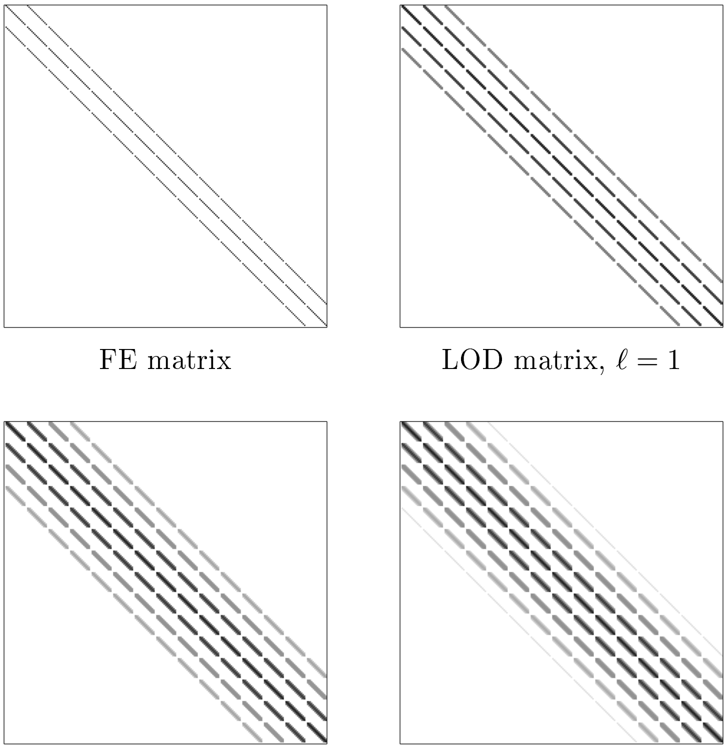
4.1.4 Reformulation of the effective model

As a next step, we discuss an alternative representation of the operator $\mathfrak{L}^{\mathrm{eff}}_{A,\ell}$ using the effective stiffness matrix corresponding to the discrete formulation (4.6). Given a coefficient $A \in \mathfrak{A}$, the corresponding LOD stiffness matrix $S_H(A, \ell)$ is defined by

$$S_H(A, \ell)[i, j] := a(\mathcal{R}^\ell \Lambda_{z_j}, \mathcal{R}^\ell \Lambda_{z_i}), \quad i,\, j \in \{1, \dots, \bar{m}\}, \tag{4.14}$$

where $i \mapsto z_i$ is a fixed ordering of the $\bar{m}$ vertices in $\mathcal{T}_H$ and $\Lambda_z \in \bar{V}_H$ denotes the classical nodal basis function associated with the vertex z of $\mathcal{T}_H$. The typical sparsity pattern of such a matrix is depicted in Figure 4.1. Next, we introduce the set of LOD stiffness matrices with oversampling parameter ℓ based on admissible coefficients, which is given by

$$\mathcal{S}(\ell, \mathcal{T}_H) := \left\{S_H(A, \ell) \in \mathbb{R}^{\bar{m}\times\bar{m}}_{\mathrm{sym}} \,:\, A \in \mathfrak{A}\right\}. \tag{4.15}$$



FE matrix

LOD matrix, $\ell = 1$

LOD matrix, $\ell = 2$

LOD matrix, $\ell = 3$

Figure 4.1: Sparsity patterns of a classical first-order FE stiffness matrix and LOD stiffness matrices for different values of ℓ on a Cartesian grid with lexicographic ordering in $D = (0,1)^2$.

For better readability, from now on we use the notation v_H (or B_H) for both the vector $v_H \in \mathbb{R}^{\bar{m}}$ (or the matrix $B_H \in \mathbb{R}^{m \times \bar{m}}$) and the corresponding function $v_H \in \bar{V}_H$ (or the mapping $B_H \colon \bar{V}_H \to V_H$). For any matrix $S_H \in \mathcal{S}(\ell, \mathcal{T}_H)$, we define the operator

$$\begin{aligned} \mathfrak{L}^{\mathrm{eff}}_{S_H} \colon X_H \times L^2(D) &\to \bar{V}_H, \\ (u_0, f) &\mapsto u_H, \text{ where } u_H \text{ solves} \\ &\quad \begin{cases} S_{H,0} R_H u_H = R_H M_H f_H - R_H S_H E^b_H u_0, \\ \qquad\qquad u_H = u_0 \text{ on } \partial D \end{cases} \end{aligned} \tag{4.16}$$

with the classical FE mass matrix M_H, the restriction $S_{H,0} = R_H S_H R_H^T$ of S_H to the inner vertices of $\mathcal{T}_H$, and $f_H = \Pi_H f$. With the above definitions, we can prove the following lemma.

Lemma 4.1.4 (Alternative representation of the effective forward operator)**.** *Let $S_H(A,\ell) \in \mathcal{S}(\ell, \mathcal{T}_H)$ be the LOD stiffness matrix corresponding to (4.6). Assume that E^b fulfills $\mathcal{C}^\ell E^b_H v_0 = \mathcal{C}^\ell E^b|_{X_H} v_0 = 0$ for any $v_0 \in X_H$. Then it holds that*

$$\mathfrak{L}^{\mathrm{eff}}_{S_H(A,\ell)}(u_0, f) = \mathfrak{L}^{\mathrm{eff}}_{A,\ell}(u_0, f) \tag{4.17}$$

for all $u_0 \in X_H$, $f \in L^2(D)$.

Remark 4.1.5. Possible choices for an extension operator E^b that fulfills the assumptions of Lemma 4.1.4 are those that extend functions in X_H to functions in $\bar{V}_H$ that are only supported on one layer of elements away from the boundary.

Proof of Lemma 4.1.4. We write $u_H = \sum_{j=1}^{\bar{m}} u_j \Lambda_{z_j}$ and observe that (4.6) is equivalent to

$$\sum_{j\,:\,z_j \notin \partial D} u_j\, a(\mathcal{R}^\ell \Lambda_{z_j}, \mathcal{R}^\ell \Lambda_{z_i}) = (f_H, \Lambda_{z_i})_{L^2(D)} - a(E_H^b u_0, \mathcal{R}^\ell \Lambda_{z_i}) \tag{4.18}$$

for all $i \in \{k \,:\, z_k \notin \partial D\}$. Inserting $f_H = \sum_{j=1}^{\bar{m}} f_j \Lambda_{z_i}$ and using the fact that

$$\begin{aligned} a(E_H^b u_0, \mathcal{R}^\ell v_H) &= a((\mathtt{id} - \mathcal{C}^\ell) E_H^b u_0, \mathcal{R}^\ell v_H) \\ &= a(\mathcal{R}^\ell E_H^b u_0, \mathcal{R}^\ell v_H) \end{aligned}$$

for any $v_H \in V_H$ and the definition (4.14), we can write equation (4.18) as

$$S_{H,0}(A,\ell) R_H u_H = R_H M_H f_H - R_H S_H(A,\ell) E_H^b u_0,$$

which proves (4.17). □

Lemma 4.1.4 and Corollary 4.1.3 show that $\mathfrak{L}_A(\cdot, f)$ and $\mathfrak{L}^{\text{eff}}_{S_H(A,\ell)}(\cdot, f)$ are close as operators from X_H to $\bar{\mathcal{V}}$ if ℓ is chosen large enough. We use this property in the next section to motivate the inverse problem. First, however, we give a brief overview of other methods that provide similar effective models as the LOD.

4.1.5 Other quasi-local approaches

In this subsection, we quantify the quasi-locality of a method with respect to the sparsity pattern of its stiffness matrix that occurs in the representation of the approach in terms of a linear system as, e.g., described in (4.16).

For the LOD, we know by the definition of $\mathcal{C}^\ell$ that the resulting stiffness matrix $S_H \in \mathcal{S}(\ell, \mathcal{T}_H)$ is included in the set

$$\mathcal{M}(\ell, \mathcal{T}_H) := \left\{ \begin{aligned} S_H \in \mathbb{R}^{\bar{m}\times\bar{m}}_{\text{sym}} \,:\, \forall\, 1 \le i \le j \le \bar{m} \,:& \\ z_i \notin \mathsf{N}^\ell(z_j) \Rightarrow S_H[i,j] = 0& \end{aligned} \right\} \tag{4.19}$$

of matrices that may only have a non-zero entry at position $[i, j]$ if the corresponding vertices z_i and z_j belong to the ℓ-neighborhood (see definition (2.13)) of each other. In other words, it holds that $\mathcal{S}(\ell, \mathcal{T}_H) \subseteq \mathcal{M}(\ell, \mathcal{T}_H)$. Standard stiffness and mass matrices arising from classical FE methods belong to the space $\mathcal{M}(0, \mathcal{T}_H)$ such that these methods can be referred to as local. Classical homogenization approaches such as the MsFEM without oversampling [HW97], the Two-Scale Finite Element Method [MS02], or the HMM [EE03, EE05] share the communication pattern of classical FE methods and therefore also lead to stiffness matrices in $\mathcal{M}(0,\mathcal{T}_H)$. Matrices arising from the MsFEM with oversampling are included in $\mathcal{M}(1,\mathcal{T}_H)$.

Concerning mathematical models that satisfactorily describe the effective behavior of physical processes on the scale of data resolution in the presence of very general coefficients, there are various other approaches that produce stiffness matrices with sparsity patterns similar to the LOD, such as the GFEM [BL11], the ALB [GGS12], RPS [OZB14], the GMsFEM [EGH13], gamblets [Owh17], CEM-GMsFEM [CEL18], and their variants. All these methods provably work in the linear elliptic setting and are based on a coarse mesh with a characteristic mesh parameter (typically the effective scale) or corresponding concepts in the setting of mesh-free methods. These methods are of Galerkin-type and thus characterized by discrete bases. To achieve optimal accuracy, a moderate price in terms of an overhead in the computational complexity has to be paid compared to a standard FE method (fixed order) on the same mesh. The overhead is either in the number of functions per mesh entity (GFEM, GMsFEM), e.g., elements or vertices, or in the support of the basis functions (LOD, RPS, gamblets, ALB). In both cases, the result is an increased communication between the DOFs which, in turn, leads to a slightly denser sparsity pattern of the corresponding system matrices. In other words, these matrices lie in $\mathcal{M}(\ell,\mathcal{T}_H)$ for some (moderate) $\ell \in \mathbb{N}$. Thus, all these methods can be referred to as quasi-local as well. It is worth noting that the set of matrices with the considered sparsity pattern includes also matrices that occur in isogeometric analysis [HCB05, CHB09]. Moreover, the set $\mathcal{M}(\ell,\mathcal{T}_H)$ also contains higher-order FE matrices with polynomial degree $p \sim \ell$ on meshes that are coarser by roughly a factor of p and matrices from peridynamics [Sil00, Lip14, Du17] with horizon $\delta \sim H\ell$. We emphasize that also the higher-order LOD approach of Chapter 3 leads to matrices in the set (4.19), with an amplified communication pattern that depends on ℓ and p.

The theoretical analysis of the methods mentioned above indicates that reliable effective models for PDEs with general microstructures are based on a controlled deviation from locality. Similar observations have been made in connection with the pollution effect in high-frequency time-harmonic wave propagation [BS97], which cannot be avoided unless the mesh size is coupled to, e.g., the polynomial degree [MS10, MS11, MPS13] or the support of the basis functions [Pet17] in a logarithmic way. Finally, we mention that non-locality is also considered in classical stochastic homogenization in connection with higher-order correctors to achieve better approximation properties; see, e.g., [DGO16].

Although the quasi-local effective models described above are purely discrete and lack a PDE representation in general, they are well-understood. This is the main motivation for the present approach of reconstructing quasi-local effective models (i.e., their matrix representation) given low-resolution measurements based on inhomogeneous boundary data in a medium with microstructures. This is further discussed in the subsequent section.

4.2 Inverse problem: reconstruction of the effective model

4.2.1 Problem setting

Let us now assume that the diffusion coefficient A is unknown. Since information about the coefficient is not available, structural assumptions such as periodicity, local periodicity, and given parameterization by few DOFs cannot be satisfied a priori. In an ideal setting, information about solutions of problem (4.3) in the form of a solution operator

$$\tilde{\mathfrak{L}} := \mathfrak{L}_A(\cdot, f)\colon \mathcal{X} \to \bar{\mathcal{V}}$$

would be given. In practical applications, however, boundary data and information about the corresponding solutions are only available on some coarse scale H, possibly much larger than the microscopic scale on which the diffusion coefficient and the corresponding solutions vary. In this case, a classical formulation of the inverse problem, for a fixed right-hand side f, consists in recovering A in (4.3) given a mapping

$$\tilde{\mathfrak{L}}^{\mathrm{eff}} := \mathfrak{L}_A^{\mathrm{eff}}(\cdot, f)\colon X_H \to \bar{V}_H$$

which comprises coarse measurements of solutions of (4.3).

If the unknown coefficient includes fine-scale features, a direct approach of recovering A by full (fine-scale) simulations is computationally unfeasible. Inspired by the ideas presented in Section 4.1, we present in this section an alternative approach to recover information about the macroscopic effective model taking into account the presence of a microscopic diffusion coefficient. Rather than reconstructing the diffusion coefficient itself, we tackle the reconstruction of an effective stiffness matrix that is able to reproduce the given data related to solutions of (4.3). We recall that such an approach is reasonable since the mapping $\tilde{\mathfrak{L}}^{\mathrm{eff}}$ can not only be characterized by the corresponding coefficient but also by the effective stiffness matrix as described in the previous section. Therefore, the alternative formulation of the inverse problem reads:

$$\text{given } \tilde{\mathfrak{L}}^{\mathrm{eff}}\colon X_H \to \bar{V}_H, \text{ find the corresponding stiffness matrix } \tilde{S}_H.$$

Note that in the case $f = 0$, the classical Calderon problem [Cal80] might be considered, where a so-called *Dirichlet-to-Neumann mapping* is given instead of the operator $\mathfrak{L}_A$. However, this problem requires information on the coefficient at the boundary, and the derivation of the method presented below needs to be adjusted accordingly.

4.2.2 The minimization problem

The inverse problem could ideally be formulated as a minimization problem for the functional

$$\tilde{\mathcal{J}}_H(S_H) = \frac{1}{2}\Big(\operatorname{dist}_f(\tilde{\mathfrak{L}}^{\mathrm{eff}}, \mathfrak{L}^{\mathrm{eff}}_{S_H})\Big)^2 \tag{4.20}$$

in the set $\mathcal{S}(\ell, \mathcal{T}_H)$ of LOD stiffness matrices based on admissible coefficients, where $\mathfrak{L}^{\mathrm{eff}}_{S_H}$ is defined in (4.16). However, since we are not able to characterize the set $\mathcal{S}(\ell, \mathcal{T}_H)$ in a way that would be suitable for optimization, we instead seek a minimizer in the linear space $\mathcal{M}(\ell, \mathcal{T}_H) \supseteq \mathcal{S}(\ell, \mathcal{T}_H)$ of matrices with prescribed sparsity pattern as defined in (4.19). That is, we enlarge the set of possible minimizers. We emphasize that with this generalization, the sole criterion in the inversion process is the sparsity pattern. From now on, we are searching for effective models with increased communication between the DOFs, including those mentioned in Section 4.1.5, without requiring any particular knowledge on the LOD or other numerical homogenization methods.

The minimization problem with respect to the space $\mathcal{M}(\ell, \mathcal{T}_H)$ reads

$$\text{find } \tilde{S}^*_H = \underset{S_H \in \mathcal{M}(\ell, \mathcal{T}_H)}{\arg\min}\ \tilde{\mathcal{J}}_H(S_H). \tag{4.21}$$

Using the previously introduced matrices, the operator $\mathfrak{L}^{\mathrm{eff}}_{S_H}(\cdot, f)\colon X_H \to \bar{V}_H$ can be interpreted as a matrix of size $\bar{m} \times n$, i.e.,

$$\mathfrak{L}^{\mathrm{eff}}_{S_H} = \left(I - R_H^T S_{H,0}^{-1} R_H S_H\right) E_H^b + R_H^T S_{H,0}^{-1} R_H M_H F_H,$$

with $F_H := [f_H, f_H, \dots, f_H] \in \mathbb{R}^{\bar{m} \times n}$ and the identity matrix $I \in \mathbb{R}^{\bar{m} \times \bar{m}}$. The matrix $\mathfrak{L}^{\mathrm{eff}}_{S_H}$ comprises full information about the forward problem in the sense that it includes the solutions of (4.16) for a complete set of basis functions of X_H. Note, however, that $\mathfrak{L}^{\mathrm{eff}}_{S_H}$ is not linear. That is, for a particular boundary condition $u_0 \in \mathbb{R}^n$, we have

$$\mathfrak{L}^{\mathrm{eff}}_{S_H}(u_0) = \left(I - R_H^T S_{H,0}^{-1} R_H S_H\right) E_H^b u_0 + R_H^T S_{H,0}^{-1} R_H M_H f_H.$$

The operator $\tilde{\mathfrak{L}}^{\mathrm{eff}}$ may also be interpreted as a matrix, so that the distance between the operators can be measured in general matrix norms. This is especially useful since a splitting of the form (4.13) is generally not known for $\tilde{\mathfrak{L}}^{\mathrm{eff}}$.

Let $\mu := \dim \mathcal{M}(\ell, \mathcal{T}_H)$. Based on the matrix representation introduced above, instead of (4.21) we consider a minimization problem for the functional $\mathcal{J}_H\colon \mathbb{R}^\mu \to \mathbb{R}$ defined by

$$\mathcal{J}_H(S_H) := \frac{1}{2}\big\|\tilde{\mathfrak{L}}^{\mathrm{eff}}\big\|^{-2}_{\mathbb{R}^{\bar{m}\times n}} \big\|\tilde{\mathfrak{L}}^{\mathrm{eff}} - \mathfrak{L}^{\mathrm{eff}}_{S_H}\big\|^2_{\mathbb{R}^{\bar{m}\times n}}. \tag{4.22}$$

At this stage, the choice of the norm in $\mathbb{R}^{\bar{m}\times n}$ in (4.22) is arbitrary. The results that we show in Section 4.3 were obtained using the Frobenius norm, which is a natural candidate.

4.2.3 Iterative minimization

To find a minimizer of (4.22), we can now apply standard minimization techniques such as the *Newton method* or the *gradient descent method.* Here, we adopt a *Gauß-Newton method* [NW06, Sec. 10.3] which, in our numerical computations, showed faster convergence in terms of number of iterations.

In order to compute the descent direction, the most important step concerns the computation of the gradient of $\mathcal{J}_H$ with respect to the *relevant entries* $\{s_i\}_{i=1}^{\mu}$ of S_H (i.e., the diagonal and the non-zero entries above the diagonal, due to symmetry). Using the chain rule, we obtain

$$\frac{\partial}{\partial s_i}\mathcal{J}_H(S_H) = -\big\|\tilde{\mathfrak{L}}^{\mathrm{eff}}\big\|^{-2}_{\mathbb{R}^{\bar{m}\times n}}\big(\tilde{\mathfrak{L}}^{\mathrm{eff}} - \mathfrak{L}^{\mathrm{eff}}_{S_H}\big) : \frac{\partial \mathfrak{L}^{\mathrm{eff}}_{S_H}}{\partial s_i} \tag{4.23}$$

with $M : \tilde{M} := \operatorname{trace}(M\tilde{M}^T)$. For the Gauß-Newton method only the derivatives of $\mathfrak{L}^{\mathrm{eff}}_{S_H}$ are needed, i.e.,

$$\begin{aligned}\frac{\partial \mathfrak{L}^{\mathrm{eff}}_{S_H}}{\partial s_i} &= -R_H^T\left(\frac{\partial S_{H,0}^{-1}}{\partial s_i}\right)R_H(S_H E_H^b - M_H F_H)\\ &\quad - R_H^T S_{H,0}^{-1} R_H\left(\frac{\partial S_H}{\partial s_i}\right)E_H^b\\ &= R_H^T S_{H,0}^{-1}\left(\frac{\partial S_{H,0}}{\partial s_i}\right)S_{H,0}^{-1}R_H(S_H E_H^b - M_H F_H)\\ &\quad - R_H^T S_{H,0}^{-1} R_H\left(\frac{\partial S_H}{\partial s_i}\right)E_H^b.\end{aligned}$$

The derivatives $\frac{\partial S_H}{\partial s_i}$ and $\frac{\partial S_{H,0}}{\partial s_i}$ are relatively easy to compute, as they are defined as global matrices that only contain at most two non-zero entries equal to 1.

For ease of notation, we interpret $\mathfrak{L}^{\mathrm{eff}}_{S_H}$ and S_H as vectors in $\mathbb{R}^{\bar{m}n}$ and $\mathbb{R}^{\bar{m}^2}$, respectively. The Gauß-Newton method to minimize the functional $\mathcal{J}_H$ is then defined by the following steps:

- Let an initial matrix $S_H^0 \in \mathcal{M}(\ell, \mathcal{T}_H)$ be given.
- For $k = 0, 1, \dots$ (until a certain stopping criterion is satisfied), solve

$$H_k p_k = \left(\nabla \mathfrak{L}^{\mathrm{eff}}_{S_H^k}\right)^T\left(\tilde{\mathfrak{L}}^{\mathrm{eff}} - \mathfrak{L}^{\mathrm{eff}}_{S_H^k}\right), \tag{4.24}$$

  where ∇ denotes the derivative with respect to the relevant entries of S_H and

$$H_k = \left(\nabla \mathfrak{L}^{\mathrm{eff}}_{S_H^k}\right)^T\left(\nabla \mathfrak{L}^{\mathrm{eff}}_{S_H^k}\right).$$

- Set $P_k \in \mathcal{M}(\ell, \mathcal{T}_H)$ as the matrix whose relevant entries are given by p_k and define

$$S_H^{k+1} = S_H^k + \delta_k P_k \tag{4.25}$$

with appropriately chosen *step size* δ_k, for example using *backtracking line search* based on the *Armijo condition*; see, e.g., [NW06, Alg. 3.1] for the details.

Due to the ill-posedness of the inverse problem, the matrix H_k might be singular. A possible approach to overcome this issue consists in replacing (4.24) with

$$(H_k + \eta I)\, p_k = \left[\nabla \mathfrak{L}^{\text{eff}}_{S_H^k}\right]^T \left[\tilde{\mathfrak{L}}^{\text{eff}} - \mathfrak{L}^{\text{eff}}_{S_H^k}\right] \tag{4.26}$$

with a given parameter $\eta > 0$, which is typically referred to as *regularization.*

Another possible strategy is to add a regularization term to the functional to be minimized, i.e., to replace (4.22) by

$$\mathcal{J}_H(S_H) = \frac{1}{2}\big\|\tilde{\mathfrak{L}}^{\text{eff}}\big\|^{-2}_{\mathbb{R}^{\bar{m}\times n}}\big\|\tilde{\mathfrak{L}}^{\text{eff}} - \mathfrak{L}^{\text{eff}}_{S_H}\big\|^2_{\mathbb{R}^{\bar{m}\times n}} + \frac{\gamma}{2}\big\|S_{\text{reg}} - S_H\big\|^2_{R^{\bar{m}\times\bar{m}}}, \tag{4.27}$$

where $\gamma > 0$ is a given regularization parameter and S_{reg} is a regularization (or stabilization) matrix. Additionally, the computations of the gradient in (4.23) need to be adapted accordingly. In the presence of multiple minimizers, this regularization forces the solution to be close (depending on the parameter γ) to the matrix S_{reg}. For instance, if the aim of the inverse problem is to find defects in an otherwise homogeneous medium, a suitable choice for S_{reg} could be a standard FE stiffness matrix for a constant diffusion coefficient. In our practical computations, the regularization approach described in (4.26) is used, which generally led to better results.

We emphasize that the presented inversion process does not need to resolve any fine scales in order to obtain an effective numerical model. Further, the information extracted by this procedure (i.e., a stiffness matrix $\tilde{S}_H$) may be used to simulate other problems subject to the same (unknown) diffusion coefficient. Finally, the information gathered can also be seen as an intermediate step towards recovering information concerning the original coefficient itself.

4.3 Numerical experiments

In this section, we present some numerical experiments that illustrate the capability of the proposed method. The inverse problem is based on synthetic data, i.e., the coarse measurements used to feed the inversion algorithm are obtained from FE functions in $\bar{V}_h$, defined on a mesh of $D = (0,1)^2$ with mesh size $h = 2^{-9}$ that resolve the fine-scale features of the diffusion coefficient. Furthermore, the data are perturbed by random noise with intensity up to 5%.

4.3.1 Example 1: full boundary data

In a first experiment, we assume to have full information on the operator (matrix) $\tilde{\mathfrak{L}}^{\text{eff}}$, i.e., we assume that measurements in D on the scale $H = 2^{-5}$ for

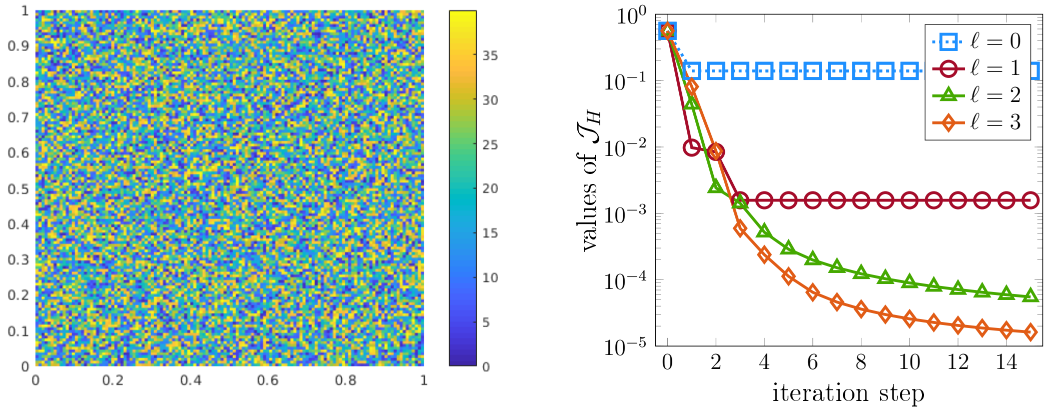


Figure 4.2: Diffusion coefficient in Example 1 (left) and values of $\mathcal{J}_H$ in the first 15 iterations of the inversion algorithm using sparsity patterns based on local matrices ($\ell = 0$) and quasi-local matrices with $\ell \in \{1, 2, 3\}$ (right).

a complete basis of X_H are available. The scalar coefficient A, for which the effective behavior should be recovered, is piecewise constant on a mesh $\mathcal{T}_\epsilon$ with $\epsilon = 2^{-7}$ and the value on each element is independently obtained as a uniformly distributed random number between 1 and 40, i.e., for any $K \in \mathcal{T}_\epsilon$ we have $A|_K \sim U(1, 40)$; see Figure 4.2 (left) for the explicit sample used here. We set $f = 1$ and start the inverse iteration with the first-order FE stiffness matrix S_H^0 based on the constant coefficient with value 1. The values of the functional $\mathcal{J}_H$ in the first 15 iterations of the inversion algorithm are given in Figure 4.2 (right). In particular, we compare the performance of a *local approach* based on matrices with the sparsity pattern of a standard first-order FE method (such as, e.g., the HMM or the Two-Scale Finite Element Method) with the proposed *quasi-local method* based on matrices in $\mathcal{M}(\ell, \mathcal{T}_H)$ for $\ell \in \{1, 2, 3\}$. One clearly sees that the quasi-local inversion leads to better results in terms of decrease and value of the error functional $\mathcal{J}_H$. In particular, with the local approach the functional seems to reach a stagnation relatively quickly, while the results significantly improve with the quasi-local approach when increasing the value of ℓ.

A necessary validation step, in order to further investigate the different methods, consists in solving a diffusion problem using the stiffness matrices reconstructed with the different approaches (local and quasi-local) and comparing the resulting numerical solutions with the FE functions from which the measurements were taken to feed the inversion algorithm. The outcome of this assessment is shown in Figure 4.3, focusing on the cross sections at $x_2 = 0.5$ (left) and at $x_1 = 0.5$ (right) of the numerical approximations corresponding to the boundary condition $u_0(x) = \sin(3\pi\, x_1)$. Figure 4.4 depicts the same cross sections when a random boundary condition $u_0 \in X_H$ is considered. As before,

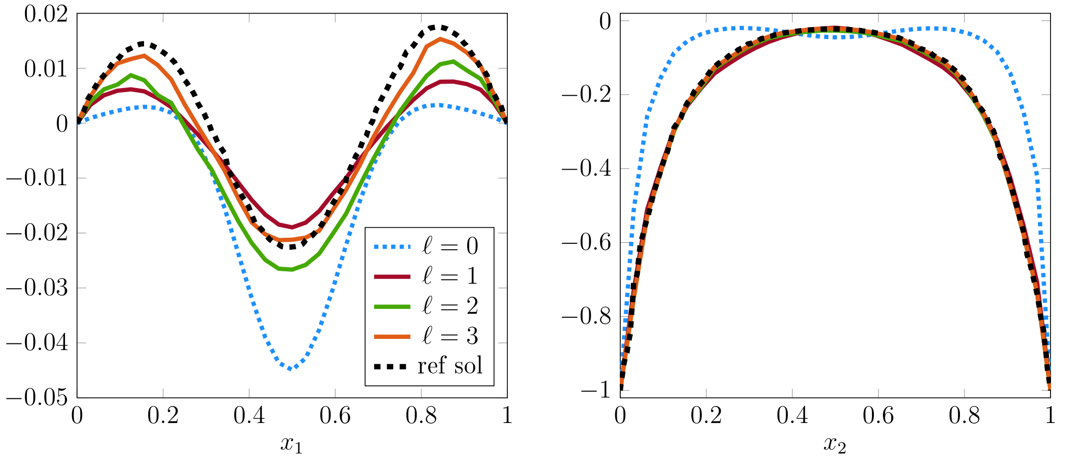


Figure 4.3: Cross sections at $x_2 = 0.5$ (left) and at $x_1 = 0.5$ (right) of reconstructed functions with the boundary condition $u_0(x) = \sin(3\pi\, x_1)$ based on local stiffness matrices ($\ell = 0$) and quasi-local ones with $\ell \in \{1, 2, 3\}$ for Example 1 obtained from full boundary data. The corresponding fine FE function is depicted as a reference.

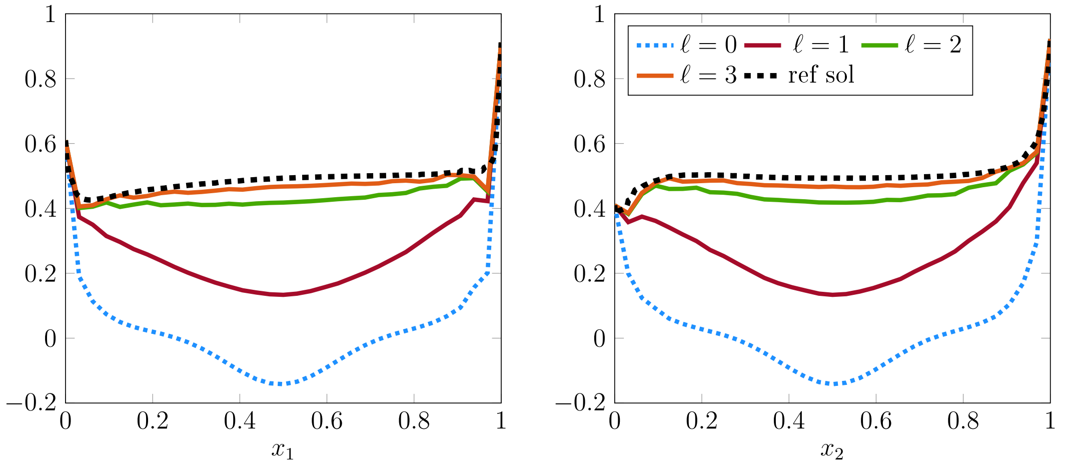


Figure 4.4: Cross sections at $x_2 = 0.5$ (left) and at $x_1 = 0.5$ (right) of reconstructed functions with a randomly chosen boundary condition $u_0 \in X_H$ based on local stiffness matrices ($\ell = 0$) and quasi-local ones with $\ell \in \{1, 2, 3\}$ for Example 1 obtained from full boundary data. The corresponding fine FE function is depicted as a reference.

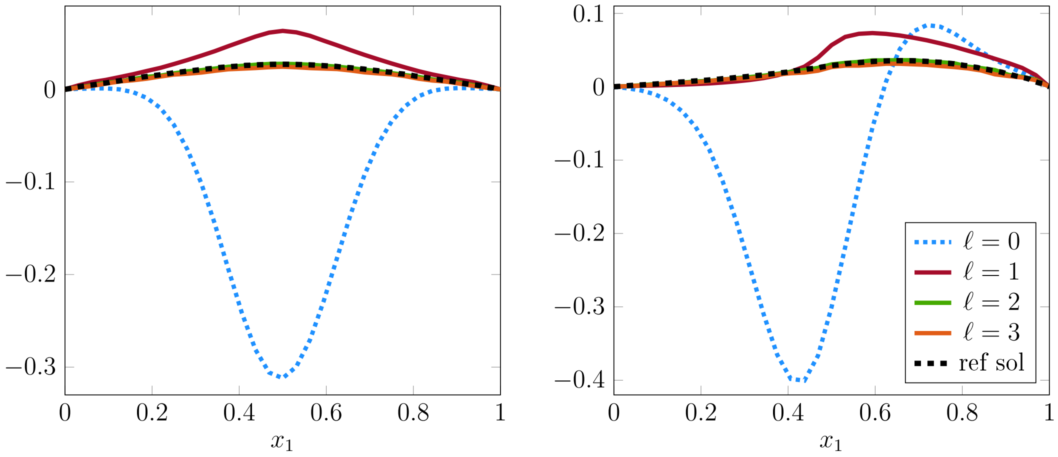


Figure 4.5: Cross sections at $x_2 = 0.5$ of reconstructed functions with homogeneous Dirichlet boundary conditions and right-hand side g_1 (left) and g_2 (right) based on local stiffness matrices ($\ell = 0$) and quasi-local ones with $\ell \in \{1, 2, 3\}$ for Example 1. The corresponding fine FE functions are given as a reference but were not part of the input data.

these results show an improved behavior if ℓ is increased, in particular in the case of the highly oscillating boundary condition considered in Figure 4.4.

Besides the accuracy of the numerical approximations based on the recovered stiffness matrices, it is also important to assess the robustness of the reconstructed effective model, i.e., to investigate to which extent the coarsened information about the diffusion coefficient encoded in the stiffness matrix can be used to simulate other scenarios.

For this purpose, we employ the reconstructed stiffness matrices to simulate a diffusion problem with two different right-hand sides, i.e.,

$$g_1(x) = 20\,(\mathbb{1}_{\{x_1<0.5\}}\,x_1 + \mathbb{1}_{\{x_1\geq 0.5\}}\,(1-x_1))(\mathbb{1}_{\{x_2<0.5\}}\,x_2 + \mathbb{1}_{\{x_2\geq 0.5\}}\,(1-x_2))$$

and

$$g_2(x) = 10\,\mathbb{1}_{\{x_1\geq 0.5\}},$$

and compare the numerical results with the corresponding fine-scale solution using the diffusion coefficient depicted in Figure 4.2 (left). In both cases, homogeneous Dirichlet boundary conditions are imposed on the outer boundaries.

Representative cross sections of the numerical approximations based on the reconstructed stiffness matrices, compared to the corresponding fine-scale solutions, are shown in Figure 4.5. The numerical results indicate that robustness can be assured only with the quasi-local inversion. Moreover, as in the previous experiments, the quality of the results improves if ℓ is increased.

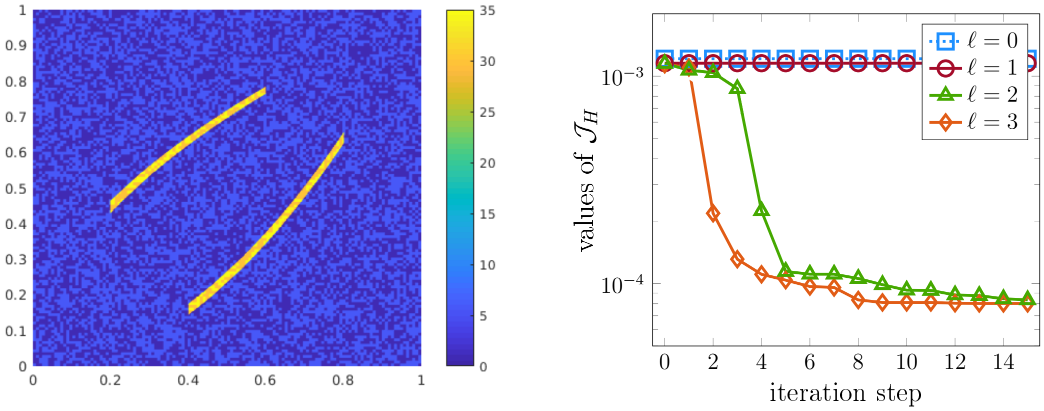


Figure 4.6: Diffusion coefficient in Example 2 (left) and values of $\mathcal{J}_H$ in the first 15 iterations of the inversion algorithm using sparsity patterns based on local matrices ($\ell = 0$) and quasi-local matrices with $\ell \in \{1, 2, 3\}$ (right).

4.3.2 Example 2: incomplete boundary data

Next, we consider a more realistic case where the operator $\tilde{\mathfrak{L}}^{\mathrm{eff}}$ is only partially known. In practice, this means that coarse measurements in D are available only for k distinct boundary conditions in X_H ($k < \dim X_H$). In this setting, the aim is to find an effective model that not only fits the given data, but that is also able to reproduce the coarse behavior for other boundary conditions not considered as input data.

The scalar coefficient A whose corresponding stiffness matrix should be recovered is shown in Figure 4.6 (left). We set $H = 2^{-5}$, $f = 1$, $k = 40$, and the initial matrix S_H^0 is defined as the first-order FE stiffness matrix based on a sample of an independent and uniformly distributed random coefficient on the coarse scale H with values between 0.1 and 10.

We adapt the *randomized approach* described in [OY19] in the context of deep learning. Namely, in each iteration step, we randomly choose half of the available data to compute the new search direction, whereas we use all available data for the line search and for the evaluation of the functional $\mathcal{J}_H$. The values of the error functional $\mathcal{J}_H$ in the first 15 iterations of the inversion algorithm are shown in Figure 4.6 (right). We observe that classical local stiffness matrices and even quasi-local ones with $\ell = 1$ cannot significantly improve the results obtained with the initial guess, while quasi-local matrices with $\ell \geq 2$ are able to reduce the values of the functional up to a certain degree.

As in the previous subsection, we validate the outcome of the inversion algorithm by solving a diffusion problem using the reconstructed stiffness matrices. Then we compare the numerical results with the corresponding fine FE solutions. The cross sections at $x_2 = 0.5$ and $x_1 = 0.5$ of the numerical approximations based on the different stiffness matrices are shown in Figure 4.7 in the case

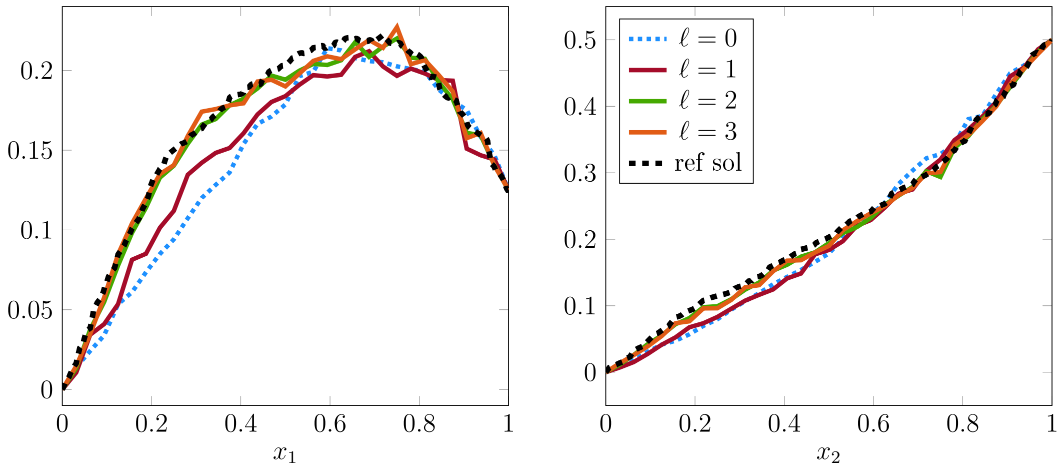


Figure 4.7: Cross sections at $x_2 = 0.5$ (left) and at $x_1 = 0.5$ (right) of reconstructed functions with the boundary condition $u_0(x) = x_1\, x_2^3$ based on local stiffness matrices ($\ell = 0$) and quasi-local ones with $\ell \in \{1, 2, 3\}$ for Example 2 obtained from incomplete boundary data and with the randomized approach. The corresponding fine FE function is depicted as a reference but was not part of the input data.

with the boundary condition $u_0(x) = x_1\, x_2^3$. We emphasize that, in this setting, neither the reference FE function (black dotted line in Figure 4.7) nor a coarse measurement from it were part of the input data. As expected from the values of $\mathcal{J}_H$, the reconstructions based on the matrices with $\ell \in \{2, 3\}$ are close and better approximate the behavior of the fine-scale solution than the matrices with $\ell \in \{0, 1\}$. The clear gap between $\ell = 1$ and $\ell = 2$ in this example may be explained by the structure of the coefficient. That is, a significant improvement of the results compared to the initial guess can only be achieved if the model is able to capture the two cracks, which probably only holds true for $\ell \geq 2$.

For a further comparison, we also present in Figure 4.8 the same cross sections of the numerical solutions obtained from the stiffness matrices using a *full-data approach*, i.e., when all available data (40 measurements) are used in every step to compute the new search direction. The reconstructed matrices behave similarly to the ones obtained with the randomized approach. However, it is worth mentioning that the randomized strategy is generally more robust in the case of incomplete boundary data and additionally requires less computational effort.

The presented inversion results demonstrate that the reconstruction of a stiffness matrix assuming a fixed local sparsity pattern of classical first-order finite elements does not allow capturing macroscopic features of solutions to a problem with underlying microscopic coefficient, while the reconstruction based on a quasi-local approach, especially with $\ell \geq 2$, is able to mimic the effective behavior quite well.

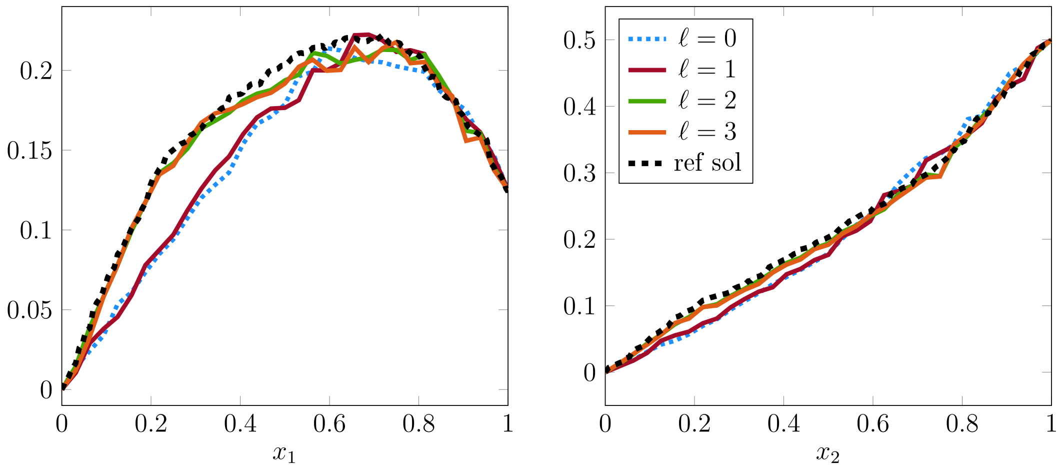


Figure 4.8: Cross sections at $x_2 = 0.5$ (left) and at $x_1 = 0.5$ (right) of reconstructed functions with the boundary condition $u_0(x) = x_1\,x_2^3$ based on local stiffness matrices ($\ell = 0$) and quasi-local ones with $\ell \in \{1, 2, 3\}$ for Example 2 obtained from incomplete boundary data and with the full-data approach. The corresponding fine FE function is depicted as a reference but was not part of the input data.

Furthermore, the quasi-local approach appears to be robust with respect to different right hand sides, a property which allows us to employ the reconstructed effective model for the simulation of other scenarios, assuming that the microscopic properties remain unchanged.

With regard to the forward setting, our experiments indicate that the use of quasi-local approaches in the presence of general multiscale coefficients is justified and maybe even necessary to obtain reasonable approximations on a coarse scale of interest. In that sense, our findings deviate from the numerical results in [GGS12], which indicate that truly local numerical homogenization might always be possible. We emphasize, however, that of course also the larger number of DOFs contributes to the better behavior of the quasi-local models in the presented inversion procedure. That is, the results by no means depreciate local homogenization approaches in general.

5 Fast Time-Explicit Multiscale Wave Propagation

In the previous chapters, we have seen how the LOD can be applied to stationary second-order linear PDEs which include microscopic features that need to be taken into account to obtain a sufficiently accurate approximation on some coarse scale of interest. Further, we have seen that the quasi-local concept of the method seems to be reasonable in the stationary setting, which becomes evident from the results of the inverse procedure described in Chapter 4.

As a next step, we extend the class of model problems and consider non-stationary problems, i.e., PDEs that depend not only on spatial variables but also on a temporal one. While the microscopic information in such a setting might depend on time as well, we restrict ourselves to the case where involved coefficients only depend on the spatial variables. The common approach to handle such problems is to apply the LOD (or any other multiscale technique for stationary equations) to the stationary part of the PDE to construct a time-independent coarse space which includes fine-scale information. Combining this spatial discretization with a suitable time-stepping approach then leads to a fully discrete method.

We emphasize that time-dependent problems allow us to exploit the full potential of the LOD. This is connected to the fact that, as described above, the technique is applied to the stationary part of the PDE and a corresponding multiscale space is only computed once in the so-called *offline stage*. Due to the coarse nature of such a space, the size of the respective system matrices is much smaller compared to an approximation space on a finer scale. This is extremely valuable for the *online stage*, where only linear systems based on the smaller matrices need to be solved in every time step without any further fine-scale simulations.

Note that although this general approach can be applied to any second-order PDE with temporal and spatial variables, the error analysis generally differs dependent on the PDE and the chosen time discretization. In this chapter, we use the approach in connection with the acoustic wave equation and an explicit time discretization scheme. The corresponding framework is introduced in the following section.

5.1 The heterogeneous wave equation

We consider the wave equation on a time interval $[0,T]$ given by

$$\begin{aligned} \partial_t^2 u - \operatorname{div}(A\nabla u) &= f && \text{in } (0,T]\times D,\\ u(0) &= u^0 && \text{in } D,\\ \partial_t u(0) &= v^0 && \text{in } D,\\ u|_\Gamma &= 0 && \text{in } (0,T],\\ \nabla u\cdot\nu|_{\partial D\setminus\Gamma} &= 0 && \text{in } (0,T], \end{aligned} \tag{5.1}$$

where $D \subseteq \mathbb{R}^d$, $d\in\{2,3\}$, is a polytopal, convex, bounded Lipschitz domain with outer normal ν and Dirichlet boundary $\Gamma\subseteq\partial D$ with $|\Gamma|>0$. Further, we assume to have initial data $u^0\in\mathcal{V}=H^1_\Gamma(D)$, $v^0\in\mathcal{H}=L^2(D)$ (cf. Section 2.1), and a time-independent rough coefficient $A\in\mathfrak{A}$ as defined in (3.2). As before, we have in mind coefficients that vary on some small scale ϵ but we do not need restrictive structural assumptions such as periodicity or scale separation.

Before we introduce and analyze the variational form corresponding to (5.1), we need to clarify some notation used in the context of time-dependent formulations.

First, we recall that by the Friedrichs inequality $\|\nabla\cdot\|_{L^2(D)}$ is a norm in $H^1_\Gamma(D)$. Moreover, we introduce the notation $H^{-1}(D):=(H^1_\Gamma(D))^*$ and write $L^p(0,T;X)$ for the *Bochner space* (see, e.g., [Eva10, Sec. 5.9.2]) with the norm

$$\begin{aligned} \|v\|_{L^p(0,T;X)} &:= \left(\int_0^T \|v\|_X^p \,\mathrm{d}t\right)^{1/p}, \quad 1\le p<\infty,\\ \|v\|_{L^\infty(0,T;X)} &:= \operatorname*{ess\,sup}_{0\le t\le T}\|v\|_X, \end{aligned}$$

where X is a Banach space equipped with the norm $\|\cdot\|_X$. The notation $v\in H^k(0,T;X)$, $k\in\mathbb{N}$, is used to denote that v and its weak time derivatives $\partial_t^j v$ for $j\in\{1,\dots,k\}$ are elements of the space $L^2(0,T;X)$. The Bochner space of functions that are continuous in time on the interval $[0,T]$ is denoted with $C([0,T];X)$ and equipped with the norm

$$\|v\|_{C([0,T];X)} := \max_{0\le t\le T}\|v\|_X.$$

As above, we write $v\in C^k([0,T];X)$, $k\in\mathbb{N}$, if v and $\partial_t^j v$ for $j\in\{1,\dots,k\}$ are elements of the space $C([0,T];X)$.

In order to compute a numerical approximation of solutions of (5.1), we write the problem in variational form, i.e., we seek a weak solution $u\in L^2(0,T;H^1_\Gamma(D))$ with $\partial_t u\in L^2(0,T;L^2(D))$ and $\partial_t^2 u\in L^2(0,T;H^{-1}(D))$ such that

$$\langle\partial_t^2 u, v\rangle_{H^{-1}(D)\times H^1_\Gamma(D)} + a(u,v) = (f,v)_{L^2(D)} \tag{5.2}$$

for all $v \in H^1_\Gamma(D)$ with initial conditions $u(0) = u^0$ and $\partial_t u(0) = v^0$, where a again denotes the bilinear form

$$a(v,w) = \int_D A\nabla v \cdot \nabla w \,\mathrm{d}x$$

and $\langle \cdot,\cdot \rangle_{H^{-1}(D)\times H^1_\Gamma(D)}$ is the dual pairing between $H^{-1}(D)$ and $H^1_\Gamma(D)$. Note that for any $u^0 \in H^1_\Gamma(D)$, $v^0 \in L^2(D)$, and $f \in L^2(0,T;L^2(D))$, there exists a unique weak solution u of problem (5.2). A proof of this property can be found, e.g., in [Eva10, Thm. 3 & 4 in Sec. 7.2.2], which also holds for non-smooth coefficients. Restricting the solution space $H^1_\Gamma(D)$ in (5.2) to a FE space V_h based on a regular and quasi-uniform mesh $\mathcal{T}_h$ of D with mesh size h (see Section 2.2) and applying the *leapfrog scheme* with step size τ in time, we obtain the following discrete problem: given $u_h^0 \in V_h$ and $u_h^1 \in V_h$, find $\{u_h^n\}_{n=0}^N$ with $u_h^n \in V_h$ such that

$$\tau^{-2}\,(u_h^{n+1} - 2u_h^n + u_h^{n-1}, v_h)_{L^2(D)} + a(u_h^n, v_h) = (f(n\tau), v_h)_{L^2(D)} \tag{5.3}$$

for all $v_h \in V_h$ and $n \in \{1,\dots,N-1\}$, where $N := T/\tau$ is the number of time steps. For simplicity, we assume that T and τ are such that $T/\tau \in \mathbb{N}$. It is well understood that the method defined in (5.3) only leads to acceptable results if the mesh size h is small enough to resolve the fine-scale features in space originating from the highly varying coefficient A. In order to obtain a sufficiently accurate approximation of the solution of (5.2), at least $h < \epsilon$ should hold. Such an h, however, may be too small to allow for reasonably fast computations. It is especially very restrictive since reducing the size of h directly leads to larger systems of linear equations that need to be solved in every time step. Furthermore, the fact that the above method (5.3) is explicit in time also introduces the so-called Courant-Friedrichs-Lewy (CFL) condition that limits the time step size τ by the mesh size h, i.e., $\tau \lesssim h$. It is, hence, too expensive to pose the discrete problem on meshes with small mesh sizes h that resolve fine-scale features.

Based on the LOD method described in the previous chapters, we introduce a way to cope with the fine-scale characteristics on an arbitrarily chosen coarse scale H which reduces the size of linear systems and enables larger time steps subject to a relaxed CFL condition $\tau \lesssim H$. Thus, the reduced computational complexity with respect to the spatial variable comes along with a complexity reduction in time. We emphasize that the results of this chapter were first presented in [MP19]. Besides, the LOD approach was already successfully applied in connection with electromagnetic waves [GHV18,Ver17] and time-harmonic wave propagation to eliminate the pollution effect [Pet17,GP15,BGP17]. Further, it was used in [PS17] for the wave equation with a constant coefficient to relax the time step restriction on adaptively refined meshes. For the wave equation (5.1) with rough coefficients, the LOD was used in combination with an implicit time discretization (Crank-Nicolson) in [AH17]. Therein, the need for additional regularity assumptions on the initial data was discussed, which is also crucial for the explicit time discretization in our case. As mentioned above, there are other possibilities to resolve the fine-scale features. The use of the HMM for the spatial discretization is for instance considered in [AG11, EHR11] or in [EHR12, AR14] in the context of wave propagation over long time. However, the corresponding analysis requires additional assumptions such as scale separation. Another method for the numerical homogenization of the wave equation can be found in [OZ08]. There, the idea is to use a harmonic coordinate transformation in order to obtain higher regularity of the weak solution. The main drawbacks of this approach are the necessary assumptions (so-called Cordes-type conditions) that are hard to verify, and the approximation of the coordinate transformation for which global fine-scale problems need to be solved. Another approach by the same authors is presented in [OZB14] based on RPS. The approach in [OZ17] based on gamblets shows the possible generalization of the present approach to a multilevel setting.

In general, any of the methods mentioned above can be used for the spatial discretization. The advantage of the LOD is that it preserves the finite element structure of the problem and is thus very convenient for practical applications. Then again, the use of an explicit time stepping scheme is motivated by its simple nature that allows for faster computations in every time step and by the fact that the discrete energy is conserved; cf. (5.7) below. Since solutions of the wave equation conserve energy in the continuous setting, such a property is very natural and desirable in the discrete setting as well.

5.2 The ideal method

In this section, we apply the ideas of Chapter 2 to the stationary part of the wave equation, i.e., we compute correctors and a corrected coarse FE space based on the bilinear form a. Since a is coercive and bounded in $H^1_\Gamma(D)$, the results of Chapter 2 may be applied without additional assumptions. Before writing down the ideal method where the classical LOD approach is combined with a time-stepping scheme, we briefly recall the main definitions and prove two auxiliary results in the following subsection.

5.2.1 Numerical upscaling by LOD

We consider, as before, a family of regular decompositions $\{\mathcal{T}_H\}_{H>0}$ of the domain D into quasi-uniform d-rectangles with mesh parameter H (cf. Section 2.2) and denote with V_H the corresponding conforming Q_1 FE space. For a linear and projective quasi-interpolation operator $\mathcal{I}_H \colon L^2(D) \to V_H$ as in Section 2.2.2, we define the fine-scale space $\mathcal{W} := \ker \mathcal{I}_H|_{H^1_\Gamma(D)}$ and the correction operator $\mathcal{C} \colon H^1_\Gamma(D) \to \mathcal{W}$ by

$$a(\mathcal{C}v, w) = a(v, w) \tag{5.4}$$

for all $w \in \mathcal{W}$. As in Chapter 2, we further define the ideal multiscale space $\tilde{V}_H := (\mathtt{id} - \mathcal{C})V_H = \mathcal{R}V_H$ and recall that $H^1_\Gamma(D) = \tilde{V}_H \oplus \mathcal{W}$ and $a(\tilde{V}_H, \mathcal{W}) = 0$ by construction.

Next, we prove that an inverse inequality holds in $\tilde{V}_H$ similarly to the classical one in the space V_H.

Lemma 5.2.1 (Inverse inequality). *For any $\tilde{v}_H \in \tilde{V}_H$, we have that*

$$\|\nabla \tilde{v}_H\|_{L^2(D)} \leq \tilde{C}_{\mathrm{inv}} H^{-1} \|\tilde{v}_H\|_{L^2(D)}. \tag{5.5}$$

The constant $\tilde{C}_{\mathrm{inv}}$ only depends on the constant C_{inv} in (2.35)*, the operator $\mathcal{I}_H$, and the contrast β/α, where α and β are the lower and upper bounds on A as quantified in Section 3.1.*

Proof. Let $\tilde{v}_H \in \tilde{V}_H$. Since $\tilde{v}_H = (1 - \mathcal{C})\mathcal{I}_H \tilde{v}_H$, we get

$$\begin{aligned}\alpha \|\nabla \tilde{v}_H\|^2_{L^2(D)} &\leq a(\tilde{v}_H, \tilde{v}_H) = a(\tilde{v}_H, \mathcal{I}_H \tilde{v}_H) \leq \beta \|\nabla \tilde{v}_H\|_{L^2(D)} \|\nabla \mathcal{I}_H \tilde{v}_H\|_{L^2(D)} \\ &\leq \beta \|\nabla \tilde{v}_H\|_{L^2(D)} C_{\mathrm{inv}} C_{\mathcal{I}_H} H^{-1} \|\tilde{v}_H\|_{L^2(D)}\end{aligned}$$

using (2.11) and the classical inverse inequality (2.35). Hence, (5.5) follows with $\tilde{C}_{\mathrm{inv}} := C_{\mathrm{inv}} C_{\mathcal{I}_H} \beta/\alpha$. □

Besides, the new space $\tilde{V}_H$ also has the following approximation property, which is a generalization of [PS17, Lem. 2.1] to the case of non-constant coefficients reusing ideas from Theorem 2.3.1.

Lemma 5.2.2. *For all $w \in H^1_\Gamma(D)$ with $\operatorname{div} A\nabla w \in L^2(D)$, it holds that*

$$\inf_{\tilde{v}_H \in \tilde{V}_H} \|\nabla(w - \tilde{v}_H)\|_{L^2(D)} \leq \alpha^{-1} C_{\mathcal{I}_H} H \,\|\operatorname{div} A\nabla w\|_{L^2(D)}.$$

Proof. Let $w \in H^1_\Gamma(D)$. Further, let $\tilde{w}_H \in \tilde{V}_H$ be the orthogonal projection with respect to the bilinear form a of w onto $\tilde{V}_H$, i.e.,

$$a(\tilde{w}_H, \tilde{v}_H) = a(w, \tilde{v}_H)$$

for all $\tilde{v}_H \in \tilde{V}_H$. Therefore, $e_H = w - \tilde{w}_H \in \mathcal{W}$ and, hence,

$$\begin{aligned}\alpha \|\nabla e_H\|^2_{L^2(D)} &\leq a(e_H, e_H) = a(w, e_H) = (-\operatorname{div} A\nabla w, e_H)_{L^2(D)} \\ &\leq \|\operatorname{div} A\nabla w\|_{L^2(D)} \|e_H\|_{L^2(D)}.\end{aligned}$$

Since $e_H \in \mathcal{W}$, it holds that

$$\|e_H\|_{L^2(D)} = \|(\mathtt{id} - \mathcal{I}_H)e_H\|_{L^2(D)} \leq C_{\mathcal{I}_H} H \|\nabla e_H\|_{L^2(D)},$$

where we use the approximation property (2.14). Combining both inequalities results in

$$\|\nabla(w - \tilde{w}_H)\|_{L^2(D)} \leq \alpha^{-1} C_{\mathcal{I}_H} H \,\|\operatorname{div} A\nabla w\|_{L^2(D)},$$

which concludes the proof. □

5.2.2 Discretization in time

Based on the adapted discrete space $\tilde{V}_H$ and the leapfrog scheme with time step τ as in (5.3), the proposed *ideal method* reads: given $\tilde{u}_H^0 = (\mathtt{id} - \mathcal{C})\mathcal{I}_H u^0$ and suitable $\tilde{u}_H^1 \in \tilde{V}_H$, find $\{\tilde{u}_H^n\}_{n=0}^N$ with $\tilde{u}_H^n \in \tilde{V}_H$ such that

$$\tau^{-2}\,(\tilde{u}_H^{n+1} - 2\tilde{u}_H^n + \tilde{u}_H^{n-1}, \tilde{v}_H)_{L^2(D)} + a(\tilde{u}_H^n, \tilde{v}_H) = (f(n\tau), \tilde{v}_H)_{L^2(D)} \tag{5.6}$$

for all $\tilde{v}_H \in \tilde{V}_H$ and $n \in \{1, \dots, N-1\}$. We emphasize that (5.6) is called ideal method because we implicitly assume that the corrector problems (5.4) can be computed exactly. In order to show stability and error estimates for this scheme, standard methods [Chr09, Jol03] can be applied. First, we introduce the *discrete energy*

$$\mathcal{E}^{n+1} := \frac{1}{2}\Big(\|D_\tau \tilde{u}_H^{n+1}\|^2_{L^2(D)} + a(\tilde{u}_H^n, \tilde{u}_H^{n+1})\Big),$$

where $D_\tau \tilde{u}_H^{n+1} := (\tilde{u}_H^{n+1} - \tilde{u}_H^n)/\tau$ denotes the discrete time derivative. Using the test function $\tilde{v}_H = \tilde{u}_H^{n+1} - \tilde{u}_H^{n-1}$ in (5.6), we derive energy conservation in the sense that

$$\begin{aligned}
&\tau\,(f(n\tau), D_\tau \tilde{u}_H^{n+1} + D_\tau \tilde{u}_H^n)_{L^2(D)} \\
&\quad = \tau^{-2}\,(\tilde{u}_H^{n+1} - 2\tilde{u}_H^n + \tilde{u}_H^{n-1}, \tilde{u}_H^{n+1} - \tilde{u}_H^{n-1})_{L^2(D)} + a(\tilde{u}_H^n, \tilde{u}_H^{n+1} - \tilde{u}_H^{n-1}) \\
&\quad = 2\,(\mathcal{E}^{n+1} - \mathcal{E}^n).
\end{aligned} \tag{5.7}$$

Therefore, if no external force is applied, i.e., $f = 0$, the discrete initial energy $\mathcal{E}^1$ is conserved over time.

Lemma 5.2.3 (Stability of the ideal method)**.** *Assume that the CFL condition*

$$1 - \frac{1}{2}\beta C_{\mathrm{inv}}^2 C_{\mathcal{I}_H}^2 H^{-2}\tau^2 \geq \delta \tag{5.8}$$

holds for some $\delta > 0$. Then the ideal method (5.6) *is stable, i.e., it holds that*

$$\begin{aligned}
&\|D_\tau \tilde{u}_H^{n+1}\|_{L^2(D)} + \|\nabla \tilde{u}_H^{n+1}\|_{L^2(D)} \\
&\quad \leq C_{\mathrm{stab}}\Big(\sum_{k=1}^{n} \tau\,\|f(k\tau)\|_{L^2(D)} + \|D_\tau \tilde{u}_H^1\|_{L^2(D)} + \|\nabla \tilde{u}_H^0\|_{L^2(D)} + \|\nabla \tilde{u}_H^1\|_{L^2(D)}\Big)
\end{aligned} \tag{5.9}$$

for all $n \in \{0, \dots, N-1\}$, where the constant C_{stab} depends on α, β, and δ only.

Proof. The proof mainly follows the ideas presented in [Chr09, Jol03], generalized to the case of arbitrary coefficients. Using the inverse inequality (2.35), the boundedness of the bilinear form a, and

$$a((\mathtt{id} - \mathcal{C})v_H, (\mathtt{id} - \mathcal{C})v_H) = a(v_H, v_H) - a(\mathcal{C}v_H, \mathcal{C}v_H) \leq \beta\,\|\nabla v_H\|^2_{L^2(D)}$$

for any $v_H \in V_H$, we have that

$$\begin{aligned}\mathcal{E}^{n+1} &= \frac{1}{2}\left(\|D_\tau \tilde{u}_H^{n+1}\|^2_{L^2(D)} + a(\tilde{u}_H^n, \tilde{u}_H^{n+1})\right)\\ &= \frac{1}{4}\,a(\tilde{u}_H^{n+1}, \tilde{u}_H^{n+1}) + \frac{1}{4}\,a(\tilde{u}_H^n, \tilde{u}_H^n)\\ &\quad - \frac{1}{4}\,a(\tilde{u}_H^{n+1} - \tilde{u}_H^n, \tilde{u}_H^{n+1} - \tilde{u}_H^n) + \frac{1}{2}\,\|D_\tau \tilde{u}_H^{n+1}\|^2_{L^2(D)}\\ &\geq \frac{1}{4}\,a(\tilde{u}_H^{n+1}, \tilde{u}_H^{n+1}) + \frac{1}{4}\,a(\tilde{u}_H^n, \tilde{u}_H^n)\\ &\quad + \frac{1}{2}\left(1 - \frac{1}{2}\beta C_{\mathrm{inv}}^2 C_{\mathcal{I}_H}^2 H^{-2}\tau^2\right)\|D_\tau \tilde{u}_H^{n+1}\|^2_{L^2(D)}.\end{aligned}$$

Therefore, the CFL condition (5.8) ensures positivity of the discrete energy since

$$\mathcal{E}^{n+1} \geq \frac{1}{4}\,a(\tilde{u}_H^{n+1}, \tilde{u}_H^{n+1}) + \frac{1}{4}\,a(\tilde{u}_H^n, \tilde{u}_H^n) + \frac{\delta}{2}\,\|D_\tau \tilde{u}_H^{n+1}\|^2_{L^2(D)}. \tag{5.10}$$

Employing (5.7) and the inequality (5.10), we get the estimate

$$\begin{aligned}\mathcal{E}^{n+1} - \mathcal{E}^n &= \frac{1}{2}\,\tau\,(f(n\tau), D_\tau \tilde{u}_H^{n+1} + D_\tau \tilde{u}_H^n)_{L^2(D)}\\ &\leq \frac{1}{2}\,\tau\,\|f(n\tau)\|_{L^2(D)}\left(\|D_\tau \tilde{u}_H^{n+1}\|_{L^2(D)} + \|D_\tau \tilde{u}_H^n\|_{L^2(D)}\right)\\ &\leq \frac{1}{\sqrt{2\delta}}\,\tau\,\|f(n\tau)\|_{L^2(D)}\left(\sqrt{\mathcal{E}^{n+1}} + \sqrt{\mathcal{E}^n}\,\right).\end{aligned}$$

This yields

$$\sqrt{\mathcal{E}^{n+1}} \leq \sqrt{\mathcal{E}^n} + \frac{1}{\sqrt{2\delta}}\,\tau\,\|f(n\tau)\|_{L^2(D)}$$

and, hence, the stability estimate

$$\sqrt{\mathcal{E}^{n+1}} \leq \sqrt{\mathcal{E}^1} + \frac{1}{\sqrt{2\delta}}\sum_{k=1}^{n}\tau\,\|f(k\tau)\|_{L^2(D)}.$$

This implies (5.9). □

Apart from the stability of $\tilde{u}_H^n$, the estimate (5.9) also provides a tool for the estimation of the error $\tilde{u}_H^n - u(t_n)$ in the following subsection. Note that the constant C_{stab}, and thus also the constant in the error bounds later on, depends on the contrast β/α. However, this dependence seems pessimistic in many cases of practical relevance; see, e.g., [PS16, HM17].

5.2.3 Error analysis

In this subsection, we derive an error estimate for the ideal method (5.6) provided that suitable regularity assumptions hold. The assumptions are met for relevant classes of problems with arbitrarily rough coefficients that are characterized by the right-hand side f and the initial conditions.

Assumption 5.2.4 (Initial regularity)**.** Suppose that the right-hand side f and the initial data in problem (5.2) satisfy the conditions

(A0) $f \in H^3(0,T;L^2(D))$,

(A1) $\partial_t u(0) = v^0 \in H^1_\Gamma(D)$,

(A2) $\partial_t^2 u(0) = f(0) + \operatorname{div} A\nabla u^0 \in H^1_\Gamma(D)$,

(A3) $\partial_t^3 u(0) = \partial_t f(0) + \operatorname{div} A\nabla v^0 \in H^1_\Gamma(D)$,

(A4) $\partial_t^4 u(0) = \partial_t^2 f(0) + \operatorname{div} A\nabla \partial_t^2 u(0) \in L^2(D)$.

Further, assume that the corresponding norms can be bounded independently of the fine scale ϵ on which A varies, i.e., that there exists a constant C_{data} (possibly dependent on T) such that

$$\|f\|_{H^3(0,T;L^2(D))} + \sum_{j=0}^{3} \|\nabla \partial_t^j u(0)\|_{L^2(D)} + \|\partial_t^4 u(0)\|_{L^2(D)} \leq C_{\text{data}}. \tag{5.11}$$

Remark 5.2.5. The regularity assumptions in Assumption 5.2.4 on the initial data and the right-hand side correspond to the conditions in [AH17] for the implicit setting and are referred to as *well-prepared and compatible of order* 3.

Under these assumptions, we can formulate an error estimate for the ideal method.

Theorem 5.2.6 (Error of the ideal method)**.** *Suppose that Assumption 5.2.4 holds and define $t_n := \tau n$ for $n \in \{0,\dots,N\}$. Then the solutions u of* (5.2) *and $\tilde{u}_H^{n+1}$ of* (5.6) *satisfy the error bound*

$$\begin{aligned} \Big\| D_\tau \tilde{u}_H^{n+1} - \frac{u(t_{n+1}) - u(t_n)}{\tau} \Big\|_{L^2(D)} + \big\|\nabla\big(\tilde{u}_H^{n+1} - u(t_{n+1})\big)\big\|_{L^2(D)} \\ \lesssim_T (H + \tau^2)\, C_{\text{data}} \end{aligned} \tag{5.12}$$

for $n \in \{0,\dots,N-1\}$.

Proof. Differentiating (5.1) with respect to time and using Assumption 5.2.4 shows that the time derivatives of u solve wave-type equations as well. As in [Eva10, Thm. 6 in Sec. 7.2.6], we get the regularity $u \in H^4(0,T;L^2(D))$ and it follows that $u \in C^3([0,T];L^2(D))$. It further holds that $u \in H^3(0,T;H^1_\Gamma(D))$ and thus $u \in C^2([0,T];H^1_\Gamma(D))$. These regularity properties are required in the estimates below.

Next, we define $\tilde{z}_H \in C^2([0,T];\tilde{V}_H)$ as the auxiliary semi-discrete solution of (5.2) which solves

$$(\partial_t^2 \tilde{z}_H(t), \tilde{v}_H)_{L^2(D)} + a(\tilde{z}_H(t), \tilde{v}_H) = (f(t), \tilde{v}_H)_{L^2(D)} \tag{5.13}$$

for all $\tilde{v}_H \in \tilde{V}_H$ and $t \in [0,T]$, with the initial conditions $\tilde{z}_H(0) = (\mathtt{id} - \mathcal{C})\mathcal{I}_H u^0$ and $\partial_t \tilde{z}_H(0) = (\mathtt{id} - \mathcal{C})\mathcal{I}_H v^0$. Similarly to the observations in [Eva10, Sec. 7.2.2], the well-posedness of (5.13) follows from standard theory for ordinary differential equations (ODEs) using the regularity assumptions on the initial data and the right-hand side. We then split the error into

$$\tilde{u}_H^n - u(t_n) = e^n + \big(\tilde{z}_H(t_n) - \Pi_{\tilde{V}_H} u(t_n)\big) - \rho(t_n) \tag{5.14}$$

with the temporal discretization error $e^n := \tilde{u}_H^n - \tilde{z}_H(t_n)$ and the spatial best-approximation error $\rho(t) := u(t) - \Pi_{\tilde{V}_H} u(t)$ for any $t \in [0,T]$. Here, $\Pi_{\tilde{V}_H} u(t)$ denotes the orthogonal projection of $u(t)$ onto $\tilde{V}_H$ with respect to the bilinear form a. First, we observe that e^n solves

$$\begin{aligned}\tau^{-2}\,(e^{n+1} - 2e^n + e^{n-1}, \tilde{v}_H)_{L^2(D)} + a(e^n, \tilde{v}_H)&\\ = \big(\partial_t^2 \tilde{z}_H(t_n) - \tau^{-2}\,\big(\tilde{z}_H(t_{n+1}) - 2\tilde{z}_H(t_n) + \tilde{z}_H(t_{n-1})\big), \tilde{v}_H\big)_{L^2(D)}&\end{aligned}$$

for all $\tilde{v}_H \in \tilde{V}_H$. Therefore, we get with Lemma 5.2.3 that

$$\begin{aligned}&\|D_\tau e^{n+1}\|_{L^2(D)} + \|\nabla e^{n+1}\|_{L^2(D)}\\ &\quad\le C_{\mathrm{stab}}\Big(\|D_\tau e^1\|_{L^2(D)} + \|\nabla e^1\|_{L^2(D)}\\ &\qquad\qquad + \sum_{k=1}^{n} \tau\,\Big\|\partial_t^2 \tilde{z}_H(t_k) - \frac{\tilde{z}_H(t_{k+1}) - 2\tilde{z}_H(t_k) + \tilde{z}_H(t_{k-1})}{\tau^2}\Big\|_{L^2(D)}\Big).\end{aligned} \tag{5.15}$$

Second, $\tilde{z}_H - \Pi_{\tilde{V}_H} u$ solves

$$(\partial_t^2 \tilde{z}_H(t) - \partial_t^2 \Pi_{\tilde{V}_H} u(t), \tilde{v}_H)_{L^2(D)} + a(\tilde{z}_H(t) - \Pi_{\tilde{V}_H} u(t), \tilde{v}_H) = (\partial_t^2 \rho(t), \tilde{v}_H)_{L^2(D)}$$

for all $\tilde{v}_H \in \tilde{V}_H$ and all $t \in [0,T]$. As in [Jol03], we thus get

$$\begin{aligned}&\|\partial_t \tilde{z}_H(t) - \partial_t \Pi_{\tilde{V}_H} u(t)\|_{L^2(D)} + \|\nabla\big(\tilde{z}_H(t) - \Pi_{\tilde{V}_H} u(t)\big)\|_{L^2(D)}\\ &\quad\le C_{\mathrm{stab}}\Big(\|\partial_t \tilde{z}_H(0) - \partial_t \Pi_{\tilde{V}_H} u(0)\|_{L^2(D)} + \|\nabla\big(\tilde{z}_H(0) - \Pi_{\tilde{V}_H} u(0)\big)\|_{L^2(D)}\\ &\qquad\qquad + \int_0^t \|\partial_t^2 \rho(s)\|_{L^2(D)}\,\mathrm{d}s\Big)\\ &\quad= C_{\mathrm{stab}} \int_0^t \|\partial_t^2 \rho(s)\|_{L^2(D)}\,\mathrm{d}s,\end{aligned} \tag{5.16}$$

where we employ the equality $\Pi_{\tilde{V}_H} u(0) = (\mathtt{id} - \mathcal{C})\mathcal{I}_H u(0) = \tilde{z}_H(0)$ as well as $\partial_t \Pi_{\tilde{V}_H} u(0) = (\mathtt{id} - \mathcal{C})\mathcal{I}_H \partial_t u(0) = \partial_t \tilde{z}_H(0)$, which follow from the definition of $\tilde{V}_H$. Further, there exists $\xi \in [t_n, t_{n+1}]$ such that

$$\frac{\tilde{z}_H(t_{n+1}) - \tilde{z}_H(t_n)}{\tau} - \frac{\Pi_{\tilde{V}_H} u(t_{n+1}) - \Pi_{\tilde{V}_H} u(t_n)}{\tau} = \partial_t \tilde{z}_H(\xi) - \partial_t \Pi_{\tilde{V}_H} u(\xi). \tag{5.17}$$

Combining (5.14)–(5.17), we get that

$$\begin{aligned}&\Big\|D_\tau \tilde{u}_H^{n+1} - \frac{u(t_{n+1})-u(t_n)}{\tau}\Big\|_{L^2(D)} + \|\nabla\big(\tilde{u}_H^{n+1} - u(t_{n+1})\big)\|_{L^2(D)}\\ &\quad\lesssim C_{\mathrm{stab}}\bigg(\|D_\tau e^1\|_{L^2(D)} + \|\nabla e^1\|_{L^2(D)} + \Big\|\frac{\rho(t_{n+1})-\rho(t_n)}{\tau}\Big\|_{L^2(D)}\\ &\qquad + \|\nabla\rho(t_{n+1})\|_{L^2(D)} + \int_0^{t_{n+1}} \|\partial_t^2\rho(s)\|_{L^2(D)}\,\mathrm{d}s\\ &\qquad + \sum_{k=1}^{n}\tau\,\Big\|\partial_t^2\tilde{z}_H(t_k) - \frac{\tilde{z}_H(t_{k+1})-2\tilde{z}_H(t_k)+\tilde{z}_H(t_{k-1})}{\tau^2}\Big\|_{L^2(D)}\bigg).\end{aligned} \tag{5.18}$$

With a Taylor expansion and an appropriate choice of $\tilde{u}_H^1$, we get

$$\|D_\tau e^1\|_{L^2(D)} + \|\nabla e^1\|_{L^2(D)} \lesssim \tau^2\,\|\tilde{z}_H\|_{C^3([0,T];H^1_\Gamma(D))}$$

and with Lemma 5.2.2, we have

$$\begin{aligned}\|\nabla\rho(t_{n+1})\|_{L^2(D)} &\lesssim H\,\|\operatorname{div} A\nabla u(t_{n+1})\|_{L^2(D)}\\ &\lesssim H\,\big(\|f\|_{C([0,T];L^2(D))} + \|u\|_{C^2([0,T];L^2(D))}\big).\end{aligned}$$

Further, we obtain the estimate

$$\Big\|\frac{\rho(t_{n+1})-\rho(t_n)}{\tau}\Big\|_{L^2(D)} + \int_0^{t_{n+1}} \|\partial_t^2\rho(s)\|_{L^2(D)}\,\mathrm{d}s \lesssim_T H\,\|u\|_{C^2([0,T];H^1(D))}$$

employing the approximation property (2.14). Note that we use $\lesssim_T$ to indicate an explicit dependence on T. Lastly, it holds that

$$\sum_{k=1}^{n}\tau\,\Big\|\partial_t^2\tilde{z}_H(t_k) - \frac{\tilde{z}_H(t_{k+1})-2\tilde{z}_H(t_k)+\tilde{z}_H(t_{k-1})}{\tau^2}\Big\|_{L^2(D)} \lesssim_T \tau^2\,\|\tilde{z}_H\|_{C^4([0,T];L^2(D))}$$

provided that $z_H \in C^4([0,T];L^2(D))$. To show this regularity of z_H, we differentiate (5.13) with respect to time as above and define suitable initial conditions

$$\partial_t^j\tilde{z}_H(0) = (\mathtt{id} - \mathcal{C})\mathcal{I}_H\partial_t^j u(0) \in \tilde{V}_H, \quad j \in \{0,\dots,4\}.$$

By standard ODE theory, solutions of equations of the form (5.13) are in $C^2([0,T];\tilde{V}_H)$. Therefore, it follows that $\tilde{z}_H \in C^4([0,T];\tilde{V}_H)$. Combining the above estimates with (5.18) and adapting the stability estimates provided in [Eva10, Sec. 7.2.3], we deduce (5.12). □

The regularity properties of the solution u that follow from Assumption 5.2.4 allow for a simplification of the method defined in (5.6) that is discussed in the following subsection.

5.2.4 A simplified method

To derive a variant of (5.6), we first observe that (5.6) can be written as an equation for standard FE functions $\{u_H^n\}_{n=0}^N$ with $u_H^n \in V_H$ using the explicit characterization $\tilde{u}_H^n = (\mathtt{id} - \mathcal{C})u_H^n = \mathcal{R}u_H^n$, i.e.,

$$\tau^{-2}\left(\mathcal{R}(u_H^{n+1} - 2u_H^n + u_H^{n-1}), \mathcal{R}v_H\right)_{L^2(D)} + a\big(\mathcal{R}u_H^n, \mathcal{R}v_H\big) = \big(f(n\tau), \mathcal{R}v_H\big)_{L^2(D)}$$

for all $v_H \in V_H$. A slightly modified method with reduced computational costs seeks $\{\bar{u}_H^n\}_{n=0}^N$ with $\bar{u}_H^n \in V_H$ such that

$$\tau^{-2}\left(\bar{u}_H^{n+1} - 2\bar{u}_H^n + \bar{u}_H^{n-1}, v_H\right) + a\big(\mathcal{R}\bar{u}_H^n, \mathcal{R}v_H\big)_{L^2(D)} = \big(f(n\tau), v_H\big)_{L^2(D)} \quad (5.19)$$

for all $v_H \in V_H$ and $n \in \{1, \ldots, N-1\}$ and given suitable initial conditions. Note that the solution of (5.19) fulfills stability properties similar to (5.9). Analogously to (5.12), we can show that

$$\Big\| D_\tau \mathcal{R}\bar{u}_H^{n+1} - \frac{u(t_{n+1}) - u(t_n)}{\tau}\Big\|_{L^2(D)} + \big\|\nabla\big(\mathcal{R}\bar{u}_H^{n+1} - u(t_{n+1})\big)\big\|_{L^2(D)} \lesssim_T (H + \tau^2)\, C_{\mathrm{data}}$$

for $n \in \{1, \ldots, N-1\}$ if the regularity properties in Assumption 5.2.4 hold. Hence, it is reasonable to use the simplified method in practice; see also Section 5.4.

Remark 5.2.7. The simplification in (5.19) might raise the question whether *mass lumping* is also a possible modification. The numerical experiments in Section 5.4 show that mass lumping generally works but might have an impact on the overall convergence rate.

5.3 The practical method

The method discussed in Section 5.2 is ideal in the sense that we implicitly assume that the corrector problems (5.4) can be solved exactly. In practice, those problems are discretized and localized as explained in the following subsections. Here, we do not replace $H^1_\Gamma(D)$ in the construction of Section 5.2 by a discrete FE space V_h on some fine mesh with parameter h as described in Section 2.4.4. Instead, we discretize the corrector problems (5.4) before we localize them and investigate the error introduced by each of these steps.

5.3.1 Discretization at the fine scale

As a first step, the problems (5.4) are discretized using classical Q_1 finite elements on a fine mesh. To quantify the error introduced by such a procedure, let $\{\tilde{u}_H^n\}_{n=0}^N$ with $\tilde{u}_H^n = (\mathtt{id} - \mathcal{C})u_H^n \in \tilde{V}_H$ be the solution of problem (5.6). Further, define for any $v_h \in V_h$, the discretized correction $\mathcal{C}_h v_h \in W_h$ as the finite element solution of (5.4), i.e.,

$$a(\mathcal{C}_h v_h, w_h) = a(v_h, w_h) \tag{5.20}$$

for all $w_h \in W_h$. This discretized version of (5.4) is posed in a discrete space $W_h \subseteq \mathcal{W}$ on a mesh $\mathcal{T}_h$ with mesh size $h < \epsilon$ that is assumed to be small enough to resolve variations of the coefficient A. Note that $W_h \subseteq V_h$, where V_h is the standard conforming Q_1 finite element space based on the mesh $\mathcal{T}_h$. Now, let $\{\tilde{u}^n_{H,h}\}_{n=0}^N$ with $\tilde{u}^n_{H,h} = (\mathtt{id} - \mathcal{C}_h)u^n_{H,h}$ be the solution of (5.6) posed in the space $\tilde{V}_{H,h} := (\mathtt{id} - \mathcal{C}_h)V_H$ instead of $\tilde{V}_H$ with suitable initial conditions. The following lemma quantifies the difference between these two solutions.

Lemma 5.3.1 (Fine-scale discretization error)**.** *Suppose that the assumptions of Lemma 5.2.3 and Assumption 5.2.4 hold. Then the discrete solutions $\tilde{u}^{n+1}_H \in \tilde{V}_H$ and $\tilde{u}^{n+1}_{H,h} \in \tilde{V}_{H,h}$ satisfy the error estimate*

$$\begin{aligned}\|D_\tau \tilde{u}^{n+1}_H - D_\tau \tilde{u}^{n+1}_{H,h}\|_{L^2(D)} &+ \|\nabla(\tilde{u}^{n+1}_H - \tilde{u}^{n+1}_{H,h})\|_{L^2(D)} \\ &\lesssim_T \left(d_{\tilde{V}_H}[V_h] + H^{-1}(d_{\tilde{V}_H}[V_h])^2\right) C_{\mathrm{data}}\end{aligned}$$

for all $n \in \{0, \dots, N-1\}$, where the approximation error $d_{\tilde{V}_H}[V_h]$ is defined by

$$d_{\tilde{V}_H}[V_h] := \sup_{v \in \tilde{V}_H} \inf_{v_h \in V_h} \frac{\|\nabla(v - v_h)\|_{L^2(D)}}{\|\nabla v\|_{L^2(D)}}.$$

Proof. Observe that the error $\tilde{e}^n = (\mathtt{id} - \mathcal{C})(u^n_H - u^n_{H,h})$ solves

$$\begin{aligned}\tau^{-2}\left(\tilde{e}^{n+1} - 2\tilde{e}^n + \tilde{e}^{n-1}, (\mathtt{id} - \mathcal{C})v_H\right)_{L^2(D)} &+ a\big(\tilde{e}^n, (\mathtt{id} - \mathcal{C})v_H\big) \\ = -\big(f(n\tau), (\mathcal{C} - \mathcal{C}_h)v_H\big)_{L^2(D)}& \\ + \tau^{-2}\left(\tilde{u}^{n+1}_{H,h} - 2\tilde{u}^n_{H,h} + \tilde{u}^{n-1}_{H,h}, (\mathcal{C} - \mathcal{C}_h)v_H\right)_{L^2(D)}& \\ + \tau^{-2}\left((\mathcal{C} - \mathcal{C}_h)(u^{n+1}_{H,h} - 2u^n_{H,h} + u^{n-1}_{H,h}), (\mathtt{id} - \mathcal{C})v_H\right)_{L^2(D)}& \\ + a\big((\mathcal{C} - \mathcal{C}_h)u^n_{H,h}, (\mathcal{C} - \mathcal{C}_h)v_H\big) =: F^n\big((\mathtt{id} - \mathcal{C})v_H\big)&\end{aligned}$$

for all $v_H \in V_H$. If $F^n|_{\tilde{V}_H} \in L^2(D)$, we can derive a bound on the error using Lemma 5.2.3. To show this, we first estimate $\|\nabla(\mathcal{C} - \mathcal{C}_h)v_H\|_{L^2(D)}$. Using the identity $W_h = V_h \cap \mathcal{W} = (\mathtt{id} - \mathcal{I}_H)V_h$ and the quasi-optimality of the solution $\mathcal{C}_h v_H$ defined in (5.20), we can show that

$$\begin{aligned}\|\nabla(\mathcal{C} - \mathcal{C}_h)v_H\|_{L^2(D)} &\leq (1 + C_{\mathcal{I}_H})\beta/\alpha \inf_{v_h \in V_h} \|\nabla(\mathcal{C}v_H - v_h)\|_{L^2(D)} \\ &\lesssim \sup_{v \in \tilde{V}_H} \inf_{v_h \in V_h} \frac{\|\nabla(v - v_h)\|_{L^2(D)}}{\|\nabla v\|_{L^2(D)}} \|\nabla \tilde{v}_H\|_{L^2(D)} \qquad (5.21) \\ &= d_{\tilde{V}_H}[V_h]\, \|\nabla \tilde{v}_H\|_{L^2(D)},\end{aligned}$$

where either $\tilde{v}_H = (\mathtt{id} - \mathcal{C})v_H$ or $\tilde{v}_H = (\mathtt{id} - \mathcal{C}_h)v_H$. Using (2.14), we further get

$$\|(\mathcal{C} - \mathcal{C}_h)v_H\|_{L^2(D)} \lesssim C_{\mathcal{I}_H} H\, d_{\tilde{V}_H}[V_h]\, \|\nabla \tilde{v}_H\|_{L^2(D)}. \tag{5.22}$$

With (5.21), (5.22), and the inverse inequality (5.5), we can derive the bound

$$\begin{aligned}\sup_{\tilde{v}_H \in \tilde{V}_H} \frac{|F^n(\tilde{v}_H)|}{\|\tilde{v}_H\|_{L^2(D)}} &\lesssim \Big(\|f(n\tau)\|_{L^2(D)} + \tau^{-2}\, \|\tilde{u}^{n+1}_{H,h} - 2\tilde{u}^n_{H,h} + \tilde{u}^{n-1}_{H,h}\|_{L^2(D)} \\ &\qquad + H^{-1} d_{\tilde{V}_H}[V_h]\, \|\nabla \tilde{u}^n_{H,h}\|_{L^2(D)}\Big)\, d_{\tilde{V}_H}[V_h] \\ &\lesssim \big(d_{\tilde{V}_H}[V_h] + H^{-1}(d_{\tilde{V}_H}[V_h])^2\big)\, C_{\mathrm{data}}.\end{aligned}$$

This implies that $F^n|_{\tilde{V}_H} \in L^2(D)$. Thus, using the above equations and the fact that, for any v_H, $w_H \in V_H$ and any suitable norm $\|\cdot\|$,

$$\|(\mathtt{id} - \mathcal{C})v_H - (\mathtt{id} - \mathcal{C}_h)w_H\| \leq \|(\mathtt{id} - \mathcal{C})(v_H - w_H)\| + \|(\mathcal{C} - \mathcal{C}_h)w_H\|,$$

it follows from Lemma 5.2.3 that

$$\begin{aligned}&\|D_\tau \tilde{u}^{n+1}_H - D_\tau \tilde{u}^{n+1}_{H,h}\|_{L^2(D)} + \|\nabla(\tilde{u}^{n+1}_H - \tilde{u}^{n+1}_{H,h})\|_{L^2(D)} \\ &\qquad\qquad \lesssim_T \big(d_{\tilde{V}_H}[V_h] + H^{-1}(d_{\tilde{V}_H}[V_h])^2\big)\, C_{\mathrm{data}}.\end{aligned}$$

Note that we employ the fact that $\tau^{-2}\,(\tilde{u}^{n+1}_{H,h} - 2\tilde{u}^n_{H,h} + \tilde{u}^{n-1}_{H,h})$ can be bounded in the L^2-norm independently of τ and H. This is due to the observation that $\{D_\tau \tilde{u}^{n+1}_{H,h}\}_{n=0}^{N-1}$ solves

$$\begin{aligned}\tau^{-2}\,\big(D_\tau \tilde{u}^{n+2}_{H,h} - 2D_\tau \tilde{u}^{n+1}_{H,h} + D_\tau \tilde{u}^n_{H,h}, \tilde{v}_H\big)_{L^2(D)} + a\big(D_\tau \tilde{u}^{n+1}_{H,h}, \tilde{v}_H\big) \\ = \tau^{-1}\,\big(f((n+1)\tau) - f(n\tau), \tilde{v}_H\big)_{L^2(D)}\end{aligned}$$

for all $\tilde{v}_H \in \tilde{V}_{H,h}$. Therefore, using Lemma 5.2.3 and the regularity conditions in Assumption 5.2.4, we can bound the L^2-norm of $\tau^{-2}\,(\tilde{u}^{n+1}_{H,h} - 2\tilde{u}^n_{H,h} + \tilde{u}^{n-1}_{H,h})$ in terms of the initial data and the right-hand side. □

5.3.2 Localized discrete corrections

As a next step, we define a fully discrete solution, which is actually computable, and quantify the error with respect to the discretized solution $\tilde{u}^n_{H,h}$ of the previous subsection. First, however, we need to introduce the localized version $\mathcal{C}^\ell_h\colon V_h \to W_h$ of the discretized correction operator $\mathcal{C}_h$ defined in (5.20). For the definition of a localized correction, we refer to Section 2.4.3. With the discretized and localized correction operator $\mathcal{C}^\ell_h$, we set $\tilde{V}^\ell_{H,h} := (\mathtt{id} - \mathcal{C}^\ell_h)V_H$ and define the *practical method* as follows: given $\tilde{u}^{\ell,0}_{H,h} = (\mathtt{id} - \mathcal{C}^\ell_h)\mathcal{I}_H u^0$ and suitable $\tilde{u}^{\ell,1}_{H,h} \in \tilde{V}^\ell_{H,h}$, find $\{\tilde{u}^{\ell,n}_{H,h}\}_{n=0}^N$ with $\tilde{u}^{\ell,n}_{H,h} = (\mathtt{id} - \mathcal{C}^\ell_h)u^{\ell,n}_{H,h} \in \tilde{V}^\ell_{H,h}$ such that

$$\tau^{-2}\,\big(\tilde{u}^{\ell,n+1}_{H,h} - 2\tilde{u}^{\ell,n}_{H,h} + \tilde{u}^{\ell,n-1}_{H,h}, \tilde{v}_H\big)_{L^2(D)} + a\big(\tilde{u}^{\ell,n}_{H,h}, \tilde{v}_H\big) = \big(f(n\tau), \tilde{v}_H\big)_{L^2(D)} \tag{5.23}$$

for all $\tilde{v}_H \in \tilde{V}^\ell_{H,h}$ and all $n \in \{1,\dots,N-1\}$.

We emphasize that the computation of the correctors is only done once during the offline stage and can be parallelized. The additional cost to solve the corrector problems is moderate and the main advantage of the method lies in the online stage, where smaller linear systems need to be solved and relatively coarse time steps (subject to the CFL condition) may be used.

Lemma 5.3.2 (Localization error)**.** *Let the assumptions of Lemma 5.2.3 and Assumption 5.2.4 hold. Then the solutions* $\tilde{u}^{n+1}_{H,h} \in \tilde{V}_{H,h}$ *and* $\tilde{u}^{\ell,n+1}_{H,h} \in \tilde{V}^\ell_{H,h}$ *satisfy*

$$\begin{aligned}\|D_\tau \tilde{u}^{n+1}_{H,h} - D_\tau \tilde{u}^{\ell,n+1}_{H,h}\|_{L^2(D)} &+ \|\nabla(\tilde{u}^{n+1}_{H,h} - \tilde{u}^{\ell,n+1}_{H,h})\|_{L^2(D)}\\ &\lesssim_T \big(\ell^{(d-1)/2}\exp(-C_{\mathrm{dec}}\,\ell) + H^{-1}\ell^{(d-1)}\exp(-2C_{\mathrm{dec}}\,\ell)\big)\,C_{\mathrm{data}}.\end{aligned}$$

for all $n \in \{0,\dots,N-1\}$.

Proof. Let $\tilde{e}^{\ell,n} = (\mathtt{id} - \mathcal{C}_h)(u^n_{H,h} - u^{\ell,n}_{H,h})$. Similarly to the findings in the proof of Lemma 5.3.1, the error $\tilde{e}^{\ell,n}$ solves

$$\begin{aligned}\tau^{-2}\big(\tilde{e}^{\ell,n+1} - 2\tilde{e}^{\ell,n} + \tilde{e}^{\ell,n-1}, (\mathtt{id}-\mathcal{C}_h)v_H\big)_{L^2(D)} &+ a\big(\tilde{e}^{\ell,n}, (\mathtt{id}-\mathcal{C}_h)v_H\big)\\ = -\big(f(n\tau), (\mathcal{C}_h - \mathcal{C}^\ell_h)v_H\big)_{L^2(D)}&\\ + \tau^{-2}\left(\tilde{u}^{\ell,n+1}_{H,h} - 2\tilde{u}^{\ell,n}_{H,h} + \tilde{u}^{\ell,n-1}_{H,h}, (\mathcal{C}_h - \mathcal{C}^\ell_h)v_H\right)_{L^2(D)}&\\ + \tau^{-2}\left((\mathcal{C}_h - \mathcal{C}^\ell_h)(u^{\ell,n+1}_{H,h} - 2u^{\ell,n}_{H,h} + u^{\ell,n-1}_{H,h}), (\mathtt{id}-\mathcal{C}_h)v_H\right)_{L^2(D)}&\\ + a\big((\mathcal{C}_h - \mathcal{C}^\ell_h)u^{\ell,n}_{H,h}, (\mathcal{C}_h - \mathcal{C}^\ell_h)v_H\big) =: F^n_h\big((\mathtt{id}-\mathcal{C}_h)v_H\big)&\end{aligned}$$

for all $v_H \in V_H$. As above, we show that $F^n_h|_{\tilde{V}_{H,h}} \in L^2(D)$. From Theorem 2.4.4 with $H^1_0(D)$ replaced by V_h, we get that

$$\|\nabla(\mathcal{C}_h - \mathcal{C}^\ell_h)v_H\|_{L^2(D)} \lesssim \ell^{(d-1)/2}\exp(-C_{\mathrm{dec}}\,\ell)\,\|\nabla v_H\|_{L^2(D)}$$

for any $v_H \in V_H$ and, additionally,

$$\|(\mathcal{C}_h - \mathcal{C}^\ell_h)v_H\|_{L^2(D)} \lesssim \ell^{(d-1)/2}\exp(-C_{\mathrm{dec}}\,\ell)C_{\mathcal{I}_H}H\,\|\nabla v_H\|_{L^2(D)},$$

see also [HP13], or [KY16,KPY18] for an alternative constructive proof. Similar to the estimates in the proof of Lemma 5.3.1, we obtain

$$\begin{aligned}&\sup_{\tilde{v}_H \in \tilde{V}_{H,h}} \frac{|F^n_h(\tilde{v}_H)|}{\|\tilde{v}_H\|_{L^2(D)}}\\ &\quad\lesssim \Big(\|f(n\tau)\|_{L^2(D)} + \tau^{-2}\,\|\tilde{u}^{\ell,n+1}_{H,h} - 2\tilde{u}^{\ell,n}_{H,h} + \tilde{u}^{\ell,n-1}_{H,h}\|_{L^2(D)}\\ &\qquad + H^{-1}\ell^{(d-1)/2}\exp(-C_{\mathrm{dec}}\,\ell)\,\|\nabla\tilde{u}^{\ell,n}_{H,h}\|_{L^2(D)}\Big)\,\ell^{(d-1)/2}\exp(-C_{\mathrm{dec}}\,\ell)\\ &\quad\lesssim \big(\ell^{(d-1)/2}\exp(-C_{\mathrm{dec}}\,\ell) + H^{-1}\ell^{(d-1)}\exp(-2\,C_{\mathrm{dec}}\,\ell)\big)\,C_{\mathrm{data}},\end{aligned}$$

and finally

$$\begin{aligned}\|D_\tau \tilde{u}^{n+1}_{H,h} - D_\tau \tilde{u}^{\ell,n+1}_{H,h}\|_{L^2(D)} &+ \|\nabla(\tilde{u}^{n+1}_{H,h} - \tilde{u}^{\ell,n+1}_{H,h})\|_{L^2(D)}\\ &\lesssim_T \big(\ell^{(d-1)/2}\exp(-C_{\mathrm{dec}}\,\ell) + H^{-1}\ell^{(d-1)}\exp(-2\,C_{\mathrm{dec}}\,\ell)\big)\,C_{\mathrm{data}},\end{aligned}$$

which concludes the proof. □

5.3.3 Error estimates

We can now formulate the following theorem using Lemma 5.3.1, Lemma 5.3.2, and the triangle inequality.

Theorem 5.3.3 (Error of the practical method)**.** *Assume that Assumption 5.2.4 holds and τ fulfills the CFL condition* (5.8)*. Further, let $u \in L^2(0,T;H^1_\Gamma(D))$ be the solution of* (5.2) *and $\tilde{u}^{\ell,n}_{H,h} \in \tilde{V}^{\ell}_{H,h}$ the solution of the practical method* (5.23) *at time step n. Then, for $n \in \{0,\dots,N-1\}$, we have*

$$\begin{aligned}\Big\|D_\tau \tilde{u}^{\ell,n+1}_{H,h} - \frac{u(t_{n+1})-u(t_n)}{\tau}\Big\|_{L^2(D)} &+ \|\nabla\big(\tilde{u}^{\ell,n+1}_{H,h} - u(t_{n+1})\big)\|_{L^2(D)} \\ \lesssim_T \big(H + \tau^2 + d_{\tilde{V}_H}[V_h] &+ H^{-1}(d_{\tilde{V}_H}[V_h])^2 \\ &+ \ell^{(d-1)/2}\exp(-C_{\mathrm{dec}}\,\ell) + H^{-1}\ell^{(d-1)}\exp(-2\,C_{\mathrm{dec}}\,\ell)\big)\,C_{\mathrm{data}}.\end{aligned} \tag{5.24}$$

If $\ell \gtrsim |\log H|$, (5.24) *simplifies to*

$$\begin{aligned}\Big\|D_\tau \tilde{u}^{\ell,n+1}_{H,h} - \frac{u(t_{n+1})-u(t_n)}{\tau}\Big\|_{L^2(D)} &+ \|\nabla\big(\tilde{u}^{\ell,n+1}_{H,h} - u(t_{n+1})\big)\|_{L^2(D)} \\ &\lesssim_T \big(H + \tau^2 + d_{\tilde{V}_H}[V_h] + H^{-1}(d_{\tilde{V}_H}[V_h])^2\big)\,C_{\mathrm{data}}.\end{aligned}$$

Theorem 5.3.3 shows that in order to obtain a reasonable error of order H, the error introduced by the discretization of the corrector problems (5.4) and thus the approximation error $d_{\tilde{V}_H}[V_h]$ need to be of order H as well. The following lemma quantifies the approximation error $d_{\tilde{V}_H}[V_h]$ under additional regularity assumptions on the coefficient A and with full homogeneous Dirichlet boundary, i.e., $\Gamma = \partial D$. Although the result does not hold for more general cases, it gives an indication on how to choose h dependent on H and ϵ.

Lemma 5.3.4. *Let $\Gamma = \partial D$ and suppose that $A \in W^{1,\infty}(D)$ is a scalar coefficient with oscillations on the scale ϵ, i.e., $\|A\|_{W^{1,\infty}(D)} \le C\epsilon^{-1}$. Further, let $\mathcal{I}_h \colon H^1_0(D) \to V_h$ be an operator with the approximation property*

$$\|\nabla(v - \mathcal{I}_h v)\|_{L^2(D)} \le C_{\mathcal{I}_h} h\,\|\nabla^2 v\|_{L^2(D)}$$

for any $v \in H^2(D) \cap H^1_0(D)$. Then it holds that

$$d_{\tilde{V}_H}[V_h] \lesssim h(H^{-1} + \epsilon^{-1}).$$

Proof. For any $\tilde{v}_H \in \tilde{V}_H$, we have

$$\begin{aligned}\inf_{v_h \in V_h} \|\nabla(\tilde{v}_H - v_h)\|_{L^2(D)} &\le \|\nabla(\mathtt{id} - \mathcal{I}_h)\tilde{v}_H\|_{L^2(D)} \le C_{\mathcal{I}_h} h\,\|\nabla^2 \tilde{v}_H\|_{L^2(D)} \\ &\le C_{\mathcal{I}_h} h\,\|\Delta \tilde{v}_H\|_{L^2(D)} \le \alpha^{-1} C_{\mathcal{I}_h} h\,\|A\Delta\tilde{v}_H\|_{L^2(D)} \\ &\le \alpha^{-1} C_{\mathcal{I}_h} h\,\big(\|\operatorname{div} A\nabla\tilde{v}_H\|_{L^2(D)} \\ &\qquad + \|A\|_{W^{1,\infty}(D)}\,\|\nabla\tilde{v}_H\|_{L^2(D)}\big) \\ &\le \alpha^{-1} C_{\mathcal{I}_h} h\,\big(\beta C_{\mathrm{inv}} C_{\mathcal{I}_H} H^{-1} + C\epsilon^{-1}\big)\,\|\nabla\tilde{v}_H\|_{L^2(D)},\end{aligned}$$

where we employ the product rule, $\|A\|_{W^{1,\infty}(D)} \leq C\epsilon^{-1}$, and

$$\|\operatorname{div} A\nabla\tilde{v}_H\|_{L^2(D)} \leq \beta C_{\mathrm{inv}} C_{\mathcal{I}_H} H^{-1}\, \|\nabla\tilde{v}_H\|_{L^2(D)}. \tag{5.25}$$

To show the last estimate, let $v \in C_c^\infty(D)$ and observe that

$$\begin{aligned}\frac{|(\operatorname{div} A\nabla\tilde{v}_H, v)_{L^2(D)}|}{\|v\|_{L^2(D)}} &= \frac{|a(\tilde{v}_H, v)|}{\|v\|_{L^2(D)}} = \frac{|a(\tilde{v}_H, \mathcal{I}_H v)|}{\|v\|_{L^2(D)}} \leq \frac{\beta\|\nabla\tilde{v}_H\|_{L^2(D)}\,\|\nabla\mathcal{I}_H v\|_{L^2(D)}}{\|v\|_{L^2(D)}}\\ &\leq \beta C_{\mathrm{inv}} C_{\mathcal{I}_H} H^{-1}\|\nabla\tilde{v}_H\|_{L^2(D)},\end{aligned}$$

where we employ the estimates (2.11) and (2.35). The inequality (5.25) then follows by the density of $C_c^\infty(D)$ in $L^2(D)$. Therefore, $d_{\tilde{V}_H}[V_h]$ can be bounded by

$$d_{\tilde{V}_H}[V_h] \lesssim h(H^{-1} + \epsilon^{-1}),$$

which is the assertion. □

Using Theorem 5.3.3 and Lemma 5.3.4, we obtain the following result.

Corollary 5.3.5. *Let Assumption 5.2.4 hold and suppose that* $A \in W^{1,\infty}(D)$, $\Gamma = \partial D$, $\tau \lesssim H$ *subject to the CFL condition* (5.8), $\ell \gtrsim |\log H|$ *and* $h \lesssim H\epsilon$. *Then, for* $n \in \{0, \ldots, N-1\}$, *we have that*

$$\Big\|D_\tau\tilde{u}_{H,h}^{\ell,n+1} - \frac{u(t_{n+1}) - u(t_n)}{\tau}\Big\|_{L^2(D)} + \|\nabla\big(\tilde{u}_{H,h}^{\ell,n+1} - u(t_{n+1})\big)\|_{L^2(D)} \lesssim_T \big(H + \tau^2\big)\, C_{\mathrm{data}}$$

with the solutions $u \in L^2(0,T; H_0^1(D))$ *of* (5.2) *and* $\tilde{u}_{H,h}^{\ell,n}$ *of* (5.23).

From Corollary 5.3.5, we directly get that, provided the additional regularity assumptions hold, the error of the method in the discrete energy norm

$$\|v\|_{N,a} := \left(\sum_{j=1}^{N} \tau\, \|A^{1/2}\nabla v(j\tau)\|_{L^2(D)}^2\right)^{1/2} \tag{5.26}$$

scales like $H + \tau^2$. While orders of convergence in space and time appear imbalanced when the error is measured in the energy norm, quadratic convergence is empirically observed for the discrete $L^2(L^2)$-norm defined by

$$\|v\|_{N,0} := \left(\sum_{j=1}^{N} \tau\, \|v(j\tau)\|_{L^2(D)}^2\right)^{1/2}, \tag{5.27}$$

see Section 5.4. In this sense, the error estimates of the explicit method are competitive with the fully implicit Crank-Nicolson approach used in [AH17] provided that the fine-scale discretization errors in [AH17] can be bounded by $\mathcal{O}((h/\epsilon)^2)$.

Remark 5.3.6. As presented in Chapter 3, the above LOD construction is not limited to approximation spaces based on Q_1 finite elements. In principle, this means that there is no restriction to combine a higher-order variant of the method in space with any higher-order time stepping approach. As the error analysis varies depending on the spatial and temporal discretization, these extensions each need to be studied separately.

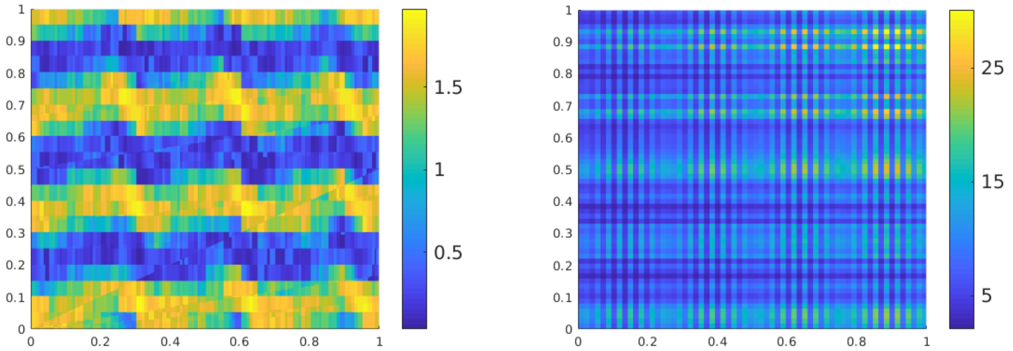


Figure 5.1: Coefficient A in Example 1 (left) and Example 2 (right).

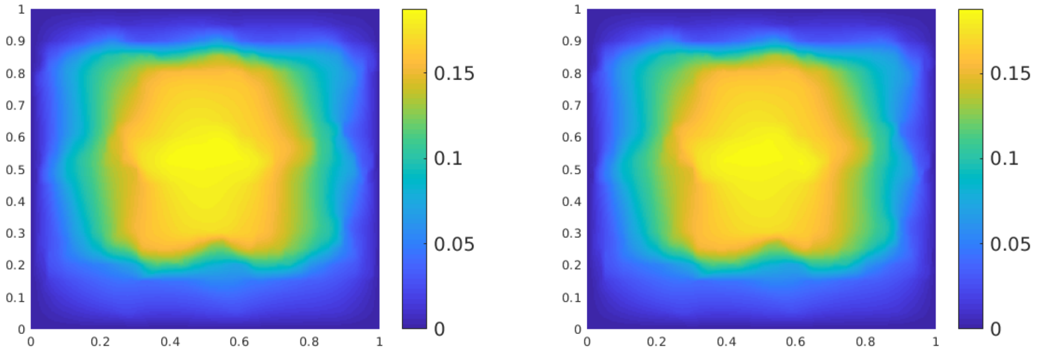


Figure 5.2: Discrete solutions of the wave equation at final time $T = 1$ for Example 1: fine-scale reference solution (left) and LOD approximation on the scale $H = 2^{-4}$ with $\ell = 2$ (right).

5.4 Numerical experiments

In this section, we present numerical experiments to illustrate the theoretical results from the previous sections. The error of the method is measured in the discrete energy norm and the discrete $L^2(L^2)$-norm as defined in (5.26) and (5.27), respectively. We set $D = (0,1)^2$ and $T = 1$ and compute a reference solution using standard finite elements paired with a leapfrog scheme in time on a mesh $\mathcal{T}_h$ with mesh size $h = 2^{-8}$, which is also the mesh parameter for the computations of the corrector problems. The fine time step size is chosen small enough subject to the standard CFL condition, i.e., $\tau_{\text{fine}} \leq C_{\text{CFL}}\, h$, where $C_{\text{CFL}} = \sqrt{2}\,\beta^{-1/2}\, C_{\text{inv}}^{-1}$. This condition can be shown similarly to (5.8) but is slightly relaxed since $C_{\mathcal{I}_H} \geq 1$ in general. Practical experiments show that $C_{\text{CFL}} = \sqrt{2}\,\beta^{-1/2}\, 0.14$ is a sufficient and rather sharp choice for the stability of both the fine FE solution and the LOD approximation. In the following

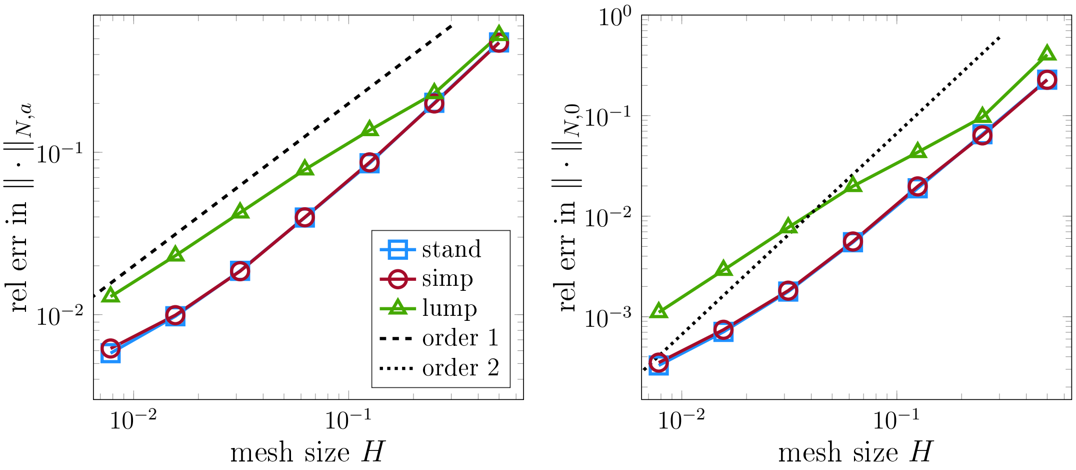


Figure 5.3: Relative errors of LOD approximations for Example 1 in the discrete energy norm (left) and the discrete $L^2(L^2)$-norm (right) with respect to the mesh size H for $\ell = 2$.

examples, we choose $\tau \leq C_{\mathrm{CFL}}\, H$ such that $N = T/\tau \in \mathbb{N}$. Note that given $\tilde{u}^{\ell,0}_{H,h} \in \tilde{V}^{\ell}_{H,h}$, $\tilde{u}^{\ell,1}_{H,h} \in \tilde{V}^{\ell}_{H,h}$ is computed using the second-order Taylor expansion

$$(\tilde{u}^{\ell,1}_{H,h}, \tilde{v}_H)_{L^2(D)} = (\tilde{u}^{\ell,0}_{H,h} + \tau\, v^0 + \tfrac{1}{2}\tau^2\, f(0), \tilde{v}_H)_{L^2(D)} - \tfrac{1}{2}\tau^2\, a(\tilde{u}^{\ell,0}_{H,h}, \tilde{v}_H)$$

for any $\tilde{v}_H \in \tilde{V}^{\ell}_{H,h}$, where v^0 and $f(0)$ may be replaced by suitable approximations. This choice is crucial in order to get optimal convergence rates.

5.4.1 Example 1

For the first example, we take the setting from [AH17, Sec. 6.2], i.e., $f = 1$, $u^0 = v^0 = 0$, $\Gamma = \partial D$, and a scalar coefficient A as depicted in Figure 5.1 (left) with $\alpha = 0.04$, $\beta = 1.96$. A detailed formula for the coefficient can be found in [AH17, Sec. 6.2]. Besides, we set $\ell = 2$ for all values of H. The remaining discretization parameters are defined above. The relative errors of the practical method in the energy norm are shown in Figure 5.3 (left). The relative errors in the $L^2(L^2)$-norm are depicted in Figure 5.3 (right). The blue curves (□) show the errors of the standard method defined in (5.23) and the red curves (○) display the errors of the simplified method based on (5.19) which uses the classical finite element mass matrix. Both curves show the expected linear convergence in the energy norm and almost second-order convergence in $L^2(L^2)$ with a commencing stagnation for smaller values of H, which can be avoided if ℓ is increased in this regime; cf. Tables 5.1 and 5.2. The fact that the curves are very close justifies the theoretical observation that the mass matrices may be exchanged. In addition, the green curves (△) display the errors if a *lumped mass matrix* is used. That is, the multiscale mass matrix is replaced by a diagonal matrix which is obtained

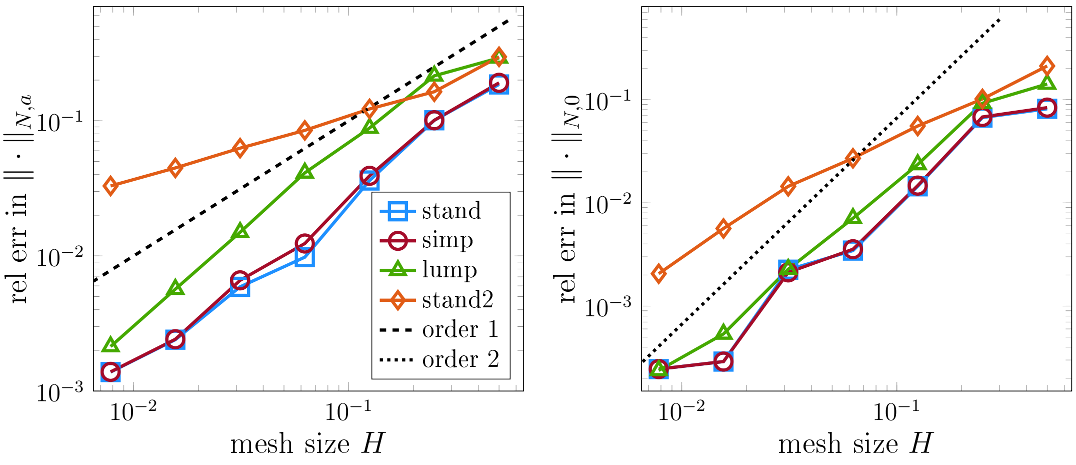


Figure 5.4: Relative errors of LOD approximations for Example 2 in the discrete energy norm (left) and the discrete $L^2(L^2)$-norm (right) with respect to the mesh size H for $\ell = 2$.

by summing up the rows of the multiscale mass matrix. The plots indicate a reduced convergence rate for the lumped version of the method. The reduced rate can also be observed in Tables 5.1 and 5.2.

Finally, the reference solution and the solution obtained with the standard method ($H = 2^{-4}$, $\ell = 2$) at final time $T = 1$ are given in Figure 5.2.

5.4.2 Example 2

In the second example, we choose $f(x,t) = \sin(4\pi\, x_1)(1-t)$ and $v^0 = 0$. Further, we set $\Gamma = \{x \in \partial D \colon x_1 = 0\}$ and let $u_0 \in H^1_\Gamma(D)$ be the solution of

$$a(u^0, v) = (5\sin(\pi\, x_1)\sin(\pi\, x_2), v)_{L^2(D)}$$

for all $v \in H^1_\Gamma(D)$. The scalar coefficient A is shown in Figure 5.1 (right), where $\alpha = 2.1$, $\beta = 30.1$, and $\epsilon = 2^{-6}$. The other discretization parameters are chosen as defined above and $\ell = 2$. The blue curves (□) in Figure 5.4 again show the relative errors of the standard method (5.23) and the red curves (○) show the relative errors of the simplified method. Both methods and even the lumped version (△) show a convergence rate in the discrete energy norm which is slightly better than one. Then again, a near second-order rate in $L^2(L^2)$ can be observed for the standard method, the simplified method, and also the lumped version up to a commencing stagnation due to localization; cf. also Tables 5.3 and 5.4.

Note that we also provide a fourth error curve (◇), which shows the relative errors of the standard method for the exact same setting but with $v^0 = 0$ replaced by $v^0 = 0.2 \cdot \mathbb{1}_{\{x_1 > 0.5\}}$. One can observe a suboptimal convergence behavior, which is possibly related to the fact that $v^0 \notin H^1_\Gamma(D)$ and, therefore, the condition (A2) in Assumption 5.2.4 is not fulfilled.

Table 5.1: Relative errors in the discrete energy norm and EOCs for Example 1 obtained with the standard approach (stand), the simplified one (simp), and the lumped version (lump).

ℓ	H	stand	simp	lump	$\mathrm{EOC}_{\mathrm{stand}}$	$\mathrm{EOC}_{\mathrm{simp}}$	$\mathrm{EOC}_{\mathrm{lump}}$
2	2^{-1}	0.47508	0.47302	0.52580	–	–	–
2	2^{-2}	0.20178	0.19994	0.23064	1.24	1.24	1.18
2	2^{-3}	0.08560	0.08668	0.13656	1.24	1.21	0.76
2	2^{-4}	0.03965	0.03978	0.07823	1.11	1.12	0.80
2	2^{-5}	0.01861	0.01852	0.04247	1.09	1.10	0.88
2	2^{-6}	0.00978	0.00993	0.02307	0.93	0.90	0.88
2	2^{-7}	0.00579	0.00619	0.01292	0.76	0.68	0.84
4	2^{-1}	0.47508	0.47302	0.52580	–	–	–
4	2^{-2}	0.20132	0.19948	0.23026	1.24	1.25	1.19
4	2^{-3}	0.08509	0.08617	0.13595	1.24	1.21	0.76
4	2^{-4}	0.03930	0.03943	0.07766	1.12	1.13	0.81
4	2^{-5}	0.01789	0.01779	0.04211	1.14	1.15	0.88
4	2^{-6}	0.00870	0.00887	0.02272	1.04	1.00	0.89
4	2^{-7}	0.00391	0.00447	0.01224	1.15	0.99	0.89

Table 5.2: Relative errors in the discrete $L^2(L^2)$-norm and EOCs for Example 1 obtained with the standard approach (stand), the simplified one (simp), and the lumped version (lump).

ℓ	H	stand	simp	lump	$\mathrm{EOC}_{\mathrm{stand}}$	$\mathrm{EOC}_{\mathrm{simp}}$	$\mathrm{EOC}_{\mathrm{lump}}$
2	2^{-1}	0.22771	0.22717	0.40377	–	–	–
2	2^{-2}	0.06536	0.06360	0.0967	1.80	1.83	2.06
2	2^{-3}	0.01895	0.01979	0.0432	1.79	1.68	1.16
2	2^{-4}	0.00551	0.00560	0.01986	1.82	1.82	1.12
2	2^{-5}	0.00178	0.00182	0.00772	1.62	1.62	1.36
2	2^{-6}	0.00071	0.00074	0.00291	1.29	1.29	1.41
2	2^{-7}	0.00035	0.00035	0.00111	1.08	1.08	1.40
4	2^{-1}	0.22771	0.22717	0.40377	–	–	–
4	2^{-2}	0.06571	0.06397	0.09673	1.79	1.83	2.06
4	2^{-3}	0.01909	0.01991	0.04310	1.78	1.68	1.16
4	2^{-4}	0.00555	0.00564	0.01978	1.78	1.82	1.12
4	2^{-5}	0.00170	0.00174	0.00774	1.70	1.70	1.35
4	2^{-6}	0.00055	0.00060	0.00292	1.61	1.55	1.41
4	2^{-7}	0.00015	0.00020	0.00110	1.84	1.61	1.41

Table 5.3: Relative errors in the discrete energy norm and EOCs for Example 2 obtained with the standard approach (stand), the simplified one (simp), and the lumped version (lump).

ℓ	H	stand	simp	lump	EOC$_{\text{stand}}$	EOC$_{\text{simp}}$	EOC$_{\text{lump}}$
2	2^{-1}	0.18571	0.19098	0.29191	–	–	–
2	2^{-2}	0.10083	0.10084	0.21508	0.88	0.92	0.44
2	2^{-3}	0.03614	0.03916	0.08837	1.48	1.36	1.28
2	2^{-4}	0.00979	0.01235	0.04117	1.88	1.66	1.10
2	2^{-5}	0.00591	0.00660	0.01496	0.73	0.91	1.46
2	2^{-6}	0.00240	0.00242	0.00568	1.30	1.45	1.40
2	2^{-7}	0.00138	0.00139	0.00214	0.80	0.80	1.41

Table 5.4: Relative errors in the discrete $L^2(L^2)$-norm and EOCs for Example 2 obtained with the standard approach (stand), the simplified one (simp), and the lumped version (lump).

ℓ	H	stand	simp	lump	EOC$_{\text{stand}}$	EOC$_{\text{simp}}$	EOC$_{\text{lump}}$
2	2^{-1}	0.08162	0.08386	0.14215	–	–	–
2	2^{-2}	0.06719	0.06809	0.09247	0.28	0.30	0.62
2	2^{-3}	0.01448	0.01471	0.02357	2.21	2.21	1.97
2	2^{-4}	0.00345	0.00356	0.00709	2.07	2.04	1.73
2	2^{-5}	0.00225	0.00212	0.00228	0.62	0.75	1.64
2	2^{-6}	0.00029	0.00029	0.00054	2.96	2.87	2.08
2	2^{-7}	0.00024	0.00025	0.00024	0.24	0.24	1.17

6 Multiscale Poroelasticity in Heterogeneous Media

In this chapter, we deal with another time-dependent PDE, known as linear poroelasticity. This problem describes the deformation of porous media saturated by an incompressible viscous fluid and is of great importance for many physical applications such as reservoir engineering in the field of geomechanics [Zob10] or the modeling of the human anatomy for medical applications [MC16, CM14]. Biot [Bio41] proposed this poroelastic model that couples a *Darcy flow* with the *linear elastic behavior* of the porous medium. The idea is to average the pressure and displacement across (infinitesimal) cubic elements such that pressure and displacement can be treated as variables on the entire domain of interest. Furthermore, the model is assumed to be quasi-static, i.e., an internal equilibrium is preserved at any time. In the poroelastic setting, this means that volumetric changes occur slowly enough for the pressure to remain basically constant throughout an infinitesimal element.

If the poroelastic coefficients at hand are homogeneous, the problem can be simulated using standard numerical methods such as the FE method, see for instance [EM09]. However, if the medium is strongly heterogeneous, the material parameters may oscillate on a fine scale. As already mentioned in the previous chapters, the classical FE method only yields acceptable results if the fine scale is resolved by the spatial discretization, which is unfeasible in practical applications. To overcome this difficulty, homogenization techniques may be applied, such as those presented in Section 1.2. Concerning these methods in the poroelastic context, the GMsFEM is, for instance, used in [BV16a, BV16b], the CEM-GMsFEM in [FAC$^+$19], or the LOD technique in [MP17] for the similar problem of linear thermoelasticity. Related work in connection with the LOD can also be found in [BP16], where porous microstructures are considered, and in [HP16] in the context of linear elasticity. All these methods aim at performing computations on a coarse scale of interest although the coefficients vary on a much finer scale. We emphasize that with respect to the physical model there exists even a third scale, namely the infinitesimal scale on which the averaging of pressure and displacement is done. This scale, however, is not treated since it is small enough and already included in the given PDE model.

In the present setting, we introduce a method that adopts ideas presented in [MP17], where a classical LOD approach is used, which is explained below. We modify this method based on structural properties which are obtained by an alternative perspective on the discretized problem. This allows us to obtain an overall simpler approach. In particular, we are able to exploit the saddle point structure of the problem in order to obtain fully symmetric and decoupled corrector problems. That is, we do not require additional corrections as in [MP17] and our correction operators are independent of the coupling term, although the corresponding coefficient may vary rapidly as well. This adapted method was first presented in [ACM$^+$20].

Before getting into the details, we introduce the PDE representation of the model and its variational formulation in the following section.

6.1 Linear poroelasticity

The problem of linear poroelasticity that we use here is posed in a bounded, convex, and polytopal Lipschitz domain $D \subseteq \mathbb{R}^d$, $d \in \{2,3\}$, and was, e.g., discussed in [Sho00]. For the sake of simplicity, we restrict ourselves to homogeneous Dirichlet boundary conditions but emphasize that an extension to Neumann boundary conditions is straightforward; see also the numerical examples in Section 6.3. This means that we seek the pressure $p\colon [0,T] \times D \to \mathbb{R}$ and the displacement field $u\colon [0,T] \times D \to \mathbb{R}^d$ up to a given final time $T > 0$ such that

$$\begin{aligned} -\nabla \cdot \bigl(\sigma(u)\bigr) + \nabla(\alpha p) &= 0 \quad \text{in } (0,T] \times D, \\ \partial_t\Bigl(\alpha \nabla \cdot u + \frac{1}{M} p\Bigr) - \nabla \cdot \Bigl(\frac{\kappa}{\nu} \nabla p\Bigr) &= f \quad \text{in } (0,T] \times D, \end{aligned} \tag{6.1}$$

with the boundary and initial conditions

$$\begin{aligned} u &= 0 \quad \text{on } (0,T] \times \partial D, \\ p &= 0 \quad \text{on } (0,T] \times \partial D, \\ p(\cdot,0) &= p^0 \quad \text{in } D. \end{aligned} \tag{6.2}$$

In the given model, the primary sources of the heterogeneities in the physical properties arise from the *stress tensor* σ, the *permeability* κ, and the *Biot-Willis fluid-solid coupling coefficient* α. Further, we denote by M the *Biot modulus* and by ν the *fluid viscosity* which are assumed to be constant. The source term f represents an injection or production process. In the case of a linear elastic stress-strain constitutive relation, we have that the *stress tensor* and *symmetric strain gradient* may be expressed as

$$\sigma(u) = 2\mu\, \varepsilon(u) + \lambda\, (\nabla \cdot u)\, I, \qquad \varepsilon(u) = \frac{1}{2}\bigl(\nabla u + (\nabla u)^T\bigr),$$

where μ and λ are the *Lamé coefficients* and I is the identity tensor. In the case of heterogeneous media, the coefficients μ, λ, κ, and α may be highly oscillatory.

We now turn our attention to the variational formulation of the poroelasticity system (6.1). To this end, we define the spaces for the displacement and the pressure by

$$\mathcal{V} := \bigl[H^1_0(D)\bigr]^d, \qquad \mathcal{Q} := H^1_0(D)$$

and write

$$\mathcal{H}_\mathcal{V} := \big[L^2(D)\big]^d, \qquad \mathcal{H}_\mathcal{Q} := L^2(D)$$

for the corresponding L^2-spaces. To obtain a variational form, we multiply the equations (6.1) with test functions from $\mathcal{V}$ and $\mathcal{Q}$, respectively, and use integration by parts as well as the boundary conditions (6.2). This leads to the following problem: find $u(\cdot,t) \in \mathcal{V}$ and $p(\cdot,t) \in \mathcal{Q}$ such that

$$\begin{aligned} a(u,v) - d(v,p) &= 0, \\ d(\partial_t u, q) + c(\partial_t p, q) + b(p,q) &= (f,q)_{\mathcal{H}_\mathcal{Q}}, \end{aligned} \tag{6.3}$$

for all $v \in \mathcal{V}$, $q \in \mathcal{Q}$ and

$$p(\cdot,0) = p^0.$$

The involved bilinear forms $a\colon \mathcal{V}\times\mathcal{V} \to \mathbb{R}$, $b,c\colon \mathcal{Q}\times\mathcal{Q} \to \mathbb{R}$, and $d\colon \mathcal{V}\times\mathcal{Q}\to\mathbb{R}$ are defined as

$$\begin{aligned} a(u,v) &:= \int_D \sigma(u) : \varepsilon(v)\,\mathrm{d}x, \qquad & b(p,q) &:= \int_D \frac{\kappa}{\nu}\,\nabla p\cdot\nabla q\,\mathrm{d}x, \\ c(p,q) &:= \int_D \frac{1}{M}\,p\,q\,\mathrm{d}x, \qquad & d(u,q) &:= \int_D \alpha\,(\nabla\cdot u)\,q\,\mathrm{d}x. \end{aligned}$$

We emphasize that the bilinear forms a, b, and c are symmetric. Note that the first equation in (6.3) can be used to define a *consistent initial condition* $u^0 := u(\cdot,0)$. Using *Korn's inequality* [Cia88, Thm. 6.3-4], we have the bounds

$$c_\sigma \|v\|_\mathcal{V}^2 \le a(v,v) \le C_\sigma \|v\|_\mathcal{V}^2 \tag{6.4}$$

for all $v \in \mathcal{V}$, where c_σ and C_σ are positive constants. Similarly, there are positive constants c_κ and C_κ such that

$$c_\kappa \|q\|_\mathcal{Q}^2 \le b(q,q) \le C_\kappa \|q\|_\mathcal{Q}^2 \tag{6.5}$$

for all $q \in \mathcal{Q}$. We write $\|\cdot\|_a$ for the energy norm induced by the bilinear form a and similarly $\|\cdot\|_b$ for the norm induced by b. Note that also the bilinear form c defines a norm $\|\cdot\|_c$, which is a weighted L^2-norm.

We conclude this section with the remark that there exist unique solutions u and p to (6.3), which was discussed and proven in [Sho00].

6.2 Numerical approximation

In this section, we present different schemes for the discretization of system (6.3): the classical FE approach analyzed in [EM09], the standard LOD approach used in [MP17], and the adapted LOD method introduced in [ACM$^+$20]. Since the classical FE ansatz is only meaningful if oscillations are resolved by the underlying mesh, this approach solely serves as a reference.

6.2.1 Fine-scale discretization with finite elements

As in Section 2.2, we define appropriate FE spaces for the poroelasticity system (6.3) based on a family $\{\mathcal{T}_h\}_{h>0}$ of quasi-uniform decompositions of D. That is, for a particular mesh parameter h, let $V_h \subseteq \mathcal{V}$ and $Q_h \subseteq \mathcal{Q}$ be the corresponding conforming Q_1 finite element spaces. For the temporal discretization, we consider a uniform time step $\tau > 0$ such that $t_n = \tau n$ for $n \in \{0, \dots, N\}$ and $T = \tau N$ as in Chapter 5.

Using the notation introduced above, we discretize system (6.3) with a *backward Euler scheme* in time and finite elements in space, i.e., for $n \in \{1, \dots, N\}$, we aim to find $u_h^n \in V_h$ and $p_h^n \in Q_h$ such that

$$\begin{aligned} a(u_h^n, v_h) - d(v_h, p_h^n) &= 0, \\ d(D_\tau u_h^n, q_h) + c(D_\tau p_h^n, q_h) + b(p_h^n, q_h) &= (f^n, q_h)_{\mathcal{H}_\mathcal{Q}} \end{aligned} \tag{6.6}$$

for all $v_h \in V_h$ and $q_h \in Q_h$. As before, D_τ denotes the discrete time derivative, i.e., $D_\tau u_h^n = (u_h^n - u_h^{n-1})/\tau$, and we set $f^n := f(t_n)$. The initial condition $p_h^0 \in Q_h$ is chosen to be a suitable approximation of p^0 and u_h^0 is uniquely determined by the variational problem

$$a(u_h^0, v_h) = d(v_h, p_h^0)$$

for all $v_h \in V_h$.

Lemma 6.2.1 (Well-posedness)**.** *Given initial data $u_h^0 \in V_h$ and $p_h^0 \in Q_h$, the system* (6.6) *is well-posed, i.e., there exists a unique solution, which is bounded in terms of the initial conditions and the source term f.*

Proof. The proof is based on [EM09, Lem. 2.1]. For the bilinear form a, it holds

$$\begin{aligned} 2\,a(u_h^n, u_h^n - u_h^{n-1}) &= a(u_h^n, u_h^n) + a(u_h^n - u_h^{n-1}, u_h^n - u_h^{n-1}) - a(u_h^{n-1}, u_h^{n-1}) \\ &\geq \|u_h^n\|_a^2 - \|u_h^{n-1}\|_a^2. \end{aligned} \tag{6.7}$$

A similar result can be shown for the bilinear form c. With $v_h = u_h^n - u_h^{n-1} \in V_h$ and $q_h = \tau p_h^n \in Q_h$ as test functions in (6.6), we obtain

$$a(u_h^n, u_h^n - u_h^{n-1}) + c(p_h^n - p_h^{n-1}, p_h^n) + \tau\, b(p_h^n, p_h^n) = \tau\,(f^n, p_h^n)_{\mathcal{H}_\mathcal{Q}} \tag{6.8}$$

when adding both equations. Inequality (6.7), an application of Young's inequality, and (6.5) then imply

$$\|u_h^n\|_a^2 + \|p_h^n\|_c^2 + \tau \|p_h^n\|_b^2 \leq \frac{\tau}{c_\kappa} \|f^n\|_{\mathcal{H}_\mathcal{Q}}^2 + \|u_h^{n-1}\|_a^2 + \|p_h^{n-1}\|_c^2.$$

A summation over all n finally leads to the stability estimate

$$\|u_h^n\|_a^2 + \|p_h^n\|_c^2 + \tau \sum_{j=1}^n \|p_h^j\|_b^2 \leq \frac{\tau}{c_\kappa} \sum_{j=1}^n \|f^j\|_{\mathcal{H}_\mathcal{Q}}^2 + \|u_h^0\|_a^2 + \|p_h^0\|_c^2.$$

This implies the uniqueness of the solutions u_h^n and p_h^n. Existence follows from the fact that system (6.6) is equivalent to a square system of linear equations and, hence, uniqueness implies existence. □

For the presented fine-scale discretization in (6.6), one can show the following stability result, which is important for the convergence proof of the alternative LOD method in Section 6.2.3.

Theorem 6.2.2 (cf. [MP17, Thm. 3.3]). *Suppose that the right-hand side fulfills $f \in L^\infty(0,T;L^2(D)) \cap H^1(0,T;H^{-1}(D))$. Then, the fully discrete solution (u_h^n, p_h^n) of (6.6) satisfies for all $n \in \{1,\dots,N\}$ the stability bound*

$$\Big(\tau \sum_{j=1}^n \|D_\tau u_h^j\|_{\mathcal{V}}^2\Big)^{1/2} + \Big(\tau \sum_{j=1}^n \|D_\tau p_h^j\|_{\mathcal{H}_{\mathcal{Q}}}^2\Big)^{1/2} + \|p_h^n\|_{\mathcal{Q}} \lesssim \|p_h^0\|_{\mathcal{Q}} + \|f\|_{L^2(0,t_n;L^2(D))}.$$

Further, in the case $p_h^0 = 0$, we have that

$$\|D_\tau u_h^n\|_{\mathcal{V}} + \|D_\tau p_h^n\|_{\mathcal{H}_{\mathcal{Q}}} + \Big(\tau \sum_{j=1}^n \|D_\tau p_h^j\|_{\mathcal{Q}}^2\Big)^{1/2} \lesssim \|f\|_{L^\infty(0,t_n;L^2(D))} + \|\partial_t f\|_{L^2(0,t_n;H^{-1}(D))}$$

and for $f = 0$, it holds that

$$\|D_\tau u_h^n\|_{\mathcal{V}} + \|D_\tau p_h^n\|_{\mathcal{H}_{\mathcal{Q}}} + t_n^{1/2}\|D_\tau p_h^n\|_{\mathcal{Q}} \lesssim t_n^{-1/2}\|p_h^0\|_{\mathcal{Q}}.$$

The following theorem states the expected order of convergence, which is $\mathcal{O}(h+\tau)$. However, the involved constant for the spatial discretization scales with the maximal $W^{1,\infty}$-norm of the coefficients, which makes this approach unfeasible in oscillatory media with period ϵ.

Theorem 6.2.3 (cf. [EM09, Thm. 3.1]). *Assume that the coefficients satisfy $\mu, \lambda, \kappa, \alpha \in W^{1,\infty}(D)$. Further, let the exact solution (u,p) of (6.1) be sufficiently smooth and (u_h^n, p_h^n) the fully discrete solution obtained by (6.6) for $n \in \{1,\dots,N\}$. Then, the error is bounded by*

$$\|u(t_n) - u_h^n\|_{\mathcal{V}} + \|p(t_n) - p_h^n\|_{\mathcal{H}_{\mathcal{Q}}} + \Big(\tau \sum_{j=1}^n \|p(t_j) - p_h^j\|_{\mathcal{Q}}^2\Big)^{1/2} \le C_\epsilon h + C\tau,$$

where the constants comprise the norms of the right-hand side f and the solutions u and p. Further, C_ϵ crucially depends on the coefficients, i.e.,

$$C_\epsilon \sim \max\{\|\mu\|_{W^{1,\infty}(D)}, \|\lambda\|_{W^{1,\infty}(D)}, \|\kappa\|_{W^{1,\infty}(D)}, \|\alpha\|_{W^{1,\infty}(D)}\}.$$

6.2.2 A classical multiscale method

Within this subsection, we review the classical LOD approach for the poroelastic problem based on a correction which is defined using the stationary version of (6.3). This method was used in [MP17] in the context of thermoelasticity but translates directly to the present setting. We note that this procedure of using the stationary PDE to define a multiscale space is here referred to as the *classical* (or *standard*) *approach* to time-dependent multiscale problems, which is generally used; see also Chapter 5 (based on [MP19]) and, e.g., [AH17, PS17, MP18] in the context of the LOD.

As before, we use the method presented in Chapter 2. However, since the fully discrete method as described in Section 2.4 is actually based on a fine FE space, we here directly define the corresponding operators based on the FE spaces V_h and Q_h from the previous subsection. Assume now that we have a coarse mesh $\mathcal{T}_H$ with mesh parameter $H > h$ that does not resolve the microscopic scale ϵ and let $V_H \subseteq V_h$ and $Q_H \subseteq Q_h$ be the corresponding conforming Q_1 spaces. Further, we define the projective quasi-interpolation operators

$$\mathcal{I}_H^u \colon \mathcal{H}_\mathcal{V} \to V_H \quad \text{and} \quad \mathcal{I}_H^p \colon \mathcal{H}_\mathcal{Q} \to Q_H,$$

which fulfill the properties (2.11) and (2.14) as in Section 2.2.2. With these operators, we define the fine-scale spaces

$$W_h^u := \ker \mathcal{I}_H^u|_{V_h} \subseteq V_h \quad \text{and} \quad W_h^p := \ker \mathcal{I}_H^p|_{Q_h} \subseteq Q_h,$$

which leads to the coupled correction $\mathcal{C}_h \colon V_h \times Q_h \to W_h^u \times W_h^p$ defined for $v_h \in V_h$ and $q_h \in Q_h$ by

$$\mathfrak{a}(\mathcal{C}_h[v_h, q_h], [w_h, r_h]) = \mathfrak{a}([v_h, q_h], [w_h, r_h]) \tag{6.9}$$

for all $w_h \in W_h^u$ and $r_h \in W_h^p$. Here $\mathfrak{a} \colon (\mathcal{V} \times \mathcal{Q}) \times (\mathcal{V} \times \mathcal{Q}) \to \mathbb{R}$ is the bilinear form corresponding to the stationary poroelastic system, i.e.,

$$\mathfrak{a}([v, q], [w, r]) := a(v, w) - d(w, q) + b(q, r).$$

One can show that (6.9) has a unique solution and, therefore, the conditions in Chapter 2 hold. This follows from the coercivity of a and b as well as the fact that we may solve the part involving the bilinear form b first and then use the result for the rest of the equation. A direct consequence of this is that the second component of $\mathcal{C}_h[v_h, q_h]$ only depends on q_h while the first one is determined by v_h and q_h.

Then again, the operator $\mathcal{C}_h^* \colon V_h \times Q_h \to W_h^u \times W_h^p$ for the correction of the test functions is given by

$$\mathcal{C}_h^*[v_h, q_h] = [\mathcal{C}_h^u v_h, \mathcal{C}_h^p q_h],$$

where $\mathcal{C}_h^u = \mathcal{C}_h^{*,u} \colon V_h \to W_h^u$ and $\mathcal{C}_h^p = \mathcal{C}_h^{*,p} \colon Q_h \to W_h^p$ are defined by

$$a(\mathcal{C}_h^u v_h, w_h) = a(v_h, w_h), \qquad b(\mathcal{C}_h^p q_h, r_h) = b(q_h, r_h) \tag{6.10}$$

for all $w_h \in W_h^u$ and $r_h \in W_h^p$. Thus, the operator $\mathcal{C}_h^*$ decouples. In [MP17], also $\mathcal{C}_h[v_h, q_h]$ is computed using the two correction operators $\mathcal{C}_h^u$ and $\mathcal{C}_h^p$ defined in (6.10). This, however, requires an auxiliary correction $\mathcal{C}_h^{\mathrm{aux}}\colon Q_h \to W_h^u$ given by

$$a(\mathcal{C}_h^{\mathrm{aux}} q_h, w_h) = -d(w_h, (\mathtt{id} - \mathcal{C}_h^p) q_h)$$

for all $w_h \in W_h^u$. With the additional correction, it holds that

$$\begin{aligned}
\mathfrak{a}([\mathcal{C}_h^u v_h + \mathcal{C}_h^{\mathrm{aux}} q_h, \mathcal{C}_h^p q_h], [w_h, r_h]) \\
&= a(\mathcal{C}_h^u v_h + \mathcal{C}_h^{\mathrm{aux}} q_h, w_h) - d(w_h, \mathcal{C}_h^p q_h) + b(\mathcal{C}_h^p q_h, r_h) \\
&= a(v_h, w_h) - d(w_h, q_h) + b(q_h, r_h) \\
&= \mathfrak{a}([v_h, q_h], [w_h, r_h])
\end{aligned}$$

for any $v_h \in V_h$ and $q_h \in Q_h$ and all $w_h \in W_h^u$ and $r_h \in W_h^p$. Therefore, the correction operator $\mathcal{C}_h$ satisfies

$$\mathcal{C}_h[v_h, q_h] = [\mathcal{C}_h^u v_h + \mathcal{C}_h^{\mathrm{aux}} q_h, \mathcal{C}_h^p q_h].$$

Next, we define the operators

$$\mathcal{R}_h^u\colon V_H \to V_h \quad \text{and} \quad \mathcal{R}_h^p\colon Q_H \to Q_h$$

defined by

$$\mathcal{R}_h^u v_H := (\mathtt{id} - \mathcal{C}_h^u) v_H \quad \text{and} \quad \mathcal{R}_h^p q_H := (\mathtt{id} - \mathcal{C}_h^p) q_H \tag{6.11}$$

for any $v_H \in V_H$ and $q_H \in Q_H$. Further, we define the corresponding multiscale spaces $\tilde{V}_H := \mathcal{R}_h^u V_H$ and $\tilde{Q}_H := \mathcal{R}_h^p Q_H$, where we omit the index h.

With these spaces, we can formulate the method presented in [MP17]: for $n \in \{1, \dots, N\}$, find $\bar{u}_H^n = \tilde{u}_H^n + u_h^{\mathrm{aux},n}$ with $\tilde{u}_H^n \in \tilde{V}_H$, $u_h^{\mathrm{aux},n} \in W_h^u$, and $\bar{p}_H^n \in \tilde{Q}_H$ such that

$$\begin{aligned}
a(\bar{u}_H^n, \tilde{v}_H) - d(\tilde{v}_H, \bar{p}_H^n) &= 0, \\
d(D_\tau \bar{u}_H^n, \tilde{q}_H) + c(D_\tau \bar{p}_H^n, \tilde{q}_H) + b(\bar{p}_H^n, \tilde{q}_H) &= (f^n, \tilde{q}_H)_{\mathcal{H}_Q}, \\
a(u_h^{\mathrm{aux},n}, w_h) + d(w_h, \bar{p}_H^n) &= 0
\end{aligned} \tag{6.12}$$

for all $\tilde{v}_H \in \tilde{V}_H$, $\tilde{q}_H \in \tilde{Q}_H$, and $w_h \in W_h^u$. Note that the initial condition is given by $\bar{p}_H^0 = \mathcal{R}_h^p p_h^0$. Moreover, we define $\bar{u}_H^0 = \tilde{u}_H^0 + u_h^{\mathrm{aux},0}$, where $u_h^{\mathrm{aux},0} \in W_h^u$ is given by the third equation of (6.12) and $\tilde{u}_H^0 \in \tilde{V}_H$ is obtained by

$$a(\bar{u}_H^0, \tilde{v}_H) = a(\tilde{u}_H^0, \tilde{v}_H) = d(\tilde{v}_H, \bar{p}_H^0)$$

for all $\tilde{v}_H \in \tilde{V}_H$. The system (6.12) is well-posed and the errors $\|u_h^n - \bar{u}_H^n\|_{\mathcal{V}}$ and $\|p_h^n - \bar{p}_H^n\|_{\mathcal{Q}}$ scale like H independently of ϵ; see [MP17, Thm. 5.2]. Together with Theorem 6.2.3, this implies that the multiscale solution $(\bar{u}_H^n, \bar{p}_H^n)$ approximates the exact solution (u, p) with an error of order $H + \tau$. Moreover, one may manipulate system (6.12) in such a way that, in practice, the additional fine-scale correction only needs to be computed in the offline stage using a set of basis functions. This keeps the coarse structure of the system in each time step at the expense of slightly more complicated systems (see [MP17]).

6.2.3 An alternative multiscale method

In this subsection, we propose an alternative approach to the method in (6.12) which does not require an additional fine-scale correction. To achieve this, we exploit some structural properties of the system. These become evident if we discretize system (6.3) in time first, i.e., if we consider

$$\begin{aligned} a(u^n, v) - d(v, p^n) &= 0, \\ d(D_\tau u^n, q) + c(D_\tau p^n, q) + b(p^n, q) &= (f^n, q)_{\mathcal{H}_Q} \end{aligned} \tag{6.13}$$

for all $v \in \mathcal{V}$, $q \in \mathcal{Q}$, and $n \in \{1, \dots, N\}$. We first prove that system (6.13) is well-posed.

Lemma 6.2.4 (Well-posedness)**.** *Let* $n \in \{1, \dots, N\}$ *and assume that* $u^{n-1} \in \mathcal{V}$ *and* $p^{n-1} \in \mathcal{Q}$ *are given. Then, system* (6.13) *is well-posed.*

Proof. We introduce the bilinear form $\mathfrak{b}\colon (\mathcal{V} \times \mathcal{Q}) \times (\mathcal{V} \times \mathcal{Q}) \to \mathbb{R}$ defined by

$$\mathfrak{b}([v, q], [w, r]) := a(v, w) - d(w, q) + d(v, r) + c(q, r) + \tau\, b(q, r) \tag{6.14}$$

for v, $w \in \mathcal{V}$ and q, $r \in \mathcal{Q}$. Note that $\mathfrak{b}$ is coercive, since

$$\mathfrak{b}([v, q], [v, q]) = \|v\|_a^2 + \|q\|_c^2 + \tau\, \|q\|_b^2.$$

Furthermore, system (6.13) is equivalent to

$$\mathfrak{b}([u^n, p^n], [v, q]) = \tau\, (f^n, q) + d(u^{n-1}, q) + c(p^{n-1}, q).$$

Thus, the well-posedness follows from the Lax-Milgram Theorem. □

We can now define an alternative correction operator based on the observation that the terms involving d in system (6.13) cancel for suitable test functions when summing both equations. Therefore, we propose to use the adapted correction operator $\tilde{\mathcal{C}}_h = \tilde{\mathcal{C}}_h^*\colon V_h \times Q_h \to W_h^u \times W_h^p$ simply defined by

$$\tilde{\mathcal{C}}_h[v_h, q_h] := [\mathcal{C}_h^u v_h, \mathcal{C}_h^p q_h]$$

for $v_h \in V_h$ and $q_h \in Q_h$, with the operators $\mathcal{C}_h^u$, $\mathcal{C}_h^p$ defined in (6.10). We show in the following that the corresponding multiscale method provides optimal orders of convergence as well. Note that with the correction operator $\tilde{\mathcal{C}}_h$, we retain the projections $\mathcal{R}_h^u$ and $\mathcal{R}_h^p$ as defined in (6.11) and the spaces $\tilde{V}_H = \mathcal{R}_h^u V_H$ and $\tilde{Q}_H = \mathcal{R}_h^p Q_H$. Here, however, we do not require an auxiliary correction as in Section 6.2.2.

Using the spaces defined above, we can formulate the alternative multiscale method. For this, we discretize system (6.13) in space and consider the problem: for $n \in \{1, \dots, N\}$, find $\tilde{u}_H^n \in \tilde{V}_H$ and $\tilde{p}_H^n \in \tilde{Q}_H$ such that

$$\begin{aligned} a(\tilde{u}_H^n, \tilde{v}_H) - d(\tilde{v}_H, \tilde{p}_H^n) &= 0, \\ d(D_\tau \tilde{u}_H^n, \tilde{q}_H) + c(D_\tau \tilde{p}_H^n, \tilde{q}_H) + b(\tilde{p}_H^n, \tilde{q}_H) &= (f^n, \tilde{q}_H)_{\mathcal{H}_Q} \end{aligned} \tag{6.15}$$

for all $\tilde{v}_H \in \tilde{V}_H$ and $\tilde{q}_H \in \tilde{Q}_H$. Note that this system is again well-posed by the arguments of Lemma 6.2.4. Given $\tilde{p}_H^0$, we define the initial condition $\tilde{u}_H^0$ as before by

$$a(\tilde{u}_H^0, \tilde{v}_H) = d(\tilde{v}_H, \tilde{p}_H^0)$$

for all $\tilde{v}_H \in \tilde{V}_H$.

Before we further investigate the method defined in (6.15), we provide an alternative characterization of the bilinear forms a and b in terms of operators. That is, we define $\mathcal{A}\colon V_h \to V_h$ and $\mathcal{B}\colon Q_h \to Q_h$ by

$$(\mathcal{A}v_h, w_h)_{\mathcal{H}_\mathcal{V}} := a(v_h, w_h), \qquad (\mathcal{B}q_h, r_h)_{\mathcal{H}_\mathcal{Q}} := b(q_h, r_h)$$

for all $w_h \in V_h$ and $r_h \in Q_h$. Note that these operators are only well-defined on the discrete spaces V_h. In the following two lemmas, we provide bounds for the projections defined above that are useful for the proof of convergence later on.

Lemma 6.2.5. *The projections $\mathcal{R}_h^u$ and $\mathcal{R}_h^p$ defined in* (6.10) *satisfy the bounds*

$$\begin{aligned} \|(\mathtt{id} - \mathcal{R}_h^u)v_h\|_{\mathcal{H}_\mathcal{V}} &\lesssim H\,\|(\mathtt{id} - \mathcal{R}_h^u)v_h\|_\mathcal{V} \lesssim H\,\|v_h\|_\mathcal{V}, \\ \|(\mathtt{id} - \mathcal{R}_h^p)q_h\|_{\mathcal{H}_\mathcal{Q}} &\lesssim H\,\|(\mathtt{id} - \mathcal{R}_h^p)q_h\|_\mathcal{Q} \lesssim H\,\|q_h\|_\mathcal{Q} \end{aligned}$$

for all $v_h \in V_h$ and $q_h \in Q_h$.

Proof. The proof is based on the arguments that are used in Theorem 2.3.1 and Theorem 3.2.6. We only show the first estimate since the second one follows analogously. Let $v_h \in V_h$. By (2.14) and the fact that $\mathcal{I}_H(\mathtt{id} - \mathcal{R}_h^u)v_h = 0$, it directly follows that

$$\|(\mathtt{id} - \mathcal{R}_h^u)v_h\|_{\mathcal{H}_\mathcal{V}} \lesssim H\,\|(\mathtt{id} - \mathcal{R}_h^u)v_h\|_\mathcal{V}.$$

The stability estimate then follows from $(\mathtt{id} - \mathcal{R}_h^u)v_h = \mathcal{C}_h^u v_h$, (6.4), and (6.10). To be more precise, it holds that

$$c_\sigma\,\|(\mathtt{id} - \mathcal{R}_h^u)v_h\|_{\mathcal{H}_\mathcal{V}}^2 \le a(\mathcal{C}_h^u v_h, \mathcal{C}_h^u v_h) = a(v_h, \mathcal{C}_h^u v_h) \le C_\sigma\,\|v_h\|_\mathcal{V}\,\|\mathcal{C}_h^u v_h\|_\mathcal{V},$$

which concludes the proof. □

Lemma 6.2.6. *The projections $\mathcal{R}_h^u$ and $\mathcal{R}_h^p$ defined in* (6.10) *are bounded in terms of $\mathcal{A}$ and $\mathcal{B}$ by*

$$\|(\mathtt{id} - \mathcal{R}_h^u)v_h\|_\mathcal{V} \lesssim H\,\|\mathcal{A}v_h\|_{\mathcal{H}_\mathcal{V}}, \qquad \|(\mathtt{id} - \mathcal{R}_h^p)q_h\|_\mathcal{Q} \lesssim H\,\|\mathcal{B}q_h\|_{\mathcal{H}_\mathcal{Q}}$$

for all $v_h \in V_h$ and $q_h \in Q_h$.

Proof. For $v_h \in V_h$, we get

$$\begin{aligned} c_\sigma\,\|v_h - \mathcal{R}_h^u v_h\|_\mathcal{V}^2 &\le a(v_h, v_h - \mathcal{R}_h^u v_h) = (\mathcal{A}v_h, v_h - \mathcal{R}_h^u v_h)_{\mathcal{H}_\mathcal{V}} \\ &\le \|\mathcal{A}v_h\|_{\mathcal{H}_\mathcal{V}}\,\|v_h - \mathcal{R}_h^u v_h\|_{\mathcal{H}_\mathcal{V}}. \end{aligned}$$

The claim then follows directly from Lemma 6.2.5. The proof of the result involving $\mathcal{B}$ follows the same lines. □

6.2.4 Convergence studies

The aim of this subsection is to prove that the solution provided by (6.15) approximates the fine-scale solution (u_h^n, p_h^n) of (6.6) up to order H. In combination with Theorem 6.2.3, this shows that the multiscale solution converges to the exact solution. More precisely, we obtain (assuming h sufficiently small) an error estimate which states that the error is bounded by $\mathcal{O}(H+\tau)$ independently of ϵ. Note that we assume here that the corrector problems are solved on the global domain. Since the localization of the corrections was already discussed in detail in the previous chapters, we omit the rigorous analysis of the localization procedure and refer to Section 2.4.3 for the details. We remark that the convergence result in Theorem 6.2.7 below remains valid if the involved localization parameter ℓ is chosen sufficiently large, i.e., $\ell \gtrsim |\log H|$.

The main result of this chapter reads as follows.

Theorem 6.2.7 (Error of the alternative multiscale method)**.** *Assume that* $f \in L^\infty(0,T;L^2(D)) \cap H^1(0,T;H^{-1}(D))$ *and consistent initial data* $u_h^0 \in V_h$, $p_h^0 \in Q_h$ *are given as well as* $\tilde u_H^0 \in \tilde V_H$ *and* $\tilde p_H^0 := \mathcal{R}_h^p p_h^0 \in \tilde Q_H$. *Then the error between the multiscale solution* $(\tilde u_H^n, \tilde p_H^n)$ *of* (6.15) *and the fine-scale solution* (u_h^n, p_h^n) *of* (6.6) *satisfies*

$$\|u_h^n - \tilde u_H^n\|_{\mathcal{V}} + \|p_h^n - \tilde p_H^n\|_{\mathcal{Q}} \lesssim H\, C_{\text{data}}^n + t_n^{-1/2} H\, \|p_h^0\|_{\mathcal{Q}}$$

for $n \in \{1, \dots, N\}$, *where* C_{data}^n *is defined by*

$$C_{\text{data}}^n := \|p_h^0\|_{\mathcal{Q}} + \|f\|_{L^2(0,t_n;L^2(D))} + \|f\|_{L^\infty(0,t_n;L^2(D))} + \|\partial_t f\|_{L^2(0,t_n;H^{-1}(D))}.$$

Proof. As in the proof of convergence for the multiscale method in [MP17], we split the errors in the displacement and pressure into two parts each, namely

$$\begin{aligned} \rho_u^n &:= u_h^n - \mathcal{R}_h^u u_h^n, & \eta_u^n &:= \mathcal{R}_h^u u_h^n - \tilde u_H^n, \\ \rho_p^n &:= p_h^n - \mathcal{R}_h^p p_h^n, & \eta_p^n &:= \mathcal{R}_h^p p_h^n - \tilde p_H^n. \end{aligned}$$

Thus, ρ_*^n contains the error of the projections and η_*^n the difference between the projection and the multiscale solution.

Step 1 (estimates of ρ_*^n): In a first step, we bound the projection error due to $\mathcal{R}_h^u$. For this, we apply Lemma 6.2.6 and use the first line of (6.6),

$$\begin{aligned} \|\rho_u^n\|_{\mathcal{V}} &= \|(\mathtt{id} - \mathcal{R}_h^u) u_h^n\|_{\mathcal{V}} \lesssim H\, \|\mathcal{A} u_h^n\|_{\mathcal{H}_\mathcal{V}} \\ &= H \sup_{v_h \in V_h} \frac{|a(u_h^n, v_h)|}{\|v_h\|_{\mathcal{H}_\mathcal{V}}} = H \sup_{v_h \in V_h} \frac{|d(v_h, p_h^n)|}{\|v_h\|_{\mathcal{H}_\mathcal{V}}} \lesssim H\, \|p_h^n\|_{\mathcal{Q}}, \end{aligned}$$

employing integration by parts in the last line. Theorem 6.2.2 then implies that $\|\rho_u^n\|_{\mathcal{V}}$ is bounded by

$$\|\rho_u^n\|_{\mathcal{V}} \lesssim H\, \big(\|p_h^0\|_{\mathcal{Q}} + \|f\|_{L^2(0,t_n;L^2(D))}\big) \lesssim H\, C_{\text{data}}^n.$$

Similarly, the projection error due to $\mathcal{R}_h^p$ can be bounded using the second line of (6.6), i.e.,

$$\begin{aligned}\|\rho_p^n\|_{\mathcal{Q}} = \|(\mathtt{id} - \mathcal{R}_h^p)p_h^n\|_{\mathcal{Q}} &\lesssim H\, \|\mathcal{B}p_h^n\|_{\mathcal{H}_\mathcal{Q}} \\ &= H \sup_{q_h \in Q_h} \frac{|b(p_h^n, q_h)|}{\|q_h\|_{\mathcal{H}_\mathcal{Q}}} \\ &= H \sup_{q_h \in Q_h} \frac{|(f^n, q_h) - d(D_\tau u_h^n, q_h) - c(D_\tau p_h^n, q_h)|}{\|q_h\|_{\mathcal{H}_\mathcal{Q}}} \\ &\lesssim H\, \big(\|f^n\|_{\mathcal{H}_\mathcal{Q}} + \|D_\tau u_h^n\|_{\mathcal{V}} + \|D_\tau p_h^n\|_{\mathcal{H}_\mathcal{Q}}\big).\end{aligned}$$

Using Theorem 6.2.2, we obtain the bounds $\|\rho_p^n\|_{\mathcal{Q}} \lesssim H\, C_{\text{data}}^n$ if $p_h^0 = 0$ and $\|\rho_p^n\|_{\mathcal{Q}} \lesssim t_n^{-1/2} H\, \|p_h^0\|_{\mathcal{Q}}$ if $f = 0$.

Step 2: In order to bound the remaining errors, we consider specific test functions within the systems (6.6) and (6.15). Using the definition of $\mathcal{R}_h^u$, we have for all $\tilde{v}_H \in \tilde{V}_H \subseteq V_h$ that

$$\begin{aligned} a(\eta_u^n, \tilde{v}_H) - d(\tilde{v}_H, \eta_p^n) &= a(\mathcal{R}_h^u u_h^n, \tilde{v}_H) - d(\tilde{v}_H, \mathcal{R}_h^p p_h^n) \\ &= a(u_h^n, \tilde{v}_H) - d(\tilde{v}_H, \mathcal{R}_h^p p_h^n) = d(\tilde{v}_H, \rho_p^n). \end{aligned} \tag{6.16}$$

Similarly, we have for all $\tilde{q}_H \in \tilde{Q}_H$ that

$$\begin{aligned} &d(D_\tau \eta_u^n, \tilde{q}_H) + c(D_\tau \eta_p^n, \tilde{q}_H) + b(\eta_p^n, \tilde{q}_H) \\ &\quad= d(D_\tau \mathcal{R}_h^u u_h^n, \tilde{q}_H) + c(D_\tau \mathcal{R}_h^p p_h^n, \tilde{q}_H) + b(p_h^n, \tilde{q}_H) - (f^n, \tilde{q}_H)_{\mathcal{H}_\mathcal{Q}} \\ &\quad= -d(D_\tau \rho_u^n, \tilde{q}_H) - c(D_\tau \rho_p^n, \tilde{q}_H), \end{aligned} \tag{6.17}$$

using the definition of $\mathcal{R}_h^p$. Combining equation (6.16) at the time steps n and $(n-1)$, we obtain

$$a(D_\tau \eta_u^n, \tilde{v}_H) - d(\tilde{v}_H, D_\tau \eta_p^n) = d(\tilde{v}_H, D_\tau \rho_p^n) \tag{6.18}$$

for any $\tilde{v}_H \in \tilde{V}_H$. Note that these equations are also valid for $n = 1$ because of the construction of u_h^0 and $\tilde{u}_H^0$. To obtain bounds for η_*^n, we consider the two cases where either $p_h^0 = 0$ or $f = 0$. This is done in the next two steps. An application of the triangle inequality then gives the stated result.

Step 3 (estimates of η_*^n if $p_h^0 = 0$): Note that $p_h^0 = 0$ also implies $u_h^0 = 0$. We now insert the test function $\tilde{v}_H = D_\tau \eta_u^n$ into (6.18) and add this to equation (6.17) with $\tilde{q}_H = D_\tau \eta_p^n$. Together, this yields

$$\begin{aligned} &a(D_\tau \eta_u^n, D_\tau \eta_u^n) + c(D_\tau \eta_p^n, D_\tau \eta_p^n) + b(\eta_p^n, D_\tau \eta_p^n) \\ &\qquad= d(D_\tau \eta_u^n, D_\tau \rho_p^n) - d(D_\tau \rho_u^n, D_\tau \eta_p^n) - c(D_\tau \rho_p^n, D_\tau \eta_p^n) \end{aligned}$$

and thus

$$\begin{aligned} &\|D_\tau \eta_u^n\|_a^2 + \|D_\tau \eta_p^n\|_c^2 + b(\eta_p^n, D_\tau \eta_p^n) \\ &\quad\le C_\alpha\, \|D_\tau \eta_u^n\|_{\mathcal{V}}\, \|D_\tau \rho_p^n\|_{\mathcal{H}_\mathcal{Q}} + C_\alpha\, \|D_\tau \rho_u^n\|_{\mathcal{V}}\, \|D_\tau \eta_p^n\|_{\mathcal{H}_\mathcal{Q}} \\ &\qquad + C_M\, \|D_\tau \rho_p^n\|_{\mathcal{H}_\mathcal{Q}}\, \|D_\tau \eta_p^n\|_{\mathcal{H}_\mathcal{Q}} \\ &\quad\le \frac{1}{2}\, \|D_\tau \eta_u^n\|_a^2 + \frac{1}{2}\, \|D_\tau \eta_p^n\|_c^2 + C\, \|D_\tau \rho_p^n\|_{\mathcal{H}_\mathcal{Q}}^2 + C'\, \|D_\tau \rho_u^n\|_{\mathcal{V}}^2. \end{aligned}$$

We can eliminate $\|D_\tau \eta_u^n\|_a$ and $\|D_\tau \eta_p^n\|_c$ on the right-hand side and multiply the estimate by 2τ. Then, a summation over n yields

$$\tau \sum_{j=1}^{n} \|D_\tau \eta_u^j\|_{\mathcal{V}}^2 + \tau \sum_{j=1}^{n} \|D_\tau \eta_p^j\|_{\mathcal{H}_\mathcal{Q}}^2 + \|\eta_p^n\|_{\mathcal{Q}}^2 \lesssim 2\tau \sum_{j=1}^{n} \|D_\tau \rho_p^j\|_{\mathcal{H}_\mathcal{Q}}^2 + 2\tau \sum_{j=1}^{n} \|D_\tau \rho_u^j\|_{\mathcal{V}}^2,$$

where we use that $\eta_p^0 = 0$. The sum including $D_\tau \rho_u^j$ can be bounded using once more Lemma 6.2.6, i.e.,

$$\begin{aligned} \|D_\tau \rho_u^j\|_{\mathcal{V}} = \|(\mathtt{id} - \mathcal{R}_h^u) D_\tau u_h^j\|_{\mathcal{V}} &\lesssim H \sup_{v_h \in V_h} \frac{|a(D_\tau u_h^j, v_h)|}{\|v_h\|_{\mathcal{H}_\mathcal{V}}} \\ = H \sup_{v_h \in V_h} \frac{|d(v_h, D_\tau p_h^j)|}{\|v_h\|_{\mathcal{H}_\mathcal{V}}} &\lesssim H \, \|D_\tau p_h^j\|_{\mathcal{Q}}. \end{aligned} \tag{6.19}$$

Together with Theorem 6.2.2, this leads to

$$\tau \sum_{j=1}^{n} \|D_\tau \rho_u^j\|_{\mathcal{V}}^2 \lesssim \tau H^2 \sum_{j=1}^{n} \|D_\tau p_h^j\|_{\mathcal{Q}}^2 \lesssim (H \, C_{\text{data}}^n)^2.$$

Then again, the sum including $D_\tau \rho_p^j$ can be bounded using

$$\|D_\tau \rho_p^j\|_{\mathcal{H}_\mathcal{Q}} = \|(\mathtt{id} - \mathcal{R}_h^p) D_\tau p_h^j\|_{\mathcal{H}_\mathcal{Q}} \lesssim H \, \|D_\tau p_h^j\|_{\mathcal{Q}},$$

which follows from Lemma 6.2.5 and results in

$$\tau \sum_{j=1}^{n} \|D_\tau \rho_p^j\|_{\mathcal{H}_\mathcal{Q}}^2 \le \tau \sum_{j=1}^{n} H^2 \, \|D_\tau p_h^j\|_{\mathcal{Q}}^2 \lesssim (H \, C_{\text{data}}^n)^2.$$

This does not only provide the bound $\|\eta_p^n\|_{\mathcal{Q}} \lesssim H \, C_{\text{data}}^n$ but also

$$\|\eta_u^n\|_{\mathcal{V}} \lesssim \|\rho_p^n\|_{\mathcal{H}_\mathcal{Q}} + \|\eta_p^n\|_{\mathcal{H}_\mathcal{Q}} \lesssim H \, C_{\text{data}}^n,$$

where we employ (6.16).

Step 4 (estimates of η_*^n if $f = 0$): We emphasize that by assumption also $\eta_p^0 = 0$ in this case. Together with (6.16), this yields

$$\|\eta_u^0\|_{\mathcal{V}}^2 \lesssim a(\eta_u^0, \eta_u^0) = d(\eta_u^0, \eta_p^0) + d(\eta_u^0, \rho_p^0) \lesssim \|\eta_u^0\|_{\mathcal{V}} \, \|\rho_p^0\|_{\mathcal{H}_\mathcal{Q}}$$

and, therefore,

$$\|\eta_u^0\|_{\mathcal{V}} \lesssim \|\rho_p^0\|_{\mathcal{H}_\mathcal{Q}} \lesssim H \, \|p_h^0\|_{\mathcal{Q}}.$$

Note that it is sufficient to bound $\|\eta_p^n\|_{\mathcal{Q}}$ in terms of $H \, C_{\text{data}}^n$ since by (6.16) $\|\eta_u^n\|_{\mathcal{V}} \lesssim \|\rho_p^n\|_{\mathcal{H}_\mathcal{Q}} + \|\eta_p^n\|_{\mathcal{H}_\mathcal{Q}}$. As in Step 3, we consider the sum of equation (6.18) with $\tilde{v}_H = D_\tau \eta_u^n$ and equation (6.17) with $\tilde{q}_H = D_\tau \eta_p^n$. Multiplying the result by 2τ, we get

$$2\tau \|D_\tau \eta_u^n\|_a^2 + 2\tau \|D_\tau \eta_p^n\|_c^2 + \|\eta_p^n\|_b^2 - \|\eta_p^{n-1}\|_b^2 \lesssim 2\tau \|D_\tau \rho_p^n\|_{\mathcal{H}_\mathcal{Q}}^2 + 2\tau \|D_\tau \rho_u^n\|_{\mathcal{V}}^2.$$

Another multiplication by t_n^2 and the estimate $t_n^2 - t_{n-1}^2 \leq 3\tau t_{n-1}$ then lead to

$$\begin{aligned} 2\tau t_n^2 \|D_\tau \eta_u^n\|_a^2 + 2\tau t_n^2 \|D_\tau \eta_p^n\|_c^2 + t_n^2 \|\eta_p^n\|_b^2 - t_{n-1}^2 \|\eta_p^{n-1}\|_b^2 \\ \lesssim 2\tau t_n^2 \|D_\tau \rho_p^n\|_{\mathcal{H}_\mathcal{Q}}^2 + 2\tau t_n^2 \|D_\tau \rho_u^n\|_{\mathcal{V}}^2 + 3\tau t_{n-1} \|\eta_p^{n-1}\|_{\mathcal{Q}}^2. \end{aligned}$$

Taking the sum, we obtain

$$\begin{aligned} &\tau \sum_{j=1}^n t_j^2 \|D_\tau \eta_u^j\|_{\mathcal{V}}^2 + t_n^2 \|\eta_p^n\|_Q^2 \\ &\qquad \lesssim \tau \sum_{j=1}^n t_j^2 \|D_\tau \eta_u^j\|_a^2 + \sum_{j=1}^n \Big(t_j^2 \|\eta_p^j\|_b^2 - t_{j-1}^2 \|\eta_p^{j-1}\|_b^2 \Big) \qquad (6.20) \\ &\qquad \lesssim \tau \sum_{j=1}^n t_j^2 \|D_\tau \rho_p^j\|_{\mathcal{H}_\mathcal{Q}}^2 + \tau \sum_{j=1}^n t_j^2 \|D_\tau \rho_u^j\|_{\mathcal{V}}^2 + \tau \sum_{j=1}^{n-1} t_j \|\eta_p^j\|_{\mathcal{Q}}^2. \end{aligned}$$

To bound the first sum on the right-hand side, we apply first Lemma 6.2.5 and then Theorem 6.2.2 and obtain

$$\tau \sum_{j=1}^n t_j^2 \|D_\tau \rho_p^j\|_{\mathcal{H}_\mathcal{Q}}^2 \lesssim \tau \sum_{j=1}^n t_j^2 H^2 \|D_\tau p_h^j\|_Q^2 \lesssim \tau \sum_{j=1}^n H^2 \|p_h^0\|_{\mathcal{H}_\mathcal{Q}}^2 = t_n H^2 \|p_h^0\|_{\mathcal{Q}}^2.$$

For the second sum, we use the estimate $\|D_\tau \rho_u^j\|_{\mathcal{V}} \lesssim H \|D_\tau p_h^j\|_{\mathcal{Q}}$ from (6.19), which is also valid for non-zero initial conditions. With Theorem 6.2.2, we further get

$$\tau \sum_{j=1}^n t_j^2 \|D_\tau \rho_u^j\|_{\mathcal{V}}^2 \lesssim \tau \sum_{j=1}^n t_j^2 H^2 \|D_\tau p_h^j\|_{\mathcal{Q}}^2 \lesssim t_n H^2 \|p_h^0\|_{\mathcal{Q}}^2.$$

Step 5 (estimate of the last sum in (6.20)): In order to bound the third sum on the right-hand side of (6.20), we consider the sum of (6.16) and (6.17). For test functions $\tilde{v}_H = D_\tau \eta_u^n$ and $\tilde{q}_H = \eta_p^n$, we get after multiplication with $2\tau t_n$ and an application of Young's inequality

$$\begin{aligned} t_n \big(\|\eta_u^n\|_a^2 - \|\eta_u^{n-1}\|_a^2\big) + t_n \big(\|\eta_p^n\|_c^2 - \|\eta_p^{n-1}\|_c^2\big) + 2\tau t_n \|\eta_p^n\|_b^2 \\ \lesssim \gamma \tau t_n^2 \|D_\tau \eta_u^n\|_{\mathcal{V}}^2 + \gamma^{-1} \tau \|\rho_p^n\|_{\mathcal{H}_\mathcal{Q}}^2 + \tau t_n^2 \|D_\tau \rho_u^n\|_{\mathcal{H}_\mathcal{V}}^2 \\ + \tau t_n^2 \|D_\tau \rho_p^n\|_{\mathcal{H}_\mathcal{Q}}^2 + \tau \|\eta_p^n\|_{\mathcal{H}_\mathcal{Q}}^2 \end{aligned}$$

for any $\gamma > 0$. We add $\tau\,\|\eta_u^{n-1}\|_a^2 + \tau\,\|\eta_p^{n-1}\|_c^2$ on both sides and take the sum over n such that we obtain

$$\begin{aligned} t_n\,\|\eta_u^n\|_{\mathcal{V}}^2 + t_n\,\|\eta_p^n\|_{\mathcal{H}_\mathcal{Q}}^2 + \sum_{j=1}^n \tau t_j\,\|\eta_p^j\|_{\mathcal{Q}}^2 &\lesssim \gamma\,\underbrace{\tau\sum_{j=1}^n t_j^2\,\|D_\tau \eta_u^j\|_{\mathcal{V}}^2}_{\textcircled{a}} + \frac{1}{\gamma}\,\underbrace{\tau\sum_{j=1}^n \|\rho_p^j\|_{\mathcal{H}_\mathcal{Q}}^2}_{\textcircled{b}} \\ &\quad \underbrace{+\,\tau\sum_{j=1}^n t_j^2\left(\|D_\tau \rho_u^j\|_{\mathcal{H}_\mathcal{V}}^2 + \|D_\tau \rho_p^j\|_{\mathcal{H}_\mathcal{Q}}^2\right)}_{\textcircled{c}} \underbrace{+\,\tau\sum_{j=1}^n \left(\|\eta_p^j\|_{\mathcal{H}_\mathcal{Q}}^2 + \|\eta_u^{j-1}\|_{\mathcal{V}}^2\right)}_{\textcircled{d}}. \end{aligned}$$

Note that the sum on the left-hand side is the term we aim to bound. For a sufficiently small γ which only depends on the generic constant of the estimates, we can eliminate ⓐ with the left-hand side in (6.20). For the remaining three parts on the right-hand side, we estimate

$$\textcircled{b} = \tau\sum_{j=1}^n \|\rho_p^j\|_{\mathcal{H}_\mathcal{Q}}^2 \lesssim \tau\sum_{j=1}^n H^2\,\|p_h^j\|_{\mathcal{Q}}^2 \lesssim \tau\sum_{j=1}^n H^2\,\|p_h^0\|_{\mathcal{Q}}^2 = t_n H^2\,\|p_h^0\|_{\mathcal{Q}}^2$$

and, with Lemma 6.2.5 and Theorem 6.2.2,

$$\begin{aligned} \textcircled{c} &\lesssim \tau\sum_{j=1}^n H^2\,t_j^2\left(\|D_\tau u_h^j\|_{\mathcal{V}}^2 + \|D_\tau p_h^j\|_{\mathcal{Q}}^2\right) \\ &\lesssim \tau\,(t_n+1)\sum_{j=1}^n H^2\,\|p_h^0\|_{\mathcal{Q}}^2 = (t_n^2 + t_n)\,H^2\,\|p_h^0\|_{\mathcal{Q}}^2. \end{aligned}$$

Finally, with the equations (6.16) and (6.17) as well as the test functions $\tilde{v}_H = \eta_u^n$ and $\tilde{q}_H = \eta_p^n$, one can show as in [MP17] that also ⓓ $\lesssim t_n H^2\,\|p_h^0\|_{\mathcal{Q}}^2$. In summary, this yields

$$\|\eta_p^n\|_{\mathcal{Q}} \lesssim (1 + t_n^{-1/2})\,H\,\|p_h^0\|_{\mathcal{Q}},$$

which concludes the proof. □

Theorem 6.2.7 shows together with Theorem 6.2.3 that the multiscale method proposed in (6.15) converges linearly, i.e., the error is bounded by $\mathcal{O}(H+\tau)$ if we consider the $L^\infty(0,T;\mathcal{V})$-norm for u and the $L^\infty(0,T;\mathcal{H}_\mathcal{Q}) \cap L^2(0,T;Q)$-norm for p. We emphasize that the involved constants are independent of derivatives of the coefficients μ, λ, κ, and α.

Remark 6.2.8. The approach of discarding the coupling term in the stationary system to obtain two decoupled projection operators is also used in [FAC$^+$19] for the problem of linear poroelasticity with high contrast employing the multiscale technique referred to as CEM-GMsFEM. Further, it is applied to more general (homogeneous) elliptic-parabolic problems in the context of semi-explicit time discretization schemes in [AMU19].

6.3 Numerical experiments

In order to assess the method numerically, we consider numerical examples in two and three space dimensions. We measure the error in the discrete $L^2(H^1)$-norm

$$\|(v,q)\|_{N,1} := \Bigg(\sum_{j=1}^{N} \tau\Big(\|\nabla v(j\tau)\|^2_{\mathcal{H}_\mathcal{V}} + \|\nabla q(j\tau)\|^2_{\mathcal{H}_\mathcal{Q}}\Big)\Bigg)^{1/2},$$

where $N = T/\tau$ is the number of time steps. Further, we set $D = (0,1)^d$ as the domain and $T = 1$ as final time with time step size $\tau = 0.01$ (and thus $N = 100$) for both the two-dimensional examples and the example in three dimensions.

The reference solution (u_h, p_h) is computed on a regular uniform mesh $\mathcal{T}_h$ consisting of elements with given mesh size h. The local corrector problems are also solved on patches with mesh size h. The parameters are chosen to be piecewise constant on elements of $\mathcal{T}_\epsilon$ and the value is obtained as a uniformly distributed random number between two given bounds, i.e., for any $K \in \mathcal{T}_\epsilon$ we have

$$\begin{aligned} \kappa|_K &\sim U(0.1, 0.3), &\quad \mu|_K &\sim U(40, 70), \\ \lambda|_K &\sim U(30, 60), &\quad \alpha|_K &\sim U(0.5, 1) \end{aligned} \tag{6.21}$$

and $M = \nu = 1$, where $\mathcal{T}_\epsilon$ is a mesh with mesh size $\epsilon > h$ to guarantee that the reference solution is reasonable. Note that we take representative global samples for the above parameters. For the second two-dimensional example, the coefficients are chosen with the pattern depicted in Figure 6.1 scaled to the respective parameter range as given in (6.21). In all numerical tests, the localization parameter is set to $\ell = 2$ which showed to be sufficient. Note, however, that the choice of the localization parameter generally needs to be increased for smaller values of H and may be decreased for larger H as quantified in Chapter 2 and [HP13].

6.3.1 Two-dimensional examples

In all two-dimensional experiments, the fine mesh size is set to $h = 2^{-8}$, and $\epsilon = 2^{-6}$.

For the first example, we set $f = 1$ and $p^0(x) = (1-x_1)\,x_1\,(1-x_2)\,x_2$. We prescribe homogeneous Dirichlet boundary conditions for p on ∂D, homogeneous Dirichlet boundary conditions for u on $\{x \in \partial D \colon x_2 = 0 \text{ or } x_2 = 1\}$ and homogeneous Neumann boundary conditions on $\{x \in \partial D \colon x_1 = 0 \text{ or } x_1 = 1\}$. The errors for different values of H are shown in Figure 6.1 (right, □). The results are in line with the theory and indicate a convergence rate even slightly better than 1 with respect to the coarse mesh size H.

In the second example, we consider $p^0(x) = \sqrt{1-x_2}$, $f = 0$, and enforce homogeneous Dirichlet boundary conditions for u and p on $\{x \in \partial D \colon x_2 = 1\}$ and homogeneous Neumann boundary conditions on the remaining part of ∂D. As mentioned above, all coefficients in this example are chosen with the pattern

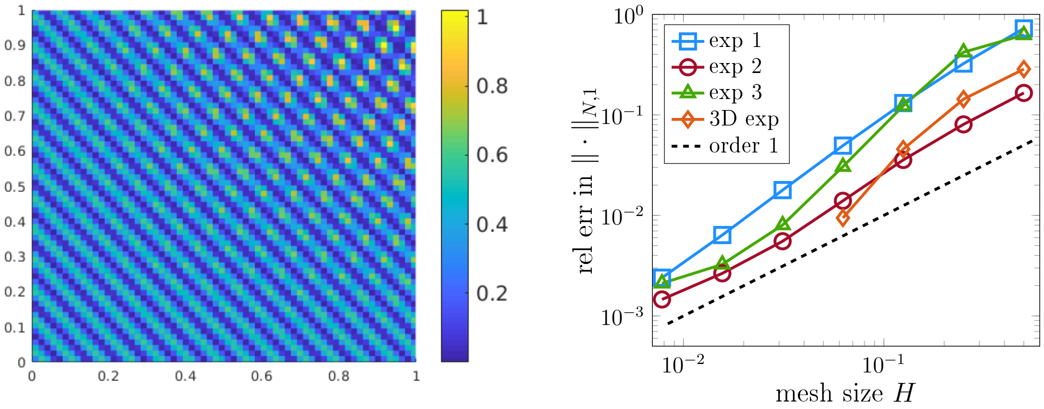


Figure 6.1: Multiscale pattern (left) and relative errors of the LOD method with respect to H in the two- and three-dimensional setting, measured in the discrete $L^2(H^1)$-norm (right).

depicted in Figure 6.1 (left). In this example, the predicted linear convergence can be observed, cf. Figure 6.1 (right, ○).

In Figure 6.1 (right, △), we also present the results of the third example, where $p^0(x) = (1 - x_2)\, x_2$ and $f(x,t) = 3\,t \cos(2\pi\, x_1) \sin(3\pi\, x_2)$. Further, we take the same boundary conditions as in the previous example. On the one hand, the errors in this example partially indicate a higher-order convergence rate. On the other hand, the error curve slightly stagnates for smaller values of H, which can be explained by the effect of the localization error.

6.3.2 Three-dimensional example

For the three-dimensional setting, we restrict ourselves to $h = 2^{-5}$ and $\epsilon = 2^{-4}$ due to the high computational complexity. We choose the coefficients as in (6.21), set $f = 0$, $p^0(x) = (1 - x_1)\, x_1\, (1 - x_2)\, x_2\, (1 - x_3)\, x_3$, and prescribe homogeneous Dirichlet boundary conditions on $\{x \in \partial D \colon x_3 = 1\}$ and homogeneous Neumann boundary conditions on the remaining part of ∂D. Further, we set $\ell = 2$ as before. The errors for this example are plotted in Figure 6.1 (right, ◇) and show at least the expected linear convergence rate. Moreover, this example indicates that the three-dimensional setting can be handled if appropriate computing capacities are available.

7 Conclusion and Outlook

7.1 Conclusion

This thesis was concerned with the coarse-scale numerical approximation of solutions of partial differential equations that involve one or more heterogeneous coefficients, possibly with oscillations on fine scales. To avoid global computations that resolve the varying coefficients, we employed the Localized Orthogonal Decomposition technique to efficiently deal with the presence of multiple scales or even a continuum of scales without restrictive structural assumptions. We presented the classical first-order approach in a relatively general setting with a rigorous analysis of the convergence behavior and illustrative examples that showed the practical performance of the method. Moreover, we extended the first-order multiscale approach to a higher-order variant in the elliptic setting based on the saddle point formulation of the classical method. The higher-order method was constructed from discontinuous finite element spaces which are favorable to extract higher-order convergence rates. In particular, the method allowed for a thorough tracing of the mesh size, the polynomial degree, and the localization parameter. We presented numerical experiments that indicate an even better dependence on the involved parameters than predicted by the theory.

The applicability of the above approach to general heterogeneous coefficients motivated a strategy to reconstruct the effective behavior of solutions to multiscale problems from given coarse measurements in connection with an inverse diffusion problem. The idea of the approach was to use the knowledge that system matrices corresponding to the Localized Orthogonal Decomposition technique, as well as other numerical homogenization approaches, obey a certain quasi-local sparsity pattern. Prescribing such a pattern then allowed us to reconstruct coarse models that recover available measurements. Since the numerical results showed that the inversion procedure favors quasi-local models with some deviation from locality, these results, in turn, emphasized the general potential of numerical homogenization methods.

Subsequently, we applied the framework of Localized Orthogonal Decomposition to two time-dependent problems. For the acoustic wave equation, we combined the method with an explicit time discretization scheme and achieved a complexity reduction in space and in time. That is, the construction of a multiscale space for the spatial discretization led to smaller systems to solve in every time step and, additionally, enabled the use of larger time steps subject to a relaxation of the time step restriction. We rigorously studied the convergence behavior of the method and presented numerical illustrations that confirmed the theoretical findings.

In connection with the multiphysics problem of linear poroelasticity, we then combined the Localized Orthogonal Decomposition method with an implicit Euler scheme in time and presented an adapted approach which was not based on the stationary equations as it is normally done. Instead, we exploited the saddle point structure of the system after a temporal discretization, which motivated a decoupling in the construction of the multiscale spaces. This construction resulted in a simple method for which we could prove first-order convergence and validate the findings with numerical experiments.

7.2 Outlook

The work presented in this thesis opens up many possibilities for future research. To start with, the decay estimates for the higher-order method in Chapter 3 are not sharp as discussed in Remark 3.3.3. Since also the numerical experiments indicate a better scaling, a natural next step is trying to improve the estimates in terms of a better decay rate with respect to the polynomial degree. Further, one could aim for a modification of the method to reduce the pollution in terms of the mesh size which occurs if the localization parameter is not increased accordingly (cf. Theorem 3.3.4). Moreover, the higher-order approach in connection with time discretization schemes could be investigated for the presented time-dependent problems.

Concerning the findings in Chapter 4, a natural next step would be to investigate how to extract information about the actual fine-scale coefficient from the reconstructed model using, e.g., additional structural knowledge if available. Besides, an application of the approach to problems beyond the elliptic framework could be studied.

With regard to time-dependent problems, multiscale approaches where fine-scale coefficients also depend on the temporal variable mark an interesting class of problems in connection with numerical homogenization not only in space but also in time. Such considerations could as well be valuable in the context of long-time wave propagation, which is an active field of research.

List of Symbols

$\mathfrak{A}$	set of admissible coefficients	36
$\mathfrak{B}$	FE basis	16
$\mathfrak{B}_K$	local FE basis for the element K	44
$\mathcal{C}$, $\mathcal{C}^*$	correction operator	15
d	spatial dimension	10
dim	dimension of a space	15
dist	Hausdorff distance	40
dist_f	distance between solution operators	66
div	divergence	36
D_τ	discrete time derivative	86
E^b	extension operator	63
E^b_H	discrete extension operator	63
$\mathcal{E}^n$	discrete energy for the wave equation	86
$\lVert \cdot \rVert_a$	energy norm with respect to the bilinear form a	29
$\lVert \cdot \rVert_{N,0}$	discrete $L^2(L^2)$-norm	96
$\lVert (\cdot,\cdot) \rVert_{N,1}$	combined discrete $L^2(H^1)$-norm for the displacement and the pressure	117
$\lVert \cdot \rVert_{N,a}$	discrete energy norm	96
ϵ	microscopic scale	13
H	coarse discretization parameter	11
h	fine discretization parameter; used as subscript for discretized operators or spaces	27
$H^{-1}(D)$	dual space of $H^1_\Gamma(D)$	82
$H^1_\Gamma(D)$	space of H^1-functions with zero traces on Γ	10
$H^k(\mathcal{T}_H)$	broken H^k-space	38
$\lvert \cdot \rvert_{H^k(\mathcal{T}_H)}$	broken H^k-seminorm	38
$\mathcal{H}$, $\mathcal{H}_\mathcal{V}$, $\mathcal{H}_\mathcal{Q}$	(general) L^2-space	10
$\mathtt{id}$	identity operator	12
$\mathbb{1}_S$	indicator function of the set S	30
$\mathcal{I}_H$	projective quasi-interpolation operator	13
$\mathcal{J}_H$	error functional for the inverse problem	71
ker	kernel of an operator	14
ℓ	localization/oversampling parameter; used as superscript for localized operators or spaces	23
$\mathfrak{L}_A$	solution operator	64
$\mathfrak{L}^{\mathrm{eff}}_{A,\ell}$	effective solution operator	64

$\mathfrak{L}^{\mathrm{eff}}_{S_H}$	effective solution operator based on the stiffness matrix S_H	67
$\lesssim$	less than or equal to up to a constant	13
$L^p(0,T;X)$	Bochner space	82
m_K	number of local basis functions in an element K	20
$\mathcal{M}(\ell,\mathcal{T}_H)$	set of matrices with prescribed sparsity pattern	68
N	number of time steps	83
$\mathsf{N}(S)$	element patch around S	14
$\mathsf{N}^\ell(S)$	ℓ-neighborhood of S	14
p	polynomial degree (Ch. 3); pressure (Ch. 6)	38
Π, Π^p_H	L^2-projection onto piecewise polynomials	39
$\mathcal{R}$, $\mathcal{R}^*$	mapping onto the ideal multiscale trial/test space	16
$\mathcal{R}^u_h$, $\mathcal{R}^p_h$	discrete mapping onto the multiscale space for displacement/pressure	109
R	restriction operator	63
R_H	discrete restriction operator	64
$\mathcal{S}(\ell,\mathcal{T}_H)$	set of LOD stiffness matrices	66
supp	support of a function	21
τ	time step	83
tr	trace operator	62
$\mathcal{T}_H$	regular and quasi-uniform mesh	11
$\mathcal{V}$, $\mathcal{Q}$	(general) H^1-space with (partially) prescribed zero traces	10
$\bar{\mathcal{V}}$	(general) H^1-space	10
V_H, Q_H	conforming first-order coarse FE space	11
$\bar{V}_H$	first-order coarse FE space without boundary conditions	63
$\tilde{V}_H$, $\tilde{Q}_H$	ideal multiscale (test) space	15
V_h, Q_h	conforming first-order fine FE space	27
V^p_H	discontinuous higher-order coarse FE space	38
$\tilde{V}^p_H$	ideal higher-order multiscale space	42
$V_{h,p'}$	conforming higher-order fine FE space	50
$\mathcal{W}$	fine-scale space	14
W_h, W^u_h, W^p_h	discrete fine-scale space	27
$\mathcal{X}$	trace space of $H^1(D)$	62
X_H	trace space of $\bar{V}_H$	63

Acronyms

ALB	Adaptive Local Basis	4
CFL	Courant-Friedrichs-Lewy	6
cG	continuous Galerkin	9
cPG	continuous Petrov-Galerkin	15
dG	discontinuous Galerkin	35
DOF	degree of freedom	3
EOC	experimental order of convergence	55
FE	finite element	2
GFEM	Generalized Finite Element Method	3
GMsFEM	Generalized Multiscale Finite Element Method	4
HMM	Heterogeneous Multiscale Method	3
LOD	Localized Orthogonal Decomposition	4
MsFEM	Multiscale Finite Element Method	3
ODE	ordinary differential equation	89
PDE	partial differential equation	1
RPS	Rough Polyharmonic Splines	4
VMM	Variational Multiscale Method	4

Bibliography

[ACM$^+$20] R. Altmann, E. Chung, R. Maier, D. Peterseim, and S.-M. Pun. Computational multiscale methods for linear heterogeneous poroelasticity. *J. Comput. Math.*, 38(1):41–57, 2020.

[AG11] A. Abdulle and M. J. Grote. Finite element heterogeneous multiscale method for the wave equation. *Multiscale Model. Simul.*, 9(2):766–792, 2011.

[AH17] A. Abdulle and P. Henning. Localized orthogonal decomposition method for the wave equation with a continuum of scales. *Math. Comp.*, 86(304):549–587, 2017.

[All92] G. Allaire. Homogenization and two-scale convergence. *SIAM J. Math. Anal.*, 23(6):1482–1518, 1992.

[All97] G. Allaire. Mathematical approaches and methods. In: U. Hornung (ed.), *Homogenization and porous media*, volume 6 of *Interdisciplinary Applied Mathematics*, pp. 225–250. Springer, New York, 1997.

[AMU19] R. Altmann, R. Maier, and B. Unger. Semi-explicit discretization schemes for weakly-coupled elliptic-parabolic problems. *ArXiv Preprint*, 1909.03497, 2019.

[AR14] D. Arjmand and O. Runborg. Analysis of heterogeneous multiscale methods for long time wave propagation problems. *Multiscale Model. Simul.*, 12(3):1135–1166, 2014.

[Arn82] D. N. Arnold. An interior penalty finite element method with discontinuous elements. *SIAM J. Numer. Anal.*, 19(4):742–760, 1982.

[Bab71] I. Babuška. Error-bounds for finite element method. *Numer. Math.*, 16:322–333, 1971.

[BBF13] D. Boffi, F. Brezzi, and M. Fortin. *Mixed finite element methods and applications.* Springer, Heidelberg, 2013.

[BCO94] I. Babuška, G. Caloz, and J. E. Osborn. Special finite element methods for a class of second order elliptic problems with rough coefficients. *SIAM J. Numer. Anal.*, 31(4):945–981, 1994.

[BG96] I. Babuška and B. Q. Guo. Approximation properties of the *hp* version of the finite element method. *Comput. Methods Appl. Mech. Engrg.*, 133(3-4):319–346, 1996.

[BGP17] D. Brown, D. Gallistl, and D. Peterseim. Multiscale Petrov-Galerkin method for high-frequency heterogeneous Helmholtz equations. In: *Meshfree methods for partial differential equations VIII*, volume 115 of *Lect. Notes Comput. Sci. Eng.*, pp. 85–115. Springer, Cham, 2017.

[Bio41] M. A. Biot. General theory of three-dimensional consolidation. *J. Appl. Phys.*, 12(2):155–164, 1941.

[BL11] I. Babuška and R. Lipton. Optimal local approximation spaces for generalized finite element methods with application to multiscale problems. *Multiscale Model. Simul.*, 9(1):373–406, 2011.

[BO83] I. Babuška and J. E. Osborn. Generalized finite element methods: their performance and their relation to mixed methods. *SIAM J. Numer. Anal.*, 20(3):510–536, 1983.

[BP16] D. L. Brown and D. Peterseim. A multiscale method for porous microstructures. *Multiscale Model. Simul.*, 14(3):1123–1152, 2016.

[Bre94] S. C. Brenner. Two-level additive Schwarz preconditioners for nonconforming finite elements. *Contemp. Math.*, 180:9–14, 1994.

[Bre03] S. C. Brenner. Poincaré-Friedrichs inequalities for piecewise H^1 functions. *SIAM J. Numer. Anal.*, 41(1):306–324, 2003.

[BS97] I. M. Babuška and S. A. Sauter. Is the pollution effect of the FEM avoidable for the Helmholtz equation considering high wave numbers? *SIAM J. Numer. Anal.*, 34(6):2392–2423, 1997.

[BS08] S. C. Brenner and L. R. Scott. *The mathematical theory of finite element methods.* Springer, New York, third edition, 2008.

[BV16a] D. L. Brown and M. Vasilyeva. A generalized multiscale finite element method for poroelasticity problems I: Linear problems. *J. Comput. Appl. Math.*, 294:372–388, 2016.

[BV16b] D. L. Brown and M. Vasilyeva. A generalized multiscale finite element method for poroelasticity problems II: Nonlinear coupling. *J. Comput. Appl. Math.*, 297:132–146, 2016.

[Cal80] A.-P. Calderón. On an inverse boundary value problem. In: *Seminar on Numerical Analysis and its Applications to Continuum Physics (Rio de Janeiro, 1980)*, pp. 65–73. Soc. Brasil. Mat, Rio de Janeiro, 1980.

[CEL18] E. T. Chung, Y. Efendiev, and W. T. Leung. Constraint energy minimizing generalized multiscale finite element method. *Comput. Methods Appl. Mech. Engrg.*, 339:298–319, 2018.

[CHB09] J. A. Cottrell, T. J. R. Hughes, and Y. Bazilevs. *Isogeometric analysis: toward integration of CAD and FEA.* John Wiley & Sons, Ltd., Chichester, 2009.

[Chr09] S. H. Christiansen. Foundations of finite element methods for wave equations of Maxwell type. In: *Applied wave mathematics*, pp. 335–393. Springer, Berlin, Heidelberg, 2009.

[Cia78] P. G. Ciarlet. *The Finite Element Method for Elliptic Problems.* North-Holland, Amsterdam, 1978.

[Cia88] P. G. Ciarlet. *Mathematical elasticity. Vol. I.* North-Holland, Amsterdam, 1988.

[CM14] A. Caiazzo and J. Mura. Multiscale modeling of weakly compressible elastic materials in the harmonic regime and applications to microscale structure estimation. *Multiscale Model. Simul.*, 12(2):514–537, 2014.

[CMP19] A. Caiazzo, R. Maier, and D. Peterseim. Reconstruction of quasi-local numerical effective models from low-resolution measurements. *WIAS Preprint*, No. 2577, 2019.

[CR72] P. G. Ciarlet and P.-A. Raviart. Interpolation theory over curved elements, with applications to finite element methods. *Comput. Methods Appl. Mech. Engrg.*, 1:217–249, 1972.

[DD76] J. Douglas and T. Dupont. Interior penalty procedures for elliptic and parabolic Galerkin methods. In: *Computing methods in applied sciences*, pp. 207–216. Springer, Berlin, 1976.

[DG75] E. De Giorgi. Sulla convergenza di alcune successioni d'integrali del tipo dell'area. *Rend. Mat. (6)*, 8:277–294, 1975.

[DG84] E. De Giorgi. G-operators and Γ-convergence. In: *Proceedings of the International Congress of Mathematicians (Warsaw, 1983)*, pp. 1175–1191. PWN, Warsaw, 1984.

[DGO16] M. Duerinckx, A. Gloria, and F. Otto. The structure of fluctuations in stochastic homogenization. *ArXiv Preprint*, 1602.01717, 2016.

[Du17] Q. Du. Nonlocal calculus of variations and well-posedness of peridynamics. In: *Handbook of Peridynamic Modeling*, pp. 63–85. CRC Press, Boca Raton, FL, 2017.

[EE03] W. E and B. Engquist. The heterogeneous multiscale methods. *Commun. Math. Sci.*, 1(1):87–132, 2003.

[EE05] W. E and B. Engquist. The heterogeneous multi-scale method for homogenization problems. In: *Multiscale methods in science and engineering*, volume 44 of *Lect. Notes Comput. Sci. Eng.*, pp. 89–110. Springer, Berlin, Heidelberg, 2005.

[EG17] A. Ern and J.-L. Guermond. Finite element quasi-interpolation and best approximation. *ESAIM Math. Model. Numer. Anal.*, 51(4):1367–1385, 2017.

[EGH13] Y. Efendiev, J. Galvis, and T. Y. Hou. Generalized multiscale finite element methods (GMsFEM). *J. Comput. Phys.*, 251:116–135, 2013.

[EGMP13] D. Elfverson, E. H. Georgoulis, A. Målqvist, and D. Peterseim. Convergence of a discontinuous Galerkin multiscale method. *SIAM J. Numer. Anal.*, 51(6):3351–3372, 2013.

[EHMP19] 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.

[EHR11] B. Engquist, H. Holst, and O. Runborg. Multi-scale methods for wave propagation in heterogeneous media. *Commun. Math. Sci.*, 9(1):33–56, 2011.

[EHR12] B. Engquist, H. Holst, and O. Runborg. Multiscale methods for wave propagation in heterogeneous media over long time. In: *Numerical analysis of multiscale computations*, volume 82 of *Lect. Notes Comput. Sci. Eng.*, pp. 167–186. Springer, Berlin, Heidelberg, 2012.

[EM09] A. Ern and S. Meunier. A posteriori error analysis of Euler-Galerkin approximations to coupled elliptic-parabolic problems. *ESAIM Math. Model. Numer. Anal.*, 43(2):353–375, 2009.

[Eva10] L. C. Evans. *Partial Differential Equations*, volume 19 of *Graduate Studies in Mathematics.* American Mathematical Society, Providence, RI, second edition, 2010.

[FAC$^+$19] S. Fu, R. Altmann, E. T. Chung, R. Maier, D. Peterseim, and S.-M. Pun. Computational multiscale methods for linear poroelasticity with high contrast. *J. Comput. Phys.*, 395:286–297, 2019.

[Geo03] E. H. Georgoulis. *Discontinuous Galerkin methods on shape-regular and anisotropic meshes.* Ph.D. thesis, University of Oxford, 2003.

[Geo08] E. H. Georgoulis. Inverse-type estimates on *hp*-finite element spaces and applications. *Math. Comp.*, 77(261):201–219, 2008.

[GGS12] L. Grasedyck, I. Greff, and S. Sauter. The AL basis for the solution of elliptic problems in heterogeneous media. *Multiscale Model. Simul.*, 10(1):245–258, 2012.

[GHS05] I. G. Graham, W. Hackbusch, and S. A. Sauter. Finite elements on degenerate meshes: inverse-type inequalities and applications. *IMA J. Numer. Anal.*, 25(2):379–407, 2005.

[GHV18] D. Gallistl, P. Henning, and B. Verfürth. Numerical homogenization of H(curl)-problems. *SIAM J. Numer. Anal.*, 56(3):1570–1596, 2018.

[GP15] 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.

[GP17] 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.

[HCB05] T. J. R. Hughes, J. A. Cottrell, and Y. Bazilevs. Isogeometric analysis: CAD, finite elements, NURBS, exact geometry and mesh refinement. *Comput. Methods Appl. Mech. Engrg.*, 194(39-41):4135–4195, 2005.

[Hel17] F. Hellman. Gridlod. `https://github.com/fredrikhellman/gridlod`, 2017. GitHub repository, commit 3e9cd20970581a32789aa1e21d7ff3f7e8f0b334.

[HFMQ98] T. J. R. Hughes, G. R. Feijóo, L. Mazzei, and J.-B. Quincy. The variational multiscale method – a paradigm for computational mechanics. *Comput. Methods Appl. Mech. Engrg.*, 166(1-2):3–24, 1998.

[HM17] F. Hellman and A. Målqvist. Contrast independent localization of multiscale problems. *Multiscale Model. Simul.*, 15(4):1325–1355, 2017.

[HMP+19] P. Hennig, R. Maier, D. Peterseim, D. Schillinger, B. Verfürth, and M. Kästner. A diffuse modeling approach for embedded interfaces in linear elasticity. *GAMM-Mitteilungen*, 2019. Online first.

[HP13] P. Henning and D. Peterseim. Oversampling for the multiscale finite element method. *Multiscale Model. Simul.*, 11(4):1149–1175, 2013.

[HP16] P. Henning and A. Persson. A multiscale method for linear elasticity reducing Poisson locking. *Comput. Methods Appl. Mech. Engrg.*, 310:156–171, 2016.

[HS07] T. J. R. Hughes and G. Sangalli. Variational multiscale analysis: the fine-scale Green's function, projection, optimization, localization, and stabilized methods. *SIAM J. Numer. Anal.*, 45(2):539–557, 2007.

[HSS02] P. Houston, C. Schwab, and E. Süli. Discontinuous *hp*-finite element methods for advection-diffusion-reaction problems. *SIAM J. Numer. Anal.*, 39(6):2133–2163, 2002.

[HSW07] P. Houston, D. Schötzau, and T. P. Wihler. Energy norm a posteriori error estimation of *hp*-adaptive discontinuous Galerkin methods for elliptic problems. *Math. Models Methods Appl. Sci.*, 17(1):33–62, 2007.

[HW97] T. Y. Hou and X.-H. Wu. A multiscale finite element method for elliptic problems in composite materials and porous media. *J. Comput. Phys.*, 134(1):169–189, 1997.

[Jol03] P. Joly. Variational methods for time-dependent wave propagation problems. In: *Topics in computational wave propagation*, volume 31 of *Lect. Notes Comput. Sci. Eng.*, pp. 201–264. Springer, Berlin, Heidelberg, 2003.

[KPY18] 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.

[KY16] R. Kornhuber and H. Yserentant. Numerical homogenization of elliptic multiscale problems by subspace decomposition. *Multiscale Model. Simul.*, 14(3):1017–1036, 2016.

[Lip14] R. Lipton. Dynamic brittle fracture as a small horizon limit of peridynamics. *J. Elasticity*, 117(1):21–50, 2014.

[MC16] J. Mura and A. Caiazzo. A two-scale homogenization approach for the estimation of porosity in elastic media. In: *Trends in differential equations and applications*, volume 8 of *SEMA SIMAI Springer Ser.*, pp. 89–105. Springer, Cham, 2016.

[Mel05] J. M. Melenk. *hp*-interpolation of nonsmooth functions and an application to *hp*-a posteriori error estimation. *SIAM J. Numer. Anal.*, 43(1):127–155, 2005.

[MP14] A. Målqvist and D. Peterseim. Localization of elliptic multiscale problems. *Math. Comp.*, 83(290):2583–2603, 2014.

[MP17] A. Målqvist and A. Persson. A generalized finite element method for linear thermoelasticity. *ESAIM Math. Model. Numer. Anal.*, 51(4):1145–1171, 2017.

[MP18] A. Målqvist and A. Persson. Multiscale techniques for parabolic equations. *Numer. Math.*, 138(1):191–217, 2018.

[MP19] R. Maier and D. Peterseim. Explicit computational wave propagation in micro-heterogeneous media. *BIT Numer. Math.*, 59(2):443–462, 2019.

[MPS13] J. M. Melenk, A. Parsania, and S. Sauter. General DG-methods for highly indefinite Helmholtz problems. *J. Sci. Comput.*, 57(3):536–581, 2013.

[MS02] A.-M. Matache and C. Schwab. Two-scale FEM for homogenization problems. *ESAIM Math. Model. Numer. Anal.*, 36(4):537–572, 2002.

[MS10] J. M. Melenk and S. Sauter. Convergence analysis for finite element discretizations of the Helmholtz equation with Dirichlet-to-Neumann boundary conditions. *Math. Comp.*, 79(272):1871–1914, 2010.

[MS11] J. M. Melenk and S. Sauter. Wavenumber explicit convergence analysis for Galerkin discretizations of the Helmholtz equation. *SIAM J. Numer. Anal.*, 49(3):1210–1243, 2011.

[MT97a] F. Murat and L. Tartar. Calculus of variations and homogenization. In: *Topics in the mathematical modelling of composite materials*, volume 31 of *Progr. Nonlinear Differential Equations Appl.*, pp. 139–173. Birkhäuser, Boston, 1997.

[MT97b] F. Murat and L. Tartar. H-convergence. In: *Topics in the mathematical modelling of composite materials*, volume 31 of *Progr. Nonlinear Differential Equations Appl.*, pp. 21–43. Birkhäuser, Boston, 1997.

[Ngu89] G. Nguetseng. A general convergence result for a functional related to the theory of homogenization. *SIAM J. Math. Anal.*, 20(3):608–623, 1989.

[NW06] J. Nocedal and S. J. Wright. *Numerical optimization*. Springer Series in Operations Research and Financial Engineering. Springer, New York, second edition, 2006.

[Osw93] P. Oswald. On a BPX-preconditioner for P1 elements. *Computing*, 51(2):125–133, 1993.

[Owh15] H. Owhadi. Bayesian numerical homogenization. *Multiscale Model. Simul.*, 13(3):812–828, 2015.

[Owh17] H. Owhadi. Multigrid with rough coefficients and multiresolution operator decomposition from hierarchical information games. *SIAM Rev.*, 59(1):99–149, 2017.

[OY19] H. Owhadi and G. R. Yoo. Kernel flows: From learning kernels from data into the abyss. *J. Comput. Phys.*, 389:22–47, 2019.

[OZ08] H. Owhadi and L. Zhang. Numerical homogenization of the acoustic wave equations with a continuum of scales. *Comput. Methods Appl. Mech. Engrg.*, 198(3-4):397–406, 2008.

[OZ17] H. Owhadi and L. Zhang. Gamblets for opening the complexity-bottleneck of implicit schemes for hyperbolic and parabolic ODEs/PDEs with rough coefficients. *J. Comput. Phys.*, 347:99–128, 2017.

[OZB14] H. Owhadi, L. Zhang, and L. Berlyand. Polyharmonic homogenization, rough polyharmonic splines and sparse super-localization. *ESAIM Math. Model. Numer. Anal.*, 48(2):517–552, 2014.

[Pet16] 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.*, pp. 341–367. Springer, Cham, 2016.

[Pet17] D. Peterseim. Eliminating the pollution effect in Helmholtz problems by local subscale correction. *Math. Comp.*, 86(305):1005–1036, 2017.

[PS12] D. Peterseim and S. Sauter. Finite elements for elliptic problems with highly varying, nonperiodic diffusion matrix. *Multiscale Model. Simul.*, 10(3):665–695, 2012.

[PS16] D. Peterseim and R. Scheichl. Robust numerical upscaling of elliptic multiscale problems at high contrast. *Comput. Meth. Appl. Mat.*, 16(4):579–603, 2016.

[PS17] D. Peterseim and M. Schedensack. Relaxing the CFL condition for the wave equation on adaptive meshes. *J. Sci. Comput.*, 72(3):1196–1213, 2017.

[Sch98] C. Schwab. *p- and hp-finite element methods. Theory and applications in solid and fluid mechanics.* Numerical Mathematics and Scientific Computation. The Clarendon Press, Oxford University Press, New York, 1998.

[Sho00] R. E. Showalter. Diffusion in poro-elastic media. *J. Math. Anal. Appl.*, 251(1):310–340, 2000.

[Sil00] S. A. Silling. Reformulation of elasticity theory for discontinuities and long-range forces. *J. Mech. Phys. Solids*, 48(1):175–209, 2000.

[Spa68] S. Spagnolo. Sulla convergenza di soluzioni di equazioni paraboliche ed ellittiche. *Ann. Scuola Norm. Sup. Pisa (3)*, 22:571–597, 1968.

[Szy06] D. B. Szyld. The many proofs of an identity on the norm of oblique projections. *Numer. Algor.*, 42(3-4):309–323, 2006.

[Tar78] L. Tartar. Quelques remarques sur l'homogénéisation. In: H. Fujita (ed.), *Functional Analysis and Numerical Analysis, Proceedings of the Japan-France Seminar (1976)*, pp. 469–482. Japan Society for the Promotion of Science, Tokyo, 1978.

[Ver17] B. Verfürth. Numerical homogenization for indefinite H(curl)-problems. In: K. Mikula, D. Sevcovic, and J. Urban (eds.), *Proceedings of Equadiff 2017 conference*, pp. 137–146. Slovak University of Technology, Bratislava, 2017.

[Wey16] M. Weymuth. *Adaptive local basis for elliptic problems with L^∞-coefficients.* Ph.D. thesis, University of Zurich, 2016.

[XZ03] J. Xu and L. Zikatanov. Some observations on Babuška and Brezzi theories. *Numer. Math.*, 94(1):195–202, 2003.

[Zla73] M. Zlámal. Curved elements in the finite element method. I. *SIAM J. Numer. Anal.*, 10:229–240, 1973.

[Zob10] M. D. Zoback. *Reservoir Geomechanics.* Cambridge University Press, New York, 2010.