Computational Multiscale Methods for Elliptic Partial Differential Equations

Habilitationsschrift zur Erlangung der Lehrbefähigung
für das Fach Mathematik

vorgelegt dem Rat der Mathematisch-Naturwissenschaftlichen
Fakultät der Humboldt-Universität zu Berlin
von Herrn Daniel Peterseim
geboren am 22.05.1980 in Mühlhausen/Thüringen

Prof. Dr. Jan-Hendrik Olbertz
Präsident der Humboldt-Universität
zu Berlin

Prof. Dr. Elmar Kulke
Dekan der Mathematisch-
Naturwissenschaftlichen Fakultät

Berlin, den 20.01.2016

Gutachterinnen und Gutachter:
1. Prof. Dr. Carsten Carstensen
2. Prof. Dr. Harry Yserentant
3. Prof. Björn Engquist

Preface

This thesis is a cumulative habilitation thesis, that is, a collection of ten selected research articles. Nine of them have already appeared in peer-reviewed scientific journals and one is submitted for publication. These papers reflect a central part of my research in the period 2010–2015 that can be summarized under the caption *Computational Multiscale Methods*. The following list provides the bibliographic information of the papers in the order they are included in the Appendices A–D of this thesis.

[A1] A. Målqvist and D. Peterseim. Localization of elliptic multiscale problems. *Mathematics Computation*, 83(290): 2583–2603, 2014.

[A2] P. Henning and D. Peterseim. Oversampling for the multiscale finite element method. *Multiscale Modeling & Simulation*, 11(4):1149–1175, 2013.

[A3] P. Henning, A. Målqvist, and D. Peterseim. A localized orthogonal decomposition method for semi-linear elliptic problems. *ESAIM: Mathematical Modelling and Numerical Analysis*, 48:1331–1349, 2014.

[B1] D. Peterseim. Eliminating the pollution effect in Helmholtz problems by local subscale correction. *ArXiv e-prints*, 1411.1944, 2014.

[B2] D. Gallistl and D. Peterseim. Stable multiscale Petrov-Galerkin finite element method for high frequency acoustic scattering. *Computer Methods in Applied Mechanics and Engineering*, 295:1–17, 2015.

[C1] A. Målqvist and D. Peterseim. Computation of eigenvalues by numerical upscaling. *Numerische Mathematik*, 130(2):337–361, 2015.

[C2] P. Henning, A. Målqvist, and D. Peterseim. Two-level discretization techniques for ground state computations of Bose-Einstein condensates. *SIAM Journal on Numerical Analysis*, 52(4):1525–1550, 2014.

[D1] D. Peterseim and C. Carstensen. Finite element network approximation of conductivity in particle composites. *Numerische Mathematik*, 124(1):73-97, 2013.

[D2] D. Peterseim. Robustness of finite element simulations in densely packed random particle composites. *Networks and Heterogeneous Media*, 7(1):113126, 2012.

[D3] D. Peterseim. Composite finite elements for elliptic interface problems. *Mathematics of Computation*, 83(290):2657-2674, 2014.

The collection of papers is preceded by four introductory chapters that provide an overview of the contributions of the above papers and the following closely related work of mine in the context of multiscale problems and their numerical simulation.

[9] D. L. Brown and D. Peterseim. A multiscale method for porous microstructures. *ArXiv e-prints*, November 2014.

[23] D. Elfverson, E. H. Georgoulis, A. Målqvist, and D. Peterseim. Convergence of a discontinuous Galerkin multiscale method. *SIAM Journal on Numerical Analysis*, 51(6):3351–3372, 2013.

[25] M. Eigel and D. Peterseim, Simulation of composite materials by a network FEM with error control. *Computational Methods in Applied Mathematics*, 15(1):21–37, 2015.

[37] P. Henning, P. Morgenstern, and D. Peterseim. Multiscale partition of unity. In Michael Griebel and Marc Alexander Schweitzer (Eds.), *Meshfree Methods for Partial Differential Equations VII*, volume 100 of *Lecture Notes in Computational Science and Engineering*, pages 185–204, Springer International Publishing, 2015.

[49] A. Målqvist and D. Peterseim. Generalized finite element methods for quadratic eigenvalue problems. *ArXiv e-prints*, 1510.05792, 2015.

[55] D. Peterseim. Variational multiscale stabilization and the exponential decay of fine-scale correctors. *ArXiv e-prints*, 1505.07611, 2015. (to appear in G.R. Barrenechea, F. Brezzi, A. Cangiani and E. Georgoulis (Eds.), *Building Bridges: Connections and Challenges in Modern Approaches to Numerical Partial Differential Equations*, Lecture Notes in Computational Science and Engineering, Springer)

[56] D. Peterseim and S. Sauter. Finite elements for elliptic problems with highly varying, nonperiodic diffusion matrix. *Multiscale Modeling & Simulation*, 10(3):665-695, 2012.

[57] D. Peterseim and R. Scheichl. Rigorous numerical upscaling at high contrast. *in preparation*, 2015+.

The four introductory chapters of this thesis coincide to a large extend with the following survey paper [55]. Although these chapters serve as a general introduction, they are not meant to simply summarize the research papers [A1]–[D3] but to put some of their central results in a new unifying perspective and, thereby, to add value to the collection of articles [A1]–[D3].

Acknowledgments. Major parts of this thesis were written when I was a group leader within the DFG Research Center Matheon in Berlin in 2010–2013. The support given by the center through the project *C33 Modeling and Simulation of Composite Materials* and the support of the hosting institution – the Institute of Mathematics at Humboldt-Universität zu Berlin – are gratefully acknowledged. Since 2013, the work was supported by the Institute of Numerical Simulation of the University of Bonn – my current affiliation – and the Hausdorff Center for Mathematics Bonn. Moreover, the funding by Deutsche Forschungsgemeinschaft in the Priority Program 1748 "Reliable simulation techniques in solid mechanics: Development of non-standard discretization methods, mechanical and mathematical analysis" under the project *Adaptive isogeometric modeling of propagating strong discontinuities in heterogeneous materials* is gratefully acknowledged.

This thesis does not only reflect my own work but is the direct or indirect result of many fruitful collaborations in the past five years. I would like to thank my co-authors Donald Brown, Martin Eigel, Daniel Elfverson, Manolis Georgoulis, Philipp Morgenstern, Stefan Sauter, Mira Schedensack and, in particular, Axel Målqvist, Patrick Henning and Dietmar Gallistl for their valuable impact. Last, but not least, I would like to thank my mentor and co-author Carsten Carstensen for his overwhelming support.

Bonn, November 2015 *Daniel Peterseim*

Contents

Chapter 1
Introduction

This thesis concerns the algorithms and mathematics that underlie the computer-aided simulation of complex processes in engineering and the sciences. Such processes are characterized by a large range of non-separable scales and we will refer to them as multiscale problems. Among the target applications are the mechanical analysis of multiphase materials such as composite and multifunctional materials, transport processes in porous media, high-frequency acoustic scattering as well as the simulation of Bose-Einstein condensates. Although mathematical physics provides sound models of partial differential equations that implicitly describe these processes, the complex interplay of effects between the scales is intractable for an analytical solution such that their understanding and control relies on numerical simulation. In many interesting applications, computers are not able to resolve all details on all relevant scales. The observation and prediction of physical phenomena from multiscale models, hence, requires insightful numerical techniques that efficiently represent unresolved scales in a numerical simulation, i.e., *computational multiscale methods*.

The design and analysis of these methods requires novel mathematical tools. In the past decades, the numerical analysis of partial differential equations (PDEs) was merely focused on the numerical approximation of sufficiently smooth solutions in the asymptotic regime of convergence. In the context of multiscale problems (and beyond), such results have only limited impact because the numerical approximation will hardly ever reach the asymptotic idealized regime under realistic conditions. Although a method performs well for sufficiently fine meshes it may fail completely on coarser (and feasible) scales of discretization. This can be seen for instance in the numerical homogenization of elliptic boundary value problems with highly varying non-smooth diffusion coefficient or high-frequency time-harmonic acoustic wave propagation, where the corresponding PDEs exhibits rough and highly oscillatory solutions.

The difficulties for the numerical approximation of such oscillatory problems by finite element methods (FEMs) or related schemes are two-fold. The pure approximation (e.g. interpolation) of the unknown solutions by finite elements already requires high spatial resolution to capture fast oscillations and heterogeneities on microscopic scales. When the function is described only implicitly as the solution of some partial differential equation, its approximation faces further scale-dependent pre-asymptotic effects caused by the under-resolution of relevant microscopic data. Examples are the poor L^2 approximation in homogenization problems (see Fig. 1.1) and the pollution effect [10] for Helmholtz problems with large wave numbers (see Fig. 1.2). We shall emphasize that, in the latter case, the existence and uniqueness of numerical approximations may not even be guaranteed in pre-asymptotic regimes.

Such situations require the stabilization of standard methods so that eventually a meaningful approximation on reasonably coarse scales of discretization becomes feasible. This introductory part of this thesis presents a general multiscale framework for the stabilization of FEMs for multiscale

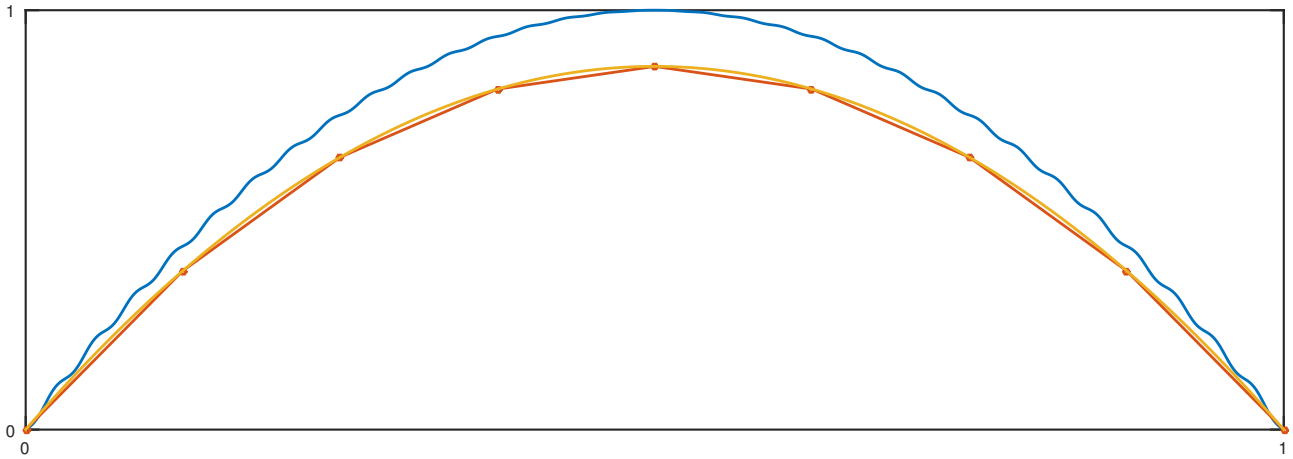


Fig. 1.1 Failure of FEM in homogenization problems: Consider the periodic problem $-\frac{d}{dx}A_\varepsilon(x)\frac{d}{dx}u_\varepsilon(x) = 1$ in the unit interval with homogeneous Dirichlet boundary condition, where $A_\varepsilon(x) := (2+\cos(2\pi x/\varepsilon))^{-1}$ for some small parameter $\varepsilon > 0$. The solution $u_\varepsilon = 4(x-x^2) - 4\varepsilon\left(\frac{1}{4\pi}\sin(2\pi\frac{x}{\varepsilon}) - \frac{1}{2\pi}x\sin(2\pi\frac{x}{\varepsilon}) - \frac{\varepsilon}{4\pi^2}\cos(2\pi\frac{x}{\varepsilon}) + \frac{\varepsilon}{4\pi^2}\right)$ is depicted in blue for $\varepsilon = 2^{-5}$. The P1-FE approximation (○) on a uniform mesh of width h interpolates the curve $x \mapsto 2\sqrt{3}(x-x^2)$ whenever h is some multiple of the characteristic length scale ε and, hence, fails to approximate u_ε in any reasonable norm in the regime $h \geq \varepsilon$.

problems with the aim to significantly reduce or even eliminate pre-asymptotic effects due to under-resolution. Our starting point will be the Variational Multiscale Method (VMS) originally introduced in [38, 39]. The method provides an abstract framework how to incorporate missing fine-scale effects into numerical problems governing coarse-scale behavior [40]. One may interpret the VMS as a Petrov-Galerkin method using standard FE trial spaces and an operator-dependent test space that needs to be precomputed in general.

The construction of this operator-dependent test space is based on some stable projection onto the standard finite element (FE) trial space and a corresponding scale decomposition of a function into its FE part given by the projection (the macroscopic/coarse-scale part) and a remainder that lies in the kernel of the projection operator (the microscopic/fine-scale part). The test functions are computed via a problem-dependent projection of the trial space into the space of fine-scale functions. This requires the solution of variational problems in the kernel of the projection – the fine-scale corrector problems. It has been observed empirically in certain applications that the Green's function associated with these fine-scale corrector problems – the so-called fine-scale Green's function [39] – may exhibit favorable exponential decay properties [44, 39] even though the decay of the classical full scale Green's function is only algebraic. It is this exponential decay property that allows one to turn the VMS into a feasible numerical method [44, 42].

The exponential decay was rigorously proved for the first time in [A1] in the context of multi-dimensional numerical homogenization. A key ingredient of the proof of [A1] is the use of a (local) quasi-interpolation operator for the scale decomposition. Although the method of [A1] still fits into the general framework of the VMS, it uses a different point of view on the method based on the orthogonalisation of coarse and fine scales with respect to the inner product associated with a symmetric and coercive model problem. This is why the method of [A1] is now referred to as the Localized Orthogonal Decomposition (LOD) method. Subsequent work showed that the ideas of [A1] can be generalized to other discretization techniques such as discontinuous Galerkin [22, 23, 26], Petrov-Galerkin formulations [21], mixed methods [34] and mesh-free methods [37]. Moreover, the method

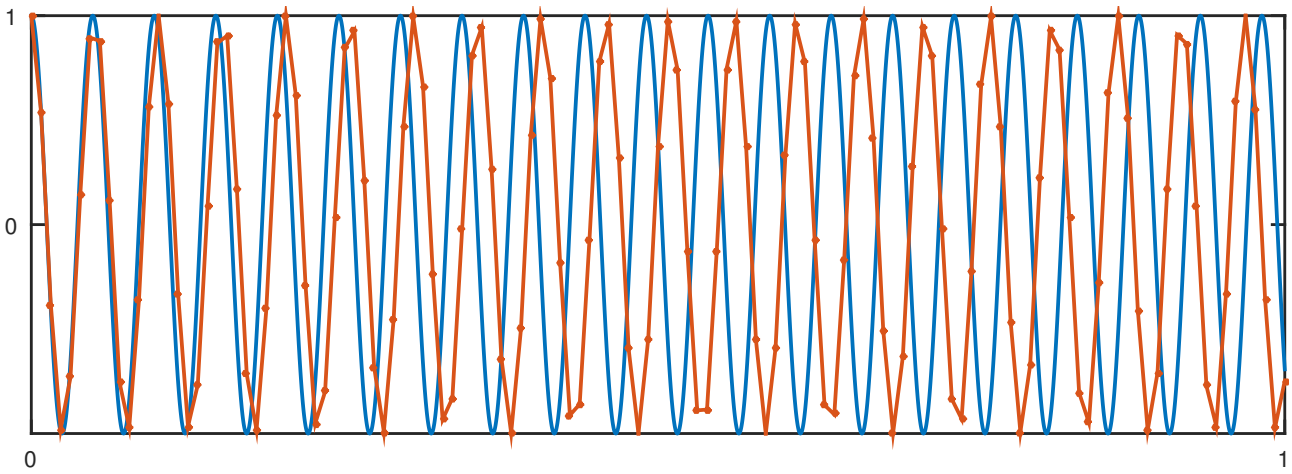


Fig. 1.2 Numerical dispersion in Helmholtz problems: Consider $-\frac{d^2}{dx^2}u_\varepsilon(x) - \kappa^2 u(x) = 0$ in the unit interval with $u(0) = 1$ and $\frac{d}{dx}u(1) = -i\kappa u(1)$ for some large parameter $\kappa > 0$. The solution $u_\kappa = \exp(-i\kappa x)$ is depicted in blue for $\kappa = 2^7$. The P1-FE approximation (○) on a uniform mesh of width $h = 2^{-7} > 6 \cdot (\text{wave length})$ fails to approximate u_κ due to the accumulation of phase errors.

can also be reinterpreted in terms of the multiscale finite element method with special oversampling [A2]. The class of problems that have been analyzed by now includes semi-linear problems [A3], high-contrast problems [57, 9], rough boundary conditions [35], problems on complicated geometries [27], linear and non-linear eigenvalue problems [C1, C2, 49], parabolic problems [48], wave propagation [3, B1, B2] and parametric problems [4].

Chapter 2 of this thesis aims to reinterpret all those results, in particular [A1]–[B2], in the abstract stabilization framework of the original VMS (see Section 2.1.1) and aims to illustrate how the exponential decay of the fine-scale Green's function can be quantified (see Section 2.2). Chapter 3 then shows how these abstract results lead to super-localized numerical homogenization [A1, A2] (see Section 3.1) and pollution-free time-harmonic acoustic scattering (see Section 3.2) [B1, B2]. Chapter 4 closes the introductory part of this thesis with a brief description of further applications including (non-linear) eigenvalue problems as they are treated in [C1, C2] as well as the challenge of high contrast (e.g. in underlying material coefficients) which is the topic of [D1, D2, D3].

Chapter 2
An abstract multiscale method

This chapter is concerned with an abstract variational problem in a complex Hilbert space V as it appears for the weak formulation of second order PDEs. In this context, V is typically some closed subspace of the Sobolev space $\mathrm{H}^1(\Omega;\mathbb{C}^m)$ for some bounded Lipschitz domain $\Omega \subset \mathbb{R}^d$. Let a denote a bounded sesquilinear form on $V \times V$ and let $F \in V'$ denote a bounded linear functional on V. We wish to find $u \in V$ satisfying the linear variational problem

$$\forall v \in V : \; a(u,v) = \overline{F(v)}. \tag{2.1}$$

We assume that the sesquilinear form a satisfies the inf-sup condition

$$\alpha := \inf_{0\neq v\in V} \sup_{0\neq w\in V} \frac{a(v,w)}{\|v\|_V \|w\|_V} = \inf_{0\neq w\in V} \sup_{0\neq v\in V} \frac{a(v,w)}{\|v\|_V \|w\|_V} > 0. \tag{2.2}$$

Under this condition, the abstract problem (2.1) is well-posed, i.e., for all $F \in V'$ there exists a unique solution $u \in V$ and the a priori bound

$$\|u\|_V \leq \alpha^{-1} \|F\|_{V'}$$

holds true; see, e.g., [5].

2.1 Finite element projections

We wish to approximate the unknown solution u of (2.1) by some computable function. The standard procedure for approximation is the Galerkin method which simply chooses a finite-dimensional subspace $V_H \subset V$ (that contains simple functions such as piecewise polynomials) and restricts the variational problem (2.1) to this subspace. Usually, V_H belongs to some family of spaces parametrized by some abstract discretization parameter H, for instance the mesh size. This parameter (or set of parameters) provides some control on the approximation properties of V_H as $H \to 0$ at the price of an increasing computational cost in the sense of $\dim V_H \to \infty$. The Galerkin method seeks a function $G_H u \in V_H$ satisfying

$$\forall v_H \in V_H : \; a(G_H u, v_H) = \overline{F(v_H)} \;\; \big(= a(u, v_H) \big). \tag{2.3}$$

Recall that the well-posedness of the original problem (2.1) does not imply the well-posedness of the discrete variational problem (2.3) but needs to be checked for the particular application via dis-

crete versions of the inf-sup condition (2.2). In many cases, such conditions are only satisfied for H sufficiently small. This means that there is some threshold complexity for computing any Galerkin approximation and this threshold can be out of reach. Even if a Galerkin solution $G_H u$ exists and is computable, it might not provide the desired accuracy or does not reflect the relevant characteristic features of the solution, as we have seen in the introduction.

Therefore, we are interested in computing projections onto the discrete space V_H other than the Galerkin projection. Let $I_H : V \to V_H$ denote such a linear surjective projection operator and let us assume that it is bounded in the sense of the space $\mathscr{L}(V)$ of linear operators from V to V with finite operator norm

$$\|I_H\|_{\mathscr{L}(V)} := \sup_{0 \neq v \in V} \frac{\|I_H v\|_V}{\|v\|_V} < \infty.$$

Implicitly, we also assume that this operator norm does not depend on the discretization parameter H in a critical way. Possible choices of I_H include the orthogonal projection onto V_H with respect to the inner product of V or any Hilbert spaces $L \supset V$ containing V and mainly (local) quasi-interpolation operators of Clément or Scott-Zhang type as they are well-established in the finite element community in the context of a posteriori error estimation [15, 59, 11, 17].

2.1.1 Petrov-Galerkin characterization of finite element projections

The Galerkin projection G_H is designed in such a way that its computation requires only the known data F associated with the unknown solution u of the abstract variation problem (2.1). This section mimics this property for a general projection $I_H \in \mathscr{L}(V)$ by characterizing it as a Petrov-Galerkin discretization using V_H as the trial space and a non-standard test space $W_H \subset V$ that depends on the problem and the projection. The definition of W_H rests on the trivial observation that, for any $v \in V$,

$$a(I_H u, v) = \overline{F(v)} - a(u - I_H u, v). \tag{2.4}$$

The choice of a test function v in the subspace

$$W_H := \{w \in V \mid \forall z \in \operatorname{Ker} I_H : a(z, w) = 0\} \tag{2.5}$$

annihilates the second term on the right-hand side of (2.4) and, hence,

$$a(I_H u, w_H) = \overline{F(w_H)}$$

holds for all $w_H \in W_H$. This shows that $I_H u$ is a solution of the Petrov-Galerkin method: Find $u_H \in V_H$ such that

$$\forall w_H \in W_H : \ a(u_H, w_H) = \overline{F(w_H)}. \tag{2.6}$$

This characterization of I_H is known from the variational multiscale method as it is presented in [40] and it is the basis of the results in Appendices A and B of this thesis, especially [A1].

The question whether or not (2.6) has a unique solution can not be answered under the general assumptions made so far. We need to assume the missing uniqueness to be able to proceed and one way of doing this is to assume that the dimensions of trial and test space are equal,

$$\dim W_H = \dim V_H. \tag{2.7}$$

In the present setting with a bounded operator I_H, this condition is equivalent to the well-posedness of the discrete variational problem (2.6), i.e., it admits a unique solution $u_H = I_H u \in V_H$ and

$$\|u_H\|_V \leq \|I_H\|_{\mathscr{L}(V)} \|u\|_V \leq \frac{\|I_H\|_{\mathscr{L}(V)}}{\alpha} \|F\|_{V'}.$$

The a priori estimate in turn implies a lower bound of the discrete inf-sup constant of the Petrov-Galerkin method by the quotient of the continuous inf-sup constant α and the continuity constant of I_H,

$$\inf_{0\neq v_H\in V_H} \sup_{0\neq w_H\in W_H} \frac{a(v_H,w_H)}{\|v_H\|_V \|w_H\|_V} \geq \frac{\alpha}{\|I_H\|_{\mathscr{L}(V)}} \leq \inf_{0\neq w_H\in W_H} \sup_{0\neq v_H\in V_H} \frac{a(v_H,w_H)}{\|v_H\|_V \|w_H\|_V}.$$

The test space W_H is the ideal test space for our purposes in the following sense. Assuming that we have access to it, the method (2.6) would enable us to compute $I_H u$ without the explicit knowledge of u. Although this will rarely be the case, we will see later that W_H can be approximated very efficiently in relevant cases. The discrete inf-sup conditions then indicate that the sufficiently accurate approximation of W_H will not harm the method, its stability properties or its subsequent error minimization properties.

The continuity of the projection operator I_H readily implies the quasi-optimality of the Petrov-Galerkin method (2.6),

$$\|u - u_H\|_V = \|(1 - I_H)u\|_V \leq \|I_H\|_{\mathscr{L}(V)} \min_{v_H\in V_H} \|u - v_H\|_V. \tag{2.8}$$

Here, we have used that $\|I_H\|_{\mathscr{L}(V)} = \|1 - I_H\|_{\mathscr{L}(V)}$; see e.g. [60]. More importantly, the same arguments show that the Petrov-Galerkin method is quasi-optimal with respect to any other Hilbert space $L \supset V$ with norm $\|\cdot\|_L$ whenever $I_H \in \mathscr{L}(L)$,

$$\|u - u_H\|_L \leq \|I_H\|_{\mathscr{L}(L)} \min_{v_H\in V_H} \|u - v_H\|_L.$$

This quasi-optimality makes the ansatz very appealing and justifies its further investigation. Hence, in the remaining part of the chapter, it is our aim to turn the method into a feasible numerical scheme while preserving these properties to a large extent. Although the discrete stability of the method depends on the stability properties of the original problem and, hence, on parameters such as the frequency in scattering problems, the quasi-optimality depends only on I_H and not necessarily on the problem.

2.1.2 Characterization of the ideal test space

A practical realization of the Petrov-Galerkin method (2.6) requires a choice of bases in the discrete trial V_H and test space W_H. As usual, these choices have big impact on the computational complexity. The underlying principle of finite elements is the locality of the bases which yields sparse linear systems and offers the possibility of linear computational complexity with respect to the number of degrees of freedom $N_H = \dim V_H$. Let

$$\{\lambda_j \mid j = 1,2,\ldots,N_H\}$$

be such a local basis of V_H.

We shall derive a basis of the test space W_H defined in (2.5) by mapping the trial basis onto a test basis via some bijective operator $\mathscr{T}$, a so-called trial-to-test operator. Due to Assumption (2.7) such an operator exists, but there are many choices and we have to make a design decision. Our choice is that

$$I_H \circ \mathscr{T} = id \tag{2.9}$$

which is consistent with almost all existing practical realizations of the method but one might as well consider distance minimization

$$\|(1-\mathscr{T})v_H\|_V = \min_{w_H \in W_H} \|v_H - w_H\|_V.$$

The condition (2.9) fixes the (macroscopic) finite element part $I_H \mathscr{T} v_H = v_H$ of $\mathscr{T} v_H$ while the fine scale remainder $(1-I_H)\mathscr{T} v_H$ is determined by the variational condition in the definition of W_H. Given $v_H \in V_H$, $(1-I_H)\mathscr{T} v_H \in \operatorname{Ker} I_H$ satisfies

$$\forall z \in \operatorname{Ker} I_H : a(z,(1-I_H)\mathscr{T} v_H) = -a(z,v_H). \tag{2.10}$$

This variational problem is referred to as the fine scale corrector problem for $v_H \in V_H$. Note that v_H can be replaced with any $v \in V$ so that $(1-I_H)\mathscr{T}$ can be understood as an operator from V into $\operatorname{Ker} I_H$. We usually denote this operator the fine scale correction operator and write $\mathscr{C} := (1-I_H)\mathscr{T}$. This operator is the Galerkin projection from V_H (or V) into $\operatorname{Ker} I_H$ related to the adjoint of the sesquilinear form a. It depends on the underlying variational problem and equips test functions with problem related features that are not present in V_H. In the context of elliptic PDEs, $\mathscr{C}$ is called the fine-scale Green's operator [38, 40].

For this construction to work we need to assume the well-posedness of the corrector problem (2.10), i.e., there is some constant $\beta > 0$ such that

$$\inf_{0\neq v\in \operatorname{Ker} I_H} \sup_{0\neq w\in \operatorname{Ker} I_H} \frac{a(v,w)}{\|v\|_V \|w\|_V} \geq \beta \leq \inf_{0\neq w\in \operatorname{Ker} I_H} \sup_{0\neq v\in \operatorname{Ker} I_H} \frac{a(v,w)}{\|v\|_V \|w\|_V}. \tag{2.11}$$

As for the Galerkin projection G_H onto V_H, these inf-sup conditions do not follow from their continuous counterparts (2.2) (unless a is coercive) and they might hold for sufficiently small H only. However, we were able to show in the context of the Helmholtz model problem of Section 3.2 that (2.11) holds in a much larger regime of the discretization parameter H than the corresponding conditions for the standard FEM do. In any case, condition (2.11) implies that the trial-to-test operator $\mathscr{T} = 1 + \mathscr{C}$ is a bounded linear projection operator from V to W_H with operator norm

$$\|\mathscr{T}\|_{\mathscr{L}(V)} = \|1-\mathscr{T}\|_{\mathscr{L}(V)} = \|\mathscr{C}\|_{\mathscr{L}(V)} \leq \frac{C_a}{\beta},$$

where C_a denotes the continuity constant of the sesquilinear form a. Moreover, $\mathscr{T}|_{V_H} : V_H \to W_H$ is invertible with $(\mathscr{T}|_{V_H})^{-1} = I_H$ and

$$\{\mathscr{T}\lambda_j \mid j = 1,2,\ldots,N_H\}$$

defines a basis of W_H with

$$\frac{1}{\|I_H\|_{\mathscr{L}(V)}}\|\lambda_j\|_V \leq \|\mathscr{T}\lambda_j\|_V \leq \tfrac{C_a}{\beta}\|\lambda_j\|_V, \quad 1 \leq j \leq N_H.$$

In general, it cannot be expected (apart from one-dimensional exceptions where $\operatorname{Ker} I_H$ is a broken Sobolev space [40]) that the $\mathscr{T}\lambda_j$ have local support. On the contrary, their support will usually be global. However, we will show in the next chapter that they decay very fast in relevant applications; for illustrations see Figures 3.1 and 3.6.

An important special case of the model problem (2.1) is the hermitian case. Note that hermiticity is preserved by the Petrov-Galerkin method in the following sense. For any $u_H, v_H \in V_H$, it holds that

$$a(u_H, \mathscr{T}v_H) = a(\mathscr{T}u_H, \mathscr{T}v_H) = \overline{a(\mathscr{T}v_H, \mathscr{T}u_H)} = \overline{a(v_H, \mathscr{T}u_H)}.$$

However, this hermiticity is typically lost once $\mathscr{T}$ is replaced with some approximation $\mathscr{T}_\ell$. In order to avoid a lack of hermiticity, the papers [A1]–[C2] use a variant of the method with W_H as the test and trial space. If hermiticity is important, one should follow this line. In this thesis, we trade hermiticity for a cheaper method that avoids any costly communication between the fine-scale correctors that is necessary in the hermitian version.

If the problem is non-hermitian, one might still consider a modified trial space based on the adjoint of $\mathscr{T}$ to improve approximation properties; see [43, 45, B1] for details. In a setting with a modified trial space, further generalizations are possible. Since V_H does not appear any more in the method, its conformity can be relaxed as it was recently proposed in [53] in the context of a multilevel solver for Poisson-type problems with L^∞ coefficients. This approach enables one to compute very general quantities of the solution such as piecewise mean values or higher moments related to elements or edges and can be linked to discontinuous Galerkin or Crouzeix-Raviart type approximations instead of conforming finite elements.

2.2 Exponential decay of fine-scale correctors

In many cases, the fine-scale correctors (i.e. the solutions of the fine-scale corrector problems (2.10)) have decay properties better than those of the Green's function associated with the underlying full-scale partial differential operator. To elaborate on this, we shall now assume that the space V is a closed subspace of $\mathrm{H}^1(\Omega)$ with a local norm (the notation $\|\cdot\|_{V,\omega}$ means that the V-norm is restricted to some subdomain $\omega \subset \Omega$). Moreover, the sesquilinear form a is assumed to be local. This is the natural setting for scalar second order PDEs. The subsequent arguments can be easily generalized to vector-valued problems.

To be more precise regarding the locality of the basis mentioned above, we shall associate the basis functions of V_H with a set of geometric entities $\mathscr{N}_H$ called nodes (e.g. the vertices of a triangulation) and assume that these nodes are well distributed in the domain Ω in the sense of local quasi-uniformity. In this context, H refers to the maximal distance between nearest neighbors (the mesh size). Given some node $z \in \mathscr{N}_H$ and the corresponding basis function $\lambda_z \in V_H$, set the corrector $\phi_z = \mathscr{C}\lambda_z$ and recall from (2.10) that

$$a(w, \phi_z) = -a(w, \lambda_z)$$

for all $w \in \operatorname{Ker} I_H$.

We aim to show that there are constants $c > 0$ and $C > 0$ independent of H and R such that

$$\|\mathscr{C}\lambda_z\|_{V,\Omega\setminus B_R(z)} = \|\phi_z\|_{V,\Omega\setminus B_R(z)} \leq C\exp\left(-c\frac{R}{H}\right)\|\mathscr{C}\lambda_z\|_V, \tag{2.12}$$

where $B_R(z)$ denotes the ball of radius $R>0$ centered at z. The proof of (2.12) is perhaps the most important contribution of this thesis and was first given in [A1].

We shall demonstrate how this result can be established and what kind of assumptions have to be made. Let $R>2H$ and $r:=R-H>H$ and let $\eta\in W^{1,\infty}(\Omega;[0,1])$ be some cut-off function with $\eta=0$ in $\Omega\setminus B_R(z)$, $\eta=1$ in $B_r(z)$, and

$$\|\nabla\eta\|_{L^\infty(\Omega)} \leq C_\eta H^{-1} \tag{2.13}$$

for some generic constant C_η. In general, the fine-scale space $\operatorname{Ker} I_H$ is not closed under multiplication by a cut-off function and we will need to project the truncated function $\eta\phi_z$ back into $\operatorname{Ker} I_H$ by the operator $1-I_H$. We assume that the concatenation of multiplication by η and $(1-I_H)$ is stable and quasi-local in the sense that

$$\forall w\in\operatorname{Ker} I_H:\ \|(1-I_H)(\eta w)\|_{V,B_R(z)\setminus B_r(z)} \leq C_{\eta,I_H}\|w\|_{V,B_{R'}(z)\setminus B_{r'(z)}} \tag{2.14}$$

holds with $r':=r-mH$ and $R':=R+mH$ and generic constants $C_{\eta,I_H}>0$ and $m\in\mathbb{N}_0$ independent of H and z. Although the multiplication by η is not a stable operation in the full space V (think of a constant function), this result is possible in the space of fine scales for example if I_H enjoys quasi-local stability and approximation properties; see Section 3.1 below for an example. The quasi-locality of I_H is also used in the next argument.

Assuming that the inf-sup condition (2.11) holds, the corrector ϕ_z satisfies

$$\begin{aligned}\|\phi_z\|_{V,\Omega\setminus B_R(z)} &= \|(1-I_H)\phi_z\|_{V,\Omega\setminus B_R(z)}\\ &\leq \|(1-I_H)((1-\eta)\phi_z)\|_V\\ &\leq \beta^{-1}a(w,(1-I_H)((1-\eta)\phi_z))\\ &= \beta^{-1}\left(a(w,\phi_z)-a(w,(1-I_H)(\eta\phi_z))\right)\end{aligned}$$

for some $w\in\operatorname{Ker} I_H$ with $\|w\|_V=1$. Since $\operatorname{supp}((1-I_H)((1-\eta)\phi_z))\subset\Omega\setminus B_r(z)$ there is a good chance to actually find a function w with

$$\operatorname{supp} w\subset\operatorname{supp}((1-I_H)((1-\eta)\phi_z))\subset\Omega\setminus B_r(z).$$

Of course, this is an assumption that needs to be verified in the particular application. Under this condition, the term $a(w,\phi_z)=a(w,\lambda_z)$ vanishes because the supports of w and λ_z have no overlap. This and (2.14) imply

$$\begin{aligned}\|\phi_z\|_{V,\Omega\setminus B_R(z)} &\leq \beta^{-1}C_aC_{\eta,I_H}\|\phi_z\|_{V,B_{R'}(z)\setminus B_{r'}(z)}\\ &= \beta^{-1}C_aC_{\eta,I_H}\left(\|\phi_z\|^2_{V,\Omega\setminus B_{r'}(z)}-\|\phi_z\|^2_{V,\Omega\setminus B_{R'}(z)}\right)^{1/2},\end{aligned}$$

where C_a denotes the continuity constant of the sesquilinear form a. Hence, the contraction

$$\|\phi_z\|^2_{V,\Omega\setminus B_{R'}(z)} \leq \frac{C'}{1+C'}\|\phi_z\|^2_{V,\Omega\setminus B_{R'-(2m+1)H}(z)}$$

holds with $C' := (\beta^{-1} C_a C_{\eta, I_H})^2$. The iterative application of this estimate with $R' \mapsto R' - (2m+1)H$ plus relabeling $R' \mapsto R$ leads to the conjectured decay result (2.12) with constants $C := (\frac{C'}{1+C'})^{-\frac{1}{2(2m+1)}}$ and $c := \left|\log(\frac{C'}{1+C'})\right| \frac{(1)}{2(2m+1)} > 0$.

The exponential decay of the ideal correctors motivates and justifies the localization of the fine-scale corrector problems to local subdomains of diameter ℓH where $\ell \in \mathbb{N}$ is a new discretization parameter, the so-called oversampling parameter. It controls the perturbation with respect to the ideal global correctors. We will explain this localization procedure on the basis of an example in Section 3.1 below. As a rule of thumb, the localization to subdomains of diameter ℓH will introduce an error of order $\mathscr{O}(\exp(-\ell))$. As long as this error is small when compared with the inf-sup constant $\alpha \|I_H\|^{-1}_{\mathscr{L}(V)}$ of the ideal method, the stability and approximation properties of the method will be largely preserved.

Chapter 3
Applications

This chapter illustrates how the abstract theory of Chapter 2 can be applied to interesting model problems such as diffusion in heterogeneous materials and time-harmonic high-frequency scattering where standard methods suffer from severe scale-dependent pre-asymptotic effects illustrated in Chapter 1.

3.1 Numerical homogenization beyond scale separation

The first prototypical model problem concerns the diffusion problem

$$-\operatorname{div} A\nabla u = f$$

in some bounded domain $\Omega \subset \mathbb{R}^d$ with homogeneous Dirichlet boundary condition. The difficulty is the strongly heterogeneous and highly varying (non-periodic) diffusion coefficient A. The heterogeneities and oscillations of the coefficient may appear on several non-separable scales. We assume that the diffusion matrix $A \in L^\infty\left(\Omega, \mathbb{R}^{d\times d}_{\mathrm{sym}}\right)$ is symmetric and uniformly elliptic with

$$0 < \alpha = \operatorname*{ess\,inf}_{x\in\Omega} \inf_{v\in\mathbb{R}^d\setminus\{0\}} \frac{(A(x)v)\cdot v}{v\cdot v}.$$

Given $f \in V' := \mathrm{H}^{-1}(\Omega)$, we wish to find the unique weak solution $u \in V := \mathrm{H}^1_0(\Omega)$ such that

$$a(u,v) := \int_\Omega (A\nabla u)\cdot\nabla v = \int_\Omega fv =: F(v) \quad \text{for all } v \in V. \tag{3.1}$$

It is well known that classical polynomial based FEMs can perform arbitrarily badly for such problems, see e.g. [8]. This is due to the fact that finite elements tend to average unresolved scales of the coefficient and the theory of homogenization shows that this way of averaging does not lead to meaningful macroscopic approximations. This was illustrated in the introduction. In the simple periodic example of Fig. 1.1, the averaging of the inverse of the diffusion coefficient A (harmonic averaging) would have led to the correct macroscopic representation.

In computational homogenization, the impact of unresolved micro-structures encoded in the rough coefficient A on the overall process is taken into account by the solution of local microscopic cell problems. While many approaches are empirically successful and robust for certain multiscale problems [24, 2], the question whether there are stable and accurate methods beyond the strong strong structural assumptions of analytical homogenization regarding scale separation or even periodicity

remained open for a long time. Only recently, the existence of an optimal approximation of the low-regularity solution space by some arbitrarily coarse generalized finite element space (that represents the homogenized problem) was shown in [7] and [32]. However, the constructions therein include prohibitively expensive global solutions of the full fine scale problem or the solution of more involved eigenvalue problems. The first efficient and feasible construction, solely based on the solution of localized microscopic cell problems, was given and rigorously justified in [A1] and later optimized [A2] and generalized in [A3, 37]. Since then, several new approaches have been developed. One approach with presumably similar properties was suggested in [54] along with the notion of sparse super-localization that reflects the locality of the discrete homogenized operator (similar to the sparsity of standard finite element matrices). Another new approach is based on the theory of iterative solvers is [41].

We shall now explain how the abstract theory of the previous sections is related to the method of [A1] and its variants. Let $\mathscr{G}_H$ denote some regular (in the sense of Ciarlet) finite element mesh into closed simplices and let $V_H := P_1(\mathscr{G}_H) \cap V$ denote the space of continuous functions that are affine when restricted to any element $T \in \mathscr{G}_H$. Let $I_H : V \to V_H$ be a quasi-interpolation operator that acts as a stable quasi-local projection in the sense that $I_H \circ I_H = I_H$ and that for any $T \in \mathscr{G}_H$ and all $v \in V$ there holds

$$H^{-1}\|v - I_H v\|_{L^2(T)} + \|I_H v\|_{V,T} \leq C_{I_H} \|\nabla v\|_{V,\Omega_T}, \tag{3.2}$$

where Ω_T refers to some neighborhood of T (typically the union of T and the adjacent elements) and $\|\cdot\|_V := \|\nabla \cdot\|_{L^2(\Omega)}$. One possible choice (among many others) is to define $I_H := E_H \circ \Pi_H$, where Π_H is the piecewise L^2 projection onto $P^1(\mathscr{G}_H)$ and E_H is the averaging operator that maps $P_1(\mathscr{G}_H)$ to V_H by assigning to each interior vertex the arithmetic mean of the corresponding function values of the adjacent elements, that is, for any $v \in P_1(\mathscr{G}_H)$ and any free vertex $z \in \mathscr{N}_H$,

$$(E_H(v))(z) = \frac{1}{\operatorname{card}\{K \in \mathscr{G}_H : z \in K\}} \sum_{T \in \mathscr{G}_H : z \in T} v|_T(z).$$

For this choice, the proof of (3.2) follows from combining the well-established approximation and stability properties of Π_H and E_H, see for example [18]. The choice of I_H in [A1, A2] was slightly different. Therein, the $L^2(\Omega)$-orthogonal projection onto V_H played the role of I_H. Since this a non-local operator, the analysis was based on the fact that the local quasi-interpolation operator of [17, Section 6] has the same kernel and, hence, induces the same method.

Following the recipe of Section 2.1.1 and taking into account the present setting with an inner product a, the ideal test space $W_H := (\operatorname{Ker} I_H)^{\perp_a}$ is simply the orthogonal complement (w.r.t. a) of the fine scale functions $\operatorname{Ker} I_H$.

Given the nodal basis of V_H, a basis of W_H is computed by means of the trial-to-test operator $\mathscr{T} = 1 + \mathscr{C}$, where

$$\forall w \in \operatorname{Ker} I_H : \; a(\mathscr{C}\lambda_z, w) = -a(\lambda_z, w). \tag{3.3}$$

It is easily checked that the assumptions made in Section 2.2 are satisfied in the present setting. In particular, formula (2.14) holds with $C_{\eta,I_H} = C_{I_H}(C_{I_H}C_\eta + 1)$ and $m = 2$. This follows from the product rule, (2.13), and the local approximation and stability properties (3.2) of I_H. This implies the exponential decay as it is stated in (2.12) with constants C and c independent of variations of the diffusion coefficient A. An example of a corrector and a test basis function are depicted in Fig. 3.1 to demonstrate the exponential decay.

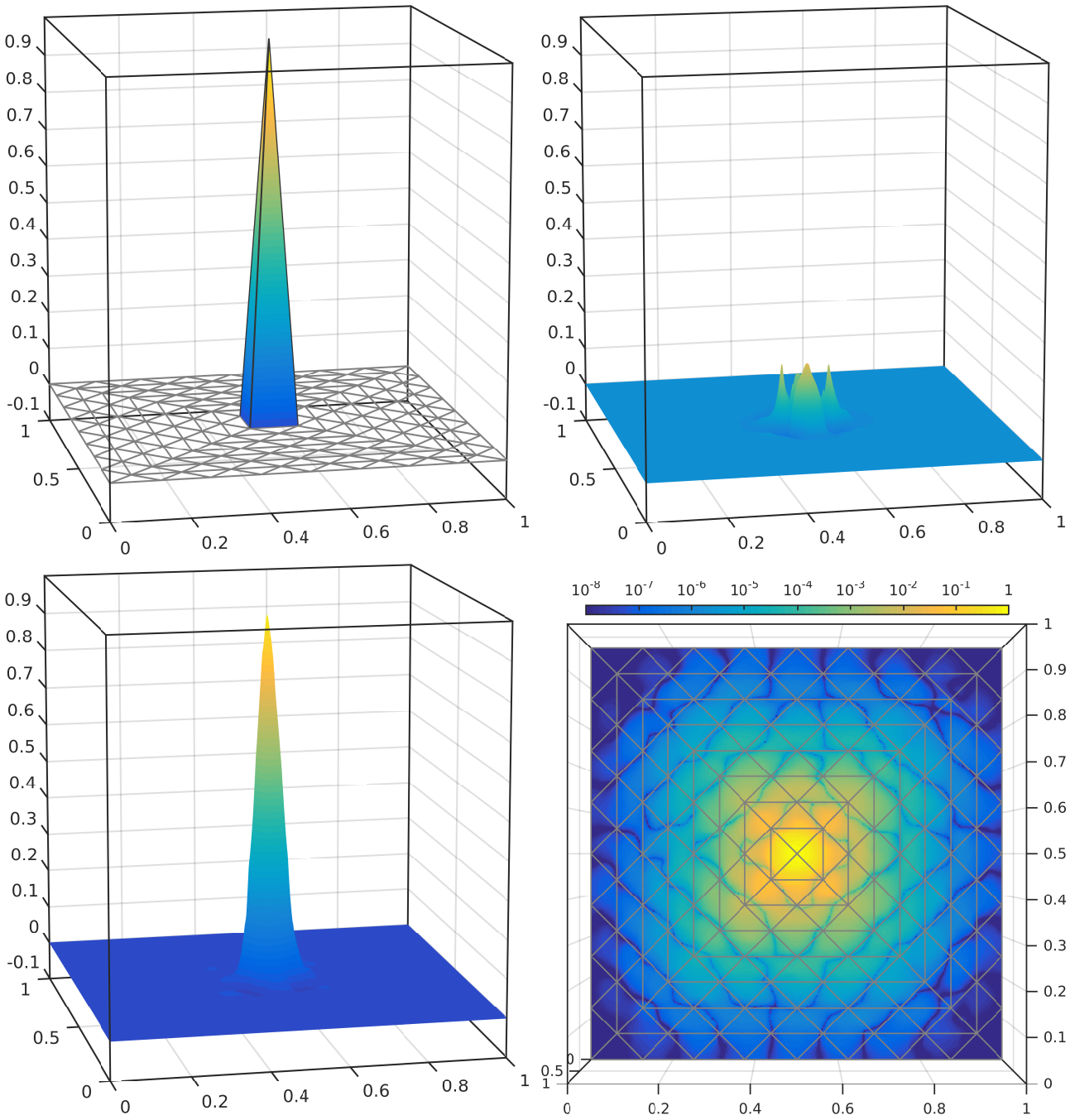


Fig. 3.1 Standard nodal basis function λ_z with respect to the coarse mesh $\mathscr{G}_H$ (top left), corresponding ideal corrector $\phi_z = \mathscr{C}\lambda_z$ (top right), and corresponding test basis function $\mathscr{T}\lambda_z = (1+\mathscr{C})\lambda_z$ (bottom left). The bottom right figure shows a top view on the modulus of test basis function $\mathscr{T}\lambda_z = (1+\mathscr{C})\lambda_z$ with logarithmic color scale to illustrate the exponential decay property. The underlying rough diffusion coefficient A is depicted in Fig. 3.4.

We truncate the computational domain of the corrector problems to local subdomains of diameter ℓH roughly. We have not yet described how to do this in practice. The obvious way would be to simply replace Ω in (3.3) with suitable neighborhoods of the nodes z. This procedure was used in [A1]. However, it turned out that it is advantageous to consider the following slightly more involved technique based on element correctors [A2, 37].

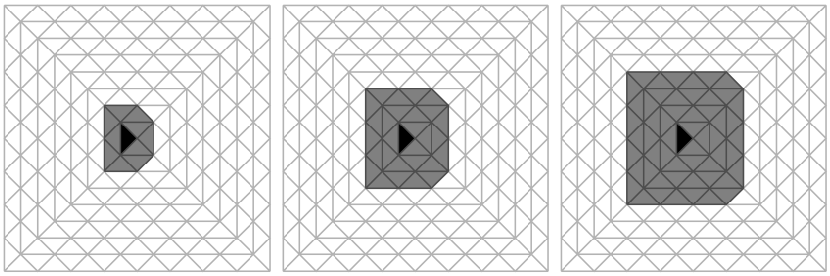


Fig. 3.2 Element patches $\Omega_{T,\ell}$ for $\ell = 1,2,3$ (from left to right) as they are used in the localized corrector problem (3.4).

We assign to any $T \in \mathscr{G}_H$ its ℓ-th order element patch $\Omega_{T,\ell}$ for a positive integer ℓ; see Fig. 3.2 for an illustration. Moreover, we define for all $v, w \in V$ and $\omega \subset \Omega$ the localized bilinear forms

$$a_\omega(v,w) := \int_\omega (A\nabla v)\cdot\nabla w.$$

Given any nodal basis function $\lambda_z \in V_H$, let $\phi_{z,\ell,T} \in \operatorname{Ker} I_H \cap \mathrm{H}^1_0(\Omega_{T,\ell})$ solve the subscale corrector problem

$$a_{\Omega_{T,\ell}}(\phi_{z,\ell,T}, w) = -a_T(\lambda_z, w) \quad \text{for all } w \in \operatorname{Ker} I_H \cap \mathrm{H}^1_0(\Omega_{T,\ell}). \tag{3.4}$$

Let $\phi_{z,\ell} := \sum_{T\in\mathscr{G}_H : z\in T} \phi_{z,\ell,T}$ and define the test function

$$\Lambda_{z,\ell} := \lambda_z + \phi_{z,\ell}.$$

The localized test basis function $\Lambda_{z,\ell}$ and the underlying correctors $\phi_{z,\ell,T}$ can be seen in Fig. 3.3. Note that we impose homogeneous Dirichlet boundary condition on the artificial boundary of the patch which is well justified by the fast decay.

More generally, we may define the localized correction operator $\mathscr{C}_\ell$ by

$$\mathscr{C}_\ell v_H := \sum_{z\in\mathscr{N}_H} v_H(z)\phi_{z,\ell}$$

as well as the localized trial-to-test operator

$$\mathscr{T}_\ell v_H := 1 + \mathscr{C}_\ell v_H = \sum_{z\in\mathscr{N}_H} v_H(z)\Lambda_{z,\ell}.$$

The space of test functions then reads

$$W_H^\ell := \mathscr{T}_\ell V_H = \operatorname{span}\{\Lambda_{z,\ell} : z \in \mathscr{N}_H\}$$

and the (localized) multiscale Petrov-Galerkin FEM seeks $u_{H,\ell} \in V_H$ such that

$$a(u_{H,\ell}, w_{H,\ell}) = (f, w_{H,\ell})_{L^2(\Omega)} \quad \text{for all } w_{H,\ell} \in W_{H,\ell}. \tag{3.5}$$

In previous papers [A1, A2, 37] we have considered the symmetric version with $W_{H,\ell}$ as trial and test space and also the reverse version with $W_{H,\ell}$ as the trial space and V_H as test space [21]. All these methods are essentially equal in the ideal case and there are no major changes in the output after localization (when only the V_H part of the discrete solution is considered). When it comes to implementation and computational complexity, the present Petrov-Galerkin version has the advantage

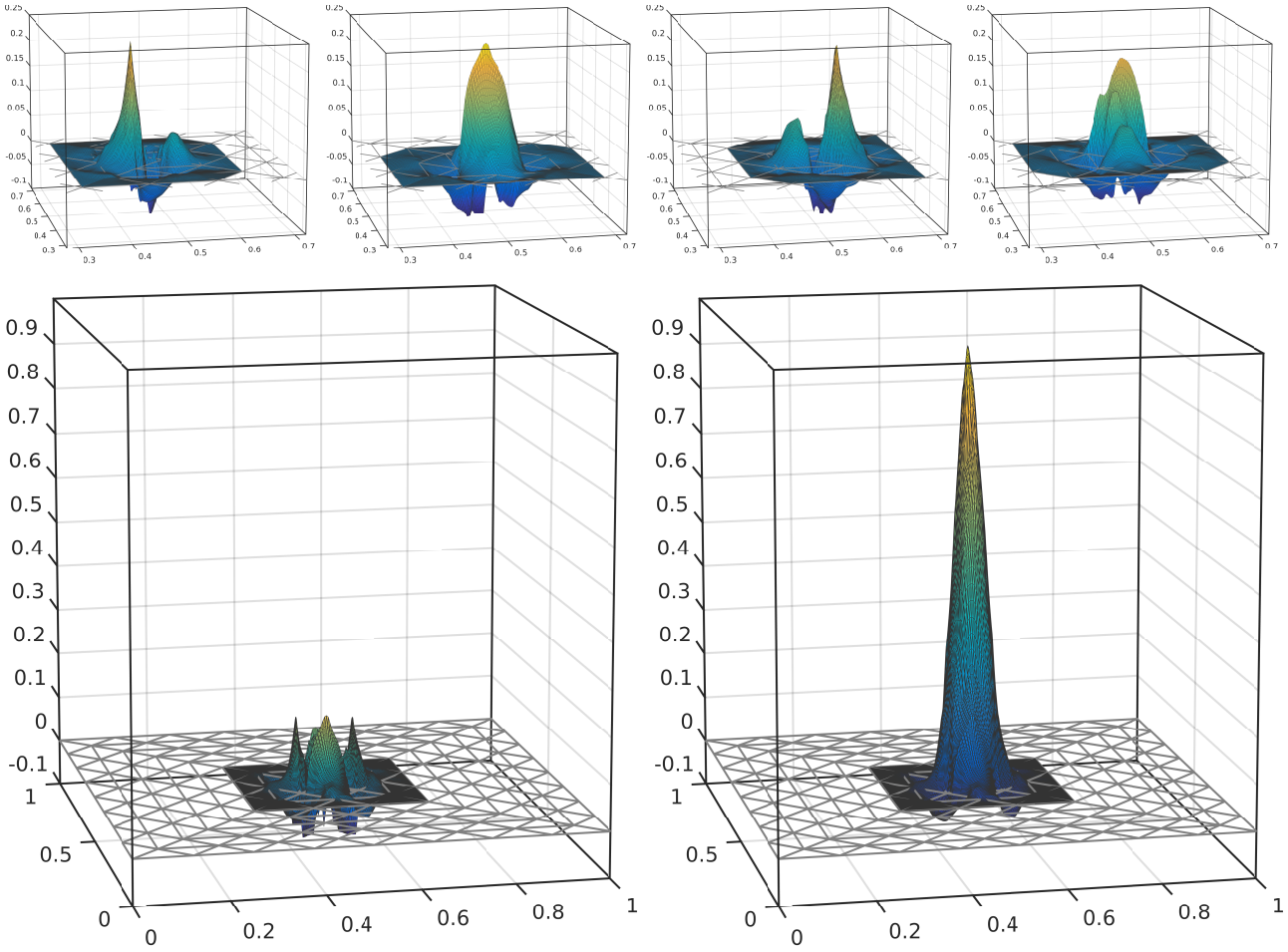


Fig. 3.3 Localized element correctors $\phi_{z,\ell,T}$ for $\ell = 2$ and all four elements T adjacent to the vertex $z = [0.5, 0.5]$ (top), localized nodal corrector $\phi_{z,\ell} = \mathscr{C}_\ell \lambda_z = \sum_{T \ni z} \phi_{z,\ell,T}$ (bottom left) and corresponding test basis function $\Lambda_{z,\ell} = \mathscr{T}_\ell \lambda_z = (1 + \mathscr{C}_\ell)\lambda_z$ (bottom right). The underlying rough diffusion coefficient is depicted in Fig. 3.4. The computations have been performed by standard linear finite elements on local fine meshes of with $h = 2^{-8}$. See Fig. 3.1 for a comparison with the ideal global corrector and basis.

that there is no communication between the correctors. This means that the fine-scale solutions of the corrector problems need not to be stored but only their interaction with the $\mathscr{O}(\ell^d)$ standard nodal basis functions in their patches; see also [21] for further discussions regarding those technical details.

The error analysis of the localized method follows similar arguments. Let $u_H \in V_H$ be the ideal Petrov-Galerkin approximation and let $e_H := u_H - u_{H,\ell} \in V_H$ denote the error with respect to the ideal method. Then there exists some $z_H \in W_H$ with $\|z_H\|_V = 1$ such that

$$\frac{\alpha}{\|I_H\|_{\mathscr{L}(V)}} \|e_H\|_V \leq a(e_H, z_H) = a(u_{H,\ell} - u, z_H - z_{H,\ell}),$$

where $z_{H,\ell} \in W_{H,\ell}$. The exponential decay property allows one to choose $z_{H,\ell}$ in such a way that $\|z_H - z_{H,\ell}\|_V \leq \tilde{C} \exp(-c\ell)$; see for instance [A2, 37]. This shows that

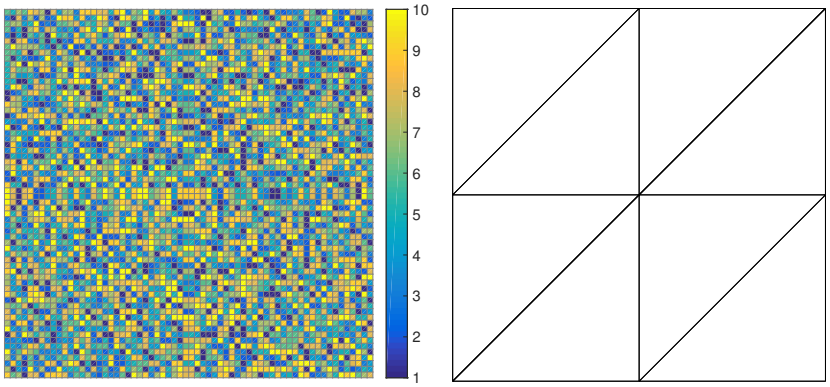


Fig. 3.4 Diffusion coefficient in the numerical experiment of Section 3.1 and coarsest mesh.

$$\begin{aligned}\|u-u_{H,\ell}\|_V &\le \|u-u_H\|_V + \|u_H-u_{H,\ell}\|_V \\ &\le \|u-u_H\|_V + \frac{\|I_H\|_{\mathscr{L}(V)}C_a}{\alpha}\tilde{C}\exp(-c\ell)\|u-u_{H,\ell}\|_V.\end{aligned}$$

We shall emphasize that, in the present context, the constants $\tilde{C}$ and c are independent of variations of the rough diffusion tensor but they may depend on the contrast (the ratio between the global upper and lower bound of A). Using (2.8), this shows that the moderate choice $\ell \ge |\log(\alpha/(2\|I_H\|_{\mathscr{L}(V)}C_a\tilde{C}))|/c = \mathscr{O}(1)$ implies the quasi-optimality (and also the well-posedness) of the Petrov-Galerkin method with respect to the V-norm

$$\|u-u_{H,\ell}\|_V \le 2\|I_H\|_{\mathscr{L}(V)} \min_{v_H\in V_H}\|u-v_H\|_V.$$

With regard to the fact that the V-best approximation may be poor and standard FE Galerkin would have provided us with an even better estimate at lower cost, this result is maybe not very impressive. Let us see if we can do something similar for the L^2-norm which appears to be the relevant measure in the context of homogenization problems. A standard duality argument shows that

$$\|e_H\|^2_{L^2(\Omega)} = a(e_H,z_H) = a(u_{H,\ell}-u, z_H - z_{H,\ell})$$

for some $z_H \in W_H$ with $\|z_H\|_V \le C_3\alpha^{-1}\|I_H\|_{\mathscr{L}(V)}\|e_H\|_{L^2(\Omega)}$ and $z_{H,\ell} := \mathscr{T}_\ell I_H z_H \in W_{H,\ell}$. Similar arguments as before yield

$$\|u-u_{H,\ell}\|_{L^2(\Omega)} \le C_1 \min_{v_H\in V_H}\|u-v_H\|_{L^2(\Omega)} + C_2\exp(-c\ell)\min_{v_H\in V_H}\|u-v_H\|_V,$$

where $C_1 := \|I_H\|_{\mathscr{L}(L^2(\Omega))}$ and $C_2 := C_a\tilde{C}C_3\alpha^{-1}\|I_H\|_{\mathscr{L}(V)}$. This shows that the method is accurate also in the L^2-norm regardless of the regularity of u. If the oversampling parameter is chosen such that $\ell \gtrsim \log H$, then the method is $\mathscr{O}(H)$ accurate in $L^2(\Omega)$ with no pre-asymptotic phenomena. This is the best worst-case rate one can expect for general $f \in V'$ and $A \in L^\infty$.

Note that the previous results hold true for general L^∞-coefficients and all constants are independent of the variations of the diffusion tensor as far as the contrast remains moderately bounded. In particular, the approach is by no means restricted to periodic coefficients or scale separation. For a more detailed discussion of high-contrast problems in this context we refer to [57].

The final step towards a fully practical method is the discretization of the fine-scale corrector problems. With regard to the possible low regularity of the solution, P_1 finite elements on a refined mesh $\mathscr{G}_h$ appears reasonable, but any other type of discretization is possible. Obviously, the fine-scale dis-

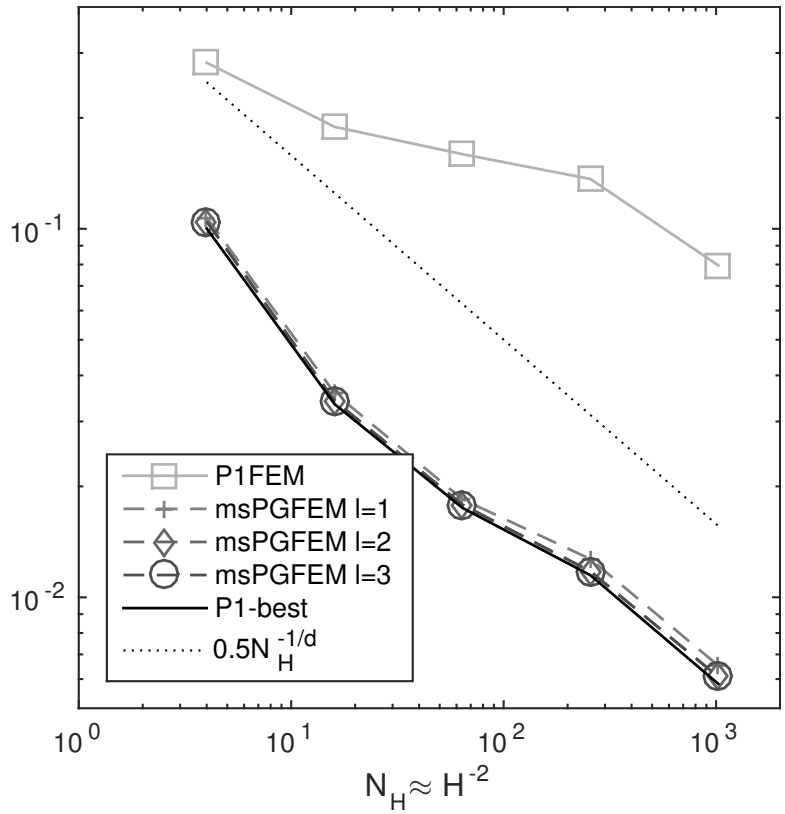


Fig. 3.5 Numerical experiment of Section 3.1. Relative L^2-errors of multiscale Petrov-Galerkin FEM (3.5) versus the number of degrees of freedom $N_H \approx H^{-2}$, where $H = 2^{-1}, \ldots, 2^{-5}$ is the uniform coarse mesh size. The localization parameter varies between $\ell = 1, \ldots, 3$. The P_1-FE solution and the best-approximation in the P_1-FE space on the same coarse meshes are depicted for comparison.

cretization parameter h has to be chosen fine enough to resolve all relevant features of the diffusion coefficient. The previous theory can be transferred to this case in a straight-forward way and we refer to [A1, A2, 35] for the technical details.

To illustrate the previous estimates, we close this section with a numerical experiment. Let Ω be the unit square and the outer force $f \equiv 1$ in Ω. Consider the coefficient A that is piecewise constant with respect to a uniform Cartesian grid of width 2^{-6}. Its values are randomly chosen between 1 and 10; see Fig. 3.4. Consider uniform coarse meshes $\mathscr{G}_H$ of size $H = 2^{-1}, 2^{-2}, \ldots, 2^{-5}$ of Ω that certainly do not resolve the rough coefficient A appropriately. The reference mesh $\mathscr{G}_h$ has width $h = 2^{-9}$. Since no analytical solutions are available, the standard finite element approximation $u_h \in V_h$ on the reference mesh $\mathscr{G}_h$ serves as the reference solution. Doing this, we assume that u_h is sufficiently accurate and, necessarily, that $\mathscr{G}_h$ resolves the discontinuities of A. The corrector problems are also solved on this scale of numerical resolution.

The numerical results, i.e. errors with respect to the reference solution u_h are depicted in Fig. 3.5. The results are in agreement with the theoretical results. They are even better in the sense that $\ell = 1$ seems to be sufficient for quasi-optimality (with respect to u_h) in the present setup and parameter regime. We expect that the true errors with respect to u would behave similar in the beginning but level off at some point when the reference error starts to dominate the upscaling error. Still, the experiment clearly indicates that numerical homogenization is possible for very general L^∞-coefficients.

We refer to [A1, A2, 37, A3, 21, 3, 9, 23, 35, C1, C2, 49] for many more numerical experiments for several model problems including nonlinear stationary and non-stationary problems as well as eigenvalue problems.

3.2 Pollution-free high-frequency acoustic scattering

This section will show that the abstract framework of Chapter 2 is indeed applicable beyond the coercive and symmetric model problem of the previous section. We consider the scattering of acoustic waves at a sound-soft scatterer modeled by the Helmholtz equation over a bounded Lipschitz domain $\Omega \subset \mathbb{R}^d$ $(d = 1,2,3)$,

$$-\Delta u - \kappa^2 u = f \quad \text{in } \Omega, \tag{3.6.a}$$

along with mixed boundary conditions of Dirichlet and Robin type

$$u = 0 \quad \text{on } \Gamma_D, \tag{3.6.b}$$

$$\nabla u \cdot \nu - i\kappa u = 0 \quad \text{on } \Gamma_R. \tag{3.6.c}$$

Here, the wave number $\kappa \gg 1$ is real and positive, i denotes the imaginary unit and $f \in L^2(\Omega,\mathbb{C})$. We assume that the boundary $\Gamma := \partial\Omega$ consists of two components

$$\partial\Omega = \overline{\Gamma_D \cup \Gamma_R}, \quad \overline{\Gamma}_D \cap \overline{\Gamma}_R = \emptyset$$

where Γ_D encloses the scatterer and Γ_R is an artificial truncation of the whole unbounded space. The vector ν denotes the unit normal vector that is outgoing from Ω.

Given $f \in L^2(\Omega,\mathbb{C})$, we wish to find $u \in V := \{v \in \mathrm{H}^1(\Omega,\mathbb{C}) \mid v = 0 \text{ on } \Gamma_D\}$ such that, for all $v \in V$,

$$a(u,v) := \int_\Omega \nabla u \cdot \nabla \bar{v} - \kappa^2 \int_\Omega u\bar{v} - i\kappa \int_{\Gamma_R} u\bar{v} = \int_\Omega f\bar{v} =: \overline{F(w)}. \tag{3.7}$$

The space V is equipped with the usual κ-weighted norm

$$\|v\|_V^2 := \kappa^2 \|v\|_{L^2(\Omega)}^2 + \|\nabla v\|_{L^2(\Omega)}^2.$$

The presence of the Robin boundary condition (3.6.c) ensures that this variational problem is well-posed in the sense of (2.2) with $\alpha = 1/C_{\mathrm{st}}(\kappa)$ for some κ-dependent stability constant $C_{\mathrm{st}}(\kappa)$; see e.g. [28]. The dependence on the wave number κ is not known in general. An exponential growth with respect to the wave number is possible [6] in non-generic domains. In most cases, the growth seems to be only polynomially, although this is an empirical rather than a theoretical statement, and sufficient geometric conditions for this to hold are rare [28, 46, 14, 47]. For the above scattering problem, we know that $C_{\mathrm{st}}(\kappa) \leq \mathcal{O}(\kappa)$ if Ω is convex and if the scatterer is star-shaped [33].

It is this κ-dependence in the stability of the problem that makes the numerical approximation by FEM or related schemes extremely difficult in the regime of large wave numbers. Any perturbation of the problem, e.g. by some discretization, can be amplified by $C_{\mathrm{st}}(\kappa)$. We have seen in the introduction that this is indeed observed in practice and causes a pre-asymptotic effect known as the pollution effect or numerical dispersion [10]. This effect puts very restrictive assumptions on the smallness of the underlying mesh that is much stronger than the minimal requirement for a meaningful representation

of highly oscillatory functions from approximation theory, that is, to have at least $5-10$ degrees of freedom per wave length and coordinate direction.

It is the aim of many newly developed methods to overcome or at least to reduce the pollution effect; see [61, 30, 31, 36, 63, 19] among many others. However, the only theoretical results regard high-order FEMs with the polynomial degree p coupled to the wave number κ via the relation $p \approx \log \kappa$ [51, 52, 50, 28]. Under this moderate assumption, those methods are stable and quasi-optimal in the regime $H\kappa/p \lesssim 1$ for certain sufficiently regular model Helmholtz problems.

The multiscale method of [B1] then showed that pollution in the numerical approximation of the Helmholtz problem can also be cured for a fairly large class of Helmholtz problems, including the acoustic scattering from convex non-smooth objects, by stabilization in the present framework. If the data of the problem (domain, boundary condition, force term) allows for polynomial-in-κ bounds of $C_{\mathrm{st}}(\kappa)$ and if the resolution condition $H\kappa \lesssim 1$ and the oversampling condition $\log(\kappa)/\ell \lesssim 1$ are satisfied, then the method is stable and quasi-optimal in the V-norm.

The recent paper [B2] interprets the method of [B1] in the present framework and we recall it here very briefly. Given the same discrete setup as in the previous section with some simplicial mesh $\mathscr{G}_H$, corresponding P_1 FE space $V_H := P_1(\mathscr{G}_H) \cap V$, and quasi-interpolation operator $I_H : V \to V_H$, the multiscale Petrov-Galerkin method is formally defined in the same way. We simply replace the inner product of Section 3.1 with the sesquilinear form a of this section.

Given any nodal basis function $\lambda_z \in V_H$, we construct a corresponding test basis function $\Lambda_{z,\ell}$ by the same procedure as in the previous section, $\Lambda_{z,\ell} := \lambda_z + \phi_{z,\ell}$, where $\phi_{z,\ell} := \sum_{T \in \mathscr{G}_H : z \in T} \phi_{z,T}$ and $\phi_{z,T}$ solves the cell problem

$$a_{\Omega_{T,\ell}}(w, \phi_{z,T}) = -a_T(w, \lambda_z) \quad \text{for all } w \in \operatorname{Ker} I_H \text{ with } \operatorname{supp} w \subset \bar{\Omega}_T.$$

Here,

$$a_\omega(u,v) := \int_{\Omega \cap \omega} \nabla u \cdot \nabla \bar{v} - \kappa^2 \int_{\Omega \cap \omega} u\bar{v} - i\kappa \int_{\Gamma_R \cap \partial\omega} u\bar{v}$$

for $\omega \in \{\Omega_{T,\ell}, T\}$. Note that the corrector problem inherits the boundary condition from the original problem when the patch boundary coincides with the boundary of Ω. On the part of the patch boundary that falls in the interior of Ω, we simply put the homogeneous Dirichlet condition. A major observation is that this corrector problem is well-posed and, in particular, coercive with $\beta = 1/3$ under the condition $H\kappa \le c_{\mathrm{res}}$ for some given constant $0 < c_{\mathrm{res}} = \mathscr{O}(1)$ that only depends on the constant in (3.2) but not on H or κ. This is because a satisfies a Gårding inequality and fine-scale functions satisfy

$$\|w\|_{L^2(\Omega)} \le C_{I_H} H \|\nabla w\|_{L^2(\Omega)}.$$

The coercivity of the sesquilinear form a on $\operatorname{Ker} I_H$ also implies the desired exponential decay of the ideal correctors so that the choice $\Omega_{T,\ell}$ is well justified. This can also be observed in Fig. 3.6.

The space of localized test functions then reads $W_{H,\ell} := \operatorname{span}\{\Lambda_{z,\ell} : z \in \mathscr{N}_H\}$ and the multiscale Petrov-Galerkin FEM seeks $u_{H,\ell} \in V_H$ such that

$$a(u_{H,\ell}, w_{H,\ell}) = \overline{F(w_{H,\ell})} \quad \text{for all } w_{H,\ell} \in W_{H,\ell}. \tag{3.8}$$

The quasi-optimality result of the previous section is easily transferred to the present setting. The resolution condition $H\kappa \le c_{\mathrm{res}}$ and the oversampling condition

$$\ell \ge |\log(\alpha/(2\|I_H\|_{\mathscr{L}(V)} C_a \tilde{C}))|/c = \mathscr{O}(\log C_{\mathrm{st}}(\kappa))$$

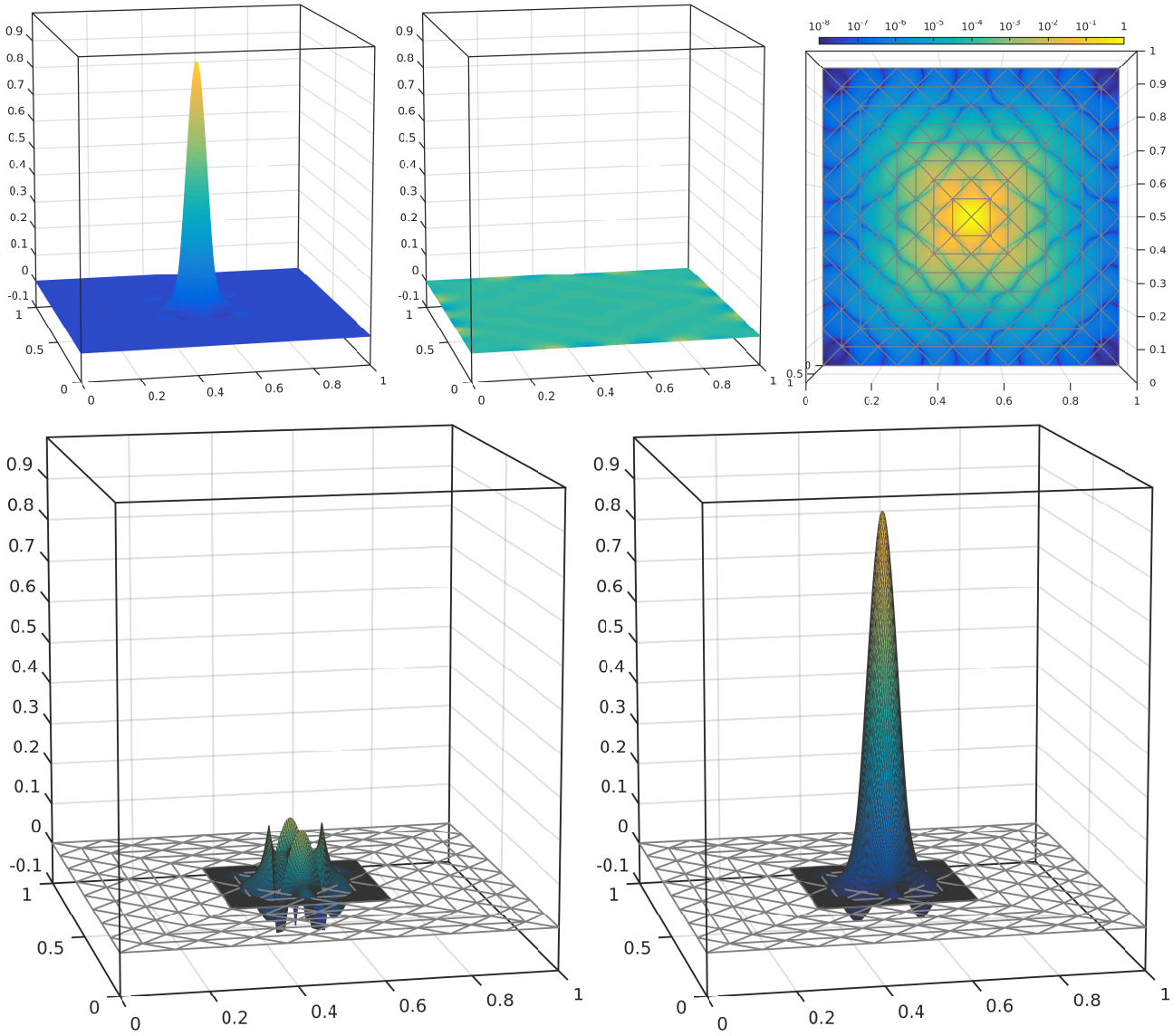


Fig. 3.6 Real and imaginary part of the ideal corrector $\mathscr{C}\lambda_z$ (top left and middle). The top right figure shows a top view on the modulus of test basis function $\mathscr{T}\lambda_z = (1+\mathscr{C})\lambda_z$ with logarithmic color scale to illustrate the exponential decay property. The underlying computational domain is the unit square with a Robin boundary condition everywhere. The wave number $\kappa = 2^4$ is chosen such that the resolution condition on the coarse mesh is just satisfied. The localized nodal corrector $\phi_{z,\ell} = \mathscr{C}_\ell\lambda_z$ (bottom left) and corresponding test basis function $\Lambda_{z,\ell} = \mathscr{T}\lambda_z$ (bottom right) are real-valued because the patch boundary doesn't touch the domain boundary. The local fine meshes used in the computation have width $h = 2^{-8}$.

imply the quasi-optimality (and stability) of the multiscale Petrov-Galerkin method with respect to the V-norm

$$\|u - u_{H,\ell}\|_V \leq 2\|I_H\|_{\mathscr{L}(V)} \min_{v_H \in V_H} \|u - v_H\|_V .$$

Here, the constants c and $\tilde{C}$ are related to the exponential decay of the test basis (cf. (2.12)) and they are independent of κ under the resolution condition. We shall emphasize that such a best-approximation property does not hold for standard FEMs which require e.g. $\kappa^2 H \lesssim 1$ for quasi-optimality [46] in the case of pure Robin boundary conditions on a convex planar domain. The FEM

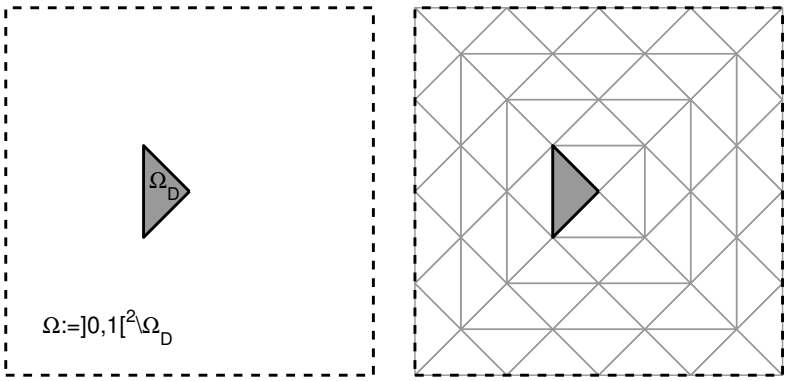


Fig. 3.7 Computational domain of the model problem of Section 3.2 and coarsest mesh.

approximation is not even known to exist unless $\kappa^{3/2}H \lesssim 1$ in the simplest model problem without a scatterer [62].

For the multiscale Petrov-Galerkin method, the result means that pollution effects do not occur. Note that the resolution condition $H\kappa \leq c_{\text{res}}$ is somewhat minimal, because any meaningful approximation of the highly oscillatory solution of (3.6) requires at least $5-10$ degrees of freedom per wave length and coordinate direction. Saying this, we assume that the fine scale corrector problems are solved sufficiently accurate; see [B2, B1] for details.

We shall present a numerical experiment taken from [B1] where this version of the method was already considered experimentally. Consider the scattering from sound-soft scatterer occupying the triangle Ω_D. The Sommerfeld radiation condition of the scattered wave is approximated by the Robin boundary condition on the boundary $\Gamma_R := \partial\Omega_R$ of the unit square so that $\Omega := (0,1)^2 \setminus \Omega_D$ is the computational domain; see Fig. 3.7. Given the wave number $\kappa = 2^7$, the incident wave $u_{inc}(x) := \exp(i\kappa\, x \cdot [\cos(0.5), \sin(0.5)]^T)$ is prescribed via an inhomogeneous Dirichlet boundary condition on $\Gamma_D := \partial\Omega_D$ and the scattered wave satisfies (3.6.a) with $f \equiv 0$ and the boundary conditions

$$\begin{aligned} u &= -u_{inc} \quad \text{on } \Gamma_D, \\ \nabla u \cdot \nu - i\kappa u &= 0 \quad \text{on } \Gamma_R. \end{aligned}$$

The error analysis extends to this setting in a straight-forward way.

We choose uniform coarse meshes of widths $H = 2^{-3}, \ldots, 2^{-7}$ as depicted in Fig. 3.7. The reference mesh $\mathscr{G}_h$ is derived by uniform mesh refinement of the coarse meshes and has mesh width $h = 2^{-9}$. The corresponding P_1 conforming finite element approximation on the reference mesh $\mathscr{G}_h$ is denoted by V_h. As in the previous section, we compare the coarse scale approximations $u_{H,\ell,h} \in V_H$ with some reference solution $u_h \in V_h$.

Fig. 3.8 depicts the results for the multiscale Petrov-Galerkin method and shows that the pollution effect that is present in the P_1 FEM is eliminated when ℓ is moderately increased. For the present wave number $\ell = 2$ is sufficient.

Further numerical experiments are reported in [B1] and [B2]. It is worth noting that the latter work also exploits the homogeneous structure of the PDE coefficients in the sense that only very few of the fine-scale corrector problems are actually solved due to translation invariance and symmetry. This makes the approach competitive.

A very natural and straight forward generalization of the method would be the case of heterogeneous media. The previous section plus the analysis of this section strongly indicate the potential of

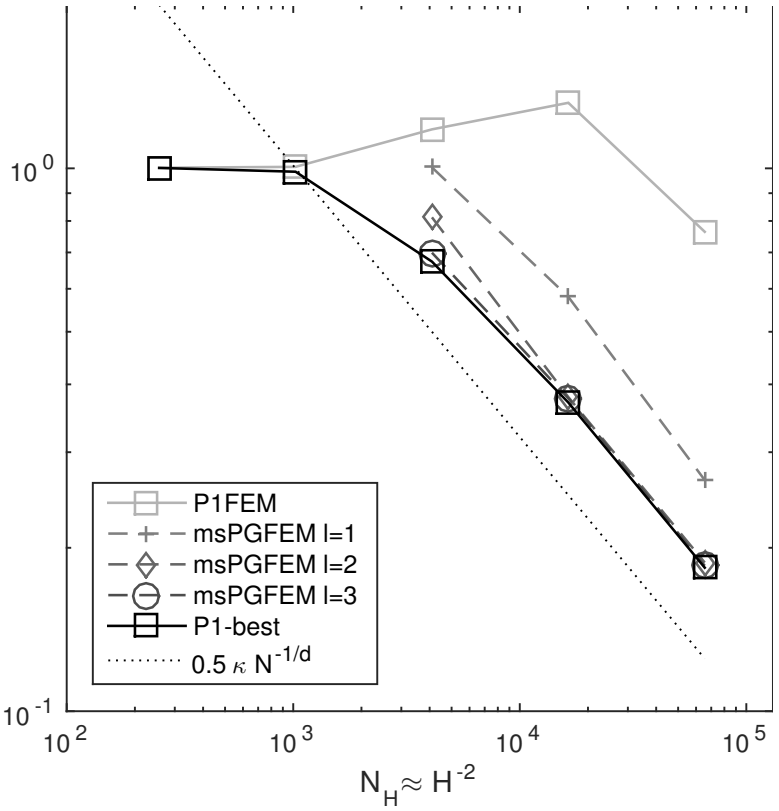


Fig. 3.8 Numerical experiment of Section 3.2: Relative V-norm errors of multiscale Petrov-Galerkin method (3.8) with wave number $\kappa = 2^7$ depending on the number of degrees of freedom $N_H \approx H^{-2}$, where $H = 2^{-5}, \ldots, 2^{-7}$ is the uniform coarse mesh size. The reference mesh size $h = 2^{-9}$ remains fixed. The oversampling parameter ℓ varies between 1 and 3. The P_1-FE solution and the best-approximation in the P_1-FE space on the same coarse meshes are depicted for comparison.

the method to treat high oscillations or jumps in the PDE coefficients and the pollution effect in one stroke.

We may close this section with some rather philosophical remark regarding the stabilization of FEMs and their inter-element continuity properties. Presently, it is believed, e.g. in the context of time-harmonic wave propagation, that stability can be increased by relaxing inter-element continuity within a discontinuous Galerkin (DG) framework. The large number of variants includes the ultra weak variational formulation [12], Trefftz methods [36], DPG [63, 19], or the continuous interior penalty method [62]. There may be some truth in this but the general impression that relaxing continuity is the only way is certainly false as one can observe from the method presented in this thesis. The multiscale Petrov-Galerkin does quite the opposite. The regularity of the test functions is increased compared to standard continuous finite elements, because they are solutions of second-order elliptic problems (at least in the ideal case). In general, test functions $w_H \in W_H$ have the property that

$$\operatorname{div} A \nabla w_H \in L^2(\Omega).$$

In the context of the Helmholtz model problem of this section where $A = 1$ this means that $\Delta W_H \subset L^2(\Omega)$. If Ω is convex and boundary conditions are appropriate, then

$$W_H \subset \mathrm{H}^2(\Omega).$$

This high regularity can be observed for one basis function in Fig. 3.6. In this respect, our methodology clearly indicates that increased differentiability might as well lead to increased stability and accuracy. Similar effects have been observed for eigenvalue computations in IGA [16] and also LOD [C1]. This shows that breaking the inter-element continuity is not at all necessary for stability.

Chapter 4
Summary and further results

In the preceding chapters, we have presented an abstract framework for the stabilization of numerical methods for multiscale partial differential equations with some focus on highly oscillatory problems. The methodology is based on the variational multiscale method and the more recent development of localized orthogonal decompositions. We have provided an abstract numerical analysis of the method which is applied to two representative model problems, a homogenization problem and a scattering problem. We have shown that the methodology can indeed eliminate critical scale-dependent pre-asymptotic effects in these cases. We expect that the framework will also be useful for convection-dominated flow, the problem that the variational method was initially designed for.

The framework leads to multiscale methods that are provably stable and accurate under moderate assumptions on the discretization parameters relative to characteristic parameters and length scales of the problem. These valuable properties require the pre-computation of the test basis on sub-grids. These pre-computations are both local and independent, but the worst-case (serial) complexity of the method can exceed the cost of a direct numerical simulation on a global sufficiently fine mesh. If the inherent parallelism of the local cell problems cannot be exploited during the computation, we still expect a significant gain with respect to computational complexity if the pre-computation can be reused several times in the context of time-dependent problems, parameter studies, coupled problems, optimal control problems or inverse problems. In many cases, there is also a lot of redundancy in the local problems which allows one to reduce the number of local problems drastically as it is shown in [B2] in the context of acoustic scattering. We expect that this technique can be generalized to far more general situations using modern techniques of model order reduction [58, 1].

In addition to the results of Appendices A–B, Appendix C shows that the multiscale framework applies to another important class of problems, i.e., eigenvalue problems. The major achievement of [C1] is that the variational multiscale method as it is interpreted in Appendix A preserves the spectrum of the linear partial differential operator in a stable and super-convergent way even in the case of rough coefficients. This observation has surprising applications, e.g., the efficient computation of ground states of Bose-Einstein condensates in quantum chemistry [C2] and also for the efficient solution of quadratic eigenvalue problems [49].

While the new multiscale methods can deal with arbitrarily fast and non-smooth oscillations in representative linear and non-linear problems [A1]–[C2] without any pre-asymptotic effects, the theory breaks down immediately for high-contrast coefficients (and also perforated domains) because the exponential decay rate of certain correctors that justify the localization degenerates with increasing contrast (cf. (2.12)). Numerical experiments in [A1, 23, 9] and also ongoing research in this direction [57] are less pessimistic but not sufficiently clear and systematic to draw any conjecture for very general classes of coefficients in the regime of high contrast.

The papers of Appendix D approach the high-contrast problem from a different direction and with a different motivation. In the case of heterogeneous materials, propagating interface cracks are among the predominant phenomena that cause degradation of the mechanical properties and – in combination with other mechanisms – eventually lead to material failure. The modeling of such effects and further interface phenomena require the resolution of the highly complex material interfaces by computational meshes aligned with interfaces. In this context, we are not heading for arbitrary coarse scales of discretization but we want to keep the resolution minimal in terms of computational complexity while still capturing the relevant properties. However, the computation of standard polygonal meshes that fulfill such a conformity constraint (at least approximately) is a major difficulty (see e.g. [20, Section 2] on constraint triangulations and references therein) that becomes even more pronounced due to the high-quality demands from finite element design (e.g. local quasi-uniformity or other angle conditions). Recent textbooks still constitute mesh generation as the bottleneck of computational partial differential equations [29, 13]. The papers [D1, D2] of Appendix D and further work [25] are devoted to this issue in the model situation of densely packed particle-reinforced composites. They establish a new structural discretization approach to the PDE models of particle composites, i.e., a new methodology which connects the finite element method with techniques from discrete network analysis. This allows one to predict macroscopic material properties very efficiently and to trace its dependence on the microscopic phase geometries, even in the regime of high contrast. Efficiency, here, refers to the fact that the finite element network model of conductivity has only $\mathcal{O}(1)$ degrees of freedom per particle independent of the actual phase geometries. This is quasi minimal. Still, the new approach allows to reliably predict the characteristic percolation of thermal conductivity as the volume fraction of particles approaches its maximum [D2, 25]. The numerical analysis of the new approach requires very precise scale-dependent stability and regularity results that are explicit with respect to the contrast and other geometric parameters such as inter-particle distances [D2]. Further results in this direction for different classes of coefficients are provided in [56] and [D3]. The latter paper also shows how to relax the requirement of exact geometric resolution of material interfaces by shifting it to the ansatz functions in a controllable way using sub-grid techniques. The methods of Appendix D are also relevant in the previous multiscale framework of Appendices A–C because, in the end, its feasibility and efficiency rests on the ability to solve local problems on the microscopic scale with minimal complexity.

Altogether, this thesis interprets computational multiscale methods as a systematic approach to the modeling and simulation of multiscale problems which includes both the derivation of detailed models adapted to all relevant scales in an efficient way [D1]–[D3] as well as the significant reduction of computational complexity by, e.g., the compression/filtering to a coarse scale of interest while still maintaining its essential features (upscaling/homogenization) [A1]–[C1], the reconstruction of information on fine scales from coarse scale computations (down-scaling) [C1], and the fast simulation of the (non-linear) fine scale problem by iterative up- and down-scaling (two- or multi-level method) [C2].

References

[A1] A. Målqvist and D. Peterseim. Localization of elliptic multiscale problems. *Mathematics of Computation*, 83(290):2583–2603, 2014.

[A2] P. Henning and D. Peterseim. Oversampling for the multiscale finite element method. *Multiscale Modeling & Simulation*, 11(4):1149–1175, 2013.

[A3] Patrick Henning, Axel Målqvist, and Daniel Peterseim. A localized orthogonal decomposition method for semi-linear elliptic problems. *ESAIM: Mathematical Modelling and Numerical Analysis*, 48:1331–1349, 2014.

[B1] D. Peterseim. Eliminating the pollution effect in Helmholtz problems by local subscale correction. *ArXiv e-prints*, 1411.1944, 2014.

[B2] D. Gallistl and D. Peterseim. Stable multiscale Petrov-Galerkin finite element method for high frequency acoustic scattering. *Computer Methods in Applied Mechanics and Engineering*, 295:1–17, 2015.

[C1] Axel Målqvist and Daniel Peterseim. Computation of eigenvalues by numerical upscaling. *Numerische Mathematik*, 130(2):337–361, 2015.

[C2] P. Henning, A. Målqvist, and D. Peterseim. Two-level discretization techniques for ground state computations of Bose-Einstein condensates. *SIAM Journal on Numerical Analysis*, 52(4):1525–1550, 2014.

[D1] D. Peterseim and C. Carstensen. Finite element network approximation of conductivity in particle composites. *Numerische Mathematik*, 124(1):73-97, 2013.

[D2] D. Peterseim. Robustness of finite element simulations in densely packed random particle composites. *Networks and Heterogeneous Media*, 7(1):113126, 2012.

[D3] D. Peterseim. Composite finite elements for elliptic interface problems. *Mathematics of Computation*, 83(290):2657-2674, 2014.

1. A. Abdulle and Y. Bai. Reduced-order modelling numerical homogenization. *Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.*, 372(2021):20130388, 2014.
2. A. Abdulle, E Weinan, B. Engquist and E. Vanden-Eijnden. The heterogeneous multiscale method. *Acta Numerica*, 21:1–87, 2012.
3. A. Abdulle and P. Henning. Localized orthogonal decomposition method for the wave equation with a continuum of scales. *ArXiv e-prints*, 1406.6325, 2014.
4. A. Abdulle and P. Henning. A reduced basis localized orthogonal decomposition. *J. Comp. Phys.*, 295:379–401, 2015.
5. I. Babuška. Error-bounds for finite element method. *Numer. Math.*, 16:322–333, 1970/1971.
6. T. Betcke, S. N. Chandler-Wilde, I. G. Graham, S. Langdon, and M. Lindner. Condition number estimates for combined potential integral operators in acoustics and their boundary element discretisation. *Numer. Methods Partial Differential Equations*, 27(1):31–69, 2011.
7. 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.
8. I. Babuška and J. E. Osborn. Can a finite element method perform arbitrarily badly? *Math. Comp.*, 69(230):443–462, 2000.
9. D. L. Brown and D. Peterseim. A multiscale method for porous microstructures. *ArXiv e-prints*, 2014.
10. I. M. Babuška and S. A. Sauter. Is the pollution effect of the FEM avoidable for the Helmholtz equation considering high wave numbers? *SIAM Rev.*, 42(3):451–484 (electronic), 2000.

11. C. Carstensen. Quasi-interpolation and a posteriori error analysis in finite element methods. *M2AN Math. Model. Numer. Anal.*, 33(6):1187–1202, 1999.
12. O. Cessenat and B. Despres. Application of an ultra weak variational formulation of elliptic PDEs to the two-dimensional Helmholtz problem. *SIAM J. Numer. Anal.*, 35(1):255–299, 1998.
13. S.-W. Cheng, T.D. Dey, and J.R. Shewchuk. *Delaunay Mesh Generation*. Chapman and Hall / CRC computer and information science series. CRC Press, 2013.
14. P. Cummings and X. Feng. Sharp regularity coefficient estimates for complex-valued acoustic and elastic Helmholtz equations. *Math. Models Methods Appl. Sci.*, 16(1):139–160, 2006.
15. P. Clément. Approximation by finite element functions using local regularization. *Rev. Française Automat. Informat. Recherche Opérationnelle Sér. RAIRO Analyse Numérique*, 9(R-2):77–84, 1975.
16. J.A. Cottrell, A. Reali, Y. Bazilevs, and T.J.R. Hughes. Isogeometric analysis of structural vibrations. *Computer Methods in Applied Mechanics and Engineering*, 195(4143):5257–5296, August 2006.
17. C. Carstensen and R. Verfürth. Edge residuals dominate a posteriori error estimates for low order finite element methods. *SIAM J. Numer. Anal.*, 36(5):1571–1587, 1999.
18. D. A. Di Pietro and A. Ern. *Mathematical aspects of discontinuous Galerkin methods*, volume 69 of *Mathématiques & Applications (Berlin)*. Springer, Heidelberg, 2012.
19. L. Demkowicz, J. Gopalakrishnan, I. Muga, and J. Zitelli. Wavenumber explicit analysis of a DPG method for the multidimensional Helmholtz equation. *Comput. Methods Appl. Mech. Engrg.*, 213/216:126–138, 2012.
20. H. Edelsbrunner. *Geometry and topology for mesh generation*. Cambridge University Press, 2006.
21. D. Elfverson, V. Ginting, and P. Henning. On multiscale methods in Petrov-Galerkin formulation. *Numer. Math.*, pages 1–40, 2015.
22. D. Elfverson, E. H. Georgoulis, and A. Målqvist. An adaptive discontinuous Galerkin multiscale method for elliptic problems. *Multiscale Model. Simul.*, 11(3):747–765, 2013.
23. 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.
24. Y. Efendiev and T. Y. Hou. *Multiscale finite element methods*, volume 4 of *Surveys and Tutorials in the Applied Mathematical Sciences*. Springer, New York, 2009. Theory and applications.
25. M. Eigel and D. Peterseim, Simulation of Composite Materials by a Network FEM with Error Control. *Computational Methods in Applied Mathematics*, 15(1):21–37, 2015.
26. D. Elfverson. A discontinuous Galerkin multiscale method for convection-diffusion problems. *ArXiv e-prints*, 2015.
27. D. Elfverson, M. G. Larson, and A. Målqvist. Multiscale methods for problems with complex geometry. *ArXiv e-prints*, 2015.
28. S. Esterhazy and J. M. Melenk. On stability of discretizations of the Helmholtz equation. In *Numerical analysis of multiscale problems*, volume 83 of *Lect. Notes Comput. Sci. Eng.*, pages 285–324. Springer, Heidelberg, 2012.
29. P. J. Frey and P.-L. George. *Mesh Generation: Application to Finite Elements*. John Wiley and Sons, 2nd edition, 2008.
30. X. Feng and H. Wu. Discontinuous Galerkin methods for the Helmholtz equation with large wave number. *SIAM J. Numer. Anal.*, 47(4):2872–2896, 2009.
31. X. Feng and H. Wu. *hp*-discontinuous Galerkin methods for the Helmholtz equation with large wave number. *Math. Comp.*, 80(276):1997–2024, 2011.
32. 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.
33. U. Hetmaniuk. Stability estimates for a class of Helmholtz problems. *Commun. Math. Sci.*, 5(3):665–678, 2007.
34. F. Hellman, P. Henning, and A. Målqvist. Multiscale mixed finite elements. *ArXiv e-prints*, 1501.05526, 2015. (to appear in DCDS-S)
35. P. Henning and A. Målqvist. Localized orthogonal decomposition techniques for boundary value problems. *SIAM J. Sci. Comput.*, 36(4):A1609–A1634, 2014.
36. R. Hiptmair, A. Moiola, and I. Perugia. Plane wave discontinuous Galerkin methods for the 2D Helmholtz equation: analysis of the *p*-version. *SIAM J. Numer. Anal.*, 49(1):264–284, 2011.
37. P. Henning, P. Morgenstern, and D. Peterseim. Multiscale partition of unity. In Michael Griebel and Marc Alexander Schweitzer (Eds.), *Meshfree Methods for Partial Differential Equations VII*, volume 100 of *Lecture Notes in Computational Science and Engineering*, pages 185–204. Springer International Publishing, 2015.

38. T. J. R. Hughes. Multiscale phenomena: Green's functions, the Dirichlet-to-Neumann formulation, subgrid scale models, bubbles and the origins of stabilized methods. *Comput. Methods Appl. Mech. Engrg.*, 127(1-4):387–401, 1995.
39. 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.
40. 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.
41. R. Kornhuber and H. Yserentant. Numerical homogenization of elliptic multiscale problems by subspace decomposition. *SFB 1114 Preprint*, 2015.
42. M. G. Larson and A. Mlqvist. Adaptive variational multiscale methods based on a posteriori error estimation: Energy norm estimates for elliptic problems. *Comput. Methods Appl. Mech. Engrg.*, 196(2124):2313 – 2324, 2007.
43. M. G. Larson and A. Målqvist. A mixed adaptive variational multiscale method with applications in oil reservoir simulation. *Math. Models Methods Appl. Sci.*, 19(07):1017–1042, 2009.
44. A. Målqvist. *Adaptive variational multiscale methods*. 2005. Thesis (Ph.D.) - Chalmers Tekniska Hgskola (Sweden).
45. A. Målqvist. Multiscale methods for elliptic problems. *Multiscale Model. Simul.*, 9:1064–1086, 2011.
46. J. M. Melenk. *On generalized finite-element methods*. ProQuest LLC, Ann Arbor, MI, 1995. Thesis (Ph.D.)–University of Maryland, College Park.
47. Ch. Makridakis, F. Ihlenburg, and I. Babuška. Analysis and finite element methods for a fluid-solid interaction problem in one dimension. *Math. Models Methods Appl. Sci.*, 06(08):1119–1141, 1996.
48. A. Målqvist and A. Persson. Multiscale techniques for parabolic equations. *ArXiv e-prints*, 2015.
49. A. Målqvist and D. Peterseim. Generalized finite element methods for quadratic eigenvalue problems. *ArXiv e-prints*, 2015.
50. J. M. Melenk, A. Parsania, and S. Sauter. General DG-methods for highly indefinite Helmholtz problems. *J. Sci. Comput.*, 57(3):536–581, 2013.
51. J. M. Melenk and S. A. Sauter. Convergence analysis for finite element discretizations of the Helmholtz equation with Dirichlet-to-Neumann boundary conditions. *Math. Comp.*, 79(272):1871–1914, 2010.
52. 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.
53. H. Owhadi. Multigrid with rough coefficients and Multiresolution operator decomposition from Hierarchical Information Games. *ArXiv e-prints*, 2015.
54. H. Owhadi, L. Zhang, and L. Berlyand. Polyharmonic homogenization, rough polyharmonic splines and sparse super-localization. *ESAIM: Math. Model. Numer. Anal.*, eFirst, 2013.
55. D. Peterseim. Variational multiscale stabilization and the exponential decay of fine-scale correctors. *ArXiv e-prints*, 1505.07611, 2015. (to appear in G.R. Barrenechea, F. Brezzi, A. Cangiani and E. Georgoulis (Eds.), *Building Bridges: Connections and Challenges in Modern Approaches to Numerical Partial Differential Equations*, Lecture Notes in Computational Science and Engineering, Springer)
56. D. Peterseim and S. Sauter. Finite Elements for Elliptic Problems with Highly Varying, Nonperiodic Diffusion Matrix. *Multiscale Model. Simul.*, 10(3):665-695, 2012.
57. D. Peterseim and R. Scheichl. Rigorous numerical upscaling at high contrast. *in preparation*, 2015+.
58. G. Rozza, D. B. P. Huynh, and A. T. Patera. Reduced basis approximation and a posteriori error estimation for affinely parametrized elliptic coercive partial differential equations: application to transport and continuum mechanics. *Arch. Comput. Methods Eng.*, 15(3):229–275, 2008.
59. L. Ridgway Scott and Shangyou Zhang. Finite element interpolation of nonsmooth functions satisfying boundary conditions. *Math. Comp.*, 54(190):483–493, 1990.
60. Daniel B. Szyld. The many proofs of an identity on the norm of oblique projections. *Numer. Algorithms*, 42(3-4):309–323, 2006.
61. R. Tezaur and C. Farhat. Three-dimensional discontinuous Galerkin elements with plane waves and Lagrange multipliers for the solution of mid-frequency Helmholtz problems. *Internat. J. Numer. Methods Engrg.*, 66(5):796–815, 2006.
62. H. Wu. Pre-asymptotic error analysis of CIP-FEM and FEM for the Helmholtz equation with high wave number. Part I: linear version. *IMA J. Numer. Anal.*, 34(3):1266–1288, 2014.
63. J. Zitelli, I. Muga, L. Demkowicz, J. Gopalakrishnan, D. Pardo, and V.M. Calo. A class of discontinuous Petrov–Galerkin methods. part IV: The optimal test norm and time-harmonic wave propagation in 1D. *J. Comput. Phys.*, 230(7):2406–2432, 2011.

Appendix A
Numerical homogenization beyond scale separation

A.1 Localization of elliptic multiscale problems

Mathematics of Computation **83**(290):2583–2603, 2014.

(with A. Målqvist)

MATHEMATICS OF COMPUTATION
Volume 83, Number 290, November 2014, Pages 2583–2603
S 0025-5718(2014)02868-8
Article electronically published on June 16, 2014

LOCALIZATION OF ELLIPTIC MULTISCALE PROBLEMS

AXEL MÅLQVIST AND DANIEL PETERSEIM

Abstract. This paper constructs a local generalized finite element basis for elliptic problems with heterogeneous and highly varying coefficients. The basis functions are solutions of local problems on vertex patches. The error of the corresponding generalized finite element method decays exponentially with respect to the number of layers of elements in the patches. Hence, on a uniform mesh of size H, patches of diameter $H\log(1/H)$ are sufficient to preserve a linear rate of convergence in H without pre-asymptotic or resonance effects. The analysis does not rely on regularity of the solution or scale separation in the coefficient. This result motivates new and justifies old classes of variational multiscale methods.

1. Introduction

This paper considers the numerical solution of second order elliptic problems with strongly heterogeneous and highly varying (non-periodic) coefficients. The heterogeneities and oscillations of the coefficient may appear on several non-separated scales. It is well known that classical polynomial based finite element methods perform arbitrarily badly for such problems; see e.g. [4]. To overcome this lack of performance, many methods that are based on general (non-polynomial) ansatz functions have been developed. Early works [1, 2], that essentially apply to one-dimensional problems, have been generalized to the multi-dimensional case in several ways during the last fifteen years; see e.g. [7, 13, 14]. In these methods the problem is split into coarse and (possibly several) fine scales. The fine scale effect on the coarse scale is either computed numerically or modeled analytically. The resulting modified coarse problem can then be solved numerically and its solution contains crucial information from the fine scales. Although many of these approaches show promising results in practice, their convergence analysis usually assumes certain periodicity and scale separation.

For problems with general L^∞ coefficient, the paper [3] gives error bounds for a generalized finite element method that involves the solutions of local eigenvalue problems. The construction in [6, 19] depends only on the solution of the original problem on certain subdomains. However, the size of these subdomains strongly depends on the mesh size. This dependence is suboptimal with respect to the

Received by the editor October 4, 2011 and, in revised form, March 22, 2012 and October 18, 2012.

2010 *Mathematics Subject Classification.* Primary 65N12, 65N30.

Key words and phrases. Finite element method, a priori error estimate, convergence, multiscale method.

The first author was supported by The Göran Gustafsson Foundation and The Swedish Research Council.

The second author was supported by the Humboldt-Universtät zu Berlin and the DFG Research Center Matheon Berlin through project C33.

theoretical statement given in [12], that is, for any shape regular mesh of size H there exist $\mathcal{O}\left((\log(1/H))^{d+1}\right)$ local (non-polynomial) basis functions per nodal point such that the error of the corresponding Galerkin solution u_H satisfies the estimate $\|u - u_H\|_{H^1(\Omega)} \leq C_g H$ with a constant C_g that depends on the right-hand side g and the global bounds of the diffusion coefficient but not on its variations. The derivation in [12] is not constructive in the sense that it involves the solution of the (global) original problem with specific right-hand sides.

In this paper, we show that such a (quasi-)optimal basis can indeed be constructed by solving only local problems on element patches. We use a modified nodal basis similar to the one presented in [16] and prove that these basis functions decay exponentially away from the node they are associated with. This exponential decay justifies an approximation using localized patches.

The precise setting of the paper is as follows. Let $\Omega \subset \mathbb{R}^d$ be a bounded Lipschitz domain with polygonal boundary and let the diffusion matrix $A \in L^\infty\left(\Omega, \mathbb{R}^{d\times d}_{\mathrm{sym}}\right)$ be uniformly elliptic:

$$
\begin{aligned}
0 < \alpha(A,\Omega) &:= \operatorname*{ess\,inf}_{x\in\Omega} \inf_{v\in\mathbb{R}^d\setminus\{0\}} \frac{(A(x)v)\cdot v}{v\cdot v},\\
\infty > \beta(A,\Omega) &:= \operatorname*{ess\,sup}_{x\in\Omega} \sup_{v\in\mathbb{R}^d\setminus\{0\}} \frac{(A(x)v)\cdot v}{v\cdot v}.
\end{aligned}
\tag{1.1}
$$

Given $g \in L^2(\Omega)$, we seek $u \in V := H^1_0(\Omega)$ such that

$$
a(u,v) := \int_\Omega (A\nabla u)\cdot\nabla v = \int_\Omega gv =: G(v) \quad \text{for all } v \in V. \tag{1.2}
$$

The bilinear form a is symmetric, coercive, bounded, and hence, (1.2) has a unique solution.

The main result of this paper (cf. Theorem 3.6) shows that the error $u - u^{\mathrm{ms}}_{H,k}$ of the generalized finite element method, which is based on our new (local) basis functions mentioned above, is bounded as follows

$$
\|A^{1/2}\nabla(u - u^{\mathrm{ms}}_{H,k})\|_{L^2(\Omega)} \leq C_g H;
$$

H being the mesh size of the underlying coarse finite element mesh and $k \approx \log(1/H)$ referring to the number of layers of coarse elements that form the support of the localized basis functions. This estimate shows that our new numerical upscaling procedure is reliable beyond strong assumptions like periodicity and scale separation. Moreover, the error bound is stable with respect to perturbations arising from the discretization of the local problems. These results give a theoretical foundation for numerous previous experiments where exponential decay of a similar modified basis have been noticed; see e.g. [18].

The outline of the paper is as follows. In Section 2, we derive a set of local basis functions and define the corresponding multiscale finite element method. The error analysis is done in Section 3. Section 4 is devoted to the discretization of the local problems. Section 5 presents numerical experiments, and Section 6 discusses the application of this theory to state-of-the-art multiscale methods.

2. Local basis

In this section, we design a set of local basis functions for the multiscale problem under consideration. The construction is based on a regular (in the sense of [10]) finite element mesh $\mathcal{T}_H$ of Ω into closed triangles ($d = 2$) or tetrahedra ($d = 3$). Subsection 2.1 recalls the classical nodal basis with respect to $\mathcal{T}_H$ and demonstrates its lack of approximation properties. Subsection 2.2 introduces a quasi-interpolation operator used in the construction of the new basis. Subsection 2.3 defines a modified (coefficient dependent) nodal basis and analyzes its approximation properties. This basis is then localized in Subsection 2.4.

2.1. **Classical nodal basis.** Let $H : \overline{\Omega} \to \mathbb{R}_{>0}$ denote the $\mathcal{T}_H$-piecewise constant mesh size function with $H|_T = \operatorname{diam}(T) =: H_T$ for all $T \in \mathcal{T}_H$. The mesh size may vary in space. In practical applications, the mesh $\mathcal{T}_H$ (resp., its size H) shall be determined by the accuracy which is desired or the computational capacity that is available but *not* by the scales of the coefficient.

The classical (conforming) P_1 finite element space is given by

$$S_H := \{v \in C^0(\bar{\Omega}) \mid \forall T \in \mathcal{T}_H, v|_T \text{ is a polynomial of total degree} \leq 1\}. \tag{2.1}$$

Let $V_H := S_H \cap V$ denote the space of finite element functions that match the homogeneous Dirichlet boundary conditions. Let $\mathcal{N}$ denote the set of interior vertices of $\mathcal{T}_H$. For every vertex $x \in \mathcal{N}$, let $\lambda_x \in S_H$ denote the corresponding nodal basis function (tent function), i.e.,

$$\lambda_x(x) = 1 \text{ and } \lambda_x(y) = 0 \quad \text{for all } y \neq x \in \mathcal{N}.$$

These nodal basis functions form a basis of V_H. The availability of such a local basis is a key property of any finite element method and ensures that the resulting system of linear algebraic equations is sparse.

The (unique) Galerkin approximation $u_H \in V_H$ satisfies

$$a(u_H, v) = G(v) \quad \text{for all } v \in V_H. \tag{2.2}$$

The above method (2.2) is optimal with respect to the energy norm $|||\cdot||| := |||\cdot|||_\Omega := \|A^{1/2}\nabla \cdot\|_{L^2(\Omega)}$ on V which is induced by a,

$$|||u - u_H||| = \min_{v_H \in V_H} |||u - v_H|||. \tag{2.3}$$

Assuming that the solution u is smooth, the combination of (2.3) and standard interpolation error estimates yields the standard a priori error estimate

$$|||u - u_H||| \leq C\|H\|_{L^\infty(\Omega)}\|\nabla^2 u\|_{L^2(\Omega)}.$$

This estimate states linear convergence of the classical finite element method (2.2) as the maximal mesh width tends to zero. However, the regularity assumption is not realistic for the problem class under consideration. Moreover, even if the coefficient is smooth, it may oscillate rapidly, say at frequency ε^{-1} for some small parameter ε. In this case, the asymptotic result is useless because $\nabla^2 u$ may oscillate at the same scale, a fact that is hidden in the constant $\|\nabla^2 u\|_{L^2(\Omega)} \approx \varepsilon^{-1}$. Unless $H \lesssim \varepsilon$, the above finite element space is unable to capture the behavior of the solution neither on the microscopic nor on the macroscopic level. In what follows, we present a new method that resolves this issue.

2.2. **Quasi-interpolation.** The key tool in our construction will be some bounded linear surjective (quasi-) interpolation operator $\mathfrak{I}_H : V \to V_H$. The choice of this operator is not unique and a different choice might lead to a different multiscale method. We have in mind the following modification of Clément's interpolation [11] which is presented and analyzed in [9, Section 6]. Given $v \in V$, $\mathfrak{I}_H v := \sum_{x\in\mathcal{N}}(\mathfrak{I}_H v)(x)\lambda_x$ defines a (weighted) Clément interpolant with nodal values

$$(\mathfrak{I}_H v)(x) := \left(\int_\Omega v\lambda_x\right) / \left(\int_\Omega \lambda_x\right) \tag{2.4}$$

for $x \in \mathcal{N}$. The nodal values are weighted averages of the function over nodal patches $\omega_x := \operatorname{supp}\lambda_x$. Since the summation is taken only with respect to interior vertices $\mathcal{N}$, this operator matches homogeneous Dirichlet boundary conditions.

Recall the (local) approximation and stability properties of the interpolation operator $\mathfrak{I}_H$ [9, Section 6]: There exists a generic constant $C_{\mathfrak{I}_H}$ such that for all $v \in V$ and for all $T \in \mathcal{T}_H$ it holds that

$$H_T^{-1}\|v - \mathfrak{I}_H v\|_{L^2(T)} + \|\nabla(v - \mathfrak{I}_H v)\|_{L^2(T)} \le C_{\mathfrak{I}_H}\|\nabla v\|_{L^2(\omega_T)}, \tag{2.5.a}$$

where $\omega_T := \bigcup\{K \in \mathcal{T}_H \mid T \cap K \neq \emptyset\}$. The constant $C_{\mathfrak{I}_H}$ depends on the shape regularity parameter ρ of the finite element mesh $\mathcal{T}_H$ (see (3.1) below) but not on H_T.

Note that the above interpolation operator is not a projection, i.e., $v_H \in V_H$ does not equal its interpolation $\mathfrak{I}_H v_H$ in general. However, the particular choice gives rise to the following lemma.

Lemma 2.1. *There exists a generic constant $C'_{\mathfrak{I}_H}$ which only depends on ρ but not on the local mesh size H, such that for all $v_H \in V_H$ there exists $v \in V$ with the properties*

$$\mathfrak{I}_H(v) = v_H, \qquad \|\nabla v\| \le C'_{\mathfrak{I}_H}\|\nabla v_H\|, \quad \textit{and} \quad \operatorname{supp} v \subset \operatorname{supp} v_H. \tag{2.5.b}$$

Proof. For every nodal basis function λ_x, $x \in \mathcal{N}$, there is some $b_x \in H_0^1(\omega_x)$ such that $\mathfrak{I}_H(b_x) = \lambda_x$ and $\|\nabla b_x\| \le C''_{\mathfrak{I}_H}\|\nabla\lambda_x\|$ with some constant $C''_{\mathfrak{I}_H}$ that does not depend on x and H. For example, b_x may be chosen as a standard cubic element bubble on an arbitrary element $T \subset \omega_x$ or a quadratic edge/face bubble related to an arbitrary edge/face of $\mathcal{T}_H$ interior to ω_x. One might as well choose b_x to be nodal interpolation of those bubbles in a finite element space that correponds to some uniform refinement of $\mathcal{T}_H$.

Given $v_H = \sum_{x\in\mathcal{N}} v_H(x)\lambda_x \in V_H$, $v := v_H + \sum_{x\in\mathcal{N}}(v_H(x) - (\mathfrak{I}_H v_H)(x))\, b_x \in V$ has the desired properties (for suitably chosen b_x). The interpolation and support properties are obvious. The stability follows from

$$\begin{aligned}
\|\nabla v\|^2 &\le C\left(\|\nabla v_H\|^2 + \sum_{x\in\mathcal{N}} |v_H(x) - (\mathfrak{I}_H v_H)(x)|^2\, \|\nabla b_x\|^2\right)\\
&\le C\left(\|\nabla v_H\|^2 + C''^2_{\mathfrak{I}_H} \sum_{x\in\mathcal{N}} |v_H(x) - (\mathfrak{I}_H v_H)(x)|^2\, \|\nabla\lambda_x\|^2\right)\\
&\le C\left(\|\nabla v_H\|^2 + C' C''^2_{\mathfrak{I}_H} \sum_{T\in\mathcal{T}_H} \|v_H - \mathfrak{I}_H v_H\|^2_{L^2(T)} H_T^{-2}\right)\\
&\le C\left(\|\nabla v_H\|^2 + C' C^2_{\mathfrak{I}_H} C''^2_{\mathfrak{I}_H} \sum_{T\in\mathcal{T}_H} \|\nabla v_H\|^2_{L^2(\omega_T)}\right)\\
&\le C'^2_{\mathfrak{I}_H}\|\nabla v_H\|^2,
\end{aligned}$$

where we use $\|\nabla\lambda_x\|^2 \approx |\operatorname{supp}\lambda_x|^{(d-2)}$, the inverse inequality $\|v_H - \mathfrak{I}_H v_H\|^2_{L^\infty(T)} \lesssim H_T^{-d}\|v_H - \mathfrak{I}_H v_H\|^2_{L^2(T)}$, and (2.5.a). □

In the forthcoming derivation of our method, the interpolation operator (2.4) may be replaced by any linear bounded surjective operator that satisfies (2.5.a)–(2.5.b). Hereby, (2.5.b) may be relaxed in the sense that $\operatorname{supp} v$ is not necessarily a subset of $\operatorname{supp} v_H$ but that $\operatorname{supp} v \setminus \operatorname{supp} v_H$ covers at most a fixed (small) number of element layers about $\operatorname{supp} v_H$.

2.3. **Multiscale splitting and modified nodal basis.** Let $\mathfrak{I}_H : V \to V_H$ be a quasi-interpolation operator according to the previous subsection. Then the kernel of $\mathfrak{I}_H$,

$$V^{\mathrm{f}} := \{v \in V \mid \mathfrak{I}_H v = 0\},$$

represents the microscopic features of V, i.e., all features that are not captured by V_H. Given $v \in V_H$, define $\mathfrak{F}v \in V^{\mathrm{f}}$ by

$$a(\mathfrak{F}v, w) = a(v, w) \quad \text{for all } w \in V^{\mathrm{f}}.$$

The finescale projection operator $\mathfrak{F} : V_H \to V^{\mathrm{f}}$ leads to an orthogonal splitting with respect to the scalar product a:

$$V = V_H^{\mathrm{ms}} \oplus V^{\mathrm{f}} \quad \text{where} \quad V_H^{\mathrm{ms}} := (V_H - \mathfrak{F}V_H).$$

Hence, any function $u \in V$ can be decomposed into $u_H^{\mathrm{ms}} \in V_H^{\mathrm{ms}}$ and $u^{\mathrm{f}} \in V^{\mathrm{f}}$, $u = u_H^{\mathrm{ms}} + u^{\mathrm{f}}$, with $a(u_H^{\mathrm{ms}}, u^{\mathrm{f}}) = 0$. Since $\dim V_H^{\mathrm{ms}} = \dim V_H$, the space V_H^{ms} can be regarded as a modified coarse space. The superscript "ms" abbreviates "multiscale" and indicates that V_H^{ms}, in addition, contains fine scale information. The corresponding Galerkin approximation $u_H^{\mathrm{ms}} \in V_H^{\mathrm{ms}}$ satisfies

$$a(u_H^{\mathrm{ms}}, v) = G(v) \quad \text{for all } v \in V_H^{\mathrm{ms}}. \tag{2.6}$$

The error $(u - u_H^{\mathrm{ms}})$ of the above method (2.6) is analyzed in Section 3.1.

Finally, we shall introduce a basis of V_H^{ms}. The image of the nodal basis function λ_x under the fine scale projection $\mathfrak{F}$ is denoted by $\phi_x = \mathfrak{F}\lambda_x \in V^{\mathrm{f}}$, i.e., ϕ_x satisfies the corrector problem

$$a(\phi_x, w) = a(\lambda_x, w) \quad \text{for all } w \in V^{\mathrm{f}}. \tag{2.7}$$

We emphasize that the corrector problem is posed in the fine scale space V^{f}, i.e., test and trial functions satisfy the constraint that their interpolation with respect to the coarse mesh vanishes.

A basis of V_H^{ms} is then given by the modified nodal basis

$$\{\lambda_x - \phi_x \mid x \in \mathcal{N}\}. \tag{2.8}$$

In general, the corrections ϕ_x of nodal basis functions λ_x, $x \in \mathcal{N}$, have global support, a fact which limits the practical use of the modified basis (2.8) and the corresponding method (2.6).

2.4. **Localization.** In Section 3.2, we will show that the correction ϕ_x decays exponentially fast away from x. Hence, simple truncation of the corrector problems to local patches of coarse elements yields localized basis functions with good approximation properties.

Let $k \in \mathbb{N}$. Define nodal patches of k-th order $\omega_{x,k}$ about $x \in \mathcal{N}$ by

$$\begin{aligned} \omega_{x,1} &:= \operatorname{supp} \lambda_x = \operatorname{int}\left(\bigcup \{T \in \mathcal{T}_H \mid x \in T\}\right), \\ \omega_{x,k} &:= \operatorname{int}\left(\bigcup \{T \in \mathcal{T}_H \mid T \cap \overline{\omega}_{x,k-1} \neq \emptyset\}\right), \quad k = 2,3,4\ldots. \end{aligned} \tag{2.9}$$

Define localized finescale spaces $V^{\mathrm{f}}(\omega_{x,k}) := \{v \in V^{\mathrm{f}} \mid v|_{\Omega\setminus\omega_{x,k}} = 0\}$, $x \in \mathcal{N}$, by intersecting V^{f} with those functions that vanish outside the patch $\omega_{x,k}$. The solutions $\phi_{x,k} \in V^{\mathrm{f}}(\omega_{x,k})$ of

$$a(\phi_{x,k}, w) = a(\lambda_x, w) \quad \text{for all } w \in V^{\mathrm{f}}(\omega_{x,k}), \tag{2.10}$$

are approximations of ϕ_x from (2.7) with local support.

We define localized multiscale finite element spaces

$$V^{\mathrm{ms}}_{H,k} = \operatorname{span}\{\lambda_x - \phi_{x,k} \mid x \in \mathcal{N}\} \subset V. \tag{2.11.a}$$

The corresponding multiscale approximation of (1.2) reads: find $u^{\mathrm{ms}}_{H,k} \in V^{\mathrm{ms}}_{H,k}$ such that

$$a(u^{\mathrm{ms}}_{H,k}, v) = G(v) \quad \text{for all } v \in V^{\mathrm{ms}}_{H,k}. \tag{2.11.b}$$

Note that $\dim V^{\mathrm{ms}}_{H,k} = |\mathcal{N}| = \dim V_H$, i.e., the number of degrees of freedom of the proposed method (2.11) is the same as for the classical method (2.2). The basis functions of the multiscale method have local support. The overlap is proportional to the parameter k. The error analysis of Section 3.2 suggests to choose $k \approx \log \frac{1}{H}$.

Remark 2.2. The localized modified basis functions could be localized further to vertex patches ω_x, $x \in \mathcal{N}$, by simply multiplying them with the classical nodal basis functions; for any $x \in \mathcal{N}$ and any $y \in \mathcal{N} \cap \omega_{x,k}$, define $\phi^y_x := \lambda_y \phi_{x,k}$. The generalized finite element space which is spanned by those $\mathcal{O}\left((\log(1/H))^d\right)$ local basis functions per vertex has similar approximation properties as $V^{\mathrm{ms}}_{H,k}$ (see [5]).

3. Error analysis

This section analyzes the proposed multiscale method in two steps. First, Subsection 3.1 presents an error bound for the idealized method (2.6). Then, Subsection 3.2 bounds the error of truncation to local patches and proves the main result, that is, an error bound for the multiscale method (2.11).

As usual, the error analysis depends on the constant $\rho > 0$ which represents shape regularity of the finite element mesh $\mathcal{T}_H$;

$$\rho := \max_{T \in \mathcal{T}_H} \rho_T \quad \text{with} \quad \rho_T := \frac{\operatorname{diam} B_T}{\operatorname{diam} T} \ \text{ for } T \in \mathcal{T}_H, \tag{3.1}$$

where B_T denotes the largest ball contained in T.

3.1. **Discretization error.**

Lemma 3.1. *Let $u \in V$ solve (1.2) and $u_H^{\mathrm{ms}} \in V_H^{\mathrm{ms}}$ solve (2.6). Then it holds that*

$$|||u - u_H^{\mathrm{ms}}||| \leq C_{\mathrm{ol}}^{1/2} C_{\mathfrak{I}_H} \alpha^{-1/2} \|Hg\|_{L^2(\Omega)}$$

with constants C_{ol} and $C_{\mathfrak{I}_H}$ that only depend on ρ.

Proof. Recall the (local) approximation and stability properties (2.5.a) of the interpolation operator $\mathfrak{I}_H$. Due to the splitting from Section 2.3, it holds that $u - u_H^{\mathrm{ms}} = u^{\mathrm{f}}$. Since $\mathfrak{I}_H u^{\mathrm{f}} = 0$, the application of (2.5.a) and Young's inequality yield

$$\begin{aligned} \big|\big|\big|u^{\mathrm{f}}\big|\big|\big|^2 = G(u^{\mathrm{f}}) &\leq \sum_{T \in \mathcal{T}_H} \|g\|_{L^2(T)} \|u^{\mathrm{f}} - \mathfrak{I}_H u^{\mathrm{f}}\|_{L^2(T)} \\ &\leq \frac{C_{\mathfrak{I}_H}^2}{2\epsilon\alpha} \|Hg\|_{L^2(\Omega)}^2 + \frac{\epsilon}{2} \sum_{T \in \mathcal{T}_H} \|A^{1/2} \nabla u^{\mathrm{f}}\|_{L^2(\omega_T)}^2 \end{aligned}$$

for any $\epsilon > 0$. Note that there exists a constant $C_{\mathrm{ol}} > 0$ that only depends on ρ such that the number of elements covered by ω_T is uniformly (w.r.t. T) bounded by C_{ol}. The choice $\epsilon = C_{\mathrm{ol}}^{-1}$ concludes the proof. □

Remark 3.2. Substituting $\mathfrak{I}_H$ by the modified Clément interpolation operator presented in [8] allows one to improve the error estimate in Lemma (3.1). The term $\|Hg\|_{L^2(\Omega)}$ can be replaced by data oscillations $(\sum_{x \in \mathcal{N}} \|H(g - g_x)\|_{L^2(\omega_x)}^2)^{1/2}$ with some weighted averages g_x of g on ω_x, $x \in \mathcal{N}$; we refer to [8, Section 2] for details. Additional smoothness of the right-hand side $g \in H^1(\Omega)$ then leads to quadratic convergence of the idealized method without localization.

3.2. **Error of localized multiscale FEM.** First, we estimate the error due to truncation to local patches. We will frequently make use of cut-off functions on element patches.

Definition 3.3. For $x \in \mathcal{N}$ and $m < M \in \mathbb{N}$, let $\eta_x^{m,M} : \Omega \to [0,1]$ be a continuous and weakly differentiable function such that

$$(\eta_x^{m,M})|_{\omega_{x,m}} = 0, \tag{3.2.a}$$

$$(\eta_x^{m,M})|_{\Omega \setminus \omega_{x,M}} = 1, \text{ and} \tag{3.2.b}$$

$$\forall T \in \mathcal{T}_H,\ \|\nabla \eta_x^{m,M}\|_{L^\infty(T)} \leq C_{\mathrm{co}} (M-m)^{-1} H_T^{-1} \tag{3.2.c}$$

with some constant C_{co} that only depends on ρ. For example, one may choose $\eta_x^{m,M} \in S_H$ with nodal values

$$\begin{aligned} &\eta_x^{m,M}(x) = 0 \quad \text{for all } x \in \mathcal{N} \cap \omega_m, \\ &\eta_x^{m,M}(x) = 1 \quad \text{for all } x \in \mathcal{N} \cap (\Omega \setminus \omega_{x,M}), \text{ and} \\ &\eta_x^{m,M}(x) = j(M-m)^{-1} \quad \text{for all } x \in \mathcal{N} \cap \partial\omega_{x,m+j},\ j = 0, 1, 2, \ldots, M-m. \end{aligned} \tag{3.3}$$

We prove the essential decay property of the corrector functions by some iterative Caccioppoli-type argument. Recall the notation $|||\cdot|||_\omega := \|A^{1/2} \nabla \cdot\|_{L^2(\omega)}$.

Lemma 3.4. *For all $x \in \mathcal{N}$, $k, \ell \geq 2 \in \mathbb{N}$, the estimate*

$$|||\phi_x - \phi_{x,\ell k}||| \leq C_2 \left(\frac{C_1}{\ell}\right)^{\frac{k-2}{2}} |||\phi_x|||_{\omega_{x,\ell}}$$

holds with constants C_1, C_2 that only depend on ρ and β/α but not on x, k, ℓ, or H.

Proof. Let $x \in \mathcal{N}$ and $\ell, k \geq 2 \in \mathbb{N}$. Observe that

$$|||\phi_x - \phi_{x,\ell k}|||^2 \leq |||\phi_x - v|||^2 = |||\phi_x - v|||^2_{\omega_{x,\ell k}} + |||\phi_x|||^2_{\Omega\setminus\omega_{x,\ell k}}\,, \tag{3.4}$$

holds for all $v \in V^{\mathrm{f}}(\omega_{x,\ell k})$ using Galerkin orthogonality.

Let $\zeta_x := 1 - \eta_x^{\ell(k-1)+1,\ell k-1}$ with a cutoff function $\eta_x^{\ell(k-1)+1,\ell k-1}$ as in Definition 3.3. According to (2.5.b), there exists $b_x \in V$ such that $\mathfrak{I}_H(b_x) = \mathfrak{I}_H(\zeta_x\phi_x)$, $|||b_x||| \leq C'_{\mathfrak{I}_H} |||\mathfrak{I}_H(\zeta_x\phi_x)|||$, and $\operatorname{supp}(b_x) \subset \omega_{x,\ell k}$. Hence, $v := \zeta_x\phi_x - b_x \in V^{\mathrm{f}}(\omega_{x,\ell k})$ and

$$\begin{aligned}|||\phi_x - v|||_{\omega_{x,\ell k}} &\leq |||\phi_x - \zeta_x\phi_x|||_{\omega_{x,\ell k}\setminus\omega_{x,\ell(k-1)+1}} + |||b_x|||_{\omega_{x,\ell k}\setminus\omega_{x,\ell(k-1)}}\\ &\leq C'_{\mathfrak{I}_H} C_{\mathfrak{I}_H}\left(|||\phi_x|||_{\omega_{x,\ell k}\setminus\omega_{x,\ell(k-1)+1}} + \sqrt{\beta}\|\nabla(\zeta_x\phi_x)\|_{L^2(\omega_{x,\ell k}\setminus\omega_{x,\ell(k-1)})}\right).\end{aligned}$$

Since $\mathfrak{I}_H\phi_x = 0$, the upper bound of the interpolation error (2.5.a) and (3.2.c) yield

$$\begin{aligned}&\|\nabla(\zeta_x\phi_x)\|^2_{L^2(\omega_{x,\ell k}\setminus\omega_{x,\ell(k-1)})}\\ &\qquad\leq C_2''' \sum_{T\in\mathcal{T}_H:\, T\subset\overline{\omega}_{x,\ell k}\setminus\omega_{x,\ell(k-1)+1}} \left(H_T^2\|\nabla\zeta_k\|^2_{L^\infty(T)} + \|\zeta_k\|^2_{L^\infty(T)}\right)\|\nabla\phi_x\|^2_{L^2(T)}\\ &\qquad\leq C_2''\alpha^{-1}\,|||\phi_x|||^2_{\omega_{x,\ell k}\setminus\omega_{x,\ell(k-1)+1}}\end{aligned}$$

with $C_2'' := 1 + C_{\mathrm{ol}}C_{\mathrm{co}}^2C_{\mathfrak{I}_H}^2$. This leads to

$$|||\phi_x - v|||_{\omega_{x,\ell k}} \leq C'_2\,|||\phi_x|||_{\omega_{x,k\ell}\setminus\omega_{x,(k-1)\ell}}\,, \tag{3.5}$$

where C'_2 depends only on ρ and $\sqrt{\beta/\alpha}$. The combination of (3.4), with $v = \zeta_x\phi_x - b_x$, and (3.5) yields

$$|||\phi_x - \phi_{x,\ell k}||| \leq C_2\,|||\phi_x|||_{\Omega\setminus\omega_{x,\ell(k-1)}}\,. \tag{3.6}$$

Further estimation of the right-hand side in (3.6) is possible using cut-off functions $\eta_j := \eta_x^{\ell(j-1)+1,\ell j}$ (cf. Definition 3.3), $j = 2, 3, \dots, k-1$. Observe that

$$\begin{aligned}\|A^{1/2}\nabla\phi_x\|^2_{L^2(\Omega\setminus\omega_{x,\ell(k-1)})} &\leq \|A^{1/2}\eta_{k-1}\nabla\phi_x\|^2_{L^2(\Omega)}\\ &= \int_\Omega (A\nabla\phi_x)\cdot\nabla(\eta_{k-1}^2\phi_x) \;-\; 2\int_\Omega \eta_{k-1}\phi_x(A\nabla\phi_x)\cdot\nabla\eta_{k-1}.\end{aligned} \tag{3.7}$$

Let, according to Lemma 2.1, $b_{x,(k-1)}$ be chosen such that $\mathfrak{I}_H b_{x,(k-1)} = \mathfrak{I}_H(\eta_{k-1}^2\phi_x)$. Then $\eta_{k-1}^2\phi_x - b_{x,(k-1)} \in V^{\mathrm{f}}$. Since $|\operatorname{supp}(\nabla\lambda_x)\cap\operatorname{supp}(\eta_{k-1})| = 0$ and $\operatorname{supp}(\nabla\eta_{k-1}) = \omega_{x,(k-1)\ell}\setminus\omega_{x,(k-2)\ell+1}$, the first term on the right-hand side of (3.7) can be rewritten as

$$\begin{aligned}&\int_\Omega (A\nabla\phi_x)\cdot\nabla(\eta_{k-1}^2\phi_x)\\ &\quad= \int_\Omega (A\nabla\phi_x)\cdot\nabla(\eta_{k-1}^2\phi_x - b_{x,(k-1)}) + \int_\Omega (A\nabla\phi_x)\cdot\nabla b_{x,(k-1)}\\ &\quad= \int_\Omega (A\nabla\phi_x)\cdot\nabla b_{x,(k-1)}\\ &\quad\leq C'_{\mathfrak{I}_H}\sqrt{\beta}\,|||\phi_x|||_{\omega_{x,(k-1)\ell}\setminus\omega_{x,(k-2)\ell+1}}\,\|\nabla\mathfrak{I}_H(\eta_{k-1}^2\phi_x)\|_{L^2(\omega_{x,(k-1)\ell}\setminus\omega_{x,(k-2)\ell+1})}.\end{aligned} \tag{3.8}$$

With $\overline{\eta_T^2} := |T|^{-1} \int_T \eta_{k-1}^2$ we have

$$\begin{aligned}
&\|\nabla \mathfrak{I}_H(\eta_{k-1}^2 \phi_x)\|_{L^2(T)} = \|\nabla \mathfrak{I}_H((\eta_{k-1}^2 - \overline{\eta_T^2})\phi_x)\|_{L^2(T)} \\
&\quad \le C_{\mathfrak{I}_H} \|\nabla((\eta_{k-1}^2 - \overline{\eta_T^2})\phi_x)\|_{L^2(T)} \\
&\quad \le C_{\mathfrak{I}_H} \left(\|\eta_{k-1}^2 - \overline{\eta_T^2}\|_{L^\infty(T)} \|\nabla \phi_x\|_{L^2(T)} + \|\nabla(\eta_{k-1}^2)\|_{L^\infty(T)} \|\phi_x\|_{L^2(T)} \right) \\
&\quad \le 2 C_{\mathfrak{I}_H} \|\nabla(\eta_{k-1})\|_{L^\infty(T)} \left(\alpha^{-1/2} \operatorname{diam}(T) \, |||\phi_x|||_T + \|\phi_x - \mathfrak{I}_H(\phi_x)\|_{L^2(T)} \right).
\end{aligned}$$

Thus, the property (3.2.c) of the cutoff function and the upper bound of the interpolation error (2.5.a) yield

$$\left|\left|\left| \mathfrak{I}_H(\eta_{k-1}^2 \phi_x) \right|\right|\right|_{\omega_{x,(k-1)\ell} \setminus \omega_{x,(k-2)\ell+1}} \le C_1' \ell^{-1} \|A^{1/2} \nabla \phi_x\|_{L^2(\Omega \setminus \omega_{x,(k-2)\ell})}, \tag{3.9}$$

where C_1' only depends on $C_{\mathfrak{I}_H}$, C_{co}, C_{ol}, and $\sqrt{\beta/\alpha}$. The same arguments allow one to bound the second term on the right-hand side in (3.7),

$$\begin{aligned}
&2 \int_\Omega \eta_{k-1} \phi_x (A \nabla \phi_x) \cdot \nabla \eta_{k-1} \\
&\quad \le 2 \sum_{T \in \mathcal{T}_H : \, T \subset \overline{\omega}_{x,(k-1)\ell} \setminus \omega_{x,(k-2)\ell+1}} \|\nabla \eta_{k-1}\|_{L^\infty(T)} \|A^{1/2} \nabla \phi_x\|_{L^2(T)} \|A^{1/2} \phi_x\|_{L^2(T)} \\
&\quad \le C_1'' \ell^{-1} \|A^{1/2} \nabla \phi_x\|^2_{L^2(\Omega \setminus \omega_{x,(k-2)\ell})},
\end{aligned} \tag{3.10}$$

where C_1'' only depends on $C_{\mathfrak{I}_H}$, C_{co}, and $\sqrt{\beta/\alpha}$. The combination of (3.7)–(3.10) yields

$$|||\phi_x|||^2_{\Omega \setminus \omega_{x,(k-1)\ell}} \le C_1 \ell^{-1} \, |||\phi_x|||^2_{\Omega \setminus \omega_{x,(k-2)\ell}}, \tag{3.11}$$

where $C_1 := C_1' + C_1''$. For $j = k-2, \dots, 2$, a similar argument (with η_{k-1} replaced by η_j) yields

$$|||\phi_x|||^2_{\Omega \setminus \omega_{x,j\ell}} \le C_1 \ell^{-1} \, |||\phi_x|||^2_{\Omega \setminus \omega_{x,(j-1)\ell}}. \tag{3.12}$$

Starting from (3.11), the successive application of (3.12) for $j = k-2, k-3, \dots, 2$ proves

$$|||\phi_x|||^2_{\Omega \setminus \omega_{x,(k-1)\ell}} \le (C_1 \ell^{-1})^{k-2} \, |||\phi_x|||^2_{\omega_{x,\ell}}. \tag{3.13}$$

Combining (3.6) and (3.13), we finally obtain the assertion. □

Lemma 3.5. *There is a constant C_3 that depends only on ρ and β/α, but not on $|\mathcal{N}|$, k, or ℓ such that*

$$\left|\left|\left| \sum_{x \in \mathcal{N}} v(x)(\phi_x - \phi_{x,\ell k}) \right|\right|\right|^2 \le C_3 (\ell k)^d \sum_{x \in \mathcal{N}} v^2(x) \, |||\phi_x - \phi_{x,\ell k}|||^2.$$

Proof. For $x \in \mathcal{N}$, let $\zeta_x = 1 - \eta_x^{\ell k+1, \ell k+2}$ (cf. Definition 3.3). By Lemma 2.1 there exists a function $b_x \in V$ such that for any $w \in V^{\mathrm{f}}$ it holds that

$$\mathfrak{I}_H b_x = \mathfrak{I}_H((1-\zeta_x)w), \operatorname{supp}(b_x) \subset \operatorname{supp}(\mathfrak{I}_H((1-\zeta_x)w)) \subset \omega_{x,\ell k+3} \setminus \omega_{x,\ell k},$$

and

$$|||b_x|||_{\omega_{x,\ell k+3} \setminus \omega_{x,\ell k}} \le C'_{\mathfrak{I}_H} \, |||\mathfrak{I}_H((1-\zeta_x)w)|||_{\omega_{x,\ell k+3} \setminus \omega_{x,\ell k}}.$$

We note that $w - \zeta_x w - b_x \in V^{\mathrm{f}}$ with support outside $\omega_{x,\ell k}$, i.e., $a(\phi_x, w - \zeta_x w - b_x) = a(\lambda_x, w - \zeta_x w - b_x) = 0$ and $a(\phi_{x,\ell k}, w - \zeta_x w - b_x) = 0$. With $w = \sum_{x\in\mathcal{N}} v(x)(\phi_x - \phi_{x,\ell k}) \in V^{\mathrm{f}}$ we have

$$
\begin{aligned}
|||w|||^2 &= \sum_{x\in\mathcal{N}} v(x)\, a(\phi_x - \phi_{x,\ell k}, \zeta_x w + b_x) \\
&\leq \sqrt{\beta} \sum_{x\in\mathcal{N}} |v(x)|\, |||\phi_x - \phi_{x,\ell k}||| \cdot \|\nabla(\zeta_x w)\|_{L^2(\Omega)} \\
&\quad + \sqrt{\beta} \sum_{x\in\mathcal{N}} |v(x)|\, |||\phi_x - \phi_{x,\ell k}||| \cdot C'_{\mathfrak{I}_H} \|\nabla(\mathfrak{I}_H((1-\zeta_x)w))\|_{L^2(\omega_{x,\ell k+3})} \\
&\leq 2\sqrt{\beta} C'_{\mathfrak{I}_H} C_{\mathfrak{I}_H} \sum_{x\in\mathcal{N}} |v(x)|\, |||\phi_x - \phi_{x,\ell k}||| \cdot \|\nabla(\zeta_x w)\|_{L^2(\Omega)} \\
&\quad + 2\sqrt{\beta} C'_{\mathfrak{I}_H} C_{\mathfrak{I}_H} \sum_{x\in\mathcal{N}} |v(x)|\, |||\phi_x - \phi_{x,\ell k}||| \cdot \|\nabla w\|_{L^2(\omega_{x,\ell k+4})} \\
&\leq 2\sqrt{\beta} C'_{\mathfrak{I}_H} C_{\mathfrak{I}_H} \sum_{x\in\mathcal{N}} |v(x)|\, |||\phi_x - \phi_{x,\ell k}||| \cdot \|(\nabla\zeta_x)(1-\mathfrak{I}_H)w)\|_{L^2(\omega_{x,\ell k+2})} \\
&\quad + 2\sqrt{\tfrac{\beta}{\alpha}} C'_{\mathfrak{I}_H} C_{\mathfrak{I}_H} \sum_{x\in\mathcal{N}} |v(x)|\, |||\phi_x - \phi_{x,\ell k}||| \cdot |||w|||_{\omega_{x,\ell k+4}} \\
&\leq 4\sqrt{\tfrac{\beta}{\alpha}} C'_{\mathfrak{I}_H} C^2_{\mathfrak{I}_H} C_{\mathrm{co}} \sum_{x\in\mathcal{N}} |v(x)|\, |||\phi_x - \phi_{x,\ell k}||| \cdot |||w|||_{\omega_{x,\ell k+4}} \\
&\leq 4\sqrt{\tfrac{\beta}{\alpha}} C'_{\mathfrak{I}_H} C^2_{\mathfrak{I}_H} C_{\mathrm{co}} C_{\mathrm{ov}} (\ell k)^{d/2} \left(\sum_{x\in\mathcal{N}} v^2(x)\, |||\phi_x - \phi_{x,\ell k}|||^2 \right)^{1/2} |||w|||\,,
\end{aligned}
$$

where $C_{\mathrm{ov}}(\ell k)^d$ represents an upper bound on the number of patches $\omega_{x,\ell k}$ that overlap a single element in the mesh. The result follows by dividing by $|||w|||$ on both sides. □

Theorem 3.6. *Let $u \in V$ solve* (1.2) *and, given $\ell, k \geq 2 \in \mathbb{N}$, let $u^{\mathrm{ms}}_{H,\ell k} \in V^{ms}_{H,\ell k}$ solve* (2.11)*. Then*

$$
\begin{aligned}
|||u - u^{\mathrm{ms}}_{H,\ell k}||| \leq & C_4 \|H_T^{-1}\|_{L^\infty(\Omega)} (\ell k)^{d/2} (C_1/\ell)^{\frac{k-2}{2}} \|g\|_{H^{-1}(\Omega)} \\
& + C_{\mathrm{ol}}^{1/2} C_{\mathfrak{I}_H} \alpha^{-1/2} \|Hg\|_{L^2(\Omega)}
\end{aligned}
$$

holds with C_1 from Lemma 3.4 *and a constant C_4 that depends on α, β and ρ but not on H, k, ℓ, g, or u.*

Proof. Let $\tilde{u}^{\mathrm{ms}}_{H,\ell k} := \sum_{x\in\mathcal{N}} u^{\mathrm{ms}}_H(x)\,(\lambda_x - \phi_{x,\ell k})$, where $u^{\mathrm{ms}}_H(x)$, $x \in \mathcal{N}$, are the coefficients in the basis representation of u^{ms}_H. Due to Galerkin orthogonality, Lemma 3.1, Lemma 3.5, and the triangle inequality,

$$
\begin{aligned}
|||u - u^{\mathrm{ms}}_{H,\ell k}||| &\leq |||u - \tilde{u}^{\mathrm{ms}}_{H,\ell k}||| = |||u - u^{\mathrm{ms}}_H + u^{\mathrm{ms}}_H - \tilde{u}^{\mathrm{ms}}_{H,\ell k}||| \\
&\leq C_{\mathrm{ol}}^{1/2} C_{\mathfrak{I}_H} \alpha^{-1/2} \|Hg\|_{L^2(\Omega)} + |||u^{\mathrm{ms}}_H - \tilde{u}^{\mathrm{ms}}_{H,\ell k}|||\,.
\end{aligned} \tag{3.14}
$$

The application of Lemma 3.4 yields

$$\begin{aligned}\left|\!\left|\!\left| u_H^{\mathrm{ms}} - \tilde{u}_{H,\ell k}^{\mathrm{ms}} \right|\!\right|\!\right|^2 &\leq C_3(\ell k)^d \sum_{x\in\mathcal{N}} u_H^{\mathrm{ms}}(x)^2 \left|\!\left|\!\left| \phi_x - \phi_{x,\ell k} \right|\!\right|\!\right|^2 \\ &\leq C_3(\ell k)^d C_2^2 (C_1/\ell)^{k-2} \sum_{x\in\mathcal{N}} u_H^{\mathrm{ms}}(x)^2 \left|\!\left|\!\left| \phi_x \right|\!\right|\!\right|_{\omega_{x,\ell}}^2 .\end{aligned}$$

Furthermore, we have

$$\begin{aligned}\sum_{x\in\mathcal{N}} u_H^{\mathrm{ms}}(x)^2 \left|\!\left|\!\left| \phi_x \right|\!\right|\!\right|_{\omega_{x,l}}^2 &\leq \beta C_{\mathrm{inv}} \sum_{T\in\mathcal{T}} H_T^{-2} \sum_{x\in T\cap\mathcal{N}} u_H^{\mathrm{ms}}(x)^2 \|\lambda_x\|_{L^2(T)}^2 \\ &\leq \beta C'_{\mathrm{inv}} \sum_{T\in\mathcal{T}} H_T^{-2} \Big\| \sum_{x\in T\cap\mathcal{N}} u_H^{\mathrm{ms}}(x)\lambda_x \Big\|_{L^2(T)}^2 \\ &= \beta C'_{\mathrm{inv}} \Big\| H^{-2} \sum_{x\in\mathcal{N}} u_H^{\mathrm{ms}}(x)\lambda_x \Big\|_{L^2(\Omega)}^2 \\ &\leq \beta C'_{\mathrm{inv}} \left(\|H^{-2} u_H^{\mathrm{ms}}\|_{L^2(\Omega)}^2 + \Big\| H^{-2} \sum_{x\in\mathcal{N}} u_H^{\mathrm{ms}}(x)(\phi_x - \mathfrak{I}_H\phi_x) \Big\|_{L^2(\Omega)}^2 \right) \\ &\leq \tfrac{\beta}{\alpha} C'_{\mathrm{inv}} (C_{\mathrm{F}} \|H_T^{-2}\|_{L^\infty(\Omega)} + C_{\mathfrak{I}_H}) \left|\!\left|\!\left| u_H^{\mathrm{ms}} \right|\!\right|\!\right|^2 ,\end{aligned}$$

where C_{inv} and C'_{inv} depend on ρ and $C_{\mathrm{F}} = C_{\mathrm{F}}(\Omega)$ is the constant from Friedrichs' inequality. This yields

$$\begin{aligned}\left|\!\left|\!\left| u_H^{\mathrm{ms}} - \tilde{u}_{H,\ell k}^{\mathrm{ms}} \right|\!\right|\!\right| &\leq C'_4 \|H_T^{-1}\|_{L^\infty(\Omega)} (\ell k)^{d/2} (C_1/\ell)^{(k-2)/2} \left|\!\left|\!\left| u_H^{\mathrm{ms}} \right|\!\right|\!\right| \\ &\leq C_4 \|H_T^{-1}\|_{L^\infty(\Omega)} (\ell k)^{d/2} (C_1/\ell)^{(k-2)/2} \|g\|_{H^{-1}(\Omega)},\end{aligned} \tag{3.15}$$

where C_4 only depends on C_2, C_3, C'_{inv}, C_{F}, $C_{\mathfrak{I}_H}$, and $\sqrt{\beta}/\alpha$. The assertion follows readily by combining (3.14) and (3.15). □

Remark 3.7. The error estimate in Theorem 3.6 contains a factor $\|H^{-1}\|_{L^\infty(\Omega)}$. However, its influence on the total error can be controlled by choosing the localization parameter k proportional to $\log(1/\|H^{-1}\|_{L^\infty(\Omega)})$. For non-uniform meshes, it is recommended to vary the choice of the localization parameter in space according to $k \approx \log \frac{1}{H}$. We neglect this opportunity to avoid overloading the paper.

4. Discretization of the fine scale computations

In this section, we focus on how to compute numerical approximations to the local basis functions $\lambda_x - \phi_{x,\ell k}$ and thereby to the multiscale solution $u_{H,\ell k}^{\mathrm{ms}}$. In order to do this, we need to extend the error analysis of Section 3 to a fully discrete setting. There is a lot of freedom in choosing different finite elements and different refinement strategies; see e.g. [16, 17]. We will focus on a very simple and natural approach. We assume that the local basis functions are computed using subgrids of a fine scale reference mesh, which is a (possibly space adaptive) refinement of the coarse grid $\mathcal{T}_H$.

More precisely, let $\mathcal{T}_h$ be the result of one uniform refinement and several conforming but possibly non-uniform refinements of the coarse mesh $\mathcal{T}_H$. We introduce $h : \overline{\Omega} \to \mathbb{R}_{>0}$ as the $\mathcal{T}_h$-piecewise constant mesh width function with $h_t := h|_t = \mathrm{diam}(t)$ for all $t \in \mathcal{T}_h$. We construct the finite element space

$$S_h := \{v \in C^0(\Omega) \mid \forall t \in \mathcal{T}_h(\Omega), v|_t \text{ is a polynomial of total degree} \leq 1\}.$$

We let $u_h \in V_h := S_h \cap H^1_0(\Omega)$ be the reference solution that satisfies

$$a(u_h, v) = G(v) \quad \text{for all } v \in V_h. \tag{4.1}$$

Locally on each patch we let

$$V_h^{\mathrm{f}}(\omega_{x,k}) := V^{\mathrm{f}}(\omega_{x,k}) \cap V_h = \{v \in V_h \mid \mathfrak{I}_H v = 0 \text{ and } v|_{\Omega\setminus\omega_{x,k}} = 0\}. \tag{4.2}$$

The numerical approximation $\phi^h_{x,k} \in V_h^{\mathrm{f}}(\omega_{x,k})$ of the corrector $\phi^h_{x,k}$ is determined by

$$a(\phi^h_{x,k}, w) = a(\lambda_x, w) \quad \text{for all } w \in V_h^{\mathrm{f}}(\omega_{x,k}).$$

We denote the discrete multiscale finite element space

$$V^{\mathrm{ms},h}_{H,k} = \operatorname{span}\{\lambda_x - \phi^h_{x,k} \mid x \in \mathcal{N}\}.$$

The corresponding discrete multiscale approximation $u^{\mathrm{ms},h}_{H,k} \in V^{\mathrm{ms},h}_{H,k}$ fulfills

$$a(u^{\mathrm{ms},h}_{H,k}, v) = G(v) \quad \text{for all } v \in V^{\mathrm{ms},h}_{H,k}. \tag{4.3}$$

Theorem 4.1. *Let $u \in V$ solve (1.2) and let $u^{\mathrm{ms},h}_{H,\ell k} \in V^{\mathrm{ms},h}_{H,k}$ solve (4.3). Then*

$$\begin{aligned}\left|\!\left|\!\left| u - u^{\mathrm{ms},h}_{H,\ell k} \right|\!\right|\!\right| &\leq \tilde{C}_4 \|H_T^{-1}\|_{L^\infty(\Omega)} (\ell k)^{d/2} (\tilde{C}_1/\ell)^{\frac{k-2}{2}} \|g\|_{H^{-1}(\Omega)} \\ &\quad + C_{\mathrm{ol}}^{1/2} C_{\mathfrak{I}_H} \alpha^{-1/2} \|Hg\|_{L^2(\Omega)} + |||u - u_h|||\,,\end{aligned}$$

where $\tilde{C}_4$ only depends on ρ, α and β.

Remark 4.2 (Multiscale splitting by nodal interpolation). Having discretized the fine scale computation, i.e., having replaced the infinite dimensional space V by some finite element space $V_h \subset C^0(\overline{\Omega})$ we are allowed to replace the Clément-type interpolation by classical nodal interpolation. This leads to the variational multiscale method in [18], which is a modification of the method first presented in [17]. Because nodal interpolation satisfies the conditions (2.5.a)–(2.5.b), Theorem 4.1 establishes an a priori error bound for the multiscale method [18]. However, the constant $C_{\mathfrak{I}_H}$ in (2.5.a) depends on the ratio H/h of the discretization scales if $d > 1$ ($C_{\mathfrak{I}_H} \approx \log(H/h)$ in $2d$ and $C_{\mathfrak{I}_H} \approx (H/h)^{-1}$ in $3d$, cf. [21]). Hence, for nodal interpolation, the constants $\tilde{C}_1, \tilde{C}_4$ in Theorem 4.1 depend on H/h in a similar fashion. In 2d this can still be acceptable because the dependence on H/h is only logarithmic.

Remark 4.3 (Estimates for the fine scale error). The finite element space V_h may be replaced by any finite element space that contains V_h, e.g., by piecewise polynomials of higher order. The third part in the error bound in Theorem 4.1 can be bounded in terms of data, mesh parameter h, and polynomial degree using standard a priori error estimates. For example, if $A \in W^{1,\infty}(\Omega)$ (bounded with bounded weak derivative) and ε is the smallest present scale, i.e., $\|\nabla A\|_{L^\infty(\Omega)} \lesssim \varepsilon^{-1}$, the third term in the error bound in Theorem 4.1 may be replaced by the worst case bound $Ch\varepsilon^{-1}$ for a first-order ansatz space V_h (see [20]). It is shown in [20] that for highly varying but smooth coefficient A, higher order ansatz spaces are superior.

Remark 4.4 (Periodic coefficient). Let Ω be some square or cube, $g \in L^2(\Omega)$, let A be smooth and periodic, $A(x) = A(x/\varepsilon)$, with some small scale parameter $\varepsilon > 0$, and let u_ε denote the corresponding solution of (1.2). Choose uniform meshes $\mathcal{T}_H$

and $\mathcal{T}_h$ with $H > \varepsilon > h$ and $k \approx \log(H^{-1})$. With regard to the previous comment, Theorem 4.1 yields the error bound

$$\left|\!\left|\!\left| u_\varepsilon - u_{H,\ell k}^{\mathrm{ms},h} \right|\!\right|\!\right| \leq C_g(H + \tfrac{h}{\varepsilon}).$$

With $h \sim \varepsilon H$ the error in the approximation becomes independent of the fine scale oscillations without any so-called resonance effects as they are observed, e.g., in [13]. We emphasize that periodicity can be exploited to reduce the number corrector problems to be solved significantly.

Remark 4.5 (Solution of the local problems). The local problems need to be solved in the spaces $V_h^{\mathrm{f}}(\omega_{x,k})$. This is a standard finite element space with the additional constraint that the trial and test functions should have no component in V_H. In practice this constraint is realized using Lagrange multipliers.

The resulting coarse scale system of equations is of the same size as the original problem, $\dim(V_{H,k}^{\mathrm{ms},h}) = \dim V_H$ and it is still sparse. The number of non-zero entries will be larger and depend on k. Note, however, that the non-zero entries in the stiffness matrix decay exponentially away from the diagonal.

Proof of Theorem 4.1. We use the triangle inequality

$$\left|\!\left|\!\left| u - u_{H,\ell k}^{\mathrm{ms},h} \right|\!\right|\!\right| \leq |||u - u_h||| + \left|\!\left|\!\left| u_h - u_H^{\mathrm{ms},h} \right|\!\right|\!\right| + \left|\!\left|\!\left| u_H^{\mathrm{ms},h} - u_{H,\ell k}^{\mathrm{ms},h} \right|\!\right|\!\right|$$

and follow the arguments from the proof of Theorem 3.6 simply replacing V by V_h and using Lemmas 4.6, 4.8, and 4.9 below (discrete versions of Lemmas 3.1, 3.4, and 3.5) to bound the last two terms. □

Lemma 4.6 (Discrete version of Lemma 3.1)**.** *Let $u_h \in V_h$ solve* (4.1) *and $u_H^{\mathrm{ms},h} \in V_H^{\mathrm{ms},h}$ solve* (4.3) *with k large enough so that $\omega_{x,k} = \Omega$ for all $x \in \mathcal{N}$. Then*

$$\left|\!\left|\!\left| u_h - u_H^{\mathrm{ms},h} \right|\!\right|\!\right| \leq C_{\mathrm{ol}}^{1/2} C_{\mathfrak{I}_H} \alpha^{-1/2} \|Hg\|_{L^2(\Omega)}$$

holds with constants C_{ol} and $C_{\mathfrak{I}_H}$ that only depend on ρ.

Proof. Note that $u_h^{\mathrm{f}} := u_h - u_H^{\mathrm{ms},h}$ is the unique element of $V_h^{\mathrm{f}} := V^{\mathrm{f}} \cap V_h$ such that $a(u_h^{\mathrm{f}}, v) = G(v)$ for all $v \in V_h^{\mathrm{f}}$. The lemma follows from the same arguments in the proof of Lemma 3.1. □

In the remaining part of this Section, $A \lesssim B$ abbreviates an inequality $A \leq C\,B$ with some generic constant $0 \leq C < \infty$ that does not depend on the mesh sizes H, h and the localization parameters. The constant may depend on the contrast β/α but not on the geometrical or topological structure of the coefficient A.

To establish discrete versions of Lemmas 3.4 and 3.5 we are facing the technical difficulty that the product of $v \in V_h$ and some cut-off function η from Definition 3.3 is not necessarily an element of V_h. However, the subsequent lemma shows that the product ηv can be approximated sufficiently well by elements from V_h.

Lemma 4.7. *For all $x \in \mathcal{N}$, $M > m \in \mathbb{N}$, and corresponding cut-off function $\eta_x^{m,M}$ defined in* (3.3) *there exists some $v \in V_h^{\mathrm{f}}(\omega_{x,M+1})$ such that*

$$\left|\!\left|\!\left| \eta_x^{m,M} \phi_x^h - v \right|\!\right|\!\right| \lesssim \frac{1}{M-m} \left|\!\left|\!\left| \phi_x^h \right|\!\right|\!\right|_{\omega_{x,M+1} \setminus \omega_{x,m-1}}.$$

Furthermore, the statement also holds if $\eta_x^{m,M}$ is replaced by $1 - \eta_x^{m,M}$ and $v \in V_h^{\mathrm{f}}(\Omega \setminus \omega_{x,m-1})$.

Proof. Let $x \in \mathcal{N}$, $M > m \in \mathbb{N}$ be fixed and define $\eta := \eta_x^{m,M}$. Let $\mathfrak{I}_h : V \cap C(\bar{\Omega}) \to V^h$ be the nodal interpolant with respect to the mesh $\mathcal{T}_h$. Recall its (local) approximation and stability properties

$$\|\nabla(v - \mathfrak{I}_h v)\|_{L^2(t)} \lesssim h_t \|\nabla^2 v\|_{L^2(t)} \quad \text{and} \quad \|\mathfrak{I}_h v\|_{L^2(t)} \lesssim \|v\|_{L^2(t)}$$

for all polynomials v. According to Lemma 2.1, there exists some $b_x \in V^h$ such that $\mathfrak{I}_H(b_x) = \mathfrak{I}_H(\mathfrak{I}_h(\eta\phi_x^h))$, $|||b_x||| \lesssim \big|||\mathfrak{I}_H(\mathfrak{I}_h(\eta\phi_x^h))\big|||$, and $\operatorname{supp}(b_x) \subset \omega_{x,M+1} \setminus \omega_{x,m-1}$. Hence, $v := \mathfrak{I}_h(\eta\phi_x^h) - b_x \in V_h^{\mathrm{f}}(\omega_{x,M+1})$. Since $\mathfrak{I}_H\mathfrak{I}_h\bar{\eta}_T\phi_x^h = \bar{\eta}_T\mathfrak{I}_H\phi_x^h = 0$ for $\bar{\eta}_T = |T|^{-1}\int_T \eta$, we get

$$\begin{aligned}
&\big|||\eta\phi_x^h - v|||^2 = \big|||\eta\phi_x^h - \mathfrak{I}_h(\eta\phi_x^h) + b_x|||^2 \\
&\lesssim \sum_{t\in\mathcal{T}_h : t\subset\bar{\omega}_{x,M}\setminus\omega_{x,m}} \|\nabla(\eta\phi_x^h - \mathfrak{I}_h(\eta\phi_x^h))\|_{L^2(t)}^2 + \big|||\mathfrak{I}_H(\mathfrak{I}_h((\eta-\bar{\eta}_T)\phi_x^h))|||^2 \\
&\lesssim \sum_{t\in\mathcal{T}_h : t\subset\bar{\omega}_{x,M}\setminus\omega_{x,m}} h_t^2\|\nabla^2(\eta\phi_x^h)\|_{L^2(t)}^2 + \sum_{T\in\mathcal{T}_H : T\subset\bar{\omega}_{x,M+1}\setminus\omega_{x,m-1}} H_T^{-2}\|\mathfrak{I}_h((\eta-\bar{\eta}_T)\phi_x^h)\|_{L^2(T)}^2 \\
&\lesssim \sum_{t\in\mathcal{T}_h : t\subset\bar{\omega}_{x,M}\setminus\omega_{x,m}} h_t^2\Big(\|\nabla^2\eta\|_{L^\infty(t)}^2\|\phi_x^h\|_{L^2(t)}^2 + \|\nabla\eta\|_{L^\infty(t)}^2\|\nabla\phi_x^h\|_{L^2(t)}^2\Big) \\
&\qquad + \sum_{T\in\mathcal{T}_H : T\subset\bar{\omega}_{x,M+1}\setminus\omega_{x,m-1}} H_T^{-2}\|\eta-\bar{\eta}_T\|_{L^\infty(T)}^2\|\phi_x^h\|_{L^2(T)}^2 \\
&\lesssim (M-m)^{-1}\big|||\phi_x^h|||^2_{\omega_{x,M+1}\setminus\omega_{x,m-1}}
\end{aligned}$$

using the property (3.2.c) of η and Poincaré's inequality. This proves the first part of the lemma.

The second part concerning $1-\eta$ follows using the same argument but with $v \in V_h^{\mathrm{f}}(\Omega \setminus \omega_{x,m-1})$. □

Lemma 4.8 (Discrete version of Lemma 3.4)**.** *For all $x \in \mathcal{N}$, $k, \ell \geq 2 \in \mathbb{N}$ the estimate*

$$\big|||\phi_x^h - \phi_{x,\ell k}^h|||_{\omega_{x,\ell}} \leq \tilde{C}_2\left(\frac{\tilde{C}_1}{\ell}\right)^{\frac{k-2}{2}}\big|||\phi_x^h|||_{\omega_{x,\ell}}$$

holds with constants $\tilde{C}_1, \tilde{C}_2$ that only depend on ρ and β/α but not on x, k, ℓ, h, or H.

Proof. Let $\zeta_x := 1 - \eta_x^{\ell(k-1)+1,\ell k-1}$ with $\eta_x^{\ell(k-1)+1,\ell k-1}$ as in equation (3.3) in Definition 3.3. Then there exists a $v \in V_h^{\mathrm{f}}(\omega_{x,\ell k})$ such that

$$\begin{aligned}
\big|||\phi_x^h - v|||_{\omega_{x,\ell k}} &\leq \big|||\phi_x^h - \zeta_x\phi_x^h|||_{\omega_{x,\ell k}} + \big|||\zeta_x\phi_x^h - v|||_{\omega_{x,\ell k}} \\
&\lesssim \big|||\phi_x^h|||_{\omega_{x,\ell k}\setminus\omega_{x,\ell(k-1)+1}} + \big|||\zeta_x\phi_x^h|||_{\omega_{x,\ell k-1}\setminus\omega_{x,\ell(k-1)+1}}.
\end{aligned}$$

Furthermore, using the same argument as in Lemma 3.4,

$$\big|||\zeta_x\phi_x^h|||_{\omega_{x,\ell k-1}\setminus\omega_{x,\ell(k-1)+1}} \lesssim \big|||\phi_x^h|||_{\omega_{x,\ell k}\setminus\omega_{x,\ell(k-1)+1}}$$

which yields

$$\big|||\phi_x^h - \phi_{x,\ell k}^h|||^2 \leq \big|||\phi_x^h - v|||^2_{\omega_{x,\ell k}} + \big|||\phi_x^h|||_{\Omega\setminus\omega_{x,\ell k}} \lesssim \big|||\phi_x^h|||_{\Omega\setminus\omega_{x,\ell(k-1)+1}}.$$

Now let $\eta_j := \eta_x^{\ell(j-1)+1,\ell j}$ (cf. Definition 3.3), $j = 2, 3, \ldots, k-1$ and note that

$$\|A^{1/2}\nabla\phi_x^h\|^2_{L^2(\Omega\setminus\omega_{x,\ell(k-1)})} \le a(\phi_x^h, \eta_{k-1}^2\phi_x^h) - 2\int_\Omega \eta_{k-1}\phi_x^h(A\nabla\phi_x^h)\cdot\nabla\eta_{k-1}.$$

The second term can be treated exactly as in Lemma 3.4 and, hence, bounded by $\ell^{-2}\left|\left|\left|\phi_x^h\right|\right|\right|^2_{\Omega\setminus\omega_{x,\ell(k-2)}}$. We make use of Lemma 4.7 to bound the first term. There exists $v \in V_h^{\mathrm{f}}(\Omega\setminus\omega_{x,\ell(k-1)+1})$ such that

$$a(\phi_x^h, \eta_{k-1}^2\phi_x^h) \le \left|\left|\left|\phi_x^h\right|\right|\right|_{\omega_{\ell(k-1)}\setminus\omega_{\ell(k-2)+1}} \left|\left|\left|\eta_{k-1}^2\phi_x^h - v\right|\right|\right| \lesssim \ell^{-2}\left|\left|\left|\phi_x^h\right|\right|\right|^2_{\Omega\setminus\omega_{x,\ell(k-2)}}.$$

The final assertion follows by similar arguments as in the proof of Lemma 3.4. □

Lemma 4.9 (Discrete version of Lemma 3.5)**.** *There is a constant $\tilde{C}_3$ depending only on ρ and β/α, but not on $|\mathcal{N}|$, k, or ℓ such that*

$$\left|\left|\left|\sum_{x\in\mathcal{N}} v(x)(\phi_x^h - \phi_{x,\ell k}^h)\right|\right|\right|^2 \le \tilde{C}_3(\ell k)^d \sum_{x\in\mathcal{N}} v^2(x)\left|\left|\left|\phi_x^h - \phi_{x,\ell k}^h\right|\right|\right|^2.$$

Proof. For $x \in \mathcal{N}$, let $\zeta_x = 1 - \eta_x^{\ell k+1,\ell k+2}$ (cf. Definition 3.3) and let $z = \sum_{x\in\mathcal{N}} v(x)\,(\phi_x - \phi_{x,\ell k})$. We have,

$$\left|\left|\left|\sum_{x\in\mathcal{N}} v(x)\,(\phi_x - \phi_{x,\ell k})\right|\right|\right|^2 = \sum_{x\in\mathcal{N}} v(x)\,a(\phi_x^h - \phi_{x,\ell k}^h, \zeta_x z + (1-\zeta_x)z) = \mathrm{I} + \mathrm{II}.$$

The first term $\mathrm{I} := \sum_{x\in\mathcal{N}} v(x)\,a(\phi_x^h - \phi_{x,\ell k}^h, \zeta_x z)$ can be treated in exactly the same way as in the proof of Lemma 3.5. We focus on the second term. Due to Lemma 4.7 there exists a $w \in V_h^{\mathrm{f}}(\Omega\setminus\omega_{\ell k})$ such that

$$\begin{aligned}
\mathrm{II} &:= \sum_{x\in\mathcal{N}} v(x)\,a(\phi_x^h - \phi_{x,\ell k}^h, (1-\zeta_x)z - w)\\
&\lesssim \left(\sum_{x\in\mathcal{N}} |v(x)|^2\,\left|\left|\left|\phi_x^h - \phi_{x,\ell k}^h\right|\right|\right|^2\right)^{1/2}\left(\sum_{x\in\mathcal{N}}\left|\left|\left|(1-\zeta_x)z - w\right|\right|\right|^2\right)^{1/2}\\
&\lesssim \left(\sum_{x\in\mathcal{N}} |v(x)|^2\,\left|\left|\left|\phi_x^h - \phi_{x,\ell k}^h\right|\right|\right|^2\right)^{1/2}\left(\sum_{x\in\mathcal{N}}\left|\left|\left|z\right|\right|\right|^2_{\omega_{x,\ell k+2}\setminus\omega_{\ell k+1}}\right)^{1/2}\\
&\lesssim (\ell k)^{d/2}\left(\sum_{x\in\mathcal{N}} |v(x)|^2\,\left|\left|\left|\phi_x^h - \phi_{x,\ell k}^h\right|\right|\right|^2\right)^{1/2}\left|\left|\left|z\right|\right|\right|.
\end{aligned}$$

The result follows immediately. □

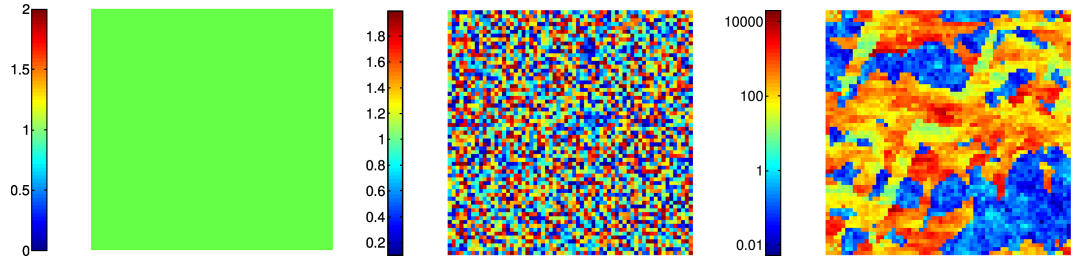


FIGURE 1. Scalar coefficient used in the numerical experiment: A_1 (left), A_2 (middle), A_3 (right).

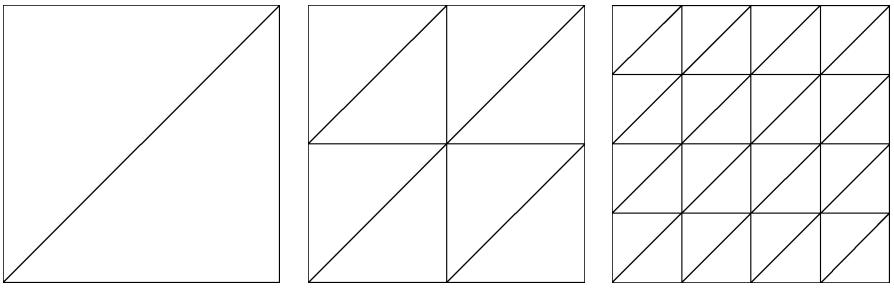


FIGURE 2. Uniform triangulations of the unit square.

5. Numerical experiments

Numerical experiments shall validate our theoretical results from the previous sections.

5.1. **Experimental setup.** Let Ω be the unit square and the outer force $g \equiv 1$ in Ω. Consider three different choices for the scalar coefficient A_1, A_2, A_3 with increasing difficulty as depicted in Figure 1. The coefficient $A_1 = 1$ is constant. The coefficient A_2 is piecewise constant with respect to a uniform Cartesian grid of width 2^{-6}. The values in each grid cell are chosen in the range $[1/20, 2]$; the contrast $\beta(A_2)/\alpha(A_2) \leq 40$ is moderate. The coefficient A_3 is piecewise constant with respect to the same uniform Cartesian grid of width 2^{-6}. Its values are taken from the data of the SPE10 benchmark; see `http://www.spe.org/web/csp/`. The contrast for A_3 is large, $\beta(A_3)/\alpha(A_3) \approx 4 \cdot 10^6$. Consider uniform coarse meshes of size $H = 2^{-1}, 2^{-2}, \ldots, 2^{-6}$ of Ω as depicted in Figure 2. Note that none of these meshes resolves the rough coefficients A_2 and A_3 appropriately.

The reference mesh $\mathcal{T}_h$ has width $h = 2^{-9}$. Since no analytical solutions are available, the standard finite element approximation $u_h \in V_h$ on the reference mesh $\mathcal{T}_h$ serves as the reference solution. All fine scale computations are performed on subsets of $\mathcal{T}_h$.

The approximations are compared with this reference solution only. Doing this, we assume that u_h is sufficiently accurate. True errors would behave similar in the beginning but level off at some point when the reference error $|||u - u_h|||$ dominates the upscaling error.

5.2. **Results for the energy error.** Figure 3 depicts the energy errors of the new multiscale method and the classical P1FEM (see (2.2)) with respect to the same coarse mesh. Depending on the coarse discretization scale H, the localization

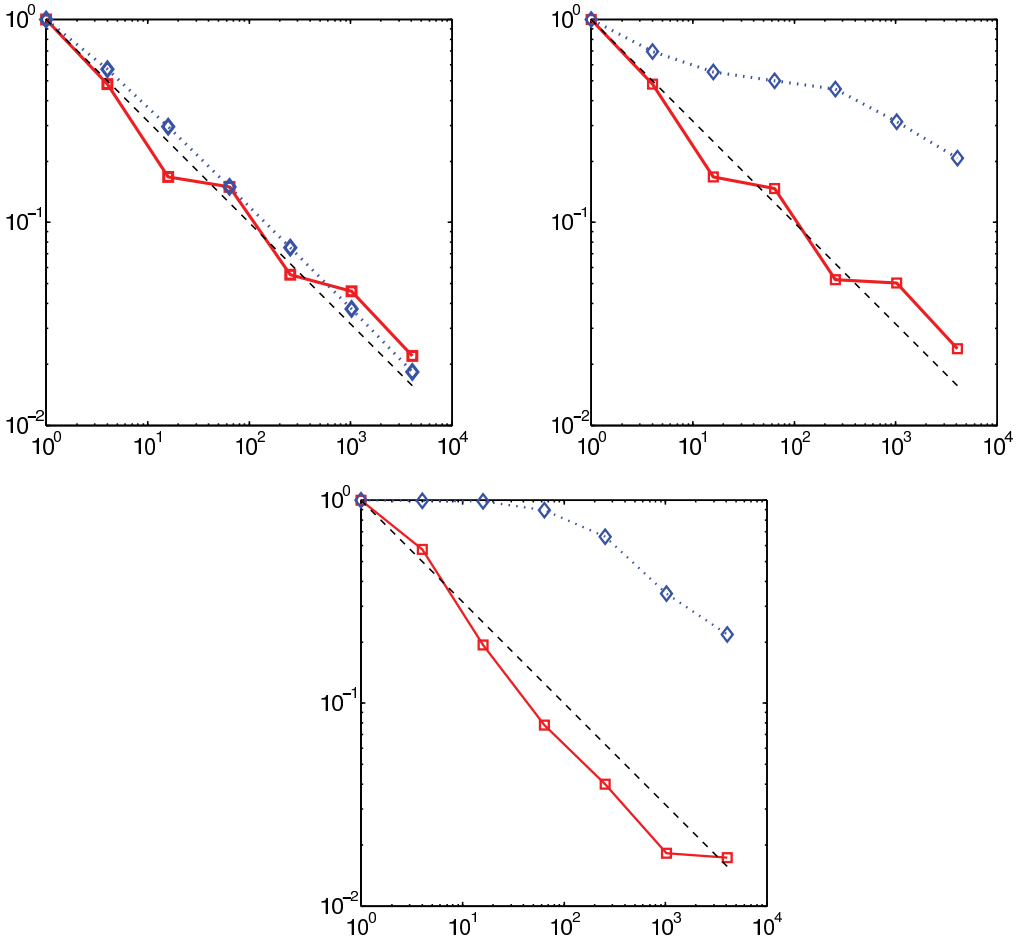


FIGURE 3. Relative energy errors $|||u_h - u_{H,k}^{\mathrm{ms},h}|||/|||u_h|||$ ($\square$ solid) with localization parameter $k = \lceil 2\log(1/H)\rceil$ and $|||u_h - u_H|||\,/\,|||u_h|||$ ($\lozenge$ dotted) vs. number of degrees of freedom $N_{\mathrm{dof}} \approx H^{-2}$ for different coefficients: A_1 (top left), A_2 (top right), A_3 (bottom). The dashed black line is $N_{\mathrm{dof}}^{-1/2}$.

parameter k is chosen to be $\lceil 2\log(1/H)\rceil$. The logarithmic dependence on $1/H$ is motivated by our a priori analysis. The choice of the constant 2 is based on numerical tests. It turns out that, in all experiments, this choice leads to the desired linear textbook convergence (rate $-1/2$) of the energy error (w.r.t. to the number of degrees of freedom $N_{\mathrm{dof}} = |\mathcal{N}| \approx H^{-2}$) related to the sequence of multiscale approximations. Pre-asymptotic effects are not observed. In particular, the performance of our method does not seem to be affected by the high contrast present in A_3. Whether our estimates on the decay of the corrector functions are sub-optimal or have worst-case character with respect to contrast is an issue of present research.

Observe that the classical P1FEM suffers from the lacks of approximability and regularity and converges only poorly for the rough coefficients A_2 and A_3.

5.3. **Results for the L^2 error.** Figure 4 shows L^2 errors of the new multiscale method and the classical P1FEM. Again, the choice of the localization parameter $k = \lceil 2\log(1/H)\rceil$ yields the optimal convergence rate -1 for our method in all

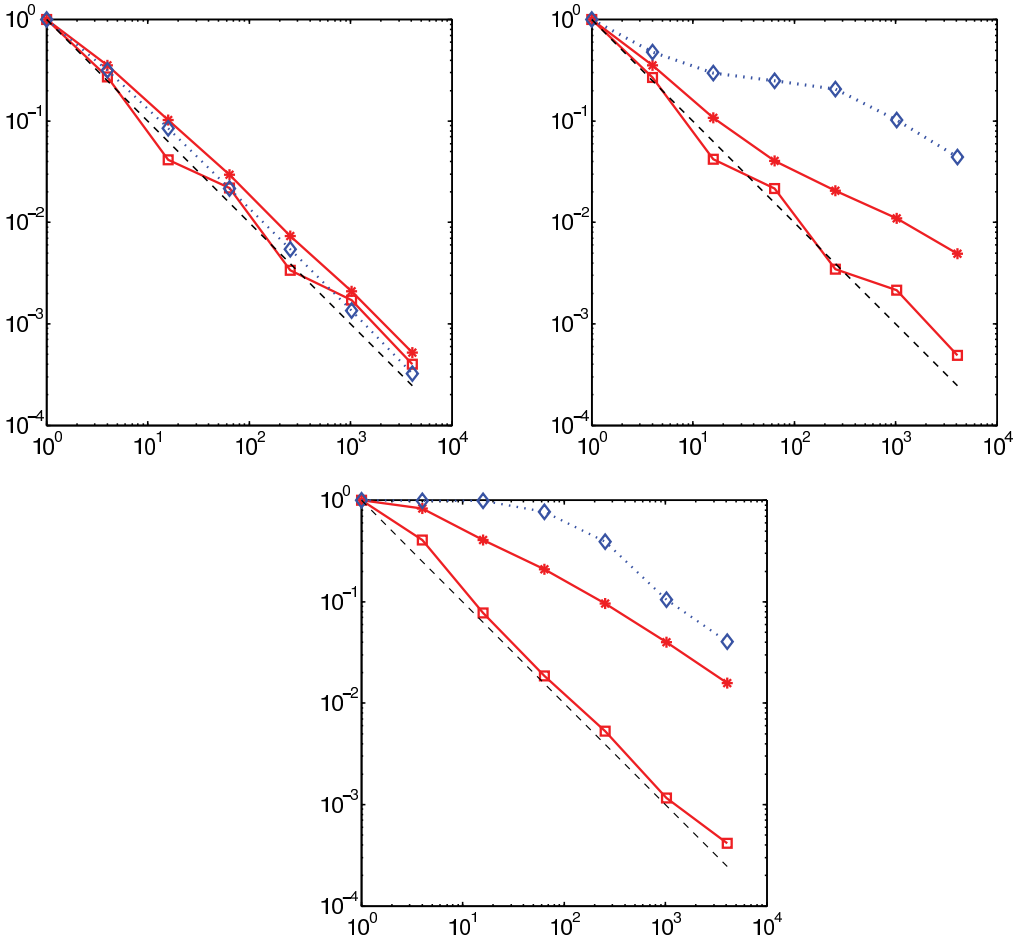


FIGURE 4. Relative L^2 errors $\|u_h - u_{H,k}^{\mathrm{ms},h}\|/\|u_h\|$ ($\square$ solid), $\|u_h - \mathfrak{I}_H u_{H,k}^{\mathrm{ms},h}\|/\|u_h\|$ ($*$ solid) with localization parameter $k = \lceil 2\log(1/H)\rceil$ and $\|u_h - u_H\|/\|u_h\|$ ($\diamond$ dotted) vs. number of degrees of freedom $N_{\mathrm{dof}} \approx H^{-2}$ for different coefficients: A_1 (top left), A_2 (top right), A_3 (bottom). The dashed black line is N_{dof}^{-1}.

experiments (w.r.t. to the number of degrees of freedom $N_{\mathrm{dof}} = |\mathcal{N}| \approx H^{-2}$) without any pre-asymptotic behavior. This observation is justified by a standard Aubin-Nitsche duality argument. Define $e := u_h - u_{H,k}^{\mathrm{ms},h} \in L^2(\Omega)$ and let $z_h \in V_h$ solve

$$a(z_h, v_h) = \int_\Omega e v_h \quad \text{for all } v_h \in V_h.$$

Galerkin orthogonality leads to

$$\|u_h - u_{H,k}^{\mathrm{ms},h}\|_{L^2(\Omega)}^2 = a(z_e - z_{H,k}^{\mathrm{ms},h}, e) \leq \left|\!\left|\!\left| z_h - z_{H,k}^{\mathrm{ms},h} \right|\!\right|\!\right| \left|\!\left|\!\left| u_h - u_{H,k}^{\mathrm{ms},h} \right|\!\right|\!\right|,$$

where $z_{H,k}^{\mathrm{ms},h} \in V_{H,k}^{\mathrm{ms},h}$ is the Galerkin projection of z_h onto the discrete multiscale finite element space $V_{H,k}^{\mathrm{ms},h}$. The estimates for the energy error (see Section 4) and the present choice of k yield the L^2 estimate

$$\|u_h - u_{H,k}^{\mathrm{ms},h}\|_{L^2(\Omega)} \lesssim H^2 \|g\|_{L^2(\Omega)}.$$

More importantly, we observe that the L^2 error between u_h and $\mathfrak{I}_H u_{H,k}^{\mathrm{ms},h}$ converges nicely at a rate close to $-3/4$ without pre-asymptotic effects. This is remarkable because $\mathfrak{I}_H u_{H,k}^{\mathrm{ms},h}$ is a truly coarse approximation. $\mathfrak{I}_H u_{H,k}^{\mathrm{ms},h}$ is an element of the coarse $P1$ finite element space. Hence, it cannot capture microscopic features of the solution. The rate of convergence (with respect to the number of degrees of freedom) is limited by $\frac{1+s}{2}$ for some $s \in [0,1]$ which is related to the regularity of the solution ($u \in H^{1+s}$ for some $s \in [0,1]$). However, $\mathfrak{I}_H u_{H,k}^{\mathrm{ms},h}$ approximates the macroscopic behavior of the solution accurately with only very few degrees of freedom. Note that the storage complexity of the modified basis is of order $\mathcal{O}(h^{-2}\log 1/H)$ whereas its interpolation can be stored in $\mathcal{O}(H^{-2}\log 1/H)$. Once the coarse system matrix of the multiscale method is assembled, $\mathfrak{I}_H u_{H,k}^{\mathrm{ms},h}$ can be computed without using any fine scale information from the modified basis whereas this would be required to represent the full multiscale approximation $u_{H,k}^{\mathrm{ms},h}$.

6. Application to multiscale methods

In this section we discuss three multiscale methods and how the presented analysis relates to each of them.

6.1. **The variational multiscale method.**

The variational multiscale method was first introduced in [14]. The function space V is here split into a coarse part (standard finite element space on a coarse mesh), in our case V_H, and a fine part, in our case V^{f}. The weak form is also decoupled into a coarse and a fine part. The method reads: find $\bar{u} \in V_H$ and $u' \in V^{\mathrm{f}}$ such that

$$\begin{aligned} a(\bar{u},\bar{v}) + a(u',\bar{v}) &= G(\bar{v}) \quad \text{for all } \bar{v} \in V_H, \\ a(u',v') &= G(v') - a(\bar{u},v') \quad \text{for all } v' \in V^{\mathrm{f}}. \end{aligned}$$

The fine scale solution is further decoupled over the coarse elements $T \in \mathcal{T}_H$ and approximated using analytical techniques. Note that the fine scale solution u' is an affine map of the coarse scale solution $\bar{u}$. If we let $u' \approx M\bar{u} + m$ and plug this in to the first equation we get a coarse stiffness matrix of the form $a(\bar{v} + M\bar{v}, \bar{w})$, i.e., a non-symmetric bilinear form for a symmetric problem.

6.2. **The multiscale finite element method.**

In [13] the multiscale finite element method was first introduced. Here modified multiscale basis functions are computed numerically on sub-grids on each coarse element individually. The corrector functions fulfill: find $\phi_{x,T} \in H_0^1(T)$,

$$a(\lambda_x - \phi_{x,T}, v) = 0 \quad \text{for all } v \in H_0^1(T) \text{ and for all } T \in \mathcal{T}_H.$$

Here homogeneous Dirichlet boundary conditions are used on the boundary of each element T, i.e., the local problems are totally decoupled. To get a more accurate method one can improve the boundary conditions using information from the data A. A larger domain can also be considered (this procedure is referred to as oversampling); see [13]. Note that since the coarse scale basis functions are modified (both trial and test space) the resulting method is symmetric.

6.3. **The adaptive variational multiscale method.** The modified basis function construction given by equation (2.7) and (2.8) was first introduced in a variational multiscale framework in [15, 16]. In these papers the Scott-Zhang interpolation was used in the analysis and nodal interpolation in the discrete setting for the numerical examples. The modified basis functions where only used for the trial functions but not for the test functions. A fine scale correction based on the right-hand side data was also included. In [18] the modified basis functions were used for both trial and test functions. The exponential decay of the modified basis functions, with respect to the number of coarse layers of elements in the vertex patches, has been demonstrated numerically in all these works; see [17, 18].

The adaptive variational multiscale method has been extended to convection dominated problems and problems in mixed form [18]. A posteriori error bounds have been derived and adaptive algorithms designed where the local mesh and patch size are chosen automatically in order to reduce the error.

6.4. **Application of the presented analysis.** The convergence proof in this paper gives a valid bound also as $h \to 0$ independent of the patch size and coarse mesh size. The proof does not rely on regularity of the solution and gives a very explicit expression for the rate of convergence. The present analysis confirms the numerical results in [17, 18] and gives the symmetric version of the method, where both trial and test space are modified, the solid theoretical foundation it has previously been missing. The analysis also justifies the use of a posteriori error bounds for adaptivity [16, 18] because we can now prove that the quantities measured on the patch boundary decays exponentially in the number of coarse layers.

For the variational multiscale method this result says that it is important to allow larger subgrid patches than just one coarse element. This will result in overlap but the local problems are totally decoupled and we have in previous works demonstrated how adaptivity can be used to only solve local problems where it is needed, see for instance [16, 18]. For the multiscale finite element method the analysis is not directly applicable since the fine scale space V^{f} is not used. It is the decay in this space which we have proven to be exponential (in number of coarse layers of elements in the subgrid). If this decay is not present, inhomogeneous boundary conditions are instead needed for the subgrid problems. To the best of our knowledge, such constructions have only been proved to be accurate in special settings, e.g., periodic coefficients.

References

[1] I. Babuška and J. E. Osborn, *Generalized finite element methods: their performance and their relation to mixed methods*, SIAM J. Numer. Anal. **20** (1983), no. 3, 510–536, DOI 10.1137/0720034. MR701094 (84h:65076)

[2] Ivo Babuška, Gabriel Caloz, and John E. Osborn, *Special finite element methods for a class of second order elliptic problems with rough coefficients*, SIAM J. Numer. Anal. **31** (1994), no. 4, 945–981, DOI 10.1137/0731051. MR1286212 (95g:65146)

[3] Ivo Babuska and Robert Lipton, *Optimal local approximation spaces for generalized finite element methods with application to multiscale problems*, Multiscale Model. Simul. **9** (2011), no. 1, 373–406, DOI 10.1137/100791051. MR2801210 (2012e:65259)

[4] Ivo Babuška and John E. Osborn, *Can a finite element method perform arbitrarily badly?*, Math. Comp. **69** (2000), no. 230, 443–462, DOI 10.1090/S0025-5718-99-01085-6. MR1648351 (2000i:65114)

[5] I. Babuška and J. M. Melenk, *The partition of unity method*, Internat. J. Numer. Methods Engrg. **40** (1997), no. 4, 727–758, DOI 10.1002/(SICI)1097-0207(19970228)40:4⟨727::AID-NME86⟩3.3.CO;2-E. MR1429534 (97j:73071)

[6] Leonid Berlyand and Houman Owhadi, *Flux norm approach to finite dimensional homogenization approximations with non-separated scales and high contrast*, Arch. Ration. Mech. Anal. **198** (2010), no. 2, 677–721, DOI 10.1007/s00205-010-0302-1. MR2721592 (2012b:35016)

[7] F. Brezzi, L. P. Franca, T. J. R. Hughes, and A. Russo, $b = \int g$, Comput. Methods Appl. Mech. Engrg. **145** (1997), no. 3-4, 329–339, DOI 10.1016/S0045-7825(96)01221-2. MR1456019 (98g:65086)

[8] Carsten Carstensen, *Quasi-interpolation and a posteriori error analysis in finite element methods*, M2AN Math. Model. Numer. Anal. **33** (1999), no. 6, 1187–1202, DOI 10.1051/m2an:1999140. MR1736895 (2001a:65135)

[9] Carsten Carstensen and Rüdiger Verfürth, *Edge residuals dominate a posteriori error estimates for low order finite element methods*, SIAM J. Numer. Anal. **36** (1999), no. 5, 1571–1587 (electronic), DOI 10.1137/S003614299732334X. MR1706735 (2000g:65115)

[10] Philippe G. Ciarlet, *The Finite Element Method for Elliptic Problems*, North-Holland Publishing Co., Amsterdam, 1978. Studies in Mathematics and its Applications, Vol. 4. MR0520174 (58 #25001)

[11] Ph. Clément, *Approximation by finite element functions using local regularization* (English, with Loose French summary), Rev. Française Automat. Informat. Recherche Opérationnelle Sér. RAIRO Analyse Numérique **9** (1975), no. R-2, 77–84. MR0400739 (53 #4569)

[12] L. Grasedyck, I. Greff, and S. Sauter, *The AL basis for the solution of elliptic problems in heterogeneous media*, Multiscale Model. Simul. **10** (2012), no. 1, 245–258, DOI 10.1137/11082138X. MR2902606

[13] Thomas Y. Hou and Xiao-Hui Wu, *A multiscale finite element method for elliptic problems in composite materials and porous media*, J. Comput. Phys. **134** (1997), no. 1, 169–189, DOI 10.1006/jcph.1997.5682. MR1455261 (98e:73132)

[14] Thomas J. R. Hughes, Gonzalo R. Feijóo, Luca Mazzei, and Jean-Baptiste Quincy, *The variational multiscale method—a paradigm for computational mechanics*, Comput. Methods Appl. Mech. Engrg. **166** (1998), no. 1-2, 3–24, DOI 10.1016/S0045-7825(98)00079-6. MR1660141 (99m:65239)

[15] Mats G. Larson and Axel Målqvist, *Adaptive variational multiscale methods based on a posteriori error estimation: duality techniques for elliptic problems*, Multiscale methods in science and engineering, Lect. Notes Comput. Sci. Eng., vol. 44, Springer, Berlin, 2005, pp. 181–193, DOI 10.1007/3-540-26444-2_9. MR2161713 (2006f:65119)

[16] Mats G. Larson and Axel Målqvist, *Adaptive variational multiscale methods based on a posteriori error estimation: energy norm estimates for elliptic problems*, Comput. Methods Appl. Mech. Engrg. **196** (2007), no. 21-24, 2313–2324, DOI 10.1016/j.cma.2006.08.019. MR2319044 (2008e:65328)

[17] Mats G. Larson and Axel Målqvist, *A mixed adaptive variational multiscale method with applications in oil reservoir simulation*, Math. Models Methods Appl. Sci. **19** (2009), no. 7, 1017–1042, DOI 10.1142/S021820250900370X. MR2553176 (2010i:65274)

[18] Axel Målqvist, *Multiscale methods for elliptic problems*, Multiscale Model. Simul. **9** (2011), no. 3, 1064–1086, DOI 10.1137/090775592. MR2831590 (2012j:65419)

[19] Houman Owhadi and Lei Zhang, *Localized bases for finite-dimensional homogenization approximations with nonseparated scales and high contrast*, Multiscale Model. Simul. **9** (2011), no. 4, 1373–1398, DOI 10.1137/100813968. MR2861243 (2012k:35037)

[20] D. Peterseim and S. Sauter, *Finite elements for elliptic problems with highly varying, nonperiodic diffusion matrix*, Multiscale Model. Simul. **10** (2012), no. 3, 665–695, DOI 10.1137/10081839X. MR3022017

[21] Harry Yserentant, *On the multilevel splitting of finite element spaces*, Numer. Math. **49** (1986), no. 4, 379–412, DOI 10.1007/BF01389538. MR853662 (88d:65068a)

Department of Mathematical Sciences, Chalmers University of Technology and University of Gothenburg, Chalmers Tvärgata 3, SE-14296 Göteborg, Sweden

E-mail address: axel@chalmers.se

Rheinische Friedrich-Wilhelms-Universität Bonn, Institute for Numerical Simulation, Wegelerstr. 6, 53115 Bonn, Germany

E-mail address: peterseim@ins.uni-bonn.de

A.2 Oversampling for the Multiscale Finite Element Method

SIAM Multiscale Modeling & Simulation **11**(4):1149-1175, 2013.

(with P. Henning)

Multiscale Model. Simul.
Vol. 11, No. 4, pp. 1149–1175

OVERSAMPLING FOR THE MULTISCALE FINITE ELEMENT METHOD*

PATRICK HENNING† AND DANIEL PETERSEIM‡

Abstract. This paper reviews standard oversampling strategies as performed in the multiscale finite element method (MsFEM). Common to those approaches is that the oversampling is performed in the full space restricted to a patch including coarse finite element functions. We suggest, by contrast, performing local computations with the additional constraint that trial and test functions be linear independent from coarse finite element functions. This approach reinterprets the variational multiscale method in the context of computational homogenization. This connection gives rise to a general fully discrete error analysis for the proposed multiscale method with constrained oversampling without any resonance effects. In particular, we are able to give the first rigorous proof of convergence for an MsFEM with oversampling.

Key words. a priori error estimate, finite element method, multiscale method, MsFEM, oversampling

AMS subject classifications. 35J15, 65N12, 65N30

DOI. 10.1137/120900332

1. Introduction.

The numerical treatment of partial differential equations with rapidly varying and strongly heterogeneous coefficient functions is still a challenging area of present research, especially with regard to applications such as porous media flow or the transport of solutes in groundwater. In such problems, the occurring permeabilities and hydraulic conductivities have rapidly changing features due to different types of soil, microscopic inclusions in the bottom, or porous subsurface rock formations. Any meaningful numerical simulation of relevant physical effects has to account for these highly heterogeneous fine scale structures in the whole computational domain. This means that the underlying computational mesh has to be sufficiently fine to resolve microscopic details. If pore scale effects become relevant or if domains spread over kilometers, then the computational load becomes extremely large and in several applications even too large to treat the problem with standard finite element or finite volume methods. This is just one instance of a so-called multiscale problem as it arises in hydrology, physics, or industrial engineering.

In recent years, many numerical methods have been designed to deal with these computational issues that come along with multiscale problems. Most of them aim to decouple the global fine scale problem into localized subproblems which can be treated independently from each other plus some global coarse problem. The list of proposed multiscale methods, meanwhile, is long. Amongst the most popular methods are the finite element heterogeneous multiscale method (HMM), initially introduced by E and Engquist [9] (see also [10, 11, 1]), the variational multiscale method (VMM) by Hughes [29] and Hughes et al. [30] (see also [32, 33, 34]), the approaches by Owhadi and Zhang [35, 36], or that of Babuska and Lipton [2].

*Received by the editors November 26, 2012; accepted for publication (in revised form) August 1, 2013; published electronically November 5, 2013.
http://www.siam.org/journals/mms/11-4/90033.html

†Department of Information Technology, Uppsala University, SE-751 05 Uppsala, Sweden (patrick.henning@uni-muenster.de).

‡Institut für Numerische Simulation, University of Bonn, D-53115 Bonn, Germany (peterseim@math.hu-berlin.de). The work of this author was supported by the Humboldt-Universität zu Berlin and the DFG Research Center Matheon Berlin through project C33.

In this paper, we deal with another popular method: the multiscale finite element method (MsFEM) proposed by Hou and Wu [26] and further investigated in several contributions [27, 16, 15, 28]. There is an ongoing development of the method to apply it to various fields and equations. For instance, an MsFEM for nonlinear elliptic problems is proposed in [13], a formulation for two phase flow problems in porous media is presented in [12], advection diffusion problems are treated in [8], and an application to elliptic interface problems with high contrast coefficients is presented in [7]. A survey on the method is given in the book by Efendiev and Hou [14]. There is a vast literature devoted to the method, but there are still open questions of strong interest. The most relevant issue is a rigorous error analysis of the method, in particular in the case of nonperiodic microstructures.

The MsFEM is related to some common finite element space with an underlying coarse grid. The essential idea is to modify the corresponding basis functions in such a way that fine scale variations on finer scales are sufficiently well captured. More specifically, local fine scale computations are performed to determine so-called *corrector functions*. These corrector functions can be added as local perturbations to the original set of basis functions of the coarse finite element space.

However, it is well known that the classical MsFEM suffers from so-called resonance errors, which are typically of order $O(\frac{\varepsilon}{H})$, where ε denotes a characteristic size of the small scale and where H denotes the mesh size of the coarse grid (cf. [27, 28]). This implies that the numerical error becomes large in regions where the coarse grid size is close to the characteristic length scale of the microscopic oscillations. There are two different explanations for this error. The first one is a mismatch between the boundary conditions imposed for the local fine scale problems and global behavior of the oscillatory exact solution (cf. [16]). The second explanation is due to the size and geometry of the sampling patch (cf. [28]). The averaged behavior in such a patch should be "representative" so that we can speak about a perfect sample size. If this is not the case, the final approximation might be distorted. In the periodic setting, for instance, the sampling domain should be some multiple of the periodic cell. On triangular patches with cathetuses of the length of a period, this patch is only half a periodic cell (i.e., the patch has bad size and geometry) and lacks essential information. This yields a completely wrong approximation (cf. [21]). In the periodic setting considerable improvements were obtained by Gloria [19, 20], who proposed a regularization of the local (patch) problems by adding a zero-order term. With this strategy, both sources of the oversampling error could be significantly reduced (cf. [19, Theorem 3.1] and [20, sections 5.3 and 5.4]).

In a lot of applications, such as oil reservoir simulations or the transport of solutes in groundwater, a characteristic microscopic length scale ε is unknown, cannot be identified, or does not exist at all. In scenarios without a clear scale separation it is often impossible to predict whether or not we are in the regime of resonance errors. It is very likely to actually hit the problematic regime. Hence, the quality of the final approximation cannot be determined unless resonance errors are eliminated. For this purpose, different *oversampling strategies* have been proposed. The fundamental idea of each of these techniques is to extend the local problems to larger patches and perform the computation on these *oversampling domains* but feed the coarse scale equation only with the information obtained within the original smaller patches. This reduces the effect of wrong boundary conditions and bad sampling sizes. In this paper, we present the two major strategies for oversampling and discuss their advantages and disadvantages. On the basis of these considerations we propose a new strategy that overcomes the issues of the existing strategies. The new approach is

closely related to the VMM-type method presented in [34]. We prove quantitative error estimates for the corresponding multiscale approximations under very general assumptions on the diffusion coefficient.

This contribution is structured as follows: In section 2 we recall the classical formulation of the MsFEM without oversampling. The most popular approaches for oversampling are discussed in section 3. In section 4 we propose a new strategy for which we present a quantitative error analysis. Numerical experiments are presented in section 5. The paper closes with a short conclusion.

2. The multiscale finite element method. In this section, we state the setting of this paper and we establish the required notation. We recall the classical multiscale finite element method (MsFEM) as initially proposed by Hou and Wu [26].

2.1. Setting and notation. Consider a bounded Lipschitz domain $\Omega \subset \mathbb{R}^d$ with a piecewise flat boundary and some matrix-valued coefficient $A \in L^\infty(\Omega, \mathbb{R}^{d\times d}_{sym})$ with uniform spectral bounds $\gamma_{\min} > 0$ and $\gamma_{\max} \geq \gamma_{\min}$,

$$\sigma(A(x)) \subset [\gamma_{\min}, \gamma_{\max}] \quad \text{for almost all } x \in \Omega. \tag{2.1}$$

Given $f \in L^2(\Omega)$, we seek the weak solution of

$$\begin{aligned} -\nabla \cdot A\nabla u &= f \quad \text{in } \Omega, \\ u &= 0 \quad \text{on } \partial\Omega; \end{aligned}$$

i.e., we seek $u \in H^1_0(\Omega) := \{v \in H^1(\Omega) \mid v|_{\partial\Omega} = 0 \text{ in the sense of traces}\}$ that satisfies

$$a(u,v) := \int_\Omega A\nabla u \cdot \nabla v = \int_\Omega fv =: F(v) \quad \text{for all } v \in H^1_0(\Omega). \tag{2.2}$$

We consider two discretization scales $H \geq h > 0$. The coarse scale H is arbitrary, whereas the small scale parameter h may be constrained by the problem. Typically, it is assumed to be smaller than the characteristic length scales of the variations of the diffusion coefficient A.

Let $\mathcal{T}_H$, $\mathcal{T}_h$ denote corresponding subdivisions of Ω into (closed) triangles (for $d = 2$) and tetrahedra (for $d = 3$), i.e., $\bar{\Omega} = \bigcup_{t\in\mathcal{T}_h} t = \bigcup_{T\in\mathcal{T}_H} T$. We assume that $\mathcal{T}_H$, $\mathcal{T}_h$ are regular in the sense that any two elements are either disjoint or share exactly one face or share exactly one edge or share exactly one vertex. For simplicity we assume that $\mathcal{T}_h$ is derived from $\mathcal{T}_H$ by some regular, possibly nonuniform, mesh refinement.

For $\mathcal{T} = \mathcal{T}_H, \mathcal{T}_h$, let

$$P_1(\mathcal{T}) = \{v \in C^0(\Omega) \mid \forall T \in \mathcal{T}, v|_T \text{ is a polynomial of total degree} \leq 1\}$$

denote the set of continuous and piecewise affine functions.

Accordingly, $V_h := P_1(\mathcal{T}_h) \cap H^1_0(\Omega)$ denotes the "high resolution" finite element space and the "coarse space" is given by $V_H := P_1(\mathcal{T}_H) \cap H^1_0(\Omega) \subset V_h$. For any given subset $\omega \subset \Omega$ we define the restriction of V_h to ω with a zero boundary condition by $\mathring{V}_h(\omega) := V_h \cap H^1_0(\omega)$. The nonconforming fine space $V_{h,\mathcal{T}_H}$ is defined by

$$V_{h,\mathcal{T}_H} := \{v_h \mid \forall T \in \mathcal{T}_H,\ (v_h)|_T \in V_h \cap H^1(T)\}.$$

A general function in $v \in V_{h,\mathcal{T}_H}$ may jump across edges of the coarse mesh $\mathcal{T}_H$ and, hence, does not belong to $H^1_0(\Omega)$. However, the $\mathcal{T}_H$-piecewise gradient $\nabla_H v$, with

$(\nabla_H v)|_T = \nabla(v|_T)$ for all $T \in \mathcal{T}_H$, exists. Typically, MsFEM approximations obtained with oversampling are nonconforming approximations of the exact solution in the sense that they do not belong to $H^1_0(\Omega)$.

In the following $x_T \in T$ denotes an arbitrary point, for instance the barycenter of T. For $\Phi_H \in V_H$ and $T \in \mathcal{T}_H$, the affine extension operator $E_T : V_H \to P_1(\Omega)$ is given by

$$E_T(\Phi_H)(x) := (x - x_T) \cdot \nabla\Phi_H(x_T) + \Phi_H(x_T).$$

Finally, by χ_T we denote the characteristic (or indicator) function with $\chi_T(x) = 1$ for $x \in T$ and $\chi_T(x) = 0$ elsewhere.

For the sake of simplicity, all fine scale computations are performed in subspaces of the fine scale finite element space V_h. The Galerkin solution $u_h \in V_h$ which satisfies

$$a_h(u_h, v) = F(v) \quad \text{for all } v \in V_h \tag{2.3}$$

may, hence, be considered as a reference approximation. Note that we never solve this large scale equation. The function u_h serves as a reference solution to compare our multiscale approximations with. The underlying assumption is that the mesh $\mathcal{T}_h$ is chosen sufficiently fine so that u_h is sufficiently accurate.

Throughout this paper, standard notation on Lebesgue and Sobolev spaces is employed and $a \lesssim b$ abbreviates an inequality $a \le C\, b$ with some generic constant $0 \le C < \infty$ that may depend on the shape regularity of finite element meshes and the contrast $\gamma_{\max}/\gamma_{\min}$ but *not* on the mesh sizes H, h and the regularity or the variations of the diffusion matrix A; $a \approx b$ abbreviates $a \lesssim b \lesssim a$.

2.2. The classical MsFEM and reformulation. We first present the classical MsFEM without oversampling as originally stated by Hou and Wu [26], and similarly by Brezzi et al. [3]. They proposed the strategy to enrich the set of standard finite element basis functions by fine scale information. The information is determined by solving local problems on the fine scale. We briefly recall the method and reformulate it in terms of a correction operator Q_h and a corresponding corrector basis.

Let N denote the dimension of the coarse space V_H, and let $\{\Phi_i \mid \ 1 \le i \le N\}$ denote the usual nodal basis of V_H. Given some basis function Φ_i, the corresponding MsFEM basis function $\Phi_i^{\mathrm{MsFEM}} \in V_h$ is uniquely determined by the condition that for all $T \in \mathcal{T}_H$ and for all $\phi_h \in \mathring{V}_h(T)$ it holds that

$$\int_T A(x)\nabla\Phi_i^{\mathrm{MsFEM}}(x) \cdot \nabla\phi_h(x)\, dx = 0 \quad \text{and } \Phi_i^{\mathrm{MsFEM}} = \Phi_i \text{ on } \partial T. \tag{2.4}$$

The span of these MsFEM functions is called the MsFEM solution space

$$V_H^{\mathrm{MsFEM}} := \mathrm{span}\{\Phi_i^{\mathrm{MsFEM}} \mid \ 1 \le i \le N\}.$$

This space is conforming in the sense of $V_H^{\mathrm{MsFEM}} \subset V_h \subset H^1_0(\Omega)$, because the set $\{\Phi_i^{\mathrm{MsFEM}} \mid \ 1 \le i \le N\}$ defines a conforming set of basis functions. The classical MsFEM in the Petrov–Galerkin (PG) formulation due to [28] reads as follows.

Definition 2.1 (MsFEM without oversampling). *The MsFEM approximation $u_H^{MsFEM} \in V_H^{MsFEM}$ is defined as the solution of*

$$\int_\Omega A(x)\nabla u_H^{MsFEM}(x) \cdot \nabla\Phi_H(x)\, dx = \int_\Omega f(x)\Phi_H(x)\, dx \quad \textit{for all } \Phi_H \in V_H. \tag{2.5}$$

In [26], the MsFEM was originally proposed in the Galerkin formulation; i.e., the test functions $\Phi_H \in V_H$ in (2.5) are replaced by test functions $\Phi_H^{\text{MsFEM}} \in V_H^{\text{MsFEM}}$. Observe that due to the orthogonality property (2.4) both formulations are almost identical (in the absence of oversampling). For structural reasons we used the PG version to introduce the MsFEM.

With regard to the general framework for oversampling that we present in the subsequent sections, we note that the MsFEM can be rewritten in the following way.

Remark 2.2. If $u_H^{\text{MsFEM}} \in V_H^{\text{MsFEM}}$ denotes the MsFEM approximation stated in Definition 2.1, then we have $u_H^{\text{MsFEM}} = u_H + Q_h(u_H)$, where $u_H \in V_H$ solves

$$\int_\Omega A\,(\nabla u_H + \nabla Q_h(u_H)) \cdot \nabla \Phi_H = \int_\Omega f \Phi_H \quad \text{for all } \Phi_H \in V_H, \tag{2.6.a}$$

with

$$Q_h(\Phi_H)(x) := \sum_{T \in \mathcal{T}_H} \sum_{i=1}^{d} \partial_{x_i} \Phi_H(x_T) w_{T,i}(x), \tag{2.6.b}$$

and $w_{T,i} \in \mathring{V}_h(T)$ is the unique solution of

$$\int_T A \nabla w_{T,i} \cdot \nabla \phi_h = -\int_T A e_i \cdot \nabla \phi_h \quad \text{for all } \phi_h \in \mathring{V}_h(T). \tag{2.6.c}$$

The set of all functions $w_{T,i}$ is what we are going to call a *local corrector basis*. From the computational point of view, it seems at first glance to be cheaper to compute the corrector basis given by (2.6.c) instead of directly computing the set of multiscale basis functions given by (2.4). The latter formally involves more problems to solve. For instance, if $d = 2$, the assembling of the corrector basis $\{w_{T,i} \mid T \in \mathcal{T}_H, i = 1, 2\}$ requires the solution of $2 \cdot |\mathcal{T}_H|$ local problems, whereas the solutions of $3 \cdot |\mathcal{T}_H|$ local problems are required to assemble $\{\Phi_i^{\text{MsFEM}} \mid \ 1 \le i \le N\}$ by using the gradients of coarse basis functions (for which we have 3 per coarse element). Still, it is possible to use the partition of unity property of the basis functions to equally decrease the costs of the original version of the MsFEM from $d \cdot |\mathcal{T}_H|$ to $(d-1) \cdot |\mathcal{T}_H|$. In particular, restricted to T, the gradients of $(d-1)$ basis functions associated with $(d-1)$ corners of the element T span the gradient of the missing dth basis function on T.

The equivalence between the formulations (2.5) and (2.6) can be easily verified by the relation $\Phi_i^{\text{MsFEM}} = \Phi_i + Q_h(\Phi_i)$. Observe that for every i, for every $T \in \mathcal{T}_H$, and for every $\phi_h \in \mathring{V}_h(T)$,

$$\begin{aligned}
&\int_T A(x)\,(\nabla \Phi_i(x) + \nabla Q_h(\Phi_i)(x)) \cdot \nabla \phi(x)\,dx \\
&= \sum_{i=1}^{d} \partial_{x_i} \Phi_H(x_T) \int_T A(x)\,\big(e_i + \nabla w_T^i(x)\big) \cdot \nabla \phi(x)\,dx = 0
\end{aligned}$$

and $\Phi_i + Q_h(\Phi_i) = \Phi_i$ on ∂T, which is the definition of Φ_i^{MsFEM}.

A symmetric formulation of (2.6.a) is given by the following: find $u_H \in V_H$ with

$$\int_\Omega A\,(\nabla u_H + \nabla Q_h(u_H)) \cdot (\nabla \Phi_H + \nabla Q_h(\Phi_H)) = \int_\Omega f \Phi_H \quad \text{for all } \Phi_H \in V_H. \tag{2.7}$$

Note that (2.6.a) and (2.7) are identical, because

$$\int_T A\,(\nabla u_H + \nabla Q_h(u_H)) \cdot \nabla \phi_h = 0 \quad \text{for all } \ \phi_h \in \mathring{V}_h(T).$$

3. Oversampling strategies. As already discussed in the introduction, the classical MsFEM in Definition 2.1 can be strongly affected or even dominated by resonance errors (cf. [14]). In the absence of scale separation or any knowledge about a suitable sample size for the local problems, the classical MsFEM needs a modification. *Oversampling* is considered to be a remedy to this issue. Oversampling means that the local problems (2.6.a) are solved on larger domains, but only the interior information (i.e., we restrict the gained fine scale information to T) is communicated to the coarse scale equation (2.6.a).

There is no unique way of extending the local problems (2.6.a) to larger patches. Different extensions lead to different oversampling strategies. In this section, we present the two common approaches for oversampling. We rephrase both approaches so that they fit into a common framework. We discuss the advantages and disadvantages of the methods, and then we propose our new oversampling strategy. Note that each of the subsequent strategies is a generalization of the case without oversampling.

We shall introduce some additional notation.

DEFINITION 3.1 (admissible patch). *For $T \in \mathcal{T}_H$, we call $U(T)$ an admissible patch of T if it is nonempty, open, and connected, if $T \subset U(T) \subset \Omega$, and if it is the union of elements of $\mathcal{T}_h$, i.e.,*

$$U(T) = \operatorname{int} \bigcup_{\tau \in \mathcal{T}_h^*} \tau, \quad \text{where } \mathcal{T}_h^* \subset \mathcal{T}_h.$$

A given set of admissible patches is given by $\mathcal{U}$, i.e.,

$$\mathcal{U} := \{U(T) \mid T \in \mathcal{T}_H \text{ and } U(T) \text{ is an admissible patch}\},$$

where $\mathcal{U}$ contains one and only one patch $U(T)$ for each $T \in \mathcal{T}_H$. The set $U(T) \setminus T$ is called an *oversampling layer*. The thickness of the oversampling layer is denoted by $d_{\mathcal{U},T} := \operatorname{dist}(T, \partial U(T))$. Furthermore, we define

$$d_{\mathcal{U}}^{\min} := \min_{T \in \mathcal{T}_H} d_{\mathcal{U},T} \quad \text{and} \quad d_{\mathcal{U}}^{\max} := \max_{T \in \mathcal{T}_H} d_{\mathcal{U},T}$$

as the minimum and maximum thickness.

In the spirit of (2.6.a) and (2.7), we now define the coarse scale equation for an arbitrary MsFEM with a chosen oversampling strategy. As we will see later on, all MsFEM realizations differ only in the correction operator Q_h that determines the oversampling strategy.

DEFINITION 3.2 (framework for oversampling strategies). *Let $\alpha = 1, 2, 3$ denote the index of the oversampling strategy to be specified later on, and let*

$$\{w_{h,T,i}^{\mathcal{U},\alpha} \mid 1 \le i \le d,\ T \in \mathcal{T}_H\}$$

denote a given local corrector basis that depends on the chosen strategy α (see (2.6.a)–(2.6.c) *for the trivial case of such a basis). Then, a (not necessarily conforming) correction operator $Q_h^{\mathcal{U},\alpha} : V_H \to V_{h,\mathcal{T}_H}$ is defined by*

$$Q_h^{\mathcal{U},\alpha}(\Phi_H)(x) := \sum_{T \in \mathcal{T}_H} \chi_T(x) \sum_{i=1}^{d} \partial_{x_i} \Phi_H(x_T)\, w_{h,T,i}^{\mathcal{U},\alpha}(x) \quad \text{for } \Phi_H \in V_H. \tag{3.1}$$

The MsFEM approximation $u_H^\alpha + Q_h^{\mathcal{U},\alpha}(u_H^\alpha)$ obtained with strategy α in the PG formulation reads as follows: find $u_H^\alpha \in V_H$ such that

$$\sum_{T \in \mathcal{T}_H} \int_T A \left(\nabla u_H^\alpha + \nabla Q_h^{\mathcal{U},\alpha}(u_H^\alpha) \right) \cdot \nabla \Phi_H = \int_\Omega f \Phi_H \quad \text{for all } \Phi_H \in V_H. \tag{3.2}$$

The MsFEM approximation $u_H^{\alpha,\mathrm{sym}} + Q_h^{\mathcal{U},\alpha}(u_H^{\alpha,\mathrm{sym}})$ *obtained with strategy* α *and a (not necessarily equivalent) symmetric formulation is given by the following: find* $u_H^{\alpha,\mathrm{sym}} \in V_H$ *with*

$$\sum_{T\in\mathcal{T}_H} \int_T A\left(\nabla u_H^{\alpha,\mathrm{sym}} + \nabla Q_h^{\mathcal{U},\alpha}(u_H^{\alpha,\mathrm{sym}})\right)\cdot\left(\nabla\Phi_H + \nabla Q_h^{\mathcal{U},\alpha}(\Phi_H)\right) = \int_\Omega f(\Phi_H + Q_h^{\mathcal{U},\alpha}(\Phi_H)) \tag{3.3}$$

for all $\Phi_H \in V_H$*. Observe that strategies can differ only in the choice of the corrector basis. The remaining structure of the methods is always the same.*

In the subsequent sections, we demonstrate how existing oversampling strategies fit into the framework presented in Definition 3.2.

3.1. Classical strategy initially introduced by Hou and Wu. The classical oversampling strategy was proposed by Hou and Wu [26] and further used and investigated in several works (cf. [13, 6, 14]).

Let $T \in \mathcal{T}_H$ be fixed, and let $\{\Phi_1^T, \Phi_2^T, \dots, \Phi_{d+1}^T\} \subset V_H$ denote the basis functions that belong to the $d+1$ nodal points in T. Hou and Wu [26] proposed the following oversampling strategy: solve for $\tilde{\Phi}_j^T \in V_h(U(T))$ with

$$\int_{U(T)} A\nabla\tilde{\Phi}_j^T \cdot \nabla\phi_h = 0 \quad \text{for all } \phi_h \in \mathring{V}_h(U(T)) \tag{3.4}$$

and the boundary condition $\tilde{\Phi}_j^T = E_T(\Phi_j^T)$ on $\partial U(T)$, where $E_T(\Phi_j^T)$ denotes the affine extension of $(\Phi_j^T)|_T$. Then, for a given coarse function $\Phi_H \in V_H$, Φ_H^{MsFEM} is defined by

$$\Phi_H^{\mathrm{MsFEM}} = \sum_{j=1}^{d+1} c_j \tilde{\Phi}_j^T,$$

where the c_j are such that $\Phi_H^{\mathrm{MsFEM}}(x_j) = \Phi_H(x_j)$ for all $d+1$ coarse nodes x_j of T. The final coarse scale equation in the PG formulation reads as follows: find $u_H^{\mathbf{1},\mathrm{MsFEM}} \in V_{h,\mathcal{T}_H}$ with

$$\int_\Omega A\nabla_H u_H^{\mathbf{1},\mathrm{MsFEM}} \cdot \nabla\Phi_H = \int_\Omega f\Phi_H \tag{3.5}$$

for all $\Phi_H \in V_H$. Observe that u_H^{MsFEM} is a nonconforming approximation of u. In [28, 16], a slightly different condition is used to define the coefficients c_i. However, it turns out that this modified condition leads to nothing but Oversampling Strategy 2 below.

We shall rephrase this multiscale method with oversampling strategy in the framework of Definition 3.2. Let $Q_T(\Phi_H) := \Phi_H^{\mathrm{MsFEM}} - E_T(\Phi_H)$ define the local corrector, i.e., an operator that communicates fine scale information to the coarse scale equation. The corresponding reduced fine scale space $\mathring{V}_h^r(U(T))$ is given by

$$\mathring{V}_h^r(U(T)) := \mathring{V}_h(U(T)) \setminus \operatorname{span}\{\Phi_1^T, \Phi_2^T, \dots, \Phi_{d+1}^T\} \tag{3.6}$$

with nodal basis functions $\{\Phi_1^T, \Phi_2^T, \dots, \Phi_{d+1}^T\} \subset V_H$. Since

$$Q_T(\Phi_H)(x_i) = \Phi_H^{\mathrm{MsFEM}}(x_i) - \Phi_H(x_i) = 0 \quad \text{for all nodes } x_i \text{ in } T,$$

$Q_T(\Phi_H) \in \mathring{V}_h^r(U(T))$. Moreover, by the definition of Φ_H^{MsFEM}, $Q_T(\Phi_H) \in \mathring{V}_h^r(U(T))$ satisfies

$$\begin{aligned}&\int_{U(T)} A\left(\nabla\Phi_H(x_T) + \nabla Q_T(\Phi_H)\right)\cdot\nabla\phi_h\\ &= \int_{U(T)} A\left(\nabla E_T(\Phi_H) + \nabla Q_T(\Phi_H)\right)\cdot\nabla\phi_h\\ &= \sum_{i=1}^{d} c_i \int_{U(T)} A\nabla\tilde{\Phi}_i^T\cdot\nabla\phi_h = 0\end{aligned}$$

for all $\phi_h \in \mathring{V}_h^r(U(T))$. Since $\nabla\Phi_H(x_T)$ is a constant in $U(T)$, we may rewrite $Q_T(\Phi_H)$ in terms of a corrector basis. This gives us the first definition of oversampling within our framework.

Oversampling Strategy 1. *Let $\mathring{V}_h^r(U(T))$ denote the reduced fine scale space given by* (3.6)*, and let $w_{h,T,i}^{\mathcal{U},\mathbf{1}} \in \mathring{V}_h^r(U(T))$ (for $i \in \{1,2,\dots,d\}$) denote the solution of*

$$\int_{U(T)} A\nabla w_{h,T,i}^{\mathcal{U},\mathbf{1}}\cdot\nabla\phi_h = -\int_{U(T)} Ae_i\cdot\nabla\phi_h \quad \text{for all } \phi_h \in \mathring{V}_h^r(U(T)). \tag{3.7}$$

For $\Phi_H \in V_H$ we define the corrector $Q_h^{\mathcal{U},\mathbf{1}}(\Phi_H) \in V_{h,\mathcal{T}_H}$ by

$$Q_h^{\mathcal{U},\mathbf{1}}(\Phi_H) := \sum_{T\in\mathcal{T}_H} \chi_T(x) \sum_{i=1}^{d} \partial_{x_i}\Phi_H(x_T)\, w_{h,T,i}^{\mathcal{U},\mathbf{1}}(x).$$

Let $u_H^{\mathbf{1}} \in V_H$ be the solution of (3.2)*, i.e.,*

$$\sum_{T\in\mathcal{T}_H}\int_T A\left(\nabla u_H^{\mathbf{1}} + \nabla Q_h^{\mathcal{U},\alpha}(u_H^{\mathbf{1}})\right)\cdot\nabla\Phi_H = \int_\Omega f\Phi_H \quad \text{for all } \Phi_H \in V_H.$$

Then, $u_H^{\mathbf{1},\mathrm{MsFEM}} := u_H^{\mathbf{1}} + Q_h^{\mathcal{U},\mathbf{1}}(u_H^{\mathbf{1}})$ defines the MsFEM approximation obtained with Oversampling Strategy 1*. Obviously, $u_H^{\mathbf{1},\mathrm{MsFEM}}$ solves* (3.5)*.*

Remark 3.3. The explicit boundary condition for the local problems (3.4) is often missing in the literature (cf. [26, 14]). However, it seems that these computations were performed for the case described above; i.e., the solution $\tilde{\Phi}_i^T$ of (3.4) takes the values of an affine function on $\partial U(T)$ (cf. [16, 28], which also refer to the numerical experiments in [26]). In some works (cf. [13]) the local problems (3.4) are formulated with the boundary condition $\tilde{\Phi}_i^T = \Phi_i^T$ on $\partial U(T)$. This seems to be a mistake, because the new basis functions will be equal to zero whenever $U(T)$ is larger than the support of the original basis functions.

3.2. Oversampling motivated from homogenization theory. The second type of oversampling is motivated from numerical homogenization theory. Assume that we regard

$$\text{find } u^\varepsilon \in H_0^1(\Omega) \text{ with } \int_\Omega A^\varepsilon\nabla u^\varepsilon\cdot\nabla\Phi = \int_\Omega f\Phi \quad \text{for all } \Phi\in H_0^1(\Omega),$$

and assume that A^ε is uniformly bounded and coercive in ε, that A^ε is H-convergent to some matrix A^0, and that $u^\varepsilon \rightharpoonup u^0$ in $H^1(\Omega)$, where $u^0 \in H_0^1(\Omega)$ is called the homogenized solution.

Then, a numerical approximation of the homogenized solution u^0 can be obtained by discretizing a more convenient equation (see (3.8) below). For this purpose, let $B(x,\eta)$ denote an open ball centered at $x \in \Omega$ with radius $\eta > 0$, and let $N(x,\eta)$ denote an open neighborhood of $x \in \Omega$ with a Lipschitz boundary. It is assumed that there exist $0 < c \le C$ so that for all $\eta > 0$ and all $x \in \Omega$ there holds $c|B(x,\eta)| \le |N(x,\eta)| \le C|B(x,\eta)|$. We seek $u^{\varepsilon,\eta,\zeta} \in H_0^1(\Omega)$ that solves

$$\int_\Omega |N(x,\eta)|^{-1} \int_{N(x,\eta)} A^\varepsilon(y) \left(\nabla u^{\varepsilon,\eta,\zeta}(x) + \nabla_y Q(u^{\varepsilon,\eta,\zeta})(x,y)\right) \cdot \nabla\Phi(x)\, dy\, dx = \int_\Omega f(x)\Phi(x)\, dx \tag{3.8}$$

for $\Phi \in H_0^1(\Omega)$, where for given $\Psi \in H_0^1(\Omega)$ the corrector $Q(\Psi)(x,\cdot) \in H_0^1(N(x,\eta+\zeta))$ is determined by

$$\int_{N(x,\eta+\zeta)} A^\varepsilon(y) \left(\nabla\Psi(x) + \nabla_y Q(\Psi)(x,y)\right) \cdot \nabla\phi(y)\, dy = 0 \text{ for all } \psi \in H_0^1(N(x,\eta+\zeta)).$$

If $\zeta = \zeta(\eta)$ and $\lim_{\eta\to 0} \frac{\zeta(\eta)}{\eta} = 0$, then it holds that

$$\lim_{\eta,\zeta\to 0} \lim_{\varepsilon\to 0} \|u^0 - u^{\varepsilon,\eta,\zeta}\|_{H^1(\Omega)} = 0.$$

As a consequence thereof, we get that

$$\lim_{\eta,\zeta\to 0} \lim_{\varepsilon\to 0} \|u^\varepsilon - u^{\varepsilon,\eta,\zeta}\|_{L^2(\Omega)} = 0.$$

This result was shown by Gloria [17, 18] in a general nonlinear setting. Since $u^{\varepsilon,\eta,\zeta}$ yields a good approximation of u^ε, the result suggests looking at discretizations of (3.8). This was exploited, for instance, in [23]. A general numerical framework that can be seen as a discretization of (3.8) was proposed in [24]. Particularly, the HMM and the MsFEM are recovered from the framework, which leads to a straightforward oversampling strategy. This strategy can be formulated as follows (cf. [18, 24, 25]).

OVERSAMPLING STRATEGY 2. *For $i \in \{1,2,\dots,d\}$, let $w_{h,T,i}^{\mathcal{U},\mathbf{2}} \in \mathring{V}_h(U(T))$ solve*

$$\int_{U(T)} A\nabla w_{h,T,i}^{\mathcal{U},\mathbf{2}} \cdot \nabla\phi_h = -\int_{U(T)} Ae_i \cdot \nabla\phi_h \quad \text{for all } \phi_h \in \mathring{V}_h(U(T)), \tag{3.9}$$

and for $\Phi_H \in V_H$, let $Q_h^{\mathcal{U},\mathbf{2}}(\Phi_H) \in V_{h,\mathcal{T}_H}$ denote the corrector given by (3.1). If $u_H^{\mathbf{2}} \in V_H$ is the solution of (3.2), then $u_H^{\mathbf{2}} + Q_h^{\mathcal{U},\mathbf{2}}(u_H^{\mathbf{2}})$ defines the MsFEM approximation obtained with Oversampling Strategy 2. We therefore denote

$$u_H^{\mathbf{2},\mathrm{MsFEM}} := u_H^{\mathbf{2}} + Q_h^{\mathcal{U},\mathbf{2}}(u_H^{\mathbf{2}}).$$

We immediately see that Oversampling Strategies 1 and 2 differ only in the fine scale trial space for the local problems and that they are identical for $U(T) = T$, even though Oversampling Strategy 2 was formulated independently of Oversampling Strategy 1. In [28, 16], Oversampling Strategy 2 is written in terms of an asymptotic expansion in the periodic case. Also note that this second approach is closely related to the finite element HMM, where the same type of oversampling is used (cf. [9, 10, 11, 22]). Notably, the HMM and the MsFEM can be reinterpreted in a common homogenization framework (cf. [17, 18]) and in a common numerical framework (cf. [24]).

3.3. Discussion of the strategies. As we just discussed, there are two widely used strategies for oversampling for the MsFEM. However, the difference between both approaches is only minor and the behavior of the resulting approximations appears to be qualitatively the same. The small difference in the local trial spaces does not seem to have a significant impact. At least, the error estimates available for Oversampling Strategies 1 and 2 are very similar. The literature does not even distinguish between these strategies. For instance, [6] (using Oversampling Strategy 1) claims to generalize the results of [16] (using Oversampling Strategy 2). Such a mixture of strategies can be observed in several works on this topic. To the best of our knowledge, even though both approaches seem to behave identically, a rigorous proof of this conjecture is still missing. Oversampling Strategy 1 suggests fixing the corrector $Q_T^{\mathbf{1}}(\Phi)$ in the corners of the coarse grid element T (forcing it to zero), whereas the corrector proposed by Oversampling Strategy 2 does not have such a restriction leaving it completely free in these corners.

Remark 3.4. As already mentioned, the MsFEM might also be considered in a symmetric formulation (cf. [16]); i.e., the coarse scale equation reads as follows: find $u_H \in V_H$ with

$$\sum_{T\in\mathcal{T}_H}\int_T A\left(\nabla u_H^\alpha + \nabla Q_h^{\mathcal{U},\alpha}(u_H^\alpha)\right)\cdot\left(\nabla\Phi_H + \nabla Q_h^{\mathcal{U},\alpha}(\Phi_H)\right) = \int_\Omega f(\Phi_H + Q_h^{\mathcal{U},\alpha}(\Phi_H))$$

for all $\Phi_H \in V_H$ and where $Q_h^{\mathcal{U},\alpha}$ is defined either with Oversampling Strategy 1 or 2. However, the theoretical and numerical results in [28] show that this version of the method still suffers from resonance errors. One explanation was suggested by Gloria [18], who proposed a simple computation:

$$\begin{aligned}
&\int_T A\left(\nabla u_H^\alpha + \nabla Q_h^{\mathcal{U},\alpha}(u_H^\alpha)\right)\cdot\left(\nabla\Phi_H + \nabla Q_h^{\mathcal{U},\alpha}(\Phi_H)\right)\\
&= \int_T A\left(\nabla u_H^\alpha + \nabla Q_h^{\mathcal{U},\alpha}(u_H^\alpha)\right)\cdot\nabla\Phi_H\\
&\quad + \int_T A\left(\nabla u_H^\alpha + \nabla Q_h^{\mathcal{U},\alpha}(u_H^\alpha)\right)\cdot\nabla Q_h^{\mathcal{U},\alpha}(\Phi_H)\\
&= \int_T A\left(\nabla u_H^\alpha + \nabla Q_h^{\mathcal{U},\alpha}(u_H^\alpha)\right)\cdot\nabla\Phi_H\\
&\quad + \int_{U(T)\setminus T} A\left(\nabla u_H^\alpha + \nabla Q_h^{\mathcal{U},\alpha}(u_H^\alpha)\right)\cdot\nabla Q_h^{\mathcal{U},\alpha}(\Phi_H).
\end{aligned}$$

This means that the effective MsFEM bilinear forms in the PG and non-PG formulations differ in the term

$$\sum_{T\in\mathcal{T}_H}\int_{U(T)\setminus T} A\left(\nabla u_H^\alpha + \nabla Q_h^{\mathcal{U},\alpha}(u_H)\right)\cdot\nabla Q_h^{\mathcal{U},\alpha}(\Phi_H),$$

which still seems to contain the problematic boundary layers that we tried to get rid of. Observe that we integrate over the layer $U(T)\setminus T$. This is exactly the region where we encounter unpleasant boundary effects of the correctors $Q_h^{\mathcal{U},\alpha}(u_H)$ and $Q_h^{\mathcal{U},\alpha}(\Phi_H)$. This might imply that preference should be given to the PG formulation. Note,

however, that uniqueness and existence of discrete solutions have not been proved for general oversampling so far.

Let us review the two essential results concerning the convergence of MsFEM approximations with oversampling. The first result is due to Gloria and is the most general result currently available for Oversampling Strategy 2.

THEOREM 3.5. *Let $f \in L^2(\Omega)$ and $A^\varepsilon \in L^\infty(\Omega, \mathbb{R}^{d\times d})$ be a sequence of (possibly nonsymmetric) matrices with uniform spectral bounds $\gamma_{\min} > 0$ and $\gamma_{\max} \geq \gamma_{\min}$,*

$$\sigma(A^\varepsilon(x)) \subset [\gamma_{\min}, \gamma_{\max}] \quad \text{for almost all } x \in \Omega \text{ and for all } \varepsilon > 0, \tag{3.10}$$

and assume that A^ε is H-convergent. Furthermore, let $u_H^\varepsilon \in V_H$ denote the corresponding MsFEM approximation obtained with Oversampling Strategy 2, and let

$$\frac{\operatorname{diam}(U(T)) - \operatorname{diam}(T)}{\operatorname{diam}(T)} \to 0 \quad \text{for } H \to 0.$$

Then we have

$$\lim_{H\to 0} \lim_{\varepsilon \to 0} \|u^\varepsilon - u_H^\varepsilon\|_{L^2(\Omega)} = 0.$$

The proof for general nonsymmetric coefficients is given in [20, Theorem 5], and the case of nonlinear problems is presented in [18, Theorem 6 and Remark 7]. At first glance the result appears counterintuitive in the sense that it suggests letting the oversampling converge to zero. However, the first limit is in ε, which makes the relative thickness $\frac{\varepsilon}{d_U^{\min}}$ of the oversampling layer grow to infinity. Hence, the correct interpretation is that for fixed ε the computational domains should blow up to infinity. In this case, the optimal corrector problem is an equation formulated on the whole $\mathbb{R}^d$. These corrector problems are exactly the cell problems known from periodic and stochastic homogenization theory. In the periodic setting the classical cell problems can be extended to the $\mathbb{R}^d$ by periodicity, and in the stochastic setting they are directly formulated in $\mathbb{R}^d$ to obtain the correct stochastic average (cf. [31]).

Theorem 3.5 gives a clear message in the case of extremely small microscopic variations. If ε (the characteristic length scale of the fine scale oscillations) is (globally) sufficiently small, then the resulting MsFEM approximation yields very good approximations. This is a very important result, but it is purely qualitative. For example, it does not answer the question of how (thick) to choose an oversampling patch. We cannot predict how the method behaves if there is a large spectrum of oscillations without a scale separation. For instance, we might encounter variations, where it is hard to tell which of them are macroscopic and which are microscopic (i.e., "ε-dependent"). In practice, we do not construct an artificial sequence in ε; we have only a given scenario and a given set of data.

The next theorem due to Hou, Wu, and Zhang is much more restrictive, but it gives a more quantitative answer than Theorem 3.5.

THEOREM 3.6. *Assume that $d = 2$, $f \in L^2(\Omega)$, and A is a bounded, elliptic, symmetric, and ε-periodic C^3-matrix, i.e., $A(x) = A_p(\frac{x}{\varepsilon})$, with $A_p \in C^3([0,1]^d, \mathbb{R}^{d\times d}_{sym})$ being periodic. Let $u_H^\varepsilon \in V_{h,\mathcal{T}_H}$ denote the MsFEM approximation obtained with Oversampling Strategy 2. Then*

$$\|u_H^\varepsilon - u^\varepsilon\|_{L^2(\Omega)} \leq C\left(\frac{\varepsilon}{d_U^{\min}} + H + \varepsilon(\log H)^{\frac{1}{2}}\right),$$

$$\left(\sum_{T\in\mathcal{T}_H} \|\nabla u_H^\varepsilon - \nabla u^\varepsilon\|^2_{L^2(\Omega)}\right)^{\frac{1}{2}} \leq C\left(\frac{\varepsilon}{d_U^{\min}} + H + \varepsilon^{\frac{1}{2}}\right).$$

A proof of this theorem is given in [28]. The assumption $d = 2$ seems to be essential for their strategy. Note that in [28] the theorem is formulated without the $\frac{1}{d_{\mathcal{U}}^{\min}}$ contribution. Instead, the authors make the assumption that the oversampling layer is sufficiently large. Following their proofs one can easily see that the generalized estimate reads as above (cf. [14] for the case $d_{\mathcal{U}}^{\min} = CH$). In particular, the $\frac{\varepsilon}{d_{\mathcal{U}}^{\min}}$-term describes the decay of the error between the exact corrector and the corrector with wrong boundary conditions in a coarse element T. The decay turns out to be inversely proportional to the thickness of the layer. Because of the ε scaling of the solution, the effective term becomes $\frac{\varepsilon}{d_{\mathcal{U}}^{\min}}$. This seems to be a sharp estimate for the decay due to the findings in [16, 5]. A proof of Theorem 3.6 for Oversampling Strategy 1 can be achieved in the same fashion as in [28]. Theorem 3.6 predicts the following: if locally $\mathrm{O}(H)$=$\mathrm{O}(\varepsilon)$, the patch size of the local problems must not be of order $\mathrm{O}(H)$ to preserve convergence. Still, the theorem gives only an answer of how to choose the oversampling patches $\mathcal{U}$ if ε is a known parameter.

If the thickness of the oversampling layer is of order $\mathrm{O}(h)$, both estimates in Theorem 3.6 receive an order $\mathrm{O}(\frac{\varepsilon}{h})$ term and the right-hand sides remain large. In general, the thickness $d_{\mathcal{U}}^{\min}$ must be large in comparison to ε. Analytically, this implies that $\mathrm{O}(H)$-oversampling might be not enough in regions where we deal with resonance errors due to $\mathrm{O}(H)$=$\mathrm{O}(\varepsilon)$. This seems to show up in the numerical experiments in [28], where the authors observe a stagnation in the convergence of the H^1-error for H entering the region with $\mathrm{O}(H)$=$\mathrm{O}(\varepsilon)$. The effect on the L^2-error is less strong. However, the value of $d_{\mathcal{U}}^{\min}$ is missing in the experiments in [28], so we can assume only that $d_{\mathcal{U}}^{\min}$ is of order H. Otherwise the computation of the MsFEM basis functions becomes quite expensive. However, we note that there also exists a modification of Oversampling Strategy 2 proposed by Gloria (cf. [19, Theorem 3.1] and [20, paragraphs 5.3 and 5.4]) where the local problems are regularized by adding the term $\kappa^{-1}(w_{h,T,i}^{\mathcal{U},\mathbf{2}}, \phi_h)_{L^2(U(T))}$ (for large $\kappa > 0$) to the left-hand side of problem (3.9). Using this modified strategy, the Theorem 3.6–type estimates can be improved enormously, even without restrictions on space dimensions and much weaker assumptions on the regularity of A^ε.

Remark 3.7. In [16], the symmetric version (3.3) of the MsFEM is considered. Here, the derived L^2-estimate reads as

$$\|u_H^\varepsilon - u^\varepsilon\|_{L^2(\Omega)} \leq C \left(\varepsilon + H^2 + \varepsilon(\log H) + \frac{\varepsilon}{d_{\mathcal{U}}^{\min}} + C_r \left(\frac{\varepsilon}{H} \right)^2 \right),$$

where u_H^ε denotes the MsFEM approximation of the symmetric problem (3.3) determined with Oversampling Strategy 1. Due to the numerical experiments in [16], C_r seems to have a considerable size so that this term is dominating the estimate if locally $\mathrm{O}(\varepsilon)$=$\mathrm{O}(H)$. We notice that the estimate is worse than the L^2-estimate for the PG version of the method, because the last term cannot be reduced even for large $d_{\mathcal{U}}^{\min}$. These observations are consistent with Remark 3.4. However, this leads to an additional problem of the MsFEM with Oversampling Strategy 1 or 2. On the one hand, the PG version should be preferred over the symmetric version (see the estimates). On the other hand, the existence and uniqueness of the corresponding MsFEM approximations have not been established so far, not to mention the stability. For the symmetric version, we can simply exploit the ellipticity of A to conclude that the method is well posed and stable. For the PG version there is no such argument. The only result is a perturbation result due to Gloria [18], saying that if the oversampling size is small enough (i.e., if the difference between the PG formulation and

symmetric formulation is small enough), then we still have existence and uniqueness. The lack of knowledge regarding the general well-posedness of the PG MsFEM with Oversampling Strategies 1 and 2 is a big issue of these approaches.

The ε-terms in the estimates that cannot be reduced with $H \gtrsim \varepsilon$ should be seen as fixed modeling errors. They describe the error between exact solution and homogenized solution. In a general nonperiodic nonstochastic scenario they cannot be quantified.

In conclusion we have two findings. First, in general, both approaches do not show clear asymptotics for a convergence to the exact solution (for $H \gtrsim \varepsilon$). There is always a remainder of order ε, even if $U(T) = \Omega$. In particular, this is a problem if ε is unknown or if the microstructure is heterogeneous. Second, if the modeling error of order ε is negligible, Theorem 3.6 still suggests that linear convergence (with respect to H) can be achieved only if the oversampling thickness scales with O(1), which makes the local problems prohibitively expensive. Since the estimate for the decay rate $\frac{\varepsilon}{d_{\mathcal{U}}^{\min}}$ of the corrector error is sharp (in the periodic setting), we cannot hope for much improvement of the final error estimates stated in Theorem 3.6.

We may summarize the following issues, which we address to solve with our new oversampling strategy to be proposed in the next section:

(a) elimination of resonance errors of any kind,
(b) clear prediction for the size of oversampling patches without explicit knowledge about the microstructure or scale separation,
(c) construction of a conforming approximation in $H_0^1(\Omega)$,
(d) a quantitative error analysis in H without restrictive regularity assumptions on the coefficients and for all space dimensions,
(e) a priori error estimation in the fully discrete setting (previous results were obtained under the assumption that the local problems are solved exactly),
(f) formulation of a stable approach for which we can guarantee existence and uniqueness of the resulting MsFEM approximation, and
(g) prevention of unstable splittings due to point evaluations as, e.g., required to implement the constraint in Oversampling Strategy 1 (cf. the definition of $\mathring{V}_h^r(U(T))$ in (3.6)).

Note that points (d) and (e) could be done in the periodic setting for Oversampling Strategies 1 and 2 by using, e.g., the techniques presented in [19, 20].

4. Constrained oversampling. In this section we introduce a third oversampling strategy for which we derive a quantitative a priori error estimate. The results are presented in subsection 4.1, and a corresponding proof is given in subsection 4.2. All the results require solely the assumptions stated in section 2.1 to be satisfied, i.e., $A \in L^\infty(\mathbb{R}^{d\times d}_{\mathrm{sym}})$ uniformly positive definite and $f \in L^2(\Omega)$.

4.1. New strategy and quantitative error estimates. In the following, let $\mathcal{N}_H$ denote the set of interior vertices of the coarse grid $\mathcal{T}_H$. For a given node $z \in \mathcal{N}_H$, $\Phi_z \in V_H$ denotes the corresponding nodal basis function as before.

Our new approach is based on some multiscale decomposition of the space V_h,

$$V_h = V_H \oplus W_h, \tag{4.1}$$

where the space W_h contains the "fine scale" functions of V_h, i.e., functions that are not captured by V_H. More precisely, we choose W_h to be the kernel of some Clément-type quasi-interpolation operator $I_H : H_0^1(\Omega) \to V_H$,

$$W_h := \{v \in V_h \mid \ I_H(v) = 0\}. \tag{4.2}$$

Several choices for I_H are possible. We refer the reader to [34] for an axiomatic characterization. In this paper, for the sake of simplicity, we choose the particular operator introduced in [4]. Given $v \in H^1_0(\Omega)$, $I_H v := \sum_{z\in\mathcal{N}_H} (I_H v)(z)\Phi_z$ is determined by the nodal values

$$(I_H v)(z) := \frac{\int_\Omega v\Phi_z}{\int_\Omega \Phi_z} \quad \text{for } z \in \mathcal{N}_H. \tag{4.3}$$

The nodal values are weighted averages of the function over nodal patches $\omega_z := \operatorname{supp}\Phi_z$. The operator is linear, surjective, bounded, and invertible on the finite element space V_H. Hence, the decomposition (4.1) exists and is stable; it is even orthogonal in $L^2(\Omega)$.

Recall the (local) approximation and stability properties of the interpolation operators I_H [4]: There exists a generic constant C such that for all $v \in H^1_0(\Omega)$ and for all $K \in \mathcal{T}_H$ it holds that

$$H_T^{-1}\|v - I_H v\|_{L^2(K)} + \|\nabla(v - I_H v)\|_{L^2(K)} \le C\|\nabla v\|_{L^2(\omega_K)}, \tag{4.4}$$

where $\omega_K := \cup\{K' \in \mathcal{T}_H \mid K' \cap K \neq \emptyset\}$. The constant C depends on the shape regularity of the finite element mesh $\mathcal{T}_H$ but not on the local mesh size $H_T := \operatorname{diam}(T)$.

Remark 4.1 (nodal interpolation). Since we consider a fully discrete setting, where corrector problems are solved in the fine scale finite element space V_h, we could have chosen nodal interpolation instead of Clément-type interpolation. The subsequent definitions and results will be almost verbatim the same. However, nodal interpolation does not satisfy the estimate (4.4) with an h-independent constant if $d > 1$. The best constant $C = C_d(h)$ reads as $C_2(h) = \log(H/h)$ and $C_3(h) = (H/h)^{-1}$ depending on the spatial dimension d (cf. [38]). Since this constant enters basically all error estimates below, we would end up with an h-dependence of the multiplicative constants in the final error estimates. In two dimensions this can still be acceptable, because the dependence on h is only logarithmic.

With the decomposition (4.1) we do not search the local correctors in the full fine scale space V_h but only in the constrained space W_h. The advantage is the following: as stated in the previous section for Oversampling Strategies 1 and 2, the standard decay for the difference between the local correctors and the global "exact" corrector is of order $\frac{1}{d_{\mathcal{U}}^{\min}}$ (see Theorem 3.6), but in the constrained space W_h we can achieve an exponential-type decay (cf. Lemma 4.9 below).

We now propose our new oversampling strategy.

Oversampling Strategy 3 (constrained oversampling). *Let W_h denote the space given by* (4.2)*, and define*

$$\mathring{W}_h(U(T)) := \{v_h \in W_h \mid v_h|_{\Omega\setminus U(T)} = 0\}. \tag{4.5}$$

The local correctors $w_{h,T,i}^{\mathcal{U},\mathbf{3}} \in \mathring{W}_h(U(T))$ (for $i \in \{1,2,\dots,d\}$) are defined as the (unique) solutions of

$$\int_{U(T)} A\nabla w_{h,T,i}^{\mathcal{U},\mathbf{3}} \cdot \nabla\phi_h = -\int_T Ae_i \cdot \nabla\phi_h \quad \text{for all } \phi_h \in \mathring{W}_h(U(T)). \tag{4.6}$$

For general $\Phi_H \in V_H$ we define the correction operator $Q_h^{\mathcal{U},\mathbf{3}} : V_H \to V_h$ by

$$Q_h^{\mathcal{U},\mathbf{3}}(\Phi_H)(x) := \sum_{T\in\mathcal{T}_H}\sum_{i=1}^{d} \partial_{x_i}\Phi_H(x_T) w_{h,T,i}^{\mathcal{U},\mathbf{3}}(x).$$

The global coarse scale approximation $u_H^{\mathbf{3}} \in V_H$ is the solution of (3.3)*; i.e., it solves*

$$\begin{aligned}\mathcal{A}^{\mathbf{3}}(u_H^{\mathbf{3}}, \Phi_H) &:= \int_\Omega A\left(\nabla u_H^{\mathbf{3}} + \nabla Q_h^{\mathcal{U},\mathbf{3}}(u_H^{\mathbf{3}})\right) \cdot \left(\nabla \Phi_H + \nabla Q_h^{\mathcal{U},\mathbf{3}}(\Phi_H)\right) \\ &= \int_\Omega f(\Phi_H + Q_h^{\mathcal{U},\mathbf{3}}(\Phi_H)) \quad \textit{for all } \Phi_H \in V_H. \end{aligned} \tag{4.7}$$

The corresponding MsFEM approximation is given by

$$u_H^{\mathbf{3},\mathrm{MsFEM}} := u_H^{\mathbf{3}} + Q_h^{\mathcal{U},\mathbf{3}}(u_H^{\mathbf{3}}).$$

Using the above definition of the localized space $\mathring{W}_h(U(T))$ does not assure that our new method boils down to the classical MsFEM in the case without oversampling. Nevertheless, this can be achieved by introducing a localized interpolation operator. Given some element $T \in \mathcal{T}_H$ and an admissible patch $U(T)$, we can define $I_H^{U(T)}$ to be the Clément-type quasi-interpolation operator with respect to the domain $U(T)$ (with extension by zero in $\Omega \setminus U(T)$). Then, the localized space $\mathring{W}_h(U(T))$ can be defined in analogy to W_h with $I_H^{U(T)}$ replacing I_H. With this modification we obtain the classical MsFEM for $U(T) = T$. This is only a subtle detail, and all results still remain valid for these modified local spaces; however, this version would generate some technicalities in the proofs later on, which is why we decided to work with the definition (4.5).

Remark 4.2. The crucial differences between the classical Oversampling Strategies 1 and 2 and Oversampling Strategy 3 are the following:

(a) The variational problem for the local corrector in Oversampling Strategy 3 is posed in the constrained space $\mathring{W}_h(U(T))$, whereas the classical corrector problem seeks the local corrector in the full space $\mathring{V}_h(U(T))$ restricted to the patch.

(b) The support of the integrals on the right-hand sides in (3.7) and (3.9) is $U(T)$. In our new version we use only the element T. This allows us to exploit nice summation properties of the local projectors, without using indicator functions χ_T that lead to discontinuities.

(c) In the classical setting, the local correctors are restricted to the corresponding elements to derive the global corrector. For Oversampling Strategy 3, we simply sum up (weighted by the coefficients of the finite element function) the local contributions to get the global corrector. Note that our global corrector is conforming in the sense that its image is a subset of $V_h \subset H_0^1(\Omega)$, whereas the classical setup leads to a nonconforming corrector.

(d) In Oversampling Strategy 3, we do not use a PG formulation for the global problem (4.7). Since A is assumed symmetric, a symmetric discretization appears more natural. Furthermore, we immediately inherit coercivity for the global bilinear form $\mathcal{A}^{\mathbf{3}}$. This gives us the existence and uniqueness of $u_H^{\mathbf{3}}$, and the arising MsFEM approximation is well posed and the method stable. The typical disadvantage of the symmetric version, which still suffers from resonance errors (which is why the PG formulation is typically preferred), does not remain for our strategy.

(e) In contrast to Oversampling Strategies 1 and 2, the corrector $Q_h^{\mathcal{U},\mathbf{3}}(\Phi_H)$ does not preserve the support of Φ_H. In other words, the set of multiscale basis functions $\Phi_z + Q_h^{\mathcal{U},\mathbf{3}}(\Phi_z)$ with $z \in \mathcal{N}_H$ has an extended support. This results in a loss of sparsity in the stiffness matrix that corresponds with the global

problem (4.7). In order to still assemble the stiffness matrix in an efficient way, one might store the intersection domain for each two given oversampling patches (in storage types with low memory requirements). This can be easily done at the same time the grids for the local patches are being generated. Once all intersection domains are available, the matrix can be assembled efficiently. A quadrature rule that resolves the microstructure is needed for each of the strategies.

Remark 4.3 (perturbation of the right-hand side). We might also replace the right-hand side of (4.7) by the term $\int_\Omega f\Phi_H$. This introduces only a perturbation of order $\|Hf\|_{L^2(\Omega)}$ in the H^1-error.

Remark 4.4 (nonsymmetric formulation). As for the classical strategies, one might also consider the nonsymmetric PG formulations: find $u_H^{\mathbf{3}} \in V_H$ such that

$$\int_\Omega A\left(\nabla(u_H^{\mathbf{3}} + Q_h^{\mathcal{U},\mathbf{3}}(u_H^{\mathbf{3}}))\right)\cdot\nabla\Phi_H = \int_\Omega f\Phi_H \quad \text{for all } \Phi_H \in V_H$$

or

$$\int_\Omega A\left(\nabla u_H^{\mathbf{3}}\right)\cdot\nabla(\Phi_H + Q_h^{\mathcal{U},\mathbf{3}}(\Phi_H)) = \int_\Omega f\Phi_H \quad \text{for all } \Phi_H \in V_H.$$

In the spirit of homogenization theory, one might pose the question of whether Theorem 3.5 still holds for MsFEM approximations obtained with Oversampling Strategy 3. At least, this seems to be likely. The reason is that Theorem 3.5 in particular covers the case without oversampling (see also [17]), and the proof given in [18] goes back to the arguments used for the case without oversampling. But for $\mathcal{U} = \mathcal{T}_H$ (no oversampling), Oversampling Strategies 1 and 2 are identical, and Oversampling Strategy 3 is at least close to the classical approach. Especially concerning Oversampling Strategy 3, if the thickness of the oversampling layer decreases faster than the coarse mesh size, we are almost in the case of Oversampling Strategy 1, up to a small perturbation of the source term that is of order $\max_{T\in\mathcal{T}_H}\frac{|U(T)\setminus T|}{|T|}$ and that converges to zero under the assumptions of Theorem 3.5. However, such arguments still need a detailed investigation. In this sense, one might carefully study whether Oversampling Strategy 3 also covers the homogenization setting established by Gloria, with $u_H^{\mathbf{3}}$ converging to the homogenized solution as in Theorem 3.5. This might be an interesting result to ensure that Oversampling Strategy 3 is not worse than the classical strategies with respect to a homogenization setting.

Besides the advantages of our new strategy mentioned previously, e.g., its conformity, stability, and unique solvability, we formulate the main error estimate, which is proved in subsection 4.2.

THEOREM 4.5 (quantitative a priori error estimates). *Assume that we have $A \in L^\infty(\Omega, \mathbb{R}^{d\times d}_{sym})$ and $f \in L^2(\Omega)$ as in the general assumptions in section* 2.1. *Let $\mathcal{T}_H$ be a given coarse triangulation, and let $\mathcal{U}$ denote a corresponding set of admissible patches, with the property $d_{\mathcal{U}}^{\min} \gtrsim H\log(H^{-1})$. By $\mathcal{T}_h$ we denote a sufficiently accurate fine triangulation of Ω and by u_h the associated finite element solution of* (2.3). *If $u_H^{\mathbf{3},\mathrm{MsFEM}}$ is the MsFEM approximation determined with Oversampling Strategy* 3 *and if $u_H^{\mathbf{3}}$ denotes the corresponding coarse part, then the following a priori error estimates holds true for arbitrary mesh sizes $H \geq h$:*

$$\begin{aligned}\|\nabla u_h - \nabla u_H^{\mathbf{3},\mathrm{MsFEM}}\|_{L^2(\Omega)} &\leq CH,\\ \|u_h - u_H^{\mathbf{3},\mathrm{MsFEM}}\|_{L^2(\Omega)} &\leq CH^2,\\ \|u_h - u_H^{\mathbf{3}}\|_{L^2(\Omega)} &\leq CH.\end{aligned}$$

Here, C denotes generic constants that depend on f, $\gamma_{\min}$, and $\gamma_{\max}$ but not *on H, h, the regularity of the exact solution, or the variations of A. Details on the constants are given in Theorems* 4.13 *and* 4.15.

4.2. Proof of the main result. Before we prove the error estimates for the MsFEM with the correctors presented in Oversampling Strategy 3, we introduce some simplifying notation for this subsection.

DEFINITION 4.6 (notation for Oversampling Strategy 3). *Let $w_{h,T,i}^{\mathcal{U},\mathbf{3}} \in \mathring{W}_h(U(T))$ denote the local corrector basis given by* (4.6)*, let $Q_h^{\mathcal{U},\mathbf{3}}$ denote the corresponding corrector operator from Oversampling Strategy* 3*, and let $u_H^{\mathbf{3}}$ denote the arising (coarse) MsFEM approximation. In the following, we skip the redundant indices and use the following notation:*

$$w_T^i := w_{h,T,i}^{\mathcal{U},\mathbf{3}}, \quad Q_h := Q_h^{\mathcal{U},\mathbf{3}}, \quad u_H := u_H^{\mathbf{3}}, \quad \text{and} \quad u^{MsFEM} := u_H + Q_h(u_H).$$

The first lemma treats the (unpractical) case of maximal oversampling.

LEMMA 4.7 (error estimate for maximal oversampling). *Let $U(T) = \Omega$ for all $T \in \mathcal{T}_H$. Then the multiscale approximation u_H that solves* (4.7) *satisfies the error estimate*

$$\|\nabla u_h - \nabla(u_H + Q_h(u_H))\|_{L^2(\Omega)} \lesssim \gamma_{\min}^{-1} \|Hf\|_{L^2(\Omega)},$$

where u_h solves the reference problem (2.3)*.*

If, moreover, $(f, w_h)_{L^2(\Omega)} = 0$ for all fine scale functions $w_h \in W_h$, then $u_H + Q_h(u_H) = u_h$.

Proof. For $U(T) = \Omega$, Q_h maps onto the fine scale space W_h. Given $\Phi_H \in V_H$, it is easily checked that $Q_h(\Phi_H) = \sum_{T\in\mathcal{T}_H} \sum_{i=1}^d \partial_{x_i}\Phi_H(x_T) w_T^i$ satisfies

$$\begin{aligned}
a(Q_h(\Phi_H), \phi_h) &= \int_\Omega A\nabla \left(\sum_{T\in\mathcal{T}_H} \sum_{i=1}^d \partial_{x_i}\Phi_H(x_T) w_T^i(x) \right) \cdot \nabla\phi_h \\
&= \int_\Omega A \left(\sum_{T\in\mathcal{T}_H} \sum_{i=1}^d \partial_{x_i}\Phi_H(x_T) \nabla w_T^i(x) \right) \cdot \nabla\phi_h \\
&= \sum_{T\in\mathcal{T}_H} \sum_{i=1}^d \partial_{x_i}\Phi_H(x_T) \int_\Omega A\nabla w_T^i(x) \cdot \nabla\phi_h \\
&= -\sum_{T\in\mathcal{T}_H} \sum_{i=1}^d \partial_{x_i}\Phi_H(x_T) \int_T Ae_i \cdot \nabla\phi_h \\
&= -\sum_{T\in\mathcal{T}_H} \int_T A\nabla\Phi_H \cdot \nabla\phi_h \\
&= -a(\Phi_H, \phi_h)
\end{aligned}$$

for all $\phi_h \in W_h$. This means that Q_h is the orthogonal projection of Φ_H onto the fine scale space W_h with respect to the scalar product $a(\cdot,\cdot)$. This yields the orthogonal decomposition

$$V_h = \tilde{V}_H \oplus_{\perp_a} W_h, \text{ where } \tilde{V}_H := \{\Phi_H + Q_h(\Phi_H) \mid \Phi_H \in V_H\}. \tag{4.8}$$

Moreover, Galerkin orthogonality holds; i.e., for $e_h := u_h - (u_H + Q_h(u_H))$ and for arbitrary $\Phi_H + Q_h(\Phi_H) \in \tilde{V}_H$,

$$a(e_h, \Phi_H + Q_h(\Phi_H)) = a(u_h, \Phi_H + Q_h(\Phi_H)) - a(u_H + Q_h(u_H), \Phi_H + Q_h(\Phi_H))$$
$$\overset{(4.7)}{=} 0. \tag{4.9}$$

The combination of (4.8) and (4.9) shows that $e_h \in W_h$, and therefore $I_H(e_h) = 0$. We obtain

$$\gamma_{\min}\|\nabla e_h\|^2_{L^2(\Omega)} \le a(e_h, e_h) = a(u_h, e_h) = \int_\Omega f e_h = \int_\Omega f(e_h - I_H(e_h)).$$

The application of the Cauchy–Schwarz inequality on the element level and the estimate (4.4) for the interpolation error yield the assertion. ∎

COROLLARY 4.8. *The new MsFEM is exact (up to the discretization error on the fine scale and oscillations $\|Hf\|_{L^2(\Omega)}$ of the right-hand side f) in the limit of maximal oversampling. This results holds true independent of the upper spectral bound $\gamma_{\max}$ and the variations of A. This is the next difference from the previous Oversampling Strategies* 1 *and* 2.

Although the error estimate in Lemma 4.7 is encouraging, maximal oversampling is not feasible. We shall study the decay of the correctors away from element they are associated with. For all $T \in \mathcal{T}_H$, define element patches in the coarse mesh $\mathcal{T}_H$ by

$$\begin{aligned} &U_0(T) := T, \\ &U_k(T) := \cup\{T' \in \mathcal{T}_H \mid T' \cap U_{k-1}(T) \neq \emptyset\} \quad k = 1, 2, \dots. \end{aligned} \tag{4.10}$$

LEMMA 4.9 (decay of the ideal correctors). *Let $U(T) = \Omega$ for all $T \in \mathcal{T}_H$ in Oversampling Strategy* 3*, and let w_T^i denote the corresponding local correctors defined in Definition* 4.6 *(and* (4.6)*). Then, for all $T \in \mathcal{T}_H$ and all $k \in \mathbb{N}$,*

$$\|A^{1/2}\nabla w_T^i\|_{L^2(\Omega\setminus U_k(T))} \lesssim e^{-rk}\|A^{1/2}\nabla w_T^i\|_{L^2(\Omega)},$$

where r is a positive constant that depends on the square root of the contrast but not on the mesh size or the variations of A.

The proof of Lemma 4.9 requires the definition of cutoff functions and an additional lemma. For $T \in \mathcal{T}_H$ and $\ell, k \in \mathbb{N}$ with $k > \ell$, define $\eta_{T,k,\ell} \in P_1(\mathcal{T}_H)$ with nodal values

$$\begin{aligned} &\eta_{T,k,\ell}(z) = 0 \quad \text{for all } z \in \mathcal{N}_H \cap U_{k-\ell}(T), \\ &\eta_{T,k,\ell}(z) = 1 \quad \text{for all } z \in \mathcal{N}_H \cap (\Omega \setminus U_k(T)), \text{ and} \\ &\eta_{T,k,\ell}(z) = \frac{m}{\ell} \quad \text{for all } x \in \mathcal{N}_H \cap \partial U_{k-\ell+m}(T),\ m = 0, 1, 2, \dots, \ell. \end{aligned} \tag{4.11}$$

For a sketch in one dimension, see Figure 4.1.

Given some $w \in W_h$, the product $\eta_{T,k,\ell}w$ is not in W_h in general. However, the distance of $\eta_{T,k,\ell}w$ and W_h is small in the following sense.

LEMMA 4.10. *Given $w \in W_h$ and some cutoff function $\eta_{T,k,\ell} \in P_1(\mathcal{T}_H)$ as in* (4.11)*, there exists some $\tilde{w} \in \mathring{W}_h(\Omega \setminus U_{k-\ell-1}(T)) \subset W_h$ such that*

$$\|\nabla(\eta_{T,k,\ell}w - \tilde{w})\|_{L^2(\Omega)} \lesssim \ell^{-1}\|\nabla w\|_{L^2(U_{k+2}(T)\setminus U_{k-\ell-2}(T))}.$$

Proof. Fix some $T \in \mathcal{T}_H$ and $k \in \mathbb{N}$, and let $\eta_\ell := \eta_{T,k,\ell}$. The operator $I_h : H^1_0(\Omega) \cap C(\bar{\Omega}) \to V^h$ denotes the nodal interpolant with respect to the mesh $\mathcal{T}_h$.

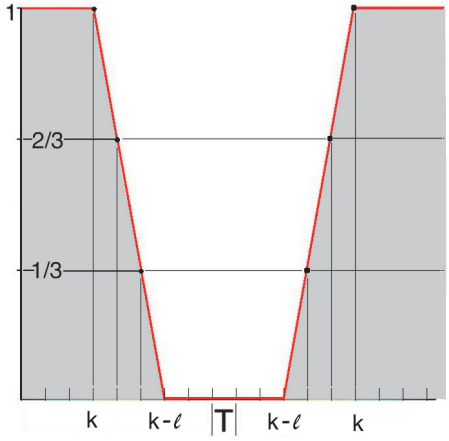


FIG. 4.1. *Sketch of $\eta_{T,k,\ell}$ (red curve) in one dimension for $k = 5$ and $\ell = 3$. $\eta_{T,k,\ell}$ is equal to zero on T and also on the first 2 $(= k - \ell)$ coarse grid layers around T. Then it grows linearly on the layers $\ell = 3$ until $k = 5$. On the remaining layers $\eta_{T,k,\ell}$ is constantly equal to 1.*

Recall that for all quadratic polynomials p and all $t \in \mathcal{T}_h$, I_h fulfills the (local) approximation and stability estimates

$$\|\nabla(p - I_h p)\|_{L^2(t)} \lesssim h_t \|\nabla^2 p\|_{L^2(t)} \quad \text{and} \quad \|\nabla(I_h p)\|_{L^2(t)} \lesssim \|\nabla p\|_{L^2(t)}. \tag{4.12}$$

We will use this estimate for the $\mathcal{T}_h$-piecewise quadratic function $p = \eta_\ell w$. Since $\nabla^2 \eta_\ell = \nabla^2 w = 0$ in every $t \in \mathcal{T}_h$, we have that $\nabla_h^2 \eta_\ell w = \nabla \eta_\ell \cdot \nabla w$ in t.

According to [34, Lemma 1], there exists some $v \in V^h$ such that

$$I_H v = I_H I_h(\eta_\ell w),\ \|\nabla v\|_{L^2(\Omega)} \lesssim \|\nabla I_H I_h(\eta_\ell w)\|_{L^2(\Omega)},\ \text{and}\ \operatorname{supp}(v) \subset \Omega \setminus U_{k-\ell-1}(T). \tag{4.13}$$

Hence, $\tilde{w} := I_h(\eta_\ell w) - v \in \mathring{W}_h(\Omega \setminus U_{k-\ell-1}(T))$. Since $I_H I_h(cw) = c I_H w = 0$ for any $c \in \mathbb{R}$, we set $c_K^\ell := |\omega_K|^{-1} \int_{\omega_K} \eta_\ell$ for $K \in \mathcal{T}_H$ and get

$$\begin{aligned}
&\|\nabla I_H I_h(\eta_\ell w)\|_{L^2(\Omega)}^2 \\
&\overset{(4.11)}{=} \sum_{\substack{K \in \mathcal{T}_H: \\ K \subset \overline{U_{k+1}(T)} \setminus U_{k-\ell-1}(T)}} \left\| \nabla I_H I_h \left(\left(\eta_\ell - c_K^\ell \right) w \right) \right\|_{L^2(K)}^2 \\
&\overset{(4.12),(4.4)}{\lesssim} \sum_{\substack{K \in \mathcal{T}_H: \\ K \subset \overline{U_{k+1}(T)} \setminus U_{k-\ell-1}(T)}} \left\| \nabla \left(\left(\eta_\ell - c_K^\ell \right) w \right) \right\|_{L^2(\omega_K)}^2 \\
&\overset{(4.2)}{\lesssim} \sum_{\substack{K \in \mathcal{T}_H: \\ K \subset \overline{U_{k+1}(T)} \setminus U_{k-\ell-1}(T)}} \|(\nabla \eta_\ell)(w - I_H w)\|_{L^2(\omega_K)}^2 + \left\| \left(\eta_\ell - c_K^\ell \right) \nabla w \right\|_{L^2(\omega_K)}^2 \\
&\overset{(4.11)}{\lesssim} \sum_{\substack{K \in \mathcal{T}_H: \\ K \subset \overline{U_k(T)} \setminus U_{k-\ell}(T)}} \|(\nabla \eta_\ell)(w - I_H w)\|_{L^2(K)}^2 \quad + \sum_{\substack{K \in \mathcal{T}_H: \\ K \subset \overline{U_{k+1}(T)} \setminus U_{k-\ell-1}(T)}} \left\| \left(\eta_\ell - c_K^\ell \right) \nabla w \right\|_{L^2(\omega_K)}^2
\end{aligned} \tag{4.14}$$

$$\lesssim \quad \|H\nabla\eta_\ell\|^2_{L^\infty(\Omega)}\|\nabla w\|^2_{L^2(U_{k+1}(T)\setminus U_{k-\ell-1}(T))} + \sum_{\substack{K\in\mathcal{T}_H:\\ K\subset\overline{U_{k+1}(T)}\setminus U_{k-\ell-1}(T)}} \big\|\big(\eta_\ell - c_K^\ell\big)\,\nabla w\big\|^2_{L^2(\omega_K)}$$

$$\lesssim \quad \|H\nabla\eta_\ell\|^2_{L^\infty(\Omega)}\|\nabla w\|^2_{L^2(U_{k+2}(T)\setminus U_{k-\ell-2}(T))}.$$

In the last step, we used the Lipschitz bound

$$\|\eta_\ell - c_K^\ell\|^2_{L^\infty(\omega_K)} \lesssim H^2\|\nabla\eta_\ell\|^2_{L^\infty(\omega_K)}.$$

In summary we get with the previous computations

$$\begin{aligned}
\|\nabla(\eta_\ell w - \tilde w)\|^2_{L^2(\Omega)} &\overset{(4.13)}{\lesssim} \|\nabla(\eta_\ell w - I_h(\eta_\ell w))\|^2_{L^2(\Omega)} + \|\nabla I_H I_h(\eta_\ell w)\|^2_{L^2(\Omega)}\\
&\overset{(4.12),(4.14)}{\lesssim} \|h\nabla\eta_\ell\cdot\nabla w\|^2_{L^2(\Omega)} + \|H\nabla\eta_\ell\|^2_{L^\infty(\Omega)}\|\nabla w\|^2_{L^2(U_{k+2}(T)\setminus U_{k-\ell-2}(T))}\\
&\overset{(4.11)}{\lesssim} \Big(\|h\nabla\eta_\ell\|^2_{L^\infty(\Omega)} + \|H\nabla\eta_\ell\|^2_{L^\infty(\Omega)}\Big)\|\nabla w\|^2_{L^2(U_{k+2}(T)\setminus U_{k-\ell-2}(T))}\\
&\overset{(4.11)}{\lesssim} \ell^{-2}\|\nabla w\|^2_{L^2(U_{k+2}(T)\setminus U_{k-\ell-2}(T))}.
\end{aligned}$$

This proves the assertion. □

Proof of Lemma 4.9. The proof exploits some recursive Caccioppoli argument as in [34]. We fix some $T\in\mathcal{T}_H$ and $k\in\mathbb{N}$. Given $\ell\in\mathbb{N}$ with $\ell < k-1$, let $\eta_\ell := \eta_{T,k-2,\ell-4}\in V_H$ be some cutoff function as in (4.11). Lemma 4.10 shows that there exists some $\tilde w_T^i\in W_h$ such that $\|\nabla(\eta_\ell w_T^i - \tilde w_T^i)\|_{L^2(\Omega)} \lesssim \ell^{-1}\|\nabla w_T^i\|_{L^2(U_k(T)\setminus U_{k-\ell}(T))}$. Since $\tilde w_T^i\in\mathring{W}_h(\Omega\setminus U_{k-\ell+1}(T))$ and, hence, $\tilde w_T^i|_T = 0$, it holds that

$$\int_{\Omega\setminus U_{k-\ell}(T)} A\nabla w_T^i\cdot\nabla\tilde w_T^i = \int_\Omega A\nabla w_T^i\cdot\nabla\tilde w_T^i = -\int_T Ae_i\cdot\nabla\tilde w_T^i = 0. \tag{4.15}$$

The definition of η_ℓ, the product rule, (4.6), and (4.2) yield

$$\begin{aligned}
\int_{\Omega\setminus U_k(T)} A\nabla w_T^i\cdot\nabla w_T^i &\le \int_{\Omega\setminus U_{k-\ell}(T)} \eta_\ell A\nabla w_T^i\cdot\nabla w_T^i\\
&= \int_{\Omega\setminus U_{k-\ell}(T)} A\nabla w_T^i\cdot\big(\nabla(\eta_\ell w_T^i) - w_T^i\nabla\eta_\ell\big)\\
&\overset{(4.15)}{=} \int_{\Omega\setminus U_{k-\ell}(T)} A\nabla w_T^i\cdot\left(\nabla(\eta_\ell w_T^i - \tilde w_T^i) - (w_T^i - \underbrace{I_H(w_T^i)}_{=0})\nabla\eta_\ell\right).
\end{aligned}$$

Observe that, by (4.11), $\|\nabla\eta_\ell\|_{L^\infty(K)} = |\nabla\eta_\ell(x_K)| \lesssim \ell^{-1}H_K^{-1}$ for all $K\in\mathcal{T}_H$. This and the estimate (4.4) for the interpolation error show that

$$\begin{aligned}
\|(w_T^i - I_H(w_T^i))\nabla\eta_\ell\|^2_{L^2(K)} &\lesssim \|\nabla\eta_\ell\|^2_{L^\infty(K)}\|w_T^i - I_H(w_T^i)\|^2_{L^2(K)}\\
&\lesssim H_K^2\|\nabla\eta_\ell\|^2_{L^\infty(K)}\|\nabla w_T^i\|^2_{L^2(\omega_K)}\\
&\lesssim \ell^{-2}\|\nabla w_T^i\|^2_{L^2(\omega_K)}
\end{aligned}$$

for any $K\in\mathcal{T}_H$. The combination of the previous estimates and Cauchy–Schwarz inequalities proves that there is some constant $C_1 > 0$ independent of T, ℓ, k, and the oscillations of A such that

$$\|A^{1/2}\nabla w_T^i\|_{L^2(\Omega\setminus U_k(T))} \le C_1\ell^{-1/2}\|A^{1/2}\nabla w_T^i\|_{L^2(\Omega\setminus U_{k-\ell-1}(T))}. \tag{4.16}$$

The choice $\ell := \lceil C_1\, e \rceil$ and the recursive application of (4.16) readily yield the assertion. ∎

This exponential decay justifies the approximation of the correctors on local patches $U_k(T)$ as proposed in (4.10). We denote by Q_h^k the corrector that corresponds to the choice $U(T) = U_k(T)$ in Oversampling Strategy 3 and by Q_h^Ω the one for $U(T) = \Omega$.

COROLLARY 4.11 (truncation/localization error). *Let $U(T) = \Omega$ for all $T \in \mathcal{T}_H$ in Oversampling Strategy 3. Then, for all $T \in \mathcal{T}_H$ and all $k \in \mathbb{N}$,*

$$\|A^{1/2}\nabla(w_T^i - w_T^{i,k})\|_{L^2(\Omega\setminus U_k(T))} \lesssim e^{-r\cdot k}\|A^{1/2}e_i\|_{L^2(T)},$$

where $r > 0$ is as in Lemma 4.9 (independent of the variations of A or the mesh size).

Proof. Galerkin orthogonality yields

$$\|A^{1/2}\nabla(w_T^i - w_T^{i,k})\|^2_{L^2(\Omega)} \le \|A^{1/2}\nabla(w_T^i - \tilde{w})\|^2_{L^2(U_{k-1}(T))} + \|A^{1/2}\nabla w_T^i\|^2_{L^2(\Omega\setminus U_{k-1}(T))},$$

where $\tilde{w} \in W_h$ is the fine scale function that corresponds to $(1-\eta_{T,k-1,1})w_T^i$ and which is constructed in the same way as $\tilde{w}$ in the proof of Lemma 4.10. Here, $\eta_{T,k-1,1}$ is some cutoff function as in (4.11). Since $\operatorname{supp}(\tilde{w}) \subset \operatorname{supp}((1-\eta_{T,k-1,1})w_T^i) \subset U_{k-1}(T)$, we have that $\tilde{w} \in \mathring{W}_h(U_k(T))$ and the use of Galerkin orthogonality is justified. Proceeding as in Lemma 4.10 shows that

$$\|A^{1/2}\nabla(w_T^i - w_T^{i,k})\|^2_{L^2(\Omega)} \lesssim \|A^{1/2}\nabla w_T^i\|^2_{L^2(\Omega\setminus U_{k-2}(T))},$$

and the application of Lemma 4.9 yields the assertion. ∎

The proof of the main theorem requires one technical result.

LEMMA 4.12. *Let $k \in \mathbb{N}_{>0}$, and let $\Phi_H \in V_H$; then*

(4.17)
$$\left\|A^{1/2}\nabla(Q_h^\Omega - Q_h^k)\Phi_H\right\|^2_{L^2(\Omega)} \lesssim k^d \sum_{T\in\mathcal{T}_H}\sum_{i=1}^d |\partial_{x_i}\Phi_H(x_T)|^2 \left\|A^{1/2}\nabla(w_T^i - w_T^{i,k})\right\|^2_{L^2(\Omega)}.$$

Proof. Let $\eta_{T,k,1}$ be as in (4.11), and define $z := (Q_h^\Omega - Q_h^k)\Phi_H \in W_h$. We decompose the error as follows:

$$\begin{aligned}\left\|A^{1/2}\nabla(Q_h^\Omega - Q_h^k)\Phi_H\right\|^2_{L^2(\Omega)} &= a(z,z)\\ &= \underbrace{\sum_{T\in\mathcal{T}_H}\sum_{i=1}^d \partial_{x_i}\Phi_H(x_T)\, a(w_T^i - w_T^{i,k}, z(1-\eta_{T,k,1}))}_{=:\mathrm{I}}\\ &\quad + \underbrace{\sum_{T\in\mathcal{T}_H}\sum_{i=1}^d \partial_{x_i}\Phi_H(x_T)\, a(w_T^i - w_T^{i,k}, z\eta_{T,k,1})}_{=:\mathrm{II}}.\end{aligned}$$

For the first term we get

$$|\mathrm{I}| \lesssim \sum_{T\in\mathcal{T}_H}\sum_{i=1}^d |\partial_{x_i}\Phi_H(x_T)|\|A^{\frac12}\nabla(w_T^i - w_T^{i,k})\|_{L^2(\Omega)}\|\nabla\left(z(1-\eta_{T,k,1})\right)\|_{L^2(U_{k+1}(T))},$$

where with $I_H(z) = 0$

$$\begin{aligned}\|\nabla\left(z(1-\eta_{T,k,1})\right)\|_{L^2(U_{k+1}(T))} &\leq \|\nabla z\|_{L^2(U_{k+1}(T))} + \|z\nabla\left(1-\eta_{T,k,1}\right)\|_{L^2(U_{k+1}(T)\setminus U_k(T))}\\ &\lesssim \|\nabla z\|_{L^2(U_{k+1}(T))} + \frac{1}{H}\|z - I_H(z)\|_{L^2(U_{k+1}(T)\setminus U_k(T))}\\ &\lesssim \|\nabla z\|_{L^2(U_{k+2}(T))},\end{aligned}$$

and therefore

$$\begin{aligned}|\mathrm{I}| &\lesssim \sum_{T\in\mathcal{T}_H}\sum_{i=1}^{d}|\partial_{x_i}\Phi_H(x_T)|\|A^{\frac{1}{2}}\nabla(w_T^i - w_T^{i,k})\|_{L^2(\Omega)}\|\nabla z\|_{L^2(U_{k+2}(T))}\\ &\lesssim k^{\frac{d}{2}}\left(\sum_{T\in\mathcal{T}_H}\sum_{i=1}^{d}|\partial_{x_i}\Phi_H(x_T)|^2\|A^{\frac{1}{2}}\nabla(w_T^i - w_T^{i,k})\|_{L^2(\Omega)}^2\right)^{\frac{1}{2}}\|\nabla z\|_{H^1(\Omega)}.\end{aligned}$$

To estimate the second term, we use Lemma 4.10, which gives us the existence of some $\tilde{z}\in\mathring{W}_h(\Omega\setminus U_{k-2}(T))$ with $a(w_T^i - w_T^{i,k},\tilde{z}) = 0$ (as in (4.15)) and the property $\|\nabla(z\eta_{T,k,1}-\tilde{z})\|_{L^2(\Omega)}\lesssim\|\nabla z\|_{L^2(U_{k+2}(T))}$. This yields

$$\begin{aligned}|\mathrm{II}| &= \left|\sum_{T\in\mathcal{T}_H}\sum_{i=1}^{d}\partial_{x_i}\Phi_H(x_T)a(w_T^i - w_T^{i,k}, z\eta_{T,k,1}-\tilde{z})\right|\\ &\leq \sum_{T\in\mathcal{T}_H}\sum_{i=1}^{d}|\partial_{x_i}\Phi_H(x_T)|\|A^{\frac{1}{2}}\nabla(w_T^i - w_T^{i,k})\|_{L^2(\Omega)}\|\nabla z\|_{L^2(U_{k+2}(T))}\\ &\lesssim k^{\frac{d}{2}}\left(\sum_{T\in\mathcal{T}_H}\sum_{i=1}^{d}|\partial_{x_i}\Phi_H(x_T)|^2\|A^{\frac{1}{2}}\nabla(w_T^i - w_T^{i,k})\|_{L^2(\Omega)}^2\right)^{\frac{1}{2}}\|\nabla z\|_{H^1(\Omega)}.\end{aligned}$$

Combining the estimates for I and II and dividing by $\|\nabla z\|_{H^1(\Omega)}\lesssim a(z,z)^{\frac{1}{2}}$ yields the assertion. ∎

THEOREM 4.13 (H^1–error estimate). *Given $k\in\mathbb{N}$, let $U(T) = U_k(T)$ for all $T\in\mathcal{T}_H$ in Oversampling Strategy 3. Then the multiscale approximation u_H^k that solves (4.7) satisfies the error estimate*

$$\|\nabla u_h - \nabla(u_H^k + Q_h^k(u_H^k))\|\lesssim\gamma_{\min}^{-1}\|Hf\|_{L^2(\Omega)} + k^{\frac{d}{2}}e^{-rk}\|f\|_{H^{-1}(\Omega)},$$

where u_h is the reference solution from (2.3) and $r>0$ as in Lemma 4.9.

Remark 4.14 (relation to the results in [34]). In the case of maximal oversampling, the new MsFEM with constrained oversampling coincides with the ideal version (without localization) of the VMM presented in [34]. The localized versions are different and allow similar, but not identical, error estimates. The upper bound obtained in [34] reads (up to some multiplicative constant) as

$$\|Hf\|_{L^2(\Omega)} + H^{-1}e^{-rk}\|f\|_{H^{-1}(\Omega)}.$$

Our new localization strategy allows for an improved estimate in the sense that the unpleasant factor H^{-1} does not appear. Note that the proof of the error estimate in Theorem 4.13 does not generalize to the localization strategy used in [34] and must therefore be seen independently. The reason is that the structure of the local problems

(4.6) gives us a nice summation property which we were able to exploit but which is not available in [34]. This observation indicates that better numerical approximations for equal sizes of oversampling patches are possible with our new approach. This will be investigated in future works.

Proof of Theorem 4.13. Using the fact that Galerkin approximation minimizes the error in the energy norm, we obtain with the definitions of u_H^k and u_h that for all $\Phi_H \in V_H$

(4.18)
$$\|A^{1/2}(\nabla u_h - \nabla(u_H^k + Q_h^k(u_H^k)))\|_{L^2(\Omega)} \leq \|A^{1/2}(\nabla u_h - \nabla(\Phi_H + Q_h^k(\Phi_H)))\|_{L^2(\Omega)}.$$

Let u_H be the solution of (4.7) with the ideal corrector $Q_h = Q_h^\Omega$. Then

$$\begin{aligned}
&\|A^{1/2}(\nabla u_h - \nabla(u_H^k + Q_h^k(u_H^k)))\|_{L^2(\Omega)} \\
&\overset{(4.18)}{\leq} \|A^{1/2}(\nabla u_h - \nabla(u_H + Q_h^k(u_H)))\|_{L^2(\Omega)} \\
&\leq \|A^{1/2}(\nabla u_h - \nabla(u_H + Q_h^\Omega(u_H)))\|_{L^2(\Omega)} \\
&\qquad + \|A^{1/2}(\nabla(u_H + Q_h^\Omega(u_H)) - \nabla(u_H + Q_h^k(u_H)))\|_{L^2(\Omega)} \\
&\lesssim \gamma_{\min}^{-1/2}\|Hf\|_{L^2(\Omega)} + \|A^{1/2}\nabla((Q_h^\Omega - Q_h^k)(u_H))\|_{L^2(\Omega)}.
\end{aligned}$$

By Corollary 4.11, we get

$$\begin{aligned}
&\|A^{1/2}\nabla((Q_h^\Omega - Q_h^k)(u_H))\|_{L^2(\Omega)}^2 \\
&= \left\| \sum_{T\in\mathcal{T}_H} \sum_{i=1}^d \partial_{x_i} u_H(x_T) A^{\frac{1}{2}} \nabla(w_T^i - w_T^{i,k}) \right\|_{L^2(\Omega)}^2 \\
&\overset{(4.17)}{\lesssim} k^d \sum_{T\in\mathcal{T}_H} \sum_{i=1}^d |\partial_{x_i} u_H(x_T)|^2 \left\| A^{\frac{1}{2}} \nabla(w_T^i - w_T^{i,k}) \right\|_{L^2(\Omega)}^2 \\
&\lesssim k^d e^{-2r\cdot k} \sum_{T\in\mathcal{T}_H} \sum_{i=1}^d |\partial_{x_i} u_H(x_T)|^2 \|A^{1/2} e_i\|_{L^2(T)}^2 \\
&\lesssim k^d e^{-2r\cdot k} \sum_{T\in\mathcal{T}_H} \sum_{i=1}^d \|A^{1/2}\nabla u_H\|_{L^2(T)}^2 \\
&\lesssim k^d e^{-2rk} \|f\|_{H^{-1}(\Omega)}^2.
\end{aligned}$$

In the last step we have used that $u_H = I_H(u_H + Q_h^\Omega(u_H))$, the stability of I_H, and the energy estimate $\|A^{1/2}\nabla(u_H + Q_h^\Omega(u_H)\|_{L^2(T)} \lesssim \gamma_{\min}^{-1/2}\|f\|_{H^{-1}(\Omega)}$. ∎

THEOREM 4.15 (L^2-estimates). *Given $k \in \mathbb{N}$, let $U(T) = U_k(T)$ for all $T \in \mathcal{T}_H$ in Oversampling Strategy* 3. *Then the multiscale approximation u_H^k that solves* (4.7) *satisfies the error estimates*

$$\|u_h - (u_H^k + Q_h^k(u_H^k))\|_{L^2(\Omega)} \lesssim (\gamma_{\min}^{-1}\|H\|_{L^\infty(\Omega)} + k^{d/2}e^{-rk})^2\|f\|_{L^2(\Omega)}$$

and

$$\|u_h - u_H^k\|_{L^2(\Omega)} \lesssim \min_{v_H\in V_H} \|u_h - v_H\|_{L^2(\Omega)} + (\gamma_{\min}^{-1}\|H\|_{L^\infty(\Omega)} + k^{d/2}e^{-rk})^2\|f\|_{L^2(\Omega)},$$

where u_h is the reference solution from (2.3) *and r is a positive constant.*

Proof. A standard Aubin–Nitsche duality argument yields the first estimate. The second estimate follows from the first one and the quasi optimality and stability of the interpolation I_H in $L^2(\Omega)$. ∎

Remark 4.16 (smooth coefficient with known smallest scale ε). Let Ω be convex, and let $f \in L^2(\Omega)$ with $\|f\|_{L^2(\Omega)} \lesssim 1$, $A \in W^{1,\infty}(\Omega)$ with $\|\nabla A\|_{L^\infty(\Omega)} \lesssim \varepsilon^{-1}$ with some small scale parameter $\varepsilon > 0$. Choose uniform meshes $\mathcal{T}_H$ and $\mathcal{T}_h$ with $H \gtrsim \varepsilon \gtrsim h$. Under these assumptions, the error of the reference solution $u_h \in V_h$ is bounded as follows:

$$\|\nabla(u - u_h)\| \lesssim h\varepsilon^{-1}.$$

We refer the reader to [37] for details. If $k \gtrsim \log(H^{-1})$, Theorems 4.13 and 4.15 yield the error bounds

$$\begin{aligned} \|\nabla(u - u_H^k - Q_h^k(u_H^k))\|_{L^2(\Omega)} &\lesssim H + \tfrac{h}{\varepsilon}, \\ \|u - u_H^k - Q_h^k(u_H^k)\|_{L^2(\Omega)} &\lesssim H^2 + \left(\tfrac{h}{\varepsilon}\right)^2, \\ \|u - u_H^k\|_{L^2(\Omega)} &\lesssim H + \left(\tfrac{h}{\varepsilon}\right)^2. \end{aligned}$$

5. Numerical experiments. In this section we present numerical experiments to confirm the derived error estimates and to compare the numerical accuracies of Oversampling Strategies 1, 2, and 3. Here we use Oversampling Strategies 1 and 2 in the PG formulation (due to the findings in [28]) and Oversampling Strategy 3 in the symmetric formulation. We consider the following model problem.

Model problem. Let $\Omega :=]0,1[^2$ and $\epsilon = 5 \cdot 10^{-2}$. We define

$$u(x_1, x_2) := \sin(2\pi x_1)\sin(2\pi x_2) + \frac{\varepsilon}{2}\cos(2\pi x_1)\sin(2\pi x_2)\sin\left(2\pi\frac{x_1}{\varepsilon}\right),$$

which is the exact solution of the problem

$$\begin{aligned} -\nabla \cdot (A\nabla u) &= f \quad \text{in } \Omega, \\ u &= 0 \quad \text{on } \partial\Omega, \end{aligned}$$

where A is given by

$$A(x_1, x_2) := \frac{1}{8\pi^2}\begin{pmatrix} 2(2 + \cos(2\pi\frac{x_1}{\varepsilon}))^{-1} & 0 \\ 0 & 1 + \frac{1}{2}\cos(2\pi\frac{x_1}{\varepsilon}) \end{pmatrix}$$

and f by

$$f(x) := -\nabla \cdot (A(x)\nabla u(x)) \approx \sin(2\pi x_1)\sin(2\pi x_2).$$

In Table 5.1 we depict the results for $h = 2^{-6}$ and various combinations of H with different numbers of oversampling layers. For a better illustration we state the number of fine grid layers and the number of coarse grid layers (k) that corresponds with that. The results in Table 5.1 match nicely with the analytically predicted behavior. In Table 5.2 we state a comparison between the L^2- and H^1-errors for the three oversampling strategies obtained for identical values of H, h, and $\mathcal{U}$. We observe that our oversampling strategy, in contrast to the classical ones, does not suffer from a loss in accuracy when H is close to the microscopic parameter ε. Moreover, the accuracy obtained for Oversampling Strategy 3 is very promising in general.

TABLE 5.1

Computations made for $h = 2^{-6}$. k denotes the number of coarse layers. u_h denotes the fine scale reference given by (2.3), and $u_H^{\mathbf{3},\mathrm{MsFEM}}$ denotes the MsFEM approximation obtained with Oversampling Strategy 3. The table depicts various errors between u_h and $u_H^{\mathbf{3},\mathrm{MsFEM}}$.

H	Fine layers	k	$\|u_h - u_H^{\mathbf{3},\mathrm{MsFEM}}\|_{L^2(\Omega)}$	$\|u_h - u_H^{\mathbf{3},\mathrm{MsFEM}}\|_{H^1(\Omega)}$
2^{-1}	16	0.5	0.490063	4.49575
2^{-2}	8	0.5	0.09491	1.66315
2^{-2}	24	1.5	0.06376	1.08960
2^{-3}	4	0.5	0.033691	1.017150
2^{-3}	8	1	0.007125	0.406317
2^{-3}	12	1.5	0.007115	0.331458
2^{-3}	16	2	0.003241	0.165703
2^{-4}	2	0.5	0.012808	0.655269
2^{-4}	4	1	0.004164	0.348814
2^{-4}	6	1.5	0.004029	0.329306
2^{-4}	8	2	0.001451	0.162747
2^{-4}	12	2.5	0.000850	0.114040
2^{-4}	16	3	0.000696	0.096378

TABLE 5.2

Computations made for $h = 2^{-6}$. k denotes the number of coarse layers. u_h denotes the fine scale reference given by (2.3), and $u_H^{\mathbf{i},\mathrm{MsFEM}}$ denotes the MsFEM approximation obtained with Oversampling Strategy 1. The error is denoted by $e_i := u_h - u_H^{\mathbf{i},\mathrm{MsFEM}}$. The second column depicts the number of fine grid layers.

		Oversampling Strategy 1		Oversampling Strategy 2		Oversampling Strategy 3	
H	k	$\|e_1\|_{L^2}$	$\|e_1\|_{H^1}$	$\|e_2\|_{L^2}$	$\|e_2\|_{H^1}$	$\|e_3\|_{L^2}$	$\|e_3\|_{H^1}$
2^{-2}	1	0.1399	1.9812	0.1399	1.9812	0.0638	1.0896
2^{-3}	1	0.0594	1.6250	0.0594	1.6250	0.0071	0.4063
2^{-3}	2	0.0593	1.6250	0.0593	1.6250	0.0032	0.1657
2^{-4}	1	0.0166	0.8067	0.0172	0.8048	0.0042	0.3488
2^{-4}	2	0.0160	0.8057	0.0168	0.7955	0.0015	0.1628
2^{-4}	3	0.0153	0.8016	0.0152	0.7937	0.0007	0.0964

6. Conclusion. In this work, we proposed a new oversampling strategy for the MsFEM, which generalizes the original method without oversampling. The new strategy is based on an additional constraint for the solution spaces of the local problems. The error analysis shows that oversampling layers of thickness $H\log(H^{-1})$ suffice to preserve the common convergence rates with respect to H without any preasymptotic effects. Moreover, this choice prevents resonance errors even for general L^∞ coefficients without any assumptions on the geometry of the microstructure or the regularity of A. In this respect, the method is reliable. The method is also efficient in the sense that structural knowledge about the coefficient, e.g., (local) periodicity or scale separation, may be exploited to reduce the number of corrector problems considerably. Whether the oversampling can be reduced to very small layers in the case of, e.g., periodicity should be investigated numerically and/or analytically in future works.

Acknowledgments. We gratefully acknowledge the helpful suggestions made by the anonymous referees, which greatly improved the presentation of the paper. Furthermore, we thank Antoine Gloria for pointing out useful references and Xiao-Hui Wu for the nice discussions at the SIAM Geoscience Conference 2013 and his helpful remarks.

REFERENCES

[1] A. Abdulle, *The finite element heterogeneous multiscale method: A computational strategy for multiscale PDEs*, in Multiple Scales Problems in Biomathematics, Mechanics, Physics and Numerics, GAKUTO Internat. Ser. Math. Sci. Appl. 31, Gakkōtosho, Tokyo, 2009, pp. 133–181.

[2] I. Babuska and R. Lipton, *Optimal local approximation spaces for generalized finite element methods with application to multiscale problems*, Multiscale Model. Simul., 9 (2011), pp. 373–406.

[3] F. Brezzi, L. P. Franca, T. J. R. Hughes, and A. Russo, $b = \int g$, Comput. Methods Appl. Mech. Engrg., 145 (1997), pp. 329–339.

[4] C. Carstensen and R. Verfürth, *Edge residuals dominate a posteriori error estimates for low order finite element methods*, SIAM J. Numer. Anal., 36 (1999), pp. 1571–1587.

[5] Z. Chen and T. Y. Hou, *A mixed multiscale finite element method for elliptic problems with oscillating coefficients*, Math. Comp., 72 (2003), pp. 541–576.

[6] Z. Chen and T. Y. Savchuk, *Analysis of the multiscale finite element method for nonlinear and random homogenization problems*, SIAM J. Numer. Anal., 46 (2008), pp. 260–279.

[7] C.-C. Chu, I. G. Graham, and T.-Y. Hou, *A new multiscale finite element method for high-contrast elliptic interface problems*, Math. Comp., 79 (2010), pp. 1915–1955.

[8] W. Deng, X. Yun, and C. Xie, *Convergence analysis of the multiscale method for a class of convection-diffusion equations with highly oscillating coefficients*, Appl. Numer. Math., 59 (2009), pp. 1549–1567.

[9] W. E and B. Engquist, *The heterogeneous multiscale methods*, Commun. Math. Sci., 1 (2003), pp. 87–132.

[10] W. E and B. Engquist, *Multiscale modeling and computation*, Notices Amer. Math. Soc., 50 (2003), pp. 1062–1070.

[11] W. E and B. Engquist, *The heterogeneous multi-scale method for homogenization problems*, in Multiscale Methods in Science and Engineering, Lect. Notes Comput. Sci. Eng. 44, Springer, Berlin, 2005, pp. 89–110.

[12] Y. Efendiev and T. Hou, *Multiscale finite element methods for porous media flows and their applications*, Appl. Numer. Math., 57 (2007), pp. 577–596.

[13] Y. Efendiev, T. Hou, and V. Ginting, *Multiscale finite element methods for nonlinear problems and their applications*, Commun. Math. Sci., 2 (2004), pp. 553–589.

[14] Y. Efendiev and T. Y. Hou, *Multiscale Finite Element Methods: Theory and Applications*, Surv. Tutor. Appl. Math. Sci. 4, Springer, New York, 2009.

[15] Y. Efendiev and A. Pankov, *Numerical homogenization of monotone elliptic operators*, Multiscale Model. Simul., 2 (2003), pp. 62–79.

[16] Y. R. Efendiev, T. Y. Hou, and X.-H. Wu, *Convergence of a nonconforming multiscale finite element method*, SIAM J. Numer. Anal., 37 (2000), pp. 888–910.

[17] A. Gloria, *An analytical framework for the numerical homogenization of monotone elliptic operators and quasiconvex energies*, Multiscale Model. Simul., 5 (2006), pp. 996–1043.

[18] A. Gloria, *An analytical framework for numerical homogenization. Part* II: *Windowing and oversampling*, Multiscale Model. Simul., 7 (2008), pp. 274–293.

[19] A. Gloria, *Reduction of the resonance error—Part* 1: *Approximation of homogenized coefficients*, Math. Models Methods Appl. Sci., 21 (2011), pp. 1601–1630.

[20] A. Gloria, *Numerical homogenization: Survey, new results, and perspectives*, in Mathematical and Numerical Approaches for Multiscale Problem, ESAIM Proc. 37, EDP Sci., Les Ulis, 2012, pp. 50–116.

[21] P. Henning, *Convergence of MsFEM approximations for elliptic, non-periodic homogenization problems*, Netw. Heterog. Media, 7 (2012), pp. 503–524.

[22] P. Henning and M. Ohlberger, *The heterogeneous multiscale finite element method for elliptic homogenization problems in perforated domains*, Numer. Math., 113 (2009), pp. 601–629.

[23] P. Henning and M. Ohlberger, *Error control and adaptivity for heterogeneous multiscale*

approximations of nonlinear monotone problems, Discrete Contin. Dyn. Syst. Ser. S, to appear.

[24] P. Henning and M. Ohlberger, *A Newton-scheme framework for multiscale methods for nonlinear elliptic homogenization problems*, in Proceedings of the Algoritmy 2012, 19th Conference on Scientific Computing, Vysoke Tatry, Podbanske, 2012, pp. 65–74.

[25] P. Henning, M. Ohlberger, and B. Schweizer, *An Adaptive Multiscale Finite Element Method*, Preprint 05/12 - N, University of Münster, Münster, Germany, 2012.

[26] 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 (1997), pp. 169–189.

[27] T. Y. Hou, X.-H. Wu, and Z. Cai, *Convergence of a multiscale finite element method for elliptic problems with rapidly oscillating coefficients*, Math. Comp., 68 (1999), pp. 913–943.

[28] T. Y. Hou, X.-H. Wu, and Y. Zhang, *Removing the cell resonance error in the multiscale finite element method via a Petrov-Galerkin formulation*, Commun. Math. Sci., 2 (2004), pp. 185–205.

[29] T. J. R. Hughes, *Multiscale phenomena: Green's functions, the Dirichlet-to-Neumann formulation, subgrid scale models, bubbles and the origins of stabilized methods*, Comput. Methods Appl. Mech. Engrg., 127 (1995), pp. 387–401.

[30] 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 (1998), pp. 3–24.

[31] V. V. Jikov, S. M. Kozlov, and O. A. Oleĭnik, *Homogenization of Differential Operators and Integral Functionals*, Springer-Verlag, Berlin, 1994.

[32] M. G. Larson and A. Målqvist, *Adaptive variational multiscale methods based on a posteriori error estimation: Energy norm estimates for elliptic problems*, Comput. Methods Appl. Mech. Engrg., 196 (2007), pp. 2313–2324.

[33] A. Målqvist, *Multiscale methods for elliptic problems*, Multiscale Model. Simul., 9 (2011), pp. 1064–1086.

[34] A. Målqvist and D. Peterseim, *Localization of elliptic multiscale problems*, Math. Comp., to appear.

[35] H. Owhadi and L. Zhang, *Metric-based upscaling*, Comm. Pure Appl. Math., 60 (2007), pp. 675–723.

[36] H. Owhadi and L. Zhang, *Localized bases for finite-dimensional homogenization approximations with nonseparated scales and high contrast*, Multiscale Model. Simul., 9 (2011), pp. 1373–1398.

[37] D. Peterseim and S. Sauter, *Finite elements for elliptic problems with highly varying, non-periodic diffusion matrix*, Multiscale Model. Simul., 10 (2012), pp. 665–695.

[38] H. Yserentant, *On the multilevel splitting of finite element spaces*, Numer. Math., 49 (1986), pp. 379–412.

A.3 A localized orthogonal decomposition method for semi-linear elliptic problems

ESAIM: Mathematical Modelling and Numerical Analysis **48**:1331–1349, 2014.

(with P. Henning and A. Målqvist)

ESAIM: M2AN 48 (2014) 1331–1349
DOI: 10.1051/m2an/2013141

ESAIM: Mathematical Modelling and Numerical Analysis
www.esaim-m2an.org

A LOCALIZED ORTHOGONAL DECOMPOSITION METHOD FOR SEMI-LINEAR ELLIPTIC PROBLEMS *, **

Patrick Henning[1], Axel Målqvist[1] and Daniel Peterseim[2]

Abstract. In this paper we propose and analyze a localized orthogonal decomposition (LOD) method for solving semi-linear elliptic problems with heterogeneous and highly variable coefficient functions. This Galerkin-type method is based on a generalized finite element basis that spans a low dimensional multiscale space. The basis is assembled by performing localized linear fine-scale computations on small patches that have a diameter of order $H|\log(H)|$ where H is the coarse mesh size. Without any assumptions on the type of the oscillations in the coefficients, we give a rigorous proof for a linear convergence of the H^1-error with respect to the coarse mesh size even for rough coefficients. To solve the corresponding system of algebraic equations, we propose an algorithm that is based on a damped Newton scheme in the multiscale space.

Mathematics Subject Classification. 35J15, 65N12, 65N30.

Received November 14, 2012. Revised June 11, 2013.
Published online August 13, 2014.

1. Introduction

This paper is devoted to the numerical approximation of solutions of semi-linear elliptic problems with rapidly oscillating and highly varying coefficient functions. We are concerned with second-order partial differential equations of the type

$$-\nabla \cdot (A\nabla u) + F(u, \nabla u) = g$$

with prescribed (zero-) Dirichlet boundary condition for the unknown function u. Here, g is a given source term, A is a given highly variable diffusion matrix and F is a given highly variable nonlinear term that represents advective and reactive processes. In particular, we have a linear term of second order and nonlinear terms of order 1 and 0. A typical application is the stationary (Kirchhoff transformed) Richards equation that describes the groundwater flow in unsaturated soils (*cf.* [1,4,5]). The corresponding equation for the unknown generalized

Keywords and phrases. Finite element method, *a priori* error estimate, convergence, multiscale method, non-linear, computational homogenization, upscaling.

* *A. Målqvist and P. Henning are supported by The Göran Gustafsson Foundation and The Swedish Research Council.*

** *The research of D. Peterseim was supported by the Humboldt-Universität and the DFG Research Center Matheon Berlin.*

[1] Department of Information Technology, Uppsala University, Box 337, 75105 Uppsala, Sweden.

[2] Institut für Numerische Simulation der Universität Bonn, Wegelerstr. 66, 53123 Bonn, Germany.
patrick.henning@uni-muenster.de

pressure u reads

$$\nabla \cdot (K \nabla u) - \nabla \cdot (K\, kr(M(u))\vec{e}) = g,$$

where K is the hydraulic conductivity in the soil, kr the relative permeability depending on the saturation, M is some nonlinearity arising from the Kirchhoff transformation and $\vec{e}$ denotes the gravity vector. If we add an infiltration process, the equation receives an additional nonlinear reaction term.

The numerical treatment of such equations is often complicated and expensive. Due to the high variability of the coefficient functions, one requires extremely fine computational grids that are able to capture all the fine scale oscillations. Using standard methods such as Finite Element or Finite Volume schemes, this results in systems of equations of enormous size and therefore in a tremendous computational demand that can not be handled in a lot of scenarios.

Multiscale methods aim to overcome this difficulty by decoupling the fine scale computations into local parts. Prominent examples of multiscale methods are the Heterogeneous Multiscale Method (HMM) by E and Engquist [13] and the Multiscale Finite Element Method (MsFEM) proposed by Hou and Wu [20]. Both methods fit into a common framework and are strongly related to numerical homogenization (*cf.* [14,15,18]). HMM and MsFEM are typically not constructed for a direct approximation of exact solutions but for homogenized solutions and corresponding correctors instead. This implies that they are only able to approximate the exact solution up to a modeling error that depends crucially on the homogenization setting (*cf.* [14]). In the absence of strong assumptions like periodicity and scale separation, accurate approximations are therefore hard to achieve.

We are concerned with a multiscale method that is based on the concept of the Variational Multiscale Method (VMM) proposed by Hughes *et al.* [21]. In comparison to HMM and MsFEM, the VMM aims to a direct approximation of the exact solution without suffering from a modeling error remainder arising from homogenization theory. The key idea of the Variational Multiscale Method is to construct a splitting of the original solution space V into the direct sum of a low dimensional space for coarse grid approximations and high dimensional space for fine scale reconstructions. In this work, we consider a modification and extension of this idea that was developed in [27,30] and that was explicitly proposed in [31]. Here, the splitting is such that we obtain an accurate but low dimensional space V^{ms} (where we are looking for our fine scale approximation instead of an approximation of a coarse part) and a high dimensional residual space V^{f}. The construction of V^{ms} involves the computation of one fine scale problem in a small patch per degree of freedom. Mesh-adaptive versions of the VMM with patch size control are discussed in [27–29,33]. The first rigorous proof of convergence was recently obtained in [31] for linear diffusion problems under minimal regularity assumptions.

In this contribution, we present an efficient way of handling semi-linear elliptic multiscale problems in the modified VMM framework, including a proof of convergence based on the techniques established in [31]. Even though the original problem is nonlinear, the local fine scale problems are purely linear that can be solved in parallel. The main result of this article is the optimal convergence of the H^1-error between exact solution u and its multiscale approximation u_H^{ms}. We show that, if the patch size is of order $H|\log(H)|$, the following error bound

$$\|u - u_H^{\mathrm{ms}}\|_{H^1(\Omega)} \leq CH$$

holds with a generic constant C independent of the mesh size of the computational grid and the oscillations of A and F.

The paper is structured as follows. In Section 2 we introduce the setting of this paper, including the assumptions on the considered semi-linear problem. In Section 3 we present and motivate our method and we state the corresponding optimal convergence result. This result is then proved in Section 4. In Section 5, we propose an algorithm for the solution of the arising nonlinear algebraic equations. This algorithm is based on a damped Newton scheme in the multiscale space. Finally, Section 6 supports the theoretical results by a numerical experiment.

2. Setting

Let $\Omega \subset \mathbb{R}^d$ be a bounded Lipschitz domain with polyhedral boundary, let $V := H_0^1(\Omega)$ and let $A \in L^\infty(\Omega, \mathbb{R}^{d\times d}_{\text{sym}})$ denote a matrix valued function with uniformly strictly positive eigenvalues. We assume that the space $H_0^1(\Omega)$ is endowed with the H^1-semi norm given by $|v|_{H^1(\Omega)} := \|\nabla v\|_{L^2(\Omega)}$ (which is equivalent to the common H^1-norm in $H_0^1(\Omega)$). By $\langle\cdot,\cdot\rangle := (\cdot,\cdot)_{L^2(\Omega)}$ we denote the inner product in $L^2(\Omega)$ and $F : \Omega\times\mathbb{R}\times\mathbb{R}^d \to \mathbb{R}$ is a nonlinear measurable function.

Given some source term $g \in L^2(\Omega) \subset H^{-1}(\Omega)$ we are concerned to find $u \in H_0^1(\Omega)$ (*i.e.* with a homogeneous Dirichlet boundary condition) with

$$\langle A\nabla u, \nabla v\rangle + \langle F(\cdot, u, \nabla u), v\rangle = \langle g, v\rangle \tag{2.1}$$

for all test functions $v \in H_0^1(\Omega)$. To simplify the notation, we define the operator $B : H_0^1(\Omega) \to H^{-1}(\Omega)$ by

$$\langle B(v), w\rangle_{H^{-1},H_0^1} := \langle A\nabla v, \nabla w\rangle + \langle F(\cdot, v, \nabla v), w\rangle \quad \text{for } v, w \in H_0^1(\Omega),$$

where $\langle\cdot,\cdot\rangle_{H^{-1},H_0^1}$ denote the dual pairing in $H_0^1(\Omega)$.

Here, the diffusion diffusion matrix A may be strongly heterogeneous and highly variable. The non-linearity $F(\cdot, \xi, \zeta)$ may as well oscillate rapidly without any assumptions on the type of the oscillations. One application can be the Richards equation, which we will discuss more in Section 6.

However, we assume implicitly that the lower-order term F does not dominate the equation. In this regime, it is sufficient to construct a multiscale space independent of the non-linearity by solving local linear problems on the fine scale. If the lower-order term is dominant, some constants in our error analysis will be large and the proposed method needs modifications with respect to the construction of the multiscale basis. A typical example where the lower-order term is dominant is the modeling of transport of solutes in groundwater where one has to deal with extremely large Péclet numbers and a corresponding scaling of the advective terms. In this case, the resolution of oscillations of F is necessary for accurate upscaled and homogenized approximation (*cf.* [16, 17]).

For the subsequent analytical considerations and in order to guarantee a unique solution of (2.1), we make the following assumptions.

Assumption 1.

(A1) $A \in L^\infty\left(\Omega, \mathbb{R}^{d\times d}_{\text{sym}}\right)$ with

$$\infty > \beta := \|A\|_{L^\infty(\Omega)} = \operatorname*{ess\,sup}_{x\in\Omega} \sup_{\zeta\in\mathbb{R}^d\setminus\{0\}} \frac{A(x)\zeta\cdot\zeta}{|\zeta|^2}.$$

and there exists α such that

$$0 < \alpha := \operatorname*{ess\,inf}_{x\in\Omega} \inf_{\zeta\in\mathbb{R}^d\setminus\{0\}} \frac{A(x)\zeta\cdot\zeta}{|\zeta|^2},$$

(A2) There exist $L_1, L_2 \in \mathbb{R}_{>0}$ such that uniformly for almost every x in Ω:

$$\begin{aligned} |F(x,\xi_1,\zeta) - F(x,\xi_2,\zeta)| &\le L_1|\xi_1-\xi_2|, \quad \text{for all } \zeta\in\mathbb{R}^d,\ \xi_1,\xi_2\in\mathbb{R},\\ |F(x,\xi,\zeta_1) - F(x,\xi,\zeta_2)| &\le L_2|\zeta_1-\zeta_2|, \quad \text{for all } \zeta_1,\zeta_2\in\mathbb{R}^d,\ \xi\in\mathbb{R},\\ F(x,0,0) &= 0. \end{aligned}$$

(A3) B is strongly monotone, *i.e.* there exist $c_0 > 0$ so that for all $u, v \in H_0^1(\Omega)$:

$$\langle B(u) - B(v), u - v\rangle_{H^{-1},H_0^1} \ge c_0|u-v|^2_{H^1(\Omega)}. \tag{2.2}$$

Under assumptions (A1)−(A3), the Browder−Minty theorem (*cf.* [36], Sect. 3, Thm. 1.5 therewithin) yields a unique solution of problem (2.1).

Typically, the validity of Assumption (A3) can be checked by looking at the properties of the nonlinear function F. For instance, if there exists a constant $\alpha_0 \geq 0$, such that $\partial_\xi F(x,\xi,\zeta) \geq \alpha_0$ for all ζ and almost every x (*i.e.* $F(x,\cdot,\zeta)$ is monotonically increasing) and if α_0 and L_2 are such that $L_2 \leq 2\alpha_0$ and $L_2 < 2\alpha$ then (A3) is fulfilled. This can be checked by a simple calculation:

$$\begin{aligned}\langle B(u) - B(v), u - v\rangle_{H^{-1},H_0^1} &\geq \alpha||\nabla u - \nabla v||^2_{L^2(\Omega)} + \alpha_0||u - v||^2_{L^2(\Omega)} - L_2(|u - v|, |\nabla u - \nabla v|)_{L^2(\Omega)} \\ &\geq \left(\alpha - \frac{L_2}{2}\right)||\nabla u - \nabla v||^2_{L^2(\Omega)} + \left(\alpha_0 - \frac{L_2}{2}\right)||u - v||^2_{L^2(\Omega)}.\end{aligned}$$

Remark 2.1. Let $C_\Omega < \operatorname{diam}\Omega$ denote the optimal constant in the Friedrichs inequality for $H_0^1(\Omega)$ functions. Observe that (A1)−(A3) imply that the solution $u \in H_0^1(\Omega)$ of (2.1) fulfills

$$\begin{aligned}&\|F(u,\nabla u)\|_{L^2(\Omega)} \leq \|F(u,\nabla u) - F(0,\nabla u)\|_{L^2(\Omega)} + \|F(0,\nabla u) - F(0,0)\|_{L^2(\Omega)} \\ &\leq (L_1 C_\Omega + L_2)|u|_{H^1(\Omega)} \leq C_\Omega \frac{L_1 C_\Omega + L_2}{c_0}\|g\|_{L^2(\Omega)}.\end{aligned} \tag{2.3}$$

Note that problem (2.1) also covers equations such as

$$-\nabla\cdot(\kappa(u)A\nabla u) + F(u,\nabla u) = g,$$

for a strictly positive and sufficiently regular function κ (independent of x). In this case, the equation can be rewritten as

$$-\nabla\cdot A\nabla u + \tilde{F}(u,\nabla u) = \tilde{g}.$$

In the remainder of this paper, we use the notation $q_1 \lesssim q_2$ if $q_1 \leq Cq_2$ where $C > 0$ is a constant that only depends on the shape regularity of the mesh, but not on the mesh size. Dependencies such as $(L_1 + L_2)\alpha^{-1}$ are always explicitly stated whereas dependencies on the contrast $\frac{\beta}{\alpha}$ are allowed to be contained in the notation $\lesssim$ for the sake of simplicity.

3. Multiscale method

In this section we propose a local orthogonal decomposition (LOD) method that is based on the concept introduced by Hughes *et al.* [21,22] and the specific constructions proposed in [27,30] for linear problems. The required multiscale (MS) basis functions are obtained with the strategy established in [31].

The main idea of the Variational Multiscale Method is to start from a finite element space $\mathcal{V}_h$ with a highly resolved computational grid and to construct a splitting of this space into the direct sum $\mathcal{V}_h = \mathcal{V}^{\mathrm{l}} \oplus \mathcal{V}^{\mathrm{f}}$ of a low dimensional space $\mathcal{V}^{\mathrm{l}}$ and a "detail space" $\mathcal{V}^{\mathrm{f}}$ containing all the missing oscillations. Then, a basis of $\mathcal{V}^{\mathrm{l}}$ is assembled and we can compute a Galerkin approximation u_{l} of u in $\mathcal{V}^{\mathrm{l}}$. However, the success of this approach strongly depends on the choice of $\mathcal{V}^{\mathrm{l}}$. On the one hand, the costs for assembling a basis of $\mathcal{V}^{\mathrm{l}}$ must be kept low. On the other hand, the basis functions somehow need to contain information about fine scale features. For instance, a standard coarse finite element space is cheap to assemble but will fail to yield reliable approximations. On the contrary, the space spanned by high resolution finite element approximations yields perfect approximations, but is as costly as the original problem that we tried to avoid. Therefore, the key is to find an optimal balance between costs and accuracy. In previous works (*cf.* [21,27,28]) the multiscale basis (MS-basis) of $\mathcal{V}^{\mathrm{l}}$ was constructed involving the full multiscale operator B that corresponds with the left hand side of the original problem. In a fully linear setting, this can be a reasonable choice. However, it gets extremely expensive if B is a nonlinear operator, since it leads to numerous nonlinear equations to solve. Furthermore it is not clear if the constructed set of basis functions leads to good approximations. One novelty of this work is

that we do not involve the full operator B in the construction of the MS-basis, but only the linear diffusive part $\langle A\nabla\cdot, \nabla\cdot\rangle$. Even though the oscillations of F are not captured by the MS-basis, we can show that we are still able to obtain accurate approximations and to preserve the optimal convergence rates.

3.1. Notation and discretization

Let $\mathcal{T}_H$ denote a regular triangulation of Ω and let $H : \overline{\Omega} \to \mathbb{R}_{>0}$ denote the $\mathcal{T}_H$-piecewise constant mesh size function with $H|_T = H_T := \operatorname{diam}(T)$ for all $T \in \mathcal{T}_H$. Additionally, let $\mathcal{T}_h$ be a regular triangulation of Ω that is supposed to be a refinement of $\mathcal{T}_H$. We assume that $\mathcal{T}_h$ is sufficiently small so that all fine scale features of B are captured by the mesh. The mesh size h denotes the maximum diameter of an element of $\mathcal{T}_h$. The corresponding classical (conforming) finite element spaces of continuous piecewise polynomials of degree 1 are given by

$$V_H := \left\{v_H \in H_0^1(\Omega) \mid \forall T \in \mathcal{T}_H : \ (v_H)|_T \text{ is affine}\right\},$$
$$V_h := \left\{v_h \in H_0^1(\Omega) \mid \forall K \in \mathcal{T}_h : \ (v_h)|_K \text{ is affine}\right\}.$$

By J, we denote the dimension of V_H and by $\mathcal{N}_H = \{z_j | \ 1 \le j \le J\}$ the set of interior vertices of $\mathcal{T}_H$. For every vertex $z_j \in \mathcal{N}_H$, let $\lambda_j \in V_H$ denote the associated nodal basis function (tent function), *i.e.* $\lambda_j \in V_H$ with the property $\lambda_j(z_i) = \delta_{ij}$ for all $1 \le i, j \le J$.

From now on, we denote by $u_h \in V_h$ the classical finite element approximation of u in the discrete (highly resolved) space V_h, *i.e.* $u_h \in V_h$ solves

$$\int_\Omega A\nabla u_h \cdot \nabla v_h + F(\cdot, u_h, \nabla u_h)v_h = \int_\Omega g v_h \tag{3.1}$$

for all $v_h \in V_h$. We assume that V_h resolves the micro structure such that the error $\|u - u_h\|_{H^1(\Omega)}$ falls below a given tolerance. For standard finite element methods the error typically scales like $C \cdot h^s$ for some $s \ge \frac{1}{2}$. However, for regular coefficients, C depends on the derivative of A with respect to the spatial variable. If A oscillates rapidly, the derivatives become very large and h must be very small to compensate the dominance of C. This is only fulfilled, when h resolves the micro structure (we refer to [34,35] for some quantitative characterization of this so-called resolution condition). We are therefore dealing with pre-asymptotic effects for the standard methods. The multiscale method that we propose in the subsequent sections is designed to approximate u_h with an error proportional to the coarse mesh size H independent of fine scale oscillations of the data or the regularity of the solution, *i.e.*, we do not have such pre-asymptotic effects.

3.2. Quasi interpolation

The key tool in our construction is a linear (quasi-)interpolation operator $\mathfrak{I}_H : V_h \to V_H$ that is continuous and surjective. The kernel of this operator is going to be our fine space (or remainder space) V_h^{f}. In [31] a weighted Clément interpolation operator was used. In this work, we do not specify the choice. Instead, we state a set of assumptions that must be fulfilled in order to derive an optimal approximation result for the constructed multiscale method.

Assumption 2. (Assumptions on the quasi-interpolation operator)**.**

(A4) $\mathfrak{I}_H \in L(V_h, V_H)$, *i.e.* $\mathfrak{I}_H$ is linear,

(A5) the restriction of $\mathfrak{I}_H$ to V_H is an isomorphism with L^2-stable inverse $(\mathfrak{I}_H|_{V_H})^{-1}$, *i.e.* $\|(\mathfrak{I}_H|_{V_H})^{-1}(v_H)\|_{L^2(\Omega)} \le C_{\mathfrak{I}_H^{-1}}\|v_H\|_{L^2(\Omega)}$ for all $v_H \in V_H$ and with a generic constant $C_{\mathfrak{I}_H^{-1}}$ only depending on the shape regularity of $\mathcal{T}_H$ and $\mathcal{T}_h$.

(A6) there exists a generic constant $C_{\mathfrak{I}_H}$, only depending on the shape regularity of $\mathcal{T}_H$ and $\mathcal{T}_h$, such that for all $v_h \in V_h$ and for all $T \in \mathcal{T}_H$ there holds

$$H_T^{-1}\|v_h - \mathfrak{I}_H(v_h)\|_{L^2(T)} + \|\nabla(v_h - \mathfrak{I}_H(v_h))\|_{L^2(T)} \le C_{\mathfrak{I}_H}\|\nabla v_h\|_{L^2(\omega_T)}$$

with

$$\omega_T := \bigcup \{K \in \mathcal{T}_H | \, \overline{K} \cap \overline{T} \neq \emptyset\}.$$

(A7) there exists a generic constant $C'_{\mathfrak{I}_H}$, only depending on the shape regularity of $\mathcal{T}_H$ and $\mathcal{T}_h$, such that for all $v_H \in V_H$ there exists $v_h \in V_h$ with

$$\mathfrak{I}_H(v_h) = v_H, \quad |v_h|_{H^1(\Omega)} \leq C'_{\mathfrak{I}_H} |v_H|_{H^1(\Omega)} \quad \text{and} \quad \operatorname{supp} v_h \subset \operatorname{supp} v_H.$$

Observe that (A6) limits the growth of the support of an $v_H \in V_H$ when $\mathfrak{I}_H$ is applied to it, *i.e.* $\operatorname{supp}(I_H(v_H)) = \bigcup \{K \in \mathcal{T}_H | \, \overline{K} \cap \operatorname{supp}(v_H) \neq \emptyset\}$. We also note that the classical nodal interpolation operator does not fulfill assumption (A6) for $d > 1$ because the constant $C_{\mathfrak{I}_H}$ blows up for $h \to 0$. Numerical experiments confirm that such a choice leads in fact to instabilities in the later method. One possibility is to choose $\mathfrak{I}_H$ as a weighted Clément interpolation operator. This construction was proposed in [31]. Given $v \in H^1_0(\Omega)$, $\mathfrak{I}_H v := \sum_{j=1}^J v_j \lambda_j$ defines a (weighted) Clément interpolant with nodal values

$$v_j := \left(\int_\Omega v \lambda_j \, \mathrm{d}x\right) / \left(\int_\Omega \lambda_j \, \mathrm{d}x\right) \tag{3.2}$$

for $1 \leq j \leq J$ (*cf.* [11]) and zero in the boundary nodes. Furthermore, there exists the desired generic constant $C_{\mathfrak{I}_H}$ (only depending on the mesh regularity parameter and in particular independent of H_T) such that for all $v \in H^1_0(\Omega)$ and for all $T \in \mathcal{T}_H$ there holds

$$H_T^{-1} \|v - \mathfrak{I}_H v\|_{L^2(T)} + \|\nabla(v - \mathfrak{I}_H v)\|_{L^2(T)} \leq C_{\mathfrak{I}_H} \|\nabla v\|_{L^2(\omega_T)}.$$

We refer to [11] for a proof of this estimate. This gives us (A6). Assumption (A4) is obvious. The validity of (A5) and (A7) was proved in [31].

Note that in certain applications, additional features (*e.g.*, orthogonality properties) of the chosen interpolation operator may be exploited for improved error estimates (see, *e.g.*, [31] Rem. 3.2 and [10]).

3.3. Multiscale splitting and modified nodal basis

In this section, we construct a splitting of the high resolution finite element space V_h into a low dimension multiscale space V^{ms} and some high dimensional remainder space V_h^{f}. From now on, we let $\mathfrak{I}_H : V_h \to V_H$ denote an interpolation operator fulfilling the properties (A4)–(A7). Recall that $V_H \subset V_h$. We start with defining V_h^{f} as the kernel of $\mathfrak{I}_H$ in V_h:

$$V_h^{\mathrm{f}} := \{v_h \in V_h \mid \mathfrak{I}_H v_h = 0\}.$$

V_h^{f} represents the features in V_h not captured by V_H. Using assumption (A5) we get

$$V_h = V_H \oplus V_h^{\mathrm{f}}, \quad \text{where } \underbrace{v_h}_{\in V_h} = \underbrace{(\mathfrak{I}_H|_{V_H})^{-1}(\mathfrak{I}_H(v_h))}_{\in V_H} + \underbrace{v_h - (\mathfrak{I}_H|_{V_H})^{-1}(\mathfrak{I}_H(v_h))}_{\in V_h^{\mathrm{f}}}. \tag{3.3}$$

Here, the property $(\mathfrak{I}_H \circ (\mathfrak{I}_H|_{V_H})^{-1}))(v_H) = v_H$ for all $v_H \in V_H$ implies the equation $\mathfrak{I}_H(v_h - (\mathfrak{I}_H|_{V_H})^{-1}(\mathfrak{I}_H(v_h))) = \mathfrak{I}_H(v_h) - (\mathfrak{I}_H \circ (\mathfrak{I}_H|_{V_H})^{-1})(\mathfrak{I}_H(v_h)) = 0$. We still need to modify the splitting of V_h, because V_H is an inappropriate space for a multiscale approximation. We therefore look for the orthogonal complement of V_h^{f} in V_h with respect to the inner product $\langle A\nabla\cdot, \nabla\cdot\rangle_{L^2(\Omega)}$. For this purpose, we define the orthogonal projection $P^{\mathrm{f}} : V_h \to V_h^{\mathrm{f}}$ as follows. For a given $v_h \in V_h$, $P^{\mathrm{f}}(v_h) \in V_h^{\mathrm{f}}$ solves

$$\left\langle A\nabla P^{\mathrm{f}}(v_h), \nabla w^{\mathrm{f}} \right\rangle = \left\langle A\nabla v_h, \nabla w^{\mathrm{f}} \right\rangle \quad \text{for all } w^{\mathrm{f}} \in V_h^{\mathrm{f}}.$$

Defining the multiscale space $V_{H,h}^{\mathrm{ms}}$ by $V_{H,h}^{\mathrm{ms}} := (1 - P^{\mathrm{f}})(V_H)$, this directly leads to the orthogonal decomposition

$$V_h = V_{H,h}^{\mathrm{ms}} \oplus V_h^{\mathrm{f}}, \tag{3.4}$$

because

$$V_h = \operatorname{kern}(P^{\mathrm{f}}) \oplus V_h^{\mathrm{f}} = (1-P^{\mathrm{f}})(V_h) \oplus V_h^{\mathrm{f}} \overset{(3.3)}{=} (1-P^{\mathrm{f}})(V_H) \oplus V_h^{\mathrm{f}} = V_{H,h}^{\mathrm{ms}} \oplus V_h^{\mathrm{f}}.$$

Hence, any function $v_h \in V_h$ can be decomposed into $v_h = v_H^{\mathrm{ms}} + v^{\mathrm{f}}$ with $v_H^{\mathrm{ms}} = (\mathfrak{I}_H|_{V_H})^{-1}(\mathfrak{I}_H(v_h)) - P^{\mathrm{f}}((\mathfrak{I}_H|_{V_H})^{-1}(\mathfrak{I}_H(v_h)))$ and $v^{\mathrm{f}} = v_h - (\mathfrak{I}_H|_{V_H})^{-1}(\mathfrak{I}_H(v_h)) + P^{\mathrm{f}}((\mathfrak{I}_H|_{V_H})^{-1}(\mathfrak{I}_H(v_h)))$. Furthermore it holds $\langle A\nabla v_H^{\mathrm{ms}}, \nabla w^{\mathrm{f}}\rangle = 0$ for all $w^{\mathrm{f}} \in V_h^{\mathrm{f}}$. The space $V_{H,h}^{\mathrm{ms}}$ is a multiscale space of the same dimension as the coarse space V_H. However, note that it is only constructed on the basis of the oscillations of A. The oscillations of F are not taken into account. We will show that $V_{H,h}^{\mathrm{ms}}$ still yields the desired approximation properties.

We now introduce a basis of $V_{H,h}^{\mathrm{ms}}$. The image of the nodal basis function $\lambda_j \in V_H$ under the fine scale projection P^{f} is denoted by $\phi_j^h = P^{\mathrm{f}}(\lambda_j) \in V_h^{\mathrm{f}}$, *i.e.*, ϕ_j^h satisfies the corrector problem

$$\langle A\nabla\phi_j^h, \nabla w\rangle = \langle A\nabla\lambda_j, \nabla w\rangle \quad \text{for all } w \in V_h^{\mathrm{f}}. \tag{3.5}$$

A basis of $V_{H,h}^{\mathrm{ms}}$ is then given by the modified nodal basis

$$\left\{\lambda_j^{\mathrm{ms}} := \lambda_j - \phi_j^h \mid 1 \le j \le J\right\}. \tag{3.6}$$

As we can see, solving (3.5) involves a fine scale computation on the whole domain Ω. However, since the right hand side has small support, we are able to localize the computations. As we will see in the next section, the correctors show an exponential decay outside of the support of the coarse shape function λ_j.

First, we define a multiscale approximation that is based on the above orthogonal decomposition of V_h, but without localization.

Definition 3.1 (Multiscale approximation without localization)**.** The Galerkin approximation $u_{H,h}^{\mathrm{ms}} \in V_{H,h}^{\mathrm{ms}}$ of the exact solution u of problem (2.1) is defined as the solution of

$$\langle A\nabla u_{H,h}^{\mathrm{ms}}, \nabla v\rangle + \langle F\left(u_{H,h}^{\mathrm{ms}}, \nabla u_{H,h}^{\mathrm{ms}}\right), v\rangle = \langle g, v\rangle \quad \text{for all } v \in V_{H,h}^{\mathrm{ms}}. \tag{3.7}$$

3.4. Localization

So far, in order to construct a suitable multiscale space, we derived a set of linear fine scale problems (3.5) that can be solved in parallel. Still, as already mentioned in the previous section, these corrector problems are fine scale equations formulated on the whole domain Ω which makes them almost as expensive as the original problem. However, in [31] it was shown that the correction ϕ_j^h decays exponentially outside of the support of the coarse basis function λ_j. We specify this feature as follows. Let $k \in \mathbb{N}_{>0}$. We define nodal patches $\omega_{j,k}$ of k coarse grid layers centered around the node $z_j \in \mathcal{N}_H$ by

$$\begin{aligned} \omega_{j,1} &:= \operatorname{supp}\lambda_j = \cup\left\{T \in \mathcal{T}_H \mid z_j \in \overline{T}\right\}, \\ \omega_{j,k} &:= \cup\left\{T \in \mathcal{T}_H \mid \overline{T} \cap \overline{\omega}_{j,k-1} \neq \emptyset\right\} \quad \text{for } k \ge 2. \end{aligned} \tag{3.8}$$

These are the truncated computational domains for the corrector problems (3.5). The fast decay is summarized by the following lemma.

Lemma 3.2 (Decay of the local correctors [31])**.** *Let assumptions* (A1) *and* (A4)–(A7) *be fulfilled. Then, for all nodes $z_j \in \mathcal{N}_H$ and for all $k \in \mathbb{N}_{>0}$, the correctors ϕ_j^h satisfy the estimates*

$$\|A^{1/2}\nabla\phi_j^h\|_{L^2(\Omega\setminus\omega_{j,k})} \lesssim \mathrm{e}^{-rk}\|A^{1/2}\nabla\phi_j^h\|_{L^2(\Omega)}$$

with a generic rate r that is proportional to $(\alpha/\beta)^{1/2}$ but independent of variations of A. Recall the definition of '$\lesssim$' at the end of Section 2.

This fast decay motivates an approximation of ϕ_j^h on the truncated nodal patches $\omega_{j,k}$. We therefore define localized fine scale spaces by intersecting V_h^{f} with those functions that vanish outside the patch $\omega_{j,k}$, *i.e.*

$$V_h^{\mathrm{f}}(\omega_{j,k}) := \left\{ v \in V_h^{\mathrm{f}} \mid v_{|\Omega\setminus\omega_{j,k}} = 0 \right\}$$

for a given node $z_j \in \mathcal{N}_H$. The solutions $\phi_{j,k}^h \in V_h^{\mathrm{f}}(\omega_{j,k})$ of

$$\langle A\nabla\phi_{j,k}^h, \nabla w\rangle = \langle A\nabla\lambda_j, \nabla w\rangle \quad \text{for all } w \in V_h^{\mathrm{f}}(\omega_{j,k}), \tag{3.9}$$

are approximations of ϕ_j^h from (3.5) with local support and therefore cheap to solve. We define localized multiscale finite element spaces by

$$V_{H,h}^{\mathrm{ms},k} = \operatorname{span}\left\{ \lambda_{j,k}^{\mathrm{ms}} := \lambda_j - \phi_{j,k}^h \mid 1 \le j \le J \right\} \subset V_h. \tag{3.10}$$

We can now define a LOD approximation by localizing the corrector problems for the basis functions.

Definition 3.3 (LOD approximation)**.** The Galerkin approximation $u_{H,h}^{\mathrm{ms},k} \in V_{H,h}^{\mathrm{ms},k}$ of the exact solution u of problem (2.1) is defined as the solution of

$$\left\langle A\nabla u_{H,h}^{\mathrm{ms},k}, \nabla v\right\rangle + \left\langle F\left(u_{H,h}^{\mathrm{ms},k}, \nabla u_{H,h}^{\mathrm{ms},k}\right), v\right\rangle = \langle g, v\rangle \quad \text{for all } v \in V_{H,h}^{\mathrm{ms},k}. \tag{3.11}$$

Note, that changing the data functions F and g does not change the multiscale basis $\{\lambda_{j,k}^{\mathrm{ms}} \mid 1 \le j \le J\}$. Once $V_{H,h}^{\mathrm{ms},k}$ is computed, it can be reused for various combinations of F and g. This makes the new problems cheap to solve.

Remark 3.4. Observe that we never need to solve a problem on the scale of the oscillations of $F(\cdot,\xi,\zeta)$ in the case that they are faster than the oscillations of $A(\cdot)$. However, we implicitly assume that the arising integrals can be computed exactly (or with high accuracy). Practically this implies that a sufficiently high quadrature rule must be used. So even if the fine grid is not fine enough to resolve the variations of F, at least the quadrature rule must be fine enough to capture the correct averaged values. From Theorem 3.5 below we deduce that the influence of the oscillations of $F(\cdot,\xi,\zeta)$ remains small, as long as we have an accurate approximation of the averages on each coarse grid element. A similar observation holds for standard finite elements, where classical convergence rates can be expected as soon as the oscillations of A are resolved by the fine grid (independent of the oscillations of F).

3.5. *A priori* error estimate

We are now prepared to state the main result of this article, namely the optimal convergence of the method for the case that the local patches $\omega_{j,k}$ have a diameter of order $H|\log(H)|$.

Theorem 3.5. *Let $u \in H_0^1(\Omega)$ denote the exact solution given by problem (2.1), let $u_h \in V_h$ denote the corresponding finite element approximation in the Lagrange space with a highly resolved computational grid (*i.e. *the solution of (3.1)) and let $u_{H,h}^{\mathrm{ms},k} \in V_{H,h}^{\mathrm{ms},k}$ be the solution of our proposed multiscale method with localization* (i.e. *the solution of (3.11)). If assumptions* (A1)–(A7) *are satisfied and if $k \gtrsim |\log(\|H\|_{L^\infty(\Omega)})|$, then the a priori error estimate*

$$\left\|u - u_{H,h}^{\mathrm{ms},k}\right\|_{H^1(\Omega)} \le C\left(L_1, L_2, \alpha, \beta, c_0\right)\left(\|H\|_{L^\infty(\Omega)} + \|u - u_h\|_{H^1(\Omega)}\right).$$

holds with a generic constant C that does not depend on mesh sizes and oscillations of A and F. A suitable choice of the localization parameter k depends on the square root of the contrast, i.e. *the multiplicative constant hidden in $k \approx |\log(\|H\|_{L^\infty(\Omega)})|$ is proportional to $\sqrt{\frac{\beta}{\alpha}}$.*

A proof of Theorem 3.5 is presented in the subsequent section. In particular, the result is a conclusion from Theorem 4.3 which is stated in Section 4 below. In Theorem 4.3 we also give details on the generic constant C. We will see that it essentially depends on $\frac{(L_1+L_2)}{\alpha}$. Recall that L_1 and L_2 denote the Lipschitz constants of F (*cf.* (A2)) and that α is the smallest eigenvalue of A. This shows the significance of assuming that the problem is not dominated by the lower order term. For instance, consider the scenario of a pollutant being transported by groundwater flow. In this case, A describes the hydraulic conductivity which changes its properties on a scale of size ϵ. On the other hand, F describes the gravity driven flow that is scaled with the so called Péclet number. However, in the described scenario the Péclet number is of order ϵ^{-1} (*cf.* Bourlioux and Majda [7]) implying that $\mathrm{O}(L_1) = \epsilon^{-1}$. So the generic constant C is of order ϵ^{-1}. This means that we need $H < \epsilon$, *i.e.* we still need to resolve the micro structure with the coarse grid $\mathcal{T}_H$ producing the same costs as the original problem. If $H \gg \epsilon$ the estimate stated in Theorem 3.5 is of no value, because the right hand side remains large.

4. Error analysis

This section is devoted to the proof of Theorem 3.5. In particular, we state a detailed version of the result (see Thm. 4.3 below), where we specify the occurring constants. The proof is splitted into several lemmata. We start with an *a priori* error estimate for the multiscale approximation without localization.

Lemma 4.1. *Let $u_h \in V_h$ denote the highly resolved finite element approximation defined* via *equation* (3.1) *and let $u_{H,h}^{\mathrm{ms}} \in V_{H,h}^{\mathrm{ms}}$ denote the LOD approximation given by equation* (3.7)*. Under assumptions* (A1)–(A7)*, the* a priori *error estimate*

$$|u_h - u_H^{\mathrm{ms}}|_{H^1(\Omega)} \lesssim \tilde{C}_0 \left(\|Hg\|_{L^2(\Omega)} + \|H\|_{L^\infty(\Omega)} C_\Omega \frac{L_1 C_\Omega + L_2}{c_0} \|g\|_{L^2(\Omega)} \right)$$

holds with

$$\tilde{C}_0 := \left(\frac{\beta + \|H\|_{L^\infty(\Omega)}(L_1 C_\Omega + L_2)}{c_0 \cdot \alpha} \right).$$

Proof. Due to (3.4), we know that there exist $\tilde{u}_{H,h}^{\mathrm{ms}} \in V_{H,h}^{\mathrm{ms}}$ and $\tilde{u}_h^{\mathrm{f}} \in V_h^{\mathrm{f}}$, such that

$$u_h = \tilde{u}_{H,h}^{\mathrm{ms}} + \tilde{u}_h^{\mathrm{f}}.$$

We use the Galerkin orthogonality obtained from the equations (3.1) and (3.7) to conclude for all $v \in V_{H,h}^{\mathrm{ms}}$,

$$\langle A\nabla(u_h - u_{H,h}^{\mathrm{ms}}), \nabla v\rangle + \langle F(u_h, \nabla u_h), v\rangle - \langle F\left(u_{H,h}^{\mathrm{ms}}, \nabla u_{H,h}^{\mathrm{ms}}\right), v\rangle = 0. \tag{4.1}$$

In particular $v = u_{H,h}^{\mathrm{ms}} - \tilde{u}_{H,h}^{\mathrm{ms}} \in V_{H,h}^{\mathrm{ms}}$ is an admissible test function in (4.1). Together with $\mathfrak{I}_H(\tilde{u}_h^{\mathrm{f}}) = 0$, this yields

$$\begin{aligned}
c_0|u_h - u_{H,h}^{\mathrm{ms}}|_{H^1(\Omega)}^2 & \\
\overset{(2.2)}{\leq} & \langle A\nabla(u_h - u_{H,h}^{\mathrm{ms}}), \nabla(u_h - u_{H,h}^{\mathrm{ms}})\rangle \\
& + \langle F(u_h, \nabla u_h) - F(u_{H,h}^{\mathrm{ms}}, \nabla u_{H,h}^{\mathrm{ms}}), u_h - u_{H,h}^{\mathrm{ms}}\rangle \\
\overset{(4.1)}{=} & \langle A\nabla(u_h - u_{H,h}^{\mathrm{ms}}), \nabla(u_h - \tilde{u}_{H,h}^{\mathrm{ms}})\rangle \\
& + \langle F(u_h, \nabla u_h) - F(u_{H,h}^{\mathrm{ms}}, \nabla u_{H,h}^{\mathrm{ms}}), u_h - \tilde{u}_{H,h}^{\mathrm{ms}}\rangle \\
= & \langle A\nabla(u_h - u_{H,h}^{\mathrm{ms}}), \nabla\tilde{u}_h^{\mathrm{f}}\rangle + \left\langle F(u_h, \nabla u_h) - F\left(u_{H,h}^{\mathrm{ms}}, \nabla u_h\right), \tilde{u}_h^{\mathrm{f}} - \mathfrak{I}_H(\tilde{u}_h^{\mathrm{f}})\right\rangle \\
& + \left\langle F\left(u_{H,h}^{\mathrm{ms}}, \nabla u_h\right) - F\left(u_{H,h}^{\mathrm{ms}}, \nabla u_{H,h}^{\mathrm{ms}}\right), \tilde{u}_h^{\mathrm{f}} - \mathfrak{I}_H(\tilde{u}_h^{\mathrm{f}})\right\rangle \\
\lesssim & \beta|u_h - u_{H,h}^{\mathrm{ms}}|_{H^1(\Omega)}|\tilde{u}_h^{\mathrm{f}}|_{H^1(\Omega)} \\
& + \|H\|_{L^\infty(\Omega)}(L_1\|u_h - u_{H,h}^{\mathrm{ms}}\|_{L^2(\Omega)} + L_2|u_h - u_{H,h}^{\mathrm{ms}}|_{H^1(\Omega)})|\tilde{u}_h^{\mathrm{f}}|_{H^1(\Omega)} \\
\lesssim & (\beta + \|H\|_{L^\infty(\Omega)}(L_1 C_\Omega + L_2)) \cdot |u_h - u_{H,h}^{\mathrm{ms}}|_{H^1(\Omega)} \cdot |\tilde{u}_h^{\mathrm{f}}|_{H^1(\Omega)}.
\end{aligned}$$

With $\langle A\nabla\tilde{u}^{\mathrm{ms}}_{H,h}, \nabla\tilde{u}^{\mathrm{f}}_h\rangle = 0$ and with $\mathfrak{I}_H(v_{\mathrm{f}}) = 0$ for all $v_{\mathrm{f}} \in V^{\mathrm{f}}$ we get

$$\begin{aligned}
\alpha|\tilde{u}^{\mathrm{f}}_h|^2_{H^1(\Omega)} &\leq \langle A\nabla\tilde{u}^{\mathrm{f}}_h, \nabla\tilde{u}^{\mathrm{f}}_h\rangle \\
&= \left\langle A\nabla u_h, \nabla\tilde{u}^{\mathrm{f}}_h\right\rangle = \left\langle g, \tilde{u}^{\mathrm{f}}_h\right\rangle - \left\langle F(u_h, \nabla u_h), \tilde{u}^{\mathrm{f}}_h\right\rangle \\
&= \left\langle g, \tilde{u}^{\mathrm{f}}_h - \mathfrak{I}_H(\tilde{u}^{\mathrm{f}}_h)\right\rangle - \left\langle F(u_h, \nabla u_h), \tilde{u}^{\mathrm{f}}_h - \mathfrak{I}_H\left(\tilde{u}^{\mathrm{f}}_h\right)\right\rangle \\
&\overset{(2.3)}{\lesssim} \left(\|Hg\|_{L^2(\Omega)} + \|H\|_{L^\infty(\Omega)} C_\Omega \frac{L_1 C_\Omega + L_2}{c_0} \|g\|_{L^2(\Omega)}\right) \cdot |\tilde{u}^{\mathrm{f}}_h|_{H^1(\Omega)}.
\end{aligned}$$

The theorem follows by combing the results. □

The subsequent lemma is a consequence of the previous one.

Lemma 4.2. *Let $u_h \in V_h$ denote the fine scale approximation obtained from equation (3.1) and let $u^{\mathrm{ms},k}_{H,h} \in V^{\mathrm{ms},k}_{H,h}$ denote the solution of problem (3.11) (fully discrete LOD approximation). If the assumptions* (A1)−(A7) *hold true we obtain the estimate*

$$|u_h - u^{\mathrm{ms},k}_{H,h}|_{H^1(\Omega)} \lesssim \tilde{C}_2 \|g\|_{L^2(\Omega)} \|H\|_{L^\infty(\Omega)} + \tilde{C}_3 \min_{v^{\mathrm{ms},k}_{H,h} \in V^{\mathrm{ms},k}_{H,h}} \left\| A^{\frac{1}{2}} \nabla \left(u^{\mathrm{ms}}_{H,h} - v^{\mathrm{ms},k}_{H,h}\right)\right\|_{L^2(\Omega)},$$

where

$$\begin{aligned}
\tilde{C}_1 &:= (\beta + (L_1 C_\Omega + L_2) C_\Omega) \cdot \left(\frac{\beta + \|H\|_{L^\infty(\Omega)}(L_1 C_\Omega + L_2)}{c_0^2 \cdot \alpha}\right), \\
\tilde{C}_2 &:= \tilde{C}_1 + \tilde{C}_1 \cdot C_\Omega \frac{L_1 C_\Omega + L_2}{c_0}, \\
\tilde{C}_3 &:= \frac{\beta^{\frac{1}{2}} + \alpha^{-\frac{1}{2}}(L_1 C_\Omega + L_2) C_\Omega}{c_0}.
\end{aligned}$$

Proof. Let $v^{\mathrm{ms},k}_{H,h} \in V^{\mathrm{ms},k}_{H,h}$ denote an arbitrary element. Using the Galerkin orthogonality obtained from (3.1) and (3.11), we start in the same way as in the proof of Lemma 4.1 to get

$$\begin{aligned}
c_0|u_h - u^{\mathrm{ms},k}_{H,h}|^2_{H^1(\Omega)} &\\
&\overset{(2.2)}{\leq} \langle A\nabla(u_h - u^{\mathrm{ms},k}_{H,h}), \nabla(u_h - u^{\mathrm{ms},k}_{H,h})\rangle \\
&\qquad + \langle F(u_h, \nabla u_h) - F(u^{\mathrm{ms},k}_{H,h}, \nabla u^{\mathrm{ms},k}_{H,h}), u_h - u^{\mathrm{ms},k}_{H,h}\rangle \\
&\overset{(4.1)}{=} \langle A\nabla(u_h - u^{\mathrm{ms},k}_{H,h}), \nabla(u_h - u^{\mathrm{ms}}_{H,h}) + \nabla(u^{\mathrm{ms}}_{H,h} - v^{\mathrm{ms},k}_{H,h})\rangle \\
&\qquad + \langle F(u_h, \nabla u_h) - F(u^{\mathrm{ms},k}_{H,h}, \nabla u^{\mathrm{ms},k}_{H,h}), (u_h - u^{\mathrm{ms}}_{H,h}) + (u^{\mathrm{ms}}_{H,h} - v^{\mathrm{ms},k}_{H,h})\rangle \\
&\leq (\beta + (L_1 C_\Omega + L_2) C_\Omega)|u_h - u^{\mathrm{ms},k}_{H,h}|_{H^1(\Omega)} |u_h - u^{\mathrm{ms}}_{H,h}|_{H^1(\Omega)} \\
&\qquad + (\beta^{\frac{1}{2}} + \alpha^{-\frac{1}{2}}(L_1 C_\Omega + L_2) C_\Omega)|u_h - u^{\mathrm{ms},k}_{H,h}|_{H^1(\Omega)} \|A^{\frac{1}{2}} \nabla(u^{\mathrm{ms}}_{H,h} - v^{\mathrm{ms},k}_{H,h})\|_{L^2(\Omega)}.
\end{aligned}$$

Dividing by $|u_h - u^{\mathrm{ms},k}_{H,h}|_{H^1(\Omega)}$ and estimating $|u_h - u^{\mathrm{ms}}_{H,h}|_{H^1(\Omega)}$ with Lemma 4.1 yields the result. □

The combination of Lemmas 3.2 and 4.2 yields the main result of this paper.

Theorem 4.3. *Let $u_h \in V_h$ be solution of (3.1) and let $u^{\mathrm{ms},k}_{H,h} \in V^{\mathrm{ms},k}_{H,h}$ be the solution of (3.11). If the assumptions* (A1)−(A7) *hold true and if the number of layers k fulfills $k \gtrsim |\log(\|H\|_{L^\infty(\Omega)})|$, then it holds*

$$\left|u_h - u^{\mathrm{ms},k}_{H,h}\right|_{H^1(\Omega)} \lesssim \tilde{C}\|H\|_{L^\infty(\Omega)}\|g\|_{L^2(\Omega)},$$

where

$$\tilde{C} := \tilde{C}_2 + C_\Omega \frac{\beta}{c_0} \tilde{C}_3$$

and with $\tilde{C}_2$ *and* $\tilde{C}_3$ *as in Lemma 4.2.*

Proof. We define $w_{H,h}^{\mathrm{ms},k} \in V_{H,k}^{\mathrm{ms}}$ by

$$w_{H,h}^{\mathrm{ms},k} := \sum_{j=1}^{J} u_{H,h}^{\mathrm{ms}}(z_j)\lambda_{j,k}^{\mathrm{ms}} = \sum_{j=1}^{J} u_{H,h}^{\mathrm{ms}}(z_j)\left(\lambda_j - \phi_{j,k}^h\right)$$

where $u_{H,h}^{\mathrm{ms}}(z_j)$, $j = 1, 2, \ldots, J$, are the coefficients in the basis representation of $u_{H,h}^{\mathrm{ms}}$ from Definition 3.1. Hence,

$$\begin{aligned} &\min_{v_{H,h}^{\mathrm{ms},k} \in V_{H,h}^{\mathrm{ms},k}} \left\| A^{\frac{1}{2}} \nabla \left(u_{H,h}^{\mathrm{ms}} - v_{H,h}^{\mathrm{ms},k} \right) \right\|_{L^2(\Omega)}^2 \\ &\leq \left\| A^{\frac{1}{2}} \nabla \left(u_{H,h}^{\mathrm{ms}} - w_{H,h}^{\mathrm{ms},k} \right) \right\|_{L^2(\Omega)}^2 \\ &\lesssim \sum_{j=1}^{J} k^d u_{H,h}^{\mathrm{ms}}(z_j)^2 \| A^{1/2} \nabla \left(\phi_j^h - \phi_{j,k}^h \right) \|_{L^2(\Omega)}^2. \end{aligned} \tag{4.2}$$

For details on the last step, we refer to Lemma 4.9 in [31]. Due to the Galerkin orthogonality for the corrector problems it is possible to show

$$\| A^{1/2} \nabla (\phi_j^h - \phi_{j,k}^h) \|_{L^2(\Omega)}^2 \lesssim \| A^{1/2} \nabla \phi_j^h \|_{L^2(\Omega \setminus \omega_{j,k-1})}^2, \tag{4.3}$$

where the idea behind the proof of (4.3) is to use the best approximation property of $\phi_{j,k}^h$ in $V_h^{\mathrm{f}}(\omega_{j,k})$ to replace it by an arbitrary other function from $V_h^{\mathrm{f}}(\omega_{j,k})$. The best choice would be $\mathbb{1}_{\omega_{j,k}} \phi_j^h$, where $\mathbb{1}_{\omega_{j,k}}$ is the indicator function of $\omega_{j,k}$ (this choice would directly give the result). However, $\mathbb{1}_{\omega_{j,k}} \phi_j^h$ is not in $V_h^{\mathrm{f}}(\omega_{j,k})$, which is why additional interpolation and projection operators are required. The rather technical details for the proof of (4.3) are therefore given in the first part of the proof of Lemma 4.8 in [31].

The application of Lemma 3.2, (3.5), (4.3) and some inverse inequality yield

$$\begin{aligned} \| A^{1/2} \nabla (\phi_j^h - \phi_{j,k}^h) \|_{L^2(\Omega)}^2 &\lesssim \mathrm{e}^{-2rk} \| A^{1/2} \nabla \phi_j^h \|_{L^2(\Omega)}^2 \\ &\leq \mathrm{e}^{-2rk} \| A^{1/2} \nabla \lambda_j \|_{L^2(\Omega)}^2 \\ &\leq \beta \mathrm{e}^{-2rk} \| H \|_\infty^{-2} \| \lambda_j \|_{L^2(\Omega)}^2, \end{aligned}$$

with a generic rate r that is proportional to $(\beta/\alpha)^{1/2}$. By choosing $k = m \cdot |\log(\|H\|_{L^\infty(\Omega)})|$ with $m \in \mathbb{N}$, we can achieve an arbitrary fast polynomial convergence of this term in H (this will also cancel the k^d term). However, we bound this by a linear convergence since this is fastest rate that we can obtain for the whole error. Finally, the combination of this estimate and (4.2) plus

$$\begin{aligned} \sum_{j=1}^{J} u_{H,h}^{\mathrm{ms}}(z_j)^2 \| \lambda_j \|_{L^2(\Omega)}^2 &\lesssim \left\| \sum_{j=1}^{J} u_{H,h}^{\mathrm{ms}}(z_j) \lambda_j \right\|_{L^2(\Omega)}^2 \\ &= \left\| \sum_{j=1}^{J} u_{H,h}^{\mathrm{ms}}(z_j) ((\mathfrak{I}_H|_{V_H})^{-1} \circ \mathfrak{I}_H)(\lambda_j - \phi_j^h) \right\|_{L^2(\Omega)}^2 \\ &= \| (\mathfrak{I}_H|_{V_H})^{-1} \circ \mathfrak{I}_H) u_{H,h}^{\mathrm{ms}} \|_{L^2(\Omega)}^2 \overset{(\mathrm{A5})+(\mathrm{A6})}{\lesssim} \| \nabla u_{H,h}^{\mathrm{ms}} \|_{L^2(\Omega)}^2 \leq C_\Omega^2 c_0^{-2} \| g \|_{L^2(\Omega)}^2 \end{aligned}$$

yields the assertion. □

5. The multiscale newton scheme

In this section we discuss a solution algorithm for handling the nonlinear multiscale problem (3.11). For this purpose, we consider a damped Newton's method in the multiscale space $V_{H,h}^{\mathrm{ms},k}$. Recall that we are looking for $u \in H_0^1(\Omega)$ with

$$\langle B(u), v\rangle_{H^{-1},H_0^1} = \langle g, v\rangle \quad \text{for all } v \in H_0^1(\Omega),$$

where we introduced the notation

$$\langle B(v), w\rangle_{H^{-1},H_0^1} := \langle A\nabla v, \nabla w\rangle + \langle F(\cdot, v, \nabla v), w\rangle.$$

Here, $B : H_0^1(\Omega) \to H^{-1}(\Omega)$ is a hemicontinuous and strongly monotone operator due to assumption (A3). As already mentioned, under these assumptions, the Browder−Minty theorem yields a unique solution of the above problem. However, we will need an additional assumption on F to guarantee that the Newton scheme converges.

Assumption 3. Let $DF(x,\cdot,\cdot)$ denote the Jacobian matrix of $F(x,\cdot,\cdot)$.

(A8) We assume that there exists some constant $L_D \geq 0$ so that for almost every x in Ω and for all $(\xi_1,\zeta_1) \in \mathbb{R}\times\mathbb{R}^d$ and $(\xi_2,\zeta_2) \in \mathbb{R}\times\mathbb{R}^d$

$$|DF(x,\xi_1,\zeta_1) - DF(x,\xi_2,\zeta_2)| \leq L_D|(\xi_1,\zeta_1) - (\xi_2,\zeta_2)|,$$

i.e. $F(x,\cdot,\cdot) \in W^{2,\infty}(\mathbb{R}\times\mathbb{R}^d)$.

For clarity of the presentation we will leave out several indices within this section. In particular, we make use of the following notation.

Definition 5.1. For simplicity, we define

$$V^{\mathrm{ms}} := V_{H,h}^{\mathrm{ms},k} \quad \text{with basis } \lambda_j^{\mathrm{ms}} := \lambda_{j,k}^{\mathrm{ms}} = \lambda_j - \phi_{j,k}^h \text{ for } 1 \leq j \leq J.$$

Furthermore, we denote $u^{\mathrm{ms}} := u_{H,h}^{\mathrm{ms},k}$. Additionally, let

$$\partial_1 F(x,\xi,\zeta) := \partial_\xi F(x,\xi,\zeta) \quad \text{and} \quad \partial_2 F(x,\xi,\zeta) := \partial_\zeta F(x,\xi,\zeta).$$

We now describe the Newton strategy in detail. The fully discrete multiscale problem is to

$$\text{find } u^{\mathrm{ms}} \in V^{\mathrm{ms}}: \quad \langle A\nabla u^{\mathrm{ms}}, \nabla\lambda_j^{\mathrm{ms}}\rangle + \langle F(\cdot, u^{\mathrm{ms}}, \nabla u^{\mathrm{ms}}), \lambda_j^{\mathrm{ms}}\rangle - \langle g, \lambda_j^{\mathrm{ms}}\rangle = 0$$

for all $1 \leq j \leq J$. Again, using Browder−Minty, u^{ms} exists and is unique. Accordingly, we get the following well posed algebraic version of the problem:

$$\text{find } \bar{\alpha} \in \mathbb{R}^J: \quad G(\bar{\alpha}) = 0$$

and where $G : \mathbb{R}^J \to \mathbb{R}^J$ is given by

$$(G(\alpha))_l := \sum_{j=1}^J \alpha_j \langle A\nabla\lambda_j^{\mathrm{ms}}, \nabla\lambda_l^{\mathrm{ms}}\rangle + \langle F\left(\cdot, \sum_{j=1}^J \alpha_j\lambda_j^{\mathrm{ms}}, \sum_{j=1}^J \alpha_j\nabla\lambda_j^{\mathrm{ms}}\right), \lambda_l^{\mathrm{ms}}\rangle - \langle g, \lambda_l^{\mathrm{ms}}\rangle. \tag{5.1}$$

We have the relation $u^{\mathrm{ms}} = \sum_{j=1}^J \bar{\alpha}_j\lambda_j^{\mathrm{ms}}$. Before we can apply the Newton method to (5.1), we need to ensure that the iterations of the scheme are well defined.

Lemma 5.2. *Let $(X, \|\cdot\|_X)$ denote a Hilbert space with dual space X'. Let furthermore $B : X \to X'$ be a hemicontinuous, Fréchet differentiable and strongly monotone operator on X,* i.e. *there exists $c_0 > 0$ so that*

$$\begin{aligned}&\langle B(v) - B(w), v - w\rangle_X \geq c_0\|v - w\|_X^2 \quad \text{for all } v, w \in X \text{ and}\\ &\quad s \mapsto \langle B(u + sv), w\rangle_X\end{aligned}$$

is a continuous function on $[0, 1]$ for all $u, v, w \in X$. Let X_N denote a finite dimensional subspace with basis $\{\psi_1, \dots, \psi_N\}$ and let $b : \mathbb{R}^N \to V_N$ define the linear bijection with $b(\alpha) := \sum_{i=1}^N \alpha_i\psi_i$. If $G(\alpha) := b^{-1}(B(b(\alpha)))$, then the Jacobi matrix $DG(\alpha) \in \mathbb{R}^{n\times n}$ has only positive eigenvalues.

Proof. Let B' denote the Fréchet derivative of B, given by

$$B'(u)(v) = \lim_{s\to 0} \frac{B(u + sv) - B(u)}{s} \quad \text{for } u, v \in X.$$

This and the strong monotonicity yield

$$\begin{aligned}\langle B'(u)(v), v\rangle_{H^{-1},H_0^1} &= \lim_{s\to 0} \frac{(B(u + sv) - B(u))(v)}{s}\\ &= \lim_{s\to 0} \frac{1}{s^2}(B(u + sv) - B(u))(u + sv - u)\\ &\geq \lim_{s\to 0} \frac{1}{s^2} c_0\|sv\|^2 = c_0\|v\|^2.\end{aligned} \tag{5.2}$$

Next, observe that b induces an inner product on $\mathbb{R}^N$ by $(\alpha_1, \alpha_2)_b := \langle b(\alpha_1), b(\alpha_2)\rangle_X$. Let $\alpha := b^{-1}(u)$ then we get

$$\begin{aligned}B'(u)(\psi_i) &= \lim_{s\to 0} \frac{B(u + s\psi_i) - B(u)}{s}\\ &= \lim_{s\to 0} \frac{(b \circ b^{-1})\left(B\left(\sum_{j=1}^N (\alpha_j + s\delta_{ij})\,\psi_j\right) - (b \circ b^{-1})\left(B\left(\sum_{j=1}^N \alpha_j\psi_j\right)\right)\right.}{s}\\ &= b\left(\lim_{s\to 0} \frac{G(\alpha + se_i) - G(\alpha)}{s}\right)\\ &= b(D_\alpha G(\alpha)e_i).\end{aligned}$$

Using this, we get for arbitrary $\xi \in \mathbb{R}^N$ and $v_\xi := b(\xi)$,

$$\begin{aligned}(D_\alpha G(\alpha)\xi, \xi)_b &= \sum_{i,j}^N \xi_i\xi_j(D_\alpha G(\alpha)e_i, e_j)_b\\ &= \sum_{i,j}^N \xi_i\xi_j(b(D_\alpha G(\alpha)e_i), b(e_j))_X\\ &= \sum_{i,j}^N \xi_i\xi_j(B'(u)(\psi_i), \psi_j)_X\\ &= (B'(u)(v_\xi), v_\xi)_X \overset{(5.2)}{\geq} c_0\|v_\xi\|_X^2 = c_0\|\xi\|_b^2.\end{aligned}$$

Since all norms in $\mathbb{R}^N$ are equivalent we have the desired result. □

Now, we can apply the Newton method for solving the nonlinear algebraic equation $G(\bar{\alpha}) = 0$. If $D_\alpha G$ denotes the Jacobian matrix of G, we get the following iteration scheme:

$$\alpha^{(n+1)} := \alpha^{(n)} + \triangle\alpha^{(n)},$$

where $\triangle\alpha^{(n)}$ solves

$$D_\alpha G\left(\alpha^{(n)}\right)\triangle\alpha^{(n)} = -G\left(\alpha^{(n)}\right). \tag{5.3}$$

Here, $D_\alpha G$ is given by

$$D_{\alpha_i}\left(G(\alpha)\right)_l := \langle A\nabla\lambda_i^{\mathrm{ms}}, \nabla\lambda_l^{\mathrm{ms}}\rangle + \left\langle \partial_1 F\left(\cdot, \sum_{j=1}^{J}\alpha_j\lambda_j^{\mathrm{ms}}, \sum_{j=1}^{J}\alpha_j\nabla\lambda_j^{\mathrm{ms}}\right)\lambda_i^{\mathrm{ms}}, \lambda_l^{\mathrm{ms}}\right\rangle$$
$$+ \left\langle \partial_2 F\left(\cdot, \sum_{j=1}^{J}\alpha_j\lambda_j^{\mathrm{ms}}, \sum_{j=1}^{J}\alpha_j\nabla\lambda_j^{\mathrm{ms}}\right)\cdot\nabla\lambda_i^{\mathrm{ms}}, \lambda_l^{\mathrm{ms}}\right\rangle.$$

Lemma 5.2 ensures that equation (5.3) has a unique solution $\triangle\alpha^{(n)}$, *i.e.* that the Newton iteration is well posed. Since $G \in C^1(\mathbb{R}^N)$ has a nonsingular Jacobian matrix $D_\alpha G$ (due to Lem. 5.2) and since we have Lipschitz-continuity of $D_\alpha G$ (due to Assumption 3), we have that the Newton scheme converges quadratically as long as the starting value is close enough to the exact solution (*cf.* [12]). However, this means that we can only guarantee local convergence of the method. In order to ensure global convergence, we can use a simple damping strategy due to Armijo [2]. Here we are looking for a damping parameter $\zeta \in (0, 1]$ so that $\alpha^{(n+1)} := \alpha^{(n)} + \zeta\triangle\alpha^{(n)}$ with the property $|G(\alpha^{(n+1)})| < (1 - \frac{\zeta}{2})|G(\alpha^{(n)})|$. In our case, the convergence of the damped Newton scheme can be guaranteed by the following lemma which is based on the results by Kelley [26].

Lemma 5.3. *Let assumptions* (A1)−(A3) *and* (A8) *be fulfilled, then the damped Newton scheme converges,* i.e. *there exists a nonempty (damping) interval* $[\zeta_0, \zeta_1] \subset (0, 1)$*, so that*

$$\left|G\left(\alpha^{(n+1)}\right)\right| < \left(1 - \frac{\zeta}{2}\right)\left|G\left(\alpha^{(n)}\right)\right| \qquad \textit{for all } \zeta \in [\zeta_0, \zeta_1].$$

Here, $\zeta_0 > 0$ *is independent of* $\alpha^{(n)}$ *and* $\triangle\alpha^{(n)}$*, which prevents* $\zeta_1 \to 0$.

Proof. The existence of a damping parameter so that $|G(\alpha^{(n+1)})| < |G(\alpha^{(n)})|$ is an easy observation if we look at the function $h(\zeta) := |G(\alpha^{(n)} + \zeta\triangle\alpha^{(n)})|^2$ which fulfills $h(0) > 0$ and $h'(0) = -2G(\alpha^{(n)}) \cdot G(\alpha^{(n)}) < 0$. The existence of a uniform lower bound $\zeta_0 > 0$ was proved by Kelley ([26], Lem. 8.2.1 and Thm. 8.2.1 therewithin). The results by Kelley require Lipschitz continuity of $D_\alpha G$ (guaranteed by Assump. (A8)) and uniform boundedness of $|(D_\alpha G(\alpha))^{-1}|$. The latter one is fulfilled since the proof of Lemma 5.2 shows that the smallest eigenvalue of $(D_\alpha G(\alpha))$ is equal or larger than c_0. This implies that the largest eigenvalue of $(D_\alpha G(\alpha))^{-1}$ is bounded by c_0^{-1}, hence $|(D_\alpha G(\alpha))^{-1}|$ is uniformly bounded. □

In summary, Lemma 5.3 guarantees globally linear convergence of the method (using damping) and locally (*i.e.* in an environment of the solution) even quadratic convergence using the classical Newton scheme without damping. With these considerations, we can state the full algorithm below. Recall that $\mathcal{N}_H$ denotes the set of interior vertices of $\mathcal{T}_H$ and for $z_j \in \mathcal{N}_H$, $\lambda_j \in V_H$ denotes the corresponding nodal basis function.

Note that in the presented algorithm, each iteration starts with the damping parameter $\zeta_n = 1$ and we do not use damping parameters from previous iterations. The advantage is that we automatically get quadratic convergence of the Newton scheme as soon as we leave the region where damping is required. Therefore, damping is only used when really necessary.

Algorithm: dampedNewtonLOD($abstol$, $reltol$, $\alpha^{(0)}$, k)

In parallel **foreach** $z_j \in \mathcal{N}_H$ **do**
compute $\phi^h_{j,k} \in V_h^{\mathrm{f}}(\omega_{j,k})$ with

$$\langle A\nabla\phi^h_{j,k}, \nabla w\rangle = \langle A\nabla\lambda_j, \nabla w\rangle \quad \text{for all } w \in V_h^{\mathrm{f}}(\omega_{j,k}).$$

end
Set $V^{\mathrm{ms},k}_{H,h} := \operatorname{span}\{\lambda_j - \phi^h_{j,k} \mid 1 \le j \le J\}$. Set $\lambda^{\mathrm{ms}}_{j,k} = \lambda_j - \phi^h_{j,k}$.

Set $\alpha^{(n)} := \alpha^{(0)}$. Set $u^{\mathrm{ms},k,(n)}_{H,h} := \sum_{j=1}^J \alpha_j^{(n)} \lambda^{\mathrm{ms}}_{j,k}$. Set

$$(G(\alpha))_i := \sum_{j=1}^J \alpha_j \langle A\nabla\lambda^{\mathrm{ms}}_{j,k}, \nabla\lambda^{\mathrm{ms}}_{i,k}\rangle + \langle F(\cdot, \sum_{j=1}^J \alpha_j\lambda^{\mathrm{ms}}_{j,k}, \sum_{j=1}^J \alpha_j \nabla\lambda^{\mathrm{ms}}_{j,k}) - g, \lambda^{\mathrm{ms}}_{i,k}\rangle.$$

Set $tol := |G(\alpha^{(0)})|_2 \cdot reltol + abstol$.

while $|G(\alpha^{(n)})|_2 > tol$ **do**
Set $u^{\mathrm{ms},k,(n)}_{H,h} := \sum_{j=1}^J \alpha_j^{(n)} \lambda^{\mathrm{ms}}_{j,k}$.
Define the entries of the stiffness matrix $M^{(n)}$ by

$$\begin{aligned} M^{(n)}_{il} &:= \langle A\nabla\lambda^{\mathrm{ms}}_{l,k}, \nabla\lambda^{\mathrm{ms}}_{i,k}\rangle + \langle \partial_1 F(\cdot, u^{\mathrm{ms},k,(n)}_{H,h}, \nabla u^{\mathrm{ms},k,(n)}_{H,h})\lambda^{\mathrm{ms}}_{l,k}, \lambda^{\mathrm{ms}}_{i,k}\rangle \\ &\quad + \langle \partial_2 F(\cdot, u^{\mathrm{ms},k,(n)}_{H,h}, \nabla u^{\mathrm{ms},k,(n)}_{H,h}) \cdot \nabla\lambda^{\mathrm{ms}}_{l,k}, \lambda^{\mathrm{ms}}_{i,k}\rangle. \end{aligned}$$

Define the entries of the right hand side by

$$F^{(n)}_i := \langle g, \lambda^{\mathrm{ms}}_{i,k}\rangle - \langle A\nabla u^{\mathrm{ms},k,(n)}_{H,h}, \nabla\lambda^{\mathrm{ms}}_{i,k}\rangle - \langle F(\cdot, u^{\mathrm{ms},k,(n)}_{H,h}, \nabla u^{\mathrm{ms},k,(n)}_{H,h}), \lambda^{\mathrm{ms}}_{i,k}\rangle.$$

Find $(\triangle\alpha)^{(n+1)} \in \mathbb{R}^J$, with

$$M^{(n)}(\triangle\alpha)^{(n+1)} = F^{(n)}.$$

Set $\zeta_n := 1$. Set $\alpha^{(n+1)} := \alpha^{(n)} + \zeta_n \triangle\alpha^{(n)}$.
while $|G(\alpha^{(n+1)})| \ge (1 - \frac{\zeta_n}{2})|G(\alpha^{(n)})|$ **do**
Set $\zeta_n := \frac{1}{2}\zeta_n$. Set $\alpha^{(n+1)} := \alpha^{(n)} + \zeta_n \triangle\alpha^{(n)}$.
end
Set $\alpha^{(n)} := \alpha^{(n+1)}$. Set $tol := |G(\alpha^{(n)})|_2 \cdot reltol + abstol$.
end

Set $u^{\mathrm{ms},k,(n)}_{H,h} := \sum_{j=1}^J \alpha_j^{(n)} \lambda^{\mathrm{ms}}_{j,k}$.

Proposition 5.4. *We use the notation stated in Definition 5.1. Let $u \in H^1_0(\Omega)$ denote the solution of (2.1), let $u_h \in V_h$ denote the solution of (3.1) and let $u^{\mathrm{ms}} \in V^{\mathrm{ms}}$ denote the solution of (3.11). Furthermore, we let $u^{\mathrm{ms},(n)} := u^{\mathrm{ms},k,(n)}_{H,h}$ define the n'th iterate from the damped Newton LOD Method stated in the algorithm. Under assumptions (A1)–(A8), the Newton step (5.3) is well posed, yields an unique solution and $u^{\mathrm{ms},(n)}$ converges at least linearly to u^{ms}. If furthermore $k \gtrsim |\log(\|H\|_{L^\infty(\Omega)})|$, the* a priori *error estimate*

$$\|u - u^{\mathrm{ms}}\|_{H^1(\Omega)} \le C\left(\|H\|_{L^\infty(\Omega)} + \|u - u_h\|_{H^1(\Omega)}\right)$$

holds with a generic constant $C = O(1)$ (see Thms. 3.5 and 4.3 for details) and

$$\left\|u^{\mathrm{ms}} - u^{\mathrm{ms},(n)}\right\|_{H^1(\Omega)} \le L_n(H)\left\|u^{\mathrm{ms}} - u^{\mathrm{ms},(n-1)}\right\|_{H^1(\Omega)}.$$

Here, we have $L_n(H) < 1$.

If $u^{\mathrm{ms},(n-1)}$ *is sufficiently close to* u^{ms}*, we even get quadratic convergence of the Newton scheme,* i.e.*,*

$$\left\|u^{\mathrm{ms}} - u^{\mathrm{ms},(n)}\right\|_{H^1(\Omega)} \leq L_n(H) \left\|u^{\mathrm{ms}} - u^{\mathrm{ms},(n-1)}\right\|^2_{H^1(\Omega)}.$$

with

$$L_n(H) \leq \frac{\|(D_\alpha G)^{-1}\|_{L^\infty(\mathbb{R}^N)}}{L},$$

where L *denotes the Lipschitz-constant of* $D_\alpha G$*. As indicated,* $L_n(H)$ *typically depends on the mesh size. However, in some cases of semi-linear problems, it is possible to bound* $L_n(H)$ *independent of the triangulation (cf.* [25]*). In particular, if* $F(x,u,\nabla u) = F(x,u)$ *(*i.e. *no dependency on* ∇u*) we get that* $L_n(H) = L_n$ *independent of the underlying mesh. The proof can be obtained analogously to the proof of Proposition 4.1 in* [25]*. The proof fails for general* $F(x,u,\nabla u)$*.*

Remark 5.5. Note that the proposed method only requires the computation of the multiscale basis $\{\lambda_j^{\mathrm{ms}} | \, 1 \leq j \leq J\}$ once at the beginning. For each iteration step of the damped Newton scheme, (5.3) is a low dimensional linear problem that can reuse the initially computed multiscale basis. If the multiscale basis was computed using the nonlinear term F, local corrector problems would have to be solved for each Newton step newly, making the whole procedure significantly more expensive. We also note that assemblation of the tangent matrix $M^{(n)}$ and the residual $F^{(n)}$ still requires a quadrature rule that captures the fine scale features. Depending on the type of the nonlinearity this might have to be done newly for each iteration step, making the quadrature rule a significant part of each Newton step.

6. Numerical experiment

As mentioned in the introduction, Richards-type equations can be an application of our LOD-Newton framework. In general, the stationary Richards equation cannot necessarily be described by a monotone operator, however depending on the chosen model and the considered hydrological effects (including hysteresis, root uptake, friction, reaction fronts, *etc.*) monotone operators can arise in certain applications. One explicit example is the (regularized) time-discretized Kirchhoff transformed Richards equation regarded in [6]. For the case that there is no Signorini boundary condition prescribed, the problem that has to be solved for each time step corresponds to a nonlinear elliptic monotone problem (on the full space) that also fulfills the required assumption of Lipschitz-continuity.

Let us now consider the stationary Kirchhoff-transformed Richards equation

$$\nabla \cdot (K \nabla u) - \nabla \cdot (K \, kr(M(u)) \vec{g}) = f, \tag{6.1}$$

where u denotes the generalized pressure, K the hydraulic conductivity and kr the relative permeability depending on the saturation. kr is a monotone increasing function with values between 0 and 1 (typically bounded away from 0 to avoid degeneracy). If we have already full saturation, water cannot be conducted anymore, if the soil is completely dry (saturation is zero), water can be perfectly conducted. Formulas for kr were *e.g.* provided by Burdine [9] and Mualem [32]. In applications the variations of the hydraulic conductivity K are assumed to be constant (or at least slow) in gravity direction $\vec{g} = (0,0,\vec{g}_z)$, where $\vec{g}_z$ denotes the gravity factor of $9.81 m/s^2$. Soil probes are often only taken once in vertical direction, but a lot of samples are required to describe the variations of conductivity in horizontal direction. As a reduction of complexity one can often assume that $\nabla \cdot (K\vec{g}) = \partial_z (K_{zz} \vec{g}_z) = 0$ to consider the reduced equation

$$\nabla \cdot (K \nabla u) - \, (kr \circ M)'(u) \, (K\vec{g}) \cdot \nabla u = f. \tag{6.2}$$

Here we have $M(u) := \theta \circ \kappa^{-1}$, where θ denotes the saturation (depending on the pressure) and κ^{-1} the inverse of the Kirchhoff transformation $\kappa(p) := \int_0^p kr(\theta(q)) \, dq$. The saturation θ can be obtained by the capillary pressure

TABLE 1. Results for fine grid with $\epsilon > h = 2^{-6} \approx 0.016 > \epsilon^{\frac{3}{2}}$ which resolves the oscillations of the linear term, but not the oscillations of the nonlinear term. The truncation parameter k determines the patch size by (3.8). We observe an average EOC of 2.37 for the L^2-error and an average EOC of 1.33 for the H^1-error.

H	k	$\left\|u_{H,h}^{\mathrm{ms},k} - u_h\right\|_{L^2(\Omega)}$	$\left\|u_{H,h}^{\mathrm{ms},k} - u_h\right\|_{H^1(\Omega)}$
2^{-2}	1	0.1455	1.6985
2^{-3}	2	0.0097	0.3737
2^{-4}	3	0.0023	0.1772
2^{-5}	3	0.0008	0.1067

relation (soil-water retention curves). Various explicit formulas for θ are available, see *e.g.* Van Genuchten [24], Brooks-Corey [8] or the Gardner model [23]. Depending on the chosen model $(kr \circ M)'$ might not be a Lipschitz continuous function, still regularization is possible. In the following numerical experiment, we consider a test problem that has the structure derived from a regularized Burdine–Brooks–Corey model. The corresponding explicit formulas for $(kr \circ M)$ are taken from [3]. Contrary to the model (6.2), we use a nonlinear advection term that is faster oscillating than the diffusion term. The reason is that we want to emphasize our claim, that the oscillations of the nonlinearity F do in fact not influence the convergence. Before stating the test problem related to (6.2), let us note that the method and the analytical results of this paper directly transfer to equations in divergence form like (6.1), *i.e.* the gradient in the weak formulation can be on the test function, as long as $F(x,u)$ does not dependent on the gradient ∇u.

We consider the following nonlinear advection-diffusion problem. Let $\Omega :=]0,1[^2$ and $\epsilon := 0.05$. Find u^ϵ with

$$\begin{aligned} -\nabla \cdot (A^\epsilon(x)\nabla u^\epsilon(x)) + \frac{1}{2}F^\epsilon(x,u^\epsilon)\partial_{x_2}u^\epsilon(x) &= -\frac{3}{10} \quad \text{in } \Omega \\ u^\epsilon(x) &= 0 \qquad \text{on } \partial\Omega, \end{aligned}$$

where A^ϵ is given by

$$A^\epsilon(x_1,x_2) := \frac{1}{8\pi^2}\begin{pmatrix} 2\left(2+\cos\left(2\pi\frac{x_1}{\epsilon}\right)\right)^{-1} & 0 \\ 0 & 1+\frac{1}{2}\cos\left(2\pi\frac{x_1}{\epsilon}\right) \end{pmatrix}$$

and

$$F^\epsilon(x,u) := \frac{1}{8\pi^2}\left(2+\cos\left(2\pi\frac{x_1}{\epsilon^{\frac{3}{2}}}\right)\right)\begin{cases} \sqrt{\frac{u}{2}+\frac{3}{2}} & \text{for } -3 \le u \le -\frac{5}{4} \\ p(u) & \text{for } -\frac{5}{4} \le u \le -1 \\ 0 & \text{for } u \ge -1 \end{cases},$$

where $p(u) = au^3 + bu^2 + cu + d$ is such that $F^\epsilon(x,\cdot) \in C^1(-3,\infty)$ for all $x \in \Omega$. The (unknown) exact solution of this problem takes values between 0 and -1.75.

The numerical experiments presented in this section were performed with a little different implementation of the localization strategy than the one described in Section 3.4. We used the localized basis functions proposed in [19], which have the completely same analytical properties than (3.9)–(3.10), with the only difference that they are computed with respect to unit vectors instead of gradients of basis functions in order to slightly stabilize the computations.

The tolerance *tol* in the Newton algorithm is set to 10^{-10}. We keep the resolution of the (uniformly refined) fine grid fixed with $h = 2^{-6} < \epsilon$. The computations were made for four different coarse grid resolutions $H = 2^{-2}, \ldots, 2^{-5}$.

For given H, we guess the truncation parameter k (according to (3.8)) by $|\log(H)|$. By log we mean the logarithm to the basis e. For $H = 2^{-l}$, $l = 2, \ldots, 5$ we obtain $\log(4) \approx 1.386$, $\log(8) \approx 2.08$, $\log(16) \approx 2.77$ and $\log(32) \approx 3.47$. Optimistically rounding we set the truncation parameter k to 1 for $H = 2^{-2}$, 2 for $H = 2^{-3}$, 3 for $H = 2^{-4}$ and 3 for $H = 2^{-5}$. The corresponding results are depicted in Table 1. We observe that the proportionality coefficient in the choice of the diameter of the patches $\mathrm{O}(\mathrm{diam}(\omega_{j,k})) \sim H|\log(\|H\|_{L^\infty(\Omega)})|$ can chosen to be on 1 without suffering from pre-asymptotic effects. In fact, we obtain an experimental order of convergence (EOC) of 2.37 for the L^2-error and an EOC of 1.33 for the H^1-error. The patches remain small and computational demand for solving the local problems remains very small. For further numerical studies of the method and the choice of patch sizes in the linear case, we refer to [31].

Acknowledgements. We would like to thank the anonymous reviewers for their valuable suggestions and their constructive criticism that helped us to improve the paper.

References

[1] H.W. Alt and S. Luckhaus, Quasilinear elliptic-parabolic differential equations. *Math. Z.* **183** (1983) 311–341.

[2] L. Armijo, Minimization of functions having Lipschitz continuous first partial derivatives. *Pacific J. Math.* **16** (1966) 1–3.

[3] H. Berninger, Domain Decomposition Methods for Elliptic Problems with Jumping Nonlinearities and Application to the Richards Equation. *Ph.D. thesis*. Freie Universität Berlin (2007).

[4] H. Berninger, Non-overlapping domain decomposition for the Richards equation *via* superposition operators. Vol. 70 of *Lect. Notes Comput. Sci. Eng.* Springer, Berlin (2009) 169–176.

[5] H. Berninger, R. Kornhuber and O. Sander, On nonlinear Dirichlet-Neumann algorithms for jumping nonlinearities. Domain decomposition methods in science and engineering XVI. Vol. 55 of *Lect. Notes Comput. Sci. Eng.* Springer, Berlin (2007) 489–496.

[6] H. Berninger, R. Kornhuber and O. Sander, Fast and robust numerical solution of the Richards equation in homogeneous soil. *SIAM J. Numer. Anal.* **49** (2011) 2576–2597.

[7] A. Bourlioux and A.J. Majda, An elementary model for the validation of flamelet approximations in non-premixed turbulent combustion. *Combust. Theory Model.* **4** (2000) 189–210.

[8] R.H. Brooks and A.T. Corey, Hydraulic properties of porous media. *Hydrol. Pap.* 4, Colo. State Univ., Fort Collins (1964).

[9] N.T. Burdine, Relative permeability calculations from pore-size distribution data. *Petr. Trans. Am. Inst. Mining Metall. Eng.* **198** (1953) 71–77.

[10] C. Carstensen, Quasi-interpolation and *a posteriori* error analysis in finite element methods. *ESAIM: M2AN* **33** (1999) 1187–1202.

[11] C. Carstensen and R. Verfürth, Edge residuals dominate *a posteriori* error estimates for low order finite element methods. *SIAM J. Numer. Anal.* **36** (1999) 1571–1587.

[12] J.E. Dennis Jr. and R.B. Schnabel, Numerical Methods for Unconstrained Optimization and Nonlinear Equations. *SIAM Classics Appl. Math.* (1996).

[13] W. E and B. Engquist, The heterogeneous multiscale methods. *Commun. Math. Sci.* **1** (2003) 87–132.

[14] A. Gloria, An analytical framework for the numerical homogenization of monotone elliptic operators and quasiconvex energies. *SIAM Multiscale Model. Simul.* **5** (2006) 996–1043.

[15] P. Henning, Convergence of MsFEM approximations for elliptic, non-periodic homogenization problems. *Netw. Heterog. Media* **7** (2012) 503–524.

[16] P. Henning and M. Ohlberger, The heterogeneous multiscale finite element method for advection-diffusion problems with rapidly oscillating coefficients and large expected drift. *Netw. Heterog. Media* **5** (2010) 711–744.

[17] P. Henning and M. Ohlberger, A Note on Homogenization of Advection-Diffusion Problems with Large Expected Drift. *Z. Anal. Anwend.* **30** (2011) 319–339.

[18] P. Henning and M. Ohlberger, Error control and adaptivity for heterogeneous multiscale approximations of nonlinear monotone problems. Preprint 01/11 – N, to appear in *DCDS-S, special issue on Numerical Methods based on Homogenization and Two-Scale Convergence* (2011).

[19] P. Henning and D. Peterseim, Oversampling for the Multiscale Finite Element Method. *SIAM Multiscale Model. Simul.* **12** (2013) 1149–1175.

[20] T. Hou and X.-H. Wu, A multiscale finite element method for elliptic problems in composite materials and porous media. *J. Comput. Phys.* **134** (1997) 169–189.

[21] 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** (1998) 3–24.

[22] 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** (2007) 539–557.

[23] W.R. Gardner, Some steady state solutions of unsaturated moisture ßow equations with application to evaporation from a water table. *Soil Sci.* **85** (1958) 228–232.

[24] M.T. van Genuchten, A closedform equation for predicting the hydraulic conductivity of unsaturated soils. *Soil Sci. Soc. Am. J.* **44** (1980) 892–898.

[25] J. Karátson, Characterizing Mesh Independent Quadratic Convergence of Newton's Method for a Class of Elliptic Problems. *J. Math. Anal.* **44** (2012) 1279–1303.

[26] C.T. Kelley, Iterative methods for linear and nonlinear equations. In vol. 16. *SIAM Frontiers in Applied Mathematics* (1996).

[27] M.G. Larson and A. Målqvist, Adaptive variational multiscale methods based on *a posteriori* error estimation: energy norm estimates for elliptic problems. *Comput. Methods Appl. Mech. Engrg.* **196** (2007) 2313–2324.

[28] M.G. Larson and A. Målqvist, An adaptive variational multiscale method for convection-diffusion problems. *Commun. Numer. Methods Engrg.* **25** (2009) 65–79.

[29] M.G. Larson and A. Målqvist, A mixed adaptive variational multiscale method with applications in oil reservoir simulation. *Math. Models Methods Appl. Sci.* **19** (2009) 1017–1042.

[30] A. Målqvist, Multiscale methods for elliptic problems. *Multiscale Model. Simul.* **9** (2011) 1064–1086.

[31] A. Målqvist and D. Peterseim, Localization of Elliptic Multiscale Problems. To appear in *Math. Comput.* (2011). Preprint `arXiv:1110.0692v4`.

[32] Y. Mualem, A New Model for Predicting the Hydraulic Conductivity of Unsaturated Porous Media. *Water Resour. Res.* **12** (1976) 513–522.

[33] J.M. Nordbotten, Adaptive variational multiscale methods for multiphase flow in porous media. *SIAM Multiscale Model. Simul.* **7** (2008) 1455–1473.

[34] D. Peterseim, Robustness of Finite Element Simulations in Densely Packed Random Particle Composites. *Netw. Heterog. Media* **7** (2012) 113–126.

[35] D. Peterseim and S.A. Sauter, Finite Elements for Elliptic Problems with Highly Varying, Non-Periodic Diffusion Matrix. *SIAM Multiscale Model. Simul.* **10** (2012) 665–695.

[36] M. Růžička, Nichtlineare Funktionalanalysis. *Oxford Mathematical Monographs*. Springer-Verlag, Berlin, Heidelberg, New York (2004).

Appendix B
Pollution-free high-frequency acoustic scattering

B.1 Eliminating the pollution effect in Helmholtz problems by local subscale correction

arXiv:1411.1944 [math.NA], 2014.
(submitted for publication in *Mathematics of Computation*)

ELIMINATING THE POLLUTION EFFECT IN HELMHOLTZ PROBLEMS BY LOCAL SUBSCALE CORRECTION

DANIEL PETERSEIM

Abstract. We introduce a new Petrov-Galerkin multiscale method for the numerical approximation of the Helmholtz equation with large wave number κ in bounded domains in $\mathbb{R}^d$. The discrete trial and test spaces are generated from standard mesh-based finite elements by local subscale corrections in the spirit of numerical homogenization. The precomputation of the corrections involves the solution of coercive cell problems on localized subdomains of size ℓH; H being the mesh size and ℓ being the oversampling parameter. If the mesh size and the oversampling parameter are such that $H\kappa$ and $\log(\kappa)/\ell$ fall below some generic constants, the method is stable and its error is proportional to H; pollution effects are eliminated in this regime.

1. Introduction

The numerical solution of the Helmholtz equation by the finite element method or related schemes in the regime of large wave numbers is still among the most challenging tasks of computational partial differential equations. The highly oscillatory nature of the solution plus a wave number dependent pollution effect puts very restrictive assumptions on the smallness of the underlying mesh. Typically, this condition is much stronger than the minimal requirement for a meaningful representation of highly oscillatory functions from approximation theory, that is, to have at least $5-10$ degrees of freedom per wave length and coordinate direction.

The wave number dependent preasymptotic effect denoted as pollution or numerical dispersion is well understood by now and many attempts have been made to overcome or at least reduce it; see [TF06, FW09, FW11, HMP11, ZMD+11, DGMZ12] among many others. However, for many standard methods, this is not possible in 2d or 3d [BS00]. A breakthrough in this context is the work of Melenk and Sauter [MS10, MS11, MPS13]. It shows that for certain model Helmholtz problems, the pollution effect can be suppressed by simply coupling the polynomial degree p of the Galerkin finite element space to the wave number κ via the relation $p \approx \log \kappa$. Under this moderate assumption, the method is stable and quasi-optimal if the mesh size H satisfies $H\kappa \lesssim 1$. It is worth noting that this result does not require the analyticity of the solution but only $W^{2,2}$-regularity and, thus, partially explains the common sense that higher-order methods are less sensitive to pollution. However, for less regular solutions as they appear for the scattering of waves from non-smooth objects, the result is not directly applicable and the existence of a pollution-free discretization scheme remained open.

Scale-dependent preasymptotic effects are also observed in simpler diffusion problems with highly oscillatory diffusion tensor and numerical homogenization provides techniques to avoid those effects. Numerical homogenization (or upscaling) refers to a class of multiscale methods for the efficient approximation on coarse meshes that do not resolve the coefficient oscillations. A novel method for this problem was recently introduced in [MP14b] and further generalized in [EGMP13, HMP14b, HP13, HMP14a]. The method is based on localizable orthogonal decompositions (LOD) into a low-dimensional coarse space (where we are looking for the approximation) and a high-dimensional remainder space. Some selectable quasi-interpolation operator serves as the basis of the decompositions. The coarse space is spanned

Date: May 6, 2015.

2000 *Mathematics Subject Classification.* 65N12, 65N15,65N30.

Key words and phrases. pollution effect, finite element, multiscale method, numerical homogenization.

by some precomputable basis functions with local support. The method provides text book convergence independent of the variations of the coefficient and without any preasymptotic effects under fairly general assumptions on the diffusion coefficient; periodicity or scale separation are not required.

This paper adapts the multiscale method of [MP14b] to cure pollution in the numerical approximation of the Helmholtz problem. To deal with the lack of hermitivity we will propose a Petrov-Galerkin version of the method (although this is not essential). We will construct a finite-dimensional trial space and corresponding test space for the approximation of the unknown solution u. The trial and test spaces are generated from standard mesh-based finite elements by local subscale corrections. The precomputation of the corrections involves the solution of H^{-d} elliptic (cell) problems on localized subdomains of size ℓH; H being the mesh size and ℓ being the adjustable oversampling parameter. If the data of the problem (domain, boundary condition, force term) allows for polynomial-in-κ bounds of the solution operator and if the mesh size and the oversampling parameter of the method are such that the resolution condition $H\kappa \lesssim 1$ and the oversampling condition $\log(\kappa)/\ell \lesssim 1$ are satisfied, then the method is stable and satisfies the error estimate

$$\kappa\|u - u_{\mathrm{msPG}}\|_{L^2(\Omega)} + \|\nabla(u - u_{\mathrm{msPG}})\|_{L^2(\Omega)} \leq C(H + \beta^\ell)\|f\|_{L^2(\Omega)}$$

with generic constants $C > 0$ and $\beta < 1$ independent of κ. For a fairly large class of Helmholtz problems, including the acoustic scattering from convex non-smooth objects, this result shows that pollution effects can be suppressed under the quasi-minimal resolution condition $H\kappa \leq \mathcal{O}(1)$ at the price of a moderate increase of the inter-element communication, i.e., logarithmic-in-κ oversampling. Using a terminology from finite difference methods, this means that the stencil is moderately enlarged. The complexity overhead due to oversampling is comparable with that of [MS10, MS11], where instead of increasing the inter-element communication, the number of degrees of freedom per element is increased via the polynomial degree which is coupled to $\log \kappa$ in a similar way.

While [BS00] shows that pollution cannot be avoided with a fixed stencil, the result shows that already a logarithmic-in-κ growths of the stencil can suffice to eliminate pollution. Although the result is constructive, its practical relevance for actual computations is not immediately clear in any case. The multiscale method presented in this paper requires precomputations on subgrids. These precomputations are both local and independent, but the worst-case (serial) complexity of the method can exceed the cost of a direct numerical simulation on a global sufficiently fine mesh. However, we expect a significant gain with respect to computational complexity in the following cases:

- The precomputation can be reused several times, e.g., if the problem (with the same geometric setting and wave number) has to be solved for a large number of force terms or incident wave directions in the context of parameter studies, coupled problems or optimal control problems.
- The (local) periodicity of the computational mesh can be exploited so that the number of local problems can be reduced drastically.

We also expect that the redundancy of the local problems can be exploited in rather general unstructured meshes by modern techniques of model order reduction [RHP08, AB14]. However, this possibility requires a careful algorithmic design and error analysis which are beyond the scope of this paper and remain a future perspective of the method. A similar statement applies to the case of heterogeneous media. This application and the generalization of the method are very natural and straight forward. Though this case is not yet covered, previous work [MP14b, EGMP13, HMP14b, HP13, HMP14a] plus the analysis of this paper strongly indicate the potential of the method to treat high oscillations or jumps in the PDE coefficients and the pollution effect in one stroke.

The remaining part of the paper is outlined as follows. Section 2 defines the model Helmholtz problem and recalls some of its fundamental properties. Section 3 introduces standard finite element spaces and corresponding quasi-interpolation operators that will be

the basis for the derivation of a prototypical multiscale method in Section 4. Sections 5 and 6 will then turn this ideal approach into a feasible method including a rigorous stability and error analysis. Finally, Section 7 demonstrates the performance of the method and one of its variants in numerical experiments.

2. Model Helmholtz problem

We consider the Helmholtz equation over a bounded Lipschitz domain $\Omega \subset \mathbb{R}^d$ $(d = 1, 2, 3)$,

$$-\Delta u - \kappa^2 u = f \quad \text{in } \Omega, \tag{2.1.a}$$

along with mixed boundary conditions of Dirichlet, Neumann and Robin type

$$u = 0 \quad \text{on } \Gamma_D, \tag{2.1.b}$$

$$\nabla u \cdot \nu = 0 \quad \text{on } \Gamma_N, \tag{2.1.c}$$

$$\nabla u \cdot \nu - i\kappa u = 0 \quad \text{on } \Gamma_R. \tag{2.1.d}$$

Here, the wave number κ is real and positive, i denotes the imaginary unit and $f \in L^2(\Omega)$ (the space of complex-valued square-integrable functions over Ω). In this paper, we assume that the boundary $\Gamma := \partial\Omega$ consists of three components

$$\partial\Omega = \overline{\Gamma_D \cup \Gamma_N \cup \Gamma_R},$$

where Γ_D, Γ_N and Γ_R are disjoint. We allow that Γ_D or Γ_N are empty but we assume that Γ_R has a positive surface measure,

$$|\Gamma_R| > 0. \tag{2.2}$$

The vector ν denotes the unit normal vector that is outgoing from Ω. To avoid overloading of the paper, we restrict ourselves to the case of homogeneous boundary conditions. Since inhomogeneous boundary data is very relevant for scattering problems, this case will be treated in the context of a numerical experiment in Section 7.2.

Given the Sobolev space $W^{1,2}(\Omega)$ (the space of complex-valued square-integrable functions over Ω with square integrable weak gradient), we introduce the subspace

$$V := \{v \in W^{1,2}(\Omega) \mid v = 0 \text{ on } \Gamma_D\}$$

along with the κ-weighted norm

$$\|v\|_V := \sqrt{\kappa^2 \|v\|_\Omega^2 + \|\nabla v\|_\Omega^2},$$

where $\|\cdot\|_\Omega$ denotes the L^2-norm over Ω. The variational formulation of the boundary value problem (2.1) seeks $u \in V$ such that, for all $v \in V$,

$$a(u, v) = (f, v)_\Omega, \tag{2.3}$$

where the sesquilinear form $a : V \times V \to \mathbb{C}$ has the form

$$a(u, v) := (\nabla u, \nabla v)_\Omega - \kappa^2 (u, v)_\Omega - i\kappa (u, v)_{\Gamma_R}. \tag{2.4}$$

Here, $(\cdot, \cdot)_\Omega := \int_\Omega u \cdot \bar{v} \, \mathrm{d}x$ abbreviates the canonical inner product of scalar or vector-valued $L^2(\Omega)$ functions and $(\cdot, \cdot)_{\Gamma_R} := \int_{\Gamma_R} u\bar{v} \, \mathrm{d}s$ abbreviates the canonical inner product of $L^2(\Gamma_R)$ (the space of complex-valued square-integrable functions over Γ_R). The sesquilinear form a is bounded, i.e., there is a constant C_a that depends only on Ω such that, for any $u, v \in V$,

$$|a(u, v)| \le C_a \|u\|_V \|v\|_V. \tag{2.5}$$

The presence of the impedance boundary condition (2.1.d) (cf. (2.2)) ensures the well-posedness of problem (2.3), i.e., there exists some constant $C_{\mathrm{st}}(\kappa)$ that may depend on κ and also on Ω and the partition of the boundary into Γ_D, Γ_N and Γ_R such that, for any $f \in L^2(\Omega)$, the unique solution $u \in V$ of (2.3) satisfies

$$\|u\|_V \le C_{\mathrm{st}}(\kappa) \|f\|_\Omega. \tag{2.6}$$

However, the stability constant $C_{\mathrm{st}}(\kappa)$ and its possible dependence on the wave number κ are not known in general. Whenever we want to quantify its effect on some parts of the error

analysis, we will assume (cf. Assumption 5.3) that there are constants $C_{\rm pst} > 0$ and $q_{\rm pst} \geq 0$ and $\kappa_0 > 0$ that may depend on Ω and the partition of the boundary into Γ_D, Γ_N and Γ_R such that, for any $\kappa \geq \kappa_0$, the stability constant $C_{\rm st}(\kappa)$ of (2.6) satisfies

$$C_{\rm st}(\kappa) \leq C_{\rm pst}\kappa^{q_{\rm pst}}. \tag{2.7}$$

This polynomial growth condition on the stability constant is certainly not satisfied in general; see [BCWG$^+$11] for the example of a so-called trapping domain that exhibits at least an exponential growth of the norm of the solution operator with respect to the wave number. Hence, the assumption (2.7) puts implicit conditions on the domain Ω and the configuration of the boundary components. Sufficient geometric conditions that ensure (2.7) with $q_{\rm pst} = 0$ are provided in [Het07, EM12, HMP14c] (see also earlier work [Mel95, CF06] that is based on the choice of a particular test function previously used in [MIB96]). Among the known admissible setups are the case of a Robin boundary condition ($\Gamma_R = \partial\Omega$) on a Lipschitz domain Ω [EM12]. Another example is the scattering of acoustic waves at a sound-soft scatterer occupying the star-shaped polygonal or polyhedral domain Ω_D where the Sommerfeld radiation condition is approximated by the Robin boundary condition on the boundary of some artificial convex polygonal or polyhedral domain $\Omega_R \supset \bar\Omega_D$; see [HMP14c].

Given some linear functional g on V, the adjoint problem of (2.3) seeks $z \in V$ such that, for any $v \in V$,

$$a(v, z) = (v, g)_\Omega. \tag{2.8}$$

Note that the adjoint problem is itself a Helmholtz problem in the sense that $S^*(g) = \overline{S(\bar f)}$, where S is the solution operator of (2.3) and S^* is the solution operator of the adjoint problem (2.8); see e.g. [MS11, Lemma 3.1]. Hence, (2.8) enjoys the same stability properties as (2.3).

According to [EM12], the stability (2.6) for $f \in L^2(\Omega)$ implies well-posedness for all bounded linear functionals f on V.

Lemma 2.1 (well-posedness)**.** *The sesquilinear form a of* (2.4) *satisfies*

$$\inf_{u\in V\setminus\{0\}} \sup_{v\in V\setminus\{0\}} \frac{\Re a(u,v)}{\|u\|_V\|v\|_V} \geq \frac{1}{2C_{\rm st}(\kappa)\kappa}. \tag{2.9}$$

Furthermore, for every $f \in V'$ (the space of bounded antilinear functionals on V) the problem (2.1) *is uniquely solvable, and its solution $u \in V$ satisfies the a priori bound*

$$\|u\|_V \leq C_{\rm st}(\kappa)\kappa\|f\|_{V'}. \tag{2.10}$$

Under the additional assumption 2.7 that the stability constant grows at most polynomially in κ, the lemma shows polynomial well-posedness in the sense of [EM12], i.e., polynomial-in-κ-bounds for the norm of the solution operator.

Proof of Lemma 2.1. The proof of (2.9) is almost verbatim the same as that of [EM12, Theorem 2.5] which covers the particular case $\Gamma_R = \partial\Omega$ and relies on a standard argument for sesquilinear forms satisfying a Gårding inequality. Given $u \in V$, define $z \in V$ as the solution of

$$2\kappa^2(v, u)_\Omega = a(v, z), \quad \text{for all } v \in V.$$

The stability (2.6) implies that

$$\|z\|_V \leq 2C_{\rm st}(\kappa)\kappa^2\|u\|_\Omega \leq 2C_{\rm st}(\kappa)\kappa\|u\|_V. \tag{2.11}$$

Set $v = u + z$ and observe that

$$\Re a(u, v) = \|u\|_V^2. \tag{2.12}$$

The combination of (2.11) and (2.12) yields (2.9). Note that an analogue inf-sup condition can be proved for the adjoint of the bilinear form a so that the Banach-Nečas-Babuška theorem yields the unique solvability of both the primal and the adjoint problem as well as the a priori estimate (2.10). □

3. Standard finite element spaces

This section recalls briefly the notions of simplicial finite element meshes and patches, standard finite element spaces and corresponding quasi-interpolation operators. In this paper, we will focus on linear finite elements based on triangles or tetrahedrons but quadrilaterals or even mesh-free approaches would be possible as well. The key property that we will exploit in the later construction is the partition of unity property of the basis; see [HMP14a].

3.1. **Finite element meshes.** We consider two discretization scales $H > h > 0$. Let $\mathcal{T}_H$ (resp. $\mathcal{T}_h$) denote corresponding regular (in the sense of [Cia78]) finite element meshes of Ω into closed simplices with mesh-size functions $0 < H \in L^\infty(\Omega)$ defined by $H|_T = \operatorname{diam} T =: H_T$ for all $T \in \mathcal{T}_H$ (resp. $0 < h \in L^\infty(\Omega)$ defined by $h|_t = \operatorname{diam} t =: h_t$ for all $t \in \mathcal{T}_h$). The mesh sizes may vary in space.

Some of the error bounds will depend on the maximal mesh size $\|H\|_{L^\infty(\Omega)}$. If no confusion seems likely, we will use H also to denote the maximal mesh size instead of writing $\|H\|_{L^\infty(\Omega)}$. For the sake of simplicity we assume that $\mathcal{T}_h$ is derived from $\mathcal{T}_H$ by some regular, possibly non-uniform, mesh refinement. However, this condition is not essential and we refer to [HMP14a] where possible generalizations are discussed in the context of a diffusion problem.

As usual, the error analysis depends on some constant $\gamma > 0$ that represents the shape regularity of the finite element mesh $\mathcal{T}_H$;

$$\gamma := \max_{T\in\mathcal{T}_H} \gamma_T \quad \text{with} \quad \gamma_T := \frac{\operatorname{diam} T}{\operatorname{diam} B_T} \ \text{ for } T \in \mathcal{T}_H, \tag{3.1}$$

where B_T denotes the largest ball inscribed in T.

3.2. **Nodal patches and element patches.** Patches are agglomerations of elements of $\mathcal{T}_H$. They will often be used in the construction of the method and its analysis. We define patches $\omega_{T,\ell}$ of variable order $\ell \in \mathbb{N}$ about an element $T \in \mathcal{T}_H$ by

$$\begin{cases} \omega_{T,1} := \cup\{T' \in \mathcal{T}_H \mid T' \cap T \neq \emptyset\}, \\ \omega_{T,\ell} := \cup \left\{T' \in \mathcal{T}_H \mid T' \cap \overline{\omega}_{T,\ell-1} \neq \emptyset\right\}, \ \ell = 2, 3, 4 \ldots. \end{cases} \tag{3.2}$$

In other words, $\omega_{T,1}$ equals the union of T and its neighbors and $\omega_{t,\ell}$ is derived from $\omega_{T,\ell-1}$ by adding one more layer of neighbors.

Note that, for a fixed $\ell \in \mathbb{N}$, the element patches have finite overlap in the following sense. There exists a constant $C_{\mathrm{ol},\ell} > 0$ such that

$$\max_{T\in\mathcal{T}_H} \#\{K \in \mathcal{T}_H \mid K \subset \omega_{T,\ell}\} \leq C_{\mathrm{ol},\ell}. \tag{3.3}$$

The constant $C_{\mathrm{ol}} := C_{\mathrm{ol},1}$ equals the maximal number of neighbors of an element plus itself and there exists some generic constant C'_{ol} such that, for any $\ell > 1$,

$$C_{\mathrm{ol},\ell} \leq \max\left\{\#\mathcal{T}_H, C'_{\mathrm{ol}}\ell^d \|H\|_{L^\infty(\omega_{T,\ell})}\|H^{-1}\|_{L^\infty(\omega_{T,\ell})}\right\}.$$

3.3. **Standard finite element spaces.** The first-order conforming finite element space with respect to the mesh $\mathcal{T}_H$ is given by

$$V_H := \{v \in V \mid \forall T \in \mathcal{T}_H, v|_T \text{ is a polynomial of total degree} \leq 1\}. \tag{3.4}$$

Let $\mathcal{N}_H$ denote the set of all vertices of $\mathcal{T}_H$ that are not elements of the Dirichlet boundary. Every vertex $z \in \mathcal{N}_H$ represents a degree of freedom via the corresponding real-valued nodal basis function $\phi_z \in V_H$ determined by nodal values

$$\phi_z(z) = 1 \ \text{ and } \ \phi_z(y) = 0 \quad \text{for all } y \neq z \in \mathcal{N}_H.$$

The ϕ_z form a basis of V_H and the dimension of V_H equals the number of vertices (excluding the Dirichlet boundary Γ_D),

$$N_H := \dim V_H = |\mathcal{N}_H|.$$

Let $V \supset V_h \supset V_H$ denote some conforming finite element space that corresponds to the fine mesh $\mathcal{T}_h$. It can be the space of continuous piecewise affine functions on the fine mesh or any other (generalized) finite element space that contains V_H, e.g., the space of continuous p-th

order piecewise polynomials as in [MS10, MS11]. By $N_h := \dim V_h$ we denote the dimension of V_h. For standard choices of V_h, this dimension is proportional to the number of vertices in the fine mesh $\mathcal{T}_h$ (excluding vertices on the Dirichlet boundary Γ_D).

3.4. **Quasi-interpolation.** A key tool in the design and the analysis of the method is some bounded linear surjective Clément-type (quasi-)interpolation operator $\mathcal{I}_H : V \to V_H$ as it is used in the a posteriori error analysis of finite element methods [CV99]. Given $v \in V$, $\mathcal{I}_H v := \sum_{z\in\mathcal{N}_H} \alpha_z(v)\phi_z$ defines a (weighted) Clément interpolant with nodal functionals

$$\alpha_z(v) := \frac{(v,\phi_z)_\Omega}{(1,\phi_z)_\Omega} \tag{3.5}$$

for $z \in \mathcal{N}_H$. Recall the (local) approximation and stability properties of the interpolation operator $\mathcal{I}_H$. There exists a generic constant $C_{\mathcal{I}_H}$ such that, for all $v \in V$ and for all $T \in \mathcal{T}_H$ and any face F of T,

$$H_T^{-1/2}\|v - \mathcal{I}_H v\|_{L^2F} + H_T^{-1}\|v - \mathcal{I}_H v\|_{L^2(T)} + \|\nabla(v - \mathcal{I}_H v)\|_{L^2(T)} \le C_{\mathcal{I}_H}\|\nabla v\|_{L^2(\omega_T)}, \tag{3.6}$$

where $\omega_T = \omega_{T,1}$ from (3.2). The constant $C_{\mathcal{I}_H}$ depends on the shape regularity parameter γ of the finite element mesh $\mathcal{T}_H$ (see (3.1) above) but not on the local mesh size H_T. The proof for the volume errors is given in [CV99]. The bound on the face error follows from those bounds and the trace inequality

$$\|v\|_F^2 \le 2\gamma_T \left(2\|\nabla v\|_T + dH_T^{-1}\|v\|_T\right)\|v\|_T. \tag{3.7}$$

The trace inequality is a consequence of the trace identity of [CF00] and the Young inequality; see [DPE12, Lemma 1.49] for a detailed proof.

Note that the space V_H is invariant under $\mathcal{I}_H$ but $\mathcal{I}_H$ is not a projection, i.e., $\mathcal{I}_H v_H \neq v_H$ for $v_H \in V_H$ in general. However, since $\mathcal{I}_H|_{V_H}$ can be interpreted as a diagonally scaled mass matrix, $\mathcal{I}_H$ is invertible on the finite element space V_H and the concatenation $(\mathcal{I}_H|_{V_H})^{-1}\circ\mathcal{I}_H : V \to V_H$ is a projection. For our particular choice of interpolation operator, one easily verifies that $(\mathcal{I}_H|_{V_H})^{-1}\circ\mathcal{I}_H$ equals the L^2-orthogonal projection $\Pi_H : V \to V_H$ onto the finite element space; see also [MP14a, Remark 3.1]. Recall that Π_H is also stable in V,

$$\|\Pi_H v\|_V \le C_{\Pi_H}\|v\|_V \quad \text{for all } v \in V, \tag{3.8}$$

where C_{Π_H} depends only on the parameter γ if the grading of the mesh is not too strong [BY14].

While $\mathcal{I}_H|_{V_H}$ is a local operator (a sparse matrix) its inverse $(\mathcal{I}_H|_{V_H})^{-1}$ is not. However, there exists some bounded right inverse $\mathcal{I}_H^{-1,\mathrm{loc}} : V_H \to V$ of $\mathcal{I}_H$ that is local. More precisely, there exists some generic constant $C'_{\mathcal{I}_H}$ depending only on γ such that, for all $v_H \in V_H$,

$$\left\{\begin{array}{rcl} \mathcal{I}_H(\mathcal{I}_H^{-1,\mathrm{loc}} v_H) &=& v_H, \\ \|\nabla \mathcal{I}_H^{-1,\mathrm{loc}} v_H\|_\Omega &\le& C'_{\mathcal{I}_H}\|\nabla v_H\|_\Omega, \\ \mathrm{supp}(\mathcal{I}_H^{-1,\mathrm{loc}} v_H) &\overset{1}{\subseteq}& \mathrm{supp}(v_H), \end{array}\right. \tag{3.9}$$

where the $\overset{1}{\subseteq}$ is a short-hand notation for $\mathrm{supp}(\mathcal{I}_H^{-1,\mathrm{loc}} v_H) \subseteq \bigcup\{\omega_{T,1} \mid T \in \mathcal{T}_H : \ T \cap \mathrm{supp}(v_H) \neq \emptyset\}$. Note that $\mathcal{I}_H^{-1,\mathrm{loc}} v_H$ is not a finite element function on $\mathcal{T}_H$ in general. An explicit construction of $\mathcal{I}_H^{-1,\mathrm{loc}}$ and a proof of the properties (3.9) can be found in [HMP14a, Lemma 1].

We shall emphasize that the choice of a quasi-interpolation operator is by no means unique and a different choice might lead to a different multiscale method. A choice that turned out to be useful in previous works [BP14, PS14] is the following one. Given $v \in V$, $\mathcal{Q}_H v := \sum_{z\in\mathcal{N}_H} \alpha_z(v)\phi_z$ defines a Clément-type interpolant with nodal functionals

$$\alpha_z(v) := (\Pi_{H,\omega_z} v)(z) \tag{3.10}$$

for $z \in \mathcal{N}_H$. Here, $\Pi_{H,\omega_z} v$ denotes the L^2-orthogonal projection of v onto standard P_1 finite elements on the patch ω_z and $\alpha_z(v)$ is the evaluation of this projection at the vertex z. We

will show in the numerical experiment of Section 7 that the choice of the interpolation can affect the practical performance of the method significantly.

4. Global wave number adapted approximation

This section introduces new (non-polynomial) approximation spaces for the model Helmholtz problem under consideration. The spaces are mesh-based in the sense that degrees of freedom (or basis functions) are associated with vertices. The support of the basis functions is not local in general but quasi-local in the sense of some very fast decay of their moduli. Their replacement by localized computable basis functions in practical computations is possible; see Sections 5 and 6.

The ideal method requires the following assumption on the numerical resolution.

Assumption 4.1 (resolution condition)**.** *Given the wave number κ and the constants $C_{\mathcal{I}_H}$ from* (3.6) *and C_{ol} from* (3.3)*, we assume that the mesh width H satisfies*

$$H\kappa \leq \frac{1}{\sqrt{2C_{\mathrm{ol}}}C_{\mathcal{I}_H}}. \tag{4.1}$$

Note that this assumption is quasi-minimal in the sense that a certain number of degrees of freedom per wave length is a necessary condition for the meaningful approximation of highly oscillatory waves.

4.1. An ideal method.

The derivation of the method follows general principles of variational multiscale methods; cf. [Hug95, HFMQ98, HS07] and [Mål11]. Our construction of the approximation space starts with the observation that the space V can be decomposed into the finite element space V_H and the remainder space

$$R_H := \operatorname{kernel} \mathcal{I}_H. \tag{4.2}$$

The particular choice of $\mathcal{I}_H$ implies that the decomposition

$$V = V_H \oplus R_H \tag{4.3}$$

is orthogonal in $L^2(\Omega)$ and, hence, stable. We shall say that this L^2-orthogonality will not be crucial in this paper and that any choice of $\mathcal{I}_H$ that allows a stable splitting of V into its image and its kernel is possible, for instance $\mathcal{Q}_H$ defined in (3.10).

The subscale corrector $\mathcal{C}_\infty$ is a linear operator that maps V onto R_H. Given $v \in V$, define the corrector $\mathcal{C}_\infty v \in R_H$ as the unique solution (cf. Lemma 4.2 below) of the variational problem

$$a(\mathcal{C}_\infty v, w) = a(v, w), \quad \text{for all } w \in R_H. \tag{4.4}$$

The subscript notation ∞ will be consistent with later modifications $\mathcal{C}_\ell$ of the corrector, where the computation is restricted to local subdomains of size ℓH.

To deal with the lack of hermitivity, we will use the adjoint corrector $\mathcal{C}^*_\infty v \in R_H$ that solves the adjoint variational problem

$$a(w, \mathcal{C}^*_\infty v) = a(w, v), \quad \text{for all } w \in R_H. \tag{4.5}$$

It turns out that

$$\mathcal{C}^*_\infty v = \overline{\mathcal{C}_\infty \bar{v}} \tag{4.6}$$

holds for the model problem under consideration. Under Assumption 4.1, the corrector problems (4.4) and (4.5) are well-posed.

Lemma 4.2 (well-posedness of the correction operator)**.** *The resolution condition of Assumption 4.1 implies that $\|\nabla \cdot\|_\Omega$ and $\|\cdot\|_V$ are equivalent norms on R_H,*

$$\|\nabla w\|_\Omega \leq \|w\|_V \leq \sqrt{\tfrac{3}{2}}\|\nabla w\|_\Omega, \quad \textit{for all } w \in R_H, \tag{4.7}$$

the sesquilinear form a is R_H-elliptic,

$$\Re a(w, w) \geq \tfrac{1}{3}\|w\|^2_V, \quad \textit{for all } w \in R_H, \tag{4.8}$$

and the correction operators $\mathcal{C}_\infty$, $\mathcal{C}_\infty^*$ are well-defined and stable,

$$\|\mathcal{C}_\infty v\|_V = \|\mathcal{C}_\infty^* v\|_V \leq C_{\mathcal{C}} \|v\|_V, \quad \text{for all } v \in V, \tag{4.9}$$

where $C_{\mathcal{C}} := 3C_a$ with C_a from (2.5).

Proof. For any $w \in R_H$, the property $\mathcal{I}_H w = 0$, the approximation property (3.6) of the quasi-interpolation operator, the bounded overlap of element patches C_{ol} and (4.1) yield

$$\begin{aligned}\kappa^2(w,w)_\Omega &= \kappa^2(w - \mathcal{I}_H w, w - \mathcal{I}_H w)_\Omega \\ &\leq C_{\mathrm{ol}} C_{\mathcal{I}_H}^2 \kappa^2 H^2 \|\nabla w\|_\Omega^2 \\ &\leq \tfrac{1}{2}\|\nabla w\|_\Omega^2.\end{aligned}$$

This implies (4.7) and (4.8). Since the sesquilinear form a is bounded (2.5), the well-posedness of (4.4) and (4.5) and the stability estimate (4.9) follow from the Lax-Milgram theorem. □

Since non-trivial projections on Hilbert spaces have the same operator norm as their complementary projections (see [Szy06] for a proof), the continuity of the projection operators $\mathcal{C}_\infty, \mathcal{C}_\infty^*$ implies the continuity of their complementary projections $(1-\mathcal{C}_\infty), (1-\mathcal{C}_\infty^*) : V \to V$, that is,

$$\|(1-\mathcal{C}_\infty)v\|_V = \|(1-\mathcal{C}_\infty^*)v\|_V \leq C_{\mathcal{C}} \|v\|_V, \quad \text{for all } v \in V, \tag{4.10}$$

where $C_{\mathcal{C}} = 3C_a$ is the constant from (4.9)

The image of the finite element space V_H under $(1-\mathcal{C}_\infty)$,

$$V_{H,\infty} := (1-\mathcal{C}_\infty)V_H, \tag{4.11}$$

defines a modified discrete approximation space. The space $V_{H,\infty}$ will be the prototypical trial space in our method. The corresponding test space is

$$V_{H,\infty}^* := (1-\mathcal{C}_\infty^*)V_H. \tag{4.12}$$

Note that R_H equals the kernel of both operators, $(1-\mathcal{C}_\infty)$ and $(1-\mathcal{C}_\infty^*)$. This implies that $V_{H,\infty}$ is the image of $(1-\mathcal{C}_\infty)$ and $V_{H,\infty}^*$ is the image of $(1-\mathcal{C}_\infty^*)$. The key properties of the spaces $V_{H,\infty}$ and $V_{H,\infty}^*$ are given in the subsequent lemma.

Lemma 4.3 (primal and dual decomposition)**.** *If the resolution condition of Assumption 4.1 is satisfied, then the decompositions*

$$V = V_{H,\infty} \oplus R_H = V_{H,\infty}^* \oplus R_H$$

are stable. More precisely, any function $v \in V$ can be decomposed uniquely into

$$v = v_{H,\infty} + r_H \quad \text{and} \quad v = w_{H,\infty} + \overline{r_H}$$

and

$$\max\{\|v_{H,\infty}\|_V, \|w_{H,\infty}\|_V, \|r_H\|_V\} \leq C_{\mathcal{C}} \|v\|_V,$$

where $v_{H,\infty} := (1-\mathcal{C}_\infty)v \in V_{H,\infty}$, $w_{H,\infty} := (1-\mathcal{C}_\infty^*)v \in V_{H,\infty}^*$ and $r_H := \mathcal{C}_\infty v \in R_H$.

The decompositions satisfy the following relations: For any $v_{H,\infty} \in V_{H,\infty}$ and any $q_H \in R_H$, it holds that

$$a(v_{H,\infty}, q_H) = 0. \tag{4.13}$$

For any $w_{H,\infty} \in V_{H,\infty}^*$ and any $q_H \in R_H$, it holds that

$$a(q_H, w_{H,\infty}) = 0, \tag{4.14}$$

Proof. The results readily follow from the construction of $\mathcal{C}_\infty$ and $\mathcal{C}_\infty^*$. □

The Petrov-Galerkin method for the approximation of (2.3) based on the trial-test pairing $(V_{H,\infty}, V_{H,\infty}^*)$ seeks $u \in V_{H,\infty}$ such that, for all $v_{H,\infty} \in V_{H,\infty}^*$,

$$a(u_{H,\infty}, v_{H,\infty}) = (f, v_{H,\infty})_\Omega. \tag{4.15}$$

We shall emphasize that we do not consider the method (4.15) for actual computations because the natural bases of the trial (resp. test) space, i.e. the image of the standard nodal

basis of the finite element space under the operator $1-\mathcal{C}_\infty$ (resp. $1-\mathcal{C}_\infty^*$) is not sparse (or local) in the sense that the basis function $(1-\mathcal{C}_\infty)\phi_z$ (resp. $(1-\mathcal{C}_\infty^*)\phi_z$) have global support in Ω in general. Moreover the corrector problems are infinite dimensional problems. We will, hence, refer to the method (4.15) as the ideal or global method. Later on (cf. Theorem 5.2), we will show that there are feasible nearby spaces with a sparse basis and we will also discretize the (localized) corrector problems and analyze the error committed by those crimes in Section 6.

4.2. **Stability and accuracy of the ideal method.** The ideal method admits a unique solution and is stable and accurate independent of κ as long as the resolution condition $H\kappa \lesssim 1$ is satisfied. The "orthogonality" relation (4.13) induces stability.

Theorem 4.4 (stability)**.** *Let Assumption 4.1 be satisfied. Then the trial space $V_{H,\infty}$ and test space $V_{H,\infty}^*$ satisfy the discrete inf-sup condition*

$$\inf_{u_{H,\infty}\in V_{H,\infty}\setminus\{0\}} \sup_{v_{H,\infty}\in V_{H,\infty}^*\setminus\{0\}} \frac{\Re a(u_{H,\infty},v_{H,\infty})}{\|u_{H,\infty}\|_V\|v_{H,\infty}\|_V} \geq \frac{1}{2C_{\mathcal{C}}C_{\mathrm{st}}(\kappa)\kappa}. \tag{4.16}$$

Proof. Observe that $(1-\mathcal{C}_\infty^*): V\to V_{H,\infty}^*$ is a Fortin operator (as in the theory of mixed methods [For77]), i.e., a bounded linear operator that satisfies

$$a(u_{H,\infty},(1-\mathcal{C}_\infty^*)v) = a(u_{H,\infty},v) - \underbrace{a(u_{H,\infty},\mathcal{C}_\infty^* v)}_{=0;\text{see (4.13)}} = a(u_{H,\infty},v),$$

for all $u_{H,\infty}\in V_{H,\infty}$ and any $v\in V$. Hence, the assertion follows from the inf-sup condition (2.9) on the continuous level and the continuity of $1-\mathcal{C}_\infty^*$ (4.9),

$$\begin{aligned}
&\inf_{u_{H,\infty}\in V_{H,\infty}\setminus\{0\}} \sup_{v_{H,\infty}\in V_{H,\infty}^*\setminus\{0\}} \frac{\Re a(u_{H,\infty},v_{H,\infty})}{\|u_{H,\infty}\|_V\|v_{H,\infty}\|_V}\\
&\qquad = \inf_{u_{H,\infty}\in V_{H,\infty}\setminus\{0\}} \sup_{v\in V\setminus\{0\}} \frac{\Re a(u_{H,\infty},(1-\mathcal{C}^*)v)}{\|u_{H,\infty}\|_V\|(1-\mathcal{C}^*)v\|_V}\\
&\qquad \geq \frac{1}{C_{\mathcal{C}}} \inf_{u\in V\setminus\{0\}} \sup_{v\in V\setminus\{0\}} \frac{\Re a(u,v)}{\|u\|_V\|v\|_V}\\
&\qquad \geq \frac{1}{2C_{\mathcal{C}}C_{\mathrm{st}}(\kappa)\kappa}.
\end{aligned}$$

□

The error estimate follows from the above discrete inf-sup condition, the "orthogonality" relation (4.14), and the Lax-Milgram theorem.

Theorem 4.5 (error of the ideal method)**.** *Let $u\in V$ solve* (2.3)*. If the resolution condition of Assumption 4.1 is satisfied, then $u_{H,\infty}=(1-\mathcal{C}_\infty)u\in V_{H,\infty}$ is the unique solution of* (4.15)*, that is, the Petrov-Galerkin approximation of u in the subspace $V_{H,\infty}$ with respect to the test space $V_{H,\infty}^*$. Moreover, it holds that*

$$\|u-u_{H,\infty}\|_V \leq 3\sqrt{C_{\mathrm{ol}}}C_{\mathcal{I}_H}\|Hf\|_\Omega. \tag{4.17}$$

Proof. The Galerkin property (4.15) of $u_{H,\infty}=(1-\mathcal{C}_\infty)u$ follows from (4.14). Hence, the error $u-u_{H,\infty}=\mathcal{C}_\infty u\in R_H$ satisfies

$$a(\mathcal{C}_\infty u,\mathcal{C}_\infty u) = a(u,\mathcal{C}_\infty u) = (f,\mathcal{C}_\infty u)_\Omega.$$

Since the sesquilinear form a is R_H-elliptic (cf. (4.8)), this yields the error estimate

$$\|u-u_{H,\infty}\|_V^2 \leq 3\left|(f,\mathcal{C}_\infty u)_\Omega\right|.$$

Since $\mathcal{I}_H\mathcal{C}_\infty u=0$, Cauchy inequalities and the interpolation error estimate (3.6) readily yield the assertion. □

Remark 4.1 (quasi-optimality). We shall say that the ideal method is also quasi-optimal in the following sense

$$\|u - u_{H,\infty}\|_V \leq 3C_a \inf_{v_{H,\infty}\in V_{H,\infty}} \|u - v_{H,\infty}\|_V. \tag{4.18}$$

Moreover, since $\Pi_H \mathcal{C}_\infty u = 0$, it holds that $\Pi_H u = \Pi_H u_{H,\infty}$. This means that the ideal method provides the L^2-best approximation in the standard finite element space V_H,

$$\|u - \Pi_H u_{H,\infty}\|_\Omega = \min_{v_H\in V_H} \|u - v_H\|_\Omega. \tag{4.19}$$

Since $\mathcal{I}_H(u - u_{H,\infty}) = 0$, $u_{H,\infty}$ also satisfies the L^2 bound

$$\|u - u_{H,\infty}\|_\Omega \leq \sqrt{C_{\rm ol}} C_{\mathcal{I}_H} H \|u - u_{H,\infty}\|_V. \tag{4.20}$$

Remark 4.2 (further stable variants of the method). We shall also mention at this point that the "orthogonality" relations (4.13) and (4.14) imply that, for any $u_H, v_H \in V_H$,

$$\begin{aligned} a((1-\mathcal{C}_\infty)u_H, v_H) &= a((1-\mathcal{C}_\infty)u_H, (1-\mathcal{C}_\infty)v_H) \\ &= a((1-\mathcal{C}_\infty)u_H, (1-\mathcal{C}^*_\infty)v_H) \\ &= a(u_H, (1-\mathcal{C}^*_\infty)v_H) = a((1-\mathcal{C}^*_\infty)u_H, (1-\mathcal{C}^*_\infty)v_H). \end{aligned}$$

This means that the Galerkin methods in $V_{H,\infty}$ or $V^*_{H,\infty}$ as well as Petrov-Galerkin methods based on the pairings $(V_{H,\infty}, V_H)$ or $(V_H, V^*_{H,\infty})$ lead to stable and accurate discretizations. The latter Petrov-Galerkin method based on $(V_H, V^*_{H,\infty})$ is closely related to a variational multiscale stabilization of the standard P_1 finite element method and seeks $u_H \in V_H$ such that, for all $v_H \in V_H$,

$$a(u_H, v_{H,\infty}) - a(u_H, \mathcal{C}^*_\infty v_H) = (f, v_H)_\Omega - (f, \mathcal{C}^*_\infty v_H)_\Omega. \tag{4.21}$$

This stabilized method will be used in the numerical experiment of Section 7.

4.3. **Exponential decay of element correctors.** Given some finite element function $v \in V$, its correction $\mathcal{C}_\infty v_H$ can be composed by element correctors $\mathcal{C}_{T,\infty}$, $T \in \mathcal{T}_H$ in the following way:

$$\mathcal{C}_\infty v = \sum_{T\in\mathcal{T}_H} \mathcal{C}_{T,\infty}(v|_T), \tag{4.22}$$

where $\mathcal{C}_{T,\infty}(v|_T) \in R_H$ solves

$$a(\mathcal{C}_{T,\infty}(v|_T), w) = a_T(v, w) := \int_T \nabla v \cdot \nabla \bar{w}\,\mathrm{d}x - \kappa^2 \int_T v\bar{w}\,\mathrm{d}x - i\kappa \int_{\partial T\cap\Gamma_R} v\bar{w}\,\mathrm{d}s, \tag{4.23}$$

for all $w \in R_H$. Dual corrections can be split into element contributions in an analogue way,

$$\mathcal{C}^* v = \sum_{T\in\mathcal{T}_H} \mathcal{C}^*_T(v|_T), \tag{4.24}$$

where $\mathcal{C}^*_{T,\infty}(v|_T) := \overline{\mathcal{C}_{T,\infty}(\bar{v}|_T)} \in R_H$.

The well-posedness of the element correctors is a consequence of Lemma 4.2. Moreover, it holds that

$$\|\mathcal{C}_{T,\infty} v\|_V = \|\mathcal{C}^*_{T,\infty} v\|_V \leq C_{\mathcal{C}} \|v\|_{V(T)}, \quad \text{for all } v \in V, \tag{4.25}$$

where $V(T)$ denotes the restriction of the space V to the element T, and $\|v\|^2_{V(T)} := \kappa^2\|v\|^2_{L^2(T)} + \|\nabla v\|^2_{L^2(T)}$.

The major observation is that the moduli of the element correctors $\mathcal{C}_T v$ and $\mathcal{C}^*_T v$ decay very fast outside T.

Theorem 4.6 (exponential decay of element correctors)**.** *If the resolution condition of Assumption 4.1 is satisfied, then there exist constants $C_{\rm dec} > 0$ and $\beta < 1$ independent of H*

and κ such that for all $v \in V$ and all $T \in \mathcal{T}_H$ and all $\ell \in \mathbb{N}$, the element corrector $\mathcal{C}_{T,\infty}v$ satisfies

$$\|\nabla \mathcal{C}_{T,\infty} v\|_{\Omega\setminus\omega_{T,\ell}} \leq C_{\mathrm{dec}}\beta^{\ell}\|\nabla v\|_T. \tag{4.26}$$

The constant β is bounded away from 1 by $\left(\frac{C_1}{C_1+\frac{1}{2}}\right)^{1/14} < 1$ and $C_{\mathrm{dec}} \leq \sqrt{C_{\mathcal{C}}\frac{C_1+\frac{1}{2}}{C_1}}$, where

$$C_1 := \tfrac{1}{2} + \tfrac{3}{2}C_{\mathcal{I}_H}C'_{\mathcal{I}_H} + (C'_{\mathcal{I}_H}C_{\mathcal{I}_H}+1)C_{\mathcal{I}_H}\sqrt{C_{\mathrm{ol}}}\gamma$$

depends only on the shape regularity parameter γ of the mesh $\mathcal{T}_H$.

According to practical experience, the bound on the decay rate β seems to be rather pessimistic. The rates observed in numerical experiments were between $\frac{1}{3}$ and $\frac{2}{3}$.

Proof of Theorem 4.6. Let $T \in \mathcal{T}_H$ be arbitrary but fixed and let $\ell \in \mathbb{N}$ with $\ell \geq 7$ and let the element patches $\omega_{T,\ell}, \omega_{T,\ell-1}, \ldots, \omega_{T,\ell-7}$ be defined as in (3.2). Set $\psi := \mathcal{C}_{T,\infty}v$.

We define the cut-off function η (depending on T and ℓ) by

$$\eta(x) := \frac{\operatorname{dist}(x,\omega_{T,\ell-4})}{\operatorname{dist}(x,\omega_{T,\ell-4}) + \operatorname{dist}(x,\Omega\setminus\omega_{T,\ell-3})}$$

for $x \in \Omega$. Note that $\eta = 0$ in the patch $\omega_{T,\ell-4}$ and $\eta = 1$ in $\Omega\setminus\omega_{T,\ell-3}$. Moreover, η is bounded between 0 and 1 and Lipschitz continuous with

$$\|H\nabla\eta\|_{L^\infty(\Omega)} \leq \gamma. \tag{4.27}$$

The choice of η implies the estimates

$$\begin{aligned}
\|\nabla\psi\|^2_{\Omega\setminus\omega_{T,\ell-3}} &= \Re\,(\nabla\psi,\nabla\psi)_{\Omega\setminus\omega_{T,\ell-3}}\\
&\leq \Re\,(\nabla\psi,\eta\nabla\psi)_\Omega\\
&= \Re\,(\nabla\psi,\nabla\,(\eta\psi))_\Omega - \Re\,(\nabla\psi,\psi\nabla\eta)_\Omega\\
&\leq |\Re\,(\nabla\psi,\nabla\left(\eta\psi - \mathcal{I}_H^{-1,\mathrm{loc}}(\mathcal{I}_H(\eta\psi))\right))_\Omega|\\
&\qquad + |\Re\,(\nabla\psi,\nabla\mathcal{I}_H^{-1,\mathrm{loc}}(\mathcal{I}_H(\eta\psi)))_\Omega| + |\Re\,(\nabla\psi,\psi\nabla\eta)_\Omega|\\
&=: M_1 + M_2 + M_3. \qquad (4.28)
\end{aligned}$$

Note that the test function $\left(\eta\psi - \mathcal{I}_H^{-1,\mathrm{loc}}(\mathcal{I}_H(\eta\psi))\right) \in R_H$ with support in $\Omega\setminus\omega_{T,\ell-6}$. If $\ell \geq 6$, then $\eta\psi - \mathcal{I}_H^{-1,\mathrm{loc}}(\mathcal{I}_H(\eta\psi))$ vanishes on T and $a_T(v,\eta\psi - \mathcal{I}_H^{-1,\mathrm{loc}}(\mathcal{I}_H(\eta\psi))) = 0$. Hence, the definition (4.5) of $\mathcal{C}_{T,\infty}$, the Cauchy-Schwarz inequality, the properties (3.6) and (3.9) of the interpolation operator $\mathcal{I}_H$ and the resolution condition Assumption 4.1 imply

$$\begin{aligned}
M_1 &:= |\Re\,(\nabla\psi,\nabla(\eta\psi - \mathcal{I}_H^{-1,\mathrm{loc}}(\mathcal{I}_H(\eta\psi))))_\Omega|\\
&= \left|\kappa^2(\psi,\eta\psi - \mathcal{I}_H^{-1,\mathrm{loc}}(\mathcal{I}_H(\eta\psi)))_\Omega\right|\\
&\leq C^2_{\mathcal{I}_H}C_{\mathrm{ol}}(H\kappa)^2\|\nabla\psi\|^2_{\Omega\setminus\omega_{T,\ell-6}} + C^3_{\mathcal{I}_H}C'_{\mathcal{I}_H}C_{\mathrm{ol}}(H\kappa)^2\|\nabla\psi\|^2_{\omega_{T,\ell}\setminus\omega_{T,\ell-7}}\\
&\leq \tfrac{1}{2}\|\nabla\psi\|^2_{\Omega\setminus\omega_{T,\ell}} + \tfrac{1}{2}(1 + C_{\mathcal{I}_H}C'_{\mathcal{I}_H})\|\nabla\psi\|^2_{\omega_{T,\ell}\setminus\omega_{T,\ell-7}}. \qquad (4.29)
\end{aligned}$$

Similar techniques and the Lipschitz bound (4.27) lead to upper bounds of the other terms on the right-hand side of (4.28),

$$\begin{aligned}
M_2 &\leq C'_{\mathcal{I}_H}C_{\mathcal{I}_H}\|\nabla(\eta\psi)\|_{\omega_{T,\ell-1}\setminus\omega_{T,\ell-6}}\|\nabla\psi\|_{\omega_{T,\ell-1}\setminus\omega_{T,\ell-6}}\\
&\leq C'_{\mathcal{I}_H}C_{\mathcal{I}_H}\left(C_{\mathcal{I}_H}\sqrt{C_{\mathrm{ol}}}\|H\nabla\eta\|_{L^\infty(\Omega)} + 1\right)\|\nabla\psi\|^2_{\omega_{T,\ell}\setminus\omega_{T,\ell-7}} \qquad (4.30)
\end{aligned}$$

and

$$M_3 \leq C_{\mathcal{I}_H}\sqrt{C_{\mathrm{ol}}}\|H\nabla\eta\|_{L^\infty(\Omega)}\|\nabla\psi\|^2_{\omega_{T,\ell-2}\setminus\omega_{T,\ell-5}}. \tag{4.31}$$

The combination of (4.28)–(4.31) readily yields the estimate

$$\tfrac{1}{2}\|\nabla\psi\|^2_{\Omega\setminus\omega_{T,\ell}} \leq C_1\|\nabla\psi\|^2_{\omega_{T,\ell}\setminus\omega_{T,\ell-7}},$$

where $C_1 := \frac{1}{2} + \frac{3}{2} C_{\mathcal{I}_H} C'_{\mathcal{I}_H} + (C'_{\mathcal{I}_H} C_{\mathcal{I}_H} + 1) C_{\mathcal{I}_H} \sqrt{C_{\rm ol}}\gamma$ depends only on the shape regularity of the coarse mesh $\mathcal{T}_H$. Since

$$\|\nabla\psi\|^2_{\omega_{T,\ell}\setminus\omega_{T,\ell-7}} = \|\nabla\psi\|^2_{\Omega\setminus\omega_{T,\ell-7}} - \|\nabla\psi\|^2_{\Omega\setminus\omega_{T,\ell}},$$

this implies the contraction

$$\|\nabla\psi\|^2_{\Omega\setminus\omega_{T,\ell}} \leq \frac{C_1}{C_1+\frac{1}{2}} \|\nabla\psi\|^2_{\Omega\setminus\omega_{T,\ell-7}}.$$

Hence,

$$\|\nabla\psi\|^2_{\Omega\setminus\omega_{T,\ell}} \leq \left(\frac{C_1}{C_1+\frac{1}{2}}\right)^{\left\lfloor\frac{\ell}{7}\right\rfloor} \|\nabla\psi\|^2_{\Omega} \leq C_{\mathcal{C}} \left(\frac{C_1}{C_1+\frac{1}{2}}\right)^{\left\lfloor\frac{\ell}{7}\right\rfloor} \|\nabla v\|^2_T,$$

and some algebraic manipulations yield the assertion. □

5. Localized approximation

This section localizes the corrector problems from Ω to subdomains of size ℓH; ℓ being a novel discretization parameter - the oversampling parameter.

5.1. **Localized correctors.** The exponential decay of the element correctors (cf. Theorem 4.6) motivates their localized approximation on element patches. Given such a patch $\omega_{T,\ell}$ for some $T \in \mathcal{T}_H$ and $\ell \in \mathbb{N}$ define the localized remainder space

$$R_H(\omega_{T,\ell}) := \{w \in R_H \mid w|_{\Omega\setminus\omega_{T,\ell}} = 0\} \tag{5.1}$$

and the localized sesquilinear form

$$a_{\omega_{T,\ell}}(u,v) := (\nabla u, \nabla v)_{\omega_{T,\ell}} - \kappa^2 (u,v)_{\omega_{T,\ell}} - i\kappa (u,v)_{\Gamma_R\cap\partial\omega_{T,\ell}}, \tag{5.2}$$

Then, given some finite element function $v_H \in V_H$, its localized primal correction $\mathcal{C}_\ell v_H$ is defined via localized element correctors in the following way:

$$\mathcal{C}_\ell v_H := \sum_{T\in\mathcal{T}_H} \mathcal{C}_{T,\ell}(v_H|_T), \tag{5.3}$$

where $\mathcal{C}_{T,\ell}(v_H|_T) \in R_H(\omega_{T,\ell})$ solves

$$a_{\omega_{T,\ell}}(\mathcal{C}_T(v_H|_T), w) \;=\; a_T(v_H,w) \;:=\; (\nabla v_H, \nabla w)_T - \kappa^2 (v_H,w)_T - i\kappa (v_H,w)_{\Gamma_R\cap\partial T} \tag{5.4}$$

for all $w \in R_H(\omega_{T,\ell})$. The localized dual correction is $\mathcal{C}^*_\ell v_H := \overline{\mathcal{C}_\ell \bar v_H}$. Note that (5.4) is truly localized insofar as the linear constraints $(w,\phi_z)_\Omega = 0$ $(z \in \mathcal{N}_H)$ that characterize an element $w \in R_H$ need to be checked only for $z \in \mathcal{N}_H \cap \omega_{T,\ell}$ and are satisfied automatically for all other nodes if $w \in R_H(\omega_{T,\ell})$ is in the localized fine space.

Though being localized, the correctors $\mathcal{C}_{T,\ell}$ and $\mathcal{C}^*_{T,\ell}$ are still somewhat ideal because their evaluation requires the solution of an infinite-dimensional variational problem in the space $R_H(\omega_{T,\ell})$. Moreover, $\mathcal{C}^*_\ell = \mathcal{C}_\ell$ whenever $\omega_{T,\ell} \cap \Gamma_R = \emptyset$. If the mesh is (locally) structured so that two patches and are equal up to translation or rotation with the same local triangulation, then also the correctors will coincide up to shift and rotation. This means that on a uniform mesh only a finite number of the interior cell problems need to be solved plus a number of cell problems that capture all possible intersections of the patches and the boundary parts. On polyhedral domains, this number will scale like the oversampling parameter ℓ times the number of boundary faces of the domain. To be fully practical, we will also have to discretize the local corrector problems (5.4). This step and the analysis of corresponding errors will be discussed Section 6.

An error bound for the localized approximation of the corrector $\mathcal{C}$ and its adjoint $\mathcal{C}^*$ is easily derived from the exponential decay property of Theorem 4.6.

Lemma 5.1 (local approximation of element correctors)**.** *If the resolution condition of Assumption 4.1 is satisfied, then, for any $T \in \mathcal{T}_H$ and any $\ell \in \mathbb{N}$, it holds that*

$$\|\nabla(\mathcal{C}_{T,\infty}v - \mathcal{C}_{T,\ell}v)\|_\Omega \leq C'_{\mathrm{dec}}\beta^\ell \|\nabla v\|_T,$$

where $\beta < 1$ is the constant from Theorem 4.6 and
$C'_{\mathrm{dec}} := \left(6C_a^2(1 + C_{\mathcal{I}_H}^2 C_{\mathcal{I}_H}'^2)\left(\frac{3}{2} + C_{\mathcal{I}_H}^2 C_{\mathrm{ol}}\gamma^2\right)\right)^{1/2} C_{\mathrm{dec}}\beta^{-6}$.

Proof. Define the cut-off function η (depending on T and ℓ)

$$\eta(x) := \frac{\operatorname{dist}(x, \Omega \setminus \omega_{T,\ell-2})}{\operatorname{dist}(x, \omega_{T,\ell-3}) + \operatorname{dist}(x, \Omega \setminus \omega_{T,\ell-2})}. \tag{5.5}$$

Note that $\eta = 1$ in $\omega_{T,\ell-3}$ and $\eta = 0$ outside $\omega_{T,\ell-2}$. Moreover, η is bounded between 0 and 1 and satisfies the Lipschitz bound (4.27). Since $\mathcal{C}_{T,\ell}v$ is the Galerkin approximation of $\mathcal{C}_{T,\infty}v$ and $\eta\mathcal{C}_{T,\infty}v - \mathcal{I}_H^{-1,\mathrm{loc}}(\mathcal{I}_H(\eta\mathcal{C}_{T,\infty}v)) \in R_H(\omega_{T,\ell})$, Céa's lemma plus Lemma 4.2, the definition of $\mathcal{C}_{T,\infty}$ (4.23), the Cauchy-Schwarz inequality, the approximation property (3.6) of the interpolation operator $\mathcal{I}_H$, the shape regularity of the mesh (3.1) (cf. (4.27)) and the resolution condition Assumption 4.1 imply

$$\begin{aligned}
\|\nabla(\mathcal{C}_{T,\infty}v - \mathcal{C}_{T,\ell}v)\|_\Omega^2 &\leq 3C_a^2\|\mathcal{C}_{T,\infty}v - (\eta\mathcal{C}_{T,\infty}v - \mathcal{I}_H^{-1,\mathrm{loc}}(\mathcal{I}_H(\eta\mathcal{C}_{T,\infty}v))\|_V^2 \\
&\leq 6C_a^2\left(\|\nabla((1-\eta)\mathcal{C}_{T,\infty}v)\|_\Omega^2 + \kappa^2\|(1-\eta)\mathcal{C}_{T,\infty}v\|_\Omega^2\right) \\
&\quad + 6C_a^2 C_{\mathcal{I}_H}^2 C_{\mathcal{I}_H}'^2\left(\|\nabla(\eta\mathcal{C}_{T,\infty}v)\|_{\omega_{T,\ell}\setminus\omega_{T,\ell-5}}^2 + \kappa^2\|\eta\mathcal{C}_{T,\infty}v\|_{\omega_{T,\ell}\setminus\omega_{T,\ell-5}}^2\right) \\
&\leq 6C_a^2(1 + C_{\mathcal{I}_H}^2 C_{\mathcal{I}_H}'^2)\Big(\|\nabla\mathcal{C}_{T,\infty}v\|_{\Omega\setminus\omega_{\ell-5}}^2 \\
&\quad + C_{\mathcal{I}_H}^2 C_{\mathrm{ol}}\|H\nabla\eta\|_{L^\infty(\Omega)}^2\|\nabla\mathcal{C}_{T,\infty}v\|_{\Omega\setminus\omega_{\ell-6}}^2 + C_{\mathcal{I}_H}^2 C_{\mathrm{ol}}(H\kappa)^2\|\nabla\mathcal{C}_{T,\infty}v\|_{\Omega\setminus\omega_{\ell-6}}^2\Big) \\
&\leq 6C_a^2(1 + C_{\mathcal{I}_H}^2 C_{\mathcal{I}_H}'^2)\left(\tfrac{3}{2} + C_{\mathcal{I}_H}^2 C_{\mathrm{ol}}\gamma^2\right)\|\nabla\mathcal{C}_{T,\infty}v\|_{\Omega\setminus\omega_{\ell-6}}^2.
\end{aligned}$$

This and Theorem 4.6 readily imply the assertion. □

Theorem 5.2 (error of the localized corrections)**.** *If the resolution condition of Assumption 4.1 is satisfied, then, for any $\ell \in \mathbb{N}$, it holds that*

$$\|\nabla(\mathcal{C}_\infty v - \mathcal{C}_\ell v)\|_\Omega \leq C_{\mathrm{loc},\ell}\beta^\ell\|\nabla v\|_\Omega,$$

where $C_{\mathrm{loc},\ell} := 3\sqrt{3}C_{\mathrm{ol},\ell+5}C_a^2(1 + C'_{\mathcal{I}_H}C_{\mathcal{I}_H})(\frac{3}{2} + C_{\mathcal{I}_H}^2 C_{\mathrm{ol}}\gamma^2)C'_{\mathrm{dec}}$.

Proof. Set $z := \mathcal{C}_\infty v - \mathcal{C}_\ell v$ and, for any $T \in \mathcal{T}_H$, set $z_T := \mathcal{C}_{T,\infty}v - \mathcal{C}_{T,\ell}v$. The R_H-ellipticity of the sesquilinear form (4.8) implies that

$$\tfrac{1}{3}\|\nabla z\|_\Omega^2 \leq \sum_{T\in\mathcal{T}_H} a(z_T, z). \tag{5.6}$$

Given some $T \in \mathcal{T}_H$, let η be the cutoff function defined by

$$\eta(x) := \frac{\operatorname{dist}(x, \omega_{T,\ell+2})}{\operatorname{dist}(x, \omega_{T,\ell+2}) + \operatorname{dist}(x, \Omega \setminus \omega_{T,\ell+3})},$$

that is $\eta = 0$ in $\omega_{T,\ell+2}$ and $\eta = 1$ outside $\omega_{T,\ell+3}$. Moreover, η is bounded between 0 and 1 and satisfies the Lipschitz bound (4.27). Since $\operatorname{supp}\mathcal{I}_H^{-1,\mathrm{loc}}(\mathcal{I}_H(\eta z)) \subset \Omega \setminus \omega_{T,\ell}$ and $\eta z - \mathcal{I}_H^{-1,\mathrm{loc}}(\mathcal{I}_H(\eta z)) \in R_H$, we have that

$$a(z_T, \eta z - \mathcal{I}_H^{-1,\mathrm{loc}}(\mathcal{I}_H(\eta z))) = a(\mathcal{C}_{T,\infty}v, \eta z - \mathcal{I}_H^{-1,\mathrm{loc}}(\mathcal{I}_H(\eta z))) = 0.$$

Hence,

$$a(z_T, z) = a(z_T, \mathcal{I}_H^{-1,\mathrm{loc}}(\mathcal{I}_H(\eta z)) + a(z_T, (1-\eta)z).$$

The properties (3.6) of the interpolation operator $\mathcal{I}_H$ and the Lipschitz bound (4.27) lead to upper bounds

$$a(z_T, z) \leq C_a(1 + C'_{\mathcal{I}_H}C_{\mathcal{I}_H})\sqrt{1 + C_{\mathcal{I}_H}^2 C_{\mathrm{ol}}\gamma^2}\|z\|_{V,\omega_{T,\ell+5}}\|z_T\|_V. \tag{5.7}$$

The combination of (5.6) and (5.7) plus a discrete Cauchy-Schwarz inequality and the bounded overlap (3.3) of the element patches leads to

$$\|\nabla z\|_\Omega \leq 2C_{\mathrm{ol},\ell+3}C_a(1+C'_{\mathcal{I}_H}C_{\mathcal{I}_H})\sqrt{1+C^2_{\mathcal{I}_H}C_{\mathrm{ol}}\gamma^2}\left(\sum_{T\in\mathcal{T}_H}\|z_T\|^2_V\right)^{1/2}. \tag{5.8}$$

This and Lemma 5.1 readily yield the assertion. □

5.2. **Localized trial and test spaces.** The localized trial space $V_{H,\ell}\subset V$ is simply defined as the image of the classical finite element space V_H under the operator $1-\mathcal{C}_\ell$,

$$V_{H,\ell} := (1-\mathcal{C}_\ell)V_H \tag{5.9}$$

and the localized test space $V^*_{H,\ell}\subset V$ reads

$$V^*_{H,\ell} := (1-\mathcal{C}^*_\ell)V_H \tag{5.10}$$

Note that both $V_{H,\ell}$ and $V^*_{H,\ell}$ are finite-dimensional with a local basis,

$$V_{H,\ell} = \mathrm{span}\{(1-\mathcal{C}_\ell)\phi_z \,|\, z\in\mathcal{N}_H\} \text{ and } V^*_{H,\ell} = \mathrm{span}\{\overline{(1-\mathcal{C}_\ell)\phi_z} \,|\, z\in\mathcal{N}_H\},$$

where ϕ_z is the (real-valued) nodal basis of V_H (cf. Section 3.3).

The Petrov-Galerkin method with respect to the trial space $V_{H,\ell}$ and the test space $V^*_{H,\ell}$ seeks $u_{H,\ell}\in V_{H,\ell}$ such that, for all $v_{H,\ell}\in V^*_{H,\ell}$,

$$a(u_{H,\ell},v_{H,\ell}) = (f,v_{H,\ell})_\Omega. \tag{5.11}$$

5.3. **Stability of the localized method.** The stability of the localized methods requires the coupling of the oversampling parameter to the stability constant which we will now assume to be polynomial with respect to the wave number.

Assumption 5.3 (polynomial-in-κ-stability and logarithmic oversampling condition)**.** *there are constants $C_{\mathrm{pst}}>0$ and $q_{\mathrm{pst}}\geq 0$ and a $\kappa_0>0$ that may depend on Ω and the partition of the boundary into Γ_D, Γ_N and Γ_R such that, for any $\kappa\geq\kappa_0$, the stability constant $C_{\mathrm{st}}(\kappa)$ of* (2.6) *satisfies* (2.7)*,*

$$C_{\mathrm{st}}(\kappa)\leq C_{\mathrm{pst}}\kappa^{q_{\mathrm{pst}}}.$$

Given the wave number κ and the constants $C_{\mathcal{I}_H}$ from (3.6) *and C_{ol} from* (3.3)*, we assume that the oversampling parameter ℓ satisfies*

$$\ell \geq \frac{(q_{\mathrm{pst}}+1)\log\kappa + \log\left(4C_{\mathcal{C}}C_{\mathrm{pst}}C_{\Pi_H}\sqrt{\tfrac{3}{2}}C_{\mathrm{loc},\ell}C_a\right)}{|\log\beta|}. \tag{5.12}$$

Since the constant $C_{\mathrm{loc},\ell}$ grows at most polynomially with ℓ (cf. (3.3)), condition (5.12) is indeed satisfiable and the proper choice of ℓ will be dominated by the logarithm $\log\kappa$ of the wave number.

The stability of the localized method follows from the fact that the ideal pairing $(V_{H,\infty},V^*_{H,\infty})$ is stable and that $(V_{H,\ell},V^*_{H,\ell})$ is exponentially close.

Theorem 5.4 (stability of the localized method)**.** *If the mesh width H is sufficiently small in the sense of Assumption 4.1 ($H\kappa\lesssim 1$) and if the oversampling parameter $\ell\in\mathbb{N}$ is sufficiently large in the sense of Assumption 5.3 ($\ell\gtrsim\log\kappa$), then the pairing of the localized spaces $V_{H,\ell}$ and $V^*_{H,\ell}$ satisfies the discrete inf-sup condition*

$$\inf_{u_{H,\ell}\in V_{H,\ell}\setminus\{0\}}\sup_{v_{H,\ell}\in V^*_{H,\ell}\setminus\{0\}}\frac{\Re a(u_{H,\ell},v_{H,\ell})}{\|u_{H,\ell}\|_V\|v_{H,\ell}\|_V}\geq\frac{1}{4C_{\mathcal{C}}C_{\mathrm{pst}}\kappa^{q_{\mathrm{pst}}+1}}. \tag{5.13}$$

This ensures that, for any $f\in V'$, there exists a unique solution of the discrete problem (5.11)*.*

Proof. Let $u_{H,\ell} \in V_{H,\ell}$ and set $u_{H,\infty} := (1-\mathcal{C})\Pi_H u_{H,\ell}$. Under the polynomial-in-κ stability of Assumption 5.3, Theorem 4.4 guarantees the existence of some $v_{H,\infty} \in V^*_{H,\infty}$ with

$$\Re a(u_{H,\infty}, v_{H,\infty}) \geq \frac{1}{2C_{\mathrm{pst}}C_{\mathcal{C}}\kappa^{q_{\mathrm{pst}}+1}}\|u_{H,\infty}\|_V\|v_{H,\infty}\|_V. \tag{5.14}$$

Set $v_{H,\ell} := (1-\mathcal{C}^*_\ell)\Pi_H v_{H,\infty} \in V^*_{H,\ell}$ and observe that (4.14) yields

$$\begin{aligned}\Re a(u_{H,\ell}, v_{H,\ell}) &= \Re a(u_{H,\ell}, v_{H,\ell} - v_{H,\infty}) + a(u_{H,\ell}, v_{H,\infty}) \\ &= \Re a(u_{H,\ell}, (\mathcal{C}^* - \mathcal{C}^*_\ell)\Pi_H v_{H,\ell}) + \Re a(u_{H,\infty}, v_{H,\infty}).\end{aligned}$$

Hence,

$$\begin{aligned}\Re a(u_{H,\ell}, v_{H,\ell}) &\geq \Re a(u_{H,\infty}, v_{H,\infty}) - C_a\|u_{H,\ell}\|_V\|(\mathcal{C}^* - \mathcal{C}^*_\ell)\Pi_H v_{H,\ell}\|_V \\ &\geq \Re a(u_{H,\infty}, v_{H,\infty}) - C_{\Pi_H}\sqrt{\tfrac{3}{2}}C_{\mathrm{loc},\ell}C_a\beta^\ell\|u_{H,\ell}\|_V\|v_{H,\ell}\|_V,\end{aligned} \tag{5.15}$$

where we have used (4.7), Theorem 5.2, and (3.8). This yields

$$\begin{aligned}\Re a(u_{H,\ell}, v_{H,\ell}) &\geq \frac{1}{C_{\mathcal{C}}C_{\mathrm{st}}\kappa^{q_{\mathrm{pst}}+1}}\|u_{H,\infty}\|_V\|v_{H,\infty}\|_V - C'\beta^\ell\|u_{H,\ell}\|_V\|v_{H,\ell}\|_V \\ &\geq \left(\frac{1}{2C_{\mathcal{C}}C_{\mathrm{pst}}\kappa^{q_{\mathrm{pst}}+1}} - C_{\Pi_H}\sqrt{\tfrac{3}{2}}C_{\mathrm{loc},\ell}C_a\beta^\ell\right)\|u_{H,\ell}\|_V\|v_{H,\ell}\|_V,\end{aligned} \tag{5.16}$$

and Assumption 5.3 readily implies the assertion. □

Theorem 5.5 (error of the localized method)**.** *If the mesh width H is sufficiently small in the sense of Assumption 4.1 ($H\kappa \lesssim 1$) and if the oversampling parameter $\ell \in \mathbb{N}$ is sufficiently large in the sense of Assumption 5.3 ($\ell \gtrsim \log\kappa$), then the localized Petrov-Galerkin approximation $u_{H,\ell} \in V_{H,\ell}$ satisfies the error estimate*

$$\|u - u_{H,\ell}\|_V \leq 6\sqrt{C_{\mathrm{ol}}}C_{\mathcal{I}_H}\|Hf\|_\Omega + 6C_aC_{\mathrm{loc},\ell}\beta^\ell C_{\Pi_H}C_{\mathrm{pst}}\kappa^{q_{\mathrm{pst}}}\|f\|_\Omega. \tag{5.17}$$

Proof. The proof is inspired by standard techniques for Galerkin methods (see [Sch74], [BS08, Thm. 5.7.6], [Sau06], [BS07]). Set $e := u - u_{H,\ell}$ and $e_{H,\ell} := (1-\mathcal{C}_\ell)\Pi_H e \in V_{H,\ell}$. The triangle inequality yields

$$\|e\|_V \leq \|e - e_{H,\ell}\|_V + \|e_{H,\ell}\|_V. \tag{5.18}$$

An Aubin-Nitsche duality argument shows that $\|e_{H,\ell}\|_V$ is controlled by some multiple of $\|e - e_{H,\ell}\|_V$. Let $z_{H,\ell} \in V^*_{H,\ell}$ be the unique solution of the discrete adjoint variational problem

$$(\nabla v_{H,\ell}, \nabla e_{H,\ell}) + \kappa^2(v_{H,\ell}, e_{H,\ell}) = a(v_{H,\ell}, z_{H,\ell}),$$

for all $v_{H,\ell} \in V_{H,\ell}$. Set $z_{H,\infty} := (1-\mathcal{C}^*)\Pi_H z_{H,\ell}$ and observe that

$$\begin{aligned}\|e_{H,\ell}\|_V^2 &= a(e_{H,\ell}, z_{H,\ell} - z_{H,\infty}) + a(e_{H,\ell}, z_{H,\infty}) \\ &= a(e_{H,\ell}, z_{H,\ell} - z_{H,\infty}) + a(e, z_{H,\infty}) \\ &= a(e_{H,\ell}, z_{H,\ell} - z_{H,\infty}) + a(e, z_{H,\infty} - z_{H,\ell}) \\ &= a(e - e_{H,\ell}, (\mathcal{C}^* - \mathcal{C}^*_\ell)\Pi_H z_{H,\ell}) \\ &\leq C_a\|e - e_{H,\ell}\|_V\|(\mathcal{C}^* - \mathcal{C}^*_\ell)\Pi_H z_{H,\ell}\|_V.\end{aligned}$$

Under Assumption (2.7), Theorem 5.2, Theorem 5.4 and (3.8) readily yield

$$\|e_{H,\ell}\|_V^2 \leq C_a^2C_{\mathrm{loc},\ell}\beta^\ell C_{\Pi_H}C_{\mathrm{pst}}\kappa^{q_{\mathrm{pst}}+1}\|e - e_{H,\ell}\|_V^2. \tag{5.19}$$

This, (5.18) and Assumption 5.3 show that

$$\|e\|_V \leq 2\|e - e_{H,\ell}\|_V. \tag{5.20}$$

Since $e - e_{H,\ell} \in R_H$, the R_H-ellipticity (4.8) yields

$$\|e - e_{H,\ell}\|_V^2 \leq 3\Re a(e - e_{H,\ell}, e - e_{H,\ell}). \tag{5.21}$$

The relation (4.13) then yields

$$a(e-e_{H,\ell},e-e_{H,\ell}) = a(u,e-e_{H,\ell}) + a((\mathcal{C}-\mathcal{C}_\ell)\Pi_H u, e-e_{H,\ell}) \\ \leq |(f,e-e_{H,\ell})_\Omega| + C_a\|(\mathcal{C}-\mathcal{C}_\ell)\Pi_H u\|_V \|e-e_{H,\ell}\|_V. \tag{5.22}$$

This, Cauchy inequalities, interpolation error estimates (3.6), Theorem 5.2 and the stability estimate (2.7) readily yield the bound

$$\|e-e_{H,\ell}\|_V \leq 3\sqrt{C_{\rm ol}}C_{\mathcal{I}_H}\|Hf\|_\Omega + 3C_aC_{{\rm loc},\ell}\beta^\ell C_{\Pi_H}C_{\rm st}\kappa^{q_{\rm pst}}\|f\|_\Omega. \tag{5.23}$$

The combination of (5.20) and (5.23) is the assertion. □

6. Fully discrete localized approximation

As already mentioned before, the localized corrector problems (4.23) are variational problems in infinite-dimensional spaces $R_H(\omega_{T,\ell})$ that require further discretization. For the ease of presentation we restrict ourselves in this paper to the classical case of piecewise affine conforming elements on simplicial meshes but we emphasize that the technique easily transfers to more general situations and can be applied to a large variety of discretization schemes.

So far, the presentation of the method was optimized with respect to theoretical aspects of the stability and error analysis. Here, we will present the method in a slightly more practical fashion.

6.1. **The fully discrete method.** For any $T\in\mathcal{T}_H$, choose an oversampling parameter ℓ_T (sufficiently large so that there is a chance that Assumption 5.3 is satisfied). Let $\mathcal{T}_h(\omega_{T,\ell})$ be a regular (and possibly adaptive) mesh of width $h\leq H$ and consider the standard finite element space $V_h(\omega_{T,\ell})\subset \mathrm{H}^1_0(\omega_{T,\ell})$ (cf. Section 3.3). For any vertex y of T, compute the element correctors $\mathcal{C}_{T,\ell,h}\phi_z\in R_H(\omega_{T,\ell})\cap V_h(\omega_{T,\ell})$ as the unique solution of the discrete cell problem

$$a(\mathcal{C}_{T,\ell,h}\phi_y, w) = a_T(\phi_y,w), \quad \text{for all } w\in R_H(\omega_{T,\ell})\cap V_h(\omega_{T,\ell}).$$

Now for every global vertex $z\in\mathcal{N}_H$, compute the correctors $\mathcal{C}_{\ell,h}\phi_z$ by

$$\mathcal{C}_{\ell,h}\phi_z = \sum_{T\in\mathcal{T}_H : z \text{ vertex of } T} \mathcal{C}_{T,\ell,h}\phi_z.$$

This leads to modified basis functions $\tilde\phi_z := \phi_z - \mathcal{C}_{\ell,h}\phi_z$ that span a discrete space

$$V_{H,\ell,h} := \operatorname{span}\{\tilde\phi_z \mid z\in\mathcal{N}_H\} \tag{6.1}$$

of the same dimension N_H as the classical finite element space V_H.

In this most general setting, the discretization of the cell problems is completely independent. In the error analysis below, however, we will restrict ourselves to the case where cell problems are synchronized in the sense that we assume there is an underlying global fine mesh $\mathcal{T}_h$ that is a regular refinement of the coarse mesh $\mathcal{T}_H$ and that local meshes $\mathcal{T}_h(\omega_{T,\ell})$ coincide with $\mathcal{T}_h$ on the patches.

The fully discrete localized Petrov-Galerkin method with respect to the trial space $V_{H,\ell,h}$ and the test space $V^*_{H,\ell,h}$ seeks $u_{H,\ell,h}\in V_{H,\ell,h}$ such that, for all $v_{H,\ell,h}\in V^*_{H,\ell,h}$,

$$a(u_{H,\ell,h}, v_{H,\ell,h}) = (f, v_{H,\ell,h})_\Omega. \tag{6.2}$$

6.2. **Error analysis of the fully discrete method.** An abstract a priori error analysis of the general method would follow the analysis of Section 5 and trace the error of the additional perturbation depending on the local choice of the approximation space. However, this will require the estimation of the error $\mathcal{C}-\mathcal{C}_{\ell,h}$ or $\mathcal{C}_\ell-\mathcal{C}_{\ell,h}$, which appears to be non-trivial and requires, for instance, regularity results for the ideal correctors. This line will be followed in future research along with an a posteriori analysis of the method.

For this paper, we will restrict ourselves to the case of synchronized cell problems in the sense that there is an underlying global fine mesh $\mathcal{T}_h$ that is a regular refinement of the coarse mesh $\mathcal{T}_H$. This implies that the global fine space V_h contains standard finite element functions (cf. Section 3.3) and the spaces for the local cell problems are derived by restriction

of V_h to the patch. In this regime, the method in fact approximates u_h, where $u_h \in V_h$ is the Galerkin approximation in the global fine scale, that is,

$$a(u_h, v_h) = (f, v_h)_\Omega, \quad \text{for all } v_h \in V_h. \tag{6.3}$$

In the remaining part of this paper, we will refer to u_h as the reference solution. It is clear that if h is sufficiently small, then the problem (6.3) is well-posed and stable. However, h underlies the typical resolution assumptions of finite element methods, for instance $h\kappa^{3/2} \lesssim 1$ for a pure Robin problem in a convex domain discretized by P_1 finite elements [Wu14].

Assumption 6.1 (well-posedness of reference problem)**.** *Assume that V_h is chosen such that, for any $f \in V'$, the reference problem* (6.3) *admits a unique solution and there is a constant C_h that may depend on the partition of the boundary into Γ_D, Γ_N and Γ_R such that*

$$\|u_h\|_V \leq C_h \|u\|_V,$$

where u denotes the solution of (2.3)*.*

Theorem 6.2 (stability and error of the fully discrete method)**.** *If the fine scale discretization space V_h is sufficiently rich so that Assumption 6.1 holds and if the coarse mesh width H is sufficiently small in the sense of Assumption 4.1 ($H\kappa \lesssim 1$) and if the oversampling parameter $\ell \in \mathbb{N}$ is sufficiently large in the sense of Assumption 5.3 ($\ell \gtrsim \log \kappa$), then the fully discrete localized Petrov-Galerkin $u_{H,\ell,h} \in V_{H,\ell,h}$ approximation satisfies the error estimate*

$$\|u_h - u_{H,\ell,h}\|_V \leq C(H + C_{\mathrm{loc},\ell}\beta^\ell \kappa^{q_{\mathrm{pst}}})\|f\|_\Omega, \tag{6.4}$$

where u_h solves the reference problem (6.3) *and C is some generic constant that does not depend on H, ℓ and κ.*

Proof. The proof follows closely the analysis of Section 5 and simply replaces the space V by V_h in the construction of the method and its error analysis. Almost all arguments remain valid. The only technical issue is that the space V_h is not closed under multiplication by cut-off functions used in the proofs of Theorem 4.6, Lemma 5.1, and Theorem 5.2. This requires minor modifications as they have already been applied successfully in previous papers [MP14b, HP13, HMP14a]. To begin with, let all cut-off functions η be replaced by their nodal interpolation $\mathcal{I}_H^{\mathrm{nodal}}\eta$ on the coarse mesh $\mathcal{T}_H$. This may affect the constant in (4.27) but not the overall results. This choice shows that $\eta\psi$ is piecewise polynomial with respect to the fine mesh $\mathcal{T}_h$ and can be approximated by nodal interpolation $\mathcal{I}_h^{\mathrm{nodal}}(\eta\psi)$ on the same mesh in a stable way. One example where such a modification is required is (4.28) in the proof of Theorem 4.6. This causes an additional term that measures the distance of $\eta\psi$ to the finite element space,

$$\begin{aligned}
\|\nabla\psi\|^2_{\Omega\setminus\omega_{T,\ell-3}} &= \Re\left(\nabla\psi, \nabla\left(\eta\psi\right)\right)_\Omega - \Re\left(\nabla\psi, \psi\nabla\eta\right)_\Omega \\
&\leq |\Re\left(\nabla\psi, \nabla\left(\mathcal{I}_h^{\mathrm{nodal}}(\eta\psi) - \mathcal{I}_H^{-1,\mathrm{loc}}(\mathcal{I}_H(\mathcal{I}_h^{\mathrm{nodal}}(\eta\psi)))\right)\right)_\Omega| \\
&\quad + |\Re\left(\nabla\psi, \nabla\mathcal{I}_H^{-1,\mathrm{loc}}(\mathcal{I}_H(\mathcal{I}_h^{\mathrm{nodal}}(\eta\psi)))\right)_\Omega| + |\Re\left(\nabla\psi, \psi\nabla\eta\right)_\Omega| \\
&\quad + |\Re\left(\nabla\psi, \nabla\left(\eta\psi - \mathcal{I}_h^{\mathrm{nodal}}(\eta\psi)\right)\right)_\Omega| \\
&=: \tilde{M}_1 + \tilde{M}_2 + \tilde{M}_3 + \tilde{M}_4.
\end{aligned} \tag{6.5}$$

The treatment of $\tilde{M}_1, \tilde{M}_2, \tilde{M}_3$ is very similar to the treatment of M_1, M_2, M_3 in the proof of Theorem 4.6 and requires only the stability of I_h on the space of piecewise polynomials. Since $I_h(\eta\psi) = \psi$ outside the support of $\nabla\eta$, $\tilde{M}_4$ can easily be bounded by

$$\tilde{M}_4 \leq (1 + C_{\mathcal{I}_H}\sqrt{C_{\mathrm{ol}}})\|H\nabla\eta\|_{L^\infty(\Omega)}\|\nabla\phi\|^2_{\omega_{T,\ell-2}\setminus\omega_{T,\ell-5}}$$

and further arguments remain valid (with a possible change of the constants involved). The proofs of Lemma 5.1 and Theorem 5.2 can be modified in similar way. □

7. Numerical Experiments

In this section we will present two numerical examples. We apply our method to model Helmholtz problems in one and two dimensions and compare the results with standard P_1 finite elements. We will demonstrate the validity of our estimates based on varying oversampling parameter ℓ, coarse mesh size H and by varying the wave number κ. A comprehensive numerical study of the algorithmic ideas proposed in this paper is topic of current and future research.

7.1. **Illustration of the theoretical results in** $1d$**.** Let $\Omega := (0,1)$, $\Gamma_R = \partial\Omega$ (solely Robin boundary condition), and let the right-hand side f defined by

$$f(x) := \begin{cases} 2\sqrt{2}, & x \in [\frac{3}{16}, \frac{5}{16}] \cup [\frac{11}{16}, \frac{13}{16}], \\ 0, & \text{elsewhere}, \end{cases} \tag{7.1}$$

represent two radiating sources. The right-hand side was normalized so that $\|f\|_{L^2(\Omega)} = 1$.

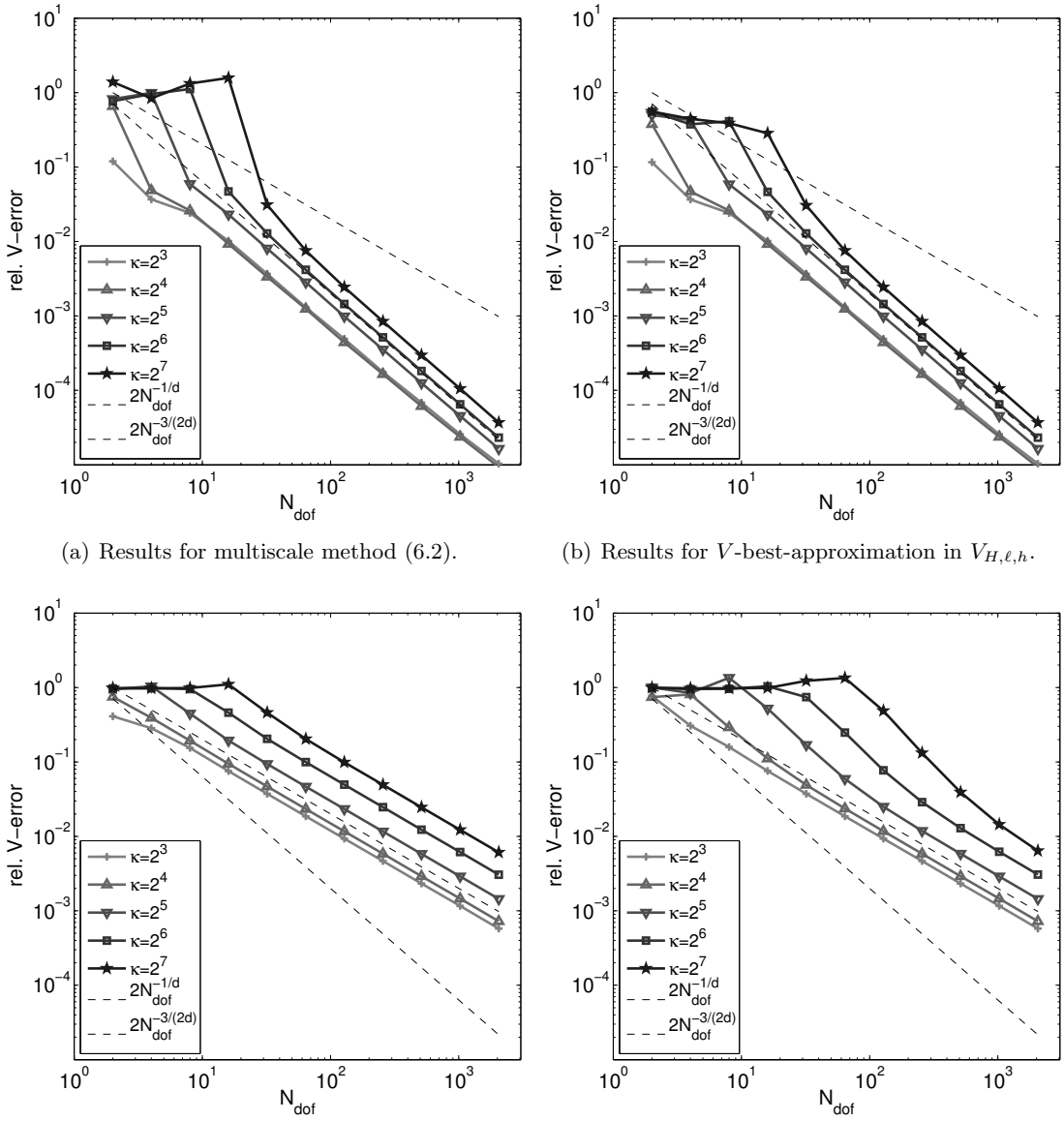
Although this one-dimensional example does not serve as a proper benchmark for the method, it nicely reflects our theoretical results for a wider range of wave numbers. Since non of our arguments depends on the space dimension (though some constants do), the $1d$ performance truly illustrates the performance in higher dimensions. We consider the following values for the wave number, $\kappa = 2^3, 2^4, \dots, 2^7$. The numerical experiment aims to study the dependence between these wave numbers and the accuracy of the numerical method.

Consider the equidistant coarse meshes with mesh widths $H = 2^{-1}, \dots, 2^{-10}$. The reference mesh $\mathcal{T}_h$ is derived by uniform mesh refinement of the coarse meshes and has maximal mesh width $h = 2^{-14}$. The corresponding P_1 conforming finite element approximation on the reference mesh $\mathcal{T}_h$ is denoted by V_h. We consider the reference solution $u_h \in V_h$ of (6.3) with data given in (7.1) and compare it with coarse scale approximations $u_{H,\ell,h} \in V_{H,\ell,h}$ (cf. Definition 6.2) depending on the coarse mesh size H and the oversampling parameter ℓ.

The results are visualized in Figures 1 and 2. Figure 1(a) shows the relative energy errors $\frac{\|u_h - u_{H,\ell,h}\|_V}{\|u_h\|_V}$ depending on the coarse mesh size H for several choices of the wave number $\kappa = 2^3, 2^4, \dots, 2^7$. The oversampling parameter ℓ is tied to H via the relation $\ell = \ell(H) = |\log_2 H|$. This choice seems to be sufficient to preserve optimal convergence as soon as $H\kappa \lesssim 1$ holds. The experimental rate of convergence $N^{\frac{3}{2d}}$ is better than predicted by Theorem 6.2. This effect is due to some unexploited L^2-orthogonality properties of the quasi-interpolation operator $\mathcal{I}_H$; see [Car99, Section 2] and [MP14b, Remark 3.2] for details. In the regime $H\kappa \lesssim 1$, the errors coincide to with those of the best approximation (with respect to the V-norm) of u_h in the space $V_{H,\ell,h}$ depicted in Figure 1(b).

We also show errors of the Petrov-Galerkin method based on the pairing $(V_H, V^*_{H,\ell,h})$ (the localized and fully discretized version of (4.21)) in Figure 1(c). The stabilization via the precomputed test functions cures pollution and the errors are comparable to those of the best possible with standard finite element test functions, whereas the pollution effect is clearly visible for the standard conforming P_1-FEM on the coarse meshes; see Figure 1(d).

Figure 2 aims to illustrate the role of the oversampling parameter. It depicts the relative energy errors $\frac{\|u_h - u_{H,\ell,h}\|_V}{\|u_h\|_V}$ of the method (6.2) and the best-approximation in $V_{H,\ell,h}$ depending on the coarse mesh size H for fixed wave number $\kappa = 2^7$ and several choices of the oversampling parameter $\ell = 1, 2, 3, \dots, 8$. (We also show errors of the standard conforming P_1-FEM on the coarse meshes for comparison.) The exponential decay of the error with respect to ℓ is observed once the mesh size reaches the regime of resolution $H\kappa \lesssim 1$. Moreover, Figure 2(b) shows that, for fixed ℓ, the approximation property of $V_{H,\ell,h}$ does not improve with decreasing H and the oversampling parameter needs to be increased with decreasing H to get any rate. By contrast, the Petrov-Galerkin method based on the trial-test-pairing $(V_H, V_{H,\ell,h})$ (which in fact computes $\Pi_H u$ for $\ell \to \infty$) allows to reduce the oversampling



(a) Results for multiscale method (6.2).

(b) Results for V-best-approximation in $V_{H,\ell,h}$.

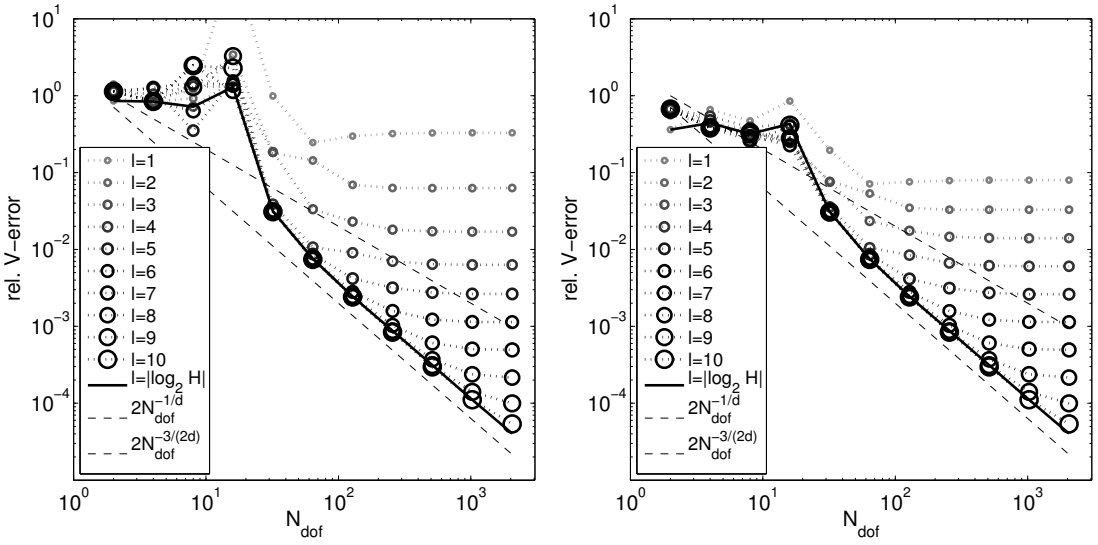
(c) Results for multiscale Petrov-Galerkin method with trial space V_H and test space $V^*_{H,\ell,h}$.

(d) Results for standard Galerkin in the space V_H.

FIGURE 1. Numerical experiment of Section 7.1: Results for the multiscale method (6.2), a modification based on the trial-test-pairing $(V_H, V^*_{H,\ell,h})$ and standard P_1-FEM with several choices of the wave number κ depending on the uniform coarse mesh size $H = N_{\mathrm{dof}}^{-1}$. The reference mesh size $h = 2^{-14}$ remains fixed. The oversampling parameter is tied to the coarse mesh size via the relation $\ell = |\log_2 H|$ in (a)-(c).

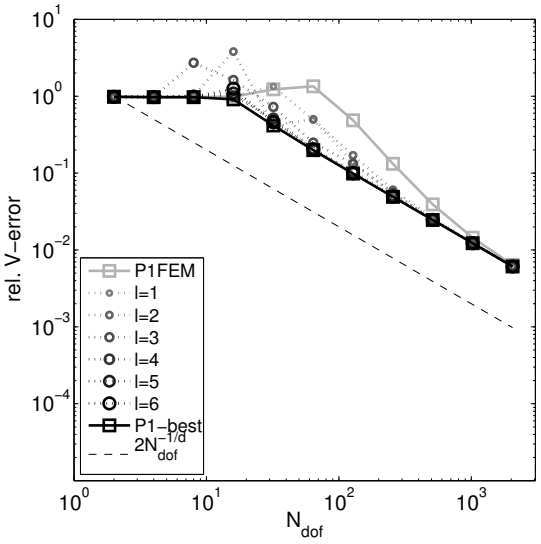
parameter with decreasing $H\kappa$ until, for $H\kappa^2 \approx 1$, the correction can be removed because P_1-FEM becomes quasi-optimal; see Figure 2(c) which depicts relative L^2-errors of the method.

Finally, we want to show that a different choice of interpolation operator in the definition (4.2) of the remainder space can lead to very different practical performance (within the range of the theoretical predictions though). Figure 3 shows the results for the above experiment when the operator $\mathcal{Q}_H$ from (3.10) is used instead of $\mathcal{I}_H$. It turns out that, for this example, the decay of the correctors is much faster so that the same accuracy is reached with more



(a) Results for multiscale method (6.2).

(b) Results for V-best-approximation in $V_{H,\ell,h}$.

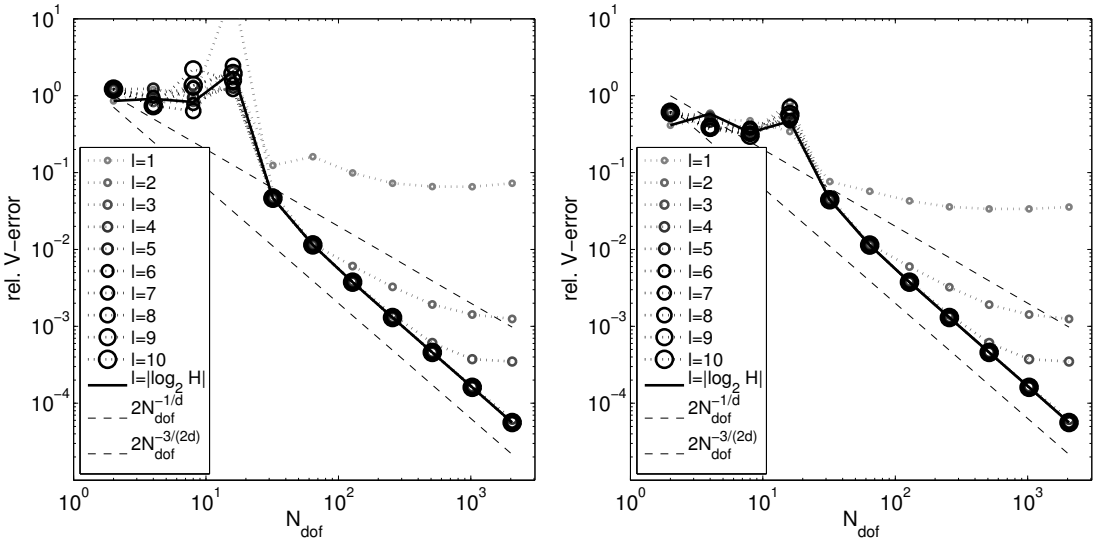


(c) Results for multiscale Petrov-Galerkin method with trial space V_H and test space $V_{H,\ell,h}$.

FIGURE 2. Numerical experiment of Section 7.1: Results for multiscale method (6.2) with wave number $\kappa = 2^8$ depending on the uniform coarse mesh size $H = N_{\mathrm{dof}}^{-1}$. The reference mesh size $h = 2^{-14}$ remains fixed. The oversampling parameter ℓ varies between 1 and 8.

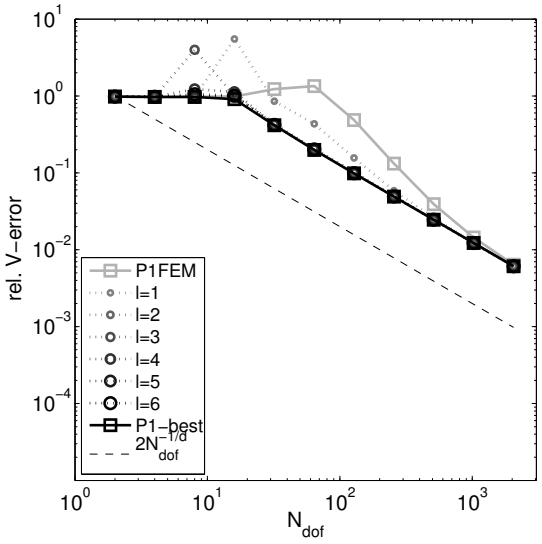
local basis functions. A similar observation has been made previously in the context of high-contrast diffusion problems [BP14, PS14]. This promising performance might be explained by the larger cost of the evaluation of $\mathcal{Q}_H$ that already involves local coarse solves on nodal patches but is not yet well understood; it cannot be explained with the existing theory and requires further investigation.

7.2. **Scattering from a triangle.** The second experiment considers the scattering from sound-soft scatterer occupying the triangle Ω_D. The Sommerfeld radiation condition of the scattered wave is approximated by the Robin boundary condition on the boundary $\Gamma_R := \partial\Omega_R$ of the artificial domain $\Omega_R :=]0,1[^2$ so that $\Omega := \Omega_R \setminus \Omega_D$ is the computational



(a) Results for multiscale method (6.2) based on quasi-interpolation $\mathcal{Q}_H$.

(b) Results for V-best-approximation in $V_{H,\ell,h}$ based on quasi-interpolation $\mathcal{Q}_H$.

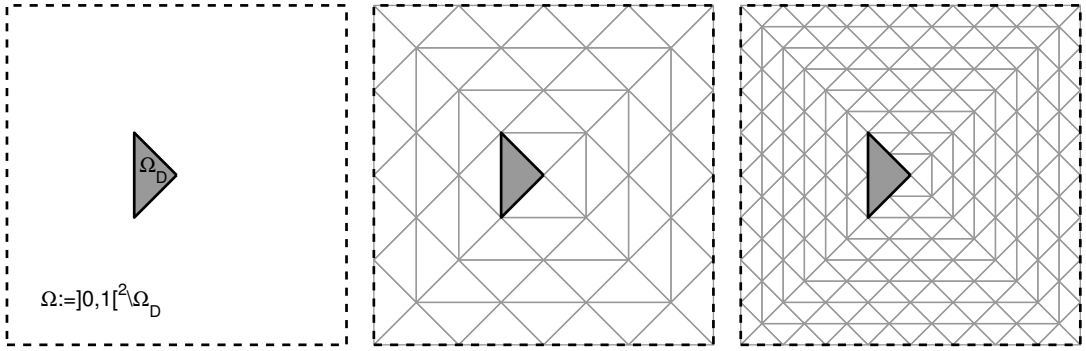


(c) Results for multiscale Petrov-Galerkin method with trial space V_H and test space $V_{H,\ell,h}$ based on quasi-interpolation $\mathcal{Q}_H$.

FIGURE 3. Numerical experiment of Section 7.1: Results for multiscale method (6.2) with interpolation operator $\mathcal{Q}_H$ for wave number $\kappa = 2^8$ depending on the uniform coarse mesh size $H = N_{\mathrm{dof}}^{-1}$. The reference mesh size $h = 2^{-14}$ remains fixed. The oversampling parameter ℓ varies between 1 and 8.

domain; see Figure 4(a). The incident wave $u_{inc}(x) := \exp\left(i\kappa\, x \cdot \begin{pmatrix}\cos(1/2)\\ \sin(1/2)\end{pmatrix}\right)$ is prescribed via an inhomogeneous Dirichlet boundary condition on $\Gamma_D := \partial\Omega_D$ and the scattered wave satisfies the model problem (2.1.a) with the boundary conditions

$$u = -u_{inc} \quad \text{on } \Gamma_D,$$
$$\nabla u \cdot \nu - i\kappa u = 0 \quad \text{on } \Gamma_R.$$



(a) Computational domain Ω with scatterer Ω_D.

(b) Initial coarse mesh.

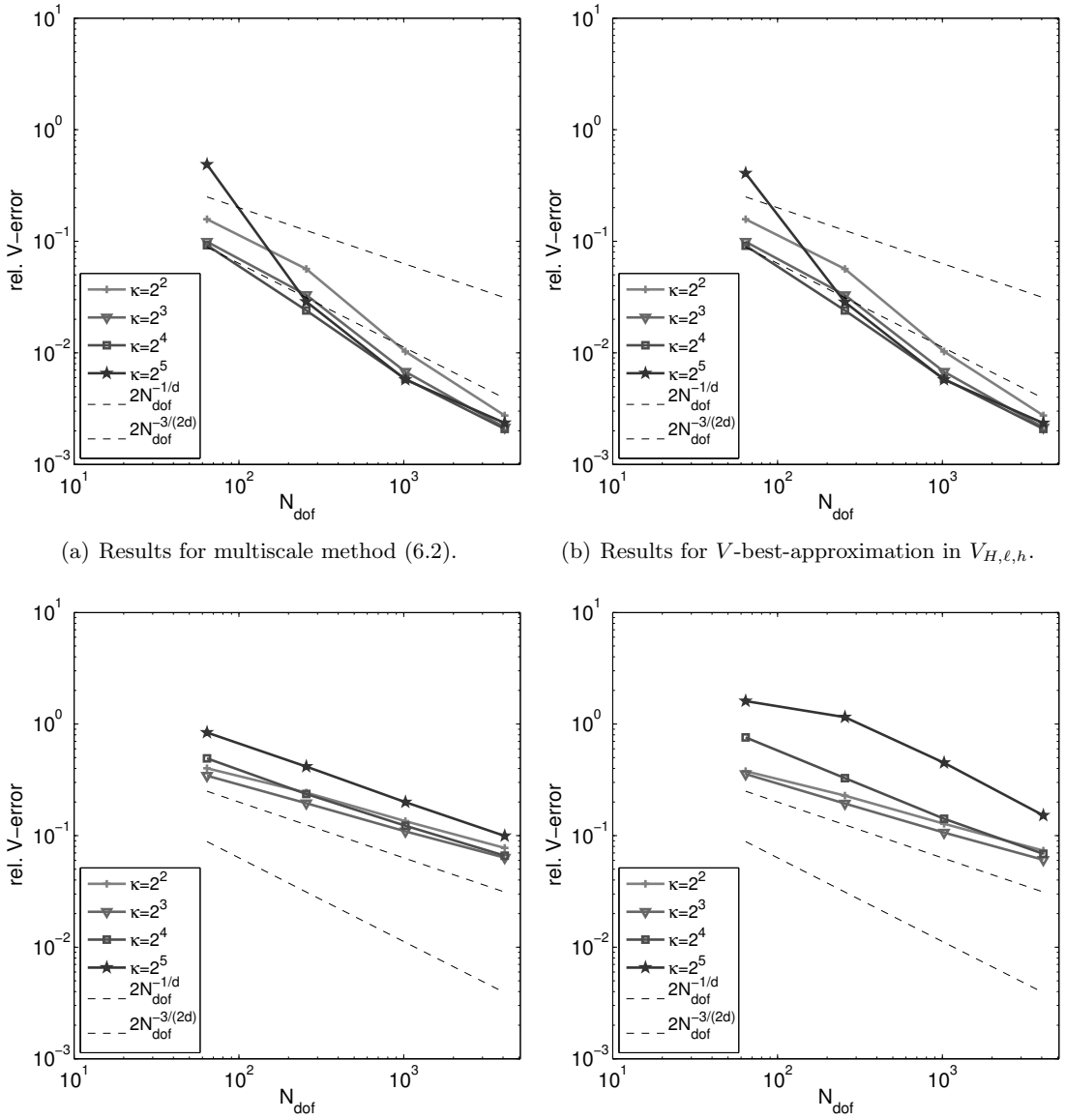
(c) Uniformly refined coarse mesh.

FIGURE 4. Computational domain of the model problem of Section 7.2 and corresponding coarse meshes.

The error analysis of the previous sections extends to this setting in a straight-forward way. By introducing some function $u_0 \in W^{1,2}(\Omega)$ that satisfies the above boundary conditions, the problem can be reformulated with homogenous boundary conditions and the additional term $-a(u_0, v)$ on the right side of (2.3). This corresponds to having the modified right hand side $f + \Delta u_0 + \kappa^2 u_0$ in the strong form (2.1.a) of the problem. If u_0 can be chosen such that $\Delta u_0 \in L^2(\Omega)$, then all error bounds of this paper remain valid. For weaker right hand sides the rates with respect to H are reduced accordingly. Note, however, that the L^2-norm of the modified right-hand side may depend on κ as it is the case in the present experiment where u_0 is related to the incident wave. The best-approximation properties of the method (cf. Remark 4.1) are not affected by this possible κ-dependence of the errors.

The numerical experiment considers the following values for the wave number, $\kappa = 2^2, 2^3, 2^4, 2^5$, and aims to study the dependence between the wave numbers and the accuracy of the numerical method. We choose uniform coarse meshes with mesh widths $H = 2^{-2}, \dots, 2^{-5}$ as depicted in Figures 4(b)–4(c). The reference mesh $\mathcal{T}_h$ is derived by uniform mesh refinement of the coarse meshes and has mesh width $h = 2^{-9}$. The corresponding P_1 conforming finite element approximation on the reference mesh $\mathcal{T}_h$ is denoted by V_h. (We disregard the possibility of adaptivity on the fine scale.) As in the previous experiment, we consider the reference solution $u_h \in V_h$ of (6.3) with the above data and compare it with coarse scale approximations $u_{H,\ell,h} \in V_{H,\ell,h}$ (cf. Definition 6.2) depending on the coarse mesh size H and the oversampling parameter ℓ. Here, we are using again the canonical quasi-interpolation $\mathcal{I}_H$.

Figures 5 and 6 show the results which conform to the theoretical predictions. If the oversampling parameter is chosen appropriately ($\ell = |\log_2 H|$) then pollution effects are eliminated for both the multiscale method (6.2) and for the Petrov-Galerkin method based on the trial-test-pairing $(V_H, V_{H,\ell,h})$ – the localized and fully discretized version of the stabilized method (4.21). Moreover, the low regularity of the solution does not affect the convergence rates of the multiscale method (6.2) when compared with the reference solution u_h, whereas slightly reduced rates are observed for the Petrov-Galerkin method based on the trial-test-pairing $(V_H, V_{H,\ell,h})$ (due to the limited approximation properties of P_1 functions in the Sobolev spaces $W^{s,2}(\Omega)$ for $s < 2$). Reduced regularity does, however, affect the accuracy of the reference solution u_h and, hence, limits the overall accuracy of our approximation. The possibility of automatic balancing the local fine scale errors of the corrector problems, the localization error, the global coarse error, and further errors due to quadrature and inexact algebraic solvers is a desirable feature of the method that needs to be addressed by future research.



(a) Results for multiscale method (6.2).

(b) Results for V-best-approximation in $V_{H,\ell,h}$.

(c) Results for multiscale Petrov-Galerkin method with trial space V_H and test space $V^*_{H,\ell,h}$.

(d) Results for standard Galerkin in the space V_H.

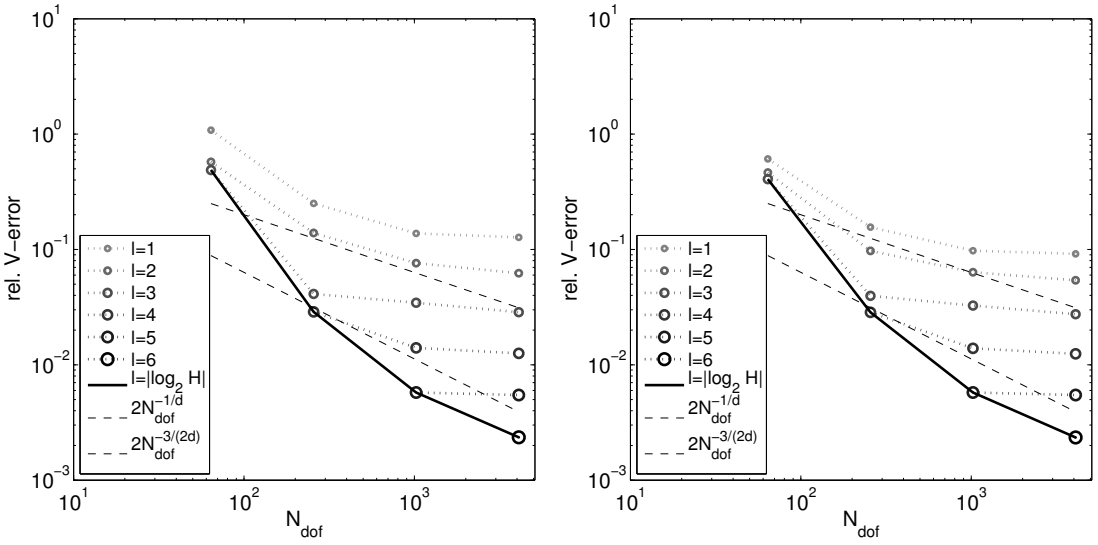
FIGURE 5. Numerical experiment of Section 7.2: Results for the multiscale method (6.2), a modification based on the trial-test-pairing $(V_H, V^*_{H,\ell,h})$ and standard P_1-FEM with several choices of the wave number κ depending on the uniform coarse mesh size $H = N_{\mathrm{dof}}^{-2}$. The reference mesh size $h = 2^{-9}$ remains fixed. The oversampling parameter is tied to the coarse mesh size via the relation $\ell = |\log_2 H|$ in (a)-(c).

REFERENCES

[AB14] A. Abdulle and Y. Bai. Reduced-order modelling numerical homogenization. *Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.*, 372(2021):20130388, 23, 2014.

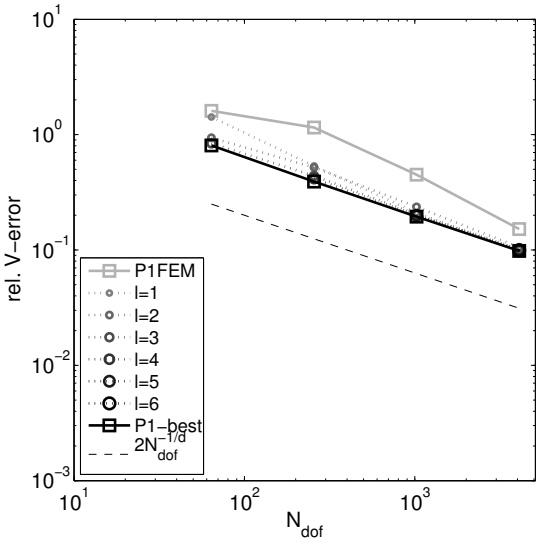
[BCWG+11] T. Betcke, S. N. Chandler-Wilde, I. G. Graham, S. Langdon, and M. Lindner. Condition number estimates for combined potential integral operators in acoustics and their boundary element discretisation. *Numer. Methods Partial Differential Equations*, 27(1):31–69, 2011.

[BP14] D. Brown and D. Peterseim. A multiscale method for porous microstructures. *ArXiv e-prints*, November 2014.



(a) Results for multiscale method (6.2).

(b) Results for V-best-approximation in $V_{H,\ell,h}$.



(c) Results for multiscale Petrov-Galerkin method with trial space V_H and test space $V^*_{H,\ell,h}$.

FIGURE 6. Numerical experiment of Section 7.2: Results for multiscale method (6.2) with wave number $\kappa = 2^5$ depending on the uniform coarse mesh size $H = N_{\mathrm{dof}}^{-2}$. The reference mesh size $h = 2^{-9}$ remains fixed. The oversampling parameter ℓ varies between 1 and 8.

[BS00] Ivo M. Babuška and Stefan A. Sauter. Is the pollution effect of the FEM avoidable for the Helmholtz equation considering high wave numbers? *SIAM Rev.*, 42(3):451–484 (electronic), 2000. Reprint of SIAM J. Numer. Anal. **3**4 (1997), no. 6, 2392–2423 [MR1480387 (99b:65135)].

[BS07] L. Banjai and S. Sauter. A refined Galerkin error and stability analysis for highly indefinite variational problems. *SIAM J. Numer. Anal.*, 45(1):37–53 (electronic), 2007.

[BS08] S. C. Brenner and L. R. Scott. *The mathematical theory of finite element methods*, volume 15 of *Texts in Applied Mathematics*. Springer, New York, third edition, 2008.

[BY14] Randolph E. Bank and Harry Yserentant. On the H^1-stability of the L_2-projection onto finite element spaces. *Numer. Math.*, 126(2):361–381, 2014.

[Car99] C. Carstensen. Quasi-interpolation and a posteriori error analysis in finite element methods. *M2AN Math. Model. Numer. Anal.*, 33(6):1187–1202, 1999.

[CF00] C. Carstensen and S. A. Funken. Constants in Clément-interpolation error and residual based a posteriori error estimates in finite element methods. *East-West J. Numer. Math.*, 8(3):153–175, 2000.

[CF06] Peter Cummings and Xiaobing Feng. Sharp regularity coefficient estimates for complex-valued acoustic and elastic Helmholtz equations. *Math. Models Methods Appl. Sci.*, 16(1):139–160, 2006.

[Cia78] Philippe G. Ciarlet. *The finite element method for elliptic problems*. North-Holland Publishing Co., Amsterdam-New York-Oxford, 1978. Studies in Mathematics and its Applications, Vol. 4.

[CV99] C. Carstensen and R. Verfürth. Edge residuals dominate a posteriori error estimates for low order finite element methods. *SIAM J. Numer. Anal.*, 36(5):1571–1587 (electronic), 1999.

[DGMZ12] L. Demkowicz, J. Gopalakrishnan, I. Muga, and J. Zitelli. Wavenumber explicit analysis of a DPG method for the multidimensional Helmholtz equation. *Comput. Methods Appl. Mech. Engrg.*, 213/216:126–138, 2012.

[DPE12] Daniele Antonio Di Pietro and Alexandre Ern. *Mathematical aspects of discontinuous Galerkin methods*, volume 69 of *Mathématiques & Applications (Berlin) [Mathematics & Applications]*. Springer, Heidelberg, 2012.

[EGMP13] D. Elfverson, E. H. Georgoulis, A. Målqvist, and D. Peterseim. Convergence of a discontinuous galerkin multiscale method. *SIAM Journal on Numerical Analysis*, 51(6):3351–3372, 2013.

[EM12] S. Esterhazy and J. M. Melenk. On stability of discretizations of the Helmholtz equation. In *Numerical analysis of multiscale problems*, volume 83 of *Lect. Notes Comput. Sci. Eng.*, pages 285–324. Springer, Heidelberg, 2012.

[For77] Michel Fortin. An analysis of the convergence of mixed finite element methods. *RAIRO Anal. Numér.*, 11(4):341–354, iii, 1977.

[FW09] Xiaobing Feng and Haijun Wu. Discontinuous Galerkin methods for the Helmholtz equation with large wave number. *SIAM J. Numer. Anal.*, 47(4):2872–2896, 2009.

[FW11] Xiaobing Feng and Haijun Wu. *hp*-discontinuous Galerkin methods for the Helmholtz equation with large wave number. *Math. Comp.*, 80(276):1997–2024, 2011.

[Het07] U. Hetmaniuk. Stability estimates for a class of Helmholtz problems. *Commun. Math. Sci.*, 5(3):665–678, 2007.

[HFMQ98] Thomas J. R. Hughes, Gonzalo R. Feijóo, Luca Mazzei, and Jean-Baptiste Quincy. The variational multiscale method—a paradigm for computational mechanics. *Comput. Methods Appl. Mech. Engrg.*, 166(1-2):3–24, 1998.

[HMP11] R. Hiptmair, A. Moiola, and I. Perugia. Plane wave discontinuous Galerkin methods for the 2D Helmholtz equation: analysis of the *p*-version. *SIAM J. Numer. Anal.*, 49(1):264–284, 2011.

[HMP14a] P. Henning, P. Morgenstern, and D. Peterseim. Multiscale Partition of Unity. In M. Griebel and M. A. Schweitzer, editors, *Meshfree Methods for Partial Differential Equations VII*, volume 100 of *Lecture Notes in Computational Science and Engineering*. Springer, 2014.

[HMP14b] Patrick Henning, Axel Målqvist, and Daniel Peterseim. A localized orthogonal decomposition method for semi-linear elliptic problems. *ESAIM: Mathematical Modelling and Numerical Analysis*, 48:1331–1349, 9 2014.

[HMP14c] Ralf Hiptmair, Andrea Moiola, and Ilaria Perugia. Trefftz discontinuous Galerkin methods for acoustic scattering on locally refined meshes. *Appl. Numer. Math.*, 79:79–91, 2014.

[HP13] P. Henning and D. Peterseim. Oversampling for the Multiscale Finite Element Method. *Multiscale Model. Simul.*, 11(4):1149–1175, 2013.

[HS07] T. 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.

[Hug95] Thomas J. R. Hughes. Multiscale phenomena: Green's functions, the Dirichlet-to-Neumann formulation, subgrid scale models, bubbles and the origins of stabilized methods. *Comput. Methods Appl. Mech. Engrg.*, 127(1-4):387–401, 1995.

[Mål11] Axel Målqvist. Multiscale methods for elliptic problems. *Multiscale Model. Simul.*, 9(3):1064–1086, 2011.

[Mel95] Jens Markus Melenk. *On generalized finite-element methods*. ProQuest LLC, Ann Arbor, MI, 1995. Thesis (Ph.D.)–University of Maryland, College Park.

[MIB96] Ch. Makridakis, F. Ihlenburg, and I. Babuška. Analysis and finite element methods for a fluid-solid interaction problem in one dimension. *Mathematical Models and Methods in Applied Sciences*, 06(08):1119–1141, 1996.

[MP14a] Axel Målqvist and Daniel Peterseim. Computation of eigenvalues by numerical upscaling. *Numerische Mathematik*, pages 1–25, 2014.

[MP14b] Axel Målqvist and Daniel Peterseim. Localization of elliptic multiscale problems. *Math. Comp.*, 83(290):2583–2603, 2014.

[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.

[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.

[PS14] D. Peterseim and R. Scheichl. Rigorous numerical upscaling at high contrast. *in preparation*, 2014+.

[RHP08] G. Rozza, D. B. P. Huynh, and A. T. Patera. Reduced basis approximation and a posteriori error estimation for affinely parametrized elliptic coercive partial differential equations: application to transport and continuum mechanics. *Arch. Comput. Methods Eng.*, 15(3):229–275, 2008.

[Sau06] S. A. Sauter. A refined finite element convergence theory for highly indefinite Helmholtz problems. *Computing*, 78(2):101–115, 2006.

[Sch74] Alfred H. Schatz. An observation concerning Ritz-Galerkin methods with indefinite bilinear forms. *Math. Comp.*, 28:959–962, 1974.

[Szy06] Daniel B. Szyld. The many proofs of an identity on the norm of oblique projections. *Numer. Algorithms*, 42(3-4):309–323, 2006.

[TF06] Radek Tezaur and Charbel Farhat. Three-dimensional discontinuous Galerkin elements with plane waves and Lagrange multipliers for the solution of mid-frequency Helmholtz problems. *Internat. J. Numer. Methods Engrg.*, 66(5):796–815, 2006.

[Wu14] Haijun Wu. Pre-asymptotic error analysis of cip-fem and fem for the helmholtz equation with high wave number. part i: linear version. *IMA Journal of Numerical Analysis*, 34(3):1266–1288, 2014.

[ZMD+11] J. Zitelli, I. Muga, L. Demkowicz, J. Gopalakrishnan, D. Pardo, and V.M. Calo. A class of discontinuous petrovgalerkin methods. part iv: The optimal test norm and time-harmonic wave propagation in 1d. *Journal of Computational Physics*, 230(7):2406 – 2432, 2011.

(D. Peterseim) Rheinische Friedrich-Wilhelms-Universität Bonn, Institute for Numerical Simulation, Wegelerstr. 6, 53115 Bonn, Germany

E-mail address: peterseim@ins.uni-bonn.de

B.2 Stable multiscale Petrov-Galerkin finite element method for high frequency acoustic scattering

Computer Methods in Applied Mechanics and Engineering **295**:1–17, 2015.

(with D. Gallistl)

Available online at www.sciencedirect.com
ScienceDirect
Comput. Methods Appl. Mech. Engrg. 295 (2015) 1–17
Computer methods in applied mechanics and engineering
www.elsevier.com/locate/cma

Stable multiscale Petrov–Galerkin finite element method for high frequency acoustic scattering

D. Gallistl, D. Peterseim*

Institut für Numerische Simulation, Universität Bonn, Wegelerstraße 6, D-53115 Bonn, Germany

Received 18 March 2015; received in revised form 24 June 2015; accepted 25 June 2015
Available online 4 July 2015

Abstract

We present and analyze a pollution-free Petrov–Galerkin multiscale finite element method for the Helmholtz problem with large wave number κ as a variant of Peterseim (2014). We use standard continuous Q_1 finite elements at a coarse discretization scale H as trial functions, whereas the test functions are computed as the solutions of local problems at a finer scale h. The diameter of the support of the test functions behaves like mH for some oversampling parameter m. Provided m is of the order of $\log(\kappa)$ and h is sufficiently small, the resulting method is stable and quasi-optimal in the regime where H is proportional to κ^{-1}. In homogeneous (or more general periodic) media, the fine scale test functions depend only on local mesh-configurations. Therefore, the seemingly high cost for the computation of the test functions can be drastically reduced on structured meshes. We present numerical experiments in two and three space dimensions.

MSC: 35J05; 65N12; 65N15; 65N30

Keywords: Multiscale method; Pollution effect; Wave propagation; Helmholtz problem; Finite element method

1. Introduction

Standard finite element methods (FEMs) for acoustic wave propagation are well known to exhibit the so-called *pollution effect* [1], which means that the stability and convergence of the scheme require a much smaller mesh-size than needed for a meaningful approximation of the wave by finite element functions. For a highly oscillatory wave at wave number κ, the typical requirement for a reasonable representation reads $\kappa H \lesssim 1$ for the mesh-size H, that is some fixed number of elements per wave-length. The standard Galerkin FEM typically requires at least $\kappa^\alpha H \lesssim 1$ where $\alpha > 1$ depends on the method and the stability and regularity properties of the continuous problem. There have been various attempts to reduce or avoid the pollution effect, e.g., discontinuous Galerkin methods [2–5], high-order finite elements [6,7], discontinuous Petrov–Galerkin methods [8,9], or the continuous interior penalty method [10] among many others. A good historical overview is provided in [8].

* Corresponding author.
E-mail addresses: gallistl@ins.uni-bonn.de (D. Gallistl), peterseim@ins.uni-bonn.de (D. Peterseim).

http://dx.doi.org/10.1016/j.cma.2015.06.017
0045-7825/

The work [11] suggested a multiscale Petrov–Galerkin method for the Helmholtz equation where standard finite element trial and test functions are modified by a local subscale correction in the spirit of numerical homogenization [12]. In the numerical experiments of [11], a variant of that method appeared attractive where only the test functions are modified while standard finite element functions are used as trial functions. In this paper, we analyze that method and reformulate it as a stabilized Q_1 method in the spirit of the variational multiscale method [13–17]. The method employs standard Q_1 finite element trial functions on a grid $\mathcal{G}_H$ with mesh-size H. The test functions are the solutions of local problems with respect to a grid $\mathcal{G}_h$ at a finer scale h which is chosen fine enough to allow for stability of the standard Galerkin FEM over $\mathcal{G}_h$. The diameter of the support of the test functions is proportional to mH for the oversampling parameter m. Under the condition that m is logarithmically coupled with the wave number κ through $m \approx \log(\kappa)$, we prove that the method is pollution-free, i.e., the resolution condition $\kappa H \lesssim 1$ is sufficient for stability and quasi-optimality under fairly general assumptions on the stability of the continuous problem. The performance of the method is illustrated in the convergence history of Fig. 1. More detailed descriptions on the numerical experiments will be given in Section 5. As the test functions only depend on local mesh-configurations, on structured meshes the number of test functions to be actually computed is much smaller than the overall number of trial and test functions on the coarse scale. In many cases, the computational cost is then dominated by the coarse solve and the overhead compared with a standard FEM on the same coarse mesh remains proportional to $m^d \approx \log(\kappa)^d$. Even if no structure of the mesh can be exploited to reduce the number of patch problems, the method may still be attractive if the problem has to be solved many times with different data (same κ though) in the context of inverse problems or parameter identification problems.

The paper is structured as follows. Section 2 states the Helmholtz problem and recalls some important results. The definition of the new Petrov–Galerkin method follows in Section 3. Stability and error analysis are carried out in Section 4. Section 5 is devoted to numerical experiments.

Standard notation on complex-valued Lebesgue and Sobolev spaces applies throughout this paper. The bar indicates complex conjugation and i is the imaginary unit. The L^2 inner product is denoted by $(v, w)_{L^2(\Omega)} := \int_\Omega v\bar{w}\,dx$. The Sobolev space of complex-valued L^p functions over a domain ω whose generalized derivatives up to order k belong to L^p is denoted by $W^{k,p}(\omega; \mathbb{C})$. The notation $A \lesssim B$ abbreviates $A \le CB$ for some constant C that is independent of the mesh-size, the wave number κ, and all further parameters in the method like the oversampling parameter m or the fine-scale mesh-size h; $A \approx B$ abbreviates $A \lesssim B \lesssim A$.

2. The Helmholtz problem

Let $\Omega \subseteq \mathbb{R}^d$, for $d \in \{1, 2, 3\}$, be an open bounded domain with polyhedral Lipschitz boundary which is decomposed into disjoint parts $\partial\Omega = \Gamma_D \cup \Gamma_R$ with Γ_D closed. The classical Helmholtz equation then reads

$$\begin{aligned} -\Delta u - \kappa^2 u &= f \quad \text{in } \Omega, \\ u &= u_D \text{ on } \Gamma_D, \\ i\kappa u - \nabla u \cdot \nu &= g \quad \text{on } \Gamma_R \end{aligned} \tag{2.1}$$

for the outer unit normal ν of Ω and the real parameter $\kappa > 0$. For the sake of a simple exposition we assume $u_D = 0$. Either of the parts Γ_D or Γ_R is allowed to be the empty set. In scattering problems, the Dirichlet boundary Γ_D typically refers to the boundary of a bounded sound-soft object whereas the Robin boundary Γ_R arises from artificially truncating the full space $\mathbb{R}^d$ to the bounded domain Ω [18]. The variational formulation of (2.1) employs the space

$$V := W^{1,2}_D(\Omega; \mathbb{C}) := \{v \in W^{1,2}(\Omega; \mathbb{C}) : v|_{\Gamma_D} = 0\}.$$

For any subset $\omega \subseteq \Omega$ we define the norm

$$\|v\|_{V,\omega} := \sqrt{\kappa^2 \|v\|^2_{L^2(\omega)} + \|\nabla v\|^2_{L^2(\omega)}} \quad \text{for any } v \in V$$

and denote $\|v\|_V := \|v\|_{V,\Omega}$. Define on V the following sesquilinear form

$$a(v, w) := (\nabla v, \nabla w)_{L^2(\Omega)} - \kappa^2 (v, w)_{L^2(\Omega)} - i\kappa (v, w)_{L^2(\Gamma_R)}.$$

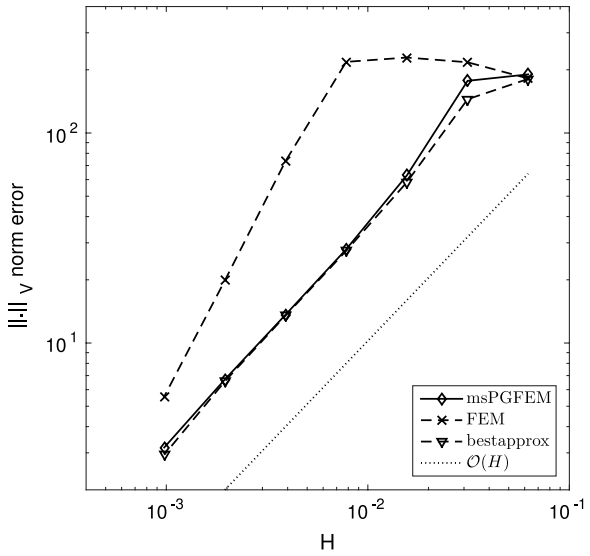


Fig. 1. Convergence history of the multiscale FEM (msPGFEM), the standard Q_1 FEM (FEM) and the best-approximation (bestapprox) in the finite element space for a two-dimensional plane wave with wave number $\kappa = 2^7$ (see also Section 5).

Although the results of this paper hold for a rather general right-hand side in the dual of V, we focus on data $f \in L^2(\Omega;\mathbb{C})$ and $g \in L^2(\Gamma_R;\mathbb{C})$ for the ease of presentation. The weak form of the Helmholtz problem then seeks $u \in V$ such that

$$a(u,v) = (f,v)_{L^2(\Omega)} + (g,v)_{L^2(\Gamma_R)} \quad \text{for all } v \in V. \tag{2.2}$$

We assume that the problem is polynomially well-posed [19] in the sense that there exists some constant $\gamma(\kappa,\Omega)$ which depends polynomially on κ such that

$$\gamma(\kappa,\Omega)^{-1} \le \inf_{v\in V\setminus\{0\}} \sup_{w\in V\setminus\{0\}} \frac{\Re a(v,w)}{\|v\|_V \|w\|_V}. \tag{2.3}$$

For instance, in the particular case of pure impedance boundary conditions $\partial\Omega = \Gamma_R$, it was proved in [20,21] by employing a technique of [22] that $\gamma(\kappa,\Omega) \lesssim \kappa$. Further setups allowing for polynomially well-posedness are described in [23,19,24]. In particular, the case of a medium described by a convex domain (with Robin boundary conditions on the outer part of the boundary) and a star-shaped scatterer (with Dirichlet boundary conditions) allows for polynomial well-posedness [23]. Another admissible setting is described in [19] where Ω is a bounded Lipschitz domain with pure Robin boundary. For general configurations, however, the dependence of the stability constant $\gamma(\kappa,\Omega)$ from (2.3) is an open question. Throughout this paper we assume that (2.3) is satisfied. The case of a possible exponential dependence [25] is excluded here.

3. The method

This section introduces the notation on finite element spaces and meshes and defines the multiscale Petrov–Galerkin method (msPGFEM) for the Helmholtz problem.

3.1. Meshes and data structures

Let $\mathcal{G}_H$ be a regular partition of Ω into intervals, parallelograms, parallelepipeds for $d = 1, 2, 3$, respectively, such that $\cup\mathcal{G}_H = \overline{\Omega}$ and any two distinct $T, T' \in \mathcal{G}_H$ are either disjoint or share exactly one lower-dimensional hyper-face (that is a vertex or an edge for $d \in \{2,3\}$ or a face for $d = 3$). We impose shape-regularity in the sense that the aspect ratio of the elements in $\mathcal{G}_H$ is uniformly bounded. Since we are considering quadrilaterals (resp. hexahedra) with parallel faces, this guarantees the non-degeneracy of the elements in $\mathcal{G}_H$. We consider this type of partitions for

the sake of a simple presentation and to exploit the structure to increase the computational efficiency. The theory of this paper carries over to simplicial triangulations or to more general quadrilateral or hexahedral partitions satisfying suitable non-degeneracy conditions or even to meshless methods based on proper partitions of unity [26].

Given any subdomain $S \subseteq \overline{\Omega}$, define its neighborhood via

$$\mathsf{N}(S) := \operatorname{int}\Big(\cup\{T \in \mathcal{G}_H \,:\, T \cap \overline{S} \neq \emptyset\}\Big).$$

Furthermore, we introduce for any $m \geq 2$ the patches

$$\mathsf{N}^1(S) := \mathsf{N}(S) \quad \text{and} \quad \mathsf{N}^m(S) := \mathsf{N}(\mathsf{N}^{m-1}(S)).$$

The shape-regularity implies that there is a uniform bound $C_{\mathrm{ol},m} = C_{\mathrm{ol},m}(d)$ on the number of elements in the mth-order patch,

$$\max_{T\in\mathcal{G}_H} \operatorname{card}\{K \in \mathcal{G}_H \,:\, K \subseteq \overline{\mathsf{N}^m(T)}\} \leq C_{\mathrm{ol},m}.$$

We abbreviate $C_{\mathrm{ol}} := C_{\mathrm{ol},1}$. Throughout this paper, we assume that the coarse-scale mesh $\mathcal{G}_H$ is quasi-uniform. This implies that $C_{\mathrm{ol},m}$ depends polynomially on m. The global mesh-size reads $H := \max\{\operatorname{diam}(T) : T \in \mathcal{G}_H\}$. Let $Q_p(\mathcal{G}_H)$ denote the space of piecewise polynomials of partial degree $\leq p$. The space of globally continuous piecewise first-order polynomials reads

$$\mathcal{S}^1(\mathcal{G}_H) := C^0(\Omega) \cap Q_1(\mathcal{G}_H).$$

The standard Q_1 finite element space reads

$$V_H := \mathcal{S}^1(\mathcal{G}_H) \cap V.$$

The set of free vertices (the degrees of freedom) is denoted by

$$\mathcal{N}_H := \{z \in \overline{\Omega} \,:\, z \text{ is a vertex of } \mathcal{G}_H \text{ and } z \notin \Gamma_D\}.$$

Let $I_H : V \to V_H$ be a surjective quasi-interpolation operator that acts as a stable quasi-local projection in the sense that $I_H \circ I_H = I_H$ and that for any $T \in \mathcal{G}_H$ and all $v \in V$ there holds

$$H^{-1}\|v - I_H v\|_{L^2(T)} + \|\nabla I_H v\|_{L^2(T)} \leq C_{I_H}\|\nabla v\|_{L^2(\mathsf{N}(T))}. \tag{3.1}$$

Under the mesh condition that $\kappa H \lesssim 1$ is bounded by a generic constant, this implies stability in the $\|\cdot\|_V$ norm

$$\|I_H v\|_V \leq C_{I_H,V}\|v\|_V \quad \text{for all } v \in V, \tag{3.2}$$

with a κ-independent constant $C_{I_H,V}$. One possible choice (which we use in our implementation of the method) is to define $I_H := E_H \circ \Pi_H$, where Π_H is the piecewise L^2 projection onto $Q_1(\mathcal{G}_H)$ and E_H is the averaging operator that maps $Q_1(\mathcal{G}_H)$ to V_H by assigning to each free vertex the arithmetic mean of the corresponding function values of the neighboring cells, that is, for any $v \in Q_1(\mathcal{G}_H)$ and any free vertex $z \in \mathcal{N}_H$,

$$(E_H(v))(z) = \sum_{\substack{T\in\mathcal{G}_H \\ \text{with } z\in T}} v|_T(z) \Big/ \operatorname{card}\{K \in \mathcal{G}_H \,:\, z \in K\}.$$

Note that $E_H(v)|_{\Gamma_D} = 0$ by construction. For this choice, the proof of (3.1) follows from combining the well-established approximation and stability properties of Π_H and E_H, see, e.g., [27].

3.2. Definition of the method

The method is determined by three parameters, namely the coarse-scale mesh-size H, and the stabilization parameters h (the fine-scale mesh-size) and m (the oversampling parameter) which are explained in the following. We assign to any $T \in \mathcal{G}_H$ its mth order patch $\Omega_T := \mathsf{N}^m(T)$ (for a positive integer m) and define for any $v, w \in V$ the localized sesquilinear forms

$$a_{\Omega_T}(v, w) := (\nabla v, \nabla w)_{L^2(\Omega_T)} - (\kappa^2 v, w)_{L^2(\Omega_T)} - i(\kappa v, w)_{L^2(\Gamma_R\cap\partial\Omega_T)}$$

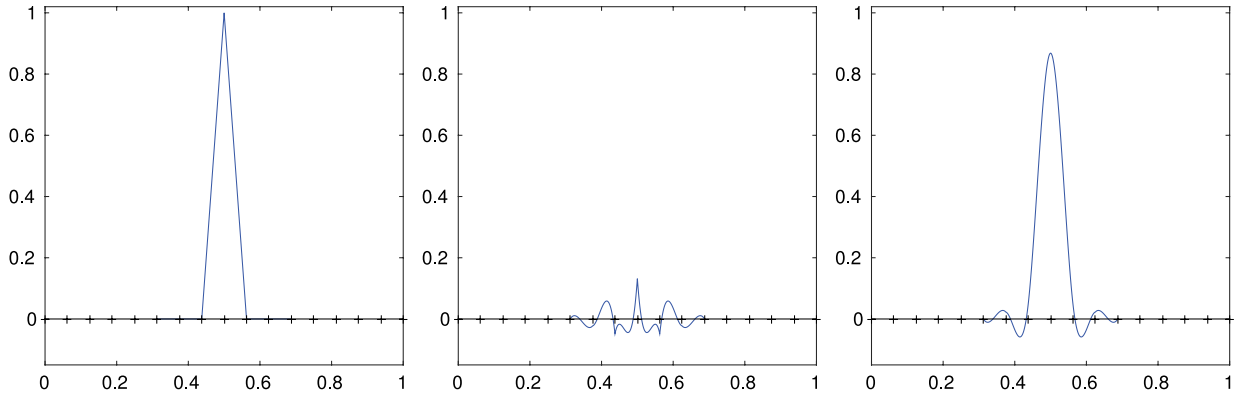


Fig. 2. Coarse-scale trial function Λ_z (left), corrector λ_z (middle), and modified test function $\widetilde{\Lambda}_z = \Lambda_z - \lambda_z$ (right) in 1D with $\kappa = 2^5$, $H = 2^{-4}$, $h = 2^{-10}$, $m = 2$.

and

$$a_T(v, w) := (\nabla v, \nabla w)_{L^2(T)} - (\kappa^2 v, w)_{L^2(T)} - i(\kappa v, w)_{L^2(\Gamma_R \cap \partial T)}.$$

Let $\mathcal{G}_h$ be a global uniform refinement of the mesh $\mathcal{G}_H$ over Ω and define

$$V_h(\Omega_T) := \{v \in Q_1(\mathcal{G}_h) \cap V \ : v = 0 \text{ outside } \Omega_T\}.$$

Define the null space

$$W_h(\Omega_T) := \{v_h \in V_h(\Omega_T) \ : \ I_H(v_h) = 0\}$$

of the quasi-interpolation operator I_H defined in the previous section. Given any nodal basis function $\Lambda_z \in V_H$, let $\lambda_{z,T} \in W_h(\Omega_T)$ solve the subscale corrector problem

$$a_{\Omega_T}(w, \lambda_{z,T}) = a_T(w, \Lambda_z) \quad \text{for all } w \in W_h(\Omega_T). \tag{3.3}$$

The well-posedness of (3.3) will be proved in Section 4. Let $\lambda_z := \sum_{T \in \mathcal{G}_H} \lambda_{z,T}$ and define the test function

$$\widetilde{\Lambda}_z := \Lambda_z - \lambda_z.$$

The space of test functions then reads

$$\widetilde{V}_H := \operatorname{span}\{\widetilde{\Lambda}_z \ : \ z \in \mathcal{N}_H\}.$$

We emphasize that the dimension $\dim V_H = \dim \widetilde{V}_H$ is independent of the parameters m and h. Figs. 2–3 display typical examples for the test functions $\widetilde{\Lambda}_z$ and correctors. The multiscale Petrov–Galerkin FEM seeks $u_H \in V_H$ such that

$$a(u_H, \tilde{v}_H) = (f, \tilde{v}_H)_{L^2(\Omega)} + (g, \tilde{v}_H)_{L^2(\Gamma_R)} \quad \text{for all } \tilde{v}_H \in \widetilde{V}_H. \tag{3.4}$$

The error analysis and the numerical experiments will show that the choice $H \lesssim \kappa^{-1}$, $m \approx \log(\kappa)$ suffices to guarantee stability and quasi-optimality properties, provided that $\kappa^\alpha h \lesssim 1$ where α depends on the stability and regularity of the continuous problem. The conditions on h are the same as for the standard Q_1 FEM on the global fine scale (e.g. $\kappa^{3/2} h \lesssim 1$ for stability [10] and $\kappa^2 h \lesssim 1$ for quasi-optimality [20] in the case of pure Robin boundary conditions on a convex domain).

3.3. Remarks on generalizations of the method

The present approach exploits additional structure in the mesh and thereby drastically decreases the cost for the computation of the test functions $(\widetilde{\Lambda}_z \ : \ z \in \mathcal{N}_H)$. Indeed, (3.3) is translation-invariant and, thus, the number of corrector problems to be solved is determined by the number of patch configurations. This number is typically much smaller than the number of elements in $\mathcal{G}_H$, see Fig. 4 for an illustration.

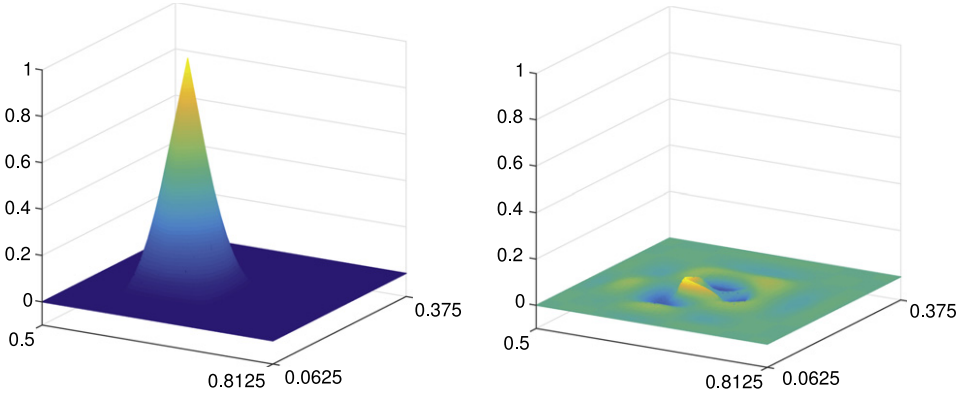


Fig. 3. Coarse-scale trial function Λ_z (left), and element corrector $\lambda_{z,T}$ (right) in 2D with $\kappa = 2^5$, $H = 2^{-4}$, $h = 2^{-7}$, $m = 2$ for the patch highlighted in Fig. 4.

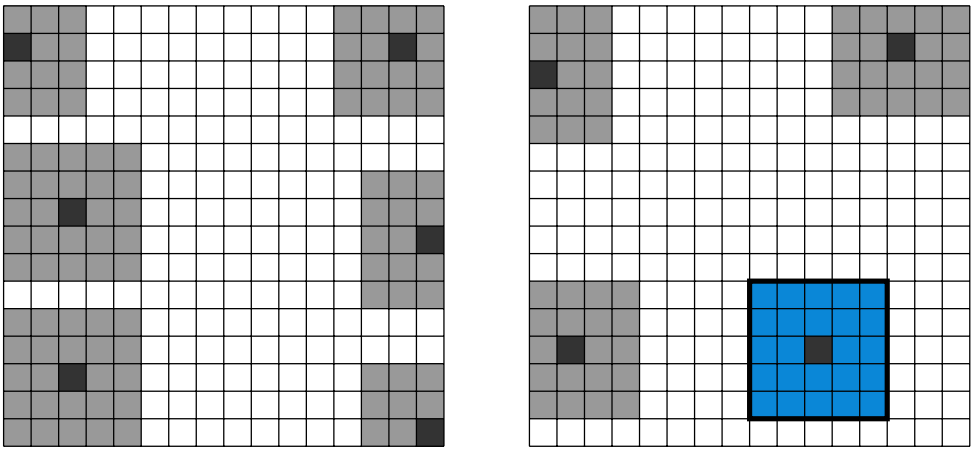


Fig. 4. All possible patch configurations (up to rotations) on a structured mesh of the square domain with pure Robin boundary with $m = 2$. A trial function and corresponding corrector for the highlighted patch are depicted in Fig. 3.

Some remarks on more general versions of the presented msPGFEM are in order.

Element shapes. As Fig. 4 illustrates, highly structured meshes are desirable as they lead to a moderate number of patch problems. The method presented in Section 3.2 considers, for simplicity, a partition of the domain in parallelepipeds. While in scattering problems the outer part of the boundary Γ_R results from a truncation of the full space and, hence, the choice of a simple geometry (e.g., a cube) is justified, it is extremely important to guarantee an accurate representation of more general scattering objects. This requires more general element shapes such as isoparametric elements or partitions in bricks and simplices with first-order ansatz functions on the reference cell (see [11] for simplicial meshes). The msPGFEM and its error analysis are also applicable to this situation. The configurations at the boundary will then determine the number of corrector problems.

Fine-scale grid. The present approach is based on a global fine-scale grid $\mathcal{G}_h$ and a particular choice of the domains Ω_T, which is convenient for the implementation of the method. It is, however, not necessary for the domains Ω_T to be aligned with the mesh $\mathcal{G}_H$. Also the spaces $W_h(\Omega_T)$ can be defined over independent fine-scale meshes over Ω_T.

Adaptive methods. For certain configurations of the domains Ω_T, for instance in the presence of re-entrant corners, it may be desirable to utilize an adaptive fine-scale mesh over Ω_T for the solution of the corrector problem (3.3). As proven in Lemma 1, the corrector problems are coercive and mesh-adaptation may improve the efficiency of the fine-scale corrector problem. As mentioned in the previous remark, it is indeed possible to employ independent fine-scale meshes over different domains Ω_T, Ω_K. The stability and error analysis for the adaptive case, which are expected to be more involved, are not discussed in this paper.

4. Error analysis

We denote the global finite element space on the fine scale by $V_h := V_h(\Omega) = \mathcal{S}^1(\mathcal{G}_h) \cap V$. We denote the solution operator of the element corrector problem (3.3) by $\mathcal{C}_{T,m}$. Then any $z \in \mathcal{N}_H$ and any $T \in \mathcal{G}_H$ satisfy $\lambda_{z,T} = \mathcal{C}_{T,m}(\Lambda_z)$ and we refer to $\mathcal{C}_{T,m}$ as element correction operator. The map $\Lambda_z \mapsto \lambda_z$ described in Section 3.2 defines a linear operator $\mathcal{C}_m$ via $\mathcal{C}_m(\Lambda_z) = \lambda_z$ for any $z \in \mathcal{N}_H$, referred to as correction operator. For the analysis we introduce idealized counterparts of these correction operators where the patch Ω_T equals Ω. Define the null space $W_h := \{v \in V_h : I_H(v) = 0\}$. For any $v \in V$, the idealized element corrector problem seeks $\mathcal{C}_{T,\infty} v \in W_h$ such that

$$a(w, \mathcal{C}_{T,\infty} v) = a_T(w, v) \quad \text{for all } w \in W_h. \tag{4.1}$$

Furthermore, define

$$\mathcal{C}_\infty v := \sum_{T \in \mathcal{G}_H} \mathcal{C}_{T,\infty} v. \tag{4.2}$$

It is proved in [6, Corollary 3.2] that the form a is continuous and there is a constant C_a such that

$$a(v, w) \le C_a \|v\|_V \|w\|_V \quad \text{for all } v, w \in V.$$

The following result implies the well-posedness of the idealized corrector problems.

Lemma 1 (*Well-posedness of the Idealized Corrector Problems*)**.** *Provided*

$$C_{I_H} \sqrt{C_{\mathrm{ol}}} H \kappa \le 1/\sqrt{2}, \tag{4.3}$$

we have for all $w \in W_h$ equivalence of norms

$$\|\nabla w\|_{L^2(\Omega)} \le \|w\|_V \le \sqrt{3/2}\, \|\nabla w\|_{L^2(\Omega)}$$

and ellipticity

$$\frac{1}{2} \|\nabla w\|^2_{L^2(\Omega)} \le \Re a(w, w).$$

Proof. For any $w \in W_h$ the property (3.1) implies

$$\kappa^2 \|w\|^2_{L^2(\Omega)} = \kappa^2 \|(1 - I_H) w\|^2_{L^2(\Omega)} \le C^2_{I_H} C_{\mathrm{ol}} H^2 \kappa^2 \|\nabla w\|^2_{L^2(\Omega)}. \quad \square$$

Lemma 1 implies that the idealized corrector problems (4.2) are well-posed and the correction operator $\mathcal{C}_\infty$ is continuous in the sense that

$$\|\mathcal{C}_\infty v_H\|_V \le C_{\mathcal{C}} \|v_H\|_V \quad \text{for all } v_H \in V_H$$

for some constant $C_{\mathcal{C}} \approx 1$. Since the inclusion $W_h(\Omega_T) \subseteq W_h$ holds, the well-posedness result of Lemma 1 carries over to the corrector problems (3.3) in the subspace $W_h(\Omega_T)$ with the sesquilinear form a_{Ω_T}.

The proof of well-posedness of the Petrov–Galerkin method (3.4) will be based on the fact that the difference $(\mathcal{C}_\infty - \mathcal{C}_m)(v)$ decays exponentially with the distance from $\mathrm{supp}(v)$. In the next theorem, we quantify the difference between the idealized and discrete correctors. The proof will be given in Appendix A of this paper and is based on the exponential decay of the corrector $\mathcal{C}_\infty \Lambda_z$ itself, see Fig. 5. That figure also illustrates that the decay requires the resolution condition (4.3), namely $\kappa H \lesssim 1$.

Theorem 1. *Under the resolution condition (4.3) there exist constants $C_1 \approx 1 \approx C_2$ and $0 < \beta < 1$ such that any $v \in V_H$, any $T \in \mathcal{G}_H$ and any $m \in \mathbb{N}$ satisfy*

$$\|\nabla(\mathcal{C}_{T,\infty} v - \mathcal{C}_{T,m} v)\|_{L^2(\Omega)} \le C_1 \beta^m \|\nabla v\|_{L^2(T)}, \tag{4.4}$$

$$\|\nabla(\mathcal{C}_\infty v - \mathcal{C}_m v)\|_{L^2(\Omega)} \le C_2 \sqrt{C_{\mathrm{ol},m}}\, \beta^m \|\nabla v\|_{L^2(\Omega)}. \quad \square \tag{4.5}$$

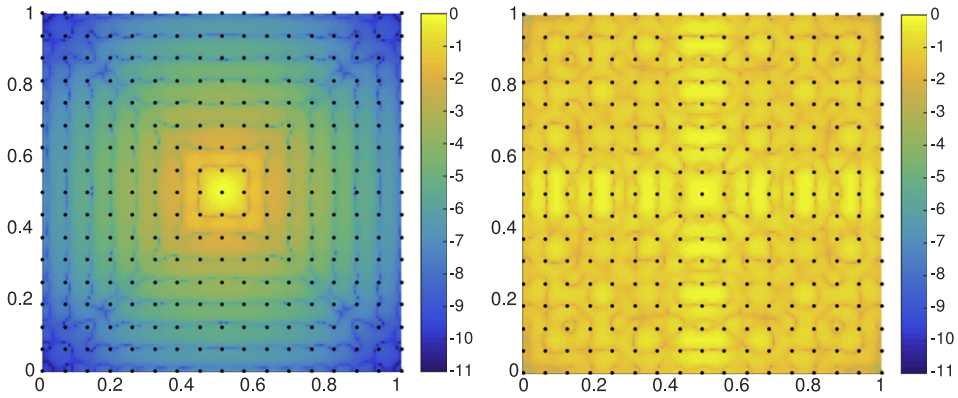


Fig. 5. Modulus of the idealized test function $\widetilde{\Lambda}_z$ for $m = \infty$, $H = 2^{-4}$, $h = 2^{-7}$ in 2D in a logarithmically scaled plot. The dots indicate the grid points of the coarse mesh. Left: $\kappa = 2^5$; right: $\kappa = 2^6$.

Provided h is chosen fine enough, the standard FEM over $\mathcal{G}_h$ is stable in the sense that there exists a constant C_{FEM} such that with $\gamma(\kappa, \Omega)$ from (2.3) there holds

$$\big(C_{\mathrm{FEM}}\gamma(\kappa, \Omega)\big)^{-1} \leq \inf_{v\in V_h\setminus\{0\}} \sup_{w\in V_h\setminus\{0\}} \frac{\Re a(v, w)}{\|v\|_V \|w\|_V}. \tag{4.6}$$

This is actually a condition on the fine-scale parameter h. In general, the requirements on h depend on the stability of the continuous problem [20].

Theorem 2 (*Well-posedness of the Discrete Problem*)**.** *Under the resolution conditions* (4.3) *and* (4.6) *and the following oversampling condition*

$$m \geq |\log\big(\sqrt{6}C_a\sqrt{C_{\mathrm{ol}}}C_{I_H}C_{I_H,V}C_2\sqrt{C_{\mathrm{ol},m}}C_{\mathrm{FEM}}\gamma(\kappa, \Omega)\big)|\big/|\log(\beta)|, \tag{4.7}$$

problem (3.4) *is well-posed and the constant* $C_{\mathrm{PG}} := 2C_{I_H,V}C_{\mathcal{C}}C_{\mathrm{FEM}}$ *satisfies*

$$\big(C_{\mathrm{PG}}\gamma(\kappa, \Omega)\big)^{-1} \leq \inf_{v_H\in V_H\setminus\{0\}} \sup_{\tilde{v}_H\in\widetilde{V}_H\setminus\{0\}} \frac{\Re a(v_H, \tilde{v}_H)}{\|v_H\|_V \|\tilde{v}_H\|_V}.$$

Proof. Let $u_H \in V_H$ with $\|u_H\|_V = 1$. From (4.6) we infer that there exists some $v \in V_h$ with $\|v\|_V = 1$ such that

$$\Re a(u_H - \overline{\mathcal{C}_\infty(\bar{u}_H)}, v) \geq \big(C_{\mathrm{FEM}}\gamma(\kappa, \Omega)\big)^{-1}\|u_H - \overline{\mathcal{C}_\infty(\bar{u}_H)}\|_V.$$

It follows from the structure of the sesquilinear form a that $\overline{\mathcal{C}_\infty(\bar{u}_H)}$ solves the following adjoint corrector problem

$$a(\overline{\mathcal{C}_\infty(\bar{u}_H)}, w) = a(u_H, w) \quad \text{for all } w \in W_h, \tag{4.8}$$

cf. [28, Lemma 3.1]. Let $\tilde{v}_H := (1 - \mathcal{C}_m)I_H v \in \widetilde{V}_H$. We have

$$a(u_H, \tilde{v}_H) = a(u_H, (1 - \mathcal{C}_\infty)I_H v) + a(u_H, (\mathcal{C}_\infty - \mathcal{C}_m)I_H v). \tag{4.9}$$

Since $\mathcal{C}_\infty$ is a projection onto W_h, we have $(1 - \mathcal{C}_\infty)(1 - I_H)v = 0$ and, thus, $(1 - \mathcal{C}_\infty)I_H v = (1 - \mathcal{C}_\infty)v$. The solution properties (4.8) of $\overline{\mathcal{C}_\infty(\bar{u}_H)}$ and (4.1)–(4.2) of $\mathcal{C}_\infty v$ prove $a(u_H, \mathcal{C}_\infty v) = a(\overline{\mathcal{C}_\infty(\bar{u}_H)}, v)$. Hence,

$$\begin{aligned}\Re a(u_H, (1 - \mathcal{C}_\infty)I_H v) &= \Re a(u_H - \overline{\mathcal{C}_\infty(\bar{u}_H)}, v)\\ &\geq \big(C_{\mathrm{FEM}}\gamma(\kappa, \Omega)\big)^{-1}\|u_H - \overline{\mathcal{C}_\infty(\bar{u}_H)}\|_V.\end{aligned}$$

Furthermore, the estimate (3.2) implies

$$1 = \|u_H\|_V = \|I_H(u_H - \overline{\mathcal{C}_\infty(\bar{u}_H)})\|_V \leq C_{I_H,V}\|u_H - \overline{\mathcal{C}_\infty(\bar{u}_H)}\|_V.$$

The second term on the right-hand side of (4.9) satisfies with $\|u_H\|_V = 1$ and Lemma 1 that

$$|a(u_H, (\mathcal{C}_\infty - \mathcal{C}_m)I_H v)| \leq \sqrt{3/2}C_a\|\nabla(\mathcal{C}_\infty - \mathcal{C}_m)I_H v\|_{L^2(\Omega)}.$$

Altogether, it follows that

$$\Re a(u_H, \tilde{v}_H) \geq \left(\frac{1}{C_{I_H,V}C_{\mathrm{FEM}}\gamma(\kappa,\Omega)} - \sqrt{\frac{3}{2}}C_a\|\nabla(\mathcal{C}_\infty - \mathcal{C}_m)I_H v\|_{L^2(\Omega)}\right).$$

Theorem 1 and (3.1) show that

$$\|\nabla(\mathcal{C}_\infty - \mathcal{C}_m)I_H v\|_{L^2(\Omega)} \leq C_2\sqrt{C_{\mathrm{ol},m}}\beta^m\|\nabla I_H v\| \leq C_2\sqrt{C_{\mathrm{ol},m}}C_{I_H}\sqrt{C_{\mathrm{ol}}}\beta^m.$$

Hence, the condition (4.7) and $\|\tilde{v}_H\|_V = \|(1-\mathcal{C}_\infty)v\|_V \leq C_{\mathcal{C}}$ imply the assertion. □

Remark 1 (*Adjoint Problem*)**.** Under the assumptions of Theorem 2, problem (3.4) is well-posed and, thus, it follows from a dimension argument that there is non-degeneracy of the sesquilinear form a over $V_H \times \widetilde{V}_H$. Thus, the adjoint problem to (3.4) is well-posed with the same stability constant as in Theorem 2.

The quasi-optimality result requires the following additional condition on the oversampling parameter m,

$$m \geq |\log\left(2C_2\sqrt{C_{\mathrm{ol},m}}C_a^2C_{\mathrm{PG}}\gamma(\kappa,\Omega)\sqrt{3/2}\right)|/|\log(\beta)|. \tag{4.10}$$

Theorem 3 (*Quasi-optimality*)**.** *The resolution conditions (4.3) and (4.6) and the oversampling conditions (4.7) and (4.10) imply that the solution u_H to (3.4) with parameters H, h, and m and the solution u_h of the standard Galerkin FEM on the mesh $\mathcal{G}_h$ satisfy*

$$\|u_h - u_H\|_V \lesssim \|(1-I_H)u_h\|_V \approx \min_{v_H \in V_H}\|u_h - v_H\|_V.$$

Proof. Let $e := u_h - u_H$. The triangle inequality and Lemma 1 yield

$$\|e\|_V \leq \|(1-I_H)u_h\|_V + \|I_H e\|_V.$$

It remains to bound the second term on the right-hand side. The proof employs a standard duality argument, the stability of the idealized method and the fact that our practical method is a perturbation of that ideal method. Let $z_H \in V_H$ be the solution to the dual problem

$$(\nabla v_H, \nabla I_H e) + \kappa^2(v_H, I_H e) = a(v_H, (1-\mathcal{C}_\infty)z_H)$$

for all $v_H \in V_H$ (cf. Remark 1). The choice of the test function $v_H = I_H e$ implies that

$$\|I_H e\|_V^2 = a(I_H e, (1-\mathcal{C}_\infty)z_H) = a(I_H e, (\mathcal{C}_m - \mathcal{C}_\infty)z_H) + a(I_H e, (1-\mathcal{C}_m)z_H).$$

The identity $I_H(\mathcal{C}_m - \mathcal{C}_\infty)z_H = 0$, the resolution condition (4.3), the estimate (4.5), and the stability of the adjoint problem imply for the first term on the right-hand side that

$$\begin{aligned} a(I_H e, (\mathcal{C}_m - \mathcal{C}_\infty)z_H) &\leq C_a\sqrt{3/2}\|I_H e\|_V\|\nabla(\mathcal{C}_m - \mathcal{C}_\infty)z_H\|_{L^2(\Omega)} \\ &\leq C_2\sqrt{C_{\mathrm{ol},m}}C_a\sqrt{3/2}\|I_H e\|_V\beta^m\|\nabla z_H\| \\ &\leq C_2\sqrt{C_{\mathrm{ol},m}}C_a^2C_{\mathrm{PG}}\gamma(\kappa,\Omega)\sqrt{3/2}\beta^m\|I_H e\|_V^2. \end{aligned}$$

The condition (4.10) implies that this is $\leq \frac{1}{2}\|I_H e\|_V^2$. The Galerkin orthogonality $a(u_h - u_H, (1-\mathcal{C}_m)z_H) = 0$, the solution property (4.2) of $\mathcal{C}_\infty z_H$, the resolution condition (4.3) and the exponential decay (4.5) imply for the second term

$$\begin{aligned} a(I_H e, (1-\mathcal{C}_m)z_H) &= a(I_H u_h - u_h, (1-\mathcal{C}_m)z_H) \\ &= a(I_H u_h - u_h, (\mathcal{C}_\infty - \mathcal{C}_m)z_H) \\ &\leq \sqrt{3/2}C_a C_2\sqrt{C_{\mathrm{ol},m}}\beta^m\|I_H u_h - u_h\|_V\|\nabla z_H\|_{L^2(\Omega)}. \end{aligned}$$

The stability of the adjoint problem implies

$$\|\nabla z_H\|_{L^2(\Omega)} \leq C_{\mathrm{PG}}\gamma(\kappa, \Omega)C_a\|I_H e\|_V.$$

Thus,

$$a(I_H e, (1-\mathcal{C}_m)z_H) \leq \sqrt{3/2}C_a^2 C_2\sqrt{C_{\mathrm{ol},m}}C_{\mathrm{PG}}\beta^m\gamma(\kappa, \Omega)\|I_H u_h - u_h\|_V\|I_H e\|_V.$$

The term $\|I_H e\|_V$ can be absorbed and the oversampling condition (4.7) implies that $\beta^m\sqrt{C_{\mathrm{ol},m}}\gamma(\kappa, \Omega)$ is controlled by some κ-independent constant. The combination with the foregoing displayed formulae concludes the proof. □

The following consequence of Theorem 3 states an estimate for the error $u - u_H$.

Corollary 1. *Under the conditions of Theorem 3, the discrete solution u_H to (3.4) satisfies with some constant $C \approx 1$ that*

$$\|u - u_H\|_V \leq \|u - u_h\|_V + C\min_{v_H\in V_H}\|u_h - v_H\|_V.$$

In particular, provided that the solution satisfies $u \in W^{1,s}(\Omega)$ for $0 < s \leq 1$, the error decays as $\|u - u_H\|_V \leq \mathcal{O}(H^s)$. ◇

Remark 2. In the idealized case that $m = \infty$, we have $u_h - I_H u_h \in W_h$ and, thus,

$$a(u_h - I_H u_h, (1-\mathcal{C}_\infty)v_H) = 0 \quad \text{for all } v_H \in V_H.$$

Therefore, problem (3.4) and the Galerkin property show that $u_H = I_H u_h$.

5. Numerical experiments

We investigate the method in three numerical experiments. The convergence history plots display the absolute error in the norm $\|\cdot\|_V$ versus the mesh size H.

5.1. Plane wave on the square domain

On the unit square $\Omega = (0, 1)^2$, we consider the pure Robin problem $\Gamma_R = \partial\Omega$ with data given by the plane wave $u(x) = \exp(-i\kappa x \cdot \binom{0.6}{0.8})$.

Fig. 6a–c displays the convergence history for $\kappa = 2^6, 2^7, 2^8$ and the fine-scale mesh parameter $h = 2^{-11}$. The best-approximation error of continuous Q_1 functions in $\|\cdot\|_V$ and the error of the standard Galerkin FEM on the same coarse mesh are plotted for comparison. As expected, the standard FEM clearly exhibits the pollution effect, and larger values of κ increase the discrepancy between the approximation error of the FEM and the theoretical best-approximation by Q_1 functions in the regime under consideration. In contrast, the approximation by the msPGFEM can compete with the best-approximation on meshes that allow a meaningful representation of the solution. We stress the fact that the convergence history plots merely take into account the coarse mesh-size H, but the computational cost in the multiscale method is moderately higher than in the standard FEM due to the increased communication caused by the coupling $m \approx \log(\kappa)$.

For the oversampling parameter $m = 2$, the number of corrector problems to be solved for the finest mesh $\mathcal{G}_H$ is 49 out of 1 048 576 when no symmetry is exploited.

Fig. 6d displays the dependence on the fine mesh parameter h for $\kappa = 2^8$ and oversampling parameter $m = 6$. Since the multiscale method based on the fine grid $\mathcal{G}_h$ computes approximations of the FEM solution on that fine grid, e.g. $u_H = I_H u_h$ for $m = \infty$ as in Remark 2, it is clear that the accuracy of the msPGFEM is limited by the accuracy of the standard FEM on the fine scale. This can be observed in Fig. 6d. It can be also seen that a finer fine-scale mesh-size h improves the error of the msPGFEM towards the best-approximation. In this two-dimensional example, the quasi-optimality constant appears to be close to 1.

Next, we study the dependence on the oversampling parameter m. Fig. 7 displays the convergence history for $\kappa = 2^7$ and $\kappa = 2^8$. The fine mesh parameter is $h = 2^{-11}$ and m varies from $m = 1$ to $m = 6$. It turns out that for the present configuration, the value $m = 2$ is sufficient for quasi-optimality. In the range where H is significantly larger

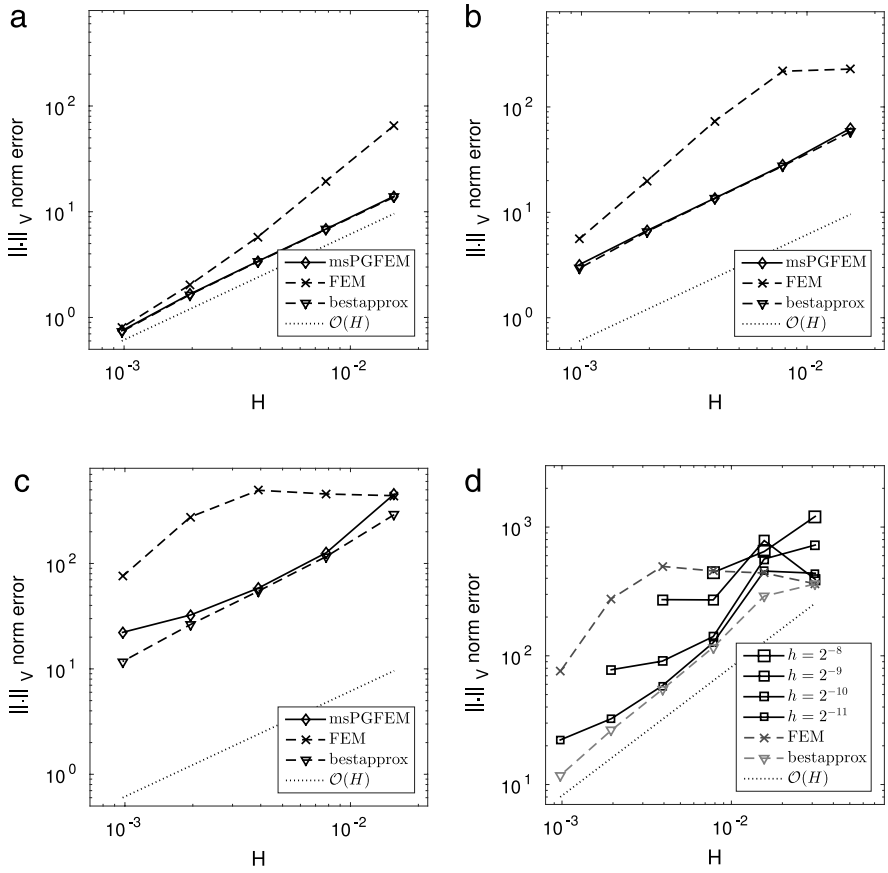


Fig. 6. (a)–(c) Comparison of the msPGFEM with the best approximation in $\|\cdot\|_V$ and the standard Galerkin FEM for the 2D plane wave example for $\kappa = 2^6, 2^7, 2^8$. (d) Dependence on the fine mesh parameter h in the 2D plane wave example with $\kappa = 2^8$.

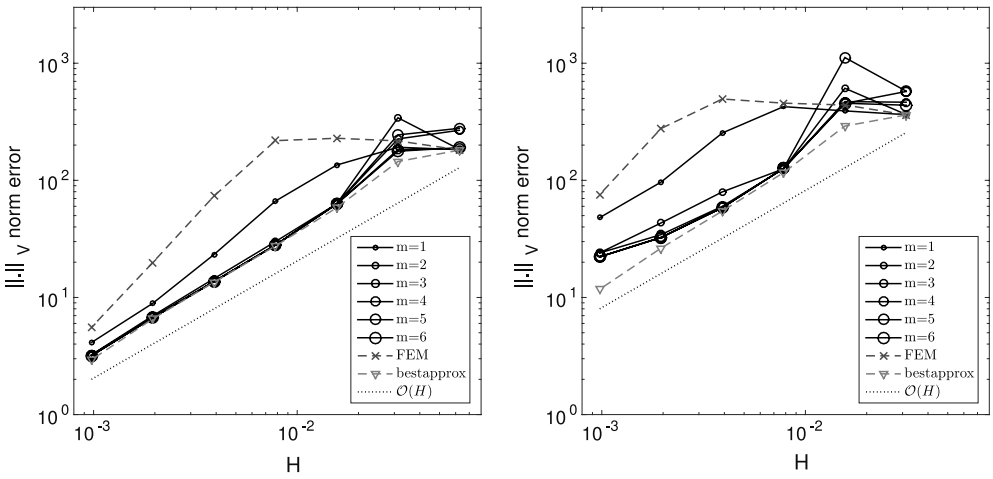


Fig. 7. Convergence history for the 2D plane wave example with $\kappa = 2^7$ (left) and $\kappa = 2^8$ (right), $h = 2^{-11}$ and varying m.

than κ^{-1} and the resolution condition is violated, larger oversampling parameters may lead to larger errors, which is not surprising in view of the lack of decay, see also Fig. 5. This, however, is no more the case as soon as H is small enough to allow for a meaningful representation of the wave.

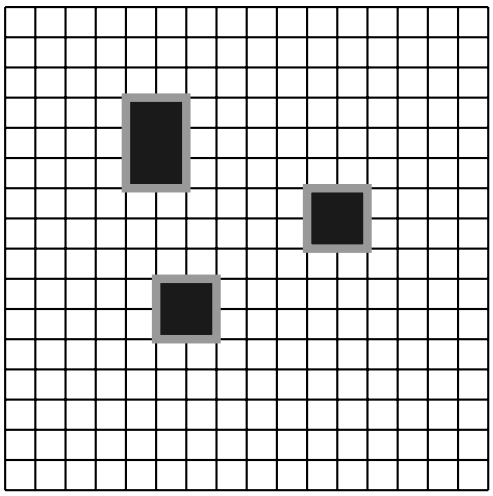


Fig. 8. Coarse mesh for the square domain with three scatterers from Section 5.2.

5.2. Multiple sound-soft scatterers in 2D

We consider the domain

$$\Omega := (0,1)^2 \setminus \left(\left[\frac{5}{16}, \frac{7}{16}\right] \times \left[\frac{5}{16}, \frac{7}{16}\right] \cup \left[\frac{10}{16}, \frac{12}{16}\right] \times \left[\frac{8}{16}, \frac{10}{16}\right] \cup \left[\frac{4}{16}, \frac{6}{16}\right] \times \left[\frac{10}{16}, \frac{13}{16}\right] \right)$$

from Fig. 8. The incident wave $u_{\mathrm{in}}(x) = \exp(-i\kappa x \cdot \binom{0.6}{0.8})$ is incorporated through the Robin boundary condition with $g := i\kappa u_{\mathrm{in}} + \partial_\nu u_{\mathrm{in}}$ on the outer boundary $\Gamma_R := \{x \in \{0,1\} \text{ or } y \in \{0,1\}\}$. On the remaining part of the boundary $\Gamma_D := \partial\Omega \setminus \Gamma_R$ we impose homogeneous Dirichlet conditions. We choose the fine mesh parameter as $h = 2^{-11}$. Since the exact solution is unknown, we compute a reference solution with the standard Q_1 FEM on the fine mesh $\mathcal{G}_h$ and we compare the coarse approximation with this reference solution. Errors committed by the fine scale are not included in the discussion. Fig. 9 displays the convergence history for $\kappa = 2^5$ and $\kappa = 2^6$. The oversampling parameter m varies from $m = 1$ to $m = 4$. As in the foregoing example, the value $m = 2$ for the oversampling parameter seems to be sufficient for the quasi-optimality and even a quasi-optimality constant close to 1 in the range of wave numbers considered here. In particular, the pollution effect that is visible for the standard Galerkin FEM is not present for the msPGFEM. Reduced convergences rates which are expected from the presence of re-entrant corners are not visible in this computational range. For the oversampling parameter $m = 2$, the number of corrector problems to be solved for the finest mesh $\mathcal{G}_H$ is 210 out of 61 952 when no symmetry is exploited.

5.3. Plane wave on the cube domain

On the unit cube $\Omega = (0,1)^3$, we consider the pure Robin problem with data given by the plane wave $u(x) = \exp(-i\kappa x \cdot \frac{1}{\sqrt{38}} \binom{2}{3}{5})$.

We choose $\kappa = 2^5$. Fig. 10 compares the error of the msPGGEM $h = 2^{-4}$ and $m \in \{1,2,3,4\}$ with the best-approximation in the $\|\cdot\|_V$ norm and the error of the standard Galerkin FEM. Also in this example, the msPGFEM is pollution-free for the oversampling parameter $m \geq 2$. The quasi-optimality constant appears slightly larger than in 2D. For the oversampling parameter $m = 2$, the number of corrector problems to be solved for the finest mesh $\mathcal{G}_H$ is 343 out of 262 144 when no symmetry is exploited.

Appendix. Proof of Theorem 1

For the sake of completeness we also present a proof of the exponential decay result Theorem 4 which is central for the method. The idea of the proof is the same as in the previous proofs of the exponential decay [12,29,30,26,31]

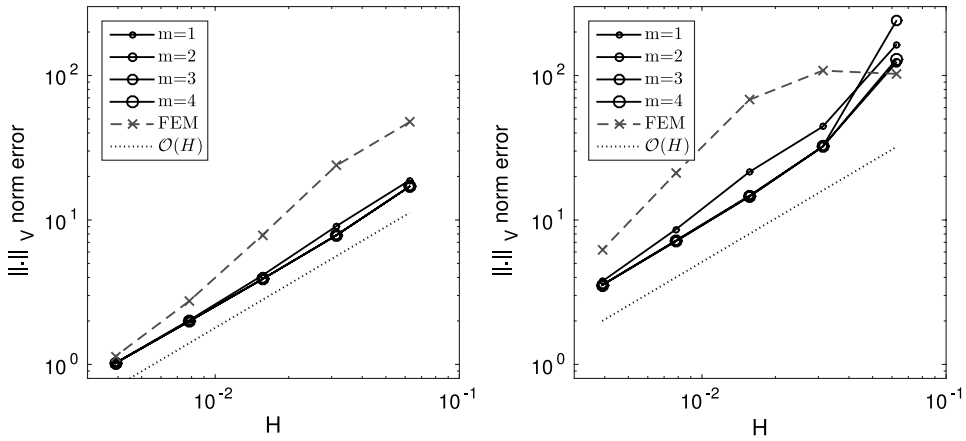


Fig. 9. Convergence history for the multiple scattering example from Section 5.2 for $\kappa = 2^5$ (left) and $\kappa = 2^6$ (right) $h = 2^{-11}$.

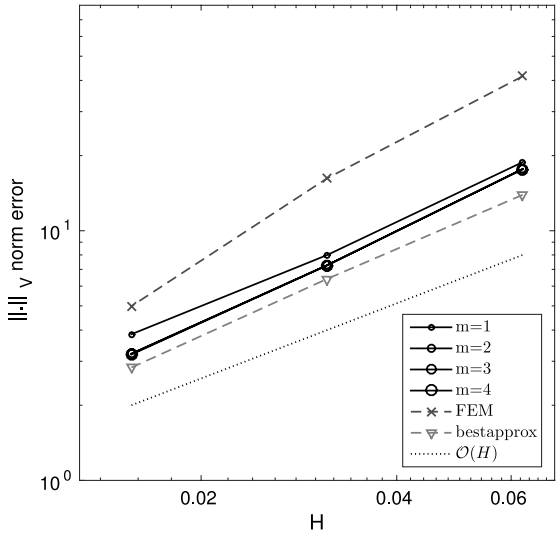


Fig. 10. Convergence history for the 3D plane wave example for $\kappa = 2^5$ and $h = 2^{-7}$.

in the context of diffusion problems. The difference especially with respect to [11] is that here the quasi-interpolation is a projection. This simplifies the proofs and leads to slightly better rates in the exponential decay that have been experimentally observed in [11].

Let $I_h : C^0(\Omega) \to V_h$ denote the nodal Q_1 interpolation operator. Standard interpolation estimates and the inverse inequality prove for any $T \in \mathcal{G}_H$ and all $q \in Q_2(T)$ the stability estimate

$$\|\nabla I_h q\|_{L^2(T)} \leq C_{I_h} \|\nabla q\|_{L^2(T)}. \tag{A.1}$$

In the proofs we will frequently make use of cut-off functions. We collect some properties in the following lemma.

Lemma 2. *Let $\eta \in \mathcal{S}^1(\mathcal{G}_H)$ be a function with values in the interval $[0, 1]$ satisfying the bound*

$$\|\nabla \eta\|_{L^\infty(\Omega)} \leq C_\eta H^{-1} \tag{A.2}$$

and let $\mathcal{R} := \operatorname{supp}(\nabla\eta)$. *Given any subset* $\mathcal{K} \subseteq \mathcal{G}_H$, *any* $\phi \in W_h$ *satisfies for* $S = \cup\mathcal{K} \subseteq \overline{\Omega}$ *that*

$$\|\phi\|_{L^2(S)} \lesssim H\|\nabla\phi\|_{L^2(\mathsf{N}(S))} \tag{A.3}$$

$$\|(1-I_H)I_h(\eta\phi)\|_{L^2(S)} \lesssim H\|\nabla(\eta\phi)\|_{L^2(\mathsf{N}(S))} \tag{A.4}$$

$$\|\nabla(\eta\phi)\|_{L^2(S)} \lesssim \|\nabla\phi\|_{L^2(S\cap\{\operatorname{supp}(\eta)\})} + \|\nabla\phi\|_{L^2(\mathsf{N}(S\cap\mathcal{R}))}. \tag{A.5}$$

Proof. The property (3.1) readily implies (A.3). Furthermore, (3.1) implies

$$\|(1-I_H)I_h(\eta\phi)\|_{L^2(S)} \le HC_{I_H}\sqrt{C_{\mathrm{ol}}}\|\nabla I_h(\eta\phi)\|_{L^2(\mathsf{N}(S))}.$$

Estimate (A.1) leads to

$$\|\nabla I_h(\eta\phi)\|_{L^2(\mathsf{N}(S))} \le C_{I_h}\|\nabla(\eta\phi)\|_{L^2(\mathsf{N}(S))}.$$

This proves (A.4). For the proof of (A.5) the product rule and (A.2) imply

$$\|\nabla(\eta\phi)\|_{L^2(S)} \le \|\nabla\phi\|_{L^2(S\cap\{\operatorname{supp}(\eta)\})} + C_\eta H^{-1}\|\phi\|_{L^2(S\cap\mathcal{R})}.$$

The combination with (A.3) concludes the proof. □

Theorem 4 (*Decay*)**.** *Under the resolution condition* (4.3), *there exists* $0 < \beta < 1$ *such that, for any* $v_H \in V_H$ *and all* $T \in \mathcal{G}_H$ *and* $m \in \mathbb{N}$,

$$\|\nabla\mathcal{C}_{T,\infty}v_H\|_{L^2(\Omega\setminus\mathsf{N}^m(T))} \le C\beta^m\|\nabla v_H\|_{L^2(T)}.$$

Proof. We define the cut-off function $\eta \in \mathcal{S}^1(\mathcal{G}_H)$ via

$$\eta \equiv 0 \quad \text{in } \mathsf{N}^{m-3}(T) \quad \text{and} \quad \eta \equiv 1 \quad \text{in } \Omega\setminus\mathsf{N}^{m-2}(T).$$

Note that η is thereby also uniquely defined on the set $\mathcal{R} := \operatorname{supp}(\nabla\eta)$. The shape-regularity implies that η satisfies (A.2). Let $v_H \in V_H$ and denote $\phi := \mathcal{C}_{T,\infty}v_H \in W_h$. Elementary estimates lead to

$$\begin{aligned}\|\nabla\phi\|^2_{\Omega\setminus\mathsf{N}^m(T)} \le |(\nabla\phi,\eta\nabla\phi)_{L^2(\Omega)}| &\le |(\nabla\phi,\nabla(\eta\phi))_{L^2(\Omega)}| + |(\nabla\phi,\phi\nabla\eta)_{L^2(\Omega)}|\\ &\le M_1+M_2+M_3+M_4\end{aligned}$$

for

$$\begin{aligned}M_1 &:= |(\nabla\phi,\nabla((1-I_h)(\eta\phi)))_{L^2(\Omega)}| & M_2 &:= |(\nabla\phi,\nabla((1-I_H)I_h(\eta\phi)))_{L^2(\Omega)}|\\ M_3 &:= |(\nabla\phi,\nabla(I_HI_h(\eta\phi)))_{L^2(\Omega)}| & M_4 &:= |(\nabla\phi,\phi\nabla\eta)_{L^2(\Omega)}|.\end{aligned}$$

The property (A.1) proves

$$M_1 \le \|\nabla\phi\|_{L^2(\mathcal{R})}\|\nabla(\eta\phi - I_h(\eta\phi))\|_{L^2(\mathcal{R})} \lesssim \|\nabla\phi\|_{L^2(\mathcal{R})}\|\nabla(\eta\phi)\|_{L^2(\mathcal{R})}.$$

Hence, it follows with (A.5) that

$$M_1 \lesssim \|\nabla\phi\|_{L^2(\mathcal{R})}\|\nabla\phi\|_{L^2(\mathsf{N}(\mathcal{R}))}.$$

Since $w := (1-I_H)I_h(\eta\phi) \in W_h$, the identity (4.1) and the fact that the support of w lies outside T imply $a(w,\phi) = a_T(w,v_H) = 0$ and therefore

$$M_2 = a(w,\phi) + \kappa^2(w,\phi) = \kappa^2(w,\phi) \le \kappa^2\|w\|_{L^2(\mathsf{N}(\mathcal{R}))}\|\phi\|_{L^2(\mathsf{N}(\mathcal{R}))}.$$

The estimates (A.3) and (A.4) and the resolution condition $\kappa H \lesssim 1$ from (4.3) imply

$$M_2 \lesssim \|\nabla\phi\|_{L^2(\mathsf{N}^2(\mathcal{R}))}\|\nabla(\eta\phi)\|_{L^2(\mathsf{N}^2(\mathcal{R}))}.$$

The application of (A.5) yields

$$M_2 \lesssim \|\nabla\phi\|_{L^2(\mathsf{N}^2(\mathcal{R}))}(\|\nabla\phi\|_{L^2(\mathsf{N}^2(\mathcal{R}))} + \|\nabla\phi\|_{L^2(\mathsf{N}(\mathcal{R}))}) \lesssim \|\nabla\phi\|^2_{L^2(\mathsf{N}^2(\mathcal{R}))}.$$

The function $I_H I_h(\eta\phi)$ vanishes outside $\mathsf{N}(\mathcal{R})$. Hence, the stability and approximation properties (3.1) and (A.1) lead to

$$\begin{aligned} M_3 &\leq \|\nabla\phi\|_{L^2(\mathsf{N}(\mathcal{R}))}\|\nabla(I_H I_h(\eta\phi))\|_{L^2(\mathsf{N}(\mathcal{R}))} \\ &\lesssim \|\nabla\phi\|_{L^2(\mathsf{N}(\mathcal{R}))}\|\nabla(\eta\phi)\|_{L^2(\mathsf{N}^2(\mathcal{R}))}. \end{aligned}$$

With (A.5) we obtain

$$M_3 \lesssim \|\nabla\phi\|_{L^2(\mathsf{N}(\mathcal{R}))}(\|\nabla\phi\|_{L^2(\mathsf{N}^2(\mathcal{R}))} + \|\nabla\phi\|_{L^2(\mathsf{N}(\mathcal{R}))}) \lesssim \|\nabla\phi\|^2_{L^2(\mathsf{N}^2(\mathcal{R}))}.$$

For the term M_4, the Lipschitz bound (A.2) and (A.3) prove

$$M_4 \leq \|\nabla\phi\|_{L^2(\mathcal{R})}\,\|\phi\|_{L^2(\mathcal{R})} C_\eta H^{-1} \lesssim \|\nabla\phi\|^2_{L^2(\mathsf{N}(\mathcal{R}))}.$$

Altogether, it follows for some constant $\widetilde{C}$ that

$$\|\nabla\phi\|^2_{L^2(\Omega\backslash\mathsf{N}^m(T))} \leq \widetilde{C}\|\nabla\phi\|^2_{L^2(\mathsf{N}^2(\mathcal{R}))}.$$

Recall that $\mathsf{N}^2(\mathcal{R}) = \mathsf{N}^m(T)\backslash\mathsf{N}^{m-5}(T)$. Since

$$\|\nabla\phi\|^2_{L^2(\Omega\backslash\mathsf{N}^m(T))} + \|\nabla\phi\|^2_{L^2(\mathsf{N}^m(T)\backslash\mathsf{N}^{m-5}(T))} = \|\nabla\phi\|^2_{L^2(\Omega\backslash\mathsf{N}^{m-5}(T))},$$

we obtain

$$(1+\widetilde{C}^{-1})\|\nabla\phi\|^2_{L^2(\Omega\backslash\mathsf{N}^m(T))} \leq \|\nabla\phi\|^2_{L^2(\Omega\backslash\mathsf{N}^{m-5}(T))}.$$

The repeated application of this argument proves for $\tilde{\beta} := (1+\widetilde{C}^{-1})^{-1} < 1$ that

$$\|\nabla\phi\|^2_{L^2(\Omega\backslash\mathsf{N}^m(T))} \leq \tilde{\beta}^{\lfloor m/5\rfloor}\|\nabla\phi\|^2_{L^2(\Omega)} \lesssim \tilde{\beta}^{\lfloor m/5\rfloor}\|\nabla v_H\|^2_{L^2(T)}.$$

This is the assertion. □

We proceed with the proof of Theorem 1.

Proof of Theorem 1. We define the cut-off function $\eta \in \mathcal{S}^1(\mathcal{G}_H)$ via

$$\eta \equiv 0 \quad \text{in } \Omega\backslash\mathsf{N}^{m-1}(T) \quad \text{and} \quad \eta \equiv 1 \quad \text{in } \mathsf{N}^{m-2}(T).$$

This function is thereby uniquely defined and satisfies the bound (A.2). Since $(1-I_H)I_h(\eta\mathcal{C}_{T,\infty}v) \in W_h(\Omega_T)$, we deduce with Céa's Lemma, the identity $I_H\mathcal{C}_{T,\infty}v = 0$ and the approximation and stability properties (3.1) and (A.1) and the resolution condition (4.3) that

$$\begin{aligned} \|\nabla(\mathcal{C}_{T,\infty}v - \mathcal{C}_{T,m}v)\|^2_{L^2(\Omega)} &\lesssim \|\mathcal{C}_{T,\infty}v - (1-I_H)I_h(\eta\mathcal{C}_{T,\infty}v)\|^2_V \\ &= \|(1-I_H)I_h(\mathcal{C}_{T,\infty}v - \eta\mathcal{C}_{T,\infty}v)\|^2_{V,\Omega\backslash\{\eta=1\}} \\ &\lesssim \|\nabla(1-\eta)\mathcal{C}_{T,\infty}v\|^2_{L^2(\mathsf{N}(\Omega\backslash\{\eta=1\}))} \\ &\lesssim \|\nabla\mathcal{C}_{T,\infty}v\|^2_{L^2(\mathsf{N}(\Omega\backslash\{\eta=1\}))}. \end{aligned}$$

Note that $\mathsf{N}(\Omega\backslash\{\eta=1\}) = \Omega\backslash\mathsf{N}^{m-3}(T)$. This and Theorem 4 prove (4.4).

Define $z := (\mathcal{C}_\infty - \mathcal{C}_m)v$ and $z_T := (\mathcal{C}_{T,\infty} - \mathcal{C}_{T,m})v$. The ellipticity from Lemma 1 proves

$$\frac{1}{2}\|\nabla z\|^2_{L^2(\Omega)} \leq \left|\sum_{T\in\mathcal{G}_H} a(z, z_T)\right|.$$

We define the cut-off function $\eta \in \mathcal{S}^1(\mathcal{G}_H)$ via

$$\eta \equiv 1 \quad \text{in } \Omega\backslash\mathsf{N}^{m+2}(T) \quad \text{and} \quad \eta \equiv 0 \quad \text{in } \mathsf{N}^{m+1}(T).$$

This function is thereby uniquely defined and satisfies the bound (A.2). For any $T \in \mathcal{G}_H$ we have $(1-I_H)I_h(\eta z) \in W_h$ with support outside Ω_T. Hence, we obtain with $z = I_h z$ that

$$a(z, z_T) = a(I_h(z-\eta z), z_T) + a(I_H I_h(\eta z), z_T).$$

The function $z - I_h(\eta z)$ vanishes on $S := \{\eta = 1\}$. Hence, the first term on the right-hand side satisfies

$$|a(I_h(z-\eta z), z_T)| \leq C_a \|I_h(z-\eta z)\|_{V,\Omega\setminus S} \|z_T\|_V .$$

The Friedrichs inequality with constant C_{F} proves together with the stability (A.1) and the estimate (A.5) applied to the cut-off function $(1-\eta)$ that

$$\|I_h(z-\eta z)\|_{V,\Omega\setminus S} \lesssim \sqrt{1+(C_{\mathrm{F}}\kappa H)^2}\|\nabla z\|_{L^2(\Omega\setminus S)} \lesssim \|\nabla z\|_{L^2(\Omega\setminus S)}.$$

Furthermore, $I_H I_h(\eta z)$ vanishes on $\Omega\setminus \mathsf{N}(\mathrm{supp}(1-\eta))$. Hence, we infer from Friedrichs' inequality and the resolution condition (4.3), the stability properties (3.1) and (A.1) and (A.5) that

$$|a(z_T, I_H I_h(\eta z))| \lesssim \|\nabla z\|_{L^2(\mathsf{N}^2(\mathrm{supp}(1-\eta)))}\|z_T\|_V .$$

The sum over all $T \in \mathcal{G}_H$ and the Cauchy inequality yield with the finite overlap of patches

$$\begin{aligned}\|\nabla z\|^2_{L^2(\Omega)} &\lesssim \sum_{T\in\mathcal{G}_H} \|\nabla z\|_{L^2(\mathsf{N}^2(\mathrm{supp}(1-\eta)))}\|z_T\|_V \\ &\lesssim \sqrt{C_{\mathrm{ol},m}}\|\nabla z\|_{L^2(\Omega)}\sqrt{\sum_{T\in\mathcal{G}_H}\|z_T\|_V^2}.\end{aligned}$$

The combination with (4.4) concludes the proof. □

References

[1] I.M. Babuska, S.A. Sauter, Is the pollution effect of the FEM avoidable for the Helmholtz equation considering high wave numbers? SIAM Rev. 42 (2000) 451–484.

[2] R. Tezaur, C. Farhat, Three-dimensional discontinuous Galerkin elements with plane waves and Lagrange multipliers for the solution of mid-frequency Helmholtz problems, Internat. J. Numer. Methods Engrg. 66 (2006) 796–815.

[3] X. Feng, H. Wu, Discontinuous Galerkin methods for the Helmholtz equation with large wave number, SIAM J. Numer. Anal. 47 (2009) 2872–2896.

[4] X. Feng, H. Wu, *hp*-discontinuous Galerkin methods for the Helmholtz equation with large wave number, Math. Comp. 80 (2011) 1997–2024.

[5] R. Hiptmair, A. Moiola, I. Perugia, Plane wave discontinuous Galerkin methods for the 2D Helmholtz equation: analysis of the *p*-version, SIAM J. Numer. Anal. 49 (2011) 264–284.

[6] J.M. Melenk, S.A. Sauter, Convergence analysis for finite element discretizations of the Helmholtz equation with Dirichlet-to-Neumann boundary conditions, Math. Comp. 79 (2010) 1871–1914.

[7] J.M. Melenk, S. Sauter, Wavenumber explicit convergence analysis for Galerkin discretizations of the Helmholtz equation, SIAM J. Numer. Anal. 49 (2011) 1210–1243.

[8] J. Zitelli, I. Muga, L. Demkowicz, J. Gopalakrishnan, D. Pardo, V. Calo, A class of discontinuous Petrov–Galerkin methods. Part IV: The optimal test norm and time-harmonic wave propagation in 1D, J. Comput. Phys. 230 (2011) 2406–2432.

[9] L. Demkowicz, J. Gopalakrishnan, I. Muga, J. Zitelli, Wavenumber explicit analysis of a DPG method for the multidimensional Helmholtz equation, Comput. Methods Appl. Mech. Engrg. 213/216 (2012) 126–138.

[10] H. Wu, Pre-asymptotic error analysis of CIP-FEM and FEM for the Helmholtz equation with high wave number. Part I: linear version, IMA J. Numer. Anal. 34 (2014) 1266–1288.

[11] D. Peterseim, Eliminating the pollution effect in Helmholtz problems by local subscale correction, 2014. ArXiv e-prints, 1411.1944.

[12] A. Målqvist, D. Peterseim, Localization of elliptic multiscale problems, Math. Comp. 83 (2014) 2583–2603.

[13] T.J.R. Hughes, Multiscale phenomena: Green's functions, the Dirichlet-to-Neumann formulation, subgrid scale models, bubbles and the origins of stabilized methods, Comput. Methods Appl. Mech. Engrg. 127 (1995) 387–401.

[14] T.J.R. Hughes, G.R. Feijóo, L. Mazzei, J.-B. Quincy, The variational multiscale method—a paradigm for computational mechanics, Comput. Methods Appl. Mech. Engrg. 166 (1998) 3–24.

[15] T. Hughes, G. Sangalli, Variational multiscale analysis: the fine-scale Green's function, projection, optimization, localization, and stabilized methods, SIAM J. Numer. Anal. 45 (2007) 539–557.

[16] A. Målqvist, Multiscale methods for elliptic problems, Multiscale Model. Simul. 9 (2011) 1064–1086.

[17] D. Peterseim, Variational multiscale stabilization and the exponential decay of fine-scale correctors, 2015. ArXiv e-prints, 1505.07611.

[18] F. Ihlenburg, Finite element analysis of acoustic scattering, in: Applied Mathematical Sciences, vol. 132, Springer-Verlag, New York, 1998, http://dx.doi.org/10.1007/b98828.

[19] S. Esterhazy, J.M. Melenk, On stability of discretizations of the Helmholtz equation, in: Numerical Analysis of Multiscale Problems, in: Lect. Notes Comput. Sci. Eng., vol. 83, Springer, Heidelberg, 2012, pp. 285–324.

[20] J.M. Melenk, On generalized finite-element methods, ProQuest LLC, Ann Arbor, MI, (Ph.D. thesis), University of Maryland, College Park, 1995.

[21] P. Cummings, X. Feng, Sharp regularity coefficient estimates for complex-valued acoustic and elastic Helmholtz equations, Math. Models Methods Appl. Sci. 16 (2006) 139–160.

[22] C. Makridakis, F. Ihlenburg, I. Babuška, Analysis and finite element methods for a fluid-solid interaction problem in one dimension, Math. Models Methods in Appl. Sci. 06 (1996) 1119–1141.
[23] U. Hetmaniuk, Stability estimates for a class of Helmholtz problems, Commun. Math. Sci. 5 (2007) 665–678.
[24] R. Hiptmair, A. Moiola, I. Perugia, Trefftz discontinuous Galerkin methods for acoustic scattering on locally refined meshes, Appl. Numer. Math. 79 (2014) 79–91.
[25] T. Betcke, S.N. Chandler-Wilde, I.G. Graham, S. Langdon, M. Lindner, Condition number estimates for combined potential integral operators in acoustics and their boundary element discretisation, Numer. Methods Partial Differential Equations 27 (2011) 31–69.
[26] P. Henning, P. Morgenstern, D. Peterseim, Multiscale partition of unity, in: M. Griebel, M.A. Schweitzer (Eds.), Meshfree Methods for Partial Differential Equations VII, in: Lecture Notes in Computational Science and Engineering, vol. 100, Springer, 2014.
[27] D.A. Di Pietro, A. Ern, Mathematical aspects of discontinuous Galerkin methods, in: Mathématiques & Applications (Berlin), vol. 69, Springer, Heidelberg, 2012.
[28] J.M. Melenk, S.A. Sauter, Wave-number explicit convergence analysis for Galerkin discretizations of the Helmholtz equation, SIAM J. Numer. Anal. 49 (2011) 1210–1243.
[29] P. Henning, D. Peterseim, Oversampling for the multiscale finite element method, Multiscale Model. Simul. 11 (2013) 1149–1175.
[30] D. Elfverson, E.H. Georgoulis, A. Målqvist, D. Peterseim, Convergence of a discontinuous Galerkin multiscale method, SIAM J. Numer. Anal. 51 (2013) 3351–3372.
[31] D. Brown, D. Peterseim, A multiscale method for porous microstructures. ArXiv e-prints, 2014.

Appendix C
Eigenvalue problems

C.1 Computation of eigenvalues by numerical upscaling

Numerische Mathematatik **130**(2):337-361, 2015.

(with A. Målqvist)

Numer. Math. (2015) 130:337–361
DOI 10.1007/s00211-014-0665-6

Numerische Mathematik

Computation of eigenvalues by numerical upscaling

Axel Målqvist · Daniel Peterseim

Received: 22 August 2013 / Revised: 31 January 2014 / Published online: 20 September 2014

Abstract We present numerical upscaling techniques for a class of linear second-order self-adjoint elliptic partial differential operators (or their high-resolution finite element discretization). As prototypes for the application of our theory we consider benchmark multi-scale eigenvalue problems in reservoir modeling and material science. We compute a low-dimensional generalized (possibly mesh free) finite element space that preserves the lowermost eigenvalues in a superconvergent way. The approximate eigenpairs are then obtained by solving the corresponding low-dimensional algebraic eigenvalue problem. The rigorous error bounds are based on two-scale decompositions of $H_0^1(\Omega)$ by means of a certain Clément-type quasi-interpolation operator.

Mathematics Subject Classification 65N30 · 65N25 · 65N15

1 Introduction

This paper presents and analyzes a novel numerical upscaling technique for computing eigenpairs of self-adjoint linear elliptic second order differential operators with arbitrary positive bounded coefficients. The precise setting of the paper is as follows.

A. Målqvist is supported by The Göran Gustafsson Foundation and The Swedish Research Council and D. Peterseim was partially supported by the DFG Research Center Matheon Berlin through project C33.

A. Målqvist
Department of Mathematical Sciences, Chalmers University of Technology and University of Gothenburg, 412 96 Göteborg, Sweden

D. Peterseim ( )
Institute for Numerical Simulation, Rheinische Friedrich-Wilhelms-Universität Bonn, Wegelerstr. 6, 53115 Bonn, Germany
e-mail: peterseim@ins.uni-bonn.de

Let $\Omega \subset \mathbb{R}^d$ be a bounded polyhedral Lipschitz domain and let $A \in L^\infty(\Omega, \mathbb{R}^{d\times d}_{\mathrm{sym}})$ be a matrix-valued coefficient with uniform spectral bounds $0 < \alpha \leq \beta < \infty$,

$$\sigma(A(x)) \subset [\alpha, \beta] \tag{1.1}$$

for almost all $x \in \Omega$. We want to approximate the eigenvalues of the prototypical operator $-\operatorname{div}(A\nabla\bullet)$. The corresponding eigenproblem in variational formulation reads: find pairs consisting of an eigenvalue $\lambda \in \mathbb{R}$ and associated non-trivial eigenfunction $u \in V := H^1_0(\Omega)$ such that

$$a(u, v) := \int_\Omega (A\nabla u) \cdot \nabla v \,\mathrm{d}x = \lambda \int_\Omega uv \,\mathrm{d}x =: \lambda(u, v)_{L^2(\Omega)} \tag{1.2}$$

for all $v \in V$. We are mainly interested in the lowermost eigenvalues of (1.2) or, more precisely, in the lowermost eigenvalues of the discretized problem: find $\lambda_h \in \mathbb{R}$ and associated non-trivial eigenfunctions $u_h \in V_h \subset V$ such that

$$a(u_h, v) = \lambda_h(u_h, v)_{L^2(\Omega)} \quad \text{for all } v \in V_h. \tag{1.2.h}$$

Here and throughout the paper, the discrete space $V_h \subset V$ shall be a conforming finite element space of dimension N_h based on some regular finite element mesh $\mathcal{T}_h$ of width h.

Popular approaches for the computation of these eigenvalues include Lanczos/Arnoldi-type iterations (as implemented, e.g., in [23]) or the QR-algorithm applied directly to the N_h-dimensional finite element matrices. If a certain structure of the discretization can be exploited (e.g., a hierarchy of finite element meshes and/or spaces) some preconditioned outer iteration for the eigenvalue approximation may be performed and linear problems are solved (approximately) in every iteration step [13,19,20]; see also [1] and [27] and references therein.

Our aim is to avoid the application of any eigenvalue solver to the fine scale discretization (1.2.h) directly. We introduce a second, coarser discretization scale $H > h$ instead. On the corresponding coarse mesh $\mathcal{T}_H$, we compute a generalized finite element space V_{c} of dimension $N_H \ll N_h$. The solutions $(\lambda_H, u_{\mathrm{c}}) \in \mathbb{R} \times V_{\mathrm{c}}$ of

$$a(u_{\mathrm{c}}, v) = \lambda_H(u_{\mathrm{c}}, v)_{L^2(\Omega)} \quad \text{for all } v \in V_{\mathrm{c}}, \tag{1.2.H}$$

then yield accurate approximations of the first N_H eigenpairs of (1.2.h) and, hence, of the first N_H eigenpairs of (1.2) (provided that V_h is properly chosen).

The computation of the coarse space V_{c} involves the (approximate) solution of N_H linear equations on the fine scale (one per coarse node). We emphasize that these linear problems are completely independent of each other. They can be computed in parallel without any communication.

The error $\lambda_H - \lambda_h$ between corresponding eigenvalues of (1.2.H) and (1.2.h), i.e., the error committed by the upscaling from the fine discretization scale h to the coarse discretization scale H, is expressed in terms of H. Without any assumptions on the smoothness of the eigenfunctions of (1.2) or (1.2.h), we prove that these errors

are at least of order H^4. Note that a standard first-order conforming finite element computation on the coarse scale yields accuracy H^2 under full $\mathrm{H}^2(\Omega)$ regularity, see e.g. [22]. Since our estimates are both, of high order (at least H^4) and independent of the underlying regularity, the accuracy of our approximation may actually suffice to fall below the error $\lambda_h - \lambda$ of the fine scale discretization which is of order Ch^{2s} where both the constant C and the exponent $s \in [0, 1]$ depend on the regularity of the data (convexity of Ω, differentiability and variability of A) in a crucial way.

The idea of employing a two-level techniques for the acceleration of eigensolvers is not new. The two-grid method of [37] allows certain post-processing (solution of linear problems on the fine scale). For standard first-order conforming finite element coarse spaces, this technique decreases the eigenvalue error from H^2 to H^4 (up to fine scale errors as above) if the corresponding eigenfunctions are $\mathrm{H}^2(\Omega)$-regular. The regularity assumption is essential and not justified on non-convex domains or for heterogeneous and highly variable coefficients. However, the post-processing technique applies as well to the generalized finite element coarse space V_c and yields eigenvalue errors of order H^6 without any regularity assumptions.

In cases with singular eigenfunctions (due to re-entrant corners in the domain or isolated jumps of the coefficient), one might as well use modern mesh-adaptive algorithms driven by some a posteriori error estimator as proposed and analyzed, e.g., in [3,5,6,10–12,22,24,26]. We are not competing with these efficient algorithms. However, adaptive mesh refinement has its limitations. For instance, if the diffusion coefficient A is highly variable on microscopic scales, the mesh width has to be sufficiently small to resolve these variations [31]. For problems in geophysics or material sciences with characteristic geometric features on microscopic length scales, this so-called resolution condition is often so restrictive that the initial mesh must be chosen very fine and further refinement exceeds computer capacity. Our method is especially designed for such situations which require coarsening rather than refinement.

A particular application of our methodology is the computation of ground states of Bose–Einstein condensates as solutions of the Gross–Pitaevskii equation. Here, certain resolution (small h) is required in order to ensure unique solvability of the discrete nonlinear eigenvalue problem. It is already exposed in [16] that our upscaling approach leads to a significant speed-up in computational time because the expensive iterative solver for the non-linear eigenproblem needs to be applied solely on a space of very low dimension.

The main tools in this paper are localizable orthogonal decompositions of $\mathrm{H}^1_0(\Omega)$ (or its subspace V_h) into coarse and fine parts. These decompositions are presented in Sect. 3. The two-level method for the approximation of eigenvalues is presented in Sect. 4. Section 5 contains its error analysis. The efficient local approximation of the coarse space, the generalization to non-nested grids, a post-processing technique, and further complexity issues are discussed in Sect. 6. Finally, Sect. 7 demonstrates the performance of the method in numerical experiments.

In the remaining part of this paper, we will frequently make use of the notation $b_1 \lesssim b_2$ which abbreviates $b_1 \leq Cb_2$, with some multiplicative constant $C > 0$ which only depends on the domain Ω and the parameter γ (cf. (2.1) below) that measures the quality of some underlying finite element mesh. We emphasize that the C does *not*

depend on the mesh sizes H, h, the eigenvalues, or the coefficient A. Furthermore, $b_1 \approx b_2$ abbreviates $b_1 \lesssim b_2 \lesssim b_1$.

2 Finite element spaces and quasi-interpolation

This section presents some preliminaries on finite element meshes, spaces, and interpolation.

2.1 Finite element meshes

We consider two discretization scales $H > h > 0$. Let $\mathcal{T}_H$ (resp. $\mathcal{T}_h$) denote corresponding regular (in the sense of [8]) finite element meshes of Ω into closed simplices with mesh-size functions $0 < H \in L^\infty(\Omega)$ defined by $H|_T = \operatorname{diam} T =: H_T$ for all $T \in \mathcal{T}_H$ (resp. $0 < h \in L^\infty(\Omega)$ defined by $h|_t = \operatorname{diam} t =: h_t$ for all $t \in \mathcal{T}_h$). The mesh sizes may vary in space but we will not exploit the possible mesh adaptivity in this paper.

The error bounds, typically, depend on the maximal mesh sizes $\|H\|_{L^\infty(\Omega)}$. If no confusion seems likely, we will use H also to denote the maximal mesh size instead of writing $\|H\|_{L^\infty(\Omega)}$. For the sake of simplicity we assume that $\mathcal{T}_h$ is derived from $\mathcal{T}_H$ by some regular, possibly non-uniform, mesh refinement. However, this condition is not essential and Sect. 6.2 will discuss possible generalizations.

As usual, the error analysis depends on the constant $\gamma > 0$ which represents the shape regularity of the finite element mesh $\mathcal{T}_H$;

$$\gamma := \max_{T \in \mathcal{T}_H} \gamma_T \quad \text{with} \quad \gamma_T := \frac{\operatorname{diam} T}{\operatorname{diam} B_T} \quad \text{for } T \in \mathcal{T}_H, \tag{2.1}$$

where B_T denotes the largest ball contained in T.

2.2 Finite element spaces

The first-order conforming finite element space corresponding to $\mathcal{T}_H$ is given by

$$V_H := \{v \in V \mid \forall T \in \mathcal{T}_H, v|_T \text{ is a polynomial of total degree} \leq 1\}. \tag{2.2}$$

Let $\mathcal{N}_H$ denote the set of interior vertices of $\mathcal{T}_H$. For every vertex $z \in \mathcal{N}_H$, let $\phi_z \in V_H$ denote the corresponding nodal basis function (tent/hat function) determined by nodal values

$$\phi_z(z) = 1 \ \text{ and } \ \phi_z(y) = 0 \quad \text{for all } y \neq z \in \mathcal{N}_H.$$

These nodal basis functions form a basis of V_H. The dimension of V_H equals the number of interior vertices,

$$N_H := \dim V_H = |\mathcal{N}_H|.$$

Let $V_h \supset V_H$ denote some conforming finite element space corresponding to the fine mesh $\mathcal{T}_h$. It can be the space of continuous piecewise affine functions on the fine mesh or any other (generalized) finite element space that contains V_H, e.g., the space of continuous p-th order piecewise polynomials as in [33]. By $N_h := \dim V_h$ we denote the dimension of V_h. For standard choices of V_h, this dimension is proportional to the number of interior vertices in the fine mesh $\mathcal{T}_h$.

2.3 Quasi-interpolation

The key tool in our construction will be the bounded linear surjective Clément-type (quasi-)interpolation operator $\mathcal{I}_H : \mathrm{H}^1_0(\Omega) \to V_H$ presented and analyzed in [9]. Given $v \in \mathrm{H}^1_0(\Omega)$, $\mathcal{I}_H v := \sum_{z\in\mathcal{N}_H} (\mathcal{I}_H v)(z)\phi_z$ defines a (weighted) Clément interpolant with nodal values

$$(\mathcal{I}_H v)(z) := \frac{(v, \phi_z)_{L^2(\Omega)}}{(1, \phi_z)_{L^2(\Omega)}} \tag{2.3}$$

for $z \in \mathcal{N}_H$. The nodal values are weighted averages of the function over nodal patches $\omega_z := \operatorname{supp}\phi_z$. Recall the (local) approximation and stability properties of the interpolation operator $\mathcal{I}_H$ [9]: There exists a generic constant $C_{\mathcal{I}_H}$ such that for all $v \in \mathrm{H}^1_0(\Omega)$ and for all $T \in \mathcal{T}_H$ it holds

$$H_T^{-1}\|v - \mathcal{I}_H v\|_{L^2(T)} + \|\nabla(v - \mathcal{I}_H v)\|_{L^2(T)} \le C_{\mathcal{I}_H}\|\nabla v\|_{L^2(\omega_T)}, \tag{2.4}$$

where $\omega_T := \cup\{K \in \mathcal{T}_H \mid T \cap K \neq \emptyset\}$. The constant $C_{\mathcal{I}_H}$ depends on the shape regularity parameter γ of the finite element mesh $\mathcal{T}_H$ (see (2.1) above) but not on H_T.

Note that there exists a constant $C_{\mathrm{ol}} > 0$ that only depends on γ such that the number of elements covered by ω_T is uniformly bounded (w.r.t. T) by C_{ol},

$$\max_{T\in\mathcal{T}_H} |\{K \in \mathcal{T}_H \mid K \subset \omega_T\}| \le C_{\mathrm{ol}}. \tag{2.5}$$

Both constant, $C_{\mathcal{I}_H}$ and C_{ol}, may be hidden in the notation "$\lesssim$" introduced at the end of Sect. 1.

3 Two-scale decompositions

Two-scale decompositions of functions $u \in V_h$ into some macroscopic/coarse part u_{c} plus some microscopic/fine part u_{f} with a certain orthogonality relation are at the very heart of this paper. The macroscopic or coarse part will be an element of a low-dimensional (classical or generalized) finite element space based on some coarse finite element mesh. The microscopic or fine part may oscillate on fine scales that cannot be represented on the coarse mesh.

We stress that all subsequent results are valid even if $h = 0$, i.e., if V_h is replaced with $V = H^1_0(\Omega)$. Actually, the structure of V_h being the space of continuous piecewise polynomials is never exploited. As far as the theory is concerned, V_h could be any space (finite or infinite dimensional) that satisfies $V_H \subsetneq V_h \subseteq H^1_0(\Omega)$.

The initial coarse space V_H may as well be generalized. This will be discussed in Sect. 6.2.

3.1 L^2-orthogonal two-scale decomposition

We define the fine scale space

$$V_{\mathrm{f}} := \mathrm{kernel}\left(\mathcal{I}_H|_{V_h}\right) \subset V_h,$$

which will take over the role of the microscopic/fine part in all subsequent decompositions.

Our particular choice of a quasi-interpolation operator gives rise to the following orthogonal decomposition. Remember that $(\bullet,\bullet)_{L^2(\Omega)} := \int_\Omega \bullet\bullet\,\mathrm{d}x$ abbreviates the canonical scalar product in $L^2(\Omega)$ and let $\|\bullet\| := \sqrt{(\bullet,\bullet)_{L^2(\Omega)}}$ abbreviate the corresponding norm of $L^2(\Omega)$.

Lemma 3.1 (L^2-orthogonal two-scale decomposition) *Any function $u \in V_h$ can be decomposed uniquely into the sum of $u_H := \mathcal{I}_H|_{V_H}^{-1}(\mathcal{I}_H u) \in V_H$ and $u_{\mathrm{f}} := u - u_H \in V_{\mathrm{f}}$ with*

$$(u_H, u_{\mathrm{f}})_{L^2(\Omega)} = 0. \tag{3.1}$$

The orthogonality implies stability in the sense of

$$\|u_H\|^2 + \|u_{\mathrm{f}}\|^2 = \|u\|^2.$$

Proof of Lemma 3.1 It is easily verified that the restriction of $\mathcal{I}_H$ on the finite element space V_H is invertible. This yields the decomposition.

For the proof of orthogonality, let $v_H = \sum_{z\in\mathcal{N}_H} v_H(z)\phi_z \in V_H$ and $v_{\mathrm{f}} \in V_{\mathrm{f}}$ be arbitrary. Since $\mathcal{I}_H v_{\mathrm{f}} = 0$, we have that $(\phi_z, v_{\mathrm{f}})_{L^2(\Omega)} = (\mathcal{I}_H v_{\mathrm{f}})(z)\int_\Omega \phi_z\,\mathrm{d}x = 0$ for all $z \in \mathcal{N}_H$. This yields

$$(v_H, v_{\mathrm{f}})_{L^2(\Omega)} = \sum_{z\in\mathcal{N}_H} v_H(z)(\phi_z, v_{\mathrm{f}})_{L^2(\Omega)} = 0$$

and shows that V_H and V_{f} are orthogonal subspaces of V_h. □

We may rewrite Lemma 3.1 as

$$V_h = V_H \oplus V_{\mathrm{f}} \quad\text{and}\quad (V_H, V_{\mathrm{f}})_{L^2(\Omega)} = 0. \tag{3.2}$$

Remark 3.1 (L^2-projection onto the finite element space) Note that the operator $\mathcal{I}_H$ is well-defined as a mapping from $L^2(\Omega)$ onto V_H. In particular, it is stable in the sense that for any $v \in L^2(\Omega)$, it holds that $\|\mathcal{I}_H v\| \lesssim \|v\|$. From the arguments of Lemma 3.1 one easily verifies that the L^2-orthogonal projection $\Pi_{V_H}^{L^2} : L^2(\Omega) \to V_H$ onto the

finite element space V_H may be characterized via the modified Clément interpolation (2.3),

$$\Pi_{V_H}^{L^2} = \mathcal{I}_H|_{V_H}^{-1}\mathcal{I}_H.$$

Furthermore, it holds $V_\mathrm{f} = \mathrm{kernel}(\Pi_{V_H}^{L^2}|_{V_h})$, i.e., V_f might as well be characterized via $\Pi_{V_H}^{L^2}$. This does not change the method. For theoretical purposes, we prefer to work with $\mathcal{I}_H$ because it is a local operator.

3.2 a-Orthogonal two-scale decomposition

The orthogonalization of the decomposition (3.2) with respect to the scalar product $a(\bullet, \bullet) := \int_\Omega (A\nabla\bullet)\cdot\nabla\bullet\,\mathrm{d}x$ yields the definition of a generalized finite element space V_c, that is the a-orthogonal complement of V_f in V_h. Given $v \in V_h$, define the a-orthogonal fine scale projection operator $\mathcal{P}_\mathrm{f} v \in V_\mathrm{f}$ by

$$a(\mathcal{P}_\mathrm{f} v, w) = a(v, w) \quad \text{for all } w \in V_\mathrm{f}.$$

We define the energy norm $|||\bullet||| := \sqrt{a(\bullet,\bullet)}$ (the norm induced by the scalar product a).

Lemma 3.2 (a-orthogonal two-scale decomposition) *Any function $u \in V_h$ can be decomposed uniquely into $u = u_\mathrm{c} + u_\mathrm{f}$, where*

$$u_\mathrm{c} := (1 - \mathcal{P}_\mathrm{f})u \in (1 - \mathcal{P}_\mathrm{f})V_H =: V_\mathrm{c}$$

and

$$u_\mathrm{f} := \mathcal{P}_\mathrm{f} u \in V_\mathrm{f} = \mathrm{kernel}(\mathcal{I}_H|_{V_h}).$$

The decomposition is orthogonal

$$a(u_\mathrm{c}, u_\mathrm{f}) = 0, \tag{3.3}$$

and, hence, stable in the sense of

$$|||u_\mathrm{c}|||^2 + |||u_\mathrm{f}|||^2 = |||u|||^2. \tag{3.4}$$

In other words,

$$V_h = V_\mathrm{c} \oplus V_\mathrm{f} \quad \text{and} \quad a(V_\mathrm{c}, V_\mathrm{f}) = 0. \tag{3.5}$$

We shall emphasize at this point that the decompositions in Lemma 3.1 and Lemma 3.2 are different in general. In particular, the fine scale part v_f may not be the same.

The orthogonalization procedure (with respect to $a(\bullet,\bullet)$) does not preserve the L^2-orthogonality. However, the key observation of this section is that the resulting decomposition (3.5) is almost orthogonal in $L^2(\Omega)$.

Theorem 3.3 (L^2-quasi-orthogonality of the a-orthogonal decomposition) *The decomposition $V_h = V_\mathrm{c} \oplus V_\mathrm{f}$ from Lemma 3.2 is L^2-quasi-orthogonal in the sense that for all $v_\mathrm{c} \in V_\mathrm{c}$ and all $v_\mathrm{f} \in V_\mathrm{f}$, it holds*

$$(v_\mathrm{c}, v_\mathrm{f})_{L^2(\Omega)} \lesssim H^2 \|\nabla v_\mathrm{c}\| \|\nabla v_\mathrm{f}\| \leq \alpha^{-1} H^2 |||v_\mathrm{c}||| |||v_\mathrm{f}|||. \tag{3.6}$$

The decomposition is stable in the sense that

$$\|v_\mathrm{c}\|^2 + \|H^{-1} v_\mathrm{f}\|^2 \lesssim \alpha^{-1} |||v_\mathrm{c} + v_\mathrm{f}|||^2. \tag{3.7}$$

Proof Given any $v_\mathrm{c} \in V_\mathrm{c}$ and $v_\mathrm{f} \in V_\mathrm{f}$, Lemma 3.1 implies that

$$(\mathcal{I}_H v_\mathrm{c}, v_\mathrm{f})_{L^2(\Omega)} = 0.$$

Since $\mathcal{I}_H v_\mathrm{f} = 0$, the Cauchy–Schwarz inequality, (2.4), and (2.5) yield

$$(v_\mathrm{c}, v_\mathrm{f})_{L^2(\Omega)} = (v_\mathrm{c} - \mathcal{I}_H v_\mathrm{c}, v_\mathrm{f} - \mathcal{I}_H v_\mathrm{f})_{L^2(\Omega)} \lesssim H^2 \|\nabla v_\mathrm{c}\| \|\nabla v_\mathrm{f}\|. \tag{3.8}$$

This is the quasi-orthogonality. The same arguments show that

$$\begin{aligned}(H^{-1} v_\mathrm{f}, H^{-1} v_\mathrm{f})_{L^2(\Omega)} &= \left(H^{-1}(v_\mathrm{f} - \mathcal{I}_H v_\mathrm{f}), H^{-1}(v_\mathrm{f} - \mathcal{I}_H v_\mathrm{f})\right)_{L^2(\Omega)} \\ &\lesssim \sum_{T \in \mathcal{T}_H} \|\nabla v_\mathrm{f}\|^2_{L^2(\omega_T)} \\ &\lesssim \alpha^{-1} |||v_\mathrm{f}|||^2.\end{aligned}$$

This, Friedrichs' inequality

$$\|v_\mathrm{c}\| \leq \pi^{-1} \operatorname{diam} \Omega \|\nabla v_\mathrm{c}\|,$$

and (3.4) readily prove the stability estimate. □

4 Upscaled approximation of eigenvalues and eigenfunctions

This section presents a new scheme for the approximation of eigenvalues and eigenfunctions of (1.2.h) or (1.2). Section 4.1 recalls the variational formulation and some characteristic properties of the problem. The new upscaled approximation is then introduced in Sect. 4.2.

4.1 Variational formulation and fine scale discretization

For problem (1.2), there exists a countable number of eigenvalues $\lambda^{(\ell)}$ ($\ell \in \mathbb{N}$) and corresponding eigenfunctions $u^{(\ell)} \in V$. Recall their characterization as solutions of

the variational problem

$$a(u^{(\ell)}, v) = \lambda^{(\ell)}(u^{(\ell)}, v)_{L^2(\Omega)} \quad \text{for all } v \in V. \tag{4.1}$$

Since A is symmetric, all eigenvalues are real and positive. They can be sorted ascending

$$0 < \lambda^{(1)} \leq \lambda^{(2)} \leq \lambda^{(3)} \leq \cdots .$$

Depending on the actual domain Ω and the coefficient A, there may be multiple eigenvalues. A multiple eigenvalue is repeated several times according to its multiplicity in the enumeration above. Let $u^{(\ell)}$ ($\ell \in \mathbb{N}$) be normalized to one in $L^2(\Omega)$, i.e., $\|u^{(\ell)}\| = 1$. It is well known that the eigenfunctions enjoy (or, in the case of multiple eigenvalues, may be chosen such that they fulfill) the orthogonality constraints

$$a(u^{(\ell)}, u^{(m)}) = (u^{(\ell)}, u^{(m)})_{L^2(\Omega)} = 0 \quad \text{if } \ell \neq m. \tag{4.2}$$

The Rayleigh–Ritz discretization of (4.1) with respect to the fine scale finite element space V_h reads: find $\lambda_h^{(\ell)} \in \mathbb{R}$ and non-trivial $u_h^{(\ell)} \in V_h$ such that

$$a(u_h^{(\ell)}, v) = \lambda_h^{(\ell)}(u_h^{(\ell)}, v)_{L^2(\Omega)} \quad \text{for all } v \in V_h. \tag{4.3}$$

Since V_h is a finite-dimensional subspace of V, we can order the discrete eigenvalues similar as the original ones

$$0 < \lambda_h^{(1)} \leq \lambda_h^{(2)} \leq \lambda_h^{(3)} \leq \cdots \leq \lambda_h^{(N_h)}.$$

Again, multiple eigenvalues are repeated according to their multiplicity. Let $u_h^{(\ell)}$ ($\ell = 1, 2, \ldots, N_h$) be normalized to one in $L^2(\Omega)$, i.e., $\|u_h^{(\ell)}\| = 1$. The discrete eigenfunctions satisfy (or, in the case of multiple eigenvalues, can be chosen such that they satisfy) the orthogonality constraints

$$a\left(u_h^{(\ell)}, u_h^{(m)}\right) = \left(u_h^{(\ell)}, u_h^{(m)}\right)_{L^2(\Omega)} = 0 \quad \text{if } \ell \neq m. \tag{4.4}$$

We do not intend to solve the fine scale eigenproblem (4.3). We aim to approximate its eigenpairs $(\lambda_h^{(\ell)}, u_h^{(\ell)})$ with the help of the coarse space V_c defined in Lemma 3.2.

4.2 Coarse scale discretization

Recall the definition of the coarse space

$$V_\mathrm{c} := (1 - \mathcal{P}_\mathrm{f})V_H$$

from Lemma 3.2. This means that V_{c} is the image of V_H under the projection operator $1-\mathcal{P}_{\mathrm{f}}$, where $\mathcal{P}_{\mathrm{f}}$ is the a-orthogonal projection onto the space

$$V_{\mathrm{f}} := \{v \in V_h \,|\, \mathcal{I}_H v = 0\}.$$

Since the intersection of V_H and V_{f} is the trivial subspace (cf. Lemma 3.1), it holds

$$\dim V_{\mathrm{c}} = \dim V_H = N_H.$$

Moreover, the images of the nodal basis functions ϕ_z ($z \in \mathcal{N}_H$) under $(1-\mathcal{P}_{\mathrm{f}})$ yield a basis of V_{c},

$$V_{\mathrm{c}} = \operatorname{span}\{(1-\mathcal{P}_{\mathrm{f}})\phi_z \mid z \in \mathcal{N}_H\}. \tag{4.5}$$

In order to actually compute those basis functions, we need to approximate N_H solutions $\psi_z = \mathcal{P}_{\mathrm{f}}\phi_z \in V_{\mathrm{f}}$ of

$$a(\psi_z, v) = a(\phi_z, v) \quad \text{for all } v \in V_{\mathrm{f}}. \tag{4.6}$$

These problems are linear. The only difference to a standard Poisson problem is that there are some linear constraints hidden in the space V_{f}, that is, the quasi-interpolation of trial and test functions vanishes. In practice, these constraints are realized using Lagrange multipliers.

The linear problems (4.6) may be solved in parallel. Moreover, Sect. 6.1 below will show that these linear problems may be restricted to local subdomains of diameter $\approx |\log(H)|H$ centered around the coarse vertex z, so that the complexity of solving all corrector problems exceeds the cost of solving one linear Poisson problem on the fine mesh only by a factor that depends algebraically on $|\log(H)|$.

The Rayleigh–Ritz discretization of (4.3) [and (4.1)] with respect to the generalized finite element space V_{c} reads: find $\lambda_H^{(\ell)} \in \mathbb{R}$ and non-trivial $u_{\mathrm{c}}^{(\ell)} \in V_{\mathrm{c}}$ such that

$$a\left(u_{\mathrm{c}}^{(\ell)}, v\right) = \lambda_H^{(\ell)} \left(u_{\mathrm{c}}^{(\ell)}, v\right)_{L^2(\Omega)} \quad \text{for all } v \in V_{\mathrm{c}}. \tag{4.7}$$

The assembly of the corresponding finite element stiffness and mass matrices requires only the evaluation of the corrector functions $\psi_z = \mathcal{P}_{\mathrm{f}}\phi_z \in V_{\mathrm{f}}$ computed previously. In general, these matrices are not sparse. However, either the dimension of the coarse problem $N_H \ll N_h$ is so small that the lack of sparsity is not an issue or the matrices may be approximated by sparse matrices with negligible loss of accuracy (see Sect. 6.1 below).

The discrete eigenvalues are ordered (multiple eigenvalues are repeated according to their multiplicity)

$$0 < \lambda_H^{(1)} \le \lambda_H^{(2)} \le \lambda_H^{(3)} \le \cdots \le \lambda_H^{(N_H)}.$$

Let also $u_{\mathrm{c}}^{(\ell)}$ ($\ell = 1, 2, \ldots, N_H$) be normalized to one in $L^2(\Omega)$, i.e., $(u_{\mathrm{c}}^{(\ell)}, u_{\mathrm{c}}^{(\ell)})_{L^2(\Omega)} = 1$. The discrete eigenfunctions satisfy (or, in the case of multiple eigenvalues, can

be chosen such that they satisfy) the orthogonality constraints

$$a\left(u_{\mathrm{c}}^{(\ell)}, u_{\mathrm{c}}^{(m)}\right) = \left(u_{\mathrm{c}}^{(\ell)}, u_{\mathrm{c}}^{(m)}\right)_{L^2(\Omega)} = 0 \quad \text{if } \ell \neq m. \tag{4.8}$$

5 Error analysis

In the subsequent paragraphs we will present error bounds for the approximate eigenvalues and eigenfunctions based on the variational techniques from [34] (which are based on [2] on their part); see also [4].

5.1 Two-scale decomposition revisited

The eigenfunctions allow a different (with respect to Sect. 3) characterization of a macroscopic function, that is, any function spanned by eigenfunctions related to the ℓ lowermost eigenvalues. Define

$$E_\ell := \operatorname{span}\left\{u_h^{(1)}, \ldots, u_h^{(\ell)}\right\}. \tag{5.1}$$

We will have a closer look at the quasi-orthogonality result of Lemma 3.2 given some macroscopic function $u \in E_\ell$.

Lemma 5.1 (L^2-quasi-orthogonality of the a-orthogonal decomposition of macroscopic functions) *Let $\ell \in \mathbb{N}$ and let $u = u_{\mathrm{c}} + u_{\mathrm{f}} \in E_\ell$ with $\|u\| = 1$, where $u_{\mathrm{c}} \in V_{\mathrm{c}}$ (resp. $u_{\mathrm{f}} \in V_{\mathrm{f}}$) denotes the coarse scale part (resp. fine scale part) of u according to the a-orthogonal decomposition in Lemma 3.2. Then it holds*

$$|||u_{\mathrm{c}}||| \leq \sqrt{\lambda_h^{(\ell)}}, \tag{5.2}$$

$$|||u_{\mathrm{f}}||| \lesssim \ell \, \frac{\left(\lambda_h^{(\ell)}\right)^{3/2}}{\alpha} \, H^2, \quad \textit{and} \tag{5.3}$$

$$|(u_{\mathrm{c}}, u_{\mathrm{f}})_{L^2(\Omega)}| \lesssim \ell \left(\frac{\lambda_h^{(\ell)}}{\alpha}\right)^2 H^4. \tag{5.4}$$

Proof Let $\delta_j \leq 1$, $j = 1, 2, \ldots, \ell$, be the coefficients in the representation of u by eigenfunctions, that is, $u = \sum_{j=1}^{\ell} \delta_j u_h^{(j)}$. Then (5.2) follows from the fact that $(1 - \mathcal{P}_{\mathrm{f}})$ is a projection and the obvious bound $|||u|||^2 \leq \lambda_h^{(\ell)}$.

For the proof of (5.3), we employ some algebraic manipulations and Eq. (4.3),

$$|||u_{\mathrm{f}}|||^2 = a(u, u_{\mathrm{f}}) = \sum_{j=1}^{\ell} \delta_j a(u_h^{(j)}, u_{\mathrm{f}}) = \sum_{j=1}^{\ell} \delta_j \lambda_h^{(j)} (u_h^{(j)}, u_{\mathrm{f}})_{L^2(\Omega)}. \tag{5.5}$$

Lemma 3.1, the Cauchy–Schwarz inequality, (2.4), and (2.5) yield

$$\left(u_h^{(j)}, u_{\mathrm{f}}\right)_{L^2(\Omega)} \lesssim \alpha^{-1} H^2 |||u_h^{(j)}||| \, |||u_{\mathrm{f}}||| \tag{5.6}$$

(cf. (3.8)). The combination of (5.5)–(5.6), $|||u_h^{(j)}|||^2 = \lambda_h^{(j)} \leq \lambda_h^{(\ell)}$ and $\delta_j \leq 1$ yields the upper bound of $|||u_{\mathrm{f}}|||$.

The inequality (5.4) follows readily from Theorem 3.3 and the bounds (5.2)–(5.3). □

Remark 5.1 (Improved L^2-quasi-orthogonality under regularity) Consider the full space $V_h = V$. Then, in certain cases, e.g., if Ω is convex and the coefficient A is constant, we have that any macroscopic function $u \in E_\ell$ is in $\mathrm{H}^2(\Omega)$ and $\|\nabla^2 u\| \lesssim \lambda^{(\ell)}/\alpha \|u\|$. Such an instance of regularity gives rise to an additional power of $H\lambda^{(\ell)}/\alpha$ in the estimates (5.3) and (5.4) in Lemma 5.1. This is due to the approximation property

$$\|v - \mathcal{I}_H v\| \lesssim H^2 \|v\|_{\mathrm{H}^2(\Omega)} \tag{5.7}$$

for $v \in V \cap \mathrm{H}^2(\Omega)$, and the possible modification

$$\left(u^{(j)}, u_{\mathrm{f}}\right)_{L^2(\Omega)} = \left(u^{(j)} - \mathcal{I}_H u^{(j)}, u_{\mathrm{f}} - \mathcal{I}_H u_{\mathrm{f}}\right)_{L^2(\Omega)} \lesssim \frac{H^3 \lambda^{(j)}}{\alpha^2} |||u_{\mathrm{f}}|||$$

of (5.6).

5.2 Estimates for approximate eigenvalues

We first introduce the Rayleigh quotient, which is defined for non-trivial $v \in V_h$ by

$$R(v) := \frac{a(v, v)}{(v, v)}.$$

Recall that the ℓth eigenvalue of (4.3) is characterized via the minmax-principle (which goes back to Poincaré [30])

$$\lambda_h^{(\ell)} = \min_{S \in \mathcal{S}_\ell(V_h)} \max_{v \in S \setminus \{0\}} R(v), \tag{5.8}$$

where $\mathcal{S}_\ell(V)$ denotes the set of ℓ-dimensional subspaces of V_h. This principle applies equally well to the coarse problem (4.7), i.e.,

$$\lambda_H^{(\ell)} = \min_{S \in \mathcal{S}_\ell(V_{\mathrm{c}})} \max_{v \in S \setminus \{0\}} R(v) \tag{5.9}$$

characterizes the ℓth discrete eigenvalue ($\ell \leq N_H$). The conformity $V_{\mathrm{c}} \subset V_h (\subseteq V)$ yields monotonicity

$$(\lambda^{(\ell)} \leq)\ \lambda_h^{(\ell)} \leq \lambda_H^{(\ell)} \quad \text{for all } \ell = 1, 2, \ldots, N_H. \tag{5.10}$$

The following theorem gives an estimate in the opposite direction.

Theorem 5.2 (Bound for the eigenvalue error) *Let H be sufficiently small so that $H \lesssim \ell^{-1/4}\sqrt{\frac{\alpha}{\lambda_h^{(\ell)}}}$. Then it holds that*

$$\frac{\lambda_H^{(\ell)} - \lambda_h^{(\ell)}}{\lambda_h^{(\ell)}} \lesssim \ell \left(\frac{\lambda_h^{(\ell)}}{\alpha}\right)^2 H^4 \quad \text{for all } \ell = 1, 2, \ldots, N_H. \tag{5.11}$$

Proof Recall the definition of E_ℓ in (5.1) and define

$$\sigma_H^{(\ell)} := \max_{u \in E_\ell : (u,u)_{L^2(\Omega)} = 1} |(u_{\mathrm{f}}, u_{\mathrm{f}})_{L^2(\Omega)} + 2(u_{\mathrm{c}}, u_{\mathrm{f}})_{L^2(\Omega)}|,$$

where $u_{\mathrm{c}} \in V_{\mathrm{c}}$ (resp. $u_{\mathrm{f}} \in V_{\mathrm{f}}$) denotes the coarse scale part (resp. fine scale part) of $u \in E_\ell$ according to the a-orthogonal decomposition in Lemma 3.2. The L^2-norm of u_{f} satisfies the estimate

$$\begin{aligned} \|u_{\mathrm{f}}\|^2 &= (u, u_{\mathrm{f}})_{L^2(\Omega)} - (u_{\mathrm{c}}, u_{\mathrm{f}})_{L^2(\Omega)} \\ &= (u - \mathcal{I}_H u, u_{\mathrm{f}} - \mathcal{I}_H u_{\mathrm{f}})_{L^2(\Omega)} - (u_{\mathrm{c}}, u_{\mathrm{f}})_{L^2(\Omega)} \\ &\lesssim \ell \left(\frac{\lambda_h^{(\ell)}}{\alpha}\right)^2 H^4 + |(u_{\mathrm{c}}, u_{\mathrm{f}})_{L^2(\Omega)}|, \end{aligned}$$

which follows from Lemma 3.1, (2.4), and (2.5). Hence, Lemma 5.1 shows that

$$\sigma_H^{(\ell)} \lesssim \ell \left(\frac{\lambda_h^{(\ell)}}{\alpha}\right)^2 H^4.$$

If H is chosen small enough so that $\sigma_H^{(\ell)} \leq \frac{1}{2}$ (i.e., $H \lesssim \ell^{-1/4}\sqrt{\frac{\alpha}{\lambda^{(\ell)}}}$), then Lemma 6.1 in [34] shows that

$$\lambda_H^{(\ell)} \leq \left(1 - \sigma_H^{(\ell)}\right)^{-1} \lambda_h^{(\ell)} \leq \left(1 + 2\sigma_H^{(\ell)}\right) \lambda_h^{(\ell)}.$$

Inserting our estimate for $\sigma_H^{(\ell)}$ readily yields the assertion. □

The triangle inequality allows to control the approximation error with respect to the continuous eigenvalues (4.1) by

$$\lambda_H^{(\ell)} - \lambda^{(\ell)} \lesssim \lambda_h^{(\ell)} - \lambda^{(\ell)} + \ell \frac{\left(\lambda_h^{(\ell)}\right)^3}{\alpha^2} H^4.$$

The first term $\lambda_h^{(\ell)} - \lambda^{(\ell)}$ depends on the choice of the space V_h and the regularity of corresponding eigenfunctions in the usual way.

Remark 5.2 (Improved eigenvalue error bound for smooth eigenfunction) With regard to Remark 5.1, the error bound in Theorem 5.2 may be improved in the ideal case $V = V_h$ provided that the first ℓ eigenfunctions are regular in the sense of $\|\nabla^2 u^{(j)}\| \lesssim \lambda^{(j)}/\alpha$. The improved bound reads

$$\frac{\lambda_H^{(\ell)} - \lambda^{(\ell)}}{\lambda^{(\ell)}} \lesssim \ell \left(\frac{\lambda^{(\ell)}}{\alpha}\right)^3 H^5 \quad \text{for all } \ell = 1, 2, \ldots, N_H. \tag{5.12}$$

This improved bound applies also to the case where V_h is a finite element space if h is sufficiently small.

The improved bound might still be pessimistic in the sense that the error in the ℓth eigenvalue/vector depends on the regularity of all previous eigenfunctions. The recent theory [21] shows that this is not necessarily true. Moreover, there might be smoothness also in the single summands of the two-scale decomposition which is not exploited.

5.3 Estimates for approximate eigenfunctions

We turn to the error in the approximate eigenfunctions. Again, we follow the receipt provided in [34].

Theorem 5.3 (Bound for the eigenfunction error) *Let $\lambda_h^{(\ell)}$ be an eigenvalue of multiplicity r, i.e., $\lambda_h^{(\ell)} = \cdots = \lambda_h^{(\ell+r-1)}$ with corresponding eigenspace spanned by the orthonormal basis $\{u_h^{(\ell+j)}\}_{j=0}^{r-1}$. Let the pairs $(\lambda_H^{(\ell)}, u_{\mathrm{c}}^{(\ell)}), \ldots, (\lambda_H^{(\ell+r-1)}, u_{\mathrm{c}}^{(\ell+r-1)})$ be the Rayleigh–Ritz approximations solving Eq. (4.7) with $\|u_{\mathrm{c}}^{(\ell+j)}\| = 1$ for $j = 0, 1, \ldots, r-1$. If $\ell + r - 1 \le N_H$ and if $H \lesssim \ell^{-1/3}(1+\rho)^{-1/3}\sqrt{\alpha/\lambda_h^{(\ell)}}$ is sufficiently small, then there exist an orthonormal basis of* $\mathrm{span}(\{u_h^{(\ell+j)}\}_{j=0}^{r-1})$*, let us denote the basis functions $\tilde{u}_h^{(\ell+j)}$, such that for all $j = 0, 1, \ldots, r-1$,*

$$|||\tilde{u}_h^{(\ell+j)} - u_{\mathrm{c}}^{(\ell+j)}||| \lesssim \sqrt{\ell}\,\frac{(\lambda_h^{(\ell)})^{3/2}}{\alpha} H^2 + \ell(1+\rho)\frac{(\lambda_h^{(\ell)})^2}{\alpha^{3/2}} H^3, \tag{5.13}$$

$$\|\tilde{u}_h^{(\ell+j)} - u_{\mathrm{c}}^{(\ell+j)}\| \lesssim \ell(1+\rho)\left(\frac{\lambda_h^{(\ell)}}{\alpha}\right)^{3/2} H^3, \tag{5.14}$$

where $\rho := \max_{j \notin \{\ell, \ell+1, \ldots, \ell+r-1\}} \frac{\lambda_h^{(\ell)}}{|\lambda_h^{(\ell)} - \lambda_H^{(j)}|}$.

Proof The analysis presented in [34, Lemma 6.4 and Theorem 6.2] shows that, for any $j = 0, 1, \ldots, r-1$, there is a function $\tilde{u}_{\mathrm{c}}^{(\ell+j)} \in \mathrm{span}(\{u_{\mathrm{c}}^{(\ell+i)}\}_{i=0}^{r-1})$ such that

$$\left\| u_h^{(\ell+j)} - \tilde{u}_{\mathrm{c}}^{(\ell+j)} \right\| \le (1+\rho)\|\mathcal{P}_{\mathrm{f}} u_h^{(\ell+j)}\|.$$

According to the a-orthogonal decomposition in Lemma 3.2, $\mathcal{P}_{\mathrm{f}} u_h^{(\ell+j)}$ is the fine scale part of $u_h^{(\ell+j)}$. Hence, the interpolation error estimate (2.4) and Lemma 5.1 yield

$$\sum_{j=1}^{r-1} \left\| u_h^{(\ell+j)} - \tilde{u}_{\mathrm{c}}^{(\ell+j)} \right\|^2 \lesssim (1+\rho)^2 \ell^2 \left(\frac{\lambda_h^{(\ell)}}{\alpha} \right)^3 H^6.$$

If the right-hand side is small enough, i.e., if the multiplicative constant hidden in $H \lesssim \ell^{-1/3}(1+\rho)^{-1/3}\sqrt{\alpha/\lambda_h^{(\ell)}}$ is sufficiently small, the linear transformation of the orthonormal basis $\{u_{\mathrm{c}}^{(\ell+j)}\}_{j=0}^{r-1}$ which defines the set of functions $\{\tilde{u}_{\mathrm{c}}^{(\ell+j)}\}_{j=0}^{r-1}$ may be replaced with an orthogonal transformation, without any harm to the estimate. In this regime, the application of the inverse orthogonal transformation to the errors proves the L^2 bound (5.14).

For the proof of (5.13), observe that for any $v \in \operatorname{span}(\{u_h^{(\ell+i)}\}_{i=0}^{r-1})$ with $\|v\| = 1$ it holds

$$\begin{aligned} |||v - u_{\mathrm{c}}^{(\ell)}|||^2 &= \lambda_h^{(\ell)} - 2\lambda_h^{(\ell)} \left(v, u_{\mathrm{c}}^{(\ell)} \right)_{L^2(\Omega)} + \lambda_H^{(\ell)} \\ &= \lambda_h^{(\ell)} \left(2 - 2 \left(v, u_{\mathrm{c}}^{(\ell)} \right)_{L^2(\Omega)} \right) + \lambda_H^{(\ell)} - \lambda_h^{(\ell)} \\ &= \lambda_h^{(\ell)} \left\| v - u_{\mathrm{c}}^{(\ell)} \right\|^2 + \lambda_H^{(\ell)} - \lambda_h^{(\ell)}. \end{aligned} \tag{5.15}$$

The assertion then follows by combining Eq. (5.15) with $v = \tilde{u}_h^{(\ell+j)}$, (5.14), and Theorem 5.2. □

6 Practical aspects

This section discusses the efficient approximation of the corrector functions $\mathcal{P}_{\mathrm{f}}\phi_z$ from (4.6) by localization, the generalization to non-nested meshes, some post-processing technique, and the overall complexity of our method.

6.1 Localization of fine scale computations

The construction of the coarse space V_{c} is based on the fine scale equations (4.6) which are formulated on the whole domain Ω. This makes them expensive to compute. However, in [25] it was shown that $\mathcal{P}_{\mathrm{f}}\phi_z$ decays exponentially fast outside of the support of the coarse basis function ϕ_z. We specify this feature as follows. Let $k \in \mathbb{N}$. We define nodal patches $\omega_{z,k}$ of k coarse grid layers centered around the node $z \in \mathcal{N}_H$ by

$$\begin{aligned} \omega_{z,1} &:= \operatorname{supp} \phi_z = \cup \{T \in \mathcal{T}_H \mid z \in T\}, \\ \omega_{z,k} &:= \cup \left\{ T \in \mathcal{T}_H \mid T \cap \omega_{z,k-1} \neq \emptyset \right\} \quad \text{for } k \geq 2. \end{aligned} \tag{6.1}$$

The result in the decay of $\mathcal{P}_{\mathrm{f}}\phi_z$ in [25] can be expressed as follows. For all vertices $z \in \mathcal{N}_H$ and for all $k \in \mathbb{N}$, it holds

$$\|A^{1/2}\nabla\mathcal{P}_{\mathrm{f}}\phi_z\|_{L^2(\Omega\setminus\omega_{z,k})} \lesssim e^{-(\alpha/\beta)^{1/2}k}|||\mathcal{P}_{\mathrm{f}}\phi_z|||. \tag{6.2}$$

For moderate contrast β/α, this motivates the truncation of the computations of the basis functions to local patches $\omega_{z,k}$. We approximate $\psi_z = \mathcal{P}_{\mathrm{f}}\phi_z \in V_{\mathrm{f}}$ from (4.6) with $\psi_{z,k} \in V_{\mathrm{f}}(\omega_{z,k}) := \{v \in V_{\mathrm{f}} \mid v|_{\Omega\setminus\omega_{x,k}} = 0\}$ such that

$$a(\psi_{z,k}, v) = a(\phi_z, v) \quad \text{for all } v \in V_{\mathrm{f}}(\omega_{z,k}). \tag{6.3}$$

We emphasize that

$$V_{\mathrm{f}}(\omega_{z,k}) = \{v \in V_h \mid v|_{\Omega\setminus\omega_{x,k}} = 0,\ \forall y \in \mathcal{N}_H \cap \omega_{z,k} : (v, \phi_y)_{L^2(\Omega)} = 0\},$$

i.e., in a practical computation with lagrangian multipliers only one linear constraint per coarse vertex in the patch $\omega_{x,k}$ needs to be considered.

The localized computations yield a modified coarse space V_{c}^k with a local basis

$$V_{\mathrm{c}}^k = \mathrm{span}\{\phi_z - \psi_{z,k} \mid z \in \mathcal{N}_H\}. \tag{6.4}$$

The number of non-zero entries of the corresponding finite element stiffness and mass matrix is proportional to $k^d N_H$ (note that we expect N_H^2 non-zero entries without the truncation). Due to the exponential decay, the very weak condition $k \approx |\log H|$ implies that the perturbation of the ideal method due to this truncation is of higher order and the estimates in Theorems 5.2 and 5.3 remain valid. We refer to [25] for details and proofs. The modified localization procedures from [15] and [18] with improved accuracy and stability properties might as well be applied.

6.2 Non-nested meshes and general coarsening

In Sect. 2.1, we have assumed that $\mathcal{T}_h$ is derived from $\mathcal{T}_H$ by some regular refinement, i.e., that the finite element meshes $\mathcal{T}_h$ and $\mathcal{T}_H$ are nested. This condition may be impracticable in relevant applications, e.g., in cases where the coefficient encodes microscopic geometric features such as jumps that require accurate resolution and the reasonable resolution can only be achieved by highly unstructured meshes (cf. Fig. 3 in Sect. 7.3 below).

A closer look to the previous error analysis shows that the nestedness of the underlying meshes is never used explicitly but enters only implicitly via the nestedness of corresponding spaces $V_H \subset V_h$. It turns out that all results generalize to the case where the standard finite element space V_H on the coarse level is replaced with some general (possibly mesh free) coarse space $\tilde{V}_H \subset V_h$ with a local basis $\{\tilde{\phi}_j\}_{j\in J}$; J being some finite index set. Precise necessary conditions for the theory read:

(a) *Local support and finite overlap.* For all $j \in J$, $\operatorname{diam}(\operatorname{supp}\tilde{\phi}_j) \lesssim H$ and there is a finite number C_{ol} independent of H such that no point $x \in \Omega$ belongs to the support of more than C_{ol} basis functions.
(b) *Non-negativity, continuity and boundedness.* For all $j \in J$, $\tilde{\phi}_j : \Omega \to [0,1]$ is continuous and $\|\nabla\tilde{\phi}_j\|_{L^\infty(\Omega)} \lesssim H^{-1}$.
(c) *Partition of unity up to a boundary strip.* For all $x \in \Omega$, it holds that $\operatorname{dist}(x, \partial\Omega) \lesssim H$ or $\sum_{j\in J} \tilde{\phi}_j(x) = 1$.

Under the conditions (a)–(c), the operator $\mathcal{I}_H$, defined by $\mathcal{I}_H v := \sum_{j\in J} \frac{(v,\tilde{\phi}_j)_{L^2(\Omega)}}{(1,\tilde{\phi}_j)_{L^2(\Omega)}}\tilde{\phi}_j$ for $v \in V$, satisfies the required stability and approximation properties. Their proofs may easily be extracted from [9], where a slightly modified operator is considered. For details regarding the generalization of the decompositions and error bounds of this paper to some general coarse space characterized by (a)–(c), we refer to [15], where everything (including the exponential decay of the coarse basis and its localization) has been worked out for a linear boundary value problem.

The conditions (a)–(c) are natural conditions for general coarse spaces used in domain decomposition methods and algebraic multigrid methods; see [36, Ch. 3.10] for an overview and [32] for a particular construction without any coarse mesh. A very simple mesh-based construction which remains very close to the standard finite element space V_H can be found in [35, Section 2.2] and works as follows. Given some regular fine mesh $\mathcal{T}_h$, consider an arbitrary regular quasi-uniform coarse mesh $\mathcal{T}_H$ with $H > h$. Let V_h (resp. V_H) be the corresponding finite element space of continuous $\mathcal{T}_h$-piecewise (resp. $\mathcal{T}_H$-piecewise) affine functions and let $I_h^{\mathrm{nodal}} : V_H \subset C^0(\Omega) \to V_h$ denote the nodal interpolation operator with respect to the fine mesh. The nodal interpolation of standard nodal basis functions of the coarse mesh defines a nested initial coarse space

$$\tilde{V}_H := \operatorname{span}\left\{ I_h^{\mathrm{nodal}}\phi_z \mid z \in \mathcal{N}_H \right\} \subset V_h \tag{6.5}$$

and $V_{\mathrm{c}} := (1 - \mathcal{P}_{\mathrm{f}})\tilde{V}_H$ is the corresponding coarse space of our method. The desired properties (a)–(c) of $\tilde{V}_H$ are proven in [35, Lemma 2.1]. Section 7.3 shows numerical results based on this construction.

6.3 Postprocessing

As already mentioned in the introduction, the two-grid method of [37] allows a certain post-processing (solution of linear problems on the fine scale) of coarse eigenpairs. So far, this method was mainly used to post-process approximate eigenpairs of standard finite element approximations on a coarse mesh, i.e., approximations with respect to the space V_H. However, the framework presented in [37] is more general and readily applies to the modified coarse space V_{c}. Given some approximate eigenpair $(\lambda_H^{(\ell)}, u_{\mathrm{c}}^{(\ell)}) \in \mathbb{R}\times V_{\mathrm{c}}$ with $\|u_{\mathrm{c}}^{(\ell)}\| = 1$ that solves (4.7), the post-processed approximate

eigenfunction $u_{\mathrm{c,post}}^{(\ell)} \in V_h$ is characterized uniquely by

$$a\left(u_{\mathrm{c,post}}^{(\ell)}, v\right) = \lambda_H^{(\ell)}(u_{\mathrm{c}}^{(\ell)}, v)_{L^2(\Omega)} \tag{6.6}$$

for all $v \in V_h$. The corresponding post-processed eigenvalue is

$$\lambda_{H,\mathrm{post}}^{(\ell)} := \frac{a(u_{\mathrm{c,post}}^{(\ell)}, u_{\mathrm{c,post}}^{(\ell)})}{(u_{\mathrm{c,post}}^{(\ell)}, u_{\mathrm{c,post}}^{(\ell)})_{L^2(\Omega)}}. \tag{6.7}$$

The error analysis of [37] relies solely on the nestedness $V_{\mathrm{c}} \subset V_h$ and, in essence, yields the error estimates

$$\begin{aligned}\left|\lambda_{H,\mathrm{post}}^{(\ell)} - \lambda_h^{(\ell)}\right| &\leq |||u_h^{(\ell)} - u_{\mathrm{c,post}}^{(\ell)}|||^2 \\ &\lesssim \left(\lambda_H^{(\ell)} - \lambda_h^{(\ell)}\right)^2 + \left(\lambda_h^{(\ell)}\right)^2 \|u_h^{(\ell)} - u_{\mathrm{c}}^{(\ell)}\|^2.\end{aligned}$$

The first estimate follows from (5.15) which remains valid for $u_{\mathrm{c}}^{(\ell)}$ and $\lambda_H^{(\ell)}$ replaced with $u_{\mathrm{c,post}}^{(\ell)}$ and $\lambda_{H,\mathrm{post}}^{(\ell)}$. The second estimate follows from the construction and standard inequalities (cf. [37, Eq. (4.3)]). Hence, with $u_h^{(\ell)}$ suitably chosen, Theorems 5.2 and 5.3 imply that the error of the post-processed eigenvalues (resp. post-processed eigenfunctions) is at least of order H^6 (resp. H^4). As for all our previous results, the rates do not depend on any regularity of the eigenfunctions. In the third numerical experiment of Sect. 7 we will also show results for this post-processing technique.

6.4 Complexity

Finally, we shall comment on the overall complexity of our approach. Consider quasi-uniform meshes of size H and h and corresponding conforming first-order finite element spaces V_H and V_h. We want to approximate the eigenvalues related to V_h.

In order to set up the coarse space V_{c}, we need to solve N_H linear problems with approximately $k^d N_h/N_H$ degrees of freedom each; the parameter k being the truncation parameter as above. Since almost linear complexity is possible (using, e.g., multilevel preconditioning techniques), the cost for solving one of these problems up to a given accuracy is proportional to the number of degrees of freedom N_h/N_H up to possible logarithmic factors. This yields an overall complexity of $k^d N_h \log(N_h)$ (resp. $N_H N_h \log(N_h)$ if $k^d \geq N_H$) for setting up the coarse problem. Note that this effort can be reduced drastically either by considering the independence of the linear problems in terms of parallelism or by exploiting a possible periodicity in the problem and the mesh. In the latter case, only very few of the problems have to be computed because all the other ones are equivalent up to translation or rotation of coordinates.

On top of the assembling, an N_H-dimensional eigenvalue problem is to be solved. The complexity of this depends only on N_H, the number of eigenvalues of interest, and the truncation parameter k but *not* on the critically large parameter N_h.

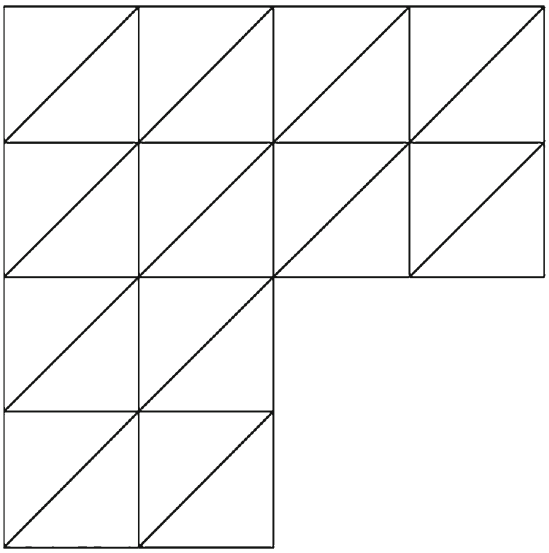


Fig. 1 Initial uniform triangulation of the L-shape domain (5 degrees of freedom)

The cost of the post-processing presented in Sect. 6.3 is proportional to one fine solve for each eigenpair of interest, i.e., proportional to N_h up to some logarithmic factor.

7 Numerical experiments

Three numerical experiments shall illustrate our theoretical results. While the first two experiments consider nested coarse and fine meshes, the third experiments uses the generalized coarsening strategy of Sect. 6.2. In all experiments, we focus on the case without localization. The localization (as discussed in Sect. 6.1) has been studied extensively for the linear problem in [14,15,25] and for semi-linear problems in [17]. In the present context of eigenvalue approximation, we are interested in observing the enormous convergence rate which is 4 or higher for the eigenvalues. In order to achieve this rate also with truncation, patches have to be large (at least 4 layers of elements) which pays off only asymptotically when H is small enough.

7.1 Constant coefficient on L-shaped domain

Let $\Omega := (-1, 1)^2 \setminus [0, 1]^2$ be the L-shaped domain. Consider the constant scalar coefficient $A_1 = 1$ and uniform coarse meshes with mesh widths $\sqrt{2}H = 2^{-1}, \ldots, 2^{-4}$ of Ω as depicted in Fig. 1.

The reference mesh $\mathcal{T}_h$ has maximal mesh width $h = 2^{-7}/\sqrt{2}$. We consider some $P1$ conforming finite element approximation of the eigenvalues on the reference mesh $\mathcal{T}_h$ and compare these discrete eigenvalues $\lambda_h^{(\ell)}$ with coarse scale approximations depending on the coarse mesh size H.

Table 1 shows results for the case without truncation, i.e., all linear problems have been solved on the whole of Ω.

Table 1 Errors $e^{(\ell)}(H) =: \frac{\lambda_H^{(\ell)} - \lambda_h^{(\ell)}}{\lambda_h^{(\ell)}}$ for $\ell = 1, \ldots, 20$, constant coefficient A_1, and various choices of the coarse mesh size H

ℓ	$\lambda_h^{(\ell)}$	$e^{(\ell)}(1/2\sqrt{2})$	$e^{(\ell)}(1/4\sqrt{2})$	$e^{(\ell)}(1/8\sqrt{2})$	$e^{(\ell)}(1/16\sqrt{2})$
1	9.6436568	0.004161918	0.000041786	0.000000696	0.000000014
2	15.1989733	0.009683715	0.000083718	0.000000888	0.000000011
3	19.7421815	0.024238729	0.000199984	0.000001930	0.000000022
4	29.5280022	0.084950011	0.000679046	0.000006309	0.000000074
5	31.9266947	0.120246865	0.001032557	0.000011298	0.000000169
6	41.4911125	–	0.002220585	0.000019622	0.000000264
7	44.9620831	–	0.002837949	0.000022540	0.000000257
8	49.3631818	–	0.003535358	0.000027368	0.000000295
9	49.3655616	–	0.004143842	0.000031434	0.000000343
10	56.7367306	–	0.006494922	0.000052862	0.000000606
11	65.4137240	–	0.013504833	0.000094150	0.000000995
12	71.0950435	–	0.013314963	0.000095197	0.000001077
13	71.6015951	–	0.011792861	0.000084001	0.000000851
14	79.0044010	–	0.021302527	0.000155038	0.000001526
15	89.3721008	–	0.038951872	0.000233603	0.000002613
16	92.3686575	–	0.042125029	0.000253278	0.000002442
17	97.4392146	–	0.033015921	0.000254700	0.000002435
18	98.7544790	–	0.039634464	0.000264156	0.000002482
19	98.7545515	–	0.046865242	0.000268012	0.000002500
20	101.6764284	–	0.045797998	0.000311683	0.000003071

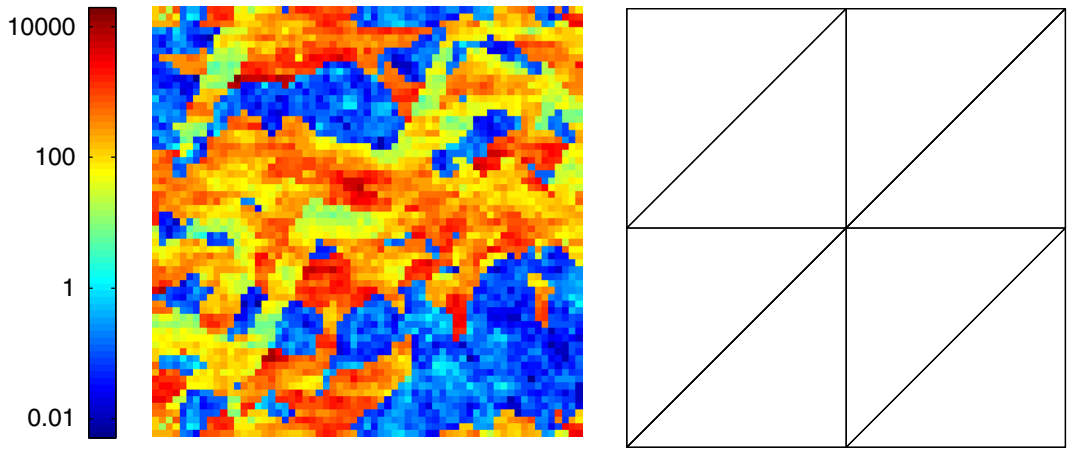


Fig. 2 Scalar coefficient A_2 used in the second numerical experiment and initial uniform triangulation of the unit square (1 degree of freedom)

For fixed ℓ, the rate of convergence of the eigenvalue error $\lambda_H^{(\ell)} - \lambda_h^{(\ell)}$ in terms of H observed in Table 1 is between 6 and 7 which is even better than predicted in Theorem 5.2 and in Remark 5.1.

Table 2 Errors $e^{(\ell)}(H) =: \frac{\lambda_H^{(\ell)} - \lambda_h^{(\ell)}}{\lambda_h^{(\ell)}}$ for $\ell = 1, \ldots, 20$, rough coefficient A_2, and various choices of the coarse mesh size H

ℓ	$\lambda_h^{(\ell)}$	$e^{(\ell)}(1/2\sqrt{2})$	$e^{(\ell)}(1/4\sqrt{2})$	$e^{(\ell)}(1/8\sqrt{2})$	$e^{(\ell)}(1/16\sqrt{2})$
1	21.4144522	5.472755371	0.237181706	0.010328293	0.000781683
2	40.9134676	–	0.649080539	0.032761482	0.002447049
3	44.1561133	–	1.687388874	0.097540102	0.004131422
4	60.8278691	–	1.648439518	0.028076168	0.002079812
5	65.6962136	–	2.071005692	0.247424446	0.006569640
6	70.1273082	–	4.265936007	0.232458016	0.016551520
7	82.2960238	–	3.632888104	0.355050163	0.013987920
8	92.8677605	–	6.850048057	0.377881216	0.049841235
9	99.6061234	–	10.305084010	0.469770376	0.026027378
10	109.1543283	–	–	0.476741452	0.005606426
11	129.3741945	–	–	0.505888044	0.062382302
12	138.2164330	–	–	0.554736550	0.039487317
13	141.5464639	–	–	0.540480876	0.043935515
14	145.7469718	–	–	0.765411709	0.034249528
15	152.6283573	–	–	0.712383825	0.024716759
16	155.2965039	–	–	0.761104705	0.026228034
17	158.2610708	–	–	0.749058367	0.091826207
18	164.1452194	–	–	0.840736127	0.118353184
19	171.1756923	–	–	0.946719951	0.111314058
20	179.3917590	–	–	0.928617606	0.119627862

7.2 Rough coefficient with multiscale features

Let $\Omega := (0, 1)^2$ be the unit square. The scalar coefficient A_2 (see Fig. 2) is piecewise constant with respect to the uniform Cartesian grid of width 2^{-6}. Its values are taken from the data of the SPE10 benchmark, see http://www.spe.org/web/csp/. The coefficient is highly varying and strongly heterogeneous. The contrast for A_2 is large, $\beta(A_2)/\alpha(A_1) \approx 4 \cdot 10^6$. Consider uniform coarse meshes of size $\sqrt{2}H = 2^{-1}, 2^{-2}, \ldots, 2^{-4}$ of Ω (cf. Fig. 2). Note that none of these meshes resolves the rough coefficient A_2 appropriately. Hence, (local) regularity cannot be exploited on coarse meshes.

Again, the reference mesh $\mathcal{T}_h$ has width $h = 2^{-7}/\sqrt{2}$ and we compare the discrete eigenvalues $\lambda_h^{(\ell)}$ (with respect to some $P1$ conforming finite element approximation of the eigenvalues on the reference mesh $\mathcal{T}_h$) with coarse scale approximations depending on the coarse mesh size H. Table 2 shows the errors and allows us to estimate the average rate around 4 which matches our expectation from the theory. We emphasize that the large contrast does not seem to affect the accuracy of our method in approximating the eigenvalues $\lambda_h^{(\ell)}$. However, the accuracy of $\lambda_h^{(\ell)}$ may be affected by the high contrast and the lack of regularity caused by the coefficient.

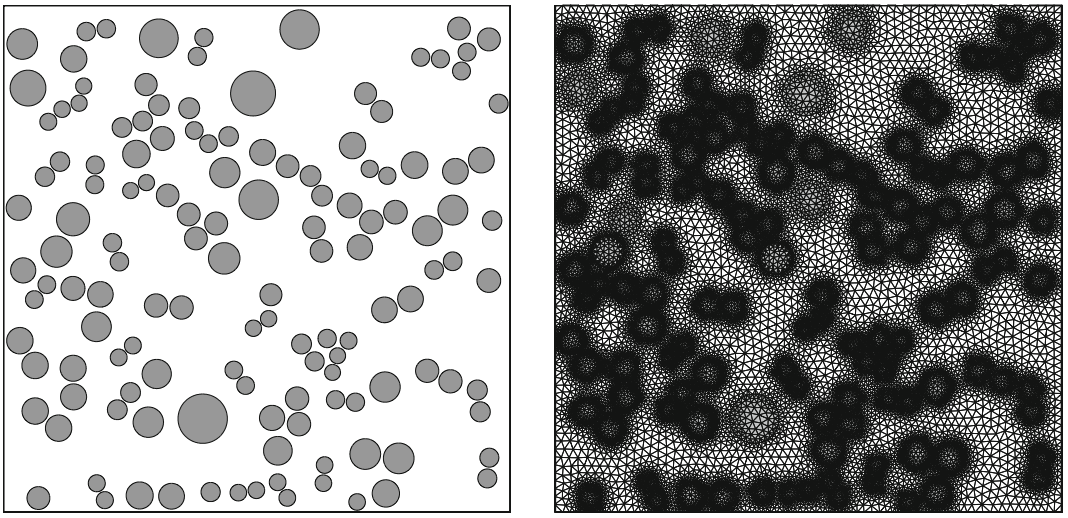


Fig. 3 *Left* Scalar coefficient A_3 used in the third numerical experiment. A_3 takes the value 100 in the *gray shaded* inclusions and the value 1 elsewhere. *Right* Un structured fine mesh $\mathcal{T}_h$ aligned with jumps of the coefficient A_3

Table 3 Errors $e^{(\ell)}(H) =: \frac{\lambda_H^{(\ell)} - \lambda_h^{(\ell)}}{\lambda_h^{(\ell)}}$ for $\ell = 1, \ldots, 20$, coefficient A_3, and various choices of the coarse mesh size H

ℓ	$\lambda_h^{(\ell)}$	$e^{(\ell)}(1/2\sqrt{2})$	$e^{(\ell)}(1/4\sqrt{2})$	$e^{(\ell)}(1/8\sqrt{2})$	$e^{(\ell)}(1/16\sqrt{2})$
1	25.6109462	0.025518831	0.000572341	0.000017083	0.000000700
2	58.9623566	–	0.005235813	0.000090490	0.000002710
3	67.5344854	–	0.006997582	0.000154850	0.000006488
4	98.2808694	–	0.023497502	0.000358178	0.000011675
5	121.2290664	–	0.052366141	0.000563438	0.000016994
6	125.2014779	–	0.066627585	0.000747688	0.000019934
7	156.0597873	–	0.145676350	0.001579177	0.000034329
8	168.2376096	–	0.095360287	0.001320185	0.000043781
9	197.4467434	–	0.343991317	0.002888471	0.000049479
10	209.4657306	–	–	0.003223901	0.000056318
11	222.4472476	–	–	0.003431462	0.000080284
12	245.5656759	–	–	0.005906282	0.000102243
13	253.7074603	–	–	0.006215809	0.000121646
14	288.0756442	–	–	0.013859535	0.000180899
15	298.8903269	–	–	0.010587124	0.000138404
16	311.4410556	–	–	0.012159268	0.000161510
17	324.6865434	–	–	0.012143676	0.000176624
18	336.7931865	–	–	0.016554437	0.000233067
19	379.5697606	–	–	0.023254268	0.000325324
20	386.9938901	–	–	0.028772395	0.000383532

Table 4 Errors $e^{(\ell)}(H) =: \frac{\lambda^{(\ell)}_{H,\mathrm{post}} - \lambda^{(\ell)}_h}{\lambda^{(\ell)}_h}$ after post-processing for $\ell = 1, \ldots, 20$, coefficient A_3, and various choices of the coarse mesh size H

ℓ	$\lambda^{(\ell)}_h$	$e^{(\ell)}(1/2\sqrt{2})$	$e^{(\ell)}(1/4\sqrt{2})$	$e^{(\ell)}(1/8\sqrt{2})$	$e^{(\ell)}(1/16\sqrt{2})$
1	25.6109462	0.001559704	0.000003765	0.000000008	3.5e−10
2	58.9623566	–	0.000191532	0.000000213	1.9e−08
3	67.5344854	–	0.000284980	0.000000474	0.000000001
4	98.2808694	–	0.002239689	0.000002253	0.000000004
5	121.2290664	–	0.007461217	0.000005065	0.000000008
6	125.2014779	–	0.011284614	0.000006826	0.000000008
7	156.0597873	–	0.042466017	0.000023867	0.000000024
8	168.2376096	–	0.025093182	0.000027547	0.000000042
9	197.4467434	–	0.186960343	0.000072471	0.000000051
10	209.4657306	–	–	0.000105777	0.000000079
11	222.4472476	–	–	0.000131569	0.000000129
12	245.5656759	–	–	0.000286351	0.000000213
13	253.7074603	–	–	0.000268463	0.000000255
14	288.0756442	–	–	0.000915102	0.000000473
15	298.8903269	–	–	0.000762135	0.000000403
16	311.4410556	–	–	0.000873769	0.000000504
17	324.6865434	–	–	0.000955392	0.000000642
18	336.7931865	–	–	0.001335246	0.000000977
19	379.5697606	–	–	0.002896202	0.000001886
20	386.9938901	–	–	0.007202657	0.000001908

7.3 Particle composite modeled by an unstructured mesh

Let $\Omega := (0, 1)^2$ be the unit square. In this experiment, the scalar coefficient A_3 models heat conductivity in some model composite material with randomly dispersed circular inclusions as depicted in Fig. 3. The coefficient A_3 takes the value 100 in the gray shaded inclusions and the value 1 elsewhere. In order to resolve the discontinuities, we simply align the fine mesh $\mathcal{T}_h$ with the boundaries of the inclusions (see Fig. 3). The mesh size of $\mathcal{T}_h$ satisfies $2^{-9} \lesssim h \lesssim 2^{-7}$. Note that this fine mesh $\mathcal{T}_h$ is solely based on geometric resolution and shape regularity. The grading towards the inclusions is not adapted to the characteristic behavior of the eigenfunctions. However, this mesh might be the actual output of some commercial mesh generator or modeling tool. Sufficient resolution could be achieved with fewer degrees of freedom, however, this would require more sophisticated discretization spaces; we refer to [7,28,29] for possible choices and further references.

As in the previous experiment, we consider uniform coarse meshes of size $\sqrt{2}H = 2^{-1}, 2^{-2}, \ldots, 2^{-4}$ of Ω (cf. Fig. 2). Note that these meshes neither resolves the coefficient A_3 appropriately nor can be refined to $\mathcal{T}_h$ in a nested way. For the construction of the upscaling approximation we employ the generalized coarse space defined in

(6.5) in Sect. 6.2. We compare the discrete eigenvalues $\lambda_h^{(\ell)}$ (with respect to some $P1$ conforming finite element approximation of the eigenvalues on the reference mesh $\mathcal{T}_h$) with coarse scale approximations depending on the coarse discretization parameter H. Table 3 shows the results which clearly support our claim that the nestedness of coarse and fine meshes is not essential and that upscaling far beyond the characteristic length scales of the problem (i.e., the radii of the inclusions and their distances) is possible.

For this problem, we have also computed the post-processed approximations according to Sect. 6.3. Table 4 shows the error for the eigenvalues which are more accurate by several orders of magnitude. The experimental rates are roughly between 5 and 6 in Table 3 without post-processing and around 9 to 10 after post-processing in Table 4.

References

1. Banjai, L., Börm, S., Sauter, S.: FEM for elliptic eigenvalue problems: how coarse can the coarsest mesh be chosen? An experimental study. Comput. Vis. Sci. **11**(4–6), 363–372 (2008)
2. Birkhoff, G., de Boor, C., Swartz, B., Wendroff, B.: Rayleigh–Ritz approximation by piecewise cubic polynomials. SIAM J. Numer. Anal. **3**, 188–203 (1966)
3. Bank, R.E., Grubišić, L., Ovall, J.S.: A framework for robust eigenvalue and eigenvector error estimation and Ritz value convergence enhancement. Appl. Numer. Math. **66**, 1–29 (2013)
4. Boffi, D.: Finite element approximation of eigenvalue problems. Acta Numer. **19**, 1–120 (2010)
5. Carstensen, C., Gedicke, J.: An oscillation-free adaptive FEM for symmetric eigenvalue problems. Numer. Math. **118**(3), 401–427 (2011)
6. Carstensen, C., Gedicke, J.: An adaptive finite element eigenvalue solver of asymptotic quasi-optimal computational complexity. SIAM J. Numer. Anal. **50**(3), 1029–1057 (2012)
7. Chu, C.-C., Graham, I.G., Hou, T.-Y.: A new multiscale finite element method for high-contrast elliptic interface problems. Math. Comput. **79**(272), 1915–1955 (2010)
8. Ciarlet, P.G.: The Finite Element Method for Elliptic Problems. North-Holland, Amsterdam (1987)
9. Carstensen, C., Verfürth, R.: Edge residuals dominate a posteriori error estimates for low order finite element methods. SIAM J. Numer. Anal. **36**(5), 1571–1587 (1999). (electronic)
10. Durán, R.G., Padra, C., Rodríguez, R.: A posteriori error estimates for the finite element approximation of eigenvalue problems. Math. Models Methods Appl. Sci. **13**(8), 1219–1229 (2003)
11. Giani, S., Graham, I.G.: A convergent adaptive method for elliptic eigenvalue problems. SIAM J. Numer. Anal. **47**(2), 1067–1091 (2009)
12. Garau, E.M., Morin, P., Zuppa, C.: Convergence of adaptive finite element methods for eigenvalue problems. Math. Models Methods Appl. Sci. **19**(5), 721–747 (2009)
13. Hackbusch, W.: On the computation of approximate eigenvalues and eigenfunctions of elliptic operators by means of a multi-grid method. SIAM J. Numer. Anal. **16**(2), 201–215 (1979)
14. Henning, P., Målqvist, A.: Localized orthogonal decomposition techniques for boundary value problems. SIAM J. Sci. Comput. **36**(4), A1609–A1634 (2014)
15. Henning, P., Morgenstern, P., Peterseim, D.: Multiscale partition of unity. In: Griebel, M., Schweitzer, M.A. (eds.) Meshfree Methods for Partial Differential Equations VII, Lecture Notes in Computational Science and Engineering, vol. 100. Springer, New York (2014)
16. Henning, P., Målqvist, A., Peterseim, D.: Two-level discretization techniques for ground state computations of Bose–Einstein condensates. SIAM J. Numer. Anal. **52**(4), 1525–1550 (2014)
17. Henning, P., Målqvist, A., Peterseim, D.: A localized orthogonal decomposition method for semi-linear elliptic problems. ESAIM. Math. Model. Numer. Anal. **48**, 1331–1349 (2014). 9
18. Henning, P., Peterseim, D.: Oversampling for the multiscale finite element method. Multiscale Model. Simul. **11**(4), 1149–1175 (2013)
19. Knyazev, A.V., Neymeyr, K.: Efficient solution of symmetric eigenvalue problems using multigrid preconditioners in the locally optimal block conjugate gradient method. Tenth Copper Mountain Conference on Multigrid Methods (Copper Mountain, CO, 2001). Electron. Trans. Numer. Anal. **15**, 38–55 (2003). (electronic)

20. Knyazev, A.V., Neymeyr, K.: A geometric theory for preconditioned inverse iteration. III. A short and sharp convergence estimate for generalized eigenvalue problems. Special issue on accurate solution of eigenvalue problems (Hagen, 2000). Linear Algebra Appl. **358**, 95–114 (2003)
21. Knyazev, A.V., Osborn, J.E.: New a priori FEM error estimates for eigenvalues. SIAM J. Numer. Anal. **43**(6), 2647–2667 (2006). (electronic)
22. Larson, M.G.: A posteriori and a priori error analysis for finite element approximations of self-adjoint elliptic eigenvalue problems. SIAM J. Numer. Anal. **38**(2), 608–625 (2000). (electronic)
23. Lehoucq, R.B., Sorensen, D.C., Yang, C.: ARPACK users' guide, volume 6 of Software, Environments, and Tools. Solution of Large-Scale Eigenvalue Problems with Implicitly Restarted Arnoldi Methods. Society for Industrial and Applied Mathematics (SIAM), Philadelphia (1998)
24. Mehrmann, V., Miedlar, A.: Adaptive computation of smallest eigenvalues of self-adjoint elliptic partial differential equations. Numer. Linear Algebra Appl. **18**(3), 387–409 (2011)
25. Målqvist, A., Peterseim, D.: Localization of elliptic multiscale problems. Math. Comput. **83**(290), 2583–2603 (2014)
26. Neymeyr, K.: A posteriori error estimation for elliptic eigenproblems. Numer. Linear Algebra Appl. **9**(4), 263–279 (2002)
27. Neymeyr, K.: Solving mesh eigenproblems with multigrid efficiency. In: Numerical Methods for Scientific Computing. Variational Problems and Applications. Internat. Center Numer. Methods Eng. (CIMNE), Barcelona, pp. 176–184 (2003)
28. Peterseim, D., Carstensen, C.: Finite element network approximation of conductivity in particle composites. Numer. Math. **124**(1), 73–97 (2013)
29. Peterseim, D.: Composite finite elements for elliptic interface problems. Math. Comput. **83**(290), 2657–2674 (2014)
30. Poincaré, H.: Sur les Equations aux Derivees Partielles de la Physique Mathematique. Am. J. Math. **12**(3), 211–294 (1890)
31. Peterseim, D., Sauter, S.: Finite elements for elliptic problems with highly varying, nonperiodic diffusion matrix. Multiscale Model. Simul. **10**(3), 665–695 (2012)
32. Sarkis, M.: Partition of unity coarse spaces and Schwarz methods with harmonic overlap. In Recent developments in domain decomposition methods (Zürich, 2001), Lect. Notes Comput. Sci. Eng., vol. 23. Springer, Berlin, pp. pages 77–94 (2002)
33. Sauter, S.: hp-finite elements for elliptic eigenvalue problems: error estimates which are explicit with respect to λ, h, and p. SIAM J. Numer. Anal. **48**(1), 95–108 (2010)
34. Strang, G., Fix, G. J.: An Analysis of the Finite Element Method, Prentice-Hall Series in Automatic Computation. Prentice-Hall Inc., Englewood Cliffs (1973)
35. Scheichl, R., Vassilevski, P.S., Zikatanov, L.T.: Weak approximation properties of elliptic projections with functional constraints. Multiscale Model. Simul. **9**(4), 1677–1699 (2011)
36. Toselli, A., Widlund, O.: Domain Decomposition Methods—Algorithms and Theory, Springer Series in Computational Mathematics, vol. 34. Springer, Berlin (2005)
37. Xu, J., Zhou, A.: A two-grid discretization scheme for eigenvalue problems. Math. Comput. **70**(233), 17–25 (2001)

C.2 Two-level discretization techniques for ground state computations of Bose-Einstein condensates

SIAM Journal on Numerical Analysis **52**(4):1525–1550, 2014.

(with P. Henning and A. Målqvist)

SIAM J. NUMER. ANAL.
Vol. 52, No. 4, pp. 1525–1550

TWO-LEVEL DISCRETIZATION TECHNIQUES FOR GROUND STATE COMPUTATIONS OF BOSE-EINSTEIN CONDENSATES*

PATRICK HENNING†, AXEL MÅLQVIST‡, AND DANIEL PETERSEIM§

Abstract. This work presents a new methodology for computing ground states of Bose–Einstein condensates based on finite element discretizations on two different scales of numerical resolution. In a preprocessing step, a low-dimensional (coarse) generalized finite element space is constructed. It is based on a local orthogonal decomposition of the solution space and exhibits high approximation properties. The nonlinear eigenvalue problem that characterizes the ground state is solved by some suitable iterative solver exclusively in this low-dimensional space, without significant loss of accuracy when compared with the solution of the full fine scale problem. The preprocessing step is independent of the types and numbers of bosons. A postprocessing step further improves the accuracy of the method. We present rigorous a priori error estimates that predict convergence rates H^3 for the ground state eigenfunction and H^4 for the corresponding eigenvalue without pre-asymptotic effects; H being the coarse scale discretization parameter. Numerical experiments indicate that these high rates may still be pessimistic.

Key words. eigenvalue, finite element, Gross–Pitaevskii equation, numerical upscaling, two-grid method, multiscale method

AMS subject classifications. 35Q55, 65N15, 65N25, 65N30, 81Q05

DOI. 10.1137/130921520

1. Introduction.

Bose–Einstein condensates (BEC) are formed when a dilute gas of trapped bosons (of the same species) is cooled down to ultra-low temperatures close to absolute zero [10, 19, 22, 38]. In this case, nearly all bosons are in the same quantum mechanical state, which means that they loose their identity and become indistinguishable from each other. The BEC therefore behaves like one "super particle" where the quantum state can be described by a single collective wave function Ψ. The dynamics of a BEC can be modeled by the time-dependent Gross–Pitaevskii equation (GPE) [26, 31, 37], which is a nonlinear Schrödinger equation given by

$$i\hbar\,\partial_t\Psi = -\frac{\hbar^2}{2m}\triangle\Psi + V_e\Psi + \frac{4\pi\hbar^2 aN}{m}|\Psi|^2\Psi. \tag{1.1}$$

Here, m denotes the atomic mass of a single boson, N is the number of bosons (typically in the span between 10^3 and 10^7), $\hbar$ is the reduced Plank's constant, and V_e is an external trapping potential that confines the system. The nonlinear term in the equation describes the effective two-body interaction between the particles. If the scattering length a is positive, the interaction is repulsive; if it is negative the interaction is attractive. For $a = 0$ there is no interaction and (1.1) becomes the

*Received by the editors May 20, 2013; accepted for publication (in revised form) April 8, 2014; published electronically July 3, 2014.

http://www.siam.org/journals/sinum/52-4/92152.html

†ANMC, Section de Mathématiques, École polytechnique fédérale de Lausanne, 1015 Lausanne, Switzerland (patrick.henning@epfl.ch). This author was supported by the Göran Gustafsson Foundation and the Swedish Research Council.

‡Department of Mathematical Sciences, Chalmers University of Technology and University of Gothenburg, 41296 Gothenburg, Sweden (axel@chalmers.se). This author was supported by the Göran Gustafsson Foundation and the Swedish Research Council.

§Institut für Numerische Simulation der Universität Bonn, 53123 Bonn, Germany (peterseim@ins.uni-bonn.de). This author was partly supported by the Humboldt-Universität and the DFG Research Center Matheon Berlin.

Schrödinger equation. The parameter a changes according to the considered species of bosons. We only consider the case $a \geq 0$ in this paper. We are mainly interested in the ground state solution of the problem. This stationary state of the BEC is of practical relevance, e.g., in the context of atom lasers [35, 30, 41]. The ansatz $\Psi(x,t) = \hat{c}e^{-i\lambda\hat{t}}u(\hat{x})$, with the unknown chemical potential of the condensate λ and a proper nondimensionalization $(x,t) \mapsto (\hat{x},\hat{t})$, reduces (1.1) to the time-independent GPE

$$-\frac{1}{2}\triangle u + Vu + \beta|u|^2u = \lambda u \qquad \text{with } \beta = \frac{4\pi aN}{x_s},$$

where x_s denotes the dimensionless length unit and where V denotes the accordingly rescaled potential. (See, e.g., [8] for a derivation of the time-independent GPE.) The ground state of the BEC is the lowest energy state of the system and is therefore stable. It minimizes the corresponding energy

$$E(v) = \int_{\mathbb{R}^d} \frac{1}{2}|\nabla v|^2 + V|v|^2 + \frac{\beta}{2}|v|^4\,dx$$

among all L^2-normalized H^1 functions. For any L^2-normalized minimizer u, $\lambda = E(u) + \frac{\beta}{2}\|u\|^4_{L^4(\mathbb{R}^d)}$ is the smallest eigenvalue of the GPE. In this paper, we shall focus on the computation of this ground state eigenvalue. Eigenfunctions whose energies are larger than the minimum energy are called excited states of the BEC and are not stable in general but may satisfy relaxed concepts of stability such as metastability (see [36]). Numerical approaches for the computation of ground states of a BEC typically involve an iterative algorithm that starts with a given initial value and diminishes the energy of the density functional E in each iteration step. Different methodologies are possible: methods related to normalized gradient flows [5, 3, 1, 2, 5, 7, 24, 6, 9, 20], methods based on a direct minimization of the energy functional [8, 11], explicit imaginary-time marching [32], the DIIS method (direct inversion in the iterated subspace) [40, 16], or the optimal damping algorithm [14, 12]. We emphasize that, in any case, the dimensionality of the underlying space discretization is the crucial factor for computational complexity because it determines the cost per iteration step. The aim of this paper is to present a low-dimensional space discretization that reduces the cost per step and, hence, speeds up the iterative solution procedure considerably. In the literature, there are only a few contributions on rigorous numerical analysis of space discretizations of the GPE. In particular, explicit orders of convergence are widely missing. In [44, 17], Zhou and coworkers proved the convergence of general finite dimensional approximations that were obtained by minimizing the energy density E in a finite dimensional subspace of $H^1_0(\Omega)$. This justifies, e.g., the direct minimization approach proposed in [8]. The iteration scheme is not specified and not part of the analysis. The results of Zhou were generalized by Cancès, Chakir, and Maday [13] allowing explicit convergence rates for finite element approximations and Fourier expansions. A priori error estimates for a conservative Crank–Nicolson finite difference method and a semi-implicit finite difference method were derived by Bao and Cai [4].

In this work, we propose a new space discretization strategy that involves a pre-processing step and a postprocessing step in standard $P1$ finite element spaces. The preprocessing step is based on the numerical upscaling procedure suggested by Målqvist and Peterseim [33] for linear eigenvalue problems. In this step, a low-dimensional approximation space is assembled. The assembling is based on some local orthogonal decomposition that incorporates problem-specific information. The constructed space exhibits high approximation properties. The nonlinear problem is

then solved in this low-dimensional space by some standard iterative scheme (e.g., the ODA [14]) with very low cost per iteration step. The postprocessing step is based on the two-grid method suggest by Xu and Zhou [42]. We emphasize that both pre- and postprocessing involve only the solution of linear elliptic Poisson-type problems using standard finite elements. We give a rigorous error analysis for our strategy to show that we can achieve convergence orders of H^4 for the computed eigenvalue approximations without any preasymptotic effects. We do not focus on the iterative scheme that is used for solving the discrete minimization problem. The various choices previously mentioned, e.g., the ODA [14], are possible. Our new strategy is particularly beneficial in experimental setups with different types of bosons, because the results of the preprocessing step can be reused over and over again independently of β. Similarly, the data gained by preprocessing can be recycled for the computation of excited states. Other applications include setups with potentials that oscillate at a very high frequency (e.g., to investigate Josephson effects [41, 43]). Here, normally very fine grids are required to resolve the oscillations, whereas our strategy still yields good approximations in low-dimensional spaces and, hence, reduces the costs within the iteration procedure tremendously.

2. Model problem. Consider the dimensionless GPE in some bounded Lipschitz domain $\Omega \subset \mathbb{R}^d$, where $d = 1, 2, 3$. Since ground state solutions show an extremely fast decay (typically exponential), the restriction to bounded domains and homogeneous Dirichlet condition are physically justified. We seek (in the sense of distributions) the minimal eigenvalue λ and corresponding L^2-normalized eigenfunction $u \in H^1_0(\Omega)$ with

$$\begin{aligned} -\operatorname{div} A\nabla u + bu + \beta|u|^2 u &= \lambda u \quad \text{in } \Omega, \\ u &= 0 \quad \text{on } \partial\Omega. \end{aligned}$$

The underlying data satisfies the following assumptions:

(a) If $d = 1$, the domain Ω is an interval. If $d = 2$ (resp., $d = 3$), Ω has a polygonal (resp., polyhedral) boundary.

(b) The diffusion coefficient $A \in L^\infty(\Omega, \mathbb{R}^{d\times d}_{sym})$ is a symmetric matrix-valued function with uniform spectral bounds $\gamma_{\max} \geq \gamma_{\min} > 0$,

$$\sigma(A(x)) \subset [\gamma_{\min}, \gamma_{\max}] \quad \text{for almost all } x \in \Omega. \tag{2.1}$$

(c) $b \in L^2(\Omega)$ is nonnegative (almost everywhere).

(d) $\beta \in \mathbb{R}$ is nonnegative.

The weak solution of the GPE minmizes the energy functional $E: H^1_0(\Omega) \to \mathbb{R}$ given by

$$E(\phi) := \frac{1}{2}\int_\Omega A\nabla\phi \cdot \nabla\phi \, dx + \frac{1}{2}\int_\Omega b\phi^2 \, dx + \frac{1}{4}\int_\Omega \beta|\phi|^4 \, dx \quad \text{for } \phi \in H^1_0(\Omega).$$

Problem 2.1 (weak formulation of the GPE). Find $u \in H^1_0(\Omega)$ such that $u \geq 0$ a.e. in Ω, $\|u\|_{L^2(\Omega)} = 1$, and

$$E(u) = \inf_{\substack{v \in H^1_0(\Omega) \\ \|v\|_{L^2(\Omega)}=1}} E(v).$$

It is well known (see, e.g., [31] and [13]) that there exists a unique solution $u \in H^1_0(\Omega)$ of Problem 2.1. This solution u is continuous in $\bar{\Omega}$ and positive in Ω. The

corresponding eigenvalue $\lambda := 2E(u) + 2^{-1}\beta\|u\|^4_{L^4(\Omega)}$ of the GPE is real, positive, and simple. Observe that the eigenpair (u, λ) satisfies

$$\int_\Omega A\nabla u \cdot \nabla\phi \, dx + \int_\Omega bu\phi \, dx + \int_\Omega \beta|u|^2 u\phi \, dx = \lambda \int_\Omega u\phi \, dx$$

for all $\phi \in H^1_0(\Omega)$. Moreover, λ is the smallest among all possible eigenvalues and satisfies the a priori bound $\lambda < 4E(u)$.

3. Discretization. This section recalls classical finite element discretizations and presents novel two-grid approaches for the numerical solution of Problem 2.1. The existence of a minimizer of the functional E in discrete spaces is easily seen. However, uniqueness does not hold in general. We note that unlike as claimed in [44] the uniqueness proof given in [31] does not generalize to arbitrary subspaces of the original solution space.

Remark 3.1 (existence of discrete solutions [13]). Let W denote a finite dimensional, nonempty subspace of $H^1_0(\Omega)$; then there exists a minimizer $u_W \in W$ with $\|u_W\|_{L^2(\Omega)} = 1$, $(u_W, 1)_{L^2(\Omega)} \geq 0$, and

$$E(u_W) = \inf_{\substack{w \in W \\ \|w\|_{L^2(\Omega)}=1}} E(w).$$

If $(W_i)_{i\in\mathbb{N}}$ represents a dense family of such subspaces, then any sequence of corresponding minimizers $(u_i)_{i\in\mathbb{N}}$ with $(u_i, 1)_{L^2(\Omega)} \geq 0$ converges to the unique solution u of Problem 2.1.

3.1. Standard finite elements. We consider two regular simplicial meshes $\mathcal{T}_H$ and $\mathcal{T}_h$ of Ω. The finer mesh $\mathcal{T}_h$ is obtained from the coarse mesh $\mathcal{T}_H$ by regular mesh refinement. The discretization parameters $h \leq H$ represent the mesh size, i.e., $h_T := \operatorname{diam}(T)$ (resp., $H_T := \operatorname{diam}(T)$) for $T \in \mathcal{T}_h$ (resp., $\mathcal{T}_H$) and $h := \max_{T\in\mathcal{T}_h}\{h_T\}$ (resp., $H := \max_{T\in\mathcal{T}_H}\{H_T\}$). For $\mathcal{T} = \mathcal{T}_H, \mathcal{T}_h$, let

$$P_1(\mathcal{T}) = \{v \in L^2(\Omega) \mid \textit{for all } T \in \mathcal{T}, v|_T \text{ is a polynomial of total degree} \leq 1\}$$

denote the set of $\mathcal{T}$-piecewise affine functions. Classical $H^1_0(\Omega)$-conforming finite element spaces are then given by

$$V_h := P_1(\mathcal{T}_h) \cap H^1_0(\Omega) \quad \text{and} \quad V_H := P_1(\mathcal{T}_H) \cap H^1_0(\Omega) \subset V_h.$$

Note that on the fine discretization scale, a different choice of polynomial degree, e.g., piecewise quadratic functions, is possible. This would be a better choice for smooth data that allows for a regular ground state. Our method and its analysis essentially require the inclusion $H^1_0(\Omega) \supset V_h \supset V_H$. The discrete problem on the fine grid $\mathcal{T}_h$ reads as follows.

Problem 3.2 (reference finite element discretization on the fine mesh). Find $u_h \in V_h$ with $(u_h, 1)_{L^2(\Omega)} \geq 0$, $\|u_h\|_{L^2(\Omega)} = 1$ and

$$E(u_h) = \inf_{\substack{v_h \in V_h \\ \|v_h\|_{L^2(\Omega)}=1}} E(v_h). \tag{3.1}$$

The corresponding eigenvalue is given by $\lambda_h := 2E(u_h) + 2^{-1}\beta\|u_h\|^4_{L^4(\Omega)}$.

According to Remark 3.1, u_h is not determined uniquely in general. Moreover, λ_h is not necessarily the smallest eigenvalue of the corresponding discrete eigenvalue

problem. In what follows, u_h refers to an arbitrary solution of Problem 3.2. It will serve as a reference to compare further (cheaper) numerical approximations with. The accuracy of u_h has been studied in [13]. Under the assumption of sufficient regularity, optimal orders of convergence are obtained (cf. (4.5)).

3.2. Preprocessing motivated by numerical homogenization. The aim of this paper is to accurately approximate the fine scale reference solution u_h of Problem 3.2 within some low-dimensional subspace of V_h. For this purpose, we introduce a two-grid upscaling discretization that was initially proposed in [34] for the treatment of multiscale problems. The framework has been applied to nonlinear problems in [27], to linear eigenvalue problems in [33], and in the context of the discontinuous Galerkin [23] and partition of unity methods [28]. This contribution aims to generalize and analyze the methodology to the case of an eigenvalue problem with an additional nonlinearity in the eigenfunction. We emphasize that the coexistence of two difficulties, the nonlinear nature of the eigenproblem itself and the additional nonlinearity in the eigenfunction, requires new essential ideas far beyond simply plugging together existing theories for the isolated difficulties.

Let $\mathcal{N}_H$ denote the set of interior vertices in $\mathcal{T}_H$. For $z \in \mathcal{N}_H$ we let $\Phi_z \in V_H$ denote the corresponding nodal basis function with $\Phi_z(z) = 1$ and $\Phi_z(y) = 0$ for all $y \in \mathcal{N}_H \backslash \{z\}$. We define a weighted Clément-type interpolation operator (cf. [15])

$$I_H : H_0^1(\Omega) \to V_H, \quad v \mapsto I_H(v) := \sum_{z \in \mathcal{N}_H} v_z \Phi_z \quad \text{with } v_z := \frac{(v, \Phi_z)_{L^2(\Omega)}}{(1, \Phi_z)_{L^2(\Omega)}}. \tag{3.2}$$

It is easily shown by Friedrichs' inequality and the Sobolev embedding $H_0^1(\Omega) \hookrightarrow L^6(\Omega)$ (for $d \leq 3$) that

$$a(v, \phi) := \int_\Omega A \nabla v \cdot \nabla \phi \, dx + \int_\Omega b v \phi \, dx \qquad \text{for } v, \phi \in H_0^1(\Omega)$$

defines a scalar product in $H_0^1(\Omega)$ and induces a norm $\| \cdot \|_{H^1(\Omega)} := \sqrt{a(\cdot, \cdot)}$ on $H_0^1(\Omega)$ which is equivalent to the standard H^1-norm. By means of the interpolation operator I_H defined in (3.2), we construct an a-orthogonal decomposition of the space V_h into a low-dimensional coarse space $V_{H,h}^{\mathrm{c}}$ (with favorable approximation properties) and a high-dimensional residual space $V_{H,h}^{\mathrm{f}}$. The residual or "fine" space is the kernel of the interpolation operator restricted to V_h,

$$V_{H,h}^{\mathrm{f}} := \operatorname{kernel}(I_H|_{V_h}). \tag{3.3.a}$$

The coarse space is simply defined as the orthogonal complement of $V_{H,h}^{\mathrm{f}}$ in V_h with respect to $a(\cdot, \cdot)$. It is characterized via the a-orthogonal projection $P^{\mathrm{f}} : H_0^1(\Omega) \to V_{H,h}^{\mathrm{f}}$ onto the fine space given by

$$a(P^{\mathrm{f}} v, \phi) = a(v, \phi) \quad \text{for all } \phi \in V_{H,h}^{\mathrm{f}}.$$

By defining $P^{\mathrm{c}} := 1 - P^{\mathrm{f}}$, the coarse space is given by

$$V_{H,h}^{\mathrm{c}} := P^{\mathrm{c}} V_H. \tag{3.3.b}$$

A basis of $V_{H,h}^{\mathrm{c}}$ is given by $(P^{\mathrm{c}} \Phi_z)_{z \in \mathcal{N}_H}$ with $\dim V_{H,h}^{\mathrm{c}} = \dim V_H$. With this definition we obtain the splitting

$$V_h = V_{H,h}^{\mathrm{c}} \oplus V_{H,h}^{\mathrm{f}}. \tag{3.3.c}$$

Some favorable properties of the decomposition, in particular its L^2-quasi-orthogonality, are discussed in section 6.2. The minimization problem in the low-dimensional space $V^{\mathrm{c}}_{H,h}$ reads as follows.

Problem 3.3 (preprocessed approximation). Find $u^{\mathrm{c}}_H \in V^{\mathrm{c}}_{H,h}$ with $(u^{\mathrm{c}}_H, 1) \geq 0$, $\|u^{\mathrm{c}}_H\|_{L^2(\Omega)} = 1$, and

$$E(u^{\mathrm{c}}_H) = \inf_{\substack{v^{\mathrm{c}} \in V^{\mathrm{c}}_{H,h} \\ \|v^{\mathrm{c}}\|_{L^2(\Omega)}=1}} E(v^{\mathrm{c}}).$$

The corresponding eigenvalue in $V^{\mathrm{c}}_{H,h}$ is given by $\lambda^{\mathrm{c}}_H := 2E(u^{\mathrm{c}}_H) + 2^{-1}\beta\|u^{\mathrm{c}}_H\|^4_{L^4(\Omega)}$.

Remark 3.4 (practical aspects of the decomposition).

(a) The assembly of the corresponding finite element matrices requires only the evaluation of $P^{\mathrm{f}}\Phi_z$, i.e., the solution to one linear Poisson-type problem per coarse vertex. This can be done in parallel. Section 3.3 below will show that these linear problems may be restricted to local subdomains centered around the coarse vertices without loss of accuracy. Hence, even in a serial computing setup, the complexity of solving all corrector problems is equivalent (up to factor $|\log(H)|$) to the cost of solving one linear Poisson problem on the fine mesh.

(b) The preprocessing step is independent of the parameter β which characterizes the species of the bosons. Hence, the method becomes considerably cheaper when experiments need to be carried out for different types and numbers of bosons. A similar argument applies to variations on the trapping potential b. Provided that this trapping potential is an element of $H^1(\Omega)$ (in practical applications it is usually even harmonic and admits the desired regularity) the bilinear form $a(\cdot,\cdot)$ (and the associated constructions of $V^{\mathrm{f}}_{H,h}$ and $V^{\mathrm{c}}_{H,h}$) can be restricted to the second order term $\int_\Omega A\nabla v \cdot \nabla\phi$ without a loss in the expected convergence rates stated in Theorems 4.1 and 4.2 below. The trapping potential may then be varied without affecting the pre-processed space $V^{\mathrm{c}}_{H,h}$.

(c) Once the coarse space has been assembled it can also be reused in computations of larger eigenvalues (i.e., not only in the ground state solution).

3.3. Sparse approximations of $V^{\mathrm{c}}_{H,h}$. The construction of the coarse space $V^{\mathrm{c}}_{H,h}$ is based on fine scale equations formulated on the whole domain Ω, which makes them expensive to compute. However, [34] shows that $P^{\mathrm{f}}\Phi_z$ decays exponentially fast away from z. We specify this feature as follows. Let $k \in \mathbb{N}$ denote the localization parameter, i.e., a new discretization parameter. We define nodal patches $\omega_{z,k}$ of k coarse grid layers centered around the node $z \in \mathcal{N}_H$ by

$$\begin{aligned} \omega_{z,1} &:= \operatorname{supp}\Phi_z = \cup\{T \in \mathcal{T}_H \mid z \in T\}, \\ \omega_{z,k} &:= \cup\{T \in \mathcal{T}_H \mid T \cap \omega_{z,k-1} \neq \emptyset\} \quad \text{for } k \geq 2. \end{aligned} \tag{3.4}$$

There exists $0 < \theta < 1$ depending on the contrast $\gamma_{\min}/\gamma_{\max}$ but not on mesh sizes h, H and fast oscillations of A such that for all vertices $z \in \mathcal{N}_H$ and for all $k \in \mathbb{N}$, it holds that

$$\|P^{\mathrm{f}}\Phi_z\|_{H^1(\Omega\setminus\omega_{z,k})} \lesssim \theta^k \|P^{\mathrm{f}}\Phi_z\|_{H^1(\Omega)}. \tag{3.5}$$

This result motivates the truncation of the computations of the basis functions to local patches $\omega_{z,k}$. We approximate $\Psi_z = P^{\mathrm{f}}\Phi_z \in V^{\mathrm{f}}_{H,h}$ from (3.3.a)–(3.3.c) with $\Psi_{z,k} \in V^{\mathrm{f}}_{H,h}(\omega_{z,k}) := \{v \in V^{\mathrm{f}}_{H,h} \mid v|_{\Omega\setminus\omega_{x,k}} = 0\}$ such that

$$a(\Psi_{z,k}, v) = a(\Phi_z, v) \quad \text{for all } v \in V^{\mathrm{f}}_{H,h}(\omega_{z,k}). \tag{3.6}$$

This yields a modified coarse space $V^{\mathrm{c}}_{H,h,k}$ with a local basis

$$V^{\mathrm{c}}_{H,h,k} = \operatorname{span}\{\Phi_z - \Psi_{z,k} \mid z \in \mathcal{N}_H\}. \tag{3.7}$$

The number of nonzero entries of the corresponding finite element matrices is proportional to $k^d N_H$. (Note that we expect N_H^2 nonzero entries without the truncation.) Due to the exponential decay, the very weak condition $k \approx |\log H|$ implies that the perturbation of the ideal method due to this truncation is of higher order and forthcoming error estimates in Theorems 4.1 and 4.2 remain valid. We refer to [34] for details and proofs. The modified localization procedure from [29] with improved accuracy and stability properties may also be applied.

3.4. Postprocessing. Although u^{c}_H and λ^{c}_H will turn out to be highly accurate approximations of the unknown solution (u, λ), the orders of convergence can be improved even further by a simple postprocessing step on the fine grid. The postprocessing applies the two-grid method originally introduced by Xu and Zhou [42] for linear elliptic eigenvalue problems to the present equation by using our upscaled coarse space on the coarse level.

Problem 3.5 (postprocessed approximation). Find $u^{\mathrm{c}}_h \in V_h$ with

$$\int_\Omega A\nabla u^{\mathrm{c}}_h \cdot \nabla\phi_h \, dx + \int_\Omega b u^{\mathrm{c}}_h \phi_h \, dx = \lambda^{\mathrm{c}}_H \int_\Omega u^{\mathrm{c}}_H \phi_h \, dx - \int_\Omega \beta |u^{\mathrm{c}}_H|^2 u^{\mathrm{c}}_H \phi_h \, dx$$

for all $\phi_h \in V_h$. Define $\lambda^{\mathrm{c}}_h := (2E(u^{\mathrm{c}}_h) + 2^{-1}\beta\|u^{\mathrm{c}}_h\|^4_{L^4(\Omega)})\|u^{\mathrm{c}}_h\|^{-2}_{L^2(\Omega)}$. Let us emphasize that this approach is different from [18], where the postprecessing problem has a different structure and where classical finite element spaces are used on both scales.

4. A-priori error estimates. This section presents the a priori error estimates for the preprocessed/upscaled approximation with and without the postprocessing step. Throughout this section, $u \in H^1_0(\Omega)$ denotes the solution of Problem 2.1, $u_h \in V_h$ the solution of reference Problem 3.2, $u^{\mathrm{c}}_H \in V^{\mathrm{c}}_{H,h}$ the solution of Problem 3.3, and u^{c}_h the postprocessed solution of Problem 3.5. The notation $f \lesssim g$ abbreviates $f \le Cg$ with some constant C that may depend on the space dimension d, Ω, $\gamma_{\min}$, $\gamma_{\max}$, $\|b\|_{L^2(\Omega)}$, β, λ and interior angles of the triangulations, but not on the mesh sizes H and h. In particular it is robust against fast oscillations of A and b.

THEOREM 4.1 (error estimates for the preprocessed approximation). *Assume that* $\|u - u_h\|_{H^1(\Omega)} \lesssim 1$. *For* u *and* u^{c}_H *as above, it holds that*

$$\|u - u^{\mathrm{c}}_H\|_{H^1(\Omega)} \lesssim H^2 + \|u - u_h\|_{H^1(\Omega)}. \tag{4.1}$$

For sufficiently small h *(in the sense of Cancès, Chakir, and Maday et al.* [13]*), we also have*

$$|\lambda - \lambda^{\mathrm{c}}_H| + \|u - u^{\mathrm{c}}_H\|_{L^2(\Omega)} \lesssim H^3 + H\,\|u - u_h\|_{H^1(\Omega)}. \tag{4.2}$$

Proof. The proof is postponed to section 6.3. ∎

The additional postprocessing improves, roughly speaking, the order of accuracy by one.

THEOREM 4.2 (error estimates for the postprocessed approximation). *Assume that h is sufficiently small. The postprocessed approximation u_h^{c} and the postprocessed eigenvalue λ_h^{c} satisfy*

$$\|u - u_h^{\mathrm{c}}\|_{H^1(\Omega)} \lesssim H^3 + \|u - u_h\|_{H^1(\Omega)}, \tag{4.3}$$

$$|\lambda - \lambda_h^{\mathrm{c}}| + \|u - u_h^{\mathrm{c}}\|_{L^2(\Omega)} \lesssim H^4 + C_{L^2}(h, H). \tag{4.4}$$

The constant $C_{L^2}(h, H)$ behaves roughly like $H^2\|u - u_h\|_{H^1(\Omega)}$ and can be extracted from the proofs in section 6.4.2.

Proof. The proof is postponed to section 6.4. ∎

Let us emphasize that both theorems remain valid for $V_{H,h}^{\mathrm{c}}$ replaced with its sparse approximation $V_{H,h,k}^{\mathrm{c}}$ (cf. section 3.3) for moderate localization parameter $k \gtrsim |\log H|$.

We shall discuss the behavior of the fine scale errors $u - u_h$ and $\lambda - \lambda_h$. Recall from [13] that for a bounded domain Ω with polygonal Lipschitz boundary, $A \in [W^{1,\infty}(\Omega)]^{d\times d}$, and sufficiently small h, the fine scale error $\|u - u_h\|_{H^1(\Omega)}$ satisfies the optimal estimate

$$\|u - u_h\|_{H^1(\Omega)} + h^{-1}\|u - u_h\|_{L^2(\Omega)} + h^{-1}|\lambda - \lambda_h| \lesssim h. \tag{4.5}$$

The proof in [13] is for constant $A = 1$ and hyperrectangle Ω but it is easily checked that the estimates remain valid for any bounded domain Ω with polygonal Lipschitz boundary and $A \in [W^{1,\infty}(\Omega)]^{d\times d}$. Under these assumptions our a priori estimates for the postprocessed approximation of the ground state eigenvalue summarize as follows:

$$|\lambda - \lambda_h^{\mathrm{c}}| \lesssim H^4 + H^2 h.$$

Hence, in this regular setting, the choice $H = h^{1/2}$ ensures that the loss of accuracy is negligible when compared to the accuracy of the expensive full fine scale approximation λ_h. However, with regard to the numerical experiment in section 5.1 below, this choice might be pessimistic.

Moreover, note that the fine scale error depends crucially on higher Sobolev regularity of the solution, whereas our estimates for the coarse scale error require only minimal regularity that holds under assumptions (a)–(d) in section 2. Thus, we believe that in a less regular setting, even coarser choices of H relative to h will balance the discretization errors on the coarse and the fine scale.

5. Numerical experiments. Any numerical approach for the computation of ground states of a BEC involves an iterative algorithm that starts with a given initial value and diminishes the energy of the density functional E in each iteration step. In this contribution, we use the optimal damping algorithm (ODA) originally developed by Cancès and Le Bris [14, 12] for the Hartree–Fock equations, since it suits our preprocessing framework. The ODA involves solving a linear eigenvalue problem in each iteration step. However, after preprocessing, these linear eigenvalue problems are very low dimensional and the precomputed basis of $V_{H,h}^{\mathrm{c}}$ can be reused for each of these problems, making the iterations extremely cheap. The approximations produced by the ODA are known to rapidly converge to a solution of the discrete minimization problem. (See [21] and [12] for a proof in the setting of the Hartree–Fock equations.) All subsequent numerical experiments have been performed using MATLAB.

5.1. Numerical results for harmonic potential. In this section, we choose the smooth experimental setup of [13, section 4, p. 109; Figure 2], i.e., $\Omega := (0,\pi)^2$, $b(x_1,x_2) := x_1^2 + x_2^2$, $A = 1$, $\beta = 1$ and with homogeneous Dirichlet boundary condition. Our method depends basically on three parameters: the coarse mesh size H, the fine mesh size h, and the localization parameter k (cf. section 3.3 and [29]). In all computations of this section we couple k to the coarse mesh size by choosing $k = 2\log_2 H$. This choice is made such that the error of localization is negligible when compared with the errors committed by the fine scale discretization and the upscaling. All approximations are computed with the ODA method as presented in [21, section 2] with accuracy parameter $\varepsilon_{\mathrm{ODA}} = 10^{-14}$.

5.1.1. Comparison with full fine scale approximation. In the first experiment, we consider uniform coarse meshes $\mathcal{T}_H$ with mesh width parameters $H = 2^{-1}\pi, 2^{-2}\pi, \ldots, 2^{-4}\pi$ of Ω. The fine mesh $\mathcal{T}_h$ for the pre- and postprocessing has width $h = 2^{-7}\pi$ and remains fixed. We study the error committed by coarsening from a fine scale h to several coarse scales H, i.e., we study the distance between the ground state (u_h, λ_h) of Problem 3.2 and either the coarse scale approximation $(u_H^{\mathrm{c}}, \lambda_H^{\mathrm{c}})$ of Problem 3.3 (with underlying fine scale h) or its postprocessed version $(u_h^{\mathrm{c}}, \lambda_h^{\mathrm{c}})$ of Problem 3.5. Our theoretical results do not allow predictions about the coarsening error. Most likely, this is an artifact of our theory and we conjecture that (u_h, λ_h) and its coarse approximations $(u_H^{\mathrm{c}}, \lambda_H^{\mathrm{c}})$ and $(u_h^{\mathrm{c}}, \lambda_h^{\mathrm{c}})$ are in fact super-close in the sense of

$$\begin{aligned} H^{-1}\|u_h - u_H^{\mathrm{c}}\|_{H^1(\Omega)} + \|u_h - u_H^{\mathrm{c}}\|_{L^2(\Omega)} + |\lambda_h - \lambda_H^{\mathrm{c}}| &\lesssim H^3, \\ H^{-1}\|u_h - u_h^{\mathrm{c}}\|_{H^1(\Omega)} + \|u_h - u_h^{\mathrm{c}}\|_{L^2(\Omega)} + |\lambda_h - \lambda_h^{\mathrm{c}}| &\lesssim H^4. \end{aligned} \tag{5.1}$$

This assertion is true in the limit $h \to 0$. Section 5.1.2 supports numerically the assertion for positive h. Figure 1 reports the numerical results. Observe that the

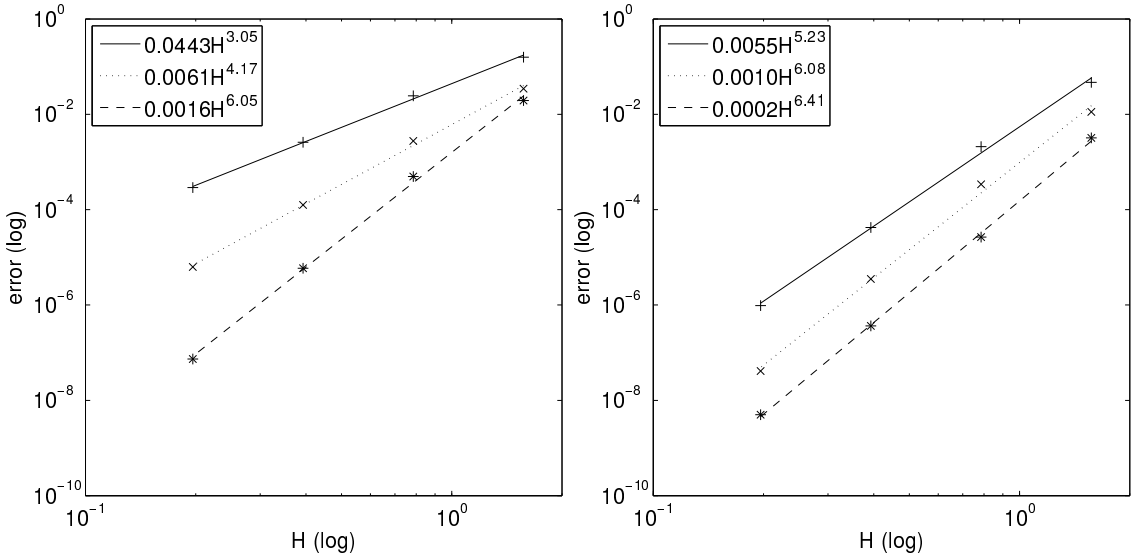


FIG. 1. *Results for harmonic potential. Left: Errors of preprocessed approximation* $\|u_h - u_H^{\mathrm{c}}\|_{H^1(\Omega)}$ *(+),* $\|u_h - u_H^{\mathrm{c}}\|_{L^2(\Omega)}$ *(×), and* $|\lambda_h - \lambda_H^{\mathrm{c}}|$ *(∗) versus coarse mesh size* H*. Right: Errors of postprocessed approximation* $\|u_h - u_h^{\mathrm{c}}\|_{H^1(\Omega)}$ *(+),* $\|u_h - u_h^{\mathrm{c}}\|_{L^2(\Omega)}$ *(×), and* $|\lambda_h - \lambda_h^{\mathrm{c}}|$ *(∗) versus coarse mesh size* H*.*

experimental rates with respect to H displayed in the figures are in fact better than the rates indicated by Theorems 4.1–4.2 and conjectured in (5.1). The reason could be the high regularity of the underlying (exact) solution $u \in H^3(\Omega)$. We do not exploit additional regularity in our error analysis. Similar observations have been made for the linear eigenvalue problem; see [33, Remark 3.3] for details and some justification of higher rates under additional regularity assumptions. Our implementation is not yet adequate for a fair comparison with regard to computational complexity and computing times between standard fine scale finite elements and our two-level techniques. However, to convince the reader of the potential savings in our new approach, let us mention that the number of iterations of the ODA was basically the same for both approaches in all numerical experiments. This statement applies as well to more challenging setups with larger values of β (see, e.g., section 5.2 below), where ODA needs many iterations to fall below some prescribed tolerance. We thus conclude that the actual speed-up of our approach is truly reflected by the dimension reduction from h^{-d} to H^{-d} up to the overhead $\mathcal{O}(k) = \mathcal{O}(\log |H|)$ induced by slightly denser (but still sparse) finite element matrices on the coarse level.

5.1.2. Comparison with high-resolution numerical approximation. In the second experiment we investigate the role of the fine scale parameter h. We consider uniform coarse meshes $\mathcal{T}_H$ with mesh width parameters $H = 2^{-1}\pi, \dots, 2^{-3}\pi$ and uniform fine meshes $\mathcal{T}_h$ for $h = H/4, \dots, 2^{-7}\pi$ for pre- and postprocessing computations. The error between the exact eigenvalue λ and coarse approximations λ_H^{c} and λ_h^{c} is estimated via a high-resolution numerical solution on a mesh of width $2^{-9}\pi$. The results are reported in Figure 2. For clarity, we show eigenvalue errors only. We conclude that it would have been sufficient to choose $H \approx h^{1/3}$ to achieve the accuracy of λ_h by our coarse approximation scheme with postprocessing.

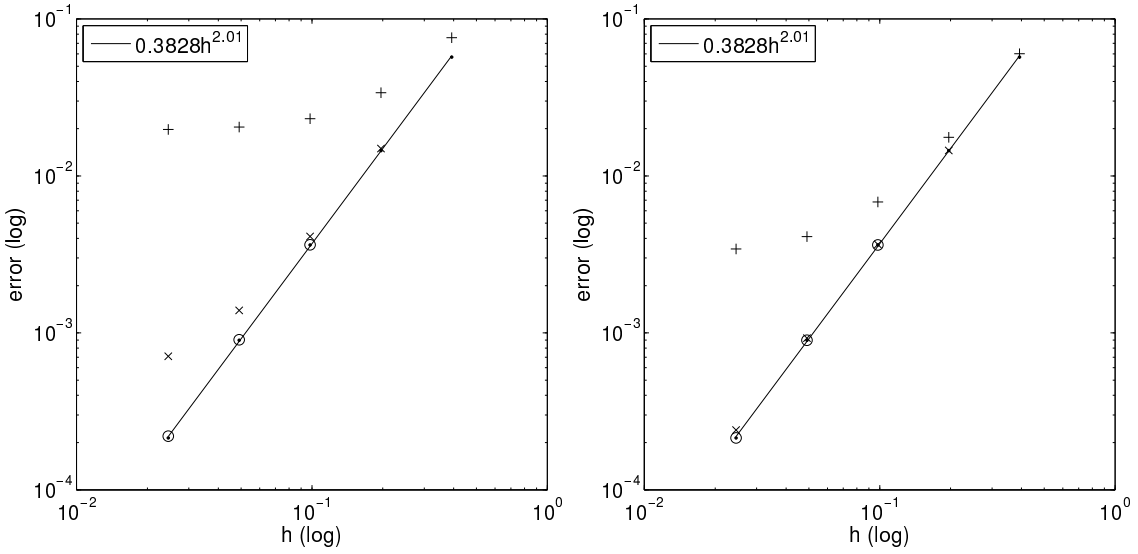


FIG. 2. *Results for harmonic potential. Left: (estimated) errors of preprocessed approximation* $|\lambda - \lambda_H^{\mathrm{c}}|$ *for fixed values* $H = 2^{-1}\pi$ (+), $H = 2^{-2}\pi$ (×), *and* $H = 2^{-3}\pi$ (○) *versus fine mesh size* h. *Right: (estimated) errors of postprocessed approximation* $|\lambda - \lambda_h^{\mathrm{c}}|$ *for fixed values* $H = 2^{-1}\pi$ (+), $H = 2^{-2}\pi$ (×), *and* $H = 2^{-3}\pi$ (○) *versus fine mesh size* h. *In both plots, the (estimated) error of the standard FEM on the fine mesh* $|\lambda - \lambda_h|$ (●) *is depicted for reference.*

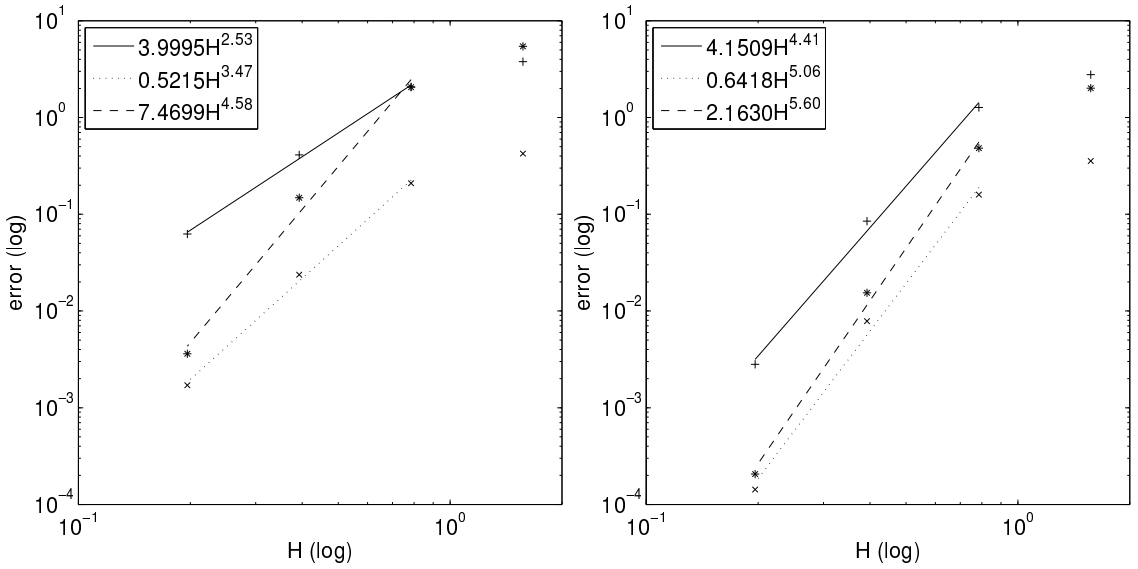


FIG. 3. *Results for periodic potential. Left: Errors of preprocessed approximation* $\|u_h - u_H^{\mathrm{c}}\|_{H^1(\Omega)}$ *(+),* $\|u_h - u_H^{\mathrm{c}}\|_{L^2(\Omega)}$ *(×), and* $|\lambda_h - \lambda_H^{\mathrm{c}}|$ *(∗) versus coarse mesh size* H*. Right: Errors of postprocessed approximation* $\|u_h - u_h^{\mathrm{c}}\|_{H^1(\Omega)}$ *(+),* $\|u_h - u_h^{\mathrm{c}}\|_{L^2(\Omega)}$ *(×), and* $|\lambda_h - \lambda_h^{\mathrm{c}}|$ *(∗) versus coarse mesh size* H*.*

5.2. Numerical results for discontinuous periodic potential. This section addresses the case of a BEC that is trapped in a periodic potential. Periodic potentials are of special interest since they can be used to explore physical phenomena such as Josephson oscillations and macroscopic quantum self-trapping of the condensate (cf. [41, 43]). Here we use a potential b that describes a periodic array of quantum wells that can be experimentally generated by the interference of overlapping laser beams (cf. [39]).

Let $\Omega = (0,\pi)^2$, $A = 1$, and $\beta = 4$. Given $b_t = 100$ and $L = 4$, define

$$b_0(x_1,x_2) := \begin{cases} 0 & \text{for } x \in \left]\frac{1}{4},\frac{3}{4}\right[^2, \\ b_t & \text{else} \end{cases}$$

and the potential $b(x) = b_0\big(L\big(x/\pi - \frac{\lfloor Lx/\pi \rfloor}{L}\big)\big)$.

Consider the same numerical setup as in section 5.1.1 (i.e., we draw our attention again to the coarsening error $u_h - u_H^{\mathrm{c}}$) with the exception that we were able to reduce the localization parameter $k = \log_2 H$ without affecting the best convergence rates possible. Figure 3 reports the errors between the fine scale reference discretization and our coarse approximations. For the discontinuous potential, the experimental rates (with respect to H) are slightly worse than those ones observed in section 5.1.1. However, they are still better than the rates indicated by Theorems 4.1–4.2 and conjectured in (5.1).

6. Proofs of the main results. In this section we are concerned with proving the main theorems.

6.1. Auxiliary results. An application of [13, Theorem 1] shows that u_h and u_H^{c} both converge to u in $H^1(\Omega)$, which guarantees stability.

Remark 6.1 (stability of discrete approximations). For sufficiently small h we have

$$\|u_h\|_{H^1(\Omega)} \leq \sqrt{\lambda_h} \lesssim \sqrt{\lambda} \qquad \text{and} \tag{6.1}$$

$$||u_h||_{L^4(\Omega)} \leq \left(\frac{\lambda_h}{\beta}\right)^{\frac{1}{4}} \lesssim \left(\frac{\lambda}{\beta}\right)^{\frac{1}{4}}. \tag{6.2}$$

The same results hold for u_h replaced by u_H^{c} and λ_h replaced by λ_H^{c} for h and H sufficiently small.

The bound (6.1) is obvious using $\|u_h\|_{L^2(\Omega)} = 1$ and the H^1-convergence $u_h \to u$ which guarantees $\lambda_h \to \lambda$. Estimate (6.2) directly follows from the definitions of λ_h and E_h which gives us $\lambda_h \geq 2E(u_h) = a(u_h, u_h) + \frac{\beta}{2}\|u_h\|_{L^4(\Omega)}^4 \geq \frac{\beta}{2}\|u_h\|_{L^4(\Omega)}^4$.

Remark 6.2 (L^∞-bound). The solution u of Problem 2.1 is in $L^\infty(\Omega)$. This follows from the uniqueness of $u \in H_0^1(\Omega)$, which shows that it is also the unique solution of the linear elliptic problem

$$\int_\Omega A\nabla u \cdot \nabla\phi + bu\phi \, dx = \int_\Omega \tilde{f}\phi \, dx \qquad \text{for all } \phi \in H_0^1(\Omega),$$

where $\tilde{f} := (\lambda u - \beta|u|^3) \in L^2(\Omega)$. Standard theory for linear elliptic problems (cf. [25, Theorem 8.15, pp. 189–193]) then yields the existence of a constant c only depending on Ω, d and $\|\gamma_{\min}^{-1} b\|_{L^2(\Omega)}$ such that

$$\|u\|_{L^\infty(\Omega)} \leq c(\|u\|_{L^2(\Omega)} + \gamma_{\min}^{-1}\|\tilde{f}\|_{L^2(\Omega)}) \lesssim 1 + \|u\|_{L^6(\Omega)}^3 \lesssim 1 + \|u\|_{H^1(\Omega)}^3. \tag{6.3}$$

6.2. Properties of the coarse space $V_{H,h}^{\mathrm{c}}$. Recall the local approximation properties of the weighted Clément-type interpolation operator I_H defined in (3.2),

$$H_T^{-1}\|v - I_H(v)\|_{L^2(T)} + \|\nabla(v - I_H(v))\|_{L^2(T)} \leq C_{I_H}\|\nabla v\|_{L^2(\omega_T)} \tag{6.4}$$

for all $v \in H_0^1(\Omega)$. Here, C_{I_H} is a generic constant that depends only on interior angles of $\mathcal{T}_H$ but not on the local mesh size and $\omega_T := \bigcup\{S \in \mathcal{T}_H | \, \overline{S} \cap \overline{T} \neq \emptyset\}$. Furthermore, for all $v \in H_0^1(\Omega)$ and for all $z \in \mathcal{N}_H$ it holds that

$$\int_{\omega_z} (v - v_z)^2 \, dx \leq C_{I_H} H^2 \|\nabla v\|_{L^2(\omega_z)}^2, \tag{6.5}$$

where $\omega_z := \mathrm{supp}(\Phi_z)$ and v_z is given by (3.2).

LEMMA 6.3 (properties of the decomposition). *The decomposition of V_h into V_H and $V_{H,h}^{\mathrm{f}}$ (stated in section 3.2) is L^2-orthogonal, i.e.,*

$$V_h = V_H \oplus V_{H,h}^{\mathrm{f}} \quad \textit{and} \quad (v_H, v^{\mathrm{f}})_{L^2(\Omega)} = 0 \ \textit{for all } v_H \in V_H, \ v^{\mathrm{f}} \in V_{H,h}^{\mathrm{f}}. \tag{6.6}$$

The decomposition of V_h in $V_{H,h}^{\mathrm{c}}$ and $V_{H,h}^{\mathrm{f}}$ is a-orthogonal

$$V_h = V_{H,h}^{\mathrm{c}} \oplus V_{H,h}^{\mathrm{f}} \quad \textit{and} \quad a(v^{\mathrm{c}}, v^{\mathrm{f}}) = 0 \ \textit{for all } v^{\mathrm{c}} \in V_{H,h}^{\mathrm{c}}, \ v^{\mathrm{f}} \in V_{H,h}^{\mathrm{f}} \tag{6.7}$$

and L^2-quasi-orthogonal in the sense that

$$(v^{\mathrm{c}}, v^{\mathrm{f}})_{L^2(\Omega)} \lesssim H^2\|\nabla v^{\mathrm{c}}\|_{L^2(\Omega)}\|\nabla v^{\mathrm{f}}\|_{L^2(\Omega)}. \tag{6.8}$$

Proof. The proof is verbatim the same as in [33]. ∎

The following lemma estimates the error of the best approximation in the modified coarse space $V_{H,h}^{\mathrm{c}}$. The lemma is also implicitly required each time that we use the abstract error estimates stated in [13, Theorem 1]. These estimates require a family of finite dimensional spaces that is dense in $H_0^1(\Omega)$. This density property is implied by the following lemma.

LEMMA 6.4 (approximation property of $V_{H,h}^{\mathrm{c}}$). *For any given $v \in H_0^1(\Omega)$ with $\operatorname{div} A\nabla v \in L^2(\Omega)$ it holds that*

$$\inf_{v_H^{\mathrm{c}} \in V_{H,h}^{\mathrm{c}}} \|v - v_H^{\mathrm{c}}\|_{H^1(\Omega)} \lesssim H \|\operatorname{div} A\nabla v + bv\|_{L^2(\Omega)} + \inf_{v_h \in V_h} \|v - v_h\|_{H^1(\Omega)}.$$

Proof. Given v, define $f_v := \operatorname{div} A\nabla v + bv \in L^2(\Omega)$ (since $v \in L^\infty(\Omega)$) and let $v_h \in V_h$ denote the corresponding finite element approximation, i.e.,

$$a(v_h, \phi_h) = (f_v, \phi_h)_{L^2(\Omega)} \qquad \text{for all } \phi_h \in V_h.$$

With $v_H^{\mathrm{c}} := P^{\mathrm{c}} v_h \in V_{H,h}^{\mathrm{c}}$, Galerkin orthogonality leads to

$$\begin{aligned} \|A^{1/2}\nabla(v_h - v_H^{\mathrm{c}})\|_{L^2(\Omega)}^2 &\overset{(6.7)}{\leq} a(v_h, P^{\mathrm{f}} v_h) = (f_v, P^{\mathrm{f}} v_h)_{L^2(\Omega)} \\ &\overset{(6.4)}{\lesssim} \gamma_{\min}^{-1/2} \|H f_v\|_{L^2(\Omega)} \|A^{1/2}\nabla(v_h - v_H^{\mathrm{c}})\|_{L^2(\Omega)}. \end{aligned}$$

This, the triangle inequality, and norm equivalences readily yield the assertion. ∎

Next, we show that there exists an element $u^{\mathrm{c}} = P^{\mathrm{c}} u_h$ in the space $V_{H,h}^{\mathrm{c}}$ that approximates u_h in the energy norm with an accuracy of order $\mathrm{O}(H^2)$.

LEMMA 6.5 (stability and approximability of the reference solution). *Let $(u_h, \lambda_h) \in V_h \times \mathbb{R}$ solve Problem 3.2. Then it holds that*

$$\begin{aligned} \|P^{\mathrm{c}} u_h\|_{H^1(\Omega)} &\leq \sqrt{\lambda_h}, \\ \|P^{\mathrm{c}} u_h - u_h\|_{H^1(\Omega)} = \|P^{\mathrm{f}} u_h\|_{H^1(\Omega)} &\lesssim H^2 + H\|u - u_h\|_{H^1(\Omega)}, \\ (P^{\mathrm{c}} u_h, P^{\mathrm{f}} u_h)_{L^2(\Omega)} &\lesssim \left(H^2 + H\|u - u_h\|_{H^1(\Omega)}\right) H^2. \end{aligned}$$

Proof. Recall $\|\cdot\|_{H^1(\Omega)} := \sqrt{a(\cdot,\cdot)}$. Since P^{c} is a projection, we have

$$\|P^{\mathrm{c}} u_h\|_{H^1(\Omega)}^2 \leq \|u_h\|_{H^1(\Omega)}^2 = \lambda_h \|u_h\|_{L^2(\Omega)}^2 - \beta \|u_h\|_{L^4(\Omega)}^4 \leq \lambda_h.$$

The a-orthogonality of (3.3.c) further yields

$$\begin{aligned} \|P^{\mathrm{f}} u_h\|_{H^1(\Omega)}^2 &= a(P^{\mathrm{f}} u_h, P^{\mathrm{f}} u_h) = a(u_h, P^{\mathrm{f}} u_h) \\ &= \lambda_h (u_h, (1 - I_H) P^{\mathrm{f}} u_h)_{L^2(\Omega)} - \beta (u^3, P^{\mathrm{f}} u_h)_{L^2(\Omega)} \\ &\quad - \beta (u_h^3 - u^3, P^{\mathrm{f}} u_h)_{L^2(\Omega)}. \end{aligned} \tag{6.9}$$

The first term on the right-hand side of (6.9) can be bounded using $I_H(P^{\mathrm{f}} u_h) = 0$, the L^2-orthogonality (6.6), and the estimates for the weighted Clément interpolation operator (6.4)

$$\begin{aligned} \lambda_h (u_h, (1 - I_H) P^{\mathrm{f}} u_h)_{L^2(\Omega)} &= \lambda_h ((1 - I_H) u_h, (1 - I_H) P^{\mathrm{f}} u_h)_{L^2(\Omega)} \\ &\lesssim \lambda_h H^2 \|u_h\|_{H^1(\Omega)} \|P^{\mathrm{f}} u_h\|_{H^1(\Omega)}. \end{aligned} \tag{6.10}$$

Since $u \in L^\infty(\Omega)$ we have $\nabla(u^3) = 3u^2\nabla u \in L^2(\Omega)$ and, hence, the second term on the right-hand side of (6.9) can be bounded as follows:

(6.11)
$$\beta(u^3, P^{\mathrm{f}}u_h) = \beta((1-I_H)u^3, (1-I_H)P^{\mathrm{f}}u_h) \overset{(6.5)}{\lesssim} H^2\|u\|^2_{L^\infty(\Omega)}\|u\|_{H^1(\Omega)}\|P^{\mathrm{f}}u\|_{H^1(\Omega)} \overset{(6.3)}{\lesssim} H^2\|u\|_{H^1(\Omega)}\|P^{\mathrm{f}}u\|_{H^1(\Omega)}.$$

Since $u_h^3 - u^3 = (u_h^2 + u_h u + u^2)(u_h - u)$, the third term on the right-hand side of (6.9) can be estimated by

$$\beta(u_h^3 - u^3, P^{\mathrm{f}}u_h)_{L^2(\Omega)} \lesssim \||u| + |u_h|\|^2_{L^6(\Omega)}\|u_h - u\|_{L^6(\Omega)}\|(1-I_H)P^{\mathrm{f}}u_h\|_{L^2(\Omega)} \lesssim H\|u - u_h\|_{H^1(\Omega)}\|P^{\mathrm{f}}u_h\|_{H^1(\Omega)}, \tag{6.12}$$

where we used (6.1) and the embedding $\||u| + |u_h|\|_{L^6(\Omega)} \lesssim \|u\|_{H^1(\Omega)} + \|u_h\|_{H^1(\Omega)}$.

The combination of (6.9)–(6.12) readily yields

$$\|P^{\mathrm{f}}u_h\|_{H^1(\Omega)} \lesssim H^2 + \|u - u_h\|^2_{H^1(\Omega)}.$$

The third assertion follows from the previous ones and

$$(P^{\mathrm{c}}u_h, P^{\mathrm{f}}u_h)_{L^2(\Omega)} = ((1-I_H)P^{\mathrm{c}}u_h, (1-I_H)P^{\mathrm{f}}u_h)_{L^2(\Omega)} \lesssim H^2\|P^{\mathrm{c}}u_h\|_{H^1(\Omega)}\|P^{\mathrm{f}}u_h\|_{H^1(\Omega)}. \qquad \square$$

6.3. Proof of Theorem 4.1. We split the proof into two parts: the estimate for the H^1-error and the estimate for the L^2-error.

6.3.1. Proof of the H^1-error estimate (4.1). We proceed similarly as in [13]. The proof is divided into four steps. In the first step, we derive an identical formulation of some energy difference. The identity is used in Step 2 to establish the inequality $\|u_h^{\mathrm{c}} - u\|^2_{H^1(\Omega)} \lesssim E(u_h^{\mathrm{c}}) - E(u)$. Since u_h^{c} is a minimizer, we can replace $E(u_h^{\mathrm{c}})$ by $E(w_h^{\mathrm{c}})$ in the estimate for an arbitrary L^2-normalized $w_h^{\mathrm{c}} \in V_{H,h}^{\mathrm{c}}$. In Step 3, we choose $w_h^{\mathrm{c}} := \frac{P^{\mathrm{c}}u_h}{\|P^{\mathrm{c}}u_h\|_{L^2(\Omega)}}$ and show that the perturbation introduced via normalization is of high order ($\approx H^3$). In Step 4, we use Step 3 to estimate $E(w_h^{\mathrm{c}}) - E(u)$.

Step 1. Given some arbitrary $w \in H^1_0(\Omega)$ with $\|w\|_{L^2(\Omega)} = 1$, we show that

$$E(w) - E(u) = \frac{1}{2}a(w-u, w-u) + \frac{\beta}{2}(|u|^2(w-u), w-u)_{L^2(\Omega)} + \frac{\beta}{4}((|u|^4 - 2|u|^2|w|^2 + |w|^4, 1)_{L^2(\Omega)} - \frac{1}{2}\lambda\|w-u\|^2_{L^2(\Omega)}. \tag{6.13}$$

First, using $\|u\|_{L^2(\Omega)} = \|w\|_{L^2(\Omega)} = 1$ we get

$$\lambda(u-w, u-w)_{L^2(\Omega)} = \lambda\|u\|^2_{L^2(\Omega)} - 2\lambda(u,w)_{L^2(\Omega)} + \lambda\|u\|^2_{L^2(\Omega)} = -2\lambda(u, w-u)_{L^2(\Omega)} = -2a(u, w-u) - 2\beta(|u|^2u, w-u)_{L^2(\Omega)}. \tag{6.14}$$

This yields

$$\begin{aligned} & a(w,w)+\beta(|u|^2w,w)_{L^2(\Omega)}-a(u,u)-\beta(|u|^2u,u)_{L^2(\Omega)} \\ \overset{(6.14)}{=}\ & a(w,w)-2a(u,w)+a(u,u) \\ & +\beta(|u|^2w,w)_{L^2(\Omega)}-2\beta(|u|^2u,w)_{L^2(\Omega)}+\beta(|u|^2u,u)_{L^2(\Omega)} \\ & -\lambda(w-u,w-u)_{L^2(\Omega)} \\ =\ & a(w-u,w-u)+\beta(|u|^2(w-u),w-u)_{L^2(\Omega)}-\lambda\|w-u\|^2_{L^2(\Omega)}. \end{aligned}$$

Plugging this last equality into the equation

$$2E(w)-2E(u)=a(w,w)+\frac{\beta}{2}(|w|^2w,w)_{L^2(\Omega)}-a(u,u)-\frac{\beta}{2}(|u|^2u,u)_{L^2(\Omega)}$$

leads to (6.13).

Step 2. Using (6.13) with $w=u_h^{\mathrm{c}}$ and the fact that there exists some c_0 (independent of H and h) such that $a(u-u_h^{\mathrm{c}},u-u_h^{\mathrm{c}})+((\beta|u|^2-\lambda)(u-u_h^{\mathrm{c}}),u-u_h^{\mathrm{c}})_{L^2(\Omega)}\geq c_0\|u-u_h^{\mathrm{c}}\|^2_{H^1(\Omega)}$ (cf. [13, Lemma 1]), we get

$$\begin{aligned} E(u_h^{\mathrm{c}})-E(u) &= \frac{1}{2}a(u_h^{\mathrm{c}}-u,u_h^{\mathrm{c}}-u)+\frac{\beta}{2}(|u|^2(u_h^{\mathrm{c}}-u),u_h^{\mathrm{c}}-u)_{L^2(\Omega)} \\ &\quad +\frac{\beta}{4}(|u|^4-2|u|^2|u_h^{\mathrm{c}}|^2+|u_h^{\mathrm{c}}|^4,1)_{L^2(\Omega)}-\frac{1}{2}\lambda\|u_h^{\mathrm{c}}-u\|^2_{L^2(\Omega)} \\ &\geq \frac{c_0}{2}\|u_h^{\mathrm{c}}-u\|^2_{H^1(\Omega)}+\frac{\beta}{4}\||u|^2-|u_h^{\mathrm{c}}|^2\|^2_{L^2(\Omega)}. \end{aligned}$$

Step 3. Using the result of step two yields

$$\|u_h^{\mathrm{c}}-u\|^2_{H^1(\Omega)}\lesssim E(u_h^{\mathrm{c}})-E(u)\leq E(w_h^{\mathrm{c}})-E(u)$$

for any L^2-normalized $w_h^{\mathrm{c}}\in V_{H,h}^{\mathrm{c}}$. We choose $w_h^{\mathrm{c}}:=\frac{P^{\mathrm{c}}u_h}{\|P^{\mathrm{c}}u_h\|_{L^2(\Omega)}}$ and observe that we get, with Lemma 6.5, that

(6.15)
$$\begin{aligned} \|P^{\mathrm{c}}u_h-w_h^{\mathrm{c}}\|_{L^2(\Omega)} &= \big|1-\|P^{\mathrm{c}}u_h\|_{L^2(\Omega)}\big|\leq\|P^{\mathrm{f}}u_h\|_{L^2(\Omega)}=\|P^{\mathrm{f}}u_h-I_H(P^{\mathrm{f}}u_h)\|_{L^2(\Omega)} \\ &\lesssim H\|P^{\mathrm{f}}u_h\|_{H^1(\Omega)}\lesssim H\|u-u_h\|^2_{H^1(\Omega)}+H^3 \end{aligned}$$

and consequently

(6.16)
$$\|P^{\mathrm{c}}u_h-w_h^{\mathrm{c}}\|_{H^1(\Omega)}=\frac{\big|1-\|P^{\mathrm{c}}u_h\|_{L^2(\Omega)}\big|}{\|P^{\mathrm{c}}u_h\|_{L^2(\Omega)}}\|P^{\mathrm{c}}u_h\|_{H^1(\Omega)}\lesssim H\|u-u_h\|^2_{H^1(\Omega)}+H^3,$$

where we used $\|u-u_h\|_{H^1(\Omega)}\lesssim 1$ (implying $\|P^{\mathrm{c}}u_h\|_{H^1(\Omega)}\lesssim 1$ and $\|P^{\mathrm{c}}u_h\|_{L^2(\Omega)}\gtrsim 1$).

Step 4. Using again (6.13) leads to

$$\begin{aligned} 2E(w_h^{\mathrm{c}})-2E(u) &= \|w_h^{\mathrm{c}}-u\|^2_{H^1(\Omega)}+\beta(|u|^2(w_h^{\mathrm{c}}-u),w_h^{\mathrm{c}}-u)_{L^2(\Omega)} \\ &\quad +\frac{\beta}{2}(|u|^4-2|u|^2|w_h^{\mathrm{c}}|^2+|w_h^{\mathrm{c}}|^4,1)_{L^2(\Omega)}-\lambda\|w_h^{\mathrm{c}}-u\|^2_{L^2(\Omega)}. \end{aligned}$$

The Hölder inequality

$$(|u|^2, |u - w_h^{\mathrm{c}}|^2)_{L^2(\Omega)} \leq \|u\|^2_{L^6(\Omega)} \|u - w_h^{\mathrm{c}}\|_{L^2(\Omega)} \|u - w_h^{\mathrm{c}}\|_{L^6(\Omega)} \tag{6.17}$$

yields the estimate

$$\begin{aligned}
&\beta(|u|^2(w_h^{\mathrm{c}} - u), w_h^{\mathrm{c}} - u)_{L^2(\Omega)} + \frac{\beta}{2} \int_\Omega \left(|u|^2 - |w_h^{\mathrm{c}}|^2\right)^2 \, dx \\
&\quad \overset{(6.17)}{\leq} \beta \|u\|^2_{L^6(\Omega)} \|u - w_h^{\mathrm{c}}\|_{L^2(\Omega)} \|u - w_h^{\mathrm{c}}\|_{L^6(\Omega)} + \frac{\beta}{2} ((|u| + |w_h^{\mathrm{c}}|)^2, |u - w_h^{\mathrm{c}}|^2)_{L^2(\Omega)} \\
&\quad \overset{(6.17)}{\leq} \beta (2\|u\|^2_{L^6(\Omega)} + \|w_h^{\mathrm{c}}\|^2_{L^6(\Omega)}) \|u - w_h^{\mathrm{c}}\|_{L^2(\Omega)} \|u - w_h^{\mathrm{c}}\|_{L^6(\Omega)} \\
&\quad \lesssim \|u - w_h^{\mathrm{c}}\|^2_{L^2(\Omega)} + \|u - w_h^{\mathrm{c}}\|^2_{H^1(\Omega)}
\end{aligned}$$

for the terms involving β. The combination of the previous results with Lemma 6.5 and estimates (6.15) and (6.16) gives us

$$\begin{aligned}
\|u_h^{\mathrm{c}} - u\|^2_{H^1(\Omega)} &\lesssim E(u_h^{\mathrm{c}}) - E(u) \leq E(w_h^{\mathrm{c}}) - E(u) \lesssim \|w_h^{\mathrm{c}} - u\|^2_{H^1(\Omega)} \\
&\lesssim \|u - P^{\mathrm{c}} u_h\|^2_{H^1(\Omega)} + \|P^{\mathrm{c}} u_h - w_h^{\mathrm{c}}\|^2_{H^1(\Omega)} \\
&\lesssim \left(\|u - u_h\|_{H^1(\Omega)} + H^2\right)^2 .
\end{aligned}$$

6.3.2. Proof of the L^2-error estimate (4.2). In the following, we let the bilinear form $c_{\lambda,u} : H^1_0(\Omega) \times H^1_0(\Omega) \to \mathbb{R}$ be given by

$$c_{\lambda,u}(v, w) := \int_\Omega A\nabla v \cdot \nabla w + bvw + 3\beta |u|^2 vw \, dx - \lambda \int_\Omega vw \, dx$$

and we define the space

$$V_u^\perp := \{v \in H^1_0(\Omega) | \, (v, u)_{L^2(\Omega)} = 0\}.$$

For $w \in H^1_0(\Omega)$ we let $\psi_w \in V_u^\perp$ denote the unique solution (see Lemma 6.6 below) of

$$c_{\lambda,u}(\psi_w, v_\perp) = (w, v_\perp)_{L^2(\Omega)} \quad \text{for all } \, v_\perp \in V_u^\perp. \tag{6.18}$$

The subsequent lemma applies the abstract L^2-error estimate, obtained by Cancès, Chakir, and Maday [13, Lemma 1, Theorem 1, and Remark 2], to our setting. Observe that Lemma 6.4 (i.e., $V^{\mathrm{c}}_{H,h}$ represents a dense family of finite dimensional subspaces of H^1) is required to apply these results.

LEMMA 6.6 (abstract approximation [13]). *Let h be sufficiently small; then*

$$|\lambda - \lambda_H^{\mathrm{c}}| \lesssim \|u - u_H^{\mathrm{c}}\|^2_{H^1(\Omega)} + \|u - u_H^{\mathrm{c}}\|_{L^2(\Omega)} \tag{6.19}$$

and

$$\|u - u_H^{\mathrm{c}}\|^2_{L^2(\Omega)} \lesssim \|u - u_H^{\mathrm{c}}\|_{H^1(\Omega)} \inf_{\psi \in V^{\mathrm{c}}_{H,h}} \|\psi_{u_H^{\mathrm{c}} - u} - \psi\|_{H^1(\Omega)}. \tag{6.20}$$

Furthermore, the bilinear form $c_{\lambda,u}(\cdot,\cdot)$ is a scalar product in $H^1_0(\Omega)$ and induces a norm that is equivalent to the standard H^1-norm.

Observe the following equivalence. If $\psi_w \in V_u^\perp$ solves

$$\int_\Omega A\nabla \psi_w \cdot \nabla v_\perp + b\psi_w v_\perp + \beta 3|u|^2 \psi_w v_\perp \, dx - \lambda \int_\Omega \psi_w v_\perp \, dx = \int_\Omega w v_\perp \, dx$$

for all $v_\perp \in V_u^\perp$, then it also solves

$$\int_\Omega A\nabla\psi_w \cdot \nabla v + b\psi_w v + \beta 3|u|^2\psi_w v\,dx - \lambda\int_\Omega \psi_w v\,dx = 2\beta(u^3,\psi_w)_{L^2(\Omega)}\int_\Omega uv\,dx + \int_\Omega (w-(w,u)_{L^2(\Omega)})v\,dx$$

for all $v \in H_0^1(\Omega)$. This can be easily seen as follows: assume $\operatorname{div} A\nabla\psi_w \in L^2(\Omega)$ (the general result follows by density arguments) and let $P^\perp : L^2(\Omega) \to V_u^\perp$ denote the L^2-orthogonal projection given by $P^\perp(v) := v - (v,u)_{L^2(\Omega)}$. Since

$$\int_\Omega \left(-\operatorname{div} A\nabla\psi_w + b\psi_w + 3\beta|u|^2\psi_w - \lambda\psi_w\right) v^\perp\,dx = \int_\Omega wv^\perp\,dx,$$

we get

$$\int_\Omega P^\perp\left(-\operatorname{div} A\nabla\psi_w + b\psi_w + 3\beta|u|^2\psi_w - \lambda\psi_w\right) v\,dx = \int_\Omega P^\perp(w)v\,dx$$

for all $v \in H_0^1(\Omega)$. By using the explicit formula for $P^\perp$ and the definition of u the reformulated equation follows. Furthermore, since $\psi_w \in H_0^1(\Omega)$ solves a standard elliptic problem, classical theory (cf. [25]) applies and we get the L^∞-estimate

$$\|\psi_w\|_{L^\infty(\Omega)} \lesssim (1+\lambda)\|\psi_w\|_{L^2(\Omega)} + |(|u|^3,\psi_w)| + \|w\|_{L^2(\Omega)} \lesssim (1+\lambda)\|w\|_{L^2(\Omega)}. \tag{6.21}$$

LEMMA 6.7 (L^2-error estimate). *Let h be sufficiently small and let u denote the solution of Problem* 2.1*, u_H^{c} the solution of Problem* 3.3*, and $\psi_{u-u_H^{\mathrm{c}}} \in V_u^\perp$ the solution of* (6.18) *for $w = u - u_H^{\mathrm{c}}$. Then*

$$\|u-u_H^{\mathrm{c}}\|_{L^2(\Omega)} \lesssim \left(\min_{\psi^h\in V_h}\frac{\|\psi_{u-u_H^{\mathrm{c}}}-\psi^h\|_{H^1(\Omega)}}{\|u-u_H^{\mathrm{c}}\|_{L^2(\Omega)}} + H\right)\|u-u_H^{\mathrm{c}}\|_{H^1(\Omega)}.$$

In Lemma 6.7, the assumption that h should be sufficiently small enters by using the L^2-estimate (6.20). Note that the coarse mesh size H remains unconstrained.

Proof. We define $e_H^{\mathrm{c}} := u - u_H^{\mathrm{c}}$. Using Lemma 6.6 (and therefore implicitly Lemma 6.4) we get

$$\frac{\|e_H^{\mathrm{c}}\|_{L^2(\Omega)}^2}{\|e_H^{\mathrm{c}}\|_{H^1(\Omega)}} \lesssim \|\psi_{u-u_H^{\mathrm{c}}} - \psi_H^{\mathrm{c}}\|_{H^1(\Omega)} \le \|\psi_{u-u_H^{\mathrm{c}}} - \psi^h\|_{H^1(\Omega)} + \|\psi_H^{\mathrm{c}} - \psi^h\|_{H^1(\Omega)} \tag{6.22}$$

for all $\psi_H^{\mathrm{c}} \in V_{H,h}^{\mathrm{c}}$ and all $\psi^h \in V_h$. It remains to properly choose ψ^h and ψ_H^{c}. The proof is structured as follows. We choose $\phi_h \in V_h$ to be the fine space approximation of the solution of the adjoint problem (6.18) and ψ_H^{c} is chosen to be the $a(\cdot,\cdot)$-orthogonal approximation of ψ^h. This guarantees that $\psi_H^{\mathrm{c}} - \psi^h$ is in the kernel of our interpolation operator (i.e., $I_H(\psi_H^{\mathrm{c}} - \psi^h) = 0$) and we can estimate the occurring terms while gaining an additional error order of H. The proof is detailed in the following.

Let us choose $\psi^h := \psi^h_{e^{\mathrm{c}}_H}$, where $\psi^h_{e^{\mathrm{c}}_H} \in V_h$ solves

$$c_{\lambda,u}(\psi^h_{e^{\mathrm{c}}_H}, v_h) = 2\beta(|u|^3, \psi^h_{e^{\mathrm{c}}_H})_{L^2(\Omega)} \int_\Omega u v_h \, dx + \int_\Omega (e^{\mathrm{c}}_H - (e^{\mathrm{c}}_H, u)_{L^2(\Omega)}) v_h \, dx$$

for all $v_h \in V_h$. The coercivity of $c_{\lambda,u}$ and reinterpretation of the equation in the sense of problem (6.18) yields that $\psi^h_{e^{\mathrm{c}}_H}$ is well defined. Next, we define

$$g(v, w, u) := -\beta 3|u|^2 v + \lambda v + 2\beta(|u|^3, v)_{L^2(\Omega)} u + (w - (w, u)_{L^2(\Omega)})$$

and solve for $\psi^{H,c}_{e^{\mathrm{c}}_H} \in V^{\mathrm{c}}_{H,h}$ with

$$\int_\Omega A \nabla \psi^{H,c}_{e^{\mathrm{c}}_H} \cdot \nabla v^{\mathrm{c}}_H + b \psi^{H,c}_{e^{\mathrm{c}}_H} v^{\mathrm{c}}_H \, dx = \int_\Omega g(\psi^h_{e^{\mathrm{c}}_H}, e^{\mathrm{c}}_H, u) v^{\mathrm{c}}_H \, dx$$

for all $v^{\mathrm{c}}_H \in V^{\mathrm{c}}_{H,h}$. Since equally $\psi^h_{e^{\mathrm{c}}_H} \in V_h$ fulfills

$$\int_\Omega A \nabla \psi^h_{e^{\mathrm{c}}_H} \cdot \nabla v_h + b \psi^h_{e^{\mathrm{c}}_H} v_h \, dx = \int_\Omega g(\psi^h_{e^{\mathrm{c}}_H}, e^{\mathrm{c}}_H, u) v_h \, dx$$

for all $v_h \in V_h$, we obtain by using the $a(\cdot,\cdot)$-orthogonality of $\psi^h_{e^{\mathrm{c}}_H}$ and $\psi^{H,c}_{e^{\mathrm{c}}_H}$

$$\begin{aligned} a(\psi^h_{e^{\mathrm{c}}_H} - \psi^{H,c}_{e^{\mathrm{c}}_H}, \psi^h_{e^{\mathrm{c}}_H} - \psi^{H,c}_{e^{\mathrm{c}}_H}) &= \int_\Omega g(\psi^h_{e^{\mathrm{c}}_H}, e^{\mathrm{c}}_H, u)(\psi^h_{e^{\mathrm{c}}_H} - \psi^{H,c}_{e^{\mathrm{c}}_H}) \, dx \\ &\le \int_\Omega g(\psi^h_{e^{\mathrm{c}}_H}, e^{\mathrm{c}}_H, u)(\mathrm{Id} - I_H)(\psi^h_{e^{\mathrm{c}}_H} - \psi^{H,c}_{e^{\mathrm{c}}_H}) \, dx \\ &\lesssim (\lambda \|\psi^h_{e^{\mathrm{c}}_H}\|_{H^1(\Omega)} + \|e^{\mathrm{c}}_H\|_{L^2(\Omega)}) H \|\nabla(\psi^h_{e^{\mathrm{c}}_H} - \psi^{H,c}_{e^{\mathrm{c}}_H})\|_{L^2(\Omega)}. \end{aligned}$$

Since

$$\|\psi^h_{e^{\mathrm{c}}_H}\|^2_{H^1(\Omega)} \lesssim c_{\lambda,u}(\psi^h_{e^{\mathrm{c}}_H}, \psi^h_{e^{\mathrm{c}}_H}) = (e^{\mathrm{c}}_H, \psi^h_{e^{\mathrm{c}}_H})_{L^2(\Omega)},$$

we get

$$\|\psi^h_{e^{\mathrm{c}}_H} - \psi^{H,c}_{e^{\mathrm{c}}_H}\|_{H^1(\Omega)} \lesssim H(\|e^{\mathrm{c}}_H\|_{L^2(\Omega)} + \lambda \|\psi^h_{e^{\mathrm{c}}_H}\|_{H^1(\Omega)}) \lesssim (1+\lambda) H \|e^{\mathrm{c}}_H\|_{L^2(\Omega)}.$$

Combining this estimate with (6.22) yields

$$\begin{aligned} \|u - u^{\mathrm{c}}_H\|_{L^2(\Omega)} &\lesssim \left(\frac{\|\psi_{u-u^{\mathrm{c}}_H} - \psi^h_{e^{\mathrm{c}}_H}\|_{H^1(\Omega)}}{\|u - u^{\mathrm{c}}_H\|_{L^2(\Omega)}} + \frac{\|\psi^h_{e^{\mathrm{c}}_H} - \psi^{H,c}_{e^{\mathrm{c}}_H}\|_{H^1(\Omega)}}{\|u - u^{\mathrm{c}}_H\|_{L^2(\Omega)}} \right) \|u - u^{\mathrm{c}}_H\|_{H^1(\Omega)} \\ &\lesssim \left(\frac{\|\psi_{u-u^{\mathrm{c}}_H} - \psi^h_{e^{\mathrm{c}}_H}\|_{H^1(\Omega)}}{\|u - u^{\mathrm{c}}_H\|_{L^2(\Omega)}} + (1+\lambda) H \right) \|u - u^{\mathrm{c}}_H\|_{H^1(\Omega)} \\ &\lesssim \left(\min_{\psi^h \in V_h} \frac{\|\psi_{u-u^{\mathrm{c}}_H} - \psi^h\|_{H^1(\Omega)}}{\|u - u^{\mathrm{c}}_H\|_{L^2(\Omega)}} + H \right) \|u - u^{\mathrm{c}}_H\|_{H^1(\Omega)}. \end{aligned}$$

In the last step we used Céa's lemma for linear elliptic problems and the fact that the H^1-best-approximation in the orthogonal space $V_u^\perp \cap V_h$ can be bounded by the H^1-best-approximation in the full space V_h (cf. [13] and equation (40) therein). ∎

Using (4.1) and Lemma 6.7 we obtain for $e_H^{\mathrm{c}} := u - u_H^{\mathrm{c}}$

$$\|e_H^{\mathrm{c}}\|_{L^2(\Omega)} \lesssim \left(\min_{\psi^h \in V_h} \frac{\|\psi_{u-u_H^{\mathrm{c}}} - \psi^h\|_{H^1(\Omega)}}{\|e_H^{\mathrm{c}}\|_{L^2(\Omega)}} + H \right) \|e_H^{\mathrm{c}}\|_{H^1(\Omega)} \lesssim (|e_h^0| + H)\,(|e_h^1| + H^2),$$

where $|e_h^1| := \min_{v_h \in V_h} \|u - v_h\|_{H^1(\Omega)}$ and $|e_h^0| := \min_{\psi^h \in V_h} \frac{\|\psi_{u-u_H^{\mathrm{c}}} - \psi^h\|_{H^1(\Omega)}}{\|u-u_H^{\mathrm{c}}\|_{L^2(\Omega)}}$. Together with (6.19) this yields

$$\begin{aligned} |\lambda - \lambda_H^{\mathrm{c}}| &\lesssim \|e_H^{\mathrm{c}}\|_{H^1(\Omega)}^2 + \|e_H^{\mathrm{c}}\|_{L^2(\Omega)} \\ &\lesssim (|e_h^1| + H^2)^2 + (|e_h^0| + H)\,(|e_h^1| + H^2) \lesssim H|e_h^1| + H^3. \end{aligned}$$

6.4. Proof of Theorem 4.2. Again, we split the proof into two subsections, one concerning the H^1-error estimate and the other the L^2-error estimate.

6.4.1. Proof of the H^1-error estimate (4.3). Due to the definitions of u_h and u_h^{c} we get for $v_h \in V_h$

$$\begin{aligned} a(u_h - u_h^{\mathrm{c}}, v_h) &= \lambda_h(u_h, v_h) - \lambda_H^{\mathrm{c}}(u_H^{\mathrm{c}}, v_h) - \beta(|u_h|^2 u_h, v_h)_{L^2(\Omega)} + \beta(|u_H^{\mathrm{c}}|^2 u_H^{\mathrm{c}}, v_h)_{L^2(\Omega)} \\ &= \lambda_h(u_h - u_H^{\mathrm{c}}, v_h) + (\lambda_h - \lambda_H^{\mathrm{c}})(u_H^{\mathrm{c}}, v_h) \\ &\quad - \beta \sum_{i=0}^{2} ((u_h)^{2-i} (u_H^{\mathrm{c}})^i (u_h - u_H^{\mathrm{c}}), v_h)_{L^2(\Omega)}. \end{aligned}$$

The treatment of the first and the second term in this error identity is obvious. The last term is treated with the Hölder inequality and the embedding $H_0^1(\Omega) \hookrightarrow L^6(\Omega)$ (for $d \le 3$):

$$\begin{aligned} &\sum_{i=0}^{2} ((u_h)^{2-i} (u_H^{\mathrm{c}})^i (u_h - u_H^{\mathrm{c}}), v_h)_{L^2(\Omega)} \\ &\quad \le \|u_h\|_{L^6(\Omega)}^2 \|u_h - u_H^{\mathrm{c}}\|_{L^2(\Omega)} \|v_h\|_{L^6(\Omega)} \\ &\qquad + \|u_h\|_{L^6(\Omega)} \|u_H^{\mathrm{c}}\|_{L^6(\Omega)} \|u_h - u_H^{\mathrm{c}}\|_{L^2(\Omega)} \|v_h\|_{L^6(\Omega)} \\ &\qquad + \|u_H^{\mathrm{c}}\|_{L^6(\Omega)}^2 \|u_h - u_H^{\mathrm{c}}\|_{L^2(\Omega)} \|v_h\|_{L^6(\Omega)} \\ &\quad \lesssim \|u_h\|_{H^1(\Omega)}^2 \|u_h - u_H^{\mathrm{c}}\|_{L^2(\Omega)} \|v_h\|_{H^1(\Omega)} \\ &\qquad + \|u_h\|_{H^1(\Omega)} \|u_H^{\mathrm{c}}\|_{H^1(\Omega)} \|u_h - u_H^{\mathrm{c}}\|_{L^2(\Omega)} \|v_h\|_{H^1(\Omega)} \\ &\qquad + \|u_H^{\mathrm{c}}\|_{H^1(\Omega)}^2 \|u_h - u_H^{\mathrm{c}}\|_{L^2(\Omega)} \|v_h\|_{H^1(\Omega)}. \end{aligned}$$

We therefore get with $v_h = u_h - u_h^{\mathrm{c}}$ and the Poincaré–Friedrichs inequality

$$\|u_h - u_h^{\mathrm{c}}\|_{H^1(\Omega)} \lesssim (\lambda_h + \lambda_H^{\mathrm{c}}) \|u_h - u_H^{\mathrm{c}}\|_{L^2(\Omega)} + |\lambda_h - \lambda_H^{\mathrm{c}}|.$$

This implies (4.3).

6.4.2. Proof of the L^2-error estimate in (4.4). We start with a lemma that allows us to formulate an error identity.

LEMMA 6.8. *Let $v \in H_0^1(\Omega)$ be an arbitrary function with $\|v\|_{L^2(\Omega)} = 1$ and let $\psi_{u-v} \in V_u^{\perp}$ denote the corresponding solution of the adjoint problem with*

$$c_{\lambda,u}(\psi_{u-v}, w_\perp) = (u - v, w_\perp)_{L^2(\Omega)}$$

for all $w_\perp \in V_u^\perp$ *(cf.* (6.18)*). Then it holds that*

$$\|u-v\|_{L^2(\Omega)}^2 = c_{\lambda,u}(v-u,\psi_{u-v}) + \|u-v\|_{L^2(\Omega)}^2 \int_\Omega |u|^2 u\psi_{u-v}\,dx + \frac{1}{4}\|u-v\|_{L^2(\Omega)}^4.$$

The lemma can be extracted from the proofs given in [13, pp. 99–100].

The following lemma treats the semidiscrete case, i.e., we assume $V_h = H_0^1(\Omega)$. The reason is that the proof of the fully discrete case becomes very technical and hard to read. We note that the proof of the semidiscrete case analogously transfers to the fully discrete case with sufficiently small h by inserting additional continuous approximations to overcome the problems produced by the missing uniform bounds for $\|u_h\|_{L^\infty(\Omega)}$ and $\|u_h^{\mathrm{c}}\|_{L^\infty(\Omega)}$. For the reader's convenience we therefore only prove the case $h = 0$.

LEMMA 6.9 (estimate (4.4) for $h = 0$). *Assume* $h = 0$*, i.e.,* $V_h = H_0^1(\Omega)$*. Accordingly we let* $u_0^{\mathrm{c}} \in H_0^1(\Omega)$ *denote the semidiscrete postprocessed approximation, i.e., the solution to the problem*

$$\int_\Omega A\nabla u_0^{\mathrm{c}}\cdot\nabla\phi\,dx + \int_\Omega b u_0^{\mathrm{c}}\phi\,dx = \lambda_H^{\mathrm{c}}\int_\Omega u_H^{\mathrm{c}}\phi\,dx - \int_\Omega \beta|u_H^{\mathrm{c}}|^2 u_H^{\mathrm{c}}\phi\,dx$$

for all $\phi \in H_0^1(\Omega)$ *(cf. Problem* 3.5*). Then it holds that*

$$\|u-u_0^{\mathrm{c}}\|_{L^2(\Omega)} \lesssim H^4.$$

Proof. We divide the proof into two steps. We want to make use of the error identity in Lemma 6.8 with $v = u_0^{\mathrm{c}}$. However, u_0^{c} is not L^2-normalized and therefore no admissible test function in the error identity. In the first step, we therefore show that the normalization only produces an error of order H^4. In the second step it remains to show that the L^2-error between u and the L^2-normalized u_0^{c} is also of order H^4.

Step 1. We show that $\big|\|u_H^{\mathrm{c}}\|_{L^2(\Omega)} - \|u_0^{\mathrm{c}}\|_{L^2(\Omega)}\big| \lesssim H^4$, which implies $1 - H^4 \lesssim \|u_0^{\mathrm{c}}\|_{L^2(\Omega)} \lesssim 1 + H^4$ (because of $\|u_H^{\mathrm{c}}\|_{L^2(\Omega)} = 1$).

First observe that $u_0^{\mathrm{c}} \in H_0^1(\Omega)$ is the solution to a classical elliptic problem, which is why we obtain

$$\|u_0^{\mathrm{c}}\|_{L^\infty} \lesssim \lambda_H^{\mathrm{c}} \lesssim \lambda. \tag{6.23}$$

Since $a(u_0^{\mathrm{c}} - u_H^{\mathrm{c}}, v_H^{\mathrm{c}}) = 0$ for all $v_H^{\mathrm{c}} \in V_{H,0}^{\mathrm{c}}$ we get $u_0^{\mathrm{c}} - u_H^{\mathrm{c}} \in V_{H,0}^{\mathrm{f}}$. Hence

$$\begin{aligned} a(u_0^{\mathrm{c}} - u_H^{\mathrm{c}}, u_0^{\mathrm{c}} - u_H^{\mathrm{c}}) &= a(u_0^{\mathrm{c}}, u_0^{\mathrm{c}} - u_H^{\mathrm{c}}) \\ &= \lambda_H^{\mathrm{c}}(u_H^{\mathrm{c}}, u_0^{\mathrm{c}} - u_H^{\mathrm{c}}) - \beta(|u_H^{\mathrm{c}}|^2 u_H^{\mathrm{c}}, u_0^{\mathrm{c}} - u_H^{\mathrm{c}}) \\ &= \lambda_H^{\mathrm{c}}(u_H^{\mathrm{c}}, u_0^{\mathrm{c}} - u_H^{\mathrm{c}}) - \beta(|u_H^{\mathrm{c}}|^2 u_H^{\mathrm{c}} - |u|^2 u, u_0^{\mathrm{c}} - u_H^{\mathrm{c}}) \\ &\quad - \beta(|u|^2 u, u_0^{\mathrm{c}} - u_H^{\mathrm{c}}). \end{aligned}$$

Using $u_0^{\mathrm{c}} - u_H^{\mathrm{c}} \in V_{H,0}^{\mathrm{f}}$ and inserting $I_H(u_H^{\mathrm{c}})$ and $I_H(u)$ several times, we get with similar arguments as above and with the previous estimate for $u_H^{\mathrm{c}} - u$

$$\|u_0^{\mathrm{c}} - u_H^{\mathrm{c}}\|_{H^1(\Omega)} \lesssim H^2$$

and

$$\|u_0^{\mathrm{c}} - u_H^{\mathrm{c}}\|_{L^2(\Omega)} = \|(u_0^{\mathrm{c}} - u_H^{\mathrm{c}}) - I_H(u_0^{\mathrm{c}} - u_H^{\mathrm{c}})\|_{L^2(\Omega)} \lesssim H\|u_0^{\mathrm{c}} - u_H^{\mathrm{c}}\|_{H^1(\Omega)} \lesssim H^3. \tag{6.24}$$

Next, we show that $\big|\|u_0^{\mathrm{c}}\|_{L^2(\Omega)} - 1\big|$ is of higher order. We start with

$$\begin{aligned}
\|u_0^{\mathrm{c}}\|_{H^1(\Omega)}^2 - \|u_H^{\mathrm{c}}\|_{H^1(\Omega)}^2 &= a(u_0^{\mathrm{c}}, u_0^{\mathrm{c}}) - a(u_H^{\mathrm{c}}, u_H^{\mathrm{c}}) \\
&= \lambda_H^{\mathrm{c}}(u_H^{\mathrm{c}}, u_0^{\mathrm{c}} - u_H^{\mathrm{c}})_{L^2(\Omega)} - \beta(|u_H^{\mathrm{c}}|^2 u_H^{\mathrm{c}}, (u_0^{\mathrm{c}} - u_H^{\mathrm{c}}))_{L^2(\Omega)} \\
&= \lambda_H^{\mathrm{c}}(u_H^{\mathrm{c}} - I_H(u_H^{\mathrm{c}}), u_0^{\mathrm{c}} - u_H^{\mathrm{c}})_{L^2(\Omega)} \\
&\quad - \beta(|u_H^{\mathrm{c}}|^2 u_H^{\mathrm{c}} - |u_0^{\mathrm{c}}|^2 u_0^{\mathrm{c}}, (u_0^{\mathrm{c}} - u_H^{\mathrm{c}}))_{L^2(\Omega)} - \beta(|u_0^{\mathrm{c}}|^2 u_0^{\mathrm{c}}, (u_0^{\mathrm{c}} - u_H^{\mathrm{c}}))_{L^2(\Omega)} \\
&\overset{(6.24)}{\lesssim} (H^4 + H^6 - \beta(|u_0^{\mathrm{c}}|^2 u_0^{\mathrm{c}}, u_0^{\mathrm{c}} - u_H^{\mathrm{c}})_{L^2(\Omega)}.
\end{aligned}$$

Using that u_0^{c} is bounded uniformly in $L^\infty(\Omega)$ we can proceed as in the proof of Lemma 6.5 to show

$$\beta(|u_0^{\mathrm{c}}|^2 u_0^{\mathrm{c}}, u_0^{\mathrm{c}} - u_H^{\mathrm{c}})_{L^2(\Omega)} \lesssim H\|u_0^{\mathrm{c}}\|_{H^1(\Omega)}\|u_0^{\mathrm{c}} - u_H^{\mathrm{c}}\|_{L^2(\Omega)} \lesssim H^4.$$

So in summary,

$$\big|\|u_0^{\mathrm{c}}\|_{H^1(\Omega)}^2 - \|u_H^{\mathrm{c}}\|_{H^1(\Omega)}^2\big| \lesssim H^4.$$

However, on the other hand,

$$\begin{aligned}
\lambda_H^{\mathrm{c}}\big(\|u_H^{\mathrm{c}}\|_{L^2(\Omega)}^2 - \|u_0^{\mathrm{c}}\|_{L^2(\Omega)}^2\big) &= \lambda_H^{\mathrm{c}}(u_H^{\mathrm{c}} - u_0^{\mathrm{c}}, u_0^{\mathrm{c}} - I_H(u_0^{\mathrm{c}}))_{L^2(\Omega)} \\
&\quad - \beta(|u_H^{\mathrm{c}}|^2 u_H^{\mathrm{c}}, (u_0^{\mathrm{c}} - u_H^{\mathrm{c}}))_{L^2(\Omega)} \\
&\quad - \|u_0^{\mathrm{c}}\|_{H^1(\Omega)}^2 + \|u_H^{\mathrm{c}}\|_{H^1(\Omega)}^2.
\end{aligned}$$

This we can treat with the previous results to get

$$\big|||u_H^{\mathrm{c}}||_{L^2(\Omega)}^2 - ||u_0^{\mathrm{c}}||_{L^2(\Omega)}^2\big| \lesssim H^4.$$

With $||u_H^{\mathrm{c}}||_{L^2(\Omega)} = 1$ we get

$$\big|||u_H^{\mathrm{c}}||_{L^2(\Omega)} - ||u_0^{\mathrm{c}}||_{L^2(\Omega)}\big| \le \big|||u_H^{\mathrm{c}}||_{L^2(\Omega)}^2 - \|u_0^{\mathrm{c}}\|_{L^2(\Omega)}^2\big| \lesssim H^4. \tag{6.25}$$

Note that in the last step we used that for any $a \ge 0$ it holds that $|1 - a| \le |1 - a^2|$.

Step 2. Step 1 justifies the definition of $\tilde{u}_0^{\mathrm{c}} := \|u_0^{\mathrm{c}}\|_{L^2(\Omega)}^{-1} u_0^{\mathrm{c}}$ which fulfills

$$\|\tilde{u}_0^{\mathrm{c}} - u_0^{\mathrm{c}}\|_{L^2(\Omega)} = \big|\|u_0^{\mathrm{c}}\|_{L^2(\Omega)} - 1\big|\ \|u_0^{\mathrm{c}}\|_{L^2(\Omega)} \lesssim H^4. \tag{6.26}$$

Next, we show $\|u - \tilde{u}_0^{\mathrm{c}}\|_{L^2(\Omega)} \lesssim H^4$. For this purpose define $\tilde{\lambda}_H^{\mathrm{c}} := \|u_0^{\mathrm{c}}\|_{L^2(\Omega)}^{-1} \lambda_H^{\mathrm{c}}$. Then $\tilde{u}_0^{\mathrm{c}} \in H_0^1(\Omega)$ solves

$$\int_\Omega A\nabla\tilde{u}_0^{\mathrm{c}} \cdot \nabla\phi\, dx + \int_\Omega b\tilde{u}_0^{\mathrm{c}}\phi\, dx = \tilde{\lambda}_H^{\mathrm{c}} \int_\Omega u_H^{\mathrm{c}}\phi\, dx - \int_\Omega \frac{\beta}{\|u_0^{\mathrm{c}}\|_{L^2(\Omega)}} |u_H^{\mathrm{c}}|^2 u_H^{\mathrm{c}}\phi\, dx.$$

We want to use Lemma 6.8 and denote $\psi := \psi_{u-\tilde{u}_0^c}$ with $\psi_{u-\tilde{u}_0^c} \in V_u^\perp$ being the solution of (6.18) for $w = u - \tilde{u}_0^c$. Before we start to estimate $c_{\lambda,u}(\tilde{u}_0^c - u, \psi)$ observe that $(u,\psi)_{L^2(\Omega)} = 0$ (by definition), which yields

(6.27)
$$\lambda_H^c(u_H^c,\psi)_{L^2(\Omega)} - \lambda(\tilde{u}_0^c,\psi)_{L^2(\Omega)} = (\lambda_H^c - \lambda)(u_H^c - u,\psi)_{L^2(\Omega)} + \lambda(u_H^c - u_0^c,\psi)_{L^2(\Omega)} + \lambda(u_0^c - \tilde{u}_0^c,\psi)_{L^2(\Omega)}.$$

We get

$$\begin{aligned}
&c_{\lambda,u}(\tilde{u}_0^c - u,\psi)\\
&\quad= a(\tilde{u}_0^c - u,\psi) + 3\beta\int_\Omega |u|^2\tilde{u}_0^c\psi\,dx - 3\beta\int_\Omega |u|^2 u\psi\,dx - \lambda(\tilde{u}_0^c,\psi)_{L^2(\Omega)} + \lambda(u,\psi)_{L^2(\Omega)}\\
&\quad= a(\tilde{u}_0^c,\psi) + 3\beta\int_\Omega |u|^2\tilde{u}_0^c\psi\,dx - 2\beta\int_\Omega |u|^2 u\psi\,dx - \lambda(\tilde{u}_0^c,\psi)_{L^2(\Omega)}\\
&\quad= \left(\frac{1-\|u_0^c\|_{L^2(\Omega)}}{\|u_0^c\|_{L^2(\Omega)}} + 1\right)\left(\lambda_H^c\int_\Omega u_H^c\psi\,dx - \beta\int_\Omega |u_H^c|^2 u_H^c\psi\,dx\right)\\
&\qquad+ 3\beta\int_\Omega |u|^2\tilde{u}_0^c\psi\,dx - 2\beta\int_\Omega |u|^2 u\psi\,dx - \lambda(\tilde{u}_0^c,\psi)_{L^2(\Omega)}\\
&\quad\overset{(6.27)}{=} \underbrace{\left(\frac{1-\|u_0^c\|_{L^2(\Omega)}}{\|u_0^c\|_{L^2(\Omega)}}\right)(\lambda_H^c u_H^c - \beta|u_H^c|^2 u_H^c,\psi)_{L^2(\Omega)}}_{=:\mathrm{I}} + \underbrace{(\lambda_H^c - \lambda)(u_H^c - u,\psi)_{L^2(\Omega)}}_{=:\mathrm{II}}\\
&\qquad+ \underbrace{\lambda(u_H^c - u_0^c,\psi - I_H(\psi))_{L^2(\Omega)}}_{=:\mathrm{III}} + \underbrace{\lambda(u_0^c - \tilde{u}_0^c,\psi)_{L^2(\Omega)}}_{=:\mathrm{IV}}\\
&\qquad+ \underbrace{3\beta(|u|^2(\tilde{u}_0^c - u_0^c),\psi)_{L^2(\Omega)}}_{=:\mathrm{V}} + \underbrace{3\beta(|u|^2(u_0^c - u_H^c),\psi)_{L^2(\Omega)}}_{=:\mathrm{VI}}\\
&\qquad\underbrace{-\beta\int_\Omega (u - u_H^c)^2(u_H^c + 2u)\psi\,dx}_{=:\mathrm{VII}}.
\end{aligned}$$

In the last step we used $(u,\psi)_{L^2(\Omega)} = 0$ and

$$a^3 - 3ab^2 + 2b^3 = (a-b)^2(a+2b) \quad \text{for } a,b \in \mathbb{R}.$$

With (6.25) we have

$$|\mathrm{I}| \lesssim \left|\frac{1-\|u_0^c\|_{L^2(\Omega)}}{\|u_0^c\|_{L^2(\Omega)}}\right|(\lambda_H^c + \|u_H^c\|_{H^1(\Omega)}^3)\|\psi\|_{L^2(\Omega)} \lesssim H^4\lambda_H^c(1 + (\lambda_H^c)^2)\|\psi\|_{H^1(\Omega)}.$$

For II we use Theorem 4.1 to obtain

$$|\mathrm{II}| \le |\lambda_H^c - \lambda|\|u_H^c - u\|_{L^2(\Omega)}\|\psi\|_{L^2(\Omega)} \lesssim H^3H^3\|\psi\|_{L^2(\Omega)} \le H^6\|\psi\|_{H^1(\Omega)}.$$

For term III we can use (6.24), which gives us

$$|\mathrm{III}| \le \lambda|(u_H^c - u_0^c,\psi - I_H(\psi))_{L^2(\Omega)}| \lesssim \lambda H^3\|\psi - I_H(\psi)\|_{L^2(\Omega)} \lesssim H^4\|\psi\|_{H^1(\Omega)}.$$

Using (6.26) we get

$$|\mathrm{IV}| \le \lambda \|u_0^{\mathrm{c}} - \tilde{u}_0^{\mathrm{c}}\|_{L^2(\Omega)} \|\psi\|_{L^2(\Omega)} \lesssim H^4 \|\psi\|_{H^1(\Omega)}.$$

Equally we get

$$|\mathrm{V}| \lesssim |(|u|^2(\tilde{u}_0^{\mathrm{c}} - u_0^{\mathrm{c}}), \psi)_{L^2(\Omega)}| \lesssim \|u\|_{L^6(\Omega)}^2 \|\tilde{u}_0^{\mathrm{c}} - u_0^{\mathrm{c}}\|_{L^2(\Omega)} \|\psi\|_{L^6(\Omega)} \lesssim \lambda H^4 \|\psi\|_{H^1(\Omega)}.$$

To estimate VI we need the L^∞-estimate given by (6.21), which reads

$$\|\psi_{u-\tilde{u}_0^{\mathrm{c}}}\|_{L^\infty(\Omega)} \lesssim \|\tilde{u}_0^{\mathrm{c}} - u\|_{L^2(\Omega)}. \tag{6.28}$$

For $z \in \mathcal{N}_H$, let the values u_z and ψ_z denote the coefficients appearing in the weighted Clément interpolation of u and ψ (cf. (3.2)). Recall that Φ_z denote the nodal basis functions of V_H. Using again (6.24), $(\Phi_z, u^{\mathrm{f}})_{L^2(\Omega)} = 0$ for all $z \in \mathcal{N}_H$, and the fact that $u_0^{\mathrm{c}} - u_H^{\mathrm{c}} \in V_{H,0}^{\mathrm{f}}$, we obtain

$$\begin{aligned}
|\mathrm{VI}| &\lesssim |(|u|^2(u_0^{\mathrm{c}} - u_H^{\mathrm{c}}), \psi)_{L^2(\Omega)}| \\
&= \Big| ((u - I_H(u))u\psi, u_0^{\mathrm{c}} - u_H^{\mathrm{c}})_{L^2(\Omega)} + \sum_{z \in \mathcal{N}_H} (u_z(u - u_z)\psi\Phi_z, u_0^{\mathrm{c}} - u_H^{\mathrm{c}})_{L^2(\Omega)} \\
&\quad + \sum_{z \in \mathcal{N}_H} (|u_z|^2(\psi - \psi_z)\Phi_z, u_0^{\mathrm{c}} - u_H^{\mathrm{c}})_{L^2(\Omega)} \Big| \\
&\lesssim \|u\|_{L^\infty(\Omega)} \left(2\|\psi\|_{L^\infty(\Omega)} \|u\|_{H^1(\Omega)} + \|u\|_{L^\infty(\Omega)} \|\psi\|_{H^1(\Omega)}\right) H \|u_0^{\mathrm{c}} - u_H^{\mathrm{c}}\|_{L^2(\Omega)} \\
&\overset{(6.28)}{\lesssim} H^4 \|\tilde{u}_0^{\mathrm{c}} - u\|_{L^2(\Omega)}.
\end{aligned}$$

For the last term Theorem 4.1 leads to

$$|\mathrm{VII}| \lesssim \|u - u_H^{\mathrm{c}}\|_{H^1(\Omega)}^2 \left(\|u_H^{\mathrm{c}}\|_{L^2(\Omega)} + 2\|u\|_{L^2(\Omega)}\right) \|\psi\|_{H^1(\Omega)} \lesssim H^4 \|\psi\|_{H^1(\Omega)}.$$

Combining the results for terms I–VII and using $\|\psi\|_{H^1(\Omega)} \lesssim \|\tilde{u}_0^{\mathrm{c}} - u\|_{L^2(\Omega)}$ we get

$$|c_{\lambda,u}(\tilde{u}_0^{\mathrm{c}} - u, \psi)| \lesssim H^4 \|\tilde{u}_0^{\mathrm{c}} - u\|_{L^2(\Omega)}.$$

Since (by using the previous estimate for $\|u - \tilde{u}_0^{\mathrm{c}}\|_{H^1(\Omega)}$)

$$\frac{1}{4}\|u - \tilde{u}_0^{\mathrm{c}}\|_{L^2(\Omega)}^4 + \|u - \tilde{u}_0^{\mathrm{c}}\|_{L^2(\Omega)}^2 \int_\Omega |u|^2 u \psi_{u-\tilde{u}_0^{\mathrm{c}}}\, dx \le CH^3 \|u - \tilde{u}_0^{\mathrm{c}}\|_{L^2(\Omega)}^2$$

we finally obtain with Lemma 6.8

$$\|u - \tilde{u}_0^{\mathrm{c}}\|_{L^2(\Omega)}^2 \lesssim |c_{\lambda,u}(\tilde{u}_0^{\mathrm{c}} - u, \psi)| \lesssim H^4 \|\tilde{u}_0^{\mathrm{c}} - u\|_{L^2(\Omega)}.$$

With (6.24) we therefore proved

$$\|u - u_0^{\mathrm{c}}\|_{L^2(\Omega)} \lesssim H^4. \qquad \square$$

PROPOSITION 6.10. *The L^2-error estimate in the fully discrete case can be proved analogously to the semidiscrete case above. We therefore get for sufficiently small h that*

$$\|u - u_h^{\mathrm{c}}\|_{L^2(\Omega)} \lesssim H^4 + C_{L^2}(h, H)$$

with $C_{L^2}(h, H)$ behaving like the term $H^2\|u - u_h\|_{H^1(\Omega)}$.

6.4.3. Proof of the eigenvalue error estimate in (4.4). From the following corollary we can conclude estimate (4.4).

COROLLARY 6.11. *Let $u_h^{\mathrm{c}} \in V_h$ denote the solution of the postprocessing step defined via Problem* 3.5 *and let $\lambda_h^{\mathrm{c}} := (2E(u_h^{\mathrm{c}}) + 2^{-1}\beta\|u_h^{\mathrm{c}}\|_{L^4(\Omega)}^4)\|u_h^{\mathrm{c}}\|_{L^2(\Omega)}^{-2}$. Then there holds*

$$|\lambda_h - \lambda_h^{\mathrm{c}}| \lesssim \|u_h - u_h^{\mathrm{c}}\|_{H^1(\Omega)}^2 + \|u_h - u_h^{\mathrm{c}}\|_{L^2(\Omega)}.$$

Proof. We have for arbitrary $v_h \in V_h$

$$\begin{aligned} &a(u_h - v_h, u_h - v_h) + \beta(|u_h|^2(u_h - v_h), u_h - v_h)_{L^2(\Omega)} - \lambda_h(u_h - v_h, u_h - v_h)_{L^2(\Omega)} \\ &\quad = a(v_h, v_h) - \lambda_h(v_h, v_h) + \beta(|u_h|^2 v_h, v_h)_{L^2(\Omega)}. \end{aligned}$$

This implies with $v_h = u_h^{\mathrm{c}}$

$$\begin{aligned} |\lambda_h^{\mathrm{c}} - \lambda_h| &= \left| \frac{a(u_h^{\mathrm{c}}, u_h^{\mathrm{c}}) + \beta(|u_h^{\mathrm{c}}|^2 u_h^{\mathrm{c}}, u_h^{\mathrm{c}})_{L^2(\Omega)} - \lambda_h \|u_h^{\mathrm{c}}\|_{L^2(\Omega)}^2}{\|u_h^{\mathrm{c}}\|_{L^2(\Omega)}^2} \right| \\ &= \left| \frac{\|u_h - u_h^{\mathrm{c}}\|_{H^1(\Omega)}^2 + \beta(|u_h|^2, (u_h - u_h^{\mathrm{c}})^2)_{L^2(\Omega)} - \lambda_h \|u_h - u_h^{\mathrm{c}}\|_{L^2(\Omega)}^2}{\|u_h^{\mathrm{c}}\|_{L^2(\Omega)}^2} \right. \\ &\qquad \left. + \frac{\beta((|u_h|^2 - |u_h^{\mathrm{c}}|^2), |u_h^{\mathrm{c}}|^2)_{L^2(\Omega)}}{\|u_h^{\mathrm{c}}\|_{L^2(\Omega)}^2} \right|. \end{aligned}$$

The remaining estimate is straightforward using $(a^2 - b^2) = (a - b)(a + b)$. Note that the last term is the dominating term. ∎

We obtain (4.4) from Corollary 6.11 and our previous estimates for $\|u - u_h^{\mathrm{c}}\|_{H^1(\Omega)}$ and $\|u - u_h^{\mathrm{c}}\|_{L^2(\Omega)}$.

REFERENCES

[1] A. AFTALION AND I. DANAILA, *Three-dimensional vortex configurations in a rotating bose-einstein condensate*, Phys. Rev. A, 68 (2003), 023603.
[2] A. AFTALION AND I. DANAILA, *Giant vortices in combined harmonic and quartic traps*, Phys. Rev. A, 69 (2004), 033608.
[3] A. AFTALION AND Q. DU, *Vortices in a rotating bose-einstein condensate: Critical angular velocities and energy diagrams in the thomas-fermi regime*, Phys. Rev. A, 64 (2001), 063603.
[4] W. BAO AND Y. CAI, *Optimal error estimates of finite difference methods for the Gross-Pitaevskii equation with angular momentum rotation*, Math. Comp., 82 (2013), pp. 99–128.
[5] W. BAO, I.-L. CHERN, AND F. Y. LIM, *Efficient and spectrally accurate numerical methods for computing ground and first excited states in Bose-Einstein condensates*, J. Comput. Phys., 219 (2006), pp. 836–854.
[6] W. BAO AND Q. DU, *Computing the ground state solution of Bose-Einstein condensates by a normalized gradient flow*, SIAM J. Sci. Comput., 25 (2004), pp. 1674–1697.
[7] W. BAO AND J. SHEN, *A generalized-Laguerre-Hermite pseudospectral method for computing symmetric and central vortex states in Bose-Einstein condensates*, J. Comput. Phys., 227 (2008), pp. 9778–9793.
[8] W. BAO AND W. TANG, *Ground-state solution of Bose-Einstein condensate by directly minimizing the energy functional*, J. Comput. Phys., 187 (2003), pp. 230–254.
[9] W. BAO, H. WANG, AND P. A. MARKOWICH, *Ground, symmetric and central vortex states in rotating Bose-Einstein condensates*, Commun. Math. Sci., 3 (2005), pp. 57–88.
[10] S. BOSE, *Plancks Gesetz und Lichtquantenhypothese*, Z. Phys., 26 (1924), pp. 178–181.
[11] M. CALIARI, A. OSTERMANN, S. RAINER, AND M. THALHAMMER, *A minimisation approach for computing the ground state of Gross-Pitaevskii systems*, J. Comput. Phys., 228 (2009), pp. 349–360.

[12] E. Cancès, *SCF algorithms for Hartree-Fock electronic calculations*, Mathematical Models and Methods for Ab Initio Quantum Chemistry, Lecture Notes in Chemistry 74, Springer, Berlin, 2000.
[13] E. Cancès, R. Chakir, and Y. Maday, *Numerical analysis of nonlinear eigenvalue problems*, J. Sci. Comput., 45 (2010), pp. 90–117.
[14] E. Cancès and C. Le Bris, *Can we outperform the DIIS approach for electronic structure calculations?*, Internat. J. Quantum Chemistry, 79 (2000), pp. 82–90.
[15] C. Carstensen, *Quasi-interpolation and a posteriori error analysis in finite element methods*, M2AN Math. Model. Numer. Anal., 33 (1999), pp. 1187–1202.
[16] M. M. Cerimele, F. Pistella, and S. Succi, *Particle-inspired scheme for the Gross-Pitaevskii equation: An application to Bose-Einstein condensation*, Comput. Phys. Comm., 129 (2000), pp. 82–90.
[17] H. Chen, X. Gong, and A. Zhou, *Numerical approximations of a nonlinear eigenvalue problem and applications to a density functional model*, Math. Methods Appl. Sci., 33 (2010), pp. 1723–1742.
[18] C.-S. Chien, H.-T. Huang, B.-W. Jeng, and Z.-C. Li, *Two-grid discretization schemes for nonlinear Schrödinger equations*, J. Comput. Appl. Math., 214 (2008), pp. 549–571.
[19] F. Dalfovo, S. Giorgini, L. P. Pitaevskii, and S. Stringari, *Theory of Bose-Einstein condensation in trapped gases*, Rev. Mod. Phys., 71 (1999), pp. 463–512.
[20] I. Danaila and P. Kazemi, *A new Sobolev gradient method for direct minimization of the Gross-Pitaevskii energy with rotation*, SIAM J. Sci. Comput., 32 (2010), pp. 2447–2467.
[21] C. M. Dion and E. Cancès, *Ground state of the time-independent Gross-Pitaevskii equation*, Comput. Phys. Comm., 177 (2007), pp. 787–798.
[22] A. Einstein, *Quantentheorie des einatomigen idealen Gases*, Sitzber. Kgl. Preuss. Akad. Wiss., 1924, pp. 261–267.
[23] D. Elfverson, E. H. Georgoulis, A. Målqvist, and D. Peterseim, *Convergence of a discontinuous Galerkin multiscale method*, SIAM J. Numer. Anal., 51 (2013), pp. 3351–3372.
[24] J. J. García-Ripoll and V. M. Pérez-García, *Optimizing Schrödinger functionals using Sobolev gradients: Applications to quantum mechanics and nonlinear optics*, SIAM J. Sci. Comput., 23 (2001), pp. 1316–1334.
[25] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, Springer-Verlag, Berlin, 1983.
[26] E. P. Gross, *Structure of a quantized vortex in boson systems*, Nuovo Cimento, 20 (1960), pp. 454–477.
[27] P. Henning, A. Målqvist, and D. Peterseim, *A localized orthogonal decomposition method for semi-linear elliptic problems*, ESAIM: Math. Model. Numer. Anal., 12 (2013).
[28] P. Henning, P. Morgenstern, and D. Peterseim, *Multiscale Partition of Unity*, Meshfree Methods for Partial Differential Equations VII, Lecture Notes in Computational Science and Engineering, 100 (2014), accepted for publication.
[29] P. Henning and D. Peterseim, *Oversampling for the multiscale finite element method*, Multiscale Model. Simul., 11 (2013), pp. 1149–1175.
[30] W. Ketterle and H.-J. Miesner, *Coherence properties of Bose-Einstein condensates and atom lasers*, Phys. Rev. A, 56 (1997).
[31] E. H. Lieb, R. Seiringer, and J. Yngvason, *Bosons in a trap: A rigorous derivation of the Gross-Pitaevskii energy functional*, Phys. Rev. A, 61 (2000), 043602.
[32] M. Tosi, M. L. Chiofalo, and S. Succi, *Ground state of trapped interacting Bose-Einstein condensates by an explicit imaginary-time algorithm*, Phys. Rev. E, 62 (2000), pp. 7438–7444.
[33] A. Målqvist and D. Peterseim, *Computation of Eigenvalues by Numerical Upscaling*, arXiv: 1212.0090, (2012).
[34] A. Målqvist and D. Peterseim, *Localization of elliptic multiscale problems*, Math. Comput. (2014), DOI: 10.1090/S0025-5718-2014-02868-8.
[35] S. Martellucci, *Bose-Einstein condensates and atom lasers*, in Proceedings of the 27th Course of the International School of Quantum Electronics on Bose-Einstein Condensates and Atom Lasers, Erice, Sicily, Italy, 1999.
[36] H.-J. Miesner, D. Stamper-Kurn, J. Stenger, S. Inouye, A. Chikkatur, and W. Ketterle, *Observation of metastable states in spinor Bose-Einstein condensates*, Phys. Rev. Lett., 82 (1999), pp. 2228–2231.
[37] L. P. Pitaevskii, *Vortex lines in an imperfect bose gas*, Soviet Phys. JETP-USSR, 13 (1961).
[38] L. P. Pitaevskii and S. Stringari, *Bose-Einstein Condensation*, Oxford University Press, Oxford, UK, 2003.

Appendix D
Scale-explicit regularity results and fine-scale discretization

D.1 Finite element network approximation of conductivity in particle composites

Numerische Mathematik **124**(1):73-97, 2013.

(with C. Carstensen)

Numer. Math. (2013) 124:73–97
DOI 10.1007/s00211-012-0509-1

Numerische Mathematik

Finite element network approximation of conductivity in particle composites

Daniel Peterseim · Carsten Carstensen

Received: 29 November 2010 / Revised: 30 August 2012 / Published online: 21 October 2012

Abstract A new finite element method computes conductivity in some unstructured particle-reinforced composite material. The 2-phase material under consideration is composed of a poorly conducting matrix material filled by highly conducting circular inclusions which are randomly dispersed. The mathematical model is a Poisson-type problem with discontinuous coefficients. The discontinuities are huge in contrast and quantity. The proposed method generalizes classical continuous piecewise affine finite elements to special computational meshes which encode the particles in a network structure. Important geometric parameters such as the volume fraction are preserved exactly. The computational complexity of the method is (almost) proportional to the number of inclusions. This is minimal in the sense that the representation of the underlying geometry via the positions and radii of the inclusions is of the same complexity. The discretization error is proportional to the distance of neighboring inclusions and independent of the conductivity contrast in the medium.

Mathematics Subject Classification (2000) 65N15 · 65N30 · 74Q20

The work of D. Peterseim was supported by the DFG Research Center Matheon Berlin through project C33.

D. Peterseim ( ) · C. Carstensen
Institut für Mathematik, Humboldt-Universität zu Berlin,
Unter den Linden 6, 10099 Berlin, Germany
e-mail: peterseim@math.hu-berlin.de

C. Carstensen
Department of CSE, Yonsei University, Seoul, Korea

1 Introduction

Composite materials (or *composites* for short) are engineered materials made from two or more constituents with significantly different physical properties. In a typical configuration, randomly distributed filler particles (inclusions) are surrounded by a second material (matrix) which binds the filler particles together.

The numerical simulation of material properties aims at a better understanding how conductivity depends on controllable variables (e.g., thermal conductivities of the material components, relative volumes, and particles shapes) and hence provides the opportunity to develop materials with enhanced performance for the particular application.

The design of efficient and reliable numerical methods for such problems is challenging. The complexity of the underlying geometry makes classical approaches hardly feasible (cf. Sect. 1.2); in the typical geometric setting, the inclusions are too big for any perturbation analysis or homogenization method, and they are too many or they are packed too densely to resolve them easily with standard finite element meshes. We face this difficulty even for simple continuum models of some material property of interest, e.g. the linear elliptic model problem of heat conduction considered in this paper (see Sect. 1.1).

Based on an efficient treatment of the microscopic geometry, the new method described in this paper (cf. Sect. 1.3) allows reliable numerical simulation of the model problem with many inclusions independent of the degree of disorder in the geometry.

1.1 Model problem

This paper considers a representative 2-dimensional model of a particle-reinforced composite occupying the nonempty open bounded convex polyhedral domain $\Omega \subset \mathbb{R}^2$. Let $\mathcal{B}_{\text{inc}}$ be a set of closed, pairwise disjoint disks of positive radii (inclusions) contained in a domain $\Omega \subset \mathbb{R}^2$, i.e.,

$$B \subset \Omega \quad \text{and} \quad dist\left(B, \tilde{B}\right) > 0 \quad \text{for all } B, \tilde{B} \in \mathcal{B}_{\text{inc}} \text{ with } B \neq \tilde{B}. \tag{1.1}$$

In the present context, the number $N := \#\mathcal{B}_{\text{inc}}$ of inclusions is a very large parameter. The two material phases are represented by the union of the inclusions Ω_{inc}, and by the so called matrix (the perforated domain) Ω_{mat},

$$\Omega_{\text{inc}} := \bigcup_{B \in \mathcal{B}_{\text{inc}}} int(B) \quad \text{and} \quad \Omega_{\text{mat}} := \Omega \backslash \overline{\Omega}_{\text{inc}}.$$

The outer boundary $\Gamma := \partial\Omega$ is partitioned into two parts Γ_D and Γ_N, where Γ_D is closed and has a positive surface measure while its relative complement $\Gamma_N := \Gamma \backslash \Gamma_D$ is relatively open, and the number of contact points $\Gamma_D \cap \overline{\Gamma}_N$ is finite.

The material geometry enters the problem through a coefficient function $c \in L^\infty(\Omega)$ which jumps between the material components. For simplicity c is chosen to

be constant with respect to each of the two phases and normalized with respect to the matrix material, i.e.,

$$c(x) = \begin{cases} 1 & \text{if } x \in \Omega_{\text{mat}}, \\ c_{\text{cont}} & \text{if } x \in \Omega_{\text{inc}}. \end{cases} \tag{1.2a}$$

The constant $c_{\text{cont}} \geq 1$ represents the conductivity contrast in the medium.

Consider the set of admissible temperature distributions

$$\mathcal{A} := u_D + V \text{ with } V := \{u \in H^1(\Omega) \mid u = 0 \text{ on } \Gamma_D\} \tag{1.2b}$$

for $u_D \in H^1(\Omega) \cap C^0(\overline{\Omega})$. Given some force density $f \in V^*$, the effective conductivity of the composite

$$c_{\text{eff}} := \min_{u \in \mathcal{A}} \mathfrak{E}(u) \tag{1.2c}$$

minimizes the energy functional $\mathfrak{E}$,

$$\mathfrak{E}(v) := \frac{1}{2} \int_\Omega c(x) |\nabla v(x)|^2 \, \mathrm{d}x - \int_\Omega f(x) v(x) \, \mathrm{d}x \quad \text{for all } v \in H^1(\Omega). \tag{1.2d}$$

1.2 Challenges to numerical simulations

In practical applications, the parameter $c_{\text{cont}} \gg 1$ is very large. In addition, the coefficient function, which is the output of certain (random) production processes (e.g. mixing of the particles within a liquid matrix material followed by hardening), has to be regarded as a statistical parameter. Corresponding to Berlyand [3], the latter two issues, random micro-structures on multiple scales and high contrast in physical properties, are the two characteristic features of general composites. They lead to major difficulties for a numerical approximation of problem (1.2).

Classical FEM A classical method for the approximate solution of (1.2) is the finite element method. However, in the present context, standard finite element approaches suffer from the fact that the material interface $\partial\Omega_{\text{inc}}$ needs to be resolved by the underlying mesh in order to get satisfactory results. The required resolution of the coefficient geometry forces even the coarsest available meshes to be very fine, i.e., the minimal mesh size has to be at most of order of the inclusion radii. Additionally, finite element methods often require high quality meshes (shape regularity) which puts even more constraints on mesh generation. Thus, the minimal number of nodes in a reasonable mesh depends critically on the distribution of the holes and their distances; Fig. 1 illustrates the problem in a model situation, which is eased for visualization purposes.

Minimal complexity Since the underlying geometry is of stochastic nature problem (1.2), typically, needs to be solved many times for different coefficient configurations

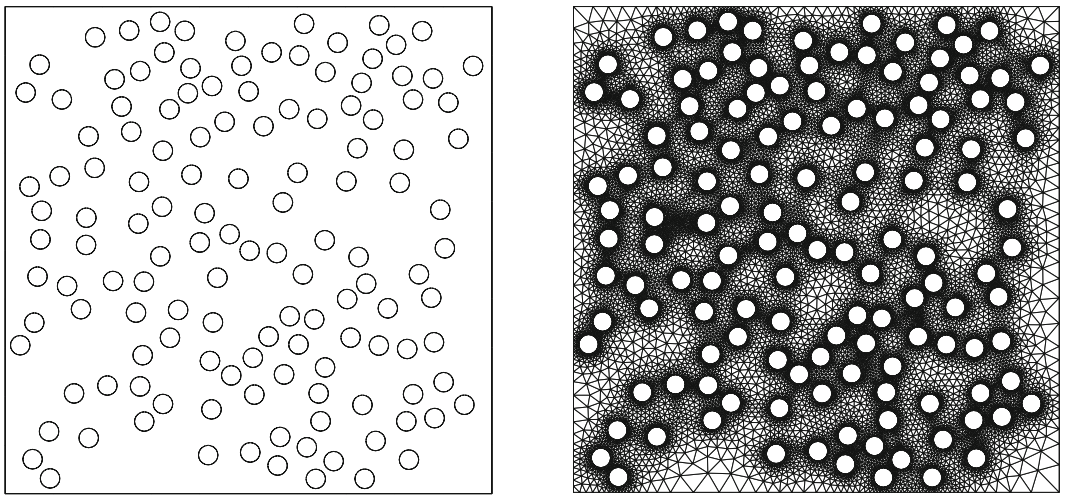


Fig. 1 Model domain (unit square) containing 133 circular inclusions with radius $r = 0.02$ (*left*) and *"coarse"* shape regular triangulation with 33903 elements (*right*)

within a statistical investigation of material properties (by a Monte Carlo method). For example, the accuracy of the approximation of the expected temperature distribution subject to a to random distribution of particles in the material, is of order $M^{-1/2}$, where M denotes the number of samples. Since the coefficient is different for different samples, meshes cannot be re-used but need to be re-computed for every single sample of the particle distribution. Hence, the computation of the finite element mesh is crucial in all complexity discussions and cannot be neglected as a precomputation (cf. Fig. 1). With regard to the possibly huge number of instances of problem (1.2) that need to be considered, this paper aims at a reasonable discrete model of minimal complexity. Minimality is determined by the data of the problem and therefore mainly by its geometry. The geometry representation requires storing the pairs of centers and radii of the N inclusions (the complexity of the representation of the outer boundary is supposed to be small compared to N). A model is considered to have minimal complexity if it provides an approximate solution in time and space complexity $\mathcal{O}(N)$. The finite element method to be presented in this paper satisfies the complexity requirement up to logarithmic factors (cf. Section below).

1.3 The new structural finite element approach

In this paper ideas from network approximations [3–5,19] are combined with nonstandard finite element methods to derive a new structural finite element method of almost minimal complexity. In particular, a special geometry treatment inspired by networks is combined with the flexibility of finite element methods. As in discrete network methods, the inclusions are modeled in a network structure. They appear as elements of the computational mesh, supplemented by channel-like objects that connect neighboring inclusions and, finally, triangles. The mesh generalizes standard Delaunay triangulations of points in the plane to sets of disks. It can be computed and represented efficiently. A generalization of continuous first-order finite ele-

ments based on the new, problem-adapted subdivisions is introduced. Its realization is conceivably simple and it provides accurate numerical approximations at almost minimal complexity. More precisely, for the solution $u \in \mathcal{A} \cap H^2(\Omega_{\text{mat}} \cup \Omega_{\text{inc}})$ of (1.2) and its structural finite element approximation u_S it holds (see Theorem 3.1, Corollaries 3.1 and 3.2).

$$\|\sqrt{c}\nabla(u - u_S)\|_{L^2(\Omega)} \leq C_{f,u_D,\mathcal{B}_{\text{inc}}} \|h\|_{L^\infty(\Omega)},$$

where h is a local mesh size parameter. The constant does *not* depend on contrast. Its dependencies on the geometry of the material (e.g., touching inclusions) are discussed in detail.

The overall motivation for the novel network approximation is its optimal complexity in the sense that the cost for a meaningful approximation remains proportional to the number of inclusions.

1.4 Numerical upscaling

The number of degrees of freedom might be reduced further by using multiscale methods, e.g., [7,11,16,17,20,21]. These methods are based on arbitrary coarse meshes that, more or less, ignore the geometric scales of the coefficient. The influence of the coefficient is instead coded in the finite element basis functions or some modified discrete operator. For this, multiscale methods require some preprocessing that involves the solution of the original problem on subdomains. The solution of these local problems, however, faces the same difficulties as the original problem, i.e., it requires submeshes fine enough to capture the heterogeneities (the influence of the microscopic geometry on macroscopic material properties can only be studied if the microscopic geometry enters the discretization). In this regard, the method presented here might be employed as an efficient fine scale solver within some multiscale numerical framework.

1.5 Outline

Section 2 defines a problem adapted generalization of triangular meshes modeling the inclusions as (vertex-like) elements of a subdivision. Based on this new type of meshes a generalized nodal basis defining a generalized conforming first-order approximation space is introduced. Contrast-independent a priori error estimates for the proposed new finite element method are given in Sect. 3. Section 4 discusses open problems and future generalizations of the method.

1.6 Notation

In this paper, capital letters $A, B, C, \ldots$ indicate sets. Calligraphic capital letters $\mathcal{B}, \mathcal{P}, \ldots$ denote sets of sets. For a given set of sets $\mathcal{B}$ the union of its elements is denoted by $\cup\mathcal{B} := \bigcup_{B\in\mathcal{B}} B$. Basic topological notations are used: For any subset X

of a metric space its closure is denoted by $\overline{X}$, its interior by $int(X)$, and its boundary by $bnd(X)$. In what follows, $dist(\cdot,\cdot)$ denotes the Euclidean distance in $\mathbb{R}^2$. The measure $|\cdot|$ is context-sensitive and refers to the volume of a set relative to its dimension, i.e., $|\cdot|$ denotes the length of a curve, or the area of a domain. The distance between nonempty subsets $A, B \subset \mathbb{R}^2$ reads

$$dist(A,B) := \inf_{x\in A, y\in B} dist(x,y)\,. \tag{1.3}$$

Given some bounded domain Ω, standard notation for (fractional) Sobolev spaces $W^m_p(\Omega)$, $m \geq 0$, $p \in \mathbb{N}\cup\{0\}$, and their corresponding norms $\|\cdot\|_{W^m_p(\Omega)}$ and seminorms $|\cdot|_{W^m_p(\Omega)}$ is used; $H^m(\Omega)$ abbreviates $W^m_2(\Omega)$ ($m \in \mathbb{N}$) and $L^p(\Omega)$ abbreviates $W^0_p(\Omega)$. Given two disjoint bounded Lipschitz domains Ω_1 and Ω_2, the space $H^m(\Omega_1\cup\Omega_2)$ denotes the space of all functions $u \in L^2(\Omega_1\cup\Omega_2)$ with $u|_{\Omega_1} \in H^m(\Omega_1)$ and $u|_{\Omega_2} \in H^m(\Omega_2)$. The dual space of a Hilbert space V is indicated by V^*. The space of $\mathbb{R}$-valued continuous functions on a set Ω is denoted by $C^0(\Omega)$.

2 A minimal conforming finite element space

This section introduces a conforming finite element space which can be regarded as a generalization of the classical continuous piecewise affine finite element space on a special mesh.

2.1 Geometric preliminaries

Cyclic polygons A convex polygon T is the closed convex hull of 2 or more distinct points. The set of vertices (corners) $\mathcal{V}(T)$ is the minimal set of points $x_1, x_2, \ldots, x_k \in \mathbb{R}^2$, such that $T = \mathrm{conv}(\{x_1, x_2, \ldots, x_k\})$. According to the number of its vertices, a convex polygon is denoted as a convex k-gon. The boundary of a convex k-gon can be described by the union of at most k line segments called *edges*. The set of edges of a convex polygon T is denoted by $\mathcal{E}(T)$. A convex polygon T is called *cyclic* if its vertices (corners) $V(T)$ are located on the boundary of a (closed) disk $CD = CD(T)$ which is denoted as the *circumdisk* of T. Examples of cyclic polygons are line segment, triangles, or rectangles.

Infinite Delaunay Triangulations A regular (possibly infinite) triangulation of a domain $\Omega \subset \mathbb{R}^2$ into cyclic polygons is a set of cyclic polygons $\mathcal{T}$ such that

$$\cup\mathcal{T} = \overline{\Omega}$$

and any two distinct cyclic polygons are either

a) disjoint, $T_1 \cap T_2 = \emptyset$, or
b) share exactly one vertex z, $T_1 \cap T_2 = V(T_1) \cap V(T_2) = \{z\}$, or
c) have one edge $E = bnd(T_1) \cap bnd(T_2) = \mathcal{E}(T_1) \cap \mathcal{E}(T_2)$ in common.

The set of all edges resp. vertices of a triangulation $\mathcal{T}$ is written as

$$\mathcal{E}(\mathcal{T}) := \bigcup_{T\in\mathcal{T}} \mathcal{E}(T) \quad \text{resp.} \quad \mathcal{V}(\mathcal{T}) := \bigcup_{T\in\mathcal{T}} \mathcal{V}(T).$$

A regular triangulation $\mathcal{T}$ is called *Delaunay* [10] if every element $T \in \mathcal{T}$ satisfies the Delaunay criterion

$$CD(T) \cap \mathcal{V}(\mathcal{T}) = \mathcal{V}(T), \tag{2.1}$$

that is, the circumdisc of T does not contain any vertices of $\mathcal{T}$ except those of T. Given a set of vertices $\mathcal{V}$, the Delaunay triangulation of $\mathrm{conv}(\mathcal{V})$ is uniquely determined (if cyclic polygons are considered). It can be constructed, e.g., by exploiting duality with respect to the Voronoi diagram [27] of $\mathcal{V}$. The uniqueness is due to the consideration of cyclic polygons instead of just triangles. In the subsequent paragraph, cyclic k-gons with $k > 3$ will further be decomposed into triangles.

2.2 Geometric modeling of particle composites

The geometry of model problem (1.2) is represented by a finite set $\mathcal{B}$ of closed disks. Every $B \in \mathcal{B}$ is described by its center $c_B = \mathrm{mid}(B)$ and its radius $r_B = \mathrm{diam}\big(B\big)/2 \geq 0$. The elements of $\mathcal{B}$ are denoted as *generalized vertices* and partitioned into the two subsets $\mathcal{B}_{\mathrm{inc}}$ and $\mathcal{B}_{\mathrm{mat}}$, i.e.,

$$\mathcal{B} = \mathcal{B}_{\mathrm{inc}} \cup \mathcal{B}_{\mathrm{mat}} \quad \text{and} \quad \mathcal{B}_{\mathrm{inc}} \cap \mathcal{B}_{\mathrm{mat}} = \emptyset.$$

The set $\mathcal{B}_{\mathrm{inc}}$ contains the inclusions of model problem (1.2), i.e., closed disks of positive radius. The set $\mathcal{B}_{\mathrm{mat}}$ contains closed disks of radius zero with

$$\mathrm{conv}(\cup\mathcal{B}_{\mathrm{mat}}) = \overline{\Omega} \quad \text{and} \quad \Gamma_D \cap \overline{\Gamma}_N \subset \cup\mathcal{B}_{\mathrm{mat}}.$$

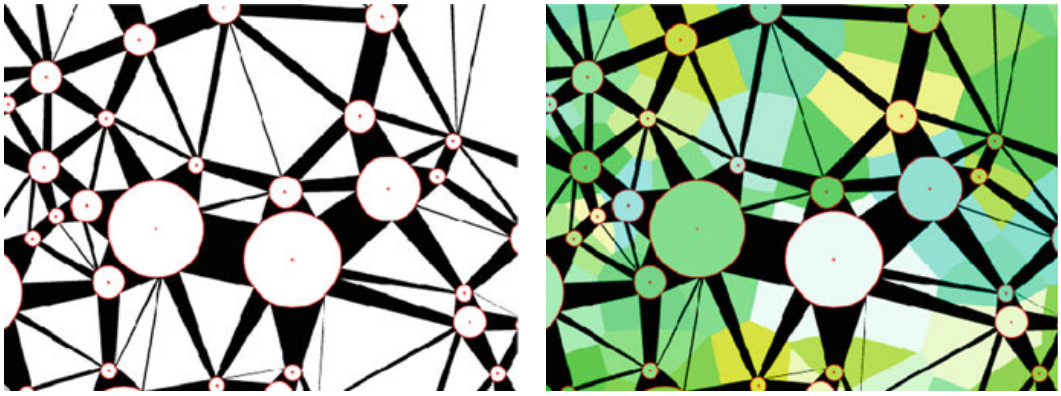
Thus $\mathcal{B}_{\mathrm{mat}}$ contains the corners of $\partial\Omega$ and all points where the type of boundary condition switches between Dirichlet and Neumann; but $\mathcal{B}_{\mathrm{mat}}$ might contain additional points (disks with vanishing radii) in the interior of the matrix Ω_{inc}, which offers the possibility of refinement and increased local resolution within the finite element framework.

By $\mathcal{T}_{\mathrm{mat}}$ we denote the Delaunay triangulation of Ω_{mat} such that

$$\mathcal{V}(\mathcal{T}_{\mathrm{mat}}) = \mathcal{B}_{\mathrm{mat}} \cup \bigcup_{B\in\mathcal{B}_{\mathrm{inc}}} \partial B.$$

Figure 2 displays a detail of $\mathcal{T}_{\mathrm{mat}}$ for some set of disks $\mathcal{B}$. Obviously, $\mathcal{T}_{\mathrm{mat}}$ consists of two classes of cyclic polygons (see [24]), namely,

a) (possibly infinitely many) cyclic 2-gons $\mathcal{T}_{\mathrm{mat}}^{|}$, i.e., line segments whose vertices are located on the circumference of exactly two distinct disks, and



(a) Generalized Delaunay triangulation

(b) Genrealized Delaunay triangulation and dual Voronoi tessellation

Fig. 2 Generalized Delaunay triangulation with respect to disks in the plane

b) (finitely many) cyclic k-gons $\mathcal{T}_{\mathrm{mat}}^{\Delta}$ for $k \geq 3$.

For simplicity we assume that $\mathcal{T}_{\mathrm{mat}}^{\Delta}$ contains exclusively triangles. This assumption can always be fulfilled if we consider a triangulation $\tilde{\mathcal{T}}_{\mathrm{mat}}$ in which the $4, 5, \ldots$ -gons of $\mathcal{T}_{\mathrm{mat}}$ are further decomposed into triangles; $\tilde{\mathcal{T}}_{\mathrm{mat}}$ is not Delaunay in the sense of (2.1) but fulfills the weaker Delaunay criterion

$$int(CD(T)) \cap \mathcal{V}(\tilde{\mathcal{T}}_{\mathrm{mat}}) = \emptyset \quad \text{for all } T \in \tilde{\mathcal{T}}_{\mathrm{mat}}, \tag{2.2}$$

that is, there are no vertices of $\tilde{\mathcal{T}}_{\mathrm{mat}}$ in the interior of the circumdisk of $T \in \tilde{\mathcal{T}}_{\mathrm{mat}}$. The subset $\mathcal{T}_{\mathrm{mat}}^{\Delta}$ of triangles of $\mathcal{T}_{\mathrm{mat}}$ provides structural (combinatorial) information about the set of inclusions $\mathcal{B}_{\mathrm{inc}}$. It induces a neighborhood relation $\mathcal{N} \subset \mathcal{B}_{\mathrm{inc}} \times \mathcal{B}_{\mathrm{inc}}$ defined by the rule: $(B_1, B_2) \in \mathcal{N}$ *if there exists a* $T \in \mathcal{T}_{\mathrm{mat}}^{\Delta}$ *such that* $V(T) \subset B_1 \cup B_2$ *and* $V(T) \cap B_1 \neq \emptyset$ *and* $V(T) \cap B_2 \neq \emptyset$. For every pair $(B_1, B_2) \in \mathcal{N}$ of neighboring disks we define the channel-like object (a bundle of line segments)

$$E(B_1, B_2) := \cup\{T \in \mathcal{T}_{\mathrm{mat}} \ : \ V(T) \subset B_1 \cup B_2\}.$$

Since $E(B_1, B_2)$ is an object that connects exactly two generalized vertices (disks) we denote $E(B_1, B_2)$ a generalized edge.

A finite subdivision $\mathcal{G}$ of Ω, which will serve as the finite element mesh later, is given by

$$\mathcal{G} = \mathcal{B}_{\mathrm{inc}} \cup \mathcal{E} \cup \mathcal{T},$$

where $\mathcal{B}_{\mathrm{inc}}$ is the given set of disks, $\mathcal{E} := \{E(B_1, B_2) : \ (B_1, B_2) \in \mathcal{N}\}$ is the set of generalized edges and $\mathcal{T} := \mathcal{T}_{\mathrm{mat}}^{\Delta}$ is the set of triangles.

Remark 1 a) The subdivision $\mathcal{G}$ can be regarded as a generalization of classical Delaunay triangulations in the sense that disks might assume the classical role of vertices while edges (i.e., objects that connect two neighboring vertices) might

generalize to channels. In the special case of equally sized inclusions such subdivisions have been used in discrete network approximations [3]. Apart from minor technical details regarding the treatment of element boundaries, the subdivision $\mathcal{G}$ fits into the framework of generalized Delaunay partitions for multidimensional sets of convex inclusions introduced in [23].

b) The subdivision $\mathcal{G}$ covers Ω while the intersection of any two of its elements is of measure zero.
c) The number of elements in $\mathcal{G}$ is proportional to the cardinality of $\mathcal{B}$ and thus is quasi minimal.
d) There is a duality concept which links generalized Delaunay triangulations to Voronoi tessellations with respect to the set of disks (see Fig. 2(b) and the next subsection). It generalizes straight-line duality between classical Voronoi tessellation and Delaunay triangulation of point sets. We refer to [23] for more insights about geometric duality and further references.
e) The generalized Delaunay triangulation $\mathcal{D}$ can be computed fast as explained subsequently. There exist algorithms of order $\mathcal{O}(\#\mathcal{B}\times\log(\#\mathcal{B}))$ for the computation of Voronoi diagrams with respect to a set of disks $\mathcal{B}$; see, e.g., [13,14,18]. These algorithms, by duality, can also be employed for the computation of the generalized Delaunay subdivision.

We refer to the recent preprint [12] for an algorithmic presentation of this construction.

2.3 Element parametrization and local mesh size

The generalized vertices $\mathcal{B}_{\text{inc}}$ and the triangles $\mathcal{T}$ form affine families and can easily be represented by reference elements and affine mappings.

A parametrization of a generalized edge can be given as follows. Let $E = E(B_1, B_2)$ in $\mathcal{E}$ be a generalized edge that connects two generalized vertices $B_1, B_2 \in \mathcal{B}$ and let

$$\Sigma_E := \left\{ y \in \mathbb{R}^2 : \ dist(y, B_1) = dist(y, B_2) \ \text{ and } dist(y, B_1) \leq dist(y, \mathcal{B}\backslash\{B_1, B_2\}) \right\}$$

denote the corresponding dual Voronoi edge, the set of points with equal distance to both B_1 and B_2. Without loss of generality we assume $r_{B_1} \geq r_{B_2}$, $c_{B_1} = (0, 0)$, $c_{B_2} = (0, \delta)$, $\delta > 0$. Note that the Voronoi dual edge might not be connected (see Fig. 4a). The same applies to the generalized edge as it can be seen in Fig. 4b. We denote the number of connected components of E by $K(E)$. The projection $\pi_{B_1} := \text{argmin}_{y\in B_1} dist(\cdot, y)$ defines angles

$$-\tfrac{\pi}{2} \leq \alpha_E^1 \leq \beta_E^1 < \alpha_E^2 \leq \beta_E^2 < \cdots < \alpha_E^{K(E)} \leq \beta_{ij}^{K(E)} \leq \frac{\pi}{2}$$

such that

$$\pi_{B_1}(\Sigma_E) = \bigcup_{k=1}^{K(E)} r_{B_1} \left[sin([\alpha_E^k, \beta_E^k]), cos([\alpha_E^k, \beta_E^k]) \right]^T .$$

In other words, the parameters $\alpha_E^1, \ldots, \alpha_E^{K(E)}, \beta_E^1, \ldots, \beta_E^{K(E)}$ are the angular values of the projections of the Voronoi vertices which are connected by Σ_E, onto B_1. Those Voronoi vertices are simply the circumcenters of triangles adjacent to E. With the reference element

$$E^{\mathrm{ref}} = E^{\mathrm{ref}}(B_1, B_2) := \left(]\alpha_E^1, \beta_E^1[\cup \cdots \cup]\alpha_E^{K(E)}, \beta_E^{K(E)}[\right) \times]0, 1[, \tag{2.3}$$

the mapping $J_E : E^{\mathrm{ref}} \to \mathrm{int}E$, given by

$$\begin{aligned} J_E(s, \lambda) &= (1-\lambda) r_{B_1} \begin{pmatrix} \sin(s) \\ \cos(s) \end{pmatrix} + \lambda \pi_{B_2} \left((\pi_{B_1}|_{\Sigma_E})^{-1} \left(r_{B_1} \begin{pmatrix} \sin(s) \\ \cos(s) \end{pmatrix} \right) \right) \\ &= \begin{pmatrix} ((1-\lambda) r_{B_1} + \lambda r_{B_2}) \sin(s) \\ ((1-\lambda) r_{B_1} - \lambda r_{B_2}) \cos(s) + \delta\lambda \end{pmatrix}, \end{aligned}$$

parametrizes E. Figure 3a visualizes the mapping J_E. Note that a generalized edge $E(B_1, B_2)$ is uniquely determined by the inclusion centers and radii, and the values of α_E, β_E, and δ.

The projection $\pi_{B_1,B_2}(\cdot) := \pi_{B_2}(\pi_{B_1}^{-1}(\cdot))$ may be rewritten as

$$\pi_{B_1,B_2}(x) := \operatorname*{argmin}_{y \in \partial B_2} \frac{dist(x, y)}{\max\{\langle (y-x)/\|y-x\|, \nu_{B_1}(y) \rangle, 0\}}, \tag{2.4}$$

where ν_{B_1} denotes the outer normal of B_1.

With

$$H(s) := \frac{\left(\delta^2 - 2\,\cos(s)\;\delta\, r_{B_1}\right) + r_{B_1}^2 - r_{B_2}^2}{\left(2\, r_{B_2} - 2\, r_{B_1}\right) + 2\,\delta\,\cos(s)}.$$

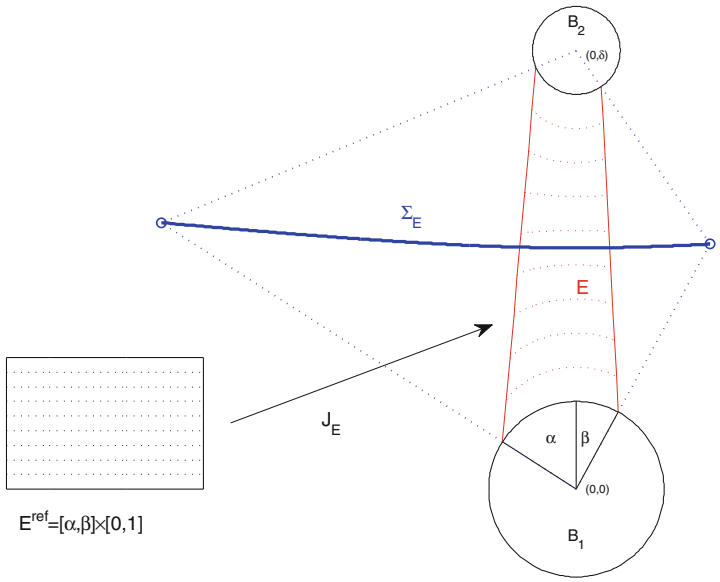
the parametrization J_E assumes the form

$$\begin{aligned} J_E(s, \lambda) = &\left((1-\lambda) r_1 + \lambda r_2 \frac{r_{B_1} + H(s)}{r_{B_2} + H(s)} \right) \begin{bmatrix} \sin(s) \\ \cos(s) \end{bmatrix} \\ &+ \delta\lambda \left(1 - \frac{r_{B_2}}{r_{B_2} + H(s)} \right) \begin{pmatrix} 0 \\ \delta \end{pmatrix}. \end{aligned} \tag{2.5}$$

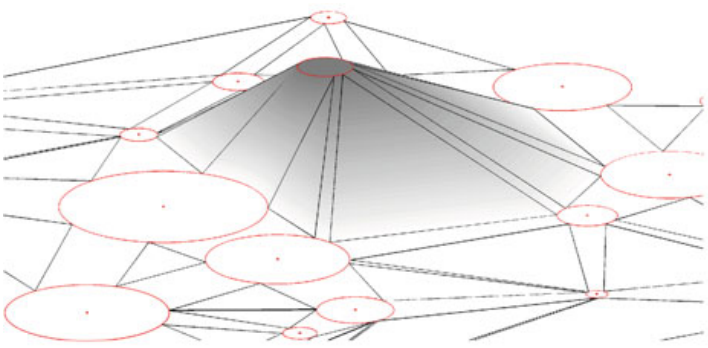
We finally introduce some $(\mathcal{T}_{\mathrm{mat}} \cup \mathcal{B})$-piecewise constant meshsize function $h : \Omega \to]0, \infty[$ by

$$h|_K = h_K := \mathrm{diam}(K) \quad \text{for } K \in \mathcal{T}_{\mathrm{mat}} \cup \mathcal{B}$$

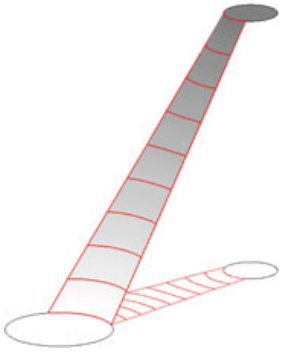
to be used in the forthcoming finite element analysis. Note that h is not constant with respect to a generalized edge (of positive measure) but captures the distance between neighboring inclusions (see (2.4)).



(a) Parametrization (isolines) of a generalized edge E that connects two generalized vertices $B_1 = \{x \in \mathbb{R}^2 | \ \|x\| \leq r_{B_1}\|$ and $B_2 = \{x \in \mathbb{R}^2 | \ \|x - (0, \delta)\| \leq r_{B_2}\|$; $\delta = 1$, $r_{B_1} = 0.2$, $r_{B_2} = 0.1$, $\alpha_E = -1$, $\beta_E = 0.5$, $K(E) = 1$



(b) A generalized nodal basis function taking value 1 on one node and zero on all the others



(c) Nodal basis function restricted to a generalized edge

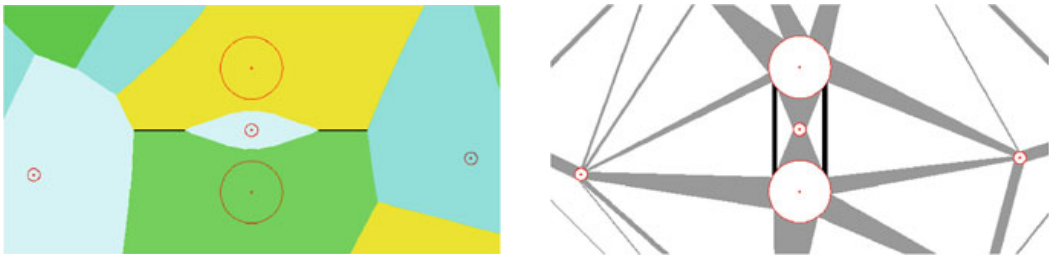
Fig. 3 Edge parametrization and nodal basis function

2.4 Finite element spaces

The degrees of freedom of the finite element spaces are assigned to the entries of $\mathcal{B}$. Every $B \in \mathcal{B}$ defines a (local) $\mathcal{T}_{\text{mat}}$-affine basis function $\lambda_B : \mathbb{R}^2 \to [0, 1]$ with

$$\lambda_B \equiv 1 \quad \text{in } B \quad \text{while } \lambda_B \equiv 0 \quad \text{in } \Omega_{\text{inc}} \setminus B.$$

More precisely, λ_B is unique continuous function with constant values on the inclusions as above and whose restriction to each element $T \in \mathcal{T}_{\text{mat}}$ is affine. This means that λ_B is affine on all triangles $T \in \mathcal{T}$. However, λ_B is not affine on generalized edges. Recall



(a) Part of a Voronoi tessellation with a multiply connected Voronoi edge (black)

(b) Part of the generalized Delaunay triangulation with a multiply connected generalized edge (black shaded)

Fig. 4 Voronoi tessellation and Generalized Delaunay triangulation of a set of disks in the plane emphasizing possible non-connectivity of its elements

that a generalized edge $E \in \mathcal{E}$ is the agglomeration of line segments. The restriction of λ_B to all those line segments is supposed to be affine. On the global level of the generalized edge E, this implies that $\lambda_B|_E$ is the image of an affine function on the rectangular reference element E^{ref} (cf. (2.3)) under the coordinate transformation J_E (cf. (2.5)). After suitable rotation of the edge as in Sect. 2.3 (with $B_1 = B$), $\lambda_B|_E$ may be written as

$$\lambda_B(x) = (1 - (J_E^{-1}(x))_2) \quad \text{for all } x \in E,$$

where $(J_E^{-1}(x))_2$ refers to the second component of the vector $J_E^{-1}(x)$.

Those basis functions generalize nodal basis functions on classical triangular meshes. In the special case of equally sized inclusions, those basis function have been used in the analysis of a network method [4]. The support of λ_B, denoted by ω_B, is given by

$$\omega_B := B \ \cup \ (\cup\{E \in \mathcal{E} : \ E \cap B \neq \emptyset\}) \ \cup \ (\cup\{T \in \mathcal{T} : \ T \cap B \neq \emptyset\}) \, .$$

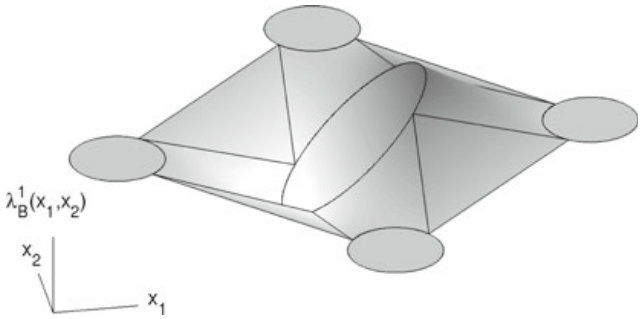
Figure 3b depicts a nodal basis function. Note that the set of nodal basis functions $\Lambda := \{\lambda_B : \ B \in \mathcal{B}\}$ forms a partition of unity in Ω. The generalized nodal basis functions which are not related to vertices on the Dirichlet boundary Γ_D span the finite element space

$$S^\infty := span(\Lambda) \cap V. \tag{2.6}$$

Obviously S^∞ has dimension $\#\mathcal{B}$ which is minimal in comparison to data complexity and will be the space of choice for very large contrast and the special case of perfectly conducting inclusions $c_{\mathrm{cont}} = \infty$. In the latter case the solution is necessarily constant with respect to every single inclusion (see Sect. 3.1), which is captured by S^∞.

If $c_{\mathrm{cont}} < \infty$ then the solution is not constant on the inclusions. Further basis functions (defined below) shall preserve sufficiently high accuracy in this setting, too.

Fig. 5 Basis function λ_B^1



Every $B \in \mathcal{B}_{\text{inc}}$ defines (local) $\mathcal{T}_{\text{mat}}$-affine basis functions $\lambda_B^1, \lambda_B^2 : \mathbb{R}^2 \to [0, 1]$ with

$$\lambda_B^k(x) = \frac{x_k - (c_B)_k}{r_B} \quad \text{if } x \in B \text{ while } \lambda_B^k \equiv 0 \text{ in } \Omega_{\text{inc}} \backslash B.$$

The subscript k refers to the kth component of a 2-dimensional vector.

This means that λ_B^k is affine on all inclusions $B \in \mathcal{B}_{\text{inc}}$ and all triangles $T \in \mathcal{T}$. After suitable rotation of the edge as in Sect. 2.3 (with $B_1 = B$), $\lambda_B|_E$ may be written as

$$\lambda_B^k(x) = (1 - (J_E^{-1}(x))_2)\lambda_B^k(J_E((J_E^{-1}(x))_1, 0)) \quad \text{for all } x \in E,$$

where the coordinate transformation J_E is given in (2.5). Note that $J_E((J_E^{-1}(\cdot))_1, 0)) \in \partial B$ and, hence, the values $\lambda_B^k(J_E((J_E^{-1}(\cdot))_1, 0))$ are given by (2.4). It holds $supp\left(\lambda_B^k\right) = \omega_B$. Figure 5 illustrates λ_B^1.

The enlarged finite element space is then given by

$$S := span\Big(\Lambda \cup \{\lambda_B^1 : \ B \in \mathcal{B}\} \cup \{\lambda_B^2 : \ B \in \mathcal{B}_{\text{inc}}\}\Big) \cap V. \tag{2.7}$$

Remark 2 a) If the radii of all inclusions are zero, the spaces S resp. S^∞ reduce to the classical conforming $\mathbb{P}^1$ finite element space with respect to the Delaunay triangulation.

b) The number of degrees of freedom in S is 3 per inclusion $B \in \mathcal{B}_{\text{inc}}$, and 1 per any other vertex $B \in \mathcal{B}_{\text{mat}}$ away from Γ_D. The overall number of degrees of freedom is bounded by $3\#\mathcal{B}_{\text{inc}} + \#\mathcal{B}_{\text{mat}}$ and, hence, proportional to data complexity.

c) Further basis functions could easily be designed by considering any continuous function on B and its $\mathcal{T}_{\text{mat}}$-affine or a more general $\mathcal{T}_{\text{mat}}$-polynomial extension to ω_B.

3 Galerkin approximation and a priori error analysis

This section considers the variational formulation of (1.2) and its Galerkin approximation and presents error estimates which are independent of the contrast parameter c_{cont}.

3.1 Variational formulation and solvability

Any minimizer $u^* \in \mathcal{A}$ of (1.2) solves the variational problem

$$\int_\Omega c\langle \nabla u^*, \nabla v\rangle \mathrm{d}x = \int_\Omega f v \mathrm{d}x \quad \text{for all } v \in V. \tag{3.1}$$

The left-hand side of (3.1) defines a symmetric bilinear form $\mathfrak{a}$,

$$\mathfrak{a}(u, v) := \int_\Omega c\langle \nabla u, \nabla v\rangle \mathrm{d}x.$$

The sum $u^* := u + u_D$ is the solution of problem (3.1); u_D denotes some extension (with finite energy) of the given inhomogeneous Dirichlet boundary data to Ω. After shifting the inhomogeneous boundary data to the right-hand side, the problem reduces to find $u \in V$ such that

$$\mathfrak{a}(u, v) = \int_\Omega f v \mathrm{d}x \; - \mathfrak{a}(u_D, v) =: F(v) \quad \text{for all } v \in V. \tag{3.2}$$

It is obvious that

$$\frac{1}{1 + C_{\mathrm{F}}} \|v\|^2_{H^1(\Omega)} \le \mathfrak{a}(v, v) \quad \text{and} \quad \mathfrak{a}(u, v) \le c_{\mathrm{cont}} \|u\|_{H^1(\Omega)} \|v\|_{H^1(\Omega)} \tag{3.3}$$

for all $u, v \in V$ with the constant from Friedrichs' inequality C_{F}. Inequality (3.3) ensures the unique solvability of the variational problem (3.2) for finite contrast $c_{\mathrm{cont}} < \infty$.

The Galerkin approximation of the solution of (3.2) with respect to the finite element space S, denoted by $u_S \in S$, is defined as the solution to the discrete variational system

$$\mathfrak{a}(u_S, v) = F(v) \quad \text{for all } v \in S. \tag{3.4}$$

Remark 3 a) The assembling of the corresponding linear system is fairly standard. It might be performed in a loop over all elements of the generalized finite element mesh (including triangles, disks, and edges), the computation of the local stiffness matrices and load vectors, and the sum of the local contributions to the global matrices. The computation of the entries of the local stiffness matrices might be done by transformation to the corresponding reference element. The only difficulty is that the transformation on the generalized edges is not affine. Still the entries of the local stiffness matrices might be precomputed as functions of the angle parameters and δ. Alternatively, numerical quadrature can be used. If two inclusions are close to each other, the basis functions are close to be singular and the quadrature rule should take the singular behavior into account.

b) The resulting stiffness matrix has a similar sparsity pattern as the stiffness matrix of the classical P_1 finite element method for the Poisson problem with respect to some regular triangulation. Hence, in the present 2-dimensional setting, sparse direct solvers offer robust, fast, and parallel solution of the linear system, even though the asymptotic complexity is not optimal (e.g. $\mathcal{O}(N^{3/2})$ for nested dissection [15]). We refer to the textbook [9] for an overview on fast direct solvers for sparse linear systems. For moderate contrast, [1] and [2] show that an iterative solver based on hierarchical factorization performs almost optimal (i.e. $\mathcal{O}(N(\log N)^k)$). In the numerical examples in [1,2], theses methods give promising results also in the high contrast regime.

This paper aims at a priori estimates of the error $u - u_S$ in energy norm $\|\cdot\|_{\mathfrak{a}} := \sqrt{\mathfrak{a}(\cdot,\cdot)}$ and therefore estimates of the error in the effective conductivity. Since u_S is the best approximation of u in energy norm we have

$$\begin{aligned} 2(\mathfrak{E}(u_S + u_D) - \mathfrak{E}(u + u_D)) &= \|(u + u_D) - (u_S + u_D)\|_{\mathfrak{a}}^2 \\ &= \|u - u_S\|_{\mathfrak{a}}^2 = \inf_{v \in S} \|u - v\|_{\mathfrak{a}}^2. \end{aligned} \tag{3.5}$$

Sections 3.3 and 3.4 will present bounds of the right hand side in (3.5). A posteriori bounds are presented in [12].

3.2 Perfectly conducting inclusions

Our analysis shall cover the case of perfectly conducting inclusions as well. The related model is a variational problem with respect to the reduced space

$$V^\infty := \{v \in V : \ v|_B = const \quad \text{for all } B \in \mathcal{B}_{\text{inc}}\} \subset V.$$

We seek $u^\infty \in V^\infty$ such that

$$\mathfrak{a}^\infty(u^\infty, v) = \int_\Omega f v \mathrm{d}x \ - \mathfrak{a}^\infty(u_D^\infty, v) =: F(v) \quad \text{for all } v \in V^\infty, \tag{3.6}$$

where $\mathfrak{a}^\infty(u,v) := \int_{\Omega_{\text{mat}}} \langle \nabla u, \nabla v \rangle \mathrm{d}x$ for $u, v \in H^1(\Omega)$ and $u_D^\infty \in H^1(\Omega)$ with $u_D^\infty|_{\Gamma_D} = u_D|_{\Gamma_D}$ and $\nabla u_D^\infty|_B = 0$ for all $B \in \mathcal{B}_{\text{inc}}$.

Since $\mathfrak{a}^\infty(u,v) \le \|u\|_{H^1(\Omega)} \|v\|_{H^1(\Omega)}$ and $\mathfrak{a}^\infty(v,v) = \|\nabla v\|_{L^2(\Omega)}^2$ for all $u, v \in V^\infty$, the variational problem (3.6) has a unique solution.

The Galerkin approximation $u_{S^\infty} \in S^\infty$ of the solution of (3.6), with respect to the finite element space S^∞ defined in (2.6), satisfies

$$\mathfrak{a}^\infty(u_{S^\infty}, v) = F(v) \quad \text{for all } v \in S^\infty. \tag{3.7}$$

The error estimate (3.5) remains valid with u replaced by u^∞ and u_S replaced by u_{S^∞}.

Mathematical justification of the limiting problem We shall justify the model problem (3.6). For fixed geometry Ω_{mat}, Dirichlet data $u_D = 0$, and force term f, let $u_{c_{\mathrm{cont}}}$ denote the solution of (3.2) associated with the contrast parameter $c_{\mathrm{cont}} \geq 1$.

Define some function $\tilde{u}_{c_{\mathrm{cont}}} \in V$ as follows. On every $B \in \mathcal{B}_{\mathrm{inc}}$, $(\tilde{u}_{c_{\mathrm{cont}}})|_B$ equals $u_{c_{\mathrm{cont}}}|_B$ minus its mean value $|B|^{-1}\int_B u_{c_{\mathrm{cont}}}\mathrm{d}x$. This defines $(\tilde{u}_{c_{\mathrm{cont}}})|_{\Omega_{\mathrm{inc}}}$. Observe that Poicareé's inequality yields $\|\tilde{u}_{c_{\mathrm{cont}}}\|_{L^2(\Omega_{\mathrm{inc}})} \leq C_1\|\nabla\tilde{u}_{c_{\mathrm{cont}}}\|_{L^2(\Omega_{\mathrm{inc}})} = C_1\|\nabla u_{c_{\mathrm{cont}}}\|_{L^2(\Omega_{\mathrm{inc}})}$ with some constant C_1 independent of contrast and the positions of the inclusions. In Ω_{mat}, $\tilde{u}_{c_{\mathrm{cont}}}$ is chosen as some bounded extension of $(\tilde{u}_{c_{\mathrm{cont}}})|_{\Omega_{\mathrm{inc}}}$ in the sense of [26], i.e., there is some constant C_2 that may depend on the geometry but not on c_{cont} such that $\|\tilde{u}_{c_{\mathrm{cont}}}\|_{H^1(\Omega)} \leq C_2\|u_{c_{\mathrm{cont}}}\|_{H^1(\Omega_{\mathrm{inc}})}$.

This construction and the classical jump relation at the interface $\partial\Omega_{\mathrm{inc}}$,

$$c_{\mathrm{cont}}\frac{\partial\left((u_{c_{\mathrm{cont}}})|_{\Omega_{\mathrm{inc}}}\right)}{\partial\nu_{\Omega_{\mathrm{inc}}}} = -\frac{\partial\left((u_{c_{\mathrm{cont}}})|_{\Omega_{\mathrm{mat}}}\right)}{\partial\nu_{\Omega_{\mathrm{mat}}}} \quad \text{in } H^{-1/2}(\partial\Omega_{\mathrm{inc}}), \tag{3.8}$$

yield

$$\begin{aligned}
&\|\nabla u_{c_{\mathrm{cont}}}\|_{L^2(\Omega_{\mathrm{inc}})}\\
&\quad= \int_{\Omega_{\mathrm{inc}}}\langle\nabla u_{c_{\mathrm{cont}}},\nabla\tilde{u}_{c_{\mathrm{cont}}}\rangle\mathrm{d}x = \int_{\partial\Omega_{\mathrm{inc}}}\frac{\partial u_{c_{\mathrm{cont}}}}{\partial\nu_{\Omega_{\mathrm{inc}}}}\tilde{u}_{c_{\mathrm{cont}}}\mathrm{d}x + c_{\mathrm{cont}}^{-1}\int_{\Omega_{\mathrm{inc}}} f\tilde{u}_{c_{\mathrm{cont}}}\mathrm{d}x\\
&\quad\leq c_{\mathrm{cont}}^{-1}\left(\left|\int_{\partial\Omega_{\mathrm{mat}}}\frac{\partial u_{c_{\mathrm{cont}}}}{\partial\nu_{\Omega_{\mathrm{mat}}}}\tilde{u}_{c_{\mathrm{cont}}}\mathrm{d}x\right| + \|f\|_{L^2(\Omega_{\mathrm{inc}})}\|\tilde{u}_{c_{\mathrm{cont}}}\|_{L^2(\Omega_{\mathrm{inc}})}\right)\\
&\quad\leq c_{\mathrm{cont}}^{-1}\left(\|\nabla u_{c_{\mathrm{cont}}}\|_{L^2(\Omega_{\mathrm{mat}})}\|\nabla\tilde{u}_{c_{\mathrm{cont}}}\|_{L^2(\Omega_{\mathrm{mat}})} + \|f\|_{L^2(\Omega_{\mathrm{mat}})}\|\tilde{u}_{c_{\mathrm{cont}}}\|_{L^2(\Omega_{\mathrm{mat}})}\right.\\
&\qquad\left.+\|f\|_{L^2(\Omega_{\mathrm{inc}})}\|\tilde{u}_{c_{\mathrm{cont}}}\|_{L^2(\Omega_{\mathrm{inc}})}\right)\\
&\quad\leq Cc_{\mathrm{cont}}^{-1}\|f\|_{L^2(\Omega)}\|\nabla u_{c_{\mathrm{cont}}}\|_{L^2(\Omega_{\mathrm{inc}})},
\end{aligned}$$

where C depends only on C_1 and C_2 but not on c_{cont}. This implies

$$\|c^{1/2}\nabla u_{c_{\mathrm{cont}}}\|_{L^2(\Omega_{\mathrm{inc}})} \leq Cc_{\mathrm{cont}}^{-1/2}\|f\|_{L^2(\Omega)}.$$

Hence, the solution $u_{c_{\mathrm{cont}}}$ of (3.2) converges (with respect to the energy norm) to the solution u^∞ of (3.6) as $c_{\mathrm{cont}} \to \infty$.

3.3 Nodal interpolation and approximability

An upper bound for the right-hand side in (3.5) is derived through the design of some finite element function based on a suitable interpolation of the solution u. The conditions

$$\int_B (u - Iu)v\mathrm{d}x = 0 \quad \text{for all } v \in \mathbb{P}^1(\mathbb{R}^2) \quad \text{and} \quad \text{for all} B \in \mathcal{B}_{\text{inc}}, \tag{3.9a}$$

$$u(b) - Iu(b) = 0 \quad \text{for all } B = \{b\} \in \mathcal{B}_{\text{mat}}, \tag{3.9b}$$

define a generalized nodal interpolation operator $I : H^2(\Omega_{\text{mat}} \cup \Omega_{\text{inc}}) \to S_0$. Since, on any inclusion $B \in \mathcal{B}_{\text{inc}}$, Iu is the $L^2(B)$ projection of u onto the space of affine functions, we have that

$$\|\nabla^m(u - Iu)\|_{L^2(B)} \leq C_I \operatorname{diam}(B)^{2-m} |u|_{H^2(B)} \quad \text{for } m = 0, 1 \tag{3.10}$$

with some universal constant C_I independent of the diameter of the disk B and $u \in H^2(\Omega_{\text{mat}} \cup \Omega_{\text{inc}})$. The estimate (3.10) already provides approximation properties of the finite element space on the inclusions. It remains to give local estimates for the interpolation error on the triangles (see Lemma 3.1) and the generalized edges (see Lemma 3.3).

As usual, the error on a triangle T depends on the aspect ratio ρ_T, i.e., the ratio between the diameters of the largest circle that can be inscribed in T and the circumcircle of T.

Lemma 3.1 *Let $u \in V \cap H^2(\Omega_{\text{mat}} \cup \Omega_{\text{inc}})$ and let $T \in \mathcal{T}$ with vertices on $B_1, B_2, B_3 \in \mathcal{B}$. Then it holds*

$$\|\nabla(u - Iu)\|^2_{L^2(T)} \leq C^2_{\mathcal{T}} \rho_T^{-2} \|h\nabla^2 u\|^2_{L^2(T \cup B_1 \cup B_2 \cup B_3)} \tag{3.11}$$

with some universal constant $C_{\mathcal{T}}$ which depends only on C_I from (3.9).

Proof A key ingredient of the proof are standard estimates for the interpolation error with respect to a triangle T. It is well known (see [8, Theorem 16.1]) that the nodal (affine) interpolant $I_T u$ of u at the vertices of T satisfies

$$|u - I_T u|_{H^m(T)} \leq C_{\text{ip}} \rho_T^{-1} \operatorname{diam}(T)^{2-m} |u|_{H^2(T)} \quad \text{for all } u \in H^2(T),\ m = 0, 1. \tag{3.12}$$

The difficulty is that Iu defined by (3.9) does not interpolate u at the vertices of T in general. Thus, the error is split into two components,

$$\|\nabla(u - Iu)\|^2_{L^2(T)} \leq \|\nabla(u - I_T u)\|^2_{L^2(T)} + \|\nabla(I_T u - Iu)\|^2_{L^2(T)}. \tag{3.13}$$

The first term on the right-hand side of (3.13) can be estimated directly with (3.12) while the second one requires further considerations.

Notice that $\nabla(I_T u - Iu)|_T$ is constant on T and the inverse estimate

$$\|\nabla q\|_{L^\infty(T)} \leq 2\rho_T^{-1} \operatorname{diam}(T)^{-1} \|q\|_{L^\infty(T)} \tag{3.14}$$

holds for all $q \in \mathbb{P}_1(T)$ on any triangle T. Thus

$$\|\nabla(I_T u - Iu)\|^2_{L^2(T)} \leq |T| |\nabla(I_T u - Iu)|^2_{L^\infty(T)} \overset{(3.14)}{\leq} 4\rho_T^{-2} \|I_T u - Iu\|^2_{L^\infty(T)}. \tag{3.15}$$

The maximal absolute value of the affine function $q := (I_T u - Iu)|_T$ on T is attained in some vertex $x_0 = V(T) \cap B_T$ for some $B_T \in \{B_1, B_2, B_3\}$. If $B_T \in \mathcal{B}_{\text{mat}}$, i.e., $B_T = x_0$, then $(I_T u - Iu)|_T = 0$. Otherwise, let $\tilde{T} \subset B_T$ be the equilateral triangle with vertices on ∂B_T and one vertex at x_0. For $q \in \mathbb{P}_1(T)$ and $p \in \mathbb{P}_1(\tilde{T})$ with $|p(x_0)| \geq |q(x_0)|$ it holds

$$\|q\|^2_{L^\infty(T)} = |q(x_0)|^2 \leq |p(x_0)|^2 \leq 2\left(|\tilde{T}|^{-1} \|p\|^2_{L^2(\tilde{T})} + \|\nabla p\|^2_{L^2(\tilde{T})}\right). \tag{3.16}$$

With the special choices $p = (I_T u - Iu)|_T$ and $q = (I_{\tilde{T}} u - Iu)|_{\tilde{T}}$ this leads to

$$\begin{aligned} \|\nabla(I_T u - Iu)\|^2_{L^2(T)} &\overset{(3.15),(3.16)}{\leq} 8\rho_T^{-2}\left(|\tilde{T}|^{-1} \|I_{\tilde{T}} u - Iu\|^2_{L^2(\tilde{T})} + \|\nabla(I_{\tilde{T}} u - Iu)\|^2_{L^2(\tilde{T})}\right) \\ &\overset{(3.12),(3.10)}{\leq} 16\rho_T^{-2}(C_I^2 + C_{\text{ip}}^2) h^2_{B_T} \|\nabla^2 u\|^2_{L^2(B_T)}. \end{aligned} \tag{3.17}$$

Together with (3.13) and (3.12) this implies (3.11) with $C_{\mathcal{T}}^2 \leq 5(C_I + C_{\text{ip}})$. □

The second step of the error analysis considers the a priori estimate of the interpolation error on the generalized edges. Every connectivity component E_k, $k = 1, 2, \ldots, K(E)$ of an edge $E \in \mathcal{E}$ is a curvilinear polygon, i.e., E_k is a simply-connected, bounded domain with the boundary $\partial E_k = \bigcup_{j=1}^4 \tau_j$, where τ_j are circular arcs. Note that all internal angles $\gamma_1(E_k), \gamma_2(E_k), \ldots, \gamma_4(E_k)$ of E_k are bounded from above by $\pi/2$. The subsequent error analysis depends on the smallest angle which is denoted γ_{E_k}. Correspondingly, $\gamma_E := \min_{k=1,2,\ldots,K(E)} \gamma_{E_k}$. The following lemma shows that all these angles are bounded from below by a positive constant.

Lemma 3.2 *There exist $\gamma_{\mathcal{E}} > 0$ such that $0 < \gamma_{\mathcal{E}} \leq \gamma_E$ for all $E \in \mathcal{E}$.*

Proof Let $E \in \mathcal{E}$ be some generalized edge connected to the inclusion $B \in \mathcal{B}_{\text{inc}}$. Let τ be one of the straight arcs that define the edge. By design, τ is an element of the infinite Delaunay triangulation $\mathcal{T}_{\text{mat}}$ (see Sect. 2.1). Since its circumdisk $CD(\tau)$ is tangential to B (due to the Delaunay criterion (2.1)), τ by itself cannot be tangential to B and the angle between τ and the circular arc $E \cap B$ is necessarily larger than zero.

Lemma 3.3 *Let $u \in V \cap H^2(\Omega_{\text{mat}} \cup \Omega_{\text{inc}})$ and let $E = E(B_1, B_2) \in \mathcal{E}$ be a generalized edge that connects two generalized vertices (inclusions) $B_1, B_2 \in \mathcal{B}_{\text{inc}}$. Then*

$$\|\nabla(u - Iu)\|^2_{L^2(E)} \leq C_{\mathcal{E}}\left(\|h\nabla^2 u\|^2_{L^2(E)} + C_E \|h\nabla^2 u\|^2_{L^2(B_1 \cup B_2)}\right)$$

holds with $C_E := \max_{k=1,2} \left\| h_{B_k}/h + h/h_{B_k} \right\|_{L^\infty(E)}$ and some universal constant $C_{\mathcal{E}}$ which depends only on γ_E.

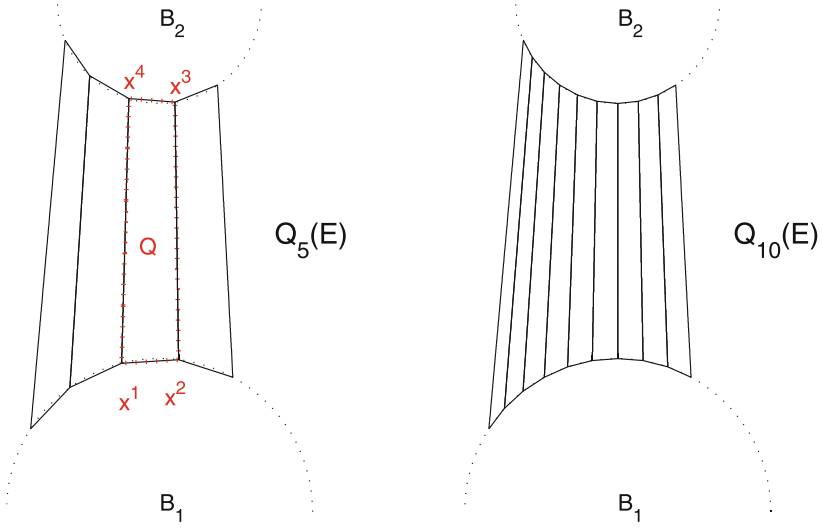


Fig. 6 Subdivisions $\mathcal{Q}_5(E)$ and $\mathcal{Q}_{10}(E)$ of some generalized edge $E = E(B_1, B_2) \in \mathcal{E}$ into quadrilaterals in the proof of Lemma 3.3

Proof The proof consists of two parts. Part *I* proves the assertion for $c_{\text{cont}} = \infty$ and prepares the proof in the case $c_{\text{cont}} < \infty$ which is complemented in part *II*.

Part I. Without loss of generality, let E be connected, $r_{B_1} \geq r_{B_2}$, and $c_{B_1} = 0$, $c_{B_2} = (0, \delta)$ for some $\delta > r_{B_1} + r_{B_2}$. The restriction $E \cap \partial B_1 = \phi([\alpha, \beta])$ of E to B_1 shall be parametrized by some angle $s \in [\alpha, \beta] \subset [-\pi/2, \pi/2]$ with $\phi(s) := r_{B_1}(sin(s), cos(s))$. The parameter interval $[\alpha, \beta]$ is subdivided by equidistributed points

$$\alpha = s^1 < s^2 < s^3 < \cdots < s^L = \beta.$$

These points are mapped by ϕ onto B_1 and by $\phi \circ \pi_{B_1}$ onto B_2 (recall (2.4) for the definition of π_{B_1}). Let

$$\begin{aligned} \mathcal{Q}_L(E) &:= \{Q_\ell \ : \ell = 1, \ldots, L-1\} \quad \text{with} \\ Q_\ell &:= \text{conv}\Big(\phi(s^\ell), \phi(s^{\ell+1}), \pi_{B_1}(\phi(s^{\ell+1})), \pi_{B_1}(\phi(s^\ell))\Big) \end{aligned}$$

be a subdivision of E into quadrilaterals (see Fig. 6).

The union of quadrilaterals on level L provides a polygonal approximation $E^L := \bigcup_{Q \in \mathcal{Q}_L(E)} Q$ of $E \subset E^L \subset \text{conv}(E)$ for all L with $|E^L \backslash E| \to 0$ as $L \to \infty$. A (bounded) extension operator $(\cdot)_E : H^2(E) \to H^2(\mathbb{R}^d)$ (see, e.g., [26]) extends $u|_E$ to $\text{conv}(E)$. The extended function is denoted by u_E.

The nodal (bilinear) interpolation operator with respect to $Q \in \mathcal{Q}_L$ is denoted by J_Q and its $\mathcal{Q}_L$-piecewise version by $J_{\mathcal{Q}_L}$. Theorem 3.8 from [22] implies

$$\|\nabla(u_E - J_Q u_E)\|_{L^2(Q)} \leq C_Q \text{diam}\big(Q\big) \|\nabla^2 u_E\|_{L^2(Q)} \tag{3.18}$$

for all $Q \in \mathcal{Q}_L$, $L \in \mathbb{N}$. The constant C_Q depends only on the interior angles of Q, i.e., C_Q can be bounded uniformly for all $Q \in \mathcal{Q}_L$ and all $L \in \mathbb{N}$ in terms of γ_E. Thus

$$\begin{aligned} \|\nabla(u_E - J_{\mathcal{Q}_L} u_E)\|^2_{L^2(E^L)} &= \sum_{Q\in\mathcal{Q}_L} \|\nabla(u_E - J_Q u_E)\|^2_{L^2(Q)} \\ &\overset{(3.18)}{\leq} \sum_{Q\in\mathcal{Q}_L} C_1 \|\operatorname{diam}(Q)\nabla^2 u_E\|^2_{L^2(Q)} \end{aligned} \tag{3.19}$$

with some constant C_1 which depends only on γ_E. Let L tend to infinity in (3.19) to verify

$$\|\nabla(u - \tilde{u})\|^2_{L^2(E)} \leq C_1 \|h\nabla^2 u\|^2_{L^2(E)} \tag{3.20}$$

for $\tilde{u} := \lim_{L\to\infty} J_{\mathcal{Q}_L} u_E$. If $c_{\mathrm{cont}} = \infty$ then $\tilde{u} = Iu$ and the proof is finished.

Part II. If otherwise $c_{\mathrm{cont}} < \infty$ then, in general, $\tilde{u} \notin S$ and $\|\nabla(Iu - \tilde{u})\|_{L^2(E)}$ needs to be estimated further. The sequence $e_L := J_{\mathcal{Q}_L}(Iu)_E - J_{\mathcal{Q}_L} u_E$ converges (in H^1) to $e := Iu - \tilde{u}$ as $L \to \infty$. Thus, bounds on $\|\nabla e_L\|_{L^2(E_L)}$ will lead to a bound on $\|\nabla(Iu - \tilde{u})\|_{L^2(E)}$. Let $Q \in \mathcal{Q}_L$ with

$$\partial Q = [x^1, x^2] \cup [x^2, x^3] \cup [x^3, x^4] \cup [x^4, x^1]$$

and $x^1, x^2 \in B_1$ and $x^3, x^4 \in B_2$ and $x^5 = x^1$ as in Fig. 6 (left). The vector $\nabla e_L|_Q$ is written as some linear combination of the vectors $(x^{k+1} - x^k)/|x^{k+1} - x^k|$ such that

$$\|\nabla e_L\|^2_{L^2(Q)} \leq |Q| \|\nabla e_L\|^2_{L^\infty(Q)} \leq C_2 |Q| \sum_{k=1}^{4} \frac{\left|\langle (\nabla e_L)|_{[x^k, x^{k+1}]}, x^{k+1} - x^k\rangle\right|^2}{|x^{k+1} - x^k|^2}$$

with a constant C_2 which depends only on the maximal angle in Q and can be bounded uniformly in terms of γ_E^{-1}. Using $\langle (\nabla e_L)|_{[x^k, x^{k+1}]}, x^{k+1} - x^k\rangle = e_L(x^{k+1}) - e_L(x^k)$ for $k = 2$ and $k = 4$, this yields

$$\begin{aligned} \|\nabla e_L\|^2_{L^2(Q)} \leq C_2 \Big(&\|h\|_{L^\infty(Q)} \|\nabla e_L\|^2_{L^2([x^1,x^2]\cup[x^3,x^4])} \\ &+ \|h^{-1}\|_{L^\infty(Q)} \|e\|^2_{L^2([x^2,x^3]\cup[x^4,x^1])} \Big). \end{aligned} \tag{3.21}$$

The summation of (3.21) over $Q \in \mathcal{Q}_L$ leads to

$$\begin{aligned} \|\nabla e_L\|^2_{L^2(E_L)} \leq C_2 \Big(&\|h\|_{L^\infty(E)} \|\nabla e_L\|^2_{L^2(\partial E_L \cap (B_1 \cup B_2))} \\ &+ \|h^{-1}\|_{L^\infty(E)} \|e_L\|^2_{L^2(\partial E_L \cap (B_1 \cup B_2))} \Big). \end{aligned}$$

In the limit $L \to \infty$ it follows

$$\|\nabla e\|^2_{L^2(E)} \leq C_2 \left(\|h\|_{L^\infty(Q)} \|\nabla e\|^2_{L^2(\partial E \cap (B_1 \cup B_2))} + \|h^{-1}\|_{L^\infty(Q)} \|e\|^2_{L^2(\partial E \cap (B_1 \cup B_2))} \right). \tag{3.22}$$

Estimate (3.10) and the trace inequality

$$\|f\|_{L^2(\partial B)} \leq \sqrt[4]{8} \left(\|f\|_{L^2(B)} + \|f\|^{1/2}_{L^2(B)} \|\nabla f\|^{1/2}_{L^2(B)} \right), \tag{3.23}$$

valid for any disk B and $f \in H^1(B)$ (see [6, Proposition 1.6.3]), imply

$$\begin{aligned} |e|_{H^m(\partial B)} &= |\tilde{u} - Iu|_{H^m(\partial B)} = |u - Iu|_{H^m(\partial B)} \\ &\overset{(3.23),(3.10)}{\leq} \sqrt[4]{8} C_I r_B^{2-m-\frac{1}{2}} |u|_{H^2(B)} \quad \text{for } m = 0, 1. \end{aligned} \tag{3.24}$$

With a universal constant C_3 which depends only on C_I and γ_E (through C_1 and C_2), this leads to

$$\|\nabla(Iu - \tilde{u})\|^2_{L^2(E)} \overset{(3.21),(3.24)}{\leq} C_3 \|h^{-1}(h|_{B_k}) + h(h|_{B_k})^{-1}\|_{L^\infty(E)} \|h\nabla^2 u\|^2_{H^2(B_1 \cup B_2)}.$$

This concludes the proof of the lemma. □

The constant C_E reflects the fact that two inclusions might touch but the corresponding affine approximations of the solution on the disks might not match at the touching point. Thus, in rare cases for $c_{\text{cont}} < \infty$, the discrete system might have infinite energy whereas the continuous solution has not. Choosing sufficiently many degrees of freedom (number of degrees of freedom per inclusion larger than or equal to the number of neighbors per inclusion) this problem disappears.

3.4 A priori error estimates

The approximation property of the finite element space S reads as follows.

Theorem 3.1 *Let $u \in V \cap H^2(\Omega_{\text{mat}} \cup \Omega_{\text{inc}})$ be the solution of (3.2) and let $u_S \in S$ be its Galerkin approximation that solves (3.4). Then it holds*

$$\|u - u_S\|^2_{\mathfrak{a}} \leq C_S^2 \left(\|h\nabla^2 u\|^2_{L^2(\Omega_{\text{mat}})} + c_{\text{cont}} \sum_{B \in \mathcal{B}_{\text{inc}}} C_B \|h\nabla^2 u\|^2_{L^2(B)} \right)$$

with $C_B := \|h_B/h + h/h_B\|_{L^\infty(\omega_B)}$ and some universal constant C_S which depends only on C_I, $C_{\mathcal{T}}$, and $C_{\mathcal{E}}$.

Proof The proof is a straight forward consequence of (3.10), Lemma 3.1, Lemma 3.3, and the equality

$$\|v\|_{\mathfrak{a}}^2 = \|\nabla v\|_{L^2(\Omega_{\mathrm{mat}})}^2 + c_{\mathrm{cont}}\|\nabla v\|_{L^2(\Omega_{\mathrm{inc}})}^2 \quad \text{for all } v \in H^1(\Omega).$$

□

By (3.3) the estimate of Theorem 3.1 is also valid for the error measured in the $H^1(\Omega)$-norm. The regularity results from [7, Appendix B] read

$$\|\nabla^2 u\|_{L^2(\Omega_{\mathrm{mat}})} \leq C_{\mathrm{reg}}\|f\|_{L^2(\Omega)}, \quad \|\nabla^2 u\|_{L^2(\Omega_{\mathrm{inc}})} \leq \frac{C_{\mathrm{reg}}}{c_{\mathrm{cont}}}\|f\|_{L^2(\Omega)}. \tag{3.25}$$

The constant C_{reg} depends solely on the geometry of the set inclusions and Ω but *not* on c_{cont}. This implies that the contrast is not a critical parameter.

Corollary 3.1 *Let $u \in V \cap H^2(\Omega_{\mathrm{mat}} \cup \Omega_{\mathrm{inc}})$ be the solution of (3.2) and $u_S \in S$ its Galerkin approximation that solves (3.4). Then it holds*

$$\|u - u_S\|_{\mathfrak{a}} \leq \tilde{C}_S\|h\|_{L^\infty(\Omega)}(\|f\|_{L^2(\Omega)} + \|\nabla u_D\|_{L^2(\Omega)}) \tag{3.26}$$

with some universal constant $\tilde{C}_S$ which depends only on C_{reg} and the constants C_S, C_B from Theorem 3.1.

The constant $\tilde{C}_S$ in (3.26) does not depend on the contrast parameter $c_{\mathrm{cont}} > 1$. However, through the constants C_B, it might depend on the term (cf. the Definition of C_B in Theorem 3.1)

$$\max_{E(B_1,B_2)\in\mathcal{E}} \frac{\max\{r_{B_1}, r_{B_2}\}}{dist(B_1, B_2)\, c_{\mathrm{cont}}}. \tag{3.27}$$

The latter constant is critical with regard to the geometry of the coefficient function. The term may blow up, whenever the distance of two inclusions relative to their size becomes very small. However, high contrast reduces this effect. In the case of perfectly conducting inclusions ($c_{\mathrm{cont}} = \infty$) it even disappears. The generalized interpolation operator from (3.9) fulfills $(u - Iu)|_B = 0$ for all $B \in \mathcal{B}$ and the proof of Lemma 3.3 consists only of part *I*. Lemma 3.1 can be simplified in a similar way which leads to the following corollary.

Corollary 3.2 *Let $c_{\mathrm{cont}} = \infty$ and let $u^\infty \in V^\infty \cap H^2(\Omega_{\mathrm{mat}} \cup \Omega_{\mathrm{inc}})$ be the solution of (3.6) and $u_{S^\infty} \in S^\infty$ its Galerkin approximation that solves (3.7). Then it holds*

$$\|\nabla(u^\infty - u_{S^\infty})\|_{L^2(\Omega)} = \|u^\infty - u_{S^\infty}\|_{\mathfrak{a}} \leq C_{S^\infty}\|h\nabla^2 u^\infty\|_{L^2(\Omega_{\mathrm{mat}})}$$

with a constant C_{S^∞} independent of u^∞, c_{cont}, and the location of the inclusions.

In the general case $c_{\text{cont}} < \infty$ the critical constant shown in (3.27) can easily be reduced with higher-order ansatz functions on the inclusions. We can therefore derive error estimates whose constants are explicit in the underlying geometry. However, in all cases the dependence on the H^2-norm of the solution remains. This issue is briefly discussed in the Sect. 4.3.

4 Concluding remarks

The main result of this paper is a numerical scheme to compute temperature distributions in composite materials with a large number of particles and high contrast. In the model situation under consideration, the method is robust and does *not* depend on the contrast $c_{\text{cont}} \to \infty$. Some of the results extend to a more general geometric setting in a straight-forward way. However, some difficulties remain open.

4.1 General inclusion geometry

For the use in practical applications it is desirable to incorporate more general inclusion shapes and 3-dimensional geometries. It is shown in [23] that the generalized partitions of Sect. 2 nicely generalize to sets of convex inclusions, e.g., ellipsoids, convex polyhedra, and line segments. Even more, the design allows inclusions to intersect. Thus, generalized Delaunay triangulations are also available for non-convex inclusions which can be represented by finite unions of convex ones. The design of according finite element methods can be done similarly as presented here. However, the complexity of the mesh and the corresponding finite element method will grow as the number of shape parameters that define a single inclusion grows. For smooth inclusions the corresponding analysis is straight-forward; non-smooth inclusions, however, require new arguments which are able to cope with lack of regularity.

4.2 Convergence

By straight forward arguments it is easy to show that the finite element solutions (the solutions of (3.2) and (3.6)) converge in H^1 to the solution of (1.2) if the meshwidth function h tends to 0.

In the matrix, the meshwidth function h can be decreased in the matrix Ω_{mat} by simply putting additional artificial inclusions (points) in the set $\mathcal{B}_{\text{mat}}$. If $c_{\text{cont}} = \infty$, this suffices to be able to construct a convergent sequence of approximation because the (energy-)error in the inclusions is always zero. The case, in which additional vertices of radius zero are added to improve the approximability properties of the finite element space, is already treated by the theory presented in this article. A different possibility is to leave the initial partition as it is and increase the polynomial degree of the shape functions. This strategy, the so-called p-refinement, is recommended for problems where geometry and data are smooth. The definition of higher-order finite element spaces is to some extent straight-forward, the corresponding analysis, however, appears more involved.

If $c_{\text{cont}} < \infty$, in addition, the error on the inclusions has to be decreased, e.g., by increasing the polynomial degree.

4.3 Geometry-explicit estimates

The method presented is stable with respect to contrast in the medium. However, the error bounds might depend on geometric parameters of the material, e.g., the distance between neighboring particles. Whether or not the dependence on the local distance is critical depends on the global distribution of particles. This can be seen already in the simplified situation of perfectly conducting ($c_{\text{cont}} = \infty$) inclusions.

Consider first two inclusions that touch but are isolated from further inclusions. Since the solution is found in H^1 the (constant) values of the solution on the two inclusions have to be equal. Provided the force term is sufficiently smooth (L^2), classical regularity theory ensures smoothness of the solution in some neighborhood of the two inclusions and the constant in the regularity estimate depends only on the distance to further inclusions or the boundary of the domain.

The critical scenario is the appearance of an almost conducting path of inclusions which connects two parts of the outer boundary with different, prescribed temperature. The temperature gap needs to be compensated in the small regions between the inclusions of the path which might cause steep gradients in the solution. If the inclusions of the path touch pairwise, the path is perfectly conducting and hence, the energy is infinite. Depending on the volume fraction of particles, the material shows a phase transition from moderate to high conductivity. Mathematically speaking, the solution operator, which maps a pair the date u_D and f to the solution of (1.2), is not uniformly bounded with respect to the geometry of the set of inclusions $\mathcal{I}$ [25, Theorem 3.5] shows that, though the energy of the solution might blow up, the error estimate in Corollary 3.2 is bounded by some generic constant independent of the distance of the particles. Thus, our method is robust with respect the such critical scenarios and allows meaningful material simulation even in densely packed composites. We refer to [12] for numerical experiments.

In the general case of high but finite contrast the situation appears more involved and a corresponding regularity theory that is explicit (and sharp) with respect to both, contrast and geometric parameters, is not yet available and has to be addressed in future research.

References

1. Bebendorf, M.: Why finite element discretizations can be factored by triangular hierarchical matrices. SIAM J. Numer. Anal. **45**(4), 1472–1494 (2007)
2. Börm, S.: Approximation of solution operators of elliptic partial differential equations by $\mathcal{H}$- and $\mathcal{H}^2$-matrices. Numer. Math. **115**(2), 165–193 (2010)
3. Berlyand, L., Kolpakov, A.: Network approximation in the limit of small interparticle distance of the effective properties of a high-contrast random dispersed composite. Arch. Ration. Mech. Anal. **159**(3), 179–227 (2001)
4. Berlyand, L., Novikov, A.: Error of the network approximation for densely packed composites with irregular geometry. SIAM J. Math. Anal. **34**(2), 385–408 (2002) (electronic)
5. Borcea, L., Papanicolaou, G.C.: Network approximation for transport properties of high contrast materials. SIAM J. Appl. Math **58**, 501–539 (1998)
6. Brenner, S.C., Scott, L.R.: The mathematical theory of finite element methods. In: Texts in Applied Mathematics, 3rd edn, vol. 15. Springer, New York (2008)

7. Chu, C.-C., Graham, I.G., Hou, T.Y.: A new multiscale finite element method for high-contrast elliptic interface problems. Math. Comput. **79**, 1915–1955 (2010)
8. Ciarlet, P.: The Finite Element Method for Elliptic Problems. North Holland, Amsterdam (1978)
9. Davis, T.A.: Direct methods for sparse linear systems. In: Fundamentals of Algorithms, vol. 2. Society for Industrial and Applied Mathematics (SIAM), Philadelphia (2006)
10. Delaunay, B.: Sur la sphère vide. Izvestia Akademii Nauk SSSR, Otdelenie Matematicheskikh i Estestvennykh Nauk **7**, 793–800 (1934)
11. Weinan, E., Engquist, B.: The heterogeneous multiscale methods. Commun. Math. Sci. **1**(1), 87–132 (2003)
12. Eigel, M., Peterseim, D.: Network FEM for Composite Materials with A Posteriori Control DFG Research Center Matheon Berlin, Preprint Series, vol. 985 (2012)
13. Fortune, S.: A sweepline algorithm for Voronoĭ diagrams. Algorithmica **2**(2), 153–174 (1987)
14. Gavrilova, M., Rokne, J.: Swap conditions for dynamic Voronoi diagrams for circles and line segments. Comput. Aided Geom. Design **16**(2), 89–106 (1999)
15. George, A., Liu, J.: Computer Solution of Large Sparse Positive Definite Systems. Prentice-Hall, Englewood Cliffs (1981)
16. Hou, T.Y., Wu, X.-H.: A multiscale finite element method for elliptic problems in composite materials and porous media. J. Comput. Phys. **134**, 169–189 (1997)
17. Hughes, T.J.R., Feijóo, G.R., Mazzei, L., Quincy, J.-B.: The variational multiscale method—a paradigm for computational mechanics. Comput. Methods Appl. Mech. Eng. **166**(1–2), 3–24 (1998)
18. Kim, D.-S., Kim, D., Sugihara, K.: Voronoi diagram of a circle set from Voronoi diagram of a point set. I. Topology. Comput. Aided Geom. Design **18**(6), 541–562 (2001)
19. Kolpakov, A.A., Kolpakov, A.G.: Capacity and transport in contrast composite structures. CRC Press, Boca Raton (2010)
20. Larson, M.G., Målqvist, A.: Adaptive variational multiscale methods based on a posteriori error estimation: energy norm estimates for elliptic problems. Comput. Methods Appl. Mech. Eng. **196**(21–24), 2313–2324 (2007)
21. Målqvist, A., Peterseim, D.: Localization of Elliptic Multiscale Problems. ArXiv e-prints, 1110.0692 (2011)
22. Mao, S., Nicaise, S., Shi, Z.-C.: On the interpolation error estimates for Q_1 quadrilateral finite elements. SIAM J. Numer. Anal. **47**(1), 467–486 (2008)
23. Peterseim, D.: Generalized Delaunay partitions and composite material modeling. DFG Research Center Matheon Berlin, Preprint Series, vol. 690 (2010)
24. Peterseim, D.: Triangulating a system of disks. In: Proceedings of the EuroCG 2010. Dortmund, Germany (2010)
25. Peterseim, D.: Robustness of Finite Element Simulations in Densely Packed Random Particle Composites. Netw. Heterog Media **7**(1), 113–126 (2012)
26. Stein, E.M.: Singular Integrals and Differentiablity Properties of Function. Priceton Univ. Press, New York (1970)
27. Voronoi, G.F.: Nouvelles applications des paramètres continus à la théorie des formes quadratiques. Journal für die Reine und Angewandte Mathematik **133**, 97–178 (1907)

D.2 Robustness of finite element simulations in densely packed random particle composites

Networks and Heterogeneous Media, 7(1):113126, 2012.

NETWORKS AND HETEROGENEOUS MEDIA doi:10.3934/nhm.2012.7.113
©American Institute of Mathematical Sciences
Volume **7**, Number **1**, March **2012** pp. **113–126**

ROBUSTNESS OF FINITE ELEMENT SIMULATIONS IN DENSELY PACKED RANDOM PARTICLE COMPOSITES

Daniel Peterseim

Humboldt-Universität zu Berlin
Institut für Mathematik
Unter den Linden 6, 10099 Berlin, Germany

(Communicated by Leonid Berlyand)

Abstract. This paper presents some weighted H^2-regularity estimates for a model Poisson problem with discontinuous coefficient at high contrast. The coefficient represents a random particle reinforced composite material, i.e., perfectly conducting circular particles are randomly distributed in some background material with low conductivity. Based on these regularity results we study the percolation of thermal conductivity of the material as the volume fraction of the particles is close to the jammed state. We prove that the characteristic percolation behavior of the material is well captured by standard conforming finite element models.

1. **Introduction.** This note studies the numerical approximability of thermal diffusion in a representative class of particle composite materials (or composites). The particles (or inclusions) are pairwise disjoint closed disks $\mathcal{I} = \{I_1, I_2, \dots, I_N\}$ with positive radii. They are randomly distributed in a background material (or matrix) that occupies some open, bounded, convex, polygonal domain $\Omega \subset \mathbb{R}^2$. The inclusions are highly conducting compared to the matrix $\Omega^{\mathrm{mat}} := \Omega \setminus \cup\mathcal{I}$, a fact which is reflected in the diffusion coefficient

$$c(x) = \begin{cases} 1 & \text{if } x \in \Omega^{\mathrm{mat}}, \\ c_{\mathrm{cont}} & \text{if } x \in \cup\mathcal{I} \end{cases} \tag{1}$$

with some contrast parameter $c_{\mathrm{cont}} \gg 1$.

The thermal diffusion in the composite is modeled by the stationary heat equation,

$$-\operatorname{div} c\nabla u = f \text{ in } \Omega, \quad u = u_D \text{ on } \partial\Omega, \tag{2}$$

with a prescribed temperature u_D at the boundary of Ω and a heat source f. If the source term f is supported in the matrix and if the inclusions are assumed to be perfectly conducting ($c_{\mathrm{cont}} = \infty$), then problem (2) reduces to an equation in the perforated domain Ω^{mat}. Consider the function spaces

$$V := \{v \in H^1(\Omega^{\mathrm{mat}}) : \ v|_{\partial I} = \text{const. for all } I \in \mathcal{I}\} \text{ and}$$
$$V_0 := \{v \in V : \ v|_{\partial\Omega} = 0 \text{ in the sense of traces}\}.$$

2000 *Mathematics Subject Classification.* Primary: 35B65, 65N15; Secondary: 65N30, 74Q20.
Key words and phrases. Perforated domain, thickness of a domain, finite element method, high contrast, percolation, phase transition.
The author is supported by the DFG Research Center Matheon Berlin through project C33.

Then the corresponding variational problem reads: *Given* $f \in L^2(\Omega^{\mathrm{mat}})$ *and* $u_D \in C^2(\partial\Omega)$, *find* $u \in V$ *such that*

$$\int_{\Omega^{\mathrm{mat}}} \nabla u(x)\nabla v(x)\,\mathrm{d}x = \int_{\Omega^{\mathrm{mat}}} f(x)v(x)\,\mathrm{d}x \quad \textit{for all } v \in V_0 \tag{3.a}$$

and

$$u(x) = u_D(x) \quad \text{for almost all } x \in \partial\Omega. \tag{3.b}$$

Since the elements of V have a constant trace on the boundary of a single inclusion, they can trivially be extended to Ω in a way that the extension $v \in H^1(\Omega)$ satisfies $\nabla v|_{(\cup\mathcal{I})} = 0$. Hence, the inequalities of Friedrichs and Schwarz yield

$$\|v\|^2_{H^1(\Omega^{\mathrm{mat}})} \leq (1 + \operatorname{diam}(\Omega)^2)\|\nabla v\|^2_{L^2(\Omega^{\mathrm{mat}})} \text{ and} \tag{4.a}$$

$$\int_{\Omega^{\mathrm{mat}}} \nabla u(x)\nabla v(x)\,\mathrm{d}x \leq \|u\|_{H^1(\Omega^{\mathrm{mat}})}\|v\|_{H^1(\Omega^{\mathrm{mat}})} \tag{4.b}$$

for all $u, v \in V_0$. The inequalities (4) ensure the unique solvability of the variational problem (3).

The major difficulty in discretizing (3) arises from the fact that the energy of the solution u, given by $\|\nabla u\|^2_{L^2(\Omega^{\mathrm{mat}})}$, might depend crucially on the geometric properties of the filler. Consider the appearance of an almost conducting path of inclusions, which connects two parts of the outer boundary $\partial\Omega$ where different temperatures are prescribed (as in Figure 1.a). The gap in the temperature needs to be compensated on the path, i.e., in the small regions (characterized by a small parameter δ_{cond} in Figure 1.a) between the inclusions of the path. Hence, the solution shows steep gradients there. If the inclusions of the path touch pairwise, the path is perfectly conducting and hence, the energy is infinite. Depending on the volume fraction of particles, the material shows a phase transition from moderate to high conductivity. Mathematically speaking, the solution operator, which maps a pair $(u_D, f) \in C^2(\partial\Omega) \times L^2(\Omega^{\mathrm{mat}})$ to the solution of (3), is not uniformly bounded with respect to the geometry of the set of inclusions $\mathcal{I}$.

In this study, we will show that standard conforming[1] finite element approximations of (3) (denoted by u^{fem}) capture such a percolation phenomenon effectively. More precisely,

$$\|\nabla(u - u^{\mathrm{fem}})\|_{L^2(\Omega^{\mathrm{mat}})} \leq C \tag{5}$$

holds with some generic constant C independent of the distance of the particles (see Theorem 4.1). This estimate is true although $\|\nabla u\|$ might blow up as described before. Thus, conforming finite element methods are robust with respect to $\delta_{\mathrm{cond}} \to 0$ and allow meaningful material simulation even in densely packed composites.

The issue of percolation and its numerical traceability in transport problems related to high (infinite) contrast particle composites was previously addressed by discrete network models [4, 2, 3]. A pioneering result [3, Theorem 3.3] is that discrete network models, for equally sized inclusions in the absence of outer forces ($f = 0$), mimic the blow-up of the energy as the volume fraction of the particles is close to the jammed state.

Compared to the analysis in [2, 3], which rests mainly on duality arguments, our analysis is built upon regularity estimates for the solution of (3) in certain weighted

[1] A finite element method is called conforming if the corresponding finite element space is contained in V. In the present context, conformity shall primarily ensure that the complicated geometry of the composite is resolved exactly – or at least sufficiently accurate compared to the geometric scales in the problem – by the underlying finite element mesh.

norms. In this context, the weight (denoted by δ) reflects the local thickness of the perforated domain Ω^{mat} (see Section 2.1). By choosing this specific weight, the constant in the regularity estimates (cf. Theorems 3.3 and 3.5) turns out to be independent of δ, i.e., they do not depend on the distances between the inclusions. The combination of the quasi-optimality of conforming finite elements, standard interpolation error estimates, and the new regularity estimates yield the general statement on robustness (5) without even specifying a discrete space precisely. Our technique generalizes in a straight forward way to problem classes beyond the model problem under consideration, e.g., to more general inclusion geometries, to the 3-dimensional case, and to general second order elliptic operators.

2. Geometric preliminaries.

This section manifests the notion of thickness of a perforated domain and a finite, problem-adapted subdivision of the perforated domain under consideration.

2.1. The thickness of a domain.

Our definition of thickness relies on a certain (infinite) triangulation of Ω^{mat}, which is first introduced.

A convex polygon T is the convex hull of 2 or more distinct points. The set of its vertices (corners) $\mathcal{V}(T)$ is the minimal set of points $x_1, x_2, \dots, x_k \in \mathbb{R}^2$, so that $T = \text{conv}(\{x_1, x_2, \dots, x_k\})$. According to the above definition, convex polygons are closed. A convex polygon T is called *cyclic* if its vertices (corners) $\mathcal{V}(T)$ are located on the boundary of its (closed) circumdisk $\text{CD}(T)$. Examples of cyclic polygons are line segments, triangles and rectangles.

Following [9], Ω^{mat} can be represented by a regular, infinite subdivision $\mathcal{T}_{\text{mat}}$ into cyclic polygons (or triangulation for short). More precisely, $\mathcal{T}_{\text{mat}}$ is a set of cyclic polygons such that its set of vertices $\mathcal{V}(\mathcal{T}_{\text{mat}})$ equals $\partial\Omega^{\text{mat}}$,

$$\mathcal{V}(\mathcal{T}_{\text{mat}}) := \bigcup_{T \in \mathcal{T}_{\text{mat}}} \mathcal{V}(T) = \partial\Omega^{\text{mat}},$$

and any two distinct cyclic polygons in $\mathcal{T}_{\text{mat}}$ are either disjoint, or share exactly one vertex, or have exactly one edge in common. Moreover, the triangulation $\mathcal{T}_{\text{mat}}$ can be chosen in a way that all of its elements $T \in \mathcal{T}_{\text{mat}}$ satisfy the so-called Delaunay criterion

$$\text{CD}(T) \cap \mathcal{V}(\mathcal{T}_{\text{mat}}) = \mathcal{V}(T). \tag{6}$$

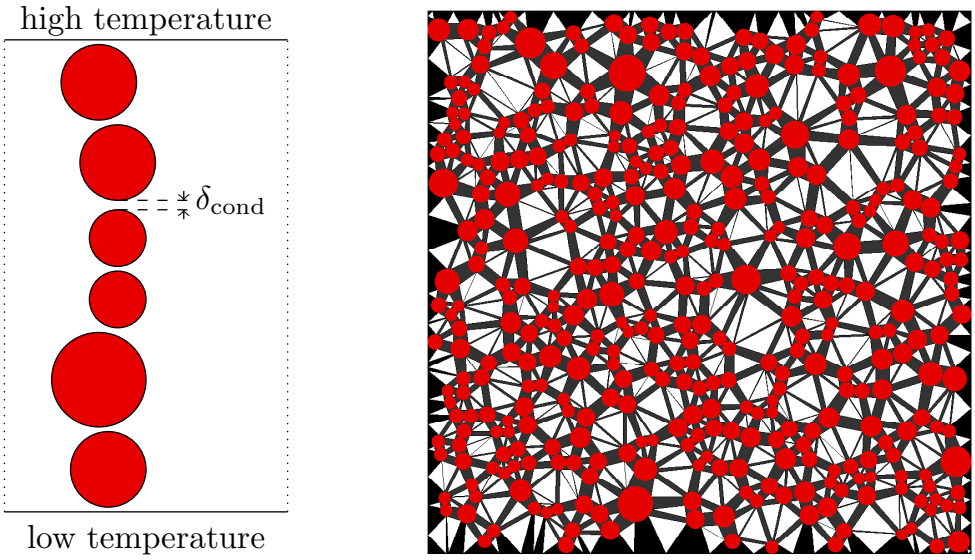
Figure 1.b depicts $\mathcal{T}_{\text{mat}}$ for some set of inclusions (the thick edges between neighboring inclusions are unions of line segments to be explained in Section 2.2; see Figure 2 for a zoom).

Remark 1. The elements of the Delaunay triangulation $\mathcal{T}_{\text{mat}}$ can be characterized locally: Let $x \in \partial\Omega^{\text{mat}}$ be any point on the boundary of Ω^{mat} and ν_x be the corresponding outer normal vector, let A be some closed subset of $\partial\Omega^{\text{mat}}$, and let

$$\Pi(x, A) := \operatorname*{argmin}_{y \in A} \frac{dist\,(x, y)}{\max\{\langle (y - x)/\,dist\,(x, y)\,, \nu_x \rangle, 0\}} \neq \emptyset \tag{7}$$

be the set of points in A which are closest to x in normal direction. Then the cyclic polygon $T_x := \text{conv}(x \cup \Pi(x, \partial\Omega^{\text{mat}})) \in \mathcal{T}_{\text{mat}}$. Moreover, for all $T \in \mathcal{T}_{\text{mat}}$ there is some $x \in \partial\Omega^{\text{mat}}$ such that $T = T_x$.

Since the Delaunay criterion (6) ensures that $int(\text{CD}(T)) \subset \Omega^{\text{mat}}$ for all $T \in \mathcal{T}_{\text{mat}}$, the diameter of T may serve as a local measure of the thickness of the perforated domain Ω^{mat}.



(a) Conducting path of inclusions.

(b) Infinite/Generalized Delaunay triangulation of a set of inclusions contained in the unit square.

FIGURE 1. Geometric aspects of problem (3).

Definition 2.1 (Thickness of a domain)**.** The $\mathcal{T}_{\mathrm{mat}}$-piecewise constant function $\delta : \Omega^{\mathrm{mat}} \to \mathbb{R}_{>0}$, given by

$$\delta|_T := \delta_T := \operatorname{diam}\big(\mathrm{CD}(T)\big) \quad \text{for } T \in \mathcal{T}_{\mathrm{mat}},$$

is denoted as the thickness of Ω^{mat}.

2.2. **A finite subdivision of perforated domains.** Inspired by [2], a finite subdivision of the perforated Ω^{mat} is extracted from the infinite triangulation $\mathcal{T}_{\mathrm{mat}}$ which was introduced in the previous subsection. Without loss of generality let us make the following technical assumption.

Assumption 2.2. An element of $\mathcal{T}_{\mathrm{mat}}$ shall either be a line segment or a triangle. In addition, every pair of triangles shall be separated by at least one line segment.

Remark 2. Assumption 2.2 is not fulfilled in general. The triangulation $\mathcal{T}_{\mathrm{mat}}$ might contain cyclic polygons with more than three vertices. Their appearance is related to the lack of uniqueness of the Delaunay triangulation (into triangles) if the given points are not in general position[2]. However, this degeneracy can be circumvented by subdividing every cyclic polygon with more than three vertices into triangles. The resulting new triangles are not separated by a line segment but share a common edge. This edge can simply be added as an element to the triangulation $\mathcal{T}_{\mathrm{mat}}$.

Let $\mathcal{H} := \{H_1, H_2 \ldots, H_M\}$ be a minimal set of shifted halfspaces that form the outer boundary of Ω, i.e.,

$$\Omega^c := \mathbb{R}^2 \setminus \Omega = \bigcup_{k=1}^{M} H_k.$$

[2]A set of points in the plane is in general position if no four points lie on a common circle.

Since the halfspaces in the set $\mathcal{H}$ can be regarded as disks with infinite radius we define an extended set of inclusions $\tilde{\mathcal{I}} := \mathcal{I} \cup \mathcal{H}$.

A cyclic polygon $T \in \mathcal{T}_{\text{mat}}$ with vertices $x_1, \ldots, x_k \in \partial\Omega^{\text{mat}}$ ($k = 2$ or 3) connects a subset of inclusions $\{I_1, \ldots, I_k\} \subset \tilde{\mathcal{I}}$ if it satisfies $x_j \in I_j$ for all $j = 1, \ldots, k$. For any $T \in \mathcal{T}_{\text{mat}}$ let $\tilde{\mathcal{I}}(T)$ denote the maximal set of inclusions that is connected by T. In this respect, $\tilde{\mathcal{I}}(\cdot)$ can be interpreted as a mapping from $\mathcal{T}_{\text{mat}}$ into the power set of $\tilde{\mathcal{I}}$. The desired finite partition of Ω^{mat} is given by the quotient modulo of this mapping $\tilde{\mathcal{I}}(\cdot)$. It is denoted as the generalized Delaunay partition $\mathcal{D}$ (see [8, 9]) and consists of curvilinear polygons, more precisely

1. (generalized) edges, i.e., channel-like objects (unions of line segments) that connect two neighboring inclusions, and
2. triangles.

According to the classification above, we distinguish between the set of edges $\mathcal{E} \subset \mathcal{D}$ and the set of triangles $\mathcal{T} = \mathcal{D} \setminus \mathcal{E}$.

We emphasize that the generalized Delaunay triangulation serves as a tool in the subsequent regularity analysis. It is a natural way to represent the geometry of particle reinforced composite materials, but it is *not* based on physical grounds.

3. Thickness-weighted regularity.

3.1. Preliminary remarks.

Recall the classical H^2-regularity result on a smooth (C^2) domain $K \subset \mathbb{R}^2$ as it is stated in every textbook on partial differential equations (e.g., [6, Theorem 6.4]): Any $u \in H_0^1(K)$ with $\Delta u \in L^2(K)$ is in $H^2(K)$ and there is a constant C that does not depend on u such that

$$\|\nabla^2 u\|_{L^2(K)} \leq C \|\Delta u\|_{L^2(K)}. \tag{8}$$

This result extends to certain domains with piecewise analytic boundary, especially to the elements of the subdivision $\mathcal{D}$ from Section 2.2. In [1], K is considered to be a curvilinear polygon, i.e., K is a simply-connected, bounded domain with the boundary $\partial K = \bigcup_{k=1}^{m} \Gamma_k$, where Γ_k are analytic simple arcs,

$$\bar{\Gamma}_k = \{\phi_k(\xi) : \xi \in [-1, 1]\}.$$

The functions ϕ_k are analytic on $[-1, 1]$ with $|\nabla \phi_k|$ being bounded away from zero. Under the assumption that all internal angles $\gamma_1, \gamma_2, \ldots, \gamma_m$ of K satisfy $0 < \gamma_k \leq \pi$, there is a constant C_{reg} such that

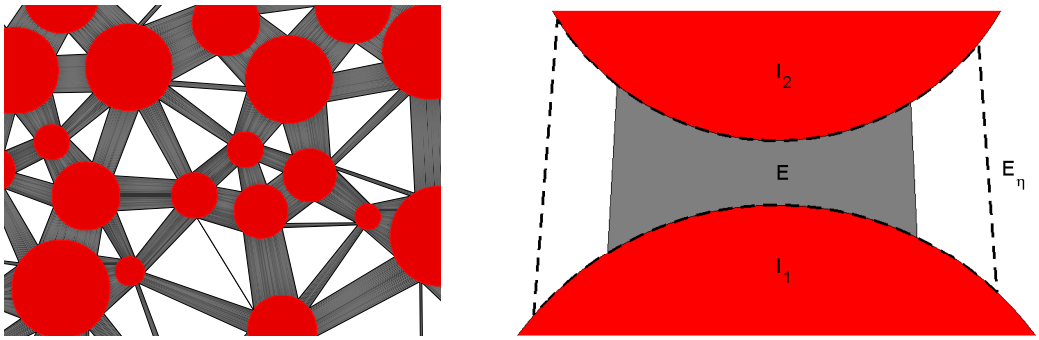
$$\|\nabla^2 u\|_{L^2(K)} \leq C_{\text{reg}} \|\Delta u\|_{L^2(K)} \tag{9}$$

holds for all $u \in H_0^1(K)$ with $\Delta u \in L^2(K)$. Let us stress that the constant C_{reg} does not depend on the scaling of K (see, e.g., [7, Remark 5.5.6]).

3.2. Local regularity.

3.2.1. *Regularity on generalized edges.* Let $E \in \mathcal{E}$, $|E| > 0$, be some generalized edge which connects two inclusions $I_1, I_2 \in \mathcal{I}$. Without loss of generality, let $I_1 = B_{r_1}([0, 0]^T)$ and $I_2 = B_{r_2}([0, d]^T)$, where $B_r(y)$ denotes the closed disk of radius r around y. Let $r_1 \geq r_2$ and $d > r_1 + r_2$. For simplicity, E is supposed to be connected (cf. Remark 3(d) in [8]); otherwise every connected component might be considered on its own.

The subsequent results require a parameterization of the edge E. The restriction of E to I_1, $E \cap \partial I_1$, shall be parameterized by some angle $s \in [\alpha, \beta] \subset]-\pi/2, \pi/2[$, i.e., $E \cap \partial I_1 = \phi([\alpha, \beta])$ with $\phi(s) := r_1(sin(s), cos(s))$. The mapping $\Pi(\cdot, \partial I_2)$



(a) Image section of Figure 1.b .

(b) A generalized edge E (gray shaded) and its neighborhood E_η (area framed by the dashed line).

FIGURE 2. Detailed views of the subdivision defined in Section 2.2.

introduced in (7) maps $E \cap \partial I_1$ onto $E \cap \partial I_2$. Based on ϕ and $\Pi(\cdot, \partial I_2)$, the generalized edge E is parameterized by the diffeomorphism

$$\mathcal{J} :]\alpha, \beta[\times]0, d[\to int(E)\,, \quad \mathcal{J}(s,\lambda) = (1-\lambda)\phi(s) + \lambda\Pi(\phi(s), \partial I_2). \tag{10}$$

For any parameter η, $0 < \eta < \eta_E^{\max} := \min\{|\alpha + \pi/2|, |\beta - \pi/2|\}$, a neighborhood of E is defined by $E_\eta := \mathcal{J}(]\alpha - \eta, \beta + \eta[\times]0, d[)$ (see Figure 2.b for an illustration).

Lemma 3.1. *There exists a constant $C'_E > 0$ which only depends on the ratios r_2/r_1, d/η, and $(\eta_E^{\max} - \eta)^{-1}$ such that for all $u \in H^1(E_\eta)$ with $\Delta u \in L^2(E_\eta)$ and $u|_{\partial(I_1 \cup I_2)} = 0$ it holds $u \in H^2(E)$ and*

$$\|\nabla^2 u\|_{L^2(E)} \le C'_E \left(\|\Delta u\|_{L^2(E_\eta)} + \eta^{-1} \|\nabla u\|_{L^2(E_\eta \setminus E)} \right).$$

Proof. We introduce a smooth cut-off function $\psi_{E,\eta} : E_\eta \to [0,1]$ with the following properties (see also Remark 3 below):

$$\begin{aligned} &(\psi_{E,\eta})|_E = 1, \\ &(\psi_{E,\eta})|_{(\partial E_\eta \setminus \partial(I_1 \cup I_2))} = 0, \text{ and} \\ &\|\nabla^k(\psi_{E,\eta})\|_{L^\infty(E_\eta)} \le C_{\mathrm{co}} \eta^k \text{ for } k \in \mathbb{N} \cup \{0\}. \end{aligned} \tag{11}$$

By construction, the product $u \cdot \psi_{E,\eta}$ vanishes on the boundary of E_η. Hence, the application of (9) and (11) yields

$$\begin{aligned} \|\nabla^2 u\|_{L^2(E)} = \|\nabla^2(u\psi_{E,\eta})\|_{L^2(E)} &\le \|\nabla^2(u\psi_{E,\eta})\|_{L^2(E_\eta)} \\ &\overset{(9)}{\le} C_{\mathrm{reg}} \|\Delta(u\psi_{E,\eta})\|_{L^2(E_\eta)} \\ &\overset{(11)}{\le} C_{\mathrm{co}} C_{\mathrm{reg}} \left(\|\Delta u\|_{L^2(E_\eta)} + 2\eta^{-1} \|\nabla u\|_{L^2(E_\eta \setminus E)} + \eta^{-2} \|u\|_{L^2(E_\eta \setminus E)} \right). \end{aligned} \tag{12}$$

Since u vanishes on $\partial E_\eta \cap \partial(I_1 \cup I_2)$, Friedrichs' inequality allows one to control the L^2 part of the right hand side of (12),

$$\|u\|_{L^2(E_\eta \setminus E)} \le d \|\nabla u\|_{L^2(E_\eta \setminus E)},$$

where $d = dist(I_1, I_2) + r_1 + r_2$ refers to the distance between the centers of I_1 and I_2 as above. Thus the assertion is proved with $C'_E = 2C_{\mathrm{co}} C_{\mathrm{reg}} \left(1 + \frac{d}{\eta}\right)$. □

Remark 3. The constant C_{co} in (11) reflects the size of the inclusions I_1 and I_2 as well as their ratio and, hence, the local uniformity of the distribution of inclusions. It depends on the ratio r_1/r_2 and on $(\eta_E^{\max}-\eta)^{-1}$, where the latter constant becomes large either if the radius r_1 tends to zero or if the ratio $\delta_T/\|\delta\|_{L^\infty(E)}$ becomes large for some adjacent triangle $T \in \mathcal{T}$. However, the dependence on $\delta_T/\|\delta\|_{L^\infty(E)}$ is only an artifact of the way we are cutting Ω^{mat} into pieces and could by avoided (e.g., replace E with some suitable sub edge $\tilde{E} \subset E$ and agglomerate the remaining part $\tilde{E} \setminus E$ and the adjacent triangles).

Lemma 3.1 will be applied to certain subdomains of the edge E (subedges) in order to derive estimates in a thickness weighted norm.

Lemma 3.2. *If $u \in H^1(E_\eta)$ with $\Delta u \in L^2(E_\eta)$ and $u|_{\partial(I_1\cup I_2)} = 0$, then it holds*

(a) $\|\delta\nabla^2 u\|_{L^2(E)} \le 4C'_E(\|\delta\|_{L^\infty(E_\eta)}\|\Delta u\|_{L^2(E_\eta)} + \eta^{-1}\|\delta\nabla u\|_{L^2(E_\eta)})$ *and*

(b) $\|\delta\nabla^2 u\|_{L^2(E)} \le C''_E(\|\delta\Delta u\|_{L^2(E_\eta)} + \|\nabla u\|_{L^2(E_\eta)})$,

where C''_E depends only on the constant C'_E from Lemma 3.1.

Proof. We assume $\alpha < \beta = -\alpha$ for simplicity. Let $0 = s_0 < s_1 < s_2 < \ldots < s_J = \beta$ induce a subdivision of $[0,\beta]$. According to $\{s_j\}_{j=1}^J$ we define subsets $E_1, E_2, \ldots, E_{J+1}$ of E by

$$\begin{aligned} E_1 &:= \mathcal{J}\,(]-s_1, s_1[\times]0,d[),\\ E_j &:= \mathcal{J}\,(]-s_j, s_j[\times]0,d[) \setminus E_{j-1} \text{ for } j = 2,3,\ldots,J, \text{ and}\\ E_{J+1} &:= E_\eta \setminus E. \end{aligned} \tag{13}$$

To prove part (a), the $\{s_j\}_{j=1}^J$ shall be chosen in such a way that

$$\begin{aligned} \delta_0 &:= \min_E \delta \quad \text{and}\\ \delta_j &:= \|\delta\|_{L^\infty(E_j)} = \min\{\|\delta\|_{L^\infty(E)}, 2\delta_{j-1}\} \text{ for } j = 1,2,\ldots,J. \end{aligned} \tag{14}$$

The application of Lemma 3.1 with E replaced by $\tilde{E}_j := \bigcup_{k=1}^j E_k$, $j = 1,2,\ldots,J$, yields

$$\|\nabla^2 u\|_{L^2(E_j)} \le \|\nabla^2 u\|_{L^2(\tilde{E}_j)} \le C'_E\left(\|\Delta u\|_{L^2(E_\eta)} + \eta^{-1}\|\nabla u\|_{L^2(E_\eta\setminus\tilde{E}_j)}\right). \tag{15.j}$$

The summation of (15.j) multiplied by δ_j over $j = 1$ to J leads to

$$\begin{aligned} \|\delta\nabla^2 u\|_{L^2(E)} &\le \sum_{j=1}^J \|\delta\nabla^2 u\|_{L^2(E_j)} \le \sum_{j=1}^J \delta_j\|\nabla^2 u\|_{L^2(E_j)}\\ &\overset{(15.j)}{\le} C'_E\sum_{j=1}^J \delta_j\left(\|\Delta u\|_{L^2(E_\eta)} + \eta^{-1}\|\nabla u\|_{L^2(E_\eta\setminus\tilde{E}_j)}\right)\\ &\le C'_E\left(2\delta_J\|\Delta u\|_{L^2(E_\eta)} + \eta^{-1}\sum_{j=1}^J \|\nabla u\|_{L^2(E_j)}\sum_{k=1}^{j-1}\delta_k\right)\\ &\overset{(14)}{\le} 4C'_E\left(\|\delta\|_{L^\infty(E_\eta)}\|\Delta u\|_{L^2(E_\eta)} + \eta^{-1}\|\delta\nabla u\|_{L^2(E_\eta)}\right). \end{aligned}$$

To prove the estimate (b) we choose $\{s_j\}_{j=1}^J$ in a different way (yielding a different partition of E_η), i.e.,

$$\begin{aligned} s_0 &:= \min_E \delta \quad \text{and}\\ s_j &:= \min\{\beta, 2s_{j-1}\} \text{ for } j = 1,2,\ldots,J. \end{aligned} \tag{16}$$

The actual choice, with regard to the inclusion geometry (convexity of the particles), implies that

$$s_j \geq \tfrac{1}{C_s}\delta_j := \|\delta\|_{L^\infty(E_j)} \quad \text{for all } j = 1, 2, \ldots, J-1. \tag{17}$$

The application of Lemma 3.1 with E replaced by E_1, and E_η replaced by $E_1 \cup E_2$ yields

$$\|\nabla^2 u\|_{L^2(E_1)} \leq C'_E \left(\|\Delta u\|_{L^2(E_1\cup E_2)} + s_1^{-1}\|\nabla u\|_{L^2(E_2)}\right). \tag{18.1}$$

The above estimate easily adapts to the case where E_1 and E_2 are replaced by some E_j and E_{j+1}, $j = 2, 3, \ldots, J$,

$$\|\nabla^2 u\|_{L^2(E_j)} \leq C'_E \left(\|\Delta u\|_{L^2(E_{j-1}\cup E_j\cup E_{j+1})} + s_{j-1}^{-1}\|\nabla u\|_{L^2(E_{j-1}\cup E_{j+1})}\right). \tag{18.j}$$

The summation of (18.j) multiplied by δ_j over $j = 1, \ldots, J$ yields

$$\begin{aligned}
\|\delta\nabla^2 u\|_{L^2(E)} &\leq \sum_{j=1}^{J} \|\delta\nabla^2 u\|_{L^2(E_j)} \leq \sum_{j=1}^{J} \delta_j \|\nabla^2 u\|_{L^2(E_j)} \\
&\overset{(18.j)}{\leq} C'_E \sum_{j=1}^{J} \delta_j \left(\|\Delta u\|_{L^2(E_{j-1}\cup E_j\cup E_{j+1})} + s_{j-1}^{-1}\|\nabla u\|_{L^2(E_{j-1}\cup E_{j+1})}\right) \\
&\overset{(16),(17)}{\leq} (16 + C_s/2) C'_E \left(\|\delta\Delta u\|_{L^2(E_\eta)} + \|\nabla u\|_{L^2(E_\eta)}\right).
\end{aligned}$$

□

Remark 4. So far, the analysis in this subsection has not considered edges that are related to parts of the outer boundary $\partial\Omega$. However, by slightly modified arguments, such cases can be treated as well. We have to distinguish two cases.

1. $E \in \mathcal{E}$ is some generalized edge that connects an inclusion $I \in \mathcal{I}$ and an artificial inclusion $H \in \mathcal{H}$ representing a part of the outer boundary $\partial\Omega$ (see Section 2.2): The previous results apply almost equally, because the boundary part can be regarded as disk with infinite radius.
2. $E \in \mathcal{E}$ connects two parts of the outer boundary $H_1, H_2 \in \mathcal{H}$: It might happen that the environment E_η is not contained in Ω^{mat} (see for instance the generalized edges in the corners in Figure 1.b). However, this issue can be cured by simply replacing E_η with $E_\eta \cap \Omega$ in the upper bounds. Since the solution is given explicitly on $\partial E_\eta \cap \partial\Omega$, Lemma 3.2 can be generalized in a straight forward way.

In general, the solution of (2) does not vanish on the boundary of the inclusions $\mathcal{I}$. We, therefore, need to face inhomogeneous boundary data in the regularity estimate. To this end, consider the affine function $q(s, \lambda) = (1-\lambda)u_1 + \lambda u_2$ on the reference edge $E_{\mathrm{ref}} := [\alpha, \beta] \times [0, d]$ with u_k being the value of $u \in V$ at the inclusion I_k, $k = 1, 2$. The transformation to E defines a function

$$U := q \circ \mathcal{J}^{-1}, \tag{19}$$

which is not affine but has a small Hessian $\nabla^2 U$ in the following sense:

$$\|\delta\nabla^2 U\|_{L^2(E)} \leq C_{\mathcal{J}}\, \eta_E^{-1} \|\delta\nabla U\|_{L^2(E)}. \tag{20}$$

The constant $C_{\mathcal{J}}\, \eta_E^{-1}$ in (20) is related to $\|\mathcal{J}^{-1}\|_{C^2(E_{\mathrm{ref}})}$. Hence, $C_{\mathcal{J}}$ depends on the ratio $\eta_E/(\eta_E^{\max} - \eta_E)$, but *not* on the local thickness δ.

3.2.2. *Interior regularity on triangles.* For some $T \in \mathcal{T}$ and $\theta \geq 0$ we denote a scaled version of T by

$$T_\theta := \{x \in T : \ dist(x, \partial T) \geq \theta\}. \tag{21}$$

We employ a cutoff function $\psi_{T,\theta}$ with

$$\begin{aligned} &(\psi_{T,\theta})|_{T_\theta} = 1, \\ &(\psi_{T,\theta})|_{\partial T} = 0, \text{ and} \\ &\|\nabla^k \psi_{T,\theta}\|_{L^\infty(T)} \leq C_{\mathrm{co}}^{\Delta} \theta^k \text{ for } k \in \mathbb{N} \cup \{0\}, \end{aligned} \tag{22}$$

to conclude that for all $u \in H^1(T)$ with $\Delta u \in L^2(T)$, it holds that $u \in H^2(T_\theta)$, and

$$\begin{aligned} \|\nabla^2 u\|_{L^2(T_\theta)} &\leq \|\nabla^2(u\psi)\|_{L^2(T)} \\ &\overset{(9),(22)}{\leq} C'_T \left(\|\Delta^2 u\|_{L^2(T)} + \theta^{-1} \|\nabla u\|_{L^2(T\setminus T_\theta)} + \theta^{-2} \|u\|_{L^2(T\setminus T_\theta)} \right), \end{aligned} \tag{23.a}$$

where $C'_T = 2C_{\mathrm{co}}^{\Delta} C_{\mathrm{reg}}$. Note that in fact

$$\|\nabla^2 u\|_{L^2(T_\theta)} \leq C_T \left(\|\Delta u\|_{L^2(T)} + \theta^{-1} \|\nabla(u - W)\|_{L^2(T\setminus T_\theta)} + \theta^{-2} \|u - W\|_{L^2(T\setminus T_\theta)} \right)$$

holds with any affine function $W : T \to \mathbb{R}$, because $\nabla^2 W \equiv 0$. Hence, the choice $W = |T|^{-1} \int_T u \, dx$ together with the Poincaré inequality yields

$$\|\nabla^2 u\|_{L^2(T_\theta)} \leq C_T \left(\|\Delta u\|_{L^2(T)} + \theta^{-1} \|\nabla(u)\|_{L^2(T)} \right) \tag{23.b}$$

with a constant C_T that depends only on C'_T and the ratio $\frac{\delta_T}{\theta}$.

3.3. **Global regularity.** We simply sum up the local estimates for the elements of $\mathcal{D} = \mathcal{E} \cup \mathcal{T}$ to derive the global bound. For every edge $E \in \mathcal{E}$ we choose a parameter $\eta = \eta_E$ so that

$$0 < \eta_E < \eta_E^{\max} \quad \text{and} \quad E_\eta \cap \Omega \subset cl\left(\Omega^{\mathrm{mat}}\right). \tag{24}$$

Accordingly, we choose parameters $\theta = \theta_T > 0$ for every triangle $T \in \mathcal{T}$ so that the union of the extended edges and the scaled triangles covers Ω^{mat},

$$\Omega^{\mathrm{mat}} \subset \bigcup_{E \in \mathcal{E}} E_{\eta/2} \quad \cup \quad \bigcup_{T \in \mathcal{T}} T_\theta. \tag{25}$$

Some $\mathcal{D}$-piecewise constant function $\sigma : \Omega^{\mathrm{mat}} \to \mathbb{R}_{>0}$ is given by

$$\begin{aligned} \sigma|_E &= \eta_E \quad \text{for } E \in \mathcal{E} \quad \text{and} \\ \sigma|_T &= \theta_T \quad \text{for } T \in \mathcal{T}. \end{aligned} \tag{26}$$

Remark 5. The results of the present section and beyond will depend locally on some negative powers of the parameter function σ defined in (26). Obviously, there exists a constant $c_{\mathcal{I}}$ such that for all $K \in \mathcal{D}$, $\sigma_K \geq c_{\mathcal{I}} \|\delta\|_{L^\infty(K)}$. Since, in this paper, we focus on the dependence of regularity on the thickness function δ we do not put any effort in the optimization of our subdivision with regard to the constants σ.

For $u \in V$ we denote its $\mathcal{T}_{\mathrm{mat}}$-piecewise affine interpolation by $\mathfrak{I}_{\mathcal{D}} u$. More precisely, $\mathfrak{I}_{\mathcal{D}} u$ is defined by (19) on every edge, and $\mathfrak{I}_{\mathcal{D}} u$ is the unique affine interpolant of u at the vertices of T on every triangle $T \in \mathcal{T}$.

Theorem 3.3. *Let $u \in V$ be the solution of (3) and $U_{\mathcal{D}} := \mathfrak{I}_{\mathcal{D}} u$ its $\mathcal{T}_{\mathrm{mat}}$-piecewise affine interpolation. Then there exists $C_{\mathcal{D}} > 0$, which only depends on the constants of Lemma 3.2 and (23.b), such that*

$$\|\delta \nabla^2 u\|_{L^2(\Omega^{\mathrm{mat}})} \leq C_{\mathcal{D}} \left(\|\delta f\|_{L^2(\Omega^{\mathrm{mat}})} + \|\sigma^{-1} \delta \nabla U_{\mathcal{D}}\|_{L^2(\Omega^{\mathrm{mat}})} \right).$$

Proof. We decompose $u = (u - u^{\mathrm{har}}) + (u^{\mathrm{har}} - U^{\mathrm{har}}) + (U^{\mathrm{har}} - U_{\mathcal{D}}) + U_{\mathcal{D}}$, where $u^{\mathrm{har}} \in H^1(\Omega^{\mathrm{mat}})$ denotes the unique harmonic function with trace $u|_{\partial\Omega^{\mathrm{mat}}}$, and U^{har} the $\mathcal{D}$-piecewise harmonic function which equals $U_{\mathcal{D}}$ on the boundary of every element $K \in \mathcal{D}$. The application of the triangle inequality yields

$$\begin{aligned}\|\delta\nabla^2 u\|_{L^2(\Omega^{\mathrm{mat}})} &\le \|\delta\nabla^2(u - u^{\mathrm{har}})\|_{L^2(\Omega^{\mathrm{mat}})} + \|\delta\nabla^2(u^{\mathrm{har}} - U^{\mathrm{har}})\|_{L^2(\Omega^{\mathrm{mat}})} \\ &\quad + \|\delta\nabla^2(U^{\mathrm{har}} - U_{\mathcal{D}})\|_{L^2(\Omega^{\mathrm{mat}})} + \|\delta\nabla^2 U_{\mathcal{D}}\|_{L^2(\Omega^{\mathrm{mat}})} \\ &=: M_1 + M_2 + M_3 + \|\delta\nabla^2 U_{\mathcal{D}}\|_{L^2(\Omega^{\mathrm{mat}})}.\end{aligned} \tag{27}$$

The estimate

$$\begin{aligned}M_1^2 &\overset{(25)}{\le} \sum_{T\in\mathcal{T}} \|\delta\nabla^2(u - u^{\mathrm{har}})\|^2_{L^2(T_\theta)} + \sum_{E\in\mathcal{E}} \|\delta\nabla^2(u - u^{\mathrm{har}})\|^2_{L^2(E_{\eta/2})} \\ &\overset{(23.\mathrm{b}),\text{Lemma } 3.2.\mathrm{b}}{\le} \sum_{T\in\mathcal{T}} C_T^2 \left(\|\delta f\|_{L^2(T)} + \|\nabla(u - u^{\mathrm{har}})\|_{L^2(T)}\right)^2 \\ &\qquad + \sum_{E\in\mathcal{E}} C_E''^2 \left(\|\delta f\|_{L^2(E_\eta)} + \|\nabla(u - u^{\mathrm{har}})\|_{L^2(E_\eta)}\right)^2 \\ &\le C_1^2 \left(\|\delta f\|_{L^2(\Omega^{\mathrm{mat}})} + \|\nabla(u - u^{\mathrm{har}})\|_{L^2(\Omega^{\mathrm{mat}})}\right)^2\end{aligned} \tag{28}$$

holds with a constant C_1 which depends only on the constants of Lemma 3.2.b and (23). Since $(u - u^{\mathrm{har}}) \in H^1_0(\Omega^{\mathrm{mat}})$, we have from (3.a) and a localized version of the Friedrichs' inequality (see Lemma A.1),

$$\|\nabla(u - u^{\mathrm{har}})\|_{L^2(\Omega^{\mathrm{mat}})} \le C_{\mathrm{F}} \|\delta f\|_{L^2(\Omega^{\mathrm{mat}})}.$$

Since $u^{\mathrm{har}} - U^{\mathrm{har}}$ is locally harmonic, the application of Lemma 3.2.a locally on $E_{\eta/2}$, $E \in \mathcal{E}$ and (23) on T_θ, $T \in \mathcal{T}$, yields

$$M_2 \le C_2' \|\sigma^{-1}\delta\nabla(u^{\mathrm{har}} - U^{\mathrm{har}})\|_{L^2(\Omega^{\mathrm{mat}})},$$

where the constant C_2 depends only on C_E' and C_T. From Lemma A.2, we also get

$$M_2 \le C_2 \|\sigma^{-1}\delta\nabla U_{\mathcal{D}}\|_{L^2(\Omega^{\mathrm{mat}})}. \tag{29}$$

Finally, the application of Lemma 3.2.b on every $E \in \mathcal{E}$, yields

$$M_3^2 \le C_3'^2 \left(\|\delta\Delta U_{\mathcal{D}}\|^2_{L^2(\Omega^{\mathrm{mat}})} + \sum_{E\in\mathcal{E}} \|U^{\mathrm{har}} - U_{\mathcal{D}}\|^2_{L^2(E)} \right)$$

where the constant C_3' depends only on C_E''. The definition of U^{har}, (20), and Lemma A.1 yield

$$M_3 \le C_3 \|\sigma^{-1}\delta\nabla U_{\mathcal{D}}\|_{L^2(\Omega^{\mathrm{mat}})}. \tag{30}$$

The assertion follows readily by combining (27), (28), (29), and (30). □

Lemma 3.4. *Let $u \in V$ be the solution of (3) and $U_{\mathcal{D}} := \mathfrak{I}_{\mathcal{D}} u$ its $\mathcal{T}_{\mathrm{mat}}$-piecewise affine interpolation. Then it holds*

$$\|\delta\nabla U_{\mathcal{D}}\|_{L^2(\Omega^{\mathrm{mat}})} \le C_{\mathfrak{I}} \left(\|f\|_{L^2(\Omega^{\mathrm{mat}})} + \|u_D\|_{L^\infty(\partial\Omega^{\mathrm{mat}})}\right)$$

with some constant $C_{\mathfrak{I}}$ that does not depend on δ.

Proof. By an inverse inequality we get

$$\|\delta\nabla U_{\mathcal{D}}\|_{L^2(\Omega^{\mathrm{mat}})} \le \|U_{\mathcal{D}}\|_{L^2(\Omega^{\mathrm{mat}})}.$$

Moreover,

$$\|U_{\mathcal{D}}\|_{L^2(\Omega^{\mathrm{mat}})} \le C'_{\mathfrak{J}}\|u\|_{L^2(\Omega^{\mathrm{mat}})} \le C'_{\mathfrak{J}}\left(\|u-u^{\mathrm{har}}\|_{L^2(\Omega^{\mathrm{mat}})} + \|u^{\mathrm{har}}\|_{L^2(\Omega^{\mathrm{mat}})}\right) \le C_{\mathfrak{J}}\left(\|f\|_{L^2(\Omega^{\mathrm{mat}})} + \|u^{\mathrm{har}}\|_{L^\infty(\partial\Omega^{\mathrm{mat}})}\right),$$

where we have used the boundedness of the interpolation operator $\mathfrak{I}_{\mathcal{D}}$, the maximum principle for second order elliptic operators (see [6, Theorem 6.4.1]) and a classical L^2 a priori estimate (see [6, Theorem 6.2.6]). □

Theorem 3.5. *Let $u \in V$ be the solution for (3). Then there exists $C_{u_D,f,\sigma} > 0$, which depends only on the data f and u_D, on σ defined in (26), and the constants of Theorem 3.3 and Lemma 3.4, such that*

$$\|\delta\nabla^2 u\|_{L^2(\Omega^{\mathrm{mat}})} \le C_{u_D,f,\sigma}.$$

Proof. The proof follows readily by combining Theorem 3.3 and Lemma 3.4. □

4. **Stable approximation close to percolation.** We now consider any appropriate conforming finite element approximation of (3). Let $V_h \subset V$ be some finite dimensional subspace of V. The corresponding discrete variational problem reads: *Find $u_h \in V_h$ such that*

$$\int_{\Omega^{\mathrm{mat}}} \nabla u_h(x)\nabla v_h(x)\,\mathrm{d}x = \int_{\Omega^{\mathrm{mat}}} f(x)v_h(x)\,\mathrm{d}x \quad \text{for all } v_h \in V_h \cap H^1_0(\Omega^{\mathrm{mat}}), \tag{31.a}$$

$$u_h = u_D \quad \text{on } \partial\Omega. \tag{31.b}$$

It is assumed for simplicity that the Dirichlet data u_D is resolved by V_h, i.e., there is some $v_h \in V_h$ such that $v_h|_{\partial\Omega} = u_D$. The discrete space V_h shall consist of functions that are piecewise smooth with respect to some mesh $\mathcal{G}$ of Ω^{mat}. The mesh $\mathcal{G}$, which consist of possibly curved elements, is supposed be conforming in the sense that $\cup\mathcal{G} = \bar{\Omega}$. Its mesh width is denoted by $h : \Omega^{\mathrm{mat}} \to \mathbb{R}_{>0}$, $h|_K := h_K := \mathrm{diam}(K)$ for all $K \in \mathcal{G}$. Clearly, there holds $h \le C_{\mathcal{G}}\delta$ with some constant $C_{\mathcal{G}}$ which is related to shape regularity of the elements, i.e., the ratio between the radius of the largest ball that can be inscribed in an element and the radius of the smallest ball that contains the element. We assume that the space V_h satisfies approximation properties locally, i.e., there exists some constant C_{appr} so that for all $K \in \mathcal{G}$ and all $u \in H^2(K)$,

$$\inf_{v_h \in V_h}\left(h_K^{-1}\|u-v_h\|_{L^2(K)} + \|\nabla(u-v_h)\|_{L^2(K)}\right) \le C_{\mathrm{appr}}h_K\|\nabla^2 u\|_{L^2(K)}. \tag{32}$$

Theorem 4.1. *If $u \in V$ is the solution for (3), and $u_h \in V_h$ its Galerkin approximation that solves (31), then*

$$\|\nabla(u-u_h)\|_{L^2(\Omega^{\mathrm{mat}})} \le C_{f,u_D,V_h}\|h/\delta\|_{L^\infty(\Omega^{\mathrm{mat}})}$$

holds with $C_{f,u_D,V_h} = C_{\mathrm{appr}}C_{u_D,f,\sigma}$ where C_{appr} is the constant from (32) and $C_{u_D,f,\sigma}$ the one from Theorem 3.5.

Proof. The optimality of the Galerkin method in energy norm together with the approximation properties of the space V_h (cf. (32)) imply that

$$\|\nabla(u-u_h)\|_{L^2(\Omega^{\mathrm{mat}})} \le C_{\mathrm{appr}}\|h\nabla^2 u\|_{L^2(\Omega^{\mathrm{mat}})}. \tag{33}$$

Using the assumption that the ratio h/δ is bounded and applying Theorem 3.5 we further estimate

$$\|h\nabla^2 u\|_{L^2(\Omega^{\mathrm{mat}})} \le \|h/\delta\|_{L^\infty(\Omega^{\mathrm{mat}})}\|\delta\nabla^2 u\|_{L^2(\Omega^{\mathrm{mat}})} \le C_{u_D,f,\sigma}\|h/\delta\|_{L^\infty(\Omega^{\mathrm{mat}})}. \tag{34}$$

The combination of (33) and (34) yields the assertion. □

In practical computations, the assumption of conformity $\cup\mathcal{G} = \bar{\Omega}$ might be relaxed. E.g., the inclusions might by approximated by linear, quadratic, or cubic splines. The resulting geometries are supported by many state-of-the-art mesh generators. However, such a perturbation of the original geometry can only lead to a meaningful approximation if it preserves the distance between neighboring inclusions very precisely.

A special choice of the mesh $\mathcal{G}$ and the corresponding space V_h which preserves conformity is discussed in [10] where

$$\mathcal{G} = \mathcal{D} \text{ and } V_h = V_{\mathcal{D}} := \{v \in C^0(\Omega^{\mathrm{mat}}) : \ v \text{ is } \mathcal{T}_{\mathrm{mat}}\text{-piecewise affine}\}.$$

Corollary 4.2. *If $u \in V$ is the solution for (3) and $u_h \in V_{\mathcal{D}}$ its Galerkin approximation that solves (31), then*

$$\|\nabla(u - u_h)\|_{L^2(\Omega^{\mathrm{mat}})} \leq C_{\mathrm{ip},\mathcal{D}} C_{u_D,f,\sigma},$$

where the constant $C_{\mathrm{ip},\mathcal{D}}$ is related to the approximation property of $V_{\mathcal{D}}$ (see [10, Theorem 3.1, Corollary 3.3]).

Proof. The proof follows readily by combining Theorem 4.1 and the approximation property of the space $V_{\mathcal{D}}$ provided by [10, Theorem 3.1, Corollary 3.3]. □

5. **Conclusion.** In this paper, we have proved that conforming finite element methods yield approximations of the temperature distribution in particle reinforced composite materials that are robust with respect to critical geometric parameters of the packing of particles. More precisely, the absolute error of such an approximation can be bounded by some universal constant that does *not* depend on the geometry of the particle distribution. The relative error scales inversely proportional to the energy of the material. Conforming finite element methods allow one to trace a possible blow-up of the energy as the thickness tends to zero on a path of inclusions that separates the domain. Hence, material simulations based on those methods are able to capture the phase transition from low conductivity to high conductivity (percolation) as the volume fraction of particles is increased.

Moreover, given a fixed sample of the geometry of the material, the regularity theory presented here shows that the use of a conforming finite element mesh with local width proportional to the local thickness of the matrix material guarantees accurate results. Therefore, finite element methods might be used to compute effective properties of a specific sample of the material. These effective properties can then be used as the basis of an numerical upscaling procedure which simulates global material behavior.

The theory presented in this paper can be extended to the case of general smooth inclusions. The same holds true for 3-dimensional setting and for the consideration of general second order elliptic differential operators.

Appendix A. **Inequalities.** We now prove a version of Friedrichs' inequality that is local with respect to the thickness of the domain.

Lemma A.1. *There is some constant C_{F} which does not depend on δ such that for all $v \in H^1_0(\Omega^{\mathrm{mat}})$, it holds that*

$$\|v\|_{L^2(\Omega^{\mathrm{mat}})} \leq C_{\mathrm{F}} \|\delta \nabla v\|_{L^2(\Omega^{\mathrm{mat}})}.$$

Proof. Let $E \in \mathcal{E}$ be some generalized edge and consider subedges E_j, $j = 1, 2, \dots, J_E$ as in (13) and (14). The classical Friedrichs' inequality is applicable (cf. Remark (A.1)) on all subedges E_j. More precisely, there holds

$$\|v\|_{L^2(E_j)} \leq \|\delta\|_{L^\infty(E_j)} \|\nabla v\|_{L^2(E_j)}.$$

Hence, by (14) we get

$$\|v\|_{L^2(E)} \leq 2\|\delta \nabla v\|_{L^2(E)}. \tag{35}$$

On the triangles $T \in \mathcal{T}$ such a result is not directly applicable, because $\partial\Omega^{\mathrm{mat}} \cap \partial T$ is of measure zero. However, the L^2-norm of v on T can be estimated together with the generalized edges $E_1, E_2, E_3 \in \mathcal{E}$ adjacent to T. Let $\tilde{T} := T \cup E_1 \cup E_2 \cup E_3$ be chosen in a way that

$$\min_{x \in \tilde{T} \cap E_k} \delta(x) \geq \tfrac{1}{2}\delta_T \quad \text{for all } k = 1, 2, 3.$$

Then

$$\|v\|_{L^2(\tilde{T})} \leq C_{\mathrm{F}} \frac{|\partial\tilde{T} \cap \partial\Omega^{\mathrm{mat}}|}{|\partial T|} \|\delta \nabla v\|_{L^2(\tilde{T})}. \tag{36}$$

The constant C_{F} does not depend on δ, the ratio $\frac{|\partial\tilde{T} \cap \partial\Omega^{\mathrm{mat}}|}{|\partial T|}$, or on v (see [5]). The assertion follows by simply summing up the local estimates (35) and (36) over all edges $E \in \mathcal{E}$ and all triangles $T \in \mathcal{T}$. ∎

We now present some thickness-weighted energy estimate.

Lemma A.2. *Let $u \in V$ be the solution of (3) and $v \in V$ be any function with trace $v|_{\partial\Omega^{\mathrm{mat}}} = u|_{\partial\Omega^{\mathrm{mat}}}$. Then there holds*

$$\|\delta\nabla(u - v)\|_{L^2(\Omega^{\mathrm{mat}})} \leq C_{\mathrm{cwe}} \left(\|\delta^2 f\|_{L^2(\Omega^{\mathrm{mat}})} + \|\delta \nabla v\|_{L^2(\Omega^{\mathrm{mat}})} \right)$$

with some constant C_{cwe} that does not depend on u, σ, or δ.

Proof. Let $\tilde{\mathcal{D}}$ denote the subdivision of Ω^{mat} which consists of the triangles $T \in \mathcal{T}$ and the subedges $E_1, \dots, E_{J_E}$ of $E \in \mathcal{E}$ as in (13) and (17). Let $\{\phi_K\}_{K \in \tilde{\mathcal{D}}}$ be the partition of unity related to $\tilde{\mathcal{D}}$ such that for all $K \in \tilde{\mathcal{D}}$, $supp(\phi_K)$ is contained in the union of K and its neighboring elements in $\tilde{\mathcal{D}}$, and

$$\|\nabla\phi_K\|_{L^\infty(\Omega^{\mathrm{mat}})} \leq C_{\tilde{\mathcal{D}}} \|\delta\|^{-1}_{L^\infty(K)} =: \delta_K^{-1}, \tag{37}$$

where $C_{\tilde{\mathcal{D}}}$ is some universal constant that does not depend on δ. Then there holds

$$\begin{aligned}
\|\delta\nabla(u - v)\|^2_{L^2(\Omega^{\mathrm{mat}})} &= \int_{\Omega^{\mathrm{mat}}} \delta^2 \nabla(u - v) \nabla \left(\sum_{K \in \tilde{\mathcal{D}}} \phi_K (u - v) \right) \mathrm{d}x \\
&\overset{(3)}{\leq} \sum_{k=1}^{K} \delta_K^2 \left(\int_{supp(\phi_K)} |f(u - v)| \,\mathrm{d}x + \delta_K^2 \int_{supp(\phi_K)} |\nabla v \nabla \left(\phi_K (u - v) \right)| \,\mathrm{d}x \right) \\
&\overset{\text{Lemma A.1,(17),(37)}}{\leq} C \sum_{K \in \tilde{\mathcal{D}}} \left(\|\delta^2 f\|_{L^2(supp(\phi_K))} \|\delta\nabla(u - v)\|_{L^2(supp(\phi_K))} \right. \\
&\qquad \left. + \|\delta\nabla v\|_{L^2(supp(\phi_K))} \|\delta\nabla(u - v)\|_{L^2(supp(\phi_K))} \right).
\end{aligned}$$

For any $\varepsilon > 0$, Young's inequality yields

$$\begin{aligned}
\|\delta\nabla(u - v)\|^2_{L^2(\Omega^{\mathrm{mat}})} \leq C^2 \varepsilon^{-1} \left(\|\delta^2 f\|_{L^2(\Omega^{\mathrm{mat}})} + \|\delta\nabla v\|_{L^2(\Omega^{\mathrm{mat}})} \right) \\
+ 2C^2 \varepsilon \|\delta\nabla(u - v)\|_{L^2(\Omega^{\mathrm{mat}})}.
\end{aligned}$$

Choosing $\varepsilon = (2C)^{-2}$ proves the assertion. □

REFERENCES

[1] I. Babuška and B. Q. Guo, *Regularity of the solution of elliptic problems with piecewise analytic data. II: The trace spaces and application to the boundary value problems with non-homogeneous boundary conditions*, SIAM J. Math. Anal., **20** (1989), 763–781.
[2] L. Berlyand and A. Kolpakov, *Network approximation in the limit of small interparticle distance of the effective properties of a high-contrast random dispersed composite*, Arch. Ration. Mech. Anal., **159** (2001), 179–227.
[3] L. Berlyand and A. Novikov, *Error of the network approximation for densely packed composites with irregular geometry*, SIAM J. Math. Anal., **34** (2002), 385–408.
[4] L. Borcea and G. C. Papanicolaou, *Network approximation for transport properties of high contrast materials*, SIAM J. Appl. Math., **58** (1998), 501–539.
[5] G. A. Chechkin, Yu. O. Koroleva and L.-E. Persson, *On the precise asymptotics of the constant in Friedrich's inequality for functions vanishing on the part of the boundary with microinhomogeneous structure*, J. Inequal. Appl., **2007**, Art. ID 34138, 13 pp.
[6] L. C. Evans, "Partial Differential Equations," 2^{nd} edition, Graduate Studies in Mathematics, **19**, American Mathematical Society, Providence, RI, 2010.
[7] J. M. Melenk, "*hp*-Finite Element Methods for Singular Perturbations," Lecture Notes in Mathematics, **1796**, Springer-Verlag, Berlin, 2002.
[8] D. Peterseim, *Generalized delaunay partitions and composite material modeling*, preprint, DFG Research Center Matheon Berlin, **690** (2010).
[9] D. Peterseim, *Triangulating a system of disks*, in "Proc. 26^{th} European Workshop on Computational Geometry," (2010), 241–244.
[10] D. Peterseim and C. Carstensen, *Finite element network approximation of conductivity in particle composites*, preprint, DFG Research Center Matheon Berlin, **807** (2010).

Received July 2011; revised October 2011.

E-mail address: peterseim@math.hu-berlin.de

D.3 Composite Finite Elements for elliptic interface problems

Mathematics of Computation **83**(290):2657-2674, 2014.

MATHEMATICS OF COMPUTATION
Volume 83, Number 290, November 2014, Pages 2657–2674
S 0025-5718(2014)02815-9
Article electronically published on February 26, 2014

COMPOSITE FINITE ELEMENTS FOR ELLIPTIC INTERFACE PROBLEMS

DANIEL PETERSEIM

Abstract. A Composite Finite Element method approximates linear elliptic boundary value problems with discontinuous diffusion coefficient at possibly high contrast. The discontinuity appears at some interface that is not necessarily resolved by the underlying finite element mesh. The method is nonconforming in the sense that shape functions preserve continuity across the interface in only an approximate way. However, the method allows balancing this non-conformity error and the error of the best approximation in such a way that the total discretization error (in energy norm) decreases linear with regard to the mesh size and independent of contrast.

1. Introduction

This research article considers the design of a *Composite Finite Element* (CFE) method for Dirichlet problems with discontinuous coefficients across an interface. The CFE method is a classical two-scale approach: The degrees of freedom are related to a possibly coarse mesh, whereas the shape of the ansatz functions is defined on a finer subgrid. In other words, finite element shape functions on a coarse scale are composite by shape functions from some finer scale.

In previous CFEs [18, 19, 22], for the treatment of essential boundary conditions on unfitted meshes (with respect to the boundary of the domain), the adaptation of shape was done in such a way that the prescribed boundary condition was fulfilled in an approximate way. Now, in the context of interface problems, finite element shape functions are adapted on a submesh such that the continuity across the interface is preserved in an approximate way. The new CFE approach has three main advantages:

(1) The definition of basis functions is explicit, i.e., no local problems have to be solved.
(2) The coarse mesh does not need to be aligned with the interface, whereas this is necessary for classical finite element methods (see [14]) to converge at an optimal rate. Moreover, the definition of the CFE method does not put any condition on the intersection of mesh cells and the interface.
(3) If the given data (domain, interface, right-hand side, etc.) allow for a (piecewise) smooth solution, the asymptotic order of convergence of the

Received by the editor October 25, 2010 and, in revised form, January 16, 2012 and February 16, 2013.

2010 *Mathematics Subject Classification.* Primary 65N30, 65N12, 35R05, 80M10.

The present paper is a full version of an extended abstract presented at the 81st Annual Meeting of the International Association of Applied Mathematics and Mechanics, Karlsruhe (Germany), 2010. The work was partially supported by the DFG Research Center MATHEON Berlin through project C33.

underlying discretization is preserved on coarse meshes which do not resolve the interface.

Alternative approaches in the literature can be found, for instance, in [24], where another CFE method is introduced, in [1, 9], where the interface condition is imposed weakly via penalization, or in [4], where special basis functions are computed by solving local problems on submeshes.

The present CFE method may be useful for problems with evolving interfaces. Because of evolution, the interface cannot be well represented by edges or faces of a stationary mesh. In classical finite element methods, an adaptation of the mesh to the interface at every time step is required. This adaptation of the mesh in time is considered to be too costly, especially in three space dimensions. The new CFE approach allows the computing of the evolution in time on a fixed (possibly coarse) mesh. It is sufficient to adapt the shape of the ansatz functions (slightly, close to the interface) in time. As we will see later, the cost for this shape adaptation is small when compared with the overall cost of updating the solution on the fixed coarse mesh.

Note finally that our method is designed to efficiently treat the singularity caused by the jump of the diffusion coefficient at the interface. Since the method does not add any degrees of freedom to the coarse finite element space to resolve the interface, it cannot be expected to resolve any singular behavior caused, e.g., by a kink in the interface. The treatment of such singularities has to be organized on top by classical techniques, e.g., by enrichment of the finite element space by certain singular functions or by mesh adaptivity. In the context of adaptivity, CFEs offer a coarse grid approximation that may serve as the initial guess for an a–posteriori–driven adaptive refinement process. They allow the adaptivity toward singularities to start long before the interface is resolved by the underlying finite element mesh.

Notation. In what follows, $\operatorname{dist}(\cdot,\cdot)$ denotes the Euclidean distance in $\mathbb{R}^2$. We use the same notation for the distance between non-empty subsets $A, B \subset \mathbb{R}^2$, $\operatorname{dist}(A,B) := \inf_{x\in A, y\in B} \operatorname{dist}(x,y)$.

The measure $|\cdot|$ is also context-sensitive and refers to the volume of a set relative to its dimension, i.e., $|\cdot|$ denotes the length of a curve, or the area of a domain.

Given some bounded domain Ω, standard notation for (fractional) Sobolev spaces $W_p^m(\Omega)$, $m \geq 0, p \in \mathbb{N}\cup\{\infty\}$, and their corresponding norms $\|\cdot\|_{W_p^m(\Omega)}$ and seminorms $|\cdot|_{W_p^m(\Omega)}$ is used; $H^m(\Omega)$ abbreviates $W_2^m(\Omega)$ ($m \in \mathbb{N}$) and $L^p(\Omega)$ abbreviates $W_p^0(\Omega)$. Given two disjoint bounded Lipschitz domains Ω_1 and Ω_2, the space $H^m(\Omega_1\cup\Omega_2)$ denotes the space of all functions $u \in L^2(\Omega_1\cup\Omega_2)$ with $u|_{\Omega_1} \in H^m(\Omega_1)$ and $u|_{\Omega_2} \in H^m(\Omega_2)$. The dual space of a Hilbert space V is indicated by V^*. The space of $\mathbb{R}$-valued continuous functions on a set Ω is denoted by $C^0(\Omega)$.

2. Composite finite element discretization of a model Poisson problem

2.1. **Model problem.** Consider Poisson's equation $-\operatorname{div}(a\nabla u) = f$ in an open, bounded, polyhedral domain $\Omega \subset \mathbb{R}^d$, $d \in \{2,3\}$, with homogeneous Dirichlet boundary conditions on $\partial\Omega$. The scalar coefficient a (permeability or conductivity) jumps across an interface $\Gamma := \bar\Omega_1 \cap \bar\Omega_2$ that separates two disjoint, open Lipschitz subdomains $\Omega_1, \Omega_2 \subset \Omega$, $\bar\Omega = \bar\Omega_1 \cup \bar\Omega_2$. The corresponding variational problem

reads: Find $u^\star \in H^1_0(\Omega)$ such that

$$\int_\Omega a\nabla u^\star \cdot \nabla v \,\mathrm{d}x = \int_\Omega f v \,\mathrm{d}x \quad \text{for all } v \in H^1_0(\Omega). \tag{2.1}$$

For simplicity, the coefficient $a : \Omega \to \mathbb{R}_{>0}$ is chosen piecewise constant,

$$a(x) = \begin{cases} 1 & \text{if } x \in \Omega_1, \\ a_{\mathrm{cont}} > 1 & \text{if } x \in \Omega_2. \end{cases}$$

The parameter a_{cont} represents the contrast which is supposed to be large in practical applications, e.g., in the modeling of heat transfer in composite materials.

The bounded bilinear form $\mathfrak{a} : H^1_0(\Omega) \times H^1_0(\Omega) \to \mathbb{R}$ given by

$$\mathfrak{a}(u,v) := \int_\Omega a\nabla u \cdot \nabla v \,\mathrm{d}x = \int_{\Omega_1} \nabla u \cdot \nabla v \,\mathrm{d}x \;+\; a_{\mathrm{cont}} \int_{\Omega_2} \nabla u \cdot \nabla v \,\mathrm{d}x$$

for $u, v \in H^1_0(\Omega)$ induces the norm $|||\cdot||| := \|\sqrt{\mathrm{a}}\nabla \cdot \|_{\mathrm{L}^2(\Omega)}$ in $H^1_0(\Omega)$, the so-called energy norm. Hence, problem (2.1) has a unique solution for all $f \in H^{-1}(\Omega) := (H^1_0(\Omega))^*$.

Usually, some finite-dimensional subspace $V_h \subset H^1_0(\Omega)$ based on piecewise polynomials replaces $H^1_0(\Omega)$ in a finite element discretization of (2.1). However, if the underlying finite element mesh is not aligned with the interface, this ansatz suffers from the lack of regularity of the solution at the interface; the solution is continuous across, but its gradient may jump.

In this paper, this issue shall be fixed by considering a discrete space V_h that violates conformity, $V_h \not\subset H^1_0(\Omega)$. We consider shape functions that are conforming with respect to each of the subdomains but possibly discontinuous across the interface, i.e.,

$$V_h \subset H^1_0(\Omega_1 \cup \Omega_2) := \{u \in H^1(\Omega_1 \cup \Omega_2) :\; u|_{\partial\Omega} = 0\}.$$

Because of the lack of Galerkin orthogonality, the discretization error of a corresponding method is not necessarily proportional to the error of the best approximation of the solution. The discretization error is bounded by the sum of the best approximation error and the error related to the violation of conformity as in (3.1). The aim of this paper is to construct a non-conforming discrete space V_h (based on piecewise affine ansatz functions) such that a balance is achieved between the errors due to non-conformity and errors due to best approximation. This balance yields linear convergence of the corresponding method with respect to the mesh size parameter h without resolving the interface by degrees of freedom.

2.2. **Construction of the finite element space.** The construction in Subsections 2.2.1–2.2.3 below follows the methodology of CFEs [8].

2.2.1. *Triangulations.* Let $\mathcal{T}$ be some regular subdivision of $\bar\Omega$ into closed nonempty simplices (or triangulation for short) according to Ciarlet [3, 5]. Two nondisjoint distinct simplices in $\mathcal{T}$ share either a common face ($d = 3$), a common edge, or a common vertex. By $V(T)$ we denote the set of vertices (corners) of a simplex $T \in \mathcal{T}$. The union of vertices in a (sub)triangulation $\mathcal{T}$ is denoted by $V(\mathcal{T}) := \bigcup_{T\in\mathcal{T}} V(T)$. The $\mathcal{T}$-piecewise mesh width function $h : \bar\Omega \to \mathbb{R}_{>0}$ is given by

$$h(x) := \max_{T\in\mathcal{T}:x\in T} \operatorname{diam}(T).$$

Note that the coarse triangulation $\mathcal{T}$ does not necessarily match the interface Γ, i.e., Γ is not the union of element edges or faces. Later on, the degrees of freedom of the CFE space will be exclusively assigned to the vertices of the (coarse) triangulation $\mathcal{T}$.

We consider the two triangulations $\mathcal{T}_1, \mathcal{T}_2 \subset \mathcal{T}$,

$$\mathcal{T}_k := \{T \in \mathcal{T} : \ T \subset \bar{\Omega}_k\}, \ k = 1, 2,$$

related to the subdomains. The union of these triangulations does not cover Ω, in general. Some neighborhood of the interface, the interface zone

$$\Omega^\Gamma := \Omega \setminus ((\bigcup \mathcal{T}_1) \cup (\bigcup \mathcal{T}_2)),$$

is not covered by elements of $\mathcal{T}_1$ or $\mathcal{T}_2$ unless the interface is resolved by $\mathcal{T}$. We introduce two triangulations of the interface zone, one associated with each subdomain. The elements $T \in \mathcal{T}$ that are contained in none of the two triangulations are collected in the set

$$\mathcal{T}_2^\Gamma := \mathcal{T} \setminus (\mathcal{T}_1 \cup \mathcal{T}_2).$$

A further fine triangulation $\mathcal{T}_1^\Gamma$ of Ω^Γ will be employed to adapt the shape of the ansatz functions in Ω_1. This fine triangulation $\mathcal{T}_1^\Gamma$ is derived by regular refinement of $\mathcal{T}_2^\Gamma$ (e.g., by red-green-refinement or newest vertex bisection) locally near the interface. The corresponding $\mathcal{T}_1^\Gamma$-piecewise mesh width function $h_1^\Gamma : \bigcup \mathcal{T}_1^\Gamma \to \mathbb{R}_{>0}$ is given by

$$h_1^\Gamma(x) := \max_{t \in \mathcal{T}_1^\Gamma : x \in t} \operatorname{diam}(t).$$

The refinement shall be done such that

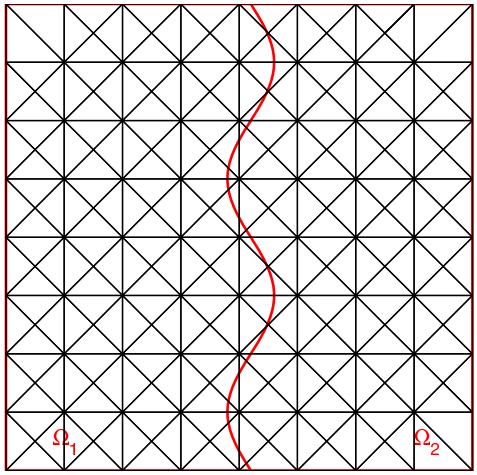
$$h_1^\Gamma|_t = \operatorname{diam}(t) \geq C_1^{-1} \operatorname{dist}(t, \Gamma) \quad \text{for all } t \in \mathcal{T}_1^\Gamma \tag{2.2}$$

holds with a universal constant C_1 independent of the h_1^Γ. This condition prevents over-refinement in the interface zone and enforces a certain grading of $\mathcal{T}_1^\Gamma$ toward the interface. This grading is essential for the stability, complexity, and accuracy of our method. The condition enters our error analysis via the external result [18, Theorem 4.4] which plays an essential role in the proof of Lemma 3.1 and, hence, in the proof of our main result Theorem 2.3.

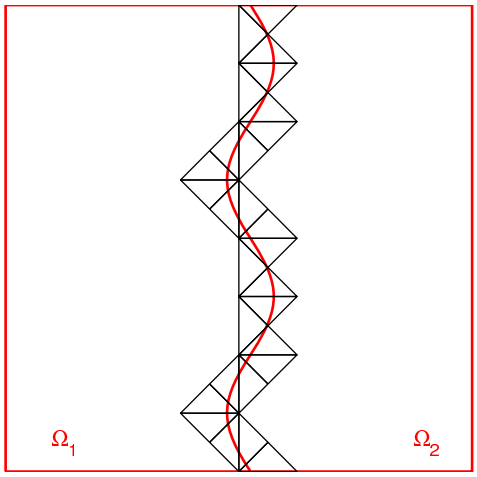
Note that condition (2.2) is satisfied with a constant $C_1 \approx 2$ if the fine triangulation $\mathcal{T}_1^\Gamma$ is computed by successive refinement of those simplices that are intersected by Γ (cf. [22, Section 2]). This shows that arbitrary small elements in a vicinity of the interface are possible in $\mathcal{T}_1^\Gamma$. Still, $\mathcal{T}_1^\Gamma$ is not aligned with Γ in general. The analysis of Section 3 will show that the mesh size $h^\Gamma|_\Gamma$ at the interface suffices to be of size $h^{3/2}$. This implies that the complexity of $\mathcal{T}_1^\Gamma$ depends only on the mesh size of the coarse mesh $\mathcal{T}$ and not on the location of the interface relative to the coarse mesh.

2.2.2. *Additional structure.* The meshes defined in the previous section cannot see the interface. However, precise information about the location of the interface is crucial for any reasonable approximation scheme. The exchange of information between the interface and the meshes shall be introduced via two mappings.

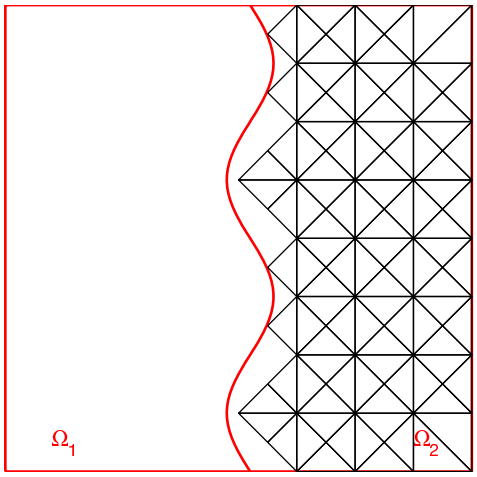
Closest inner simplex. The mapping $T^1_{(\cdot)} : V(\mathcal{T}_1^\Gamma) \to \mathcal{T}_1$ is chosen such that $T^1_x \in \operatorname{argmin}_{T \in \mathcal{T}_1} \operatorname{dist}(x, T)$, i.e., $T^1_{(\cdot)}$ assigns a closest inner simplex (fully contained in Ω_1) to every vertex $x \in V(\mathcal{T}_1^\Gamma)$. $\mathcal{I}_{T^1_x} u \in \mathbb{P}_1(\mathbb{R}^2)$ denotes the globally affine



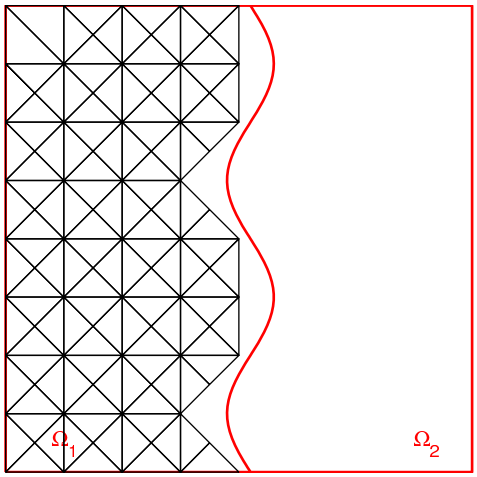
(a) The coarse triangulation $\mathcal{T}$.



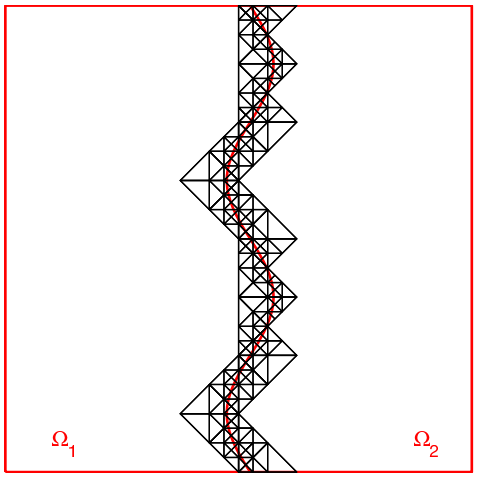
(b) The triangulation $\mathcal{T}_2^\Gamma$ of Ω^Γ.



(c) The triangulation $\mathcal{T}_2$ related to Ω_2.



(d) The triangulation $\mathcal{T}_1$ related to Ω_1.



(e) The refined triangulation $\mathcal{T}_1^\Gamma$ of the interface zone Ω^Γ.

FIGURE 1. The triangulations introduced in Section 2.2.1. The interface is not well represented in $\mathcal{T}$ shown in (a). It is better represented by $\mathcal{T}_1^\Gamma$ in (e) but still not resolved.

function which interpolates u in the vertices of T_x^1. Accordingly, $T^2_{(\cdot)} : V(\mathcal{T}_2^\Gamma) \to \mathcal{T}_2$ and $\mathcal{I}_{T_x^2} u$ are defined.

Interface projection. The projection operator $(\cdot)^\Gamma : \mathbb{R}^d \to \Gamma$ is chosen such that $x^\Gamma \in \operatorname{argmin}_{y\in\Gamma} \operatorname{dist}(x,y)$. This projection encodes the geometrical information about the interface that is required by our method.

2.2.3. *The CFE space.* By S_k, $k = 1, 2$, we denote the finite element space of continuous $\mathcal{T}_k$-piecewise affine functions

$$S_k := \{u : C^0(\textstyle\bigcup \mathcal{T}_k) \ : \ u|_T \in \mathbb{P}^1 \text{ for all } T \in \mathcal{T}_k,\ u|_{\partial\Omega\cap\partial(\bigcup \mathcal{T}_k)}\} \tag{2.3}$$

with the homogeneous Dirichlet boundary condition on $\partial\Omega$ built in. These spaces represent the degrees of freedom of the method, in that CFE shape functions are derived by extending elements from S_1 resp. S_2 to the interface zone.

In other words, CFE shape functions are certain elements of the target space

$$\begin{aligned} S^\Gamma := \{u \in H^1_0(\Omega_1 \cap \Omega_2) \ : \ & u|_{T\cap\Omega_1} \in \mathbb{P}^1 \text{ for all } T \in \mathcal{T}_1 \cup \mathcal{T}_1^\Gamma, \\ & u|_{T\cap\Omega_2} \in \mathbb{P}^1 \text{ for all } T \in \mathcal{T}_2 \cup \mathcal{T}_2^\Gamma\}. \end{aligned} \tag{2.4}$$

Definition of shape functions via extensions. The CFE space S^{cfe} is given as the image of $S_1 \times S_2$ under the bounded linear injective operator $\mathcal{P}^{\mathrm{cfe}} : S_1 \times S_2 \to S^\Gamma$, i.e., $S^{\mathrm{cfe}} := \mathcal{P}^{\mathrm{cfe}}(S)$. The definition of $\mathcal{P}^{\mathrm{cfe}}$ is based on two mappings that relate the different meshes and the interface. The projection operator $\mathcal{P}^{\mathrm{cfe}}$ is defined in the two subdomains as follows:

$$\mathcal{P}^{\mathrm{cfe}}(u_1,u_2)(x) \ := \ \begin{cases} \mathcal{P}_1^{\mathrm{cfe}}(u_1,u_2)(x) & \text{if } x \in \Omega_1, \\ \mathcal{P}_2^{\mathrm{cfe}} u_2(x) & \text{if } x \in \Omega_2 \end{cases} \tag{2.5}$$

with $\mathcal{P}_1^{\mathrm{cfe}}$ and $\mathcal{P}_2^{\mathrm{cfe}}$ given subsequently.

Definition of $\mathcal{P}^{\mathrm{cfe}_2}$. The operator $\mathcal{P}_2^{\mathrm{cfe}}$ extends functions defined in $\bigcup \mathcal{T}_2$ to the interface zone $\Omega\Gamma$. Given $u_2 \in S_2$, the continuous $(\mathcal{T}_2 \cup \mathcal{T}_2^\Gamma)$-piecewise affine function $\mathcal{P}_2^{\mathrm{cfe}} u_2$ is uniquely defined by nodal values

$$(\mathcal{P}_2^{\mathrm{cfe}} u_2)(x) := \begin{cases} u_2(x) & \text{if } x \in V(\mathcal{T}_2), \\ \mathcal{I}_{T_x^2} u(x) & \text{if } x \in V(\mathcal{T}_2^\Gamma) \setminus V(\mathcal{T}_2) \text{ and } x \notin \partial\Omega^\Gamma \cap \partial\Omega, \\ 0 & x \in \partial\Omega^\Gamma \cap \partial\Omega, \end{cases} \tag{2.6}$$

with $\mathcal{I}_{T_x^2} u$ defined in Section 2.2.2 above.

Definition of $\mathcal{P}^{\mathrm{cfe}}{}_1$. The operator $\mathcal{P}_1^{\mathrm{cfe}}$ extends functions defined in $\bigcup \mathcal{T}_1$ to the interface zone Ω^Γ in such a way that its trace on Γ approximately coincides with $(\mathcal{P}_2^{\mathrm{cfe}} u_2)|_\Gamma$. Given $u_1 \in S_1$, $u_2 \in S_2$, $\mathcal{P}_1^{\mathrm{cfe}}(u_1,u_2)$ is the unique continuous $(\mathcal{T}_1 \cup \mathcal{T}_1^\Gamma)$-piecewise affine function which takes the following values at vertices $x \in V(\mathcal{T}_1 \cup \mathcal{T}_1^\Gamma)$:

$$\begin{aligned} &\mathcal{P}_1^{\mathrm{cfe}}(u_1,u_2)(x) \\ &\quad := \begin{cases} u_1(x) & \text{if } x \in V(\mathcal{T}_1) \\ & \quad \text{or } x \in V(\mathcal{T}_1^\Gamma) \cap (\bigcup \mathcal{T}_1), \\ \mathcal{P}_2^{\mathrm{cfe}} u_2)(x^\Gamma) + \left\langle \nabla \mathcal{I}_{T_x^1} u_1, x - x^\Gamma \right\rangle & \text{if } x \in V(\mathcal{T}_1^\Gamma) \setminus (\bigcup \mathcal{T}_1) \\ & \quad \text{and } x \notin \partial\Omega^\Gamma \cap \partial\Omega, \\ 0 & \text{if } x \in \partial\Omega^\Gamma \cap \partial\Omega, \end{cases} \end{aligned} \tag{2.7}$$

with $\mathcal{I}_{T_x^1} u$ defined in Section 2.2.2 above. Note that the definition of $\mathcal{P}_1^{\mathrm{cfe}}$ ensures continuity of its images although $\mathcal{T}_1 \cap \mathcal{T}_1^\Gamma$ is not necessarily a regular triangulation in the sense that hanging nodes may appear (as in Figure 1(d)–(e)).

Although the one-dimensional case (Ω is an interval and Γ is some point in Ω) does not share the numerical difficulties of the multi-dimensional setting (because the interface can easily be resolved by adding the vertex Γ to any mesh), it clearly illustrates the definition of $\mathcal{P}^{\mathrm{cfe}}$ and the derivation of our shape functions (see Figures 2–3). Note that, in one dimension, our construction ensures continuity across the interface and the method is conforming. In general, conformity is only achieved in the limit $h_1^\Gamma|_\Gamma \to 0$. However, the discontinuity of shape functions across the interface (see Figure 4) is sufficiently small to preserve stability and accuracy of our method.

Remark 2.1. There is some algorithmic freedom in the above construction:

(1) It is not essential that the subtriangulations $\mathcal{T}_1, \mathcal{T}_2, \mathcal{T}_2^\Gamma$ form a partition of some regular triangulation $\mathcal{T}$. They could have been chosen to be non-matching overlapping triangulations representing Ω_1, Ω_2, and some neighborhood Ω^Γ of the interface Γ.
(2) It is not essential that the definitions of the mappings $T^1_{(\cdot)}$ and $(\cdot)^\Gamma$ in the above construction are based on the minimality of certain distances; any point or simplex sufficiently close (distance proportional to local mesh size) would do the job as well.

2.2.4. *A local basis of the CFE space.* The degrees of freedom of the method are function values at vertices

$$V_{\mathrm{dof}}(\mathcal{T}) := V(\mathcal{T}_1) \cup V(\mathcal{T}_2) \subset V(\mathcal{T}).$$

Hence, degrees of freedom are solely assigned to vertices in the coarse (interface independent) mesh $\mathcal{T}$ and every vertex in $\mathcal{T}$ represents at most one basis function of S^{cfe}.

The images of the nodal basis functions $\lambda_z \in S_1 \cup S_2$ for $z \in V_{\mathrm{dof}}(\mathcal{T})$ yield a basis of S^{cfe}, i.e.,

$$S^{\mathrm{cfe}} = \operatorname{span}\left(\{\mathcal{P}^{\mathrm{cfe}}(\lambda_z, 0) \mid z \in V(\mathcal{T}_1)\} \cup \{\mathcal{P}^{\mathrm{cfe}}(0, \lambda_z) \mid z \in V(\mathcal{T}_2)\}\right),$$

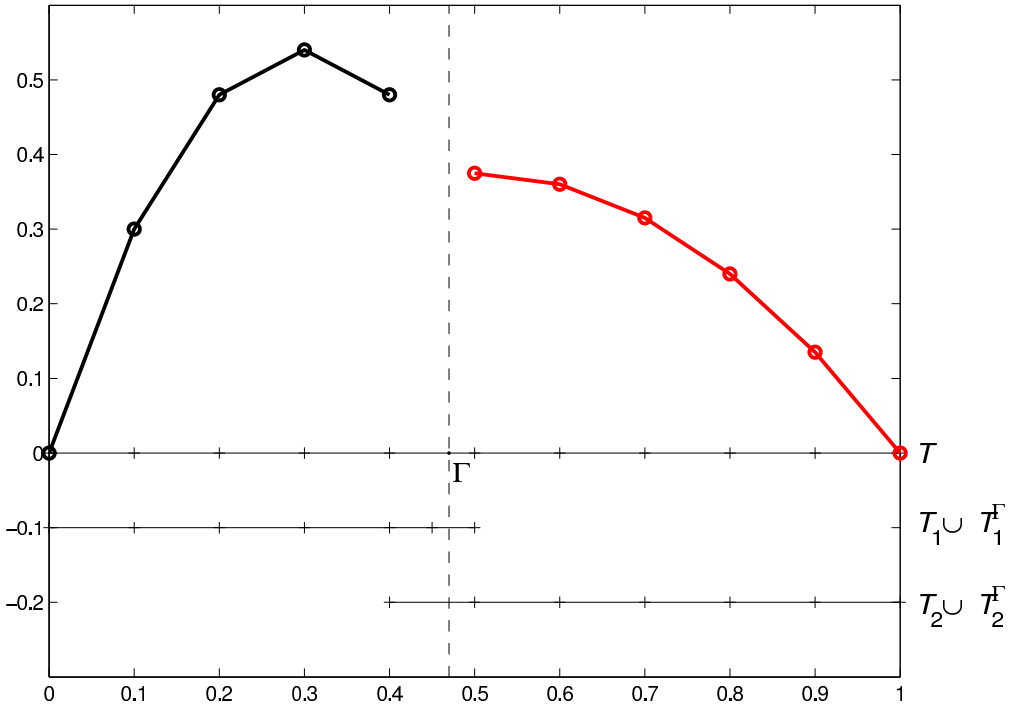
where $\mathcal{P}^{\mathrm{cfe}}\lambda_z$ and $\mathcal{P}^{\mathrm{cfe}}\lambda_y$ are linearly independent if $z \neq y$.

Most of the basis functions are standard nodal basis functions. More precisely, $\mathcal{P}^{\mathrm{cfe}}$ has no effect on functions that vanish in Ω^Γ plus one layer of coarse elements $T \in \mathcal{T}$. Only a few basis functions are manipulated via the explicit linear operator $\mathcal{P}^{\mathrm{cfe}}$. Those basis functions have slightly enlarged supports when compared with standard nodal basis functions on $\mathcal{T}$. However, the supports remain local in the sense that their diameters remain proportional to the local coarse mesh size h. Thus, the supports have finite overlap independent of the mesh size h.

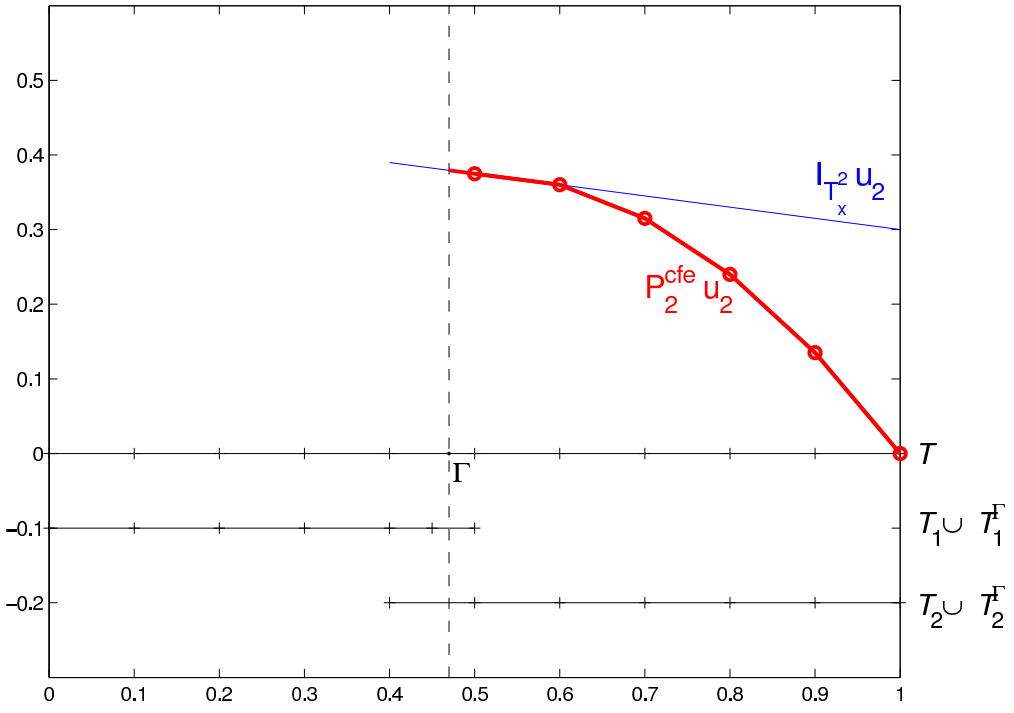
2.3. **Discrete problem.** The discrete variational formulation of (2.1) reads: Find $u^{\mathrm{cfe}} \in S^{\mathrm{cfe}}$ such that

$$\int_\Omega a \nabla u^{\mathrm{cfe}} \cdot \nabla v \,\mathrm{d}x = \int_\Omega f v \,\mathrm{d}x \quad \text{for all } v \in S^{\mathrm{cfe}}. \tag{2.8}$$

Note that the basis given in the previous section turns this variational problem into a system of linear algebraic equations. Since those basis functions have local support, sparsity of the corresponding stiffness matrix is ensured.

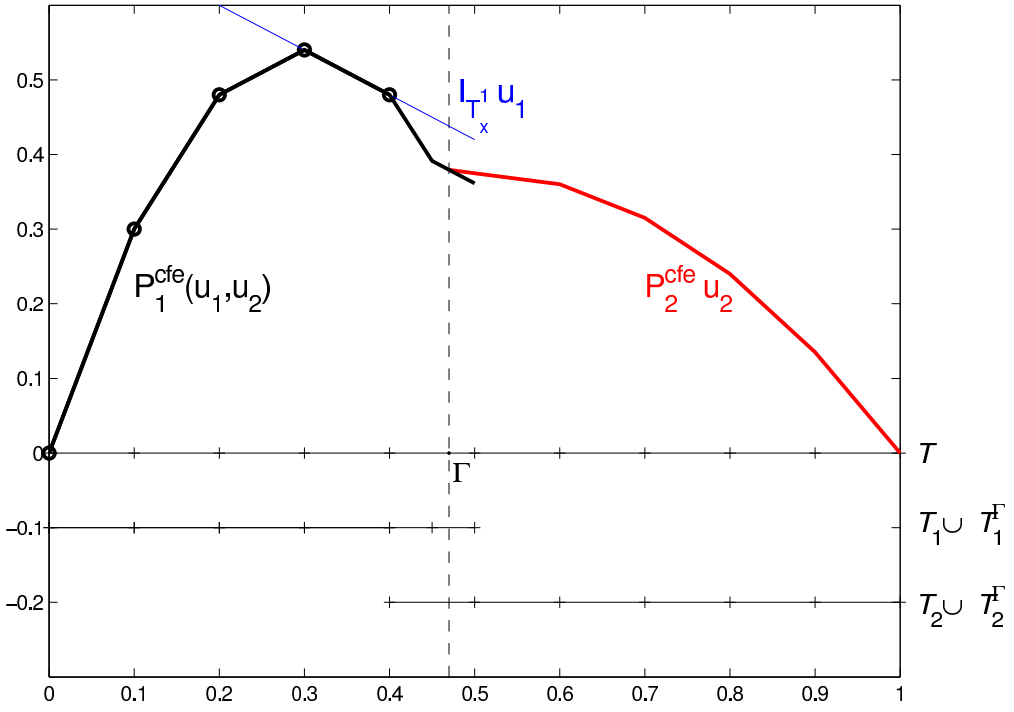


(a) Some functions $u_1 \in S_1$ (left) and $u_2 \in S_2$ (right) representing the degrees of freedom.

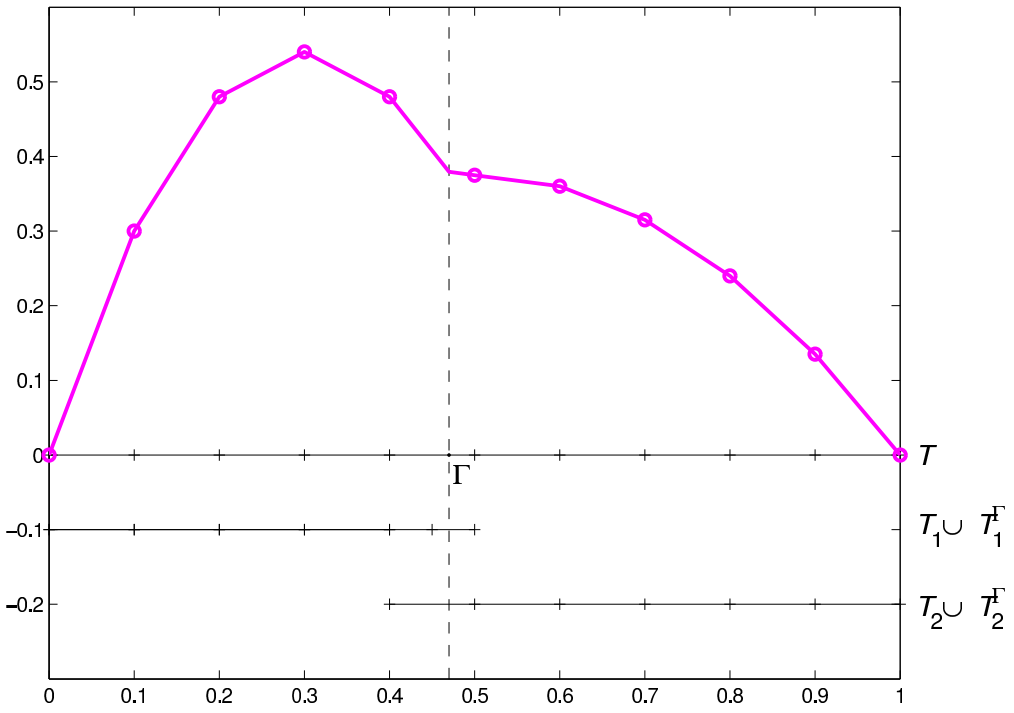


(b) Extension $\mathcal{P}_2^{\mathrm{cfe}} u_2$ of u_2 to the interface zone Ω^Γ.

FIGURE 2. Illustration of the definition of the CFE space in Section 2.2.3.

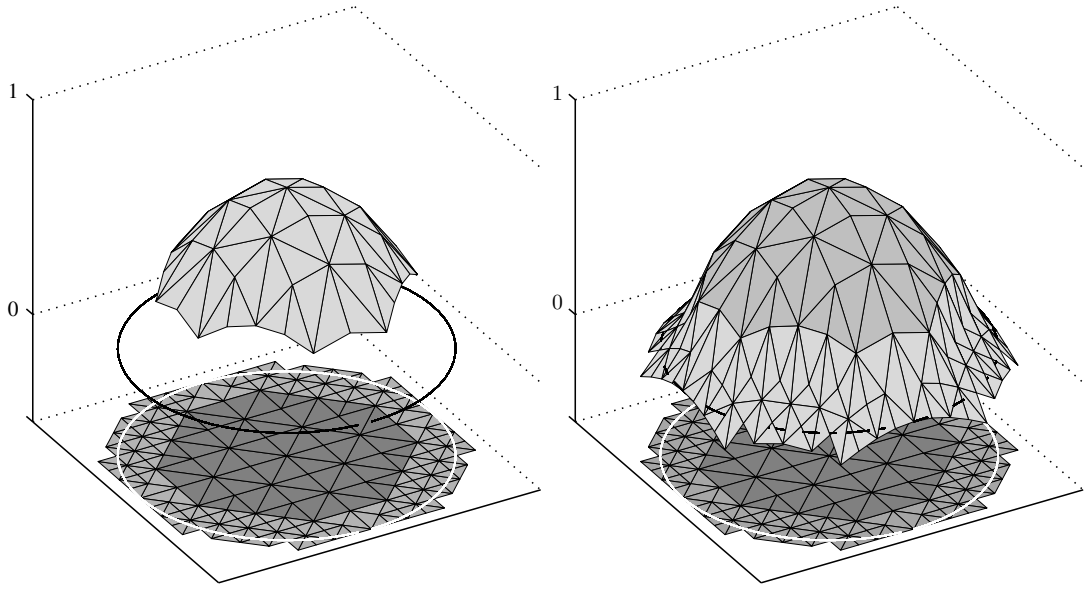


(a) Extension $\mathcal{P}_1^{\text{cfe}} u_1$ of u_1 to the interface zone Ω^Γ.



(b) The corresponding CFE function $\mathcal{P}^{\text{cfe}}(u_1, u_2) \in S^{\text{cfe}}$ with degrees of freedom (○).

FIGURE 3. Illustration (continued from Figure 2) of the definition of the CFE space in Section 2.2.3.



(a) Some function $u_1 \in S_1$ defined on $\bigcup \mathcal{T}_1$. (b) Extension $\mathcal{P}_1^{\mathrm{cfe}}(u_1, 0)$ of u_1.

FIGURE 4. Illustration of the approximate trace matching across the interface: Γ is the unit circle and Ω_1 its interior, u_2 and, hence, $\mathcal{P}_2^{\mathrm{cfe}} u_2|_\Gamma$ are chosen zero.

Remark 2.2. The implementation of the method is similar to previous CFE methods and we refer to [7, 16, 21, 22] for computational insights.

One issue is that the solution of (2.8) requires the evaluation of integrals over intersections $T \cap \Omega_k$ which is beyond the scope of this note. The forthcoming theoretical results assume that all integrals are evaluated exactly. We refer to [7, 15, 22] for a practical resolution of this issue.

2.4. **Error estimates.** The following theorem addresses the solvability of (2.8). Moreover, assuming $H^2(\Omega_1 \cup \Omega_2)$-regularity, an optimal a priori error bound in energy norm is given. Besides parameters already mentioned in the construction, the constant in the error estimate depends on ρ_T, which is the ratio between the diameter of the largest ball that can be inscribed in $T \in \mathcal{T}$ and $\operatorname{diam}(T)$. The triangulations $\mathcal{T}$ and $\mathcal{T}_1^\Gamma$ are assumed to be non-degenerate, i.e., $\rho_{\mathcal{T}} := \min_{T\in\mathcal{T}} \rho_T > 0$ (resp. $\rho_{\mathcal{T}_1^\Gamma} := \min_{T\in\mathcal{T}_1^\Gamma} \rho_T > 0$).

Theorem 2.3 (Linear convergence with respect to mesh size)**.** *The discrete problem* (2.8) *always has a unique solution* $u^{\mathrm{cfe}} \in S^{\mathrm{cfe}}$.

If, in addition, the solution of (2.1)*,* $u^\star \in H_0^1(\Omega)$*, is piecewise smooth,* $u^\star \in H^2(\Omega_1 \cup \Omega_2)$*, and if*

$$\|h_1^\Gamma / h^{3/2}\|_{L^\infty(\bigcup\{t\in\mathcal{T}_1^\Gamma : t\cap\Gamma\neq\emptyset\})} \leq C_2 \tag{2.9}$$

for some generic constant C_2*, then the following a priori error estimate holds:*

$$|||u^\star - u^{\mathrm{cfe}}||| \leq C\|h\|_{L^\infty(\Omega)} \left\|\sqrt{a}\nabla^2 u\right\|_{L^2(\Omega_1\cup\Omega_2)}. \tag{2.10}$$

The constant $C = C(\rho_{\mathcal{T}}, \rho_{\mathcal{T}_1^\Gamma}, C_1, C_2)$ *does not depend on the mesh width functions* h, h^Γ *and the contrast parameter* a_{cont}.

Proof. The unique solvability of (2.8) follows from the fact that $|||\cdot|||$ is a norm in S^{cfe}. Since, in the limit $h_1^\Gamma|_\Gamma \to 0$, S^{cfe} is conforming, the latter is quite obvious if $h_1^\Gamma|_{(\bigcup\{t\in\mathcal{T}_1^\Gamma : t\cap\Gamma\neq\emptyset\})}$ is sufficiently small. Otherwise, this property can be proven along the lines of [18, Lemma 4.10].

The proof of the error estimate will be given in Section 3. □

The error estimate in the above theorem rests on the regularity of the solution $u^\star \in H^2(\Omega_1\cup\Omega_2)$. In general, this regularity does not hold for solutions of problem (2.1). Moreover, even though the constant in the error estimate does not depend on the contrast a_{cont}, the $H^2(\Omega_1\cup\Omega_2)$ seminorm of the solution on the right-hand side of estimate (2.10) may do. In Section 4 we will prove that $f \in L^2(\Omega)$, the Lipschitz properties of the subdomains Ω_k, and, in addition, convexity of $\Omega \subset \mathbb{R}^2$ and the assumption $\Gamma \in C^{1,1}$ imply $u^\star \in H^2(\Omega_1\cup\Omega_2)$ and

$$\left\|\nabla^2 u^\star\right\|_{L^2(\Omega_1)} \le C_{\mathrm{reg}}\,\|f\|_{L^2(\Omega)}\,, \quad \left\|\nabla^2 u^\star\right\|_{L^2(\Omega_2)} \le \frac{C_{\mathrm{reg}}}{a_{\mathrm{cont}}}\,\|f\|_{L^2(\Omega)} \tag{2.11}$$

with some universal constant C_{reg} that depends only on the geometry of the subdomains Ω_k and the interface Γ but not on f and a_{cont}. Hence, under those assumptions on the geometry, the error of the CFE method does not depend on the contrast parameter a_{cont}.

Theorem 2.4 (Contrast independence)**.** *If $\Omega \subset \mathbb{R}^2$ is convex, $\Gamma \in C^{1,1}$, and if* (2.9) *is satisfied, then the following a priori error estimate holds:*

$$|||u^\star - u^{\mathrm{cfe}}||| \le C\|h\|_{L^\infty(\Omega)}\,\|f\|_{L^2(\Omega)}\,.$$

The constant $C = C(\rho_{\mathcal{T}}, \rho_{\mathcal{T}_1^\Gamma}, C_1, C_2, C_{\mathrm{reg}})$ does not depend on f, the mesh width functions h, h_1^Γ, and the contrast a_{cont}.

Remark 2.5. As already mentioned in the introduction, our method is designed to capture the kink of the solution across the interface. Further lack of regularity, caused, e.g., by singularities at kinks of the interface, is not addressed by the proposed method and leads to reduced convergence rates. The actual rate depends on the strength of the singularities as usual, i.e., if $u^\star \in H^{1+s}(\Omega_1\cup\Omega_2)$ for some $s \in [0,1[$, then standard interpolation theory of Sobolev spaces allows one to estimate

$$|||u^\star - u^{\mathrm{cfe}}||| \le C\|h\|^s_{L^\infty(\Omega)}.$$

Standard techniques may be applied to improve the convergence rate of the method for singular solutions, e.g., adding certain singular functions to the approximation space, or adaptive refinement of the coarse mesh $\mathcal{T}$ toward the singularity.

2.5. **Complexity.** Let us briefly discuss the complexity of our method. Considering a uniform coarse-scale grid of width h, the number of degrees of freedom in our method is proportional to h^{-d}, where $d \in \{2,3\}$ denotes the dimension of the physical space as before. The cost of setting up and storing the basis functions can be estimated as follows. Because of Theorems 2.3 and 2.4, it is sufficient to adapt the shape functions on a submesh with minimal mesh size $h_1^\Gamma|_\Gamma \approx h^{3/2}$ close to the interface. Since the dimension of the interface Γ is $d-1$ and owing to (2.2), the number of elements in the submesh is proportional to $h^{-3/2(d-1)}$, i.e., $h^{-3/2}$ for $d=2$ and h^{-3} for $d=3$. Hence, the cost caused by the adaptation of the shape functions is at most proportional to the number of degrees of freedom h^{-d}.

3. Detailed error analysis

This section proves the error estimate in Theorem 2.3. The error of the CFE approximation can be estimated as in [3, Lemma 10.1.7] by

$$|||u^\star - u^{\mathrm{cfe}}||| \leq \inf_{v\in S^{\mathrm{cfe}}} |||u^\star - v||| + \sup_{0\neq v\in S^{\mathrm{cfe}}} \frac{\left|\mathfrak{a}(u^\star - u^{\mathrm{cfe}}, v)\right|}{|||v|||}. \tag{3.1}$$

The first term in the above estimate reflects the best approximation error which is further addressed in Section 3.1. The additional second term is due to nonconformity (see Section 3.2).

3.1. **Approximation property.** For $\mathcal{G} \in \{\mathcal{T}, \mathcal{T}_1, \mathcal{T}_2, \mathcal{T}_1^\Gamma, \mathcal{T}_2\}$, let $\mathcal{I}_\mathcal{G} u$ denote the unique $\mathcal{G}$-piecewise affine function that interpolates a sufficiently smooth function u at the vertices of $\mathcal{G}$. The solution $u^\star \in H^2(\Omega_1 \cup \Omega_2)$ of (2.1) is well approximated by the discontinuous function u_h with $u_h|_{\Omega_k} = (\mathcal{I}_{\mathcal{T}_k} u)|_{\Omega_k}$, $k = 1, 2$. The error in the energy norm is proportional to h. This approximation property is preserved if u_h is suitably mapped onto the finite element space S^{cfe} as the following lemma states.

For the ease of notation, observe that $\mathcal{P}^{\mathrm{cfe}} u := \mathcal{P}^{\mathrm{cfe}}(\mathcal{I}_{\mathcal{T}_1} u, \mathcal{I}_{\mathcal{T}_2} u)$ defines $\mathcal{P}^{\mathrm{cfe}}$ for arguments $u \in H^1_0(\Omega) \cap H^2(\Omega_1 \cup \Omega_2)$. Accordingly, $\mathcal{P}_2^{\mathrm{cfe}} u := \mathcal{P}_2^{\mathrm{cfe}} \mathcal{I}_{\mathcal{T}_2} u$ (resp. $\mathcal{P}_1^{\mathrm{cfe}} u := \mathcal{P}_1^{\mathrm{cfe}}(\mathcal{I}_{\mathcal{T}_1} u, \mathcal{I}_{\mathcal{T}_2} u)$) extends $\mathcal{P}_2^{\mathrm{cfe}}$ (resp. $\mathcal{P}_1^{\mathrm{cfe}}$) to $H^1_0(\Omega) \cap H^2(\Omega_1 \cup \Omega_2)$.

Lemma 3.1 (Approximation property of S^{cfe})**.** *There is a constant $C > 0$ which may depend on $\rho_\mathcal{T}, \rho_{\mathcal{T}_1^\Gamma}, C_1, C_2$ but* not *on h and h_1^Γ such that for all $u \in H^1_0(\Omega) \cap H^2(\Omega_1 \cup \Omega_2)$ it holds that*

$$|||u - \mathcal{P}^{\mathrm{cfe}} u||| \leq C \left\| \sqrt{a} h \nabla^2 u \right\|_{L^2(\Omega_1 \cup \Omega_2)}.$$

Proof. The proof picks up some standard techniques for CFEs as they are used, e.g., in the proof of Theorem 4.4 in [18]. In addition, we will frequently make use of classical error estimates of nodal interpolation with respect to simplices. Following [5, Theorem 16.1], there exists a universal constant C_{ip} such that

$$|u - \mathcal{I}_t u|_{W^m_p(t)} \leq \frac{C_{\mathrm{ip}}}{\rho_t} \operatorname{diam}(t)^{2 - \frac{d}{2} - \frac{d}{p} - m} |u|_{H^2(t)} \tag{3.2}$$

for all $u \in H^2(t)$, $m \in \{0, 1\}$, provided $W^m_p(t) \subset H^2(t)$; $\mathcal{I}_t u$ denotes the affine interpolant of u at the vertices of a triangle t.

The main tool for exploiting the piecewise regularity is a suitable extension operator. It is known that, since Ω_k is assumed bounded and Lipschitz, there exists a continuous, linear extension operator $\mathfrak{E}_k : H^2(\Omega_k) \cap H^1_0(\Omega) \to H^2(\Omega) \cap H^1_0(\Omega)$, $k \in \{1, 2\}$, such that for all $u \in H^2(\Omega_k) \cap H^1_0(\Omega)$ there holds

$$\mathfrak{E}_k u|_{\Omega_k} = u \quad \text{and} \quad \left\| \nabla^2 \mathfrak{E} u \right\|_{L^2(\Omega)} \leq C_{\mathrm{ext}} \left\| \nabla^2 u \right\|_{L^2(\Omega_k)} \tag{3.3}$$

with a constant C_{ext} that depends only on Ω_k and Ω [25]. Moreover, C_{ext} is moderately small under mild assumptions on the geometry [23]. Throughout the rest of the paper, u_k abbreviates $\mathfrak{E}_k u$, $k = 1, 2$.

Our proof rests upon the splitting

$$u = u_2 + (u - u_2)$$

and the observation that $(u-u_2)|_{\Omega_1} \in H^1_0(\Omega_1)\cap H^2(\Omega_1)$ and $(u-u_2)|_{\Omega_2}=0$. The splitting and the linearity of $\mathcal{P}^{\mathrm{cfe}}$ lead to the upper bound

$$|||u-\mathcal{P}^{\mathrm{cfe}}u|||^2 \le |||u_2-\mathcal{P}^{\mathrm{cfe}}u_2|||^2 + \left\|\nabla\left(u-u_2-\mathcal{P}^{\mathrm{cfe}}(u-u_2)\right)\right\|^2_{L^2(\Omega_1)}. \tag{3.4}$$

The second term on the right–hand side of (3.4) can be bounded by classical techniques for the analysis of CFEs. In particular, [18, Theorem 4.4] and (3.3) show that

$$\left\|\nabla\left((u-u_2)-\mathcal{P}^{\mathrm{cfe}}(u-u_2)\right)\right\|^2_{L^2(\Omega_1)} \le C\left\|h\nabla^2 u\right\|_{H^2(\Omega_1)} \tag{3.5}$$

with some constant C that depends only on $\rho_{\mathcal{T}}$, $\rho_{\mathcal{T}_1^\Gamma}$, C_1, C_2, and C_{ext}.

Thus, we are left to bound the first term on the right–hand side of (3.4). The advantage of the splitting is that, compared to the initial assertion, we can now make use of the fact that $u_2\in H^2(\Omega)$ regardless of the interface.

Throughout the rest of the proof, $a\lesssim b$ abbreviates $a\le Cb$ with some constant C that depends only on the constants C_1, C_2, C_{ip}, C_{ext}, $\rho_{\mathcal{T}}$, and $\rho_{\mathcal{T}_1^\Gamma}$.

By repeated use of the triangle inequality we separate the elements where standard estimates apply from those where more involved techniques are required:

$$\begin{aligned}
|||u_2-\mathcal{P}^{\mathrm{cfe}}u_2|||^2 &= \left\|\nabla(u_2-\mathcal{P}^{\mathrm{cfe}}u_2)\right\|^2_{L^2(\Omega_1)} + a_{\mathrm{cont}}\left\|\nabla(u_2-\mathcal{P}^{\mathrm{cfe}}u_2)\right\|^2_{L^2(\Omega_2)}\\
&\overset{(2.6),(2.7)}{\le} \|\nabla(u_2-\mathcal{I}_{\mathcal{T}}u_2)\|^2_{L^2(\Omega_1)} + a_{\mathrm{cont}}\|\nabla(u_2-\mathcal{I}_{\mathcal{T}}u_2)\|^2_{L^2(\Omega_2)}\\
&\quad + \left\|\nabla(\mathcal{I}_{\mathcal{T}}u_2-\mathcal{P}_1^{\mathrm{cfe}}u_2)\right\|^2_{L^2(\Omega^\Gamma)} + a_{\mathrm{cont}}\left\|\nabla(\mathcal{I}_{\mathcal{T}}u_2-\mathcal{P}_2^{\mathrm{cfe}}u_2)\right\|^2_{L^2(\Omega^\Gamma)}\\
&\overset{(3.2),(3.3)}{\lesssim} \left\|h\nabla^2u_2\right\|^2_{L^2(\Omega_1)} + a_{\mathrm{cont}}\left\|h\nabla^2u_2\right\|^2_{L^2(\Omega_2)}\\
&\quad + \left\|\nabla(\mathcal{I}_{\mathcal{T}}u_2-\mathcal{P}_1^{\mathrm{cfe}}u_2)\right\|^2_{L^2(\Omega^\Gamma)} + a_{\mathrm{cont}}\left\|\nabla(\mathcal{I}_{\mathcal{T}}u_2-\mathcal{P}_2^{\mathrm{cfe}}u_2)\right\|^2_{L^2(\Omega^\Gamma)}.
\end{aligned} \tag{3.6}$$

Let $t\in\mathcal{T}_1^\Gamma$ and $T\in\mathcal{T}_2^\Gamma$ such that $t\subset T$ (recall that $\mathcal{T}_1^\Gamma$ was derived from $\mathcal{T}_2^\Gamma\subset\mathcal{T}$ by refinement). Then, by an inverse estimate,

$$\left\|\nabla(\mathcal{I}_T u_2-\mathcal{P}_1^{\mathrm{cfe}}u_2)\right\|_{L^2(t)} \le 2\,\mathrm{diam}(t)^{d/2}\,\mathrm{diam}(T)^{-1}\left\|\mathcal{I}_T u_2-\mathcal{P}_1^{\mathrm{cfe}}u_2\right\|_{L^\infty(t)}. \tag{3.7}$$

We fix $x\in V(t)$ with $\left\|\mathcal{I}_T u_2-\mathcal{P}_1^{\mathrm{cfe}}u_2\right\|_{L^\infty(t)} = \left|\mathcal{I}_T u_2(x)-\mathcal{P}_1^{\mathrm{cfe}}u_2(x)\right|$ and define $T_t^1:=T_x^1$, $T_t^2:=T_x^2$. In addition we introduce neighborhoods

$$\omega_T := \bigcup\{K\in\mathcal{T}:\ T\cap K\neq\emptyset\}$$

containing both coarse elements T_t^1 and T_t^2. The definition of $\mathcal{P}_1^{\mathrm{cfe}}$ in (2.7) and the application of Lemma 4.1 from [18] lead to

$$\begin{aligned}
&\left\|\mathcal{I}_T u_2-\mathcal{P}_1^{\mathrm{cfe}}u_2\right\|_{L^\infty(t)} \overset{(2.7)}{=} \left|\mathcal{I}_T u_2(x)-\mathcal{I}_{T_t^2}u_2(x^\Gamma)-\langle\nabla\mathcal{I}_{T_t^1}u_2, x-x^\Gamma\rangle\right|\\
&\lesssim \left|\mathcal{I}_T u_2(x)-\mathcal{I}_{T_t^2}u_2(x)\right| + \left|\langle\nabla(\mathcal{I}_{T_t^2}u_2-\mathcal{I}_{T_t^1}u_2), x-x^\Gamma\rangle)\right|\\
&\lesssim \mathrm{diam}(T)^{-d/2}\left(\left\|\mathcal{I}_T u_2-\mathcal{I}_{T_t^2}u_2\right\|_{L^2(T)} + \mathrm{diam}(t)\left\|\nabla\left(\mathcal{I}_T u_2-\mathcal{I}_{T_t^2}u_2\right)\right\|_{L^2(T)}\right.\\
&\quad \left. + \mathrm{diam}(t)\left\|\nabla\left(\mathcal{I}_{T_t^2}u_2-\mathcal{I}_{T_t^1}u_2\right)\right\|_{L^2(T)}\right)\\
&\overset{\mathrm{diam}(t)\le\mathrm{diam}(T)}{\lesssim} \mathrm{diam}(T)^{2-d/2}\left\|\nabla^2u_2\right\|_{L^2(\omega_T)}.
\end{aligned} \tag{3.8}$$

The summation over all $t \in \mathcal{T}_1^\Gamma$ yields

$$
\begin{aligned}
&\left\|\nabla(\mathcal{I}_{\mathcal{T}} u_2 - \mathcal{P}_1^{\mathrm{cfe}} u_2)\right\|^2_{L^2(\Omega^\Gamma)} \leq \sum_{t\in\mathcal{T}_1^\Gamma} \left\|\nabla(\mathcal{I}_t u_2 - \mathcal{P}_1^{\mathrm{cfe}} u_2)\right\|^2_{L^2(t)} \\
&\leq \sum_{T\in\mathcal{T}} \sum_{t\in\mathcal{T}_1^\Gamma : t\subset T} \left\|\nabla(\mathcal{I}_T u_2 - \mathcal{P}_1^{\mathrm{cfe}} u_2)\right\|^2_{L^2(t)} \\
&\overset{(3.7),(3.8)}{\lesssim} \sum_{T\in\mathcal{T}} \Big(\sum_{t\in\mathcal{T}_1^\Gamma : t\subset T} |t| \Big) \operatorname{diam}(T)^{2-d} \left\|\nabla^2 u_2\right\|^2_{L^2(\omega_T)} \\
&\overset{(3.3)}{\lesssim} \left\|h\nabla^2 u\right\|^2_{L^2(\Omega_2)}.
\end{aligned}
\tag{3.9}
$$

Similar arguments as in (3.7), (3.8), and (3.9) lead to an estimate of the last term on the right-hand side of (3.6),

$$
\left\|\nabla(\mathcal{I}_{\mathcal{T}_2} u_2 - \mathcal{P}_2^{\mathrm{cfe}} u_2)\right\|^2_{L^2(\cup\mathcal{T}_2^\Gamma)} \lesssim \left\|h\nabla^2 u\right\|^2_{L^2(\Omega_2)}. \tag{3.10}
$$

The combination of (3.4), (3.5), (3.6), (3.9), and (3.10) proves the assertion. □

3.2. **Non-conformity.** If the solution is sufficiently smooth, i.e., $u^\star \in H^{3/2}(\Omega_1 \cup \Omega_2)$, the second term in (3.1) can be estimated using Greens's identity, (2.1), (2.8), the classical jump relation, and the Cauchy-Schwarz inequality as follows:

$$
\sup_{0\neq v\in S^{\mathrm{cfe}}} \frac{\left|\mathfrak{a}(u^\star - u^{\mathrm{cfe}}, v)\right|}{|||v|||} \leq C \left\| \frac{\partial u^\star}{\partial \nu_{\Omega_1}} \right\|_{L^2(\Gamma)} \sup_{0\neq v\in S^{\mathrm{cfe}}} \frac{\|[[v]]_\Gamma\|_{L^2(\Gamma)}}{|||v|||}. \tag{3.11}
$$

Here, ν_{Ω_1} denotes the outer normal of Ω_1 and $[[v]]_\Gamma$ denotes the jump of v across Γ. By picking up ideas from [18, Lemma 4.9], one checks that the discontinuity $[[v]]_\Gamma$ is in fact small.

Lemma 3.2 (Non-conformity)**.** *There is a constant $C = C(C_1, C_3) > 0$ with $C_3 := \max_{T\in\mathcal{T}_1^\Gamma : T\cap\Gamma\neq\emptyset} |\Gamma\cap T|/\operatorname{diam}(T)^{(d-1)}$, such that*

$$
\|[[v]]_\Gamma\|_{L^2(\Gamma)} \leq C\|h\|_{L^\infty(\Omega)} \, \|h_1^\Gamma / h^{3/2}\|_{L^\infty(\bigcup\{T\in\mathcal{T}_1^\Gamma : T\cap\Gamma\neq\emptyset\})} |||v|||
$$

for all $v \in S^{\mathrm{cfe}}$.

Proof. Let $v = \mathcal{P}^{\mathrm{cfe}}(\mathcal{I}_{\mathcal{T}_1} v, \mathcal{I}_{\mathcal{T}_2} v) \in S^{\mathrm{cfe}}$. Let $t \in \mathcal{T}_1^\Gamma$ with $t \cap \Gamma \neq \emptyset$. We start with some pointwise estimate of the jump of v on t:

$$
\|[[v]]\|_{L^\infty(\Gamma\cap t)} = \|v|_{\Omega_2} - v|_{\Omega_1}\|_{L^\infty(\Gamma\cap t)} \leq \|v|_{\Omega_2} - v|_{\Omega_1}\|_{L^\infty(t)},
$$

where $v|_{\Omega_1}$ (resp. $u|_{\Omega_2}$) is identified with its unique affine extension onto t. The definitions (2.6) and (2.7) yield

$$
\begin{aligned}
\|[[v]]\|_{L^\infty(\Gamma\cap t)} &= \max_{y\in V(t)} \left| v|_{\Omega_2}(y) - v|_{\Omega_2}(y^\Gamma) - \left\langle \nabla \mathcal{I}_{T_y^1} v, y - y^\Gamma \right\rangle \right| \\
&\lesssim \operatorname{diam}(t) \left\| h^{-d/2} \nabla v \right\|_{L^2(T_y^1 \cup T_y^2)}.
\end{aligned}
\tag{3.12}
$$

Hence, the L^2-norm of v on $\partial\Omega$ is estimated as follows:

$$\begin{aligned}\|v\|^2_{L^2(\Gamma)} &\leq \sum_{T_1\in\mathcal{T}_1, T_2\in\mathcal{T}_2} \sum_{t\in\mathcal{T}_1^\Gamma : t\cap\Gamma\neq\emptyset, T_t^k = T_k} |\Gamma\cap t|\, \|[[v]]\|^2_{L^\infty(t)} \\ &\overset{(3.12)}{\lesssim} \sum_{T_1\in\mathcal{T}_1, T_2\in\mathcal{T}_2} \sum_{t\in\mathcal{T}_1^\Gamma : t\cap\Gamma\neq\emptyset, T_t^k = T_k} \frac{|\Gamma\cap t|\,\operatorname{diam}(t)^2}{\operatorname{diam}(T)^d}\, \|\nabla v\|^2_{L^2(T_1\cup T_2)} \\ &\lesssim \|h_1^\Gamma/\sqrt{h}\|^2_{L^\infty(\cup\{t\in\mathcal{T}_1^\Gamma:\, t\cap\Gamma\neq\emptyset\})}\, \|\nabla v\|^2_{L^2(\Omega)}\,.\end{aligned}$$

□

If $h_1^\Gamma|_\Gamma$ is chosen proportional to $h^{(3/2)}$, as it is assumed in (2.9), Theorem 2.3 follows from (3.1), Lemma 3.1, and Lemma 3.2.

Remark 3.3. The constant C_3 introduced in Lemma 3.2 reflects the smoothness of the interface Γ. Note that C_3 may be large if Γ is highly oscillating. However, the proof of Lemma 3.2 shows that a possibly large constant can be controlled by simply choosing $h^\Gamma|_\Gamma$ appropriately small. This modification concerns only the submesh $\mathcal{T}_1^\Gamma$ and does not affect the overall number of degrees of freedom.

4. Regularity

This section proves the regularity result (2.11) under the following assumptions on the geometrical setting:

(R1) $\Omega\subset\mathbb{R}^2$ is a convex polygon,
(R2) $\Omega_1,\Omega_2\subset\Omega$ are disjoint open Lipschitz domains with $\overline{\Omega}=\overline{\Omega}_1\cup\overline{\Omega}_2$, and
(R3) $\Gamma:=\overline{\Omega}_1\cap\overline{\Omega}_2$ is a $C^{1,1}$ curve that separates Ω_1 and Ω_2.

The conditions (R1)–(R3) guarantee that both subdomains Ω_k have a piecewise smooth boundary with interior angles less than π. In particular, the interface is not tangential to $\partial\Omega$ in intersection points $\Gamma\cap\partial\Omega$. Two relevant cases covered by these conditions are depicted in Figure 5.

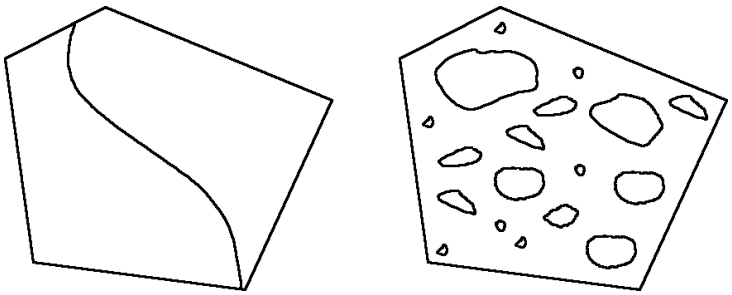


Figure 5. Two geometric situations in which conditions (R1)–(R3) are satisfied: (left) interface cuts through the boundary of Ω at some positive angle, (right) smooth separated inclusions dispersed in some matrix.

Under the conditions (R1)–(R3) [13] shows piecewise H^2 regularity in the sense that $f\in L^2(\Omega)$ implies $u^\star\in H^2(\Omega_1\cup\Omega_2)$ and

$$\|\nabla^2 u^\star\|_{L^2(\Omega_1)} + \|\nabla^2 u^\star\|_{L^2(\Omega_2)} \leq C\|f\|_{L^2(\Omega)}.$$

The subsequent theorem clarifies the dependence of the constant in the estimate above on the contrast parameter a_{cont}. In this respect, the theorem generalizes the previous result [4, Theorem B.1], which assumes a smoother interface and, more importantly, the inclusion $\bar{\Omega}_1 \subset \Omega$ with some positive distance between Ω_1 and $\partial\Omega$.

Theorem 4.1. *Under the assumptions (R1)–(R3) the unique solution $u^\star$ of* (2.1) *is piecewise smooth, $u^\star \in H^2(\Omega_1 \cup \Omega_2)$. Moreover, the estimates* (2.11) *hold with a generic constant C_{reg} that does not depend on f and a_{cont}.*

Proof. Let $u_k^\star := u^\star|_{\Omega_k}$ for $k = 1, 2$. As discussed earlier, the assumptions (R1)–(R3) yield $u_k^\star \in H^2(\Omega_k)$ for $k = 1, 2$. Since Ω_k, $k = 1, 2$, is piecewise smooth with interior angles less than π, classical a priori bounds yield

$$\left\|\nabla^2 u_1^\star\right\|_{L^2(\Omega_1)} \leq C'_{\mathrm{reg}} \left(\|f\|_{L^2(\Omega_1)} + \|u_1^\star\|_{H^{3/2}(\Gamma)} \right),$$
$$\left\|\nabla^2 u_2^\star\right\|_{L^2(\Omega_2)} \leq C''_{\mathrm{reg}} \left(a_{\mathrm{cont}}^{-1} \|f\|_{L^2(\Omega_2)} + \left\| \frac{\partial u_2^\star}{\partial \nu_{\Omega_2}} \right\|_{H^{1/2}(\Gamma)} \right).$$

Since the above estimates are solely performed in the subdomains, the constants C'_{reg} and C''_{reg} do not depend on a_{cont}. The classical jump relations at the interface imply $\|u_1^\star\|_{H^{3/2}(\Gamma)} = \|u_2^\star\|_{H^{3/2}(\Gamma)}$ and $\left\| \frac{\partial u_1^\star}{\partial \nu_{\Omega_1}} \right\|_{H^{1/2}(\Gamma)} = a_{\mathrm{cont}} \left\| \frac{\partial u_2^\star}{\partial \nu_{\Omega_2}} \right\|_{H^{1/2}(\Gamma)}$. Hence,

$$\left\|\nabla^2 u_1^\star\right\|_{L^2(\Omega_1)} \leq C'_{\mathrm{reg}} \left(\|f\|_{L^2(\Omega_1)} + \|u_2^\star\|_{H^{3/2}(\Gamma)} \right), \tag{4.1}$$
$$\left\|\nabla^2 u_2^\star\right\|_{L^2(\Omega_2)} \leq C''_{\mathrm{reg}} a_{\mathrm{cont}}^{-1} \left(\|f\|_{L^2(\Omega_2)} + \left\| \frac{\partial u_1^\star}{\partial \nu_{\Omega_1}} \right\|_{H^{1/2}(\Gamma)} \right). \tag{4.2}$$

The combination of (4.2), the trace theorem $\left\| \frac{\partial u_1^\star}{\partial \nu_{\Omega_1}} \right\|_{H^{1/2}(\Gamma)} \leq C \|u_1^\star\|_{H^2(\Omega_1)}$, (4.1), and the trace theorem $\|u_2^\star\|_{H^{3/2}(\Gamma)} \leq C \|u_2^\star\|_{H^2(\Omega_2)}$ leads to

$$\left\|\nabla^2 u_2^\star\right\|_{L^2(\Omega_2)} \leq C'''_{\mathrm{reg}} a_{\mathrm{cont}}^{-1} \left(\|f\|_{L^2(\Omega)} + \|u^\star\|_{H^1(\Omega)} + \|\nabla^2 u_2^\star\|_{L^2(\Omega_2)} \right).$$

Since C'''_{reg} does not depend on a_{cont}, coercivity of the bilinear form $\mathfrak{a}$ and the energy estimate $|||u^\star||| \leq \|f\|_{L^2(\Omega)}$ prove the estimate

$$\left\|\nabla^2 u_2^\star\right\|_{L^2(\Omega_2)} \leq C_{\mathrm{reg}} a_{\mathrm{cont}}^{-1} \|f\|_{L^2(\Omega)},$$

provided $a_{\mathrm{cont}} \geq C'''_{\mathrm{reg}}/2$. Since for small contrast $a_{\mathrm{cont}} < C'''_{\mathrm{reg}}/2$ nothing is to show, one assertion is proved. The estimate for $\left\|\nabla^2 u_2^\star\right\|_{L^2(\Omega_2)}$ is analogous by interchanging the application of (4.2) and (4.1) as well as the corresponding trace inequalities. □

For a characterization of the singularities that may appear if the conditions (R1)–(R3) are not satisfied, we refer the reader to [2, 6, 10, 11] among many others. A comprehensive regularity analysis for the three-dimensional case is more technical and beyond the scope of this paper; we refer to [12] for necessary conditions under which $H^2(\Omega_1 \cup \Omega_2)$-regularity is achieved. If the geometric setting allows $H^2(\Omega_1 \cup \Omega_2)$-regularity, then the proof of (2.11) could be treated in a similar way as in Theorem 4.1.

We shall emphasize that the above result is not explicit with respect to the geometric setting, e.g., the constants C'_{reg} and C''_{reg} may depend on oscillations of the interface, minimal distances between inclusions, the distance between inclusions

and the boundary, etc. The dependence on the geometry is involved and has been studied only for special cases, e.g., the case of densely packed, perfectly conducting, circular inclusions in $2d$ [17]. We further mention that reularity estimates for the case of diffusive interfaces may be found in [20].

5. Conclusion

We have described a finite element method for the Poisson equation with discontinuous diffusion coefficient across some interface. The method does not require the underlying finite element mesh to resolve the interface exactly. Overlapping, and possibly structured, simplicial meshes can be used instead. Moreover, the definition of the basis functions is explicit, and no local problems have to be solved. On a quasi-uniform coarse grid of width h, the complexity of our method is proportional to h^{-d}, whereas the error is proportional to h. This is optimal in comparison with the approximation of a Poisson problem with overall constant coefficient on the same mesh.

This paper focuses on the difficulty of treating discontinuous coefficients. To keep notation and technicalities at a minimum, the simplest possible setting has been chosen. Generalizations, not only to general linear elliptic problems but also saddle point problems such as Stokes' problem, are straightforward with regard to the previous work [18, 19].

References

[1] Roland Becker, Erik Burman, and Peter Hansbo, *A Nitsche extended finite element method for incompressible elasticity with discontinuous modulus of elasticity*, Comput. Methods Appl. Mech. Engrg. **198** (2009), no. 41-44, 3352–3360, DOI 10.1016/j.cma.2009.06.017. MR2571349 (2011b:74023)

[2] Matthias Blumenfeld, *The regularity of interface-problems on corner-regions* (Oberwolfach, 1983), Lecture Notes in Math., vol. 1121, Springer, Berlin, 1985, pp. 38–54, DOI 10.1007/BFb0076261. MR806385 (87a:35063)

[3] Susanne C. Brenner and L. Ridgway Scott, *The mathematical theory of finite element methods*, 3rd ed., Texts in Applied Mathematics, vol. 15, Springer, New York, 2008. MR2373954 (2008m:65001)

[4] C.-C. Chu, I. G. Graham, and T.-Y. Hou, *A new multiscale finite element method for high-contrast elliptic interface problems*, Math. Comp. **79** (2010), no. 272, 1915–1955, DOI 10.1090/S0025-5718-2010-02372-5. MR2684351 (2011j:65267)

[5] Philippe G. Ciarlet, *The finite element method for elliptic problems*, North-Holland Publishing Co., Amsterdam, 1978. Studies in Mathematics and its Applications, Vol. 4. MR0520174 (58 #25001)

[6] Monique Dauge and Serge Nicaise, *Oblique derivative and interface problems on polygonal domains and networks*, Comm. Partial Differential Equations **14** (1989), no. 8-9, 1147–1192, DOI 10.1080/03605308908820649. MR1017069 (91a:35046)

[7] W. Hackbusch and S. A. Sauter, *Composite finite elements for problems containing small geometric details. Part II: Implementation and numerical results*, Computing and Visualization in Science **1** (1997), 15–25.

[8] W. Hackbusch and S. A. Sauter, *Composite finite elements for the approximation of PDEs on domains with complicated micro-structures*, Numer. Math. **75** (1997), no. 4, 447–472, DOI 10.1007/s002110050248. MR1431211 (97k:65251)

[9] Anita Hansbo and Peter Hansbo, *An unfitted finite element method, based on Nitsche's method, for elliptic interface problems*, Comput. Methods Appl. Mech. Engrg. **191** (2002), no. 47-48, 5537–5552, DOI 10.1016/S0045-7825(02)00524-8. MR1941489 (2003i:65105)

[10] R. B. Kellogg, *Singularities in interface problems*, In Numerical Solution of Partial Differential Equations, II (SYNSPADE 1970) (Proc. Sympos., Univ. of Maryland, College Park, Md., 1970), Academic Press, New York, 1971, pp. 351–400. MR0289923 (44 #7108)

[11] D. Leguillon and E. Sánchez-Palencia, *Computation of singular solutions in elliptic problems and elasticity*, John Wiley & Sons Ltd., Chichester, 1987. MR995254 (90m:73015)
[12] K. Lemrabet, *An interface problem in a domain of* $\mathbf{R}^3$, J. Math. Anal. Appl. **63** (1978), no. 3, 549–562, DOI 10.1016/0022-247X(78)90059-8. MR493687 (81c:35036)
[13] Keddour Lemrabet, *Régularité de la solution d'un problème de transmission* (French), J. Math. Pures Appl. (9) **56** (1977), no. 1, 1–38. MR0509312 (58 #23016)
[14] Jingzhi Li, Jens Markus Melenk, Barbara Wohlmuth, and Jun Zou, *Optimal a priori estimates for higher order finite elements for elliptic interface problems*, Appl. Numer. Math. **60** (2010), no. 1-2, 19–37, DOI 10.1016/j.apnum.2009.08.005. MR2566075 (2011a:65370)
[15] M. Ohlberger and M. Rumpf, *Hierarchical and adaptive visualization on nested grids*, Computing **59** (1997), no. 4, 365–385, DOI 10.1007/BF02684418. MR1486386 (98h:65058)
[16] D. Peterseim, *The composite mini element: A mixed FEM for the Stokes equations on complicated domains*, Ph.D. thesis, Universität Zürich, http://www.dissertationen.uzh.ch/, 2007.
[17] Daniel Peterseim, *Robustness of finite element simulations in densely packed random particle composites*, Netw. Heterog. Media **7** (2012), no. 1, 113–126, DOI 10.3934/nhm.2012.7.113. MR2908612
[18] Daniel Peterseim and Stefan A. Sauter, *The composite mini element-coarse mesh computation of Stokes flows on complicated domains*, SIAM J. Numer. Anal. **46** (2008), no. 6, 3181–3206, DOI 10.1137/070704356. MR2439507 (2009e:76107)
[19] Daniel Peterseim and Stefan A. Sauter, *Finite element methods for the Stokes problem on complicated domains*, Comput. Methods Appl. Mech. Engrg. **200** (2011), no. 33-36, 2611–2623, DOI 10.1016/j.cma.2011.04.017. MR2812028 (2012e:76074)
[20] D. Peterseim and S. Sauter, *Finite Elements for Elliptic Problems with Highly Varying, Nonperiodic Diffusion Matrix*, Multiscale Model. Simul. **10** (2012), no. 3, 665–695, DOI 10.1137/10081839X. MR3022017
[21] M. Rech, *Composite finite elements: An adaptive two-scale approach to the non-conforming approximation of Dirichlet problems on complicated domains*, Ph.D. thesis, Universität Zürich, 2006.
[22] M. Rech, S. Sauter, and A. Smolianski, *Two-scale composite finite element method for Dirichlet problems on complicated domains*, Numer. Math. **102** (2006), no. 4, 681–708, DOI 10.1007/s00211-005-0654-x. MR2207285 (2006m:65283)
[23] S. A. Sauter and R. Warnke, *Extension operators and approximation on domains containing small geometric details*, East-West J. Numer. Math. **7** (1999), no. 1, 61–77. MR1683936 (2000c:65105)
[24] S. A. Sauter and R. Warnke, *Composite finite elements for elliptic boundary value problems with discontinuous coefficients*, Computing **77** (2006), no. 1, 29–55, DOI 10.1007/s00607-005-0150-2. MR2207954 (2006k:65335)
[25] Elias M. Stein, *Singular integrals and differentiability properties of functions*, Princeton Mathematical Series, No. 30, Princeton University Press, Princeton, N.J., 1970. MR0290095 (44 #7280)

Institut für Mathematik, Humboldt-Universität zu Berlin, Unter den Linden 6, 10099 Berlin, Germany

Current address: Institut für Numerische Simulation der Universität Bonn, Wegelerstr. 6, 53115 Bonn, Germany

E-mail address: peterseim@ins.uni-bonn.de