UPPSALA
UNIVERSITET

Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology 1287

Multiscale Methods and Uncertainty Quantification

DANIEL ELFVERSON

ACTA
UNIVERSITATIS
UPSALIENSIS
UPPSALA
2015

ISSN 1651-6214
ISBN 978-91-554-9336-3
urn:nbn:se:uu:diva-262354

Dissertation presented at Uppsala University to be publicly examined in Room 2446, Polacksbacken, Lägerhyddsvägen 2, Uppsala, Friday, 30 October 2015 at 10:15 for the degree of Doctor of Philosophy. The examination will be conducted in English. Faculty examiner: Professor Frédéric Legoll (École Nationale des Ponts et Chaussées, Paris).

Abstract
Elfverson, D. 2015. Multiscale Methods and Uncertainty Quantification. *Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology* 1287. 32 pp. Uppsala: Acta Universitatis Upsaliensis. ISBN 978-91-554-9336-3.

In this thesis we consider two great challenges in computer simulations of partial differential equations: *multiscale data*, varying over multiple scales in space and time, and *data uncertainty*, due to lack of or inexact measurements.

We develop a multiscale method based on a coarse scale correction, using localized fine scale computations. We prove that the error in the solution produced by the multiscale method decays independently of the fine scale variation in the data or the computational domain. We consider the following aspects of multiscale methods: continuous and discontinuous underlying numerical methods, adaptivity, convection-diffusion problems, Petrov-Galerkin formulation, and complex geometries.

For uncertainty quantification problems we consider the estimation of p-quantiles and failure probability. We use spatial a posteriori error estimates to develop and improve variance reduction techniques for Monte Carlo methods. We improve standard Monte Carlo methods for computing p-quantiles and multilevel Monte Carlo methods for computing failure probability.

Keywords: multiscale methods, finite element method, discontinuous Galerkin, Petrov-Galerkin, a priori, a posteriori, complex geometry, uncertainty quantification, multilevel Monte Carlo, failure probability

Daniel Elfverson, Department of Information Technology, Division of Scientific Computing, Box 337, Uppsala University, SE-751 05 Uppsala, Sweden. Department of Information Technology, Numerical Analysis, Box 337, Uppsala University, SE-75105 Uppsala, Sweden.

ISSN 1651-6214
ISBN 978-91-554-9336-3
urn:nbn:se:uu:diva-262354 (http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-262354)

List of papers

This thesis is based on the following papers, which are referred to in the text by their Roman numerals.

I 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).

II 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).

III D. Elfverson. *A Discontinuous Galerkin Multiscale Method for Convection-Diffusion Problems.* Available as arXiv:1509.03523 e-print (submitted).

IV D. Elfverson, V. Ginting, and P. Henning. *On Multiscale Methods in Petrov-Galerkin Formulation.* Numer. Math. (2015).

V D. Elfverson, M. G. Larson, and A. Målqvist. *Multiscale Methods for Problems with Complex Geometry.* Available as arXiv:1509.03991 e-print (submitted).

VI D. Elfverson, D. J. Estep, F. Hellman, and A. Målqvist. *Uncertainty Quantification for Approximate p-Quantiles for Physical Models with Stochastic Inputs.* SIAM/ASA J. Uncertainty Quantification, 2(1), 826–850 (2014).

VII D. Elfverson, F. Hellman, and A. Målqvist. *A Multilevel Monte Carlo Method for Computing Failure Probabilities.* Available as arXiv:1408.6856 e-print (submitted).

Contents

1 Introduction 7

2 Model problem 9
- 2.1 The Poisson equation 9
- 2.2 The finite element method 10
- 2.3 The discontinuous Galerkin method 11

3 Multiscale problems 13
- 3.1 Multiscale methods 13
- 3.2 Continuous and discontinuous Galerkin method 15
- 3.3 Complex domain 16
- 3.4 Petrov-Galerkin formulation 16
- 3.5 Adaptivity for discontinuous Galerkin multiscale method. 17

4 Uncertainty quantification 18
- 4.1 Selective refinement 18
- 4.2 Multilevel Monte Carlo with selective refinement 19

5 Future works 21

6 Summary of papers 22
- 6.1 Paper I 22
- 6.2 Paper II 22
- 6.3 Paper III 23
- 6.4 Paper IV 24
- 6.5 Paper V 24
- 6.6 Paper VI 25
- 6.7 Paper VII 26

7 Summary in Swedish 27

8 Acknowledgments 29

References 30

1. Introduction

The focus of this thesis is twofold: we consider both partial differential equations (PDE) where the solution varies on several different scales, *multiscale problems*, and PDEs with uncertain data, *uncertainty quantification*. Modeling and simulation of this type of problems are very challenging and appear in most areas of science and engineering. A prominent example is flow in a porous medium. To apply standard one scale numerical methods and Monte Carlo (MC) simulation for multiscale and uncertainty quantification problems is in many cases intractable and in other cases impossible due to the immense cost. We will discuss how to address the difficulties in multiscale and uncertainty quantification problems separately.

Standard (one scale) numerical methods applied to multiscale problems fail to perform when the data is rough or the finest scale features of the data are not resolved by the underlying mesh. We will consider both when the coefficients and the computational domain have multiscale features. The main challenge in constructing numerical methods for multiscale problems is to reduce the computational complexity and still remain accurate. We propose a multiscale method where the coarse basis functions spanning the trial and/or test spaces are corrected using fine scale computations. Using a corrected basis the multiscale method has the same order of accuracy as a standard one scale method for smooth problems. The corrector problems are global, however the correctors decay exponentially away from the support of the coarse basis and the computation can be localized to patches. The size of the patches is chosen such that the accuracy is not affected. The corrector problems can be computed independently of each other, which makes them perfectly suited for parallel computation. The correctors can also be reused in e.g. time stepping and nonlinear iterations. For further discussion regarding numerical methods for multiscale problems see Section 3.

We consider applications where the model parameters are uncertain and random. We want to compute statistical properties of a quantity of interest of the solution of the PDE, in particular p-quantiles and failure probability. Failure probability is defined as the probability that a given functional or quantity of interest of the model solution is below some predetermined value. The estimation of p-quantiles is the inverse problem, i.e., determine the value such that a given functional of the solution is below that value with the predetermined probability p. Since we are interested in problems with high stochastic dimension, we consider sample based methods. When considering this type of problems we have two error sources: the numerical discretization of the

model and the stochastic sampling. To efficiently estimate p-quantiles or failure probability the two error sources need to be balanced. In this thesis we use spatial a posteriori error estimates within variance reductions techniques to reduce the computational cost and to balance the two error sources. For further discussion regarding failure probability see Section 4.

The main results of this thesis are the following:

- Adaptivity and convergence analysis for a Discontinuous Galerkin multiscale method.
- Multiscale methods in Petrov-Galerkin formulation.
- Extension of multiscale analysis to complex geometries.
- Improvment of Monte Carlo methods for p-quantiles and multilevel Monte Carlo method for failure probability, using selective refinement.

2. Model problem

In this chapter we present some notations, a model problem, and give a short introduction to the finite element method (FEM) and discontinuous Galerkin (DG) method.

2.1 The Poisson equation

We consider the boundary value problem

$$\begin{aligned} -\nabla \cdot A\nabla u &= f \quad \text{in } D, \\ u &= 0 \quad \text{on } \partial D, \end{aligned} \tag{2.1}$$

where D is a spatial domain with boundary ∂D, f is an external forcing, and A is a diffusion matrix. For multiscale problems A, ∂D, and f varies over several different scales that are not necessarily resolved by the computational mesh. For uncertainty quantification $A = A(\omega)$ and $f = f(\omega)$ are realizations from a given sample space Ω. In subsurface flow the physical interpretation of A is permability, illustrated in Figure 2.1.

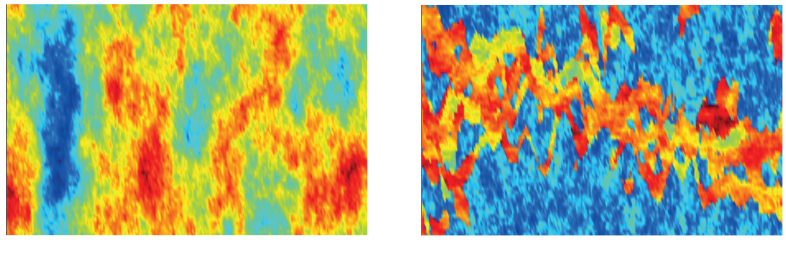


Figure 2.1. Examples of the permability in subsurface flow simulations.

Two function spaces which will be frequently used are $L^2(D)$ and $H^1(D)$. Both of the spaces are Hilbert spaces [2], i.e., both are complete inner product spaces with their inner products defined as

$$(u,v)_{L^2(D)} = \int_D uv\,\mathrm{d}x \quad \text{and} \quad (u,v)_{H^1(D)} = (\nabla u, \nabla v)_{L^2(D)} + (u,v)_{L^2(D)}, \tag{2.2}$$

respectively. We will denote

$$\|v\|_{L^2(D)} = \sqrt{(v,v)_{L^2(D)}} \quad \text{and} \quad \|v\|_{H^1(D)} = \sqrt{(v,v)_{H^1(D)}}, \tag{2.3}$$

the $L^2(D)$ and $H^1(D)$ norms induced by their inner products. Let us consider the function space $V_0 = \{v \in H^1(D) \mid v|_{\partial D} = 0\}$, i.e., all functions in $H^1(D)$ that vanishes on the boundary ∂D. The weak form of (2.1) reads: find $u \in V_0$ such that

$$a(u,v) := \int_D A\nabla u \cdot \nabla v \,\mathrm{d}x = \int_D fv \,\mathrm{d}x =: F(v) \quad \text{for all } v \in V_0. \tag{2.4}$$

By the Lax-Milgram lemma, there exists a unique solution $u \in V_0$ to (2.4) if the bilinear form $a(\cdot,\cdot)$ is coercive and continuous and the forcing function $F(\cdot)$ is a bounded linear functional. For a bilinear form to be coercive and continuous it has to fulfill

$$a(v,v) \geq C_1 \|v\|^2_{H^1(D)}, \quad \text{and} \quad |a(v,w)| \leq C_2 \|v\|_{H^1(D)} \|w\|_{H^1(D)}, \tag{2.5}$$

for all $v, w \in V_0$.

2.2 The finite element method

Since there typically is no closed form solution to (2.4) it needs to be approximated by a numerical method. A powerful numerical method is the FEM which has a strong mathematical foundation from functional analysis, that can be used to derive analytic error estimates/bounds [6].

The FEM seeks the solution in a finite dimensional subset $V_h \subset V$ of continuous piecewise polynomials defined on a mesh $\mathscr{T}_h$ covering the computational domain. The mesh typically consists of triangles/quadrilaterals in 2D and tetrahedras/prisms in 3D. Let $h : \Omega \to \mathbb{R}$ be a mesh-size function defined elementwise as $h|_T = \mathrm{diam}(T)$, i.e., the diameter of smallest circle containing T. The FEM approximation reads: find $u_h \in V_h$ such that

$$a(u_h, v) = F(v) \quad \text{for all } v \in V_h. \tag{2.6}$$

For the solution u_h to be a good approximation of u, the space V_h needs to resolve the variation in A. For many realistic problems this assumption is very computationally demanding to fulfill.

There are two main classes of error estimates or bounds for the FEM, *a priori* and *a posteriori*. The a priori error bound depends on the data and smoothness of the exact solution u, i.e.,

$$|||u - u_h||| := \|A^{1/2}\nabla(u - u_h)\|_{L^2(D)} \leq Ch^{s-1}|u|_{H^s(D)}, \tag{2.7}$$

where $h = \max_{T \in \mathscr{T}_h} h|_T$ and $H^s(D)$ is a function space containing all functions with bounded weak derivatives of total degree s in $L^2(D)$. To achieve linear convergence the smoothness constraint $u \in H^2(D)$ must be fulfilled. Higher order convergence can be obtain if both higher order polynomials are used and

the exact solution is smoother, $s > 2$. However even if $u \in H^2(D)$, $|u|_{H^2(D)}$ depends on the variation of the coefficient A and there is a pre-asymptotic regime where no convergence occur until the variations are resolved. The a posteriori error bound depends on the data and the numerical solution. Hence, a posteriori error bounds can be used in an adaptive algorithm to improve the numerical solution iteratively. For the standard FEM the a posteriori error bounds have the form

$$\begin{aligned} |||u-u_h|||^2 \leq C \sum_{T\in\mathscr{T}_h} \Big(& h|_K^2 \|f+\nabla\cdot A\nabla u_h\|^2_{L^2(T)} \\ & + h|_K \|\nu\cdot[A\nabla U]\|^2_{L^2(\partial T)} \Big), \end{aligned} \tag{2.8}$$

where $[\cdot]$ is the jump in function value and ν is a unit normal on ∂T.

2.3 The discontinuous Galerkin method

An interesting alternative to the standard (conforming) FEM is the DG method. In DG methods there is no continuity constraint imposed on the approximation spaces. Instead the continuity is imposed weakly in the bilinear form, i.e., the DG method allows for jumps in the numerical solution between different elements in the mesh. The first DG method was introduced in [34] for numerical approximations of first order hyperbolic problems and analyzed in [26, 24]. For some early work for DG method for elliptic problems see [38, 8, 3]. See also [30] for some preliminary work and [18, 33, 35] for a literature review.

We will use the same notation for the bilinear form, energy norm, and for the discrete function spaces for the DG method as for the FEM, however, with different definitions. The approximation space for the DG method, V_h, is the space of piecewise polynomials, i.e., DG methods uses a non-conforming ansatz $V_h \not\subset V$. The DG method has a higher number of degrees of freedom than the standard (conforming) FEM, but has the advantages that non-conforming meshes can be used and that it does not suffer from stability issues for first order or convection dominated PDEs. Also, the DG method is perfectly suited for hp-adaptivity, where both the mesh size and the order of the polynomial's degree can vary over the domain, see e.g. [21]. Since the DG method seeks the solution in a space which consists of piecewise polynomials without any continuity constraints, a modified bilinear form has to be used. In the bilinear form the continuity is imposed weakly, i.e., there is a penalty which forces the jump in the approximate solution to decrease when the mesh-size decreases. Let $\mathscr{T}_h$ be a given mesh and $\mathscr{E}_h$ be the skeleton of the mesh, i.e., the set of all edges of the elements in $\mathscr{T}_h$. For two elements T^+ and T^- sharing a common edge, $e := T^+ \cap T^-$, the jump and average on e are defined as

$$\{v\} = \frac{1}{2}(v|_{T^+} + v|_{T^-}) \quad \text{and} \quad [v] = v|_{T^+} - v|_{T^-} \tag{2.9}$$

in the interior and as $\{v\} = [v] = v_T$ on the boundary. We let ν_e be the unit normal pointing from T^+ to T^- and σ_e be an edge-wise constant depending on A. The bilinear form for the DG method is defined as

$$\begin{aligned} a(u,v) = &\sum_{T\in\mathscr{T}_h}\int_T A\nabla u\cdot\nabla v\,dx \\ &-\sum_{e\in\mathscr{E}_h}\int_e\left(\nu_e\cdot\{A\nabla u\}[v]+\nu_e\cdot\{A\nabla v\}[u]-\frac{\sigma_e}{h}[u][v]\right)ds. \end{aligned} \tag{2.10}$$

where σ_e is chosen large enough to make the bilinear form coercive in the standard DG energy norm, which is defined as

$$|||v|||^2 = \left(\sum_{T\in\mathscr{T}_h}\|A^{1/2}\nabla v\|^2_{L^2(T)}+\sum_{E\in\mathscr{E}_h}\frac{\sigma_e}{h}\|[v]\|^2_{L^2(e)}\right)^{1/2}. \tag{2.11}$$

The DG method reads: find $u_h \in V_h$ such that

$$a(u_h,v) = F(v) \quad \text{for all } v\in V_h. \tag{2.12}$$

Discontinuous Galerkin methods, as well as conforming finite element methods, perform badly when the smallest length scale of the data is not resolved. However, DG methods have the advantage in treating discontinuous coefficients, convection dominated problems, mass conservation, and flexibility of the underling mesh, all which are crucial issues in many multiscale problems, including e.g. porous media flow.

3. Multiscale problems

For *multiscale problems* standard numerical techniques fail to perform when the data is not resolved by the computational mesh [4]. A remedy for this, when the roughness is local in space, is adaptive techniques [37]. However, this is not the case for many multiscale problems. We will consider both when there are multiscale features in the coefficients and in the computational domain. In particular we consider multiscale diffusion, domains with cracks, and rough boundaries.

In the last two decades there have been a lot of research on multiscale methods treating some of these difficulties, see e.g. [20, 19, 13, 12, 9, 10, 11, 22, 23, 25, 27, 28]. Common for this these approaches is that local problems are solved on subgrid patches which resolves the data variation. The solutions to the subproblems are then used to modify a coarse scale space or equation.

We consider the local orthogonal decomposition method (LOD) first presented in [28]. See [25, 27, 23] for some preliminary work and [14, 15, 29, 32] for further development. In the LOD method the test and trial space are decomposed into coarse and fine scale subspaces using a quasi-interpolation operator. The coarse space is then corrected using fine scale information such that the corrected basis takes the fine scale behavior of the data into account. The corrected basis is constructed to be orthogonal to the kernel of the quasi-interpolation operator in the scalar product induced by the bilinear form.

3.1 Multiscale methods

In this section we will not explicitly define the function spaces and bilinear form, instead we use an abstract formulation that fits both the FEM and the DG method. We let V_H and V_h, where H, h are mesh-size functions, be the finite dimensional spaces where V_H does not and V_h does resolve the data. We assume that $V_H \subset V_h$ and $H > h$. The space V_h is referred to as the reference space and the reference solution u_h solves: find $u_h \in V_h$ such that

$$a(u_h, v) = F(v) \quad \text{for all } v \in V_h. \tag{3.1}$$

We assume u_h to be a sufficiently good approximation of u. We split the reference space V_h into a coarse and a fine scale contribution. Let $\mathscr{I}_H : L^2(\Omega) \to V_H$ be a quasi-interpolation operator onto the coarse space V_H with range$(\mathscr{I}_H) = V_H$, i.e., $V_H = \mathscr{I}_H V_h$. To simplify the analysis, we will only consider when

$\mathscr{I}_H$ is a projection, $\mathscr{I}_H^2 = \mathscr{I}_H$. The interpolation operator needs to satisfy the following local approximation and stability estimate. For any $K \in \mathscr{T}_H$ and $v \in V_h$

$$\|A^{1/2}H^{-1}(v - \mathscr{I}_H v)\|_{L^2(K)} + \|A^{1/2}\nabla \mathscr{I}_H v\|_{L^2(K)} \leq C|||v|||_{\omega_K}, \tag{3.2}$$

holds, where $\omega_K = \text{int}\{S \in \mathscr{T}_H \mid \overline{S} \cap \overline{K} \neq 0\}$ and $|||\cdot|||_{\omega_K}$ is the energy norm restricted to ω_K. We define a fine correction space to be the kernel of the interpolation operator

$$V^{\text{f}} = (1 - \mathscr{I}_H)V_h = \{v \in V_h \mid \mathscr{I}_H v = 0\}. \tag{3.3}$$

Any function $v_h \in V_h$ can be decomposed into a coarse contribution, $v_H \in V_H$, and fine scale remainder, $v^{\text{f}} \in V^{\text{f}}$, i.e., $v_h = v_H + v^{\text{f}}$ where $v_H = \mathscr{I}_H v_H$ and $v^{\text{f}} = (1 - \mathscr{I}_H)v_h$. Choosing V_H as the coarse space the fine scale remainder v^{f} is large and oscillatory and does not decay until the variations in the data are resolved. A remedy is to correct the space V_H such that the coarse basis takes the fine scale into account. We define the corrected space by $V_H^{\text{ms}} = (1+Q)V_H$ where $Q : V_H \to V^{\text{f}}$ is defined as: given $v_H \in V_H$ find $Q(v_H) \in V^{\text{f}}$ such that

$$a(Q(v_H), w) = -a(v_H, w) \quad \text{for all } w \in V^{\text{f}}. \tag{3.4}$$

We can write the reference space as the direct sum $V_h = V_H^{\text{ms}} \oplus V^{\text{f}}$. By correcting the basis functions spanning the space $V_H = \text{span}\{\varphi_i\}$ we can write the corrected space as the span of corrected basis function $V_H^{\text{ms}} = \text{span}\{\varphi_i + Q(\varphi_i)\}$. To compute the correctors is a global computation which is as expensive as solving the original reference problem. Instead, each of the correctors of the basis φ_i are computed on localized patches

$$\begin{aligned} \omega_i^0 &:= \text{int}\left(\cup(\bar{T} \in \mathscr{T}_H \mid \bar{T} \cap \{x\} \neq \emptyset)\right) \cap \Omega, \\ \omega_i^\ell &:= \text{int}\left(\cup(\bar{T} \in \mathscr{T}_H \mid \bar{T} \cap \bar{\omega}_T^{\ell-1} \neq \emptyset)\right) \cap \Omega, \quad \text{for } \ell = 1, \ldots, L. \end{aligned} \tag{3.5}$$

See Figure 3.1 for a graphical illustration of a localized patch. Let us define

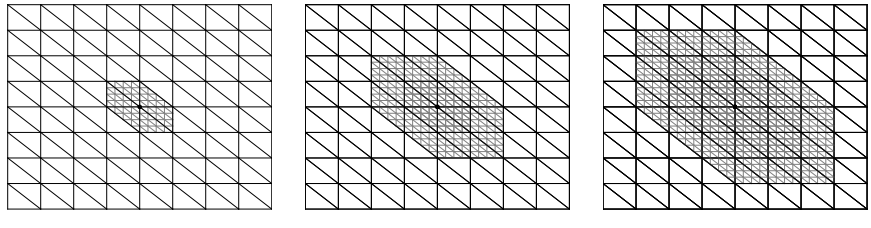


Figure 3.1. An example of 0, 1, and 2 level patches, i.e., ω_i^0, ω_i^0, and ω_i^2.

the localized corrected space by $V_H^{ms,L} = \text{span}\{\varphi_i + Q^L(\varphi_i)\}$ where Q^L solves: given ϕ_i find $Q(\phi_i) \in V^{\text{f}}(\omega_i^L) = \{v \in V^{\text{f}}(\omega_i^L) \mid v|_{D\setminus\omega_i^L=0}\}$ such that

$$a(Q^L(\phi_i), w) = -a(\phi_i, w) \quad \text{for all } v \in V^{\text{f}}(\omega_i^L). \tag{3.6}$$

The multiscale method posed in $V_H^{ms,L}$ reads: find $u_H^{ms,L} \in V_H^{ms,L}$ such that

$$a(u_H^{\text{ms}}, v) = F(v) \quad \text{for all } v \in V_H^{ms,L}. \tag{3.7}$$

The solution $u_H^{ms,L}$ fulfills the a priori error estimate

$$|||u - u_H^{\text{ms}}||| \leq |||u - u_h||| + CH, \tag{3.8}$$

choosing $L = \mathcal{O}(\log(H^{-1}))$, where H is the coarse mesh size and C is a constant independent of the mesh sizes h, H and the variations in A. We have that $\text{diam}(\omega_i^L) = \mathcal{O}(H\log(H^{-1}))$. See Paper II and IV for a more elaborate discussion and Paper III for a generalization toward convection-diffusion problems.

3.2 Continuous and discontinuous Galerkin method

The difference between the continuous and discontinuous Galerkin multiscale method is the choice of reference space V_h, bilinear form $a(\cdot,\cdot)$, and quasi-interpolation $\mathcal{I}_H$. The choices we make for the reference space and bilinear form are given in Section 2.2 for the FEM (continuous Galerkin) and in Section 2.3 for the DG multiscale method. The choice for the quasi-interpolation is not unique and different operators can be chosen depending on the application. Let $\mathcal{T}_H$ be the coarse mesh on which V_H is defined and $\mathcal{N}$ be the set of all vertices in $\mathcal{T}_H$.

For the continuous Galerkin method we choose $\mathcal{I}_H^{cG} : L^2(D) \to V_H$ to be defined by

$$\mathcal{I}_H^{cG} v = \sum_{x \in \mathcal{N}} (P_x v)(x) \varphi_x \tag{3.9}$$

where $P_x u \in V_H|_{\omega_x^0}$ solves

$$(P_x u, v)_{L^2(\omega_x^0)} = (u, v)_{L^2(\omega_x^0)} \quad \text{for all } v \in V_H|_{\omega_x^0}. \tag{3.10}$$

The space $V_H|_{\omega_x^0}$ is the restriction of V_H to the patch ω_x^0. See Paper V for a more elaborate discussion and Paper IV for an other choice of quasi-interpolation operator.

For the discontinuous Galerkin method we choose an elementwise L^2-projection $\mathcal{I}_H^{dG} : L^2(D) \to V_H$ defined by

$$\mathcal{I}_H^{dG} v = \sum_{T \in \mathcal{T}_H} \Pi_T v \tag{3.11}$$

where $\Pi_T u \in V_H|_T$ solves

$$(\Pi_T u, v)_{L^2(T)} = (u, v)_{L^2(T)} \quad \text{for all } v \in V_H|_T. \tag{3.12}$$

See Paper II for a more elaborate discussion.

3.3 Complex domain

So far most of the work in multiscale community has been focused for treating multiscale coefficients and less on treating complex domains. However, many multiscale applications involve voids, cracks, and rough interfaces. We extend the analysis for multiscale methods when there are multiscale features in the domain that are not resolved by the mesh. For simplicity we consider the case $A = 1$ in a complex domain. Then it is only necessary to compute the corrector problems close to the complex boundary and not in the entire domain D, see Figure 3.2. In Figure 3.2 the domain boundary cuts some of the coarse elements, however this does not affect the convergence and conditioning of the multiscale method. The condition number κ of the linear system obtained

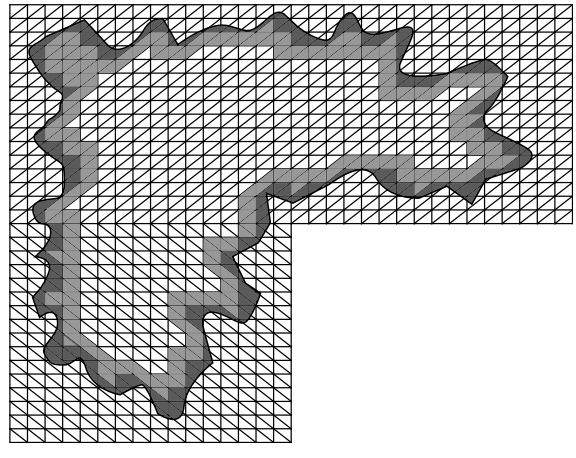


Figure 3.2. Example of complex domain embedded in a coarse mesh. The fine scale correctors only needs to be computed in the gray area. The dark Gray elements mark a 1-layer and the light gray a 2-layer patch of element round the complex boundary.

from (3.7) scales like

$$\kappa \leq CH^{-2}, \tag{3.13}$$

in the coarse mesh size H where C is a constant independent on the mesh-size and how the elements are cut by the domain boundary. See Paper V for a more elaborate discussion.

3.4 Petrov-Galerkin formulation

It is also possible to use a Petrov-Galerkin (P-G) formulation of the proposed multiscale method, i.e., using different test and trial spaces. The P-G formula-

tion reads: find $u_H^{ms,L} \in V_H^{ms,L}$ such that

$$a(u_H^{\text{ms}}, v) = F(v) \quad \text{for all } v \in V_H. \tag{3.14}$$

The P-G formulation still has the same convergence rate as the standard symmetric formulation. If global patches are used for the fine scale computations the two version are identical up to a perturbation of the right hand side,

$$a(u_H^{\text{ms}}, v) = a(u_H^{\text{ms}}, v + Q^L(v)) \tag{3.15}$$

for $v \in V_H$. However, the P-G formulation can reduce the computational complexity since no communication between the correctors is needed when assembling the matrices for the corrected coarse problem, i.e., the assembling is $a(\varphi_i + Q^L(\varphi_i), \varphi_j)$ for the P-G formulation and $a(\varphi_i + Q^L(\varphi_i), \varphi_j + Q^L(\varphi_j))$ for the standard symmetric formulation. See Paper IV for a more elaborate discussion.

3.5 Adaptivity for discontinuous Galerkin multiscale method.

For porous media flow problems the permeability in the ground can vary with several orders of magnitudes over the entire domain. This motivates the use of an adaptive multiscale method to tune the method parameters in order to obtain an efficient and reliable solution. For adaptive multiscale methods, see e.g. [25, 31, 17, 1]. Given a uniform or possibly an adaptive coarse mesh $\mathscr{T}_H$, the adaptive discontinuous Galerkin multiscale method balances the error caused by truncation of the patches and the fine scale discretization error. The a posteriori error bound takes the form

$$|||u - u_H^{\text{ms}}||| \leq C_1 \left(\sum_{S \in \mathscr{T}_h} \rho_S^2(u_H^{\text{ms}}) \right)^{1/2} + C_2 \left(\sum_{T \in \mathscr{T}_H} \rho_{\omega_T}^2(u_H^{\text{ms}}) \right)^{1/2}, \tag{3.16}$$

where ρ_S^2 and $\rho_{\omega_T}^2$ are error indicators which measure the effect of the fine scale mesh size and of the truncated patches, respectively. Since using general nonconforming meshes is allowed using DG, it is easy to construct a global reference grid for the localized fine scale computations. This takes advantage of the error cancellation between different fine scale patches. See Paper I for a more elaborate discussion.

4. Uncertainty quantification

We consider PDEs with uncertain data which have high stochastic dimension. More specifically, we consider the estimation of failure probability and p-quantiles. Given some physical model

$$\mathscr{M}(\omega, u) = 0 \tag{4.1}$$

where ω is a random parameter, let $X = X(u)$ be a quantity of interest of the model solution $u = u(\omega)$, i.e., $X : \omega \to \mathbb{R}$. The estimation of failure probability reads: given $y \in \mathbb{R}$ find $p \in [0,1]$ such that

$$p = \Pr(X \leq y) \tag{4.2}$$

holds. The estimation of p-quantiles is the inverse problem: given $p \in [0,1]$ find $y \in \mathbb{R}$ such that (4.2) holds. For simplicity we will only consider failure probabilitiy here, for p-quantiles see Paper VI. Because of the high stochastic dimension we consider MC approaches using different variance reduction techniques. The curse of dimensionality does not effected MC based methods. This is a consequence of the central limit theorem which states that mean value of a sequence consisting of independent and identically distributed random variables with size n, where the random variables have mean value μ and variance σ^2, tends to the normal distribution with mean μ and variance σ^2/n, independent of the stochastic dimension [36].

The key idea is to use a posteriori error estimates/bounds to improve existing MC methods and variance reduction techniques for MC methods. We will consider the MC method and multilevel Monte Carlo method (MLMC) [16, 5, 7].

4.1 Selective refinement

We want to estimate the probability that a quantity of interest X is below a critical value y. Let us define $Q = \mathbb{1}(X \leq y)$ where $\mathbb{1}(\text{true}) = 1$ and $\mathbb{1}(\text{false}) = 0$. Then the failure probability can be expressed as the expected value of Q, i.e., $p = \mathbb{E}[Q]$. Since the quantity of interest X is a functional of the model solution it needs to be approximated using some numerical method. We will make the following assumption on the numerical approximation X_ℓ of X. Let X_ℓ satisfy

$$|X - X_\ell| \leq \left(\frac{1}{2}\right)^\ell \quad \text{or} \quad |X - X_\ell| \leq |X_\ell - y|, \tag{4.3}$$

for all ℓ. The approximation of Q reads

$$Q_\ell = \mathbb{1}(X_\ell \leq y). \tag{4.4}$$

The Monte Carlo estimator using selective refinement reads

$$\widehat{Q}^{MC} = \frac{1}{N}\sum_{i=1}^{N} Q_L(\omega_i) \tag{4.5}$$

where ω_i is a realization of the model data and L is fixed. Compared to a standard Monte Carlo estimator where all samples are solved to the same tolerance, the selective refinement estimator only refines samples, to the finest tolerance level, that are close to the failure, see Figure 4.1. This can significantly reduce the cost since most samples are computed on coarse model resolution and hence at smaller cost than for the standard Monte Carlo estimator. See Paper VI for a discussion towards estimating p-quantiles and Paper VII for more elaborate discussion towards failure probability.

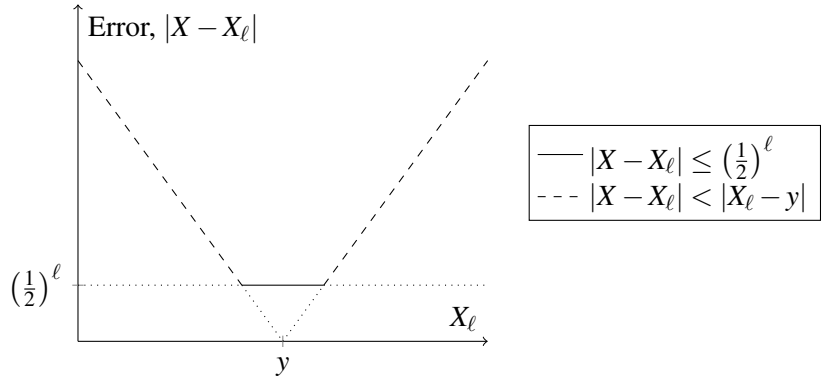


Figure 4.1. Illustration of the selective refinement condition (4.3). The numerical error is allowed to be larger far away from y.

4.2 Multilevel Monte Carlo with selective refinement

The MLMC method is a variance reduction technique that splits the estimator into different levels. On low levels many samples are used where the samples are cheap to compute and on high levels few samples are used where the samples are more expensive. The multilevel Monte Carlo estimator reads

$$\widehat{Q}^{ML} = \sum_{\ell=1}^{L} \frac{1}{N_\ell}\sum_{i=1}^{N_\ell} \left(Q_\ell(\omega_i) - Q_{\ell-1}(\omega_i)\right), \tag{4.6}$$

where $Q_0(\omega) = 0$. Note that

$$\mathbb{E}[\widehat{Q}^{ML}] = \sum_{\ell=1}^{L} \frac{1}{N_\ell} \sum_{i=1}^{N_\ell} (\mathbb{E}[Q_\ell(\omega_i)] - \mathbb{E}[Q_{\ell-1}(\omega_i)]) = \mathbb{E}[Q_L(\omega_i)] \tag{4.7}$$

because of the telescopic sum. The variance of the multilevel Monte Carlo estimator is

$$\mathbb{V}[\widehat{Q}^{ML}] = \sum_{\ell=1}^{L} \frac{1}{N_\ell} \mathbb{V}[Q_\ell(\omega_i) - Q_{\ell-1}(\omega_i)]. \tag{4.8}$$

The level dependent approximation Q_ℓ of the random variable Q satisfy

$$\begin{aligned} |\mathbb{E}[Q_\ell - Q]| &\leq C_1 \left(\frac{1}{2}\right)^\ell, \\ \mathbb{V}[Q_\ell - Q_{\ell-1}] &\leq C_2 \left(\frac{1}{2}\right)^\ell \end{aligned} \tag{4.9}$$

if (4.3) holds. For the root mean square error to satisfy

$$e(\widehat{Q}^{ML}) = \left(\mathbb{V}[\widehat{Q}^{ML}] + (\mathbb{E}[\widehat{Q}^{ML} - Q])^2\right)^{1/2} \leq \varepsilon, \tag{4.10}$$

for some tolerance ε, the total cost required to compute the MLMC estimator with selective refinement is

$$\text{Cost}\left(\widehat{Q}^{ML}\right) \leq C \begin{cases} N & q < 2, \\ \text{Cost}\,(Q_L) & q > 2. \end{cases} \tag{4.11}$$

where C is a generic constant independent of ε and q. The constant q typically depends on the dimension of the problem, convergence rate of the numerical approximation, and the linear solvers. This is a huge improvement compared to the standard Monte Carlo estimator which has the cost $N \cdot \text{Cost}\,(Q_L)$, i.e., the computational complexity for the MLMC estimator is either solving all problems att cost 1 or one problem at the highest cost. See Paper VII for a more elaborate discussion.

5. Future works

There is a rapid development in numerical techniques for both multiscale and uncertainty quantification problems. Some natural extension of the work considered in this thesis is the following.

- Combine the multiscale and uncertainty quantification algorithms. Many uncertainty quantification problems has multiscale structure.
- Use the selective refinement to create a multilevel subset simulation for rare event estimation, i.e., a small failure probabilities.
- Apply and extend the analysis to more realistic engineering problems. The numerical experiments in this thesis are based on model problems.

6. Summary of papers

6.1 Paper I

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).

In this paper we present an adaptive discontinuous Galerkin multiscale method driven by an energy norm a posteriori error bound. In the multiscale method the problem is split into a coarse and fine scale, $V_h = V_H^{\text{ms}} \oplus V^f$. Localized fine scale problems to correct the coarse basis are solved on truncated patches of the domain and are used to construct the coarse space, V_H^{ms}. The coarse space has considerably less degrees of freedom than the fine scale reference problem. The a posteriori error bound is used within an adaptive algorithm to tune the critical parameters,

$$|||u - u_H^{\text{ms}}||| \leq C_1 \left(\sum_{S \in \mathcal{T}_h} \rho_h^2(u_H^{\text{ms}}) \right)^{1/2} + C_2 \left(\sum_{T \in \mathcal{T}_H} \rho_{\omega_T}^2(u_H^{\text{ms}}) \right)^{1/2}, \qquad (6.1)$$

i.e., the error indicator of the refinement level ρ_h^2 and of the patch sizes $\rho_{\omega_T}^2$ on which the truncated fine scale problems are solved. The fine scale computations are completely parallelizable, since no communication between different processors is required when computing the basis for the multiscale space.

Contribution: The author of this thesis was main responsible for the analysis, writing, and performed all the numerical experiment. The idea was developed in close collaboration between the authors.

6.2 Paper II

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).

In this paper we derive a convergence result for a discontinuous Galerkin multiscale method for second order elliptic problems. We consider a heterogeneous and highly varying diffusion matrix in $L^\infty(\Omega, \mathbb{R}^{d\times d}_{sym})$ with uniform spectral bounds without any assumption on scale separation or periodicity. The multiscale method uses a space V_H^{ms} spanned by corrected basis that is constructed by correcting a coarse basis on truncated patches. The error, due to

truncation of the corrected basis, decreases exponentially with the size of the patches. Hence, we achieve the linear convergence rate

$$|||u - u_H^{\text{ms}}||| \leq H, \tag{6.2}$$

in energy norm of the multiscale solution on a uniform mesh with mesh size H, by choosing the patch sizes to be $\mathcal{O}(H|\log H|)$. Improved convergence rate can be achieved depending on the piecewise regularity of the forcing function. Also, quadratic convergence

$$||u - u_H^{\text{ms}}||_{L^2(D)} \leq H^2, \tag{6.3}$$

in the L^2-norm is obtained for arbitrary forcing function in L^2.

Contribution: The author of this thesis was main responsible for the analysis, writing, and performed all the numerical experiment. The idea was developed in close collaboration between the authors.

6.3 Paper III

D. Elfverson. *A Discontinuous Galerkin Multiscale Method for Convection-Diffusion Problems.* Available as arXiv:1509.03523 e-print (submitted).

In this paper we consider convection-diffusion problems with rough, heterogeneous, and highly varying coefficients. We propose a generalization of the discontinuous Galerkin multiscale method presented Paper II to convection-diffusion problems. The properties of the multiscale method and the discontinuous Galerkin method allow us to better cope with multiscale features as well as interior/boundary layers in the solution. The coarse trial and test spaces are corrected using fine scale computation on localized patches of size $\mathcal{O}(H\log(H^{-1}))$, where H is the mesh size. For convection-diffusion it is better to have directed patches, i.e., increase them in the direction of the convection. Linear convergence in energy norm,

$$|||u - u_H^{\text{ms}}||| \leq H, \tag{6.4}$$

is obtain under the assumption that the ratio between the size of the convection and diffusion coefficients scales like

$$\mathcal{O}\left(\frac{\|H\mathbf{b}\|_{L^\infty(\Omega)}}{\|A^{-1}\|_{L^\infty(\Omega)}}\right) \leq 1, \tag{6.5}$$

where $\mathbf{b}$ is the convection and A is the diffusion coefficient. However the convergence rates are independent of the variation in the coefficients.

Contribution: The author of this thesis is the sole author.

6.4 Paper IV

D. Elfverson, V. Ginting, and P. Henning. *On Multiscale Methods in Petrov-Galerkin Formulation,* Numer. Math. (2015).

In this work we investigate the advantages of multiscale methods in P-G formulation in a general framework which both fit multiscale methods based on the continuous and discontinuous Galerkin method. The framework splits the high dimensional reference space into a low dimensional corrected coarse space and high dimensional corrector space. The high dimensional corrector space only contains negligible fine scale information. The corrected coarse space V_H^{ms} can then be used to obtain accurate Galerkin approximations in P-G formulation: find $u_H^{\text{ms}} \in V_H^{\text{ms}}$ such hat

$$a(u_H^{\text{ms}}, v) = F(v) \quad \text{for all } v \in V_H. \tag{6.6}$$

Thus a Petrov-Galerkin formulation preservs the convergence rate,

$$|||u - u_H^{\text{ms}}||| \leq H, \tag{6.7}$$

with only a slightly lager constant compared to original symmetric multiscale method. However, P-G method can decrease the computational complexity significantly, allowing for more efficient solution algorithms. This makes the P-G method more preferable compared to the symmetric method. We prove inf-sup stability of a P-G continuous and a discontinuous Galerkin multiscale method. As another application of the framework, we show how the Petrov-Galerkin framework can be used to construct a locally mass conservative solver for two-phase flow simulation that employs the Buckley-Leverett equation. To achieve this, we couple a Petrov-Galekin discontinuous Galerkin finite element method with an upwind scheme for a hyperbolic conservation law.

Contribution: The author of this thesis was main responsible for the analysis, writing, numerical experiments regarding the discontinuous Galerkin part of the multiscale method.

6.5 Paper V

D. Elfverson, M. G. Larson, and A. Målqvist. *Multiscale Methods for Problems with Complex Geometry.* Available as arXiv:1509.03991 e-print (submitted).

In this paper we extend the analysis for the LOD to problems on complex domains, i.e., domains with voids, cracks, and complicated boundaries. The multiscale method uses a corrected test and trial space V_H^{ms}, where correctors for the basis function are computed on truncated patches. The correctors do

only need to be computed close to the boundary. We achieve linear convergence rate

$$|||u - u_H^{\text{ms}}||| \leq H, \tag{6.8}$$

in energy norm for the multiscale solution, even if the computational mesh does not resolve the fine features of the domain D. The conditioning of the multiscale method is not affected by how the domain boundary cuts the elements in the mesh.

Contribution: The author of this thesis was main responsible for the analysis, writing, and performed all the numerical experiments. The idea was developed in close collaboration between the authors.

6.6 Paper VI

D. Elfverson, D. J. Estep, F. Hellman, and A. Målqvist. *Uncertainty Quantification for Approximate p-Quantiles for Physical Models with Stochastic Inputs.* SIAM/ASA J. Uncertainty Quantification, 2(1), 826–850 (2014).

In this paper we consider the estimation of p-quantiles for a given functional evaluated on numerical solutions of a deterministic model in which model input is subject to stochastic variation. We derive upper and lower bounding estimators of the p-quantile, i.e., given p and $0 < \beta < 1$ find the computational bounds y^- and y^+ such that

$$\Pr(y \in [y^-, y^+]) > 1 - \beta. \tag{6.9}$$

The main idea is to perform an a posteriori error analysis for the p-quantile estimators that takes into account the effects of both the stochastic sampling error and the deterministic numerical solution error. We propose a selective refinement algorithm for computing an estimate of the p-quantile with a desired accuracy in a computationally efficient fashion. In the selective refinement only samples that can effect the p-quantile are refined, i.e., different samples are solved to different accuracy. Only a relatively small subset of samples significantly affects the accuracy of a p-quantile estimator and need to be solved to full accuracy. The algorithm leads to significant computational gain. For instance, if the numerical model is a first order discretization of a partial differential equation with spatial dimension greater than one, the reduction in computational work (compared to standard Monte Carlo using n samples) is asymptotically proportional to $n^{1/2}$.

Contribution: The author of this thesis did the writing and performed the numerical experiment in close collaboration with the third author. The analysis was done in close collaboration with the author of this thesis, the third, and the fourth author. The idea was developed in close collaboration between all the authors.

6.7 Paper VII

D. Elfverson, F. Hellman, and A. Målqvist. *A Multilevel Monte Carlo method for Computing Failure Probabilities.* Available as arXiv:1408.6856 e-print (submitted).

In this paper we propose and analyze a method for computing failure probabilities of systems modeled as numerical deterministic models (e.g., PDEs) with uncertain input data. Failure probability is defined as the probability that a functional of the solution to the model is below some critical value, i.e., given y find p such that

$$p = \Pr(X < y) \tag{6.10}$$

where X is a quantity of interest. By combining selective refinement with a multilevel Monte Carlo method we develop a method which reduces computational cost without loss of accuracy compared to the standard multilevel Monte Carlo method. We prove how the computational cost of the method relates to the root mean square error of the failure probability and is asymptotically proportional to solving a single accurate realization of the numerical model (independent of the number of samples) or a standard Monte Carlo method where all samples have cost 1 (independent of the numerical cost) which is the optimal rate.

Contribution: The author of this thesis did the analysis, writing, and performed the numerical experiment in close collaboration with the second author. The idea was developed in close collaboration between the authors.

7. Summary in Swedish

I denna avhandling fokuserar vi både på partiella differentialekvationer med data som varierar över flera olika skalor i rummet, *multiskalproblem*, samt som har osäkerhet i datat, *osäkerhetskvantifiering*. Modellering och simulering av denna typ av problem är mycket utmanande och förekommer i de flesta områden inom vetenskap och teknik. Några exempel är flöden i porösa medier och kompositmaterial. Vanliga numeriska metoder, t.ex enskaliga numeriska metoder samt Monte Carlo simuleringar för multiskal och osäkerhetkvantifieringsproblem är i många fall olämpliga och i andra fall omöjligt att använda på grund av deras höga kostnad. I denna avhandling behandlar vi problemen som dyker upp i multiskal- och osäkerhetskvantifieringsproblem separat.

Standardmetoderna för numeriska beräkningar fungerar dåligt för multiskalproblem när man har snabbt varierande data och när den finskaliga informationen i data inte löses upp av beräkningsnätet. Vi behandlar problem där både koefficienterna och beräkningsdomänen har multiskalstruktur. Den huvudsakliga utmaningen i att konstruera numeriska metoder för multiskalproblem är att minska beräkningskomplexiteten utan att förlora noggrannhet i lösningen. Vi utvecklar en multiskalmetod där grova basfunktioner som spänner upp lösningen korrigeras med hjälp av lokaliserade finskaliga beräkningar. Lösningen för multiskalmetoden har samma konvergenshastighet som standard metoderna har för problem utan multiskaldata. De finskaliga korrektionsproblemen avtar exponentiell bort från stödet av den ursprungliga basfunktionen och beräkningarna kan därför lokaliseras till små områden. Storleken av beräkningsområderna kan väljas så att konvergensen för multiskalmetoden inte påverkas. Korrektionsproblemen kan lösas helt oberoende av varandra, vilket gör metoden perfekt lämpad för parallella beräkning. Det är också möjligt att återanvända de finskaliga beräkningarna i t.ex. tidsstegning och icke-linjära iterationer.

I osäkerhetkvantifiering fokuserar vi på tillämpningar där modellparametrarna beror på stokastiska variationer. Vi vill beräkna statistiska egenskaper hos en kvantitet av lösningen till modellen, mer exakt så vill vi beräkna p-kvantiler och felsannolikheter. Felsannolikheter definieras som sannolikheten att en given kvantiteten av lösningen till modellen är mindre än något kritiskt värde. Uppskattningen av p-kvantiler är det inversa problemet, dvs bestäm värdet så att en givna kvantiteten av lösningen är större eller mindre än ett givet värde med sannolikhet p. Beräkningar av den här typen av problem har två felkällor, ett numeriskt fel från diskretiseringen av modellen och ett statistiskt fel från ett ändligt antal stickprov. För att uppskatta p-kvantiler eller

felsannolikheter effektivt så måste de båda felkällorna balanseras. Vi utvecklar tekniker för att uppskatta/beräkna det nummeriska felet tillammans med existerande variansreducerande metoder för att minska beräkningskostnaden samt för att balansera de båda felkällorna.

8. Acknowledgments

First and foremost I would like to give a huge thanks to my advisor Axel Målqvist for introducing me to the subject, for all his help and guidance, for sharing his contacts in the academic world, and support during my PhD. I would also like to thank all my co-authors Donald Estep, Emmanuil Georgoulis, Victor Ginting, Axel Målqvist, Fredrik Hellman, Patrick Henning, Mats Larson and Daniel Peterseim for all their help, ideas, advice, and expertise.

Thanks to all my fellow PhD students and friends here at the the department that made these years go by way too fast. There are a few people that I especially would like to mention. Fredrik Hellman for being great friend and office mate and for great collaborations and discussions. My old office mate Josefin Ahlkrona for being a great friend. To my awesome former and present flatmates Andreas Löscher and Slobodan Milovanović. Finally, I would like to thank my closest family Göran, Hanna, Ingrid, and Johannes. The list can be made much longer and even if you are not mentioned, I love you all.

References

[1] A. Abdulle and A. Nonnenmacher. A posteriori error analysis of the heterogeneous multiscale method for homogenization problems. *C. R. Math. Acad. Sci. Paris*, 347(17-18):1081–1086, 2009.

[2] R. A. Adams. *Sobolev spaces*. Academic Press, New York, 1975.

[3] D. N. Arnold. An interior penalty finite element method with discontinuous elements. *SIAM J. Numer. Anal.*, 19(4):pp. 742–760, 1982.

[4] I. Babuška and J. E. Osborn. Can a finite element method perform arbitrarily badly? *Math. Comp.*, 69(230):443–462, 2000.

[5] A. Barth, C. Schwab, and N. Zollinger. Multi-level Monte Carlo finite element method for elliptic PDEs with stochastic coefficients. *Numer. Math.*, 119(1):123–161, 2011.

[6] S. C. Brenner and L. R. Scott. *The Mathematical Theory of Finite Element Methods*. Springer Verlag, 1979.

[7] K. A. Cliffe, M. B. Giles, R. Scheichl, and A. L. Teckentrup. Multilevel Monte Carlo methods and applications to elliptic PDEs with random coefficients. *Comput. Vis. Sci.*, 14(1):3–15, 2011.

[8] J. Douglas and T. Dupont. Interior penalty procedures for elliptic and parabolic Galerkin methods. In *Computing methods in applied sciences (Second Internat. Sympos., Versailles, 1975)*, pages 207–216. Lecture Notes in Phys., Vol. 58. Springer, Berlin, 1976.

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

[10] W. E and B. Engquist. Multiscale modeling and computation. *Notices Amer. Math. Soc.*, 1(1):1062–1070, 2003.

[11] W. E and B. Engquist. The heterogeneous multi-scale method for homogenization problems. In *Multiscale methods in science and engineering*, volume 44 of *Lect. Notes Comput. Sci. Eng.*, pages 89–110. Springer, Berlin, 2005.

[12] 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.

[13] Y. R. Efendiev, T. Y. Hou, and X.-H. Wu. Convergence of a nonconforming multiscale finite element method. *SIAM J. Numer. Anal.*, 37(3):888–910, 2000.

[14] 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.

[15] 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.

[16] M. B. Giles. Multilevel Monte Carlo path simulation. *Oper. Res.*, 56(3):607–617, 2008.

[17] P. Henning and M. Ohlberger. The heterogeneous multiscale finite element method for elliptic homogenization problems in perforated domains. *Numer. Math.*, 113(4):601–629, 2009.
[18] J. S. Hesthaven and T. Warburton. *Nodal discontinuous Galerkin methods*, volume 54 of *Texts in Applied Mathematics*. Springer, New York, 2008.
[19] T. Hou, X.-H. Wu, and Z. Cai. Convergence of a multiscale finite element method for elliptic problems with rapidly oscillating coefficients. *Math. Comput.*, 68(227):913–943, July 1999.
[20] T. Y. Hou and X.-H. Wu. A multiscale finite element method for elliptic problems in composite materials and porous media. *J. Comput. Phys.*, 134(1):169–189, 1997.
[21] P. Houston and E. Süli. *hp*-adaptive discontinuous Galerkin finite element methods for first-order hyperbolic problems. *SIAM J. Sci. Comput.*, 23(4):1226–1252, 2001.
[22] T. Hughes. Multiscale phenomena: Green's functions, the Dirichlet-to-Neumann formulation, subgrid scale models, bubbles and the origins of stabilized methods. *Computer Methods in Applied Mechanics and Engineering*, 127(1-4):387–401, 1995.
[23] T. Hughes, G. Feijoo, L. Mazzei, and J.-B. Quincy. The variational multiscale method–a paradigm for computational mechanics. *Computer Methods in Applied Mechanics and Engineering*, 166(1-2):3 – 24, 1998.
[24] C. Johnson and J. Pitkäranta. An analysis of the discontinuous Galerkin method for a scalar hyperbolic equation. *Math. Comp.*, 46(173):1–26, 1986.
[25] 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(21-24):2313–2324, 2007.
[26] P. Lesaint and P.-A. Raviart. On a finite element method for solving the neutron transport equation. In *Mathematical aspects of finite elements in partial differential equations (Proc. Sympos., Math. Res. Center, Univ. Wisconsin, Madison, Wis., 1974)*, pages 89–123. Publication No. 33. Math. Res. Center, Univ. of Wisconsin-Madison, Academic Press, New York, 1974.
[27] A. Målqvist. Multiscale methods for elliptic problems. *Multiscale Modeling & Simulation*, 9(3):1064–1086, 2011.
[28] A. Målqvist and D. Peterseim. Localization of elliptic multiscale problems. *Math. Comp.*, 83(290):2583–2603, 2014.
[29] A. Målqvist and D. Peterseim. Computation of eigenvalues by numerical upscaling. *Numerische Mathematik*, 130(2):337–361, 2015.
[30] J. Nitsche. Über ein Variationsprinzip zur Lösung von Dirichlet-Problemen bei Verwendung von Teilräumen, die keinen Randbedingungen unterworfen sind. *Abh. Math. Sem. Univ. Hamburg*, 36:9–15, 1971.
[31] M. Ohlberger. A posteriori error estimates for the heterogeneous multiscale finite element method for elliptic homogenization problems. *Multiscale Model. Simul.*, 4(1):88–114 (electronic), 2005.
[32] D. Peterseim. Eliminating the pollution effect in Helmholtz problems by local subscale correction. *ArXiv e-prints*, Nov. 2014.
[33] D. A. Pietro and A. Ern. *Mathematical Aspects of Discontinuous Galerkin Methods*, volume 69 of *Mathématiques et Applications*. Springer, 2012.

[34] W. H. Reed and T. R. Hill. Triangular mesh methods for the neutron transport equation. Technical report, Los Alamos Scientific Laboratory, 1973.
[35] B. Rivière. *Discontinuous Galerkin methods for solving elliptic and parabolic equations*, volume 35 of *Frontiers in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2008.
[36] C. P. Robert and G. Casella. *Monte Carlo Statistical Methods (Springer Texts in Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2005.
[37] R. Verfürth. *A review of a posteriori error estimation and adaptive mesh-refinement techniques*. Wiley-Teubner series in advances in numerical mathematics. Wiley-Teubner, 1996.
[38] M. Wheeler. An elliptic collocation-finite element method with interior penalties. *SIAM Journal on Numerical Analysis*, 15(1):pp. 152–161, 1978.

Acta Universitatis Upsaliensis

Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology 1287

Editor: The Dean of the Faculty of Science and Technology

A doctoral dissertation from the Faculty of Science and Technology, Uppsala University, is usually a summary of a number of papers. A few copies of the complete dissertation are kept at major Swedish research libraries, while the summary alone is distributed internationally through the series Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology. (Prior to January, 2005, the series was published under the title "Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology".)

Distribution: publications.uu.se
urn:nbn:se:uu:diva-262354

ACTA
UNIVERSITATIS
UPSALIENSIS
UPPSALA
2015

Paper I

MULTISCALE MODEL. SIMUL.
Vol. 11, No. 3, pp. 747–765

AN ADAPTIVE DISCONTINUOUS GALERKIN MULTISCALE METHOD FOR ELLIPTIC PROBLEMS*

DANIEL ELFVERSON†, EMMANUIL H. GEORGOULIS‡, AND AXEL MÅLQVIST†

Abstract. An adaptive discontinuous Galerkin multiscale method driven by an energy norm a posteriori error bound is proposed. The method is based on splitting the problem into a coarse and fine scale. Localized fine scale constituent problems are solved on patches of the domain and are used to obtain a modified coarse scale equation. The coarse scale equation has considerably less degrees of freedom than the original problem. The a posteriori error bound is used within an adaptive algorithm to tune the critical parameters, i.e., the refinement level and the size of the different patches on which the fine scale constituent problems are solved. The fine scale computations are completely parallelizable, since no communication between different processors is required for solving the constituent fine scale problems. The convergence of the method, the performance of the adaptive strategy, and the computational effort involved are investigated through a series of numerical experiments.

Key words. multiscale, discontinuous Galerkin, a posteriori error bound

AMS subject classifications. 65N30, 65N15

DOI. 10.1137/120863162

1. Introduction.

Problems involving features on several different scales, usually termed multiscale problems, can be found in many branches of the engineering sciences. Examples include the modeling of flow in a porous medium and of composite materials. Multiscale problems involving partial differential equations are often impossible to simulate with an acceptable accuracy using standard (single mesh) numerical methods. A different approach, usually coming under the general term of multiscale methods, consists of considering coarse and fine scale contributions to the solution, with the fine scale contributions approximated on localized patches. The fine scale contributions are then used to upscale the problem in order to obtain an approximation to the global multiscale solution.

1.1. Previous work.

Numerous multiscale methods have been developed during the last three decades; see, e.g., [8, 7] for early works, or [16, 29, 15] and the references therein for exposition and recent developments. An important development is the Multiscale Finite Element Method (MsFEM) by Hou and Wu [21], which was further developed in [12], with the introduction of oversampling to reduce resonance effects. Another approach is the so-called Variational Multiscale method (VMS) of Hughes and co-workers [22, 23]. The idea in VMS is to decompose the solution space into coarse and fine scale contributions. A modified coarse scale problem is then solved (using a finite element approach), so that the fine scale contribution is taken into account. To maintain the conformity of the resulting modified finite element space, homogeneous Dirichlet boundary conditions are imposed on each fine-problem

*Received by the editors January 23, 2012; accepted for publication (in revised form) May 14, 2013; published electronically August 1, 2013.
http://www.siam.org/journals/mms/11-3/86316.html
†Department of Information Technology, Uppsala University, SE-751 05, Uppsala, Sweden (daniel.elfverson@it.uu.se, axel.malqvist@it.uu.se). The first author was supported by The Göran Gustafsson Foundation and The Swedish Research Council. The third author was supported by The Göran Gustafsson Foundation.
‡Department of Mathematics, University of Leicester, Leicester LE1 7RH, UK (emmanuil.georgoulis@le.ac.uk).

patch boundary. The Adaptive variational multiscale method (AVMS) using the VMS framework, introduced by Larson and Målqvist [27], makes use of multiscale-type a posteriori error bound to adapt the coarse and fine scale mesh sizes as well as the fine-problem patch-sizes automatically. A priori error analysis can be found in [30].

An interesting alternative to conforming finite element methods is the class of discontinuous Galerkin (DG) methods, whereby the approximation spaces are elementwise discontinuous; the continuity of the underlying exact solutions is imposed weakly. DG methods appeared in the 1970s and in the early 1980s [32, 28, 9, 5, 24] and have recently received renewed interest; we refer to the volumes [13, 14, 20, 33] and the references therein for a literature review. DG methods admit good conservation properties of the state variable and, due to the lack of interelement continuity requirements, are ideally suited for application to complex and/or irregular meshes. Also, there has been work to better cope with the case of high contrast diffusion; see e.g., [19] where a DG method based on weighted average is proposed and analyzed. Discontinuous Galerkin methods for solving multiscale problems have been discussed using the framework of the MsFEM [1] and of the Heterogeneous Multiscale Method (HMM) [2]; see also [37, 36, 35, 34]. An a priori error analysis for the class of discontinuous Galerkin multiscale method studied in this paper can be found in [17].

1.2. New contributions. In this work, we propose an *Adaptive Discontinuous Galerkin MultiScale method* (ADG-MS) using the framework of VMS. The underlying DG method is based on weighted averages across the element interfaces. The adaptivity is driven by energy norm a posteriori error bounds. The multiscale method is based on solving localized problems on patches, which are then upscaled to solve a coarse scale equation. The lack of any interelement continuity requirements of the approximate solution allows for very general meshes, which is very common in multiscale applications; i.e., meshes that contain several types of elements and/or hanging nodes. The split between the coarse and fine scale is realized using the elementwise L^2-projection onto the coarse mesh. This is more natural in a multiscale setting than, e.g., using the nodal interpolant as in [27]. It is also much easier and efficient to construct an L^2 orthogonal split using DG as opposed to conforming multiscale methods. The ADG-MS inherits a local conservation property from DG on the coarse scale, which is crucial in many applications such as porous media flow. The fine scale problems can be solved independently with localized right-hand sides, and it is known that the solutions decay exponentially [17], which allows for small patches. In this case the ADG-MS converges to the reference solution, thereby taking full advantage of cancellation between patches; this is not the case for the original AVMS [27] since hanging nodes are not allowed. In the a posteriori error bound, the error is bounded in terms of the size of the different fine-scale patches and on both the fine-scale and the coarse-scale mesh sizes. An adaptive algorithm to tune all these parameters automatically is proposed. The numerical experiments show good performance of the algorithm for a number of benchmark problems.

1.3. Outline. The rest of this work is structured as follows. Section 2 is devoted to setting up the model problem, the basic DG discretization, and some notation. A general framework for multiscale problems along with the discontinuous Galerkin multiscale method is derived in section 3, and the a posteriori error bound is derived in section 4. The implementation of the method and the adaptive algorithm are discussed in section 5. In section 6, a number of numerical experiments are presented, and finally some conclusions are drawn in section 7.

2. Preliminaries. In this section we define some notations and the underlying DG method is presented.

2.1. Notation. Let $\omega \subseteq \mathbb{R}^d$, $d = 2, 3$, be an open polygonal domain. Denote the $L^2(\omega)$-inner product by $(\cdot,\cdot)_{L^2(\omega)}$, and the corresponding norm by $\|\cdot\|_{L^2(\omega)}$. Also, let $H^1(\omega)$ be the Sobolev space with norm $\|\cdot\|_{H^1(\omega)} := (\|\cdot\|^2_{L^2(\omega)} + \|\nabla\cdot\|^2_{L^2(\omega)})^{1/2}$, and let $H^s(\omega)$ be the standard Hilbertian Sobolev space of index $s \in \mathbb{R}$. We shall also make use of the space $L^\infty(\omega)$ consisting of almost everywhere bounded functions, with norm $\|\cdot\|_{L^\infty(\omega)} := \operatorname{ess\,sup}_\omega|\cdot|$; see, e.g., [3] for details. Finally, the d-dimensional Lebesgue measure will be denoted by $\mu_d(\cdot)$.

2.2. The model problem. Let $\Omega \subset \mathbb{R}^d$ be an open polygonal domain with Lipschitz boundary $\partial\Omega$, $d = 2, 3$, and consider the elliptic boundary value problem find $u \in \{v \in H^1(\Omega) : v|_{\partial\Omega} = 0\}$ fulfilling

$$-\nabla \cdot A\nabla u = f, \qquad u \in \Omega, \tag{2.1}$$

$$u = 0, \qquad u \in \partial\Omega, \tag{2.2}$$

with $f \in L^2(\Omega)$ and $A \in L^\infty(\Omega, \mathbb{R}^{d,d}_{\mathrm{sym}})$ such that A has uniform spectral bounds, bounded below by $\alpha > 0 \in \mathbb{R}$ almost everywhere.

2.3. Discretization and subdivision. The domain Ω is subdivided into a partition $\mathcal{K} = \{K\}$ of shape-regular and closed elements K with boundaries ∂K; i.e., $\bar\Omega = \cup_{K\in\mathcal{K}}\bar K$. On the partition $\mathcal{K}$, let $h : \cup_{K\in\mathcal{K}}K \to \mathbb{R}$ be a mesh-function defined elementwise by $h|_K := \operatorname{diam}(K)$, $K \in \mathcal{K}$. The partition is allowed to be irregular (i.e., hanging nodes are allowed) and it is locally quasi uniform in the sense that the ratio of the mesh function h for neighboring elements is uniformly bounded from above and below. Let Γ^B be the set of all boundary edges, and let Γ^I be the set of all interior edges (or faces when $d = 3$) such that $\Gamma = \Gamma^B \cap \Gamma^I$ is the set of all edges in the partition $\mathcal{K}$. Associated with the diffusion tensor, we consider the elementwise constant functions $A^0, A_0 : \cup_{K\in\mathcal{K}}K \to \mathbb{R}$ defined by the biggest and smallest eigenvalue of A, respectively, on each element K. For $K_i, K_j \in \mathcal{K}$, with $\mu_{d-1}(\partial K_i \cap \partial K_j) > 0$, let K_i, K_j be denoted by K^+ and K^-, where K^+ is the element with the higher index. On interior element interfaces $e \in \Gamma^I$ we shall make use of the shorthand notation $v^+ := v|_{K^+}$, $v^- := v|_{K^-}$; on boundary edges we set $v^+ := v|_K$. We also define the weighted mean value by

$$\{v\}_w := w_{K^+(e)}v^+ + w_{K^+(e)}v^-, \tag{2.3}$$

where

$$w_{K^+(e)} := \frac{A^0|_{K^-}}{A^0|_{K^+} + A^0|_{K^-}}, \qquad w_{K^-(e)} := \frac{A^0|_{K^+}}{A^0|_{K^+} + A^0|_{K^-}} \tag{2.4}$$

for each $e \in \Gamma^I$ and

$$w_{K^+(e)} = 1, \qquad w_{K^+(e)} = 0 \tag{2.5}$$

for $e \in \Gamma^B$. Further, the jump across element interfaces is defined by

$$[v] := v^+ - v^- \text{ for } e \in \Gamma^I \qquad \text{and} \qquad [v] := v^+ \text{ for } e \in \Gamma^B, \tag{2.6}$$

and the harmonic mean value γ_e by

$$\gamma_e := \frac{2A^0|_{K^+} \cdot A^0|_{K^-}}{A^0|_{K^+} + A^0|_{K^-}}. \tag{2.7}$$

Also, n will denote the outward unit normal to ∂K^+ when $\mu_{d-1}(\partial K^+ \cap \partial K^-) > 0$. When $\mu_{d-1}(\partial K \cap \partial\Omega) > 0$, n will be the outward unit normal to $\partial\Omega$.

2.4. The discontinuous Galerkin method. For a nonnegative integer r, we denote by $\mathcal{P}_r(\hat{K})$ the set of all polynomials on $\hat{K}$ of total degree at most r if $\hat{K}$ is the reference d-simplex, or of degree at most r in each variable if $\hat{K}$ is the reference d-hypercube.

Consider the space $\mathcal{V} := \mathcal{V}_h + H^{1+\epsilon}(\Omega)$ with $\epsilon > 0$ but arbitrary small, and let the discontinuous finite element space be given by

$$\mathcal{V}_h := \{v \in L^2(\Omega) : v \circ F_K|_K \in \mathcal{P}_r(\hat{K}),\ \hat{K} \in \mathcal{K}\}, \tag{2.8}$$

where $F_K : \hat{K} \to K$ is the respective elemental map for $K \in \mathcal{K}$, which is allowed to be nonaffine, provided its Jacobian remains nonsingular and uniformly bounded from above and below with respect to all meshes.

The discontinuous Galerkin method then reads as follows: find $u_h \in \mathcal{V}_h$ such that

$$a(u_h, v) = \ell(v) \quad \forall v \in \mathcal{V}_h, \tag{2.9}$$

where the bilinear form $a(\cdot,\cdot) : \mathcal{V} \times \mathcal{V} \to \mathbb{R}$ and the linear form $\ell(\cdot) : \mathcal{V} \to \mathbb{R}$ are given by

$$\begin{aligned} a(v,z) := \sum_{K\in\mathcal{K}} (A\nabla v, \nabla z)_{L^2(K)} - \sum_{e\in\Gamma} \Big(&(n \cdot \{A\Pi\nabla v\}_w, [z])_{L^2(e)} \\ &+ (n \cdot \{A\Pi\nabla z\}_w, [v])_{L^2(e)} - \frac{\sigma_e\gamma_e}{h_e}([v],[z])_{L^2(e)}\Big), \end{aligned} \tag{2.10}$$

$$\ell(v) := (f, v)_{L^2(\Omega)}, \tag{2.11}$$

respectively. Here $\Pi : (L^2(\Omega))^d \to (\mathcal{V}_h)^d$ denotes the orthogonal L^2-projection operator onto $(\mathcal{V}_h)^d$, $h_e := \operatorname{diam}(e)$, and $\sigma_e \in \mathbb{R}$ is a positive constant. The bilinear form (2.11) is coercive with respect to the natural energy norm,

$$|||v||| = \left(\sum_{K\in\mathcal{K}} \|A^{1/2}\nabla v\|^2_{L^2(K)} + \sum_{e\in\Gamma} \frac{\sigma_e\gamma_e}{h_e} \|[v]\|^2_{L^2(e)} \right)^{1/2} \tag{2.12}$$

if σ_e is chosen to be large enough. We refer, e.g., to [14, 6] and the references therein for details on the analysis of DG methods for elliptic problems. Discontinuous Galerkin methods with weighted averages were introduced in [10, 19].

Remark 2.1. For all $v \in \mathcal{V}_h$, we have $\Pi\nabla v = \nabla v$; therefore, the bilinear form (2.10) with $v, z \in \mathcal{V}_h$ is reduced to the more familiar form

$$\begin{aligned} a(v,z) = \sum_{K\in\mathcal{K}} (A\nabla v, \nabla z)_{L^2(K)} - \sum_{e\in\Gamma} \Big(&(n \cdot \{A\nabla v\}_w, [z])_{L^2(e)} \\ &+ (n \cdot \{A\nabla z\}_w, [v])_{L^2(e)} - \frac{\sigma_e\gamma_e}{h_e}([v],[z])_{L^2(e)}\Big). \end{aligned} \tag{2.13}$$

3. The multiscale method. In the VMS framework, the finite element solution space $\mathcal{V}_h$ is decoupled into coarse and fine scale contributions, viz., $\mathcal{V}_h = \mathcal{V}_H \oplus \mathcal{V}_f$, with $\mathcal{V}_H \subset \mathcal{V}_h$. To this end, let $\Pi_H : L^2(\Omega) \to \mathcal{V}_H$ be the (orthogonal) L^2-projection onto the coarse mesh. The split between the coarse and fine scales is then determined by $\mathcal{V}_H := \Pi_H\mathcal{V}_h$ and $\mathcal{V}_f := (I - \Pi_H)\mathcal{V}_h = \{v \in \mathcal{V}_h : \Pi_H v = 0\}$, where I is the identity operator.

The multiscale map $\mathcal{T} : \mathcal{V}_H \to \mathcal{V}_f$ from the coarse to the fine scale is defined as

$$a(\mathcal{T}v_H, v_f) = -a(v_H, v_f) \quad \forall v_H \in \mathcal{V}_H \text{ and } \forall v_f \in \mathcal{V}_f. \tag{3.1}$$

The next step is to decompose u_h and v in (2.9) into coarse and fine scale components. In particular, we have

$$u_h = u_H + \mathcal{T}u_H + u_f, \tag{3.2}$$

and $v = v_H + v_f$, with $u_H, v_H \in \mathcal{V}_H$ and $\mathcal{T}u_H, v_f \in \mathcal{V}_f$ for some $u_f \in \mathcal{V}_f$. Equation (2.9) is equivalent to the following problem: find $u_H \in \mathcal{V}_H$ and $v_f \in \mathcal{V}_f$ such that

$$a(u_H + \mathcal{T}u_H + u_f, v_H + v_f) = \ell(v_H + v_f) \quad \forall v_H \in \mathcal{V}_H \text{ and } \forall v_f \in \mathcal{V}_f. \tag{3.3}$$

The fine scale component u_f can be computed by letting $v_H = 0$ in (3.3) and using the multiscale map (3.1). We obtain the fine scale problem driven by the right-hand side data f: find $u_f \in \mathcal{V}_f$ such that

$$a(u_f, v_f) = \ell(v_f) \quad \forall v_f \in \mathcal{V}_f. \tag{3.4}$$

The coarse scale solution is obtained by letting $v_f = 0$ in (3.3): find $u_H \in \mathcal{V}_H$ such that

$$a(u_H + \mathcal{T}u_H, v_H) = \ell(v_H) - a(u_f, v_H) \quad \forall v_H \in \mathcal{V}_H. \tag{3.5}$$

In (3.5), $\mathcal{T}v_H$ and u_f are unknown and obtained by solving (3.1) and (3.4). Note that the linear system (3.5) has $\dim(\mathcal{V}_H)$ unknowns.

3.1. Localization and discretization. The bilinear form is characterized by more local behavior in $\mathcal{V}_f$ than in $\mathcal{V}_h$ [30, 17]. This motivates us to solve the fine scale equations on (localized) overlapping patches, instead of the whole domain Ω. The patches are chosen large enough to ensure sufficiently accurate computations of $\mathcal{T}v_H$ and u_f. The computations of the fine scale components of the solution can be done in parallel with localized right-hand sides. To define the coarse space $\mathcal{V}_H$, we begin by fixing a coarse mesh $\mathcal{K}_H$. Then, $\mathcal{V}_H$ is defined as

$$\mathcal{V}_H := \{v \in L^2(\Omega) : v \circ F_K|_K \in \mathcal{P}_r(\hat{K}),\ \hat{K} \in \mathcal{K}_H\}. \tag{3.6}$$

DEFINITION 3.1. *For all $K \in \mathcal{K}_H$, define element patches of size L patch as*

$$\begin{aligned} \omega_K^1 &= int(K), \\ \omega_K^L &= int(\cup\{K' \in \mathcal{K}_H \mid K' \cap \bar{\omega}_K^L\}), \quad L = 2, 3, \ldots. \end{aligned} \tag{3.7}$$

The patch ω_K^L will be refered to as a L-layer patch. This is illustrated in Figure 1.

On each L-layer patch, we let $\mathcal{K}(\omega_K^L)$ be a restriction of $\mathcal{K}$ to ω_K^L, such that $\cup_{K \in \mathcal{K}(\omega_K^L)} = \bar{\omega}_K^L$. Also let $\Gamma^I(\omega_K^L)$ and $\Gamma^B(\omega_K^L)$ be the interior, respectively, boundary edges on $\mathcal{K}(\omega_K^L)$. Moreover, we assume that $\mathcal{K}_H|_{\omega_K^L}$ and $\mathcal{K}(\omega_K^L)$ are nested; that is, every coarse element $K_H \in \mathcal{K}_H|_{\omega_K^L}$ coincides with a union of fine elements $K \in \mathcal{K}(\omega_K^L)$. Also, the fine test spaces $\mathcal{V}_f(\omega_K^L)$ are defined by

$$\mathcal{V}_f(\omega_K^L) := \{v \in \mathcal{V}_f : v|_{\Omega \setminus \omega_K^L} = 0\}. \tag{3.8}$$

Finally, let the indicator function be $\chi_K = 1$ on element K and 0 otherwise, and let $\mathcal{M}_K$ be the index set of all basis functions $\phi_j \in \mathcal{V}_H$ that have support on K; i.e., $\chi_K = \sum_{j \in \mathcal{M}_K} \phi_j$.

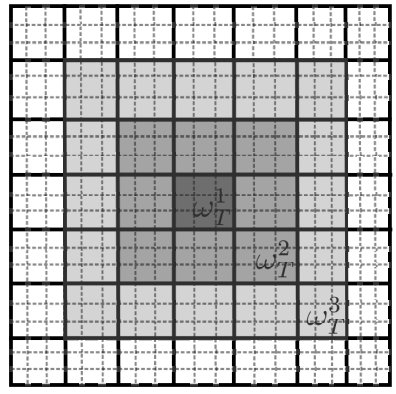


FIG. 1. *Example of one ω_K^1, two ω_K^2, and three ω_K^3 layer patches around element T in a quadrilateral mesh.*

3.2. The discontinuous Galerkin multiscale method. For each $K \in \mathcal{K}_H$ the following local problems need to be solved: find $\tilde{\mathcal{T}}\phi_j \in \mathcal{V}_f(\omega_K^L)$ $\forall j \in \mathcal{M}_K$ and $U_{f,K} \in \mathcal{V}_f(\omega_K^L)$ such that

$$a(\tilde{\mathcal{T}}\phi_j, v_f) = -a(\phi_j, v_f) \quad \forall v_f \in \mathcal{V}_f(\omega_K^L), \tag{3.9}$$

$$a(U_{f,K}, v_f) = \ell(\chi_K v_f) \quad \forall v_f \in \mathcal{V}_f(\omega_K^L). \tag{3.10}$$

The modified coarse scale problem is formulated as follows: find $U_H \in \mathcal{V}_H$ such that

$$a(U_H + \tilde{\mathcal{T}}U_H, v_H) = \ell(v_H) - a(U_f, v_H), \quad \forall v_H \in \mathcal{V}_H, \tag{3.11}$$

where $U_f := \sum_{K \in \mathcal{K}_H} U_{f,K}$. The approximate solution to the multiscale problem is given by

$$U = U_H + \tilde{\mathcal{T}}U_H + U_f. \tag{3.12}$$

The above procedure will be referred to as the *discontinuous Galerkin multiscale method*.

We note that the approximation U is *not* equal to u_h, in general, since the domains of the fine scale problems are truncated. However, as discussed above, it is expected that U is a good approximation to u_h, due to the decaying nature of the fine scale solutions away from the respective patch. For the approximation U to converge to the exact solution u of (2.1) in the limit, the support of the local problems should be gradually extended to both the whole computational domain and the fine scale meshsize h should converge to 0. The multiscale method proposed here differs from the one proposed in [17] in that a right-hand side correction is present. Using the formulation without the presence of a right-hand side correction, the multiscale solution converges to an H-perturbation of the exact solution u, uniformly with respect to the diffusion coefficient structure.

Remark 3.2. Note that for a nonuniform mesh $\mathcal{K}$ (and/or $\mathcal{K}_H$), the convergence results presented in [17] still hold if the corrected basis functions are computed on patches of a common reference mesh $\mathcal{K}$. On the other hand, if the adaptive algorithm

is used so that the overlap between different corrected basis functions are computed on different meshes (cf., e.g., [27]), less cancellation of the error will occur and convergence can no longer be guaranteed by the argument in [17].

3.3. Local conservation property. The DG methods are known to have good local conservation properties in that the normal fluxes are conservative. The ADGMS inherits this property on the coarse scale. To see this, we introduce the normal fluxes on element $K_H \in \mathcal{K}_H$ as

$$\hat{\sigma}(U) := (\{n \cdot A\nabla U\}_w - \sigma_e \gamma_e h_e^{-1}[U])[\chi_{K_H}], \quad e \in \partial K_H, \tag{3.13}$$

where $U = U_H + \tilde{\mathcal{T}} U_H + U_f$, $\chi_{K_H} = 1$ on element K_H and $\chi_{K_H} = 0$ otherwise ($[\chi_{K_H}]$ is either 1 or -1), and each interface e is a face of a fine scale element $K \in \mathcal{K}$, i.e., the number of edges can exceed the number of faces for each element K_H. By setting $w \in \mathcal{V}_H$ to be $w = \chi_{K_H}$ in (2.10), (2.11), and by using the discrete normal fluxes defined in (3.13), we arrive to the discrete elementwise conservation law

$$(f, 1)_{L^2(K_H)} + (\hat{\sigma}(U), 1)_{L^2(\partial K_H)} = 0 \tag{3.14}$$

for all $K_H \in \mathcal{K}_H$.

4. A posteriori error bound in energy norm. Let the constant $0 \leq C < \infty$ be any generic constant neither depending on H, h, L, nor A; let $a \lesssim b$ abbreviate the inequality $a \leq Cb$. The following approximation results will be used frequently throughout this section. Let π be the orthogonal L^2-projection operator onto elementwise constant functions. Then π satisfies the following approximation properties: for an element K, we have

$$||v - \pi v||_{L^2(K)} \lesssim \frac{h_K}{\sqrt{A_0}} ||A^{1/2} \nabla v||_{L^2(K)} \quad \forall v \in H^1(K), \tag{4.1}$$

$$||v - \pi v||_{L^2(\partial K)} \lesssim \sqrt{\frac{h_K}{A_0}} ||A^{1/2} \nabla v||_{L^2(K)} \quad \forall v \in H^1(K). \tag{4.2}$$

LEMMA 4.1. *Let $\mathcal{I}_h^c : \mathcal{V}_h \to \mathcal{V}_h \cap H^1(\Omega)$ be a averaging interpolation operator defined pointwise as*

$$\mathcal{I}_h^c v_h(\tilde{x}) = \frac{1}{|\mathcal{K}_{\tilde{x}}|} \sum_{K \in \mathcal{K}_{\tilde{x}}} v_h(\tilde{x})|_K, \tag{4.3}$$

where $\mathcal{K}_{\tilde{x}}$ is the set of elements in $\mathcal{K}$ for which $\tilde{x}$ belong, with the cardinal $|\mathcal{K}_{\tilde{x}}|$. Then,

$$||v_h - \mathcal{I}_h^c v_h||_{L^2(K)}^2 \lesssim ||\sqrt{h_e}[v_h]||_{L^2(\partial K)}^2, \tag{4.4}$$

$$||A^{1/2} \nabla (v_h - \mathcal{I}_h^c v_h)||_{L^2(K)}^2 \lesssim A^0 \left|\left| \frac{1}{\sqrt{h_e}} [v_h] \right|\right|_{L^2(\partial K)}^2 \tag{4.5}$$

holds for all $v_h \in \mathcal{V}_h$ and $K \in \mathcal{K}$.

The proof, omitted here, follows closely that of [25]. Lemma 4.1 can also be extended to irregular meshes. There a hierarchical refinement of the mesh is performed to eliminate the hanging nodes; we refer to [26] for details. For irregular meshes the constant in the bounds of Lemma 4.1 also depends on the number of hanging nodes on each face.

Remark 4.2. The result in Lemma 4.1 can be sharpened if the diffusion tensor is isotropic and a locally quasi-monotone [31] distribution is assumed to hold. Then $A^0|_K$ can be replaced by the harmonic mean value γ_e on face e; see [11].

First we derive a posteriori error bound for the underlying (one scale) DG method.

THEOREM 4.3. *Let u, u_h be given by* (2.1)–(2.2) *and* (2.9)*, respectively. Also let $\mathcal{I}_h^c u_h \in \mathcal{V}_h \cap H^1(\Omega)$ be given by* (4.3)*. Moreover, let $\mathcal{E} := \mathcal{E}_c + \mathcal{E}_d$, where $\mathcal{E}_c := u - \mathcal{I}_h^c u_h$ and $\mathcal{E}_d := \mathcal{I}_h^c u_h - u_h$. Then*

$$|||\mathcal{E}||| \lesssim \left(\sum_{K \in \mathcal{K}} \varrho_K^2 \right)^{1/2} + \left(\sum_{K \in \mathcal{K}} \zeta_K^2 \right)^{1/2}, \tag{4.6}$$

where

$$\varrho_K = \frac{h_K}{\sqrt{A_0}} ||(1-\Pi)(f + \nabla \cdot A \nabla u_h)||_{L^2(K)} + \sqrt{\frac{h_K}{A_0}} \left(||(1 - w_{K(e)}) n \cdot [A \nabla u_h]||_{L^2(\partial K \setminus \Gamma^B)} + \left|\left| \frac{\sigma_e \gamma_e}{h_e} [u_h] \right|\right|_{L^2(\partial K)} \right), \tag{4.7}$$

$$\zeta_K^2 = ||A^{1/2} \nabla (u_h - \mathcal{I}_h^c u_h)||_{L^2(K)}^2 + \left|\left| \sqrt{\frac{\sigma_e \gamma_e}{h_e}} [u_h] \right|\right|_{L^2(\partial K)}^2. \tag{4.8}$$

Remark 4.4. Using $\mathcal{I}_h^c u_h$ as the conforming part of u_h, we arrive to an a posteriori bound whereby $\mathcal{I}_h^c u_h$ can either be evaluated directly or bounded using Lemma 4.1. Another possible choice is a weighted averaging interpolation operator with the weights depending on the diffusion tensor [4].

Remark 4.5. Concerning the lower efficiency bounds, the term (4.7) is robust with respect to the diffusion tensor; see [18]. But to prove that (4.8) is robust with respect to the diffusion tensor, to the best of the authors' knowledge, the diffusion tensor has to be isotropic and satisfy a locally quasi-monotone property [31, 11].

Proof. Note that

$$|||\mathcal{E}||| \leq |||\mathcal{E}_c||| + |||\mathcal{E}_d|||, \tag{4.9}$$

where the first part can be bounded by

$$|||\mathcal{E}_c|||^2 \lesssim a(\mathcal{E}_c, \mathcal{E}_c) = a(\mathcal{E}, \mathcal{E}_c) - a(\mathcal{E}_d, \mathcal{E}_c) \lesssim a(\mathcal{E}, \mathcal{E}_c) + |||\mathcal{E}_d||| |||\mathcal{E}_c|||. \tag{4.10}$$

Let π_h be the L^2-orthogonal projection onto the elementwise constant functions and define $\eta := \mathcal{E}_c - \pi_h \mathcal{E}_c$. We then have

$$a(\mathcal{E}, \mathcal{E}_c) = a(u, \mathcal{E}_c) - a(u_h, \mathcal{E}_c) = \ell(\mathcal{E}_c) - a(u_h, \mathcal{E}_c) = \ell(\eta) - a(u_h, \eta), \tag{4.11}$$

which implies

$$|||\mathcal{E}_c|||^2 = a(\mathcal{E}_c, \mathcal{E}_c) = \big(\ell(\eta) - a(u_h, \eta)\big) - a(\mathcal{E}_d, \mathcal{E}_c). \tag{4.12}$$

Upon integration by parts and using the identity $[vw] = \{v\}_w [w] + \{w\}_{\bar{w}} [v]$, where $\bar{w}$ is the skew-weighted average given by

$$\{v\}_{\bar{w}} := w_{K^-(e)} v^+ + w_{K^+(e)} v^-, \tag{4.13}$$

the first term on the right-hand side of (4.12) yields

$$\begin{aligned}(4.14)\qquad &\ell(\eta)-a(u_h,\eta)\\ &=\sum_{K\in\mathcal{K}}(f+\nabla\cdot A\nabla u_h,\eta)_{L^2(K)}+\sum_{e\in\Gamma}\Big(-(n\cdot[A\nabla u_h],\{\eta\}_{\bar{w}})_{L^2(e\setminus\Gamma^B)}\\ &\quad+(n\cdot\{A\Pi\nabla\eta\}_w,[u_h])_{L^2(e)}-\sigma\gamma_e h_e^{-1}([u_h],[\eta])_{L^2(e)}\Big).\end{aligned}$$

The first term on the right-hand side of (4.14) can be bounded as follows:

$$\sum_{K\in\mathcal{K}}(f+\nabla\cdot A\nabla u_h,\eta)_{L^2(K)}\lesssim\sum_{K\in\mathcal{K}}\frac{h_K}{\sqrt{A_0}}||(1-\Pi)(f+\nabla\cdot A\nabla u_h)||_{L^2(K)}||A^{1/2}\nabla\mathcal{E}_c||_{L^2(K)},$$

using (4.1). The second term on the right-hand side of (4.14) gives

$$\begin{aligned}(4.15)\qquad &\sum_{e\in\Gamma\setminus\Gamma^B}(n\cdot[A\nabla u_h],\{\eta\}_{\bar{w}})_{L^2(e)}\\ &\lesssim\sum_{K\in\mathcal{K}}\sqrt{\frac{h_K}{A_0}}||(1-w_{K(e)})n\cdot[A\nabla u_h]||_{L^2(\partial K\setminus\Gamma^B)}||A^{1/2}\nabla\mathcal{E}_c||_{L^2(K)}\end{aligned}$$

using (4.2). For the third term on the right-hand side of (4.14), noting that $\nabla\eta=\nabla\mathcal{E}_c$, we deduce that

$$\sum_{e\in\Gamma}(n\cdot\{A\Pi\nabla\mathcal{E}_c\}_w,[\mathcal{E}_d])_{L^2(e)}\lesssim\sum_{K\in\mathcal{K}}\frac{1}{\sqrt{h_KA_0}}||\gamma_e[\mathcal{E}_d]||_{L^2(\partial K)}||A^{1/2}\nabla\mathcal{E}_c||_{L^2(K)},$$

using an inverse estimate and the L^2-stability of Π. For the last term on the right-hand side of (4.14), we have

$$\sum_{e\in\Gamma}\frac{\sigma_e\gamma_e}{h_e}([u_h],[\eta])_{L^2(e)}\lesssim\sum_{K\in\mathcal{K}}\sqrt{\frac{h_K}{A_0}}\left|\left|\frac{\sigma_e\gamma_e}{h_e}[u_h]\right|\right|_{L^2(\partial K\setminus\Gamma^B)}||A^{1/2}\nabla\mathcal{E}_c||_{L^2(K)}.$$

The last term on the right-hand side of (4.12) is bounded using the continuity of the bilinear form. Combining all of the above bounds and using Lemma 4.1 to bound the nonconforming part, the result follows. ∎

A posteriori error estimate for the ADG-MS is given below.

THEOREM 4.6. *Let u, U be defined in* (2.1)–(2.2) *and* (3.12)*, respectively, and set $\mathcal{I}_h^cU\in H^1(\Omega)$. Set $\mathscr{E}:=\mathscr{E}_c+\mathscr{E}_d$, where $\mathscr{E}_c:=u-\mathcal{I}_h^cU$ and $\mathscr{E}_d:=\mathcal{I}_h^cU-U$. Define $U_{K_H}:=\sum_{j\in\mathcal{M}_{K_H}}U_j(\phi_j+\tilde{\mathcal{T}}\phi_j)+U_{f,K_H}$, where U_j are the nodal values calculated by* (3.11) *for all $\mathcal{K}_H$. Then, $\mathscr{E}$ satisfies the estimate*

$$(4.16)\qquad |||\mathscr{E}|||\lesssim\left(\sum_{K\in\mathcal{K}}\varrho_K^2\right)^{1/2}+\left(\sum_{K\in\mathcal{K}}\zeta_K^2\right)^{1/2}+\left(\sum_{K_H\in\tilde{\mathcal{K}}_H}\rho^2_{\omega_{K_H}^L}\right)^{1/2},$$

where

$$(4.17)\quad\rho^2_{\omega_{K_H}^L}=\sum_{e\in\Gamma^B(\omega_{K_H}^L)}\left(\frac{H^2_{K_H^O}}{h_{K^O}A_0|_{K_H^O}}\right)\left(||n\cdot\{A\nabla U_i\}_w||_{L^2(e)}+\frac{\sigma_e\gamma_e}{h_e}||[U_i]||_{L^2(e)}\right)^2$$

measures the effect of the truncated patches, K^O, K_H^O are from outside of $\omega_{K_H}^L$, and

$$\varrho_K = \frac{h_K}{\sqrt{A_0}}||(1-\Pi)(f+\nabla\cdot A\nabla U)||_{L^2(K)} + \sqrt{\frac{h_K}{A_0}}\left(||(1-w_{K(e)})n\cdot[A\nabla U]||_{L^2(\partial K)} + \left|\left|\frac{\sigma_e\gamma_e}{h_e}[U]\right|\right|_{L^2(\partial K)}\right), \tag{4.18}$$

$$\zeta_K^2 = ||\sqrt{A}\nabla(U-\mathcal{I}_h^c U)||_{L^2(K)}^2 + \left|\left|\sqrt{\frac{\sigma_e\gamma_e}{h_e}}[U]\right|\right|_{L^2(\partial K)}^2 \tag{4.19}$$

measuring the refinement level of the fine scale.

Remark 4.7. One possible adaptive strategy would be to refine the coarse mesh as much one can afford, using a standard a posteriori error bound (e.g., using Theorem 4.3), and then further improve the approximation using Theorem 4.6. Note that fine scale problems do not have to be solved everywhere.

Remark 4.8. For the estimator $\rho_{\omega_{K_H}^L}$ to retain its optimality with respect to the mesh-sizes, one should assume that $H_{K_H}^2 \lesssim h_K$. We note that this is not an unreasonable requirement, for, otherwise, each fine scale problem would be more expensive to solve than the coarse scale problem.

Proof. Using the same idea as in Theorem 4.3. We first note that

$$|||\mathscr{E}_c|||^2 = a(\mathscr{E}_c,\mathscr{E}_c) = a(\mathscr{E},\mathscr{E}_c) - a(\mathscr{E}_d,\mathscr{E}_c). \tag{4.20}$$

Then, using (2.9) and the fine scale equations (3.9)–(3.10), we have

$$a(\mathscr{E},\mathscr{E}_c) = \ell(\mathscr{E}_c) - a(U,\mathscr{E}_c) \tag{4.21}$$

$$= \ell(\mathscr{E}_c - v_H) - a(U,\mathscr{E}_c - v_H) \tag{4.22}$$

$$= \ell(\mathscr{E}_c - v_H - v_f) - a(U,\mathscr{E}_c - v_H - v_f) + \ell(v_f) - a(U,v_f) \tag{4.23}$$

for any $v_H \in \mathcal{V}_H$ and $v_f \in \mathcal{V}_f$. Note that,

$$\ell(v_f) - a(U,v_f) = \sum_{K_H\in\tilde{\mathcal{K}}_H} \ell(\chi_{K_H}v_f) - a(U_{K_H},v_f) \tag{4.24}$$

$$= \sum_{K_H\in\tilde{\mathcal{K}}_H}\sum_{e\in\Gamma^B(\omega_{K_H}^L)}\Big((n\cdot\{A\nabla U_i\}_w,[\xi_{K_H}^L v_f])_{L^2(e)} + (n\cdot\{A\nabla\xi_{K_H}^L v_f\}_w,[U_i])_{L^2(e)} - \frac{\sigma_e\gamma_e}{h_e}([U_i],[\xi_{K_H}^L v_f])_{L^2(e)}\Big), \tag{4.25}$$

where $\xi_{K_H}^L = 0$ on $\omega_{K_H}^L$ and $\xi_{K_H}^L = 1$ otherwise; that is, $v_f = \xi_{K_H}^L v_f + (1-\xi_{K_H}^L)v_f$, where $(1-\xi_{K_H}^L)v_f \in \mathcal{V}_f(\omega_{K_H}^L)$. Then, applying (4.25), we deduce

$$\begin{aligned} a(\mathscr{E},\mathscr{E}_c) &= \Big(\ell(\mathscr{E}_c - v_H - v_f) - a(U,\mathscr{E}_c - v_H - v_f)\Big) \\ &\quad + \sum_{K_H\in\tilde{\mathcal{K}}_H}\sum_{e\in\Gamma^B(\omega_{K_H}^L)}\Big((n\cdot\{A\nabla U_i\}_w,[\xi_{K_H}^L v_f])_{L^2(e)} \\ &\quad + (n\cdot\{A\nabla\xi_{K_H}^L v_f\}_w,[U_i])_{L^2(e)} - \frac{\sigma_e\gamma_e}{h_e}([U_i],[\xi_{K_H}^L v_f])_{L^2(e)}\Big) \\ &=: I + II. \end{aligned} \tag{4.26}$$

Term I can be estimated as in the proof of Theorem 4.3, upon selecting $v_H := \pi_H \mathscr{E}_c$ and $v_f = \pi_f(\mathscr{E}_c - \pi_H \mathscr{E}_c) = \pi_f \mathscr{E}_c$, where π_H and π_f are the elementwise constant L^2-orthogonal projections onto the coarse space $\mathcal{V}_H$ on the fine space $\mathcal{V}_f$, respectively. We note that, by construction, $\pi_f \pi_H v = 0 \ \forall v \in \mathcal{V}_h$.

Since v_f is chosen to be piecewise constant the second term in II is equal to zero. For each $K \in \mathcal{K}$, and for each $e \in \Gamma^B(\omega^L_{K_H}) \backslash \Gamma^B$, we have

$$\begin{aligned} &\left|(n \cdot \{A\nabla U_i\}_w, [\xi^L_{K_H} v_f])_{L^2(e)} - \frac{\sigma_e \gamma_e}{h_e}([U_i], [\xi^L_{K_H} v_f])_{L^2(e)}\right| \\ &\quad \lesssim \left(||n \cdot \{A\nabla U_i\}_w||_{L^2(e)} + \frac{\sigma_e \gamma_e}{h_e}|||[U_i]|||_{L^2(e)}\right)|||[\xi^L_{K_H} v_f]|||_{L^2(e)} \end{aligned} \tag{4.27}$$

using (4.28) and the Cauchy–Schwarz inequality, for $e \in \Gamma^B$, the first term in (4.27) disappears. Note that, $|||[\xi^L_{K_H} v_f]|||_{L^2(e)}$ is either $|||[v_f^+]|||_{L^2(e)}$ or $|||[v_f^-]|||_{L^2(e)}$ depending on $\xi^L_{K_H}$. To bound the term involving v_f, for simplicity let v_f be either v_f^+ or v_f^-. We note that

$$\begin{aligned} ||v_f||_{L^2(e)} &\lesssim \frac{1}{\sqrt{h_K}}\|v_f\|_{L^2(K)} \lesssim \frac{1}{\sqrt{h_K}}\|v_f\|_{L^2(K_H)} \\ &\lesssim \frac{1}{\sqrt{h_K}}||\mathscr{E}_c - \pi_H \mathscr{E}_c||_{L^2(K_H)} \lesssim \frac{H_{K_H}}{\sqrt{h_K}}||\nabla \mathscr{E}_c||_{L^2(K_H)} \\ &\lesssim \frac{H_{K_H}}{\sqrt{h_K A_0}}||\sqrt{A}\nabla \mathscr{E}_c||_{L^2(K)} \end{aligned} \tag{4.28}$$

using a trace inequality and the L^2-stability of π_f, viz., $||\pi_f v||_{L^2(K_H)} \leq ||v||_{L^2(K_H)}$.

Combining the above and summing over all patches, using the discrete version of the Cauchy–Schwarz inequality, the proof is concluded. ∎

5. Implementation and adaptivity. The system of equations arising from the discretization of the modified coarse multiscale problem (3.11) is given by

$$KU = b - d, \tag{5.1}$$

where $K_{i,j} = a(\phi_j + \tilde{\mathcal{T}}\phi_j, \phi_i)$, $b_i = \ell(\phi_i)$, and $d_i = a(U_f, \phi_i)$. To assemble the right- and left-hand sides of (5.1), $\tilde{\mathcal{T}}\phi_i$ and $U_{f,i}$ need to be computed for all $i \in \mathcal{N}$. This can be done in parallel since no communication is required between the different fine scale problems. For each fine scale problem it is also possible to assemble $K_{i,j} = a(\phi_j + \tilde{\mathcal{T}}\phi_j, \phi_i)$, $b_i = \ell(\phi_i)$, and $d_i = \sum_{j \in \mathcal{N}} a(U_{f,j}, \phi_i)$ for a fixed i and for all j such that $\mu_d(\text{supp}(\phi_j) \cap \bar{\omega}_K) > 0$. The constraints needed on the fine scale test spaces to solve $\tilde{\mathcal{T}}\phi_i$ and $U_{f,i}$ are $\mathcal{V}_f = \{v \in \mathcal{V}^h : \Pi_H v = 0\}$, which are implemented using Lagrange multipliers. The spaces $\mathcal{V}_f$ and $\mathcal{V}_H$ are orthogonal with respect to the L^2-inner product.

Let $\mathcal{V}_H = \text{span}\{\phi_i\}$ and $\mathcal{V}_f = \text{span}\{\varphi_i\}$. Then, the system of equations to be solved on the fine scale is given by

$$\begin{pmatrix} K & P^T \\ P & 0 \end{pmatrix} \xi = \begin{pmatrix} b \\ 0 \end{pmatrix}, \tag{5.2}$$

where

$$P = \begin{pmatrix} (\phi_1, \varphi_1) & (\phi_1, \varphi_2) & \dots & (\phi_1, \varphi_N) \\ (\phi_2, \varphi_1) & (\phi_2, \varphi_2) & \dots & (\phi_2, \varphi_N) \\ \vdots & \vdots & \ddots & \vdots \\ (\phi_M, \varphi_1) & (\phi_M, \varphi_2) & \dots & (\phi_M, \varphi_N) \end{pmatrix}, \tag{5.3}$$

with $K_{k,l} = a_i(\varphi_k, \varphi_l)$ and b either $b_k = l_i(\varphi_k)$ for (3.10) or $b_k = -a_i(\phi_i, \varphi_k)$ for (3.9).

Using the a posteriori error estimate above, it is possible to design an adaptive algorithm that automatically tunes the fine mesh size and the size of the patches. In the numerical experiments below, we have implemented Algorithm 1, which extends the patches in all directions and uses a uniform mesh refinement of the fine scale on each coarse element. A more elaborate algorithm, which only extends in the direction where the error is large and uses adaptive mesh refinement, would be a possible extension, since the a posteriori indicators above contain local contributions of each individual patch-boundary face and of each fine scale element residual.

Algorithm 1. Adaptive discontinuous Galerkin multiscale method.

Initialize the coarse mesh, $\mathcal{K}_H$ with mesh function H, and a fine mesh, $\mathcal{K}_h$ with meshfunction h, by using to uniform refinements of $\mathcal{K}_H$; i.e., $h = H/4$.
For all $\mathcal{K}_H$ let the size of the patches be $\omega^3_{K_H}$.
Set the mesh refinement level to $X\%$.
while $(\sum_{K\in\mathcal{K}} \varrho^2_{h,K})^{1/2} + (\sum_{K\in\mathcal{K}} \zeta^2_{h,K})^{1/2} + (\sum_{K_H\in\mathcal{K}_H} \rho^2_{\omega^L_{K_H}})^{1/2} > TOL$ **do**
    for $K \in \tilde{\mathcal{K}}_H$ **do**
        Solve the fine scale problems (3.1) and (3.10).
        Compute the matrix and vector entries on the coarse scale (5.1).
    end for
    Solve the modified coarse scale problem (3.11).
    Mark the indicator with $X\%$ largest error in $\{\varrho^2_{h,K} + \zeta^2_{h,K}, \rho^2_{L,\omega_i}\}$.
    for $K_H \in K_H$ **do**
        if ρ^2_{L,ω_i} is marked **then**
            $\omega^L_{K_H} := \omega^{L+1}_{K_H}$
        end if
        if $\rho^2_{h,K} + \zeta^2_{h,K}$ is marked **then**
            $h|_{K_H} := h|_{K_H}/2$
        end if
    end for
end while

6. Numerical examples. We present some numerical experiments where the converge of the method as well as the performance of the adaptive algorithm is investigated.

6.1. Convergence. We consider the model problem (2.1)–(2.2) on the L-shaped domain constructed by removing the lower right quadrant in the unit square, with forcing function $f = 1$. We consider a coarse quadrilateral mesh of size $H = 2^{-4}$. Furthermore, each coarse element $K \in \mathcal{K}_H$ is further subdivided using two uniform refinements to construct the fine mesh. The error is measured in the relative energy norm, (2.12), where u_h is the DG solution on the fine mesh; i.e., there is only a truncation error (due to the fine scale patch size) between the multiscale solution and the DG solution. The permeabilities *One* and *SPE*,[1] illustrated in Figure 2, are used. In *One*, we have $A = 1$, and in *SPE* the data is taken from the tenth SPE comparative solution project and is projected into the fine mesh. Exponential decay is observed with respect to the number of layers for the different permeabilities *One* and *SPE*,

[1] Data is taken from the tenth SPE comparative solution project http://www.spe.org/web/csp/.

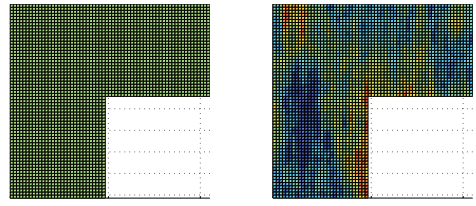


FIG. 2. *Permeability structure of* One *and* SPE *in log scale on an L-shaped domain.*

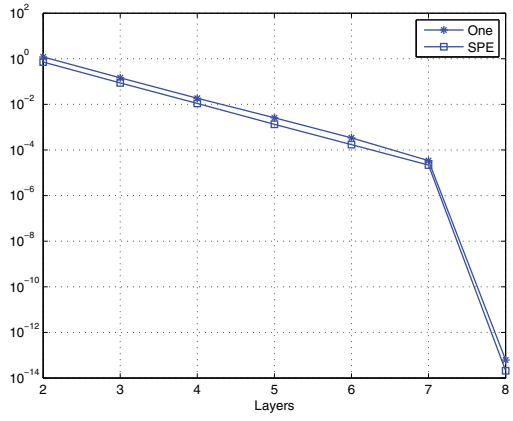


FIG. 3. *Convergence in relative energy norm* (2.12)*, on an L-shaped domain when the number of layers are increased using the different permeabilities* One *and* SPE.

until the patches cover the whole domain when $L = 8$; this is illustrated in Figure 3. As expected, when $L = 8$, only round off error between the multiscale solution and the reference solution is observed. Note that, by including the right-hand side fine scale correction, convergence of U to u_h itself is observed.

6.2. Adaptivity for a problem with analytic solution. Let us consider the model problem (2.1)–(2.2) on a unit square, using the permeability $A = 1$ and the forcing function $f = 4a^2(1 - ar^2)e^{-ar^2}$ for some $a > 0$. Using $a = 400$, the analytic solution can be approximated sufficiently well by the Gaussian pulse $u = ae^{-ar^2}$, centered in the middle of the domain. We consider a coarse quadrilateral mesh of size $H = 2^{-4}$, and a fine mesh of size $h = 2^{-6}$. The adaptive algorithm (Algorithm 1) with 10% refinement level is used. The starting values for L and h used are $L = 3$ layers, and the fine scale mesh is uniformly refined two times. Figure 4 shows the error and the error indicators decay after each iteration of the adaptive algorithm, while Figure 5 shows the locations where the adaptive algorithm has chosen to concentrate the computational effort, which indeed coincides with the position of the pulse.

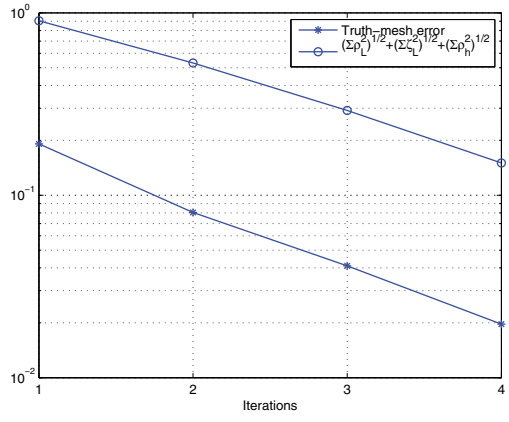


FIG. 4. *Convergence in relative energy norm* (2.12)*, using the adaptive algorithm on the unit square with Gaussian pulse in the middle.*

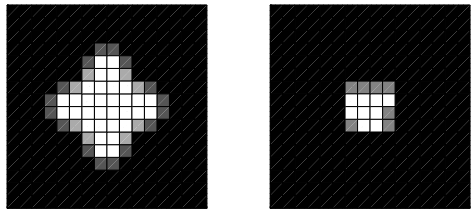


FIG. 5. *The level of refinement and the size of the patches L illustrated in the left, respectively right plots, using the adaptive algorithm on a unit square with Gaussian pulse in the middle. White is where the most refinements, respectively bigger L, are used and black is where the least refinements, respectively smallest patches, are used.*

6.3. Adaptivity on an *L*-shaped domain. Consider the model problem and the same data as in section 6.1. The solution produced by the adaptive algorithm is compared to a reference solution computed with the standard (one scale) DG method on a uniform quadrilateral mesh with mesh-size $h = 2^{-9}$; see Figure 6. Consider a coarse mesh consisting of a uniform quadrilateral mesh of size $H = 2^{-4}$. The starting values in the adaptive algorithm (Algorithm 1) are $L = 3$ and the fine scale mesh is derived by two uniform refinements of the coarse mesh. In each iteration, a refinement level of 30% is used. Figure 7 shows the error decays after each iteration of the adaptive algorithm. Also, the adaptive algorithm chooses to increase the patches in the beginning since the error from the truncation is initially larger than the discretization error and after a few iterations it is starting to refine the fine scale mesh more and more. When the patch sizes are increased, the error, due to truncation, decays exponentially independent of the regularity of the solution as shown theoretically in [17]. This is not true for the discretization error. This motivates the use of an adaptive algorithm which tunes the error between the truncation and discretization. Figure 8

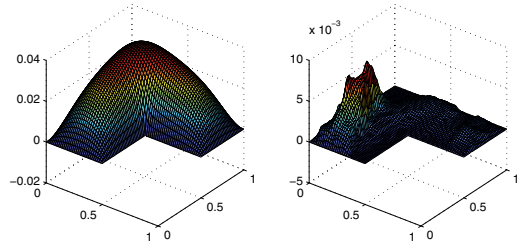


FIG. 6. *Reference solution for the different permeabilities computed onto a mesh with size* $h = 2^{-9}$ *and projected onto a mesh with size* $h = 2^{-6}$.

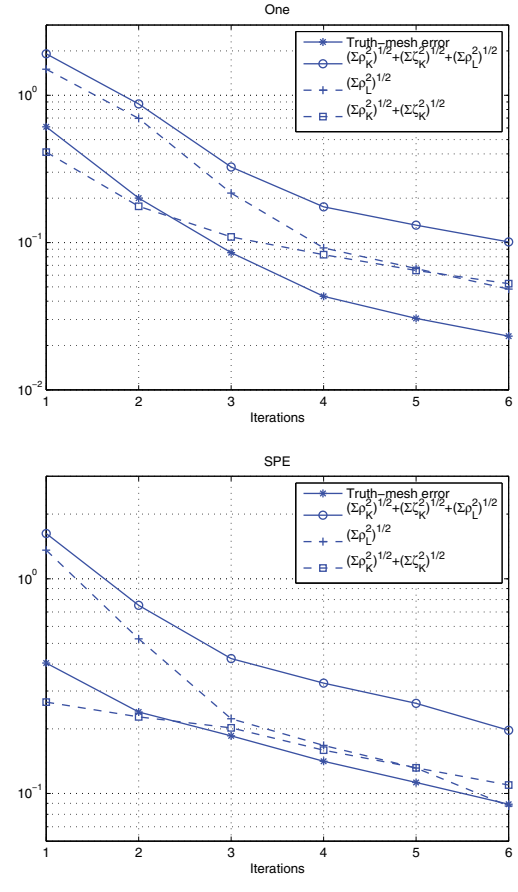


FIG. 7. *The relative energy norm error for the multiscale solution using the adaptive algorithm;* ρ_L *denotes the truncation error indicator and* ϱ_K *and* ζ_K *are the discretization error indicators.*

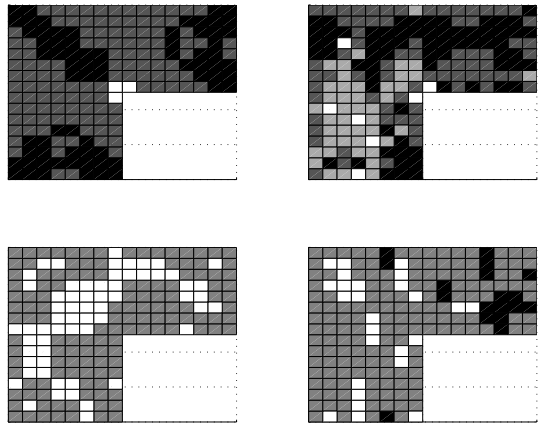
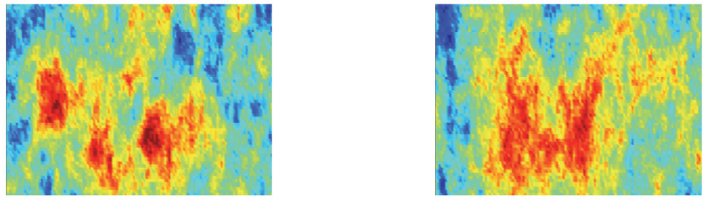


FIG. 8. *The level of refinement and size of the patches illustrated in the upper, respectively lower plots, for the different permeability One (left) and SPE (right). White is where most refinements, respectively larger patch, are used and black is where least refinements, respectively smallest patches, are used.*



(a) SPE11, $\alpha_{max}/\alpha_{min} = 6.1765e-5$. (b) SPE21, $\alpha_{max}/\alpha_{min} = 5.0193e-5$.

FIG. 9. *Permeabilities projection in log scale.*

shows where the adaptive algorithm put the most computational effort.

6.4. Adaptivity for a porous media flow problem. We consider the problem (2.1)–(2.2) on the unit square $\Omega = [0,1]^2$, with forcing function $f = -1$ in the lower left corner $\{0 \leq x, y \leq 1/128\}$, $f = 1$ in the upper right corner $\{127/128 \leq x, y \leq 1\}$, and $f = 0$ otherwise. The following permeabilities $SPE11$ and $SPE21$ are used and projected into a mesh with 64×64 elements; see Figure 9. The computational domain Ω is split into 32×32 coarse square elements $K_H \in \mathcal{K}_H$. The error is measured in the relative energy norm, with the reference solution u_h being the DG solution computed on a 512×512-element mesh. The adaptive algorithm (Algorithm 1) with refinement level 30% is used. In Iteration 1 the multiscale problem is solved using two refinements on each coarse element and each fine scale problem is solved with $L = 3$, and so on. Even though complicated permeabilities with $\alpha_{max}/\alpha_{min} \sim 10^5$ are used, the proposed adaptive algorithm is able to reduce relative error considerably; this is shown in Figure 10.

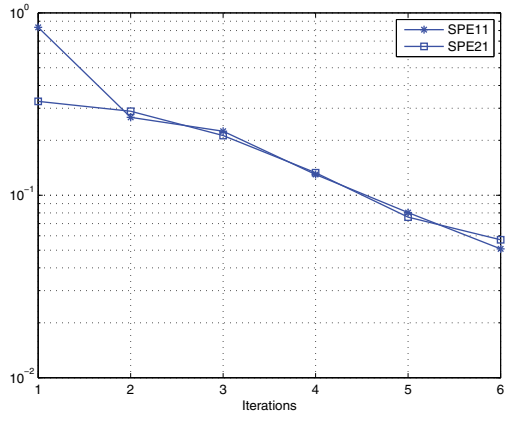


FIG. 10. *Relative in error broken energy norm against the number of iterations using the adaptive algorithm for flow in porous media.*

7. Concluding remarks. An adaptive multiscale method based on discontinuous Galerkin discretization has been proposed and assessed in practice. There are several different advantages in using the proposed multiscale method. The possibility to allow a global underlying reference grid (using the DG framework including hanging nodes) is crucial. This does not only account for cancellation of the error between different fine scale problems in the a posteriori error bound, it also fits the method into the convergence framework presented in [17]. It admits a local conservation of the state variable, which is crucial in many applications, e.g., porous media flow. The multiscale method and the adaptive algorithm admit naturally parallel implementation, which results in further savings in computational time.

An adaptive algorithm for which the coarse scale, the fine scale, and the size of the different patches are taken into account, based on an energy norm a posteriori bound has been proposed. Using the proposed multiscale method, together with the adaptive algorithm, leads to substantial computational savings, while maintaining a good performance when applied to challenging benchmark problems.

REFERENCES

[1] J. AARNES AND B.-O. HEIMSUND, *Multiscale discontinuous Galerkin methods for elliptic problems with multiple scales*, in Multiscale Methods in Science and Engineering, Lect. Notes Comput. Sci. Eng. 44, Springer, Berlin, 2005, pp. 1–20.

[2] A. ABDULLE, *Discontinuous Galerkin finite element heterogeneous multiscale method for elliptic problems with multiple scales*, Math. Comp., 81 (2012), pp. 687–713.

[3] R. A. ADAMS, *Sobolev Spaces*, Academic Press, New York, 1975.

[4] M. AINSWORTH, *Robust a posteriori error estimation for nonconforming finite element approximation*, SIAM J. Numer. Anal., 42 (2005), pp. 2320–2341.

[5] D. N. ARNOLD, *An interior penalty finite element method with discontinuous elements*, SIAM J. Numer. Anal., 19 (1982), pp. 742–760.

[6] D. N. ARNOLD, F. BREZZI, B. COCKBURN, AND L. D. MARINI, *Unified analysis of discontinuous Galerkin methods for elliptic problems*, SIAM J. Numer. Anal., 39 (2002), pp. 1749–1779.

[7] I. BABUŠKA, G. CALOZ, AND J. E. OSBORN, *Special finite element methods for a class of second order elliptic problems with rough coefficients*, SIAM J. Numer. Anal., 31 (1994),

pp. 945–981.

[8] I. Babuška and J. E. Osborn, *Generalized finite element methods: Their performance and their relation to mixed methods*, SIAM J. Numer. Anal., 20 (1983), pp. 510–536.

[9] G. A. Baker, *Finite element methods for elliptic equations using nonconforming elements*, Math. Comp., 31 (1977), pp. 45–59.

[10] E. Burman and P. Zunino, *A domain decomposition method based on weighted interior penalties for advection-diffusion-reaction problems*, SIAM J. Numer. Anal., 44 (2006), pp. 1612–1638.

[11] Z. Cai, X. Ye, and S. Zhang, *Discontinuous Galerkin finite element methods for interface problems: A priori and a posteriori error estimations*, SIAM J. Numer. Anal., 49 (2011), pp. 1761–1787.

[12] J. Chu, Y. Efendiev, V. Ginting, and T. Y. Hou, *Flow based oversampling technique for multiscale finite element methods*, Advances in Water Resources, 31 (2008), pp. 599–608.

[13] B. Cockburn, G. E. Karniadakis, and C.-W. Shu, *The development of discontinuous Galerkin methods*, in Discontinuous Galerkin Methods (Newport, RI, 1999), Lect. Notes Comput. Sci. Eng. 11, Springer, Berlin, 2000, pp. 3–50.

[14] D. A. Di Pietro and A. Ern, *Mathematical Aspects of Discontinuous Galerkin Methods*, Math. Appl. (Berlin) 69, Springer, Heidelberg, 2012.

[15] W. E, *Principles of Multiscale Modeling*, Cambridge University Press, Cambridge, UK, 2011.

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

[17] D. Elfverson, E. H. Georgoulis, A. Målqvist, and D. Peterseim, *Convergence of a discontinuous Galerkin multiscale method*, arXiv:1211.5524, 2012.

[18] A. Ern and A. F. Stephansen, *A posteriori energy-norm error estimates for advection-diffusion equations approximated by weighted interior penalty methods*, J. Comput. Math., 26 (2008), pp. 488–510.

[19] A. Ern, A. F. Stephansen, and P. Zunino, *A discontinuous Galerkin method with weighted averages for advection-diffusion equations with locally small and anisotropic diffusivity*, IMA J. Numer. Anal., 29 (2009), pp. 235–256.

[20] J. S. Hesthaven and T. Warburton, *Nodal Discontinuous Galerkin Methods. Algorithms, Analysis, and Applications*, Texts Appl. Math. 54, Springer, New York, 2008.

[21] 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.

[22] 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.

[23] T. J. R. Hughes, G. 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.

[24] C. Johnson and J. Pitkäranta, *An analysis of the discontinuous Galerkin method for a scalar hyperbolic equation*, Math. Comp., 46 (1986), pp. 1–26.

[25] O. A. Karakashian and F. Pascal, *A posteriori error estimates for a discontinuous Galerkin approximation of second-order elliptic problems*, SIAM J. Numer. Anal., 41 (2003), pp. 2374–2399.

[26] O. A. Karakashian and F. Pascal, *Adaptive discontinuous Galerkin approximation of secound-order elliptic problems*, in Proceedings of the European Congress on Computational Methods in Applied Sciences and Engineering ECCOMAS 2004, P. Neittaanmäki et al., eds., Jyväsklä, 2004.

[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), pp. 2313–2324.

[28] P. Lasaint and P.-A. Raviart, *On a finite element method for solving the neutron transport equation*, in Mathematical Aspects of Finite Elements in Partial Differential Equations (Proc. Sympos., Math. Res. Center, Univ. Wisconsin, Madison, Wis., 1974), Math. Res. Center, Univ. of Wisconsin-Madison, Academic Press, New York, 1974, pp. 89–123.

[29] A. Målqvist, *Multiscale methods for elliptic problems*, Multiscale Model. Simul., 9 (2011), pp. 1064–1086.

[30] A. Målqvist and D. Peterseim, *Localization of elliptic multiscale problems*, arXiv:1110.0692, 2011.

[31] M. Petzoldt, *A posteriori error estimators for elliptic equations with discontinuous coefficients*, Adv. Comput. Math., 16 (2002), pp. 47–75.

[32] W. H. Reed and T. R. Hill, *Triangular Mesh Methods for the Neutron Transport Equation*, Technical report, Los Alamos Scientific Laboratory, Los Alamos, NM, 1973.
[33] B. Rivière, *Discontinuous Galerkin Methods for Solving Elliptic and Parabolic Equations. Theory and Implementation*, Frontiers Appl. Math. 35, SIAM, Philadelphia, 2008.
[34] W. Wang, *Multiscale Discontinuous Galerkin Methods and Applications*, ProQuest LLC, Ann Arbor, MI, 2008, Ph.D. thesis, Brown University, Providence, RI.
[35] W. Wang, J. Guzmán, and C.-W. Shu, *The multiscale discontinuous Galerkin method for solving a class of second order elliptic problems with rough coefficients*, Int. J. Numer. Anal. Model., 8 (2011), pp. 28–47.
[36] L. Yuan and C.-W. Shu, *Discontinuous Galerkin method based on non-polynomial approximation spaces*, J. Comput. Phys., 218 (2006), pp. 295–323.
[37] L. Yuan and C.-W. Shu, *Discontinuous Galerkin method for a class of elliptic multi-scale problems*, Internat. J. Numer. Methods Fluids, 56 (2008), pp. 1017–1032.

Paper II

SIAM J. NUMER. ANAL.
Vol. 51, No. 6, pp. 3351–3372

CONVERGENCE OF A DISCONTINUOUS GALERKIN MULTISCALE METHOD*

DANIEL ELFVERSON†, EMMANUIL H. GEORGOULIS‡, AXEL MÅLQVIST†, AND DANIEL PETERSEIM§

Abstract. We present a discontinuous Galerkin multiscale method for second order elliptic problems and prove convergence. We consider a heterogeneous and highly varying diffusion coefficient in $L^\infty(\Omega, \mathbb{R}^{d\times d}_{sym})$ with uniform spectral bounds without any assumption on scale separation or periodicity. The multiscale method uses a corrected basis that is computed on patches/subdomains. The error, due to truncation of the corrected basis, decreases exponentially with the size of the patches. Hence, to achieve an algebraic convergence rate of the multiscale solution on a uniform mesh with mesh size H to a reference solution, it is sufficient to choose the patch sizes $\mathcal{O}(H|\log H|)$. We also discuss a way to further localize the corrected basis to elementwise support. Improved convergence rate can be achieved depending on the piecewise regularity of the forcing function. Linear convergence in energy norm and quadratic convergence in the L^2-norm is obtained independently of the forcing function. A series of numerical experiments confirms the theoretical rates of convergence.

Key words. multiscale method, discontinuous Galerkin, a priori error estimate, convergence

AMS subject classifications. 65N12, 65N30

DOI. 10.1137/120900113

1. Introduction.

This work considers the numerical solution of second order elliptic problems with a heterogeneous and highly varying (nonperiodic) diffusion coefficient. The heterogeneities and oscillations of the coefficient may appear on several nonseparated scales. More specifically, let $\Omega \subset \mathbb{R}^d$ be a bounded Lipschitz domain with polygonal boundary Γ. The boundary Γ may be partitioned into some subset Γ_D (the Dirichlet boundary) with positive measure and its complement $\Gamma_N := \Gamma \setminus \Gamma_D$ (the, possibly empty, Neumann boundary). We assume that the diffusion matrix $A \in L^\infty\left(\Omega, \mathbb{R}^{d\times d}_{\mathrm{sym}}\right)$ has uniform spectral bounds $0 < \alpha, \beta < \infty$, defined by

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

Given $f \in L^2(\Omega)$, we seek the weak solution to the boundary-value problem

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

*Received by the editors November 26, 2012; accepted for publication (in revised form) October 3, 2013; published electronically December 11, 2013.
http://www.siam.org/journals/sinum/51-6/90011.html

†Information Technology, Uppsala University, Uppsala, SE-751 05, Sweden (daniel.elfverson@it.uu.se, axel.malqvist@it.uu.se). These authors were supported by the Swedish Research Council and the Göran Gustafsson Foundation.

‡Department of Mathematics, University of Leicester, Leicester, LE1 7RH, UK (emmanuil.georgoulis@le.ac.uk).

§Institut für Mathematik, Humboldt-Universität zu Berlin, Berlin 10099, Germany (peterseim@math.hu-berlin.de). This author was supported by the DFG Research Center Matheon Berlin through project C33.

i.e., we seek $u \in H^1_D(\Omega) := \{v \in H^1(\Omega) \mid v|_{\Gamma_D} = 0\}$ such that

$$a(u,v) := \int_\Omega A\nabla u \cdot \nabla v \,\mathrm{d}x = \int_\Omega fv \,\mathrm{d}x =: F(v) \quad \text{for all } v \in H^1_D(\Omega). \tag{1.2}$$

Many methods have been developed in recent years to overcome the lack of performance of classical finite element methods when A is rough, meaning that A has discontinuities and/or high variation; we refer to [6, 4, 19, 8, 1, 2], among others. Common to all the aforementioned approaches is the idea to solve problems on small subdomains and to use the results to construct a better basis for some Galerkin method or to modify the coarse scale operator. However, apart from the one-dimensional setting, the performance of those methods correlates strongly with periodicity and scale separation of the diffusion coefficient. There has also been work to design a hierarchical basis such that the multigrid convergence rate does not depended on the variation in the coefficients, e.g., [29], where they assume that the diffusion coefficient fulfills a so-called quasi-monotone property.

Other approaches [7, 28, 5, 11, 12] perform well without any assumptions on periodicity or scale separation in the diffusion coefficient at the price of a high computational cost: in [7, 28] the support of the modified basis functions is large and in [5, 11, 12] the computation of the basis functions involves the solutions of local eigenvalue problems.

In the framework of the variational multiscale method (VMS), introduced in [21, 22], the space for which the solution is sought is split into a coarse and a fine scale contribution. Writing the fine scale contribution in terms of the coarse scale residuals eliminates it from the coarse scale equation. This was first employed in an adaptive setting in the adaptive VMS [24], where the basic idea is to split the fine scale residuals into localized contributions solved on element patches, possibly larger than a single element, with the Dirichlet boundary condition. Using the solution from the fine scale patches a modified nonsymmetric (Petrov–Galerkin) formulation is obtained on the coarse scale. An a posteriori error bound is derived and used within an adaptive algorithm to automatically tune the coarse and fine mesh size as well as the size of the patches.

An abstract framework for constructing multiscale methods for elliptic partial differential equations using the VMS framework is derived in [27]. Both symmetric and nonsymmetric (Petrov–Galerkin) formulations are considered and an a posteriori error bound is derived both for convection-diffusion-reaction problems and for the Poisson's equations on mixed form.

Only recently in [25] the first rigorous a priori error bound for a VMS was derived, which allows for textbook convergence with respect to the mesh size H, $\|u - u_H\|_{H^1(\Omega)} \leq C_{f,\beta/\alpha} H$ with a constant $C_{f,\beta/\alpha}$ that depends on f and the global bounds of the diffusion coefficient but not on its variations. This result, which is a natural extension of [24, 27], is achieved using a local orthogonal decomposition technique, where an operator dependent modification of the classical nodal basis is constructed using the solution of local problems on vertex patches of diameter $\mathcal{O}(H|\log H|)$. In the error analysis, the size of the patches depends on the global bound of the diffusion coefficient. This indicates that it may not be suited for high contrast (or degenerate) problems; however, numerical experiments show promising results also for these cases [25]. The methodology has been extended to semilinear elliptic problems [16] and (non)linear eigenvalue problems [26, 15]. In [17] it is shown that the approach may

also be interpreted as a multiscale finite element method, in the sense of [18], with some novel oversampling strategy.

In this work, we present a discontinuous Galerkin (dG) multiscale method with similar performance as [25]. We show that the error between the exact solution and the solution obtained by the dG multiscale method converge as $\tilde{C}_{f,\beta/\alpha}H$ in standard dG energy norm for $f \in L^2(\Omega)$. Higher convergence rates (up to $\tilde{C}_{f,\beta/\alpha}H^3$) can be obtained depending on the elementwise regularity of f. We also give an error bound for a quantity of interest (a linear functional of the solution) and the convergence rate $C'_{f,\beta/\alpha}H^2$ (up to $C'_{f,\beta/\alpha}H^4$) in the L^2-norm follows. Adaptivity for the dG multiscale method is considered in [13] and an extension of the a priori analysis to convection dominated convection-diffusion-reaction problems is considered in [14]. Since the dG method seeks the solution in a nonconforming space, the elementwise L^2-projection as the split between the coarse and fine contribution is now admissible. This is a more natural choice than, e.g., the nodal interpolant used in [24] for multiscale applications and may lead to better performance of the dG-based multiscale method (compared to conforming variants) for eigenvalue computations [26, 15].

The dG finite element method admits good conservation properties of the state variable and also offers the use of very general meshes due to the lack of interelement continuity requirements, e.g., meshes that contain several different types of elements and/or hanging nodes. Both these features are crucial in many multiscale applications.

Although the error analysis presented in this work is restricted to regular simplicial or quadrilateral/hexahedral meshes, we stress that all the results appear to be extendable for the case of irregular meshes (i.e., meshes containing hanging nodes). We refrained from presenting these extensions here for simplicity of the current presentation. Under these assumptions, we provide a complete a priori error analysis of this method including errors caused by the approximation of basis functions.

In this dG multiscale method and in previous related methods [25, 13], the accuracy is ensured by enlarging the support of basis functions appropriately. Hence, supports of basis functions overlap and the communication is no longer restricted to neighboring elements but is present also between elements at a certain distance. This overlap leads to a slight decrease of sparsity of the coarse stiffness matrix. We will show that the overhead is acceptable in the sense that it scales only logarithmically with respect to the coarse mesh size.

In order to retain the dG-typical sparse structure of the stiffness matrix with communication restricted to neighboring elements only, we discuss the possibility of localizing the multiscale basis functions to single elements. Instead of having $\mathcal{O}(1)$ basis functions per element with $\mathcal{O}(H|\log H|)$ support, we would then have $\mathcal{O}(|\log H|)$ basis functions per element with element support. The elementwise application of an eigenvalue decomposition easily prevents ill-conditioning of the element stiffness matrices, while simultaneously offering further compression of the multiscale basis.

The outline of the paper is as follows. In section 2, we recall the dG finite element method. Section 3 defines our multiscale method, which is then analyzed in section 4. Section 5 presents numerical experiments confirming the theoretical developments. Finally, in section 6 we draw some conclusions.

Throughout this paper, standard notation for Lebesgue and Sobolev spaces is employed. Let $0 \leq C < \infty$ be any generic constant that depends on neither the mesh size nor the diffusion matrix A; $a \lesssim b$ abbreviates an inequality $a \leq C\,b$ and $a \approx b$ abbreviates $a \lesssim b \lesssim a$. Also, let the constant $C_{\beta/\alpha}$ depend on the minimum and maximum bounds (α and β) of the diffusion matrix A in (1.1).

2. Fine scale discretization.

2.1. Finite element meshes and spaces. Let $\mathcal{T}$ denote a subdivision of Ω into (closed) regular simplices or into quadrilaterals (for $d = 2$) or hexahedra (for $d = 3$), i.e., $\bar{\Omega} = \cup_{T\in\mathcal{T}} T$. We assume that $\mathcal{T}$ is conforming in the sense that any two elements are either disjoint or share exactly one edge or vertex.

Let $\mathcal{E}$ denote the set of edges (or faces for $d = 3$) of $\mathcal{T}$; $\mathcal{E}(\Omega)$ denotes the set of interior edges; and $\mathcal{E}(\Gamma)$, $\mathcal{E}(\Gamma_D)$, and $\mathcal{E}(\Gamma_N)$ refer to the set of edges on the boundary of Ω, on the Dirichlet, and on the Neumann boundary, respectively. Let $\hat{T}$, denote the reference simplex or (hyper)cube and let $\mathcal{P}_p(\hat{T})$ and $\mathcal{Q}_p(\hat{T})$ denote the spaces of polynomials of degree less than or equal to p in all and on each variable, respectively. We define the set of piecewise polynomials

$$P_p(\mathcal{T}) := \{v : \Omega \to \mathbb{R} \mid \text{for all } T \in \mathcal{T}, v|_T \circ F_T \in \mathcal{R}_p(\hat{T})\}$$

with $\mathcal{R}_p \in \{\mathcal{P}_p, \mathcal{Q}_p\}$, where $F_T : \hat{T} \to T$, $T \in \mathcal{T}$ is a family of element maps. Let also

$$\Pi_p(\mathcal{T}) : L^2(\Omega) \to P_p(\mathcal{T}) \tag{2.1}$$

denote the L^2-projection onto $\mathcal{T}$-piecewise polynomial functions of order p. In particular, we have $(\Pi_0(\mathcal{T})f)|_T = |T|^{-1}\int_T f\,\mathrm{d}x$, $T \in \mathcal{T}$, for all $f \in L^2(\Omega)$. Note that $v \in P_p(\mathcal{T})$ does not necessarily belong to $H^1(\Omega)$. The $\mathcal{T}$-piecewise gradient $\nabla_{\mathcal{T}} v$, with $(\nabla_{\mathcal{T}} v)|_T = \nabla(v|_T)$ for all $T \in \mathcal{T}$, is well-defined and $\nabla_{\mathcal{T}} v \in (P_{p-1}(\mathcal{T}))^d$.

For any interior edge/face $e \in \mathcal{E}(\Omega)$ there are two adjacent elements T^- and T^+ with $e = \partial T^- \cap \partial T^+$. We define ν to be the normal vector of e that points from T^- to T^+. For boundary edges/faces $e \in \mathcal{E}(\Gamma)$ let ν be the outward unit normal vector of Ω.

Define the jump of $v \in P_k(\mathcal{T})$ across $e \in \mathcal{E}(\Omega)$ by $[v] := v|_{T^-} - v|_{T^+}$ and define $[v] := v|_e$ for $e \in \mathcal{E}(\Gamma)$. The average of $v \in P_p(\mathcal{T})$ across $e \in \mathcal{E}(\Omega)$ is defined by $\{v\} := (v|_{T^-} + v|_{T^+})/2$ and for boundary edges $e \in \mathcal{E}(\Gamma)$ by $\{v\} := v|_e$. Also, we make the shorthand notation $\mathcal{E}(\Omega \cup \Gamma) = \mathcal{E}(\Omega) \cup \mathcal{E}(\Gamma)$.

In the remaining part of this work, we consider two different meshes: a coarse mesh $\mathcal{T}_H$ and a fine mesh $\mathcal{T}_h$, with respective definitions for the edges/faces $\mathcal{E}_H$ and $\mathcal{E}_h$. We denote the $\mathcal{T}_H$-piecewise gradient by $\nabla_H v := \nabla_{\mathcal{T}_H} v$ and, respectively, $\nabla_h v := \nabla_{\mathcal{T}_h} v$ for the $\mathcal{T}_h$-piecewise gradient. We assume that the fine mesh $\mathcal{T}_h$ is the result of one or more refinements of the coarse mesh $\mathcal{T}_H$. The subscripts h, H refer to the corresponding mesh sizes; in particular, we have $H \in P_0(\mathcal{T}_H)$ with $H|_T = \operatorname{diam}(T) =: H_T$ for all $T \in \mathcal{T}_H$, $H_e = \operatorname{diam} e$ for all $e \in \mathcal{E}_H$, $h \in P_0(\mathcal{T}_h)$ with $h|_T = \operatorname{diam}(T) =: h_T$ for all $T \in \mathcal{T}_h$, and $h_e = \operatorname{diam} e$ for all $e \in \mathcal{E}_h$. Obviously, $h \le H$. For simplicity we assume that the discontinuities in A are aligned with the fine mesh $\mathcal{T}_h$.

2.2. Discretization by the symmetric interior penalty method. We consider the symmetric interior penalty method (SIP) dG method [9, 3, 20]. We seek an approximation in the space $\mathcal{V}_h := P_1(\mathcal{T}_h)$. Given some positive penalty parameter $\sigma > 0$, we define the symmetric bilinear form $a_h : \mathcal{V}_h \times \mathcal{V}_h \to \mathbb{R}$ by

$$\begin{aligned} a_h(u,v) := (A\nabla_h u, \nabla_h v)_{L^2(\Omega)} - \sum_{e\in\mathcal{E}_h(\Omega\cup\Gamma_D)} \Big(&(\{\nu\cdot A\nabla u\}, [v])_{L^2(e)} \\ &+ (\{\nu\cdot A\nabla v\}, [u])_{L^2(e)} - \frac{\sigma}{h_e}([u],[v])_{L^2(e)} \Big). \end{aligned} \tag{2.2}$$

The jump-seminorm associated with the space $\mathcal{V}_h$ is defined by

$$|\bullet|_h^2 := \sum_{e\in\mathcal{E}_h(\Omega\cup\Gamma_D)} \frac{\sigma}{h_e}\|[\bullet]\|_{L^2(e)}^2, \tag{2.3}$$

while the energy norm in $\mathcal{V}_h$ is then given by

$$|||\bullet|||_h := (\|A^{1/2}\nabla_h \bullet\|_{L^2(\Omega)}^2 + |\bullet|_h^2)^{1/2}. \tag{2.4}$$

If the penalty parameter is chosen sufficiently large, the dG bilinear form (2.2) is coercive and bounded with respect to the energy norm (2.4).

Remark 1. The penalty parameter, σ, depends of the arithmetic mean of diffusion coefficient on edge e. If a SIP with a weighted average would be used [10], the penalty parameter, σ, would instead depend on the harmonic average of the diffusion coefficient. For simplicity of the presentation and since this choice suffices for our purposes, here we consider the standard SIP dG formulation. Hence, there exists a (unique) dG approximation $u_h \in \mathcal{V}_h$, satisfying

$$a_h(u_h, v) = F(v) \quad \text{for all } v \in \mathcal{V}_h. \tag{2.5}$$

We assume that (2.5) is computationally intractable for practical problems, so we shall never seek to solve for u_h directly. Instead, u_h will serve as a reference solution to compare our low dimensional coarse grid multiscale dG approximation with. The underlying assumption is that the mesh $\mathcal{T}_h$ is chosen sufficiently fine so that u_h is sufficiently accurate. The aim of this work is to devise and analyze a multiscale dG discretization with coarse scale H in such a way that the accuracy of u_h is preserved up to an $\mathcal{O}(H)$ perturbation independent of the variation of the coefficient A.

3. Discontinuous Galerkin multiscale method. As mentioned above, the choice of the reference mesh $\mathcal{T}_h$ is not directly related to the desired accuracy but is instead strongly affected by the roughness and variation of the coefficient A. The corresponding coarse mesh $\mathcal{T}_H$, with mesh width function $H \geq h$, is assumed to be completely independent of A. In the spirit of [21, 22] the test space is divided into coarse and fine components, where the fine scale components are computed on the patches (submeshes) of the reference mesh. To encapsulate the fine scale information in the coarse mesh, we shall design coarse generalized finite element spaces based on $\mathcal{T}_H$.

3.1. Multiscale decompositions. We introduce a two-scale splitting for the space $\mathcal{V}_h$. To this end, let $\Pi_H := \Pi_1(\mathcal{T}_H)$, from (2.1), and define $\mathcal{V}_H := \Pi_H\mathcal{V}_h = P_1(\mathcal{T}_H)$ and

$$\mathcal{V}^{\mathrm{f}} := (1-\Pi_H)\mathcal{V}_h = \{v \in \mathcal{V}_h \mid \Pi_H v = 0\}.$$

LEMMA 2 (L^2-orthogonal multiscale decomposition). *The decomposition*

$$\mathcal{V}_h = \mathcal{V}_H \oplus \mathcal{V}^{\mathrm{f}}$$

is orthogonal in $L^2(\Omega)$.

Proof. The proof is immediate, as any $v \in \mathcal{V}_h$ can be decomposed uniquely into a coarse finite element function $v_H := \Pi_H v \in \mathcal{V}_H$ and a (possibly highly oscillatory) remainder $v^{\mathrm{f}} := (1-\Pi_H)v \in \mathcal{V}^{\mathrm{f}}$ with $\|v\|_{L^2(\Omega)}^2 = \|v_H\|_{L^2(\Omega)}^2 + \|v^{\mathrm{f}}\|_{L^2(\Omega)}^2$. ∎

We now orthogonalize the above splitting with respect to the dG scalar product a_h; we keep the space of fine scale oscillations $\mathcal{V}^{\mathrm{f}}$ and simply replace $\mathcal{V}_H$ with the orthogonal complement of $\mathcal{V}^{\mathrm{f}}$ in $\mathcal{V}_h$. We define the fine scale projection $\mathfrak{F} : \mathcal{V}_h \to \mathcal{V}^{\mathrm{f}}$ by

$$a_h(\mathfrak{F}v, w) = a_h(v, w) \quad \text{for all } w \in \mathcal{V}^{\mathrm{f}}. \tag{3.1}$$

Using the fine scale projection, we can define the coarse scale approximation space by

$$\mathcal{V}_H^{ms} := (1 - \mathfrak{F})\mathcal{V}_H.$$

LEMMA 3 (a_h-orthogonal multiscale decomposition). *The decomposition*

$$\mathcal{V}_h = \mathcal{V}_H^{ms} \oplus \mathcal{V}^{\mathrm{f}}$$

is orthogonal with respect to a_h, i.e., any function v in $\mathcal{V}_h$ can be decomposed uniquely into some function $v_H^{ms} \in \mathcal{V}_H^{ms}$ plus $v^{\mathrm{f}} \in \mathcal{V}^{\mathrm{f}}$ with $C^{-1}|||v|||_h^2 \leq |||v_H^{ms}|||_h^2 + |||v^{\mathrm{f}}|||_h^2 \leq C|||v|||_h^2$, where the constant C only depends on the coercivity and continuity constants of the bilinear form. The functions $v_H^{ms} \in \mathcal{V}_H^{ms}$ and $v^{\mathrm{f}} \in \mathcal{V}^{\mathrm{f}}$ are the Galerkin projections of $v \in \mathcal{V}_h$ onto the subspaces $\mathcal{V}_H^{ms}$ and $\mathcal{V}^{\mathrm{f}}$, i.e.,

$$\begin{aligned} a_h(v_H^{ms}, w) &= a_h(v, w) \quad \textit{for all } w \in \mathcal{V}_H^{ms}, \\ a_h(v^{\mathrm{f}}, w) &= a_h(v, w) \quad \textit{for all } w \in \mathcal{V}^{\mathrm{f}}. \end{aligned}$$

The unique Galerkin approximation $u_H^{ms} \in \mathcal{V}_H^{ms}$ of $u \in \mathcal{V}$ solves

$$a_h(u_H^{ms}, v) = F(v) \quad \text{for all } v \in \mathcal{V}_H^{ms}. \tag{3.2}$$

We shall see in the error analysis (cf. Theorem 9) that the orthogonality yields error estimates (with respect to a reference solution) for the Galerkin approximation $u_H^{ms} \in \mathcal{V}_H^{ms}$ of (3.2) that are independent of the regularity of the solution and of the variation in the diffusion coefficient A. However, the space $\mathcal{V}_H^{ms}$ is not suitable for practical computations as a local basis for this space is not easily available. Indeed, given a basis of $\mathcal{V}_H$, e.g., the elementwise Lagrange basis functions $\{\lambda_{T,j} \mid T \in \mathcal{T}_H,\ j = 1, \dots, r\}$, where $r = (1 + d)$ for regular simplices or $r = 2^d$ for quadrilaterals/hexahedra, the space $\mathcal{V}_H^{ms}$ is spanned by the corrected basis functions $(1 - \mathfrak{F})\lambda_{T,j}$, $T \in \mathcal{T}_H$, $j = 1, \dots, r$. Although $\lambda_{T,j}$ has local support $\operatorname{supp} \lambda_{T,j} = T$, its corrected version $(1 - \mathfrak{F})\lambda_{T,j}$ has global support in Ω, as (3.1) is a variational problem on the whole domain Ω. Fortunately, as we shall prove below, the corrector functions $\phi_{T,j}$ decay quickly away from T. (See previous numerical results in [13] and a similar observation for a related conforming method [25].) This decay motivates the local approximation of the corrector functions at the expense of introducing small perturbations in the method's accuracy.

3.2. Localization and computational method. The localized approximations of the corrector functions are supported on element patches in the coarse mesh $\mathcal{T}_H$.

DEFINITION 4. *For all $T \in \mathcal{T}_H$, define element patches with size L as*

$$\begin{aligned} \omega_T^1 &:= \operatorname{int}(T), \\ \omega_T^L &:= \operatorname{int}(\cup\{T' \in \mathcal{T}_H \mid T' \cap \bar{\omega}_T^{L-1} \neq \emptyset\}), \quad L = 1, 2, \dots. \end{aligned}$$

We refer to Figure 3.1 *for an illustration.*

We introduce a new discretization parameter $0 < L \in \mathbb{N}$ and define localized corrector functions $\phi_{T,j}^L \in \mathcal{V}^{\mathrm{f}}(\omega_T^L) := \{v \in \mathcal{V}^{\mathrm{f}} \mid v|_{\Omega \setminus \omega_T^L} = 0\}$ by

$$a_h(\phi_{T,j}^L, w) = a_h(\lambda_{T,j}, w) \quad \text{for all } w \in \mathcal{V}^{\mathrm{f}}(\omega_T^L). \tag{3.3}$$

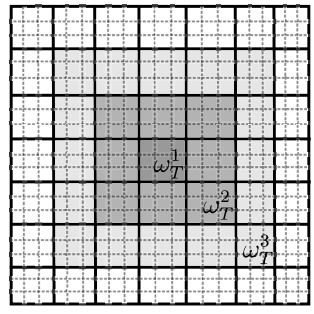


FIG. 3.1. *Example of a one-layer patch ω_T^1, a two-layer patch ω_T^2, and a three-layer patch ω_T^3 on a quadrilateral mesh.*

Further, we define the multiscale approximation space

$$\mathcal{V}_H^{ms,L} = \operatorname{span}\{\lambda_{T,j} - \phi_{T,j}^L \mid T \in \mathcal{T}_H,\ j = 1, \dots, r\}. \tag{3.4}$$

The dG multiscale method seeks $u_H^{ms,L} \in \mathcal{V}_H^{ms,L}$ such that

$$a_h(u_H^{ms,L}, v) = F(v) \quad \text{for all } v \in \mathcal{V}_H^{ms,L}. \tag{3.5}$$

Since $\mathcal{V}_H^{ms,L} \subset \mathcal{V}_h$, this method is a Galerkin method in the Hilbert space $\mathcal{V}_h$ (with scalar product a_h) and hence inherits well-posedness from the reference discretization (2.5).

Moreover, the proposed basis $\{\lambda_{T,j} - \phi_{T,j}^L \mid T \in \mathcal{T}_H,\ j = 1, \dots, r\}$ is stable with respect to the fine scale parameter h, as we shall see in Lemma 8 below.

3.3. Compressed dG multiscale method. The basis functions in the above multiscale method have enlarged supports (element patches) when compared with standard dG methods (elements). We can decompose the corrector functions into its element contributions

$$\phi_{T,j}^L = \sum_{T' \in \mathcal{T}_H : T' \subset \omega_T^L} \phi_{T,j}^L \chi_{T'},$$

where $\chi_{T'}$ is the indicator function of the element $T' \in \mathcal{T}_H$.

This motivates the coarse approximation space

$$\begin{aligned} \mathcal{W}_H^{ms,L} = \operatorname{span}\big(&\{\lambda_{T,j} \,|T \in \mathcal{T}_H,\ j = 1, \dots, r\} \\ &\cup \{\phi_{T,j}^L \chi_{T'} \,|T, T' \in \mathcal{T}_H,\ T' \subset \omega_T^L,\ j = 1, \dots, r\}\big). \end{aligned}$$

This space offers the advantage of a known basis with elementwise support which leads to improved (localized) connectivity in the corresponding stiffness matrix. This is at the expense of a slight increase in the dimension of the space

$$\dim(\mathcal{W}_H^{ms,L}) \approx L^d \dim(\mathcal{V}_H^{ms,L}). \tag{3.6}$$

The corresponding localized dG multiscale method seeks $w_H^{ms,L} \in \mathcal{W}_H^{ms,L}$ such that

$$a_h(w_H^{ms,L}, v) = F(v) \quad \text{for all } v \in \mathcal{W}_H^{ms,L}. \tag{3.7}$$

Since $\mathcal{V}_H^{ms,L} \subset \mathcal{W}_H^{ms,L} \subset \mathcal{V}_h$, Galerkin orthogonality yields

$$|||u_h - w_H^{ms,L}|||_h \leq |||u_h - u_H^{ms,L}|||_h, \tag{3.8}$$

i.e., the new localized version (3.7) is never worse than the previous multiscale approximation in terms of accuracy. However, it may lead to very ill-conditioned element stiffness matrices. (See Lemma 11, which shows that $\phi_{T,j}^L \chi_{T'}$ may be very small if the distance between T and T' relative to their sizes is large.)

To circumvent ill-conditioning, one may choose a reduced local approximation space based on an eigendecomposition of the element stiffness matrix. The eigenfunctions which correspond to sufficiently large eigenvalues (principal components) are used as basis functions for the reduced space. Since the dimension of the element stiffness matrix is small (at most proportional to $L^d \times L^d$), the cost of this additional preprocessing step is negligible when compared with the cost of solving the local problems for the corrector functions.

To determine an acceptable level of truncation of the localized basis functions, we can use the a posteriori error estimator contribution of the local problem from [13], which is an estimation of the local fine scale error. Using an adaptive algorithm in [13] to determine the size of the patches may additionally lead to large reduction of the dimension of the local approximation spaces (3.6), since in (3.6) all the patches are assumed to have the same size L.

4. Error analysis. We present an a priori error analysis for the proposed multiscale method (3.5). In view of (3.8), this analysis applies immediately to the modified versions presented in section 3.3. The error analysis will be split into a number of steps. First, in section 4.1, we present some properties of the coarse scale projection operator Π_H. In section 4.2, an error bound for dG multiscale method u_H^{ms} from (3.2) (Theorem 9) is shown, whereby the corrected basis functions are solved globally. Results for the decay of the localized corrected basis function (Lemmas 11 and 12) are shown, along with an error bound for the dG multiscale method $u_H^{ms,L}$ from (3.5) (Theorem 13), where the corrected basis functions are solved locally on element patches. Finally, in section 4.3, we show an error bound given a quantity of interest (Theorem 15), leading to an error bound in the L^2-norm (Corollary 16).

We shall make use of the following (semi)norms. The jump-seminorm and energy norms, associated with the coarse space $\mathcal{V}_H$, are defined by

$$|\bullet|_H^2 := \sum_{e \in \mathcal{E}_H(\Omega \cup \Gamma_D)} \frac{\sigma}{H_e} \|[\bullet]\|_{L^2(e)}^2,$$
$$|||\bullet|||_H := (\|A^{1/2}\nabla_H \bullet\|_{L^2(\Omega)}^2 + |\bullet|_H^2)^{1/2},$$

respectively, along with a localized version of the local jump and energy norms (2.3) and (2.4) on a patch $\omega \subseteq \Omega$, where ω is aligned with the mesh $\mathcal{T}_h$, given by

$$|\bullet|_{h,\omega}^2 := \sum_{\substack{e \in \mathcal{E}_h(\Omega \cup \Gamma_D): \\ e \cap \bar{\omega} \neq 0}} \frac{\sigma}{h_e} \|[\bullet]\|_{L^2(e)}^2,$$
$$|||\bullet|||_{h,\omega} := (\|A^{1/2}\nabla_h \bullet\|_{L^2(\omega)}^2 + |\bullet|_{h,\omega}^2)^{1/2}.$$

The shape-regularity assumptions $h_T \approx h_e$ for all $e \in \partial T : T \in \mathcal{T}_h$ and $H_T \approx H_e$ for all $T \in \partial T : T \in \mathcal{T}_H$ will also be used.

4.1. Properties of the coarse scale projection operator Π_H. The following lemma gives stability and approximation properties of the operator Π_H.

LEMMA 5. *For any $v \in \mathcal{V}_h$, the estimate*

$$H^{-1}\|v - \Pi_H v\|_{L^2(T)} \lesssim \alpha^{-1/2}|||v|||_{h,T}$$

is satisfied for all $T \in \mathcal{T}_H$. Moreover, it holds that

$$\beta^{-1/2}|||\Pi_H v|||_H + \|H^{-1}(v - \Pi_H v)\|_{L^2(\Omega)} \lesssim \alpha^{-1/2}|||v|||_h,$$

where α and β are defined in (1.1).

Proof. Theorem 2.2 in [23] implies that for each $v \in \mathcal{V}_h$, there exists a bounded linear operator $\mathcal{I}_h^{\mathrm{c}} : \mathcal{V}_h \to \mathcal{V}_h \cap H^1(\Omega)$ such that

$$\beta^{-1/2}\|A^{1/2}\nabla_H(v - \mathcal{I}_h^{\mathrm{c}}v)\|_{L^2(T)} + \|h^{-1}(v - \mathcal{I}_h^{\mathrm{c}}v)\|_{L^2(T)} \lesssim \alpha^{-1/2}|v|_{h,T}. \tag{4.1}$$

We split $v = v^{\mathrm{c}} + v^{\mathrm{d}} \in \mathcal{V}_h$ into a conforming, $v^{\mathrm{c}} = \mathcal{I}_h^{\mathrm{c}}v$, and a nonconforming, $v^{\mathrm{d}} = v - \mathcal{I}_h^{\mathrm{c}}v$, part and obtain

$$\begin{aligned} H^{-1}\|v - \Pi_H v\|_{L^2(T)} &\le H^{-1}(\|v^{\mathrm{c}} - \Pi_H v^{\mathrm{c}}\|_{L^2(T)} + \|v^{\mathrm{d}} - \Pi_H v^{\mathrm{d}}\|_{L^2(T)}) \\ &\lesssim \|\nabla_h v\| + \|\nabla_h(v - v^{\mathrm{c}})\|_{L^2(T)} + H^{-1}\|v^{\mathrm{d}}\|_{L^2(T)}) \\ &\lesssim \alpha^{-1/2}|||v|||_{h,T} \end{aligned} \tag{4.2}$$

using the triangle inequality, stability of the L^2-projection, and (4.1). Furthermore,

$$\begin{aligned} |||\Pi_H v|||_H^2 &= \sum_{T\in\mathcal{T}_H} \|\sqrt{A}\nabla(\Pi_H v - \Pi_0(\mathcal{T}_H)v)\|_{L^2(T)}^2 + \sum_{e\in\mathcal{E}_H(\Omega\cup\Gamma_D)} \frac{\sigma}{H}\|[v_c - \Pi_H v]\|_{L^2(e)}^2 \\ &\lesssim \sum_{T\in\mathcal{T}_H} \beta\left(\frac{1}{H^2}\|v - \Pi_0(\mathcal{T}_H)v\|_{L^2(T)}^2 + \frac{1}{H^2}\|v_c - \Pi_H v\|_{L^2(T)}^2\right) \\ &\lesssim C_{\beta/\alpha}^2|||v|||_h^2 \end{aligned}$$

using the triangle inequality, (4.1), and (4.2), which concludes the proof. ∎

The operator Π_H is surjective. The next lemma shows that given some $v_H \in \mathcal{V}_H$ in the image of Π_H there exists a H^1-conforming preimage $v \in \Pi_H^{-1}v_H \subset \mathcal{V}_h$ with comparable support.

LEMMA 6. *For each $v_H \in \mathcal{V}_H$, there exists a $v \in \mathcal{V}_h \cap H^1(\Omega)$ such that $\Pi_H v = v_H$, $|||v|||_h \lesssim C_{\beta/\alpha}|||v_H|||_H$, and $\operatorname{supp}(v) \subseteq \operatorname{supp}(\mathcal{I}_h^{\mathrm{c}}v_H)$. Note that the support of $\mathcal{I}_h^{\mathrm{c}}v_H$ is one layer of coarse element larger than the support of v_H.*

Proof. Using Theorem 2.2 in [23] but on the space $\mathcal{V}_H$ gives for each $v \in \mathcal{V}_H$ that there exists a bounded linear operator $\mathcal{I}_H^{\mathrm{c}} : \mathcal{V}_H \to \mathcal{V}_H \cap H^1(\Omega)$ such that

$$\beta^{-1/2}\|A^{1/2}\nabla_H(v - \mathcal{I}_H^{\mathrm{c}}v)\|_{L^2(T)} + \|H^{-1}(v - \mathcal{I}_H^{\mathrm{c}}v)\|_{L^2(T)} \lesssim \alpha^{-1/2}|v|_{H,T}. \tag{4.3}$$

We define

$$v := \mathcal{I}_H^{\mathrm{c}}v_H + \sum_{T\in\mathcal{T}_H,\ j=1,\dots,r} (v_H(x_j) - \mathcal{I}_H^{\mathrm{c}}v_H(x_j))\,\theta_{T,j},$$

where $\theta_{T,j} \in \mathcal{V}_h \cap H_0^1(T)$ are coarse scale bubble functions, supported on each element T, with $\Pi_H\theta_{T,j} = \lambda_{T,j}$ and $|||\theta_{T,j}|||_h^2 \lesssim \beta H^{d-2}$. Observe that $\operatorname{supp}(v) \subseteq \operatorname{supp}(\mathcal{I}_h^{\mathrm{c}}v_H)$. The interpolation property follows from

$$\begin{aligned}\Pi_H v &= \mathcal{I}_H^{\mathrm{c}} v_H + \Pi_H \sum_{T\in\mathcal{T}_H,\ j=1,\dots,r} (v_H(x_j) - \mathcal{I}_H^{\mathrm{c}} v_H(x_j))\,\theta_{T,j},\\ &= \mathcal{I}_H^{\mathrm{c}} v_H + \sum_{T\in\mathcal{T}_H,\ j=1,\dots,r} (v_H(x_j) - \mathcal{I}_H^{\mathrm{c}} v_H(x_j))\lambda_{T,j} = v_H.\end{aligned}$$

To prove stability, we estimate $|||v|||_h$ as follows:

$$\begin{aligned}|||v|||_h^2 &\le \|A^{1/2}\nabla \mathcal{I}_H^{\mathrm{c}} v_H\|_{L^2(\Omega)}^2 + \sum_{T\in\mathcal{T}_H,\ j=1,\dots,r} |v_H(x_j) - \mathcal{I}_H^{\mathrm{c}} v_H(x_j)|^2\, |||\theta_j|||_h^2\\ &\lesssim \|A^{1/2}\nabla_H \mathcal{I}_H^{\mathrm{c}} v_H\|_{L^2(\Omega)}^2 + \beta\|H^{-1}(v_H - \mathcal{I}_H^{\mathrm{c}} v_H)\|_{L^2(\Omega)}^2\\ &\lesssim C_{\beta/\alpha}^2 |||v_H|||_H^2,\end{aligned}$$

using the inverse estimate $\|v\|_{L^\infty(T)} \le H^{-d/2}\|v\|_{L^2(T)}$ for all $v \in \mathcal{V}_H$ and using the estimate (4.3). ∎

Remark 7. Note that $\theta_{T,j} \in \mathcal{V}_h \cap H_0^1(T)$ for all $T \in \mathcal{T}_H$ (fulfilling the conditions in Lemma 6) can be constructed using two (or more) refinements of the coarse scale parameter H. We can let $\theta_{T,j} \in \mathcal{V}_{h'} \cap H_0^1(T)$, where $\mathcal{V}_{h'} \subset \mathcal{V}_h$ and $h \le h' \le 2^{-2}H$. This does not put a big restriction on h since the mesh $\mathcal{T}_h$ is assumed to be sufficiently fine to resolve the variation in the coefficient A, while the parameter H does not need to resolve A.

The following lemma says that the corrected basis is stable with respect to the fine scale parameter h in the energy norm (2.4); this is not a trivial result since the basis function $\{\lambda_{T,j} | T \in \mathcal{T}_H,\ j = 1,\dots,r\}$ is discontinuous.

LEMMA 8 (stability of the corrected basis functions). *For all $T \in \mathcal{T}_H$, $j = 1,\dots,r$, and $L > 0 \in \mathbb{N}$, the estimate*

$$|||\lambda_{T,j} - \phi_{T,j}^L|||_h \lesssim C_{\beta/\alpha}|||\lambda_{T,j}|||_H$$

is satisfied, independently of the fine scale parameter h.

Proof. For any $T \in \mathcal{T}_H$, $j = 1,\dots,r$, by Lemma 6 there exists a b such that $v = \lambda_{T,j} - b \in \mathcal{V}_h^f(\omega_T^L)$, and $|||b|||_h \lesssim C_{\beta/\alpha}|||\lambda_{T,j}|||_H$. We have

$$\begin{aligned}|||\lambda_{T,j} - \phi_{T,j}^L|||_h^2 &\lesssim a_h(\lambda_{T,j} - \phi_{T,j}^L, \lambda_{T,j} - \phi_{T,j}^L) = a_h(\lambda_{T,j} - \phi_{T,j}^L, \lambda_{T,j} - v),\\ &\lesssim a_h(\lambda_{T,j} - \phi_{T,j}^L, b) \lesssim C_{\beta/\alpha}|||\lambda_{T,j} - \phi_{T,j}^L|||_h|||\lambda_{T,j}|||_H,\end{aligned}$$

which concludes the proof. ∎

4.2. A priori estimates. The following theorem gives an error bound for the idealized dG multiscale method, whereby the correctors for the basis are solved globally.

THEOREM 9. *Let $u_h \in \mathcal{V}_h$ solve* (2.5) *and let $u_H^{ms} \in \mathcal{V}_H^{ms}$ solve* (3.5)*; then the estimate*

$$|||u_h - u_H^{ms}|||_h \le C_1\alpha^{-1/2}||H(f - \Pi_H f)||_{L^2(\Omega)}$$

is satisfied, where C_1 depends on neither the mesh (h or H) size nor the diffusion matrix A.

Proof. Let $e := u_h - u_H^{ms} = u_f \in \mathcal{V}^{\mathrm{f}}$; then

$$\begin{aligned}|||e|||_h^2 &\lesssim a_h(e,e) = (f,e)_{L^2(\Omega)} = (f - \Pi_H f, e - \Pi_H e)_{L^2(\Omega)}\\ &\le ||H(f - \Pi_H f)||_{L^2(\Omega)}||H^{-1}(e - \Pi_H e)||_{L^2(\Omega)}\\ &\lesssim \frac{1}{\sqrt{\alpha}}||H(f - \Pi_H f)||_{L^2(T)}|||e|||_h\end{aligned}$$

using Lemma 3, Lemma 2, the Cauchy–Schwarz inequality, and Lemma 5, respectively. ∎

DEFINITION 10. *The cutoff functions* $\zeta_T^{d,D} \in P_0(\mathcal{T}_h)$ *are defined by the conditions*

$$\begin{aligned}\zeta_T^{d,D}|_{\omega_T^d} &= 1,\\ \zeta_T^{d,D}|_{\Omega\setminus\omega_T^D} &= 0,\\ \|[\zeta_T^{d,D}]\|_{L^\infty(\mathcal{E}_h(T))} &\lesssim \frac{h_T}{(D-d)H_T} \quad \text{for all } T\in\mathcal{T}_H,\end{aligned}$$

and $\zeta_T^{d,D}$ *is constant on the boundary* $\partial(\omega_T^D\setminus\omega_T^d)$.

The next lemma shows the exponential decay in the corrected basis; this is a key result in the analysis.

LEMMA 11. *For all* $T\in\mathcal{T}_H$, $j=1,\dots,r$, *the estimate*

$$|||(\lambda_{T,j}-\phi_{T,j})-(\lambda_{T,j}-\phi_{T,j}^L)|||_h = |||\phi_{T,j}-\phi_{T,j}^L|||_h \le C_3\gamma^L|||\phi_{T,j}-\lambda_{T,j}|||_h$$

is satisfied with $C_3 = CC_{\beta/\alpha}^3$, $0<\gamma<1$ *given by* $\gamma := (\frac{C_2}{\ell-1})^{\frac{k-1}{2\ell}}$, $C_2 = C'C_{\beta/\alpha}^2$, *and* $L=k\ell$, $k,\ell\ge 2\in\mathbb{N}$, *noting that* C *and* C' *are positive constants that are independent of the mesh* (h *or* H), *of the patch size* L, *and of the diffusion matrix* A.

Proof. Define $e := \phi_{T,j}-\phi_{T,j}^L = \phi_{T,j}-\phi_{T,j}^{\ell k}$. We have

$$|||e|||_h^2 \lesssim a_h(e,\phi_{T,j}-\phi_{T,j}^{\ell k}) = a_h(e,\phi_{T,j}-v) \lesssim |||e|||_h\cdot|||\phi_{T,j}-v|||_h \tag{4.4}$$

for $v\in\mathcal{V}_h^f(\omega_T^{\ell k})$. Let $\zeta := \zeta_T^{\ell k-1,\ell k}$; then by Lemma 6 there exists a b such that $v=\zeta\phi_{T,j}-b\in\mathcal{V}_h^f(\omega_T^{\ell k})$, $\Pi_H b=\Pi_H\zeta\phi_{T,j}$, $|||b|||_h\lesssim C_{\beta/\alpha}|||\Pi_H\zeta\phi_{T,j}|||_H$, and $\operatorname{supp}(b)\subseteq\operatorname{supp}(\mathcal{I}_h^c\Pi_H\zeta\phi_{T,j})$. We have

$$\begin{aligned}|||\phi_{T,j}-v|||_h &= |||\phi_{T,j}-(\zeta\phi_{T,j}-b)|||_h\\ &\le |||\phi_{T,j}-\zeta\phi_{T,j}|||_h+|||b|||_h\\ &\lesssim |||\phi_{T,j}-\zeta\phi_{T,j}|||_h+C_{\beta/\alpha}|||\Pi_H(\zeta\phi_{T,j}-\phi_{T,j})|||_H\\ &\lesssim C_{\beta/\alpha}^2|||\phi_{T,j}-\zeta\phi_{T,j}|||_h.\end{aligned} \tag{4.5}$$

Furthermore, using the properties of ζ we have

$$\|\sqrt{A}\nabla_h(1-\zeta)\phi_{T,j}\|_{L^2(\Omega)} \le \|\sqrt{A}\nabla_h\phi_{T,j}\|_{L^2(\Omega\setminus\omega_T^{\ell k-1})} \tag{4.6}$$

and

$$\begin{aligned}|(1-\zeta)\phi_{T,j}|_h^2 &= \sum_{e\in\mathcal{E}_h(\Omega\cup\Gamma_D)}\frac{\sigma}{h_e}\|[(1-\zeta)\phi_{T,j}]\|_{L^2(e)}^2\\ &\le \sum_{e\in\mathcal{E}_h(\Omega\cup\Gamma_D)}\frac{\sigma}{h_e}\left(\|\{1-\zeta\}[\phi_{T,j}]\|_{L^2(e)}^2+\|\{\phi_{T,j}\}[1-\zeta]\|_{L^2(e)}^2\right)\\ &\le \sum_{\substack{e\in\mathcal{E}_h(\Omega\cup\Gamma_D):\\ e\cap\Omega\setminus\omega_T^{\ell k-1}\neq 0}}\left(\frac{\sigma}{h_e}\|[\phi_{T,j}]\|_{L^2(e)}^2+\frac{\sigma h_T^2}{h_eH_T^2}\|\{\phi_{T,j}\}\|_{L^2(e)}^2\right)\\ &\le \sum_{\substack{e\in\mathcal{E}_h(\Omega\cup\Gamma_D):\\ e\cap\Omega\setminus\omega_T^{\ell k-1}\neq 0}}\frac{\sigma}{h_e}\|[\phi_{T,j}]\|_{L^2(e)}^2+\frac{\sigma}{H_T^2}\|\phi_{T,j}-\Pi_H\phi_{T,j}\|_{L^2(\Omega\setminus\omega_T^{\ell k-1})}^2\\ &\lesssim C_{\beta/\alpha}^2|||\phi_{T,j}|||_{h,\Omega\setminus\omega_T^{\ell k-1}}^2\end{aligned} \tag{4.7}$$

using a trace inequality and Lemma 5, respectively. Combining (4.4), (4.5), (4.6), and (4.7) yields

$$|||e|||_h \lesssim C_{\beta/\alpha}^2 |||\phi_{T,j} - \zeta\phi_{T,j}|||_h \lesssim C_{\beta/\alpha}^3 |||\phi_{T,j}|||_{h,\Omega\setminus\omega_T^{\ell k-1}}. \tag{4.8}$$

To simplify notation, let $m := \ell(k-1)-1$ and $M := \ell k - 1$. For $\eta_T := 1 - \zeta_T^{m+1,M}$, we obtain

$$|||\phi_{T,j}|||^2_{h,\Omega\setminus\omega_T^M} \leq |||\eta_T\phi_{T,j}|||_h^2 \lesssim a_h(\eta_T\phi_{T,j}, \eta_T\phi_{T,j}), \tag{4.9}$$

where

$$\begin{aligned} a_h(\eta_T\phi_{T,j}, \eta_T\phi_{T,j}) &= (A\nabla_h\eta_T\phi_{T,j}, \nabla_h\eta_T\phi_{T,j})_{L^2(\Omega)} \\ &\quad + \sum_{e\in\mathcal{E}_h(\Omega\cup\Gamma_D)} \left(-2(\{\nu\cdot A\nabla\eta_T\phi_{T,j}\}, [\eta_T\phi_{T,j}]) + \frac{\sigma}{h_e}([\eta_T\phi_{T,j}], [\eta_T\phi_{T,j}])\right). \end{aligned} \tag{4.10}$$

For the first term on the right-hand side of (4.10), we have

$$(A\nabla_h\eta_T\phi_{T,j}, \nabla_h\eta_T\phi_{T,j})_{L^2(\Omega)} = (A\nabla_h\phi_{T,j}, \nabla_h\eta_T^2\phi_{T,j})_{L^2(\Omega)} \tag{4.11}$$

since η_T is constant on each element $T\in\mathcal{T}_h$; for the other terms we use (A.3) and (A.4) (with $v=\eta_T$, $w=\nu\cdot A\nabla\phi_{T,j}$ and $u=\phi_{T,j}$). We can thus arrive at

$$\begin{aligned} |||\phi_{T,j}|||^2_{h,\Omega\setminus\omega_T^M} \leq a_h(\eta_T\phi_{T,j}, \eta_T\phi_{T,j}) &= a_h(\phi_{T,j}, \eta_T^2\phi_{T,j}) \\ &\quad + \sum_{e\in\mathcal{E}_h(\Omega)} \big(1/2(\{\nu\cdot A\nabla\phi_{T,j}\}, [\eta_T]^2[\phi_{T,j}])_{L^2(e)} \\ &\quad - 1/4([\nu\cdot A\nabla\phi_{T,j}], [\eta_T]^2\{\phi_{T,j}\})_{L^2(e)} \\ &\quad - \frac{\sigma}{4h_e}([\eta_T]^2, [\phi_{T,j}]^2)_{L^2(e)} + \frac{\sigma}{h_e}([\eta_T]^2, \{\phi_{T,j}\}^2)_{L^2(e)}\big) \end{aligned} \tag{4.12}$$

using (4.9), (4.10), and (4.11). Note that

$$\begin{aligned} &\sum_{e\in\mathcal{E}_h(\Omega)} \big(1/2(\{\nu\cdot A\nabla\phi_{T,j}\}, [\eta_T]^2[\phi_{T,j}])_{L^2(e)} - 1/4([\nu\cdot A\nabla\phi_{T,j}], [\eta_T]^2\{\phi_{T,j}\})_{L^2(e)} \\ &\quad - \frac{\sigma}{4h_e}([\eta_T]^2, [\phi_{T,j}]^2)_{L^2(e)} + \frac{\sigma}{h_e}([\eta_T]^2, \{\phi_{T,j}\}^2)_{L^2(e)}\big) \\ &\lesssim \sum_{\substack{e\in\mathcal{E}_h(\Omega):\\ e\cap\omega_T^M\setminus\omega_T^{m+1}\neq 0}} \frac{h_T^2}{\ell^2 H_T^2}\Big(\|\{\nu\cdot A\nabla\phi_{T,j}\}\|_{L^2(e)}\|[\phi_{T,j}]\|_{L^2(e)} \\ &\quad + \|[\nu\cdot A\nabla\phi_{T,j}]\|_{L^2(e)}\|\{\phi_{T,j}\}\|_{L^2(e)} + \frac{\sigma}{h_e}\left(\|[\phi_{T,j}]\|^2_{L^2(e)} + \|\{\phi_{T,j}\}\|^2_{L^2(e)}\right)\Big) \\ &\lesssim \sum_{\substack{e\in\mathcal{E}_h(\Omega):\\ e\cap\omega_T^M\setminus\omega_T^{m+1}\neq 0}} \left(\frac{h_T}{\ell^2 H_T^2}\|A\nabla\phi_{T,j}\|_{L^2(T^+\cup T^-)}\|\phi_{T,j}\|_{L^2(T^+\cup T^-)} + \frac{\sigma}{\ell^2 H_T^2}\|\phi_{T,j}\|^2_{L^2(T^+\cup T^-)}\right) \\ &\lesssim \beta\ell^{-2}\|H_T^{-1}(\phi_{T,j} - \Pi_H\phi_{T,j})\|^2_{L^2(\omega_T^M\setminus\omega_T^{m+1})} \leq C_{\beta/\alpha}^2\ell^{-2}|||\phi_{T,j}|||^2_{h,\omega_T^M\setminus\omega_T^{m+1}}. \end{aligned} \tag{4.13}$$

Using that there exist a b such that $\Pi_H b = \Pi_H \eta_T^2 \phi_{T,j}$, $|||b|||_h \lesssim C_{\beta/\alpha} |||\Pi_H \eta_T^2 \phi_{T,j}|||_H$, and $\operatorname{supp}(v) \subseteq \operatorname{supp}(\mathcal{I}_h^c v_H)$ from Lemma 6, we have

$$
\begin{aligned}
a_h(\phi_{T,j}, \eta_T^2 \phi_{T,j}) &= a_h(\phi_{T,j}, \eta_T^2 \phi_{T,j} - b) + a_h(\phi_{T,j}, b) = a_h(\phi_{T,j}, b) \\
&\lesssim |||\phi_{T,j}|||_{h,\omega_T^{M+1}\setminus\omega_T^m} |||b|||_{h,\omega_T^{M+1}\setminus\omega_T^m} \\
&\leq C_{\beta/\alpha} |||\phi_{T,j}|||_{h,\omega_T^{M+1}\setminus\omega_T^m} |||\Pi_H \eta_T^2 \phi_{T,j}|||_{H,\omega_T^M\setminus\omega_T^m}.
\end{aligned} \tag{4.14}
$$

Furthermore, we have that

$$
\begin{aligned}
|||\Pi_H \eta_T^2 \phi_{T,j}|||^2_{H,\omega_T^M\setminus\omega_T^m} &= |||\Pi_H(\eta_T^2 - \Pi_0(\mathcal{T}_H)\eta_T^2)\phi_{T,j}|||^2_{H,\omega_T^M\setminus\omega_T^m} \\
&= \|\sqrt{A}\nabla_H \Pi_H(\eta_T^2 - \Pi_0(\mathcal{T}_H)\eta_T^2)\phi_{T,j}\|^2_{L^2(\omega_T^M\setminus\omega_T^m)} \\
&\quad + \sum_{\substack{e\in\mathcal{E}_h(\Omega\cup\Gamma_D):\\ e\cap\omega_T^M\setminus\omega_T^m\neq 0}} \frac{\sigma}{H_e} \|[\Pi_H(\eta_T^2 - \Pi_0(\mathcal{T}_H)\eta_T^2)\phi_{T,j}]\|^2_{L^2(e)} \\
&\lesssim \beta \|H_e^{-1}\Pi_H(\eta_T^2 - \Pi_0(\mathcal{T}_H)\eta_T^2)\phi_{T,j}\|^2_{L^2(\omega_T^M\setminus\omega_T^m)} \\
&\leq \sum_{T\in\mathcal{T}_h(\omega_T^M\setminus\omega_T^m)} \beta \|H_e^{-1}(\eta_T^2 - \Pi_0(\mathcal{T}_H)\eta_T^2)\|^2_{L^\infty(T)} \|\phi_{T,j}\|^2_{L^2(T)} \\
&\lesssim \beta\ell^{-2} \|H_e^{-1}(\phi_{T,j} - \Pi_H\phi_{T,j})\|^2_{L^2(\omega_T^M\setminus\omega_T^m)} \\
&\lesssim C^2_{\beta/\alpha}\ell^{-2} |||\phi_{T,j}|||^2_{h,\omega_T^M\setminus\omega_T^m}
\end{aligned} \tag{4.15}
$$

using a trace inequality, an inverse inequality, and Lemma 5, respectively. Combining the inequalities (4.12), (4.13), (4.14), and (4.15) yields

$$
|||\phi_{T,j}|||^2_{h,\Omega\setminus\omega_T^M} \leq \frac{C_2}{\ell-1} |||\phi_{T,j}|||^2_{h,\omega_T^{M+1}\setminus\omega_T^m} \leq \frac{C_2}{\ell-1} |||\phi_{T,j}|||^2_{h,\Omega\setminus\omega_T^m},
$$

where $C_2 = C' C^2_{\beta/\alpha}$. Substituting back to ℓ and k and using a cutoff function with a slightly different argument yields

$$
\begin{aligned}
|||\phi_{T,j}|||^2_{h,\Omega\setminus\omega_T^{\ell k-1}} &\leq \frac{C_2}{\ell-1} |||\phi_{T,j}|||^2_{h,\Omega\setminus\omega_T^{\ell(k-1)-1}} \leq \left(\frac{C_2}{\ell-1}\right)^2 |||\phi_{T,j}|||^2_{h,\Omega\setminus\omega_T^{\ell(k-2)-1}} \\
&\leq \cdots \leq \left(\frac{C_2}{\ell-1}\right)^{k-1} |||\phi_{T,j}|||^2_{h,\omega^\ell\setminus\omega_T^{\ell-1}},
\end{aligned}
$$

which together with (4.8) concludes the proof. ∎

LEMMA 12. *For all $T \in \mathcal{T}_H$, $j = 1, \ldots, r$, the estimate*

$$
\left|\!\left|\!\left| \sum_{T\in\mathcal{T}_H,\, j=1,\ldots,r} v_j(\phi_{T,j} - \phi_{T,j}^L) \right|\!\right|\!\right|^2_h \leq C_4 L^d \sum_{T\in\mathcal{T}_H,\, j=1,\ldots,r} |v_j|^2 |||\phi_{T,j} - \phi_{T,j}^L|||^2_h
$$

is satisfied with $C_4 = C C^3_{\beta/\alpha}$ and C being a positive constant independent of the mesh (h or H), of the patch size L, and of the diffusion matrix A.

Proof. Let $w = \sum_{T\in\mathcal{T}_H,\, j=1,\ldots,r} v_j(\phi_{T,j} - \phi_{T,j}^L)$, and note that

$$
\begin{aligned}
a_h(\phi_{T,j} - \lambda_{T,j}, w - \zeta_T w + b_T) &= 0, \\
a_h(\phi_{T,j}^L - \lambda_{T,j}, w - \zeta_T w + b_T) &= 0,
\end{aligned} \tag{4.16}
$$

where $\zeta_T := \zeta_T^{L+1,L+2}$, using Lemma 6 and the property of the cutoff function. We obtain

(4.17)
$$\begin{aligned}
\Big|\Big|\Big| \sum_{T\in\mathcal{T}_H,\, j=1,\dots,r} v_j(\phi_{T,j}-\phi_{T,j}^L) \Big|\Big|\Big|_h &\lesssim \sum_{T\in\mathcal{T}_H,\, j=1,\dots,r} v_j a_h(\phi_{T,j}-\phi_{T,j}^L, w)\\
&= \sum_{T\in\mathcal{T}_H,\, j=1,\dots,r} v_j a_h(\phi_{T,j}-\phi_{T,j}^L, \zeta_T w - b_T)\\
&\lesssim \sum_{T\in\mathcal{T}_H,\, j=1,\dots,r} |v_j|\cdot |||\phi_{T,j}-\phi_{T,j}^L|||_h \left(|||\zeta_T w|||_h + |||b_T|||_h\right)\\
&\lesssim \sum_{T\in\mathcal{T}_H,\, j=1,\dots,r} |v_j|\cdot |||\phi_{T,j}-\phi_{T,j}^L|||_h\\
&\quad \times \left(|||\zeta_T w|||_h + C_{\beta/\alpha}|||\Pi_H \zeta_T w|||_H\right)\\
&\lesssim \sum_{T\in\mathcal{T}_H,\, j=1,\dots,r} |v_j|\cdot |||\phi_{T,j}-\phi_{T,j}^L|||_h C_{\beta/\alpha}^2 |||\zeta_T w|||_h.
\end{aligned}$$

From (4.6) and (4.7), we have

$$|||\zeta_T w|||_h = |||\zeta_T w|||_{h,\omega_T^{L+2}} \lesssim C_{\beta/\alpha}|||w|||_{h,\omega_T^{L+2}}. \tag{4.18}$$

Then, further estimation of (4.17) can be achieved using (4.18) and the discrete Cauchy–Schwarz inequality:

$$\begin{aligned}
&\Big|\Big|\Big| \sum_{T\in\mathcal{T}_H,\, j=1,\dots,r} v_j(\phi_{T,j}-\phi_{T,j}^L) \Big|\Big|\Big|\\
&\quad\le C_{\beta/\alpha}^3 \left(\sum_{T\in\mathcal{T}_H,\, j=1,\dots,r} |v_j|^2 |||\phi_{T,j}-\phi_{T,j}^L|||_h^2\right)^{1/2} \left(\sum_{T\in\mathcal{T}_H,\, j=1,\dots,r} |||w|||_{h,\omega_T^{L+2}}^2\right)^{1/2}\\
&\quad\le C_{\beta/\alpha}^3 L^{d/2}\cdot \left(\sum_{T\in\mathcal{T}_H,\, j=1,\dots,r} |v_j|^2 |||\phi_{T,j}-\phi_{T,j}^L|||_h^2\right)^{1/2} \cdot |||w|||_h.
\end{aligned}$$

Dividing by w on both sides concludes the proof. ∎

The following theorem gives an error bound for the dG multiscale method.

THEOREM 13. *Let $u \in H_D^1(\Omega)$ solve (1.2) and let $u_H^{ms,L} \in \mathcal{V}_H^{ms,L}$ solve (3.5). Then, the estimate*

$$\begin{aligned}
|||u-u_H^{ms,L}|||_h \le &|||u-u_h|||_h + C_1\alpha^{-1/2}||H(f-\Pi_H f)||_{L^2(\Omega)}\\
&+ C_5\|H^{-1}\|_{L^\infty(\Omega)} L^{d/2}\gamma^L \|f\|_{L^2(\Omega)}
\end{aligned}$$

is satisfied with $0<\gamma<1$, L from Lemma 11, C_1 from Theorem 9, and $C_5 = CC_{\beta/\alpha}^2 C_4^{1/2} C_3$, where C_3 is from Lemma 11 and C_4 is from Lemma 12; C is a positive constant independent of the mesh (h or H), of the patch size L, and of the diffusion matrix A.

Remark 14. To counteract the factor $\|H^{-1}\|_{L^\infty(\Omega)}$ in the error bound in Theorem 13, we can choose the localization parameter as $L = \lceil C\log(||H^{-1}||_{L^\infty(\Omega)})\rceil$. On adaptively refined meshes it is recommended to choose $L = \lceil C\log(H^{-1})\rceil$.

Proof. We define $\tilde{u}_H^{ms,L} := \sum_{T\in\mathcal{T}_H,\, j=1,\dots,r} u_{H,T}^{ms}(x_j)\phi_{T,j}^L$. Then, we obtain

$$
\begin{aligned}
(4.19)\quad |||u - u_H^{ms,L}|||_h &\le |||u - \tilde{u}_H^{ms,L}|||_h \\
&\le |||u - u_h|||_h + |||u_h - u_H^{ms}|||_h + |||u_H^{ms} - \tilde{u}_H^{ms,L}|||_h \\
&\le |||u - u_h|||_h + |||u_h - u_H^{ms}|||_h + \Big|\Big|\Big| \sum_{T\in\mathcal{T}_H,\, j=1,\dots,r} u_{H,T}^{ms}(x_j)(\phi_{T,j} - \phi_{T,j}^L)\Big|\Big|\Big|_h.
\end{aligned}
$$

Now, estimating the terms in (4.19), we have

$$|||u_h - u_H^{ms}|||_h \le C_1\alpha^{-1/2}\|H(f - \Pi_H f)\|_{L^2(\Omega)}$$

using Theorem 9 and

$$
\begin{aligned}
(4.20)\quad &\Big|\Big|\Big| \sum_{T\in\mathcal{T}_H,\, j=1,\dots,r} u_{H,T}^{ms}(x_j)(\phi_{T,j} - \phi_{T,j}^L)\Big|\Big|\Big|_h^2 \\
&\le C_4 L^d \sum_{T\in\mathcal{T}_H,\, j=1,\dots,r} |u_{H,T}^{ms}(x_j)|^2 |||\phi_{T,j} - \phi_{T,j}^L|||_h^2. \\
&\le C_4 C_3^2 L^d \gamma^{2L} \sum_{T\in\mathcal{T}_H,\, j=1,\dots,r} |u_{H,T}^{ms}(x_j)|^2 |||\phi_{T,j} - \lambda_j|||_h^2
\end{aligned}
$$

using Lemmas 12 and 11, respectively. Further estimation, using Lemma 8, yields

$$
\begin{aligned}
(4.21)\quad &\sum_{T\in\mathcal{T}_H,\, j=1,\dots,r} |u_{H,T}^{ms}(x_j)|^2 |||\phi_{T,j} - \lambda_{T,j}|||_h^2 \\
&\lesssim C_{\beta/\alpha}^2 \sum_{T\in\mathcal{T}_H,\, j=1,\dots,r} |u_{H,T}^{ms}(x_j)|^2 |||\lambda_{T,j}|||_H^2 \\
&\lesssim C_{\beta/\alpha}^2 \beta \sum_{T\in\mathcal{T}_H,\, j=1,\dots,r} |u_{H,T}^{ms}(x_j)|^2 H_T^{-2} \|\lambda_{T,j}\|_{L^2(T)}^2 \\
&= C_{\beta/\alpha}^2 \beta \sum_{T\in\mathcal{T}_H,\, j=1,\dots,r} \|H_T^{-1} u_{H,T}^{ms}(x_j)\lambda_{T,j}\|_{L^2(T)}^2 \\
&\lesssim C_{\beta/\alpha}^2 \beta \Big\| \sum_{T\in\mathcal{T}_H,\, j=1,\dots,r} H_T^{-1} u_{H,T}^{ms}(x_j)\lambda_{T,j}\Big\|_{L^2(\Omega)}^2.
\end{aligned}
$$

Furthermore, using a Poincaré–Friedrichs inequality for piecewise H^1 functions, we deduce

$$
\begin{aligned}
(4.22)\quad &\Big\| \sum_{T\in\mathcal{T}_H,\, j=1,\dots,r} H_T^{-1} u_{H,T}^{ms}(x_j)\lambda_{T,j}\Big\|_{L^2(\Omega)}^2 \\
&\lesssim \Big\| \sum_{T\in\mathcal{T}_H,\, j=1,\dots,r} H_T^{-1} u_{H,T}^{ms}(x_j)\Pi_H(\lambda_{T,j} - \phi_{T,j})\Big\|_{L^2(\Omega)}^2 \\
&\lesssim \alpha^{-1} |||H^{-1} u_H^{ms}|||_h^2 \\
&\lesssim \alpha^{-1} \|H^{-1}\|_{L^\infty(\Omega)}^2 \|f\|_{L^2(\Omega)}^2.
\end{aligned}
$$

Combining (4.20), (4.21), and (4.22), we arrive at

$$|||u_H^{ms} - u_H^{ms,L}|||_h \lesssim C_{\beta/\alpha}^2 C_4^{1/2} C_3 \|H^{-1}\|_{L^\infty(\Omega)} L^{d/2}\gamma^L \|f\|_{L^2(\Omega)}. \qquad \square$$

4.3. Error in a quantity of interest. In engineering applications, we are often interested in a quantity of interest, usually a functional $G(v)$ of the solution. To this end, consider the dual reference solution (2.5): find $\phi_h \in \mathcal{V}_h$ such that

$$a_h(v, \phi_h) = G(v) \quad \text{for all } v \in \mathcal{V}_h; \tag{4.23}$$

and consider the dual multiscale solution (3.5): find $\phi_H^{ms,L} \in \mathcal{V}_H^{ms,L}$ such that

$$a_h(v, \phi_H^{ms,L}) = G(v) \quad \text{for all } v \in \mathcal{V}_H^{ms,L}. \tag{4.24}$$

THEOREM 15. *Let $u \in H_D^1(\Omega)$ solve (1.2), let $u_H^{ms,L} \in \mathcal{V}_H^{ms,L}$ solve (3.5), and let $G(v)$ be the quantity of interest. Then, the estimate*

$$|G(u) - G(u_H^{ms,L})| \lesssim |G(u) - G(u_h)| + |||u_h - u_H^{ms,L}|||_h |||\phi_h - \phi_H^{ms,L}|||_h$$

is satisfied.

Proof. From (4.23) and (4.24), we obtain the Galerkin orthogonality

$$a_h(v, \phi_h - \phi_H^{ms,L}) = 0 \quad \text{for all } v \in \mathcal{V}_H^{ms,L}. \tag{4.25}$$

Using the triangle inequality, we have

$$|G(u) - G(u_H^{ms,L})| \leq |G(u) - G(u_h)| + |G(u_h) - G(u_H^{ms,L})|.$$

Finally, observing that

$$\begin{aligned} |G(u_h) - G(u_H^{ms,L})| &= |a_h(u_h - u_H^{ms,L}, \phi_h)| \\ &= |a_h(u_h - u_H^{ms,L}, \phi_h - \phi_H^{ms,L})| \\ &\lesssim |||u_h - u_H^{ms,L}|||_h |||\phi_h - \phi_H^{ms,L}|||_h, \end{aligned}$$

using (4.25), concludes the proof. ∎

COROLLARY 16. *For $G(v) = (u_h - u_H^{ms,L}, v)_{L^2(\Omega)}$, the following L^2-norm error estimates hold:*

$$\|u - u_H^{ms,L}\|_{L^2(\Omega)} \lesssim \|u - u_h\|_{L^2(\Omega)} + |||u_h - u_H^{ms,L}|||_h^{1/2} |||\phi_h - \phi_H^{ms,L}|||_h^{1/2}$$

and

$$\|u - u_H^{ms,L}\|_{L^2(\Omega)} \lesssim \|u - u_h\|_{L^2(\Omega)} + H|||u_h - u_H^{ms,L}|||_h \tag{4.26}$$

for $L = \lceil C \log(H^{-1}) \rceil$ with C a sufficiently large positive constant independent of the mesh parameters.

Remark 17. As expected, if we are interested in a bounded linear functional with additional smoothness, a higher convergence rate is obtained for $|G(u_h) - G(u_H^{ms,L})|$. For example, given the forcing function for the primal problem $f \in H^m(\mathcal{T}_H)$, a quantity of interest $G(v) = (g, v)_{L^2(\Omega)}$, where $g \in H^n(\mathcal{T}_H)$ (with $H^0(\mathcal{T}_H)$ denoting the $L^2(\Omega)$ space), and choosing $L = \lceil C \log(H^{-1}) \rceil$ with large enough C gives

$$\begin{aligned} |G(u) - G(u_H^{ms,L})| \lesssim{}& |G(u) - G(u_h)| \\ &+ H^{2+m+n} \left(\sum_{T \in \mathcal{T}_H} |f|_{H^m(T)}^2 \right)^{1/2} \left(\sum_{T \in \mathcal{T}_H} |g|_{H^n(T)}^2 \right)^{1/2} \end{aligned}$$

for $2 \geq m, n \in \mathbb{N}$.

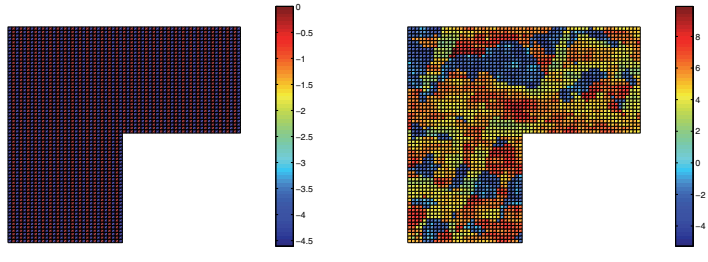
5. Numerical experiments. Let Ω be an L-shaped domain (constructed by removing the lower right quadrant in the unit square) and let the forcing function be $f = 1 + \cos(2\pi x)\cos(2\pi y)$ for $(x, y) \in \Omega$. The boundary Γ is divided into the Neumann boundary $\Gamma_N := \Gamma \cap (\{(x, y) : y = 0\} \cup \{(x, y) : x = 1\})$ and the Dirichlet boundary $\Gamma_D = \Gamma \setminus \Gamma_N$. We shall consider three different permeabilities: constant $A_1 = 1$, $A_2 = A_2(x)$, which is piecewise constant with periodic values of 1 and 0.01 with respect to a Cartesian grid of width 2^{-6} in the x-direction, and $A_3 = A_3(x, y)$, which is piecewise constant with respect to a Cartesian grid of width 2^{-6} in both the x- and y-directions and has a maximum ratio $\beta/\alpha = 4 \cdot 10^6$. The data for A_3 are taken from layer 64 in the SPE benchmark problem (see http://www.spe.org/web/csp). The permeabilities A_2 and A_3 are illustrated in Figure 5.1. For the periodic problem many of the corrected basis functions will be identical. For instance, all the local corrected bases in the interior are solved on identical patches, reducing the computational effort considerably. In the extreme case of a problem with periodic coefficients on a unit hypercube, with period boundary conditions, the correctors $\phi_{T,j}$, $j = 1, \ldots, r$, will be identical for all $T \in \mathcal{T}_H$.

Consider the uniform (coarse) quadrilateral mesh $\mathcal{T}_H$ with size $H = 2^{-i}$, $i = 1, \ldots, 6$. The convergence rate $-p/2$ corresponds to $\mathcal{O}(H^p)$ since the number of degrees of freedom $\approx H^{-2}$. The corrector functions (3.3) are solved on a subgrid of a (fine) quadrilateral mesh $\mathcal{T}_h$ with mesh size 2^{-8}. The mesh $\mathcal{T}_h$ will also act as a reference grid on which we shall compute a reference solution $u_h \in \mathcal{V}_h$ (2.5). Note that the mesh $\mathcal{T}_h$ is chosen so that it resolves the fine scale features of A_i, $i = 1, 2, 3$; hence we assume that the solution u_h is sufficiently accurate.

5.1. Localization parameter. If $f \in H^m(\mathcal{T}_H)$ we have the bound

$$||H(f - \Pi_H f))||_{L^2(\Omega)} \lesssim \left(\sum_{T \in \mathcal{T}_H} H^{2k+2} |f|^2_{H^k(T)} \right)^{1/2}, \tag{5.1}$$

where $k = 0$ for $m = 0$, $k = 1$ for $m = 1$, and $k = 2$ for $m > 1$. Hence, to balance the error between the terms on the right-hand side of the estimate in Theorem 13, a different constant C has to be used for the localization parameter, $L = \lceil C \log(H^{-1}) \rceil$, depending on the elementwise regularity of the forcing function f on $\mathcal{T}_H$. Figure 5.2 shows the relative error in the energy norm $|||u_h - u_H^{ms,L}|||_h / |||u_h|||_h$ and Figure 5.3



(a) $\beta/\alpha = 10^2$ (b) $\beta/\alpha \approx 4 \cdot 10^6$

FIG. 5.1. *The permeability structure of* (a) A_2 *and* (b) A_3 *in log scale.*

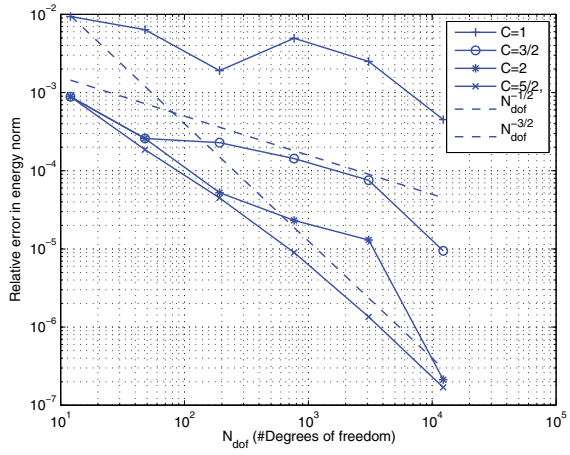


FIG. 5.2. *Diffusion coefficient* $A_1 = 1$. *Relative energy-norm error against* N_{dof} *for different values of* C *for the localization parameter* L.

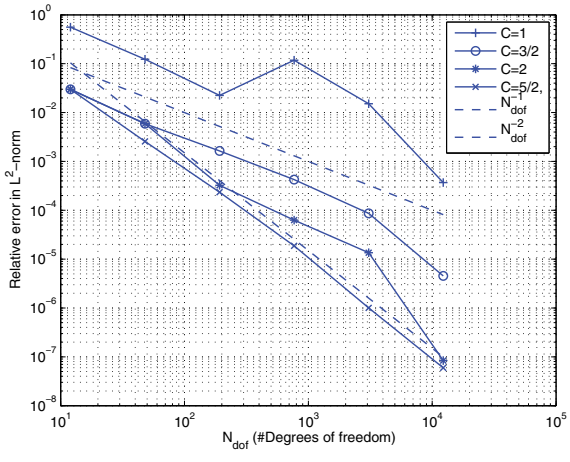


FIG. 5.3. *Diffusion coefficient* $A_1 = 1$. *Relative* L^2*-norm error against* N_{dof} *for different values of* C *for the localization parameter* L.

the relative error in the L^2-norm $\|u_h - u_H^{ms,L}\|_{L^2(\Omega)}/\|u_h\|_{L^2(\Omega)}$ between u_h and $u_H^{ms,L}$ against the number of degrees of freedom $N_{\text{dof}} \approx O(H^{-2})$, using different constants $C = 1, 3/2, 2, 5/2$. With the choice $C = 5/2$ the errors due to the localization can be neglected compared to the errors from the forcing function for both the energy and the L^2-norm. For $f \notin H^1(\mathcal{T}_h)$, $C = 3/2$ is sufficient since (5.1) gives linear convergence. In the remaining numerical experiments we use $C = 2$, since this value seems to balance the error sufficiently. Note that the numerical overhead increases with C as

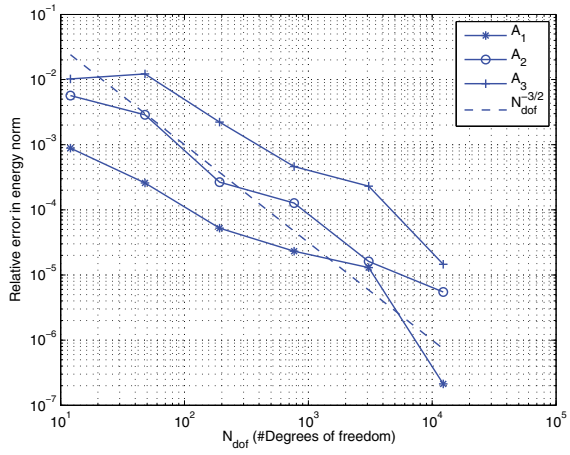


FIG. 5.4. *Relative energy-norm error against N_{dof} for $C = 2$ in the localization parameter L for the the diffusion coefficients A_1, A_2, and A_3.*

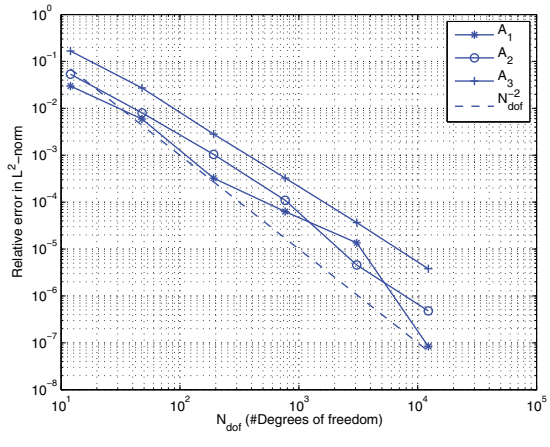


FIG. 5.5. *Relative L^2-norm error against N_{dof} for $C = 2$ in the localization parameter L for the diffusion coefficients A_1, A_2, and A_3.*

the sizes of the patches ω_T^L, $T \in \mathcal{T}_H$, increase with $L = \lceil C \log(H^{-1}) \rceil$. This results in both increased computational cost to compute the corrector functions and reduced sparseness in the coarse scale stiffness matrix.

5.2. Energy-norm convergence. Let the localization parameter be given by $L = \lceil 2 \log(H^{-1}) \rceil$. Figure 5.4 shows the relative error in the energy norm plotted against the number of degrees of freedom. The different permeabilities A_i, $i = 1, 2, 3$, and the singularity arising from the L-shaped domain do not appear to have a

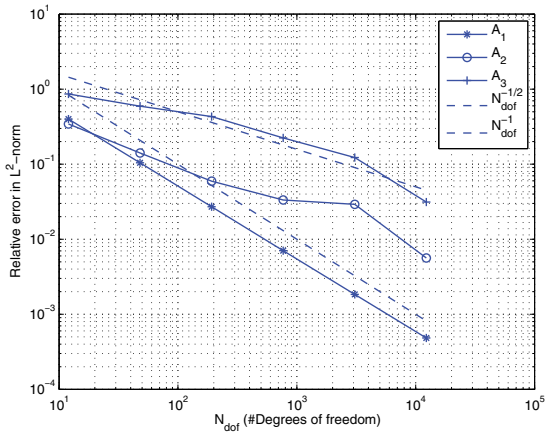


FIG. 5.6. *Relative L^2-norm error against N_{dof} for $C = 2$ in the localization parameter L for the diffusion coefficients A_1, A_2, and A_3.*

substantial impact on the convergence rate, which is about $-3/2$, as expected. We note in passing that using standard dG on the coarse mesh only admits poor convergence behavior for A_2 and for A_3. This is to be expected, since standard dG on the coarse mesh does not resolve the fine scale features.

5.3. L^2-norm convergence. Again, set $L = \lceil 2\log(H^{-1})\rceil$. Figures 5.5 and 5.6 show the relative L^2-norm error against the number of degrees of freedom between u_h and $u_H^{ms,L}$ and between u_h and $\Pi_H u_H^{ms,L}$, viz., $\|u_h - \Pi_H u_H^{ms,L}\|_{L^2(\Omega)}/\|u_h\|_{L^2(\Omega)}$, respectively. In Figure 5.5 we see that the L^2-norm error between u_h and $u_H^{ms,L}$ converges at a faster rate than in the energy norm (convergence rate -2 compared to $-3/2$, respectively,) as expected from (4.26). In Figure 5.6 only the coarse part of $u_H^{ms,L}$ is used (i.e., $\Pi_H u_H^{ms,L}$); nevertheless it appears to have a faster convergence rate than $-1/2$, except for the case of the permeability A_3.

6. Concluding remarks. We present a dG multiscale method for second order elliptic problem with heterogeneous and highly varying diffusion coefficients in $L^\infty(\Omega, \mathbb{R}^{d\times d}_{sym})$ with uniform spectral bounds. For $f \in L^2(\Omega)$, the method seeks the solution $u_H^{ms,L}$ (from (3.5)) in a space of corrected basis function (3.4) calculated on patches of size $\mathcal{O}(H\log(H^{-1}))$ (i.e., $L = \lceil C\log(H^{-1})\rceil$). We have shown the error bounds

$$|||u - u_H^{ms,L}|||_h \leq |||u - u_h|||_h + C_{\alpha/\beta,f}H$$

in the energy norm (Theorem 13) and

$$\|u - u_H^{ms,L}\|_{L^2(\Omega)} \leq \|u - u_h\|_{L^2(\Omega)} + \tilde{C}_{\alpha/\beta,f}H^2$$

in the $L^2(\Omega)$-norm (Corollary 16), where u_h (from (2.5)) is the reference solution on the fine scale and the constants $C_{\alpha/\beta,f}$ and $\tilde{C}_{\alpha/\beta,f}$ depend on the forcing function f and the global bound of the diffusion matrix, but not on its variations. Numerical

experiments show that choosing the localization parameter as $L = \lceil 2\log(H^{-1})\rceil$ is sufficient to achieve good convergence for the diffusion coefficients (A_1, A_2, and A_3).

Appendix A. Equalities for averages and jump operators. We derive equalities for averages and jump operators across interfaces where the functions v and w have discontinuities. Using $[vw] = \{v\}[w] + [v]\{w\}$ and $\{v\}\{w\} = \{vw\} - 1/4[v][w]$, we have

$$\begin{aligned}\{vw\}[vu] &= \{w\}\{v\}[vu] + 1/4[v][w][vu]\\ &= \{w\}[v^2u] - \{w\}[v]\{vu\} + 1/4[v][w][vu]\\ &= \{w\}[v^2u] - [v]\{w\}\{v\}\{u\} - 1/4[v]^2\{w\}[u]\\ &\quad + 1/4[v]^2[w]\{u\} + 1/4[v]\{v\}[w][u]\end{aligned} \tag{A.1}$$

and

$$\begin{aligned}\{vw\}[vu] &= \{v\}\{vw\}[u] + \{vw\}[v]\{u\}\\ &= \{v^2w\}[u] - 1/4[v][vw][u] + \{vw\}[v]\{u\}\\ &= \{v^2w\}[u] - 1/4[v]^2\{w\}[u] - 1/4[v]\{v\}[w][u]\\ &\quad + [v]\{v\}\{w\}\{u\} + 1/4[v]^2[w]\{u\}.\end{aligned} \tag{A.2}$$

Combining (A.1) and (A.2) we obtain

$$2\{vw\}[vu] = \{w\}[v^2u] + \{v^2w\}[u] + 1/2[v]^2[w]\{u\} - 1/2[v]^2\{w\}[u]. \tag{A.3}$$

Also,

$$\begin{aligned}[vu][vu] &= [u]\{v\}[vu] + [v]\{u\}[vu]\\ &= [u][v^2u] - [v][u]\{vu\} + [v]\{u\}[vu]\\ &= [u][v^2u]) - [v][u]\{v\}\{u\} - 1/4[v][u][v][u]\\ &\quad + [v]\{u\}[v]\{u\} + [v]\{u\}\{v\}[u]\\ &= [u][v^2u] - 1/4[v]^2[u]^2 + [v]^2\{u\}^2.\end{aligned} \tag{A.4}$$

REFERENCES

[1] T. Arbogast, G. Pencheva, M. F. Wheeler, and I. Yotov, *A multiscale mortar mixed finite element method*, Multiscale Model. Simul., 6 (2007), pp. 319–346.

[2] T. Arbogast and H. Xiao, *A multiscale mortar mixed space based on homogenization for heterogeneous elliptic problems*, SIAM J. Numer. Anal., 51 (2013), pp. 377–399.

[3] D. N. Arnold, *An interior penalty finite element method with discontinuous elements*, SIAM J. Numer. Anal., 19 (1982), pp. 742–760.

[4] I. Babuška, G. Caloz, and J. E. Osborn, *Special finite element methods for a class of second order elliptic problems with rough coefficients*, SIAM J. Numer. Anal., 31 (1994), pp. 945–981.

[5] I. Babuška 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.

[6] I. Babuška and J. E. Osborn, *Generalized finite element methods: Their performance and their relation to mixed methods*, SIAM J. Numer. Anal., 20 (1983), pp. 510–536.

[7] L. Berlyand and H. Owhadi, *Flux norm approach to finite dimensional homogenization approximations with non-separated scales and high contrast*, Arch. Ration. Mech. Anal., 198 (2010), pp. 677–721.

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

[9] J. Douglas and T. Dupont, *Interior penalty procedures for elliptic and parabolic Galerkin methods*, in Computing Methods in Applied Sciences, Lecture Notes in Phys., 58, Springer, Berlin, 1976, pp. 207–216.

[10] M. Dryja, *On discontinuous Galerkin methods for elliptic problems with discontinuous coefficients*, Comput. Methods Appl. Math., 3 (2003), pp. 76–85.

[11] Y. Efendiev, J. Galvis, and T. Y. Hou, *Generalized Multiscale Finite Element Methods (GMsFEM)*, J. Comput. Phys., 251 (2013), pp. 116–135.

[12] Y. Efendiev, J. Galvis, R. Lazarov, M. Moon, and M. Sarkis, *Generalized Multiscale Finite Element Method. Symmetric Interior Penalty Coupling*, J. Comput. Phys., 255 (2013), pp. 1–15.

[13] D. Elfverson, E. Georgoulis, and A. Målqvist, *An adaptive discontinuous Galerkin multiscale method for elliptic problems*, Multiscale Model. Simul., 11 (2013), pp. 747–765.

[14] D. Elfverson and A. Målqvist, *Discontinuous Galerkin Multiscale Methods for Convection Dominated Problems*, Tech. report 2013-011, Department of Information Technology, Uppsala University, Sweden, 2013.

[15] P. Henning, A. Målqvist, and D. Peterseim, *Two-Level Discretization Techniques for Ground State Computations of Bose-Einstein Condensates*, arXiv:1305.4080, 2013.

[16] P. Henning, A. Målqvist, and D. Peterseim, *A localized orthogonal decomposition method for semi-linear elliptic problems*, arXiv:1211.3551.

[17] P. Henning and D. Peterseim, *Oversampling for the Multiscale Finite Element Method*, Multiscale Model. Simul., 11(2013), pp. 1149–1175.

[18] T. Y. Hou, T. Y. Hou, and X. Wu, *A multiscale finite element method for elliptic problems in composite materials and porous media*, J. Comput. Phys., 134 (1997), pp. 169–189.

[19] 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.

[20] P. Houston, D. Schötzau, and T. P. Wihler, *Energy norm a posteriori error estimation of hp-adaptive discontinuous Galerkin methods for elliptic problems*, Math. Models Methods Appl. Sci., 17 (2007), pp. 33–62.

[21] T. 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.

[22] T. Hughes, G. 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.

[23] O. Karakashian and F. Pascal, *A posteriori error estimates for a discontinuous Galerkin approximation of second-order elliptic problems*, SIAM J. Numer. Anal., 41 (2003), pp. 2374–2399.

[24] 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.

[25] A. Målqvist and D. Peterseim, *Localization of elliptic multiscale problems*, Math. Comp., arXiv:1110.0692.

[26] A. Målqvist and D. Peterseim, *Computation of Eigenvalues by Numerical Upscaling*, arXiv:1212.0090, 2012.

[27] A. Målqvist, *Multiscale methods for elliptic problems*, Multiscale Model. Simul., 9 (2011), pp. 1064–1086.

[28] 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.

[29] R. Scheichl, P. S. Vassilevski, and L. T. Zikatanov, *Multilevel methods for elliptic problems with highly varying coefficients on nonaligned coarse grids*, SIAM J. Numer. Anal., 50 (2012), pp. 1675–1694.

Paper III

A discontinuous Galerkin multiscale method for convection-diffusion problems

Daniel Elfverson

Abstract

We propose an discontinuous Galerkin local orthogonal decomposition multiscale method for convection-diffusion problems with rough, heterogeneous, and highly varying coefficients. The properties of the multiscale method and the discontinuous Galerkin method allows us to better cope with multiscale features as well as interior/boundary layers in the solution. In the proposed method the trail and test spaces are spanned by a corrected basis computed on localized patches of size $\mathcal{O}(H\log(H^{-1}))$, where H is the mesh size. We prove convergence rates independent of the variation in the coefficients and present numerical experiments which verify the analytic findings.

1 Introduction

In this paper we consider numerical approximation of convection-diffusion problems with possible strong convection and with rough, heterogeneous, and highly varying coefficients, without assumption on scale separation or periodicity. This class of problems, normally refereed to as multiscale problem, are know to be very computational demanding and arise in many different areas of the engineering sciences, e.g., in porous media flow and composite materials. More precisely, we consider the following convection-diffusion equation: given any $f \in L^2(\Omega)$ we seek $u \in H_0^1(\Omega) = \{v \in H^1(\Omega) \mid v|_\Gamma = 0\}$ such that

$$-\nabla \cdot A\nabla u + \mathbf{b} \cdot \nabla u = f \quad \text{in } \Omega, \tag{1}$$

is fulfilled in a weak sense, where Ω is the computational domain with boundary Γ. The multiscale coefficients $A, \mathbf{b}$ will be specified later. There are two key issues which make classical conforming finite element methods perform badly for these kind of problems,

- the multiscale features of the coefficient need to be resolved by the finite element mesh and

- strong convection leads to boundary and interior layers in the solution which need to be resolved.

To overcome the lack of performance using classical finite element methods in the case of multiscale features in the coefficient many different so called multiscale methods have been proposed, see [25, 26, 23, 7, 13, 12, 10, 11] among others, which perform localized fine scale computations to construct a different basis or a modified coarse scale operator. Common to the aforementioned approaches is that the performance of the method rely strongly on scale separation or periodicity of the diffusion coefficients. There is also approaches which perform well without scale separation or periodicity in the diffusion coefficient but to high computational cost by either having to solve eigenvalue problems [2] or where the support of the localized patches is large [37, 4]. See also [38].

In the variational multiscale method (VMS) framework [25, 26] the solution space is split into coarse and fine scale contribution. This idea was employed for multiscale problems in a adaptive setting for classical finite element in [31, 34, 32] and to the discontinuous Galerkin (DG) method in [14]. A further development is the local orthogonal decomposition (LOD) method, see [36, 20, 19] for classical finite element and [15] for DG methods. The LOD operates in linear complexity without any assumptions on scale separation or periodicity and the trail and test spaces are spanned by a corrected basis function computed on patches of size $\mathcal{O}(H\log(H^{-1}))$. The LOD has e.g. been applied to eigenvalue problems [35], non-linear elliptic problems [21], non-linear Schrödinger equation [17], and in Petrov-Galerkin formulation [16].

There is a vast literature on numerical methods for convection dominated problems, we reefer to [28, 24, 27], among others. There has also been a lot of work on DG methods, we refer to [39, 33, 3, 29] for some early work and to [8, 22, 40, 9] and references therein for recent development and a literature review. DG methods exhibit attractive properties for convection dominated problems, e.g., they have enhanced stability properties, good conservation property of the state variable, and the use of complex and/or irregular meshes are admissible. For multiscale methods for convection-diffusion problems, see e.g. [1, 41, 18].

In this paper we extended the analysis of the discontinuous Galerkin local orthogonal decomposition (DG-LOD) [15] to convection-diffusion problems. For problems with strong convection using the standard LOD won't suffice, since convergence can no longer be guarantied. Instead we propose to include the convective term in the computations of the corrected basis functions. We prove convergence results under some assumptions of the magnitude of the

convection and present a series of numerical experiment to verify the analytic findings. For problems with weak convection it is not necessary to include the convective part [21].

The outline of this paper is as follows. In section 2 the discrete setting and underlying DG method is presented. In section 3 the multiscale decomposition, the DG-LOD, and the corresponding convergence result are stated. In Section 4 numerical experiments are presented. Finally, the proofs for some of the theoretical results are given in Section 5.

2 Preliminaries

In this section we present some notations and properties frequently used in the paper.

2.1 Setting

Let $\Omega \subset \mathbb{R}^d$ for $d = 2, 3$ be a polygonal domain with Lipschitz boundary Γ. We assume that: the diffusion coefficients, $A \in L^\infty(\Omega, \mathbb{R}^{d\times d}_{sym})$, has uniform spectral bounds $0 < \alpha, \beta < \infty$, defined by

$$0 < \alpha := \operatorname*{ess\,inf}_{x\in\Omega} \inf_{v\in\mathbb{R}^d\setminus\{0\}} \frac{(A(x)v)\cdot v}{v\cdot v} \leq \operatorname*{ess\,sup}_{x\in\Omega} \sup_{v\in\mathbb{R}^d\setminus\{0\}} \frac{(A(x)v)\cdot v}{v\cdot v} =: \beta < \infty, \quad (2)$$

and the convective coefficient, $\mathbf{b} \in [W^1_\infty(\Omega)]^d$, is divergence free

$$\nabla \cdot \mathbf{b}(x) = 0 \text{ a.e. } x \in \Omega. \quad (3)$$

We denote $C_A = (\beta/\alpha)^{1/2}$.

We will consider a coarse and a fine mesh, with mesh function h and H respectively. Furthermore, we assume that the fine mesh resolve and that the coarse mesh do not resolve the fine scale features in the coefficients. Let $\mathcal{T}_k$, for $k = \{h, H\}$, denote a shape-regular subdivision of Ω into (closed) regular simplexes or into quadrilaterals/hexahedras ($d = 2/d = 3$), given a mesh function $k : \mathcal{T}_k \to \mathbb{R}$ defined as $k := \operatorname{diam}(T) \in P_0(\mathcal{T}_k)$ for all $T \in \mathcal{T}_k$. Also, let $\nabla_k v$ denote the $\mathcal{T}_k$-broken gradient defined as $(\nabla v)|_T = \nabla v|_T$ for all $T \in \mathcal{T}_k$. For simplicity we will also assume that $\mathcal{T}_k$ is conforming in the sense that no hanging nodes are allowed, but the analysis can easily be extend to non-conforming meshes with a finite number of hanging nodes on each edge. Let $\hat{T}$ be the reference simplex or (hyper)cube. We define $\mathcal{P}_p(\hat{T})$ to be the space of polynomials of degree less than or equal to p if $\hat{T}$ is a simplex, or the space of polynomials of degree less than or equal to p, in each variable, if $\hat{T}$

is a (hyper)cube. The space of discontinuous piecewise polynomial function is defined by

$$P_p(\mathcal{T}_k) := \{v : \Omega \to \mathbb{R} \mid \forall T \in \mathcal{T}_k,\, v|_T \circ F_T \in \mathcal{P}_p(\hat{T})\}, \tag{4}$$

where $F_T : \hat{T} \to T$, $T \in \mathcal{T}_k$ is a family of element maps. We will work with the spaces $\mathcal{V}_k := P_1(\mathcal{T}_k)$. Let $\Pi_p(\mathcal{T}_k) : L^2(\Omega) \to P_p(\mathcal{T}_k)$ denote the L^2-projection onto $P_p(\mathcal{T}_k)$. Also, let $\mathcal{E}_k$ denote the set of all edges in $\mathcal{T}_k$ where $\mathcal{E}_k(\Omega)$ and $\mathcal{E}_k(\Gamma)$ denote the set of interior and boundary edges, respectively. Given that T^+ and T^- are two adjacent elements in $\mathcal{T}_k$ sharing an edge $e = T^+ \cap T^- \in \mathcal{E}_k(\Omega)$, let ν_e be the the unit normal vector pointing from T^- to T^+, and for $e \in \mathcal{E}_k(\Gamma)$ let ν_e be outward unit normal of Ω. For any $v \in P_p(\mathcal{T}_k)$ we denote the value on edge $e \in \mathcal{E}(\Omega)$ as $v^\pm = v|_{e \cap T^\pm}$. The jump and average of $v \in P_p(\mathcal{T}_k)$ is defined as, $[v] = v^- - v^+$ and $\{v\} = (v^- + v^+)/2$ respectively for $e \in \mathcal{E}_k(\Omega)$, and $[v] = \{v\} = v|_e$ for $e \in \mathcal{E}_k(\Gamma)$. For a real number x we define its negative part as $x^\ominus = 1/2(|x| - x)$.

Let $0 \le C < \infty$ denote any generic constant that neither depends on the mesh size or the variables A and $\mathbf{b}$; then $a \lesssim b$ abbreviates the inequality $a \le Cb$.

2.2 Discontinuous Galerkin discretization

For simplicity let the bilinear form $a_h(\cdot,\cdot) : \mathcal{V}_h \times \mathcal{V}_h \to \mathbb{R}$, given any mesh function $h : \Omega \to P_0(\mathcal{T}_h)$, be split into two parts

$$a_h(u,v) := a_h^{\mathrm{d}}(u,v) + a_h^{\mathrm{c}}(u,v), \tag{5}$$

where $a_h^{\mathrm{d}}(\cdot,\cdot)$ represents the diffusion part and $a_h^{\mathrm{c}}(\cdot,\cdot)$ represents the convection part. The diffusion part is approximated using a symmetric interior penalty method

$$\begin{aligned} a_h^{\mathrm{d}}(u,v) := (A\nabla_h u, \nabla_h v)_{L^2(\Omega)} + \sum_{e \in \mathcal{E}_h} \Big(& \frac{\sigma_e}{h_e}([u],[v])_{L^2(e)} \\ & - (\{\nu_e \cdot A\nabla u\},[v])_{L^2(e)} - (\{\nu_e \cdot A\nabla v\},[u]_{L^2(e)}) \Big), \end{aligned} \tag{6}$$

where σ_e is a constant, depending on the diffusion, large enough to make $a_h^{\mathrm{d}}(\cdot,\cdot)$ coercive. The convective part is approximated by

$$\begin{aligned} a_h^{\mathrm{c}}(u,v) := (\mathbf{b} \cdot \nabla_h u, v)_{L^2(\Omega)} + & \sum_{e \in \mathcal{E}_h(\Omega)} (b_e[u],[v])_{L^2(e)} \\ - \sum_{e \in \mathcal{E}_h(\Omega)} (\nu_e \cdot \mathbf{b}[u],\{v\})_{L^2(e)} + & \sum_{e \in \mathcal{E}_h(\Gamma)} ((\nu_e \cdot \mathbf{b})^\ominus u, v)_{L^2(e)}, \end{aligned} \tag{7}$$

where upwind is imposed choosing the stabilization term as $b_e = |\mathbf{b} \cdot \nu_e|/2$ [5].

The following definitions and results are needed both on the fine and coarse scale, for this sake let $k = \{h, H\}$. The energy norm on $\mathcal{V}_k$ is given by

$$\begin{aligned} |||v|||_{k,\mathrm{d}}^2 &= \|A^{1/2}\nabla_k v\|_{L^2(\Omega)}^2 + \sum_{e\in\mathcal{E}_k} \frac{\sigma_e}{k}\|[v]\|_{L^2(e)}^2, \\ |||v|||_{k,\mathrm{c}}^2 &= \sum_{e\in\mathcal{E}_k} \|b_e^{1/2}[v]\|_{L^2(e)}^2, \\ |||v|||_k^2 &= |||v|||_{k,\mathrm{d}}^2 + |||v|||_{k,\mathrm{c}}^2. \end{aligned} \tag{8}$$

From Theorem 2.2 in [30] we have that for each $v \in \mathcal{V}_k$, there exist an averaging operator $\mathcal{I}_k^c : \mathcal{V}_k \to \mathcal{V}_k \cap H^1(\Omega)$ with the following property

$$\|\nabla_k(v - \mathcal{I}_k^c v)\|_{L^2(\Omega)} + \|k^{-1}(v - \mathcal{I}_k^c v)\|_{L^2(\Omega)} \lesssim \sum_{e\in\mathcal{E}_k} \frac{1}{k}\|[v]\|_{L^2(e)}^2. \tag{9}$$

In the error analysis we will also need a localized energy norm, defined in a domain $\omega \subset \Omega$ (aligned with the mesh $\mathcal{T}_k$) as

$$\begin{aligned} |||v|||_{k,\mathrm{d},\omega}^2 &= \|A^{1/2}\nabla_k v\|_{L^2(\omega)}^2 + \sum_{\substack{e\in\mathcal{E}_k \\ e\cap\bar{\omega}\neq 0}} \frac{\sigma_e}{k}\|[v]\|_{L^2(e)}^2, \\ |||v|||_{k,\mathrm{c},\omega}^2 &= \sum_{\substack{e\in\mathcal{E}_k \\ e\cap\bar{\omega}\neq 0}} \|b_e^{1/2}[v]\|_{L^2(e)}^2, \\ |||v|||_{k,\omega}^2 &= |||v|||_{k,\mathrm{d},\omega}^2 + |||v|||_{k,\mathrm{c},\omega}^2. \end{aligned} \tag{10}$$

3 Multiscale method

In this section we preset the multiscale decomposition and extend the results in [15] to convection-diffusion problems. For the constants in the convergence results to be stable we assume the following relation of the convective term

$$\mathcal{O}\left(\frac{\|H\mathbf{b}\|_{L^\infty(\Omega)}}{\alpha}\right) \leq 1 \tag{11}$$

How the magnitude of (11) affects the convergence of the method is investigated in the numerical experiments.

3.1 Multiscale decomposition

In order to do the multiscale decomposition the problem is divided into a coarse and a fine scale. To this end let $\mathcal{T}_H$ and $\mathcal{T}_h$, with the respective mesh function H and h, denote the two different subdivisions, where $\mathcal{T}_h$ is constructed by some (possible adaptive) refinements of $\mathcal{T}_H$.

The aim of this section is to construct a coarse finite element space based on $\mathcal{T}_H$, which takes the fine scale behavior of the data into account. We assume that the mesh $\mathcal{T}_h$ resolves the variation in the data, i.e., the solution to: find $u_h \in \mathcal{V}_h$ such that

$$a_h(u_h, v) = F(v) \quad \text{for all } v \in \mathcal{V}_h, \tag{12}$$

gives a sufficiently good approximation of the weak solution u to (1). Note however that u_h never have to be computed in practice, it only acts as a reference solution. We introduce a coarse projection operator $\Pi_H := \Pi_1(\mathcal{T}_H)$ and let the fine scale reminder space be defined by the kernel of Π_H, i.e.,

$$\mathcal{V}^{\mathrm{f}} := \{v \in \mathcal{V}_h \mid \Pi_H v = 0\} \subset \mathcal{V}_h. \tag{13}$$

The coarse projection operator has the following approximation and stability properties.

Lemma 1. *For any $v \in \mathcal{V}_h$ and $T \in \mathcal{T}_H$, the approximation property*

$$H|_T^{-1}\|v - \Pi_H v\|_{L^2(T)} \lesssim \alpha^{-1/2}|||v|||_{h,T}, \tag{14}$$

and stability estimate

$$|||\Pi_H v|||_H \lesssim C_s|||v|||_{\mathrm{h}}, \tag{15}$$

is satisfied, with

$$C_s = \left(C_A^2 + \frac{\|H\mathbf{b}\|_{L^\infty(\Omega)}}{\alpha}\right)^{1/2}. \tag{16}$$

Proof. The approximation property follows directly from [15, Lemma 5]. Let $\mathcal{C}_H : H^1 \to H^1 \cap \mathcal{V}_H$ be a Clément type interpolation operator proposed in [6, Section 6] which satisfy

$$\|\nabla \mathcal{C}_H u\|_{L^2(T)} + \|H^{-1}(u - \mathcal{C}_H u)\|_{L^2(T)} \lesssim \|\nabla u\|_{L^2(\omega_T^1)}, \tag{17}$$

where $\omega_T^1 = \mathrm{int}(\cup\{T' \in T_H \mid T \cap T \neq 0\})$ are the union of all elements that share a edge with T. We define the conforming function $v_c = \mathcal{C}_H \mathcal{I}_h^c v$ using

averaging operator in (9). We obtain

$$\begin{aligned}
|||\Pi_H v|||_H^2 &= \sum_{T\in\mathcal{T}_H} \|A^{1/2}\nabla(\Pi_H v - \Pi_0 v)\|_{L^2(T)}^2 \\
&\quad + \sum_{e\in\mathcal{E}_h(\Omega\cup\Gamma_D)} \left(\frac{\sigma}{H}\|[v_c - \Pi_H v]\|_{L^2(e)}^2 + \|b_e^{1/2}[v_c - \Pi_H v]\|^2\right) \\
&\lesssim \sum_{T\in\mathcal{T}_H} \beta\left(\frac{1}{H^2}\|v - \Pi_0 v\|_{L^2(T)}^2 + \left(\frac{1}{H^2} + \frac{\|\mathbf{b}\|_{L^\infty(T)}}{H}\right)\|v_c - v\|_{L^2(T)}^2\right)
\end{aligned} \tag{18}$$

using that $\Pi_0 := \Pi_0(\mathcal{T}_H)$ is the L^2-projection onto constants, a trace inequality, and stability of Π_H. Next, using that

$$\begin{aligned}
\|\mathcal{C}_H\mathcal{I}_h^c v - v\|_{L^2(\Omega)} &\leq \|\mathcal{C}_H\mathcal{I}_h^c v - \mathcal{I}_h^c v\|_{L^2(\Omega)} + \|\mathcal{I}_h^c v - v\|_{L^2(\Omega)} \\
&\lesssim H\|\nabla\mathcal{I}_h^c v\|_{L^2(\Omega)} + \|\mathcal{I}_h^c v - v\|_{L^2(\Omega)} \\
&\lesssim \alpha^{-1/2} H|||v|||_{\mathrm{h}}
\end{aligned} \tag{19}$$

in (18) concludes the proof. □

The following lemma shows that for every $v_H \in \mathcal{V}_H$ there exist a (non-unique) $v \in \Pi_H^{-1} v_H \in \mathcal{V}_h$ in the preimage of Π_H which is $H^1(\Omega)$ conforming.

Lemma 2. *For each $v_H \in \mathcal{V}_H$, there exist a $v \in \mathcal{V}_h \cap H^1(\Omega)$ such that $\Pi_H v = v_H$, $|||v|||_{\mathrm{h}} \lesssim C_A|||v_H|||_H$, and $\operatorname{supp}(v) \subset \operatorname{supp}(\mathcal{I}_H^c v_H)$.*

Proof. Follows directly from [15, Lemma 6], since $v \in H^1(\Omega)$. □

The next step is to split any $v \in \mathcal{V}_h$ into some coarse part based on $\mathcal{T}_H$, such that the fine scale reminder in the space $\mathcal{V}^{\mathrm{f}}$ is sufficiently small. A naive way to do this splitting is to use a L^2-orthogonal split. An alternative definition of the coarse space is $\mathcal{V}_H = \Pi_H\mathcal{V}_h$. A set of basis functions that span $\mathcal{V}_H$ is the element-wise Lagrange basis functions $\{\lambda_{T,j} \mid T \in \mathcal{T}_H, j = 1,\dots,r\}$ where $r = (1+d)$ for simplexes or $r = 2^d$ for quadrilaterals/hexahedra. The space $\mathcal{V}_H$ is known to give poor approximation properties if $\mathcal{T}_H$ does not resolve the variable coefficients in (1). We will use another choice, see [36, 15], based on $a_h(\cdot,\cdot)$, to construct a space of corrected basis functions. To this end, we define a fine scale projection operator $\mathfrak{F} : \mathcal{V}_h \to \mathcal{V}^{\mathrm{f}}$ by

$$a_h(\mathfrak{F}v, w) = a_h(v, w) \quad \text{for all } w \in \mathcal{V}^{\mathrm{f}}, \tag{20}$$

and let the corrected coarse space be defined as

$$\mathcal{V}_H^{ms} := (1 - \mathfrak{F})\mathcal{V}_H. \tag{21}$$

The corrected space are spanned by corrected basis functions $\mathcal{V}_H^{ms} := \{\lambda_{T,j} - \phi_{T,j} \mid T \in \mathcal{T}_H,\, j = 1, \ldots, r\}$ which can be computed as: for all $T \in \mathcal{T}_H$, $j = 1, \ldots, r$ find $\phi_{T,j} \in \mathcal{V}^{\mathrm{f}}$ such that

$$a_h(\phi_{T,j}, v) = a_h(\lambda_{T,j}, v) \quad \text{for all } v \in \mathcal{V}^{\mathrm{f}}. \tag{22}$$

Note that, $\dim(\mathcal{V}_H^{ms}) = \dim(\mathcal{V}_H)$. From (21) we have that any $v_h \in \mathcal{V}_h$ can be decomposed into a coarse $v_H^{ms} \in \mathcal{V}_H^{ms}$ and a fine $v^f \in \mathcal{V}^{\mathrm{f}}$ scale contribution, $v_h = v_H^{ms} + v^f$.

Lemma 3 (Stability of the corrected basis function)**.** *For all $T \in \mathcal{T}_H$, $j = 1, \ldots, r$, the following estimate*

$$|||\phi_{T,h} - \lambda_{T,j}|||_{\mathrm{h}} \lesssim C_\phi \beta^{1/2} \|H^{-1}\lambda_{T,j}\|_{L^2(\Omega)} \tag{23}$$

holds, where $C_\phi = (C_A^2 + \|H\mathbf{b}\|_{L^\infty(\Omega)}\alpha^{-1})^{1/2}$.

Proof. Let $v = \lambda_{T,j} - b_{T,j} \in \mathcal{V}^{\mathrm{f}}$, where $b_{T,j} \in H_0^1(T)$, $\Pi_H b_{T,j} = \lambda_{T,j}$, $|||b_{T,j}|||_{\mathrm{h}} \leq C_A |||\lambda_{T,j}|||_H$ from Lemma 2. We have

$$\begin{aligned} |||\phi_{T,h} - \lambda_{T,j}|||_{\mathrm{h}}^2 &\lesssim a_h(\phi_{T,h} - \lambda_{T,j}, \phi_{T,h} - \lambda_{T,j}) \\ &= a_h(\phi_{T,h} - \lambda_{T,j}, v - \lambda_{T,j}) = a_h(\phi_{T,h} - \lambda_{T,j}, b_{T,j}) \\ &= a_h^d(\phi_{T,h} - \lambda_{T,j}, b_{T,j}) + (\mathbf{b} \cdot \nabla_h(\phi_{T,h} - \lambda_{T,j}), b_{T,j})_{L^2(\Omega)}\,. \end{aligned} \tag{24}$$

Using that the diffusion part in (24) of the bilinear form is continuous in $(\mathcal{V}_h \times \mathcal{V}_h)$ with the constant C_A, Lemma 2, and a inverse inequality, we get

$$\begin{aligned} a_h^d(\phi_{T,h} - \lambda_{T,j}, b_{T,j}) &\lesssim C_A |||\phi_{T,h} - \lambda_{T,j}|||_{\mathrm{h}} |||b_{T,j}|||_{\mathrm{h}} \\ &\lesssim C_A^2 |||\phi_{T,h} - \lambda_{T,j}|||_{\mathrm{h}} |||\lambda_{T,j}|||_H \\ &\lesssim C_A^2 \beta^{1/2} |||\phi_{T,h} - \lambda_{T,j}|||_{\mathrm{h}} \|H^{-1}\lambda_{T,j}\|_{L^2(T)}. \end{aligned} \tag{25}$$

For the convection part in (24), we have

$$\begin{aligned} &(\mathbf{b} \cdot \nabla_h(\phi_{T,h} - \lambda_{T,j}), b_{T,j})_{L^2(\Omega)} \\ &\lesssim \|H\mathbf{b} \cdot \nabla_h(\phi_{T,h} - \lambda_{T,j})\|_{L^2(\Omega)} \|H^{-1} b_{T,j}\|_{L^2(\Omega)} \\ &\lesssim \|H\mathbf{b}\|_{L^\infty(\Omega)} \|\nabla_h(\phi_{T,h} - \lambda_{T,j})\|_{L^2(\Omega)} \|H^{-1}\lambda_{T,j}\|_{L^2(\Omega)}, \end{aligned} \tag{26}$$

and obtain

$$|||\phi_{T,h} - \lambda_{T,j}|||_{\mathrm{h}} \leq C_\phi \beta^{1/2} \|H^{-1}\lambda_{T,j}\|_{L^2(\Omega)}. \tag{27}$$

with $C_\phi = (C_A^2 + \|H\mathbf{b}\|_{L^\infty(\Omega)}\alpha^{-1})^{1/2}$. □

3.2 Ideal discontinuous Galerkin multiscale method

An ideal multiscale method seeks $u_H^{ms} \in \mathcal{V}_H^{ms}$ such that

$$a_h(u_H^{ms}, v) = F(v) \quad \text{for all } v \in \mathcal{V}_H^{ms}. \tag{28}$$

Note that, to construct in the space $\mathcal{V}_H^{ms}$ a variational problem has to be solved on the whole domain Ω for each basis function, which is not feasible for real computations. The following theorem shows the convergence of the ideal (non-realistic) multiscale method.

Theorem 4. *Let $u_h \in \mathcal{V}_h$ be the solution to* (12)*, and $u_H^{ms} \in \mathcal{V}_H^{ms}$ be the solution to* (28)*, then*

$$|||u_h - u_H^{ms}||| \lesssim C_1 \alpha^{-1/2} ||H(f - \Pi_H f)||_{L^2(\Omega)} \tag{29}$$

holds, with $C_1 = C_A + \|H\mathbf{b}\|_{L^\infty(\Omega)}\alpha^{-1}$

Proof. See Section 5. □

3.3 Discontinuous Galerkin multiscale method

The fast decay of the corrected basis functions (Lemma 6), motivates us to solve the corrector functions on localized patches. This introduces a localization error, but choosing the patch size as $\mathcal{O}(H \log(H^{-1}))$ (Theorem 7) the localization error has the same convergence rate as the ideal multiscale method in Theorem 4. The corrector functions are solved on element patches, defined as follows.

Definition 5. *For all $T \in \mathcal{T}_H$, let ω_T^L be a patch centered around element T with size L, defined as*

$$\begin{aligned} &\omega_T^0 := \text{int}(T), \\ &\omega_T^L := \text{int}(\cup\{T' \in T_H \mid T \cap \bar{\omega}_T^{L-1} \neq 0\}), \quad L = 1, 2, \ldots. \end{aligned} \tag{30}$$

See Figure 1 for an illustration.

The localized corrector functions are calculated as follows: for all $\{T \in \mathcal{T}_H, j = 1, \ldots, r\}$ find $\phi_{T,j}^L \in \mathcal{V}^f(\omega_T^L) = \{v \in \mathcal{V}^{\text{f}} \mid v|_{\Omega \setminus \omega_T^L} = 0\}$ such that

$$a_h(\phi_{T,j}^L, v) = a_h(\lambda_{T,j}, v), \quad \text{for all } v \in \mathcal{V}^f(\omega_T^L). \tag{31}$$

The decay of the corrected basis function is given in the following lemma.

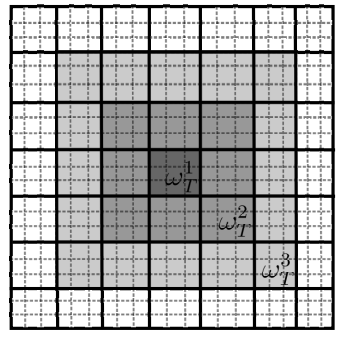


Figure 1: Example of a patch with one layer, ω_T^1, two layers ω_T^2, and three layers ω_T^3, centered around element T.

Lemma 6. *For all $T \in \mathcal{T}_H$, $j = 1, \ldots, r$ where $\phi_{T,j}$ is the solution to* (22) *and $\phi_{T,j}^L$ is the solution to* (36)*, the following estimate*

$$\left|\left|\left|\phi_{T,j} - \phi_{T,j}^L\right|\right|\right|_h \lesssim C_2 \gamma^L |||\lambda_{T,j} - \phi_{T,j}^L|||_{\mathrm{h}} \tag{32}$$

holds, where $L = \ell k$ is the size of the patch, $0 < \gamma = (\ell^{-1} C_3)^{\frac{\ell(k-1)-1}{2\ell k(\ell+1)}} < 1$, $C_2 = C_c C_\zeta (1 + C_A C_s)$, and $C_3 = C(C_A^2 + \|H\mathbf{b}\|_{L^\infty(\Omega)} \alpha^{-1})$, where C is a generic constants neither depending on the mesh size, the size of the patches, or the problem data.

Proof. See Section 5. □

The space of localized corrected basis function is defined by $\mathcal{V}_H^{ms,L} := \{\phi_{T,j}^L - \lambda_{T,j} \mid T \in \mathcal{T}_H,\, r = 1, \ldots, r\}$. The DG multiscale method reads: find $u_H^{ms,L} \in \mathcal{V}_H^{ms,L}$ such that

$$a_h(u_H^{ms,L}, v) = F(v) \quad \text{for all } v \in \mathcal{V}_H^{ms,L}. \tag{33}$$

An error bound for the DG multiscale method using a localized corrected basis is given in Theorem 7. Note that it is only the first term $|||u - u_h|||_{\mathrm{h}}$ in Theorem 7 that depends on the regularity of u.

Theorem 7. *Let $u_h \in \mathcal{V}_h$ and $u_H^{ms,L} \in \mathcal{V}_H^{ms,L}$ be the solutions to* (12) *and* (33)*, respectively. Then*

$$\begin{aligned} |||u - u_H^{ms,L}|||_{\mathrm{h}} \leq & |||u - u_h|||_{\mathrm{h}} + C_c \alpha^{-1/2} \|H(f - \Pi_H f)\|_{L^2(\Omega)} \\ & + C_5 \|H^{-1}\|_{L^\infty(\Omega)} L^{d/2} \gamma^L \|f\|_{L^2(\Omega)} \end{aligned} \tag{34}$$

holds, where L is the size of the patches, C_1 is a constant defined in Theorem 4, $0 < \gamma < 1$ and $C_5 = C_4^{1/2} C_2 C_\phi C_A$, where $C_4 = C_c^2 C_\zeta^2 (1 + C_A C_s)^2$ is defined in Lemma 13, and C_2 and γ are defined in Lemma 6.

Proof. See Section 5. □

Remark 8. *Theorem 7 is simplified to,*

$$|||u - u_H^{ms,L}|||_{\mathrm{h}} \leq |||u - u_h|||_{\mathrm{h}} + C_1 \|H\|_{L^\infty(\Omega)}. \tag{35}$$

given that the patch size is chosen as $L = \lceil C \log(H^{-1}) \rceil$ with an appropriate C and $\|f\|_{L^2} = 1$. In the numerical experiments we choose $C = 2$.

Remark 9. *If the convective term is small it is not necessary to include it in the computation of the correctors [21]. Instead the following correctors can be used: for all $\{T \in \mathcal{T}_H, j = 1, \ldots, r\}$ find $\hat{\phi}_{T,j}^L \in \mathcal{V}^f(\omega_T^L)$ such that*

$$a_h^{\mathrm{d}}(\hat{\phi}_{T,j}^L, v) = a_h^{\mathrm{d}}(\lambda_{T,j}, v), \quad \text{for all } v \in \mathcal{V}^f(\omega_T^L). \tag{36}$$

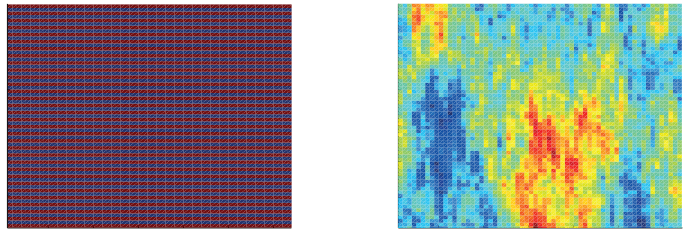
This gives the right convergence results if

$$\mathcal{O}\left(\frac{\|\mathbf{b}\|_{L^\infty(\Omega)}}{\alpha}\right) = 1 \tag{37}$$

compared to (11) *if the convective term is included.*

4 Numerical experiment

We consider the domain $\Omega = [0,1] \times [0,1]$ and the forcing function $f = 1 + \cos(2\pi x)\cos(2\pi y)$. The localization parameter which determine the size of the patches is chosen as $L = \lceil 2\log(H^{-1}) \rceil$, i.e., the size of the patches are $2H\log(H^{-1})$. Consider a coarse quadrilateral mesh, $\mathcal{T}_H$, of size $H = 2^{-i}$, $i = 2, 3, 4, 5$. The corrector functions are solved on sub-grids of the quadrilateral mesh, $\mathcal{T}_h$, where $h = 2^{-7}$. We consider three different permeabilities: $A_1 = 1$, $A_2 = A_2(y)$ which is piecewise constant with respect to a Cartesian grid of width 2^{-6} in y-direction taking the values 1 or 0.01, and $A_3 = A_3(x, y)$ which is piecewise constant with respect to a Cartesian grid of width 2^{-6} both in the x- and y-directions, bounded below by $\alpha = 0.05$ and has a maximum ratio $\beta/\alpha = 4 \cdot 10^5$. The permeability A_3 is taken from the 31 layer in the SPE 10 benchmark problem, see `http:www.spe.org/web/csp/`. The diffusion coefficients A_2 and A_3 are illustrated in Figure 2. For the convection term we consider: $\mathbf{b} = [C, 0]$, for different values of C.



(a) A_2 (b) A_3

Figure 2: The diffusion coefficients A_2 and A_3 in log scale.

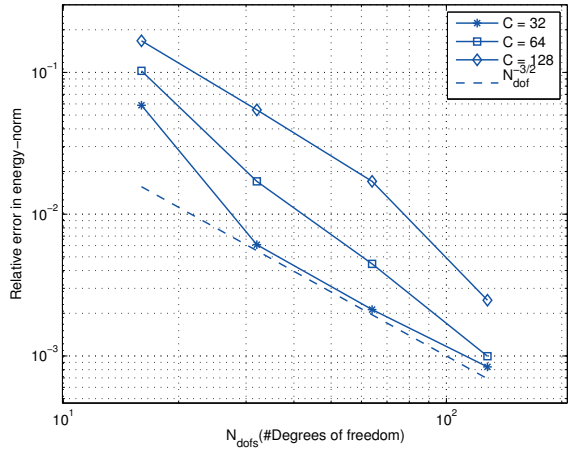


Figure 3: The number degrees of freedom (N_{dof}) vs. the relative error in energy-norm, for different sizes of the convection term, C.

To investigate how the error in relative energy-norm, $|||u_h - u_H^{ms,L}|||/|||u_h|||$, depends on the magnitude of the convection we consider: A_1 and $\mathbf{b} = [C, 0]$ with $C = \{32, 64, 128\}$. Figure 3 shows the convergence in energy-norm as a function of the coarse mesh size H for the different values of C.

Also, to see the effect of heterogeneous diffusion of the error in the relative energy-norm, $|||u_h - u_H^{ms,L}|||/|||u_h|||$, we consider: Figure 4 which shows the error in relative energy-norm using A_2 and $\mathbf{b} = [1, 0]$ and Figure 5 which

shows the error in relative energy-norm using A_3 and $\mathbf{b} = [512, 0]$.

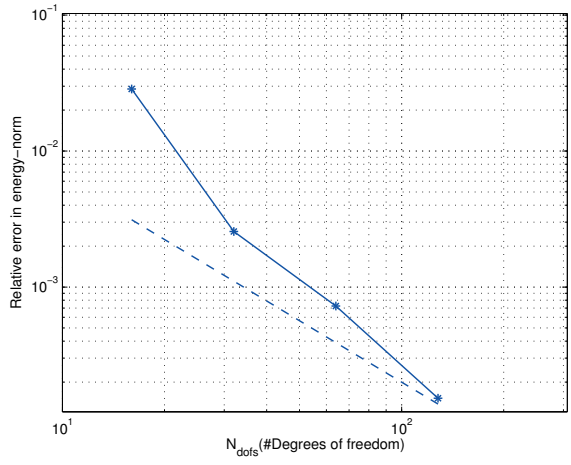


Figure 4: The number degrees of freedom (N_{dof}) vs. the relative error in energy-norm, using a high contrast diffusion coefficients A_2 and $\mathbf{b} = [1, 0]$. The dotted line corresponds to $N_{dof}^{-3/2}$.

We obtain H^3 convergence of the DG multiscale method to a reference solution in the relative energy-norm, $|||u_h - u_H^{ms,L}|||/|||u_h|||$, independent of the variation in the coefficients or regularity of the underlying solution.

5 Proofs from Section 3

In this section we state the proofs of the main results which was postponed from in section 3. To this end we start by proving some technical lemmas in Section 5.1 which we use to prove the main results in Section 5.2.

5.1 Technical lemmas

In the proofs of the main results, Theorem 4, Lemma 6, and Theorem 7, we will need some definitions and technical lemmas stated below.

Continuity of the DG bilinear form for convective problems can be proven on a orthogonal subset of $\mathcal{V}_h$. The space $\mathcal{V}^{\mathrm{f}}$ is an orthogonal subset of $\mathcal{V}_h$ but on a coarse scale.

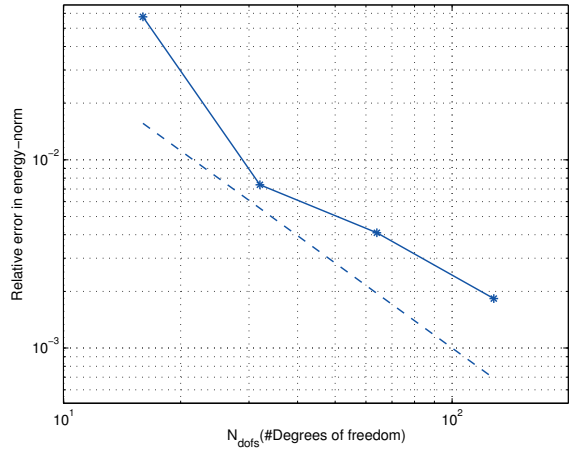


Figure 5: The number degrees of freedom (N_{dof}) vs. the relative error in energy-norm, using a high contrast diffusion coefficients A_3 and $\mathbf{b} = [512, 0]$. The dotted line corresponds to $N_{dof}^{-3/2}$.

Lemma 10 (Continuity in $(\mathcal{V}_h \times \mathcal{V}^{\mathrm{f}})$ and $(\mathcal{V}^{\mathrm{f}} \times \mathcal{V}_h)$)**.** *For all,* $(u, v) \in \mathcal{V}^{\mathrm{f}} \times \mathcal{V}_h$ *or in* $\mathcal{V}_h \times \mathcal{V}^{\mathrm{f}}$ *, it holds*

$$a(v, w) \lesssim C_c |||v|||_{\mathrm{h}} |||w|||_{\mathrm{h}} \tag{38}$$

where

$$C_c = C_A + \|H\mathbf{b}\|_{L^\infty(\Omega)} \alpha^{-1}. \tag{39}$$

Proof. Since a_h^d is continuous in $(\mathcal{V}_h \times \mathcal{V}_h)$ with the constant C_A, continuity in $(\mathcal{V}^{\mathrm{f}} \times \mathcal{V}_h)$ follows from $\mathcal{V}^{\mathrm{f}} \subset \mathcal{V}_h$. For the convective part a_h^{c}, we have

$$\begin{aligned} a^{\mathrm{c}}(v, w) &= \sum_{T \in \mathcal{T}_h} (\mathbf{b} \cdot \nabla v, w)_{L^2(T)} + \sum_{e \in \mathcal{E}_k} (b_e[v], [w])_{L^2(e)} \\ &\quad - \sum_{e \in \mathcal{E}_k(\Omega)} (\nu_e \cdot \mathbf{b}[v], \{w\})_{L^2(e)} + \sum_{e \in \mathcal{E}_k(\Gamma)} ((\nu_e \cdot \mathbf{b})^{\ominus} v, w)_{L^2(e)} \\ &\lesssim \sum_{T \in \mathcal{T}_h} \left(\|\mathbf{b}\|_{L^\infty(T)} \|\nabla v\|_{L^2(T)} \|w - \Pi_H w\|_{L^2(T)} \right) \\ &\quad + \sum_{e \in \mathcal{E}_k} \left(\|\mathbf{b}\|_{L^\infty(e)} h^{-1/2} \|[v]\|_{L^2(e)} \|w\|_{L^2(S^+ \cup S^-)} \right). \end{aligned} \tag{40}$$

where $S^+, S^- \in \mathcal{T}_h$ and $e = S^+ \cap S^-$. Using a discrete Cauchy-Schwartz inequality and summing over the coarse elements, we get

$$\begin{aligned} a^{\mathrm{c}}(v,w) &\lesssim \alpha^{-1/2}\|H\mathbf{b}\|_{L^\infty(\Omega)}|||v|||_{\mathrm{h}}\|H^{-1}(w-\Pi_H w)\|_{L^2(\Omega)}, \\ &\lesssim \|H\mathbf{b}\|_{L^\infty(\Omega)}\alpha^{-1}|||v|||_{\mathrm{h}}|||w|||_{\mathrm{h}}, \end{aligned} \tag{41}$$

which concludes the proof for $(\mathcal{V}_h \times \mathcal{V}^{\mathrm{f}})$. The proof of $(\mathcal{V}^{\mathrm{f}} \times \mathcal{V}_h)$ is obtained by first integrating $(\mathbf{b}\cdot\nabla u, v)_{L^2(T)}$ by parts. □

The following cut-off function will be frequently used in the proof of the main results.

Definition 11. *The function $\zeta^{d,D} \in P_o(\mathcal{T}_h)$, for $D > d$, is a cut off function fulfilling the following condition*

$$\begin{aligned} \zeta_T^{d,D}|_{\omega_T^d} &= 1, \\ \zeta_T^{d,D}|_{\Omega\setminus\omega_T^D} &= 0, \\ \|[\zeta_T^{d,D}]\|_{L^\infty(\mathcal{E}_h(T))} &\lesssim \frac{\|h\|_{L^\infty(T)}}{(D-d)H|_T}, \end{aligned} \tag{42}$$

and $|||[\zeta^{d,D}]|||_{L^\infty(\partial(\omega_T^D\setminus\omega_T^d))} = 0$, for all $T \in \mathcal{T}_H$.

For the cut off function has the following stability property.

Lemma 12. *For any $v \in \mathcal{V}_h$ and $\zeta_T^{d,D}$ from Definition 11, the estimate,*

$$|||\zeta_T^{d,D} v|||_{\mathrm{h}} \lesssim C_\zeta |||v|||_{h,\omega_T^D}, \tag{43}$$

holds, where $C_\zeta = (C_A^2 + \|h\mathbf{b}\|_{L^\infty(\Omega)}/\alpha)^{1/2}$.

Proof. For the diffusion part we use the following result from [15],

$$|||(1-\zeta_T^{d,D})v|||_{\mathrm{h,d}} \lesssim C_A |||v|||_{h,\Omega\setminus\omega_T^{L-1}} \tag{44}$$

and focus on the convective part. We obtain

$$\begin{aligned} &|||(1-\zeta_T^{d,D})v|||_{\mathrm{h,c}}^2 \\ &= \sum_{e\in\mathcal{E}_h} \|b_e^{1/2}[(1-\zeta_T^{d,D})v]\|_{L^2(e)}^2 \\ &\le \sum_{\substack{e\in\mathcal{E}_h: \\ e\cap\omega_T^{L-1}\neq 0}} \left(\|b_e^{1/2}[v]\|_{L^2(e)}^2 + \|h\|_{L^\infty(S^+\cup S^-)}^2 \|H^{-1}b_e^{1/2}\{v\}\|_{L^2(e)}^2\right) \\ &\lesssim \sum_{\substack{T\in\mathcal{T}_H: \\ e\cap\omega_T^{L-1}\neq 0}} \|h\mathbf{b}\|_{L^\infty(T)} \left(\|h^{-1/2}[v]\|_{L^2(e)}^2 + \|H^{-1}(v-\Pi_H v)\|_{L^2(T)}^2\right) \\ &\lesssim \frac{\|h\mathbf{b}\|_{L^\infty(\Omega)}}{\alpha} |||v|||_{h,\Omega\setminus\omega^{L-1}}^2, \end{aligned} \tag{45}$$

using $[vw] = \{v\}[w] + \{w\}[v]$, the triangle inequality, and a trace inequality. The proof is concluded using (44) and (45). □

The following lemmas will be necessary in order to prove Theorem 7.

Lemma 13. *The following estimate,*

$$|||\sum_{T\in\mathcal{T}_H,\, j=1,\dots,r} v_j(\phi_{T,j}-\phi_{T,j}^L)|||_{\mathrm{h}}^2 \lesssim C_4 L^d \sum_{T\in\mathcal{T}_H,\, j=1,\dots,r} |v_j|^2 |||\phi_{T,j}-\phi_{T,j}^L|||_{\mathrm{h}}^2, \quad (46)$$

holds, where $C_4 = C_c^2 C_\zeta^2 (1 + C_A C_s)^2$.

Proof. The proof is analogous with the proof of Lemma 12 in [15]. □

5.2 Proof of main results

We are now ready to prove, Theorem 4, Lemma 6, and Theorem 7.

Theorem 4. Let us decompose u_h into a coarse contribution, $v_H^{ms} \in \mathcal{V}_H^{ms}$, and a fine scale remainder, $v^f \in \mathcal{V}^{\mathrm{f}}$, i.e., $u_h = v_H^{ms} + v^f$. For v^f we have

$$\begin{aligned} |||v^f|||_{\mathrm{h}}^2 &\lesssim a_h(v^f, v^f) = a_h(u_h, v^f) = (f, v^f)_{L^2(\Omega)} \\ &= (f - \Pi_H f, v^f - \Pi_H v^f)_{L^2(\Omega)} \\ &\leq \|H(f - \Pi_H f)\|_{L^2(\Omega)} \|H^{-1}(v^f - \Pi_H v^f)||_{L^2(\Omega)} \\ &\leq \alpha^{-1/2} \|H(f - \Pi_H f)\|_{L^2(\Omega)} |||v^f|||_{\mathrm{h}}. \end{aligned} \quad (47)$$

Using continuity, we get

$$\begin{aligned} |||u_h - u_H^{ms}|||_{\mathrm{h}}^2 &\lesssim a_h(u_h - u_H^{ms}, u_h - u_H^{ms}) = a_h(u_h - u_H^{ms}, u_h - v_H^{ms}) \\ &\lesssim C_c |||u_h - u_H^{ms}|||_{\mathrm{h}} |||u_h - v_H^{ms}|||_{\mathrm{h}}, \end{aligned} \quad (48)$$

which concludes the proof together with (47). □

Lemma 6. Define $e := \phi_{T,j} - \phi_{T,j}^L$ where $\phi_{T,j} \in \mathcal{V}^{\mathrm{f}}$ and $\phi_{T,j}^L \in \mathcal{V}^f(\omega_T^L)$. We have

$$|||e|||_h^2 \lesssim a_h(e, \phi_{T,j} - \phi_{T,j}^L) = a_h(e, \phi_{T,j} - v) \lesssim C_c |||e|||_{\mathrm{h}} |||\phi_{T,j} - v|||_{\mathrm{h}}. \quad (49)$$

Furthermore from Lemma 2, there exist a $v = \zeta_T^{L-1,L}\phi_{T,j} - b_T \in \mathcal{V}^f(\omega_T^L)$ such that $\Pi_H b_T = \Pi_H(\zeta_T^{L-1,L}\phi_{T,j})$ and $|||b_T|||_{\mathrm{h}} \lesssim C_A |||\Pi_H(\zeta_T^{L-1,L}\phi_{T,j})|||_H$, we have

$$|||e|||_h \lesssim C_c \left(|||(1 - \zeta_T^{L-1,L})\phi_{T,j}|||_{\mathrm{h}} + |||b_T|||_{\mathrm{h}} \right), \quad (50)$$

where

$$\begin{aligned}|||b_T|||_{\mathrm{h}} &\lesssim C_A|||\Pi_H\zeta_T^{L-1,L}\phi_{T,j}|||_H = C_A|||\Pi_H(1-\zeta_T^{L-1,L})\phi_{T,j}|||_H \\ &\lesssim C_AC_s|||(1-\zeta_T^{L-1,L})\phi_{T,j}|||_{\mathrm{h}} \lesssim C_AC_sC_\zeta|||\phi_{T,j}|||_{h,\Omega\setminus\omega_T^{L-1}}.\end{aligned} \tag{51}$$

using Lemma 2, Lemma 1, and Lemma 12. We obtain,

$$|||e|||_h \lesssim C_2|||\phi_{T,j}|||_{h,\Omega\setminus\omega_T^{L-1}}, \tag{52}$$

where $C_2 = C_cC_\zeta(1+C_AC_s)$ from (50) and (51).

The next step in the proof is to construct a recursive relation which will be used to prove the decay of the correctors. To this end, let $\ell k = L-1$, and define another the cut off function, $\eta_T^m := (1-\zeta^{\ell(k-m-1)-m,\ell(k-m)-m})$ and the patch $\widetilde{\omega}_T^m := \omega_T^{\ell(k-m+1)-m}$, for $m = 0,1,\ldots,\lfloor \ell k/(\ell+1)-1\rfloor$. Note that $\widetilde{\omega}_T^{m+1} \subset \widetilde{\omega}_T^m$. We obtain

$$|||\phi_{T,j}|||_{h,\Omega\setminus\widetilde{\omega}_T^m} \leq |||\eta_T^m\phi_{T,j}|||_{\mathrm{h}} \lesssim a_h(\eta_T^m\phi_{T,j},\eta_T^m\phi_{T,j}). \tag{53}$$

To shorten the proof we refer to the following inequality

$$a^{\mathrm{d}}(\eta_T^m\phi_{T,j},\eta_T^m\phi_{T,j}) \lesssim a^{\mathrm{d}}(\phi_{T,j},(\eta_T^m)^2\phi_{T,j}-b_T) + \frac{C_A^2}{\ell}|||\phi_{T,j}|||^2_{h,\widetilde{\omega}_T^m\setminus\widetilde{\omega}_T^{m+1}}. \tag{54}$$

where $(\eta_T^m)^2\phi_{T,j}-b_T \in \mathcal{V}^{\mathrm{f}}$, in the proof of Lemma 10 in [15]. We focus on the convection term, since the cut of function is piecewise constant it follows that

$$(\mathbf{b}\cdot\nabla\eta_T^m\phi_{T,j},\eta_T^m\phi_{T,j})_{L^2(S)} = (\mathbf{b}\cdot\nabla\phi_{T,j},(\eta_T^m)^2\phi_{T,j})_{L^2(S)} \tag{55}$$

for all $S \in \mathcal{T}_h$. Using the following equalities from (Appendix A in [15])

$$\begin{aligned}\{vw\}[vw] &= \{w\}[v^2w] - [v]\{w\}\{v\}\{w\} + 1/4[v]\{v\}[w][w], \\ [vw][vw] &= [w][v^2w] - 1/4[v]^2[w]^2 + [v]^2\{w\}^2,\end{aligned} \tag{56}$$

and (55), we obtain

$$\begin{aligned}a^{\mathrm{c}}(\eta_T^m\phi_{T,j},\eta_T^m\phi_{T,j}) &= a^{\mathrm{c}}(\phi_{T,j},(\eta_T^m)^2\phi_{T,j}) \\ &+ \sum_{e\in\mathcal{E}_h(\Omega)}\Big((\nu_e\cdot\mathbf{b}[\eta_T^m]\{\phi_{T,j}\},\{\eta_T^m\}\{\phi_{T,j}\})_{L^2(e)} \\ &\quad - 1/4(\nu_e\cdot\mathbf{b}[\eta_T^m]\{\phi_{T,j}\},\{\eta_T^m\}[\phi_{T,j}])_{L^2(e)} \\ &\quad - 1/4(b_e[\eta_T^m]^2,[\phi_{T,j}]^2)_{L^2(e)} + (b_e[\eta_T^m]^2,\{\phi_{T,j}\}^2)_{L^2(e)}\Big).\end{aligned} \tag{57}$$

The sum over the edges terms can be bounded using that $\|[\eta_T^m]\|_{L^\infty(T)} \lesssim \|h\|_{L^\infty(T)}/H|_T$, $\|\{\eta_T^m\}\|_{L^\infty(\Omega)} \lesssim 1$, $\|h\|_{L^\infty(T)}/H|_T \ell < 1$, and a trace inequality. We obtain

$$\begin{aligned}
&\sum_{\substack{e\in\mathcal{E}_h(\Omega):\\ e\cap(\widetilde{\omega}_T^m\setminus\widetilde{\omega}_T^m)\neq 0}} \frac{\|H^{-1}\mathbf{b}\|_{L^\infty(e)}}{\ell}\Big(\|h^{1/2}\{\phi_{T,j}\}\|_{L^2(e)}\|h^{1/2}\{\phi_{T,j}\}\|_{L^2(e)} + \\
&\qquad \|h^{1/2}\{\phi_{T,j}\}\|_{L^2(e)}\|h^{1/2}[\phi_{T,j}]\|_{L^2(e)} + \|h^{1/2}[\phi_{T,j}]\|_{L^2(e)}^2 \\
&\qquad + \|h^{1/2}\{\phi_{T,j}\}\|_{L^2(e)}^2\Big) \\
&\lesssim \sum_{\substack{e\in\mathcal{E}_H(\Omega):\\ e\cap(\widetilde{\omega}_T^m\setminus\widetilde{\omega}_T^m)\neq 0}} \frac{\|H^{-1}\mathbf{b}\|_{L^\infty(e)}}{\ell}\|\phi_{T,j}\|_{L^2(T^+\cup T^-)}^2 \qquad (58)\\
&\lesssim \sum_{\substack{T\in\mathcal{T}_H:\\ T\cap(\widetilde{\omega}_T^m\setminus\widetilde{\omega}_T^m)\neq 0}} \frac{\|H\mathbf{b}\|_{L^\infty(T)}}{\ell}\|H^{-1}(\phi_{T,j}-\Pi_H\phi_{T,j})\|_{L^2(T)}^2 \\
&\lesssim \frac{\|H\mathbf{b}\|_{L^\infty(\Omega)}}{\ell\alpha}|||\phi_{T,j}|||_{h,(\widetilde{\omega}_T^m\setminus\widetilde{\omega}_T^{m+1})}^2.
\end{aligned}$$

Combining the results, we have

$$\begin{aligned}
|||\phi_{T,j}|||_{h,\Omega\setminus\widetilde{\omega}_T^m}^2 &\lesssim a(\phi_{T,j},(\eta_T^m)^2\phi_{T,j}-b_T)+a(\phi_{T,j},b_T)\\
&\quad+\ell^{-1}\left(C_A^2+\frac{\|H\mathbf{b}\|_{L^\infty(\Omega)}}{\alpha}\right)|||\phi_{T,j}|||_{h,(\widetilde{\omega}_T^m\setminus\widetilde{\omega}_T^{m+1})}^2, \qquad (59)
\end{aligned}$$

where b_T has support in $\widetilde{\omega}_T^m\setminus\widetilde{\omega}_T^{m+1}$, such that $(\eta_T^m)^2\phi_{T,j}-b_T\in\mathcal{V}^{\mathrm{f}}$ and $|||b_T|||_{\mathrm{h}}\lesssim C_A|||\Pi_H((\eta_T^m)^2\phi_{T,j})|||_H$, see Lemma 2. We have

$$a(\phi_{T,j},(\eta_T^m)^2\phi_{T,j}-b_T)=0. \qquad (60)$$

For all $T\in\mathcal{T}_H$ the operator Π_H is stable in the $L^2(T)$-norm, we have

$$\begin{aligned}
|||b_T|||_{h,\widetilde{\omega}_T^m\setminus\widetilde{\omega}_T^{m+1}}^2 &\lesssim C_s^2|||\Pi_H((\eta_T^m)^2\phi_{T,j})|||_H^2\\
&= C_s^2\left(|||\Pi_H((\eta_T^m)^2\phi_{T,j})|||_{\mathrm{d},H}^2+|||\Pi_H((\eta_T^m)^2\phi_{T,j})|||_{\mathrm{a},H}^2\right). \qquad (61)
\end{aligned}$$

For the first term in (61) we refer to the result

$$|||\Pi_H((\eta_T^m)^2\phi_{T,j})|||_{\mathrm{d},H}^2\lesssim\frac{C_A^2}{\ell^2}|||\phi_{T,j}|||_{h,\widetilde{\omega}_T^m\setminus\widetilde{\omega}_T^{m+1}}^2, \qquad (62)$$

from [15] and for the second them we have

$$\begin{aligned}
&|||\Pi_H((\eta_T^m)^2\phi_{T,j})|||_{\mathrm{a},H}^2 = |||\Pi_H((\eta_T^m-\Pi_0\eta_T^m)^2\phi_{T,j})|||_{\mathrm{a},H}^2\\
&= \sum_{e\in\mathcal{E}_H(\Omega)} \|b_e^{1/2}[\Pi_H((\eta_T^m)^2-\Pi_0(\eta_T^m)^2)\phi_{T,j})]\|_{L^2(e)}^2\\
&= \sum_{T\in\mathcal{T}_H} \|H^{-1}\mathbf{b}\|_{L^\infty(T)}\|(\eta_T^m)^2-\Pi_0(\eta_T^m)^2\|_{L^\infty(T)}^2\|\phi_{T,j}-\Pi_H\phi_{T,j}\|_{L^2(T)}^2\\
&\lesssim \frac{\|H\mathbf{b}\|_{L^\infty(T)}}{\alpha\ell^2}|||\phi_{T,j}|||_{h,\widetilde{\omega}_T^m\setminus\widetilde{\omega}_T^{m+1}}^2.
\end{aligned} \tag{63}$$

We obtain

$$\begin{aligned}
|||\phi_{T,j}|||_{h,\Omega\setminus\omega_T^m}^2 &\lesssim \ell^{-1}\left(C_A^2+\frac{\|H\mathbf{b}\|_{L^\infty(T)}}{\alpha}\right)|||\phi_{T,j}|||_{h,\Omega\setminus\omega_T^{m+1}}^2\\
&= C_3\ell^{-1}|||\phi_{T,j}|||_{h,\Omega\setminus\widetilde{\omega}_T^{m+1}}^2
\end{aligned} \tag{64}$$

where $C_3 = C(C_A^2+\|H\mathbf{b}\|_{L^\infty(\Omega)}\alpha^{-1})$ and C is the generic constant hidden in '$\lesssim$'. We have

$$|||\phi_{T,j}|||_{h,\Omega\setminus\widetilde{\omega}_T^m}^2 \lesssim C_3\ell^{-1}|||\phi_{T,j}|||_{h,\Omega\setminus\widetilde{\omega}_T^{m+1}}^2, \tag{65}$$

for any $m=0,1,\ldots,\lfloor \ell k/(\ell+1)\rfloor-1$, which we can use recursively as

$$\begin{aligned}
|||\phi_{T,j}|||_{h,\Omega\setminus\widetilde{\omega}_T^1}^2 &\lesssim (C_3\ell^{-1})^{k-1}|||\phi_{T,j}|||_{h,\Omega\setminus\widetilde{\omega}_T^k}^2\\
&= (C_3\ell^{-1})^{\lfloor \ell k/(\ell+1)\rfloor-1}|||\phi_{T,j}-\lambda_{T,j}|||_{h,\Omega}^2.
\end{aligned} \tag{66}$$

Note that $k/2$ is a lower bound of $\ell k/(\ell+1)$. Equation (52) together with (65), gives

$$|||\phi_{T,j}-\phi_h^L|||_{\mathrm{h}} \lesssim C_2(C_3\ell^{-1})^{\frac{1}{2}(\ell k/(\ell+1)-1)}|||\phi_{T,j}-\lambda_{T,j}|||_{\mathrm{h}}. \tag{67}$$

which concludes the proof is concluded. □

Theorem 7. Using the triangle inequality, we have

$$|||u-u_H^{ms,L}|||_{\mathrm{h}} \le |||u-u_h|||_{\mathrm{h}} + |||u_h-u_H^{ms,L}|||_{\mathrm{h}}. \tag{68}$$

Note that, $u_h\in\mathcal{V}_h$, can be decomposed into a coarse, $v_H^{ms}\in\mathcal{V}_H^{ms}$, and a fine, $v^f\in\mathcal{V}^{\mathrm{f}}$, scale contribution, i.e., $u_h = v_H^{ms}+u^f$. Also, let $v_H^{ms,L}\in\mathcal{V}_H^{ms,L}$ be chosen such that $\Pi_H v_H^{ms,L} = \Pi_H v_H^{ms}$. We have

$$\begin{aligned}
|||u_h-u_H^{ms,L}|||_{\mathrm{h}} &\lesssim a_h(u_h-u_H^{ms,L},u_h-u_H^{ms,L})\\
&= a_h(u_h-u_H^{ms,L},u_h-v_H^{ms,L})\\
&\lesssim C_{\mathrm{c}}|||u_h-u_H^{ms,L}|||_{\mathrm{h}}|||u_h-v_H^{ms,L}|||_{\mathrm{h}},
\end{aligned} \tag{69}$$

and obtain

$$\begin{aligned}|||u - u_H^{ms,L}|||_{\mathrm{h}} \leq & |||u - u_h|||_{\mathrm{h}} \\ & + C_c \left(|||u_h - v_H^{ms}|||_{\mathrm{h}} + |||v_H^{ms} - v_H^{ms,L}|||_{\mathrm{h}} \right).\end{aligned} \tag{70}$$

The first term in (70) implies that the reference mesh need to be sufficiently fine to get a sufficient approximation. The second term is approximated using (47), i.e.

$$|||u_h - v_H^{ms}|||_{\mathrm{h}} \lesssim \alpha^{-1/2} \|H(1 - \Pi_H) f\|_{L^2(\Omega)}, \tag{71}$$

and for the last term in we have,

$$\begin{aligned}|||v_H^{ms} - v_H^{ms,L}|||_{\mathrm{h}}^2 &= ||| \sum_{T \in \mathcal{T}_H,\, j=1,\ldots,r} v_{H,T}^{ms}(x_j)(\phi_{T,h} - \phi_{T,j}^L)|||_{\mathrm{h}}^2 \\ &\lesssim C_4 L^d \sum_{T \in \mathcal{T}_H,\, j=1,\ldots,r} |v_{H,T}^{ms}(x_j)|^2 |||\phi_{T,h} - \phi_{T,j}^L|||_{\mathrm{h}}^2 \\ &\lesssim C_4 C_2^2 L^d \gamma^{2L} \sum_{T \in \mathcal{T}_H,\, j=1,\ldots,r} |v_{H,T}^{ms}(x_j)|^2 |||\phi_{T,h} - \lambda_{T,j}|||_{\mathrm{h}}^2,\end{aligned} \tag{72}$$

using Lemma 13 and Lemma 6.

We obtain, using Lemma 3, that

$$\begin{aligned}&\sum_{T \in \mathcal{T}_H,\, j=1,\ldots,r} |v_{H,T}^{ms}(x_j)|^2 |||\phi_{T,h} - \lambda_{T,j}|||_{\mathrm{h}}^2 \\ &\leq C_\phi^2 \sum_{T \in \mathcal{T}_H,\, j=1,\ldots,r} \|H^{-1} v_{H,T}^{ms}(x_j) \lambda_{T,j}\|_{L^2(\Omega)}^2 \\ &\lesssim C_\phi^2 \beta \| \sum_{T \in \mathcal{T}_H,\, j=1,\ldots,r} H^{-1} v_{H,T}^{ms}(x_j) \lambda_{T,j} \|_{L^2(\Omega)}^2 \\ &= C_\phi^2 \beta \| \sum_{T \in \mathcal{T}_H,\, j=1,\ldots,r} H^{-1} v_{H,T}^{ms}(x_j) \Pi_H(\lambda_{T,j} - \phi_{T,j}) \|_{L^2(\Omega)}^2 \\ &\leq C_\phi^2 \beta \|H^{-1}\|_{L^\infty(\Omega)} \|\Pi_H(v_H^{ms} + u^f)\|_{L^2(\Omega)}^2 \\ &\leq C_\phi^2 \beta \|H^{-1} u_h\|_{L^2(\Omega)}^2 \\ &\leq C_\phi^2 C_A^2 \|H^{-1}\|_{L^\infty(\Omega)} |||u_h|||_{\mathrm{h}}.\end{aligned} \tag{73}$$

holds and we conclude the proof.

□

References

[1] A. Abdulle and M. E. Huber. Discontinuous Galerkin finite element heterogeneous multiscale method for advection-diffusion problems with multiple scales. *Numer. Math.*, 126(4):589–633, 2014.

[2] I. Babuska 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.

[3] G. A. Baker. Finite element methods for elliptic equations using nonconforming elements. *Math. Comp.*, 31:45–59, 1977.

[4] L. Berlyand and H. Owhadi. Flux norm approach to finite dimensional homogenization approximations with non-separated scales and high contrast. *Arch. Ration. Mech. Anal.*, 198:677–721, 2010.

[5] F. Brezzi, L. D. Marini, and E. Süli. Discontinuous Galerkin methods for first-order hyperbolic problems. *Mathematical Models and Methods in Applied Sciences*, 14:1893–1903, 2004.

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

[7] J. Chu, Y. Efendiev, V. Ginting, and T. Y. Hou. Flow based oversampling technique for multiscale finite element methods. *Advances in Water Resources*, 31:599–608, 2008.

[8] B. Cockburn, G. E. Karniadakis, and C-W. Shu, editors. *Discontinuous Galerkin methods*, volume 11 of *Lecture Notes in Computational Science and Engineering*. Springer-Verlag, Berlin, 2000.

[9] D. A. Di Pietro and A. Ern. *Mathematical aspects of discontinuous Galerkin methods*, volume 69 of *Mathématiques & Applications (Berlin) [Mathematics & Applications]*. Springer, Heidelberg, 2012.

[10] W. E and B. Engquist. The heterogeneous multiscale methods. *Commun. Math. Sci.*, 1:87–132, 2003.

[11] W. E and B. Engquist. Multiscale modeling and computation. *Notices Amer. Math. Soc.*, 1:1062–1070, 2003.

[12] 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.

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

[14] 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.

[15] 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.

[16] D. Elfverson, V. Ginting, and P. Henning. On multiscale methods in petrov–galerkin formulation. *Numerische Mathematik*, pages 1–40, 2015.

[17] P. Henning, A. Målqvist, and D. Peterseim. Two-level discretization techniques for ground state computations of bose-einstein condensates. *SIAM J. on Numer. Anal.*, 52(4):1525–1550, 2014.

[18] 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(4):711–744, 2010.

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

[20] Patrick Henning and Axel Målqvist. Localized orthogonal decomposition techniques for boundary value problems. *SIAM J. Sci. Comput.*, 36(4):A1609–A1634, 2014.

[21] 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.

[22] J. S. Hesthaven and T. Warburton. *Nodal discontinuous Galerkin methods*, volume 54 of *Texts in Applied Mathematics*. Springer, New York, 2008.

[23] T. Y. Hou and X.-H. Wu. A multiscale finite element method for elliptic problems in composite materials and porous media. *J. Comput. Phys.*, 134:169–189, 1997.

[24] M. Hughes, T. Mallet and A. Mizukami. A new finite element formulation for computational fluid dynamics. II. Beyond SUPG. *Comput. Methods Appl. Mech. Engrg.*, 54:341–355, 1986.

[25] T. Hughes. Multiscale phenomena: Green's functions, the Dirichlet-to-Neumann formulation, subgrid scale models, bubbles and the origins of stabilized methods. *Computer Methods in Applied Mechanics and Engineering*, 127:387–401, 1995.

[26] T. Hughes, G. Feijóo, L. Mazzei, and J.-B. Quincy. The variational multiscale method—a paradigm for computational mechanics. *Comput. Methods Appl. Mech. Engrg.*, 166:3–24, 1998.

[27] T. Hughes, L. P. Franca, and G. Hulbert. A new finite element formulation for computational fluid dynamics. VIII. The Galerkin/least-squares method for advective-diffusive equations. *Comput. Methods Appl. Mech. Engrg.*, 73:173–189, 1989.

[28] C. Johnson and U. Nävert. An analysis of some finite element methods for advection-diffusion problems. In *Analytical and numerical approaches to asymptotic problems in analysis (Proc. Conf., Univ. Nijmegen, Nijmegen, 1980)*, volume 47 of *North-Holland Math. Stud.*, pages 99–116. North-Holland, Amsterdam, 1981.

[29] C. Johnson and J. Pitkäranta. An analysis of the discontinuous Galerkin method for a scalar hyperbolic equation. *Math. Comp.*, 46:1–26, 1986.

[30] O. Karakashian and F. Pascal. A posteriori error estimates for a discontinuous Galerkin approximation of second-order elliptic problems. *SIAM J. Numer. Anal.*, 41:2374–2399, June 2003.

[31] 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.*, 196:2313–2324, 2007.

[32] M. G. Larson and A. Målqvist. An adaptive variational multiscale method for convection-diffusion problems. *Comm. Numer. Methods Engrg.*, 25:65–79, 2009.

[33] P. Lesaint and P.-A. Raviart. On a finite element method for solving the neutron transport equation. In *Mathematical aspects of finite elements in partial differential equations (Proc. Sympos., Math. Res. Center, Univ. Wisconsin, Madison, Wis., 1974)*, pages 89–123. Publication No. 33. Math. Res. Center, Univ. of Wisconsin-Madison, Academic Press, New York, 1974.

[34] A. Målqvist. Multiscale methods for elliptic problems. *Multiscale Modeling & Simulation*, 9:1064–1086, 2011.

[35] A. Målqvist and D. Peterseim. Computation of eigenvalues by numerical upscaling. *Numer. Math.*, 2014.

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

[37] H. Owhadi and L. Zhang. Localized bases for finite-dimensional homogenization approximations with nonseparated scales and high contrast. *Multiscale Modeling & Simulation*, 9(4):1373–1398, 2011.

[38] H. Owhadi, L. Zhang, and L. Berlyand. Polyharmonic homogenization, rough polyharmonic splines and sparse super-localization. *ESAIM Math. Model. Numer. Anal.*, 48(2):517–552, 2014.

[39] W. H. Reed and T. R. Hill. Triangular mesh methods for the neutron transport equation. Technical report, Los Alamos Scientific Laboratory, 1973.

[40] B. Rivière. *Discontinuous Galerkin methods for solving elliptic and parabolic equations*, volume 35 of *Frontiers in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2008. Theory and implementation.

[41] R. Söderlund. *Finite element methods for multiscale/multiphysics problems*. PhD thesis, Umeå University, Department of Mathematics and Mathematical Statistics, 2011.

Paper IV

Numer. Math.
DOI 10.1007/s00211-015-0703-z

Numerische Mathematik

On multiscale methods in Petrov–Galerkin formulation

Daniel Elfverson · Victor Ginting · Patrick Henning

Received: 22 May 2014 / Revised: 17 November 2014

Abstract In this work we investigate the advantages of multiscale methods in Petrov–Galerkin (PG) formulation in a general framework. The framework is based on a localized orthogonal decomposition of a high dimensional solution space into a low dimensional multiscale space with good approximation properties and a high dimensional remainder space, which only contains negligible fine scale information. The multiscale space can then be used to obtain accurate Galerkin approximations. As a model problem we consider the Poisson equation. We prove that a Petrov–Galerkin formulation does not suffer from a significant loss of accuracy, and still preserve the convergence order of the original multiscale method. We also prove inf-sup stability of a PG continuous and a discontinuous Galerkin finite element multiscale method. Furthermore, we demonstrate that the Petrov–Galerkin method can decrease the com-

D. Elfverson and P. Henning were partially supported by the Göran Gustafsson Foundation and the Swedish Research Council.

D. Elfverson
Department of Information Technology, Uppsala University,
Box 337, 751 05 Uppsala, Sweden
e-mail: daniel.elfverson@it.uu.se

V. Ginting
Department of Mathematics, University of Wyoming, Laramie, WY 82071, USA
e-mail: vginting@uwyo.edu

P. Henning
Section de Mathématiques, École polytechnique fédérale de Lausanne,
1015 Lausanne, Switzerland

Present Address:
P. Henning (✉)
University of Münster, 48149 Münster, Germany
e-mail: patrick.henning@uni-muenster.de

Published online: 11 January 2015

putational complexity significantly, allowing for more efficient solution algorithms. As another application of the framework, we show how the Petrov–Galerkin framework can be used to construct a locally mass conservative solver for two-phase flow simulation that employs the Buckley–Leverett equation. To achieve this, we couple a PG discontinuous Galerkin finite element method with an upwind scheme for a hyperbolic conservation law.

Mathematics Subject Classification 35J15 · 65N12 · 65N30 · 76S05

1 Introduction

In this contribution we consider linear elliptic problems with a heterogenous and highly variable diffusion coefficient A as arisen often in hydrology or in material sciences. In the following, we are looking for u which solves

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

in a weak sense. Here, we denote

(A1) $\Omega \subset \mathbb{R}^d$, $d = 1, 2, 3$, a bounded Lipschitz domain with a piecewise polygonal boundary,

(A2) $f \in L^2(\Omega)$ a source term, and,

(A3) $A \in L^\infty(\Omega, \mathbb{R}^{d\times d}_{sym})$ a symmetric matrix-valued function with uniform spectral bounds $\beta_0 \geq \alpha_0 > 0$, $\sigma(A(x)) \subset [\alpha_0, \beta_0]$ for almost all $x \in \Omega$. We call the ratio β_0/α_0 the *contrast* of A.

Under assumptions (A1)–(A3) and by the Lax–Milgram theorem, there exists a unique weak solution $u \in H^1_0(\Omega)$ to

$$a(u, v) = (f, v) \quad \text{for all } v \in H^1_0(\Omega), \tag{1}$$

where

$$a(v, w) := \int_\Omega A\nabla v \cdot \nabla w \quad \text{and} \quad (v, w) := (v, w)_{L^2(\Omega)}.$$

The problematic term in the equation is the diffusion matrix A, which is known to exhibit very fast variations on a very fine scale (i.e. it has a multiscale character). These variations can be highly heterogenous and unstructured, which is why it is often necessary to resolve them globally by an underlying computational grid that matches the said heterogeneity. Using standard finite element methods, this results in high dimensional solution spaces and hence an enormous computational demand, which often cannot be handled even by today's computing technology. Consequently,

there is a need for alternative methods, so called multiscale methods, which can either operate below linear computational complexity by using local representative elements (cf. [1,2,17,18,23,36,40]) or which can split the original problem into very localized subproblems that cover Ω but that can be solved cheaply and independent from each other (cf. [5,8,11,12,16,25,27,28,31,33,37,38]).

In this paper, we focus on a rather recent approach called localized orthogonal decomposition (LOD) that was introduced by Målqvist and Peterseim [35] and further generalized in [19,24].

We consider a coarse space V_H, which is low-dimensional but possibly inadequate for finding a reliable Galerkin approximation to the multiscale solution of problem (1). The idea of the method is to start from this coarse space and to update the corresponding set of basis functions step-by-step to improve the approximation properties of the space. In a summarized form, this can be described in four steps: (1) define a (quasi) interpolation operator I_H from $H_0^1(\Omega)$ onto V_H, (2) information in the kernel of the interpolation operator is considered to be negligible (having a small L^2-norm), (3) hence define the space of negligible information by the kernel of this interpolation, i.e. $W := \mathrm{kern}(I_H)$, and (4) find the orthogonal complement of W with respect to a scalar product $a_h(\cdot,\cdot)$, where $a_h(\cdot,\cdot)$ describes a discretization of the problem to solve. In many cases, it can be shown, that this (low dimensional) orthogonal complement space has very accurate approximation properties with respect to the exact solution. Typically, the computation of the orthogonal decomposition is localized to small patches in order to reduce the computational complexity.

So far, the concept of the LOD has been successfully applied to nonlinear elliptic problems [20], eigenvalue problems [34] and the nonlinear Schrödinger equation [21]. Furthermore, it was combined with a discontinuous Galerkin method [13,14] and extended to the setting of partition of unity methods [22].

In this work, we are concerned with analyzing the LOD framework in Petrov–Galerkin formulation, i.e. for the case that the discrete trial and test spaces are not identical. We show that an LOD method in Petrov–Galerkin formulations still preserves the convergence rates of the original formulation of the method. At the same time, the new method can exhibit significant advantages, such as decreased computational complexity and mass conservation properties. In this paper, we discuss these advantages in detail; we give examples for realizations and present numerical experiments. In particular, we apply the proposed framework to design a locally conservative multiscale solver for the simulation of two-phase flow models as governed by the Buckley–Leverett equation. We remark that employing Petrov–Galerkin variational frameworks in the construction and analysis of multiscale methods for solving elliptic problems in heterogeneous media has been investigated in the past, see for example [16,26].

The rest of the paper is organized as follows. Section 2 lays out the setting and notation for the formulation of the multiscale methods that includes the description of two-grid discretization and the LOD. In Sect. 3, we present the multiscale methods based on the LOD framework, starting from the usual Galerkin variational equation and concentrating further on the Petrov–Galerkin variational equation that is the main contribution of the paper. We establish in this section that the Petrov–Galerkin LOD (PG-LOD) exhibits the same convergence behavior as the usual Galerkin LOD (G-

LOD). Furthermore, we draw a contrast in the aspect of practical implementation that makes up a strong advantage of PG-LOD in relative comparison to G-LOD. The other advantage of the PG-LOD which cannot be achieved with G-LOD is the ability to produce a locally conservative flux field at the elemental level when discontinuous finite element is utilized. We also discuss in this section an application of the PG-LOD for solving the pressure equation in the simulation of two-phase flow models to demonstrate this particular advantage. Section 4 gives two sets of numerical experiment: one that confirms the theoretical finding and the other demonstrating the application of PG-LOD in the two-phase flow simulation. We present the proofs of the theoretical findings in Sect. 5.

2 Discretization

In this section we introduce notations that are required for the formulation of the multiscale methods.

2.1 Abstract two-grid discretization

We define two different meshes on Ω. The first mesh is a 'coarse mesh' and is denoted by $\mathcal{T}_H$, where $H > 0$ denote the maximum diameter of all elements of $\mathcal{T}_H$. The second mesh is a 'fine mesh' denoted by $\mathcal{T}_h$ with h representing the maximum diameter of all elements of $\mathcal{T}_h$. By 'fine' we mean that any variation of the coefficient A is resolved within this grid, leading to a high dimensional discrete space that is associated with this mesh. The mesh $\mathcal{T}_h$ is assumed to be a (possibly non-uniform) refinement of $\mathcal{T}_H$. Furthermore, both grids are shape-regular and conforming partitions of Ω and we assume that $h < H/2$. For the subsequent methods to make sense, we also assume that each element of $\mathcal{T}_H$ is at least twice uniformly refined to create $\mathcal{T}_h$. The set of all Lagrange points (vertices) of $\mathcal{T}_\star$ is denoted by $\mathcal{N}_\star$, and the set of interior Lagrange points is denoted by $\mathcal{N}_\star^0$, where $\star$ is either H or h.

Now we consider an abstract discretization of the exact problem (1). For this purpose, we let V_h denote a high dimensional discrete space in which we seek an approximation u_h of u. A simple example would be the classical $P1$ Lagrange finite element space associated with $\mathcal{T}_h$. However, note that we do not assume that V_h is a subspace of $H_0^1(\Omega)$. In fact, later we give an example for which V_h consists of non-continuous piecewise linear functions. Next, we assume that we are interested in solving a fine scale problem, that can be characterized by a scalar product $a_h(\cdot,\cdot)$ on V_h. Accordingly, a method on the coarse scale can be described by some $a_H(\cdot,\cdot)$, which we specify by assuming

(A4) $a_\star(\cdot,\cdot)$ is a scalar product on $V_\star$ where $\star$ is either h or H.

This allows us to define the abstract reference problem stated below.

Definition 1 (*Fine scale reference problem*) We call $u_h \in V_h$ the fine scale reference solution if it solves

$$a_h(u_h, v_h) = (f, v_h)_{L^2(\Omega)} \quad \text{for all } v_h \in V_h, \tag{2}$$

where $a_h(\cdot,\cdot)$ 'describes the method'. It is implicitly assumed that problem (2) is of tremendous computational complexity and cannot be solved by available computing resources i n a convenient time.

A simple example of $a_h(\cdot,\cdot)$ is $a_h(v_h, w_h) = a_H(v_h, w_h) = a(v_h, w_h)$. A more complex example is the $a_h(\cdot,\cdot)$ that stems from a discontinuous Galerkin approximation, in which case $a_h(\cdot,\cdot)$ is different from $a_H(\cdot,\cdot)$. The goal is to approximate problem (2) by a new problem that reaches a comparable accuracy but one that can be solved with a significantly lower computational demand.

2.2 Localized orthogonal decomposition

In this subsection, we introduce the notation that is required in the formulation of the multiscale method. In particular, we introduce an orthogonal decomposition of the high dimensional solution space V_h into the orthogonal direct sum of a low dimensional space with good approximation properties and a high dimensional remainder space. For this purpose, we make the following abstract assumptions.

(A5) $|||\cdot|||_h$ denotes a norm on V_h that is equivalent to the norm that is induced by $a_h(\cdot,\cdot)$, hence there exist generic constants $0 < \alpha \leq \beta$ such that

$$\alpha |||v_h|||_h^2 \leq a_h(v_h, v_h) \quad \text{and} \quad a_h(v_h, w_h) \leq \beta |||v_h|||_h |||w_h|||_h$$

for all $v_h, w_h \in V_h$. In the same way, $|||\cdot|||_H$ denotes a norm on V_H (equivalent to the norm induced by $a_H(\cdot,\cdot)$). Furthermore, we let $C_{H,h}$ denote the constant with $|||v|||_H \leq C_{H,h}|||v|||_h$ for all $v \in V_h$. Note that $C_{H,h}$ might degenerate for $h \to 0$.

(A6) The coarse space $V_H \subset V_h$ is a low dimensional subspace of V_h that is associated with $\mathcal{T}_H$.

(A7) Let $I_H \colon V_h \to V_H$ be an L^2-stable quasi-interpolation (or projection) operator with the properties

- there exists a generic constant C_{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 $v_H \in V_H$ it holds $\|v_h - I_H(v_h)\|_{L^2(\Omega)} \leq C_{I_H} H |||v_h|||_h$; $|||I_H(v_h)|||_H \leq C_{I_H}|||v_h|||_h$; $\|v_H - I_H(v_H)\|_{L^2(\Omega)} \leq C_{I_H} H |||v_H|||_H$ and $\|I_H(v_H)\|_{L^2(\Omega)} \leq C_{I_H}|||v_H|||_H$,
- the restriction $(I_H)_{|V_H}$ is an isomorphism with $|||\cdot|||_H$-stable inverse, i.e. we have $v_H = (I_H \circ (I_H|_{V_H})^{-1})(v_H)$ for $v_H \in V_H$ and there exists a generic $C_{I_H^{-1}}$ such that for all $v_H \in V_H$ it holds $|||(I_H|_{V_H})^{-1}(v_H)|||_H \leq C_{I_H^{-1}}|||v_H|||_H$.

Typically, L^2-projections onto V_H can be verified to fulfill assumption (A7). Similarly, I_H can be a quasi-interpolation of the Clément-type that is related to the L^2-projection. An example for this case is given in Eq. (13) below. Alternatively, I_H can be also constructed from local L^2-projections as it is done for the classical Clément interpolation. Nodal interpolations typically do not satisfy (A7).

Using the assumption that $(I_H)_{|V_H}: V_H \to V_H$ is an isomorphism (i.e. assumption (A7)), a splitting of the space V_h is given by the direct sum

$$V_h = V_H \oplus W_h, \quad \text{with } W_h := \{v_h \in V_h | I_H(v_h) = 0\}. \tag{3}$$

Observe that the 'remainder space' W_h contains all fine scale features of V_h that cannot be expressed in the coarse space V_H.

Next, consider the $a_h(\cdot,\cdot)$-orthogonal projection $P_h: V_h \to W_h$ that fulfills:

$$a_h(P_h(v_h), w_h) = a_h(v_h, w_h) \quad \text{for all } w_h \in W_h. \tag{4}$$

Since $V_h = V_H \oplus W_h$, we have that $V_\Omega^{\text{ms}} := \text{kern}(P_h) = (1 - P_h)(V_H)$ induces the $a_h(\cdot,\cdot)$-orthogonal splitting

$$V_h = V_\Omega^{\text{ms}} \oplus W_h.$$

Note that V_Ω^{ms} is a low dimensional space in the sense that it has the same dimension as V_H. As shown for several applications (cf. [20,21,34]) the space V_Ω^{ms} has very rich approximation properties in the $|||\cdot|||_h$-norm. However, it is very expensive to assemble V_Ω^{ms}, which is why it is practically necessary to localize the space W_h (respectively localize the projection). This is done using admissible patches of the following type.

Definition 2 (*Admissible patch*) For any coarse element $T \in \mathcal{T}_H$, we say that the open and connected set $U(T)$ is an *admissible patch* of T, if $T \subset U(T) \subset \Omega$ and if it consists of elements from the fine grid, i.e.

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

It is now relevant to define the restriction of W_h to an admissible patch $U(T) \subset \Omega$ by

$$\mathring{W}_h(U(T)) := \{v_h \in W_h | v_h = 0 \quad \text{in } \Omega \backslash U(T)\}.$$

A general localization strategy for the space V_Ω^{ms} can be described as follows (see [19] for a special case of this localization and [35] for a different localization strategy).

Definition 3 (*Localization of the solution space*) Let the bilinear form $a_h^T(\cdot,\cdot)$ be a localization of $a_h(\cdot,\cdot)$ on $T \in \mathcal{T}_H$ in the sense that

$$a_h(v_h, w_h) = \sum_{T \in \mathcal{T}_H} a_h^T(v_h, w_h), \tag{5}$$

where $a_h^T(\cdot,\cdot)$ acts only on T or a small environment of T. Let furthermore $U(T)$ be an admissible patch associated with $T \in \mathcal{T}_H$. Let $Q_h^T\colon V_h \to \mathring{W}_h(U(T))$ be a local correction operator that is defined as finding $Q_h^T(\phi_h) \in \mathring{W}_h(U(T))$ satisfying

$$a_h(Q_h^T(\phi_h), w_h) = -a_h^T(\phi_h, w_h) \quad \text{for all } w_h \in \mathring{W}_h(U(T)), \tag{6}$$

where $\phi_h \in V_h$. The global corrector is given by

$$Q_h(\phi_h) := \sum_{T\in\mathcal{T}_H} Q_h^T(\phi_h). \tag{7}$$

A (localized) generalized finite element space is defined as

$$V^{\mathrm{ms}} := \{\Phi_H + Q_h(\Phi_H) |\ \Phi_H \in V_H\}.$$

The variational formulation (6) is called the corrector problem associated with $T \in \mathcal{T}_H$. Solvability of each of these problems is guaranteed by the Lax–Milgram theorem. By its nature, the system matrix corresponding to (6) is localized to the patch $U(T)$ since the support of w_h is in $U(T)$. Furthermore, each of (6) pertaining to $T \in \mathcal{T}_H$ is designed to be elementally independent and thus attributing to its immediate parallelizability. The corrector problems are solved in a preprocessing step and can be reused for different source terms and for different realization of the LOD methods. Since V^{ms} is a low dimensional space with locally supported basis functions, solving a problem in V^{ms} is rather inexpensive. Normally, the solutions $Q_h^T(\phi_h)$ of (6) decays exponentially to zero outside of T. This is the reason why we can hope for good approximations even for small patches $U(T)$. Later, we quantify this decay by an abstract assumption (which is known to hold true for many relevant applications).

Remark 1 If $U(T) = \Omega$ for all $T \in \mathcal{T}_H$, then $Q_h = -P_h$, where P_h is the orthogonal projection given by (4). In this sense, V^{ms} is localization of the space V_Ω^{ms}. This can be verified using (5), which yields for all $w_h \in W_h$

$$a_h(\phi_h + Q_h(\phi_h), w_h) = \sum_{T\in\mathcal{T}_H} \Big(a_h^T(\phi_h, w_h) + a_h(Q_h^T(\phi_h), w_h)\Big) = 0.$$

By uniqueness of the projection, we conclude $Q_h = -P_h$.

The above setting is used to construct the multiscale methods utilizing the LOD method as e.g. done in [19,35] for the standard finite element formulation and a corresponding Petrov–Galerkin formulation.

3 Methods and properties

In this section, we state the LOD in Galerkin and in Petrov–Galerkin formulation along with their respective a priori error estimates and the inf-sup stability. In the last

part of this section, we give two explicit examples and discuss the advantages of the Petrov–Galerkin formulation. Subsequently we use the notation $a \lesssim b$ to abbreviate $a \leq Cb$, where C is a constant that is independent of the mesh sizes H and h; and which is independent of the possibly rapid oscillations in A.

In order to state proper a priori error estimates, we describe the notion of 'patch size' and how the size of $U(T)$ affects the final approximation. All the stated theorems on the error estimates of the LOD methods are proved in Sect. 5.

Definition 4 (*Patch size*) Let $k \in \mathbb{N}_{>0}$ be fixed. We define patches $U(T)$ that consist of the element T and k-layers of coarse element around it. For all $T \in \mathcal{T}_H$, we define element patches in the coarse mesh $\mathcal{T}_H$ by

$$\begin{aligned} U_0(T) &:= T, \\ U_k(T) &:= \cup \left\{ T' \in \mathcal{T}_H | T' \cap U_{k-1}(T) \neq \emptyset \right\} \quad k = 1, 2, \ldots . \end{aligned} \tag{8}$$

The above concept of patch sizes and patch shapes can be also generalized. See for instance [22] for a LOD that is purely based on partitions of unity. Using Definition 4, we make an abstract assumption on the decay of the local correctors $Q_h^T(\Phi_H)$ for $\Phi_H \in V_H$:

(A8) Let $Q_h^{\Omega,T}(\Phi_H)$ be the *optimal* local corrector using $U(T) = \Omega$ that is defined according to (6) and let $Q_h^{\Omega}(\Phi_H) := \sum_{T \in \mathcal{T}_H} Q_h^{\Omega,T}(\Phi_H)$. Let $k \in \mathbb{N}_{>0}$ and for all $T \in \mathcal{T}_H$ let $U(T) = U_k(T)$ as in Definition 4. Then there exists $p \in \{0, 1\}$ and a generic constant $0 < \theta < 1$ that can depend on the contrast, but not on H, h or the variations of A such that for all $\Phi_H \in V_H$,

$$\left|\left|\left|(Q_h - Q_h^{\Omega})(\Phi_H)\right|\right|\right|_h^2 \lesssim k^d \theta^{2k} (1/H)^{2p} \left|\left|\left|\Phi_H + Q_h^{\Omega}(\Phi_H)\right|\right|\right|_h^2, \tag{9}$$

where $Q_h(\Phi_H)$ is given by (7) for $U(T) = U_k(T)$.

Assumption (A8) quantifies the decay of local correctors, by stating that the solutions of the local corrector problems decay exponentially to zero outside of T. This is central for all a priori error estimates. For continuous Galerkin methods, we can obtain the optimal order $p = 0$ for the exponent in (9). This means, that the $(1/H)$-term fully vanishes. However, depending on the localization strategy [i.e. how $Q_h(\Phi_H)$ is computed] it is also possible that p takes the value 1 and that hence a pollution term of order $(1/H)$ arises (see [19, Remark 3.8] for a discussion). For discontinuous Galerkin methods, the optimal known order is $p = 1$. However, even for this case it is known that the $(1/H)$-term is rapidly overtaken by the decay, leading purely to slightly larger patch sizes (see e.g. [35]).

3.1 Galerkin LOD

This method was originally proposed in [35]: find $u_H^{\text{G-LOD}} \in V^{\text{ms}}$ that satisfies

$$a_h\left(u_H^{\text{G-LOD}}, \Phi^{\text{ms}}\right) = (f, \Phi^{\text{ms}}) \quad \text{for all } \Phi^{\text{ms}} \in V^{\text{ms}}. \tag{10}$$

Theorem 1 (A priori error estimate for Galerkin LOD) *Assume (A1)–(A8). Given a positive $k \in \mathbb{N}_{>0}$, let for all $T \in \mathcal{T}_H$ the patch $U(T) = U_k(T)$ be defined as in* (8) *and let $u_H^{\text{G-LOD}} \in V^{\text{ms}}$ be as governed by* (10). *Let $u_h \in V_h$ be the fine scale reference solution governed by* (2). *Then, the following a priori error estimate holds true*

$$\begin{aligned} &\left\| u_h - ((I_H|_{V_H})^{-1} \circ I_H)(u_H^{G\text{-}LOD}) \right\|_{L^2(\Omega)} + \left|\left|\left| u_h - u_H^{G\text{-}LOD} \right|\right|\right|_h \\ &\quad \lesssim (H + (1/H)^p k^{d/2} \theta^k) \| f \|_{L^2(\Omega)}, \end{aligned} \tag{11}$$

where $0 < \theta < 1$ and $p \in \{0, 1\}$ are the generic constants in (A8).

The term $((I_H|_{V_H})^{-1} \circ I_H)(u_H^{\text{G-LOD}})$ describes the coarse part (resulting from V_H) of $u_H^{\text{G-LOD}}$ and thus is numerically homogenized (the oscillations are averaged out). In this sense, we can say that $u_H^{\text{G-LOD}}$ is an H^1-approximation of u_h and $((I_H|_{V_H})^{-1} \circ I_H)(u_H^{\text{G-LOD}})$ an L^2-approximation of u_h, respectively. Furthermore, because $k^{\frac{d}{2}}\theta^k$ converges with exponential order to zero, the error $|||u_h - u_H^{\text{G-LOD}}|||_h$ is typically dominated by the first term of order $\mathrm{O}(H)$. This was observed in various numerical experiments in different works, cf. [19,20,35]. In particular, a specific choice $k \gtrsim (p+1)|\log(H)|$ leads to a $\mathrm{O}(H)$ convergence for the total H^1-error, see also [19,20,35].

3.2 Petrov–Galerkin LOD

In a straightforward manner, we can now state the LOD in Petrov–Galerkin formulation: find $u_H^{\text{PG-LOD}} \in V^{\text{ms}}$ that satisfies

$$a_h\left(u_H^{\text{PG-LOD}}, \Phi_H\right) = (f, \Phi_H) \quad \text{for all } \Phi_H \in V_H. \tag{12}$$

A unique solution of (12) is guaranteed by the inf-sup stability. In practice, inf-sup stability is clearly observable in numerical experiments (see Sect. 4). Analytically we can make the following observations.

Remark 2 (*Quasi-orthogonality and inf-sup stability*) The inf-sup stability of the LOD in Petrov–Galerkin formulation is a natural property to expect, since we have quasi-orthogonality in $a_h(\cdot,\cdot)$ of the spaces V^{ms} and W_h. This can be verified by a simple computation. Let $\Phi^{\text{ms}} = \Phi_H + Q_h(\Phi_H) \in V^{\text{ms}}$, let $w_h \in W_h$ and let $Q_h^{\Omega}(\Phi_H)$ the optimal corrector as in assumption (A8), then

$$\begin{aligned} a_h(\Phi^{\text{ms}}, w_h) &= a_h(\Phi_H + Q_h(\Phi_H), w_h) \\ &= a_h\left(Q_h(\Phi_H) - Q_h^{\Omega}(\Phi_H), w_h\right) \\ &\leq \left|\left|\left| Q_h(\Phi_H) - Q_h^{\Omega}(\Phi_H) \right|\right|\right|_h |||w_h|||_h \\ &\lesssim k^{d/2}\theta^k (1/H)^p \left|\left|\left| \Phi_H + Q_h^{\Omega}(\Phi_H) \right|\right|\right|_h |||w_h|||_h, \end{aligned}$$

with generic constants $0 < \theta < 1$ and $p \in \{0, 1\}$ as in (A8). This means that $a_h(\Phi^{\text{ms}}, w_h)$ converges exponentially in k to zero, and it is identical to zero for all

sufficiently large k [because then $Q_h(\Phi_H) = Q_h^{\Omega}(\Phi_H)$]. Writing the PG-LOD bilinear form as

$$\begin{aligned} &a_h(\Phi_H + Q_h(\Phi_H), \Psi_H) \\ &\quad = a_h(\Phi_H + Q_h(\Phi_H), \Psi_H + Q_h(\Psi_H)) + a_h(\Phi_H + Q_h(\Phi_H), Q_h(\Psi_H)), \end{aligned}$$

we see that it is only a small perturbation of the symmetric (coercive) G-LOD version, where the difference can be bounded by the quasi-orthogonality.

Even though the quasi-orthogonality *suggests* inf-sup stability, the given assumptions (A1)–(A8) do not seem to be sufficient for rigorously proving it. Here, it seems necessary to leave the abstract setting and to prove the inf-sup stability result for the various LOD realizations separately. For simplification, we therefore make the inf-sup stability to be an additional assumption [see (A9) below]. Later we give an example how to prove this assumption for a certain realization of the method. We also note that the inf-sup stability can be always verified numerically (for a given k) by investigating the system matrix $S^{\text{PG-LOD}}$ given by the entries

$$(S^{\text{PG-LOD}})_{ij} = a_h(\Phi_j + Q_h(\Phi_j), \Phi_i)$$

for $1 \leq i, j \leq N_H$ where N_H denotes the dimension of V_H and where $\{\Phi_i | \, 1 \leq i \leq N_H\}$ denotes a basis of V_H. To check the inf-sup stability we must compute the eigenvalues of $S^{\text{PG-LOD}}$. If their real parts are all strictly positive, we have inf-sup stability and the inf-sup constant is identical to the smallest real part of an eigenvalue. Standard approaches for computing the eigenvalues of a non-symmetric matrix are the Arnoldi method, the Jacobi–Davidson method and the non-symmetric Lanczos algorithm (cf. [39] for a comprehensive overview). Since N_H is moderately small, the cost for applying one of the methods are still feasible.

(A9) We assume that the LOD in Petrov–Galerkin formulation is inf-sup stable in the following sense: there exists a sequence of constants $\alpha(k)$ and a generic limit $\alpha_0 > 0$ (independent of H, h, k or the oscillations of A) such that $\alpha(k)$ converges with *exponential speed* to α_0, i.e. there exist constants $C(H)$ (possibly depending on H, but not on h, k or the oscillations of A) and a generic $\theta \in (0, 1)$ such that $|\alpha(k) - \alpha_0| \leq C(H)k^{d/2}\theta^k$. Furthermore it holds $\alpha(\bar{k}) = \alpha_0$ for all sufficiently large $\bar{k}$ and

$$\frac{a_h(\Phi^{\text{ms}}, \Phi_H)}{|||\Phi_H|||_H} \geq \alpha(k)|||\Phi^{\text{ms}}|||_h,$$

for all $\Phi^{\text{ms}} \in V^{\text{ms}}$ and $\Phi_H := ((I_H|_{V_H})^{-1} \circ I_H)(\Phi^{\text{ms}}) \in V_H$.

The following result states that the approximation quality of the LOD in Petrov–Galerkin formulation is of the same order as for the Galerkin LOD, up to a possible pollution term depending on $C_{H,h}$, but which still converges exponentially to zero.

Theorem 2 (A priori error estimate for PG-LOD) *Assume (A1)–(A9). Given a positive $k \in \mathbb{N}_{>0}$, let for all $T \in \mathcal{T}_H$ the patch $U(T) = U_k(T)$ be defined as in* (8) *and large*

enough so that the inf-sup constant in (A9) fulfills $\alpha(k) \geq \bar{\alpha}$ *for some* $\bar{\alpha} > 0$ *and let* $u_H^{PG\text{-}LOD}$ *be the unique solution of* (12). *Let* $u_h \in V_h$ *be the fine scale reference solution governed by* (2). *Then, the following a priori error estimate holds true*

$$\begin{aligned}&\left\|u_h - ((I_H|_{V_H})^{-1} \circ I_H)(u_H^{PG\text{-}LOD})\right\|_{L^2(\Omega)} + |||u_h - u_H^{PG\text{-}LOD}|||_h \\ &\lesssim (H + (1/H)^p(1 + (1/\bar{\alpha}))(1 + C_{H,h})k^{d/2}\theta^k)\|f\|_{L^2(\Omega)},\end{aligned}$$

where $0 < \theta < 1$ *and* $p \in \{0, 1\}$ *are the generic constants from assumption* (*A8*) *and* $C_{H,h}$ *as in* (*A5*).

3.3 Example 1: continuous Galerkin finite element method

The previous subsection showed that the Petrov–Galerkin formulation of the LOD does not suffer from a loss in accuracy with respect to the symmetric formulation. In this subsection, we give the specific example of the LOD for the continuous Galerkin finite element method. In particular, we discuss the advantage of the PG formulation over the symmetric formulation. Let us first introduce the specific setting and the corresponding argument about the validity of (A4)–(A9) on this setting.

In addition to the assumptions that we made on the shape regular partitions $\mathcal{T}_H$ and $\mathcal{T}_h$ in Sect. 2.1, we assume that $\mathcal{T}_H$ and $\mathcal{T}_h$ are either triangular or quadrilateral meshes. Accordingly, for $\mathcal{T} = \mathcal{T}_H, \mathcal{T}_h$ we denote

$$\begin{aligned}P_1(\mathcal{T}) &:= \left\{v \in C^0(\Omega) | \forall T \in \mathcal{T}, v|_T \text{ is a polynomial of total degree} \leq 1\right\} \quad \text{and} \\ Q_1(\mathcal{T}) &:= \left\{v \in C^0(\Omega) | \forall T \in \mathcal{T}, v|_T \text{ is a polynomial of partial degree} \leq 1\right\}\end{aligned}$$

and define $V_h := P_1(\mathcal{T}_h) \cap H_0^1(\Omega)$ if $\mathcal{T}_h$ is simplicial and $V_h := Q_1(\mathcal{T}_h) \cap H_0^1(\Omega)$ if it is a quadrilation. The coarse space $V_H \subset V_h$ is defined in the same fashion and since $\mathcal{T}_h$ is a refinement of $\mathcal{T}_H$, assumption (A6) is obviously fulfilled. For simplicity, we also assume that the coarse mesh $\mathcal{T}_H$ is quasi-uniform (which is the typical choice in applications).

The bilinear form $a_h(\cdot, \cdot)$ is defined by the standard energy scalar product on $H_0^1(\Omega)$ that belongs to the elliptic problem to solve, i.e.

$$a_h(v, w) := \int_\Omega A\nabla v \cdot \nabla w \quad \text{for } v, w \in H_0^1(\Omega).$$

Accordingly, we set $|||v|||_h := |||v|||_H := \|A^{1/2}\nabla v\|_{L^2(\Omega)}$ for $v \in H^1(\Omega)$. Hence, assumptions (A5) and (A6) are fulfilled and the solution $u_h \in V_h$ of (2) is nothing but the standard continuous Galerkin finite element solution on the fine grid $\mathcal{T}_h$.

Next, we specify $I_H : V_h \to W_h$ in (A7). For this purpose, let $\Phi_z \in V_H$ be the nodal basis function associated with the coarse grid node $z \in \mathcal{N}_H$, i.e., $\Phi_z(y) = \delta_{yz}$. Let I_H be the weighted Clément-type quasi-interpolation operator as defined in [9,10]:

$$I_H \colon H_0^1(\Omega) \to V_H, \quad v \mapsto I_H(v) := \sum_{z \in \mathcal{N}_H^0} v_z \Phi_z \quad \text{with } v_z := \frac{(v, \Phi_z)_{L^2(\Omega)}}{(1, \Phi_z)_{L^2(\Omega)}}. \quad (13)$$

First we note that it was shown in [35] that $(I_H)_{|V_H} \colon V_H \to V_H$ is an isomorphism [but not a projection, i.e. $(I_H|_{V_H})^{-1} \neq I_H|_{V_H}$]. Hence, $(I_H)_{|V_H}^{-1}$ exists. This is one of the properties in (A7). The L^2- and H^1-stability of I_H, as well as corresponding approximation properties, were proved in [9]. It only remains to check the H^1-stability of $(I_H)_{|V_H}^{-1}$. Unfortunately, this property is not trivial to fulfill. First, we note that it was shown in [34] that the mapping $(I_H)_{|V_H}^{-1} \circ I_H$ is nothing but the L^2-projection $P_{L^2} \colon H_0^1(\Omega) \to V_H$ (see also Remark 5 below). Consequently, the question of H^1-stability of $(I_H)_{|V_H}^{-1}$ is equivalent to the question of H^1-stability of the L^2-projection. This result is well-established for quasi uniform grids (cf. [6]) as assumed at the beginning of this section. However it is still open for arbitrary refinements. The most recent results on this issue can be found in [7,15,29], where the desired H^1-stability was shown for certain types of adaptively refined meshes. To avoid complicated mesh assumptions in this paper, we simply assume $\mathcal{T}_H$ to be quasi-uniform. This is not very restrictive since adaptive refinements should typically take place on the fine mesh $\mathcal{T}_h$. Alternatively, in light of [7,15,29], we could also directly assume that the L^2-projection on V_H is H^1-stable to allow more general coarse meshes.

It remains to specify $a_h^T(\cdot, \cdot)$, which we define by

$$a_h^T(v, w) := \int_T A \nabla v \cdot \nabla w \quad \text{for } v, w \in H_0^1(\Omega).$$

Let us for simplicity denote $|||\cdot|||_{h,T} := \|A^{1/2} \nabla \cdot \|_{L^2(T)}$. The decay assumption (A8) was essentially proved in [19, Lemma 3.6], which established the existence of a generic constant $0 < \theta < 1$ with the properties as in (A8) such that

$$\left|\left|\left|(Q_h - Q_h^\Omega)(\Phi_H)\right|\right|\right|_h^2 \lesssim k^d \theta^{2k} \sum_{T \in \mathcal{T}_H} \left|\left|\left| Q_h^{\Omega,T}(\Phi_H) \right|\right|\right|_h^2, \quad (14)$$

for all $\Phi_H \in V_H$. On the other hand we have by $|||\cdot|||_{h,T} = \|A^{1/2} \nabla \cdot \|_{L^2(T)}$ and Eq. (6) that

$$\begin{aligned} \left|\left|\left| Q_h^{\Omega,T}(\Phi_H) \right|\right|\right|_h^2 &\lesssim a_h \left(Q_h^{\Omega,T}(\Phi_H), Q_h^{\Omega,T}(\Phi_H) \right) \\ &= -a_h^T \left(\Phi_H, Q_h^{\Omega,T}(\Phi_H) \right) \\ &\lesssim |||\Phi_H|||_{h,T} \left|\left|\left| Q_h^{\Omega,T}(\Phi_H) \right|\right|\right|_h . \end{aligned} \quad (15)$$

Hence, by plugging this result into (14):

$$\begin{aligned}
\big|\!\big|\!\big|(Q_h - Q_h^\Omega)(\Phi_H)\big|\!\big|\!\big|_h^2 &\lesssim k^d\theta^{2k} \sum_{T\in\mathcal{T}_H} |||\Phi_H|||_{h,T}^2 \\
&\lesssim k^d\theta^{2k}|||\Phi_H|||_h^2 = k^d\theta^{2k}|||((I_H|_{V_H})^{-1}\circ I_H)(\Phi_H + Q_h^\Omega(\Phi_H))|||_h^2 \\
&\overset{(A7)}{\lesssim} k^d\theta^{2k}|||\Phi_H + Q_h^\Omega(\Phi_H)|||_h^2,
\end{aligned}$$

which proves that assumption (A8) holds even with $p = 0$. The remaining assumption (A9) is less obvious and requires a proof. We give this proof for the continuous Galerkin PG-LOD in Sect. 5. We summarize the result in the following lemma.

Lemma 1 (Inf-sup stability of continuous Galerkin PG-LOD) *For all $T \in \mathcal{T}_H$ let $U(T) = U_k(T)$ for $k \in \mathbb{N}$. Then there exist generic constants C_1, C_2 (independent of H, h, k or the oscillations of A) and $0 < \theta < 1$ as in assumption (A8), so that it holds*

$$\inf_{\Phi_H\in V_H}\sup_{\Phi^{\mathrm{ms}}\in V^{\mathrm{ms}}} \frac{a(\Phi^{\mathrm{ms}}, \Phi_H)}{|||\Phi^{\mathrm{ms}}|||_h|||\Phi_H|||_h} \geq \alpha(k),$$

for $\alpha(k) := C_1\alpha - C_2k\theta^k\omega(\Phi^{\mathrm{ms}})$ and

$$0 \leq \omega(\Phi^{\mathrm{ms}}) := \inf_{w_h\in W_h^T} \frac{\big\|\nabla\Phi^{\mathrm{ms}} - \nabla((I_H|_{V_H})^{-1}\circ I_H)(\Phi^{\mathrm{ms}}) - \nabla w_h\big\|}{\big\|\nabla\Phi^{\mathrm{ms}} - \nabla((I_H|_{V_H})^{-1}\circ I_H)(\Phi^{\mathrm{ms}})\big\|} \leq 1,$$

where $W_h^T := \{w_h \in W_h |\ w_h|_T \in W_h(T)\}$, i.e. the space of all functions from W_h that are zero on the boundary of the coarse grid elements. Observe that $\alpha(k)$ converges with exponential speed to αC_1. Furthermore we have $\alpha(0) = C_1\alpha$ [because $\omega(\Phi^{\mathrm{ms}}) = 0$] and also $\alpha(\ell) = C_1\alpha$ for all sufficiently large ℓ.

Remark 3 Let $U(T) = U_k(T)$ for $k \in \mathbb{N}$ with $k \gtrsim |\log(H)|$, then the CG-LOD in Petrov–Galerkin formulation is inf-sup stable for sufficiently small H. In particular, there exists a unique solution of problem (12).

Remark 4 Lemma 1 does not allow to conclude to inf-sup stability for the regime $0 < k \ll |\log(H)|$. However, even though this regime is not of practical relevance, it is interesting to note that we could not observe a violation of the inf-sup stability for any value of k and in any numerical experiment that we set up so far.

Since assumptions (A1)–(A9) are fulfilled for this setting, Theorems 1 and 2 hold true for the arising method. Furthermore, we have $p = 0$ and $C_{H,h} = 1$ in the estimates, meaning that the $(1/H)$-pollution in front of the decay term vanishes. We can summarize the result in the following conclusion.

Conclusion 3 *Assume the (continuous Galerkin) setting of this subsection and let $u_H^{PG\text{-}LOD}$ denote a Petrov–Galerkin solution of* (12)*. If $k \gtrsim mH|\log(H)|$ for $m \in \mathbb{N}$, then it holds*

$$\big\|u_h - u_H^{PG\text{-}LOD}\big\|_{H^1(\Omega)} \lesssim (H + H^m)\|f\|_{L^2(\Omega)}.$$

In particular, the bound is independent of $C_{H,h}$.

3.4 Discussion of advantages

The central disadvantage of the Galerkin LOD is that it requires a communication between solutions of different patches. Consider for instance the assembly of the system matrix that belongs to problem (10). Here it is necessary to compute entries of the type

$$\int_\Omega A\nabla(\Phi_i + Q_h(\Phi_i)) \cdot \nabla(\Phi_j + Q_h(\Phi_j)),$$

which particularly involves the computation of the term

$$\sum_{\substack{T\in\mathcal{T}_H \\ T\subset\omega_i}} \sum_{\substack{K\in\mathcal{T}_H \\ K\subset\omega_j}} \int_{U(T)\cap U(K)} A\nabla Q_h^T(\Phi_i) \cdot \nabla Q_h^K(\Phi_j), \tag{16}$$

where $\Phi_i, \Phi_j \in V_H$ denote two coarse nodal basis functions and ω_i and ω_j its corresponding supports. The efficient computation of (16) requires information about the intersection area of any two patches $U(T)$ and $U(K)$. Even if T and K are not adjacent or close to each other, the intersection of the corresponding patches can be complicated and non-empty. The drawback becomes obvious: first, these intersection areas must be determined, stored and handled in an efficient way and second, the number of relevant entries of the stiffness matrix (i.e. the non-zeros) increases considerably. Note that this also leads to a restriction in the parallelization capabilities, in the sense that the assembly of the stiffness matrix can only be 'started' if the correctors $Q_h(\Phi_i)$ are already computed. Another disadvantage is that the assembly of the right hand side vector associated with (f, Φ^{ms}) in (10) is much more expensive since it involves the computation of entries $(f, \Phi_i + Q_h(\Phi_i))_{L^2(\Omega)}$. First, the integration area is $\cup\{U(T)|\, T \in \mathcal{T}_H,\ T \subset \omega_i\}$ instead of typically ω_i. This increases the computational costs. At the same time, it is also hard to assemble these entries by performing (typically more efficient) element-wise computations (for which each coarse element has to be visited only once). Second, $(f, \Phi_i + Q_h(\Phi_i))_{L^2(\Omega)}$ involves a quadrature rule of high order, since $Q_h(\Phi_i)$ is rapidly oscillating. These oscillations must be resolved by the quadrature rule, even if f is a purely macroscopic function that can be handled exactly by a low order quadrature. Hence, the costs for computing $(f, \Phi_i + Q_h(\Phi_i))_{L^2(\Omega)}$ depend indirectly on the oscillations of A. Finally, if the LOD shall be applied to a sequence of problems of type (1), which only differ in the source term f (or a boundary condition), the system matrix can be fully reused, but the complications that come with the right hand side have to be addressed each time again.

The Petrov–Galerkin formulation of the LOD clearly solves these problems without suffering from a loss in accuracy. In particular:

– The PG-LOD does not require any communication between two different patches and the resulting stiffness matrix is sparser than the one for the symmetric LOD. In particular, the entries of the system matrix S can be computed with the following algorithm:

Let S denote the empty system matrix with entries S_{ij}.

Algorithm: assembleSystemMatrix($\mathcal{T}_H, \mathcal{T}_h, k$)

In parallel **foreach** $T \in \mathcal{T}_H$ **do**
  foreach $z_i \in \mathcal{N}_H^0$ with $z_i \in \overline{T}$ **do**
    compute $Q_h^T(\Phi_{z_i}) \in W_h(U_k(T))$ with

$$a(Q_h^T(\Phi_{z_i}), w_h) = -\int_T A\nabla\Phi_{z_i} \cdot \nabla w_h \quad \text{for all } w_h \in W_h(U_k(T)).$$

    foreach $z_j \in \mathcal{N}_H^0$ with $z_j \in \overline{U(T)}$ **do**
      update the system matrix:

$$S_{ji} \mathrel{+}= \int_{\omega_j} A\left(\Phi_{z_i} + \nabla Q_h^T(\Phi_{z_i})\right) \cdot \nabla\Phi_{z_j}.$$

    end
  end
end

Observe that it is possible to add the local terms $a(\Phi_{z_i} + Q_h^T(\Phi_{z_i}), \Phi_{z_j})$ directly to the system matrix S, i.e. the assembling of the matrix is parallelized in a straightforward way and does not rely on the availability of other results.

– Replacing the source term f in (1), only involves the re-computation of the terms $(f, \Phi_i)_{L^2(\omega_i)}$ for coarse nodal basis functions Φ_i, i.e. the same costs as for the standard FE method on the coarse scale. Furthermore, the choice of the quadrature rule relies purely on f, but not on the oscillations of A.

Besides the previously mentioned advantages, there is still a memory consuming issue left: the storage of the local correctors $Q_h^T(\Phi_{z_i})$. These local correctors need to be saved in order to express the final approximation $u_H^{\text{PG-LOD}}$ which is spanned by the multiscale basis functions $\Phi_i + Q_h(\Phi_i)$. As long as we are interested in a good H^1-approximation of the solution, this problem seems to be unavoidable. However, in many applications we can even overcome this difficulty by exploiting another very big advantage of the PG-LOD: Theorem 2 predicts that alone the 'coarse part' of $u_H^{\text{PG-LOD}}$, denoted by $u_H := ((I_H|_{V_H})^{-1} \circ I_H)(u_H^{\text{PG-LOD}})) \in V_H$, already exhibits very good L^2-approximation properties, i.e. if $k \gtrsim |\log(H)|$ we have essentially

$$\|u_h - u_H\|_{L^2(\Omega)} \leq \mathrm{O}(H).$$

In contrast to $u_H^{\text{PG-LOD}}$, the representation of u_H does only require the classical coarse finite element basis functions. Hence, we can use the algorithm presented earlier, with the difference that we can immediately delete $Q_h^T(\Phi_i)$ after updating the stiffness matrix. Observe that even if computations have to be repeated for different

source terms f, this stiffness matrix can be reused again and again. Also, if a user is interested in the fine scale behavior in a local region [but the $Q_h^T(\Phi_i)$ were already dropped], it is still possible to quickly re-compute the desired local corrector for the region.

As an application, consider for instance the case that the problem

$$\int_\Omega A\nabla u \cdot \nabla v = \int_\Omega fv$$

describes the diffusion of a pollutant in groundwater. Here, u describes the concentration of the pollutant, A the (rapidly varying) hydraulic conductivity and f a source term describing the injection of the pollutant. In such a scenario, there is typically not much interest in finding a good approximation of the (locally fluctuating) gradient ∇u, but rather in the macroscopic behavior of pollutant u, i.e. in purely finding a good L^2-approximation that allows to conclude where the pollutant spreads. A similar scenario is the investigation of the properties of a composite material, where A describes the heterogenous material and f some external force. Again, the interest is in finding an accurate L^2-approximation. Besides, the corresponding simulations are typically performed for a variety of different source terms f, investigating different scenarios. In this case, the PG-LOD yields reliable approximations with very low costs, independent of the structure of A.

Remark 5 (*Relation to the L^2-projection*) Assume the setting of this subsection. In [34] it was shown that $(v_H, w_h)_{L^2(\Omega)} = 0$ for all $v_H \in V_H$ and $w_h \in W_h$, i.e. V_H and W_h are L^2-orthogonal. This implies that

$$(I_H|_{V_H})^{-1} \circ I_H = P_{L^2},$$

with P_{L^2} denoting the L^2-projection on V_H. To verify this, let $v_h \in V_h$ be arbitrary. Then due to $V_h = V_H \oplus W_h$ we can write $v_h = v_H + w_h$ (with $v_H \in V_H$ and $w_h \in W_h$) and observe for all $\Phi_H \in V_H$

$$\begin{aligned}\int_\Omega P_{L^2}(v_h) \quad \Phi_H &= \int_\Omega v_h \quad \Phi_H \overset{V_H \perp_{L^2} W_h}{=} \int_\Omega v_H \quad \Phi_H \\ &= \int_\Omega ((I_H|_{V_H})^{-1} \circ I_H)(v_H) \quad \Phi_H \overset{I_H(w_h)=0}{=} \int_\Omega ((I_H|_{V_H})^{-1} \circ I_H)(v_h) \quad \Phi_H.\end{aligned}$$

Hence, $u_H^{\text{PG-LOD}} = u_H + Q_h(u_H)$ with $u_H = P_{L^2}(u_H^{\text{PG-LOD}})$.

Conclusion 4 (Application to homogenization problems) *Assume the setting of this subsection and let P_{L^2} denote the L^2-projection on V_H as in Remark* 5. *We consider now a typical homogenization setting with $(\epsilon)_{>0} \subset \mathbb{R}_{>0}$ being a sequence of positive parameters that converges to zero. Let $Y := [0,1]^d$ denote the unique cube in $\mathbb{R}^d$ and let $A^\epsilon(x) = A_p(x, \frac{x}{\epsilon})$ for a function $A_p \in W^{1,\infty}(\Omega \times Y)$ that is Y-periodic in the second argument (hence A^ϵ is rapidly oscillating with frequency ϵ). The corresponding exact solution of problem* (1) *shall be denoted by $u_\epsilon \in H_0^1(\Omega)$. It is well known* (*cf.* [3])

that u_ϵ converges weakly in H^1 (but not strongly) to some unique function $u_0 \in H_0^1(\Omega)$. Furthermore, if $\|f\|_{L^2(\Omega)} \lesssim 1$ it holds $\|u_\epsilon - u_0\|_{L^2(\Omega)} \lesssim \epsilon$. With Theorem 2 together with Remark 5 and standard error estimates for FE problems, we hence obtain:

$$\|u_0 - u_H\|_{L^2(\Omega)} \lesssim \epsilon + \left(\frac{h}{\epsilon}\right)^2 + H,$$

for $u_H = P_{L^2}(u_H^{PG\text{-}LOD})$. Homogenization problems are typical problems, where one is often purely interested in the L^2-approximation of the exact solution u_ϵ, meaning one is interested in the homogenized solution u_0.

As discussed in this section, the PG-LOD can have significant advantages over the (symmetric) G-LOD with respect to computational costs, efficiency and memory demand. In Sect. 4.1 we additionally present a numerical experiment to demonstrate that the approximations produced by the PG-LOD are in fact very close to the ones produced by (symmetric) G-LOD, i.e. not only of the same order as predicted by the theorems, but also of the same quality.

Remark 6 (*Nonlinear problems*) The above results suggest that the advantages can become even more pronounced for certain types of nonlinear problems. For instance, consider a well-posed problem of the type

$$-\nabla \cdot A\nabla u + c(u) = f,$$

for a nonlinear function c. Here, it is intuitively reasonable to construct $Q_h(\Phi_H)$ as before using only the linear elliptic part of the problem. This is a preprocessing step that is done once and can be immediately deleted stiffness matrix is calculated and saved. Then we solve for $u_H \in V_H$ that satisfies

$$(A\nabla(u_H + Q_h(u_H)), \nabla\Phi_H)_{L^2(\Omega)} + (c(u_H), \Phi_H)_{L^2(\Omega)} = (f, \Phi_H)_{L^2(\Omega)}$$

for all $\Phi_H \in V_H$. Clearly, typical iterative solvers can be utilized to solve this variational problem. This iteration is inexpensive because it is done in V_H and the preconstructed stiffness matrix can be fully reused within every iteration and since the other contributions are independent of Q_h. Performing iterations on the coarse space for solving nonlinear problems within the framework of multiscale finite element has been investigated (see for example [12,16]).

3.5 Example 2: discontinuous Galerkin finite element method

In this subsection, we apply the results of Sect. 3.2 to a LOD Method that is based on a discontinuous Galerkin approach. The DG-LOD was originally proposed in [14] and fits into the framework proposed in Sect. 2.2. First, we show that the setting fulfills assumptions (A4)–(A8) and after we discuss the advantage of the PG DG-LOD over the symmetric DG-LOD. For simplification, we assume that A is piecewise constant with respect to the fine mesh $\mathcal{T}_h$ so that all of the subsequent traces are well-defined.

Again, we make the same assumptions on the partitions $\mathcal{T}_H$ and $\mathcal{T}_h$ as in Sect. 2.1 and additionally assume that $\mathcal{T}_H$ and $\mathcal{T}_h$ are either triangular or quadrilateral meshes. The corresponding total sets of edges (or faces for $d = 3$) are denoted by $\mathcal{E}_h$ (for $\mathcal{T}_h$), where $\mathcal{E}_h(\Omega)$ and $\mathcal{E}_h(\partial\Omega)$ denotes the set of interior and boundary edges, respectively.

Furthermore, for $\mathcal{T} = \mathcal{T}_H, \mathcal{T}_h$ we denote the spaces of discontinuous functions with total, respectively partial, polynomial degree equal to or less than 1 by

$$\mathcal{P}_1(\mathcal{T}) := \left\{ v \in L^2\Omega) | \forall T \in \mathcal{T},\, v|_T \text{ is a polynomial of total degree} \leq 1 \right\} \quad \text{and}$$
$$\mathcal{Q}_1(\mathcal{T}) := \left\{ v \in L^2(\Omega) | \forall T \in \mathcal{T},\, v|_T \text{ is a polynomial of partial degree} \leq 1 \right\}$$

and define $V_h := P_1(\mathcal{T}_h)$ if $\mathcal{T}_h$ is a triangulation and $V_h := Q_1(\mathcal{T}_h)$ if it is a quadrilation. The coarse space $V_H \subset V_h$ is defined in the same fashion with $\mathcal{T}_H$ instead of $\mathcal{T}_h$. Note that these spaces are no subspaces of $H^1(\Omega)$ as in the previous example. For this purpose, we define ∇_h to be the $\mathcal{T}_h$-piecewise gradient [i.e. $(\nabla_h v_h)|t := \nabla(v_h|t)$ for $v_h \in V_h$ and $t \in \mathcal{T}_h$].

For every edge/face $e \in \mathcal{E}_h(\Omega)$ there are two adjacent elements $t^-, t^+ \in \mathcal{T}_h$ with $e = \partial t^- \cap \partial t^+$. We define the jump and average operators across $e \in \mathcal{E}_h(\Omega)$ by

$$[v] := (v|t^- - v|t^+) \quad \text{and} \quad \{A\nabla v \cdot n\} := \frac{1}{2}((A\nabla v)|t^- + (A\nabla v)|t^+) \cdot n,$$

where n be the unit normal on e that points from t^- to t^+, and on $e \in \mathcal{E}_h(\partial\Omega)$ by

$$[v] := w|t \quad \text{and} \quad \{A\nabla v \cdot n\} := (A\nabla v)|t \cdot n$$

where n is the outwards unit normal of $t \in \mathcal{T}_h$ (and Ω). Observe that flipping the roles of t^- and t^+ leads to the same terms in the bilinear form defined below.

With that, we can define the typical bilinear form that characterizes the discontinuous Galerkin method:

$$a_h(v_h, w_h) := (A\nabla_h v_h, \nabla_h w_h)_{L^2(\Omega)} + \sum_{e\in\mathcal{E}_h} \frac{\sigma}{h_e}([v_h], [w_h])_{L^2(e)} - \sum_{e\in\mathcal{E}_h} \left((\{A\nabla v_h \cdot n\}, [w_h])_{L^2(e)} + (\{A\nabla w_h \cdot n\}, [v_h])_{L^2(e)} \right).$$

Here, σ is a penalty parameter that is chosen sufficiently large and $h_e = \operatorname{diam}(e)$. The coarse bilinear form $a_H(\cdot,\cdot)$ is defined analogously with coarse scale quantities. It is well known, that $a_h(\cdot,\cdot)$ [respectively $a_H(\cdot,\cdot)$] is a scalar product on V_h (respectively V_H). Consequently (A4) is fulfilled. As a norm on V_h that fulfills assumption (A5), we can pick

$$|||v|||_h := \left\| A^{1/2} \nabla_h v \right\|_{L^2(\Omega)} + \left(\sum_{e \in \mathcal{E}_h} \frac{\sigma}{h_e} \|[v]\|^2_{L^2(e)} \right)^{1/2}.$$

Analogously, we define $|||v|||_H$ to be a norm on V_H. In this case we obtain the constant $C_{H,h} = \sqrt{H/h}$. Assumption (A6) is obviously fulfilled.

As the operator in assumption (A7) we pick the L^2-projection on V_H, i.e. for $v_h \in V_h$ we have

$$(I_h(v_h), \Phi_H)_{L^2(\Omega)} = (v_h, \Phi_H)_{L^2(\Omega)} \quad \text{for all } \Phi_H \in V_H.$$

In [14, Lemma 5] it was proved that the operator fulfills the desired approximation and stability properties. Since I_H is a projection, we have $I_H = (I_H|_{V_H})^{-1}$ and hence obviously also $|||\cdot|||_H$-stability of the inverse on V_H.

The localized bilinear form $a_h^T(\cdot,\cdot)$ in (5) is defined by $a_h^T(v_h, w_h) := a_h(\chi_T v_h, w_h)$ where $\chi_T = 1$ in T and 0 otherwise, is the element indicator function. Obviously we have for all $v_h, w_h \in V_h$ that

$$a_h(v_h, w_h) = \sum_{T \in \mathcal{T}_H} a_h^T(v_h, w_h).$$

In [14] the DG-LOD is presented in a slightly different way, in the sense that there exists no general corrector operator Q_h. Instead, 'basis function correctors' are introduced. However, it is easily checkable that each of these 'basis function correctors' is nothing but the corrector operator, defined via (6), applied to an original coarse basis function. Therefore, the correctors given by (6) are just an extension of the definition to arbitrary coarse functions. Hence, both methods coincide and are just presented in a different way.

Next, we discuss (A8). This property was shown in [14, Lemmas 11 and 12], however not explicitly for the setting that we established in Definition 3. It was only shown for $\Phi_H = \lambda_{T,j}$, where $\lambda_{T,j} \in V_H$ denotes a basis function on T associated with the j'th node. However, the proofs in [14] directly generalize to the local correctors $Q_h^T(\Phi_H)$ given by Eq. (6). More precisely, following the proofs in [14] it becomes evident that the availability of the required decay property (A8) purely relies on the fact, that the right hand side in the local problems is only locally supported (with a support that remains fixed, even if the patch size decreases). Therefore (A8) can be proved analogously.

Finally, assumption (A9) is not easy to verify. It is obviously fulfilled for the case $U(T) = \Omega$, but the generalized result is harder to verify. The following result holds under some restrictions on the meshes $\mathcal{T}_H$ and $\mathcal{T}_h$.

Lemma 2 (Inf-sup stability of discontinuous Galerkin PG-LOD) *Assume that $\mathcal{T}_H$ is quasi-uniform and that there exists an exponent $m \in \mathbb{R}$ with $m > 1$ such that for all $T \in \mathcal{T}_H$*

$$\operatorname{diam}(T)^m \lesssim \min \left\{ h_e |\, e \in \mathcal{E}_h \text{ and } e \subset \overline{T} \right\}$$

(i.e. if $\mathcal{T}_h$ is also quasi-uniform we assume $H^m \lesssim h$). If $k \in \mathbb{N}$ is such that $k \gtrsim \frac{(m+3)}{2}|\log(H)|$ then, for sufficiently small H, there exist generic positive constants C_1, C_2 such that

$$\inf_{\Phi_H \in V_H} \sup_{\Phi^{\mathrm{ms}} \in V^{\mathrm{ms}}} \frac{a_h(\Phi^{\mathrm{ms}}, \Phi_H)}{|||\Phi^{\mathrm{ms}}|||_h |||\Phi_H|||_H} \geq C_1(\alpha - C_2 H).$$

Hence, we have inf-sup stability for sufficiently small H.

The proof is given in Sect. 5. We note that the inf-sup stability can be observed numerically already under weaker assumptions (see Sect. 4) and that it is in general 'a reasonable thing to expect' as discussed in Remark 2.

In conclusion, the discontinuous Galerkin LOD in Petrov–Galerkin formulation fulfills the assumptions of our framework [up to a discussion on (A9)]. The advantages that we discussed in the previous subsection for the Petrov–Galerkin continuous finite element method in terms of memory and efficiency remains true. However, for the PG DG-LOD there is a very important additional advantage. It is known that the classical DG method has the feature of local mass conservation with respect to the elements of the underlying mesh. This can be easily checked by testing with the indicator function of an element T in the variational formulation of the method. The local mass conservation is a highly desired property for various flow and transport problems. However, the DG-LOD does not preserve this property, since the indicator function of an element (whether coarse or fine) is not in the space V^{ms}. This problem is solved in the PG DG-LOD, where we can test with any element from V_H and in particular with the indicator function of a coarse element. Hence, in contrast to the symmetric DG-LOD, the PG DG-LOD is locally mass conservative with respect to coarse elements $T \in \mathcal{T}_H$. This allows for example the coupling of the PG DG-LOD for an elliptic problem with the solver for a hyperbolic conservation law, which was not possible before without relinquishing the mass conservation. We discuss this further in the next subsection.

3.6 Perspectives towards two-phase flow

In this subsection, we investigate an application of the Petrov–Galerkin DG-LOD in the simulation of two-phase flow as governed by the Buckley–Leverett equation. Specifically, the LOD framework is utilized to solve the pressure equation, which is an elliptic boundary value problem, and is coupled with a solver for a hyperbolic conservation law. The Buckley–Leverett equation can be used to model two-phase flow in a porous medium. Generally, the flow of two immiscible and incompressible fluids is driven by the law of mass balance for the two fluids:

$$\Theta \partial_t S_\alpha + \nabla \cdot \boldsymbol{v}_\alpha = q_\alpha \quad \text{in } \Omega \times (0, T_{end}] \quad \text{for } \alpha = w, n. \tag{17}$$

Here, Ω is a computational domain, $(0, T_{end}]$ a time interval, the unknowns $S_w, S_n : \Omega \to [0, 1]$ describe the saturations of a wetting and a non-wetting fluid and $\boldsymbol{v}_w$ and $\boldsymbol{v}_n$ are the corresponding fluxes. Furthermore, Θ describes the porosity and q_w and q_n

are two source terms. Darcy's law relates the fluxes with the two unknown pressures p_n and p_w by

$$\boldsymbol{v}_\alpha = -K\frac{k_\alpha(S_\alpha)}{\mu_\alpha}(\nabla p_\alpha - \rho_\alpha \boldsymbol{g}) \quad \text{for } \alpha = w, n.$$

Here, K denotes the hydraulic conductivity, k_w and k_n the relative permeabilities depending on the saturations, μ_w and μ_n the viscosities, ρ_w and ρ_n the densities and $\boldsymbol{g}$ the gravity vector. The saturations are coupled via $S_n + S_w = 1$ and a relation between the two pressures is typically given by the capillary pressure relation $P_c(S_w) = p_n - p_w$ for a monotonically decreasing capillary pressure curve P_c. In this case, we obtain the full two-phase flow system, which consists of two strongly coupled, possibly degenerate parabolic equations. However, if we neglect the gravity and the capillary pressure [i.e. assume that $P_c(S_w) = 0$], the system reduces to the so called Buckley–Leverett system with an elliptic pressure equation and an hyperbolic equation for the saturation:

$$-\nabla \cdot (K\lambda(S)\nabla p) = q \quad \text{and} \quad \Theta \partial_t S + \nabla \cdot (f(S)\boldsymbol{v}) = q_w, \tag{18}$$

where we have $S = S_w$, $p = p_w = p_n$, the total mobility $\lambda(S) := \frac{k_w(S)}{\mu_w} + \frac{k_n(1-S)}{\mu_n} > 0$, the flux $\boldsymbol{v} := -K\lambda(S)\nabla p$ and the flux function $f(S) := \frac{k_w(S)}{\mu_w \lambda(S)}$. The total source is given by $q := \frac{q_w + q_n}{2}$. Observe that (18) is obtained from (17) by summing up the equations for the saturations, using $\partial_t(s_n + s_w) = \partial_t 1 = 0$.

An application for which neglecting the capillary pressure is typically justified are oil recovery processes. Here, a replacement fluid, such as water or liquid carbon dioxide, is injected with very high rates into a reservoir to move oil towards a production well. However, often oil is trapped at interfaces of a low and a high conductivity region. This oil would become inaccessible which is why detailed simulations are required before the replacement fluid can be actually injected.

Depending on the choice for the mobilities, the hyperbolic Buckley–Leverett problem can have one or more weak solutions (cf. [32]). One approach for solving the problem numerically is to use an operator splitting technique as proposed in [4], which is more well-known as the *implicit pressure explicit saturation* (IMPES). Here, the hyperbolic Buckley–Leverett problem is treated with an explicit time stepping method where the flux velocity $\boldsymbol{v}$ is kept constant for a certain time interval and then updated by solving the elliptic problem with the saturation from the previous time step (see Fig. 1 for an illustration). Alternatively, depending on the type of the flux function f, the hyperbolic problem can be also solved implicitly with a suitable numerical scheme for conservation laws (cf. [30]) where the flux $\boldsymbol{v}$ arising from the Darcy equation is, as in the previous case, only updated every fixed number of time steps.

Observe that the difficulties produced by the multiscale character of the problem are primarily related to the elliptic part of the problem. Once the Darcy problem is solved to update the flux velocity, the grid for solving the hyperbolic problem can be significantly coarsened. The reason is that $\boldsymbol{v} = -K\lambda(S)\nabla p$ is possibly still rapidly

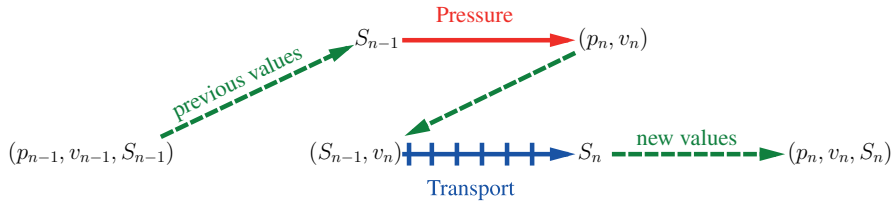


Fig. 1 A schematic of operator splitting (IMPES) for system (18)

oscillating, but the relative amplitude of the oscillations is expected to remain small. In other words, just like for standard elliptic homogenization problems, $\boldsymbol{v}$ behaves like an upscaled quantity $-K_0\lambda(S_0)\nabla p_0$ with effective/homogenized functions K_0, S_0 and p_0.

Remark 7 Any realization of the LOD involves to solve a number of local problems that help us to construct the low dimensional space V^{ms}. One might consider to update this space every time that the Darcy problem has to be solved with a new saturation. However, since $\lambda(S)$ is essentially macroscopic, it is generally sufficient to construct the space only once for $\lambda = 1$ and reuse the result for every time step. This makes solving the elliptic multiscale problem much cheaper after the multiscale space is assembled. A justification for this reusing of the basis can be e.g. found in [20] where it was shown that oscillations coming from advective terms can be often neglected in the construction of a multiscale basis. Under certain assumptions, the relative permeability $\lambda(S)$ can in fact be interpreted as a pure enforcement by an additional advection term.

4 Numerical experiments

In this section we present two different model problems. The first one involves a LOD methods for the continuous Galerkin method. Here, we compare the results obtained with the symmetric version of the method with the results obtained for the Petrov–Galerkin version. In the second model problem, we use a PG DG-LOD for solving the Buckley–Leverett system.

4.1 Continuous Galerkin PG-LOD for elliptic multiscale problems

In this section, we use the setting established in Sect. 3.3. All experiments were performed with the G-LOD and PG-LOD for the continuous finite element method.

In order to be more flexible in the choice of the localization patches $U(T)$, we make subsequently use of "half" or "quarter coarse layers", i.e. $k \in \mathbb{Q}_{\geq 0}$. This can be easily accomplished by extending Definition 4 straightforwardly to fine grid layers, i.e. for $k \in \mathbb{Q}_{\geq 0}$ and $T \in \mathcal{T}_H$ we define the number of fine layers by $\ell := \lfloor \frac{kH}{h} \rfloor \in \mathbb{N}$ and the corresponding (broken layer) patch by $U_k(T) := U_{\mathrm{f},\ell}(T)$, where iteratively

$U_{\mathrm{f},\ell}(T) := \cup\overline{\{t \in \mathcal{T}_h \mid t \cap U_{\mathrm{f},\ell-1}(T) \neq \emptyset\}}$ and $U_{\mathrm{f},0}(T) := \overline{T}$. This allows us a more careful investigation of the decay behavior.

Let u_h be the solution of (2). In the following we denote by $\|\cdot\|^{\mathrm{rel}}_{L^2(\Omega)}$ and $\|\cdot\|^{\mathrm{rel}}_{H^1(\Omega)}$ the corresponding relative error norms defined by

$$\|u_h - v_h\|^{\mathrm{rel}}_{L^2(\Omega)} := \frac{\|u_h - v_h\|_{L^2(\Omega)}}{\|u_h\|_{L^2(\Omega)}} \quad \text{and}$$
$$\|u_h - v_h\|^{\mathrm{rel}}_{H^1(\Omega)} := \frac{\|u_h - v_h\|_{H^1(\Omega)}}{\|u_h\|_{H^1(\Omega)}}$$

for any $v_h \in V_h$. The coarse part ('the V_H-part') of an LOD approximation $u^{\text{G-LOD}}$ (respectively $u^{\text{PG-LOD}}$) is subsequently denoted by $P_{L^2}(u^{\text{G-LOD}})$ [respectively $P_{L^2}(u^{\text{PG-LOD}})$], where P_{L^2} denotes the L^2-projection on V_H (see also Remark 5).

We consider the following model problem. Let $\Omega :=]0,1[^2$ and $\varepsilon := 0.05$. Find $u_\varepsilon \in H^1(\Omega)$ with

$$-\nabla \cdot (A_\varepsilon(x)\nabla u_\varepsilon(x)) = x_1 - \frac{1}{2} \quad \text{in } \Omega$$
$$u_\varepsilon(x) = 0 \quad \text{on } \partial\Omega.$$

The scalar diffusion term A_ε is shown in Fig. 2. It is given by

$$A_\varepsilon(x) := (h \circ c_\varepsilon)(x) \quad \text{with } h(t) := \begin{cases} t^4 & \text{for } \frac{1}{2} < t < 1 \\ t^{\frac{3}{2}} & \text{for } 1 < t < \frac{3}{2} \\ t & \text{else} \end{cases} \tag{19}$$

and where

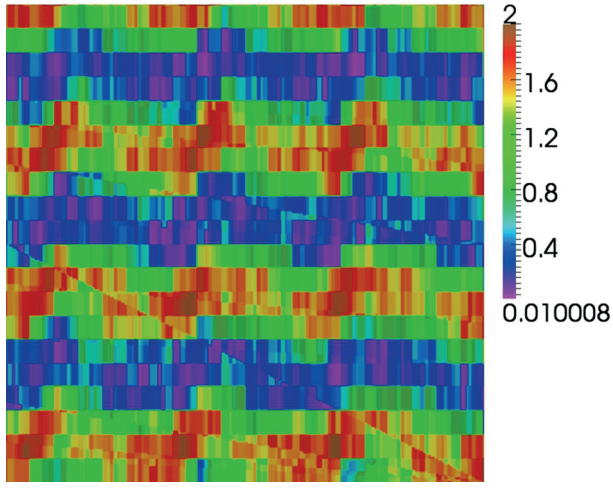


Fig. 2 Sketch of heterogeneous diffusion coefficient A_ε defined according to Eq. (19)

Table 1 Results for the errors between LOD approximations and reference solutions

H	k	$\|e_H\|^{\text{rel}}_{L^2(\Omega)}$	$\|e_h\|^{\text{rel}}_{L^2(\Omega)}$	$\|e_h\|^{\text{rel}}_{H^1(\Omega)}$	$\|e_H^{\text{PG}}\|^{\text{rel}}_{L^2(\Omega)}$	$\|e_h^{\text{PG}}\|^{\text{rel}}_{L^2(\Omega)}$	$\|e_h^{\text{PG}}\|^{\text{rel}}_{H^1(\Omega)}$
2^{-2}	0	0.3794	0.3772	0.6377	0.3778	0.3755	0.6375
2^{-2}	1/2	0.2756	0.2381	0.5312	0.2588	0.2269	0.5628
2^{-2}	1	0.2523	0.1445	0.3637	0.2544	0.1504	0.3642
2^{-2}	3/2	0.2514	0.1355	0.3125	0.2518	0.1380	0.3162
2^{-3}	0	0.2039	0.2037	0.5048	0.2037	0.2036	0.5048
2^{-3}	1	0.1100	0.0526	0.2278	0.1139	0.0619	0.2345
2^{-3}	2	0.1073	0.0423	0.1761	0.1078	0.0453	0.1807
2^{-3}	3	0.1070	0.0366	0.1567	0.1077	0.0399	0.1600
2^{-4}	0	0.0874	0.0873	0.3563	0.0874	0.0873	0.3563
2^{-4}	2	0.0353	0.0105	0.0932	0.0357	0.0123	0.0994
2^{-4}	4	0.0351	0.0082	0.0653	0.0353	0.0093	0.0680
2^{-4}	6	0.0351	0.0080	0.0634	0.0353	0.0091	0.0662

We define $e_h := u_h - u^{\text{G-LOD}}$ and $e_h^{\text{PG}} := u_h - u^{\text{PG-LOD}}$. Accordingly we define the errors between the reference solution and the coarse parts of the LOD approximations by $e_H := u_h - P_{L^2}(u^{\text{G-LOD}})$ (for the symmetric case) and $e_H^{\text{PG}} := u_h - P_{L^2}(u^{\text{PG-LOD}})$ (for the Petrov–Galerkin case). The reference solution u_h was obtained on a fine grid of mesh size $h = 2^{-6} \approx 0.0157 < \varepsilon$ which just resolves the micro structure of the coefficient A_ε. The number of 'coarse grid layers' is denoted by k and determines the patch size $U_k(T)$

$$c_\varepsilon(x_1, x_2) := 1 + \frac{1}{10}\sum_{j=0}^{4}\sum_{i=0}^{j}\left(\frac{2}{j+1}\cos\left(\left\lfloor i x_2 - \tfrac{x_1}{1+i}\right\rfloor + \left\lfloor \tfrac{i x_1}{\varepsilon}\right\rfloor + \left\lfloor \tfrac{x_2}{\varepsilon}\right\rfloor\right)\right).$$

The goal of the experiments is to investigate the accuracy of the PG-LOD, compared to the classical symmetric LOD. Moreover, we investigate the accuracy of the coarse part of the LOD approximation in terms of L^2-approximation properties (see Sect. 3.3 for a corresponding discussion).

In Table 1 we can see the results for a fine grid $\mathcal{T}_h$ with resolution $h = 2^{-6} < \varepsilon$ which just resolves the micro structure of the coefficient A_ε. Comparing the relative L^2- and H^1-errors for the G-LOD and the PG-LOD respectively (with the reference solution u_h), we observe that the errors are of similar size in each case. In general, we obtain slightly worse results for the PG-LOD, however the difference is so small that is does not justify the usage of the more memory-demanding (and more expensive) symmetric LOD. For both methods we observe the same nice error decay (in terms of the patch size) that was already predicted by the theoretical results. Comparing the relative L^2-errors between u_h and the coarse parts of the LOD-approximations, we observe that they already yield very good approximations. We also observe that they seem to be much more dominated by H-error contribution than by the θ^k-error contribution (i.e. the error coming from the decay). Using patches consisting of more than 8 fine element layers did not lead to any significant improvement, while there were still clear improvements visible for the other errors for the full G-LOD approximations.

Table 2 Results for the errors between LOD approximations and reference solutions

H	k	$\|e_H\|^{\text{rel}}_{L^2(\Omega)}$	$\|e_h\|^{\text{rel}}_{L^2(\Omega)}$	$\|e_h\|^{\text{rel}}_{H^1(\Omega)}$	$\|e_H^{\text{PG}}\|^{\text{rel}}_{L^2(\Omega)}$	$\|e_h^{\text{PG}}\|^{\text{rel}}_{L^2(\Omega)}$	$\|e_h^{\text{PG}}\|^{\text{rel}}_{H^1(\Omega)}$
2^{-2}	0	0.3840	0.3815	0.6434	0.3820	0.3796	0.6432
2^{-2}	1/8	0.2985	0.2781	0.5486	0.2957	0.2753	0.5513
2^{-2}	1/4	0.2852	0.2592	0.5578	0.2718	0.2472	0.5774
2^{-2}	1/2	0.2769	0.2392	0.5386	0.2607	0.2291	0.5722
2^{-2}	3/4	0.2676	0.2052	0.4784	0.2577	0.1972	0.4956
2^{-3}	0	0.2106	0.2103	0.5190	0.2103	0.2100	0.5190
2^{-3}	1/4	0.1480	0.1375	0.4510	0.1569	0.1469	0.4486
2^{-3}	1/2	0.1372	0.1163	0.3957	0.1305	0.1089	0.4029
2^{-3}	1	0.1138	0.0535	0.2308	0.1176	0.0628	0.2372
2^{-3}	3/2	0.1117	0.0399	0.1710	0.1126	0.0437	0.1761
2^{-4}	0	0.0988	0.0984	0.3854	0.0987	0.0983	0.3854
2^{-4}	1/2	0.0637	0.0592	0.2896	0.0500	0.0442	0.2934
2^{-4}	1	0.0406	0.0211	0.1613	0.0431	0.0263	0.1690
2^{-4}	2	0.0381	0.0109	0.0957	0.0385	0.0130	0.1017
2^{-4}	3	0.0380	0.0087	0.0726	0.0382	0.0099	0.0753

The errors are defined as in Table 1. The reference solution u_h was obtained on a fine grid of mesh size $h = 2^{-8} \approx 0.0039 \ll \varepsilon$ which fully resolves the micro structure of the coefficient A_ε. Again, the number of 'coarse grid layers' is denoted by k and determines the patch size $U_k(T)$

Furthermore, the linear convergence in H is clearly visible for $\|e_H\|^{\text{rel}}_{L^2(\Omega)}$ (respectively $\|e_H^{\text{PG}}\|^{\text{rel}}_{L^2(\Omega)}$) showing that the obtained error estimates seem to be indeed optimal.

The same observations can be made for the errors depicted in Table 2 for a fine grid $\mathcal{T}_h$ with resolution $h = 2^{-8} \ll \varepsilon$. Again, the results for the (symmetric) G-LOD are slightly better than the ones for the PG-LOD, but always of the same order. The exponential convergence in k for both realization is visualized in Fig. 3. It is clearly observable that there is no argument for using the G-LOD when dealing with patch communication issues which are storage demanding.

These findings are confirmed in the Figs. 4 and 5. In Fig. 4 we can see a visual comparison of the reference solution with the corresponding full LOD approximations (symmetric and Petrov–Galerkin). Both are almost not distinguishable for the investigated setting with $(h, H, k) = (2^{-8}, 2^{-4}, 2)$. Also the coarse parts of the LOD approximations already capture all the essential behavior of the reference solution. In Fig. 5 this is emphasized. Here, we compare the isolines between the reference solution and PG-LOD approximation (respectively its coarse part) and we observe that they are highly matching.

4.2 PG DG-LOD for the Buckley–Leverett equation

In this subsection we present the results of a two-phase flow simulation, based on solving the Buckley–Leverett equation as discussed in Sect. 3.6. Recall that, the Buckley–

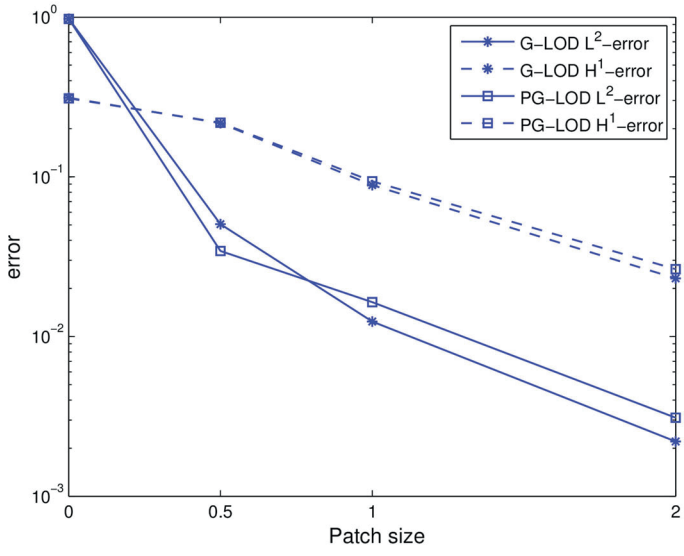


Fig. 3 The graphic visualizes the error decay in k. The results correspond to the results of Table 2 for $(h, H) = (2^{-8}, 2^{-4})$. We include $\|e_h\|^{\text{rel}}_{L^2(\Omega)}$, $\|e_h\|^{\text{rel}}_{H^1(\Omega)}$, $\|e_h^{\text{PG}}\|^{\text{rel}}_{L^2(\Omega)}$ and $\|e_h^{\text{PG}}\|^{\text{rel}}_{H^1(\Omega)}$. The x-axis depicts the localization parameter k and the y-axis the error "$\|e(k)\| - \|e(3)\|$" on the log-scale, where $\|e(k)\|$ denotes an error for k-layers (the error $\|e(3)\|$ is hence the limit reference)

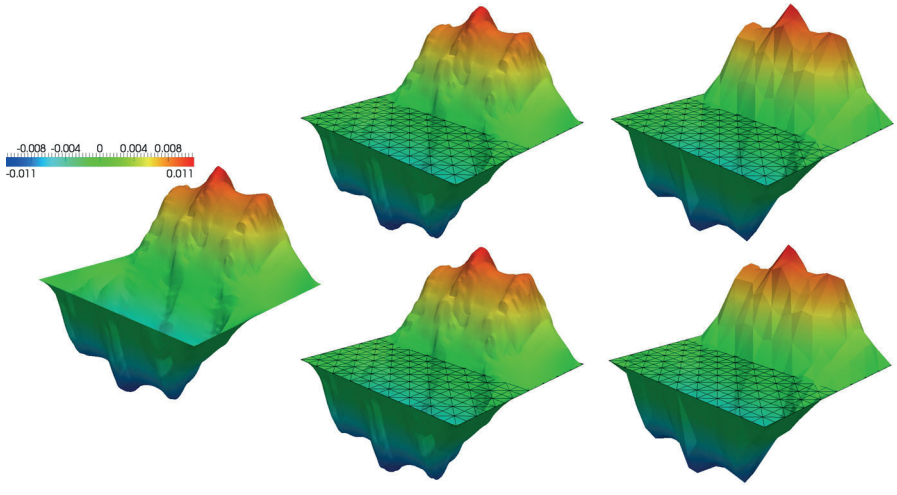


Fig. 4 The *left picture* shows the finite element reference solution u_h for $h = 2^{-8}$. The *remaining pictures* show LOD approximations for the case $(H, k) = (2^{-4}, 2)$, where k denotes the (*broken*) number of coarse layers. The two *top row pictures* show the full G-LOD approximation $u^{\text{G-LOD}}$ (*left*) and the coarse part of it, i.e. $P_{L^2}(u^{\text{G-LOD}})$ (*right*). The *bottom row* shows the full Petrov–Galerkin LOD approximation $u^{\text{PG-LOD}}$ (*left*) and the corresponding coarse part, i.e. $P_{L^2}(u^{\text{PG-LOD}})$ (*right*). The grid that is added to each of the pictures shows the coarse grid $\mathcal{T}_H$

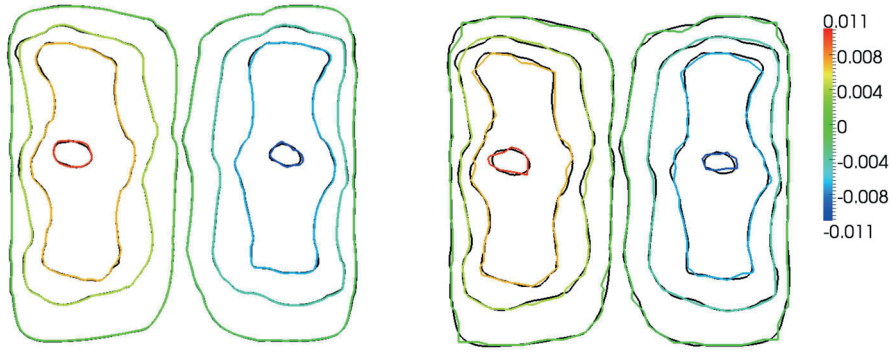


Fig. 5 The *pictures* depict a comparison of isolines. The *black lines* belong to the reference solution u_h for $h = 2^{-8}$. The *colored isolines in the left picture* belong to the PG-LOD approximation $u^{\text{PG-LOD}}$ and match almost perfectly with the one from the reference solution. The *right picture* shows the coarse part of $u^{\text{PG-LOD}}$, i.e. $P_{L^2}(u^{\text{PG-LOD}})$. We observe that the isolines still match nicely (color figure online)

Leverett equation has two parts, a hyperbolic equation for the saturation and a elliptic equation for the pressure. For that reason, we use the operator splitting technique IMPES, that we stated in Sect. 3.6. The elliptic pressure equation is solved by the PG DG-LOD for which a discontinuous linear finite element method is utilized that allows for recovering an elemental locally conservative normal flux. We emphasize that having a locally conservative flux is typically central for numerical schemes for solving hyperbolic partial differential equations. In this experiment we use an upwinding scheme.

Employing PG DG-LOD in this simulation proves to be a very efficient since the local correctors for the generalized basis functions only have to be computed once in a preprocessing step, this follows from the fact the saturation only influence the permeability on the macroscopic scale. The time stepping in the IMPES scheme using the PG DG-LOD for the is realized through Algorithm 2 below.

Set the end time T_{end}, number of update of the pressure n, number of explicit updates on each implicit step update m.

Algorithm 2: solveBuckleyLeverett($\mathcal{T}_H$, $\mathcal{T}_h$, T_{end}, n, m)

Set the initial values: $S = S_0$ and $i = 1$
Preprocessing step: Compute local corrections Q_h^T for all $T \in \mathcal{T}_H$ with $\lambda(S) = 1$
while $t \leq T_{end}$ **do**
  Compute pressure p using PG DG-LOD at $(t + T_{\text{end}}/(n))$
  Extract conservative flux **v**
  while $t \leq i\,T_{end}/n$ **do**
    Compute saturation S at $(t + T_{\text{end}}/(nm))$
    Update time: $t + T_{\text{end}}/(nm) \mapsto t$
  end
  $i + 1 \mapsto i$
end

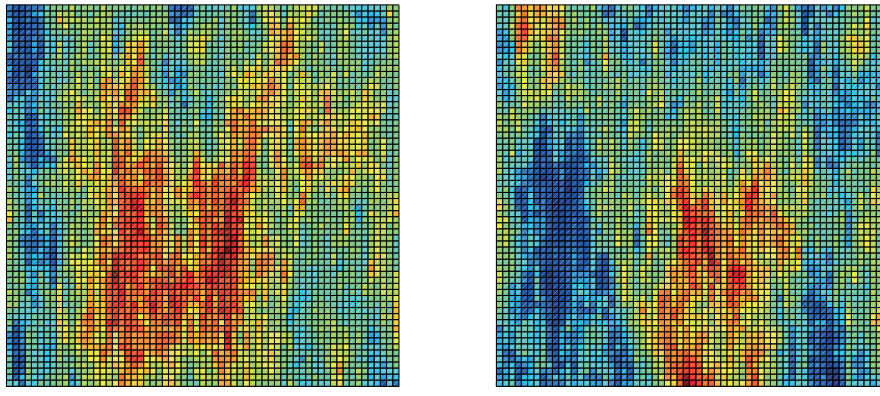


Fig. 6 The permeability structure of K_i in log scale with, $\beta_0/\alpha_0 \approx 5 \times 10^5$ for $i = 1$ (*left*) and $\beta_0/\alpha_0 \approx 4 \times 10^5$ for $i = 2$ (*right*)

Table 3 The resulting error in relative L^2-norm between S and S^{ref}, where S is obtained using PG DG-LOD for the pressure computed on $\mathcal{T}_H$ and S^{ref} is the reference solution computed on $\mathcal{T}_h$

Data	$\|e(T_1)\|_{L^2(\Omega)}$	$\|e(T_2)\|_{L^2(\Omega)}$	$\|e(T_3)\|_{L^2(\Omega)}$
K_1	0.088	0.073	0.070
K_2	0.058	0.087	0.079

We have $T_1 = 0.05$, $T_2 = 0.25$ and $T_3 = 0.45$

In the numerical experiment we consider the domain Ω to be the unit square. The permeability K_i for $i = 1, 2$ is given by layer 21 and 31 of the Society of Petroleum Engineering comparative permeability data (available on http://www.spe.org/web/csp), projected on a uniform mesh with resolution 2^{-6} as illustrated in Fig. 6.

We consider a microscopic partition $\mathcal{T}_h$ with mesh size size $h = 2^{-8}$ and a macroscopic partition $\mathcal{T}_H$ with mesh size $H = 2^{-i}$ for $i = 3, 4, 5, 6$. The patch size is chosen such that the overall H convergence for the PG DG-LOD is not effected. A reference solution to the Buckley–Leverett equation is obtained when both the pressure and saturation equation are computed on $\mathcal{T}_h$, compared to using Algorithm 2 where both the pressure and saturation equation are computed on $\mathcal{T}_H$. We consider the following setup. For the pressure equation we use the boundary condition $p = 1$ for the left boundary, $p = 0$ for the right boundary, $K\lambda(S)\nabla p = 0$ otherwise, and the source terms $q_w = q_n = 0$. For the saturation the initial value is $S = 1$ on the left boundary and 0 elsewhere. The error is defined by $e(\cdot, t) := S(\cdot, t) - S^{\text{rel}}(\cdot, t)$, where $S(\cdot, t)$ is the solution obtained by Algorithm 2 (at time t) and $S^{\text{rel}}(\cdot, t)$ is the reference solution (at time t). The errors are measured in the L^2-norm. In Table 3 we fix the coarse mesh size to be $H = 2^{-5}$, and compute the error for the permeabilities K_1 and K_2 at the times $T_1 := 0.05$, $T_2 := 0.25$ and $T_3 := 0.45$. A graphical comparison is shown in Figs. 7 and 8. The errors in the L^2-norm is less than 0.1 for both permeabilities at all times which is quite remarkable since the coarse mesh $\mathcal{T}_H$ for $H = 2^{-5}$ does not resolve the data. In Table 4 we consider the test case involving permeability K_1. We

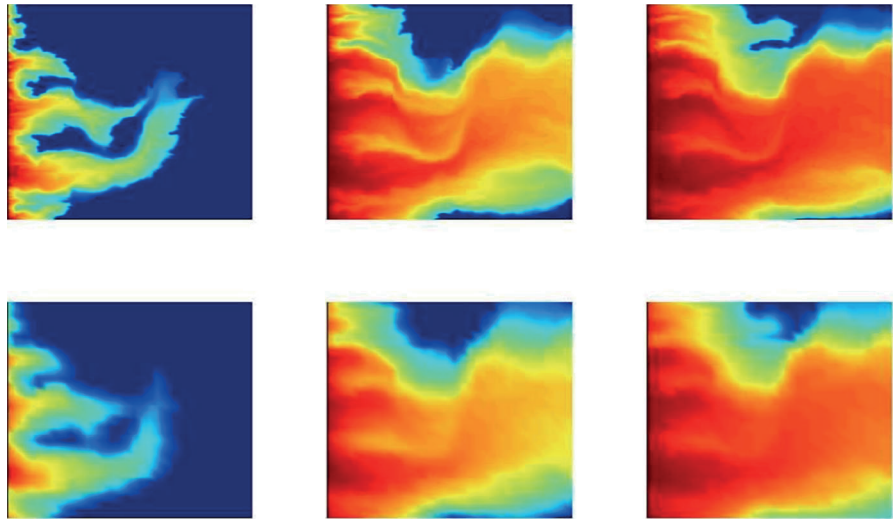


Fig. 7 The saturation profile using PG DG-LOD for the pressure equation on the grid $\mathcal{T}_H$ (*bottom*) and the reference solution on the grid $\mathcal{T}_h$ (*upper*) at time $T_1 = 0.05$ (*left*), $T_2 = 0.25$ (*middle*), and $T_3 = 0.45$ (*right*) using permeability K_1

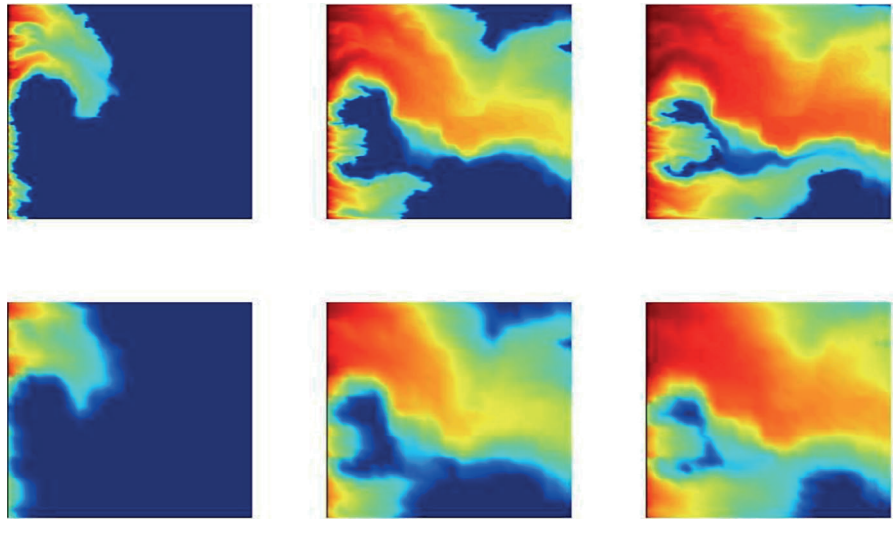


Fig. 8 The saturation profile using PG DG-LOD for the pressure equation on the grid $\mathcal{T}_H$ (*bottom*) and the reference solution on the grid $\mathcal{T}_h$ (*upper*) at time $T_1 = 0.05$ (*left*), $T_2 = 0.25$ (*middle*), and $T_3 = 0.45$ (*right*) using permeability K_2

present the L^2-errors at $t = T_2$ for different values of H. We basically observe a linear convergence rate in H/h (for fixed h) which is just what we would expect (since we only use the coarse part of the LOD pressure approximation).

Table 4 We consider the test case involving K_1

H	$\|e(T_2)\|_{L^2(\Omega)}$
2^{-3}	0.220
2^{-4}	0.113
2^{-5}	0.073
2^{-6}	0.048

The table depicts relative L^2-errors between S and S^{ref} at $T_2 = 0.25$ for different values of the coarse mesh size H. Here, S^{ref} denotes the reference solution computed on T_h with $h = 2^{-8}$ and S denotes the numerical approximation obtained with the IMPES scheme, using the PG DG-LOD for solving the pressure equation (with coarse mesh $\mathcal{T}_H$). We pick $k = \lceil 2|\log(H)|\rceil$

5 Proofs of the main results

In this proof section we will frequently exploit the estimate

$$\|v_h\|_{L^2(\Omega)} \lesssim |||v_h|||_h \quad \text{for all } v_h \in V_h, \tag{20}$$

which is a conclusion from assumption (A7). Let $I_H^{-1} := (I_H|_{V_H})^{-1}$, then (20) can be verified as follows by using (A7).

$$\begin{aligned}
\|v_h\|_{L^2(\Omega)} &\leq \|v_h - I_H(v_h)\|_{L^2(\Omega)} + \|I_H(v_h)\|_{L^2(\Omega)} \\
&\lesssim H|||v_h|||_h + \|(I_H \circ I_H^{-1} \circ I_H)(v_h)\|_{L^2(\Omega)} \\
&\lesssim H|||v_h|||_h + |||(I_H^{-1} \circ I_H)(v_h)|||_H \lesssim H|||v_h|||_h + |||I_H(v_h)|||_H \\
&\lesssim H|||v_h|||_h + |||v_h|||_h.
\end{aligned}$$

5.1 Proof of Theorem 1

The arguments for establishing the error estimate in $|||\cdot|||_h$-norm is analogous to the standard case, see for example [35] or [19]. We only recall the main arguments.

Proof (Proof of Theorem 1) Let $u_H^{\text{G-LOD}} = (u_H + Q_h(u_H)) \in V^{\text{ms}}$ be the Galerkin LOD solution governed by (10). Utilizing the notation in (A8), we set $u_{H,\Omega} \in V_H$ to satisfy

$$a_h(u_{H,\Omega} + Q_h^\Omega(u_{H,\Omega}), \Phi_H + Q_h^\Omega(\Phi_H)) = (f, \Phi_H + Q_h^\Omega(\Phi_H)) \quad \text{for all } \Phi_H \in V_H$$

and define $e_h := u_{H,\Omega} + Q_h^\Omega(u_{H,\Omega}) - u_h$. Using Galerkin orthogonality, we obtain $a_h(e_h, \Phi) = 0$ for all $\Phi \in V_\Omega^{\text{ms}}$ and hence $e_h \in W_h$ (i.e. $I_H(e_h) = 0$). This implies $|||e_h|||_h^2 \lesssim a_h(e_h, e_h) = (f, e_h) = (f, e_h - I_H(e_h)) \lesssim H\|f\|_{L^2(\Omega)} |||e_h|||_h$ and

consequently by energy minimization

$$\begin{aligned}|||u_H^{\text{G-LOD}} - u_h|||_h = |||u_H + Q_h(u_H) - u_h|||_h &\lesssim |||u_{H,\Omega} + Q_h(u_{H,\Omega}) - u_h|||_h \\ &\leq |||e_h|||_h + \left|\left|\left| Q_h^\Omega(u_{H,\Omega}) - Q_h(u_{H,\Omega})\right|\right|\right|_h \\ &\overset{\text{(A8)}}{\lesssim} H\|f\|_{L^2(\Omega)} + (1/H)^p k^{d/2}\theta^k \left|\left|\left| u_{H,\Omega} + Q_h^\Omega(u_{H,\Omega})\right|\right|\right|_h .\end{aligned}$$

The bound $|||u_{H,\Omega} + Q_h^\Omega(u_{H,\Omega})|||_h \lesssim \|f\|_{L^2(\Omega)}$ finishes the energy-norm estimate. The estimate in the L^2-norm is established in a similar fashion using (20). □

5.2 Proof of Theorem 2

We begin with stating and proving a lemma that is required to establish the a priori error estimate.

Lemma 3 *For all $v^{\text{ms}} \in V_\Omega^{\text{ms}}$ with $v^{\text{ms}} = v_H + v^f$, where $v_H \in V_H$ and $v^f \in W_h$, we have*

$$\|v^f\|_{L^2(\Omega)} \lesssim H|||v^{\text{ms}}|||_h. \tag{21}$$

Proof Because of $I_H(v^f) = 0$ and $(I_H^{-1} \circ I_H)(v_H) = v_H$,

$$v^f = v^f - I_H(v^f) + v_H - (I_H^{-1} \circ I_H)(v_H + v^f) + I_H(v_H + v^f) - I_H(v_H),$$

and therefore with $I_H = I_H \circ I_H^{-1} \circ I_H$ and (A7),

$$\begin{aligned}\|v^f\|_{L^2(\Omega)} &\leq \|v^{\text{ms}} - I_H(v^{\text{ms}})\|_{L^2(\Omega)} + \left\| \left(I_H^{-1} \circ I_H\right)(v^{\text{ms}}) - I_H(v^{\text{ms}})\right\|_{L^2(\Omega)} \\ &\lesssim H|||v^{\text{ms}}|||_h + \left\| \left(I_H^{-1} \circ I_H\right)(v^{\text{ms}}) - \left(I_H \circ I_H^{-1} \circ I_H\right)(v^{\text{ms}})\right\|_{L^2(\Omega)} \\ &\lesssim H|||v^{\text{ms}}|||_h + H\left|\left|\left| \left(I_H^{-1} \circ I_H\right)(v^{\text{ms}})\right|\right|\right|_H \\ &\lesssim H|||v^{\text{ms}}|||_h.\end{aligned}$$

In the last step we used again the stability estimates for I_H^{-1} and I_H in (A7). □

Proof (Proof of Theorem 2) Let $u_{H,\Omega}^{\text{G-LOD}}$ and $u_{H,\Omega}^{\text{PG-LOD}}$ be respectively the solution of (10) and (12) for $U(T) = \Omega$. As in the statement of the theorem, $u_H^{\text{PG-LOD}}$ is the solution of (12) for $U(T) = U_k(T)$. By adding and subtracting appropriate terms and applying triangle inequality, we arrive at

$$\left|\left|\left| u_h - u_H^{\text{PG-LOD}}\right|\right|\right|_h \leq \text{I}_1 + \text{I}_2 + \text{I}_3,$$

where we set $\text{I}_1 = |||u_h - u_{H,\Omega}^{\text{G-LOD}}|||_h$, $\text{I}_2 = |||u_{H,\Omega}^{\text{G-LOD}} - u_{H,\Omega}^{\text{PG-LOD}}|||_h$, and $\text{I}_3 = |||u_{H,\Omega}^{\text{PG-LOD}} - u_H^{\text{PG-LOD}}|||_h$. In the following, we estimate these three terms. Because

$e^{(1)} := (u_h - u_{H,\Omega}^{\text{G-LOD}}) \in W_h$ (cf. proof of Theorem 1) and by applying the Galerkin orthogonality, we get

$$\begin{aligned} \mathrm{I}_1^2 \lesssim a_h(e^{(1)}, e^{(1)}) &= a_h(u_h, e^{(1)}) \\ &= (f, e^{(1)} - I_H(e^{(1)})) \lesssim H\|f\|_{L^2(\Omega)} \quad |||e^{(1)}|||_h \leq H\|f\|_{L^2(\Omega)} \quad \mathrm{I}_1, \end{aligned} \tag{22}$$

i.e. $\mathrm{I}_1 \lesssim H\|f\|$. Furthermore, $e^{(2)} := (u_{H,\Omega}^{\text{PG-LOD}} - u_{H,\Omega}^{\text{G-LOD}}) \in V_\Omega^{\text{ms}}$ and the splitting $e^{(2)} = e_H^{(2)} + e_f^{(2)}$ with $e_H^{(2)} \in V_H$ and $e_f^{(2)} \in W_h$ (i.e. $I_H(e_f^{(2)}) = 0$) holds true. Because $a_h(u_{H,\Omega}^{\text{PG-LOD}}, e_f^{(2)}) = 0$, we obtain

$$\begin{aligned} \mathrm{I}_2^2 &\lesssim a_h(e^{(2)}, e^{(2)}) \\ &= a_h\left(u_{H,\Omega}^{\text{PG-LOD}}, e_H^{(2)}\right) - a_h\left(u_{H,\Omega}^{\text{G-LOD}}, e^{(2)}\right) = \left(f, e_H^{(2)} - e^{(2)}\right) = -\left(f, e_f^{(2)}\right), \end{aligned} \tag{23}$$

where $(f, e_f^{(2)}) \leq \|f\|_{L^2(\Omega)} \|e_f^{(2)}\|_{L^2(\Omega)} \lesssim \|f\|_{L^2(\Omega)} H|||e^{(2)}|||_h = H\|f\|_{L^2(\Omega)} \mathrm{I}_2$ by Lemma 3. Again, we conclude that $\mathrm{I}_2 \lesssim H\|f\|_{L^2(\Omega)}$. It remains to estimate I_3 for which we define $e^{(3)} := u_{H,\Omega}^{\text{PG-LOD}} - u_H^{\text{PG-LOD}}$. To simplify the notation, we subsequently denote (according to the definitions of V^{ms} and V_Ω^{ms})

$$u_H^{\text{PG-LOD}} = u_H + Q_h(u_H) \quad \text{and} \quad u_{H,\Omega}^{\text{PG-LOD}} = u_H^\Omega + Q_h^\Omega\left(u_H^\Omega\right),$$

where $u_H \in V_H$ and $u_H^\Omega \in V_H$. By the definition of problem (12) we have

$$a_h\left(u_H^{\text{PG-LOD}}, \Phi_H\right) = (f, \Phi_H) = a_h\left(u_{H,\Omega}^{\text{PG-LOD}}, \Phi_H\right). \tag{24}$$

On the other hand, by the definition of $Q_h^\Omega = -P_h$ (see Remark 1) and since $Q_h(\Phi_H) \in W_h$ we get

$$a_h\left(u_{H,\Omega}^{\text{PG-LOD}}, Q_h(\Phi_H)\right) = 0. \tag{25}$$

Combining (24) and (25) we get the equality

$$\begin{aligned} a_h\left(u_H^{\text{PG-LOD}}, \Phi_H + Q_h(\Phi_H)\right) = a_h&\left(u_H^{\text{PG-LOD}}, Q_h(\Phi_H)\right) \\ &+ a_h\left(u_{H,\Omega}^{\text{PG-LOD}}, \Phi_H + Q_h(\Phi_H)\right). \end{aligned}$$

We use this equality cast u_H as a unique solution of a self-adjoint variational equation expressed as

$$a_h(u_H + Q_h(u_H), \Phi_H + Q_h(\Phi_H)) = F_{u_H, u_H^\Omega}(\Phi_H) \quad \text{for all } \Phi_H \in V_H,$$

where F_{u_H,u_H^Ω} is a given fixed data function written as

$$F_{u_H,u_H^\Omega}(\Phi_H)=a_h(u_H+Q_h(u_H),\, Q_h(\Phi_H))+a_h\left(u_H^\Omega+Q_h^\Omega\left(u_H^\Omega\right),\, \Phi_H+Q_h(\Phi_H)\right).$$

Since this problem is self-adjoint, we get that u_H is equally the minimizer in V_H of the functional

$$\begin{aligned} J(\Phi_H) := a_h&\left(\Phi_H+Q_h(\Phi_H)-u_H^\Omega-Q_h^\Omega\left(u_H^\Omega\right),\, \Phi_H+Q_h(\Phi_H)-u_H^\Omega-Q_h^\Omega\left(u_H^\Omega\right)\right)\\ &-2a_h(u_H+Q_h(u_H),\, Q_h(\Phi_H)). \end{aligned}$$

Hence we obtain

$$\begin{aligned} \alpha \mathrm{I}_3^2 &= \alpha |||e^{(3)}|||_h^2\\ &\le a_h(e^{(3)}, e^{(3)})\\ &= J(u_H) + 2a_h(u_H + Q_h(u_H),\, Q_h(u_H))\\ &\le J\left(u_H^\Omega\right) + 2a_h(u_H + Q_h(u_H),\, Q_h(u_H))\\ &= a_h\left(Q_h\left(u_H^\Omega\right) - Q_h^\Omega\left(u_H^\Omega\right),\, Q_h\left(u_H^\Omega\right) - Q_h^\Omega\left(u_H^\Omega\right)\right)\\ &\quad - 2a_h\left(u_H + Q_h(u_H),\, Q_h(u_H) - Q_h\left(u_H^\Omega\right)\right)\\ &= \mathrm{I}_{31} + \mathrm{I}_{32}, \end{aligned} \tag{26}$$

where

$$\begin{aligned} \mathrm{I}_{31} &= a_h\left(Q_h\left(u_H^\Omega\right) - Q_h^\Omega\left(u_H^\Omega\right),\, Q_h\left(u_H^\Omega\right) - Q_h^\Omega\left(u_H^\Omega\right)\right)\\ \mathrm{I}_{32} &= a_h\left(Q_h(u_H) - Q_h^\Omega(u_H),\, Q_h(u_H) - Q_h\left(u_H^\Omega\right)\right). \end{aligned}$$

By the boundedness of $a_h(\cdot,\cdot)$ and applying (9) we get

$$\mathrm{I}_{31} \lesssim \left|||Q_h\left(u_H^\Omega\right) - Q_h^\Omega\left(u_H^\Omega\right)|||_h^2 \lesssim k^p\theta^{2k}(1/H)^{2p}\left|||u_H^\Omega + Q_h^\Omega\left(u_H^\Omega\right)|||_h^2. \tag{27}$$

We now need to estimate $u_{H,\Omega}^{\text{PG-LOD}} = u_H^\Omega + Q_h^\Omega(u_H^\Omega)$. By the inf-sup condition and Lemma 3,

$$\begin{aligned} \left|\left|\left|u_{H,\Omega}^{\text{PG-LOD}}\right|\right|\right|_h^2 &\lesssim a_h\left(u_{H,\Omega}^{\text{PG-LOD}}, u_{H,\Omega}^{\text{PG-LOD}}\right)\\ &= a\left(u_{H,\Omega}^{\text{PG-LOD}}, u_H^\Omega\right)\\ &= \left(f, u_H^\Omega\right)\\ &= \left(f, u_{H,\Omega}^{\text{PG-LOD}}\right) - \left(f,\, Q_h^\Omega(u_H^\Omega)\right)\\ &\lesssim (1+H)\|f\|_{L^2(\Omega)}\left|\left|\left|u_{H,\Omega}^{\text{PG-LOD}}\right|\right|\right|_h, \end{aligned} \tag{28}$$

and thus combining it with (27) yields

$$\mathrm{I}_{31} \lesssim k^d \theta^{2k} (1/H)^{2p} \|f\|^2_{L^2(\Omega)} \tag{29}$$

Furthermore, in a similar fashion we use the boundedness of $a_h(\cdot,\cdot)$ and (9) to get

$$\begin{aligned} \mathrm{I}_{32} &\lesssim \left|\left|\left| Q_h(u_H) - Q_h^\Omega(u_H) \right|\right|\right|_h \ \left|\left|\left| Q_h(u_H) - Q_h\left(u_H^\Omega\right) \right|\right|\right|_h \\ &\lesssim k^{d/2}\theta^k (1/H)^p \left|\left|\left| u_H^{\text{PG-LOD}} \right|\right|\right|_h \ \left|\left|\left| Q_h(u_H) - Q_h\left(u_H^\Omega\right) \right|\right|\right|_h \end{aligned} \tag{30}$$

By adding and subtracting appropriate terms and applying triangle inequality

$$\begin{aligned} &\left|\left|\left| Q_h(u_H) - Q_h\left(u_H^\Omega\right) \right|\right|\right|_h \\ &\quad \le \left|\left|\left| \left(Q_h - Q_h^\Omega\right)(u_H) \right|\right|\right|_h + \left|\left|\left| Q_h^\Omega\left(u_H - u_H^\Omega\right) \right|\right|\right|_h + \left|\left|\left| \left(Q_h^\Omega - Q_h\right)\left(u_H^\Omega\right) \right|\right|\right|_h . \end{aligned} \tag{31}$$

We use (9) to estimate the first and last terms in (31) to yield

$$\begin{aligned} &\left|\left|\left| \left(Q_h - Q_h^\Omega\right)(u_H) \right|\right|\right|_h + \left|\left|\left| \left(Q_h^\Omega - Q_h\right)\left(u_H^\Omega\right) \right|\right|\right|_h \\ &\quad \lesssim k^{d/2}\theta^k (1/H)^p \left(\left|\left|\left| u_H^{\text{PG-LOD}} \right|\right|\right|_h + \left|\left|\left| u_{H,\Omega}^{\text{PG-LOD}} \right|\right|\right|_h \right). \end{aligned} \tag{32}$$

Moreover, by the $|||\cdot|||_h$-stability of Q_h^Ω [which holds true since $Q_h^\Omega = -P_h$ with P_h being the orthogonal projection defined in (4)], we have

$$\begin{aligned} \left|\left|\left| Q_h^\Omega\left(u_H - u_H^\Omega\right) \right|\right|\right|_h &\lesssim \left|\left|\left| u_H - u_H^\Omega \right|\right|\right|_h \\ &= \left|\left|\left| \left(\left(I_H|_{V_H}\right)^{-1} \circ I_H \right)\left(e^{(3)}\right) \right|\right|\right|_h \lesssim C_{H,h} |||e^{(3)}|||_h. \end{aligned} \tag{33}$$

Putting back (33) and (32) to (31) and place it in (30) gives

$$\begin{aligned} \mathrm{I}_{32} &\lesssim k^d \theta^{2k} (1/H)^{2p} \left|\left|\left| u_H^{\text{PG-LOD}} \right|\right|\right|_h \left(\left|\left|\left| u_H^{\text{PG-LOD}} \right|\right|\right|_h + \left|\left|\left| u_{H,\Omega}^{\text{PG-LOD}} \right|\right|\right|_h \right) \\ &\quad + k^{d/2}\theta^k (1/H)^p \left|\left|\left| u_H^{\text{PG-LOD}} \right|\right|\right|_h \ C_{H,h} |||e^{(3)}|||_h \\ &\lesssim k^d \theta^{2k} (1/H)^{2p} \left(\left|\left|\left| u_H^{\text{PG-LOD}} \right|\right|\right|_h^2 + \left|\left|\left| u_{H,\Omega}^{\text{PG-LOD}} \right|\right|\right|_h^2 \right) \\ &\quad + \frac{C_{H,h}^2}{\delta} k^d \theta^{2k} (1/H)^{2p} \left|\left|\left| u_H^{\text{PG-LOD}} \right|\right|\right|_h^2 + \frac{\delta}{4} |||e^{(3)}|||_h^2, \end{aligned} \tag{34}$$

where in the last step we use the Young's inequality for both terms, and in particular for the second term, inserting a sufficiently small $\delta > 0$ so that we can later on hide the term $\frac{\delta}{4}|||e^{(3)}|||_h^2$ in the left hand side of (26). Note that the choice of δ is independent

of H, h or k. Rearranging and collecting common terms in the last inequality gives

$$\mathrm{I}_{32} \lesssim k^d \theta^{2k} (1/H)^{2p} \left(\left(1 + \frac{C_{H,h}^2}{\delta} \right) \left| \left| \left| u_H^{\text{PG-LOD}} \right| \right| \right|^2 + \left| \left| \left| u_{H,\Omega}^{\text{PG-LOD}} \right| \right| \right|^2 \right) + \frac{\delta}{4} |||e^{(3)}|||_h^2,$$

so that we need to estimate $|||u_H^{\text{PG-LOD}}|||_h$ and $|||u_{H,\Omega}^{\text{PG-LOD}}|||_h$, respectively. The stability of the second piece was established in (28), while the stability of the first piece is achieved by employing (A9) and (A7) in

$$\bar{\alpha} \left| \left| \left| u_H^{\text{PG-LOD}} \right| \right| \right|_h |||u_H|||_H \lesssim a_h \left(u_H^{\text{PG-LOD}}, u_H \right) = (f, u_H) \lesssim \|f\|_{L^2(\Omega)} |||u_H|||_H.$$

From which we conclude that

$$\mathrm{I}_{32} \lesssim k^d \theta^{2k} (1/H)^{2p} \left(\left(1 + \frac{C_{H,h}^2}{\delta} \right) (1 + \bar{\alpha}^{-1}) \|f\|^2 \right) + \frac{\delta}{4} \mathrm{I}_3^2.$$

To summarize, putting this last inequality and (29)–(26) and choosing sufficiently small δ gives

$$\mathrm{I}_3 \lesssim k^{d/2} \theta^k (1/H)^p \left(\left(1 + \frac{C_{H,h}}{\delta} \right) (1 + \bar{\alpha}^{-1}) \|f\| \right),$$

combining it with the existing estimates for I_1 and I_2 proves the error estimate in $|||\cdot|||_h$. Moreover, the estimate in the L^2-norm is established in a similar fashion. This completes the proof of the theorem. □

5.3 Proof of Lemmas 1 and 2

Next, we prove the inf-sup stability of the continuous Galerkin LOD in Petrov–Galerkin formulation.

Proof (Proof of Lemma 1) Let $\Phi^{\text{ms}} \in V^{\text{ms}}$ be an arbitrary element. To prove the inf-sup condition, we aim to show that

$$\frac{a_h(\Phi^{\text{ms}}, \Phi_H)}{|||\Phi_H|||_h} \geq \alpha(k) |||\Phi^{\text{ms}}|||_h \quad \text{for } \Phi_H = \left((I_H|_{V_H})^{-1} \circ I_H \right) (\Phi^{\text{ms}}). \tag{35}$$

Let therefore $U(T) = U_k(T)$ for fixed $k \in \mathbb{N}$. By the definitions of V^{ms} and Φ_H, we have $\Phi^{\text{ms}} = \Phi_H + Q_h(\Phi_H)$, where $Q_h(\Phi_H)$ denotes the corresponding corrector given by (7). By $Q_h^{\Omega}(\Phi_H)$ we denote the corresponding global corrector for the case $U(T) = \Omega$ and the local correctors are denoted by $Q_h^{\Omega,T}(\Phi_H)$. First, we observe that by $|||\cdot|||_h = |||\cdot|||_H$

$$|||\Phi_H|||_h = \left| \left| \left| \left((I_H|_{V_H})^{-1} \circ I_H \right) (\Phi^{\text{ms}}) \right| \right| \right|_h \lesssim |||\Phi^{\text{ms}}|||_h, \tag{36}$$

where we used the $|||\cdot|||_h$-stability of I_H and $(I_H|_{V_H})^{-1}$ according to (A7). Consequently, Eq. (36) implies

$$|||Q_h(\Phi_H)|||_h \leq |||\Phi^{\text{ms}}|||_h + |||\Phi_H|||_h \lesssim |||\Phi^{\text{ms}}|||_h, \tag{37}$$

and thus

$$\begin{aligned} a_h(\Phi^{\text{ms}}, \Phi_H) &= a_h(\Phi^{\text{ms}}, \Phi^{\text{ms}}) - a_h\left(\Phi^{\text{ms}}, Q_h(\Phi_H)\right) \\ &\geq \alpha|||\Phi^{\text{ms}}|||_h^2 - a_h\left(\Phi^{\text{ms}}, Q_h(\Phi_H)\right) \\ &\geq C\alpha|||\Phi_H|||_h\,|||\Phi^{\text{ms}}|||_h - a_h\left(\Phi^{\text{ms}}, Q_h(\Phi_H)\right), \end{aligned} \tag{38}$$

where we have used (36) again to bound $|||\Phi^{\text{ms}}|||_h$ from below. Note here that C denotes a generic constant. It remains to bound $a_h(\Phi^{\text{ms}}, Q_h(\Phi_H))$. By the orthogonality of V_Ω^{ms} and W_h we have

$$a_h\left(\Phi_H + Q_h^\Omega(\Phi_H), Q_h(\Phi_H)\right) = 0, \tag{39}$$

and since $a_h(\cdot,\cdot)$ is such that $a_h(v_h, w_h) = 0$ for all $v_h, w_h \in V_h$ with the property $\text{supp}(v_h)\cap\text{supp}(w_h) = \emptyset$ we get by the definition of $Q_h(\Phi_H)$ for every $w_h^T \in W_h(T)$

$$\begin{aligned} a_h\left(\Phi_H + Q_h(\Phi_H), w_h^T\right) &= \sum_{K\in\mathcal{T}_H}\left(a_h^K\left(\Phi_H, w_h^T\right) + a_h\left(Q_h(\Phi_H), w_h^T\right)\right) \\ &= \left(\sum_{K\in\mathcal{T}_H} a_h^K\left(\Phi_H, w_h^T\right)\right) + a_h\left(Q_h^T(\Phi_H), w_h^T\right) \\ &= a_h\left(\Phi_H + Q_h^T(\Phi_H), w_h^T\right) \\ &= 0. \end{aligned} \tag{40}$$

Using both equalities above and by the boundedness of $a_h(\cdot,\cdot)$ and applying (37) yields

$$\begin{aligned} &a_h\left(\Phi^{\text{ms}}, Q_h(\Phi_H)\right) \\ &\quad= a_h\left(\Phi_H + Q_h^\Omega(\Phi_H), Q_h(\Phi_H)\right) + a_h\left(Q_h(\Phi_H) - Q_h^\Omega(\Phi_H), Q_h(\Phi_H)\right) \\ &\quad= a_h\left(Q_h(\Phi_H) - Q_h^\Omega(\Phi_H), Q_h(\Phi_H) - w_h\right) \\ &\quad\leq \left|\left|\left|Q_h(\Phi_H) - Q_h^\Omega(\Phi_H)\right|\right|\right|_h \frac{|||Q_h(\Phi_H) - w_h|||_h}{|||Q_h(\Phi_H)|||_h}|||\Phi^{\text{ms}}|||_h. \end{aligned} \tag{41}$$

We next estimate $|||Q_h(\Phi_H) - Q_h^\Omega(\Phi_H)|||_h$ by applying (14) and establishing an analog of (15) for $Q_h^{\Omega,T}(\Phi_H)$ expressed as

$$\left|\left|\left|Q_h^{\Omega,T}(\Phi_H)\right|\right|\right|_h^2 \lesssim |||\Phi_H|||_{h,T}\left|\left|\left|Q_h^{\Omega,T}(\Phi_H)\right|\right|\right|_h, \tag{42}$$

giving (for $k > 0$)

$$\begin{aligned}\left|\left|\left|Q_h(\Phi_H) - Q_h^{\Omega}(\Phi_H)\right|\right|\right|_h &\lesssim k^{d/2}\theta^k \left(\sum_{T\in\mathcal{T}_H} \left|\left|\left|Q_h^{\Omega,T}(\Phi_H)\right|\right|\right|_h^2\right)^{1/2} \\ &\lesssim k^{d/2}\theta^k \left(\sum_{T\in\mathcal{T}_H} |||\Phi_H|||_{h,T}^2\right)^{1/2} \\ &\lesssim k^{d/2}\theta^k |||\Phi_H|||_h. \end{aligned} \tag{43}$$

Thus we end up with

$$a_h(\Phi^{\mathrm{ms}}, Q_h(\Phi_H)) \lesssim \left(\frac{|||Q_h(\Phi_H) - w_h|||_h}{|||Q_h(\Phi_H)|||_h}\right) k^{d/2}\theta^k |||\Phi_H|||_h \, |||\Phi^{\mathrm{ms}}|||_h, \tag{44}$$

which when combined with (38) implies that there exist positive generic constants C_1, C_2 (independent of H and k) such that

$$\frac{a_h(\Phi^{\mathrm{ms}}, \Phi_H)}{|||\Phi_H|||_h \;\; |||\Phi^{\mathrm{ms}}|||_h} \geq C_1\alpha - C_2 k^{d/2}\theta^k \inf_{w_h\in W_h^T} \frac{|||Q_h(\Phi_H) - w_h|||_h}{|||Q_h(\Phi_H)|||_h}. \tag{45}$$

Since $\inf_{w_h\in W_h^T} \frac{|||Q_h(\Phi_H)-w_h|||_h}{|||Q_h(\Phi_H)|||_h} = 0$ for $k = 0$, estimate (45) holds for all $k \in \mathbb{N}$ and the condition $k > 0$ is not required. The relation $Q_h(\Phi_H) = \Phi^{\mathrm{ms}} - ((I_H|_{V_H})^{-1} \circ I_H)(\Phi^{\mathrm{ms}})$ finishes the proof. □

Finally, we prove the inf-sup stability of the discontinuous Galerkin LOD in Petrov–Galerkin formulation.

Proof (Proof of Lemma 2) The main arguments are similar as in the proof of Lemma 1. Set $n := (m + 3)/2$. Let $\Phi^{\mathrm{ms}} = \Phi_H + Q_h(\Phi_H) \in V^{\mathrm{ms}}$ be an arbitrary element and let $U(T) = U_k(T)$ for fixed $k \gtrsim n|\log(H)|$. By the assumptions on $\mathcal{T}_H$ and $\mathcal{T}_h$ and by the definitions of $|||\cdot|||_h$ and $|||\cdot|||_H$ it is easy to see that

$$|||\Phi_H|||_h \lesssim H^{(1-m)/2}|||\Phi_H|||_H \quad \text{and} \quad |||\Phi_H|||_H \lesssim |||\Phi^{\mathrm{ms}}|||_h.$$

Consequently we get

$$|||Q_h(\Phi_H)|||_h \leq |||\Phi^{\mathrm{ms}}|||_h + |||\Phi_H|||_h \lesssim (1 + H^{(1-m)/2})|||\Phi^{\mathrm{ms}}|||_h. \tag{46}$$

Thus

$$\begin{aligned} a_h\left(\Phi^{\mathrm{ms}}, \Phi_H\right) &= a_h(\Phi^{\mathrm{ms}}, \Phi^{\mathrm{ms}}) - a_h\left(\Phi^{\mathrm{ms}}, Q_h(\Phi_H)\right) \\ &\geq \alpha |||\Phi^{\mathrm{ms}}|||_h^2 - a_h\left(\Phi^{\mathrm{ms}}, Q_h(\Phi_H)\right) \\ &= \alpha |||\Phi^{\mathrm{ms}}|||_h^2 - a_h\left(Q_h(\Phi_H) - Q_h^\Omega(\Phi_H), Q_h(\Phi_H)\right) \\ &\geq \alpha |||\Phi^{\mathrm{ms}}|||_h^2 - \left|||Q_h(\Phi_H) - Q_h^\Omega(\Phi_H)\right|||_h \; |||Q_h(\Phi_H)|||_h \\ &\overset{(28)}{\geq} \alpha |||\Phi^{\mathrm{ms}}|||_h^2 - \left|||Q_h(\Phi_H) - Q_h^\Omega(\Phi_H)\right|||_h \; (1 + H^{(1-m)/2}) |||\Phi^{\mathrm{ms}}|||_h. \end{aligned} \tag{47}$$

Using

$$\begin{aligned} \left|||Q_h(\Phi_H) - Q_h^\Omega(\Phi_H)\right|||_h &\leq C(1/H) k^{d/2} \theta^k \left|||\Phi_H + Q_h^\Omega(\Phi_H)\right|||_h \\ &\leq C(1/H) k^{d/2} \theta^k \left(|||\Phi^{\mathrm{ms}}|||_h + \left|||Q_h(\Phi_H) - Q_h^\Omega(\Phi_H)\right|||_h\right) \\ &\leq C H^{n-1} \left(|||\Phi^{\mathrm{ms}}|||_h + \left|||Q_h(\Phi_H) - Q_h^\Omega(\Phi_H)\right|||_h\right) \end{aligned}$$

we obtain that we have for small enough H

$$\left|||Q_h(\Phi_H) - Q_h^\Omega(\Phi_H)\right|||_h \lesssim H^{n-1} |||\Phi^{\mathrm{ms}}|||_h.$$

Inserting this into (47) gives us

$$\begin{aligned} a_h\left(\Phi^{\mathrm{ms}}, \Phi_H\right) &\geq \alpha |||\Phi^{\mathrm{ms}}|||_h^2 - (1 + H^{(1-m)/2}) H^{n-1} |||\Phi^{\mathrm{ms}}|||_h^2 \\ &\geq C_1(\alpha - C_2 H) |||\Phi^{\mathrm{ms}}|||_h^2. \end{aligned}$$

If H is small enough so that $(\alpha - C_2 H)$ is positive, the stability estimate $|||\Phi_H|||_H \lesssim |||\Phi^{\mathrm{ms}}|||_h$ concludes the inf-sup estimate. □

Acknowledgments We would like to thank the anonymous referees for their valuable comments and their constructive feedback on the original manuscript which helped us to improve this article.

References

1. Abdulle, A.: On a priori error analysis of fully discrete heterogeneous multiscale FEM. Multiscale Model. Simul. **4**(2), 447–459 (2005)
2. Abdulle, A., E,W., Engquist, B., Vanden-Eijnden, E.: The heterogeneous multiscale method. Acta Numer. **21**, 1–87 (2012)
3. Allaire, G.: Homogenization and two-scale convergence. SIAM J. Math. Anal. **23**(6), 1482–1518 (1992)
4. Aziz, K., Settari, A.: Petroleum Reservoir Simulation. Applied Science Publishers, London (1997)
5. Babuska, I., Lipton, R.: Optimal local approximation spaces for generalized finite element methods with application to multiscale problems. Multiscale Model. Simul. **9**(1), 373–406 (2011)
6. Bank, R.E., Dupont, T.: An optimal order process for solving finite element equations. Math. Comput. **36**(153), 35–51 (1981)
7. Bank, R.E., Yserentant, H.: On the H^1-stability of the L_2-projection onto finite element spaces. Numer. Math. **126**(2), 361–381 (2014)

8. Bush, L., Ginting, V., Presho, M.: Application of a conservative, generalized multiscale finite element method to flow models. J. Comput. Appl. Math. **260**, 395–409 (2014)
9. Carstensen, C.: Quasi-interpolation and a posteriori error analysis in finite element methods. M2AN. Math. Model. Numer. Anal. **33**(6), 1187–1202 (1999)
10. 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)
11. Efendiev, Y., Ginting, V., Hou, T., Ewing, R.: Accurate multiscale finite element methods for two-phase flow simulations. J. Comput. Phys. **220**(1), 155–174 (2006)
12. Efendiev, Y., Hou, T., Ginting, V.: Multiscale finite element methods for nonlinear problems and their applications. Commun. Math. Sci. **2**(4), 553–589 (2004)
13. Elfverson, D., Georgoulis, E.H., Målqvist, A.: An adaptive discontinuous Galerkin multiscale method for elliptic problems. Multiscale Model. Simul. **11**(3), 747–765 (2013)
14. Elfverson, D., Georgoulis, E.H., Målqvist, A., Peterseim, D.: Convergence of a discontinuous Galerkin multiscale method. SIAM J. Numer. Anal. **51**(6), 3351–3372 (2013)
15. Gaspoz, F.D., Heine, C.-J., Siebert, K.G.: Optimal grading of the newest vertex bisection and H^1-stability of the L^2-projection. SimTech Universität, Stuttgart (2014)
16. Ginting, V.: Analysis of two-scale finite volume element method for elliptic problem. J. Numer. Math. **12**(2), 119–141 (2004)
17. Gloria, A.: An analytical framework for the numerical homogenization of monotone elliptic operators and quasiconvex energies. Multiscale Model. Simul. **5**(3), 996–1043 (2006)
18. Gloria, A.: Reduction of the resonance error-part 1: approximation of homogenized coefficients. Math. Models Methods Appl. Sci. **21**(8), 1601–1630 (2011)
19. Henning, P., Målqvist, A.: Localized orthogonal decomposition techniques for boundary value problems. SIAM J. Sci. Comput. **36**(4), A1609–A1634 (2014)
20. Henning, P., Målqvist, A., Peterseim, D.: A localized orthogonal decomposition method for semi-linear elliptic problems. ESAIM Math. Model. Numer. Anal. **48**(5), 1331–1349 (2014)
21. 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)
22. Henning, P., Morgenstern, P., Peterseim, D.: Multiscale partition of unity. In: Meshfree Methods for Partial Differential Equations VII. Lecture Notes in Computational Science and Engineering, vol. 100. Springer, Berlin (2015)
23. Henning, P., Ohlberger, M.: The heterogeneous multiscale finite element method for elliptic homogenization problems in perforated domains. Numer. Math. **113**(4), 601–629 (2009)
24. Henning, P., Peterseim, D.: Oversampling for the multiscale finite element method. Multiscale Model. Simul. **11**(4), 1149–1175 (2013)
25. Hou, T.Y., Wu, X.-H.: A multiscale finite element method for elliptic problems in composite materials and porous media. J. Comput. Phys. **134**(1), 169–189 (1997)
26. Hou, T.Y., Wu, X.-H., Zhang, Y.: Removing the cell resonance error in the multiscale finite element method via a Petrov–Galerkin formulation. Commun. Math. Sci. **2**(2), 185–205 (2004)
27. Hughes, T.J.R.: Multiscale phenomena: Green's functions, the Dirichlet-to-Neumann formulation, subgrid scale models, bubbles and the origins of stabilized methods. Comput. Methods Appl. Mech. Eng. **127**(1–4), 387–401 (1995)
28. 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)
29. Karkulik, M., Pfeiler, C.-M., Praetorius, D.: L^2-orthogonal projections onto finite elements on locally refined meshes are H^1-stable (2013). arXiv:1307.0917
30. Kröner, D.: Numerical schemes for conservation laws. In: Wiley-Teubner Series Advances in Numerical Mathematics. Wiley, Chichester (1997)
31. Larson, M.G., Målqvist, A.: Adaptive variational multiscale ethods based on a posteriori error estimation: duality techniques for elliptic problems. In: Multiscale Methods in Science and Engineering. Lecture Notes in Computational Science and Engineering, vol. 44, pp. 181–193. Springer, Berlin (2005)
32. LeFloch, P.G.: Hyperbolic systems of conservation laws. In: Lectures in Mathematics ETH Zürich. Birkhäuser Verlag, Basel (2002). (The theory of classical and nonclassical shock waves)
33. Målqvist, A.: Multiscale methods for elliptic problems. Multiscale Model. Simul. **9**(3), 1064–1086 (2011)
34. Målqvist, A., Peterseim, D.: Computation of eigenvalues by numerical upscaling. Numer. Math. (2014). doi:10.1007/s00211-014-0665-6

35. Målqvist, A., Peterseim, D.: Localization of elliptic multiscale problems. Math. Comput. **83**(290), 2583–2603 (2014)
36. Ohlberger, M.: A posteriori error estimates for the heterogeneous multiscale finite element method for elliptic homogenization problems. Multiscale Model. Simul. **4**(1), 88–114 (2005)
37. Owhadi, H., Zhang, L.: Localized bases for finite-dimensional homogenization approximations with nonseparated scales and high contrast. Multiscale Model. Simul. **9**(4), 1373–1398 (2011)
38. Owhadi, H., Zhang, L., Berlyand, L.: Polyharmonic homogenization, rough polyharmonic splines and sparse super-localization. ESAIM Math. Model. Numer. Anal. **48**(2), 517–552 (2014)
39. van der Vorst, H.A.: Computational methods for large eigenvalue problems. In: Handbook of Numerical Analysis, vol. VIII, pp. 3–179. North-Holland, Amsterdam (2002)
40. Weinan, E., Engquist, B.: The heterogeneous multiscale methods. Commun. Math. Sci. **1**(1), 87–132 (2003)

Paper V

Multiscale methods for problems with complex geometry

Daniel Elfverson*, Mats G. Larson†, and Axel Målqvist‡

September 14, 2015

Abstract

In this paper we extend the multiscale analysis to elliptic problems on complex domains, e.g. domains with cracks or complicated boundary. We construct corrected coarse test and trail spaces which takes the fine scale features of the domain into account. The corrections only needs to be computed in regions effected by the fine scale geometrical information of the domain. We achieve linear convergence rate in energy norm for the multiscale solution. Moreover, the conditioning of the multiscale method is not affected by how the domain boundary cuts the coarse elements in the background mesh. The analytical findings are verified in a series of numerical experiments.

1 Introduction

Partial differential equations with data varying on multiple scales in space and time, so called *multiscale problems*, appear in many areas of science and engineering. Two of the most prominent examples are composite materials and flow in a porous medium. Standard numerical techniques may perform arbitrarily badly for multiscale problems, since the convergence rely on smoothness of the solution, see [3]. Also adaptive techniques [21], where local singularities are resolved by local mesh refinement fail for multiscale problems since the roughness of the data is often not localized in space. As a remedy against this issue generalized finite element methods [2] and other related multiscale techniques [12, 13, 11, 6, 13, 14, 15, 17] have been developed. So far these techniques have focused on multiscale coefficients in general and multiscale diffusion in particular. Significantly less work within the multiscale community has been directed towards handling a computational domain with multiscale boundary. However, in many applications including voids and cracks in materials and rough surfaces, multiscale behavior emanates from the complex geometry of the computational domain. Furthermore, the classical multiscale methods mentioned above aim at, in different ways, upscaling the multiscale data to a coarse scale where the equation is possible to solve at a reasonable computational cost. However, these techniques typically assume that the representation of the computational domain is the same on the coarse and fine scale. In practice this is very difficult to achieve unless the computational domain has a very simple shape, which is not the case in many practical applications.

Other techniques for handling complex geometry include e.g. the extended finite element method (XFEM) [10] and the cut finite element method (CutFEM) [5]. In XFEM the polynomial approximation space is enriched with non-polynomial functions and CutFEM uses a robust Nitsche's formulation to weakly enforce the boundary/interface conditions on so called cut elements that have more general shape compared to standard elements. However, none of these approaches handles boundaries which varies on a smaller scale then the computational mesh.

*Department of Information Technology, Uppsala University, Box 337, SE-751 05 Uppsala, Sweden.

†Department of Mathematics, Umeå University, SE-901 87 Umeå, Sweden. Supported by SSF.

‡Department of Mathematical Sciences, Chalmers University of Technology and University of Gothenburg SE-412 96 Göoteborg, Sweden. Supported by the Swedish research council.

In this paper we consider the problem of designing a multiscale method for problems with complex computational domain. In order to simplify the presentation we will neglect multiscale coefficients in the analysis even though the methodology directly extends to this situation. The proposed algorithm is based on the localized orthogonal decomposition (LOD) technique presented in [15] and further developed in [7, 8, 16, 19]. In LOD both test and trail spaces are decomposed into a multiscale space and a reminder space that are orthogonal with respect to the scalar product induced by the bilinear form appearing in the weak form of the equation. In this paper we propose and analyze how the modified multiscale basis functions can be blended with standard finite element basis functions, allowing them to be used only close to the complex boundary. We prove optimal convergence and show that the condition number of the resulting coarse system of equations scales at an optimal rate with the mesh sizes.

The gain of this approach is that the global solution is computed on the coarse scale, with the accuracy of the fine scale. Also, all localized fine scale computations needed to enrich the standard finite element basis are localized and can thus be done in parallel.

The outline of the paper is as follows. In Section 2 we present the model problem and introduce some notation. In Section 3 we formulate a multiscale method for problems where the mesh does not resolve the boundary. In Section 4 we analyse the proposed method in several different steps and finally prove a bound of the error in energy norm, which shows that the error is of the same order as the standard finite element method on the coarse mesh for a smooth problem. In Section 5 we shortly describe the implementation of the method and we prove a bound of the condition number of the stiffness matrix, which is optimal compared to standard finite elements on the coarse scale. Finally, in Section 6 we present some numerical experiment to verify the convergence rate and conditioning of the proposed method.

2 Preliminaries

In this section we present a model problem, introduce some notation, and define a reference finite element discretization of the model problem.

2.1 Model problem

We consider the Poisson equation in a bounded polygonal/polyhedral domain $\Omega \subset \mathbb{R}^d$ for $d = 2, 3$, with a complex/fine scale boundary $\partial\Omega = \Gamma_D \cup \Gamma_R$. That is, we consider

$$\begin{aligned} -\Delta u &= f && \text{in } \Omega, \\ \nu \cdot \nabla u + \kappa u &= 0 && \text{on } \Gamma_R, \\ u &= 0 && \text{on } \Gamma_D, \end{aligned} \tag{2.1}$$

where ν is the exterior unit normal of Ω, $0 \leq \kappa \in \mathbb{R}$, and $f \in L^2(\Omega)$. For simplicity we assume that, if $\kappa = 0$ then $\Gamma_D \neq \emptyset$ to ensure existence and uniqueness of the solution u. The weak form of the partial differential equation reads: find $u \in \mathcal{V} := \{v \in H^1(\Omega) \mid v|_{\Gamma_D} = 0\}$ such that

$$a(u, v) := \int_\Omega \nabla u \cdot \nabla v \,\mathrm{d}x + \int_{\Gamma_R} \kappa uv \,\mathrm{d}S = \int_\Omega fv \,\mathrm{d}x =: F(v), \tag{2.2}$$

for all $v \in \mathcal{V}$. Throughout the paper we use standard notation for Sobolev spaces [1]. We denote the local energy and L^2-norm in a subset $\omega \subset \Omega$ by

$$|||v|||_\omega = \left(\int_\omega |\nabla u|^2 \,\mathrm{d}x + \int_{\Gamma_R \cap \partial\omega} \kappa u^2 \,\mathrm{d}S \right)^{1/2}, \tag{2.3}$$

and

$$\|v\|_\omega = \left(\int_\omega u^2 \, \mathrm{d}x \right)^{1/2}, \tag{2.4}$$

respectively. Moreover, if $\omega = \Omega$ we omit the subscript, $|||v||| := |||v|||_\Omega$ and $\|v\| := \|v\|_\Omega$.

2.2 Reference finite element method

We embed the domain Ω within a polygonal domain Ω_0 equipped with a quasi-uniform and shape regular mesh $\mathcal{T}_{H,0}$, i.e., $\Omega \subset \Omega_0$ and $\overline{\Omega}_0 = \sum_{T \in \mathcal{T}_{H,0}} \overline{T}$. We let $\mathcal{T}_H$ be the sub mesh of $\mathcal{T}_{H,0}$ consisting of elements that are cut or covered by the physical domain Ω, i.e,

$$\mathcal{T}_H = \{T \in \mathcal{T}_{H,0} \mid T \cap \Omega \neq \emptyset\}. \tag{2.5}$$

The finite element space on $\mathcal{T}_H$ is defined by

$$\mathcal{V}_H = \{v \in \mathcal{C}^0(\Omega) \mid \forall T \in \mathcal{T}_H,\, v|_T \in \mathcal{P}_1(T)\}, \tag{2.6}$$

where $\mathcal{P}_1(T)$ is the space of polynomial of total degree ≤ 1 on T. We have $\mathcal{V}_H = \operatorname{span}\{\varphi_x\}_{x \in \mathcal{N}}$, where $\mathcal{N}$ is the set of all nodes in the mesh $\mathcal{T}_H$ and φ_x is the linear nodal basis function associated with node $x \in \mathcal{N}$.

The space $\mathcal{V}_H$ will not be sufficiently fine to represent the boundary and the boundary data. We therefore enrich the space $\mathcal{V}_H$ close to the complex boundary $\partial\Omega$. In order to construct the enrichment we define L-layer patches around the boundary recursively as follows

$$\begin{aligned} \omega^0 &:= \operatorname{int}\left((\bar{T} \in \mathcal{T}_H \mid T \cap \Omega \neq T) \cap \Omega \right), \\ \omega^\ell &:= \operatorname{int}\left((\bar{T} \in \mathcal{T}_H \mid \bar{T} \cap \bar{\omega}_T^{\ell-1} \neq \emptyset) \cap \Omega \right), \quad \text{for } \ell = 1, \ldots, L. \end{aligned} \tag{2.7}$$

note that ω^0 is the set all elements which are cut by the domain boundary $\partial\Omega$. An illustration of Ω, $\mathcal{T}_{H,0}$, ω^0, and ω^1 is given in Figure 1. We will later see in Lemma 4.8 that the appropriate number of layers is determined by the decay of the H^2-norm of the exact solution away from the boundary.

Let $\mathcal{T}_h$ be a fine mesh defined on ω^k, obtained by refining the coarse mesh in ω^k. On the interior part of the boundary $\partial\omega^k \setminus \partial\Omega$ we allow hanging nodes. Let $\mathcal{V}_h(\omega^k) = \{v \in \mathcal{C}^0(\omega^k) \mid \forall T \in \mathcal{T}_h,\, v|_T \in \mathcal{P}_1(T)\}$. We define a reference finite element space by

$$\mathcal{V}_h^\Gamma := (\mathcal{V}_H + \mathcal{V}_h(\omega^k)) \cap H^1_{\Gamma_D}(\Omega), \tag{2.8}$$

which consists of the standard finite element space enriched with a locally finer finite element space in ω^k. We assume that the space $\mathcal{V}_h^\Gamma$ is fine enough to resolve the fine scale features of the boundary, i.e., we assume that the boundary $\partial\Omega$ is exactly represented by the fine mesh $\mathcal{T}_h$.

The finite element method posed in the enriched space $\mathcal{V}_h^\Gamma$ reads: find $u_h \in \mathcal{V}_h^\Gamma$ such that

$$a(u_h, v) = F(v) \quad \text{for all } v \in \mathcal{V}_h^\Gamma. \tag{2.9}$$

We call the solution to (2.9) the reference solution and we have the following standard a priori error estimate

$$|||u - u_h||| \leq C(H|u|_{H^2(\Omega \setminus \omega^{k-1})} + h^{s-1}|u|_{H^s(\omega^{k-1})}) \tag{2.10}$$

where $1 \leq s \leq 2$ depends on the regularity of u in ω^k.

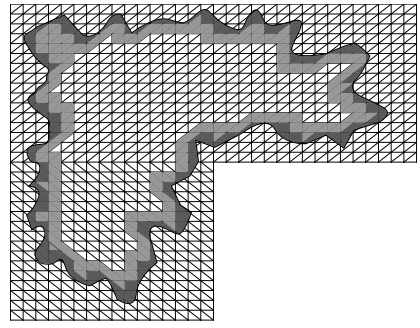


Figure 1: The computational domain Ω embedded in the mesh $\mathcal{T}_{H,0}$. The dark grey area is ω^0, and the union of the dark and light grey area is ω^1.

3 The multiscale method

In the multiscale method we want to construct a coarse scale approximation of u_h, which can be computed cheaply. We present the method in two steps:

- First, we construct a global multiscale method using a corrected coarse basis which takes the fine scale variation of the boundary into account.
- Then, we construct a localized multiscale method where the corrected basis is computed on localized patches.

3.1 Global multiscale method

For each $x \in \mathcal{N}$ (the set of free nodes) we define a L-layer nodal patch recursively as

$$\begin{aligned} \omega_x^0 &=: \operatorname{int}\left(\left(\bar{T} \in \mathcal{T}_H \mid \bar{T} \cap x \neq \emptyset\right) \cap \Omega\right), \\ \omega_x^\ell &=: \operatorname{int}\left(\left(\bar{T} \in \mathcal{T}_H \mid \bar{T} \cap \bar{\omega}_T^{\ell-1} \neq \emptyset\right) \cap \Omega\right), \quad \text{for } \ell = 1, \ldots, L. \end{aligned} \tag{3.1}$$

We consider a projective Clément interpolation operator defined by

$$\mathcal{I}_H v = \sum_{x \in \mathcal{N}_I} (P_x v)(x) \varphi_x, \tag{3.2}$$

where $\mathcal{N}_I$ is the index set of all interior nodes in Ω and P_x is a local L^2-projection defined by: find $P_x v \in \{v \in \mathcal{V}_H \mid \operatorname{supp}(v) \cap \omega_x^0 \neq \emptyset\}$ such that

$$(P_x v, w)_{\omega_x^0} = (v, w)_{\omega_x^0} \quad \text{for all } w \in \{v \in \mathcal{V}_H \mid \operatorname{supp}(v) \cap \omega_x^0 \neq \emptyset\}. \tag{3.3}$$

Using the interpolation operator we split the space $\mathcal{V}_h^\Gamma$ into the range and the kernel of the interpolation operator, i.e., $\mathcal{V}_H = \mathcal{I}_H \mathcal{V}_h^\Gamma$ and $\mathcal{V}^f = (1 - \mathcal{I}_H)\mathcal{V}_h^\Gamma$. However, the space $\mathcal{V}_H$ does

not have the satisfactory approximation properties. Instead we use the same idea as in LOD and construct an orthogonal splitting with respect to the bilinear form. We define the corrected coarse space as

$$\mathcal{V}_H^\Gamma = (1+Q)\mathcal{I}_H\mathcal{V}_h^\Gamma \tag{3.4}$$

where the operator Q is defined as follows: given $v_H \in \mathcal{V}_H$ find $Q(v_H) \in \{v \in \mathcal{V}^f \mid Q(v_H)|_{\Gamma_D} = -v_H\}$ such that

$$a(Q(v_H), w) = -a(v_H, w) \quad \text{for all } v \in \{v \in \mathcal{V}^f \mid Q(v_H)|_{\Gamma_D} = 0\}. \tag{3.5}$$

Note that $\mathcal{V}_H \not\subset \mathcal{V}_h^\Gamma$ but $\mathcal{V}_H^\Gamma \subset \mathcal{V}_h^\Gamma$ because the correctors are solved with boundary conditions that compensates for the nonconformity of the space $\mathcal{V}_H$. Also, from (3.5) we have the orthogonality $a(\mathcal{V}_H^\Gamma, \mathcal{V}^f) = 0$ and can write the reference space as the direct sum $\mathcal{V}_h^\Gamma = \mathcal{V}_H^\Gamma \oplus_a \mathcal{V}^f$.

The multiscale method posed in the space $\mathcal{V}_H^\Gamma$ reads: find $u_H \in \mathcal{V}_H^\Gamma$ such that

$$a(u_H, v) = F(v) \quad \text{for all } v \in \mathcal{V}_H^\Gamma. \tag{3.6}$$

3.2 Localized multiscale method

Finally, we localize the computation of the corrected basis functions to nodal patches instead of solving them globally on ω^k. Using linearity of the operator $Q(v_H)$, we obtain

$$Q(v_H) = \sum_{x \in \mathcal{N}_I} v_H(x) Q(\varphi_x). \tag{3.7}$$

We denote the localized corrector by

$$Q^L(v_H) = \sum_{x \in \mathcal{N}_I} v_H(x) Q_x^L(\varphi_x). \tag{3.8}$$

where $Q_x^L(\varphi_x)$ is the localization of $Q_x(\varphi_x)$ computed on an L-layer patch. The local correctors are computed as: given $x \in \mathcal{N}_I$ find $Q_x^L(\varphi_x) \in \{v \in \mathcal{V}^{\mathrm{f}} \mid v|_{\Omega\setminus\omega_x^L} = 0 \text{ and } v|_{\Gamma_D} = -\varphi_x\}$ such that,

$$a(Q_x^L(\varphi_x), w) = -a(\varphi_x, w) \quad \text{for all } v \in \{v \in \mathcal{V}^{\mathrm{f}} \mid v|_{\Omega\setminus\omega_x^L} = 0 \text{ and } v|_{\Gamma_D} = 0\}. \tag{3.9}$$

The localized multiscale method reads: find $u_H^L \in \mathcal{V}_H^{\Gamma,L} := \operatorname{span}\{\varphi_x + Q_x^L(\varphi_x)\}_{x\in\mathcal{N}_I}$ such that

$$a(u_H^L, v) = F(v) \quad \text{for all } v \in \mathcal{V}_H^{\Gamma,L}. \tag{3.10}$$

The space $\mathcal{V}_H^{\Gamma,L}$ has the same dimension as the coarse space $\mathcal{V}_H$ but the basis functions have slightly larger support. The multiscale solution $u_H^L \in \mathcal{V}_H^{\Gamma,L}$ has better approximation properties compared to the standard finite element solution on the same mesh, i.e, $u_H^L \in \mathcal{V}_H^{\Gamma,L}$ satisfy

$$|||u_h - u_H^L||| \leq CH, \tag{3.11}$$

where H is the mesh size and the constant C is independent of the fine scale features of the boundary Γ, which is proven in in Theorem 4.12.

In the next section we will show that the multiscale space has better approximation properties that the standard finite element space.

4 Error estimates

In this section we derive our main error estimates. First we present the following technical tools needed to prove the main result

- We present an explicit way to compute an upper bound for Poincaré-Friedrich constants on complex domains.
- We prove approximation properties of the interpolation operator on these domains.

The main result is obtained in four steps:

- We bound the difference between the analytic solution and the reference finite element solution $|||u - u_h|||$.
- We bound the difference between the reference finite element solution and the global multiscale method $|||u_h - u_H|||$.
- We bound the difference between a function $v \in \mathcal{V}_H$ modified by the global corrector and localized corrector $|||Q(v) - Q^L(v)|||$.
- Together these properties are used to estimate the error between the analytic solution and the localized multiscale method.

Furthermore, let $a \lesssim b$ abbreviate the inequality $a \leq Cb$ where C is any generic positive constant independent on the domain Ω and of the the coarse and fine mesh sizes H, h.

4.1 Poincaré-Friedrichs' inequality on complex domains

A crucial part of the proof is a Poincaré-Friedrichs inequality with a constant of moderate size. The Poincaré inequality reads: for all $u \in H^1(\omega)$,

$$\inf_{c\in\mathbb{R}} \|u - c\|_\omega \leq C(\omega)\mathrm{diam}(\omega)\|\nabla u\|_\omega, \tag{4.1}$$

holds where the optimal constant is

$$c = \frac{1}{|\omega|}\int_\omega u \,\mathrm{d}x. \tag{4.2}$$

Following [18], we consider inequalities of the following type: for all $u \in H^1(\omega)$

$$\|u - \lambda_\gamma(u)\|_\omega \leq C(\omega)\mathrm{diam}(\omega)\|\nabla u\|_\omega, \tag{4.3}$$

where $\gamma \subset \partial\omega$ is a $(d-1)$-dimensional manifold and

$$\lambda_\gamma(u) = \frac{1}{|\gamma|}\int_\gamma u \,\mathrm{d}S. \tag{4.4}$$

We introduce the notation $C_{\mathrm{PF}} = C(\omega)$, and refer to C_{PF} as the Poincaré-Friedrichs constant, which depends on the domain ω but not on its diameter.

A direct consequence of (4.3) is the following inequality

$$\|u\|_\omega \lesssim C_{\mathrm{PF}}\mathrm{diam}(\omega)\|\nabla u\|_\omega + \mathrm{diam}(\omega)^{1/2}\|u\|_\gamma, \tag{4.5}$$

if $\operatorname{diam}(\omega)^{d-1} \lesssim \operatorname{meas}_{d-1}(\gamma)$, i.e., the average is taken over a large enough manifold $\gamma \subset \partial\omega$. A short proof is given by

$$\begin{aligned}\|u\|_\omega &\leq \|u - \lambda_\gamma(u)\|_\omega + \|\lambda_\gamma(u)\|_\omega \\ &\leq C_{\mathrm{PF}}\operatorname{diam}(\omega)\|\nabla u\|_\omega + |\lambda_\gamma(u)|\operatorname{meas}_d(\omega)^{1/2} \\ &\leq C_{\mathrm{PF}}\operatorname{diam}(\omega)\|\nabla u\|_\omega + \|\lambda_\gamma(u)\|_\gamma \operatorname{meas}_{d-1}(\gamma)^{-1/2}\operatorname{meas}_d(\omega)^{1/2} \\ &\lesssim C_{\mathrm{PF}}\operatorname{diam}(\omega)\|\nabla u\|_\omega + \operatorname{diam}(\omega)^{1/2}\|\lambda_\gamma(u)\|_\gamma.\end{aligned} \tag{4.6}$$

Furthermore, from [18] we have the bound $C_{\mathrm{PF}} \leq 1$ for the Poincaré constant on a d-dimensional simplex where γ is one of the facets.

Next we will review some results given in [18] applied to domains with complex boundary. In [18] the notion of quasi-monotone paths is use to prove weighted Poincaré-Friedrichs type inequalities using average on $(d-1)$-dimensional manifolds $\gamma \subset \omega$. These results have also been discussed for perforated domains in [4].

Definition 4.1. For simplicity we assume that ω is a polygonal domain which is subdivided into a quasi-uniform partition of simplices $\tau = \{T_\ell\}_{\ell=1}^n$. We call the region $P_{\ell_1,\ell_2} = (\bar{T}_{\ell_1} \cup \bar{T}_{\ell_2} \cup \cdots \cup \bar{T}_{\ell_s})$ a path, if $\bar{T}_{\ell_i}$ and $\bar{T}_{\ell_{i+1}}$ share a common $(d-1)$-dimensional manifold, $\operatorname{meas}_{d-1}(T_{\ell_i} \cap T_{\ell_{i+1}}) > 0$. We will call $s_{\ell_1,\ell_s} := s$ the length of the path P_{ℓ_1,ℓ_s} and $\eta = \max_{T\in\tau}\{\operatorname{diam}(T)\}$.

Lemma 4.2. *Given τ from Definition 4.1 and define the index set $\mathcal{J} = \{\ell : \partial T_\ell \cap \gamma \neq \emptyset\}$. Then*

$$C_{\mathrm{PF}}^2 \lesssim \frac{s_{\max} r_{\max} \eta^{d+1}}{|\gamma| H^2}, \tag{4.7}$$

holds where $s_{\max} = \max(s_{k,j})$ is the length of the longest path and $r_{\max} = \max_{i\in\mathcal{I}} |\{(s,k) \in \mathcal{I}\times\mathcal{J} \mid T_i \in P_{k,j}\}|$ is the maximum times the paths intersect. For Friedrichs inequality we need the extra condition $u|_\gamma = 0$.

Proof. See [18] for a proof. □

We will now use Lemma 4.2 to show some cases when C_{PF} can be bounded independent of the complex/fine scale boundary $\partial\Omega$.

Fractal domain. We consider the fractal shaped domain given in Figure 2. First we compute $s_{\max}$. The number of T_ℓ on γ is then proportional to 2^k, where k is the total number of uniform refinements of the domain and we bound the maximum path length as

$$s_{\max} \sim \sum_{i=0}^{k} \frac{2^k}{2^i} \leq 2 \cdot 2^k, \tag{4.8}$$

i.e., the maximum length of a path is proportional to 2^k. Next we compute the maximum times a simplex is in a path, $r_{\max}$. First we show how many times the elements that are in γ are in a path and then we show that this number is larger than on any other γ_i, see Figure 2. The number of paths on each T_ℓ is the total number of elements in the domain. We get

$$r_\gamma \sim \sum_{i=0}^{k} n_\square(i) e_\square(i) = \sum_{i=0}^{k} 3^i \frac{(2^k)^2}{4^i} \leq 4^k \sum_{i=0}^{k} \left(\frac{3}{4}\right)^i \leq 4 \cdot 4^k, \tag{4.9}$$

where $n_\square(i)$ is the number sub domains with index i and $e_\square(i)$ is the number of elements inside a single sub domain with index i. Next we show that there is no T in other parts of the domain

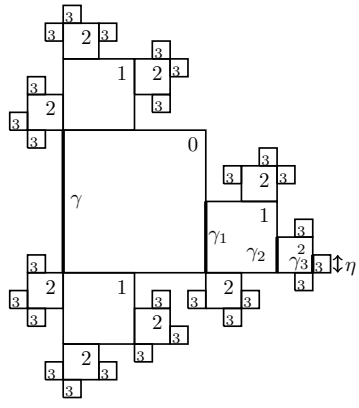


Figure 2: A fractal domain that has a bounded Poincaré-Friedrich constant.

where the number of paths are proportional to something with a stronger dependence on n than r_γ. We obtain,

$$r_{\gamma^j} \sim \frac{2^j}{3^j} \sum_{i=j}^{k} n_\square(i) e_\square(i) < r_\gamma, \tag{4.10}$$

where 2^j comes from that $2^j n_\square(j) = n_\square(0)$ and that only $1/3^j$ of the domain affects boundary r_{γ^j}. This proves that $r_{\max} \sim r_\gamma$, choosing L-type paths paths in the interior of the squares. To finish the argument we note that $H/\eta = 2^k$ and $|\gamma| = H$, and we obtain

$$C_{\mathrm{PF}}^2 \lesssim \frac{s_{\max} r_{\max} \eta^{d+1}}{|\gamma| H^2} \lesssim 1. \tag{4.11}$$

Saw tooth domain. An other example of a complex geometry is the saw domain given in Figure 3. Let the width of the saw teeth be $\eta = 2^{-k}$. A mesh constructed using 2^k uniform

Figure 3: Saw domain that has a bounded Poincaré-Friedrich constant. Here η is the width of one of the saw teeth.

refinements are needed to resolve the saw teeth. It is clear that $s_{\max} \sim 2^k$ and choosing L-shaped paths we have that $r_{\max} \sim (2^k)^2$. Again we have that $C_{\mathrm{PF}} \lesssim 1$ as long the length of the saw teeth are fixed.

An example of a domain with a non-bounded Poincaré-Friedrich constant is e.g. a dumbbell domain.

4.2 Estimation of the interpolation error

In this section we compute the interpolation error for a class of fine scale functions needed in the analysis. For each $T \in \mathcal{T}_H$ we define an L-layer element patch recursively as

$$\begin{aligned} \omega_T^0 &=: T \cap \Omega, \\ \omega_T^\ell &=: \operatorname{int}\left((\bar{T} \in \mathcal{T}_H \mid \bar{T} \cap \bar{\omega}_T^{\ell-1} \neq 0) \cap \Omega\right), \quad \text{for } \ell = 1, \ldots, L. \end{aligned} \tag{4.12}$$

Lemma 4.3. *The projective Clément type operator inherits the local approximation and stability properties for all interior elements, i.e., for all $v \in H^1(\Omega)$*

$$\|H^{-1}(v - \mathcal{I}_H v)\|_T + \|\nabla \mathcal{I}_H v\|_T \lesssim \|\nabla v\|_{\omega_T^1}, \tag{4.13}$$

holds for all interior elements $T \in \mathcal{T}_H$.

Proof. Follows directly from the standard proof [4] since $\sum_{i \in \mathcal{N}} \varphi_i$ is a partition of unity on interior elements. □

The trace of a function $v \in \mathcal{V}^f$ is "small" since the function v is in the kernel of an averaging operator,

$$\mathcal{V}^f = \{v \in \mathcal{V}_h^\Gamma \mid \mathcal{I}_H v = 0\}. \tag{4.14}$$

Lemma 4.4. *Given an interior element $T \in \mathcal{T}_H$, let $\gamma \subset \partial T$ be one of its faces, then*

$$\|v\|_\gamma^2 \lesssim H^{1/2} \|\nabla v\|_{\omega_T^1}, \tag{4.15}$$

holds for all $v \in \mathcal{V}^f$.

Proof. For an interior element T the standard approximation property of the Clemént type interpolation operator holds, i.e.,

$$\|v\|_T = \|v - \mathcal{I}_H v\|_T \lesssim H \|\nabla v\|_{\omega_T^1}. \tag{4.16}$$

since $v \in \mathcal{V}^f$. Using a trace inequality and (4.16) we obtain

$$\|v\|_\gamma^2 \lesssim H^{-1} \|v\|_T^2 + H \|\nabla v\|_T^2 \lesssim H \|\nabla v\|_{\omega_T^1}^2, \tag{4.17}$$

where T is an interior element. □

We will now make an assumption which is a sufficient condition to prove the main results of the paper and which will also simplify the analysis.

Assumption 4.5. All elements in $S \in \mathcal{T}_H$ share a vertex with an *interior element*, i.e., an element $T \in \mathcal{T}_H$ such that $T \cap \Omega = T$.

Lemma 4.6. *Let T be an element that is cut by the boundary $\partial\Omega$. Under Assumption 4.5 the following Poincaré-Friedrich like inequality holds*

$$\|v\|_T \lesssim H \|\nabla v\|_{\omega_T^2} \quad \textit{for all } v \in \mathcal{V}^f. \tag{4.18}$$

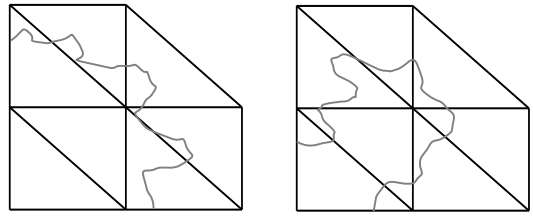


Figure 4: Admissible (left) and non-admissible (right) mesh according to Assumption 4.5. The line is where the elements are cut by the outer boundary. In this case a uniform refinement would make the (right) mesh admissible.

Proof. Let $\widetilde{T}$ be an element which share the vertex x with T. Using (4.5) we have that

$$\|v\|_T \leq \|v\|_{\omega_x^0} \lesssim H\|\nabla v\|_{\omega_x^0} + H^{1/2}\|u\|_\gamma \lesssim H\|\nabla v\|_{\omega_{\widetilde{T}}^1} \lesssim H\|\nabla v\|_{\omega_T^2}, \tag{4.19}$$

holds, since $\operatorname{diam}(\omega_T^{k-1}) \lesssim H$ and $\omega_x^0 \subset \omega_{\widetilde{T}}^1$. □

Next we prove local approximation and stability properties for functions which are in a larger space than $\mathcal{V}^f$.

Lemma 4.7. *Let $\mathcal{I}_H : L^2(\Omega_H) \to \mathcal{V}_H$ be the Clément interpolation operator defined by* (3.2)*. If $v = \eta w$, where η and w satisfies $0 \leq \eta \leq 1$, $\|\nabla\eta\|_{L^\infty} \lesssim H^{-1}$, and $w \in \mathcal{V}^f$, then the following estimate holds*

$$\|H^{-1}(v - \mathcal{I}_H v)\|_T + \|\nabla\mathcal{I}_H v\|_T \lesssim \|\nabla w\|_{\omega_T^2}, \tag{4.20}$$

for all $T \in \mathcal{T}_H$.

Proof. The local approximation and stability properties for an interior element follows directly from Lemma 4.16 together with

$$\|\nabla\eta w\|_T \lesssim H^{-1}\|w\|_T + \|\nabla w\|_T \lesssim \|\nabla w\|_{\omega_T^1}. \tag{4.21}$$

Next we investigate the local approximation and stability properties for elements on the boundary. Let T be an element cut by the boundary and $\widetilde{T}$ an interior element sharing vertex x with T. Then the L^2-stability follows directly from the stability of the interior elements, i.e,

$$\begin{aligned}\|(P_x v)(x)\|_T &= |(P_x v)(x)|\frac{\|1\|_T}{\|1\|_{\widetilde{T}}}\|1\|_{\widetilde{T}} \lesssim |(P_x v)(x)|\|1\|_{\widetilde{T}} = \|(P_x v)(x)\|_{\widetilde{T}} \\ &\lesssim H^{d/2}\|(P_x v)(x)\|_{L^\infty(\widetilde{T})} \leq H^{d/2}\|P_x v\|_{L^\infty(\widetilde{T})} \lesssim \|P_x v\|_{\widetilde{T}} \\ &\leq \|P_x v\|_{\omega_x^0} \leq \|v\|_{\omega_x^0} \leq \|v\|_{\omega_T^1}.\end{aligned} \tag{4.22}$$

We obtain

$$\|v - \mathcal{I}_H v\|_T \lesssim \|v\|_{\omega_T^1} \lesssim \|w\|_{\omega_T^1} \lesssim H\|\nabla w\|_{\omega_T^2}, \tag{4.23}$$

since $w \in \mathcal{V}^f$ using Lemma 4.4, L^2-stability of the interpolation operator, and $w \in \mathcal{V}^f$. Similar argument yields

$$\|\nabla\mathcal{I}_H v\|_T \lesssim \sum_{x\in\mathcal{N}_T} H^{-1}\|(P_x v)(x)\|_T \lesssim \|\nabla w\|_{\omega_T^2}, \tag{4.24}$$

where $\mathcal{N}_T$ is all vertices in element T. □

4.3 Estimation of the error in the reference finite element solution

The reference finite element solution $u_h \in \mathcal{V}_h^\Gamma$ has the following approximation property.

Lemma 4.8. *Let $u \in \mathcal{V}$ and $u_h \in \mathcal{V}_h^\Gamma$ be the solutions to* (2.2) *and* (2.9) *respectively, then*

$$|||u-u_h||| \lesssim \inf_{v_H \in \mathcal{V}_H} \left(\|H^{-1}(u-v_H)\|_{\Omega \setminus \omega_\Gamma^{k-1}} + |||u-v_H|||_{\Omega \setminus \omega_\Gamma^{k-1}} \right) + \inf_{v_h \in \mathcal{V}_h(\omega_\Gamma^k)} |||u-v_h|||_{\omega_\Gamma^{k-1}}, \tag{4.25}$$

holds.

Proof. We split Ω into the different parts $\Omega \setminus \omega^k$, $\omega_\Gamma^k \setminus \omega_\Gamma^{k-1}$, and ω_Γ^{k-1}. Since $\mathcal{V}_h^\Gamma \subset \mathcal{V}$, we have the best approximation result

$$|||u-u_h||| \lesssim |||u-w|||, \qquad \text{for all } v \in \mathcal{V}_h^\Gamma. \tag{4.26}$$

Let $\eta \in V_H$ be a cut off function, where $\eta|_{\Omega \setminus \omega^k} = 0$, $\eta|_{\omega^{k-1}} = 1$, and $\|\nabla \eta\|_{L^\infty(T)} \lesssim H^{-1}$. We construct $w = v_H + \pi_h \eta (v_h - v_H) \in \mathcal{V}_h^\Gamma$ where $v_H \in \mathcal{V}_H$, $v_h \in \mathcal{V}_h(\omega_\Gamma^k)$, and π_h is the nodal interpolant onto the finite element space $\mathcal{V}_h$ and obtain

$$|||u-w|||^2 = |||u-v_H|||^2_{\Omega \setminus \omega^k} + |||u-v_H - \pi_h \eta (v_h - v_H)|||^2_{\omega^k \setminus \omega_\Gamma^{k-1}} + |||u-v_h|||^2_{\omega_\Gamma^{k-1}}. \tag{4.27}$$

The first and third term are in the right form, see the statement of Lemma 4.8. Next we turn to the second term. Using the fact that the nodal interpolant π_h is H^1-stable for finite polynomial degrees (2 in our case) we obtain

$$\begin{aligned} |||\pi_h \eta (v_h - v_H)|||^2_{\omega^k \setminus \omega^{k-1}} &\lesssim |||\eta (v_h - v_H)|||^2_{\omega^k \setminus \omega^{k-1}} \\ &= ||\nabla(\eta(v_h - v_H))||^2_{\omega^k \setminus \omega^{k-1}} + \kappa ||\eta (v_h - v_H)||^2_{\partial(\omega^k \setminus \omega^{k-1}) \cap \partial \Gamma_R}. \end{aligned} \tag{4.28}$$

We have

$$\begin{aligned} &||\nabla(\eta(v_h - v_H))||_{\omega^k \setminus \omega^{k-1}} \\ &\quad \le ||(v_h - v_H)\nabla \eta||_{\omega^k \setminus \omega^{k-1}} + ||\eta \nabla (v_h - v_H)||_{\omega^k \setminus \omega^{k-1}} \\ &\quad \lesssim H^{-1} ||v_h - v_H||_{\omega^k \setminus \omega^{k-1}} + ||\nabla (v_h - v_H)||_{\omega^k \setminus \omega^{k-1}} \\ &\quad \lesssim H^{-1} ||v_h - u + u - v_H||_{\omega^k \setminus \omega^{k-1}} + ||\nabla (v_h - u + u - v_H)||_{\omega^k \setminus \omega^{k-1}} \\ &\quad \lesssim H^{-1} ||u - v_H||_{\omega^k \setminus \omega^{k-1}} + ||\nabla (u - v_H)||_{\omega^k \setminus \omega^{k-1}} \\ &\qquad + H^{-1} ||u - v_h||_{\omega^k \setminus \omega^{k-1}} + ||\nabla (u - v_h)||_{\omega^k \setminus \omega^{k-1}}. \end{aligned} \tag{4.29}$$

We also have that

$$\begin{aligned} \kappa^{1/2} ||(1-\eta)(v_h - v_H)||_{\partial(\omega^k \setminus \omega^{k-1}) \cap \partial \Omega} &\lesssim \kappa^{1/2} ||v - v_H||_{\partial(\omega^k \setminus \omega^{k-1}) \cap \partial \Omega} \\ &\lesssim \kappa^{1/2} ||v - v_H||^{1/2}_{\omega^k \setminus \omega^{k-1}} ||\nabla (v - v_H)||^{1/2}_{\omega^k \setminus \omega^{k-1}} \lesssim |||v - v_H|||_{\omega^k \setminus \omega^{k-1}}. \end{aligned} \tag{4.30}$$

Taking the infimum and using that

$$\begin{aligned} &\inf_{v_h \in \mathcal{V}_h} \left(H^{-1} ||u - v_h||_{\omega^k \setminus \omega^{k-1}} + ||\nabla (u - v_h)||_{\omega^k \setminus \omega^{k-1}} \right) \\ &\qquad \le \inf_{v_H \in \mathcal{V}_H} \left(H^{-1} ||u - v_H||_{\omega^k \setminus \omega^{k-1}} + ||\nabla (u - v_H)||_{\omega^k \setminus \omega^{k-1}} \right), \end{aligned} \tag{4.31}$$

since $h < H$ and $\mathcal{V}_H \subset V_h$, concludes the proof. □

The analysis extends to a non-polygonal boundary if we assume that h is fine enough to approximate the boundary using interpolation onto piecewise affine functions, see e.g. [20].

4.4 Estimation of the error in the global multiscale method

In this section we present and analyze the method with non-localized correctors.

Lemma 4.9. *Let $u_h \in \mathcal{V}_h^\Gamma$ solve (2.9) and $u_H \in \mathcal{V}_H^\Gamma$ solve (3.6), then*

$$|||u_h - u_H||| \lesssim H\|f\|_{\omega^k} \tag{4.32}$$

holds.

Proof. Any $u_h \in \mathcal{V}_h^\Gamma$ can be uniquely written as $u_h = u_H + u^{\mathrm{f}}$ where $u_H \in \mathcal{V}_H^\Gamma$ and $u^{\mathrm{f}} \in \mathcal{V}^{\mathrm{f}}$. This follows from the result from functional analysis, that if we have a projection $\mathcal{P} : u_h \to u_H$ onto a closed subspace, we have the unique split $u_h = \mathcal{P}u_h + (1-\mathcal{P})u_h$. For $\mathcal{P} = (1+Q)\mathcal{I}_H$, we have $\mathcal{P}\mathcal{V}_h^\Gamma = \mathcal{V}_H^\Gamma$ and

$$\mathcal{P}^2 = (1+Q)\mathcal{I}_H(1+Q)\mathcal{I}_H = (1+Q)\mathcal{I}_H\mathcal{I}_H = (1+Q)\mathcal{I}_H = \mathcal{P}. \tag{4.33}$$

We obtain

$$|||u^{\mathrm{f}}|||^2 = a(u^{\mathrm{f}}, u^{\mathrm{f}}) = (f, u^{\mathrm{f}})_{L^2(\Omega)} = (f, u^{\mathrm{f}} - \mathcal{I}_H u^{\mathrm{f}})_{L^2(\omega^k)} \lesssim H\|f\|_{\omega^k}|||u^{\mathrm{f}}|||, \tag{4.34}$$

which concludes the proof. □

4.5 Estimation of the error between global and localized correction

The correctors fulfill the following decay property.

Lemma 4.10 (Decay of correctors)**.** *For any $x \in \mathcal{N}_I$ there exist a $0 < \gamma < 1$ such that the local corrector $Q_x^L(\varphi_x) \in \mathcal{V}_L^{\mathrm{f}}(\varphi_x)$ and the global corrector $Q(\varphi_x) \in \mathcal{V}^{\mathrm{f}}(\varphi_x)$, which solves (3.9) and (3.5) respectively, fulfills the decay property*

$$|||(Q - Q^L)(v_H)|||^2 \leq L^d \gamma^{\lfloor (L-3)/3 \rfloor} \sum_{x \in \mathcal{N}_I} |||Q_x v|||^2, \tag{4.35}$$

where $\lfloor \cdot \rfloor$ is the floor function which maps a real number to the largest previous integer.

Proof. See Appendix A □

The localized corrected basis functions fulfill the following stability property.

Lemma 4.11. *Under Assumption 4.5 we have the stability*

$$|||\varphi_x + Q^L(\varphi_x)||| \lesssim H^{-1}||\varphi_x||, \tag{4.36}$$

for the corrected basis function given any $x \in \mathcal{N}_I$.

Proof. First we will prove that there exist a (non-unique) function $g_x \in \mathcal{V}^f(\omega_x^L)$ such that $(g_x - \varphi_x)|_{\Gamma_D} = 0$ and $|||g_x||| \lesssim |||\varphi_x|||$ for all $x \in \mathcal{N}_I$. Given any x define $w|_T = g_x - \varphi_x$ and $w|_{\Omega \setminus T_I} = 0$ where T is an interior element. The function w have to fulfill the following restriction

$$\mathcal{I}_H w = \mathcal{I}_H \varphi_x = \varphi_x, \tag{4.37}$$

which is equivalent to

$$P_y w(y) = \delta_{xy}, \tag{4.38}$$

where δ_{xy} is the Kronecker delta function. In order to construct w we perform two a uniform refinements in 2D. A similar construction is possible in 3D using two red-green refinements. Then we have three free nodes in T for a function that is zero on the boundary ∂T. We write w as $w = \sum_{j=1}^{d+1} \alpha_j \hat{\varphi}_j$ where φ_j are the P_1 Lagrange basis function associated with the three interior nodes. We can determine w by letting it fulfill

$$\sum_{i,j=1}^{d+1} (P_{y_i} \alpha_j \hat{\varphi}_j)(y_i) = \delta_{x,y_i}. \tag{4.39}$$

The value $P_y \hat{\varphi}_i(x)$ can be computed as

$$P_y \hat{\varphi}_j(y) = \delta_y^T (\Pi^T M_{H/4} \Pi)^{-1} \Pi^T M_{H/4} \varphi_j, \tag{4.40}$$

where $M_{H/4}$ is a local mass matrix computed on ω_x^0, $\delta_x^T = 1$ for index x and 0 otherwise, and $\Pi : \mathcal{V}_H \to \mathcal{V}_{H/4}$ is a projection from the finer space onto the coarse. On a quasi-uniform mesh $P_y \hat{\varphi}_j(y)$ is independent of H. Therefore, α_j is independent of H and there exist a constant such that

$$||w|| \leq C||\varphi_x||. \tag{4.41}$$

This yields

$$|||w||| \lesssim 4H^{-1}||w|| \lesssim H^{-1}||\varphi_x||. \tag{4.42}$$

Using the triangle inequality we have

$$|||g_x||| \leq |||\varphi_x||| + |||w||| \lesssim H^{-1}||\varphi_x||. \tag{4.43}$$

Next we consider the problem: find $Q_0 \in V^f(\omega_x^L)$ such that

$$a(Q_0, w) = a(\varphi_x - g_x, w) \quad w \in \mathcal{V}^f(\omega_x^L), \tag{4.44}$$

where g satisfy $g \in \mathcal{V}^f(\omega_x^L)$, $(\varphi_x - g)|_{\Gamma_D} = 0$, and $|||g_x||| \lesssim |||\varphi_x|||$. It is clear that $Q^L(\varphi_x) = Q_0 + g$. For the stability we obtain

$$\begin{aligned} |||\varphi_x + Q^L(\varphi_x)|||^2 &\leq a(\varphi_x + Q^L(\varphi_x), \varphi_x + Q^L(\varphi_x)) = a(\varphi_x + Q^L(\varphi_x), \varphi_x + Q_0 + g) \\ &= a(\varphi_x + Q^L(\varphi_x), \varphi_x + g) \leq |||\varphi_x + Q^L(\varphi_x)|||(|||\varphi_x||| + |||g_x|||) \\ &\lesssim |||\varphi_x + Q^L(\varphi_x)||| \cdot |||\varphi_x|||, \end{aligned} \tag{4.45}$$

which concludes the proof. □

4.6 Estimation of the error for the localized multiscale method

The a priori results for the localized multiscale method reads.

Theorem 4.12. *Let $u \in \mathcal{V}$ solve (2.2) and $u_H^L \in \mathcal{V}_H^{\Gamma,L}$ solve (3.6). Then under Assumption 4.5 the bound*

$$\begin{aligned} |||u - u_H^L||| \lesssim \inf_{v \in \mathcal{V}_H} &\left(\|H^{-1}(u - v)\|_{\Omega \setminus \omega_\Gamma^0} + |||u - v|||_{\Omega \setminus \omega_\Gamma^0} \right) \\ &+ \inf_{v \in \mathcal{V}_h(\omega_\Gamma^k)} |||u - v|||_{\omega_\Gamma^{k-1}} + \|Hf\|_{\omega^k} + L^{d/2} H^{-1} \gamma^L \|f\|_\Omega, \end{aligned} \tag{4.46}$$

holds for $k \geq 2$.

Proof. Since $\mathcal{V}_H^{\Gamma,L} \subset \mathcal{V}$ we have the best approximation result

$$|||u - u_H^L||| \leq |||u - v_H||| \quad \text{for all } v_H \in \mathcal{V}_H^{\Gamma,L}. \tag{4.47}$$

We obtain

$$|||u - u_H^L||| \leq |||u - u_h||| + |||u_h - u_H||| + |||u_H - v_H|||, \tag{4.48}$$

using the triangle inequality. The first and second term is bounded using Lemma 4.8 and 4.9. For the third term we choose $v_H = \mathcal{I}_H u_H + Q(\mathcal{I}_H u_H)$ which gives

$$|||u_H - v_H|||^2 = |||Q(\mathcal{I}_H u_H) - Q^L(\mathcal{I}_H u_H)|||^2 \lesssim L^d \gamma^{2L} \sum_{x \in \mathcal{N}_I} |||Q_x(\mathcal{I}_H u_H)|||^2, \tag{4.49}$$

using Lemma 4.10. Using Lemma 4.11 we obtain

$$\begin{aligned} |||u_H - v_H|||^2 &\lesssim H^{-2} L^d \gamma^{2L} \sum_{x \in \mathcal{N}_I} u_H(x)^2 \|\varphi_x\|^2 \lesssim L^d \gamma^{2L} \|\mathcal{I}_H u_H\|^2 \\ &\lesssim H^{-2} L^d \gamma^{2L} ||\mathcal{I}_H u_H||^2 \lesssim H^{-2} L^d \gamma^{2L} ||u_H||^2 \\ &\lesssim H^{-2} L^d \gamma^{2L} |||u_H|||^2 \leq H^{-2} L^d \gamma^{2L} ||f||^2, \end{aligned} \tag{4.50}$$

using a Poincaré-Friedrich inequality. Combining (4.48) and (4.50) concludes the proof. □

5 Implementation and conditioning

In this section we will shortly discuss how to implement the method and analyze the conditioning of the matrices.

5.1 Implementation

To compute $Q^L(\varphi_x)$ in (3.9) we impose the extra condition $\mathcal{I}_H v = 0$ using Lagrangian multipliers. Let n_x and N_x be the number of fine and coarse degrees of freedom in the patch ω_x^0. Let M_x and K_x denote the local mass and stiffness matrix on ω_x^0 satisfying

$$(v, w)_{\omega_x^0} \quad \Leftrightarrow \quad \hat{w}^T M_x \hat{v}, \tag{5.1}$$

and

$$a(v, w)|_{\omega_x^0} \quad \Leftrightarrow \quad \hat{w}^T K_x \hat{v}, \tag{5.2}$$

where $v, w \in \mathcal{V}_h|_{\omega_x^0}$ and $\hat{w}, \hat{v} \in \mathbb{R}^{n_x}$ are the nodal values of v, w. We also define the projection matrix $\Pi_x : \{v \in \mathcal{V}_H(\omega_x^0) \mid \text{supp}(v) \cap \omega_x^0 \neq 0\} \to \mathcal{V}_h(\omega_x^0)$ of size $(n_x \times N_x)$ which project a coarse function onto the fine mesh and a Kronecker delta vector of size $N_x \times 1$ as $\delta_x = (0, \ldots, 0, 1, 0, \ldots, 0)$ where 1 is in node x. We obtain

$$P_x v(x) = 0 \quad \Leftrightarrow \quad \lambda_x^T \hat{v} = \delta_x^T (\Pi_x^T \widehat{M}_x \Pi_x)^{-1} \Pi_x^T M_x \hat{v} = 0. \tag{5.3}$$

Set $\Lambda = [\lambda_{y_1}, \lambda_{y_1}, \ldots, \lambda_{y_{N_x}}]$, then (3.9) is equivalent to solving the linear system

$$\begin{bmatrix} K_x & \Lambda \\ \Lambda^T & 0 \end{bmatrix} \begin{bmatrix} \widehat{Q}^L(\varphi_x) \\ \mu \end{bmatrix} = \begin{bmatrix} -K_x \Pi_x \widehat{\varphi}_x \\ 0 \end{bmatrix}, \tag{5.4}$$

where $\widehat{Q}^L(\varphi_x) \in \mathbb{R}^{n_x}$ contains the nodal values of $Q^L(\varphi_x)$ and μ is a Lagrange multiplier. For each coarse node x in ω_x^0 we need to invert $\Pi_x^T \widehat{M}_x \Pi_x$ to assemble (5.4), however since the size is

only $(N_x \times N_x)$ they are cheap to invert. The coarse scale stiffness matrix $\widehat{K}$ in (3.10) is given element wise by

$$\widehat{K}_{i,j} = a(\varphi_j + Q^L(\varphi_j), \varphi_i + Q^L(\varphi_i)). \tag{5.5}$$

We save the nodal values of the corrected basis as

$$\Phi = \left[\varphi_{y_1} + \widehat{Q}^L(\varphi_{y_1}), \varphi_{y_2} + \widehat{Q}^L(\varphi_{y_2}) \ldots, \varphi_{y_N} + \widehat{Q}^L(\varphi_{y_N}).\right], \tag{5.6}$$

Given the stiffness matrix on the fine scale K and the collection of corrected basis functions Φ, we can compute the coarse stiffness matrix as

$$\widehat{K} = \Phi^T K \Phi. \tag{5.7}$$

and the linear system (3.10) is computed as

$$\widehat{K}\hat{u}_H^{\Gamma,L} = b, \tag{5.8}$$

where $\hat{u}_H^{\Gamma,L}$ is the nodal values of $\hat{u}_H^{\Gamma,L}$ and b correspond to some right hand side $b_{y_1} = (f, \varphi_{y_1} + \widehat{Q}^L(\varphi_{y_1}))$. However, the fine stiffness matrix does not have to be assembled globally. If a Petrov-Galerkin formulation is used, further savings can be made [9].

5.2 Conditioning

The Euclidean matrix norm is defined as

$$||A||_N = \sup_{0 \neq v \in \mathbb{R}^N} \frac{|Av|_N}{|v|_N}, \tag{5.9}$$

where $\langle v, w \rangle = \sum_{i=1}^N v(x_i)w(x_i)$ and $|v|_N = \sqrt{\langle v, v \rangle}$.

Theorem 5.1. *The bound*

$$\kappa = \|\widehat{K}\|_N \|\widehat{K}^{-1}\|_N \lesssim H^{-2}, \tag{5.10}$$

on the condition number κ holds.

Proof. To prove the condition number we use the following three properties.

1. An inverse type inequality for the modified basis functions. We have

$$|||\varphi_i + Q(\varphi_i)||| \lesssim H^{-1}\|\varphi_i\|, \tag{5.11}$$

   from Lemma 4.11.

2. A Poincaré-Friedrich type inequality on the full domain, see Section 4.1.

3. An equivalence between the Euclidean norm and the L^2-norm. We have that

$$\begin{aligned} \|v\|^2 &\leq \sum v_i^2 \|\varphi_i + Q(\varphi_i)\|^2 \lesssim \sum v_i^2 (\|\varphi_i\| + \|Q(\varphi_i) - \mathcal{I}_H Q(\varphi_i)\|)^2 \\ &\lesssim \sum v_i^2 (\|\varphi_i\| + H\|\nabla Q(\varphi_i)\|)^2 \\ &\lesssim \sum v_i^2 (\|\varphi_i\| + H\|\nabla \varphi_i\| + H\|\nabla(\varphi_i - Q(\varphi_i))\|)^2 \\ &\lesssim \sum v_i^2 \|\varphi_i\|^2 \lesssim H^d |v|_N, \end{aligned} \tag{5.12}$$

and

$$|v|_N^2 = \sum_{i=1}^N v_i^2 \lesssim H^{-d} \sum_{i=1}^N v_i^2 \|\varphi_i\|^2 \lesssim H^{-d} \|\sum_{i=1}^N v_i \varphi_i\|^2 = H^{-d} \|\mathcal{I}_H v\|^2 \lesssim H^{-d} \|v\|^2. \quad (5.13)$$

holds, hence $|v|_N \sim H^{-d/2}\|v\|$.

We have

$$|\widehat{K}v|_N = \sup_{0\neq w\in\mathbb{R}^N} \frac{|\langle \widehat{K}v, w\rangle|}{|w|_N} = \sup_{0\neq w\in\mathbb{R}^N} \frac{|a(v,w)|}{|w|_N} = \sup_{0\neq w\in\mathbb{R}^N} \frac{|||v||| \cdot |||w|||}{|w|_N} \leq H^{d-2}|v|_N, \quad (5.14)$$

using property 1) and 3). Also

$$\begin{aligned} |\widehat{K}^{-1}v|_N^2 &= H^{-d}\|\widehat{K}^{-1}v\| \leq C_{\mathrm{PF}} H^{-d} |||\widehat{K}^{-1}v||| \\ &\leq H^{-d}\langle \widehat{K}\widehat{K}^{-1}v, \widehat{K}^{-1}v\rangle \leq H^{-d}|v|_N \cdot |\widehat{K}^{-1}v|_N, \end{aligned} \quad (5.15)$$

using property 2) and 3). The proof is concluded by taking the supremum over v. □

6 Numerical experiments

In the following section we present some numerical experiments which verifies our analytical results. In all the following experiment we fix the right hand side to $f = 1$.

6.1 Accuracy on fractal shaped domain

We consider the domain in Figure 2. We use homogeneous Dirichlet boundary condition on the most left, down, and right hand side boundaries and Robin boundary condition with $\kappa = 10$ on the rest. Correctors are computed in the full domain. The reference solution is computed using $h = 2^{-9}$. As seen in Figure 5, a even higher convergence rate than linear convergence to the reference solution and the correct scaling of the condition number are observed with respect to the coarse mesh H.

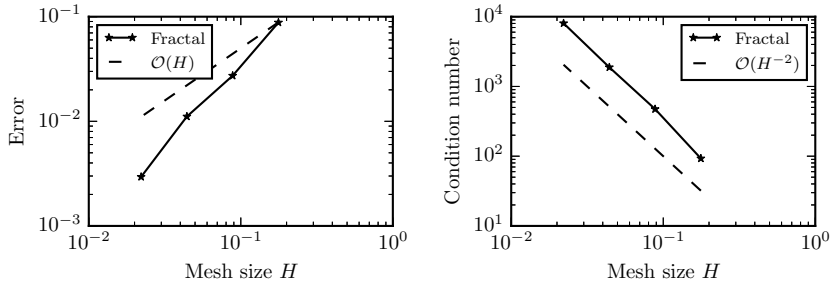


Figure 5: The convergence rate to the reference solution in relative energy norm (left) and the scaling of the condition number (right) for the fractal domain in Figure 5.

6.2 Locally added correctors around singularities

Let us consider two different domains, the L-shaped domain $L = ([0,1]\times[0,1])\backslash([0.5,1]\times[0,0.5])$ and a slit-domain $L = ([0,1]\times[0,1])\backslash([0.5,0.5]\times[0,0.5])$ with homogeneous Dirichlet boundary condition. We only compute correctors in the vicinity of the singularities, see Figure 6. As seen in Figure 7, the correct convergence rates to the reference solutions and the correct scaling of the condition number are observed for both singularities.

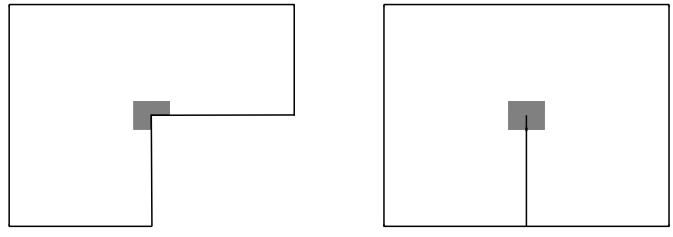


Figure 6: The marked area is where the finite element space is enriched, for the L-shaped domain (left) and a slit domain (right).

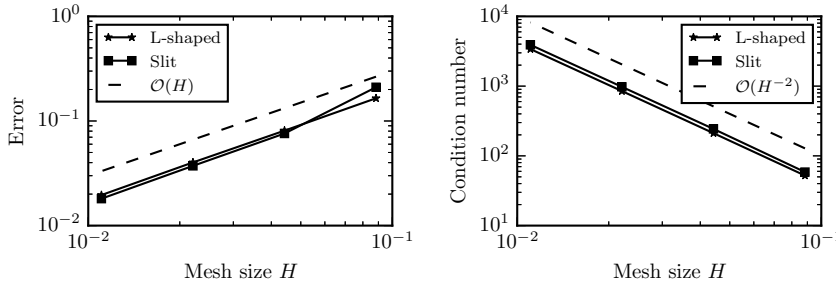


Figure 7: The convergence rate to the reference solution in relative energy norm (left) and scaling of the condition number (right) for a L-shaped and a slit domain.

6.3 Locally add correctors around saw tooth boundary

Let us consider a unit square where one of the boundaries are cut as a saw tooth and where correctors are only computed in the vicinity of the saw teeth, see Figure 8. On all the non saw tooth boundaries we use homogeneous Dirichlet boundary conditions. On the saw tooth boundary we test both homogeneous Dirichlet and Neumann boundary condition. We observe the correct convergence and scaling of the conditoin number, see Figure 9.

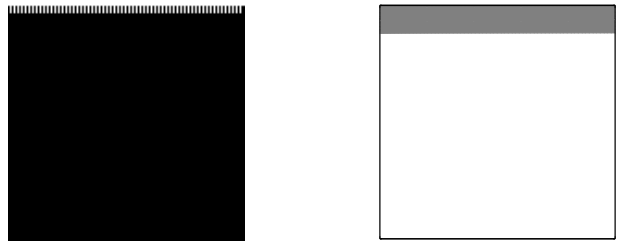


Figure 8: We consider the saw tooth domain (left). The marked area (right) is where the finite element space is enriched.

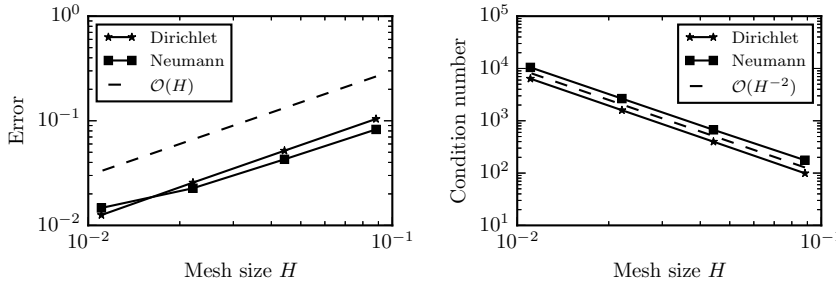


Figure 9: The convergence rate to the reference solution in relative energy norm (left) and the scaling of the condition number (right) for the saw tooth shaped boundary using different boundary condition on the saw tooths.

6.4 Accuracy and conditioning on cut domains

Let us consider an L-shaped domain $\Omega = ([0,1]\times[0,1])\backslash([0.5,1]\times[0,0.5])$. We want to investigate how sensitive the accuracy in the solution and the conditioning of the coarse stiffness matrix are to how the coarse mesh are cut for different boundary condition. We fix the coarse $H = 2^{-3}$ and fine $h = 2^{-8}$ mesh sizes and consider three different setups of boundary condition, homogeneous Dirichlet on the whole boundary, homogeneous Dirichlet on the cut elements and homogeneous Neumann otherwise, and homogeneous Neumann on the cut elements and homogeneous Dirichlet otherwise. We will cut the coarse mesh in two different ways, 1) with a horizontal cut and 2) a circular cut around the reentered corner, see Figure 10. The errors are measured in energy norm. The results are presented in Table 1 and 2. We conclude that neither the error nor the conditioning are sensitive to how the domain is cut.

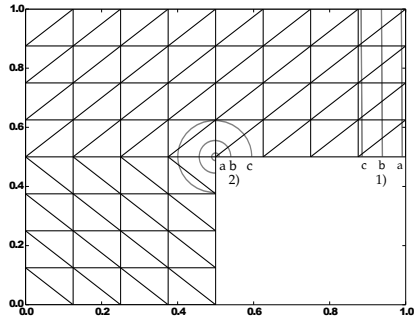


Figure 10: A given background mesh which is cut in two different ways 1) and 2) with different size of the cut a), b), and c).

Γ_{cut}	$\partial\Omega \setminus \Gamma_{cut}$	$e_{\text{rel}}(a)$	$e_{\text{rel}}(b)$	$e_{\text{rel}}(c)$
D	D	0.059	0.057	0.056
D	N	0.018	0.019	0.020
N	D	0.063	0.055	0.053
		cond(a)	cond(b)	cond(c)
D	D	9.85	10.10	13.63
D	N	299.75	282.26	353.27
N	D	10.537	10.79	11.47

Table 1: For cut 1 we have $L \cap [0, 1-r] \times [0, 1]$, where $r = \{h, 0.5H, H-h\}$ in a), b), and c), respectively. The errors measured in relative energy norm and condition number of the coarse stiffness matrix are presented. We try the different boundary conditions, D (Dirichlet) and N (Neumann), on the boundary segement Γ_{cut}, which cuts the elements.

Γ_{cut}	$\partial\Omega \setminus \Gamma_{cut}$	$e_{\text{rel}}(a)$	$e_{\text{rel}}(b)$	$e_{\text{rel}}(c)$
D	D	0.060	0.064	0.073
D	N	0.0205	0.035	0.048
N	D	0.060	0.057	0.059
		cond(a)	cond(b)	cond(c)
D	D	9.80	9.23	7.03
D	N	246.99	107.52	59.67
N	D	9.90	11.44	12.16

Table 2: For cut 2 we have $L \setminus B(x_0, r)$ for a ball B centered in $x_0 = (0.5, 0.5)$ and with radius r, where $r = \{h, 0.5H, H\}$ in a), b), and c), respectively. The errors measured in energy norm and condition number of the coarse stiffness matrix are presented. We try the different boundary conditions, D (Dirichlet) and N (Neumann), on the boundary segement Γ_{cut}, which cuts the elements.

A Proofs

In the appendix we collect proofs of the more technical results.

Proof of Lemma 4.10. We will make frequent use of the cut off function $\eta_x^{k-1,k}$ which satisfy $\eta_x^{k-1,k} = 0$ in ω_x^{k-1}, $\eta_x^{k-1,k} = 1$ in $\Omega \setminus \omega_x^k$, and $\|\nabla \eta_x^{k-1,k}\|_{L^\infty(\Omega)} \lesssim H^{-1}$.

Let $e = (Q - Q^L)(v)$, we obtain

$$|||e|||^2 \lesssim a(e,e) = \sum_{x \in \mathcal{N}_I} a\left((Q_x - Q_x^L)(v), e\right) = \sum_{x \in \mathcal{N}_I} \left(a((Q_x - Q_x^L)(v), e - \tilde{v}_x\right) \tag{A.1}$$

where we choose $\tilde{v}_x = \eta_x^{L+2,L+1} e - \mathcal{I}_H \eta_x^{L+2,L+1} e$ which satisfy

$$a((Q_x - Q_x^L)(v), \tilde{v}_x) = 0. \tag{A.2}$$

since $\tilde{v}_x \in \mathcal{V}^{\mathrm{f}}$ and $\operatorname{supp}(\tilde{v}_x) \cap \operatorname{supp}(Q_x^L(v)) = 0$. For each $x \in \mathcal{N}_I$, we obtain

$$a\left((Q_x - Q_x^L)(v), e - \tilde{v}_x\right) \leq |||(Q_x - Q_x^L)(v)||| \cdot |||e - \tilde{v}_x||| \tag{A.3}$$

where we split

$$|||e - \tilde{v}_x||| \leq |||e - \eta_x^{L+2,L+1} e||| + |||\eta_x^{L+2,L+1} e - \tilde{v}_x|||. \tag{A.4}$$

The first term in (A.4) can be bounded as

$$\begin{aligned} |||(1 - \eta_T^{L+2,L+1})e|||_{\omega_T^{L+2}}^2 &= \|\nabla(1 - \eta_T^{L+2,L+1})e\|_{\omega_T^{L+2}}^2 + \|\kappa(1 - \eta_T^{L+2,L+1})e\|_{\Gamma_T^{L+2}}^2 \\ &\leq \|\nabla e\|_{\omega_T^{L+2}}^2 + H^{-1}\|e\|_{\omega_T^{L+2}}^2 + \|\kappa(1 - \eta_T^{L+2,L+1})e\|_{\Gamma_T^{L+2}}^2 \\ &\leq |||e|||_{\omega_T^{L+3}}^2 \end{aligned} \tag{A.5}$$

using the product rule, inverse estimates, and interpolation estimates. The second term in (A.4) can be bounded as

$$|||\eta_x^{L+2,L+1} e - \tilde{v}|||^2 = \|\nabla \mathcal{I}_H(\eta_T^L e)\|^2 + \|\kappa \mathcal{I}_H(\eta_T^L e)\|_\Gamma^2 \lesssim |||e|||_{\omega_x^{L+4}}^2 \tag{A.6}$$

where we used

$$\begin{aligned} \|\nabla \mathcal{I}_H(\eta_x^{L+2,L+1} e)\|^2 = \|\nabla(\mathcal{I}_H \eta_x^{L+2,L+1} e)\|_{\omega_x^{L+3} \setminus \omega_x^L}^2 &\lesssim H^{-2}\|e\|_{\omega_x^{L+3} \setminus \omega_x^L}^2 \\ &\leq |||e|||_{\omega_x^{L+4} \setminus \omega_x^{L-1}}^2 \leq |||e|||_{\omega_x^{L+4}}^2 \end{aligned} \tag{A.7}$$

and

$$\begin{aligned} \|\kappa \mathcal{I}_H(\eta_x^{L+2,L+1} e)\|_{\Gamma_x^L}^2 = \sum_{E \in \Gamma_x^L} \|\kappa \mathcal{I}_H(\eta_x^{L-2,L-1} e)\|_E^2 &\leq \kappa H^{-1}\|\mathcal{I}_H(\eta_{\omega_x^{L-1}} e)\|_{\omega_x^{L+3}}^2 \\ &\leq \kappa H^{-1}\|e - \mathcal{I}_H e\|_{\omega_x^{L+3}}^2 \leq \kappa H \|\nabla e\|_{\omega_x^{L+4}}^2 \leq |||e|||_{\omega_x^{L+4}}^2 \end{aligned} \tag{A.8}$$

which follows from Lemma 4.7 and that $\kappa \lesssim H^{-1}$. Hence, combining (A.1), (A.4), (A.5), and (A.6) we obtain

$$\begin{aligned} |||e|||^2 &\lesssim \sum_{x \in \mathcal{N}_I} |||(Q_x - Q_x^L)(v)||| \cdot |||e|||_{\omega_x^{L+4}} \\ &\lesssim L^d \left(\sum_{x \in \mathcal{N}_I} |||(Q_x - Q_x^L)(v)|||^2 \right)^{1/2} |||e||| \end{aligned} \tag{A.9}$$

that is

$$|||e|||^2 \lesssim L^d \sum_{x\in\mathcal{N}_I} ||||(Q_x - Q_x^L)(v)|||^2 \tag{A.10}$$

Let $e_x = (Q_x - Q_x^L)(v)$ and again use $\eta_x^{L+2,L+1}$ and $\tilde{w}_x = (1-\eta_x^{L-1,L-2})Q_x v + \mathcal{I}_H \eta_x^{L-1,L-2} Q_x v \in \mathcal{V}^f(\omega_x^L)$, we obtain

$$\begin{aligned} |||e_x|||^2 \le a(e_x, e_x) &= a(e_x, Q_x v - \tilde{w}_x) \\ &= a(e_x, \eta_x^{L-1,L-2} Q_x v - \mathcal{I}_H \eta_x^{L-1,L-2} Q_x v) \\ &\le |||e_x|||(|||\eta_x^{L-1,L-2} Q_x v||| + |||\mathcal{I}_H \eta_x^{L-1,L-2} Q_x v|||) \\ &\lesssim |||e_x||| \cdot |||Q_x v|||_{\Omega\setminus\omega_x^{L-3}} \end{aligned} \tag{A.11}$$

Next we construct a recursive scheme which will be used to show the decay. We obtain

$$\begin{aligned} |||Q_T v|||^2_{\Omega\setminus\omega_x^k} &\le \int_\Omega \eta_T^{k,k-1} \nabla Q_T v \nabla Q_T v \,\mathrm{d}x + \int_{\Gamma_R} \eta_T^{k,k-1} \kappa Q_T v Q_T v \,\mathrm{d}S \\ &= \int_\Omega \nabla Q_T v \nabla(\eta_T^{k,k-1} Q_T v) \,\mathrm{d}x + \int_{\Gamma_R} \kappa Q_T v (\eta_T^{k,k-1} Q_T v) \,\mathrm{d}S \\ &\quad - \int_\Omega Q_T v \nabla Q_T v \nabla \eta_T^{k,k-1} \,\mathrm{d}x \\ &= a(Q_T v, \eta_T^{k,k-1} Q_T v) - \int Q_T v \nabla Q_T v \nabla \eta_T^{k,k-1} \,\mathrm{d}x. \end{aligned} \tag{A.12}$$

The first term in (A.12) can be bounded as

$$\begin{aligned} a(Q_x v, \eta_x^{k,k-1} Q_x v) &= a(Q_x v, \eta_x^{k,k-1} Q_x v - \mathcal{I}_H \eta_x^{k,k-1} Q_x v) + a(Q_x v, \mathcal{I}_H \eta_x^{k,k-1} Q_x v) \\ &= a(Q_x v, \mathcal{I}_H \eta_x^{k,k-1} Q_x v) \lesssim |||Q_x v|||_{\omega_x^k \setminus \omega_x^{k-1}} |||\mathcal{I}_H \eta_x^{k,k-1} Q_x v||| \\ &\lesssim |||Q_x v|||_{\omega_x^k\setminus\omega_x^{k-1}} |||Q_x v|||_{\omega_x^{k+1}\setminus\omega_x^{k-2}} |||Q_x v|||^2_{\omega_x^{k+1}\setminus\omega_x^{k-2}} \end{aligned} \tag{A.13}$$

using Lemma 4.7. The second term is bounded as

$$\begin{aligned} \int_\Omega Q_x v \nabla Q_x v \nabla \eta_x^{k,k-1} \,\mathrm{d}x &\le H^{-1} \|Q_x v - \mathcal{I}_H Q_x v\|_{\omega_x^k\setminus\omega_x^{k-1}} |||Q_x v|||_{\omega_x^k\setminus\omega_x^{k-1}} \\ &\lesssim |||Q_x v|||_{\omega_x^{k+1}\setminus\omega_x^{k-2}} |||Q_x v||||_{\omega_x^{k+1}\setminus\omega_x^{k-2}} \le |||Q_x v|||^2_{\omega_x^{k+1}\setminus\omega_x^{k-2}} \end{aligned} \tag{A.14}$$

Combining (A.12), (A.13), and (A.14) we obtain

$$\begin{aligned} |||Q_T v|||^2_{\Omega\setminus\omega_x^k} &\le C_1 |||Q_x v|||^2_{\omega_x^{k+1}\setminus\omega_x^{k-2}} = C_1 \left(|||Q_x v|||^2_{\Omega\setminus\omega_x^{k-2}} - |||Q_x v|||^2_{\Omega_x\setminus\omega_x^{k+1}} \right) \\ &\le C_1 \left(|||Q_x v|||^2_{\Omega\setminus\omega_x^{k-2}} - |||Q_x v|||^2_{\Omega_x\setminus\omega_x^k} \right) \end{aligned} \tag{A.15}$$

and hence

$$|||Q_x v|||^2_{\Omega\setminus\omega_x^k} \le \gamma |||Q_x v|||^2_{\Omega\setminus\omega_x^{k-2}}, \tag{A.16}$$

where $\gamma = \frac{C_1}{1+C_1}$. Using (A.16) recursively we obtain

$$\begin{aligned} & |||Q_x v|||^2_{\Omega\setminus\omega^{L-3}} \le \gamma^k |||Q_x v|||^2_{\Omega\setminus\omega_x^{L-3(k+1)}} \\ \Leftrightarrow \quad & |||Q_x v|||^2_{\Omega\setminus\omega^{L-3}} \le \gamma^{\lfloor (L-3)/3 \rfloor} |||Q_x v|||^2_{\Omega\setminus\omega_x^0} \le \gamma^{\lfloor (L-3)/3 \rfloor} |||Q_x v|||^2. \end{aligned} \tag{A.17}$$

Combing (A.9), (A.11), and, (A.17) we get

$$|||e|||^2 \lesssim L^d \gamma^{\lfloor (L-3)/3 \rfloor} \sum_{x \in \mathcal{N}_I} |||Q_x v|||^2 \tag{A.18}$$

which concludes the lemma. □

References

[1] R. A. Adams. *Sobolev spaces*. Academic Press, 1975. Pure and Applied Mathematics, Vol. 65.

[2] I. Babuška, G. Caloz, and J. E. Osborn. Special finite element methods for a class of second order elliptic problems with rough coefficients. *SIAM J. Numer. Anal.*, 31(4):945–981, 1994.

[3] I. Babuška and J. E. Osborn. Can a finite element method perform arbitrarily badly? *Math. Comp.*, 69(230):443–462, 2000.

[4] D. Brown and D. Peterseim. A multiscale method for porous microstructures. *ArXiv e-prints*, Nov. 2014. Also available as INS Preprint No. 1410.

[5] E. Burman, S. Claus, P. Hansbo, M. G. Larson, and A. Massing. Cutfem: Discretizing geometry and partial differential equations. *International Journal for Numerical Methods in Engineering*, 2014.

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

[7] 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.

[8] 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.

[9] D. Elfverson, V. Ginting, and P. Henning. On multiscale methods in petrovgalerkin formulation. *Numerische Mathematik*, pages 1–40, 2015.

[10] T.-P. Fries and T. Belytschko. The extended/generalized finite element method: an overview of the method and its applications. *Internat. J. Numer. Methods Engrg.*, 84(3):253–304, 2010.

[11] T. Y. Hou and X.-H. Wu. A multiscale finite element method for elliptic problems in composite materials and porous media. *J. Comput. Phys.*, 134(1):169–189, 1997.

[12] 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.

[13] 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(21-24):2313–2324, 2007.

[14] A. Målqvist. Multiscale methods for elliptic problems. *Multiscale Model. Simul.*, 9(3):1064–1086, 2011.

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

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

[17] H. Owhadi, L. Zhang, and L. Berlyand. Polyharmonic homogenization, rough polyharmonic splines and sparse super-localization. *ESAIM Math. Model. Numer. Anal.*, 48(2):517–552, 2014.

[18] C. Pechstein and R. Scheichl. Weighted Poincaré inequalities. *IMA J. Numer. Anal.*, 33(2):652–686, 2013.

[19] D. Peterseim. Eliminating the pollution effect in Helmholtz problems by local subscale correction. *ArXiv e-prints*, Nov. 2014. Also available as INS Preprint No. 1411.

[20] R. Scott. Interpolated boundary conditions in the finite element method. *SIAM J. Numer. Anal.*, 12:404–427, 1975.

[21] R. Verfürht. *A review of a posteriori error estimation and adaptive mesh-refinement techniques.* Advances in numerical mathematics. Wiley, 1996.

Paper VI

SIAM/ASA J. UNCERTAINTY QUANTIFICATION
Vol. 2, pp. 826–850

Uncertainty Quantification for Approximate p-Quantiles for Physical Models with Stochastic Inputs*

Daniel Elfverson†, Donald J. Estep‡, Fredrik Hellman†, and Axel Målqvist§

Abstract. We consider the problem of estimating the p-quantile for a given functional evaluated on solutions of a deterministic model in which model input is subject to stochastic variation. We derive upper and lower bounding estimators of the p-quantile. We perform an a posteriori error analysis for the p-quantile estimators that takes into account the effects of both the stochastic sampling error and the deterministic numerical solution error and yields a computational error bound for the estimators. We also analyze the asymptotic convergence properties of the p-quantile estimator bounds in the limit of large sample size and decreasing numerical error and describe algorithms for computing an estimator of the p-quantile with a desired accuracy in a computationally efficient fashion. One algorithm exploits the fact that the accuracy of only a subset of sample values significantly affects the accuracy of a p-quantile estimator resulting in a significant gain in computational efficiency. We conclude with a number of numerical examples, including an application to Darcy flow in porous media.

Key words. p-quantile, a posteriori error bound, stochastic parameters, physical models, selective refinement

AMS subject classifications. 65N15, 65N30, 65C05, 65C30

DOI. 10.1137/140967039

1. Introduction.

The general setting for this paper is the computation of information about a given functional evaluated on solutions of a deterministic model in which model input, e.g., parameters and/or data, is subject to stochastic variation, e.g., arising from experimental error. If we assume the stochastic input is a random vector associated with a given probability space and typical conditions on the continuity of the model solution, the output functional is a random variable associated with the induced probability measure. In this case, the goal is

*Received by the editors April 29, 2014; accepted for publication (in revised form) November 4, 2014; published electronically December 23, 2014.

http://www.siam.org/journals/juq/2/96703.html

†Department of Information Technology, Uppsala University, Uppsala, SE-751 05, Sweden (daniel.elfverson@it.uu.se, fredrik.hellman@it.uu.se). The first author's research was supported by the Swedish Research Council and the Göran Gustafsson Foundation. The third author's research was supported by the Centre for Interdisciplinary Mathematics (CIM).

‡Department of Statistics, Colorado State University, Fort Collins, CO 80523 (estep@stat.colostate.edu). This author's research was supported in part by the Defense Threat Reduction Agency (HDTRA1-09-1-0036), the Department of Energy (DE-FG02-04ER25620, DE-FG02-05ER25699, DE-FC02-07ER54909, DE-SC0001724, DE-SC0005304, INL00120133, DE0000000SC9279), the Dynamics Research Corporation (PO672TO001), Idaho National Laboratory (00069249, 00115474), Lawrence Livermore National Laboratory (B573139, B584647, B590495), the National Science Foundation (DMS-0107832, DMS-0715135, DGE-0221595003, MSPA-CSE-0434354, ECCS-0700559, DMS-1065046, DMS-1016268, DMS-FRG-1065046, DMS-1228206), and the National Institutes of Health (R01GM096192).

§Department of Mathematical Sciences, Chalmers University of Technology and University of Gothenburg, SE-412 96 Göteborg, Sweden (axel@chalmers.se). This author's research was supported by the Swedish Research Council and the Göran Gustafsson Foundation.

to compute information about the stochastic properties of the output quantity.

As a concrete example, we consider an elliptic partial differential equation modeling incompressible single-phase Darcy flow in porous media. The problem is posed on a fixed domain with specified boundary conditions and with a stochastic permeability field. The output functional is an integral of the normal flux of the pressure on one segment of the boundary of the domain of the problem.

A common numerical problem in this setting is the approximation of the cumulative distribution function for the output functional. But there are other important statistical quantities that may be targeted. In this paper, we consider the problem of estimating the p-quantile for the output quantity. Quantiles, such as the median, provide important statistical information about complex probability distributions. For example, they are used in formulating engineering problems involving failure probabilities and they are important in a number of hypothesis tests. Quantiles are also relatively insensitive to the effects arising from a long-tailed distribution (a form of heavy-tailed distribution) and outliers in data, which makes them useful measures in those situations [11].

There are two primary sources of error affecting a p-quantile estimator in a practical setting, namely, finite sampling and numerical solution error. In a Monte Carlo approach, we compute a p-quantile estimate using model solutions for a finite sample of input parameter values chosen at random. Moreover, the typical physical model must be solved numerically, which means that the sample model values are only approximations of the true model outputs. These two sources of error have a complex interdependency, with numerical errors of sample solutions varying significantly as the input parameters vary in general.

Therefore, uncertainty quantification for the estimation of the p-quantile for a deterministic model with stochastic input involves not only computing a p-quantile estimate, but also estimating the effects of finite sampling and numerical solution on the accuracy of a p-quantile estimator. That is the subject of this paper. In particular, the main goal of this paper is a posteriori error analysis for a p-quantile estimator that takes into account the effects of both the stochastic error arising from finite sampling and the deterministic error arising from numerical solution of the model and yields a computational error bound for the estimator.

In [8, 9], we carry out the analogous a posteriori error analysis for an approximate cumulative distribution function. However, the fact that the p-quantile is determined by an inequality condition on the cumulative distribution function complicates analysis of the effects of numerical sample error on the accuracy of an estimator. Our approach involves computing upper and lower bounding quantities for the p-quantile that individually are estimators. The difference between the bounds provides an estimate of the accuracy of either estimator.

The model treatment is carried out on an abstract level, requiring only a computational a posteriori bound on the error of any given numerical solution that can be made arbitrarily small by suitable adjustment of discretization parameters. Under general assumptions, we analyze the asymptotic convergence properties of the p-quantile estimator bounds in the limit of large sample size and decreasing numerical error. We also describe two algorithms for computing an estimate of the p-quantile with a desired accuracy in a computationally efficient fashion, i.e., by approximately minimizing the number of samples and maximizing the sample error while still achieving the desired accuracy. One algorithm exploits the fact that the accuracy of only a subset of sample values significantly affects the accuracy of a p-quantile

estimator. Under the assumption of a model for computational "work," we show that this algorithm leads to a significant gain in computational efficiency. Finally, we investigate the performance of the p-quantile bounding estimators as well as various issues affecting the accuracy of the bounds in a set of numerical examples.

The paper is organized as follows. In section 2 we set up the problem, and in section 3 we derive error bounds for the approximate cumulative distribution function useful for our purposes. Section 4 presents the main theoretical results, giving the bounding estimators of the p-quantile and the error analysis for the estimators. Section 5 is devoted to presenting and analyzing algorithms for computing p-quantile estimates of a desired accuracy in an efficient way. We present some observations about p-quantile estimates in section 6. We present numerical examples in section 7. Finally, we present proofs of several results in section 8.

2. Problem formulation. The deterministic model is expressed as

$$M(u;\omega) = 0,$$

where $\omega \in \Omega$ is a vector of parameters and/or data valued in domain Ω, and $u = u(\omega)$ denotes the solution of the model. We assume the model has a unique solution for a given parameter value and also assume continuous dependence on the parameter values in Ω. Note that the model solution may also depend on other data or parameters that are held fixed. We let V denote the solution space of the model. In a common situation, M is an integral or differential equation and V is an appropriate Sobolev space. We assume that the object of solving the model is to compute a specified *Quantity of Interest (QoI)* expressed as a continuous (non)linear functional $Q : V \to \mathbb{R}$. We set $x(\omega) = Q(u(\omega))$, which is a continuous function of ω. We note that in the case of a differential equation in space and/or time, the application of the functional removes all explicit dependence on the independent variables other than the parameters.

We assume that Ω is the sample space for a probability space (Ω, Σ, P). This implies that the output $X(\omega) = Q(u(\omega))$ is a real-valued random variable with the induced measure on the Borel σ-algebra of $\mathbb{R}$. We let $F(x)$ denote the cumulative distribution function associated with X, and the p-quantile y is defined as

$$y = F^{-1}(p) = \inf\{x \in \mathbb{R} : F(x) \geq p\}.$$

We seek an estimator of y, along with a computable bound on the accuracy of the estimator.

As an example, we consider a model for incompressible single-phase Darcy flow for the pressure field u,

$$-\nabla \cdot A(\omega)\nabla u = 0, \quad x \in \mathscr{D}, \tag{2.1}$$

posed on the unit square $\mathscr{D} = [-1, 1] \times [-1, 1]$ with specified boundary conditions. The QoI is the normal flux through the left-hand boundary Γ_1,

$$Q(u(\omega)) = \int_{\Gamma_1} n \cdot A(\omega)\nabla u \, \mathrm{d}s.$$

We assume a stochastic permeability field $k : \mathscr{D} \to \mathbb{R}$ constructed using Layer 30 of the Society of Petroleum Engineering comparative permeability data (which are available online

from http://www.spe.org/web/csp). We introduce a conforming triangulation $\mathcal{T}^{h_0}$ of $\mathscr{D}$, with elements having diameter $h_0 = 0.2$ and vertices $p_j \in \mathcal{N}_0$. We let $\omega = (\omega_1, \ldots, \omega_{N_0})$ be a vector of independent random variables of standard normal distribution ($\mathcal{N}(0,1)$), where N_0 is the number of points in $\mathcal{N}_0$. For $j = 1, \ldots, N_0$, we let λ_j denote the linear Lagrange basis function for which $\lambda_j(p_\ell) = \delta_{j\ell}$, $\ell = 1, \ldots, N_0$. We define

$$A(\omega, \mathcal{N}_0) = A_0 + \sum_{j=1}^{N_0} e^{\omega_j} k(p_j) \lambda_j,$$

where $0 < A_0$ is chosen to guarantee coercivity. Thus, A is a continuous, piecewise linear polynomial on $\mathscr{D}$ that is affine on each $T \in \mathcal{T}^{h_0}$.

To estimate the p-quantile, we employ a finite number of random approximate sample values. Thus, the accuracy of the p-quantile estimate is affected both by stochastic sampling error and deterministic numerical error. We let $\{\omega_i\}_{i=1}^n$ be an independent and identically distributed (i.i.d.) sample of size n from Ω, for which the true QoIs are $x_i = Q(u(\omega_i))$ for $i = 1, \ldots, n$. We assume that numerical approximations $x_i^\epsilon = Q(u^\epsilon(\omega_i))$ are computed by solving an approximate model,

$$M^\Delta(u^\epsilon(\omega_i), \omega_i) = 0,$$

for an approximate solution $u^\epsilon(\omega_i) \approx u(\omega_i)$, where Δ denotes some discretization parameter. We assume that the error of the approximate value x_i^ϵ can be made as small as desired by adjusting Δ.

The computational problem we address is as follows: Given p and $0 < \beta < 1$, find computable bounds $y_{n,\epsilon}^-$ and $y_{n,\epsilon}^+$ for y such that

$$\Pr\left(y \in [y_{n,\epsilon}^-, y_{n,\epsilon}^+]\right) > 1 - \beta,$$

for all n sufficiently large and ϵ sufficiently small, and

$$y_{n,\epsilon}^- \to y, \quad y_{n,\epsilon}^+ \to y \text{ as } n \to \infty, \epsilon \to 0.$$

We note that the error of any estimator $\hat{y}_{n,\epsilon}$ satisfying $y_{n,\epsilon}^- \le \hat{y}_{n,\epsilon} \le y_{n,\epsilon}^+$ of y is bounded,

$$\Pr(|y - \hat{y}_{n,\epsilon}| \le |y_{n,\epsilon}^+ - y_{n,\epsilon}^-|) > 1 - \beta,$$

which provides the desired estimate on the accuracy of any such estimator.

3. Error analysis of the approximate cumulative distribution function. Computing the p-quantiles estimates involves computing approximate cumulative distribution functions (cdfs) using a finite number of samples of approximate model solutions. The error in the approximate cdf in turn affects the accuracy of the p-quantile estimates.

We begin by decomposing the error of a computed cdf into statistical and numerical contributions by introducing the empirical distribution function,

$$F_n(x) = \frac{\#\{i = 1, \ldots, n : x_i \le x\}}{n}, \quad x \in \mathbb{R},$$

and its numerical approximation,

$$F_{n,\epsilon}(x) = \frac{\#\{i = 1, \ldots, n : x_i^\epsilon \leq x\}}{n}, \quad x \in \mathbb{R},$$

where # denotes cardinality. The error decomposition is then

$$F(x) - F_{n,\epsilon}(x) = \underbrace{F(x) - F_n(x)}_{\text{statistical error}} + \underbrace{F_n(x) - F_{n,\epsilon}(x)}_{\text{numerical error}}.$$

We note that F_n cannot be computed.

3.1. Bounds on the statistical error contribution. The nature of the error introduced by stochastic sampling means that we employ an asymptotic bound rather than an a posteriori estimate in the sense used for differential equations. There are a number of ways to derive such bounds [8, 9]. The statistical bounds needed in this paper are formulated in the following assumption.

Assumption 3.1 (computable bound on statistical error). There exist a positive continuous function $G : [0, 1] \to \mathbb{R}$ and constant $\widetilde{C}_1 > 0$, independent of x and n, such that for any given $0 < \beta < 1$,

$$\Pr\left(|F(x) - F_n(x)| \leq G(F_n(x))n^{-1/2} + \widetilde{C}_1 n^{-1}\right) > 1 - \beta/2 \tag{3.1}$$

for $x \in \mathbb{R}$ for all n sufficiently large.

The $\widetilde{C}_1 n^{-1}$ is generally required in order to derive a bound independent of the unknown cdf. We note that (3.1) implies that there is a constant C_1 such that

$$G(F_n(x))n^{-1/2} + \widetilde{C}_1 n^{-1} \leq C_1 n^{-1/2}. \tag{3.2}$$

We actually need the following assumption.

Lemma 3.2. *Under Assumption* 3.1*,* (3.1) *holds for any two points* $x_1, x_2 \in \mathbb{R}$ *simultaneously with probability* $1 - \beta$*.*

Proof. This is a consequence of Bonferroni's inequality $\Pr(E_1 \cap E_2) \geq \Pr(E_1) + \Pr(E_2) - 1$ for two events E_1 and E_2. Let E_1 and E_2 be the events that (3.1) is satisfied pointwise at two points x_1 and x_2 with confidence level for (3.1) such that $\Pr(E_1) = \Pr(E_2) = 1 - \beta/2$. Bonferroni's inequality implies (3.1) holds with simultaneous probability at least $1 - \beta$. ∎

A standard way to derive (3.1) uses the fact that the distribution of $nF_n(x)$ is binomial. Consequently, Chebyshev's inequality implies

$$\Pr\left(|F_n(x) - F(x)| \geq kn^{-1/2}F(x)^{1/2}(1 - F(x))^{1/2}\right) \leq 1/k^2.$$

We use the expansion $F(x)(1 - F(x)) = F_n(x)(1 - F_n(x)) + (F(x) - F_n(x))$; then we set $G(q) = (2/\beta)^{1/2}q^{1/2}(1 - q)^{1/2}$ and $\widetilde{C}_1 = 2\beta^{-1}$.

Alternatively, we can use the DKW inequality [4], which states that for all $K > 0$,

$$\Pr\left(\sup_{x \in \mathbb{R}} |F_n(x) - F(x)| > Kn^{-1/2}\right) \leq 2e^{-2K^2}. \tag{3.3}$$

This is a uniform confidence bound, and we let $G(q) = \sqrt{2^{-1}\ln(2/\beta)}$ and $\widetilde{C}_1 = 0$.

Assumption 3.1 defines an interval for F that is symmetric around F_n. We can also handle an "asymmetric" interval. We now assume there is an affine transformation $T : \mathbb{R} \to \mathbb{R}$ such that

$$|F(x) - T(F_n(x))| \le G(T(F_n(x)))n^{-1/2} + \widetilde{C}_1 n^{-1}. \tag{3.4}$$

Any subsequent results for F_n or any numerical approximation that depend on Assumption 3.1 hold for T applied to F_n or any approximation.

For example, the Agresti–Coull interval [1, 3] is an asymmetric approximate confidence interval for binomial parameter $F(x)$ that is recommended over other common bounds. It reads as

$$\Pr\left(|F(x) - \tilde{p}| \le \kappa \tilde{p}^{1/2}(1-\tilde{p})^{1/2}\tilde{n}^{-1/2}\right) > 1 - \beta/2, \tag{3.5}$$

with $\kappa = \Phi^{-1}(1-\beta/4)$, $\Phi(z) \sim N(0,1)$, $\tilde{n} = n + \kappa^2$, and $\tilde{p} = (nF_n(x) + \frac{1}{2}\kappa^2)/\tilde{n}$. We let $T(F_n(x)) = \tilde{p}$ and define $G(q) = \kappa q^{1/2}(1-q)^{1/2}$ and $\widetilde{C}_1 = 0$ to satisfy (3.4).

3.2. Bounds on the numerical error contribution. Depending on how approximate solutions of the physical model are computed, there are generally several approaches for computing estimates and bounds on the error of computed information obtained from a numerical solution. We assume the following.

Assumption 3.3 (computable bound on QoI). There is a computational procedure for computing a numerical bound ϵ_i for each sample numerical solution x_i, $i = 1, \ldots, n$, such that

$$|x_i - x_i^\epsilon| \le \epsilon_i, \tag{3.6}$$

where ϵ_i can be made as small as desired by adjusting Δ.

We discuss a particular approach for computing numerical error estimates and bounds in section 6.

3.3. Error bounds for the approximate cdf. We now derive error estimates for various approximate numerical cdfs. The central issue is that error in the sample values leads to miscounts in the computation of the cdf. The following two approximate cdfs can be considered "worst case" approximations:

$$F^-_{n,\epsilon}(x) = \frac{\#\{i = 1, \ldots, n : x_i^\epsilon + \epsilon_i \le x\}}{n}, \quad F^+_{n,\epsilon}(x) = \frac{\#\{i = 1, \ldots, n : x_i^\epsilon - \epsilon_i \le x\}}{n}.$$

These definitions assume that the errors always have the disadvantageous sign and are the size of the bounding quantities. However, we note that only the values of the samples in a relatively small region affect the computation of p-quantile estimates.

We define the computable bound on the statistical error contribution,

$$\mathcal{E}^{\text{stat}}_{n,\epsilon}(x) = \max_{F^-_{n,\epsilon}(x) \le q \le F^+_{n,\epsilon}(x)} G(q)n^{-1/2} + \widetilde{C}_1 n^{-1}, \tag{3.7}$$

and the computable bound on the numerical error contribution,

$$\mathcal{E}^{\text{num}}_{n,\epsilon}(x) = \max\left(F^+_{n,\epsilon}(x) - F_{n,\epsilon}(x), F_{n,\epsilon}(x) - F^-_{n,\epsilon}(x)\right). \tag{3.8}$$

These definitions yield the following theorem.

Theorem 3.4 (bound on the error in the cdf). *Under Assumptions* 3.1 *and* 3.3, *given* $0 < \beta < 1$, *for any two* $x_j \in \mathbb{R}$, $j = 1, 2$,

$$\Pr\left(|F(x) - F_{n,\epsilon}(x_j)| \leq \mathcal{E}^{\text{stat}}_{n,\epsilon}(x_j) + \mathcal{E}^{\text{num}}_{n,\epsilon}(x_j)\right) > 1 - \beta \tag{3.9}$$

for all sufficiently large n.

Proof. Since for every $x \in \mathbb{R}$ the number of elements in $\{x_i^\epsilon + \epsilon_i\}_{i=1}^n$ less than x is smaller than or equal to the number of elements in $\{x_i\}_{i=1}^n$ less than x, $F_n(x) \geq F^-_{n,\epsilon}(x)$. Using a similar argument, we conclude that $F_n(x) \leq F^+_{n,\epsilon}(x)$. Therefore, $|F_n(x) - F_{n,\epsilon}(x)| \leq \mathcal{E}^{\text{num}}_{n,\epsilon}(x)$. Next, we combine Lemma 3.2, (3.8), and these inequalities to reach (3.9). ∎

4. p-quantile bounding estimators and convergence rates. In this section, we derive computable bounds for the p-quantile which are used as estimators. We use the notation $\epsilon = (\epsilon_i)_{i=1}^n$, $\epsilon_{\max} = \max_{i=1,\ldots,n} \epsilon_i$, $\epsilon_{\min} = \min_{i=1,\ldots,n} \epsilon_i$. We analyze the convergence properties of the bounds in the limits $\epsilon_{\max} \to 0$ and $n \to \infty$.

4.1. Computable error bounds for the p-quantile. The two bounding estimators handle the "worst case" scenario,

$$y^+_{n,\epsilon} = \inf\{x \in \mathbb{R} : F^-_{n,\epsilon}(x) - \mathcal{E}^{\text{stat}}_{n,\epsilon}(x) \geq p\}, \quad y^-_{n,\epsilon} = \inf\{x \in \mathbb{R} : F^+_{n,\epsilon}(x) + \mathcal{E}^{\text{stat}}_{n,\epsilon}(x) \geq p\}. \tag{4.1}$$

With these definitions, we have the following theorem.

Theorem 4.1 (existence of the p-quantile bounding estimators). *The computable quantities* $y^+_{n,\epsilon}, y^-_{n,\epsilon}$ *exist, and given* $0 < \beta < 1$,

$$\Pr\left(y \in [y^-_{n,\epsilon}, y^+_{n,\epsilon}]\right) > 1 - \beta$$

for all sufficiently large n.

Proof. We define $Y = \{x \in \mathbb{R} : F^-_{n,\epsilon}(x) - \mathcal{E}^{\text{stat}}_{n,\epsilon}(x) \geq p\}$. We start by showing that Y is nonempty and $\inf Y$ exists. The assumption on n implies $\mathcal{E}^{\text{stat}}_{n,\epsilon}(x) \leq C_1 n^{-1/2} < 1 - p$ for all x. For a fix n, and for all $x > \max_{i=1,\ldots,n}(x_i^\epsilon + \epsilon_i)$, we have $F^-_{n,\epsilon}(x) = 1$ and $F^-_{n,\epsilon}(x) - \mathcal{E}^{\text{stat}}_{n,\epsilon}(x) > 1 - 1 + p = p$, rendering Y nonempty. Since $p > 0$, $\mathcal{E}^{\text{stat}}_{n,\epsilon}$ is nonnegative, $F^-_{n,\epsilon}$ is nondecreasing, and $F^-_{n,\epsilon}(x) = 0$ for some finite x, we can conclude that Y is bounded from below, implying $y^+_{n,\epsilon} = \inf Y$ exists. Further, Theorem 3.4 and the inequalities used in its proof apply to $y^+_{n,\epsilon}$, and we conclude that $y \leq y^+_{n,\epsilon}$ from

$$y = \inf\{x \in \mathbb{R} : F(x) \geq p\} \leq \inf\{x \in \mathbb{R} : F^-_{n,\epsilon}(x) - \mathcal{E}^{\text{stat}}_{n,\epsilon}(x) \geq p\} = y^+_{n,\epsilon}.$$

Similarly, $y \geq y^-_{n,\epsilon}$. The results hold with probability greater than $1 - \beta$ for both bounds simultaneously from Theorem 3.4. ∎

The minimization problems (4.1) in Theorem 4.1 form the basis for a practically feasible procedure to compute the bounding p-quantile estimators $y^-_{n,\epsilon}$ and $y^+_{n,\epsilon}$; see section 4.3.

4.2. Convergence of the bounding p-quantile estimators. We next analyze the convergence properties of $y^+_{n,\epsilon}, y^-_{n,\epsilon}$. We define

$$y^- = \lim_{\eta\to 0^+} \inf\{x \in \mathbb{R} : F(x) + \eta \geq p\} \quad \text{and} \quad y^+ = \lim_{\eta\to 0^-} \inf\{x \in \mathbb{R} : F(x) + \eta \geq p\},$$

which bound the quantile, $y^- \leq y \leq y^+$, by definition. The lower bound y^- is actually equal to y. However, y^+ is not necessarily equal to y in the case when F is "flat." When $y^- \neq y^+$, the problem of finding y is ill-conditioned, since small perturbations in the data p or F cause large variations in the solution y and the quantile bounds converge to either y^- or y^+, or cycles between them, as n approaches infinity and numerical error approaches zero (see [10]).

On the other hand, when the p-quantile is unique (i.e., $y^- = y = y^+$) and F is continuous, then we have the following theorem.

Theorem 4.2 (convergence of the bounding p-quantile estimators). *If F is continuous, then with probability* 1*,*

$$\min(|y^+_{n,\epsilon} - y^+|, |y^+_{n,\epsilon} - y^-|) \to 0 \text{ and } \min(|y^-_{n,\epsilon} - y^+|, |y^-_{n,\epsilon} - y^-|) \to 0$$

as $n \to \infty$ and $\epsilon \to 0$.

The proof is given in section 8.

Furthermore, for unique p-quantiles, we have the following asymptotic convergence rate result, proved in section 8.

Theorem 4.3 (convergence rate of the bounding p-quantile estimators). *For a fixed $n > 0$ and $0 < p < 1$, choose $K > 0$ such that $(K + C_1)n^{-1/2} < p < 1 - (K + C_1)n^{-1/2}$; then if F is absolutely continuous and $\ell = \inf_{\{x\in\mathbb{R}:|F(x)-p|\leq(K+C_1)n^{-1/2}\}} F'(x) > 0$, we have*

$$|y^+_{n,\epsilon} - y^-_{n,\epsilon}| \leq 2\ell^{-1}(K + C_1)n^{-1/2} + 4\epsilon_{\max}$$

with probability at least $1 - 2e^{-2K^2}$.

4.3. An algorithm for computing the bounding p-quantile estimates. We describe how $y^-_{n,\epsilon}$ and $y^+_{n,\epsilon}$ can be computed in practice. We first note that the functions $F^-_{n,\epsilon}$, $F^+_{n,\epsilon}$ are piecewise constant on $n + 1$ intervals. From (3.7), we observe that $\mathcal{E}^{\text{stat}}_{n,\epsilon}$ has discontinuities only at the points of discontinuity of $F^-_{n,\epsilon}$ and $F^+_{n,\epsilon}$ and hence is piecewise constant on at most $2n + 1$ intervals. The sums $F^+_{n,\epsilon} + \mathcal{E}^{\text{stat}}_{n,\epsilon}$ and $F^-_{n,\epsilon} - \mathcal{E}^{\text{stat}}_{n,\epsilon}$ have $2n + 1$ intervals of constant value to be searched when solving (4.1). The procedure is described in Algorithm 1. Note that the conditions in Theorem 4.1 need to hold for the obtained values in Algorithm 1 to make sense (or even exist). The computational time complexity is dominated by sorting and is $\mathcal{O}(n \log n)$.

Algorithm 1. Algorithm for computing the bounding p-quantile estimates.

1: Let $z = (z_i)_{i=1}^{2n} \leftarrow$ sort $(x^\epsilon_1 + \epsilon_1, x^\epsilon_1 - \epsilon_1, \ldots, x^\epsilon_n + \epsilon_n, x^\epsilon_n - \epsilon_n)$ (requires sorting $2n$ values)
2: Compute $F^+_{n,\epsilon}$ and $F^-_{n,\epsilon}$ at all points in z (requires sorting n values twice)
3: Compute $\mathcal{E}^{\text{stat}}_{n,\epsilon}$ at all points in z (using $F^+_{n,\epsilon}$ and $F^-_{n,\epsilon}$ at z)
4: Let $y^-_{n,\epsilon} \leftarrow$ smallest z_i for which $F^+_{n,\epsilon}(z_i) + \mathcal{E}^{\text{stat}}_{n,\epsilon}(z_i) \geq p$
5: Let $y^+_{n,\epsilon} \leftarrow$ smallest z_i for which $F^-_{n,\epsilon}(z_i) - \mathcal{E}^{\text{stat}}_{n,\epsilon}(z_i) \geq p$

5. Algorithms for control of the error of the bounding p-quantile estimators. In a practical situation, an important goal is to determine the number of samples and the accuracy of the samples required to guarantee a given level of accuracy, i.e., $|y^+_{n,\epsilon} - y^-_{n,\epsilon}| \leq \text{TOL}$, in a computationally efficient way. By computational efficiency, we mean that the numerical samples should not be overly accurate and the number of numerical samples should not be overly large.

Equation (3.9) shows that the bound on the error in the cdf is decomposed in terms of $\mathcal{E}^{\text{stat}}_{n,\epsilon}$ and $\mathcal{E}^{\text{num}}_{n,\epsilon}$. However, such a decomposition cannot be perfect. This complicates the selection of the number of samples and the accuracy of each sample. Since a priori selection is difficult, an a posteriori approach is employed. Such an approach is based on the following cycle: the computation of an estimate, the estimation of the accuracy of the computed estimate, and adjustment of computational parameters for the next cycle. There are a number of ways to organize an algorithm for controlling the error following this basic idea.

From the definitions in (3.8) it is apparent that the statistical error bound $\mathcal{E}^{\text{stat}}_{n,\epsilon}$ can be bounded independently of ϵ, so a value of n can be determined a priori. With this choice, we can use a computational error estimate on the error of the approximate samples to achieve a "balance" in the stochastic and deterministic contributions to the error. The following theorem shows that balancing the error indicators lead to a p-quantile interval length dependent only on n.

Theorem 5.1. *Given ϵ such that $\mathcal{E}^{\text{num}}_{n,\epsilon}(x) - \mathcal{E}^{\text{stat}}_{n,\epsilon}(x) \leq 0$ for all $y^-_{n,\epsilon} \leq x \leq y^+_{n,\epsilon}$ and $n > 9C_1^2 \max\left((1-p)^{-2}, p^{-2}\right)$, it holds that*

$$y^+_{n,\epsilon} - y^-_{n,\epsilon} \leq F_n^{-1}(p + 3C_1 n^{-1/2}) - F_n^{-1}(p - 3C_1 n^{-1/2}).$$

The proof is given in section 8.

In a practical procedure to reach a specified error tolerance TOL, an initial n is chosen and the numerical error tolerance parameters ϵ are reduced until the balance condition $(\mathcal{E}^{\text{num}}_{n,\epsilon}(x) - \mathcal{E}^{\text{stat}}_{n,\epsilon}(x) \leq 0)$ in Theorem 5.1 is satisfied. The p-quantile interval length is then checked against the tolerance, and possibly a larger n is chosen. We now focus only on the problem of finding ϵ for balancing the two error indicators at minimal computational cost, given a fixed n.

5.1. A full refinement algorithm for control of sample accuracy. We first present a straightforward algorithm for computing approximate p-quantile bounding estimates to within a prescribed accuracy. The algorithm employs a sequence of refinements, by which we mean the discretization actions required to decrease the numerical error estimate or bound. For example, refinement of a realization might be mesh refinement of the discretization for that realization.

The convergence rate result in Theorem 4.3 is based on uniform refinement for all realizations so that $\epsilon_{\max} \to 0$. Using the balance criterion in Theorem 5.1 as a termination criterion, we construct an algorithm which refines all realizations to the same numerical error tolerance in each iteration. This *full refinement algorithm* is given in Algorithm 2. To state the algorithm, we use δ to denote another vector of numerical error tolerance parameters when two different vectors are needed simultaneously. Approximate quantities based on δ instead of ϵ are indicated with a superscript δ.

The full refinement algorithm refines all realizations to the same numerical error tolerance

Algorithm 2. Algorithm for full refinement.

Pick p, β, n, and δ_{init}
Set $\delta = (\delta_{\text{init}}, \dots, \delta_{\text{init}})$
Compute x_i^δ satisfying Assumption 3.3 for all $i = 1, \dots, n$
Let $j = 0$ be an iteration counter
while $\sup_{x \in [y^-_{n,\delta}, y^+_{n,\delta}]} \left(\mathcal{E}^{\text{num}}_{n,\delta}(x) - \mathcal{E}^{\text{stat}}_{n,\delta}(x) \right) > 0$ **do**
    Set $j \leftarrow j + 1$
    Set $\delta_i \leftarrow \delta_{\text{init}} 2^{-j}$ for all $i = 1, \dots, n$
    Recompute x_i^δ (satisfying Assumption 3.3) for all $i = 1, \dots, n$
    Save $\delta^{(j)} \leftarrow \delta$
end while

$\delta_{\text{init}} 2^{-j}$ in each iteration j until the errors are balanced. Here δ_{init} is the initial numerical error tolerance. Following the algorithm listing, initially all n numerical error tolerance parameters δ are set to δ_{init}. Before entering the main loop, n realizations are generated satisfying Assumption 3.3 with the initial numerical error tolerance. The balance criterion $\mathcal{E}^{\text{num}}_{n,\delta}(x) - \mathcal{E}^{\text{stat}}_{n,\delta}(x) \le 0$ is checked and the main loop is entered if it is not satisfied. In each iteration, all realizations are refined to the same numerical error tolerance $\delta_{\text{init}} 2^{-j}$, where j is the iteration number. Then x_i^δ are recomputed before checking the termination criterion again.

5.2. A selective refinement algorithm for control of sample accuracy. The second algorithm is based on the observation that it is not necessary to refine all realizations as called for in the full refinement algorithm. The bound $|y^+_{n,\epsilon} - y^-_{n,\epsilon}|$ can be made as small as desired even when there is a significant number of realizations that have a large numerical error bound ϵ_i. In each iteration in the full refinement algorithm, it is possible to identify the set of realizations whose accuracy may affect the interval $[y^-_{n,\epsilon}, y^+_{n,\epsilon}]$, while the complement of this set consists of realizations with no potential to affect the interval. Hence, only a subset of the realizations needs to be considered for further refinement in each iteration. We propose a *selective refinement algorithm*, Algorithm 3, that exploits this fact. The criterion for a realization to be refined is that any further refinement of the realization *might* affect the interval $[y^-_{n,\epsilon}, y^+_{n,\epsilon}]$, which can be determined computationally.

The following theorem shows that the result of selective refinement is at least as accurate as the result of full refinement at the same iteration count. We assume without loss of generality that the QoI values and numerical error tolerances are scaled so that $\delta_{\text{init}} = \epsilon_{\text{init}} = 1$.

Theorem 5.2. *For any $0 < p < 1$ and $j \in \mathbb{N}$, if we let $\delta = \delta^{(j)} = (2^{-j}, \dots, 2^{-j})$ from Algorithm 2 and $\epsilon = \epsilon^{(j)}$ from Algorithm 3 (with $\epsilon_{\min} = 2^{-j}$), then*

$$y^-_{n,\delta} \le y^-_{n,\epsilon} \quad \text{and} \quad y^+_{n,\epsilon} \le y^+_{n,\delta}.$$

As a direct consequence, Theorem 4.3 holds with $\epsilon_{\max}$ replaced by $\epsilon_{\min}$ for ϵ chosen according to Algorithm 3.

The proof is presented in section 8.

It is easy to see that selective refinement always performs fewer refinements than full refinement. In the cases where the computations due to refinements are the dominant part

Algorithm 3. Algorithm for selective refinement.

Pick p, β, n, and ϵ_{init}
Set $\epsilon = (\epsilon_{\text{init}}, \ldots, \epsilon_{\text{init}})$ and $I = \{1, \ldots, n\}$
Compute x_i^ϵ for all $i \in I$
Let $j = 0$ be an iteration counter
while $\sup_{x \in [y_{n,\epsilon}^-, y_{n,\epsilon}^+]} \left(\mathcal{E}_{n,\epsilon}^{\text{num}}(x) - \mathcal{E}_{n,\epsilon}^{\text{stat}}(x) \right) > 0$ **do**
    Set $j \leftarrow j + 1$
    Compute $I \leftarrow \{i = 1, \ldots, n : y_{n,\epsilon}^- - \epsilon_i < x_i^\epsilon \leq y_{n,\epsilon}^+ + \epsilon_i\} \cap I$
    Set $\epsilon_i \leftarrow \epsilon_{\text{init}} 2^{-j}$ for all $i \in I$
    Recompute x_i^ϵ (satisfying Assumption 3.3) for all $i \in I$
    Save $\epsilon^{(j)} \leftarrow \epsilon$, and $I^{(j)} = \{i = 1, \ldots, n : \epsilon_i^{(j)} = 2^{-j}\}$
end while

of the computational work, there is always a gain from using selective refinement. The next section is devoted to quantifying this gain.

5.3. Quantification of the gain in computational complexity by selective refinement. In order to quantify the gain in computational complexity in terms of n obtained by selective refinement in comparison to full refinement, we need an estimate of the work required by the two algorithms.

For this, we make an additional assumption,

Assumption 5.3 (model of work). The work W for computing x_i^ϵ satisfying (3.6) depends on the numerical error tolerance and satisfies

$$C_2 \epsilon_i^{-q} \leq W(\epsilon_i) \leq \epsilon_i^{-q}, \tag{5.1}$$

where $C_2 \leq 1$ and $q > 0$ are independent of i.

As motivation, consider the situation in which the QoI is a functional of a finite element solution to a d-dimensional elliptic partial differential equation. On a uniform mesh of maximum element size h, the accuracy of the solution is proportional to h^λ for some $\lambda > 0$. The linear system to produce the approximate solution is solved in linear time in the number of degrees of freedom N. The numerical error bound ϵ_i is determined by an a posteriori error bound for the functional value. Neglecting constants, we have $W \approx N$, $N \approx h^{-d}$, and $\epsilon_i \approx h^\lambda$, i.e., the work to compute a solution with accuracy ϵ_i is $W \approx \epsilon_i^{-d/\lambda}$, that is, with $q = d/\lambda$ in Assumption 5.3.

Inequality (5.1) implies there is a minimum amount of work $C_2 \epsilon_i^{-q}$ required to achieve a tolerance ϵ_i. It is possible to construct cases when there is no minimum work for specific realizations or a class of realizations. For example, for a differential equation with a piecewise linear finite element discretization, all realizations rendering solutions to the model that can be exactly represented in the discretization give no discretization error and hence require no additional work to achieve any lower numerical error tolerance. We assume that the class of such realizations occurs with probability 0.

In this analysis, the computations used for the refinement algorithm itself are not considered in the work estimate. This is motivated by the fact that the most computationally

demanding work (complexitywise) associated with the selective algorithm itself is computing $y^-_{n,\epsilon}$ and $y^+_{n,\epsilon}$ (see Algorithm 1). This amounts to sorting $\mathcal{O}(n^{1/2})$ number of elements (see Theorem 5.1) in each iteration, with a complexity of $\mathcal{O}(n^{1/2}\log(n))$. In each iteration, at least $\mathcal{O}(n^{1/2})$ realizations need to be refined and the amount of required work is $\mathcal{O}(n^{1/2}W(\epsilon_i))$. When the errors are balanced, $\mathcal{O}(\epsilon_i) = \mathcal{O}(n^{-1/2})$, the work for refining is $\mathcal{O}(n^{(1+q)/2})$. This means the work for the selective algorithm itself can be neglected for large n.

In Algorithm 3, the numerical error tolerance is reduced by a factor of two in each iteration, so that $\epsilon^{(j)} = 2^{-j}$ for iteration j. The amount of work $W^{(j)}$ performed in iteration $j = 0, 1, 2, \ldots$ in the algorithm is then (see Assumption 5.3)

$$W^{(j)} = W(2^{-j})\#I^{(j)}, \quad j = 0, 1, 2, \ldots.$$

Note that $\#I^{(0)} = n$. The work for an iteration in the full refinement algorithm is

$$\widehat{W}^{(j)} = W(2^{-j})n, \quad j = 0, 1, 2, \ldots.$$

The computational complexity for selective refinement compared to full refinement is given in Theorem 5.4.

Theorem 5.4. *For a fixed $n \geq 1$, if the cdf F associated to the QoI is Lipschitz continuous, and assuming that $J = \lceil \frac{1}{2}\log_2 n - \log_2 C_3 \rceil$ iterations are required for Algorithm 3 to terminate (see Remark 5.5), the ratio between the required work, $\sum_{j=0}^{J} W^{(j)}$, using selective selective refinement (Algorithm 3) and the required work, $\sum_{j=0}^{J} \widehat{W}^{(j)}$, using full refinement (Algorithm 2), is bounded above by*

$$\frac{\sum_{j=0}^{J} W^{(j)}}{\sum_{j=0}^{J} \widehat{W}^{(j)}} \leq \min\left(1, KC_4 \begin{cases} n^{-q/2} & \text{if } q < 1 \\ n^{-1/2}\log_2 n & \text{if } q = 1 \\ n^{-1/2} & \text{if } q > 1 \end{cases}\right) \tag{5.2}$$

with probability at least $1-2e^{-2K^2}$, where C_4 depends on the cdf F, the statistical error constant C_1 (3.2), the model of work constants C_2 and q (5.1), and the error balance constant C_3.

Proof. See section 8. ∎

Remark 5.5. The termination criterion is satisfied when $\epsilon_{\min} = C_3 n^{-1/2}$ (Theorem 4.3) for some constant C_3, depending on the specific sample, but not asymptotically on n. Then the number of iterations required to balance the error is $J = \lceil \frac{1}{2}\log_2 n - \log_2 C_3 \rceil$, since $\epsilon_{\min} = 2^{-J}$. If more iterations are required, selective refinement provides greater gain in the comparison to full refinement.

Remark 5.6. The rates in (5.2) are limited by the rate of convergence of $\mathcal{E}^{\text{stat}}_{n,\epsilon}$ in terms of n. If $\mathcal{E}^{\text{stat}}_{n,\epsilon} \leq C_1 n^{-1}$ through a different sampling technique, e.g., quasi Monte Carlo, then the rates can be replaced by n^{-q}, $n^{-1}\log_2 n$, and n^{-1} for the three cases, respectively. In the last case, this means the cost for Algorithm 3 is asymptotically independent of the number of realizations. The probability for the result to hold is also affected, since the DKW inequality has to be replaced by the improved confidence interval.

6. Some additional observations. In this section, we comment briefly on the use of a posteriori error estimates instead of bounds and the potential cancellation of errors in the cdf due to miscounts. In this section, we simplify notation by setting $\epsilon_i = \epsilon$. We denote the true (signed) error in the quantity of interest by e_i, i.e., $e_i = x_i - x_i^\epsilon$.

6.1. Using accurate error estimates instead of bounds for numerical sample error. There are approaches to error estimation that yield accurate error estimates $\bar{e}_i$ rather than bounds; i.e., for each sample numerical solution, $i = 1, \ldots, n$,

$$x_i - x_i^\epsilon \approx \bar{e}_i. \tag{6.1}$$

It is natural to consider the use of such estimates (6.1) in the estimation of the p-quantile. We discuss this briefly.

An important issue is that in practice, accurate error estimates are only approximations to the true error. Issues affecting accuracy of an error estimate include the fact that the derivation often involves neglecting terms that cannot be estimated (though may be provably smaller than the error) and because of various numerical approximations used in the computation of an estimate. Consequently, an estimate may be smaller or larger than the error. One difficulty in estimating the effects of sample errors on the computation of a p-quantile is the fact that small errors in sample values can lead to an $O(1)$ miscount in the computation of the cdf, which in turn affects the evaluation of the inequality defining the p-quantile. This is a main motivation for using an error bound on the error of each sample value in Assumption 3.3.

In many situations, it is possible to derive a bound on the accuracy of the error estimate of the form

$$|x_i - x_i^\epsilon - \bar{e}_i| \leq C(\bar{e}_i)^\lambda$$

for some constant C and λ depending on the accuracy of the error estimate. In this case, we can use the accurate error estimate to "correct" the approximate sample values, and exploit all of the previous analysis to define p-quantile bounds using $\{x_i^\epsilon + \bar{e}_i\}$ in place of $\{x_i^\epsilon\}$ and by setting $\epsilon = C(\bar{e}_i)^\lambda$. This results in a gain in computational efficiency, since we can expect to use a coarser discretization parameter Δ in the numerical approximation while still achieving the specified numerical error tolerance.

Accurate a posteriori error estimates can be used to define another p-quantile estimator. Specifically, the numerical error $|F_n(x) - F_{n,\epsilon}(x)|$ can be estimated by defining a "corrected" cdf, based on accurate a posteriori error estimates $\bar{e}_i$, such that $|e_i - \bar{e}_i| \leq \epsilon^\lambda$, for some $\lambda > 1$,

$$\bar{F}_{n,\epsilon}(x) = \frac{\#\{i = 1, \ldots, n : x_i^\epsilon + \bar{e}_i \leq x\}}{n}, \quad x \in \mathbb{R},$$

which generates a presumably more accurate numerical cdf. If the a posteriori error estimates are accurate and reliable, we can approximate

$$|F_n - F_{n,\epsilon}| \approx |\bar{F}_{n,\epsilon} - F_{n,\epsilon}|$$

and use the alternative definitions,

$$F_{n,\epsilon}^+(x) = \max(F_{n,\epsilon}(x), \bar{F}_{n,\epsilon}(x)), \quad F_{n,\epsilon}^-(x) = \min(F_{n,\epsilon}(x), \bar{F}_{n,\epsilon}(x)).$$

This gives

$$|F_{n,\epsilon} - \bar{F}_{n,\epsilon}| = |F_{n,\epsilon}^+ - F_{n,\epsilon}^-|.$$

Now we could use all of the results in the paper starting with these definitions.

6.2. The effect of miscount cancellation on the numerical error in the cdf. Up to this point, the only assumption used on the numerical error in the QoI is $e_i \leq \epsilon$. The following discussion shows there can be a miscount cancellation effect in the numerical error $|F_n(x) - F_{n,\epsilon}(x)|$ in the cdf.

We consider $e_i = e_i(\omega_i)$ to be a random variable and define $Y_i^\epsilon(x) = \mathbb{1}(x - x_i) - \mathbb{1}(x - x_i^\epsilon)$, where $\mathbb{1}(x)$ is zero for $x < 0$ and one for $x \geq 0$, and we note that

$$F_n(y) - F_{n,\epsilon}(y) = \frac{1}{n} \sum_{i=1}^{n} Y_i^\epsilon(y).$$

The random variable $Y_i^\epsilon(y)$ takes the values $\{-1, 0, 1\}$ with probabilities $\{p_{-1}$, p_0, and $p_1\}$, respectively. The case -1 corresponds to $x_i - e_i \leq y < x_i$; the case 1 to $x_i \leq y < x_i - e_i$; and the case 0 otherwise. It is apparent that the probabilities p_i depend on both the distributions of x_i and e_i. The expected value and variance of $Y_i^\epsilon(y)$ obey

$$E[Y_i^\epsilon(y)] = -1p_{-1} + 1p_1 \quad \text{and} \quad V[Y_i^\epsilon(y)] \leq E[(Y_i^\epsilon(y))^2] = (-1)^2 p_{-1} + 1^2 p_1.$$

Since

$$|p_1 - p_{-1}| \leq p_1 + p_{-1} \leq \Pr(|y - x_i| \leq \epsilon) \leq C_L \epsilon,$$

where C_L depends on the Lipschitz constant of F, we obtain

$$E[F_n(y) - F_{n,\epsilon}(y)] = p_1 - p_{-1} \leq C_L \epsilon, \quad Var[F_n(y) - F_{n,\epsilon}(y)] = n^{-1}(p_1 + p_{-1}) \leq C_L n^{-1} \epsilon. \tag{6.2}$$

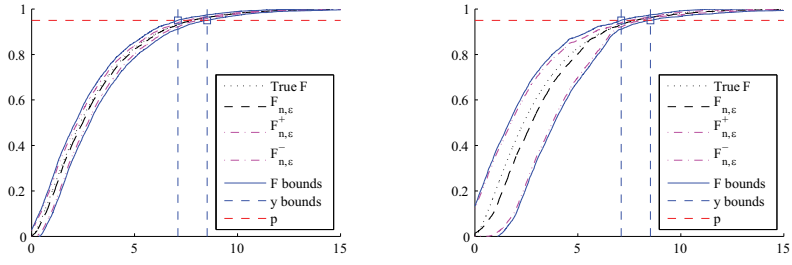
Thus, in the case $p_{-1} = p_1$, the numerical error in the cdf is in the order of $n^{-1/2}\epsilon^{1/2}$, since the expected value is zero. Thus, no refinements are necessary, i.e., we can let $\epsilon \approx 1$ and still balance the statistical and numerical errors in the cdf, thanks to cancellations in the miscounts. However, the case $p_{-1} = p_1$ is rather unrealistic. Assuming $F(y)$ is differentiable, we still need e_i to be median-unbiased given x_i, which cannot be expected from errors in numerical simulations in general. The effect of miscounts is investigated numerically in section 7.4.

7. Numerical experiments. This section presents a few numerical experiments demonstrating the selective refinement algorithm and its gain in computational complexity compared to full refinement. The last numerical example illustrates the discussion in section 6 on how miscounts affect the convergence with respect to the numerical error.

7.1. Demonstration in principle. In this experiment, we let the QoI be sampled directly from a χ^2-distribution with three degrees of freedom, i.e., $X \sim \chi^2(3)$. For a sample $\{x_i\}_{i=1}^n$ from $\chi^2(3)$, the approximate sample $\{x_i^\epsilon\}_{i=1}^n$ is computed as follows. For a given ϵ_i, x_i^ϵ is computed as $x_i^\epsilon = x_i + 2/3(\sin(100 \times \epsilon_i \times i) + 1/2) \times \epsilon_i$, to simulate some solution procedure generating approximate values with a systematic error within the error bound. We use the Agresti–Coull interval. With this setup, both Assumptions 3.1 and 3.3 are satisfied. We pick $n = 10000$, $p = 0.95$, $\beta = 0.99$, and $\epsilon_{\text{init}} = 1$. These values are chosen to illustrate the performance of the selective refinement algorithm.

Algorithms 2 and 3 are executed with the described setup. The resulting functions $F_{n,\epsilon}$; $F_{n,\epsilon}^+$; $F_{n,\epsilon}^-$; lower and upper bounds of F; and lower and upper bounds, $y_{n,\epsilon}^-$ and $y_{n,\epsilon}^+$, respectively, of y are plotted after termination of the two algorithms in Figures 1a and 1b, respectively. (Note that all functions $F_{\cdot,\cdot}$ are transformed via the affine transformation $T(F_n(x)) = \tilde{p}$

defined by the Agresti–Coull confidence interval in (3.5), i.e., the figure actually shows $T(F^+_{n,\epsilon})$, and so on.) The figures illustrate how the numerical error in samples away from the 95%-quantile is larger after selective refinement than after full refinement. Both algorithms executed two iterations before the error balance was achieved. The p-quantile bounding estimates are identical for both algorithms, with $y^-_{n,\epsilon} = 7.1055$ and $y^+_{n,\epsilon} = 8.5244$. This is in accordance with Theorem 5.2. The true 95%-quantile is $y = 7.8147$.



(a) Full refinement. (b) Selective refinement.

Figure 1. 99% *confidence band of F with* 95%*-quantile bounds after* (a) *full refinement and* (b) *selective refinement. Note how the numerical error (distance between dash-dotted, magenta lines) is larger for samples away from the p-quantile with selective refinement.*

7.2. Computational complexity experiment. Theorem 5.4 predicts the following computational complexity reduction for selective refinement versus full refinement:

$$\frac{\sum_{j=0}^{J} W^{(j)}}{\sum_{j=0}^{J} \widehat{W}^{(j)}} \leq \min\left(1, KC_4 \begin{cases} n^{-q/2} & \text{if } q < 1 \\ n^{-1/2}\log_2 n & \text{if } q = 1 \\ n^{-1/2} & \text{if } q > 1 \end{cases}\right),$$

with different values of C_4 for the three different cases. In this experiment, we use exactly the same setup as in the previous experiment. This means X and x^ϵ_i are defined as in section 7.1. Additionally, for the model of work, we assume $C_2 = 1$, i.e., $W(\epsilon_i) = \epsilon_i^{-q}$, and we consider three different values of q: $q = 3$, 1, and 1/3 in order to try the three cases above. We pick $p = 0.95$, $\beta = 0.99$, $\epsilon_{\text{init}} = 1$ and execute Algorithms 2 and 3. The resulting work ratio is presented in Figure 2. The solid lines show the value of the work ratio, i.e., $\frac{\sum_{j=0}^{N} W^{(j)}}{\sum_{j=0}^{N} \widehat{W}^{(j)}}$, for the three different values of q. The constants 6, 2, and 3 in the definition of the dashed lines are selected manually to make the slope comparison easy. The slopes of the experimental data verify Theorem 5.4.

7.3. An engineering application. We return to the model for Darcy flow (2.1). We complete the problem formulation by applying the boundary conditions

$$\begin{aligned} u &= 0 \quad \text{on } \Gamma_1, \\ u &= 1 \quad \text{on } \Gamma_2, \\ n \cdot A(\omega)\nabla u &= 0 \quad \text{on } \Gamma_3 \cup \Gamma_4, \end{aligned}$$

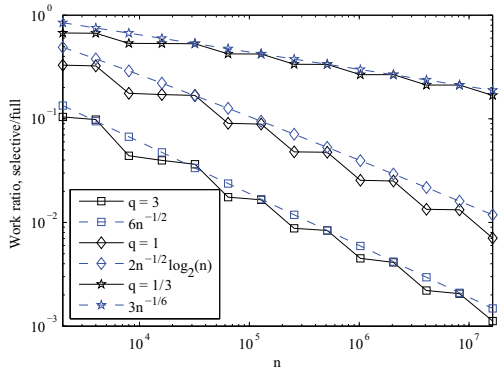


Figure 2. *Work reduction; full vs. selective refinement as a function of sample size.*

where n denotes the outward normal on the boundary of $\mathscr{D}$ and Γ_1, Γ_2, Γ_3, Γ_4 are the left, right, upper, and lower boundaries, respectively. We define $\Gamma_D = \Gamma_1 \cup \Gamma_2$ and $\Gamma_N = \Gamma_3 \cup \Gamma_4$ to denote the Dirichlet and Neumann boundaries, respectively.

We define $H^1_D(\mathscr{D}) = \{v \in H^1(\mathscr{D}) : v|_{\Gamma_1} = 0 \text{ and } v|_{\Gamma_2} = 1\}$ and $H^1_0(\mathscr{D}) = \{v \in H^1(\mathscr{D}) : v|_{\Gamma_D} = 0\}$ to be function spaces that satisfy the boundary condition and vanishing on the Dirichlet boundary, respectively. Let $V^h \subset H^1(\mathscr{D})$ be the space of continuous functions on $\mathscr{D}$ that are also affine on all triangles $T \in \mathcal{T}^h$, $\mathcal{T}^h$ being a conforming triangulation of $\mathscr{D}$, where $h = \max_{T\in\mathcal{T}^h} \operatorname{diam}(T)$. We assume that the finite element triangulation is a refinement of $\mathcal{T}^{h_0}$ used in the definition of the diffusion coefficient. The finite element discretization is then as follows: Find $u^h \in V^h \cap H^1_D(\mathscr{D})$ such that

$$a(\omega; u^h, v) = \int_{\mathscr{D}} A(\omega)\nabla u \cdot \nabla v \, \mathrm{d}x = 0 \quad \text{for all } v \in V^h \cap H^1_0(\mathscr{D}).$$

We use an adjoint-based approach to error estimation [6, 5, 7, 2]. The QoI (normal flux through Γ_1) is approximated by the linear functional

$$Q(u^h(\omega)) = a(\omega; u^h, v) \quad \text{for all } v \in V^h \cap H^1_D(\mathscr{D}). \tag{7.1}$$

We solve for a corresponding numerical adjoint solution: Find $\phi^k \in V^k \cap H^1_D(\mathscr{D})$ such that

$$a(\omega; v, \phi^k) = 0 \quad \text{for all } v \in V^k \cap H^1_0(\mathscr{D}), \tag{7.2}$$

where $k < h$. We use $k = h/2$ to approximate the adjoint solution. We define $\pi^h : H^1_D(\mathscr{D}) \to V^h \cap H^1_D(\mathscr{D})$ to be a (quasi-)interpolation operator.

With this framework, we can produce both accurate a posteriori error estimates and a posteriori error bounds.

1. For an accurate estimate, we use [6, 5, 7, 2]

$$Q(u) - Q^h(u^h) \approx a(\omega; u^h, \phi^k - \pi^h\phi^k) = e^{\mathrm{EST}}(u^h).$$

This estimate is exact if $\phi = \phi^k$. We approximate the QoI as $x_i^\epsilon = Q(u^h(\omega_i)) + e^{\mathrm{EST}}(u^h(\omega_i))$ (note the correction) for an h_i satisfying $|e^{\mathrm{EST}}(u^h(\omega_i))| < \epsilon_i$ in order to reach a numerical error tolerance of ϵ_i. The procedure to reach the tolerance is to halve h_i until the error estimate is less than numerical error tolerance.

2. We derive an (dual or adjoint weighted) a posteriori error bound from the a posteriori error estimate by integration by parts over each element in the mesh and accumulating quantity values on common element boundaries to obtain

$$|Q(u) - Q^h(u^h)| \leq \sum_{T \in \mathcal{T}^k} R_T(u^h) \cdot w_T + r_T(u^h) \cdot w_{\partial T} = e^{\mathrm{DWR}}(u^h), \tag{7.3}$$

where the residuals R_T and r_T are defined by

$$R_T(u^h) = \|\nabla \cdot A(\omega)\nabla u^h\|_{L^2(T)},$$
$$r_T^2(u^h) = \frac{1}{2}\|h^{1/2}[A(\omega)\nabla u^h]\|^2_{L^2(\partial T \setminus (\Gamma_D \cup \Gamma_N))} + \|h^{1/2} A(\omega)\nabla u^h\|^2_{L^2(\partial T \cap \Gamma_N)},$$

respectively, where $[\cdot]$ denotes the jump in normal direction, and h is a piecewise constant function $h|_T = \mathrm{diam}(T)$. The adjoint weights (w_T and $w_{\partial T}$) are defined by

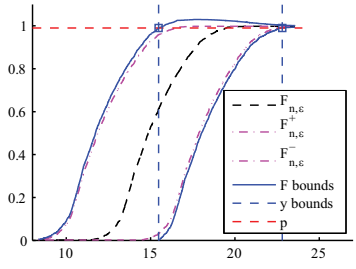
$$w_T = \|\phi^k - \pi^h \phi^k\|_{L^2(T)}, \qquad w_{\partial T} = \|h^{-1/2}(\phi^k - \pi^h \phi^k)\|_{L^2(\partial T)},$$

respectively. For a given realization ω_i, we approximate the QoI as $x_i^\epsilon = Q(u^h(\omega_i))$ for an h_i satisfying $e^{\mathrm{DWR}}(u^h(\omega_i)) < \epsilon_i$. In order to find such an h_i, we start with an initial h_i and halve it until the bound is less than the numerical error tolerance.

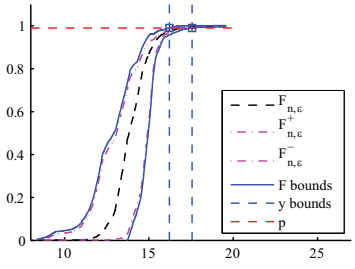
The statistical error, $\mathcal{E}^{\mathrm{stat}}_{n,\epsilon}$, is approximated using the Agresti–Coull confidence interval (see (3.4), (3.5), and (3.7)). We pick $n = 2000$, $p = 0.99$, $\beta = 0.99$, $\epsilon_{\mathrm{init}} = 3$, $A_0 = 1$ and execute Algorithm 3 (selective refinement) using the two error bounding and estimation methods introduced above.

For both error bounding and estimation methods, four iterations were performed until the errors were balanced and the algorithm terminated. Figures 3a and 3b illustrate the initial and final p-quantile bounding estimates, respectively, for the adjoint-based error bounds (method 2 above). It is evident that realizations close to the p-quantile have been refined to a larger extent than those far from the p-quantile. Figure 3c shows a zoomed-in version of Figure 3b, where the balance of numerical and statistical error can be observed. Also, the interval defined by the final p-quantile bounding estimates can be read from Figure 3c and is approximately $[16.8, 18.1]$.

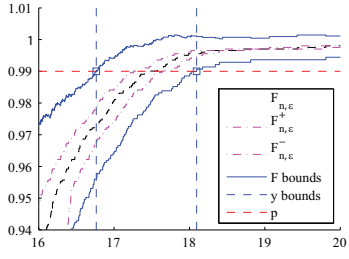
As in the previous section, we compare the ratio of required work between selective and full refinement. In this example, we use the following model of work, $W(h_i) = h_i^{-2}$, where the exponent is -2, since we have a uniform triangulation of a two-dimensional domain and solve the linear equation systems in linear time complexity. However, in this example it is too expensive to perform the full algorithm to yield values of h_i. Instead, h_i for the full algorithm is estimated from the error estimates in the resulting selective algorithm solution using the numerically verified rate of convergence 1. That is, for each realization, the number of times the numerical error has to be halved to reach the numerical error tolerance is computed, and the corresponding h_i is halved accordingly. This leaves a set of h_i values that is used to



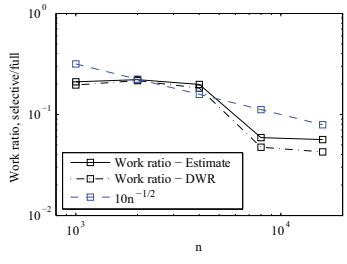
(a) Initial state, before iteration 1.



(b) Final state, after iteration 4.



(c) Zoom-in of Figure 3b of the region close to the p-quantile.

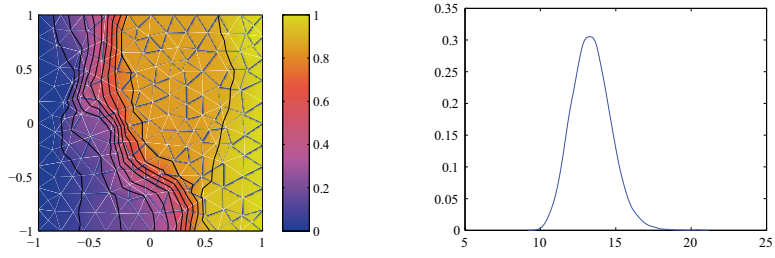


(d) Work reduction; full vs. selective refinement as a function of sample size.

Figure 3. *Plots illustrating performance of the selective algorithm for the boundary flux problem.*

estimate the work for the full algorithm. The ratio between the required work for the two algorithms, for $n = 1000$, 2000, 4000, 8000, and 16000, is shown in Figure 3d. The figure shows work savings in the order of 10 for this span of n, and the work reduction rate in Theorem 5.4 is observed in practice. The jump between $n = 4000$ and $n = 8000$ is explained by the fact that an additional iteration was required to balance the errors for the latter case. This causes a substantial increase of work for the full algorithm. These jumps are present also in Figure 2. For illustration purposes, Figure 4a contains a solution plot for a single realization on the coarsest mesh and Figure 4b shows an estimated probability density function for Q based on 4000 realizations with error tolerance 0.1 using the adjoint-based error estimate (method 2 above).

7.4. Effect of miscounts on numerical error. Following the discussion in section 6, we illustrate how miscounts in the computation of the cdf affect the "exact" numerical error $|F_n - F_{n,\epsilon}|$. We let $X \sim \mathcal{N}(0,1)$ and $\{\psi_i\}_{i=1}^n$ be an i.i.d. sample of the uniform distribution $\mathcal{U}(0,1)$. We consider two cases for the numerical error: (a) no systematic error, $x_i^\epsilon = x_i + \epsilon(2\psi_i - 1)$; (b) systematic error, $x_i^\epsilon = x_i + \epsilon(2(\psi_i)^2 - 1)$. Given a value of n, we pick $\epsilon = n^{-1/2}$ (simulating balance between numerical and statistical error), generate a random sample of size n from X and $\mathcal{U}$ to compute x_i, ψ_i, and x_i^ϵ, and compute the numerical error $|F_n(y) - F_{n,\epsilon}(y)|$ for



(a) Solution for single realization, $h = 0.2$.

(b) Estimated density function for Q, with $n = 4000$ and $\epsilon = 0.1$.

Figure 4.

$y = 1$ for the two cases. This is done for a range of values of n. A simple moving average with respect to n is used in order to increase the readability of the resulting graphs, which can be found in Figure 5. From the figure, we can see that there is a cancellation effect of miscounts in the numerical error in case (a), where we gain a factor $n^{-1/4}$ in the numerical error. However, from case (b) we see that when systematic errors are present, the miscounts do not affect the order of convergence of the numerical error. This means the "worst case" bounds give an overly pessimistic bound of the numerical error when no systematic error in the numerical approximations is present.

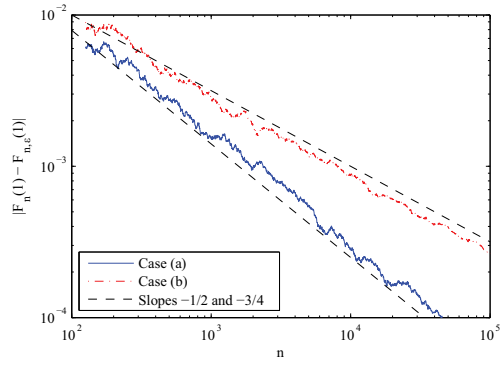


Figure 5. *Convergence of numerical error for* (a) *no systematic error and* (b) *systematic error in numerical approximation.*

8. Technical results and proofs.

In this section, we collect technical results and proofs.

Lemma 8.1. *If F is continuous, then with probability* 1*,*

$$\sup_{x\in\mathbb{R}} \mathcal{E}^{\text{stat}}_{n,\epsilon}(x) + \mathcal{E}^{\text{num}}_{n,\epsilon}(x) \to 0 \quad \textit{and} \quad \sup_{x\in\mathbb{R}} |F_{n,\epsilon}(x) - F(x)| \to 0 \tag{8.1}$$

as $n \to \infty$ and $\epsilon \to 0$.

Proof of Lemma 8.1. First, Assumption 3.3 implies

$$F^+_{n,\epsilon}(x) - F^-_{n,\epsilon}(x) \le \frac{\#\{i = 1, \dots, n : x_i - 2\epsilon_i \le x < x_i + 2\epsilon_i\}}{n}. \tag{8.2}$$

By the continuity of F, $x_i - x_j \neq 0$ almost surely for all $i \neq j$. Let $\tilde{\epsilon} = \min_{i \neq j} |x_i - x_j| > 0$. For all i, choose $\epsilon_i = \tilde{\epsilon}/4$. Continuing from (8.2),

$$F^+_{n,\epsilon}(x) - F^-_{n,\epsilon}(x) \le \frac{\#\{i = 1, \dots, n : x_i - \tilde{\epsilon}/2 \le x < x_i + \tilde{\epsilon}/2\}}{n} \le 1/n.$$

This implies $\sup_{x \in \mathbb{R}}(F^+_{n,\epsilon}(x) - F^-_{n,\epsilon}(x)) \to 0$ as $n \to \infty$ and $\epsilon \to 0$.

From Lemma 8.1, we have

$$\sup_x \left(\mathcal{E}^{\text{num}}_{n,\epsilon}(x) + \mathcal{E}^{\text{stat}}_{n,\epsilon}(x)\right) \le \sup_x (F^+_{n,\epsilon} - F^-_{n,\epsilon}(x)) + C_1 n^{-1/2} \to 0$$

as $n \to \infty$ and $\epsilon \to 0$. The Glivenko–Cantelli theorem implies (see, for example, page 61 of [11] and the references therein)

$$\sup_x |F_{n,\epsilon}(x) - F(x)| \le \sup_x (F^+_{n,\epsilon}(x) - F^-_{n,\epsilon}(x)) + \sup_x |F_n(x) - F(x)| \to 0$$

as $n \to \infty$ and $\epsilon \to 0$. ∎

Proof of Theorem 4.2. We set $\eta(x) = -\mathcal{E}^{\text{stat}}_{n,\epsilon}(x) - \mathcal{E}^{\text{num}}_{n,\epsilon}(x) + F_{n,\epsilon}(x) - F(x)$. By Lemma 8.1,

$$\sup_x |\eta(x)| \le \sup_x \mathcal{E}^{\text{stat}}_{n,\epsilon}(x) + \mathcal{E}^{\text{num}}_{n,\epsilon}(x) + \sup_x |F_{n,\epsilon}(x) - F(x)| \to 0$$

as $n \to \infty$ and $\epsilon \to 0$. Now, from the definition of $y^+_{n,\epsilon}$, y^-, and y^+, $|\eta(x)| \to 0$ implies the result. If we let $\eta(x) = \mathcal{E}^{\text{stat}}_{n,\epsilon}(x) + \mathcal{E}^{\text{num}}_{n,\epsilon}(x) + F_{n,\epsilon}(x) - F(x)$ instead, we can show the same result for $y^-_{n,\epsilon}$. ∎

Proof of Theorem 4.3. Using the definition of $y^-_{n,\epsilon}$ and $y^+_{n,\epsilon}$, and (3.3), we obtain

$$\begin{aligned} y^+_{n,\epsilon} - y^-_{n,\epsilon} &\le \inf\{x \in \mathbb{R} : F(x - 2\epsilon_{\max}) - (K + C_1)n^{-1/2} \ge p\} \\ &\qquad - \inf\{x \in \mathbb{R} : F(x + 2\epsilon_{\max}) + (K + C_1)n^{-1/2} \ge p\} \\ &\le F^{-1}(p + (K + C_1)n^{-1/2}) - F^{-1}(p - (K + C_1)n^{-1/2}) + 4\epsilon_{\max} \\ &\le 2\ell^{-1}(K + C_1)n^{-1/2} + 4\epsilon_{\max}, \end{aligned}$$

with probability at least $1 - 2e^{-2K^2}$. ∎

Proof of Theorem 5.2. First we show that, for $y^-_{n,\epsilon} \le x \le y^+_{n,\epsilon}$,

$$F^-_{n,\epsilon}(x) \ge F^-_{n,\delta}(x), \quad F^+_{n,\epsilon}(x) \le F^+_{n,\delta}(x). \tag{8.3}$$

We let I^-, I, and I^+ be the following partition of $\{1,\dots,n\}$:

$$\begin{aligned} I^- &= \{i = 1,\dots,n : \qquad x^\epsilon_i \le y^-_{n,\epsilon} - \epsilon_i\}, \\ I &= \{i = 1,\dots,n : y^-_{n,\epsilon} - \epsilon_i < x^\epsilon_i \le y^+_{n,\epsilon} + \epsilon_i\}, \\ I^+ &= \{i = 1,\dots,n : y^+_{n,\epsilon} + \epsilon_i < x^\epsilon_i\}. \end{aligned}$$

Defining the predicate $P_{\epsilon,i}(x) = [x^\epsilon_i + \epsilon \le x]$, we have

$$nF^-_{n,\epsilon}(x) = \#\{i = 1,\dots,n : P_{\epsilon,i}(x)\} = \#I^- + \#\{i \in I : P_{\epsilon,i}(x)\}; \tag{8.4}$$

i.e., all elements in I^-, some from I, and none from I^+ satisfy the predicate and contribute to the value of $F^-_{n,\epsilon}(x)$. From (3.6) we have

$$x^\zeta_i - \zeta_i \le x_i \le x^\eta_i + \eta_i \quad \text{for any} \quad 0 \le \zeta_i, \eta_i \le 1. \tag{8.5}$$

We investigate how $P_{\delta,\cdot}(x)$ (i.e., the predicate with numerical error tolerance parameter δ) acts on elements in I^+:

$$\#\{i \in I^+ : P_{\delta,i}(x)\} \le \#\{i \in I^+ : x^\epsilon_i - \epsilon_i \le x\} \le \#\{i \in I^+ : y^+_{n,\epsilon} < x\} = 0, \tag{8.6}$$

where (8.5) and the definition of I^+ were used in the inequalities. Now consider $P_{\delta,\cdot}(x)$ on elements in I:

$$\#\{i \in I : P_{\delta,i}(x)\} = \#\{i \in I : x^\epsilon_i + \epsilon \le x\} = \#\{i \in I : P_{\epsilon,i}(x)\}, \tag{8.7}$$

since $\epsilon_i = \delta_i$ for $i \in I$. Finally, for $P_{\delta,\cdot}(x)$ on elements in I^-, we obviously have

$$\#\{i \in I^- : P_{\delta,i}(x)\} \le \#I^- = \#\{i \in I^- : P_{\epsilon,i}(x)\}. \tag{8.8}$$

Combining (8.4), (8.6), (8.7), and (8.8) we get that

$$nF^-_{n,\epsilon}(x) \ge \#\{i = I^- \cup I \cup I^+ : P_{\delta,i}(x)\} = nF^-_{n,\delta}(x),$$

which proves the first inequality in (8.3). A similar argument can be used for the second one.

Now we can continue with the main result. The following argument shows $y^+_{n,\delta} \ge y^+_{n,\epsilon}$. An analogous argument is used to show $y^-_{n,\delta} \le y^-_{n,\epsilon}$. First, note that

$$\mathcal{E}^{\text{stat}}_{n,\epsilon}(x) \le \mathcal{E}^{\text{stat}}_{n,\delta}(x) \tag{8.9}$$

for all $y^-_{n,\epsilon} \le x \le y^+_{n,\epsilon}$ satisfying (8.3) by the definition of $\mathcal{E}^{\text{stat}}_{n,\epsilon}$ in (3.8), since the maximum over a subset is not greater than the maximum over its superset. From the definition of $y^+_{n,\epsilon}$ and inequalities (8.3) and (8.9) we have that for $y^-_{n,\epsilon} \le x < y^+_{n,\epsilon}$,

$$p > F^-_{n,\epsilon}(x) - \mathcal{E}^{\text{stat}}_{n,\epsilon}(x) \ge F^-_{n,\delta}(x) - \mathcal{E}^{\text{stat}}_{n,\delta}(x). \tag{8.10}$$

Further, we obviously have $y^-_{n,\epsilon} \le y_n \le y^+_{n,\delta}$. Now, if $y^-_{n,\epsilon} \le y^+_{n,\delta} < y^+_{n,\epsilon}$, then considering (8.10), there must exist an $0 \le \eta < y^+_{n,\epsilon} - y^+_{n,\delta}$ such that

$$F^-_{n,\delta}(y^+_{n,\delta}+\eta) - \mathcal{E}^{\text{stat}}_{n,\delta}(y^+_{n,\delta}+\eta) \ge p > F^-_{n,\delta}(y^+_{n,\delta}+\eta) - \mathcal{E}^{\text{stat}}_{n,\delta}(y^+_{n,\delta}+\eta),$$

which is a contradiction. Hence, $y^+_{n,\delta} \ge y^+_{n,\epsilon}$. ■

Proof of Theorem 5.4. The work using selective refinement is always less than or equal to the work using full refinement. This is obvious, since the full refinement is equivalent to using $\{i = 1, \dots, n : x_i\}$ as the set of realizations to refine in each iteration, i.e., realizations that do not affect the values are refined, whereas the selective algorithm refines the realizations in $I^{(j)}$, whose cardinality is at most n.

Next, we find a set $\hat{I}^{(j)}$ defined by a priori information only, with the property $I^{(j)} \subseteq \hat{I}^{(j)}$; i.e., $\hat{I}^{(j)}$ is a superset of the realizations refined in each iteration. We make use of the following bounds:

$$\begin{aligned} y^-_{n,\epsilon} &\ge \inf\{x \in \mathbb{R} : \#\{i : x_i - 2\epsilon_{\max} \le x\}/n + C_1 n^{-1/2} \ge p\} \\ &= F_n^{-1}(p - C_1 n^{-1/2}) - 2\epsilon_{\max} = y^{--}_{n,\epsilon} \end{aligned} \tag{8.11}$$

and

$$\begin{aligned} y^+_{n,\epsilon} &\le \inf\{x \in \mathbb{R} : \#\{i : x_i \le x - 2\epsilon_{\max}\}/n - C_1 n^{-1/2} \ge p\} \\ &= \inf\{x \in \mathbb{R} : F_n(x - 2\epsilon_{\max}) - C_1 n^{-1/2} \ge p\} \\ &= F_n^{-1}(p + C_1 n^{-1/2}) + 2\epsilon_{\max} = y^{++}_{n,\epsilon}. \end{aligned} \tag{8.12}$$

Further, let $\delta = (2^{-j}, \dots, 2^{-j})$ and $\epsilon = \epsilon^{(j)}$, i.e., the numerical error tolerance parameters for full and selective refinement, respectively, after j iterations. For $i \in I^{(j)}$, we have $\epsilon_i = \delta_i = 2^{-j}$ and the set $I^{(j)}$ cannot be made smaller (but possibly larger) by replacing ϵ with δ, which implies the first set relation in (8.13). For the second set relation in (8.13), we have used (3.6) together with inequality (8.11) and (8.12). We define $\hat{I}^{(j)}$ as

$$\begin{aligned} I^{(j)} &= \{i = 1, \dots, n : y^-_{n,\epsilon} - \epsilon_i < x_i^\epsilon \le y^+_{n,\epsilon} + \epsilon_i\} \\ &\subseteq \{i = 1, \dots, n : y^-_{n,\epsilon} - \delta_i < x_i^\delta \le y^+_{n,\epsilon} + \delta_i\} \\ &\subseteq \{i = 1, \dots, n : y^{--}_{n,\epsilon} - 2^{-j+1} < x_i \le y^{++}_{n,\epsilon} + 2^{-j+1}\} = \hat{I}^{(j)}. \end{aligned} \tag{8.13}$$

The cardinality of this set can be expressed as

$$\begin{aligned} \#\hat{I}^{(j)} &= n(F_n(y^{++}_{n,\epsilon} + 2^{-j+1}) - F_n(y^{--}_{n,\epsilon} - 2^{-j+1})) \\ &= n(F_n(F_n^{-1}(p + C_1 n^{-1/2}) + 2^{-j+2}) - F_n(F_n^{-1}(p - C_1 n^{-1/2}) - 2^{-j+2})). \end{aligned}$$

Using the DKW inequality (see (3.3)), we obtain for a Lipschitz continuous F with Lipschitz constant L that the following holds with probability at least $1 - 2e^{-2K^2}$: if $x \ge 0$,

$$\begin{aligned} F_n(F_n^{-1}(p) + x) &\le F(F_n^{-1}(p) + x) + Kn^{-1/2} \\ &\le F_n(F_n^{-1}(p)) + \int_{F_n^{-1}(p)}^{F_n^{-1}(p)+x} F'(y)\,\mathrm{d}y + 2Kn^{-1/2} \\ &\le p + n^{-1} + Lx + 2Kn^{-1/2}, \end{aligned} \tag{8.14a}$$

and similarly, if $x \leq 0$,

$$F_n(F_n^{-1}(p) + x) \geq p + Lx - 2Kn^{-1/2}. \tag{8.14b}$$

Using (8.14), the total work for j iterations can be bounded from above by

$$\begin{aligned}
\sum_{j=0}^{J} W^{(j)} &\leq n + 2^q \sum_{j=0}^{J-1} 2^{qj} \#\hat{I}^{(j)} \\
&= n\Bigg(1 + 2^q \sum_{j=0}^{J-1} 2^{qj} \Big(F_n\left(F_n^{-1}\left(p + C_1 n^{-1/2}\right) + 4 \times 2^{-j}\right) \\
&\qquad\qquad - F_n\left(F_n^{-1}\left(p - C_1 n^{-1/2}\right) - 4 \times 2^{-j}\right)\Big)\Bigg) \\
&\leq n\left(1 + 2^q \sum_{j=0}^{J-1} 2^{qj}\left((2C_1 + 4K)n^{-1/2} + 8L2^{-j} + n^{-1}\right)\right) = T.
\end{aligned}$$

We use the assumption that $J < \frac{1}{2}\log_2 n - \log_2 C_3 + 1$ and observe that we need to consider three different cases for the geometric sums: case (1) for $q < 1$,

$$\begin{aligned}
T &= n\left(1 + 2^q\left(((2C_1 + 4K)n^{-1/2} + n^{-1})\frac{2^{qJ} - 1}{2^q - 1} + 8L\frac{2^{(q-1)J} - 1}{2^{(q-1)} - 1}\right)\right) \\
&\leq n\left(1 + 2^q\left(((2C_1 + 4K)n^{-1/2} + n^{-1})\frac{2^{q(1-\log_2 C_3)}n^{q/2} - 1}{2^q - 1} + 8L\frac{1}{1 - 2^{(q-1)}}\right)\right) \\
&\leq \left((D_1 + KD_2)C_3^{-q} + LD_3\right) n^{-q/2} n^{1+q/2};
\end{aligned}$$

case (2) for $q = 1$,

$$\begin{aligned}
T &= n\left(1 + 2^q\left(((2C_1 + 4K)n^{-1/2} + n^{-1})\frac{2^{qJ} - 1}{2^q - 1} + 8LJ\right)\right) \\
&\leq n\Bigg(1 + 2^q\Bigg(((2C_1 + 4K)n^{-1/2} + n^{-1})\frac{2^{q(1-\log_2 C_3)}n^{q/2} - 1}{2^q - 1} \\
&\qquad\qquad + 8L\left((1/2)\log_2 n - \log_2 C_3 + 1\right)\Bigg)\Bigg) \\
&\leq \left((D_1 + KD_2)C_3^{-q} + LD_3\right) n^{-1/2}(\log_2 n) n^{1+q/2};
\end{aligned}$$

and case (3) for $q > 1$,

$$\begin{aligned} T &= n\left(1 + 2^q\left(((2C_1 + 4K)n^{-1/2} + n^{-1})\frac{2^{qJ} - 1}{2^q - 1} + 8L\frac{2^{(q-1)J} - 1}{2^{(q-1)} - 1}\right)\right) \\ &\le n\left(1 + 2^q\left(((2C_1 + 4K)n^{-1/2} + n^{-1})\frac{2^{q(1-\log_2 C_3)}n^{q/2} - 1}{2^q - 1}\right.\right. \\ &\qquad \left.\left. + 8L\frac{2^{(q-1)(1+\log_2 C_3)}n^{-1/2}n^{q/2} - 1}{2^{(q-1)} - 1}\right)\right) \\ &\le \left((D_1 + KD_2)C_3^{-q} + LD_3C_3^{1-q}\right)n^{-1/2}n^{1+q/(2\lambda)}, \end{aligned}$$

with probability at least $1 - 2e^{-2K^2}$, where D_1, D_2, and D_3 depend on C_1 and q. The total work $\sum_{j=0}^{J}\widehat{W}^{(j)}$ (using that $1/2\log_2 n - \log_2 C_3 - 1 \le J$) for full refinement can be bounded below by

$$C_2^{-1}\sum_{j=0}^{J}\widehat{W}^{(j)} \ge n + 2^q\sum_{j=0}^{J-1}2^{qj}n = n\left(1 + 2^q\frac{2^{qJ} - 1}{2^q - 1}\right) \ge D_4C_3^{-q}n^{1+q/2},$$

where D_4 depends on q. The ratio between the required work for the selective refinement and the full refinement can be bounded above by

$$\frac{\sum_{j=0}^{J}W^{(j)}}{\sum_{j=0}^{J}\widehat{W}^{(j)}} \le \min\left(1, KC_4(F, C_1, C_2, C_3, q)\begin{cases} n^{-q/2} & \text{if } q < 1 \\ n^{-1/2}\log_2 n & \text{if } q = 1 \\ n^{-1/2} & \text{if } q > 1 \end{cases}\right)$$

with probability at least $1 - 2e^{-2K^2}$ and with different constant $KC_4(F, C_1, C_2, C_3, q)$ in the three different cases. ∎

9. Conclusion. In this paper, we consider the problem of estimating the p-quantile for a given functional evaluated on numerical solutions of a deterministic model in which the model input is subject to stochastic variation. Assuming a computational a posteriori error bound for the functional computed from a specific numerical solution, we derive a computational a posteriori error bound for the p-quantile estimators that takes into account the effects of both the stochastic sampling error and the deterministic numerical solution error. Under general assumptions, we prove asymptotic convergence of the p-quantile estimator bounds in the limit of large sample size and decreasing numerical error.

The a posteriori error bound provides the capability of quantifying the effect of the numerical accuracy of each sample on the computed p-quantile. We propose a selective refinement algorithm for computing an estimate of the p-quantile with a desired accuracy in a computationally efficient fashion. The algorithm exploits the fact that the accuracy of a relatively small subset of sample values significantly affects the accuracy of a p-quantile estimator. The algorithm calls for refinement of the discretization in order to achieve the necessary accuracy for only those solutions in the subset. The algorithm can lead to significant computational

gain. For instance, if the numerical model is a first order discretization of a partial differential equation with spatial dimension greater than one, the reduction in computational work (compared to standard Monte Carlo using n samples) is asymptotically proportional to $n^{1/2}$. The numerical experiments presented in the paper support this conclusion.

Acknowledgment. D. Estep gratefully acknowledges Chalmers University of Technology for the support provided by the appointment as Chalmers Jubilee Professor.

REFERENCES

[1] A. Agresti and B. A. Coull, *Approximate is better than "exact" for interval estimation of binomial proportions*, Amer. Statist., 52 (1998), pp. 119–126.
[2] W. Bangerth and R. Rannacher, *Adaptive Finite Element Methods for Differential Equations*, Lectures Math. ETH Zürich, Birkhäuser Verlag, Basel, 2003.
[3] L. D. Brown, T. T. Cai, and A. DasGupta, *Interval estimation for a binomial proportion*, Statist. Sci., 16 (2001), pp. 101–133.
[4] A. Dvoretzky, J. Kiefer, and J. Wolfowitz, *Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator*, Ann. Math. Statist., 27 (1956), pp. 642–669.
[5] K. Eriksson, D. Estep, P. Hansbo, and C. Johnson, *Introduction to adaptive methods for differential equations*, in Acta Numerica, 1995, Acta Numer., Cambridge University Press, Cambridge, UK, 1995, pp. 105–158.
[6] K. Eriksson, D. Estep, P. Hansbo, and C. Johnson, *Computational Differential Equations*, Cambridge University Press, New York, 1996.
[7] D. Estep, M. G. Larson, and R. D. Williams, *Estimating the error of numerical solutions of systems of reaction-diffusion equations*, Mem. Amer. Math. Soc., 146 (2000), 696.
[8] D. Estep, A. Målqvist, and S. Tavener, *Nonparametric density estimation for randomly perturbed elliptic problems* I: *Computational methods, a posteriori analysis, and adaptive error control*, SIAM J. Sci. Comput., 31 (2009), pp. 2935–2959.
[9] D. Estep, A. Målqvist, and S. Tavener, *Nonparametric density estimation for randomly perturbed elliptic problems*. II. *Applications and adaptive modeling*, Internat. J. Numer. Methods Engrg., 80 (2009), pp. 846–867.
[10] D. Feldman and H. G. Tucker, *Estimation of non-unique quantiles*, Ann. Math. Statist., 37 (1966), pp. 451–457.
[11] R. Serfling, *Approximation Theorems of Mathematical Statistics*, Wiley-Interscience, New York, 2001.

Paper VII

A multilevel Monte Carlo method for computing failure probabilities

Daniel Elfverson* Fredrik Hellman† Axel Målqvist‡

September 9, 2015

Abstract

We propose and analyze a method for computing failure probabilities of systems modeled as numerical deterministic models (e.g., PDEs) with uncertain input data. A failure occurs when a functional of the solution to the model is below (or above) some critical value. By combining recent results on quantile estimation and the multilevel Monte Carlo method we develop a method which reduces computational cost without loss of accuracy. We show how the computational cost of the method relates to error tolerance of the failure probability. For a wide and common class of problems, the computational cost is asymptotically proportional to solving a single accurate realization of the numerical model, i.e., independent of the number of samples. Significant reductions in computational cost are also observed in numerical experiments.

1 Introduction

This paper is concerned with the computational problem of finding the probability for failures of a modeled system. The model input is subject to uncertainty with known distribution and a failure is the event that a functional (quantity of interest, QoI) of the model output is below (or above) some

*Information Technology, Uppsala University, Box 337, SE-751 05, Uppsala, Sweden (daniel.elfverson@it.uu.se). Supported by the Göran Gustafsson Foundation.

†Information Technology, Uppsala University, Box 337, SE-751 05, Uppsala, Sweden (fredrik.hellman@it.uu.se). Supported by the Centre for Interdisciplinary Mathematics, Uppsala University.

‡Department of Mathematical Sciences, Chalmers University of Technology and University of Gothenburg, SE-412 96 Göteborg, Sweden (axel@chalmers.se). Supported by the Swedish Research Council.

critical value. The goal of this paper is to develop an efficient and accurate multilevel Monte Carlo (MLMC) method to find the failure probability. We focus mainly on the case when the model is a partial differential equation (PDE) and we use terminology from the discipline of numerical methods for PDEs. However, the methodology presented here is also applicable in a more general setting.

A multilevel Monte Carlo method inherits the non-intrusive and non-parametric characteristics from the standard Monte Carlo (MC) method. This allows the method to be used for complex black-box problems for which intrusive analysis is difficult or impossible. The MLMC method uses a hierarchy of numerical approximations on different accuracy levels. The levels in the hierarchy are typically directly related to a grid size or timestep length. The key idea behind the MLMC method is to use low accuracy solutions as control variates for high accuracy solutions in order to construct an estimator with lower variance. Savings in computational cost are achieved when the low accuracy solutions are cheap and are sufficiently correlated with the high accuracy solutions. MLMC was first introduced in [10] for stochastic differential equations as a generalization of a two-level variance reduction technique introduced in [17]. The method has been applied to and analyzed for elliptic PDEs in [3, 5, 4, 19]. Further improvements of the MLMC method, such as work on optimal hierarchies, non-uniform meshes and more accurate error estimates can be found in [15, 6]. In the present paper, we are not interested in the expected value of the QoI, but instead a failure probability, which is essentially a single point evaluation of the cumulative distribution function (cdf). For extreme failure probabilities, related methods include importance sampling [14], importance splitting [13], and subset simulations [1]. Works more related to the present paper include the results on MLMC methods for computing payoffs of binary options [2] and non-parameteric density estimation for PDE models in [9], and in particular [8]. In the latter, the selective refinement method for quantiles was formulated and analyzed.

In this paper, we seek to compute the cdf at a given critical value. The cdf at the critical value can be expressed as the expectation value of a binomially distributed random variable Q that is equal to 1 if the QoI is smaller than the critical value, and 0 otherwise. The key idea behind selective refinement is that realizations with QoI far from the critical value can be solved to a lower accuracy than those close to the critical value, and still yield the same value of Q. The random variable Q lacks regularity with respect to the uncertain input data, and hence we are in an unfavorable situation for application of the MLMC method. However, with the computational savings from the selective refinement it is still possible to obtain an asymptotic result for the computational cost where the cost for the full estimator is proportional to

the cost for a single realization to the highest accuracy.

The paper is structured as follows. Section 2 presents the necessary assumptions and the precise problem description. It is followed by Section 3 where our particular failure probability functional is defined and analyzed for the MLMC method. In Section 4 and Section 5 we revisit the multilevel Monte Carlo and selective refinement method adapted to this problem and in Section 6 we show how to combine multilevel Monte Carlo with the selective refinement to obtain optimal computational cost. In Section 7 we give details on how to implement the method in practice. The paper is concluded with two numerical experiments in Section 8.

2 Problem formulation

We consider a model problem $\mathcal{M}$, e.g., a (non-)linear differential operator with uncertain data. We let u denote the solution to the model

$$\mathcal{M}(\omega, u) = 0,$$

where the data ω is sampled from a space Ω. In what follows we assume that there exists a unique solution u given any $\omega \in \Omega$ almost surely. It follows that the solution u to a given model problem $\mathcal{M}$ is a random variable which can be parameterized in ω, i.e., $u = u(\omega)$.

The focus of this work is to compute failure probabilities, i.e., we are not interested in some pointwise estimate of the expected value of the solution, $\mathbb{E}[u]$, but rather the probability that a given QoI expressed as a functional, $X(u)$ of the solution u, is less (or greater) than some given critical value y. We let F denote the cdf of the random variable $X = X(\omega)$. The failure probability is then given by

$$p = F(y) = \Pr(X \leq y). \tag{1}$$

The following example illustrates how the problem description relates to real world problems.

Example 1. *As an example, geological sequestration of carbon dioxide (CO_2) is performed by injection of CO_2 in an underground reservoir. The fate of the CO_2 determines the success or failure of the storage system. The CO_2 propagation is often modeled as a PDE with random input data, such as a random permeability field. Typical QoIs include reservoir breakthrough time or pressure at a fault. The value y corresponds to a critical value which the QoI may not exceed or go below. In the breakthrough time case, low values are considered failure. In the pressure case, high values are considered failure.*

In that case one should negate the QoI to transform the problem to the form of equation (1).

The only regularity assumption on the model is the following Lipschitz continuity assumption of the cdf, which is assumed to hold throughout the paper.

Assumption 2. *For any* $x, y \in \mathbb{R}$,

$$|F(x) - F(y)| \leq C_L |x - y|. \tag{2}$$

To compute the failure probability we consider the binomially distributed variable $Q = \mathbb{1}(X \leq y)$ which takes the value 1 if $X \leq y$ and 0 otherwise. The cdf can be expressed as the expected value of Q, i.e., $p = F(y) = \mathbb{E}[Q]$. In practice we construct an estimator $\widehat{Q}$ for $\mathbb{E}[Q]$, based on approximate sample values from X. As such, $\widehat{Q}$ often suffers from numerical bias from the approximation in the underlying sample. Our goal is to compute the estimator $\widehat{Q}$ to a given root mean square error (RMSE) tolerance ϵ, i.e., to compute

$$e\Big[\widehat{Q}\Big] = \left(\mathbb{E}\left[\left(\widehat{Q} - \mathbb{E}[Q]\right)^2\right]\right)^{1/2} = \left(\mathbb{V}\Big[\widehat{Q}\Big] + \left(\mathbb{E}\Big[\widehat{Q} - Q\Big]\right)^2\right)^{1/2} \leq \epsilon$$

to a minimal computational cost. The equality above shows a standard way of splitting the RMSE into a stochastic error and numerical bias contribution.

The next section presents assumptions and results regarding the numerical discretization of the particular failure probability functional Q.

3 Approximate failure probability functional

We will not consider a particular approximation technique for computing $\widehat{Q}$, but instead make some abstract assumptions on the underlying discretization. We introduce a hierarchy of refinement levels $\ell = 0, 1, \ldots$ and let X'_ℓ and $Q'_\ell = \mathbb{1}(X'_\ell \leq y)$ be an approximate QoI of the model, and approximate failure probability, respectively, on level ℓ. One possible and natural way to define the accuracy on level ℓ is by assuming

$$|X - X'_\ell| \leq \gamma^\ell, \tag{3}$$

for some $0 < \gamma < 1$. This means the error of all realizations on level ℓ are uniformly bounded by γ^ℓ. In a PDE setting, typically an a priori error bound or a posteriori error estimate,

$$|X(\omega) - X_h(\omega)| \leq C(\omega) h^s,$$

can be derived for some constants $C(\omega)$, s, and a discretization parameter h. Then we can choose $X'_\ell = X_h$ with $h = \left(C(\omega)^{-1}\gamma^\ell\right)^{1/s}$ to fulfill (3).

For an accurate value of the failure probability functional the condition in (3) is unnecessarily strong. This functional is very sensitive to perturbations of values close to y, but insensitive to perturbations for values far from y. This insensitivity can be exploited. We introduce a different approximation X_ℓ, and impose the following, relaxed, assumption on this approximation of X, which allows for larger errors far from the critical value y. This assumption is illustrated in Figure 1.

Assumption 3. *The numerical approximation X_ℓ of X satisfies*

$$|X - X_\ell| \leq \gamma^\ell \quad \text{or} \quad |X - X_\ell| < |X_\ell - y| \tag{4}$$

for a fix $0 < \gamma < 1$.

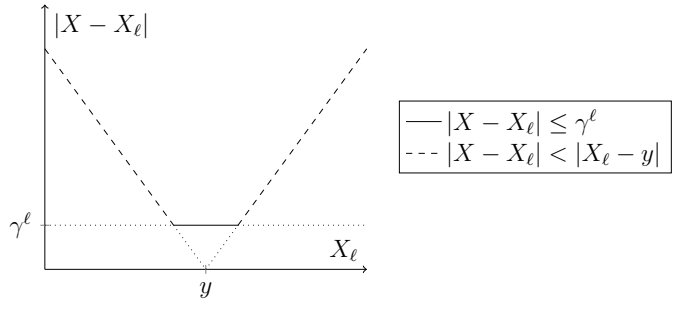


Figure 1: Illustration of condition (4). The numerical error is allowed to be larger than γ^ℓ far away from y.

We define $Q_\ell = \mathbb{1}(X_\ell \leq y)$ analogously to Q'_ℓ. Let us compare the implications of the two conditions (3) and (4) on the quality of the two respective approximations. Denote by X'_ℓ and Q'_ℓ stochastic variables obeying the error bound (3) and its corresponding approximate failure functional, respectively, and let X_ℓ obey (4). In a practical situation, Assumption 3 is fulfilled by iterative refinements of X_ℓ until condition (4) is satisfied. It is natural to use a similar procedure to achieve the stricter condition (3) for X'_ℓ. We express this latter assumption of using similar procedures for computing X_ℓ and X'_ℓ as

$$|X - X_\ell| \leq \gamma^\ell \text{ implies } X'_\ell = X_\ell, \tag{5}$$

i.e., for outcomes where X_ℓ is solved to accuracy γ^ℓ, X'_ℓ is equal to X_ℓ. Under that assumption, the following lemma shows that it is not less probable that Q_ℓ is correct than that Q'_ℓ is.

Lemma 4. *Let X'_ℓ and X_ℓ fulfill* (3) *and* (4)*, respectively, and assume* (5) *holds. Then* $\Pr(Q_\ell = Q) \geq \Pr(Q'_\ell = Q)$.

Proof. We split Ω into the events $A = \{\omega \in \Omega : |X - X_\ell| \leq \gamma^\ell\}$ and its complement $\Omega \setminus A$. For $\omega \in A$, using (5), we conclude that $Q'_\ell = Q_\ell$, hence

$$\Pr(Q_\ell = Q \mid A) = \Pr(Q'_\ell = Q \mid A).$$

For $\omega \notin A$, we have $|X - X_\ell| > \gamma^\ell$, and from (4) that $|X - X_\ell| < |X_\ell - y|$, i.e., $Q_\ell = Q$ and hence

$$\Pr(Q_\ell = Q \mid \Omega \setminus A) = 1.$$

Since $\Pr(Q'_\ell = Q \mid \Omega \setminus A) \leq 1$, we get $\Pr(Q_\ell = Q) \geq \Pr(Q'_\ell = Q)$. □

Under Assumption 3 we can prove the following lemma on the accuracy of the failure probability function Q_ℓ.

Lemma 5. *Under Assumption 2 and 3, the statements*

M1 $|\mathbb{E}[Q_\ell - Q]| \leq C_1\gamma^\ell$,

M2 $\mathbb{V}[Q_\ell - Q_{\ell-1}] \leq C_2\gamma^\ell$ *for* $\ell \geq 1$,

are satisfied where C_1 and C_2 do not depend on ℓ.

Proof. We split Ω into the events $B = \{\omega \in \Omega : \gamma^\ell \geq |X_\ell - y|\}$ and its complement $\Omega \setminus B$. In $\Omega \setminus B$, we have $Q_\ell = Q$, since $|X - X_\ell| < |X_\ell - y|$ from (4). Also, we note that the event B implies $|X - X_\ell| \leq \gamma^\ell$, hence $|X - y| \leq 2\gamma^\ell$. Then,

$$\begin{aligned}
|\mathbb{E}[Q_\ell - Q]| &= \left| \int_B Q_\ell(\omega) - Q(\omega) \, \mathrm{d}P(\omega) \right| \leq \int_B 1 \, \mathrm{d}P(\omega) \\
&\leq \Pr(|X - y| \leq 2\gamma^\ell) = F(y - 2\gamma^\ell) - F(y + 2\gamma^\ell) \\
&\leq 4C_L\gamma^\ell,
\end{aligned}$$

which proves **M1**. **M2** follows directly from **M1**, since

$$\begin{aligned}
\mathbb{V}[Q_\ell - Q_{\ell-1}] &= \mathbb{E}\big[(Q_\ell - Q_{\ell-1})^2\big] - \mathbb{E}[Q_\ell - Q_{\ell-1}]^2 \\
&\leq \mathbb{E}[Q_\ell - 2Q_\ell Q_{\ell-1} + Q_{\ell-1}] \\
&\leq |\mathbb{E}[Q_\ell - Q]| + |2\mathbb{E}[Q_\ell Q_{\ell-1} - Q]| + |\mathbb{E}[Q_{\ell-1} - Q]| \\
&\leq 2|\mathbb{E}[Q_\ell - Q]| + 2|\mathbb{E}[Q_{\ell-1} - Q]| \\
&\leq C_2\gamma^\ell,
\end{aligned}$$

where $(Q_\ell)^2 = Q_\ell$ was used. □

Interesting to note with this particular failure probability functional is that the convergence rate in **M2** cannot be improved if the rate in **M1** is already sharp, as the following lemma shows.

Lemma 6. *Let $0 < \gamma < 1$ be fixed. If there is a $0 < c \leq C_1$ such that the failure probability functional satisfies*

$$c\gamma^\ell \leq |\mathbb{E}[Q_\ell - Q]| \leq C_1\gamma^\ell$$

for all $\ell = 0, 1, \ldots$, then

$$\mathbb{V}[Q_\ell - Q_{\ell-1}] \leq C_2\gamma^{\beta\ell},$$

where $\beta = 1$ is sharp in the sense that the relation will be violated for sufficiently large ℓ, if $\beta > 1$.

Proof. Assume that $\mathbb{V}[Q_\ell - Q_{\ell-1}] \leq C\gamma^{\beta\ell}$ for some constant C and $\beta > 1$. For two levels $k < \ell$, such that $c\gamma^k > C_1\gamma^\ell$ we have that

$$|\mathbb{E}[Q_\ell - Q_k]| \geq ||\mathbb{E}[Q_\ell - Q]| - |\mathbb{E}[Q_k - Q]|| \geq \big(c - C_1\gamma^{\ell-k}\big)\gamma^k = \tilde{c}\gamma^k,$$

with $\tilde{c} = c - C_1\gamma^{\ell-k} > 0$. For such ℓ and k, we have

$$\begin{aligned}
\tilde{c}\gamma^k \leq |\mathbb{E}[Q_\ell - Q_k]| &\leq \sum_{j=k}^{\ell-1} |\mathbb{E}[Q_{j+1} - Q_j]| \leq \sum_{j=k}^{\ell-1} \mathbb{E}\big[(Q_{j+1} - Q_j)^2\big] \\
&= \sum_{j=k}^{\ell-1} \big(\mathbb{V}[Q_{j+1} - Q_j] + (\mathbb{E}[Q_{j+1} - Q_j])^2\big) \\
&\leq \sum_{j=k}^{\ell-1} \big(C\gamma^{\beta j} + \mathcal{O}(\gamma^{2j})\big) \leq \widetilde{C}\gamma^{\beta k} + \mathcal{O}(\gamma^{2k}).
\end{aligned}$$

For $\ell, k \to \infty$ (keeping $\ell - k$ constant) we have a contradiction due to the mismatching rates and hence $\beta \leq 1$, which proves that the bound can not be improved. □

4 Multilevel Monte Carlo method

In this section, we present the multilevel Monte Carlo method in a general context. Because of the low convergence rate of the variance in **M2**, the MLMC method does not perform optimally for the failure probability functional. The results presented here will be combined with the results from

Section 5 to derive a new method to compute failure probabilities efficiently in Section 6.

The (standard) MC estimator at refinement level ℓ of $\mathbb{E}[Q]$ using a sample $\{\omega_\ell^i\}_{i=1}^{N_\ell}$, reads

$$\widehat{Q}_{N_\ell,\ell}^{MC} = \frac{1}{N_\ell}\sum_{i=1}^{N_\ell} Q_\ell(\omega_\ell^i).$$

Note that the subscripts N_ℓ and ℓ control the statistical error and numerical bias, respectively. The expected value and variance of the estimator $\widehat{Q}_{N_\ell,\ell}^{MC}$ are $\mathbb{E}\Big[\widehat{Q}_{N_\ell,\ell}^{MC}\Big] = \mathbb{E}[Q_\ell]$ and $\mathbb{V}\Big[\widehat{Q}_{N_\ell,\ell}^{MC}\Big] = N_\ell^{-1}\mathbb{V}[Q_\ell]$, respectively. Referring to the goal of the paper, we want the MSE (square of the RMSE) to satisfy

$$e\Big[\widehat{Q}_{N_\ell,\ell}^{MC}\Big]^2 = N_\ell^{-1}\mathbb{V}[Q_\ell] + (\mathbb{E}[Q_\ell - Q])^2 \leq \epsilon^2/2 + \epsilon^2/2 = \epsilon^2,$$

i.e., both the statistical error and the numerical error should be less than $\epsilon^2/2$. The MLMC method is a variance reduction technique for the MC method. The MLMC estimator $\widehat{Q}_{\{N_\ell\},L}^{ML}$ at refinement level L is expressed as a telescoping sum of L MC estimator correctors:

$$\widehat{Q}_{\{N_\ell\},L}^{ML} = \sum_{\ell=0}^{L}\frac{1}{N_\ell}\sum_{i=1}^{N_\ell}\big(Q_\ell(\omega_\ell^i) - Q_{\ell-1}(\omega_\ell^i)\big),$$

where $Q_{-1} = 0$. There is one corrector for every refinement level $\ell = 0, \ldots, L$, each with a specific MC estimator sample size N_ℓ. The expected value and variance of the MLMC estimator are

$$\begin{aligned}\mathbb{E}\Big[\widehat{Q}_{\{N_\ell\},L}^{ML}\Big] &= \sum_{\ell=0}^{L}\mathbb{E}[Q_\ell - Q_{\ell-1}] = \mathbb{E}[Q_L] \quad \text{and} \\ \mathbb{V}\Big[\widehat{Q}_{\{N_\ell\},L}^{ML}\Big] &= \sum_{\ell=0}^{L} N_\ell^{-1}\mathbb{V}[Q_\ell - Q_{\ell-1}],\end{aligned} \tag{6}$$

respectively. Using (6) the MSE for the MLMC estimator can be expressed as

$$e\Big[\widehat{Q}_{\{N_\ell\},L}^{ML}\Big]^2 = \sum_{\ell=0}^{L} N_\ell^{-1}\mathbb{V}[Q_\ell - Q_{\ell-1}] + (\mathbb{E}[Q_L - Q])^2,$$

and can be computed at expected cost

$$\mathcal{C}\Big[\widehat{Q}_{\{N_\ell\},L}^{ML}\Big] = \sum_{\ell=0}^{L} N_\ell c_\ell,$$

where $c_\ell = \mathcal{C}[Q_\ell] + \mathcal{C}[Q_{\ell-1}]$. Here, by $\mathcal{C}[\cdot]$ we denote the expected computational cost to compute a certain quantity. Given that the variance of the MLMC estimator is $\epsilon^2/2$ the expected cost is minimized by choosing

$$N_\ell = 2\epsilon^{-2}\sqrt{\mathbb{V}[Q_\ell - Q_{\ell-1}]/c_\ell}\sum_{k=0}^{L}\sqrt{\mathbb{V}[Q_k - Q_{k-1}]c_k} \tag{7}$$

(see Appendix A), and hence the total expected cost is

$$\mathcal{C}\left[\widehat{Q}^{ML}_{\{N_\ell\},L}\right] = 2\epsilon^{-2}\left(\sum_{\ell=0}^{L}\sqrt{\mathbb{V}[Q_\ell - Q_{\ell-1}]c_\ell}\right)^2. \tag{8}$$

If the product $\mathbb{V}[Q_\ell - Q_{\ell-1}]c_\ell$ increases (or decreases) with ℓ then dominating term in (8) will be $\ell = L$ (or $\ell = 0$). The values N_ℓ can be estimated on the fly in the MLMC algorithm using (7) while the cost c_ℓ can be estimated using an a priori model. The computational complexity to obtain a RMSE less than ϵ of the MLMC estimator for the failure probability functional is given by the theorem below. In the following, the notation $a \lesssim b$ stands for $a \leq Cb$ with some constant C independent of ϵ and ℓ.

Theorem 7. *Let Assumption 2 and 3 hold (so that Lemma 5 holds) and $\mathcal{C}[Q_\ell] \lesssim \gamma^{-r\ell}$. Then there exists a constant L and a sequence $\{N_\ell\}$ such that the RMSE is less than ϵ, and the expected cost of the MLMC estimator is*

$$\mathcal{C}\left[\widehat{Q}^{ML}_{\{N_\ell\},L}\right] \lesssim \begin{cases} \epsilon^{-2} & r < 1 \\ \epsilon^{-2}(\log \epsilon^{-1})^2 & r = 1 \\ \epsilon^{-1-r} & r > 1. \end{cases} \tag{9}$$

Proof. For a proof see, e.g., [5, 10]. □

The most straight-forward procedure to fulfill Assumption 3 in practice is to refine all samples on level ℓ uniformly to an error tolerance γ^ℓ, i.e., to compute X'_ℓ introduced in Section 3, for which $|X - X'_\ell| \leq \gamma^\ell$. Typical numerical schemes for computing X'_ℓ include finite element, finite volume, or finite difference schemes. Then the expected cost $\mathcal{C}[Q'_\ell]$ typically fulfill

$$\mathcal{C}[Q'_\ell] = \gamma^{-q\ell}, \tag{10}$$

where q depends on the physical dimension of the computational domain, the convergence rate of the solution method, and computational complexity for assembling and solving the linear system. Note that one unit of work is normalized according to equation (10). Using Theorem 7, with Q'_ℓ instead

of Q_ℓ (which is possible, since Q'_ℓ trivially fulfills Assumption 3) we obtain a RMSE of the expected cost less than $\epsilon^{-1-q} = \epsilon^{-1}\mathcal{C}[Q'_\ell]$ for the case $q > 1$.

In the next section we describe how the selective refinement algorithm computes X_ℓ (hence Q_ℓ) that fulfills Assumption 3 to a lower cost than its fully refined equivalent X'_ℓ. The theorem above can then be applied with $r = q - 1$ instead of $r = q$.

5 Selective refinement algorithm

In this section we modify the selective refinement algorithm proposed in [8] for computing failure probabilities (instead of quantiles) and for quantifying the error using the RMSE. The selective refinement algorithm computes X_ℓ so that

$$|X - X_\ell| \leq \gamma^\ell \quad \text{or} \quad |X - X_\ell| < |X_\ell - y|$$

in Assumption 3 is fulfilled without requiring the stronger (full refinement) condition

$$|X - X_\ell| \leq \gamma^\ell.$$

In contrast to the selective refinement algorithm in [8], Assumption 3 can be fulfilled by iterative refinement of realizations over all realizations independently. This allows for an efficient totally parallell implementation. We are particularly interested in quantifying the expected cost required by the selective refinement algorithm, and showing that the X_ℓ resulting from the algorithm fulfills Assumption 3.

Algorithm 1 exploits the fact that $Q_\ell = Q$ for realizations satisfying $|X - X_\ell| < |X_\ell - y|$. That is, even if the error of X_ℓ is greater than γ^ℓ, it might be sufficiently accurate to yield the correct value of Q_ℓ. The algorithm works on a per-realization basis, starting with an error tolerance 1. The realization is refined iteratively until Assumption 3 is fulfilled. The advantage is that many samples can be solved only with low accuracy and hence the average cost per Q_ℓ is reduced. Lemma 8 shows that X_ℓ computed using Algorithm 1 satisfies Assumption 3.

Lemma 8. *Approximations X_ℓ computed using Algorithm 1 satisfy Assumption 3.*

Proof. At each iteration in the while-loop of Algorithm 1, γ^j is the error tolerance of $X_\ell(\omega_\ell^i)$, i.e., $|X(\omega_\ell^i) - X_\ell(\omega_\ell^i)| \leq \gamma^j$. The stopping criterion hence implies Assumption 3 for $X_\ell(\omega_\ell^i)$. □

The expected cost for computing Q_ℓ using Algorithm 1 is given by the following lemma.

Algorithm 1 Selective refinement algorithm

1: Input arguments: level ℓ, realization i, critical value y, and tolerance factor γ
2: Compute $X'_0(\omega_\ell^i)$
3: Let $j = 0$
4: **while** $j < \ell$ and $\gamma^j \geq |X'_j(\omega_\ell^i) - y|$ **do**
5: Let $j = j + 1$
6: Compute $X'_j(\omega_\ell^i)$
7: **end while**
8: Let $X_\ell(\omega_\ell^i) = X'_j(\omega_\ell^i)$

Lemma 9. *The expected cost to compute the failure probability functional using Algorithm 1 can be bounded as*

$$\mathcal{C}[Q_\ell] \lesssim \sum_{j=0}^{\ell} \gamma^{(1-q)j}.$$

Proof. Consider iteration j, i.e., when $X_\ell(\omega_\ell^i)$ has been computed to tolerance γ^{j-1}. We denote by E_j the probability that a realization enters iteration j. For $j \leq \ell$,

$$\begin{aligned}\Pr(E_j) &= \Pr(y - \gamma^{j-1} \leq X_\ell \leq y + \gamma^{j-1}) \\ &\leq \Pr(y - 2\gamma^{j-1} \leq X \leq y + 2\gamma^{j-1}) \\ &= F(y + 2\gamma^{j-1}) - F(y - 2\gamma^{j-1}) \\ &\leq 4C_L\gamma^{j-1}.\end{aligned}$$

Every realization is initially solved to tolerance 1. Using that the cost for solving a realization to tolerance γ^j is γ^{-qj}, we get that the expected cost is

$$\mathcal{C}[Q_\ell] = 1 + \sum_{j=1}^{\ell} \Pr(E_j)\gamma^{-qj} \leq 1 + \sum_{j=1}^{\ell} 4C_L\gamma^{j-1}\gamma^{-qj} \lesssim \sum_{j=0}^{\ell} \gamma^{(1-q)j}$$

which concludes the proof. □

6 Multilevel Monte Carlo using the selective refinement strategy

Combining the MLMC method with the algorithm for selective refinement there can be further savings in computational cost. We call this method

multilevel Monte Carlo with selective refinement (MLMC-SR). In particular, for $q > 1$ we obtain from Lemma 9 that the expected cost for one sample can be bounded as

$$\mathcal{C}[Q_\ell] \lesssim \sum_{j=0}^{\ell} \gamma^{(1-q)j} \lesssim \gamma^{(1-q)\ell}. \tag{11}$$

Applying Theorem 7 with $r = q - 1$ yields the following result.

Theorem 10. *Let Assumption 2 and Assumption 3 hold (so that Lemma 5 holds) and suppose that Algorithm 1 is executed to compute Q_ℓ. Then there exists a constant L and a sequence $\{N_\ell\}$ such that the RMSE is less than ϵ, and the expected cost for the MLMC estimator with selective refinement is*

$$\mathcal{C}\left[\widehat{Q}^{ML}_{\{N_\ell\},L}\right] \lesssim \begin{cases} \epsilon^{-2} & q < 2 \\ \epsilon^{-2}(\log \epsilon^{-1})^2 & q = 2 \\ \epsilon^{-q} & q > 2. \end{cases} \tag{12}$$

Proof. For $q > 1$, follows directly from Theorem 7 since Lemma 5 holds with $r = q - 1$. For $q \leq 1$, we use the rate ϵ^{-2} from the case $1 < q < 2$, since the cost cannot be worsened by making each sample cheaper to compute. □

In a standard MC method we have $\epsilon^{-2} \sim N$ where N is the number of samples and $\epsilon^{-q} \sim \mathcal{C}[Q'_L]$ where $\mathcal{C}[Q'_L]$ is the expected computational cost for solving one realization on the finest level without selective refinement. The MLMC-SR method then has the following cost,

$$\mathcal{C}\left[\widehat{Q}^{ML}_{\{N_\ell\},L}\right] \lesssim \begin{cases} N & q < 2 \\ \mathcal{C}[Q'_L] & q > 2. \end{cases} \tag{13}$$

A comparison of MC, MLMC with full refinement (MLMC), and MLMC with selective refinement (MLMC-SR), is given in Table 1. To summarize, the best possible scenario is when the cost is ϵ^{-2}, which is equivalent with a standard MC method where all samples can be obtained with cost 1. This complexity is obtained for the MLMC method when $q < 1$ and for the MLMC-SR method when $q < 2$. For $q > 2$ the MC method has the same complexity as solving N problem on the finest level $N\mathcal{C}[Q'_L]$, MLMC has the same cost as $N^{1/2}$ problem on the finest level $N^{1/2}\mathcal{C}[Q'_L]$, and MLMC-SR method as solving one problem on the finest level $\mathcal{C}[Q'_L]$.

Method	$0 \leq q < 1$	$1 < q < 2$	$q > 2$
MC	ϵ^{-2-q}	ϵ^{-2-q}	ϵ^{-2-q}
MLMC	ϵ^{-2}	ϵ^{-1-q}	ϵ^{-1-q}
MLMC-SR	ϵ^{-2}	ϵ^{-2}	ϵ^{-q}

Table 1: Comparison of work between MC, MLMC with full refinement (MLMC), and MLMC with selective refinement (MLMC-SR) for different q.

7 Heuristic algorithm

In this section, we present a heuristic algorithm for the MLMC method with selective refinement. Contrary to Theorem 10, this algorithm does not guarantee that the RMSE is $\mathcal{O}(\epsilon)$, since we in practice lack a priori knowledge of the constants C_1 and C_2 in Lemma 5. Instead, the RMSE needs to be estimated. Recall the split of the MSE into a numerical and statistical contribution:

$$\left(\mathbb{E}\Big[Q - \widehat{Q}\Big]\right)^2 \leq \frac{1}{2}\epsilon^2 \quad \text{and} \quad \mathbb{V}\Big[\widehat{Q}\Big] \leq \frac{1}{2}\epsilon^2. \tag{14}$$

With $\widehat{Q}$ being the multilevel Monte Carlo estimator $\widehat{Q}^{ML}_{\{N_\ell\},L}$, we here present heuristics for estimating the numerical and statistical error of the estimator.

For both estimates and $\ell \geq 1$, we make use of the trinomially distributed variable $Y_\ell(\omega) = Q_\ell(\omega) - Q_{\ell-1}(\omega)$. We denote the probabilities for Y_ℓ to be -1, 0 and 1 by p_{-1}, p_0 and p_1, respectively. For convenience, we drop the index ℓ for the probabilities, however, they do depend on ℓ. In order to estimate the numerical bias $\mathbb{E}\Big[Q - \widehat{Q}^{ML}_{\{N_\ell\},L}\Big] = \mathbb{E}[Q - Q_L]$, we assume that **M1** holds approximately with equality, i.e., $|\mathbb{E}[Q - Q_\ell]| \approx C_1\gamma^\ell$. Then the numerical bias can be overestimated, $|\mathbb{E}[Q - Q_\ell]| \leq |\mathbb{E}[Y_\ell]|(\gamma^{-1} - 1)^{-1}$, since

$$\begin{aligned}
|\mathbb{E}[Y_\ell]| &= |\mathbb{E}[Q_\ell - Q] - \mathbb{E}[Q_{\ell-1} - Q]| \\
&\geq ||\mathbb{E}[Q_\ell - Q]| - |\mathbb{E}[Q_{\ell-1} - Q]|| \\
&\approx \left|C_1\gamma^\ell - C_1\gamma^{\ell-1}\right| \\
&= C_1\gamma^\ell(\gamma^{-1} - 1).
\end{aligned}$$

Hence, we concentrate our effort on estimating $|\mathbb{E}[Y_\ell]|$.

It has been observed that the accuracy of sample estimates of mean and variance of Y_ℓ might deteriorate for deep levels $\ell \gg 1$, and a continuation multilevel Monte Carlo method was proposed in [6] as a remedy for this. That

idea could be applied and specialized for this functional to obtain more accurate estimates. However, in this work we use the properties of the trinomially distributed Y_ℓ to construct a method with optimal asymptotic behavior, possibly with increase of computational cost by a constant.

We consider the three binomial distributions $[Y_\ell = 1]$, $[Y_\ell = -1]$ and $[Y_\ell \neq 0]$ which have parameters p_1, p_{-1} and $p_1 + p_{-1}$, respectively ($[\cdot]$ is the Iverson bracket notation). These parameters can be used in estimates for both the expectation value and variance of the trinomially distributed Y_ℓ. Considering a general binomial distribution $B(n, p)$, we want to estimate p. For our distributions, as the level ℓ increases, p approaches zero, why we are concerned with finding stable estimates for small p. It is important that the parameter is not underestimated, since it is used to control the numerical bias and statistical error and could then cause premature termination. We propose an estimation method that is easy to implement, and that will overestimate the parameter in case of accuracy problems, rather than underestimate it, while keeping the asymptotic rates given in Lemma 5 for the estimators.

The standard unbiased estimator of p is $\hat{p} = xn^{-1}$, where x is the number of observed successes. The proposed alternative (and biased) estimator is $\tilde{p} = (x + k)(n + k)^{-1}$ for a $k > 0$. This corresponds to a Bayesian estimate with prior beta distribution with parameters $(k + 1, 1)$. Observing that

$$\begin{aligned} |\mathbb{E}[Y_\ell]| &= |p_1 - p_{-1}|, \\ \mathbb{V}[Y_\ell] &= p_1 + p_{-1} - (p_1 - p_{-1})^2 \end{aligned} \tag{15}$$

and considering Lemma 5 (assuming equality with the rates), we conclude that all three parameters $p \propto \gamma^\ell$ (where $\propto$ means asymptotically proportional to, for $\ell \gg 1$). With the standard estimator $\hat{p}$, the relative variance can be expressed as $\mathbb{V}[\hat{p}](\mathbb{E}[\hat{p}])^{-2}$. This quantity should be less than one for an accurate estimate. We now examine its asymptotic behavior. The parameter n is the optimal number of samples at level ℓ (equation (7)) and can be expressed as

$$n \propto \gamma^{\frac{1}{2}\ell q - \frac{1}{2}L(2+q)}, \tag{16}$$

where we used that $\epsilon \propto \gamma^L$, $\mathcal{C}[Y_\ell] \propto \gamma^{(1-q)\ell}$ and $\mathbb{V}[Y_\ell] \propto \gamma^\ell$. Then we have

$$\frac{\mathbb{V}[\hat{p}]}{\mathbb{E}[\hat{p}]^2} = \frac{n^{-1}p(1-p)}{p^2} = \frac{1-p}{np} \propto \gamma^{\frac{2+q}{2}(L-\ell)}.$$

In particular, for $\ell = L$, the relative variance is asymptotically constant, but we don't know a priori how big this constant is. When it is large (greater

than 1), the relative variance of $\hat{p}$ might be very large. An analogous analysis on $\tilde{p}$ yields

$$\frac{\mathbb{V}[\tilde{p}]}{\mathbb{E}[\tilde{p}]^2} = \frac{(n+k)^{-2}np(1-p)}{(n+k)^{-2}(np+k)^2} = \frac{np(1-p)}{(np+k)^2} \leq \frac{np}{(np+k)^2}. \tag{17}$$

Maximizing the bound in (17) with respect to np, gives

$$\frac{\mathbb{V}[\tilde{p}]}{\mathbb{E}[\tilde{p}]^2} \leq \frac{1}{4k}.$$

Choosing for instance $k = 1$ gives a maximum relative variance of $1/4$. Choosing a larger k gives larger bias, but smaller relative variance. The bias of this estimator is significant if $np \ll k$, however, that is the case when we have too few samples to estimate the parameter accurately, and then $\tilde{p}$ instead acts as a bound. The estimate $\tilde{p}$ keeps the asymptotic behavior $\mathbb{E}[\tilde{p}] \propto \gamma^\ell$, since

$$\begin{aligned}\mathbb{E}[\tilde{p}] &= \frac{np+k}{n+k} \propto \frac{np+k}{n} = p + \frac{k}{n} \\ &\propto \gamma^\ell + \gamma^{-\frac{1}{2}\ell q + \frac{1}{2}L(2+q)} = \gamma^\ell(1 + \gamma^{\frac{1}{2}(L-\ell)(2+q)}) \leq 2\gamma^\ell \propto p\end{aligned}$$

where we used that $\ell < L$ and k is constant.

Now, estimating the parameters p_1, p_{-1} and $p_1 + p_{-1}$ as $\tilde{p}_1$, $\tilde{p}_{-1}$ and $\tilde{p}_{\pm 1}$, respectively, using the estimator $\tilde{p}$ above (note that the sum $p_1 + p_{-1}$ is estimated separately from p_1 and p_{-1}) we can bound (approximately) the expected value and variance of Y_ℓ in (15):

$$|\mathbb{E}[Y_\ell]| \leq \max(p_1, p_{-1}) \approx \max(\tilde{p}_1, \tilde{p}_{-1}) \tag{18}$$

and

$$\mathbb{V}[Y_\ell] \leq p_1 + p_{-1} \approx \tilde{p}_{\pm 1} \tag{19}$$

for $\ell \geq 1$. For $\ell = 0$, the sample size is usually large enough to use the sample mean and variance as accurate estimates. Since the asymptotic behavior of $\tilde{p}$ is γ^ℓ, the rates in Lemma 5 still holds and Theorem 10 applies (however, with approximate quantities).

The algorithm for the MLMC method using selective refinement is presented in Algorithm 2. The termination criterion is the same as was used in the standard MLMC algorithm [10], i.e.,

$$\max(\gamma|\mathbb{E}[Y_{L-1}]|, |\mathbb{E}[Y_L]|) < \frac{1}{\sqrt{2}}(\gamma^{-1} - 1)\epsilon, \tag{20}$$

where $|\mathbb{E}[Y_{L-1}]|$ and $|\mathbb{E}[Y_L]|$ are estimated using the methods presented above. A difference from the standard MLMC algorithm is that the initial sample

size for level L is $N_L = N\gamma^{-L}$ instead of $N_L = N$, for some N. This is what is predicted by equation (16) and is necessary to provide accurate estimates of the expectation value and variance of Y_ℓ for deep levels. Other differences from the standard MLMC algorithm is that the selective refinement algorithm (Algorithm 1) is used to compute $\widehat{Q}^{MC}_{N_\ell,L}$, and that the estimates of expectation value and variance of Y_ℓ are computed according to the discussion above.

Algorithm 2 MLMC method using selective refinement

1: Pick critical value y, cost model parameter q, tolerance factor γ, initial number of samples N, parameter k, and final tolerance ϵ
2: Set $L = 0$
3: **loop**
4: Let $N_L = N\gamma^{-L}$ and compute $\widehat{Q}^{MC}_{N_\ell,L}$ using selective refinement (Algorithm 1)
5: Estimate $\mathbb{V}[Q_\ell - Q_{\ell-1}]$ using (18)
6: Compute the optimal $\{N_\ell\}_{\ell=0}^{L}$ using (7) and cost model (11)
7: Compute $\widehat{Q}^{MC}_{N_\ell,\ell}$ for all levels $\ell = 0, \ldots, L$ using selective refinement (Algorithm 1)
8: Estimate $\mathbb{E}[Q_\ell - Q_{\ell-1}]$ using (19)
9: Terminate if converged by checking inequality (20)
10: Set L = L + 1
11: **end loop**
12: The MLMC-SR estimator is $\widehat{Q}^{ML}_{\{N_\ell\},L} = \sum_{\ell=0}^{L} \widehat{Q}^{MC}_{N_\ell,\ell}$

8 Numerical experiments

Two types of numerical experiments are presented in this section. The first experiment (in Section 8.1) is performed on a simple and cheap model $\mathcal{M}$ so that the asymptotic results of the computational cost, derived in Theorem 10, can be verified. The second experiment (in Section 8.2) is performed on a PDE model $\mathcal{M}$ to show the method's applicability to realistic problems. In our experiments we made use of the software FEniCS [18] and SciPy [16].

8.1 Failure probability of a normal distribution

In this first demonstrational experiment, we let the quantity of interest X belong to the standard normal distribution and we seek to find the probability of $X \leq y = 0.8$. The true value of this probability is $\Pr(X \leq 0.8) =$

$\Phi(0.8) \approx 0.78814$ and we hence have a reliable reference solution. We define approximations X_h of X as follows. First, we let our input data ω belong to the standard normal distribution, and let $X(\omega) = \omega$. Then, we let $X_h(\omega) = \omega + h(2U(\omega, h) - 1 + b)/(1 + b)$, where $b = 0.1$ and $U(\omega, h)$ is a uniformly distributed random number between 0 and 1. Since we have an error bound $|X_h - X| \leq h$, the selective refinement algorithm (Algorithm 1) can be used to construct a function X_ℓ satisfying Assumption 3. With this setup it is very cheap to compute X_h to any accuracy h, however, for illustrational purposes we assume a cost model $\mathcal{C}[X_h] = h^{-q}$ with $q = 1$, 2, and 3 to cover the three cases in Theorem 10.

For the three values of q, and eight logarithmically distributed values of ϵ between 10^{-3} and 10^{-1}, we performed 100 runs of Algorithm 2. All parameters used in the simulations are presented in Table 2.

Parameter	Value
y	0.8
q	1, 2, 3
γ	0.5
N	10
k	1
ϵ	$(10^{-3}, 10^{-1})$

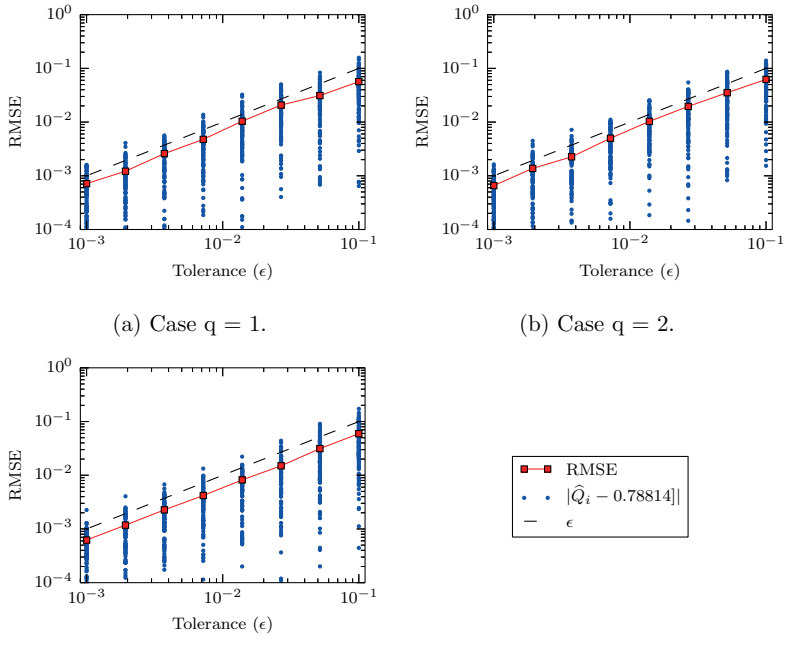
Table 2: Parameters used for the demonstrational experiment.

For convenience, we denote by $\widehat{Q}_i$ the MLMC-SR estimator $\widehat{Q}^{ML}_{\{N_\ell\},L}$ of the failure probability from run $i = 1, \ldots, M$ with $M = 100$. For each tolerance ϵ and cost parameter q, we estimated the RMSE of the MLMC-SR estimator by

$$e\Big[\widehat{Q}^{ML}_{\{N_\ell\},L}\Big] = \left(\mathbb{E}\left[\left(\widehat{Q}^{ML}_{\{N_\ell\},L} - \mathbb{E}[Q]\right)^2\right]\right)^{1/2} \approx \left(\frac{1}{M}\sum_{i=1}^{M}\Big(\widehat{Q}_i - \mathbb{E}[Q]\Big)^2\right)^{1/2}.$$

Also, for each of the eight tolerances ϵ, we computed the run-specific estimation errors $|\widehat{Q}_i - \mathbb{E}[Q]|$, $i = 1, \ldots, M$. In Figure 2 we present three plots of the RMSE vs. ϵ, one for each value of q. We can see that the method yields solutions with the correct accuracy.

In order to verify Theorem 10, we estimated the expected cost for each tolerance ϵ and value of q by computing the mean of the total cost over the 100 runs. The cost for each realization was computed using the cost model



(a) Case q = 1.

(b) Case q = 2.

(c) Case q = 3.

(d) Legend.

Figure 2: RMSE (square markers and line) plotted vs. tolerance for the experiment described in Section 8.1. The dashed line is the tolerance ϵ and the dots are the individual errors for the 100 runs at each tolerance.

in equation (10). The cost for realizations differs not only between levels ℓ, but also within a level ℓ owing to the selective refinement algorithm. For each run i, the costs of all realizations were summed to obtain the total cost for that run. We computed a mean of the total costs for the 100 runs. A plot of the result can be found in Figure 3. As the tolerance ϵ decreases the expected cost approaches the rates given in Theorem 10. The reference costs are multiplied by constants to align well with the estimated expected costs.

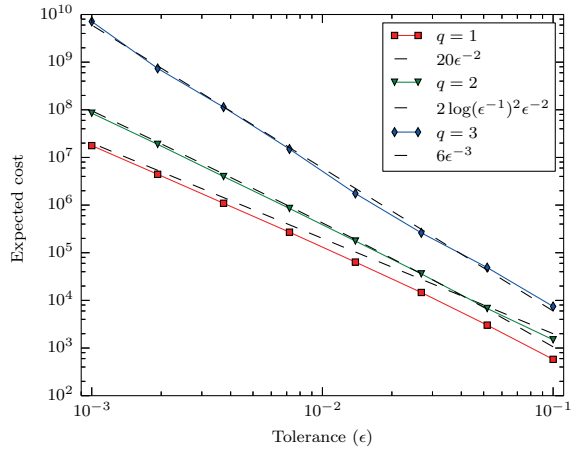


Figure 3: Computed mean total cost (diamond, triangle, square markers and lines) plotted with theoretical reference cost (dashed lines) for the experiment described in Section 8.1. The reference costs for the three values of q are: $20\epsilon^{-2}$ for $q = 1$; $2\log(\epsilon^{-1})^2\epsilon^{-2}$ for $q = 2$; and $6\epsilon^{-3}$ for $q = 3$.

8.2 Single-phase flow in media with lognormal permeability

We consider Darcy's law on a unit square $[0, 1]^2$ on which we have impearmeable upper and lower boundaries, high pressure on the left boundary (Γ_1) and low pressure on the right boundary (Γ_2). We define the spaces $H^1_f(\mathscr{D}) = \{v \in H^1(\mathscr{D}) : v|_{\Gamma_1} = f \text{ and } v|_{\Gamma_2} = 0\}$, and let n denote the unit normal of $\mathscr{D}$.

The weak form of the partial differential equation reads: find $u \in H^1_1(\mathscr{D})$ such that

$$(a(\omega, \cdot)\nabla u, \nabla v) = 0 \quad \text{in } \mathscr{D}, \tag{21}$$

for all $v \in H^1_0(\mathscr{D})$, and a is a stationary log-normal distributed random field

$$a(\omega, \cdot) = \exp(\kappa(\omega, \cdot)), \tag{22}$$

over $\mathscr{D}$, where $\kappa(\cdot, x)$ has zero mean and is normal distributed with exponential covariance, i.e., for all $x_1, x_2 \in \mathscr{D}$ we have that

$$\mathbb{V}[\kappa(\cdot, x_1)\kappa(\cdot, x_2)] = \sigma^2 \exp\left(\frac{-\|x_1 - x_2\|_2}{\rho}\right). \tag{23}$$

We choose $\sigma = 1$ and $\rho = 0.1$ in the numerical experiment.

We are interested in the boundary flux on the right boundary, i.e., the functional $X(\omega) = \int_{\Gamma_2} n \cdot a(\omega)\nabla u \, \mathrm{d}x = (a(\omega, \cdot)\nabla u, \nabla g)$, for any $g \in H^1(\mathscr{D})$, $g|_{\Gamma_1} = 0$ and $g|_{\Gamma_2} = 1$. The last equality comes by a generalized Green's identity, see [12, Chp. 1, Corollary 2.1].

To generate realizations of $a(\omega, \cdot)$, the circulant embedding method introduced in [7] is employed. The mesh resolution for the input data of the realizations generated on level ℓ in the MLMC-SR algorithm is chosen such that the finest mesh needed on level ℓ is not finer than the chosen mesh. For a fixed realization on level ℓ we don't know how fine data we need, because of the selective refinement procedure. This means that the complexity obtained for the MLMC-SR algorithm do not apply for the generation of data. The circulant embedding method has log-linear complexity. A remedy for the complexity of generating realizations is to use a truncated Karhunen-Loève expansion that can easily be refined. However, numerical experiments show that we are in a regime where the time spent on generating realizations using circulant embedding is negligible compared to the time spent in the linear solvers.

The PDE is discretized using a FEM-discretization with linear Lagrange elements. We have a family of structured nested meshes $\mathcal{T}_{h_m}$, where a mesh h_m is the maximum element diameter of the given mesh. The data $a(\omega, \cdot)$ is defined in the grid points of the meshes. Using the circulant embedding we get an exact representation of the stochastic field in the grid points of the given mesh. This can be interpreted as not making any approximation of the stochastic field but instead making a quadrature error when computing the bilinear form.

The functional for a discretization on mesh m is defined as $X_{h_m}(\omega) = (a(\omega, \cdot)\nabla u_{h_m}, \nabla g)$. The convergence rates in energy norm for log-normal data is $h^{1/2-\delta}$ for any $\delta > 0$ [4]. Using postprocessing, it can be shown that the error in the functional converges twice as fast [11], i.e, $|X_{h_m} - X_{h_m}(\omega)| \leq Ch^{s-2\delta}$ for $s = 1$. We use a multigrid solver that has linear $\alpha = 1$ (up to log-factors) complexity. The work for one sample can then be computed as $\gamma^{-q\ell}$ where γ^ℓ is the numerical bias tolerance for the sample and $q \approx 2\alpha/s = 2$, which was also verified numerically. The error is estimated using the dual solution computed on a finer mesh. Since it can be quite expensive to solve a dual problem for each realization of the data, the error in the functional can also be computed by estimating the constant C and s either numerically or theoretically.

We choose $\gamma = 0.5$, $N = 10$, and $k = 1$ in the MLMC-SR algorithm, see Section 7 for more information on the choices of parameters. The problem reads: find the probability p for $X \leq y = 1.5$ to the given RMSE ϵ. We

Parameter	Value
y	1.5
q	2
γ	0.5
N	10
k	1
ϵ	$10^{-1}, 10^{-1.5}, 10^{-2}$
ρ	0.1
σ	1

Table 3: Parameters used for the single-phase flow experiment. The parameters $y, q, \gamma, N, k, \epsilon$ are used in the MLMC-SR algorithm and ρ, σ to define the log-normal field.

ϵ	Mean p	Sample std	Target std ($\epsilon/\sqrt{2}$)
10^{-1}	0.8834	$6.472 \cdot 10^{-2}$	$7.071 \cdot 10^{-2}$
$10^{-1.5}$	0.8890	$1.873 \cdot 10^{-2}$	$2.236 \cdot 10^{-2}$
10^{-2}	0.8933	$5.557 \cdot 10^{-3}$	$7.071 \cdot 10^{-3}$

Table 4: The mean failure probability p and sample standard deviation (std) is computed using 100 MLMC-SR estimators and compared to the target std which is the statistical part of the RMSE error ϵ.

compute p for $\epsilon = 10^{-1}$, $10^{-1.5}$, and 10^{-2}. All parameters used in the simulation are presented in Table 3. To verify the accuracy of the estimator we compute 100 simulations of the MLMC-SR estimator for each RMSE ϵ and present the sample standard deviation (square root of the sample variance) of the MLMC-SR estimators in Table 4. We see that in all the three cases the sample standard deviation is smaller than the statistical contribution $\epsilon/\sqrt{2}$ of the RMSE ϵ. Since the exact flux is unknown, the numerical contribution in the estimator has to be approximated to be less than $\epsilon/\sqrt{2}$ as well, which is done in the termination criterion of the MLMC-SR algorithm so it is not presented here. The mean number of samples computed to the different tolerances on each level of the MLMC-SR algorithm is computed from 100 simulations of the MLMC-SR estimator for $\epsilon = 10^{-2}$ and are shown in Table 5. The table shows that the selective refinement algorithm only refines a fraction of all problems to the highest accuracy level $j = \ell$. Using a

ℓ	0	1	2	3	4
Mean N_ℓ	16526.81	9045.41	4524.83	1471.63	738.63
$j=0$	16526.81	4520.99	2265.23	734.21	366.90
$j=1$		4524.42	1486.62	484.11	244.69
$j=2$			772.98	232.33	116.77
$j=3$				20.98	9.76
$j=4$					0.51

Table 5: The distribution of realizations solved to different tolerance levels j for the case $\epsilon = 10^{-2}$. The table is based on the mean of 100 runs.

MLMC method (without selective refinement) N_ℓ problem would be solved to the highest accuracy level. Using the cost model $\gamma^{-q\ell}$ for $\epsilon = 10^{-2}$ we gain a factor ~ 6 in computational cost for this particular problem using MLMC-SR compared to MLMC. From Theorem 10 the computational cost for MLMC-SR and MLMC increase as $\epsilon^{-2}\log(\epsilon^{-1})^2$ and ϵ^{-3}, respectively.

A Derivation of optimal level sample size

To determine the optimal sample level size N_ℓ in equation (7), we minimize the total cost keeping the variance of the MLMC estimator equal to $\epsilon^2/2$, i.e.,

$$\begin{aligned} &\min \sum_{\ell=0}^{L} N_\ell c_\ell \\ &\text{subject to} \sum_{\ell=0}^{L} N_\ell^{-1}\mathbb{V}[Y_\ell] = \epsilon^2/2, \end{aligned} \tag{24}$$

where $Y_\ell = Q_\ell - Q_{\ell-1}$. We reformulate the problem using a Lagrangian multiplier μ for the constraint. Define the objective function

$$g(N_\ell, \mu) = \sum_{\ell=0}^{L} N_\ell c_\ell + \mu\left(\sum_{\ell=0}^{L} N_\ell^{-1}\mathbb{V}[Y_\ell] - \epsilon^2/2\right). \tag{25}$$

The solution is a stationary point (N_ℓ, μ) such that $\nabla_{N_\ell,\mu} g(N_\ell, \mu) = 0$. Denoting by $\hat{N}_\ell$ and $\hat{\mu}$ the components of the gradient, we obtain

$$\nabla_{N_\ell,\mu} g(N_\ell, \mu) = \left(c_\ell - \mu N_\ell^{-2}\mathbb{V}[Y_\ell]\right)\hat{N}_\ell + \left(\sum_{\ell=0}^{L} N_\ell^{-1}\mathbb{V}[Y_\ell] - \epsilon^2/2\right)\hat{\mu}. \tag{26}$$

Choosing $N_\ell = \sqrt{\mu \mathbb{V}[Y_\ell]/c_\ell}$ makes the $\hat{N}_\ell$ components zero. The $\hat{\mu}$ component is zero when $\sum_{\ell=0}^{L} N_\ell^{-1} \mathbb{V}[Y_\ell] = \epsilon^2/2$. Plugging in N_ℓ yields $2\epsilon^{-2} \sum_{\ell=0}^{L} \sqrt{\mathbb{V}[Y_\ell] c_\ell} = \sqrt{\mu}$ and hence the optimal sample size is

$$N_\ell = 2\epsilon^{-2} \sqrt{\mathbb{V}[Y_\ell]/c_\ell} \sum_{k=0}^{L} \sqrt{\mathbb{V}[Y_k] c_k}. \tag{27}$$

References

[1] S.-K. Au and J. L. Beck. Estimation of small failure probabilities in high dimensions by subset simulation. *Probabilistic Engineering Mechanics*, 16(4):263–277, 2001.

[2] R. Avikainen. On irregular functionals of SDEs and the Euler scheme. *Finance Stoch.*, 13(3):381–401, 2009.

[3] A. Barth, C. Schwab, and N. Zollinger. Multi-level Monte Carlo finite element method for elliptic PDEs with stochastic coefficients. *Numer. Math.*, 119(1):123–161, 2011.

[4] J. Charrier, R. Scheichl, and A. L. Teckentrup. Finite element error analysis of elliptic PDEs with random coefficients and its application to multilevel Monte Carlo methods. *SIAM J. Numer. Anal.*, 51(1):322–352, 2013.

[5] K. A. Cliffe, M. B. Giles, R. Scheichl, and A. L. Teckentrup. Multilevel Monte Carlo methods and applications to elliptic PDEs with random coefficients. *Comput. Vis. Sci.*, 14(1):3–15, 2011.

[6] N. Collier, A.-L. Haji-Ali, F. Nobile, E. von Schwerin, and R. Tempone. A continuation multilevel Monte Carlo algorithm. *BIT*, 55:399–432, 2015.

[7] C. Dietrich and G. Newsam. Fast and exact simulation of stationary gaussian processes through circulant embedding of the covariance matrix. *SIAM J. Sci. Comput.*, 18(4):1088–1107, 1997.

[8] D. Elfverson, D. Estep, F. Hellman, and A Målqvist. Uncertainty quantification for approximate p-quantiles for physical models with stochastic inputs. *SIAM/ASA J. Uncertain. Quantif.*, 2(1):826–850, 2014.

[9] D. Estep, A. Målqvist, and S. Tavener. Nonparametric density estimation for randomly perturbed elliptic problems. I. Computational methods, a posteriori analysis, and adaptive error control. *SIAM J. Sci. Comput.*, 31(4):2935–2959, 2009.

[10] M. B. Giles. Multilevel Monte Carlo path simulation. *Oper. Res.*, 56(3):607–617, 2008.

[11] M. B. Giles and E. Süli. Adjoint methods for PDEs: a posteriori error analysis and postprocessing by duality. *Acta Numer.*, 11:145–236, 2002.

[12] V. Girault and P.-A. Raviart. *Finite element methods for Navier-Stokes equations*, volume 5 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, 1986. Theory and algorithms.

[13] P. Glasserman, P. Heidelberger, P. Shahabuddin, and T. Zajic. Splitting for rare event simulation: analysis of simple cases. In *Proceedings of the 1996 Winter Simulation Conference*, pages 302–308, 1996.

[14] P. Glynn. Importance sampling for monte carlo estimation of quantiles. In *Mathematical Methods in Stochastic Simulation and Experimental Design: Proc. 2nd St. Petersburg Workshop on Simulation (Publishing House of Saint Petersburg University)*, pages 180–185, 1996.

[15] A.-L. Haji-Ali, F. Nobile, E. von Schwerin, and R. Tempone. Optimization of mesh hierarchies in multilevel Monte Carlo samplers. *Stoch. Partial Differ. Equ. Anal. Comput.*, pages 1–37, 2015.

[16] E. Jones, T. Oliphant, P. Peterson, et al. SciPy: Open source scientific tools for Python, 2001. [Online; accessed 2014-08-22].

[17] A. Kebaier. Statistical Romberg extrapolation: a new variance reduction method and applications to options pricing. *Annals of Applied Probability*, 14(4):2681–2705, 2005.

[18] A. Logg, K.-A. Mardal, and G. Wells. *Automated Solution of Differential Equations by the Finite Element Method*, volume 84 of *Lecture Notes in Computational Science and Engineering*. Springer, Berlin Heidelberg, 2012.

[19] A. L. Teckentrup, R. Scheichl, M. B. Giles, and E. Ullmann. Further analysis of multilevel Monte Carlo methods for elliptic PDEs with random coefficients. *Numer. Math.*, 125(3):569–600, 2013.