UPPSALA
UNIVERSITET

IT Licentiate theses
2015-003

Multiscale and Multilevel Methods for Porous Media Flow Problems

FREDRIK HELLMAN

UPPSALA UNIVERSITY
Department of Information Technology

UPPSALA
UNIVERSITET

Multiscale and Multilevel Methods for Porous Media Flow Problems

Fredrik Hellman
fredrik.hellman@it.uu.se

September 2015

Division of Scientific Computing
Department of Information Technology
Uppsala University
Box 337
SE-751 05 Uppsala
Sweden

http://www.it.uu.se/

Dissertation for the degree of Licentiate of Technology in Scientific Computing with specialization in Numerical Analysis

ISSN 1404-5117
Printed by the Department of Information Technology, Uppsala University, Sweden

Abstract

We consider two problems encountered in simulation of fluid flow through porous media. In macroscopic models based on Darcy's law, the permeability field appears as data.

The first problem is that the permeability field generally is not entirely known. We consider forward propagation of uncertainty from the permeability field to a quantity of interest. We focus on computing p-quantiles and failure probabilities of the quantity of interest. We propose and analyze improved standard and multilevel Monte Carlo methods that use computable error bounds for the quantity of interest. We show that substantial reductions in computational costs are possible by the proposed approaches.

The second problem is fine scale variations of the permeability field. The permeability often varies on a scale much smaller than that of the computational domain. For standard discretization methods, these fine scale variations need to be resolved by the mesh for the methods to yield accurate solutions. We analyze and prove convergence of a multiscale method based on the Raviart–Thomas finite element. In this approach, a low-dimensional multiscale space based on a coarse mesh is constructed from a set of independent fine scale patch problems. The low-dimensional space can be used to yield accurate solutions without resolving the fine scale.

Acknowledgments

First, I am most grateful to my main adviser Axel Målqvist for introducing me to the subject, generously sharing his expertise and connections, and providing excellent advice about my work. I would like to thank Auli Niemi and Fritjof Fagerlund for guidance in the field of geohydrology and fruitful discussions. I also thank Daniel Elfverson, Patrick Henning and Don Estep for great discussions and collaborations. I gratefully acknowledge all sources of funding for my work and travels, particularly the Centre for Interdisciplinary Mathematics. Thank you all wonderful colleagues at TDB for providing a top-notch working environment! I want to especially mention my great friends Josefin Ahlkrona, Daniel Elfverson, Patrick Henning, Hanna Holmgren, and Martin Tillenius. Finally, I want to thank my family: Birgitta, Anna, Åsa, Maj-Britt, Håkan, and Ayanori for their unfailing support.

List of Papers

This thesis is based on the following papers.

I 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):828–850, 2014.

II D. Elfverson, F. Hellman, and A. Målqvist. A multilevel Monte Carlo method for computing failure probabilities. *Submitted.*

III F. Fagerlund, F. Hellman, A. Målqvist, and A. Niemi. Improved Monte Carlo methods for computing failure probabilities of porous media flow systems. Technical Report 2015-025, Department of Information Technology, Uppsala University, 2015.

IV F. Hellman, P. Henning, and A. Målqvist. Multiscale mixed finite elements. *To appear in Discrete Contin. Dyn. Syst. Ser. S.*

Reprints were made with permission from the publishers.

Contents

Chapter 1

Introduction

The field of numerical simulation of fluid flow through porous media covers a large range of topics. In this thesis, we focus on two computational challenges for simulation of flows governed by Darcy's law, both stemming from properties of the permeability field. In applications such as underground carbon dioxide sequestration, oil recovery and groundwater flow, the permeability field is determined by the properties of the porous medium.

First, the medium properties are generally not known. Even if direct measurements (e.g. core samples or borehole flowmeter tests) and indirect meaurements (e.g. geological facies identification) are available, they cannot be used to reconstruct the properties of the medium in every point of the domain of interest. Instead, we consider the medium properties uncertain. We are concerned only with the forward propagation problem of uncertainty, where the uncertain data is random and follows a known distribution from which random samples can be generated. The forward propagation problem is important, being a crucial component in the solution of inverse problems. We develop improved Monte Carlo type methods for estimating p-quantiles and failure probabilities where numerical simulations are used to solve the deterministic problems. The methods are applied to and evaluated for porous media flow problems. This topic is covered in Chapter 2 and Papers I–III.

Secondly, if the medium properties are known or drawn from some distribution, they generally vary spatially at distances in the order of meters, while the size of the computational domain is in the order of 10–100 km. Fractures, channel structures, and random spatial variability of the field have characteristic scale (fine scale) several orders

of magnitude smaller than that of the domain (coarse scale). The fine scale generally needs to be resolved by the numerical method for it to yield an accurate numerical solution even for coarse scale features. This can be very computationally demanding and upscaling approaches or multiscale methods are employed to reduce the computational effort. We analyze and evaluate a multiscale method for Poisson's equation on mixed form. This equation occurs frequently as the pressure equation in many discretizations for fluid flows through porous media. This topic is covered in Chapter 3 and Paper IV.

A porous medium consists of a solid material with a network of cavities called pores through which fluid can flow. Figure 1.1 shows a

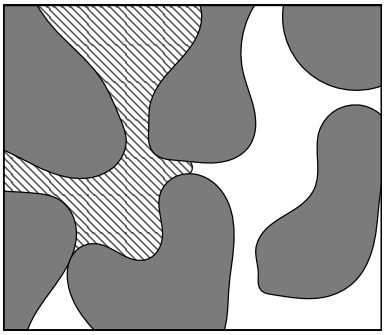


Figure 1.1: A cross section of a porous medium with two phases filling the pore space. Gray, white, and striped fields represent the solid, wetting phase, and non-wetting phase, respectively.

schematic illustration of a cross section of a porous medium depicting pore space, solid and two phases with a sharp interface. While fluid flow thorugh a porous medium can be modeled using the Navier–Stokes equation or network models at microscopic level, this thesis discusses macroscopic models of single or two-phase flows. In a macroscopic model, medium properties and state variables are averaged over a representative elementary volume (REV). For flows in this regime, Darcy's law is the most common form of the momentum equation in the literature: for phases $\alpha = n, w$ (non-wetting and wetting, respectively) in domain Ω,

$$\mathbf{u}_\alpha = -\frac{k_{r\alpha}(s_\alpha)}{\mu_\alpha}\mathbf{K}(\nabla p_\alpha - \rho_\alpha \mathbf{g}), \tag{1.1}$$

where $\mathbf{u}_\alpha$ is Darcy flux (or velocity, bold-face denotes a vector quantity), $k_{r\alpha}$ is relative permeability, μ_α is dynamic viscosity, s_α is saturation, $\mathbf{K}$

is (intrinsic) permeability, p_α is pressure, ρ_α is fluid density, and $\mathbf{g}$ is gravitational acceleration. All quantities above are considered over an REV. The parameters permeability $\mathbf{K}$, saturation s_α and relative permeability $k_{r\alpha}$ are specific to the macroscopic model. The permeability $\mathbf{K}$ is a symmetric positive definite matrix reflecting to what degree the pore structure in the REV allows for fluid flow in different directions. The saturation $0 \leq s_\alpha \leq 1$ of a phase is the fraction of the pore space that is occupied by that phase. This means that at the macroscopic level we do not recognize the sharp interface between phases. Relative permeability $0 \leq k_{r\alpha}(s_\alpha) \leq 1$ is a nonlinear function of the saturation which in product with intrinsic permeability forms the effective permeability. It models the reduction in effective permeability when the pore space is blocked by the presence of the other phase. In addition to Darcy's law, we have mass conservation for each phase,

$$\phi \frac{\partial s_\alpha}{\partial t} + \nabla \cdot \mathbf{u}_\alpha = f_\alpha, \tag{1.2}$$

where ϕ is porosity (pore space fraction of REV) and f_α is a source or sink mass flux. Here we assumed incompressibility of the fluids and solid. In addition to (1.1) and (1.2) we define a relation between the phase pressures by a capillary pressure curve and further let $s_w + s_n = 1$ to close the system. In a single-phase system we omit subscript α, use $s \equiv 1$ to make (1.1) and (1.2) form the pressure equation,

$$\begin{aligned} \mathbf{A}^{-1}\mathbf{u} + \nabla p &= 0, \\ \nabla \cdot \mathbf{u} &= f, \end{aligned} \tag{1.3}$$

where $\mathbf{A} = \frac{k_r}{\mu}\mathbf{K}$. The gravitational force was discarded for simplicity. We refer to [8, 20] for texts on multiphase flow modelling.

The nonlinear PDE in (1.1) and (1.2) is used as model in Paper III and the linear PDE (1.3) is used as model in examples in Papers I, II and as principal object of investigation in Paper IV.

Chapter 2

Quantiles and failure probabilities

We consider the problem of forward propagation of uncertainties. The permeability field $\mathbf{K}$ is modeled as a random field with a known high-dimensional distribution, from which we can generate independent realizations. We define a quantity of interest X (a functional of the solution $\mathbf{u}_\alpha$, s_α and p_α) which also follows a distribution, however unknown, being a response to $\mathbf{K}$ of the multiphase flow model. We focus entirely on the two problems of estimating failure probabilities and quantiles for the distribution of X. More precisely, the first problem is to find the failure probability p for a critical value y of X,

$$p = \Pr(X \leq y), \tag{2.1}$$

or equivalently, $p = F(y)$ where F is the cumulative distribution function (cdf) of X. The second problem is the inverse: given p find the p-quantile, $y = F^{-1}(p)$. This inverse is defined as the smallest y satisfying (2.1). Quantities of interest considered in this thesis are: the flow over a part Γ of the domain boundary for single-phase flow in Paper I and II, and sweep efficiency for two-phase flow in Paper III. The flow over Γ for single-phase flow is defined as

$$X = \int_\Gamma \mathbf{u} \cdot \mathbf{n} \, \mathrm{d}\gamma, \tag{2.2}$$

where $\mathbf{n}$ is the outgoing unit normal vector of Γ. For two-phase flow, the sweep efficiency is the fraction of the domain Ω that is swept by the

non-wetting phase at time T, i.e.

$$X = |\Omega|^{-1} \int_\Omega \chi_{(0,1]}(s_n(T))\,\mathrm{d}x, \tag{2.3}$$

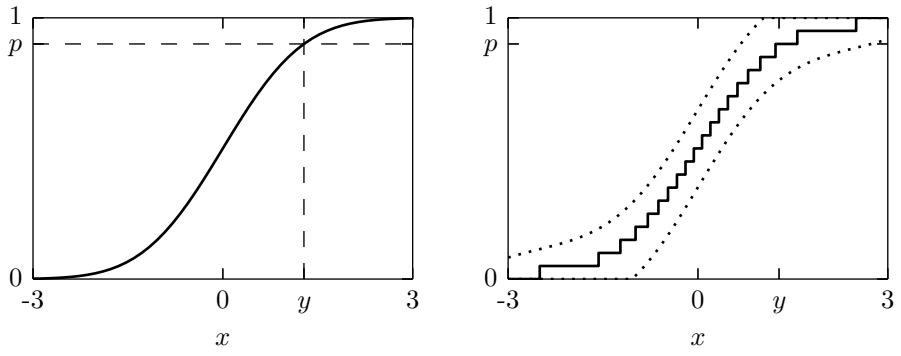
where χ_A is the indicator function for the set A.

The literature on methods for propagating uncertainty through numerical models is vast. One approach is the stochastic Galerkin method [16] which discretizes physical dimensions (space and time) together with stochastic dimensions and is thus an intrusive method. Stochastic collocation [7, 31], stochastic point collocation [9] and pseudospectral projection [28] are non-intrusive methods, based on interpolation, response surfaces and projections in the stochastic space. Further, there are deterministic methods for numerical quadrature in high dimensions: sparse grids [30, 15], lattice rules [29], and quasi Monte Carlo methods [26]. There is also the Monte Carlo method and multilevel Monte Carlo method [17, 19] that perform independently of stochastic dimension, but converge slowly. The specific problems of computing p-quantiles and pointwise evaluation of a cdf have also been studied recently. A multilevel Monte Carlo method for cdf estimation is proposed and analyzed in [5] for a direct approach and also more recently in [18] where smoothness in the cdf is exploited to lower the costs by regularization. In this work, we consider non-intrusive standard and multilevel Monte Carlo methods [17]. In particular, we focus on how to use a posteriori error bounds of the quantity of interest to reduce the computational cost of Monte Carlo methods. The main idea is selective refinement (introduced in Paper I), which makes use of a hierarchy of approximations for X (of increasing accuracy and cost) and exploits error bounds on X to refine only a subset of all generated realizations to the most accurate and costly level.

We turn our attention to estimation of p-quantiles and failure probabilities. Here, the cdf F is central. Figure 2.1a illustrates how the cdf relates with the (failure) probability p and p-quantile y. An empirical cdf F_N (illustrated in Figure 2.1b) can be constructed using a sample X^i of size N. Each jump in F_N corresponds to a realization X^i. Confidence bands F_N^- and F_N^+ for the true F can be formed using e.g. the Dvoretzky–Kiefer–Wolfowitz-inequality or pointwise confidence intervals for $N \cdot F_N(x)$ (which is binomially distributed). Then for some prescribed probability, it holds that

$$F_N^-(x) \leq F(x) \leq F_N^+(x), \tag{2.4}$$

for all (or a few) x simultaneously. This confidence band can be used to give a measure of the statistical error in p or y, depending on which of them is unknown.



(a) Cdf F (solid), probability p and p-quantile y (dashed).

(b) Empirical cdf F_N (solid) and confidence band F_N^- and F_N^+ (dotted).

Figure 2.1: Illustration of a cdf, the empirical cdf and its confidence band.

In addition to the statistical error giving rise to the confidence bands, there is generally a numerical error from computing the sample approximately. That is, neither F nor F_N are computable. We consider situations where the numerical error is controllable, e.g. by solving a PDE numerically to a certain error tolerance in the quantity of interest by adaptive mesh refinement, or by introducing a hierarchy of uniform meshes for which the quantity of interest converges. More precisely, we let X_ℓ be approximations of X with exponentially decreasing error with level ℓ, i.e.

$$|X - X_\ell| \leq C\gamma^\ell, \tag{2.5}$$

for a positive C and $0 < \gamma < 1$. To fix ideas, consider X_ℓ as a functional of a discrete solution to a PDE with mesh size $h = 2^{-\ell}$. Naturally, the cost to compute X_ℓ is greater for large ℓ. This hierarchy is used in the multilevel Monte Carlo method, and also in the proposed selective refinement algorithm. In particular, the selective refinement algorithm (presented in slightly different versions in Papers I–III), exploits that less accurate but cheaper approximations of X can be used when $|y - X|$ is large without sacrificing accuracy of the failure probability or quantile estimate. This allows for a reduction in the computational cost for

estimating p or y.

In Paper I, the proposed algorithm to compute a p-quantile y is to take a sample of random input and initially only compute approximations on level 0. An approximation of y with numerical and statistical bounds is obtained. A subset of all realizations are selected for refinement to the next level. Only the realizations that can potentially affect the p-quantile bound are included in the subset, and a confidence band similar to (2.4) is used to determine this subset. The selective refinement is repeated until the numerical and statistical errors of the p-quantile are balanced. We compare this with full refinement of all realizations. Under mild assumptions on F, the asymptotic ratio of computational cost between selective refinement and full refinement is $\mathcal{O}(N^{-1/2})$ as $N \to \infty$. This holds if the numerical and statistical errors are balanced.

Regarding the problem of computing a failure probability p given y, a standard Monte Carlo (MC) method can be applied directly to the random variable $Q = \chi_{(-\infty,y]}(X)$ (where χ_A is the indicator function for the set A), since

$$p = \mathbb{E}\left[Q\right], \tag{2.6}$$

where $\mathbb{E}\left[\cdot\right]$ denotes expected value. Note that Q attains only the values 1 and 0 with probability p and $1-p$, respectively. This method basically amounts to counting the number of failures obtained in the sample. In practice, we can only compute approximations Q_ℓ (defined in the natural way based on the hierarchy X_ℓ). The MC estimator $\widehat{Q}_{N,L}^{\mathrm{MC}}$ of $\mathbb{E}\left[Q_L\right]$ is the mean

$$\widehat{Q}_{N,L}^{\mathrm{MC}} = \frac{1}{N}\sum_{i=1}^{N} Q_L^i \tag{2.7}$$

of a sample Q_L^i of N independent random variables with the same distribution as Q_L. We call $\mathbb{E}\left[Q - Q_L\right]$ the numerical bias, an error introduced by using an approximation Q_L rather than the true Q. The computational cost for this estimator is N times the cost for computing Q_L. To quantify this, we use a cost model for Q_L: the cost to compute Q_L with numerical bias TOL is $\mathcal{O}(\mathrm{TOL}^{-q})$ for some $q > 0$. For an MC estimator, we have that the standard deviation converges with rate $N^{-1/2}$. Now, choosing N and L to balance the standard deviation with numerical bias, we get that the computational cost is $\mathcal{O}(\mathrm{TOL}^{-2-q})$.

The multilevel Monte Carlo (MLMC) method [17, 19] is a variance reduction technique that uses the full hierarchy Q_ℓ $(0 \leq \ell \leq L)$ to estimate $\mathbb{E}\left[Q_L\right]$. It can be interpreted as a recursive application of control

variates. A new random variable Z is formed using Q and an a priori known random variable R correlated with Q and with known expected value:

$$Z = \mathbb{E}[R] + Q - R. \tag{2.8}$$

The variance $\mathrm{Var}[Z] = \mathrm{Var}[Q - R]$ is smaller than $\mathrm{Var}[Q]$ if Q and R are sufficiently correlated. Now, Z (with lower variance) is sampled instead of Q. A smaller sample is needed to obtain the same variance of the estimator. For this to be beneficial, it is required that i) realizations of R are sufficiently cheap to generate and ii) the expected value of R is known (or can be estimated). In the multilevel Monte Carlo method, less accurate and cheap approximations are used as control variates for more accurate and expensive approximations. We expand $\mathbb{E}[Q_L]$ in a telescoping sum of L levels (and level 0):

$$\mathbb{E}[Q_L] = \mathbb{E}[Q_0] + \sum_{\ell=1}^{L} \mathbb{E}[Q_\ell - Q_{\ell-1}]. \tag{2.9}$$

The expected values on the right hand side in (2.9) are estimated using standard MC estimators, with individual sample sizes N_ℓ for each level ℓ:

$$\widehat{Q}^{\mathrm{ML}}_{\{N_\ell\},L} = \frac{1}{N_0} \sum_{i=1}^{N_0} Q_0^i + \sum_{\ell=1}^{L} \frac{1}{N_\ell} \sum_{i=1}^{N_\ell} Y_\ell^i, \tag{2.10}$$

where $Y_\ell = Q_\ell - Q_{\ell-1}$. Then N_ℓ are chosen such that the expected computational cost is minimized with constraint that the standard deviation of the estimator is equal to TOL. Now if the deepest level L is chosen to balance the numerical bias with the standard deviation (like in the Monte Carlo case), we obtain that the expected cost to realize the estimator is $\mathcal{O}(\mathrm{TOL}^{-1-q})$. This rate holds for approximations Q_ℓ whose cost increases fast enough with ℓ (more precisely, $q > 1$). Comparing with the standard Monte Carlo method, this is an improvement of a factor TOL^{-1}.

In Paper II, we again use the idea of selective refinement but in the context of MLMC. Random samples of sizes N_ℓ are drawn for the MC estimators for all levels $0 \leq \ell \leq L$. Starting with all realizations on level 0 (i.e. by computing X_0), the realizations that could potentially switch from success to failure after refinement are solved on a deeper level, to at most level ℓ. The computable error bounds are used to determine which realizations have this potential. For Lipschitz continuous

F, the expected cost to realize the MLMC estimator with selective refinement is $\mathcal{O}(\mathrm{TOL}^{-q})$. A factor TOL^{-1} is gained by this compared to MLMC without selective refinement. In fact, the cost is asymptotically proportional to that of solving a single realization at level L.

In Paper III, we quantify the performance gains possible by using selective refinement in combination with both the standard and multilevel Monte Carlo method for computing failure probabilities. The methods are applied to a two-phase flow scenario with sweep efficiency (2.3) as quantity of interest. In particular, we construct a probabilistic bound (that holds to a certain probability) rather than using the guaranteed error bound in (2.5), which can be difficult to prove in general. This paper shows that savings of one order of magnitude are possible by using selective refinement in practical applications.

Chapter 3

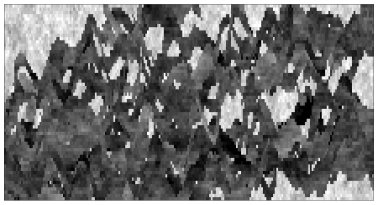
The multiscale problem

We now consider the single-phase flow pressure equation (1.3) on mixed form, i.e. we seek flux $\mathbf{u}$ and pressure p, such that in Ω,

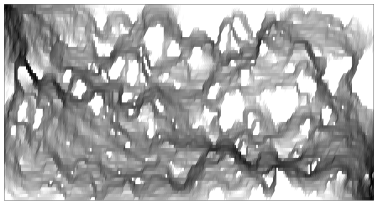
$$\begin{aligned} \mathbf{A}^{-1}\mathbf{u} + \nabla p &= 0, \\ \nabla \cdot \mathbf{u} &= f. \end{aligned} \tag{3.1}$$

These equations can also be written on standard form $-\nabla \cdot \mathbf{A}\nabla p = f$, however, the mixed formulation is very common in applications to flows in porous media, since the flux solution $\mathbf{u}$ is of particular interest. Also, mass conservative flux follows directly from the second equation if the discrete function spaces are chosen properly.

In many porous media flow applications the coefficient $\mathbf{A}$ varies rapidly on a fine scale compared to the scale of the domain. For general boundary conditions and source functions, such coefficients render variations in the flux solution $\mathbf{u}$ at the fine scale. This is illustrated in Figure 3.1 where a rapidly varying permeability field (85th permeability layer of model 2 in the SPE 10 benchmark [13], piecewise constant on a 220×60 grid) and the corresponding flux solution are plotted. One can see that the fine scale variations from the permeability data are present in the solution. Even if we only seek a solution on the coarse scale, it is well-known that fine scale variations in the data need to be resolved by the mesh in standard methods for the coarse features of the solution to be accurate [6]. Another problem is that the computational requirements (memory and time) for a full-scale simulation on the fine scale can exceed the available resources. During the last few decades multiscale methods for elliptic equations have been developed as a remedy for these problems, for example, the variational multiscale method (VMM) [10, 23, 24],



(a) $\log \|\mathbf{A}\|_2$



(b) $\log \|\mathbf{u}\|$

Figure 3.1: (a) Permeability field and (b) flux solution, with source in upper left and sink in lower right corner. The color scales range around six orders of magnitude from white (low) to black (high).

the multiscale finite element method (MsFEM) [14, 21] and subgrid upscaling [4]. These multiscale methods have been further developed to include the mixed formulation and mass conservation [1, 2, 3, 11, 25]. They all have in common that a low-dimensional discretization with high approximation properties is constructed, based on a coarse mesh on a suitable scale. Local problems are solved on the fine scale to incorporate the fine scale features into the low-dimensional representation. In Paper IV we study a multiscale method for the Raviart–Thomas finite element based on VMM and the work [25]. This method is mass conservative on the fine scale up to coarse elements containing non-zero source. Full mass conversation on the fine scale is possible by source correction in those coarse elements.

In the VMM framework (see e.g. [22] for a description of the abstract VMM framework), a fine scale Green's operator is defined by a set of fine scale equations. It is used to incorporate fine scale effects in a set of coarse scale equations. The procedure in Paper IV follows the works [23, 24, 25], where instead of a fine scale Green's operator, we define a corrector operator, which solves a fine scale problem without source function. Here follows an abstract description of the method. We let Y be the function space in which the equation can be solved accurately. Further, let Y_c and Y^f be the coarse and fine spaces, respectively, so that $Y = Y_\mathrm{c} \oplus Y^\mathrm{f}$. (To fix ideas, Y_c is typically the range of a projection to a low-dimensional function space defined on coarse mesh, e.g. piecewise linears, and Y^f is the kernel of this projection). The equation is defined by a symmetric bilinear form B and a linear functional F: find $x \in Y$ such that for all $y \in Y$, $B(x, y) = F(y)$. The VMM idea starts by splitting this equation in two, testing with coarse and fine functions

separately,

$$\begin{aligned} B(x, y_{\mathrm{c}}) &= F(y_{\mathrm{c}}) \qquad \text{for all } y_{\mathrm{c}} \in Y_{\mathrm{c}}, \\ B(x, y^{\mathrm{f}}) &= F(y^{\mathrm{f}}) \qquad \text{for all } y^{\mathrm{f}} \in Y^{\mathrm{f}}. \end{aligned} \tag{3.2}$$

We split the solution in the fine scale equation, i.e. $B(x_{\mathrm{c}}+x^{\mathrm{f}}, y^{\mathrm{f}}) = F(y^{\mathrm{f}})$. If this equation is well-posed, the fine scale component x^{f} is a function of x_{c} (and F). In the VMM framework, this fine scale component is called the fine scale Green's operator and depends on the coarse scale residual, i.e. including F. Here, we depart from VMM and omit F to define the fine scale corrector $Gx_{\mathrm{c}} \in Y^{\mathrm{f}}$ by

$$B(Gx_{\mathrm{c}}, y^{\mathrm{f}}) = B(x_{\mathrm{c}}, y^{\mathrm{f}}). \tag{3.3}$$

This establishes an orthogonality with respect to B (in a generalized sense) between Y^{f} and $Y^{\mathrm{ms}} = (1-G)Y_{\mathrm{c}}$, the multiscale space. Now, splitting x into components of Y^{ms} and Y^{f} in the original equation, and testing in the multiscale space only, gives

$$B((1-G)x_{\mathrm{c}}, (1-G)y_{\mathrm{c}}) = F((1-G)y_{\mathrm{c}}) \qquad \text{for all } y_{\mathrm{c}} \in Y_{\mathrm{c}}, \tag{3.4}$$

which is to solve the equation in the multiscale space. When expanding x_{c} and y_{c} into linear combinations of coarse base functions ϕ, we note that we need to solve the global fine scale corrector problems (3.3) to obtain $G\phi$ for every ϕ. However, for many B and choices of space splits, $G\phi$ decays exponentially with the distance to the support of ϕ. Thus, localization to subdomains around the support of the basis functions (vertex patches) allows for efficient solution of (3.3) with little loss of accuracy.

In the setting of the mixed formulation of the pressure equation, we have a fine triangular or tetrahedral mesh with mesh size h, resolving all the fine scales. We use the lowest order Raviart–Thomas [27] finite element space V_h and the space Q_h of piecewise constants for the flux and pressure, respectively. They form a stable pair for the (weak) mixed formulation of (3.1): find $\mathbf{u}_h \in V_h$ and $p_h \in Q_h$, such that

$$\underbrace{(\mathbf{A}^{-1}\mathbf{u}_h, \mathbf{v}_h) - (\nabla \cdot \mathbf{v}_h, p_h) + (\nabla \cdot \mathbf{u}_h, q_h)}_{B(x,y)} = \underbrace{(f, q_h)}_{F(y)}, \tag{3.5}$$

for all $\mathbf{v}_h \in V_h$ and $q_h \in Q_h$. Here $(\cdot, \cdot)$ denotes the L^2-scalar product. The equation fits into the framework above with B and F defined in (3.5), and $Y = V_h \times Q_h$. A similar coarse mesh with mesh size $H > h$

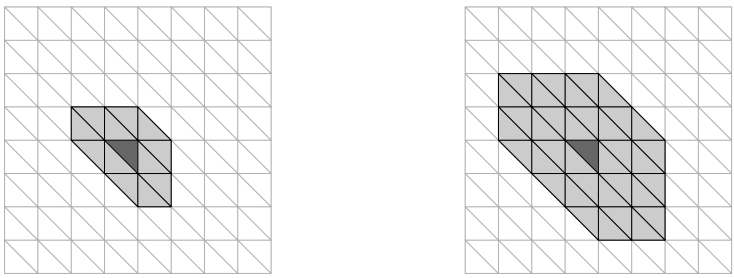
defines $Y_c = V_H \times Q_H$. The fine space Y^f is defined by the kernel of the interpolation from $Y \to Y_c$ (Raviart–Thomas interpolator for V_H and L^2-projection for Q_H). For these particular interpolation operators, we have a commuting property: the divergence of the Raviart–Thomas interpolation is equal to the L^2-projection of the divergence. This property allows for significant simplifications in the corrector equations (3.3). For example, we only need to compute the flux component G_h of the fine scale correction. The multiscale problem (3.4) becomes the following low-dimensional problem: find $\mathbf{u}_H \in V_H$ and $p_H \in Q_H$, such that

$$(\mathbf{A}^{-1}(1 - G_h)\mathbf{u}_H, (1 - G_h)\mathbf{v}_H) - (\nabla \cdot \mathbf{v}_H, p_H) + (\nabla \cdot \mathbf{u}_H, q_H) = (f, q_H) \tag{3.6}$$

for all $\mathbf{v}_H \in V_H$ and $q_H \in Q_H$. Without localization, the following error estimate in energy norm $|||\cdot|||$ holds

$$|||\mathbf{u}_h - (1 - G_h)\mathbf{u}_H||| \lesssim H\|f - P_H f\|_{L^2(\Omega)}, \tag{3.7}$$

where P_H is the L^2-projection onto Q_H. (Here $a \lesssim b$ means $a \leq Cb$ for a positive constant C independent of the fine scale variations and mesh size parameters). Now, if we localize the corrector problems (3.3) to k-coarse-



(a) One-coarse-layer patch, $k = 1$. (b) Two-coarse-layer patch, $k = 2$.

Figure 3.2: Illustration of k-coarse-layer patches around a triangle.

layer patches around the support of the basis functions (see Figure 3.2) and compute the localized flux correction operator G_h^k instead we can save memory and time. Keeping h fixed, choosing k appropriately, we obtain the error estimate

$$\left|\left|\left|\mathbf{u}_h - (1 - G_h^k)\mathbf{u}_H\right|\right|\right| \lesssim H\|f\|_{L^2(\Omega)}, \tag{3.8}$$

when solving (3.6) with the k-coarse-layer patch correctors G_h^k instead of G_h. This relies on the exponential decay of G_h with the distance to the support of the basis functions. The method suffers from an instability for small h: the L^2 stability constant for the Raviart–Thomas interpolation operator increases logarithmically (in 2D and 3D) as h decreases. However, this can be compensated by increasing patch size k. The instability can be removed completely from the method, if the Raviart–Thomas interpolation operator is replaced by a stable interpolation operator that satisfies the commuting property. Such operators have been shown to exist [12], but seem not be easy to compute. In Paper IV, the exponential decay of the fine scale correctors and the a priori error estimates are proved. We investigate the instabilities and convergence numerically.

Chapter 4

Future work

Below is a list of four possible future projects related to the work in this thesis.

- Develop sensitivity analysis for failure probability with respect to a set of control parameters for use in for example an optimization process.
- Investigate alternative approaches to exploit regularity of the cdf in the estimation of p-quantiles and failure probabilities.
- Develop mixed finite element methods for the nonlinear Richard's equation for water movement in unsaturated soils.
- Combine the forward propagation problem with multiscale methods by letting fine scale features (e.g. cap rock cracks or faults) be randomly located.

Chapter 5

Author's contributions

Paper I. The author of this thesis prepared the manuscript and performed the computations in close collaboration with the first author. The analysis was done in collaboration with the first and fourth author. The ideas were developed in collaboration between all authors.

Paper II. The author of this thesis prepared the manuscript, performed the computations and analysis in close collaboration with the first author. The ideas were developed in collaboration between all authors.

Paper III. The author of this thesis had the main responsibility for preparing the manuscript and performed all computations and analysis. The ideas were developed in collaboration between all authors.

Paper IV. The author of this thesis prepared the manuscript and performed all computations. The analysis was done in collaboration with the second author. The ideas were developed in collaboration between all authors.

Bibliography

[1] J. E. Aarnes. On the use of a mixed multiscale finite element method for greater flexibility and increased speed or improved accuracy in reservoir simulation. *Multiscale Model. Simul.*, 2(3):421–439, 2004.

[2] J. E. Aarnes, S. Krogstad, and K.-A. Lie. A hierarchical multiscale method for two-phase flow based upon mixed finite elements and nonuniform coarse grids. *Multiscale Model. Simul.*, 5(2):337–363, 2006.

[3] T. Arbogast and K. J. Boyd. Subgrid upscaling and mixed multiscale finite elements. *SIAM J. Numer. Anal.*, 44(3):1150–1171, 2006.

[4] T. Arbogast, S. E. Minkoff, and P. T. Keenan. An operator-based approach to upscaling the pressure equation. In V. N. Burganos et al., editors, *Computational Methods in Contamination and Remediation of Water Resources*, volume 1 of *Computational Methods in Water Resources XII*, pages 404–412. Computational Mechanics Publications, Southhampton, U.K., 1998.

[5] R. Avikainen. On irregular functionals of SDEs and the Euler scheme. *Finance Stoch.*, 13(3):381–401, 2009.

[6] I. Babuška and J. E. Osborn. Generalized finite element methods: their performance and their relation to mixed methods. *SIAM J. Numer. Anal.*, 20(3):510–536, 1983.

[7] V. Barthelmann, E. Novak, and K. Ritter. High dimensional polynomial interpolation on sparse grids. *Adv. Comput. Math.*, 12(4):273–288, 2000.

[8] J. Bear. *Dynamics of fluids in porous media.* Elsevier, New York, 1972.

[9] M. Berveiller, B. Sudret, and M. Lemaire. Stochastic finite element: a non intrusive approach by regression. *Eur. J. Comput. Mech.*, 15:81–92, 2006.

[10] F. Brezzi, L. P. Franca, T. J. R. Hughes, and A. Russo. $b = \int g$. *Comput. Methods Appl. Mech. Engrg*, 145(3-4):329–339, 1997.

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

[12] S. H. Christiansen and R. Winther. Smoothed projections in finite element exterior calculus. *Math. Comp.*, 77(262):813–829, 2008.

[13] M. A. Christie. Tenth SPE comparative solution project: A comparison of upscaling techniques. *SPE Reservoir Eval. Eng.*, 4:308–317, 2001.

[14] 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.

[15] T. Gerstner and M. Griebel. Numerical integration using sparse grids. *Numer. Algorithms*, 18(3–4):209–232, 1998.

[16] R. G. Ghanem and P. D. Spanos. *Stochastic finite elements: a spectral approach*. Springer-Verlag, New York, 1991.

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

[18] M. B. Giles, T. Nagapetyan, and K. Ritter. Multilevel Monte Carlo approximation of distribution functions and densities. *SIAM/ASA J. Uncertain. Quantif.*, 3(1):267–295, 2015.

[19] S. Heinrich. Multilevel Monte Carlo methods. In *Large-Scale Scientific Computing*, volume 2179 of *Lecture Notes in Computer Science*, pages 58–67. Springer Berlin Heidelberg, 2001.

[20] R. Helmig. *Multiphase flow and transport processes in the subsurface*. Springer-Verlag, Berlin Heidelberg, 1997.

[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(1):169 – 189, 1997.

[22] T. J. R. Hughes and G. Sangalli. Variational multiscale analysis: the fine-scale green's function, projection, optimization, localization, and stabilized methods. *SIAM J. Numer. Anal.*, 45(2):539–557, 2007.

[23] 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.

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

[25] A. Målqvist. Multiscale methods for elliptic problems. *Multiscale Model. Simul.*, 9(3):1064–1086, 2011.

[26] H. Niederreiter. *Random number generation and quasi-Monte Carlo methods*. SIAM, Philadelphia, 1992.

[27] P. A. Raviart and J. M. Thomas. A mixed finite element method for 2-nd order elliptic problems. In I. Galligani and E. Magenes, editors, *Mathematical Aspects of Finite Element Methods*, volume 606 of *Lecture Notes in Mathematics*, pages 292–315. Springer Berlin Heidelberg, 1977.

[28] M. T. Reagan, H. N. Najm, R. G. Ghanem, and O. M. Knio. Uncertainty quantification in reacting-flow simulations through non-intrusive spectral projection. *Combustion and Flame*, pages 545–555, 2003.

[29] I. H. Sloan and S. Joe. *Lattice methods for multiple integration*. Oxford University Press, Oxford, 1994.

[30] S. A. Smolyak. Quadrature and interpolation formulas for tensor products of certain classes of functions. *Dokl. Akad. Nauk*, 4:240–243, 1963.

[31] D. Xiu and J. S. Hesthaven. High-order collocation methods for differential equations with random inputs. *SIAM J. Sci. Comput.*, 27(3):1118–1139, 2005.

Paper I

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, $i = 1, \dots, n$, 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:

$$F^-_{n,\epsilon}(x) = \frac{\#\{i = 1, \dots, n : x_i^\epsilon + \epsilon_i \le x\}}{n}, \quad F^+_{n,\epsilon}(x) = \frac{\#\{i = 1, \dots, n : x_i^\epsilon - \epsilon_i \le x\}}{n}.$$

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* $x_j \in \mathbb{R}$, $j = 1, 2$,

$$\Pr\left(|F(x) - F_{n,\epsilon}(x_j)| \le \mathcal{E}^{\text{stat}}_{n,\epsilon}(x_j) + \mathcal{E}^{\text{num}}_{n,\epsilon}(x_j)\right) > 1 - \beta \tag{3.9}$$

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, $F_n(x) \ge F^-_{n,\epsilon}(x)$. Using a similar argument, we conclude that $F_n(x) \le F^+_{n,\epsilon}(x)$. Therefore, $|F_n(x) - F_{n,\epsilon}(x)| \le \mathcal{E}^{\text{num}}_{n,\epsilon}(x)$. 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 $\epsilon = (\epsilon_i)_{i=1}^n$, $\epsilon_{\max} = \max_{i=1,\dots,n} \epsilon_i$, $\epsilon_{\min} = \min_{i=1,\dots,n} \epsilon_i$. 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,

$$y^+_{n,\epsilon} = \inf\{x \in \mathbb{R} : F^-_{n,\epsilon}(x) - \mathcal{E}^{\text{stat}}_{n,\epsilon}(x) \ge p\}, \quad y^-_{n,\epsilon} = \inf\{x \in \mathbb{R} : F^+_{n,\epsilon}(x) + \mathcal{E}^{\text{stat}}_{n,\epsilon}(x) \ge p\}. \tag{4.1}$$

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 $Y = \{x \in \mathbb{R} : F^-_{n,\epsilon}(x) - \mathcal{E}^{\text{stat}}_{n,\epsilon}(x) \ge p\}$. We start by showing that Y is nonempty and $\inf Y$ exists. The assumption on n implies $\mathcal{E}^{\text{stat}}_{n,\epsilon}(x) \le C_1 n^{-1/2} < 1 - p$ for all x. For a fix n, and for all $x > \max_{i=1,\dots,n}(x_i^\epsilon + \epsilon_i)$, 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 $y \le y^+_{n,\epsilon}$ from

$$y = \inf\{x \in \mathbb{R} : F(x) \ge p\} \le \inf\{x \in \mathbb{R} : F^-_{n,\epsilon}(x) - \mathcal{E}^{\text{stat}}_{n,\epsilon}(x) \ge p\} = y^+_{n,\epsilon}.$$

Similarly, $y \ge y^-_{n,\epsilon}$. 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 $\mathcal{E}^{\text{num}}_{n,\delta}(x) - \mathcal{E}^{\text{stat}}_{n,\delta}(x) \leq 0$ 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*

$$y^-_{n,\delta} \leq y^-_{n,\epsilon} \quad \text{and} \quad y^+_{n,\epsilon} \leq y^+_{n,\delta}.$$

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 $\epsilon = (\epsilon_{\text{init}}, \dots, \epsilon_{\text{init}})$ and $I = \{1, \dots, n\}$
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 $I \leftarrow \{i = 1, \dots, n : y_{n,\epsilon}^- - \epsilon_i < x_i^\epsilon \le y_{n,\epsilon}^+ + \epsilon_i\} \cap I$
    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 $I^{(j)} = \{i = 1, \dots, n : \epsilon_i^{(j)} = 2^{-j}\}$
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

$$C_2 \epsilon_i^{-q} \le W(\epsilon_i) \le \epsilon_i^{-q}, \tag{5.1}$$

where $C_2 \le 1$ 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

$$E[F_n(y) - F_{n,\epsilon}(y)] = p_1 - p_{-1} \leq C_L\epsilon, \quad Var[F_n(y) - F_{n,\epsilon}(y)] = n^{-1}(p_1 + p_{-1}) \leq C_L n^{-1}\epsilon. \tag{6.2}$$

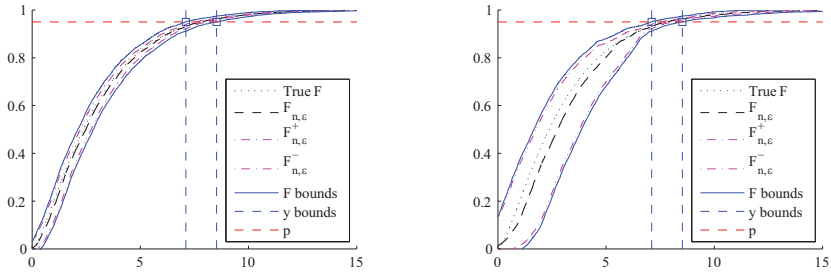
Thus, in the case $p_{-1} = p_1$, the numerical error in the cdf is in the order of $n^{-1/2}\epsilon^{1/2}$, since the expected value is zero. Thus, no refinements are necessary, i.e., we can let $\epsilon \approx 1$ and still balance the statistical and numerical errors in the cdf, thanks to cancellations 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

$$\begin{aligned} u &= 0 \quad \text{on } \Gamma_1, \\ u &= 1 \quad \text{on } \Gamma_2, \\ n \cdot A(\omega) \nabla u &= 0 \quad \text{on } \Gamma_3 \cup \Gamma_4, \end{aligned}$$

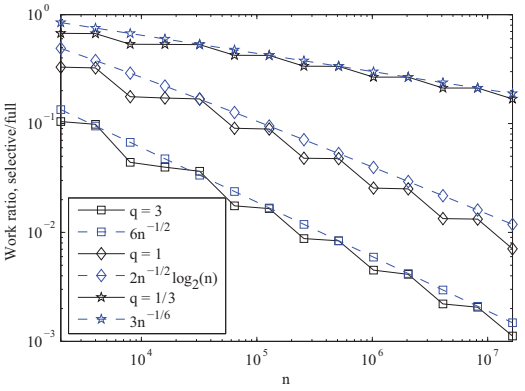


Figure 2. *Work reduction; full vs. selective refinement as a function of sample size.*

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

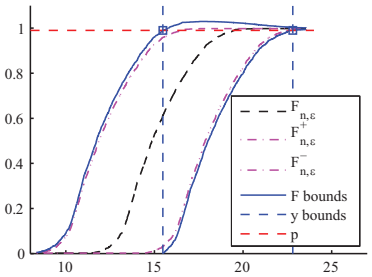
$$w_T = \|\phi^k - \pi^h\phi^k\|_{L^2(T)}, \qquad w_{\partial T} = \|h^{-1/2}(\phi^k - \pi^h\phi^k)\|_{L^2(\partial T)},$$

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.

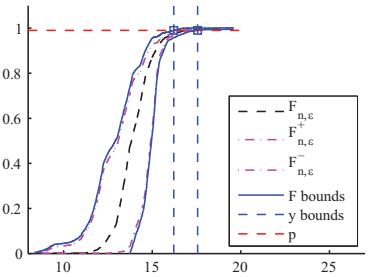
The statistical error, $\mathcal{E}^{\mathrm{stat}}_{n,\epsilon}$, is approximated using the Agresti–Coull confidence interval (see (3.4), (3.5), and (3.7)). We pick $n = 2000$, $p = 0.99$, $\beta = 0.99$, $\epsilon_{\mathrm{init}} = 3$, $A_0 = 1$ and execute Algorithm 3 (selective refinement) using 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 $[16.8, 18.1]$.

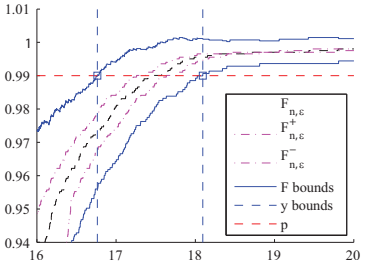
As in the previous section, we compare the ratio of required work between selective and full refinement. In this example, we use the following model of work, $W(h_i) = h_i^{-2}$, where the exponent is -2, since we have a uniform triangulation of a two-dimensional domain and solve the linear equation 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



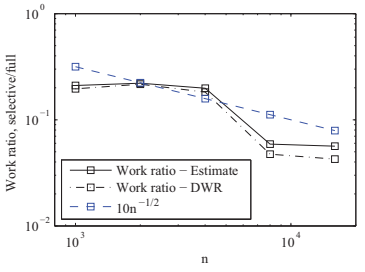
(a) Initial state, before iteration 1.



(b) Final state, after iteration 4.



(c) Zoom-in of Figure 3b of the region close to the p-quantile.

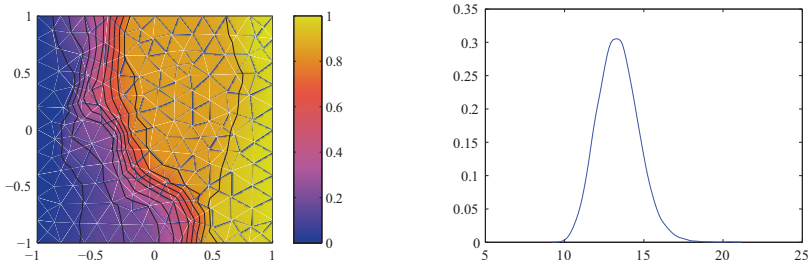


(d) Work reduction; full vs. selective refinement as a function of sample size.

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.

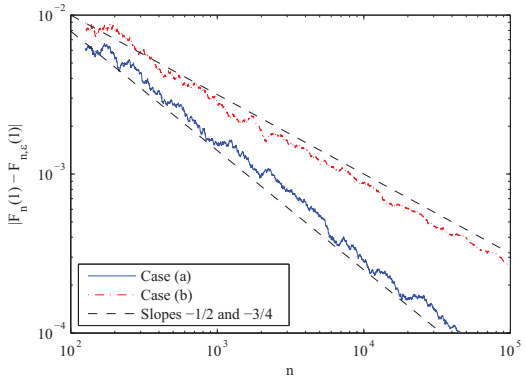


Figure 5. *Convergence of numerical error for* (a) *no systematic error and* (b) *systematic error in numerical approximation.*

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

$$\begin{aligned} y^+_{n,\epsilon} - y^-_{n,\epsilon} &\le \inf\{x \in \mathbb{R} : F(x - 2\epsilon_{\max}) - (K + C_1)n^{-1/2} \ge p\} \\ &\quad - \inf\{x \in \mathbb{R} : F(x + 2\epsilon_{\max}) + (K + C_1)n^{-1/2} \ge p\} \\ &\le F^{-1}(p + (K + C_1)n^{-1/2}) - F^{-1}(p - (K + C_1)n^{-1/2}) + 4\epsilon_{\max} \\ &\le 2\ell^{-1}(K + C_1)n^{-1/2} + 4\epsilon_{\max}, \end{aligned}$$

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 $\{1,\dots,n\}$:

$$\begin{aligned} I^- &= \{i = 1,\dots,n : \quad x^\epsilon_i \le y^-_{n,\epsilon} - \epsilon_i\},\\ I &= \{i = 1,\dots,n : y^-_{n,\epsilon} - \epsilon_i < x^\epsilon_i \le y^+_{n,\epsilon} + \epsilon_i\},\\ I^+ &= \{i = 1,\dots,n : y^+_{n,\epsilon} + \epsilon_i < x^\epsilon_i\}. \end{aligned}$$

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

$$\begin{aligned} I^{(j)} &= \{i = 1, \dots, n : y^-_{n,\epsilon} - \epsilon_i < x^\epsilon_i \le y^+_{n,\epsilon} + \epsilon_i\} \\ &\subseteq \{i = 1, \dots, n : y^-_{n,\epsilon} - \delta_i < x^\delta_i \le y^+_{n,\epsilon} + \delta_i\} \\ &\subseteq \{i = 1, \dots, n : y^{--}_{n,\epsilon} - 2^{-j+1} < x_i \le y^{++}_{n,\epsilon} + 2^{-j+1}\} = \hat{I}^{(j)}. \end{aligned} \tag{8.13}$$

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 II

A multilevel Monte Carlo method for computing failure probabilities

Daniel Elfverson* Fredrik Hellman† Axel Målqvist‡

August 31, 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 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.

*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.

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 $0 < \gamma < 1$.

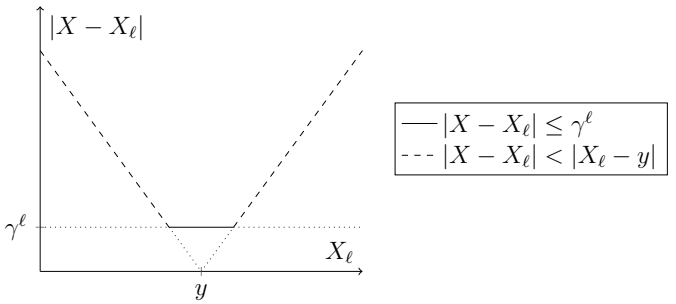


Figure 1: Illustration of condition (4). The numerical error is allowed to be larger than γ^ℓ far away from y.

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 $\widehat{Q}^{ML}_{\{N_\ell\},L}$ at refinement level L is expressed as a telescoping sum of L MC estimator correctors:

$$\widehat{Q}^{ML}_{\{N_\ell\},L} = \sum_{\ell=0}^{L} \frac{1}{N_\ell} \sum_{i=1}^{N_\ell} \big(Q_\ell(\omega_\ell^i) - Q_{\ell-1}(\omega_\ell^i)\big),$$

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

$$\begin{aligned} \mathbb{E}\Big[\widehat{Q}^{ML}_{\{N_\ell\},L}\Big] &= \sum_{\ell=0}^{L} \mathbb{E}[Q_\ell - Q_{\ell-1}] = \mathbb{E}[Q_L] \quad \text{and} \\ \mathbb{V}\Big[\widehat{Q}^{ML}_{\{N_\ell\},L}\Big] &= \sum_{\ell=0}^{L} N_\ell^{-1} \mathbb{V}[Q_\ell - Q_{\ell-1}], \end{aligned} \tag{6}$$

respectively. Using (6) the MSE for the MLMC estimator can be expressed as

$$e\Big[\widehat{Q}^{ML}_{\{N_\ell\},L}\Big]^2 = \sum_{\ell=0}^{L} N_\ell^{-1} \mathbb{V}[Q_\ell - Q_{\ell-1}] + (\mathbb{E}[Q_L - Q])^2,$$

and can be computed at expected cost

$$\mathcal{C}\Big[\widehat{Q}^{ML}_{\{N_\ell\},L}\Big] = \sum_{\ell=0}^{L} N_\ell c_\ell,$$

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

$$\mathcal{C}\Big[\widehat{Q}^{ML}_{\{N_\ell\},L}\Big] = 2\epsilon^{-2} \left(\sum_{\ell=0}^{L} \sqrt{\mathbb{V}[Q_\ell - Q_{\ell-1}]c_\ell} \right)^2. \tag{8}$$

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.

Algorithm 1 Selective refinement algorithm

Input arguments: level ℓ, realization i, critical value y, and tolerance factor γ
Compute $X'_0(\omega_\ell^i)$
Let $j = 0$
while $j < \ell$ and $\gamma^j \geq |X'_j(\omega_\ell^i) - y|$ **do**
    Let $j = j + 1$
    Compute $X'_j(\omega_\ell^i)$
end while
Let $X_\ell(\omega_\ell^i) = X'_j(\omega_\ell^i)$

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.

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 $\mathcal{C}[Q'_L]$.

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

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.

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	$(10^{-3}, 10^{-1})$

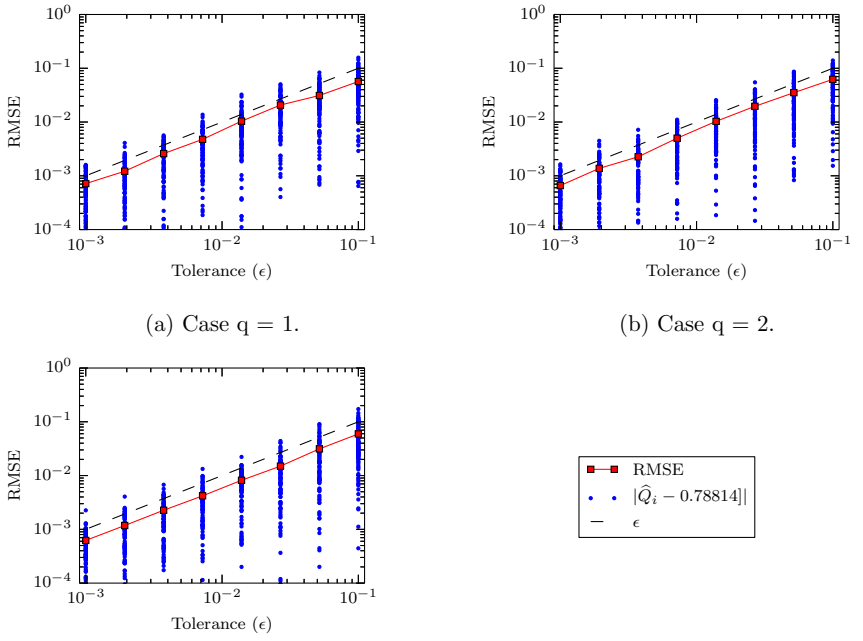
Table 2: Parameters used for the demonstrational experiment.

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

$$e\Big[\widehat{Q}^{ML}_{\{N_\ell\},L}\Big] = \left(\mathbb{E}\left[\left(\widehat{Q}^{ML}_{\{N_\ell\},L} - \mathbb{E}[Q]\right)^2\right]\right)^{1/2} \approx \left(\frac{1}{M}\sum_{i=1}^{M}\left(\widehat{Q}_i - \mathbb{E}[Q]\right)^2\right)^{1/2}.$$

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 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.



(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.

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.

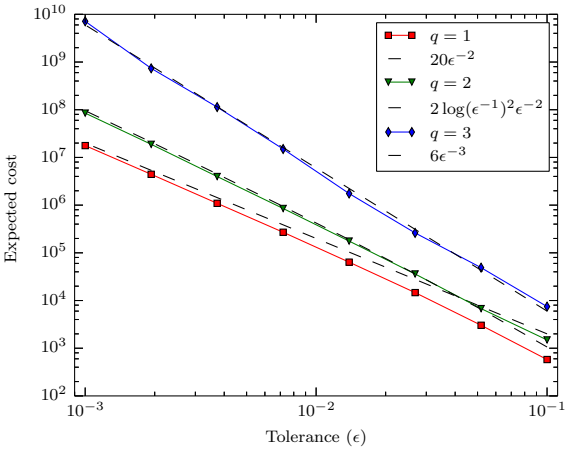


Figure 3: Computed mean total cost (diamond, triangle, square markers and lines) plotted with theoretical reference cost (dashed lines) for the experiment described in Section 8.1. The reference costs for the three values of q are: $20\epsilon^{-2}$ for $q = 1$; $2\log(\epsilon^{-1})^2\epsilon^{-2}$ for $q = 2$; and $6\epsilon^{-3}$ for $q = 3$.

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

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.

[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 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

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.

shows that the selective refinement algorithm only refines a fraction of all problems to the highest accuracy level $j = \ell$. Using a 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.

Paper III

Improved Monte Carlo methods for computing failure probabilities of porous media flow systems

Fritjof Fagerlund[1], Fredrik Hellman[2], Axel Målqvist[3], Auli Niemi[1]

August 31, 2015

Abstract

We study improvements of standard and multilevel Monte Carlo methods for point evaluation of the cumulative distribution function (failure probability) applied to porous media two-phase flow simulations with uncertain permeability. In an injection scenario with sweep efficiency of the injected phase as quantity of interest, we seek the probability that this quantity of interest is smaller than a critical value. In the sampling procedure, we use computable error bounds on the sweep efficiency functional to solve only a subset of all realizations to highest accuracy by means of what we call selective refinement. We quantify the performance gains possible by using selective refinement in combination with both the standard and multilevel Monte Carlo method. We also identify issues in the process of practical implementation of the methods. We conclude that significant savings (one order of magnitude) in computational cost are possible for failure probability estimation in a realistic setting using the selective refinement technique, both in combination with standard and multilevel Monte Carlo.

1 Introduction

Engineering systems are often subject to uncertain conditions that might reduce the performance or function of the system. Monte Carlo methods for quantification of the failure probability of porous media flow systems by numerical simulation is the topic of this report. We focus entirely on an application to two-phase flow where the permeability field is uncertain and modeled as a random field. Given the uncertainty in the permeability, we seek the probability that sweep efficiency is less than a given critical value. In other words, the failure probability is the value of the cumulative distribution function (cdf) for the sweep efficiency functional at the critical value. In this work, we quantify how error bounds on this functional can be used to improve the performance of standard and multilevel Monte Carlo methods. This is a continuation of the works [11, 12] where the selective refinement technique was introduced and asymptotic cost rate results for (multilevel) Monte Carlo methods using error bounds were derived for the point evaluation of a cdf. Other approaches for point evaluation of the cdf in a multilevel Monte Carlo context can

[1]Department of Earth Sciences, Uppsala University, Box 337, SE-751 05 Uppsala, Sweden. Supported by EU FP7 project TRUST.

[2]Department of Information Technology, Uppsala University, Box 337, SE-751 05 Uppsala, Sweden. Supported by the Centre for Interdisciplinary Mathematics, Uppsala University.

[3]Department of Mathematical Sciences, Chalmers University of Technology and University of Gothenburg, SE-412 96 Göteborg, Sweden. Supported by the Swedish Research Council.

be found in [1, 16]. The considered methods (in total four different Monte Carlo methods) are non-intrusive, i.e. we pick realizations from a distribution of input data for which a partial differential equation (PDE) is solved by numerical simulation and the value of the quantity of interest is computed. Each realization is considered either a success or a failure based on the value of this functional in the numerical solution. It can be solved on meshes of varying resolution depending on desired accuracy. The four Monte Carlo setups are briefly discussed below.

For a sample of independent realizations, the sequence of failures and successes can be used to compute an estimate of the failure probability in a Monte Carlo (MC) estimator by computing the mean of the sequence where failure is counted as 1 and success as 0. An improvement of the MC estimator for application on numerical simulations of controllable numerical accuracy is the multilevel Monte Carlo (MLMC) estimator [19] first introduced in the context of differential equations in [15]. It has since then been applied to elliptic PDEs in [2, 6, 21, 23], to density estimation in [1, 16] and has been further analyzed and extended in [7, 17, 18]. It exploits the convergence of numerical solutions with respect to some discretization parameter h (typically mesh size) and uses a series of corrector estimators of increasing cost but decreasing variance to allow for redistribution of the variance reduction effort to cheap low-accuracy problems. Another (independent) improvement is Monte Carlo with selective refinement (MC-SR, [11]) for estimation of p-quantiles or point evaluation of a cdf. Selective refinement uses error bounds of the quantity of interest to determine which realizations need to be solved on a fine mesh, and which realizations can be solved on a coarser mesh to a smaller cost. Selective refinement can be combined with multilevel Monte Carlo (MLMC-SR, [12]). The computational cost of the four setups (MC, MC-SR, MLMC and MLMC-SR) are estimated for the sweep efficiency in a two-phase flow scenario where the failure probability is of magnitude 5–10% and an absolute accuracy of this probability in the order of a few percent is required. This work is to a large extent experimental and also aims at identifying problems and difficulties in the practical implementation of the mentioned improved Monte Carlo techniques. In particular, the two issues of estimating the variance of the correctors for the multilevel Monte Carlo method, and the establishment of an error bound required for selective refinement are addressed in this work.

The report is structured as follows. The problem setting and the continuous two-phase flow model is described in Section 2. In Section 3 we introduce a mesh hierarchy, the space and time discretization used, and the procedure for generating random permeability fields. Section 4 gives an overview of the four Monte Carlo setups and presents the numerical experiments, their results and a discussion. The conclusion is found in Section 5.

2 Continuous model

The problem is to estimate the failure probability $p = F(y) = \Pr(X \leq y)$, where y is a given value, called critical value, and F is the cdf of the sweep efficiency X. The random sweep efficiency is modeled as a functional of the solution to a nonlinear PDE with random inputs modeling two-phase flow. This section describes the continuous model for the two-phase flow system, introducing the PDE in Section 2.1, random input in Section 2.2 and sweep efficiency in Section 2.3.

2.1 Fractional-flow formulation for two-phase flow

We use the fractional flow equations as model for the two-phase flow in porous media and assume isotropic permeability, immiscibility, incompressibility, no capillary forces and that the flow is perpendicular to the gravitaional field. Let the domain be the two-dimensional unit square and denoted by $\mathcal{D} = [0,1]^2$ and its boundary by Γ. We denote an arbitrary phase by α and the two particular phases by $\alpha = w$ and $\alpha = n$ for the wetting and non-wetting phase, respectively. For each phase, we have a mass conservation equation

$$\rho_\alpha \phi \frac{\partial s_\alpha}{\partial t} + \rho_\alpha \nabla \cdot \mathbf{u}_\alpha = \nu_\alpha, \quad \alpha = w, n, \tag{1}$$

in $\mathcal{D}$, where ρ_α is density, ϕ is porosity, s_α is saturation, $\mathbf{u}_\alpha$ is volumetric flux and ν_α is a source term. The pore space is occupied only by the two fluids, i.e. $s_w + s_n = 1$. For convenience, we denote the wetting phase saturation by $s = s_w = 1 - s_n$ and refer to it simply by saturation. The flux is coupled with pressure and saturation via the relative permeabilities in Darcy's law,

$$\mathbf{u}_\alpha = -\frac{k_{r,\alpha}(s) K}{\mu_\alpha} \nabla p, \quad \alpha = w, n. \tag{2}$$

Here, K is the isotropic permeability field, $k_{r,\alpha}$ is the relative permeability, p is pressure (assumed equal for both phases), and μ_α is the dynamic viscosity. For the relative permeability, we use

$$k_{r,w} = (s_e)^3 \quad \text{and} \quad k_{r,n} = \zeta (1 - s_e)^3 \tag{3}$$

for the wetting and non-wetting phase, respectively, where s_e is the effective wetting fluid saturation $s_e = (s_w - s_{r,w})(1 - s_{r,w} - s_{r,n})^{-1}$. Here, $s_{r,\alpha}$ is the residual saturation for the two phases and ζ is a parameter, which we set to 1.

We now present the fractional flow formulation. We denote the total fluid flux by $\mathbf{u} = \mathbf{u}_w + \mathbf{u}_n$ and the phase mobilities for the two phases by

$$\lambda_\alpha(s) = \frac{k_{r,\alpha}(s)}{\mu_\alpha}, \quad \alpha = w, n. \tag{4}$$

The total mobility is defined as $\lambda(s) = \lambda_w(s) + \lambda_n(s)$ and the fractional flow function as $f(s) = \lambda(s)^{-1} \lambda_w(s)$. The wetting phase fluxes can be expressed in terms of total flux using the fractional flow function $\mathbf{u}_w = f(s)\mathbf{u}$. Summing the Darcy equations (2) and mass conservation equations (1) (using that $\frac{\partial}{\partial t}(s_w + s_n) = 0$) yields the pressure equation:

$$\begin{aligned} \mathbf{u} + \lambda(s) K \nabla p &= 0, \\ \nabla \cdot \mathbf{u} &= \rho_w^{-1} \nu_w + \rho_n^{-1} \nu_n. \end{aligned} \tag{5}$$

Finally, we use (1) for $\alpha = w$, and obtain the saturation equation:

$$\phi \frac{\partial s}{\partial t} + \nabla \cdot (f(s)\mathbf{u}) = \rho_w^{-1} \nu_w. \tag{6}$$

The pressure and saturation equations form a non-linear system of equations in the unknowns $\mathbf{u}$, p and s.

The pore space is initiallly filled with the wetting phase, i.e. we have at $t = 0$,

$$s = 1 \quad \text{in } \mathcal{D}.$$

The pressure and flux depend only on s and need not be assigned initial values. Regarding boundary conditions, we let the upper and lower boundary segments $\Gamma_N \subset \Gamma$ of the square be impermeable; the left boundary $\Gamma_L \subset \Gamma$ be assigned high pressure and zero saturation; and the right boundary $\Gamma_R \subset \Gamma$ be assigned low pressure. The pressure gradient makes it necessary only to pose boundary conditions for the saturation on the left boundary, since inward flux will be present only along the left boundary. More precisely, we have for $t \geq 0$,

$$\begin{aligned} \mathbf{u} &= 0 \quad \text{on } \Gamma_N, \\ p = 1, s &= 0 \quad \text{on } \Gamma_L, \\ p &= 0 \quad \text{on } \Gamma_R. \end{aligned} \tag{7}$$

The flow is driven by the boundary conditions exclusively, and we let the source functions be zero, $\nu_\alpha = 0$.

2.2 Lognormal permeability field with exponential covariance

The permeability field K is considered random input data. It is common in the subsurface hydrology literature to model the random permeability fields as lognormal with exponential covariance, possibly at multiple correlation scales (see e.g. [13]). We use this model, but with a single correlation scale of a tenth of the size of the computational domain. More precisely, we let

$$K(x) = \exp(\kappa(x)),$$

over $\mathcal{D}$, where $\kappa(x)$ has zero mean and is normal distributed with exponential covariance, i.e. for all $x_1, x_2 \in \mathcal{D}$ we have that

$$\text{Var}\left[\kappa(x_1)\kappa(x_2)\right] = \sigma^2 \exp\left(\frac{-\|x_1 - x_2\|_2}{\rho}\right) =: C(\|x_1 - x_2\|_2). \tag{8}$$

For this stationary field, the covariance function $C(r)$ in (8) depends only on the distance r between two points. Two realizations of this field are shown in Figure 1.

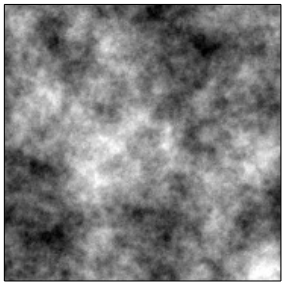
2.3 Sweep efficiency

Our quantity of interest is sweep efficiency. Sweep efficiency is the proportion of the domain covered by the non-wetting fluid at some time after injection, or the proportion of the domain covered at steady-state. Here, since we neglect capillary forces and steady-state is full coverage of the non-wetting fluid, we consider the swept proportion after a fixed time T. The functional is expressed as

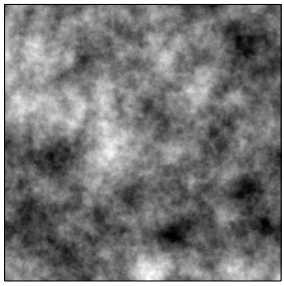
$$X = X(s) = \int_{\mathcal{D}} \chi_{(0,1]}(1 - s(T))\,\mathrm{d}x,$$

where χ_A is the indicator function

$$\chi_A(x) = \begin{cases} 1 & \text{if } x \in A, \\ 0 & \text{otherwise.} \end{cases}$$



(a) Realization 1.



(b) Realization 2.

Figure 1: The 10-logarithm of two realizations of a lognormally distributed random field with exponential covariance. The colormap spans from black to white the range $[-4, 4]$. The parameter values are $\rho = 0.1$ and $\sigma = 3$.

Table 1: Parameter values for the continuous model.

Parameter name (and symbol)	Value
Porosity (ϕ)	1
Relative permeability function parameter (ζ)	1
Residual saturations ($s_{r,w}$ and $s_{r,n}$)	0
Dynamic viscosity (μ_w and μ_n)	1
Source terms (ν_w and ν_n)	0
Standard deviation (σ)	3
Correlation length (ρ)	0.1
Stop time (T)	0.2

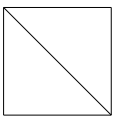
We conclude this section with a table (Table 1) of all parameter values used in this work for the continuous model. Note that the two phases have identical properties.

3 Discretization

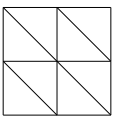
This section is a description of the discretizations used to solve the continuous model numerically. A hierarchy of meshes is introduced in Section 3.1. This is necessary for the multilevel Monte Carlo method and selective refinement, which both require a hierarchy of solutions in a convergent regime to be efficient. A sequential splitting scheme is presented in Section 3.2 describing the simulation procedure, once grid and data are given. Section 3.3 describes how the circulant embedding technique is used to generate permeability data, and how such a field is truncated to the meshes on the lower levels in the mesh hierarchy. An approximation of the sweep efficiency functional is presented in Section 3.4.

3.1 Mesh hierarchy

The domain $\mathcal{D}$ is meshed with a family of uniform and conforming triangulations $\mathcal{T}_h$ depicted in Figure 2. Here h is the vertical and horizontal vertex spacing. The coordinates of the vertices are (ih, jh) for $i, j = 0, \ldots, h^{-1}$ and every square in the grid is split into two triangles by connecting the upper-left and lower-right corners. We introduce a mesh hierarchy consisting of all meshes $\mathcal{T}_h$ with $h = 2^{-(L_0+\ell)}$ for hierarchy levels $\ell = 0, 1, \ldots, L$. Note that from one level ℓ to the next $\ell + 1$, new vertices are added to the grid, but none are removed. The values of L_0 (coarsest mesh size) and L (the number of levels) are to be determined below. In general, the lower bound L_0 is determined by the coarsest mesh that is part of a convergent regime for the quanity of interest, while the number of levels L depends on the limitations of computational resources and desired accuracy.



(a) The mesh with $h = 1$.



(b) The mesh with $h = 1/2$.

Figure 2: Two members of $\mathcal{T}_h$.

We choose $L_0 = 4$, which means the coarsest mesh has a mesh size of $h = 1/16$. This is slightly smaller than the correlation length $\sigma = 0.1$ of K. In, for example [4], it has been indicated that a finite element discretization of the pressure equation does not converge in regimes where the correlation length is not resolved by the mesh for a standard finite element method, why we require $h < \sigma$. The number of levels L is chosen to 4, which means the finest mesh size is $h = 1/256$. This was determined based on our available computational resources.

3.2 Sequential splitting

We use a sequential splitting scheme to split the pressure and saturation equation similarly to the improved Implicit Pressure Explicit Saturation (improved IMPES) scheme presented in [5]. This split renders the pressure equation a linear elliptic equation, and the saturation equation a nonlinear hyperbolic transport equation with fixed total flux field. The pressure equation (5) is solved implicitly for a given time step, keeping saturation fixed from the previous time step. The flux solution from the pressure equation is used in the saturation equation (6), from which the saturation for the next time step is computed using an explicit method.

Let $0 = t^0 < t^1 < \cdots < t^M = T$ be the M outer time steps for which a pressure iteration is performed (additional inner time steps will be needed for solving the saturation equation). We use a constant outer time step length $\tau = t^{m+1} - t^m$. We denote by $\mathbf{u}^m$, p^m and s^m the solutions at outer time step t^m, and state the semi-discrete and mixed form of the pressure equation:

$$\begin{aligned} (\lambda(s^m)K)^{-1}\mathbf{u}^{m+1} + \nabla p^{m+1} &= 0, \\ \nabla \cdot \mathbf{u}^{m+1} &= 0, \end{aligned} \tag{9}$$

with the boundary conditions given in (7).

The saturation equation is discretized using the explicit forward Euler method on a finer grid in time to ensure stability. We subdivide each interval $[t^m, t^{m+1}]$ into K_m inner time steps $t^m = t^{m,0} < t^{m,1} < \cdots < t^{m,K_m} = t^{m+1}$, with an inner time step length $\tau_m = t^{m,k+1} - t^{m,k}$. Note that $K_m = \tau/\tau_m$. The saturation equation is then discretized in time using forward Euler:

$$\phi s^{m,k+1} = \phi s^{m,k} - \tau_m \nabla \cdot \mathbf{u}_w^{m,k}, \quad k = 0, \ldots, K_m - 1. \tag{10}$$

where

$$\mathbf{u}_w^{m,k} = f(s^{m,k})\mathbf{u}^{m+1}. \tag{11}$$

Equations (9) and (10) are solved in sequence, so that $\mathbf{u}^{m+1}$ is available as data for the saturation equation, and $s^{m,K_m} = s^{m+1}$ is available for the pressure equation. The number of inner time steps K_m is chosen adaptively for each outer time step to match the CFL condition for the full discretization (more details can be found below in this section).

For the spatial discretization of the pressure equation, we use the zeroth order Raviart–Thomas finite elements that yield a conservative flux field $\mathbf{u}_h^{m+1}$. For the saturation equation, we use a donor cell upwind finite volume scheme (see e.g. [20]) on a triangular mesh and the saturation is approximated by piecewse constants $s_h^{m,k}$.

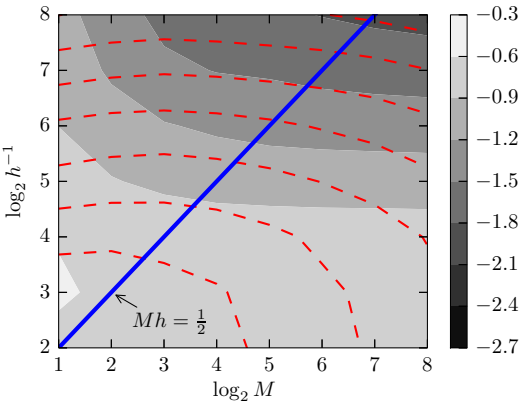


Figure 3: The filled (gray) contour plot shows the 10-logarithm of the estimated relative error. The dashed (red) contour plot shows computational cost iso-lines. The cost increases as h^{-1} and M increases, while the error decreases. The solid (blue) line is selected by hand and is one possible and efficient relation between M and h.

While the description of the discretization above is done for meshes of any size h and any number of time steps M, we use special notation for solutions obtained using the hierarchical meshes introduced in Section 3.1 obeying the relation $Mh = 0.5$. (This relation was determined in an experiment where errors from the spatial and temporal discretizations were balanced to minimize the computational cost, see Figure 3). We define the approximate saturation solution at time T at level ℓ as

$$s_\ell = s_h^M,$$

where $h = 2^{-(L_0+\ell)}$ and $Mh = 0.5$. The remainder of this section is a detailed description of the discretization schemes used.

The set of interior edges in the mesh $\mathcal{T}_h$ are denoted by $\mathcal{F}_h^i$, i.e. $F \in \mathcal{F}_h^i$ is an edge in the triangulation, but $F \cap \partial\mathcal{D} = \emptyset$. The edge normals are denoted by $\mathbf{n}$. The edges on the domain boundaries have outgoing normals. The L^2-scalar product on $\mathcal{D}$ is defined by $(u, v) = \int_{\mathcal{D}} uv \,\mathrm{d}x$ for scalar-valued functions u and v, and $(\mathbf{u}, \mathbf{v}) = \int_{\mathcal{D}} \mathbf{u} \cdot \mathbf{v} \,\mathrm{d}x$ for vector-valued functions $\mathbf{u}$ and $\mathbf{v}$. Let $\mathcal{RT}_0(\mathcal{T}_h)$ denote the zeroth order vector-valued Raviart–Thomas finite element space [22] on $\mathcal{T}_h$ and let $\mathcal{P}_0(\mathcal{T}_h)$ denote the space of piecewise constants on $\mathcal{T}_h$. We then look for a flux solution in $\mathcal{RT}_0(\mathcal{T}_h)$ and for a pressure solution in $\mathcal{P}_0(\mathcal{T}_h)$ (see e.g. [3]). Multiplying (9) with test functions $\mathbf{v}_h$ and q_h from these spaces and integrating over the domain yields the following fully discrete mixed system in $\mathbf{u}_h^{m+1}$ and p_h^{m+1},

$$\begin{aligned} ((\lambda(s^m)K)^{-1}\mathbf{u}_h^{m+1}, \mathbf{v}_h) - (p_h^{m+1}, \nabla \cdot \mathbf{v}_h) = 0 \quad &\text{for all } \mathbf{v}_h \in \mathcal{RT}_0(\mathcal{T}_h), \\ (\nabla \cdot \mathbf{u}_h^{m+1}, q_h) = 0 \quad &\text{for all } q_h \in \mathcal{P}_0(\mathcal{T}_h). \end{aligned} \tag{12}$$

The saturation equation (10) is discretized using a donor cell upwind finite volume method or lowest order discontinuous Galerkin method (see e.g. [9]). We seek a saturation solution $s_h^{m,k+1} \in \mathcal{P}_0(\mathcal{T}_h)$. We multiply (10) by a test function r_h and integrate, and do integration by parts for every triangle on the rightmost term to get the following equation in $s_h^{m,k+1}$, for all $k = 0, \ldots, K_m - 1$,

$$(\phi s_h^{m,k+1}, r_h) = (\phi s_h^{m,k}, r_h) - \tau \sum_{U \in \mathcal{T}_h} (\mathbf{u}_{h,w}^{m,k} \cdot \mathbf{n}, r_h)_{\partial U} \quad \text{for all } r_h \in \mathcal{P}_0(\mathcal{T}_h). \tag{13}$$

Since $s_h^{m,k}$ is only piecewise constant, $\mathbf{u}_{h,w}^{m,k}(s_h^{m,k}) = f(s_h^{m,k})\mathbf{u}_h^{m+1}$ is undefined on the triangle edges. We define the following upwind numerical flux operator A_h^{upw}:

$$\begin{aligned} (A_h^{\mathrm{upw}} s, r) = &\int_{\partial\Omega} (\mathbf{u}^{m+1} \cdot \mathbf{n})^{\ominus} sr \,\mathrm{d}\gamma - \sum_{F \in \mathcal{F}_h^i} \int_F (\mathbf{u}^{m+1} \cdot \mathbf{n}) [\![s]\!] \langle r \rangle \,\mathrm{d}\gamma \\ &+ \sum_{F \in \mathcal{F}_h^i} \int_F \frac{1}{2} |\mathbf{u}^{m+1} \cdot \mathbf{n}| [\![s]\!] [\![r]\!] \,\mathrm{d}\gamma. \end{aligned}$$

The symbol $(\cdot)^{\ominus}$ indicates the negative part, i.e. $(a)^{\ominus} = \frac{1}{2}(|a| - a)$. Further, the symbols $[\![\cdot]\!]$ and $\langle\cdot\rangle$ are defined on interior edges and denote the jump and average operators respectively. For every edge F, let the normal $\mathbf{n}$ be directed towards triangle U_2 from triangle U_1. Then we define, for any $r \in \mathcal{P}_0(\mathcal{T}_h)$,

$$[\![r]\!] = r_{|U_1} - r_{|U_2} \quad \text{and} \quad \langle r \rangle = \frac{r_{|U_1} + r_{|U_2}}{2}.$$

The final form of the discrete saturation equation is

$$(\phi s_h^{m,k+1}, r_h) = (\phi s_h^{m,k}, r_h) - \tau(A_h^{\mathrm{upw}} f(s_h^{m,k}), r_h)$$

for all $r_h \in \mathcal{P}_0(\mathcal{T}_h)$ and $k = 0, \ldots, K_m - 1$.

The number of inner time steps K_m is determined by ensuring that the following CFL condition holds, for all $U \in \mathcal{T}_h$,

$$\frac{\tau_m}{|U|} \int_{\partial U} (\mathbf{u}^{m+1} \cdot \mathbf{n}_U)^{\ominus} L_f \, \mathrm{d}\gamma \leq 1. \tag{14}$$

Here $|U|$ is the area of triangle U, $\mathbf{n}_U$ is the outgoing edge normal of ∂U, and $L_f = 3$ is the Lipschitz constant for $f(s)$ for $0 \leq s \leq 1$. Since $K_m = \tau/\tau_m$, a minimum value of K_m can be determined from equation (14). The value of K_m is determined after the pressure equation has been solved. This ensures that the value of $s_h^{m,k+1}$ in triangle U is a convex combination of the values of $s_h^{m,k}$ in the adjacent triangles, U itself and the boundary conditions. A maximum principle and hence stability immediately follows.

3.3 Circulant embedding and spectral truncation

The multilevel Monte Carlo method requires gradual refinement of the permeability field based on the gradual refinement of the mesh in the mesh hierarchy. In this section, we briefly describe how the permeability data is generated and how it is truncated for different levels ℓ in the mesh hierarchy to avoid aliasing effects.

The random field K is realized using the circulant embedding technique [10]. This is an efficient way to realize a stationary random field on a cartesian grid with equispaced points. The computational complexity is $\mathcal{O}(n^2 \log n)$ if n^2 is the number of vertices in the grid. A detailed description of circulant embedding can be found in Appendix A. Here we proceed with a brief description of the truncation of the fields to coarser meshes.

One way to do the truncation is to realize the field on the finest mesh (level L) and then do linear pointwise interpolation for all coarser meshes (levels $\ell < L$). This approach has the disadvantage that aliasing effects become apparent for the coarsest meshes (see e.g. [14]), i.e. non-negligible variance of the omitted high frequencies are folded onto the low frequencies and the low frequency modes get too large variance.

Instead, we perform the truncation in the spectral domain. We use that the circulant embedding permits a Fourier diagonalization of an extended covariance matrix. The field is realized by generating a random eigenvalue-weighted linear combination of the Fourier modes. If to generate a realization on a mesh with n^2 grid points, we include only the n^2 lowest (ℓ^∞ mixed) Fourier modes in the linear combination. This way, the aliasing effect is avoided. Also, this allows for gradual refinement of a mesh by gradually increasing the number of included Fourier modes. We refer to Appendix A and [10] for details.

3.4 Approximate sweep efficiency

The functional X is not suitable for direct application to the discrete approximation s_h^M of $s(T)$ for two reasons: First, $s(T)$ is equal to 1 in large parts of the domain, and so should a good approximation s_h^M. However, the functional is very sensitive for perturbations of saturation values close to 1 due to the step at that point. The approximation s_h^M might deviate from 1 due to rounding or discretization errors and cause large errors in the functional value. Secondly, even if the plume front is sharp in the continuous solution $s(T)$, it is smoothed out in s_h^M. A smoothed out front can cause the sweep efficiency to be

Table 2: Parameter values for the discretization.

Parameter name (and symbol)	Value
Coarsest mesh (L_0)	4
Deepest level (L)	4
Lipschitz constant for f (L_f)	3
Sweep efficiency threshold (c)	0.5

overestimated. Instead, motivated by the appearance of a Buckley–Leverett plume front (see Figure 4), we define the following functional with a threshold $0 < c \leq 1$,

$$X_h^M = \int_{\mathcal{D}} \chi_{(c,1]}(1 - s_h^M)\,\mathrm{d}x.$$

Further, we define

$$X_\ell = X_h^M,$$

when $h = 2^{-(L_0+\ell)}$ and $Mh = 0.5$. Figure 4 shows the wetting saturation in a continuous solution and an upwind discretized solution to the one dimensional Buckley–Leverett equation in a drainage scenario with relative permeability functions given in (3). The figure suggests that a value of $c = 1$ would overestimate the sweep efficiency. Guided by the figure, we choose $c = 0.5$.

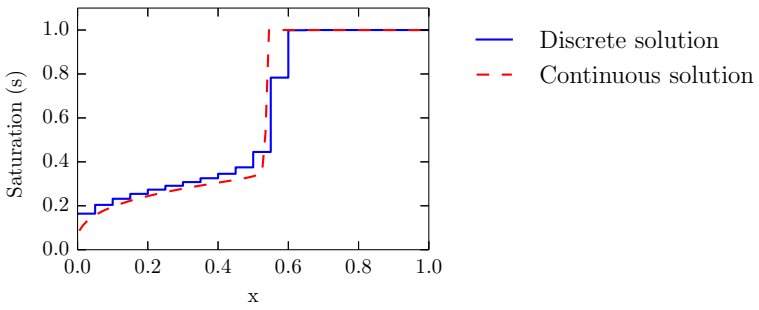


Figure 4: Wetting saturation in continuous and discrete (upwind) solution to Buckley–Leverett solution under drainage.

We conclude this section with a table (Table 2) of all parameter values used for the discretization.

4 Monte Carlo methods for estimating failure probability

We are interested in computing the probability p for the sweep efficiency X to be less than a critical value y, i.e. $p = \Pr(X \leq y)$. We define the failure probability functional

$$Q = \chi_{(-\infty,y]}(X)$$

which is equal to 1 if $X \leq y$, and equal to 0 otherwise. The basis for the Monte Carlo methods in this work is that the failure probability p can be written $p = \mathbb{E}[Q]$, where $\mathbb{E}[\cdot]$ denotes expected value. Similarly to the numerical approximations X_ℓ at level ℓ of X using the numerical scheme presented in Section 3 we define $Q_\ell = \chi_{(-\infty,y]}(X_\ell)$ as the approximate failure probability functional at level ℓ. All estimators $\widehat{Q}$ will be based on discretizations on levels no deeper than L. We use the mean squared error (MSE) (and root mean squared error, RMSE) as a measure of error:

$$e\left[\widehat{Q}\right]^2 = \mathbb{E}\left[\left(\widehat{Q} - \mathbb{E}[Q]\right)^2\right] = \mathrm{Var}\left[\widehat{Q}\right] + \left(\mathbb{E}\left[\widehat{Q} - Q\right]\right)^2.$$

The variance $\mathrm{Var}\left[\widehat{Q}\right]$ of the estimator will be referred to as statistical error and the bias $\mathbb{E}\left[\widehat{Q} - Q\right]$ as numerical bias. Note that this latter bias includes error from truncation of the permeability field as well as discretization of space and time.

We review four different Monte Carlo simulation setups. We use either standard Monte Carlo or multilevel Monte Carlo. For each of them, we either use or do not use selective refinement, rendering four possibilites. The asymptotic computational cost rates for (multilevel) Monte Carlo methods in terms of total RMSE tolerance $e\left[\widehat{Q}\right] \leq \mathrm{TOL}$ for the failure probability functional with and without selective refinement have been studied in [11, 12], based on the analysis in [15]. Convergence rates for the strong error and expected cost to compute a realization are assumed to satisfy the following assumption.

Assumption 1. *Assume that for some $0 < \gamma < 1$ and $C > 0$ independent of ℓ, it holds*

A1 $\mathbb{E}[|Q_\ell - Q|] \leq C\gamma^\ell$,

A2 $\mathcal{C}[Q_\ell] = \gamma^{-q\ell}$,

where $\mathcal{C}[Q_\ell]$ denotes the expected cost to compute a realization of Q_ℓ. Assume additionally that the cdf F is Lipschitz continuous.

A summary of the asymptotic computational cost rates as $\mathrm{TOL} \to 0$ for each setup is given in Table 3, in a scenario where the two sources of error (statistical error and numerical bias) are balanced, i.e. $\mathrm{Var}\left[\widehat{Q}\right] = \left(\mathbb{E}\left[\widehat{Q} - Q\right]\right)^2 = \frac{1}{2}\mathrm{TOL}^2$ (so that $e\left[\widehat{Q}\right] = \mathrm{TOL}$), and the cost for computing a better approximation Q_ℓ increases more than twice as fast as the strong error $\mathbb{E}[|Q_\ell - Q|]$ decreases, i.e. $q > 2$. (It is shown below that this is indeed the case for our application). It is clear from the table that the combination of multilevel Monte Carlo and selective refinement has the lowest asymptotic cost. In fact, the asymptotic cost in this case is proportional to the cost for a single simulation at the deepest level.

In this section, we quantify the time savings possible in a regime where $\mathrm{TOL} \approx 0.01$ and $p \approx 0.07$ for the four setups. The critical value is chosen as $y = 0.08$, i.e. the failure event reads: "the sweep efficiency is less than 8%".

Table 3: Asymptotic cost rates for the failure probability estimators as TOL $\to 0$ for different Monte Carlo method setups when $q > 2$.

	No selective refinement	Selective refinement
Monte Carlo	TOL^{-2-q}	TOL^{-1-q}
Multilevel Monte Carlo	TOL^{-1-q}	TOL^{-q}

4.1 Standard Monte Carlo method

The standard Monte Carlo (MC) estimator $\widehat{Q}^{\text{MC}}_{N,L}$ estimates $\mathbb{E}[Q_L]$ by computing the mean of an i.i.d. N-sample of Q_L,

$$\widehat{Q}^{\text{MC}}_{N,L} = \frac{1}{N}\sum_{i=0}^{N} Q_L^i. \tag{15}$$

Each Q_L^i is computed using the procedure described in Section 3. An MC estimator is unbiased $\mathbb{E}\left[\widehat{Q}^{\text{MC}}_{N,L}\right] = \mathbb{E}[Q_L]$, so the statistical error and numerical bias are

$$\text{Var}\left[\widehat{Q}^{\text{MC}}_{N,L}\right] = \frac{\text{Var}[Q_L]}{N} \quad \text{and} \quad \mathbb{E}\left[\widehat{Q}^{\text{MC}}_{N,L} - Q\right] = \mathbb{E}[Q_L - Q],$$

respectively. Note that we already fixed the level L, while N is to be chosen. The asymptotic cost rate in terms of TOL for balanced statistical and numerical error is TOL^{-2-q} for the MC method.

4.2 Multilevel Monte Carlo method

The multilevel Monte Carlo [15] estimator $\widehat{Q}^{\text{ML}}_{\{N_\ell\},L}$ estimates $\mathbb{E}[Q_L]$ by expanding it in a telescoping sum,

$$\mathbb{E}[Q_L] = \mathbb{E}[Q_0] + \sum_{\ell=1}^{L} \mathbb{E}[Q_\ell - Q_{\ell-1}],$$

and using standard Monte Carlo estimators for the expected values. We introduce correctors $Y_\ell = Q_\ell - Q_{\ell-1}$, and the MLMC estimator reads

$$\widehat{Q}^{\text{ML}}_{\{N_\ell\},L} = \frac{1}{N_0}\sum_{i=1}^{N_0} Q_0^i + \sum_{\ell=1}^{L}\frac{1}{N_\ell}\sum_{i=1}^{N_\ell} Y_\ell^i, \tag{16}$$

where N_ℓ are the sample sizes for the MC estimators and Q_0^i and Y_ℓ^i are realizations of the lowest level and correctors, respectively. A corrector realization Y_ℓ^i is computed by generating one realization i of the random input data and compute both Q_ℓ^i and $Q_{\ell-1}^i$ using that one realization. Due to the telescoping sum, the MLMC estimator is again an unbiased estimator of Q_L. The variance, however, depends on all N_ℓ. The statistical error and numerical bias are,

$$\text{Var}\left[\widehat{Q}^{\text{ML}}_{\{N_\ell\},L}\right] = \frac{\text{Var}[Q_0]}{N_0} + \sum_{\ell=1}^{L}\frac{\text{Var}[Y_\ell]}{N_\ell} \quad \text{and} \quad \mathbb{E}\left[\widehat{Q}^{\text{ML}}_{\{N_\ell\},L} - Q\right] = \mathbb{E}[Q_L - Q],$$

respectively. The number of samples N_ℓ on each level is determined by minimizing the expected cost $\mathcal{C}\left[\widehat{Q}^{\mathrm{ML}}_{\{N_\ell\},L}\right]$ for the estimator constraining the variance, $\mathrm{Var}\left[\widehat{Q}^{\mathrm{ML}}_{\{N_\ell\},L}\right] = \frac{1}{2}\mathrm{TOL}^2$, yielding

$$N_0 = 2C_L\mathrm{TOL}^{-2}\sqrt{\mathrm{Var}\left[Q_0\right]/\mathcal{C}\left[Q_0\right]} \quad \text{and} \quad N_\ell = 2C_L\mathrm{TOL}^{-2}\sqrt{\mathrm{Var}\left[Y_\ell\right]/\mathcal{C}\left[Y_\ell\right]}, \tag{17}$$

where $C_L = \sqrt{\mathrm{Var}\left[Q_0\right]\mathcal{C}\left[Q_0\right]} + \sum_{\ell=1}^{L}\sqrt{\mathrm{Var}\left[Y_\ell\right]\mathcal{C}\left[Y_\ell\right]}$. See [15] for details. The variances of Q_0 and Y_ℓ and the expected cost need to be estimated to determine N_ℓ. The asymptotic cost rate in terms of TOL for balanced statistical and numerical error is TOL^{-1-q} (if $q > 1$) for the MC method for the failure probability functional [12].

An additional comment on MLMC for failure probability is that there are difficulties estimating the variance $\mathrm{Var}\left[Y_\ell\right]$ for deep levels ℓ, due to the discrete distribution of Y_ℓ. The sample size N_L for the deepest level L in the MLMC estimator is typically too small for estimating $\mathrm{Var}\left[Y_L\right]$ reliably using sample variance. (See [7, 12] for elaborate discussions about this). The approach used here is to estimate the variance for the lower levels and extrapolate it to the deeper levels.

4.3 Selective refinement

Selective refinement uses a posteriori error bounds to reduce the expected cost to compute a realization of Q_ℓ. Suppose we are provided with a level j-specific error bound E_j such that

$$|X_\ell - X_j| \leq E_j, \tag{18}$$

for all $\ell > j$. Then if $|y - X_j| > E_j$, the approximate failure probability Q_j is equal to Q_ℓ,

$$Q_j = \chi_{(-\infty,y]}(X_j) = \chi_{(-\infty,y]}(X_\ell) = Q_\ell,$$

since $|y - X_j| > |X_\ell - X_j|$. The selective refinement idea is to approximate Q_ℓ by $Q_\ell^S := Q_j$ for the smallest $j \geq 0$ such that $|y - X_j| > E_j$ or $j = \ell$. An algorithm for computing Q_ℓ^S is presented in Algorithm 1. Using this algorithm (under Assumption 1) makes the

Algorithm 1 Selective refinement algorithm

```
Input arguments: level ℓ, realization i, critical value y
Compute X_0^i and E_0^i
Let j = 0
while j < ℓ and |y − X_j^i| ≤ E_j^i do
    Let j = j + 1
    Compute X_j^i and E_j^i
end while
Let Q_ℓ^{S,i} = χ_{(−∞,y]}(X_j^i).
```

asymptotic cost rate $\mathcal{C}\left[Q_\ell^S\right] \leq \tilde{C}\gamma^{-(q-1)\ell}$ for some constant $\tilde{C}$ independent of ℓ (depending however, on the distribution F). Compared with the rate for $\mathcal{C}\left[Q_\ell\right]$ in Assumption 1, the cost rate is decreased by 1 (see [12]). Thus, using selective refinement in combination with MC and MLMC, yields the rates TOL^{-1-q} and TOL^{-q}, respectively. See Table 3 for a summary of the asymptotic cost rates.

It is often difficult to provide a guaranteed error bound E_j. However, as will be seen Section 4.3.1, in our case it is possible to give a probabilistic bound that holds with probability at least $1-\alpha$, i.e. for any $\ell > j$

$$\Pr(|X_\ell - X_j| \leq E_j) \geq 1 - \alpha. \tag{19}$$

Denote the event that the bound is broken for some $j < \ell$ by $A_\ell = \bigcup_{0\leq j<\ell}\{|X_\ell - X_j| > E_j\}$, so that $\Pr(A_\ell) \leq \sum_{j=0}^{\ell-1} \Pr(|X_\ell - X_j| > E_j) \leq \ell\alpha$. Using a probabilistic bound (19) in place of (18) to compute Q_ℓ^S introduces a bias,

$$\mathbb{E}\left[|Q_\ell^S - Q_\ell|\right] = \mathbb{E}\left[|Q_\ell^S - Q_\ell| \mid A_\ell\right] \Pr(A_\ell) + \mathbb{E}\left[|Q_\ell^S - Q_\ell| \mid \overline{A_\ell}\right] \Pr(\overline{A_\ell}) \leq \ell\alpha,$$

where we denote the complement of A_ℓ by $\overline{A_\ell}$. This bias on the last level L should be of comparable size to the numerical bias and stochastic error, i.e. $L\alpha \leq \text{TOL}$.

The Monte Carlo and multilevel Monte Carlo estimator using selective refinement are denoted by

$$\widehat{Q}_{N,L}^{\text{MC,S}} = \frac{1}{N_0}\sum_{i=0}^{N} Q_L^{S,i} \quad \text{and} \quad \widehat{Q}_{\{N_\ell\},L}^{\text{ML,S}} = \frac{1}{N_0}\sum_{i=1}^{N_0} Q_0^{S,i} + \sum_{\ell=1}^{L}\frac{1}{N_\ell}\sum_{i=1}^{N_\ell} Y_\ell^{S,i},$$

respectively, where $Y_\ell^S = Q_\ell^S - Q_{\ell-1}^S$.

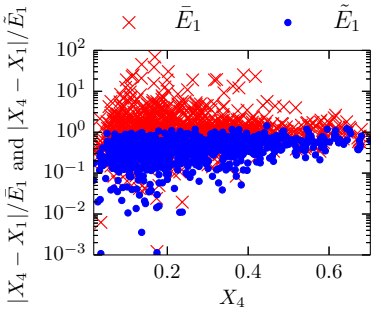
4.3.1 Error bound for sweep efficiency

This section motivates a choice of E_j that satisfies the bound in (19) with probability $1-\alpha$. Starting with the absolute difference in sweep efficiency between two consecutive levels, we define $\bar{E}_j$ and its upper bound $\tilde{E}_j$:

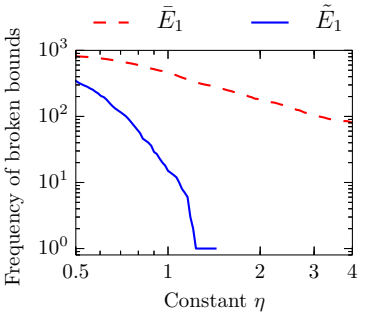
$$\begin{aligned} |X_j - X_{j-1}| &= \left|\int_{\mathcal{D}} \chi_{(c,1]}(s_j) - \chi_{(c,1]}(s_{j-1})\,\mathrm{d}x\right| =: \bar{E}_j \\ &\leq \int_{\mathcal{D}} \left|\chi_{(c,1]}(s_j) - \chi_{(c,1]}(s_{j-1})\right|\,\mathrm{d}x =: \tilde{E}_j. \end{aligned}$$

We now consider $\eta\bar{E}_j$ and $\eta\tilde{E}_j$ as two candidates for a probabilistic bound E_j, where η is a parameter to be determined.

For a sample of size 1000, we computed $X_1, \ldots, X_4$, $\tilde{E}_1$, and $\bar{E}_1$. The ratio between the approximate error and the error bounds $|X_4 - X_1|/\tilde{E}_1$ and $|X_4 - X_1|/\bar{E}_1$ were computed and plotted in Figure 5a as functions of X_4. We also counted number of samples for which the bounds $|X_4 - X_1| \leq \eta\tilde{E}_1$ and $|X_4 - X_1| \leq \eta\bar{E}_1$ are broken for values of η in the range $0.5 \leq \eta \leq 4$. The frequency as function of η is shown in Figure 5b. (The corresponding experiments for $|X_2 - X_1| \leq \eta\tilde{E}_1$ and $|X_3 - X_1| \leq \eta\tilde{E}_1$ were performed. The figures only present the results for $\ell = L = 4$, which was the least optimistic case). The means of $\bar{E}_1$ and $\tilde{E}_1$ were estimated to 0.028 and 0.056, respectively. It is thus clear from the figures and those mean values that $\tilde{E}_1$ is more suitable as an error bound, since a much larger constant η must be used to get the same error frequency with $\bar{E}_1$ than with $\tilde{E}_1$. This would be worthwhile if $\bar{E}_1$ was in average much smaller than $\tilde{E}_1$, however, the ratio between the mean estimates determines this is not the case.



(a) Scatter plot of ratio between approximate error and error bound.



(b) Frequency of $|X_4 - X_1| \not\leq \eta\bar{E}_1$ and $\not\leq \eta\tilde{E}_1$ among the 1000 realizations for $0.5 \leq \eta \leq 4$.

Figure 5: Results from experiment to determine error bound and η.

Guided by Figure 5b, we see that for $\eta = 2$ the error bound was not broken for a single realization. The behaviour of the tail of the distribution of $|X_4 - X_1|/\tilde{E}_1$ indicates that the probability for breaking the error bound $2\tilde{E}_1$ is less than 10^{-3}. Based on this experiment, we choose

$$E_j = \eta\tilde{E}_j$$

with $\eta = 2$ as our error bound with the estimate $\alpha \approx 10^{-3}$. Note that the computational cost for the error bound is proportional to that of the quantity of interest itself. We extrapolate the results to hold for all $j = 1, 2$ and 3.

4.4 Evaluation

We are now ready to evaluate the four combinations Monte Carlo (MC), Monte Carlo with selective refinement (MC-SR), multilevel Monte Carlo (MLMC) and multilevel Monte Carlo with selective refinement (MLMC-SR) with respect to computational cost with the aim to obtain estimators of $\mathbb{E}[Q_L]$ with variance $\approx 0.5 \cdot 10^{-4}$.

We summarize the two-phase flow setting. Q_L is the failure probability functional of X_L, the sweep efficiency functional on discretization level L, i.e. $\mathbb{E}[Q_L] = \Pr(X_L \leq y)$ with $y = 0.08$ in our case. All four setups rely on generating realizations of sweep efficiency X_ℓ for different levels ℓ. This procedure starts by generating the lognormal permeability field on a mesh ℓ, where we use the procedure described in Section 3.3, then solving the PDE numerically using the scheme in Section 3.2. The approximation X_ℓ of the sweep efficiency is computed as described in Section 3.4. For the selective refinement procedure, an error bound for X_ℓ is computed using $\eta\tilde{E}_\ell$ in Section 4.3.1.

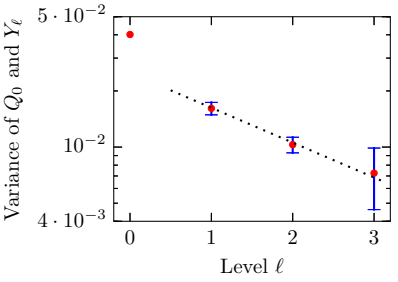
The easiest method to apply is the MC method, which requires knowledge only of $\mathrm{Var}[Q_L]$ (which still is unknown) to determine N. Selective refinement requires an error bound, where we use the error bound developed in Section 4.3.1. MLMC requires a cost model and variance estimates of the correctors $\mathrm{Var}[Y_\ell]$ to determine N_ℓ in (17). As mentioned previously, the variance become increasingly costly to estimate as ℓ increases.

Before evaluating the four setups, we present experiments performed to estimate variances and cost models.

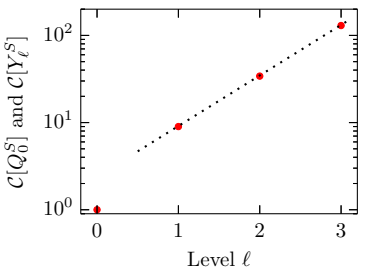
Variance estimates were computed based on a sample of size $4 \cdot 10^4$ for $\ell = 0, 1, 2$ and based on a sample of size $4 \cdot 10^3$ for $\ell = 3$. 95% confidence intervals for the variance were computed for $\ell \geq 1$. The results are presented in Figure 6a. The dotted line is the graph of $\ell \mapsto C\gamma^\ell$ with $C = 0.025$ and $\gamma = 0.65$ and is a convergence rate estimate based on this figure. While the last level ($\ell = 3$) does not verify the rate due to the large confidence interval, it still does not contradict it. Based on this experiment, we extrapolate the rate to hold for all ℓ and base our variance estimate for $\ell \geq 1$ on the function graphed by the dotted line.

As a basis for the cost model we use that the expected computational cost $\mathcal{C}[Q_\ell]$ for a realization follows very closely the relation $\mathcal{C}[Q_\ell] = 2^{3\ell}$ since the mesh size parameter and time step are halved for every level. For the range of meshes used, the computational cost scales linearly with the number of degrees of freedom in the discretization. As is mentioned in Appendix A, the cost to generate random permeability is negligible compared to the cost to do the two-phase flow simulation. The cost for an MLMC-corrector is $\mathcal{C}[Y_\ell] = \mathcal{C}[Q_\ell] + \mathcal{C}[Q_{\ell-1}]$. The costs for a selectively refined realization of Q_ℓ^S and corrector Y_ℓ^S are equal (since $Q_{\ell-1}^S$ comes for free when computing Q_ℓ^S) and are $\mathcal{C}\left[Y_\ell^S\right] = \mathcal{C}\left[Q_\ell^S\right] = \mathcal{C}[Q_0] + \mathcal{C}[Q_1] + \ldots + \mathcal{C}[Q_J]$ for some $J \leq \ell$ where J is random.

The mean computational cost $\mathcal{C}\left[Q_0^S\right]$ and $\mathcal{C}\left[Y_\ell^S\right]$ at each level $\ell = 1, 2$ and 3 were computed. The results are shown in Figure 6b. Based on this experiment, the cost model for the correctors was chosen as $\mathcal{C}\left[Y_\ell^S\right] = C\gamma^{-q\ell}$ with $C = 2.4$ and $q = 3.1$ (where $\gamma = 0.65$ from the variance experiment).



(a) Variance estimates for Q_0 and Y_ℓ as function of level. Red points are variance estimates. Blue bars are confidence intervals. Dotted line is graph $(\ell, C\gamma^\ell)$ with $C = 0.025$ and $\gamma = 0.65$.



(b) Expected computational cost for selective refinement alorithm as function of level. Red points are mean costs. Dotted line is graph $(\ell, C\gamma^{-q\ell})$ with $C = 2.4$, $\gamma = 0.65$ and $q = 3.1$.

Figure 6: Results from experiment in Section 4.4.

In summary, we use the following models for variance and expected cost, with $\gamma = 0.65$

and $\ell \geq 1$:

$$\begin{aligned} \mathrm{Var}\left[Q_0\right] &= \mathrm{Var}\left[Q_0^S\right] \approx 0.0401, \\ \mathrm{Var}\left[Y_\ell\right] &= \mathrm{Var}\left[Y_\ell^S\right] \approx 0.025 \cdot \gamma^\ell, \\ \mathcal{C}\left[Q_0\right] &= \mathcal{C}\left[Q_0^S\right] = 1 \quad \text{(this defines unit time)}, \\ \mathcal{C}\left[Q_\ell\right] &\approx 2^{3\ell}, \\ \mathcal{C}\left[Y_\ell\right] &\approx (2^3+1)\cdot 2^{3(\ell-1)} \approx 1.125 \cdot \gamma^{-4.8\ell}, \\ \mathcal{C}\left[Q_\ell^S\right] &= \mathcal{C}\left[Y_\ell^S\right] \approx 2.4 \cdot \gamma^{-3.1\ell}. \end{aligned} \tag{20}$$

To save computational resources, selective refinement was used to do this experiment, whose total cost was $2.32 \cdot 10^6$.

4.4.1 Monte Carlo

We picked a sample of size $N = 2000$ and computed two estimates of $\widehat{Q}_{N,L}^{\mathrm{MC}}$. The sample mean was 0.0755 and sample variance $3.49 \cdot 10^{-5}$ working as estimate for $\mathrm{Var}\left[\widehat{Q}_{N,L}^{\mathrm{MC}}\right]$. The expected computational cost for this estimator is $\mathcal{C}\left[\widehat{Q}_{N,L}^{\mathrm{MC}}\right] = 2000 \cdot 2^{3\cdot 4} = 8.19 \cdot 10^6$.

4.4.2 Monte Carlo with selective refinement

Based on the same sample as for the MC method, we computed $\widehat{Q}_{N,L}^{\mathrm{MC,S}}$ using the error bound $\eta \tilde{E}_k$ presented in Section 4.3.1. Note that $\tilde{E}_0$ is not available, since the error bound relies on a simulation on a coarser mesh. Starting with $j = 1$ in Algorithm 1, the MC-SR estimator still uses realizations on level 0 in order to compute $\tilde{E}_1$.

The estimate results (number of failures) were identical to those of MC. The computational cost (including the cost to compute error bounds) for this estimator was estimated using the cost model in (20) to $\mathcal{C}\left[\widehat{Q}_{N,L}^{\mathrm{MC}}\right] = 2000 \cdot 2.4 \cdot \gamma^{-3.1\cdot 4} = 1.00 \cdot 10^6$.

4.4.3 Multilevel Monte Carlo

Using the equations for N_ℓ in (17), we choose $\mathrm{TOL} = 10^{-2}$ and compute the sample sizes (rounded up) for each level. The total cost for this estimator was estimated using the cost model in (20) to $\mathcal{C}\left[\widehat{Q}_{\{N_\ell\},L}^{\mathrm{ML}}\right] = 1.28 \cdot 10^6$.

4.4.4 Multilevel Monte Carlo with selective refinement

Using (17) with $\mathrm{TOL} = 10^{-2}$ the sample sizes (rounded up) were computed for each level. The total cost (including the cost to compute error bounds) for this estimator was estimated using the cost model in (20) to $\mathcal{C}\left[\widehat{Q}_{\{N_\ell\},L}^{\mathrm{ML,S}}\right] = 2.65 \cdot 10^5$.

4.4.5 Variance of MLMC estimators

The performance of the four estimators $\widehat{Q}$ can be compared by examining the product $\mathcal{C}\left[\widehat{Q}\right] \cdot \mathrm{Var}\left[\widehat{Q}\right]$. Given a deepest level L, this product is constant for all methods under

Table 4: Expected cost estimates, variance estimates and their product for the four setups.

Estimator ($\widehat{Q}$)	$\mathcal{C}\left[\widehat{Q}\right]$	$\mathrm{Var}\left[\widehat{Q}\right]$	$\mathcal{C}\left[\widehat{Q}\right] \cdot \mathrm{Var}\left[\widehat{Q}\right]$
MC ($\widehat{Q}^{\mathrm{MC}}_{N,L}$)	$8.19 \cdot 10^6$	$3.49 \cdot 10^{-5}$	286
MC-SR ($\widehat{Q}^{\mathrm{MC,S}}_{N,L}$)	$1.00 \cdot 10^6$	$3.49 \cdot 10^{-5}$	34.9
MLMC ($\widehat{Q}^{\mathrm{ML}}_{\{N_\ell\},L}$)	$1.28 \cdot 10^6$	$3.23 \cdot 10^{-5}$	41.3
MLMC-SR ($\widehat{Q}^{\mathrm{ML,S}}_{\{N_\ell\},L}$)	$2.65 \cdot 10^5$	$3.23 \cdot 10^{-5}$	8.56

Table 5: Cost (relative to the cost of MLMC-SR) to realize the four estimators with equal variance.

	No selective refinement	Selective refinement
Monte Carlo	33.4	4.08
Multilevel Monte Carlo	4.82	1.00

investigation, since they are all Monte Carlo methods where sample size and variance are inversely proportional. Based on the expected cost estimates and variance estimates, this product is presented in Table 4 for each setup. A description of the estimation of the variance of the MLMC (and MLMC-SR) estimator follows.

To determine the variance of the MLMC estimators, a sample of 50 MLMC-SR estimates was computed and the sample variance of these 50 estimators was $\mathrm{Var}\left[\widehat{Q}^{\mathrm{ML,S}}_{\{N_\ell\},L}\right] \approx 3.23 \cdot 10^{-5}$. The variance of the MLMC estimator without selective refinement is assumed to be similar, based on the argument that the selective refinement procedure introduces only a bias in the order of 10^{-3} cannot have a significant impact on this variance estimate.

The values of the 50 estimates are plotted at $\ell = L = 4$ in Figure 7 together with a 95% confidence interval for the estimator. Additionally, for all $\ell = 0, \ldots, 4$, the mean values of the 50 truncated estimators (i.e. where only the first ℓ correctors in (16) are included) are plotted in the figure together with a 95% confidence interval for these mean values.

In this experiment, we have assumed normality of the MLMC estimator. Normality of the MLMC estimator does hold asymptotically using a generalized central limit theorem, under the condition that the numerical bias decreases as $\mathrm{TOL}^{1-\epsilon}$ for some $0 < \epsilon < 1$ while the statistical error decreases as TOL when $\mathrm{TOL} \to 0$ (see Lemma 2). For this particular experiment, a Q–Q-plot was used to verify that the estimator was close to normally distributed.

4.4.6 Discussion

The obtained products $\mathcal{C}\left[\widehat{Q}\right] \cdot \mathrm{Var}\left[\widehat{Q}\right]$ normalized with respect to the cheapest setup for all combinations are presented in Table 5. The results show that there are significant gains in computational cost using multilevel Monte Carlo and selective refinement in addition to standard Monte Carlo. Also, the combination of multilevel Monte Carlo and selective refinement gives further gains, in total a factor 33.

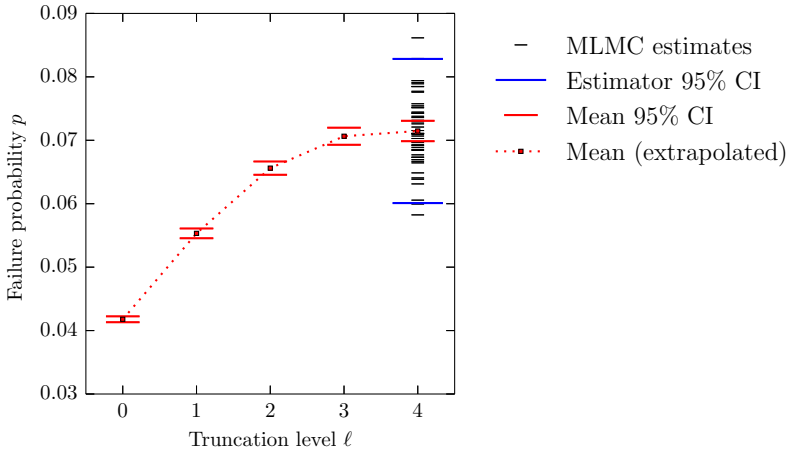


Figure 7: Narrow (black) bars at $\ell = 4$ are 50 MLMC-SR estimates. Wide (blue) bars at $\ell = 4$ cover the 95% confidence interval (CI) for the estimator. Squares (red) at $\ell = 0, \ldots, 4$ are the mean values of 50 the level ℓ-truncated estimates. Bars (red) indicate the 95% CI for these mean values.

These gains depend on the choice of L and they will increase as L increases. For the MLMC estimators constructed above, the goal tolerance TOL was chosen as $\text{TOL} = 10^{-2}$ and the number of samples were chosen based on that. It is clear from Figure 7 that there is a large numerical bias for low values of ℓ. However, the bias improvement obtained by going from $\ell = 2$ to $\ell = 3$ or 4 is dominated by the statistical error of the estimator, indicated by the 95% confidence interval for the estimator at $\ell = 4$. A better balance between the two sources of error at this level L would be attained if TOL was chosen smaller.

In addition to the cost to realize an estimator, the cost to construct the estimators differ between the setups. For selective refinement we established the reliability of an error bound by numerical experiments in Section 4.3.1 (which is not necessary if the error bound is already trusted). For multilevel Monte Carlo, we performed an additional experiment to determine the variance of the correctors and to determine the cost model for selective refinement.

5 Conclusion

It is evident from the experiments on the two-phase flow simulations that if a good error bound on the quantity of interest (in this case sweep efficiency) is available, computational savings of one order of magnitude is possible by using selective refinement. Significant savings in computational cost can most likely be expected from other similar applications by using selective refinement in combination with standard and multilevel Monte Carlo estimators.

A Circulant embedding

The circulant embedding technique [10] is a fast method to generate stationary random fields on a cartesian grid with equispaced points for a given covariance function C.

Let n be the number of rows and columns in the mesh $\mathcal{T}_h$, i.e. $n = h^{-1}$. The total number of vertices is n^2. Let x_k denote the coordinates of vertex k, where the vertices are ordered lexicographically, for example row-by-row. The covariance matrix $\mathbf{C} = (c_{ij})_{ij}$ for the vertices can be written

$$c_{ij} = C(\|x_i - x_j\|_2),$$

where C was defined in Section 2.2. The lexicographical ordering of the vertices in the cartesian grid gives $\mathbf{C}$ a symmetric Toeplitz block structure of symmetric Toeplitz blocks (each block corresponding to a row in the grid, and each column within a block to a column in the row). We define a $(2n-2) \times n$ extension matrix $\mathbf{E}_n$ such that for any $\mathbf{y}$, we have $\tilde{\mathbf{y}} = \mathbf{E}_n\mathbf{y}$ where $\tilde{y}_i = y_i$ for $1 \leq i \leq n$ and $\tilde{y}_{n+i} = y_{n-i}$ for $1 \leq i \leq n-2$. We further define $\mathbf{E} = \mathbf{E}_n \otimes \mathbf{E}_n$. Additionally, we define an $n \times (2n-2)$ restriction matrix $\mathbf{R}_n$ such that for any $\tilde{\mathbf{y}}$, it holds $\mathbf{y} = \mathbf{R}_n\tilde{\mathbf{y}}$ where $y_i = \tilde{y}_i$ for $1 \leq i \leq n$. Also, $\mathbf{R} = \mathbf{R}_n \otimes \mathbf{R}_n$. Note that $\mathbf{RE} = \mathbf{I}$, the identity matrix. Finally, we define the 2D discrete Fourier transform matrix $\mathbf{F} = \mathbf{F}_{2n-2} \otimes \mathbf{F}_{2n-2}$, where $\mathbf{F}_{2n-2}$ is the discrete Fourier transform matrix of size $2n-2$.

The covariance matrix $\mathbf{C}$ has an extension $\widetilde{\mathbf{C}} = \mathbf{ECE}^T$ that is a block circulant matrix with circulant blocks where each block is of size $2n-2$ and there are $2n-2$ blocks per dimension. (We assume this minimal embedding is nonnegative definite, which it always is for our choice of correlation lengths and grid sizes). This block structure allows $\widetilde{\mathbf{C}}$ to be diagonalized by the 2D discrete Fourier transform (see e.g. [8]),

$$\widetilde{\mathbf{C}} = \mathbf{F}^H\Lambda\mathbf{F},$$

where $\cdot^H$ denotes conjugate transpose and Λ is a diagonal matrix with the eigenvalues on its diagonal. The 2D discrete Fourier transform of the first row $\tilde{\mathbf{c}}_1$ of $\widetilde{\mathbf{C}}$ determines the eigenvalues,

$$\Lambda = \operatorname{diag}((2n-2)\mathbf{F}\tilde{\mathbf{c}}_1).$$

Now, let $\epsilon = \epsilon_1 + \mathrm{i}\epsilon_2$, with $\epsilon_1, \epsilon_2 \in \mathcal{N}(\mathbf{0}, \mathbf{I})$ be a random complex vector of length $(2n-2)^2$ and let $\tilde{\mathbf{y}} = \mathbf{F}^H\Lambda^{1/2}\epsilon$. Due to the linear relation with ϵ, both the real and imaginary part of $\tilde{\mathbf{y}}$ are normally distributed with zero mean with covariance $\widetilde{\mathbf{C}}$. Let $\mathbf{y}$ be the vector of interest embedded in $\tilde{\mathbf{y}}$, i.e. $\mathbf{y} = \mathbf{R}\tilde{\mathbf{y}}$. The real part of $\mathbf{y}$ can be written $\Re(\mathbf{y}) = (\mathbf{y} + \overline{\mathbf{y}})/2$, and the covariance matrix of the real part is $\mathbb{E}\left[\Re(\mathbf{y})\Re(\mathbf{y})^T\right] = \mathbb{E}\left[\mathbf{y}\mathbf{y}^H + \mathbf{y}\mathbf{y}^T\right]/2$. We have

$$\mathbb{E}\left[\mathbf{y}\mathbf{y}^H\right] = \mathbb{E}\left[\mathbf{R}\tilde{\mathbf{y}}\tilde{\mathbf{y}}^H\mathbf{R}^T\right] = \mathbf{RF}^H\Lambda^{1/2}\mathbb{E}\left[\epsilon\epsilon^H\right]\Lambda^{1/2}\mathbf{FR}^T = 2\mathbf{RF}^H\Lambda\mathbf{FR}^T = 2\mathbf{C},$$

and $\mathbb{E}\left[\mathbf{y}\mathbf{y}^T\right] = \mathbf{0}$, since $\mathbb{E}\left[\epsilon\epsilon^H\right] = \mathbf{I}$ and $\mathbb{E}\left[\epsilon\epsilon^T\right] = \mathbf{0}$. That is, the real part of $\mathbf{y}$ has covariance matrix $\mathbf{C}$. This also holds for the imaginary part, and the two parts are independent. Thus, for each vector $\mathbf{y}$, two fields are generated. However, the additional data generated (roughly 3/4) due to the extension cannot be used.

In terms of computational cost, computing Λ is an $\mathcal{O}(h^{-2}\log(h^{-2}))$ operation and generating a realization of $\mathbf{y}$ is also an $\mathcal{O}(h^{-2}\log(h^{-2}))$ operation. This is cheaper than

using e.g. Cholesky factorization of $\mathbf{C}$, which is $\mathcal{O}(h^{-4})$ for factorization and generating a realization.

The mesh on which the fields are generated need to be chosen fine enough to avoid aliasing effects (see e.g. [14]). For the lower levels the meshes are coarse enough for the aliasing effect to be significant for pointwise sampling of the random field. Our approach is as follows. We compute Λ based on the finest mesh with $h = 1/256$ (with 65536 vertices) to make the aliasing effect negligible. When a random field is to be computed for a coarser mesh with $\tilde{n}^2 = \tilde{h}^{-2}$ vertices ($\tilde{n} < n$), we truncate Λ and keep only the lowest $\tilde{n}^2$ frequencies (the $\tilde{n}$ lowest for each dimension and their combinations). Random numbers are generated only for the kept eigenvalues. Refinement of a field to a finer mesh is done by keeping the coefficients for the low frequency eigenfunctions and generating new random numbers for higher frequency eigenfunctions to match the number of used eigenvalues with the number of vertices. The computational cost to compute Λ is still $\mathcal{O}(h^{-2}\log(h^{-2}))$, however, this only needs to be done once. The computational cost to generate a new realization depends only on the coarse mesh for which it is generated and is $\mathcal{O}(\tilde{h}^{-2}\log(\tilde{h}^{-2}))$. Thus, the computational cost of generating random data for a simulation is asymptotically negligible in comparison to performing the simulation itself. Solving the PDE numerically costs at least $\mathcal{O}(\tilde{h}^{-3})$.

B Asymptotic normality of MLMC estimator

The asymptotic normality holds for $\widehat{Q}^{\mathrm{MC}}_{N,L}$ by the central limit theorem (CLT) and for $\widehat{Q}^{\mathrm{ML}}_{\{N_\ell\},L}$ with fixed L as $\mathrm{TOL} \to 0$ by applying CLT to each MC estimator in the MLMC estimator individually. However, if numerical bias and statistical error are to be reduced simultaneously, a result from [7] applied to our problem states that asymptotic normality of $\widehat{Q}^{\mathrm{ML}}_{\{N_\ell\},L}$ as $\mathrm{TOL} \to 0$ holds if the convergence rate of the numerical bias is strictly smaller than the rate of the statistical error, i.e. new levels are added to the estimator slightly slower than necessary for even balance between the two error sources.

Lemma 2. *Assume $\mathbb{E}[|Y_\ell|] \geq C\gamma^\ell$ for some $C > 0$. If letting $L = \log(\mathrm{TOL}^{1-\epsilon})/\log(\gamma)$ for some $0 < \epsilon < 1$ and choosing N_ℓ according to* (17)*, then for $\mathcal{A} = \widehat{Q}^{\mathrm{ML}}_{\{N_\ell\},L}$,*

$$\lim_{\mathrm{TOL}\to 0} \Pr\left(\frac{\mathcal{A} - \mathbb{E}[\mathcal{A}]}{\sqrt{\mathrm{Var}[\mathcal{A}]}} \leq z\right) = \Phi(z),$$

where Φ is the cdf for the standard normal distribution.

Proof. We apply [7, Lemma 7.1] which is based on the Lindeberg–Feller theorem on asymptotic normality for sums of independent but not necessarilly identically distributed random variables.

Considering the assumptions for [7, Lemma 7.1], we first establish $C_1\gamma^\ell \leq \mathbb{E}[(Y_\ell - \mathbb{E}[Y_\ell])^2]$. Since Y_ℓ is discrete with outcomes -1, 0 and 1, we can set $\mathbb{E}[Y_\ell] = p_1 - p_{-1}$ for some $p_1, p_{-1} \geq 0$. Then, using $p_1 + p_{-1} = \mathbb{E}[|Y_\ell|] \geq C\gamma^\ell$, we get

$$\mathbb{E}\left[(Y_\ell - \mathbb{E}[Y_\ell])^2\right] = p_1 + p_{-1} - (p_1 - p_{-1})^2 \geq p_1 + p_{-1} - (p_1 + p_{-1})^2 \geq C_1\gamma^\ell$$

for some $C_1 > 0$. Next, since $|Y_\ell - \mathbb{E}[Y_\ell]|$ is bounded by 2, we can bound

$$\mathbb{E}\left[|Y_\ell - \mathbb{E}[Y_\ell]|^3\right] = C_2\mathbb{E}\left[|Y_\ell - \mathbb{E}[Y_\ell]|^2\right] \leq C_3\gamma^\ell$$

for some $C_2, C_3 > 0$. Referring to the notation in [7, Lemma 7.1] we let $\beta = \gamma^{-1}$, $\delta = q_2 = q_3 = \tau = 1$ and apply the lemma to obtain the result. □

Acknowledgement

We are grateful to Andreas Hellander and Salman Toor for assistance with computational parallelization tools (MOLNs) and resources (SMOG Cloud).

References

[1] R. Avikainen. On irregular functionals of SDEs and the Euler scheme. *Finance Stoch.*, 13(3):381–401, 2009.

[2] 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.

[3] D. Boffi, F. Brezzi, and M. Fortin. *Mixed finite element methods and applications*, volume 44 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin Heidelberg, 2nd edition, 2013.

[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] Z. Chen, G. Huan, and B. Li. An improved IMPES method for two-phase flow in porous media. *Transp. Porous Media*, 54(3):361–376, 2004.

[6] 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.

[7] 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.

[8] P. J. David. *Circulant matrices*. John Wiley & Sons, Inc., New York, 1979.

[9] D. A. Di Pietro and A. Ern. *Mathematical Aspects of Discontinuous Galerkin Methods*. Springer-Verlag Berlin Heidelberg, 2012. Mathématiques et Applications, Vol. 69.

[10] 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.

[11] 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.*, 2:826–850, 2014.

[12] D. Elfverson, F. Hellman, and A. Målqvist. A multilevel Monte Carlo method for computing failure probabilities. *ArXiv e-prints:1408.6856*, 2014.

[13] L. W. Gelhar. Stochastic subsurface hydrology from theory to applications. *Water Resources Research*, 22(9S):135S–145S, 1986.

[14] L. W. Gelhar. *Stochastic subsurface hydrology*. Prentice Hall, New Jersey, 1993.

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

[16] M. B. Giles, T. Nagapetyan, and K. Ritter. Multilevel Monte Carlo approximation of distribution functions and densities. *SIAM/ASA J. Uncertain. Quantif.*, 3(1):267–295, 2015.

[17] A.-L. Haji-Ali, F. Nobile, and R. Tempone. Multi-index Monte Carlo: when sparsity meets sampling. *Numer. Math*, pages 1–40, 2015.

[18] 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.

[19] S. Heinrich. Multilevel monte carlo methods. In *Large-Scale Scientific Computing*, volume 2179 of *Lecture Notes in Computer Science*, pages 58–67. Springer Berlin Heidelberg, 2001.

[20] R. J. LeVeque. *Finite volume methods for hyperbolic problems*. Cambridge University Press, New York, 2002. Cambridge Texts in Applied Mathematics.

[21] M. Park and A. Teckentrup. Improved multilevel Monte Carlo methods for finite volume discretisations of darcy flow in randomly layered media. *ArXiv e-prints:1506.04694*, 2015.

[22] P. A. Raviart and J. M. Thomas. A mixed finite element method for 2-nd order elliptic problems. In Ilio Galligani and Enrico Magenes, editors, *Mathematical Aspects of Finite Element Methods*, volume 606 of *Lecture Notes in Mathematics*, pages 292–315. Springer Berlin Heidelberg, 1977.

[23] 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.

Paper IV

Multiscale mixed finite elements

Fredrik Hellman[1], Patrick Henning[2], Axel Målqvist[3]

August 31, 2015

Abstract

In this work, we propose a mixed finite element method for solving elliptic multiscale problems based on a localized orthogonal decomposition (LOD) of Raviart–Thomas finite element spaces. It requires to solve local problems in small patches around the elements of a coarse grid. These computations can be perfectly parallelized and are cheap to perform. Using the results of these patch problems, we construct a low dimensional multiscale mixed finite element space with very high approximation properties. This space can be used for solving the original saddle point problem in an efficient way. We prove convergence of our approach, independent of structural assumptions or scale separation. Finally, we demonstrate the applicability of our method by presenting a variety of numerical experiments, including a comparison with an MsFEM approach.

1 Introduction

In this work we study the mixed formulation of Poisson's equation with a multiscale diffusion coefficient, i.e. where the diffusion coefficient is highly varying on a continuum of different scales. For such coefficients, the solution is typically also highly varying and standard Galerkin methods fail to converge to the correct solution, unless the features on the finest scale are resolved by the underlying computational mesh. A classical application is the flow in a porous medium, modeled by Darcy's law. In this case, the multiscale coefficient describes a permeability field, which is heterogeneous, rapidly varying and has high contrast. Classical discretizations that involve the full fine scale often lead to a vast number of degrees of freedom, which limits the performance and feasibility of corresponding computations. In this paper, we address this kind of problems in the context of mixed finite elements.

We will interpret the mixed formulation of Poisson's equation in a Darcy flow setting, referring to the vector component as flux, and the scalar component as pressure. In Darcy flow applications the flux solution is of particular interest since it tells us how a fluid is transported through the medium. It is desirable and common to use flux conservative discretization schemes. The proposed method is based on the Raviart–Thomas finite element [29] which is locally flux conservative. Concerning the mixed formulation

[1]Department of Information Technology, Uppsala University, Box 337, SE-751 05 Uppsala, Sweden. Supported by Centre for Interdisciplinary Mathematics (CIM), Uppsala University.

[2]Institute for Computational and Applied Mathematics, University of Münster, Einsteinstrasse 62, Germany.

[3]Department of Mathematical Sciences, Chalmers University of Technology and University of Gothenburg SE-412 96 Göteborg, Sweden. Supported by the Swedish Research Council.

of Poisson's equation, corresponding multiscale methods were for instance proposed in [1, 3, 4, 7]. These methods are based on the Raviart–Thomas finite element and fit into the framework of the Multiscale Finite Element Method (MsFEM, cf. [17]). Another family of multiscale methods is derived from the framework of the Variational Multiscale Method (VMS) [18, 19, 20, 22, 27]. A multiscale method for mixed finite elements based on VMS is proposed and studied in [23, 26]. Inspired by the results presented in [26], a new multiscale framework arose [24]. We refer to this framework as Localized Orthogonal Decomposition (LOD). It is based on the idea that a finite element space is decomposed into a low dimensional space that incorporates multiscale features and a high dimensional remainder space which is given as the kernel of an interpolation or quasi-interpolation operator. The multiscale space can be used for Galerkin-approximations and allows for cheap computations. Various realizations have been proposed so far. For corresponding formulations and rigorous convergence results for elliptic multiscale problems, we refer to [12, 13, 16, 24] for Galerkin finite element methods, to [11, 12] for discontinuous Galerkin methods and to [15] for Galerkin Partition of Unity methods. Among the various applications we refer to the realizations for eigenvalue problems [25], for semilinear equations [14], for the wave equation [2] and for the Helmholtz equation [28].

In this paper we introduce a two level discretization of the mixed problem, that is we work with two meshes: A fine mesh (mesh size h) which resolves all the fine scale features in the solution and a coarse mesh (mesh size H) which is of computationally feasible size. This gives us a fine and a coarse Raviart–Thomas function space for the flux. We denote them respectively by V_h (high dimensional) and V_H (low dimensional). The kernel of the (standard) nodal Raviart–Thomas interpolation operator Π_H onto V_H is the detail space V_h^{f}. This space can be interpreted as all fine scale features that can not be captured in the coarse space V_H. A low dimensional ideal multiscale space is constructed as the orthogonal complement to the divergence free fluxes in V_h^{f}. We prove that this space has good approximation properties in the sense that the energy norm of the error converges with H without pre-asymptotic effects due to the multiscale features. However, the basis functions of the ideal multiscale space have global support and are expensive to compute. We show exponential decay of these basis functions allowing them to be truncated to localized patches with a preserved order of accuracy for the convergence. The resulting space is called the localized multiscale space. The problems that are associated with the localized basis functions have a small number of degrees of freedom and can be solved in parallel with reduced computational cost and memory requirement. Once computed, the low dimensional localized multiscale space can be reused in a nonlinear or time iterative scheme.

We prove inf-sup stability and a priori error estimates (of linear order in H) for both the ideal and the localized method. The local L^2-instability of the nodal Raviart–Thomas interpolation operator leads to instabilities as h decreases for the localized method. We show that these instabilities can be compensated by increasing the patch size or using Clément-type interpolators instead. In the numerical examples we verify that the localized method has the theoretically derived order of accuracy. We confirm our theoretical findings by performing experiments on the unit square and an L-shaped domain, as well as using a diffusion coefficient with high contrast noise and channel structures. The proposed method is also compared numerically with results from an MsFEM-based approach using a permeability field from the SPE10 benchmark problem.

2 Preliminaries

We consider a bounded Lipschitz domain $\Omega \subset \mathbb{R}^d$ (dimension $d = 2$ or 3) with a piecewise polygonal boundary $\partial\Omega$ and let $\mathbf{n}$ denote the outgoing normal vector of $\partial\Omega$. For any subdomain $\omega \subseteq \Omega$, we shall use standard notation for Lebesgue and Sobolev spaces, i.e. for $r \in [1, \infty]$, $L^r(\omega)$ consists of measurable functions with bounded L^r-norm and the space $H^1(\omega)$ consists of L^2-bounded weakly differentiable functions with L^2-bounded partial derivatives. The full norm on $H^1(\omega)$ shall be denoted by $\|\cdot\|_{H^1(\omega)}$, whereas the semi-norm is denoted by $|\cdot|_{H^1(\omega)} := \|\nabla \cdot\|_{L^2(\omega)}$.

For scalar functions p and q we denote by $(p, q)_\omega := \int_\omega p\, q$ the L^2-scalar product on ω. When $\omega = \Omega$, we omit the subscript, i.e. $(p, q) := (p, q)_\Omega$. For d-dimensional vector valued functions $\mathbf{u}$ and $\mathbf{v}$, we define $(\mathbf{u}, \mathbf{v})_\omega := \int_\omega \mathbf{u} \cdot \mathbf{v}$ with $(\mathbf{u}, \mathbf{v}) = (\mathbf{u}, \mathbf{v})_\Omega$. Observe that we use the same notation for norms and scalar products in L^2 without distinguishing between scalar and vector valued functions. This is purely for simplicity, since the appropriate definition is always clear from the context. We use, however, bold face letters for vector valued quantities.

In the following, we define the Sobolev space of functions with L^2-bounded weak divergence by $H(\mathrm{div}, \omega) := \{\mathbf{v} \in [L^2(\omega)]^d : \nabla \cdot \mathbf{v} \in L^2(\omega)\}$. We equip this space with the usual norm $\|\cdot\|_{H(\mathrm{div},\omega)}$, where $\|\mathbf{v}\|^2_{H(\mathrm{div},\omega)} := \|\nabla \cdot \mathbf{v}\|^2_{L^2(\omega)} + \|\mathbf{v}\|^2_{L^2(\omega)}$. Additionally, for $\omega = \Omega$, we introduce the subspace $H_0(\mathrm{div}, \Omega) := \{\mathbf{v} \in H(\mathrm{div}, \Omega) : \mathbf{v} \cdot \mathbf{n}|_{\partial\Omega} = 0\}$ of functions with zero flux on the boundary, where $\mathbf{v} \cdot \mathbf{n}|_{\partial\Omega}$ should be interpreted in the sense of traces. We denote by $L^2(\Omega)/\mathbb{R} := \{q \in L^2(\Omega) : \int_\Omega q = 0\}$ the quotient space of $L^2(\Omega)$ by $\mathbb{R}$. The continuous dual space of a Banach space X is denoted by X'.

2.1 Continuous problem

With these definitions we are ready to state the continuous problem, which is Poisson's equation in mixed form with Neumann boundary conditions on the full boundary.

Definition 1 (Continuous problem)**.** *Find* $\mathbf{u} \in V := H_0(\mathrm{div}, \Omega)$, $p \in Q := L^2(\Omega)/\mathbb{R}$ *such that*

$$\begin{aligned} \left(\mathbf{A}^{-1}\mathbf{u}, \mathbf{v}\right) + (\nabla \cdot \mathbf{v}, p) &= 0, \\ (\nabla \cdot \mathbf{u}, q) &= -(f, q), \end{aligned} \tag{1}$$

for all $\mathbf{v} \in V$, $q \in Q$.

We pose the following assumptions on the coefficient and data.

Assumption A (Assumptions on coefficients, data and domain)**.**

(A1) $\mathbf{A} \in [L^\infty(\Omega)]^{d\times d}$ *is a diffusion coefficient, possibly with rapid fine scale variations. Its value is an almost everywhere symmetric matrix and bounded in the sense that there exist real numbers* α *and* β *such that for almost every* x *and any* $\mathbf{v} \in \mathbb{R}^d/\{0\}$

$$0 < \alpha \leq \frac{(\mathbf{A}(x)^{-1}\mathbf{v}) \cdot \mathbf{v}}{\mathbf{v} \cdot \mathbf{v}} \leq \beta < \infty.$$

(A2) $f \in L^2(\Omega)$ *is a source function that fulfills the compatibility condition* $\int_\Omega f = 0$.

(A3) *The domain Ω is a bounded Lipschitz domain with polygonal (or polyhedral) boundary.*

We introduce the following bilinear forms and norms. Let

$$a(\mathbf{u},\mathbf{v}) := (\mathbf{A}^{-1}\mathbf{u},\mathbf{v}) \quad \text{and} \quad b(\mathbf{v},q) := (\nabla\cdot\mathbf{v},q)$$

and, further,

$$\|\mathbf{v}\|_V := \|\mathbf{v}\|_{H(\mathrm{div},\Omega)} \quad \text{and} \quad \|q\|_Q := \|q\|_{L^2(\Omega)}.$$

The energy norm is defined as the following weighted flux L^2-norm,

$$|||\mathbf{v}|||^2 := \|\mathbf{A}^{-1/2}\mathbf{v}\|^2_{L^2(\Omega)} = a(\mathbf{v},\mathbf{v})$$

The energy norm can be subscripted with a subdomain $\omega \subseteq \Omega$, for example $|||\cdot|||^2_\omega$, to indicate that the integral is taken only over that subdomain.

The following lemma gives the conditions for existence and uniqueness of a solution to the mixed formulation in (1) for subspaces $\mathcal{V} \subseteq V$ and $\mathcal{Q} \subseteq Q$. This lemma is useful for establishing existence and uniqueness for all discretizations presented in this paper, since all presented discretizations are conforming.

Lemma 2 (Existence and uniqueness of solution to mixed formulation)**.** *Let $\mathcal{V} \subseteq V$ and $\mathcal{Q} \subseteq Q$. Denote by $\mathcal{K} = \{\mathbf{v} \in \mathcal{V} : b(\mathbf{v},q) = 0 \ \forall q \in \mathcal{Q}\}$. If $a(\cdot,\cdot)$ is coercive on $\mathcal{K}$ with constant $\tilde{\alpha} > 0$, i.e. $a(\mathbf{v},\mathbf{v}) \geq \tilde{\alpha}\|\mathbf{v}\|_V^2$ for $\mathbf{v} \in \mathcal{K}$, and bounded with constant $\tilde{\beta} > 0$, i.e. $|a(\mathbf{v},\mathbf{w})| \leq \tilde{\beta}\|\mathbf{v}\|_V\|\mathbf{w}\|_V$ for all $\mathbf{v},\mathbf{w} \in \mathcal{V}$, and additionally $b(\cdot,\cdot)$ is inf-sup stable with constant $\tilde{\gamma} > 0$, i.e.*

$$\inf_{q\in\mathcal{Q}} \sup_{\mathbf{v}\in\mathcal{V}} \frac{b(\mathbf{v},q)}{\|\mathbf{v}\|_V\|q\|_Q} \geq \tilde{\gamma},$$

then the problem $a(\mathbf{u},\mathbf{v}) + b(\mathbf{v},p) - b(\mathbf{u},q) = (f,q)$ for all $(\mathbf{v},q) \in \mathcal{V}\times\mathcal{Q}$ has a unique solution $(\mathbf{u},p) \in \mathcal{V}\times\mathcal{Q}$ bounded by

$$\|\mathbf{u}\|_V \leq \frac{2\tilde{\beta}^{1/2}}{\tilde{\alpha}^{1/2}\tilde{\gamma}}\|f\|_{L^2(\Omega)} \quad \textit{and} \quad \|p\|_Q \leq \frac{\tilde{\beta}}{\tilde{\gamma}^2}\|f\|_{L^2(\Omega)}.$$

Proof. See e.g. [6, Theorem 4.2.3]. □

Under Assumptions (A1)–(A3), the conditions for Lemma 2 are satisfied for $\mathcal{V} = V$ and $\mathcal{Q} = Q$ with $\tilde{\alpha} = \alpha$, $\tilde{\beta} = \beta$ and $\tilde{\gamma}$ being a constant that depends only on the computation domain. The lemma then yields a unique solution to the continuous problem (1).

2.2 Discretization with the Raviart–Thomas element

Regarding the discretization, we introduce two conforming families of simplicial (i.e. triangular or tetrahedral) meshes $\{\mathcal{T}_h\}$ and $\{\mathcal{T}_H\}$ of Ω where h and H are the maximum element diameters. Throughout the paper we refer to $\mathcal{T}_h$ as the fine mesh and to $\mathcal{T}_H$ as the coarse mesh. Hence, we indirectly assume $h < H$. We pose the following assumptions on the meshes.

Assumption B (Assumptions on meshes)**.**

(B1) *The fine mesh $\mathcal{T}_h$ is the result of one or more conforming (but possibly non-uniform) refinements of the coarse mesh $\mathcal{T}_H$ such that $\mathcal{T}_h \cap \mathcal{T}_H = \emptyset$.*

(B2) *Both meshes $\mathcal{T}_h$ and $\mathcal{T}_H$ are shape regular. In particular the positive shape regularity constant ρ for the coarse mesh $\mathcal{T}_H$ will be referred to below and is defined as $\rho = \min_{T\in\mathcal{T}_H} \frac{\operatorname{diam} B_T}{\operatorname{diam} T}$ where B_T is the largest ball contained in the element $T \in \mathcal{T}_H$.*

(B3) *The coarse family of meshes $\{\mathcal{T}_H\}$ is quasi-uniform, whereas $\{\mathcal{T}_h\}$ could be obtained from an arbitrary adaptive refinement.*

Remark 3 (Quadrilateral or hexahedral elements)**.** *Affine quadrilateral (or hexahedral) elements can also be used. However, the definition of the Raviart–Thomas element presented below in this paper is based on triangular (or tetrahedral) meshes.*

We denote by t and T an element of $\mathcal{T}_h$ or $\mathcal{T}_H$, respectively. Similarly e and E denote an edge (for $d = 2$) or a face (for $d = 3$) of the elements of $\mathcal{T}_h$ and $\mathcal{T}_H$. Further, $\mathbf{n}_e$ (respectively $\mathbf{n}_E$) is the outward normal vector of an edge (or face) e (respectively E). We continue this section by discussing finite element discretization using the two meshes.

We denote all polynomials of degree $\leq k$ on a subdomain ω by $\mathbb{P}^k(\omega)$ and a d-dimensional vector of such polynomials by $[\mathbb{P}^k(\omega)]^d$. We introduce the $H_0(\operatorname{div},\Omega)$-conforming lowest (zeroth) order Raviart–Thomas finite element. For each fine element $t \in \mathcal{T}_h$ and coarse element $T \in \mathcal{T}_H$, the spaces of Raviart–Thomas shape functions are given by

$$\mathcal{RT}_h(t) = \{\mathbf{v}|_t = [\mathbb{P}^0(t)]^d + x\mathbb{P}^0(t)\} \text{ and}$$
$$\mathcal{RT}_H(T) = \{\mathbf{v}|_T = [\mathbb{P}^0(T)]^d + x\mathbb{P}^0(T)\},$$

respectively, where $x = (x_1, \ldots, x_d)$ is the space coordinate vector. The Raviart–Thomas finite element spaces on $\mathcal{T}_h$ and $\mathcal{T}_H$ are then defined as

$$V_h = \{\mathbf{v} \in H_0(\operatorname{div},\Omega) : \mathbf{v}|_t \in \mathcal{RT}_h(t) \quad \forall t \in \mathcal{T}_h\} \text{ and}$$
$$V_H = \{\mathbf{v} \in H_0(\operatorname{div},\Omega) : \mathbf{v}|_T \in \mathcal{RT}_H(T) \quad \forall T \in \mathcal{T}_H\}.$$

The degrees of freedom (in the coarse and fine Raviart–Thomas spaces) are given by the averages of the normal fluxes over the edges (respectively faces for $d = 3$). We denote the degrees of freedom by

$$N_e(\mathbf{v}) := \frac{1}{|e|}\int_e \mathbf{v}\cdot\mathbf{n}_e \quad \text{and} \quad N_E(\mathbf{v}) := \frac{1}{|E|}\int_E \mathbf{v}\cdot\mathbf{n}_E$$

for the fine and coarse discretization, respectively. The direction of the normal $\mathbf{n}_e$ (respectively $\mathbf{n}_E$) can be fixed arbitrarily for each edge (respectively face). Here, N_e and N_E are bounded linear functionals on the space $W := H_0(\operatorname{div},\Omega) \cap L^s(\Omega)$, for some $s > 2$. Note, that the additional regularity (i.e. $L^s(\Omega)$ for $s > 2$) is necessary for the edge integrals to be well-defined (cf. [6]). We introduce the (standard) nodal Raviart–Thomas interpolation operators $\Pi_h : W \to V_h$ and $\Pi_H : W \to V_H$ by fixing the degrees of freedom in the natural way, i.e. Π_h and Π_H are defined such that

$$N_e(\Pi_h\mathbf{v}) = N_e(\mathbf{v}) \quad \text{and} \quad N_E(\Pi_H\mathbf{v}) = N_E(\mathbf{v}).$$

Additionally, we let $Q_H \subset Q_h \subset Q$ be the space of all piecewise constant functions on $\mathcal{T}_H$ and $\mathcal{T}_h$ with zero mean. We denote by P_h and P_H the L^2-projections onto Q_h and Q_H, respectively. Using the fine spaces, we define the fine scale discretization of (1), which will be referred to as the reference problem.

Definition 4 (Reference problem)**.** *Find* $\mathbf{u}_h \in V_h$ *and* $p_h \in Q_h$*, such that*

$$\begin{aligned} a(\mathbf{u}_h, \mathbf{v}_h) + b(\mathbf{v}_h, p_h) &= 0, \\ b(\mathbf{u}_h, q_h) &= -(f, q_h), \end{aligned} \tag{2}$$

for all $\mathbf{v}_h \in V_h$ *and* $q_h \in Q_h$*.*

A similar problem can be stated with the coarse spaces V_H and Q_H with flux solution $\mathbf{u}_H$. The remainder of this section treats only the fine discretization. However, all results hold also for the coarse discretization.

We denote the space of divergence free functions on the fine grid by

$$K_h := \{\mathbf{v} \in V_h : \nabla \cdot \mathbf{v} = 0\}. \tag{3}$$

Remark 5 (Kernel of divergence operator)**.** *A natural definition of* K_h *for our purposes is* $K_h = \{\mathbf{v} \in V_h : (\nabla \cdot \mathbf{v}, q_h) = 0 \ \forall q_h \in Q_h\}$*. However, since we have* $\nabla \cdot \mathbf{v} \in Q_h$ *for all* $\mathbf{v} \in V_h$ *(due to the definition of the Raviart–Thomas element), we can characterize* K_h *equivalently as done in* (3)*.*

To establish existence and uniqueness of a solution to the reference problem, we use that Π_h is divergence compatible, i.e. we have the commuting property $\nabla \cdot \Pi_h \mathbf{v} = P_h \nabla \cdot \mathbf{v}$ for $\mathbf{v} \in W$, and that Π_h is bounded on W (but not on V!), i.e. there exists a generic h-independent constant C_W such that $\|\Pi_h \mathbf{v}\|_V \leq C_W \|\mathbf{v}\|_W$ for $\mathbf{v} \in W$. Using this, the inf-sup stability of $b(\cdot, \cdot)$ with respect to V_h and Q_h follows: For $q \in Q_h$,

$$\begin{aligned} \sup_{\mathbf{v} \in V_h} \frac{b(\mathbf{v}, q)}{\|\mathbf{v}\|_V} &= \sup_{\mathbf{v} \in W} \frac{(\nabla \cdot \Pi_h \mathbf{v}, q)}{\|\Pi_h \mathbf{v}\|_V} \geq \sup_{\mathbf{v} \in W} \frac{(\nabla \cdot \mathbf{v}, q)}{C_W \|\mathbf{v}\|_W} \\ &\geq \frac{(\nabla \cdot \mathbf{w}, q)}{C_W \|\mathbf{w}\|_W} \geq \frac{(q, q)}{C_W C_\Omega \|q\|_{L^2(\Omega)}} = C_W^{-1} C_\Omega^{-1} \|q\|_{L^2(\Omega)}, \end{aligned} \tag{4}$$

where $\mathbf{w} \in W$ is chosen such that $\nabla \cdot \mathbf{w} = q$ and $\|\mathbf{w}\|_W \leq C_\Omega \|q\|_{L^2(\Omega)}$. This is possible by letting $\mathbf{w} = \nabla \phi$ for a solution ϕ to $\Delta \phi = q$ in Ω with homogeneous Neumann boundary conditions. Now, applying Lemma 2 with $\mathcal{V} = V_h$, $\mathcal{Q} = Q_h$, $\mathcal{K} = K_h$, we can derive the constants $\tilde{\alpha} = \alpha$, $\tilde{\beta} = \beta$ and $\tilde{\gamma} = \gamma := C_W^{-1} C_\Omega^{-1}$ and establish existence and uniqueness of a solution to the reference problem (2). Note that the inf-sup stability constant γ is independent of h and hence also holds for the pair of spaces V_H and Q_H.

In the following, we are mainly interested in approximating the flux component $\mathbf{u}_h$ of the solution. We treat $\mathbf{u}_h$ as a reliable reference to the exact solution. Note that the L^2-norm of the divergence error is controlled by the data

$$\|\nabla \cdot \mathbf{u} - \nabla \cdot \mathbf{u}_h\|_{L^2(\Omega)} \leq \|f - P_h f\|_{L^2(\Omega)}.$$

For the energy norm of the flux error, we have the following error estimate in the energy norm for the lowest order Raviart–Thomas element:

$$|||\mathbf{u} - \mathbf{u}_h||| \leq Ch |\mathbf{u}|_{H^1(\Omega)},$$

where C is independent of h. For a problem with a coefficient $\mathbf{A}$ that has fast variations at a scale of size ϵ, we have in general that $|\mathbf{u}|_{H^1(\Omega)} \approx \epsilon^{-1}$. Hence, we require $h \ll \epsilon$ before

we can observe the linear convergence in h numerically. We call the regime with $h \geq \epsilon$ a pre-asymptotic regime. The goal of this work is the construction of a discrete space which does not suffer from such pre-asymptotic effects triggered by $\mathbf{A}$. In the following, we assume that the fine mesh is fine enough (in the sense that $h \ll \epsilon$) so that $|||\mathbf{u} - \mathbf{u}_h|||$ is sufficiently small and hence $\mathbf{u}_h$ a sufficiently accurate reference solution. With the same argument, the accuracy of the coarse solution $\mathbf{u}_H$ will not be satisfying as long as $H > \epsilon$. Note that reference problem (2) never needs to be solved. It just serves as a reference.

In the next section, we will construct the ideal multiscale space of the same (low) dimension as V_H, but which yields approximations that are of similar accuracy as the reference solution $\mathbf{u}_h$ (in particular in the regime $H \gg \epsilon$). Throughout the paper, we do not consider errors that arise from numerical quadrature. For simplicity, we assume that all integrals can be computed exactly.

3 Ideal multiscale problem

In this section, we construct a low dimensional space that can capture the fine scale features of the true multiscale solution. We focus on constructing a good multiscale representation of the flux solution $\mathbf{u}$ only. We call it ideal since the reference flux solution is in this space for all $f \in Q_H$. This should be contrasted to a localized multiscale space to be introduced in Section 4. In addition to the spaces V_h and V_H defined above we introduce the following detail space as the intersection of the fine space and the kernel of the coarse Raviart–Thomas interpolation operator,

$$V_h^{\mathrm{f}} = \{\mathbf{v} \in V_h : \Pi_H \mathbf{v} = 0\}.$$

Since V_h^{f} is the kernel of a projection, it induces the splitting $V_h = V_H \oplus V_h^{\mathrm{f}}$, where V_H is low dimensional and V_h^{f} is high dimensional. We refer to V_h^{f} as the detail space. In this section we aim at constructing a modified splitting, where V_H is replaced by a multiscale space which incorporates fine scale features.

3.1 Ideal multiscale space

We will construct the ideal multiscale space by applying fine scale correctors to all coarse functions in V_H, i.e. so that $(\mathrm{Id} - G_h)(V_H)$ is the desired multiscale space for a linear corrector operator G_h. The corrector operator is constructed using information from the coefficient $\mathbf{A}$, and has divergence free range in order to keep the flux conservation property of the coarse space.

The definition of the corrector requires us to construct the splitting $K_h = K_H \oplus K_h^{\mathrm{f}}$ with

$$K_h^{\mathrm{f}} := \{\mathbf{v} \in K_h : \Pi_H \mathbf{v} = 0\}, \quad \text{and} \quad K_H := \mathrm{Range}((\Pi_H)_{|K_h}).$$

Next, we introduce an ideal corrector operator. We distinguish between local (element-wise) correctors and a global corrector.

Definition 6 (Ideal corrector operators)**.** *Let $a^T(\mathbf{u}, \mathbf{v}) := (\mathbf{A}^{-1}\mathbf{u}, \mathbf{v})_T$ for $T \in \mathcal{T}_H$. For each such $T \in \mathcal{T}_H$, we define an* ideal element corrector operator $G_h^T : V \to K_h^{\mathrm{f}}$ *by the equation*

$$a(G_h^T \mathbf{v}, \mathbf{v}^{\mathrm{f}}) = a^T(\mathbf{v}, \mathbf{v}^{\mathrm{f}}) \tag{5}$$

for all $\mathbf{v}^{\mathrm{f}} \in K_h^{\mathrm{f}}$. *Furthermore, we define the* ideal global corrector operator *by summing the local contributions, i.e.* $G_h := \sum_{T \in \mathcal{T}_H} G_h^T$.

The ideal corrector operators are well-defined since equation (5) is guaranteed a unique solution by the Lax–Milgram theorem due to the coercivity and boundedness of a on K_h^{f}. Using the ideal global corrector operator, we can define the discrete multiscale function space by

$$V_{H,h}^{\mathrm{ms}} := (\mathrm{Id} - G_h)(V_H),$$

where Id is the identity operator. This space has the same dimension as V_H. Furthermore, it allows for the splitting $V_h = V_{H,h}^{\mathrm{ms}} \oplus V_h^{\mathrm{f}}$. Note that the ideal multiscale space is the orthogonal complement of K_h^{f} with respect to $a(\cdot,\cdot)$, i.e.

$$a(\mathbf{v}_{H,h}^{\mathrm{ms}}, \mathbf{v}^{\mathrm{f}}) = 0 \tag{6}$$

for all $\mathbf{v}_{H,h}^{\mathrm{ms}} \in V_{H,h}^{\mathrm{ms}}$ and $\mathbf{v}^{\mathrm{f}} \in K_h^{\mathrm{f}}$.

3.2 Ideal multiscale problem formulation

In this section, we use the previously defined ideal multiscale space to define a (preliminary) multiscale approximation. The ideal multiscale problem reads as follows.

Definition 7 (Ideal multiscale problem)**.** *Find* $\mathbf{u}_{H,h}^{\mathrm{ms}} \in V_{H,h}^{\mathrm{ms}}$ *and* $p_H \in Q_H$*, such that*

$$\begin{aligned} a(\mathbf{u}_{H,h}^{\mathrm{ms}}, \mathbf{v}_h) + b(\mathbf{v}_h, p_H) &= 0, \\ b(\mathbf{u}_{H,h}^{\mathrm{ms}}, q_H) &= -(f, q_H), \end{aligned} \tag{7}$$

for all $\mathbf{v}_h \in V_{H,h}^{\mathrm{ms}}$ *and* $q_H \in Q_H$.

Lemma 8 (Unique solution of the ideal multiscale problem)**.** *Under Assumptions (A1)–(A3) and (B1)–(B3), the ideal multiscale problem* (7) *has a unique solution. In particular, we have*

$$\gamma(1 + \alpha^{-1}\beta)^{-1} \leq \inf_{q \in Q_H} \sup_{\mathbf{v} \in V_{H,h}^{\mathrm{ms}}} \frac{b(\mathbf{v}, q)}{\|q\|_Q \|\mathbf{v}\|_V},$$

i.e. inf-sup stability independent of h *and* H.

Proof. We let $K_{H,h}^{\mathrm{ms}} = \{\mathbf{v} \in V_{H,h}^{\mathrm{ms}} : \nabla \cdot \mathbf{v} = 0\}$. The coercivity of $a(\cdot,\cdot)$ on $K_{H,h}^{\mathrm{ms}}$ follows immediately from its coercivity on K_h since $K_{H,h}^{\mathrm{ms}} \subset K_h$. The operator $\mathrm{Id} - G_h$ is stable in V with constant $1 + \alpha^{-1}\beta$, since $\nabla \cdot G_h \mathbf{v} = 0$ and

$$\begin{aligned} \|G_h \mathbf{v}\|_{L^2(\Omega)}^2 &\leq \alpha^{-1} a(G_h \mathbf{v}, G_h \mathbf{v}) \\ &= \alpha^{-1} a(\mathbf{v}, G_h \mathbf{v}) \\ &\leq \alpha^{-1} \beta \|\mathbf{v}\|_{L^2(\Omega)} \|G_h \mathbf{v}\|_{L^2(\Omega)} \end{aligned}$$

for all $\mathbf{v} \in V$. Combining these results with the inf-sup stability of $b(\cdot,\cdot)$ on V_H and Q_H, we get

$$\begin{aligned}\gamma &\leq \inf_{q\in Q_H}\sup_{\mathbf{v}\in V_H}\frac{b(\mathbf{v},q)}{\|q\|_Q\,\|\mathbf{v}\|_V}\\ &\leq (1+\alpha^{-1}\beta)\inf_{q\in Q_H}\sup_{\mathbf{v}\in V_H}\frac{(\nabla\cdot(\mathrm{Id}-G_h)\mathbf{v},q)}{\|q\|_Q\,\|(\mathrm{Id}-G_h)\mathbf{v}\|_V}\\ &= (1+\alpha^{-1}\beta)\inf_{q\in Q_H}\sup_{\mathbf{v}\in V^{\mathrm{ms}}_{H,h}}\frac{(\nabla\cdot\mathbf{v},q)}{\|q\|_Q\,\|\mathbf{v}\|_V},\end{aligned} \tag{8}$$

i.e. $b(\cdot,\cdot)$ is inf-sup stable with constant $\gamma(1+\alpha^{-1}\beta)^{-1}$ independent of H and h. We note that $K^{\mathrm{ms}}_{H,h} = \{\mathbf{v} \in V^{\mathrm{ms}}_{H,h} : b(\mathbf{v},q_H) = 0\ \forall q \in Q_H\}$, since $\nabla\cdot\mathbf{v} \in Q_H$ (see Remark 5). Finally, we apply Lemma 2 with $\mathcal{V} = V^{\mathrm{ms}}_{H,h}$, $\mathcal{Q} = Q_H$, $\mathcal{K} = K^{\mathrm{ms}}_{H,h}$ and constants $\tilde{\alpha} = \alpha$, $\tilde{\beta} = \beta$ and $\tilde{\gamma} = \gamma(1+\alpha^{-1}\beta)^{-1}$. □

3.3 Error estimate for ideal problem

In this section, we show that the flux solution of the ideal multiscale problem above converges in the energy norm with linear order in H to the reference solution. This convergence is independent of the variations of $\mathbf{A}$, i.e. we do not have any pre-asymptotic effects from the multiscale features.

Lemma 9 (Error estimate for ideal solution)**.** *Under Assumptions (A1)–(A3) and (B1)–(B3), let* $\mathbf{u}_h$ *solve* (2) *and* $\mathbf{u}^{\mathrm{ms}}_{H,h}$ *solve* (7)*, then*

$$\left|\!\left|\!\left|\mathbf{u}_h - \mathbf{u}^{\mathrm{ms}}_{H,h}\right|\!\right|\!\right| \leq \beta^{1/2} C_{\widehat{\Pi}} C_{\rho,d} H \|f - P_H f\|_{L^2(\Omega)}$$

where $C_{\rho,d}$ *and* $C_{\widehat{\Pi}}$ *are independent of* h *and* H*.*

Proof. Parametrizing the solutions $\mathbf{u}_h(f)$ and $\mathbf{u}^{\mathrm{ms}}_{H,h}(f)$ by the data f, we use the triangle inequality to obtain

$$\begin{aligned}&\left|\!\left|\!\left|\mathbf{u}_h(f) - \mathbf{u}^{\mathrm{ms}}_{H,h}(f)\right|\!\right|\!\right|\\ &\quad\leq \left|\!\left|\!\left|\mathbf{u}_h(f) - \mathbf{u}_h(P_H f)\right|\!\right|\!\right| + \left|\!\left|\!\left|\mathbf{u}_h(P_H f) - \mathbf{u}^{\mathrm{ms}}_{H,h}(P_H f)\right|\!\right|\!\right| + \left|\!\left|\!\left|\mathbf{u}^{\mathrm{ms}}_{H,h}(P_H f) - \mathbf{u}^{\mathrm{ms}}_{H,h}(f)\right|\!\right|\!\right|.\end{aligned}$$

The two last terms will be shown to equal zero.

For the first term, we proceed in several steps. Let us define $\tilde{\mathbf{u}}_h := \mathbf{u}_h(f) - \mathbf{u}_h(P_H f) = \mathbf{u}_h(f - P_H f)$, which is the flux solution for the data $f - P_H f$. The corresponding pressure solution shall be denoted by $\tilde{p}_h$. First, we observe

$$\left|\!\left|\!\left|\tilde{\mathbf{u}}_h\right|\!\right|\!\right|^2 = (f - P_H f, \tilde{p}_h) = (f - P_H f, \tilde{p}_h - P_H\tilde{p}_h) \leq \|f - P_H f\|_{L^2(\Omega)}\|\tilde{p}_h - P_H\tilde{p}_h\|_{L^2(\Omega)}. \tag{9}$$

In order to bound the term $\|\tilde{p}_h - P_H\tilde{p}_h\|_{L^2(\Omega)}$, we let $\phi \in H^1_0(\Omega)$ be the weak solution to $\Delta\phi = \tilde{p}_h - P_H\tilde{p}_h$. Then we have

$$|\phi|^2_{H^1(\Omega)} = (\tilde{p}_h - P_H\tilde{p}_h, \phi - P_H\phi) \leq C_{\rho,d} H \|\tilde{p}_h - P_H\tilde{p}_h\|_{L^2(\Omega)}|\phi|_{H^1(\Omega)}.$$

Defining $\mathbf{w} := \nabla\phi$ we get $\nabla\cdot\mathbf{w} = \tilde{p}_h - P_H\tilde{p}_h$ and $\|\mathbf{w}\|_{L^2(\Omega)} \leq C_{\rho,d} H\|\tilde{p}_h - P_H\tilde{p}_h\|_{L^2(\Omega)}$. Next, we use a pair of projection operators $\widehat{\Pi}_h : V \to V_h$ and $\widehat{P}_h : Q \to Q_h$ that commute with

respect to the divergence operator, allows for $\mathcal{T}_h$ to be non quasi-uniform, and where $\widehat{\Pi}_h$ is L^2-stable, i.e. $\widehat{P}_h \nabla\cdot \mathbf{w} = \nabla\cdot\widehat{\Pi}_h\mathbf{w}$ and $\|\widehat{\Pi}_h\mathbf{w}\|_{L^2(\Omega)} \leq C_{\widehat{\Pi}}\|\mathbf{w}\|_{L^2(\Omega)}$, with $C_{\widehat{\Pi}}$ independent of h. The existence of such operators is proved in [9]. Exploiting this stability and the fact that $\tilde{p}_h - P_H\tilde{p}_h = \widehat{P}_h(\nabla\cdot\mathbf{w})$ (since $\widehat{P}_h$ is a projection on Q_h and $\tilde{p}_h - P_H\tilde{p}_h \in Q_h$), we obtain

$$\begin{aligned}\|\tilde{p}_h - P_H\tilde{p}_h\|^2_{L^2(\Omega)} &= (\tilde{p}_h - P_H\tilde{p}_h, \tilde{p}_h) = (\widehat{P}_h(\nabla\cdot\mathbf{w}), \tilde{p}_h)\\ &= (\nabla\cdot\widehat{\Pi}_h\mathbf{w}, \tilde{p}_h) = -(\mathbf{A}^{-1}\tilde{\mathbf{u}}_h, \widehat{\Pi}_h\mathbf{w}) \leq |||\tilde{\mathbf{u}}_h|||\,\|\mathbf{A}^{-1/2}\widehat{\Pi}_h\mathbf{w}\|_{L^2(\Omega)}\\ &\leq \beta^{1/2}C_{\widehat{\Pi}}|||\tilde{\mathbf{u}}_h|||\,\|\mathbf{w}\|_{L^2(\Omega)} \leq \beta^{1/2}C_{\widehat{\Pi}}C_{\rho,d}H|||\tilde{\mathbf{u}}_h|||\,\|\tilde{p}_h - P_H\tilde{p}_h\|_{L^2(\Omega)}.\end{aligned}$$

Combining this estimate with (9) yields

$$|||\tilde{\mathbf{u}}_h|||^2 \leq \beta^{1/2}C_{\widehat{\Pi}}C_{\rho,d}H\|f - P_Hf\|_{L^2(\Omega)}|||\tilde{\mathbf{u}}_h|||.$$

For the second term, since the correctors are divergence free, we have $\nabla\cdot\mathbf{u}^{\mathrm{ms}}_{H,h}(P_Hf) \in Q_H$. This implies $\nabla\cdot\mathbf{u}^{\mathrm{ms}}_{H,h}(P_Hf) = -P_Hf$, hence

$$\nabla\cdot\mathbf{u}^{\mathrm{ms}}_{H,h}(P_Hf) - \nabla\cdot\mathbf{u}_h(P_Hf) = 0,$$

i.e. $\mathbf{u}^{\mathrm{ms}}_{H,h}(P_Hf) - \mathbf{u}_h(P_Hf) \in K_h$. Now, from first the equations in (2) and (7) in combination with the $a(\cdot,\cdot)$-orthogonality between $V^{\mathrm{ms}}_{H,h}$ and K^{f}_h, we get

$$\begin{aligned} a(\mathbf{u}_h(P_Hf), \mathbf{v}) &= 0, \quad && \mathbf{v}\in V_h, && \nabla\cdot\mathbf{v} = 0, \text{ and}\\ a(\mathbf{u}^{\mathrm{ms}}_{H,h}(P_Hf), \mathbf{v}) &= 0, && \mathbf{v}\in V^{\mathrm{ms}}_{H,h}, && \nabla\cdot\mathbf{v} = 0, \text{ and}\\ a(\mathbf{u}^{\mathrm{ms}}_{H,h}(P_Hf), \mathbf{v}) &= 0, && \mathbf{v}\in V^{\mathrm{f}}_h, && \nabla\cdot\mathbf{v} = 0.\end{aligned}$$

Since $V_h = V^{\mathrm{ms}}_{H,h} \oplus V^{\mathrm{f}}_h$, we obtain

$$a(\mathbf{u}_h(P_Hf) - \mathbf{u}^{\mathrm{ms}}_{H,h}(P_Hf), \mathbf{v}) = 0,$$

for all $\mathbf{v}\in K_h$. Choosing $\mathbf{v} = \mathbf{u}_h(P_Hf) - \mathbf{u}^{\mathrm{ms}}_{H,h}(P_Hf)$, we see that $\mathbf{u}_h(P_Hf) = \mathbf{u}^{\mathrm{ms}}_{H,h}(P_Hf)$, thus the second term equals zero.

To show that the third term is zero, it is sufficient to show that $\mathbf{u}^{\mathrm{ms}}_{H,h}(f - P_Hf) = 0$. The data $f - P_Hf$ is L^2-orthogonal to the test space Q_H and it enters the equation (7) only in an L^2 scalar product with test functions. Hence $\mathbf{u}^{\mathrm{ms}}_{H,h}(f - P_Hf) = \mathbf{u}^{\mathrm{ms}}_{H,h}(P_H(f - P_Hf)) = 0$. □

4 Localized multiscale method

The ideal corrector problems (5) are at least as expensive to solve as the original reference problem. Hence, we require to localize these problems to very small patches, without sacrificing the good approximation properties. If we can achieve this, the corrector problems can be solved with low computational costs and fully in parallel. In this section, we show that this is indeed possible. We prove that we can truncate the computational domain Ω in the local corrector problems (5) to a small environment of a coarse element T. This is possible, since the solutions of (5) decay with exponential rate outside the coarse element T. We obtain a new localized corrector operator which can be used analogously to

the ideal corrector operator to construct a localized multiscale space. This localization reduces the computational effort for assembling the multiscale space significantly.

In addition to the assumptions (A1)–(A3) and (B1)–(B3), we require additional assumptions on the computational domain and the mesh. More precisely we assume the following.

(A4) The domain Ω is simply-connected.

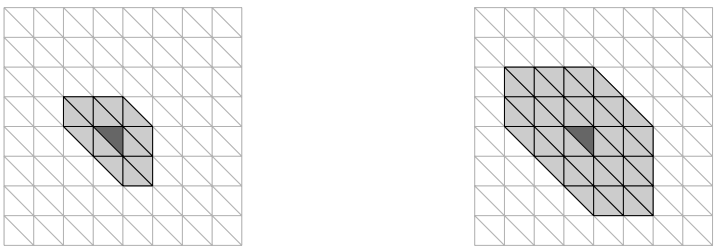
(B4) The fine grid $\mathcal{T}_h$ is quasi-uniform, i.e. the ratio between the maximum and the minimum diameter of a grid element is bounded by a generic constant.

We note that assumption (A4) is crucial for our proof. Assumption (B4) on the other hand could be dropped with a more careful analysis. In this case the estimates (and in particular the decay) will depend on the inverse of the minimum mesh size of the fine grid in a patch $U(T)$. For simplicity of the presentation, we do not elaborate this case and restrict ourselves to quasi-uniform meshes, i.e. to (B4).

In order to localize the detail space K_h^{f}, we use admissible patches. We call this restriction to patches *localization*. For each $T \in \mathcal{T}_H$ we pick a connected patch $U(T)$ consisting of coarse grid elements and containing T. More precisely, for positive $k \in \mathbb{N}$ we define k-coarse-layer patches iteratively in the following way. For all $T \in \mathcal{T}_H$ (which are assumed to be closed sets), we define the element patch $U_k(T)$ in the coarse mesh $\mathcal{T}_H$ by

$$\begin{aligned} U_0(T) &:= T, \\ U_k(T) &:= \bigcup \{T' \in \mathcal{T}_H : T' \cap U_{k-1}(T) \neq \emptyset\} \quad k = 1, 2, \ldots. \end{aligned} \tag{10}$$

See Figure 1 for an illustration of patches. For a given patch $U(T)$, we define the



(a) One-coarse-layer patch, $k = 1$. (b) Two-coarse-layer patch, $k = 2$.

Figure 1: Illustration of k-coarse-layer patches. Dark gray subdomain is T. Light gray subdomain is $U_k(T)$.

restriction of V_h^{f} to $U(T)$ by

$$V_h^{\mathrm{f}}(U(T)) := \{\mathbf{w} \in V_h^{\mathrm{f}} : \mathbf{w} = 0 \text{ in } \Omega \setminus U(T)\}.$$

Accordingly, we also define

$$K_h^{\mathrm{f}}(U(T)) := \{\mathbf{w} \in V_h^{\mathrm{f}}(U(T)) : \nabla \cdot \mathbf{w} = 0\}.$$

Using this localized space, we define the localized corrector operators. Localized quantities are indexed by the patch layer size k.

Definition 10 (Localized corrector operators)**.** *For each $T \in \mathcal{T}_H$ and $k \geq 1$ layers, we define a* localized element corrector operator $G^T_{h,k} : V \to K^{\mathrm{f}}_h(U_k(T))$*:*

$$a(G^T_{h,k}\mathbf{v}, \mathbf{w}) = a^T(\mathbf{v}, \mathbf{w}) \tag{11}$$

for all $\mathbf{w} \in K^{\mathrm{f}}_h(U_k(T))$. Further, we define the localized global corrector operator $G_{h,k} := \sum_{T\in\mathcal{T}_H} G^T_{h,k}$.

The localized corrector operators are again well-defined by the Lax-Milgram theorem, exploiting that $a(\cdot,\cdot)$ is a weighted L^2-scalar product. Note that the definition of $K^{\mathrm{f}}_h(U_k(T))$ implies Neumann boundary conditions on the localized corrector problems (11). We define a localized multiscale function space by

$$V^{\mathrm{ms},k}_{H,h} := (\mathrm{Id} - G_{h,k})(V_H)$$

and state the localized multiscale problem as follows.

Definition 11 (Localized multiscale problem)**.** *The localized multiscale problem reads: find $\mathbf{u}^{\mathrm{ms},k}_{H,h} \in V^{\mathrm{ms},k}_{H,h}$ and $p_H \in Q_H$, such that*

$$\begin{aligned} a(\mathbf{u}^{\mathrm{ms},k}_{H,h}, \mathbf{v}_h) + b(\mathbf{v}_h, p_H) &= 0, \\ b(\mathbf{u}^{\mathrm{ms},k}_{H,h}, q_H) &= -(f, q_H), \end{aligned} \tag{12}$$

for all $\mathbf{v}_h \in V^{\mathrm{ms},k}_{H,h}$ and $q_H \in Q_H$.

Definitions 10 and 11 constitute the proposed multiscale method. Next, we show that the above stated problem is well-posed.

Lemma 12 (Unique solution of localized multiscale problem)**.** *Under Assumptions (A1)–(A3) and (B1)–(B3), the localized multiscale problem* (12) *has a unique solution for all k, h and H.*

Proof. We use similar arguments as in Lemma 8. The basic difference is that we need to show stability for the localized corrector operator $G_{h,k}$. We start with the stability of the localized element corrector operators. Here we have for arbitrary $\mathbf{v} \in V$

$$\left|\!\left|\!\left| G^T_{h,k}\mathbf{v} \right|\!\right|\!\right|^2 = a(G^T_{h,k}\mathbf{v}, G^T_{h,k}\mathbf{v}) = a^T(\mathbf{v}, G^T_{h,k}\mathbf{v}) \leq |\!|\!|\mathbf{v}|\!|\!|_T \left|\!\left|\!\left| G^T_{h,k}\mathbf{v} \right|\!\right|\!\right|. \tag{13}$$

Now, we can prove L^2-stability of the localized global operator. We get

$$\begin{aligned} \|G_{h,k}\mathbf{v}\|^2_{L^2(\Omega)} &= \left\| \sum_{T\in\mathcal{T}_H} G^T_{h,k}\mathbf{v} \right\|^2_{L^2(\Omega)} \leq \alpha^{-1} a\left(\sum_{T\in\mathcal{T}_H} G^T_{h,k}\mathbf{v}, \sum_{T'\in\mathcal{T}_H} G^{T'}_{h,k}\mathbf{v} \right) \\ &= \alpha^{-1} \sum_{T\in\mathcal{T}_H} \sum_{T'\subset U_k(T)} a(G^T_{h,k}\mathbf{v}, G^{T'}_{h,k}\mathbf{v}) \\ &\leq \frac{1}{2}\alpha^{-1} \sum_{T\in\mathcal{T}_H} \sum_{T'\subset U_k(T)} \left(\left|\!\left|\!\left| G^T_{h,k}\mathbf{v} \right|\!\right|\!\right|^2 + \left|\!\left|\!\left| G^{T'}_{h,k}\mathbf{v} \right|\!\right|\!\right|^2 \right) \\ &\leq \alpha^{-1} C_\rho k^d \sum_{T\in\mathcal{T}_H} \left|\!\left|\!\left| G^T_{h,k}\mathbf{v} \right|\!\right|\!\right|^2 \overset{(13)}{\leq} \alpha^{-1} C_\rho k^d \sum_{T\in\mathcal{T}_H} |\!|\!|\mathbf{v}|\!|\!|^2_T \leq \alpha^{-1}\beta C_\rho k^d \|\mathbf{v}\|^2_{L^2(\Omega)}, \end{aligned}$$

where C_ρ is a constant only depending on the shape regularity constant ρ of the coarse mesh. Similar to (8) we derive inf-sup stability with

$$\begin{aligned}\gamma &\leq \inf_{q\in Q_H}\sup_{\mathbf{v}\in V_H}\frac{b(\mathbf{v},q)}{\|q\|_Q\,\|\mathbf{v}\|_V}\\ &\leq (1+\alpha^{-1/2}\beta^{1/2}C_\rho^{1/2}k^{d/2})\inf_{q\in Q_H}\sup_{\mathbf{v}\in V_{H,h}^{\mathrm{ms},k}}\frac{b(\mathbf{v},q)}{\|q\|_Q\,\|\mathbf{v}\|_V}.\end{aligned}$$

Observe that the inf-sup stability constant $\gamma_k^0 := \gamma(1+\alpha^{-1/2}\beta^{1/2}C_\rho^{1/2}k^{d/2})^{-1}$ depends on k this time. □

The inf-sup stability constant γ_k^0 depends on k due to overlapping patches. We come back to another estimate of the inf-sup stability constant in Section 4.3 after proving the decay of the correctors.

It is important to note that in the localized case we do not have orthogonality between $V_{H,h}^{\mathrm{ms},k}$ and K_h^{f} as in the ideal case (cf. equation (6)). This orthogonality was crucial in the error estimate for the ideal method presented in Lemma 9. In the localized case, we rely on the exponential decay of the localized element correctors, which justifies localization to patches.

4.1 Error estimate for localized problem

In this section we state the main result of this paper, which is an a priori error estimate in the energy norm between the reference solution and the localized multiscale approximation. We first present a logarithmic stability result for the nodal Raviart–Thomas interpolation operator Π_H for fine scale functions and then state a lemma on the exponential decay of the correctors. Then the main theorem follows. The proof of the exponential decay is contained in Section 4.2. The notation $a \lesssim b$ stands for $a \leq Cb$ with some constant C that might depend on d, Ω, α, β and coarse and fine mesh regularity constants, but not on the mesh sizes h and H. In particular it does not depend on the possibly rapid oscillations in $\mathbf{A}$.

We recall a well known stability result for the nodal Raviart–Thomas interpolation operator.

Lemma 13 (Logarithmic stability of the nodal interpolation operator for divergence free functions)**.** *Assume (B1)–(B4). For any given element $T \in \mathcal{T}_H$ there exists a constant C that only depends on the regularity of T and the quasi-uniformity of $\mathcal{T}_h$, such that*

$$\|\Pi_H \mathbf{v}_h\|^2_{L^2(T)} \leq C\lambda(H/h)^2\|\mathbf{v}_h\|^2_{L^2(T)},$$

with $\lambda(H/h) := (1+\log(H/h))^{1/2}$ for all $\mathbf{v}_h \in V_h$ with $\nabla\cdot\mathbf{v}_h = 0$.

A proof for this can be found in [32, Lemma 4.1]. This result holds for both $d = 2$ and 3.

Remark 14. *There exist unconditionally L^2-stable Clément-type interpolation operators for which we could define $\lambda(H/h) := 1$ for all h and H instead, see [5, 8, 9, 30]. In particular, the operators introduced in [5, 9] are projections and were used as a technical tool in the proof of Lemma 9 above. However, these operators are hard to implement in practice and hence are not used in the proposed numerical method.*

Lemma 15 (Exponential decay of correctors)**.** *Under Assumptions (A1)–(A4) and (B1)–(B4), there exists a generic constant $0 < \theta < 1$ depending on the contrast β/α, but not on h or H such that for all positive $k \in \mathbb{N}$:*

$$\left|\!\left|\!\left| \sum_{T\in\mathcal{T}_H} \left(G_h^T\mathbf{v} - G_{h,k}^T\mathbf{v}\right) \right|\!\right|\!\right|^2 \lesssim k^d \lambda(H/h)^2 \theta^{2k/\lambda(H/h)} \sum_{T\in\mathcal{T}_H} \left|\!\left|\!\left| G_h^T\mathbf{v} \right|\!\right|\!\right|^2 \tag{14}$$

for all $\mathbf{v} \in V$.

Proof. The lemma is a direct consequence of Lemma 21 in Section 4.2. □

Now, combining the error estimate for the ideal multiscale method in Lemma 9 and Lemma 15 we get the following a priori error estimate of the localized multiscale method.

Theorem 16 (Error estimate for localized multiscale solution)**.** *Under Assumptions (A1)–(A4) and (B1)–(B4), for a positive $k \in \mathbb{N}$, let $\mathbf{u}_h$ solve (2) and $\mathbf{u}_{H,h}^{\mathrm{ms},k}$ solve (12), then*

$$\left|\!\left|\!\left| \mathbf{u}_h - \mathbf{u}_{H,h}^{\mathrm{ms},k} \right|\!\right|\!\right| \lesssim H\|f - P_H f\|_{L^2(\Omega)} + k^{d/2}\lambda(H/h)^2\theta^{k/\lambda(H/h)}\|f\|_{L^2(\Omega)}, \tag{15}$$

for some $0 < \theta < 1$ depending on the contrast β/α, but not on k, h and H.

Before stating the proof, we discuss the role and choice of k. The second term in the error estimate (15) is an effect of the localization. This term can be made small by choosing large values of k, i.e. large patch sizes. A natural question is how to choose k to make the second term of order H to some power.

We write $\lambda = \lambda(H/h)$ for convenience. Let $\tilde{k} = 2d^{-1}\log(\theta)\lambda^{-1}k = -C_\theta\lambda^{-1}k$, where $C_\theta = -2d^{-1}\log(\theta) > 0$ is a constant independent of H and h. We are interested in the asymptotic behavior, so we consider $H \ll 1$. Setting the second term in (15) equal to $H\|f\|_{L^2(\Omega)}$ yields

$$\tilde{k}e^{\tilde{k}} = -C_\theta\lambda^{-4/d-1}H^{2/d},$$

that is $\tilde{k} = W(-C_\theta\lambda^{-4/d-1}H^{2/d})$, where W is the Lambert W-function. In terms of the number of layers k, we get $k = -C_\theta^{-1}\lambda W(-C_\theta\lambda^{-4/d-1}H^{2/d})$. This equation has two solutions for sufficiently small H. Since we require $k \geq 1$, we pick the branch $W \leq -1$.

Another, more practical option is to choose $k = R\lambda\log(1/H)$ for some constant R. Then the expression $k^{d/2}\lambda\theta^{k/\lambda}$ will be asymptotically (as $H \to 0$) dominated by the power $H^{-R\log\theta}$. Choosing R sufficiently large yields arbitrary order of accuracy of the term. The fine mesh size h is often fixed and we can choose

$$k = (1 + |\log_r(H)|)^{1/2}\log_s(1/H) \tag{16}$$

for some bases r and s of the two logarithms.

Remark 17. *If Clément-type interpolation operators are used, we have $\lambda \equiv 1$ independent of H/h. Choosing $k = C\log(1/H)$ makes the second term in (15) proportional to $\log(1/H)^{d/2}H^{-C\log\theta}$. For an appropriate C we can make the first term in (15) dominate the error estimate.*

Proof of Theorem 16. Let $\tilde{\mathbf{u}}_{H,h}^{\mathrm{ms},k} := ((\mathrm{Id} - G_{h,k}) \circ \Pi_H)\mathbf{u}_{H,h}^{\mathrm{ms}} \in V_{H,h}^{\mathrm{ms},k}$, then $\tilde{\mathbf{u}}_{H,h}^{\mathrm{ms},k} - \mathbf{u}_{H,h}^{\mathrm{ms},k}$ is divergence free. Hence, by Galerkin orthogonality we have

$$a(\mathbf{u}_h - \mathbf{u}_{H,h}^{\mathrm{ms},k}, \mathbf{u}_h - \mathbf{u}_{H,h}^{\mathrm{ms},k}) = a(\mathbf{u}_h - \mathbf{u}_{H,h}^{\mathrm{ms},k}, \mathbf{u}_h - \tilde{\mathbf{u}}_{H,h}^{\mathrm{ms},k})$$

and obtain

$$\left|\!\left|\!\left| \mathbf{u}_h - \mathbf{u}_{H,h}^{\mathrm{ms},k} \right|\!\right|\!\right| \leq \left|\!\left|\!\left| \mathbf{u}_h - \tilde{\mathbf{u}}_{H,h}^{\mathrm{ms},k} \right|\!\right|\!\right| \leq \left|\!\left|\!\left| \mathbf{u}_h - \mathbf{u}_{H,h}^{\mathrm{ms}} \right|\!\right|\!\right| + \left|\!\left|\!\left| \mathbf{u}_{H,h}^{\mathrm{ms}} - \tilde{\mathbf{u}}_{H,h}^{\mathrm{ms},k} \right|\!\right|\!\right|.$$

The first term can be bounded by $\beta^{1/2} C_{\widehat{\mathrm{I}}} C_{\rho,d} H \|f - P_H f\|_{L^2(\Omega)}$ by Lemma 9. Regarding the second term, using [32, Lemma 4.1] and stability of the ideal multiscale solution, we get

$$\sum_{T \in \mathcal{T}_H} \left|\!\left|\!\left| G_h^T \Pi_H \mathbf{u}_{H,h}^{\mathrm{ms}} \right|\!\right|\!\right|^2 \leq \sum_{T \in \mathcal{T}_H} \left|\!\left|\!\left| \Pi_H \mathbf{u}_{H,h}^{\mathrm{ms}} \right|\!\right|\!\right|_T^2 = \left|\!\left|\!\left| \Pi_H \mathbf{u}_{H,h}^{\mathrm{ms}} \right|\!\right|\!\right|^2 \lesssim \lambda(H/h)^2 \|f\|_{L^2(\Omega)}^2$$

and can combine this with Lemma 15 to get

$$\begin{aligned}
\left|\!\left|\!\left| \mathbf{u}_{H,h}^{\mathrm{ms}} - \tilde{\mathbf{u}}_{H,h}^{\mathrm{ms},k} \right|\!\right|\!\right| &= \left|\!\left|\!\left| (G_{h,k} - G_h) \Pi_H \mathbf{u}_{H,h}^{\mathrm{ms}} \right|\!\right|\!\right| \\
&= \left|\!\left|\!\left| \sum_{T \in \mathcal{T}_H} (G_{h,k}^T - G_h^T) \Pi_H \mathbf{u}_{H,h}^{\mathrm{ms}} \right|\!\right|\!\right| \\
&\lesssim k^{d/2} \lambda(H/h) \theta^{k/\lambda(H/h)} \left(\sum_{T \in \mathcal{T}_H} \left|\!\left|\!\left| G_h^T \Pi_H \mathbf{u}_{H,h}^{\mathrm{ms}} \right|\!\right|\!\right|^2 \right)^{1/2} \\
&\lesssim k^{d/2} \lambda(H/h)^2 \theta^{k/\lambda(H/h)} \|f\|_{L^2(\Omega)}.
\end{aligned}$$

□

4.2 Proof of exponential decay of correctors

This section consists of four lemmas, Lemma 18–21, of which the last one is the main result. The two first lemmas are auxiliary and are motivated by steps in the proofs of the latter two. Before starting, we need to set some notation and introduce some tools. We use the notation $W^{1,2}_{\mathrm{loc}}(\mathbb{R}^d) = \{f : f \in H^1(\omega)\ \forall \text{ compact subsets } \omega \subset \mathbb{R}^d\}$. Note that we will use the letter K to denote arbitrary triangles of the coarse mesh $\mathcal{T}_H$. The first lemma says that every divergence free function $\mathbf{w}$ in $H(\mathrm{div}, \Omega)$ is the divergence of a skew-symmetric matrix.

Lemma 18. *Let Ω be a simply connected domain with Lipschitz boundary and let $\mathbf{w} \in H(\mathrm{div}, \Omega)$ with $\nabla \cdot \mathbf{w} = 0$ in Ω. Then there exists a skew-symmetric matrix $\psi \in [W^{1,2}_{\mathrm{loc}}(\mathbb{R}^d)]^{d \times d}$ with $\nabla \psi_{ij} \in [L^2(\mathbb{R}^d)]^d$ and $\int_\Omega \psi = 0$ such that*

$$\mathbf{w} = \nabla \cdot \psi \quad \text{in } \Omega \qquad \text{and} \qquad \|\nabla \psi_{ij}\|_{L^2(\omega)} \lesssim \|\mathbf{w}\|_{L^2(\omega)} \qquad \text{for } \omega \subset \Omega. \tag{17}$$

Here, the divergence of ψ is defined along the rows.

Proof. The result is a combination of well-known results. First, we extend the divergence-free vector field $\mathbf{w} \in H(\mathrm{div}, \Omega)$ to a divergence-free vector field $\tilde{\mathbf{w}} \in H(\mathrm{div}, \mathbb{R}^d)$. In particular we have $\tilde{\mathbf{w}} \in [L^2(\mathbb{R}^d)]^d$ and $\tilde{\mathbf{w}} = \mathbf{w}$ in Ω. Note that the extension of $\mathbf{w}$ to $\mathbb{R}^d$ will be typically not zero outside of Ω. The existence of such an extension operator was proved in [31, Proposition 3.8]. It is well known that there exists a skew-symmetric matrix $\psi \in [W^{1,2}_{\mathrm{loc}}(\mathbb{R}^d)]^{d\times d}$ with $\nabla\psi_{ij} \in [L^2(\mathbb{R}^d)]^d$, such that $\tilde{\mathbf{w}} = \nabla \cdot \psi$ (see [21, Lemma 2.3]). The matrix is only unique up to a constant, so we fix the constant by $\int_\Omega \psi = 0$ (which gives us a Poincaré inequality). The inequality $\|\nabla\psi_{ij}\|_{L^2(\omega)} \lesssim \|\tilde{\mathbf{w}}\|_{L^2(\omega)}$ (for $\omega \subset \mathbb{R}^d$) can be extracted from the proof given for [21, Lemma 2.3] and is based on the observation that it holds $\nabla\psi_{ij} = \nabla\Delta^{-1}(\partial_j\tilde{\mathbf{w}}_i - \partial_i\tilde{\mathbf{w}}_j)$. Lets consider the case $d = 2$. Obviously, if $i = j$ we obtain $\nabla\psi_{ii} = \nabla\psi_{jj} = 0$ and estimate (17) is trivial. If $i \neq j$, we obtain by using the skew-symmetry

$$\begin{aligned}\|\mathbf{w}\|^2_{L^2(\omega)} &= \|\nabla\cdot\psi\|^2_{L^2(\omega)} = \|\partial_1\psi_{11} + \partial_2\psi_{12}\|^2_{L^2(\omega)} + \|\partial_1\psi_{21} + \partial_2\psi_{22}\|^2_{L^2(\omega)}\\ &= \|\partial_2\psi_{12}\|^2_{L^2(\omega)} + \|\partial_1\psi_{21}\|^2_{L^2(\omega)} = \|\partial_2\psi_{12}\|^2_{L^2(\omega)} + \|\partial_1\psi_{12}\|^2_{L^2(\omega)}\\ &= \|\nabla\psi_{12}\|^2_{L^2(\omega)} = \|\nabla\psi_{21}\|^2_{L^2(\omega)},\end{aligned}$$

i.e. we obtain even equality in estimate (17). □

We also require suitable cut-off functions that are central for the proof. For $T \in \mathcal{T}_H$ and positive $k \in \mathbb{N}$, we let the function $\eta_{T,k} \in P_1(\mathcal{T}_H)$ (globally continuous and piecewise linear w.r.t. $\mathcal{T}_H$) be defined as

$$\begin{aligned}&\eta_{T,k}(x) = 0 \quad \text{for } x \in U_{k-1}(T),\\ &\eta_{T,k}(x) = 1 \quad \text{for } x \in \Omega \setminus U_k(T).\end{aligned} \tag{18}$$

We start with the following lemma, which enables us to approximate truncated functions from K^{f}_h.

Lemma 19. *Let $\mathbf{w}_h \in K^{\mathrm{f}}_h$ and let $\psi \in [W^{1,2}_{\mathrm{loc}}(\Omega)]^{d\times d}$ with $\mathbf{w}_h = \nabla\cdot\psi$ denote the corresponding skew-symmetric matrix as in Lemma 18. Let furthermore $\psi_K := |K|^{-1}\int_K \psi$ denote the average on $K \in \mathcal{T}_H$ and let $\psi_H \in [L^2(\Omega)]^{d\times d}$ denote the corresponding piecewise constant matrix with $\psi_H(x) = \psi_K$ for $x \in K$. The broken divergence-operator $\nabla_H \cdot$ is given by $\nabla_H \cdot v := \nabla\cdot v|_K$ for $K \in \mathcal{T}_H$. The function $\eta_{T,k} \in P_1(\mathcal{T}_H)$ is a given cut-off function as defined in (18) for $k > 0$. Then, we have that the function $\tilde{\mathbf{w}}_h := \Pi_h(\nabla\cdot(\eta_{T,k}\psi)) - (\Pi_H \circ \Pi_h)(\nabla\cdot(\eta_{T,k}\psi)) \in K^{\mathrm{f}}_h$ fulfills the following estimate for any $K \in \mathcal{T}_H$:*

$$\|\nabla\cdot(\eta_{T,k}\psi) - \nabla_H\cdot(\eta_{T,k}\psi_H) - \tilde{\mathbf{w}}_h\|_{L^2(K)} \lesssim \begin{cases}\lambda(H/h)\|\mathbf{w}_h\|_{L^2(K)} & K \subset U_k(T)\setminus U_{k-1}(T)\\ 0 & \text{otherwise.}\end{cases}$$

Obviously we also have $\mathrm{supp}(\tilde{\mathbf{w}}_h) \subset \Omega \setminus U_{k-1}(T)$.

Proof. First, we observe that the skew-symmetric matrix ψ must be a polynomial of maximum degree 2 on each fine grid element. We use this in the following without mentioning.

We fix the element $T \in \mathcal{T}_H$ and $k \in \mathbb{N}$ and denote $\eta := \eta_{T,k}$. Furthermore, we define for $K \in \mathcal{T}_H$

$$c_K := |K|^{-1} \int_K \eta \qquad \text{and} \qquad \psi_K := |K|^{-1} \int_K \psi.$$

We define $\tilde{\mathbf{w}}_h := \Pi_h (\nabla \cdot (\eta\psi)) - (\Pi_H \circ \Pi_h)(\nabla \cdot (\eta\psi))$ and observe that $\tilde{\mathbf{w}}_h \in K_h^{\mathrm{f}}$ and $\mathbf{w}_h = \tilde{\mathbf{w}}_h$ on $\Omega \backslash U_k(T)$. The property $\Pi_H(\tilde{\mathbf{w}}_h) = 0$ is clear. The property $\nabla \cdot \tilde{\mathbf{w}}_h = 0$ follows from the fact that $\eta\psi$ is still skew symmetric and that $\nabla \cdot (\Pi_H \circ \Pi_h)(\cdot) = (P_H \circ P_h)(\nabla \cdot)$. Since ψ_K and c_K are constant on K we have

$$\Pi_h (\nabla \cdot (c_K \psi_K)) = \nabla \cdot (c_K \psi_K) = 0 \quad \text{on } K. \tag{19}$$

Furthermore, since $\Pi_H(\mathbf{v}_H) = \mathbf{v}_H$ for all $\mathbf{v}_H \in V_H$ and since $\nabla \cdot (\eta\psi_K) \in V_H$ we also have

$$(\Pi_H \circ \Pi_h)(\nabla \cdot (\eta\psi_K)) = \Pi_h (\nabla \cdot (\eta\psi_K)) \quad \text{on } K. \tag{20}$$

Finally, we also have on K,

$$(\Pi_H \circ \Pi_h)(\nabla \cdot (c_K\psi)) = c_K (\Pi_H \circ \Pi_h)(\nabla \cdot \psi) = c_K \Pi_H(\mathbf{w}_h) = 0. \tag{21}$$

Combining (19), (20), and (21) we obtain for every $K \in \mathcal{T}_H$

$$\begin{aligned}
&\|(\Pi_H \circ \Pi_h)(\nabla \cdot (\eta\psi)) - \Pi_h(\nabla \cdot (\eta\psi_K))\|_{L^2(K)} \\
&\quad = \|(\Pi_H \circ \Pi_h)(\nabla \cdot (\eta\psi) - \nabla \cdot (c_K\psi) - \nabla \cdot (\eta\psi_K) + \nabla \cdot (c_K\psi_K))\|_{L^2(K)} \\
&\quad = \|(\Pi_H \circ \Pi_h)(\nabla \cdot ((\eta - c_K)(\psi - \psi_K)))\|_{L^2(K)}.
\end{aligned} \tag{22}$$

Now, we consider the quantity we want to estimate. For any $K \in \mathcal{T}_H$,

$$\begin{aligned}
&\|\nabla \cdot (\eta\psi) - \nabla_H \cdot (\eta\psi_H) - \tilde{\mathbf{w}}_h\|_{L^2(K)} \\
&\quad \le \|\nabla \cdot (\eta(\psi - \psi_K)) - \Pi_h (\nabla \cdot (\eta(\psi - \psi_K)))\|_{L^2(K)} \\
&\qquad + \|\Pi_h (\nabla \cdot (\eta(\psi - \psi_K))) - \Pi_h (\nabla \cdot (\eta\psi)) + (\Pi_H \circ \Pi_h)(\nabla \cdot (\eta\psi))\|_{L^2(K)} \\
&\quad = \|\nabla \cdot (\eta(\psi - \psi_K)) - \Pi_h (\nabla \cdot (\eta(\psi - \psi_K)))\|_{L^2(K)} \\
&\qquad + \|(\Pi_H \circ \Pi_h)(\nabla \cdot (\eta\psi)) - \Pi_h(\nabla \cdot (\eta\psi_K))\|_{L^2(K)} \\
&\quad \overset{(22)}{=} \|\nabla \cdot ((\eta - c_K)(\psi - \psi_K)) - \Pi_h (\nabla \cdot ((\eta - c_K)(\psi - \psi_K)))\|_{L^2(K)} \\
&\qquad + \|(\Pi_H \circ \Pi_h)(\nabla \cdot ((\eta - c_K)(\psi - \psi_K)))\|_{L^2(K)} \\
&\quad \lesssim \lambda(H/h) \|\nabla \cdot ((\eta - c_K)(\psi - \psi_K))\|_{L^2(K)}.
\end{aligned} \tag{23}$$

In the last step we used Lemma 13, the property that $\Pi_h \nabla \cdot ((\eta - c_K)(\psi - \psi_K))$ is divergence free and the fact that Π_h is locally L^2-stable when applied to functions of small fixed polynomial degree, i.e. for fixed $t \in \mathcal{T}_h$ and $r \in \mathbb{N}$ there exists a constant $C(r)$ that only depends on r and the shape regularity of t such that

$$\|\Pi_h(\mathbf{v})\|_{L^2(t)} \le C(r) \|\mathbf{v}\|_{L^2(t)} \qquad \text{for all } \mathbf{v} \in [\mathbb{P}^r(t)]^d.$$

Continuing from (23) we obtain

$$\begin{aligned}
&\|\nabla\cdot((\eta-c_K)(\psi-\psi_K)\|^2_{L^2(K)}\\
&\quad\lesssim\ \|(\eta-c_K)\nabla\cdot\psi\|^2_{L^2(K)}+\|(\psi-\psi_K)\nabla\eta\|^2_{L^2(K)}\\
&\quad\lesssim\ H^2\|\nabla\eta\|^2_{L^\infty(K)}\|\nabla\psi\|^2_{L^2(K)}\\
&\quad\overset{(17)}{\lesssim}\begin{cases}\|\mathbf{w}\|^2_{L^2(K)} & K\subset U_k(T)\setminus U_{k-1}(T)\\ 0 & \text{otherwise.}\end{cases}
\end{aligned}\tag{24}$$

Note that we used the properties of η to obtain the Lipschitz bound $\|\eta-c_K\|_{L^\infty(K)}\lesssim H\|\nabla\eta\|_{L^\infty(K)}\lesssim 1$ and that $\nabla\eta$ has no support outside $U_k(T)\setminus U_{k-1}(T)$. We also used the Poincaré inequality for $\eta-c_K$ which has a zero average on K. Combining (23) and (24) yields the sought result. □

We continue with a lemma showing the exponential decay of solutions to problems of the form in (5).

Lemma 20. *Now, let $\mathbf{w}^T\in K_h^{\mathrm{f}}$ be the solution of*

$$\int_\Omega \mathbf{A}^{-1}\mathbf{w}^T\cdot\mathbf{v}_h=F_T(\mathbf{v}_h)\qquad\text{for all }\mathbf{v}_h\in K_h^{\mathrm{f}}\tag{25}$$

where $F_T\in(K_h^{\mathrm{f}})'$ is such that $F_T(\mathbf{v}_h)=0$ for all $\mathbf{v}_h\in K_h^{\mathrm{f}}(\Omega\setminus T)$. Then, there exists a generic constant $0<\theta<1$ (depending on the contrast β/α) such that for all positive $k\in\mathbb{N}$:

$$\left|\!\left|\!\left|\mathbf{w}^T\right|\!\right|\!\right|_{\Omega\setminus U_k(T)}\lesssim\theta^{k/\lambda(H/h)}\left|\!\left|\!\left|\mathbf{w}^T\right|\!\right|\!\right|_\Omega.\tag{26}$$

Proof. The proof exploits similar arguments as in [28]. Let us fix $k\in\mathbb{N}$. We denote again $\eta:=\eta_{T,k}\in P_1(\mathcal{T}_H)$ (as in (18)). We apply Lemma 19 to $\mathbf{w}^T\in K_h^{\mathrm{f}}$. The corresponding skew symmetric matrix shall again be denoted by $\psi=\psi(\mathbf{w}^T)$ and we define

$$\tilde{\mathbf{w}}^T:=\Pi_h(\nabla\cdot(\eta\psi))-(\Pi_H\circ\Pi_h)(\nabla\cdot(\eta\psi)).$$

We obtain that $\nabla\cdot(\eta\psi)-\nabla_H\cdot(\eta\psi_H)-\tilde{\mathbf{w}}^T$ is zero outside $U_k(T)\setminus U_{k-1}(T)$ and

$$\|\nabla\cdot(\eta\psi)-\nabla_H\cdot(\eta\psi_H)-\tilde{\mathbf{w}}^T\|_{L^2(U_k(T)\setminus U_{k-1}(T))}\ \lesssim\ \lambda(H/h)\|\mathbf{w}^T\|_{L^2(U_k(T)\setminus U_{k-1}(T))}.\tag{27}$$

First observe that

$$\int_{\Omega\setminus U_{k-1}(T)}\mathbf{A}^{-1}\mathbf{w}^T\cdot\tilde{\mathbf{w}}^T=\int_\Omega\mathbf{A}^{-1}\mathbf{w}^T\cdot\tilde{\mathbf{w}}^T=F_T(\tilde{\mathbf{w}}^T)=0\tag{28}$$

and

$$\eta\mathbf{w}^T=\eta\nabla\cdot\psi=\nabla\cdot(\eta\psi)-\psi\nabla\eta.\tag{29}$$

With that we have

$$\begin{aligned}
\int_{\Omega\setminus U_k(T)} \mathbf{A}^{-1}\mathbf{w}^T\cdot\mathbf{w}^T &\leq \int_{\Omega\setminus U_{k-1}(T)} \mathbf{A}^{-1}\mathbf{w}^T\cdot(\eta\mathbf{w}^T)\\
&\overset{(29)}{=} \int_{\Omega\setminus U_{k-1}(T)} \mathbf{A}^{-1}\mathbf{w}^T\cdot(\nabla\cdot(\eta\psi)-\psi\nabla\eta)\\
&\overset{(28)}{=} \int_{\Omega\setminus U_{k-1}(T)} \mathbf{A}^{-1}\mathbf{w}^T\cdot\big(\nabla\cdot(\eta\psi)-\psi\nabla\eta-\tilde{\mathbf{w}}^T\big)\\
&= \underbrace{\int_{\Omega\setminus U_{k-1}(T)} \mathbf{A}^{-1}\mathbf{w}^T\cdot\big(\nabla\cdot(\eta\psi)-\nabla_H\cdot(\eta\psi_H)-\tilde{\mathbf{w}}^T\big)}_{=:\mathrm{I}}\\
&\quad+\underbrace{\int_{\Omega\setminus U_{k-1}(T)} \mathbf{A}^{-1}\mathbf{w}^T\cdot(\nabla_H\cdot(\eta\psi_H)-\psi\nabla\eta)}_{=:\mathrm{II}}.
\end{aligned}$$

For I we use (27) to obtain

$$\mathrm{I}\lesssim\lambda(H/h)\big|\!\big|\!\big|\mathbf{w}^T\big|\!\big|\!\big|^2_{U_k(T)\setminus U_{k-1}(T)}$$

and for II we obtain

$$\begin{aligned}
\mathrm{II} &= \int_{\Omega\setminus U_{k-1}(T)} \mathbf{A}^{-1}\mathbf{w}^T\cdot((\psi_H-\psi)\nabla\eta)\\
&\lesssim \sum_{\substack{K\in\mathcal{T}_H\\ K\subset U_k(T)\setminus U_{k-1}(T)}} \big|\!\big|\!\big|\mathbf{w}^T\big|\!\big|\!\big|_K H\|\nabla\eta\|_{L^\infty(K)}\|\nabla\psi\|_{L^2(K)}\\
&\lesssim \big|\!\big|\!\big|\mathbf{w}^T\big|\!\big|\!\big|^2_{U_k(T)\setminus U_{k-1}(T)}.
\end{aligned}$$

Now, denote by $L:=C\lambda(H/h)$, and we get

$$\big|\!\big|\!\big|\mathbf{w}^T\big|\!\big|\!\big|^2_{\Omega\setminus U_k(T)}\leq L\big|\!\big|\!\big|\mathbf{w}^T\big|\!\big|\!\big|^2_{U_k(T)\setminus U_{k-1}(T)}\leq L\left(\big|\!\big|\!\big|\mathbf{w}^T\big|\!\big|\!\big|^2_{\Omega\setminus U_{k-1}(T)}-\big|\!\big|\!\big|\mathbf{w}^T\big|\!\big|\!\big|^2_{\Omega\setminus U_k(T)}\right)$$

where C is independent of T, k and $\mathbf{A}$, but can depend on the contrast. We obtain

$$\big|\!\big|\!\big|\mathbf{w}^T\big|\!\big|\!\big|^2_{\Omega\setminus U_k(T)}\leq(1+L^{-1})^{-1}\big|\!\big|\!\big|\mathbf{w}^T\big|\!\big|\!\big|^2_{\Omega\setminus U_{k-1}(T)}.$$

A recursive application of this inequality and $\big|\!\big|\!\big|\mathbf{w}^T\big|\!\big|\!\big|_{\Omega\setminus U_0(T)}\leq\big|\!\big|\!\big|\mathbf{w}^T\big|\!\big|\!\big|_\Omega$ yields

$$\big|\!\big|\!\big|\mathbf{w}^T\big|\!\big|\!\big|^2_{\Omega\setminus U_k(T)}\leq e^{-\log(1+L^{-1})k}\big|\!\big|\!\big|\mathbf{w}^T\big|\!\big|\!\big|^2_\Omega\leq e^{-\log(1+C^{-1})k/\lambda(H/h)}\big|\!\big|\!\big|\mathbf{w}^T\big|\!\big|\!\big|^2_\Omega,$$

where we used Bernoulli's inequality and that $0<L^{-1}\leq C^{-1}$ in the last step. The choice $\theta:=(1+C^{-1})^{-1}$ proves the lemma. □

The following lemma is the main result of this subsection. It can be directly applied to the localized corrector problems (11) with $F_T(\mathbf{v}_h)=a^T(\mathbf{v},\mathbf{v}_h)$, $G^T_{h,k}\mathbf{v}=\mathbf{w}^{T,k}$ and $G^T_h\mathbf{v}=\mathbf{w}^T$ for any $\mathbf{v}\in V$.

Lemma 21. *Let the setting of Lemma 20 hold true and let additionally* $\mathbf{w}^{T,k} \in K_h^{\mathrm{f}}(U_k(T))$ *denote the solution of*

$$\int_{U_k(T)} \mathbf{A}^{-1}\mathbf{w}^{T,k} \cdot \mathbf{v}_h = F_T(\mathbf{v}_h) \qquad \text{for all } \mathbf{v}_h \in K_h^{\mathrm{f}}(U_k(T)). \tag{30}$$

Then, there exists a generic constant $0 < \theta < 1$ *(depending on the contrast) such that for all positive* $k \in \mathbb{N}$*:*

$$\left|\!\left|\!\left| \sum_{T\in\mathcal{T}_H} \left(\mathbf{w}^T - \mathbf{w}^{T,k}\right) \right|\!\right|\!\right|_\Omega^2 \lesssim k^d \lambda(H/h)^2 \theta^{2k/\lambda(H/h)} \sum_{T\in\mathcal{T}_H} |||\mathbf{w}^T|||_\Omega^2. \tag{31}$$

Proof. Let $\eta_{T,k}$ be defined according to (18) and denote $\mathbf{z} := \sum_{T\in\mathcal{T}_H}(\mathbf{w}^T - \mathbf{w}^{T,k}) \in K_h^{\mathrm{f}}$. We obtain

$$|||\mathbf{z}|||_\Omega^2 = \sum_{T\in\mathcal{T}_H} \underbrace{(\mathbf{A}^{-1}(\mathbf{w}^T - \mathbf{w}^{T,k}), (1-\eta_{T,k+1})\mathbf{z})}_{=:\mathrm{I}} + \underbrace{(\mathbf{A}^{-1}(\mathbf{w}^T - \mathbf{w}^{T,k}), \eta_{T,k+1}\mathbf{z})}_{=:\mathrm{II}}.$$

The first term is estimated by

$$\mathrm{I} \le \left|\!\left|\!\left|\mathbf{w}^T - \mathbf{w}^{T,k}\right|\!\right|\!\right|_\Omega |||\mathbf{z}(1-\eta_{T,k+1})|||_{U_{k+1}(T)} \le \left|\!\left|\!\left|\mathbf{w}^T - \mathbf{w}^{T,k}\right|\!\right|\!\right|_\Omega |||\mathbf{z}|||_{U_{k+1}(T)}.$$

For the second term we have $\mathbf{z} \in K_h^{\mathrm{f}}$, hence there exists again a skew-symmetric matrix $\psi = \psi(\mathbf{z})$ with the properties as in Lemma 18 with

$$\eta_{T,k+1}\mathbf{z} = \eta_{T,k+1}\nabla\cdot\psi = \nabla\cdot(\eta_{T,k+1}\psi) - \psi\nabla\eta_{T,k+1}.$$

We define $\tilde{\mathbf{z}} := \Pi_h(\nabla\cdot(\eta_{T,k+1}\psi)) - (\Pi_H \circ \Pi_h)(\nabla\cdot(\eta_{T,k+1}\psi))$. Using Lemma 19 and $\mathrm{supp}(\eta_{T,k+1}\mathbf{z}) \cap \mathrm{supp}(\mathbf{w}^{T,k}) = \emptyset$ we get

$$\begin{aligned}
(\mathbf{A}^{-1}(\mathbf{w}^T - \mathbf{w}^{T,k}), \eta_{T,k+1}\mathbf{z}) &= (\mathbf{A}^{-1}\mathbf{w}^T, \eta_{T,k+1}\mathbf{z}) \\
&\overset{(28)}{=} \int_{\Omega\setminus U_k(T)} \mathbf{A}^{-1}\mathbf{w}^T \cdot (\nabla\cdot(\eta_{T,k+1}\psi) - \psi\nabla\eta_{T,k+1} - \tilde{\mathbf{z}}) \\
&= \int_{\Omega\setminus U_k(T)} \mathbf{A}^{-1}\left(\mathbf{w}^T - \mathbf{w}^{T,k}\right) \cdot (\nabla\cdot(\eta_{T,k+1}\psi) - \psi\nabla\eta_{T,k+1} - \tilde{\mathbf{z}}).
\end{aligned}$$

Now proceed as in Lemma 20 to obtain

$$\mathrm{II} \lesssim \lambda(H/h)\left|\!\left|\!\left|\mathbf{w}^T - \mathbf{w}^{T,k}\right|\!\right|\!\right|_\Omega |||\mathbf{z}|||_{U_{k+1}(T)}.$$

Combining the estimates for I and II and applying Hölder's inequality finally yields, for $k \ge 1$,

$$\begin{aligned}
|||\mathbf{z}|||_\Omega^2 &\lesssim \lambda(H/h) \sum_{T\in\mathcal{T}_H} \left|\!\left|\!\left|\mathbf{w}^T - \mathbf{w}^{T,k}\right|\!\right|\!\right|_\Omega |||\mathbf{z}|||_{U_{k+1}(T)} \\
&\lesssim k^{\frac{d}{2}} \lambda(H/h) \left(\sum_{T\in\mathcal{T}_H} \left|\!\left|\!\left|\mathbf{w}^T - \mathbf{w}^{T,k}\right|\!\right|\!\right|_\Omega^2 \right)^{\frac{1}{2}} |||\mathbf{z}|||_\Omega.
\end{aligned} \tag{32}$$

It remains to bound $\left|\left|\left|\mathbf{w}^T - \mathbf{w}^{T,k}\right|\right|\right|_\Omega^2$. In order to do this, we use Galerkin orthogonality for the local problems, which gives us

$$\left|\left|\left|\mathbf{w}^T - \mathbf{w}^{T,k}\right|\right|\right|_\Omega^2 \leq \inf_{\tilde{\mathbf{w}}^{T,k} \in K_h^{\mathrm{f}}(U_k(T))} \left|\left|\left|\mathbf{w}^T - \tilde{\mathbf{w}}^{T,k}\right|\right|\right|_\Omega^2.$$

Again, we use Lemma 20 to show

$$\left|\left|\left|\mathbf{w}^T - \mathbf{w}^{T,k}\right|\right|\right|_\Omega^2 \lesssim \theta^{2k/\lambda(H/h)} \left|\left|\left|\mathbf{w}^T\right|\right|\right|_\Omega^2. \tag{33}$$

Combining (32) and (33) proves the lemma. □

4.3 Inf-sup stability revisited

The decay results can be used to prove another inf-sup stability constant γ_k^1 in addition to γ_k^0 from Lemma 12 for the bilinear form $b(\cdot,\cdot)$ with the localized multiscale space. Using Lemma 21, we obtain

$$\begin{aligned}
\|G_{h,k}\mathbf{v} - G_h\mathbf{v}\|_{L^2(\Omega)}^2 &= \left\| \sum_{T \in \mathcal{T}_H} (G_{h,k}^T\mathbf{v} - G_h^T\mathbf{v}) \right\|_{L^2(\Omega)}^2 \\
&\lesssim k^d \lambda(H/h)^2 \theta^{2k/\lambda(H/h)} \sum_{T \in \mathcal{T}_H} \|G_h^T\mathbf{v}\|_{L^2(\Omega)}^2 \\
&\lesssim k^d \lambda(H/h)^2 \theta^{2k/\lambda(H/h)} \|\mathbf{v}\|_{L^2(\Omega)}^2.
\end{aligned}$$

We get the following stability

$$\begin{aligned}
\|G_{h,k}\mathbf{v}\|_{L^2(\Omega)} &\leq \|G_{h,k}\mathbf{v} - G_h\mathbf{v}\|_{L^2(\Omega)} + \|G_h\mathbf{v}\|_{L^2(\Omega)} \\
&\lesssim (k^{d/2}\lambda(H/h)\theta^{k/\lambda(H/h)} + 1)\|\mathbf{v}\|_{L^2(\Omega)}.
\end{aligned}$$

Using the same technique as in Lemma 12, we obtain an inf-sup stability constant $\gamma_k^1 := \gamma(2 + k^{d/2}\lambda(H/h)\theta^{k/\lambda(H/h)})^{-1}$.

For the nodal Raviart–Thomas interpolation operator Π_H, $\lambda(H/h)$ depends on h and H, and we cannot obtain a uniform bound on the constant for this estimate either. However, for L^2-stable Clément-type interpolation operators (discussed in Remark 14), we have $\lambda(H/h) \equiv 1$, independently of h and H. If using such an interpolator in place of Π_H, the inf-sup stability constant γ_k^1 can be bounded from below by a positive constant independent of h and H, since $k^{d/2}\theta^k$ is bounded from above with respect to k.

5 Numerical experiments

Four numerical experiments are presented in this section. Their purpose is to show that the error estimate for the localized multiscale method presented in Theorem 16 is valid and useful for determining the patch sizes and that the method is competitive.

A brief overview of the implementation of the method follows. The two dimensional Raviart–Thomas finite element is used. For all free degrees of freedom e (interior edges), the localized global corrector $G_{h,k}\Phi_e$ for the corresponding basis function Φ_e is computed

according to equation (11). The additional constraints on the test and trial functions to be in the kernel to the coarse Raviart–Thomas projection operator are implemented using Lagrange multipliers (in addition to those already there due to the mixed formulation). The corrector problems are cheap since they are solved only on small patches. This can be done in parallel over all basis functions. Finally, problem (12) is solved. Regarding the linear system arising here, we compare it with the linear system arising from a standard Raviart–Thomas discretization (using V_H for the flux) of the mixed formulation on the corase mesh:

$$\begin{pmatrix} \mathbf{K} & \mathbf{B}^T \\ \mathbf{B} & \mathbf{0} \end{pmatrix} = \begin{pmatrix} \mathbf{0} \\ \mathbf{b} \end{pmatrix},$$

for matrices $\mathbf{K}$ and $\mathbf{B}$ and a vector $\mathbf{b}$. The difference with the multiscale method is that matrix corresponding to the bilinear form $a(\cdot,\cdot)$ is computed using the low dimensional modified localized multiscale basis $\{\Phi_E - G_{h,k}\Phi_E\}_E$ spanning $V_{H,h}^{\mathrm{ms},k}$. Since the correctors are divergence free, $\mathbf{K}$ is replaced by a different matrix $\tilde{\mathbf{K}}$ in the system above, whereas $\mathbf{B}$ and $\mathbf{b}$ are left intact.

In all numerical experiments below, the diffusion matrix is diagonal with identical diagonal elements, $\mathbf{A}(x) = A(x)\mathbf{I}$, with $\mathbf{I}$ being the identity matrix, for a scalar-valued function A.

5.1 Investigation of error from localization

In this experiment, we investigate how the error in energy norm of the localized multiscale solution is affected by the localization to patches of the correctors. The error due to localization is bounded by the second term in the estimate in Theorem 16. This term will be the focus of this experiment.

The computational domain is the unit square $\Omega = [0,1]^2$ and the source function is given by

$$f(x) = \begin{cases} 1 & \text{if } x \in [0,1/4]^2, \\ -1 & \text{if } x \in [3/4,1]^2, \\ 0 & \text{otherwise.} \end{cases}$$

We consider three different diffusion coefficients A:

1. Constant: $A(x) = 1$ in the whole domain.
2. Noise: $A(x)$ is piecewise constant on a $2^7 \times 2^7$ uniform rectangular grid. In each grid cell, the value of A is equal to a realization of $\exp(10\omega)$, where ω is a cell-specific standard uniformly distributed variable.
3. Channels: $A(x)$ is as is shown in Figure 2. It is piecewise constant on a $2^7 \times 2^7$ uniform rectangular grid. The coefficient $A(x) = 1$ for x in black cells and $A(x) = \exp(10)$ for x in white cells.

Figure 3 shows the mesh used in the experiment. Both fine and coarse meshes are constructed as shown in the figure. A reference solution $\mathbf{u}_h$ was computed with the standard Raviart–Thomas spaces V_h and Q_h with $h = 2^{-8}$. Solutions $\mathbf{u}_{H,h}^{\mathrm{ms},k}$ to the localized multiscale problem were computed using $H = 2^{-2}, 2^{-3}, \ldots, 2^{-6}$. The patch size k was chosen

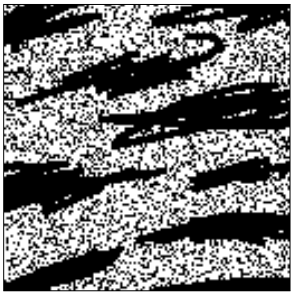
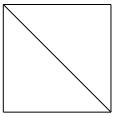
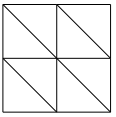


Figure 2: Coefficient A defined on a $2^7 \times 2^7$ grid of Ω.



(a) Coarsest mesh, $h = 1$.



(b) One refinement, $h = 1/2$.

Figure 3: Family of triangulations of the unit square.

as

$$k = C(1 + \log_2(H/h))^{1/2} \log_2(H^{-1})$$

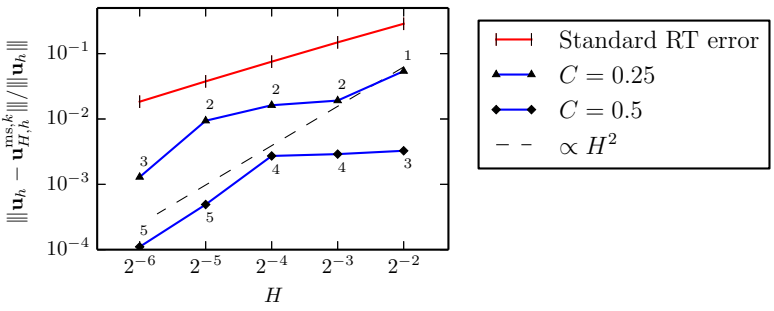
rounded to the nearest integer with $C = 0.25$ and $C = 0.5$. The relative error (using the reference solution in place of the exact solution) in energy norm, i.e. $\left|\!\left|\!\left| \mathbf{u}_h - \mathbf{u}_{H,h}^{\mathrm{ms},k} \right|\!\right|\!\right| / |\!|\!| \mathbf{u}_h |\!|\!|$ was computed. See Figure 4 for the resulting convergence of this error with respect to H for the two values of C. Note that since $f \in Q_H$ for all examples, the first term in (15) vanishes. The error is hence bounded by $k^{d/2}\lambda(H/h)^2\theta^{k/\lambda(H/h)}\|f\|_{L^2(\Omega)}$, which allows for a careful investigation of the influence of k, H and h. A reference line proportional to H^2 is plotted for guidance. We can see that we achieve convergence for both choices of C. However, since k is rounded to an integer, the convergence plots have a staggered appearance. This example shows that the error due to localization can be kept small and decreasing with H. The plots also show the relative error in energy norm for the standard Raviart–Thomas discretization on the coarse mesh. It is evident that the localized multiscale space has good approximation properties since it permits convergence while the standard space of the same dimension does not.

5.2 Investigation of instability

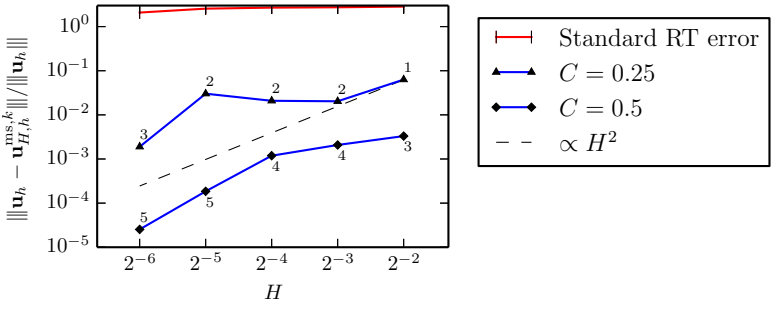
In this experiment we show how singularity-like features can appear in the solution, probably as a result of high contrast in combination with the L^2-instability of the nodal Raviart–Thomas interpolator.

Again, we consider the unit square $\Omega = [0, 1]^2$. The diffusion coefficient A is chosen according to Figure 5. In other words, A is defined as

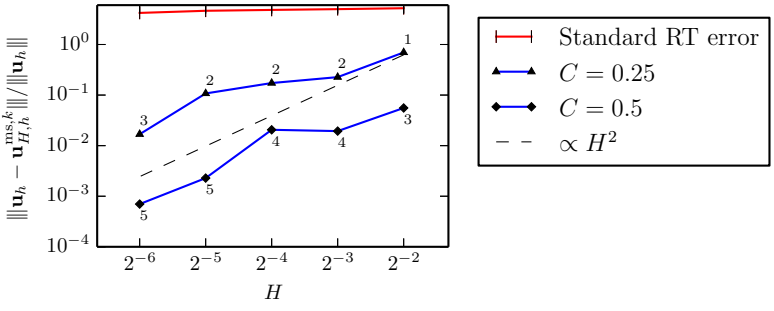
$$A(x) = \begin{cases} \exp(10) & \text{if } x_2 < 1/2 \text{ or } x \in [\frac{1}{2} - 2^{-5}, \frac{1}{2} + 2^{-5}] \times [\frac{1}{2}, \frac{1}{2} + 2^{-5}], \\ 1 & \text{otherwise.} \end{cases}$$



(a) Diffusion coefficient is constant.



(b) Diffusion coefficient is noisy.



(c) Diffusion coefficient has channel structures.

Figure 4: Convergence plots for localization error experiment. Relative error in energy norm for three choices of A, for different values of the constant C determining the patch size. The number adjacent to a point is the actual value of k for the specific simulation corresponding to that point.

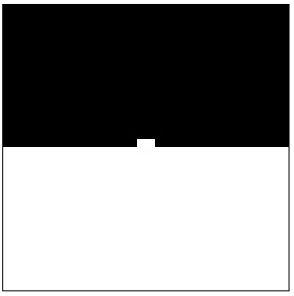


Figure 5: Coefficient A defined on a $2^5 \times 2^5$ grid of Ω where $A(x) = 1$ for x in black cells and $A(x) = \exp(10)$ for x in white cells.

The source function is chosen as

$$f(x) = \begin{cases} -1 & \text{if } x_2 < 1/2, \\ 1 & \text{otherwise.} \end{cases}$$

This particular choice of A and f yields a localized multiscale solution with a clear singularity-like feature at $x = (x_1, x_2) = (1/2, 1/2)$ in the localized multiscale solution.

We use the family of triangulations presented in Figure 3 and fix $H = 1/4$ so that f is resolved on the coarse scale. Then $f \in Q_H$ and all error stems from localization (see Theorem 16). We let the resolution h of the fine space be $h = 2^{-5}, 2^{-6}, \ldots, 2^{-9}$. Choosing $k = 2$, we compute the localized multiscale solution $\mathbf{u}_{H,h}^{\text{ms},k}$ and reference solution $\mathbf{u}_h$ for the given values of h.

From the error estimate in Theorem 16, we expect to have

$$\begin{aligned} \left|\!\left|\!\left| \mathbf{u}_h - \mathbf{u}_{H,h}^{\text{ms},k} \right|\!\right|\!\right| &\lesssim k^{d/2} \lambda (H/h)^2 \theta^{k/\lambda(H/h)} \|f\|_{L^2(\Omega)} \\ &\propto \log\left(h^{-1}\right) \quad \text{as } h \to 0. \end{aligned}$$

The energy norm of the error is plotted in Figure 6. We can see that for this particular problem and range of h, the error increases with h and with rate $\log(h^{-1})$ as predicted by the error estimate. However, the error estimate seems not to be sharp for this particular example. Figure 7 shows the reference and multiscale flux solutions. The magnitude of the reference solution is in the range $[0, 3]$, while the multiscale solution has a spike reaching magnitude 30 at $x = (1/2, 1/2)$. Interesting to note is that the singularities vanish for the ideal multiscale method, i.e. without localization, see Lemma 9.

5.3 Convergence in an L-shaped domain

Next, we consider an L-shaped domain with noisy diffusion coefficient A (case 2. in Section 5.1) and with $f \notin Q_H$. In this experiment, we show that the localization error investigated in the previous section can be dominated by errors from projecting f.

We use the domain $\Omega = [0, 1]^2 \setminus [1/2, 1] \times [0, 1/2]$ and the triangulation presented in Figure 8. Both fine and coarse meshes are constructed as shown in the figure. Further,

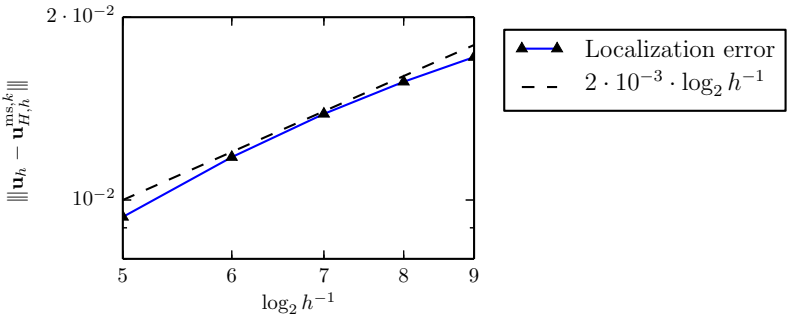
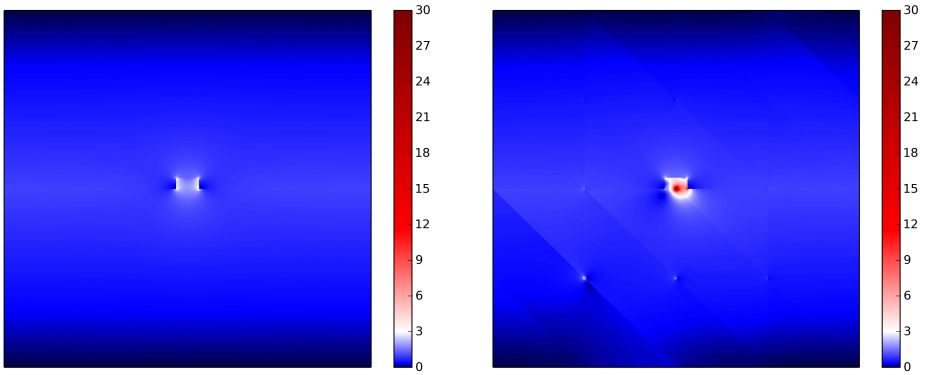


Figure 6: Divergence of the energy norm of the localization error of a particular multiscale solution as h decreases.



(a) Reference solution, $h = 2^{-9}$.

(b) Multiscale solution, $h = 2^{-9}$, $H = 2^{-2}$, $k = 2$.

Figure 7: Magnitude of flux at the centroid of the triangles.

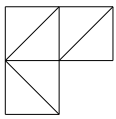
we choose source function as

$$f(x) = \begin{cases} 1/2 + x_1 - x_2 & \text{if } x_2 < 1/2, \\ -(1/2 + x_1 - x_2) & \text{if } x_1 > 1/2, \\ 0 & \text{otherwise.} \end{cases}$$

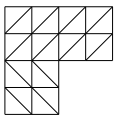
Note that $f \notin Q_H$ and $\|f - P_H f\|_{L^2(\Omega)} \lesssim H\|f\|_{L^2(\Omega)}$. A reference solution $\mathbf{u}_h$ was computed with the standard Raviart–Thomas spaces V_h and Q_h with $h = 2^{-8}$. Solutions $\mathbf{u}_{H,h}^{\mathrm{ms},k}$ to the localized multiscale problem were computed using $H = 2^{-2}, 2^{-3}, \ldots, 2^{-6}$. The patch size k was chosen as

$$k = C(1 + \log_2(H/h))^{1/2} \log_2(H^{-1})$$

rounded to the nearest integer, with $C = 0.25$ and $C = 0.5$. The relative error in energy norm was recorded for the solutions corresponding to the values of H. The resulting



(a) Coarsest mesh, $h = 1/2$.



(b) One refinement, $h = 1/4$.

Figure 8: Family of triangulations of the L-shaped domain.

convergence plot can be found in Figure 9. We expect the first term in the error estimate,

$$\left|\!\left|\!\left| \mathbf{u}_h - \mathbf{u}_{H,h}^{\mathrm{ms},k} \right|\!\right|\!\right| \lesssim H\|f - P_H f\|_{L^2(\Omega)} + k^{d/2}\lambda(H/h)\theta^{k/\lambda(H/h)}\|f\|_{L^2(\Omega)} \tag{34}$$

to be of order H^2. From the convergence plots we can see that $C = 0.25$ is not sufficient to make the localization error of at least order H^2, however, $C = 0.5$ is.

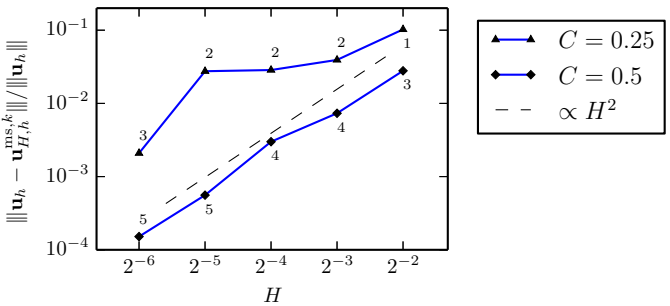


Figure 9: Convergence plot for experiment with L-shaped domain. Shows relative error in energy norm for two values of C and a series of values of H. The number adjacent to a point is the actual value of k for the specific simulation corresponding to that point.

5.4 Comparison with MsFEM

We compare the proposed method with the results obtained using the Multiscale Finite Element Method (MsFEM) based approach in [3]. The domain is $\Omega = [0, 1.2] \times [0, 2.2]$ and the permeability coefficient A is given in a uniform rectangular grid of size 60×220 by the 85th permeability layer in model 2 of SPE10 [10].

The method proposed in [3] is based on a fine and a coarse mesh with quadrilateral elements. The fine mesh is uniform 60×220, i.e. aligned with the permeability data, and the coarse mesh is 6×22, so that each coarse element is subdivided into 10×10 fine elements. The implementation of the method proposed in this work uses triangular meshes, which is why we divide each of the rectangular elements into two triangular elements by a diagonal line drawn from the upper left corner to the lower right corner. As coarse mesh, we use a similar triangular mesh that is constructed from a 6×22 rectangular mesh such that the fine mesh is a conforming refinement of the coarse mesh.

The (quasi-singular) source data f is equal to 1 in the lower left and -1 in the upper right fine quadrilateral element. Note that such f is a discretization of point sources

that model production wells. In particular, the source terms on the continuous level are mathematically described by Dirac delta functions. Hence, for $h \to 0$, we only have $f \in W^{-m,2}(\Omega)$ for $m > \frac{d}{2}$, opposed to $f \in L^2(\Omega)$ as is required for our analysis. To account for this difference, we follow [26] and compute the localized source corrections $F_h^{T,\ell} f \in V_h^{\mathrm{f}}(U_\ell(T))$ on ℓ-coarse-layer patches for $T \in \mathcal{T}_H$,

$$a(F_h^{T,\ell} f, \mathbf{v}_h^{\mathrm{f}}) + b(\mathbf{v}_h^{\mathrm{f}}, \tilde{F}_h^{T,\ell} f) + b(F_h^{T,\ell} f, q_h^{\mathrm{f}}) = -(f, q_h^{\mathrm{f}})_T,$$

for all $\mathbf{v}_h^{\mathrm{f}} \in V_h^{\mathrm{f}}(U_\ell(T))$ and $q_h^{\mathrm{f}} \in Q_h^{\mathrm{f}}(U_\ell(T))$, where $Q_h^{\mathrm{f}}(U_\ell(T))$ is the restriction of Q_h^{f} to $U_\ell(T)$, analogous to the definition of $V_h^{\mathrm{f}}(U_\ell(T))$. (The pressure solution $\tilde{F}_h^{T,\ell} f$ is not needed for correcting the flux and is discarded after its use as Lagrange multiplier). Since f is non-zero only for the two triangles T_1 and T_2 in the lower left and upper right corners, only two such corrector problems need to be solved. The total localized source correction is $F_h^\ell f = F_h^{T_1,\ell} f + F_h^{T_2,\ell} f \in V_h^{\mathrm{f}}$.

The localized corrector problems (11) are unaffected by the source correction. The right hand side of the localized multiscale problem (12) is appended with the localized source corrections and instead reads: find $\mathbf{u}_{H,h}^{\mathrm{ms},k,\ell}$ such that

$$a(\mathbf{u}_{H,h}^{\mathrm{ms},k,\ell}, \mathbf{v}_h) + b(\mathbf{v}_h, p_H) + b(\mathbf{u}_{H,h}^{\mathrm{ms},k,\ell}, q_H) = -(f, q_H) - a(F_h^\ell f, \mathbf{v}_h).$$

Using a value of $\ell = 0$ will be referred to as an ad-hoc source correction, since we do not expect to have any decay of the correction already within the source triangle itself.

We emphasize that the need for source correctors for singular source terms is not an exclusive drawback for our approach, but it is a common necessity shared by all comparable multiscale methods in this setting. In particular they are also used for the MsFEM-based approach in [3] that we use for our comparative study.

The proposed localized multiscale method was used to solve for the flux in the described problem for three corrector patch sizes: $k = 1, 2$, and 3. Three variants of source correction were used: i) without source correction, i.e. $\mathbf{u}_{H,h}^{\mathrm{ms},k}$, ii) with ad-hoc source correction, i.e. $\mathbf{u}_{H,h}^{\mathrm{ms},k,\ell}$ for $\ell = 0$ (without interpolation constraint), and iii) with source correction, i.e. $\mathbf{u}_{H,h}^{\mathrm{ms},k,\ell}$ for $\ell = k, k+1, \infty$. A reference solution $\mathbf{u}_h$ was computed on the fine mesh. Table 1 shows the relative energy norm and L^2-norm of the difference between the localized multiscale solution and the reference solution for the different values of k and ℓ. The corresponding L^2-norm of the error for the MsFEM method with oversampling HE0-OS proposed in [3] is also presented in the table. Note that HE0-OS is based on a discretization with roughly 33% less degrees of freedom than the proposed method, since it uses quadrilaterals instead of triangles (however, since this holds for both the fine and the coarse mesh, the relative change in the amount of degrees of freedom with respect to the reference solution is the same). The flux solutions are plotted in Figure 10.

The results show that the proposed method even without error correction compares favorably with the homogenization based approach. Ad-hoc error correction gives small errors for this problem in both norms. For source correction with patch size $\ell = k$, instabilities similar to that studied in Section 5.2 cause the error to increase. However, letting $\ell = k + 1$ is enough to get errors that compare favorably with [3].

Table 1: Relative error in energy norm and L^2-norm for the SPE10-85 problem.

Method	k	ℓ	Energy norm	L^2-norm
Proposed method without source correction	1	–	0.7863	0.4069
	2	–	0.7856	0.3369
	3	–	0.7855	0.3325
Proposed method with ad-hoc source correction ($\ell = 0$)	1	0	0.1541	0.2700
	2	0	0.1515	0.1467
	3	0	0.1537	0.1379
Proposed method with source correction ($\ell = k, k+1, \infty$)	1	1	0.1090	0.8292
	1	2	0.0459	0.2703
	1	∞	0.0350	0.2504
	2	2	0.0549	0.7453
	2	3	0.0185	0.0517
	2	∞	0.0150	0.0490
	3	3	0.0080	0.0178
	3	4	0.0051	0.0424
	3	∞	0.0041	0.0088
HE0-OS [3]	–	–	–	0.3492

References

[1] J. Aarnes. On the use of a mixed multiscale finite element method for greater flexibility and increased speed or improved accuracy in reservoir simulation. *Multiscale Model. Simul.*, 2(3):421–439, 2004.

[2] A. Abdulle and P. Henning. Localized orthogonal decomposition method for the wave equation with a continuum of scales. *arXiv:1406.6325*, 2014.

[3] T. Arbogast. Homogenization-based mixed multiscale finite elements for problems with anisotropy. *Multiscale Model. Simul.*, 9(2):624–653, 2011.

[4] T. Arbogast and K. Boyd. Subgrid upscaling and mixed multiscale finite elements. *SIAM J. Numer. Anal.*, 44(3):1150–1171, 2006.

[5] D. N. Arnold, R. S. Falk, and R. Winther. Finite element exterior calculus, homological techniques, and applications. *Acta Numer.*, 15:1–155, 2006.

[6] D. Boffi, F. Brezzi, and M. Fortin. *Mixed Finite Element Methods and Applications*, volume 44 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin Heidelberg, 2nd edition, 2013.

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

[8] S. H. Christiansen. Stability of Hodge decompositions in finite element spaces of differential forms in arbitrary dimension. *Numer. Math.*, 107(1):87–106, 2007.

[9] S. H. Christiansen and R. Winther. Smoothed projections in finite element exterior calculus. *Math. Comp.*, 77(262):813–829, 2008.

[10] M. A. Christie. Tenth SPE comparative solution project: A comparison of upscaling techniques. *SPE Reservoir Eval. Eng.*, 4:308–317, 2001.

[11] 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.

[12] D. Elfverson, V. Ginting, and P. Henning. On multiscale methods in Petrov–Galerkin formulation. *Numer. Math. (Online First)*, 2015.

[13] P. Henning and A. Målqvist. Localized orthogonal decomposition techniques for boundary value problems. *SIAM J. Sci. Comput.*, 36(4):A1609–A1634, 2014.

[14] P. Henning, A. Målqvist, and D. Peterseim. A localized orthogonal decomposition method for semi-linear elliptic problems. *ESAIM Math. Model. Numer. Anal.*, 48(5):1331–1349, 2014.

[15] P. Henning, P. Morgenstern, and D. Peterseim. Multiscale partition of unity. In M. Griebel and M. A. Schweitzer, editors, *Meshfree Methods for Partial Differential Equations VII*, volume 100 of *Lecture Notes in Computational Science and Engineering*, pages 185–204. Springer International Publishing, 2015.

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

[17] 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.

[18] T. Hughes and G. Sangalli. Variational multiscale analysis: the fine-scale Green's function, projection, optimization, localization, and stabilized methods. *SIAM J. Numer. Anal.*, 45(2):539–557, 2007.

[19] T. J. R. Hughes. Multiscale phenomena: Green's functions, the Dirichlet-to-Neumann formulation, subgrid scale models, bubbles and the origins of stabilized methods. *Comput. Methods Appl. Mech. Engrg.*, 127(1-4):387–401, 1995.

[20] 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.

[21] D. Iftimie, G. Karch, and C. Lacave. Asymptotics of solutions to the Navier–Stokes system in exterior domains. *J. London Math. Soc. (Online First)*, 2014.

[22] 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.

[23] M. G. Larson and A. Målqvist. A mixed adaptive variational multiscale method with applications in oil reservoir simulation. *Math. Models Methods Appl. Sci.*, 19(07):1017–1042, 2009.

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

[25] A. Målqvist and D. Peterseim. Computation of eigenvalues by numerical upscaling. *Numer. Math. (Online First)*, 2014/15.

[26] A. Målqvist. Multiscale methods for elliptic problems. *Multiscale Model. Simul.*, 9(3):1064–1086, 2011.

[27] J. Nolen, G. Papanicolaou, and O. Pironneau. A framework for adaptive multiscale methods for elliptic problems. *Multiscale Model. Simul.*, 7(1):171–196, 2008.

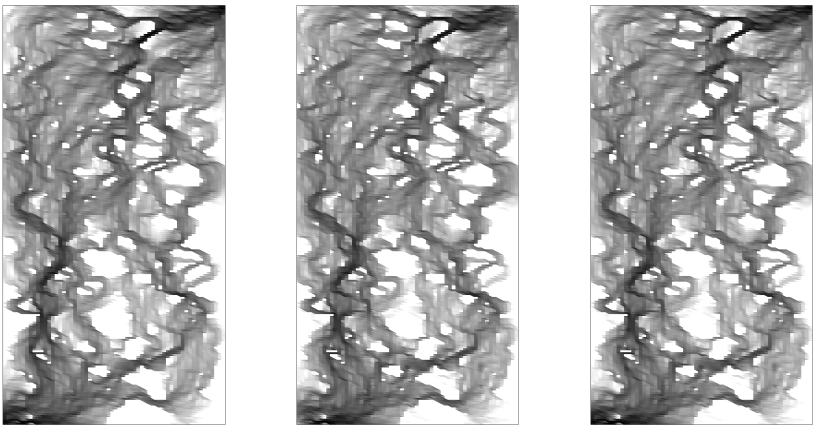
[28] D. Peterseim. Eliminating the pollution effect in Helmholtz problems by local subscale correction. *arXiv:1411.7512*, 2014.

[29] P. A. Raviart and J. M. Thomas. A mixed finite element method for 2-nd order elliptic problems. In I. Galligani and E. Magenes, editors, *Mathematical Aspects of Finite Element Methods*, volume 606 of *Lecture Notes in Mathematics*, pages 292–315. Springer Berlin Heidelberg, 1977.

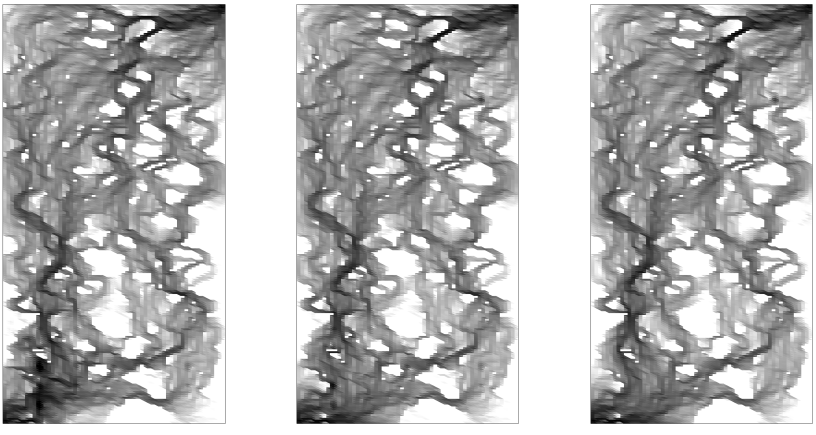
[30] J. Schöberl. A posteriori error estimates for Maxwell equations. *Math. Comp.*, 77(262):633–649, 2008.

[31] H. Wendland. Divergence-free kernel methods for approximating the Stokes problem. *SIAM J. Numer. Anal.*, 47(4):3158–3179, 2009.

[32] B. Wohlmuth, A. Toselli, and O. Widlund. An iterative substructuring method for Raviart–Thomas vector fields in three dimensions. *SIAM J. Numer. Anal.*, 37(5):1657–1676, 2000.



(a) Reference solution (b) $k = 1, \ell = -1$ (c) $k = 1, \ell = 0$



(d) $k = 1, \ell = 1$ (e) $k = 1, \ell = 2$ (f) $k = 2, \ell = 3$

Figure 10: Flux solutions for the SPE10-85 problem. Figure (a) shows the reference flux solution and (b–f) show the multiscale flux solutions for $k = 1$ and 2, and different source corrections ($\ell = -1$ means no source correction and $\ell = 0$ means ad-hoc error correction). The color maps to the magnitude of the flux at the midpoint of the triangular elements. The colors map from 10^{-5} (white) to 10^{-2} (black) and is saturated at white and black for lower and higher values, respectively.