Multilevel Monte Carlo methods and uncertainty quantification

submitted by

Aretha Leonore Teckentrup

for the degree of Doctor of Philosophy

of the

University of Bath

Department of Mathematical Sciences

June 2013

Summary

We consider the application of multilevel Monte Carlo methods to elliptic partial differential equations with random coefficients. Such equations arise, for example, in stochastic groundwater flow modelling. Models for random coefficients frequently used in these applications, such as log-normal random fields with exponential covariance, lack uniform coercivity and boundedness with respect to the random parameter and have only limited spatial regularity.

To give a rigorous bound on the cost of the multilevel Monte Carlo estimator to reach a desired accuracy, one needs to quantify the bias of the estimator. The bias, in this case, is the spatial discretisation error in the numerical solution of the partial differential equation. This thesis is concerned with establishing bounds on this discretisation error in the practically relevant and technically demanding case of coefficients which are not uniformly coercive or bounded with respect to the random parameter.

Under mild assumptions on the regularity of the coefficient, we establish new results on the regularity of the solution for a variety of model problems. The most general case is that of a coefficient which is piecewise Hölder continuous with respect to a random partitioning of the domain. The established regularity of the solution is then combined with tools from classical discretisation error analysis to provide a full convergence analysis of the bias of the multilevel estimator for finite element and finite volume spatial discretisations. Our analysis covers as quantities of interest several spatial norms of the solution, as well as point evaluations of the solution and its gradient and any continuously Fréchet differentiable functional.

Lastly, we extend the idea of multilevel Monte Carlo estimators to the framework of Markov chain Monte Carlo simulations. We develop a new multilevel version of a Metropolis Hastings algorithm, and provide a full convergence analysis.

Acknowledgements

First of all, I would like to thank my supervisor Rob Scheichl - for his encouragement and support over the last years, for his continuous enthusiasm about our work and for providing me with numerous opportunities to discuss my work with others.

My thanks also go to Professor Mike Giles at the University of Oxford, with whom I've had many helpful discussions, and who has often given me the opportunity to present my work at international conferences.

Finally, I would also like to thank all my other collaborators, with whom I have very much enjoyed working over the last few years. In no particular order, they are Andrew Cliffe, Julia Charrier, Elisabeth Ullmann, Minho Park, Panayot Vassilevski and Christian Ketelsen. I would in particular like to thank Panayot Vassilevski for giving me the opportunity to visit the Lawrence Livermore National Lab on two occasions.

Contents

List of Figures

Chapter 1

Introduction

1.1 Motivation

There are many situations in which modelling and computer simulation are indispensable tools and where the mathematical models employed have been demonstrated to give adequate representations of reality. However, the parameters appearing in the models often have to be estimated from measurements and are, therefore, subject to uncertainty. This uncertainty propagates through the simulations and quantifying its impact on the results is frequently of great importance.

A good example is provided by the problem of assessing the safety of a potential deep geological repository for radioactive waste. Any radionuclides leaking from such a repository could be transported back to the human environment by groundwater flowing through the rocks beneath the earth's surface. The very long timescales involved mean that modelling and simulation are essential in evaluating repository performance. The study of groundwater flow is well established, and there is general scientific consensus that in many situations Darcy's Law can be expected to lead to an accurate description of the flow [21, 16]. The classical equations governing (steady state) single phase subsurface flow consist of Darcy's law coupled with an incompressibility condition:

$$\mathbf{q} + \mathbf{A}\nabla u = \mathbf{g} \quad \text{and} \quad \nabla \cdot \mathbf{q} = 0, \quad \text{in} \;\; D \subset \mathbb{R}^d, \; d = 1, 2, 3, \tag{1.1}$$

subject to suitable boundary conditions. In physical terms, u denotes the pressure head of the fluid, $\mathbf{A}$ is the permeability tensor, $\mathbf{q}$ is the filtration velocity (or Darcy

flux) and $\mathbf{g}$ are the source terms.

The main parameter appearing in Darcy's Law is the permeability $\mathbf{A}$, which characterises how easily water can flow through the rock under a given pressure gradient. In practice it is only possible to measure the permeability at a limited number of spatial locations, but it is required at all points of the computational domain for the simulation. This fact is the primary source of uncertainty in groundwater flow calculations. Understanding and quantifying the impact of this uncertainty on predictions of radionuclide transport is essential for reliable repository safety assessments.

A widely used approach for dealing with uncertainty in groundwater flow is to represent the permeability $\mathbf{A}$ as a random field over a probability space $(\Omega, \mathcal{E}, \mathbb{P})$, with a mean and covariance structure that has to be inferred from the data [23, 22]. This means that (1.1) becomes a system of PDEs with random coefficients, which can be written in second order form as

$$-\nabla \cdot (\mathbf{A}(\omega, x)\nabla u(\omega, x)) = f(\omega, x), \qquad \text{in} \quad D, \tag{1.2}$$

with $f = -\nabla \cdot \mathbf{g}$, and subject to appropriate boundary conditions. This means that the solution u will also be a random field.

In this general form solving (1.2) is extremely challenging computationally, and in practice it is common to use relatively simple models for $\mathbf{A}$ that are as faithful as possible to the measured data. One model that has been studied extensively is a log-normal distribution for $\mathbf{A}$, i.e. replacing the permeability tensor by a scalar valued field a whose log is Gaussian. It guarantees that $a > 0$ almost surely (a.s.) in Ω, and it allows the permeability to vary over many orders of magnitude, which is typical in subsurface flow.

When modelling a whole aquifer, a whole oil reservoir, or a sufficiently large region around a potential radioactive waste repository, the correlation length scale for a is typically significantly smaller than the size of the computational region D. In addition, typical sedimentation processes lead to fairly irregular structures and pore networks, and faithful models should therefore also only assume limited spatial regularity of a.

In applications, one is then usually interested in finding the expected value (or higher order moments) of some functional $Q = \mathcal{G}(u)$ of the solution u to (1.2).

This could for example be the value of the pressure u or the Darcy flux $-\mathbf{A}\nabla u$ at or around a given point in the computational domain, or the outflow over parts of the boundary. In the context of radioactive waste disposal, it could also be something more complicated, such as positions and travel times of particles released somewhere in the computational domain [39].

Since realistic random field models often need a rather large number of stochastic degrees of freedom ($>$ 100s) for their accurate representation, stochastic Galerkin and stochastic collocation approaches [54, 67] are impractical and standard Monte Carlo (MC) simulation is the method of choice. Since the individual realisations of the random field have low spatial regularity and significant spatial variation, obtaining samples of the pressure field is very costly. The notoriously slow rate of convergence of the standard MC method means that many such realisations are required to obtain accurate results, and the standard MC approach quickly becomes unfeasible. The computational cost of solving elliptic PDEs with random coefficient is a major challenge in uncertainty quantification for groundwater flow studies.

In this thesis, we address the problem of the large cost of solving elliptic PDEs with random coefficients. Our approach is based on a novel variance reduction technique for the standard MC method, called the multilevel Monte Carlo (MLMC) method. The basic idea was introduced by Heinrich to accelerate Monte-Carlo computations of high-dimensional, parameter dependent integrals and to solve integral equations [47]. Similar ideas were used by Brandt and his co-workers to accelerate statistical mechanical calculations [5, 6]. The method was extended by Giles [33, 32] to infinite-dimensional integration related to stochastic differential equations in finance. Since then, it has been applied in many areas of mathematics related to differential equations, in particular stochastic differential equations [24, 32, 50, 53] and several types of partial differential equations (PDEs) with random forcing [34, 40] or random coefficients [4, 12, 15, 37, 66, 65].

The main challenge in the rigorous numerical analysis of MLMC methods for elliptic PDEs with random coefficients, is the quantification of the numerical discretisation error, or in other words the *bias* of the estimator. Models for the random coefficient frequently used in applications, such as log–normal random fields, are not uniformly coercive and bounded, making the numerical analysis challenging. Indeed, if one does assume uniform coercivity and boundedness of

the random coefficient, as well as (spatial) differentiability, the analysis of the discretisation error is classical, and follows immediately from the analysis in the deterministic setting (see e.g [2, 4, 27]).

As such, this thesis builds on and complements [10, 36, 28], which are all concerned with the well-posedness and numerical approximation of elliptic PDEs with infinite dimensional stochastic coefficients that are not uniformly bounded and coercive, such as log-normal coefficients. The novel approach to the discretisation error analysis in this thesis crucially makes use of the observation that for each *fixed* ω, we have a uniformly coercive and bounded problem (in x). The standard tools from (deterministic) discretisation error analysis are hence applicable, albeit with special attention to how these results depend on the data $\mathbf{A}(\omega, x)$.

1.2 Aims, achievements and structure of thesis

The main aim of this thesis is to give a rigorous numerical analysis of the MLMC algorithm applied to elliptic PDEs such as model problem (1.2), under minimal assumptions on the random coefficient $\mathbf{A}$. In particular, we will not assume uniform coercivity or boundedness, and require only limited spatial regularity.

The achievements of this thesis are the following:

- Under minimal assumptions on the coefficient, we prove new regularity results on weak solutions to model problem (1.2) in certain Bochner spaces. We first consider problems posed on smooth domains (Theorem 2.7). By analysing the singularities, we are able to extend the regularity results to Lipschitz polygonal domains (Theorems 2.12) and discontinuous coefficients (Theorem 2.17).

- Using these new regularity results, we then prove bounds on (moments of) the finite element discretisation error in the natural H^1-norm (Theorem 3.3). From this, error estimates in the L^2-norm (Corollary 3.4) and output functionals (Lemma 3.7) follow from a duality argument. We further prove estimates of the finite element error in the L^∞- and $W^{1,\infty}$-norms (Theorem 3.11), and extend the discretisation error analysis to cover also some finite volume schemes (Theorem 3.25 and Lemma 3.26).

- We apply the discretisation error analysis to bound the bias of MLMC estimators applied to model problem (1.2) (Propositions 4.3 - 4.7). Using a generalised complexity theorem (Theorem 4.1), we then give a rigorous bound on the cost of the multilevel estimator, and establish its superiority over standard MC. We show that the cost of the multilevel estimator can be reduced further by using level-dependent estimators (section 4.5.3).

- Finally, we develop a new multilevel estimator in the setting of Markov chain Monte Carlo simulations (Algorithm 2), where the probability distribution of interest (the *posterior* distribution) is generally intractable. We show that moments with respect to the posterior distribution can be bounded in terms of moments with respect to the *prior* distribution (Lemma 5.9). With the prior distribution fulfilling the assumptions for the discretisation error analysis carried out previously, we are then able to prove rigorously the convergence of the new multilevel Markov chain Monte Carlo estimator (Theorems 5.8 and 5.14).

The general structure of the thesis is as follows. We begin by proving regularity results for several variations of model problem (1.2) in chapter 2. We then move on to a (spatial) discretisation error analysis in chapter 3, which includes (but is not limited to) the type of model problems considered in chapter 2. We finally consider the application and analysis of multilevel Monte Carlo methods in chapter 4 and (new) multilevel Markov chain Monte Carlo methods in chapter 5. We finish with some concluding remarks in chapter 6.

Parts of the material in this thesis has been published, or submitted for publication, in the references [15, 12, 66, 65, 52].

1.3 Notation

Given a probability space $(\Omega, \mathcal{E}, \mathbb{P})$ and a bounded Lipschitz domain $D \subset \mathbb{R}^d$, we introduce the following notation. For more details on any of the introduced function spaces, see e.g. [44].

The space of all Lebesgue–measurable functions which are square integrable on D (with respect to the Lebesgue measure) is denoted by $L^2(D)$, with the norm

defined by

$$\|v\|_{L^2(D)} = \left(\int_D |v|^2 \, \mathrm{d}x \right)^{1/2}.$$

For two functions $v, w \in L^2(D)$, we define the $L^2(D)$-inner product

$$(v, w)_{L^2(D)} = \int_D v \, w \, \mathrm{d}x.$$

For any $k \in \mathbb{N}$, the Sobolev space $H^k(D) \subset L^2(D)$ consists of all functions having weak derivatives of order $|\alpha| \leq k$ in $L^2(D)$,

$$H^k(D) = \{v \in L^2(D) : D^\alpha v \in L^2(D) \text{ for } |\alpha| \leq k\}.$$

We identify $H^0(D)$ with $L^2(D)$. We define the following semi-norm and norm on $H^k(D)$:

$$|v|_{H^k(D)} = \left(\int_D \sum_{|\alpha|=k} |D^\alpha v|^2 \, \mathrm{d}x \right)^{1/2} \quad \text{and} \quad \|v\|_{H^k(D)} = \left(\int_D \sum_{|\alpha|\leq k} |D^\alpha v|^2 \, \mathrm{d}x \right)^{1/2}.$$

With $C_0^\infty(D)$ the space of infinitely differentiable functions with compact support on D, the completion of $C_0^\infty(D)$ in $L^2(D)$ with respect to the norm $\|\cdot\|_{H^k(D)}$ is denoted by $H_0^k(D)$. We recall that, since D is bounded, the semi-norm $|\cdot|_{H^k(D)}$ defines a norm equivalent to the norm $\|\cdot\|_{H^k(D)}$ on the subspace $H_0^k(D)$ of $H^k(D)$.

For any real $r \geq 0$, with $r \notin \mathbb{N}$, set $r = k + s$ with $k \in \mathbb{N}$ and $0 < s < 1$, and denote by $|\cdot|_{H^r(D)}$ and $\|\cdot\|_{H^r(D)}$ the Sobolev–Slobodetskii semi-norm and norm, respectively, defined for $v \in H^k(D)$ by

$$|v|_{H^r(D)} = \left(\iint_{D\times D} \sum_{|\alpha|=k} \frac{[D^\alpha v(x) - D^\alpha v(y)]^2}{|x-y|^{d+2s}} \, \mathrm{d}x \, \mathrm{d}y \right)^{1/2} \text{ and}$$

$$\|v\|_{H^r(D)} = \left(\|v\|^2_{H^k(D)} + |v|^2_{H^r(D)} \right)^{1/2}.$$

The Sobolev space $H^r(D)$ is then defined as the space of functions v in $H^k(D)$ such that the integral $|v|^2_{H^r(D)}$ is finite. For $0 < s \leq 1$, the space $H^{-s}(D)$ denotes the dual space to $H_0^s(D)$ with the dual norm.

The space of essentially bounded measurable functions is denoted by $L^\infty(D)$,

with the norm defined as

$$\|v\|_{L^\infty(D)} = \operatorname*{ess\,sup}_{x\in D} |v(x)|.$$

In a similar fashion, we define for $k \in \mathbb{N}$ the Sobolev space $W^{k,\infty}(D)$ containing all functions having weak derivatives of order $|\alpha| \leq k$ in $L^\infty(D)$,

$$W^{k,\infty}(D) = \{v \in L^\infty(D) : D^\alpha v \in L^\infty(D) \text{ for } |\alpha| \leq k\}.$$

We define the following semi-norm and norm on $W^{k,\infty}(D)$:

$$|v|_{W^{k,\infty}(D)} = \max_{|\alpha|=k} \operatorname*{ess\,sup}_{x\in D} |D^\alpha v(x)| \quad \text{and} \quad \|v\|_{W^{k,\infty}(D)} = \max_{0\leq|\alpha|\leq k} \operatorname*{ess\,sup}_{x\in D} |D^\alpha v(x)|.$$

In addition to the above Sobolev spaces, we also make use of Hölder spaces. For $k \in \mathbb{N} \cup \{0\}$, $C^k(\overline{D})$ denotes the space of continuous functions which are k times continuously differentiable, with semi–norm and norm

$$|v|_{C^k(\overline{D})} = \max_{|\alpha|=k} \sup_{x\in\overline{D}} |D^\alpha v(x)| \quad \text{and} \quad \|v\|_{C^k(\overline{D})} = \sum_{0\leq|\alpha|\leq k} \sup_{x\in\overline{D}} |D^\alpha v(x)|$$

For any real $r > 0$, with $r \notin \mathbb{N}$, we set $r = k + s$, and define the following semi-norm and norm

$$|v|_{C^r(\overline{D})} = \max_{|\alpha|=k} \sup_{x,y\in\overline{D}:\, x\neq y} \frac{|D^\alpha v(x) - D^\alpha v(y)|}{|x-y|^s} \quad \text{and} \quad \|v\|_{C^r(\overline{D})} = \|v\|_{C^k(\overline{D})} + |v|_{C^r(\overline{D})}.$$

Similarly, we define $|\cdot|_{C^k(\overline{D},\mathbb{R}^{d\times d})}$ and $|\cdot|_{C^r(\overline{D},\mathbb{R}^{d\times d})}$, for k and r as above, by

$$\|V\|_{C^k(\overline{D},\mathbb{R}^{d\times d})} = \sum_{0\leq|\alpha|\leq k} \sup_{x\in\overline{D}} \|D^\alpha V(x)\|_{d\times d},$$

and

$$|V|_{C^r(\overline{D},\mathbb{R}^{d\times d})} = \max_{|\alpha|=k} \sup_{x,y\in\overline{D}:\, x\neq y} \frac{\|D^\alpha V(x) - D^\alpha V(y)\|_{d\times d}}{|x-y|^s},$$

where $\|\cdot\|_{d\times d}$ denotes a suitable matrix norm on $\mathbb{R}^{d\times d}$.

Finally, we will also require spaces of Bochner integrable functions. To this end, let $\mathcal{B}$ be a separable Banach space with norm $\|\cdot\|_{\mathcal{B}}$, and $v : \Omega \to \mathcal{B}$ be

measurable. With the norm $\|\cdot\|_{L^p(\Omega,\mathcal{B})}$ defined by

$$\|v\|_{L^p(\Omega,\mathcal{B})} = \begin{cases} \left(\int_\Omega \|v\|_{\mathcal{B}}^p \,\mathrm{d}\mathbb{P}\right)^{1/p}, & \text{for} \quad p < \infty, \\ \operatorname{ess\,sup}_{\omega\in\Omega} \|v\|_B, & \text{for} \quad p = \infty, \end{cases}$$

the space $L^p(\Omega,\mathcal{B})$ is defined as the space of all strongly measurable functions on which this norm is finite. In particular, we denote by $L^p(\Omega, H_0^k(D))$ the space where the norm on $H_0^k(D)$ is chosen to be the seminorm $|\cdot|_{H^k(D)}$. For simplicity we write $L^p(\Omega)$ for $L^p(\Omega,\mathbb{R})$.

A key task in this thesis is to keep track of how the constants in the bounds and estimates depend on the coefficient $\mathbf{A}(\omega, x)$ and on the mesh size h. Hence, we will almost always be stating constants explicitly. Constants that do not depend on $\mathbf{A}(\omega, x)$ or h will not be explicitly stated. Instead, we will write $b \lesssim c$ for two positive quantities b and c, if b/c is uniformly bounded by a constant independent of $\mathbf{A}(\omega, x)$ and of h.

Chapter 2

Regularity

The convergence rate of numerical methods is usually governed by the regularity of the function being approximated. The smoother the function is, the better it can be approximated by piecewise polynomials. To rigorously prove convergence of numerical methods, it is essential to establish the regularity of the problem under consideration. This section is therefore devoted to a study of the regularity of model problems such as (1.2) in Section 1.1.

Given a probability space $(\Omega, \mathcal{E}, \mathbb{P})$ and $\omega \in \Omega$, we consider the following linear elliptic partial differential equation (PDE) with random coefficients, posed on a bounded, Lipschitz polygonal/polyhedral domain $D \subset \mathbb{R}^d$, $d = 1, 2, 3$, and subject to Dirichlet boundary conditions: Find $u : \Omega \times D \to \mathbb{R}$ such that

$$\begin{aligned} -\mathrm{div}\,(\mathbf{A}(\omega, x)\nabla u(\omega, x)) &= f(\omega, x), \qquad \text{for } x \in D, \\ u(\omega, x) &= \phi_j(\omega, x), \qquad \text{for } x \in \Gamma_j\,. \end{aligned} \tag{2.1}$$

The differential operators div and ∇ are with respect to $x \in D$, and $\Gamma := \cup_{j=1}^m \overline{\Gamma}_j$ denotes the boundary of D, partitioned into straight line segments in 2D and into planar polygonal panels in 3D. We assume that the boundary conditions are compatible, i.e. $\phi_j(x) = \phi_k(x)$, if $x \in \overline{\Gamma}_j \cap \overline{\Gamma}_k$. We also let $\phi \in H^1(D)$ be an extension of the boundary data $\{\phi_j\}_{j=1}^m$ to the interior of D whose trace coincides with ϕ_j on Γ_j. The existence of such ϕ is guaranteed by Theorem 2.6 for $\phi_j \in H^{1/2}(\Gamma_j)$. Denote by $H^1_\phi(D)$ the subspace of $H^1(D)$ consisting of functions whose trace on Γ is ϕ.

The restriction to Dirichlet conditions in (2.1) is for ease of presentation only,

and the results in this chapter can be extended to the case of Neumann or mixed Dirichlet/Neumann conditions. It is also possible to include lower order terms in the differential operator, provided these are regular enough (cf assumptions A1-A3).

The coefficient tensor $\mathbf{A}(\omega, \cdot)$ is assumed to take values in the space of real–valued, symmetric $d \times d$ matrices. Given the usual vector norm $|v| := (v \cdot v)^{1/2}$ on $\mathbb{R}^d$, we choose the norm $\|\cdot\|_{d\times d}$ on $\mathbb{R}^{d\times d}$ as the norm induced by $|\cdot|$, or any matrix norm equivalent to it.

For all $\omega \in \Omega$, let now $\mathbf{A}_{\min}(\omega)$ and $\mathbf{A}_{\max}(\omega)$ be such that

$$\mathbf{A}_{\min}(\omega)|\xi|^2 \leq \mathbf{A}(\omega, x)\xi \cdot \xi \leq \mathbf{A}_{\max}(\omega)|\xi|^2, \tag{2.2}$$

for all $\xi \in \mathbb{R}^d$, uniformly in $x \in D$. If the trajectories of $\mathbf{A}$ are continuous, appropriate choices are

$$\mathbf{A}_{\min}(\omega) := \min_{x\in\overline{D}} \|\mathbf{A}^{-1}(\omega, x)\|_{d\times d}^{-1}, \text{ and } \mathbf{A}_{\max}(\omega) := \max_{x\in\overline{D}} \|\mathbf{A}(\omega, x)\|_{d\times d}. \tag{2.3}$$

In the special case of scalar coefficients $\mathbf{A}(\omega, x) = a(\omega, x)I_d$, for some $a : \Omega \times D \to \mathbb{R}$, we will denote $\mathbf{A}_{\min}(\omega)$ and $\mathbf{A}_{\max}(\omega)$ by $a_{\min}(\omega)$ and $a_{\max}(\omega)$, respectively. The quantities in (2.2) in this case reduce to

$$a_{\min}(\omega) := \min_{x\in\overline{D}} a(\omega, x), \text{ and } a_{\max}(\omega) := \max_{x\in\overline{D}} a(\omega, x).$$

We make the following assumptions on the input data:

A1. $\mathbf{A}_{\min} > 0$ almost surely and $1/\mathbf{A}_{\min} \in L^p(\Omega)$, for all $p \in [1, \infty)$.

A2. $\mathbf{A} \in L^p(\Omega, C^t(\overline{D}, \mathbb{R}^{d\times d}))$, for some $0 \leq t \leq 1$ and for all $p \in [1, \infty)$.

A3. $f \in L^{p_*}(\Omega, H^{t-1}(D))$ and $\phi_j \in L^{p_*}(\Omega, H^{t+1/2}(\Gamma_j))$, for all $j = 1, \ldots, m$ and for some $p_* \in [1, \infty]$, with t as in A2.

The Hölder continuity of $\mathbf{A}$ in assumption A2 implies that the quantities in (2.3) are well defined, that $\mathbf{A}_{\max} = \|\mathbf{A}\|_{C^0(\overline{D},\mathbb{R}^{d\times d})} \in L^p(\Omega)$ and (together with assumption A1) that $0 < \mathbf{A}_{\min}(\omega) < \mathbf{A}_{\max}(\omega) < \infty$, for almost all $\omega \in \Omega$. Note that assumption A2 also implies that $\mathbf{A}_{i,j} \in C^t(\overline{D})$ almost surely, for $i, j \in 1, \ldots, d$, with $\|\mathbf{A}_{i,j}(\omega, \cdot)\|_{C^t(\overline{D})} \leq \|\mathbf{A}(\omega, \cdot)\|_{C^t(\overline{D},\mathbb{R}^{d\times d})}$.

We will here not make the assumption that we can bound $\mathbf{A}_{\min}(\omega)$ away from zero and $\mathbf{A}_{\max}(\omega)$ away from infinity, uniformly in ω, and shall instead work with the quantities $\mathbf{A}_{\min}(\omega)$ and $\mathbf{A}_{\max}(\omega)$ directly. As we will see in Remark 2.13, we could even weaken assumptions A1 and A2 and assume that $\|\mathbf{A}\|_{C^t(\overline{D},\mathbb{R}^{d\times d})}$ and $1/\mathbf{A}_{\min}$ have only a finite number of bounded moments, i.e. $0 < p \leq p_a$, for some fixed $p_a > 0$, but in order not to complicate the presentation we did not choose to do this.

We will study the PDE (2.1) in weak (or variational) form, for fixed $\omega \in \Omega$. This is not possible uniformly in Ω, since we have not assumed uniform boundedness of $\mathbf{A}_{\min}(\omega)$ and $\mathbf{A}_{\max}(\omega)$, but it is possible almost surely. In the following we will not explicitly write this each time. With $f(\omega,\cdot) \in H^{t-1}(D)$ and $0 < \mathbf{A}_{\min}(\omega) \leq \mathbf{A}_{\max}(\omega) < \infty$, for all $x \in D$, the variational formulation of (2.1), parametrised by $\omega \in \Omega$, is

$$b_\omega\big(u(\omega,\cdot), v\big) = L_\omega(v)\,, \quad \text{for all} \quad v \in H^1_0(D), \tag{2.4}$$

where the bilinear form b_ω and the linear functional L_ω (both parametrised by $\omega \in \Omega$) are defined as usual, for all $u, v \in H^1(D)$, by

$$b_\omega(u,v) := \int_D \mathbf{A}(\omega,x)\nabla u(x)\cdot\nabla v(x)\,\mathrm{d}x \quad \text{and} \tag{2.5}$$

$$L_\omega(v) := \langle f(\omega,\cdot), v\rangle_{H^{t-1}(D),H^{1-t}_0(D)}\,. \tag{2.6}$$

We say that for any $\omega \in \Omega$, $u(\omega,\cdot)$ is a weak solution of (2.1) iff $u(\omega,\cdot) \in H^1_\phi(D)$ and satisfies (2.4).

The following result is classical. It is based on the Lax-Milgram Lemma (see e.g. [44]).

Lemma 2.1. *For almost all $\omega \in \Omega$, the bilinear form b_ω is bounded and coercive in $H^1_0(D)$ with respect to $|\cdot|_{H^1(D)}$, with constants $\mathbf{A}_{\max}(\omega)$ and $\mathbf{A}_{\min}(\omega)$, respectively. Moreover, there exists a unique solution $u(\omega,\cdot) \in H^1_\phi(D)$ to the variational problem* (2.4)*, with*

$$\|u(\omega,\cdot)\|_{H^1(D)} \lesssim \frac{\|f(\omega,\cdot)\|_{H^{-1}(D)} + \mathbf{A}_{\max}(\omega)\|\phi\|_{H^1(D)}}{\mathbf{A}_{\min}(\omega)} =: C_{2.1}(\omega).$$

The following proposition is a direct consequence of Lemma 2.1.

Theorem 2.2. *Let assumptions A1-A3 hold with $t = 0$. Then the weak solution u of (2.1) is unique and belongs to*
$L^p(\Omega, H^1_\phi(D))$, for all $p < p_*$.

Proof. First note that $u : \Omega \to H^1_\phi(D)$ is measurable, since u is a continuous function of $\mathbf{A}$. The result then follows directly from Lemma 2.1 and assumptions A1- A3, together with Hölder's inequality. □

The aim of this chapter is now to establish more (spatial) regularity of the solution u. This is necessary to prove convergence of numerical approximations of (2.1) in chapter 3. To finish this introductory section, let us give a summary of the main results in this chapter. Detailed proofs are provided later on.

For smooth domains $D \subset \mathbb{R}^d$, for any $d \geq 1$, the regularity of u depends solely on the regularity of the input data $\mathbf{A}$, $\{\phi_j\}_{j=1}^m$ and f. With t and p_* as in assumptions A1-A3, we will prove in Theorem 2.7 that $u \in L^p(\Omega, H^{1+s}(D))$, for any $s < t$ and $p < p_*$. If $t = 1$, we have $u \in L^p(\Omega, H^2(D))$.

In the case of Lipschitz polygonal/polyhedral domains, the regularity of u depends on the regularity of the input data $\mathbf{A}$, $\{\phi_j\}_{j=1}^m$ and f, as well as on the geometry of D, so we need the following definition in addition to assumptions A1-A3.

Definition 2.3. *Let $0 < \lambda_\Delta(D) \leq 1$ be such that for any $0 < s \leq \lambda_\Delta(D), s \neq \frac{1}{2}$, the Laplace operator Δ is surjective as an operator from $H^{1+s}(D) \cap H^1_0(D)$ to $H^{s-1}(D)$. In other words, let $\lambda_\Delta(D)$ be no larger than the order of the strongest singularity of the Laplace operator with homogeneous Dirichlet boundary conditions on D.*

The number $\lambda_\Delta(D)$ exists for any Lipschitz polygonal/polyhedral domain, see e.g. [42, Remarks 2.4.6 and 2.6.7]. We will come back to specific values of $\lambda_\Delta(D)$ in section 2.3. For convex domains, we have $\lambda_\Delta(D) = 1$.

Combining the regularity result for smooth domains with an analysis of the corner singularities in D, we will prove in Theorem 2.12 that for any Lipschitz polygonal domain $D \subset \mathbb{R}^2$, we have $u \in L^p(\Omega, H^{1+s}(D))$, for any $s < t$ such that $s \leq \lambda_\Delta(D)$ and any $p < p_*$. If $t = \lambda_\Delta(D) = 1$, we again have $u \in L^p(\Omega, H^2(D))$.

In the special case of scalar coefficients and Lipschitz polygonal domains $D \subset \mathbb{R}^2$, we further extend the regularity results to coefficients which are only

piecewise Hölder continuous with respect to a (possibly random) partitioning of D. The regularity result from Theorem 2.12 in this case applies locally on each subdomain. However, global regularity is limited due to the interfaces. If all subdomains are convex, and no more than two subdomains meet at any point in $\overline{D}$, then it follows from Theorem 2.17 that $u \in L^p(\Omega, H^{1+s}(D))$, for any $s < \min(t, 1/2)$ and $p < p_*$.

The tools we will use to prove the above results are classical. However, since we did not assume uniform ellipticity or boundedness of b_ω, it is essential that we track exactly how the constants appearing in the regularity estimates depend on $\mathbf{A}$.

The structure of the remainder of this chapter is as follows. In §2.1 we give some preliminary estimates which will be useful in the subsequent analysis. We then begin the regularity analysis by proving a regularity result for elliptic problems posed on smooth domains in §2.2. We extend this result to polygonal domains in §2.3 by analysing the corner singularities of u. In §2.4, we further extend the regularity results by relaxing assumption A2 and considering coefficients which are only piecewise Hölder continuous. In §2.5, we analyse some dual problems which will later be used in chapter 3 to prove optimal convergence rates for functionals. §2.6 gives a regularity result in the framework of Hölder spaces. Finally, in §2.7, we give examples of random fields which satisfy the assumptions needed for the regularity results in the earlier sections.

2.1 Preliminary estimates

In this section we present a collection of results on functions in Hölder and Sobolev spaces which will be frequently used in the regularity analysis in the remainder of this chapter.

Lemma 2.4. *Let $D \subset \mathbb{R}^d$, and let s, t be such that either $0 < s < t < 1$ or $s = t = 1$. If $b \in C^t(\overline{D})$ and $v \in H^s(D)$, then $bv \in H^s(D)$ and*

$$\|bv\|_{H^s(D)} \lesssim |b|_{C^t(\overline{D})} \|v\|_{L^2(D)} + \|b\|_{C^0(\overline{D})} \|v\|_{H^s(D)}.$$

The hidden constants depend only on t, s and d.

Proof. This is a classical result, but we require the exact dependence of the bound

on b. First note that trivially $\|bv\|_{L^2(D)} \leq \|b\|_{C^0(\overline{D})}\|v\|_{L^2(D)}$. The case $s = t = 1$ follows from the product rule, since

$$\begin{aligned}|bv|^2_{H^1(D)} &= \int_D |\nabla(bv)|^2 \,\mathrm{d}x = \int_D |b\nabla v + v\nabla b|^2 \,\mathrm{d}x \\ &\lesssim (|b|_{C^1(\overline{D})}\,\|v\|_{L^2(D)} \,+\, \|b\|_{C^0(\overline{D})}\,\|v\|_{H^1(D)})^2.\end{aligned}$$

For the case $0 < s < t < 1$, we have, for any $x, y \in D$,

$$|b(x)v(x) - b(y)v(y)|^2 \leq 2(b(x)^2|v(x) - v(y)|^2 + v(y)^2|b(x) - b(y)|^2).$$

Denoting by $\tilde{v}$ the extension of v by 0 on $\mathbb{R}^d$, this implies

$$\begin{aligned}&\iint_{D^2} \frac{|b(x)v(x) - b(y)v(y)|^2}{\|x-y\|^{d+2s}} \,\mathrm{d}x\,\mathrm{d}y \\ &\leq\; 2\|b\|^2_{C^0(\overline{D})}\|v\|^2_{H^s(D)} + 2\iint_{D^2} \frac{v(y)^2|b(x)-b(y)|^2}{\|x-y\|^{d+2s}} \\ &\leq\; 2\|b\|^2_{C^0(\overline{D})}\|v\|^2_{H^s(D)} + \\ &\quad \iint_{\substack{x,y\in D\\ \|x-y\|\geq 1}} 8\|b\|^2_{C^0(\overline{D})}\frac{v(y)^2}{\|x-y\|^{d+2s}} \;+\; 2|b|^2_{C^t(\overline{D})}\frac{v(y)^2}{\|x-y\|^{d+2(s-t)}}\,\mathrm{d}x\,\mathrm{d}y \\ &\leq\; 2\|b\|^2_{C^0(\overline{D})}\|v\|^2_{H^s(D)} + \\ &\quad \left(8\|b\|^2_{C^0(\overline{D})}\left\|\frac{\mathbb{1}_{\|z\|\geq 1}}{\|z\|^{d+2s}}\right\|_{L^1(\mathbb{R}^d)} + \;2|b|^2_{C^t(\overline{D})}\left\|\frac{\mathbb{1}_{\|z\|\leq 1}}{\|z\|^{d+2(s-t)}}\right\|_{L^1(\mathbb{R}^d)}\right)\|\tilde{v}\|^2_{L^2(\mathbb{R}^d)} \\ &\lesssim\; \|b\|^2_{C^0(\overline{D})}\|v\|^2_{H^s(D)} \;+\; |b|^2_{C^t(\overline{D})}\|v\|^2_{L^2(D)}.\end{aligned}$$

The result then follows. □

Lemma 2.5. *Let $D \subset \mathbb{R}^d$ and $0 < s < 1$, $s \neq 1/2$. Then*

$$|||v|||_s := \left(\|v\|^2_{L^2(D)} + \sum_{i=1}^d \left\|\frac{\partial v}{\partial x_i}\right\|^2_{H^{s-1}(D)}\right)^{1/2}$$

defines a norm on $H^s(D)$ that is equivalent to $\|v\|_{H^s(D)}$.

Proof. This is [44, Lemma 9.1.12]. □

Note that for the case $s = 1$, the equivalence of the norms defined in Lemma 2.5 follows directly from the definition of $\|v\|_{H^1(D)}$.

Theorem 2.6. *Let D be a Lipschitz domain with boundary $\Gamma := \bigcup_{j=1}^m \Gamma_j$, with each Γ_j of class C^2.*

a) Let $1/2 < r \leq 1$. For each $w \in H^{r-1/2}(\Gamma)$, there exists an extension $\tilde{w} \in H^r(D)$ with trace on Γ equal to w and $\|\tilde{w}\|_{H^r(D)} \lesssim \|w\|_{H^{r-1/2}(\Gamma)}$, where the hidden constant is independent of w.

b) Let $1 < r \leq 2$ and $D \subset \mathbb{R}^2$. Suppose $w_j \in H^{r-1/2}(\Gamma_j)$ and $w_j(x) = w_k(x)$ for $x \in \overline{\Gamma}_j \cap \overline{\Gamma}_k$. Then there exists an extension $\tilde{w} \in H^r(D)$ with trace on Γ_j equal to w_j and $\|\tilde{w}\|_{H^r(D)} \lesssim \sum_{j=1}^m \|w_j\|_{H^{r-1/2}(\Gamma_j)}$, where the hidden constant is independent of $w_j, j = 1, \ldots, m$.

Proof. This follows directly from the assumptions that Γ is Lipschitz continuous and Γ_j, for $j = 1, \cdots, m$, is of class C^2, together with [44, Theorem 6.2.40], [42, section 1.4] and the Sobolev embedding Theorem [1]. □

2.2 Regularity in smooth domains

We start by proving a regularity result for problems posed on smooth domains. Since the proofs are rather long and technical, we give here only the main ideas of the proofs. Full proofs can be found in appendix A. The main result in this section is the following.

Theorem 2.7. *Suppose D is a C^2 domain, and consider the boundary value problem*

$$\begin{aligned} -\text{div}\,(\mathbf{A}(\omega, x)\nabla w(\omega, x)) &= f(\omega, x), \qquad \text{for } x \in D, \\ w(\omega, x) &= 0, \qquad \text{for } x \in \partial D\,. \end{aligned} \tag{2.7}$$

Let assumptions A1-A2 hold with $0 < t \leq 1$, and suppose $f \in L^{p_*}(\Omega, H^{t-1}(D))$, for some $p_* \in (0, \infty]$. Then $w(\omega, \cdot) \in H^{1+s}(D)$ and

$$\|w(\omega, \cdot)\|_{H^{1+s}(D)} \lesssim C_{2.7}(\omega),$$

for almost all $\omega \in \Omega$ *and all* $0 < s < t$*, where*

$$C_{2.7}(\omega) := \frac{\mathbf{A}_{\max}(\omega)\|\mathbf{A}(\omega,\cdot)\|_{C^t(\overline{D},\mathbb{R}^{d\times d})}}{\mathbf{A}_{\min}(\omega)^3}\,\|f(\omega,\cdot)\|_{H^{t-1}(D)}.$$

Moreover $w \in L^p(\Omega, H^{1+s}(D))$*, for all* $p < p_*$*. If the assumptions hold with* $t = 1$*, then* $w \in L^p(\Omega, H^2(D))$ *and the above bound holds with* $s = 1$*.*

The proof of Proposition 2.7 consists of three main steps, and follows the proof in Hackbusch [44]. We formulate the first two steps as separate lemmas and then give the final step following these lemmas. We fix $\omega \in \Omega$, and to simplify the notation we do not specify the dependence on ω anywhere in the proof. We will only consider the case $0 < t < 1$ in detail. A full proof in the case of scalar coefficients with $t = 1$ can be found in [9].

In the first step of the proof we consider the case $D = \mathbb{R}^d$.

Lemma 2.8. *Let* $0 < t < 1$ *and* $D = \mathbb{R}^d$*, and let* $\mathbf{T} = (\mathbf{T}_{ij})_{i,j=1}^d$ *be a symmetric, uniformly positive definite* $d \times d$ *matrix-valued function from* $\mathbb{R}^d$ *to* $\mathbb{R}^{d\times d}$*, i.e. there exists* $\mathbf{T}_{\min} > 0$ *such that* $\mathbf{T}(x)\xi \cdot \xi \geq \mathbf{T}_{\min}|\xi|^2$ *uniformly in* $x \in \mathbb{R}^d$ *and* $\xi \in \mathbb{R}^d$*, and let* $\mathbf{T}_{ij} \in C^t(\mathbb{R}^d)$*, for all* $i, j = 1, \ldots, d$*. Consider*

$$-\mathrm{div}(\mathbf{T}(x)\nabla w(x)) = F(x), \qquad \text{for } \; x \in \mathbb{R}^d, \tag{2.8}$$

with $F \in H^{s-1}(\mathbb{R}^d)$*, for some* $0 < s < t$*. Any weak solution* $w \in H^1(\mathbb{R}^d)$ *of* (2.8) *is in* $H^{1+s}(\mathbb{R}^d)$ *and*

$$\|w\|_{H^{1+s}(\mathbb{R}^d)} \lesssim \frac{1}{\mathbf{T}_{\min}}\Big(|\mathbf{T}|_{C^t(\mathbb{R}^d,\mathbb{R}^{d\times d})}\,|w|_{H^1(\mathbb{R}^d)} + \|F\|_{H^{s-1}(\mathbb{R}^d)}\Big) \;+\; \|w\|_{H^1(\mathbb{R}^d)}.$$

Proof. This is essentially [44, Theorem 9.1.8] with the dependence on $\mathbf{T}$ made explicit, and it can be proved using the representation of the norm on $H^{1+s}(\mathbb{R}^d)$ via Fourier coefficients, as well as a fractional difference operator R_h^i, $i = 1, \ldots, d$, on a Cartesian mesh with mesh size $h > 0$ (similar to the classical Nirenberg translation method for proving H^2 regularity). For a definition of R_h^i and more details see [44, Theorem 9.1.8] or section A.1 in the appendix. □

The second step consists in treating the case where $D = \mathbb{R}^d_+ := \{y = (y_1, ..., y_d) : y_d > 0\}$.

Lemma 2.9. *Let $0 < t < 1$ and $D = \mathbb{R}^d_+$, and let $\mathbf{T} : D \to \mathbb{R}^{d\times d}$ be as in Lemma 2.8. Consider now (2.8) on $D = \mathbb{R}^d_+$ subject to $w = 0$ on ∂D with $F \in H^{t-1}(\mathbb{R}^d_+)$. Then any weak solution $w \in H^1(\mathbb{R}^d_+)$ of this problem is in $H^{1+s}(\mathbb{R}^d_+)$ and*

$$\|w\|_{H^{1+s}(\mathbb{R}^d_+)} \lesssim \frac{\mathbf{T}_{\max}}{\mathbf{T}^2_{\min}}\Big(|\mathbf{T}|_{C^t(\overline{\mathbb{R}^d_+},\mathbb{R}^{d\times d})}\,|w|_{H^1(\mathbb{R}^d_+)}+\|F\|_{H^{t-1}(\mathbb{R}^d_+)}\Big)+\frac{\mathbf{T}_{\max}}{\mathbf{T}_{\min}}\,\|w\|_{H^1(\mathbb{R}^d_+)}\,.$$

Proof. This is essentially [44, Theorem 9.1.11] with the dependence on $\mathbf{T}$ made explicit. It uses Lemmas 2.4 and 2.5. For details see [44, Theorem 9.1.11] or section A.2 in the appendix. □

Proof of Theorem 2.7. We are now ready to prove Theorem 2.7 using Lemmas 2.8 and 2.9. The third and last step consists in using a covering of D by $r+1$ bounded regions $(D_i)_{0\le i\le r}$, such that

$$\overline{D}_0 \subset D, \quad \overline{D} \subset \bigcup_{i=0}^{r} D_i \quad \text{and} \quad \partial D = \bigcup_{i=1}^{r}(D_i \cap \partial D).$$

Using a (non-negative) partition of unity $\{\chi_i\}_{0\le i\le r} \subset C^\infty(\mathbb{R}^d)$ subordinate to this cover, it is possible to reduce the proof to bounding $\|\chi_i u\|_{H^{1+s}(D)}$, for all $0 \le i \le r$.

For $i = 0$ this reduces to an application of Lemma 2.8 with w and F chosen to be extensions by 0 from D to $\mathbb{R}^d$ of $\chi_0 u$ and of $f\chi_0 + \mathbf{A}\nabla u\cdot\nabla\chi_0 + \operatorname{div}(\mathbf{A}u\nabla\chi_0)$, respectively. The tensor $\mathbf{T}$ is $\overline{\mathbf{A}}(x)$, where $\overline{\mathbf{A}}$ is a smooth extension of $\mathbf{A}(x)$ on D_0 to $\mathbf{A}_{\min}I_d$ on $\mathbb{R}^d\backslash D$, and so $\mathbf{T}_{\min} \gtrsim \mathbf{A}_{\min}$ and $\|\mathbf{T}\|_{C^t(\mathbb{R}^d,\mathbb{R}^{d\times d})} \lesssim \|\mathbf{A}\|_{C^t(\mathbb{R}^d,\mathbb{R}^{d\times d})}$.

For $1 \le i \le r$, the proof reduces to an application of Lemma 2.9. As for $i = 0$, we can see that $\chi_i u \in H^1_0(D \cap D_i)$ is the weak solution of the problem $-\operatorname{div}(\mathbf{A}\nabla u_i) = f_i$ on $D \cap D_i$ with $f_i := f\chi_i + \mathbf{A}\nabla u\cdot\nabla\chi_i + \operatorname{div}(\mathbf{A}u\nabla\chi_i)$. To be able to apply Lemma 2.9 to the weak form of this PDE, we define now a twice continuously differentiable bijection α_i (with α_i^{-1} also in C^2) from D_i to the cylinder

$$Q_i := \{y = (y_1,\dots,y_d) : |(y_1,\dots,y_{d-1})| < 1 \ \text{ and } \ |y_d| < 1\},$$

such that $D_i\cap D$ is mapped to $Q_i\cap\mathbb{R}^d_+$, and $D_i\cap\partial D$ is mapped to $Q_i\cap\{y : y_d = 0\}$.

We use α_i^{-1} to map all the functions defined above on $D_i \cap D$ to $Q_i \cap \mathbb{R}^d_+$, and then extend them suitably to functions on $\mathbb{R}^d_+$ to finally apply Lemma 2.9. The tensor $\mathbf{T}$ in this case depends on the mapping α_i. However, since ∂D was assumed to be C^2, we get $\mathbf{T}_{\min} \gtrsim \mathbf{A}_{\min}$, $\mathbf{T}_{\max} \lesssim \mathbf{A}_{\max}$ and $\|\mathbf{T}\|_{C^t(\mathbb{R}^d, \mathbb{R}^{d\times d})} \lesssim \|\mathbf{A}\|_{C^t(\mathbb{R}^d, \mathbb{R}^{d\times d})}$, with hidden constants that only depend on α_i, α_i^{-1} and their Jacobians. For details see [44, Theorem 9.1.16] or section A.3 in the appendix. □

2.3 Corner singularities

For polygonal/polyhedral domains D, the solution u can have singularities near the non–smooth parts of the boundary Γ, i.e. near corners in 2D and near corners and edges in 3D. These singularities can reduce the global regularity of u, and hence need to be analysed. However, we will see that under assumptions A1-A2, this question can be reduced to analysing the singularities of the Laplace operator on D. We will follow [41, §5.2], and as in the previous section we will again establish the result first path wise, almost surely in $\omega \in \Omega$. The key technicality will again be to track how the constants in all the necessary estimates, in particular in the semi-Fredholm property of the underlying random differential operator, depend on ω.

As in [41, §5.2], for simplicity we actually consider D to be a piecewise C^2 domain and restrict ourselves to $\mathbb{R}^2$. We again write the boundary Γ as $\Gamma = \cup_{j=1}^m \Gamma_j$, where now in 2D each Γ_j is an open arc of curve of class C^2, and $\overline{\Gamma}_j$ meets $\overline{\Gamma}_{j+1}$ at S_j (where we identify Γ_{m+1} and Γ_1). We consider only domains with boundaries that are rectilinear near the corners, which of course includes Lipschitz polygonal/polyhedral domains. This means that at each corner S_j, we can find a polygonal domain $W_j \subset D$ such that the boundary ∂W_j coincides with Γ near S_j.

For a given $\omega \in \Omega$, with $0 < \mathbf{A}_{\min}(\omega) \leq \mathbf{A}_{\max}(\omega) < \infty$, we define the differential operator

$$A_\omega v = -\mathrm{div}(\mathbf{A}(\omega, \cdot)\nabla v)).$$

The following key result, which is based on [51, §4, Theorem 5.26], is proved via a homotopy method in the proof of [41, Lemma 5.2.5], for $s = 1$. The proof for $s < 1$ is analogous.

Lemma 2.10. *Let $m = 1$ and $\omega \in \Omega$. If $0 < s \leq \lambda_\Delta(D)$ and if there exists $C_{\text{semi}}(\omega) > 0$ such that*

$$\|v\|_{H^{1+s}(D)} \leq C_{\text{semi}}(\omega)\|A_\omega v\|_{H^{s-1}(D)}, \quad \text{for all } v \in H^{1+s}(D) \cap H_0^1(D), \tag{2.9}$$

then A_ω is surjective from $H^{1+s}(D) \cap H_0^1(D)$ to $H^{s-1}(D)$.

Thus, if we can establish (2.9), which essentially means that A_ω is semi-Fredholm as an operator from $H^{1+s}(D) \cap H_0^1(D)$ to $H^{s-1}(D)$, for some $s \leq \lambda_\Delta(D)$, then we can also conclude on the regularity of solutions of the stochastic variational problem (2.1). The following lemma essentially follows [41, Lemma 5.2.3]. However, in the case of a random coefficient, we crucially need to make sure that the constant $C_{\text{semi}}(\omega)$ in (2.9) has sufficiently many moments as a random field on Ω. To ensure this we need to carefully track the dependence on $\mathbf{A}$ in the bounds in [41, Lemma 5.2.5].

Lemma 2.11. *Let $m \in \mathbb{N}$ and let assumptions A1 –A2 hold for some $0 < t \leq 1$. Then (2.9) holds for all $0 < s < t$ s.t. $s \leq \lambda_\Delta(D)$, $s \neq \frac{1}{2}$, with*

$$C_{\text{semi}}(\omega) \;=\; \frac{\mathbf{A}_{\max}(\omega)\|\mathbf{A}(\omega,\cdot)\|^2_{C^t(\overline{D},\mathbb{R}^{2\times 2})}}{\mathbf{A}_{\min}(\omega)^4} := C_{2.11}(\omega). \tag{2.10}$$

In the case $t = \lambda_\Delta(D) = 1$, (2.9) also holds for $s = 1$, i.e. for the $H^2(D)$-norm.

Proof. We first consider the case where $m = 1$ and $t = \lambda_\Delta(D) = 1$. For ease of notation, we suppress the dependence on ω in the coefficient, and denote A_ω simply by A and $\mathbf{A}(\omega, x)$ by $\mathbf{A}(x)$. Furthermore, we denote by A_1 the operator A with coefficients frozen at S_1, i.e. $A_1 v = -\text{div}\,(\mathbf{A}(S_1)\nabla v)$.

We will prove (2.9) by combining the regularity results for A in C^2 domains with regularity results of the constant coefficient operator A_1 on polygonal domains. Since we assume that Γ is rectilinear near S_1, we can find a polygonal domain W such that $W \subset D$ and ∂W coincides with Γ near S_1. Let $v \in H^2(D) \cap H_0^1(D)$ and let η be a smooth cut-off function with support in W, such that $\eta \equiv 1$ near S_1, and then consider ηv and $(1-\eta)v$ separately. We start with ηv.

Let $w \in H^2(W) \cap H_0^1(W)$. We first establish the estimate

$$\|w\|_{H^2(W)} \lesssim \frac{1}{\mathbf{A}_{\min}} \|A_1 w\|_{L^2(W)}. \tag{2.11}$$

A proof of this estimate in the special case where $A_1 = -\Delta$ can be found in [42]. We will follow the same steps. Firstly, by the Poincaré inequality, we have that

$$\|w\|^2_{H^2(W)} \lesssim |w|^2_{H^2(W)} + |w|^2_{H^1(W)}. \tag{2.12}$$

Using integration by parts and the fact that $w = 0$ on ∂W, we further have

$$\mathbf{A}_{\min} |w|^2_{H^1(W)} \lesssim \int_W \mathbf{A}(S_1)\nabla w \cdot \nabla w \, \mathrm{d}x = \int_W w \nabla \cdot (\mathbf{A}(S_1)\nabla w) \, \mathrm{d}x \tag{2.13}$$

and so via the Cauchy-Schwarz and the Poincaré inequalities

$$|w|_{H^1(W)} \lesssim \frac{1}{\mathbf{A}_{\min}} \|A_1 w\|_{L^2(W)}. \tag{2.14}$$

It remains to prove a bound on $|w|^2_{H^2(W)}$. This is easily done by noting that

$$\begin{aligned} |w|^2_{H^2(W)} &:= \int_W \left(\sum_{1 \le i,j \le d} \frac{\partial^2 w}{\partial x_i \partial x_j} \right)^2 \mathrm{d}x \\ &\lesssim \frac{1}{\mathbf{A}^2_{\min}} \int_W \left(\sum_{1 \le i,j \le d} \mathbf{A}_{i,j}(S_1) \frac{\partial^2 w}{\partial x_i \partial x_j} \right)^2 \mathrm{d}x \\ &= \frac{1}{\mathbf{A}^2_{\min}} \|A_1 w\|^2_{L^2(W)}. \end{aligned} \tag{2.15}$$

The estimate (2.11) now follows from (2.12)–(2.15), with the hidden constant only depending on the shape of W. Using (2.11), we have

$$\begin{aligned} \mathbf{A}_{\min} \|w\|_{H^2(W)} &\lesssim \|Aw\|_{L^2(W)} + \|Aw - A_1 w\|_{L^2(W)} \\ &= \|Aw\|_{L^2(W)} + \|\mathrm{div}(\mathbf{A}(\cdot) - \mathbf{A}(S_1))\nabla w\|_{L^2(W)} \\ &= \|Aw\|_{L^2(W)} + \left\| \sum_{i,j=1}^{2} \frac{\partial}{\partial x_i} \big(\mathbf{A}_{i,j}(\cdot) - \mathbf{A}_{i,j}(S_1) \frac{\partial w}{\partial x_j} \big) \right\|_{L^2(W)}. \end{aligned}$$

Now, using Lemma 2.4 we get

$$\begin{aligned}&\mathbf{A}_{\min}\,\|w\|_{H^2(W)}\\&\leq C\,\left(\|Aw\|_{L^2(W)}+|\mathbf{A}|_{C^1(\overline{W},\mathbb{R}^{2\times 2})}|w|_{H^1(W)}+\|\mathbf{A}-\mathbf{A}(S_1)\|_{C^0(\overline{W},\mathbb{R}^{2\times 2})}|w|_{H^2(W)}\right).\end{aligned}\tag{2.16}$$

Denote now by C the best constant such that (2.16) holds. Since $\mathbf{A}$ was assumed to be in $C^1(\overline{W},\mathbb{R}^{2\times 2})$, we can choose W (and hence the support of η) small enough so that

$$C\|\mathbf{A}-\mathbf{A}(S_1)\|_{C^0(\overline{W},\mathbb{R}^{2\times 2})}\leq\frac{1}{2}\,\mathbf{A}_{\min}\,.\tag{2.17}$$

By replacing $\mathbf{A}(S_1)$ by $\mathbf{A}$ in (2.13), one can show $|w|_{H^1(W)}\lesssim\|Aw\|_{L^2(W)}/\mathbf{A}_{\min}$. Substituting this and (2.17) into (2.16) and using $\mathbf{A}_{\min}\leq\mathbf{A}_{\max}$ we have

$$\begin{aligned}\mathbf{A}_{\min}\,\|w\|_{H^2(W)}&\leq 2C\left(1+\frac{|\mathbf{A}|_{C^1(\overline{W},\mathbb{R}^{2\times 2})}}{\mathbf{A}_{\min}}\right)\|Aw\|_{L^2(W)}\\&\lesssim\frac{\|\mathbf{A}\|_{C^1(\overline{W},\mathbb{R}^{2\times 2})}}{\mathbf{A}_{\min}}\|Aw\|_{L^2(W)}\,.\end{aligned}\tag{2.18}$$

Since $v\in H^2(D)\cap H_0^1(D)$ and W contains the support of η, we have $\eta v\in H^2(W)\cap H_0^1(W)$ and so estimate (2.18) applies to ηv. Thus

$$\|\eta v\|_{H^2(D)}\lesssim\frac{\|\mathbf{A}\|_{C^1(\overline{W},\mathbb{R}^{2\times 2})}}{\mathbf{A}_{\min}^2}\|A(\eta v)\|_{L^2(W)}\,.$$

Let us move on to $(1-\eta)v$. Let $D'\subset D$ be a C^2 domain that coincides with D outside of the region where $\eta=1$. This is always possible due to our assumptions on the geometry of D near S_1. Then $(1-\eta)v\in H^2(D')\cap H_0^1(D')$, and using Theorem 2.7 we have

$$\|(1-\eta)v\|_{H^2(D)}\lesssim\frac{\mathbf{A}_{\max}\|\mathbf{A}\|_{C^1(\overline{D}',\mathbb{R}^{2\times 2})}}{\mathbf{A}_{\min}^3}\,\|A\left((1-\eta)v\right)\|_{L^2(D')}\,.$$

Adding the last two estimates together and using the triangle inequality, we have

$$\|v\|_{H^2(D)}\;\lesssim\;\frac{\|\mathbf{A}\|_{C^1(\overline{D},\mathbb{R}^{2\times 2})}}{\mathbf{A}_{\min}^2}\left(\|A(\eta v)\|_{L^2(W)}\;+\;\frac{\mathbf{A}_{\max}}{\mathbf{A}_{\min}}\,\|A((1-\eta)v)\|_{L^2(D')}\right).\tag{2.19}$$

It remains to bound the term in the bracket on the right hand side of (2.19) in terms of $\|Av\|_{L^2(D)}$. Note that

$$A(\eta v) = \eta(Av) + 2\mathbf{A}\nabla\eta \cdot \nabla v + (A\eta)v\,.$$

Thus, applying the triangle inequality and using the fact that η was assumed to be smooth with $0 \le \eta \le 1$, we get

$$\|A(\eta v)\|_{L^2(W)} \lesssim \|Av\|_{L^2(W)} + \mathbf{A}_{\max}|v|_{H^1(W)} + \|\mathbf{A}\|_{C^1(\overline{D},\mathbb{R}^{2\times 2})}\|v\|_{L^2(W)}\,. \tag{2.20}$$

The hidden constant depends on $\|\nabla\eta\|_{L^\infty(W)}$ and on $\|\Delta\eta\|_{L^\infty(W)}$. Finally using Poincaré's inequality on all of D, as well as an elliptic estimate similar to (2.14) for v, i.e. $|v|_{H^1(D)} \le \|Av\|_{L^2(D)}/\mathbf{A}_{\min}$, leads to

$$\|A(\eta v)\|_{L^2(W)} \lesssim \frac{\|\mathbf{A}\|_{C^1(\overline{D},\mathbb{R}^{2\times 2})}}{\mathbf{A}_{\min}}\|Av\|_{L^2(D)}\,.$$

Substituting this and the corresponding bound for $\|A((1-\eta)v)\|_{L^2(D')}$ into (2.19), we finally get

$$\|v\|_{H^2(D)} \;\lesssim\; \frac{\mathbf{A}_{\max}\|\mathbf{A}\|^2_{C^1(\overline{D},\mathbb{R}^{2\times 2})}}{\mathbf{A}^4_{\min}}\,\|Av\|_{L^2(D)}\,,$$

for all $v \in H^2(D) \cap H^1_0(D)$. This completes the proof for the case $m = 1$ and $t = \lambda_\Delta = 1$.

The proof for $t < 1$ and/or $\lambda_\Delta(D) < 1$ follows exactly the same lines. As in (2.14), one can prove the inequality $\|w\|_{H^1(W)} \lesssim \|A_1 w\|_{H^{-1}(W)}/\mathbf{A}_{\min}$. Together with (2.11) and an interpolation argument, this gives the estimate

$$\|w\|_{H^{1+s}(W)} \lesssim \frac{1}{\mathbf{A}_{\min}}\,\|A_1 w\|_{H^{s-1}(W)}\,, \tag{2.21}$$

where the hidden constant again only depends on W. Using Lemma 2.5, one can derive the following equivalent of (2.16), for any $\frac{1}{2} \neq s < t$:

$$\begin{aligned}&\mathbf{A}_{\min}\,\|w\|_{H^{1+s}(W)}\\ &\lesssim\; \|Aw\|_{H^{s-1}(W)} + |\mathbf{A}|_{C^t(\overline{W},\mathbb{R}^{2\times 2})}|w|_{H^1(W)} + \|\mathbf{A} - \mathbf{A}(S_1)\|_{C^0(\overline{W},\mathbb{R}^{2\times 2})}\|\nabla w\|_{H^s(W)}\,.\end{aligned}$$

As before, the $|w|_{H^1(W)}$ term can be bounded using (2.13), Hölder's inequality and the Poincaré inequality:

$$\begin{aligned}\mathbf{A}_{\min}\,|w|^2_{H^1(W)} &\le \|w\|_{H^{1-s}(W)}\|Aw\|_{H^{s-1}(W)} \le \|w\|_{H^1(W)}\|Aw\|_{H^{s-1}(W)}\\ &\lesssim |w|_{H^1(W)}\|Aw\|_{H^{s-1}(W)}\,.\end{aligned}$$

The remainder of the proof requires only minor modifications.

The case $m > 1$ is treated by repeating the above procedure with a different cut–off function η_j at each corner S_j. Estimate (2.18) applies to $\eta_j v$, for all $j = 1,\dots,m$, and the regularity estimate in Theorem 2.7 applies to $(1-\sum_{j=1}^n \eta_j)v$. □

We are now ready to prove regularity for Lipschitz polygonal domains $D \subset \mathbb{R}^2$.

Theorem 2.12. *Let assumptions A1-A3 hold for some $0 < t \le 1$, and let $D \subset \mathbb{R}^2$ be Lipschitz polygonal. Then $u(\omega,\cdot) \in H^{1+s}(D)$ and*

$$\|u(\omega,\cdot)\|_{H^{1+s}(D)} \lesssim C_{2.12}(\omega),$$

for almost all $\omega \in \Omega$ and for all $0 < s < t$ such that $s \le \lambda_\Delta(D)$, where

$$C_{2.12}(\omega) := \frac{\mathbf{A}_{\max}(\omega)\|\mathbf{A}(\omega,\cdot)\|^2_{C^t(\overline{D},\mathbb{R}^{2\times 2})}}{\mathbf{A}_{\min}(\omega)^4} \quad \times$$
$$\left[\|f(\omega,\cdot)\|_{H^{t-1}(D)} + \|\mathbf{A}(\omega,\cdot)\|_{C^t(\overline{D},\mathbb{R}^{2\times 2})}\sum_{j=1}^m \|\phi_j(\omega,\cdot)\|_{H^{t+1/2}(\Gamma_j)}\right].$$

Moreover, $u \in L^p(\Omega, H^{1+s}(D))$, for all $p < p_*$. If $t = \lambda_\Delta(D) = 1$, then $u \in L^p(\Omega, H^2(D))$ and the above bound holds with $s = 1$.

Proof. Let us first consider the case $\phi \equiv 0$. In this case, the fact that $u \in H^{1+s}(D) \cap H^1_0(D)$ and the bound on $\|u\|_{H^{1+s}(D)}$ follow immediately from Lemmas 2.10 and 2.11, for any $s < t$ and $s \le \lambda_\Delta(D)$, as well as for $s = 1$ if $t = \lambda_\Delta(D) = 1$, since $f = Au$.

The case $\phi \neq 0$ now follows from Theorem 2.6. We will only show the proof for $t = \lambda_\Delta(D) = 1$ in detail. Due to assumption A3 we can choose $\phi \in H^2(D)$ with $\|\phi\|_{H^2(D)} \lesssim \sum_{j=1}^m \|\phi_j\|_{H^{3/2}(\Gamma_j)}$, and so $f_0 := f - A\phi \in L_2(D)$. Since $u_0 := u - \phi \in H^1_0(D)$ we can apply the result we just proved for the case $\phi \equiv 0$ to the problem $Au_0 = f_0$ to get

$$\begin{aligned}\|u_0\|_{H^2(D)} &\lesssim C_{2.11}(\omega)\left(\|Au_0\|_{L^2(D)} + \|A\phi\|_{L^2(D)}\right)\\ &\lesssim C_{2.11}(\omega)\left(\|f\|_{L^2(D)} + \|\mathbf{A}\|_{C^1(\overline{D},\mathbb{R}^{2\times 2})}\|\phi\|_{H^2(D)}\right),\end{aligned}$$

where in the last step we have used Lemma 2.5. The bound $C_{2.12}$ then follows by Minkowski's inequality. That $u \in L^p(\Omega, H^2(D))$, for any $p < p_*$, then follows from assumptions A1-A3, together with Minkowski's and Hölder's inequality, since

$$\begin{aligned}&\|C_{2.12}\|_{L^p(\Omega)}\\ &\leq \|C_{2.11}\|_{L^{p_1}(\Omega)}\|f\|_{L^{p_*}(\Omega,L^2(D))} + \left\|C_{2.11}\,\|\mathbf{A}\|_{C^1(\overline{D},\mathbb{R}^{2\times 2})}\right\|_{L^{p_1}(\Omega)}\sum_{j=1}^{m}\|\phi\|_{L^{p*}(\Omega,H^{3/2}(D))},\end{aligned}$$

where $p_1 = \frac{p_*-p}{p_*p}$. □

Remark 2.13. (a) In order to establish $u \in L^p(\Omega, H^{1+s}(D))$, for some fixed $1 \leq p < p_*$, it would have been sufficient to assume that the constants $C_{2.11}$ and $C_{2.11}\,\|\mathbf{A}\|_{C^t(\overline{D},\mathbb{R}^{2\times 2})}$ in Theorem 2.12 are in $L^q(\Omega)$, for $q = \frac{p_*-p}{p_*p}$. In the case $p_* = \infty$, $q = p$ is sufficient, which in turn implies that we can weaken assumption A1 to $1/\mathbf{A}_{\min} \in L^q(\Omega)$ with $q = 4p$, or assumption A2 to $\mathbf{A} \in L^q(\Omega, C^t(\overline{D}, \mathbb{R}^{2\times 2}))$ with $q = 4p$, or both assumptions to L^q with $q = 8p$.

(b) In the case $\lambda_\Delta(D) = 1$, it is also possible to weaken assumption A2 to $\mathbf{A} \in L^p(\Omega, C^{0,1}(\overline{D}, \mathbb{R}^{2\times 2}))$, i.e. to assume only Lipschitz continuity instead of differentiability of the trajectories of $\mathbf{A}$, and still conclude $u \in L^p(\Omega, H^2(D))$.

Remark 2.14. (a) The behaviour of the Laplace operator near corners is described in detail in [41, 42]. In particular, in the pure Dirichlet case for convex domains we always get $\lambda_\Delta(D) = 1$. For non-convex domains $\lambda_\Delta(D) = \min_{j=1}^m \pi/\theta_j$, where θ_j is the angle at corner S_j. Hence, $\lambda_\Delta(D) \geq 1/2$ for any Lipschitz polygonal domain.

(b) In a similar manner one can prove regularity of u also in the case of Neumann and mixed Dirichlet/Neumann boundary conditions provided the boundary conditions are compatible. For example, in order to apply the same proof technique used here at a point where a Dirichlet and a homogeneous Neumann boundary meet, we can first reflect the problem and the

solution across the Neumann boundary. Then we apply the above theory on the union of the original and the reflected domain. The regularity for the Laplacian is in general lower in the mixed Dirichlet/Neumann case than in the pure Dirichlet case. In particular, full regularity (i.e. $\lambda_\Delta(D) = 1$) is only possible, if all angles where the type of boundary condition changes are less than $\pi/2$. For an arbitrary Lipschitz polygonal domain we can only guarantee $\lambda_\Delta(D) \geq 1/4$.

(c) The 3D case is similar, but in addition to singularities at corners (for which the analysis is identical to the above) we also need to consider edge singularities. This is a bit more involved and we refer to [41, §8.2.1] for more details. However, provided D is convex, we obtain again $\lambda_\Delta(D) = 1$ always in the pure Dirichlet case.

2.4 Transmission problems

We now shift our attention to more general random coefficients. In practice, one is often interested in models with discontinuous coefficients, e.g. modelling different rock strata in the subsurface. Such coefficients do not satisfy assumption A2, and the regularity results from Theorem 2.12 can not be applied directly. However, the loss of regularity is confined to the interface between different strata, and it is still possible to prove a limited amount of regularity even globally.

Since the results in this section again crucially make use of regularity results for operators with (piecewise) constant coefficients, and the known results in this area are mostly restricted to the case of scalar coefficients, we will for the remainder of this section assume that $\mathbf{A}(\omega, x) = a(\omega, x)I_d$, for some scalar random field $a(\omega, x) : \Omega \times \overline{D} \to \mathbb{R}$. Let us consider (2.1) on a Lipschitz polygonal domain $D \subset \mathbb{R}^2$ that can be decomposed into disjoint Lipschitz polygonal subdomains D_k, $k = 1, \ldots, K$. Let $PC^t(\overline{D}) \subset L^\infty(D)$ denote the space of piecewise C^t functions with respect to the partition $\{D_k\}_{k=1}^K$ (up to the boundary of each region D_k). We replace assumption A2 by the following milder assumption on the coefficient function a:

A2*. $a \in L^p(\Omega, PC^t(\overline{D}))$, for some $0 < t \leq 1$ and for all $p \in [1, \infty)$.

Our regularity results for discontinuous coefficients rely on the following result from [41, 58].

Lemma 2.15. *Let $v \in H^1(D)$ and $s < 1/2$, and suppose that $v \in H^{1+s}(D_k)$, for all $k = 1, \ldots, K$. Then $v \in H^{1+s}(D)$ and*

$$\|v\|_{H^{1+s}(D)} = \|v\|_{H^1(D)} + \sum_{k=1}^{K} |v|_{H^{1+s}(D_k)} .$$

The proof of this result uses the fact that for $0 \leq s < 1/2$, $w \in H^s(D_k)$ if and only if the extension $\tilde{w}$ of w by zero is in $H^s(\mathbb{R}^d)$. Thus, we cannot expect more than $H^{3/2-\delta}(D)$ regularity globally in the discontinuous case. However, as in the case of continuous fields, the regularity of the solution will also depend on the parameter t in assumptions A2* and A3, as well as on the behaviour of the operator A_ω at any singular points. Since Lemma 2.15 restricts us to $s < 1/2$ and since $\lambda_\Delta(D) \geq 1/2$ for any Lipschitz polygonal $D \subset \mathbb{R}^2$ in the case of a pure Dirichlet problem, we do not have to worry about corners. Instead we define the set of *singular* (or *cross*) *points* $\mathcal{S}^\times := \{S_\ell^\times : \ell = 1, \ldots, L\}$ to consist of all points $S_\ell^\times$ in D where three or more subdomains meet, as well as all those points $S_\ell^\times$ on ∂D where two or more subdomains meet. By the same arguments as in section 2.3, the behaviour of A_ω at these singular points is again fully described by studying transmission problems for the Laplace operator, i.e. elliptic problems with piecewise constant coefficients, locally near each singular point (cf. [57, 17, 58]).

Definition 2.16. Denote by $T(\alpha_1, \ldots, \alpha_K)$ the operator corresponding to the transmission problem for the Laplace operator with (constant) material parameter α_k on subdomain D_k, $k = 1, \ldots, K$. Let $0 \leq \lambda_T(D) < 1/2$ be such that $T(\alpha_1, \ldots, \alpha_K)$ is a surjective operator from $H^{1+s}(D) \cap H_0^1(D)$ to $H^{s-1}(D)$, for any choice of $\alpha_1, \ldots, \alpha_K$ and for $s \leq \lambda_T(D)$. In other words, $\lambda_T(D)$ is a bound on the order of the strongest singularity of $T(\alpha_1, \ldots, \alpha_K)$.

Without any assumptions on the partition $\{D_k\}_{k=1}^K$ or any bounds on the constants $\{\alpha_k\}_{k=1}^K$ it is in general not possible to choose $\lambda_T(D) > 0$. However, if no more than three regions meet at every interior singular point and no more than two at every boundary singular point, then we can choose $0 < \lambda_T(D) \leq$

1/4. If in addition each of the subregions D_k is convex, then we can choose any $0 < \lambda_T(D) < 1/2$, which due to the restrictions in Lemma 2.15 is the maximum we can achieve anyway. See for example [57, 17, 58] for details.

The following is an analogue of Theorem 2.12 on the regularity of the solution u of (2.1) for piecewise C^t coefficients.

Theorem 2.17. *Let $D \subset \mathbb{R}^2$ be a Lipschitz polygonal domain and let $\lambda_T(D) > 0$. Suppose assumptions A1, A2* and A3 hold with $0 < t \leq 1$. Then, the solution u of (2.1) is in $L^p(\Omega, H^{1+s}(D))$, for any $0 < s < t$ such that $s \leq \lambda_T(D)$ and for all $p < p_*$.*

Proof. Let us first consider $\phi \equiv 0$ again. Then, the existence of a unique solution $u(\omega, \cdot) \in H^1(D)$ of (2.1) follows again from the Lax-Milgram Lemma, for almost all $\omega \in \Omega$. Also note that restricted to D_k the transmission operator $T(\alpha_1, \ldots, \alpha_K) = \alpha_k \Delta$, for all $k = 1, ..., K$. Therefore, using assumption A2* we can prove as in section 2.3 via a homotopy method that $u(\omega, \cdot)$ restricted to D_k is in $H^{1+s}(D_k)$, for any $s < t$ and $s \leq \lambda_T(D)$, for almost all $\omega \in \Omega$. The result then follows from Lemma 2.15 and an application of Hölder's inequality. The case $\phi \not\equiv 0$ follows as in the proof to Theorem 2.12 via a trace estimate. □

Remark 2.18. The results in this section can easily be extended to the case where also the partitioning $\{D_k\}_{k=1}^K$ is random (i.e. depends on ω). The value of $\lambda_T(D)$ will in this case be such that $T(\alpha_1, \ldots, \alpha_K)$ is a surjective operator from $H^{1+s}(D) \cap H_0^1(D)$ to $H^{s-1}(D)$, for any choice of $\alpha_1, \ldots, \alpha_K$, for almost all $\omega \in \Omega$ and for $s \leq \lambda_T(D)$. See section 3.6 for an example of such a random partitioning.

2.5 A dual problem

In this section we briefly discuss some dual problems to the variational problem (2.4), and show how the regularity results from the previous sections apply in this case. The dual problems will be used in chapter 3 to prove optimal convergence rates of the discretisation error.

Denote by $M_\omega : H^1(D) \to \mathbb{R}$ some measurable functional on $H^1(D)$. Like the bilinear form $b_\omega(\cdot, \cdot)$, the functional $M_\omega(\cdot)$ is again parametrised by ω, and the analysis is done almost surely in ω. When the functional does not depend on ω, we will simply write M instead of M_ω.

To give the basic idea, let us assume for the moment that M_ω is linear and bounded on $H^1(D)$, i.e. $M_\omega(v) \lesssim \|v\|_{H^1(D)}$, for all $v \in H^1(D)$. Now, let us associate with our *primal problem* (2.4) the following *dual problem*: find $z(\omega,\cdot) \in H_0^1(D)$ such that

$$b_\omega\big(v, z(\omega,\cdot)\big) = M_\omega(v)\,, \quad \text{for all} \quad v \in H_0^1(D),$$

where the bilinear form $b_\omega(\cdot,\cdot)$ again is as in (2.5). Since M_ω is linear and bounded, we can apply the Lax-Milgram Lemma to ensure existence and uniqueness of a weak solution $z(\omega,\cdot) \in H_0^1(D)$, for almost all ω.

For nonlinear functionals, following [59], the dual problem is not defined as above. Instead, a different functional is chosen on the right hand side (which reduces to M_ω in the linear case). It is related to the derivative of the functional of interest and so we need to assume a certain differentiability of M_ω. We will assume here that M_ω is continuously Fréchet differentiable. In particular, this implies that M_ω is also Gateaux differentiable, with the two derivatives being the same.

Let $v, w \in H^1(D)$. Then the Gateaux derivative of M_ω at w and in the direction v is defined as

$$D_v M_\omega(w) := \lim_{\varepsilon\to 0} \frac{M_\omega(w+\varepsilon v) - M_\omega(w)}{\varepsilon}.$$

For $w_1, w_2 \in H^1(D)$, we define

$$\overline{D_v M_\omega}(w_1, w_2) := \int_0^1 D_v M_\omega(w_1 + \theta(w_2 - w_1))\,\mathrm{d}\theta,$$

which is in some sense an average derivative of M_ω on the path from w_1 to w_2, and define the dual problem now as: find $z(\omega,\cdot) \in H_0^1(D)$ such that

$$b_\omega\big(v, z(\omega,\cdot)\big) = \overline{D_v M_\omega}(w_1, w_2), \quad \text{for all} \quad v \in H_0^1(D). \tag{2.22}$$

Note that, for any linear functional M_ω, we have $\overline{D_v M_\omega}(w_1, w_2) = M_\omega(v)$, for all $v \in H_0^1(D)$.

For our further analysis, we need to make the following assumption on M_ω.

F1. Let $w_1(\omega), w_2(\omega) \in H^1(D)$. Let M_ω be continuously Fréchet differentiable,

and suppose that there exists $t_* \in [0,1]$, $q_* \in [1,\infty]$ and $C_{\mathrm{F1}} \in L^{q_*}(\Omega)$, such that

$$|\overline{D_v M_\omega}(w_1(\omega), w_2(\omega))| \lesssim C_{\mathrm{F1}}(\omega)\|v\|_{H^{1-t_*}(D)},$$

for all $v \in H_0^1(D)$ and for almost all $\omega \in \Omega$.

The random variable C_{F1} in assumption F1 may depend on ω through the functional M_ω, as well as $w_1(\omega)$ and $w_2(\omega)$.

As in the linear case, it is sufficient to assume that $|\overline{D_v M_\omega}(w_1, w_2)|$ is bounded in $H^1(D)$, i.e. that assumption F1 holds with $t_* = 0$, to ensure existence and uniqueness of the dual solution $z(\omega,\cdot) \in H_0^1(D)$, for almost all $\omega \in \Omega$. As in Lemma 2.1, we then have

$$\|z(\omega,\cdot)\|_{H^1(D)} \lesssim \frac{C_{\mathrm{F1}}(\omega)}{\mathbf{A}_{\min}(\omega)},$$

for almost all $\omega \in \Omega$. However, in order to apply Theorem 2.12 to prove stronger spatial regularity for z, we need to assume boundedness of $|\overline{D_v M_\omega}(w_1, w_2)|$ in $H^{1-t_*}(D)$, for some $t_* > 0$. In particular, if assumptions A1–A3 and F1 are satisfied with $t \in (0,1]$ and $t_* \in (0,1]$, then by Theorem 2.12 we have for almost all $\omega \in \Omega$,

$$\|z(\omega,\cdot)\|_{H^{1+s}(D)} \lesssim \frac{\mathbf{A}_{\max}(\omega)\|\mathbf{A}(\omega,\cdot)\|^2_{C^t(\overline{D},\mathbb{R}^{d\times d})}}{\mathbf{A}_{\min}(\omega)^4} C_{\mathrm{F1}}(\omega),$$

for any $0 < s < \min(t, t_*)$ such that $s \le \lambda_\Delta(D)$ and for almost all $\omega \in \Omega$.

To finish, let us give some examples of output functionals which fit into the framework described above. We start with linear functionals.

(a) **Point evaluations:** Since $\mathbf{A}(\omega,\cdot) \in C^t(\overline{D}, \mathbb{R}^{d\times d})$, we know that trajectories of the solution u are in $C^{1+t}(\overline{D})$ (see e.g. [31]), and it is meaningful to consider point values. Consider $M^{(1)}(u) := u(x^*)$, for some $x^* \in D$. For $D \subset \mathbb{R}$, i.e. in one space dimension, we have the compact embedding $H^{1/2+\delta}(D) \hookrightarrow C^\delta(\overline{D})$, for any $\delta > 0$, and so

$$M^{(1)}(v) = v(x^*) \;\le\; \|v\|_{C^0(\overline{D})} \;\lesssim\; \|v\|_{H^{1/2+\delta}(D)}, \quad \text{for all} \;\; v \in H^1(D).$$

Hence, assumption F1 is satisfied for any $t_* < \min(\frac{1}{2}, t)$ with $C_{\mathrm{F1}} = 1$ and

$q_* = \infty$.

In space dimensions higher than one, point evaluation of the pressure u is not a bounded functional on $H^1(D)$. One often regularises this type of functional by approximating the point value by a local average,

$$M^{(2)}(v) := \frac{1}{|D^*|}\int_{D^*} v \,\mathrm{d}x \quad \Big[\approx \; v(x^*) \Big],$$

where D^* is a small subdomain of D that contains x^* [35]. Here, $M^{(2)}$ satisfies F1 with $C_{\mathrm{F1}} = 1$, $t_* = 1$ and $q_* = \infty$, due to the Cauchy-Schwarz inequality.

Similarly, point evaluations of the flux $-\mathbf{A}\nabla v$ can be approximated by a local average. However, in this case F1 only holds for $t_* = 0$ with $C_{\mathrm{F1}}(\omega) = \mathbf{A}_{\max}(\omega)$ and $q_* = \infty$.

Next we give some examples of non–linear functionals. The first obvious example is to estimate higher order moments of linear functionals.

(b) **Second moment of average local pressure:** Let M_ω be an arbitrary linear functional and let $q > 1$. Then

$$D_v\big(M_\omega(\tilde{v})^q\big) = \lim_{\varepsilon\to 0}\frac{M_\omega(\tilde{v}+\varepsilon v)^q - M_\omega(\tilde{v})^q}{\varepsilon} \;=\; qM_\omega(\tilde{v})^{q-1}M_\omega(v).$$

Thus, in case of the second moment of the average local pressure $M^{(3)}(v) := M^{(2)}(v)^2$, this gives

$$D_v M^{(3)}(\tilde{v}) = \frac{2}{|D^*|^2}\left(\int_{D^*} v\,\mathrm{d}x\right)\left(\int_{D^*}\tilde{v}\,\mathrm{d}x\right),$$

and so

$$\begin{aligned}
&|\overline{D_v M^{(3)}}(w_1(\omega), w_2(\omega))| \\
&\quad= \frac{2}{|D^*|^2}\left|\left(\int_{D^*} v\,\mathrm{d}x\right)\left(\int_0^1\int_{D^*}(w_1(\omega)+\theta(w_2(\omega)-w_1(\omega)))\,\mathrm{d}x\mathrm{d}\theta\right)\right| \\
&\quad= \frac{1}{|D^*|^2}\left|\left(\int_{D^*} v\,\mathrm{d}x\right)\left(\int_{D^*} w_1(\omega)+w_2(\omega)\,\mathrm{d}x\right)\right| \\
&\quad\lesssim \left(\|w_1(\omega)\|_{L^2(D)} + \|w_2(\omega)\|_{L^2(D)}\right)\|v\|_{L^2(D)}\,.
\end{aligned}$$

So assumption F1 is satisfied for all $t_* \leq 1$ and with q_* such that $C_{\mathrm{F1}}(\omega) = \big(\|w_1(\omega)\|_{L^2(D)} + \|w_2(\omega)\|_{L^2(D)}\big) \in L^{q_*}(\Omega)$.

(c) **Outflow through boundary:** Consider $M_\omega^{(4)}(v) := L_\omega(\psi) - b_\omega(\psi, v)$, for some given function $\psi \in H^1(D)$. Note that for the solution u of (2.4), by Green's formula, we have

$$\begin{aligned} M_\omega^{(4)}(u) &= \int_D \psi(x) f(x,\omega)\,\mathrm{d}x - \int_D \mathbf{A}(\omega,x)\nabla\psi(x)\cdot\nabla u(\omega,x)\,\mathrm{d}x \\ &= -\int_D \psi(x)\,\nabla\cdot(\mathbf{A}(\omega,x)\nabla u(\omega,x))\;\mathrm{d}x - \int_D \mathbf{A}(\omega,x)\nabla\psi(x)\cdot\nabla u(\omega,x)\,\mathrm{d}x \\ &= -\int_\Gamma \psi(x)\mathbf{A}(\omega,x)\nabla u(\omega,x)\cdot\nu\,\mathrm{d}s\,. \end{aligned}$$

Thus, $M_\omega^{(4)}(u)$ is equal to the outflow through the boundary Γ weighted by ψ, and so $M_\omega^{(4)}$ can be used to approximate the flux through a part $\Gamma_{\mathrm{out}} \subset \Gamma$ of the boundary, by setting $\psi|_{\Gamma_{\mathrm{out}}} \approx 1$ and $\psi|_{\Gamma\setminus\Gamma_{\mathrm{out}}} \approx 0$, see e.g. [3, 26, 35].

Note that for $f \not\equiv 0$ this functional is only affine, not linear. When $f \equiv 0$, then it is linear. In any case,

$$\begin{aligned} D_v M_\omega^{(4)}(\tilde v) &:= \lim_{\varepsilon\to 0} \frac{M_\omega^{(4)}(\tilde v + \varepsilon v) - M_\omega^{(4)}(\tilde v)}{\varepsilon} \\ &= \lim_{\varepsilon\to 0} \frac{-\int_D \mathbf{A}(\omega,x)\nabla\psi(x)\cdot\nabla(\varepsilon v(\omega,x))\,\mathrm{d}x}{\varepsilon} \\ &= -\int_D \mathbf{A}(\omega,x)\nabla\psi(x)\cdot\nabla v(x)\,\mathrm{d}x \\ &= \int_D v(x)\,\nabla\cdot(\mathbf{A}(\omega,x)\nabla\psi(x))\;\mathrm{d}x\,, \end{aligned}$$

for $v, \tilde v \in H_0^1(D)$. Since this is independent of $\tilde v$, we have in particular

$$\overline{D_v M_\omega^{(4)}}(w_1(\omega), w_2(\omega)) = \int_D v(x)\,\nabla\cdot(\mathbf{A}(\omega,x)\nabla\psi(x))\;\mathrm{d}x,$$

for any $w_1(\omega), w_2(\omega) \in H^1(D)$. If we now assume that assumptions A1-A3 are satisfied for some $0 < t \leq 1$ and that $\psi \in H^{1+t}(D)$, then using Lemmas

2.4 and 2.5, we have $\nabla\psi \in H^t(D)$ and for any $t^* < t$,

$$\begin{aligned}|\overline{D_v M_\omega^{(4)}}(w_1(\omega), w_2(\omega))| &\leq \|\nabla\cdot(\mathbf{A}(\omega,\cdot)\nabla\psi)\,\|_{H^{t^*-1}(D)}\|v\|_{H^{1-t^*}(D)}\\ &\lesssim \|\,(\mathbf{A}(\omega,\cdot)\nabla\psi)\,\|_{H^{t^*}(D)}\|v\|_{H^{1-t^*}(D)}\\ &\lesssim \|\mathbf{A}(\omega,\cdot)\|_{C^t(\overline{D},\mathbb{R}^{d\times d})}\|\nabla\psi\|_{H^{t^*}(D)}\|v\|_{H^{1-t^*}(D)}.\end{aligned} \tag{2.23}$$

Hence, assumption F1 is satisfied, for any $q_* < \infty$ and $t_* < t$, with $C_{\mathrm{F1}}(\omega) = \|\mathbf{A}(\omega,\cdot)\|_{C^t(\overline{D},\mathbb{R}^{d\times d})}$. If $t = 1$, then estimate (2.23) holds with $t^* = t = 1$, and assumption F1 is satisfied with $t_* = 1$. Our assumption on ψ is satisfied for example if ψ is linear, which is a suitable choice for the numerical test in the next section.

The functional $\frac{1}{\Gamma_{\mathrm{out}}}\int_{\Gamma_{\mathrm{out}}}\mathbf{A}(\omega,x)\nabla u(\omega,x)\cdot\nu\,\mathrm{d}s$ (or its regularised equivalent over a narrow region near Γ_{out}), which also approximates the flux through Γ_{out}, can only be bounded in $H^1(D)$, and will hence satisfy assumption F1 only with $t_* = 0$.

2.6 Regularity in Hölder spaces

Some error bounds in chapter 3 will require results about the spatial regularity of u in Hölder spaces. In the special case of scalar coefficients $\mathbf{A}(\omega,x) = a(\omega,x)I_d$, for some $a(\omega,x) : \Omega\times\overline{D}\to\mathbb{R}$, the following result was proved in [11].

Proposition 2.19. *Suppose $D\subset\mathbb{R}^d$ is a C^2–domain, and consider the boundary value problem*

$$\begin{aligned}-\mathrm{div}\,(a(\omega,x)\nabla w(\omega,x)) &= f(\omega,x), \qquad \text{for } x\in D,\\ w(\omega,x) &= 0, \qquad \text{for } x\in\partial D\,.\end{aligned}$$

Let assumptions A1-A2 hold with $0 < t < 1$, and suppose $f\in L^{p_*}(\Omega,L^q(D))$, for some $p_*\in(0,\infty]$ and $q > d/(1-t)$, $q\geq 2$. Then, $w(\omega,\cdot)\in C^{1+t}(\overline{D})$ for almost all $\omega\in\Omega$, and

$$\|w(\omega,\cdot)\|_{C^{1+t}(\overline{D})} \lesssim C_{2.19}(\omega),$$

where

$$C_{2.19}(\omega) := \|a(\omega,\cdot)\|_{C^t(\overline{D})}^{\frac{2+p+8t-(2+p)t^2}{2t(1-t)}} \, a_{\min}(\omega)^{-\frac{2+p+(2-p)t-2t^2}{2t(1-t)}} \, \|f(\omega,\cdot)\|_{L^q(D)}.$$

It follows that $u \in L^p(\Omega, C^{1+t}(\overline{D}))$*, for any* $p < p_*$*.*

Note that it is in general not true that $u(\omega,\cdot) \in C^2(\overline{D})$ if $a(\omega,\cdot) \in C^1(\overline{D})$ and $f(\omega,\cdot) \in C^0(\overline{D})$.

To conclude on the Hölder regularity of solutions to problems posed on polygonal domains, one would again have to analyse possible corner singularities, as is done in section 2.3. The regularity of u will again depend on t, as well as the angles in D. In particular, we have $u(\omega,\cdot) \in C^{1+t}(\overline{D})$, for almost all $\omega \in \Omega$, if all the angles are less than $\pi/(1+t)$ (see [31, Theorem 6.2.10]).

Alternatively, one can use the Sobolev embedding theorem (see e.g. [1]), to conclude from Theorem 2.12 that $u(\omega,\cdot) \in C^s(\overline{D})$, for any $s < t$ such that $s \leq \lambda_\Delta(D)$, with $\|u(\omega,\cdot)\|_{C^s(\overline{D})} \leq C_{2.12}(\omega)$. However, this approach does not lead to sharp bounds on the Hölder regularity of u.

2.7 Log–normal random fields

A coefficient of particular interest in subsurface flow applications of (2.1) is a scalar log-normal random field $a(\omega,x)$, where $a(\omega,x) = \exp[g(\omega,x)]$, with $g : \Omega \times \overline{D} \to \mathbb{R}$ denoting a Gaussian field. We consider homogeneous Gaussian fields with Lipschitz continuous covariance kernel

$$C(x,y) := \mathbb{E}\Big[(g(\omega,x) - \mathbb{E}[g(\omega,x)])(g(\omega,y) - \mathbb{E}[g(\omega,y)])\Big] = k\big(\|x-y\|\big), \quad (2.24)$$

for some $k \in C^{0,1}(\mathbb{R}^+)$ and some norm $\|\cdot\|$ in $\mathbb{R}^d$.

With this type of covariance function, it follows from Kolmogorov's Theorem [60] that, for all $t < 1/2$, the trajectories of g belong to $C^t(\overline{D})$ almost surely. More precisely, Kolmogorov's Theorem ensures the existence of a version $\tilde{g}$ of g (i.e. for any $x \in D$, we have $g(\cdot,x) = \tilde{g}(\cdot,x)$ almost surely) such that $\tilde{g}(\omega,\cdot) \in C^t(\overline{D})$, for almost all $\omega \in \Omega$. In particular, we have for almost all ω, that $g(\omega,\cdot) = \tilde{g}(\omega,\cdot)$ almost everywhere. We will identify g with $\tilde{g}$ in what follows.

Built on the Hölder continuity of the trajectories of g and using Fernique's

Theorem [60], it was shown in [10] that assumption A1 holds and that $a \in L^p(\Omega, C^0(\overline{D}))$, for all $p \in [1, \infty)$. We will here prove that assumption A2 holds for log–normal random fields, i.e. that $a \in L^p(\Omega, C^t(\overline{D}))$, for some $0 < t \leq 1$ and for all $p \in [1, \infty)$.

Lemma 2.20. *Let g be a Gaussian field with covariance* (2.24)*. Then the trajectories of the log-normal field $a = \exp[g]$ belong to $C^t(\overline{D})$ almost surely, for all $t < 1/2$, and*

$$\|a(\omega, \cdot)\|_{C^t(\overline{D})} \leq \left(1 + 2\,|g(\omega, \cdot)|_{C^t(\overline{D})}\right) a_{\max}(\omega)\,.$$

Proof. Fix $\omega \in \Omega$ and $t < 1/2$. Since the trajectories of g belong to $C^t(\overline{D})$ almost surely, we have

$$\begin{aligned} |e^{g(\omega,x)} - e^{g(\omega,y)}| &\leq |g(\omega, x) - g(\omega, y)| \left(e^{g(\omega,x)} + e^{g(\omega,y)}\right) \\ &\leq 2\, a_{\max}(\omega)\, |g(\omega)|_{C^t(\overline{D})}\, |x - y|^t. \end{aligned}$$

for any $x, y \in \overline{D}$. Now, $a_{\max}(\omega)\, |g(\omega)|_{C^t(\overline{D})} < \infty$ almost surely, and so the result follows by taking the supremum over all $x, y \in \overline{D}$. □

Lemma 2.20 can in fact be generalised from the exponential function to any smooth function of g.

Proposition 2.21. *Let g be a mean zero Gaussian field with covariance* (2.24)*. Then assumptions A1–A2 are satisfied for the log-normal field $a = \exp[g]$ with any $t < \frac{1}{2}$.*

Proof. Clearly by definition $a_{\min} \geq 0$. The proof that $1/a_{\min} \in L^p(\Omega)$, for all $p \in (0, \infty)$, is based on an application of Fernique's Theorem [60] and can be found in [10, Proposition 2.3]. To prove assumption A2 note that, for all $t < 1/2$ and $p \in (0, \infty)$, $g \in L^p(\Omega, C^t(\overline{D}))$ (cf. [10, Proposition 3.8]) and $a_{\max} \in L^p(\Omega)$ (cf. [10, Proposition 2.3]). Thus the result follows from Lemma 2.20 and an application of Hölder's inequality. □

The results in Proposition 2.21 can be extended to log-normal random fields for which the underlying Gaussian field $g(\omega, x)$ does not have mean zero, under the assumption that this mean is sufficiently regular. Adding a mean $\mu(x)$ to g,

we have $a(\omega, x) = \exp[\mu(x)] \exp[g(\omega, x)]$, and assumptions A1-A2 are satisfied if $\mu(x) \in C^t(\overline{D})$.

A typical example of a covariance function used in practice (cf. [49]) that is only Lipschitz continuous is the exponential covariance function given by

$$k(\|x - y\|_p) \;=\; \sigma^2 \exp\left(-\|x - y\|_p/\lambda\right), \qquad x, y \in \overline{D}, \tag{2.25}$$

where $\|\cdot\|_p$ denotes the ℓ_p-norm in $\mathbb{R}^d$ and typically $p = 1$ or 2. The parameters σ^2 and λ denote the *variance* and the *correlation length*, respectively, and in subsurface flow applications typically only $\sigma^2 \geq 1$ and $\lambda \leq \operatorname{diam} D$ will be of interest. Smoother covariance functions, such as the Gaussian covariance kernel

$$k(\|x - y\|_p) = \sigma^2 \exp(-\|x - y\|_p^2/\lambda^2), \qquad x, y \in \overline{D}, \tag{2.26}$$

which is analytic on $\overline{D} \times \overline{D}$, or more generally the covariance functions in the Matérn class with $\nu > 1$, all lead to $g \in C^1(\overline{D})$ and thus assumption A2 is satisfied for all $t \leq 1$.

As an example of a random coefficient that satisfies assumption A2*, we can consider a piecewise log-normal random field $a = \exp(g)$ such that $g|_{D_k} := g_k$, for all $k = 1, \ldots, K$, where $\{g_k\}_{k=1}^K$ is a set of independent Gaussian random fields. If g_k has mean $\mu_k \in C^{1/2}(\overline{D})$ and exponential covariance function (2.25), then assumption A2* is satisfied for all $t < 1/2$. Similarly, if we let g_k be a Gaussian field with mean $\mu_k \in C^1(\overline{D})$ and Gaussian covariance function (2.26), we have assumption A2* satisfied for any $t \leq 1$. The mean $\mu_k(x)$, the variance σ_k^2 and the correlation length λ_k can be vastly different from one subregion to another.

An example of a random tensor $\mathbf{A}(\omega, x)$ that satisfies assumptions A1 and A2, for all $p \in [1, \infty)$, is a tensor of the form $\mathbf{A} = \exp(g_1)K_1 + \exp(g_2)K_2$, where g_1 and g_2 are scalar Gaussian random fields with a Hölder–continuous mean and a Lipschitz continuous covariance function, and K_1 and K_2 are deterministic tensors satisfying (deterministic versions of) assumptions A1–A2.

Chapter 3

Discretisation Error Analysis

We now turn to a (spatial) discretisation error analysis of the solution u to model problem (2.1). As in the previous chapter, we will first establish results for trajectories of u, i.e. for a fixed $\omega \in \Omega$, and from this deduce estimates for the moments of the error. We will mostly consider approximations using standard, continuous, piecewise linear finite elements on Lipschitz polygonal/polyhedral domains.

Denote by $\{\mathcal{T}_h\}_{h>0}$ a shape-regular family of simplicial triangulations of the Lipschitz polygonal/polyhedral domain D, parametrised by its mesh width $h := \max_{\tau\in\mathcal{T}_h} \operatorname{diam}(\tau)$. Associated with each triangulation $\mathcal{T}_h$ we define the space

$$V_{h,\phi} := \big\{ v_h \in C^0(\overline{D}) \,:\, v_h|_T \text{ linear } \forall T \in \mathcal{T}_h, \text{ and } v_h|_{\Gamma_j} = \phi_j, \text{ for all } j = 1, \ldots, m \big\}$$

of continuous, piecewise linear functions on D that satisfy the boundary conditions in (2.1). For simplicity we assume that the functions $\{\phi_j\}_{j=1}^m$ are piecewise linear with respect to the triangulation $\mathcal{T}_h$ restricted to Γ_j. To deal with more general boundary conditions is a standard exercise in finite element analysis (see e.g. [7, §10.2]).

The finite element approximation of u in $V_{h,\phi}$, denoted by u_h, is now found by solving

$$b_\omega\big(u_h(\omega,\cdot), v_h\big) = L_\omega(v)\,, \quad \text{for all} \quad v_h \in V_{h,0}, \tag{3.1}$$

where the bilinear form $b_\omega(\cdot,\cdot)$ and the linear functional $L_\omega(\cdot)$ are as in (2.5) and (2.6), respectively. Since b_ω is bounded and coercive on $H_0^1(D)$ (cf Lemma 2.1),

we have

$$\|u_h(\omega,\cdot)\|_{H^1(D)} \lesssim C_{2.1}(\omega).$$

The remainder of this chapter is devoted to quantifying the error committed in approximating u by u_h. We make the following assumption.

R1. There exist $0 < s_u \leq 1$ and $1 \leq p_u \leq \infty$ such that for almost all $\omega \in \Omega$,

$$\|u(\omega,\cdot)\|_{H^{1+s_u}(D)} \leq C_{\mathrm{R1}}(\omega),$$

for some $C_{\mathrm{R1}} \in L^{p_u}(\Omega)$.

Assumption R1 holds under reasonable assumptions on $\mathbf{A}$, f and $\{\phi_j\}_{j=1}^m$, as discussed in chapter 2. For examples, see Theorems 2.12 and 2.17.

We start the error analysis by deriving bounds on (moments of) $|u-u_h|_{H^1(D)}$ and $\|u-u_h\|_{L^2(D)}$ in §3.1. In §3.2, we then use these results to prove a bound on $M_\omega(u) - M_\omega(u_h)$, for any continuously differentiable functional M_ω. In §3.3, we establish bounds on $\|u-u_h\|_{W^{1,\infty}(D)}$ and $\|u-u_h\|_{L^\infty(D)}$ using techniques similar to those in §3.1. Since the exact computation of the finite element solution u_h is in general not possible in practice, we analyse some variational crimes, such as quadrature and truncation errors, in §3.4. §3.5 shows how the results from the previous sections can be used to also prove convergence of finite volume methods. Some of the results proved in this chapter are then finally demonstrated numerically in §3.6.

3.1 H^1 and L^2 error estimates

The key tools in proving convergence of the finite element method are Céa's lemma and a best approximation result (cf [7]):

Lemma 3.1 (Céa's Lemma)**.** *Let u and u_h be the solutions to* (2.4) *and* (3.1)*, respectively, and let $0 < \mathbf{A}_{\min}(\omega) \leq \mathbf{A}_{\max}(\omega) < \infty$. Then,*

$$|(u-u_h)(\omega,\cdot)|_{H^1(D)} \leq \left(\frac{\mathbf{A}_{\max}(\omega)}{\mathbf{A}_{\min}(\omega)}\right)^{1/2} \inf_{v_h \in V_{h,\phi}} |u(\omega,\cdot) - v_h|_{H^1(D)}.$$

Lemma 3.2. *Let $v \in H^{1+s}(D) \cap H^1_0(D)$, for some $0 < s \leq 1$. Then*

$$\inf_{v_h \in V_{h,\phi}} |v - v_h|_{H^1(D)} \lesssim \|v\|_{H^{1+s}(D)}\, h^s,$$

where the hidden constant is independent of v and h.

An estimate of the error $|u - u_h|_{H^1(D)}$ now follows.

Theorem 3.3. *Let assumptions A1–A2 hold with $t = 0$, and let assumption R1 hold for some $0 < s_u \leq 1$ and $1 \leq p_u \leq \infty$. Then, for all $p < p_u$ and $h > 0$, we have*

$$\|u - u_h\|_{L^p(\Omega, H^1_0(D))} \lesssim C_{3.3}\, h^{s_u}, \quad \text{with } C_{3.3} := \left\| \left(\frac{\mathbf{A}_{\max}}{\mathbf{A}_{\min}} \right)^{1/2} \right\|_{L^q(\Omega)} \|C_{\mathrm{R1}}\|_{L^{p_u}(\Omega)},$$

where $q = \frac{p_u p}{p_u - p}$.

Proof. It follows directly from Lemmas 3.2 and 3.1, together with assumption R1, that for almost all $\omega \in \Omega$,

$$|(u - u_h)(\omega, \cdot)|_{H^1(D)} \lesssim \left(\frac{\mathbf{A}_{\max}(\omega)}{\mathbf{A}_{\min}(\omega)} \right)^{1/2} C_{\mathrm{R1}}(\omega)\, h^{s_u}. \tag{3.2}$$

The claim of the Theorem then follows from assumptions A1-A2 and R1, together with Hölder's inequality:

$$\begin{aligned}
\|u - u_h\|_{L^p(\Omega, H^1_0(D))} &\lesssim \left\| \left(\frac{\mathbf{A}_{\max}}{\mathbf{A}_{\min}} \right)^{1/2} C_{\mathrm{R1}}\, h^{s_u} \right\|_{L^p(\Omega)} \\
&\lesssim \left\| \left(\frac{\mathbf{A}_{\max}}{\mathbf{A}_{\min}} \right)^{1/2} \right\|_{L^q(\Omega)} \|C_{\mathrm{R1}}\|_{L^{p_u}(\Omega)}\, h^{s_u} \\
&\lesssim \|\mathbf{A}_{\max}\|^{1/2}_{L^{q_1}(\Omega)} \|\mathbf{A}_{\min}\|^{1/2}_{L^{q_2}(\Omega)} \|C_{\mathrm{R1}}\|_{L^{p_u}(\Omega)}\, h^{s_u},
\end{aligned}$$

where $q = \frac{p_u p}{p_u - p}$ and $q_1^{-1} + q_2^{-1} = q^{-1}$. Since we can choose any $q_1, q_2 \in [1, \infty)$, it follows that we can choose any $p < p_u$. □

The usual duality (or Aubin–Nitsche) trick leads to a bound on the L^2–error. To this end, denote by $z_1(\omega, \cdot) \in H^1_0(D)$ the solution to dual problem (2.22) with $M_\omega(v) = (u - u_h, v)_{L^2(D)}$, i.e.

$$b_\omega\big(v, z_1(\omega, \cdot)\big) = \big((u - u_h)(\omega, \cdot), v\big)_{L^2(D)}, \quad \text{for all} \quad v \in H^1_0(D). \tag{3.3}$$

We make the following assumption on the regularity of z_1.

R2. For s_u as in assumption R1, we have that for almost all $\omega \in \Omega$,

$$\|z_1(\omega,\cdot)\|_{H^{1+s_u}(D)} \leq C_{\mathrm{R2}}(\omega)\, \|(u-u_h)(\omega,\cdot)\|_{L^2(D)},$$

for some $C_{\mathrm{R2}} \in L^p(\Omega)$, for all $1 \leq p < \infty$.

Note that this assumption is more specific than assumption R1, due to the explicit dependence of the bound on $\|(u-u_h)(\omega,\cdot)\|_{L^2(D)}$. The constant C_{R2} is assumed to be in $L^p(\Omega)$, for any $p < \infty$, since it will typically only depend on the coefficient $\mathbf{A}$. It follows again from Theorem 2.12 that assumption R2 is satisfied for the model problems considered in chapter 2.

We then have the following estimate on $\|u-u_h\|_{L^2(D)}$.

Corollary 3.4. *Let assumptions A1–A2 hold with $t = 0$, let assumption R1 hold for some $0 < s_u \leq 1$ and $1 \leq p_u \leq \infty$ and let assumption R2 hold. Then, for all $p < p_u$ and $h > 0$, we have*

$$\|u-u_h\|_{L^p(\Omega,L^2(D))} \lesssim C_{3.4}\, h^{2s_u}, \quad \text{with } C_{3.4} = \left\| \frac{C_{\mathrm{R2}}\, \mathbf{A}_{\max}^{3/2}}{\mathbf{A}_{\min}^{1/2}} \right\|_{L^q(\Omega)} \|C_{\mathrm{R1}}\|_{L^{p_u}(\Omega)},$$

where $q = \frac{p_u - p}{p_u\, p}$.

Proof. Dropping for brevity the dependence on ω, and using the dual problem (3.3), Galerkin orthogonality and the boundedness of $b_\omega(\cdot,\cdot)$, we have

$$\begin{aligned}\|u-u_h\|_{L^2(D)}^2 &= (u-u_h, u-u_h)_{L^2(D)} = \inf_{z_h \in V_{h,0}} b_\omega(u-u_h, z_1 - z_h) \\ &\lesssim \mathbf{A}_{\max}\, |u-u_h|_{H^1(D)} \inf_{z_h \in V_{h,0}} |z_1 - z_h|_{H^1(D)}.\end{aligned}$$

As in (3.2), we have $|u-u_h|_{H^1(D)} \lesssim (\mathbf{A}_{\max}/\mathbf{A}_{\min})^{1/2}\, C_{\mathrm{R1}} h^{s_u}$. It follows from Lemma 3.2, together with assumption R2, that $\inf_{z_h \in V_{h,0}} |z_1 - z_h|_{H^1(D)} \lesssim C_{\mathrm{R2}}\, \|u - u_h\|_{L^2(D)} h^{s_u}$. Combining the estimates gives

$$\|u-u_h\|_{L^2(D)} \lesssim \mathbf{A}_{\max} \left(\frac{\mathbf{A}_{\max}}{\mathbf{A}_{\min}}\right)^{1/2} C_{\mathrm{R1}}\, C_{\mathrm{R2}}\, h^{2s_u}.$$

The claim of the Corollary then follows by an application of Hölder's inequality, together with assumption A1-A2, R1 and R2. □

3.2 Error estimates for functionals

In practical applications, one is often interested in the expected value of functionals of the solution. A standard technique to prove convergence for finite element approximations of output functionals is to use a duality argument, similar to the classic Aubin-Nitsche trick used to prove optimal convergence rates for the L^2-norm.

Denote by $M_\omega(\cdot) : H^1(D) \to \mathbb{R}$ the functional of interest, and assume M_ω satisfies assumption F1 with $w_1 = u, w_2 = u_h$ and $t_* = 0$. Denote by $z_2(\omega, \cdot) \in H_0^1(D)$ the solution to dual problem (2.22) with $w_1 = u$ and $w_2 = u_h$, i.e.

$$b_\omega\big(v, z_2(\omega,\cdot)\big) = \overline{D_v M_\omega}(u(\omega,\cdot), u_h(\omega,\cdot)), \quad \text{for all} \quad v \in H_0^1(D).$$

Denote correspondingly by $z_{2,h}(\omega,\cdot)$ the finite element approximation in $V_{h,0}$. We make the following assumption on the regularity of z_2.

R3. There exist $0 \le s_z \le 1$ and $1 \le p_z \le \infty$ such that for almost all $\omega \in \Omega$,

$$\|z_2(\omega,\cdot)\|_{H^{1+s_z}(D)} \le C_{\mathrm{R3}}(\omega)\,,$$

for some $C_{\mathrm{R3}} \in L^{p_z}(\Omega)$.

Assumptions R3 again holds under reasonable assumptions on the data $\mathbf{A}$, f and $\{\phi_j\}_{j=1}^m$, and for a large class of functionals M_ω, as discussed in section 2.5.

Recall that we assumed that the boundary data $\{\phi_j\}_{j=1}^m$ are piecewise linear with respect to $\mathcal{T}_h$, and so $u - u_h \in H_0^1(D)$. From the Fundamental Theorem of Calculus for Fréchet derivatives, it follows that

$$\begin{aligned} M_\omega(u) - M_\omega(u_h) &= \int_0^1 D_{u-u_h} M_\omega(u + \theta(u_h - u))\,\mathrm{d}\theta = \overline{D_{u-u_h} M_\omega}(u, u_h) \\ &= b_\omega(u - u_h, z_2). \end{aligned} \tag{3.4}$$

We then have the following error bound.

Lemma 3.5. *For almost all $\omega \in \Omega$,*

$$\begin{aligned}&|M_\omega\left(u(\omega,\cdot)\right) - M_\omega\left(u_h(\omega,\cdot)\right)| \\ &\qquad \leq \mathbf{A}_{\max}(\omega)\,|u(\omega,\cdot) - u_h(\omega,\cdot)|_{H^1(D)} \inf_{z_h \in V_{h,0}} |z_2(\omega,\cdot) - z_h(\omega,\cdot)|_{H^1(D)}\,. \quad (3.5)\end{aligned}$$

Proof. Dropping for brevity the dependence on ω and using (3.4), as well as Galerkin orthogonality for the primal problem (2.4) and the boundedness of b_ω, we have

$$\begin{aligned}|M_\omega(u) - M_\omega(u_h)| = |b_\omega(u-u_h, z_2)| &= \inf_{z_h \in V_{h,0}} |b_\omega(u-u_h, z_2 - z_h)| \\ &\leq \mathbf{A}_{\max}(\omega)\,|u-u_h|_{H^1(D)} \inf_{z_h \in V_{h,0}} |z_2 - z_h|_{H^1(D)}\,.\end{aligned}$$

□

This simple argument will be crucial to obtain optimal convergence rates for functionals.

Remark 3.6. Continuous Fréchet differentiability is in fact not a necessary condition to obtain the bound in Lemma 3.5. It is sufficient to assume only Lipschitz continuity of M_ω (see e.g.[13]).

We are now ready to prove optimal convergence rates for moments of the finite element error for Fréchet differentiable (and thus also for linear) functionals as defined above.

Lemma 3.7. *Let assumptions A1–A2 hold with $t = 0$, let assumption R1 hold with some $0 < s_u \leq 1$ and $1 \leq p_u \leq \infty$, and let assumption R3 hold with some $0 \leq s_z \leq 1$ and $1 \leq p_z \leq \infty$. Then, for any $p < \frac{p_u p_z}{p_u + p_z}$, we have*

$$\|M_\omega(u) - M_\omega(u_h)\|_{L^p(\Omega)} \lesssim C_{3.7} h^{s_u + s_z},$$

with

$$C_{3.7} := \left\| \frac{\mathbf{A}_{\max}^{3/2}}{\mathbf{A}_{\min}^{1/2}} \right\|_{L^q(\Omega)} \|C_{\mathrm{R1}}\|_{L^{p_u}(\Omega)}\, \|C_{\mathrm{R3}}\|_{L^{p_z}(\Omega)},$$

where $q = \frac{p_u\, p_z\, p}{p_u p_z - p_u p - p_z p}$.

Proof. Dropping for brevity the dependence on ω, we have from Lemma 3.5 that for almost all $\omega \in \Omega$,

$$|M_\omega(u) - M_\omega(u_h)| \leq \mathbf{A}_{\max} |u - u_h|_{H^1(D)} \inf_{z_h \in V_{h,0}} |z_2 - z_h|_{H^1(D)}.$$

As in (3.2), we have $|u - u_h|_{H^1(D)} \lesssim (\mathbf{A}_{\max}/\mathbf{A}_{\min})^{1/2} C_{\mathrm{R1}} h^{s_u}$. It follows from Lemma 3.2, together with assumption R3, that $\inf_{z_h \in V_{h,0}} |z_2 - z_h|_{H^1(D)} \lesssim C_{\mathrm{R3}}\, h^{s_z}$. Combining the estimates together gives

$$|M_\omega(u) - M_\omega(u_h)| \leq \mathbf{A}_{\max} \left(\frac{\mathbf{A}_{\max}}{\mathbf{A}_{\min}}\right)^{1/2} C_{\mathrm{R1}}\, C_{\mathrm{R3}}\, h^{s_u}\, h^{s_z}.$$

The claim of the Lemma then follows from assumptions A1-A2, R1 and R3, together with Hölder's inequality. □

3.3 L^∞ and $W^{1,\infty}$ error estimates

The aim of this section is to derive bounds on moments of $\|u - u_h\|_{L^\infty(D)}$ and $|u - u_h|_{W^{1,\infty}(D)}$. A classical method used to derive these estimates, is the method of weighted Sobolev spaces by Nitsche. The results presented in this section are specific to continuous, linear finite elements on quasi-uniform triangulations $\mathcal{T}_h$ in $\mathbb{R}^2$, but extensions to higher spatial dimensions and/or higher order elements can be proved in a similar way (see e.g. [14]).

As in the previous sections, the convergence rate of the finite element error depends on the spatial regularity of u. However, instead of requiring a certain Sobolev regularity of u, we will now require a certain Hölder regularity. We make the following assumption.

R4. There exist $0 < s_u \leq 1$ and $1 \leq p_u \leq \infty$ such that for almost all $\omega \in \Omega$,

$$\|u(\omega, \cdot)\|_{C^{1+s_u}(\overline{D})} \leq C_{\mathrm{R4}}(\omega),$$

for some $C_{\mathrm{R4}} \in L^{p_u}(\Omega)$.

Assumption R4 holds for model problem (2.1) under reasonable assumptions on the data $\mathbf{A}$, f and $\{\phi_j\}_{j=1}^m$, as discussed in section 2.6.

As for the H^1 error in section 3.1, the key ingredients in proving the error bounds in the L^∞-norm and $W^{1,\infty}$-norm are a quasi-optimality result (similar to Lemma 3.1) and a best approximation result (similar to Lemma 3.2).

To prove quasi-optimality, we will use the method of weighted norms by Nitsche. For the case $\mathbf{A} = I_d$, a full proof can be found in [14, §3.3]. For simplicity, we shall restrict ourselves to the case $u \in H_0^1(D)$, i.e. $\phi = 0$ in model problem (2.1).

Denote by $P_h : H_0^1(D) \to V_{h,0}$ the projection operator associated with the inner product b_ω, defined for all $v \in H_0^1(D)$ by the relations

$$P_h v \in V_{h,0} \quad \text{and} \quad \forall w_h \in V_{h,0}, \quad b_\omega(v - P_h v, w_h) = 0.$$

We hence in particular have $P_h u = u_h$. The main idea of the proof of the quasi-optimality is then to show that the projection operators P_h can be bounded independently of h, provided the norms on $H_0^1(D)$ and $V_{h,0}$ are chosen in a particular, h–dependent way. This gives the following bound.

Lemma 3.8. *Let $0 < \mathbf{A}_{\min}(\omega) \leq \mathbf{A}_{\max}(\omega) < \infty$. For all $v \in H_0^1(D) \cap W^{1,\infty}(D)$ and h sufficiently small, we have*

$$\begin{aligned} |\ln h|^{-1/2} \|P_h v\|_{L^\infty(D)} \,+\, h\, |P_h v|_{W^{1,\infty}(D)} \\ \lesssim \frac{\mathbf{A}_{\max}(\omega)}{\mathbf{A}_{\min}(\omega)} \left(\|v\|_{L^\infty(D)} \,+\, h\, |\ln h|\, |v|_{W^{1,\infty}(D)} \right). \end{aligned} \tag{3.6}$$

Proof. The proof is identical to that of [14, Theorem 3.3.6], with the dependence of the bound on $\mathbf{A}$ made explicit. In particular, the dependence on $\mathbf{A}$ enters in the proof of [14, Theorem 3.3.5], which uses the coercivity and boundedness of the bilinear form b_ω. □

This now allows us to prove the following quasi–optimality result.

Lemma 3.9. *For all h sufficiently small,*

$$\begin{aligned} |\ln h|^{-1/2} \|(u - u_h)(\omega, \cdot)\|_{L^\infty(D)} \,+\, h\, |(u - u_h)(\omega, \cdot)|_{W^{1,\infty}(D)} \\ \lesssim \frac{\mathbf{A}_{\max}(\omega)}{\mathbf{A}_{\min}(\omega)} \inf_{v_h \in V_{h,0}} \left(\|u(\omega, \cdot) - v_h\|_{L^\infty(D)} \,+\, h |\ln h|\, |u(\omega, \cdot) - v_h|_{W^{1,\infty}(D)} \right). \end{aligned}$$

Proof. We first note that for any $v_h \in V_{h,0}$, we have the identity

$$u - u_h = u - P_h u = (I - P_h)(u - v_h), \tag{3.7}$$

where I denotes the identity operator.

Next, we note that the norm of the identity mapping from $H_0^1(D) \cap W^{1,\infty}(D)$ equipped with the norm $v \to \|v\|_{L^\infty(D)} + h\,|\ln h|\,|v|_{W^{1,\infty}(D)}$, into the same space equipped with the norm $v \to |\ln h|^{-1/2}\,\|v\|_{L^\infty(D)} + h\,|v|_{W^{1,\infty}(D)}$, can be bounded by $|\ln h_0|^{-1/2}$ for all $h \leq \min\{h_0, \mathrm{e}^{-1}\}$, for some $h_0 > 0$. The claim of the Lemma then follows from (3.7), together with Lemma 3.8. □

We have the following result on the best approximation error (see e.g [14, 63]).

Lemma 3.10. *Let $v \in C^r(\overline{D})$, for some $1 < r \leq 2$. Then*

$$\inf_{v_h \in V_{h,0}} \left(\|v - v_h\|_{L^\infty(D)} + h|v - v_h|_{W^{1,\infty}(D)}\right) \lesssim h^r\,\|v\|_{C^r(\overline{D})},$$

where the hidden constant is independent of v and h.

Proof. Firstly note that for $v \in C^r(\overline{D})$, the norms $\|v\|_{C^r(\overline{D})}$ and $\|v\|_{W^{r,\infty}(D)}$ are equivalent (see e.g. [56]). It follows from [14, Theorem 3.1.6] that

$$\inf_{v_h \in V_{h,0}} \left(\|v - v_h\|_{L^\infty(D)} + h|v - v_h|_{W^{1,\infty}(D)}\right) \lesssim h^2\,\|v\|_{W^{2,\infty}(D)},$$

and

$$\inf_{v_h \in V_{h,0}} \left(\|v - v_h\|_{L^\infty(D)} + h|v - v_h|_{W^{1,\infty}(D)}\right) \lesssim h\,\|v\|_{W^{1,\infty}(D)}.$$

By interpolation (see e.g [7, §14]), it then follows that

$$\inf_{v_h \in V_{h,0}} \left(\|v - v_h\|_{L^\infty(D)} + h|v - v_h|_{W^{1,\infty}(D)}\right) \lesssim h^r\,\|v\|_{W^{r,\infty}(D)},$$

for any $1 < r \leq 2$. The claim of the lemma then follows. □

Combining Lemmas 3.9 and 3.10, we have the following convergence result.

Theorem 3.11. *Suppose $\mathcal{T}_h$ is a quasi-uniform triangulation of $D \subset \mathbb{R}^2$, and suppose $\phi = 0$. Let assumptions A1-A2 hold with $t = 0$, and let assumption R4*

hold with some $0 < s_u \leq 1$ *and* $1 \leq p_u \leq \infty$. *Then, for all* $p < p_u$, $s < s_u$ *and* h *sufficiently small, we have*

$$\|u - u_h\|_{L^p(\Omega, L^\infty(D))} + h|u - u_h|_{L^p(\Omega, W^{1,\infty}(D))} \lesssim C_{3.11}\, h^{1+s},$$

with

$$C_{3.11} := \left\| \frac{\mathbf{A}_{\max}}{\mathbf{A}_{\min}} \right\|_{L^q(\Omega)} \|C_{\mathrm{R4}}\|_{L^{p_u}(\Omega)},$$

where $q = \frac{p - p_u}{p_u\, p}$.

Proof. It follows from Lemmas 3.9 and 3.10, together with $|\ln h| < h^\delta$ for any $\delta > 0$, that, for almost all $\omega \in \Omega$,

$$\|(u - u_h)(\omega, \cdot)\|_{L^\infty(D)} \lesssim \frac{\mathbf{A}_{\max}(\omega)}{\mathbf{A}_{\min}(\omega)}\, h^{1+s}\, \|u(\omega, \cdot)\|_{C^{1+s_u}(\overline{D})},$$

and

$$|(u - u_h)(\omega, \cdot)|_{W^{1,\infty}(D)} \lesssim \frac{\mathbf{A}_{\max}(\omega)}{\mathbf{A}_{\min}(\omega)}\, h^{s}\, \|u(\omega, \cdot)\|_{C^{1+s_u}(\overline{D})},$$

for any h sufficiently small and any $s < s_u$. The claim then follows from assumptions A1-A2 and R4, together with Minkowski's and Hölder's inequality. □

3.4 Variational crimes

In practice, the exact computation of the finite element solution u_h, as defined in (3.1), is in general not possible, and further approximations are often required for the numerical computation. In this section, we discuss two important such approximations: the use of approximate bilinear forms, resulting for example from quadrature or other approximations of the coefficient, and the approximation of non–polygonal boundaries.

3.4.1 Quadrature error

The integrals appearing in the bilinear form b_ω and in the linear functional L_ω involve realisations of random fields. It will in general be impossible to evaluate these integrals exactly, and so one generally uses quadrature instead. We will only explicitly analyse the quadrature error in b_ω, but the quadrature error in

approximating L_ω can be analysed analogously. We will for simplicity restrict our attention to scalar coefficients $\mathbf{A}(\omega, x) = a(\omega, x)\,\mathbf{I}_d$, for some $a : \Omega \times \overline{D} \to \mathbb{R}$, and to homogeneous boundary conditions $\phi = 0$.

We will analyse the midpoint rule, approximating the integrand by its value at the midpoint x_τ of each simplex $\tau \in \mathcal{T}_h$. The trapezoidal rule, and indeed any other rule which uses a linear combination of point evaluations of a in τ, can be analysed analogously. Let us denote the resulting bilinear form that approximates b_ω on the grid $\mathcal{T}_h$ by

$$\tilde{b}_\omega(w_h, v_h) = \sum_{\tau \in \mathcal{T}_h} a(\omega, x_\tau) \int_\tau \nabla w_h(x) \cdot \nabla v_h(x)\,\mathrm{d}x\,,$$

and let $\tilde{u}_h(\omega, \cdot) \in V_{h,0}$ denote the corresponding solution to

$$\tilde{b}_\omega(\tilde{u}_h(\omega, \cdot), v_h) = L_\omega(v_h)\,, \quad \text{for all} \quad v_h \in V_{h,0}.$$

Clearly the bilinear form $\tilde{b}_\omega$ is bounded and coercive, with the same constants as the exact bilinear form b_ω and so we can apply the following classical result [14] (with explicit dependence of the bound on the coefficients).

Lemma 3.12 (First Strang Lemma)**.** *Let $0 < a_{\min}(\omega) \le a_{\max}(\omega) < \infty$. Then*

$$|(u - \tilde{u}_h)(\omega, \cdot)|_{H^1(D)} \le \inf_{v_h \in V_{h,0}} \left\{ \left(1 + \frac{a_{\max}(\omega)}{a_{\min}(\omega)}\right) |u(\omega, \cdot) - v_h|_{H^1(D)} + \frac{1}{a_{\min}(\omega)} \sup_{w_h \in V_{h,0}} \frac{|b_\omega(v_h, w_h) - \tilde{b}_\omega(v_h, w_h)|}{|w_h|_{H^1(D)}} \right\}.$$

This gives the following convergence for the approximate finite element solution $\tilde{u}_h$.

Proposition 3.13. *Suppose $\mathbf{A}(\omega, x) = a(\omega, x) I_d$, and suppose $\phi = 0$. Let assumption R1 hold with some $0 < s_u \le 1$ and $1 \le p_u \le \infty$, and let assumptions A1–A3 hold with $t \ge s_u$ and $p_* \ge p_u$. Then, for all $p < p_u$, we have*

$$\|u - \tilde{u}_h\|_{L^p(\Omega, H^1_0(D))} \lesssim C_{3.13}\, h^{s_u},$$

with

$$C_{3.13} := \left\| \left(\frac{a_{\max}}{a_{\min}} \right)^{3/2} \right\|_{L^q(\Omega)} \|C_{\mathrm{R1}}\|_{L^{p_u}(\Omega)} + \left\| \frac{|a|_{C^{s_u}(\overline{D})}}{a_{\min}^2} \right\|_{L^q(\Omega)} \|f\|_{L^{p_u}(\Omega, H^{-1}(D))},$$

where $q = \frac{p-p_u}{p_u\, p}$.

Proof. We first note that, for all $v_h, w_h \in V_{h,0}$,

$$\begin{aligned} \left| b_\omega(v_h, w_h) - \tilde{b}_\omega(v_h, w_h) \right| &= \left| \sum_{\tau \in \mathcal{T}_h} \int_\tau \left(a(\omega, x) - a(\omega, x_\tau) \right) \nabla v_h \cdot \nabla w_h \, \mathrm{d}x \right| \\ &\leq \sum_{\tau \in \mathcal{T}_h} \int_\tau \frac{|a(\omega, x) - a(\omega, x_\tau)|}{|x - x_\tau|^{s_u}} \, |x - x_\tau|^{s_u} \, |\nabla v_h \cdot \nabla w_h| \, \mathrm{d}x \\ &\leq |a(\omega)|_{C^{s_u}(\overline{D})} \, h^{s_u} \, |v_h|_{H^1(D)} \, |w_h|_{H^1(D)} . \end{aligned}$$

Dropping for brevity the dependence on ω, it follows from Lemma 3.12 and $a_{\min} \leq a_{\max}$ that, for almost all $\omega \in \Omega$,

$$|u - \tilde{u}_h|_{H^1(D)} \leq \inf_{v_h \in V_{h,0}} \left\{ \left(\frac{a_{\max}}{a_{\min}} \right) |u - v_h|_{H^1(D)} + h^{s_u} \, \frac{|a|_{C^{s_u}(\overline{D})}}{a_{\min}} \, |v_h|_{H^1(D)} \right\}.$$

Let us now make the particular choice $v_h := u_h \in V_{h,0}$, i.e. the solution of (3.1). Then it follows from (3.2) that

$$|u - \tilde{u}_h|_{H^1(D)} \lesssim \left(\left(\frac{a_{\max}}{a_{\min}} \right)^{3/2} C_{\mathrm{R1}} \; + \; \frac{|a|_{C^{s_u}(\overline{D})}}{a_{\min}} \, C_{2.1} \right) h^{s_u}.$$

The result then follows from assumptions A1-A3 and R1, together with Hölder's inequality. □

Proposition 3.13 shows that the convergence rates of the finite element error in the H^1-norm are not influenced by the use of quadrature. However, the use of the mesh-dependent bilinear forms $\tilde{b}_\omega$ crucially leads to the loss of Galerkin orthogonality, and the duality arguments to prove higher convergence rates for the L^2-norm and the error in functionals (cf Corollary 3.4 and Lemma 3.7) are hence no longer applicable. However, since $H^1(D) \subset L^2(D)$, it follows immediately from Proposition 3.13 that

$$\|u - \tilde{u}_h\|_{L^p(\Omega, L^2(D))} \lesssim C_{3.13}\, h^{s_u},$$

for all $p < p_u$. Similarly, using (3.4) and assumption F1 with $t_* = 0$ and $1 \leq q_* \leq \infty$, we have

$$\|M_\omega(u) - M_\omega(\tilde{u}_h)\|_{L^p(\Omega)} \lesssim \|C_{F1}\|_{L^{q_*}(\Omega)}\, C_{3.13}\, h^{s_u},$$

for any $p < \frac{p_u\, q_*}{p_u + q_*}$.

To recover the faster convergence rates for the L^2-error and the error in functionals M_ω in the case of quadrature, we require additional regularity of the coefficient a.

3.4.2 Truncation error

We will in this section restrict our attention to the case $\mathbf{A}(\omega, x) = a(\omega, x)\, \mathbf{I}_d$, where $a : \Omega \times \overline{D} \to \mathbb{R}$ is a log–normal random field as described in section 2.7.

A starting point for many numerical schemes for PDEs with random coefficients is the approximation of the random field $a(\omega, x)$ as a function of a finite number of random variables, $a(\omega, x) \approx a(\xi_1(\omega), ..., \xi_R(\omega), x)$. This is true, for example, for the stochastic collocation and stochastic Galerkin methods. Sampling methods, such as Monte–Carlo type methods discussed in chapter 4, do not rely on such a finite–dimensional approximation as such, but may make use of such approximations as a way of producing samples of the input random field.

A popular choice to achieve good approximations of this kind for log–normal fields is the truncated Karhunen-Loève (KL) expansion. Let $g : \Omega \times \overline{D} \to \mathbb{R}$ be a Gaussian random field such that $a = \exp[g]$. Denote by $\mu(x)$ the mean (or *expected value*) of g, and by $C(x, y)$ its two-point covariance function (as in (2.24)). The KL–expansion of g is then given by

$$g(x, \omega) = \mu(x) + \sum_{n=1}^{\infty} \sqrt{\mu_n}\, \xi_n(\omega)\, b_n(x), \tag{3.8}$$

where $\{\mu_n\}_{n \in \mathbb{N}}$ are the eigenvalues and $\{b_n\}_{n \in \mathbb{N}}$ the $L^2(D)$ orthonormalised eigenfunctions of the covariance operator with kernel function $C(x, y)$ defined in (2.24). $\{\xi_n\}_{n \in \mathbb{N}}$ is a set of independent, standard Gaussian random variables. For a more detailed derivation of the KL–expansion, see e.g. [30]. We will here only sum-

marise some of its main properties.

- With the kernel function $C(x, y)$ in (2.24), the covariance operator is self-adjoint, non-negative and compact on $L^2(D)$, which implies that the eigenvalues $\{\mu_n\}_{n\in\mathbb{N}}$ are real, non-negative and tend to 0 as $n \to \infty$ and the eigenfunctions $\{b_n\}_{n\in\mathbb{N}}$ form an orthonormal basis of $L^2(D)$.

- For the homogeneous random field g, we have

$$\sigma^2 := \mathbb{V}[g] = \mathbb{E}\left[\left(\sum_{n=1}^{\infty} \sqrt{\mu_n}\, \xi_n(\omega)\, b_n(x)\right)^2\right] = \sum_{n=1}^{\infty} \mu_n\, b_n^2(x)$$

  since the random variables ξ_n are independent with $\xi_n \tilde{N}(0, 1)$. Since the eigenfunctions b_n are orthonormalised in $L^2(D)$, it follows that

$$\sum_{n=1}^{\infty} \mu_n = \|\sigma^2\|_{L^2(D)} = \sigma^2\, \mathrm{meas}(D),$$

  where $\mathrm{meas}(D) := \int_D \mathrm{d}x$.

We shall write the random field $a(\omega, x)$ as

$$a(\omega, x) = \exp[\mu(x)]\, \exp\left[\sum_{n=1}^{\infty} \sqrt{\mu_n}\, \xi_n(\omega)\, b_n(x)\right]$$

In practice we have to truncate the expansion (3.8) after a finite number R of terms. Let g^R denote the KL-expansion of g truncated after R terms, and let $a^R := \exp[g^R]$. Moreover, we denote by $u_{R,h} \in V_{h,\phi}$ the solution to the variational problem

$$b_\omega^R\big(u_{R,h}(\omega, \cdot), v\big) = L_\omega(v)\,, \quad \text{for all} \quad v \in V_{h,0}, \tag{3.9}$$

where the linear functional $L_\omega(\cdot)$ is as in (2.6), and the bilinear form $b_\omega^R(\cdot, \cdot)$ is defined as

$$b_\omega^R(u, v) := \int_D a^R(\omega, x) \nabla u(x) \cdot \nabla v(x)\, \mathrm{d}x, \tag{3.10}$$

i.e. the bilinear form (2.5) with a replaced by its R-term approximation a^R.

Since the bilinear form $b_\omega^R(\cdot, \cdot)$ is bounded and coercive on $H_0^1(D)$ with constants $a_{\max}^R(\omega) := \max_{x\in\overline{D}} a^R(\omega, x)$ and $a_{\min}^R(\omega) := \min_{x\in\overline{D}} a^R(\omega, x)$ respectively

(cf Lemma 2.1), we have

$$\|u_{R,h}\|_{H^1(D)} \lesssim \frac{\|f(\omega,\cdot)\|_{H^{-1}(D)} + a^R_{\max}(\omega)\|\phi\|_{H^1(D)}}{a^R_{\min}(\omega)} =: C^R_{2.1}(\omega).$$

The aim of this section is now to analyse the error committed by approximating a by a^R. More precisely, we will derive a bound on $\|u_h - u_{R,h}\|_{L^p(\Omega,H^1_0(D))}$.

We make the following assumptions on $\{\mu_n, b_n\}_{n\in\mathbb{N}}$:

B1. The eigenfunctions are continuously differentiable, i.e. $b_n \in C^1(\overline{D})$ for any $n \in \mathbb{N}$.

B2. We have

$$\sum_{n=1}^{\infty} \mu_n \|b_n\|^2_{L^\infty(D)} < +\infty.$$

B3. There exists an $r \in (0,1)$ such that,

$$\sum_{n=1}^{\infty} \mu_n \|b_n\|^{2(1-r)}_{L^\infty(D)} \|\nabla b_n\|^{2r}_{L^\infty(D)} < +\infty.$$

For r as in assumption B3, let us define

$$E^r_R := \max\left(\sum_{n\geq R} \mu_n \|b_n\|^2_{L^\infty(D)}, \sum_{n\geq R} \mu_n \|b_n\|^{2(1-r)}_{L^\infty(D)} \|\nabla b_n\|^{2r}_{L^\infty(D)}\right). \tag{3.11}$$

Assumptions B1–B3 are fulfilled, among other cases, for the analytic covariance function (2.26) as well as for the exponential covariance function (2.25) with 1-norm $\|x\| = \sum_{i=1}^d |x_i|$, since then the eigenvalues and eigenfunctions can be computed explicitly and we have explicit decay rates for the KL–eigenvalues. For details see [10, section 7]. In the latter case, on non-rectangular domains D, we need to use a KL-expansion on a bounding box containing D to get again explicit formulae for the eigenvalues and eigenfunctions. Strictly speaking this is not a KL–expansion on D.

The following results were proven in [10].

Proposition 3.14. *Let assumptions B1–B3 hold for some $0 < r < 1$. Then*

$\|a(\omega,\cdot) - a^R(\omega,\cdot)\|_{C^0(\overline{D})} \to 0$ *for almost all* $\omega \in \Omega$, *and*

$$\|a - a^R\|_{L^p(\Omega, C^0(\overline{D}))} \lesssim \sqrt{E_R^r},$$

for any $p \in [1,\infty)$. *Furthermore,* $\|a^R_{\min}\|_{L^q(\Omega)}$ *and* $\|1/a^R_{\min}\|_{L^q(\Omega)}$ *can be bounded by a constant independent of* R, *for any* $q \in [1,\infty)$.

Proposition 3.15. *We have the following bound on* $\sqrt{E_R^r}$*:*

$$\sqrt{E_R^r} \lesssim \begin{cases} R^{-\rho}, & \text{for 1-norm exponential covariance,} \\ R^{\frac{d-1}{2d}} \exp\big(-c_1 R^{1/d}\big), & \text{for Gaussian covariance,} \end{cases}$$

for some constant $c_1 > 0$ *and for any* $0 < \rho < 1/2$. *The hidden constants are independent of* R.

In [10, 12], the results in Proposition 3.14 were used to bound the truncation error in the solution u to (2.4), i.e. to bound the error $\|u - u_R\|_{L^p(\Omega,H^1(D))}$, where u_R is the solution to the variational problem (2.4) with the bilinear form $b_\omega(\cdot,\cdot)$ replaced by its truncated version $b^R_\omega(\cdot,\cdot)$. Following the same lines, one can prove a bound on $\|u_h - u_{R,h}\|_{L^p(\Omega,H^1_0(D))}$.

Proposition 3.16. *Let assumptions B1–B3 hold for some* $0 < r < 1$. *Then for all* $p \in [1,\infty)$, *we have*

$$\|u_h - u_{R,h}\|_{L^p(\Omega,H^1_0(D))} \lesssim C_{3.16}\sqrt{E_R^r}, \quad \text{with } C_{3.16} := \left\| \frac{C_{2.1}}{a^R_{\min}} \right\|_{L^q(\Omega)},$$

for some $q < p$. *Similarly,* $\|u_h - u_{R,h}\|_{L^p(\Omega,L^2(D))} \lesssim C_{3.16}\sqrt{E_R^r}$. *The hidden constants are independent of* R.

Proof. The proof follows that of [10, Proposition 4.1 and Theorem 4.2]. Dropping for brevity the dependence on ω, and using the variational problems (2.4) and (3.9), we have, for almost all $\omega \in \Omega$,

$$\begin{aligned} a^R_{\min}\,|u_h - u_{R,h}|^2_{H^1(D)} &\le \int_D a^R\,|\nabla(u_h - u_{R,h})|^2\,\mathrm{d}x \\ &= \int_D (a^R - a)\,\nabla u_h\,\nabla(u_h - u_{R,h})\,\mathrm{d}x \\ &\le \|a - a^R\|_{C^0(\overline{D})}\,|u_h|_{H^1(D)}\,|u_h - u_{R,h}|_{H^1(D)} \end{aligned}$$

It follows that

$$|u_h - u_{R,h}|_{H^1(D)} \lesssim \frac{1}{a^R_{\min}} |u_h|_{H^1(D)} \|a - a^R\|_{C^0(\overline{D})}.$$

The claim of the Proposition then follows by Hölder's inequality, Lemma 2.1 and Proposition 3.14. □

As in the previous sections this result can again be extended in a straightforward way to functionals.

Corollary 3.17. *Let assumptions B1-B3 hold for some $0 < r < 1$, and let $M_\omega(\cdot)$ satisfy assumption F1 with $t_* = 0$ for $w_1 = u_h$ and $w_2 = u_{R,h}$, i.e. there exists $C'_{\mathrm{F1}} \in L^{q_*}(\Omega)$ s.t. $\overline{D_v M_\omega}(u_h, u_{R,h}) \lesssim C'_{\mathrm{F1}}(\omega)|v|_{H^1(D)}$ for any $v \in H^1_0(D)$. Then for all $p < q_*$,*

$$\|M_\omega(u_h) - M_\omega(u_{R,h})\|_{L^p(\Omega)} \lesssim C_{3.17}\sqrt{E^r_R}, \text{ with } C_{3.17} := \left\| \frac{C_{2.1}}{a^R_{\min}} \right\|_{L^q(\Omega)} \|C'_{F1}\|_{L^{q_*}(\Omega)},$$

where $q < \frac{q_* - p}{q_* p}$. The hidden constant is again independent of R.

Proof. Dropping for brevity the dependence on ω, it follows by assumption that for almost all $\omega \in \Omega$,

$$M_\omega(u_h) - M_\omega(u_{R,h}) = \overline{D_{u_h - u_{R,h}} M_\omega}(u_h, u_{R,h}) \lesssim C'_{F1} |u_h - u_{R,h}|_{H^1(D)}.$$

The claim of the Corollary then follows from Proposition 3.16 and Hölder's inequality. □

Note that in Corollary 3.17, we cannot exploit Galerkin orthogonality to get a doubling of the convergence rate with respect to R, since u_h and $u_{R,h}$ are solutions to problems with different bilinear forms.

Combining the truncation error analysis with the discretisation error analysis from sections 3.1–3.2, we get the following total error bounds.

Theorem 3.18. *Let a be a log–normal random field s.t. $\log[a]$ has 1-norm exponential covariance, and suppose that assumption R1 is satisfied for some $1 \le p_u \le \infty$ and all $s_u < 1/2$. Then for all $p < p_u$ and $0 < s, \rho < 1/2$, we have*

$$\|u - u_{R,h}\|_{L^p(\Omega, H^1_0(D))} \lesssim \left(C_{3.3}\, h^s + C_{3.16}\, R^{-\rho} \right).$$

Similarly, $\|u - u_{R,h}\|_{L^p(\Omega,L^2(D))} \lesssim (C_{3.4}h^{2s} + C_{3.16}\,R^{-\rho})$. *The hidden constants are independent of* h *and* R.

Proof. This follows directly from Theorem 3.3, Corollary 3.4 and Propositions 3.16 and 3.15, together with Hölder's inequality. □

Corollary 3.19. *Let* a *be a log–normal random field s.t.* $\log[a]$ *has 1-norm exponential covariance, and suppose* M_ω *satisfies the assumptions of Corollary 3.17 and Lemma 3.7, for all* $s_u < 1/2$, *some* $0 \leq s_z < 1/2$ *and some* $1 \leq p_u, p_z \leq \infty$. *Then*

$$\|M_\omega(u) - M_\omega(u_{R,h})\|_{L^p(\Omega)} \lesssim \left(C_{3.7}\,h^{s+s_z} + C_{3.17}\,R^{-\rho}\right),$$

for all $p < \frac{p_u\,p_z}{p_u+p_z}$ *and* $0 < s, \rho < 1/2$.

Proof. This follows directly from Lemma 3.7, Corollary 3.17 and Proposition 3.15, together with Hölder's inequality. □

Remark 3.20. *As already stated in Proposition 3.14, the quantities* $\|a^R_{\max}\|_{L^q(\Omega)}$ *and* $\|1/a^R_{\min}\|_{L^q(\Omega)}$ *can be bounded by a constant independent of* R, *for any* $q \in [1,\infty)$. *Furthermore, it follows from Lemma 2.20 and [10, Proposition 3.8], that* $\|a\|_{L^q(\Omega,C^t(\overline{D}))}$ *can be bounded independently of* R, *for any* $t < 1/2$ *and* $q \in [1,\infty)$. *It hence follows under the same assumptions as in Corollary 3.19, that*

$$\|M_\omega(u_R) - M_\omega(u_{R,h})\|_{L^p(\Omega)} \lesssim C_{3.7}\,h^{s+s_z},$$

for all $p < \frac{p_u\,p_z}{p_u+p_z}$ *and* $0 < s < 1/2$.

3.4.3 Boundary approximation

We now consider the application of finite element methods to problems posed on smooth domains D. Since finite element methods use triangulations of the computational domain, and are hence naturally linked to polygonal/polyhedral domains, this involves the approximation of D by polygonal domains D_h. For simplicity, we again consider the case of homogeneous boundary conditions $\phi = 0$.

We denote by $\{\widehat{\mathcal{T}}_h\}_{h>0}$ a shape-regular family of simplicial triangulations of the domain D, parametrised by its mesh width $h := \max_{\tau\in\widehat{\mathcal{T}}_h} \operatorname{diam}(\tau)$, such that, for any $h > 0$,

- $\overline{D} \subset \bigcup_{\tau \in \widehat{\mathcal{T}}_h} \tau$, i.e. the triangulation covers all of $\overline{D}$, and
- the vertices $x_1^\tau, \dots, x_{d+1}^\tau$ of any $\tau \in \widehat{\mathcal{T}}_h$ lie either all in $\overline{D}$ or all in $\mathbb{R}^d \backslash D$.

Let $\overline{D}_h$ denote the union of all simplices that are interior to $\overline{D}$ and D_h its interior, so that $D_h \subset D$.

Associated with each triangulation $\widehat{\mathcal{T}}_h$ we define the space

$$\widehat{V}_{h,0} := \left\{ v_h \in C^0(\overline{D}) : v_h|_\tau \text{ linear } \forall \tau \in \widehat{\mathcal{T}}_h \text{ with } \tau \subset \overline{D}_h \,, \quad \text{and} \quad v_h|_{\overline{D} \backslash D_h} = 0 \right\} \tag{3.12}$$

of continuous, piecewise linear functions on D_h that vanish on the boundary of D_h and in $D \backslash D_h$. Let us recall the following standard interpolation result (cf Lemma 3.2).

Lemma 3.21. *Let $v \in H^{1+s}(D_h)$, for some $0 < s \leq 1$. Then*

$$\inf_{v_h \in \widehat{V}_{h,0}} |v - v_h|_{H^1(D_h)} \lesssim \|v\|_{H^{1+s}(D_h)} \, h^s. \tag{3.13}$$

The hidden constant is independent of h and v.

This can easily be extended to an interpolation result for functions $v \in H^{1+s}(D) \cap H_0^1(D)$, by estimating the residual over $D \backslash D_h$. However, when D is not convex it requires local mesh refinement in the vicinity of any non-convex parts of the boundary. We make the following assumption on $\widehat{\mathcal{T}}_h$:

T1. For all $\tau \in \widehat{\mathcal{T}}_h$ with $\tau \cap D_h = \emptyset$ and $x_1^\tau, \dots, x_{d+1}^\tau \in \overline{D}$, we have $\operatorname{diam}(\tau) \lesssim h^2$.

Lemma 3.22. *Let $v \in H^{1+s}(D) \cap H_0^1(D)$, for some $0 < s \leq 1$, and let assumption T1 hold. Then*

$$\inf_{v_h \in \widehat{V}_{h,0}} |v - v_h|_{H^1(D)} \lesssim \|v\|_{H^{1+s}(D)} \, h^s. \tag{3.14}$$

Proof. This result is classical (for parts of the proof see [44, section 8.6] or [68]). Set $D_\delta := D \backslash \overline{D}_h$ where δ denotes the maximum width of D_δ, and let first $s = 1$. Since $v_h = 0$ on D_δ it suffices to show that

$$|v|_{H^1(D_\delta)} \lesssim \|v\|_{H^2(D)} \, h \,. \tag{3.15}$$

The result then follows for $s = 1$ with Lemma 3.21. The result for $s < 1$ follows by interpolation, since trivially, $|v|_{H^1(D_\delta)} \leq \|v\|_{H^1(D)}$,

To show (3.15), let $w \in H^2(D)$. Using a trace result we get

$$\|w\|_{H^1(D_\delta)} \leq \|w\|_{H^1(S_\delta)} \lesssim \delta^{1/2} \|w\|_{H^2(D)},$$

where $S_\delta = \{x \in D : \mathrm{dist}(x, \partial D) \leq \delta\} \subset D$ is the boundary layer of width δ. It follows from assumption T1 that $\mathrm{diam}(\tau) \lesssim h^2$ wherever the boundary is not convex. In regions where D is convex it follows from the smoothness assumption on ∂D that the width of D_δ is $\mathcal{O}(h^2)$. Hence $\delta \lesssim h^2$, which completes the proof of (3.15). □

Hence, the bounds on the finite element error hold for smooth domains also, provided assumption T1 is satisfied.

Proposition 3.23. *Let the assumptions of Theorem 3.3 be satisfied, and suppose assumption T1 holds. Then the bound given in Theorem 3.3 holds also for the smooth domain D.*

Proof. This follows as in Theorem 3.3, using Lemma 3.22 instead of Lemma 3.2. □

As in sections 3.1 and 3.2, Proposition 3.23 can now be used to prove optimal convergence rates of the L^2 error and the error in functionals $M_\omega(\cdot)$.

3.5 Application to finite volume methods

We will now use the finite element error analysis developed earlier in this chapter to prove convergence of some finite volume discretisations of model problem (2.1). For simplicity, we shall again restrict our attention to scalar coefficients $a(\omega, x)$ and homogeneous boundary conditions $\phi \equiv 0$.

The starting point of finite volume discretisations is a non-overlapping partitioning of the domain D into boxes (or *volumes*) $\mathcal{B}$. Equation (2.1) is then integrated over each box $B \in \mathcal{B}$, leading to the set of algebraic equations

$$-\int_B \mathrm{div}(a(\omega, x)\nabla u(\omega, x))\,\mathrm{d}x = \int_B f(\omega, x)\,\mathrm{d}x, \qquad \forall B \in \mathcal{B}.$$

The volume integral on the left hand side is transformed into a boundary integral using the Divergence Theorem:

$$-\int_{\partial B} a(\omega,x)\frac{\partial u}{\partial \mathbf{n}}(\omega,x)\,\mathrm{d}s = \int_B f(\omega,x)\,\mathrm{d}x \qquad \forall B \in \mathcal{B}. \tag{3.16}$$

The specific finite volume scheme is now determined by the choice of volumes $\mathcal{B}$, as well as how the integrals in (3.16) are computed. We will here consider two finite volume schemes: the box method as considered by Hackbusch in [43] (§3.5.1), and a finite volume method on uniform rectangular meshes (§3.5.2).

3.5.1 Triangular meshes

As before, let $\{\mathcal{T}_h\}_{h>0}$ be a shape-regular family of simplicial triangulations of the Lipschitz polygonal/polyhedral domain D, parametrised by its mesh width $h := \max_{\tau\in\mathcal{T}_h} \mathrm{diam}(\tau)$. Following [43], we restrict ourselves to the case $D \subset \mathbb{R}^2$ and $\phi = 0$, and define the corresponding box meshes $\mathcal{B}_h$ as dual meshes to the triangulations $\mathcal{T}_h$. We introduce the following notation:

$$\begin{aligned}
&C(\mathcal{T}_h) := \{P \in \overline{D} :\ P \text{ is a corner of some } \tau \in \mathcal{T}_h\},\\
&C_0(\mathcal{T}_h) := \{P \in D :\ P \text{ is an (interior) corner of some } \tau \in \mathcal{T}_h\},\\
&\mathcal{T}_h^P := \{\tau \in \mathcal{T}_h :\ P \text{ is a corner in } \tau\}, \quad \text{for } P \in C(\mathcal{T}_h).
\end{aligned}$$

The polygonal boxes $B \in \mathcal{B}_h$ are now defined by associating one box, denoted by B_P, to each point $P \in C(\mathcal{T}_h)$ in the following way:

- $P \in B_P$ for all $P \in C(\mathcal{T}_h)$,
- B_P and B_Q, for $P \neq Q$, intersect at most at their boundaries,
- $\bigcup_{P\in C(\mathcal{T}_h)} B_P = \overline{D}$,
- $B_P \subset \bigcup_{\tau\in\mathcal{T}_h^P} \tau$,
- the boundary ∂B_P intersects the edges of $\tau \in \mathcal{T}_h^P$ emanating from P at their midpoints.

Denote now by $u_h^{\mathrm{FV}} \in V_{h,0}$ the solution to (3.16) with the box mesh described above, i.e.

$$-\int_{\partial B_P} a(\omega,x)\frac{\partial u_h^{\mathrm{FV}}}{\partial \mathbf{n}}(\omega,x)\,\mathrm{d}s = \int_{B_P} f(\omega,x)\,\mathrm{d}x \qquad \forall P \in C_0(\mathcal{T}_h).$$

We assume that the integrals above are computed exactly. The following result is proved in [43].

Lemma 3.24. *Let $a_{\min}(\omega) > 0$ and $f(\omega,\cdot) \in L^2(D)$. Then*

$$\|u_h(\omega,\cdot) - u_h^{\mathrm{FV}}(\omega,\cdot)\|_{H^1(D)} \leq \frac{\|f(\omega,\cdot)\|_{L^2(D)}}{\sqrt{2}a_{\min}(\omega)}\,h.$$

We then immediately have the following convergence result.

Theorem 3.25. *Suppose $D \subset \mathbb{R}^2$, $\phi = 0$ and $\mathbf{A}(\omega,x) = a(\omega,x)I_d$. Let the assumptions of Theorem 3.3 hold, for some $0 < s_u \leq 1$ and $1 \leq p_u \leq \infty$, and suppose $f \in L^{p_u}(\Omega, L^2(D))$. Then, for all $p < p_u$ and $h > 0$, we have*

$$\|u - u_h^{\mathrm{FV}}\|_{L^p(\Omega,H_0^1(D))} \lesssim C_{3.25}\,h^{s_u},$$

with

$$C_{3.25} := \max\left[C_{3.3}, \left\|\frac{1}{a_{\min}}\right\|_{L^q(\Omega)} \|f\|_{L^{p_u}(\Omega,L^2(D))}\right],$$

where $q = \frac{p-p_u}{p\,p_u}$.

Proof. This follows immediately from Theorem 3.3 and Lemma 3.24, together with Minkowski's and Hölder's inequality. □

3.5.2 Uniform rectangular meshes

We now consider a finite volume scheme on a uniform square/cubic mesh. For simplicity, we only consider the case $D = (0,1)^2$ in detail. For some mesh width $h \leq 1$, the boxes B in $\mathcal{B}_h$ are in this case defined by

$$B_{i,j} := [x_1^{(i-1)}, x_1^{(i)}] \otimes [x_2^{(j-1)}, x_2^{(j)}], \quad \text{for } i,j \in \{1,\ldots,h^{-1}\},$$

where $x_1^{(k)} = x_2^{(k)} = kh$. To compute the boundary integral on the left hand side of (3.16), we approximate the permeability a on each of the edges of $\partial B_{i,j}$ by a constant. The value of this constant is taken to be the permeability at the midpoint of this edge. Denote the resulting approximation of the permeability by $a^{\mathrm{FV}}(\omega, x)$, and denote by $u_h^{\mathrm{FV}} \in V_{h,\phi}$ the corresponding solution to

$$-\int_{\partial B_{i,j}} a^{\mathrm{FV}}(\omega, x) \frac{\partial u_h^{\mathrm{FV}}}{\partial \mathbf{n}}(\omega, x)\,\mathrm{d}s = \int_{B_{i,j}} f(\omega, x)\,\mathrm{d}x \qquad \forall i, j = 1, \ldots, h^{-1}. \quad (3.17)$$

Denote now by u_h the bilinear finite element solution on the mesh $\mathcal{B}_h$, as defined in (3.1). The quasi-optimality result in Lemma 3.1 (Cea's Lemma) and the best approximation results in Lemmas 3.2 and 3.12 hold also for the square mesh $\mathcal{B}_h$, and the error estimate on the finite element error with quadrature in Proposition 3.13 is hence applicable.

It turns out that the finite volume solution u_h^{FV} defined in (3.17) is equivalent to the (approximate) finite element solution on $\mathcal{B}_h$ with a particular quadrature scheme that is a mixture of the midpoint and trapezoidal rules. In particular, recall that the computation of the finite element solution requires evaluation of the integrals

$$\int_{B_{i,j}} a(\omega, x) \nabla w_h(x) . \nabla v_h(x)\,\mathrm{d}x,$$

for $w_h, v_h \in V_{h,0}$. Expanding the dot product, the above integral splits into the sum of two integrals. We approximate the first integral, which involves derivatives with respect to the first coordinate direction x_1, by the midpoint rule in x_1 and the trapezoidal rule in x_2:

$$\int_{B_{i,j}} f(x)\,\mathrm{d}x \approx \frac{h^2}{2}\left(f(x_1^{(i-1/2)}, x_2^{(j)}) + f(x_1^{(i-1/2)}, x_2^{(j+1)})\right),$$

where $x_1^{(i-1/2)} = (i-1/2)h$. The second integral, involving derivatives with respect to x_2, is similarly approximated by the midpoint rule in x_2 and the trapezoidal rule in x_1. Explicit computations now show that this leads to the same set of equations as (3.17) (see [14, Exercise 4.1.8] and [8, §3.3] for more details).

We then have the following convergence result.

Lemma 3.26. *Let the assumptions of Proposition 3.13 hold for some* $0 < s_u \leq 1$ *and* $1 \leq p_u \leq \infty$. *Then, for all* $p < p_u$, *we have*

$$\|u - u_h^{\mathrm{FV}}\|_{L^p(\Omega, H^1(D))} \lesssim C_{3.13}\, h^{s_u}.$$

Proof. This follows immediately from Proposition 3.13, since the quadrature scheme described in this section uses a linear combination of point evaluations of a and hence fits into the framework of quadrature schemes discussed in section 3.4.1. □

Remark 3.27. The results in this section can easily be extended to three (or one) spatial dimensions. The quadrature scheme which makes the approximate finite element and the finite volume solution equivalent, is the one which uses the midpoint rule in the coordinate direction in which the derivatives are taken, and the trapezoidal rule in the remaining coordinate directions. See e.g. [8, §3.3] for more details.

3.6 Numerics

In this section, we want to confirm numerically some of the results proved in earlier sections. As model the permeability, we choose a scalar (piecewise) continuous log-normal random field as described in section 2.7. We consider two different model problems in 2D, both in the unit square $D = (0,1)^2$: either (2.1) with $f \equiv 1$ and $\phi \equiv 0$, i.e.

$$-\nabla \cdot (a(\omega, x)\nabla u(\omega, x)) = 1, \quad \text{for } x \in D, \quad \text{and} \quad u(\omega, x) = 0 \quad \text{for} \quad x \in \partial D, \tag{3.18}$$

or the mixed boundary value problem

$$\begin{aligned} -\nabla \cdot (a(\omega, x)\nabla u(\omega, x)) &= 0, \qquad \text{for} \quad x \in D, \\ \text{and} \quad u\big|_{x_1=0} = 1, \quad u\big|_{x_1=1} = 0, \quad \frac{\partial u}{\partial \mathbf{n}}\Big|_{x_2=0} &= 0, \quad \frac{\partial u}{\partial \mathbf{n}}\Big|_{x_2=1} = 0. \end{aligned} \tag{3.19}$$

To produce samples of fields with exponential covariance function (2.25), we use a circulant embedding technique [25, 39]. In contrast to a truncated KL-expansion, this technique gives exact samples of the full field $g(\omega, x)$ at the vertices of our spatial grid. Fields with Gaussian covariance (2.26) are approximated by Karhunen–Loève expansions truncated after $R^* = 170$ terms. The eigenpairs of the covariance operator are computed numerically using a spectral collocation

method. Similar to the analysis in section 3.4.1, we then use the trapezoidal rule to approximate the integrals in the stiffness matrix. To estimate the errors, we approximate the exact solution u by a reference solution u_{h^*} on a grid with mesh width $h^* = 1/256$.

Let us start with a discontinuous model of the permeability on a fixed (deterministic) partitioning of D. A rock formation which is often encountered in applications is a channelised medium. To simulate this, we divide $D = (0,1)^2$ into 3 horizontal layers, and model the permeabilities in the 3 layers by 2 different log–normal distributions. The middle layer occupies the region $\{1/3 \leq x_2 \leq 2/3\}$. The parameters in the top and bottom layer are taken to be $\mu_1 = 0$, $\lambda_1 = 0.3$ and $\sigma_1^2 = 1$, and for the middle layer we take $\mu_2 = 4$, $\lambda_2 = 0.1$ and $\sigma_2^2 = 1$ (assuming no correlation across layers). As a test problem we choose the flow cell model problem (3.19).

We start with the norms $\|u_{h^*} - u_h\|_{L^2(D)}$ and $|u_{h^*} - u_h|_{H^1(D)}$. Figures 3-1 and 3-2 show results for fields with exponential and Gaussian covariance functions, respectively. For comparison, we have added the graphs for the case where there is no "channel", i.e. where the permeability field is one continuous log–normal field with $\mu = \mu_1 = 0$, $\lambda = \lambda_1 = 0.3$ and $\sigma^2 = \sigma_1^2 = 1$. As expected from the global regularity results in Theorems 2.12 and 2.17, we observe the same convergence rates for both the continuous and the discontinuous permeability fields in the case of an exponential covariance in Figure 3-1. We observe $O(h^{1/2})$ convergence of the $H^1(D)$–seminorm of the error, and linear convergence of the $L^2(D)$–norm of the error. The quadrature error seems not to be dominant (cf section 3.4.1). For the Gaussian covariance, however, we indeed observe slower convergence rates for the layered medium (Figure 3-2). Whereas we observe $O(h^{1/2})$ convergence of the $H^1(D)$–seminorm, and linear convergence of the $L^2(D)$–norm for the layered medium, we have linear convergence of the $H^1(D)$–seminorm, and quadratic convergence of the $L^2(D)$–norm for the continuous permeability field. Since the slower convergence rates are caused by singularities at the interfaces, one could of course use local mesh refinement near the interfaces in order to recover the faster convergence rates also for the layered medium.

For the continuous permeability field described above, with 2-norm exponential covariance function with $\mu = 0, \lambda = 0.3$ and $\sigma^2 = 1$, let us now consider some of the functionals from section 2.5. First, we consider the approximation of

the second moment of the pressure at the centre of the domain for the Dirichlet model problem (3.18). As described in §2.5 for functional $M^{(2)}$, we approximate it by the average of u_h over the region D^*, which is chosen to consist of the six elements (of a uniform grid with $h^* = 1/256$) adjacent to the node at $(1/2, 1/2)$. The results for the estimation of the second moment are shown in the right plot in Figure 3-3. We see that $\left|\mathbb{E}\left[M^{(3)}(u_{h^*}) - M^{(3)}(u_h)\right]\right|$ converges linearly in h, as predicted by Lemma 4.5 for the exact FE solution. The quadrature error seems to again not be dominant.

For the mixed model problem (3.19), we consider an approximation of the average outflow through the boundary $\Gamma_{\text{out}} := \{x_1 = 1\}$ computed via the functional $M_\omega^{(4)}$ in §2.5. As the weight function, we choose the linear function $\psi(x) = x_1$, which is equal to 1 at all nodes on Γ_{out} and equal to 0 at all other Dirichlet nodes. Thus, $M_\omega^{(4)}(u)$ is exactly equal to the flow through Γ_{out}. As predicted we see again linear convergence in h for $\left|\mathbb{E}\left[M_\omega^{(4)}(u_{h^*}) - M_\omega^{(4)}(u_h)\right]\right|$ in the left plot in Figure 3-3.

The convergence of $\|u_{h^*} - u_h\|_{L^\infty(D)}$ for model problem (3.18) with continuous permeability field with 2-norm exponential covariance with $\lambda = 0.3$ and $\sigma^2 = 1$ is shown in Figure 3-4. Although Proposition 2.19 suggests a convergence rate of $\mathcal{O}(h^{3/2})$, we observe a slightly slower convergence which is still better than linear convergence. This lower convergence rate might be due to quadrature error.

Finally, let us finish this section with a model problem where the permeability is piecewise constant on a random partitioning of the domain. As before, we divide $D = (0, 1)^2$ into three horizontal regions. For a given ω, the regions are constructed by sampling from uniform random variables $y_1 \sim \text{Unif}(0.8, 0.9)$, $y_2 \sim \text{Unif}(0.6, 0.7)$, $y_3 \sim \text{Unif}(0.2, 0.3)$ and $y_4 \sim \text{Unif}(0.4, 0.5)$, and then drawing straight lines between the points $(0, y_1)$ and $(1, y_2)$, and $(0, y_3)$ and $(1, y_4)$, respectively. This ensures that the subregions are always convex. Furthermore, we then sample from three independent, standard normal random variables z_1, z_2 and z_3, and then set the permeability values in the three subregions to $\exp[z_1], \exp[z_2]$ and $\exp[z_3]$, respectively. In Figure 3-5, we observe $O(h^{1/2})$ convergence of the $H^1(D)$–seminorm of the error, and (slightly faster than) linear convergence of the $L^2(D)$–norm of the error, as predicted by the global regularity results in Theorem 2.17.

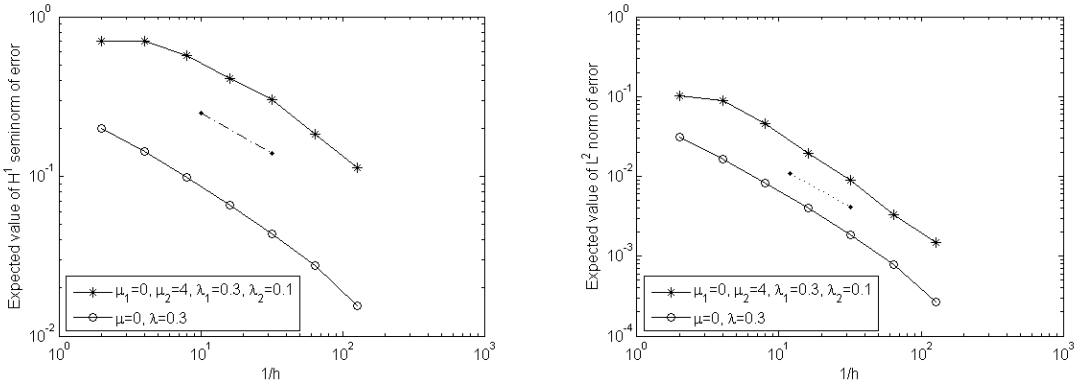

Figure 3-1: Left plot: $\mathbb{E}\big[|u_{h^*} - u_h|_{H^1(D)}\big]$ versus $1/h$ for model problem (3.19) with $d = 2$ and 2–norm exponential covariance, with $\mu = \mu_1 = 0$, $\mu_2 = 4$, $\lambda = \lambda_1 = 0.3$, $\lambda_2 = 0.1$, $\sigma^2 = \sigma_1^2 = \sigma_2^2 = 1$ and $h^* = 1/256$. Right plot: $\mathbb{E}\big[\|u_{h^*} - u_h\|_{L^2(D)}\big]$. The gradient of the dash–dotted (resp. dotted) line is $-1/2$ (resp. -1).

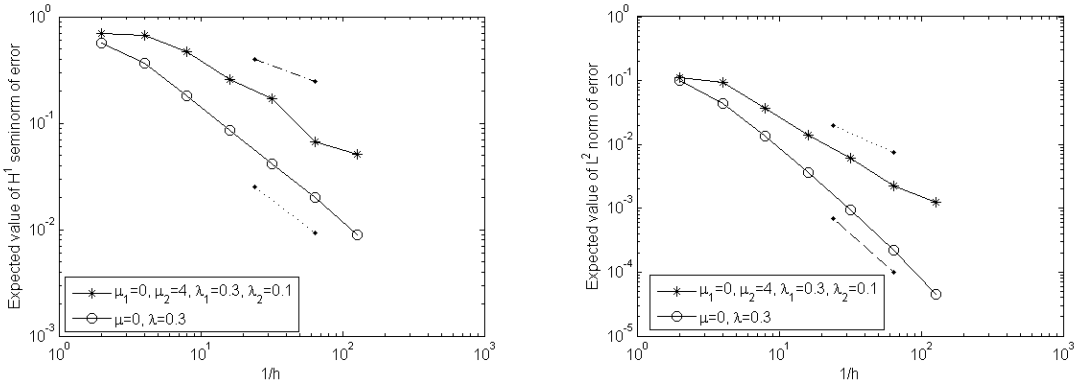

Figure 3-2: Left plot: $\mathbb{E}\big[|u_{h^*} - u_h|_{H^1(D)}\big]$ versus $1/h$ for model problem (3.19) with $d = 2$ and Gaussian covariance, with $\mu = \mu_1 = 0$, $\mu_2 = 4$, $\lambda = \lambda_1 = 0.3$, $\lambda_2 = 0.1$, $\sigma^2 = \sigma_1^2 = \sigma_2^2 = 1$, $h^* = 1/256$ and $K^* = 170$. Right plot: $\mathbb{E}\big[\|u_{h^*} - u_h\|_{L^2(D)}\big]$. The gradient of the dash–dotted (resp. dotted and dashed) line is $-1/2$ (resp. -1 and-2).

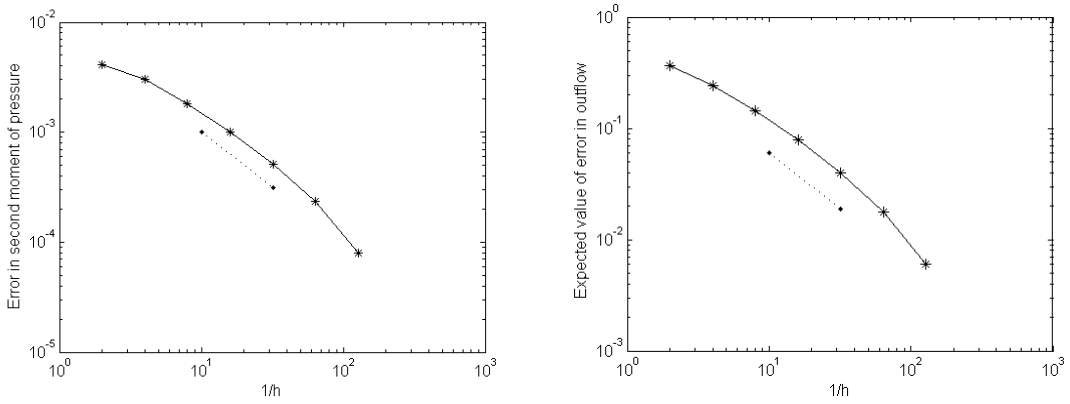

Figure 3-3: Left plot: $\left|\mathbb{E}\left[M^{(3)}(u_{h^*}) - M^{(3)}(u_h)\right]\right|$ versus $1/h$ for model problem (3.18) with $d = 2$ and 2-norm exponential covariance with $\mu = 0, \lambda = 0.3$, $\sigma^2 = 1$ and $h^* = 1/256$. Right plot: $\left|\mathbb{E}\left[M^{(4)}_{\omega}(u_{h^*}) - M^{(4)}_{\omega}(u_h)\right]\right|$ versus $1/h$ for model problem (3.19) with $d = 2$ and 2-norm exponential covariance with $\mu = 0, \lambda = 0.3$, $\sigma^2 = 1$, $\psi = x_1$ and $h^* = 1/256$. The gradient of the dotted line is -1.

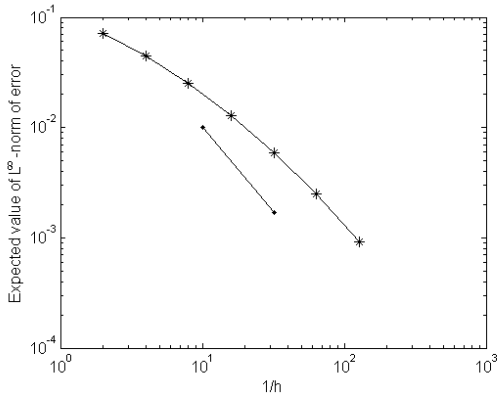

Figure 3-4: $\mathbb{E}\left[\|u_{h^*} - u_h\|_{L^\infty(D)}\right]$ versus $1/h$ for model problem (3.18) with $d = 2$ and 2-norm exponential covariance with $\mu = 0, \lambda = 0.3$, $\sigma^2 = 1$ and $h^* = 1/256$. The gradient of the (non-marked) solid line is $-3/2$.

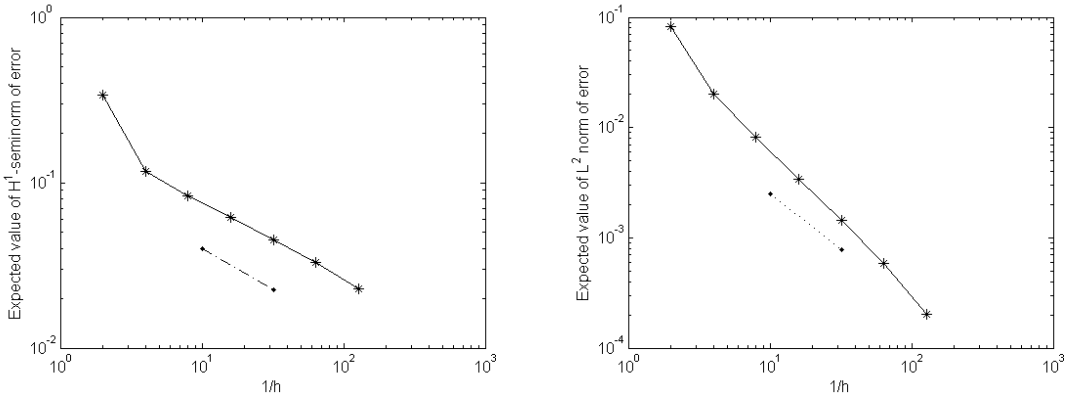

Figure 3-5: Left plot: $\mathbb{E}\big[|u_{h^*} - u_h|_{H^1(D)}\big]$ versus $1/h$ for model problem (3.19) with $d = 2$ and piecewise constant permeability on random subdomains. Right plot: $\mathbb{E}\big[\|u_{h^*} - u_h\|_{L^2(D)}\big]$. The gradient of the dash–dotted (resp. dotted) line is $-1/2$ (resp. -1).

Chapter 4

Multilevel Monte Carlo methods

We will now apply the discretisation error analysis in chapter 3, to give a rigorous bound on the cost of the multilevel Monte Carlo method applied to (2.1), and to establish its superiority over the classical Monte Carlo method. We start by describing the classical Monte Carlo (MC) and multilevel Monte Carlo (MLMC) algorithms for PDEs with random coefficients.

In the Monte Carlo framework, we are usually interested in finding the expected value of some functional $Q = \mathcal{G}(u)$ of the solution u to our model problem (2.1). This may be a single component or a norm of u, or it may be a more complicated nonlinear functional (e.g. a higher order moment). Since u is not easily accessible, Q is often approximated by the quantity $Q_h := \mathcal{G}(u_h)$, where u_h denotes a discretised solution on a sufficiently fine spatial grid $\mathcal{T}_h$. This could for example be the finite element or finite volume solution described in chapter 3, and can include further approximations, such as the truncated solution $u_{R,h}$ described in section 3.4.2.

We assume that the expected value $\mathbb{E}[Q_h] \to \mathbb{E}[Q]$ as $h \to 0$, and that (in mean) the order of convergence is α, i.e.

$$|\mathbb{E}[Q_h - Q]| \lesssim h^\alpha.$$

Thus, to estimate $\mathbb{E}[Q]$, we compute approximations (or *estimators*) $\widehat{Q}_h$ to $\mathbb{E}[Q_h]$, and quantify the accuracy of our approximations via the root mean square error (RMSE)

$$e(\widehat{Q}_h) := \Big(\mathbb{E}\big[(\widehat{Q}_h - \mathbb{E}(Q))^2\big]\Big)^{1/2}.$$

The computational cost of the estimator, denoted by $\mathcal{C}(\widehat{Q}_h)$, is then quantified by the number of floating point operations that are needed to compute it. The computational cost of the estimator to achieve a RMSE of $e(\widehat{Q}_h) \leq \varepsilon$ will be referred to as the ε–cost and denoted by $\mathcal{C}_\varepsilon(\widehat{Q}_h)$.

4.1 Standard Monte Carlo simulation

The classical Monte Carlo (MC) estimator for $\mathbb{E}[Q_h]$ is

$$\widehat{Q}_{h,N}^{\mathrm{MC}} := \frac{1}{N} \sum_{i=1}^{N} Q_h^{(i)}, \tag{4.1}$$

where $Q_h^{(i)}$ is the ith sample of Q_h and N independent samples are computed in total. We assume that the cost to compute one sample $Q_h^{(i)}$ of Q_h is

$$\mathcal{C}(Q_h^{(i)}) \lesssim h^{-\gamma}, \qquad \text{for some } \gamma > 0.$$

There are two sources of error in the estimator (4.1): the approximation of Q by Q_h, which is related to the spatial discretisation in the case of our PDE application, and the sampling error due to replacing the expected value by a finite sample average. The contribution of both of these errors becomes clear when we expand the mean square error (MSE):

$$\begin{aligned} e(\widehat{Q}_{h,N}^{\mathrm{MC}})^2 &= \mathbb{E}\left[\left(\widehat{Q}_{h,N}^{\mathrm{MC}} - \mathbb{E}[\widehat{Q}_{h,N}^{\mathrm{MC}}] \;+\; \mathbb{E}[\widehat{Q}_{h,N}^{\mathrm{MC}}] - \mathbb{E}[Q]\right)^2\right] \\ &= \mathbb{E}\left[(\widehat{Q}_{h,N}^{\mathrm{MC}} - \mathbb{E}[\widehat{Q}_{h,N}^{\mathrm{MC}}])^2\right] + \left(\mathbb{E}[\widehat{Q}_{h,N}^{\mathrm{MC}}] - \mathbb{E}[Q]\right)^2 \\ &= \mathbb{V}[\widehat{Q}_{h,N}^{\mathrm{MC}}] + \left(\mathbb{E}[\widehat{Q}_{h,N}^{\mathrm{MC}}] - \mathbb{E}[Q]\right)^2. \end{aligned} \tag{4.2}$$

It is well known for standard MC that

$$\mathbb{E}[\widehat{Q}_{h,N}^{\mathrm{MC}}] \;=\; \mathbb{E}[Q_h] \qquad \text{and} \qquad \mathbb{V}[\widehat{Q}_{h,N}^{\mathrm{MC}}] \;=\; N^{-1}\,\mathbb{V}[Q_h]\,.$$

Substituting this in (4.2) we get

$$e(\widehat{Q}_{h,N}^{\mathrm{MC}})^2 \;=\; N^{-1}\mathbb{V}[Q_h] + \left(\mathbb{E}[Q_h - Q]\right)^2. \tag{4.3}$$

So the first term in the MSE is the variance of the MC estimator, which represents the sampling error and decays inversely with the number of samples. The second term is the square of the error in mean between Q_h and Q.

Hence, a sufficient condition to achieve a RMSE of ε with this estimator is that both of the terms are less than $\varepsilon^2/2$. Under the assumption that $\mathbb{V}(Q_h)$ is constant independent of h, this can be achieved by choosing $N \gtrsim \varepsilon^{-2}$ and $h \lesssim \varepsilon^{1/\alpha}$, where the convergence rate α is as defined previously and problem dependent. In other words, we need to take a large enough number of samples N, as well as a small enough value for h, so that $\widehat{Q}_{h,N}^{\mathrm{MC}}$ is a sufficiently accurate approximation of our quantity of interest $\mathbb{E}[Q]$.

Since the cost to compute one sample of Q_h was assumed to satisfy $\mathcal{C}(Q_h^{(i)}) \lesssim h^{-\gamma}$, we have $\mathcal{C}(\widehat{Q}_{h,N}^{\mathrm{MC}}) \lesssim N h^{-\gamma}$ and so the total computational cost of achieving a RMSE of $O(\varepsilon)$ is

$$\mathcal{C}_\varepsilon(\widehat{Q}_{h,N}^{\mathrm{MC}}) \lesssim \varepsilon^{-2-\gamma/\alpha}.$$

4.2 Multilevel Monte Carlo simulation

The main idea of multilevel Monte Carlo (MLMC) simulation is very simple. We sample not just from one approximation Q_h of Q, but from several. Let us recall the main ideas and the main theorem from [33, 15].

Let $\{h_\ell\}_{\ell=0,\dots,L}$ be the mesh widths of a sequence of increasingly fine triangulations $\mathcal{T}_{h_\ell}$ with $h := h_L$, the finest mesh width, and assume for simplicity that there exists an $s \in \mathbb{N}\backslash\{1\}$ such that

$$h_\ell = s^{-1}\, h_{\ell-1}\,, \qquad \text{for all } \ell = 1, \dots, L. \tag{4.4}$$

As for multigrid methods applied to discretised (deterministic) PDEs, the key is to avoid estimating $\mathbb{E}[Q_{h_\ell}]$ directly on level ℓ, but instead to estimate the correction with respect to the next lower level, i.e. $\mathbb{E}[Y_\ell]$ where $Y_\ell := Q_{h_\ell} - Q_{h_{\ell-1}}$. Linearity of the expectation operator then implies that

$$\mathbb{E}[Q_h] = \mathbb{E}[Q_{h_0}] + \sum_{\ell=1}^{L} \mathbb{E}[Q_{h_\ell} - Q_{h_{\ell-1}}] = \sum_{\ell=0}^{L} \mathbb{E}[Y_\ell], \tag{4.5}$$

where for simplicity we have set $Y_0 := Q_{h_0}$.

Hence, the expectation on the finest level is equal to the expectation on the coarsest level, plus a sum of corrections adding the difference in expectation between simulations on consecutive levels. The multilevel idea is now to independently estimate each of these expectations such that the overall variance is minimised for a fixed computational cost.

Let now $\widehat{Y}_\ell$ be an unbiased estimator for $\mathbb{E}[Y_\ell]$. The *multilevel* estimator is then simply defined as

$$\widehat{Q}_h^{\mathrm{ML}} \;:=\; \sum_{\ell=0}^{L} \widehat{Y}_\ell. \tag{4.6}$$

A possible choice for the estimators $\widehat{Y}_\ell$ are the standard Monte Carlo estimators

$$\widehat{Y}_{0,N_0}^{\mathrm{MC}} := \widehat{Q}_{h_0,N_0}^{\mathrm{MC}}, \;\text{ and }\; \widehat{Y}_{\ell,N_\ell}^{\mathrm{MC}} \;:=\; \frac{1}{N_\ell}\sum_{i=1}^{N_\ell}\Big(Q_{h_\ell}^{(i)} - Q_{h_{\ell-1}}^{(i)}\Big), \;\text{ for } \ell \geq 1\,. \tag{4.7}$$

The resulting multilevel estimator is denoted by $\widehat{Q}_{h,\{N_\ell\}}^{\mathrm{MLMC}}$ and referred to as the *multilevel Monte Carlo* estimator. It is important to note that the quantity $Q_{h_\ell}^{(i)} - Q_{h_{\ell-1}}^{(i)}$ in (4.7) comes from using the same random sample $\omega^{(i)} \in \Omega$ on both levels ℓ and $\ell - 1$.

Since all the expectations $\mathbb{E}[Y_\ell]$ are estimated independently, the variance of the MLMC estimator is $\mathbb{V}[\widehat{Q}_h^{\mathrm{ML}}] = \sum_{\ell=0}^{L}\mathbb{V}[\widehat{Y}_\ell]$, and expanding as in (4.2-4.3) in the previous section leads again to the following form for the MSE:

$$e(\widehat{Q}_{h,\{N_\ell\}}^{\mathrm{MLMC}})^2 \;:=\; \mathbb{E}\Big[\big(\widehat{Q}_{h,\{N_\ell\}}^{\mathrm{MLMC}} - \mathbb{E}[Q]\big)^2\Big] \;=\; \sum_{\ell=0}^{L} N_\ell^{-1}\,\mathbb{V}[Y_\ell] \;+\; \Big(\mathbb{E}[Q_h - Q]\Big)^2. \tag{4.8}$$

As in the single level MC case before, we see that the MSE consists of two terms, the variance of the estimator and the approximation error. Note that the second term is exactly the same as before in (4.2), and so it is again sufficient to choose $h = h_L \lesssim \varepsilon^{1/\alpha}$. To then achieve an overall RMSE of ε, the first term in (4.8) has to be less than $\varepsilon^2/2$ as well. We claim that this is cheaper to achieve in MLMC for two reasons:

- If Q_h converges to Q not just in mean, but also in mean square, then $\mathbb{V}[Y_\ell] = \mathbb{V}[Q_{h_\ell} - Q_{h_{\ell-1}}] \to 0$ as $\ell \to \infty$, and so it is possible to choose $N_\ell \to 1$ as $\ell \to \infty$.

- The coarsest level $\ell = 0$ and thus h_0 can be kept fixed for all ε, and so the cost per sample on level $\ell = 0$ does not grow as $\varepsilon \to 0$.

In practical applications, h_0 must be chosen sufficiently small to provide a minimal level of resolution of the problem. In our PDE application, this cut-off point is related to the spatial regularity of the solution u, which in turn depends on the regularity of the covariance function of the random coefficient and on the correlation length λ. We will return to this point in section 4.4.

The computational cost of the multilevel Monte Carlo estimator is

$$\mathcal{C}(\widehat{Q}^{\mathrm{MLMC}}_{h,\{N_\ell\}}) \;=\; \sum_{\ell=0}^{L} N_\ell \, \mathcal{C}_\ell .$$

where $\mathcal{C}_\ell := \mathcal{C}(Y_\ell^{(i)})$ represents the cost of a single sample of Y_ℓ. Treating the N_ℓ as continuous variables, the variance of the MLMC estimator is minimised for a fixed computational cost by choosing

$$N_\ell \;\eqsim\; \sqrt{\mathbb{V}[Y_\ell]/\mathcal{C}_\ell}\,, \tag{4.9}$$

with the constant of proportionality chosen so that the overall variance is $\varepsilon^2/2$. The total cost on level ℓ is then proportional to $\sqrt{\mathbb{V}[Y_\ell]\,\mathcal{C}_\ell}$ and hence

$$\mathcal{C}(\widehat{Q}^{\mathrm{MLMC}}_{h,\{N_\ell\}}) \;\lesssim\; \sum_{\ell=0}^{L} \sqrt{\mathbb{V}[Y_\ell]\,\mathcal{C}_\ell}.$$

If the variance $\mathbb{V}[Y_\ell]$ decays faster with ℓ than $\mathcal{C}_\ell$ increases, the dominant term will be on level 0. Since $N_0 \eqsim \varepsilon^{-2}$, the cost savings compared to standard MC will in this case be approximately $\mathcal{C}_0/\mathcal{C}_L \eqsim (h_L/h_0)^\gamma \eqsim \varepsilon^{\gamma/\alpha}$, reflecting the ratio of the costs of samples on level 0 compared to samples on level L.

If the variance $\mathbb{V}[Y_\ell]$ decays slower than the cost $\mathcal{C}_\ell$ increases, the dominant term will be on the finest level L, and the cost savings compared to standard MC will be approximately $\mathbb{V}[Y_L]/\mathbb{V}[Y_0]$ which is $O(\varepsilon^2)$, if we have truncated the telescoping sum in (4.5) with h_0 such that $\mathbb{V}[Y_0] \eqsim \mathbb{V}[Q_0]$. Hence, in both cases we have a significant gain.

This outline analysis is made more precise in the following section. Let us finish this section by discussing how the MLMC algorithm can be implemented

in practice.

The (optimal) values of L and $\{N_\ell\}_{\ell=0}^L$ can be computed "on the fly" from the sample averages and the (unbiased) sample variances of Y_ℓ. To do this we need to assume further that there exists an $h' \in \mathbb{N}$ such that the decay in $|\mathbb{E}[Q_h - Q]|$ is actually monotonic for $h \leq h'$ and satisfies

$$|\mathbb{E}[Q_h - Q]| \eqsim h^\alpha.$$

This ensures (via the triangle inequality) that $|\mathbb{E}[Y_L]| \eqsim h^\alpha$ (since $s > 1$ in (4.4)), and thus $|\widehat{Y}_L| \eqsim h^\alpha$ for N_L sufficiently large, providing us with a computable error estimator to determine whether h is sufficiently small or whether the number of levels L needs to be increased. It can in fact even be used to further improve the MLMC estimate by eliminating the leading order bias term via Richardson extrapolation (see [33, §4.2] for details).

Putting these ideas together, the MLMC algorithm can be implemented in practice as follows:

1) Start with L=0.

2) Estimate $\mathbb{V}[Y_L]$ by the sample variance of an initial number of samples.

3) Calculate the optimal N_ℓ, $\ell = 0, 1, \ldots, L$ using (4.9).

4) Evaluate extra samples at each level as needed for the new N_ℓ.

5) If $L \geq 1$, test for convergence using $\widehat{Y}_L \eqsim h^\alpha$.

6) If not converged, set $L = L + 1$ and go back to 2.

Note that in the above algorithm, step 3 aims to make the variance of the MLMC estimator less than $\frac{1}{2}\varepsilon^2$, while step 5 tries to ensure that the remaining bias is less than $\frac{1}{\sqrt{2}}\varepsilon$.

4.3 Convergence analysis

4.3.1 Abstract convergence theorem

We give a convergence analysis for general multilevel estimators $\widehat{Q}_h^{\mathrm{ML}}$ based on unbiased estimators $\widehat{Y}_\ell$ on each level. Theorem 4.1 below contains a parameter

δ, which determines the convergence of the variance of the estimator $\widehat{Y}_\ell$ with respect to the total number of samples N_ℓ. More precisely, δ is such that $\mathbb{V}[\widehat{Y}_\ell] = N_\ell^{-1/\delta}\mathbb{V}[Y_\ell]$. In the case of the standard Monte Carlo estimators defined in (4.7), we have $\delta = 1$ (cf (4.8)). If the multilevel estimator $\widehat{Q}_h^{\mathrm{ML}}$ is built using Quasi-Monte Carlo estimators on each level, then it possible to achieve any $\delta \in (1/2, 1]$ (see e.g. [39, 38] for more details).

Theorem 4.1. *Let $\varepsilon < \exp[-1]$. Suppose the sequence $\{h_\ell\}_{\ell=0,1,\dots}$ satisfies (4.4), and suppose there are constants $\alpha, \beta, \gamma, \delta, c_{\mathrm{M1}}, c_{\mathrm{M2}}, c_{\mathrm{M4}} > 0$ such that $\alpha \geq \frac{1}{2}\min(\beta, \delta^{-1}\gamma)$ and $\delta \in (1/2, 1]$. Under the assumptions*

M1. $|\mathbb{E}[Q_{h_\ell} - Q]| \leq c_{\mathrm{M1}}\, h_\ell^\alpha$

M2. $\mathbb{V}[\widehat{Y}_\ell] \leq c_{\mathrm{M2}}\, N_\ell^{-1/\delta}\, h_\ell^\beta$

M3. $\mathbb{E}[\widehat{Y}_\ell] = \begin{cases} \mathbb{E}[Q_0], & \ell = 0 \\ \mathbb{E}[Q_{h_\ell} - Q_{h_{\ell-1}}], & \ell > 0 \end{cases}$

M4. $\mathcal{C}(\widehat{Y}_\ell) \leq c_{\mathrm{M4}}\, N_\ell\, h_\ell^{-\gamma}$

there exists a sequence $\{N_\ell\}_{\ell=0}^L$ such that

$$e(\widehat{Q}_h^{\mathrm{ML}})^2 := \mathbb{E}\left[\left(\widehat{Q}_h^{\mathrm{ML}} - \mathbb{E}[Q]\right)^2\right] < \varepsilon^2,$$

where $h = h_L$, and

$$\mathcal{C}(\widehat{Q}_h^{\mathrm{ML}}) \leq c \begin{cases} \varepsilon^{-2\delta}, & \text{if } \delta\beta > \gamma, \\ \varepsilon^{-2\delta}(\log\varepsilon)^{1+\delta}, & \text{if } \delta\beta = \gamma, \\ \varepsilon^{-2\delta-(\gamma-\delta\beta)/\alpha}, & \text{if } \delta\beta < \gamma. \end{cases}$$

The constant c depends on $c_{\mathrm{M1}}, c_{\mathrm{M2}}$ and c_{M4}.

Proof. Recall that, by (4.4), we have

$$h_\ell = s^{-1}\, h_{\ell-1}\,, \qquad \text{for all } \ell = 1, \dots, L,$$

for some $s \in \mathbb{N}\setminus\{1\}$. Without loss of generality, we shall also assume that $h_0 = 1$. If this is not the case, this will only scale the constants c_{M1}, c_{M2} and c_{M4}.

Then, using the notation $\lceil x \rceil$ to denote the unique integer n satisfying the inequalities $x \leq n < x+1$, we start by choosing L to be

$$L = \left\lceil \alpha^{-1} \log_s(\sqrt{2}\, c_{\text{M1}}\, \varepsilon^{-1}) \right\rceil < \alpha^{-1} \log_s(\sqrt{2}\, c_{\text{M1}}\, \varepsilon^{-1}) + 1\,, \tag{4.10}$$

so that

$$s^{-\alpha} \frac{\varepsilon}{\sqrt{2}} \;<\; c_{\text{M1}}\, s^{-\alpha\, L} \;\leq\; \frac{\varepsilon}{\sqrt{2}}, \tag{4.11}$$

and hence, due to assumptions M1 and M3,

$$\left(\mathbb{E}[\widehat{Q}_h^{\text{ML}}] - \mathbb{E}[Q] \right)^2 \leq \tfrac{1}{2}\, \varepsilon^2.$$

This $\frac{1}{2}\varepsilon^2$ upper bound on the square of the bias error, together with the $\frac{1}{2}\varepsilon^2$ upper bound on the variance of the estimator to be proved later, gives an ε^2 upper bound on the estimator MSE.

Using the left-hand inequality in (4.11), we obtain the following inequality which will be used later,

$$\sum_{\ell=0}^{L} s^{\gamma\, \ell} \;<\; \frac{s^{\gamma L}}{1-s^{-\gamma}} \;<\; \frac{s^{\gamma}\, (\sqrt{2}\, c_{\text{M1}})^{\gamma/\alpha}}{1-s^{-\gamma}} \;\; \varepsilon^{-\gamma/\alpha}. \tag{4.12}$$

We now need to consider the different possible values for β.

a) If $\delta\beta = \gamma$, we set $N_\ell = \left\lceil 2^\delta\, \varepsilon^{-2\delta}\, (L+1)^\delta\, c_{\text{M2}}^\delta\, s^{-\beta\delta\,\ell} \right\rceil$ so that

$$\mathbb{V}[\widehat{Q}_h^{\text{ML}}] \;=\; \sum_{\ell=0}^{L} \mathbb{V}[\widehat{Y}_\ell] \;\leq\; \sum_{\ell=0}^{L} c_{\text{M2}}\, N_\ell^{-1/\delta} s^{-\beta\,\ell} \;\leq \tfrac{1}{2}\, \varepsilon^2 \frac{1}{L+1} \sum_{\ell=0}^{L} s^{(\beta-\beta)\ell} \;\leq\; \tfrac{1}{2}\, \varepsilon^2,$$

which is the required upper bound on the variance of the estimator. Since

$$N_\ell \leq 2^\delta\, \varepsilon^{-2\delta}\, (L+1)^\delta\, c_{\text{M2}}^\delta\, s^{-\beta\delta\,\ell} + 1,$$

the computational complexity is bounded by

$$\begin{aligned}\mathcal{C}(\widehat{Q}_h^{\mathrm{ML}}) &\le c_{\mathrm{M4}} \sum_{\ell=0}^{L} N_\ell \, s^{\gamma\,\ell} \\ &\lesssim \varepsilon^{-2\delta}(L+1)^{1+\delta} \, + \, \sum_{\ell=0}^{L} s^{\gamma\,\ell}\end{aligned}$$

For $\varepsilon < e^{-1} < 1$ we have $1 < \log \varepsilon^{-1}$ and $\varepsilon^{-\gamma/\alpha} \le \varepsilon^{-2\delta} \le \varepsilon^{-2\delta}(\log \varepsilon)^{1+\delta}$ since $\alpha \ge \frac{1}{2}\delta^{-1}\gamma$. Hence, using the inequalities in (4.10) and (4.12), it follows that $\mathcal{C}(\widehat{Q}_h^{\mathrm{ML}}) \lesssim \varepsilon^{-2\delta}(\log \varepsilon)^{1+\delta}$.

b) For $\delta\beta > \gamma$, we set $N_\ell = \left\lceil 2^\delta \, \varepsilon^{-2\delta} \, c_{\mathrm{M2}}^\delta \, \left(1 - s^{-(\delta\beta-\gamma)/(\delta+1)}\right)^{-\delta} s^{-(\beta+\gamma)\delta\ell/(\delta+1)} \right\rceil$ so that

$$\sum_{\ell=0}^{L} \mathbb{V}[\widehat{Y}_\ell] \le \tfrac{1}{2}\,\varepsilon^2 \, \left(1 - s^{-(\delta\beta-\gamma)/(\delta+1)}\right) \, \sum_{\ell=0}^{L} s^{-(\delta\beta-\gamma)\ell/(\delta+1)} \le \tfrac{1}{2}\,\varepsilon^2.$$

Since

$$N_\ell < 2^\delta \, \varepsilon^{-2\delta} \, c_{\mathrm{M2}}^\delta \, \left(1 - s^{-(\delta\beta-\gamma)/(\delta+1)}\right)^{-\delta} s^{-(\beta+\gamma)\delta\ell/(\delta+1)} + 1,$$

the computational complexity is bounded by

$$\begin{aligned}\mathcal{C}(\widehat{Q}_h^{\mathrm{ML}}) &\le c_{\mathrm{M4}} \left(2^\delta \, \varepsilon^{-2\delta} \, c_{\mathrm{M2}}^\delta \, \left(1 - s^{-(\delta\beta-\gamma)/(\delta+1)}\right)^{-\delta} \sum_{l=0}^{L} s^{-(\beta+\gamma)\delta\ell/(\delta+1)} \, s^{\gamma\,\ell} + \sum_{l=0}^{L} s^{\gamma\,\ell} \right) \\ &\lesssim \varepsilon^{-2\delta} + \sum_{l=0}^{L} s^{\gamma\,\ell}.\end{aligned}$$

Again for $\varepsilon < e^{-1} < 1$, we have $\varepsilon^{-\gamma/\alpha} \le \varepsilon^{-2\delta}$ since $\alpha \ge \frac{1}{2}\delta^{-1}\gamma$, and hence due to inequality (4.12) we have $\mathcal{C}(\widehat{Q}_h^{\mathrm{ML}}) \lesssim \varepsilon^{-2\delta}$.

c) For $\delta\beta < \gamma$, we set

$$N_\ell = \left\lceil 2^\delta \, \varepsilon^{-2\delta} \, c_{\mathrm{M2}}^\delta \, s^{(\gamma-\delta\beta)\delta L/(\delta+1)} \left(1 - s^{-(\gamma-\delta\beta)/(\delta+1)}\right)^{-\delta} s^{-(\beta+\gamma)\delta\ell/(\delta+1)} \right\rceil$$

so that

$$\sum_{\ell=0}^{L} \mathbb{V}[\widehat{Y}_\ell] \;\le\; \tfrac{1}{2}\,\varepsilon^2 \, s^{-(\gamma-\delta\beta)L/(\delta+1)} \left(1 - s^{-(\gamma-\delta\beta)/(\delta+1)}\right) \, \sum_{\ell=0}^{L} s^{(\gamma-\delta\beta)\delta\ell/(\delta+1)} \;\le\; \tfrac{1}{2}\,\varepsilon^2.$$

Since

$$N_\ell < 2^\delta\, \varepsilon^{-2\delta}\, c_{\mathrm{M2}}^\delta s^{(\gamma-\delta\beta)\delta L/(\delta+1)} \left(1-s^{-(\gamma-\delta\beta)/(\delta+1)}\right)^{-\delta} s^{-(\beta+\gamma)\delta\ell/(\delta+1)} + 1,$$

the computational complexity is bounded by

$$\begin{aligned}
&\mathcal{C}(\widehat{Q}_h^{\mathrm{ML}}) \\
&\leq c_{\mathrm{M4}} \left(\sum_{\ell=0}^{L} 2^\delta\, \varepsilon^{-2\delta}\, c_{\mathrm{M2}}^\delta\, s^{(\gamma-\delta\beta)\delta/L(\delta+1)} \left(1-s^{-(\gamma-\delta\beta)/(\delta+1)}\right)^{-\delta} \sum_{\ell=0}^{L} s^{(\gamma-\delta\beta)\ell/(\delta+1)} \right. \\
&\qquad \left. + \sum_{\ell=0}^{L} s^{\gamma\ell} \right) \\
&\lesssim \varepsilon^{-2\delta} s^{(\gamma-\delta\beta)L} + \sum_{\ell=0}^{L} s^{\gamma\ell}
\end{aligned}$$

Using the first inequality in (4.11),

$$s^{(\gamma-\delta\beta)L} < \left(\sqrt{2}\, c_{\mathrm{M1}}\right)^{(\gamma-\delta\beta)/\alpha} s^{(\gamma-\delta\beta)}\, \varepsilon^{-(\gamma-\delta\beta)/\alpha}.$$

Also, for $\varepsilon < e^{-1} < 1$ we have $\varepsilon^{-\gamma/\alpha} \leq \varepsilon^{-2\delta-(\gamma-\delta\beta)/\alpha}$ since $\alpha \geq \frac{1}{2}\beta$. Hence, due to inequality (4.12), we have $\mathcal{C}(\widehat{Q}_h^{\mathrm{ML}}) \lesssim \varepsilon^{-2\delta-(\gamma-\delta\beta)/\alpha}$. □

Remark 4.2. The geometric growth condition (4.4) is not necessary, and the conclusions of Theorem 4.1 hold provided $\{h_\ell\}_{\ell=0,\dots,L}$ satisfies $k_1 \leq h_{\ell-1}/h_\ell \leq k_2$, for all $\ell = 1, \dots, L$ and some $1 < k_1 \leq k_2 < \infty$.

4.3.2 Application of abstract convergence theorem

We will now verify assumptions M1 and M2 in Theorem 4.1 for a variety of functionals $\mathcal{G}(u)$. For this, we use the results from chapter 3 on the discretisation error. Since for the multilevel Monte Carlo estimator $\widehat{Q}_{h,\{N_\ell\}}^{\mathrm{MLMC}}$, it is well known that $\mathbb{V}[\widehat{Y}_{\ell,N_\ell}^{\mathrm{MC}}] = N_\ell^{-1}\mathbb{V}[Y_\ell]$, to prove assumption M1 and M2 we only need to prove $|\mathbb{E}[Q - Q_\ell]| \leq c_{\mathrm{M1}}\, h_\ell^\alpha$ and $\mathbb{V}[Y_\ell] \leq c_{\mathrm{M2}}\, h_\ell^\beta$, for some $\alpha, \beta, c_{\mathrm{M1}}, c_{\mathrm{M2}} > 0$.

We will start with the simple functionals $\|u\|_{L^2(D)}$ and $|u|_{H^1(D)}$.

Proposition 4.3. *Suppose that the assumptions of Theorem 3.3 hold, for some* $0 < s_u \leq 1$ *and* $p_u > 2$*, and let* $Q = |u|_{H^1(D)}^q$*, for some* $1 \leq q < p_u/2$*. Then*

assumptions M1–M2 in Theorem 4.1 hold with $\alpha = s_u$ and $\beta = 2s_u$.

Proof. Let $Q_{h_\ell} := |u_{h_\ell}|^q_{H^1(D)}$. Using the expansion $a^q - b^q = (a-b)\sum_{j=0}^{q-1} a^j b^{q-1-j}$, for $a, b \in \mathbb{R}$ and $q \in \mathbb{N}$, we get

$$\begin{aligned}
&|Q(\omega) - Q_{h_\ell}(\omega)| \\
&\lesssim |u(\omega,\cdot) - u_{h_\ell}(\omega,\cdot)|_{H^1(D)} \max\left\{|u(\omega,\cdot)|^{q-1}_{H^1(D)}, |u_{h_\ell}(\omega,\cdot)|^{q-1}_{H^1(D)}, 1\right\},
\end{aligned}$$

almost surely. This also holds for non-integer values of $q > 1$. Now, it follows from Lemma 2.1 and Theorem 3.3 that

$$|Q(\omega) - Q_{h_\ell}(\omega)| \lesssim \left(\frac{\mathbf{A}_{\max}(\omega)}{\mathbf{A}_{\min}(\omega)}\right)^{1/2} C_{\mathrm{R1}}(\omega)\, C^{q-1}_{2.1}(\omega)\, h_\ell^{s_u}, \quad \text{almost surely.}$$

Taking the expectation on both sides and applying Hölder's inequality, since $q < p_u$, it follows from assumptions A1–A2 and R1 that assumption M1 holds with $\alpha = s_u$.

To prove assumption M2, let us consider $Y_\ell = Q_{h_\ell} - Q_{h_{\ell-1}}$. As above, it follows from Lemma 2.1 and Theorem 3.3 together with the triangle inequality that

$$\begin{aligned}
&|Y_{h_\ell}(\omega)| \\
&\lesssim |u_{h_\ell}(\omega,\cdot) - u_{h_{\ell-1}}(\omega,\cdot)|_{H^1(D)} \max\left\{|u_{h_\ell}(\omega,\cdot)|^{q-1}_{H^1(D)}, |u_{h_{\ell-1}}(\omega,\cdot)|^{q-1}_{H^1(D)}, 1\right\} \\
&\lesssim \left(\frac{\mathbf{A}_{\max}(\omega)}{\mathbf{A}_{\min}(\omega)}\right)^{1/2} C_{\mathrm{R1}}(\omega)\, C^{q-1}_{2.1}(\omega)\, h_\ell^{s_u}, \quad \text{almost surely.}
\end{aligned}$$

where the hidden constant depends on s from (4.4). Since $q < p_u/2$ and $\mathbb{V}[Y_{h_\ell}] \leq \mathbb{E}[Y^2_{h_\ell}]$, it follows again from assumptions A1-A2 and R1, together with Hölder's inequality, that assumption M2 holds with $\beta = 2s_u$. □

Proposition 4.4. *Suppose that the assumptions of Corollary 3.4 hold, for some $0 < s_u \leq 1$ and $p_u > 2$, and let $Q := \|u\|^q_{L^2(D)}$, for some $1 \leq q < p_u/2$. Then assumptions M1 and M2 in Theorem 4.1 hold with $\alpha = 2s_u$ and $\beta = 4s_u$.*

Proof. This can be shown in the same way as Proposition 4.3, using Corollary 3.4 instead of Theorem 3.3. □

Let us now consider functionals as discussed in section 3.2.

Proposition 4.5. *Suppose the assumptions of Lemma 3.7 hold for some* $0 < s_u \leq 1$, $p_u > 2$, $0 \leq s_z \leq 1$ *and* $p_z > \frac{2p_u}{p_u - 2}$, *and let* $Q = M_\omega(u)$. *Then assumptions M1 and M2 in Theorem 4.1 hold with* $\alpha = s_u + s_z$ *and* $\beta = 2(s_u + s_z)$.

Proof. This follows immediately from Lemma 3.7, together with the triangle inequality and $\mathbb{V}[Y_{h_\ell}] \leq \mathbb{E}[Y_{h_\ell}^2]$. □

Finally, let us consider $\|u\|_{L^\infty(D)}$ and $|u|_{W^{1,\infty}(D)}$.

Proposition 4.6. *Suppose that the assumptions of Theorem 3.11 hold, for some* $0 < s_u \leq 1$ *and* $p_u > 2$, *and let* $Q := \|u\|_{L^\infty(D)}$. *Then assumptions M1 and M2 in Theorem 4.1 hold with* $\alpha = 1 + s$ *and* $\beta = 2(1 + s)$, *for any* $s < s_u$.

Proof. Using the reverse triangle inequality, we have almost surely

$$|Q(\omega) - Q_h(\omega)| \lesssim \|(u - u_h)(\omega, \cdot)\|_{L^\infty(D)}.$$

The claim then follows from Theorem 3.11, the triangle inequality and $\mathbb{V}[Y_{h_\ell}] \leq \mathbb{E}[Y_{h_\ell}^2]$. □

It is easy to show that the convergence rates in Proposition 4.6 hold not only for the L^∞-norm, but in fact for any point evaluation $u(x^*)$, for some $x^* \in D$. Since both u and u_h are almost surely continuous ([31]), it is meaningful to consider point evaluations, and it follows from the reverse triangle inequality that

$$||u(\omega, x^*)| - |u_h(\omega, x^*)|| \leq |u(\omega, x^*) - u_h(\omega, x^*)| \leq \|(u - u_h)(\omega, \cdot)\|_{L^\infty(D)}.$$

Proposition 4.7. *Suppose that the assumptions of Theorem 3.11 hold, for some* $0 < s_u \leq 1$ *and* $p_u > 2$, *and let* $Q := |u|_{W^{1,\infty}(D)}$. *Then assumptions M1 and M2 in Theorem 4.1 hold with* $\alpha = s$ *and* $\beta = 2s$, *for any* $s < s_u$.

Proof. This follows exactly as in Proposition 4.6. □

Similar to point evaluations of u, Proposition 4.7 can be used to prove convergence rates for point evaluations of the norm of the Darcy flux $\mathbf{A}\nabla u$, or the value of the Darcy flux in any given coordinate direction. Since u is almost surely continuously differentiable, and u_h is continuously differentiable in the interior of each element, it is meaningful to consider point evaluations of the fluxes $\mathbf{A}\nabla u$

	$\alpha = 1/2, \beta = 1$		$\alpha = 1, \beta = 2$		$\alpha = 3/2, \beta = 3$		$\alpha = 2, \beta = 4$	
d	MC	MLMC	MC	MLMC	MC	MLMC	MC	MLMC
1	ε^{-4}	ε^{-2}	ε^{-3}	ε^{-2}	$\varepsilon^{-8/3}$	ε^{-2}	$\varepsilon^{-5/2}$	ε^{-2}
2	ε^{-6}	ε^{-4}	ε^{-4}	ε^{-2}	$\varepsilon^{-10/3}$	ε^{-2}	ε^{-3}	ε^{-2}
3	ε^{-8}	ε^{-6}	ε^{-5}	ε^{-3}	ε^{-4}	ε^{-2}	$\varepsilon^{-7/2}$	ε^{-2}

Table 4.1: Theoretical upper bounds for the ε-costs of classical and multilevel Monte Carlo from Theorem 4.1. (For simplicity we wrote ε^{-p}, instead of $\varepsilon^{-p-\rho}$ with $\rho > 0$.)

and $\mathbf{A}\nabla u_h$ at any point $x^* \in D$ which is not on the boundary of any element $\tau \in \mathcal{T}_h$. The reverse triangle inequality again gives

$$||\mathbf{A}\nabla u(\omega, x^*)| - |\mathbf{A}\nabla u_h(\omega, x^*)|| \ \leq \ \mathbf{A}_{\max}(\omega)\|(u - u_h)(\omega, \cdot)\|_{W^{1,\infty}(D)}.$$

Substituting the convergence rates from Propositions 4.3–4.7 into Theorem 4.1, we can get theoretical upper bounds for the ε-costs of classical and multilevel Monte Carlo, as shown in Table 4.1. We assume here that $s_u = 1/2 - \delta$, for any $\delta > 0$, as is the case for log-normal random fields with exponential covariance. We assume that we can obtain individual samples in optimal cost $\mathcal{C}_\ell \lesssim h_\ell^{-d} \log(h_\ell^{-1})$ via a multigrid solver, i.e. $\gamma = d + \rho$ for any $\rho > 0$. We clearly see the advantages of the multilevel Monte Carlo method. Note that for small values of α and β, even though the actual computational cost of the estimator is larger, the savings from using the multilevel estimator are also larger, especially in low spatial dimensions.

Note also that since $\beta = 2\alpha$, we have that the cost of the multilevel estimator in the case $\beta < \gamma$ is proportional to $\varepsilon^{-\gamma/\alpha}$. This is of the same order as the cost of obtaining one sample on the finest grid, i.e. solving one deterministic PDE with the same regularity properties to accuracy ε. This implies that the method is optimal.

4.4 Level dependent estimators

The key ingredient in the multilevel Monte Carlo algorithm is the telescoping sum (4.5),

$$\mathbb{E}[Q_h] = \mathbb{E}[Q_{h_0}] + \sum_{\ell=1}^{L} \mathbb{E}[Q_{h_\ell} - Q_{h_{\ell-1}}].$$

We are free to choose how to approximate Q on the different levels, without violating the above identity, as long as the approximation of Q_{h_ℓ} is the same in the two terms in which it appears on the right hand side, for $\ell = 0, ..., L-1$. In particular, this implies that we can approximate the coefficient $a(\omega, x)$ differently on each level. Even though this strategy does not introduce any additional bias in the final result $\mathbb{E}[Q_h]$, it may influence the values of the convergence rates α and β in Theorem 4.1. One has to be careful not to introduce any additional model/approximation errors that decay at a slower rate than the discretisation error.

It is particularly useful when the random field $a(\omega, x)$ is highly oscillatory and varies on a fine scale. Coarse grids will not be able to resolve the coefficient well. As a consequence of this, one needs to choose the coarsest grid size h_0 smaller than a certain threshold to get the MLMC estimator with the smallest absolute cost. Numerical investigations in [15], for example, show that for log-normal random fields with underlying exponential, 1-norm covariance function and correlation length λ, the optimal choice is $h_0 \approx \lambda$. This limits the benefit that the MLMC estimator potentially offers. A possible solution to this problem is to use smoother approximations of the coefficient on the coarser levels. We will present one way of doing this, by using level-dependent truncations of the Karhunen-Lòeve expansion of $a(\omega, x)$.

As an exemplary case, let us now consider log-normal random fields a, where $\log[a]$ has exponential, 1-norm covariance, i.e. covariance function (2.25) with $\|x\| = \|x\|_1 := \sum_{i=1}^{d} |x_i|$.

Since the convergence with respect to R is quite slow (see §3.4.2), to get a good approximation to $\mathbb{E}[Q_h]$ we need to include a large number of terms on the finest grid, both in the case of the standard and the MLMC estimator. The eigenvalues $\{\mu_n\}_{n\in\mathbb{N}}$ are all non–negative with $\sum_{n\geq 1} \mu_n < +\infty$, and if they are ordered in decreasing order of magnitude, the corresponding eigenfunctions

$\{b_n\}_{n\in\mathbb{N}}$ will be ordered in increasing order of oscillations over D. By truncating the KL-expansion after fewer terms, we are hence disregarding the contributions of the most oscillatory eigenfunctions, leading to smoother problems that can be solved more accurately on the coarser levels. In order to determine a suitable strategy for the level dependent truncation of the KL–expansion, we make use of Theorem 3.18 and Corollary 3.19.

These results suggest that to balance out the two error contributions, we should choose R_ℓ as a power of h_ℓ. Note that a similar strategy was also suggested in the context of the related Brinkman problem in [37]. However, there, a certain decay rate for the error with respect to the number of KL-modes R was assumed. Here we make no such assumption and instead use Corollary 3.19 for the 1-norm exponential covariance. We have the following results for the multilevel Monte Carlo convergence rates in Theorem 4.1.

Proposition 4.8. *Provided assumption R3 is satisfied with $s_z \geq \frac{1}{2}$, and $R_\ell \gtrsim h_\ell^{-2}$, for all $\ell = 0, \dots, L$, then the convergence rate of the multilevel Monte Carlo method in §4.2 does not deteriorate when approximating the functional $M_\omega(u_{h_\ell})$ by $Q_{h_\ell} := M_\omega(u_{R_\ell,h_\ell})$ on each level ℓ. In particular, let the assumptions of Corollary 3.19 be satisfied with $p_u > 2$ and $p_z > \frac{2p_u}{p_u-2}$. Then assumptions M1–M2 in Theorem 4.1 hold for any $\alpha < 1$ and $\beta < 2$. If assumption R3 is satisfied only for some $s_z < 1/2$, then $R_\ell \gtrsim h_\ell^{-(1+2s_z)}$ is a sufficient condition.*

Proof. The proof is analogous to that of Proposition 4.5, using Corollary 3.19. □

As before, in the presence of quadrature error (cf. §3.4.1), we will not be able to get $\mathcal{O}(h^{s_u+s_z})$ convergence for the (spatial) discretisation error for the approximate finite element solution $\tilde{u}_{R,h}$. Due to the loss of Galerkin orthogonality for the primal problem, it is in general only possible to prove

$$\|M_\omega(u) - M_\omega(\tilde{u}_{R,h})\|_{L^p(\Omega)} = \mathcal{O}\left(h^{s_u} + R^{-\rho}\right),$$

for any $s_u, \rho < 1/2$. Thus with the quadrature error taken into account the optimal choice is $R_\ell \gtrsim h_\ell^{-1}$ for all functionals which satisfy assumption R3 with $s_z \geq 1/2$ and we will always use that in our numerical tests in section 4.5.3.

However, these results are asymptotic results, as $h_\ell \to 0$, and thus they only guarantee that level-dependent truncations do not deteriorate the performance

of the multilevel Monte Carlo method asymptotically as the tolerance $\varepsilon \to 0$. The real benefit of using level-dependent truncations is in absolute terms for a fixed tolerance ε, since the smoother fields potentially allow the use of coarser levels and thus significant gains in the absolute cost of the algorithm. In section 4.5.3, we see that this is in fact the case and we show the gains that are possible, especially for covariance functions with short correlation length λ.

4.5 Numerics

In this section, we want to confirm numerically some of the results proved in this chapter. We first confirm the values of α and β in Theorem 4.1, and then examine the cost of the various multilevel estimators. We again consider the two model problems (3.18) and (3.19) on $D = (0,1)^2$. Throughout this section, we model the permeability as a continuous log-normal random field $a(\omega, x)$, s.t. $\log[a]$ has mean zero and exponential covariance function (2.25).

4.5.1 Convergence rates

We want to confirm the rate of decay of $|\mathbb{E}[Q - Q_h]|$ and $\mathbb{V}[Q_h - Q_{2h}]$, for various quantities Q as considered in section 4.3.2. To estimate the error, we again approximate the solution u by a reference solution u_{h^*} on a fine mesh of width $h^* = 1/256$. We choose the 2-norm exponential covariance function for $\log[a]$ with $\lambda = 0.3$ and $\sigma^2 = 1$.

We start with the simple functionals $Q := \|u\|_{L^2(D)}$ and $Q = |u|_{H^1(D)}$ for model problem (3.18). We observe linear convergence for $|\mathbb{E}\left[\|u_{h^*}\|_{L^2(D)} - \|u_h\|_{L^2(D)}\right]|$, and quadratic convergence for $\mathbb{V}\left[\|u_h\|_{L^2(D)} - \|u_{2h}\|_{L^2(D)}\right]$ in Figure 4-1, as predicted by Proposition 4.4. For $Q = |u|_{H^1(D)}$, we in fact observe convergence rates which are slightly better that those proved in Proposition 4.3. We observe slightly faster than $\mathcal{O}(h^{1/2})$ convergence for $|\mathbb{E}\left[|u_{h^*}|_{H^1(D)} - |u_h|_{H^1(D)}\right]|$, and slightly faster than linear convergence for $\mathbb{V}\left[|u_h|_{H^1(D)} - |u_{2h}|_{H^1(D)}\right]$.

Let us now move on to more complicated functionals. As already described in section 3.6, we consider the approximation of the second moment of the pressure at the centre of the domain for the Dirichlet model problem (3.18), and the approximation of the average outflow through the boundary $\Gamma_{\text{out}} := \{x_1 = 1\}$

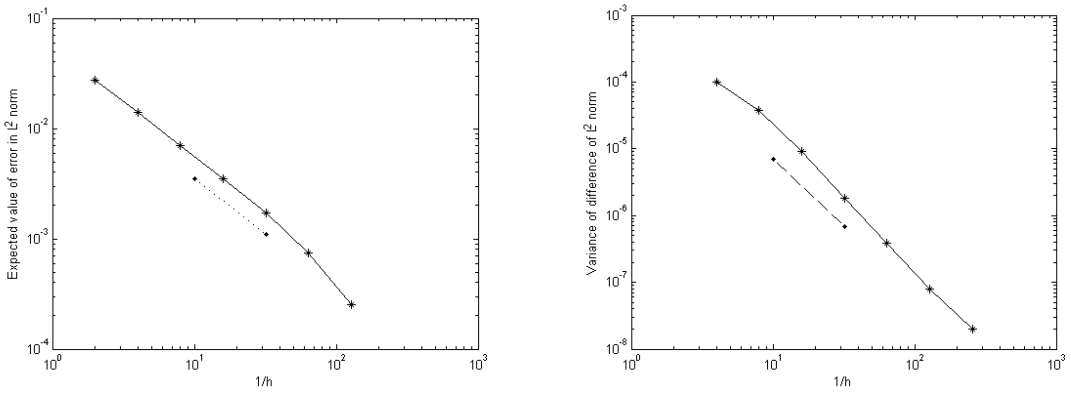

Figure 4-1: Left: Plot of $|\mathbb{E}\left[\|u_{h^*}\|_{L^2(D)} - \|u_h\|_{L^2(D)}\right]|$, for model problem (3.18) with $d = 2$, $\lambda = 0.3$, $\sigma^2 = 1$ and $h^* = 1/256$. Right: Corresponding plot of the variance $\mathbb{V}\left[\|u_h\|_{L^2(D)} - \|u_{2h}\|_{L^2(D)}\right]$. The gradient of the dotted (resp. dashed) line is -1 (resp. -2).

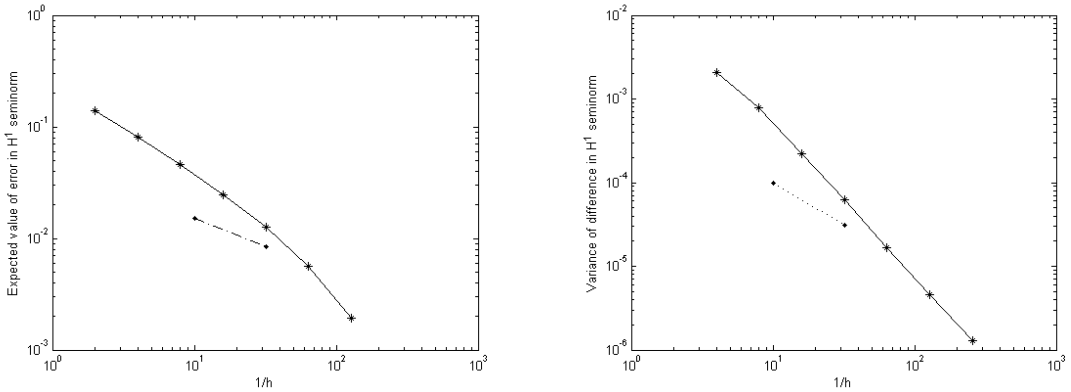

Figure 4-2: Left plot: $|\mathbb{E}\left[|u_{h^*}|_{H^1(D)} - |u_h|_{H^1(D)}\right]|$, for model problem (3.18) with $d = 2$ and 2–norm exponential covariance with $\lambda = 0.3$, $\sigma^2 = 1$ and $h^* = 1/256$. Right plot : $\mathbb{V}\left[|u_h|_{H^1(D)} - |u_{2h}|_{H^1(D)}\right]$. The gradient of the dash-dotted (resp. dotted) line is $-1/2$ (resp. -1).

computed via the functional $M_\omega^{(4)}$ for the mixed problem (3.19). It follows from the plots in Figure 3-3 that $\alpha = 1$ is observed numerically in both cases. In Figure 4-3, we further see that both $\mathbb{V}\big[M^{(3)}(u_h) - M^{(3)}(u_{2h})\big]$ for the Dirichlet problem and $\mathbb{V}\big[M_\omega^{(4)}(u_h) - M_\omega^{(4)}(u_{2h})\big]$ for the mixed problem converge quadratically in h, as predicted by Proposition 4.5.

Finally, let us consider the point evaluation of the horizontal Darcy flux $a\frac{\partial u}{\partial x_1}$ at the point $x^* = (\frac{3}{4} + \frac{1}{512}, \frac{3}{4} + \frac{1}{512})$. This is not a grid point in any of the meshes used for computation. The gradient $\frac{\partial u_h}{\partial x_1}$ at x^* is easily computed using u_h, and the permeability a is approximated using the average of the permeability values at the nodes of the element containing x^*. Proposition 4.7 suggests that we should observe $\alpha = 1/2$ and $\beta = 1$. However, we see in Figure 4-4, that both $\big|\mathbb{E}\big[a\frac{\partial u_{h^*}}{\partial x_1}(x^*) - a\frac{\partial u_h}{\partial x_1}(x^*)\big]\big|$ and $\mathbb{V}\big[a\frac{\partial u_h}{\partial x_1}(x^*) - a\frac{\partial u_{2h}}{\partial x_1}(x^*)\big]$ converge linearly in h.

4.5.2 Computational cost

We now want to compare the cost of the standard (single-level) MC estimator and the multilevel MC estimator as described in section 4.2. We again consider model problem (3.18) and the simple functional $Q = \|u\|_{L^2(D)}$, and choose the 2-norm exponential covariance function for $\log[a]$. The grid hierarchy in the multilevel estimator is chosen as $h_0 = 1/8$, and $h_\ell = h_{\ell-1}/2$, for $\ell = 1, \ldots, L$. The performance of the MC and MLMC estimators in estimating $\|u\|_{L^2(D)}$ is shown in Figure 4-5. The accuracy ε is scaled by the expected value of the quantity of interest, $\mathbb{E}\left[\|u_{1/256}\|_{L^2(D)}\right] \approx 0.045$, for $\lambda = 0.3, \sigma^2 = 1$. We see a clear advantage of the multilevel Monte Carlo method. The cost on the vertical axis of the right plot is calculated as $N_0 + \sum_{\ell=1}^{L} N_\ell \frac{M_\ell + M_{\ell-1}}{M_0}$, where $M_\ell = h_\ell^{-2}$, and so represents a standardised cost assuming an optimal linear solver ($\gamma = d = 2$). Figure 4-5 confirms the ε-cost of order ε^{-2} predicted by Table 4.1 for $\alpha = 1, \beta = 2$.

We now want to analyse the gains of increasing the number of levels in the MLMC algorithm in more detail. The quantity of interest is again $Q = \|u\|_{L^2(D)}$, but for model problem (3.18) with (rougher) 2-norm exponential covariance with $\lambda = 0.1, \sigma^2 = 1$. We quantify the cost of the different estimators with CPU-times, calculated using a MATLAB implementation running on a 3GHz Intel Core 2 Duo E8400 processor with 3.2GByte of RAM. As linear solver we use the standard backslash operation, which we numerically found to have $\gamma \approx 2.4$ in 2D. The

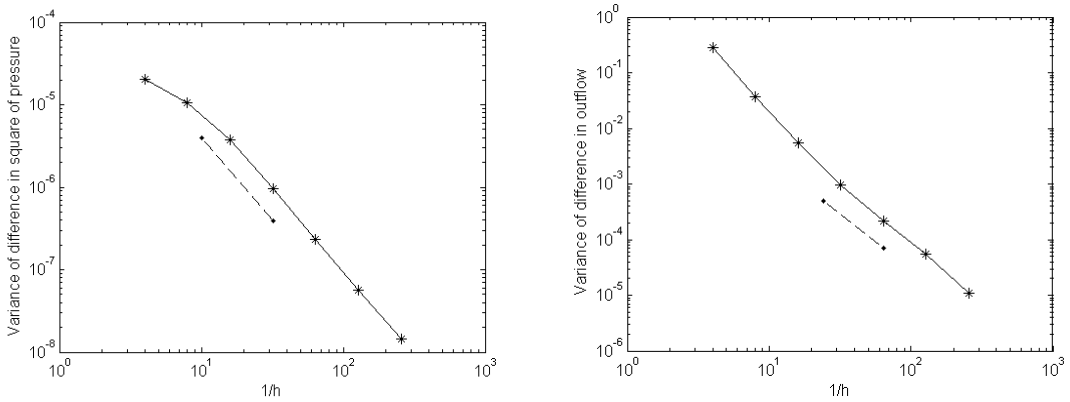

Figure 4-3: Left plot: $\mathbb{V}\big[M^{(3)}(u_h) - M^{(3)}(u_{2h})\big]$ versus $1/h$ for model problem (3.18) with $d = 2$ and 2-norm exponential covariance with $\lambda = 0.3$, $\sigma^2 = 1$ and $h^* = 1/256$. Right plot: $\mathbb{V}\big[M^{(4)}_\omega(u_h) - M^{(4)}_\omega(u_{2h})\big]$ versus $1/h$ for model problem (3.19) with $d = 2$ and 2-norm exponential covariance with $\lambda = 0.3$, $\sigma^2 = 1$, $\psi = x_1$ and $h^* = 1/256$. The gradient of the dashed line is -2.

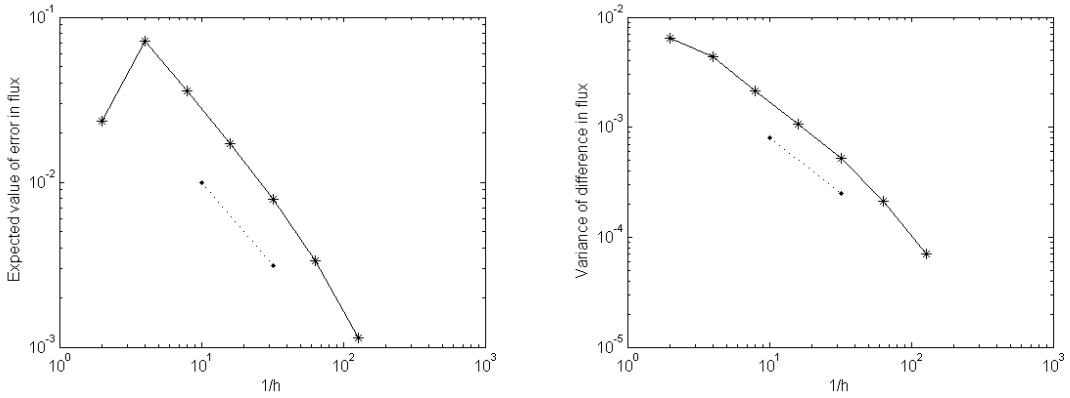

Figure 4-4: Left plot: $\big|\mathbb{E}\big[a\frac{\partial u_{h^*}}{\partial x_1}(x^*) - a\frac{\partial u_h}{\partial x_1}(x^*)\big]\big|$ for model problem (3.18) with $d = 2$ and 2-norm exponential covariance with $\lambda = 0.3$, $\sigma^2 = 1$ and $h^* = 1/256$. Right plot: $\mathbb{V}\big[a\frac{\partial u_h}{\partial x_1}(x^*) - a\frac{\partial u_{2h}}{\partial x_1}(x^*)\big]$. The gradient of the dotted line is -1.

results are shown in Figure 4-6.

In the left plot, we fix the standard deviation of the estimators (scaled by $\mathbb{E}\left[\|u_{1/256}\|_{L^2(D)}\right] \approx 0.041$ for $\lambda = 0.1, \sigma^2 = 1$) to $10^{-4}/0.041 = 2.4 * 10^{-3}$, and study how the cost of the different estimators grows with $M = h^{-2}$. To achieve the required standard deviation, the standard MC estimator is as costly on a grid of size $M = 32^2$ as the 5-level method on a grid of size $M = 256^2$. The standard MC method would be 100 times more costly on the $M = 256^2$ grid.

In the right plot in Figure 4-6, we fix the spatial discretisation to $M = 256^2$, and study how the cost of the estimators increases as we decrease the required scaled standard deviation of the estimator. The horizontal line represents the scaled discretisation error at $h = 1/256$. We see that the time needed to get a standard deviation of the same size as the spatial discretisation error is 20 minutes for the standard MC estimator, while it is only 20 seconds for the 4-level estimator.

4.5.3 Level dependent estimators

Finally, let us now present some numerical results with the level-dependent estimators described in section 4.4. To be able to deal with very short correlation lengths in a reasonable time, we start with the 1D equivalent of model problem (3.18), on $D = (0, 1)$. We take a to be a log–normal random field, with $\log[a]$ having exponential covariance function (2.25) with $\lambda = 0.01$ and $\sigma^2 = 1$. The results in Figure 4-7 are for point evaluation of the pressure at $x^* = 2049/4096$. Similar gains can be obtained for other quantities of interest.

In the left plot in Figure 4-7, we study the behaviour of $\mathbb{V}[Q_{h_\ell} - Q_{h_{\ell-1}}]$ and $\mathbb{V}[Q_{h_\ell}]$. When $\mathbb{V}[Q_{h_\ell} - Q_{h_{\ell-1}}] \geq \mathbb{V}[Q_{h_\ell}]$, there is no benefit including level $\ell - 1$ in the multilevel estimator, since it would only increase the cost of the estimator. We can see that if we approximate a with a (large) fixed number of modes on each level (labelled "keep" in Figure 4-7), we should not include any levels coarser than $h_0 = 1/64\,(\approx \lambda)$ in the estimator, as was already observed in [15]. With the level–dependent regime (labelled "drop"), however, it is viable to include levels as coarse as $h_0 = 1/2$. This leads to significant reductions in computational cost, as is shown in the right plot in Figure 4-7.

In Figure 4-7, we fix the required tolerance for the sampling error (i.e. the

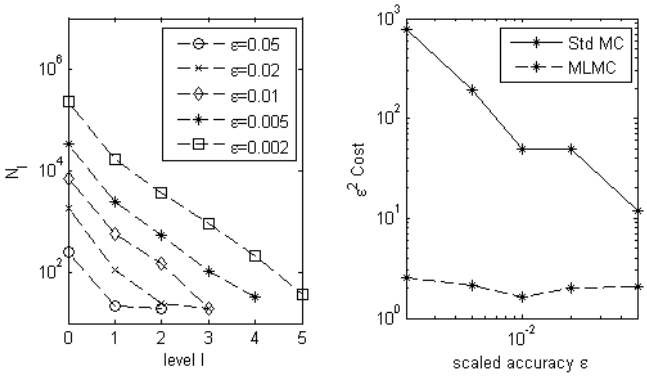

Figure 4-5: Left: Number of samples N_ℓ per level. Right: Plot of the cost scaled by ε^{-2} of the MLMC and standard MC estimators for $d = 2$, with $\lambda = 0.3$ and $\sigma^2 = 1$. The coarsest mesh size in all tests is $h_0 = 1/8$.

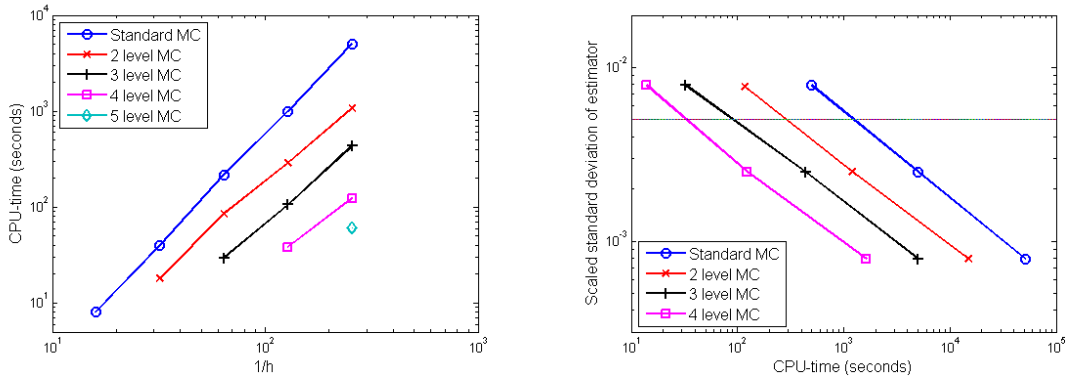

Figure 4-6: Left plot: CPU-time versus $1/h$ for a fixed maximum scaled standard deviation $2.4 * 10^{-3}$, for model problem (3.18) with $d = 2$ and 2-norm exponential covariance with $\lambda = 0.1$ and $\sigma^2 = 1$. Right plot: Standard deviation versus CPU-time, for a fixed finest mesh $h = h_L = 1/256$.

standard deviation of the estimator) at $\delta = 10^{-3}$, and look at how the cost of the different estimators grows as we decrease the mesh size $h := h_L$ of the finest grid, with each line in the plot using a fixed number of grid levels in the multilevel simulation (e.g. 4L means 4 levels). The computational cost of the multilevel estimator is calculated as $N_0 h_0^{-1} + \sum_{\ell=1}^{L} N_\ell (h_\ell^{-1} + h_{\ell-1}^{-1})$ work units, since we know that $\gamma = 1$ in M4 for $d = 1$. To make the estimators comparable, for each finest grid h_L, the standard Monte Carlo estimator is computed with $R_L = h_L^{-1}$ modes, the "MLMC keep" estimator is computed with $R_\ell = h_L^{-1}$ modes on all levels, and the "MLMC drop" estimator is computed with a varying number $R_\ell = h_\ell^{-1}$ modes on the levels. We clearly see the benefit of using the level–dependent multilevel estimator. For example, on the grid of size $h = 1/2048$, the cheapest multilevel estimator with a fixed number of modes is the 4 level estimator, which has a cost of 8.6×10^5 work units. The cheapest level–dependent multilevel estimator, on the other hand, is the 7 level estimator, whose computational cost is only 1.8×10^5 units. For comparison, the cost of the single–level MC estimator on this grid is 2.8×10^6 units.

An important point we would like to make here, is that not only do the level–dependent estimators have a smaller absolute cost than the estimators with a fixed number of modes, they are also a lot more robust with respect to the coarse grids included. On the $h = 1/2048$ grid, the 11 level estimator (i.e. $h_0 = 1/2$) with fixed R, costs 1.1×10^7 units, which is 4 times the cost of the standard MC estimator. The 11 level estimator with level–dependent R_ℓ costs 2.4×10^5 units, which is only marginally more than the best level–dependent estimator (the 7 level estimator).

For practical purposes, the real advantage of the level–dependent approach is evident on coarser grids. We see in Figure 4-7 that on grids coarser than $h = 1/256$, all multilevel estimators with a fixed number of modes are more expensive than the standard MC estimator. With the level–dependent multilevel estimators on the other hand, we can make use of (and benefit from) multilevel estimators on grids as coarse as $h = 1/64$. This is very important, especially in the limit as the correlation length $\lambda \to 0$, as eventually all computationally feasible grids will be "coarse" with respect to λ. With the level–dependent estimators, we can benefit from the multilevel approach even for very small values of λ.

Let us now move on to a model problem in $d = 2$. We will study the flow

cell model problem (3.19), and take the outflow functional $M_\omega^{(4)}(u)$ with weight function $\psi = x_1$ as our quantity of interest. We choose a to be a log–normal random field s.t. $\log[a]$ has 1-norm exponential covariance function (2.25), with $\lambda = 0.1$ and $\sigma^2 = 1$.

The left plot in Figure 4-8 is similar to the left plot in Figure 4-7. We again see that the level–dependent regime allows for much coarser grids. In the right plot, we see the gains in computational cost that are possible with the level–dependent estimators. Since we do not know the value of γ in (M4) theoretically, we quantify the cost of the estimators by the CPU–time. The results shown are again calculated with a MATLAB implementation on a 3GHz Intel Core 2 Duo E8400 processor with 3.2GByte of RAM, using the sparse direct solver provided in Matlab through the standard backslash operation to solve the linear systems for each sample. On the finest grid $h = 1/256$, we clearly see a benefit from the level–dependent estimators. The cheapest multilevel estimator with a fixed number of modes is the 5 level estimator, with takes 13.5 minutes. The cheapest level–dependent estimator is the 7 level estimator, which takes only 2.5 minutes. For comparison, the standard MC estimator takes more than 7.5 hours.

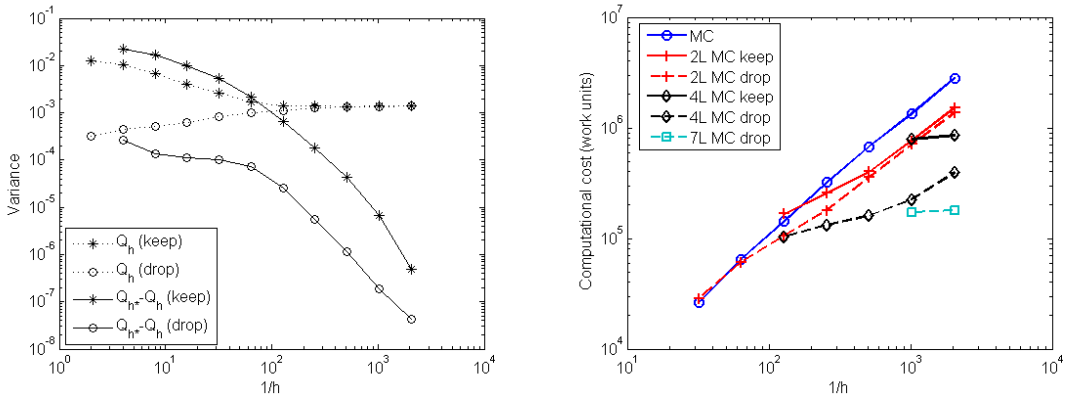

Figure 4-7: Left: Plot of $\mathbb{V}\big[M^{(1)}(u_h)\big]$ and $\mathbb{V}\big[M^{(1)}(u_h) - M^{(1)}(u_{2h})\big]$, for (3.18) with $d = 1$, $\lambda = 0.01$, $\sigma^2 = 1$, $R_\ell = h_\ell^{-1}$, $h^* = 1/4096$, $R^* = 4096$ and $x^* = 2049/4096$. Right: Plot of cost versus $1/h$ for a fixed tolerance of the sampling error of $\delta = 10^{-3}$, for the same model problem.

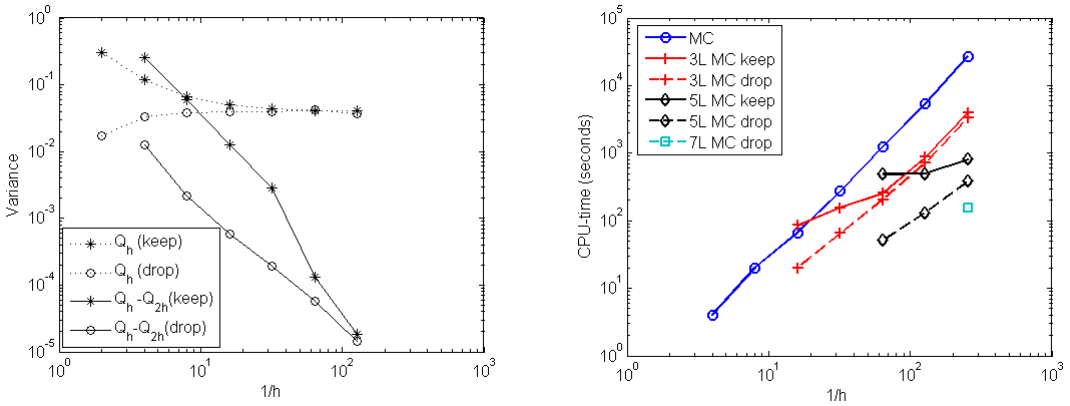

Figure 4-8: Left: Plot of $\mathbb{V}\big[M_\omega^{(4)}(u_h)\big]$ and $\mathbb{V}\big[M_\omega^{(4)}(u_h) - M_\omega^{(4)}(u_{2h})\big]$, for (3.19) with $d = 2$, $\lambda = 0.1$, $\sigma^2 = 1$, $\psi = x_1$, $R_\ell = 4h_\ell^{-1}$, $h^* = 1/256$ and $R^* = 1024$. Right: Plot of cost CPU-time versus $1/h$ for a fixed tolerance of the sampling error of $\delta = 10^{-3}$, for the same model problem.

Chapter 5

Multilevel Markov chain Monte Carlo methods

In this chapter, we consider the setting where we have some real-world dynamic data (or *observations*) F_{obs} available, and want to incorporate this information into our simulation in order to reduce the overall uncertainty. We will use the Bayesian framework, and assign a *prior* distribution to the model input. To get a better representation of the input, we condition the prior on the data F_{obs}, leading to the *posterior* distribution.

In most situations, the posterior distribution is intractable in the sense that exact sampling from it is unavailable. One way to circumvent this problem, is to generate samples using a Metropolis–Hastings type Markov chain Monte Carlo (MCMC) approach [46, 55, 61], which consists of two main steps: (i) given the previous sample, a new sample is generated according to some proposal distribution, such as a random walk; (ii) the likelihood of this new sample (i.e. the model fit to F_{obs}) is compared to the likelihood of the previous sample. Based on this comparison, the proposed sample is then either accepted and used for inference, or it is rejected and we use instead the previous sample again, leading to a Markov chain.

A major problem with MCMC is the high cost of the likelihood calculation for large–scale applications, which involves the (accurate) numerical solution of model problem (2.1). Due to the slow convergence of Monte Carlo averaging, the number of samples is also large and moreover, the likelihood has to be calculated not only for the samples that are eventually used for inference, but also for the

samples that end up being rejected. Altogether, this leads to an often impossibly high overall complexity, particularly in the context of high-dimensional parameter spaces (typically needed in subsurface flow applications), where the acceptance rate of the algorithm can be very low. We show here how the computational cost of the standard Metropolis-Hastings algorithm can be reduced significantly by using the multilevel approach already introduced in chapter 4.

Before we go on to describe the standard MCMC and new multilevel MCMC methods, let us describe the mathematical setting for the remainder of this chapter. We will assume that a is a (scalar) log-normal random field as described in section 2.7, such that assumptions B1-B3 are satisfied. Recall the Karhunen-Loève expansion of the Gaussian random field $g = \log[a]$ from (3.8):

$$g(x,\omega) \;=\; \sum_{n=1}^{\infty} \sqrt{\mu_n}\, \xi_n(\omega)\, b_n(x)\,,$$

where we for simplicity have assumed that $g(\omega, x)$ has mean zero. Denote $\vartheta := \{\xi_n\}_{n\in\mathbb{N}} \in \mathbb{R}^{\mathbb{N}}$. We will work with prior and posterior measures on the space $\mathbb{R}^{\mathbb{N}}$. To this end, we equip $\mathbb{R}^{\mathbb{N}}$ with the product sigma algebra $\mathcal{B} := \bigotimes_{n\in\mathbb{N}} \mathcal{B}^1(\mathbb{R})$, where $\mathcal{B}^1(\mathbb{R})$ denotes the sigma algebra of Borel sets of $\mathbb{R}$. We denote by ρ_0 the prior measure on $\mathbb{R}^{\mathbb{N}}$, defined by $\{\xi_n\}_{n\in\mathbb{N}}$ being independent and identically distributed (i.i.d) $N(0,1)$ random variables,

$$\rho_0 = \bigotimes_{n\in\mathbb{N}} g(\xi_n)\, \mathrm{d}\xi_n, \tag{5.1}$$

where $g : \mathbb{R} \to \mathbb{R}^+$ is the Lebesgue density of a $N(0,1)$ random variable and $d\xi_n$ denotes the one dimensional Lebesgue measure.

We assume that the observed data F_{obs} is finite dimensional, with $F_{\text{obs}} \in \mathbb{R}^m$, for some $m \in \mathbb{N}$. We further assume that

$$F_{\text{obs}} = \mathcal{F}(u(\vartheta)) + \eta, \tag{5.2}$$

where $\mathcal{F} : H^1(D) \to \mathbb{R}^m$ is continuous, u is the (weak) solution model problem (2.1) which depends on ϑ through a, and the observational noise η is assumed to be a realisation of a $N(0, \sigma_F^2 I_m)$ random variable (independent of ϑ). The parameter σ_F^2 is a fidelity parameter that indicates the level of observational

noise present in F_{obs}.

With ρ_0 as in (5.1), we have $\rho_0(\mathbb{R}^{\mathbb{N}}) = 1$. Furthermore, it follows from the assumption that $\mathcal{F}$ is continuous, together with assumptions B1-B3 and the continuous dependence of u on ϑ (see [10, Propositions 3.6 and 4.1] or Lemmas 2.20 and 5.13), that the map $\mathcal{F}(u(\cdot)) : \mathbb{R}^{\mathbb{N}} \to \mathbb{R}^m$ is continuous. The posterior distribution, which we will denote by ρ, is then known to be absolutely continuous with respect to the prior and satisfies

$$\frac{\partial \rho}{\partial \rho_0}(\vartheta) \eqsim \exp\left[-\frac{\|F_{\text{obs}} - \mathcal{F}(u(\vartheta))\|^2}{2\sigma_F^2}\right] =: \exp\left[-\Phi(\vartheta; F_{\text{obs}})\right], \tag{5.3}$$

where $\|\cdot\|$ denotes the Euclidean norm on $\mathbb{R}^m$, and the hidden constant depends only on F_{obs} and is generally not known (see [64] and the references therein for more detail). The right hand side of (5.3) is often referred to as the *likelihood*.

In practical computations, since the exact solution $u(\vartheta)$ is not available, the likelihood $\exp[-\Phi(\vartheta; F_{\text{obs}})]$ needs to be approximated. We will use the approximation $\mathcal{F}(u(\vartheta)) \approx \mathcal{F}(u_{R,h}(\vartheta))$, where $u_{R,h}$ is the finite element approximation of the truncated model problem defined in (3.9). We may also change the value of σ_F^2 to $\sigma_{F,h}^2$ (see section 5.2.3 for a reason why we might want to do this). We denote the resulting approximate posterior measure correspondingly by $\rho^{h,R}$, with

$$\frac{\partial \rho^{h,R}}{\partial \rho_0}(\vartheta) \eqsim \exp\left[-\frac{\|F_{\text{obs}} - \mathcal{F}(u_{R,h}(\vartheta))\|^2}{2\sigma_{F,h}^2}\right] =: \exp\left[-\Phi^{h,R}(\vartheta; F_{\text{obs}})\right]. \tag{5.4}$$

Since $\mathcal{F}(u_{R,h}(\vartheta))$ only depends on $\theta := \{\xi_n\}_{n=1}^R$, the first R components of ϑ, and since the prior measure can be factorised as $\rho_0 = \rho_0^R \otimes \rho_0^\perp$, the approximate posterior measure can be factorised as $\rho^{h,R} = \nu^{h,R} \otimes \rho^\perp$, where

$$\frac{\partial \nu^{h,R}}{\partial \rho_0^R}(\theta) \eqsim \exp\left[-\Phi^{h,R}(\theta; F_{\text{obs}})\right], \tag{5.5}$$

and $\rho^\perp = \rho_0^\perp$ [20]. Note that $\nu^{h,R}$ is a measure on the finite dimensional space $\mathbb{R}^R$. Denoting by $\pi^{h,R}$ and π_0^R the densities with respect to the R dimensional Lebesgue measure of $\nu^{h,R}$ and ρ_0^R, respectively, it follows from (5.5) that

$$\pi^{h,R}(\theta) \eqsim \exp\left[-\Phi^{h,R}(\theta; F_{\text{obs}})\right] \pi_0^R(\theta). \tag{5.6}$$

ALGORITHM 1. (Metropolis Hastings MCMC)
Choose θ^0. For $n \geq 0$:

- Given θ^n, generate a proposal θ' from a given proposal density $q(\theta'|\theta^n)$.
- Accept θ' as a sample with probability

$$\alpha^{h,R}(\theta'|\theta^n) = \min\left\{1, \frac{\pi^{h,R}(\theta')\, q(\theta^n|\theta')}{\pi^{h,R}(\theta^n)\; q(\theta'|\theta^n)}\right\} \tag{5.7}$$

  i.e. $\theta^{n+1} = \theta'$ with probability $\alpha^{h,R}$ and $\theta^{n+1} = \theta^n$ with probability $1 - \alpha^{h,R}$.

We are interested in approximating the expected value (with respect to the posterior ρ) of a quantity $Q = \mathcal{G}(u(\vartheta))$, for some continuous $\mathcal{G} : H^1(D) \to \mathbb{R}$. We denote this expected value by $\mathbb{E}_\rho[Q] := \int_{\mathbb{R}^\mathbb{N}} \mathcal{G}(u(\vartheta))\, \rho(\mathrm{d}\vartheta)$. We assume that, as $h \to 0$ and $R \to \infty$,

$$\mathbb{E}_{\nu^{h,R}}[Q_{h,R}] \to \mathbb{E}_\rho[Q],$$

where $\mathbb{E}_{\nu^{h,R}}[Q_{h,R}] := \int_{\mathbb{R}^R} \mathcal{G}(u_{R,h}(\theta))\, \nu^{h,R}(\mathrm{d}\theta)$ is a finite dimensional integral. To estimate $\mathbb{E}_\rho[Q]$, we hence construct estimators of $\mathbb{E}_{\nu^{h,R}}[Q_{h,R}]$, for h sufficiently small and R sufficiently large.

5.1 Standard Markov chain Monte Carlo

We will start in this section with a review of the standard Metropolis Hastings algorithm. As already mentioned, the posterior measure $\nu^{h,R}$ is usually intractable. In order to generate samples for inference on $\mathbb{E}_{\nu^{h,R}}[Q_{h,R}]$, we will use the Metropolis Hastings MCMC algorithm in Algorithm 1.

Algorithm 1 creates a Markov chain $\{\theta^n\}_{n\in\mathbb{N}}$, and the states θ^n are used in the usual way as samples for inference in a Monte Carlo sampler. The proposal density $q(\theta'|\theta^n)$ is what defines the algorithm. A common choice is a simple random walk. However, as outlined in [45], the basic random walk does not lead to dimension R independent convergence, and a better choice is a preconditioned Crank-Nicolson (pCN) algorithm [19]. Below we will see that the proposal density

is also the crucial ingredient in our multilevel Metropolis-Hastings algorithm. When the proposal density is symmetric, i.e. when $q(\theta^n|\theta') = q(\theta'|\theta^n)$, then the formula for $\alpha^{h,R}(\theta'|\theta^n)$ in (5.7) simplifies. Note that the acceptance probability $\alpha^{h,R}$ is computable, since the unknown normalising constant in (5.6) appears in both the numerator and denominator and hence cancels.

Under reasonable assumptions, one can show that sample averages computed with these samples converge to expected values with respect to the desired target distribution $\nu^{h,R}$ (see Theorem 5.2). The first several samples of the chain $\{\theta^n\}_{n\in\mathbb{N}}$, say $\theta^0, \dots, \theta^{n_0}$, are not usually used for inference, since the chain needs some time to get close to the target distribution $\nu^{h,R}$. This is referred to as the *burn–in* of the MCMC algorithm. Although the length of the burn-in is crucial for practical purposes, and largely influences the behaviour of the resulting MCMC estimator for finite sample sizes, statements about the asymptotics of the estimator are usually independent of the burn-in. We will therefore denote our MCMC estimator by

$$\widehat{Q}_N^{\mathrm{MC}} := \frac{1}{N}\sum_{n=n_0+1}^{N+n_0} Q_{h,R}^{(n)} = \frac{1}{N}\sum_{n=n_0+1}^{N+n_0} \mathcal{G}\left(u_{R,h}(\theta^n)\right), \tag{5.8}$$

for any $n_0 \geq 0$, and only explicitly state the dependence on n_0 where needed.

5.1.1 Abstract convergence analysis

We will now give a brief overview of the convergence properties of the Metropolis-Hastings algorithm, which we will need below in the analysis of the multilevel variant. For more details we refer the reader, e.g., to [61]. Let

$$K(\theta'|\theta^n) := \alpha^{h,R}(\theta'|\theta^n)\, q(\theta'|\theta^n) + \left(1 - \int_{\mathbb{R}^R} \alpha^{h,R}(\theta''|\theta^n)\, q(\theta''|\theta^n)\, \mathrm{d}\theta''\right) \delta(\theta^n - \theta')$$

denote the transition kernel of the Markov chain $\{\theta^n\}_{n\in\mathbb{N}}$, with $\delta(\cdot)$ the Dirac delta function, and denote

$$\begin{aligned}\mathcal{E} &= \{\theta : \pi^{h,R}(\theta) > 0\},\\ \mathcal{D} &= \{\theta' : q(\theta'|\theta) > 0 \text{ for some } \theta \in \mathcal{E}\}.\end{aligned}$$

The set $\mathcal{E}$ contains all values of θ which have a positive posterior probability, and is the set that Algorithm 1 should sample from. The set $\mathcal{D}$, on the other hand, consists of all samples which can be generated by the proposal density q, and hence contains the set that Algorithm 1 will actually sample from. For the algorithm to fully explore the target distribution, we therefore crucially require $\mathcal{E} \subset \mathcal{D}$. The following results are classical, and can be found in [61].

Lemma 5.1. *Provided $\mathcal{E} \subset \mathcal{D}$, $\nu^{h,R}$ is a stationary distribution of the chain $\{\theta^n\}_{n\in\mathbb{N}}$.*

Note that the condition $\mathcal{E} \subset \mathcal{D}$ is sufficient for the transition kernel $K(\cdot|\cdot)$ to satisfy the usual detailed balance condition $K(\theta'|\theta^n)\,\pi^{h,R}(\theta^n) = K(\theta^n|\theta')\,\pi^{h,R}(\theta')$.

Theorem 5.2. *Suppose that $\mathbb{E}_{\nu^{h,R}}\left[|Q_{h,R}|\right] < \infty$ and*

$$q(\theta'|\theta) > 0, \ \textit{for all } (\theta,\theta') \in \mathcal{E}\times\mathcal{E}. \tag{5.9}$$

Then

$$\lim_{N\to\infty} \widehat{Q}_N^{\mathrm{MC}} = \mathbb{E}_{\nu^{h,R}}\left[Q_{h,R}\right], \qquad \textit{for any } \theta^0 \in \mathcal{E} \textit{ and } n_0 \geq 0.$$

The condition (5.9) is sufficient for the chain $\{\theta^n\}_{n\in\mathbb{N}}$ to be *irreducible*, and it is satisfied for example for the random walk sampler or for the pCN algorithm (cf. [45]). Lemma 5.1 and Theorem 5.2 above ensure that asymptotically, sample averages computed with samples generated by Algorithm 1 converge to the desired expected value. In particular, we note that stationarity of $\{\theta^n\}_{n\in\mathbb{N}}$ is not required for Theorem 5.2, and the above convergence results hence hold true for any burn–in $n_0 \geq 0$, and for all initial values $\theta^0 \in \mathcal{E}$.

Now that we have established the (asymptotic) convergence of the MCMC estimator (5.8), let us establish a bound on the cost of this estimator. We will quantify the accuracy of our estimator via the root mean square error (RMSE)

$$e(\widehat{Q}_N^{\mathrm{MC}}) := \left(\mathbb{E}_{\boldsymbol{\Theta}}\left[\left(\hat{Q}_N^{\mathrm{MC}} - \mathbb{E}_\rho(Q)\right)^2\right]\right)^{1/2}, \tag{5.10}$$

where $\mathbb{E}_{\boldsymbol{\Theta}}$ denotes the expected value not with respect to the target measure $\nu^{h,R}$, but with respect to the joint distribution of $\boldsymbol{\Theta} := \{\theta^n\}_{n\in\mathbb{N}}$ as generated by Algorithm 1. We denote by $\mathcal{C}_\varepsilon(\widehat{Q}_N^{\mathrm{MC}})$ the computational ε-cost of the estimator,

that is the number of floating point operations that are needed to achieve a RMSE of $e(\widehat{Q}_N^{\mathrm{MC}}) < \varepsilon$.

Classically, the mean square error (MSE) can be written as the sum of the variance of the estimator and its bias squared,

$$e(\widehat{Q}_N^{\mathrm{MC}})^2 = \mathbb{V}_{\boldsymbol{\Theta}}\left[\widehat{Q}_N^{\mathrm{MC}}\right] + \left(\mathbb{E}_{\boldsymbol{\Theta}}\left[\widehat{Q}_N^{\mathrm{MC}}\right] - \mathbb{E}_{\rho}\left[Q\right]\right)^2.$$

Here, $\mathbb{V}_{\boldsymbol{\Theta}}$ is again the variance with respect to the approximating measure generated by Algorithm 1. Using the triangle inequality and linearity of expectation, we can further write this as

$$\begin{aligned} &e(\widehat{Q}_N^{\mathrm{MC}})^2 \\ &\leq \mathbb{V}_{\boldsymbol{\Theta}}\left[\widehat{Q}_N^{\mathrm{MC}}\right] + 2\left(\mathbb{E}_{\boldsymbol{\Theta}}\left[\widehat{Q}_N^{\mathrm{MC}}\right] - \mathbb{E}_{\nu^{h,R}}\left[\widehat{Q}_N^{\mathrm{MC}}\right]\right)^2 + 2\left(\mathbb{E}_{\nu^{h,R}}\left[Q_{h,R}\right] - \mathbb{E}_{\rho}\left[Q\right]\right)^2. \end{aligned} \tag{5.11}$$

The three terms in (5.11) correspond to the three sources of error in the MCMC estimator. The third (and last) term in (5.11) is the discretisation error due to approximating Q by $Q_{h,R}$ and ρ by $\nu^{h,R}$. The other two terms are the errors introduced by using an MCMC estimator for the expected value; the first term is the error due to using a finite sample average and the second term is due to the samples in the estimator not all being perfect (i.i.d.) samples from the target distribution $\nu^{h,R}$.

Let us first consider the two MCMC related error terms. Quantifying, or even bounding, the variance and bias of an MCMC estimator in terms of the number of samples N is not an easy task, and is in fact still a very active area of research. The main issue with bounding the variance is that the samples used in the MCMC estimator are not independent, which means that knowledge of the covariance structure is required in order to bound the variance of the estimator. Asymptotically, the behaviour of the MCMC related errors (i.e. Terms 1 and 2 on the right hand side of (5.11)) can be described using the following Central Limit Theorem, which can be found in [61, Theorem 4.7.7].

Let $\tilde{\theta}^0 \sim \nu^{h,R}$. Then the auxiliary chain $\widetilde{\boldsymbol{\Theta}} := \{\tilde{\theta}^n\}_{n\in\mathbb{N}}$ constructed by Algorithm 1 starting from $\tilde{\theta}^0$ is stationary, i.e. $\tilde{\theta}^n \sim \nu^{h,R}$ for all $n \geq 0$. Note that the covariance structure of $\widetilde{\boldsymbol{\Theta}}$ is still implicitly defined by Algorithm 1 as for

$\boldsymbol{\Theta}$. However, now $\mathbb{V}_{\widetilde{\boldsymbol{\Theta}}}[\tilde{Q}^n_{h,R}] = \mathbb{V}_{\nu^{h,R}}[\tilde{Q}_{h,R}]$ and $\mathbb{E}_{\widetilde{\boldsymbol{\Theta}}}[\tilde{Q}^n_{h,R}] = \mathbb{E}_{\nu^{h,R}}[\tilde{Q}_{h,R}]$, for any $n \geq 0$, and

$$\mathrm{Cov}_{\widetilde{\boldsymbol{\Theta}}}\left[\tilde{Q}^0_{h,R},\, \tilde{Q}^n_{h,R}\right] = \mathbb{E}_{\widetilde{\boldsymbol{\Theta}}}\left[\left(\tilde{Q}^0_{h,R} - \mathbb{E}_{\nu^{h,R}}[Q_{h,R}]\right)\left(\tilde{Q}^n_{h,R} - \mathbb{E}_{\nu^{h,R}}[Q_{h,R}]\right)\right],$$

where $\tilde{Q}^n_{h,R} := \mathcal{G}(u_{R,h}(\tilde{\theta}^n))$. We now define the so called *asymptotic variance* of the MCMC estimator

$$\sigma^2_Q := \mathbb{V}_{\nu^{h,R}}\left[\tilde{Q}_{h,R}\right] + 2\sum_{n=1}^{\infty} \mathrm{Cov}_{\widetilde{\boldsymbol{\Theta}}}\left[\tilde{Q}^0_{h,R},\, \tilde{Q}^n_{h,R}\right].$$

Note that stationarity of the chain is assumed only in the definition of σ^2_Q, i.e. for $\widetilde{\boldsymbol{\Theta}}$, and it is not necessary for the samples $\boldsymbol{\Theta}$ actually used in the computation of $\widehat{Q}^{\mathrm{MC}}_N$.

Theorem 5.3 (Central Limit Theorem)**.** *Suppose* $\sigma^2_Q < \infty$, (5.9) *holds, and*

$$\mathbb{P}\left[\alpha^{h,R} = 1\right] < 1. \tag{5.12}$$

Then we have

$$\frac{1}{\sqrt{N}}\left(\widehat{Q}^{\mathrm{MC}}_N - \mathbb{E}_{\nu^{h,R}}\left[Q_{h,R}\right]\right) \xrightarrow{D} \mathcal{N}(0, \sigma^2_Q),$$

where $\xrightarrow{D}$ *denotes convergence in distribution.*

The condition (5.12) is sufficient for the chain $\boldsymbol{\Theta}$ to be *aperiodic*. It is difficult to prove theoretically. In practice, however, this condition is always satisfied, since not all proposals in Algorithm 1 will agree with the observed data and thus be accepted.

Theorem 5.3 holds again for any burn-in $n_0 \geq 0$ and any starting value $\theta^0 \in \mathcal{E}$. It shows that asymptotically, the sampling error of the MCMC estimator decays at the same rate as the sampling error of an estimator based on i.i.d. samples. Note that this includes both sampling errors, and so the constant σ^2_Q is in general larger than in the i.i.d. case where it is simply $\mathbb{V}_{\nu^{h,R}}\left[Q_{h,R}\right]$.

Since we are interested in a bound on the MSE of our MCMC estimator for a fixed number of samples N, we make the following assumption:

C1. For any $N \in \mathbb{N}$,

$$\mathbb{V}_{\boldsymbol{\Theta}}\left[\widehat{Q}_N^{\mathrm{MC}}\right] + \left(\mathbb{E}_{\boldsymbol{\Theta}}\left[\widehat{Q}_N^{\mathrm{MC}}\right] - \mathbb{E}_{\nu^{h,R}}\left[\widehat{Q}_N^{\mathrm{MC}}\right]\right)^2 \lesssim \frac{\mathbb{V}_{\nu^{h,R}}[Q_{h,R}]}{N}, \tag{5.13}$$

with a constant that is independent of h, R and N.

Non-asymptotic bounds such as in assumption C1 are difficult to obtain, but have recently been proved for certain Metropolis–Hastings algorithms, see e.g. [45, 62, 48]. These results require that the chain is sufficiently burnt–in. The hidden constant usually depends on quantities such as the covariances appearing in the asymptotic variance σ_Q^2.

To complete the error analysis, let us now consider the last term in the MSE (5.11), the discretisation bias. As before, we assume $\mathbb{E}_{\nu^{h,R}}[Q_{h,R}] - \mathbb{E}_{\rho}[Q] \to 0$ as $h \to 0$ and $R \to \infty$, and we furthermore assume that we have a certain order of convergence, i.e.

$$|\mathbb{E}_{\nu^{h,R}}[Q_{h,R}] - \mathbb{E}_{\rho}[Q]| \lesssim h^{\alpha} + R^{-\alpha'}, \tag{5.14}$$

for some $\alpha, \alpha' > 0$. The rates α and α' will be problem dependent. Let now $R = h^{-\alpha/\alpha'}$, such that the two error contributions in (5.14) are balanced. Then it follows from (5.11), (5.13) and (5.14) that the MSE of the MCMC estimator can be bounded by

$$e(\widehat{Q}_N^{\mathrm{MC}})^2 \lesssim \frac{\mathbb{V}_{\nu^{h,R}}[Q_{h,R}]}{N} + h^{\alpha}. \tag{5.15}$$

Under the assumption that $\mathbb{V}_{\nu^{h,R}}[Q_{h,R}] \approx$ constant, independent of h and R, it is hence sufficient to choose $N \gtrsim \varepsilon^{-2}$ and $h \lesssim \varepsilon^{1/\alpha}$ to get a RMSE of $\mathcal{O}(\varepsilon)$.

Let us now give a bound on the computational cost to achieve this error, the so called *ε-cost*. For this, assume that the cost to compute one sample $Q_{h,R}^n$ satisfies $\mathcal{C}(Q_{h,R}^n) \lesssim h^{-\gamma}$, for some $\gamma > 0$. Thus, with $N \gtrsim \varepsilon^{-2}$ and $h \lesssim \varepsilon^{1/\alpha}$, the ε–cost of our MCMC estimator can be bounded by

$$\mathcal{C}_{\varepsilon}(\widehat{Q}_N^{\mathrm{MC}}) \lesssim N h^{-\gamma} \lesssim \varepsilon^{-2-\gamma/\alpha}. \tag{5.16}$$

In practical applications, especially in subsurface flow, the discretisation parameter h needs to very small and the dimension R needs to be very large in order for $\mathbb{E}_{\nu^{h,R}}[Q_{h,R}]$ to be a good approximation to $\mathbb{E}_{\rho}[Q]$. Moreover, from the analysis above, we see that we need to use a large number of samples N in order to get

an accurate MCMC estimator with a small MSE. Since each sample requires the evaluation of the likelihood $\exp\left[-\Phi^{h,R}(\theta; F_{\text{obs}})\right]$, and this is very expensive when h is very small and R is very large, the standard MCMC estimator (5.8) is often extraordinarily expensive in practical situations. Additionally, the acceptance rate of the algorithm can be very low when R is very large. This means that the covariance between the different samples will decay more slowly, which again makes the hidden constant in assumption C1 larger, and the number of samples we have to take in order to get a certain accuracy increases even further.

To overcome the prohibitively large computational cost of the standard MCMC estimator (5.8), we will now introduce a new multilevel version of the estimator.

5.2 Multilevel Markov chain Monte Carlo

As in section 4.2, let now $\{h_\ell : \ell = 0, \dots, L\}$ be a sequence of mesh widths satisfying the geometric growth condition (4.4), for some $s \in \mathbb{N} \setminus \{1\}$:

$$h_\ell = s^{-1} h_{\ell-1}, \qquad \text{for all } \ell = 1, \dots, L.$$

In addition, we choose a (not necessarily strictly) increasing sequence $\{R_\ell\}_{\ell=0}^L \subset \mathbb{N}$, i.e. $R_\ell \geq R_{\ell-1}$, for all $\ell = 1, \dots, L$. For each level ℓ, we denote by $\theta_\ell := \{\xi_n\}_{n=1}^{R_\ell}$ the first R_ℓ entries of ϑ. We denote correspondingly the coefficient by $a_\ell := a^{R_\ell}$, the solution by $u_\ell := u_{R_\ell,h_\ell}$, the quantity of interest by $Q_\ell := Q_{h_\ell,R_\ell}$ and the resulting posterior distribution on $\mathbb{R}^{R_\ell}$ by $\nu^\ell := \nu^{h_\ell,R_\ell}$, with density π^ℓ.

Since in the context of MCMC simulations, the target distribution ν^ℓ depends on ℓ, the new multilevel MCMC (MLMCMC) estimator has to be defined carefully. We will use the identity

$$\mathbb{E}_{\nu^L}[Q_L] \;=\; \mathbb{E}_{\nu^0}[Q_0] \;+\; \sum_{\ell=1}^{L} \left(\mathbb{E}_{\nu^\ell}[Q_\ell] - \mathbb{E}_{\nu^{\ell-1}}[Q_{\ell-1}]\right) \tag{5.17}$$

as a basis. Note that in the case where all the distributions are the same, the above reduces to the telescoping sum (4.5) used for Monte Carlo estimators based on i.i.d samples.

As before, the idea of the multilevel estimator is now to estimate each of the terms on the right hand side of (5.17) independently, in a way that minimises

the variance of the estimator for a fixed computational cost. In particular, we will estimate each term in (5.17) by an MCMC estimator. The first term $\mathbb{E}_{\nu^0}[Q_0]$ can be estimated using the standard MCMC estimator described in Algorithm 1, i.e. $\widehat{Q}^{\mathrm{MC}}_{0,N_0}$ as in (5.8) with N_0 samples. We need to be more careful in estimating the differences $\mathbb{E}_{\nu^\ell}[Q_\ell] - \mathbb{E}_{\nu^{\ell-1}}[Q_{\ell-1}]$, and build an effective two-level version of Algorithm 1. For $\ell \geq 1$, we again denote $Y_\ell := Q_\ell(\theta_\ell) - Q_{\ell-1}(\Theta_{\ell-1})$ and define the estimator on level ℓ as

$$\widehat{Y}^{\mathrm{MC}}_{\ell,N_\ell} := \frac{1}{N_\ell} \sum_{n=n_0^\ell+1}^{n_0^\ell+N_\ell} Y_\ell^{(n)} = \frac{1}{N_\ell} \sum_{n=n_0^\ell+1}^{n_0^\ell+N_\ell} Q_\ell(\theta_\ell^n) - Q_{\ell-1}(\Theta_{\ell-1}^n),$$

where n_0^ℓ again denotes the burn-in of the estimator, N_ℓ is the number of samples on level ℓ and $\Theta_{\ell-1}$ has the same dimension as $\theta_{\ell-1}$. The main ingredient in this two level estimator is a judicious choice of the two Markov chains θ_ℓ^n and $\Theta_{\ell-1}^n$ (to be described later). The full MLMCMC estimator is now defined as

$$\widehat{Q}^{\mathrm{ML}}_{L,\{N_\ell\}} \;:=\; \widehat{Q}^{\mathrm{MC}}_{0,N_0} \;+\; \sum_{\ell=1}^{L} \widehat{Y}^{\mathrm{MC}}_{\ell,N_\ell}, \tag{5.18}$$

where it is important (i) that the $L+1$ estimators in (5.18) are independent, and (ii) that the two chains $\{\theta_\ell^n\}_{n\in\mathbb{N}}$ and $\{\Theta_\ell^n\}_{n\in\mathbb{N}}$, that are used in $\widehat{Y}^{\mathrm{MC}}_{\ell,N_\ell}$ and in $\widehat{Y}^{\mathrm{MC}}_{\ell+1,N_{\ell+1}}$ respectively, are drawn from the same posterior distribution ν^ℓ, so that $\widehat{Q}^{\mathrm{ML}}_{L,\{N_\ell\}}$ is a consistent estimator of $\mathbb{E}_{\nu^L}[Q_L]$.

There are two main ideas underlying the reduction in computational cost associated with the multilevel estimator. Firstly, samples of Q_ℓ, for $\ell < L$, are cheaper to compute than samples of Q_L, reducing the cost of the estimators on the coarser levels for any fixed number of samples. Secondly, if the variance of $Y_\ell = Q_\ell(\theta_\ell) - Q_{\ell-1}(\Theta_{\ell-1})$ tends to 0 as $\ell \to \infty$, we need only a small number of samples to obtain a sufficiently accurate estimate of the expected value of Y_ℓ on the fine grids, and so the computational effort on the fine grids is also greatly reduced.

By using the telescoping sum (5.17) and by sampling from the posterior distribution ν^ℓ on level ℓ, we ensure that a sample of Q_ℓ, for $\ell < L$, is indeed cheaper to compute than a sample of Q_L. It remains to ensure that the variance of $Y_\ell = Q_\ell(\theta_\ell) - Q_{\ell-1}(\Theta_{\ell-1})$ tends to 0 as $\ell \to \infty$. This will be ensured by the

ALGORITHM 2. (Metropolis Hastings MCMC for $Q_\ell - Q_{\ell-1}$)

Choose initial states $\Theta^0_{\ell-1}$ and $\theta^0_\ell := [\Theta^0_{\ell-1}\,,\,\theta^0_{\ell,F}]$. For $n \geq 0$:

- **On level $\ell-1$:** Given $\Theta^n_{\ell-1}$ generate $\Theta^{n+1}_{\ell-1}$ using Algorithm 1 with some proposal density $q^{\ell,C}(\Theta'_{\ell-1}\,|\,\Theta^n_{\ell-1})$ and acceptance probability

$$\alpha^{\ell,C}(\Theta'_{\ell-1}\,|\,\Theta^n_{\ell-1}) \;=\; \min\left\{1, \frac{\pi^{\ell-1}(\Theta'_{\ell-1})\,q^{\ell,C}(\Theta^n_{\ell-1}|\Theta'_{\ell-1})}{\pi^{\ell-1}(\Theta^n_{\ell-1})\,q^{\ell,C}(\Theta'_{\ell-1}|\Theta^n_{\ell-1})}\right\}.$$

- **On level ℓ:** Given θ^n_ℓ generate θ^{n+1}_ℓ using Algorithm 1 with the specific proposal density $q^\ell(\theta'_\ell\,|\,\theta^n_\ell)$ induced by taking $\theta'_{\ell,C} := \Theta^{n+1}_{\ell-1}$ and by generating a proposal for $\theta'_{\ell,F}$ from some proposal density $q^{\ell,F}(\theta'_{\ell,F}\,|\,\theta^n_{\ell,F})$. The acceptance probability is

$$\alpha^{\ell}(\theta'_\ell\,|\,\theta^n_\ell) \;=\; \min\left\{1, \frac{\pi^{\ell}(\theta'_\ell)\,q^{\ell}(\theta^n_\ell|\theta'_\ell)}{\pi^{\ell}(\theta^n_\ell)\,q^{\ell}(\theta'_\ell|\theta^n_\ell)}\right\}.$$

choice of θ^n_ℓ and $\Theta^n_{\ell-1}$.

5.2.1 The estimator for $Y_\ell = Q_\ell - Q_{\ell-1}$

Let us for the moment fix $1 \leq \ell \leq L$. The challenge is now to generate the chains $\{\theta^n_\ell\}_{n\in\mathbb{N}}$ and $\{\Theta^n_{\ell-1}\}_{n\in\mathbb{N}}$ such that the variance of Y_ℓ is small. To this end, we partition the chain θ_ℓ into two parts: the entries which are present already on level $\ell-1$ (the "coarse" modes), and the new entries on level ℓ (the "fine" modes):

$$\theta_\ell = [\theta_{\ell,C}\,,\,\theta_{\ell,F}],$$

where $\theta_{\ell,C}$ has length $R_{\ell-1}$, i.e. the same length as $\Theta_{\ell-1}$. $\theta_{\ell,F}$ has length $R_\ell - R_{\ell-1}$.

An easy way to construct θ^n_ℓ and $\Theta^n_{\ell-1}$ such that $\mathbb{V}[Y_\ell]$ is small, would be to generate θ^n_ℓ first, and then simply use $\Theta^n_{\ell-1} = \theta^n_{\ell,C}$. This is what was done in section 4.2. However, since we require $\Theta^n_{\ell-1}$ to come from a Markov chain with stationary distribution $\nu^{\ell-1}$, and θ^n_ℓ comes from the distribution ν^ℓ, this approach is not permissible. We will, however, use this general idea in Algorithm 2.

The coarse sample $\Theta^{n+1}_{\ell-1}$ is generated using the standard MCMC algorithm

given in Algorithm 1, using, e.g., a random walk or the pCN proposal density [19] for $q^{\ell,C}$. Based on the outcome on level $\ell-1$, we then generate θ_ℓ^{n+1}, using a new two-level proposal density in conjunction with the usual accept/reject step from Algorithm 1. The proposal density $q^{\ell,F}$ for the fine modes in that step can again be a simple random walk or the pCN algorithm.

At each step in Algorithm 2, there are four different outcomes, depending on whether we accept on both, one or none of the levels. The different possibilities are given in Table 5.1. Observe that when we accept on level ℓ, we always have $\theta_{\ell,C}^{n+1} = \Theta_{\ell-1}^{n+1}$, i.e. the coarse modes are the same. If, on the other hand, we reject on level ℓ, we crucially return to the previous state θ_ℓ^n on that level, which means that the coarse modes of the two states may differ. They will definitely differ if we accept on level $\ell-1$ and reject on level ℓ. If both proposals are rejected then it depends on the decision made at the previous state whether the coarse modes differ or not.

Level $\ell-1$ test	Level ℓ test	$\Theta_{\ell-1}^{n+1}$	$\theta_{\ell,C}^{n+1}$
reject	accept	$\Theta_{\ell-1}^{n}$	$\Theta_{\ell-1}^{n}$
accept	accept	$\Theta_{\ell-1}'$	$\Theta_{\ell-1}'$
reject	reject	$\Theta_{\ell-1}^{n}$	$\theta_{\ell,C}^{n}$
accept	reject	$\Theta_{\ell-1}'$	$\theta_{\ell,C}^{n}$

Table 5.1: Possible states of $\Theta_{\ell-1}^{n+1}$ and $\theta_{\ell,C}^{n+1}$ in Algorithm 2.

In general, this "divergence" of the coarse modes could mean that the variance of Y_ℓ is not small. For the setting considered in this chapter, however, we will prove in section 5.2.3 that the variance of Y_ℓ does in fact go to 0 as $\ell \to \infty$.

The specific proposal density q^ℓ in Algorithm 2 can be computed very easily and at no additional cost, leading to a simple formula for the "two-level" acceptance probability α^ℓ.

Lemma 5.4. *Let $\ell \geq 1$. Then*

$$\alpha^\ell(\theta_\ell' \,|\, \theta_\ell^n) \;=\; \min\left\{1, \frac{\pi^\ell(\theta_\ell')\,\pi^{\ell-1}(\theta_{\ell,C}^n)\,q^{\ell,F}(\theta_{\ell,F}^n|\theta_{\ell,F}')}{\pi^\ell(\theta_\ell^n)\,\pi^{\ell-1}(\theta_{\ell,C}')\,q^{\ell,F}(\theta_{\ell,F}'|\theta_{\ell,F}^n)}\right\}.$$

If we further suppose that the proposal densities $q^{\ell,C}$ and $q^{\ell,F}$ are symmetric, then

$$\alpha^{\ell,C}(\Theta'_{\ell-1} \,|\, \Theta^n_{\ell-1}) = \min\left\{1, \frac{\pi^{\ell-1}(\Theta'_{\ell-1})}{\pi^{\ell-1}(\Theta^n_{\ell-1})}\right\}$$

and

$$\alpha^{\ell}(\theta'_{\ell} \,|\, \theta^n_{\ell}) = \min\left\{1, \frac{\pi^{\ell}(\theta'_{\ell})\, \pi^{\ell-1}(\theta^n_{\ell,C})}{\pi^{\ell}(\theta^n_{\ell})\, \pi^{\ell-1}(\theta'_{\ell,C})}\right\}.$$

Proof. Let θ^a_ℓ and θ^b_ℓ be any two admissible states on level ℓ. Since the proposals for the coarse modes $\theta_{\ell,C}$ and for the fine modes $\theta_{\ell,F}$ are generated independently, the transition probability $q^\ell(\theta^b_\ell \,|\theta^a_\ell)$ can be written as a product of transition probabilities on the two parts of θ_ℓ. For the coarse level transition probability, we have to take into account the decision that was made on level $\ell - 1$. Hence,

$$q^\ell(\theta^b_\ell|\theta^a_\ell) = \alpha^{\ell,C}(\theta^b_{\ell,C}|\theta^a_{\ell,C})\, q^{\ell,C}(\theta^b_{\ell,C}|\theta^a_{\ell,C})\, q^{\ell,F}(\theta^b_{\ell,F}|\theta^a_{\ell,F}). \tag{5.19}$$

and so

$$\begin{aligned}
\frac{q^\ell(\theta^a_\ell|\theta^b_\ell)}{q^\ell(\theta^b_\ell|\theta^a_\ell)} &= \frac{\min\left\{1, \frac{\pi^{\ell-1}(\theta^a_{\ell,C})q^{\ell,C}(\theta^b_{\ell,C}|\theta^a_{\ell,C})}{\pi^{\ell-1}(\theta^b_{\ell,C})q^{\ell,C}(\theta^a_{\ell,C}|\theta^b_{\ell,C})}\right\} q^{\ell,C}(\theta^a_{\ell,C}|\theta^b_{\ell,C}) q^{\ell,F}(\theta^a_{\ell,F}|\theta^b_{\ell,F})}{\min\left\{1, \frac{\pi^{\ell-1}(\theta^b_{\ell,C})q^{\ell,C}(\theta^a_{\ell,C}|\theta^b_{\ell,C})}{\pi^{\ell-1}(\theta_{\ell,C})q^{\ell,C}(\theta^b_{\ell,C}|\theta^a_{\ell,C})}\right\} q^{\ell,C}(\theta^b_{\ell,C}|\theta^a_{\ell,C}) q^{\ell,F}(\theta^b_{\ell,F}|\theta^a_{\ell,F})} \\
&= \frac{\pi^{\ell-1}(\theta^a_{\ell,C}) q^{\ell,F}(\theta^a_{\ell,F}|\theta^b_{\ell,F})}{\pi^{\ell-1}(\theta^b_{\ell,C}) q^{\ell,F}(\theta^b_{\ell,F}|\theta^a_{\ell,F})}.
\end{aligned}$$

This completes the proof of the first result, if we choose $\theta^a_\ell := \theta^n_\ell$ and $\theta^b_\ell := \theta'_\ell$. The corollary for symmetric densities $q^{\ell,C}$ and $q^{\ell,F}$ follows by definition. □

Remark 5.5 (Recursive algorithm)**.** Note that one particular choice for the coarse level proposal density in Step 1 of Algorithm 2 on each of the levels $\ell \geq 1$ is $q^{\ell,C} := q^{\ell-1}$, i.e. the "two-level" proposal density defined in Step 2 of Algorithm 2 on level $\ell - 1$. We can apply this strategy recursively on every level and set q^0 to be, e.g., the pCN proposal density. So proposals for $Q_{\ell-1}$ and for Q_ℓ get "pre-screened" at all coarser levels, starting always at level 0. The formula for the acceptance probability α^ℓ in Lemma 5.4 does not depend on $q^{\ell,C}$ and so it remains the same. However, this choice did not prove advantageous in practice. It requires $\ell + 1$ evaluations of the likelihood on level ℓ instead of two and it does not improve the acceptance probability. Instead, we found that choosing

the pCN algorithm for $q^{\ell,C}$ (as well as for $q^{\ell,F}$) worked better.

A simplified version of Algorithm 2, making use of the symmetry of the pCN proposal density and of the formulae derived in Lemma 5.4, is given in Section 5.3 and will be used for the numerical computations.

5.2.2 Abstract convergence analysis

Let us now move on to convergence properties of the multilevel estimator. As in the standard MCMC case, let

$$K_\ell(\theta'_\ell \,|\, \theta^n_\ell) := \alpha^\ell(\theta'_\ell \,|\, \theta^n_\ell)\, q^\ell(\theta'_\ell \,|\, \theta^n_\ell) + \left(1 - \int_{\mathbb{R}^{R_\ell}} \alpha^\ell(\theta''_\ell \,|\, \theta^n_\ell)\, q^\ell(\theta''_\ell \,|\, \theta^n_\ell)\, \mathrm{d}\theta''_\ell\right) \delta(\theta^n_\ell - \theta'_\ell),$$

denote the transition kernel of $\{\theta^n_\ell\}_{n\in\mathbb{N}}$, and define, for all $\ell = 0, \ldots, L$, the sets

$$\begin{aligned}
\mathcal{E}^\ell &= \{\theta_\ell : \pi^\ell(\theta_\ell) > 0\}, \\
\mathcal{D}^\ell &= \{\theta'_\ell : q^\ell(\theta'_\ell \,|\, \theta_\ell) > 0 \text{ for some } \theta_\ell \in \mathcal{E}^\ell\}.
\end{aligned}$$

The following convergence results follow from the classical results, due to the telescoping sum property (5.17) and the algebra of limits.

Lemma 5.6. *Provided $\mathcal{E}^\ell \subset \mathcal{D}^\ell$, ν^ℓ is a stationary distribution of the chain $\{\theta^n_\ell\}_{n\in\mathbb{N}}$.*

Theorem 5.7. *Suppose that for all $\ell = 0, \ldots, L$, $\mathbb{E}_{\nu^\ell}[|Q_\ell|] < \infty$ and*

$$q^\ell(\theta_\ell \,|\, \theta'_\ell) > 0, \quad \text{for all } \ (\theta_\ell, \theta'_\ell) \in \mathcal{E}^\ell \times \mathcal{E}^\ell. \tag{5.20}$$

Then

$$\lim_{\{N_\ell\}\to\infty} \widehat{Q}^{\mathrm{ML}}_{L,\{N_\ell\}} = \mathbb{E}_{\nu^L}[Q_L], \qquad \text{for any } \ \theta^0_\ell \in \mathcal{E}^\ell \ \text{ and } \ n^\ell_0 \geq 0.$$

Let us have a closer look at the irreducibility condition (5.20). As in (5.19), we have

$$q^\ell(\theta_\ell|\theta'_\ell) = \alpha^{\ell,C}(\theta_{\ell,C}|\theta'_{\ell,C})\, q^{\ell,C}(\theta_{\ell,C}|\theta'_{\ell,C})\, q^{\ell,F}(\theta_{\ell,F}|\theta'_{\ell,F})$$

and thus (5.20) holds, if and only if, for all $(\theta_\ell, \theta'_\ell) \in \mathcal{E}^\ell \times \mathcal{E}^\ell$, $\pi^{\ell-1}(\theta_{\ell,C})$, $q^{\ell,C}(\theta'_{\ell,C}|\theta_{\ell,C})$, $q^{\ell,C}(\theta_{\ell,C}|\theta'_{\ell,C})$ and $q^{\ell,F}(\theta_{\ell,F}|\theta'_{\ell,F})$ are all positive. The final three

terms are positive for common choices of proposal distributions, such as the random walk sampler or the pCN algorithm. The first term is assured to be positive by our choice of prior density $\pi_0^\ell := \pi_0^{R_\ell}$.

We finish the abstract discussion of the new, hierarchical multilevel Metropolis-Hastings MCMC algorithm with the main theorem that establishes a bound on the ε-cost of the multilevel estimator under certain assumptions on the MCMC error, on the (weak) model error, on the strong error between the states on level ℓ and on level $\ell - 1$ (in the two-level estimator for Y_ℓ), as well as on the cost $\mathcal{C}_\ell$ to advance Algorithm 2 by one state from n to $n + 1$ (i.e. one evaluation of the likelihood on level ℓ and one on level $\ell - 1$). This is the equivalent of Theorem 4.1 in the case of Markov chain Monte Carlo estimators.

To state our assumption on the MCMC error and to define the mean square error of the estimator, we introduce the following notation. We define $\mathbf{\Theta}_\ell := \{\theta_\ell^n\}_{n\in\mathbb{N}} \cup \{\Theta_{\ell-1}^n\}_{n\in\mathbb{N}}$, for $\ell \geq 1$, and $\mathbf{\Theta}_0 := \{\theta_0^n\}_{n\in\mathbb{N}}$, and define by $\mathbb{E}_{\mathbf{\Theta}_\ell}$ (respectively $\mathbb{V}_{\mathbf{\Theta}_\ell}$) the expected value (respectively variance) with respect to the distribution of $\mathbf{\Theta}_\ell$ generated by Algorithm 2. Furthermore, let us for $\ell \geq 1$ denote by $\nu^{\ell,\ell-1}$ the joint stationary distribution of θ_ℓ and $\Theta_{\ell-1}$. $\nu^{\ell,\ell-1}$ is defined by the marginals of θ_ℓ and $\Theta_{\ell-1}$ being ν^ℓ and $\nu^{\ell-1}$, respectively, and the correlation being determined by Algorithm 2. For $\ell = 0$ define $\nu^{0,-1} := \nu^0$.

Theorem 5.8. *Let $\varepsilon < \exp[-1]$. Suppose the sequence $\{h_\ell\}_{\ell=0,1,\dots}$ satisfies (4.4), suppose there are positive constants $\alpha, \alpha', \beta, \beta', \gamma, c_{\mathrm{M1}}, c_{\mathrm{M2}}, c_{\mathrm{M3}}, c_{\mathrm{M4}} > 0$ such that $\alpha \geq \frac{1}{2}\min(\beta, \gamma)$ and $R_\ell \gtrsim h_\ell^{-\max\{\alpha/\alpha', \beta/\beta'\}}$. Under the following assumptions,*

M1. $|\mathbb{E}_{\nu^\ell}[Q_\ell] - \mathbb{E}_\rho[Q]| \;\leq\; c_{\mathrm{M1}} \;\left(h_\ell^\alpha + R_\ell^{-\alpha'}\right)$

M3. $\mathbb{V}_{\nu^{\ell,\ell-1}}[Y_\ell] \;\leq\; c_{\mathrm{M2}} \;\left(h_{\ell-1}^\beta + R_{\ell-1}^{-\beta'}\right)$

M3. $\mathbb{V}_{\mathbf{\Theta}_\ell}[\widehat{Y}_{\ell,N_\ell}^{\mathrm{MC}}] + (\mathbb{E}_{\mathbf{\Theta}_\ell}[\widehat{Y}_{\ell,N_\ell}^{\mathrm{MC}}] - \mathbb{E}_{\nu^{\ell,\ell-1}}[\widehat{Y}_{\ell,N_\ell}^{\mathrm{MC}}])^2 \;\leq\; c_{\mathrm{M3}}\; N_\ell^{-1}\, \mathbb{V}_{\nu^{\ell,\ell-1}}[Y_\ell]$

M4. $\mathcal{C}_\ell \;\leq\; c_{\mathrm{M4}}\; h_\ell^{-\gamma}$,

there exists a number of levels L and a sequence $\{N_\ell\}_{\ell=0}^L$ such that

$$e(\widehat{Q}_{L,\{N_\ell\}}^{\mathrm{ML}})^2 \;:=\; \mathbb{E}_{\cup_\ell \mathbf{\Theta}_\ell}\left[\left(\widehat{Q}_{L,\{N_\ell\}}^{\mathrm{ML}} - \mathbb{E}_\rho[Q]\right)^2\right] < \varepsilon^2,$$

and

$$\mathcal{C}_\varepsilon(\widehat{Q}^{\mathrm{ML}}_{L,\{N_\ell\}}) \lesssim \begin{cases} \varepsilon^{-2}\,|\log\varepsilon|, & \text{if } \beta > \gamma, \\ \varepsilon^{-2}\,|\log\varepsilon|^3, & \text{if } \beta = \gamma, \\ \varepsilon^{-2-(\gamma-\beta)/\alpha}\,|\log\varepsilon|, & \text{if } \beta < \gamma. \end{cases}$$

Proof. The proof of this theorem is very similar to the proof of the complexity theorem in the case of multilevel estimators based on i.i.d samples (cf. Theorem 4.1). First note that by assumption we have $R_\ell^{-\alpha'} \lesssim h_\ell^\alpha$ and $R_\ell^{-\beta'} \lesssim h_\ell^\beta$.

Furthermore, similar to (5.11), we can expand

$$e(\widehat{Q}^{\mathrm{ML}}_{L,\{N_\ell\}})^2 \leq \mathbb{V}_{\cup_\ell \mathbf{\Theta}_\ell}\left[\widehat{Q}^{\mathrm{ML}}_{L,\{N_\ell\}}\right] + 2\left(\mathbb{E}_{\cup_\ell \mathbf{\Theta}_\ell}\left[\widehat{Q}^{\mathrm{ML}}_{L,\{N_\ell\}}\right] - \mathbb{E}_{\nu^L}\left[\widehat{Q}^{\mathrm{ML}}_{L,\{N_\ell\}}\right]\right)^2 + 2\Big(\mathbb{E}_{\nu^L}[Q_L] - \mathbb{E}_\rho[Q]\Big)^2.$$

Since the second term in the MSE above can be bounded by

$$\begin{aligned}\left(\mathbb{E}_{\cup_\ell \mathbf{\Theta}_\ell}\left[\widehat{Q}^{\mathrm{ML}}_{L,\{N_\ell\}}\right] - \mathbb{E}_{\nu^L}\left[\widehat{Q}^{\mathrm{ML}}_{L,\{N_\ell\}}\right]\right)^2 &= \left(\sum_{l=0}^{L}\left(\mathbb{E}_{\mathbf{\Theta}_\ell}\left[\widehat{Y}^{\mathrm{MC}}_{\ell,N_\ell}\right] - \mathbb{E}_{\nu^{\ell,\ell-1}}[\widehat{Y}^{\mathrm{MC}}_{\ell,N_\ell}]\right)\right)^2 \\ &\leq (L+1)\sum_{l=1}^{L}\left(\mathbb{E}_{\mathbf{\Theta}_\ell}\left[\widehat{Y}^{\mathrm{MC}}_{\ell,N_\ell}\right] - \mathbb{E}_{\nu^{\ell,\ell-1}}[\widehat{Y}^{\mathrm{MC}}_{\ell,N_\ell}]\right)^2,\end{aligned}$$

it follows from assumption M3 that

$$e(\widehat{Q}^{\mathrm{ML}}_{L,\{N_\ell\}})^2 \lesssim (L+1)\sum_{\ell=0}^{L} N_\ell^{-1}\,\mathbb{V}_{\nu^{\ell,\ell-1}}[Y_\ell] + \Big(\mathbb{E}_{\nu^L}[Q_L] - \mathbb{E}_\rho[Q]\Big)^2. \tag{5.21}$$

In contrast to the MSE for multilevel estimators based on i.i.d samples considered in section 4.2, we hence have a factor $(L+1)$ multiplying the sampling error term on the right hand side of (5.21). This implies that in order to make this term less than $\varepsilon^2/2$, the number of samples N_ℓ needs to be increased by a factor of $(L+1)$ compared to the i.i.d. case. The cost of the multilevel estimator is correspondingly also increased by a factor of $(L+1)$. The remainder of the proof remains identical.

Since L is chosen such that the second term in (5.21) (the bias of the multilevel

estimator) is less than $\varepsilon^2/2$ (cf (4.10)), it follows from assumptions M1 that $L + 1 \lesssim \log \varepsilon^{-1}$. The bounds on the ε-cost then follow as in Theorem 4.1 with $\delta = 1$, but with an extra $|\log \varepsilon|$ factor. □

Assumptions M1 and M4 are the same assumptions as in the single level case, and are related to the bias in the model (due to discretisation) and to the cost per sample, respectively. Assumption M3 is similar to assumption C1, in that it is a non-asymptotic bound for the sampling errors of the MCMC estimator $\widehat{Y}^{\mathrm{MC}}_{\ell,N_\ell}$. For this assumption to hold, it is in general necessary that the chains have been sufficiently burnt in, i.e. that the values n_0^ℓ are sufficiently large.

5.2.3 Application of abstract convergence analysis

Let us now move on to quantifying the cost of the multilevel MCMC estimator, by verifying that the assumptions in Theorem 5.8 hold for our model problem. As mentioned earlier, assumption M3 involves bounding the mean square error of an MCMC estimator, and a proof of M3 is beyond the scope of this thesis. Results of this kind can be found in e.g. [62, 45]. We will also not address M4, which is an assumption on the cost of obtaining one sample of Q_ℓ. In the best case, with an optimal linear solver to solve the discretised (finite element) equations for each sample, M4 is satisfied with $\gamma \approx d$.

We will address assumptions M1 and M2, which are the assumptions related to the discretisation errors in the quantity of interest Q. However, assumption M1 also involves the discretisation error in the measure ρ. Using the triangle inequality, we have

$$|\mathbb{E}_{\nu^\ell}[Q_\ell] - \mathbb{E}_\rho[Q]| \leq \left|\mathbb{E}_{\nu^\ell}[Q_\ell - Q^{R_\ell}]\right| + \left|\mathbb{E}_{\nu^\ell}[Q^{R_\ell}] - \mathbb{E}_\rho[Q]\right|, \tag{5.22}$$

where $Q^{R_\ell} := \mathcal{G}(u_{R_\ell}(\theta_\ell))$ is the approximation of Q resulting from the approximation $a = a_\ell$ in the (exact) solution to model problem (2.1). We make the following assumption on the second term on the right hand side of (5.22).

C2. There exist constants $\eta, \eta', c_{\mathrm{C2}} > 0$ such that

$$\left|\mathbb{E}_{\nu^\ell}[Q^{R_\ell}] - \mathbb{E}_\rho[Q]\right| \leq c_{\mathrm{C2}} \left(h_\ell^\eta + R_\ell^{-\eta'}\right)$$

Assumption C2 can be proved by proving that $\mathrm{d}_{\mathrm{Hell}}(\nu^\ell, \rho) \leq c_{\mathrm{C2}} \left(h_\ell^\eta + R_\ell^{-\eta'} \right)$, where $\mathrm{d}_{\mathrm{Hell}}$ denotes the Hellinger distance [64]. This proof is again beyond the scope of the thesis, but bounds of this type have been proven for similar model problems in [64, 18, 20, 48]. In particular, it follows from [48] that for model problem (2.1) with uniformly bounded and coercive coefficients $a(\omega, x)$, assumption C2 holds with the same convergence rates as for $\mathcal{F}(u(\vartheta)) - \mathcal{F}(u_\ell(\theta_\ell))$, which for the case considered in this section would suggest the rates given in Lemma 3.7 and Corollary 3.17.

For ease of presentation, we will for the remainder of this section assume that $g = \log[a]$ has mean zero and exponential covariance function (2.25) with $p = 1$, and that $\{\phi_j\}_{j=1}^m$ and f in (2.1) are deterministic, with $\phi_j \in H^1(\Gamma_j)$ and $f \in H^{-1/2}(D)$. This implies that the solution u to (2.1) is in $L^p(\Omega, H^{1/2-\delta})$, for any $\delta > 0$. Accordingly, we shall assume that the functionals $\mathcal{G}$ and $\mathcal{F}$, where $\mathcal{G}$ is such that $Q = \mathcal{G}(u(\vartheta))$ and $\mathcal{F}$ is as in (5.2), satisfy assumption F1 in section 2.5 for some $t_* > 1/2$ and any $q_* < \infty$. In particular, this implies that $|\mathcal{G}(u) - \mathcal{G}(u_\ell)| \leq C_{\mathcal{G}}(\omega) \|(u - u_\ell)(\omega, \cdot)\|_{H^1(D)}$ and $|\mathcal{F}(u) - \mathcal{F}(u_\ell)| \leq C_{\mathcal{F}}(\omega) \|(u - u_\ell)(\omega, \cdot)\|_{H^1(D)}$, for some $C_{\mathcal{G}}(\omega), C_{\mathcal{F}}(\omega) \in L^q(\Omega)$, for any $1 \leq q < \infty$.

Under the above assumptions, it follows from Remark 3.20 that for $\rho_0^\ell := \rho_0^{R_\ell}$ (defined in (5.1) and (5.5)), we have

$$\mathbb{E}_{\rho_0^\ell} \left[|Q^{R_\ell} - Q_\ell|^q \right]^{1/q} \leq C_{a,f,\phi,q} \, h_\ell^{1-\delta}, \tag{5.23}$$

for any $\delta > 0$, where the (generic) constant $C_{a,f,\phi,q}$ (here and below) depends on the data a, f, ϕ and on q, but is independent of any other parameters.

The aim is now to generalise the convergence result in (5.23) to include the framework of the new MLMCMC estimator. There are two issues which need to be addressed. Firstly, the bounds in assumptions M1 and M2 in Theorem 5.8 involve moments with respect to the posterior distributions ν^ℓ, which are not known explicitly, but are related to the prior distribution ρ_0^ℓ through (5.6). Secondly, the samples which are used to compute the differences $Y_\ell = Q_\ell - Q_{\ell-1}$ are generated by Algorithm 2, and may differ not only due to the truncation order, but also because they come from different Markov chains (i.e. $\Theta_{\ell-1}^n$ is not necessarily equal to $\theta_{\ell,C}^n$, as seen in Table 5.1).

To circumvent the problem of the intractability of the posterior distribution,

we have the following lemma, which relates moments with respect to the posterior distribution ν^ℓ to moments with respect to the prior distribution ρ_0^ℓ.

Lemma 5.9. *For any random variable $Z = Z(\theta_\ell)$ and for any q s.t. $\mathbb{E}_{\rho_0^\ell}[|Z|^q] < \infty$, we have*

$$|\,\mathbb{E}_{\nu^\ell}[Z^q]\,| \lesssim \mathbb{E}_{\rho_0^\ell}[|Z|^q]\,.$$

Proof. Using (5.6), we have

$$\begin{aligned}|\,\mathbb{E}_{\nu^\ell}[Z^q]\,| &\eqsim \left|\int_{\mathbb{R}^{R_\ell}} Z^q(\theta_\ell)\exp\left[-\Phi^{h,R}(\theta_\ell;F_{\text{obs}})\right]\,\pi_0^\ell(\theta_\ell)\,\mathrm{d}\theta_\ell\right| \\ &\eqsim \sup_{\theta_\ell}\left\{\exp\left[-\Phi^{h,R}(\theta_\ell;F_{\text{obs}})\right]\right\}\int_{\mathbb{R}^{R_\ell}}|Z(\theta_\ell)|^q\,\pi_0^\ell(\theta_\ell)\,\mathrm{d}\theta_\ell.\end{aligned}$$

The claim of the Lemma then follows, since the above supremum can be bounded by 1. □

Note that a bound on the first term on the right hand side of (5.22) follows immediately from Lemma 5.9, together with (5.23): $\left|\mathbb{E}_{\nu^\ell}[Q_\ell - Q^{R_\ell}]\right| \leq C_{a,f,\phi}h_\ell^{1-\delta}$, for any $\delta > 0$. In order to prove M3, we further have to analyse the situation where the two samples θ_ℓ^n and $\Theta_{\ell-1}^n$ used to compute Y_ℓ^n "diverge", i.e. when $\Theta_{\ell-1}^n \neq \theta_{\ell,C}^n$.

Recall that the coarser levels in our multilevel estimator are introduced only to accelerate the convergence and that the multilevel estimator is still a consistent estimator of the expected value of Q_L with respect to the posterior ν^L on the finest level L. Hence, the posterior distributions on the coarser levels ν^ℓ, $\ell = 0,\ldots,L-1$, do not have to model the measured data as faithfully as ν^L. In particular, this means that we can choose larger values of the fidelity parameter $\sigma_{F,\ell}^2 := \sigma_{F,h_\ell}^2$ on the coarse levels, which will increase the acceptance probability on the coarser levels, since it is easier to match the model response $\mathcal{F}(u_\ell(\theta_\ell))$ with the data F_{obs}. As we will see below (cf. assumption C3), the growth in $\sigma_{F,\ell}^2$ has to be controlled. Typically, we will choose $\sigma_{F,L}^2 = \sigma_F^2$.

We need to make the following two assumptions on the parameters $\sigma_{F,\ell}^2$ in the likelihood and on the growth of the dimension R_ℓ.

C3. The dimension $R_\ell \to \infty$ as $\ell \to \infty$ and

$$(R_\ell - R_{\ell-1})(2\pi)^{-\frac{R_\ell - R_{\ell-1}}{2}} \lesssim R_{\ell-1}^{-1/2+\delta}, \quad \text{for all } \delta > 0.$$

C4. The sequence of fidelity parameters $\{\sigma_{F,\ell}^2\}_{\ell=0}^{\infty}$ satisfies

$$\sigma_{F,\ell}^{-2} - \sigma_{F,\ell-1}^{-2} \lesssim \max\left(R_{\ell-1}^{-1/2+\delta}, h_{\ell-1}^{1-\delta}\right), \quad \text{for all } \delta > 0.$$

For C3 to be satisfied it suffices that $R_\ell - R_{\ell-1}$ grows logarithmically with $R_{\ell-1}$. Assumption C4 holds for example, if we choose the fidelity parameter to be equal to σ_F^2 for all $\ell \geq \ell_0$, for some $\ell_0 \geq 0$. Note that for assumption C2 to hold, one usually requires $\sigma_{F,\ell}^2 \to \sigma_F^2$ as $\ell \to \infty$.

Under these assumptions we can now prove that assumption M2 in Theorem 5.8 is satisfied, with $\beta = 1 - \delta$ and $\beta' = 1/2 - \delta$, for any $\delta > 0$.

Lemma 5.10. *Let θ_ℓ and $\Theta_{\ell-1}$, with joint distribution $\nu^{\ell,\ell-1}$, be such that $Y_\ell = Q_\ell(\theta_\ell) - Q_{\ell-1}(\Theta_{\ell-1})$. Let (5.23), as well as assumptions C3 and C4 hold. Then*

$$\mathbb{V}_{\nu^{\ell,\ell-1}}[Y_\ell] \leq C_{a,f,\phi}\left(h_{\ell-1}^{1-\delta} + R_{\ell-1}^{-1/2+\delta}\right), \quad \textit{for any } \delta > 0.$$

To prove Lemma 5.10, we first need some preliminary results. Firstly, note that for $\Theta_{\ell-1}^n \neq \theta_{\ell,C}^n$ to be the case, the proposal generated for θ_ℓ^n had to be rejected. Given the proposal θ_ℓ' and the previous state θ_ℓ^{n-1}, the probability of this rejection is given by $1 - \alpha^\ell(\theta_\ell'|\theta_\ell^{n-1})$. We need to quantify this probability. Before we can do so, we need to specify the (marginal) distribution of the proposal θ_ℓ'. When θ_ℓ and $\Theta_{\ell-1}$ are jointly distributed as $\nu^{\ell,\ell-1}$, it follows from the construction of Algorithm 2 that the first $R_{\ell-1}$ entries of θ_ℓ' are distributed as $\nu^{\ell-1}$, since they come from $\Theta_{\ell-1}$. The remaining $R_\ell - R_{\ell-1}$ dimensions are (independent of the first $R_{\ell-1}$ dimensions) distributed according to the prior distribution ρ_0 restricted to these dimensions. This is in fact the posterior distribution $\rho^{\ell-1} := \rho^{h_{\ell-1},R_{\ell-1}}$ (on $\mathbb{R}^{\mathbb{N}}$) restricted to $\mathbb{R}^{R_\ell}$, and we shall denote this distribution by $\rho_\ell^{\ell-1}$. Using the same proof technique as in Lemma 5.9, together with the relation (5.4), we establish the following.

Lemma 5.11. *For any random variable $Z = Z(\theta_\ell)$ and for any q s.t. $\mathbb{E}_{\rho_0^\ell}[|Z|^q] < \infty$, we have*

$$\left| \mathbb{E}_{\rho_\ell^{\ell-1}}[Z^q] \right| \lesssim \mathbb{E}_{\rho_0^\ell}[|Z|^q].$$

We now have the following crucial result.

Theorem 5.12. *Suppose θ'_ℓ and θ''_ℓ have joint distribution $f(\theta'_\ell, \theta''_\ell)$, with marginal distributions $f(\theta''_\ell) = \nu^\ell$ and $f(\theta'_\ell) = \rho_\ell^{\ell-1}$. Suppose C3 and C4 hold. Then*

$$\lim_{\ell\to\infty} \alpha^\ell(\theta'_\ell \,|\, \theta''_\ell) = 1, \qquad \text{for almost all } \theta'_\ell, \theta''_\ell.$$

Furthermore,

$$\mathbb{E}_f\left[(1-\alpha^\ell)^q\right]^{1/q} \le C_{a,f,\phi,q}\Big(h_{\ell-1}^{1-\delta} + R_{\ell-1}^{-1/2+\delta}\Big),$$

for any $q \in [1,\infty)$ and $\delta > 0$.

Proof. We will first derive a bound on $1-\alpha^\ell(\theta'_\ell \,|\, \theta''_\ell)$, for $\ell > 1$ and for θ'_ℓ and θ''_ℓ given. First note that if $\frac{\pi^\ell(\theta'_\ell)\,\pi^{\ell-1}(\theta''_{\ell,C})}{\pi^\ell(\theta''_\ell)\,\pi^{\ell-1}(\theta'_{\ell,C})} \ge 1$, then $1-\alpha^\ell(\theta'_\ell \,|\, \theta''_\ell) = 0$. Otherwise, we have

$$\begin{aligned} 1-\alpha^\ell(\theta'_\ell \,|\, \theta''_\ell) &= \left(1 - \frac{\pi^\ell(\theta'_\ell)}{\pi^{\ell-1}(\theta'_{\ell,C})}\right) + \left(\frac{\pi^\ell(\theta'_\ell)\,\pi^{\ell-1}(\theta''_{\ell,C})}{\pi^\ell(\theta''_\ell)\,\pi^{\ell-1}(\theta'_{\ell,C})}\right)\left(1 - \frac{\pi^\ell(\theta''_\ell)}{\pi^{\ell-1}(\theta''_{\ell,C})}\right) \\ &\le \left|1 - \frac{\pi^\ell(\theta'_\ell)}{\pi^{\ell-1}(\theta'_{\ell,C})}\right| + \left|1 - \frac{\pi^\ell(\theta''_\ell)}{\pi^{\ell-1}(\theta''_{\ell,C})}\right|. \end{aligned} \tag{5.24}$$

Let us consider either of these two terms and set $\theta_\ell = (\xi_j)_{j=1}^{R_\ell}$ to be either θ'_ℓ or θ''_ℓ. Using (5.6), as well the prior (5.1) we have

$$\begin{aligned} \frac{\pi^\ell(\theta_\ell)}{\pi^{\ell-1}(\theta_{\ell,C})} &= \frac{\pi_0^\ell(\theta_\ell)\,\exp[-\Phi_\ell(\theta_\ell; F_{\text{obs}})]}{\pi_0^{\ell-1}(\theta_{\ell,C})\,\exp[-\Phi_{\ell-1}(\theta_{\ell,C}; F_{\text{obs}})]} \\ &= \exp\Bigg(-(2\pi)^{-\frac{R_\ell - R_{\ell-1}}{2}} \sum_{j=R_{\ell-1}+1}^{R_\ell} \frac{\xi_j^2}{2} - \frac{\|F_{\text{obs}} - \mathcal{F}(u_\ell(\theta_\ell))\|^2}{\sigma_{F,\ell}^2} \\ &\qquad + \frac{\|F_{\text{obs}} - \mathcal{F}(u_{\ell-1}(\theta_{\ell,C}))\|^2}{\sigma_{F,\ell-1}^2}\Bigg). \end{aligned} \tag{5.25}$$

Denoting $F_\ell := \mathcal{F}(u_\ell(\theta_\ell))$ and $F_{\ell-1} := \mathcal{F}(u_{\ell-1}(\theta_{\ell,C}))$, and using the triangle

inequality, we have that

$$\begin{aligned}
&\frac{\|F_{\text{obs}} - F_\ell\|^2}{\sigma_{F,\ell}^2} - \frac{\|F_{\text{obs}} - F_{\ell-1}\|^2}{\sigma_{F,\ell-1}^2} \\
&\leq \frac{\left(\|F_{\text{obs}} - F_{\ell-1}\| + \|F_\ell - F_{\ell-1}\|\right)^2}{\sigma_{F,\ell}^2} - \frac{\|F_{\text{obs}} - F_{\ell-1}\|^2}{\sigma_{F,\ell-1}^2} \\
&= \|F_{\text{obs}} - F_{\ell-1}\|^2 \left(\sigma_{F,\ell}^{-2} - \sigma_{F,\ell-1}^{-2}\right) + \frac{2\|F_{\text{obs}} - F_{\ell-1}\| + \|F_\ell - F_{\ell-1}\|}{\sigma_{F,\ell}^2}\|F_\ell - F_{\ell-1}\|.
\end{aligned}$$

By our assumptions on $\mathcal{F}$, it follows from Proposition 3.16 that

$$\|F_\ell - F_{\ell-1}\| \lesssim C(\theta_\ell)\left(\|a_\ell - a_{\ell-1}\|_{C^0(\overline{D})} + h_{\ell-1}^{1-\delta}\right),$$

for almost all θ_ℓ and for a constant $C(\theta_\ell) < \infty$ that depends on θ_ℓ only through a_ℓ. Since $\|F_{\ell-1}\|$ can be bounded independently of ℓ, for almost all θ_ℓ (again by assumption on $\mathcal{F}$), and since $\|F_{\text{obs}} - F_{\ell-1}\| \leq \|F_{\text{obs}}\| + \|F_{\ell-1}\|$, we can deduce that

$$\begin{aligned}
&\frac{\|F_{\text{obs}} - F_\ell\|^2}{\sigma_{F,\ell}^2} - \frac{\|F_{\text{obs}} - F_{\ell-1}\|^2}{\sigma_{F,\ell-1}^2} \\
&\qquad \lesssim C(\theta_\ell)\left((\sigma_{F,\ell}^{-2} - \sigma_{F,\ell-1}^{-2}) + \|a_\ell - a_{\ell-1}\|_{C^0(\overline{D})} + h_{\ell-1}^{1-\delta}\right).
\end{aligned}$$

Finally, substituting this into (5.25) and using the inequality $|1 - \exp(x)| \leq |x| \exp|x|$ we have

$$\begin{aligned}
&\left|1 - \frac{\pi^\ell(\theta_\ell)}{\pi^{\ell-1}(\theta_{\ell,C})}\right| \\
&\lesssim C(\theta_\ell)\left((2\pi)^{-\frac{R_\ell - R_{\ell-1}}{2}}\zeta_\ell + (\sigma_{F,\ell}^{-2} - \sigma_{F,\ell-1}^{-2}) + \|a_\ell - a_{\ell-1}\|_{C^0(\overline{D})} + h_{\ell-1}^{1-\delta}\right), \quad (5.26)
\end{aligned}$$

for almost all θ_ℓ, where $\zeta_\ell := \sum_{j=R_{\ell-1}+1}^{R_\ell} \xi_j^2$, i.e. a realisation of a χ^2-distributed random variable with $R_\ell - R_{\ell-1}$ degrees of freedom.

Now as $\ell \to \infty$, by assumption C3 we have $R_\ell \to \infty$ and $(2\pi)^{-(R_\ell - R_{\ell-1})/2}\zeta_\ell \to 0$, almost surely. Moreover, $h_\ell \to 0$ and it follows from Proposition 3.14 that

$\|a_\ell - a_{\ell-1}\|_{C^0(\overline{D})} \to 0$, almost surely. Hence, using also C4 we have

$$\lim_{\ell\to\infty} \left| 1 - \frac{\pi^\ell(\theta_\ell)}{\pi^{\ell-1}(\theta_{\ell,C})} \right| = 0, \qquad \text{for almost all } \theta_\ell.$$

The first claim of the Theorem then follows immediately from (5.24).

For the bound on the moments of $1-\alpha_\ell$, we use that all finite moments of $C(\theta_\ell)$ can be bounded independently of ℓ (cf. Proposition 3.14). It also follows from Propositions 3.14 and 3.15 that

$$\mathbb{E}_{\rho_0^\ell}\left[\|a_\ell - a_{\ell-1}\|^q_{C^0(\overline{D})} \right]^{1/q} \lesssim R_{\ell-1}^{-1/2+\delta}, \quad \text{for any } \ \delta > 0, \ \ q < \infty.$$

Finally, since ζ_ℓ under the prior ρ_0^ℓ is χ^2-distributed with $R_\ell - R_{\ell-1}$ degrees of freedom, we have

$$\mathbb{E}_{\rho_0^\ell}[\zeta_\ell^q] = 2^q \frac{\Gamma\left(\frac{1}{2}(R_\ell - R_{\ell-1}) + q\right)}{\Gamma\left(\frac{1}{2}(R_\ell - R_{\ell-1})\right)} \lesssim (R_\ell - R_{\ell-1})^q, \quad \text{for any } \ \delta > 0, \ \ q < \infty.$$

To bound the qth moment of $1-\alpha_\ell$, we now use (5.24). Since $(a+b)^q \lesssim a^q + b^q$, where the hidden constant depends only on q, it suffices to prove bounds on the qth moments of the two terms on the right hand side of (5.24). Minkowski's inequality, together with the definition of f, gives

$$\begin{aligned}
\mathbb{E}_f\left[(1-\alpha^\ell)^q\right]^{1/q} &\lesssim \mathbb{E}_f\left[\left| 1 - \frac{\pi^\ell(\theta'_\ell)}{\pi^{\ell-1}(\theta'_{\ell,C})} \right|^q \right]^{1/q} + \mathbb{E}_f\left[\left| 1 - \frac{\pi^\ell(\theta''_\ell)}{\pi^{\ell-1}(\theta''_{\ell,C})} \right|^q \right]^{1/q} \\
&= \mathbb{E}_{\rho_\ell^{\ell-1}}\left[\left| 1 - \frac{\pi^\ell(\theta'_\ell)}{\pi^{\ell-1}(\theta'_{\ell,C})} \right|^q \right]^{1/q} + \mathbb{E}_{\nu^\ell}\left[\left| 1 - \frac{\pi^\ell(\theta''_\ell)}{\pi^{\ell-1}(\theta''_{\ell,C})} \right|^q \right]^{1/q}.
\end{aligned}$$

The bound on the qth moment of $1-\alpha^\ell$ then follows from (5.26), assumptions C3 and C4, Minkowski's and Hölder's inequality and Lemmas 5.9 and 5.11. $\square$

We will further need the following result.

Lemma 5.13. *For any θ_ℓ, let $a_\ell(\theta_\ell) := \exp\left(\sum_{j=1}^{R_\ell} \sqrt{\mu_j}\phi_j\xi_j\right)$ and $\kappa(\theta_\ell) :=$*

$\min_{x\in\overline{D}} a_\ell(\cdot,x)$. *Let* $\theta'_\ell, \theta''_\ell$ *be as in Theorem 5.12. Then*

$$|u_\ell(\theta'_\ell) - u_\ell(\theta''_\ell)|_{H^1(D)} \lesssim \frac{C_{2.1}^{R_\ell}(\theta''_\ell)}{\kappa(\theta'_\ell)} \|a_\ell(\theta'_\ell) - a_\ell(\theta''_\ell)\|_{\mathcal{C}^0(\overline{D})}, \quad \textit{for almost all } \theta'_\ell, \theta''_\ell, \tag{5.27}$$

and

$$\mathbb{E}_f\left[|u_\ell(\theta'_\ell) - u_\ell(\theta''_\ell)|^q_{H^1(D)}\right]^{1/q} \lesssim 1, \tag{5.28}$$

for any $q < \infty$*, where the hidden constants are independent of* ℓ *and* u_ℓ*.*

Proof. Using the definition of $\kappa(\theta'_\ell)$, as well as the identity

$$\int_D a_\ell(\theta'_\ell)\nabla u_\ell(\theta'_\ell)\cdot\nabla v\,\mathrm{d}x = \int_D fv\,\mathrm{d}x = \int_D a_\ell(\theta''_\ell)\,\nabla u_\ell(\theta''_\ell)\cdot\nabla v\,\mathrm{d}x,$$

for all $v \in H^1_0(D)$, (deduced from (2.4)) we have

$$\begin{aligned}
&\kappa(\theta'_\ell)|u_\ell(\theta'_\ell) - u_\ell(\theta''_\ell)|^2_{H^1(D)} \\
&\leq \int_D a_\ell(\theta'_\ell)\,\nabla\left(u_\ell(\theta'_\ell) - u_\ell(\theta''_\ell)\right)\cdot\nabla\left(u_\ell(\theta'_\ell) - u_\ell(\theta''_\ell)\right)\,\mathrm{d}x \\
&\leq \int_D \left(a_\ell(\theta'_\ell) - a_\ell(\theta''_\ell)\right)\,\nabla u_\ell(\theta''_\ell)\cdot\nabla\left(u_\ell(\theta'_\ell) - u_\ell(\theta''_\ell)\right)\,\mathrm{d}x.
\end{aligned}$$

Due to the estimate $|u_\ell(\theta''_\ell)|_{H^1(D)} \lesssim C_{2.1}^{R_\ell}(\theta''_\ell)$ this implies (5.27).

It follows from Proposition 3.14 that $\mathbb{E}_f\left[\|a_\ell(\theta''_\ell) - a_\ell(\theta'_\ell)\|^q_{\mathcal{C}^0(\overline{D})}\right]$ and $\mathbb{E}_f\left[\kappa(\theta''_\ell)^{-q}\right]$ and can be bounded independently of ℓ. The result then follows from an application of the Minkowski and Hölder's inequalities, together with Lemmas 5.9 and 5.11. □

Using Theorem 5.12 and Lemma 5.13, we are now ready to prove Lemma 5.10.

Proof of Lemma 5.10. Firstly, we have

$$\mathbb{V}_{\nu^{\ell,\ell-1}}\left[Q_\ell(\theta^n_\ell) - Q_{\ell-1}(\Theta^n_{\ell-1})\right] \lesssim \mathbb{E}_{\nu^{\ell,\ell-1}}\left[\left(Q_\ell(\theta^n_\ell) - Q_{\ell-1}(\Theta^n_{\ell-1})\right)^2\right]. \tag{5.29}$$

Let us denote by θ'_ℓ the proposal generated for θ^n_ℓ by Algorithm 2, with $\theta'_{\ell,C} = \Theta^n_{\ell-1}$ and with some $\theta'_{\ell,F}$. Note that $\theta^n_\ell \neq \theta'_\ell$ only if this proposal was rejected. It

follows from (5.29), together with Minkowski's inequality, that

$$\begin{aligned}\mathbb{V}_{\nu^{\ell,\ell-1}}&\left[Q_\ell(\theta_\ell^n) - Q_{\ell-1}(\Theta_{\ell-1}^n)\right] \\ &\lesssim \mathbb{E}_f\left[\left(Q_\ell(\theta_\ell^n) - Q_\ell(\theta_\ell')\right)^2\right] + \mathbb{E}_{\rho_\ell^{\ell-1}}\left[\left(Q_\ell(\theta_\ell') - Q_{\ell-1}(\theta_{\ell,C}')\right)^2\right], \end{aligned} \tag{5.30}$$

where, by construction, the joint distribution $f(\theta_\ell^n, \theta_\ell')$ is such that the marginal distributions are $f(\theta_\ell^n) = \nu^\ell$ and $f(\theta_\ell') = \rho_\ell^{\ell-1}$, and we have used that $\theta_{\ell,C}' = \Theta_{\ell-1}^n$. A bound on the second term follows immediately from (5.23) and Lemma 5.11, i.e.

$$\mathbb{E}_{\rho_\ell^{\ell-1}}\left[\left(Q_\ell(\theta_\ell') - Q_{\ell-1}(\Theta_{\ell-1}^n)\right)^2\right] \;\leq\; C_{a,f,\phi}\,\left(h_{\ell-1}^{2-\delta} + R_{\ell-1}^{-1+\delta}\right). \tag{5.31}$$

The first term in (5.30) is nonzero only if $\theta_\ell^n \neq \theta_\ell'$. We will now use Theorem 5.12 and Lemma 5.13, as well as the characteristic function $\mathbb{I}_{\{\theta_\ell^n \neq \theta_\ell'\}} \in \{0,1\}$ to bound it. Firstly, Hölder's inequality gives

$$\begin{aligned}\mathbb{E}_f\left[\left(Q_\ell(\theta_\ell^n) - Q_\ell(\theta_\ell')\right)^2\right] \;&=\; \mathbb{E}_f\left[\left(Q_\ell(\theta_\ell^n) - Q_\ell(\theta_\ell')\right)^2 \mathbb{I}_{\{\theta_\ell^n \neq \theta_\ell'\}}\right] \\ &\leq\; \mathbb{E}_f\left[\left(Q_\ell(\theta_\ell^n) - Q_\ell(\theta_\ell')\right)^{2q_1}\right]^{1/q_1} \mathbb{E}_f\left[\mathbb{I}_{\{\theta_\ell^n \neq \theta_\ell'\}}\right]^{1/q_2}, \end{aligned} \tag{5.32}$$

for any q_1, q_2 s.t. $q_1^{-1} + q_2^{-1} = 1$. By our assumptions on $\mathcal{G}$, it follows from Lemma 5.13 that the term $\mathbb{E}_f\big[\,\left(Q_\ell(\theta_\ell^n) - Q_\ell(\theta_\ell')\right)^{2q_1}\,\big]^{1/q_1}$ in (5.32) can be bounded by a constant independent of ℓ, for any $q_1 < \infty$. Moreover, using the law of total expectation, we have

$$\mathbb{E}_f\left[I_{\{\theta_\ell^n \neq \theta_\ell'\}}\right] = \mathbb{E}_f\Big[\mathbb{P}\left[\theta_\ell^n \neq \theta_\ell' \,|\, \theta_\ell^n, \theta_\ell'\right]\Big].$$

Since $\theta_\ell^n \neq \theta_\ell'$ only if the proposal θ_ℓ' has been rejected on level ℓ at the nth step, and in this case $\theta_\ell^{n-1} = \theta_\ell^n$, the probability that this happens can be bounded by $1 - \alpha^\ell(\theta_\ell'|\theta_\ell^n)$, and so it follows by Theorem 5.12 that

$$\mathbb{E}_f\left[I_{\{\theta_\ell^n \neq \theta_\ell'\}}\right] \leq \mathbb{E}_f\left[1 - \alpha^\ell(\theta_\ell'|\theta_\ell^n)\right] \;\lesssim\; h_{\ell-1}^{1-\delta} + R_{\ell-1}^{-1/2+\delta} \tag{5.33}$$

Combining (5.30)-(5.33) the claim of the Lemma then follows. □

We now collect the results in the preceding lemmas to state our main result of this section.

Theorem 5.14. *Let a, f, ϕ, $\mathcal{F}$ and $\mathcal{G}$ be as described at the beginning of this section, and suppose assumption C2 holds with $\eta = 1 - \delta$ and $\eta' = 1/ - \delta$, for any $\delta > 0$. Under the same assumptions as in Lemma 5.10, the assumptions M1 and M2 in Theorem 5.8 are satisfied, with $\alpha = \beta = 1 - \delta$ and $\alpha' = \beta' = 1/2 - \delta$, for any $\delta > 0$.*

If we assume that we can obtain individual samples in optimal cost $\mathcal{C}_\ell \lesssim h_\ell^{-d} \log(h_\ell^{-1})$, e.g. via a multigrid solver, we can satisfy assumption M5 with $\gamma = 1 + \delta$, for any $\delta > 0$. We assume that assumptions M1 and M4 hold, with $\alpha' = \beta' = 1 - \delta$. Then it follows from Theorems 5.8 and 5.14, as well as equation (5.16), that we can get the following theoretical upper bounds for the ε-costs of classical and multilevel MCMC applied to model problem (2.1) with log-normal coefficients a, respectively:

$$\mathcal{C}_\varepsilon(\widehat{Q}_N^{\mathrm{MC}}) \lesssim \varepsilon^{-(d+2)-\delta} \quad \text{and} \quad \mathcal{C}_\varepsilon(\widehat{Q}_{L,\{N_\ell\}}^{\mathrm{ML}}) \lesssim \varepsilon^{-(d+1)-\delta}, \quad \text{for any } \delta > 0. \tag{5.34}$$

We clearly see the advantages of the multilevel method, which gives a saving of one power of ε compared to the standard MCMC method. Note that for multilevel estimators based on i.i.d samples, the savings of the multilevel method over the standard method are two powers of ε for $d = 2, 3$. The larger savings stem from the fact that $\beta = 2\alpha$ in this case, compared to $\beta = \alpha$ in the MCMC analysis above. The numerical results in the next section for $d = 2$ show that in practice we do seem to observe $\beta \approx 1 \approx 2\alpha$, suggesting $\mathcal{C}_\varepsilon(\widehat{Q}_{L,\{N_\ell\}}^{\mathrm{ML}}) = \mathcal{O}(\varepsilon^{-d})$. However, we do not believe that this is a lack of sharpness in our theory, but rather a pre-asymptotic phase. The constant in front of the leading order term in the bound of $\mathbb{V}_{\nu^{\ell,\ell-1}}[Y_\ell]$, namely the term $\mathbb{E}_f\big[\,(Q_\ell(\theta_\ell^n) - Q_\ell(\theta_\ell'))^{2q_1}\,\big]^{1/q_1}$ in (5.32), depends on the difference between $Q_\ell(\theta_\ell^n)$ and $Q_\ell(\theta_\ell')$. In the case of the pCN algorithm for the proposal distributions $q^{\ell,C}$ and $q^{\ell,F}$ (as used in Section 5.3 below) this difference will be small, since θ_ℓ^n and θ_ℓ' will in general be very close to each other. However, the difference is bounded from below and so we should eventually see the slower convergence rate for the variance as predicted by our theory.

5.3 Numerics

In this section we describe the implementation details of the MLMCMC algorithm and examine the performance of the method in estimating the expected value of some quantity of interest. We start by presenting in Algorithm 3 a simplified version of Algorithm 2 given in Section 5.2 using symmetric proposal distributions for $q^{\ell,C}$ and $q^{\ell,F}$, describing in some more detail the evolution of the multilevel Markov chains used to compute samples of Y_ℓ.

Implementation Details

Given the general description of the multilevel sampling in Algorithm 3, it remains to describe several computational details of the method, such as the choice of the symmetric transition densities $q^{\ell,C}(\Theta'_{\ell-1}|\Theta^n_{\ell-1})$ and $q^{\ell,F}(\theta'_{\ell,F}|\theta^n_{\ell,F})$, the values R_ℓ defining the partition of the KL modes over the multilevel hierarchy, as well as various MCMC tuning parameters.

For all our symmetric proposal densities $q(\theta'|\theta^n)$ we use the so-called preconditioned Crank-Nicholson (pCN) random walk proposed by Cotter et al. in [19]. Given the current state θ^n, the j^{th} entry of the proposal is obtained by

$$\theta'_j = \sqrt{1-\beta^2}\,\theta^n_j \;+\; \beta\,\zeta_j,$$

where $\zeta_j \sim \mathcal{N}(0,1)$ and β is a tuning parameter used to control the size of the step in the proposal, that may be chosen level dependent, i.e. $\beta = \beta_\ell$. In the numerical experiments, we typically choose $\beta_\ell < \beta_0$ for $\ell = 1, \ldots, L$.

The other free parameters in Algorithm 3 are the parameters $\sigma^2_{F,\ell}$ found in the likelihood model described in (5.4). The value of $\sigma^2_{F,\ell}$ controls the fidelity with which we require the model response to match the observed data on level ℓ. In our implementation we fix the fine-level likelihood variance $\sigma^2_{F,L}$ to a value consistent with traditional single level MCMC simulations (i.e. the measurement error associated with F_{obs} in a practical application), and then allow the remaining parameters to increase on coarser levels. In particular, we choose

$$\sigma^2_{F,\ell} = (1+h_l)\,\sigma^2_{F,\ell+1}, \quad \ell = 0, \ldots, L-1.$$

To reduce dependence of the simulation on the initial state of the Markov

ALGORITHM 3. (Simplified Metropolis Hastings MCMC for Y_ℓ, $\ell > 0$)

Choose initial states $\Theta^0_{\ell-1}$ and θ^0_ℓ. For $n \geq 0$:

- On level $\ell - 1$:
  - Given $\Theta^n_{\ell-1}$, generate $\Theta'_{\ell-1}$ from a symmetric distribution $q^{\ell,C}(\Theta'_{\ell-1}|\Theta^n_{\ell-1})$.
  - Compute
$$\alpha^{\ell,C}(\Theta'_{\ell-1}|\Theta^n_{\ell-1}) = \min\left\{1, \frac{\pi^{\ell-1}(\Theta'_{\ell-1})}{\pi^{\ell-1}(\Theta^n_{\ell-1})}\right\}.$$
  - Set $\Theta^{n+1}_{\ell-1} = \begin{cases} \Theta'_{\ell-1} & \text{with probability } \alpha^{\ell,C}(\Theta'_{\ell-1}|\Theta^n_{\ell-1}) \\ \Theta^n_{\ell-1} & \text{with probability } 1 - \alpha^{\ell,C}(\Theta'_{\ell-1}|\Theta^n_{\ell-1}). \end{cases}$
- On level ℓ:
  - Given θ^n_ℓ, let $\theta'_{\ell,C} = \Theta^{n+1}_{\ell-1}$ and draw $\theta'_{\ell,F}$ from a symmetric distribution $q^{\ell,F}(\theta'_{\ell,F}|\theta^n_{\ell,F})$.
  - Compute
$$\alpha^\ell(\theta'_\ell|\theta^n_\ell) = \min\left\{1, \frac{\pi^\ell(\theta'_\ell)}{\pi^\ell(\theta^n_\ell)}\frac{\pi^{\ell-1}(\theta^n_{\ell,C})}{\pi^{\ell-1}(\theta'_{\ell,C})}\right\}.$$
  - Set $\theta^{n+1}_\ell = \begin{cases} \theta'_\ell \equiv [\Theta^{n+1}_{\ell-1}, \theta'_{\ell,F}] & \text{with probability } \alpha^\ell(\theta'_\ell|\theta^n_\ell) \\ \theta^n_\ell & \text{with probability } 1 - \alpha^\ell(\theta'_\ell|\theta^n_\ell). \end{cases}$
- Compute
$$Y^{n+1}_\ell = Q_\ell\left(\theta^{n+1}_\ell\right) - Q_{\ell-1}\left(\Theta^{n+1}_{\ell-1}\right)$$

chain, and to aid in the exploration of the potentially multi-modal stochastic space, we simulate multiple parallel chains simultaneously. The variance of the multilevel estimator $\mathbb{V}_{\boldsymbol{\Theta}_\ell}[\widehat{Y}^{\mathrm{MC}}_{\ell,N_\ell}]$ is approximated on each grid level by $s^2_{\ell,N}$ using the method of Gelman and Rubin [29]. Finally, due to the very high-dimensional parameter space in our numerical experiments, both the single-level and multilevel samplers displayed poor mixing properties. As such, we use a thinning process to decrease the correlation between consecutive samples, whereby we include only every T^{th} sample in the approximation of the level-dependent estimator, where T is some integer thinning parameter [61]. Then, after discarding n_0 initial burn-in samples, the approximation of $\mathbb{E}_{\nu^{\ell,\ell-1}}[Y_\ell]$ is computed by

$$\widehat{Y}^{\mathrm{MC}}_{\ell,N_\ell} := \frac{1}{N_\ell} \sum_{n=n_0^\ell+1}^{n_0+N_\ell} Y_\ell^{(nT)}.$$

After the initial burn-in phase, the multilevel MCMC simulation is run until the sum of the sample variances of the $L+1$ estimators satisfies

$$\sum_{\ell=0}^{L} \frac{s^2_{\ell,N_\ell}}{N_\ell} \;\leq\; \frac{\varepsilon^2}{2}$$

for some user prescribed tolerance ε. The number of samples on each level is chosen to satisfy

$$N_\ell \propto \sqrt{\mathbb{V}_{\nu^{\ell,\ell-1}}\left[Y_\ell\right]/\mathcal{C}_\ell} \approx \sqrt{s^2_{\ell,N_\ell}/\mathcal{C}_\ell}, \tag{5.35}$$

as described in (4.9), where $\mathcal{C}_\ell$ is the cost of generating a single sample of Y_ℓ on level ℓ. We assume this cost can be expressed as

$$\mathcal{C}_\ell = \mathrm{C}^* \eta_\ell^\gamma h_\ell^{-\gamma},$$

where the constant C^* may depend on the parameters σ^2 and λ in (2.25), but does not depend on ℓ. The factors η_ℓ reflect the additional cost for the auxiliary coarse solve required on grid $\ell-1$. For the experiments presented below, with geometric coarsening by a factor of 2, we have $\eta_0 = 1$ and $\eta_\ell = 1.25$, for $\ell = 1, \ldots, L$. When an optimal linear solver (e.g. algebraic multigrid) is used to perform the forward solves in the simulation we can take $\gamma \approx d$. For a given accuracy ε, the

(standardised) total cost of the multilevel estimator can be written as

$$\mathcal{C}_\varepsilon\left(\widehat{Q}^{\mathrm{ML}}_{L,\{N_\ell\}}\right) := N_0 + \sum_{\ell=1}^{L} N_\ell\, \eta_\ell^\gamma \left(\frac{M_\ell}{M_0}\right)^\gamma, \tag{5.36}$$

where $M_\ell = h_\ell^{-2}$.

Numerical experiments

We consider the mixed problem

$$\begin{aligned} -\nabla\cdot(a(\omega,x)\nabla u(\omega,x)) &= 1, \qquad \text{for } \ x\in D, \\ \text{and} \quad u\big|_{x_1=0} = 1, \quad u\big|_{x_1=1} &= 0, \quad \frac{\partial u}{\partial \mathbf{n}}\Big|_{x_2=0} = 0, \quad \frac{\partial u}{\partial \mathbf{n}}\Big|_{x_2=1} = 0, \end{aligned} \tag{5.37}$$

defined on the domain $D = (0,1)^2$. The quantity of interest is the average outflow through the boundary $\{x_1 = 1\}$ computed via the functional $M_\omega^{(4)}$ in section 2.5.

The (prior) conductivity field is modelled as a log-normal random field with 1-norm exponential covariance function (2.24). The "observed" data F_{obs} is obtained synthetically by generating a reference conductivity field from the prior, solving the forward problem, and evaluating the pressure at 9 randomly selected points in the domain. The grid hierarchy in the multilevel estimator is chosen as $h_0 = 1/16$, and $h_\ell = h_{\ell-1}/2$, for $\ell = 1,\ldots,L$. Five parallel chains are used in each estimator $\widehat{Y}^{\mathrm{MC}}_{\ell,N_\ell}$.

Figure 5-1 shows the results of a four-level simulation with $\lambda = 0.5$ and $\sigma^2 = 1$. The partitioning of the KL modes was such that $R_0 = 96$, $R_1 = 121$, $R_2 = 153$, and $R_3 = 169$. The fidelity parameter in the likelihood on the finest grid was taken to be $\sigma^2_{F,L} = 10^{-4}$. The top two plots show the variance (respectively the mean) of Q_ℓ and Y_ℓ on each level. The variance and mean of Y_ℓ seem to decay with $\mathcal{O}(h_\ell^2)$ and $\mathcal{O}(h_\ell)$, respectively. This suggests that at least in the pre-asymptotic phase our theoretical result on the variance which predicts $\mathcal{O}(h_\ell)$ (in Theorem 5.14) is not sharp (see comments at the end of Section 5.2). The result on the bias seems to be confirmed.

The bottom right plot in Figure 5-1 shows the number of samples N_ℓ required on each level of the multilevel MCMC sampler. The bottom left plot compares the (standardised) computational cost of the standard and multilevel MCMC

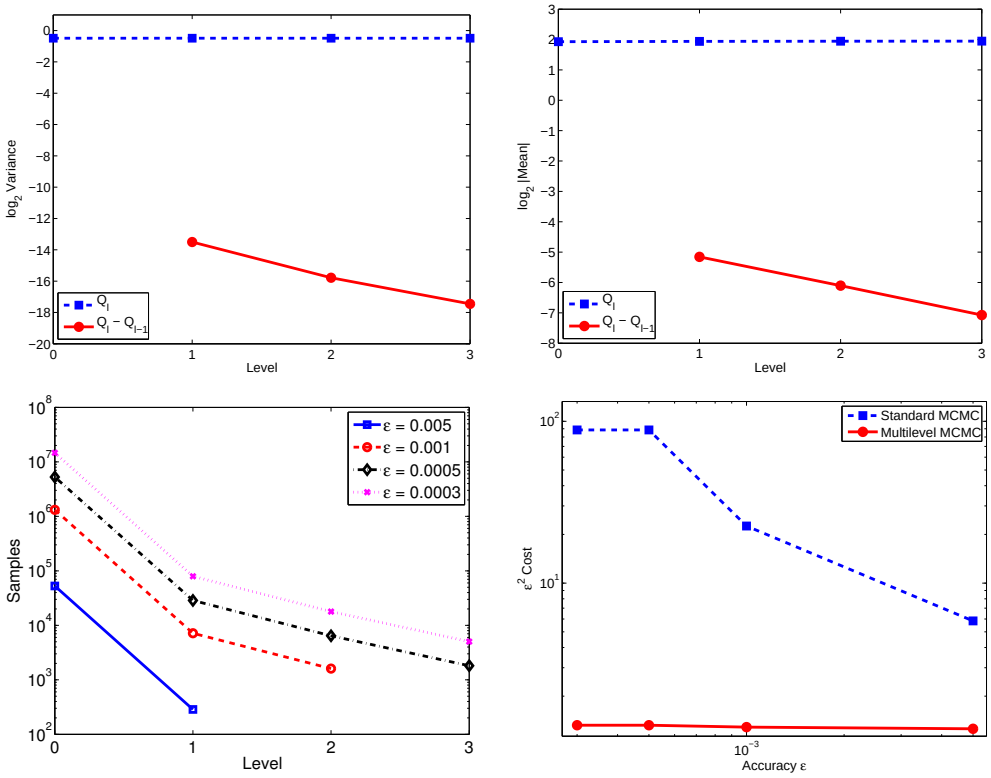

Figure 5-1: Performance plots for $\lambda = 0.5$, $\sigma^2 = 1$, $R_L = 169$, and $h_0 = 1/16$.

samplers for varying values of accuracy ε. The vertical axis is scaled by ε^2. It is clear that the multilevel sampler attains a dramatic reduction in computational cost over the standard MCMC sampler. The precise speedup of the multilevel over the standard algorithm can be evaluated by taking the ratio of the total cost of the respective estimators, as defined by (5.35)-(5.36). When an optimal linear solver (such as AMG, with $\gamma \approx d$) is used for the forward solves in the four-level simulation with $\varepsilon = 8 \times 10^{-4}$ (as in Figure 5-1), the computational cost of the simulation is reduced by a factor of 50. When a suboptimal linear solver is used (say, $\gamma \approx 1.5d$ for a sparse direct method) the computational cost is reduced by a factor of 275 for the same value of ε.

Figure 5-2 (left) confirms that the average acceptance rates α^ℓ of the fine-level samplers – the three right most data points in Figure (5-2) (left) – tend to 1 as ℓ increases, and $\mathbb{E}[1 - \alpha^\ell] \approx \mathcal{O}(h_\ell)$, as predicted in Theorem 5.12. Finally, the results in Figure 5-2 (right) demonstrate the good agreement between the MLMCMC estimate $\widehat{Q}^{\mathrm{ML}}_{L,\{N_\ell\}}$ and the standard MCMC estimate $\widehat{Q}^{\mathrm{MC}}_N$ of the quantity of interest $M^{(4)}_\omega(u)$ for nine distinct sets of reference data with three levels of fine-grid resolution. As before, the coarse grid in each case was defined with $h_0 = 1/16$, the tolerance for both estimators was $\varepsilon = 8 \times 10^{-4}$ and the model for the log-normal conductivity field is parametrised by $\lambda = 0.5$, $\sigma^2 = 1$ and $R_L = 169$ on the finest grid.

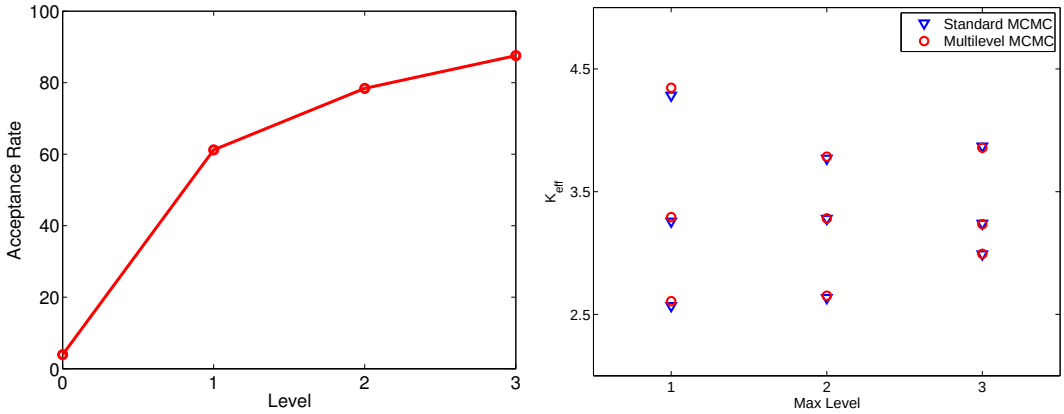

Figure 5-2: Average acceptance rate α^{ℓ} of the multilevel sampler (left figure) and estimates for the outflow $M_{\omega}^{(4)}(u)$ for nine reference data sets (right figure) for $\lambda = 0.5$, $\sigma^2 = 1$, $R_L = 169$, and $h_0 = 1/16$.

Chapter 6

Conclusion

Multilevel Monte Carlo methods have the potential to significantly outperform standard Monte Carlo methods in a variety of contexts. In this thesis, we considered the application of multilevel Monte Carlo methods to elliptic PDEs with random coefficients, in the practically relevant and technically demanding case where realisations of the diffusion coefficient have limited spatial regularity and are not uniformly bounded or elliptic. Our setting includes, for example, log-normal coefficients with very short correlation lengths.

Since the analysis of the discretisation error of the multilevel Monte Carlo estimator requires knowledge of the regularity of the solution, the first chapter of this thesis was devoted to establishing regularity results for the solution of a linear, second-order elliptic PDE with random coefficients. For a wide class of coefficients, we showed that the solution lies in the Bochner space $L^p(\Omega, H^{1+s}(D))$, where the value of s is determined by the Hölder regularity of the coefficient, as well as the geometry of the spatial domain. For log-normal coefficients with exponential covariance, we can for example choose any $s < 1/2$ for any Lipschitz polygonal domain. Our regularity results are optimal in the sense that the value of s is the same as if we would replace the random coefficient by a deterministic coefficient with the same spatial regularity. There is no loss in regularity due to the randomness.

Using these new regularity results, we then furnished a complete finite element error analysis using classical tools such as Cea's lemma and a best approximation result. We proved that all finite moments of the error in the natural H^1-norm converge with $\mathcal{O}(h^s)$, where s is as in the regularity estimate above. This rate is

optimal with respect to the regularity of the solution. Rates of convergence were established also for the error in other spatial norms and the error in functionals.

The discretisation error analysis was finally used to give a rigorous bound on the computational cost of multilevel Monte Carlo estimators. In the case of log-normal random coefficients with exponential covariance function, for example, we showed that the savings of multilevel Monte Carlo compared to standard Monte Carlo are in most cases 2 powers of ε, if the aim is to achieve a mean square error of ε^2. For the multilevel Markov chain Monte Carlo estimator, with log-normal random coefficient with exponential covariance function chosen as prior distribution, the corresponding savings are 1 power of ε.

Appendix A

Detailed Regularity proof

In this appendix, we give more detailed proofs of Theorem 2.7 and Lemmas 2.8 and 2.9. The proofs follow those of Hackbusch [44, Theorems 9.1.8, 9.1.11 and 9.1.16], making explicit the dependence of all the constants that appear on the PDE coefficient $\mathbf{A}$. The proof follows the classical Nirenberg translation method and consists of three main steps.

A.1 Step 1 – The Case $D = \mathbb{R}^d$

Proof of Lemma 2.8. In this proof we will use the norm on $H^s(\mathbb{R}^d)$ provided by the Fourier transform, $\|u\|^2_{H^s(\mathbb{R}^d)} := \|\hat{u}(\xi)(1+|\xi|^2)^{s/2}\|_{L^2(\mathbb{R}^d)}$, which is equivalent to the norm defined previously and defines the same space.

For any $h > 0$, we define the fractional difference operator (in direction $i = 1, \dots, d$) by

$$R^i_h(v)(x) := h^{-s} \sum_{\mu=0}^{+\infty} e^{-\mu h} (-1)^\mu \binom{s}{\mu} v(x + \mu h e_i),$$

where e_i is the ith unit vector in $\mathbb{R}^d$,

$$\binom{s}{0} = 1 \quad \text{and} \quad \binom{s}{\mu} (-1)^\mu = \frac{-s(1-s)(2-s)...(\mu-1-s)}{\mu!}.$$

Let us recall here some properties of R^i_h from [44, Proof of Theorem 9.1.8]:

- $(R^i_h)^*(v)(x) = h^{-s} \sum_{\mu=0}^{+\infty} e^{-\mu h} (-1)^\mu \binom{s}{\mu} v(x - \mu h e_i)$.

- For any $\tau \in \mathbb{R}$ and $v \in H^{\tau+s}(\mathbb{R}^d)$,

$$\|R_h^i v\|_{H^\tau(\mathbb{R}^d)} \leq \|v\|_{H^{\tau+s}(\mathbb{R}^d)} \text{ and } \|(R_h^i)^* v\|_{H^\tau(\mathbb{R}^d)} \leq \|v\|_{H^{\tau+s}(\mathbb{R}^d)}. \qquad (A.1)$$

- $\widehat{R_h^i v}(\xi) = [(1-e^{-h+i\xi_j h})/h]^s \hat{v}(\xi)$ and $\widehat{(R_h^i)^* v}(\xi) = [(1-e^{-h-i\xi_j h})/h]^s \hat{v}(\xi)$.

We define for $u, v \in H^1(\mathbb{R}^d)$ the bilinear form

$$\begin{aligned} d(u,v) &:= \int_{\mathbb{R}^d} \mathbf{T}\, \nabla u \nabla R_h^i(v)\, \mathrm{d}x - \int_{\mathbb{R}^d} \mathbf{T}\, \nabla (R_h^i)^* u \nabla v\, \mathrm{d}x \\ &= \sum_{\mu=1}^{\infty} h^{-s} e^{-\mu h} (-1)^\mu \binom{s}{\mu} \int_{\mathbb{R}^d} (\mathbf{T}(x-\mu h e_i) - \mathbf{T}(x)) \nabla u(x-\mu h e_i) \nabla v\, \mathrm{d}x. \end{aligned}$$

Hence,

$$|d(u,v)| \lesssim |\mathbf{T}|_{\mathcal{C}^t(\mathbb{R}^d, \mathbb{R}^{d\times d})} |u|_{H^1(\mathbb{R}^d)} |v|_{H^1(\mathbb{R}^d)},$$

where the hidden constant is proportional to

$$\sum_{\mu=1}^{\infty} h^{-s} e^{-\mu h} (-1)^\mu \binom{s}{\mu} (\mu h)^t,$$

which is finite, since $\binom{s}{\mu} = \mathcal{O}(\mu^{-s-1})$ and thus $\sum_{\mu=1}^{\infty} e^{-\mu h} \mu^{t-s-1} = \mathcal{O}(h^{s-t})$. The three spaces $H^{1-s}(\mathbb{R}^d) \subset L^2(\mathbb{R}^d) \subset H^{s-1}(\mathbb{R}^d)$ form a Gelfand triple, so that we can deduce, using $(A.1)$, that

$$\begin{aligned} \mathbf{T}_{\min} |(R_h^i)^* w|^2_{H^1(\mathbb{R}^d)} &\leq \int_{\mathbb{R}^d} \mathbf{T}\, \nabla (R_h^i)^* w \nabla (R_h^i)^* w\, \mathrm{d}x \\ &= -d(w, (R_h^i)^* w) + \langle F, R_h^i (R_h^i)^* w \rangle_{H^{s-1}(\mathbb{R}^d), H^{1-s}(\mathbb{R}^d)} \\ &\leq |d(w, (R_h^i)^* w)| + \|F\|_{H^{s-1}(\mathbb{R}^d)} \|R_h^i (R_h^i)^* w\|_{H^{1-s}(\mathbb{R}^d)} \\ &\lesssim |\mathbf{T}|_{\mathcal{C}^t(\mathbb{R}^d, \mathbb{R}^{d\times d})} |w|_{H^1(\mathbb{R}^d)} |(R_h^i)^* w|_{H^1(\mathbb{R}^d)} + \|F\|_{H^{s-1}(\mathbb{R}^d)} \|(R_h^i)^* w\|_{H^1(\mathbb{R}^d)}, \end{aligned}$$

therefore we get

$$\begin{aligned}
&\mathbf{T}_{\min}\|(R_h^i)^*w\|_{H^1(\mathbb{R}^d)}^2 \\
&\quad\lesssim |\mathbf{T}|_{\mathcal{C}^t(\mathbb{R}^d,\mathbb{R}^{d\times d})}\,|w|_{H^1(\mathbb{R}^d)}\,|(R_h^i)^*w|_{H^1(\mathbb{R}^d)} \;+\; \|F\|_{H^{s-1}(\mathbb{R}^d)}\,\|(R_h^i)^*w\|_{H^1(\mathbb{R}^d)} \\
&\qquad\qquad +\; \mathbf{T}_{\min}\|(R_h^i)^*w\|_{L^2(\mathbb{R}^d)}^2 \\
&\quad\lesssim |\mathbf{T}|_{\mathcal{C}^t(\mathbb{R}^d,\mathbb{R}^{d\times d})}\,|w|_{H^1(\mathbb{R}^d)}\,|(R_h^i)^*w|_{H^1(\mathbb{R}^d)} \;+\; \|F\|_{H^{s-1}(\mathbb{R}^d)}\,\|(R_h^i)^*w\|_{H^1(\mathbb{R}^d)} \\
&\qquad\qquad +\; \mathbf{T}_{\min}\|(R_h^i)^*w\|_{H^{-1}(\mathbb{R}^d)}\|(R_h^i)^*w\|_{H^1(\mathbb{R}^d)},
\end{aligned}$$

and finally, using $(A.1)$ once more,

$$\|(R_h^i)^*w\|_{H^1(\mathbb{R}^d)} \;\lesssim\; \frac{1}{\mathbf{T}_{\min}}\Big(|\mathbf{T}|_{\mathcal{C}^t(\mathbb{R}^d,\mathbb{R}^{d\times d})}\,|w|_{H^1(\mathbb{R}^d)} + \|F\|_{H^{s-1}(\mathbb{R}^d)}\Big) + \|w\|_{L^2(\mathbb{R}^d)}.$$

For any $1 \geq h > 0$, since $|1-e^{-h-i\xi_i h}|^2 \geq |\mathrm{Im}(1-e^{-h-i\xi_i h})|^2 = e^{-2h}\sin(\xi_i h)^2 \geq e^{-2}\sin(\xi_i h)^2$, and since $\sin^2(\xi h) \geq \big(\frac{2}{\pi}\xi h\big)^2$, for all $|\xi| \leq 1/h$, we have conversely that

$$\begin{aligned}
\sum_{i=1}^d \|(R_h^i)^*w\|_{H^1(\mathbb{R}^d)}^2 \;&\geq\; \int_{|\xi|\leq 1/h}(1+|\xi|^2)\sum_{i=1}^d |\widehat{(R_h^i)^*w}(\xi)|^2 d\xi \\
&=\; \int_{|\xi|\leq 1/h}(1+|\xi|^2)\sum_{i=1}^d \left|\frac{1-e^{-h-i\xi_i h}}{h}\right|^{2s}|\hat{w}(\xi)|^2 d\xi \\
&\geq\; e^{-2}\int_{|\xi|\leq 1/h}(1+|\xi|^2)\sum_{i=1}^d \left|\frac{\sin(\xi_i h)}{h}\right|^{2s}|\hat{w}(\xi)|^2 d\xi \\
&\gtrsim\; \int_{|\xi|\leq 1/h}(1+|\xi|^2)|\xi|^{2s}|\hat{w}(\xi)|^2 d\xi.
\end{aligned}$$

Hence, for any $0 < h \leq 1$, we obtain

$$\begin{aligned}
\|w\|_{H^{1+s}(\mathbb{R}^d)}^2 \;&\leq\; \int_{\mathbb{R}^d}(1+|\xi|^2)|\xi|^{2s}|\hat{w}(\xi)|^2 d\xi \;+\; \int_{\mathbb{R}^d}(1+|\xi|^2)|\hat{w}(\xi)|^2 d\xi \\
&\leq\; \sum_{i=1}^d \|(R_h^i)^*w\|_{H^1(\mathbb{R}^d)}^2 \;+\; \|w\|_{H^1(\mathbb{R}^d)}^2
\end{aligned}$$

and so

$$\begin{aligned}\|w\|_{H^{1+s}(\mathbb{R}^d)} &\lesssim \frac{1}{\mathbf{T}_{\min}}\Big(|\mathbf{T}|_{\mathcal{C}^t(\mathbb{R}^d,\mathbb{R}^{d\times d})}\,|w|_{H^1(\mathbb{R}^d)} + \|F\|_{H^{s-1}(\mathbb{R}^d)}\Big) + \|w\|_{H^1(\mathbb{R}^d)}\\ &< +\infty.\end{aligned}$$

□

A.2 Step 2 – The Case $D = \mathbb{R}^d_+$

Proof of Lemma 2.9. First we extend the solution w by 0 on $\mathbb{R}^d \setminus \mathbb{R}^d_+$ and denote the extension $\tilde{w} \in H^1(\mathbb{R}^d)$. Take $1 \le i \le d-1$. Similarly to the previous section, we define for $u, v \in H^1(\mathbb{R}^d)$

$$d(u,v) := \int_{\mathbb{R}^d_+} \mathbf{T}\,\nabla u \nabla R^i_h(v)\,\mathrm{d}x - \int_{\mathbb{R}^d_+} \mathbf{T}\,\nabla (R^i_h)^* u \nabla v\,\mathrm{d}x$$

and deduce again that

$$|d(u,v)| \lesssim |\mathbf{T}|_{\mathcal{C}^t(\overline{\mathbb{R}^d_+},\mathbb{R}^{d\times d})}\,|u|_{H^1(\mathbb{R}^d_+)}\,|v|_{H^1(\mathbb{R}^d_+)}\,.$$

We now note that, since $i \neq d$, $(R^i_h)^* w \in H^1_0(\mathbb{R}^d_+)$ and $(R^i_h)^*\tilde{w} \in H^1(\mathbb{R}^d)$ is equal to the extension by 0 on $\mathbb{R}^d \setminus \mathbb{R}^d_+$ of $(R^i_h)^* w$. We deduce, similarly to the proof in Section A.1 using $(A.1)$, that

$$\begin{aligned}\|(R^i_h)^*\tilde{w}\|_{H^1(\mathbb{R}^d)} &\lesssim \frac{1}{\mathbf{T}_{\min}}\Big(|\mathbf{T}|_{\mathcal{C}^t(\overline{\mathbb{R}^d_+},\mathbb{R}^{d\times d})}\,|w|_{H^1(\mathbb{R}^d_+)} + \|F\|_{H^{s-1}(\mathbb{R}^d_+)}\Big) + \|w\|_{H^1(\mathbb{R}^d_+)}\\ &=: \mathcal{B}(w).\end{aligned}$$

(Note that we added $|w|_{H^1(\mathbb{R}^d_+)}$ to the bound to simplify the notation later.) Hence, by the same token as in the previous section, we get

$$\int_{\mathbb{R}^d} (1+|\xi|^2)(|\xi_1|^2 + ... + |\xi_{d-1}|^2)^s |\widehat{\tilde{w}}(\xi)|^2 d\xi \lesssim \mathcal{B}(w)^2. \tag{A.2}$$

In particular, this implies that, for $1 \leq i \leq d$ and $1 \leq j \leq d-1$, we have

$$\begin{aligned}\int_{\mathbb{R}^d} \left| \widehat{\frac{\partial^2 \tilde{w}}{\partial x_i \partial x_j}}(\xi) \right|^2 (1+|\xi|^2)^{s-1} d\xi &= \int_{\mathbb{R}^d} |\xi_i|^2 |\xi_j|^2 |\widehat{\tilde{w}}(\xi)|^2 (1+|\xi|^2)^{s-1} d\xi \\ &\lesssim \mathcal{B}(w)^2,\end{aligned}$$

which means that $\frac{\partial^2 \tilde{w}}{\partial x_i \partial x_j} \in H^{s-1}(\mathbb{R}^d)$ and $\left\| \frac{\partial^2 \tilde{w}}{\partial x_i \partial x_j} \right\|_{H^{s-1}(\mathbb{R}^d)} \lesssim \mathcal{B}(w)$. In particular, for all $(i,j) \neq (d,d)$, this further implies that $\frac{\partial^2 w}{\partial x_i \partial x_j} \in H^{s-1}(\mathbb{R}^d_+)$ and that $\left\| \frac{\partial^2 w}{\partial x_i \partial x_j} \right\|_{H^{s-1}(\mathbb{R}^d_+)} \lesssim \mathcal{B}(w)$. Using Lemma 2.5 we deduce that $\frac{\partial w}{\partial x_j} \in H^s(\mathbb{R}^d_+)$ and that

$$\left\| \frac{\partial w}{\partial x_j} \right\|_{H^s(\mathbb{R}^d_+)} \lesssim \mathcal{B}(w), \qquad \text{for all } \ 1 \leq j \leq d-1. \tag{A.3}$$

It remains to bound $\left\| \frac{\partial w}{\partial x_d} \right\|_{H^s(\mathbb{R}^d_+)}$, which is rather technical. To achieve it we will use the PDE (2.7), Lemma 2.4 and the following result.

Lemma A.1. *For almost all $x_d \in \mathbb{R}$, we have $\frac{\partial \tilde{w}}{\partial x_d}(., x_d) \in H^s(\mathbb{R}^{d-1})$ and*

$$\int_{\mathbb{R}} \left\| \frac{\partial \tilde{w}}{\partial x_d}(., x_d) \right\|^2_{H^s(\mathbb{R}^{d-1})} \mathrm{d}x_d = \int_{\mathbb{R}^d} (1+|\xi'|^2)^s |\xi_d|^2 \left| \widehat{\tilde{w}}(\xi) \right|^2 d\xi \lesssim \mathcal{B}(w)^2.$$

Proof. This follows from Fubini's theorem and Plancherel's formula, together with (A.2). □

From this we deduce that $\frac{\partial w}{\partial x_d}(., x_d) \in H^s(\mathbb{R}^{d-1})$, for almost all $x_d \in \mathbb{R}_+$, and that

$$\int_{\mathbb{R}_+} \left\| \frac{\partial w}{\partial x_d}(., x_d) \right\|^2_{H^s(\mathbb{R}^{d-1})} \mathrm{d}x_d \lesssim \mathcal{B}(w)^2.$$

Let $1 \leq i \leq d-1$. Using Lemma 2.4 we deduce that $\mathbf{T}_{id} \frac{\partial w}{\partial x_d}(., x_d) \in H^s(\mathbb{R}^{d-1})$, for almost all $x_d \in \mathbb{R}_+$, and that

$$\begin{aligned}\left\| \left(\mathbf{T}_{id} \frac{\partial w}{\partial x_d} \right)(., x_d) \right\|_{H^s(\mathbb{R}^{d-1})} &\lesssim |\mathbf{T}_{id}(., x_d)|_{\mathcal{C}^t(\mathbb{R}^{d-1})} \left\| \frac{\partial w}{\partial x_d}(., x_d) \right\|_{L^2(\mathbb{R}^{d-1})} \\ &+ \|\mathbf{T}_{id}(., x_d)\|_{\mathcal{C}^0(\mathbb{R}^{d-1})} \left\| \frac{\partial w}{\partial x_d}(., x_d) \right\|_{H^s(\mathbb{R}^{d-1})}.\end{aligned}$$

Therefore, since by definition $\|\mathbf{T}_{id}\|_{\mathcal{C}^0(\mathbb{R}^d)} \leq \mathbf{T}_{\max}$, we get

$$\begin{aligned}\int_{\mathbb{R}_+} \left\|\mathbf{T}_{id}\frac{\partial w}{\partial x_d}\right\|^2_{H^s(\mathbb{R}^{d-1})} \,\mathrm{d}x_d &\lesssim |\mathbf{T}_{id}|^2_{\mathcal{C}^t(\overline{\mathbb{R}^d_+})}\,|w|^2_{H^1(\mathbb{R}^d_+)} + \mathbf{T}^2_{\max}\,\mathcal{B}(w)^2 \\ &\lesssim \mathbf{T}^2_{\max}\,\mathcal{B}(w)^2.\end{aligned}$$

Since $\frac{\partial}{\partial x_i}$ is linear continuous from $H^{1-s}(\mathbb{R}^{d-1})$ to $H^{-s}(\mathbb{R}^{d-1})$ (cf. [44, Remark 6.3.14(b)]) we can deduce from this that $\frac{\partial}{\partial x_i}\left(\mathbf{T}_{id}\frac{\partial w}{\partial x_d}\right) \in H^{s-1}(\mathbb{R}^d_+)$ and that

$$\left\|\frac{\partial}{\partial x_i}\left(\mathbf{T}_{id}\frac{\partial w}{\partial x_d}\right)\right\|_{H^{s-1}(\mathbb{R}^d_+)} \lesssim \mathbf{T}_{\max}\,\mathcal{B}(w), \quad \text{for all } 1\leq i\leq d-1\,. \quad (A.4)$$

To see this take $\varphi \in \mathcal{D}(\mathbb{R}^d_+)$. Then

$$\begin{aligned}&\left|\left\langle \frac{\partial}{\partial x_i}\left(\mathbf{T}_{id}\frac{\partial w}{\partial x_d}\right), \varphi\right\rangle_{\mathcal{D}'(\mathbb{R}^d_+),\mathcal{D}(\mathbb{R}^d_+)}\right| = \left\|\mathbf{T}_{id}\frac{\partial w}{\partial x_d}(x',x_d)\,\frac{\partial \varphi}{\partial x_i}(x',x_d)\right\|_{L^2(\mathbb{R}_+,H^s(\mathbb{R}^{d-1}))} \\ &\leq \left\|\mathbf{T}_{id}\frac{\partial w}{\partial x_d}\right\|_{L^2(\mathbb{R}_+,H^s(\mathbb{R}^{d-1}))}\left\|\frac{\partial \varphi}{\partial x_i}(x',x_d)\right\|_{L^2(\mathbb{R}_+,H^{-s}(\mathbb{R}^{d-1}))} \\ &\leq \left\|\mathbf{T}_{id}\frac{\partial w}{\partial x_d}\right\|_{L^2(\mathbb{R}_+,H^s(\mathbb{R}^{d-1}))}\|\varphi\|_{L^2(\mathbb{R}_+,H^{1-s}(\mathbb{R}^{d-1}))}.\end{aligned}$$

Using $(A.3)$ and Lemma 2.4, we deduce in a similar way that $\frac{\partial}{\partial x_i}\left(\mathbf{T}_{ij}\frac{\partial w}{\partial x_j}\right) \in H^{s-1}(\mathbb{R}^d_+)$ and that for all $1\leq i\leq d$ and $1\leq j\leq d-1$,

$$\left\|\frac{\partial}{\partial x_i}\left(\mathbf{T}_{ij}\frac{\partial w}{\partial x_j}\right)\right\|_{H^{s-1}(\mathbb{R}^d_+)} \lesssim \mathbf{T}_{\max}\,\mathcal{B}(w). \quad (A.5)$$

We can now use the PDE (2.7) to get a similar bound for $(i,j)=(d,d)$. Since $F\in H^{s-1}(\mathbb{R}^d_+)$, it follows from $(A.4)$ and $(A.5)$ that $\frac{\partial}{\partial x_d}\left(\mathbf{T}_{dd}\frac{\partial w}{\partial x_d}\right) \in H^{s-1}(\mathbb{R}^d_+)$ and that

$$\left\|\frac{\partial}{\partial x_d}\left(\mathbf{T}_{dd}\frac{\partial w}{\partial x_d}\right)\right\|_{H^{s-1}(\mathbb{R}^d_+)} \lesssim \mathbf{T}_{\max}\,\mathcal{B}(w) \;+\; \|F\|_{H^{s-1}(\mathbb{R}^d_+)} \lesssim \mathbf{T}_{\max}\,\mathcal{B}(w)\,.$$

Analogously to $(A.4)$ we can prove that

$$\left\| \frac{\partial}{\partial x_i} \left(\mathbf{T}_{dd} \frac{\partial w}{\partial x_d} \right) \right\|_{H^{s-1}(\mathbb{R}^d_+)} \lesssim \mathbf{T}_{\max} \mathcal{B}(w), \qquad \text{for all } 1 \leq i \leq d-1.$$

Hence, we can finally apply Lemma 2.5 to get that $\mathbf{T}_{dd} \frac{\partial w}{\partial x_d} \in H^s(\mathbb{R}^d_+)$ and that

$$\begin{aligned} \left\| \mathbf{T}_{dd} \frac{\partial w}{\partial x_d} \right\|^2_{H^s(\mathbb{R}^d_+)} &\lesssim \sum_{i=1}^{d} \left\| \frac{\partial}{\partial x_i} \left(\mathbf{T}_{dd} \frac{\partial w}{\partial x_d} \right) \right\|^2_{H^{s-1}(\mathbb{R}^d_+)} + \left\| \mathbf{T}_{dd} \frac{\partial w}{\partial x_d} \right\|^2_{L^2(\mathbb{R}^d_+)} \\ &\lesssim \mathbf{T}^2_{\max} \mathcal{B}(w)^2 . \end{aligned}$$

By applying Lemma 2.4 again, this time with $b := 1/\mathbf{T}_{dd}$ and $v := \mathbf{T}_{dd} \frac{\partial w}{\partial x_d}$, we deduce that $\frac{\partial w}{\partial x_d} \in H^s(\mathbb{R}^d_+)$ and that

$$\begin{aligned} \left\| \frac{\partial w}{\partial x_d} \right\|_{H^s(\mathbb{R}^d_+)} &\lesssim \left| \frac{1}{\mathbf{T}_{dd}} \right|_{\mathcal{C}^t(\overline{\mathbb{R}^d_+})} \left\| \mathbf{T}_{dd} \frac{\partial w}{\partial x_d} \right\|_{L^2(\mathbb{R}^d_+)} + \frac{1}{\mathbf{T}_{\min}} \left\| \mathbf{T}_{dd} \frac{\partial w}{\partial x_d} \right\|_{H^s(\mathbb{R}^d_+)} \\ &\lesssim \frac{|\mathbf{T}_{dd}|_{\mathcal{C}^t(\overline{\mathbb{R}^d_+})}}{\mathbf{T}^2_{\min}} \mathbf{T}_{\max} \, |w|_{H^1(\mathbb{R}^d_+)} + \frac{1}{\mathbf{T}_{\min}} \left\| \mathbf{T}_{dd} \frac{\partial w}{\partial x_d} \right\|_{H^s(\mathbb{R}^d_+)} \\ &\lesssim \frac{\mathbf{T}_{\max}}{\mathbf{T}_{\min}} \mathcal{B}(w) . \end{aligned}$$

To finish the proof we use this bound together with $(A.3)$ and apply once more Lemma 2.5 to show that $w \in H^{1+s}(\mathbb{R}^d_+)$ and

$$\begin{aligned} \|w\|_{H^{1+s}(\mathbb{R}^d_+)} \lesssim \frac{\mathbf{T}_{\max}}{\mathbf{T}^2_{\min}} \Big(|\mathbf{T}|_{\mathcal{C}^t(\overline{\mathbb{R}^d_+}, \mathbb{R}^{d \times d})} \, |w|_{H^1(\mathbb{R}^d_+)} + \|F\|_{H^{s-1}(\mathbb{R}^d_+)} \Big) \\ + \frac{\mathbf{T}_{\max}}{\mathbf{T}_{\min}} \|w\|_{H^1(\mathbb{R}^d_+)}. \end{aligned}$$

□

A.3 Step 3 – The Case D Bounded

We can now prove Theorem 2.7 using Lemmas 2.8 and 2.9 in two successive steps. We recall that D was supposed to be $\mathcal{C}^2$. Let $(D_i)_{0 \leq i \leq p}$ be a covering of D such that the $(D_i)_{0 \leq i \leq p}$ are open and bounded, $\overline{D} \subset \cup_{i=0}^{p} D_i$, $\cup_{i=1}^{p} (D_i \cap \partial D) = \partial D$, $\overline{D_0} \subset D$.

Let $(\chi_i)_{0\le i\le p}$ be a partition of unity subordinate to this cover, i.e. we have $\chi_i \in \mathcal{C}^\infty(\mathbb{R}^d, \mathbb{R}_+)$ with compact support $\mathrm{supp}(\chi_i) \subset D_i$, such that $\sum_{i=0}^p \chi_i = 1$ on $\overline{D}$. We denote by u the solution of (2.4) and split it into $u = \sum_{i=0}^p u_i$, with $u_i = u\chi_i$. We treat now separately u_0 and then u_i, $1 \le i \le p$, using Lemma 2.8 and 2.9, respectively.

Lemma A.2. *u_0 belongs to $H^{1+s}(D)$ and*

$$\|u_0\|_{H^{1+s}(D)} \lesssim \frac{\|\mathbf{A}\|_{\mathcal{C}^t(\overline{D},\mathbb{R}^{d\times d})}}{\mathbf{A}_{\min}^2}\|f\|_{H^{s-1}(D)}\,.$$

Proof. Since $\mathrm{supp}(u_0) \subset D_0$, we have that $u_0 \in H_0^1(D)$ and it is the weak solution of the new equation $-\mathrm{div}(\mathbf{A}\nabla u_0) = F$ on D, where

$$F \;:=\; f\chi_0 + \mathbf{A}\nabla u\cdot\nabla\chi_0 + \mathrm{div}(u\mathbf{A}\nabla\chi_0) \quad \text{on } D.$$

To apply Lemma 2.8 we will now extend all terms to $\mathbb{R}^d$, but continue to denote them the same. The terms u_0 and $f\chi_0 + \mathbf{A}\nabla u\cdot\nabla\chi_0$ can both be extended by 0. Their extensions will belong to $H^1(\mathbb{R}^d)$ and $H^{s-1}(\mathbb{R}^d)$, respectively. It follows from Lemma A.2 in [12] that every element of $u\mathbf{A} \in H^s(D)$ and so if we continue every element of $u\mathbf{A}\nabla\chi_0$ by 0 on $\mathbb{R}^d$, the extension belongs to $H^s(\mathbb{R}^d)$, since $\mathrm{supp}(\chi_0)$ is compact in D. Using the fact that div is linear and continuous from $H^s(\mathbb{R}^d)$ to $H^{s-1}(\mathbb{R}^d)$ (cf. [44, Remark 6.3.14(b)]), we can deduce that the divergence of the extension of $u\mathbf{A}\nabla\chi_0$ is in $H^{s-1}(\mathbb{R}^d)$, leading to an extension of F on $\mathbb{R}^d$, which belongs to $H^{s-1}(\mathbb{R}^d)$.

Let $\psi \in \mathcal{C}^\infty(\mathbb{R}^d, [0,1])$ such that $\psi = 0$ on D_0 and $\psi = 1$ on $\tilde{D}^c$, where $\tilde{D}$ is an open set such that $\overline{D}_0 \subset \tilde{D}$ and $\overline{\tilde{D}} \subset D$. We use the following extension of $\mathbf{A}$ from D_0 to all of $\mathbb{R}^d$:

$$\overline{\mathbf{A}}(x) := \begin{cases} \mathbf{A}(x)(1-\psi(x)) + \mathbf{A}_{\min}\psi(x)\, I_d, & \text{if } \; x \in D, \\ \mathbf{A}_{\min}\psi(x)\, I_d, & \text{otherwise.} \end{cases}$$

This implies that $\overline{\mathbf{A}} \in \mathcal{C}^t(\mathbb{R}^d, \mathbb{R}^{d\times d})$, with $\|\overline{\mathbf{A}}\|_{\mathcal{C}^t(\overline{D},\mathbb{R}^{d\times d})} \lesssim \|\mathbf{A}\|_{\mathcal{C}^t(\overline{D},\mathbb{R}^{d\times d})}$, and for any $\xi \in \mathbb{R}^d$, $\overline{\mathbf{A}}(x)\xi.\xi \gtrsim \mathbf{A}_{\min}|\xi|^2$.

Using these extensions, we have that $-\mathrm{div}(\overline{\mathbf{A}}\nabla u_0) = F$ in $\mathcal{D}'(\mathbb{R}^d)$. Indeed,

for any $v \in \mathcal{D}(\mathbb{R}^d)$,

$$\int_{\mathbb{R}^d} \overline{\mathbf{A}}(x)\nabla u_0(x)\nabla v(x)\,\mathrm{d}x \;=\; \int_D \mathbf{A}(x)\nabla u_0(x)\nabla v(x)\,\mathrm{d}x, \text{ for any } \; v \in \mathcal{D}(\mathbb{R}^d),$$

since $\mathrm{supp}(u_0)$ is included in the open bounded set D_0, which implies that $\nabla u_0 = 0$ on D_0^c and $\mathbf{A} = \overline{\mathbf{A}}$ on D_0. Since $u \in H_0^1(D)$, we have by Poincaré's inequality that $\|u\|_{L^2(D)} \lesssim |u|_{H^1(D)}$. Therefore it follows from the Lemma 2.1 that

$$|u_0|_{H^1(\mathbb{R}^d)} \;\leq\; |u|_{H^1(D)}\|\chi_0\|_\infty + \|u\|_{L^2(D)}\|\nabla\chi_0\|_\infty \;\lesssim\; \frac{\|f\|_{H^{s-1}(D)}}{\mathbf{A}_{\min}}.$$

By the triangle inequality, we have

$$\|F\|_{H^{s-1}(\mathbb{R}^d)} \leq \|f\chi_0\|_{H^{s-1}(\mathbb{R}^d)} + \|\mathbf{A}\nabla u\cdot\nabla\chi_0\|_{H^{s-1}(\mathbb{R}^d)} + \|\mathrm{div}(u\mathbf{A}\nabla\chi_0)\|_{H^{s-1}(\mathbb{R}^d)}$$

Since $\chi_0 \in \mathcal{C}^\infty(\mathbb{R}^d)$ and $|\chi_0| \leq 1$ in D, we have $\|f\chi_0\|_{H^{s-1}(\mathbb{R}^d)} \leq \|f\|_{H^{s-1}(D)}$. Using the definition of $\mathbf{A}_{\max}$ and the fact that $|A_{i,j}(x)| \leq \|A(x)\|_{d\times d}$, we have

$$\begin{aligned}\|\mathbf{A}\nabla u\cdot\nabla\chi_0\|_{H^{s-1}(\mathbb{R}^d)} \leq \|\mathbf{A}\nabla u\cdot\nabla\chi_0\|_{L^2(\mathbb{R}^d)} &= \|\sum_{i,j=1}^d \mathbf{A}_{i,j}\frac{\partial u}{\partial x_j}\frac{\partial\chi_0}{\partial x_i}\|_{L^2(\mathbb{R}^d)}\\ &\lesssim \mathbf{A}_{\max}|u|_{H^1(D)}.\end{aligned}$$

Finally, using Lemma 2.5, the linearity and continuity of div from $H^s(\mathbb{R}^d)$ to $H^{1-s}(\mathbb{R}^d)$ and $|\mathbf{A}_{i,j}|_{\mathcal{C}^t(\overline{D})} \leq |\mathbf{A}|_{\mathcal{C}^t(\overline{D},\mathbb{R}^{d\times d})}$, we further get

$$\begin{aligned}\|\mathrm{div}(u\mathbf{A}\nabla\chi_0)\|_{H^{s-1}(\mathbb{R}^d)} &= \|\sum_{i,j=1}^d \frac{\partial}{\partial x_i}\big(u\mathbf{A}_{i,j}\frac{\partial\chi_0}{x_j}\big)\|_{H^{s-1}(\mathbb{R}^d)}\\ &\lesssim \|\sum_{i,j=1}^d \big(u\mathbf{A}_{i,j}\frac{\partial\chi_0}{x_j}\big)\|_{H^s(\mathbb{R}^d)}\\ &\lesssim |\mathbf{A}|_{\mathcal{C}^t(\overline{D},\mathbb{R}^{d\times d})}\|u\|_{L^2(D)} + \mathbf{A}_{\max}\|u\|_{H^s(D)}\end{aligned}$$

Putting these estimates together, we have

$$\|F\|_{H^{s-1}(\mathbb{R}^d)} \lesssim \frac{\|\mathbf{A}\|_{\mathcal{C}^t(\overline{D},\mathbb{R}^{d\times d})}}{\mathbf{A}_{\min}}\|f\|_{H^{s-1}(D)}.$$

We can now apply Lemma 2.8 with $\mathbf{T} = \overline{\mathbf{A}}$ and $w = u_0$ to show that $u_0 \in H^{1+s}(\mathbb{R}^d)$ and

$$\begin{aligned}\|u_0\|_{H^{1+s}(\mathbb{R}^d)} &\lesssim \frac{1}{\mathbf{A}_{\min}}\Big(|\overline{\mathbf{A}}|_{\mathcal{C}^t(\overline{D},\mathbb{R}^{d\times d})}|u_0|_{H^1(\mathbb{R}^d)} + \|F\|_{H^{s-1}(\mathbb{R}^d)}\Big) + \|u_0\|_{H^1(\mathbb{R}^d)} \\ &\lesssim \frac{\|\mathbf{A}\|_{\mathcal{C}^t(\overline{D},\mathbb{R}^{d\times d})}}{\mathbf{A}_{\min}^2}\,\|f\|_{H^{s-1}(D)}\,.\end{aligned}$$

The hidden constant depends on the choices of χ_0 and ψ and on the constant in Poincaré's inequality, which depends on the shape and size of D, but not on $\mathbf{A}$. □

Let us now treat the case of u_i, $1 \le i \le p$.

Lemma A.3. *For $1 \le i \le p$, $u_i \in H^{1+s}(D)$ and*

$$\|u_i\|_{H^{1+s}(D)} \lesssim \frac{\mathbf{A}_{\max}\|\mathbf{A}\|_{\mathcal{C}^t(\overline{D},\mathbb{R}^{d\times d})}}{\mathbf{A}_{\min}^3}\|f\|_{H^{s-1}(D)}.$$

Proof. Similarly to the proof of the previous Lemma, $u_i \in H_0^1(D \cap D_i)$ is the weak solution of a new problem $-\mathrm{div}(\mathbf{A}\nabla u_i) = g_i$ in $\mathcal{D}'(D \cap D_i)$ with

$$g_i := f\chi_i + \mathbf{A}\nabla u \cdot \nabla\chi_i + \mathrm{div}(u\mathbf{A}\nabla\chi_i)$$

As in Lemma A.2, we can establish that $g_i \in H^{s-1}(D \cap D_i)$ and

$$\|g_i\|_{H^{s-1}(D\cap D_i)} \lesssim \frac{\|\mathbf{A}\|_{\mathcal{C}^t(\overline{D},\mathbb{R}^{d\times d})}}{\mathbf{A}_{\min}}\,\|f\|_{H^{s-1}(D)}.$$

Now let $Q = \{(y', y_d) \in \mathbb{R}^{d-1} \times \mathbb{R} : |y'| < 1 \text{ and } |y_d| < 1\}$, $Q_0 = \{(y', y_d) \in \mathbb{R}^{d-1} \times \{0\} |\ \|y'\| < 1\}$ and $Q_+ = Q \cap \mathbb{R}^d_+$. For $1 \le i \le p$, let α_i be a bijection from D_i to Q such that $\alpha_i \in \mathcal{C}^2(\overline{D}_i)$, $\alpha_i^{-1} \in \mathcal{C}^2(\overline{Q})$, $\alpha_i(D_i \cap D) = Q_+$ and $\alpha_i(D_i \cap \partial D) = Q_0$.

For all $y \in Q_+$, we define $w_i(y) := u_i(\alpha_i^{-1}(y)) \in H_0^1(Q_+)$ with $\nabla w_i(y) = J_i^{-T}(y)\nabla u_i(\alpha_i^{-1}(y))$, where $J_i(y) := D\alpha_i(\alpha_i^{-1}(y)))$ is the Jacobian of α_i. Furthermore, for $x \in D_i \cap D$ and $\varphi \in H_0^1(Q_+)$, we define $v(x) := \varphi(\alpha_i(x))$. Then

$v \in H_0^1(D_i \cap D)$ and $\nabla v(\alpha_i^{-1}(y)) = J_i^T(y)\nabla\varphi(y)$, for all $y \in Q_+$, so that

$$\begin{aligned}
&\int_{D_i\cap D} \mathbf{A}(x)\nabla u_i(x)\cdot\nabla v(x)\,\mathrm{d}x \\
&= \int_{Q_+} \mathbf{A}(\alpha_i^{-1}(y))\,\nabla u_i(\alpha_i^{-1}(y))\cdot\nabla v(\alpha_i^{-1}(y))\,|\mathrm{det}J_i(y))|^{-1}\,\mathrm{d}y \\
&= \int_{Q_+} \mathbf{T}_i(y)\,\nabla w_i(y)\cdot\nabla\varphi(y)\;\mathrm{d}y\,,
\end{aligned}$$

where

$$\mathbf{T}_i(y) := |\mathrm{det}J_i(y))|^{-1}\,J_i(y)\,\mathbf{A}(\alpha_i^{-1}(y))\,J_i^T(y) \in S_d(\mathbb{R}).$$

We define $F_i \in H^{s-1}(Q_+)$ by

$$\langle F_i,\varphi\rangle_{H^{s-1}(Q_+),H_0^{1-s}(Q_+)} \;:=\; \langle g_i,\varphi\circ\alpha_i\rangle_{H^{s-1}(D_i\cap D),H_0^{1-s}(D_i\cap D)},$$

for all $\varphi \in H_0^{1-s}(Q_+)$. Indeed, since we assumed that α_i and α_i^{-1} are in $\mathcal{C}^2$, we have $\varphi\circ\alpha_i \in H_0^{1-s}(D_i\cap D)$ and moreover $\|\varphi\circ\alpha_i\|_{H^{1-s}(D_i\cap D)} \lesssim \|\varphi\|_{H^{1-s}(Q_+)}$ (cf. [44, Theorems 6.2.17 and 6.2.25(g)]), which implies that $F_i \in H^{s-1}(Q_+)$ and

$$\|F_i\|_{H^{s-1}(Q_+)} \;\lesssim\; \|g_i\|_{H^{s-1}(D\cap D_i)}.$$

We finally get that $v_i \in H_0^1(Q_+)$ solves

$$\int_{Q_+} \mathbf{T}_i\nabla v_i\cdot\nabla\varphi\;\mathrm{d}y \;=\; \langle F_i,\varphi\rangle_{H^{s-1}(Q_+),H_0^{1-s}(Q_+)} \qquad \text{for all}\;\; \varphi\in H_0^1(Q_+).$$

In order to apply Lemma 2.9 we check first that $\mathbf{T}_i \in \mathcal{C}^t(\overline{Q_+},\mathbb{R}^{d\times d})$ and that it is coercive, and then define an extension of $\mathbf{T}_i$ to $\mathbb{R}^d_+$. Recalling that α_i is a $\mathcal{C}^2$–diffeomorphism from $D_i\cap D$ to Q_+, with $\alpha_i^{-1}\in\mathcal{C}^2(\overline{Q_+})$, we have for any $y\in Q_+$ and $\xi\in\mathbb{R}^d$:

- Coercivity: Using the compatibility of $|\cdot|$ and $\|\cdot\|_{d\times d}$, we have

$$\begin{aligned}
\mathbf{T}_i(y)\xi\cdot\xi \;&=\; |\mathrm{det}J_i(y)|^{-1}\,J_i(y)\,\mathbf{A}(\alpha_i^{-1}(y))\,J_i^T(y)\xi\cdot\xi \\
&=\; |\mathrm{det}J_i(y)|^{-1}\,\mathbf{A}(\alpha_i^{-1}(y))\,J_i^T(y)\xi\cdot J_i^t(y)\xi \\
&\gtrsim\; |\mathrm{det}J_i(y)|^{-1}\mathbf{A}_{\min}|J_i^T(y)\xi|^2 \\
&\gtrsim\; \mathbf{A}_{\min}|\xi|^2.
\end{aligned}$$

Hence $\mathbf{T}_{\min} \gtrsim \mathbf{A}_{\min}$.

- Boundedness: Using the sub–multiplicativity of $\|\cdot\|_{d\times d}$, we have

$$\begin{aligned}\mathbf{T}_{\max} &:= \|\mathbf{T}_i\|_{\mathcal{C}^0(\overline{Q_+},\mathbb{R}^{d\times d})}\\ &= \sup_{y\in\overline{Q}_+} \||\det J_i(y)|^{-1}\, J_i(y)\, \mathbf{A}(\alpha_i^{-1}(y))\, J_i^T(y)\|_{d\times d}\\ &\le \sup_{y\in\overline{Q}_+} \left(|\det J_i(y)|^{-1}\, \|J_i(y)\|_{d\times d}\, \|\mathbf{A}(\alpha_i^{-1}(y))\|_{d\times d}\, \|J_i^T(y)\|_{d\times d}\right)\\ &\lesssim \mathbf{A}_{\max}.\end{aligned}$$

- Regularity: Since $\mathbf{A}\in\mathcal{C}^t(\overline{D},\mathbb{R}^{d\times d})$ and $\alpha_i\in\mathcal{C}^2(\overline{D}_i)$, $\mathbf{T}_i\in\mathcal{C}^t(\overline{Q_+},\mathbb{R}^{d\times d})$. Using the fact that $\|MN\|_{\mathcal{C}^t(\overline{Q_+},\mathbb{R}^{d\times d})} \le \|M\|_{\mathcal{C}^t(\overline{Q_+},\mathbb{R}^{d\times d})}\|N\|_{\mathcal{C}^t(\overline{Q_+},\mathbb{R}^{d\times d})}$, for any $M,N\in\mathbb{R}^{d\times d}$, we further have

$$\|\mathbf{T}_i\|_{\mathcal{C}^t(\overline{Q_+},\mathbb{R}^{d\times d})} = \||\det J_i|^{-1} J_i(\mathbf{A}\circ\alpha_i^{-1})J_i^T\|_{\mathcal{C}^t(\overline{Q_+},\mathbb{R}^{d\times d})} \lesssim \|\mathbf{A}\|_{\mathcal{C}^t(\overline{D},\mathbb{R}^{d\times d})}.$$

We now extend $\mathbf{T}_i$ to $\mathbb{R}^d_+$. Since we assumed that $\mathrm{supp}(\chi_i)$ is compact in D_i, we can choose Q_i and $\tilde{Q}_i$ such that $\mathrm{supp}(v_i)\subset Q_i\subset\overline{Q_i}\subset\widetilde{Q}_i\subset\overline{\widetilde{Q}_i}\subset Q$ and consider $\psi\in\mathcal{C}^\infty(\mathbb{R}^d,[0,1])$ such that $\psi=0$ on Q_i and $\psi=1$ on $\overline{\widetilde{Q}_i}^c$. We define the extension $\overline{\mathbf{T}_i}$ of $\mathbf{T}_i$ on $\mathbb{R}^d_+$ by

$$\overline{\mathbf{T}_i}(x) := \begin{cases} \mathbf{T}_i(x)(1-\psi(x)) + \mathbf{A}_{\min}\psi(x) I_d & \text{if } x\in Q_+\\ \mathbf{A}_{\min}\psi(x) I_d & \text{if } x\in Q_+^c\end{cases}$$

Analogously to the case of $\overline{\mathbf{A}}$ in Lemma A.2, we can deduce that, for any $\xi\in\mathbb{R}^d$,

$$\overline{\mathbf{T}_i}(y)\xi\cdot\xi \gtrsim \mathbf{A}_{\min}|\xi|^2,\ \ \overline{\mathbf{T}}_{\max}\lesssim\mathbf{A}_{\max}\ \text{ and }\ \|\overline{\mathbf{T}_i}\|_{\mathcal{C}^t(\overline{\mathbb{R}^d_+},\mathbb{R}^{d\times d})}\lesssim\|\mathbf{A}\|_{\mathcal{C}^t(\overline{D},\mathbb{R}^{d\times d})}. \quad (A.6)$$

We now define an extension of F_i on $\mathbb{R}^d_+$. Note again that we can choose an open set G_i such that $\mathrm{supp}(F_i)\subset G_i\subset\overline{G}_i\subset Q$ and extend F_i to all of $\mathbb{R}^d_+$ such that $\|F_i\|_{H^{s-1}(\mathbb{R}^d_+)}\lesssim\|F_i\|_{H^{s-1}(Q_+)}$.

Finally we continue v_i by 0 on $\mathbb{R}^d_+$, which yields $v_i\in H^1_0(\mathbb{R}^d_+)$. Moreover, since $\overline{\mathbf{T}_i}=\mathbf{T}_i$ on $\mathrm{supp}(v_i)\subset Q_i$, v_i is then the weak solution on $\mathbb{R}^d_+$ of

$$-\mathrm{div}(\overline{\mathbf{T}_i}(x)\nabla v_i(x)) = F_i(x),$$

which enables us to apply Lemma 2.9 and to obtain that $v_i \in H^{1+s}(\mathbb{R}^d_+)$ and

$$\begin{aligned}
&\|v_i\|_{H^{1+s}(\mathbb{R}^d_+)} \\
&\lesssim \frac{\mathbf{A}_{\max}}{\mathbf{A}^2_{\min}} \Big(|\overline{\mathbf{T}_i}|_{\mathcal{C}^t(\overline{\mathbb{R}^d_+}, S_d(\mathbb{R}))} \, |v_i|_{H^1(\mathbb{R}^d_+)} + \|F_i\|_{H^{s-1}(\mathbb{R}^d_+)} \Big) \; + \; \frac{\mathbf{A}_{\max}}{\mathbf{A}_{\min}} \|v_i\|_{H^1(\mathbb{R}^d_+)} \, .
\end{aligned}$$

Recalling that $u_i(x) = v_i(\alpha_i(x))$ for any $x \in D \cap D_i$ and using the bounds in (A.6), as well as the transformation theorem [44, Theorem 6.2.17], we finally get

$$\begin{aligned}
&\|u_i\|_{H^{1+s}(D)} \\
&\lesssim \frac{\mathbf{A}_{\max}}{\mathbf{A}^2_{\min}} \Big(\|\mathbf{A}\|_{\mathcal{C}^t(\overline{D}, S_d(\mathbb{R}))} \, \|u\|_{H^1(D \cap D_i)} + \|g_i\|_{H^{s-1}(D \cap D_i)} \Big) \; + \; \frac{\mathbf{A}_{\max}}{\mathbf{A}_{\min}} \|u\|_{H^1(D)} \\
&\lesssim \frac{\mathbf{A}_{\max} \, \|\mathbf{A}\|_{\mathcal{C}^t(\overline{D}, S_d(\mathbb{R}))}}{\mathbf{A}^3_{\min}} \, \|f\|_{H^{s-1}(D)} \, .
\end{aligned}$$

□

The result in Theorem 2.7 follows directly from Lemmas A.2 and A.3, if we recall that $u = \sum_{i=0}^{m} u_i$.

Bibliography

[1] R.A. Adams. *Sobolev Spaces*. Academic Press, 1975.

[2] I. Babuška, F. Nobile, and R. Tempone. A stochastic collocation method for elliptic partial differential equations with random input data. *SIAM J. Numer. Anal.*, 45(3):1005–1034, 2007.

[3] J.W. Barrett and C.E. Elliott. Total flux estimates for a finite-element approximation of elliptic equations. *IMA J. Numer. Anal.*, 7:129–148, 1987.

[4] A. Barth, Ch. Schwab, and N. Zollinger. Multi–level Monte Carlo finite element method for elliptic PDE's with stochastic coefficients. *Numer. Math.*, 119(1):123–161, 2011.

[5] A. Brandt, M. Galun, and D. Ron. Optimal multigrid algorithms for calculating thermodynamic limits. *J. Stat. Phys.*, 74(1-2):313–348, 1994.

[6] A. Brandt and V. Ilyin. Multilevel Monte Carlo methods for studying large scale phenomena in fluids. *J. Mol. Liq.*, 105(2-3):245–248, 2003.

[7] S.C. Brenner and L.R. Scott. *The Mathematical Theory of Finite Element Methods*, volume 15 of *Texts in Applied Mathematics*. Springer, third edition, 2008.

[8] S.D. Buckeridge. *Numerical Solution of Weather and Climate Systems*. PhD thesis, University of Bath, 2010.

[9] J. Charrier. *Numerical analysis of partial differential equations with random coefficients, applications to hydrogeology*. PhD thesis, ENS Cachan, 2011.

[10] J. Charrier. Strong and weak error estimates for the solutions of elliptic partial differential equations with random coefficients. *SIAM J. Numer. Anal*, 50(1):216–246, 2012.

[11] J. Charrier and A. Debussche. Weak truncation error estimates for elliptic partial differential equations with lognormal coefficients. *Stochastic Partial Differential Equations: Analysis and Computations*, 1(1):63–93, 2013.

[12] 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.

[13] X. Chen, Z. Nashed, and L. Qi. Smoothing methods and semismooth methods for nondifferentiable operator equations. *SIAM J. Numer. Anal.*, 38(4):1200–1216, 2001.

[14] P. G. Ciarlet. *The Finite Element Method for Elliptic Problems*. North–Holland, 1978.

[15] 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.

[16] K.A. Cliffe, I.G. Graham, R. Scheichl, and L. Stals. Parallel computation of flow in heterogeneous media using mixed finite elements. *J. Comput. Phys.*, 164:258–282, 2000.

[17] M. Costabel, M. Dauge, and S. Nicaise. Singularities of Maxwell interface problems. *M2AN Math. Model. Numer. Anal.*, 33(3):627–649, 1999.

[18] S.L. Cotter, M. Dashti, and A.M. Stuart. Approximation of bayesian inverse problems. *SIAM J. Numer. Anal.*, 48:322–345, 2010.

[19] S.L. Cotter, M. Dashti, and A.M. Stuart. Variational data assimilation using targetted random walks. *Int. J. Numer. Meth. Fluids.*, 68:403–421, 2012.

[20] M. Dashti and A. Stuart. Uncertainty quantification and weak approximation of an elliptic inverse problem. *SIAM J. Numer. Anal.*, 49(6):2524–2542, 2011.

[21] G. de Marsily. *Quantitative Hydrogeology.* Academic Press, 1986.

[22] G. de Marsily, F. Delay, J. Goncalves, P. Renard, V. Teles, and S. Violette. Dealing with spatial heterogeneity. *Hydrogeol. J*, 13:161–183, 2005.

[23] P. Delhomme. Spatial variability and uncertainty in groundwater flow param- eters, a geostatistical approach. *Water Resourc. Res.*, pages 269–280, 1979.

[24] S. Dereich and F. Heidenreich. A multilevel Monte Carlo algorithm for Lévy-driven stochastic differential equations. *Stochastic Process. Appl.*, 121(7):1565–1587, 2011.

[25] C.R. Dietrich and G.N. 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.

[26] J. Douglas, T. Dupont, and M.F. Wheeler. A Galerkin procedure for approximating the flux on the boundary for elliptic and parabolic boundary value problems. *RAIRO Modèl. Math. Anal. Numèr.*, 2:47–59, 1974.

[27] P. Frauenfelder, Ch. Schwab, and R.A. Todor. Finite elements for elliptic problems with stochastic coefficients. *Comput. Methods Appl. Mech. Engrg.*, 194(2-5):205–228, 2005.

[28] J. Galvis and M. Sarkis. Approximating infinity-dimensional stochastic Darcy's equations without uniform ellipticity. *SIAM J. Numer. Anal.*, 47:3624–3651, 2009.

[29] A. Gelman and D.B. Rubin. Inference from iterative simulation using multiple sequences. *Statistical Sciences*, 7(4):457–511, 1992.

[30] R.G. Ghanem and P.D. Spanos. *Stochastic Finite Elements: A Spectral Approach.* Springer, New York, 1991.

[31] D. Gilbarg and N. S. Trudinger. *Elliptic Partial Differential Equations of Second Order.* Classics in Mathematics. Springer, Berlin, 2001. Reprint of the 1998 edition.

[32] M.B. Giles. *Improved multilevel Monte Carlo convergence using the Milstein scheme*, pages 343–358. Monte Carlo and Quasi-Monte Carlo methods 2006. Springer, 2007.

[33] M.B. Giles. Multilevel Monte Carlo path simulation. *Oper. Res.*, 256:981–986, 2008.

[34] M.B. Giles and C. Reisinger. Stochastic finite differences and multilevel Monte Carlo for a class of SPDEs in finance. *SIFIN*, 1(3):575–592, 2012.

[35] M.B. Giles and E. Süli. *Adjoint methods for PDEs: a posteriori error analysis and postprocessing by duality*, volume 11 of *Acta Numer.*, pages 145–236. Cambridge University Press, 2002.

[36] C. J. Gittelson. Stochastic Galerkin discretization of the log-normal isotropic diffusion problem. *Math. Models Methods Appl. Sci.*, 20(2):237–263, 2010.

[37] C.J. Gittelson, J. Könnö, Ch. Schwab, and R. Stenberg. The multilevel Monte Carlo finite element method for a stochastic Brinkman problem. *Numer. Math.*, published online March 2013.

[38] I.G. Graham, F.Y. Kuo, J.A Nichols, R. Scheichl, Ch. Schwab, and I.H. Sloan. Quasi-Monte Carlo finite element methods for elliptic PDEs with log–normal random coefficients. SAM Research Report 2013–14, ETH Zurich, 2013.

[39] I.G. Graham, F.Y. Kuo, D. Nuyens, R. Scheichl, and I.H. Sloan. Quasi-Monte Carlo methods for elliptic PDEs with random coefficients and applications. *J. Comput. Phys.*, 230(10):3668–3694, 2011.

[40] S. Graubner. Multi-level Monte Carlo Methoden für stochastische partielle Differentialgleichungen. Diploma thesis, TU Darmstadt, 2008.

[41] P. Grisvard. *Elliptic Problems in Non–smooth Domains*. Pitman, 1985.

[42] P. Grisvard. *Singularities in Boundary Value Problems*. Res. Notes Math. Springer, 1992.

[43] W. Hackbusch. On first and second order box schemes. *Computing*, 41:277–296, 1989.

[44] W. Hackbusch. *Elliptic Differential Equations*, volume 18 of *Spr. S. Comp.* Springer, 2010.

[45] M. Hairer, A.M. Stuart, and S.J. Vollmer. Spectral gaps for a Metropolis–Hastings Algorithm in Infinite Dimensions. Technical report, arxiv, 2011. Available at `http://arxiv.org/abs/1112.1392`.

[46] W.K. Hastings. Monte-Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

[47] S. Heinrich. *Multilevel Monte Carlo Methods*, volume 2179 of *Lecture notes in Comput. Sci.*, pages 3624–3651. Springer, 2001.

[48] V.H. Hoang, Ch. Schwab, and A.M. Stuart. Sparse MCMC GPC finite element methods for Bayesian inverse problems. Technical report, arxiv, 2012. Available at `http://arxiv.org/abs/1207.2411`.

[49] R.J. Hoeksema and P.K. Kitanidis. Analysis of the spatial structure of properties of selected aquifers. *Water Resour. Res.*, 21:536–572, 1985.

[50] H. Hoel, E. von Schwerin, A. Szepessy, and R. Tempone. Adaptive multilevel Monte Carlo simulation. In B. Engquist, O. Runborg, and Y.-H. R. Tsai, editors, *Numerical Analysis of Multiscale Computations*, volume 82 of *Lecture Notes in Comput. Sci.*, pages 217–234. Springer, 2012.

[51] T. Kato. *Perturbation Theory for Linear Operators*. Springer, 1966.

[52] C. Ketelsen, R. Scheichl, and A.L. Teckentrup. A hierarchical multilevel Markov chain Monte Carlo algorithm with applications to uncertainty quantification in subsurface flow. Submitted, 2013.

[53] P. Kloeden, A. Neuenkirch, and R. Pavani. Multilevel Monte Carlo for stochastic differential equations with additive fractional noise. *Ann. App. Probab.*, 189:255–276, 2011.

[54] O.P. Le Maître and O.M. Kino. *Spectral Methods for Uncertainty Quantification, With Applications to Fluid Dynamics*. Springer, 2010.

[55] N. Metropolis, A.W. Rosenbluth, M.N.Rosenbluth, A.H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21:1087, 1953.

[56] E. Di Nezza, G. Palatucci, and E. Valdinoci. Hitchhiker's guide to the fractional Sobolev spaces. *B. Sci. Math.*, 136:521–573, 2012.

[57] S. Nicaise and A.M. Sändig. General interface problems – I. *Math. Methods Appl. Sci.*, 17(6):395–429, 1994.

[58] M. Petzoldt. *Regularity and error estimators for elliptic problems with discontinuous coefficients*. PhD thesis, WIAS, 2001. Available at `http://webdoc.sub.gwdg.de/ebook/diss/2003/fu-berlin/2001/111/`.

[59] N.A. Pierce and M.B. Giles. Adjoint and defect error bounding and correction for functional estimates. *J. Comput. Phys.*, 200:769–794, 2004.

[60] G. Da Prato and J. Zabczyk. *Stochastic Equations in Infinite Dimensions*, volume 44 of *Encyclopedia Math. Appl.* Cambridge University Press, Cambridge, 1992.

[61] C. Robert and G. Casella. *Monte Carlo Statistical Methods*. Springer, 1999.

[62] D. Rudolf. *Explicit error bounds for Markov chain Monte Carlo*. PhD thesis, Friedrich–Schiller–Universität Jena, 2011. Available at `http://tarxiv.org/abs/1108.3201`.

[63] A. Schatz. A weak discrete maximum principle and stability of the finite element method in L_∞ on plane polygonal domains. I. *Math. Comput.*, 34(149):77–91, 1980.

[64] A.M. Stuart. *Inverse problems*, volume 19 of *Acta Numer.*, pages 451–559. Cambridge University Press, 2010.

[65] A. L. Teckentrup. Multilevel Monte Carlo methods for highly heterogeneous media. In *Proceedings of the Winter Simulation Conference 2012*, 2012. Available at `http://informs-sim.org`.

[66] 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.*, published online March 2013.

[67] D. Xiu. *Numerical Methods for Stochastic Computations, A Spectral Method Approach*. Princeton University Press, 2010.

[68] S. Zhang. Analysis of finite element domain embedding methods for curved domains using uniform grids. *SIAM J. Numer. Anal.*, 46:2843–2866, 2008.