A Hierarchical Multilevel Markov Chain Monte Carlo Algorithm with Applications to Uncertainty Quantification in Subsurface Flow

A Hierarchical Multilevel Markov Chain Monte Carlo Algorithm with Applications to Uncertainty Quantification in Subsurface Flow*

T.J. Dodwell[1], C. Ketelsen[2], R. Scheichl[3] and A.L. Teckentrup[4]

[1] Dept of Mechanical Engineering, University of Bath, Bath BA2 7AY, UK

[2] Dept of Applied Mathematics, 526 UCB, University of Colorado at Boulder, CO 80309-0526, USA

[3] Dept of Mathematical Sciences, University of Bath, Bath BA2 7AY, UK. Email: R.Scheichl@bath.ac.uk

[4] Mathematics Institute, Zeeman Building, University of Warwick, Coventry CV4 7AL, UK

Abstract

In this paper we address the problem of the prohibitively large computational cost of existing Markov chain Monte Carlo methods for large–scale applications with high dimensional parameter spaces, e.g. in uncertainty quantification in porous media flow. We propose a new multilevel Metropolis-Hastings algorithm, and give an abstract, problem dependent theorem on the cost of the new multilevel estimator based on a set of simple, verifiable assumptions. For a typical model problem in subsurface flow, we then provide a detailed analysis of these assumptions and show significant gains over the standard Metropolis-Hastings estimator. Numerical experiments confirm the analysis and demonstrate the effectiveness of the method with consistent reductions of more than an order of magnitude in the cost of the multilevel estimator over the standard Metropolis-Hastings algorithm for tolerances $\varepsilon < 10^{-2}$.

Keywords. Elliptic PDES with random coefficients, log-normal coefficients, finite element analysis, Bayesian approach, Metropolis-Hastings algorithm, multilevel Monte Carlo.

Mathematics Subject Classification (2000). 35R60, 62F15, 62M05, 65C05, 65C40, 65N30

1 Introduction

The parameters in mathematical models for many physical processes are often impossible to determine fully or accurately, and are hence subject to uncertainty. It is of great importance to quantify the uncertainty in the model outputs based on the (uncertain) information that is available on the model inputs. A popular way to achieve this is stochastic modelling. Based on the available information, a probability distribution (the *prior* in the Bayesian framework) is assigned to the input parameters. If in addition, some dynamic data (or *observations*) F_{obs} related to the model outputs are available, it is possible to reduce the overall uncertainty and to get a better representation of the model by conditioning the prior distribution on this data (leading to the *posterior*).

In most situations, however, the posterior distribution is intractable in the sense that exact sampling from it is impossible. One way to circumvent this problem, is to generate samples using a Metropolis–Hastings–type Markov chain Monte Carlo (MCMC) approach [22, 28, 30], 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 either accepted

*Part of this work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07A27344. LLNL-JRNL-630212-DRAFT

and used for inference, or rejected and the previous sample is used again, leading to a Markov chain. A major problem with MCMC is the high cost of the likelihood calculation for large–scale applications, e.g. in subsurface flow where, for accuracy reasons, a partial differential equation (PDE) with highly varying coefficients needs to be solved numerically on a fine spatial grid. Due to the slow convergence of Monte Carlo averaging, the number of samples is also large and moreover, the likelihood has to be calculated also for all the samples that are rejected in the end. Altogether, this often leads to an intractably high overall complexity, particularly in the context of high-dimensional parameter spaces (typical in subsurface flow), where the acceptance rate of MCMC methods can be very low.

We show here how the computational cost of the standard Metropolis-Hastings algorithm can be reduced significantly by using a multilevel approach. This has already proved highly successful in the context of standard Monte Carlo estimators based on independent and identically distributed (i.i.d.) samples [9, 1, 19, 6, 34] for subsurface flow problems. The multilevel Monte Carlo (MLMC) method was first introduced by Heinrich for the computation of high-dimensional, parameter-dependent integrals [25], and then rediscovered by Giles [18] in the context of stochastic differential equations in finance. Similar ideas were also used in [2, 3] to accelerate statistical mechanics calculations. The basic ideas are to (i) exploit the linearity of expectation, (ii) introduce a hierarchy of computational models that converge (with increasing model resolution) to some limit model (e.g. the original PDE), and (iii) build estimators for the differences of output quantities instead of the quantities themselves. In the context of PDEs with random coefficients, the multilevel estimators use a hierarchy of spatial grids and exploit that the numerical solution of a PDE, and thus the evaluation of the likelihood, is computationally much cheaper on coarser spatial grids. In that way, the individual estimators will either have small variance, since differences of output quantities from consecutive models go to zero with increased model resolution, or they will require significantly less computational work per sample for low model resolutions. Either way the cost of all the individual estimators is significantly reduced, easily compensating for the cost of having to compute $L + 1$ estimators instead of one, where L is the number of levels.

However, the application of the multilevel approach in the context of MCMC is not straightforward. The posterior distribution, which depends on the likelihood, has to be level-dependent, since otherwise the cost on all levels would be dominated by the evaluation of the likelihood on the finest level, leading to no real cost reduction. In order to avoid introducing extra bias in the estimator, we construct instead two parallel Markov chains $\{\theta_\ell^n\}_{n\geq 0}$ and $\{\Theta_{\ell-1}^n\}_{n\geq 0}$ on levels ℓ and $\ell - 1$ each from the correct posterior distribution on the respective level. The coarser of the two chains is constructed using the standard Metropolis–Hastings algorithm, for example using a (preconditioned) random walk. The main innovation is a new proposal distribution for the finer of the two chains $\{\theta_\ell^n\}_{n\geq 0}$. Although similar two-level sampling strategies have been investigated in other applications [7, 15, 16], the computationally cheaper coarse models were only used to accelerate the MCMC sampling and not as a variance reduction technique in the estimator. Some ideas on how to obtain a multilevel version of the MCMC estimator can also be found in the recent work [26] on sparse MCMC finite element methods.

The central result of the paper is a complexity theorem (cf. Theorem 3.4) that quantifies, for an abstract large–scale inference problem, the gains in the ε-cost of the multilevel Metropolis–Hastings algorithm over the standard version, i.e. the cost to achieve a root mean square error less than ε, in terms of powers of the tolerance ε. For a particular application in stationary, single phase subsurface flow with log-normal permeability prior and exponential covariance, we then verify the assumptions of Theorem 3.4. We show that the ε-cost of our new multilevel version is indeed one order of ε lower than its single-level counterpart (cf. Theorem 4.9), i.e. $\mathcal{O}(\varepsilon^{-(d+1)-\delta})$ instead of $\mathcal{O}(\varepsilon^{-(d+2)-\delta})$, for any $\delta > 0$, where d is the spatial dimension of the problem. The numerical experiments for $d = 2$ in Section 5 confirm the theoretical results. In fact, in practice the cost for the multilevel estimator grows only like $\mathcal{O}(\varepsilon^{-d})$, but this seems to be a pre–asymptotic effect. The absolute cost is about $\mathcal{O}(10\text{–}50)$ times lower than for the standard estimator for values of ε around 10^{-3}, which is a vast improvement and brings the cost of the multilevel MCMC estimator down to a similar order of the cost of standard multilevel

MC estimators based on i.i.d. samples. This provides real hope for practical applications of MCMC analyses in subsurface flow and other large scale PDE applications.

The outline of the rest of the paper is as follows. In Section 2, we recall, in a very general context, the Metropolis Hastings algorithm, together with results on its convergence. In Section 3, we then present a new multilevel version and give a general convergence analysis under a set of problem-dependent, but verifiable assumptions. A typical model problem arising in subsurface flow modelling is then presented in Section 4. We briefly describe the application of the new multilevel algorithm to this application, and give a rigorous convergence analysis and cost estimate of the new multilevel estimator by verifying the abstract assumptions from Section 3. Finally, in Section 5, we present some numerical results for the model problem discussed in Section 4.

2 Standard Markov chain Monte Carlo

We will start with a review of the standard Metropolis Hastings algorithm, described in a general context. For ease of presentation, we leave a precise mathematical description of our model problem until Section 4. We denote by $\theta := (\theta_i)_{i=1}^R$ the $\mathbb{R}^R$–valued random input vector to the model, and denote by $X := (X_j)_{j=1}^M = X(\theta)$ the $\mathbb{R}^M$–valued random output. Let further $Q_{M,R} = \mathcal{G}(X)$ be some linear or non–linear functional of X. In the context of groundwater flow modelling, this could for example be the value of the pressure or the Darcy flux at or around a given point in the computational domain, or the outflow over parts of the boundary. In practice, both θ and X are often finite dimensional approximations of infinite dimensional objects, and an underlying "true" model is recovered as $M, R \to \infty$. We shall therefore refer to M as the *discretisation level* of the model. For more details see Section 4.

We consider the setting where we have some real-world data (or *observations*) F_{obs} available, and want to incorporate this information into our simulation in order to reduce the overall uncertainty. The data F_{obs} is assumed to be finite dimensional, with $F_{\text{obs}} \in \mathbb{R}^m$ for some $m \in \mathbb{N}$, and usually corresponds to another linear or non-linear functional $\mathcal{F}(X)$ of the model output.

Let us denote the density of the conditional distribution of θ given F_{obs} by $\mathcal{P}(\theta \,|\, F_{\text{obs}})$. Using Bayes' Theorem, we have

$$\mathcal{P}(\theta \,|\, F_{\text{obs}}) = \frac{\mathcal{L}(F_{\text{obs}} \,|\, \theta)\, \pi_0^R(\theta)}{\mathcal{P}(F_{\text{obs}})} \propto \mathcal{L}(F_{\text{obs}} \,|\, \theta)\, \pi_0^R(\theta).$$

In the Bayesian framework, one usually refers to the conditional distribution $\mathcal{P}(\theta \,|\, F_{\text{obs}})$ as the *posterior distribution*, to $\mathcal{L}(F_{\text{obs}} \,|\, \theta)$ as the *likelihood* and to $\pi_0^R(\theta)$ as the *prior distribution*. Since the normalising constant $\mathcal{P}(F_{\text{obs}})$ is not known in general, the conditional distribution $\mathcal{P}(\theta \,|\, F_{\text{obs}})$ is generally intractable and exact sampling not available.

The likelihood $\mathcal{L}(F_{\text{obs}} \,|\, \theta)$ gives the probability of observing the data F_{obs} given a particular value of θ. In practice, this usually involves computing the *model response* $F_{M,R} := \mathcal{F}\left(X(\theta)\right)$ and comparing this to the observed data F_{obs}. Note that since the model response depends on the discretisation parameter M, in practice we compute an approximation $\mathcal{L}_M(F_{\text{obs}} \,|\, \theta)$ of the true likelihood $\mathcal{L}(F_{\text{obs}} \,|\, \theta)$. We will denote the corresponding density of the approximate posterior distribution by

$$\pi^{M,R}(\theta) \propto \mathcal{L}_M(F_{\text{obs}} \,|\, \theta)\, \pi_0^R(\theta).$$

Let now $\nu^{M,R}(\theta) := \pi^{M,R}(\theta)\, \mathrm{d}\theta$ denote the probability measure corresponding to the density $\pi^{M,R}$. We assume that as $M, R \to \infty$, we have $\mathbb{E}_{\nu^{M,R}}\left[Q_{M,R}\right] \to \mathbb{E}_\rho\left[Q\right]$, for some (inaccessible) random variable Q and measure ρ. The goal of the simulation is to estimate $\mathbb{E}_{\nu^{M,R}}\left[Q_{M,R}\right]$, for M, R sufficiently large. Hence, we compute approximations (or *estimators*) $\widehat{Q}_{M,R}$ of $\mathbb{E}_{\nu^{M,R}}\left[Q_{M,R}\right]$. To estimate this with a Monte Carlo type estimator, or in other words by a finite sample average, we need to generate samples from the conditional distribution $\nu^{M,R}$, which is usually intractable, as already mentioned. Instead, we will use the Metropolis Hastings MCMC algorithm in Algorithm 1.

ALGORITHM 1. (Metropolis Hastings MCMC)

Choose θ^0. For $n \geq 0$:

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

$$\alpha^{M,R}\left(\theta'|\theta^n\right) = \min\left\{1, \frac{\pi^{M,R}(\theta')\, q(\theta^n|\theta')}{\pi^{M,R}(\theta^n)\; q(\theta'|\theta^n)}\right\} \tag{2.1}$$

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

Algorithm 1 creates a Markov chain $\{\theta^n\}_{n\in\mathbb{N}}$, and the states θ^n are used as samples for inference in a Monte Carlo sampler in the usual way. The proposal distribution $q(\theta'|\theta^n)$ is what defines the algorithm. A common choice is a simple random walk. However, as outlined in [21], the basic random walk does not lead to a convergence that is independent of the input dimension R. A better choice is a preconditioned Crank-Nicholson (pCN) algorithm [11], which is also a crucial ingredient in the multilevel Metropolis-Hastings algorithm applied to the subsurface flow model problem below.

Under reasonable assumptions, one can show that $\theta^n \sim \nu^{M,R}$, as $n \to \infty$, and that sample averages computed with these samples converge to expected values with respect to the desired target distribution $\nu^{M,R}$ (see Theorem 2.2). The first few samples of the chain, say $\theta^0, \ldots, \theta^{n_0}$, are not usually used for inference to allow the chain to get close to the target distribution $\nu^{M,R}$. 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, asymptotic statements about the estimator are usually independent of the burn-in. We will denote our MCMC estimator by

$$\widehat{Q}_N^{\mathrm{MC}} := \frac{1}{N}\sum_{n=n_0+1}^{N+n_0} Q_{M,R}^n = \frac{1}{N}\sum_{n=n_0+1}^{N+n_0} \mathcal{G}\left(X(\theta^n)\right), \tag{2.2}$$

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

2.1 Convergence analysis of standard Metropolis-Hastings MCMC

Let us give a brief overview of the convergence properties of Algorithm 1, which we will need below in the analysis of the multilevel variant. For more details we refer the reader, e.g., to [30]. Let

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

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

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

The set $\mathcal{E}$ contains all parameter vectors 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 distribution 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 [30].

Lemma 2.1. *Provided $\mathcal{E} \subset \mathcal{D}$, $\nu^{M,R}$ is a stationary distribution of the chain $\{\theta^n\}_{n\in\mathbb{N}}$.*

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

Theorem 2.2. *Suppose that $\mathbb{E}_{\nu^{M,R}}\left[|Q_{M,R}|\right] < \infty$ and*

$$q(\theta|\theta^*) > 0, \text{ for all } (\theta, \theta^*) \in \mathcal{E} \times \mathcal{E}. \tag{2.3}$$

Then

$$\lim_{N\to\infty} \widehat{Q}_N^{\mathrm{MC}} = \mathbb{E}_{\nu^{M,R}}\left[Q_{M,R}\right], \qquad \text{for any } \theta^0 \in \mathcal{E} \text{ and } n_0 \geq 0.$$

The condition (2.3) 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. [21]). Lemma 2.1 and Theorem 2.2 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 in Theorem 2.2, and the estimator converges 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 (2.2), let us bound its cost. We will quantify the accuracy of our estimator via the mean square error (MSE)

$$e(\widehat{Q}_N^{\mathrm{MC}})^2 := \mathbb{E}_{\boldsymbol{\Theta}}\left[\left(\widehat{Q}_N^{\mathrm{MC}} - \mathbb{E}_\rho(Q)\right)^2\right], \tag{2.4}$$

where $\mathbb{E}_{\boldsymbol{\Theta}}$ denotes the expected value with respect to the joint distribution of $\boldsymbol{\Theta} := \{\theta^n\}_{n\in\mathbb{N}}$ as generated by Algorithm 1 (not with respect to the target measure $\nu^{M,R}$). We denote by $\mathcal{C}_\varepsilon(\widehat{Q}_N^{\mathrm{MC}})$ the computational ε-cost of the estimator, i.e. the number of floating point operations needed to achieve a MSE $e(\widehat{Q}_N^{\mathrm{MC}})^2 < \varepsilon^2$.

Classically, the 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 bound this by

$$e(\widehat{Q}_N^{\mathrm{MC}})^2 \leq \mathbb{V}_{\boldsymbol{\Theta}}\left[\widehat{Q}_N^{\mathrm{MC}}\right] + 2\left(\mathbb{E}_{\boldsymbol{\Theta}}\left[\widehat{Q}_N^{\mathrm{MC}}\right] - \mathbb{E}_{\nu^{M,R}}\left[\widehat{Q}_N^{\mathrm{MC}}\right]\right)^2 + 2\left(\mathbb{E}_{\nu^{M,R}}\left[Q_{M,R}\right] - \mathbb{E}_\rho\left[Q\right]\right)^2 \tag{2.5}$$

The three terms in (2.5) correspond to the three sources of error in the MCMC estimator. The third (and last) term in (2.5) is the discretisation error due to approximating Q by $Q_{M,R}$ and ρ by $\nu^{M,R}$. 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 number of samples and the second term is due to the samples not all being perfect (i.i.d.) samples from the target distribution $\nu^{M,R}$.

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 (2.5)) can be described using the following Central Limit Theorem, which can again be found in [30].

Let $\tilde{\theta}^0 \sim \nu^{M,R}$. 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. $\tilde{\theta}^n \sim \nu^{M,R}$ for all $n \geq 0$. The covariance structure of $\widetilde{\boldsymbol{\Theta}}$ is still implicitly defined by Algorithm 1 as for $\boldsymbol{\Theta}$. However, now $\mathbb{V}_{\widetilde{\boldsymbol{\Theta}}}[\tilde{Q}_{M,R}^n] = \mathbb{V}_{\nu^{M,R}}[\tilde{Q}_{M,R}]$, $\mathbb{E}_{\widetilde{\boldsymbol{\Theta}}}[\tilde{Q}_{M,R}^n] = \mathbb{E}_{\nu^{M,R}}[\tilde{Q}_{M,R}]$ and

$$\mathrm{Cov}_{\widetilde{\boldsymbol{\Theta}}}\left[\tilde{Q}_{M,R}^0,\, \tilde{Q}_{M,R}^n\right] = \mathbb{E}_{\widetilde{\boldsymbol{\Theta}}}\left[\left(\tilde{Q}_{M,R}^0 - \mathbb{E}_{\nu^{M,R}}[Q_{M,R}]\right)\left(\tilde{Q}_{M,R}^n - \mathbb{E}_{\nu^{M,R}}[Q_{M,R}]\right)\right],$$

for any $n \geq 0$, where $\tilde{Q}^n_{M,R} := \mathcal{G}(X(\tilde{\theta}^n))$. The so-called *asymptotic variance* of the MCMC estimator is now defined as

$$\sigma_Q^2 := \mathbb{V}_{\nu^{M,R}}\left[\tilde{Q}_{M,R}\right] + 2\sum_{n=1}^{\infty} \mathrm{Cov}_{\widetilde{\boldsymbol{\Theta}}}\left[\tilde{Q}^0_{M,R}, \tilde{Q}^n_{M,R}\right]. \tag{2.6}$$

Note that stationarity of the chain is assumed only in the definition of σ_Q^2, 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 2.3 (Central Limit Theorem)**.** *Suppose* (2.3) *holds,* $\sigma_Q^2 < \infty$*, and*

$$\mathbb{P}\left[\alpha^{M,R} = 1\right] < 1. \tag{2.7}$$

Then we have, for any $n_0 \geq 0$ *and* $\theta^0 \in \mathcal{E}$*,*

$$\sqrt{N}\left(\widehat{Q}^{\mathrm{MC}}_N - \mathbb{E}_{\nu^{M,R}}\left[Q_{M,R}\right]\right) \xrightarrow{D} \mathcal{N}(0, \sigma_Q^2),$$

where $\xrightarrow{D}$ *denotes convergence in distribution.*

The condition (2.7) 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 2.3 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_Q^2 is in general larger than in the i.i.d. case where it is simply $\mathbb{V}_{\nu^{M,R}}\left[Q_{M,R}\right]$.

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:

A1. For any $N \in \mathbb{N}$,

$$\mathbb{V}_{\boldsymbol{\Theta}}\left[\widehat{Q}^{\mathrm{MC}}_N\right] + \left(\mathbb{E}_{\boldsymbol{\Theta}}\left[\widehat{Q}^{\mathrm{MC}}_N\right] - \mathbb{E}_{\nu^{M,R}}\left[\widehat{Q}^{\mathrm{MC}}_N\right]\right)^2 \lesssim \frac{\mathbb{V}_{\nu^{M,R}}[Q_{M,R}]}{N}, \tag{2.8}$$

with a constant that is independent of M, N and R.

Such non-asymptotic bounds on the sampling errors are difficult to obtain, but have recently been proved for certain Metropolis–Hastings algorithms, see e.g. [21, 31, 26], provided the chain is sufficiently burnt–in. The implied constant in Assumption A1 usually depends on quantities such as the covariances appearing in the asymptotic variance σ_Q^2 and will in general only be independent of the dimension R for judiciously chosen proposal distributions such as the pCN algorithm. For the simple random walk, for example, the hidden constant grows linearly in R. It is possible to relax Assumption A1 and prove convergence for algorithms also in this case, but we choose not to do this for ease of presentation.

To complete the error analysis, let us now consider the last term in the MSE (2.5), the discretisation bias. As before, we assume $\mathbb{E}_{\nu^{M,R}}\left[Q_{M,R}\right] \to \mathbb{E}_{\rho}\left[Q\right]$ for $M, R \to \infty$ with a certain order of convergence

$$\left|\mathbb{E}_{\nu^{M,R}}\left[Q_{M,R}\right] - \mathbb{E}_{\rho}\left[Q\right]\right| \lesssim M^{-\alpha} + R^{-\alpha'}, \tag{2.9}$$

for some $\alpha, \alpha' > 0$. The rates α and α' will be problem dependent. Let now $R = M^{\alpha/\alpha'}$, such that the two error contributions in (2.9) are balanced. Then it follows from (2.5), (2.8) and (2.9) that the MSE of the MCMC estimator can be bounded by

$$e(\widehat{Q}^{\mathrm{MC}}_N)^2 \lesssim \frac{\mathbb{V}_{\nu^{M,R}}[Q_{M,R}]}{N} + M^{-\alpha}. \tag{2.10}$$

Under the assumption that $\mathbb{V}_{\nu^{M,R}}[Q_{M,R}]$ is approximately constant, independent of M and R, it is hence sufficient to choose $N \gtrsim \varepsilon^{-2}$ and $M \gtrsim \varepsilon^{-1/\alpha}$ to get a MSE of $\mathcal{O}(\varepsilon^2)$.

To bound the computational cost to achieve this error, the so called *ε-cost*, we assume that one sample $Q^n_{M,R}$ can be obtained at cost $\mathcal{C}(Q^n_{M,R}) \lesssim M^\gamma$, for some $\gamma > 0$. Thus, with $N \gtrsim \varepsilon^{-2}$ and $M \gtrsim \varepsilon^{-1/\alpha}$, the ε–cost of our MCMC estimator can be bounded by

$$\mathcal{C}_\varepsilon(\widehat{Q}^{\mathrm{MC}}_N) \lesssim NM^\gamma \lesssim \varepsilon^{-2-\gamma/\alpha}. \tag{2.11}$$

In many practical applications, especially in subsurface flow, both the discretisation parameter M and the length of the input R need to be very large in order for $\mathbb{E}_{\nu^{M,R}}[Q_{M,R}]$ to be a good approximation to $\mathbb{E}_\rho[Q]$. Moreover, as outlined, 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 $\mathcal{L}_M(F_{\mathrm{obs}}|\theta^n)$, and this is very expensive when M and R are large, the standard MCMC estimator (2.2) is often too expensive in practical situations. Additionally, the acceptance rate of the algorithm can be very low when R is large. This means that the covariance between the different samples will decay more slowly, which again makes the hidden constant in Assumption A1 larger, and the number of samples we have to take increases even further.

To overcome the prohibitively large computational cost of the standard MCMC estimator (2.2), we will now introduce a new multilevel version of the estimator.

3 Multilevel Markov chain Monte Carlo algorithm

The main idea of multilevel Monte Carlo (MLMC) simulation is very simple. We sample not just from one approximation $Q_{M,R}$ of Q, but from several. Let us recall the main ideas from [18, 9].

Let $\{M_\ell\}_{\ell=0}^L \subset \mathbb{N}$ be an increasing sequence in $\mathbb{N}$, i.e. $M_0 < M_1 < \ldots < M_L =: M$, and assume for simplicity that there exists an $s \in \mathbb{N}\backslash\{1\}$ such that

$$M_\ell = s\, M_{\ell-1}\,, \qquad \text{for all } \ell = 1, \ldots, L. \tag{3.1}$$

We also 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 $\ell = 1, \ldots, L$. For each level ℓ, denote correspondingly the parameter vector by $\theta_\ell \in \mathbb{R}^{R_\ell}$, the quantity of interest by $Q_\ell := Q_{M_\ell,R_\ell}$, the posterior distribution by $\nu^\ell := \nu^{M_\ell,R_\ell}$ and the posterior density by $\pi^\ell := \pi^{M_\ell,R_\ell}$. For simplicity we assume that the parameter vectors $\{\theta_\ell\}_{\ell=0}^L$ are nested, i.e. that $\theta_{\ell-1}$ is a subset of θ_ℓ, and that the elements of θ_ℓ are independent.

As for multigrid methods applied to discretised (deterministic) PDEs, the key is to avoid estimating the expected value of Q_ℓ directly on level ℓ, but instead to estimate the correction with respect to the next lower level. 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

$$\mathbb{E}_{\nu^L}[Q_L] \;=\; \mathbb{E}_{\nu^0}[Q_0] \;+\; \sum_{\ell=1}^{L} \left(\mathbb{E}_{\nu^\ell}[Q_\ell] - \mathbb{E}_{\nu^{\ell-1}}[Q_{\ell-1}]\right) \tag{3.2}$$

as a basis. Note that in the case where the distributions are the same, the above reduces to the telescoping sum used for multilevel Monte Carlo estimators based on i.i.d samples.

The idea is now to estimate each of the terms on the right hand side of (3.2) separately, in such a way that the variance of the resulting multilevel estimator is small. In particular, we will estimate each term in (3.2) by an MCMC estimator. The first term $\mathbb{E}_{\nu^0}[Q_0]$ can be estimated using the standard MCMC estimator in Algorithm 1, i.e. $\widehat{Q}^{\mathrm{MC}}_{0,N_0}$ as in (2.2) 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 every $\ell \geq 1$, we denote $Y_\ell := Q_\ell - Q_{\ell-1}$ and define the estimator on level ℓ as

$$\widehat{Y}^{\mathrm{MC}}_{\ell,N_\ell} := \frac{1}{N_\ell} \sum_{n=n_0^\ell+1}^{n_0^\ell+N_\ell} Y_\ell^n = \frac{1}{N_\ell} \sum_{n=n_0^\ell+1}^{n_0^\ell+N_\ell} Q_\ell(\theta_\ell^n) - Q_{\ell-1}(\Theta_{\ell-1}^n), \tag{3.3}$$

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 $\{\theta_\ell^n\}$ and $\{\Theta_{\ell-1}^n\}$ (see Section 3.1). The full MLMCMC estimator is defined as

$$\widehat{Q}^{\mathrm{ML}}_{L,\{N_\ell\}} \;:=\; \widehat{Q}^{\mathrm{MC}}_{0,N_0} \;+\; \sum_{\ell=1}^{L} \widehat{Y}^{\mathrm{MC}}_{\ell,N_\ell}, \tag{3.4}$$

where it is important 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 an unbiased estimator of $\mathbb{E}_{\nu^L}[Q_L]$.

There are two main ideas in [18, 9] 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 (3.2) 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 choice of θ_ℓ and $\Theta_{\ell-1}$. Note that crucially, this requires the two chains $\{\theta_\ell^n\}$ and $\{\Theta_{\ell-1}^n\}$ to be correlated. However, as long as the stationary marginal distributions of $\{\theta_\ell^n\}$ and $\{\Theta_{\ell-1}^n\}$ are ν^ℓ and $\nu^{\ell-1}$ respectively, this correlation does not introduce any bias in the telescoping sum (3.2).

3.1 The estimator for $Q_\ell - Q_{\ell-1}$

Let us fix $1 \leq \ell \leq L$. The challenge is now to generate the chains $\{\theta_\ell^n\}_{n\in\mathbb{N}}$ and $\{\Theta_{\ell-1}^n\}_{n\in\mathbb{N}}$ 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 $\Theta_{\ell-1}$. The vector $\theta_{\ell,F}$ has length $R_\ell - R_{\ell-1}$.

An easy way to construct θ_ℓ^n and $\Theta_{\ell-1}^n$ such that the variance of Y_ℓ is small, would be to generate θ_ℓ^n first, and then simply use $\Theta_{\ell-1}^n = \theta_{\ell,C}^n$. However, since we require $\Theta_{\ell-1}^n$ to come from a Markov chain with stationary distribution $\nu^{\ell-1}$, and θ_ℓ^n comes from the distribution ν^ℓ, this approach would lead to additional bias. We do, however, use a similar idea in Algorithm 2.

Let us for the moment assume that we have a way of producing i.i.d. samples from the posterior distribution $\nu^{\ell-1}$. Since the distributions $\nu^{\ell-1}$ and ν^ℓ are both approximations of the true posterior distribution ρ, and differ only in the choice of approximation parameters M and R, the distributions $\nu^{\ell-1}$ and ν^ℓ will, for sufficiently large ℓ, be very similar. The distribution $\nu^{\ell-1}$ is hence an ideal candidate for the proposal distribution on level ℓ, and this is what is used in Algorithm 2. First, we generate a sample $\Theta_{\ell-1}^{n+1}$ from the distribution $\nu^{\ell-1}$, which is independent of the previous sample $\Theta_{\ell-1}^n$. We will use the independence of these samples in Lemma 3.1. Based on $\Theta_{\ell-1}^{n+1}$, we then generate θ_ℓ^{n+1} using a new

ALGORITHM 2. (Metropolis Hastings MCMC for $Q_\ell - Q_{\ell-1}$)

Choose initial states $\Theta^0_{\ell-1} \sim \nu^{\ell-1}$ and $\theta^0_\ell := [\Theta^0_{\ell-1}\,,\, \theta^0_{\ell,F}]$. For $n \geq 0$:

- **On level $\ell-1$:** Generate an independent sample $\Theta^{n+1}_{\ell-1}$ from the distribution $\nu^{\ell-1}$.
- **On level ℓ:** Given θ^n_ℓ and $\Theta^{n+1}_{\ell-1}$, generate θ^{n+1}_ℓ using Algorithm 1 with the specific proposal distribution $q^\ell_{\mathrm{ML}}(\theta'_\ell \,|\, \theta^n_\ell)$ induced by taking $\theta'_{\ell,C} := \Theta^{n+1}_{\ell-1}$ and by generating a proposal for $\theta'_{\ell,F}$ from some proposal distribution $q^{\ell,F}_{\mathrm{ML}}(\theta'_{\ell,F} \,|\, \theta^n_{\ell,F})$ that is independent of the coarse modes. The acceptance probability is

$$\alpha^\ell_{\mathrm{ML}}(\theta'_\ell \,|\, \theta^n_\ell) \;=\; \min\left\{1, \frac{\pi^\ell(\theta'_\ell)\, q^\ell_{\mathrm{ML}}(\theta^n_\ell|\theta'_\ell)}{\pi^\ell(\theta^n_\ell)\, q^\ell_{\mathrm{ML}}(\theta'_\ell|\theta^n_\ell)}\right\}.$$

two-level proposal density q^ℓ_{ML} in conjunction with the usual Metropolis-Hastings accept/reject step in Algorithm 1. In particular, to make a proposal on level ℓ, we take $\theta'_{\ell,C} = \Theta^{n+1}_{\ell-1}$ and independently generate $\theta'_{\ell,F}$ from a proposal distribution $q^{\ell,F}_{\mathrm{ML}}$ for the fine modes, which can again be a simple random walk or the pCN algorithm.

At each step in Algorithm 2, there are two different outcomes, depending on whether we accept or reject on level ℓ. The different possibilities are given in Table 1. Observe that when we accept on level ℓ, we have $\theta^{n+1}_{\ell,C} = \Theta^{n+1}_{\ell-1}$, 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^n_ℓ on that level, which means that the coarse modes of the two states may differ.

Level ℓ test	$\Theta^{n+1}_{\ell-1}$	$\theta^{n+1}_{\ell,C}$
accept	$\Theta^{n+1}_{\ell-1}$	$\Theta^{n+1}_{\ell-1}$
reject	$\Theta^{n+1}_{\ell-1}$	$\theta^n_{\ell,C}$

Table 1: Possible states of $\Theta^{n+1}_{\ell-1}$ and $\theta^{n+1}_{\ell,C}$ in Algorithm 2.

In general, this "divergence" of the coarse modes may mean that the variance of Y_ℓ does not go to 0 as $\ell \to \infty$ for a particular application. But provided the modes are ordered according to their relative "influence" on the likelihood $\mathcal{L}(F_{\mathrm{obs}}|\theta)$, we can guarantee that $\alpha^\ell_{\mathrm{ML}}(\theta'_\ell|\theta^n_\ell) \to 1$ and thus that the variance of Y_ℓ does in fact tend to 0 as $\ell \to \infty$. We will show this for a subsurface flow application in Section 4.

The specific proposal distribution q^ℓ_{ML} in Algorithm 2 can be computed very easily and at no additional cost, leading to a simple formula for the "two-level" acceptance probability $\alpha^\ell_{\mathrm{ML}}$.

Lemma 3.1. *Let $\ell \geq 1$. Then*

$$\alpha^\ell_{\mathrm{ML}}(\theta'_\ell \,|\, \theta^n_\ell) \;=\; \min\left\{1, \frac{\pi^\ell(\theta'_\ell)\, \pi^{\ell-1}(\theta^n_{\ell,C})\, q^{\ell,F}_{\mathrm{ML}}(\theta^n_{\ell,F}|\theta'_{\ell,F})}{\pi^\ell(\theta^n_\ell)\, \pi^{\ell-1}(\theta'_{\ell,C})\, q^{\ell,F}_{\mathrm{ML}}(\theta'_{\ell,F}|\theta^n_{\ell,F})}\right\}$$

and the induced transition kernel K^ℓ_{ML} satisfies detailed balance.

Furthermore, if the distribution $q_{\mathrm{ML}}^{\ell,F}$ is either (i) symmetric, or (ii) the pCN proposal distribution, then

$$\alpha_{\mathrm{ML}}^{\ell}(\theta'_\ell \,|\, \theta_\ell^n) \;=\; \begin{cases} \min\left\{1, \dfrac{\pi^\ell(\theta'_\ell)\,\pi^{\ell-1}(\theta_{\ell,C}^n)}{\pi^\ell(\theta_\ell^n)\,\pi^{\ell-1}(\theta'_{\ell,C})}\right\}, & \textit{Case (i)}, \\ \min\left\{1, \dfrac{\mathcal{L}_\ell(F_{\mathrm{obs}} \,|\, \theta'_\ell)\,\mathcal{L}_{\ell-1}(F_{\mathrm{obs}} \,|\, \theta_{\ell,C}^n)}{\mathcal{L}_\ell(F_{\mathrm{obs}} \,|\, \theta_\ell^n)\,\mathcal{L}_{\ell-1}(F_{\mathrm{obs}} \,|\, \theta'_{\ell,C})}\right\}, & \textit{Case (ii)}. \end{cases}$$

Proof. Since the proposals for the coarse modes $\theta_{\ell,C}$ and for the fine modes $\theta_{\ell,F}$ are generated independently, the proposal density $q_{\mathrm{ML}}^{\ell}(\theta'_\ell \,|\theta_\ell^n)$ can be written as a product of densities on the two parts of θ_ℓ, i.e. $q_{\mathrm{ML}}^{\ell,C}$ and $q_{\mathrm{ML}}^{\ell,F}$. For the coarse part of the proposal distribution, we simply have $q_{\mathrm{ML}}^{\ell,C}(\theta'_{\ell,C}|\theta_{\ell,C}^n) = \pi^{\ell-1}(\theta'_{\ell,C})$ and $q_{\mathrm{ML}}^{\ell,C}(\theta_{\ell,C}^n|\theta'_{\ell,C}) = \pi^{\ell-1}(\theta_{\ell,C}^n)$.

This completes the proof of the first result. Detailed balance for K_{ML}^{ℓ} follows trivially due to the Metropolis-Hastings construction. The corollary for symmetric distributions $q_{\mathrm{ML}}^{\ell,F}$ follows by definition. The corollary for pCN proposals follows from the identity $q_{\mathrm{ML}}^{\ell,F}(\theta_{\ell,F}^n|\theta'_{\ell,F})/q_{\mathrm{ML}}^{\ell,F}(\theta'_{\ell,F}|\theta_{\ell,F}^n) = \pi_0^{\ell,F}(\theta_{\ell,F}^n)/\pi_0^{\ell,F}(\theta'_{\ell,F})$ (see, e.g. [11]), together with the factorisation $\pi_0^{\ell}(\theta_\ell) = \pi_0^{\ell-1}(\theta_{\ell,C})\,\pi_0^{\ell,F}(\theta_{\ell,F})$. □

3.2 Recursive sub-sampling to generate i.i.d. samples from $\nu^{\ell-1}$

In practice, it will not be possible to generate independent samples of the coarse level posterior distribution $\nu^{\ell-1}$ directly. We instead suggest approximating independent samples of $\nu^{\ell-1}$ using Algorithm 1 in the following manner: After a sufficiently long burn-in period, Algorithm 1 will produce samples which are (approximately) distributed according to $\nu^{\ell-1}$. Although the samples produced in this way are correlated, the correlation between the nth and $(n+j)$th sample decays as j increases, and for sufficiently large j, the samples $\Theta_{\ell-1}^n$ and $\Theta_{\ell-1}^{n+j}$ will be nearly uncorrelated. Hence, an i.i.d sequence of samples of $\nu^{\ell-1}$ can be approximated by subsampling a chain $\{\Theta_{\ell-1}^n\}_{n\in\mathbb{N}}$ generated by Algorithm 1 with, e.g., the pCN proposal distribution.

This procedure can be applied very naturally in a recursive manner. Starting on the coarsest level, burning in a Markov chain of samples and subsampling this chain to produce (nearly) independent samples from ν^0 we can then apply Algorithm 2 to produce a Markov chain of samples from ν^1. This can then be subsampled again to apply Algorithm 2 on level 2. Continuing in this way, we can recursively produce independent samples from $\nu^{\ell-1}$ for any $\ell > 0$. See Algorithm 3 in Section 5 for details.

Although, in general the i.i.d. samples of $\nu^{\ell-1}$ will in practice have to be approximated, for the analysis of our multilevel algorithm we will assume that the chains $\{\Theta_{\ell-1}^n\}_{n\in\mathbb{N}}$ and $\{\theta_\ell^n\}_{n\in\mathbb{N}}$ are generated as in Algorithm 2. The additional bias introduced in the practical Algorithm 3 below is in fact so small that we did initially not detect it in our numerical experiments, even for very short subsampling rates.

3.3 Convergence analysis of the multilevel MCMC estimator

Let us now move on to convergence properties of the multilevel estimator. As in Section 2.1, we define, for all $\ell = 0, \ldots, L$, the sets

$$\mathcal{E}^\ell = \{\theta_\ell : \pi^\ell(\theta_\ell) > 0\},$$
$$\mathcal{D}^\ell = \{\theta_\ell : q_{\mathrm{ML}}^{\ell}(\theta_\ell \,|\, \theta_\ell^*) > 0 \text{ for some } \theta_\ell^* \in \mathcal{E}^\ell\}.$$

The following convergence results follow from the classical results, due to the telescoping sum property (3.2) and the algebra of limits.

Lemma 3.2. *Provided $\mathcal{E}^\ell \subset \mathcal{D}^\ell$, ν^ℓ is a stationary marginal distribution of the chain $\{\theta_\ell^n\}_{n\in\mathbb{N}}$.*

Theorem 3.3. *Suppose that for all* $\ell = 0, \ldots, L$, $\mathbb{E}_{\nu^\ell}[|Q_\ell|] < \infty$ *and*

$$q^\ell_{\mathrm{ML}}(\theta_\ell \,|\, \theta^*_\ell) > 0, \quad \textit{for all} \;\; \theta_\ell, \theta^*_\ell \in \mathcal{E}^\ell. \tag{3.5}$$

Then

$$\lim_{\{N_\ell\}\to\infty} \widehat{Q}^{\mathrm{ML}}_{L,\{N_\ell\}} = \mathbb{E}_{\nu^L}[Q_L], \qquad \textit{for any} \;\; \theta^0_\ell \in \mathcal{E}^\ell \;\; \textit{and} \;\; n^\ell_0 \geq 0.$$

Let us have a closer look at the irreducibility condition (3.5). As in the proof of Lemma 3.1, we have

$$q^\ell_{\mathrm{ML}}(\theta_\ell|\theta^*_\ell) = \pi^{\ell-1}(\theta_{\ell,C})\, q^{\ell,F}_{\mathrm{ML}}(\theta_{\ell,F}|\theta^*_{\ell,F})$$

and thus (3.5) holds, if and only if $\pi^{\ell-1}(\theta_{\ell,C})$ and $q^{\ell,F}_{\mathrm{ML}}(\theta_{\ell,F}|\theta^*_{\ell,F})$ are both positive, for all $(\theta_\ell, \theta^*_\ell) \in \mathcal{E}^\ell \times \mathcal{E}^\ell$. Both terms are positive for common choices of likelihood, prior and proposal distributions.

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 $\ell-1$). As in the case of the standard MCMC estimator, this bound is obtained by quantifying and balancing the decay of the bias and the sampling errors of the estimator.

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 $\mathbf{\Theta}_\ell := \{\theta^n_\ell\}_{n\in\mathbb{N}} \cup \{\Theta^n_{\ell-1}\}_{n\in\mathbb{N}}$, for $\ell \geq 1$, and $\mathbf{\Theta}_0 := \{\theta^n_0\}_{n\in\mathbb{N}}$, 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 denote by $\nu^{\ell,\ell-1}$ the joint distribution of θ_ℓ and $\Theta_{\ell-1}$, for $\ell \geq 1$, which 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 convenience, we define $Y_0 := Q_0$, $\nu^{0,-1} := \nu^0$ and $M_{-1} = R_{-1} = 1$.

Theorem 3.4. *Let* $\varepsilon < \exp[-1]$ *and suppose there are positive constants* $\alpha, \alpha', \beta, \beta', \gamma > 0$ *such that* $\alpha \geq \frac{1}{2}\min(\beta, \gamma)$. *Under the following assumptions, for* $\ell = 0, \ldots, L$,

M1. $|\mathbb{E}_{\nu^\ell}[Q_\ell] - \mathbb{E}_\rho[Q]| \;\leq C_{\mathrm{M1}} \left(M_\ell^{-\alpha} + R_\ell^{-\alpha'} \right)$

M2. $\mathbb{V}_{\nu^{\ell,\ell-1}}[Y_\ell] \;\leq C_{\mathrm{M2}} \left(M_{\ell-1}^{-\beta} + R_{\ell-1}^{-\beta'} \right)$

M3. $\mathbb{V}_{\mathbf{\Theta}_\ell}[\widehat{Y}^{\mathrm{MC}}_{\ell,N_\ell}] + \left(\mathbb{E}_{\mathbf{\Theta}_\ell}[\widehat{Y}^{\mathrm{MC}}_{\ell,N_\ell}] - \mathbb{E}_{\nu^{\ell,\ell-1}}[\widehat{Y}^{\mathrm{MC}}_{\ell,N_\ell}] \right)^2 \;\leq C_{\mathrm{M3}}\; N_\ell^{-1}\, \mathbb{V}_{\nu^{\ell,\ell-1}}[Y_\ell]$

M4. $\mathcal{C}_\ell \;\leq C_{\mathrm{M4}}\; M_\ell^\gamma$,

and provided $R_\ell \gtrsim M_\ell^{\max\{\alpha/\alpha', \beta/\beta'\}}$, *there exists a number of levels* L *and a sequence* $\{N_\ell\}_{\ell=0}^L$ *such that*

$$e(\widehat{Q}^{\mathrm{ML}}_{L,\{N_\ell\}})^2 \;:=\; \mathbb{E}_{\cup_\ell \mathbf{\Theta}_\ell}\left[\left(\widehat{Q}^{\mathrm{ML}}_{L,\{N_\ell\}} - \mathbb{E}_\rho[Q] \right)^2 \right] < \varepsilon^2,$$

and

$$\mathcal{C}_\varepsilon(\widehat{Q}^{\mathrm{ML}}_{L,\{N_\ell\}}) \;\leq C_{\mathrm{ML}} \begin{cases} \varepsilon^{-2}\,|\log\varepsilon|, & \textit{if} \;\; \beta > \gamma, \\ \varepsilon^{-2}\,|\log\varepsilon|^3, & \textit{if} \;\; \beta = \gamma, \\ \varepsilon^{-2-(\gamma-\beta)/\alpha}\,|\log\varepsilon|, & \textit{if} \;\; \beta < \gamma. \end{cases}$$

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. [9, Theorem 1]), which can be found in the appendix of [9]. First note that by assumption we have $R_\ell^{-\alpha'} \lesssim M_\ell^{-\alpha}$ and $R_\ell^{-\beta'} \lesssim M_\ell^{-\beta}$.

Furthermore, in the same way as in (2.5), we can expand

$$e(\widehat{Q}^{\mathrm{ML}}_{L,\{N_\ell\}})^2 \;\leq\; \mathbb{V}_{\cup_\ell \boldsymbol{\Theta}_\ell}\left[\widehat{Q}^{\mathrm{ML}}_{L,\{N_\ell\}}\right] + 2\underbrace{\left(\mathbb{E}_{\cup_\ell \boldsymbol{\Theta}_\ell}\left[\widehat{Q}^{\mathrm{ML}}_{L,\{N_\ell\}}\right] - \mathbb{E}_{\nu^L}\left[\widehat{Q}^{\mathrm{ML}}_{L,\{N_\ell\}}\right]\right)^2}_{(I)} + 2\Big(\mathbb{E}_{\nu^L}[Q_L] - \mathbb{E}_\rho[Q]\Big)^2.$$

It follows from the Cauchy Schwarz inequality that

$$\mathbb{V}_{\cup_\ell \boldsymbol{\Theta}_\ell}\left[\widehat{Q}^{\mathrm{ML}}_{L,\{N_\ell\}}\right] = \sum_{l=0}^{L} \mathbb{V}_{\boldsymbol{\Theta}_\ell}[\widehat{Y}^{\mathrm{MC}}_{\ell,N_\ell}] + 2\sum_{0\leq \ell<\ell'\leq L} \mathrm{Cov}_{\cup_\ell \boldsymbol{\Theta}_\ell}[\widehat{Y}^{\mathrm{MC}}_{\ell,N_\ell}, \widehat{Y}^{\mathrm{MC}}_{\ell',N_{\ell'}}] \lesssim (L+1)\sum_{l=0}^{L} \mathbb{V}_{\boldsymbol{\Theta}_\ell}[\widehat{Y}^{\mathrm{MC}}_{\ell,N_\ell}].$$

We can bound the second term in the MSE above by

$$(I) \;=\; \left(\sum_{l=0}^{L}\left(\mathbb{E}_{\boldsymbol{\Theta}_\ell}\left[\widehat{Y}^{\mathrm{MC}}_{\ell,N_\ell}\right] - \mathbb{E}_{\nu^{\ell,\ell-1}}\left[\widehat{Y}^{\mathrm{MC}}_{\ell,N_\ell}\right]\right)\right)^2 \;\leq\; (L+1)\sum_{l=1}^{L}\left(\mathbb{E}_{\boldsymbol{\Theta}_\ell}\left[\widehat{Y}^{\mathrm{MC}}_{\ell,N_\ell}\right] - \mathbb{E}_{\nu^{\ell,\ell-1}}\left[\widehat{Y}^{\mathrm{MC}}_{\ell,N_\ell}\right]\right)^2,$$

and thus it follows from Assumption M3 that

$$e(\widehat{Q}^{\mathrm{ML}}_{L,\{N_\ell\}})^2 \lesssim (L+1)\sum_{\ell=0}^{L} N_\ell^{-1}\,\mathbb{V}_{\nu^{\ell,\ell-1}}[Y_\ell] + \Big(\mathbb{E}_{\nu^L}[Q_L] - \mathbb{E}_\rho[Q]\Big)^2. \tag{3.6}$$

In contrast to i.i.d case, we have an additional factor $(L+1)$ multiplying the sampling error term on the right hand side of (3.6). Hence, 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, which also increases the cost of the multilevel estimator by a factor of $(L+1)$. The remainder of the proof remains identical.

Since L is chosen such that the second term in (3.6) (the bias of the multilevel estimator) is less than $\varepsilon^2/2$, it follows from Assumption M1 that $L+1 \lesssim |\log\varepsilon|$. The bounds on the ε-cost then follow as in [9, Theorem 1], but with an extra $|\log\varepsilon|$ factor. □

Note that in our proof we do not require the estimators $\widehat{Y}^{\mathrm{MC}}_{\ell,N_\ell}$, $\ell = 0,\ldots,L$, to be independent. However, in practice we found that independent estimators lead to a faster absolute performance of the multilevel estimator (in terms of cost versus error).

Assumptions M1 and M4 are the same assumptions as in the single–level case, and are related to the bias in the model (e.g. due to discretisation) and to the cost per sample, respectively. Assumption M3 is similar to assumption A1, 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.

4 Model Problem

In this section, we will apply the proposed MLMCMC algorithm to a simple model problem arising in subsurface flow modelling. Probabilistic uncertainty quantification in subsurface flow is of interest in a number of situations, as for example in risk analysis for radioactive waste disposal or in oil reservoir simulation. The classical equations governing (steady state) single–phase subsurface flow consist of Darcy's law coupled with an incompressibility condition (see e.g. [14, 10]):

$$w + k\nabla p = g \quad \text{and} \quad \text{div } w = 0, \quad \text{in } \; D \subset \mathbb{R}^d,\; d = 1,2,3, \tag{4.1}$$

subject to suitable boundary conditions. In physical terms, p denotes the pressure head of the fluid, k is the permeability tensor, w is the filtration velocity (or Darcy flux) and g is the source term.

4.1 Uncertainty quantification

A typical approach to quantify uncertainty in p and w is to model the permeability as a random field $k = k(x, \omega)$ on $D \times \Omega$, for some probability space $(\Omega, \mathcal{A}, \mathbb{P})$. The mean and covariance structure of k has to be inferred from the (limited) geological information available. This means that (4.1) becomes a system of PDEs with random coefficients, which can be written in second order form as

$$-\nabla \cdot (k(x,\omega)\nabla p(x,\omega)) = f(x), \qquad \text{in} \quad D, \tag{4.2}$$

with $f := -\text{div}\, g$. This means that the solution p itself will also be a random field on $D \times \Omega$. For simplicity, we shall restrict ourselves to Dirichlet conditions $p(\omega, x) = \psi(x)$ on ∂D, and assume that the boundary data ψ and the source term g are known (and thus deterministic).

In this general form solving (4.2) is extremely challenging computationally, and so in practice it is common to use relatively simple models for k that are as faithful as possible to the measurements. One model that has been studied extensively is a log-normal distribution for k, i.e. replacing the permeability tensor by a scalar valued field whose log is Gaussian. It guarantees that $k > 0$ almost surely (a.s.) in Ω, and it allows the permeability to vary over many orders of magnitude, which is typically the case.

When modelling a whole oil reservoir or a sufficiently large region around a potential radioactive waste repository, the correlation length scale for k is typically significantly smaller than the size of the computational region. In addition, typical sedimentation processes lead to fairly irregular structures and pore networks. Faithful models should therefore also only assume limited spatial regularity of k. A covariance function that has been proposed in the application literature (cf. [27]) is the following exponential two-point covariance function for $\log k$:

$$C(x,y) \;:=\; \sigma^2 \exp\left(-\frac{\|x-y\|_r}{\lambda}\right), \qquad x, y \in D, \tag{4.3}$$

where $\|\cdot\|_r$ denotes the ℓ_r-norm in $\mathbb{R}^d$ and typically $r = 1$ or 2. The parameters σ^2 and λ denote *variance* and *correlation length*, respectively. In subsurface flow applications typically only $\sigma^2 \geq 1$ and $\lambda \leq \text{diam}\, D$ will be of interest. The choice of covariance function in (4.3) implies that k is *homogeneous* and it follows from Kolmogorov's theorem [29] that $k(\cdot, \omega) \in C^{0,t}(D)$ a.s., for any $t < 1/2$.

For the purpose of this paper, we will assume that k is a log-normal random field, where $\log k$ has mean zero and exponential covariance function (4.3) with $r = 1$. However, other models for k are possible, and the required theoretical results can be found in [6, 34, 33].

Let us now put model problem (4.2) into context for the MCMC and MLMCMC methods described in sections 2 and 3. The quantity of interest Q is in this case some functional $\mathcal{G}$ of the PDE solution p, and $Q_{M,R}$ is the same functional $\mathcal{G}$ evaluated at a discretised solution $p_{M,R}$. The discretisation level M denotes the number of degrees of freedom for the numerical solution of (4.2) for a given sample and the parameter R denotes the number of random variables used to model the permeability k. The random vector X will contain the M degrees of freedom of the discrete pressure $p_{M,R}$.

For the spatial discretisation of model problem (4.2), we will use standard, continuous, piecewise linear finite elements (FEs), see e.g. [4, 8] for more details. Other spatial discretisation schemes are possible, see for example [9] for a numerical study with finite volume methods and [20] for a theoretical treatment of mixed finite elements. We choose a regular triangulation $\mathcal{T}_h$ of mesh width h of our spatial domain D, which results in $M = \mathcal{O}(h^{-d})$ degrees of freedom for the numerical approximation.

In order to apply the proposed MCMC methods to model problem (4.2), we need to represent the permeability k in terms of a set of random variables. For this, we will use the Karhunen-Loève (KL-) expansion. For the Gaussian field $\log k$, this is an expansion in terms of a countable set of independent, standard Gaussian random variables $\{\xi_n\}_{n\in\mathbb{N}}$. It is given by

$$\log k(\omega, x) = \sum_{n=1}^{\infty} \sqrt{\mu_n}\phi_n(x)\xi_n(\omega),$$

where $\{\mu_n\}_{n\in\mathbb{N}}$ are the eigenvalues and $\{\phi_n\}_{n\in\mathbb{N}}$ the corresponding L^2-normalised eigenfunctions of the covariance operator with kernel function $C(x,y)$. For more details on its derivation and properties, see e.g. [17]. We will here only mention that the eigenvalues $\{\mu_n\}_{n\in\mathbb{N}}$ are non–negative with $\sum_{n\geq 0}\mu_n < \infty$. For the particular covariance function (4.3) with $r = 1$, we have $\mu_n \lesssim n^{-2}$ and hence there is an intrinsic ordering of importance in the KL-expansion. Truncating the KL-expansion after R terms, gives an approximation of k in terms of R standard normal random variables,

$$k_R(\omega, x) = \exp\left[\sum_{n=1}^{R}\sqrt{\mu_n}\phi_n(x)\xi_n(\omega)\right]. \tag{4.4}$$

Denote by $\vartheta := \{\xi_n\}_{n\in\mathbb{N}} \in \mathbb{R}^{\mathbb{N}}$ the vector of independent random variables appearing in the KL-expansion of $\log k$. 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) $\mathcal{N}(0,1)$ random variables, such that

$$\rho_0 = \bigotimes_{n\in\mathbb{N}} g(\xi_n)\,\mathrm{d}\xi_n, \tag{4.5}$$

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

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

$$F_{\text{obs}} = \mathcal{F}(p(\vartheta)) + \eta, \tag{4.6}$$

where $\mathcal{F} : H^1(D) \to \mathbb{R}^m$ is a continuous function of p, the (weak) solution to model problem (4.1) which depends on ϑ through k. The observational noise η is assumed to be a realisation of a $\mathcal{N}(0, \sigma_F^2 I_m)$ 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 (4.5), we have $\rho_0(\mathbb{R}^{\mathbb{N}}) = 1$. Furthermore, since p depends continuously on ϑ (see [5, Propositions 3.6 and 4.1] or [35, Lemmas 2.20 and 5.13]), the map $\mathcal{F}\circ p : \mathbb{R}^{\mathbb{N}} \to \mathbb{R}^m$ is also continuous (by assumption). The posterior distribution, which we will denote by ρ, is then known to be absolutely continuous with respect to the prior and satisfies

$$\frac{\partial\rho}{\partial\rho_0}(\vartheta) \eqsim \exp\left[-\frac{\|F_{\text{obs}} - \mathcal{F}(p(\vartheta))\|^2}{2\sigma_F^2}\right] =: \exp\left[-\Phi(\vartheta; F_{\text{obs}})\right], \tag{4.7}$$

where $\|\cdot\|$ denotes the Euclidean norm on $\mathbb{R}^m$. The hidden constant depends only on F_{obs} and is generally not known (for more details see [32] and the references therein). The right hand side of (4.7) is referred to as the *likelihood*.

Since the exact solution $p(\vartheta)$ is not available, the likelihood $\exp\left[-\Phi(\vartheta; F_{\text{obs}})\right]$ needs to be approximated in practical computations. We use a truncation of the KL-expansion of $\log k$ after R terms and a spatial approximation $p_{M,R}$ of $p(\vartheta)$ by piecewise linear FEs. The value of σ_F^2 may also be changed to $\sigma_{F,M}^2$. We denote the resulting approximate posterior measure correspondingly by $\rho^{M,R}$, with

$$\frac{\partial\rho^{M,R}}{\partial\rho_0}(\vartheta) \eqsim \exp\left[-\frac{\|F_{\text{obs}} - \mathcal{F}(p_{M,R}(\vartheta))\|^2}{2\sigma_{F,M}^2}\right] =: \exp\left[-\Phi^{M,R}(\vartheta; F_{\text{obs}})\right]. \tag{4.8}$$

Since $\mathcal{F}\circ p_{R,M}$ only depends on $\theta := \{\xi_n\}_{n=1}^R$, the first R components of ϑ, and since the prior measure factorises as $\rho_0 = \rho_0^R \otimes \rho_0^\perp$, the approximate posterior measure also factorises as $\rho^{M,R} = \nu^{M,R}\otimes\rho^\perp$, where

$$\frac{\partial\nu^{M,R}}{\partial\rho_0^R}(\theta) \eqsim \exp\left[-\Phi^{M,R}(\theta; F_{\text{obs}})\right], \tag{4.9}$$

and $\rho^\perp = \rho_0^\perp$ [12]. Note that $\nu^{M,R}$ is a measure on the finite dimensional space $\mathbb{R}^R$. Denoting by $\pi^{M,R}$ and π_0^R the densities with respect to the R dimensional Lebesgue measure of $\nu^{M,R}$ and ρ_0^R, respectively, it follows from (4.9) that

$$\pi^{M,R}(\theta) \eqsim \exp\left[-\Phi^{M,R}(\theta; F_{\text{obs}})\right] \pi_0^R(\theta)\,. \tag{4.10}$$

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

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

where $\mathbb{E}_{\nu^{M,R}}[Q_{M,R}] := \int_{\mathbb{R}^R} \mathcal{G}(p_{M,R}(\theta))\, \nu^{M,R}(\mathrm{d}\theta)$ is a finite dimensional integral.

Finally, let us set the notation for our MLMCMC algorithm. To achieve a level-dependent representation of k, we simply truncate the KL-expansion after a sufficiently large, level-dependent number of terms R_ℓ, such that the truncation error on each level is bounded by the discretisation error, and set $\theta_\ell := \{\xi_n\}_{n=1}^{R_\ell}$. A sequence of discretisation levels M_ℓ satisfying (3.1) can be constructed by choosing a coarsest mesh width h_0 for the spatial approximation, and choosing $h_\ell := s^{-\ell} h_0$. A common (but not necessarily optimal) choice is $s = 2$ and uniform refinement between the levels. We denote the resulting (truncated) FE solution by $p_\ell := p_{M_\ell, R_\ell}$.

The prior density π_0^ℓ of θ_ℓ is simply a standard R_ℓ-dimensional Gaussian:

$$\pi_0^\ell(\theta_\ell) = \frac{1}{(2\pi)^{R_\ell/2}} \exp\left[-\sum_{j=1}^{R_\ell} \frac{\xi_j^2}{2}\right]. \tag{4.11}$$

For the likelihood, we have

$$\mathcal{L}_\ell(F_{\text{obs}} \,|\, \theta_\ell) \eqsim \exp\left[\frac{-\|F_{\text{obs}} - F^\ell(\theta_\ell)\|^2}{2\sigma_{F,\ell}^2}\right], \tag{4.12}$$

where $F^\ell(\theta_\ell) = \mathcal{F}(p_\ell(\theta_\ell))$. Recall that the coarser levels in our multilevel estimator are introduced only to accelerate the convergence and that the multilevel estimator is still an unbiased 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 $\sigma_{F,\ell}^2$ on the coarse levels, which will increase the acceptance probability on the coarser levels. The growth in $\sigma_{F,\ell}^2$ has to be controlled, as we will see below (cf. Assumption A3).

4.2 Convergence analysis

We now perform a rigorous convergence analysis of the MLMCMC estimator $\widehat{Q}^{\text{ML}}_{L,\{N_\ell\}}$ introduced in Section 3 applied to model problem (4.1). We will first verify that the multilevel estimator is indeed an unbiased estimator of $\mathbb{E}_{\nu^L}[Q_L]$. To achieve this, we only need to verify the irreducibility condition (3.5) in Theorem 3.3. As already noted, for common choices of proposal distributions, the condition holds true if $\pi^{\ell-1}(\theta_{\ell,C}) > 0$, for all θ_ℓ s.t. $\pi^\ell(\theta_\ell) > 0$. The conclusion follows, since both the prior and the likelihood were chosen as normal distributions and normal distributions have infinite support.

Theorem 4.1. *Suppose that for all $\ell = 0, \ldots, L$, $\mathbb{E}_{\nu^\ell}[|Q_\ell|] < \infty$. Then*

$$\lim_{\{N_\ell\}\to\infty} \widehat{Q}^{\text{ML}}_{L,\{N_\ell\}} = \mathbb{E}_{\nu^L}[Q_L]\,, \qquad \textit{for any } \ \theta_\ell^0 \in \mathcal{E}^\ell \ \textit{ and } \ n_0^\ell \geq 0.$$

Let us now move on to quantifying the cost of the multilevel estimator, and verify the assumptions in Theorem 3.4 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 paper. Results of this kind can be found in e.g. [31, 21]. 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 (FE) equations for each sample, M4 is satisfied with $\gamma = 1$.

We will address assumptions M1 and M2, which are the assumptions related to the discretisation errors in the quantity of interest Q and the measure ρ. For ease of presentation, we will for the remainder of this section assume that $\log k$ has mean zero and exponential covariance function (4.3) with $r = 1$, and that ψ and f in (4.2) are deterministic, with $\psi \in H^1(\partial D)$ and $f \in H^{-1/2}(D)$. This implies that the solution p to (4.2) is in $L^q(\Omega, H^{3/2-\delta})$, for any $\delta > 0$ and $q < \infty$ (cf. [34]). In the Metropolis-Hastings algorithm we will only consider symmetric proposal distributions or the pCN algorithm.

Since they will become useful later, let us recall some of the main results in the convergence analysis of ("plain vanilla") multilevel Monte Carlo estimators based on independent and identically distributed (i.i.d.) samples. An extensive convergence analysis of FE multilevel estimators based on i.i.d. samples for model problem (4.2) with log–normal coefficients can be found in [6, 34, 33]. We firstly have the following result on the convergence of the FE error in the natural H^1–norm.

Theorem 4.2. *Let g be a Gaussian field with constant mean and covariance function (4.3) with $r = 1$, and let $k = \exp[g]$ in model problem (4.2). Suppose $D \subset \mathbb{R}^d$ is Lipschitz polygonal (polyhedral). Then*

$$\mathbb{E}_{\rho_0}\left[|p - p_\ell|^q_{H^1(D)}\right]^{1/q} \leq C_{k,f,\psi,q}\,(M_\ell^{-1/2d+\delta} + R_\ell^{-1/2+\delta}),$$

for any $q < \infty$ and $\delta > 0$, where the (generic) constant $C_{k,f,\psi,q}$ (here and below) depends on the data k, f, ψ and on q, but is independent of any other parameters.

Proof. This follows from [34, Proposition 4.1]. □

Convergence results for functionals of the solution p can now be derived from Theorem 4.2 using a duality argument. We will here for simplicity only consider bounded, linear functionals, but the results extend to continuously Frèchet differentiable functionals (see [34, §3.2]). We make the following assumption on the functional $\mathcal{G}$ (cf. Assumption F1 in [34]).

A2. Let $\mathcal{G} : H^1(D) \to \mathbb{R}$ be linear, and suppose there exists $C_{\mathcal{G}} \in \mathbb{R}$, such that

$$|\mathcal{G}(v)| \leq C_{\mathcal{G}} \|v\|_{H^{1/2-\delta}}, \qquad \text{for all } \delta > 0.$$

An example of a functional which satisfies A2 is a local average of the pressure, $\frac{1}{|D^*|}\int_{D^*} p\,\mathrm{d}x$ for some $D^* \subset D$. The main result on the convergence for functionals is the following.

Corollary 4.3. *Let the assumptions of Theorem 4.2 be satisfied, and suppose $\mathcal{G}$ satisfies A2. Then*

$$\mathbb{E}_{\rho_0}\left[|\mathcal{G}(p) - \mathcal{G}(p_\ell)|^q\right]^{1/q} \leq C_{k,f,\psi,q}\left(M_\ell^{-1/d+\delta} + R_\ell^{-1/2+\delta}\right),$$

for any $q < \infty$ and $\delta > 0$.

Proof. This follows from [34, Corollary 4.1]. □

Note that assumption A2 is crucial in order to get the faster convergence rates of the spatial discretisation error in Corollary 4.3. For multilevel estimators based on i.i.d. samples, it follows immediately from Corollary 4.3 that the (corresponding) assumptions M1 and M2 are satisfied, with $\alpha = 1/d + \delta$, $\alpha' = 1/2 + \delta$ and $\beta = 2\alpha$, $\beta' = 2\alpha'$, for any $\delta > 0$ (see [34] for details).

The aim is now to generalise the result in Corollary 4.3 to the new MLMCMC estimator. Two issues need to be addressed. Firstly, the bounds in assumptions M1 and M2 in Theorem 3.4 involve moments with respect to the posterior distributions ν^ℓ and ρ, which are not known explicitly, but are related to the prior distributions ρ_0^ℓ and ρ_0 through Bayes' Theorem. Secondly, the samples on levels ℓ and $\ell-1$ that are used to compute samples of the differences $Y_\ell = Q_\ell - Q_{\ell-1}$ are generated by Algorithm 2, and may differ not only due to discretisation and 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 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 to moments with respect to the prior distribution.

Lemma 4.4. *For any random variable $Z = Z(\theta_\ell)$ and for any q s.t. $\mathbb{E}_{\rho_0^\ell}\left[|Z|^q\right] < \infty$, we have*

$$\left|\,\mathbb{E}_{\nu^\ell}\left[Z^q\right]\,\right| \lesssim \mathbb{E}_{\rho_0^\ell}\left[|Z|^q\right].$$

Similarly, for any random variable $Z = Z(\vartheta)$ and for any q s.t. $\mathbb{E}_{\rho_0}\left[|Z|^q\right] < \infty$, we have

$$\left|\,\mathbb{E}_{\rho^\ell}\left[Z^q\right]\,\right| \lesssim \mathbb{E}_{\rho_0}\left[|Z|^q\right].$$

Proof. Using (4.10), we have

$$\begin{aligned}\left|\,\mathbb{E}_{\nu^\ell}\left[Z^q\right]\,\right| &\eqsim \left|\int_{\mathbb{R}^{R_\ell}} Z^q(\theta_\ell)\exp\left[-\Phi^{M,R}(\theta_\ell;F_{\text{obs}})\right]\,\pi_0^\ell(\theta_\ell)\,\mathrm{d}\theta_\ell\right| \\ &\lesssim \sup_{\theta_\ell}\left\{\exp\left[-\Phi^{M,R}(\theta_\ell;F_{\text{obs}})\right]\right\}\int_{\mathbb{R}^{R_\ell}}|Z(\theta_\ell)|^q\,\pi_0^\ell(\theta_\ell)\,\mathrm{d}\theta_\ell.\end{aligned}$$

The first claim of the Lemma then follows, since the above supremum can be bounded by 1. The proof of the second claim is analogous, using the Radon-Nikodym derivative (4.7). □

We are now ready to prove assumption M1, under the following assumption on the parameters $\sigma_{F,\ell}^2$ in the likelihood model (4.12):

A3. The sequence of fidelity parameters $\{\sigma_{F,\ell}^2\}_{\ell=0}^\infty$ satisfies

$$\sigma_F^{-2} - \sigma_{F,\ell}^{-2} \lesssim \max\left(R_\ell^{-1/2+\delta}, M_\ell^{-1/d+\delta}\right), \quad \text{for all } \delta > 0.$$

Lemma 4.5. *Let the assumptions of Corollary 4.3 be satisfied. Suppose $\mathcal{F}$ satisfies A2, and A3 holds. Then*

$$|\mathbb{E}_{\nu^\ell}[Q_\ell] - \mathbb{E}_\rho[Q]| \le C_{k,f,\psi}\left(M_\ell^{-1/d+\delta} + R_\ell^{-1/2+\delta}\right).$$

Proof. Since Q_ℓ only depends on θ_ℓ we have $\mathbb{E}_{\nu^\ell}[Q_\ell] = \mathbb{E}_{\rho^\ell}[Q_\ell]$ and so, using the triangle inequality,

$$|\mathbb{E}_{\nu^\ell}[Q_\ell] - \mathbb{E}_\rho[Q]| \le |\mathbb{E}_{\rho^\ell}[Q_\ell] - \mathbb{E}_{\rho^\ell}[Q]| + |\mathbb{E}_{\rho^\ell}[Q] - \mathbb{E}_\rho[Q]|. \tag{4.13}$$

The first term can be bounded using Corollary 4.3 and Lemma 4.4, i.e.

$$|\mathbb{E}_{\rho^\ell}[Q_\ell] - \mathbb{E}_{\rho^\ell}[Q]| \le C_{k,f,\psi}\left(M_\ell^{-1/d+\delta} + R_\ell^{-1/2+\delta}\right).$$

For the second term, we will prove a bound on the Hellinger distance $d_{\text{Hell}}(\rho,\rho^\ell)$. This proof follows closely the proof of [26, Proposition 10]. Denote by Z and Z_ℓ the normalising constants of ρ and ρ^ℓ:

$$Z = \int_{\mathbb{R}^{\mathbb{N}}}\exp\left[-\frac{1}{2}\Phi(\vartheta;F_{\text{obs}})\right]\mathrm{d}\rho_0(\vartheta) \quad \text{and} \quad Z_\ell = \int_{\mathbb{R}^{\mathbb{N}}}\exp\left[-\frac{1}{2}\Phi^\ell(\vartheta;F_{\text{obs}})\right]\mathrm{d}\rho_0(\vartheta), \quad \text{respectively.}$$

Since $\mathcal{F}$ satisfies Assumption A2, it follows from the results in [32] that both Z and Z_ℓ can be bounded away from zero. Next, we have

$$2\, d^2_{\mathrm{Hell}}(\rho,\rho^\ell) = \int_{\mathbb{R}^{\mathbb{N}}} \left(Z^{-1/2} \exp\left[-\frac{1}{2}\Phi(\vartheta;F_{\mathrm{obs}})\right] - Z_\ell^{-1/2} \exp\left[-\frac{1}{2}\Phi^\ell(\vartheta;F_{\mathrm{obs}})\right]\right)^2 \mathrm{d}\rho_0(\vartheta) \leq I + II,$$

where

$$I := \frac{2}{Z}\int_{\mathbb{R}^{\mathbb{N}}} \left(\exp\left[-\frac{1}{2}\Phi(\vartheta;F_{\mathrm{obs}})\right] - \exp\left[-\frac{1}{2}\Phi^\ell(\vartheta;F_{\mathrm{obs}})\right]\right)^2 \mathrm{d}\rho_0(\vartheta),$$
$$II := 2\,|Z^{-1/2} - Z_\ell^{-1/2}|^2 \int_{\mathbb{R}^{\mathbb{N}}} \exp\left[-\Phi^\ell(\vartheta;F_{\mathrm{obs}})\right] \mathrm{d}\rho_0(\vartheta).$$

To estimate I, note that both $\exp\left[-\frac{1}{2}\Phi(\vartheta;F_{\mathrm{obs}})\right]$ and $\exp\left[-\frac{1}{2}\Phi^\ell(\vartheta;F_{\mathrm{obs}})\right]$ are bounded above by 1, so that

$$\exp\left[-\frac{1}{2}\Phi(\vartheta;F_{\mathrm{obs}})\right] - \exp\left[-\frac{1}{2}\Phi^\ell(\vartheta;F_{\mathrm{obs}})\right] \leq |\Phi(\vartheta;F_{\mathrm{obs}}) - \Phi^\ell(\vartheta;F_{\mathrm{obs}})|.$$

Denoting $F := \mathcal{F}(p(\vartheta))$ and $F_\ell := \mathcal{F}(p_\ell(\theta))$, and using the triangle inequality, we have that

$$\left|\frac{\|F_{\mathrm{obs}}-F\|^2}{\sigma_F^2} - \frac{\|F_{\mathrm{obs}}-F_\ell\|^2}{\sigma_{F,\ell}^2}\right| \;\leq\; \left|\frac{\left(\|F_{\mathrm{obs}}-F_\ell\| + \|F-F_\ell\|\right)^2}{\sigma_F^2} - \frac{\|F_{\mathrm{obs}}-F_\ell\|^2}{\sigma_{F,\ell}^2}\right|$$
$$= \;\|F_{\mathrm{obs}}-F_\ell\|^2\left(\sigma_F^{-2} - \sigma_{F,\ell}^{-2}\right) \;+\; \frac{2\|F_{\mathrm{obs}}-F_\ell\| + \|F-F_\ell\|}{\sigma_F^2}\|F-F_\ell\|.$$

Since $\mathcal{F}$ was assumed to satisfy A2, it follows from Corollary 4.3 that

$$\mathbb{E}_{\rho_0}[\|F-F_\ell\|^q]^{1/q} \;\leq\; C_{k,f,\psi}\left(M_\ell^{-1/d+\delta} \;+\; R_\ell^{-1/2+\delta}\right).$$

Moreover, since $\|F_\ell\|$ can be bounded independently of ℓ (again courtesy of Assumption A2), and since $\|F_{\mathrm{obs}}-F_\ell\| \leq \|F_{\mathrm{obs}}\| + \|F_\ell\|$, we can deduce that

$$I \;\lesssim\; \mathbb{E}_{\rho_0}[|\Phi(\vartheta;F_{\mathrm{obs}}) - \Phi^\ell(\vartheta;F_{\mathrm{obs}})|^2] \;\leq\; C_{k,f,\psi}\left(M_\ell^{-1/d+\delta} \;+\; R_\ell^{-1/2+\delta}\right)^2.$$

using Assumption A3. For the second term II, we note that $|Z^{-1/2}-Z_\ell^{-1/2}|^2 \lesssim \max\{Z^{-3}, Z_\ell^{-3}\}\,|Z-Z_\ell|^2$, and an analysis similar to the above shows that

$$II \;\lesssim\; \mathbb{E}_{\rho_0}[|\Phi(\vartheta;F_{\mathrm{obs}}) - \Phi^\ell(\vartheta;F_{\mathrm{obs}})|]^2 \;\leq\; C_{k,f,\psi}\left(M_\ell^{-1/d+\delta} \;+\; R_\ell^{-1/2+\delta}\right)^2.$$

The claim of the Theorem then follows, since $|\mathbb{E}_{\rho^\ell}[Q] - \mathbb{E}_\rho[Q]| \leq C_{k,f,\psi} d_{\mathrm{Hell}}(\rho,\rho^\ell)$. □

In order to prove M2, we further have to analyse the situation where the two samples θ_ℓ^{n+1} and $\Theta_{\ell-1}^{n+1}$ used to compute Y_ℓ^{n+1} "diverge", i.e. when $\Theta_{\ell-1}^{n+1} \neq \theta_{\ell,C}^{n+1}$.

For the remainder we will consider only symmetric or pCN proposal distributions.

Lemma 4.6. *Let θ_ℓ^{n+1} and $\Theta_{\ell-1}^{n+1}$ have joint distribution $\nu^{\ell,\ell-1}$, and set $Y_\ell^{n+1} = Q_\ell(\theta_\ell^{n+1}) - Q_{\ell-1}(\Theta_{\ell-1}^{n+1})$. If $q_{\mathrm{ML}}^{\ell,F}$ is a pCN proposal distribution, then*

$$\mathbb{V}_{\nu^{\ell,\ell-1}}\left[Y_\ell^{n+1}\right] \leq C_{k,f,\psi}\left(M_{\ell-1}^{-1/d+\delta} + R_{\ell-1}^{-1/2+\delta}\right), \quad \textit{for any } \; \delta > 0.$$

This bound also holds for a symmetric proposal distribution $q_{\mathrm{ML}}^{\ell,F}$ under the additional assumption that

$$(R_\ell - R_{\ell-1})(2\pi)^{-\frac{R_\ell - R_{\ell-1}}{2}} \;\lesssim\; R_{\ell-1}^{-1/2+\delta}, \quad \textit{for all } \; \delta > 0. \tag{4.14}$$

For the growth condition (4.14) to be satisfied, it suffices that $R_\ell - R_{\ell-1}$ grows logarithmically with $R_{\ell-1}$. To prove Lemma 4.6, we first need some preliminary results. Firstly, note that $\Theta_{\ell-1}^{n+1} \neq \theta_{\ell,C}^{n+1}$ only if the proposal θ_ℓ' generated for θ_ℓ^{n+1} was rejected. Given the states θ_ℓ^n and θ_ℓ', the probability of this rejection is given by $1 - \alpha_{\mathrm{ML}}^\ell(\theta_\ell'|\theta_\ell^n)$. The total probability of a rejection is then $\mathbb{E}_{\boldsymbol{\zeta}}[(1 - \alpha_{\mathrm{ML}}^\ell]$, where ζ denotes the joint distribution of the two variables. We need to quantify this probability.

Before we can do so, we need to specify the (marginal) distribution of the proposal θ_ℓ', which we denote by ζ_ℓ'. 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 distributed according to the proposal density $q_{\mathrm{ML}}^{\ell,F}(\theta_{\ell,F}' \,|\, \theta_{\ell,F}^n)$ (independent of the first $R_{\ell-1}$ dimensions). The same proof technique as in Lemma 4.4 shows again that $|\mathbb{E}_{\zeta_\ell'}[Z^q]| \lesssim \mathbb{E}_{\rho_0^\ell}[|Z|^q]$, for any random variable $Z = Z(\theta_\ell')$.

Lemma 4.7. *Let θ_ℓ^n and θ_ℓ' be as generated by Algorithm 2 at the $(n+1)$th step. Denote their joint distribution by $\boldsymbol{\zeta}$, with marginal distributions ν^ℓ and ζ_ℓ', respectively. Suppose $\mathcal{F}$ satisfies A2, and A3 and the assumptions of Corollary 4.3 hold. If $q_{\mathrm{ML}}^{\ell,F}$ is a pCN proposal distribution, then*

$$\mathbb{E}_{\boldsymbol{\zeta}}\Big[(1 - \alpha_{\mathrm{ML}}^\ell(\theta_\ell'|\theta_\ell^n))\Big] \leq C_{k,f,\psi}\Big(M_{\ell-1}^{-1/d+\delta} + R_{\ell-1}^{-1/2+\delta}\Big), \quad \textit{for any } \ \delta > 0.$$

This bound also holds for a symmetric proposal distribution $q_{\mathrm{ML}}^{\ell,F}$ under the additional assumption (4.14).

Proof. We will start by assuming that $q_{\mathrm{ML}}^{\ell,F}$ is a pCN proposal distribution. For brevity, denote $\mathcal{L}_\ell(F_{\mathrm{obs}} \,|\, \cdot) =: \mathcal{L}_\ell(\cdot)$. We will first derive a bound on $1 - \alpha_{\mathrm{ML}}^\ell(\theta_\ell' \,|\, \theta_\ell^n)$, for $\ell > 1$ and for θ_ℓ' and θ_ℓ^n given. First note that if $\frac{\mathcal{L}_\ell(\theta_\ell')\,\mathcal{L}_{\ell-1}(\theta_{\ell,C}^n)}{\mathcal{L}_\ell(\theta_\ell^n)\,\mathcal{L}_{\ell-1}(\theta_{\ell,C}')} \geq 1$, then $1 - \alpha^\ell(\theta_\ell' \,|\, \theta_\ell^n) = 0$. Otherwise, we have

$$\begin{aligned} 1 - \alpha_{\mathrm{ML}}^\ell(\theta_\ell' \,|\, \theta_\ell^n) &= \left(1 - \frac{\mathcal{L}_\ell(\theta_\ell')}{\mathcal{L}_{\ell-1}(\theta_{\ell,C}')}\right) + \left(\frac{\mathcal{L}_\ell(\theta_\ell')\,\mathcal{L}_{\ell-1}(\theta_{\ell,C}^n)}{\mathcal{L}_\ell(\theta_\ell^n)\,\mathcal{L}_{\ell-1}(\theta_{\ell,C}')}\right)\left(1 - \frac{\mathcal{L}_\ell(\theta_\ell^n)}{\mathcal{L}_{\ell-1}(\theta_{\ell,C}^n)}\right) \\ &\leq \left|1 - \frac{\mathcal{L}_\ell(\theta_\ell')}{\mathcal{L}_{\ell-1}(\theta_{\ell,C}')}\right| + \left|1 - \frac{\mathcal{L}_\ell(\theta_\ell^n)}{\mathcal{L}_{\ell-1}(\theta_{\ell,C}^n)}\right|. \end{aligned} \tag{4.15}$$

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^n. Using the definition (4.12) of the likelihood, we have

$$\frac{\mathcal{L}_\ell(\theta_\ell)}{\mathcal{L}_{\ell-1}(\theta_{\ell,C})} \;=\; \exp\left(-\,\frac{\|F_{\mathrm{obs}} - F_\ell(\theta_\ell)\|^2}{\sigma_{F,\ell}^2} \;+\; \frac{\|F_{\mathrm{obs}} - F_{\ell-1}(\theta_{\ell,C})\|^2}{\sigma_{F,\ell-1}^2}\right). \tag{4.16}$$

Denoting $F_\ell := F(\theta_\ell)$ and $F_{\ell-1} := F(\theta_{\ell,C})$, we get as in the proof of Lemma 4.5 that

$$\begin{aligned} \left|\frac{\|F_{\mathrm{obs}} - F_\ell\|^2}{\sigma_{F,\ell}^2} - \frac{\|F_{\mathrm{obs}} - F_{\ell-1}\|^2}{\sigma_{F,\ell-1}^2}\right| \leq\; & \|F_{\mathrm{obs}} - F_{\ell-1}\|^2 \left|\sigma_{F,\ell}^{-2} - \sigma_{F,\ell-1}^{-2}\right| \\ & +\; \frac{2\|F_{\mathrm{obs}} - F_{\ell-1}\| + \|F_\ell - F_{\ell-1}\|}{\sigma_{F,\ell}^2}\|F_\ell - F_{\ell-1}\|. \end{aligned} \tag{4.17}$$

Using the inequality $|1 - \exp(x)| \leq |x|$, for $0 \leq |x| \leq 1$, it follows immediately from (4.17), Assumption A3, Corollary 4.3, Lemma 4.4 and Hölders inequality that

$$\mathbb{E}_{\boldsymbol{\zeta}}\left[\left|1 - \frac{\mathcal{L}_\ell(\theta_\ell)}{\mathcal{L}_{\ell-1}(\theta_{\ell,C})}\right|\right] \;\leq\; C_{k,f,\psi}\left(M_{\ell-1}^{-1/d+\delta} \;+\; R_{\ell-1}^{-1/2+\delta}\right). \tag{4.18}$$

A bound on the expected value of $1 - \alpha_{\mathrm{ML}}^\ell(\theta_\ell' \,|\, \theta_\ell^n)$ now follows from Minkowski's inequality.

The proof in the case of a symmetric proposal distribution is analogous. The bound (4.15) is replaced by

$$1 - \alpha^{\ell}_{\mathrm{ML}}(\theta'_\ell \,|\, \theta^n_\ell) \leq \left| 1 - \frac{\pi^\ell(\theta'_\ell)}{\pi^{\ell-1}(\theta'_{\ell,C})} \right| + \left| 1 - \frac{\pi^\ell(\theta^n_\ell)}{\pi^{\ell-1}(\theta^n_{\ell,C})} \right|.$$

Using the definition of π^ℓ in (4.10), as well as the models (4.11) and (4.12) for the prior and the likelihood, respectively, we have instead of (4.16) that

$$\begin{aligned} \frac{\pi^\ell(\theta_\ell)}{\pi^{\ell-1}(\theta_{\ell,C})} &= \frac{\pi^\ell_0(\theta_\ell)}{\pi^{\ell-1}_0(\theta_{\ell,C})} \frac{\mathcal{L}_\ell(\theta_\ell)}{\mathcal{L}_{\ell-1}(\theta_{\ell,C})} \\ &= \exp\left(-(2\pi)^{-\frac{R_\ell - R_{\ell-1}}{2}} \sum_{j=R_{\ell-1}+1}^{R_\ell} \frac{\xi_j^2}{2} - \frac{\|F_{\mathrm{obs}} - F_\ell(\theta_\ell)\|^2}{\sigma^2_{F,\ell}} + \frac{\|F_{\mathrm{obs}} - F_{\ell-1}(\theta_{\ell,C})\|^2}{\sigma^2_{F,\ell-1}} \right). \end{aligned} \tag{4.19}$$

Since $\sum_{j=R_{\ell-1}+1}^{R_\ell} \xi_j^2$ is χ^2-distributed with $R_\ell - R_{\ell-1}$ degrees of freedom, we have

$$\mathbb{E}_{\rho^\ell_0}\left[\textstyle\sum_j \xi_j^2\right] = 2\frac{\Gamma\left(\frac{1}{2}(R_\ell - R_{\ell-1}) + 1\right)}{\Gamma\left(\frac{1}{2}(R_\ell - R_{\ell-1})\right)} \lesssim R_\ell - R_{\ell-1}.$$

Together with the assumption in (4.14) this implies that the expected value of the additional term in (4.19) is bounded by $R_{\ell-1}^{-1/2+\delta}$. The proof then reduces to that in the pCN case above. □

We will further need the following result.

Lemma 4.8. *For any θ_ℓ, let $k_\ell(\theta_\ell) := \exp\left(\sum_{j=1}^{R_\ell} \sqrt{\mu_j}\phi_j(\theta_\ell)_j\right)$ and $\kappa(\theta_\ell) := \min_{x\in\overline{D}} k_\ell(\cdot, x)$. Then*

$$|p_\ell(\theta_\ell) - p_\ell(\theta^*_\ell)|_{H^1(D)} \lesssim \frac{\|f\|_{H^{-1}(D)}}{\kappa(\theta_\ell)\kappa(\theta^*_\ell)} \|k_\ell(\theta_\ell) - k_\ell(\theta^*_\ell)\|_{\mathcal{C}^0(\overline{D})}, \quad \textit{for almost all } \theta_\ell, \theta^*_\ell, \tag{4.20}$$

and

$$\mathbb{E}_{\rho^\ell_0}\left[|p_\ell(\theta_\ell)|^q_{H^1(D)}\right] \leq \text{constant}, \tag{4.21}$$

for any $q < \infty$, where the hidden constants are independent of ℓ and p_ℓ.

Proof. Using the definition of $\kappa(\theta_\ell)$, as well as the identity

$$\int_D k_\ell(\theta_\ell)\nabla p_\ell(\theta_\ell)\cdot\nabla v\,\mathrm{d}x = \int_D fv\,\mathrm{d}x = \int_D k_\ell(\theta^*_\ell)\,\nabla p_\ell(\theta^*_\ell)\cdot\nabla v\,\mathrm{d}x, \qquad \text{for all } v \in H^1_0(D),$$

we have

$$\begin{aligned} \kappa(\theta_\ell)|p_\ell(\theta_\ell) - p_\ell(\theta^*_\ell)|^2_{H^1(D)} &\leq \int_D k_\ell(\theta_\ell)\,\nabla\left(p_\ell(\theta_\ell) - p_\ell(\theta^*_\ell)\right)\cdot\nabla\left(p_\ell(\theta_\ell) - p_\ell(\theta^*_\ell)\right)\,\mathrm{d}x \\ &\leq \int_D \left(k_\ell(\theta_\ell) - k_\ell(\theta^*_\ell)\right)\,\nabla p_\ell(\theta^*_\ell)\cdot\nabla\left(p_\ell(\theta_\ell) - p_\ell(\theta^*_\ell)\right)\,\mathrm{d}x. \end{aligned}$$

Due to the standard estimate $|p_\ell(\theta^*_\ell)|_{H^1(D)} \leq \|f\|_{H^{-1}(D)}/\kappa(\theta^*_\ell)$, (4.20) follows from an application of the Cauchy-Schwarz inequality, and (4.21) follows from the fact that $\mathbb{E}_{\rho^\ell_0}\left[\kappa(\cdot)^{-q}\right]$ is bounded independent of ℓ ([5, Prop. 3.10]). □

Using Lemmas 4.7 and 4.8, we are now ready to prove Lemma 4.6.

Proof of Lemma 4.6. Let θ_ℓ^{n+1} and $\Theta_{\ell-1}^{n+1}$ be as generated by Algorithm 2 at the $(n+1)$th step, with joint distribution $\nu^{\ell,\ell-1}$. As before, denote the proposal generated for θ_ℓ^n by θ_ℓ'. Firstly, since $\theta_{\ell,C}' = \Theta_{\ell-1}^{n+1}$, it follows from Minkowski's inequality that

$$\begin{aligned}\mathbb{V}_{\nu^{\ell,\ell-1}}\left[Y_\ell^{n+1}\right] &\leq \mathbb{E}_{\nu^{\ell,\ell-1}}\left[\left(Q_\ell(\theta_\ell^{n+1}) - Q_{\ell-1}(\Theta_{\ell-1}^{n+1})\right)^2\right] \\ &\lesssim \mathbb{E}_{\widetilde{\boldsymbol{\zeta}}}\left[\left(Q_\ell(\theta_\ell^{n+1}) - Q_\ell(\theta_\ell')\right)^2\right] + \mathbb{E}_{\zeta_\ell'}\left[\left(Q_\ell(\theta_\ell') - Q_{\ell-1}(\theta_{\ell,C}')\right)^2\right]. \end{aligned} \tag{4.22}$$

Here, $\widetilde{\boldsymbol{\zeta}}$ denotes the joint distribution of θ_ℓ' and θ_ℓ^{n+1} and ζ_ℓ' is the marginal distribution of θ_ℓ'. A bound on the second term follows immediately from Corollary 4.3 and Lemma 4.4, i.e.

$$\mathbb{E}_{\zeta_\ell'}\left[\left(Q_\ell(\theta_\ell') - Q_{\ell-1}(\theta_{\ell,C}')\right)^2\right] \lesssim \mathbb{E}_{\rho_0^\ell}\left[\left(Q_\ell(\theta_\ell') - Q_{\ell-1}(\theta_{\ell,C}')\right)^2\right] \leq C_{k,f,\psi}\left(M_{\ell-1}^{-1/d+\delta} + R_{\ell-1}^{-1+\delta}\right). \tag{4.23}$$

The first term in (4.22) is nonzero only if $\theta_\ell^{n+1} \neq \theta_\ell'$. We will now use Lemmas 4.7 and 4.8, as well as the characteristic function $\mathbb{I}_{\{\theta_\ell^{n+1}\neq\theta_\ell'\}} \in \{0,1\}$ to bound it. Firstly, Hölder's inequality gives

$$\begin{aligned}\mathbb{E}_{\widetilde{\boldsymbol{\zeta}}}\left[\left(Q_\ell(\theta_\ell^{n+1}) - Q_\ell(\theta_\ell')\right)^2\right] &= \mathbb{E}_{\widetilde{\boldsymbol{\zeta}}}\left[\left(Q_\ell(\theta_\ell^{n+1}) - Q_\ell(\theta_\ell')\right)^2 \mathbb{I}_{\{\theta_\ell^{n+1}\neq\theta_\ell'\}}\right] \\ &\leq \mathbb{E}_{\widetilde{\boldsymbol{\zeta}}}\left[\left(Q_\ell(\theta_\ell^{n+1}) - Q_\ell(\theta_\ell')\right)^{2q_1}\right]^{1/q_1} \mathbb{E}_{\widetilde{\boldsymbol{\zeta}}}\left[\mathbb{I}_{\{\theta_\ell^{n+1}\neq\theta_\ell'\}}\right]^{1/q_2}, \end{aligned} \tag{4.24}$$

for any q_1, q_2 s.t. $q_1^{-1} + q_2^{-1} = 1$. Since $\mathcal{G}$ satisfies assumption A2, it follows from Lemmas 4.4 and 4.8 that the term $\mathbb{E}_{\widetilde{\boldsymbol{\zeta}}}\left[\left(Q_\ell(\theta_\ell^{n+1}) - Q_\ell(\theta_\ell')\right)^{2q_1}\right]^{1/q_1}$ in (4.24) can be bounded by a constant independent of ℓ, for any $q_1 < \infty$:

$$\mathbb{E}_{\widetilde{\boldsymbol{\zeta}}}\left[\left(Q_\ell(\theta_\ell^n) - Q_\ell(\theta_\ell')\right)^{2q_1}\right] \lesssim \mathbb{E}_{\nu^\ell}\left[(Q_\ell(\theta_\ell^{n+1}))^{2q_1}\right] + \mathbb{E}_{\zeta_\ell'}\left[(Q_\ell(\theta_\ell'))^{2q_1}\right] \lesssim \mathbb{E}_{\rho_0^\ell}\left[|p_\ell(\theta_\ell)|_{H^1(D)}^{2q_1}\right] \leq \text{constant}.$$

Since $\theta_\ell^{n+1} \neq \theta_\ell'$ only if the proposal θ_ℓ' has been rejected on level ℓ at the $(n+1)$th step, the probability that this happens can be bounded by $\mathbb{E}_{\boldsymbol{\zeta}}[1 - \alpha_{\mathrm{ML}}^\ell(\theta_\ell'|\theta_\ell^n)]$, where the joint distribution $\boldsymbol{\zeta}$ is as in Lemma 4.7. It follows by Lemma 4.7 that

$$\mathbb{E}_{\widetilde{\boldsymbol{\zeta}}}\left[\mathbb{I}_{\{\theta_\ell^{n+1}\neq\theta_\ell'\}}\right] = \mathbb{P}[\theta_\ell^{n+1} \neq \theta_\ell'] \leq C_{k,f,\psi}\left(M_{\ell-1}^{-1/d+\delta} + R_{\ell-1}^{-1+\delta}\right). \tag{4.25}$$

Combining (4.22)-(4.25) 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 4.9. *Under the same assumptions as in Lemma 4.6, Assumptions M1 and M2 of Theorem 3.4 are satisfied, with $\alpha = \beta = 1/d - \delta$ and $\alpha' = \beta' = 1/2 - \delta$, for any $\delta > 0$.*

If we assume that we can obtain individual samples in optimal cost $\mathcal{C}_\ell \lesssim M_\ell \log(M_\ell)$, e.g. via a multigrid solver, we can satisfy Assumption M4 with $\gamma = 1 + \delta$, for any $\delta > 0$. It follows from Theorems 3.4 and 4.9 that we can get the following theoretical upper bounds for the ε-costs of classical and multilevel MCMC applied to model problem (4.2) with log-normal coefficients k:

$$\mathcal{C}_\varepsilon(\widehat{Q}_N^{\mathrm{MC}}) \lesssim \varepsilon^{-(d+2)-\delta} \quad \text{and} \quad \mathcal{C}_\varepsilon(\widehat{Q}_{L,\{N_\ell\}}^{\mathrm{ML}}) \lesssim \varepsilon^{-(d+1)-\delta}, \quad \text{for any } \delta > 0. \tag{4.26}$$

We clearly see the advantages of the multilevel method, which gives a saving of one power of ε^{-1} 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^{-1}, 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 $\beta \approx 1 \approx 2\alpha$, leading to $\mathcal{C}_\varepsilon(\widehat{Q}^{\mathrm{ML}}_{L,\{N_\ell\}}) = \mathcal{O}(\varepsilon^{-2})$. 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 $\mathbb{V}_{\nu^{\ell,\ell-1}}[Y^n_\ell]$, namely the term $\mathbb{E}_{\tilde{\zeta}}\big[\,\big(Q_\ell(\theta^{n+1}_\ell) - Q_\ell(\theta'_\ell)\big)^{2q_1}\,\big]^{1/q_1}$ in (4.24), depends on the difference between $Q_\ell(\theta^{n+1}_\ell)$ and $Q_\ell(\theta'_\ell)$. In the case of the pCN algorithm for the proposal distributions $q^{\ell-1}$ and $q^{\ell,F}_{\mathrm{ML}}$ (as used in Section 5 below) this difference will be small, since θ^n_ℓ 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 Numerics

In this section we describe the implementation details of the MLMCMC algorithm and examine the performance of the method in estimating the posterior expectation of some quantity of interest for our model problem (4.2). We consider (4.2) on the domain $D = (0,1)^2$ with $f \equiv 1$. On the lateral boundaries of the domain we choose Dirichlet boundary conditions; on the top and bottom we choose Neumann conditions:

$$p|_{x_1=0} = 0, \quad p|_{x_1=1} = 1, \quad \frac{\partial p}{\partial \mathbf{n}}\Big|_{x_2=0} = 0 \quad \text{and} \quad \frac{\partial p}{\partial \mathbf{n}}\Big|_{x_2=1} = 0. \tag{5.1}$$

The quantity of interest is the flux across the boundary at $x_1 = 1$, given by

$$Q := -\int_0^1 k \frac{\partial p}{\partial x}\Big|_{x_1=1} dx_2. \tag{5.2}$$

The (prior) permeability field k is modelled as a log-normal random field, with covariance function (4.3) with $r = 1$, $\sigma^2 = 1$ and $\lambda = 0.5$. The log-normal distribution is approximated using truncated KL-expansion (4.4) with an increasing number R_ℓ of terms as ℓ increases. For $r = 1$, the KL eigenfunctions in (4.4) are known explicitly [9].

The model problem is discretised using piecewise linear FEs on a uniform triangular mesh. The coarsest mesh consists of $m_0 + 1$ grid points in each direction, with refined meshes containing $m_\ell + 1 = 2^\ell m_0 + 1$ points, so that the total number of grid points on level ℓ is $M_\ell = (m_\ell + 1)^2$. All our algorithms have been implemented within `freeFEM++` [23]. As the linear solver for the resulting linear equation system for each sample we used `UMFPACK` [13].

5.1 Implementation Details

Let us first define two important quantities for the convergence analysis of Metropolis-Hastings MCMC.

Effective sample size and integrated autocorrelation time. Let $\{\theta^n\}_{n\geq 0}$ be the Markov chain produced by Algorithm 1 and $\widehat{Q}^{\mathrm{MC}}_N$ the resulting MCMC estimator defined in (2.2). The integrated autocorrelation time τ_Q of the correlated samples $Q^n_{M,R} := \mathcal{G}(X(\theta^n))$ produced by Algorithm 1 is defined to be the ratio of the asymptotic variance σ^2_Q of the MCMC estimator $\widehat{Q}^{\mathrm{MC}}_N$, defined in (2.6), and the actual variance $\mathbb{V}_{\nu^{M,R}}[Q_{M,R}]$ of $Q_{M,R}$. If

$$s^2_Q := \frac{1}{N}\sum_{j=0}^{N}\Big(Q^n_{M,R} - \widehat{Q}^{\mathrm{MC}}_N\Big)^2$$

denotes the sample variance, then a good estimate for τ_Q, used e.g. in `R`, is given by $\tau_Q = s^2_Q/\rho(0)$, where $\rho(0)$ is the so-called spectral density at frequency zero. Details of a method for approximating the spectral density are given in [24] (included in `R` under the package `'coda'`). The effective sample

ALGORITHM 3. (Recursive independence sampling)

Choose initial states $\Theta^0_{\ell-1} = \widetilde{\Theta}^0_{\ell-1}, \ldots, \widetilde{\Theta}^0_0$ such that $\widetilde{\Theta}^0_{k,C} = \widetilde{\Theta}^0_{k-1}$ and subsampling rates t_k, for all $k = 1, \ldots, \ell-1$. Then, for $j \geq 0$:

- On level 0:
  - Given $\widetilde{\Theta}^j_0$, generate $\widetilde{\Theta}'_0$ from a pCN proposal distribution.
  - Compute
$$\alpha^0(\widetilde{\Theta}'_0|\widetilde{\Theta}^j_0) = \min\left\{1, \frac{\mathcal{L}_0(F_{\text{obs}}\,|\,\widetilde{\Theta}'_0)}{\mathcal{L}_0(F_{\text{obs}}\,|\,\widetilde{\Theta}^j_0)}\right\}.$$
  - Set $\widetilde{\Theta}^{j+1}_0 = \widetilde{\Theta}'_0$ with probability $\alpha^0(\widetilde{\Theta}'_0|\widetilde{\Theta}^j_0)$. Set $\widetilde{\Theta}^{j+1}_0 = \widetilde{\Theta}^j_0$ otherwise.
- On level $k = 1, \ldots, \ell-1$:
  - Given $\widetilde{\Theta}^j_k$, let $\widetilde{\Theta}'_{k,C} = \widetilde{\Theta}^{(j+1)t_{k-1}}_{k-1}$ and generate $\widetilde{\Theta}'_{k,F}$ from a pCN proposal distribution.
  - Compute
$$\alpha^\ell_{\text{ML}}(\widetilde{\Theta}'_k\,|\,\widetilde{\Theta}^j_k) \;=\; \min\left\{1, \frac{\mathcal{L}_k(F_{\text{obs}}\,|\,\widetilde{\Theta}'_k)\,\mathcal{L}_{k-1}(F_{\text{obs}}\,|\,\widetilde{\Theta}^j_{k,C})}{\mathcal{L}_k(F_{\text{obs}}\,|\,\widetilde{\Theta}^j_k)\,\mathcal{L}_{k-1}(F_{\text{obs}}\,|\,\widetilde{\Theta}'_{k,C})}\right\}.$$
  - Set $\widetilde{\Theta}^{j+1}_k = \widetilde{\Theta}'_k$ with probability $\alpha^k_{\text{ML}}(\widetilde{\Theta}'_k|\widetilde{\Theta}^j_k)$. Set $\widetilde{\Theta}^{j+1}_k = \widetilde{\Theta}^j_k$ otherwise.
- Set $\Theta^{j+1}_{\ell-1} = \widetilde{\Theta}^{(j+1)t_{\ell-1}}_{\ell-1}$.

size is defined as $N^{\text{eff}} := N/\tau_Q$. It represents the number of i.i.d. samples from $\nu_{M,R}$ that would lead to a Monte Carlo estimator with the same variance as $\widehat{Q}^{\text{MC}}_N$.

Recursive independence sampling. The final ingredient for our hierarchical multilevel MCMC algorithm is an efficient practical algorithm to obtain independent samples $\Theta^n_{\ell-1}$ from the coarse posterior $\nu^{\ell-1}$ which we need in Algorithm 2 in Section 3 to estimate $\mathbb{E}_{\nu^\ell}[Q_\ell] - \mathbb{E}_{\nu^{\ell-1}}[Q_{\ell-1}]$. The algorithm is summarised in Algorithm 3.

We start on level 0 by creating a sufficiently long Markov chain $\{\widetilde{\Theta}^j_0\}_{j\geq 0}$ using Algorithm 1 with pCN proposal distribution q^0 [11] (see (5.3) below for details). Let $\widetilde{Q}^j_0 := \mathcal{G}(p_0(\widetilde{\Theta}^j_0))$ be the sample of the output quantity of interest associated with the jth sample of the auxiliary chain $\{\widetilde{\Theta}^j_0\}_{j\geq 0}$ on level 0. The samples in this chain are correlated, but by subsampling it with a sufficiently large rate $t_0 \in \mathbb{N}$, we obtain independent samples. The typical rule in statistics to achieve independence is to choose t_0 to be twice the integrated autocorrelation time $\widetilde{\tau}_0$ of the Markov chain $\{\widetilde{Q}^j_0\}_{j\geq 0}$. In practice, we found that much shorter subsampling rates were sufficient (see below).

Then, on level $0 < k \leq \ell-1$, we use the independent samples created on level $k-1$ in Algorithm 2, to recursively create a Markov chain $\{\widetilde{\Theta}^j_k\}_{j\geq 0}$ on level k. The proposal distribution $q^{k,F}$ for the modes that are added on level k is again chosen to be a pCN random walk (see (5.4) below for details). We subsample this chain again with sufficiently large rate $t_k \in \mathbb{N}$ to obtain independent samples on level k. Finally, we set $\Theta^n_{\ell-1} := \widetilde{\Theta}^{nt_{\ell-1}}_{\ell-1}$. In summary, to produce one independent sample $\Theta^n_{\ell-1}$ on level $\ell-1$, we need to compute $T_k := \prod_{k'=k}^{\ell-1} t_{k'}$ samples, on each of levels $k = 0, \ldots, \ell-1$. Since the

acceptance probability $\alpha^k_{\mathrm{ML}}(\widetilde{\Theta}'_k|\widetilde{\Theta}^j_k)$ converges to 1, as k increases (cf. Lemma 4.7), and since we are using independent proposals from level $k-1$, the integrated autocorrelation times $\widetilde{\tau}_k$ of the auxiliary chains $\{\widetilde{\Theta}^j_k\}_{j\geq 0}$, $k=1,\ldots,\ell-1$, converge to 1, i.e. the samples are essentially independent for large k. As a consequence T_k is actually of the same order as the autocorrelation time of samples that Algorithm 1 with pCN proposals would produce on level k (see below for more details).

At the jth state of the auxiliary chain on level 0, the pCN proposal from the standard multivariate normal prior distribution is generated as follows:

$$(\widetilde{\Theta}'_0)_i = \sqrt{1-\beta_0^2}\,(\widetilde{\Theta}^j_0)_i + \beta_0\,\Psi_i\,, \quad i=1,\ldots,R_0\,. \tag{5.3}$$

Here, $\Psi_i \sim \mathcal{N}(0,1)$ and β_0 is a tuning parameter used to control the size of the step in the pCN random walk [11]. Similarly, the proposal $\widetilde{\Theta}'_{k,F}$ for the fine modes at the jth state of the auxiliary chain on level $k\in\{1,\ldots,L\}$ is generated by

$$(\widetilde{\Theta}'_{k,F})_i = \sqrt{1-\beta_k^2}\,(\widetilde{\Theta}^j_{k,F})_i + \beta_k\,\Psi_i\,, \quad i=1,\ldots,R_k-R_{k-1}\,. \tag{5.4}$$

The actual values of $\beta_k = 0.1$, for all $k=0,\ldots,L$, that are used in all the calculations that follow were chosen after carrying out a series of preliminary tests to achieve "good" mixing properties.

As in (2.2), in practice, the first j^0_k samples from each of the auxiliary chains are discarded by prescribing a "burn-in" period. We choose the length j^0_k of the "burn-in" period on level k to be twice the integrated autocorrelation time $\widetilde{\tau}_k$.

Multilevel estimator. We can now use the independent samples $\Theta^n_{\ell-1}\sim\nu^{\ell-1}$ produced by Algorithm 3 above in Algorithm 2 to produce samples θ^n_ℓ of the fine chain on level ℓ, and thus samples $Y^n_\ell := \mathcal{G}(p_\ell(\theta^n_\ell)) - \mathcal{G}(p_{\ell-1}(\Theta^n_{\ell-1}))$ for the estimator $\widehat{Y}^{\mathrm{MC}}_{\ell,N_\ell}$ of $\mathbb{E}_{\nu^\ell}[Q_\ell] - \mathbb{E}_{\nu^{\ell-1}}[Q_{\ell-1}]$ in (3.3). The samples for the estimator $\widehat{Q}^{\mathrm{MC}}_{0,N_0}$ on level 0 are produced with Algorithm 1 using again pCN-proposals. This completes the definition of the multilevel MCMC estimator $\widehat{Q}^{\mathrm{ML}}_{L,\{N_\ell\}}$ in (3.4). It only remains to decide on an optimal sample size N_ℓ on each level that will ensure that the total sampling error is below the prescribed tolerance and that the total cost of the estimator is minimised.

Let τ_ℓ be the integrated autocorrelation time of the chain Y^n_ℓ (resp. Q^n_0), for $\ell=1,\ldots,L$ (resp. $\ell=0$), and let s^2_ℓ be the sample variance on level ℓ. Then $N^{\mathrm{eff}}_\ell := N_\ell/\tau_\ell$ is the effective sample size on level ℓ and $s^2_\ell/N^{\mathrm{eff}}_\ell$ is an estimate of the variance of the estimator $\widehat{Y}^{\mathrm{MC}}_{\ell,N_\ell}$. Our aim is to achieve the following bound on the total sampling error for the multilevel MCMC estimator:

$$\sum_{\ell=0}^{L} \frac{s^2_\ell}{N^{\mathrm{eff}}_\ell} \leq \frac{\varepsilon^2}{2}, \tag{5.5}$$

for some prescribed tolerance ε. In what follows, we will choose ε such that the bias error on level L is $\frac{\varepsilon^2}{2}$ and thus the two contributions to the mean square error in (3.6) are balanced.

To decide on a cost-optimal strategy for the choice of the N_ℓ, we first need to discuss the cost per sample. Recall that $\mathcal{C}_\ell$ denotes the cost to evaluate Q_ℓ for a single sample Θ_ℓ from the prior on level ℓ. However, to quantify the cost of the estimator $\widehat{Y}^{\mathrm{MC}}_{\ell,N_\ell}$ on level ℓ, we also need to take all the samples in the auxiliary chains on the coarser levels in Algorithm 3 into account, as well as the integrated autocorrelation time τ_ℓ of the chain $\{Y^n_\ell\}$. Recalling that t_k is the subsampling rate on level k in Algorithm 3 and that $T_k = \prod_{k'=k}^{\ell-1} t_{k'}$, the total cost to produce one independent (effective) sample is

$$\mathcal{C}^{\mathrm{eff}}_\ell := \lceil\tau_\ell\rceil \left(\mathcal{C}_\ell + \sum_{k=1}^{\ell-1} T_k\,\mathcal{C}_k\right). \tag{5.6}$$

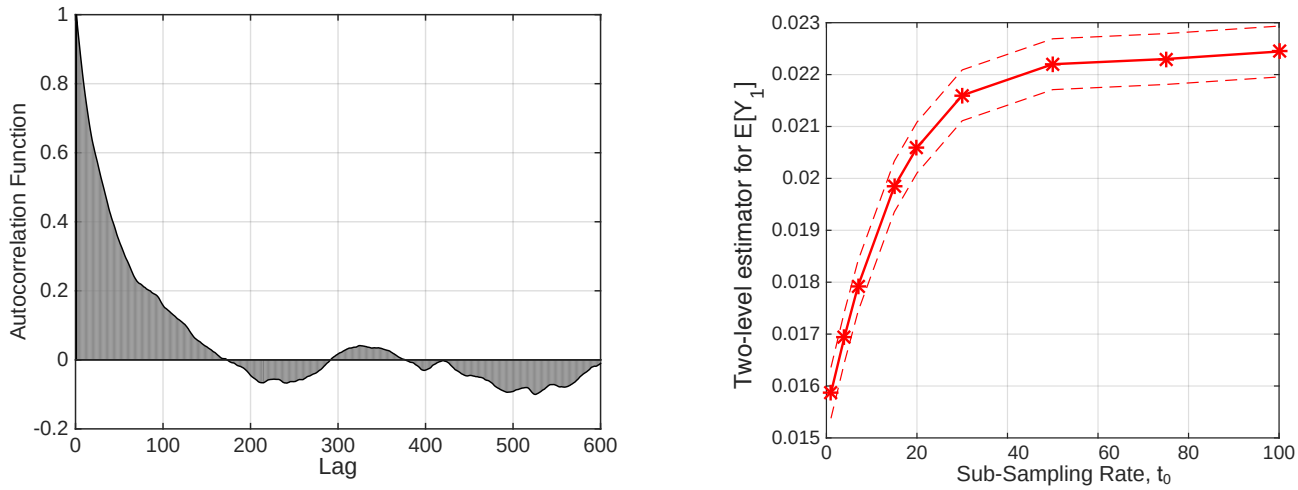


Figure 1: Left: Autocorrelation function for a typical (burnt-in) coarse level chain $\{Q_0^n\}$ with an integrated autocorrelation time of $\tau_0 \approx 86$. Right: Plot of $\mathbb{E}\big[\widehat{Y}_1^{\mathrm{MC}}\big]$ against subsampling rate t_0; the solid line shows the computed results whilst the dashed lines give the two-sided 95% confidence interval.

As in the case of standard multilevel MC with i.i.d. samples, the total cost of the multilevel estimator is minimised, subject to the constraint (5.5), when the effective number of samples on each level satisfies

$$N_\ell^{\mathrm{eff}} = \frac{2}{\varepsilon^2}\left(\sum_{\ell=0}^{L}\sqrt{s_\ell^2 \mathcal{C}_\ell^{\mathrm{eff}}}\right)\sqrt{\frac{s_\ell^2}{\mathcal{C}_\ell^{\mathrm{eff}}}} \tag{5.7}$$

as described in [18, 9]. In practice, the optimal number of samples can be estimated adaptively after an initial number of samples to get an estimate for s_ℓ^2 (see again [18, 9] for standard MLMC).

In all calculations which follow we simultaneously run P parallel chains. This allows for an efficient parallelisation and aids exploration of multi-modal posterior distributions. Furthermore the calculation of the total sampling error (5.5) is simplified. The parallel chains provide P independent estimates for $\widehat{Y}_\ell^{\mathrm{MC}}$. Therefore, using standard statistical tools, the sampling error on each level can be calculated without the need for accurate estimates of the integrated autocorrelation times. For the implementation considered here we chose $P = 128$ and distributed the computations across 128 processors.

5.2 Two-Level Results

We start with a two level test to investigate the additional bias created in Algorithm 2 due to the dependence of the coarse samples from the recursive subsampling procedure in Algorithm 3 and how that bias depends on the subsampling rate t_k. We choose two grids with $m_0 = 8$ and $m_1 = 16$ and fix the numbers of KL modes to be $R_0 = R_1 = 20$. The data is generated synthetically from a single random sample from the prior distribution computed on grid level 4, i.e. with $m_4 = 128$. The observations F_{obs} are taken to be the pressure values at 16 uniformly spaced points interior to the domain. The data fidelity is set to $\sigma_F^2 = 10^{-4}$ on both levels.

We first computed the autocorrelation function for a typical (burnt-in) coarse level chain $\{Q_0^n\}$ (see Fig. 1(left)) and note that the integrated autocorrelation time is approximately $\tau_0 \approx 86$ in this case. We then ran Algorithms 2 and 3, for different subsampling rates from $t_0 = 1$ to $100 > \tau_0$, until the standard error for the estimator $\widehat{Y}_1^{\mathrm{MC}}$ reached a prescribed tolerance of $\varepsilon = 2.5 \times 10^{-4}$. Fig. 1(right) shows the expected value of $\mathbb{E}_{\boldsymbol{\Theta}_1}[\widehat{Y}_1^{\mathrm{MC}}]$ as a function of t_0, as well as the two-sided 95% confidence interval, i.e. $\mathbb{E}_{\boldsymbol{\Theta}_1}[\widehat{Y}_1^{\mathrm{MC}}] \pm 1.96\,\varepsilon$. We note that $\mathbb{E}_{\nu^1}[Q_1] - \mathbb{E}_{\nu^0}[Q_0] \approx \mathbb{E}_{\{\Theta_1^n\}}[\widehat{Q}_1^{\mathrm{MC}}] - \mathbb{E}_{\{\Theta_0^n\}}[\widehat{Q}_0^{\mathrm{MC}}] \approx 0.0222$, calculated from two independent standard MCMC runs to a tolerance of $\varepsilon = 2.5 \times 10^{-5}$ on each level.

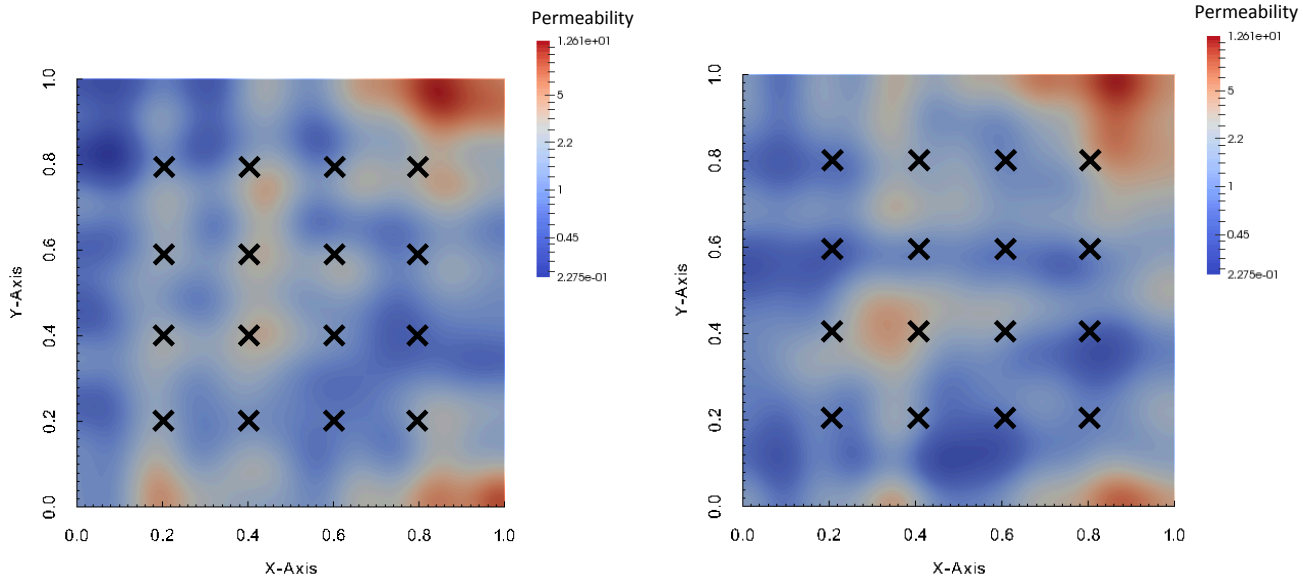


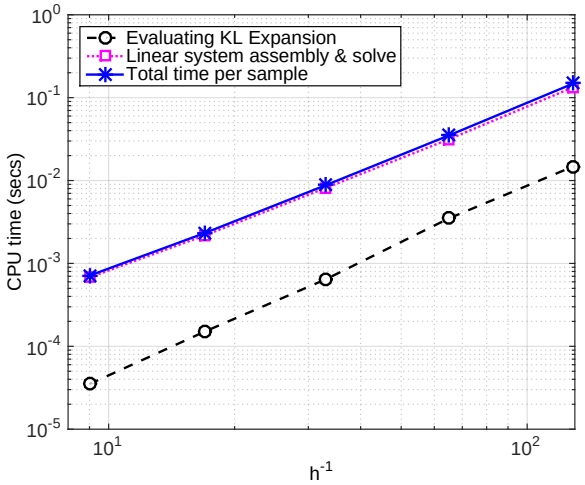
Figure 2: Left: Synthetic data used in Section 5.3. Right: Posterior sample created by our algorithm on grid level 4. For both plots, data points are marked by crosses.

We note that, for the example considered here, the additional bias error due to the dependence of the samples is less than 30% even if no subsampling is used (i.e. $t_0 = 1$). In practice, a value of $t_0 = 50$ would be sufficient to reduce the bias to a negligible amount ($< 1\%$), given all the other bias errors due to FE discretisation, KL truncation and Metropolis-Hastings sampling. However, to be on the safe side for all the calculations that follow we take the subsampling rate equal to the smallest integer that is bigger than our estimate of the integrated autocorrelation time, i.e. $t_\ell = \lceil \widetilde{\tau}_\ell \rceil$.

5.3 Comparison of MLMCMC with a standard single-level MCMC estimator

We now test the full MLMCMC Algorithm, using the same coarsest grid with $m_0 = 8$ and considering up to five levels in our method with a uniformly increasing number of KL modes across the levels from $R_0 = 50$ to $R_4 = 150$. As for the two level example, the data is generated synthetically from a single random sample from the prior distribution on level 4, see Fig. 2(left). We note that since $R_4 = 150$ here, the data differs slightly from that used in the two-level results in Sect. 5.2 (although we used the same random numbers for the first 20 KL modes). The fidelity parameter was again chosen to be $\sigma^2_{F,\ell} = 10^{-4}$, for all $\ell = 0, \ldots, 4$. A typical sample from the posterior distribution on grid level 4, produced by our multilevel algorithm, is shown in Fig. 2(right).

We compare the performance of our new multilevel method to standard Metropolis-Hastings MCMC with pCN proposal distribution (again with tuning parameter $\beta_\ell = 0.1$). The cost $\mathcal{C}_\ell$ to compute one individual sample of Q_ℓ on level ℓ with our code is shown in actual CPU time in Fig. 3(left), obtained on a 2.4GHz Intel Core i7 processor. The cost in `FreeFEM++` is dominated by the assembly of the FE stiffness matrix and so it grows like $\mathcal{O}(h_\ell^{-2}) = \mathcal{O}(M_\ell)$. We believe that this behaviour is representative for problems of this size when the uniform grid structure is not exploited in the assembly process and that these CPU times are competitive. For larger problem sizes, the cost of the linear solver will become the dominant part. However, for the MLMCMC algorithm we are really interested in the cost $\mathcal{C}_\ell^{\text{eff}}$ defined in (5.6) to compute one independent sample on level ℓ using Algorithms 2 and 3 with $t_k = \lceil \widetilde{\tau}_k \rceil$. These times are shown in Fig. 3(right). They are compared to the cost to produce one independent sample on level ℓ using the standard MCMC Algorithm 1. The integrated autocorrelation times $\widetilde{\tau}_\ell$ for the auxiliary chains $\{\widetilde{Q}_\ell^n\}$ on each level in our example are given in Tab. 2. Note that since



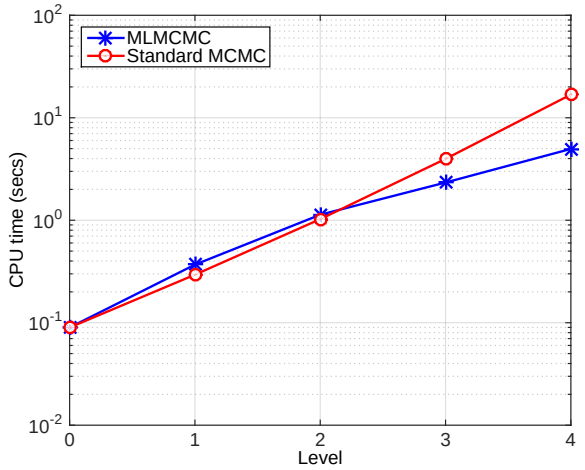


Figure 3: Left: Cost (CPU time in seconds) to compute one sample of Q_h as a function of h. Right: Cost $\mathcal{C}_\ell^{\text{eff}}$ per independent sample on level ℓ.

the coarse samples are (essentially) independent, the integrated autocorrelation times τ_ℓ for the chains $\{Y_\ell^n\}$ are almost identical, i.e. $\tau_\ell \approx \widetilde{\tau}_\ell$.

Level	0	1	2	3	4
$\widetilde{\tau}_\ell$	136.23	3.66	2.93	1.46	1.23

Table 2: Integrated autocorrelation times of the auxiliary chains $\{\widetilde{Q}_\ell^n\}$ on levels $\ell = 0, \ldots, 4$.

In Fig. 4 we now compare the performance of our MLMCMC method with finest level L varying from 1 to 4 with standard MCMC on the same level. The tolerance ε_L for each of the cases is chosen such that the the bias error is less than $\varepsilon_L/\sqrt{2}$, leading to $\varepsilon_1 = 0.04$, $\varepsilon_2 = 0.017$, $\varepsilon_3 = 0.013$ and $\varepsilon_4 = 0.0067$, respectively. The estimated bias error decays with about $\mathcal{O}(h)$ which is faster than what we would expect for the functional in (5.2) which does not satisfy Assumption A2 (see [34]). It is likely that this is because the second term in (4.13), i.e. the bias error in the posterior distribution, dominates. That bias error is due to the FE approximation of pressure evaluations at points here, which are expected to converge with $\mathcal{O}(h \log |h|))$ (see [35]). The slight variation in the convergence rate could mean that some features in the posterior were only picked up on a sufficiently fine grid. The optimal numbers N_ℓ^{eff} of (independent) samples on each level are chosen according to formula (5.7). They are plotted in Fig. 4(left). Please note that these are numbers of independent samples. The total number of samples computed on the coarser levels is much larger. For example, for the four level estimator we needed about 4×10^7 actual PDE solves for all the auxiliary chains on level 0 combined. However, each of these solves is about 250 times cheaper than a solve on level 4. Because $\tau_4 \approx \widetilde{\tau}_4 = 1.23$, we see from Fig. 4(left) that we need only about 562 PDE solves on level 4. These are huge savings against standard MCMC which requires about 4×10^6 solves on level 4 to achieve the same sampling error. We can see this clearly in the overall cost comparison in Fig. 4(right). The gains are even more pronounced if we relax the overly conservative choice of $t_k = \lceil \widetilde{\tau}_k \rceil$ for the subsampling rates.

In our final Fig. 5, we confirm our theoretical results and plot our estimates for $\mathbb{V}_{\nu^{\ell,\ell-1}}\left[Y_\ell^n\right]$ (left) and for $\mathbb{E}_{\boldsymbol{\zeta}}\left[(1 - \alpha_{\text{ML}}^\ell(\theta_\ell'|\theta_\ell^n))\right]$ (right). Ignoring the last data point in each of the plots, which seem to be outliers, the variance seems to converge with almost $\mathcal{O}(h^2)$ and the multilevel rejection probability slightly faster than $\mathcal{O}(h)$. We are not sure whether this means that the bounds in Lemma 4.6 and in Lemma 4.7 are both slightly pessimistic or whether this is just some pre-asymptotic behaviour.

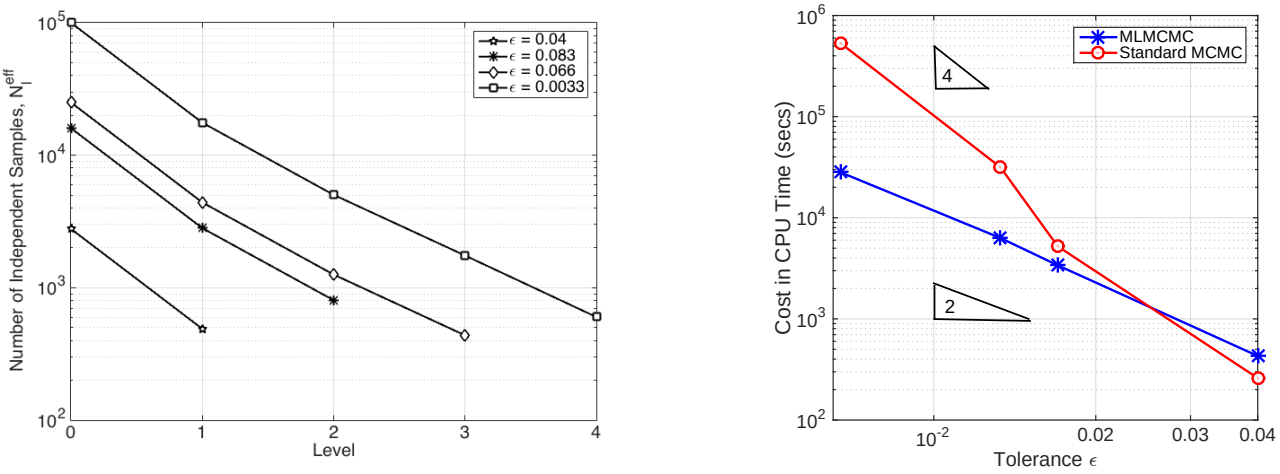


Figure 4: Left: Number of independent samples N_ℓ^{eff} on each level for four different tolerances. Right: Total cost (in seconds) for the multilevel and the single-level estimators plotted against tolerance ϵ.

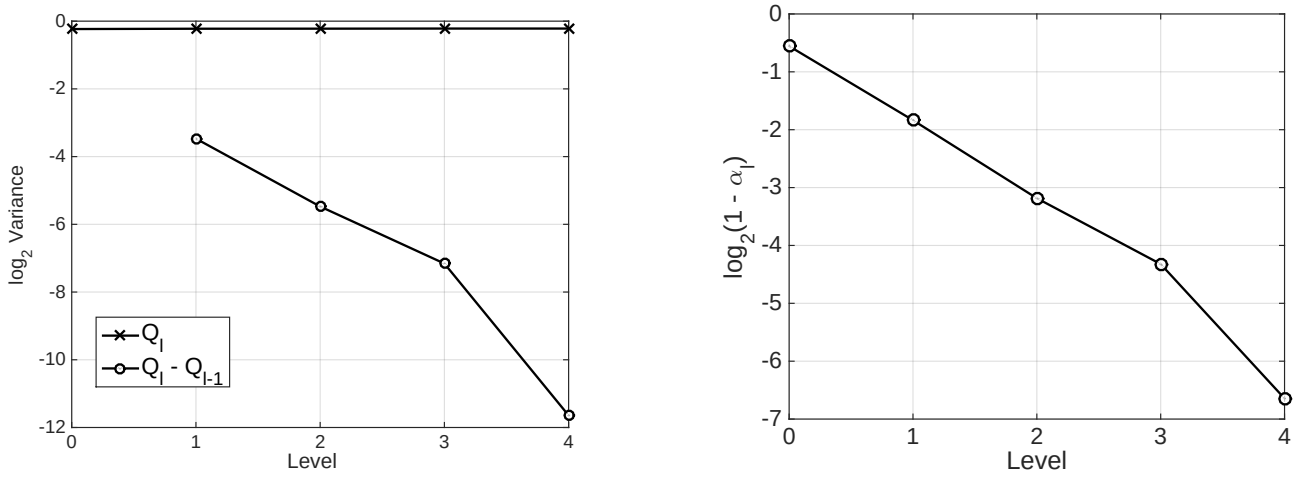


Figure 5: Convergence plots for $\mathbb{V}_{\nu^{\ell,\ell-1}}\left[Y_\ell^n\right]$ and $\mathbb{E}_{\boldsymbol{\zeta}}\left[(1-\alpha_{\text{ML}}^\ell(\theta_\ell'|\theta_\ell^n))\right]$.

Remark 5.1. It is worth to point out that the recursive independence sampling in Algorithm 3 also brings significant savings if used to produce proposals for a standard MCMC algorithm, as the comparison of the cost per independent sample in Fig. 3(right) clearly shows. This is related to the delayed acceptance method of [7]. The multilevel approach also provides a very efficient burn-in method, due to the significantly reduced integrated autocorrelation times on the finer levels and since most of the burn-in happens on the coarsest level. This is related to the approach in [15].

6 Conclusion

Bayesian inverse problems in large scale applications are often too costly to solve using conventional Metropolis-Hastings MCMC algorithms due to the high dimension of the parameter space and the large cost of computing the likelihood. In this paper, we employed a hierarchy of computational models to define a novel multilevel version of a Metropolis-Hastings algorithm, leading to significant reductions in computational cost. The main idea underlying the cost reduction is to build estimators for the difference in the quantity of interest between two successive models in the hierarchy, rather than estimators for

the quantity itself. The new algorithm was then analysed and implemented for a single-phase Darcy flow problem in groundwater modelling, confirming the effectiveness of the algorithm.

The algorithm presented in this paper is not reliant on the specific computational model underlying the simulations, and is generally applicable. The underlying computational model will in general influence the convergence rates α, α', β and β' of the discretisation errors, and the growth rate γ of the cost of the likelihood computation (cf Theorem 3.4), which in turn govern the cost of the standard and multilevel Metropolis-Hastings algorithms. The gain to be expected from employing the multilevel algorithm is always significant, and the gain is in fact larger for more challenging model problems, where the values of α, α', β and β' are small and γ is large.

The algorithm also allows for the use of a variety of proposal distributions. The crucial result in this context is the convergence of the multilevel acceptance probability to 1 (cf. Lemma 4.7), which in general has to be verified for each proposal distribution individually, but is expected to hold for most proposal distributions.

Acknowledgement. Big thanks go to Panayot Vassilevski who initiated and financially supported this work during two visits of Scheichl and Teckentrup at Lawrence Livermore National Labs (LLNL), California. He was involved in most of the original discussions about this method. Christian Ketelsen was postdoctoral researcher under his supervision at LLNL under Contract DE-AC52-07A27344 at the time. We would also like to particularly thank Finn Lindgren and Rob Jack for spotting an error in our original version of Lemma 3.1 and for helping us to find a fix.

References

[1] 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.

[2] A. Brandt, M. Galun, and D. Ron. Optimal multigrid algorithms for calculating thermodynamic limits. *J. Stat. Phys.*, 74(1-2):313–348, 1994.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] J.A. Christen and C. Fox. MCMC using an approximation. *J. Comput. Graph. Stat.*, 14(4):795–810, 2005.

[8] P. G. Ciarlet. *The Finite Element Method for Elliptic Problems*. North–Holland, 1978.

[9] 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.*, 14:3–15, 2011.

[10] 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.

[11] 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.

[12] M. Dashti and A. Stuart. Uncertainty quantification and weak approximation of an elliptic inverse problem. *SIAM J. Numer. Anal.*, 49(6):2524–2542, 2011.

[13] T. A. Davis. Algorithm 832: Umfpack v4.3–an unsymmetric-pattern multifrontal method. *ACM Transactions on Mathematical Software (TOMS)*, 30(2):196–199, 2004.

[14] G. de Marsily. *Quantitative Hydrogeology*. Academic Press, 1986.

[15] Y. Efendiev, T. Hou, and W. Lou. Preconditioning Markov chain Monte Carlo simulations using coarse–scale models. *Water Resourc. Res.*, pages 1–10, 2005.

[16] M.A.R. Ferreira, Z. Bi, M. West, H. Lee, and D. Higdon. Multi-scale Modelling of 1-D Permeability Fields. In *Bayesian Statistics 7*, pages 519–527. Oxford University Press, 2003.

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

[18] M.B. Giles. Multilevel Monte Carlo path simulation. *Oper. Res.*, 256:981–986, 2008.

[19] C.J. Gittelson, J. Könnö, Ch. Schwab, and R. Stenberg. The multilevel Monte Carlo finite element method for a stochastic Brinkman problem. *Numer. Math.*, 125:347–386, 2013.

[20] I.G. Graham, R. Scheichl, and E. Ullmann. Mixed finite element analysis of lognormal diffusion and multilevel Monte Carlo methods. *Stoch. PDE Anal. Comp.*, pages 1–35. published online June 12, 2015.

[21] M. Hairer, A.M. Stuart, and S.J. Vollmer. Spectral gaps for a Metropolis–Hastings algorithm in infinite dimensions. *Ann. Appl. Probab.*, 24(6):2455–2490, 2014.

[22] W.K. Hastings. Monte-Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

[23] F. Hecht. New developments in freeFem++. *J. Numer. Math.*, 20(3-4):251–265, 2012.

[24] P. Heidelberger and P. D. Welch. A spectral method for confidence interval generation and run length control in simulations. *Communications of the ACM*, 24(4):233–245, 1981.

[25] S. Heinrich. Multilevel Monte Carlo methods. volume 2179 of *Lecture notes in Comput. Sci.*, pages 3624–3651. Springer, 2001.

[26] V.H. Hoang, Ch. Schwab, and A.M. Stuart. Complexity analysis of accelerated MCMC methods for Bayesian inversion. *Inverse Probl.*, 29(8):085010, 2013.

[27] R.J. Hoeksema and P.K. Kitanidis. Analysis of the spatial structure of properties of selected aquifers. *Water Resour. Res.*, 21:536–572, 1985.

[28] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *The J. of Chemical Physics*, 21:1087, 1953.

[29] G. Da Prato and J. Zabczyk. *Stochastic equations in infinite dimensions*, volume 44 of *Encyclopedia Math. Appl.* Cambridge University Press, Cambridge, 1992.

[30] C. Robert and G. Casella. *Monte Carlo Statistical Methods*. Springer, 1999.

[31] 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`.

[32] A.M. Stuart. *Inverse problems*, volume 19 of *Acta Num.*, pages 451–559. Cambridge University Press, 2010.

[33] A. L. Teckentrup. Multilevel Monte Carlo methods for highly heterogeneous media. In *Proceedings of the Winter Simulation Conference 2012*, number Article Nr. 32, 2012. Available at `http://informs-sim.org`.

[34] A. L. Teckentrup, R. Scheichl, M. B. Giles, and E. Ullmann. Further analysis of multilevel Monte Carlo methods for elliptic PDEs with random coefficients. *Numer. Math.*, 125(3):569–600, 2013.

[35] A.L. Teckentrup. *Multilevel Monte Carlo methods and uncertainty quantification*. PhD thesis, University of Bath, 2013. Available at `http://people.bath.ac.uk/masrs/Teckentrup_PhD.pdf`.