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

C. Ketelsen[2], R. Scheichl[1] and A.L. Teckentrup[1]

[1] Dept of Mathematical Sciences, University of Bath, Bath BA2 7AY, UK
R.Scheichl@bath.ac.uk, A.L.Teckentrup@bath.ac.uk

[2] Center for Applied Scientific Computing, Lawrence Livermore National Laboratory, P.O. Box 808, L-561, Livermore, CA 94551, U.S.A. ketelsen1@llnl.gov

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 a factor of $\mathcal{O}(10\text{–}50)$ in the ε-cost of the multilevel estimator over the standard Metropolis-Hastings algorithm for tolerances ε around 10^{-3}.

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 unavailable. One way to circumvent this problem, is to generate samples using a Metropolis–Hastings type Markov chain Monte Carlo (MCMC) approach [21, 25, 27], which consists of two main steps: (i) given the previous sample, a new sample is generated according to some proposal distribution, such as a random walk; (ii) the likelihood of this new sample (i.e. the model fit to F_{obs}) is compared to the likelihood of the previous sample. Based on this comparison, the proposed sample is then either accepted and used for inference, or it is rejected and we use instead the previous sample again, leading to a Markov chain. A major problem with MCMC is the high cost of the likelihood calculation for large–scale applications, e.g. in subsurface flow where it involves the numerical solution of a partial differential equation (PDE) with highly varying

*Part of this work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344. LLNL-JRNL-XXXXXX

coefficients on a – for accuracy reasons – very 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 not only for the samples that are eventually used for inference, but also for the samples that end up being rejected. Altogether, this leads to an often impossibly high overall complexity, particularly in the context of high-dimensional parameter spaces (typically needed in subsurface flow applications), where the acceptance rate of the algorithm can be very low.

We show here how the computational cost of the standard Metropolis-Hastings algorithm can be reduced significantly by using a multilevel approach. This has already proved highly successful for subsurface flow problems in the context of standard Monte Carlo estimators based on independent and identically distributed (i.i.d.) samples [9, 1, 18, 6, 30]. The multilevel Monte Carlo (MLMC) method was first introduced by Heinrich for the computation of high-dimensional, parameter-dependent integrals [22], and then rediscovered by Giles [17] in the context of infinite-dimensional integration in stochastic differential equations in finance. Similar ideas were also used by Brandt and his co-workers to accelerate statistical mechanics calculations [2, 3]. The basic ideas are to (i) exploit the linearity of expectation, (ii) introduce a hierarchy of computational models that are assumed to converge (as the model resolution is increased) to some limit model (e.g. the original PDE), and (iii) build estimators for differences of output quantities instead of estimators for 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 on a coarser spatial grid, and thus the evaluation of the likelihood, is computationally much cheaper than on a fine grid. In that way each individual estimator will either have a smaller variance, since differences of output quantities from two consecutive models go to zero with increased model resolution, or it will require significantly less computational work per sample for low model resolutions. Either way the cost of each of 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 will be dominated by the evaluation of the likelihood on the finest level leading to no real cost reduction. Instead, and in order to avoid introducing extra bias in the estimator, we construct 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, 13, 14], 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 [23] on sparse MCMC finite element methods.

The central result of the paper is a complexity theorem (cf. Theorem 3.5) 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 in terms of powers of the tolerance ε. For a particular application in stationary, single phase subsurface flow (with a lognormal permeability prior with exponential covariance), we then verify all the assumptions in Theorem 3.5. 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 4.8), 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 all these theoretical results. In fact, in practice it seems that 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 than 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 for 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 certain, 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 in this section with a review of the standard Metropolis Hastings algorithm, described in a general context. 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. We shall often refer to M as the *discretisation level* of the model.

We consider the setting where we have some real-world dynamic data (or *observations*) F_{obs} available, and want to incorporate this information into our simulation in order to reduce the overall uncertainty. The data F_{obs} usually corresponds to another linear or non-linear functional $\mathcal{F}$ of the model output. 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.

Let us denote the conditional distribution of θ given F_{obs} by $\pi^{M,R}(\theta)$. We assume that as $M, R \to \infty$, we have $\mathbb{E}_{\pi^{M,R}}[Q_{M,R} - Q] \to 0$ for some (inaccessible) random variable Q. We are interested in estimating $\mathbb{E}_{\pi^{M,R}}[Q]$, for M, R sufficiently large. Hence, we compute approximations (or *estimators*) $\widehat{Q}_{M,R}$ of $\mathbb{E}_{\pi^{M,R}}[Q_{M,R}]$. 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 $\pi^{M,R}$. Using Bayes' Theorem, we have

$$\pi^{M,R}(\theta) := \mathcal{P}(\theta \,|\, F_{\text{obs}}) = \frac{\mathcal{L}(F_{\text{obs}} \,|\, \theta)\, \mathcal{P}(\theta)}{\mathcal{P}_F(F_{\text{obs}})} \approx \mathcal{L}(F_{\text{obs}} \,|\, \theta)\, \mathcal{P}(\theta). \tag{2.1}$$

Since the normalising constant $\mathcal{P}_F(F_{\text{obs}})$ is not known in general, the conditional distribution $\pi^{M,R}$ is generally intractable and exact sampling not available.

For the remainder of the paper, we will refer to the conditional distribution $\pi^{M,R}(\theta)$ as the *posterior distribution*, to $\mathcal{L}(F_{\text{obs}} \,|\, \theta)$ as the *likelihood* and to $\mathcal{P}(\theta)$ as the *prior distribution*. The likelihood gives the probability of observing the data F_{obs} given a particular value of θ, and usually involves computing the *model response* $F_{M,R} := \mathcal{F}(X(\theta))$ and comparing this to the observed data F_{obs}. Note that since the model output depends on the discretisation parameter M, the likelihood and hence the posterior distribution $\pi^{M,R}$ will in general also depend on M. As already mentioned, the posterior distribution $\pi^{M,R}$ is usually intractable. In order to generate samples for inference, we will use the Metropolis Hastings MCMC algorithm in Algorithm 1.

Algorithm 1 creates a Markov chain $\{\theta^n\}_{n\in\mathbb{N}}$, and the states θ^n are used in the usual way as samples for inference in a Monte Carlo sampler. The proposal distribution $q(\theta'|\theta^n)$ is what defines the algorithm. A common choice is a simple random walk. However, as outlined in [20], the basic random walk does not lead to dimension R independent convergence, and a better choice

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

$$\alpha^{M,R}\left(\theta'|\theta^n\right) = \min\left\{1, \frac{\pi^{M,R}(\theta')\, q(\theta^n|\theta')}{\pi^{M,R}(\theta^n)\; q(\theta'|\theta^n)}\right\} \tag{2.2}$$

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

is a preconditioned Crank-Nicholson (pCN) algorithm [11]. Below we will see that it is also the crucial ingredient in our multilevel Metropolis-Hastings algorithm. When the proposal distribution is symmetric, i.e. when $q(\theta^n|\theta') = q(\theta'|\theta^n)$, then the formula for $\alpha^{M,R}(\theta'|\theta^n)$ in (2.2) simplifies.

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

$$\widehat{Q}_N^{\mathrm{MC}} := \frac{1}{N}\sum_{n=n_0}^{N+n_0} Q_{M,R}^{(n)} = \frac{1}{N}\sum_{n=n_0}^{N+n_0} \mathcal{G}\left(X(\theta^n)\right), \tag{2.3}$$

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

2.1 Convergence analysis

We will now give a brief overview of the convergence properties of the Metropolis-Hastings algorithm, which we will need below in the analysis of the multilevel variant. For more details we refer the reader, e.g., to [27]. 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 [27].

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

Note that the condition $\mathcal{E} \subset \mathcal{D}$ is 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 $\mathbb{E}_{\pi^{M,R}}\left[|Q_{M,R}|\right] < \infty$ and*

$$q(\theta|\theta') > 0, \text{ for all } (\theta,\theta') \in \mathcal{E}\times\mathcal{E}. \tag{2.4}$$

Then

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

The condition (2.4) 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. [20]). 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 for Theorem 2.2, and the above convergence results hence hold true for any burn–in $n_0 \geq 0$, and for all initial values $\theta^0 \in \mathcal{E}$.

Now that we have established the (asymptotic) convergence of the MCMC estimator (2.3), let us establish a bound on the cost of this estimator. We will quantify the accuracy of our estimator via the root mean square error (RMSE)

$$e(\widehat{Q}_N^{\mathrm{MC}}) := \left(\mathbb{E}_{\boldsymbol{\Theta}}\left[\left(\hat{Q}_N^{\mathrm{MC}} - \mathbb{E}_{\pi^{M,R}}(Q)\right)^2\right]\right)^{1/2}, \tag{2.5}$$

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

Classically, the mean square error (MSE) can be written as the sum of the variance of the estimator and its bias squared,

$$e(\widehat{Q}_N^{\mathrm{MC}})^2 = \mathbb{V}_{\boldsymbol{\Theta}}\left[\widehat{Q}_N^{\mathrm{MC}}\right] + \left(\mathbb{E}_{\boldsymbol{\Theta}}\left[\widehat{Q}_N^{\mathrm{MC}}\right] - \mathbb{E}_{\pi^{M,R}}\left[Q\right]\right)^2.$$

Here, $\mathbb{V}_{\boldsymbol{\Theta}}$ is again the variance with respect to the approximating measure generated by Algorithm 1. Using the triangle inequality and linearity of expectation, we can further write this as

$$e(\widehat{Q}_N^{\mathrm{MC}})^2 \leq \mathbb{V}_{\boldsymbol{\Theta}}\left[\widehat{Q}_N^{\mathrm{MC}}\right] + 2\left(\mathbb{E}_{\boldsymbol{\Theta}}\left[\widehat{Q}_N^{\mathrm{MC}}\right] - \mathbb{E}_{\pi^{M,R}}\left[\widehat{Q}_N^{\mathrm{MC}}\right]\right)^2 + 2\left(\mathbb{E}_{\pi^{M,R}}\left[Q_{M,R} - Q\right]\right)^2 \tag{2.6}$$

The three terms in (2.6) correspond to the three sources of error in the MCMC estimator. The third (and last) term in (2.6) is the discretisation error due to approximating Q by $Q_{M,R}$. 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 sample average and the second term is due to the samples in the estimator not all being perfect (i.i.d.) samples from the target distribution $\pi^{M,R}$.

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 (2.6)) can be described using the following Central Limit Theorem, which can again be found in [27].

Let $\tilde{\theta}^0 \sim \pi^{M,R}$. Then the auxiliary chain $\widetilde{\mathbf{\Theta}} := \{\tilde{\theta}^n\}_{n\in\mathbb{N}}$ constructed by Algorithm 1 starting from $\tilde{\theta}^0$ is stationary, i.e. $\tilde{\theta}^n \sim \pi^{M,R}$ for all $n \geq 0$. Note that the covariance structure of $\widetilde{\mathbf{\Theta}}$ is still implicitly defined by Algorithm 1 as for $\mathbf{\Theta}$. However, now $\mathbb{V}_{\widetilde{\mathbf{\Theta}}}[\tilde{Q}^n_{M,R}] = \mathbb{V}_{\pi^{M,R}}[\tilde{Q}_{M,R}]$ and $\mathbb{E}_{\widetilde{\mathbf{\Theta}}}[\tilde{Q}^n_{M,R}] = \mathbb{E}_{\pi^{M,R}}[\tilde{Q}_{M,R}]$, for any $n \geq 0$, and

$$\mathrm{Cov}_{\pi^{M,R},\pi^{M,R}}\left[\tilde{Q}^0_{M,R},\, \tilde{Q}^n_{M,R}\right] = \mathbb{E}_{\pi^{M,R},\pi^{M,R}}\left[\left(\tilde{Q}^0_{M,R} - \mathbb{E}_{\pi^{M,R}}[Q_{M,R}]\right)\left(\tilde{Q}^n_{M,R} - \mathbb{E}_{\pi^{M,R}}[Q_{M,R}]\right)\right],$$

where $\tilde{Q}^n_{M,R} := \mathcal{G}(X(\tilde{\theta}^n))$ and $\mathbb{E}_{\pi,\pi'}[Z] = \int_{\mathbb{R}^R}\int_{\mathbb{R}^R} Z(\theta,\theta')\,\mathrm{d}\pi(\theta)\,\mathrm{d}\pi'(\theta')$, for a random variable Z that depends on θ and θ'. We now define the so called *asymptotic variance* of the MCMC estimator

$$\sigma^2_Q := \mathbb{V}_{\pi^{M,R}}\left[\tilde{Q}_{M,R}\right] + 2\sum_{n=1}^{\infty}\mathrm{Cov}_{\pi^{M,R},\pi^{M,R}}\left[\tilde{Q}^0_{M,R},\, \tilde{Q}^n_{M,R}\right].$$

Note that stationarity of the chain is assumed only in the definition of σ^2_Q, i.e. for $\widetilde{\mathbf{\Theta}}$, and it is not necessary for the samples $\mathbf{\Theta}$ actually used in the computation of $\widehat{Q}^{\mathrm{MC}}_N$.

Theorem 2.3 (Central Limit Theorem)**.** *Suppose* $\sigma^2_Q < \infty$, (2.4) *holds, and*

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

Then we have

$$\frac{1}{\sqrt{N}}\left(\widehat{Q}^{\mathrm{MC}}_N - \mathbb{E}_{\pi^{M,R}}\left[Q_{M,R}\right]\right) \xrightarrow{D} \mathcal{N}(0,\sigma^2_Q),$$

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

The condition (2.7) is sufficient for the chain $\mathbf{\Theta}$ 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 holds again for any burn-in $n_0 \geq 0$ and any starting value $\theta^0 \in \mathcal{E}$. It shows that asymptotically, the sampling error of the MCMC estimator decays at a similar rate to the sampling error of an estimator based on i.i.d. samples. Note that this includes both sampling errors, and so the constant σ^2_Q is in general larger than in the i.i.d. case where it is simply $\mathbb{V}_{\pi^{M,R}}[Q_{M,R}]$.

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 $N \in \mathbb{N}$,

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

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

Non-asymptotic bounds such as in Assumption A1 are difficult to obtain, but have recently been proved for certain Metropolis–Hastings algorithms, see e.g. [20, 28, 23]. These results require that the chain is sufficiently burnt–in. The hidden constant usually depends on quantities such as the covariances appearing in the asymptotic variance σ^2_Q.

To complete the error analysis, let us now consider the last term in the MSE (2.6), the discretisation bias. As before, we assume $\mathbb{E}_{\pi^{M,R}}\left[Q_{M,R} - Q\right] \to 0$ as $M, R \to \infty$, and we furthermore assume that we have a certain order of convergence, i.e.

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

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 (2.6), (2.8) and (2.9) that the MSE of the MCMC estimator can be bounded by

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

Under the assumption that $\mathbb{V}_{\pi^{M,R}}[Q_{M,R}] \approx$ 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 RMSE of $\mathcal{O}(\varepsilon)$.

Let us now give a bound on the computational cost to achieve this error, the so called *ε-cost*. For this, assume that the cost to compute one sample $Q_{M,R}^n$ satisfies $\mathcal{C}(Q_{M,R}^n) \lesssim M^\gamma$, 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

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

In practical applications, especially in subsurface flow, both the discretisation parameter M and the length of the input random vector R usually need to be very large in order for $\mathbb{E}_{\pi^{M,R}}[Q_{M,R}]$ to be a good approximation to $\mathbb{E}_{\pi^\infty}[Q]$. Moreover, from the analysis above, we see that we need to use a large number of samples N in order to get an accurate MCMC estimator with a small MSE. Since each sample requires the evaluation of the likelihood $\mathcal{L}(F_{\mathrm{obs}}|\theta^n)$, and this is very expensive when M and R are very large, the standard MCMC estimator (2.3) is often extraordinarily expensive in practical situations. Additionally, the acceptance rate of the algorithm can be very low when R is very large. This means that the covariance between the different samples will decay more slowly, which again makes the hidden constant in Assumption A1 larger, and the number of samples we have to take in order to get a certain accuracy increases even further.

To overcome the prohibitively large computational cost of the standard MCMC estimator (2.3), 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 [17, 9].

Let $\{M_\ell : \ell = 0, \ldots, L\}$ 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 $Q_\ell := Q_{M_\ell,R_\ell}$ and the posterior distribution by $\pi^\ell := \pi^{M_\ell,R_\ell}$.

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 π^ℓ depends on ℓ, the new multilevel MCMC (MLMCMC) estimator has to be defined carefully. We will use the identity

$$\mathbb{E}_{\pi^L}[Q_L] \;=\; \mathbb{E}_{\pi^0}[Q_0] \;+\; \sum_{\ell=1}^{L} \mathbb{E}_{\pi^\ell,\pi^{\ell-1}}[Q_\ell - Q_{\ell-1}] \tag{3.2}$$

as a basis, where by the linearity of expectation

$$\begin{aligned}\mathbb{E}_{\pi^\ell,\pi^{\ell-1}}[Q_\ell - Q_{\ell-1}] &= \mathbb{E}_{\pi^\ell}[Q_\ell] \int_{\mathbb{R}^{R_{\ell-1}}} \mathrm{d}\pi^{\ell-1}(\theta_{\ell-1}) \;-\; \mathbb{E}_{\pi^{\ell-1}}[Q_{\ell-1}] \int_{\mathbb{R}^{R_\ell}} \mathrm{d}\pi^\ell(\theta_\ell) \\ &= \mathbb{E}_{\pi^\ell}[Q_\ell] - \mathbb{E}_{\pi^{\ell-1}}[Q_{\ell-1}].\end{aligned} \tag{3.3}$$

The idea of the multilevel estimator is now to estimate each of the terms on the right hand side of (3.2) independently, in a way that minimises the variance of the estimator for a fixed computational cost. In particular, we will estimate each term in (3.2) by an MCMC estimator. The first term $\mathbb{E}_{\pi^0}[Q_0]$ can be estimated using the standard MCMC estimator described in Algorithm 1, i.e. $\widehat{Q}^{\mathrm{MC}}_{0,N_0}$ as in (2.3) with N_0 samples. We need to be more careful in estimating the differences $\mathbb{E}_{\pi^\ell,\pi^{\ell-1}}[Q_\ell - Q_{\ell-1}]$, 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

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

where n_0^ℓ again denotes the burn-in of the estimator and N_ℓ is the number of samples on level ℓ. The main ingredient in this two level estimator is a judicious choice of the two input vectors θ_ℓ^n and $\Theta_{\ell-1}^n$ (see Section 3.1). The full MLMCMC estimator is now defined as

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

where it is important (i) that all the chains that are used to produce the $L+1$ estimators in (3.4) are independent, and (ii) that the two chains $\{\theta_\ell^n\}_{n\in\mathbb{N}}$ and $\{\Theta_\ell^n\}_{n\in\mathbb{N}}$, that are used in $\widehat{Y}^{\mathrm{MC}}_{\ell,N_\ell}$ and in $\widehat{Y}^{\mathrm{MC}}_{\ell+1,N_{\ell+1}}$ respectively, are drawn from the same posterior distribution π^ℓ, so that $\widehat{Q}^{\mathrm{ML}}_{L,\{N_\ell\}}$ is an unbiased estimator of $\mathbb{E}_{\pi^L}[Q_L]$.

There are two main ideas in [17, 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 $\mathbb{V}_{\pi^\ell,\pi^{\ell-1}}[Y_\ell] \to 0$ as $\ell \to \infty$, we need only a small number of samples to obtain a sufficiently accurate estimate of $\mathbb{E}_{\pi^\ell,\pi^{\ell-1}}[Y_\ell]$ on the fine grids, and so the computational effort on the fine grids is also greatly reduced. Here,

$$\mathbb{V}_{\pi^\ell,\pi^{\ell-1}}[Y_\ell] \;:=\; \mathbb{E}_{\pi^\ell,\pi^{\ell-1}}\left[\left(Y_\ell - \mathbb{E}_{\pi^\ell,\pi^{\ell-1}}[Y_\ell]\right)^2\right], \tag{3.5}$$

where the expectation $\mathbb{E}_{\pi^\ell,\pi^{\ell-1}}$ is as in (3.3).

By using the telescoping sum (3.2) and by sampling from the posterior distribution π^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 $\mathbb{V}_{\pi^\ell,\pi^{\ell-1}}[Y_\ell] \to 0$ as $\ell \to \infty$.

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

Let us for the moment fix $1 \leq \ell \leq L$. The challenge is now to generate the chains $\{\theta_\ell^n\}_{n\in\mathbb{N}}$ and $\{\Theta_{\ell-1}^n\}_{n\in\mathbb{N}}$ such that $\mathbb{V}_{\pi^\ell,\pi^{\ell-1}}[Y_\ell]$ is small. To this end, we partition the input vector 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 $\mathbb{V}_{\pi^\ell,\pi^{\ell-1}}[Y_\ell]$ 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 $\pi^{\ell-1}$, and θ_ℓ^n comes from the distribution π^ℓ, this approach is not permissible. We will, however, use this general idea in Algorithm 2.

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

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

- **On level $\ell - 1$:** Given $\Theta^n_{\ell-1}$ generate $\Theta^{n+1}_{\ell-1}$ using Algorithm 1 with some proposal distribution $q^{\ell,C}(\Theta'_{\ell-1} \,|\, \Theta^n_{\ell-1})$ and acceptance probability

$$\alpha^{\ell,C}(\Theta'_{\ell-1} \,|\, \Theta^n_{\ell-1}) \;=\; \min\left\{1, \frac{\pi^{\ell-1}(\Theta'_{\ell-1})\, q^{\ell,C}(\Theta^n_{\ell-1}|\Theta'_{\ell-1})}{\pi^{\ell-1}(\Theta^n_{\ell-1})\, q^{\ell,C}(\Theta'_{\ell-1}|\Theta^n_{\ell-1})}\right\}.$$

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

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

The coarse sample $\Theta^{n+1}_{\ell-1}$ is generated using the standard MCMC algorithm given in Algorithm 1, using, e.g., a random walk or the pCN proposal distribution [11] for $q^{\ell,C}$. Based on the outcome on level $\ell - 1$, we then generate θ^{n+1}_ℓ, using a new two-level proposal distribution in conjunction with the usual accept/reject step from Algorithm 1. The proposal distribution $q^{\ell,F}$ for the fine modes in that step can again be via a simple random walk or the pCN algorithm.

At each step in Algorithm 2, there are four different outcomes, depending on whether we accept on both, one or none of the levels. The different possibilities are given in Table 1. Observe that when we accept on level ℓ, we always 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. They will definitely differ if we accept on level $\ell - 1$ and reject on level ℓ. If both proposals are rejected then it depends on the decision made at the previous state whether the coarse modes differ or not. In general, this

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

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

"divergence" of the coarse modes may mean that $\mathbb{V}_{\pi^\ell,\pi^{\ell-1}}[Y_\ell]$ 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 $\mathcal{L}(F_{\text{obs}}|\theta)$, we can guarantee that $\alpha^\ell(\theta'_\ell|\theta^n_\ell) \to 1$ and thus that $\mathbb{V}_{\pi^\ell,\pi^{\ell-1}}[Y_\ell] \to 0$ as $\ell \to \infty$. We will show this for a subsurface flow application in Section 4.

The specific proposal distribution q^ℓ in Algorithm 2 can be computed very easily and at no additional cost, leading to a simple formula for the "two-level" acceptance probability α^ℓ.

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

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

If we further suppose that the proposal distributions $q^{\ell,C}$ and $q^{\ell,F}$ are symmetric, then

$$\alpha^{\ell,C}(\Theta'_{\ell-1} \,|\, \Theta^n_{\ell-1}) = \min\left\{1, \frac{\pi^{\ell-1}(\Theta'_{\ell-1})}{\pi^{\ell-1}(\Theta^n_{\ell-1})}\right\} \quad \textit{and} \quad \alpha^\ell(\theta'_\ell \,|\, \theta^n_\ell) = \min\left\{1, \frac{\pi^\ell(\theta'_\ell)\,\pi^{\ell-1}(\theta^n_{\ell,C})}{\pi^\ell(\theta^n_\ell)\,\pi^{\ell-1}(\theta'_{\ell,C})}\right\}.$$

Proof. Let θ^a_ℓ and θ^b_ℓ be any two admissible states on level ℓ. Since the proposals for the coarse modes $\theta_{\ell,C}$ and for the fine modes $\theta_{\ell,F}$ are generated independently, the transition probability $q^\ell(\theta^b_\ell\,|\theta^a_\ell)$ can be written as a product of transition probabilities on the two parts of θ_ℓ. For the coarse level transition probability, we have to take into account the decision that was made on level $\ell-1$. Hence,

$$q^\ell(\theta^b_\ell|\theta^a_\ell) = \alpha^{\ell,C}(\theta^b_{\ell,C}|\theta^a_{\ell,C})\,q^{\ell,C}(\theta^b_{\ell,C}|\theta^a_{\ell,C})\,q^{\ell,F}(\theta^b_{\ell,F}|\theta^a_{\ell,F}). \tag{3.6}$$

and so

$$\frac{q^\ell(\theta^a_\ell|\theta^b_\ell)}{q^\ell(\theta^b_\ell|\theta^a_\ell)} = \frac{\min\left\{1, \frac{\pi^{\ell-1}(\theta^a_{\ell,C})q^{\ell,C}(\theta^b_{\ell,C}|\theta^a_{\ell,C})}{\pi^{\ell-1}(\theta^b_{\ell,C})q^{\ell,C}(\theta^a_{\ell,C}|\theta^b_{\ell,C})}\right\} q^{\ell,C}(\theta^a_{\ell,C}|\theta^b_{\ell,C})q^{\ell,F}(\theta^a_{\ell,F}|\theta^b_{\ell,F})}{\min\left\{1, \frac{\pi^{\ell-1}(\theta^b_{\ell,C})q^{\ell,C}(\theta^a_{\ell,C}|\theta^b_{\ell,C})}{\pi^{\ell-1}(\theta_{\ell,C})q^{\ell,C}(\theta^b_{\ell,C}|\theta^a_{\ell,C})}\right\} q^{\ell,C}(\theta^b_{\ell,C}|\theta^a_{\ell,C})q^{\ell,F}(\theta^b_{\ell,F}|\theta^a_{\ell,F})} = \frac{\pi^{\ell-1}(\theta^a_{\ell,C})q^{\ell,F}(\theta^a_{\ell,F}|\theta^b_{\ell,F})}{\pi^{\ell-1}(\theta^b_{\ell,C})q^{\ell,F}(\theta^b_{\ell,F}|\theta^a_{\ell,F})}.$$

This completes the proof of the first result, if we choose $\theta^a_\ell := \theta^n_\ell$ and $\theta^b_\ell := \theta'_\ell$. The corollary for symmetric distributions $q^{\ell,C}$ and $q^{\ell,F}$ follows by definition. □

Remark 3.2 (Recursive algorithm)**.** Note that one particular choice for the coarse level proposal distribution in Step 1 of Algorithm 2 on each of the levels $\ell \geq 1$ is $q^{\ell,C} := q^{\ell-1}$, i.e. the "two-level" proposal distribution defined in Step 2 of Algorithm 2 on level $\ell-1$. We can apply this strategy recursively on every level and set q^0 to be, e.g., the pCN algorithm [11]. So proposals for $Q_{\ell-1}$ and for Q_ℓ get "pre-screened" at all coarser levels, starting always at level 0. The formula for the acceptance probability α^ℓ in Lemma 3.1 does not depend on $q^{\ell,C}$ and so it remains the same. However, this choice did not prove advantageous in practice. It requires $\ell+1$ evaluations of the likelihood on level ℓ instead of two and it does not improve the acceptance probability. Instead, we found that choosing the pCN algorithm for $q^{\ell,C}$ (as well as for $q^{\ell,F}$) worked better.

A simplified version of Algorithm 2, making use of the symmetry of the pCN proposal distribution and of the formulae derived in Lemma 3.1, is given in Section 5 and will be used for the numerical computations.

3.2 Convergence analysis

Let us now move on to convergence properties of the multilevel estimator. As in Section 2.1, let

$$K_\ell(\theta_\ell \,|\, \theta'_\ell) := \alpha^\ell(\theta_\ell \,|\, \theta'_\ell)\,q^\ell(\theta_\ell \,|\, \theta'_\ell) + \left(1 - \int_{\mathbb{R}^{R_\ell}} \alpha^\ell(\theta_\ell \,|\, \theta'_\ell)\,q^\ell(\theta_\ell \,|\, \theta'_\ell)\,\mathrm{d}\theta'_\ell\right)\delta(\theta_\ell - \theta'_\ell),$$

denote the transition kernel of $\{\theta^n_\ell\}_{n\in\mathbb{N}}$, and define, for all $\ell = 0, \ldots, L$, the sets

$$\mathcal{E}^\ell = \{\theta_\ell : \pi^\ell(\theta_\ell) > 0\},$$
$$\mathcal{D}^\ell = \{\theta_\ell : q^\ell(\theta_\ell \,|\, \theta'_\ell) > 0 \text{ for some } \theta'_\ell \in \mathcal{E}^\ell\}.$$

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

Lemma 3.3. *Provided $\mathcal{E}^\ell \subset \mathcal{D}^\ell$, π^ℓ is a stationary distribution of the chain $\{\theta_\ell^n\}_{n\in\mathbb{N}}$.*

Theorem 3.4. *Suppose that for all $\ell = 0, \dots, L$, $\mathbb{E}_{\pi^\ell}[|Q_\ell|] < \infty$ and*

$$q^\ell(\theta_\ell \,|\, \theta'_\ell) > 0, \quad \text{for all } \ (\theta_\ell, \theta'_\ell) \in \mathcal{E}^\ell \times \mathcal{E}^\ell. \tag{3.7}$$

Then

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

Let us have a closer look at the irreducibility condition (3.7). As in (3.6), we have

$$q^\ell(\theta_\ell|\theta'_\ell) = \alpha^{\ell,C}(\theta_{\ell,C}|\theta'_{\ell,C})\, q^{\ell,C}(\theta_{\ell,C}|\theta'_{\ell,C})\, q^{\ell,F}(\theta_{\ell,F}|\theta'_{\ell,F})$$

and thus (3.7) holds, if and only if, for all $(\theta_\ell, \theta'_\ell) \in \mathcal{E}^\ell \times \mathcal{E}^\ell$, $\pi^{\ell-1}(\theta_{\ell,C})$, $q^{\ell,C}(\theta'_{\ell,C}|\theta_{\ell,C})$, $q^{\ell,C}(\theta_{\ell,C}|\theta'_{\ell,C})$ and $q^{\ell,F}(\theta_{\ell,F}|\theta'_{\ell,F})$ are all positive. The final three terms are positive for common choices of proposal distributions, such as the random walk sampler or the pCN algorithm. The first term can also be assured to be positive by appropriate choices for the likelihood and prior 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 define $\mathbf{\Theta}_\ell := \{\theta_\ell^n\}_{n\in\mathbb{N}} \cup \{\Theta_{\ell-1}^n\}_{n\in\mathbb{N}}$, for $\ell \geq 1$, and $\mathbf{\Theta}_0 := \{\theta_0^n\}_{n\in\mathbb{N}}$.

Theorem 3.5. *Let $\varepsilon < \exp[-1]$ and suppose there are positive constants $\alpha, \alpha', \beta, \beta', \gamma > 0$ such that $\alpha \geq \frac{1}{2}\min(\beta, \gamma)$ and $R_\ell \gtrsim M_\ell^{\max\{\alpha/\alpha', \beta/\beta'\}}$. Under the following assumptions,*

M1. $|\mathbb{E}_{\pi^\ell}[Q_\ell - Q]| \ \lesssim \ (M_\ell^{-\alpha} + R_\ell^{-\alpha'})$

M2. $\mathbb{V}_{\pi^\ell, \pi^{\ell-1}}[Y_\ell] \ \lesssim \ M_{\ell-1}^{-\beta} + R_{\ell-1}^{-\beta'}$

M3. $\mathbb{V}_{\mathbf{\Theta}_\ell}[\widehat{Y}^{\mathrm{MC}}_{\ell,N_\ell}] + (\mathbb{E}_{\mathbf{\Theta}_\ell}[\widehat{Y}^{\mathrm{MC}}_{\ell,N_\ell}] - \mathbb{E}_{\pi^\ell,\pi^{\ell-1}}[\widehat{Y}^{\mathrm{MC}}_{\ell,N_\ell}])^2 \ \lesssim \ N_\ell^{-1}\, \mathbb{V}_{\pi^\ell,\pi^{\ell-1}}[Y_\ell]$

M4. $\mathcal{C}_\ell \ \lesssim \ M_\ell^\gamma$,

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

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

and

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

Proof. The proof of this theorem is very similar to the proof of the complexity theorem in the case of multilevel estimators based on i.i.d samples (cf. [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, similar to (2.6), we can expand

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

Since the second term in the MSE above can be bounded by

$$\begin{aligned}\left(\mathbb{E}_{\cup_\ell \boldsymbol{\Theta}_\ell}\left[\widehat{Q}^{\mathrm{ML}}_{L,\{N_\ell\}}\right] - \mathbb{E}_{\pi^L}\left[\widehat{Q}^{\mathrm{ML}}_{L,\{N_\ell\}}\right]\right)^2 &= \left(\sum_{l=0}^{L}\left(\mathbb{E}_{\boldsymbol{\Theta}_\ell}\left[\widehat{Y}^{\mathrm{MC}}_{\ell,N_\ell}\right] - \mathbb{E}_{\pi^\ell,\pi^{\ell-1}}[\widehat{Y}^{\mathrm{MC}}_{\ell,N_\ell}]\right)\right)^2 \\ &\leq (L+1)\sum_{l=1}^{L}\left(\mathbb{E}_{\boldsymbol{\Theta}_\ell}\left[\widehat{Y}^{\mathrm{MC}}_{\ell,N_\ell}\right] - \mathbb{E}_{\pi^\ell,\pi^{\ell-1}}[\widehat{Y}^{\mathrm{MC}}_{\ell,N_\ell}]\right)^2,\end{aligned}$$

where we have set $Y_0 := Q_0$ and $\mathbb{E}_{\pi^0,\pi^{-1}}[\widehat{Y}^{\mathrm{MC}}_{0,N_0}] := \mathbb{E}_{\pi^0}[\widehat{Q}^{\mathrm{MC}}_{0,N_0}]$, it follows from Assumption M3 that

$$e(\widehat{Q}^{\mathrm{ML}}_{L,\{N_\ell\}})^2 \lesssim (L+1)\sum_{\ell=0}^{L} N_\ell^{-1}\,\mathbb{V}_{\pi^\ell,\pi^{\ell-1}}[Y_\ell] + \Big(\mathbb{E}_{\pi^L}[Q_L - Q]\Big)^2. \tag{3.8}$$

In contrast to the MSE for multilevel estimators based on i.i.d samples, we hence have a factor $(L+1)$ multiplying the sampling error term on the right hand side of (3.8). This implies that in order to make this term less than $\varepsilon^2/2$, the number of samples N_ℓ needs to be increased by a factor of $(L+1)$ compared to the i.i.d. case. The cost of the multilevel estimator is correspondingly also increased by a factor of $(L+1)$. The remainder of the proof remains identical.

Since L is chosen such that the second term in (3.8) (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. □

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. [12, 10]):

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

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 are the source terms.

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 $p(\omega, x) = p_0(x)$ on ∂D, and assume that the boundary conditions p_0 and the sources 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 typical in subsurface flow.

When modelling a whole aquifer, a whole oil reservoir, or a sufficiently large region around a potential radioactive waste repository, the correlation length scale for 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, and faithful models should therefore also only assume limited spatial regularity of k. A covariance function that has been proposed in the application literature (cf. [24]) 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. This choice of covariance function implies that k is *homogeneous* and it follows from Kolmogorov's theorem [26] that $k(\cdot, \omega) \in C^{0,t}(D)$ a.s., for any $t < 1/2$.

For the purpose of this paper, we will for the remainder of this section 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, 30, 29].

Let us briefly 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 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 (e.g. grid nodes for standard piecewise linear finite elements) 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 $X_{M,R}$ will contain approximate values of the pressure p at M given points in the spatial domain D.

In order to apply the proposed MCMC methods to model problem (4.2), we hence need to represent the permeability k in terms of a vector θ 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. [16]. We will here only mention that the eigenvalues $\{\mu_n\}_{n\in\mathbb{N}}$ are all 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 a finite number R of terms gives an approximation of k in

terms of R standard normal random variables,

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

The coefficients $\{\xi_n\}_{n=1}^R$ will be our input random vector θ in the MCMC algorithms. 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 $\theta_\ell := \{\xi_n\}_{n=1}^{R_\ell}$.

For the spatial discretisation of model problem (4.2), we will use standard, continuous, piecewise linear finite elements (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 [19] 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. A sequence of discretisation levels M_ℓ satisfying (3.1) can then be constructed by choosing a coarsest mesh width h_0, and choosing $h_\ell := s^{-\ell}h_0$. A common (but not necessarily the optimal) choice is $s = 2$ and uniform refinement between the levels. We will denote the finite element solution on level ℓ by p_ℓ.

Let us finally specify the prior distribution and likelihood model that we will assume for the remainder of this paper. The prior distribution $\mathcal{P}_\ell$ of θ_ℓ is simply a standard R_ℓ-dimensional Gaussian:

$$\mathcal{P}_\ell(\theta_\ell) \propto \frac{1}{(2\pi)^{R_\ell/2}} \exp\left[-\sum_{j=1}^{R_\ell} \frac{\xi_j^2}{2}\right]. \tag{4.4}$$

For the likelihood we also choose a normal distribution, centred around the model response $F^\ell(\theta_\ell) = \mathcal{F}(p_\ell(\theta_\ell))$ and with variance $\sigma_{F,\ell}^2$:

$$\mathcal{L}_\ell(F_{\text{obs}} \,|\, \theta_\ell) \propto \exp\left[\frac{-\|F_{\text{obs}} - F^\ell(\theta_\ell)\|^2}{2\sigma_{F,\ell}^2}\right]. \tag{4.5}$$

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 π^L on the finest level L. Hence, the posterior distributions on the coarser levels π^ℓ, $\ell = 0, \ldots, L-1$, do not have to model the measured data as faithfully as π^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, since it is easier to match the model response $F^\ell(\theta_\ell)$ with the data F_{obs}. As we will see below (cf. Assumption A3), the growth in $\sigma_{F,\ell}^2$ has to be controlled.

4.1 Convergence analysis

We now perform a rigorous convergence analysis of the MLMCMC estimator $\widehat{Q}^{\text{ML}}_{L,\{N_\ell\}}$ introduced in Section 3 to the described model problem (4.1). Using Theorem 3.4, we will first verify that indeed this multilevel estimator is an unbiased estimator of $\mathbb{E}_{\pi^L}[Q_L]$, before we go on to quantify its computational cost by verifying the assumptions of Theorem 3.5.

To conclude that the multilevel estimator converges to the correct expected value $\mathbb{E}_{\pi^L}[Q_L]$ as the number of samples tends to infinity, we only need to verify the irreducibility condition (3.7) in Theorem 3.4. As already noted in Section 3, for common choices of proposal distribution, the condition holds true if we have $\pi^{\ell-1}(\theta_{\ell,C}) > 0$ for all θ_ℓ s.t. $\pi^\ell(\theta_\ell) > 0$. Since both the prior and

the likelihood were chosen as normal distributions, and normal distributions have infinite support, the conclusion then follows.

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

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

Let us now move on to quantifying the cost of the multilevel estimator, and verify that the assumptions in Theorem 3.5 hold for our model problem. We will prove M1 and M2. 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. [28, 20]. We will also not address M4, which is an assumption on the cost of obtaining one sample of Q_ℓ. In the best case, with an optimal linear solver to solve the discretised (finite element) equations for each sample, M4 is satisfied with $\gamma = 1$.

Since they will become useful later, let us recall some of the main results in the convergence analysis of multilevel Monte Carlo estimators based on independent and identically distributed (i.i.d.) samples, rather than samples generated by Algorithm 2. An extensive convergence analysis of finite element multilevel estimators based on i.i.d. samples for model problem (4.1) with log–normal coefficients can be found in [6, 30, 29]. We firstly have the following result on the convergence of the finite element 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*

$$\mathbb{E}_{\mathcal{P}_\ell}\left[|p - p_\ell|^q_{H^1(D)}\right]^{1/q} \leq C_{k,f,p_0,q}\,(M_\ell^{-1/2d+\delta} + R_\ell^{-1/2+\delta}),$$

for any $q < \infty$ and $\delta > 0$, where the (generic) constant $C_{k,f,p_0,q}$ (here and below) depends on the data k, f, p_0 and on q, but is independent of any other parameters.

Proof. This follows from [30, Proposition 4.1]. □

Convergence results for functionals $\mathcal{G}$ 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 can easily be extended to any continuously Frèchet differentiable functional (see [30, §3.2]). We make the following assumption on the functional $\mathcal{G}$ (cf. Assumption F1 in [30]).

A2. For given $\omega \in \Omega$, let $\mathcal{G} : H^1(D) \to \mathbb{R}$ be linear, and suppose that, for any $q < \infty$, there exists $C_\mathcal{G} \in L^q(\Omega)$, such that

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

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

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

$$\mathbb{E}_{\mathcal{P}_\ell}[|\mathcal{G}(p) - \mathcal{G}(p_\ell)|^q]^{1/q} \leq C_{k,f,p_0,q}\left(M_\ell^{-1/d+\delta} + R_\ell^{-1/2+\delta}\right),$$

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

Proof. This follows from [30, Corollary 4.1]. □

Note that assumption A2 is crucial in order to get the faster convergence rates of the spatial discretisation error in Lemma 4.3. For multilevel estimators based on i.i.d. samples, it follows immediately from Lemma 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 [30] for details).

The aim is now to generalise the result in Lemma 4.2 to include the framework of the new MLMCMC estimator. There are two issues which need to be addressed. Firstly, the bounds in assumptions M1 and M2 in Theorem 3.5 involve moments with respect to the posterior distributions π^ℓ, which are not known explicitly, but are related to the prior distributions $\mathcal{P}_\ell$ through Bayes' Theorem. Secondly, the samples which are used to compute the differences $Q_\ell^n - Q_{\ell-1}^n$ are generated by Algorithm 2, and may differ not only due to the truncation order, but also because they come from different Markov chains (i.e. $\Theta_{\ell-1}^n$ is not necessarily equal to $\theta_{\ell,C}^n$, as seen in Table 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 π^ℓ to moments with respect to the prior distribution $\mathcal{P}_\ell$.

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

$$|\,\mathbb{E}_{\pi^\ell}[Z^q]\,| \lesssim \mathbb{E}_{\mathcal{P}_\ell}[|Z|^q]\,.$$

Similarly, for any random variable $Z = Z(\theta_\ell, \Theta_{\ell-1})$ and for any q s.t. $\mathbb{E}_{\mathcal{P}_\ell,\mathcal{P}_{\ell-1}}[|Z|^q] < \infty$, we have

$$\left|\,\mathbb{E}_{\pi^\ell,\pi^{\ell-1}}[Z^q]\,\right| \lesssim \mathbb{E}_{\mathcal{P}_\ell,\mathcal{P}_{\ell-1}}[|Z|^q]\,.$$

Proof. Using Bayes' Theorem (2.1), we have

$$|\,\mathbb{E}_{\pi^\ell}[Z^q]\,| = \left|\int_{\mathbb{R}^{R_\ell}} Z^q(\theta_\ell)\frac{\mathcal{L}_\ell(F_{\text{obs}}\,|\,\theta_\ell)\,\mathcal{P}_\ell(\theta_\ell)}{\mathcal{P}_F(F_{\text{obs}})}\,\mathrm{d}\theta_\ell\right| \leq \frac{\sup_{\theta_\ell}[\mathcal{L}_\ell(F_{\text{obs}}\,|\,\theta_\ell)]}{\mathcal{P}_F(F_{\text{obs}})}\int_{\mathbb{R}^{R_\ell}}|Z(\theta_\ell)|^q\,\mathcal{P}_\ell(\theta_\ell)\,\mathrm{d}\theta_\ell.$$

Since the likelihood $\mathcal{L}_\ell$ is not a discrete probability measure, so that $\sup_{\theta_\ell}[\mathcal{L}_\ell(F_{\text{obs}}\,|\,\theta_\ell)] < \infty$, the first claim of the Lemma then follows, since $\mathcal{P}_F(F_{\text{obs}})$ is a constant. The second claim of the Lemma can be proved analogously. □

Note that it follows immediately from Lemmas 4.3 and 4.4 and the linearity of expectation that assumption M1 in Theorem 3.5 is satisfied, with $\alpha = 1/d - \delta$ and $\alpha' = 1/2 - \delta$, for any $\delta > 0$. In order to prove M2, we further have to analyse the situation where the two samples θ_ℓ^n and $\Theta_{\ell-1}^n$ used to compute Y_ℓ^n "diverge", i.e. when $\Theta_{\ell-1}^n \neq \theta_{\ell,C}^n$.

We need to make the following two assumptions on the parameters $\sigma_{F,\ell}^2$ in the likelihood model (4.5) and on the growth of the dimension R_ℓ.

A3. The dimension $R_\ell \to \infty$ as $\ell \to \infty$ and

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

A4. The sequence of fidelity parameters $\{\sigma_{F,\ell}^2\}_{\ell=0}^\infty$ satisfies

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

For A3 to be satisfied it suffices that $R_\ell - R_{\ell-1}$ grows logarithmically with $R_{\ell-1}$. Assumption A4 holds for example, if we choose the fidelity parameter to be constant for all $\ell \geq \ell_0$, for some $\ell_0 \geq 0$.

Under these assumptions we can now prove that assumption M2 in Theorem 3.5 is satisfied, with $\beta = 1/d - \delta$ and $\beta' = 1/2 - \delta$, for any $\delta > 0$.

Lemma 4.5. *For $n \in \mathbb{N}$, let θ_ℓ^n and $\Theta_{\ell-1}^n$ be the nth states of the Markov chains generated by Algorithm 2. Let the assumptions of Theorem 4.2, as well as Assumptions A3 and A4 hold, and suppose that $\mathcal{F}$ and $\mathcal{G}$ both satisfy Assumption A2. Denote $Y_\ell^n = Q_\ell(\theta_\ell^n) - Q_{\ell-1}(\Theta_{\ell-1}^n)$. Then*

$$\mathbb{V}_{\pi^\ell,\pi^{\ell-1}}\left[Y_\ell^n\right] \leq C_{k,f,p_0}\left(M_{\ell-1}^{-1/d+\delta} + R_{\ell-1}^{-1/2+\delta}\right), \quad \text{for any } \ \delta > 0.$$

To prove Lemma 4.5, we first need some preliminary results. Firstly, note that for $\Theta_{\ell-1}^{n+1} \neq \theta_{\ell,C}^{n+1}$ to be the case, the proposal on level ℓ at state $n+1$ had to be rejected. Given the proposal θ_ℓ' and the previous state θ_ℓ^n, the probability of this rejection is given by $1 - \alpha^\ell(\theta_\ell'|\theta_\ell^n)$. We need to quantify this probability, and this leads to the following crucial result.

Theorem 4.6. *Suppose $\mathcal{F}$ satisfies A2 and A3 and A4 hold. Then*

$$\lim_{\ell\to\infty} \alpha^\ell(\theta_\ell' \,|\, \theta_\ell'') = 1, \qquad \text{for } \mathcal{P}_\ell - \text{almost all } \theta_\ell', \theta_\ell''.$$

Furthermore,

$$\mathbb{E}_{\mathcal{P}_\ell,\mathcal{P}_\ell}\left[(1-\alpha^\ell)^q\right]^{1/q} \leq C_{k,f,p_0,q}\Big(M_{\ell-1}^{-1/d+\delta} + R_{\ell-1}^{-1/2+\delta}\Big),$$

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

Proof. We will first derive a bound on $1 - \alpha^\ell(\theta_\ell' \,|\, \theta_\ell'')$, for $\ell > 1$ and for θ_ℓ' and θ_ℓ'' given. First note that if $\frac{\pi^\ell(\theta_\ell')\,\pi^{\ell-1}(\theta_{\ell,C}'')}{\pi^\ell(\theta_\ell'')\,\pi^{\ell-1}(\theta_{\ell,C}')} \geq 1$, then $1 - \alpha^\ell(\theta_\ell' \,|\, \theta_\ell'') = 0$. Otherwise, we have

$$\begin{aligned} 1 - \alpha^\ell(\theta_\ell' \,|\, \theta_\ell'') &= \left(1 - \frac{\pi^\ell(\theta_\ell')}{\pi^{\ell-1}(\theta_{\ell,C}')}\right) + \left(\frac{\pi^\ell(\theta_\ell')\,\pi^{\ell-1}(\theta_{\ell,C}'')}{\pi^\ell(\theta_\ell'')\,\pi^{\ell-1}(\theta_{\ell,C}')}\right)\left(1 - \frac{\pi^\ell(\theta_\ell'')}{\pi^{\ell-1}(\theta_{\ell,C}'')}\right) \\ &\leq \left|1 - \frac{\pi^\ell(\theta_\ell')}{\pi^{\ell-1}(\theta_{\ell,C}')}\right| + \left|1 - \frac{\pi^\ell(\theta_\ell'')}{\pi^{\ell-1}(\theta_{\ell,C}'')}\right|. \end{aligned} \tag{4.6}$$

Let us consider either of these two terms and set $\theta_\ell = (\xi_j)_{j=1}^{R_\ell}$ to be either θ_ℓ' or θ_ℓ''. Using the definition of π^ℓ in (2.1), as well as the models (4.4) and (4.5) for the prior and the likelihood, respectively, we have

$$\begin{aligned} \frac{\pi^\ell(\theta_\ell)}{\pi^{\ell-1}(\theta_{\ell,C})} &= \frac{\mathcal{P}_\ell(\theta_\ell)}{\mathcal{P}_{\ell-1}(\theta_{\ell,C})}\,\frac{\mathcal{L}_\ell(F_{\text{obs}}|\theta_\ell)}{\mathcal{L}_{\ell-1}(F_{\text{obs}}|\theta_{\ell,C})} \\ &= \exp\left(-(2\pi)^{-\frac{R_\ell - R_{\ell-1}}{2}} \sum_{j=R_{\ell-1}+1}^{R_\ell} \frac{\xi_j^2}{2} \;-\; \frac{\|F_{\text{obs}} - F_\ell(\theta_\ell)\|^2}{\sigma_{F,\ell}^2} \;+\; \frac{\|F_{\text{obs}} - F_{\ell-1}(\theta_{\ell,C})\|^2}{\sigma_{F,\ell-1}^2}\right). \end{aligned} \tag{4.7}$$

Denoting $F_\ell := F(\theta_\ell)$ and $F_{\ell-1} := F(\theta_{\ell,C})$, and using the triangle inequality, we have that

$$\begin{aligned} \frac{\|F_{\text{obs}} - F_\ell\|^2}{\sigma_{F,\ell}^2} - \frac{\|F_{\text{obs}} - F_{\ell-1}\|^2}{\sigma_{F,\ell-1}^2} &\leq \frac{\Big(\|F_{\text{obs}} - F_{\ell-1}\| + \|F_\ell - F_{\ell-1}\|\Big)^2}{\sigma_{F,\ell}^2} - \frac{\|F_{\text{obs}} - F_{\ell-1}\|^2}{\sigma_{F,\ell-1}^2} \\ &= \|F_{\text{obs}} - F_{\ell-1}\|^2\left(\sigma_{F,\ell}^{-2} - \sigma_{F,\ell-1}^{-2}\right) + \frac{2\|F_{\text{obs}} - F_{\ell-1}\| + \|F_\ell - F_{\ell-1}\|}{\sigma_{F,\ell}^2}\|F_\ell - F_{\ell-1}\|. \end{aligned}$$

Since $\mathcal{F}$ was assumed to satisfy A2, it follows from the theory in [5, 30] (for the particular covariance function $C(x,y)$ in (4.3) with $r = 1$) that

$$\|F_\ell - F_{\ell-1}\| \;\lesssim\; C_{k_\ell,f,p_0}(\theta_\ell)\left(\|k_\ell - k_{\ell-1}\|_{C^0(\overline{D})} \;+\; M_\ell^{-1/d+\delta}\right),$$

for almost all θ_ℓ and for a constant $C_{k_\ell,f,p_0}(\theta_\ell) < \infty$ that depends on θ_ℓ only through $k_\ell := \exp\left(\sum_{j=1}^{R_\ell} \sqrt{\mu_j}\phi_j\xi_j\right)$. Since $\|F_{\ell-1}\|$ can be bounded independently of ℓ, for almost all θ_ℓ (again courtesy of Assumption A2), and since $\|F_{\text{obs}} - F_{\ell-1}\| \leq \|F_{\text{obs}}\| + \|F_{\ell-1}\|$, we can deduce that

$$\frac{\|F_{\text{obs}} - F_\ell\|^2}{\sigma_{F,\ell}^2} - \frac{\|F_{\text{obs}} - F_{\ell-1}\|^2}{\sigma_{F,\ell-1}^2} \lesssim C_{k_\ell,f,p_0}(\theta_\ell)\left((\sigma_{F,\ell}^{-2} - \sigma_{F,\ell-1}^{-2}) + \|k_\ell - k_{\ell-1}\|_{C^0(\overline{D})} + M_\ell^{-1/d+\delta}\right).$$

Finally, substituting this into (4.7) and using the inequality $|1 - \exp(x)| \leq |x| \exp|x|$ we have

$$\left|1 - \frac{\pi^\ell(\theta_\ell)}{\pi^{\ell-1}(\theta_{\ell,C})}\right| \lesssim C_{k_\ell,f,p_0}(\theta_\ell)\left((2\pi)^{-\frac{R_\ell - R_{\ell-1}}{2}}\zeta_\ell + (\sigma_{F,\ell}^{-2} - \sigma_{F,\ell-1}^{-2}) + \|k_\ell - k_{\ell-1}\|_{C^0(\overline{D})} + M_\ell^{-1/d+\delta}\right), \tag{4.8}$$

for almost all θ_ℓ, where $\zeta_\ell := \sum_{j=R_{\ell-1}+1}^{R_\ell} \xi_j^2$, i.e. a realisation of a χ^2-distributed random variable with $R_\ell - R_{\ell-1}$ degrees of freedom.

Now as $\ell \to \infty$, due to Assumption A3 we have $R_\ell \to \infty$ and $(2\pi)^{-(R_\ell - R_{\ell-1})/2}\zeta_\ell \to 0$, almost surely. Moreover, $M_\ell \to \infty$ and it follows from [5, Prop. 3.6 & §7.1] that $\|k_\ell - k_{\ell-1}\|_{C^0(\overline{D})} \to 0$, almost surely. Hence, using also A4 we have

$$\lim_{\ell\to\infty}\left|1 - \frac{\pi^\ell(\theta_\ell)}{\pi^{\ell-1}(\theta_{\ell,C})}\right| = 0, \quad \text{for almost all } \theta_\ell.$$

The first claim of the Theorem then follows immediately from (4.6).

For the bound on the moments of $1 - \alpha_\ell$, we use that all finite moments of $C_{k_\ell,f,p_0}(\theta_\ell)$ can be bounded independently of ℓ (cf. [5, 30]). It also follows from [5, Prop. 3.11 & §7.1] that

$$\mathbb{E}_{\mathcal{P}_\ell}\left[\|k_\ell - k_{\ell-1}\|_{C^0(\overline{D})}^q\right]^{1/q} \lesssim R_\ell^{-1/2+\delta}, \quad \text{for any } \ \delta > 0, \ \ q < \infty.$$

Finally, since ζ_ℓ is χ^2-distributed with $R_\ell - R_{\ell-1}$ degrees of freedom, we have

$$\mathbb{E}_{\mathcal{P}_\ell}[\zeta_\ell^q] = 2^q\frac{\Gamma\left(\frac{1}{2}(R_\ell - R_{\ell-1}) + q\right)}{\Gamma\left(\frac{1}{2}(R_\ell - R_{\ell-1})\right)} \lesssim (R_\ell - R_{\ell-1})^q, \quad \text{for any } \ \delta > 0, \ \ q < \infty.$$

Thus, the bound on the qth moment of $1 - \alpha_\ell$ follows immediately from (4.8), Assumptions A3 and A4 and Hölder's inequality. □

We will further need the following result.

Lemma 4.7. *For any θ_ℓ, let $k_\ell(\theta_\ell) := \exp\left(\sum_{j=1}^{R_\ell} \sqrt{\mu_j}\phi_j\xi_j\right)$ and $\kappa(\theta_\ell) := \min_{x\in\overline{D}} k_\ell(\cdot, x)$. Then*

$$|p_\ell(\theta_\ell) - p_\ell(\theta'_\ell)|_{H^1(D)} \ \lesssim \ \frac{\|f\|_{H^{-1}(D)}}{\kappa(\theta_\ell)\kappa(\theta'_\ell)}\,\|k_\ell(\theta_\ell) - k_\ell(\theta'_\ell)\|_{\mathcal{C}^0(\overline{D})}, \quad \textit{for almost all } \ \theta_\ell, \theta'_\ell, \tag{4.9}$$

and

$$\mathbb{E}_{\mathcal{P}_\ell,\mathcal{P}_\ell}\left[|p_\ell(\theta_\ell) - p_\ell(\theta'_\ell)|_{H^1(D)}^q\right]^{1/q} \ \leq \ \text{constant}, \tag{4.10}$$

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

$$\int_D k_\ell(\theta_\ell)\nabla p_\ell(\theta_\ell)\cdot\nabla v\,\mathrm{d}x = \int_D fv\,\mathrm{d}x = \int_D k_\ell(\theta'_\ell)\,\nabla p_\ell(\theta'_\ell)\cdot\nabla v\,\mathrm{d}x, \qquad \text{for all } v \in H_0^1(D),$$

(deduced from (4.2)) we have

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

Due to the standard estimate $|p_\ell(\theta_\ell)|_{H^1(D)} \lesssim \|f\|_{H^{-1}(D)}/\kappa(\theta_\ell)$ this implies (4.9)

It follows from [5, Prop. 3.10] that $\mathbb{E}_{\mathcal{P}_\ell}\left[\kappa(\theta_\ell)^{-q}\right]$ and $\mathbb{E}_{\mathcal{P}_\ell,\mathcal{P}_\ell}\left[\|k_\ell(\theta_\ell)\|^q_{\mathcal{C}^0(\overline{D})}\right]$ can be bounded independently of ℓ. The result then follows from an application of the Minkowski inequality to $\mathbb{E}_{\mathcal{P}_\ell,\mathcal{P}_\ell}\left[\|k_\ell(\theta_\ell) - k_\ell(\theta'_\ell)\|^q_{\mathcal{C}^0(\overline{D})}\right]^{1/q}$, as well as Hölder's inequality. □

Using Theorem 4.6 and Lemma 4.7, we are now ready to prove Lemma 4.5.

Proof of Lemma 4.5. Let θ^n_ℓ and $\Theta^n_{\ell-1}$ be the nth states of the Markov chains generated by Algorithm 2 on level ℓ. It follows from Lemma 4.4 and the fact that $\mathbb{V}_\pi[X] \leq \mathbb{E}_\pi[X^2]$, for any random variable X and any measure π, that

$$\mathbb{V}_{\pi^\ell,\pi^{\ell-1}}\left[Q_\ell(\theta^n_\ell) - Q_{\ell-1}(\Theta^n_{\ell-1})\right] \lesssim \mathbb{E}_{\mathcal{P}_\ell,\mathcal{P}_{\ell-1}}\left[\left(Q_\ell(\theta^n_\ell) - Q_{\ell-1}(\Theta^n_{\ell-1})\right)^2\right]. \tag{4.11}$$

Now, to simplify the presentation let us set $\theta := \theta^n_\ell$, $\theta_C = \theta^n_{\ell,C}$ and $\theta_F = \theta^n_{\ell,F}$, and denote by $\theta' = \theta'_\ell$ the proposal generated at the nth step of Algorithm 2 with $\theta'_C = \Theta^n_{\ell-1}$ and with some θ'_F. Note that $\theta' \neq \theta$ only if this proposal has been rejected at the nth step. It follows from (4.11) by the triangle inequality that

$$\mathbb{V}_{\pi^\ell,\pi^{\ell-1}}\left[Q_\ell(\theta) - Q_{\ell-1}(\theta'_C)\right] \lesssim \mathbb{E}_{\mathcal{P}_\ell,\mathcal{P}_\ell}\left[\left(Q_\ell(\theta) - Q_\ell(\theta')\right)^2\right] + \mathbb{E}_{\mathcal{P}_\ell,\mathcal{P}_{\ell-1}}\left[\left(Q_\ell(\theta') - Q_{\ell-1}(\theta'_C)\right)^2\right]. \tag{4.12}$$

A bound on the second term follows immediately from Lemma 4.3, i.e.

$$\mathbb{E}_{\mathcal{P}_\ell,\mathcal{P}_{\ell-1}}\left[\left(Q_\ell(\theta') - Q_{\ell-1}(\theta'_C)\right)^2\right] \leq C_{k,f,p_0}\left(M_\ell^{-2/d+\delta} + R_\ell^{-1+\delta}\right). \tag{4.13}$$

The first term in (4.12) is nonzero only if $\theta \neq \theta'$. We will now use Theorem 4.6 and Lemma 4.7, as well as the characteristic function $\mathbb{I}_{\{\theta\neq\theta'\}} \in \{0,1\}$ to bound it. Firstly, Hölder's inequality gives

$$\mathbb{E}_{\mathcal{P}_\ell,\mathcal{P}_\ell}\left[\left(Q_\ell(\theta) - Q_\ell(\theta')\right)^2\right] = \mathbb{E}_{\mathcal{P}_\ell,\mathcal{P}_\ell}\left[\left(Q_\ell(\theta) - Q_\ell(\theta')\right)^2 \mathbb{I}_{\{\theta\neq\theta'\}}\right]$$
$$\leq \mathbb{E}_{\mathcal{P}_\ell,\mathcal{P}_\ell}\left[\left(Q_\ell(\theta) - Q_\ell(\theta')\right)^{2q_1}\right]^{1/q_1} \mathbb{E}_{\mathcal{P}_\ell,\mathcal{P}_\ell}\left[\mathbb{I}_{\{\theta\neq\theta'\}}\right]^{1/q_2}, \tag{4.14}$$

for any q_1, q_2 s.t. $q_1^{-1} + q_2^{-1} = 1$. Since the functional $\mathcal{G}$ was assumed to be linear and bounded on $H^1(D) \subset H^{1/2-\delta}$, for all $\delta > 0$ (Assumption A2), it follows from Lemma 4.7 that the term $\mathbb{E}_{\mathcal{P}_\ell,\mathcal{P}_\ell}\left[\left(Q_\ell(\theta) - Q_\ell(\theta')\right)^{2q_1}\right]$ in (4.14) can be bounded by a constant independent of ℓ, for any $q_1 < \infty$. Moreover, using the law of total expectation, we have

$$\mathbb{E}_{\mathcal{P}_\ell,\mathcal{P}_\ell}\left[I_{\{\theta\neq\theta'\}}\right] = \mathbb{E}_{\mathcal{P}_\ell,\mathcal{P}_\ell}\Big[\mathbb{P}\left[\theta \neq \theta' \,|\, \theta, \theta'\right]\Big].$$

Since $\theta \neq \theta'$ only if the proposal θ' has been rejected on level ℓ at the nth step, the probability that this happens can be bounded by $1 - \alpha^\ell(\theta'|\theta)$, and so it follows by Theorem 4.6 that

$$\mathbb{E}_{\mathcal{P}_\ell,\mathcal{P}_\ell}\left[I_{\{\theta\neq\theta'\}}\right] \leq \mathbb{E}_{\mathcal{P}_\ell,\mathcal{P}_\ell}\left[1 - \alpha^\ell(\theta'|\theta)\right] \lesssim M_\ell^{-1/d+\delta} + R_\ell^{-1/2+\delta} \tag{4.15}$$

Combining (4.12)-(4.15) 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 4.8. *Under the same assumptions as in Lemma 4.5, the Assumptions M1 and M2 in Theorem 3.5 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 $\mathcal{C}_\ell \lesssim h_\ell^{-d} \log(h_\ell^{-1})$, e.g. via a multigrid solver, we can satisfy Assumption M4 with $\gamma = 1 + \delta$, for any $\delta > 0$. Then it follows from Theorems 3.5 and 4.8, as well as equation (2.11), 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, respectively:

$$\mathcal{C}_\varepsilon(\widehat{Q}_N^{\mathrm{MC}}) \lesssim \varepsilon^{-(d+2)-\delta} \quad \text{and} \quad \mathcal{C}_\varepsilon(\widehat{Q}_{L,\{N_\ell\}}^{\mathrm{ML}}) \lesssim \varepsilon^{-(d+1)-\delta}, \quad \text{for any } \delta > 0. \tag{4.16}$$

We clearly see the advantages of the multilevel method, which gives a saving of one power of ε compared to the standard MCMC method. Note that for multilevel estimators based on i.i.d samples, the savings of the multilevel method over the standard method are two powers of ε for $d = 2, 3$. The larger savings stem from the fact that $\beta = 2\alpha$ in this case, compared to $\beta = \alpha$ in the MCMC analysis above. The numerical results in the next section for $d = 2$ show that in practice we do seem to observe $\beta \approx 1 \approx 2\alpha$, suggesting $\mathcal{C}_\varepsilon(\widehat{Q}_{L,\{N_\ell\}}^{\mathrm{ML}}) = \mathcal{O}(\varepsilon^{-d})$. However, we do not believe that this is a lack of sharpness in our theory, but rather a pre-asymptotic phase. The constant in front of the leading order term in the bound of $\mathbb{V}_{\pi^\ell,\pi^{\ell-1}}[Y_\ell^n]$, namely the term $\mathbb{E}_{\mathcal{P}_\ell,\mathcal{P}_\ell}\big[\,(Q_\ell(\theta_\ell^n) - Q_\ell(\theta_\ell'))^{2q_1}\,\big]^{1/q_1}$ in (4.14), depends on the difference between $Q_\ell(\theta_\ell^n)$ and $Q_\ell(\theta_\ell')$. In the case of the pCN algorithm for the proposal distributions $q^{\ell,C}$ and $q^{\ell,F}$ (as used in Section 5 below) this difference will be small, since θ and θ' 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 expected value of some quantity of interest for our model problem (4.2). We start by presenting in Algorithm 3 a simplified version of Algorithm 2 given in Section 3 using symmetric proposal distributions for $q^{\ell,C}$ and $q^{\ell,F}$, describing in some more detail the evolution of the multilevel Markov chain used to approximate $\mathbb{E}_{\pi^\ell,\pi^{\ell-1}}[Y_\ell]$.

5.1 Implementation Details

Given the general description of the multilevel sampling in Algorithm 3, it remains to describe several computational details of the method, such as the choice of the symmetric transition probabilities $q^{\ell,C}(\Theta'_{\ell-1}|\Theta^n_{\ell-1})$ and $q^{\ell,F}(\theta'_{\ell,F}|\theta^n_{\ell,F})$, the values R_ℓ defining the partition of the KL modes over the multilevel hierarchy, as well as various MCMC tuning parameters.

For all our symmetric proposal distributions $q^{\ell,C}$ and $q^{\ell,F}$, $\ell = 1, \ldots, L$, we use the so-called precondition Crank-Nicholson (pCN) random walk proposed by Cotter et al. in [11]. Given the current state θ^n, the j^{th} entry of the proposal is obtained by

$$\theta'_j = \sqrt{1 - \beta^2}\, \theta_j^n \,+\, \beta\, \xi_j, \tag{5.4}$$

where $\xi_j \sim \mathcal{N}(0, 1)$ and β is a tuning parameter used to control the size of the step in the proposal, that may be chosen level dependent, i.e. $\beta = \beta_\ell$. In the numerical experiments, we typically choose $\beta_\ell < \beta_0$ for $\ell = 1, \ldots, L$.

ALGORITHM 3. (Simplified Metropolis Hastings MCMC for Y_ℓ, $\ell > 0$)

Choose initial states $\Theta^0_{\ell-1}$ and θ^0_ℓ. For $n \geq 0$:

- On level $\ell - 1$:
  - Given $\Theta^n_{\ell-1}$, generate $\Theta'_{\ell-1}$ from a symmetric distribution $q^{\ell,C}(\Theta'_{\ell-1}|\Theta^n_{\ell-1})$.
  - Compute $$\alpha^{\ell,C}(\Theta'_{\ell-1}|\Theta^n_{\ell-1}) = \min\left\{1, \frac{\pi^{\ell-1}(\Theta'_{\ell-1})}{\pi^{\ell-1}(\Theta^n_{\ell-1})}\right\}. \tag{5.1}$$
  - Set $\Theta^{n+1}_{\ell-1} = \begin{cases} \Theta'_{\ell-1} & \text{with probability } \alpha^{\ell,C}(\Theta'_{\ell-1}|\Theta^n_{\ell-1}) \\ \Theta^n_{\ell-1} & \text{with probability } 1 - \alpha^{\ell,C}(\Theta'_{\ell-1}|\Theta^n_{\ell-1}). \end{cases}$
- On level ℓ:
  - Given θ^n_ℓ, let $\theta'_{\ell,C} = \Theta^{n+1}_{\ell-1}$ and draw $\theta'_{\ell,F}$ from a symmetric distribution $q^{\ell,F}(\theta'_{\ell,F}|\theta^n_{\ell,F})$.
  - Compute $$\alpha^\ell(\theta'_\ell|\theta^n_\ell) = \min\left\{1, \frac{\pi^\ell(\theta'_\ell)}{\pi^\ell(\theta^n_\ell)}\frac{\pi^{\ell-1}(\theta^n_{\ell,C})}{\pi^{\ell-1}(\theta'_{\ell,C})}\right\}. \tag{5.2}$$
  - Set $\theta^{n+1}_\ell = \begin{cases} \theta'_\ell \equiv [\Theta^{n+1}_{\ell-1}, \theta'_{\ell,F}] & \text{with probability } \alpha^\ell(\theta'_\ell|\theta^n_\ell) \\ \theta^n_\ell & \text{with probability } 1 - \alpha^\ell(\theta'_\ell|\theta^n_\ell). \end{cases}$
- Compute $$Y^{n+1}_\ell = Q_\ell\left(\theta^{n+1}_\ell\right) - Q_{\ell-1}\left(\Theta^{n+1}_{\ell-1}\right) \tag{5.3}$$

The other free parameters in Algorithm 3 are the parameters $\sigma^2_{F,\ell}$ found in the likelihood model described in (4.5). The value of $\sigma^2_{F,\ell}$ controls the fidelity with which we require the model response to match the observed data on level ℓ. In our implementation we fix the fine-level likelihood variance $\sigma^2_{F,L}$ to a value consistent with traditional single level MCMC simulations (i.e. the measurement error associated with F_{obs} in a practical application), and then allow the remaining parameters to increase on coarser levels. This is done for two reasons. First, because the coarse simulations do not include all stochastic modes of the model, and so the coarse approximation will not necessarily agree exactly with the observed data. Second, since the coarse approximations necessarily include a higher level of discretisation error, it makes sense to relax the restrictions on the agreement between the model response and the observed data. Due to the consistency of the multilevel estimator, the choices of $\sigma^2_{F,\ell}$, $\ell < L$, will only influence the overall cost of the estimator and not the bias. The particular values used in the presented numerical experiments are chosen so that the values of likelihood variance increase with the characteristic mesh size in the hierarchy. Specifically we fix the value of $\sigma^2_{F,L}$ on the finest grid, and then set

$$\sigma^2_{F,\ell} = (1 + \kappa h_l)\,\sigma^2_{F,\ell+1}, \quad \ell = 0, \ldots, L-1, \tag{5.5}$$

where h_l is the mesh size on level l and κ is a tuning parameter. This choice ensures assumption **A4**.

To reduce dependence of the simulation on the initial state of the Markov chain, and to aid in the exploration of the potentially multi-modal stochastic space, we simulate multiple parallel chains simultaneously. The variance of the multilevel estimator $\mathbb{V}_{\boldsymbol{\Theta}_\ell}[\widehat{Y}^{\text{MC}}_{\ell,N_\ell}]$ is approximated on each grid level by $s^2_{\ell,N}$ using the method of Gelman and Rubin [15]. Finally, due to the very high-dimensional parameter space in our numerical experiments, both the single-level and multilevel

samplers displayed poor mixing properties. As such, we use a thinning process to decrease the correlation between consecutive samples, whereby we include only every T^{th} sample in the approximation of the level-dependent estimator, where T is some integer thinning parameter [27]. Then, after discarding n_0 initial burn-in samples, the approximation of $\mathbb{E}_{\pi^\ell,\pi^{\ell-1}}[Y_\ell]$ is computed by

$$\widehat{Y}^{\text{MC}}_{\ell,N_\ell} := \frac{1}{N_\ell}\sum_{n=n_0}^{n_0+N_\ell} Y_\ell^{(nT)}.$$

After the initial burn-in phase, the multilevel MCMC simulation is run until the weighted sum of the estimators from the $L+1$ grid levels satisfies

$$\sum_{\ell=0}^{L}\frac{s^2_{\ell,N_\ell}}{N_\ell} \;\leq\; \frac{\varepsilon^2}{2} \tag{5.6}$$

for some user prescribed tolerance ε. The total cost of the multilevel estimator is minimised when the number of samples on each level is chosen to satisfy

$$N_\ell \propto \sqrt{\mathbb{V}_{\pi^\ell,\pi^{\ell-1}}\left[Y_\ell\right]/\mathcal{C}_\ell} \approx \sqrt{s^2_{\ell,N_\ell}/\mathcal{C}_\ell}, \tag{5.7}$$

as described in [9], where $\mathcal{C}_\ell$ is the cost of generating a single sample of Y_ℓ on level ℓ. We assume this cost to be expressed as

$$\mathcal{C}_\ell = \mathrm{C}^\star \eta_\ell^\gamma M_\ell^\gamma, \tag{5.8}$$

where the constant $\mathrm{C}^\star$ may depend on the parameters σ^2 and λ in (4.3), but does not depend on ℓ. The factors η_ℓ reflect the additional cost for the auxiliary coarse solve required on grid $\ell-1$. For the experiments presented below, with geometric coarsening by a factor of 4, we have $\eta_0 = 1$ and $\eta_\ell = 1.25$ for $j = 1,\ldots,L$. When an optimal linear solver (e.g. algebraic multigrid) is used to perform the forward solves in the simulation we can take $\gamma \approx 1$. For a given accuracy ε, the total cost of the multilevel estimator can be written as

$$\mathcal{C}_\varepsilon\left(\widehat{Q}^{\text{ML}}_{L,\{N_\ell\}}\right) := \sum_{\ell=0}^{L}\mathcal{C}_\ell N_\ell. \tag{5.9}$$

5.2 Numerical Experiments

We consider (4.2) defined on the domain $D = (0,1)^2$ with $f \equiv 1$. The boundary conditions are taken to be Dirichlet on the lateral boundaries of the domain, and Neumann on the top and bottom:

$$p|_{x_1=0} = 1, \quad p|_{x_1=1} = 0, \quad \left.\frac{\partial p}{\partial \mathbf{n}}\right|_{x_2=0} = 0, \quad \left.\frac{\partial p}{\partial \mathbf{n}}\right|_{x_2=1} = 0.$$

The quantity of interest we approximate in our numerical simulation is the flux through the "outflow" part of the boundary, given by

$$q_{\text{out}} := -\int_0^1 k\left.\frac{\partial p}{\partial x_1}\right|_{x_1=1} dx_2. \tag{5.10}$$

The (prior) conductivity field is modelled as a log-normal random field with covariance function (4.3) with $r = 1$. The "observed" data F_{obs} is obtained synthetically by generating a reference conductivity field from the prior, solving the forward problem, and evaluating the pressure at 9 randomly selected points in the domain. The domain is discretised using piecewise linear finite

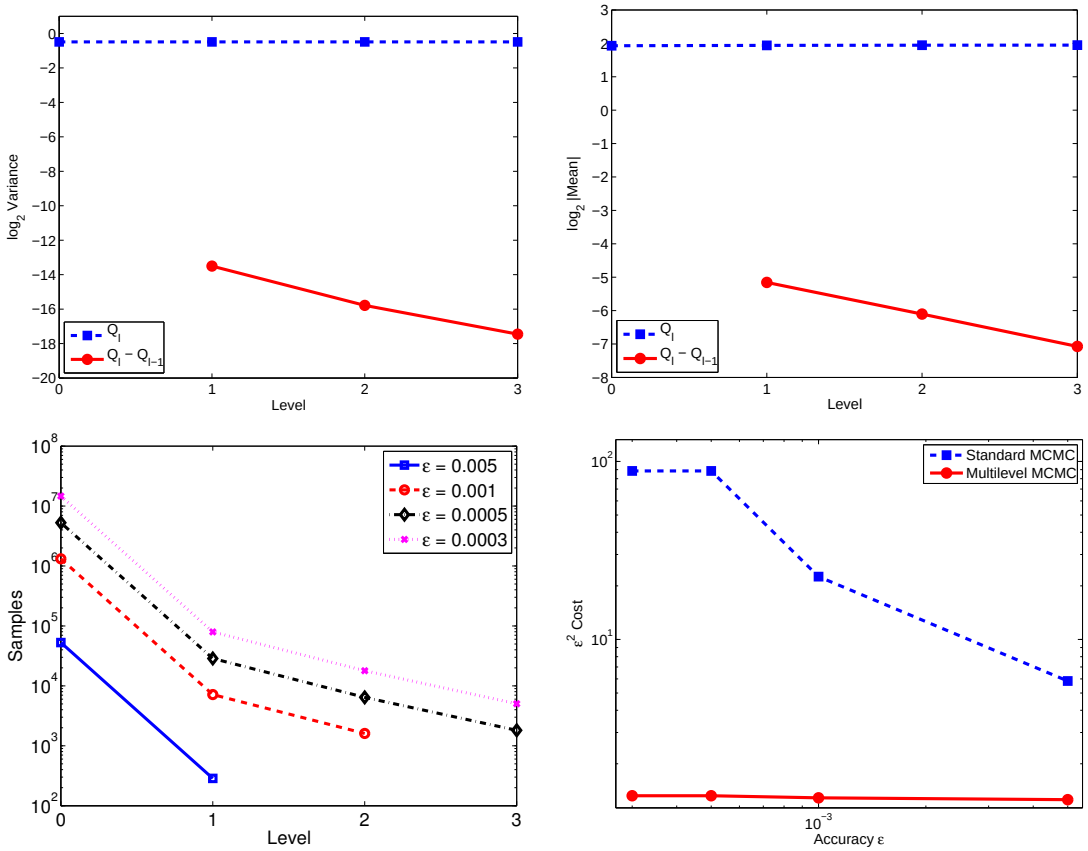

Figure 1: Performance plots for $\lambda = 0.5$, $\sigma^2 = 1$, $R_L = 169$, and $m_0 = 16$.

elements on a 2D uniform triangular mesh. The coarsest mesh contains $m_0 = 16$ grid points in each direction, with subsequently refined meshes containing $m_\ell = 2^\ell m_0$ in each direction, with the total number of grid points on level ℓ defined as $M_\ell = m_\ell^2$. Five parallel chains are used in each multilevel estimator.

The top two plots in Figure 1 show the results of a four-level simulation with $\lambda = 0.5$, $\sigma^2 = 1$, and $m_0 = 16$. The partition of the KL modes was such that $R_0 = 96$, $R_1 = 121$, $R_2 = 153$, and $R_3 = 169$. Five parallel chains were used for each level-dependent estimator. The fidelity parameter in the likelihood on the finest grid was taken to be $\sigma_{F,L}^2 = 10^{-4}$. The fidelity parameters on the other levels were obtained by equation (5.5) with $\kappa = 1$. The simulation was stopped when the variance of the multilevel estimator reached $\varepsilon^2/2$ with $\varepsilon = 8 \times 10^{-4}$. The top left plot compares the variance of quantities Q_ℓ and Y_ℓ on each level. The top right plot compares the mean of quantities Q_ℓ and Y_ℓ on each level. The plots for Y_ℓ seem to decay with $\mathcal{O}(h_\ell^2)$ and $\mathcal{O}(h_\ell)$, respectively. This suggests that at least in the pre-asymptotic phase our theoretical result on the variance which predicts $\mathcal{O}(h_\ell)$ (in Theorem 4.8) is not sharp (see comments at the end of Section 4). The result on the bias seems to be confirmed.

The bottom two plots in Figure 1 show the number of samples N_ℓ required on each level of the multilevel MCMC sampler and compare the computational cost of the standard and multilevel MCMC samplers for varying values of accuracy ε, respectively. Note that for larger values of ε fewer grid levels are required to attain a reduction in variance equivalent to the spatial discretisation error. The total cost of the simulation is given in terms of the cost of one forward solve on the coarsest grid (which is the same in each case). The y-axis is scaled by ε^2. It is clear that the multilevel sampler attains a dramatic reduction in computational cost over the standard MCMC

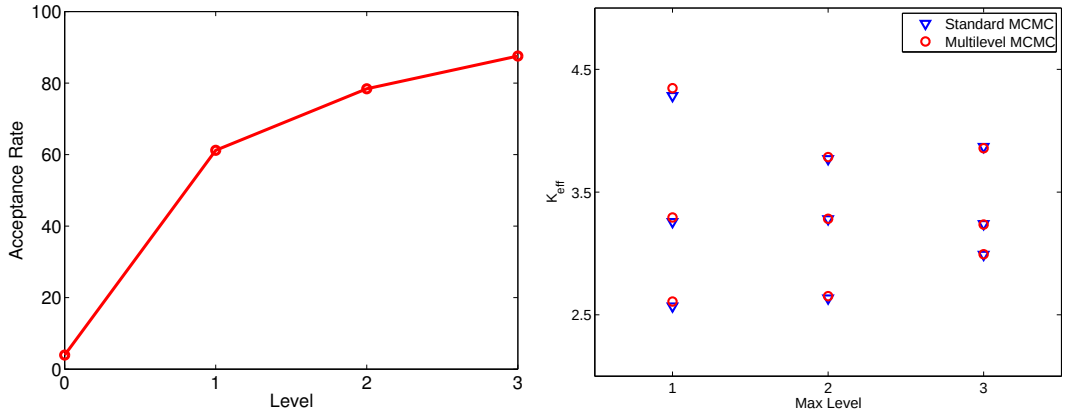

Figure 2: Average acceptance rate α^ℓ of the multilevel sampler (left figure) and estimates for q_{out} for nine reference data sets (right figure) for $\lambda = 0.5$, $\sigma^2 = 1$, $R_L = 169$, and $m_0 = 16$.

sampler. The precise speedup of the multilevel over the standard algorithm can be evaluated by taking the ratio of the total cost of the respective estimators, as defined by (5.7)-(5.9). When an optimal linear solver (such as AMG, with $\gamma \approx 1$) is used for the forward solves in the four-level simulation with $\varepsilon = 8 \times 10^{-4}$ (as in the top plots of Figure 1), the computational cost of the simulation is reduced by a factor of 50. When a suboptimal linear solver is used (say, $\gamma \approx 1.5$ for a sparse direct method) the computational cost is reduced by a factor of 275.

Figure (2) (left) confirms that the average acceptance rates α_ℓ of the fine-level samplers – the last three dots in Figure (2) (left) – tend to 1 as ℓ increases, and $\mathbb{E}[1-\alpha_\ell] \approx \mathcal{O}(h_\ell)$, as predicted in Theorem 4.6. Finally, the results in Figure (2) (right) demonstrate the good agreement between the MLMCMC estimate $\widehat{Q}^{\text{ML}}_{L,\{N_\ell\}}$ and the standard MCMC estimate $\widehat{Q}^{\text{MC}}_N$ of the quantity of interest q_{out} for nine distinct sets of reference data with three levels of fine-grid resolution. As before, the coarse grid in each case was defined with $m_0 = 16$, the tolerance for both estimators was $\varepsilon = 8 \times 10^{-4}$ and the model for the log-normal conductivity field is parametrised by $\lambda = 0.5$, $\sigma^2 = 1$ and $R_L = 169$ on the finest grid.

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

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

[14] 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.

[15] A. Gelman and D.B. Rubin. Inference from iterative simulation using multiple sequences. *Statistical Sciences*, 7(4):457–511, 1992.

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

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

[18] C.J. Gittelson, J. Könnö, Ch. Schwab, and R. Stenberg. The multilevel Monte Carlo finite element method for a stochastic Brinkman problem. SAM Report 2011–31, ETH Zurich, 2011.

[19] I.G. Graham, R. Scheichl, and E. Ullmann. Finite element error analysis for mixed formulations of elliptic PDEs with lognormal coefficients. In preparation, 2012.

[20] M. Hairer, A.M. Stuart, and S.J. Vollmer. Spectral gaps for a Metropolis–Hastings Algorithm in Infinite Dimensions. Technical Report arXiv:1112.1392, 2011. Available at `arxiv.org`.

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

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

[23] V.H. Hoang, Ch. Schwab, and A.M. Stuart. Sparse MCMC GPC finite element methods for Bayesian inverse problems. Technical Report arXiv:1207.2411, 2012. Available at `arxiv.org`.

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

[25] 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.

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

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

[28] D. Rudolf. *Explicit error bounds for Markoc chain Monte Carlo*. PhD thesis, Friedrich–Schiller–Universität Jena, 2011. Available at `http://tarxiv.org/abs/1108.3201`.

[29] A. L. Teckentrup. Multilevel Monte Carlo methods for highly heterogeneous media. Technical Report arXiv:1206:1479, 2012. To appear in the Proceedings of the Winter Simulation conference 2012, available at `http://informs-sim.org`.

[30] 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.*, pages 1–32. Published online March 12th, 2013.