MLMCMC – Multilevel Markov Chain Monte Carlo
Robert Scheichl
(joint work with Christian Ketelsen, Aretha Teckentrup, Panayot Vassilevski)

In this talk we address the problem of the prohibitively large computational cost of existing Markov chain Monte Carlo (MCMC) methods for large–scale applications with high dimensional parameter spaces, e.g. 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.

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*) 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 MCMC approach [9], 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 (the data fit) 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, since it commonly involves the numerical solution of a partial differential equation (PDE) with highly varying coefficients (for accuracy reasons usually) on a 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 infeasibly high overall complexity, particularly in the context of high-dimensional parameter spaces, typical in realistic subsurface flow problems, 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 in the context of standard Monte Carlo estimators based on independent and identically distributed (i.i.d.) samples [7, 5], in particular for subsurface flow problems [3, 1, 2, 10]. The basic ideas are to exploit the linearity of expectation, to introduce (in an unbiased way) 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 to build estimators for differences of output quantities instead of estimators for the quantities themselves. In that way each individual estimator will either (i) have a smaller variance, since the differences of the output quantities from two consecutive models go to zero with increased model resolution, or (ii) require significantly less computational work per sample, if the model resolution is low. Either way the cost of an individual estimator is significantly reduced, easily compensating for the extra cost of having to compute L 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 is dominated by the evaluation of the likelihood in the finest model leading to no real cost reduction on the coarser levels. 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 $\{\theta_\ell^n\}_{n\geq 0}$. A similar two-level proposal distribution has been investigated before in [4], but only for standard single-level Metropolis-Hastings.

Let us describe the new algorithm for the following model problem of stationary, single phase flow in a porous medium:

$$-\nabla \cdot (k(x,\omega)\nabla p(x,\omega)) = f(x), \qquad \text{in} \quad D \subset \mathbb{R}^d, \tag{1}$$

subject to the Dirichlet boundary condition $p(\omega,x) = p_0(x)$ on ∂D, with a lognormal distribution for the input random field, the permeability $k(x,\omega)$, with covariance function $C(x,y) = \sigma^2 \exp(-\|x-y\|_1/\lambda)$. We discretise (1) using standard, continuous, piecewise linear finite elements (FEs) on a sequence of grids $\{\mathcal{T}_\ell\}_{\ell\geq 1}$, with mesh width $h_\ell = h_0 2^{-\ell}$, and we sample from the input random field on level ℓ using a truncated Karhunen-Loève (KL) expansion of $\log k$,

$$k_\ell(\theta_\ell(\omega), x) = \exp\Big(\sum_{j=1}^{R_\ell} \sqrt{\mu_j}\phi_j(x)\xi_j(\omega)\Big), \tag{2}$$

with R_ℓ terms. The KL-eigenpairs (μ_j, ϕ_j) are known explicitly for the above (exponential) covariance function $C(x,y)$. The prior of our model on level ℓ is thus the R_ℓ-dimensional random vector $\theta_\ell = (\xi_j)_{j=1}^{R_\ell}$ with multivariate standard normal $N(0,I)$ distribution $\mathcal{P}_\ell$. Using Bayes' Theorem, the posterior distribution, conditioned on observations $F_{\rm obs}$ of some functional $\mathcal{F}(p)$ of the PDE solution, is

$$\pi_\ell(\theta_\ell) = \mathbb{P}(\theta_\ell | F_{\rm obs}) \propto \mathcal{L}_\ell(F_{\rm obs}|\theta_\ell)\, \mathcal{P}_\ell(\theta_\ell). \tag{3}$$

with an (in general) intractable normalising constant $\mathcal{P}_F(F_{\text{obs}})$. The data fit is modelled to be Gaussian, i.e. $\mathcal{L}_\ell(F_{\text{obs}}|\theta_\ell) \propto \exp\big(-\|F_{\text{obs}} - \mathcal{F}(p_\ell(\theta_\ell))\|^2/(2\sigma_F^2)\big)$, where p_ℓ is the PDE solution on $\mathcal{T}_\ell$ and σ_F^2 is the *fidelity*. The output quantity of interest is the expected value of another functional $Q = \mathcal{G}(p)$ of the PDE solution.

We start by choosing a tolerance $\varepsilon > 0$ and a grid $\mathcal{T}_L$, $L \in \mathbb{N}$, such that the bias $|\mathbb{E}_{\pi_L}[Q - Q_L]| \le \varepsilon/\sqrt{2}$, where $Q_\ell = \mathcal{G}(p_\ell)$. Now, as indicated, we use linearity of expectation to define the following unbiased multilevel estimator for $\mathbb{E}_{\pi_L}[Q_L]$:

$$\widehat{Q}_L^{\text{ML}} \;=\; \tfrac{1}{N_1}\sum_{n=1}^{N_1} Q_1(\theta_1^n) \;+\; \sum_{\ell=2}^{L} \tfrac{1}{N_\ell}\sum_{n=1}^{N_\ell} Q_\ell(\theta_\ell^n) - Q_\ell(\Theta_{\ell-1}^n). \tag{4}$$

The two Markov chains $\{\theta_\ell^n\}_{n\ge0}$ and $\{\Theta_\ell^n\}_{n\ge0}$, for $1 \le \ell \le L-1$, are independent, but drawn from the same posterior distribution π_ℓ. Clearly, the multilevel estimator coincides with the standard MCMC estimator on level L (in the limit as $N_1, \ldots, N_L \to \infty$), since all other terms cancel.

All the Markov chains $\{\theta_\ell^n\}_{n\ge0}$ and $\{\Theta_\ell^n\}_{n\ge0}$ in (4) are constructed via the following (standard) Metropolis-Hastings algorithm (for details see [9, 8]):

Algorithm 1. Choose θ^0 (from the prior $\mathcal{P}$ or from some "burnt-in" chain)

1. Given θ^n, generate a new proposal θ' from a proposal distribution $q(\theta'|\theta^n)$
2. Evaluate $\alpha(\theta'|\theta^n) = \min\left\{1, \dfrac{\pi(\theta')\,q(\theta^n|\theta')}{\pi(\theta^n)\,q(\theta'|\theta^n)}\right\}$.
3. Set $\theta^{n+1} = \begin{cases} \theta' & \text{with probability } \alpha(\theta'|\theta^n), \\ \theta^n & \text{with probability } 1 - \alpha(\theta'|\theta^n). \end{cases}$

In practice this is randomised by averaging over several such chains on each level. For the "coarse" chains $\{\theta_1^n\}$, $\{\Theta_1^n\}$, $\ldots$, $\{\Theta_{L-1}^n\}$ we use a standard proposal distribution $q = q_\ell^{\text{RW}}$ based on a preconditioned random walk [6].

The important new ingredient is a novel two-level proposal distribution $q = q_\ell^{\text{TL}}$ for the fine chains $\{\theta_2^n\}$, $\ldots$, $\{\theta_L^n\}$. We want the chains $\{\theta_\ell^n\}$ and $\{\Theta_{\ell-1}^n\}$ to be close, so that the variance of $Q_\ell(\theta_\ell^n) - Q_\ell(\Theta_{\ell-1}^n)$ is small and thus the variance of the estimator on level ℓ in (4) is small. To ensure this, we use $\theta'_\ell = [\Theta_{\ell-1}^{n+1}, \theta'_{\ell,F}]$, where $\theta'_{\ell,F}$ contains the last $R_\ell - R_{\ell-1}$ ("fine") components of θ'_ℓ that are not active on level $\ell - 1$ and is obtained again by a random walk from $\theta_{\ell,F}^n$. It turns out that q_ℓ^{TL} is computable. It depends on the acceptance probability for $\Theta_{\ell-1}^{n+1}$, and so it follows via some algebra that the two-level acceptance probability is

$$\alpha_\ell^{\text{TL}}(\theta'_\ell|\theta_\ell^n) = \min\left\{1, \frac{\pi_\ell(\theta'_\ell)\;\;\pi^{\ell-1}(\theta_{\ell,C}^n)}{\pi_\ell(\theta_\ell^n)\pi^{\ell-1}(\Theta_{\ell-1}^{n+1})}\right\}, \tag{5}$$

where $\theta_{\ell,C}^n$ denotes the first $R_{\ell-1}$ ("coarse") components of θ_ℓ^n.

For linear (or Fréchet differentiable) functionals $\mathcal{F}$ and $\mathcal{G}$ we have the following theoretical results. The first lemma is a consequence of the decay, as $j \to \infty$, of the KL-eigenvalues μ_j in (2).

Lemma 1 ([8, Theorem 4.6])**.** *Let $R_\ell \gtrsim h_\ell^{-2}$. Then all finite moments of $1 - \alpha_\ell^{\text{TL}}(\theta'_\ell|\theta_\ell)$ are $\mathcal{O}(h_{\ell-1}^{1-\delta})$, for any $\delta > 0$.*

This means that on the finer levels we accept almost all samples. Using this together with the theory for standard Multilevel MC based on i.i.d. samples in [2, 10], it is possible to establish the following main result.

Theorem 2 ([8, Thm. 4.1 & 4.8])**.** *For any* ε, $\delta > 0$ *and* $\{\theta_\ell^0\}_{\ell=1}^L$ *with* $\pi_\ell(\theta_\ell^0) > 0$, *we have* $\lim_{\min\{N_\ell\}\to\infty} \widehat{Q}_L^{\mathrm{ML}} = \mathbb{E}_{\pi^L}[Q_L]$ *and there exists* $L \in \mathbb{N}$, $(N_\ell)_{\ell=1}^L \in \mathbb{N}^L$ *s.t.*

$$\mathbb{E}_{\mathrm{ML}}\left[(\widehat{Q}_L^{\mathrm{ML}} - \mathbb{E}_{\pi^L}[Q])^2\right] \le \varepsilon^2 \quad \textit{and} \quad \textit{Cost}(\widehat{Q}_L^{\mathrm{ML}}) = \mathcal{O}(\varepsilon^{-(d+1)-\delta}), \tag{6}$$

where $\mathbb{E}_{\mathrm{ML}}[\cdot]$ *is expectation w.r.t. the joint distribution of all the chains in* (4).

Note that in comparison, the standard Metropolis-Hastings algorithm with q_L^{RW} instead of q_L^{TL} (but with the same L so that the bias is again less than $\varepsilon/\sqrt{2}$) has an ε-cost of $\mathcal{O}(\varepsilon^{-(d+2)-\delta})$, i.e. a whole power of ε more than the multilevel approach.

The numerical experiments for $d = 2$ in [8] confirm all these theoretical results. In fact, in practice it seems that (at least in the pre–asymptotic phase) the cost seems to grow only like $\mathcal{O}(\varepsilon^{-d})$ and the absolute cost is between 10 and 100 times lower than for the standard estimator, 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 practically relevant MCMC analyses for many large scale PDE applications.

Note also that there is nothing special about the model problem above and that the algorithm is applicable in any other MCMC application, provided the input parameters can be ordered according to their "importance" for the functionals $\mathcal{F}$ and $\mathcal{G}$. In [8, Theorem 3.5] we formulate the above theoretical results in abstract terms and show that under certain assumptions – that need to be verified for any new application – the multilevel estimator always leads to a reduction in the ε-cost over the standard Metropolis-Hastings algorithm.

References

[1] A Barth, C 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] J Charrier, R Scheichl and AL Teckentrup, Finite element error analysis of elliptic PDEs with random coefficients and its application to multilevel Monte Carlo methods, *SIAM J. Num. Anal.*, **51** (1):322-352, 2013.

[3] KA Cliffe, MB Giles, R Scheichl and AL Teckentrup, Multilevel Monte Carlo methods and applications to elliptic PDEs with random coefficients, *Comput. Vis. Sci.*, **14**:3-15, 2011.

[4] Y Efendiev, T Hou and W Lou, Preconditioning Markov chain Monte Carlo simulations using coarse–scale models, *SIAM J. Sci. Comput.*, **28**(2):776–803, 2006.

[5] MB Giles, Multilevel Monte Carlo path simulation, *Oper. Res.*, **256** (3):981–986, 2008.

[6] M Hairer, AM Stuart and SJ Vollmer, Spectral gaps for a Metropolis–Hastings algorithm in infinite dimensions, preprint at `http://arxiv.org/abs/1112.1392`, submitted, 2011.

[7] S Heinrich, Multilevel Monte Carlo methods, in *Lecture Notes in Comput. Sci.*, Vol. 2179, Springer, pp. 3624-3651, 2001.

[8] C Ketelsen, R Scheichl, AL Teckentrup, A hierarchical multilevel Markov chain Monte Carlo algorithm with applications to uncertainty quantification in subsurface flow, submitted, 2013

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

[10] AL Teckentrup, R Scheichl, MB Giles and E Ullmann, Further analysis of multilevel MC methods for elliptic PDEs with random coefficients, *Numer. Math.*, publ. online March 2013.