

---

# Multilevel Delayed Acceptance MCMC with an Adaptive Error Model in PyMC3

---

**Mikkel B. Lykkegaard**  
Centre for Water Systems and  
Institute for Data Science and Artificial Intelligence  
University of Exeter  
EX4 4QF, United Kingdom  
m.lykkegaard@exeter.ac.uk

**Grigorios Mingas**  
The Alan Turing Institute  
NW1 2DB, United Kingdom  
gmingas@turing.ac.uk

**Robert Scheichl**  
Institute for Applied Mathematics and  
Interdisciplinary Center for Scientific Computing  
Ruprecht-Karls-Universität Heidelberg  
69120 Heidelberg, Germany  
r.scheichl@uni-heidelberg.de

**Colin Fox**  
Department of Physics  
University of Otago  
Dunedin 9016, New Zealand  
colin.fox@otago.ac.nz

**Tim J. Dodwell**  
Institute for Data Science and Artificial Intelligence  
University of Exeter  
EX4 4QF, United Kingdom  
t.dodwell@exeter.ac.uk

## Abstract

Uncertainty Quantification through Markov Chain Monte Carlo (MCMC) can be prohibitively expensive for target probability densities with expensive likelihood functions, for instance when the evaluation it involves solving a Partial Differential Equation (PDE), as is the case in a wide range of engineering applications. Multilevel Delayed Acceptance (MLDA) with an Adaptive Error Model (AEM) is a novel approach, which alleviates this problem by exploiting a hierarchy of models, with increasing complexity and cost, and correcting the inexpensive models on-the-fly. The method has been integrated within the open-source probabilistic programming package PyMC3 and is available in the latest development version. In this paper, the algorithm is presented along with an illustrative example.

## 1 Introduction

Sampling from an unnormalised posterior distribution  $\pi(\cdot)$  using Markov Chain Monte Carlo (MCMC) methods is a central task in computational statistics. This can be a particularly challenging problem when the evaluation of  $\pi(\cdot)$  is computationally expensive and the parameter space  $\theta$  and data  $\mathbf{d}$  defining  $\pi(\cdot)$  are high-dimensional. The sequential (highly) correlated nature of a Markov chain and the slow converge rates of Monte Carlo sampling, means that many MCMC samples are often required to obtain a sufficient representation of a posterior distribution  $\pi(\cdot)$ . Examples of such problems frequently occur in Bayesian inverse problems, image reconstruction and probabilistic machine learning, where simulations of the measurements (required to calculate a likelihood) depend on the evaluation of complex mathematical models (e.g. a system of partial differential equations) or the evaluation of prohibitively large data sets.

In this paper a MCMC approach capable of accelerating existing sampling methods is proposed, where a hierarchy (or sequence)  $\pi_0(\cdot), \dots, \pi_{L-1}(\cdot)$  of computationally cheaper approximations to the ‘full’ posterior density  $\pi(\cdot) \equiv \pi_L(\cdot)$  are available. As with the original delayed acceptance algorithm, proposed by Christen and Fox [1], the idea is to generate MCMC proposals for the next step in the chain from runs of MCMC subchains targeting the computationally cheaper, approximate densities. The original DA method proposed the approach for just two levels. In this paper, the approach is extended to recursively apply delayed acceptance across a complete hierarchy of model approximations, a method termed *multilevel delayed acceptance* (MLDA). There are close connections to and similarities with multilevel variance reduction techniques, first proposed by Giles [2], widely studied for forward uncertainty propagation problems and importantly extended to Multilevel Markov Chain Monte Carlo approach by Hoang et al. [3] and Dodwell *et al.* [4], and further to a Multi-Index setting by Jasra *et al.* [5]. As in other multilevel approaches, the subchains in MLDA can be exploited for variance reduction, but this is beyond the scope of this paper.

The increase in use of Bayesian probabilistic tools has naturally coincided with the development of user-friendly computational packages, allowing users to focus on model development and testing, rather than algorithm development of sampling methods and post-processing diagnostics. Various high quality packages are available. Examples include: MUQ, STAN and Pyro.<sup>1</sup> A guiding principle of our work and of this contribution was to ensure that the MLDA implementation is easily accessible, well supported and gives flexibility to users to define complex models in a friendly language. To achieve this we embed our sampler into the widely used open-source probabilistic programming package PyMC3 [6]. The method and implementation have been accepted in the development version, and will be made available with the next full release (version 3.9.4).

## 2 Adaptive Multilevel Delayed Acceptance (MLDA)

### 2.1 Preliminaries: Metropolis-Hastings MCMC Algorithms

Here, a typical Bayesian inverse problem is considered. Given are (limited) observations  $d \in \mathbb{R}^M$  of a system and a mathematical model  $\mathcal{F}(\theta) : \mathbb{R}^R \mapsto \mathbb{R}^M$ , which maps from a set of model parameters  $\theta \in \mathbb{R}^R$  to the space of model predictions of the data. The connection between model and data is then, in the simplest case, described by the additive model

$$d = \mathcal{F}(\theta) + \epsilon \quad (1)$$

(but it can also be more general). Here,  $\epsilon$  is a random variable, which can depend on  $\theta$  and captures the uncertainty of the model’s reproduction of the data. It might include measurement uncertainty of the recorded data, uncertainty due to model mis-specification and/or uncertainties due to sing in practice a numerical approximation of the mathematical model. The distribution of the random variable  $\epsilon$  defines the *likelihood*, i.e. the probability distribution  $\mathcal{L}(d|\theta)$ . For simplicity it is assumed to be Gaussian, i.e.  $\epsilon \sim \mathcal{N}(\mu_\epsilon, \Sigma_\epsilon)$  and  $\mathcal{L}(d|\theta) \sim \mathcal{N}(d - \mathcal{F}(\theta) - \mu_\epsilon, \Sigma_\epsilon)$ , but it does not have to be.

Given *prior* information  $\pi(\theta)$  on the distribution of the model parameters  $\theta$ , the aim is to condition this distribution on the observations, i.e. to obtain samples from the *posterior* distribution  $\pi(\theta|d)$ . Through Bayes’ theorem, it follows that

$$\pi(\theta|d) = \frac{\mathcal{L}(d|\theta)\pi(\theta)}{\pi(d)} \propto \mathcal{L}(d|\theta)\pi(\theta). \quad (2)$$

Since the normalising constant  $\pi(d)$  (the *evidence*) is not typically known, the conditional distribution  $\pi(\theta|d)$  is generally intractable and exact sampling is not possible. There are various computational strategies for generating samples from  $\pi(\theta|d)$ . This paper focuses on the Metropolis-Hastings MCMC algorithm, described in Algorithm 1. It creates a Markov chain  $\{\theta^j\}_{j \in \mathbb{N}}$  of correlated parameter states  $\theta^j$  that (in the limit) target the exact posterior distribution  $\pi(\theta|d)$  (cf. e.g. [7]). The efficiency of the algorithm is determined by the choice of the proposal distribution  $q(\cdot|\cdot)$ .

Whilst MCMC methods are the gold-standard for sampling from complex posterior distributions, for many types of models and data they come with significant practical challenges. Firstly, each cycle of Alg. 1 requires the evaluation of the model  $\mathcal{F}(\theta')$  which may be computationally very expensive. Secondly, the samples generated in the chain are correlated, and therefore many cycles of Alg. 1 are

<sup>1</sup>MUQ: <http://muq.mit.edu>, STAN: <https://mc-stan.org>, Pyro: <https://pyro.ai>

**Algorithm 1 (Metropolis-Hastings MCMC):** Choose  $\theta^0$ . Then, for  $j = 0, \dots, J - 1$ :

1. Given  $\theta^j$ , generate a proposal  $\theta'$  from a given proposal distribution  $q(\theta'|\theta^j)$ ,
2. Accept proposal  $\theta'$  as the next sample with probability

$$\alpha(\theta'|\theta^j) = \min \left\{ 1, \frac{\mathcal{L}(d|\theta') \pi(\theta') q(\theta^j|\theta')}{\mathcal{L}(d|\theta^j) \pi(\theta^j) q(\theta'|\theta^j)} \right\},$$

i.e. set  $\theta^{j+1} = \theta'$  with probability  $\alpha$ , and  $\theta^{j+1} = \theta^j$  with probability  $1 - \alpha$ .

often required to produce a sufficient number of "independent" (or *effective*) samples from  $\pi(\theta|\mathbf{d})$ . The ideal proposal distribution generates cheap candidate proposals  $\theta'$ , which have a high probability of being accepted, and are independent of the previous sample  $\theta^j$ .

In this paper, efficient, Metropolis-style proposal strategies are developed that exploit a hierarchy of approximations  $\mathcal{F}_\ell(\theta)$ , for  $\ell = 0, \dots, L - 1$ , to the full model  $\mathcal{F}_L := \mathcal{F}$ , which are assumed to be ordered according to increasing accuracy and computational cost.

## 2.2 Multilevel Delayed Acceptance

Delayed Acceptance (DA) is an approach first introduced by Christen and Fox [1], exploiting a simple, but highly effective idea. The original DA approach is a two-level method that assumes a computationally cheaper approximation  $\mathcal{F}^*$  for the forward map  $\mathcal{F}$  is available. The idea is that for any chosen proposal  $\theta'$ , a standard Metropolis accept/reject step (as given in Alg. 1) is performed with the approximate forward map  $\mathcal{F}^*(\theta')$  before the expensive forward model  $\mathcal{F}(\theta')$  is evaluated. Only if accepted, a second accept/reject step with the original forward map  $\mathcal{F}(\theta')$  and with acceptance probability  $\alpha = \min \left\{ 1, \frac{\mathcal{L}(d|\theta') \mathcal{L}^*(d|\theta^j)}{\mathcal{L}(d|\theta^j) \mathcal{L}^*(d|\theta')} \right\}$  is carried out. Here,  $\mathcal{L}^*(d|\cdot)$  denotes the posterior distribution with the likelihood defined by  $\mathcal{F}^*$ . The validity of this approach as a proposal method, yielding a convergent MCMC algorithm, is provided in [1].

The basic DA approach can be extended in two ways. First, instead of doing a single check for the proposal that comes from the fine level, a subchain of length  $J$  can be ran on the coarse level [8, 9]. This does not affect the theory, but has the advantage of decorrelating samples passed back as proposals to the fine level. Second, and this is the main, novel algorithmic contribution, DA is extended to a general multilevel setting, exploiting links to the Multilevel Markov Chain Monte Carlo (MLMCMC) Method proposed by Dodwell *et al.* [4].

The subtle differences between the approaches are apparent when comparing the schematics of the two multilevel proposal processes shown in Fig. 1. Algorithmically, Multilevel Delayed Acceptance (MLDA) can be seen as a recursion of Delayed Acceptance over multiple levels  $\ell = \{0, 1, \dots, L\}$ . Crucially, if  $\theta_\ell^i$  is the current state at level  $\ell$ , and a proposal  $\theta'$  from the coarse subchain on level  $\ell - 1$  is rejected at level  $\ell$ , the coarse subchain to generate the subsequent proposal for level  $\ell$  is again initiated from  $\theta_\ell^i$ . For MLMCMC, even if the coarse proposal is rejected, the coarse chain continues independently of the fine chain and does not revert to the state  $\theta_\ell^i$  (see Fig. 1, right). As a result, coarse and fine chains will detach, and only align once a coarse proposal is accepted at the fine level.

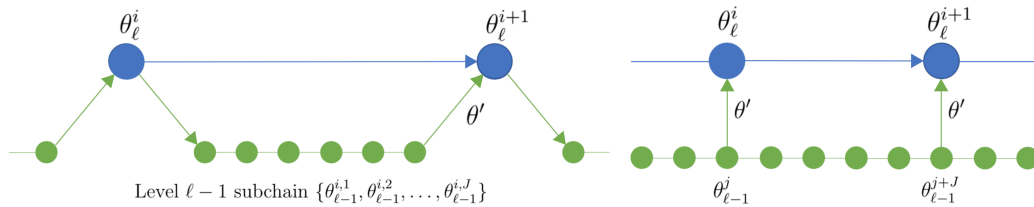


Figure 1: Schematic for generating a proposal  $\theta'$  on level  $\ell$  in MLDA (left) and in MLMCMC (right).

The new MLDA algorithm with subchain length  $J_\ell \in \mathbb{N}$  on level  $0 \leq \ell < L$  is described in Algorithm 2.

**Algorithm 2 (Multilevel Delayed Acceptance MCMC):**

Choose  $\theta^0$  and set the states of all subchains  $\theta_0^0 = \dots = \theta_{L-1}^0 = \theta^0$ . Then, for  $j = 0, \dots, J - 1$ :

1. Given  $\theta^j$  and  $\theta_\ell^{j_\ell}$  such that  $j_\ell < J_\ell$  for all  $1 \leq \ell < L$ , generate a subchain of length  $J_0$  with Alg. 1 on level 0, starting from  $\theta_0^0 = \theta_1^{j_1}$  and using the transition kernel  $q(\theta'_0|\theta_0^{j_1})$ .
2. Let  $\ell = 1$  and  $\theta'_1 = \theta_0^{j_1}$ .
3. If  $\ell = L$  go to Step 7. Otherwise compute the delayed acceptance probability on level  $\ell$ , i.e.,

$$\alpha_\ell = \min \left\{ 1, \frac{\mathcal{L}_\ell(d|\theta'_\ell) \mathcal{L}_{\ell-1}(d|\theta_\ell^{j_\ell})}{\mathcal{L}_\ell(d|\theta_\ell^{j_\ell}) \mathcal{L}_{\ell-1}(d|\theta'_\ell)} \right\}.$$

4. Set  $\theta_\ell^{j_\ell+1} = \theta'_\ell$  with probability  $\alpha_\ell$  and  $\theta_\ell^{j_\ell+1} = \theta_\ell^{j_\ell}$  otherwise. Increment  $j_\ell \rightarrow j_\ell + 1$ .
5. If  $j_\ell = J_\ell$  set  $\theta'_{\ell+1} = \theta_\ell^{j_\ell}$ , increment  $\ell \rightarrow \ell + 1$  and return to Step 3.
6. Otherwise set  $j_k = 0$  and  $\theta_k^0 = \theta_\ell^{j_\ell}$ , for all  $0 \leq k < \ell$ , and return to Step 1.
7. Compute the delayed acceptance probability on level  $L$ , i.e.,

$$\alpha_L = \min \left\{ 1, \frac{\mathcal{L}_L(d|\theta'_L) \mathcal{L}_{L-1}(d|\theta^j)}{\mathcal{L}_L(d|\theta^j) \mathcal{L}_{L-1}(d|\theta'_L)} \right\}.$$

Set  $\theta^{j+1} = \theta'_L$  with probability  $\alpha_L$  and  $\theta^{j+1} = \theta^j$  otherwise. Increment  $j \rightarrow j + 1$ .

8. Set  $j_\ell = 0$  and  $\theta_\ell^0 = \theta^j$ , for all  $0 \leq \ell < L$ , and return to Step 1.

**2.3 Adaptive correction of the approximate posteriors**

While the approach outlined above does guarantee sampling from the exact posterior, there are situations when convergence can be prohibitively slow. When the model approximation is poor, the delayed acceptance probability is low, and many proposals are rejected. This will result in suboptimal acceptance rates and low effective sample sizes. The leftmost panel in Fig. 2 shows a contrived example, where the approximate likelihoods (red/orange isolines) are offset from the likelihood on the finest level (blue contours) and their scales, shapes and orientations are incorrect. Thus, as an additional modification, an Adaptive Error Model (AEM) is introduced to account for discrepancies between model levels.

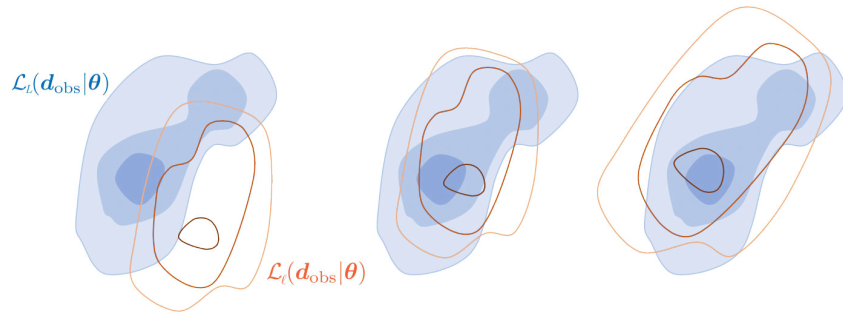


Figure 2: Effect of applying the Gaussian Adaptive Error Model (AEM). The first panel shows the initial state before adaptation, where the coarse likelihoods  $\mathcal{L}_\ell(d|\theta)$  (red/orange isolines) approximate the fine likelihood  $\mathcal{L}_L(d|\theta)$  (blue contours) poorly. The second panel shows the effect of shifting the likelihoods by the mean of the bias. The third panel shows the effect of additionally incorporating estimates of the covariance of the bias. (Adapted from [9].)

Let  $\mathcal{F}_\ell$  denote a coarse forward map of level  $\ell$  and  $\mathcal{F}_L$  denote the forward map on the finest level  $L$ . To obtain a better approximation of the data  $d$  using  $\mathcal{F}_\ell$ , the two-level AEM suggested in [10, 11] and analysed in [12] is extended by adding a telescopic sum of the differences in the forward model

output across all levels from  $\ell$  to  $L$ :

$$d = \mathcal{F}_L(\theta) + \epsilon = \mathcal{F}_\ell(\theta) + \mathcal{B}_\ell(\theta) + \epsilon \quad \text{with} \quad \mathcal{B}_\ell(\theta) := \sum_{k=\ell}^{L-1} \underbrace{\mathcal{F}_{k+1}(\theta) - \mathcal{F}_k(\theta)}_{:=B_k(\theta)}, \quad (3)$$

denoting the bias on level  $\ell$  at  $\theta$ . The trick in the context of MLDA is that, since  $\mathcal{B}_\ell$  is just a simple sum, the individual bias terms  $B_k$  from pairs of adjacent model levels can be estimated independently, so that new information can be exploited each time *any* set of adjacent levels are evaluated for the same parameter value  $\theta$ . Approximating each individual bias term  $B_k = \mathcal{F}_{k+1} - \mathcal{F}_k$  with a multivariate Gaussian  $B_k^* \sim \mathcal{N}(\mu_k, \Sigma_k)$ , the total bias  $\mathcal{B}_\ell$  can be approximated by the Gaussian  $\mathcal{B}_\ell^* \sim \mathcal{N}(\mu_{\mathcal{B},\ell}, \Sigma_{\mathcal{B},\ell})$  with  $\mu_{\mathcal{B},\ell} = \sum_k \mu_k$  and  $\Sigma_{\mathcal{B},\ell} = \sum_k \Sigma_k$ .

The bias-corrected likelihood function for level  $\ell$  is then proportional to

$$\mathcal{L}_\ell^*(d|\theta) \propto \exp\left(-\frac{1}{2}(d - \mathcal{F}_\ell(\theta) - \mu_\epsilon - \mu_{\mathcal{B},\ell})^T (\Sigma_\epsilon + \Sigma_{\mathcal{B},\ell})^{-1} (d - \mathcal{F}_\ell(\theta) - \mu_\epsilon - \mu_{\mathcal{B},\ell})\right). \quad (4)$$

One way to construct the AEM is offline, by sampling from the prior before running the MCMC, as suggested in [10]. However, this approach requires a significant overhead prior to sampling, and may result in a suboptimal error model, since the bias in the posterior may differ substantially from the bias in the prior. Instead, as suggested by [11], an estimate for the  $B_k$  can be constructed iteratively during sampling, using the following recursive formulae for sample mean and sample covariance [13]:

$$\mu_{k,i+1} = \frac{1}{i+1} \left( i\mu_{k,i} + B_k(\theta^{i+1}) \right) \quad \text{and} \quad (5)$$

$$\Sigma_{k,i+1} = \frac{i-1}{i} \Sigma_{k,i} + \frac{1}{i} \left( i\mu_{k,i} \mu_{k,i}^T - (i+1)\mu_{k,i+1} \mu_{k,i+1}^T + B_k(\theta^{i+1}) B_k(\theta^{i+1})^T \right) \quad (6)$$

While this approach in theory compromises ergodicity in the strict sense, the recursively constructed sample moments exhibit *diminishing adaptation* [13].

### 3 Implementation and Demonstration

The Multilevel Delayed Acceptance MCMC algorithm (Alg. 2) has been implemented in PyMC3 [6], an open-source probabilistic programming package for Python built on top of the Theano library [14]. The code is available in the development version of PyMC3.<sup>2</sup> In the following section, we present a numerical experiment, in which we compare the ‘‘vanilla’’ MLDA sampler to the AEM-activated MLDA sampler. To demonstrate the effect of the AEM, we have chosen models of very low resolution on the coarse levels. It is important to stress, however, that the AEM is not a strict requirement for MLDA in cases, where the coarse models are better approximations of the fine.

#### 3.1 Example: Estimation of Soil Permeability in Subsurface Flow

In this example, a simple model problem arising in subsurface flow modelling is considered. Probabilistic uncertainty quantification is of interest in various situations, for example in risk assessment of radioactive waste repositories. Moreover, this simple PDE model is often used as a benchmark for MCMC algorithms in the applied mathematics literature. The classical equations which govern steady-state single-phase subsurface flow are Darcy’s law coupled with an incompressibility constraint

$$w + k\nabla p = g \quad \text{and} \quad \nabla \cdot w = 0, \quad \text{in} \quad D \subset \mathbb{R}^d \quad (7)$$

for  $d = 1, 2$  or  $3$ , subject to suitable boundary conditions. Here  $p$  denotes the hydraulic head of the fluid,  $k$  the permeability tensor,  $w$  the flux and  $g$  is the source term.

A typical approach to treat the inherent uncertainty in this problem is to model the permeability as a random field  $k = k(x, \omega)$  on  $D \times \Omega$ , for some probability space  $(\Omega, \mathcal{A}, \mathbb{P})$ . Therefore, (7) can be written as the following PDE with random coefficients:

$$-\nabla \cdot k(x, \omega) \nabla p(x, \omega) = f(x), \quad \text{for all} \quad x \in D, \quad (8)$$

<sup>2</sup><https://github.com/pymc-devs/pymc3>

where  $f := -\nabla \cdot g$ . As a synthetic example, consider the domain  $D := [0, 1]^2$  with  $f \equiv 0$  and deterministic boundary conditions

$$p|_{x_1=0} = 0, \quad p|_{x_1=1} = 1 \quad \text{and} \quad \partial_n p|_{x_2=0} = \partial_n p|_{x_2=1} = 0. \quad (9)$$

A widely used model for the prior distribution of the permeability in hydrology is a log-Gaussian random field, characterised by the mean of  $\log k$ , here chosen to be 0, and by its covariance function, here chosen to be

$$C(x, y) := \sigma^2 \exp\left(-\frac{\|x - y\|_2^2}{2\lambda^2}\right), \quad \text{for } x, y \in D, \quad (10)$$

with  $\sigma = 2$  and  $\lambda = 0.3$ . The log-Gaussian random field is parametrised using a truncated Karhunen-Loève (KL) expansion of  $\log k$ , i.e., an expansion in terms of a finite set of independent, standard Gaussian random variables  $\theta_i \sim \mathcal{N}(0, 1)$ ,  $i = 1, \dots, R$ , given by

$$\log k(x, \omega) = \sum_{i=1}^R \sqrt{\mu_i} \phi_i(x) \theta_i(\omega). \quad (11)$$

Here,  $\{\mu_i\}_{i \in \mathbb{N}}$  are the sequence of strictly decreasing real, positive eigenvalues, and  $\{\phi_i\}_{i \in \mathbb{N}}$  the corresponding  $L^2$ -orthonormal eigenfunctions of the covariance operator with kernel  $C(x, y)$ . Thus, the prior distribution on the parameter  $\theta = (\theta_i)_{i=1}^R$  in the stochastic PDE problem (8) is  $\mathcal{N}(0, I_R)$ .

The aim is to infer the posterior distribution of  $\theta$ , conditioned on measurements of  $p$  at  $M = 25$  discrete locations  $x^j \in D$ ,  $j = 1, \dots, M$ , stored in the vector  $d_{obs} \in \mathbb{R}^M$ . Thus, the forward operator is  $\mathcal{F} : \mathbb{R}^R \rightarrow \mathbb{R}^M$  with  $\mathcal{F}_j(\theta_\omega) = p(x^j, \omega)$ .

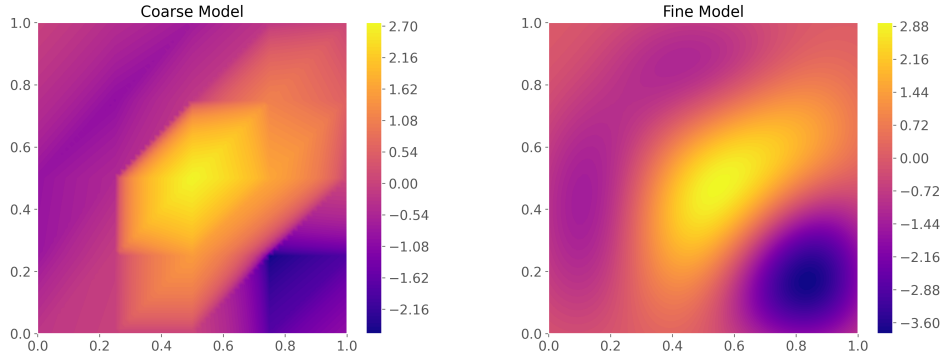


Figure 3: True log-conductivity field of the coarsest model with  $m_0$  grid points (left) and the finest model with  $m_2$  grid points (right).

All finite element (FE) calculations were carried out with FEniCS [15], using piecewise linear FEs on a uniform triangular mesh. The coarsest mesh  $\mathcal{T}_0$  consisted of  $m_0 = 5$  grid points in each direction, while subsequent levels were constructed by two steps of uniform refinement of  $\mathcal{T}_0$ , leading to  $m_\ell = 4^\ell(m_0 - 1) + 1$  grid points in each direction on the three grids  $\mathcal{T}_\ell$ ,  $\ell = 0, 1, 2$  (Fig. 3).

To demonstrate the excellent performance of MLDA with the AEM, synthetic data was generated by drawing a sample  $\theta^{ex}$  from the prior distribution and solving (8) with the resulting realisation of  $k$  on  $\mathcal{T}_2$ . To construct  $d_{obs}$ , the computed discrete hydraulic head values at  $(x^j)_{j=1}^M$ , were then perturbed by independent Gaussian random variables, i.e. by a sample  $\epsilon^* \sim \mathcal{N}(0, \Sigma_\epsilon)$  with  $\Sigma_\epsilon = 0.01^2 I_M$ .

To compare the “vanilla” MLDA approach to the AEM-enhanced version, we sampled the same model using identical sampling parameters, with and without AEM activated. For each approach, we sampled four independent chains, each initialised at a random point from the prior. For each independent chain, we drew 5000 samples plus a burn-in of 2000. We used subchain lengths  $J_0 = J_1 = 5$ , since that produced the best trade-off between computation time and effective sample size for MLDA with the AEM. Note that the cost of computing the subchains on the coarser levels only leads to about a 50% increase in the total cost for drawing a sample on level  $L$ . The PyMC3 non-blocked Random Walk Metropolis Hastings (RWMH) sampler was employed on the coarsest level with automatic step-size tuning during burn-in to achieve an acceptance rate between 0.2 and 0.5. All other sampling parameters were maintained at the default setting of the MLDA method.

To assess the performance of the two approaches the Effective Sample Size (ESS) for each parameter was computed [16]. Since the coarsest model was quite a poor approximation of the finest, running MLDA without the Adaptive Error Model (AEM), yielded very poor results. None of the four chains converged, there was poor mixing, a sub optimal acceptance rate of 0.019 on level  $L$ , and an ESS of 4 out of 20000 samples, meaning that each independent chain was only capable of producing a single independent sample. When the AEM was employed and otherwise using the exact same sampling parameters, we observed convergence for every chain, good mixing, an acceptance rate of 0.66 on level  $L$  and an ESS of 3319 out of 20000 samples (Fig. 4). In comparison, a single-level non-blocked RWMH sampler on grid  $\mathcal{T}_2$  with automatic step-size tuning during burn-in produced an ESS of 19 out of 5000 samples with an acceptance rate of 0.26.

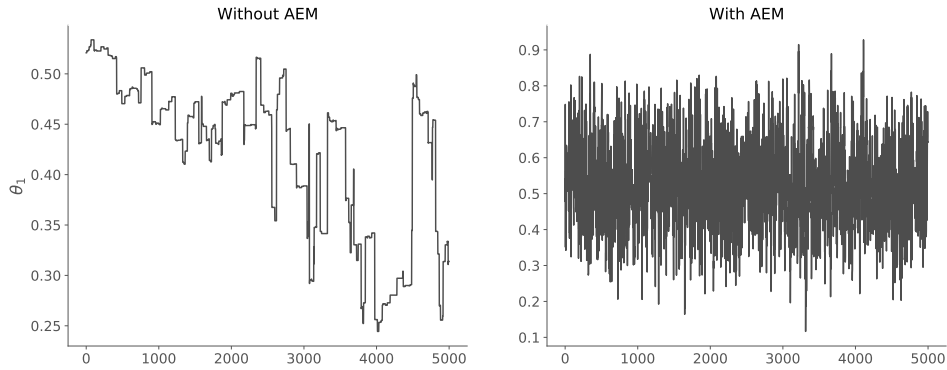


Figure 4: Traces of  $\theta_1$  on level  $\ell = 2$ , for MLDA without (left) and with AEM (right).

Note that the particular numerical experiment was chosen to demonstrate the dramatic effect that employing the AEM can have in MLDA. Thus, making it possible to use multilevel sampling strategies with very crude approximate models. A FE mesh with 25 degrees of freedom is extremely coarse for a Gaussian random field with correlation length  $\lambda = 0.3$ , yet using the AEM it still provides an excellent surrogate for delayed acceptance. Typically much finer models are used in real applications with longer subchains on the coarser levels (cf. [4]). The AEM will be less critical in that case and MLDA will also produce good ESS without the AEM. In a future journal paper, this topic will be carefully studied along with a comparison with other samplers on the finest level and an analysis of the multilevel variance reduction capabilities of MLDA.

## Broader Impact

This research has the potential to make unbiased uncertainty quantification of expensive models available to a greater audience, including engineers employed in risk assessment and reliability engineering. Since many engineering problems involve solving PDEs, multi-level hierarchies can easily be introduced using grid refinement, making this method exceptionally well suited for engineering applications.

## Acknowledgements

The work was funded by a Turing AI fellowship (2TAFPP100007) and the Water Informatics Science and Engineering Centre for Doctoral Training (WISE CDT) under a grant from the Engineering and Physical Sciences Research Council (EPSRC), grant number EP/L016214/1.

## References

- [1] J. A. Christen and C. Fox. Markov chain Monte Carlo using an approximation. *J. Comput. Graph. Stat.*, 14(4):795–810, 2005.
- [2] M. B. Giles. Multilevel Monte Carlo path simulation. *Oper. Res.*, 56(3):607–617, 2008.

- [3] V. H. Hoang, C. Schwab, and A. M. Stuart. Complexity analysis of accelerated MCMC methods for Bayesian inversion. *Inverse Probl.*, 29(8):085010, 2013.
- [4] T. J. Dodwell, C. Ketelsen, R. Scheichl, and A. L. Teckentrup. A hierarchical multilevel Markov chain Monte Carlo algorithm with applications to uncertainty quantification in subsurface flow. *SIAM/ASA J. Uncertain. Q.*, 3(1):1075–1108, 2015.
- [5] A. Jasra, K. Kamatani, K. Law, and Y. Zhou. A multi-index Markov chain Monte Carlo method. *Int. J. Uncertain. Quant.*, 8(1):61–73, 2018.
- [6] J. Salvatier, T. V. Wiecki, and C. Fonnesbeck. Probabilistic programming in Python using PyMC3. *PeerJ. Comput. Sci.*, 2:e55, 2016.
- [7] Gareth O. Roberts and Jeffrey S. Rosenthal. General state space Markov chains and MCMC algorithms. *Probability Surveys*, 1(0):20–71, 2004.
- [8] J. S. Liu. *Monte Carlo Strategies in Scientific Computing*. Springer, New York, 2004.
- [9] M. B. Lykkegaard, T. J. Dodwell, and D. Moxey. Accelerating uncertainty quantification of groundwater flow modelling using deep neural networks. *arXiv:2007.00400*, 2020. Manuscript submitted for publication.
- [10] J. Kaipio and E. Somersalo. Statistical inverse problems: Discretization, model reduction and inverse crimes. *J. Comput. Appl. Math.*, 198(2):493–504, 2007.
- [11] T. Cui, C. Fox, and M. J. O’Sullivan. Bayesian calibration of a large-scale geothermal reservoir model by a new adaptive delayed acceptance Metropolis Hastings algorithm. *Water. Resour. Res.*, 47:W10521, 2011.
- [12] T. Cui, C. Fox, and M. J. O’Sullivan. A posteriori stochastic correction of reduced models in delayed-acceptance MCMC, with application to multiphase subsurface inverse problems. *Int. J. Numer. Meth. Eng.*, 118(10):578–605, 2019.
- [13] H. Haario, E. Saksman, and J. Tamminen. An adaptive Metropolis algorithm. *Bernoulli*, 7(2):223, 2001.
- [14] Theano Development Team: Rami Al-Rfou et al. Theano: A Python framework for fast computation of mathematical expressions. *arXiv e-prints*, abs/1605.02688, 2016.
- [15] H. P. Langtangen and A. Logg. *Solving PDEs in Python – The FEniCS tutorial Volume I*. Simula SpringerBriefs on Computing. Springer International Publishing, 2017.
- [16] Aki Vehtari, Andrew Gelman, Daniel Simpson, Bob Carpenter, and Paul-Christian Bürkner. Rank-normalization, folding, and localization: An improved  $\widehat{R}$  for assessing convergence of MCMC. *arXiv:1903.08008 [stat]*, May 2020. arXiv: 1903.08008.