INTERNATIONAL JOURNAL FOR NUMERICAL METHODS IN ENGINEERING
Int. J. Numer. Meth. Engng 0000; **00**:1–31
Published online in Wiley Online Library (www.onlinelibrary.wiley.com). DOI: 10.1002/nme.6028

A posteriori stochastic correction of reduced models in delayed acceptance MCMC, with application to multiphase subsurface inverse problems

Tiangang Cui[1],Colin Fox*[2],Michael J. O'Sullivan[3]

[1]*School of Mathematical Sciences, Monash University, Melbourne,Australia*
[2]*Department of Physics, University of Otago, Dunedin, New Zealand*
[3]*Department of Engineering Science, The University of Auckland, Auckland, New Zealand*

SUMMARY

Sample-based Bayesian inference provides a route to uncertainty quantification in the geosciences, and inverse problems in general, though is very computationally demanding in the naïve form that requires simulating an accurate computer model at each iteration. We present a new approach that constructs a stochastic correction to the error induced by a reduced model, with the correction improving as the algorithm proceeds. This enables sampling from the correct target distribution at reduced computational cost per iteration, as in existing delayed-acceptance schemes, while avoiding appreciable loss of statistical efficiency that necessarily occurs when using a reduced model. Use of the stochastic correction significantly reduces the computational cost of estimating quantities of interest within desired uncertainty bounds. In contrast, existing schemes that use a reduced model directly as a surrogate do not actually improve computational efficiency in our target applications. We build on recent simplified conditions for adaptive Markov chain Monte Carlo algorithms to give practical approximation schemes and algorithms with guaranteed convergence. The efficacy of this new approach is demonstrated in two computational examples, including calibration of a large-scale numerical model of a real geothermal reservoir, that show good computational and statistical efficiencies on both synthetic and measured data sets.

Received . . .

KEY WORDS: adaptive MCMC; delayed acceptance; reduced model; geothermal reservoir modelling; inverse problem

1. INTRODUCTION

Characterizing subsurface flow in aquifers, geothermal and petroleum reservoirs often requires inferring unobserved subsurface properties from indirect observations made on the underlying system. This is a typical inverse problem. There are several fundamental difficulties associated with this process: data are corrupted by noise in the measurement process and often sparsely

*Correspondence to: C. Fox, Department of Physics, University of Otago, PO Box 56, Dunedin, New Zealand.
Email: colin.fox@otago.ac.nz

measured, the forward model only has a limited range of accuracy in representing the underlying system, and the parameters of interest are usually spatially distributed and highly heterogeneous, e.g., permeabilities and conductivities.

These features make the inverse problem ill-posed [1, 2]. This implies that there exists a range of feasible parameters that are consistent with the data, and hence a range of possible model predictions. Ill-posedness also causes the correlations between fitted parameters to be extremely high, and the model output to occupy some low dimensional manifold in data space. These properties make the traditional, deterministic solution to the inverse problem very sensitive to measurement error and model error.

In this paper, we develop the 'solution' of the inverse problem in the framework of Bayesian inference, and present algorithmic advances for computing inferential solutions. Solution of inverse problems under the Bayesian framework is well established, with comprehensive examples including characterization of subsurface properties of aquifers and petroleum reservoirs such as [3, 4, 5, 6], remote sensing such as [7, 8] and impedance imaging such as [9, 10]. By incorporating prior information and modelling uncertainties in the data observation process, all information regarding the solution of the inverse problem is coded into the posterior distribution over unknown parameters conditioned on the measured data.

Sample-based inference proceeds by computing Monte Carlo estimates of posterior statistics, to give 'solutions' and quantified uncertainties, using samples drawn from the posterior distribution via Markov chain Monte Carlo (MCMC) sampling. However, the applicability of this approach is fundamentally limited by the computational cost of the forward map and errors associated with the data modelling process. Difficulties arise from the high dimensionality of the model parameters, and computational demands of the forward map, which is usually dictated by numerical solutions of nonlinear partial differential equations (PDEs). In practice, it is computationally prohibitive to run standard MCMC for many iterations in realistic geophysical inverse problems.

Computational efficiency can be significantly improved by exploiting reduced models and recent advances in model errors. We present a new adaptive delayed acceptance (ADA) algorithm that is based on the delayed acceptance (DA) MCMC algorithm [11] that takes advantage of a reduced model and the associated approximate posterior to speed up the computation. Coupled with recent advances in stochastic models for model errors [12, 13] and state-of-the-art techniques in adaptive MCMC algorithms [14], ADA dynamically improves the accuracy of the approximate posterior distribution.

The adaptation used here is substantially different from the classical adaptive mesh refinement of finite difference or finite element methods. Rather, we start with a given reduced model for the forward map, in addition to an expensive, exact forward map that defines the desired target distribution and solutions. We implement a new usage of adaptive MCMC methods by learning a stochastic correction to the reduced model, thereby bringing the corrected approximate posterior closer to the exact posterior, as well as the usual adaptation of the proposal distribution.

The use of reduced, deterministic models is well developed in large-scale inverse problems, particularly in optimization for evaluating regularized inverses. Our interest is in using reduced models to improve efficiency of MCMC for fitting Bayesian hierarchical formulations of inverse problems. A reduced model induces an approximated likelihood function, and hence approximate posterior distribution. The approximate posterior distribution may be used to modify a proposal distribution by first performing an accept-reject step using the approximate posterior distribution. Subsequently using the modified proposal in a standard Metropolis–Hastings accept-reject step with the exact posterior distribution guarantees convergence to the desired, exact target distribution. When the reduced model has reduced computational cost, compared to the exact computational model, computational expense is potentially reduced since only those proposals that are accepted by the approximation are passed on for the exact, expensive model computation.

The simplest case of using a fixed reduced model, that induces a fixed 'surrogate' target distribution, was introduced by Liu [15] as the 'surrogate transition method', and later reintroduced as preconditioned or two-stage MCMC [16]. This method uses Approximation 1 in Sec. 3. Christen and Fox [11] introduced the more general 'delayed acceptance' algorithm that allows the

approximation to depend on the state of the MCMC, and demonstrated a speed-up in an inverse problem using a local linearization as the state-dependent reduced model. The delayed acceptance algorithm is presented in Section 2.4.

The use of an approximate model in delayed acceptance MCMC necessarily reduces statistical efficiency [11], compared to the MCMC that uses the exact target only. That is, more iterations are required to calculate quantities of interest to within a desired accuracy, asymptotic in the length of the chain. Without care, this increase in required number of iterations can offset the decrease in computation per iteration to the point where computational cost is not actually reduced, despite the extra programming effort. In the computed example in Section 5.2 we find that the use of a state-dependent reduced model is critical in improving computational efficiency in our target application; the same conclusion was reached in the setting of a large-data inverse problem in economics [17].

The main novel contribution in this work is extending delayed acceptance MCMC to allow correcting the deterministic reduced model by learning a stochastic correction to the *a posteriori* error induced by the reduced model, i.e., the difference between exact and approximated likelihood functions over states distributed as the posterior distribution. Stochastic corrections to the likelihood defined by a reduced model have been used previously, though the construction has been made *a priori*, that is, before exploration of the posterior distribution is performed. Here, we show that the construction may be performed *a posteriori*, as the posterior distribution is being explored, resulting in significant computational gains. In Section 3.1 we follow the development of the approximation error model (AEM) introduced by Kaipio & Somersalo for Bayesian-inspired regularization [2], though the same construction may be found in the field of computer model calibration [12]. We also draw on the techniques of adaptive MCMC [18, 14], that allows the transition kernel of the Markov chain to be modified as iterations progress, in a way that guarantees ergodicity for the desired posterior distribution. The AEM is presented in Section 3.1, with improvements developed subsequently in Section 3. Standard adaptive MCMC is presented in Section 2.5, with our novel adaptation of the stochastic correction presented in Section 4.

In Section 3 we present details of five approximation schemes: a fixed reduced model giving the surrogate transition method (Approx. 1); a fixed reduced model with AEM built over the prior distribution as in [2] (Approx. 2); a fixed reduced model with AEM built adaptively over the posterior distribution (Approx. 3); a zeroth-order correction of the fixed reduced model that gives a state-dependent approximation (Approx. 4); and the state-dependent approximation with stochastic correction built adaptively over the posterior distribution (Approx. 5). Approximations 3, 4, and 5 are new methods. Each of these five approximations have essentially the same computational cost per iteration, as the cost of the corrections are negligible in our target applications. Yet the statistical efficiency of the resulting MCMC is very different. Compared to the unmodified, exact MCMC, the surrogate transition method (Approx. 1) requires many times more iterations to estimate parameters to within a desired accuracy, while Approximation 5 requires virtually the same number of iterations. Statistical efficiency and computational efficiency are formally defined in Sections 2.3 and 2.4, respectively. This quantification of the trade-off between statistical efficiency and computing time of MCMC is applicable to all sampled-based solutions of inverse problems. A comparative study of computational cost and efficiency is presented in a simulated analysis of a well discharge test in Section 5.2 that demonstrates the progression in computational efficiency from Approximation 1 to Approximation 5.

The new adaptive delayed acceptance (ADA) algorithm, that builds the *a posteriori* stochastic correction and can implement all approximations, is presented in Section 4. A proof that ADA is ergodic for the target distribution is given in Section 4.1, to guarantee that Monte Carlo estimates evaluated over the chain converge to the true, posterior value. Proofs of ergodicity are crucial for correctness of an MCMC since convergence is in distribution so cannot be diagnosed from a single state, unlike optimization algorithms where a local check of zero gradient is possible. Ergodicity of adaptive MCMC algorithms, such as ADA, is even more delicate since it is easy to specify an adaptive MCMC that is incorrect, that is, Monte Carlo estimates fail to converge to the correct value. To emphasize this point [14] presents examples of seemingly straightforward adaptive MCMC algorithms that are actually incorrect. Our proof of ergodicity utilizes the simplified

requirements of simultaneous uniform ergodicity and diminishing adaptation introduced in [14]; see Theorem 1. The general framework allowed by the proof of ergodicity in Theorems 2 and 3 is not limited by the current form of reduced models and posterior approximations used in this paper. The general regularity conditions of our ergodicity theorem open the door to building other goal-oriented reduced models for high-dimensional inverse problems.

A further innovation presented in Section 2.5.2 is the grouped-component adaptive Metropolis (GCAM) proposal distribution that is suitable for high-dimensional inverse problems with black-box forward models.

We validate ADA on two models of geothermal reservoirs. The first is a 1D homogeneous model with 7 unknown parameters, using synthetic transient data. This moderate-sized inverse problem allows extensive calculation of statistics providing a comparative study of efficiencies of algorithms and approximations. In the second example we predict the hot plume of a 3D multi-phase geothermal reservoir model by estimating the heterogeneous and anisotropic permeability distribution and the heterogeneous boundary conditions [19]. This model has $10,049$ unknown parameters, and each model simulation requires about 30 to 50 minutes of CPU time. In both studies, ADA shows superior computational performance than the counterparts that do not use adaptation.

This paper is structured as follows: Section 2 briefly sets out the Bayesian formulation of inverse problems, reviews existing sample-based inference including delayed-acceptance and adaptive MCMC algorithms, discusses computational and statistical efficiency, and presents the GCAM proposal used in this paper. Section 3 presents deterministic and stochastic approximations to the forward map and posterior distribution that are used to reduce computational cost per iteration. Section 3.2 *et seq.* constitute the new contributions in this paper. Section 4 presents the ADA algorithm that utilizes the approximations developed in Section 3, and gives a proof of convergence. Section 5 presents two case studies of ADA on calibrating geothermal reservoir models. Section 6 offers some conclusions and discussion.

2. BAYESIAN FORMULATION AND POSTERIOR EXPLORATION

In this section, we review existing sample-based Bayesian methods for inverse problems that are relevant to the ADA algorithm developed in Section 4. We also discuss the computational efficiency of MCMC, that is the total computing cost required to evaluate a quantity of interest to within a given error tolerance, and present a new adaptive MCMC proposal used in our numerical examples.

2.1. Bayesian formulation

Suppose a physical system is simulated by a forward model $F(\cdot)$ and the unknowns of the physical system are parametrized by model parameters $\mathbf{x}$. The measurable features of the system are the image of the forward map at the true but unknown parameters $\mathbf{x}$, $\mathbf{d} = F(\mathbf{x})$, called the true, noise-free data. Practical measurements $\tilde{\mathbf{d}}$ are a noisy version of $\mathbf{d}$ subject to measurement errors and other sources of imprecision. In an inverse problem, we wish to recover the unknown $\mathbf{x}$ from measurements $\tilde{\mathbf{d}}$, and to make predictions about properties of the physical system, such as future, unobserved data.

Uncertainty in measured data $\tilde{\mathbf{d}}$ and in the model $F(\cdot)$ leads to uncertain estimates of parameters $\mathbf{x}$, and in subsequent predictions. What then is the range of permissible values, that is, the distribution over resulting estimates? The Bayesian formulation assigns probability distributions to each of the sources of error, and tracks the resulting distributions over quantities of interest. In our experience, this is the essential feature of the Bayesian method.

The Bayesian framework quantifies the distribution over parameters, consistent with the measured data, by the posterior distribution with the unnormalized probability density function

$$\pi_{\text{post}}(\mathbf{x}, \gamma, \delta \,|\, \tilde{d}) \propto L(\tilde{\mathbf{d}} \,|\, \mathbf{x}, \gamma)\, \pi_{\text{prior}}(\mathbf{x} \,|\, \delta)\, \pi_{\text{prior}}(\gamma)\, \pi_{\text{prior}}(\delta), \tag{1}$$

where $L(\tilde{\mathbf{d}} \,|\, \mathbf{x}, \gamma)$ is the likelihood function, giving the probability density of measuring data $\tilde{\mathbf{d}}$ in identical measurement setups defined by $\mathbf{x}$ and $\pi_{\text{prior}}(\mathbf{x} \,|\, \delta)$ is the prior probability density

function for model parameters $\mathbf{x}$. In Eqn. 1, we also introduce hyperparameters γ and δ to represent modelling assumptions in the likelihood function and the stochastic model of $\mathbf{x}$, respectively. In a typical hierarchical Bayesian setting, these hyperparameters follow independent prior distributions $\pi_{\text{prior}}(\gamma)$ and $\pi_{\text{prior}}(\delta)$.

The likelihood function is determined by modelling the stochastic process that results in measured data for a given state of the system. Measured data is commonly assumed to be related to noise-free data by the additive error model

$$\tilde{\mathbf{d}} = F(\mathbf{x}) + \mathbf{e}, \tag{2}$$

where the random vector $\mathbf{e}$ captures the measurement noise and other uncertainties such as model error. When $\mathbf{e}$ follows a zero mean multivariate Gaussian distribution the resulting likelihood function has the form

$$L(\tilde{\mathbf{d}} \,|\, \mathbf{x}, \Sigma_{\mathbf{e}}) \propto \exp\left[-\frac{1}{2}\big(F(\mathbf{x}) - \tilde{\mathbf{d}}\big)^{\top} \boldsymbol{\Sigma}_{\mathbf{e}}^{-1} \big(F(\mathbf{x}) - \tilde{\mathbf{d}}\big)\right], \tag{3}$$

where the hyperparameter $\gamma = \boldsymbol{\Sigma}_{\mathbf{e}}$ is the covariance matrix of the noise vector $\mathbf{e}$, with uncertainty in the covariance being modelled by the (hyper)prior distribution $\pi_{\text{prior}}(\gamma)$. Note that evaluating the likelihood function for a particular $\mathbf{x}$ requires simulating the forward model $F(\mathbf{x})$, which is a computationally taxing numerical simulation of a mathematical model of the physical system.

The contribution of model error to the noise vector $\mathbf{e}$ is usually non-negligible, in practice. This may be caused by discretization error in the computer implementation of the mathematical forward model and/or wrong assumptions in the mathematical model. We follow [5] who observe that it may not possible to separate the measurement noise and model error in the case that only single set of data is available. Thus, it is necessary to incorporate the modeller's judgments about the appropriate size and nature of the noise term $\mathbf{e}$. We also adopt the assumption of [5] that the noise $\mathbf{e}$ follows a zero mean Gaussian distribution, giving the likelihood function in Eqn. 3.

The primary unknowns $\mathbf{x}$ of a physical system can often be represented by spatially distributed parameters, for example, the coefficients in a partial differential equation modelling energy and/or mass propagation. Correspondingly, representations and prior models are usefully drawn from spatial statistics, often in the form of hierarchical models [20]; see [21] for a review. Since the particulars of prior modelling are problem specific, we delay presenting detailed functional forms to the examples in Section 5.

Because the interaction between the primary parameter $\mathbf{x}$ and hyperparameters $\Sigma_{\mathbf{e}}$ and δ introduces significant computational difficulties [22], here we set the value of hyperparameters based on expert opinion and field measurements. This yields the posterior density function

$$\pi_{\text{post}}(\mathbf{x} \,|\, \tilde{\mathbf{d}}) \propto \exp\left[-\frac{1}{2}\big(F(\mathbf{x}) - \tilde{\mathbf{d}}\big)^{\top} \boldsymbol{\Sigma}_{\mathbf{e}}^{-1} \big(F(\mathbf{x}) - \tilde{\mathbf{d}}\big)\right] \pi_{\text{prior}}(\mathbf{x}), \tag{4}$$

that is used throughout the remainder of this paper.

We summarize solutions to the inverse problem and the associated uncertainties as the expectation of some quantities of interest over the posterior. For instance, the mean of parameters, the variance of model states, and the credible intervals of model predictions. Given a quantity of interest $g(\mathbf{x})$, we calculate the Monte Carlo estimate, denoted by $\overline{g}_n$, of the posterior expectation as

$$\mathbb{E} = \int g(\mathbf{x}) \pi_{\text{post}}(\mathbf{x}|\tilde{\mathbf{d}}) \, \mathrm{d}\mathbf{x} \approx \overline{g}_n = \frac{1}{n} \sum_{i=1}^{n} g(\mathbf{x}_i), \tag{5}$$

using n samples drawn from the posterior distribution, i.e., $\mathbf{x}_i \sim \pi_{\text{post}}(\mathbf{x} \,|\, \tilde{\mathbf{d}})$. Thus, the task of estimating parameters or predictive values is reduced to the task of drawing samples from the posterior distribution; this defines sample-based Bayesian inference. We present algorithms for sampling from π_{post} in Sections 2.2 and 2.4, and novel efficient methods in Section 4.

Algorithm 1 Metropolis-Hastings

At iteration n, given $\mathbf{x}_n = \mathbf{x}$, then $\mathbf{x}_{n+1}$ is determined in the following way:

1. Propose new state $\mathbf{y}$ from some distribution $q(\mathbf{x}, \cdot)$.
2. With probability

$$\alpha(\mathbf{x}, \mathbf{y}) = \min\left\{1, \frac{\pi_{\text{post}}\left(\mathbf{y} \mid \tilde{\mathbf{d}}\right) q(\mathbf{y}, \mathbf{x})}{\pi_{\text{post}}\left(\mathbf{x} \mid \tilde{\mathbf{d}}\right) q(\mathbf{x}, \mathbf{y})}\right\}, \tag{6}$$

set $\mathbf{x}_{n+1} = \mathbf{y}$, otherwise $\mathbf{x}_{n+1} = \mathbf{x}$.

2.2. Metropolis-Hastings Dynamics

The basis of the sampling methods we develop is the Metropolis Hasting (MH) algorithm [23, 24]. One initializes the Markov chain at some starting state $\mathbf{x}_0$ and then iterate as in Alg. 1.

The only choice one has in Alg. 1 is the choice of proposal distribution $q(\mathbf{x}, \mathbf{y})$. Together with the acceptance probability $\alpha(\mathbf{x}, \mathbf{y})$, Alg. 1 defines a transition kernel

$$K(\mathbf{x}, \mathbf{y}) = q(\mathbf{x}, \mathbf{y})\ \alpha(\mathbf{x}, \mathbf{y}) + \left[1 - \int q(\mathbf{x}, \mathbf{z})\ \alpha(\mathbf{x}, \mathbf{z})\, d\mathbf{z}\right] \delta_{\mathbf{x}}(\mathbf{y}), \tag{7}$$

that produces a Markov chain of correlated samples. In Eqn. 7, $\delta_{\mathbf{x}}(\mathbf{y})$ is the Dirac delta function and the second term represents the probability of remaining at current state $\mathbf{x}$. The choice of the proposal is largely arbitrary as long as it satisfies reversibility, recurrence, and irreducibility [25], and thus the resulting MH algorithm is *ergodic* and has $\pi_{\text{post}}(\cdot \mid \tilde{\mathbf{d}})$ as the *unique stationary distribution*. An ergodic MH algorithm produces samples that converge in distribution to the posterior as the number of iterations $n \to \infty$. After a burn-in period, in which the Markov chain effectively loses dependency on the starting state, samples from the chain may be substituted directly into the Monte Carlo estimate in Eqn. 5 to produce the estimate $\overline{g}_n$ of quantity g.

The choice of the proposal has a significant influence on the rate of convergence of $\overline{g}_n$ to g, that depends on the degree of correlation [26, 27]. Markov chains that are fast to converge have lower correlation between adjacent samples. Traditionally, the proposal distribution is chosen from some simple family of distributions, e.g., multivariate Gaussian, then manually tuned in a "trial and error" manner to optimize the rate of convergence. We present automatic, adaptive methods for tuning the proposal in Sections 2.5 and 4.

2.3. Convergence and Statistical Efficiency

Sample-based inference returns $\overline{g}_n$ in Eqn. 5, which is an estimate of the quantity of interest with some Monte Carlo error due to n being finite. Under mild conditions, convergence of the Monte Carlo estimate $\overline{g}_n$, in Eqn. 5, is guaranteed by a central limit theorem (CLT) [28] that gives $\overline{g}_n - \mathbb{E}[g] \sim \mathcal{N}(0, \mathrm{Var}(\overline{g}_n))$ as $n \to \infty$. Here, $\mathcal{N}\left(\mu, \sigma^2\right)$ denotes the normal (or Gaussian) distribution with mean μ and standard deviation σ. Hence, asymptotic in sample size n, $\overline{g}_n \to \mathbb{E}[g]$ almost surely, with accuracy $\sqrt{\mathrm{Var}(\overline{g}_n)}$ when n samples are used. When the samples $\mathbf{x}_i$ are independent, it follows that

$$\mathrm{Var}(\overline{g}_n) = \frac{\mathrm{Var}(g)}{n}.$$

Since $\mathrm{Var}(g)$ depends only on the quantity being estimated, this sets the fewest number of iterations that a practical MCMC algorithm will require to achieve a given accuracy.

When the $\mathbf{x}_i$ are samples from a correlated Markov chain, as generated by any of the sampling algorithms discussed in this paper, instead (for large n)

$$\mathrm{Var}(\overline{g}_n) = \frac{\mathrm{Var}(g)}{n}\left(1 + 2\sum_{i=1}^{\infty} \rho_{gg}(i)\right),$$

where $\rho_{gg}(i)$ is the autocorrelation coefficient for the chain in g at lag i. Thus, the rate of variance reduction, compared to independent samples, is reduced by the factor

$$\tau = \left(1 + 2\sum_{i=1}^{\infty} \rho_{gg}(i)\right),$$

which is the integrated autocorrelation time (IACT) for the statistic g [26]. We can think of $\tau \geq 1$ being the length of the correlated chain that produces the same variance reduction as one independent sample. We call the quantity $1/\tau$ the *statistical efficiency*, while n/τ gives the number of effective (independent) samples for a Markov chain of length n.

2.4. Delayed Acceptance and Computational Efficiency

We define the *computational efficiency* of MH to be the effective sample size per unit CPU time. Hence, as shown in the previous section, this is proportional to the rate of variance reduction in estimates per CPU time.

For inverse problems, applying standard MH can be computationally costly as the cost is dominated by the evaluation of the posterior density, which involves simulating of the forward map $F(\mathbf{x}')$ at proposed parameters. Instead, we consider using an approximate posterior, obtained by using some reduced model of the forward model, to improve the computational efficiency. We embed the approximation in the DA algorithm shown in Alg. 2 to obtain asymptotically unbiased MCMC estimates.

Algorithm 2 Delayed Acceptance (DA)

At iteration n, given $\mathbf{x}_n = \mathbf{x}$ and a (possibly state-dependent) approximation to the target distribution $\pi^*_{\mathbf{x}}(\cdot \mid \tilde{\mathbf{d}})$, then $\mathbf{x}_{n+1}$ is determined in the following way:

1. Propose new state $\mathbf{y}$ from some distribution $q(\mathbf{x}, \cdot)$. With probability

$$\alpha(\mathbf{x}, \mathbf{y}) = \min\left\{1, \frac{\pi^*_{\mathbf{x}}\left(\mathbf{y} \mid \tilde{\mathbf{d}}\right) q(\mathbf{y}, \mathbf{x})}{\pi^*_{\mathbf{x}}\left(\mathbf{x} \mid \tilde{\mathbf{d}}\right) q(\mathbf{x}, \mathbf{y})},\right\}$$

   promote $\mathbf{y}$ to be used in Step 2, otherwise set $\mathbf{y} = \mathbf{x}$ (or, equivalently, set $\mathbf{x}_{n+1} = \mathbf{x}$ and exit).
2. The effective proposal distribution at the second step is

$$q^*(\mathbf{x}, \mathbf{y}) = q(\mathbf{x}, \mathbf{y})\ \alpha(\mathbf{x}, \mathbf{y}) + \left[1 - \int q(\mathbf{x}, \mathbf{z})\ \alpha(\mathbf{x}, \mathbf{z})\ \mathrm{d}\mathbf{z}\right] \delta_{\mathbf{x}}(\mathbf{y}).$$

   With probability

$$\beta(\mathbf{x}, \mathbf{y}) = \min\left\{1, \frac{\pi_{\text{post}}\left(\mathbf{y} \mid \tilde{\mathbf{d}}\right) q^*(\mathbf{y}, \mathbf{x})}{\pi_{\text{post}}\left(\mathbf{x} \mid \tilde{\mathbf{d}}\right) q^*(\mathbf{x}, \mathbf{y})}\right\}$$

   set $\mathbf{x}_{n+1} = \mathbf{y}$, otherwise $\mathbf{x}_{n+1} = \mathbf{x}$.

The DA algorithm allows the approximation to depend on the state of the MCMC. Hence DA generalizes the surrogate transition method of [15] that uses a fixed approximate target distribution, or 'surrogate'.

DA uses two accept-reject steps. The first step uses an approximation to the target distribution, while the second accept-reject step ensures that the Markov chain correctly targets the desired distribution (see [11] for details).

The resulting transition kernel is

$$K^*(\mathbf{x}, \mathbf{y}) = q(\mathbf{x}, \mathbf{y})\, \alpha(\mathbf{x}, \mathbf{y})\, \beta(\mathbf{x}, \mathbf{y}) + \left[1 - \int q(\mathbf{x}, \mathbf{z})\, \alpha(\mathbf{x}, \mathbf{z})\, \beta(\mathbf{x}, \mathbf{z}) d\mathbf{z}\right] \delta_{\mathbf{x}}(\mathbf{y}), \tag{8}$$

that composes the proposal density, the first step acceptance probability, and the second step acceptance probability. Given a proposal $q(\mathbf{x}, \cdot)$ for which a single level MH satisfies reversibility, recurrence, and irreducibility, DA is ergodic under mild conditions and has $\pi_{\text{post}}(\cdot \,|\, \tilde{\mathbf{d}})$ as the unique stationary distribution. [11]

In Alg. 2, there is never need to evaluate the integral in the definition of q^*. The computational cost per iteration is reduced because only those proposals that are accepted using the approximation $\pi^*_{\mathbf{x}}(\cdot \,|\, \tilde{\mathbf{d}})$ go on to evaluation of the posterior distribution $\pi_{\text{post}}(\cdot \,|\, \tilde{\mathbf{d}})$, that requires evaluating the full, expensive forward map. However, DA necessarily has lower statistical efficiency than the unmodified counterpart in Alg. 1, because of the additional acceptance/rejection in Step 2 [11]. That is, $\tau_{\text{DA}} \geq \tau$ for any quantity g. Fortunately, DA may still be more computationally efficient that the standard MH.

Here we analyze the potential speed-up factor in computational efficiency of DA compared to the standard MH. Let the CPU time to evaluate the approximate posterior density and the exact posterior density be t^* and t, respectively. Suppose that the average acceptance probability in Step 1 of DA is $\bar{\alpha}$, set by the choice of proposal. Since evaluation of the posterior density dominates the CPU time of MH, simulating n iterations of the standard MH asymptotically costs $n\,t$ CPU time. Using the same CPU time, DA can be simulated for $n\,t/(\bar{\alpha}t + t^*)$ iterations. This way, the effective sample size of DA is given by

$$\text{ESS}_{DA} = \frac{n\,t}{\tau_{DA}\,(\bar{\alpha}t + t^*)} = \frac{n}{\tau_{DA}\,(\bar{\alpha} + t^*/t)},$$

whereas the effective sample size of the standard MH is

$$\text{ESS} = \frac{n}{\tau}.$$

Thus, the speed-up factor of DA compared to the standard MH is

$$\frac{\text{ESS}_{\text{DA}}}{\text{ESS}} = \frac{\tau}{\tau_{\text{DA}}} \frac{1}{\bar{\alpha} + t^*/t}. \tag{9}$$

The factor $\tau/\tau_{\text{DA}} \leq 1$ counts the decrease in statistical efficiency by using DA instead of MH, while the factor $\bar{\alpha} + t^*/t$ gives the decrease in average compute cost per iteration. It is necessary to address both factors if one is to improve computational efficiency. That is, we want τ/τ_{DA} to be close to one and t^*/t to be as small as possible.

Several forms of reduced models have been used in DA to approximate the posterior. For example, local linearized models have been used in [11] and model coarsening based on multiscale finite element is used in [16]. It can be challenging to balance the reduction in CPU time against accuracy of the reduced model. Using a lower accuracy reduced model, which runs faster compared to a more accurate one, will reduce the CPU time of MCMC per iteration, but at the risk of lower statistical efficiency since DA is more likely to reject the proposal in Step 2.

Here we present a systematic way to modify the statistical model of the likelihood function without changing the reduced model. By including the statistics of the numerical error of the reduced model in the likelihood, our corrected approximate posterior can potentially improve the acceptance rate in step 2, at no extra cost. This way, the statistical efficiency, and hence the speed-up factor, of the DA can be improved by using the same reduced model.

2.5. Adaptive MCMC

A key ingredient of our modified likelihood is the use of adaptive MCMC to estimate posterior error statistics of the reduced model. Towards this goal, we design an adaptive delayed acceptance algorithm that both adaptively adjusts the proposal distribution and also adapts to the posterior error statistics in the likelihood using the MCMC sample history. The adaptation of the likelihood is significantly different from existing adaptive MCMC algorithms which only focus on adapting the proposal. This offers new insights in the adaptive construction of goal-oriented approximations to

the likelihood for posterior exploration. In this paper, we extend the regularity conditions established by [14, 29] to our case, to guarantee the ergodicity of our adaptive scheme.

We first review classical adaptive MCMC methods that focuses on tuning proposal distributions using past MCMC samples. This includes the classical adaptive Metropolis (AM) algorithm [18] and our modification that suits our case studies.

2.5.1. Adaptive Metropolis An important task in practical MCMC is identifying good proposal distributions that can minimize autocorrelation. As the forward model used in our case studies does not have adjoint capabilities, advanced MCMC proposals, such as the stochastic Newton [30] and the likelihood-informed dimension-independent proposal [31] that rely on derivatives of the posterior, cannot be used here. We consider the Gaussian random walk proposal

$$q(\mathbf{x}, \cdot) = \mathcal{N}(\mathbf{x}, \sigma^2 \boldsymbol{\Sigma}),$$

where $\boldsymbol{\Sigma}$ is some covariance matrix and σ is a scalar that dictates the jump size. Under certain technical assumptions, the covariance of the posterior can be a good choice for $\boldsymbol{\Sigma}$. There are two ways of choosing an optimal scaling σ, see Roberts and Rosenthal [32] and references therein for detailed analysis. One is to set $\sigma = 2.38/\sqrt{d}$, where d is the dimension of the parameter. Alternatively, one can choose an scaling σ such that the acceptance rate of MCMC is about 0.234. The values 2.38 and 0.234 are numerical approximations to analytical results. These two choices are equivalent under strict technical assumptions [32], however the latter often demonstrates better efficiency in practice.

Since the posterior covariance is unknown before carrying posterior sampling, Harrio et al. [18] introduced the adaptive Metropolis that estimates the posterior covariance on-the-fly using past posterior samples generated during the MCMC simulation. AM uses $\boldsymbol{\Sigma} \approx \boldsymbol{\Sigma}_n$ where $\boldsymbol{\Sigma}_n$ is the empirical posterior covariance evaluated over n iterations, with the scale is chosen as $\sigma = 2.38/\sqrt{d}$. The scaled empirical posterior covariance is mixed with a fixed Gaussian distribution, $\mathcal{N}(\mathbf{x}, (0.1^2/d)\mathbf{I}_d)$ to avoid the situation that $\boldsymbol{\Sigma}_n$ is singular. The algorithm is shown in Alg. 3.

Algorithm 3 Adaptive Metropolis

At iteration n, given $\mathbf{x}_n = \mathbf{x}$, then $\mathbf{x}_{n+1}$ is determined in the following way:

1. Given the empirical covariance estimate Σ_n of the target distribution up to step n and a small positive constant β, propose a new state $\mathbf{y}$ from the proposal

$$q_n(\mathbf{x}, \cdot) = \begin{cases} \mathcal{N}\big(\mathbf{x}, \frac{0.1^2}{d}\mathbf{I}_d\big) & n \leq 2d \\ \mathcal{N}\big(\mathbf{x}, (1-\beta)\frac{2.38^2}{d}\boldsymbol{\Sigma}_n + \beta\frac{0.1^2}{d}\mathbf{I}_d\big) & n > 2d \end{cases}, \tag{10}$$

   where d is the dimension of the parameter.
2. With probability $\min\big\{1, \pi_{\text{post}}(\mathbf{y} \mid \tilde{\mathbf{d}})/\pi_{\text{post}}(\mathbf{x} \mid \tilde{\mathbf{d}})\big\}$, $\mathbf{x}_{n+1} = \mathbf{y}$, otherwise $\mathbf{x}_{n+1} = \mathbf{x}$.

Remark 1

Note that the proposal is modified in every iteration of adaptive MCMC sampling. Thus, it becomes non-trivial to establish ergodicity of the resulting algorithm. For a non-adaptive proposal the transition kernel of MH satisfies the detailed balance condition

$$\pi_{\text{post}}(\mathbf{x}_n \mid \tilde{\mathbf{d}})K(\mathbf{x}_n, \mathbf{x}_{n+1}) = \pi_{\text{post}}(\mathbf{x}_{n+1} \mid \tilde{\mathbf{d}})K(\mathbf{x}_{n+1}, \mathbf{x}_n),$$

and hence reversibility. For an adaptive proposal, detailed balance is clearly violated as the proposal and the transition kernel depend on the iteration number n.

Results by [18, 33, 14, 34, 29] and others have established ergodicity for some adaptive MCMC proposals using differing techniques.

Roberts and Rosenthal [14] provided simplified regularity conditions required for ergodicity, namely simultaneous ergodicity and diminishing adaptation, that provides a viable route to establishing ergodicity of many adaptive MCMC algorithms, and is the route we take here.

2.5.2. Grouped components adaptive Metropolis As demonstrated by a set of numerical examples [35], the scaling $2.38/\sqrt{d}$ often shows suboptimal statistical performance, as the technical assumptions used to derive the scale are typically too restrictive for highly non-Gaussian posterior distributions with correlated parameters.

In addition, it may be computationally costly to estimate the covariance Σ for problems with high parameter dimensions.

To overcome these limitations, we present the grouped components adaptive Metropolis (GCAM) proposal that separately estimates the covariance Σ and the scale σ on different groups in a partition of the parameter coordinates.

Suppose we have L groups of components $\mathcal{I}_1, \ldots, \mathcal{I}_L$, $\mathcal{I}_j \subseteq \{1, 2, \ldots, d\}$ with each group $\mathcal{I}_j$ associated with a scale variable σ_j. Let d_j be the number of elements of group $\mathcal{I}_j$, and $-\mathcal{I}_j = \{1, 2, \ldots, d\} \setminus \mathcal{I}_j$.

The GCAM proposal is shown in Alg. 4.

Algorithm 4 Grouped Components Adaptive Metropolis

At iteration n, given $\mathbf{x}_n = \mathbf{x}$, then $\mathbf{x}_{n+1}$ is determined in the following way:

1. Initialize $\mathbf{y} = \mathbf{x}$, for all $j = 1 \ldots, L$:
   - Given the empirical covariance $\mathbf{\Sigma}_{n,\mathcal{I}_j}$ for the components $\mathcal{I}_j$ estimated from past samples and a small positive constant β, draw a d_j dimensional random variable $\mathbf{z}$ from the proposal

$$q_n(\mathbf{x}_{\mathcal{I}_j}, \cdot) = \begin{cases} \mathcal{N}\big(\mathbf{x}_{\mathcal{I}_j}, \frac{0.1^2}{d_j}\mathbf{I}_{d_j}\big) & n \le 2d_j \\ \mathcal{N}\Big(\mathbf{x}_{\mathcal{I}_j}, \frac{\sigma_j^2}{\max_{i\in\mathcal{I}_j}\{\mathbf{\Sigma}_{n,\mathcal{I}_j}(i,i)\}}\big(\mathbf{\Sigma}_{n,\mathcal{I}_j} + \beta\mathbf{I}_{d_j}\big)\Big) & n > 2d_j \end{cases} \tag{11}$$

   - With probability $\min\Big\{1, \pi_{\text{post}}(\mathbf{y}_{-\mathcal{I}_j}, \mathbf{z} \,|\, \tilde{\mathbf{d}})/\pi_{\text{post}}(\mathbf{y} \,|\, \tilde{\mathbf{d}})\Big\}$, set $\mathbf{y}_{\mathcal{I}_j} = \mathbf{z}$, otherwise $\mathbf{y}_{\mathcal{I}_j}$ unchanged.
2. Then $\mathbf{x}_{n+1} = \mathbf{y}$ after updating all the L groups of components.
3. For a pre-specified batch number N, if $n \bmod N = 0$, for all $j = 1 \ldots, L$:
   - Calculate the acceptance rate $\hat{\alpha}_j$ from the past N updates on jth group of components,
   - If $\hat{\alpha}_j > 0.234$, $\sigma_j = \sigma_j \exp(\delta)$, otherwise $\sigma_j = \sigma_j \exp(-\delta)$.

   Here, $\delta = \min\{0.01, \sqrt{N/n}\}$.

In the proposal distribution (11), $1/\max_{i\in\mathcal{I}_j}\{\mathbf{\Sigma}_{n,\mathcal{I}_j}(i,i)\}$ gives the inverse of the largest variance among parameters in the group $\mathcal{I}_j$. This factor approximately normalizes the covariance matrix $\mathbf{\Sigma}_{n,\mathcal{I}_j}$, and hence avoids the interaction of $\mathbf{\Sigma}_{n,\mathcal{I}_j}$ and σ_j during the adaptation. Such numerical treatment makes the scale variable σ_j stabilize faster.

The ergodicity of GCAM can be shown using the simplified conditions of Roberts and Rosenthal. [14] We will discuss this further in Section 4.

3. APPROXIMATIONS TO THE FORWARD MAP AND POSTERIOR DISTRIBUTION

Apart from the high dimensionality and complex nature of the posterior, a significant computational challenge arises from the high computational cost of evaluating the posterior density that entails computationally demanding numerical schemes used by the forward model $F(\cdot)$. For example, the 3D geothermal reservoir model presented in Section 5 has about ten thousand parameters, and each model evaluation takes about 30 to 50 minutes CPU time.

To improve the computational efficiency of MCMC as discussed previously, we approximate the likelihood, and hence the posterior, by employing reduced models. Efficiency of the MCMC requires the reduced model to be accurate and cheap, though these requirements are application specific.

The backbone of the approximation developed here is a given reduced model built using existing techniques, including grid coarsening [36, 13, 16, 35], linearization of the forward model [11], and projection-based methods [37, 38, 39, 40, 41].

Our key contribution here is to present a new way to improve the approximation to the likelihood function by considering the posterior statistics of the numerical error of the reduced model.

This leads to the adaptive delayed acceptance algorithm with substantial improvement in the computational efficiency compared to the classical delayed acceptance or surrogate transition algorithms.

In this section, we first present posterior approximation using reduced models and the classical approximation error models, then discuss various ways to estimate posterior error statistics and correct the approximated posteriors.

3.1. Approximation Error Models

Let $F^*(\cdot)$ denote the reduced model. In the computed examples in Section 5 $F^*(\cdot)$ is defined by a coarse discretization, however, in other applications $F^*(\cdot)$ may be a linearization of $F(\cdot)$, or a projection-based reduced model, or any other approximation; the only restriction on $F^*(\cdot)$ are the conditions required for DA to be valid [11], essentially that $F^*(\cdot)$ is finite valued where $F^*(\cdot)$ is finite valued, which is virtually no restriction in practice.

We start with the common approximation to the posterior distribution that simply uses $F^*(\cdot)$ in place of the true forward map $F(\cdot)$:

Approximation 1
Approximate posterior distribution using $F^*(\cdot)$ in place of $F(\cdot)$, which has the probability density function

$$\pi^*_{\text{post}}(\mathbf{x}\,|\,\tilde{\mathbf{d}}) \propto \exp\left[-\frac{1}{2}\big(F^*(\mathbf{x})-\tilde{\mathbf{d}}\big)^\top \boldsymbol{\Sigma}_{\mathbf{e}}^{-1}\big(F^*(\mathbf{x})-\tilde{\mathbf{d}}\big)\right]\pi_{\text{prior}}(\mathbf{x}). \tag{12}$$

Using Approximation 1 in DA, Alg. 2, gives the surrogate transition method [15, 16].

The reduced model usually has a non-negligible discrepancy with the forward model, and hence the approximation in Eqn. (12), by itself, can result in biased estimates while producing uncertainty intervals that are too small. This effect was investigated by [2, 13], who noted that this is one way to perform an "inverse crime". This indicates that the approximate posterior has displaced support and is too narrow to include the support of the true posterior.

One possible remedy is afforded by considering the statistics of the numerical error of the reduced model. Given a reduced model $F^*(\cdot)$, Eqn. (2) can be expressed as

$$\begin{aligned}\tilde{\mathbf{d}} &= F^*(\mathbf{x}) + \big(F(\mathbf{x}) - F^*(\mathbf{x})\big) + \mathbf{e}\\ &= F^*(\mathbf{x}) + B(\mathbf{x}) + \mathbf{e},\end{aligned} \tag{13}$$

where $B(\mathbf{x})$ is the model reduction error between the true forward model and the reduced model. Assuming that the model reduction error is independent of the model parameters and normally distributed gives the approximation error model (AEM) [13]

$$\tilde{\mathbf{d}} = F^*(\mathbf{x}) + B + \mathbf{e},$$

where $B \sim N(\boldsymbol{\mu}_B, \boldsymbol{\Sigma}_B)$. Kaipio & Somersalo [13] showed that this improved the approximation of the posterior distribution, compared to Approximation 1, in their applications, using an *a priori* construction of the AEM. That is, before utilizing data and solving the inverse problem, the AEM was empirically estimated over the prior distribution. The resulting estimates for mean and covariance of B are

$$\mu_B = \int_{\mathcal{X}} B(\mathbf{x})\pi_{\text{prior}}(\mathbf{x})d\mathbf{x} \approx \frac{1}{L}\sum_{i=1}^{L} B(\mathbf{x}_i), \tag{14}$$

$$\Sigma_B = \int_{\mathcal{X}} \big(B(\mathbf{x}) - \mu_B\big)\big(B(\mathbf{x}) - \mu_B\big)^\top \pi_{\text{prior}}(\mathbf{x}) d\mathbf{x} \approx \frac{1}{L-1}\sum_{i=1}^{L} \big(B(\mathbf{x}_i) - \mu_B\big)\big(B(\mathbf{x}_i) - \mu_B\big)^\top, \tag{15}$$

where $B(\mathbf{x})$ is defined by Eqn. 13, and $\mathbf{x}_i \sim \pi_{\text{prior}}(\mathbf{x}), i = 1, \cdots, L$, are L samples drawn from the prior distribution.

Using $\boldsymbol{\mu}_B$ and $\boldsymbol{\Sigma}_B$ estimated from the prior distribution, we have the following approximate posterior.

Approximation 2
Approximate posterior distribution using the reduced model $F^*(\cdot)$ and AEM estimated from the prior, which has the probability density function

$$\pi^*_{\text{post}}(\mathbf{x} \,|\, \tilde{\mathbf{d}}) \propto \exp\left[-\frac{1}{2}\big(F^*(\mathbf{x}) + \boldsymbol{\mu}_B - \tilde{\mathbf{d}}\big)^\top \big(\boldsymbol{\Sigma}_B + \boldsymbol{\Sigma}_{\mathbf{e}}\big)^{-1}\big(F^*(\mathbf{x}) + \boldsymbol{\mu}_B - \tilde{\mathbf{d}}\big)\right] \pi_{\text{prior}}(\mathbf{x}). \tag{16}$$

3.2. Posterior Approximation Error Models

There are two major drawbacks when using the AEM estimated from the prior.

Firstly,

the support of the prior is typically quite different to the support of the posterior, as the observed data in the likelihood necessarily make the posterior concentrate compared to the prior.

Hence, the AEM may be reasonable over the prior distribution but the L samples may not include any samples with appreciable posterior probability so the AEM could be far from optimal over the support of the posterior distribution.

Secondly, the *a priori* approach requires appreciable pre-computation to construct the AEM.

We improve the AEM, and hence make a better approximate posterior distribution, by empirically estimating the AEM over the posterior distribution. That is, we estimate the mean and covariance of the AEM by

$$\mu_B = \int_{\mathcal{X}} B(\mathbf{x}) \pi_{\text{post}}(\mathbf{x} \,|\, \tilde{\mathbf{d}}) d\mathbf{x} \qquad , \tag{17}$$

$$\Sigma_B = \int_{\mathcal{X}} \big(B(\mathbf{x}) - \mu_B\big)\big(B(\mathbf{x}) - \mu_B\big)^\top \pi_{\text{post}}(\mathbf{x} \,|\, \tilde{\mathbf{d}}) d\mathbf{x}. \tag{18}$$

For each accepted posterior sample of DA (Alg. 2), we can calculate the model error $B(\mathbf{x})$ between full model and the reduced model, since both are evaluated at each acceptance. Since DA can sample the full posterior, we can adaptively estimate μ_B and Σ_B from the posterior samples during the simulation of DA. A carefully designed adaptive MCMC can make the *a posteriori* estimates of μ_B and Σ_B converge to the true values given in Eqns 17 and 18. Let n denote the MCMC iteration, the estimated mean and covariance of AEM denoted by $\bar{\boldsymbol{\mu}}_{B,n}$ and $\bar{\boldsymbol{\Sigma}}_{B,n}$, respectively, can be iteratively updated as

$$\bar{\boldsymbol{\mu}}_{B,n} = \frac{1}{n}\big[(n-1)\bar{\boldsymbol{\mu}}_{B,n-1} + B_{\mathbf{x}_{n-1}}(\mathbf{x}_n)\big] \qquad , \tag{19}$$

$$\bar{\boldsymbol{\Sigma}}_{B,n} = \frac{1}{n-1}\big[(n-2)\bar{\boldsymbol{\Sigma}}_{B,n-1} + B_{\mathbf{x}_{n-1}}(\mathbf{x}_n) B_{\mathbf{x}_{n-1}}(\mathbf{x}_n)^\top\big]. \tag{20}$$

This defines an approximate posterior that changes adaptively over MCMC simulations.

Approximation 3
Adaptive approximate posterior distribution using the reduced model $F^*(\cdot)$ and AEM adaptively estimated over the posterior, giving the posterior probability density function at iteration n

$$\pi^*_{n,\text{post}}(\mathbf{x} \,|\, \tilde{\mathbf{d}}) \propto \exp\left[-\frac{1}{2}\big(F^*(\mathbf{x}) + \bar{\boldsymbol{\mu}}_{B,n} - \tilde{\mathbf{d}}\big)^\top \big(\bar{\boldsymbol{\Sigma}}_{B,n} + \boldsymbol{\Sigma}_{\mathbf{e}}\big)^{-1}\big(F^*(\mathbf{x}) + \bar{\boldsymbol{\mu}}_{B,n} - \tilde{\mathbf{d}}\big)\right] \pi_{\text{prior}}(\mathbf{x}). \tag{21}$$

We can use the adaptive posterior in Approximation 3 within DA to speed-up MCMC sampling, while using the resulting posterior samples to simultaneously update the AEM in the approximation. A further advantage of this adaptive approach is that it does not require any pre-computation to estimate the AEM before setting up a MCMC simulation. Indeed, in all computational experiments we find that building the AEM over the posterior leads to better statistical efficiency in the MCMC, than when the AEM is estimated over the prior, and so gives a method that both does not require precomputation and also gives a more computationally efficient MCMC.

A critical issue arises because the approximate posterior changes over iterations in the MCMC simulation. As with the adaptive Metropolis case, it becomes unclear whether DA using this adaptive approximation is ergodic. We extend the current framework for adaptive MCMC [14, 18], in Section 4, to establish ergodicity of the resulting DA scheme using adaptive approximate posteriors.

3.3. State-dependent Approximation Error Models

Christen & Fox [11] demonstrated DA using a local linearization of the forward map as the approximate forward map, which is a local reduced model that depends on the current state $\mathbf{x}$ of the MCMC. For the applications we present in Section 5, we utilize the existing Fortran code TOUGH2 [42] to simulate the forward map. This package does not give access to derivatives (nor adjoints, etc) and so we form a reduced model $F^*(\cdot)$ by using a coarsened discretization. This reduced model depends only on the point $\mathbf{y}$ at which it is evaluated, but not the state of the MCMC.

In many cases, including in the applications we consider here, both the true forward map and reduced model are P-Hölder continuous, i.e., $\|F(\mathbf{y}) - F(\mathbf{x})\| \leq C\|\mathbf{y} - \mathbf{x}\|^P$ for some $C > 0$ and $P > 0$. Then, when using DA, a local improvement to the state-independent reduced model can be made, for little additional computational cost, by using the values of $F(\mathbf{y})$ and $F^*(\mathbf{y})$ for points $\mathbf{y}$ that are accepted, and hence become the state of the chain. We define a deterministic state-dependent reduced model using a zeroth-order correction to the reduced model, as follows.

Approximation 4
State-dependent reduced model and approximate posterior distribution: Suppose that at iteration n, the Markov chain has state $\mathbf{x}_n = \mathbf{x}$. For a proposed state $\mathbf{y} \sim q(\mathbf{x}, \cdot)$, the state-dependent reduced model $F^*_{\mathbf{x}}(\cdot)$ is

$$F^*_{\mathbf{x}}(\mathbf{y}) = F^*(\mathbf{y}) + \big(F(\mathbf{x}) - F^*(\mathbf{x})\big). \tag{22}$$

The resulting approximate posterior density function is

$$\pi^*_{\mathbf{x}}(\mathbf{y} \mid \tilde{\mathbf{d}}) \propto \exp\left[-\frac{1}{2}\big(F^*_{\mathbf{x}}(\mathbf{y}) - \tilde{\mathbf{d}}\big)^\top \boldsymbol{\Sigma}_{\mathbf{e}}^{-1}\big(F^*_{\mathbf{x}}(\mathbf{y}) - \tilde{\mathbf{d}}\big)\right] \pi_{\text{prior}}(\mathbf{y}). \tag{23}$$

It is worth mentioning that the state-dependent reduced model (22) has the desirable property that $F^*_{\mathbf{x}}(\mathbf{x}) = F(\mathbf{x})$.

The error structure of the state-dependent reduced model (22) can also be estimated by employing the AEM. In particular, Approximation 3 and 4 can be combined, at no significant increase in computational cost, to give a more accurate approximation to the posterior distribution. The state-dependent model reduction error, for the zeroth-order correction (22), is

$$B_{\mathbf{x}}(\mathbf{y}) = F(\mathbf{y}) - F^*_{\mathbf{x}}(\mathbf{y}),$$

where $B_{\mathbf{x}}(\mathbf{x}) = \mathbf{0}$.

This way, the mean and covariance of the AEM in Approximation 5 with reduced model (22) are

$$\boldsymbol{\mu}_B = \quad \mathrm{E}_{\pi_{\text{post}}}\left\{\int_{\mathcal{X}} B_{\mathbf{x}}(\mathbf{y}) K(\mathbf{x}, \mathbf{y}) d\mathbf{y}\right\}, \tag{24}$$

$$\boldsymbol{\Sigma}_B = \mathrm{Cov}_{\pi_{\text{post}}}\left\{\int_{\mathcal{X}} B_{\mathbf{x}}(\mathbf{y}) K(\mathbf{x}, \mathbf{y}) d\mathbf{y}\right\}, \tag{25}$$

where $K(\mathbf{x}, \mathbf{y})$ is the transition kernel implemented by the MCMC iteration.

The mean of the AEM (24) for reduced model (22) can be shown to be $\mathbf{0}$, as follows.

$$\mathbb{E}_{\pi_{\text{post}}}\left[\int B_{\mathbf{x}}(\mathbf{y})K(\mathbf{x},\mathbf{y})d\mathbf{y}\right] = \int_{\mathcal{X}}\int_{\mathcal{X}}\big(F(\mathbf{y})-F^*(\mathbf{y})\big)\pi_{\text{post}}(\mathbf{x}\,|\,\tilde{\mathbf{d}})K(\mathbf{x},\mathbf{y})d\mathbf{y}d\mathbf{x} - \int_{\mathcal{X}}\int_{\mathcal{X}}\big(F(\mathbf{x})-F^*(\mathbf{x})\big)\pi_{\text{post}}(\mathbf{x}\,|\,\tilde{\mathbf{d}})K(\mathbf{x},\mathbf{y})d\mathbf{y}d\mathbf{x}$$

with the two terms on the right canceling because the kernel K satisfies the detailed balance condition $\pi_{\text{post}}(\mathbf{x}\,|\,\tilde{\mathbf{d}})K(\mathbf{x},\mathbf{y}) = \pi_{\text{post}}(\mathbf{y}\,|\,\tilde{\mathbf{d}})K(\mathbf{y},\mathbf{x})$. Since we have $\boldsymbol{\mu}_B = \mathbf{0}$, the covariance of the model reduction error $B_{\mathbf{x}}(\mathbf{y})$ can be computed adaptively at iteration n by the inductive formula

$$\hat{\boldsymbol{\Sigma}}_{B,n} = \frac{1}{n-1}\left[(n-2)\hat{\boldsymbol{\Sigma}}_{B,n-1} + B_{\mathbf{x}_{n-1}}(\mathbf{x}_n)B_{\mathbf{x}_{n-1}}(\mathbf{x}_n)^\top\right]. \tag{26}$$

The approximate posterior distribution based on state-dependent reduced model 22 and AEM estimated from the posterior takes the following form.

Approximation 5
AEM built over the posterior distribution with a state-dependent reduced model: Suppose that at iteration n the Markov chain is at state $\mathbf{x}_n = \mathbf{x}$ and a proposed state is $\mathbf{y} \sim q(\mathbf{x},\cdot)$. The state-dependent approximate posterior density function is given by

$$\pi^*_{n,\mathbf{x}}(\mathbf{y}\,|\,\tilde{\mathbf{d}}) \propto \exp\left[-\frac{1}{2}\big(F^*_{\mathbf{x}}(\mathbf{y})-\tilde{\mathbf{d}}\big)^\top\big(\hat{\boldsymbol{\Sigma}}_{B,n}+\boldsymbol{\Sigma}_{\mathbf{e}}\big)^{-1}\big(F^*_{\mathbf{x}}(\mathbf{y})-\tilde{\mathbf{d}}\big)\right]\pi_{\text{prior}}(\mathbf{y}). \tag{27}$$

As for Approximation 3 we have to show ergodicity of the resulting adaptive MCMC scheme.

4. ADAPTIVE DELAYED ACCEPTANCE ALGORITHM

The ADA algorithm combines the adaptive AEM built over the posterior, suggested in Section 3, and the adaptation of the proposal distribution as in existing adaptive algorithms in Section 2.5. Adaptive estimates of the AEM are made possible by using the DA algorithm, which provides the basic structure of ADA, and also the mechanism for reduction in compute cost per iteration. The use of adaptively improved stochastic approximations means the reduction in statistical efficiency, which is an unavoidable consequence of using DA, may be minimized so the reduction of compute cost per iteration is transferred directly into computational efficiency of the resulting algorithm.

We summarize ADA in Alg. 5. In this algorithm, the proposal $q_n(\cdot,\cdot)$ and its adaptive update in Step 4 may have the form of any of the classical adaptive MCMC algorithms, such as the AM in Alg. 3 and the GCAM in Alg. 4. We use the general notation $\pi^*_{n,\mathbf{x}}(\,\cdot\,|\,\tilde{\mathbf{d}})$ to denote the approximate posterior. It includes the state-dependent, state-independent, adaptive, and non-adaptive cases. When the adaptive approximate posteriors 3 and 5 are used, the corresponding adaptive error models have to be updated in Step 3 according to Eqn 19 and 20 and Eqn. 26, respectively.

We note that ADA in Alg. 5 is not restricted to the form of the approximate posteriors used here. It offers a general framework for constructing other forms of posterior approximations in Step 3 using the forward model evaluated at past posterior samples. In the rest of this section, we present regularity conditions on the adaptive approximation and adaptive proposal that guarantee ergodicity of ADA. These regularity conditions provide useful guidelines to the construction of other posterior approximations in future research.

4.1. Ergodicity conditions and main result

In this section, we follow the notation in [14] to formalize ADA. Suppose the parameter space $\mathcal{X}$ is equipped with σ-algebra $\mathcal{B}(\mathcal{X})$. Without loss of generality, we define all the proposal densities, target densities and its approximations with respect to the Lebesgue measure here.

To simplify the notation, we use $\pi(\cdot) \equiv \pi_{\text{post}}(\,\cdot\,|\,\tilde{\mathbf{d}})$ denote the exact posterior density. We parametrize (potentially state-dependent) adaptive approximate posteriors and adaptive proposals

Algorithm 5 Adaptive delayed acceptance Metropolis-Hastings (ADA)

At iteration n, given $\mathbf{x}_n = \mathbf{x}$, adaptive proposal $q_n(\mathbf{x}, \cdot)$, and approximate posterior distribution $\pi^*_{n,\mathbf{x}}(\cdot \mid \tilde{\mathbf{d}})$, then $\mathbf{x}_{n+1}$ and updated distributions are determined as follows:

1. Generate a proposal $\mathbf{y} \sim q_n(\mathbf{x}, \cdot)$. With probability
$$\alpha_n(\mathbf{x}, \mathbf{y}) = \min\left\{1, \frac{\pi^*_{n,\mathbf{x}}(\mathbf{y} \mid \tilde{\mathbf{d}})\, q_n(\mathbf{y}, \mathbf{x})}{\pi^*_{n,\mathbf{x}}(\mathbf{x} \mid \tilde{\mathbf{d}})\, q_n(\mathbf{x}, \mathbf{y})}\right\}$$
promote $\mathbf{y}$ to be used as a proposal for the following step. Otherwise set $\mathbf{y} = \mathbf{x}$ and proceed.
2. The proposal distribution at this step is
$$q^*_n(\mathbf{x}, \mathbf{y}) = \alpha_n(\mathbf{x}, \mathbf{y}) q_n(\mathbf{x}, \mathbf{y}) + \Big[1 - \int_{\mathcal{X}} \alpha_n(\mathbf{x}, \mathbf{y}) q_n(\mathbf{x}, \mathbf{y})\, \mathrm{d}\mathbf{y}\Big] \delta_{\mathbf{x}}(\mathbf{y}),$$
where $\delta_{\mathbf{x}}(\cdot)$ denotes the Dirac mass at $\mathbf{x}$. With probability
$$\min\left\{1, \frac{\pi_{\text{post}}(\mathbf{y} \mid \tilde{\mathbf{d}}) q^*_n(\mathbf{y}, \mathbf{x})}{\pi_{\text{post}}(\mathbf{x} \mid \tilde{\mathbf{d}}) q^*_n(\mathbf{x}, \mathbf{y})}\right\}$$
set $\mathbf{x}_{n+1} = \mathbf{y}$. Otherwise set $\mathbf{x}_{n+1} = \mathbf{x}$.
3. Form the updated approximation $\pi^*_{n+1,\mathbf{x}}(\cdot \mid \tilde{\mathbf{d}})$.
4. Form the updated adaptive proposal $q_{n+1}(\mathbf{x}, \cdot)$.

Here we use the general notation $\pi^*_{n,\mathbf{x}}(\cdot \mid \tilde{\mathbf{d}})$ to denote the approximate posterior. It can be either state-dependent or state-independent, either adaptive or non-adaptive.

using adaptation indices $\boldsymbol{\nu} \in \mathcal{V}$ and $\boldsymbol{\xi} \in \mathcal{E}$, respectively. Here $\mathcal{V}$ and $\mathcal{E}$ denote the space of adaptation indices. We use the adaptation indices to replace the iteration number n in Algorithm 5, since the former provide unique notations for functions. This way, we denote the adaptive approximate posterior as $\pi^*_{\boldsymbol{\xi},\mathbf{x}}(\cdot)$ and the adaptive proposal as $q_{\boldsymbol{\nu}}(\mathbf{x}, \cdot)$ in this section. For example, in Approximation 3, the adaptation index is $\boldsymbol{\xi} = (\bar{\boldsymbol{\mu}}_{B,n}, \bar{\boldsymbol{\Sigma}}_{B,n})$ that define the AEM, while in the GCAM algorithm 4, the adaptation indices are $\boldsymbol{\nu} = \{\sigma_i, \boldsymbol{\Sigma}_{n,\mathcal{I}_j}\}_{i=1}^{L}$.

We first introduce the transition kernels involved in ADA.

Definition 1
For each $\boldsymbol{\nu} \in \mathcal{V}$, we can define the transition kernel of a single level MH algorithm with target density $\pi(\cdot)$ as

$$K_{\boldsymbol{\nu}}(\mathbf{x}, \mathbf{y}) = q_{\boldsymbol{\nu}}(\mathbf{x}, \mathbf{y})\, \alpha_{\boldsymbol{\nu}}(\mathbf{x}, \mathbf{y}) + \Big[1 - \int_{\mathcal{X}} q_{\boldsymbol{\nu}}(\mathbf{x}, \mathbf{z})\, \alpha_{\boldsymbol{\nu}}(\mathbf{x}, \mathbf{z}) d\mathbf{z}\Big] \delta_{\mathbf{x}}(\mathbf{y}),$$

where

$$\alpha_{\boldsymbol{\nu}}(\mathbf{x}, \mathbf{y}) = \min\left\{1, \frac{\pi(\mathbf{y})\, q_{\boldsymbol{\nu}}\,(\mathbf{y}, \mathbf{x})}{\pi(\mathbf{x})\, q_{\boldsymbol{\nu}}\,(\mathbf{x}, \mathbf{y})}\right\}.$$

This way, $\{K_{\boldsymbol{\nu}}\}_{\boldsymbol{\nu} \in \mathcal{V}}$ defines a family of Markov chain transition kernels (associated with MH) on $\mathcal{X}$ with the target $\pi(\cdot)$.

Definition 2
For each pair of $\boldsymbol{\nu} \in \mathcal{V}$ and $\boldsymbol{\xi} \in \mathcal{E}$, we have the first step and second step acceptance probabilities

$$\alpha_{\boldsymbol{\nu},\boldsymbol{\xi}}(\mathbf{x}, \mathbf{y}) = \min\left\{1, \frac{\pi^*_{\boldsymbol{\xi},\mathbf{x}}(\mathbf{y})\, q_{\boldsymbol{\nu}}\,(\mathbf{y}, \mathbf{x})}{\pi^*_{\boldsymbol{\xi},\mathbf{x}}(\mathbf{x})\, q_{\boldsymbol{\nu}}\,(\mathbf{x}, \mathbf{y})}\right\}, \text{ and } \beta_{\boldsymbol{\nu},\boldsymbol{\xi}}(\mathbf{x}, \mathbf{y}) = \min\left\{1, \frac{\pi(\mathbf{y})\, q^*_{\boldsymbol{\nu},\boldsymbol{\xi}}\,(\mathbf{y}, \mathbf{x})}{\pi(\mathbf{x})\, q^*_{\boldsymbol{\nu},\boldsymbol{\xi}}\,(\mathbf{x}, \mathbf{y})}\right\},$$

respectively, where the effective proposal in the second step has the density

$$q^*_{\boldsymbol{\nu},\boldsymbol{\xi}}(\mathbf{x},\mathbf{y}) = q_{\boldsymbol{\nu}}(\mathbf{x},\mathbf{y})\ \alpha_{\boldsymbol{\nu},\boldsymbol{\xi}}(\mathbf{x},\mathbf{y}) + \Big[1 - \int_{\mathcal{X}} q_{\boldsymbol{\nu}}(\mathbf{x},\mathbf{z})\ \alpha_{\boldsymbol{\nu},\boldsymbol{\xi}}(\mathbf{x},\mathbf{z})\,\mathrm{d}\mathbf{z}\Big]\delta_{\mathbf{x}}(\mathbf{y}).$$

We can define the transition kernel of ADA algorithm with target $\pi(\cdot)$ as

$$K_{\boldsymbol{\nu},\boldsymbol{\xi}}(\mathbf{x},\mathbf{y}) = q_{\boldsymbol{\nu}}(\mathbf{x},\mathbf{y})\,\alpha_{\boldsymbol{\nu},\boldsymbol{\xi}}(\mathbf{x},\mathbf{y})\,\beta_{\boldsymbol{\nu},\boldsymbol{\xi}}(\mathbf{x},\mathbf{y}) + \Big[1 - \int_{\mathcal{X}} q_{\boldsymbol{\nu}}(\mathbf{x},\mathbf{z})\,\alpha_{\boldsymbol{\nu},\boldsymbol{\xi}}(\mathbf{x},\mathbf{z})\,\beta_{\boldsymbol{\nu},\boldsymbol{\xi}}(\mathbf{x},\mathbf{z})d\mathbf{z}\Big]\delta_{\mathbf{x}}(\mathbf{y}).$$

This way, $\{K_{\boldsymbol{\nu},\boldsymbol{\xi}}\}_{\boldsymbol{\nu}\in\mathcal{V},\boldsymbol{\xi}\in\mathcal{E}}$ defines a family of Markov chain transition kernels (associated with ADA) on $\mathcal{X}$ with target $\pi(\cdot)$.

In ADA, the adaptation indices $\boldsymbol{\nu}$ and $\boldsymbol{\xi}$ are respectively updated by a $\mathcal{V}$-valued random variable $\boldsymbol{\Gamma}_n$ and a $\mathcal{E}$-valued random variable $\boldsymbol{\Xi}_n$ at each step. In contrast, DA only employs fixed parameters $\boldsymbol{\nu}$ and $\boldsymbol{\xi}$. Although for each pair of $\boldsymbol{\nu} \in \mathcal{V}$ and $\boldsymbol{\xi} \in \mathcal{E}$, the resulting DA scheme can be ergodic, there is no guarantee that the ADA scheme will be ergodic if the adaptation on $\boldsymbol{\nu}$ and $\boldsymbol{\xi}$ are not carefully constructed.

Utilizing ergodic theory of standard adaptive MCMC that only adapts on the proposal, we aim to establish regularity conditions for ADA to be ergodic. The key is to analyze the behaviour of the effective proposal $q^*_{\boldsymbol{\nu},\boldsymbol{\xi}}(\mathbf{x},\mathbf{y})$ in Definition 2 – which involves both the proposal and the approximate posterior – and the associated transition kernel $K_{\boldsymbol{\nu},,\boldsymbol{\xi}}(\mathbf{x},\mathbf{y})$ during adaptation. By considering only the proposal adaptation (indexed by $\boldsymbol{\nu} \in \mathcal{V}$), Roberts and Rosenthal [14] provide conditions for constructing ergodic adaptive MCMC algorithms. We first restate Theorem 1 of Roberts and Rosenthal [14] with a small extension required for ADA, including both adaptation indices $\boldsymbol{\nu} \in \mathcal{V}$ and $\boldsymbol{\xi} \in \mathcal{E}$.

Theorem 1
Suppose we have a target density $\pi(\cdot)$ defined on a parameter space $\mathcal{X}$. Suppose an ADA algorithm with $\mathcal{V}$-valued proposal adaptation index and $\mathcal{E}$-valued approximation adaptation index is ergodic for $\pi(\cdot)$ for given $\boldsymbol{\nu} \in \mathcal{V}$ and $\boldsymbol{\xi} \in \mathcal{E}$. Under the following conditions, ADA is ergodic:

1. (Simultaneous uniform ergodicity.) For all $\epsilon > 0$, there exist $n = n(\epsilon) \in \mathbb{N}$ such that

$$\|K^n_{\boldsymbol{\nu},\boldsymbol{\xi}}(\mathbf{x},\cdot) - \pi(\cdot)\|_{TV} \leq \epsilon,$$

   for any $\mathbf{x} \in \mathcal{X}$, $\boldsymbol{\nu} \in \mathcal{V}$ and $\boldsymbol{\xi} \in \mathcal{E}$. Here $K^n_{\boldsymbol{\nu},\boldsymbol{\xi}}(\mathbf{x},\mathbf{y})$ is the n-step transition kernel defined as

$$K^n_{\boldsymbol{\nu},\boldsymbol{\xi}}(\mathbf{x},\mathbf{y}) = \int_{\mathcal{X}} K^{n-1}_{\boldsymbol{\nu},\boldsymbol{\xi}}(\mathbf{x},\mathbf{z})K_{\boldsymbol{\nu},\boldsymbol{\xi}}(\mathbf{z},\mathbf{y})d\mathbf{z}.$$

2. (Diminishing adaptation.) In two consecutive iterations n and $n+1$, the transition kernels satisfy

$$\lim_{n\to\infty} \sup_{\mathbf{x}\in\mathcal{X}} \|K_{\boldsymbol{\nu}_{n+1},\boldsymbol{\xi}_{n+1}}(\mathbf{x},\cdot) - K_{\boldsymbol{\nu}_n,\boldsymbol{\xi}_n}(\mathbf{x},\cdot)\|_{TV} = 0.$$

Here $\|\lambda_1(\cdot) - \lambda_2(\cdot)\|_{TV} = \sup_{\mathbf{A}\in\mathcal{B}(\mathcal{X})} \|\lambda_1(\mathbf{A}) - \lambda_2(\mathbf{A})\|$ is the total variational distance. †

Proof
The proof directly follows from the Theorem 1 of Roberts and Rosenthal. [14]. □

As sufficient conditions for an MCMC with adaptive proposals satisfying simultaneous uniform ergodicity and diminishing adaptation are well understood in the literature, here we focus on establishing sufficient conditions on the adaptation of posterior approximations to make the resulting ADA ergodic. We prove the following two theorems that separately show ADA can satisfy simultaneous uniform ergodicity and diminishing adaptation by imposing mild regularity conditions.

†We denote $\lambda(\mathbf{A}) = \int_{\mathbf{A}} \lambda(\mathbf{x})d\mathbf{x}$ for $\mathbf{A} \in \mathcal{B}(\mathcal{X})$.

Theorem 2
Suppose an ADA algorithm with the target density $\pi(\cdot)$ has a family of first step proposal densities $\{q_{\boldsymbol{\nu}}(\mathbf{x},\cdot)\}_{\boldsymbol{\nu}\in\mathcal{V}}$, a family of approximate target densities $\{\pi^*_{\boldsymbol{\xi},x}(\cdot)\}_{\boldsymbol{\nu}\in\mathcal{V},\boldsymbol{\xi}\in\mathcal{E}}$, and a family of transition kernels $\{K_{\boldsymbol{\nu},\boldsymbol{\xi}}(\mathbf{x},\cdot)\}_{\boldsymbol{\nu}\in\mathcal{V},\boldsymbol{\xi}\in\mathcal{E}}$.

The transition kernels satisfy simultaneous uniform ergodicity given the following sufficient conditions.

1. The spaces $\mathcal{X}$, $\mathcal{V}$, and $\mathcal{E}$ are compact.
2. The target density is Lipschitz continuous in $\mathbf{x}$.
3. For any $\boldsymbol{\nu}\in\mathcal{V}$, if one applies the proposal $q_{\boldsymbol{\nu}}(\mathbf{x},\cdot)$ in a single-level MH with target $\pi(\cdot)$, then each transition kernel $K_{\boldsymbol{\nu}}$ (as defined in Definition 1) is ergodic for $\pi(\cdot)$.
4. For any $\boldsymbol{\nu}\in\mathcal{V}$, the proposal $q_{\boldsymbol{\nu}}(\mathbf{x},\cdot)$ is uniformly bounded, and for each fixed $\mathbf{y}\in\mathcal{X}$, the mapping $(\mathbf{x},\boldsymbol{\nu})\mapsto q_{\boldsymbol{\nu}}(\mathbf{x},\mathbf{y})$ is Lipschitz continuous in $\mathbf{x}$ and $\boldsymbol{\nu}$.
5. The mapping $(\mathbf{x},\mathbf{y},\boldsymbol{\xi})\mapsto \log\pi^*_{\boldsymbol{\xi},\mathbf{x}}(\mathbf{y})$ is Lipschitz continuous in $\mathbf{x}$, $\mathbf{y}$, and $\boldsymbol{\xi}$.

Proof
We extend the result of Corollary 5 of Roberts and Rosenthal [14] to prove this theorem. First note that, since $\mathcal{X}$, $\mathcal{Y}$, and $\mathcal{E}$ are compact, all product spaces are compact in the product topology. Since for a given pair of indices $(\boldsymbol{\nu},\boldsymbol{\xi})\in\mathcal{V}\times\mathcal{E}$, the ADA becomes a standard DA, employing Theorem 1 of DA [11], conditions (1), (3), and (5) imply that, an Markov chain induced by the transition kernel $K_{\boldsymbol{\nu},\boldsymbol{\xi}}$ is ergodic for $\pi(\cdot)$ for any pair of $(\boldsymbol{\nu},\boldsymbol{\xi})\in\mathcal{V}\times\mathcal{E}$.

Recall that the effective proposal in the second step of ADA has density

$$q^*_{\boldsymbol{\nu},\boldsymbol{\xi}}(\mathbf{x},\mathbf{y}) = q_{\boldsymbol{\nu}}(\mathbf{x},\mathbf{y})\ \alpha_{\boldsymbol{\nu},\boldsymbol{\xi}}(\mathbf{x},\mathbf{y}) + r_{\boldsymbol{\nu},\boldsymbol{\xi}}(\mathbf{x})\delta_{\mathbf{x}}(\mathbf{y}).$$

where $r_{\boldsymbol{\nu},\boldsymbol{\xi}}(\mathbf{x})$ is the probability of remaining at $\mathbf{x}$ after the first step, which is defined as

$$r_{\boldsymbol{\nu},\boldsymbol{\xi}}(\mathbf{x}) = 1 - \int_{\mathcal{X}} q_{\boldsymbol{\nu}}(\mathbf{x},\mathbf{y})\ \alpha_{\boldsymbol{\nu},\boldsymbol{\xi}}(\mathbf{x},\mathbf{z})\,\mathrm{d}\mathbf{z}.$$

For each fixed $\mathbf{y}\in\mathcal{X}$, it follows from conditions (4) and (5) that the map $(\mathbf{x},\boldsymbol{\nu},\boldsymbol{\xi})\mapsto q_{\boldsymbol{\nu}}(\mathbf{x},\mathbf{y})\alpha_{\boldsymbol{\nu},\boldsymbol{\xi}}(\mathbf{x},\mathbf{y})$ is Lipschitz continuous, as $q_{\boldsymbol{\nu}}(\mathbf{x},\mathbf{y})\,\alpha_{\boldsymbol{\nu},\boldsymbol{\xi}}(\mathbf{x},\mathbf{y})$ takes the form

$$q_{\boldsymbol{\nu}}(\mathbf{x},\mathbf{y})\ \alpha_{\boldsymbol{\nu},\boldsymbol{\xi}}(\mathbf{x},\mathbf{y}) = q_{\boldsymbol{\nu}}(\mathbf{x},\mathbf{y})\ \min\left\{1,\frac{\pi^*_{\boldsymbol{\xi},\mathbf{x}}(\mathbf{y})\,q_{\boldsymbol{\nu}}(\mathbf{y},\mathbf{x})}{\pi^*_{\boldsymbol{\xi},\mathbf{x}}(\mathbf{x})\,q_{\boldsymbol{\nu}}(\mathbf{x},\mathbf{y})}\right\} = \min\left\{q_{\boldsymbol{\nu}}(\mathbf{x},\mathbf{y})\ ,\frac{\pi^*_{\boldsymbol{\xi},\mathbf{x}}(\mathbf{y})}{\pi^*_{\boldsymbol{\xi},\mathbf{x}}(\mathbf{x})}\,q_{\boldsymbol{\nu}}(\mathbf{y},\mathbf{x})\right\}.$$

We also have the map $(\mathbf{x},\boldsymbol{\nu},\boldsymbol{\xi})\mapsto r_{\boldsymbol{\nu},\boldsymbol{\xi}}(\mathbf{x})$ is Lipschitz continuous by the bounded convergence theorem.

Thus, for each fixed $\mathbf{y}\neq\mathbf{x}$, the second step acceptance probability of ADA

$$\beta_{\boldsymbol{\nu},\boldsymbol{\xi}}(\mathbf{x},\mathbf{y}) = \min\left\{1,\frac{\pi(\mathbf{y})\,q^*_{\boldsymbol{\nu},\boldsymbol{\xi}}(\mathbf{y},\mathbf{x})}{\pi(\mathbf{x})\,q^*_{\boldsymbol{\nu},\boldsymbol{\xi}}(\mathbf{x},\mathbf{y})}\right\} = \min\left\{1,\frac{\pi(\mathbf{y})\,q_{\boldsymbol{\nu}}(\mathbf{y},\mathbf{x})\ \alpha_{\boldsymbol{\nu},\boldsymbol{\xi}}(\mathbf{y},\mathbf{x})}{\pi(\mathbf{x})\,q_{\boldsymbol{\nu}}(\mathbf{x},\mathbf{y})\ \alpha_{\boldsymbol{\nu},\boldsymbol{\xi}}(\mathbf{x},\mathbf{y})}\right\},$$

is jointly continuous in $\mathbf{x}$, $\boldsymbol{\nu}$ and $\boldsymbol{\xi}$.

We also have the probability of remaining at $\mathbf{x}$ after both steps 1 and 2 of ADA, which has the form

$$\rho_{\boldsymbol{\nu},\boldsymbol{\xi}}(\mathbf{x}) = 1 - \int_{\mathcal{X}} q_{\boldsymbol{\nu}}(\mathbf{x},\mathbf{z})\,\alpha_{\boldsymbol{\nu},\boldsymbol{\xi}}(\mathbf{x},\mathbf{z})\,\beta_{\boldsymbol{\nu},\boldsymbol{\xi}}(\mathbf{x},\mathbf{z})\,\mathrm{d}\mathbf{z},$$

which is jointly continuous in $\mathbf{x}$, $\boldsymbol{\nu}$ and $\boldsymbol{\xi}$ by the bounded convergence theorem.

Denoting $k_{\boldsymbol{\nu},\boldsymbol{\xi}}(\mathbf{x},\mathbf{y}) = q_{\boldsymbol{\nu}}(\mathbf{x},\mathbf{y})\,\alpha_{\boldsymbol{\nu},\boldsymbol{\xi}}(\mathbf{x},\mathbf{y})\,\beta_{\boldsymbol{\nu},\boldsymbol{\xi}}(\mathbf{x},\mathbf{y})$, we can decompose the transition kernel of ADA as

$$K_{\boldsymbol{\nu},\boldsymbol{\xi}}(\mathbf{x},\mathbf{y}) = k_{\boldsymbol{\nu},\boldsymbol{\xi}}(\mathbf{x},\mathbf{y}) + \rho_{\boldsymbol{\nu},\boldsymbol{\xi}}(\mathbf{x})\delta_{\mathbf{x}}(\mathbf{y}).$$

Note that the transition function $k_{\boldsymbol{\nu},\boldsymbol{\xi}}(\mathbf{x},\mathbf{y})$ is jointly continuous in $\mathbf{x}$, $\boldsymbol{\nu}$ and $\boldsymbol{\xi}$ for each fixed $\mathbf{y}\in\mathcal{X}$ following the above derivations.

Since the Dirac delta $\delta_{\mathbf{x}}(\mathbf{y})$ is a point mass, $\delta_{\mathbf{x}}(\mathbf{y})$ and the Lebesgue measure are orthogonal measures.

This way, iterating the transition kernel, we have the n-step transition kernel

$$K^n_{\boldsymbol{\nu},\boldsymbol{\xi}}(\mathbf{x},\mathbf{y}) = k^n_{\boldsymbol{\nu},\boldsymbol{\xi}}(\mathbf{x},\mathbf{y}) + \rho_{\boldsymbol{\nu},\boldsymbol{\xi}}(\mathbf{x})^n \delta_{\mathbf{x}}(\mathbf{y}),$$

in which the n-step transition function $k^n_{\boldsymbol{\nu},\boldsymbol{\xi}}(\mathbf{x},\mathbf{y})$ is also joint continuous in $\mathbf{x}$, $\boldsymbol{\nu}$ and $\boldsymbol{\xi}$. Furthermore, we have

$$\|K^n_{\boldsymbol{\nu},\boldsymbol{\xi}}(\mathbf{x},\cdot) - \pi(\cdot)\|_{TV} = \rho_{\boldsymbol{\nu},\boldsymbol{\xi}}(\mathbf{x})^n + \frac{1}{2}\int_{\mathcal{X}} \left|k^n_{\boldsymbol{\nu},\boldsymbol{\xi}}(\mathbf{x},\mathbf{y}) - \pi(\mathbf{y})\right| d\mathbf{y},$$

following the property of total variation distance that $\|\lambda_1(\cdot) - \lambda_2(\cdot)\|_{TV} = \frac{1}{2}\int_{\mathcal{X}} |\lambda_1(\mathbf{x}) - \lambda_2(\mathbf{x})| d\mathbf{x}$.

This quantity is again joint continuous in $\mathbf{x}$, $\boldsymbol{\nu}$ and $\boldsymbol{\xi}$ by the bounded convergence theorem.

In addition, for each fixed pair of $\boldsymbol{\nu}$ and $\boldsymbol{\xi}$, $\lim_{n\to\infty} \|K^n_{\boldsymbol{\nu},\boldsymbol{\xi}}(\mathbf{x},\cdot) - \pi(\cdot)\|_{TV}$ uniformly converges to zero in $\mathbf{x}$, $\boldsymbol{\nu}$ and $\boldsymbol{\xi}$ by ergodicity and compactness.

Therefore, the simultaneous uniform ergodicity condition holds given uniformly convergence and continuity. □

Theorem 3

Suppose an ADA algorithm with the target density $\pi(\cdot)$ has a family of transition kernels $\{K_{\boldsymbol{\nu},\boldsymbol{\xi}}(\mathbf{x},\cdot)\}_{\boldsymbol{\nu}\in\mathcal{V},\boldsymbol{\xi}\in\mathcal{E}}$.

Suppose further Conditions 1–5 of Theorem 2 hold.

The transition kernel satisfies the diminishing adaptation condition given the following conditions:

1. The proposal satisfies diminishing adaptation, that is,
$$\lim_{n\to\infty} \sup_{\mathbf{x}} \|q_{\boldsymbol{\nu}_{n+1}}(\mathbf{x},\cdot) - q_{\boldsymbol{\nu}_n}(\mathbf{x},\cdot)\|_{TV} = 0 \quad \text{in probability,}$$
2. The approximation adaptation index satisfies diminishing adaptation, that is, $\lim_{n\to\infty} \|\boldsymbol{\Xi}_{n+1} - \boldsymbol{\Xi}_n\| = 0$ in probability.

Proof

The first half of our proof uses the result of Lemma 4.21 in [29].

For simplicity, we define the transition probability of ADA as $p_{\boldsymbol{\nu},\boldsymbol{\xi}}(\mathbf{x},\mathbf{y}) = q_{\boldsymbol{\nu}}(\mathbf{x},\mathbf{y})\,\alpha_{\boldsymbol{\nu},\boldsymbol{\xi}}(\mathbf{x},\mathbf{y})$, which is the probability propose $\mathbf{y}$ from $\mathbf{x}$ and accept $\mathbf{y}$ in the first step of ADA. This way, the transition kernel of the ADA has the form

$$K_{\boldsymbol{\nu},\boldsymbol{\xi}}(\mathbf{x},\mathbf{y}) = p_{\boldsymbol{\nu},\boldsymbol{\xi}}(\mathbf{x},\mathbf{y})\,\beta_{\boldsymbol{\nu},\boldsymbol{\xi}}(\mathbf{x},\mathbf{y}) + \rho_{\boldsymbol{\nu},\boldsymbol{\xi}}(\mathbf{x})\delta_{\mathbf{x}}(\mathbf{y}), \quad \text{where } \rho_{\boldsymbol{\nu},\boldsymbol{\xi}}(\mathbf{x}) = 1 - \int_{\mathcal{X}} p_{\boldsymbol{\nu},\boldsymbol{\xi}}(\mathbf{x},\mathbf{z})\,\beta_{\boldsymbol{\nu},\boldsymbol{\xi}}(\mathbf{x},\mathbf{z})\,\mathrm{d}\mathbf{z}.$$

Then, for any $\mathbf{x}\in\mathcal{X}$ and $\mathbf{A}\in\mathcal{B}(\mathcal{X})$, transition kernels in two consecutive iterations n and $n+1$ satisfy

$$\begin{aligned}
&\left|K_{\boldsymbol{\nu}_{n+1},\boldsymbol{\xi}_{n+1}}(\mathbf{x},\mathbf{A}) - K_{\boldsymbol{\nu}_n,\boldsymbol{\xi}_n}(\mathbf{x},\mathbf{A})\right| \\
&\quad= \left|\int_{\mathbf{A}} \left[p_{\boldsymbol{\nu}_{n+1},\boldsymbol{\xi}_{n+1}}(\mathbf{x},\mathbf{y})\,\beta_{\boldsymbol{\nu}_{n+1},\boldsymbol{\xi}_{n+1}}(\mathbf{x},\mathbf{y}) - p_{\boldsymbol{\nu}_n,\boldsymbol{\xi}_n}(\mathbf{x},\mathbf{y})\,\beta_{\boldsymbol{\nu}_n,\boldsymbol{\xi}_n}(\mathbf{x},\mathbf{y})\right] d\mathbf{y}\right. \\
&\qquad\left. + \mathbb{1}_{\mathbf{x}\in\mathbf{A}}\left[\rho_{\boldsymbol{\nu}_{n+1},\boldsymbol{\xi}_{n+1}}(\mathbf{x}) - \rho_{\boldsymbol{\nu}_n,\boldsymbol{\xi}_n}(\mathbf{x})\right]\right| \\
&\quad\le \int_{\mathbf{A}} \left|p_{\boldsymbol{\nu}_{n+1},\boldsymbol{\xi}_{n+1}}(\mathbf{x},\mathbf{y})\,\beta_{\boldsymbol{\nu}_{n+1},\boldsymbol{\xi}_{n+1}}(\mathbf{x},\mathbf{y}) - p_{\boldsymbol{\nu}_n,\boldsymbol{\xi}_n}(\mathbf{x},\mathbf{y})\,\beta_{\boldsymbol{\nu}_n,\boldsymbol{\xi}_n}(\mathbf{x},\mathbf{y})\right| d\mathbf{y} \\
&\qquad + \chi_{\mathbf{x}\in\mathbf{A}} \int_{\mathcal{X}} \left|p_{\boldsymbol{\nu}_{n+1},\boldsymbol{\xi}_{n+1}}(\mathbf{x},\mathbf{y})\,\beta_{\boldsymbol{\nu}_{n+1},\boldsymbol{\xi}_{n+1}}(\mathbf{x},\mathbf{y}) - p_{\boldsymbol{\nu}_n,\boldsymbol{\xi}_n}(\mathbf{x},\mathbf{y})\,\beta_{\boldsymbol{\nu}_n,\boldsymbol{\xi}_n}(\mathbf{x},\mathbf{y})\right| d\mathbf{y},
\end{aligned} \tag{28}$$

where $\chi_{\mathbf{x}\in\mathbf{A}}$ is the indicator function. For $\mathbf{x}\neq\mathbf{y}$, we have

$$p_{\boldsymbol{\nu},\boldsymbol{\xi}}(\mathbf{x},\mathbf{y})\,\beta_{\boldsymbol{\nu},\boldsymbol{\xi}}(\mathbf{x},\mathbf{y}) = \min\left\{p_{\boldsymbol{\nu},\boldsymbol{\xi}}(\mathbf{x},\mathbf{y}), \frac{\pi(\mathbf{y})}{\pi(\mathbf{x})}\,p_{\boldsymbol{\nu},\boldsymbol{\xi}}(\mathbf{y},\mathbf{x})\right\}.$$

Given the identity $|\min\{a,b\} - \min\{c,d\}| \leq |a-c| + |b-d|$, we have

$$\begin{aligned}&\left|p_{\boldsymbol{\nu}_{n+1},\boldsymbol{\xi}_{n+1}}(\mathbf{x},\mathbf{y})\,\beta_{\boldsymbol{\nu}_{n+1},\boldsymbol{\xi}_{n+1}}(\mathbf{x},\mathbf{y}) - p_{\boldsymbol{\nu}_n,\boldsymbol{\xi}_n}(\mathbf{x},\mathbf{y})\,\beta_{\boldsymbol{\nu}_n,\boldsymbol{\xi}_n}(\mathbf{x},\mathbf{y})\right| \\ &\quad\leq \left|p_{\boldsymbol{\nu}_{n+1},\boldsymbol{\xi}_{n+1}}(\mathbf{x},\mathbf{y}) - p_{\boldsymbol{\nu}_n,\boldsymbol{\xi}_n}(\mathbf{x},\mathbf{y})\right|\left(1+\frac{\pi(\mathbf{y})}{\pi(\mathbf{x})}\right). \qquad (29)\end{aligned}$$

The compactness of parameter space $\mathcal{X}$ and continuity of the target $\pi(\cdot)$ imply that

$$\left(1+\frac{\pi(\mathbf{y})}{\pi(\mathbf{x})}\right) < C_1,\ \forall \mathbf{x},\mathbf{y}\in\mathcal{X}, \qquad (30)$$

for some constant $C_1 < \infty$.

Recall the property $\|\lambda_1(\cdot) - \lambda_2(\cdot)\|_{TV} = \frac{1}{2}\int_{\mathcal{X}} |\lambda_1(\mathbf{y}) - \lambda_2(\mathbf{y})| d\mathbf{y}$, Eqn. 28–30 imply that

$$\left|K_{\boldsymbol{\nu}_{n+1},\boldsymbol{\xi}_{n+1}}(\mathbf{x},\mathbf{A}) - K_{\boldsymbol{\nu}_n,\boldsymbol{\xi}_n}(\mathbf{x},\mathbf{A})\right| \leq C_2\ \left\|p_{\boldsymbol{\nu}_{n+1},\boldsymbol{\xi}_{n+1}}(\mathbf{x},\cdot) - p_{\boldsymbol{\nu}_n,\boldsymbol{\xi}_n}(\mathbf{x},\cdot)\right\|_{TV},$$

for some constant $C_2 < \infty$. Thus, we have

$$\left\|K_{\boldsymbol{\nu}_{n+1},\boldsymbol{\xi}_{n+1}}(\mathbf{x},\cdot) - K_{\boldsymbol{\nu}_n,\boldsymbol{\xi}_n}(\mathbf{x},\cdot)\right\|_{TV} \leq C_2\ \left\|p_{\boldsymbol{\nu}_{n+1},\boldsymbol{\xi}_{n+1}}(\mathbf{x},\cdot) - p_{\boldsymbol{\nu}_n,\boldsymbol{\xi}_n}(\mathbf{x},\cdot)\right\|_{TV}.$$

Since we can bound the total variation distance between transition kernels by the total variation distance between transition densities in the form of $p_{\boldsymbol{\nu},\boldsymbol{\xi}}(\mathbf{x},\mathbf{y}) = q_{\boldsymbol{\nu}}(\mathbf{x},\mathbf{y})\,\alpha_{\boldsymbol{\nu},\boldsymbol{\xi}}(\mathbf{x},\mathbf{y})$, diminishing adaptation of the overall transition kernel follows from diminishing adaptation of $q_{\boldsymbol{\nu}}(\mathbf{x},\mathbf{y})\,\alpha_{\boldsymbol{\nu},\boldsymbol{\xi}}(\mathbf{x},\mathbf{y})$.

Given the triangle inequality

$$\begin{aligned}\left\|p_{\boldsymbol{\nu}_{n+1},\boldsymbol{\xi}_{n+1}}(\mathbf{x},\cdot) - p_{\boldsymbol{\nu}_n,\boldsymbol{\xi}_n}(\mathbf{x},\cdot)\right\|_{TV} \leq{}& \left\|p_{\boldsymbol{\nu}_{n+1},\boldsymbol{\xi}_{n+1}}(\mathbf{x},\cdot) - p_{\boldsymbol{\nu}_n,\boldsymbol{\xi}_{n+1}}(\mathbf{x},\cdot)\right\|_{TV} \\ &+ \left\|p_{\boldsymbol{\nu}_n,\boldsymbol{\xi}_{n+1}}(\mathbf{x},\cdot) - p_{\boldsymbol{\nu}_n,\boldsymbol{\xi}_n}(\mathbf{x},\cdot)\right\|_{TV},\end{aligned}$$

we can treat adaptation indices $\boldsymbol{\nu}$ and $\boldsymbol{\xi}$ in separate steps.

Using a similar argument in the first part of this proof, we can show that for any fixed approximation adaptation index $\boldsymbol{\xi}$, diminishing adaptation of $p_{\boldsymbol{\nu},\boldsymbol{\xi}}(\mathbf{x},\mathbf{y})$ with respect to proposal adaptation in $\boldsymbol{\nu}$ follows directly from the above condition 1

In the rest of this proof, we establish the diminishing adaptation of $p_{\boldsymbol{\nu},\boldsymbol{\xi}}(\mathbf{x},\mathbf{y})$ with respect to approximation adaptation in $\boldsymbol{\xi}$. For any fixed proposal adaptation index $\boldsymbol{\nu}$, we have the following inequality

$$\begin{aligned}\left\|p_{\boldsymbol{\nu},\boldsymbol{\xi}_{n+1}}(\mathbf{x},\cdot) - p_{\boldsymbol{\nu},\boldsymbol{\xi}_n}(\mathbf{x},\cdot)\right\|_{TV} &= \frac{1}{2}\int_{\mathcal{X}} q_{\boldsymbol{\nu}}(\mathbf{x},\mathbf{y})\ \left|\alpha_{\boldsymbol{\nu},\boldsymbol{\xi}_{n+1}}(\mathbf{x},\mathbf{y}) - \alpha_{\boldsymbol{\nu},\boldsymbol{\xi}_n}(\mathbf{x},\mathbf{y})\right| d\mathbf{y} \\ &= \frac{1}{2}\int_{\mathcal{X}} \left|\min\left\{q_{\boldsymbol{\nu}}(\mathbf{x},\mathbf{y}), \frac{\pi^*_{\boldsymbol{\xi}_{n+1},\mathbf{x}}(\mathbf{y})}{\pi^*_{\boldsymbol{\xi}_{n+1},\mathbf{x}}(\mathbf{x})} q_{\boldsymbol{\nu}}(\mathbf{y},\mathbf{x})\right\}\right. \\ &\quad \left. - \min\left\{q_{\boldsymbol{\nu}}(\mathbf{x},\mathbf{y}), \frac{\pi^*_{\boldsymbol{\xi}_n,\mathbf{x}}(\mathbf{y})}{\pi^*_{\boldsymbol{\xi}_n,\mathbf{x}}(\mathbf{x})} q_{\boldsymbol{\nu}}(\mathbf{y},\mathbf{x})\right\}\right| d\mathbf{y} \\ &\leq \frac{1}{2}\int_{\mathcal{X}} \left|\frac{\pi^*_{\boldsymbol{\xi}_{n+1},\mathbf{x}}(\mathbf{y})}{\pi^*_{\boldsymbol{\xi}_{n+1},\mathbf{x}}(\mathbf{x})} - \frac{\pi^*_{\boldsymbol{\xi}_n,\mathbf{x}}(\mathbf{y})}{\pi^*_{\boldsymbol{\xi}_n,\mathbf{x}}(\mathbf{x})}\right| q_{\boldsymbol{\nu}}(\mathbf{y},\mathbf{x})\, d\mathbf{y}.\end{aligned}$$

Given the compactness of $\mathcal{X}\times\mathcal{E}$ and Lipschitz continuity of $\log \pi^*_{\boldsymbol{\xi},\mathbf{x}}(\mathbf{y})$, it follows that the RHS uniformly converges to zero as $\|\boldsymbol{\xi}_{n+1} - \boldsymbol{\xi}_n\| \to 0$. Thus, diminishing adaptation holds.

□

Remark 2

Conditions required in Theorems 2 and 3 are not restrictive for many practical applications.

Condition (1) of Theorem 2, that parameter space and the adaptation spaces are compact, is often a consequence of physical bounds on the parameters and bounded model outputs. In practical computation, one could argue that this assumption always holds as computers are finite dimensional, though one does not want to explore the full range of numerical representations if the algorithm is to be efficient! Conditions (3) and (4) of Theorem 2 and Condition (1) of Theorem 3 are conditions on on the proposal distribution, and depend on the choice of proposal and adaptation that is used. Thus, these conditions can be satisfied by making suitable choices.

Conditions 2 and 5 of Theorem 2, of Lipschitz continuity, is satisfied by the forward model and its reduced model in most inverse problems; Indeed, the more ill-posed is the inverse problem, the more well-posed is the forward model and the higher the order of continuity satisfied by the forward model. Finite-dimensional reduced models are Lipschitz continuous because stiffness matrices are not singular when the reduced model is well posed.

Then, we can use Theorems 2 and 3 to establish ergodicity of ADA using the approximation schemes in Section 3.

Corollary 1
Suppose that the forward map $F(\cdot)$ and its reduced model $F^*(\cdot)$ are continuous functions, then the ADA Algorithm 5 with either of the adaptive proposals in Section 2.5, and using any of Approximation 1 to 5 in Section 3 is ergodic for $\pi(\cdot)$.

Proof
We treat Approximations 3 and 5, since Approximations 1, 2, and 4 are non-adaptive special cases.

Without loss of generality, we may assume that the parameter space $\mathcal{X}$ is compact, as we can always define bounds on the input parameter to the computer model. Since $F(\cdot)$ and $F^*(\cdot)$ are continuous, it follows that the model reduction error $B(\cdot)$ or $B_{\mathbf{x}}(\cdot)$ is compact. Thus, the space of possible $\boldsymbol{\mu}_{B,n}$ and $\boldsymbol{\Sigma}_{B,n}$ is compact, that is, $\mathcal{E}$ is compact. Compactness of $\mathcal{V}$ follows from the form of proposal adaptation in Algorithms 3 and 4, and existing results for such proposals, such as Corollary 6 of [14]. This establishes Condition (1) of Theorem 2.

Since $F(\cdot)$ and $F^*(\cdot)$ are continuous, Condition (5) of Theorem 2 follows from the form of Equations 21 and 27 when $\boldsymbol{\Sigma}_{B,n} + \boldsymbol{\Sigma}_{\mathbf{e}}$ is nonsingular. Since $\boldsymbol{\Sigma}_{\mathbf{e}}$ is positive definite and $\boldsymbol{\Sigma}_{B,n}$ is positive semi-definite, it follows that $\boldsymbol{\Sigma}_{B,n} + \boldsymbol{\Sigma}_{\mathbf{e}}$ is actually positive definite.

The AEM updating rules Eqns 19–20 and Eqn. 26 satisfy the diminishing adaptation condition (Condition (2) of Theorem 3), because the empirical estimates change $O(1/n)$, in probability, at the n-th iteration. Similarly, the proposal diminishing adaptation condition (Condition (1) of Theorem 3) follows from the form of adaptation in Algorithms 3 and 4.

The conditions in Theorems 2 and 3 hold, and so the result follows. □

5. APPLICATIONS IN FITTING GEOTHERMAL RESERVOIR MODELS

In this section we apply ADA to two calibration problems for geothermal reservoir models. First, a one dimensional radial symmetry model of the feedzone of a geothermal reservoir with synthetic data is presented. This example is small enough that extensive statistics can be computed to study relative efficiencies of algorithms. Then ADA is applied to sample a 3D model with measured data. We begin with a description of the governing equations of the geothermal reservoir and its numerical simulator.

5.1. Data simulation

Consider a two phase geothermal reservoir (water and vapour) governed by the general mass balance and energy balance equations[43, 44]

$$\frac{d}{dt}\int_{\Omega} M_\alpha \, dV = \int_{\partial\Omega} Q_\alpha \cdot \mathbf{n} \, d\Gamma + \int_{\Omega} q_\alpha \, dV, \quad \alpha \in \{\mathrm{m}, \mathrm{e}\}, \tag{31}$$

where Ω is the control volume and $\partial\Omega$ is its boundary. The accumulation term q_α represents the mass (q_{m}) and heat (q_{e}) sources or sinks in Ω, and Q_α denotes the mass (Q_{m}) or energy (Q_{e}) flux through $\partial\Omega$. The mass and energy within Ω are represented by M_{m} and M_{e}.

A complex set of nonlinear partial differential equations including non-isothermal, multiphase Darcy's law are used to model M_α and Q_α.

Using a two-phase flow as the example, the mass and energy per unit volume can be modelled by

$$M_{\mathrm{m}} = \phi\left[\rho_{\mathrm{l}}(1-S_{\mathrm{v}}) + \rho_{\mathrm{v}}S_{\mathrm{v}}\right] \quad , \tag{32}$$

$$M_{\mathrm{e}} = (1-\phi)\,\rho_{\mathrm{r}}c_{\mathrm{r}}T + \phi\left[\rho_{\mathrm{l}}u_{\mathrm{l}}(1-S_{\mathrm{v}}) + \rho_{\mathrm{v}}u_{\mathrm{v}}S_{\mathrm{v}}\right], \tag{33}$$

where ϕ is the porosity of rock, T is the temperature, and S_{v} represent the vapour saturation.

We use the subscripts l, v and r represent the liquid phase, the vapour phase, and the rock, respectively.

There are physical constants involved in the above equations: $c_{(\cdot)}$ is the specific heat, $\rho_{(\cdot)}$ denotes the density, and $u_{(\cdot)}$ is the specific internal energy. The mass and energy fluxes are modelled by

$$Q_{\mathrm{m}} = \sum_{\beta=\mathrm{l,v}} \frac{k k_{\mathrm{r}\beta}}{\nu_\beta}(\nabla p - \rho_\beta \vec{g}), \tag{34}$$

$$Q_{\mathrm{e}} = \sum_{\beta=\mathrm{l,v}} \frac{k k_{\mathrm{r}\beta}}{\nu_\beta}(\nabla p - \rho_\beta \vec{g})h_\beta - K\nabla T, \tag{35}$$

where k is a diagonal second order permeability tensor in 3-dimensions. Physical constants $h_{(\cdot)}$ and K denote specific enthalpy and the thermal conductivity in a saturated medium, respectively. In the above equations, relative permeabilities k_{rl} and k_{rv}—which are empirically derived functions—are introduced to account for the interference between liquid and vapour phases. Here, we use the van Genuchten-Mualem model [45]:

$$(k_{\mathrm{rl}}, k_{\mathrm{rv}}) = f_{\mathrm{vGM}}(S_{\mathrm{v}}; m, S_{\mathrm{rl}}, S_{\mathrm{ls}}),$$

which is a function of the saturation S_{v}. The van Genuchten-Mualem model also depends on explicitly provided parameters m, S_{rl}, and S_{ls} to characterize the behaviour of relative permeabilities.

In the system of equations 31–35, the system states are the spatially distributed quantities (p, T, S_{v}). We can make partial observation either on these quantities directly or some measurements that are typically nonlinear function of these quantities.

The system of equations contains unknown parameters including porosity, permeability, initial and boundary conditions, and parameters of the relative permeability model, are fit to data using sample-based inference.

We use the package TOUGH2 [42] to numerically solve the system of equations 31–35, for a given set of parameters, using the finite volume method with first order accuracy in space.

TOUGH2 carries time integration using the implicit first order scheme, in which the implicit time integration is solved by the Newton–Krylov method. In TOUGH2, time stepping is automatically adjusted according to the number of Newton iterations in each time step.

This way, numerical models based on grids with different resolutions can be naturally used to construct the forward model and its reduced model. We do not control the time step in the discretization.

5.2. Well discharge test analysis

5.2.1. Parameter and data Well discharge analysis is usually used to interpret the near-well properties of the reservoir from pressure and enthalpy data measured during a short period of field production. Based on the typical assumption that all flows into the well come through a single layer feedzone, we build an one-dimensional radially-symmetric forward model $F(\cdot)$ with 640 blocks as shown in Fig. 1 (a). A high resolution grid is used immediately outside the well, with cell thickness

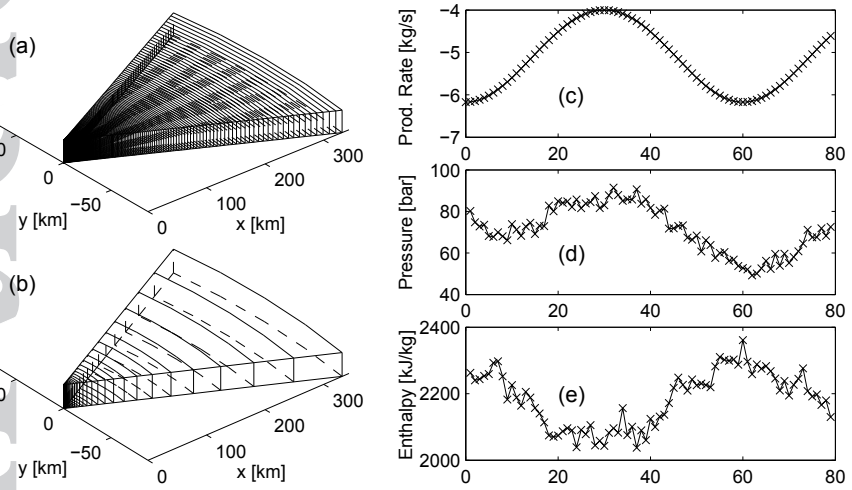

Figure 1. Finite volume grids used for well discharge test analysis and data sets used for well discharge test. (a): the forward model (640 blocks), (b): the reduced model (40 blocks), (c): the production rate (kg/second), (d): the pressure (bar), and (e): the flowing enthalpy (kJ/kg).

increasing exponentially outside this region. The reduced model $F^*(\cdot)$ is built by coarsening the grid of the forward model to a coarse grid with 40 blocks, see Figure 1 (b). Based on 1,000 simulations with different set of parameters on a DELL T3400 workstation, we estimate the CPU time for evaluating the forward model is 2.60 seconds. The computing time of the reduced model is 0.15 seconds, which is 5.8% of that for the forward model.

Table I. Prior constraints for parameters of the well test model, and parameter values for generating synthetic data.

	ϕ[-]	$\log_{10}(k)[\mathrm{m}^2]$	p_0 [bar]	$S_{\mathrm{v}0}$ [-]	m [-]	S_{rl} [-]	S_{ls} [-]
Lower bound	0	-Inf	0.5	0	0.7	0	100
Upper bound	0.3	Inf	1	0.5	1	0.3	150
True value	0.12	-14.82	120	0.1	0.65	0.25	0.91

The parameters of interest are the porosity, (horizontal) permeability, the parameters in the van Genuchten-Mualem relative permeability model, as well as the initial vapour saturation ($S_{\mathrm{v}0}$) and initial pressure (p_0) that are used to represent the initial thermodynamic state of the two-phase system. These make up the seven unknowns required for data simulation:

$$\mathbf{x} = \{\phi, \log_{10}(k), p_0, S_{\mathrm{v}0}, m, S_{\mathrm{rl}}, S_{\mathrm{ls}}\}\,.$$

Note that the permeability k is represented on a base 10 logarithmic scale. These parameters are assumed to be independent and follow non-informative prior distributions with bounds given in Table I. Table I also gives "true" values of model parameters used to simulate the synthetic data. The model is simulated over 80 days with production rates varying smoothly from about 4 kg/second to about 6 kg/second (see Figure 1 (c)).

For well test experiments, the observations include the pressure and flowing enthalpy measured at N discrete time points $t_1, t_2, \ldots, t_N$ up to day 30, where the flowing enthalpy h_f is a nonlinear function of pressure and vapour saturation, i.e., $h_f(p, S_{\mathrm{v}})$. This way, we can denote the observed data as

$$\tilde{\mathbf{d}} = \left[\tilde{p}(t_1), \tilde{p}(t_2), \ldots, \tilde{p}(t_N), \tilde{h}_f(t_1), \tilde{h}_f(t_2), \ldots, \tilde{h}_f(t_N)\right]^\top.$$

For each realization of the model parameter, we can simulate the forward model $F(\mathbf{x})$ to generate observable model outputs corresponding to the observed data.

We assume the measurement noise follows i.i.d. zero mean Gaussian distribution with standard deviations $\sigma_p = 3$ bar for pressure and $\sigma_h = 30$ kJ/kg for the flowing enthalpy. The noise corrupted pressure and flowing enthalpy data are plotted in Figure 1 (d) and (e), respectively. This yields the posterior distribution

$$\pi_{\text{post}}(\mathbf{x} \,|\, \tilde{\mathbf{d}}) \propto \exp\left[-\frac{1}{2}\left(F(\mathbf{x}) - \tilde{\mathbf{d}}\right)^{\top} \mathbf{\Sigma}_{\mathbf{e}}^{-1}\left(F(\mathbf{x}) - \tilde{\mathbf{d}}\right)\right] \chi(\mathbf{x})$$

where $\chi(\mathbf{x})$ is the prior distribution which we take to be the indicator function that implements the bounds in Table I and

$$\mathbf{\Sigma}_{\mathbf{e}} = \begin{bmatrix} \sigma_{\text{h}}^2 \mathbf{I}_n & 0 \\ 0 & \sigma_{\text{p}}^2 \mathbf{I}_n \end{bmatrix}$$

is the covariance of measurement noise.

Remark 3
To check the assumptions of Theorems 2 and 3, we note that the compactness of the parameter space is satisfied as we impose bounds on parameters.

Lipschitz continuity of the forward model and reduced model, and hence the continuity of the posterior and its approximations based on reduced model, is satisfied as the system of equations 31–35 are continuous.

Compactness of the approximation adaptation space $\mathcal{E}$ directly follows from the compactness of the parameter space and the continuity of the forward model and reduced model.

5.2.2. MCMC sampling To benchmark various approximate posterior distributions, we first run GCAM with one group of all 7 parameters and targeting the exact posterior distribution. Based on several short runs (not reported here), we find that an acceptance rate of 13% gives optimal efficiency for this problem. This configuration of GCAM is run for 10^5 iterations (after discarding 5×10^4 burn-in steps) giving an IACT of the log-likelihood function estimated as 84.4.

Then we then run ADA using Approximation 1, Approximation 2 (with the AEM calculated *a priori*), Approximation 3 (with the AEM calculated adaptively over the posterior), and Approximation 5 with the AEM calculated adaptively over the posterior. All cases used GCAM for the proposal with the target acceptance rate of 13%. The acceptance rate in step 2 of ADA, $\bar{\beta}$, and the IACT of the likelihood function are shown in Table II.

We note that when Approximation 1 is used in ADA, the method is the equivalent method in [16], except that we use an adaptive proposal here.

Table II. Performance summary of various approximations for the well test model. The abbreviations used are: $\bar{\beta}$ – the second step acceptance rate, and IACT – IACT of the log-likelihood function.

	Approx. 1	Approx. 2 (prior)	Approx. 3 (posterior)	Approx. 5	Standard MH
$\bar{\beta}$	0.12	0.31	0.77	0.93	1
IACT	-	-	208	153	169

Approximation 1 (approximation uses reduced model directly) only produces $\bar{\beta} = 12\%$. (This agrees closely with the equivalent method in [16].) Approximation 2 with AEM built over the prior, increases the acceptance rate in step 2 of ADA to $\bar{\beta} = 31\%$. However, both Approximation 1 and the AEM constructed over the prior cannot produce a well mixed Markov chain, even after 2×10^5 iterations, so the IACT for the log-likelihood function could not be estimated, and is not reported for these cases.

Approximation 3 with the AEM calculated adaptively over the posterior produces significantly better mixing, with an estimated $\bar{\beta} = 77\%$, and the IACT of the log-likelihood function is 208. Approximation 5, with AEM calculated adaptively over the posterior and with state-dependent reduced model (22), improves the performance further, achieving $\bar{\beta} = 93\%$, and the IACT of the

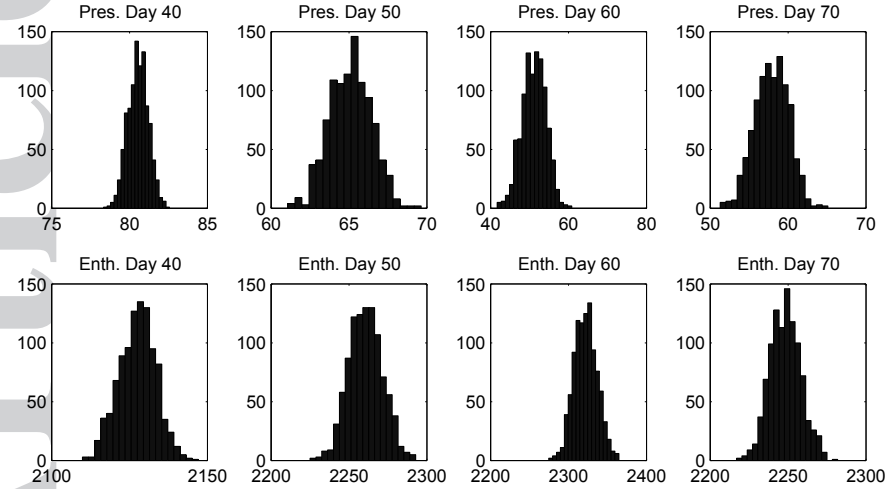

Figure 2. Histograms of the predictive density, at day 40, 50, 60 and 70. Top row: Pressure, bottom row: flowing enthalpy.

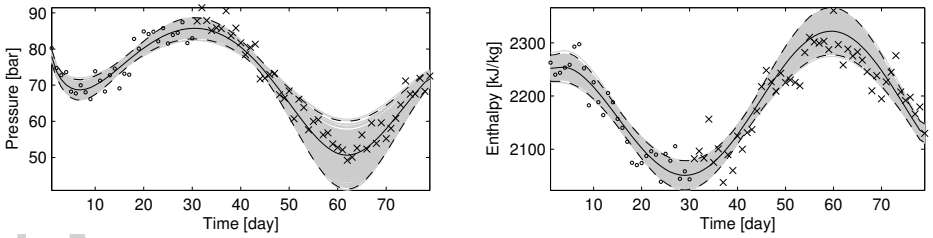

Figure 3. Predictions for pressure and flowing enthalpy. Left column: pressure, Right column: flowing enthalpy. The circles and the crosses are the training data and validation data, respectively; the solid line and dashed lines are the mean prediction and 95% credible interval, respectively; and the shaded lines represent the predictions made by samples.

log-likelihood function is 153. For Approximation 3 and 5, we ran both chains for 2×10^6 iterations to calculate further summary statistics, with the first 5×10^5 steps discarded as burn-in.

By using formula (9), we can estimate that the factor by which computational efficiency is improved for ADA with Approximation 3 (posterior) and 5 is about 4.3 and 5.9, respectively. In contrast, the use of Approximation 2 (prior) only improves computational efficiency marginally, while the use of the naïve Approximation 1 appears to actually reduce computational efficiency. We also notice that the IACTs of the log-likelihood function suggest that the ADA with Approximation 5 is more statistically efficient than the standard MH, which cannot be the case as discussed in Section 2.3. This effect could be caused by finite-sampling error in the IACT estimate. However, this result suggests that the decrease of statistical efficiency may be negligible in this particular case.

The histograms of the model predictions computed on several different time points are given in Figure 2, with pressure in the top row and enthalpy in the bottom row. We can observe that the predictions at these observation times follow uni-modal distributions. The model predictions and the 95% credible intervals over an 80-day period are shown in Figure 3. For both predictions, the means follow the observed data reasonably well.

The histograms of the marginal distributions (first two rows of Figure 4) of the parameter $\mathbf{x}$ show skewness in porosity and two of the parameters of the van Genuchten-Mualem relative permeability model (m and S_{rl}). The scatter plots between parameters show strong negative correlations between the permeability (on base 10 logarithmic scale) and the initial pressure; see the left plot of last row of Figure 4. There is also a strong negative correlation between the initial saturation and one of the hyperparameters of the van Genuchten-Mualem (S_{ls}); see the right plot of last row of Figure 4.

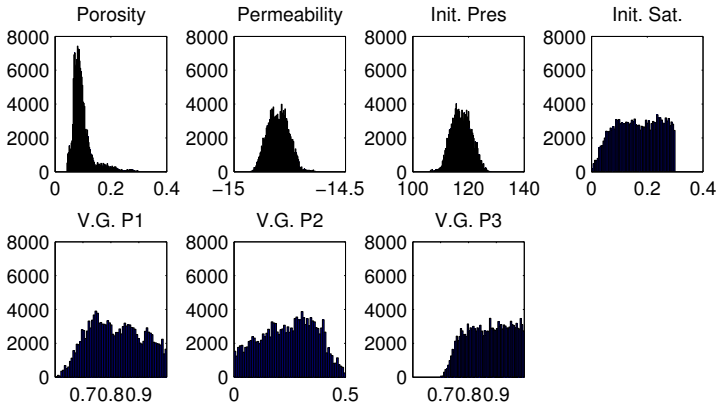

Figure 4. Histograms of the marginal distributions and scatter plots between parameters, for the synthetic data set.

5.3. Natural state modelling

5.3.1. Parameter and data We now present an application of ADA to a 3D geothermal reservoir model with measured field data. We aim to infer the permeability structure and mass input at the bottom of the reservoir from temperature data measured from wells. We also wish to predict the size and shape of the hot plume of the reservoir, which is only sparsely measured in wells. The forward model covers a volume of 12.0 km by 14.4 km extending down to 3050 meters below sea level. Relatively large blocks were used near the outside of the model and then were progressively refined near the wells to achieve a well-by-well allocation to the blocks. The 3D structure of the forward model $F(\cdot)$ has $26,005$ blocks, and is shown in Figure 5 (a), where the blue lines in the middle of the grid show wells drilled into the reservoir. To speed up the computation a reduced model $F^*(\cdot)$ based on a coarse grid with $3,335$ blocks is constructed by combining adjacent blocks in the x, y and z directions of the forward model; see Figure 5 (b). Each simulation of the forward model takes about 30 to 50 minutes CPU time on a DELL T3400 workstation, and the computing time for the reduced model is about 1 to 1.5 minutes (roughly 3% of the forward model). The computing time for these models is sensitive to the input parameters.

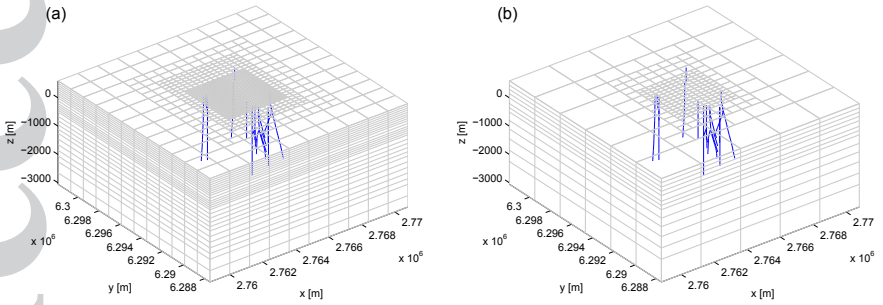

Figure 5. The fine grid (left) and the coarse grid (right) used for natural state modelling.

Let s denote the spatial coordinate and $\Omega \subset \mathbb{R}^3$ denote the domain of the reservoir. In the natural state modelling, we are interested in estimating spatially distributed and heterogeneous permeabilities $k(s)$ for $s \in \Omega$ and the mass input $q_{\mathrm{m}}(r)$ from the bottom boundary $\partial\Omega_b \subset \mathbb{R}^2$, where $r \in \partial\Omega_b$.

We model the permeability tensor by scalar-valued permeability functions in the form of

$$k(s) = \begin{bmatrix} k_1(s) & & \\ & k_2(s) & \\ & & k_3(s) \end{bmatrix},$$

and discretize permeability functions $k_1(s)$, $k_2(s)$, and $k_3(s)$ on the coarse grid using piecewise linear representation. Given the centres of the finite volume cells in the coarse model, $s_1, s_2, \ldots, s_{3335}$, we have the discretized permeabilities

$$\mathbf{k} = \left[k_1(s_1), k_1(s_2), \ldots, k_1(s_{3335}), k_2(s_1), k_2(s_2), \ldots, k_2(s_{3335}), k_3(s_1), k_3(s_2), \ldots, k_3(s_{3335})\right]^\top,$$

which is about $10,005$ dimensional. The spatial correlation of permeabilities in the base 10 logarithmic scale are modelled by a Gaussian Markov random field prior [22]. We further impose that the permeability are bounded between 10^{-2} and 10^3 millidarcy. This leads to the prior distribution of $\mathbf{k}$ in the form of

$$\pi_{\text{prior}}(\mathbf{k}) \propto \exp\left(-\frac{1}{2}\log_{10}(\mathbf{k})^\top \mathbf{Q}\, \log_{10}(\mathbf{k})\right) \chi(\mathbf{k}),$$

where $\mathbf{Q}$ is a symmetric positive definite sparse matrix constructed from the Gaussian Markov random field. The mass input from the bottom boundary is modeled by the linear combination of radial basis functions with the squared-exponential kernel function

$$q_{\text{m}}(r) = \sum_{i=1}^{41} w_i \exp\left[-\lambda \left\|r - r_i\right\|^2\right],$$

centred at pre-specified control points $\mathbf{r}_i, i = 1, \ldots, 41$. This way, the unknowns in this parametrization are the 41 dimensional weighting variable

$$\mathbf{w} = (w_1, \ldots, w_{41})^\top,$$

associated with the control points. Since this weighting vector $\mathbf{w}$ controls the distribution of the mass input, the constraints $\mathbf{w} > 0$ and $\sum \mathbf{w} = constant$ are imposed to ensure that the mass input is positive and has a fixed total amount. Overall, we have unknown parameters $\mathbf{x} = [\mathbf{k}, \mathbf{w}]^\top$ with prior

$$\begin{aligned} \pi_{\text{prior}}(\mathbf{x}) &= \pi_{\text{prior}}(\mathbf{k})\, \pi_{\text{prior}}(\mathbf{w}) \propto \quad \exp\left(-\frac{1}{2}\log_{10}(\mathbf{k})^\top \mathbf{Q}\, \log_{10}(\mathbf{k})\right) \chi(\mathbf{k})\, \chi(\mathbf{w}), \qquad (36) \\ &\text{subject to} \qquad \sum \mathbf{w} = \text{constant and } \mathbf{w} > 0, \end{aligned}$$

where $\chi(\mathbf{k})$ and $\chi(\mathbf{w})$ impose bounds on parameters. We refer interested readers to [19] for a further details of the prior modelling of this problem.

Steady state temperature measured at discrete locations along the well bores (as shown by blues in Figure 5) are used for estimating permeability and mass input at the depth. Given measurement locations $s_1, s_2, \ldots, s_N$, we have the observed data

$$\tilde{\mathbf{d}} = \left[\tilde{T}(s_1), \tilde{T}(s_2), \ldots, \tilde{T}(s_N)\right]^\top,$$

as shown by the crosses in Figure 6. Empirical estimation of the noise vector [19] suggests an i.i.d. Gaussian distribution with standard deviation $\sigma_{\text{T}} = 7.5°\text{C}$ to be used in the likelihood function. Simulating the forward model F with realizations of the parameter $\mathbf{x}$ produces observable model outputs that are temperatures at measurement locations. This yields the posterior distribution

$$\pi_{\text{post}}(\mathbf{x}|\tilde{\mathbf{d}}) \propto \exp\left[-\frac{1}{2\sigma_{\text{T}}{}^2}\left\|F(\mathbf{x}) - \tilde{\mathbf{d}}\right\|_2^2\right] \pi_{\text{prior}}(\mathbf{x}) \qquad (37)$$

where $\pi_{\text{prior}}(\mathbf{x})$ is defined in Eqn. 36. Similar to the well test case, compactness assumptions and continuity assumptions (of the posterior and its approximations) in Theorems 2 and 3 are satisfied.

5.3.2. MCMC Sampling In the first step of ADA, the new permeability state is proposed using GCAM, and the weights controlling the mass input are proposed separately by an adaptive reversible jump move to meet the specific constraints. These proposals produce a first step acceptance rate of $\bar{\alpha} = 0.1$ in ADA; see [19] for details. Since the forward model is computationally very demanding, we first simulate the chain with the reduced model only, for about 200 sweeps of updates to get through the initial burn-in period. Then, we start ADA with Approximation 5 to sample the exact posterior distribution.

We are able to sample the posterior distribution for about $11,200$ iterations in 40 days, and ADA achieves about $\bar{\alpha} = 74\%$ acceptance rate in the second accept-reject step. After discarding the first $2,800$ iterations as burn-in steps, the estimated IACT of the log-likelihood function is about 5.6. This is only a rough estimate because the chain has not been running long enough, however all the samples are consistent with measured data, as shown by the model outputs of the realizations (Figure 6). Using the computational cost of the fine model and the IACT estimated from MCMC simulations over the posterior distribution defined using the coarse model, we can estimate that the adaptive delayed acceptance achieves a speed-up factor of 7.7.

The mean and standard deviation of the temperature profiles are estimated as posterior sample averages. We compare these estimates with the measured data in Figure 6. The solid black lines are the estimated mean temperatures, dashed black lines denote the 95% percent credible interval, and measured data are shown as red crosses. The green and gray lines represent outputs of the forward model and various reduced model realizations, respectively. Figure 6 shows that the forward model and the reduced model produce significantly different temperature profiles, and the forward model is hotter than the reduced model in average. This suggests that the model outputs are defined on some low dimensional manifold of the data space, and the forward model and reduced model produce outputs that are located on substantially different manifolds. The mean model reduction error at each of the measurement positions span a range of $[-33.64, 54.98]$, and hence the noise level of these model reduction errors are more significant than the zero mean normal distribution with standard deviation $\sigma_e = 7.5\,^{\circ}\mathrm{C}$. Therefore, the stochastic modeling of the model reduction error in Section 3 is essential for efficient and accurate inference when using the reduced model.

The samples of the permeability distributions (we only report the permeabilities in the vertical direction for brevity, see Figure 7) are highly variable, even though each of these samples produces reasonable match to the measured data. In comparison the predicted temperature distribution (Figure 8) has very small uncertainty intervals. It reveals that even though the parameters are not well identified, the predicted temperature field is relatively well determined.

6. DISCUSSION

We considered sample-based uncertainty quantification for inverse problems within the Bayesian formulation. Our primary contribution has been to advance computational methods that utilize a cheap approximation to the forward map, or reduced model, to improve computational efficiency of sample-based inference, that is the computational rate of reducing Monte Carlo errors in sample-based estimates. We presented state-dependent and stochastic corrections to reduced models, and gave a novel algorithm for adapting to the AEM, along with regularity conditions that guarantee ergodicity for the correct posterior distribution. We validated this algorithm and the improved approximations in two examples of parameter estimation for multi-phase flow in geothermal reservoirs, including a large-scale natural-state system using field data.

A straightforward method is to use a reduced model, such as a coarse PDE solve, in place of the correct forward map, and employ the DA algorithm to ensure the correct target distribution. We presented this as Approximation 1 and computed results using this approximation. In our test cases, we found that the acceptance rate, $\bar{\beta}$ in the second accept-reject step in DA (or ADA) was never more than 0.2. This indicates that the reduction in statistical efficiency, that necessarily occurs when using an approximation, may effectively cancel the reduction in compute time per iteration, resulting in negligible improvement in computational efficiency for the multi-phase flow problems we considered.

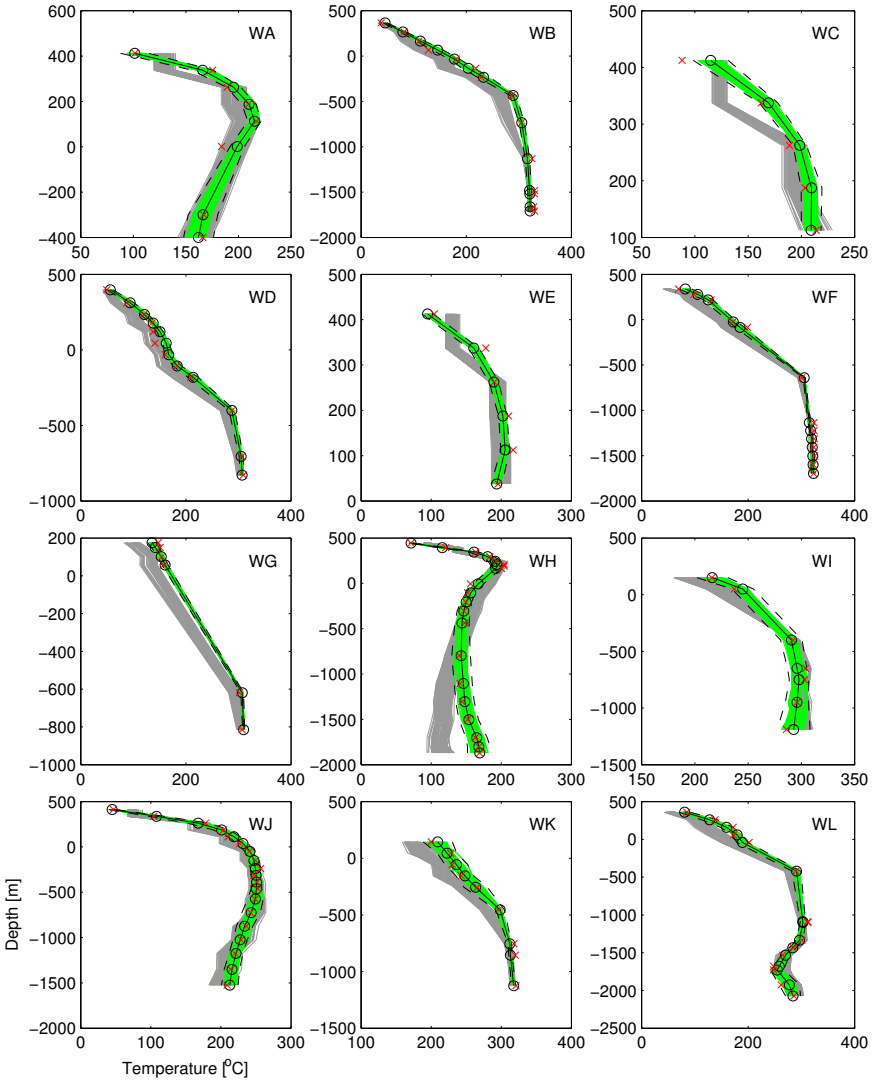

Figure 6. Comparison of estimated temperatures and measured data. The solid black lines are the estimated mean temperatures, dashed black lines are the 95% percent credible interval, and measured data are shown as red crosses. The green and gray lines represent outputs of the forward model and the reduced model various realizations, respectively.

We improved the approximation to the target distribution, at no significant increase in computational cost, by introducing a state-dependent correction to the coarse PDE solve, and also a stochastic model for the model reduction error. The former was made possible by using the DA algorithm [11], while posterior estimation of the AEM was made possible by using adaptive MCMC techniques [14]. We validated the sequence of approximations in a stylized seven-dimensional inverse problem in multi-phase flow, that allowed extensive evaluation of diagnostics.

Of particular importance was adaptive construction of the AEM over the posterior distribution; construction of the AEM over the prior, as in the original formulation of [2], improves statistical efficiency noticeably by increasing $\bar{\beta}$ from 0.12 to 0.31, but the construction over the posterior made a much greater improvement to $\bar{\beta} = 0.77$, while also not needing the expensive pre-calculation required for prior construction of the AEM. Including both the state-dependent correction and posterior AEM further gave further improvement with to $\bar{\beta} = 0.93$, indicating that statistical efficiency is only slightly reduced in comparison to standard MH, so reduction in computational cost per iteration is transferred directly to improvement in computational efficiency.

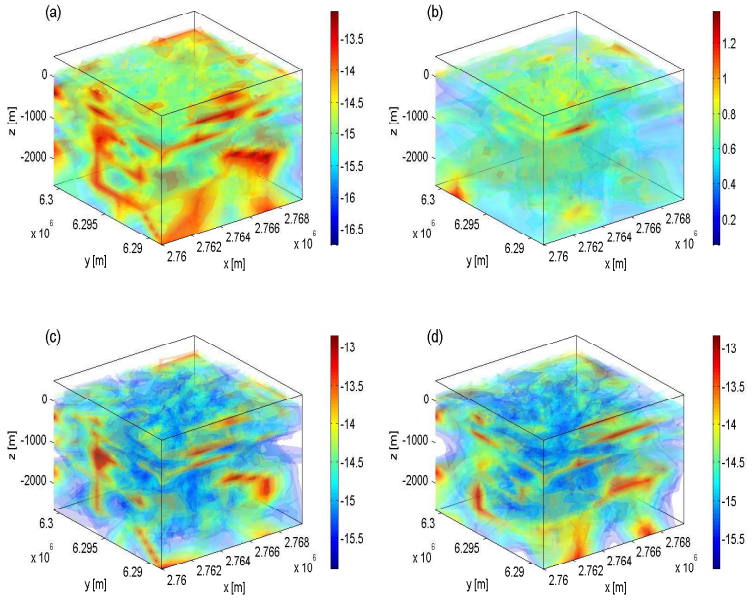

Figure 7. Permeability distribution in the vertical direction on base 10 logarithmic scale. (a): the sample mean, (b): the sample standard deviation, (c): one realization from the Markov chain, and (d): another realization from the Markov chain.

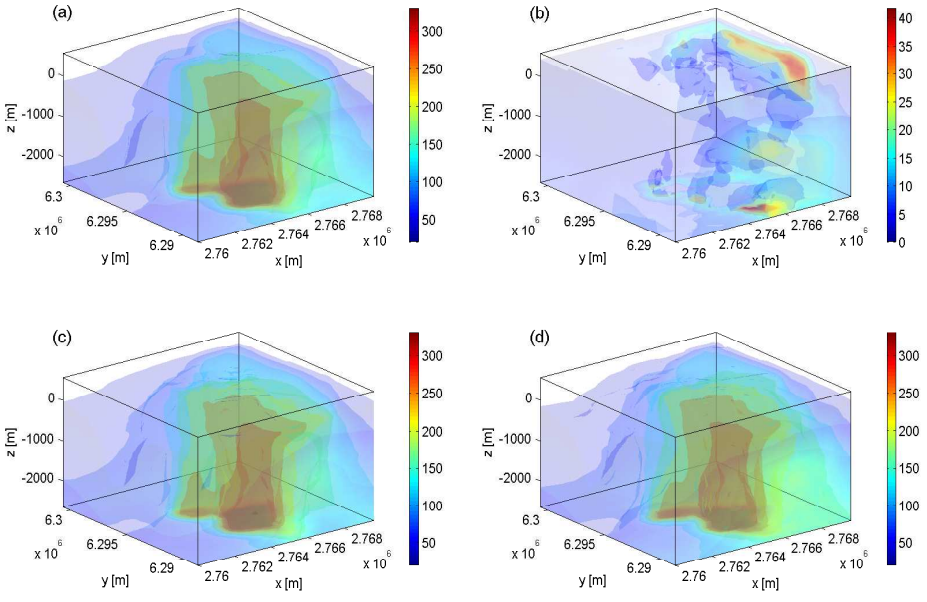

Figure 8. Distributions of model temperatures, unit is $^{\circ}C$. (a): the sample mean, (b): the sample standard deviations, (c): one realization from the Markov chain, and (d): another realization from the Markov chain.

Finally, we reported results from a large-scale inverse problem in calibration and prediction using a comprehensive numerical model of an actual geothermal field, using measured field data. For

that example, the novel ADA algorithm with Approximation 5 gave a increase in computational efficiency by a factor of about 7.7. Given the very large-scale nature of the computation required in this example, the computational savings by using the novel ADA algorithm are significant. We expect that the computational approach we have developed here will also be of significant value for uncertainty quantification in other large-scale inverse problems in science and engineering.

The approximations used in this paper are based on grid coarsening, due to the blackbox nature of our forward model. The power of our adaptive delayed acceptance framework is not limited by this form of reduced models. The regularity conditions on the approximate posteriors established in Theorems 2 and 3 offer further insights in designing new goal-oriented reduced models for solving inverse problems. For example, methods such as projection-based model reduction[37, 38, 39, 40, 41], which often replies on expensive offline training phase by repeatedly evaluating froward models over prior samples, can also be adaptively constructed using the ADA framework. This can potentially lead to significant speed-up and accuracy improvement for quantifying uncertainties of inverse problems.

REFERENCES

1. Hadamard J. Sur les problèmes aux dérivées partielles et leur signification physique. [On the problems about partial derivatives and their physical significance].. *Princeton University Bulletin* 1902; 13: 49–52.
2. Kaipio J, Somersalo E. *Statistical and Computational Inverse Problems*. Springer-Verlag . 2004.
3. Oliver DS, Cunha LB, Reynolds AC. Markov chain Monte Carlo methods for conditioning a permeability field to pressure data. *Mathematical Geology* 1997; 29(1): 61–91. http://dx.doi.org/10.1007/BF02769620 doi: 10.1007/BF02769620
4. Higdon D, Lee H, B. Z. A Bayesian approach to characterizing uncertainty in inverse problems using coarse and fine-scale information. *IEEE Transactions on Signal Processing* 2002; 50(2): 389–399.
5. Higdon D, Lee H, Holloman C. Markov chain Monte Carlo-based approaches for inference in computationally intensive inverse problems. In: Bernardo JM, Bayarri MJ, Berger JO, et al. , eds. *Bayesian Statistics 7*Oxford University Press; 2003: 181–197.
6. Mondal A, Efendiev Y, Mallick B, Datta-Gupta A. Bayesian uncertainty quantification for flows in heterogeneous porous media using reversible jump Markov chain Monte Carlo methods. *Advances in Water Resources* 2010; 33(3): 241–256. http://dx.doi.org/10.1016/j.advwatres.2009.10.010 doi: 10.1016/j.advwatres.2009.10.010
7. Cornford D, Csató L, Evans DJ, Opper M. Bayesian Analysis of the Scatterometer Wind Retrieval Inverse Problem: Some New Approaches. *Journal of the Royal Statistical Society. Series B* 2004; 66(3): 609-652.
8. Haario H, Laine M, Lehtinen M, Saksman E, Tamminen J. Markov Chain Monte Carlo Methods for High Dimensional Inversion in Remote Sensing. *Journal of the Royal Statistical Society. Series B* 2004; 66(3): 591-607.
9. Andersen KE, Brooks SP, Hansen MB. Bayesian Inversion of Geoelectrical Resistivity Data. *Journal of the Royal Statistical Society: Series B* 2004; 65: 619-644.
10. Watzenig D, Fox C. A review of statistical modelling and inference for electrical capacitance tomography. *Measurement Science and Technology* 2009; 20(5). http://dx.doi.org/10.1088/0957-0233/20/5/052002 doi: 10.1088/0957-0233/20/5/052002
11. Christen JA, Fox C. Markov chain Monte Carlo using an approximation. *Journal of Computational and Graphical Statistics* 2005; 14(4): 795-810.
12. Kennedy MC, O'Hagan A. Bayesian calibration of computer models (with discussion). *Journal of the Royal Statistical Society: Series B* 2001; 63: 425-464.
13. Kaipio JP, Somersalo E. Statistical inverse problems: Discretization, model reduction and inverse crimes. *Journal of Computational and Applied Mathematics* 2007; 198(2): 493-504.
14. Roberts GO, Rosenthal JS. Coupling and Ergodicity of Adaptive Markov chain Monte Carlo Algorithms. *Journal of Applied Probability* 2007; 44: 458-475.
15. Liu JS. *Monte Carlo Strategies in Scientific Computing*. Springer-Verlag . 2001.
16. Efendiev Y, Hou T, Luo W. Preconditioning Markov Chain Monte Carlo Simulations Using Coarse-Scale Models. *SIAM Journal on Scientific Computing* 2006; 28(2): 776-803.
17. Quiroz M, Tran MN, Villani M, Kohn R. Speeding up MCMC by delayed acceptance and data subsampling. *Journal of Computational and Graphical Statistics* 2018; 27(1): 12–22.
18. Haario H, Saksman E, Tamminen J. An adaptive Metropolis algorithm. *Bernoulli* 2001; 7: 223-242.
19. Cui T, Fox C, O'Sullivan MJ. Bayesian Calibration of a Large Scale Geothermal Reservoir Model by A New Adaptive Delayed Acceptance Metropolis Hastings Algorithm. *Water Resource Research* 2011; 47. 26 pp.http://dx.doi.org/10.1029/2010WR010352 doi: 10.1029/2010WR010352
20. Banerjee S, Gelfand AE, Carlin BP. *Hierarchical Modeling and Analysis for Spatial Data.* Chapman and Hall/CRC . 2003.
21. Hurn MA, Husby O, Rue H. Advances in Bayesian Image Analysis. In: Green PJ, Hjort NL, Richardson S., eds. *Highly Structured Stochastic Systems*Oxford University Press. 2003 (pp. 301-322).
22. Rue H, Held L. *Gaussian Markov Random Fields: Theory and Applications*. Chapman & Hall . 2005.
23. Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E. Equation of state calculations by fast computing machines. *Journal of chemical physics* 1953; 21: 1087-1092.
24. Hastings W. Monte Carlo sampling using Markov chains and their applications. *Biometrika* 1970; 57: 97-109.

25. Robert CP, Casella G. *Monte Carlo Statistical Methods*. Springer-Verlag . 1999.
26. Sokal A. Monte Carlo methods in statistical mechanics: foundations and new algorithms. 1989. In Course de Troisième Cycle de la Physique en Suisse Romande.
27. Geyer CJ. Practical Markov Chain Monte Carlo. *Statistical Science* 1992; 7(4).
28. Kipnis C, Varadhan SRS. Central limit theorem for additive functionals of reversible Markov processes and applications to simple exclusions. *Communications in Mathematical Physics* 1986; 104(1): 1-19.
29. Łatuszyński K, Gareth O. Roberts GO, Rosenthal JS. Adaptive Gibbs samplers and related MCMC methods. *The Annals of Applied Probability* 2013; 23(1): 66–98. http://dx.doi.org/10.1214/11-AAP8 doi: 10.1214/11-AAP8
30. Martin J, Wilcox LC, Burstedde C, Ghattas O. A Stochastic Newton MCMC method for large-scale statistical inverse problems with application to seismic inversion. *SIAM Journal on Scientific Computing* 2012; 34(3): A1460–A1487.
31. Cui T, Law KJH, Marzouk YM. Dimension-independent likelihood-informed MCMC. *Journal of Computational Physics* 2016; 304: 109–137.
32. Roberts GO, Rosenthal JS. Optimal scaling for various Metropolis-Hastings algorithms. *Statistical Science* 2001; 16: 351–367.
33. Andrieu C, Moulines E. On the ergodicity properties of some adaptive Markov chain Monte Carlo algorithms. *Annals of Applied Probability* 2006; 16(1): 462-505.
34. Atchade YF, Rosenthal JS. On Adaptive Markov Chain Monte Carlo Algorithms. *Bernoulli* 2005; 11: 815-828.
35. Cui T. *Bayesian Calibration of Geothermal Reservoir Models via Markov Chain Monte Carlo*. PhD thesis. The University of Auckland, 2010.
36. Christie MA, Blunt MJ. Tenth SPE comparative solution project: A comparison of upscaling techniques. *SPE Reservoir Engineering and Evaluation* 2001; 4: 308–317.
37. Benner P, Gugercin S, Willcox K. A survey of projection-based model reduction methods for parametric dynamical systems. *SIAM Review* 2015; 57(4): 483–531.
38. Ghasemi M, Yang E, Efendiev YR, Calo VM. Fast multiscale reservoir simulations using POD-DEIM model reduction. In: Society of Petroleum Engineers; 2015.
39. Lieberman C, Willcox K, Ghattas O. Parameter and state model reduction for large-scale statistical inverse problems. *SIAM Journal on Scientific Computing* 2015; 32(5): 2523–2542.
40. Cui T, Marzouk Y, Willcox K. Data-driven model reduction for the Bayesian solution of inverse problems. *International Journal for Numerical Methods in Engineering* 2015; 102(5): 966–990.
41. Andrea M, Pagani S, Lassila T. Accurate solution of Bayesian inverse uncertainty quantification problems combining reduced basis methods and reduction error models. *SIAM/ASA Journal on Uncertainty Quantification* 2016; 4(1): 380–412.
42. Pruess K. *TOUGH2 - A General-Purpose Numerical Simulator for Multiphase Fluid and Heat Flow*. Lawrence Berkeley National Laboratory; Berkeley, California: 1991.
43. Grant MA, Donaldson IG, Bixley PF. *Geothermal Reservoir Engineering*. Academic Press . 1982.
44. O'Sullivan MJ. Geothermal reservoir simulation. *International Journal of Energy Research* 1985; 9(3): 319–332.
45. Genuchten vMT. A closed-form equation for predicting the hydraulic conductivity of unsaturated soils. *Soil Science Society of America Journal* 1980; 44: 892-898.