## Introduction

The objective of seismic inversion is to provide information about the internal structure of the Earth from observed data. In the Bayesian framework (Tarantola, 2005; Fichtner et al., 2018), firstly the parameters of interest are estimated by combining the information from observed data with a priori information on the parameters. The information from the observed data is incorporated in the likelihood function, and the a priori information on the parameters is incorporated in the prior distribution. The product of the two determines the posterior distribution, which is the solution to the inverse problem within the Bayesian framework. The uncertainty analysis can be performed through sampling the posterior distribution with the family of Monte Carlo methods. Through the accepted samples, we can infer the parameter fields and conduct the uncertainty analysis.

In this abstract, we will review the gradient-based Markov Chain Monte Carlo (MCMC) and demonstrate its applicability in inferring the uncertainty in seismic inversion. There are many flavours of gradient-based MCMC (Welling and Teh, 2011; Girolami and Calderhead, 2011; Martin et al., 2012; Simsekli et al., 2016; Tong et al., 2019); here we will only focus on the Unadjusted Langevin algorithm (ULA) and Metropolis-Adjusted Langevin algorithm (MALA). We propose an adaptive step-length $\tau$ based on the Lipschitz condition within ULA to automate the tuning of step-length and suppress the Metropolis-Hastings acceptance step in MALA. We consider the linear seismic travel-time tomography problem as a numerical example to demonstrate the applicability of both methods.

## Theoretical Aspects of MCMC with Langevin Dynamics

Consider a probability distribution for a model parameter $\mathbf{m}$ with density function $c\pi(\mathbf{m})$, where $c$ is an unknown normalisation constant, and $\pi$ is a known function. The Langevin diffusion $\mathbf{m}(t)$ for $t \geq 0$, associated with $\pi$ is the solution to the following stochastic differential equation:

$$d\mathbf{m}(t) = \Sigma\nabla \log \pi(\mathbf{m}(t))dt + \sqrt{2}\Sigma^{\frac{1}{2}}dW(t) \tag{1}$$

where $W_t$ for $t \geq 0$ is a standard $n$-dimensional Brownian motion, and given $\Sigma$ is a symmetric positive definite matrix. The Langevin dynamics in equation (1) is ergodic with unique invariant distribution $\pi$ under appropriate assumptions, and if one could solve equation (1) analytically and in the limit as time $t$ goes to infinity then it would be possible to generate samples from a distribution $\pi$ (Brooks et al., 2011; Girolami and Calderhead, 2011).

In a discrete setting, a standard approach is to discretise equation (1) using the Euler-Maruyama discretisation and this produces the Unadjusted Langevin algorithm (ULA) MCMC proposal

$$\mathbf{m}_{t+1} = \mathbf{m}_t + \tau_t\Sigma\nabla \log \pi(\mathbf{m}_t) + \sqrt{2\tau_t}\Sigma^{\frac{1}{2}}\xi, \tag{2}$$

where $\xi_t \sim \mathcal{N}(0, \mathbf{I}_{n\times n})$ is a standard Gaussian random variable in $\mathbf{R}^n$, and $\tau$ is the step-length which can be set to be fixed or varies each iteration. From equation (2), a sophisticated sampler can be designed, and Table 1 shows an overview of main gradient-based MCMC algorithms.

ULA is simple in its implementation, yet it introduces a bias. To tackle this issue, we need to introduce the acceptance-rejection step through the Metropolis-Hastings (M-H) algorithm. The idea is to construct a Markov chain at each step $t$, given $\mathbf{m}_t$, a new candidate $\mathbf{m}_{t+1}$ is generated from a proposal density $q(\mathbf{m}_t, \cdot)$. This candidate is then accepted with probability $\alpha(\mathbf{m}_t, \mathbf{m}_{t+1})$ given by

| Methods | Step-Length, $\tau$ | Pre-Conditioning, $\Sigma$ | M-H |
|---|---|---|---|
| Unadjusted Langevin Algorithm (ULA) | Constant | $\Sigma = \mathbf{I}_{n\times n}$ | No |
| Metropolis-Adjusted Langevin algorihtm (MALA) | Constant | $\Sigma = \mathbf{I}_{n\times n}$ | Yes |
| Riemann manifold MALA (RM-MALA) | Constant | $\Sigma = \mathbf{P}_{n\times n}$ | Yes |
| Stochastic Newton (SN) | $\tau = 1$ | $\Sigma = \mathbf{P}_{n\times n}$ | Yes |
| Stochastic Gradient Langevin Dynamics (SGLD) | Adaptive | $\Sigma = \mathbf{I}_{n\times n}$ or $\Sigma = \mathbf{P}_{n\times n}$ | No |
| Lipschitz-ULA (LULA) | Adaptive | $\Sigma = \mathbf{I}_{n\times n}$ or $\Sigma = \mathbf{P}_{n\times n}$ | No |

***Table 1*** *Overview of main algorithms in gradient-based MCMC.*

$$\alpha(\mathbf{m}_t, \mathbf{m}_{t+1}) = \min\left(\frac{\pi(\mathbf{m}_{t+1})q(\mathbf{m}_t, \mathbf{m}_{t+1})}{\pi(\mathbf{m}_t)q(\mathbf{m}_{t+1}, \mathbf{m}_t)}\right), \tag{3}$$

and rejected with probability $1 - \alpha(\mathbf{m}_t, \mathbf{m}_{t+1})$, i.e., let $\mathbf{m}_{t+1} = \mathbf{m}_{t+1}$ with probability $\alpha(\mathbf{m}_t, \mathbf{m}_{t+1})$, and $\mathbf{m}_{t+1} = \mathbf{m}_t$ with probability $1 - \alpha(\mathbf{m}_t, \mathbf{m}_{t+1})$. The resulting Markov chain is reversible with respect to distribution $\pi$ and under mild assumption is ergodic (Brooks et al., 2011). By introducing Metropolis-Hastings (M-H) algorithm into ULA, we will obtain the Metropolis-Adjusted Langevin algorithm (MALA) MCMC, and the pseudocode for MALA MCMC is presented below.

**Algorithm 1** Metropolis Adjusted Langevin algorithm (MALA) MCMC

```
Input: Initial model $\mathbf{m}_1$, step-length $\tau > 0$
Output: $N$ number of samples
  for $t = 1$ to $N - 1$ do
    Draw diffusion vector $\xi_t \sim \mathcal{N}(0, \mathbf{I}_{n\times n})$
    Propose $\mathbf{m}' = \mathbf{m}_t + \tau_t\Sigma\nabla_{\mathbf{m}_t}\pi(\mathbf{m}_t) + \sqrt{2\tau_t}\Sigma^{\frac{1}{2}}\xi_t$
    Compute the accept-reject probability
    $\alpha(\mathbf{m}'|\mathbf{m}_t) = \min\left(1, \frac{\pi(\mathbf{m}')\mathcal{N}(\mathbf{m}'|\mathbf{m}_t+\tau_t\Sigma\nabla_{\mathbf{m}_t}\pi(\mathbf{m}_t), 2\tau_t\Sigma)}{\pi(\mathbf{m}_t)\mathcal{N}(\mathbf{m}_t|\mathbf{m}'+\tau_t\Sigma\nabla_{\mathbf{m}'}\pi(\mathbf{m}'), 2\tau_t\Sigma)}\right)$
    Draw $u$ from a uniform distribution of a range $[0, 1]$.
    if $\alpha_t > u$ then
      Accept: $\mathbf{m}_{t+1} = \mathbf{m}'$
    else
      Reject: $\mathbf{m}_{t+1} = \mathbf{m}_t$
    end if
  end for
```

The main advantage of gradient-based MCMC is that it directs the proposed moves towards areas of high probability for the distribution $\pi$, using the gradient of the logarithm of distribution $\pi$, $\nabla \log \pi(\mathbf{m})$. However, one issue with MCMC samplers, in general, is that their step-length, $\tau$, needs to be tuned. In particular that $\tau$ must decrease with dimension, $n$. By fixing the acceptance ratio with small $\tau$ leads to large acceptance ratio; however, all accepted steps are small (on average on the order of $\tau$ ), so that the sampler moves often, but slowly. In general, it takes $\mathcal{O}(1/\tau)$ iterations to move through the support of the target distribution after the burn-in period (Roberts and Rosenthal, 1998; Neal and Roberts, 2006). Besides, by introducing the Metropolis-Hastings (M-H) algorithm, this will introduce substantial computational costs, and choosing an appropriate proposal distribution for the Metropolis-Hastings algorithm is non-trivial. For high dimensions, many MCMC samplers, including gradient-based MCMC, are thus slow to converge and computationally costly. We propose ULA with the step-length, $\tau$, based on the Lipschitz condition

$$\tau_t = \frac{1}{2}\frac{||\mathbf{m}_{t+1} - \mathbf{m}_t||_2}{||\nabla \log \pi(\mathbf{m}_{t+1}) - \nabla \log \pi(\mathbf{m}_t)||_2}, \tag{4}$$

an extension of ULA and similar in spirit with Stochastic Gradient Langevin Dynamics (SGLD) algorithm proposed by (Welling and Teh, 2011) by suppressing the Metropolis-Hastings acceptance steps. In this algorithm, the step-length is adaptive to the local geometry, with convergence guarantees depending only on smoothness in a neighbourhood of a solution. Thus, it shares the benefits of convergence as in optimization and sampling from the posterior as in MCMC without to manually tune the step-length $\tau$ and to undergo computationally demanding Metropolis-Hastings acceptance steps.

In summary, in designing or choosing an MCMC sampler for a given distribution, one typically considers the following three criteria. First, the type of proposal distribution is chosen based on or in relative to how much information about the target distribution is available. Second, the step-length which acts as a tuning parameter to control how far the proposed sample deviates from the current MCMC state. The choice of step-length needs to be done wisely as it influences the mixing rate. The optimal choices of the step-length are problem dependent, and it may depend, among other things, on the choice of proposal distribution, the computational resources available, the effective dimension of the problem, and the overall desired accuracy of the MCMC computation. Lastly, one may propose an n-dimensional update via an n-dimensional proposal, or one can propose the Gibbs sampler, which at each step in the chain, an update for an n/m-dimensional "block" of variables.

## Numerical Example

In this numerical example, we consider the linear seismic travel-time tomography problem as in (Gazzola et al., 2019). We consider a square domain of size $[0,30] \times [0,30]$ with 15 equally spaced sources located at the right side of the domain, and 25 equally spaced receivers located on the top of the surface and the left side of the domain. For sampling configurations, we start the algorithms from a constant initial model, and for prior distribution, we consider Gaussian prior with covariance matrix $\mathbf{C}_{prior} = (\mathbf{L}^T\mathbf{L})^{-1}$, where $\mathbf{L}$ is the Laplacian matrix. Then, we perform 4 chains of MCMC sampling with $N = 1,000,000$. For comparison, we implement ULA, MALA, and Lipschitz-ULA (LULA). We fix the step-length $\tau = 1/\sigma_1(\mathbf{H})$, where $\sigma_1(\mathbf{H})$ is the first singular values of the Hessian, for ULA and MALA. As for LULA, the step-length $\tau$ equals equation (4). Here, we consider $\Sigma = \mathbf{I}_{n\times n}$. The results are illustrated in the figures below.

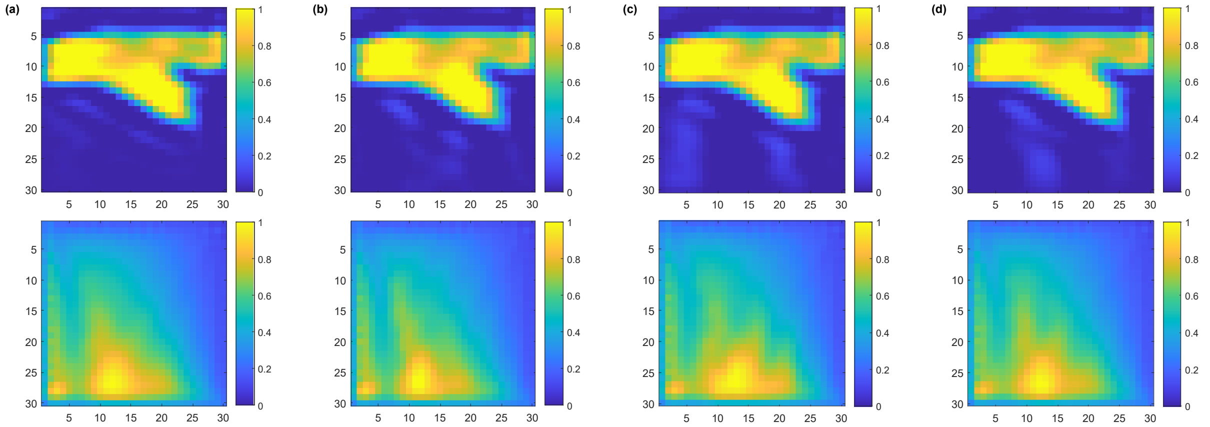

*Figure 1* *Columns: Posterior mean and normalised variance for (a) Maximum a posteriori (MAP) model, (b) samples from ULA, (c) samples from MALA, and (d) samples from LULA. In this experiment, we consider $2 \times 10^3$ as burn-in period based on the expected error plot in Figure 2.*

## Conclusions

In this extended abstract, we have provided a short introduction to gradient-based MCMC algorithms, and we proposed the usage of adaptive step-length $\tau$, based on the Lipschitz condition within ULA (LULA) as an extension to ULA to automate the step-length tuning, and to suppress the Metropolis-Hastings acceptance step in MALA. The smallest step-length indicates a slow mixing around the density, while the too large step-length causes the sampler to stay put for long periods of times with an adverse

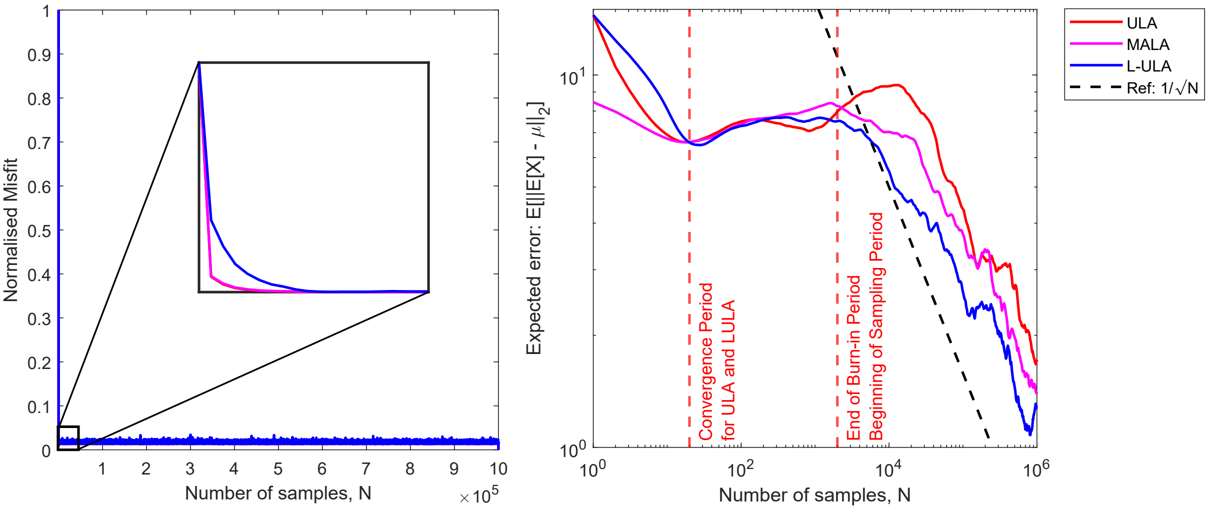

*Figure 2 Left: Normalised misfit curve. Right: Expected error* $\mathbf{E}[||\mathbf{E}[\mathbf{m}]-\mu||_2]$*. The expectation on number of chains for the* $\ell_2$*-norm of the difference between the sample mean* $\mathbf{E}[\mathbf{m}]$ *and the mean* $\mu$*.*

effect on the independence of the samples. In general, gradient-based MCMC is still a relatively new class of Monte Carlo algorithms compared to traditional MCMC methods, and there remain many open problems and opportunities for further research in this area. Some critical areas for future development in gradient-based MCMC include scalable MCMC algorithms for large-scale and non-linear problems, tuning techniques and pre-conditioning to increase the mixing and convergence rates.

## References

Brooks, S., Gelman, A., Jones, G. and Meng, X.L. [2011] *Handbook of Markov Chain Monte Carlo*. CRC press.

Fichtner, A., Zunino, A. and Gebraad, L. [2018] Hamiltonian Monte Carlo solution of tomographic inverse problems. *Geophysical Journal International*, **216**(2), 1344–1363.

Gazzola, S., Hansen, P.C. and Nagy, J.G. [2019] IR Tools: a MATLAB package of iterative regularization methods and large-scale test problems. *Numerical Algorithms*, **81**(3), 773–811.

Girolami, M. and Calderhead, B. [2011] Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **73**(2), 123–214.

Martin, J., Wilcox, L., Burstedde, C. and Ghattas, O. [2012] A Stochastic Newton MCMC Method for Large-Scale Statistical Inverse Problems with Application to Seismic Inversion. *SIAM Journal on Scientific Computing*, **34**(3), A1460–A1487.

Neal, P. and Roberts, G.O. [2006] Optimal Scaling For Partially Updating Mcmc Algorithms.

Roberts, G.O. and Rosenthal, J.S. [1998] Optimal scaling of discrete approximations to Langevin diffusions.

Simsekli, U., Badeau, R., Cemgil, A.T. and Richard, G. [2016] Stochastic Quasi-Newton Langevin Monte Carlo. In: *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16. JMLR.org, 642–651.

Tarantola, A. [2005] *Inverse Problem Theory and Methods for Model Parameter Estimation*. Society for Industrial and Applied Mathematics.

Tong, X.T., Morzfeld, M. and Marzouk, Y.M. [2019] MALA-within-Gibbs samplers for high-dimensional distributions with sparse conditional structure.

Welling, M. and Teh, Y.W. [2011] Bayesian Learning via Stochastic Gradient Langevin Dynamics. In: *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11. Omnipress, USA, 681–688.