2021 EAGE ANNUAL
82ND CONFERENCE & EXHIBITION
AMSTERDAM | THE NETHERLANDS

APPROXIMATE LANGEVIN MONTE CARLO WITH ADAPTATION FOR BAYESIAN FULL-WAVEFORM INVERSION

M. Izzatullah[1], T. Van Leeuwen[2], D. Peter[1]

[1] King Abdullah University of Science and Technology; [2] Utrecht University

Summary

In this work, we present a proof of concept for Bayesian full-waveform inversion (FWI) in 2-D. This is based on approximate Langevin Monte Carlo sampling with a gradient-based adaptation of the posterior distribution. We apply our method to the Marmousi model, and it reliably recovers important aspects of the posterior, including the statistical moments, and 1-D and 2-D marginals. Depending on the variations of seismic velocities, the posterior can be significantly non-Gaussian, which directly suggest that using a Hessian approximation for uncertainty quantification in FWI may not be sufficient.

Approximate Langevin Monte Carlo with Adaptation for Bayesian Full-waveform Inversion

Introduction

Seismic full-waveform inversion (FWI) addresses the geophysical inverse problem of estimating subsurface model parameters from observed seismic data. The common approach to quantify and to analyse the uncertainties is through Bayesian inference framework (e.g., Fichtner and Zunino, 2019; Fichtner et al., 2019; Gebraad et al., 2020; Izzatullah et al., 2020). In this work, we focus on a scalable and computationally sophisticated MCMC algorithm based on the Langevin diffusion process for a large-scale Bayesian inference such as FWI. Inspired by the work of Kingma and Ba (2015) and Nemeth and Fearnhead (2020), we aim to bridge the fields of optimization and Bayesian inference through approximate Langevin Monte Carlo algorithm. This algorithm has advantageous computational costs in large-scale problems because of its fast sampling, due to the absence of the Metropolis-Hastings acceptance criterion.

We present a proof of concept for Bayesian full-waveform inversion in 2-D. This is based on approximate Langevin Monte Carlo sampling with a gradient-based adaptation of the posterior distribution. We apply our method to the Marmousi model, and it reliably recovers important aspects of the posterior, including the statistical moments, and 1-D and 2-D marginals. Depending on the variations of seismic velocities, the posterior can be significantly non-Gaussian, which suggests that using a Hessian approximation for uncertainty quantification in FWI may not be sufficient.

Bayesian Inference Framework

To quantify uncertainties in FWI, we reformulate it within Bayesian inference framework. In Bayesian framework, we introduce the notion of the prior probability density and the likelihood function. The prior probability density $\pi_{prior}(\mathbf{m})$ encodes the confidence we have in the prior information on the unknown subsurface model parameters $\mathbf{m}$, and the likelihood function $\pi_{like}(\mathbf{D}|\mathbf{m})$ describes the conditional probability density that the subsurface model parameters give rise to the actual seismic data $\mathbf{D}$. Based on Bayes' theorem, we obtain the posterior probability density $\pi_{post}(\mathbf{m}|\mathbf{D})$ by combining the prior probability density and the likelihood function

$$\pi_{post}(\mathbf{m}|\mathbf{D}) \propto \pi_{like}(\mathbf{D}|\mathbf{m})\pi_{prior}(\mathbf{m}), \tag{1}$$

where the posterior probability density $\pi_{post}(\mathbf{m}|\mathbf{D})$ can be evaluated up to its normalizing constant. Note that to evaluate the equation above, particularly in high dimensions, maybe intractable to compute and impossible to interpret. Thus, we refer to the MCMC algorithms for its evaluation.

Approximate Langevin Monte Carlo

At the basis of our proposed algorithm is a class of MCMC algorithms known as Langevin Monte Carlo (LMC). LMC is originally derived from the Langevin diffusion process, which can be described by

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

where W_t for $t \geq 0$ is a standard d-dimensional Brownian motion, and given Σ is a symmetric positive definite preconditioning matrix. The evolution of $\mathbf{m}(t)$ is controlled by a deterministic drift term proportional to the gradient of log-density $\pi(\mathbf{m}(t))$. The Langevin diffusion in equation (2) is ergodic with unique invariant distribution π, and if one could solve equation (2) analytically in the limit as time t goes to infinity, then it would be possible to generate samples from a distribution π. However, in practice, to simulate the Langevin diffusion, it is necessary to use a discrete approximation, such as Euler-Maruyama discretization. This produces the Unadjusted Langevin algorithm (ULA) MCMC proposal

$$\mathbf{m}_{t+1} = \mathbf{m}_t + \lambda\Sigma\nabla \log \pi(\mathbf{m}_t) + \sqrt{2\lambda}\Sigma^{\frac{1}{2}}\xi, \tag{3}$$

where $\xi_t \sim \mathcal{N}(0, \mathbf{I}_{n\times n})$ is d-dimensional vector of standard Gaussian random variable, and $\lambda > 0$ is a temporal step size. ULA resembles the gradient descent algorithm but with injected Gaussian noise. It

also belongs to the approximate MCMC family as it has no Metropolis-Hastings acceptance criterion in its procedure. In addition, ULA with the acceptance criterion is known as Metropolis-adjusted Langevin algorithm (MALA). ULA and MALA shares $\mathcal{O}(d)$ computational cost due to the gradient evaluation. However, this computational cost will be a bottleneck for MALA as sample rejections rate increase, especially in large-scale inference problems. This will give ULA an advantage as all the samples will be accepted with probability one. In this work, we focus on ULA and introduce an adaption using a *diagonal* adaptive preconditioning matrix and an *adaptive drift vector* by exploiting the previously-computed gradients in an online fashion to improve algorithm's performance and efficiency. To emphasize the understanding of the proposed algorithm, we use the following notation: $\odot$, $\oslash$, and $^{\circ\frac{1}{2}}$ are element-wise multiplication, division, and square root, respectively. Consider a set of samples $\mathbf{m}_1,\ldots,\mathbf{m}_t$ and its gradient $\mathbf{g}_1,\ldots,\mathbf{g}_t$ produced by ULA. We may first forms an *accumulation* vector as:

$$\gamma_t = \sum_{\tau=1}^{t} \beta^{\tau-1} + (1-\beta)\mathbf{g}_\tau \odot \mathbf{g}_\tau \tag{4}$$

The constant $\beta \in [0,1]$ results in accumulation with exponentially decaying weights. Next, we use γ_t to approximate the preconditioning matrix at $\mathbf{m}_t$ as:

$$\Sigma_t = diag\left(1 \oslash \left(10^{-6} + (\gamma_t)^{\circ\frac{1}{2}}\right)\right) \tag{5}$$

We observe that the computation of the preconditioning matrix Σ_t only uses previously computed gradients, and both computation and factorization only using simple *scalar* operations which save computational budgets. In addition to the preconditioning matrix, we introduce an adaptive drift vector in ULA which can be computed as:

$$\mu_t = \sum_{\tau=1}^{t} \alpha^{\tau-1} + (1-\alpha)\mathbf{g}_\tau \tag{6}$$

where $\alpha \in [0,1]$ controls the exponential decay of older gradients. Implementing the diagonal preconditioning matrix and adaptive drift vector into ULA gives us an approximate Langevin Monte Carlo algorithm with gradient-based adaptation. We present the pseudocode in Algorithm 1 below.

Algorithm 1 Approximate Langevin Monte Carlo with gradient-based adaptation

Input: Initial model $\mathbf{m}_0$, step size $\lambda > 0$, exponential decay rates $\alpha, \beta \in [0,1]$
Output: N number of samples
  Initialize $\gamma_t = 0$ and $\mu_t = 0$
  for $t = 0$ to $N-1$ **do**
    Compute gradient: $\mathbf{g}_t = \nabla \log \pi(\mathbf{m}_t)$
    Compute *accumulation* vector: $\gamma_{t+1} = \beta\gamma_t + (1-\beta)\mathbf{g}_t \odot \mathbf{g}_t$
    Compute *preconditioning* matrix: $\Sigma_{t+1} = diag\left(1 \oslash \left(10^{-6} + (\gamma_{t+1})^{\circ\frac{1}{2}}\right)\right)$
    Compute *adaptive drift vector*: $\mu_{t+1} = \alpha\mu_t + (1-\alpha)\mathbf{g}_t$
    Draw diffusion vector: $\xi_t \sim \mathcal{N}(0, \mathbf{I}_{n\times n})$
    Sample: $\mathbf{m}_{t+1} = \mathbf{m}_t + \lambda\Sigma_{t+1} \odot \mu_{t+1} + \sqrt{2\lambda}(\Sigma_{t+1})^{\circ\frac{1}{2}} \odot \xi_t$
  end for

Numerical Example

We consider the Marmousi model with a domain size $3,000 \times 11,000$ m as in Figure 1(a). We discretize the model with a grid spacing of 50 m, yielding $13,420$ unknown parameters. At the surface, we place 55 sources and 110 receivers with a horizontal sampling interval of 100 m and 50 m, respectively. The signal-to-noise ratio in the data is 0.059 dB, and the relative standard deviation of the observation noise is 5%. We use a frequency content from 1 Hz to 4 Hz with uniform frequency sampling of 1 Hz. All frequencies are used simultaneously in sampling procedure, no multi-scale strategy is applied.

For this problem, we set the data error covariance matrix $\mathbf{C_D} = \sigma_D^2 I_\mathbf{D}$ with $\sigma_D = 0.003$ and $I_\mathbf{D}$ the identity matrix. For the model prior, we use uniform distributions within certain bounds. The width of the prior

reflects the minimum and maximum velocity values of the Marmousi model. We started sampling with an initial model obtained by smoothing the true model with Gaussian kernel. In this numerical example, we set $\alpha = 0.9$, $\beta = 0.999$, and we consider a fixed step size $\lambda = 0.0001$. We perform the proposed algorithm with $50,000$ iterations. The number of iterations is set to be very large to guarantee we can sample the target distribution neighbourhood.

The results of the proposed method are statistical assessments. We plot the statistical moments in Figure 1 below. The first statistical moment model is similar and very close to the true model. The variance model quantifies the variations for the Bayesian inference. We observe small variation for the shallow part of the model. This is because we have good data coverage and model illumination. However, as we go deeper and towards the corners, we observe large variations. Those regions are poorly illuminated, and the inferred velocities spread out over a wider range of values. The third statistical moment model measures asymmetric or "non-Gaussianness" of a distribution. Nonzero values indicate a non-Gaussian behaviour in the posterior. We observe that for many parameters, especially at the shallow region, indicating a strong non-Gaussian posterior. In addition to the statistical moments, sampling also allows us to visualize marginal and conditional distributions. We also plot the 1-D and 2-D marginal plots in Figure 2. As a consequence of the nonlinearity of FWI, the marginal is significantly non-Gaussian. This suggests that the use of a Hessian approximation for uncertainty quantification in FWI may not be sufficient.

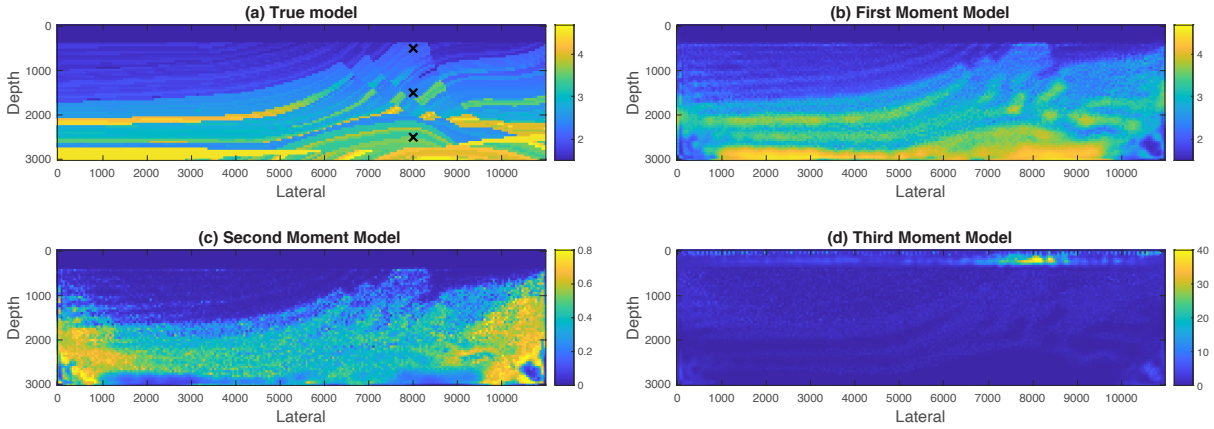

Figure 1 *Statistical moments obtained from Bayesian FWI. (a) True model with the black crosses represent the chosen elements for selected posterior distributions, the* 9709*-th,* 9729*-th, and* 9749*-th elements, respectively. (b) First statistical moment model (mean), (c) second statistical moment model (variance), and (d) third statistical moment model (skewness).*

Conclusions

We have introduced a scalable and computationally sophisticated MCMC algorithm for a large-scale Bayesian inference such as FWI. We demonstrated the proposed algorithm for a Bayesian FWI in 2-D Marmousi model. Our key ingredients are (1) diagonal preconditioning and (2) adaptive drift vector, computed using only past gradients in an online fashion to improve the algorithm's performance and efficiency. The proposed algorithm recovers important aspects of the posterior, which can be significantly non-Gaussian. Consequently, this suggests that using a Hessian approximation for uncertainty quantification in FWI may not be sufficient. In general, approximate LMC is still a relatively new class of Monte Carlo algorithms compared to the traditional MCMC methods. There remain many open problems and opportunities for further research in this area.

Acknowledgments

The first author would like to thank Tristan van Leeuwen at Utrecht University for visiting his research lab, which led to this work, and his continuous support. The research visits and the work reported here

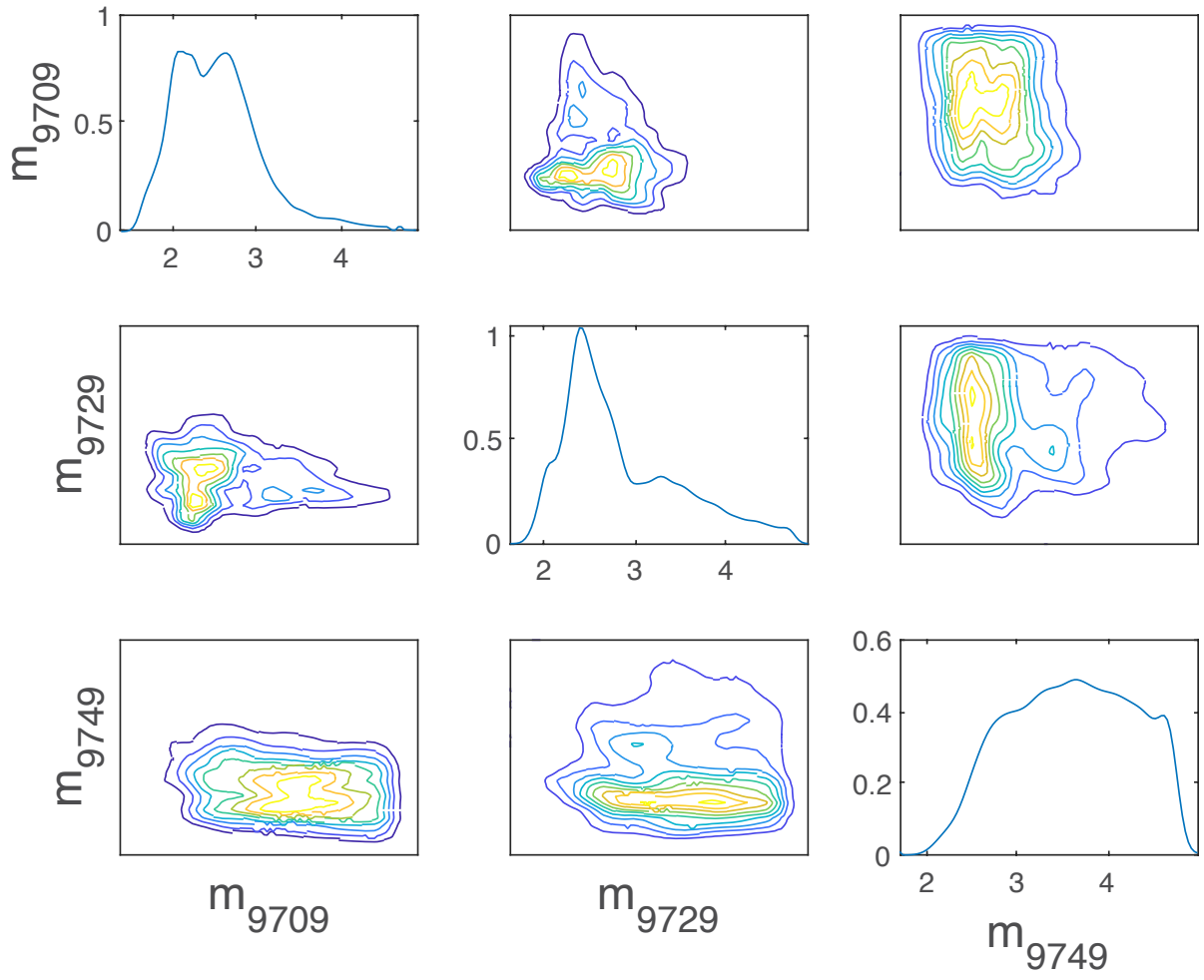

Figure 2 *Selected posterior distributions.* ***Diagonal****: 1-D distribution.* ***Off-diagonal:*** *2-D bivariate joint distributions. These 2-D marginals are extreme cases of non-Gaussian behavior in the obtained posteriors. This suggests that using a Hessian approximation for uncertainty quantification in FWI may not be sufficient.*

were supported by funding from King Abdullah University of Science and Technology (KAUST).

References

Fichtner, A. and Zunino, A. [2019] Hamiltonian Nullspace Shuttles. *Geophysical Research Letters*, **46**(2), 644–651.

Fichtner, A., Zunino, A. and Gebraad, L. [2019] Hamiltonian Monte Carlo solution of tomographic inverse problems. *Geophysical Journal International*, **216**(2), 1344–1363.

Gebraad, L., Boehm, C. and Fichtner, A. [2020] Bayesian Elastic Full-Waveform Inversion Using Hamiltonian Monte Carlo. *Journal of Geophysical Research: Solid Earth*, **125**(3), e2019JB018428.

Izzatullah, M., Van Leeuwen, T. and Peter, D. [2020] Langevin Dynamics Markov Chain Monte Carlo Solution for Seismic Inversion. **2020**(1), 1–5.

Kingma, D.P. and Ba, J. [2015] Adam: A Method for Stochastic Optimization. In: Bengio, Y. and LeCun, Y. (Eds.) *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Nemeth, C. and Fearnhead, P. [2020] Stochastic gradient Markov chain Monte Carlo. *Journal of the American Statistical Association*, **0**(ja), 1–47.