Geophysical Journal International

Geophys. J. Int. (2020) **222,** 406–411
Advance Access publication 2020 April 11
GJI Seismology

doi: 10.1093/gji/ggaa170

EXPRESS LETTER

Variational full-waveform inversion

Xin Zhang and Andrew Curtis

School of GeoSciences, University of Edinburgh, Edinburgh EH93FE, UK. E-mail: X.Zhang2@ed.ac.uk

Accepted 2020 April 5. Received 2020 March 30; in original form 2020 February 13

SUMMARY

Seismic full-waveform inversion (FWI) can produce high-resolution images of the Earth's subsurface. Since full-waveform modelling is significantly nonlinear with respect to velocities, Monte Carlo methods have been used to assess image uncertainties. However, because of the high computational cost of Monte Carlo sampling methods, uncertainty assessment remains intractable for larger data sets and 3-D applications. In this study, we propose a new method called variational FWI, which uses Stein variational gradient descent to solve FWI problems. We apply the method to a 2-D synthetic example and demonstrate that the method produces accurate approximations to those obtained by Hamiltonian Monte Carlo. Since variational inference solves the problem using optimization, the method can be applied to larger data sets and 3-D applications by using stochastic optimization and distributed optimization.

Key words: Inverse theory; Probability distributions; Waveform inversion.

1 INTRODUCTION

Seismic full-waveform inversion (FWI) is a method that characterizes properties of the Earth's subsurface by exploiting information throughout recorded seismic waveforms (Tarantola 1984; Gauthier *et al.* 1986; Pratt 1999; Tromp *et al.* 2005). The method has been used successfully from industrial scale (Prieux *et al.* 2013; Warner *et al.* 2013), regional scale (Chen *et al.* 2007; Fichtner *et al.* 2009; Tape *et al.* 2009) to global scale (French & Romanowicz 2014; Bozdağ *et al.* 2016; Fichtner *et al.* 2018).

The FWI problem is often solved using optimization by minimizing a misfit function between observed and predicted seismograms. Since the problem is highly nonlinear with multimodal objective functions, a poor starting model can cause convergence to incorrect solutions. Apart from finding an adequate starting model, numerous misfit functions that can reduce multimodalities have been proposed (Luo & Schuster 1991; Fichtner *et al.* 2008; Brossier *et al.* 2010; Van Leeuwen & Mulder 2010; Bozdağ *et al.* 2011; Métivier *et al.* 2016). Nevertheless, although optimization has been used widely in practical applications, the method cannot provide accurate uncertainty estimations which makes it difficult to assess and interpret the results of FWI.

Monte Carlo sampling methods provide a procedure to solve general nonlinear problems and quantify uncertainties (Brooks *et al.* 2011). The methods have been applied to traveltime tomography (Bodin & Sambridge 2009; Galetti *et al.* 2015; Zhang *et al.* 2018, 2019) and FWI (Ray *et al.* 2016; Biswas & Sen 2017; Ray *et al.* 2017; Gebraad *et al.* 2020). However, Monte Carlo sampling methods are computationally expensive and remain intractable for large data sets due to the curse of dimensionality (Curtis & Lomax 2001).

To extend nonlinear uncertainty analysis to larger systems, Nawaz & Curtis (2018, 2019) and Zhang & Curtis (2019) introduced variational inference methods to geophysics, and Zhang & Curtis (2019) applied them to seismic traveltime tomography. By optimizing a different formulation of the inverse problem, variational inference methods can be more efficient than Monte Carlo sampling methods (Bishop 2006; Blei *et al.* 2017), can be applied to larger systems by using methods like stochastic optimization (Robbins & Monro 1951; Kubrusly & Gravier 1973) and distributed optimization, and provide uncertainties in the form of marginal probability distributions on parameters (Nawaz & Curtis 2018, 2019; Zhang & Curtis 2019; Nawaz *et al.* 2020).

In this study, we apply variational inference methods to FWI, which we refer as variational FWI (VFWI). Specifically we use Stein variational gradient descent (SVGD) to solve FWI problems because SVGD can produce accurate approximations to the results of Monte Carlo sampling methods (Zhang & Curtis 2019). In section 2, we provide a brief overview of SVGD and FWI. In section 3, we apply the method to a 2-D synthetic test and compare the results with those obtained by Hamiltonian Monte Carlo (HMC). We then provide a discussion about the possibility to apply the method to larger systems and 3-D applications.

2 METHODS

2.1 SVGD

Bayesian methods update a *prior* probability density function (pdf) $p(\mathbf{m})$ with new information from the data to produce a probability distribution of model parameters post inversion, which is often

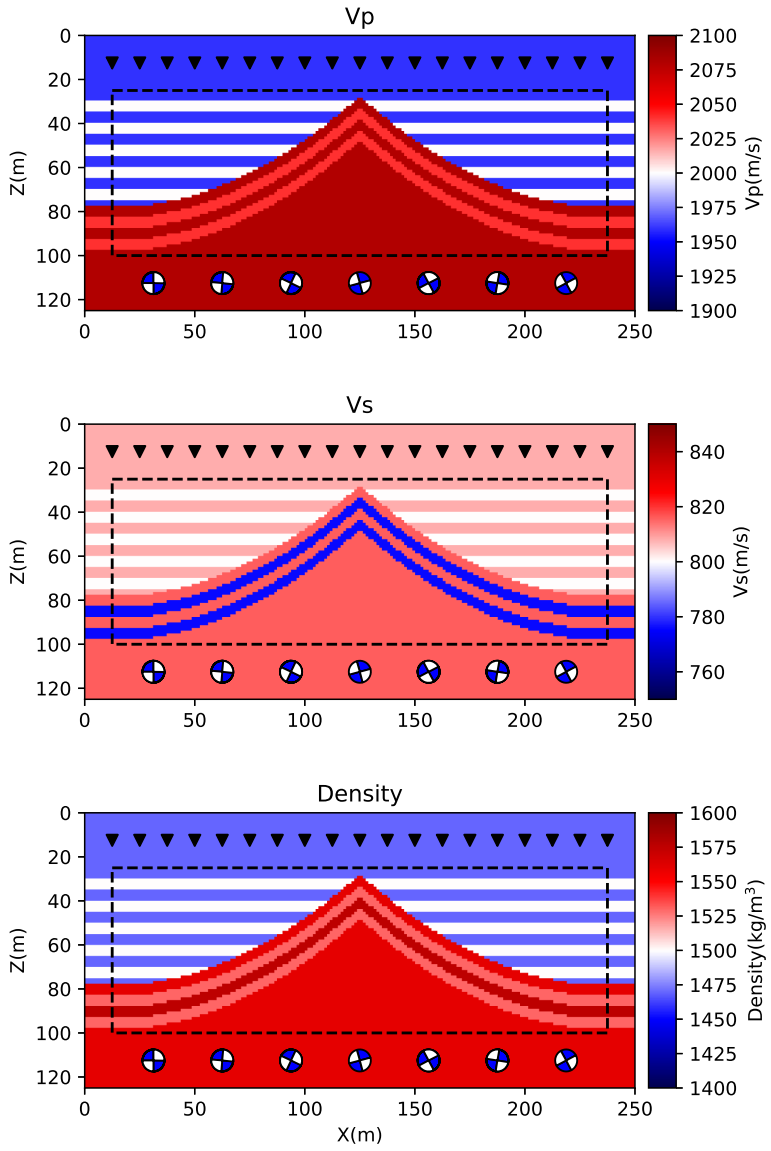

Figure 1. The true model for Vp, Vs and density. The dashed black line indicates the study region within which parameters are inverted. Sources are located at the bottom which are represented by beachballs and receivers are shown with black triangles.

called a *posterior* pdf, written as $p(\mathbf{m}|\mathbf{d}_{obs})$. According to Bayes' theorem,

$$p(\mathbf{m}|\mathbf{d}_{\text{obs}}) = \frac{p(\mathbf{d}_{\text{obs}}|\mathbf{m})p(\mathbf{m})}{p(\mathbf{d}_{\text{obs}})}, \tag{1}$$

where $p(\mathbf{d}_{\text{obs}}|\mathbf{m})$ is the *likelihood*, which is the probability of observing data $\mathbf{d}_{\text{obs}}$ if model $\mathbf{m}$ was true, and $p(\mathbf{d}_{\text{obs}})$ is a normalization factor called the *evidence*. The likelihood function is often represented as the exponential of a misfit function $L(\mathbf{d}_{\text{obs}}, \mathbf{m})$,

$$p(\mathbf{d}_{\text{obs}}|\mathbf{m}) = \frac{1}{C}\exp(-L(\mathbf{d}_{\text{obs}}, \mathbf{m})) \tag{2}$$

where C is the normalization factor. This process is called Bayesian inference.

Bayesian inference is often solved by using Markov chain Monte Carlo (McMC) methods. However, due to the high computational expense of Monte Carlo methods, they cannot easily be applied to large dats sets that are often expensive to simulate given a set of model parameters. Variational inference provides a different way to solve Bayesian inference problems: the method seeks an optimal approximation to the posterior pdf within a pre-defined family of distributions by minimizing the Kullback–Leibler (KL) divergence (Kullback & Leibler 1951) between the approximate probability distribution and the posterior probability distribution (Blei *et al.* 2017). Since variational inference solves Bayesian inference problems using optimization, it can be more efficient than Monte Carlo sampling methods (Blei *et al.* 2017; Zhang & Curtis 2019).

Stein variational gradient descent (SVGD) is one such algorithm based on iterative incremental transforms of the prior pdf (Liu & Wang 2016). In SVGD, a smooth transform $T(\mathbf{m}) = \mathbf{m} + \epsilon\boldsymbol{\phi}(\mathbf{m})$ is used, where $\mathbf{m} = [m_1, ..., m_d]$ and m_i is the ith parameter, and $\boldsymbol{\phi}(\mathbf{m}) = [\phi_1, ..., \phi_d]$ is a smooth vector function that describes the perturbation direction and where ϵ is the magnitude of the perturbation. Say $q_T(\mathbf{m})$ is the transformed probability distribution of the initial distribution $q(\mathbf{m})$. The gradient of KL divergence with respect to ϵ can be calculated as (Liu & Wang 2016)

$$\nabla_\epsilon \text{KL}[q_T||p]\,|_{\epsilon=0} = -\text{E}_q\left[\text{trace}\left(A_p\boldsymbol{\phi}(\mathbf{m})\right)\right] \tag{3}$$

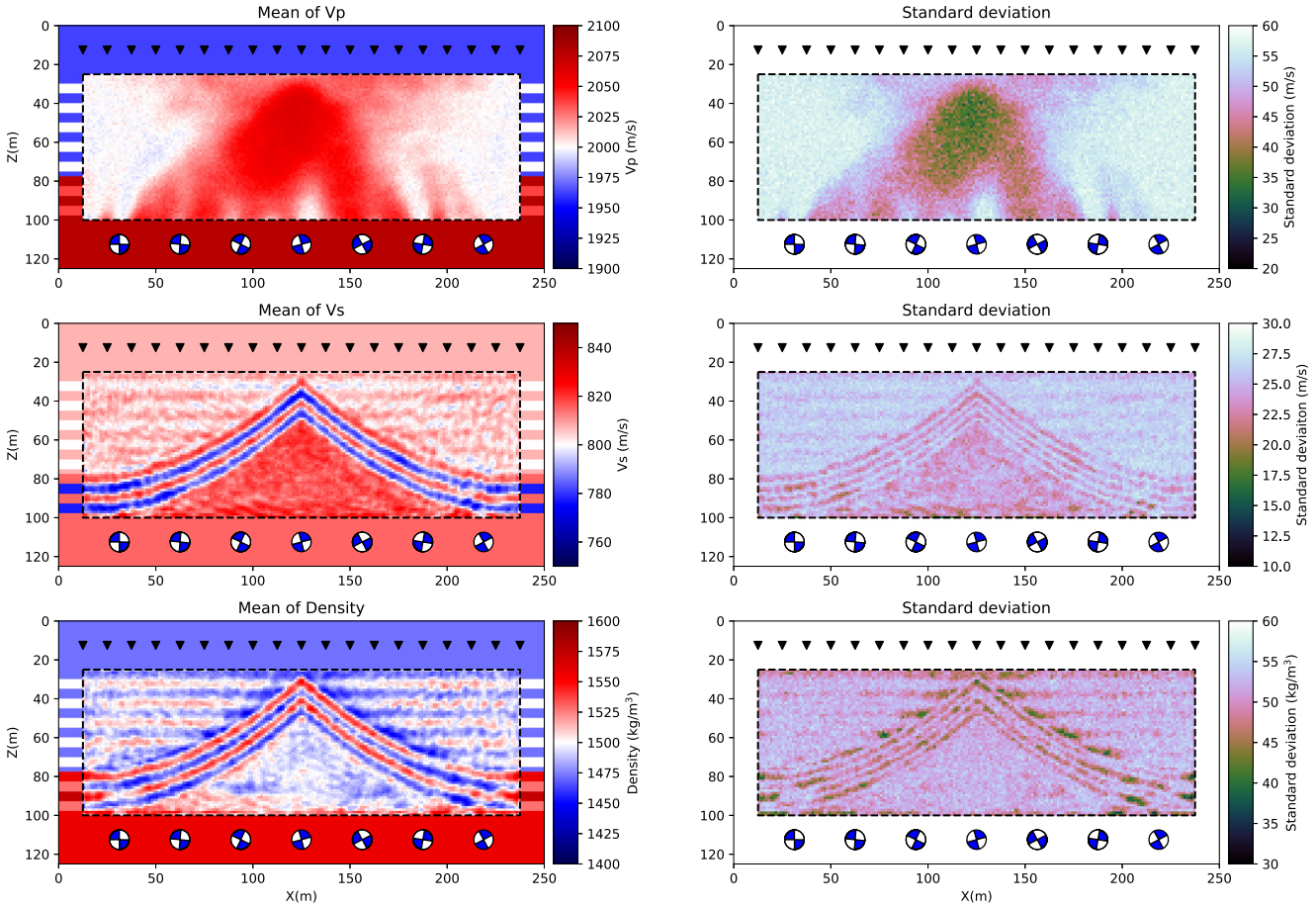

Figure 2. The mean (left) and standard deviation (right) for Vp, Vs and density obtained using SVGD.

where A_p is the Stein operator such that $A_p\boldsymbol{\phi}(\mathbf{m}) = \nabla_{\mathbf{m}}\log p(\mathbf{m})\boldsymbol{\phi}(\mathbf{m})^T + \nabla_{\mathbf{m}}\boldsymbol{\phi}(\mathbf{m})$. It has been found that the right-hand expectation is maximized when

$$\boldsymbol{\phi}^* = \boldsymbol{\phi}^*_{q,p}(\mathbf{m})/||\boldsymbol{\phi}^*_{q,p}(\mathbf{m})||_{H^d} \tag{4}$$

and

$$\boldsymbol{\phi}^*_{q,p}(\mathbf{m}) = \mathrm{E}_{\{\mathbf{m}'\sim q\}}[A_p k(\mathbf{m}', \mathbf{m})] \tag{5}$$

where H^d is a reproducing kernel Hilbert space and $k(\mathbf{m}', \mathbf{m})$ is a kernel function (see details in Liu & Wang 2016). The expectation in eq. (5) can be calculated using a set of particles.

Given the above solution, SVGD minimizes the KL divergence by iteratively applying the optimal transform $\boldsymbol{\phi}^*$ to the current approximate probability distribution represented by a set of particles, and eventually converges to an approximation to the true posterior. The method has been introduced to geophysics to solve 2-D seismic traveltime tomographic problems by Zhang & Curtis (2019). In this study we use SVGD to solve VFWI problems.

2.2 FWI

FWI uses full-waveform information to image the Earth's subsurface. In this study, we solve a *P–SV* wave system along a 2-D vertical cross-section of isotropic wave velocities and density. The wave equation is solved by using a fourth-order variant of the staggered-grid finite-difference scheme (Virieux 1986; Gebraad *et al.* 2020). The gradients with respect to velocities and density are calculated using the adjoint method (Tarantola 1988; Fichtner *et al.* 2006; Liu & Tromp 2006; Plessix 2006) and are used to transform the pdf in the SVGD algorithm. For the misfit function, we choose the L_2 waveform difference:

$$L = \frac{1}{2}\sum_i \left(\frac{d_i^{\mathrm{obs}} - d_i(\mathbf{m})}{\sigma_i}\right)^2 \tag{6}$$

where i is the index of time samples and σ_i is the standard deviation of each data point. Since the L_2 misfit is dominated by large amplitude shear waves, it is probably more sensitive to shear velocities than to *P*-wave velocities.

3 RESULTS

We apply the above method to a 2-D synthetic example identical to that in Gebraad *et al.* (2020) who used a particularly efficient MC method, so that the results can be fairly compared. Fig. 1 shows the true model for Vp, Vs and density. Sources are located at the bottom of the region and have random moment tensors. For source-time function, we use a Ricker wavelet with dominant frequency of 50 Hz. Receivers are located at the depth of 10 m near the surface. The data are simulated using the staggered-grid finite-difference scheme over a 220 × 110 gridded discretization in space, within which a 180 × 60 subgrid of cells has free parameters (region within the dashed black box in Fig. 1).

To reduce the complexity of the inverse problem, we use strong prior information as in Gebraad *et al.* (2020): Uniform distributions

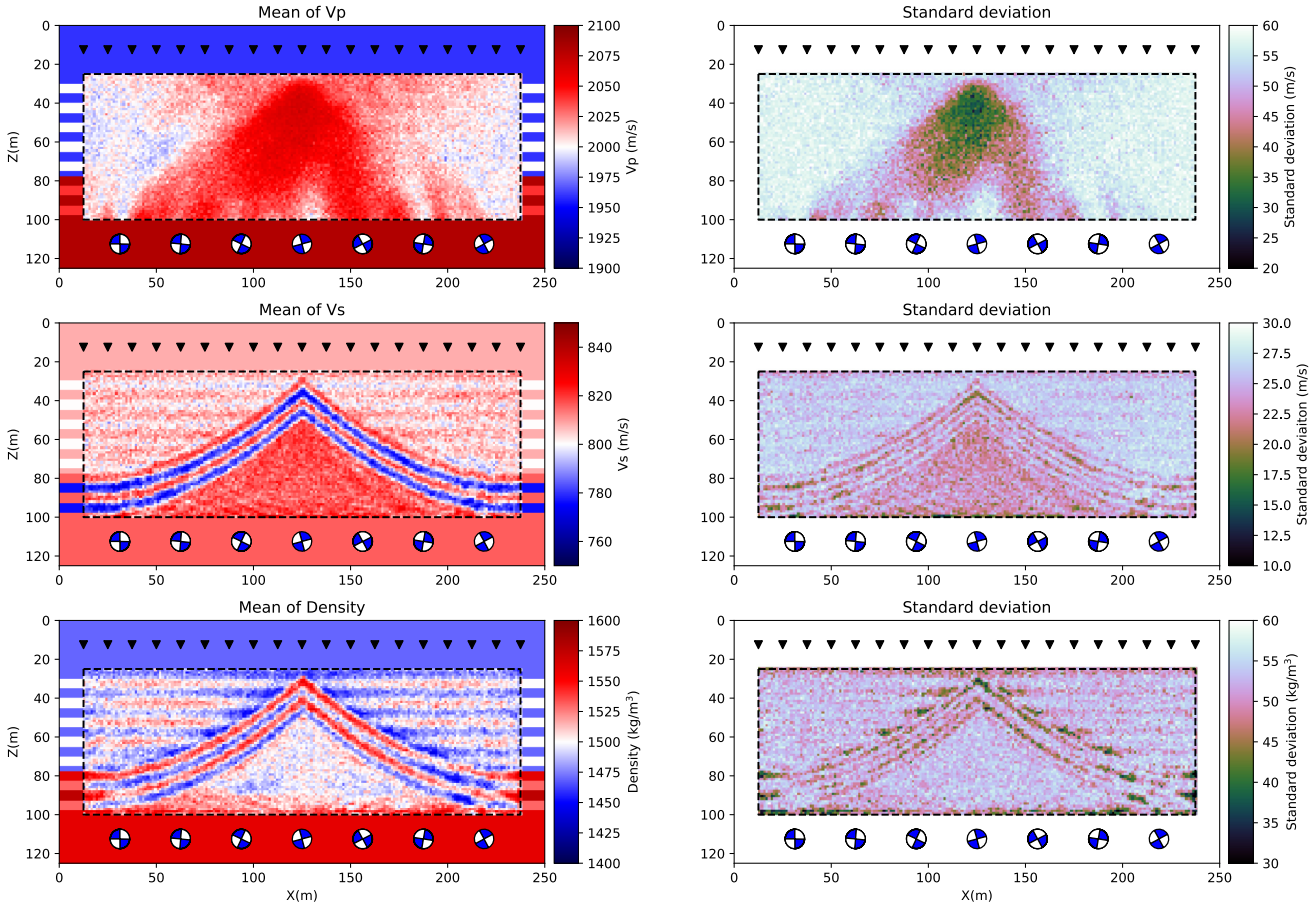

Figure 3. The mean (left) and standard deviation (right) for Vp, Vs and density obtained by Gebraad *et al.* (2020) using Hamiltonian Monte Carlo.

in the interval of 2000 ± 100 m s^{-1} for Vp, 800 ± 50 m s^{-1} for Vs and 1500 ± 100 kg m^{-3} for density. For the noise level, we use a fixed data variance of 1 μm^2 as this variance produces a more accurate model when using HMC (Gebraad *et al.* 2020). For SVGD, we use 600 particles that are initially generated from the prior probability distribution. The particles are first transformed to an unconstrained space as in Zhang & Curtis (2019) and updated using 600 iterations. The final particles are transformed back to the original space and are used to calculated mean and standard deviations.

Fig. 2 shows the mean and standard deviation models for Vp, Vs and density obtained using SVGD. The mean model of Vs successfully recovers the true model, whereas the mean model of Vp provides a significantly different image to the true model. This is probably because the waveforms are more sensitive to Vs than to Vp, so that large-scale structure of Vp can be recovered. The mean model of density shows similar features to the true model near discontinuities, which is likely because waveforms are primarily sensitive to density gradients. In comparison, the bottom high density structure is not present in the result.

The standard deviation of Vs shows similar features to the velocity structure. For example, the horizontal higher velocity layers and the bottom high-velocity structure have smaller standard deviations. There are higher standard deviations at the boundary of tilted layers which have been observed previously in traveltime tomography (Galetti *et al.* 2015; Zhang *et al.* 2018; Zhang & Curtis 2019). This suggests that the location of velocity boundaries are not well constrained. The standard deviation of Vp shows similar features to the mean model, for example, high velocities are associated with lower standard deviations. Similar to the results of shear velocity, the standard deviations of density are lower at the horizontal lower density layers and the boundary of the tilted layers has higher standard deviations. Due to the fact that waveforms are more sensitive to density gradients, the bottom constant higher density structure is not well constrained and has higher standard deviations.

To validate the method we compare the results with those obtained using HMC (Fig. 3) by Gebraad *et al.* (2020). Overall the results from HMC are very similar to those obtained using SVGD except for slightly different magnitudes. Since the same solution is found by completely different methods, it is likely to be the true solution to the full-waveform Bayesian inference problem. Note that the results from SVGD appear to be smoother than those from HMC, which is probably caused by undersampling and lack of convergence of HMC as noted by Gebraad *et al.* (2020).

4 DISCUSSION

We first compare the computational cost of the two methods. SVGD involves $600 \times 600 = 360\,000$ forward and adjoint simulations, whereas HMC involves approximately 130 000 forward and adjoint simulations. While in this case it thus appears that HMC is more efficient than SVGD, in the above example HMC was conducted using only one chain which had not fully converged (Gebraad *et al.* 2020). Since in practice multiple chains are usually required to produce an accurate result, HMC may need more computational cost. Also, in

contrast to HMC, the simulations in SVGD can easily be parallelized which could make the method more efficient in real time (Zhang & Curtis 2019). A Markov chain cannot be easily parallelized due to dependence between successive Markov samples (Neiswanger *et al.* 2013). In practice, HMC often requires deliberate and tedious tuning to achieve an efficient Markov chain (e.g. see discussions in Gebraad *et al.* 2020) so HMC may incur a significantly higher computational cost than that reported above, whereas SVGD requires less effort to achieve an efficient algorithm by using available optimization techniques, for example, ADAGRAD (Duchi *et al.* 2011; Liu & Wang 2016). Note that instead of tuning HMC manually some adaptive methods may also be used (Hoffman & Gelman 2014). To give an overall idea about the computational cost of SVGD, the above example takes about 6 d parallelized using 16 CPU cores.

Although in this study we applied the method to a simple 2-D example with only seven sources, the method can possibly be applied to larger data sets and to 3-D applications by using stochastic optimization (Robbins & Monro 1951; Kubrusly & Gravier 1973) and distributed optimization by dividing large data sets into random minibatches. In comparison the same technique cannot easily be applied to McMC methods since it breaks the reversibility property of Markov chains which is required by most Monte Carlo methods. It is nevertheless not entirely obvious how SVGD will perform on real 3-D applications, so further work is required to compare the efficiency of the methods in a range of practical applications.

In this study, we used a simple L_2 misfit function which may cause multimodality in the likelihood function. Although SVGD can approximate arbitrary probability distributions, the absence of local minima may improve the efficiency of convergence and require fewer particles. Therefore in practice other misfit functions that measure similarity of waveforms may be used to reduce multimodality—assuming that the definitions of those misfit functions are derived in a Bayesian formulation from the forward function and noise statistics (Luo & Schuster 1991; Fichtner *et al.* 2008; Brossier *et al.* 2010; Van Leeuwen & Mulder 2010; Bozdağ *et al.* 2011; Métivier *et al.* 2016). In the example we used a fixed noise level from Gebraad *et al.* (2020). In practice, the noise level may be estimated from the data (Sambridge 2014; Ray *et al.* 2016) or estimated in the inversion in a hierarchical way (Malinverno & Briggs 2004; Bodin *et al.* 2012; Ranganath *et al.* 2016; Zhang *et al.* 2018, 2019).

5 CONCLUSION

In this study, we introduced a new method called VFWI that uses SVGD to solve FWI problems and provide accurate uncertainty estimation. We applied the method to a 2-D synthetic example and compared the results with those obtained using HMC. The results show that SVGD can produce accurate approximations to the probabilistic results obtained by HMC. Although in the simple 2-D example SVGD is less efficient than HMC, the method can easily be parallelized and applied to larger data sets by taking advantage of methods like stochastic optimization and distributed optimization. This can make the method more efficient in practice, allowing it to be applied to larger dats sets and 3-D applications.

ACKNOWLEDGEMENTS

The authors thank the Edinburgh Interferometry Project sponsors (Schlumberger, Equinor and Total) for supporting this research. This work has made use of the resources provided by the Edinburgh Compute and Data Facility (ECDF) (http://www.ecdf.ed.ac.uk/).

REFERENCES

Bishop, C.M., 2006. *Pattern Recognition and Machine Learning,* springer.

Biswas, R. & Sen, M., 2017. 2D full-waveform inversion and uncertainty estimation using the reversible jump Hamiltonian Monte Carlo, in *SEG Technical Program Expanded Abstracts 2017,* pp. 1280–1285, Society of Exploration Geophysicists.

Blei, D.M., Kucukelbir, A. & McAuliffe, J.D., 2017. Variational inference: A review for statisticians, *J. Am. Stat. Assoc.,* **112**(518), 859–877.

Bodin, T. & Sambridge, M., 2009. Seismic tomography with the reversible jump algorithm, *Geophys. J. Int.,* **178**(3), 1411–1436.

Bodin, T., Sambridge, M., Tkalčić, H., Arroucau, P., Gallagher, K. & Rawlinson, N., 2012. Transdimensional inversion of receiver functions and surface wave dispersion, *J. geophys. Res.,* **117**(B2), doi:10.1029/2011JB008560.

Bozdağ, E., Trampert, J. & Tromp, J., 2011. Misfit functions for full waveform inversion based on instantaneous phase and envelope measurements, *Geophys. J. Int.,* **185**(2), 845–870.

Bozdağ, E., Peter, D., Lefebvre, M., Komatitsch, D., Tromp, J., Hill, J., Podhorszki, N. & Pugmire, D., 2016. Global adjoint tomography: first-generation model, *Geophys. J. Int.,* **207**(3), 1739–1766.

Brooks, S., Gelman, A., Jones, G. & Meng, X.-L., 2011. *Handbook of Markov chain Monte Carlo,* CRC Press.

Brossier, R., Operto, S. & Virieux, J., 2010. Which data residual norm for robust elastic frequency-domain full waveform inversion?, *Geophysics,* **75**(3), R37–R46.

Chen, P., Zhao, L. & Jordan, T.H., 2007. Full 3D tomography for the crustal structure of the Los Angeles region, *Bull. seism. Soc. Am.,* **97**(4), 1094–1120.

Curtis, A. & Lomax, A., 2001. Prior information, sampling distributions, and the curse of dimensionality, *Geophysics,* **66**(2), 372–378.

Duchi, J., Hazan, E. & Singer, Y., 2011. Adaptive subgradient methods for online learning and stochastic optimization, *J. Mach. Learning Res.,* **12,** 2121–2159.

Fichtner, A., Bunge, H.-P. & Igel, H., 2006. The adjoint method in seismology: I. theory, *Phys. Earth planet. Inter.,* **157**(1–2), 86–104.

Fichtner, A., Kennett, B.L., Igel, H. & Bunge, H.-P., 2008. Theoretical background for continental-and global-scale full-waveform inversion in the time–frequency domain, *Geophys. J. Int.,* **175**(2), 665–685.

Fichtner, A., Kennett, B.L., Igel, H. & Bunge, H.-P., 2009. Full seismic waveform tomography for upper-mantle structure in the Australasian region using adjoint methods, *Geophys. J. Int.,* **179**(3), 1703–1725.

Fichtner, A. *et al.*, 2018. The collaborative seismic earth model: generation 1, *Geophys. Res. Lett.,* **45**(9), 4007–4016.

French, S. & Romanowicz, B., 2014. Whole-mantle radially anisotropic shear velocity structure from spectral-element waveform tomography, *Geophys. J. Int.,* **199**(3), 1303–1327.

Galetti, E., Curtis, A., Meles, G.A. & Baptie, B., 2015. Uncertainty loops in travel-time tomography from nonlinear wave physics, *Phys. Res. Lett.,* **114**(14), 148501, doi:10.1103/PhysRevLett.114.148501.

Gauthier, O., Virieux, J. & Tarantola, A., 1986. Two-dimensional nonlinear inversion of seismic waveforms: Numerical results, *Geophysics,* **51**(7), 1387–1403.

Gebraad, L., Boehm, C. & Fichtner, A., 2020. Bayesian elastic full-waveform inversion using Hamiltonian Monte Carlo, *J. geophys. Res.,* **125**(3), e2019JB018428.

Hoffman, M.D. & Gelman, A., 2014. The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo., *J. Mach. Learning Res.,* **15**(1), 1593–1623.

Kubrusly, C. & Gravier, J., 1973. Stochastic approximation algorithms and applications, in *1973 IEEE Conference on Decision and Control Including the 12th Symposium on Adaptive Processes,* pp. 763–766, IEEE, San Diego, CA.

Kullback, S. & Leibler, R.A., 1951. On information and sufficiency, *Ann. Math. Stat.,* **22**(1), 79–86.

Liu, Q. & Tromp, J., 2006. Finite-frequency kernels based on adjoint methods, *Bull. seism. Soc. Am.,* **96**(6), 2383–2397.

Liu, Q. & Wang, D., 2016. Stein variational gradient descent: a general purpose Bayesian inference algorithm, in *Advances in Neural Information Processing Systems,* pp. 2378–2386, eds Lee, D.D., Sugiyama, M., Luxburg, U.V., Guyon, I. & Garnett, R., Curran Associates, Inc..

Luo, Y. & Schuster, G.T., 1991. Wave-equation traveltime inversion, *Geophysics,* **56**(5), 645–653.

Malinverno, A. & Briggs, V.A., 2004. Expanded uncertainty quantification in inverse problems: Hierarchical Byes and empirical Byes, *Geophysics,* **69**(4), 1005–1016.

Métivier, L., Brossier, R., Mérigot, Q., Oudet, E. & Virieux, J., 2016. Measuring the misfit between seismograms using an optimal transport distance: application to full waveform inversion, *Geophys. J. Int.,* **205**(1), 345–377.

Nawaz, M. & Curtis, A., 2019. Rapid discriminative variational Bayesian inversion of geophysical data for the spatial distribution of geological properties, *J. geophys. Res.,* **124,** 5867–5887.

Nawaz, M.A. & Curtis, A., 2018. Variational Bayesian inversion (VBI) of quasi-localized seismic attributes for the spatial distribution of geological facies, *Geophys. J. Int.,* **214**(2), 845–875.

Nawaz, M.A., Curtis, A., Shahraeeni, M.S. & Gerea, C., 2020. Variational Bayesian inversion of seismic attributes jointly for geological facies and petrophysical rock properties, *Geophysics,* 1–78.

Neiswanger, W., Wang, C. & Xing, E., 2013. Asymptotically exact, embarrassingly parallel MCMC, arXiv:1311.4780.

Plessix, R.-E., 2006. A review of the adjoint-state method for computing the gradient of a functional with geophysical applications, *Geophys. J. Int.,* **167**(2), 495–503.

Pratt, R.G., 1999. Seismic waveform inversion in the frequency domain, part 1: Theory and verification in a physical scale model, *Geophysics,* **64**(3), 888–901.

Prieux, V., Brossier, R., Operto, S. & Virieux, J., 2013. Multiparameter full waveform inversion of multicomponent ocean-bottom-cable data from the Valhall field. Part 1: Imaging compressional wave speed, density and attenuation, *Geophys. J. Int.,* **194**(3), 1640–1664.

Ranganath, R., Tran, D. & Blei, D., 2016. Hierarchical variational models, in *International Conference on Machine Learning,* eds Balcan , M.F. & QWeinberger, K., Proceedings of Machine Learning Research, pp. 324–333.

Ray, A., Sekar, A., Hoversten, G.M. & Albertin, U., 2016. Frequency domain full waveform elastic inversion of marine seismic data from the Alba field using a Bayesian trans-dimensional algorithm, *Geophys. J. Int.,* **205**(2), 915–937.

Ray, A., Kaplan, S., Washbourne, J. & Albertin, U., 2017. Low frequency full waveform seismic inversion within a tree based Bayesian framework, *Geophys. J. Int.,* **212**(1), 522–542.

Robbins, H. & Monro, S., 1951. A stochastic approximation method, *Ann. Math. Stat.,* **22**(3), 400–407.

Sambridge, M., 2014. A parallel tempering algorithm for probabilistic sampling and multimodal optimization, *Geophys. J. Int.* **196,** 357–374.

Tape, C., Liu, Q., Maggi, A. & Tromp, J., 2009. Adjoint tomography of the southern California crust, *Science,* **325**(5943), 988–992.

Tarantola, A., 1984. Inversion of seismic reflection data in the acoustic approximation, *Geophysics,* **49**(8), 1259–1266.

Tarantola, A., 1988. Theoretical background for the inversion of seismic waveforms, including elasticity and attenuation, in *Scattering and Attenuations of Seismic Waves, Part I,* eds Aki , K. & Wu , R.S., Springer, pp. 365–399, .

Tromp, J., Tape, C. & Liu, Q., 2005. Seismic tomography, adjoint methods, time reversal and banana-doughnut kernels, *Geophys. J. Int.,* **160**(1), 195–216.

Van Leeuwen, T. & Mulder, W., 2010. A correlation-based misfit criterion for wave-equation traveltime tomography, *Geophys. J. Int.,* **182**(3), 1383–1394.

Virieux, J., 1986. P-SV wave propagation in heterogeneous media: Velocity-stress finite-difference method, *Geophysics,* **51**(4), 889–901.

Warner, M. *et al.*, 2013. Anisotropic 3D full-waveform inversion, *Geophysics,* **78**(2), R59–R80.

Zhang, X. & Curtis, A., 2020. Seismic tomography using variational inference methods, *J. geophys. Res.,* **125,** e2019JB018589., doi:10.1029/2019JB018589.

Zhang, X., Curtis, A., Galetti, E. & de Ridder, S., 2018. 3-D Monte Carlo surface wave tomography, *Geophys. J. Int.,* **215**(3), 1644–1658.

Zhang, X., Hansteen, F., Curtis, A. & de Ridder, S., 2020. 1D, 2D and 3D Monte Carlo ambient noise tomography using a dense passive seismic array installed on the North Sea seabed, *J. geophys. Res.,* **125,** e2019JB018552. , doi:10.1029/2019JB018552.