Evolutionary full-waveform inversion

Dirk Philip van Herwaarden, Michael Afanasiev, Solvi Thrastarson and Andreas Fichtner

Department of Earth Sciences, Institute of Geophysics, ETH Zurich, 8092 Zurich, Switzerland. E-mail: dirkphilip.vanherwaarden@erdw.ethz.ch

Received 2020 September 8; in original form 2020 July 8

SUMMARY
We present a new approach to full-waveform inversion (FWI) that enables the assimilation of data sets that expand over time without the need to reinvert all data. This evolutionary inversion rests on a reinterpretation of stochastic Limited-memory Broyden–Fletcher–Goldfarb–Shanno (L-BFGS), which randomly exploits redundancies to achieve convergence without ever considering the data set as a whole. Specifically for seismological applications, we consider a dynamic mini-batch stochastic L-BFGS, where the size of mini-batches adapts to the number of sources needed to approximate the complete gradient. As an illustration we present an evolutionary FWI for upper-mantle structure beneath Africa. Starting from a 1-D model and data recorded until 1995, we sequentially add contemporary data into an ongoing inversion, showing how (i) new events can be added without compromising convergence, (ii) a consistent measure of misfit can be maintained and (iii) the model evolves over times as a function of data coverage. Though applied retrospectively in this example, our method constitutes a possible approach to the continuous assimilation of seismic data volumes that often tend to grow exponentially.

Key words: Inverse theory; Waveform inversion; Computational seismology; Seismic tomography.

1 INTRODUCTION
Seismic tomography has come a long way since its early applications (e.g. Aki & Lee 1976; Dziewonski et al. 1977). Increasing computational resources enable increasingly sophisticated methods, for instance, finite-frequency tomography (e.g. Dahlen et al. 2000; Friederich 2003; Yoshizawa & Kennett 2005), joint inversions of multiple wave types (e.g. Ritsema et al. 1999; Chang et al. 2010; Koelemeijer et al. 2017) or Bayesian approaches (e.g. Trampert et al. 2004; Bodin & Sambridge 2009; Mosca et al. 2012; Gebraad et al. 2020). Recently, it has become increasingly feasible to perform 3-D full-waveform inversion (FWI) at various scales (e.g. Chen et al. 2007; Tape et al. 2010; French & Romanowicz 2014; Bozdag et al. 2016; Simute et al. 2016; Fichtner et al. 2018).

While FWI may eliminate most simplifications of wave physics and thereby improve resolution, it also comes with significant computational cost because each iterative model update requires forward and adjoint simulations, at a cost scaling with frequency to the power of 4 (Virieux & Operto 2009; Piri 2017). Additionally, Fresnel zones of higher-frequency waves are narrower, and as a result more data are needed to constrain the earth model sufficiently.

As increasing computational power allows us to invert higher-frequency data, also the number of data increases roughly exponentially. For example, the IRIS data archive grows ∼25 per cent per year (http://ds.iris.edu/data/distribution/). Hence, earth models should ideally update regularly to incorporate this new wealth of information. For ray tomography, such evolutionary approaches were pioneered by Debayle et al. (2016) at global scale and Tong et al. (2017) at regional scale. These authors expand their data set sequentially and then reinvert the larger data set to obtain an update. While evolutionary inversion would also be desirable for FWI, periodic reinversions of a growing data set are computationally out of scale.

Here we propose a strategy to efficiently incorporate information from new seismic recordings into an evolving FWI model. This is based on a stochastic L-BFGS method where newly acquired data contribute to quasi-random mini-batches, driving the next update. We demonstrate this strategy with an example, where we start from a 1-D model and a small data set from the early 1990s, and show how our approach can be used to dynamically expand this data set as new recordings become available. The result is an evolving and up-to-date model of the Earth which retains previous information while consistently incorporating new data.

2 METHODOLOGY
Primarily developed for machine learning applications where data sets are too large to be assimilated all at once (e.g. Bottou 2010; Byrd et al. 2011), stochastic gradient descent (SGD) has recently been adapted for seismic data inversion with the main goal to reduce computational cost (e.g. van Leeuwen & Herrmann 2013). Stochastic
L-BFGS is an extension of SGD which utilizes curvature information and has been applied in medical imaging (Boehm et al. 2018). In the following paragraphs we demonstrate how stochastic L-BFGS can be reinterpreted as a tool that enables evolutionary model updating. Because of the unique challenges encountered with regional-to-continental scale data sets, such as irregular data coverage, we will focus on the dynamic mini-batch stochastic L-BFGS proposed by van Herwaarden et al. (2020), which is specifically tuned to harness redundancies in seismological data sets, and which we briefly recapitulate for completeness in the following paragraphs.

2.1 Stochastic L-BFGS for evolutionary model construction

In most FWI approaches, a master data set of events is carefully curated and kept constant over the course of an inversion. In contrast, dynamic mini-batch FWI quasi-randomly selects for each iteration, a smaller subset (a mini-batch) of events which contribute significant information to the gradient direction. This is illustrated schematically in Fig. 1. To maximize coverage, the first mini-batch is constructed with Mitchell’s best-candidate algorithm (Mitchell 1991), choosing the next event to be furthest from the previously selected ones.

To consistently monitor waveform misfit over the course of an inversion where the mini-batch composition changes continuously, a control group subset is carried over to the next iteration. Control group events are selected among the mini-batch events based on the requirement that the angle between the control group and the mini-batch gradient, defined through the scalar product of the two, must be below a defined threshold, empirically set to $22.5^\circ$ (van Herwaarden et al. 2020). The number of control group events needed to meet the threshold may vary, thereby indicating if the mini-batch is too small or too large. Also, the size of the subsequent mini-batch is empirically set to twice the size of the current control group, meaning that the mini-batch size adapts dynamically. For subsequent iterations, events not in the current control group are replaced, again using Mitchell’s best-candidate algorithm to incorporate new data. If no unseen data exists, events are selected quasi-randomly (van Herwaarden et al. 2020), whereby events that had a significant contribution to a past mini-batch’s gradient direction have a higher probability of being selected again. The quasi-random event selection essentially penalizes frequent iterations with events that have either comparably few recordings or have converged already.

Once a mini-batch and a control group have been selected, the model updating proceeds with the help of a trust-region L-BFGS method, as illustrated in the lower panel of Fig. 1. The original L-BFGS method (Nocedal 1980; Nocedal & Wright 1999) approximates the inverse Hessian for the complete data set in each iteration using the gradients of the previous $\ell$ iterations, where $\ell$ is typically below 10. The inverse Hessian approximation is then used for a quasi-Newton iteration. While the original L-BFGS sequence of gradients is deterministic, that is, fully controlled by the initial model, the dynamic mini-batch approach produces a random sequence of gradients for the approximation of the inverse Hessian. This is because quasi-random subsets of the complete data set are used to compute gradient approximations.

Although the dynamic mini-batch approach was originally designed as a method to improve computational efficiency, it also allows us to implement evolutionary FWI. In fact, within the dynamic mini-batch framework there is no difference between an event held out of previous mini-batch iterations and an event which had not yet been included in the master data set. Hence, as data from new events become available, we can simply add the event to the collection of candidate data of which the dynamic mini-batch algorithm is able to sample. By generally using newly added events as candidates for the next mini-batch, the inclusion of previously unseen data is prioritized. In this sense, the quasi-random acquisition of new seismic data, for instance, through the installation of new stations, becomes part of the quasi-random selection of mini-batch for the next iteration.

2.2 Example problem and setup

To illustrate how dynamic mini-batch methods can perform on an expanding data set we use the example of a continental-scale FWI of the African continent, consider a time frame from the early 1990s to the present, and show how the model and data set evolve over time. The initial data set contains 46 events and 1452 unique source-receiver pairs. It results from the selection of events that occurred between 1990 and 1995 and have a magnitude range of 5.5–6.7, which provides useful signal-to-noise ratios while avoiding finite-source effects. Adhering to the same magnitude range, we gradually expand the data set over the course of the iteration to finally include 428 events and 147 903 unique source–receiver pairs for events between 1990 and 2017. The majority of the data has been recorded with three-component seismometers. All of the data have been acquired through the publicly available International Federation of Digital Seismograph Networks (FDSN; Romanowicz & Dziewonski 1986). Both the initial and final data sets are shown in Fig. 2.

We perform forward and adjoint modelling using the spectral-element wave propagation solver Salvus (Afanasiev et al. 2019). The starting model is the transversely isotropic Preliminary Reference Earth Model (PREM) (Dziewonski & Anderson 1981), and topography and ellipticity are accounted for. We apply Gaussian smoothing to the model in the radial direction with a standard deviation of 50 km. This is done for two reasons: (1) We do not wish to incorporate a strong prior on the presence and location of discontinuities and (2) it helps to reduce the computational cost due to a reduction in the number of small crustal elements, without introducing cycle-skipping problems for the long period (65–120 s) waveforms that we will model and invert. While our method is not exclusive to a specific misfit function, the inversion presented herein employs time–frequency phase misfits (Fichtner et al. 2008). The obtained gradients are smoothed to prevent subwavelength structure from entering the model by effectively convolving the gradients with a Gaussian filter with a standard deviation of 200 km.

The stochastic gradient inversion initiates with the selection of 10 out of 46 events for the first mini-batch. We then add data at several stages as indicated by the vertical lines in Fig. 3. The number of source-receiver pairs contributing to the model grows with the amount of available data. Furthermore, as shown by the blue dots in Fig. 3, the composition of the mini-batches changes in each iteration, with priority given to new data that are more likely to bring new independent information. Towards the end of the iteration, the average mini-batch size has grown to around 50 events, indicating that more events are needed to approximate the gradient of the whole data set.

3 RESULTS

Since we have simulated a journey of an evolutionary inversion through time, we show the model at several stages in Fig. 4. Each
Figure 1. Inversion flowchart of the stochastic-gradient mini-batch inversion. In passing from iteration $n$ to $n+1$ the data set may be extended, and a new mini-batch is selected, giving preference to new data. After computing misfits and gradients for the mini-batch, a new control group is selected to monitor misfit reduction of updates. In our applications, presented in the following sections, we empirically chose that mini-batch twice as large as the control group. Other choices are possible, in case they produce satisfactory results.

Figure 2. Data summary in the form of great-circle paths connecting earthquakes (beach balls) and stations (black triangles). Left: The initial data set with events from 1990–1995 contains 1452 unique source–receiver pairs. Right: The final data set with events from 1990–2017 has grown to 147 903 unique source–receiver pairs over the course of the inversion.

Stage has been influenced by a unique number of seismic recordings available until that point in time through a varying number of mini-batch iterations. As the inversion progresses and more data are incorporated, several geologic features, such as the West African Craton and the Cameroon Volcanic Line become more pronounced.

The blue line in Fig. 5 shows the $L_2$ norm difference between $SV$ velocity ($V_{SV}$) in the current and initial models. This difference increases quickly as large-scale velocity structure is recovered during the first few iterations. Afterwards, there are less changes, except around iteration 200, when significantly more stations started to record earthquakes in Africa. The inversion is carried out in a fixed frequency band. This was a deliberate choice, intended to emphasize the impact of data volumes on the resulting model.

Stochastic gradient descent in general, and the dynamic mini-batch approach in particular, are primarily designed to assimilate data sets that are too large to be considered at once. Hence, by the purpose of their existence, they do not permit an analysis of the total misfit of the entire data set. Instead, to assess convergence, we study the behaviour of the control group, that is, the subset of a mini-batch that is kept for the next iteration to enable a direct misfit comparison between iterations. The red line in Fig. 5 shows the relative control group misfit reduction as a function of iteration count. After control group misfit initially decreases by nearly 10 per cent per iteration, it plateaus around 3 per cent. However, when new events are added, control group misfit reductions can be more significant because their waveforms may still need to be matched.
4 DISCUSSION AND CONCLUSIONS

We have shown results of a workflow that enables an evolutionary mode of inversion without reinversion of a complete data set. The workflow enables an integration of new data, thereby adding information to constrain the model. For the purpose of illustration, the example shown in the previous sections is retrospective, successively assimilating data from the early 1990s to 2017. The resulting suite of models provides snapshots of our knowledge of African upper-mantle structure as a function of time and station coverage. Yet, the potential of the method primarily lies in the continued incorporation of future data volumes that tend to grow exponentially.

In principle, it would be possible to expand the data set in other ways too. One could, for example, simply expand the data set and restart from the latest model. While being straightforward, this is unlikely to be efficient because previously matched data would be modelled without contributing significantly to the new search direction. Alternatively, one may perform a few iterations only with the new data and then iterate again with all the previous data. Also this strategy poses its own challenges. How many iterations would need to be performed and which data should be selected?

A major advantage of the stochastic gradient approach is that none of these choices need to be made, and that, by design, a large data set can be assimilated without having to consider the data set as a whole. The stochastic gradient approximation combined with the dynamically chosen mini-batches automatically ‘learns’ which events are relevant and how many are required to make a decent approximation of the larger data set.

The redundancy that is typical for seismological data sets, where events often occur in similar locations, allows the dynamic mini-batch approach to achieve significant computational savings, compared to inversions of a complete data set (up to 80 per cent in the examples shown by van Herwaarden et al. 2020). As more data are incorporated into an inversion, it becomes more likely that the data set will contain redundant information. As a result, it seems plausible that the aforementioned computational savings will increase as the data sets become larger.

As computational power increases, it becomes possible to exploit higher-frequency data. Our workflow in principle allows model updates using events modelled at varying frequencies. In practice, however, this adds additional complexities. Higher frequencies must be modelled on more finely discretized meshes, and otherwise simple operations such as summing eventwise contributions to the total...
gradient become non-trivial (Thrastarson et al. 2020). In addition, higher frequencies are sensitive to smaller-scale structures, thus requiring changes in regularization and parametrization. The exact point at which these changes become important is a problem that still requires further research. In this study, we deliberately chose to keep the regularization and the frequency content constant, in order to isolate and show the effect that an evolving data set has on the model. As such a model continues to incorporate more and higher frequency data, it will be important to change the regularization accordingly.

ACKNOWLEDGEMENTS

We gratefully acknowledge editor Joerg Renner, as well as François Lavoué and an anonymous reviewer for their constructive comments that helped to improve the paper. In addition, we would like to thank Christian Boehm for the helpful discussions and advice. This work was supported by the European Research Council (ERC) under the EU’s Horizon 2020 Framework Programme (grant no. 714069). We also gratefully acknowledge support from the Swiss National Supercomputing Centre (CSCS) in the form of computing time grants c13 and s868. In addition, we would like to thank IRIS DMC, whose FDSN webservices provided the majority of the data.

REFERENCES

Afanasiev, M., Boehm, C., van Driel, M., Krischer, L., Rietmann, M., May, D. A., Knepley, M. G. & Fichtner, A., 2019. Modular and flexible spectral-element waveform modelling in two and three dimensions, *Geophys. J. Int.*, 216(3), 1675–1692.

Aki, K. & Lee, W.H.K., 1976. Determination of three-dimensional velocity anomalies under a seismic array using first P arrival times from local earthquakes: 1. A homogeneous initial model, *J. geophys. Res.*, 81(23), 4381–4399.

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

Boehm, C., Martiuart, N.K., Vinard, N., Balic, I.J. & Fichtner, A., 2018. Time-domain spectral-element ultrasound waveform tomography using a stochastic quasi-Newton method, *Proc. SPIE*, 10580, 105800H.

Bottou, L., 2010. Large-Scale Machine Learning with Stochastic Gradient Descent, in *Proceedings of COMPSTAT'2010*, pp. 177–186, Physica-Verlag HD, Heidelberg.

Bozdog, 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), 1759–1766.

Byrd, R., Chin, G., Neveitt, W. & Nocedal, J., 2011. On the use of stochastic Hessian information in optimization methods for machine learning, *SIAM J. Optim.*, 21(3), 977–995.

Chang, S.-J. et al., 2010. Joint inversion for three-dimensional S-velocity mantle structure along the Tethyan margin, *J. geophys. Res.*, 115, doi:10.1029/2009JB007204.

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, 1094–1120.

Dahlen, F.A., Hung, S.-H. & Nolet, G., 2000. Fréchet kernels for finite-frequency traveltimes—I. Theory, *Geophys. J. Int.*, 141(1), 157–174.

Debayle, E., Dubuffet, F. & Durand, S., 2016. An automatically updated S-wave model of the upper mantle and the depth extent of azimuthal anisotropy, *Geophys. Res. Lett.*, 43, 674–682.

Dziewonski, A.M. & Anderson, D.L., 1981. Preliminary reference Earth model, *Phys. Earth Planet. Inter.*, 25, 297–356.

Dziewonski, A.M., Hager, B.H. & O’Connell, R.J., 1977. Large-scale heterogeneities in the lower mantle, *J. geophys. Res.*, 82(2), 239–255.

Fichtner, A., Kennett, B. L. N., 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, 665–685.

Fichtner, A. et al., 2018. The Collaborative Seismic Earth Model: Generation 1, *Geophys. Res. Lett.*, 45(9), 4007–4016.

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

Friederich, W., 2003. The S-velocity structure of the East Asian mantle from inversion of shear and surface waveforms, *Geophys. J. Int.*, 153, 88–102.

Gebrad, L., Boehm, C. & Fichtner, A., 2020. Bayesian elastic full-waveform inversion using hamiltonian monte carlo, *J. geophys. Int.*, 125(3), e2019JB018428 , doi:10.1029/2019JB018428.

Kooi, P., Deuss, A. & Ritsema, J., 2017. Density structure of Earth’s lowermost mantle from Stony Lee model splitting observations, *Nat. Commun.*, 8, 15241, doi:10.1038/ncomms15241.

Mitchell, D.P., 1991. Spectrally optimal sampling for distribution ray tracing, *SIGGRAPH Comput. Graph.*, 25(4), 157–164.

Mosca, I., Cobden, L., Deuss, A., Ritsema, J. & Trampert, J., 2012. Seismic and mineralogical structures of the lower mantle from probabilitistic tomography, *J. geophys. Res.*, 117, doi:10.1029/2011JB008851.

Nocedal, J., 1980. Updating quasi-Newton matrices with limited storage, *Math. Comp.*, 35, 773–782.

Nocedal, J. & Wright, S., 1999. *Numerical Optimization*, Springer, New York.

Pirli, M., 2017. Heiner Igel: computational seismology: a practical introduction, *J. Seismol.*, 21(3), 567–570.

Ritsema, J., Heijst, H.J.V. & Woodhouse, J.H., 1999. Complex shear wave velocity structure imaged beneath Africa and Iceland, *Science*, 286(5446), 1925–1928.

Romanowicz, B.A. & Dziewonski, A.M., 1986. Toward a federation of broadband seismic networks, *EOS, Trans. Am. geophys. Un.*, 67(25), 541–542.

Simute, S., Steptoe, H., Cobden, L., Goldberg, A. & Fichtner, A., 2016. Full-waveform inversion of the Japanese islands region, *J. geophys. Res.*, 121(5), 3722–3741.

Tape, C., Liu, Q., Maggi, A. & Tromp, J., 2010. Seismic tomography of the southern California crust based upon spectral-element and adjoint methods, *Geophys. J. Int.*, 180, 433–462.

Thrastarson, S., van Driel, M., Krischer, L., Boehm, C., Afanasiev, M., van Herwaarden, D.-P. & Fichtner, A., 2020. Accelerating numerical wave
propagation by wavefield adapted meshes. Part II: full-waveform inversion, *Geophys. J. Int.*, 221(3), 1591–1604.

Tong, P., Yang, D., Li, D. & Liu, Q., 2017. Time-evolving seismic tomography: The method and its application to the 1989 Loma Prieta and 2014 South Napa earthquake area, California, *Geophys. Res. Lett.*, 44, 3165–3175.

Trampert, J., Deschamps, F., Resovsky, J. & Yuen, D., 2004. Probabilistic tomography maps chemical heterogeneities throughout the lower mantle, *Science*, 306, 853–856.

van Herwaarden, D.-P., Boehm, C., Afanasiev, M., Thrastarson, S., Krischer, L., Trampert, J. & Fichtner, A., 2020. Accelerating full-waveform inversion using dynamic mini-batches, *Geophys. J. Int.*, 221, 1427–1438.

van Leeuwen, T. & Herrmann, F.J., 2013. Fast waveform inversion without source-encoding, *Geophys. Prospect.*, 61(s1), 10–19.

Virieux, J. & Operto, S., 2009. An overview of full-waveform inversion in exploration geophysics, *Geophysics*, 74(6), WCC1–WCC26.

Yoshizawa, K. & Kennett, B.L.N., 2005. Sensitivity kernels for finite-frequency surface waves, *Geophys. J. Int.*, 162(3), 910–926.