Received: 14 March 2023 | Accepted: 5 July 2023

DOI: 10.1111/1365-2478.13405

Geophysical Prospecting EAGE EUROPEAN ASSOCIATION OF GEOSCIENTISTS & ENGINEERS

ORIGINAL ARTICLE

Wave-based inversion at scale on graphical processing units with randomized trace estimation

Mathias Louboutin[1,*] | **Felix J. Herrmann**[1,2,*]

[1]School of Earth and Atmospheric Sciences, Georgia Institute of Technology, Atlanta, Georgia, USA

[2]School of Computational Science and Engineering, Georgia Institute of Technology, Atlanta, Georgia, USA

Correspondence
Mathias Louboutin, School of Earth and Atmospheric Sciences, Georgia Institute of Technology, 756 W Peachtree St NW, Atlanta, GA 30332, USA.
Email: mathias.louboutin@gmail.com

Abstract

Thanks to continued performance improvements in software and hardware, wave-equation-based imaging technologies, such as full-waveform inversion and reverse-time migration, are becoming more commonplace. However, widespread adaptation of these advanced imaging modalities has not yet materialized because current implementations are not able to reap the full benefits from accelerators, in particular those offered by memory-scarce graphics processing units. Through the use of randomized trace estimation, we overcome the memory bottleneck of this type of hardware. At the cost of limited computational overhead and controllable incoherent errors in the gradient, the memory footprint of adjoint-state methods is reduced drastically. Thanks to this relatively simple to implement memory reduction via an approximate imaging condition, we are able to benefit from graphics processing units without memory offloading. We demonstrate the performance of the proposed algorithm on acoustic two- and three-dimensional full-waveform inversion examples and on the formation of image gathers in transverse tilted isotropic media.

KEYWORDS

full-waveform inversion, finite differences, inversion, reverse-time migration, stochastic

INTRODUCTION

With the advance of high-performance computing, wave-equation-based inversions such as full-waveform inversion (FWI) and reverse-time migration (Lions, 1971; Tarantola, 1984; Virieux & Operto, 2009) have become pivotal research topics with academic and industrial applications. While powerful, these wave-based inversion methods come at high computational and memory costs, which explains their relatively limited application to real-world problems. The fundamental limitation of wave-equation-based inversion lies in the excessive memory footprint of the time-domain adjoint-state method (Lions, 1971; Tarantola, 1984), which requires access to the complete time history of the forward modelled wavefield when applying the imaging condition. In its simplest form, this imaging condition entails, for each source experiment, on-the-fly accumulation of a spatial cross correlation between the (stored) forward wavefield and time snapshots of solutions of the adjoint wave equation as they become available. Because three-dimensional (3D) forward modelled wavefields require terabytes of storage, memory usage and input/output (I/O) bandwidth demand continue to be major bottlenecks. While dedicated high-memory hardware may address this issue, it precludes the use of modern accelerators (e.g., graphical processing units (GPUs)), which generally do not have access to large amounts of low-latency memory.

To tackle the high-memory requirements of the adjoint-state method, several solutions have been proposed. Griewank and Walther (2000) and Symes (2007) presented optimal checkpointing, which avoids excessive memory usage by balancing I/O and computational overhead optimally. This approach, which was initially developed for generic

*Equally contributing authors.

adjoint-state methods on CPUs (Griewank & Walther, 2000), has been used successfully in seismic imaging (Symes, 2007) and machine learning (Chen et al., 2016) and has recently been extended to multi-stage timestepping (Zhang & Constantinescu, 2022). By adding on-the-fly compression and decompression of the checkpointed forward wavefields, Kukreja et al. (2020) further reduced the computational overhead of optimal checkpointing. Instead of checkpointing the forward wavefield, McMechan (1983), Mittet (1994), and Raknes and Weibull (2016) rely on time reversibility to reconstruct forward wavefields during backpropagation from values stored on the boundary. Unfortunately, this type of wavefield reconstruction is only stable for attenuation-free wave equations, which limits its applicability. Finally, optimized implementations of accelerators of these approaches quickly become involved, especially in situations where the wave physics becomes more complex, for example, when dealing with elastic or tilted transversely isotropic media. This added complexity explains the lack of native implementations of the time-domain adjoint-state method on accelerators including GPUs.

While recomputing or decompressing forward wavefields as part of memory-footprint mitigation certainly has its merits, we put forward algorithmically much simpler randomized approaches where memory use and accuracy are traded against computational overhead. Unlike lossless approaches, which aim to compute gradients exactly, we propose to approximate gradients with randomized estimates that balance computational gains and loss of accuracy. Examples of trading computational cost and accuracy include working with random subsets of shots (Friedlander & Schmidt, 2012), with simultaneous shots (Haber et al., 2015; Krebs et al., 2009; van Leeuwen & Herrmann, 2013; Moghaddam et al., 2013; Romero et al., 2000), or with randomized singular value decompositions (van Leeuwen et al., 2017; Yang et al., 2021) and trace estimation (Halko et al., 2011). The latter random trace estimation technique was used by Haber et al. (2015) to analyse computational speedups of FWI with computational simultaneous sources. As long as errors are controlled (Friedlander & Schmidt, 2012; van Leeuwen and Herrmann, 2014), these approximate methods all lead to inversion results that in stochastic expectation are equivalent to the original problem but at a fraction of the computational costs.

Inspired by these contributions, and ideas from randomized trace estimation, we propose an approximate adjoint-state method that leads to major memory improvements (Louboutin & Herrmann, 2021, 2022) is unbiased, relatively easy to implement, and supported by rigorous theory (Avron & Toledo, 2011; Meyer et al., 2020). Unlike methods based on lossy compression (Kukreja et al., 2020) or on the on-the-fly Fourier transform, artefacts introduced by our proposed randomized trace estimation appear as incoherent Gaussian-like noise, which can be handled easily by stacking, sparsity promotion (Witte, Louboutin, Luporini, et al., 2019) or constrained optimization (Peters et al., 2018; Peters & Herrmann, 2019). Below, we will support this claim empirically by means of carefully selected seismic inversion examples.

Our paper is organized as follows. First, we introduce the method of randomized-trace estimation and derive how computing gradients with the adjoint-state method can be recast in terms of trace estimation. We show that random trace estimates allow for approximations with a low memory footprint and low computational overhead. Next, we describe how to increase the accuracy of randomized-trace estimation with data-informed probing vectors. After comparing the computational costs of our method with traditional memory-saving approaches, we show how our method leads to significant cost reductions when computing image volumes and complex imaging conditions. Performance of our method on two realistic seismic inversion problems will be demonstrated. We conclude by showcasing a 3D FWI example produced with a purely GPU-native implementation.

ADJOINT-STATE METHOD WITH RANDOMIZED PROBING

To arrive at our low-memory wave-equation-based inversion formulation, we first describe the main theoretical features of randomized-trace estimation. Next, we show how randomized-trace estimation can be used to reduce the memory footprint of time-domain gradient (isotropic an anisotropic) and subsurface-offset image volume calculations. For comparison with existing state-of-the-art memory reduction approaches, we will also derive estimates for memory use and computational overhead.

Randomized-trace estimation

With the increasing demand for large-scale data-driven applications, randomized algorithms have steadily gained popularity especially in situations where memory access comes at a premium and where access to compute cycles is relatively abundant. Unlike conventional techniques in linear algebra, which aim to carry out accurate calculations at the price of high-memory pressure, randomized algorithms (Halko et al., 2011; Yang et al., 2021) limit their memory footprint at the cost of a controllable error. The technique we consider relies on an unbiased estimator based on a randomized probing technique that yields estimates for the trace (sum of the diagonal elements) of a matrix with errors that average to zero in stochastic expectation. Instead of forming the matrix explicitly, randomize-trace estimation (Avron & Toledo, 2011; Meyer et al., 2020) relies on matrix-free actions on random probing vectors. As long as these matrix–vector products are available and cheap, the trace can be estimated even for matrices that are too large to fit into memory. Contrary to earlier

work where random-trace estimation was used to reduce the number of wave-equation solves (Haber et al., 2015), we use randomized-trace estimation to reduce the memory cost of computing gradients of wave-equation themselves.

At its heart, randomized-trace estimation (Avron & Toledo, 2011; Meyer et al., 2020) derives from an unbiased approximation of the identity, $\mathbf{I} = \mathbb{E}[\mathbf{z}\mathbf{z}^\top]$ – that is, we have

$$\begin{aligned} \mathrm{tr}(\mathbf{A}) &= \mathrm{tr}\big(\mathbf{A}\mathbb{E}\big[\mathbf{z}\mathbf{z}^\top\big]\big) = \mathbb{E}\big[\mathrm{tr}\big(\mathbf{A}\mathbf{z}\mathbf{z}^\top\big)\big] \\ &= \mathbb{E}\big[\mathbf{z}^\top\mathbf{A}\mathbf{z}\big] \\ &\approx \frac{1}{r}\sum_{i=1}^{r}\big[\mathbf{z}_i^\top\mathbf{A}\mathbf{z}_i\big] = \frac{1}{r}\mathrm{tr}\big(\mathbf{Z}^\top\mathbf{A}\mathbf{Z}\big). \end{aligned} \tag{1}$$

In these expressions, the $\mathbf{z}_i$s are the random probing vectors collected as columns in the matrix $\mathbf{Z}$. The operator $\mathbb{E}$ stands for stochastic expectation with respect to these random vectors. By ensuring $\mathbb{E}(\mathbf{z}^\top\mathbf{z}) = 1$, the above trace estimator is unbiased (exact in expectation) and converges to the true trace of the matrix $\mathbf{A}$, that is, $\mathrm{tr}(A) = \sum_i \mathbf{A}_{ii}$, with an error that decays to zero for increasing r. Compared to the original computation of the trace, randomized-trace estimation does not need access to the entries of the matrix $\mathbf{A}$. Only actions of the matrix $\mathbf{A}$ on probing vectors are needed. To improve the computational performance of the estimator, we follow Graff-Kray et al. (2017) and Meyer et al. (2020) and use a partial `qr` factorization (Trefethen & Bau, 1997) to derive the probing vectors collected in the matrix $\big[\mathbf{Q}, \sim\big] = \mathrm{qr}(\mathbf{A}\mathbf{Z})$. These orthogonal probing vectors are computed from the $n \times n$ matrix $\mathbf{A}$ with random probing vectors collected in the tall $n \times r$ (with $r \ll n$) Rademacher matrix $\mathbf{Z}$ with ± 1 entries (1 or -1 with probability 0.5 each). In the ensuing sections, we will exploit this randomized-trace estimation technique to reduce the memory footprint of gradient calculations for the adjoint-state method of wave-equation-based inversion.

Approximate gradient calculations

While this may sound controversial but non-convex optimization problems such as full-waveform inversion (FWI) (Tarantola, 1984; Virieux & Operto, 2009) benefit from stochastic errors in their gradients whether these are due to working with randomized minibatches, as in stochastic gradient descent, a technique widely employed by machine learning, or with sub-samplings in terms of randomized (super)shots as in FWI. In either case, computational costs are reduced and the optimization is less prone to local minima thanks to an annealing effect (Neelakantan et al., 2015). As shown by van Leeuwen and Herrmann (2014), it can also be computationally advantageous to allow for errors in the gradient calculation themselves, which is the approach taken here.

For this purpose, let us consider the standard adjoint-state FWI problem, which aims to minimize the misfit between recorded field data and numerically modelled synthetic data (Lions, 1971; Louboutin et al., 2017; Louboutin, Witte, Lange, et al., 2018; Tarantola, 1984; Virieux & Operto, 2009). In its simplest form, the data misfit objective for a single shot record is given by

$$\underset{\mathbf{m}}{\text{minimize}}\ \frac{1}{2}||\mathbf{F}(\mathbf{m};\mathbf{q}) - \mathbf{d}_{\text{obs}}||_2^2. \tag{2}$$

In this expression, the vector $\mathbf{m}$ represents the unknown physical model parameter (e.g. the squared slowness in the isotropic acoustic case), $\mathbf{q}$ is assumed to be the known source, and $\mathbf{d}_{\text{obs}}$ is the observed data. The symbol $\mathbf{F}$ denotes the nonlinear forward modelling operator. This data misfit is typically minimized with gradient-based optimization methods such as gradient descent (Plessix, 2006) or Gauss–Newton (Li et al., 2016). Without loss of generality, let us first consider scalar isotropic acoustic wave physics where the gradient $\delta\mathbf{m}$ can be written as the sum over n_t timesteps – that is, we have

$$\delta\mathbf{m} = \sum_{t=1}^{n_t} \ddot{\mathbf{u}}[t] \odot \mathbf{v}[t], \tag{3}$$

where the vectors $\mathbf{u}[t]$ and $\mathbf{v}[t]$ denote the vectorized (along space) forward and reverse-time solutions of the wave equation at time index t. The symbols $\ddot{}$ and $\odot$ represent second-order time derivative and elementwise (Hadamard) product, respectively. To arrive at a form where randomized-trace estimation can be invoked, we rewrite the above zero-lag crosscorrelations over time for each space index $\mathbf{x}$ separately in terms of the trace of the outer product. By combining the dot product property, $\mathbf{a}^\top\mathbf{b} = \mathrm{tr}(\mathbf{a}\mathbf{b}^\top)$, for vectors $\mathbf{a}$ and $\mathbf{b}$, with Equation (1), we approximate the exact gradient in Equation (3) by

$$\delta\mathbf{m}[\mathbf{x}] = \mathrm{tr}\big(\ddot{\mathbf{u}}[t,\mathbf{x}]\mathbf{v}[t,\mathbf{x}]^\top\big) \approx \frac{1}{r}\mathrm{tr}\big((\mathbf{Q}^\top\ddot{\mathbf{u}}[\mathbf{x}])(\mathbf{v}[\mathbf{x}]^\top\mathbf{Q})\big). \tag{4}$$

We added parentheses and made dependence on the spatial coordinates (collected in the spatial index vector $\mathbf{x}$) explicit to show that the matrix–vector products between the forward and adjoint wavefields and the adjoint of the time-probing matrix $\mathbf{Q} \in \mathbb{R}^{n_t \times r}$ can be computed separately, independently along all spatial locations. This property is essential because it allows us to on-the-fly accumulate, $\mathbf{Q}^\top\ddot{\mathbf{u}}[\mathbf{x}]$, the second time derivative of the forward wavefield in the variable $\ddot{\mathbf{u}}$. Compared to the original wavefield, the dimension of this wavefield is reduced to $N \times r \ll N \times n_t$, where N is the

number of spatial gridpoints in $\mathbf{x}$ and n_t is the number of time samples employed by the forward solver. As before, r represents the number of probing vectors (columns) of $\mathbf{Q}$. Similarly, the dimensionality reduced adjoint wavefield, $\overline{\mathbf{v}}$, can be computed after the forward sweep is completed via $\mathbf{v}[\mathbf{x}]^\top\mathbf{Q}$. To avoid the build-up of coherent errors in the gradient due to the randomized probing, we repeat this process for each separate gradient with a different probing matrix, $\mathbf{Q}$. We compute this matrix with a different random realization of the Rademacher matrix used in the `qr` factorization, which we use to improve the accuracy of the random-trace estimator.

Practical choice for the probing matrix Q

While the proposed random-trace estimator works for strictly random probing vectors, for example, vectors with random ± 1 entries, as in Rademacher matrices, or matrices with independently and identically distributed Gaussian entries (Avron & Toledo, 2011), its accuracy can according to Meyer et al. (2020) be improved. This leads to a reduction in the number of probing vectors, r, and associated memory footprint needed to attain certain accuracy. However, this improvement in performance calls for an extra orthogonalization step that involves a `qr` factorization of $\mathbf{AZ}$, which reduces errors due to 'cross-talk' – that is $\mathbf{ZZ}^\top \neq \mathbf{I}$. Unfortunately, we do not have easy access to matrix–vector multiplications with $\mathbf{A}$ during gradient calculations. Moreover, factorization costs become prohibitively expensive when carried out for each of the N gridpoints separately.

Shot data informed QR factorization

To overcome computational costs and lack of access to matrix–vector products, we propose to work with a single factorization for each shot record. We derive this factorization from observed shot data. To this end, single shot data collected in the vector, $\mathbf{d}_{\text{obs}}$, are reshaped into a matrix, $\mathbf{D}_{\text{obs}}$ with the time index arranged along the rows and the receiver coordinate(s) along the columns. Because the observed shot data contain the wavefield along the receiver coordinate(s), we form the outer product, $\mathbf{A} = \mathbf{D}_{\text{obs}}\mathbf{D}_{\text{obs}}^\top$ and argue that the resulting $n_t \times n_t$ matrix can serve as a proxy for the temporal characteristics of the wavefield everywhere. For each shot record, the r probing vectors are computed as follows:

$$\left[\mathbf{Q}, \sim\right] = \text{qr}(\mathbf{AZ}) \quad \text{with} \quad \mathbf{A} = \mathbf{D}_{\text{obs}}\mathbf{D}_{\text{obs}}^\top. \tag{5}$$

Remember, as before we never form the matrix $\mathbf{A}$. We only compute its action on the Rademacher matrix $\mathbf{Z}$.

To demonstrate the benefits of the additional orthogonalization step, we include Figure 1 where comparisons are made between crosstalk produced by probing with Rademacher vectors (Figure 1 first row); with orthogonalized vectors derived from Rademacher probing according to Equation (5) (Figure 1 second row); and probing with vectors selected randomly from the Fourier matrix (Figure 1 third row). The orthogonalized vectors are computed using the same shot record from the 2D overthrust model with $n_t = 751$ samples and a 4-ms sampling rate (3 s recording). As expected, the crosstalk – that is, energy leakage away from the main diagonal, for these different cases, varies but decreases with increasing r for all. However, the frequency content and coherence of the errors do differ. Because the outer product converges the fastest the identity matrix within the seismic frequency band, we argue that the orthogonalized probing vectors perform the best.

ALGORITHM 1 Approximate gradient calculation with random trace estimation

Draw probing matrix $\mathbf{Q}$ with Equation (5), set initial condition $\mathbf{u}[0], \mathbf{u}[1]$ and final conditions $\mathbf{v}[n_t], \mathbf{v}[n_t - 1]$.

0. **for** $t = 1 : n_t - 1$ # forward propagation
1. $\mathbf{u}[t+1] = \text{forward}(\mathbf{u}[t], \mathbf{u}[t-1], \mathbf{m}, \mathbf{q}[t])$
2. **for** $i = 1 : r$

$\overline{\mathbf{u}}[i]$ += $\mathbf{Q}[i,t]\ddot{\mathbf{u}}[t]$

end for

3. **end for**
4. **for** $t = n_t - 1 : -1 : 1$ # back propagation
5. $\mathbf{v}[t-1] = \text{backward}(\mathbf{v}[t], \mathbf{v}[t+1], \mathbf{m}, \delta\mathbf{d}[t])$
6. **for** $i = 1 : r$

$\overline{\mathbf{v}}[i]$ += $\mathbf{v}[t]\mathbf{Q}[i,t]$

end for

7. **end for**
8. output: $\frac{1}{r}\text{tr}(\overline{\mathbf{u}}\,\overline{\mathbf{v}}^\top) = \frac{1}{r}\sum_{i=1}^{r}\overline{\mathbf{u}}[i] \odot \overline{\mathbf{v}}[i]$

Algorithmic details and validation

Based on the above practical and computational considerations, we propose the implementation as outlined in Algorithm 1. This algorithm runs for each source independently (possibly in parallel) and redraws a new probing matrix $\mathbf{Q}$ for each gradient computation. By propagating the forward wavefield with a single timestep (line 1), $\text{forward}(\mathbf{u}[t], \mathbf{u}[t-1], \mathbf{m}, \mathbf{q}[t])$, followed by probing with r vectors (line 2), the second derivative of the dimensionality reduced forward wavefield is accumulated for each time index (the 'for loop' starting at line 2). Notice that we suppressed the loop over the spatial index $\mathbf{x}$, which is implied. After the forward loop is completed, a similar process is followed when accumulating the dimensionality reduced adjoint wavefield after backpropagation with a single timestep via $\mathbf{v}[t-1] = \text{backward}(\mathbf{v}[t], \mathbf{v}[t+1], \mathbf{m}, \delta\mathbf{d}[t])$. After the second loop is completed, Algorithm 1 produces an estimate for the trace via $\frac{1}{r}\text{tr}(\overline{\mathbf{u}}\,\overline{\mathbf{v}}^\top) = \frac{1}{r}\sum_r \overline{\mathbf{u}}[r,:] \odot \overline{\mathbf{v}}[r,:]$ in which use is made of the Matlab-style notation.

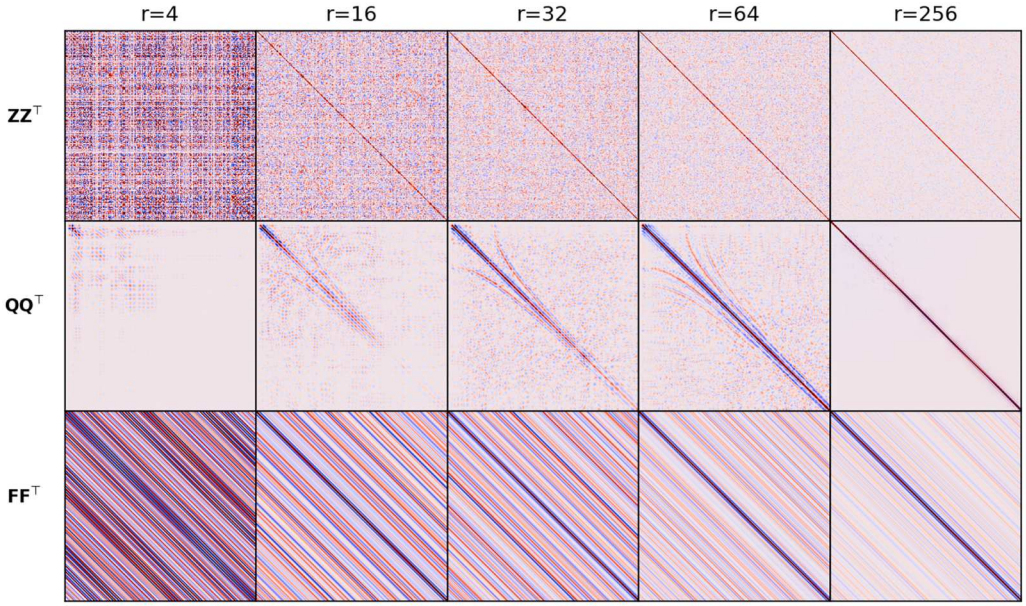

FIGURE 1 Crosstalk for different implementations of randomized probing as a function of increasing probing size $r = 4, 16, 32, 64, 256$. Compared to probing with Fourier vectors (third column), crosstalk for probing with Rademacher (first column) and orthogonalized Rademacher (second column) is less coherent with energy converging to the main diagonal faster. As expected, this effect is the strongest for the orthogonalized probing vectors within the seismic frequency band.

Before reviewing predicted memory savings, let us first make a comparison between single-source gradients computed with the three different probing vectors juxtaposed in Figure 1. To assess the accuracy of results with randomized-trace estimation, we set side by side the approximate gradients as a function of the probing size, $r = 4, 16, 32, 64, 256$ and compare these gradients with the true gradient. We make these comparisons for two-dimensional (2D) gradients computed from the overthrust model (Lecomte et al., 1994) with an experimental setting detailed in the Numerical case studies section.

The results of this exercise are summarized in Figure 2, which includes difference plots between the true and approximate gradients. The following observations can be made. First, in accordance with the results in Figure 1, the accuracy of the approximate gradient calculations improves for increasing r. Second, probing with Rademacher vectors (Figure 2a) yields gradients that contain noisy relatively high-spatial frequency artefacts that extend across the model and that decay relatively slowly as r increases. Results obtained with the orthogonalization (Figure 2b) and Fourier (Figure 2c) build up the gradient more slowly as a function of increasing r, capturing the small amplitudes far away from the source only for relatively large r. Because both the orthogonalized and Fourier approaches act within the data's temporal frequency spectrum, they do not contain high-frequency artefacts. Third, as expected results from the orthogonalized probing vectors converge the fastest with the smallest errors.

To further analyse the error, we consider its theoretical bound. As per Avron and Toledo (2011), error estimates for the trace decrease as $\mathcal{O}(\frac{1}{r})$ when Rademacher or Gaussian probing vectors are used on semi-definite positive matrices, where r represents the number of random probing. This can be expressed as

$$\|\widehat{\mathrm{tr}}(\mathbf{A}) - \mathrm{tr}(\mathbf{A})\| \leq \frac{\mathrm{const}}{r}, \tag{6}$$

where $\widehat{\mathrm{tr}}(\mathbf{A})$ represents the randomized-trace estimate and $\|.\|$ is the absolute value. In our case, we approximate the trace of the rank one matrix $\mathbf{A} = \ddot{\mathbf{u}}[t, \mathbf{x}]\mathbf{v}[t, \mathbf{x}]^\top$, whose only non-zero singular value is $\lambda = \ddot{\mathbf{u}}[t, \mathbf{x}]^\top \mathbf{v}[t, \mathbf{x}]$. Therefore, $\mathbf{A}$ or $-\mathbf{A}$ is always semi-definite positive ($\lambda \geq 0$ or $\lambda \leq 0$), satisfying the convergence-bound hypothesis introduced by Avron and Toledo (2011). The decay of our estimator's error is shown in Figure 3, where expected accuracy gains are achieved when increasing the number of probing vectors, r, irrespective of whether Rademacher or Gaussian probing vectors are used. Although Meyer et al. (2020) introduced a tighter error bound for probing vectors based on the QR factorization, we cannot expect to achieve this theoretical convergence rate because we use the observed data $\mathbf{D}_{\mathrm{obs}}$ as a proxy for the full-space wavefield, which is too large to manipulate. Nevertheless, we observe that the error for QR probing decays faster and is significantly lower, supporting the aforementioned claims regarding its benefits. We also note that the mathematical definition of the estimator and its upper bound are independent of the data's and wavefield's frequency content data. Consequently, the required number of probing vectors for an accurate estimate remains the same for various frequency and

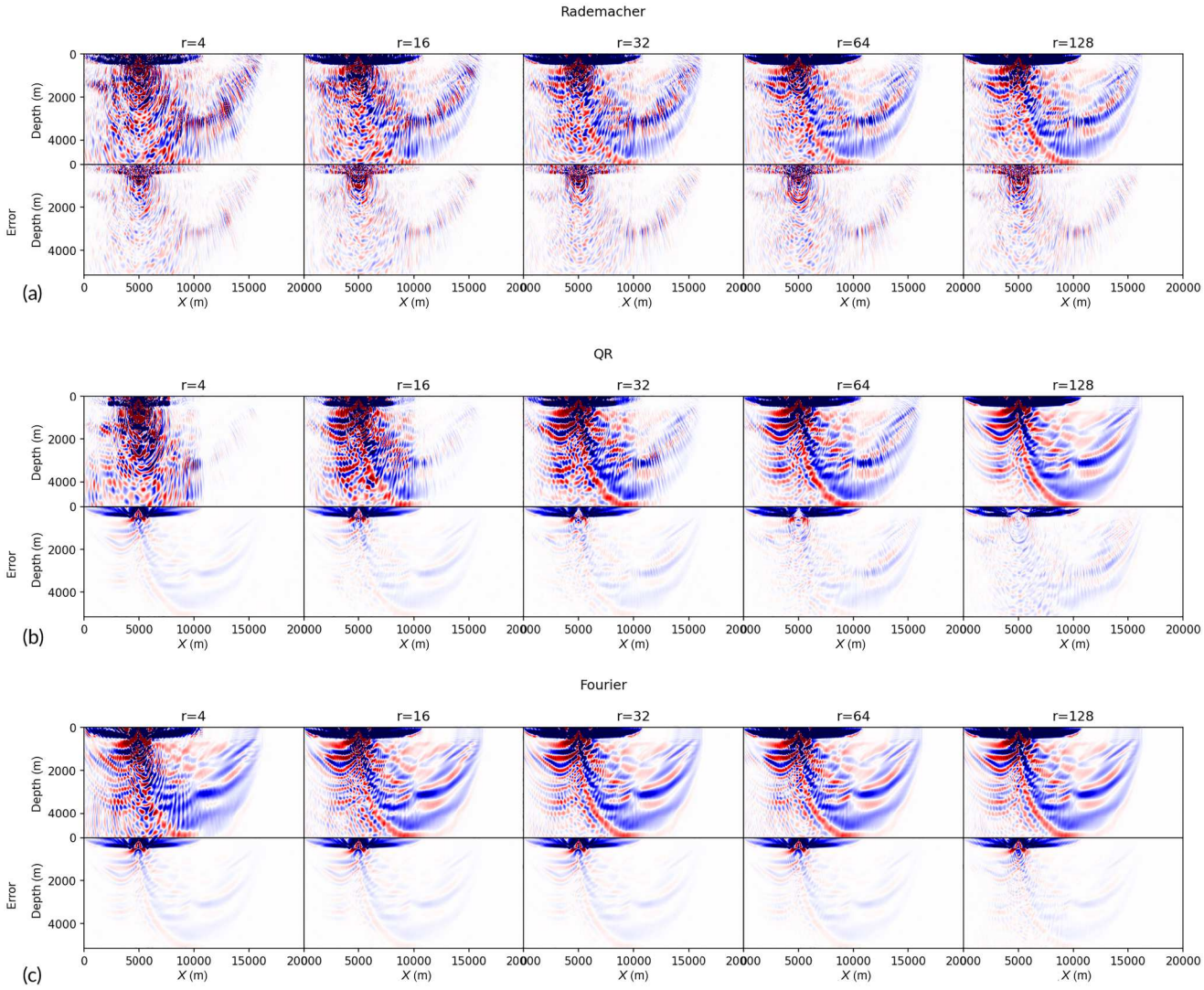

FIGURE 2 Comparison between approximate gradients of the 2D overthrust model for a single source and increasing numbers of probing vectors: (a) contains approximate gradients and errors obtained by probing with Rademacher vectors, (b) the same but with orthogonalized Rademacher probing vectors and (c) the same but with probing vectors randomly selected from the Fourier matrix.

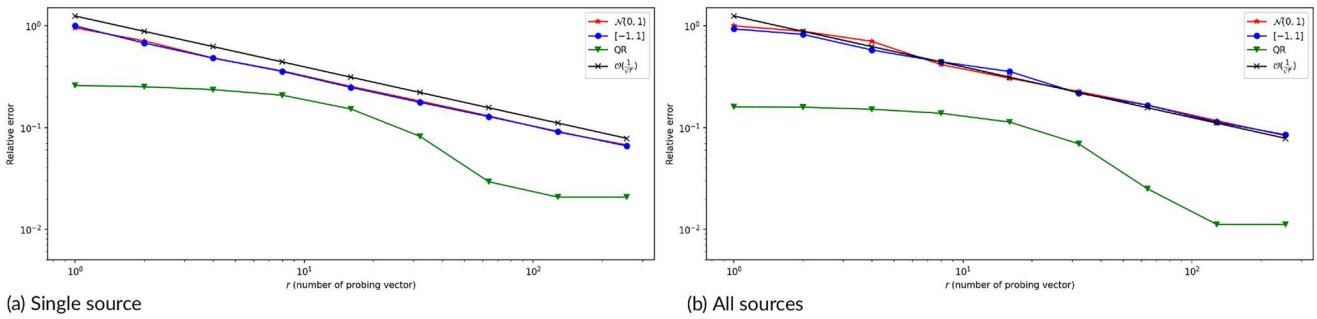

FIGURE 3 Relative error decay of our estimator compared to the theoretical upper bound of $\mathcal{O}(\frac{1}{r})$ as a function of the number of probing vectors. Errors averaged over all gridpoints for the gradient of the 2D overthrust model for a single source (left) and 25 sources (right) are plotted.

recording time settings. Finally, Figure 3b includes the decay of relative errors in the gradient approximations for the different probing vectors as a function of r for the single gradient case as (Figure 3a) as well as for the case where the gradients are stacked (Figure 3b). From these plots, we make the following observations. First, while probing with Gaussian and Rademacher probing vectors follow the theoretical bound, the relative errors for the QR-based probing are always smaller. Second, the errors for QR probing decay faster after the number of probing vectors is large enough to compensate for the approximation of the QR factorization the proxy derived from the observed data. Third, the relative errors after stacking the gradients decrease significantly while the relative errors for the other probing vectors stay close to the convergence bound. These improvements in relative errors not only justify the proposed algorithm but also demonstrate the validity of our implementation.

Before conducting more rigorous tests and emonstrating our claims on realistic two- and three-dimensional models, we will first discuss projected memory savings and various extensions involving more elaborate imaging conditions and derived products such as subsurface-offset image volumes.

Estimates for the memory footprint

As we mentioned before, the excessive memory usage of adjoint-state methods constitutes a major impediment to the implementation of wave-based inversion technology on modern accelerators where memory access comes at a premium. For this reason, we proposed approximate calculations with randomized probing where the dimensionality of the forward and adjoint wavefield is reduced by the method described in Algorithm 1. Theoretical estimates of the memory imprint can be computed easily from Equation (4). We compare these estimates with the memory footprint and computational overhead associated with other low-memory approaches, including optimal checkpointing (Griewank & Walther, 2000; Kukreja et al., 2020; Symes, 2007), reconstruction from wavefields on the boundary (McMechan, 1983; Mittet, 1994; Raknes & Weibull, 2016) and the closely related method based on probing with the discrete Fourier transform (DFT) (Nihei & Li, 2007; Sirgue et al., 2010; Witte, Louboutin, Luporini, et al., 2019). Without loss of generality, we will make these comparisons for the scalar acoustic wave equation in 3D, where $N = N_x \times N_y \times N_z$ is the total number of grid points and N_x, N_y, and N_z are the number of gridpoints in the x-, y- and z-directions. In that setting, the total memory requirement of conventional FWI is $N \times n_t$ (single-precision) floating point values, which is prohibitive in practice.

Table 1 lists estimates for memory use and computational overhead to achieve the anticipated memory savings. Values for optimal checkpointing and reconstruction from the boundary are taken from the literature. From these figures, we observe that optimal checkpointing could in principle achieve the largest memory savings at the expense of computational overhead and a relatively complex implementation. While memory savings achieved with the boundary reconstruction method do limit memory usage, this approach scales unfavourably with the number of timesteps, n_t, compared to the methods based on probing with Fourier or the proposed orthogonalized data-adaptive vectors. Because probing with the DFT involves complex numbers, its memory use and computational overhead doubles. Our method, on the other hand, probes with $\mathbf{Q} \in \mathbb{R}^{n_t \times r}$ and thus requires storage of only $N \times r$ floating point values during each of the forward and backward passes, which results in a total storage of $2 \times N \times r$ values and a memory reduction by a factor of, $n_t/2r$. This memory reduction corresponds to approximating the gradient with $\frac{r}{2}$ Fourier modes and puts the DFT approach at a relative disadvantage. Because the errors decay more slowly with r compared to our randomized-trace method, this drawback is made worse for the Fourier-based method.

Aside from its relative simplicity and favourable (n_t-independent) memory scaling, probing methods can, as we will show below, relatively easily be extended to different imaging conditions and vector-valued wave equations. In addition, the proposed method also works for wave propagation in attenuating media, which renders wavefield reconstruction from the boundaries unstable.

EXTENSIONS

In this section, we will show how the proposed randomized probing technique also leads to computationally efficient implementations of more involved imaging conditions including formating of subsurface offset image gathers. Both instances benefit from reductions in computational costs by a factor of $n_t/r \ll 1$.

Imaging conditions

So far, we limited ourselves to the scalar isotropic acoustic wave equation with the standard zero-offset imaging condition. While adequate in some applications, seismic imaging and inversion methodologies often call for more sophisticated imaging conditions designed to bring out certain features in migrated images or to make full-waveform inversion (FWI) more sensitive to reflections. Because imposing various imaging conditions requires manipulations with the forward and adjoint wavefields, we stand to benefit from replacing these wavefields with their dimensionality reduced counterparts. After incurring the computational overhead of probing, by mapping $\{\ddot{\mathbf{u}}, \mathbf{v}\} \rightarrow \{\overline{\mathbf{u}}, \overline{\mathbf{v}}\}$, these wavefield manipulations come almost at no additional costs. Below, we showcase a number of illustrative examples that underline this important feature.

TABLE 1 Memory estimates and computational overhead of different seismic inversion methods for n_t time steps and N grid points. Analytical estimates are extracted from the literature for the methods listed in the table.

	FWI	Randomized-trace	DFT (Witte, Louboutin, Luporini, et al., 2019)	Optimal checkpointing (Symes, 2007)	Boundary reconstruction (Mittet, 1994)
Memory	$N \times n_t$	$r \times N$	$2r \times N$	$\mathcal{O}(\log(n_t)) \times N$	$n_t \times N^{\frac{2}{3}}$
Compute	0	$\mathcal{O}(r) \times n_t \times N$	$\mathcal{O}(2r) \times n_t \times N$	$\mathcal{O}(\log(n_t)) \times n_t \times N$	$n_t \times N$

Abbreviations: DFT, discrete Fourier transform; FWI, full-waveform inversion.

Spatial differential operations

To improve reverse-time migration or FWI, different contributions of the gradient of the adjoint-state method can be (de)emphasized by changing the imaging condition. For instance, tomographic artefacts can be removed from reverse-time migrated images by imposing the inverse scattering imaging condition (Stolk et al., 2009; Op't Root et al., 2012; Whitmore & Crawley, 2012; Witte et al., 2017). In a related but different approach, in reflection FWI (Chang et al., 2020; Irabor & Warner, 2016; Liu et al., 2011), tomographic contributions to the gradient can be emphasized via wavefield separation. As with most imaging conditions, the inverse scattering condition does not entail manipulations along time and involves (differential) operators acting along the spatial coordinates only. Because imaging conditions are often linear, these operations commute with probing, which allows for direct application of imaging conditions on the dimensionality reduced wavefield by using the following identity:

$$\mathbf{Q}^\top\big(\mathbf{D}_x\mathbf{u}[\cdot,\mathbf{x}]\big) = \mathbf{D}_x\big(\mathbf{Q}^\top\mathbf{u}[\cdot,\mathbf{x}]\big), \tag{7}$$

where the symbol $\mathbf{D}_x$ represents a linear differential operator acting along the spatial coordinates. By virtue of this identity, numerically expensive operations with $\mathbf{D}_x$ can be factored out, reducing the number of applications of this operator from n_t to r. Because $r \ll n_t$, this can lead to significant computational savings, especially in the common situation where imposing imaging conditions may become almost as computationally expensive as solving the wave equation itself.

Subsurface-offset image gathers

Another benefit of approximating gradients via the trace (cf. Equation 4) is that it makes it possible to compute subsurface-offset image volumes directly on graphical processing units by working with the dimensionality reduced wavefields – that is, we have

$$\delta\mathcal{M}[\mathbf{x},\mathbf{h}] \approx \frac{1}{r}\mathrm{tr}\big(\bar{\tilde{\mathbf{u}}}[\cdot,\mathbf{x}+\mathbf{h}]\bar{\mathbf{v}}[\cdot,\mathbf{x}-\mathbf{h}]^\top\big). \tag{8}$$

In this expression, the symbol $\mathbf{h}$ corresponds to the (horizontal) subsurface offset and $\delta\mathcal{M}[\mathbf{h}]$ to the subsurface image volume. As with computing the zero-offset imaging condition (Equation 4), the cost of computing these extended image volumes is reduced by a factor of r/n_t. Finally, note that $\delta\mathbf{m}[\mathbf{x}] = \delta\mathcal{M}[\mathbf{x},h]|_{\mathbf{h}=0}$.

Coupled vector-valued wave equation

Adequate representation of the wave physics balanced by computational considerations are prerequisites to the success of seismic inversion on three-dimensional (3D) field data. A good example where such a balance is struck is wave modelling with the acoustic tilted transverse isotropic (TTI) wave-equation, where elastic anisotropic behaviour of the subsurface is modelled by an acoustic approximation (Thomsen, 1986) that is computationally feasible. However, compared to the isotropic scalar acoustic wave equation, the TTI wave equation requires the solution of two coupled partial differential equation (PDEs). Because the gradient with respect to the squared slowness and anisotropic parameters now involves four wavefields, this increases memory pressure. According to Bube et al. (2016), Louboutin, Witte, & Herrmann (2018) and Zhang et al. (2011), the gradient for the squared slowness in TTI media reads

$$\delta\mathbf{m} = \sum_t \ddot{\mathbf{p}}[t] \odot \bar{\mathbf{p}}[t] + \ddot{\mathbf{r}}[t] \odot \bar{\mathbf{r}}[t] = \mathrm{tr}(\ddot{\mathbf{p}}\bar{\mathbf{p}}^\top) + \mathrm{tr}(\ddot{\mathbf{r}}\bar{\mathbf{r}}^\top), \tag{9}$$

where $\mathbf{p}$ and $\mathbf{r}$ are solutions of two coupled PDEs and $\bar{\mathbf{p}}$ and $\bar{\mathbf{r}}$ are solutions of the adjoint of these coupled PDEs. When implemented naively, the memory footprint would effectively double when the above gradients are approximated with separate randomized-trace estimations for $\ddot{\mathbf{p}}[t] \odot \bar{\mathbf{p}}[t]$ and $\ddot{\mathbf{r}}[t] \odot \bar{\mathbf{r}}[t]$. However, these extra costs can be avoided if we make use of the following identity:

$$\mathrm{tr}(\ddot{\mathbf{p}}\bar{\mathbf{p}}^\top) + \mathrm{tr}(\ddot{\mathbf{r}}\bar{\mathbf{r}}^\top) = \mathrm{tr}(\begin{bmatrix}\ddot{\mathbf{p}}\\ \ddot{\mathbf{r}}\end{bmatrix}\begin{bmatrix}\bar{\mathbf{p}}\\ \bar{\mathbf{r}}\end{bmatrix}^\top), \tag{10}$$

which holds for the trace of vector-valued wavefields. When cast in this form, the above gradient can be approximated by

$$\begin{aligned}\delta\mathbf{m} = \mathrm{tr}(\begin{bmatrix}\ddot{\mathbf{p}}\\ \ddot{\mathbf{r}}\end{bmatrix}\begin{bmatrix}\bar{\mathbf{p}}\\ \bar{\mathbf{r}}\end{bmatrix}^\top) &\approx \frac{1}{r}\mathrm{tr}(\begin{bmatrix}\mathbf{Q}\\ \mathbf{Q}\end{bmatrix}^\top\begin{bmatrix}\ddot{\mathbf{p}}\\ \ddot{\mathbf{r}}\end{bmatrix}\begin{bmatrix}\bar{\mathbf{p}}\\ \bar{\mathbf{r}}\end{bmatrix}^\top\begin{bmatrix}\mathbf{Q}\\ \mathbf{Q}\end{bmatrix})\\ &\approx \frac{1}{r}\mathrm{tr}\big((\mathbf{Q}^\top(\ddot{\mathbf{p}}+\ddot{\mathbf{r}}))((\bar{\mathbf{p}}+\bar{\mathbf{r}})^\top\mathbf{Q})\big)\end{aligned} \tag{11}$$

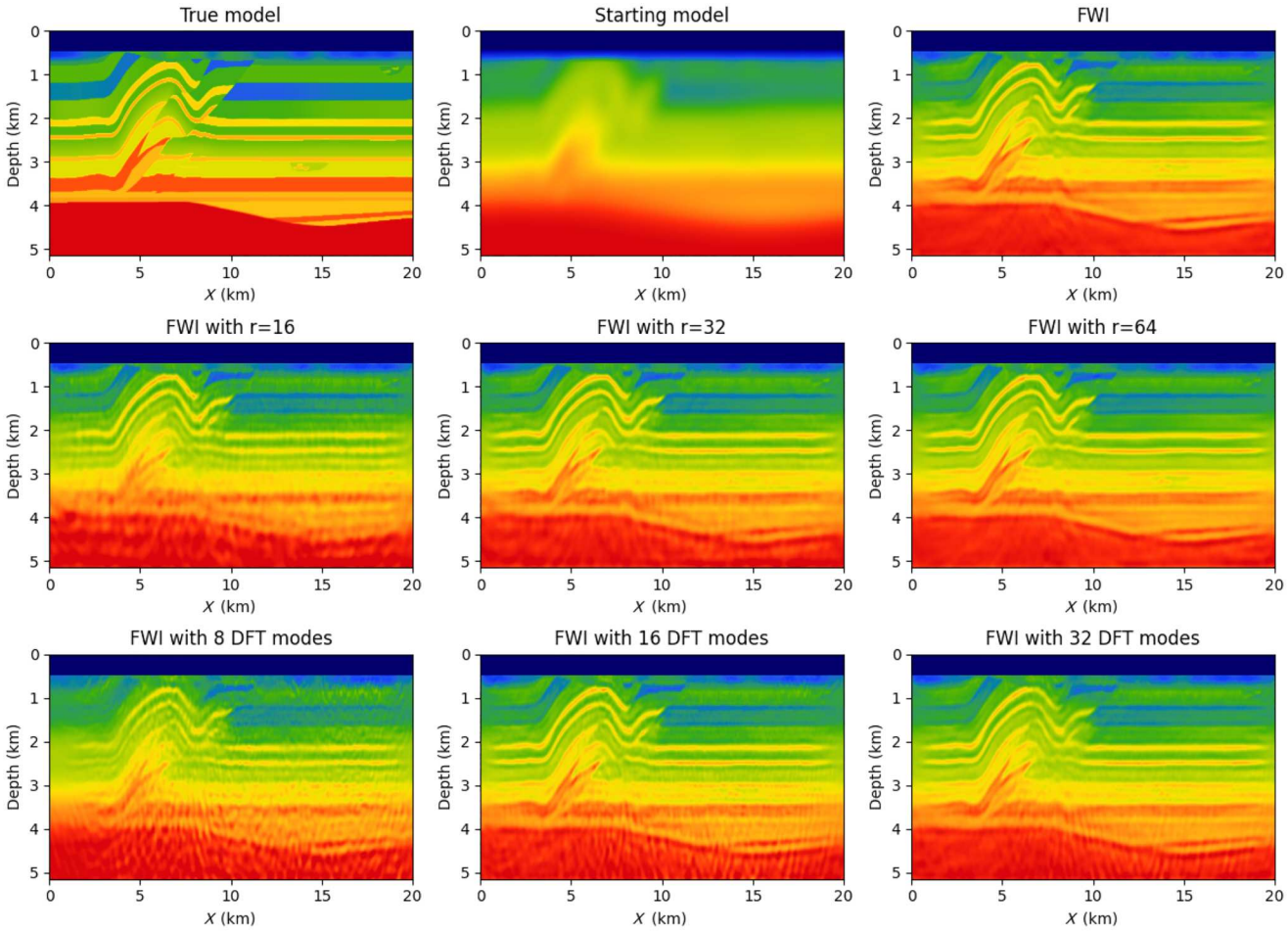

FIGURE 4 Comparison of FWI on the 2D overthrust model between our proposed probed method with 16, 32 and 64 orthogonalized probing vectors and results obtained with on-the-fly DFTs with an equivalent memory imprint, which corresponds to 8, 16 and 32 Fourier modes.

if the same probing vectors are used for each component of the vector-valued wavefield. Using this expression reduces the memory cost to that of a single probed wavefield and, consequently, its memory use remains the same as that of gradients of isotropic acoustic media, which we consider as a major advantage of our method. In practice, we observe that the accuracy of this approximation does not decrease when the same probing matrix $\mathbf{Q}$ is used. With significant computational and memory savings established, we will now validate its performance on realistic numerical experiments of varying complexity and problem size.

NUMERICAL CASE STUDIES

While the presented methodology has the potential to unlock the usage of memory-scarce accelerators, its performance needs to be validated on realistic wave-equation-based inversion examples. For this purpose, we consider three synthetic examples that vary in complexity of the wave physics. First, we revisit the two-dimensional overthrust model and compare conventional full-waveform inversion (FWI) results with inversions obtained with randomized-trace estimation for increasing numbers of probing vectors. In the second example, we demonstrate that our probing method is capable of producing high-quality subsurface-offset image volumes at a significantly reduced computational cost. We conclude by showcasing a three-dimensional FWI example. We limit our considerations to synthetic data because it allows us to make informed comparisons between exact and approximate gradient calculations.

TWO-DIMENSIONAL FULL-WAVEFORM INVERSION

As part of validating our random-trace estimation technique, we consider a synthetic two-dimensional (2D) acoustic full-waveform inversion (FWI) example with a geometry representing a sparse ocean bottom nodes (OBN) acquisition while applying source–receiver reciprocity (coarse sources, dense receivers). The data are simulated for a 20-km by

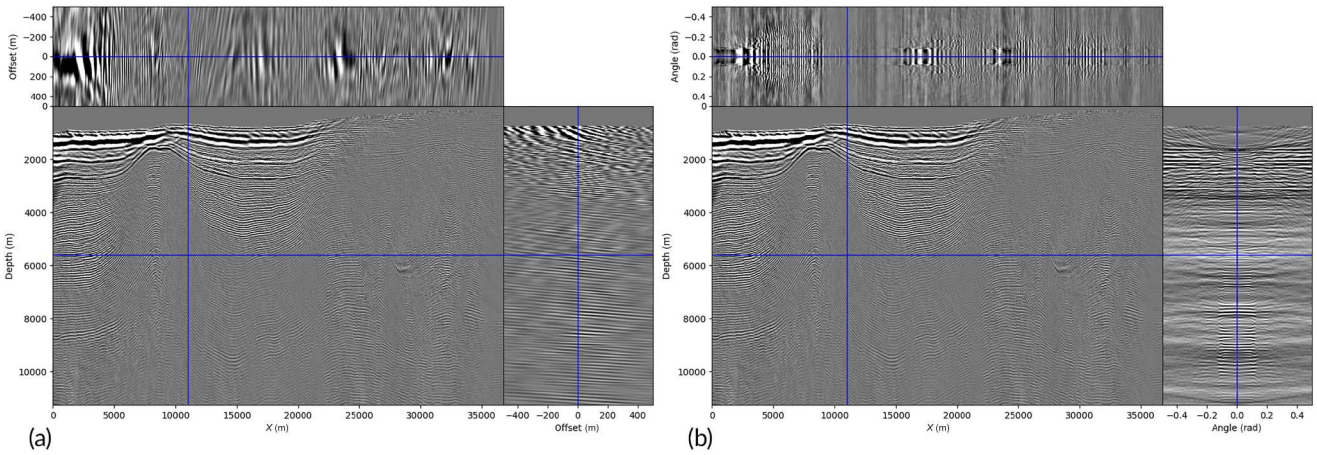

FIGURE 5 Subsurface offset (a) and angle b) gathers with −500 m to 500 m horizontal subsurface offset and −0.5 to 0.5 rad subsurface angles.

5-km section taken from the overthrust model (Lecomte et al., 1994) and plotted in Figure 4 (top left). For the FWI experiment, we work with 97 shot records sampled 200-m apart, mimicking sparse OBNs sampled at one source position per wavelength. Each shot record contains between 127 and 241 receivers 50-m apart, yielding a maximum offset of 6-km. The data are modelled with an 8-Hz Ricker wavelet and 3-s recording.

For reference, we first conduct conventional FWI given the smooth starting model depicted in Figure 4 (top middle). We compare this conventional FWI result plotted in Figure 4 (top tight) with results yielded by approximate gradient calculations where the memory footprint is kept the same experiment by experiment – that is, $r = 16$, 32, $r = 64$ for randomized-trace estimation with orthogonalized probing vectors, and 8, 16, 32 for probing with randomly selected Fourier modes. Results of these experiments are included in the second and third rows of Figure 4. In all cases, FWI results are computed with 20 iterations of the spectral projected gradient method (Schmidt et al., 2009), which imposes box and total-variation (TV) constraints on the inverted velocity model. Computational costs are limited by working with subsets of eight randomly selected (without replacement) shots (Aravkin et al., 2012). From the approximations plotted in Figure 4, we can make the following observations. First, compared to results obtained with 16 Fourier modes our result for the same number of probing vectors contains fewer coherent steeply dipping artefacts especially at deeper areas of the inverted velocity model. Second, when memory use is kept constant, for example, by choosing $r = 32$ orthogonalized probing vector and 16 complex-valued discrete Fourier transform modes, our method produces results that are more accurate and less noisy. This observation is consistent with results presented in Figure 1.

TWO-DIMENSIONAL EXTENDED TILTED TRANSVERSE ISOTROPIC IMAGING

To illustrate our ability to handle more realistic physics, we show that it is possible to create high-fidelity subsurface-offset image gathers with the proposed randomized-trace estimation technique. For this purpose, shot data provided with the 2007 BP tilted transverse isotropic model is migrated using our randomized approximation. The resulting image – that is, the zero offset/angle section, is shown in Figure 5 and shows the accurate location and continuity of reflectors compared to the existing literature (Sun et al., 2016; Louboutin, Witte, & Herrmann , 2018). We also computed subsurface image gathers using the approximation given in Equation (11). We show the computed subsurface-offset image gathers in Figure 5a and corresponding subsurface-angle gathers in Figure 5b. Even though only a limited number of probing vectors ($r = 64 \ll n_t = 4600$, 70× memory reduction) were used, the resulting image gathers are properly focused and nearly noise-free thanks to the relatively high fold. Each image volume, of size $n_x \times n_y \times n_{\text{offsets}}$, consists of 81 offsets between −500 and 500 m sampled at 12.5 m. Remark that formation of these image volumes requires more memory, namely 81 model-size arrays, than carrying out the probing itself, which involves only 64 model-size probed wavefields for $\overline{\mathbf{u}}, \overline{\mathbf{v}}$. This highlights how memory frugal our proposed randomized-trace estimation method really is.

THREE-DIMENSIONAL FULL-WAVEFORM INVERSION

Finally, to demonstrate scalability we run three-dimensional (3D) full-waveform inversion (FWI) on the overthrust model.

Geophysical Prospecting EAGE EUROPEAN ASSOCIATION OF GEOSCIENTISTS & ENGINEERS

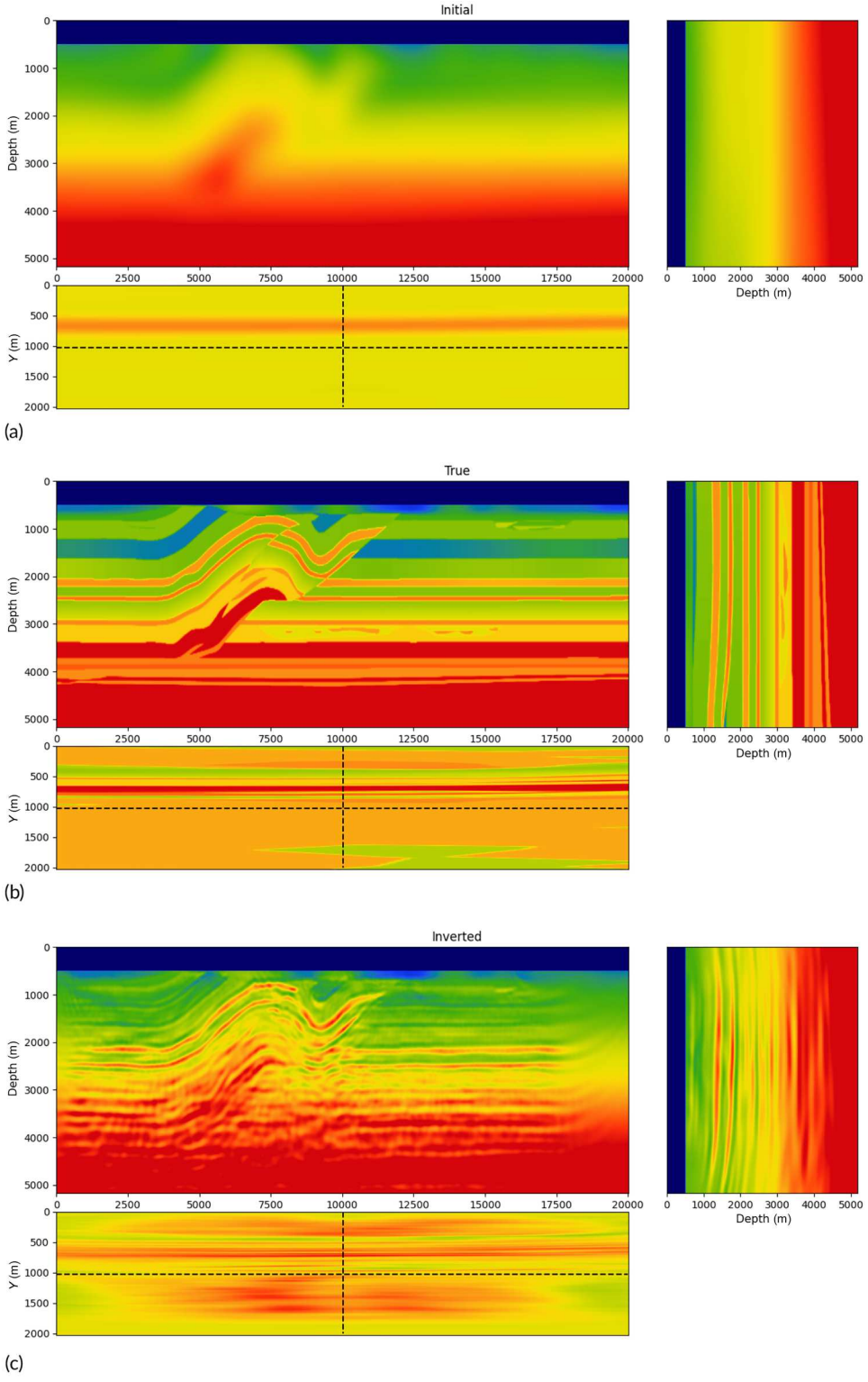

FIGURE 6 3D FWI of the 3D overthrust model on Azure using 50 K80 (NVIDIA Kepler, 8 Gb memory) nodes. The inversion was performed with 32 probing vectors ($r = 32$), which approximate each gradient natively on the GPUs without relying on any form of checkpointing.

The computational resources needed for this inversion exceed available memory on Azure's Standard_NC6 instances whose NVIDIA K80 GPUs are limited to 8 Gb each, rendering FWI implementations without checkpointing/offloading/streaming impractical. Because our method only requires a fraction of memory, we are actually able to run 3D FWI with randomized-trace estimation (for $r = 32 \ll n_t = 2001$, 30× memory reduction) on 50 instances.

Specifically, we consider a narrow azimuth towed streamer acquisition on a 20 km × 2 km × 5 km (inline × crossline × depth) subsection of the overthrust model. The acquisition consists of 1902 sources (50 m inline spacing and 200 m crossline spacing) with six 8 km long cables 100 m apart with a receiver spacing of 12.5 m per. We simulate the shot data with a 12.5-Hz Ricker wavelet. To avoid unrealistic low frequencies, frequencies below 3 Hz are removed with a high-pass filter. Given these simulations, 20 iterations of FWI are performed using $r = 32$ orthogonalized probing vectors and 400 sources per iteration. As before, FWI is carried out with projected quasi-Newton (Schmidt et al., 2009) imposing both box and total-variation (TV) constraints (Peters et al., 2018; Peters & Herrmann, 2019). The box constraints guarantee physical velocities, while the TV constraint removes noisy artefacts due to the randomized-trace estimation. The inverted velocity model is included in Figure 6.

From this experiment, we see that we recover an accurate velocity model that contains most of the main features of the true model and recovered most of the fine layers at depth. Swing artefacts are being observed towards the edges of the model. However, these are more likely to be associated with the marine acquisition rather than with the proposed gradient approximation. As shown in the two-dimensional (2D) example (Figure 4), our method only introduces incoherent noise instead of coherent structural artefacts. This result shows that the proposed method scales to a realistic three-dimensional model without the need for additional probing vectors to compensate for the added dimension (cf. the 2D FWI result).

DISCUSSION AND CONCLUSIONS

By approximating the gradient of wave-based inversion with randomized trace estimation, we were able to drastically reduce the memory footprint of time-domain full-waveform inversion and reverse-time migration with the adjoint-state method. Through careful design of data-adaptive probing vectors, memory reductions of about 50× were achieved without tangible loss in accuracy. These attained memory reductions, in turn, facilitate accelerator-native software implementations for the time-domain adjoint-state method, which benefit maximally from graphics processing units with limited computational overhead. To achieve these results, we controlled the approximation errors due to randomized-trace estimation by increasing the number of probing vectors, the fold and by imposing additional constraints, for example, the total-variation norm, during inversion. Because of its relative simplicity, the proposed method can be extended readily to more complicated wave physics, including vector-valued wavefields in transversely isotropic media. By virtue of the memory footprint reduction, the proposed method is also capable of efficient calculation of extended (subsurface-offset) image volumes with computational gains that are proportional to the reductions in memory usage. In future work, we plan to expand this work to include extended Born least-squares migration and extended full-waveform inversion.

ACKNOWLEDGEMENTS

This research was carried out with the support of the Georgia Research Alliance and partners of the ML4Seismic Center. We thank John Washbourne for the constructive discussions.

DATA AVAILABILITY STATEMENT

Our implementation and examples are available as open-source software, TimeProbeSeismic.jl, at https://github.com/slimgroup/TimeProbeSeismic.jl , which extends our Julia inversion framework, JUDI.jl (https://github.com/slimgroup/JUDI.jl) (Louboutin et al., 2022; Witte, Louboutin, Kukreja, et al., 2019). Our code is available at https://github.com/slimgroup , and since it is built on Devito (https://www.devitoproject.org) (Louboutin et al., 2019; F. Luporini et al., 2020; Luporini et al., 2022) it supports more complicated wave physics.

ORCID


Mathias Louboutin https://orcid.org/0000-0002-1255-2107

REFERENCES

Aravkin, A.Y., Friedlander, M.P., Herrmann, F.J. & van Leeuwen, T. (2012) Robust inversion, dimensionality reduction, and randomized sampling. *Mathematical Programming*, 134(1), 101–125. http://www.springerlink.com/content/35rwr101h5736340/

Avron, H. & Toledo, S. (2011) Randomized algorithms for estimating the trace of an implicit symmetric positive semi-definite matrix. *Journal of ACM*, 58(2), 1–34.https://doi.org/10.1145/1944345.1944349

Bube, K., Washbourne, J., Ergas, R. & Nemeth, T. (2016) Self-adjoint, energy-conserving second-order pseudoacoustic systems for VTI and TTI media for reverse time migration and full-waveform inversion. In *SEG Technical Program Expanded Abstracts 2016*, Housten, TX: Society of Exploration Geophysicists, pp. 1110–1114. https://library.seg.org/doi/10.1190/segam2016-13878451.1

Chang, K., Song, C., Alkhalifah, T. & Zhang, H. (2020) In *SEG Technical Program Expanded Abstracts 2020* Housten, TX: Society of Exploration Geophysicists, pp. 770–774. https://library.seg.org/doi/abs/10.1190/segam2020-3427081.1

Chen, T., Xu, B., Zhang, C. & Guestrin, C. (2016) Training deep nets with sublinear memory cost. CoRR. abs/1604.06174. http://arxiv.org/abs/1604.06174

Friedlander, M.P. & Schmidt, M. (2012) Hybrid deterministic-stochastic methods for data fitting. *SIAM Journal on Scientific Computing*, 34(3), A1380–A1405. https://doi.org/10.1137/110830629

Graff-Kray, M., Kumar, R. & Herrmann, F.J. (2017) Low-rank representation of omnidirectional subsurface extended image volumes. https://slim.gatech.edu/Publications/Public/Conferences/SINBAD/2017/Fall/graff2017SINBADFlrp/graff2017SINBADFlrp.pdf

Griewank, A. & Walther, A. (2000) Algorithm 799: Revolve: An implementation of checkpointing for the reverse or adjoint mode of computational differentiation. *ACM Transactions on Mathematical Software*, 26(1), 19–45. http://doi.acm.org/10.1145/347837.347846

Haber, E., van den Doel, K. & Horesh, L. (2015) Optimal design of simultaneous source encoding. *Inverse Problems in Science and Engineering*, 23(5), 780–797. https://doi.org/10.1080/17415977.2014.934821

Halko, N., Martinsson, P.G. & Tropp, J.A. (2011) Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review*, 53(2), 217–288. https://doi.org/10.1137/090771806

Irabor, K. & Warner, M. (2016) Reflection FWI. In *SEG Technical Program Expanded Abstracts 2016*. Housten, TX: Society of Exploration Geophysicists, pp. 1136–1140. https://library.seg.org/doi/abs/10.1190/segam2016-13944219.1

Krebs, J.R., Anderson, J.E., Hinkley, D., Neelamani, R., Lee, S., Baumstein, A. & Lacasse, M.-D. (2009) Fast full-wavefield seismic inversion using encoded sources. *Geophysics*, 74(6), WCC177–WCC188. https://doi.org/10.1190/1.3230502

Kukreja, N., Hückelheim, J., Louboutin, M., Washbourne, J., Kelly, P.H.J. & Gorman, G.J. (2020) Lossy checkpoint compression in full waveform inversion. *Geoscientific Model Development Discussions*, 2020, 1–26. https://gmd.copernicus.org/preprints/gmd-2020-325/

Lecomte, J.-C., Campbell, E. & Letouzey, J. (1994) Building the SEG/EAGE overthrust velocity macro model. In *Conference Proceedings*. Houten, the Netherlands: European Association of Geoscientists and Engineers. https://www.earthdoc.org/content/papers/10.3997/2214-4609.201407587

Li, X., Esser, E. & Herrmann, F.J. (2016) Modified Gauss-Newton full-waveform inversion explained—why sparsity-promoting updates do matter. *Geophysics*, 81(3), R125–R138. https://slim.gatech.edu/Publications/Public/Journals/Geophysics/2016/li2015GEOPmgn/li2015GEOPmgn.pdf

Lions, J.L. (1971) *Optimal control of systems governed by partial differential equations*, 1st ed. Berlin: Springer-Verlag.

Liu, F., Zhang, G., Morton, S.A. & Leveille, J.P. (2011) An effective imaging condition for reverse-time migration using wavefield decomposition. *Geophysics*, 76(1), S29–S39. https://doi.org/10.1190/1.3533914

Louboutin, M. & Herrmann, F. (2022) Enabling wave-based inversion on GPUs with randomized trace estimation. In *Conference proceedings*, volume 2022. Houten, the Netherlands: European Association of Geoscientists and Engineers, pp. 1–5. https://www.earthdoc.org/content/papers/10.3997/2214-4609.202210531

Louboutin, M. & Herrmann, F.J. (2021) Ultra-low memory seismic inversion with randomized trace estimation. In *SEG Technical Program Expanded Abstracts* (IMAGE, Denver). Houston, TX: Society of Exploration Geophysicists, pp. 787–791. https://slim.gatech.edu/Publications/Public/Conferences/SEG/2021/louboutin2021SEGulm/louboutinp.html

Louboutin, M., Lange, M., Luporini, F., Kukreja, N., Witte, P.A., Herrmann, F.J., Velesko, P. & Gorman, G.J. (2019) Devito (v3.1.0): an embedded domain-specific language for finite differences and geophysical exploration. *Geoscientific Model Development*, 12(3), 1165–1187. https://www.geosci-model-dev.net/12/1165/2019/

Louboutin, M., Witte, P. & Herrmann, F.J. (2018) Effects of wrong adjoints for RTM in TTI media. In *SEG Technical Program Expanded Abstracts 2018*, Houston, TX: Society of Exploration Geophysicists, pp. 331–335. https://library.seg.org/doi/abs/10.1190/segam2018-2996274.1

Louboutin, M., Witte, P., Yin, Z.F., Modzelewski, H. & Herrmann, F.J. (2022) slimgroup/judi.jl: v3.2.1. https://doi.org/10.5281/zenodo.7429592

Louboutin, M., Witte, P.A., Lange, M., Kukreja, N., Luporini, F., Gorman, G. & Herrmann, F.J. (2017) Full-waveform inversion - part 1: forward modeling. *The Leading Edge*, 36(12), 1033–1036. https://slim.gatech.edu/Publications/Public/Journals/TheLeadingEdge/2017/louboutin2017fwi/louboutin2017fwi.html

Louboutin, M., Witte, P.A., Lange, M., Kukreja, N., Luporini, F., Gorman, G. & Herrmann, F.J. (2018) Full-waveform inversion - part 2: adjoint modeling. *The Leading Edge*, 37(1), 69–72. https://slim.gatech.edu/Publications/Public/Journals/TheLeadingEdge/2018/louboutin2017fwip2/louboutin2017fwip2.html

Luporini, F., Louboutin, M., Lange, M., Kukreja, N., rhodrin, Bisbas, G., Pandolfo, V., Cavalcante, L., Burgess, T., Gorman, G. & Hester, K. (2022) devitocodes/devito: v4.7.1. https://doi.org/10.5281/zenodo.6958070

Luporini, F., Louboutin, M., Lange, M., Kukreja, N., Witte, P., Hückelheim, J., Yount, C., Kelly, P. H.J., Herrmann, F.J. & Gorman, G.J. (2020) Architecture and performance of Devito, a system for automated stencil computation. *ACM Transactions on Mathematical Software*, 46(1), 1–28. https://doi.org/10.1145/3374916

McMechan, G.A. (1983) Migration by extrapolation of time-dependent boundary values. *Geophysical Prospecting*, 31(3), 413–420. http://doi.org/10.1111/j.1365-2478.1983.tb01060.x

Meyer, R.A., Musco, C., Musco, C. & Woodruff, D.P. (2020) Hutch++: optimal stochastic trace estimation. arXiv. arXiv:2010.09649.

Mittet, R. (1994) Implementation of the Kirchhoff integral for elastic waves in staggered-grid modeling schemes. *Geophysics*, 59(12), 1894–1901. http://doi.org/10.1190/1.1443576

Moghaddam, P.P., Keers, H., Herrmann, F.J. & Mulder, W.A. (2013) A new optimization approach for source-encoding full-waveform inversion. *Geophysics*, 78(3), R125–R132. https://doi.org/10.1190/geo2012-0090.1

Neelakantan, A., Vilnis, L., Le, Q.V., Sutskever, I., Kaiser, L., Kurach, K. & Martens, J. (2015) Adding gradient noise improves learning for very deep networks. https://arxiv.org/abs/1511.06807

Nihei, K.T. & Li, X. (2007) Frequency response modelling of seismic waves using finite difference time domain with phase sensitive detection (TD-PSD). *Geophysical Journal International*, 169(3), 1069–1078. https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-246X.2006.03262.x

Op't Root, T.J., Stolk, C.C. & Maarten, V. (2012) Linearized inverse scattering based on seismic reverse time migration. *Journal de mathématiques pures et appliquées*, 98(2), 211–238.

Peters, B. & Herrmann, F.J. (2019) Generalized Minkowski sets for the regularization of inverse problems. https://arxiv.org/abs/1903.03942

Peters, B., Smithyman, B.R. & Herrmann, F.J. (2018) Projection methods and applications for seismic nonlinear inverse problems with multiple constraints. *Geophysics*, 84(2), R251–R269. https://slim.gatech.edu/Publications/Public/Journals/Geophysics/2018/peters2018pmf/peters2018pmf.html

Plessix, R.-E. (2006) A review of the adjoint-state method for computing the gradient of a functional with geophysical applications. *Geophysical Journal International*, 167(2), 495–503. http://doi.org/10.1111/j.1365-246X.2006.02978.x

Raknes, E.B. & Weibull, W. (2016) Efficient 3D elastic full-waveform inversion using wavefield reconstruction methods. *Geophysics*, 81(2), R45–R55. http://geophysics.geoscienceworld.org/content/81/2/R45

Romero, L.A., Ghiglia, D.C., Ober, C.C. & Morton, S.A. (2000) Phase encoding of shot records in prestack migration. *Geophysics*, 65(2), 426–436. https://doi.org/10.1190/1.1444737

Schmidt, M., Berg, E., Friedlander, M. & Murphy, K. (2009) Optimizing costly functions with simple constraints: a limited-memory projected quasi-Newton algorithm. In *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics*, Vol. 5, 456–463. Cambridge, MA: MIT Press. http://proceedings.mlr.press/v5/schmidt09a.html

Sirgue, L., Etgen, J., Albertin, U. & Brandsberg-Dahl, S. (2010) System and method for 3D frequency domain waveform inversion based on 3D time-domain forward modeling. US Patent 7,725,266. http://www.google.ca/patents/US7725266

Stolk, C., De Hoop, M. & Op't Root, T. (2009) Inverse scattering of seismic data in the reverse time migration (RTM) approach. In *Proceedings of the Project Review, Geo-Mathematical Imaging Group*, vol. 1. 91–108.

Sun, J., Fomel, S. & Ying, L. (2016) Low-rank one-step wave extrapolation for reverse time migration. *Geophysics*, 81(1), S39–S54. https://doi.org/10.1190/geo2015-0183.1

Symes (2007) Reverse time migration with optimal checkpointing. *Geophysics*, 72(5), SM213–SM221. http://library.seg.org/doi/abs/10.1190/1.2742686

Tarantola, A. (1984) Inversion of seismic reflection data in the acoustic approximation. *Geophysics*, 49(8), 1259. http://doi.org/10.1190/1.1441754

Thomsen, L. (1986) Weak elastic anisotropy. *Geophysics*, 51(10), 1954–1966.

Trefethen, L.N. & Bau, D., III. (1997) *Numerical linear algebra*. 1st ed.. Philadelphia, PA: SIAM.

van Leeuwen, T. & Herrmann, F.J. (2013) Fast waveform inversion without source-encoding. *Geophysical Prospecting*, 61(1), 10–19. https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-2478.2012.01096.x

van Leeuwen, T. & Herrmann, F.J. (2014) 3D frequency-domain seismic inversion with controlled sloppiness. *SIAM Journal on Scientific Computing*, 36(5), S192–S217.

van Leeuwen, T., Kumar, R. & Herrmann, F.J. (2017) Enabling affordable omnidirectional subsurface extended image volumes via probing. *Geophysical Prospecting*, 65(2), 385–406. https://slim.gatech.edu/Publications/Public/Journals/GeophysicalProspecting/2016/vanleeuwen2015GPWEMVA/vanleeuwen2015GPWEMVA.pdf

Virieux, J. & Operto, S. (2009) An overview of full-waveform inversion in exploration geophysics. *Geophysics*, 74(5), WCC1–WCC26. http://library.seg.org/doi/abs/10.1190/1.3238367

Whitmore, N.D. & Crawley, S. (2012) Applications of RTM inverse scattering imaging conditions. In *SEG Technical Program Expanded Abstracts 2012*. Houston, TX: Society of Exploration Geophysicists, pp. 1–6. https://library.seg.org/doi/abs/10.1190/segam2012-0779.1

Witte, P., Louboutin, M., Luporini, F., Gorman, G.J. & Herrmann, F.J. (2019) Compressive least-squares migration with on-the-fly Fourier transforms. *Geophysics*, 84(5), R655–R672. https://slim.gatech.edu/Publications/Public/Journals/Geophysics/2019/witte2018cls/witte2018cls.pdf

Witte, P., Yang, M. & Herrmann, F. (2017) Sparsity-promoting least-squares migration with the linearized inverse scattering imaging condition. In *79th EAGE Conference and Exhibition 2017*, volume 2017, Houten, the Netherlands: European Association of Geoscientists and Engineers, pp. 1–5. https://www.earthdoc.org/content/papers/10.3997/2214-4609.201701125

Witte, P.A., Louboutin, M., Kukreja, N., Luporini, F., Lange, M., Gorman, G.J. & Herrmann, F.J. (2019) A large-scale framework for symbolic implementations of seismic inversion algorithms in Julia. *Geophysics*, 84(3), F57–F71. https://doi.org/10.1190/geo2018-0174.1

Yang, M., Graff, M., Kumar, R. & Herrmann, F.J. (2021) Low-rank representation of omnidirectional subsurface extended image volumes. *Geophysics*, 86(3), 1–41. https://slim.gatech.edu/Publications/Public/Journals/Geophysics/2021/yang2020lrpo/Paper_final.html

Zhang, H. & Constantinescu, E.M. (2022) Optimal checkpointing for adjoint multistage time-stepping schemes. *Journal of Computational Science*, 101913.

Zhang, Y., Zhang, H. & Zhang, G. (2011) A stable TTI reverse time migration and its implementation. *Geophysics*, 76(3), WA3–WA11. https://doi.org/10.1190/1.3554411

How to cite this article: Louboutin, M. & Herrmann, F.J. (2023) Wave-based inversion at scale on graphical processing units with randomized trace estimation. *Geophysical Prospecting*, 1–14. https://doi.org/10.1111/1365-2478.13405