Towards time-domain extended source waveform inversion

Pengliang Yang\textsuperscript{1} and Wei Zhou\textsuperscript{2}

\textsuperscript{1} Harbin Institute of Technology, China. E-mail: ypl.2100@gmail.com
\textsuperscript{2} KFUPM, Saudi Arabia. E-mail: joey1986@gmail.com

November 14, 2022

Abstract

Full waveform inversion (FWI) updates the subsurface model from an initial model by comparing observed and synthetic seismograms. Despite the potential of FWI for high resolution seismic imaging, this technique is a nonlinear optimization and may be trapped into local minima. Wavefield reconstruction inversion (WRI) relaxes the condition that the wave equation associated with a true physical source does need to be accurately satisfied during the iteration process. Alternatively, extended source waveform inversion (ESI) introduces a non-physical source deployed into the whole computing domain for matching the observed data by synthetic data. The iterative process will move this extended source to the physical source. WRI and ESI are formulated within the same mathematical framework using Lagrangian-based adjoint-state method with a special focus on time-domain formulation using extended sources, while putting connections between classical FWI, WRI and ESI: both WRI and ESI can be viewed as weighted versions of classic FWI. Due to symmetric positive definite Hessian, the conjugate gradient is explored to efficiently solve the normal equation in a matrix free manner, while both time and frequency domain wave equation solvers are feasible.

The most significant challenge comes from the huge storage demand to store time-domain wavefields through iterations. We describe two possible workaround strategies by extracting sparse frequencial wavefields or by considering time-domain data instead of wavefields for reducing such challenge. We suggest that these options should be explored more intensively for tractable workflows.

1 Introduction

Full waveform inversion (FWI), proposed by Lailly (1983) and Tarantola (1984a,b), is a high-resolution seismic imaging technique that iteratively minimizes the misfit between the observed seismic data and the synthetic data modelled according to the wave equation. The non-linearity of FWI objective function demands a good starting model with accurate background information and high-quality data including sufficient low-frequency content and good signal-to-noise ratio to ensure that the numerical optimization converges to the true solution (Bunks et al., 1995). By verifying exactly the wave-equation constraints at the current inexact velocity model, such reduced-model optimization suffers from local minima, preventing convergence to the global minimum for various reasons. For example, once the phase error is larger than half of the local wavelength, FWI will be trapped into a local minimum, which is known as cycle-skipping issue (Virieux and Operto, 2010).

In order to overcome these issues, significant effort has been performed, with the aim to enlarge the basin of the attraction of the misfit function while mitigating the strong dependence of the inversion to an accurate initial model. In order to obtain a reliable initial macro-model before moving to a high-resolution reconstruction, Luo and Schuster (1991) developed the wave-equation travel-time inversion using the correlation between the observed data and synthetic data to measure and decrease time shifts of the seismic phase events, while the pairing is supposed to be correct. Bunks et al. (1995) highlighted the importance of low-frequency content in driving FWI out of local minima, therefore suggesting a multiscale inversion strategy which has become a common practice of the processing workflow in seismic imaging. Shin and Cha (2008) developed the Laplace domain FWI which highlights the early arrivals and damped the late reflections in retrieving a reliable macro model (Shin and Ha, 2008). Dynamic wrapping (Hale, 2013; Ma and Hale, 2013) was proposed to utilize the structural similarity to better identify and match seismic events. This can be achieved at each step of the optimization.

Intensive work by different research groups has been done by modifying the misfit functional to enlarge the width of the convexity valley of the FWI misfit functional. Bozdag et al. (2011) have studied FWI using instantaneous phase and envelope measurements. A deconvolution-based misfit function for FWI, coined as adaptive waveform inversion has been proposed by Warner and Guasch (2014) to mitigate the cycle-skipping issue, which has been shown to be equivalent to the introduction of a matching filter (Zhu and Fomel, 2016). Wu...
et al. (2014) show that the envelope-based misfit functional is also a potential candidate for retrieving reliable velocity models in tomographic style, which could be further improved by cascading a classic FWI. Optimal transport (OT) distance (Engquist et al., 2011; Métivier et al., 2016) measures the data difference in a more global manner (which again means a systematic global search of phase pairing at each iterative step), allowing translation along different axes (time axis, for example) of the data to be compared. This distance, sometimes called Earth Mover’s Distance (EMD), makes the FWI workflow less sensitive to the initial model design.

Another quite often used approach for overcoming local minima in linearized optimization is an arbitrary extension of the search space, moving from a reduced-model approach to a full-space one. van Leeuwen and Herrmann (2013) introduced a wavefield reconstruction inversion (WRI) method, for which the misfit includes the wavefield through the wave equation as the penalty. This relaxes the condition that the wave equation does need to be exactly satisfied through the whole numerical optimization process. The objective is then a two-variables optimization including both physical parameters and wavefield values instead of only physical parameters in the classical FWI. The penalty formulation leads to a normal equation for the wavefield involving the Hessian regularized by penalty parameters. Such equation could be solved in the frequency domain by direct or iterative solvers, thanks to the drastic memory reduction for storing frequencial wavefields. Unfortunately, the use of the time-stepping solver would require, among other issues, a huge storage of time-sampled wavefields (Aghamiry et al., 2020), preventing its application to 3D datasets.

Following the same strategy of space extension and inspired by the idea of differential semblance optimization (DSO) to achieve global convergence (Symes, 1998; Symes, 2015) developed the so-called extended domain FWI (ESI). The key idea is adding extra degrees of freedom which may be unphysical, but effectively fit the observed data through optimization. The misfit design switches from a data landscape to a model landscape, providing alternative strategies for updating the model. The optimization should decrease the contribution of these added degrees in the extended model space to the synthetic data similar to observed data along iterations. The source extension (Huang et al., 2016), offset extension (Fu and Symes, 2017), source and receiver extension (Huang et al., 2017) and many others (Biondi and Almomin, 2013; Barnier and Biondi, 2020), have been shown to be successful in mitigating local minima issues under different settings.

In this paper, we shall formulate WRI and ESI in one framework, using the Lagrangian-based adjoint-state method. By doing so, we shade a new light on some misunderstanding among WRI, ESI, and FWI approaches: the misfit functions of both WRI and ESI can be expressed as the weighted norm of the data residual, where the weighting matrix is parameter dependent. Since the regularized Hessian with penalty parameter prevents direct use of time-stepping solver, we present three approaches to implement extended FWI in the time domain. We made another attempt to design computationally efficient and memory-frugal methods, which fulfills the low memory promises while showing the inconsistency with the theory. This study aims to steer the focus of the community from algorithmic issue to the development of low-memory extended domain methods with good efficiency.

2  

FWI, WRI, and ESI formulations revisited

Although previous attempts have been performed for highlighting connections between these different approaches (van Leeuwen, 2019; Symes, 2020), we present here these three methods in a common notation.

2.1 Least-squares Full Waveform Inversion

Conventional FWI formulation in least-squares sense can be expressed by minimizing the following misfit function with constraints,

$$J_{FWI}(m) = \frac{1}{2} \| Ru - d \|^2 \quad s.t. \quad A(m)u = f,$$

where the quantity $d := d(x_r; t; x_s)$ is the observed data from the receivers in the field, while $u := u(x; t; x_s)$ is the modeled wavefield based on the wave operator $A(m)$ and the source $f; m$ is the target parameter to be inverted. The projection operator $R$ restricts the wavefield $u$ at the receiver location, i.e., $Ru = \int_X u(x; t; x_s) \delta(x - x_r)dx = u(x_r; t; x_s), \quad x \in X, \quad t \in [0, T]$. Consider the 2nd-order acoustic equation with a constant density, we have

$$A(m)u = (\frac{1}{c^2} \partial_t^2 u - A)u = f.$$  (2)

Let us note that the initial condition $(u|_{t=0} = \partial_t u|_{t=0} = 0)$ and boundary condition $(u|_{x \in \partial \Omega} = 0)$ must be satisfied by the wavefield $u$. The minimization of the misfit function with wave equation constraint in [1]
can be translated into an unconstrained optimization through Lagrangian-based adjoint state method (Plessix 2006). The Lagrangian function for the least-squares FWI is specified by

\[ \mathcal{L}_{FWI}(m, u, \lambda) = \frac{1}{2} \| Ru - d \|^2 + \langle \lambda, A(m)u - f \rangle_{X \times [0,T]}, \]

where the Lagrangian multiplier \( \lambda \) is also call the adjoint wavefield. Here, we have equipped with the inner product between two functions \( h(x, t) \) and \( g(x, t) \) defined as

\[ \langle h, g \rangle_{X \times [0,T]} = \int_0^T \int_X dxh(x,t)g(x,t). \]

2.2 Wavefield Reconstruction Inversion

In order to expand the basin of the attraction of the misfit functional, van Leeuwen and Herrmann (2013) proposed wavefield reconstruction inversion (WRI) by taking the wave equation for penalty, leading to the following misfit function

\[ J_{WRI}(m, u) = \frac{1}{2} \| Ru - d \|^2 + \frac{\beta}{2} \| A(m)u - f \|^2, \]

where both model parameters \( m \) and the wavefield \( u \) are considered as parameters of the optimization problem. During the iterations, \( \beta \) should be increased such that the wavefield \( u \) converges to the solution of the wave equation \( (Au = f) \) as \( \beta \to \infty \). The main deviation of WRI compared with the classic FWI is that the wave equation is a penalty and therefore does not need to be fully satisfied in WRI, while classic FWI requires that the wavefield satisfies wave equation exactly. van Leeuwen and Herrmann (2013) have shown that WRI is more immune to cycle skipping by reducing the existence of possible local minima.

The WRI misfit functional with penalty can be rewritten as

\[ J_{WRI}(m, u) = \frac{1}{2} \| \sqrt{\beta} A(m) \| u - \left( \frac{d}{\sqrt{\beta} f} \right) \|^2, \]

and the solution of \( u \) should be obtained by solving normal equation due to \( \partial J/\partial u = 0 \):

\[ \langle \beta A^H A + R^H R \rangle u = R^H d + \beta A^H f, \]

where the symbol \( ^H \) denotes the adjoint (Hermitian transpose) operator. It is therefore natural to see that the above augmented wave equation could be solved in frequency domain by direct or iterative solver. However, it prevents the use of time-stepping wave equation solver, since the penalty parameter \( \beta \) poses significant challenges for time-domain WRI under penalty-based formulation.

Let us introduce a new field \( q(x, t; x_s) \) to represent the mismatch of the wave equation based on the given source, that is,

\[ q := A(m)u - f. \]

The misfit then becomes

\[ J_{WRI}(m, u, q) = \frac{1}{2} \| Ru - d \|^2 + \frac{\beta}{2} \| q \|^2. \]

Using the standard adjoint-state method (Plessix 2006) with the equality (8) as the constraint, we have the Lagrangian

\[ \mathcal{L}_{WRI}(m, u, q, \lambda) = \frac{1}{2} \| Ru - d \|^2 + \frac{\beta}{2} \| q \|^2 + \langle \lambda, A(m)u - f - q \rangle_{X \times [0,T]}, \]

The gradient of the Lagrangian with respect to the model parameter is

\[ \frac{\partial \mathcal{L}_{WRI}}{\partial m} = \frac{\partial \mathcal{L}_{WRI}}{\partial u} \frac{\partial u}{\partial m} + \frac{\partial \mathcal{L}_{WRI}}{\partial q} \frac{\partial q}{\partial m} + \frac{\partial \mathcal{L}_{WRI}}{\partial \lambda} \frac{\partial \lambda}{\partial m} + \langle \lambda, \frac{\partial A}{\partial m}u \rangle_{[0,T]}. \]

Zeroring the derivative of \( \mathcal{L} \) with respect to variable \( \lambda \), i.e. \( \partial \mathcal{L}/\partial \lambda = 0 \) ensures the forward equation (8). Nullifying the derivative of \( \mathcal{L} \) with respect to state variables \( u \) and \( q \) \( (\partial \mathcal{L}/\partial u = \partial \mathcal{L}/\partial q = 0) \) leads to two adjoint equations

\[ \begin{align*}
\frac{\partial \mathcal{L}}{\partial u} &= 0 \quad \Rightarrow \quad R^H (Ru - d) + A^H \lambda = 0 \\
\frac{\partial \mathcal{L}}{\partial q} &= 0 \quad \Rightarrow \quad \beta q - \lambda = 0.
\end{align*} \]
At the saddle point, the gradient of the misfit functional with respect to model parameters \( m \) will be given by
\[
\partial_m J_{\text{WRI}} = \partial_m \mathcal{L} = \langle \lambda, \partial_A \partial_m u \rangle_{[0,T]}.
\] (14)

According to (13), we can insert \( q = \lambda/\beta \) into (8)
\[
Au = f + \frac{1}{\beta} \lambda.
\] (15)

Multiplying \( A^H \) on both sides of the above equation while plugging in the expression in (12) recovers the normal equation (7). From (15)
\[
\lambda = \frac{\beta}{\lambda} (Au - f),
\]
we then have another expression for the gradient expression in (14)
\[
\partial_m J_{\text{WRI}} = \frac{\beta}{\lambda} \sum_{\omega} \langle Au - f, \partial_A \partial_m u \rangle_{[0,T]},
\] (16)

where we have applied Parseval’s theorem to translate from time domain to frequency to obtain the last equality. When the squared slowness is chosen as the model parameter \( m = 1/v^2 \), we have \( \partial A/\partial m = -\omega^2 \) in frequency domain. It turns out to be the same as the one from van Leeuwen and Herrmann (2013, their equation 7) except the scaling constant \( \beta \).

Let us define the data residual \( \delta d := d - Ru \) and the operator \( S := RA^{-1} \). From equations (12) and (13), we obtain
\[
q = \frac{1}{\beta} \lambda = \frac{1}{\beta} A^{-H} R^H (d - Ru) = \frac{1}{\beta} S^H \delta d.
\] (17)

The misfit functional then becomes
\[
J_{\text{WRI}} = \frac{1}{2} \| \delta d \|^2 + \frac{1}{2\beta} \| S^H \delta d \|^2 = \frac{1}{2} \delta d^H \left( I + \frac{1}{\beta} SS^H \right) \delta d = \frac{1}{2} \| \delta d \|_{W_1}^2,
\] (18)

which clearly shows that the WRI approach is still a least-squares FWI based on weighted data residual. This has been concluded by van Leeuwen (2019) and Symes (2020, equation 22).

Note the forward equation implies \( u = A^{-1} (f + q) \). The misfit function of WRI also reads
\[
J_{\text{WRI}} = \frac{1}{2} \| Ru - d \|^2 + \frac{\beta}{2} \| q \|^2 = \frac{1}{2} \| Sq - (d - Sf) \|^2 + \frac{\beta}{2} \| q \|^2,
\] (19)

leading to the normal equation for the extended source
\[
(S^H S + \beta I) q = S^H (d - Sf).
\] (20)

Let us assume that the initial background model can reproduce the first arrivals (diving waves) by modelling using physical source \( f \) \( (Sf = RA^{-1} f) \), then the remaining data \( d - Sf \) are mainly reflections. Under such a hypothesis, the objective in (19) can be interpreted as reflection waveform inversion based on the penalty formulation.

Although Peters et al. (2014); Esser et al. (2015) have highlighted that WRI is different from FWI based on the adjoint-state method, such approach still does not overcome local minima issues, even if the misfit landscape is different. The above formulation illustrates that WRI is a specific form of FWI which is consistent with the adjoint-state method: the resulting gradient expressions are equivalent as well. Since WRI is equivalent to classic FWI using weighted data residual, it naturally cycle-skips. A simple 1D example has been given by Symes (2020), besides many positive stories of WRI (van Leeuwen and Herrmann, 2013; Peters et al., 2014; Aghamiry et al., 2020; Aghamiry et al., 2021) have shown that WRI is mathematically equivalent to contrast source inversion (Abubakar et al., 2009), which is known to have local minimum issue (Van den Berg and Abubakar, 2001).

### 2.3 Extended Source Inversion

The key idea of extended source waveform inversion (ESI) proposed by Symes (2015) is slightly different from the one embedded into the WRI. An extended source \( q(x,t;x_s) \) is introduced such that
\[
q := Au,
\] (21)
which is almost identical to equation (8) in WRI except that the physical source \( f \) has been replaced by the extended source \( q \). This leads to a source-independent inversion scheme. The corresponding misfit functional is defined as (Huang et al., 2018)

\[
J_{ESI} = \frac{1}{2}\|Ru - d\|^2 + \frac{\beta}{2}\|Bq\|^2,
\]

where the penalty term is different from the one in WRI misfit (9) as it involves a weighting operator \( B \) - the so-called annihilator named by Symes (2008). Possible options such as \( B = \|x - x_a\| \) and \( B = t \) in the form of pseudo-differential operator are propositions to achieve the global convergence (Symes et al., 2020). The resulting Lagrangian is

\[
\mathcal{L}(m, u, q, \lambda) = \frac{1}{2}\|Ru - d\|^2 + \frac{\beta}{2}\|Bq\|^2 + \langle \lambda, Au - q \rangle_{X \times [0, T]}. \tag{23}
\]

The gradient of the Lagrangian with respect to the model parameter \( m \) is then

\[
\frac{\partial \mathcal{L}}{\partial m} = \frac{\partial \mathcal{L}}{\partial u} \frac{\partial u}{\partial m} + \frac{\partial \mathcal{L}}{\partial \lambda} \frac{\partial \lambda}{\partial m} + \frac{\partial \mathcal{L}}{\partial q} \frac{\partial q}{\partial m} + \langle \lambda, \frac{\partial A}{\partial m}u \rangle_{[0, T]}. \tag{24}
\]

Following the same procedure, we set \( \frac{\partial \mathcal{L}}{\partial u} = \frac{\partial \mathcal{L}}{\partial \lambda} = \frac{\partial \mathcal{L}}{\partial q} = 0 \), yielding

\[
Au = q \tag{25a}
\]

\[
A^H \lambda = R^H (d - Ru) \tag{25b}
\]

\[
\lambda = \beta B^H Bq. \tag{25c}
\]

At the saddle point, both the forward and adjoint equations are satisfied. This allows the gradient of the misfit functional for ESI to be computed via

\[
\partial_m J_{ESI} = \langle \lambda, \frac{\partial A}{\partial m}u \rangle_{[0, T]}. \tag{26}
\]

According to (25b) and (25c), we can express the extended source as

\[
q = (\beta B^H B)^{-1} \lambda = (\beta B^H B)^{-1} A^{-} R^H (d - Ru) = (\beta B^H B)^{-1} S^H \delta d, \tag{27}
\]

where we assume \( \beta B^H B \) is invertible. It leads to the misfit functional of ESI expressed as

\[
J_{ESI} = \frac{1}{2}\|\delta d\|^2 + \frac{\beta}{2}\|B(\beta B^H B)^{-1} S^H \delta d\|^2 = \frac{1}{2}\|\delta d\|^2 + \frac{1}{2}\|S(\beta B^H B)^{-1} S^H \delta d\|^2 = \frac{1}{2}\|\delta d\|^2 + \frac{1}{2}\|Sd\|^2_{W_2}, \tag{28}
\]

which is also a weighted-data residual as in classic FWI, and may also suffer from the curse of local minima, such as the cycle-skipping issue. Eliminating \( \lambda \) and \( u \) gives the normal equation in the same form as Huang et al. (2018, equation 10)

\[
(S^H S + \beta B^H B)q = S^H d \tag{29}
\]

This equation shares the similarity to the one obtained by WRI: the regularized Hessian has to be inverted for the solution of \( q \), preventing the use of a time-domain wave equation solver. This is the reason why Huang et al. (2018) have conducted frequency-domain ESI.

The above development puts classic least-squares FWI, WRI and FWI into the same framework thanks to the adjoint state method. This allows us to better catch the similarity and distinction among them. Table 1 lists the expressions of the misfit function, the Lagrangian, the adjoint source and the normal equation for least-squares FWI, WRI and ESI for comparison.

| Method | Misfit \((J)\) | Lagrangian \((\mathcal{L})\) | Adjoint field \(\lambda\) | Normal equation |
|--------|----------------|-----------------|-----------------|----------------|
| FWI    | 0.5\|Ru - d\|^2 | \(J + \langle \lambda, Au - f \rangle\) | \(S^H \delta d\) | - |
| WRI    | 0.5\|Ru - d\|^2 + 0.5\|q\|^2 | \(J + \langle \lambda, Au - f - q \rangle\) | \(S^H \delta d(= \beta q)\) | \(S^H S + \beta I)q = S^H (d - S f)\) |
| ESI    | 0.5\|Ru - d\|^2 + 0.5\|Bq\|^2 | \(J + \langle \lambda, Au - q \rangle\) | \(S^H \delta d(= \beta B^H Bq)\) | \(S^H S + \beta B^H Bq = S^H d\) |

Table 1: The misfit function, Lagrangian, adjoint field and normal equations for FWI, WRI and ESI.
2.4 The role of the penalty parameter

Let us highlight the important role of the penalty parameter $\beta$ in WRI and ESI. The misfit function is the summation over the data error $J^e$ and the penalty $J^p$ scaled by the penalty coefficient $\beta$:

$$J_{WRI/ESI} = J^e + \beta J^p. \quad (30)$$

Fu and Symes (2017) have shown that

$$\frac{dJ^e}{d\beta} \geq 0, \quad \frac{dJ^p}{d\beta} \leq 0, \quad (31)$$

implying the misfit $J^e$ is a monotonic increasing function of $\beta$, while the misfit $J^p$ is a monotonic decreasing function of $\beta$. According to Fu and Symes (2017), with small $\beta$ the objective can converge globally from any initial guess, but may converge very slowly when approaching the minimum. Assuming a good initial guess, large $\beta$ will force the extended wave equation to be satisfied better and give the penalty term small contribution to the misfit functional. In short, the tuning of $\beta$ should start from small value and increase through iterations. Based on the discrepancy principle, Fu and Symes (2017) designed an automatic scheme to choose a proper $\beta$, starting the outer iterations for nonlinear optimization from $\beta = 0$ (this implies the knowledge of the noise level for the data fitting must be specified). It is noteworthy that the discrepancy principle applies equally well to WRI.

3 Solving normal equation by conjugate gradient method

As can be seen above, the regularized Hessian prevents direct use of time stepping solvers for WRI and ESI. To move WRI into time domain, the first workaround was given by Wang et al. (2016): the idea is to use approximate wavefield by neglecting higher order terms, so that the quantities are computable via the solution of wave equation using time stepping. Another attempt was made by Rizzuti et al. (2020) based on a dual formulation, however, similar approximation has to be introduced in order to make computation feasible, while twice of the modeling is required to compute all field quantities. Instead of solving the normal equation for the wavefield, Gholami et al. (2020) proposed to solve an equivalent normal equation for a data vector.

Unfortunately, the key problem introduced by penalty formulation still stands. To proceed, they made drastic approximation by introducing a scalar to replace the inverse of the regularized Hessian, while assuming the Green’s function between two iterates are invariant. By adding the quadratic term associated with the penalty, however, similar approximation has to be introduced in order to make computation feasible, while

$$\alpha \approx \frac{1}{2 \beta B}.$$ \quad (32)

and to solve it efficiently by linear conjugate gradient (CG) method in a matrix free fashion, as presented in Algorithm 1:

- The right hand side, $b = S^H d = A^{-H} R^H d$, can be computed by first projecting the data $d$ back to the wavefield $(R^H d)$, and then performing an adjoint modelling to get the adjoint field $b$.

- The key to CG method is how to perform the matrix vector product. Given an input vector $p$, the matrix vector product, $Ap = S^H Sp + \beta B^H B p$, can be computed by summing over two parts. The first part, $S^H Sp = A^{-H} R^H RA^{-1} p$, is obtained by performing two modelings: first do a forward modelling using $p$ as the source; then extract the data at receiver, and project it back to the wavefield; finally, use the extracted data to do an adjoint modelling. The evaluation of the second part, $\beta B^H B p$, is straightforward and could be added to the first part.

- The solution of the unknown vector after the convergence of CG serves as the extended source: $z := q$.

- In terms of the storage request, each CG iteration needs at least four vectors (the solution vector $z$, direction vector $p$, residual vector $r$ and the matrix vector product $Ap$) to drive the algorithm into the next iteration. Here these vectors are built by the wavefields at every time step at every grid point, demanding huge amount of storage requirement. For the moment, we consider the strategy by disk IO, and discuss two possible alternatives after the numerical test.

Algorithm 1:

- The right hand side, $b = S^H d = A^{-H} R^H d$, can be computed by first projecting the data $d$ back to the wavefield $(R^H d)$, and then performing an adjoint modelling to get the adjoint field $b$.

- The key to CG method is how to perform the matrix vector product. Given an input vector $p$, the matrix vector product, $Ap = S^H Sp + \beta B^H B p$, can be computed by summing over two parts. The first part, $S^H Sp = A^{-H} R^H RA^{-1} p$, is obtained by performing two modelings: first do a forward modelling using $p$ as the source; then extract the data at receiver, and project it back to the wavefield; finally, use the extracted data to do an adjoint modelling. The evaluation of the second part, $\beta B^H B p$, is straightforward and could be added to the first part.

- The solution of the unknown vector after the convergence of CG serves as the extended source: $z := q$.

- In terms of the storage request, each CG iteration needs at least four vectors (the solution vector $z$, direction vector $p$, residual vector $r$ and the matrix vector product $Ap$) to drive the algorithm into the next iteration. Here these vectors are built by the wavefields at every time step at every grid point, demanding huge amount of storage requirement. For the moment, we consider the strategy by disk IO, and discuss two possible alternatives after the numerical test.
The CG method does not require the explicit expression of $(B^H B)^{-1}$. The method converges for any $\beta \geq 0$. This is advantageous over some other possibilities such as Gauss-Seidel iteration and recursion assisted by surrogate function [Aghamiry et al. 2020], see Appendix A. Thanks to the matrix free formulation, the above CG procedure can equally be applied to WRI by simply choosing $B = I$ with a different right hand side. In terms of equation (20), the right-hand-side term of the normal equation for WRI is $b = S^H(d - Sf) = A^{-H} R^H (d - RA^{-1} f)$, which can be computed by first extracting the forward wavefield excited by the true source $f$, then use the data difference $(d - RA^{-1} f)$ as the virtual source to compute the adjoint wavefield $b$. Both time and frequency domain wave equation solvers can be applied when solving normal equation using CG.

Algorithm 1 Conjugate gradient algorithm for solving $A x = b (A^H = A)$

1: $x_0 := 0$
2: $r_0 := b - A x_0 = b$
3: $p_0 := r_0$
4: for $k = 0, \cdots, N_{CG} - 1$ do
5: $\alpha_k = \frac{r_k^H r_k}{p_k^H A p_k}$
6: $x_{k+1} = x_k + \alpha_k p_k$
7: $r_{k+1} = r_k - \alpha_k A p_k$
8: if $||r_{k+1}|| < tol \cdot ||r_0||$ or $k == N_{CG} - 1$ then
9: exit loop
10: end if
11: $\beta_k = \frac{r_{k+1}^H r_{k+1}}{r_k^H r_k}$
12: $p_{k+1} = r_{k+1} + \beta_k p_k$
13: end for

A special case comes with $\beta = 0$. Then the normal equation becomes

$$S^H S q = S^H d.$$  \hspace{1cm} (33)

When $q$ is restricted to the true source location, finding $q$ based on the given data $d$ is exactly the source wavelet estimation process, which is solvable using CG since $S^H S$ is SPD. In this case, the problem degenerates to classic FWI: we simply use the observed data $d$ to estimate a wavelet $q$ at the true source location, then minimize the misfit function iteratively.

The optimization is consequently performed using the variable projection method [Golub and Pereyra 1973] due to dramatic difference in the sensitivity of the misfit with respect to extended sources $q$ and model parameters $m$. There are nested loops with the nonlinear outer loop for searching model parameters $m$ and the linear inner loop for updating extended source $q$ [Huang et al. 2018]. In the frequency formulation, [Aghamiry et al. 2019] has proposed a linear outer loop over the wavefield $u$ and a linear inner loop over specific model parameters $m$. He has taken benefit from the linear property of the scalar wave equation with respect to the square of slowness (Lamé coefficients would be those parameters for the elastic wave equation). Still the inner loop has to be solved through direct solver or iterative solver in this frequency formulation. In our strategy for time formulation, as sketched in Algorithm 2, the outer loop is the iterative update for model parameters $m$, while the inner loop is the solution of the normal equation with respect to extend source $q$. In order to mitigate the potential non-linearity of the inverse problem, the inner loop for the linear optimization, i.e. the CG iteration, should converge with good precision. The recursion strategy illustrates that the inner loop should be solved quite accurately (Aghamiry et al. 2019).

Algorithm 2 Nested loops for variable projection between $m$ and $q$

1: Given initial model: $m_0$
2: for $i = 0, \cdots, N_{outer} - 1$ (outer loop: $\min_m J_{ESI}(m, q)$) do
3: solve $(S^H S + \beta B^H B)q = S^H d$ using Algorithm 1 (inner loop: $\min_q J_{ESI}(m, q)$)
4: compute forward field: $u = A^{-1} q$
5: compute adjoint field: $\lambda = \beta B^H B q$
6: calculate gradient: $\partial_u J =\sum_{j} \lambda H \frac{\partial A}{\partial m} u$
7: estimate descent direction: $\delta m = - \tilde{H}^{-1} \partial_u J$ ($\tilde{H}$=approximate Hessian by L-BFGS algorithm)
8: find a good step size $\alpha$ by line search to update model $m_{i+1} := m_i + \alpha \delta m$
9: end for
4 Numerical tests

We now numerically access the feasibility of time-domain WRI and ESI using CG method. Schematically, we always first invert for the extended source, then deduce the forward and adjoint fields to build the gradient of the misfit functional. It should be highlighted that the first stage in conventional WRI actually focuses on inverting the wavefield $u$ instead of extended source $q$, which is different from what the WRI computing scheme presented here. In all the numerical experiments as follows, we use 10 CG iterations to solve the normal equation for WRI and ESI. The annihilator is chosen to be $B = \|x - x_s\|$.

4.1 Data fitting in reflection case

We start with a 2D synthetic model including two layers of different velocities (1500 m/s and 1800 m/s), as shown in Figure 1a. The initial model displayed in Figure 1b is homogeneous with velocity equal to the first layer of the true model. The locations of the source and the receiver has been marked up in Figure 1b. The purpose of this test is to understand the data fitting and the gradient of FWI, WRI and ESI.

The observed data from the true model has been displayed in Figure 2a. Since the velocity of the initial model is the same as the true model, the synthetic data in classic least-squares FWI reproduces direct arrivals of the observed data perfectly; however, the reflections from the interface between two layers are completely missing, as can be seen from Figure 2b. Figures 2c and 2d shows that both WRI and ESI generate data very close to observed data, mimicking that the simulation is performed in the true model such that the data can be fitted almost perfectly. By domain extension and recursive data assimilation, the information from the observed data are added back to the modelled data, behaving that the true physics has been taken into account.

As displayed in Figure 2a, classic FWI only considers information in the starting model; thus is blind to the information in the data, which is missing in the starting model, leading to the absence of reflection events that should be present by modeling in the true model.

The gradients from FWI, WRI and ESI have been plotted in Figure 3. The FWI gradient in Figure 3a shows better source receiver symmetry in the banana donut of the energy isochrone, while the ESI gradient in Figure 3c shows more energy comes from receiver side, including a lot of low frequency energy in the first Fresnel zone. Figure 3b is in between the previous two, with more energy from receiver side than the source side. This is quite understandable as WRI uses both the physical source and the extended source. In both WRI and ESI, the extended source extracts information based on the data from the receiver location.

4.2 Misfit landscape in transmission case

In order to better understand the capability of WRI and ESI methods in getting rid of spurious local minima inherent in classic FWI, we now focus now the study of the landscape of the misfit function, based on a simple 1D layered model shown in Figure 4. We consider the surface acquisition plus two vertical lines of receivers distributed on two sides of the model. The source has been deployed at the top left corner coinciding with the receiver position, such that there are sufficient transmission energy. We first generate a number initial
velocity model according to the formula \( m_0 = (1 + \epsilon)m_{\text{true}} \) \((-0.5 \leq \epsilon \leq 0.5)\), where \( m_{\text{true}} \) is the true layered model. We then compute the misfit functions by taking different values of the velocity error \( \epsilon \). The basin of the attraction of FWI, WRI and ESI are shown in Figures 5, 6 and 7. We examine the WRI and ESI using a large penalty \( \beta = 100 \) and a small penalty \( \beta = 0.01 \). The classic FWI misfit plotted in Figure 5 shows multiple local minimum, while the convexity domain is rather narrow. The WRI misfit in Figure 6 has larger convexity domain than FWI, but still suffers from local minimum issue, especially when \( \epsilon > +0.2 \) for \( \beta = 100 \). Figure 7 confirms that ESI further extends the search space, possessing wider basin of attraction, compared to WRI and least-squares FWI. It is interesting to note that due to the absence of the knowledge on the physical source, the observed data can never be fit perfectly. Hence, the ESI misfit is not zero even if the input model has no error compared to the true model. For both WRI and ESI, the landscape of misfit function becomes flattened when the penalty is small. This indicates that the penalty must be judiciously chosen, increasing from a small value with wider basin of attraction of the misfit, to a large value for fast convergence when the model gets close to the ground truth.

4.3 Marmousi-II test: preliminary validation of time-domain ESI

Finally, we validate our time-domain ESI implementation based on the Marmousi-II model. As shown in Figure 8a, the starting model has been highly smoothed based on the true model in Figure 8a. The model has the water depth of 500 m. We consider a surface acquisition geometry: 28 sources are evenly distributed at 25 m of the water depth, while 641 receivers are uniformly deployed at the same depth. The observed data was generated using a Ricker wavelet of 5 Hz peak frequency. The tuning of the penalty parameter \( \beta \) was automatically determined through iterations using discrepancy principle.

We observed that our CG algorithm enjoys a fast convergence rate, as can be seen from Figure 9. Figure 10 computed the amplitude of the extended source at 2 Hz for the 1st shot in the 1st CG loop, using discrete time Fourier transform. It illustrates that the extended source varies dramatically at early iterations, while reach a steady state after a number of CG iterations: the energy of extended source indeed focuses quickly to the true source location (the 1st shot resides in the top left part of the model) through CG iterations, even though the synthetic data in ESI inversion is purely deciphered from the recursion process according to the observed data without the knowledge of the true physical source.

We managed to complete 6 iterations of the ESI inversion. The updated models after limited number of iterations are displayed in Figure 11, showing that the model is evolving to the true model. Unfortunately, the whole inversion still converges very slowly so that we have to terminate it on the way, since the 6 updates of the 2D test has consumed more than 72 hours of CPU time due to intensive computational cost and slow IO in order to read and write huge amount of wavefield snapshot to the disk. This motivates us to focus on the...
Figure 3: The sensitivity kernels of (a) Classic FWI, (b) WRI and (c) ESI. The sensitivity kernels of ESI is one-sided, significantly different from least-squares FWI and WRI.
Figure 4: The 1D layered velocity model used for generating the initial model to study the variation of the misfit function with respect to velocity errors. The dashed line indicates the location of the receivers, while the source coincides with the receiver in the top left corner.

Figure 5: The variation of the misfit functional with respect to velocity errors for classic FWI.
Figure 6: The variation of the misfit functional with respect to velocity errors for WRI

Figure 7: The variation of the misfit functional with respect to velocity errors for ESI
Figure 8: Marmousi-II test: (a) True model; (b) initial model.

Figure 9: The convergence history of CG for ESI in the 1st CG loop.

development of low storage ESI approach without disk IO.

5 Mitigating the storage demand

Algorithmically speaking, the above methods should work nicely, as long as the wavefield vector can be stored. This can be doable for frequency domain approach, since very few frequencies are needed. However, in the time domain, the wavefield at every time step has to be stored in memory or on disk before the two parts of the matrix vector product are assembled. Consequently, the huge storage requirement becomes one of the major bottlenecks to push time-domain WRI and ESI forward to 3D practical applications. Other bottlenecks such as nested loop for expensive computation and the need for good preconditioning are out of scope of the discussion for the moment. A natural idea is to compress the wavefield, which is however not practical for 3D large scale applications: Most of the lossless compression technique cannot achieve an extremely high compression ratio such that the time domain wavefields fits into memory. Our numerical experience shows also that the lossy compression technique breaks the orthogonality of the residual vectors between two iterates, making CG algorithm infeasible to work. We suspect similar difficulties will be encountered in Gauss-Seidel iterations and the recursion scheme assisted by surrogate function.

We have made two attempts to bypass the storage issue while still using efficient time domain modelling (the ability to model multiple frequencies in one go without heavy matrix store and solve): the data space approach and frequency domain inversion by time domain modelling. We discuss them here to highlight the significant
Figure 10: The evolution of extended source at 2 Hz for the 1st shot in the 1st CG loop using discrete Fourier transform. Red: amplitude=0; Green: high amplitude.
Figure 11: The model updated at different iterations.
challenge to solve the memory issue, even though both proposals are not successful.

## 5.1 Switch from wavefield to data space

Let us define the synthetic data extracted at receiver location as \( w := Ru \). Then we have

\[
    w = Ru = RA^{-1} q \quad (u = A^{-1} q \text{ from equation } 25a)
\]

\[
    = Sq = S(\beta B^H B)^{-1}\lambda \quad (q = (\beta B^H B)^{-1}\lambda \text{ from equation } 25c)
\]

\[
    = S(\beta B^H B)^{-1}A^{-H}R^H(d - Ru) \quad (\lambda = A^{-H}R^H(d - Ru) \text{ from equation } 25b)
\]

\[
    = S(\beta B^H B)^{-1}S^H(d - w) \quad (S = RA^{-1}, w = Ru)
\]

yielding

\[
    (I + S(\beta B^H B)^{-1}S^H)w = S(\beta B^H B)^{-1}S^H d
\]

\[
\Leftrightarrow (\beta I + S(B^H B)^{-1}S^H)w = S(B^H B)^{-1}S^H d
\]

where we have defined

\[
    v := S(B^H B)^{-1}S^H d
\]

so that equation (35) becomes

\[
(\beta I + S(B^H B)^{-1}S^H)w = v.
\]

The data vector \( v \) can be computed by first performing an adjoint modelling using \( d \) as the adjoint source, then apply the operator \((B^H B)^{-1}\) on the adjoint field, and finally extract the data at receiver locations after a forward modelling using scaled adjoint field. Thanks to the symmetric positive definite matrix \((\beta I + S(B^H B)^{-1}S^H)\), the solution of linear system for \( w \) in (37) can also be found using conjugate gradient method in a matrix free manner. Once the synthetic data \( w \) is computed, the adjoint field can be computed according to equation (25b). Then the extended source is obtained from equation (25c): \( q = (\beta B^H B)^{-1}\lambda \). The field \( u \) is immediately available after a forward modelling based on the extended source \( q: u = A^{-1}q \). Finally, the ESI gradient is built by cross-correlation between the forward and the adjoint fields via (26).

In fact, the above computing workflow can also be discovered according to Sherman-Morrison-Woodbury (SMW) formula (Golub 1996, p.51, equation 2.1.4)

\[
(P + UV^H)^{-1} = P^{-1} - P^{-1}(I_k + V^H P^{-1}U)^{-1}V^H P^{-1},
\]

where \( P \in \mathbb{R}^{n \times n} \) is invertible while \( U \in \mathbb{R}^{n \times k}, V \in \mathbb{R}^{n \times k} \). Assume \( \beta B^H B \) is invertible. We can express the inverse of the normal operator in the following:

\[
(S^H S + \beta B^H B)^{-1} = (\beta B^H B)^{-1} - (\beta B^H B)^{-1}S^H (I + S(\beta B^H B)^{-1}S^H)^{-1}S(\beta B^H B)^{-1}.
\]

The extended source is then

\[
q = (S^H S + \beta B^H B)^{-1}S^H d
\]

\[
= (\beta B^H B)^{-1}S^H d - (\beta B^H B)^{-1}S^H (I + S(\beta B^H B)^{-1}S^H)^{-1}S(\beta B^H B)^{-1}S^H d
\]

\[
= (\beta B^H B)^{-1}S^H \left( d - \underbrace{(\beta I + S(B^H B)^{-1}S^H)^{-1}S(B^H B)^{-1}S^H d}_{v} \right)
\]

The dimensionality of \( w \) living in data space is much lower than the dimensionality of the wavefield space in the time domain. It can therefore be directly stored in memory. Finally, the extended source will be given by

\[
q = (\beta B^H B)^{-1}S^H (d - w).
\]

Since the synthetic data \( w \) reads

\[
w := (\beta I + S(B^H B)^{-1}S^H)^{-1}S(B^H B)^{-1}S^H d,
\]

it is therefore natural to see that the synthetic data in ESI is able to fit the observed data precisely as \( w \to d \) (\( \beta \to 0 \)).
A pure time domain implementation of the above algorithm is still difficult due to the need to access the adjoint field as the source for forward modelling (applying operator $S$ after application of $(B^HB)^{-1}S^H$). The scaled adjoint field has opposite direction in time during the simulation compared to the forward field. This creates also challenges due to gigantic memory overhead and immense amount of IO traffic, particularly for 3D practical applications. In fact, building the FWI gradient faces up to the same issue as the recursion process. For ESI in non-attenuating medium, we can reconstruct the wavefield on the fly by decimation and interpolation of the stored boundaries (Yang et al., 2016b). For ESI in attenuative medium we may have to resort to the optimal checkpointing (Symes, 2007) or CARFS algorithm (Yang et al., 2016a), since recomputing wavefield by decimation and interpolation of the stored boundaries is no more stable.

The above computational strategy applies equally to WRI: what we need is just replacing $d$ with $d - Sf$ and set $B = I$ in computing $v$ and $q$. That is,

$$q = (S^H S + \beta I)^{-1}S^H (d - Sf) = \beta^{-1} S^H (d - Sf) - \beta^{-1} S^H (I + \beta^{-1} S S^H)^{-1} \beta^{-1} S S^H (d - Sf)$$

where

$$w = (\beta I + S S^H)^{-1} S S^H (d - Sf),$$

yielding

$$w = \beta^{-1} S (B^H B)^{-1} S^H (d - Sf - w) = S (\beta B^H B)^{-1} S^H (d - Sf - w) = Sq = Ra^{-1}(Au - f) = Ru - Sf$$

so that the synthetic data in WRI is $Ru = w + Sf$, and the data residual is $\delta d = d - Ru = d - Sf - w$ which can also be very small as $w \rightarrow (d - Sf)$ if $\beta \rightarrow 0$ according to (44). The wavefield $u$ in WRI is obtained from the forward modelling based on the source $f + q$ instead of only $q$.

By choosing the annihilator $B = I$ and the penalty parameter $\beta = 10^{-2}$, we have experimented the above data space approach based on the Marmousi-II model shown above. We compare the synthetic data $w$ after CG iterations in the data space with the observed data, which looks quite similar, as shown in Figure 12. This is in agreement with the theory, as we expect ESI can always fit the observed data. Using this synthetic data, we then can compute the data residual and therefore the extended source $q$. After a new forward modelling using the extended source, we obtain the modelled forward field $u$. It is then natural to extract again the synthetic data at receivers which is then denoted as $\tilde{Ru}$. To our surprise, the extracted data $\tilde{Ru}$ is significantly different from the synthetic data $w$ after CG iterations. The amplitude of $\tilde{Ru}$ is several orders of magnitude higher than $w$, which is expected to be the same as $Ru$. We attribute this problem to the numerical instability of the application of the SMW formula. Hao and Simoncini (2020) have pointed out that “the updating method cannot be expected to be numerically stable in all cases. In particular, problems will arise when the initial problem is more ill-conditioned than the modified one”.

5.2 Frequency domain inversion by time domain modelling

The second idea is to do frequency domain inversion with time domain modelling, inspired by Sirgue et al. (2008). The main merit of this approach is that all (forward and adjoint) wavefields needed are stored at few frequencies, while the time domain wavefield can be computed on the fly during time stepping, via discrete time Fourier transform (DTFT):

$$u(x, f_m) = \sum_{n=0}^{N_t-1} u(x, n\Delta t) \exp(j2\pi f_m n\Delta t), \quad m = 1, \cdots, N_f,$$

where $f_m$ refers to the $m$-th discrete frequencies. The modelling of the adjoint field in the time domain requires the adjoint source (the data residual) as a time series, meaning that all the frequencies are needed. It is unfortunate that the data and the wavefield are only accessible at limited number of discrete frequencies. To supply such an adjoint source $s$, we use a time series computed by inverse DTFT based on few discrete frequencies available:

$$s(x, n\Delta t) = \frac{1}{N_t} \sum_{m=1}^{N_f} s(x, f_m) \exp(-j2\pi f_m n\Delta t), \quad n = 0, \cdots, N_t - 1.$$
Note that the number of discrete frequencies $N_f$ is much less than total number of time steps $N_t$ in the simulation. The stringent storage demand is therefore resolved.

Indeed, equation (47) should be understood as the adjoint of the DTFT scaled by the factor $1/N_t$, which makes the CG feasible. One may consider apply the same scaling factor $1/\sqrt{N_t}$ for both forward DTFT and its inverse, to make it fully adjoint. This is in fact unnecessary as each CG iteration requires one matrix vector product. The matrix vector product operates on a frequency domain wavefield, which is first transformed into time domain by inverse DTFT using (47), then performs time domain adjoint and forward modelling, and finally the forward field are transformed back into frequency domain using DTFT (46). Although the constant scaling factor $1/N_t$ only appears in inverse DTFT, the resulting matrix vector product will be the same as normalizing DTFT and inverse DTFT with the same factor $1/\sqrt{N_t}$.

Let us consider a frequency domain formulation of ESI.

$$J = \frac{1}{2} \|d - FRu\|^2 + \frac{\beta}{2} \|Bq\|^2 = \frac{1}{2} \|d - FSF^Hq\|^2 + \frac{\beta}{2} \|BF^Hq\|^2$$

such that the normal equation is given by

$$((FSF^H)^HFSF^H + \beta(BF^H)^HBF^H)\bar{q} = (FSF^H)^H\bar{d}$$

Due to the fact that $F^H = I_{N_f \times N_f}$ and the spatial operator $B$ and the temporal operator $F$ commute, we immediately reduce the above normal equation to

$$(FSF^H + \beta BF^H B)\bar{q} = FSF^H \bar{d}$$

Let us denote DTFT in (46) as $F \in \mathbb{C}^{N_f \times N_t}$ and the inverse DTFT in (47) as $F^H \in \mathbb{C}^{N_t \times N_f}$. The solution of normal equation using CG can then be understood as applying an additional DTFT operator on both sides of the equation (50).

$$F(S^H S + \beta B^H B)^F H q = F S^H F^H \bar{d}$$

where the time domain wavefield and the observed data in (29) have been represented by frequency domain counterpart through inverse DTFT, that is,

$$q(x,t) = F^H \bar{q}(x,f), \quad d(x_r,t) = F^H \bar{d}(x_r,f).$$

It is clear that $F(S^H S + \beta B^H B)^F H$ is SPD, and therefore can be solved using CG. In particular, when the number of frequencies are the same as the number of time steps ($N_f = N_t$), the solution of normal equation in (51) using CG becomes standard preconditioned CG (Nocedal and Wright, 2006, section 5.1 preconditioning).
Once the extended source is obtained, we immediately obtain the gradient of the misfit function by applying Parseval’s theorem to (25):

$$\partial_m J_{ESI} = \Re \sum_\omega \lambda H \frac{\partial A}{\partial m} u, \quad (52)$$

where the forward field $u = A^{-1} q$ should be computed by a forward modelling using the estimated extended source after CG, while the adjoint field is given by $\lambda = \beta B^H B q$. A pure frequency domain implementation through the solution of Helmholtz equation has been done by Huang et al. (2018), while frequency domain inversion using time domain modelling is another alternative to achieve the same goal thanks to DTFT (Sirgue et al., 2008).

As a proof of concept, we generate a time domain signal $x(t) = \sum_{m=1}^{N_f} \sin(2\pi f_m t), t \in [0, 1]$ up to $N_t = 200$ temporal samples, where $N_f = 5$ frequencies are specified, $[f_1, f_2, f_3, f_4, f_5] = [1, 3, 5, 7, 9]$ Hz. A Toeplitz matrix of size $N_t \times N_t$ has been created to mimic the role of regularized Hessian operator in the time domain. We can then construct the right hand side of linear system $b = Ax$. By converting $b$ into frequency domain using DTFT, we use CG to solve the linear inverse problem to obtain a frequency domain signal which is only of length $N_f = 5$ rather than $N_t = 200$. Finally, we convert the frequency domain solution back to time domain. The reconstructed signal matches the true one perfectly in the time domain, as can be seen in Figure 13.

6 Conclusion

This paper formulates the WRI and ESI in one framework using Lagrangian-based adjoint state method. We show both WRI and ESI may be expressed as a weighted norm of least-squares FWI, hence potentially suffering from local minima such as cycle-skipping issue. Due to penalty formulation, the regularized Hessian in WRI and ESI methods prevents a direct use of time domain modelling. We have considered the conjugate gradient method to efficiently solve the normal equation, allowing both time and frequency domain wave equation used in a matrix free fashion. After the time domain implementation, the numerical tests show that ESI has larger convexity domain compared to WRI and classic FWI using least-squares norm. A pure time domain
implementation is highly feasible from algorithmic point of view, but suffers heavily from storage issue. We made two additional attempts to alleviate memory problem, which have not been fully completed yet. To make ESI a viable candidate for practical application, smart strategies must be designed to mitigate storage request for time domain ESI, while efficient preconditioner should be developed to reduce the number of inner CG iterations in the nested loop.

Acknowledgement

Pengliang Yang thanks the inspiring communications with Jean Virieux and William W. Symes, in particular on the understanding of WRI approach.

A Another two methods for time domain ESI

A.1 Gauss-Seidel iterations

The relation in (25) derived using adjoint state method gives us several coupled equations equivalent to the normal equation (29), but shows possibility to compute wavefield by time stepping wave equation solvers if the right hand side is known. It inspires us to consider a recursive strategy to solve the normal equation, from a starting guess of the wavefield solution $u^0$. We can then recursively compute a new iterate of the field $\lambda^k$ by (25b), hence a new $q^k$ by (25c), and then a new iterate of the field $u^{k+1}$ by (25a). That is,

\[
\lambda^k = A^{-H}RH(d - Ru^k)
\]

\[
q^k = \frac{1}{\beta} (B^H B)^{-1} \lambda^k
\]

\[
u^{k+1} = A^{-1}q^k,
\]

which can be symbolically described in the following:

\[\ldots u^k \xrightarrow{A^{-H}} \lambda^k \xrightarrow{(B^H B)^{-1}} q^k \xrightarrow{A^{-1}} u^{k+1} \xrightarrow{A^{-H}} \ldots\]

By eliminating $q^k$, we have

\[
\begin{split}
\begin{pmatrix}
\lambda^k = A^{-H}RH(d - Ru^k) \\
u^{k+1} = \frac{1}{\beta} A^{-1}(B^H B)^{-1} \lambda^k
\end{pmatrix}
\iff
\begin{bmatrix}
I & 0 \\
\frac{1}{\beta} A^{-1}(B^H B)^{-1} & I
\end{bmatrix}
\begin{bmatrix}
\lambda^k \\
u^{k+1}
\end{bmatrix}
= \begin{bmatrix}
0 & -A^{-H}RH
0 & 0
\end{bmatrix}
\begin{bmatrix}
\lambda^{k-1}
\\u^k
\end{bmatrix}
+ \begin{bmatrix}
-A^{-H}RHd
0
\end{bmatrix}
\end{split}
\]

The diagonal and lower triangle of the matrix $M$, as well as the upper triangle of the matrix $N$ are none zero. The iterative scheme of this structure is exactly the well-known Gauss-Seidel iteration [Golub 1996 section 10.1.2]. The Gauss-Seidel scheme may be costly but makes the use of both time and frequency domain solvers feasible for FWI using extended source.

In terms of (53a), we plug in $\lambda^k = A^{-H}RH(d - Ru^k)$ into (53c), yielding

\[
Au^{k+1} = \frac{1}{\beta} (B^H B)^{-1} A^{-H}RH(d - Ru^k) = -\frac{1}{\beta} (B^H B)^{-1} A^{-H}RH Ru^k + \frac{1}{\beta} (B^H B)^{-1} A^{-H}RHd
\]

Multiplying $A^{-1}$ on both side of (55), the Gauss-Seidel recursion can be written as the fixed point iteration for single variable $u$

\[
u^{k+1} = -\frac{1}{\beta} A^{-1}(B^H B)^{-1} A^{-H}RH Ru^k + \frac{1}{\beta} A^{-1}(B^H B)^{-1} A^{-H}RHd.
\]

Assume $u^*$ is the true solution after convergence, hence

\[
u^* = -\frac{1}{\beta} A^{-1}(B^H B)^{-1} A^{-H}RH Ru^* + \frac{1}{\beta} A^{-1}(B^H B)^{-1} A^{-H}RHd.
\]

Define $e^k = u^k - u^*$ the error between current iterate with the true solution. It is then easy to establish the relation of the errors between two iterates

\[
\|e^{k+1}\| = \|u^{k+1} - u^*\| = \|\frac{1}{\beta} A^{-1}(B^H B)^{-1} A^{-H}RH R(u^k - u^*)\| \leq \|\frac{1}{\beta} A^{-1}(B^H B)^{-1} A^{-H}RH R\| \cdot \|e^k\|
\]
We conclude that regardless any initial value $u^0$, the Gauss-Seidel iterations are guaranteed to converge provided that the spectral radius satisfies
\[ \left\| \frac{1}{\beta}A^{-1}(B^H B)^{-1}A^{-H}R^HR \right\| < 1, \quad (59) \]
This establishes a quantitative requirement to choose a proper penalty coefficient $\beta > \left\| A^{-1}(B^H B)^{-1}A^{-H}R^HR \right\|$. Assuming $A$ is invertible, for any positive $\beta$ the regularized Hessian $\beta(BA)^H(BA) + R^HR$ in (29) is always positive definite, therefore invertible for the solution of the linear system of $u$. Unfortunately, the Gauss-Seidel scheme requires the penalty coefficient $\beta$ must be larger than certain values to converge. It rules out the Gauss-Seidel iterations to seek for the solution of the misfit functional if $\beta$ is chosen very small, even though the solution of the resulting normal equation does exist.

A.2 Recursion assisted by surrogate function

To bypass the above limitation, we may introduce a surrogate function of the objective based on the wavefield at iterate $k$, inspired by the work of Aghamiry et al. (2020)

\[ J'(u; u^k) = J(u) + \frac{1}{2}(u - u^k)^H(\gamma(BA)^HBA - R^HR)(u - u^k), \quad (60) \]
where the parameter $\gamma$ has to be chosen such that $\gamma(BA)^HBA - R^HR$ is positive definite, thus $J'(u; u^k) \geq J(u)$ holds for any $u$. Denote $u^{k+1} = \arg \min_u J'(u; u^k)$. It can be seen that

\[ J(u^{k+1}) \leq J'(u^{k+1}; u^k) \leq J'(u^k; u^k) = J(u^k). \quad (61) \]

Consequently, the minimizers of the surrogate function at each iteration constructs a sequence of wavefield solution that approaches the minimizer of the original misfit function $J(u)$.

The surrogate function can be rewritten as
\[ J'(u; u^k) = \frac{1}{2}\|d - Ru^k - R(u - u^k)\|^2 + \frac{\beta}{2}\|BA(u - u^k + u^k)\|^2 + \frac{\gamma}{2}\|BA(u - u^k)\|^2 - \frac{1}{2}\|R(u - u^k)\|^2 \]
\[ = \frac{1}{2}\|d - Ru^k\|^2 - \langle d - Ru^k, R(u - u^k) \rangle + \frac{\beta + \gamma}{2}\|BA(u - u^k)\|^2 + \frac{\beta}{2}\langle BA(u - u^k), BAu^k \rangle + \frac{\beta}{2}\|BAu^k\|^2 \quad (62) \]

Minimization of the surrogate function $J'(u; u^k)$ requires $\partial_u J' = 0$, yielding the normal equation
\[ -R^HR(d - Ru^k) + (\beta + \gamma)A^HBA(u - u^k) + \beta A^HBAu^k = 0 \Leftrightarrow (\beta + \gamma)u = \gamma u^k + A^{-1}(B^H B)^{-1}A^{-H}R^HR(d - Ru^k) \quad (63) \]

In terms of (25), we can define the following intermediate fields to complete each step of the recursion
\[ \lambda^k := A^{-H}R^H(d - Ru^k) \quad (64a) \]
\[ q^k := \frac{1}{\beta}(B^H B)^{-1}\lambda^k \quad (64b) \]
\[ u^k := A^{-1}q^k \quad (64c) \]
such that the minimizer of $J'(u; u^k)$ is given by
\[ u^{k+1} = \frac{1}{\beta + \gamma}(\beta^k + \gamma u^k) \quad (65) \]
The field $q$ can be stored using the same memory unit as $\lambda$ to avoid extra memory allocation. Each step of the above recursion requires at least two modellings (one adjoint modelling and one forward modelling). It is then symbolically described in the following:
\[ \cdots \to u^k \to A^{-H} \to \lambda^k \to (B^H B)^{-1} \to q^k \to A^{-1} \to u^k \to (\beta + \gamma)u^k \to u^{k+1} \to \cdots \]

The surrogate function is a quadratic form associated with operator $BA$ (its Hessian is $(BA)^HBA$). The convergence of the recursion will finally be dictated by the choice of $\gamma$ and the condition number
\[ \text{cond}(BA) = \frac{\sigma_{\text{max}}}{\sigma_{\text{min}}}, \]
where $\sigma_{\text{max}}$ and $\sigma_{\text{min}}$ stand for maximum and minimum of the eigenvalues of $BA$. The choice of the annihilator $B$ will drastically affect the convergence rate. If $\gamma$ is too large, then $J'$ will significantly differ from $J$, the resulting convergence to the solution of $J$ would be arbitrarily slow. If $\gamma$ is too small, we have no guarantee that the surrogate function satisfies $J'(u; u^k) \geq J(u)$, hence the solution of surrogate function may not converge to the solution of $J(u)$. 

21
References

Abubakar, A., Hu, W., Habashy, T. M., and van den Berg, P. M. (2009). Application of the finite-difference contrast-source inversion algorithm to seismic full-waveform data. *Geophysics*, 74(6):WCC47–WCC58.

Aghamiry, H. S., Gholami, A., and Operto, S. (2019). Improving full-waveform inversion by wavefield reconstruction with the alternating direction method of multipliers. *Geophysics*, 84(1):R125–R148.

Aghamiry, H. S., Gholami, A., and Operto, S. (2020). Accurate and efficient data-assimilated wavefield reconstruction in the time domain. *Geophysics*, 85(2):A7–A12.

Aghamiry, H. S., Gholami, A., and Operto, S. (2021). On efficient frequency-domain full-waveform inversion with extended search space. *Geophysics*, 86(2):R237–R252.

Barnier, G. and Biondi, E. (2020). Full waveform inversion by model extension using a model-space multi-scale approach. In *SEG International Exposition and Annual Meeting*. OnePetro.

Biondi, B. and Almomim, A. (2013). Tomographic full waveform inversion (TFWI) by combining FWI and wave-equation migration velocity analysis. *The Leading Edge*, September, special section: full waveform inversion:1074–1080.

Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J. (2010). Distributed optimization and statistical learning via the alternating direction of multipliers. *Foundations and trends in machine learning*, 3(1):1–122.

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

Bunks, C., Salek, F. M., Zaleski, S., and Chavent, G. (1995). Multiscale seismic waveform inversion. *Geophysics*, 60(5):1457–1473.

Engquist, B., Poulson, J., and Ying, L. (2011). Sweeping preconditioner for the 3D Helmholtz equation. In *SEG Technical Program Expanded Abstracts*, pages 3158–3163.

Esser, E., Guasch, L., Leeuwen, T. v., Aravkin, A. Y., and Herrmann, F. J. (2015). Automatic salt delineation—wavefield reconstruction inversion with convex constraints. In *SEG Technical Program Expanded Abstracts 2015*, pages 1337–1343. Society of Exploration Geophysicists.

Fu, L. and Symes, W. W. (2017). A discrepancy-based penalty method for extended waveform inversion. *Geophysics*, 82(5):R287–R298.

Gholami, A., Aghamiry, H. S., and Operto, S. (2020). Extended full waveform inversion in the time domain by the augmented lagrangian method. *arXiv preprint arXiv:2011.14102*

Golub, G. H. (1996). *Matrix Computation, third edition*. Johns Hopkins Studies in Mathematical Sciences.

Hao, Y. and Simoncini, V. (2020). The sherman–morrison–woodbury formula for generalized linear matrix equations and applications. *Numerical Linear Algebra with Applications*.

Huang, G., Nammour, R., and Symes, W. (2017). Full-waveform inversion with source-receiver extension. *Geophysics*, 82(3):R153–R171.

Huang, G., Nammour, R., and Symes, W. W. (2018). Source-independent extended waveform inversion based on space-time source extension: Frequency-domain implementation. *Geophysics*, 83(5):R449–R461.

Huang, G., Symes, W., and Nammour, R. (2016). Matched source waveform inversion: Space-time extension. In *SEG Technical Program Expanded Abstracts*, pages 1426–1431.

Lailly, P. (1983). The seismic inverse problem as a sequence of before stack migrations. In Bednar, R. and Weglein, editors, *Conference on Inverse Scattering, Theory and application*, Society for Industrial and Applied Mathematics, Philadelphia, pages 206–220.
Luo, Y. and Schuster, G. T. (1991). Wave-equation traveltime inversion. *Geophysics*, 56(5):645–653.

Ma, Y. and Hale, D. (2013). Wave-equation reflection traveltime inversion with dynamic warping and full waveform inversion. *Geophysics*, 78(6):R223–R233.

Métivier, L., Brossier, R., Mérigot, Q., Oudet, E., and Virieux, J. (2016). Measuring the misfit between seismograms using an optimal transport distance: Application to full waveform inversion. *Geophysical Journal International*, 205:345–377.

Nocedal, J. and Wright, S. J. (2006). *Numerical Optimization*. Springer, 2nd edition.

Peters, B., Herrmann, F. J., and van Leeuwen, T. (2014). Wave-equation based inversion with the penalty method-adjoint-state versus wavefield-reconstruction inversion. In *76th EAGE Conference and Exhibition 2014*, volume 2014, pages 1–5. European Association of Geoscientists & Engineers.

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.

Rizzuti, G., Louboutin, M., Wang, R., and Herrmann, F. J. (2020). A dual formulation of wavefield reconstruction inversion for large-scale seismic inversion. *arXiv preprint arXiv:2010.09033*.

Shin, C. and Cha, Y. H. (2008). Waveform inversion in the Laplace domain. *Geophysical Journal International*, 173(3):922–931.

Shin, C. and Ha, W. (2008). A comparison between the behavior of objective functions for waveform inversion in the frequency and laplace domains. *Geophysics*, 73(5):VE119–VE133.

Sirgue, L., Etgen, J. T., and Albertin, U. (2008). 3D Frequency Domain Waveform Inversion using Time Domain Finite Difference Methods. In *Proceedings 70th EAGE, Conference and Exhibition, Roma, Italy*, page F022.

Symes, W. (2015). Algorithmic aspects of extended waveform inversion. In *77th EAGE Conference and Exhibition-Workshops*, volume 2015, pages 1–5. European Association of Geoscientists & Engineers.

Symes, W. W. (1998). All stationary points of differential semblance are asymptotic global minimizers: layered acoustics. Technical report, The Rice Inversion Project.

Symes, W. W. (2007). Reverse time migration with optimal checkpointing. *Geophysics*, 72(5):SM213–SM221.

Symes, W. W. (2008). Migration velocity analysis and waveform inversion. *Geophysical Prospecting*, 56:765–790.

Symes, W. W. (2020). Wavefield reconstruction inversion: an example. *Inverse Problems*, 36(10):105010.

Symes, W. W., Chen, H., and Minkoff, S. E. (2020). Full-waveform inversion by source extension: Why it works. In *SEG Technical Program Expanded Abstracts 2020*, pages 765–769. Society of Exploration Geophysicists.

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

Tarantola, A. (1984b). Linearized inversion of seismic reflection data. *Geophysical Prospecting*, 32:998–1015.

Van den Berg, P. M. and Abubakar, A. (2001). Contrast source inversion method: State of art. *Progress in Electromagnetics Research*, 34:189–218.

van Leeuwen, T. (2019). A note on extended full waveform inversion. *arXiv preprint arXiv:1904.00363*.

van Leeuwen, T. and Herrmann, F. J. (2013). Mitigating local minima in full-waveform inversion by expanding the search space. *Geophysical Journal International*, 195(1):661–667.

Virieux, J. and Operto, S. (2010). *An overview of full waveform inversion in exploration geophysics*, page ISBN: 9781560802266 (13). Society of Exploration Geophysicists.

Wang, C., Yingst, D., Farmer, P., and Leveille, J. (2016). Full-waveform inversion with the reconstructed wavefield method. In *SEG Technical Program Expanded Abstracts 2016*, pages 1237–1241. Society of Exploration Geophysicists.
Warner, M. and Guasch, L. (2014). Adaptative waveform inversion - FWI without cycle skipping - theory. In *76th EAGE Conference and Exhibition 2014*, page We E106 13.

Wu, R.-S., Luo, J., and Wu, B. (2014). Seismic envelope inversion and modulation signal model. *Geophysics, 79*(3):WA13–WA24.

Yang, P., Brossier, R., Métivier, L., and Virieux, J. (2016a). Wavefield reconstruction in attenuating media: A checkpointing-assisted reverse-forward simulation method. *Geophysics, 81*(6):R349–R362.

Yang, P., Brossier, R., and Virieux, J. (2016b). Wavefield reconstruction from significantly decimated boundaries. *Geophysics, 80*(5):T197–T209.

Zhu, H. and Fomel, S. (2016). Building good starting models for full-waveform inversion using adaptive matching filtering misfit. *Geophysics, 81*(5):U61–U72.