Parsimonious truncated Newton method for time-domain full waveform inversion based on Fourier-domain full-scattered-field approximation

Peng Yong, Romain Brossier, Ludovic Métivier

Parsimonious truncated Newton method for time-domain full waveform inversion based on Fourier-domain full-scattered-field approximation

Peng Yong [1], Romain Brossier[1] and Ludovic Métivier[2,1]

[1] **Univ. Grenoble Alpes, ISTerre, F-38000 Grenoble, France**

[2] **Univ. Grenoble Alpes, CNRS, LJK, F-38000 Grenoble, France**

Email address: peng.yong@univ-grenoble-alpes.fr

romain.brossier@univ-grenoble-alpes.fr

ludovic.metivier@univ-grenoble-alpes.fr

(September 16, 2021)

Running head: **Parsimonious truncated Newton method**

ABSTRACT

In order to exploit Hessian information in Full Waveform Inversion (FWI), the matrix-free truncated Newton method can be used. In such a method, Hessian-vector product computation is one of the major concerns due to the huge memory requirements and demanding computational cost. Using the adjoint-state method, the Hessian-vector product can be estimated by zero-lag cross-correlation of the first-order/second-order incident wavefields and the second-order/first-order adjoint wavefields. Different from the implementation in frequency-domain FWI, Hessian-vector product construction in the time domain becomes much more challenging as it is not affordable to store the entire time-dependent wavefields. The widely used wavefield recomputation strategy leads to computationally intensive tasks. We present an efficient alternative approach to computing the Hessian-vector product

for time-domain FWI. In our method, discrete Fourier transform is applied to extract frequency-domain components of involved wavefields, which are used to compute wavefield cross-correlation in the frequency domain. This makes it possible to avoid reconstructing the first-order and second-order incident wavefields. In addition, a full-scattered-field approximation is proposed to efficiently simplify the second-order incident and adjoint wavefields computation, which enables us to refrain from repeatedly solving the first-order incident and adjoint equations for the second-order incident and adjoint wavefields (re)computation. With the proposed method, the computational time can be reduced by 70% and 80% in viscous media for Gauss-Newton and full-Newton Hessian-vector product construction, respectively. The effectiveness of our method is also verified in the frame of a 2D multi-parameter inversion, in which the proposed method almost reaches the same iterative convergence of the conventional time-domain implementation.

INTRODUCTION

Full waveform inversion (FWI) (Lailly, 1983; Tarantola, 1984) has been widely used in exploration and global seismology for high-resolution parameter estimation (Virieux and Operto, 2009; Tromp, 2020). With the success of the application of mono-parameter FWI to field data (Sirgue et al., 2010; Warner et al., 2013; Operto et al., 2015; Shen et al., 2018), it becomes more and more attractive to study multiple parameters (Operto et al., 2013), to account for the effect of attenuation (Kamei and Pratt, 2013; Fabien-Ouellet et al., 2017; da Silva et al., 2019; Kamath et al., 2021), density (Yang et al., 2016a; Operto and Miniussi, 2018), anisotropy (Prieux et al., 2011; Alkhalifah and Plessix, 2014), or elastic parameters (Brossier et al., 2009; Köhn et al., 2012; Vigh et al., 2014; Pan et al., 2018; Trinh et al., 2019; Wang et al., 2021).

FWI formulates seismic inversion into a PDE-constrained optimization problem. The optimal model parameters are usually obtained by gradient-based optimization methods through minimizing the objective function (van Leeuwen and Herrmann, 2015; Virieux et al., 2017). The Hessian matrix describes the local curvature of the objective function, which can be used to correct for geometrical spreading and second-order scattering effects (Pratt et al., 1998; Virieux and Operto, 2009; Métivier et al., 2013; Liu et al., 2020). For multi-parameter FWI, the Hessian matrix plays an important role to accelerate convergence rate and mitigate cross-talk between different parameters (Pratt et al., 1998; Métivier et al., 2015; Pan et al., 2016). As the size of Hessian is the square of the size of the gradient vector, it is impractical to store and explicitly use it. The truncated Newton strategy provides a matrix-free fashion to take into account the Hessian information during the descent direction computation, through a linear conjugate-gradient-based system, and has been shown to be a powerful tool in FWI (Epanomeritakis et al., 2008; Métivier et al., 2013; Yang et al., 2018; Matharu and Sacchi, 2019; Liu et al., 2020). Theoretically, there are many benefits of second-order optimization method

(Métivier et al., 2013; Pan et al., 2016), while truncated Newton method has not been widely adopted in FWI. An important part of the problem is the expensive computational cost of Hessian-vector product construction.

In fact, most applications of truncated Newton method to FWI are mainly in the frequency domain (Métivier et al., 2013, 2014; Liu et al., 2020), because storing few frequency-domain wavefields in memory is relatively cheap even for large-scale problems, and the cost of Hessian-vector product construction is not so expensive, in particular when direct solvers are involved for forward problem (Métivier et al., 2013). However, direct solvers for 3D realistic frequency-domain FWI applications require drastic memory for the matrix decomposition (Operto et al., 2007; Li et al., 2020). Most FWI applications are thus performed in the time domain, with the additional advantage that selecting specific arrivals (diving/transmitted waves, reflected phases) is possible and easily implemented through time-windowing. However, Hessian-vector product construction in the time-domain FWI becomes much more expensive since storing an entire time-dependent wavefield in memory is challenging, and the widely used recomputation strategies lead to high computational cost (Yang et al., 2018). Therefore, there is an interest for developing strategies reducing Hessian-vector computation cost for time-domain implementation of truncated Newton methods.

Source encoding (Castellanos et al., 2015) and subsampling shot strategy (Matharu and Sacchi, 2019) have been applied to reduce the computational cost in a "coarse-grained" way. The main idea is to reduce the number of seismic shots involved in the Hessian-vector product computation. The Hessian operator is indeed the summation over shots of Hessian-like operators associated with each shot taken separately.

We propose a parsimonious approach for time-domain FWI in a "fine-grained" way, using all the shots for gradient, but relying on two approximations to significantly decrease the computation effort

of Hessian-vector product. The developed method can almost reach the same iterative convergence rate of time-domain method while using greatly reduced computational time, which has not been achieved by these "coarse-grained" methods to the best of our knowledge.

Considering that frequency-domain wavefields are relatively cheap to store, and relying on the same kind of "on-the-fly" Fourier transform than the ones proposed for gradient building with phase-sensitive detection (Nihei and Li, 2007) or discrete Fourier transform (Sirgue et al., 2008), we can extract the Fourier-domain components of the wavefields during wavefield extrapolation and save them for several frequencies due to the band-limited nature of seismic data. Therefore, we approximate Hessian-vector products with a few frequencies thanks to Fourier-domain compression. This approximation enables us to avoid the computationally intensive task of reconstructing the first and second-order incident wavefields (Nguyen and McMechan, 2015), in particular when viscous media are involved (Yang et al., 2016d).

In addition to this Fourier-domain approximation, a second improvement relies on the fact that the second-order incident and adjoint wavefields can be interpreted as first-order Born scattering wavefields generated from interactions between model perturbations and first-order incident and adjoint wavefields. It is challenging to implement such equivalent first-order Born modeling in 3D, as source terms of the second-order incident and adjoint wavefields are volumetric, time-dependent and related to the first-order wavefields. As it is not feasible to store entire first-order incident and adjoint wavefields, it is mandatory to solve first-order incident and adjoint equations for second-order incident and adjoint wavefields (re)computation. Note that the physical meaning of the second-order incident and adjoint wavefields is related to linear wavefield changes with a medium perturbation (Schuster, 2017), which is known as first-order Born scattering wavefield in least-squares migration (Dai and Schuster, 2013; Yong et al., 2019). In fact, the original gradient of FWI is derived from the first-order Born assumption (Tarantola, 1984), which uses the first-order Born wavefield to approximate the

difference between the observed and predicted data. Here, we consider reversely applying the first-order Born assumption, approximating second-order incident and adjoint wavefields with the total wavefield change generated by a small perturbation. Using this "full scattered field" approximation, one can avoid repeatedly solving first-order incident and adjoint equations for second-order incident and adjoint wavefields computation.

We first give a brief introduction of truncated Newton algorithm for FWI with adjoint-state method. Then, we introduce how to integrate Fourier-domain compression and full-scattered-field approximation into Hessian-vector product construction. The error analysis and computational complexity are also discussed in this theory part. Following, we give numerical tests on Born wavefield comparison and Hessian-vector product construction to illustrate the accuracy and efficiency of the proposed approximation. Next, we apply the developed parsimonious truncated Newton method to multi-parameter inversion on a 2D synthetic Valhall model to test its performance. Discussion and conclusion are presented in the last two sections.

TRUNCATED NEWTON METHOD IN FWI

We first give the general theory about the application of truncated Newton method to least-squares inverse problem. Then, we briefly introduce how to calculate gradient and Hessian-vector product using adjoint-state method. Finally, we discuss the time-domain implementation with wavefield reconstruction for FWI.

General theory

Time-domain FWI updates the model by minimizing the difference between the observed and predicted data, which can be formulated as a PDE-constrained optimization problem of the form

$$\min_{\mathbf{m}} \chi(\mathbf{m}) = \frac{1}{2} \int_0^T \mathrm{d}t (R\mathbf{w}(\mathbf{m}) - \mathbf{d})^\dagger (R\mathbf{w}(\mathbf{m}) - \mathbf{d}), \tag{1}$$

where $\mathbf{m}$ is the model parameters of interest in model space $\mathcal{M} \subset \mathbf{R}^n$, $\mathbf{d} := \mathbf{d}(\mathbf{x}_r, t)$ is the observed data at receiver location $\mathbf{x}_r$, R is a receiver sampling operator and $R\mathbf{w}$ represents the predicted data, which is extracted from the wavefield $\mathbf{w}(\mathbf{x}, t)$ at the receiver location. $\mathbf{x} \in \Omega$ and $t \in [0, T]$ denote the spatial and time domain in the physical world. For parameter estimation in the geophysical inverse problem, the model space $\mathcal{M}$ usually relies on the spatial space Ω. $\dagger$ denotes complex conjugate transport.

The wavefield $\mathbf{w}$ is governed by

$$A(\mathbf{m})\mathbf{w} = \mathbf{s}, \quad \mathbf{w}(\mathbf{x}, t)|_{t=0} = 0, \tag{2}$$

where $A(\mathbf{m})$ represents the forward modeling operator and $\mathbf{s}$ is the source term. At this point, we keep the derivation as general as possible and therefore do not specify any particular time-domain

wave equation. Later, we will present the wave equation used in the numerical studies.

In the framework of truncated Newton method, the optimization solution can be found by iteratively updating the model parameter following the scheme

$$\mathbf{m}^{k+1} = \mathbf{m}^k + \alpha^k \Delta \mathbf{m}^k, \tag{3}$$

where α^k represents step length at the k-th iteration, which can be determined by the linesearch method with the classical Wolfe conditions (Nocedal and Wright, 2006; Métivier and Brossier, 2016). $\Delta \mathbf{m}^k$ is the model update direction, and is obtained by solving the following equation

$$H(\mathbf{m}^k)\Delta \mathbf{m}^k = -\nabla \chi(\mathbf{m}^k), \tag{4}$$

where $\nabla \chi(\mathbf{m}) \in \mathbf{R}^n$ and $H(\mathbf{m}) = \nabla^2 \chi(\mathbf{m}) \in \mathbf{R}^{n \times n}$ are the gradient vector and Hessian matrix, which respectively represent the first and second-order derivatives of the misfit function with respect to model parameters $\mathbf{m}$. For realistic 3D FWI applications, the size of model parameters n can reach $O(10^9) - O(10^{12})$, which prohibits explicit storing the Hessian matrix. In general, the matrix-free conjugate gradient method is applied to solve the system of linear equations (4) (Nash, 2000; Knoll and Keyes, 2004). A basic description of the truncated Newton method with preconditioned CG is given in the algorithm 1, in which efficient construction of gradient and Hessian-vector product via adjoint-state method are two key ingredients in applying truncated Newton method to FWI (Métivier et al., 2013; Virieux et al., 2017). Figure 1 presents the workflow of the truncated Newton method, also includes the main contribution in this paper that efficiently and accurately compute Hessian-vector product with Fourier-domain full-scattered-field approximation, which will be delivered in the following sections.

[Figure 1 about here.]

Note that truncated Newton method is based on the second-order Taylor expansion of the misfit function. Since this approximation may be inaccurate for highly nonlinear FWI problem, there is no need to solve the Newton equations accurately at each iteration (Nash, 2000). To avoid over-solving the Newton equations, the parameter η in the algorithm 1, determined by Eisenstat and Walker forcing-term formula (Eisenstat and Walker, 1996; Métivier et al., 2017), is used in this paper to automatically control the number of CG iterations depending on the accuracy of the local second-order Taylor expansion of the misfit function. The forcing-term formula used in this study is given by

$$\eta^k = \frac{||\nabla\chi(\mathbf{m}^k) - \nabla\chi(\mathbf{m}^{k-1}) - \alpha^{k-1}H(\mathbf{m}^{k-1})\Delta\mathbf{m}^{k-1}||_2}{||\nabla\chi(\mathbf{m}^{k-1})||_2}. \tag{5}$$

Here, α^{k-1} is the step length for the latest model update and $||\cdot||_2$ is L_2 norm. In practice, the following two additional safeguards are usually applied to ensure an effective and stable value of η:

- If $(\eta^{k-1})^{(1+\sqrt{5})/2} > 0.1$, then $\eta^k = \max\{\eta^k, (\eta^{k-1})^{(1+\sqrt{5})/2}\}$
- If $\eta^k > 1$, then $\eta^k = 0.9$

For more details about the forcing-term formula in FWI applications, it can be found at the 4.3 section in the paper (Métivier et al., 2017).

Gradient computation via first-order adjoint-state method

We will introduce how to compute the gradient with first-order adjoint-state method. For a constrained optimization problem (1) with the PDE constraint (2), one can apply Lagrange multiplier method to

Algorithm 1 Truncated Newton method with preconditioned conjugate gradient

Input: initial model $\mathbf{m}^0$, observed data $\mathbf{d}$, tolerance error ϵ, forcing term η
Output: $\min_{\mathbf{m}} \chi(\mathbf{m})$

while $\chi(\mathbf{m}) > \epsilon$ **do**
  compute gradient $\nabla\chi(\mathbf{m})$ via first-order adjoint-state method
  set $\Delta\mathbf{m} = 0$, $\mathbf{g} = \nabla\chi(\mathbf{m})$, $\mathbf{p} = P\mathbf{g}$, $\mathbf{r} = -\mathbf{p}$
  while $\|H(\mathbf{m})\Delta\mathbf{m} + \nabla\chi(\mathbf{m})\| > \eta\|\nabla\chi(\mathbf{m})\|$ **do**
    compute Hessian-vector product $H(\mathbf{m})\mathbf{r}$ via second-order adjoint-state method
    $\beta_1 = \langle H(\mathbf{m})\mathbf{r}, \mathbf{r}\rangle$
    if $\beta_1 < 0$ **then**
      stop the inner iterations
    else
      $\beta_2 = \langle \mathbf{p}, \mathbf{g}\rangle$
      $\Delta\mathbf{m} = \Delta\mathbf{m} + (\beta_2/\beta_1)\,\mathbf{r}$
      $\mathbf{g} = \mathbf{g} + (\beta_2/\beta_1)\,H(\mathbf{m})\mathbf{r}$
      $\mathbf{p} = P\mathbf{g}$
      $\mathbf{r} = -\mathbf{p} + (\langle \mathbf{p}, \mathbf{g}\rangle/\beta_2)\,\mathbf{r}$
    end if
  end while
  compute step length α (globalization method)
  $\mathbf{m} = \mathbf{m} + \alpha\Delta\mathbf{m}$
  update η with the chosen Eisenstat and Walker forcing-term formula
end while

convert it into an unconstrained problem.

$$\min_{\mathbf{m},\mathbf{w},\boldsymbol{\lambda}} \mathcal{L}_1(\mathbf{m}, \mathbf{w}, \boldsymbol{\lambda}) = \chi(\mathbf{m}) + \langle \boldsymbol{\lambda}, A(\mathbf{m})\mathbf{w} - \mathbf{s}\rangle_{\mathcal{W}} \tag{6}$$

where $\langle\cdot,\cdot\rangle_{\mathcal{W}} =: \Omega\times[0,T] \to \mathbf{R}$, denotes the integral of inner product over space and time $\int_T \int_\Omega \langle\cdot,\cdot\rangle \mathrm{d}t\mathrm{d}\mathbf{x}$ in wavefield space, and $\boldsymbol{\lambda}$ is the Lagrange multiplier or adjoint-state variable. Here, to make the derivation readable, we do not include the initial condition of the forward modeling equation (2) in the Lagrangian (6). In fact, the final condition of the adjoint-state equation is derived from this initial condition. We will directly give the final condition of the adjoint-state equation. For a more rigorous mathematical derivation, please refer to the review of the adjoint-state method (Plessix, 2006).

Note that simultaneously updating (and hence storing) all the variables $(\mathbf{m}, \mathbf{w}, \boldsymbol{\lambda})$ in (6) is

usually unfeasible for the large-scale geophysical applications. In general, one considers a reduced formulation to only update $\mathbf{m}$ during each iteration by zeroing the derivative of $\mathcal{L}_1$ with respect to the adjoint-state variable $\boldsymbol{\lambda}$ and state variable $\mathbf{w}$.

$$\frac{\partial \mathcal{L}_1(\mathbf{m}, \mathbf{w}, \boldsymbol{\lambda})}{\partial \boldsymbol{\lambda}} = 0 \Rightarrow A(\mathbf{m})\mathbf{w} = \mathbf{s}. \tag{7}$$

$$\frac{\partial \mathcal{L}_1(\mathbf{m}, \mathbf{w}, \boldsymbol{\lambda})}{\partial \mathbf{w}} = 0 \Rightarrow A^{\dagger}(\mathbf{m})\boldsymbol{\lambda} = R^{\dagger}(\mathbf{d} - R\mathbf{w}). \tag{8}$$

Equations (7) and (8) are known as the state equation and adjoint-state equation. $\mathbf{w}$ and $\boldsymbol{\lambda}$ are often called as the first-order incident and adjoint wavefields, and both of them now depend on the model parameter $\mathbf{m}$. $A^{\dagger}(\mathbf{m})$ is the adjoint operator of the forward modeling operator $A(\mathbf{m})$. It is important to point out that, different from the initial condition in the equation (2), the first-order adjoint-state equation contains a final condition

$$\boldsymbol{\lambda}(\mathbf{x}, t)|_{t=T} = 0, \tag{9}$$

which indicates back-propagating the data residuals at the receiver location (Tarantola, 1984; Tromp et al., 2005; Virieux et al., 2017).

Thanks to the two conditions above ($\frac{\partial \mathcal{L}_1}{\partial \mathbf{w}(\mathbf{m})} = \frac{\partial \mathcal{L}_1}{\partial \boldsymbol{\lambda}(\mathbf{m})} = 0$), the gradient $\nabla\chi(\mathbf{m})$ can be written as

$$\nabla\chi(\mathbf{m}) = \langle \boldsymbol{\lambda}, \frac{\partial A(\mathbf{m})}{\partial \mathbf{m}}\mathbf{w} \rangle_T = \int_0^T \mathrm{d}t \boldsymbol{\lambda}^{\dagger} \frac{\partial A(\mathbf{m})}{\partial \mathbf{m}}\mathbf{w}, \tag{10}$$

where $\frac{\partial A(\mathbf{m})}{\partial \mathbf{m}}$ is the so-called first-order Born scattering operator, $\langle \cdot, \cdot \rangle_T$ denotes the integral of inner product over time $\int_T \langle \cdot, \cdot \rangle \mathrm{d}t$.

Hessian-vector product computation via second-order adjoint-state method

Following a similar procedure of the adjoint-state method for the gradient, we can compute the Hessian-vector product $H(\mathbf{m})\mathbf{r}$, without the explicit expression of Hessian matrix, via the so-called second-order adjoint-state method (Fichtner and Trampert, 2011; Métivier et al., 2013). Here, $\mathbf{r} \in \mathcal{M}$ is related to the CG update in the Algorithm 1, and does not depend on the model parameter $\mathbf{m}$.

To construct the Hessian-vector product $H(\mathbf{m})\mathbf{r}$, we can consider the following objective function (Métivier et al., 2013)

$$\phi_{\mathbf{r}}(\mathbf{m}) := \langle \nabla\chi(\mathbf{m}), \mathbf{r}\rangle_{\mathcal{M}}. \tag{11}$$

One can easily find out that the derivative of $\phi_{\mathbf{r}}(\mathbf{m})$ with respect to $\mathbf{m}$ is exactly the Hessian-vector product $H(\mathbf{m})\mathbf{r}$

$$\nabla\phi_{\mathbf{r}}(\mathbf{m}) = \frac{\partial\phi_{\mathbf{r}}(\mathbf{m})}{\partial\mathbf{m}} = H(\mathbf{m})\mathbf{r}. \tag{12}$$

Before applying second-order adjoint-state method to compute the Hessian-vector product, we should be aware of the three existing constraints in (11):

- First-order state equation: $A(\mathbf{m})\mathbf{w} = \mathbf{s}$,
- First-order adjoint-state equation: $A^{\dagger}(\mathbf{m})\boldsymbol{\lambda} = R^{\dagger}(\mathbf{d} - R\mathbf{w})$,
- Gradient: $\nabla\chi(\mathbf{m}) = \langle\boldsymbol{\lambda}, \frac{\partial A(\mathbf{m})}{\partial\mathbf{m}}\mathbf{w}\rangle_T$.

Using Lagrange multiplier method again yields

$$\begin{aligned}\mathcal{L}_2(\mathbf{m}, \mathbf{w}, \boldsymbol{\lambda}, \nabla_{\mathbf{m}}\chi, \nu, \mathbf{u}, \boldsymbol{\mu}) \quad &= \langle\nabla_{\mathbf{m}}\chi, \mathbf{r}\rangle_{\mathcal{M}} + \langle\nu, \nabla_{\mathbf{m}}\chi - \langle\boldsymbol{\lambda}, \tfrac{\partial A(\mathbf{m})}{\partial\mathbf{m}}\mathbf{w}\rangle_T\rangle_{\mathcal{M}} \\ &\quad + \langle\mathbf{u}, A^{\dagger}(\mathbf{m})\boldsymbol{\lambda} + R^{\dagger}(R\mathbf{w} - \mathbf{d})\rangle_{\mathcal{W}} + \langle\boldsymbol{\mu}, A(\mathbf{m})\mathbf{w} - \mathbf{s}\rangle_{\mathcal{W}},\end{aligned} \tag{13}$$

where $\mathbf{w}$, $\nabla_{\mathbf{m}}\chi$ and $\boldsymbol{\lambda}$ are the state variables, and the auxiliary variables $\mathbf{u}$, $\boldsymbol{\mu}$ and ν are adjoint-state

variables in the second-order adjoint-state method. Please keep in mind that it is not feasible to update all the variables in $\mathcal{L}_2(\mathbf{m}, \mathbf{w}, \boldsymbol{\lambda}, \nabla_{\mathbf{m}}\chi, \nu, \mathbf{u}, \boldsymbol{\mu})$ at the same time for large-scale geophysical applications. We therefore apply again the procedure of zeroing the derivative of $\mathcal{L}_2$ with respect to all the state and adjoint-state variables, which gives the second-order adjoint-state and state equations, respectively.

- Differentiating $\mathcal{L}_2$ with respect to adjoint-state variables ν, $\boldsymbol{\mu}$ and $\mathbf{u}$ leads to the second-order state equations:

$$\frac{\partial \mathcal{L}_2}{\partial \nu} = 0 \Rightarrow \nabla_{\mathbf{m}}\chi = \langle \boldsymbol{\lambda}, \frac{\partial A(\mathbf{m})}{\partial \mathbf{m}}\mathbf{w} \rangle_T, \tag{14a}$$

$$\frac{\partial \mathcal{L}_2}{\partial \boldsymbol{\mu}} = 0 \Rightarrow A(\mathbf{m})\mathbf{w} = \mathbf{s}, \tag{14b}$$

$$\frac{\partial \mathcal{L}_2}{\partial \mathbf{u}} = 0 \Rightarrow A^{\dagger}(\mathbf{m})\boldsymbol{\lambda} = R^{\dagger}(\mathbf{d} - R\mathbf{w}). \tag{14c}$$

- Differentiating $\mathcal{L}_2$ with respect to state variables $\nabla_{\mathbf{m}}\chi$, $\boldsymbol{\lambda}$ and $\mathbf{w}$, then rearranging terms gives the second-order adjoint-state equations:

$$\frac{\partial \mathcal{L}_2}{\partial (\nabla_{\mathbf{m}}\chi)} = 0 \Rightarrow \nu = -\mathbf{r}, \tag{15a}$$

$$\frac{\partial \mathcal{L}_2}{\partial \boldsymbol{\lambda}} = 0 \Rightarrow A(\mathbf{m})\mathbf{u} = \sum_{m_i \in \mathbf{m}} \nu_i \frac{\partial A(\mathbf{m})}{\partial m_i}\mathbf{w}, \tag{15b}$$

$$\frac{\partial \mathcal{L}_2}{\partial \mathbf{w}} = 0 \Rightarrow A^{\dagger}(\mathbf{m})\boldsymbol{\mu} = -R^{\dagger}R\mathbf{u} + \sum_{m_i \in \mathbf{m}} \nu_i \left(\frac{\partial A(\mathbf{m})}{\partial m_i} \right)^{\dagger} \boldsymbol{\lambda}. \tag{15c}$$

Using the relationship of $\nu = -\mathbf{r}$, the second-order adjoint-state equations can be simplified as

$$A(\mathbf{m})\mathbf{u} = -\sum_{m_i \in \mathbf{m}} r_i \frac{\partial A(\mathbf{m})}{\partial m_i}\mathbf{w}, \quad \mathbf{u}(\mathbf{x}, t)|_{t=0} = 0, \tag{16}$$

$$A^{\dagger}(\mathbf{m})\boldsymbol{\mu} = -R^{\dagger}R\mathbf{u} - \sum_{m_i \in \mathbf{m}} r_i \left(\frac{\partial A(\mathbf{m})}{\partial m_i}\right)^{\dagger} \boldsymbol{\lambda}, \quad \boldsymbol{\mu}(\mathbf{x}, t)|_{t=T} = 0. \tag{17}$$

Here, we also directly give the initial and final conditions. Following the name of the first-order incident and adjoint wavefields $\mathbf{w}$ and $\boldsymbol{\lambda}$, $\mathbf{u}$ and $\boldsymbol{\mu}$ are often called as the second-order incident and adjoint wavefields, which are forward and backward propagating, respectively. Both of them depend on not only the model parameter $\mathbf{m}$ but also the given vector $\mathbf{r}$. Thanks to the zero-valued derivatives in the equations (14) and (15), the Hessian-vector product can be obtained by taking differentiation of $\mathcal{L}_2$ with respect to model parameters $\mathbf{m}$.

$$H(\mathbf{m})\mathbf{r} = \langle \mathbf{u}, \frac{\partial A^{\dagger}(\mathbf{m})}{\partial \mathbf{m}} \boldsymbol{\lambda} \rangle_T + \langle \boldsymbol{\mu}, \frac{\partial A(\mathbf{m})}{\partial \mathbf{m}} \mathbf{w} \rangle_T - \langle \boldsymbol{\lambda}, \left(\frac{\partial^2 A(\mathbf{m})}{\partial \mathbf{m}^2} \mathbf{w}\right) \nu \rangle_T, \tag{18}$$

In mathematics, the conjugate gradient method in the Newton-CG Algorithm 1 is designed for the numerical solution of symmetric positive definite systems of linear equations (Shewchuk et al., 1994; Saad, 2003). However, the symmetric full-Newton (FN) Hessian matrix might be indefinite, which is an indication of an inaccurate local quadratic approximation of the misfit function. The truncated Newton method adapts the accuracy with which it solves the inner Newton linear system through the forcing term of Eisenstat and Walker (1996), to take into account this information on the quality of the local quadratic approximation. Besides, as soon as a negative curvature is detected in the CG algorithm, the inner CG iterations are stopped (Métivier et al., 2013). Nocedal and Wright (2006) proves that this stopping criterion guarantees to always provide a descent direction with the truncated Newton procedure. In the following section, we will introduce Gauss-Newton (GN) Hessian, a symmetric and positive-definite approximation of the FN Hessian, which makes this additional stopping criterion becomes unnecessary.

Gauss-Newton Hessian-vector product computation

The GN approximation is interesting because of a reduced computational cost (Pratt et al., 1998; Epanomeritakis et al., 2008). It is a pertinent approximation as soon as the residuals become small as well as the magnitude of second-order derivatives of the calculated data with respect to the model parameters. In FWI application, the latter condition is satisfied when no strong contrasts, producing high amplitude multi-scattered events are met. When this kind of contrasts are present in investigated media, the GN approximation becomes inaccurate (Métivier et al., 2013; Liu et al., 2020). The corresponding Hessian-vector product is given by

$$H(\mathbf{m})\mathbf{r} = \langle \boldsymbol{\mu}_1, \frac{\partial A(\mathbf{m})}{\partial \mathbf{m}} \mathbf{w} \rangle_T, \tag{19}$$

where $\boldsymbol{\mu}_1$ is defined by

$$A^{\dagger}(\mathbf{m})\boldsymbol{\mu}_1 = -R^{\dagger}R\mathbf{u}, \quad \boldsymbol{\mu}_1(\mathbf{x}, t)|_{t=T} = 0. \tag{20}$$

The relationships between different wavefields involved in the computation of Hessian-vector production are shown in Figure 2. In addition, a brief analysis on relationship between Hessian-vector product and gradient, and derivation of GN Hessian-vector product can be found in APPENDIX A.

[Figure 2 about here.]

Preconditioner

Both Métivier et al. (2013) and Yang et al. (2018) have pointed out that FN and GN Hessian matrices are both ill-conditioned, thus preconditioner for the Newton equation is of critical importance

(Innanen, 2014; Métivier et al., 2015). A preconditioner based on source energy illumination is used in this study, which can make a compensation for unbalanced illumination for different parameters.

$$diag\overline{H} = \sqrt{\int_0^T \mathrm{d}t \left(\mathbf{w}(t)\frac{\partial A(\mathbf{m})}{\partial \mathbf{m}_i}\mathbf{w}(t)\right)^{\dagger} \left(\mathbf{w}(t)\frac{\partial A(\mathbf{m})}{\partial \mathbf{m}_i}\mathbf{w}(t)\right)}. \tag{21}$$

A tunable scaling strategy (Kamei and Pratt, 2013; Yang et al., 2018) is adopted to balance physical sensitivity of different parameters. In a three-parameter case (velocity, density, and attenuation), the preconditioner can be written as

$$P = \begin{bmatrix} s_1 diag\overline{H}^{(1)} & & \\ & s_2 diag\overline{H}^{(2)} & \\ & & s_3 diag\overline{H}^{(3)} \end{bmatrix}^{-1}. \tag{22}$$

Here, $s_i (i = 1, 2, 3)$ is determined according to the sensitivities of seismic data to the different model parameters. For instance, seismic data is most sensitive to velocity, we set s_1 as the smallest value. We think seismic data is least sensitive to density, and set s_2 for the density as the largest value. The value of s_3 for attenuation is between s_1 and s_3.

Wavefield recomputation techniques

In order to build the gradient vector (equation (10)) and Hessian-vector product (equations (18) and (19)), the forward and the backward wavefields have to be accessed simultaneously for the cross-correlation computation. One can implement this through one of the following four approaches in the time-domain FWI.

- The many numbers of incident wavefields snapshots are stored at Nyquist sampling rate during the incident wavefield extrapolation, and then loaded during solving the adjoint equation. It is

time-efficient but requires large storage, which is not feasible for large-scale applications.

- The final state and boundaries points for all time steps of incident wavefields are stored and then recomputed backwards in time, together with the adjoint wavefield using the reversibility property of the wave equation. However, this is not applicable when the medium is dissipative as the wave equation is no more reversible (see equation (15) in Yang et al. (2016b)).

- The checkpointing algorithms (Griewank and Walther, 2000; Symes, 2007) provide a balance between storage and time efficiency, in which a smaller number of snapshots are stored, called checkpoints. By recursive forward recomputation starting from checkpoints, it is possible to stably reconstruct incident fields in dissipative media.

- A checkpointing-assisted reverse-forward simulation (CARFS) combines the efficiency of the recomputation strategy and the stability of the checkpointing strategy for dissipative media. In the CARFS algorithm, the choice of forward modeling using checkpoints or reverse propagation is based on the minimum timestepping cost and an energy measure. Within tolerate accuracy loss, it is less computationally demanding than the checkpointing strategy (Yang et al., 2016b).

We will use the CARFS algorithm to test the time consumption of visco-acoustic FWI. The main computationally expensive steps for time-domain FN and GN Hessian-vector product construction are shown in Algorithm 2 and 3, respectively.

Algorithm 2 FN Hessian-vector product construction in time domain

for $it = 1$ to nt **do**
  update incident fields $\mathbf{w}(\mathbf{x}, t)$ and $\mathbf{u}(\mathbf{x}, t)$
end for
for $it = nt$ to 1 **do**
  update adjoint fields $\boldsymbol{\lambda}(\mathbf{x}, t)$ and $\boldsymbol{\mu}(\mathbf{x}, t)$
  reconstruct incident fields $\mathbf{w}(\mathbf{x}, t)$ and $\mathbf{u}(\mathbf{x}, t)$
  build the Hessian-vector product $H(\mathbf{m})\mathbf{r} = \langle \mathbf{u}, \frac{\partial A^{\dagger}(\mathbf{m})}{\partial \mathbf{m}} \boldsymbol{\lambda} \rangle_T + \langle \boldsymbol{\mu}, \frac{\partial A(\mathbf{m})}{\partial \mathbf{m}} \mathbf{w} \rangle_T + \langle \boldsymbol{\lambda}, \left(\frac{\partial^2 A(\mathbf{m})}{\partial \mathbf{m}^2} \mathbf{w} \right) \mathbf{r} \rangle_T$
end for

Algorithm 3 GN Hessian-vector product construction in time domain

for $it = 1$ to nt **do**
  update incident fields $\mathbf{w}(\mathbf{x}, t)$ and $\mathbf{u}(\mathbf{x}, t)$
end for
for $it = nt$ to 1 **do**
  update adjoint field $\boldsymbol{\mu}_1(\mathbf{x}, t)$
  reconstruct incident field $\mathbf{w}(\mathbf{x}, t)$
  build the Hessian-vector product $H(\mathbf{m})\mathbf{r} = \langle \boldsymbol{\mu}_1, \frac{\partial A(\mathbf{m})}{\partial \mathbf{m}} \mathbf{w} \rangle_T$
end for

FOURIER-DOMAIN FULL-SCATTERED-FIELD APPROXIMATION

We will first introduce discrete Fourier transform (DFT) to extract frequency-domain wavefields. Then we will show how to use full-scattered-field approximation to further reduce the computational complexity. Following subsection presents an error analysis of full-scattered-field approximation. Finally, we will make a comparison of computational complexity of Hessian-vector product construction among three methods, namely the time-domain implementation of Yang et al. (2018), the frequency-domain compression only and the frequency-domain full-scattered-field approximation.

On-the-fly Fourier transform

Considering the Parseval's identity with two arbitrary 1D time-domain signals $g_1(t)$ and $g_2(t)$, we have

$$\int_{\mathbb{R}} (g_1(t))^\dagger g_2(t) \mathrm{d}t = \int_{\mathbb{R}} (\tilde{g}_1(f))^\dagger \tilde{g}_2(f) \mathrm{d}f, \tag{23}$$

where $\tilde{g}_1(f)$ and $\tilde{g}_2(f)$ denote the frequency-domain representations of $g_1(t)$ and $g_2(t)$, respectively. Thus, the FN Hessian-vector product can be accurately expressed in the frequency domain as

$$H(\mathbf{m})\mathbf{r} = 2\int_0^{f_{max}} \mathcal{R}\left(\widetilde{\mathbf{u}}^\dagger \frac{\partial \widetilde{A}^\dagger(\mathbf{m})}{\partial \mathbf{m}} \widetilde{\boldsymbol{\lambda}} + \widetilde{\boldsymbol{\mu}}^\dagger \frac{\partial \widetilde{A}(\mathbf{m})}{\partial \mathbf{m}} \widetilde{\mathbf{w}} + \widetilde{\boldsymbol{\lambda}}^\dagger \left(\frac{\partial^2 \widetilde{A}(\mathbf{m})}{\partial \mathbf{m}^2} \widetilde{\mathbf{w}} \right) \mathbf{r} \right) \mathrm{d}f, \tag{24}$$

where $\widetilde{A}(\mathbf{m})$ is the forward modeling operator in frequency domain. $\mathcal{R}$ is the real part operator. $\widetilde{\mathbf{w}}(\mathbf{x}, f)$, $\widetilde{\boldsymbol{\lambda}}(\mathbf{x}, f)$, $\widetilde{\mathbf{u}}(\mathbf{x}, f)$, and $\widetilde{\boldsymbol{\mu}}(\mathbf{x}, f)$ denote frequency-domain wavefields.

In the following part, we consider approximating the Hessian-vector product in the frequency domain with a small number of frequencies. Thus, we can avoid the computationally intensive tasks of recomputing the incident wavefields for Hessian-vector product in the time domain. The

frequency-domain wavefields can be obtained by discrete Fourier transform.

Although Fast Fourier transform (FFT) has a high efficiency to obtain the frequency-domain representation of a signal, it requires storing the entire time-dependent wavefields ahead, which can not be satisfied in our case. Instead, DFT allows us to numerically implement the Fourier integral by accumulative summation over the time-loop of the finite difference scheme (Sirgue et al., 2008), which is adopted in this work and the on-the-fly DFT of the forward wavefield $\mathbf{w}(\mathbf{x}, t)$ is given by

$$\widetilde{\mathbf{w}}(\mathbf{x}, f) = \sum_{k=1}^{nt} exp(-2\pi i f k \Delta t)\mathbf{w}(\mathbf{x}, k\Delta t)\Delta t, \tag{25}$$

where $i = \sqrt{-1}$. Note that the temporal sampling interval Δt used for DFT is defined by the Nyquist theorem, not the one constrained by Courant-Friedrichs-Lewy (CFL) condition for the numerical stability of the finite-difference time scheme. In fact, the temporal sampling interval given by the Nyquist theorem is much larger than the one determined by CFL condition (see the analysis in APPENDIX C). This sub-sampling makes it possible to greatly reduce computational cost by using DFT to extract Fourier-domain wavefield instead of time-domain wavefield extrapolation.

After obtaining the four frequency-domain wavefields, we can approximate Hessian-vector product in the frequency domain.

$$H(\mathbf{m})\mathbf{r} \approx 2 \sum_{0}^{f_{max}} \mathcal{R}\left(\widetilde{\mathbf{u}}^{\dagger} \frac{\partial \widetilde{A}^{\dagger}(\mathbf{m})}{\partial \mathbf{m}} \widetilde{\boldsymbol{\lambda}} + \widetilde{\boldsymbol{\mu}}^{\dagger} \frac{\partial \widetilde{A}(\mathbf{m})}{\partial \mathbf{m}} \widetilde{\mathbf{w}} + \widetilde{\boldsymbol{\lambda}}^{\dagger} \left(\frac{\partial^2 \widetilde{A}(\mathbf{m})}{\partial \mathbf{m}^2} \widetilde{\mathbf{w}} \right) \mathbf{r} \right) \Delta f. \tag{26}$$

We compute the frequency-domain wavefields with an equal interval Δf. The workflows for FN and GN Hessian-vector product construction are presented in Algorithm 4 and Algorithm 5, respectively.

To perfectly reconstruct the causal time-domain signal $\mathbf{w}(\mathbf{x}, t), t \in (0, T)$ from frequency-

domain signal via inverse DFT, the frequency sampling interval needs to satisfy

$$\Delta f \leq \frac{1}{T}. \tag{27}$$

Theoretically, the number of frequencies should satisfy $nf \geq (f_{max} - f_{min})T$ to ensure no accuracy loss. In practice, we can not store all frequencies especially for 3D application. In this paper, we approximate the Hessian-vector product in Fourier domain using a small number of frequencies. For large-scale application, we have to balance between accuracy and efficiency. From the point view of signal analysis, the accuracy of undersampling Fourier-domain representation at one specific point is related to the sparsity of the time signal at this point. In our case, the accuracy of the Hessian-vector product, in the Fourier-domain approximation, depends on the medium complexity. For smooth models, the corresponding band-limited wavefield can be represented accurately with a relatively small number of discrete frequencies. In more complex models (for instance after FWI updates of the initial model) the complexity of the wavefield increases and the number of frequencies to consider to represent it with the same accuracy also increases.

One may consider using inverse DFT to obtain $\mathbf{w}(\mathbf{x}, t)$ and $\boldsymbol{\lambda}(\mathbf{x}, t)$ instead of solving wave equations in the Algorithm 4. However, for wavefields (re)computation of $\mathbf{u}$ and $\boldsymbol{\mu}$, we need time-domain wavefields of $\mathbf{w}(\mathbf{x}, t)$ and $\boldsymbol{\lambda}(\mathbf{x}, t)$ at the CFL interval, thus it can not benefit from the sub-sampling. In addition, we can not store a lot of frequencies to perfectly reconstruct time-domain wavefields. Therefore, it is mandatory to solve the wave equation again. In the next part, we will introduce full-scattered-field approximation to further simplify the computation.

Algorithm 4 FN Hessian-vector Product with DFT

$\widetilde{\mathbf{w}}(\mathbf{x}, f)$ and $\widetilde{\boldsymbol{\lambda}}(\mathbf{x}, f)$ are computed and stored in the gradient construction
for $it = 1$ to nt **do**
    update incident fields $\mathbf{w}(\mathbf{x}, t)$ and $\mathbf{u}(\mathbf{x}, t)$
    apply DFT to $\mathbf{u}(x, t)$ for obtaining $\widetilde{\mathbf{u}}(\mathbf{x}, f)$
end for
for $it = nt$ to 1 **do**
    compute adjoint source terms for $\boldsymbol{\mu}(\mathbf{x}, t)$
    update adjoint fields $\boldsymbol{\lambda}(\mathbf{x}, t)$ and $\boldsymbol{\mu}(\mathbf{x}, t)$
    apply DFT to $\boldsymbol{\mu}(x, t)$ for obtaining $\widetilde{\boldsymbol{\mu}}(\mathbf{x}, f)$
end for
build the Hessian-vector product $H(\mathbf{m})\mathbf{r} \approx 2\sum_0^{f_{max}} \mathcal{R}\left(\widetilde{\mathbf{u}}^\dagger \frac{\partial \widetilde{A}^\dagger(\mathbf{m})}{\partial \mathbf{m}} \widetilde{\boldsymbol{\lambda}} + \widetilde{\boldsymbol{\mu}}^\dagger \frac{\partial \widetilde{A}(\mathbf{m})}{\partial \mathbf{m}} \widetilde{\mathbf{w}} + \widetilde{\boldsymbol{\lambda}}^\dagger \left(\frac{\partial^2 \widetilde{A}(\mathbf{m})}{\partial \mathbf{m}^2} \widetilde{\mathbf{w}}\right) \mathbf{r}\right) \Delta f$

Algorithm 5 GN Hessian-vector Product with DFT

$\widetilde{\mathbf{w}}(\mathbf{x}, f)$ are computed and stored in the gradient construction
for $it = 1$ to nt **do**
    update incident fields $\mathbf{w}(\mathbf{x}, t)$ and $\mathbf{u}(\mathbf{x}, t)$
end for
for $it = nt$ to 1 **do**
    compute adjoint source terms for $\boldsymbol{\mu}_1(\mathbf{x}, t)$
    update adjoint field $\boldsymbol{\mu}_1(\mathbf{x}, t)$
    apply DFT to $\boldsymbol{\mu}_1(x, t)$ for obtaining $\widetilde{\boldsymbol{\mu}}_1(\mathbf{x}, f)$
end for
build the Hessian-vector product $H(\mathbf{m})\mathbf{r} \approx 2\sum_0^{f_{max}} \mathcal{R}\left(\widetilde{\boldsymbol{\mu}}_1{}^\dagger \frac{\partial \widetilde{A}(\mathbf{m})}{\partial \mathbf{m}} \widetilde{\mathbf{w}}\right) \Delta f$

Full-scattered-field approximation

It is clear that the first-order incident and adjoint wavefields ($\mathbf{w}$ and $\boldsymbol{\lambda}$) would not change during the inner loop of CG algorithm for solving Newton equation. However, we still need to solve the first-order incident equation (2) and adjoint equation (8) as they are involved in the source terms of the second-order incident and adjoint equations. In order to avoid the computationally intensive task of repeatedly solving first-order incident equation (2) and adjoint equation (8) when computing second-order incident $\mathbf{u}$ and adjoint $\boldsymbol{\mu}$, we propose to rely on subtraction of two wavefields in frequency domain to approximate the original Born modeling and further reduce the computational cost.

Let us consider wavefield difference generated by a small model perturbation $\mathbf{r}$

$$\begin{cases} A(\mathbf{m})\mathbf{w} = \mathbf{s}, \\ A(\mathbf{m}+\mathbf{r})\mathbf{w}' = \mathbf{s}, \\ \delta\mathbf{w} = \mathbf{w}' - \mathbf{w}. \end{cases} \tag{28}$$

$\mathbf{w}'$ is the total wavefield in the perturbed medium, which includes the background wavefield $\mathbf{w}$ and the full-scattered-field $\delta\mathbf{w}$ generated from the model perturbation $\mathbf{r}$. This full-scattered-field $\delta\mathbf{w}$ can be decomposed as

$$\delta\mathbf{w} = \delta\mathbf{w}_1 + \delta\mathbf{w}_2 + ..., \tag{29}$$

where $\delta\mathbf{w}_j$ ($j = 1, 2, ...$) is the jth-order Born scattered wavefield, which is recursively given by Schuster (2017):

$$\begin{cases} A(\mathbf{m})\delta\mathbf{w}_j = -\sum\limits_{m_i \in \mathbf{m}} r_i \frac{\partial A(\mathbf{m})}{\partial m_i}(\delta\mathbf{w}_{j-1}), \quad j = 1, 2, ... \\ \delta\mathbf{w}_0 = \mathbf{w}, \end{cases} \tag{30}$$

One can recognize that, for $j = 1$, equation (30) is similar to equation (16), meaning that $\mathbf{u}$ is the first-order Born wavefield $\delta\mathbf{w}_1$ caused by the model perturbation $\mathbf{r}$. It can be observed that $\mathbf{u}$ ($\delta\mathbf{w}_1$) is linearly related to the model perturbation $\mathbf{r}$, which is the reason why least-squares migration can be regarded as a linearized waveform inversion (Dai and Schuster, 2013; Yong et al., 2019).

From the previous development, assuming that $\mathbf{r}$ is small enough to neglect second and higher order scattered terms, we can approximate $\mathbf{u} \approx \delta\mathbf{w}$, and therefore compute $\mathbf{u}$ as the difference between $\mathbf{w}'$ and $\mathbf{w}$, which gives

$$\mathbf{u}(\mathbf{x}) \approx \mathbf{w}'(\mathbf{x}) - \mathbf{w}(\mathbf{x}). \tag{31}$$

By implementing this full-scattered-field approximation of $\mathbf{u}(\mathbf{x})$ in the Fourier domain, we only need to solve one additional wave equation for each Hessian-vector product, considering that $\widetilde{\mathbf{w}}(\mathbf{x}, f)$ is already known and stored. The second-order adjoint wavefields $\boldsymbol{\mu}_1$ in GN approximation can be obtained by back-propagating the approximate first-order Born wavefield at the receiver place.

For the second-order adjoint wavefields $\boldsymbol{\mu}$ in the FN method, the same strategy can be used, but leads to a small inaccuracy because of the $R^\dagger R\mathbf{u}$ in the perturbed model instead of the originally unperturbed model:

$$\begin{cases} A^\dagger(\mathbf{m})\boldsymbol{\lambda}(\mathbf{x},t) = R^\dagger(\mathbf{d} - R\mathbf{w}), \\ A^\dagger(\mathbf{m}+\mathbf{r})\boldsymbol{\lambda}'(\mathbf{x},t) = R^\dagger(\mathbf{d} - R\mathbf{w}) - R^\dagger R\mathbf{u}, \\ \boldsymbol{\mu}(\mathbf{x}) \approx \boldsymbol{\lambda}'(\mathbf{x}) - \boldsymbol{\lambda}(\mathbf{x}). \end{cases} \tag{32}$$

By using the Fourier-domain approximation of the wavefields, we end up with the Algorithm 6 and 7 for the main steps of our full-scattered-field approximation. In the original time-domain implementation, we have to carefully treat the Born source term in the second-order incident and adjoint equations. Using the proposed method, we can avoid this complex process, which makes code implementation easier.

Algorithm 6 FN Hessian-vector product with Fourier-domain full-scattered-field approximation

$\widetilde{\mathbf{w}}(\mathbf{x}, f)$ and $\widetilde{\boldsymbol{\lambda}}(\mathbf{x}, f)$ are computed and stored in the gradient construction
for $it = 1$ to nt **do**
    update incident fields by solving $A(\mathbf{m} + \mathbf{r})\mathbf{w}'(\mathbf{x}, t) = \mathbf{s}$
    apply DFT to $\mathbf{w}'(\mathbf{x}, t)$ for obtaining $\widetilde{\mathbf{w}}'(\mathbf{x}, f)$
end for
approximate second-order incident wavefield $\widetilde{\mathbf{u}}(\mathbf{x}, f) \approx \widetilde{\mathbf{w}}'(\mathbf{x}, f) - \widetilde{\mathbf{w}}(\mathbf{x}, f)$
approximate adjoint source for second-order adjoint wavefield $R^\dagger R\mathbf{u} \approx R^\dagger R\mathbf{w}_t - R^\dagger R\mathbf{w}$
for $it = nt$ to 1 **do**
    update adjoint fields by solving $A^\dagger(\mathbf{m} + \mathbf{r})\boldsymbol{\lambda}'(\mathbf{x}, t) = R^\dagger(\mathbf{d} - R\mathbf{w}) - R^\dagger R\mathbf{u}$
    apply DFT to $\boldsymbol{\lambda}'(\mathbf{x}, t)$ for obtaining $\widetilde{\boldsymbol{\lambda}}'(\mathbf{x}, f)$
end for
approximate second-order adjoint wavefield $\widetilde{\boldsymbol{\mu}}(\mathbf{x}, f) \approx \widetilde{\boldsymbol{\lambda}}'(\mathbf{x}, f) - \widetilde{\boldsymbol{\lambda}}(\mathbf{x}, f)$
build the Hessian-vector product $H(\mathbf{m})\mathbf{r} \approx 2\sum_0^{f_{max}} \mathcal{R}\left(\widetilde{\mathbf{u}}^\dagger \frac{\partial \widetilde{A}^\dagger(\mathbf{m})}{\partial \mathbf{m}} \widetilde{\boldsymbol{\lambda}} + \widetilde{\boldsymbol{\mu}}^\dagger \frac{\partial \widetilde{A}(\mathbf{m})}{\partial \mathbf{m}} \widetilde{\mathbf{w}} + \widetilde{\boldsymbol{\lambda}}^\dagger \left(\frac{\partial^2 \widetilde{A}(\mathbf{m})}{\partial \mathbf{m}^2} \widetilde{\mathbf{w}}\right) \mathbf{r}\right) \Delta f$

Algorithm 7 GN Hessian-vector product with Fourier-domain full-scattered-field approximation

$\widetilde{\mathbf{w}}(\mathbf{x}, f)$ are computed and stored in the gradient construction
for $it = 1$ to nt **do**
    update incident fields by solving $A(\mathbf{m} + \mathbf{r})\mathbf{w}'(\mathbf{x}, t) = \mathbf{s}$
    apply DFT to $\mathbf{w}'(\mathbf{x}, t)$ for obtaining $\widetilde{\mathbf{w}}'(\mathbf{x}, f)$
end for
approximate adjoint source for second-order adjoint wavefield $R^\dagger R\mathbf{u} \approx R^\dagger R\mathbf{w}' - R^\dagger R\mathbf{w}$
for $it = nt$ to 1 **do**
    update adjoint fields by solving $A^\dagger(\mathbf{m})\boldsymbol{\mu}_1(\mathbf{x}, t) = -R^\dagger R\mathbf{u}$
    apply DFT to $\widetilde{\boldsymbol{\mu}}_1(\mathbf{x}, f)$ for obtaining $\widetilde{\boldsymbol{\mu}}_1(\mathbf{x}, f)$
end for
build the Hessian-vector product $H(\mathbf{m})\mathbf{r} \approx 2\sum_0^{f_{max}} \mathcal{R}\left(\widetilde{\boldsymbol{\mu}}_1{}^\dagger \frac{\partial \widetilde{A}(\mathbf{m})}{\partial \mathbf{m}} \widetilde{\mathbf{w}}\right) \Delta f$

Error analysis of full-scattered-field approximation

Combining equations (29) and (30), the error of $\mathbf{u}$ can be expressed as

$$e_{\mathbf{u}} = \delta\mathbf{w} - \mathbf{u} = \sum_{j=2}^{\infty} \delta\mathbf{w}_j, \tag{33}$$

The error $e_{\mathbf{u}}$ is equal to the summation of all j-th ($j \geq 2$) scattering wavefields of the original incident wavefield $\mathbf{w}$. Using the recursive formula (30), $e_{\mathbf{u}}$ can be written as

$$\begin{aligned} e_{\mathbf{u}} &= \left((A(\mathbf{m}+\mathbf{r}))^{-1} - (A(\mathbf{m}))^{-1} \right) \left(- \sum_{m_i \in \mathbf{m}} r_i \frac{\partial A(\mathbf{m})}{\partial m_i} \mathbf{u} \right) \\ &= \underbrace{\mathbf{u}(\mathbf{m}+\mathbf{r}, \mathbf{r}) - \mathbf{u}(\mathbf{m}, \mathbf{r})}_{\delta\mathbf{u}}. \end{aligned} \tag{34}$$

Considering the definition of the full scattered field of the first-order incident wavefields (see the equation (28)), we can find that the error of $\mathbf{u}$ is the full scattered field of the second-order incident wavefields. Namely, the error $e_{\mathbf{u}}$ ($\delta\mathbf{u}$) can be understood as that one computes $\mathbf{u}$ using the second-order incident wave equation (16) with model parameter as $\mathbf{m}+\mathbf{r}$ instead of the correct parameter $\mathbf{m}$.

The error of second-order adjoint wavefield $\boldsymbol{\mu}_1$ in GN method comes from the second-order adjoint source, which can be given by

$$e_{\boldsymbol{\mu}_1} = \left(A^{\dagger}(\mathbf{m}) \right)^{-1} (-R^{\dagger} R \mathbf{e}_{\mathbf{u}}). \tag{35}$$

It has to be recalled that the original second-order adjoint wavefield in FN method is related to the sum of two terms: back-propagating $\mathbf{u}$ at the receivers position in the unperturbed media and the source term related to the perturbation interacting with the first-order adjoint wavefield. The error of

second-order adjoint wavefield $\boldsymbol{\mu}$ in FN method can be decomposed into three parts:

- using approximate adjoint source $R^\dagger R\mathbf{u}$,
- back-propagating the $R^\dagger R\mathbf{u}$ in the perturbed model instead of the unperturbed model,
- approximating first-order Born wavefield of first-order adjoint wavefield $\boldsymbol{\lambda}$ with the same subtraction strategy for $\mathbf{u}$.

Thus $e_{\boldsymbol{\mu}}$ in FN method can be expressed as

$$\begin{aligned} e_{\boldsymbol{\mu}} &= \left(A^\dagger(\mathbf{m}+\mathbf{r})\right)^{-1}(-R^\dagger R\mathbf{e}_{\mathbf{u}}) + \left(\left(A^\dagger(\mathbf{m}+\mathbf{r})\right)^{-1} - \left(A^\dagger(\mathbf{m})\right)^{-1}\right)(-R^\dagger R\mathbf{u}) + \sum_{i=2}^{\infty}\delta\boldsymbol{\lambda}_i \\ &= e_{\boldsymbol{\mu}_1} + \left(\left(A^\dagger(\mathbf{m}+\mathbf{r})\right)^{-1} - \left(A^\dagger(\mathbf{m})\right)^{-1}\right)\left(-R^\dagger R\mathbf{u} - \sum_{m_i\in\mathbf{m}} r_i\frac{\partial A^\dagger(\mathbf{m})}{\partial m_i}\boldsymbol{\lambda}\right) \\ &= e_{\boldsymbol{\mu}_1} + \underbrace{\boldsymbol{\mu}(\mathbf{m}+\mathbf{r},\mathbf{r}) - \boldsymbol{\mu}(\mathbf{m},\mathbf{r})}_{\delta\boldsymbol{\mu}}. \end{aligned} \tag{36}$$

Here, $\delta\boldsymbol{\mu}$ denotes the full scattered field of the second-order adjoint wavefields μ. It can also been understood as that one generates $\boldsymbol{\mu}$ using the second-order adjoint wave equation with model parameter $\mathbf{m}+\mathbf{r}$ instead of the correct model parameter $\mathbf{m}$.

In summary, the errors of second-order incident and adjoint wavefields introduced by full-scattered-field approximation is directly related to the model perturbation $\mathbf{r}$. The smaller the perturbation, the more accurate the approximation. In truncated Newton method, the model perturbation used to build Hessian-vector product is generally small, thus full-scattered-field approximation should be effective.

Computational complexity analysis

The computationally intensive steps of GN and FN Hessian-vector product construction are the wavefield simulations and the time-domain cross-correlation steps. Thanks to Fourier-domain compression, we do not need reconstructing the incident wavefields. In addition, with the help of full-scattered-field approximation, we successfully avoid first-order incident and adjoint wavefields (re)computation when obtaining second-order incident (GN and FN methods) and adjoint (only FN method) wavefields.

A comparison of the number of wavefields to be computed following the different mentioned strategies is proposed in the Table 1. By analyzing algorithm 2, we can observe that the original time-domain formulation leads to 6 wavefield simulations per FN Hessian-vector product: 2 forward fields, 2 backward fields and 2 recomputation of forward fields backward in time or with CARFS strategy. In algorithm 4, with the help of DFT, 2 forward and 2 backward simulations per FN Hessian-vector product are required, and no any recomputation steps, which can be quite intensive in particular in viscous media. Analyzing algorithm 6 shows that our frequency-domain full-scattered-field approximation only needs 1 forward and 1 backward wavefield simulations per FN Hessian-vector product. For GN Hessian-vector product, the time-domain formulation requires 4 wavefield simulations. Only relying on DFT, we need to compute 3 wavefield simulations. With full-scattered-field approximation, only 1 forward field and 1 backward field are required for per GN Hessian-vector product construction.

[Table 1 about here.]

NUMERICAL EXAMPLES

2D Valhall model and data

The numerical test is carried out with a 2D synthetic Valhall model. This synthetic model is built based on the geology of the Valhall oil field, which is a gas field located in the North Sea. Successful 3D FWI applications on the field data have made the possible high-resolution construction of 3D velocity, density and attenuation models (Sirgue et al., 2010; Operto et al., 2015; Kamath et al., 2021). With the local geological interpretation, the 2D synthetic models have been made to represent the shallow water environment, which contain horizontally stratified structures and gas bearing sediments. The presence of low-velocity gas layer yields a strong attenuation effect (amplitude decrease and dispersion) on wave propagation, which makes the imaging at the reservoir depth challenging.

The 2D multi-parameter Valhall models shown in Figure 3 (the first row) are defined on a regular grid with a size of $nz = 281, nx = 704$. The spatial interval is set to 12.5 m. Here, we consider velocity, density and attenuation. A fixed-spread acquisition is used, with 32 equally spaced sources and 351 equally spaced receivers with interval of 25 m placed on the surface. To increase the illumination for parameter inversion, we also position two vertical lines of 69 receivers with interval of 25 m close to the left and right boundaries of the model (▼ in Figure 3(a)).

The forward modeling operator $A(\mathbf{m})$ is a 2D visco-acoustic VTI time-domain wave equation (Yang et al., 2018) in an attenuative medium (see the APPENDIX E). The generalized Maxwell body (GMB) is applied to simulate attenuation effects, in which the number of the relaxation mechanisms is set as 3. The numerical tests are performed with the TOYxDAC_TIME package (Yang et al., 2018). We parallelly simulate all 32 shots on one node of our local cluster using the entire 32 cores. The synthetic data is generated with second order in time and fourth order in space finite-difference modeling. The source function is a Ricker wavelet with peak frequency of 5 Hz. The time

discretization step is set to 1.5 ms. The maximum frequency considered here is about 12.5 Hz and frequency-domain wavefields are equally sampled from 0-12.5 Hz.

The reconstruction of incident wavefields in time-domain method is based on the CARFS strategy (Yang et al., 2016b). We choose the tolerance error of 10^{-5} in CARFS to detect the deviation from the recorded energy when reconstructing the wavefield in the conventional time-domain implementation. In our numerical test, the ratio between reconstructing incident fields and direct forward modeling is about 2.5 for the CARFS, which could be 3.2 for the standard check-point method (Griewank and Walther, 2000; Symes, 2007). In the next part, we will use the result generated by the CARFS as a reference.

The initial models used to implement inversion are presented in Figure 3 (the second row). The initial velocity and density models are obtained by applying a Gaussian smoother with a radius of 20 points to the true models. The value of initial Q model below water layer is a constant (Q=200), and Q value in the water layer is fixed as the true one ($Q = 1000$) in the inversion. The values of velocity, density and Q have different magnitudes, besides they have different physical units. A unity-based normalization (Yang et al., 2018) is applied to handle this issue in the inversion.

$$\tilde{m}_i = \frac{m_i - m_i^{min}}{m_i^{max} - m_i^{min}}, \tag{37}$$

where $i = 1, 2, 3$ respectively denotes the parameter class of velocity, density and Q^{-1}. m_i^{min} and m_i^{max} are the lower and upper bounds of the physical parameter. Using this normalization strategy, the values of all physical parameter are restricted in the range $[0, 1]$. The normalized models are displayed in Figure 4 (the first row). Based on the chain rule, the gradient and Hessian-vector product will be scaled accordingly. The corresponding gradients are presented in the second row of Figure 4. The tuning parameters in the preconditioner (22) are set as $s_1 = 1$, $s_2 = 8$, $s_3 = 2$ for velocity,

density and Q^{-1}, respectively.

Theoretically, as the perturbation $\mathbf{r}$ decreases, full-scattered-field approximation becomes more accurate. In numerical implementation, it would be more robust to first scale the perturbation $\mathbf{r}$ to a small size by multiplying a scalar number when generating wavefield in perturbed media. To obtain the required Hessian-vector product, we have to rescale the wavefields $\mathbf{u}$ and $\boldsymbol{\mu}$. Note that scaling perturbation is only to make the wavefield more accurate and robust, and we do not scale the Hessian-vector product. In the following test, we scale the norm of the perturbation $\mathbf{r}$ to $0.1 \times ||\mathbf{g_0}||$, and $\mathbf{g_0}$ is the first gradient.

[Figure 3 about here.]

[Figure 4 about here.]

Effectiveness of full-scattered-field approximation

The error analysis on full-scattered-field approximation has been presented in the theory part. We will compare the time-domain wavefields to further test the effectiveness of full-scattered-field approximation. In this part, the negative preconditioned gradient is used as the model perturbation to obtain the second-order incident and adjoint wavefields, which in fact are used to construct the Hessian-vector product of the first iteration in truncated Newton method.

Figure 5 (a) presents the second-order incident wavefield at the receiver position. To clearly compare the difference of wavefields generated by exact first-order modeling and full-scattered-field approximation, we extract some traces from Figure 5 (a) and plot them in Figure 5 (b). It can be found that the error brought by full-scattered-field approximation is small. We also present snapshots of second-order incident wavefield ($\mathbf{u}$) and adjoint wavefields (GN $\boldsymbol{\mu}_1$ and FN $\boldsymbol{\mu}$). From Figure 6, we

can find that all wavefields and corresponding errors roughly differ by two orders of magnitude. The quantified L_1 errors by by full-scattered-field approximation in the wavefields above are displayed in Table 2.

[Figure 5 about here.]

[Figure 6 about here.]

[Table 2 about here.]

Efficiency and accuracy on Hessian-vector product construction

We first compare computational time for Hessian-vector products constructed by three approaches, namely the time-domain implementation of Yang et al. (2018), the frequency-domain compression, and frequency-domain full-scattered-field approximation, in which the negative gradient is used as the vector $\mathbf{r}$ for Hessian-vector product. Then we discuss the accuracy and efficiency of the two approximation methods and study the effect of frequencies used to build Hessian-vector product on the accuracy.

Figures 7 and 8 show the GN and FN Hessian-vector product provided by the three schemes for velocity, density and attenuation parameters. As Hessian plays the role of convolution operator, $H\mathbf{r}$ is a blurred version of the negative gradient. It can be noted that the two approximate method provides solutions very close to the reference version in the time-domain with 25 frequencies considered here. Since double scattering is considered in FN Hessian, we can observe that FN Hessian-vector product shown in Figure 8 contain more high-wavenumber information compared to GN Hessian-vector product.

The computational times of constructing one Hessian-vector product via the three approaches are listed in Table 3. The number of 75 frequencies is determined by Nyquist sampling theorem (equation (27)) to theoretically avoid aliasing. We compute and store the frequency-domain component of first-order incident and adjoint wavefields during gradient construction. We also give the elapsed times of gradient construction. Thanks to sub-sampling (equation (B-4)), the elapsed time increases slowly with frequencies. As expected, the two approximations proposed in this work makes it possible to reduce significantly the recomputation tasks and therefore reduce the "time-to-solution". It has to be noted that in attenuative media, the recomputation effort would be even bigger for the time-domain implementation, while frequency-domain approximation involve only forward-time propagation which are much less demanding. Relying on frequency-domain compression, the elapsed time can be roughly reduced by 40% and 50% for GN and FN Hessian-vector product construction, respectively. Thanks to full-scattered-field approximation, the reduction ratio can further reach 70% and 80% for GN and FN Hessian-vector product construction,respectively.

Table 4 lists the normalized L_1 errors of Hessian-vector product obtained by two approximate approaches. Here, the results obtained using time-domain formulation are used as reference. Compared to full-scattered-field approximation, the number of frequencies used to construct Hessian-vector product has a larger effect on the accuracy. Among three parameters, density is most sensitive to the number of the used frequencies. In addition, velocity is most robust to the approximate method. FN Hessian-vector product is less accurate than GN Hessian-vector product, which is consistent with the error analysis. Overall, we can approximate reference result using Fourier-domain full-scattered-field approximation within an acceptable error level.

[Figure 7 about here.]

[Figure 8 about here.]

[Table 3 about here.]

[Table 4 about here.]

Convergence performance comparison

We compare the convergence rate using the truncated Newton method based on the different Hessian-vector product construction method presented in this study. We terminate inversion when misfit reduces to 1 percent of the initial objective function. The maximum inner iteration number is set as 10 with initial $\eta = 0.9$ in the inner loop for solving the Newton equation. The SEISCOPE OPTIMIZATION TOOLBOX is employed regarding the overall truncated Newton strategy.

Figure 9 presents that the cumulative CG updates increase with outer iteration. From Figure 9, we can observe that, when 25 frequencies are used , the inversion convergence obtained with the truncated Newton method based on the two approximate methods for the Hessian-vector product are similar to the one obtained when no approximation is made (the original time-domain implementation). However, to reach the same error level, two approximate methods with 15 frequencies need more Hessian-vector product constructions. This again illustrates that frequencies used to do this approximation play an important role on accuracy. It can also been observed that the trend difference occurs earlier in FN method compared with GN method, which is consistent with the analysis that FN Hessian is more sensitive to approximation errors.

Figure 10 displays the data residual convergence rate with outer iteration. At the beginning, all approaches have similar convergence rate. With error accumulation, the convergence rate gradually changes. Overall, the time-domain method needs the least iterations to reach the stopping condition. However, the time-domain method takes the most elapsed time due to the expensive cost of Hessian-vector product construction. In comparison, Fourier-domain full-scattered-field method only takes

40% and 30% computational time of time-domain GN and FN method, respectively (see Figure 11). Figure 12 gives the comparison of convergence rate between ℓ-BFGS method and truncated Newton methods. Here, the same preconditioner is used in ℓ-BFGS method ($\ell = 10$). It can be observed that truncated GN methods take the least outer iterations to reach final misfit error level (Figure 12(a)). The ℓ-BFGS method requires the most outer iterations. Since ℓ-BFGS method does not need to compute Hessian-vector product, each iteration of ℓ-BFGS is much cheaper than truncated Newton method. Thanks to Fourier-domain compression and full-scattered-field approximation, our proposed method is competitive with ℓ-BFGS method, in term of elapsed time (Figure 12(b-c)).

[Figure 9 about here.]

[Figure 10 about here.]

[Figure 11 about here.]

[Figure 12 about here.]

Inversion result analysis

Figure 13 displays the final inverted results by truncated GN method. The final inverted results by truncated FN method are shown in Figure 14. It can be found that results generated by two approximate methods are similar to that obtained by the time-domain method, and the difference of results obtained by GN and FN are small. We also plot final results generated by ℓ-BFGS method shown in Figure 15. The extracted vertical profiles from final results are displayed in Figure 16. The main geological structures have been recovered by all methods. Since seismic data is most sensitive to velocity parameter, compared to density and attenuation, the reconstructed velocity

models match the true model best. We notice that the inversion results of the middle part in density model is distorted, which may be improved by using edge-preserving regularization method (Lin and Huang, 2014; Yong et al., 2018). Although the initial attenuation model is constant, FWI can recover the main features. We can also see that the reconstructed attenuation model has a low resolution, which has also been observed in the previous studies (Groos et al., 2014; Pan and Innanen, 2019). In realistic constant Q model, attenuative effect on seismic data increases with frequency (Wang, 2009; Zhu et al., 2013), therefore FWI using high-frequency data may improve the edge characteristic of attenuation model.

To better understand the parameter coupling, we plot the L_1 model reduction with L_2 data misfit. From Figure 17, it is clear that the decrease of data misfit does not guarantee a decrease of model misfit for all parameters. The model error of the dominating velocity parameter almost monotonically decreases with the data misfit, while the density and $1/Q$ model misfit may become larger with iterations. The use of truncated Newton method helps to mitigate this over-fitting issue to some extend. Since the proposed approximate methods have similar convergence rate, in terms of elapsed time, with ℓ-BFGS method, the developed method can be a promising alternative to ℓ-BFGS method.

[Figure 13 about here.]

[Figure 14 about here.]

[Figure 15 about here.]

[Figure 16 about here.]

[Figure 17 about here.]

DISCUSSION

We focus on the computational cost of Hessian-vector product construction. With on-the-fly DFT, we can approximate Hessian-vector product in frequency domain using a small number of frequencies, thus we do not need to reconstruct the incident wavefields to implement wavefield cross-correlation, which is a quite expensive step in time-domain method. Thank to full-scattered-field approximation, we can avoid solving the first-order incident and adjoint wave equations when simulating second-order incident and adjoint wavefields. Our method not only reduces computational cost but also simplifies the numerical implementation, which can promote the application of truncated Newton method to multi-parameter waveform inversion.

In the two approximations, frequencies have a more important influence on the accuracy. To perfectly represent time signal in frequency domain, the frequency interval depends on the length of the time series (equation (27)). When more physical parameter classes are inverted and geological structures become more complex, the interaction between different layers will become more complex and last for longer time, which makes the approximation less accurate as the frequency number is fixed. Numerical examples show that using 1/3 of the Nyquist frequencies can yield a convergence rate of the inversion method similar to what would have been obtained with the purely time-domain formula to reach 1% data residual. To reach 5% data residual, using 1/5 of the Nyquist frequencies is enough to get similar convergence rates (see Figure 10). The proposed method is naturally adapted for the practically used multi-scale strategy in FWI application. For the beginning of background velocity building, we can use a small number of frequencies, then gradually increase the numbers of frequencies. In realistic 3D application, the number of frequencies are limited and should be carefully determined by users according to the study case and the available memory resources. Since five grid points per minimum wavelength is generally used in 4th-order FD scheme forward modeling, we

can reduce memory requirement by space sub-sampling strategy that extracting frequency-domain wavefield on a coarse grid (two grid points per minimum wavelength), then interpolating the coarse-grid Hessian-vector product built in frequency domain to the fine-grid one used for model update. This can also further reduce computational time.

Numerical study exhibits that our "fine-grained" method can almost reach the same iterative convergence rate of time-domain with 70%-80% elapsed time save, which has not been achieved by these "coarse-grained" source encoding (Castellanos et al., 2015) and shot subsampling strategy (Matharu and Sacchi, 2019). Certainly, the proposed method can also combine with these "coarse-grained" methods to further reduce computational cost of Hessian-vector product construction.

It is clear that truncated Newton method has a fast iterative convergence rate compared with ℓ-BFGS method. With two approximations, the proposed methods requires a similar (slightly less) computational time as the ℓ-BFGS method to reach a target misfit, which, to our best knowledge, has not been achieved before. In addition, the Hessian information estimated by ℓ-BFGS method depends on the ℓ previous iterations. In large-scale applications, the shot subsampling strategy is often used in practice (Warner et al., 2013; Kamath et al., 2021). To combine subsampling strategy and ℓ-BFGS method, we have to keep the same pool of sources for a few iterations to make ℓ-BFGS effective. Truncated Newton methods are free from this kind of limitation.

Previous studies have also indicated that truncated FN method may be more robust than truncated GN method and ℓ-BFGS method when data contain contain strong imprint of multi-scattered events (Métivier et al., 2013). Our observation that truncated Newton method has the advantage to mitigate over-fitting issue over ℓ-BFGS inversion could be related here to a better behavior of such algorithms in the frame of multi-parameter reconstruction (Métivier et al., 2015). In addition, double scattering information in FN Hessian may help to exploit prismatic events for large-contrast salt inversion (Liu

et al., 2020).

To make the most of Hessian information, it is important to develop effective preconditioners (Martens et al., 2010; Yang et al., 2018) and study other matrix-free iteration methods for faster and more robustly solving the ill-posed Newton equation (Xu et al., 2020; Roosta et al., 2018). The proposed Hessian-vector product construction method can also been used in resolution analysis (Fichtner and Leeuwen, 2015) and uncertainty analysis (Tarantola, 2005; Virieux et al., 2017). In addition, the proposed approximations can also been applied to other large-scale PDE-constrained optimization problems.

CONCLUSION

A parsimonious truncated Newton approach has been developed for time-domain FWI thanks to Fourier-domain compression and full-scattered-field approximation. Discrete Fourier transform has been applied to extract frequency-domain component of wavefields, which allow us to avoid reconstructing incident wavefields. Using full-scattered-field approximation, we do not need solving first-order incident and adjoint equations for second-order incident and adjoint waveflieds computation. The computational time of Hessian-vector product construction by the proposed method, within affordable additional memory cost, can be reduced by about 70% and 80% in viscous media for GN and FN method, respectively. The effectiveness and efficiency of the developed method has also been verified in the multi-parameter inversion tests on a 2D realistic Valhall model.

ACKNOWLEDGMENTS

We are thankful to the editors, the reviewer Tom Dickens, and other two anonymous reviewers for the constructive comments and suggestions that helped to improve this manuscript. This study was partially funded by the SEISCOPE consortium (*http://seiscope2.osug.fr*), sponsored by AKERBP, CGG, CHEVRON, EQUINOR, EXXON-MOBIL, JGI, PETROBRAS, SCHLUMBERGER, SHELL, SINOPEC, SISPROBE and TOTAL. This study was granted access to the HPC resources of the Dahu platform of the CIMENT infrastructure (*https://ciment.ujf-grenoble.fr*), which is supported by the Rhône-Alpes region (GRANT CPER07_13 CIRA), the OSUG@2020 labex (reference ANR10 LABX56) and the Equip@Meso project (reference ANR-10-EQPX-29-01) of the programme Investissements d'Avenir supervised by the Agence Nationale pour la Recherche, and the HPC resources of CINES/IDRIS/TGCC under the allocation 046091 made by GENCI.

REFERENCES

Alkhalifah, T., and R.-É. Plessix, 2014, A recipe for practical full-waveform inversion in anisotropic media: An analytical parameter resolution study: Geophysics, **79**, R91–R101.

Brossier, R., S. Operto, and J. Virieux, 2009, Seismic imaging of complex onshore structures by 2d elastic frequency-domain full-waveform inversion: Geophysics, **74**, WCC105–WCC118.

Castellanos, C., L. Métivier, S. Operto, R. Brossier, and J. Virieux, 2015, Fast full waveform inversion with source encoding and second-order optimization methods: Geophysical Journal International, **200**, 720–744.

da Silva, N. V., G. Yao, and M. Warner, 2019, Semiglobal viscoacoustic full-waveform inversion: Geophysics, **84**, R271–R293.

Dai, W., and G. T. Schuster, 2013, Plane-wave least-squares reverse-time migration: Geophysics, **78**, S165–S177.

Eisenstat, S. C., and H. F. Walker, 1996, Choosing the forcing terms in an inexact newton method: SIAM Journal on Scientific Computing, **17**, 16–32.

Epanomeritakis, I., V. Akçelik, O. Ghattas, and J. Bielak, 2008, A newton-cg method for large-scale three-dimensional elastic full-waveform seismic inversion: Inverse Problems, **24**, 034015.

Fabien-Ouellet, G., E. Gloaguen, and B. Giroux, 2017, Time domain viscoelastic full waveform inversion: Geophysical Journal International, **209**, 1718–1734.

Fichtner, A., and T. v. Leeuwen, 2015, Resolution analysis by random probing: Journal of Geophysical Research: Solid Earth, **120**, 5549–5573.

Fichtner, A., and J. Trampert, 2011, Hessian kernels of seismic data functionals based upon adjoint techniques: Geophysical Journal International, **185**, 775–798.

Fornberg, B., 1988, Generation of finite difference formulas on arbitrarily spaced grids: Mathematics of computation, **51**, 699–706.

Griewank, A., and A. Walther, 2000, Algorithm 799: revolve: an implementation of checkpointing for the reverse or adjoint mode of computational differentiation: ACM Transactions on Mathematical Software (TOMS), **26**, 19–45.

Groos, L., M. Schäfer, T. Forbriger, and T. Bohlen, 2014, The role of attenuation in 2d full-waveform inversion of shallow-seismic body and rayleigh waves: Geophysics, **79**, R247–R261.

Innanen, K. A., 2014, Seismic avo and the inverse hessian in precritical reflection full waveform inversion: Geophysical Journal International, **199**, 717–734.

Kamath, N., R. Brossier, L. Métivier, A. Pladys, and P. Yang, 2021, Multiparameter full-waveform inversion of 3d ocean-bottom cable data from the valhall field: Geophysics, **86**, B15–B35.

Kamei, R., and R. Pratt, 2013, Inversion strategies for visco-acoustic waveform inversion: Geophysical Journal International, **194**, 859–884.

Knoll, D. A., and D. E. Keyes, 2004, Jacobian-free newton–krylov methods: a survey of approaches and applications: Journal of Computational Physics, **193**, 357–397.

Köhn, D., D. De Nil, A. Kurzmann, A. Przebindowska, and T. Bohlen, 2012, On the influence of model parametrization in elastic full waveform tomography: Geophysical Journal International, **191**, 325–345.

Lailly, P., 1983, The seismic inverse problem as a sequence of before stack migrations: Conference on Inverse Scattering, Theory and application, Society for Industrial and Applied Mathematics, Philadelphia, Conference on Inverse Scattering, Theory and application, Society for Industrial and Applied Mathematics, Philadelphia, 206–220.

Li, Y., R. Brossier, and L. Métivier, 2020, 3d frequency-domain elastic wave modeling with the spectral element method using a massively parallel direct solver: Geophysics, **85**, T71–T88.

Lin, Y., and L. Huang, 2014, Acoustic-and elastic-waveform inversion using a modified total-variation regularization scheme: Geophysical Journal International, **200**, 489–502.

Liu, Y., Z. Wu, H. Kang, and J. Yang, 2020, Use of prismatic waves in full-waveform inversion with the exact hessian: Geophysics, **85**, R325–R337.

Martens, J., et al., 2010, Deep learning via hessian-free optimization.: ICML, 735–742.

Matharu, G., and M. Sacchi, 2019, A subsampled truncated-newton method for multiparameter full-waveform inversion: Geophysics, **84**, R333–R340.

Métivier, L., F. Bretaudeau, R. Brossier, S. Operto, and J. Virieux, 2014, Full waveform inversion and the truncated newton method: quantitative imaging of complex subsurface structures: Geophysical Prospecting, **62**, 1353–1375.

Métivier, L., and R. Brossier, 2016, The seiscope optimization toolbox: A large-scale nonlinear optimization library based on reverse communication: Geophysics, **81**, F1–F15.

Métivier, L., R. Brossier, S. Operto, and J. Virieux, 2015, Acoustic multi-parameter fwi for the reconstruction of p-wave velocity, density and attenuation: Preconditioned truncated newton approach, *in* SEG Technical Program Expanded Abstracts 2015: Society of Exploration Geophysicists, 1198–1203.

———, 2017, Full waveform inversion and the truncated newton method: SIAM review, **59**, 153–195.

Métivier, L., R. Brossier, J. Virieux, and S. Operto, 2013, Full waveform inversion and the truncated newton method: SIAM Journal on Scientific Computing, **35**, B401–B437.

Nash, S. G., 2000, A survey of truncated-newton methods: Journal of computational and applied mathematics, **124**, 45–59.

Nguyen, B. D., and G. A. McMechan, 2015, Five ways to avoid storing source wavefield snapshots in 2d elastic prestack reverse time migration: Geophysics, **80**, S1–S18.

Nihei, K. T., and X. Li, 2007, Frequency response modelling of seismic waves using finite difference time domain with phase sensitive detection (td—psd): Geophysical Journal International, **169**, 1069–1078.

Nocedal, J., and S. Wright, 2006, Numerical optimization: Springer Science & Business Media.

Operto, S., Y. Gholami, V. Prieux, A. Ribodetti, R. Brossier, L. Metivier, and J. Virieux, 2013, A guided tour of multiparameter full-waveform inversion with multicomponent data: From theory to practice: The leading edge, **32**, 1040–1054.

Operto, S., and A. Miniussi, 2018, On the role of density and attenuation in three-dimensional multiparameter viscoacoustic vti frequency-domain fwi: An obc case study from the north sea: Geophysical Journal International, **213**, 2037–2059.

Operto, S., A. Miniussi, R. Brossier, L. Combe, L. Métivier, V. Monteiller, A. Ribodetti, and J. Virieux, 2015, Efficient 3-d frequency-domain mono-parameter full-waveform inversion of ocean-bottom cable data: application to valhall in the visco-acoustic vertical transverse isotropic approximation: Geophysical Journal International, **202**, 1362–1391.

Operto, S., J. Virieux, P. Amestoy, J.-Y. L'Excellent, L. Giraud, and H. B. H. Ali, 2007, 3d finite-difference frequency-domain modeling of visco-acoustic wave propagation using a massively parallel direct solver: A feasibility study: Geophysics, **72**, SM195–SM211.

Pan, W., and K. A. Innanen, 2019, Amplitude-based misfit functions in viscoelastic full-waveform inversion applied to walk-away vertical seismic profile data: Geophysics, **84**, B335–B351.

Pan, W., K. A. Innanen, and Y. Geng, 2018, Elastic full-waveform inversion and parametrization analysis applied to walk-away vertical seismic profile data for unconventional (heavy oil) reservoir characterization: Geophysical Journal International, **213**, 1934–1968.

Pan, W., K. A. Innanen, G. F. Margrave, M. C. Fehler, X. Fang, and J. Li, 2016, Estimation of elastic constants for hti media using gauss-newton and full-newton multiparameter full-waveform inversion: Geophysics, **81**, R275–R291.

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**, 495–503.

Pratt, R. G., C. Shin, and G. Hick, 1998, Gauss–newton and full newton methods in frequency–space seismic waveform inversion: Geophysical journal international, **133**, 341–362.

Prieux, V., R. Brossier, Y. Gholami, S. Operto, J. Virieux, O. Barkved, and J. Kommedal, 2011, On the footprint of anisotropy on isotropic full waveform inversion: the valhall case study: Geophysical Journal International, **187**, 1495–1515.

Roosta, F., Y. Liu, P. Xu, and M. W. Mahoney, 2018, Newton-mr: Newton's method without smoothness or convexity: arXiv preprint arXiv:1810.00303.

Saad, Y., 2003, Iterative methods for sparse linear systems: SIAM.

Schuster, G., 2017, Seismic inversion: Society of exploration geophysicists: Tusla, Ok.

Shen, X., I. Ahmed, A. Brenders, J. Dellinger, J. Etgen, and S. Michell, 2018, Full-waveform inversion: The next leap forward in subsalt imaging: The Leading Edge, **37**, 67b1–67b6.

Shewchuk, J. R., et al., 1994, An introduction to the conjugate gradient method without the agonizing pain.

Sirgue, L., O. I. Barkved, J. Dellinger, J. Etgen, U. Albertin, and J. H. Kommedal, 2010, Full waveform inversion: the next leap forward in imaging at Valhall: First Break, **28**, 65–70.

Sirgue, L., J. Etgen, and U. Albertin, 2008, 3d frequency domain waveform inversion using time domain finite difference methods: 70th EAGE Conference and Exhibition incorporating SPE EUROPEC 2008, European Association of Geoscientists & Engineers, cp–40.

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

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

———, 2005, Inverse problem theory and methods for model parameter estimation: SIAM.

Trinh, P.-T., R. Brossier, L. Métivier, L. Tavard, and J. Virieux, 2019, Efficient time-domain 3d

elastic and viscoelastic full-waveform inversion using a spectral-element method on flexible cartesian-based mesh: Geophysics, **84**, R61–R83.

Tromp, J., 2020, Seismic wavefield imaging of earth's interior across scales: Nature Reviews Earth & Environment, **1**, 40–53.

Tromp, J., C. Tape, and Q. Liu, 2005, Seismic tomography, adjoint methods, time reversal and banana-doughnut kernels: Geophysical Journal International, **160**, 195–216.

van Leeuwen, T., and F. J. Herrmann, 2015, A penalty method for pde-constrained optimization in inverse problems: Inverse Problems, **32**, 015007.

Vigh, D., K. Jiao, D. Watts, and D. Sun, 2014, Elastic full-waveform inversion application using multicomponent measurements of seismic data collection: Geophysics, **79**, R63–R77.

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

Virieux, J., A. Asnaashari, R. Brossier, L. Métivier, A. Ribodetti, and W. Zhou, 2017, An introduction to full waveform inversion, *in* Encyclopedia of exploration geophysics: Society of Exploration Geophysicists, R1–1.

Virieux, J., and S. Operto, 2009, An overview of full-waveform inversion in exploration geophysics: Geophysics, **74**, WCC1–WCC26.

Wang, H., O. Burtz, P. Routh, D. Wang, J. Violet, R. Lu, and S. Lazaratos, 2021, Anisotropic 3d elastic full-wavefield inversion to directly estimate elastic properties and its role in interpretation: The Leading Edge, **40**, 277–286.

Wang, Y., 2009, Seismic inverse q filtering: John Wiley & Sons.

Warner, M., A. Ratcliffe, T. Nangoo, J. Morgan, A. Umpleby, N. Shah, V. Vinje, I. Štekl, L. Guasch, C. Win, et al., 2013, Anisotropic 3d full-waveform inversion: Geophysics, **78**, R59–R80.

Xu, P., F. Roosta, and M. W. Mahoney, 2020, Newton-type methods for non-convex optimization

under inexact hessian information: Mathematical Programming, **184**, 35–70.

Xu, S., D. Wang, F. Chen, G. Lambaré, and Y. Zhang, 2012, Inversion on reflected seismic wave, *in* SEG Technical Program Expanded Abstracts 2012: Society of Exploration Geophysicists, 1–7.

Yang, J., Y. Liu, and L. Dong, 2016a, Simultaneous estimation of velocity and density in acoustic multiparameter full-waveform inversion using an improved scattering-integral approach: Geophysics, **81**, R399–R415.

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

———, 2016c, Wavefield reconstruction in attenuating media: A checkpointing-assisted reverse-forward simulation method: Geophysics, **81**, R349–R362.

Yang, P., R. Brossier, L. Métivier, J. Virieux, and W. Zhou, 2018, A time-domain preconditioned truncated newton approach to visco-acoustic multiparameter full waveform inversion: SIAM Journal on Scientific Computing, **40**, B1101–B1130.

Yang, P., R. Brossier, and J. Virieux, 2016d, Wavefield reconstruction by interpolating significantly decimated boundaries: Geophysics, **81**, T197–T209.

Yong, P., J. Huang, Z. Li, W. Liao, and L. Qu, 2019, Least-squares reverse time migration via linearized waveform inversion using a wasserstein metric: Geophysics, **84**, S411–S423.

Yong, P., W. Liao, J. Huang, and Z. Li, 2018, Total variation regularization for seismic waveform inversion using an adaptive primal dual hybrid gradient method: Inverse Problems, **34**, 045006.

Zhu, T., J. M. Carcione, and J. M. Harris, 2013, Approximating constant-q seismic propagation in the time domain: Geophysical prospecting, **61**, 931–940.

APPENDIX A

ANALYSIS AND REASSEMBLY OF HESSIAN-VECTOR PRODUCT

We will first briefly build the connection between gradient and Hessian-vector product, then give the derivation of GN and FN Hessian-vector product. Most of the proofs are not entirely mathematically rigorous to keep the derivation readable.

The relation between gradient and Hessian-vector product

As Hessian is the derivative of the gradient (10) with respect to model parameter $\mathbf{m}$, we can intuitively infer that Hessian matrix can be decomposed into three terms:

$$H(\mathbf{m}) = \nabla^2_{\mathbf{m}}\chi(\mathbf{m}) = \langle \boldsymbol{\lambda}, \frac{\partial A(\mathbf{m})}{\partial \mathbf{m}} \frac{\partial \mathbf{w}}{\partial \mathbf{m}} \rangle_T + \langle \frac{\partial \boldsymbol{\lambda}}{\partial \mathbf{m}}, \frac{\partial A(\mathbf{m})}{\partial \mathbf{m}} \mathbf{w} \rangle_T + \langle \boldsymbol{\lambda}, \frac{\partial^2 A(\mathbf{m})}{\partial \mathbf{m}^2} \mathbf{w} \rangle_T. \tag{A-1}$$

To understand the relationships between the three terms here and those terms in the equation (18), we directly differentiate $\phi_{\mathbf{r}}(\mathbf{m}, \mathbf{w}(\mathbf{m}), \boldsymbol{\lambda}(\mathbf{m}))$ with respect to model parameters $\mathbf{m}$, and the Hessian-vector product with chain rules can be written as

$$\begin{aligned}(H(\mathbf{m})\mathbf{r})_j = & \underbrace{\sum_{m_i \in \mathbf{m}} r_i \int_0^T \mathrm{d}t \boldsymbol{\lambda}^\dagger \frac{\partial A}{\partial m_j} \frac{\partial \mathbf{w}}{\partial m_i}}_{H_1} + \underbrace{\sum_{m_i \in \mathbf{m}} r_i \int_0^T \mathrm{d}t \left(\frac{\partial \boldsymbol{\lambda}}{\partial m_i} \right)^\dagger \frac{\partial A}{\partial m_j} \mathbf{w}}_{H_2} \\ & + \underbrace{\sum_{m_i \in \mathbf{m}} r_i \int_0^T \mathrm{d}t \boldsymbol{\lambda}^\dagger \left(\frac{\partial^2 A}{\partial m_j \partial m_i} \right) \mathbf{w}}_{H_3},\end{aligned} \tag{A-2}$$

The element in FN Hessian-vector product (18) can be denoted by

$$(H(\mathbf{m})\mathbf{r})_j = \underbrace{\int_0^T \mathrm{d}t\mathbf{u}^\dagger \frac{\partial A^\dagger(\mathbf{m})}{\partial m_j}\boldsymbol{\lambda}}_{P_1} + \underbrace{\int_0^T \mathrm{d}t\boldsymbol{\mu}^\dagger \frac{\partial A(\mathbf{m})}{\partial m_j}\mathbf{w}}_{P_2} + \underbrace{\sum_{m_i\in\mathbf{m}} r_i \int_0^T \mathrm{d}t\boldsymbol{\lambda}^\dagger \left(\frac{\partial^2 A(\mathbf{m})}{\partial m_j \partial m_i}\right)\mathbf{w}}_{P_3} \tag{A-3}$$

It is obvious that the H_3 in equation (A-2) is exactly equivalent to the P_3 in equation (A-3). We will give a brief proof to demonstrate that the H_1 and H_2 in equation (A-2) are equivalent to the P_1 and P_2 in equation (A-3), respectively.

With the state equation (2) in the first-order adjoint method, we have

$$\frac{\partial A(\mathbf{m})}{\partial \mathbf{m}}\mathbf{w} + A(\mathbf{m})\frac{\partial \mathbf{w}}{\partial \mathbf{m}} = 0 \Rightarrow \frac{\partial \mathbf{w}}{\partial \mathbf{m}} = -\left(A(\mathbf{m})\right)^{-1}\frac{\partial A(\mathbf{m})}{\partial \mathbf{m}}\mathbf{w} \tag{A-4}$$

Inserting equation (A-4) into the H_1 in the equation (A-2), then combining the equation (16) yields

$$\begin{aligned}(H_1)_j &= \int_0^T \mathrm{d}t \left(\sum_{m_i\in\mathbf{m}} r_i \frac{\partial \mathbf{w}}{\partial m_i}\right)^\dagger \frac{\partial A^\dagger(\mathbf{m})}{\partial m_j}\boldsymbol{\lambda} \\ &= \int_0^T \mathrm{d}t \left((A(\mathbf{m}))^{-1}\sum_{m_i\in\mathbf{m}} -r_i \frac{\partial A(\mathbf{m})}{\partial m_i}\mathbf{w}\right)^\dagger \frac{\partial A^\dagger(\mathbf{m})}{\partial m_j}\boldsymbol{\lambda} \\ &= \underbrace{\int_0^T \mathrm{d}t\mathbf{u}^\dagger \frac{\partial A^\dagger(\mathbf{m})}{\partial m_j}\boldsymbol{\lambda}}_{P_1}\end{aligned} \tag{A-5}$$

Considering the adjoint equation (8) in the first-order adjoint method, we have

$$\frac{\partial A^\dagger(\mathbf{m})}{\partial \mathbf{m}}\boldsymbol{\lambda} + A^\dagger(\mathbf{m})\frac{\partial \boldsymbol{\lambda}}{\partial \mathbf{m}} = -R^\dagger R\frac{\partial \mathbf{w}}{\partial \mathbf{m}} \Rightarrow \frac{\partial \boldsymbol{\lambda}}{\partial \mathbf{m}} = \left(A^\dagger(\mathbf{m})\right)^{-1}\left(-R^\dagger R\frac{\partial \mathbf{w}}{\partial \mathbf{m}} - \frac{\partial A^\dagger(\mathbf{m})}{\partial \mathbf{m}}\boldsymbol{\lambda}\right). \tag{A-6}$$

Inserting equation (A-6) into the H_2 in the equation (A-2), then combining the equation (17) yields

$$\begin{aligned}(H_2)_j &= \sum_{m_i \in \mathbf{m}} r_i \int_0^T \mathrm{d}t \left(\frac{\partial \boldsymbol{\lambda}}{\partial m_i}\right)^\dagger \frac{\partial A(\mathbf{m})}{\partial m_j}\mathbf{w} \\ &= \int_0^T \mathrm{d}t \left((A(\mathbf{m}))^{-1}\left(-R^\dagger R\mathbf{u} - \sum_{m_i \in \mathbf{m}} r_i \frac{\partial A^\dagger(\mathbf{m})}{\partial m_i}\boldsymbol{\lambda}\right)\right)^\dagger \frac{\partial A(\mathbf{m})}{\partial m_j}\mathbf{w} \\ &= \underbrace{\int_0^T \mathrm{d}t \boldsymbol{\mu}^\dagger \frac{\partial A(\mathbf{m})}{\partial m_j}\mathbf{w}}_{P_2}\end{aligned} \tag{A-7}$$

Now it is clear that the P_1, P_2, and P_3 in the equation (A-2) are respectively related to the derivatives of $\mathbf{w}(\mathbf{m})$, $\boldsymbol{\lambda}(\mathbf{m})$, and $A(\mathbf{m})$ in the gradient with respect to $\mathbf{m}$.

Gauss-Newton and full-Newton Hessian-vector product

The assembly of Hessian-vector product above is directly obtained by the second-order adjoint-state method. In fact, there is an another way to build the Hessian-vector product when one wants a semi-positive Hessian approximation. Let us further go back to the original misfit (1), the gradient can be rewritten as

$$\nabla_{\mathbf{m}}\chi(\mathbf{m}) = \int_0^T \mathrm{d}t \left(\frac{\partial \mathbf{w}}{\partial \mathbf{m}}\right)^\dagger R^\dagger(R\mathbf{w} - \mathbf{d}), \tag{A-8}$$

and the Hessian can be also rewritten as

$$\nabla^2_{\mathbf{m}}\chi(\mathbf{m}) = \int_0^T \mathrm{d}t \left(\frac{\partial \mathbf{w}}{\partial \mathbf{m}}\right)^\dagger R^\dagger R \frac{\partial \mathbf{w}}{\partial \mathbf{m}} + \left(\frac{\partial^2 \mathbf{w}}{\partial \mathbf{m}^2}\right)^\dagger R^\dagger(R\mathbf{w} - \mathbf{d}) \tag{A-9}$$

The first term in the right side of the equation (A-9) represents the product of two first-order scattering, while the second term denotes double scattering (Fichtner and Trampert, 2011). Note that the first term is semi-positive, which is a good property for solving systems of linear equations. If the problem is weakly nonlinear, we can neglect the second term, and full Hessian reduces to GN Hessian.

Applying the relation in the equation (A-4), GN Hessian-vector product can be given by

$$
\begin{aligned}
(H_{GN}(\mathbf{m})\mathbf{r})_j &= \int_0^T \mathrm{d}t \left(\frac{\partial \mathbf{w}}{\partial m_j}\right)^\dagger R^\dagger R \left(\sum_{m_i \in \mathbf{m}} r_i \frac{\partial \mathbf{w}}{\partial m_i}\right) \\
&= \int_0^T \mathrm{d}t \left(-\left(A(\mathbf{m})\right)^{-1} \frac{\partial A(\mathbf{m})}{\partial m_j}\mathbf{w}\right)^\dagger R^\dagger R \underbrace{\left(\left(A(\mathbf{m})\right)^{-1} \sum_{m_i \in \mathbf{m}} -r_i \frac{\partial A(\mathbf{m})}{\partial m_i}\mathbf{w}\right)}_{\mathbf{u}} \\
&= \int_0^T \mathrm{d}t \left(\frac{\partial A(\mathbf{m})}{\partial m_j}\mathbf{w}\right)^\dagger \underbrace{\left(-A^\dagger(\mathbf{m})\right)^{-1} R^\dagger R \mathbf{u}}_{\boldsymbol{\mu}_1} \\
&= \int_0^T \mathrm{d}t \boldsymbol{\mu}_1^\dagger \frac{\partial A(\mathbf{m})}{\partial m_j}\mathbf{w}
\end{aligned} \tag{A-10}
$$

Here, the adjoint-state variable $\boldsymbol{\mu}_1$ obtained by solving

$$
A(\mathbf{m})\mathbf{u} = -\sum_i r_i \frac{\partial A(\mathbf{m})}{\partial m_i}\mathbf{w}, \tag{A-11}
$$

$$
A^\dagger(\mathbf{m})\boldsymbol{\mu}_1 = -R^\dagger R\mathbf{u}. \tag{A-12}
$$

Compared with FN Hessian-vector product in the equation (A-3), GN Hessian-vector product has the same structure with the second term. Note that the source term in the equation (A-12) does not contain the Born scattering of the first-order adjoint wavefield $\boldsymbol{\lambda}$ in the equation (17), which means that we only need to solve three equations to build the GN Hessian-vector product.

For strongly nonlinear problem, the second term in the right side of the equation (A-9) plays an important roles. However, most of studies focus on the GN term. Here, we take a look at the second term. To obtain the second-order derivative $\frac{\partial^2 \mathbf{w}}{\partial \mathbf{m}^2}$, we differentiate the state equation (2) in the first-order adjoint method with respect to $\mathbf{m}$ twice. Thus we get

$$\frac{\partial^2 \mathbf{w}}{\partial m_i \partial m_j} = -\left(A(\mathbf{m})\right)^{-1}\left(\frac{\partial^2 A(\mathbf{m})}{\partial m_i \partial m_i}\mathbf{w} + \frac{\partial A(\mathbf{m})}{\partial m_i}\frac{\partial \mathbf{w}}{\partial m_j} + \frac{\partial A(\mathbf{m})}{\partial m_j}\frac{\partial \mathbf{w}}{\partial m_i}\right) \tag{A-13}$$

Substituting the equation (A-13) into the second-order term in the equation (A-9) yields

$$\begin{aligned}
(H_{2nd}(\mathbf{m})\mathbf{r})_j &= \sum_{m_i \in \mathbf{m}} r_i \int_0^T \mathrm{d}t \left(\frac{\partial^2 A(\mathbf{m})}{\partial m_j \partial m_i}\mathbf{w}\right)^{\dagger} \underbrace{\left(A(\mathbf{m})^{\dagger}\right)^{-1} R^{\dagger}(\mathbf{d} - R\mathbf{w})}_{\boldsymbol{\lambda}} \\
&+ \sum_{m_i \in \mathbf{m}} r_i \int_0^T \mathrm{d}t \left(\frac{\partial A(\mathbf{m})}{\partial m_i}\frac{\partial \mathbf{w}}{\partial m_j} + \frac{\partial A(\mathbf{m})}{\partial m_j}\frac{\partial \mathbf{w}}{\partial m_i}\right)^{\dagger} \underbrace{\left(A(\mathbf{m})^{\dagger}\right)^{-1} R^{\dagger}(\mathbf{d} - R\mathbf{w})}_{\boldsymbol{\lambda}} \\
&= \underbrace{\int_0^T \mathrm{d}t \boldsymbol{\lambda}^{\dagger} \left(\frac{\partial A(\mathbf{m})}{\partial m_j}\left(A(\mathbf{m})\right)^{-1} \sum_{m_i \in \mathbf{m}} -r_i \frac{\partial A(\mathbf{m})}{\partial m_i}\mathbf{w}\right)}_{P_1} + \underbrace{\sum_{m_i \in \mathbf{m}} r_i \int_0^T \mathrm{d}t \boldsymbol{\lambda}^{\dagger} \frac{\partial^2 A(\mathbf{m})}{\partial m_j \partial m_i}\mathbf{w}}_{P_3} \\
&+ \int_0^T \mathrm{d}t \left(\frac{\partial A(\mathbf{m})}{\partial m_j}\mathbf{w}\right)^{\dagger} \left(\left(A^{\dagger}(\mathbf{m})\right)^{-1} \sum_{m_i \in \mathbf{m}} -r_i \frac{\partial A^{\dagger}(\mathbf{m})}{\partial m_i}\boldsymbol{\lambda}\right) \\
&= \int_0^T \mathrm{d}t (\boldsymbol{\mu} - \boldsymbol{\mu}_1)^{\dagger} \frac{\partial A(\mathbf{m})}{\partial m_j}\mathbf{w} + P_1 + P_3
\end{aligned} \tag{A-14}$$

It is obvious that two Hessian-vector product assemblies above are equivalent. In fact, Hessian information also has strong relationships with reflection FWI (Xu et al., 2012). $H_{GN}(\mathbf{m})\mathbf{r}$ and $H_{2nd}(\mathbf{m})\mathbf{r}$ can be connected to migration and tomography kernels in reflection FWI.

APPENDIX B

FROM CFL TO NYQUIST SAMPLING

Due to the numerical stability condition, the temporal interval defined by Courant-Friedrichs-Lewy (CFL) condition for wavefield extrapolation is much less than that determined by the Nyquist sampling theorem. In practice, we compute Fourier-domain wavefield with the Nyquist temporal interval instead of CFL one. In this appendix, we will compute the ratio between Nyquist temporal interval and CFL one to show that how much we can benefit from the subsampling strategy to implement DFT.

To sample a band-limited signal with maximum frequency f_{max}, according to Nyquist sampling theorem, the temporal sampling interval Δt_1 should meet

$$\Delta t_1 \leq \frac{1}{2f_{max}}. \tag{B-1}$$

When solving the wave equation with finite-difference method, the temporal sampling interval Δt_2 are defined by the CFL condition:

$$\Delta t_2 \leq \frac{\Delta x}{v_{\max}} \frac{1}{\sqrt{D}\sum_{i=1}^{N}|a_i|}, \tag{B-2}$$

where a_i is the finite difference (FD) coefficients on the staggered grid of order $2N$ (Virieux, 1986; Fornberg, 1988); $v_{\max}$ is the maximum velocity; D is the number of dimension.

Here we consider a typical 3D case with 4th-order spatial FD scheme. Generally, five grid points per minimum wavelength can ensure accurate wave propagation simulation with 4th-order FD scheme. Thus, we have

$$\Delta x = \frac{v_{\min}}{5f_{\max}} \Rightarrow \Delta t_2 \leq \frac{0.49487\Delta x}{v_{\max}} \approx \frac{1}{10f_{\max}}\frac{v_{\min}}{v_{\max}}, \tag{B-3}$$

where $v_{\min}$ is minimum velocity of the media. Comparing the Nyquist theorem (B-1) and CFL condition (B-3), one can find the practical implementation of DFT on the fly can be downsampled with a ratio ζ without accuracy loss (Yang et al., 2016d).

$$\zeta_{3D} = \left(\frac{0.5}{f_{\max}}\right) / \left(\frac{0.1}{f_{\max}}\frac{v_{\min}}{v_{\max}}\right) = \frac{5v_{\max}}{v_{\min}}, \quad \zeta_{2D} = \sqrt{\frac{2}{3}}\zeta_{3D} \approx \frac{4v_{\max}}{v_{\min}}. \tag{B-4}$$

The down-sample rate depends on the minimum and maximum velocities. In general, $v_{\max}/v_{\min} \geq 3$ for acoustic media. This means that we are able to theoretically reduce at least 15 times computational complexity using Nyquist temporal interval instead of the temporal interval for wavefield simulation. In 3D elastic media, the subsample rate can reach 50 due to the high ratio between the minimum S-wave and maximum P-wave velocities (Yang et al., 2016d), which can significantly save computational time.

In this paper, we implement acoustic wavefield extrapolation with finite difference method, which is one of the most widely used methods in exploration seismology. Therefore, we discuss the ratio between Nyquist temporal interval and CFL temporal interval in finite difference method. One may use other numerical methods (e.g., spectral element method, finite volume method, discontinuous galerkin method, and so) to implement wavefield extrapolation, the value of ratio may change but we still will benefit a lot when using Nyquist temporal sampling interval for DFT computation.

APPENDIX C

GRADIENT AND HESSIAN-VECTOR PRODUCT IN VTI VISCOACOUSTIC MEDIA

In the previous sections, we use generic gradient and Hessian-vector expression to explain our algorithm. In this paper, we will apply the proposed algorithm to FWI in VTI viscoacoustic media.

Now, we will give the formula to compute gradient and Hessian-vector product, and related wave equations involved in the numerical studies. Following the symbols used in the previous study (Yang et al., 2018), the VTI viscoacoustic wave equation can be given by

$$
\begin{cases}
\rho\partial_t \underbrace{\begin{bmatrix} v_x \\ v_y \\ v_z \end{bmatrix}}_{\mathbf{v}} = \underbrace{\begin{bmatrix} \partial_x & 0 \\ \partial_y & 0 \\ 0 & \partial_z \end{bmatrix}}_{D^T} \underbrace{\begin{bmatrix} g \\ q \end{bmatrix}}_{\boldsymbol{\sigma}} + \underbrace{\begin{bmatrix} f_{v_x} \\ f_{v_y} \\ f_{v_z} \end{bmatrix}}_{\mathbf{f_v}} \\
\partial_t \underbrace{\begin{bmatrix} g \\ q \end{bmatrix}}_{\boldsymbol{\sigma}} = \underbrace{\begin{bmatrix} c_{11} & c_{13} \\ c_{13} & c_{33} \end{bmatrix}}_{C} \underbrace{\begin{bmatrix} \partial_x & \partial_y & 0 \\ 0 & 0 & \partial_z \end{bmatrix}}_{D} \underbrace{\begin{bmatrix} v_x \\ v_y \\ v_z \end{bmatrix}}_{\mathbf{v}} - \sum_{\ell=1}^{L} Y_\ell \underbrace{\begin{bmatrix} c_{11} & c_{13} \\ c_{13} & c_{33} \end{bmatrix}}_{C} \underbrace{\begin{bmatrix} \xi_\ell^g \\ \xi_\ell^q \end{bmatrix}}_{\boldsymbol{\xi}_\ell} + \underbrace{\begin{bmatrix} f_g \\ f_q \end{bmatrix}}_{\mathbf{f}_\sigma} \\
\partial_t \underbrace{\begin{bmatrix} \xi_\ell^g \\ \xi_\ell^q \end{bmatrix}}_{\boldsymbol{\xi}_\ell} = -\omega_\ell \underbrace{\begin{bmatrix} \xi_\ell^g \\ \xi_\ell^q \end{bmatrix}}_{\boldsymbol{\xi}_\ell} + \omega_\ell \underbrace{\begin{bmatrix} \partial_x & \partial_y & 0 \\ 0 & 0 & \partial_z \end{bmatrix}}_{D} \underbrace{\begin{bmatrix} v_x \\ v_y \\ v_z \end{bmatrix}}_{\mathbf{v}}.
\end{cases}
\tag{C-1}
$$

Here, $\mathbf{v} = (v_x, v_y, v_z)^T$ are particle velocities. g and q are stresses. $\boldsymbol{\xi}_\ell$ are known as memory variables with the frequency ω_ℓ, and $Y_\ell \approx y_\ell Q_{inv}$ ($Q_{inv} = 1/Q$) with the separable approximation for the anelastic coefficients (Yang et al., 2016c), and y_ℓ is obtained before forward simulation by solving

$$
\min_{y_\ell} \frac{1}{2} \int_{\omega \in \bar{\Omega}} \left(\sum_{\ell=1}^{n} y_\ell \frac{\omega \omega_\ell}{\omega_\ell^2 + \omega^2} - 1 \right)^2 \mathrm{d}\omega \tag{C-2}
$$

where $\bar{\Omega} = [\omega_{min}, \omega_{max}]$, and ω_ℓ are often chosen as ω_{min}, ω_{max}, and $\sqrt{\omega_{min}\omega_{max}}$. The element of matrix C is related to the elastic coefficients

$$c_{11} = \rho V_p^2(1+2\epsilon) = \rho V_h^2 \quad c_{33} = \rho V_p^2 = \kappa \quad c_{13} = c_{33}\sqrt{1+2\delta} = \rho V_p^2\sqrt{1+2\delta}, \tag{C-3}$$

where ϵ and δ are Thomsen's anisotropy parameters. In this study, they are simply set as zero for isotropic media. For the sake of simplicity, we rewrite the first-order incident equation in a compact form as

$$\mathbf{w}: \begin{cases} \rho\partial_t \mathbf{v} = -D^\dagger \boldsymbol{\sigma} + \mathbf{f}_\mathbf{v} \\ C^{-1}\partial_t \boldsymbol{\sigma} = D\mathbf{v} - \sum_{\ell=1}^{L} Y_\ell \boldsymbol{\xi}_\ell + C^{-1}\mathbf{f}_\sigma \\ \frac{1}{\omega_\ell}\partial_t \boldsymbol{\xi}_\ell = -\boldsymbol{\xi}_\ell + D\mathbf{v}, \ell = 1, \cdots, L \end{cases} . \tag{C-4}$$

With first-order adjoint-state method (Plessix, 2006), the first-order adjoint equation can be given by

$$\boldsymbol{\lambda}: \begin{cases} \rho\partial_t \bar{\mathbf{v}} + D^\dagger \bar{\boldsymbol{\sigma}} + \sum_{\ell=1}^{L} D^\dagger \bar{\boldsymbol{\xi}}_\ell = \Delta d_\mathbf{v} \\ C^{-1}\partial_t \bar{\boldsymbol{\sigma}} - D\bar{\mathbf{v}} = \Delta d_\sigma \\ \frac{1}{\omega_\ell}\partial_t \bar{\boldsymbol{\xi}}_\ell - \bar{\boldsymbol{\xi}}_\ell - Y_\ell \bar{\boldsymbol{\sigma}} = 0, \ell = 1, \cdots, L. \end{cases} \tag{C-5}$$

Under the parameterization $\mathbf{m} = (\kappa_{inv}, \rho, Q_{inv})$ ($\kappa_{inv} := 1/\kappa$), the gradients of the misfit in (10) can be written as

$$\frac{\partial \chi}{\partial \rho} = \int_0^T \mathrm{d}t \bar{\mathbf{v}}^\dagger \partial_t \mathbf{v}, \quad \frac{\partial \chi}{\partial \kappa_{inv}} = \int_0^T \mathrm{d}t \bar{\boldsymbol{\sigma}}^\dagger \frac{\partial C^{-1}}{\partial \kappa_{inv}} \partial_t \boldsymbol{\sigma}, \quad \frac{\partial \chi}{\partial Q_{inv}} = \int_0^T \mathrm{d}t \bar{\boldsymbol{\sigma}}^\dagger \sum_{\ell=1}^{L} y_\ell \boldsymbol{\xi}_\ell. \tag{C-6}$$

Using the second-order adjoint-state method (Fichtner and Trampert, 2011; Métivier et al., 2013), the second-order incident and adjoint wavefields can be obtained by solving

$$\mathbf{u}:\begin{cases}\rho\partial_t\bar{\bar{\mathbf{v}}}^1+D^\dagger\bar{\bar{\boldsymbol{\sigma}}}^1=-r_\rho\partial_t\mathbf{v}\\ C^{-1}\partial_t\bar{\bar{\boldsymbol{\sigma}}}^1-D\bar{\bar{\mathbf{v}}}^1+\sum_{\ell=1}^{L}Y_\ell\bar{\bar{\boldsymbol{\xi}}}_\ell^1=-r_{\kappa_{inv}}\frac{\partial C^{-1}}{\partial\kappa_{inv}}\partial_t\boldsymbol{\sigma}-r_{Q_{inv}}\sum_{\ell=1}^{L}\frac{\partial Y_\ell}{\partial Q_{inv}}\boldsymbol{\xi}_\ell\\ \frac{1}{\omega_\ell}\partial_t\bar{\bar{\boldsymbol{\xi}}}_\ell^1+\bar{\bar{\boldsymbol{\xi}}}_\ell^1-D\bar{\bar{\mathbf{v}}}^1=0,\ell=1,\cdots,L\end{cases}\tag{C-7}$$

and

$$\boldsymbol{\mu}:\begin{cases}\rho\partial_t\bar{\bar{\mathbf{v}}}^2+D^\dagger\bar{\bar{\boldsymbol{\sigma}}}^2+\sum_{\ell=1}^{L}D^\dagger\bar{\bar{\boldsymbol{\xi}}}_\ell^2=-r_\rho\partial_t\bar{\mathbf{v}}+f_{\bar{\bar{\mathbf{v}}}^1}\\ C^{-1}\partial_t\bar{\bar{\boldsymbol{\sigma}}}^2-D\bar{\bar{\mathbf{v}}}^2=-r_{\kappa_{inv}}\frac{\partial C^{-1}}{\partial\kappa_{inv}}\partial_t\bar{\boldsymbol{\sigma}}+f_{\bar{\bar{\boldsymbol{\sigma}}}^1}\\ \frac{1}{\omega_\ell}\partial_t\bar{\bar{\boldsymbol{\xi}}}_\ell^2-\bar{\bar{\boldsymbol{\xi}}}_\ell^2-Y_\ell\bar{\bar{\boldsymbol{\sigma}}}^2=r_{Q_{inv}}\frac{\partial Y_\ell}{\partial Q_{inv}}\bar{\boldsymbol{\sigma}},\ell=1,\cdots,L.\end{cases}\tag{C-8}$$

Here, $f_{\bar{\bar{\mathbf{v}}}^1}$ and $f_{\bar{\bar{\boldsymbol{\sigma}}}^1}$ are the data extracted at receiver locations from $\mathbf{u}$:

$$R^\dagger R\mathbf{u}=(f_{\bar{\bar{\mathbf{v}}}^1},f_{\bar{\bar{\sigma}}^1},0,\cdots,0)^\dagger.\tag{C-9}$$

It is easy to find that $\frac{\partial^2 A(\mathbf{m})}{\partial\mathbf{m}^2}=0$ with the parameterization $\mathbf{m}=(\kappa_{inv},\rho,Q_{inv})$ ($\kappa_{inv}:=1/\kappa$). Thus the Hessian-vector product with respect to the selected parameters can be efficiently computed by Yang et al. (2018)

$$\begin{aligned}(H(\mathbf{m})\mathbf{r})|_{m_j=\rho}&=\int_0^T\mathrm{d}t(-\bar{\bar{\mathbf{v}}}^{1\dagger}\partial_t\bar{\mathbf{v}}+\bar{\bar{\mathbf{v}}}^{2\dagger}\partial_t\mathbf{v}),\\ (H(\mathbf{m})\mathbf{r})|_{m_j=\kappa_{inv}}&=\int_0^T\mathrm{d}t(-\bar{\bar{\boldsymbol{\sigma}}}^{1\dagger}\frac{\partial C^{-1}}{\partial\kappa_{inv}}\partial_t\bar{\boldsymbol{\sigma}}+\bar{\bar{\boldsymbol{\sigma}}}^{2\dagger}\frac{\partial C^{-1}}{\partial\kappa_{inv}}\partial_t\boldsymbol{\sigma}),\\ (H(\mathbf{m})\mathbf{r})|_{m_j=Q_{inv}}&=\int_0^T\mathrm{d}t(\sum_{\ell=1}^{L}y_\ell\bar{\bar{\boldsymbol{\xi}}}_\ell^{1\dagger}\bar{\boldsymbol{\sigma}}+\bar{\bar{\boldsymbol{\sigma}}}^{2\dagger}\sum_{\ell=1}^{L}y_\ell\boldsymbol{\xi}_\ell).\end{aligned}\tag{C-10}$$

With Fourier transform, one can easily obtain the frequency-domain formula on Hessian-vector product construction. In addition, after obtaining the gradients with equation (C-6) and Hessian-vector

products with equation (C-10), one can transform them to obtain the gradients and Hessian-vector products with other parameterizations by the chain rule.

LIST OF FIGURES

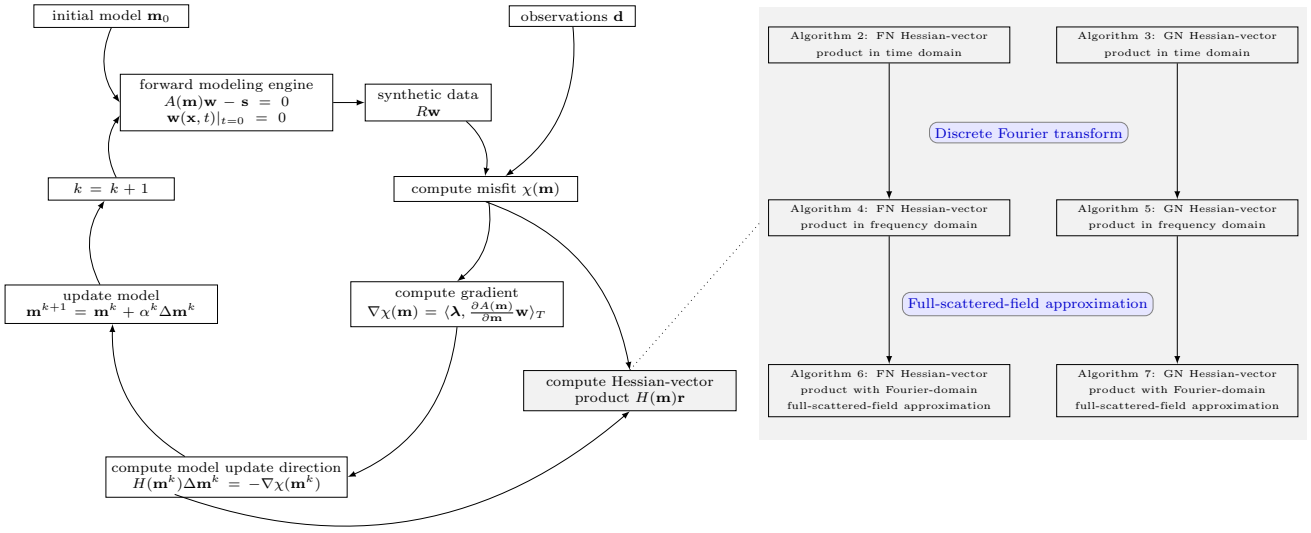
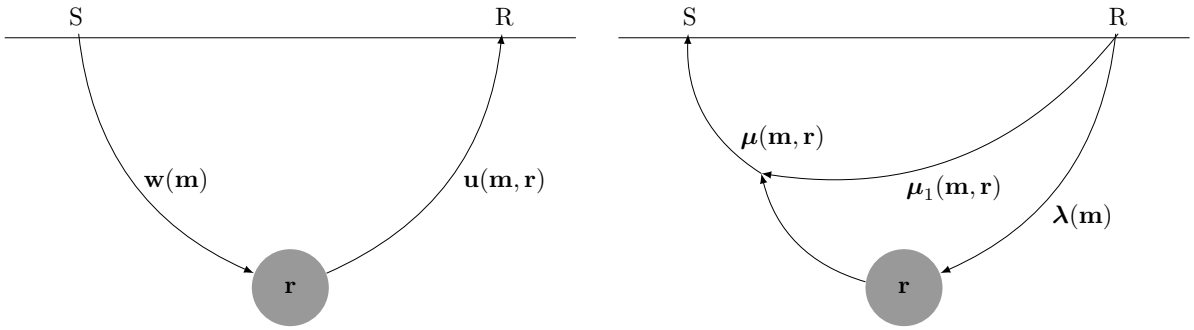


Figure 1: A brief workflow of the truncated Newton method. Note that the vector $\mathbf{r}$ used in Hessian-vector product construction is given by the CG algorithm during solving the Newton equation.



(a) incident wavefields (b) adjoint wavefields

Figure 2: The first-order incident wavefield $\mathbf{w}$ starts from the source location S, and the second-order incident wavefield $\mathbf{u}$ is generated by the first-order Born interaction between $\mathbf{w}$ and the given vector $\mathbf{r}$. The second-order adjoint wavefields $\boldsymbol{\mu}$ in FN method includes two parts. The first part is $\boldsymbol{\mu}_1$, used in the GN method, which is computed by back-propagating $-\mathbf{u}$ at the receiver location R. The second part is related to the first-order Born interaction between the first-order adjoint wavefield $\boldsymbol{\lambda}$ and the given vector $\mathbf{r}$.

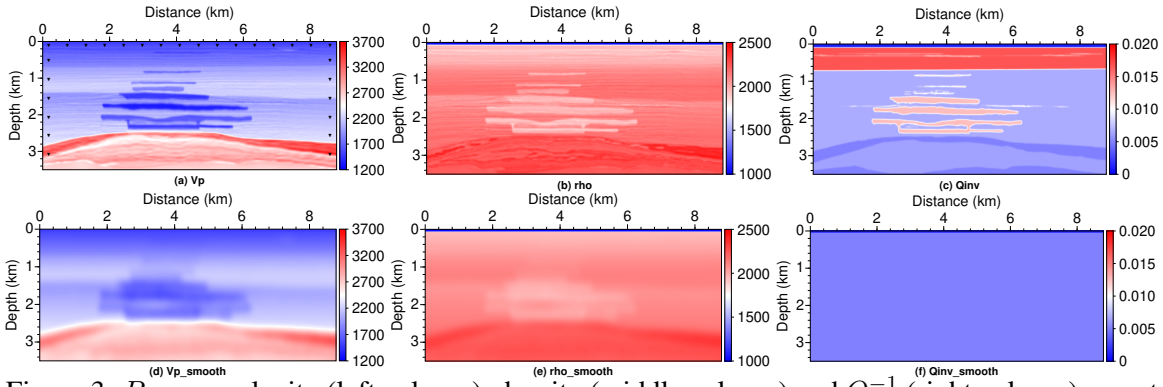


Figure 3: P-wave velocity (left column), density (middle column) and Q^{-1} (right column): exact models (first row), and initial models (second row).

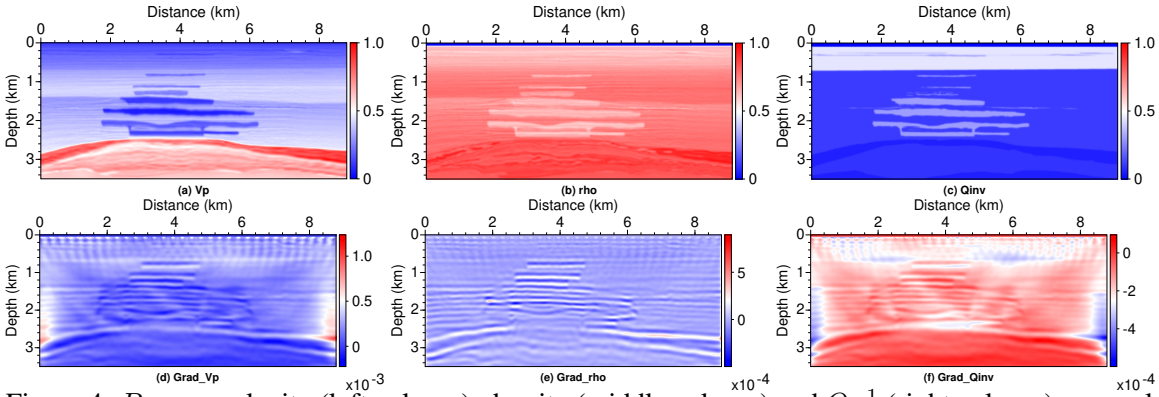


Figure 4: P-wave velocity (left column), density (middle column) and Q^{-1} (right column): normalized models (first row) and preconditioned gradients of the 1st iteration (second row).

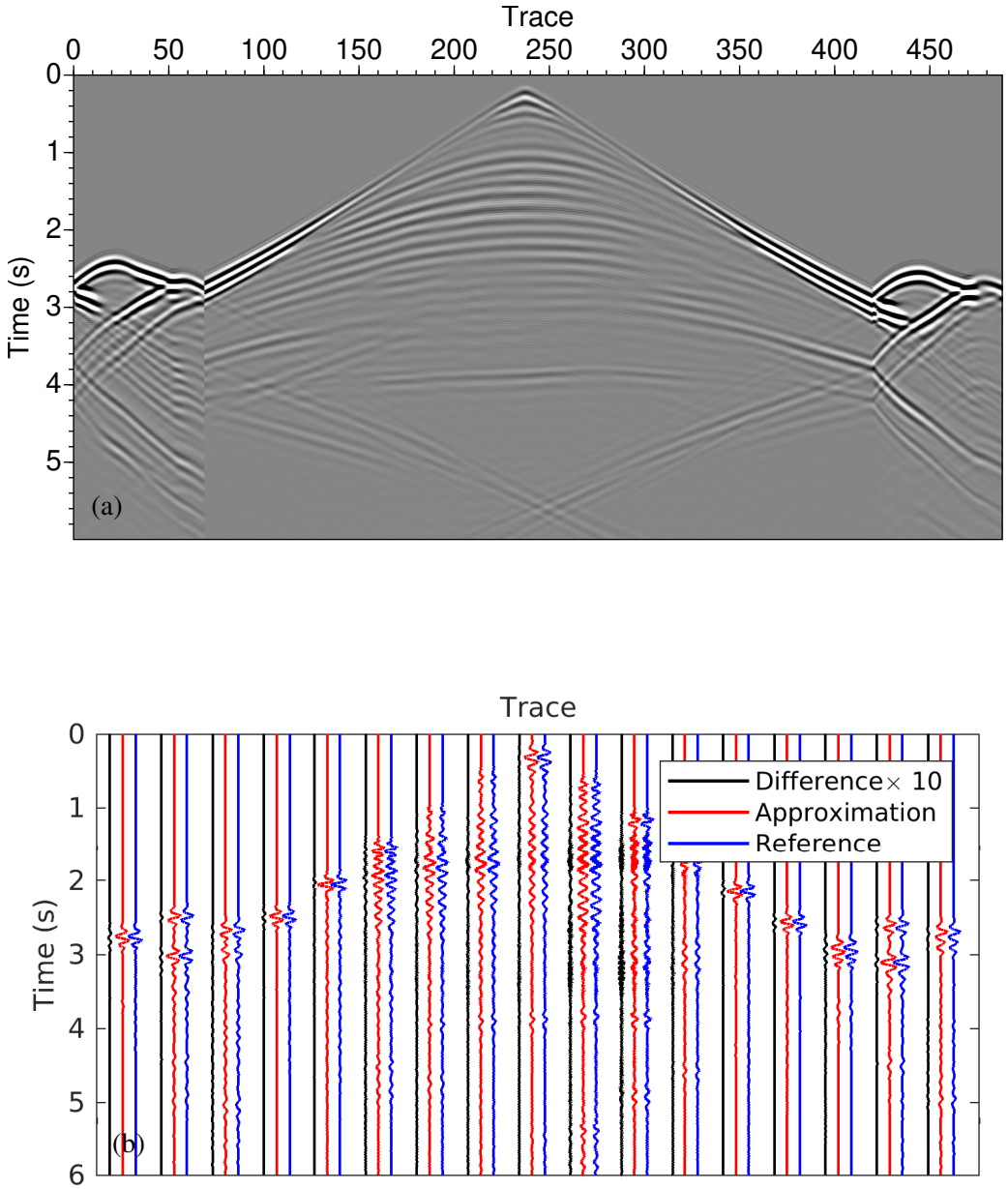


Figure 5: Comparison of $R\mathbf{u}$ between first-order Born modeling and full-scattered-field approximation: (a) seismic record of the second-order incident wavefields $R\mathbf{u}$. The source is located at the middle of the surface, (b) red lines represent wavefields obtained by first-order Born modeling, and blue lines represent wavefields generated by full-scattered-field approximation. The 10-time magnified difference (black lines) is small. Receivers are put on the left, top and right sides.

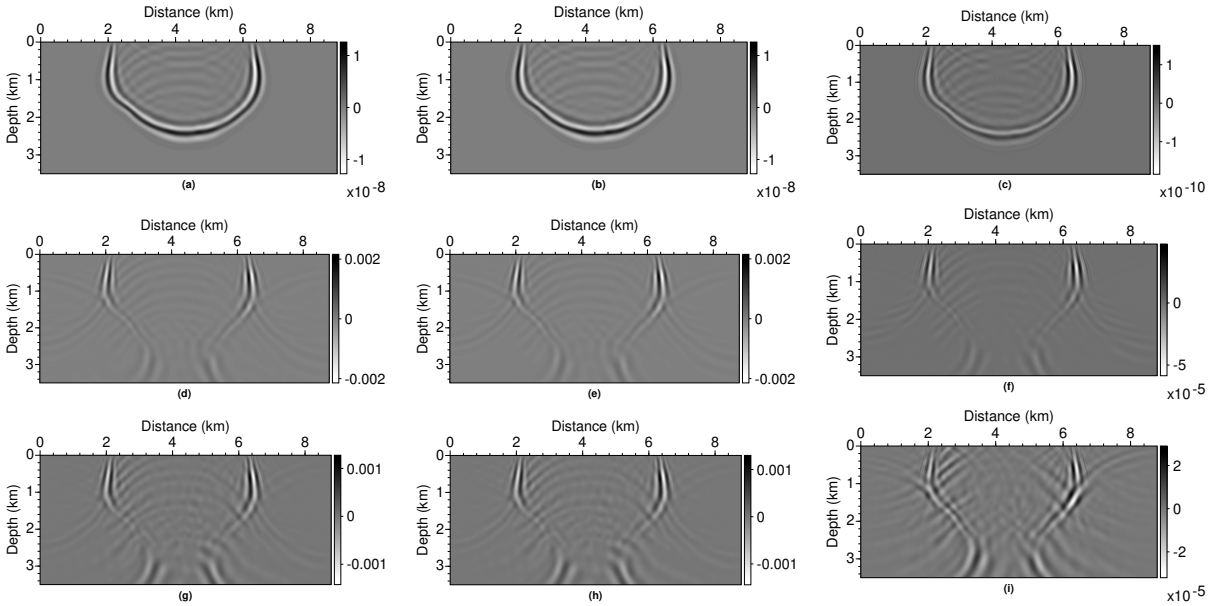


Figure 6: Comparison of the wavefields between first-order Born modeling (left column) and full-scattered-field approximation (middle column) using snapshots at $t = 1.5s$, and the corresponding differences are presented in the right column: second-order incident wavefields $\mathbf{u}$ (first row), second-order adjoint wavefields in GN method $\boldsymbol{\mu}_1$, and second-order adjoint wavefields in FN method $\boldsymbol{\mu}$.

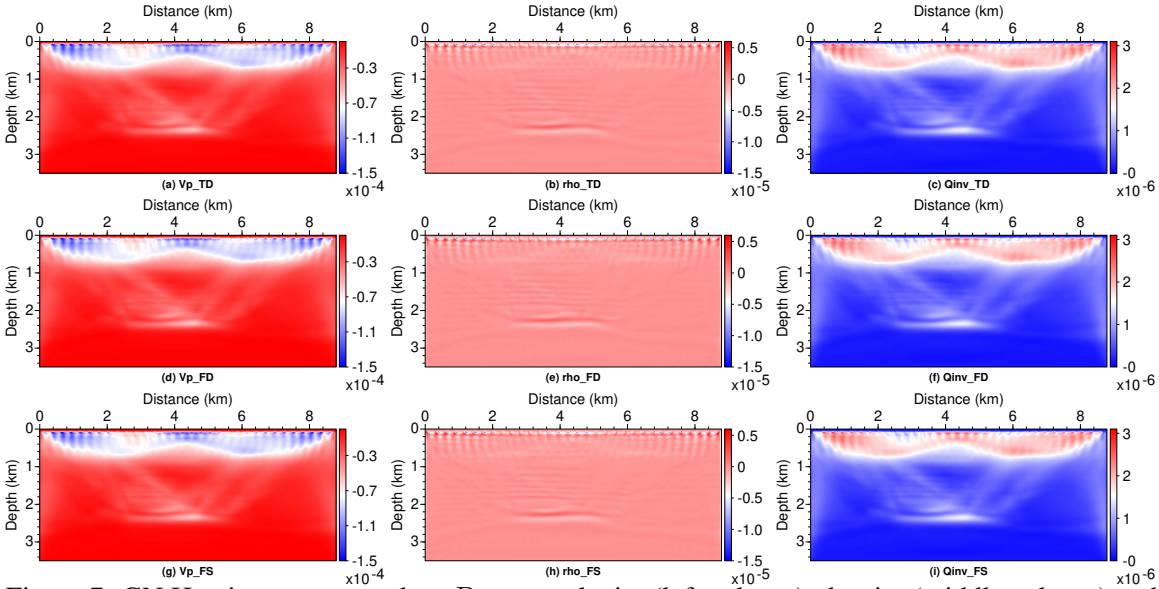


Figure 7: GN Hessian-vector product:P-wave velocity (left column), density (middle column) and Q^{-1} (right column). Time-domain implementation (first row), frequency-domain approximation (second row) and full-scattered-field approach in frequency domain (third row). 25 frequencies are used in two approximate methods.

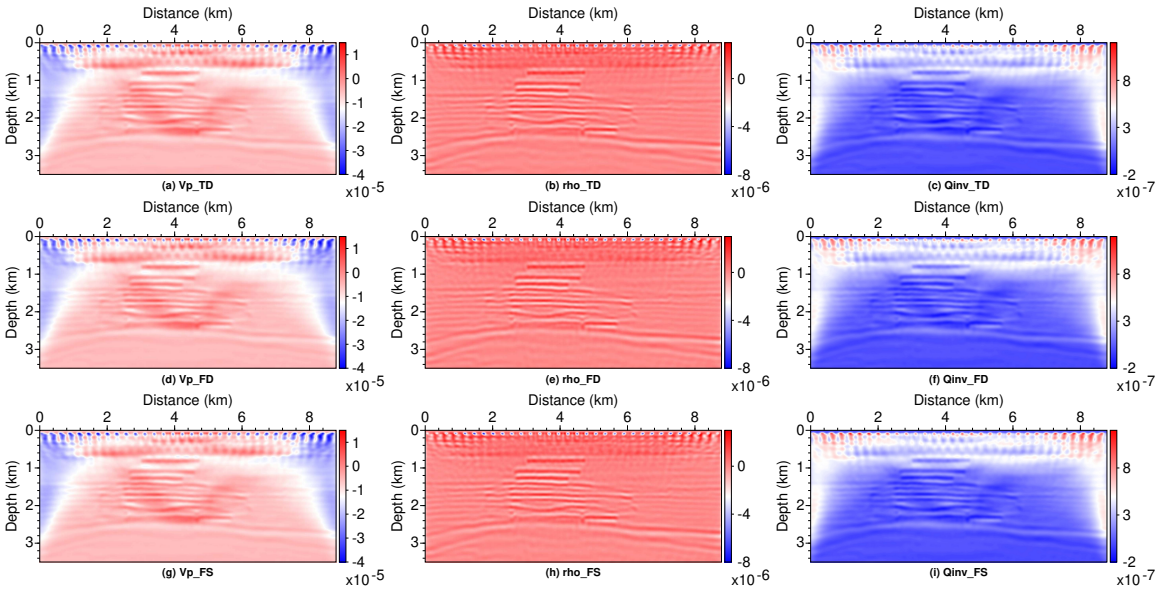


Figure 8: FN Hessian-vector product:P-wave velocity (left column), density (middle column) and Q^{-1} (right column). Time-domain implementation (first row), frequency-domain approximation (second row) and full-scattered-field approach in frequency domain (third row). 25 frequencies are used in two approximate methods.

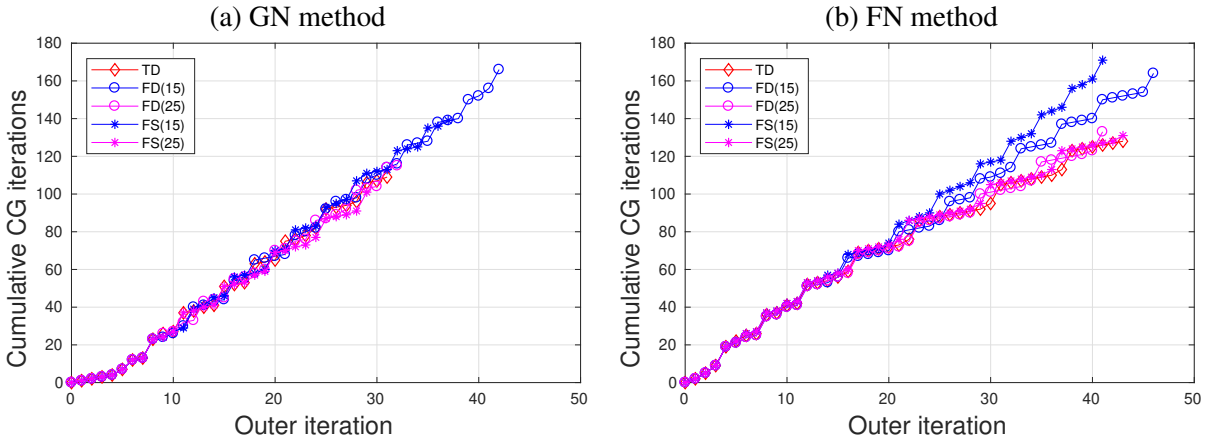


Figure 9: Comparison of cumulative CG updates via outer iteration. TD: time-domain formulation, FD: frequency-domain approximation, FS: full-scattered-field approach. The numbers in parentheses denote how many frequencies are used to build the Hessian-vector product.

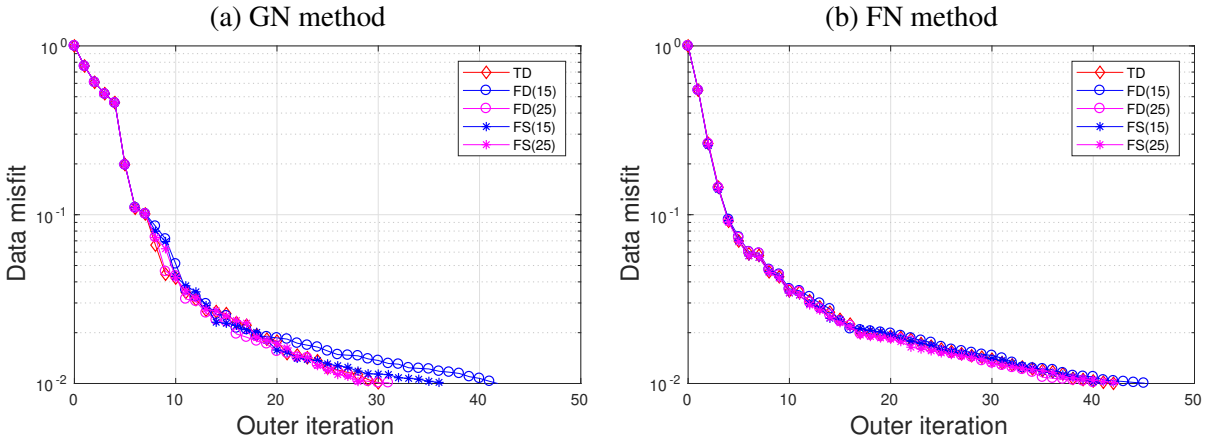


Figure 10: Comparison of convergence rate of data residual decrease with outer iteration.

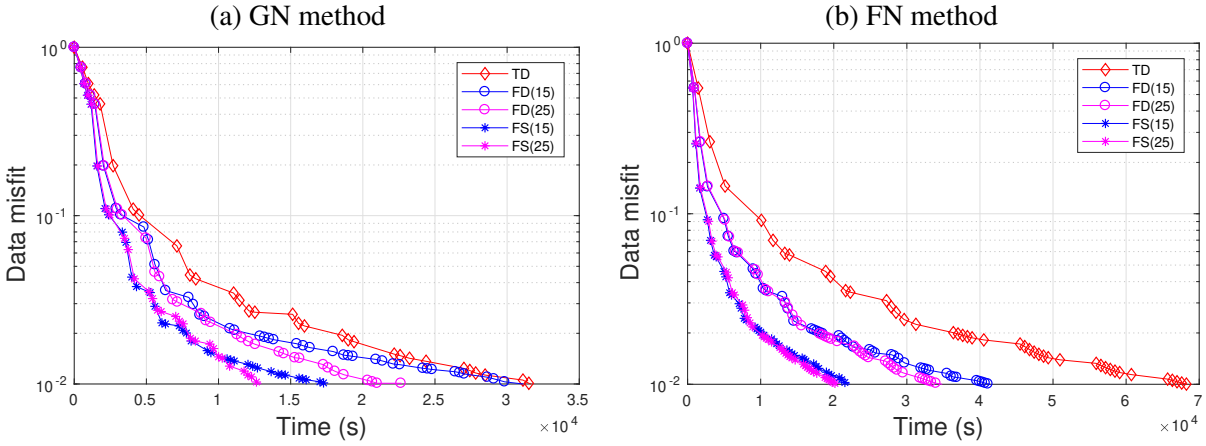
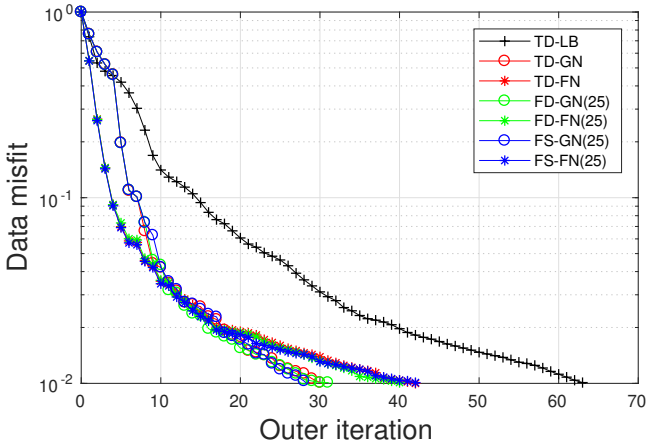
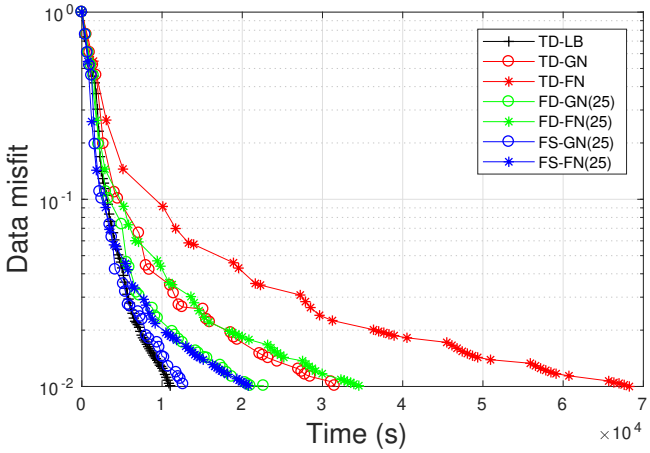


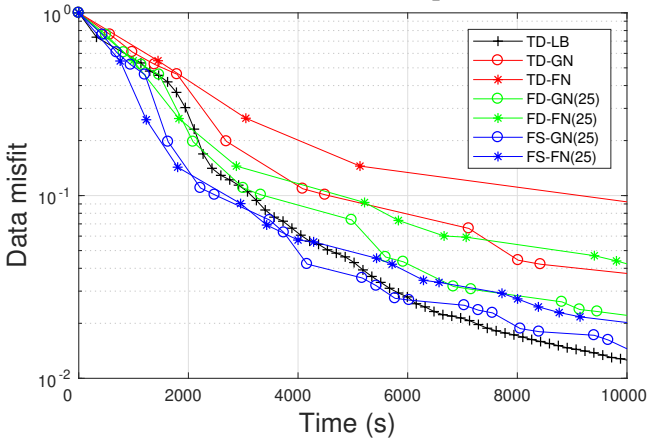
Figure 11: Comparison of convergence rate of data residual decrease with elapsed time.



(a) misfit evolution with outer iteration



(b) misfit evolution with elapsed time



(c) zoom of the first part of (b)

Figure 12: Comparison of performance of truncated Newton method and ℓ-BFGS method. Truncated Newton methods have a faster iterative convergence rate over ℓ-BFGS method. With two approximations, the proposed truncated Newton methods can provide a competitive elapsed-time convergence rate with ℓ-BFGS method.

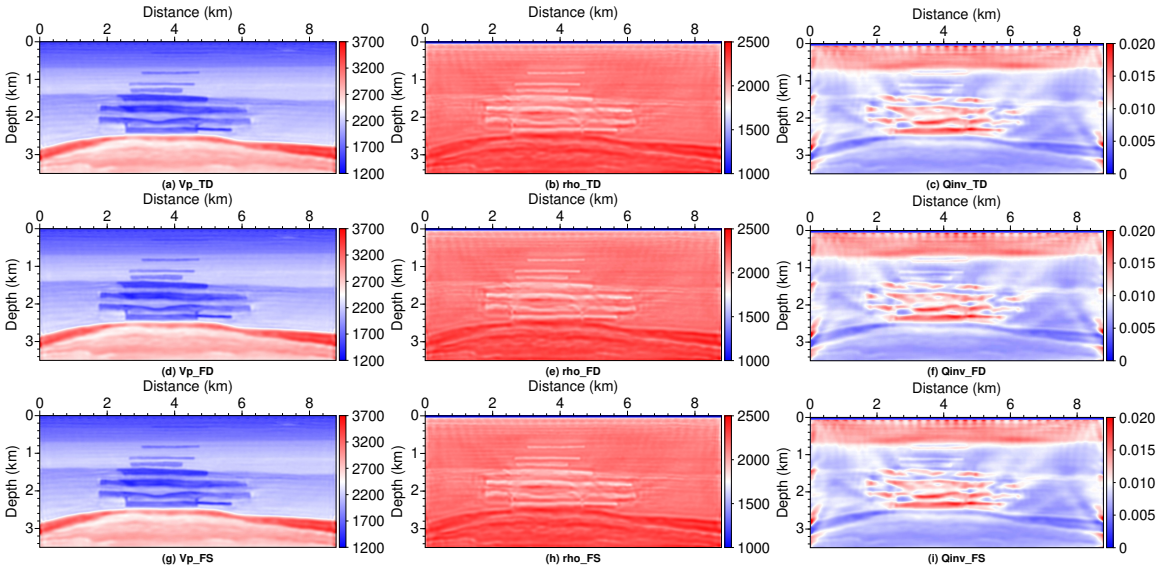


Figure 13: Final inversion results with truncated GN method: P-wave velocity (left column), density (middle column) and Q^{-1} (right column). Time-domain implementation (first row), frequency-domain approximation (second row) and full-scattered-field approach in frequency domain (third row). 25 frequencies are used in two approximate methods.

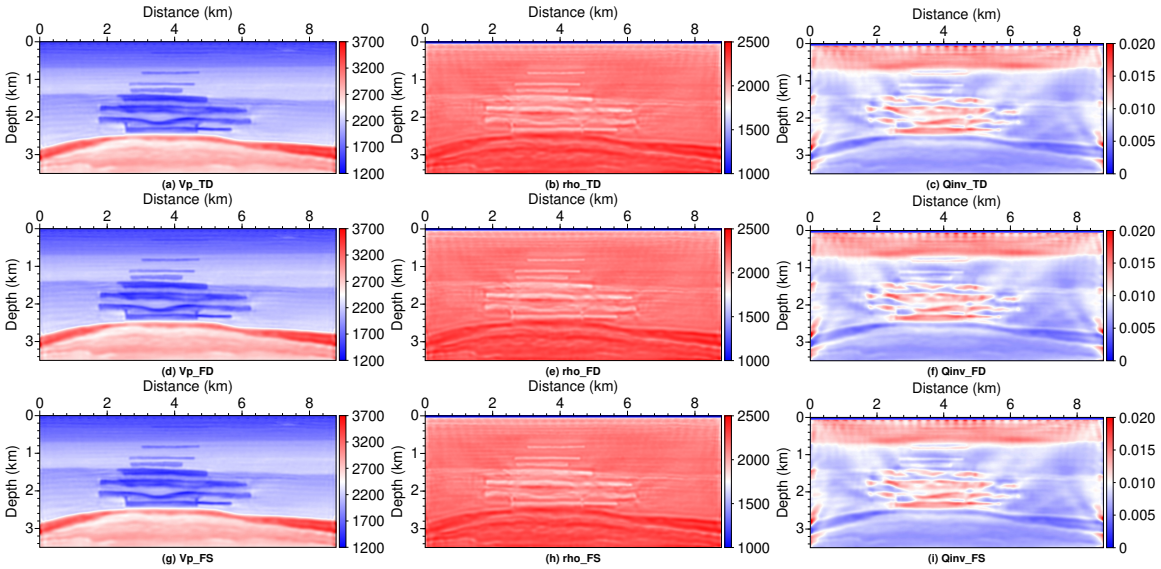


Figure 14: Final inversion results with truncated FN method: P-wave velocity (left column), density (middle column) and Q^{-1} (right column). Time-domain implementation (first row), frequency-domain approximation (second row) and full-scattered-field approach in frequency domain (third row). 25 frequencies are used in two approximate methods.

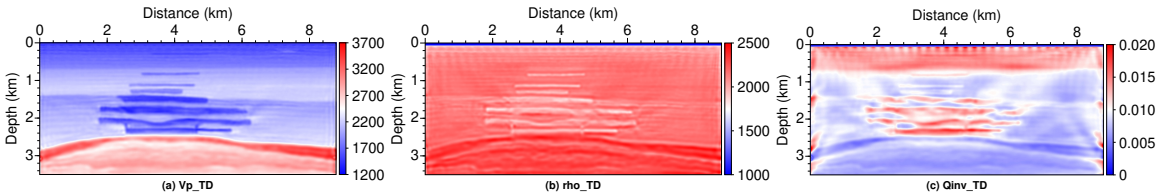


Figure 15: Final inversion results with ℓ-BFGS method:P-wave velocity (left column), density (middle column) and Q^{-1} (right column).

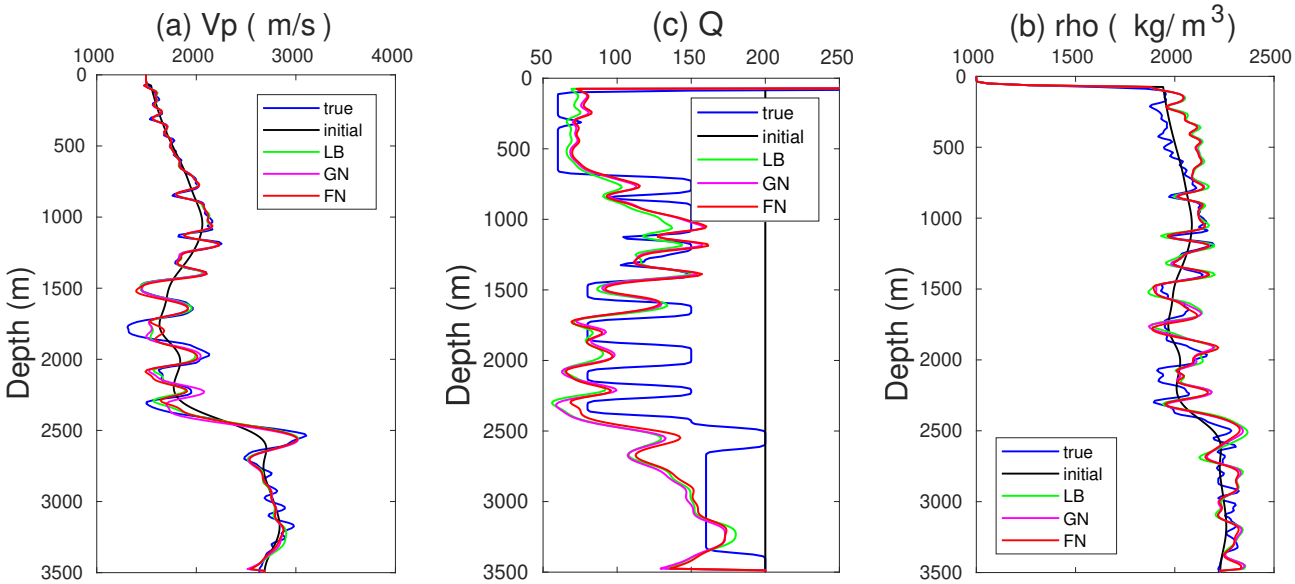


Figure 16: The extracted vertical profiles from final inverted velocity, density, and Q at the distance $x = 4 \quad km$. The results of truncated Newton method are obtained with Fourier-domain full-scattered-field approximation.

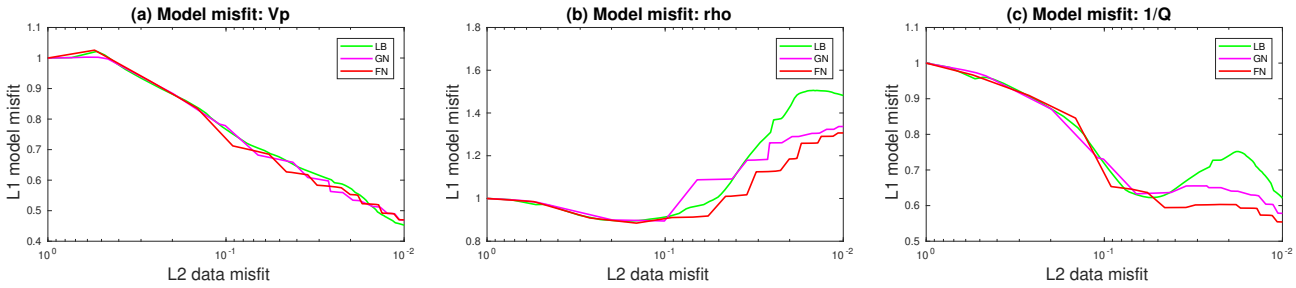


Figure 17: The normalized data misfit and the normalized model misfit for Vp (a), ρ (b), and Q^{-1} (c) using the ℓ-BFGS and truncated Newton methods. Due to strong interparameter trade-off, the monotonic decrease of data misfit can not guarantee the monotonic decrease of model misfit for each parameter. The over-fitting phenomenon occurs in ρ and Q^{-1} inversion, which can be mitigated by truncated Newton methods.

LIST OF TABLES

Table 1: Comparison of Hessian-vector product computation time in terms of wave equation solution depending on the selected approach. TD:Time-domain formulation, FD:Fourier-domain approximation, FS:full-scattered-field approximation in Fourier domain.

Method	forward modeling	backward modeling	reconstruction of incident field
GN (TD)	2	1	1
GN (FD)	2	1	0
GN (FS)	1	1	0
FN (TD)	2	2	2
FN (FD)	2	2	0
FN (FS)	1	1	0

Table 2: Normalized L_1 errors of wavefields obtained by full-scattered-field approximation.

	$R\mathbf{u}$	$\mathbf{u}$	$\boldsymbol{\mu}_1$	$\boldsymbol{\mu}$
Relative error	3.01 percent	0.99 percent	2.54 percent	3.53 percent

Table 3: Elapsed time for gradient and Hessian-vector product construction via different approaches. The numbers in parentheses denote how many frequencies used to build Hessian-vector product. The unit of the elapsed time in the table is second.

	FD (10)	FS (10)	FD (15)	FS (15)	FD (25)	FS (25)	FD (75)	FS (75)	TD
Grad (GN)	167	167	171	171	175	175	204	204	162
Grad (FN)	172	172	177	177	188	188	237	237	162
Hv (GN)	143	78	145	80	148	83	169	104	246
Hv (FN)	203	83	208	88	215	96	250	132	480

Table 4: Quantified errors of Hessian-vector product with two approximate approaches. The numbers in parentheses denote how many frequencies used to build Hessian-vector product. The unit of the values in the table is percent.

	FD (10)	FS (10)	FD (15)	FS (15)	FD (25)	FS (25)	FD (75)	FS (75)
Vp (GN)	1.54	1.59	1.23	1.23	0.83	0.86	0.70	0.72
rho (GN)	26.06	26.16	18.66	18.66	9.86	9.41	6.58	4.82
Q_{inv} (GN)	1.43	1.48	1.29	1.29	0.51	1.10	0.30	1.11
Vp (FN)	9.40	11.56	6.28	8.75	3.28	6.36	1.94	5.75
rho (FN)	34.46	34.45	23.28	23.18	10.36	9.90	4.59	3.24
Q_{inv} (FN)	7.32	13.01	5.76	12.02	4.32	11.35	3.97	11.18