Full-waveform inversion in the frequency-ray parameter domain

Wenyong Pan, Kris Innanen

ABSTRACT

Full-waveform inversion (FWI) promises high-resolution estimates of the subsurface model properties by iteratively minimizing the difference between modeled and observed data. Its computational cost remains an obstacle in practical applications, and research is active in developing efficient FWI implementations. We describe an efficient frequency-ray parameter (f-p) domain FWI equipped with linear phase-encoding in this paper. The linear phase-encoding is performed by constructing the super-gathers by summing densely distributed individual shots with linear phase shifts. A slant update strategy with varied ray parameters is proposed to further reduce the computation burden. The proposed strategies can reduce the computation burden significantly but also unfortunately introduce strong cross-talk artifacts. We demonstrate that a partial overlap-frequency strategy is important to suppress these cross-talk artifacts. The frequency-ray parameter domain FWI is implemented with gradient-based methods, quasi-Newton l-BFGS method and a truncated Gauss-Newton method. The f-p domain FWI is then enacted on a Marmousi-II model to demonstrate the effectiveness and efficiency of the combined strategies on reconstructing the velocity model. Different optimization methods with the proposed strategies are examined and compared. The resistivity to noisy data is finally analyzed and discussed.

INTRODUCTION

Full-waveform inversion (FWI) is becoming increasingly popular for building velocity models in oil and gas exploration (Lailly, 1983; Tarantola, 1984; Pratt et al., 1998; Virieux and Operto, 2009). It promises high-resolution estimates of the model parameters, but this comes at high computational cost. Aiming at reducing the computational cost for FWI, an efficient frequency-ray parameter (f-p) domain FWI with a slant update strategy is developed in this research.

Formulated as a non-linear optimization problem, FWI estimates the subsurface parameters by iteratively minimizing the difference between the modeled data and observed data. With the adjoint-state method (Plessix, 2006), the gradient of the least-squares objective function can be constructed by zero-lag cross-correlation between the forward modeled wavefield and back-propagated data residual wavefield. For modern seismic acquisitions, dense sources and receivers generate massive seismic data sets, and therefore, iterative solution of the seismic inverse problem with shot-profile methods is computationally very costly.

One popular approach for reducing the computational cost of FWI is to employ the phase-encoding technique, which involves the formation of super-gathers from summation of individual shots. Phase-encoding method was first introduced in pre-stack migration (Morton and Ober, 1998; Romero et al., 2000) but was subsequently adopted to construct the gradient and Hessian in FWI (Krebs et al., 2009; Boonyasiriwat et al., 2009; Tang, 2009; Pan et al., 2015a). Simulation of simultaneous sources can reduce computational cost significantly but at the expense of introducing cross-talk artifacts, which arise from

undesired interactions between unrelated source and receiver wavefields (Liu et al., 2006). Furthermore, a random phase-encoding is considered to be sensitive to noisy data set (Ben-Hadj-Ali et al., 2011). Linear phase-encoding strategy is performed by applying linear phase shifts at the source locations and transforming the shot gathers into plane-wave gathers with different ray parameters (Zhang et al., 2005; Liu et al., 2006; Tao and Sen, 2013). Compared to random phase-encoding, linear phase-encoding is not sensitive to random noise and shows limited amount of noise with dense acquisition arrangements (Vigh and Starr, 2008). Linear phase-encoding still produces undesired artifacts due to the simultaneous extrapolation of multiple sources. In traditional linear-phase encoding (TLPE) method, the cross-talk artifacts are mitigated by stacking a set of ray parameters at each iteration (Tao and Sen, 2013; Wu et al., 2015). Because the number of ray parameter gathers is much smaller than the number of common shot gathers in shot-profile (SP) method, the computational cost can be significantly reduced (Kwon et al., 2015). In this paper, to reduce the computational cost further, we develop a slant update (SU) strategy with linear-phase encoding, in which the gradient is constructed with single ray parameter but the ray parameter varies sequentially or randomly at each iteration. We give 2D numerical examples to show that the SU strategy can reconstruct the velocity model very well with reducing the computational cost significantly. Furthermore, the proposed strategies are expected to reduce the computation burden greatly for 3D large-scale inverse problems and they are applicable for both fixed-spread (land or ocean bottom survey) and non-fixed spread acquisitions (marine-streamer survey).

The gradient in FWI is equivalent to a reverse time migration (RTM) image constructed with a cross-correlation imaging condition. This means that the gradient is poorly-scaled and imperfectly focused as a consequence of geometrical spreading and finite-frequency effects (Pratt et al., 1998; Shin et al., 2001a; Pan et al., 2014b). The search direction can be greatly enhanced by multiplying the gradient with the inverse Hessian. In the steepest-descent (SD) method, the search direction is aligned with the negative of the gradient. In a non-linear conjugate-gradient (NCG) approach, the search direction is a linear combination of the gradient with the previous search direction (Hu et al., 2011). These gradient-based methods, by approximating the Hessian matrix as an identity matrix, experience a slow local convergence rate.

In Newton-type methods (e.g., full Newton and Gauss-Newton methods), the model update is calculated by multiplying the gradient with the inverse Hessian, which leads to a quadratic convergence rate. Unfortunately, the calculation, storage and inversion of the Hessian at each iteration greatly increases the computation burden, especially for large-scale inverse problems. In response, various Hessian approximations are developed to precondition the gradient. Shin et al. (2001a) introduced the pseudo-Hessian, in which the Fréchet derivative wavefield is replaced with the virtual source during the auto-correlation process. In Gauss-Newton methods, an approximate Hessian is adopted by neglecting the second-order term that accounts for the non-linearity of the Hessian (Pratt et al., 1998). Tang (2009) introduced a phase-encoded Hessian by constructing the Green's functions using a phase-encoding method.

The truncated-Newton method employs second-order adjoint-state formulas to compute the Hessian-vector products instead of calculating the Hessian explicitly (Métivier

et al., 2012, 2013). The search direction is obtained by solving the Newton linear system approximately with a conjugate gradient (CG) algorithm. In this research, a truncated Gauss-Newton (TGN) method is also implemented by constructing the gradient and Hessian with the linear phase-encoding. Furthermore, a phase-encoded diagonal pseudo-Hessian is employed to precondition the CG algorithm. Quasi-Newton methods do not construct the Hessian explicitly, but update the inverse Hessian approximations by storing the information from previous iterations. The BFGS method, due to Broyden (1970), Fletcher (1970), Goldfarb (1970) and Shanno (1970), is a popular quasi-Newton method to iteratively approximate the inverse Hessian. The limited-memory BFGS (l-BFGS) method stores the changes of the gradient and model from a limited number of previous iterations and uses the stored information to implicitly form an inverse of the approximated Hessian (Nocedal, 1980; Byrd et al., 1995; Nocedal and Wright, 2006). In this research, the l-BFGS optimization strategy is implemented with a "two-loop recursion" scheme (Nocedal and Wright, 2006; Anagaw, 2014; Wu et al., 2015).

FWI can be carried out in the time domain (Lailly, 1983; Tarantola, 1984; Mora, 1987) or frequency domain (Pratt et al., 1998; Sirgue and Pratt, 2004). When considering computational efficiency, the frequency domain FWI is more attractive because it involves only a set of discrete frequencies for inversion. In this research, we implement frequency domain FWI with a multiscale approach by sequentially increasing the discrete frequencies from low to high (Pratt and Chapman, 1992; Bunks et al., 1995; Sirgue and Pratt, 2004). This multiscale strategy can help to mitigate cycle-skipping problem. The most common frequency selection strategy for this process, is mono-frequency strategy, involves choosing single frequencies for inversion sequentially (Kim et al., 2011). In this research, we implement a partial overlap-frequency approach, in which multiple frequencies are used for inversion simultaneously and the frequency band expands with overlapping frequencies (Anagaw and Sacchi, 2014). This is demonstrated numerically to attenuate cross-talk noise and improve the inversion results obviously based on slant updates. Another challenge of FWI with phase-encoding is the sensitivity to noisy data (van Leeuwen et al., 2011). We add Gaussian noise to the data set for examining the noise resistance ability of the proposed strategies. We show that the f-p domain FWI with slant update strategy is robust and stable for noisy data, especially when using low frequencies for inversion.

We organize the results as follows. First, the wave equation forward modelling problem and the basic theory of FWI are reviewed. Then, we explain the traditional linear phase-encoding method and the proposed slant update strategy in this paper. We then introduce different optimization methods for FWI and describe the line search method with weak Wolfe condition. The partial overlap-frequency strategy employed in this research is then introduced. In the numerical modeling section, we first apply the l-BFGS f-p domain FWI with slant update strategy on a Marmousi-II model. The numerical examples are illustrated to examine the effectiveness and efficiency of the proposed strategy in reconstructing the velocity model. We then compare the performances of different optimizations methods with the proposed encoding strategy. Finally, we give examples to analyze the sensitivity of different encoding methods to noisy data.

THEORY AND METHODS

In this section, first, the forward modelling problem in the f-p domain and basic principle of full-waveform inversion are reviewed, and then we describe the proposed slant update strategy with linear phase-encoding. The different numerical optimization methods for FWI and the partial overlap-frequency strategy are then introduced.

Forward Modelling in the f-p domain

Frequency domain forward modelling in acoustic medium is governed by the following wave equation (Marfurt, 1984):

$$\nabla^2 \mathbf{u}(\mathbf{x}, \mathbf{x}_s, \omega) + \mathbf{m}(\mathbf{x})\omega^2 \mathbf{u}(\mathbf{x}, \mathbf{x}_s, \omega) = -f_s(\omega)\delta\left(\mathbf{x} - \mathbf{x}_s\right), \tag{1}$$

where $\mathbf{x} = (x, y, z)$ denotes the subsurface location with Cartesian coordinates, ω is the angular frequency, ∇^2 is the Laplacian operator, $\mathbf{m}\left(\mathbf{x}\right)$ is the square of slowness, $\mathbf{u}(\mathbf{x}, \mathbf{x}_s, \omega)$ denotes the pressure wavefield at position $\mathbf{x}$, $\delta\left(\mathbf{x} - \mathbf{x}_s\right)$ is the Dirac delta function, and $f_s(\omega)$ means the source signature at position $\mathbf{x}_s$. The solution of equation (1) can be written as the convolution of the source $f_s(\omega)$ with the Green's function $G\left(\mathbf{x}, \mathbf{x}_s, \omega\right)$:

$$\mathbf{u}\left(\mathbf{x}, \mathbf{x}_s, \omega\right) = f_s\left(\omega\right) G\left(\mathbf{x}, \mathbf{x}_s, \omega\right). \tag{2}$$

In this research, we use a 9-point finite difference scheme to discretize the model (Jo et al., 1996) and a Engquist-Majda boundary condition is applied on all of the boundaries of the model (Engquist and Majda, 1977). After discretization, equation (1) can be written as:

$$\mathbf{L}\left(\mathbf{x}, \omega\right) \mathbf{u}(\mathbf{x}, \mathbf{x}_s, \omega) = -f_s(\omega)\delta\left(\mathbf{x} - \mathbf{x}_s\right), \tag{3}$$

where $\mathbf{L}\left(\mathbf{x}, \omega\right)$ is the impedance matrix, which is typically sparse and symmetric.

The linear phase-encoding is performed by applying linear phase shifts (or time delays in the time domain) at densely distributed sources. The phase shift function $\gamma\left(x_s, p, \omega\right) = \omega p\left(x_s - x_0\right)$ is associated with ray parameter (or slant parameter) p and source location x_s. A common-receiver gather can be transformed into ray parameter super-gather from a line source wavefield, which is known as τ-p transform (Zhang et al., 2005):

$$\tilde{\mathbf{u}}\left(\mathbf{x}_g, p, \omega\right) = \sum_{\mathbf{x}_s} \mathbf{u}\left(\mathbf{x}_g, \mathbf{x}_s, \omega\right) \mathbf{exp}\left(i\omega p(x_s - x_0)\right), \tag{4}$$

where x_0 indicates the horizontal location of the initial source and $\mathbf{x}_g$ indicate the receiver locations. The synthetic ray parameter gather in frequency domain is generated with a line source by applying phase shifts at the source locations. Thus, the corresponding wave equation is given as:

$$\mathbf{L}\left(\mathbf{x}, \omega\right) \tilde{\mathbf{u}}(\mathbf{x}, p, \omega) = -\sum_{\mathbf{x}_s} \mathbf{exp}\left(i\omega p\left(\mathbf{x}_s - \hat{\mathbf{x}}_s\right)\right) f_s(\omega)\delta\left(\mathbf{x} - \mathbf{x}_s\right), \tag{5}$$

where when $p \geq 0$, $\hat{\mathbf{x}}_s$ indicates the location of initial source $\mathbf{x}_s^0$, if $p < 0$, $\hat{\mathbf{x}}_s$ indicates the location of the right most source $\mathbf{x}_s^{\text{end}}$. The solution of equation (5) with a line source can be written as:

$$\tilde{\mathbf{u}}(\mathbf{x}, p, \omega) = \sum_{\mathbf{x}_s} f_s(\omega) G\left(\mathbf{x}, \mathbf{x}_s, \omega\right) \mathbf{exp}\left(i\omega p\left(\mathbf{x}_s - \hat{\mathbf{x}}_s\right)\right). \tag{6}$$

In this research, the linear equations (equations (3) and (5)) are solved with a direct solver based on multi-frontal Lower Upper (LU) decomposition (Davis and Duff, 1997), which is efficient for a multi-source problem with forward and backward substitutions (Tao and Sen, 2013). For equation (5), because the number of ray parameters is generally much smaller than the number of sources, the number of backward substitutions is considerably reduced (Wu et al., 2015).

Full-waveform inversion: Review

As a non-linear least-squares optimization problem, FWI seeks to estimate the subsurface parameters through an iterative process by minimizing the difference between the synthetic data $\mathbf{u}_{\text{syn}}$ and observed data $\mathbf{u}_{\text{obs}}$ (Lailly, 1983; Tarantola, 1984; Virieux and Operto, 2009). The misfit function Φ is formulated in a least-squares form:

$$\Phi\left(\mathbf{m}\right)=\frac{1}{2}\sum_{\mathbf{x}_s}\sum_{\mathbf{x}_g}\sum_{\omega}\left\|\mathbf{u}_{\text{obs}}\left(\mathbf{x}_g,\mathbf{x}_s,\omega\right)-\mathbf{u}_{\text{syn}}\left(\mathbf{x}_g,\mathbf{x}_s,\omega\right)\right\|^2, \tag{7}$$

where $\|\cdot\|$ means the ℓ-2 norm. The Newton optimization approach is developed based on the second-order Taylor-Lagrange expansion of the misfit function Φ:

$$\Phi\left(\mathbf{m}+\Delta\mathbf{m}\right)\approx\Phi\left(\mathbf{m}\right)+\mathbf{g}^{\dagger}\Delta\mathbf{m}+\frac{1}{2}\Delta\mathbf{m}^{\dagger}\mathbf{H}\Delta\mathbf{m}, \tag{8}$$

where the symbol "†" means transpose, $\Delta\mathbf{m}$ is the search direction, $\mathbf{g}=\nabla_{\mathbf{m}}\Phi\left(\mathbf{m}\right)$ and $\mathbf{H}=\nabla_{\mathbf{m}}\nabla_{\mathbf{m}}\Phi\left(\mathbf{m}\right)$ indicate gradient and Hessian respectively.

To minimize the quadratic approximation of the misfit function, the updated model at the $(n+1)$th iteration can be written as the sum of the model at the nth iteration and the search direction $\Delta\mathbf{m}_n$:

$$\mathbf{m}_{n+1}=\mathbf{m}_n+\mu_n\Delta\mathbf{m}_n, \tag{9}$$

where μ_n is the step length, a scalar constant calculated through a line search method (Gauthier et al., 1986; Pica et al., 1990; Nocedal and Wright, 2006). Within a Newton optimization framework, the search direction $\Delta\mathbf{m}_n$ is the solution of the Newton linear system:

$$\mathbf{H}_n\Delta\mathbf{m}_n=-\mathbf{g}_n. \tag{10}$$

The gradient is the first-order partial derivative of the misfit function with respect to the model parameter and it indicates the direction in which the misfit function is increasing most rapidly (Pratt et al., 1998). It can be constructed by zero-lag correlation between the Fréchet derivative wavefields with complex conjugate of the data residuals $\Delta\mathbf{d}$:

$$\mathbf{g}\left(\mathbf{x}\right)=\sum_{\mathbf{x}_s}\sum_{\mathbf{x}_g}\sum_{\omega}\Re\left(\frac{\partial\mathbf{u}_{\text{syn}}^{\dagger}\left(\mathbf{x}_g,\mathbf{x}_s,\omega\right)}{\partial\mathbf{m}\left(\mathbf{x}\right)}\Delta\mathbf{d}^{*}\left(\mathbf{x}_g,\mathbf{x}_s,\omega\right)\right), \tag{11}$$

where the symbol "$*$" means complex conjugate and $\Re\left(\cdot\right)$ denotes the real part. The gradient can be efficiently constructed by applying a zero-lag cross-correlation between the forward modeled wavefield and back-propagated data residuals wavefield, using the adjoint-state technique (Plessix, 2006). With the adjoint formalism, the FWI gradient is given by

(Sirgue and Pratt, 2004; Plessix and Mulder, 2004):

$$\mathbf{g}\left(\mathbf{x}\right)=\sum_{\mathbf{x}_s}\sum_{\mathbf{x}_g}\sum_{\omega}\Re\left(\omega^2 f_s(\omega)G(\mathbf{x},\mathbf{x}_s,\omega)G^*(\mathbf{x},\mathbf{x}_g,\omega)\Delta\mathbf{d}^*\left(\mathbf{x}_g,\mathbf{x}_s,\omega\right)\right), \tag{12}$$

where $G(\mathbf{x},\mathbf{x}_s,\omega)$ and $G(\mathbf{x},\mathbf{x}_g,\omega)$ indicate the source-side and receiver-side Green's functions in the reference medium respectively (Stolt and Benson, 1986; Sirgue and Pratt, 2004). The gradient is poorly-scaled due to geometrical spreading, and it is also contaminated by spurious correlations because of finite-frequency effects. The Hessian operator is the second-order partial derivative of the misfit function with respect to the model parameter:

$$\begin{aligned}\mathbf{H}\left(\mathbf{x},\mathbf{x}'\right)&=\nabla_{\mathbf{m}(\mathbf{x})}\nabla_{\mathbf{m}(\mathbf{x}')}\Phi\left(\mathbf{m}\right)\\&=\sum_{\mathbf{x}_g}\sum_{\mathbf{x}_s}\sum_{\omega}\Re\left(\frac{\partial\mathbf{u}^{\dagger}_{\text{syn}}\left(\mathbf{x}_g,\mathbf{x}_s,\omega\right)}{\partial\mathbf{m}\left(\mathbf{x}\right)}\frac{\partial\mathbf{u}^{*}_{\text{syn}}\left(\mathbf{x}_g,\mathbf{x}_s,\omega\right)}{\partial\mathbf{m}\left(\mathbf{x}'\right)}+\frac{\partial^2\mathbf{u}^{\dagger}_{\text{syn}}\left(\mathbf{x}_g,\mathbf{x}_s,\omega\right)}{\partial\mathbf{m}\left(\mathbf{x}\right)\partial\mathbf{m}\left(\mathbf{x}'\right)}\Delta\mathbf{d}^*\left(\mathbf{x}_g,\mathbf{x}_s,\omega\right)\right),\end{aligned} \tag{13}$$

where $\mathbf{H}\left(\mathbf{x},\mathbf{x}'\right)$ denotes one element in Hessian $\mathbf{H}$ corresponding to two model perturbations at $\mathbf{x}$ and $\mathbf{x}'$. Preconditioning the gradient with the Hessian can greatly enhance the model update, which provides a quadratic convergence rate.

A slant update strategy with linear phase-encoding

The simultaneous source method was first proposed in pre-stack depth migration for addressing the obstacle of high computational cost (Morton and Ober, 1998; Romero et al., 2000; Stoffa et al., 2006). This technique was then applied in FWI for gradient and Hessian calculation (Vigh and Starr, 2008; Krebs et al., 2009; Ben-Hadj-Ali et al., 2011; Tao and Sen, 2013; Castellanos et al., 2015; Pan et al., 2015a). The linear phase-encoding strategy is performed by decomposing the densely seismic data into plane-wave domain and choosing a set of ray parameters for modeling, migration and inversion. Liu et al. (2006) proved the equivalence between the shot-profile migration image and plane-wave migration image and showed that the cross-talk artifacts, arising from the interactions between unrelated source and receiver wavefields, can be suppressed reasonably well by stacking sufficient ray parameters.

The ray parameter is controlled by the take-off angle and top surface velocity (Zhang et al., 2005):

$$p=\frac{\sin\theta}{c}, \tag{14}$$

where θ is the take-off angle and c indicates the surface velocity. Different ray parameters account for illuminating subsurface layers with different dip angles and the pre-knowledge of the geological structures in the target area can be used to determine the ray parameter range (Wang et al., 2006; Vigh and Starr, 2008; Pan et al., 2015a). Following equation (4), the data residual vector $\Delta\tilde{\mathbf{d}}\left(\mathbf{x}_g,p,\omega\right)$ with ray parameter p can be obtained by τ-p transform:

$$\Delta\tilde{\mathbf{d}}\left(\mathbf{x}_g,p,\omega\right)=\sum_{\mathbf{x}'_s}\Delta\mathbf{d}\left(\mathbf{x}_g,\mathbf{x}'_s,\omega\right)\mathbf{exp}\left(i\omega p(\mathbf{x}'_s-\hat{\mathbf{x}}_s)\right). \tag{15}$$

The adjoint data residual wavefield can be obtained by convolving the data residual vector $\Delta\tilde{\mathbf{d}}\left(\mathbf{x}_g, p, \omega\right)$ with the receiver-side Green's function $G\left(\mathbf{x}, \mathbf{x}_g, \omega\right)$:

$$\tilde{\mathbf{v}}^*\left(\mathbf{x}_g, p, \omega\right) = \sum_{\mathbf{x}'_s} \Delta\mathbf{d}^*\left(\mathbf{x}_g, \mathbf{x}'_s, \omega\right) \mathbf{exp}\left(i\omega p(\hat{\mathbf{x}}_s - \mathbf{x}'_s)\right) G^*\left(\mathbf{x}, \mathbf{x}_g, \omega\right). \tag{16}$$

Cross-correlating the forward modeled wavefield $\tilde{\mathbf{u}}\left(\mathbf{x}, p, \omega\right)$ (equation (6)) with the back-propagated wavefield $\tilde{\mathbf{v}}^*\left(\mathbf{x}_g, p, \omega\right)$ (equation (16)) gives the slant gradient with ray parameter p:

$$\begin{aligned}\tilde{\mathbf{g}}(\mathbf{x}, p) = \sum_{\mathbf{x}_s}\sum_{\mathbf{x}_g}\sum_{\mathbf{x}'_s}\sum_{\omega} \Re\Big(&\omega^2 f_s(\omega) G\left(\mathbf{x}, \mathbf{x}_s, \omega\right) G^*\left(\mathbf{x}, \mathbf{x}_g, \omega\right) \Delta\mathbf{d}^*\left(\mathbf{x}_g, \mathbf{x}'_s, \omega\right) \\ &\times A^2\left(\omega\right) \mathbf{exp}\left(i\omega p(\mathbf{x}_s - \mathbf{x}'_s)\right)\Big),\end{aligned} \tag{17}$$

where $A\left(\omega\right)$ is a weighting function depending on angular frequency ω (Liu et al., 2006; Tang, 2009). When $\mathbf{x}_s = \mathbf{x}'_s$, the slant gradient $\tilde{\mathbf{g}}(\mathbf{x}, p)$ is equal to shot-profile gradient $\mathbf{g}(\mathbf{x})$ (equation (11)). When $\mathbf{x}_s \neq \mathbf{x}'_s$, the slant gradient $\tilde{\mathbf{g}}(\mathbf{x}, p)$ becomes the cross-talk artifacts term $\mathbf{g}_{\text{cross}}$. Thus, the slant gradient is the summation of shot-profile gradient with the cross-talk term. These cross-talk artifacts can be suppressed reasonably by stacking a group of ray parameters:

$$\begin{aligned}\tilde{\mathbf{g}}(\mathbf{x}) = \sum_{\mathbf{x}_s}\sum_{\mathbf{x}_g}\sum_{\mathbf{x}'_s}\sum_{\mathbf{p}}\sum_{\omega} \Re\Big(&\omega^2 f_s(\omega) G\left(\mathbf{x}, \mathbf{x}_s, \omega\right) G^*\left(\mathbf{x}, \mathbf{x}_g, \omega\right) \Delta\mathbf{d}^*\left(\mathbf{x}_g, \mathbf{x}'_s, \omega\right) \\ &\times A^2\left(\omega\right) \mathbf{exp}\left(i\omega\mathbf{p}\cdot(\mathbf{x}_s - \mathbf{x}'_s)\right)\Big),\end{aligned} \tag{18}$$

where $\mathbf{p}$ is the ray parameter vector and "$\cdot$" means inner product. Equation (18) is named as phase-encoded gradient in this paper. Extracting the encoding function $\psi\left(\mathbf{x}_s, \mathbf{x}'_s, \omega\right)$ from equation (18):

$$\psi\left(\mathbf{x}_s, \mathbf{x}'_s, \omega\right) = \sum_{\mathbf{p}} A^2\left(\omega\right) \mathbf{exp}\left(i\omega\mathbf{p}\cdot(\mathbf{x}_s - \mathbf{x}'_s)\right). \tag{19}$$

If the number of ray parameters in $\mathbf{p}$ is large enough, integrating over ray parameters yields a Dirac delta function in space (Liu et al., 2006; Tang, 2009):

$$\psi\left(\mathbf{x}_s, \mathbf{x}'_s, \omega\right) = \frac{A^2\left(\omega\right)}{\omega^2}\delta\left(\mathbf{x}_s - \mathbf{x}'_s\right). \tag{20}$$

Inserting equation (20) into equation (18) gives:

$$\tilde{\mathbf{g}}(\mathbf{x}) = \sum_{\mathbf{x}_s}\sum_{\mathbf{x}_g}\sum_{\mathbf{x}'_s}\sum_{\omega} \Re\left(\omega^2 f_s(\omega) G\left(\mathbf{x}, \mathbf{x}_s, \omega\right) G^*\left(\mathbf{x}, \mathbf{x}_g, \omega\right) \Delta\mathbf{d}^*\left(\mathbf{x}_g, \mathbf{x}'_s, \omega\right) \frac{A^2\left(\omega\right)}{\omega^2}\delta(\mathbf{x}_s - \mathbf{x}'_s)\right). \tag{21}$$

Considering the sifting property of the delta function and making $A^2\left(\omega\right) = \omega^2$, the phase-encoded gradient becomes the shot-profile gradient (equation (11)) and the cross-talk artifacts are suppressed completely:

$$\tilde{\mathbf{g}}(\mathbf{x}) = \mathbf{g}(\mathbf{x}) = \sum_{\mathbf{x}_s}\sum_{\mathbf{x}_g}\sum_{\omega} \Re\Big(\omega^2 f_s(\omega) G\left(\mathbf{x}, \mathbf{x}_s, \omega\right) G^*\left(\mathbf{x}, \mathbf{x}_g, \omega\right) \Delta\mathbf{d}^*\left(\mathbf{x}_g, \mathbf{x}_s, \omega\right)\Big). \tag{22}$$

To mitigate the cross-talk artifacts with linear phase-encoding method, stacking a set of ray parameters at each iteration with ray parameter spacing Δp is usually performed (Tao and Sen, 2013; Kwon et al., 2015). Zhang et al. (2005) gave the suggestion for determining the number of ray parameters:

$$N_p \geq \frac{2N_s \Delta x_s f_{\max} \sin\theta_{\max}}{c}, \tag{23}$$

where $f_{\max}$ is the maximum frequency, N_s and Δx_s are the number of sources and source spacing and $\theta_{\max}$ indicates the maximum take-off angle, which can be obtained with maximum ray parameter $p_{\max}$: $\theta_{\max} = \sin^{-1}(p_{\max} c)$ following equation (14). Here, we define this strategy as traditional linear phase-encoding (TLPE) method, in which number of $2 \times N_p \times N_f$ forward modelling problems need to be solved at each iteration. N_f indicates the number of frequencies used for inversion at each iteration. In shot-profile (SP) method, number of $2 \times N_s \times N_f$ forward modelling problems are required for gradient calculation, which is proportional to the number of sources N_s. Because the number of ray parameters N_p is far smaller than the number of sources N_s, the computation burden at each iteration in TLPE method is greatly reduced compared to SP method. In this research, for reducing the computation burden further, we develop a slant update (SU) strategy, in which the slant gradient with single ray parameter (equation (17)) is used to update the model instead of the phase-encoded gradient with stacking ray parameters (equation (18)). Furthermore, the ray parameter needs to be changed sequentially or randomly iteration by iteration. Thus, number of N_f forward modelling problems are required for gradient construction at each non-linear iteration.

Gauss-Newton and full Newton methods

The Newton-type optimization methods (e.g., Gauss-Newton (GN) and full Newton (FN) methods) use the quadratic search direction and converge fast for inverting a limited number of unknown parameters. In exact GN method, the search direction is calculated by multiplying the gradient with an inverse approximate Hessian:

$$\Delta \mathbf{m}_n = -\left(\mathbf{H}_a^n\right)^{-1} \mathbf{g}_n, \tag{24}$$

where the approximate Hessian $\mathbf{H}_a$ is formed by correlating two Fréchet derivative wavefields, which accounts for the first-order scattering effects, as indicated by the first term of equation (13):

$$\mathbf{H}_a(\mathbf{x}, \mathbf{x}') = \sum_{\mathbf{x}_s} \sum_{\mathbf{x}_g} \sum_{\omega} \Re \left(\frac{\partial \mathbf{u}^{\dagger}_{\text{syn}}(\mathbf{x}_g, \mathbf{x}_s, \omega)}{\partial \mathbf{m}(\mathbf{x})} \frac{\partial \mathbf{u}^{*}_{\text{syn}}(\mathbf{x}_g, \mathbf{x}_s, \omega)}{\partial \mathbf{m}(\mathbf{x}')} \right), \tag{25}$$

where $\mathbf{x}'$ can be considered as the neighboring point around the imaging point $\mathbf{x}$ in subsurface (Valenciano, 2008). Because of band-limited seismic data, the two Fréchet derivative wavefields are often to a great degree uncorrelated, meaning that the approximate Hessian is diagonally dominant and banded (Pratt et al., 1998; Valenciano, 2008; Tang, 2009; Pan et al., 2014a). Thus, the approximate Hessian in Gauss-Newton method can de-blur the gradient and remove the finite-frequency effects (Pratt et al., 1998). Furthermore, the doubly scattered energy in the data residuals can result in second-order scattering artifacts in

the gradient. The second term in equation (13) predicts the second-order non-linear effects by correlating the second-order partial derivative wavefields with the data residuals. With this second-order preconditioner, the second-order scattering artifacts in the gradient can be suppressed effectively (Pratt et al., 1998; Pan et al., 2015b). Multiplying the gradient with the inverse of the full Hessian $\mathbf{H}$ (equation (13)) forms one full Newton step. For these Newton-type methods, explicit calculating and inversing the Hessian at each iteration are required. Considering the subsurface model with number of N_m model parameters, the Hessian is a $N_m \times N_m$ square and symmetric matrix. Even though, the Newton-type methods benefit from fast convergence rates, the computation, storage and inversion of the second derivative Hessian at each iteration are prohibitively expensive, which limits their applications for large-scale inverse problems in exploration geophysics.

Gradient-based methods

The gradient-based methods (e.g., steepest-descent (SD) and non-linear conjugate-gradient (NCG) methods) assume the Hessian matrix $\mathbf{H}$ as an identity matrix $\mathbf{I}$ and they are more attractive than the Newton-type ones when inversing a large number of unknown model parameters. The steepest-descent method simply determines the search direction to be the negative of the gradient: $\Delta\mathbf{m}_n = -\mathbf{g}_n$. In mathematics, conjugate-gradient method seeks the solution of a linear system. The non-linear conjugate-gradient method generalizes the conjugate-gradient method to non-linear optimization and obtain the local minimum of a non-linear function using its gradient alone (Hu et al., 2011). The search direction is just a linear combination of current gradient and previous search direction, which can be expressed as:

$$\Delta\mathbf{m}_n = -\mathbf{g}_n + \beta_n \Delta\mathbf{m}_{n-1}, \tag{26}$$

where β_n is a scalar, which makes $\Delta\mathbf{m}_n$ and $\Delta\mathbf{m}_{n-1}$ conjugate. There are a lot of approaches for determining parameter β_n. In this research, the Fletcher-Reeves method (Fletcher and Reeves, 1964) is used to obtain the parameter β_n:

$$\beta_n = \frac{\mathbf{g}_n^\dagger \mathbf{g}_n}{\mathbf{g}_{n-1}^\dagger \mathbf{g}_{n-1}}. \tag{27}$$

The truncated-Newton methods

The truncated-Newton methods (Hessian-free optimization methods or inexact-Newton methods), instead of constructing the Hessian approximations or inverse Hessian approximations, solve the Newton linear system (equation (10)) approximately with a conjugate-gradient (CG) method (Saad, 2003; Métivier et al., 2014). The CG method is an optimal algorithm for solving a symmetric and positive definite system and it only requires computing the Hessian-vector products $\mathbf{H}v$ instead of forming the Hessian matrix explicitly, where v is an arbitrary vector in model space. In truncated-Newton methods, the full Hessian (equation (13)) is always replaced with Gauss-Newton Hessian (equation (25)), which makes sure that the linear system is always positive definite (Nash, 2000):

$$\left(\mathbf{H}_a^n + \varepsilon\mathcal{A}\right)\Delta\mathbf{m}_n = -\mathbf{g}_n, \tag{28}$$

where $\epsilon\mathcal{A}$ is the damping term, ε is a small constant value and $\mathcal{A}$ indicates the maximum value of the Gauss-Newton Hessian. Thus, the truncated Gauss-Newton (TGN) method

is implemented in a double-iterative scheme: the outer loop is iteratively updating the model parameters for the non-linear optimization problem, and the inner loop is solving the linear system (equation(28)) iteratively with the CG algorithm. The inner iteration is typically stopped or "truncated" before the solution of the Newton equation is obtained. The truncated Gauss-Newton method can be accelerated by preconditioning the CG algorithm (PCG method) in the inner iteration:

$$\mathcal{M}^{-1}\left(\mathbf{H}_a^n + \varepsilon\mathcal{A}\right)\Delta\mathbf{m}_n = -\mathcal{M}^{-1}\mathbf{g}_n, \tag{29}$$

where the preconditioner $\mathcal{M}$ is devised as Hessian approximation or $\mathcal{M}^{-1}$ is the inverse Hessian approximation.

In this research, to implement the truncated Gauss-Newton (TGN) method, the gradient is obtained by the phase-encoding method with slant update strategy. The Gauss-Newton Hessian-vector product is calculated using second-order adjoint-state method with linear phase-encoding (Métivier et al., 2014), which forms the phase-encoded Gauss-Newton Hessian $\tilde{\mathbf{H}}_a$ (equation (44)). Furthermore, a phase-encoded diagonal pseudo-Hessian $\tilde{\mathbb{H}}_{\text{diag}}$ (equation (55)) is employed as the preconditioner. The search direction is obtained by solving the following linear system:

$$\left(\tilde{\mathbb{H}}_{\text{diag}}^n + \lambda\mathcal{B}\right)^{-1}\left(\tilde{\mathbf{H}}_a^n + \varepsilon\mathcal{A}\right)\Delta\mathbf{m}_n = -\left(\tilde{\mathbb{H}}_{\text{diag}}^n + \lambda\mathcal{B}\right)^{-1}\tilde{\mathbf{g}}_n, \tag{30}$$

where the phase-encoded Gauss-Newton Hessian $\tilde{\mathbf{H}}_a$ and phase-encoded diagonal pseudo-Hessian $\tilde{\mathbb{H}}_{\text{diag}}$ are derived in Appendix A and Appendix B respectively. $\lambda\mathcal{B}$ is the stabilization term for the preconditioner, where λ is a small constant value and $\mathcal{B}$ indicates the maximum value of the diagonal pseudo-Hessian. The Gauss-Newton Hessian-vector products construction with linear phase-encoding is equal to

The inner CG algorithm should be terminated with an appropriate stopping criteria (Nash, 2000). We define the maximum inner iteration $\tilde{n}_{max}$ and relative residual $\gamma_{\tilde{n}}$:

$$\gamma_{\tilde{n}} = \frac{\|\tilde{\mathbf{H}}_a^{\tilde{n}}\Delta\mathbf{m}_{\tilde{n}} + \tilde{\mathbf{g}}_{\tilde{n}}\|}{\|\tilde{\mathbf{g}}_{\tilde{n}}\|}, \tag{31}$$

where $\tilde{n}$ indicates the CG inner iteration index. The inner iteration is stopped when $\gamma_{\tilde{n}} < \gamma_{min}$, where γ_{min} indicates the relative residual tolerance.

Quasi-Newton methods

The quasi-Newton methods provide an attractive alternative to Newton-type and gradient-based methods by approximating the inversion Hessian iteratively instead of constructing the Hessian matrix explicitly (Brossier et al., 2009; Ma and Hale, 2012). BFGS method, named after Broyden (1970), Fletcher (1970), Goldfarb (1970) and Shanno (1970), is one popular quasi-Newton strategy to approximate the inverse Hessian iteratively using the changes of the model and gradient (Nocedal and Wright, 2006). The model update with BFGS formula can be written as:

$$\Delta\mathbf{m}_n = -\mathcal{H}_n\mathbf{g}_n, \tag{32}$$

where $\mathcal{H}_n$ indicates the inverse Hessian:

$$\mathcal{H}_{n+1} = \left(\mathbf{I} - \frac{\mathbf{s}_n \mathbf{y}_n^\dagger}{\mathbf{y}_n^\dagger \mathbf{s}_n}\right) \mathcal{H}_n \left(\mathbf{I} - \frac{\mathbf{y}_n \mathbf{s}_n^\dagger}{\mathbf{y}_n^\dagger \mathbf{s}_n}\right) + \frac{\mathbf{s}_n \mathbf{s}_n^\dagger}{\mathbf{y}_n^\dagger \mathbf{s}_n}, \tag{33}$$

where $\mathbf{s}_n = \mathbf{m}_{n+1} - \mathbf{m}_n$ and $\mathbf{y}_n = \mathbf{g}_{n+1} - \mathbf{g}_n$ indicate the changes of the model and gradient respectively. The initial approximation $\mathcal{H}_0$ is important to BFGS method and it is usually set as an identity matrix to make sure that the updated matrix maintains positive definiteness (Wu et al., 2015).

Even though, BFGS method can reduce the computational cost to approximate the Hessian, the memory requirement for storage and computational cost for preconditioning remain to be challenging, especially for large-scale optimization problem. To overcome this difficulty, a limited-memory BFGS (l-BFGS) method is developed by storing the model and gradient changes from a limited number M of previous iterations (typically $M < 10$) (Nocedal, 1980). The stored information is then used to construct an approximated inverse Hessian. A "two-loop recursion" scheme is implemented in this research to obtain the search direction using the information of previous updates (Nocedal and Wright, 2006). The pseudo-code of this "two-loop recursion" scheme is illustrated in Table 1.

The inverse Hessian approximation employed in l-BFGS method is based on the gradient and model changes from previous iterations. In TLPE and SU methods, the shot-profile gradient is replaced by the phase-encoded gradient and slant gradient, which impacts the inverse Hessian approximations. The inverse Hessian approximations (equation (33)) $\tilde{\mathcal{H}}(\mathbf{p})$ and $\tilde{\mathcal{H}}(p_{n+1}, p_n)$ in TLPE and SU methods can be expressed as:

$$\tilde{\mathcal{H}}_{n+1}(\mathbf{p}) = \left(\mathbf{I} - \frac{\mathbf{s}_n \tilde{\mathbf{y}}_n^\dagger(\mathbf{p})}{\tilde{\mathbf{y}}_n^\dagger(\mathbf{p})\,\mathbf{s}_n}\right) \tilde{\mathcal{H}}_n(\mathbf{p}) \left(\mathbf{I} - \frac{\tilde{\mathbf{y}}_n(\mathbf{p})\,\mathbf{s}_n^\dagger}{\tilde{\mathbf{y}}_n^\dagger(\mathbf{p})\,\mathbf{s}_n}\right) + \frac{\mathbf{s}_n \mathbf{s}_n^\dagger}{\tilde{\mathbf{y}}_n^\dagger(\mathbf{p})\,\mathbf{s}_n}, \tag{34}$$

$$\tilde{\mathcal{H}}_{n+1}(p_{n+1}, p_n) = \left(\mathbf{I} - \frac{\mathbf{s}_n \tilde{\mathbf{y}}_n^\dagger(p_{n+1}, p_n)}{\tilde{\mathbf{y}}_n^\dagger(p_{n+1}, p_n)\,\mathbf{s}_n}\right) \tilde{\mathcal{H}}_n(p_{n+1}, p_n) \left(\mathbf{I} - \frac{\tilde{\mathbf{y}}_n(p_{n+1}, p_n)\,\mathbf{s}_n^\dagger}{\tilde{\mathbf{y}}_n^\dagger(p_{n+1}, p_n)\,\mathbf{s}_n}\right) + \frac{\mathbf{s}_n \mathbf{s}_n^\dagger}{\tilde{\mathbf{y}}_n^\dagger(p_{n+1}, p_n)\,\mathbf{s}_n}, \tag{35}$$

where $\tilde{\mathbf{y}}_n(\mathbf{p}) = \tilde{\mathbf{g}}_{n+1}(\mathbf{p}) - \tilde{\mathbf{g}}_n(\mathbf{p})$ and $\tilde{\mathbf{y}}_n(p_{n+1}, p_n) = \tilde{\mathbf{g}}_{n+1}(p_{n+1}) - \tilde{\mathbf{g}}_n(p_n)$. The convergence rate of the l-BFGS optimization method will be impacted using the changes of the phase-encoded gradient and slant gradient to approximate the inverse Hessian. Similarly, the line search method for calculating step length will also be influenced with phase-encoding strategy. Studying the influences of phase-encoding method on the convergence rate of the optimization method is beyond the scope of this research.

Line search with weak Wolfe condition

In this paper, we use a line search method to get the step length μ_n (equation (9)) satisfying the Wolfe condition, which can be measured by the following inequalities (Nocedal and Wright, 2006):

$$\Phi(\mathbf{m}_n + \mu_n \Delta\mathbf{m}_n) \leq \Phi(\mathbf{m}_n) + \mu_n c_1 \Delta\mathbf{m}_n^\dagger \nabla\Phi(\mathbf{m}_n), \tag{36}$$

$$\nabla\Phi(\mathbf{m}_n + \mu_n \Delta\mathbf{m}_n) \geq c_2 \nabla\Phi(\mathbf{m}_n), \tag{37}$$

Table 1 Two-loop recursion scheme for l-BFGS FWI.

1. Given initial inverse Hessian approximation $\mathcal{H}_0 = \mathbf{I}$;
2. $\Delta\mathbf{m} = \mathbf{g}$;
3. **For** $i = n-1, n-2, ..., n-M$
4. $\mathbf{x}_i$=$\mathbf{s}_i^\dagger \Delta\mathbf{m} / \mathbf{y}_i^\dagger \mathbf{s}_i$;
5. $\Delta\mathbf{m}$=$\Delta\mathbf{m}$-$\mathbf{x}_i\mathbf{y}_i$;
6. $\mathbf{v} = \mathcal{H}_0 \Delta\mathbf{m}$;
7. **End**
8. **For** $i = n-M, n-M+1, ..., n-1$
9. $\mathbf{z}$=$\mathbf{y}_i^\dagger \mathbf{x} / \mathbf{y}_i^\dagger \mathbf{s}_i$;
10. $\mathbf{v}$=$\mathbf{v} - \mathbf{s}_i(\mathbf{x}_i - \mathbf{z})$;
11. **End**
12. $\mathcal{H}_n \mathbf{g}_n = \mathbf{v}$;
13. $\Delta\mathbf{m}_n$=$\mathbf{g}_n$.

Table 2 Frequency selection strategies.

Strategies	Frequency Group (FG)	Iteration Times Per FG
Mono-frequency	$\underbrace{f_1}_{\text{FG}_1}, \underbrace{f_2}_{\text{FG}_2}, \ldots, \underbrace{f_{\max}}_{\text{FG}_{\max}}$	N_{FG}
Partial overlap-frequency	$\underbrace{f_1, f_2, f_3}_{\text{FG}_1}, \underbrace{f_2, f_3, f_4}_{\text{FG}_2}, \ldots, \underbrace{f_{\max-2}, f_{\max-1}, f_{\max}}_{\text{FG}_{\max}}$	N_{FG}

where the constant parameters c_1 and c_2 satisfy $0 < c_1 < c_2 < 1$. In practice, parameter c_1 can be chosen as $c_1 = 10^{-4}$ and parameter c_2 should be much larger $c_2 = 0.9$ (Nocedal and Wright, 2006). Equation (36) is the Armijo condition, which ensures that the step length μ_n decreases the misfit function sufficiently (Armijo, 1966). Equation (37) is known as the weak Wolfe condition, which ensures that the slope has been reduced sufficiently (Nocedal and Wright, 2006). The initial step length μ_0 is always chosen as 1 and then after a set of trial step lengths, the optimal one will be accepted to satisfy the above conditions.

Frequency selection strategy

The common frequency selection strategy in frequency domain FWI is the sequential inversion of single frequencies (Sirgue and Pratt, 2004), which is named as mono-frequency strategy in this research. This mono-frequency selection strategy is performed by increasing the single frequencies from low to high sequentially without overlapping the frequencies, as shown in Table 2. Each frequency group only contains one single frequency and for each single frequency, number of N_{FG} iterations are performed. We employ a partial overlap-frequency strategy, in which number of $N_f = 3$ frequencies are used for inversion simultaneously. The frequency group increases from low to high with 2 frequencies overlapped and for each frequency band, number of N_{FG} iterations are performed. In this research, we demonstrate that inversing multiple frequencies simultaneously is important to reduce the cross-talk artifacts for the slant update strategy.

NUMERICAL EXPERIMENTS

In this section, we first apply the l-BFGS f-p domain FWI with slant update strategy on a Marmousi-II model and verify the effectiveness and efficiency of proposed strategy, compared to shot-profile (SP) and traditional linear-phase encoding (TLPE) methods. Then, we illustrate the inversion results by different optimization methods with the slant update strategy and demonstrate the computation efficiency of l-BFGS method. Finally, we analyze the sensitivity of different encoding methods to the noisy data.

Examining the f-p domain FWI with slant update strategy

In this numerical section, the l-BFGS f-p domain FWI with the slant update strategy is applied on a Marmousi-II model. We examine the effectiveness of the proposed strategies in reconstructing the velocity model and the importance of the partial overlap-frequency strategy. The Marmousi-II model has 244×681 grid cells with a grid interval of 10 m in both horizontal and vertical directions. We deploy 67 sources from 100 m to 6700 m with a source interval of 100 m and a depth of 20 m. A total of 681 receivers are distributed from 10 m to 6810 m with a receiver interval of 10 m and a depth of 20 m. A Ricker wavelet with a 30 Hz dominant frequency is used as the source function. Figures 1a and 1b show the true Marmousi-II P-wave velocity model and initial P-wave velocity model respectively. The initial velocity model is obtained by smoothing the true model with a Gaussian function.

Figures 2a and 2b show the monochromatic seismic data sets ($f = 20$ Hz) generated using the true Marmousi-II model in source-receiver domain for shot-profile FWI ($N_s = 67$) and ray parameter-receiver domain for f-p domain FWI with increasing the ray parameter from -0.5 s/km to 0.5 s/km ($\Delta p = 0.01$ s/km). For SP method, the whole data set shown in Figure 2a should be used for inversion in each iteration. While in TLPE and SU methods, the data sets with several ray parameters and single ray parameter are extracted for inversion. To evaluate the quality of the inversion results, we use the relative least-squares error (RLSE) ϵ:

$$\epsilon = \frac{\|\mathbf{m}_n - \mathbf{m}_t\|}{\|\mathbf{m}_0 - \mathbf{m}_t\|}, \tag{38}$$

where $\mathbf{m}_0$, $\mathbf{m}_t$ and $\mathbf{m}_n$ indicate the initial model, true model and the inverted model at the nth iteration. The RLSE of the initial model is $\epsilon_0 = 1$. Smaller RLSE means better inversion result. If the model parameters are reconstructed completely, RLSE ϵ approaches 0.

We perform two different frequency selection strategies for comparison with l-BFGS optimization method. In the mono-frequency strategy (as shown in Table 2), the frequency is increased sequentially from 1 Hz to 36 Hz by 1 Hz every 11 iterations. In the partial overlap-frequency selection strategy (as shown in Table 2), a group of 3 frequencies are used for inversion simultaneously and the frequency band expands every 11 iterations with overlapping 2 frequencies. For the SU strategy, we compare two different ray parameter selections. The random SU is implemented by randomly selecting the ray parameter in the range of [-0.05 s/km, 0.05 s/km] at each iteration. While for the sequential SU, the ray parameter is sequentially changed from -0.05 s/km to 0.05 s/km with an interval $\Delta p = 0.01$ s/km for each frequency group.

First, we obtain the inversion results by slant update strategy with fixed ray parameters $p = -0.01$ s/km and $p = -0.03$ s/km, as shown in Figures 3a and 3b respectively. As we can see, the inversion results with fixed ray parameters are contaminated by artifacts seriously and the main geological structures of the model are obscure. This is because single ray parameter is not enough to illuminate the subsurface layers with different dip angles. Figures 3c and 3e show the inversion results by random and sequential SU with mono-frequency strategy. Even though, the inversion results become much better in comparison with Figures 3a and 3b, they are still contaminated by strong cross-talk artifacts.

We next carry out the partial overlap-frequency strategy for inversion. Figures 3d and 3f show the inversion results by random and sequential SU with partial overlap-frequency strategy. Compared to Figures 3c and 3e, it can be observed that the artifacts have been suppressed effectively and the inversion results have been improved significantly, which demonstrate the importance of inverting multiple frequencies simultaneously for slant update strategy. Furthermore, the sequential SU strategy appears to provide a better inversion result than random SU strategy.

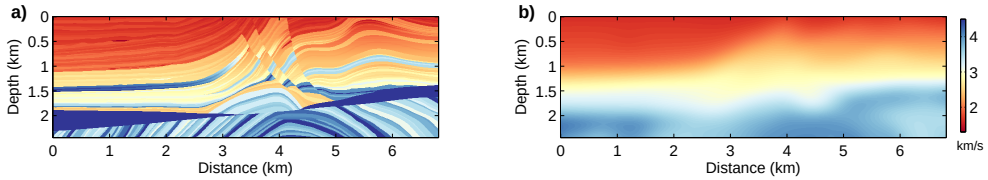

FIG. 1. (a) shows the true Marmousi-II P-wave velocity model; (b) shows the initial P-wave velocity model.

Computational cost comparison of different encoding methods

Next, we give the inversion results using TLPE and SP methods with l-BFGS optimization method for comparison to examine the computational efficiency of the slant update strategy. We perform TLPE method with two ray parameter settings. For the first one, the ray parameter is varied from -0.03 s/km to 0.03 s/km with an interval of 0.01 s/km ($N_p = 7$). For the second one, the ray parameter is varied from -0.05 s/km to 0.05 s/km with the same interval ($N_p = 11$).

Figures 4a and 4c show the inverted velocity models by mono-frequency TLPE method with $N_p = 7$ and $N_p = 11$ respectively. As we can see, compared to Figures 3c and 3e, the inversion results have been improved obviously by stacking ray parameters at each iteration with mono-frequency strategy. Furthermore, we observe that by stacking more ray parameters at each iteration, the inversion result more closely approaches the true model and the SP method provides best result, as shown in Figure 4e, but at the cost of extensive computation. Figures 4b, 4d and 4f are the inverted models using TLPH ($N_p = 7$), TLPH ($N_p = 11$) and SP methods with partial overlap-frequency strategy, which are better than those obtained with mono-frequency strategy. We also note that with partial overlap-frequency strategy, SU methods can get inversion results (Figures 3d and 3f) comparable to those by TLPE and SP methods (Figures 4b, 4d and 4f).

We show the RLSE vs. Iterations for different methods with mono-frequency strategy and partial overlap-frequency strategy in Figures 5a and 5b respectively. In Figure 5a, we see that the random and sequential SU methods fail to converge. While the SU methods with partial overlap-frequency strategy converge efficiently, as shown in Figure 5b. Figure 6 show the RLSE vs. Data amount for different methods. Figure 6b gives the enlarged view of red box area in Figure 6a. Table 3 lists the CPU running time for different methods with partial overlap-frequency strategy to achieve comparable quality inversion results. As can be seen in Table 3, to achieve the same quality inversion results (RLSE $\epsilon = 0.7$, 0.6 and 0.5), the CPU running times using sequential SU method are reduced approximately by 49, 47 and 30 times, compared to SP method. These illustrations verify that the proposed strategies can get comparable inversion results with reducing the computation burden con-

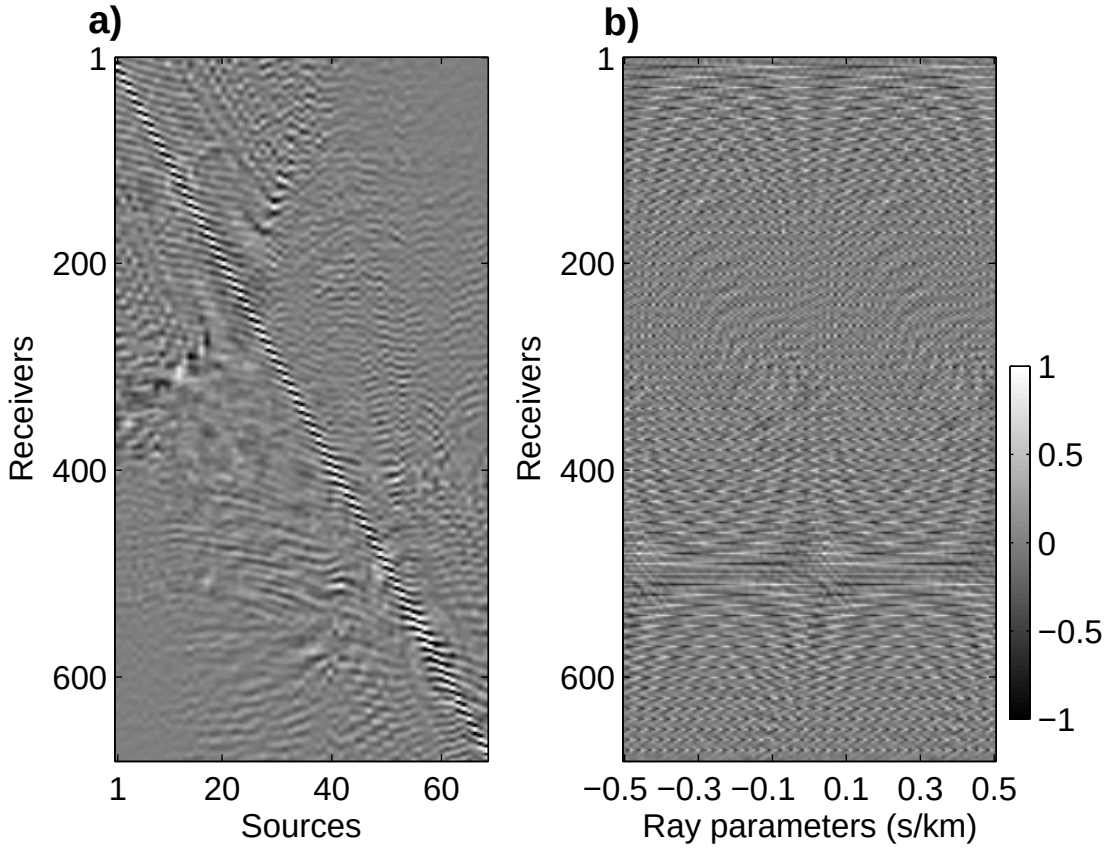

FIG. 2. The monochromatic seismic data set ($f = 20$ Hz) in source-receiver domain for shot-profile FWI (a) and ray parameter-receiver domain for f-p domain FWI (b).

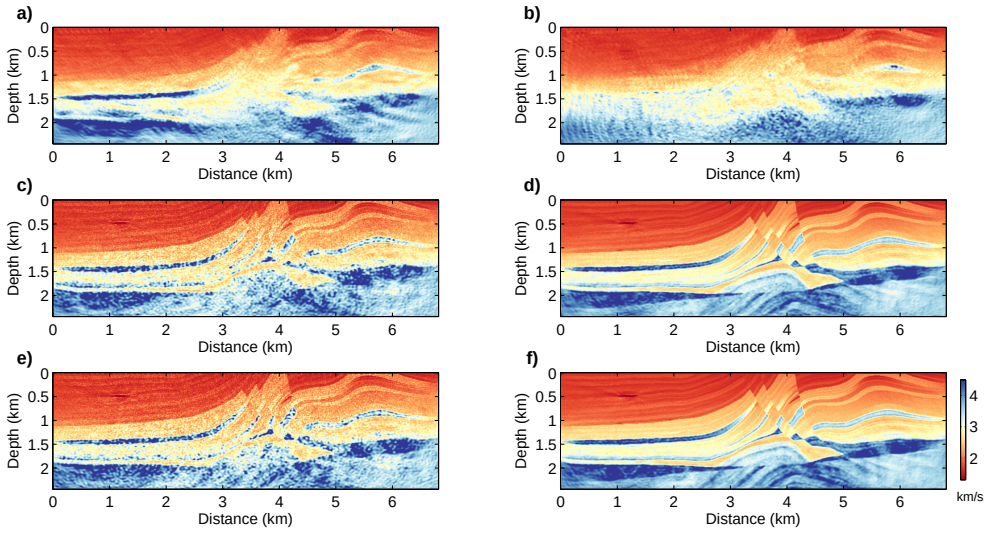

FIG. 3. (a) Mono-frequency slant update with $p = -0.01$ s/km; (b) Mono-frequency slant update with $p = -0.03$ s/km; (c) Mono-frequency random slant update; (d) Partial overlap-frequency random slant update ($\epsilon = 0.4862$); (e) Mono-frequency sequential slant update; (f) Partial overlap-frequency sequential slant update ($\epsilon = 0.4596$).

siderably.

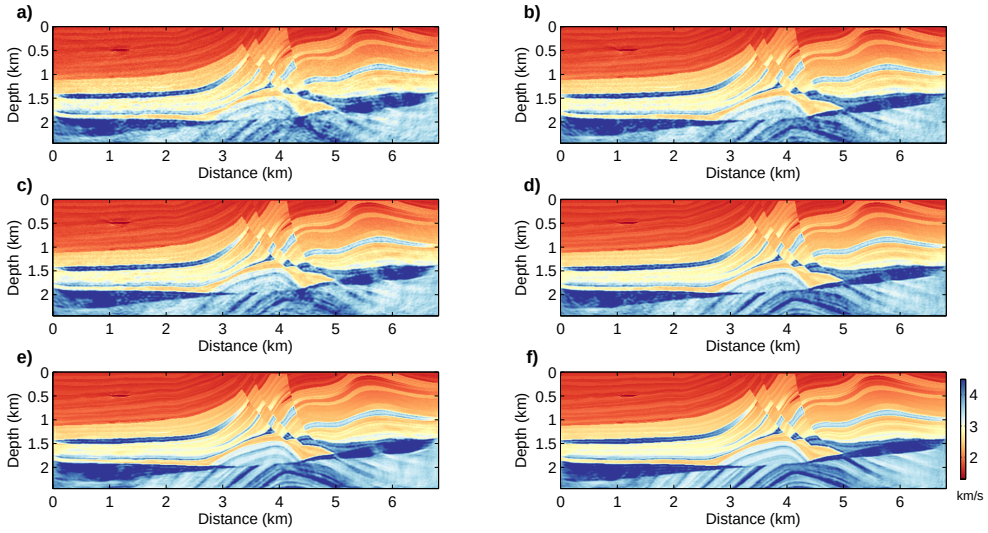

FIG. 4. (a) Mono-frequency TLPE with $N_p = 7$ ($\epsilon = 0.5305$); (b) Partial overlap-frequency TLPE with $N_p = 7$ ($\epsilon = 0.4251$); (c) Mono-frequency TLPE with $N_p = 11$ ($\epsilon = 0.4472$); (d) Partial overlap-frequency TLPE with $N_p = 11$ ($\epsilon = 0.3598$); (e) Mono-frequency SP ($\epsilon = 0.3777$); (f) Partial overlap-frequency SP ($\epsilon = 0.3064$).

Comparison of different optimization methods

In this numerical section, we implement the proposed slant update strategy with different optimization methods, including SD, NCG and truncated Gauss-Newton (TGN) methods, in comparison with l-BFGS method. In the TGN method, the Gauss-Newton Hessian

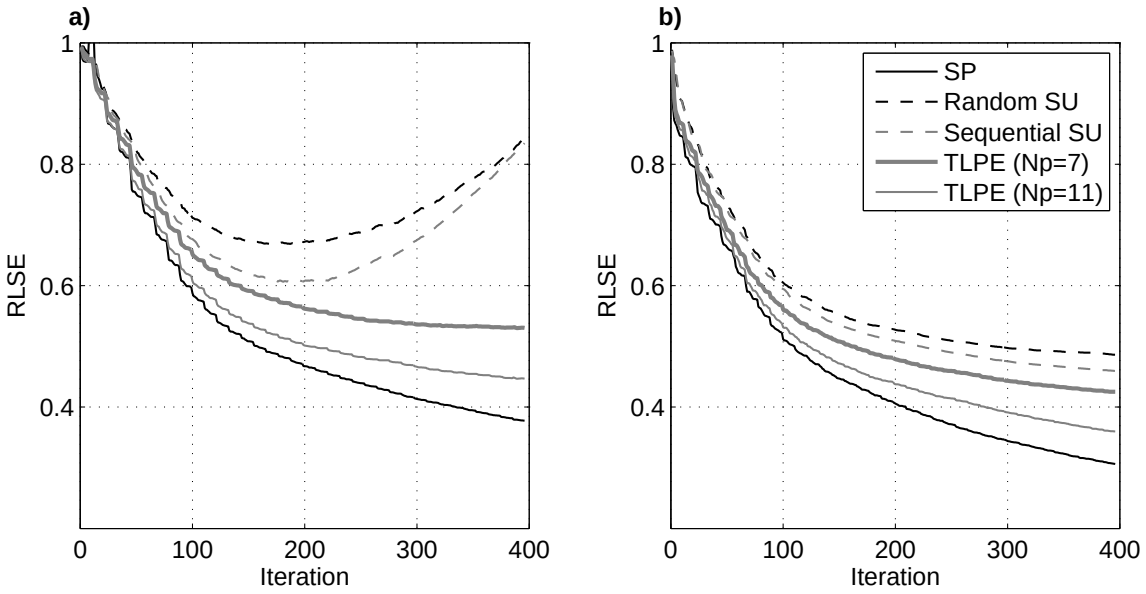

FIG. 5. RLSE vs. Iteration for different methods with mono-frequency strategy (a) and partial overlap-frequency strategy (b). The black-solid, black-dash, grey-dash, grey-bold and grey-thin lines indicate the RLSE as iteration proceeds for SP, random SU, sequential SU, TLPE ($N_p = 7$) and TLPE ($N_p = 11$) methods.

$\tilde{\mathbf{H}}_a$ (equation (44)) and diagonal pseudo-Hessian preconditioner (equation (55)) are both constructed with linear phase-encoding. The source-side ray parameter p_s and receiver-side ray parameter p_g both range from -0.01 s/km to 0.01 s/km with $\Delta p = 0.01$ s/km. The stopping criteria for the inner iteration is $\tilde{n}_{max} = 10$ and $\gamma_{min} = 2.0e$-1. The stabilization parameters are $\epsilon = 1.0e$-2 and $\lambda = 1.0e$-2 (equation (30))

Figures 7a and 7b show the inversion results by sequential SU with SD and NCG methods respectively. The partial overlap-frequency strategy is adopted. Compared to the inversion result obtained using l-BFGS method (Figure 3f), the inversion results obtained using SD and NCG methods are seen to be contaminated by more noise, even though, the velocity model can be reconstructed well. The green, blue and red lines in Figure 8 illustrate the RLSE for l-BFGS, NCG and SD optimization methods as the iteration proceeds. These numerical experiments verify that the l-BFGS method for f-p domain FWI with slant update strategy provides a faster convergence rate.

Sensitivity to random noise

Another major challenge of FWI with phase-encoding technique is the sensitivity of the method to noisy data (van Leeuwen et al., 2011). To test the resistance ability of the proposed strategies to random noise, we add Gaussian noise to the seismic data set with SNR $= 3$ and SNR $= 5$. SNR means signal to noise ratio and smaller SNR indicates stronger random noise. We use the RLSE change $\Delta\epsilon$ of the inverted model to evaluate the sensitivity of the inversion result to noisy data:

$$\Delta\epsilon = \frac{\tilde{\epsilon} - \epsilon}{\epsilon} \times 100\%, \tag{39}$$

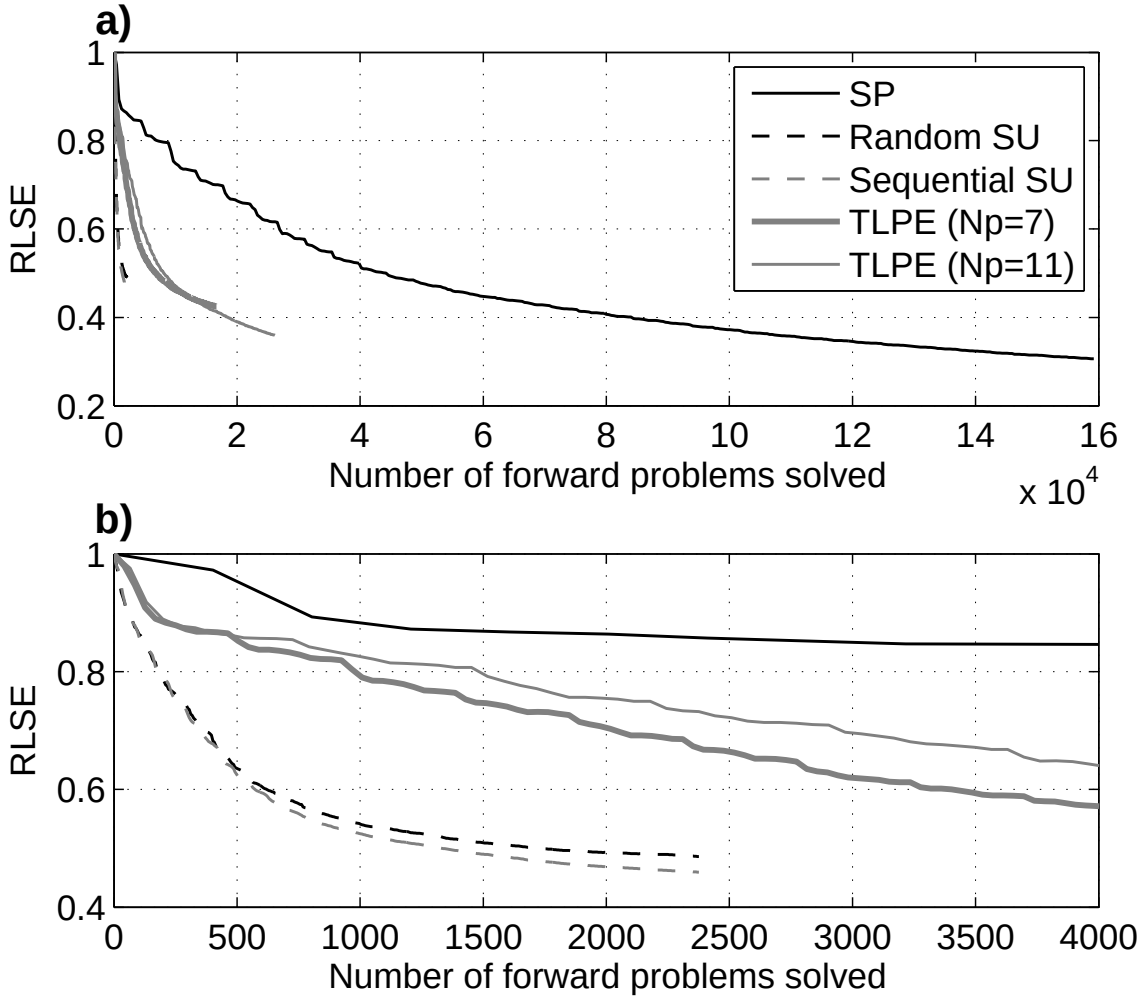

FIG. 6. RLSE vs. Number of forward problems solved. The black-solid, black-dash, grey-dash, grey-bold and grey-thin lines indicate the RLSE for SP, random SU, sequential SU, TLPE ($N_p = 7$) and TLPE ($N_p = 11$) methods. (b) is the enlarged view of (a).

where $\tilde{\epsilon}$ indicates the RLSE of the inverted model with noisy data. Smaller $\Delta\epsilon$ means less sensitivity to random noise.

The reconstructed models by partial overlap-frequency SP with SNR $= 3$ and SNR $= 5$ are presented in Figures 9a and 9b respectively. Figure 10a shows the RLSE vs. Iterations for the SP method with noisy data. Even though, noise effects are not obvious in the inverted models for SP method, the qualities of the final inversion results are decreased, compared to Figure 4f. Figures 9c and 9d are the inverted models by TLPE ($N_p = 7$) method with SNR $= 3$ and SNR $= 5$. The reconstructed models by SU method with SNR $= 3$ and SNR $= 5$ are shown in Figures 9e and 9f. Figure 10b and 10c shows the corresponding RLSE as iteration proceeds for TLPE and SU methods, respectively. For TLPE and SU methods, the artifacts caused by random noise become more obvious, compared to the inversion results with noise-free data (Figure 4b and Figure 3f).

We illustrate the RLSE ϵ of the inversion results and RLSE changes $\Delta\epsilon$ in Table 4

Table 3 Comparison of the CPU running time for different methods.

Methods	CPU time (s)	CPU time (s)	CPU time (s)
Partial overlap-frequency	$\epsilon = 0.7$	$\epsilon = 0.6$	$\epsilon = 0.5$
SP	5.424e3	8.653e3	13.044e3
TLPE ($N_p = 7$)	0.647e3	1.093e3	2.172e3
TLPE ($N_p = 11$)	0.954e3	1.506e3	2.608e3
Random SU	0.120e3	0.199e3	0.553e3
Sequential SU	0.116e3	0.183e3	0.428e3

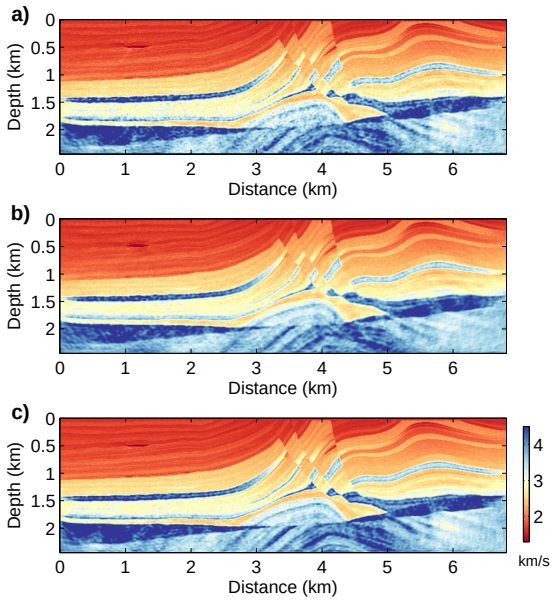

FIG. 7. (a) and (b) show the inversion results using partial overlap-frequency SU strategy with SD ($\epsilon = 0.5572$), NCG ($\epsilon = 0.5062$) and Hessian-free Gauss-Newton ($\epsilon = 0.3616$) methods respectively.

to compare the sensitivities of different methods to random noise at iterations $n = 100$, $n = 200$ and $n = 300$, which correspond to frequency groups 9-11 Hz, 18-20 Hz and 27-29 Hz. We can observe that at early iterations when using low frequencies for inversion, the inversion results are less sensitive to the random noise, compared to using high frequencies. Furthermore, we notice that SU method is not very sensitive to noisy data, especially when using low frequencies for inversion. Its random noise resistance ability is better than TLPE method ($N_p = 7$).

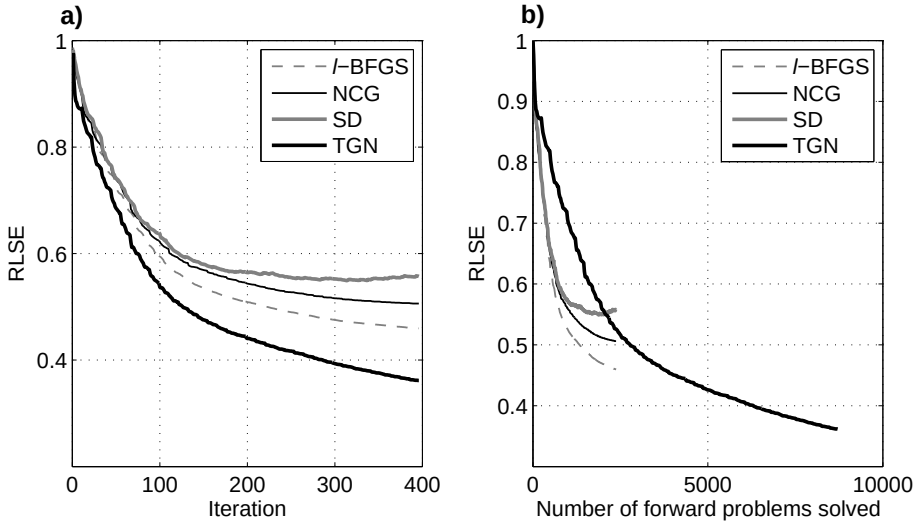

FIG. 8. (a)RLSE vs. Iterations for different optimization methods; (b) RLSE vs. Number of forward problems solved. The gray-dash, black-solid, gray-bold and black-bold lines indicate the l-BFGS, non-linear conjugate-gradient (NCG), steepest-descent (SD), and Hessian-free Gauss-Newton methods.

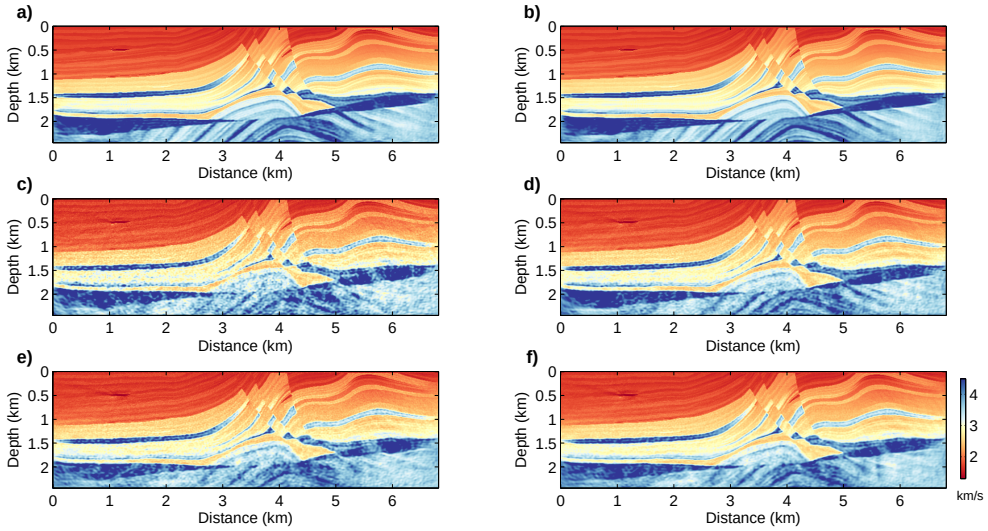

FIG. 9. Inversion results obtained by different methods with partial overlap-frequency strategy and noisy data. (a) SP method with SNR $= 3$ ($\tilde{\epsilon} = 0.3064$); (b) SP method with SNR $= 5$ ($\tilde{\epsilon} = 0.3251$); (c) TLPE ($N_p = 7$) method with SNR $= 3$ ($\tilde{\epsilon} = 0.5503$); (d) TLPE ($N_p = 7$) method with SNR $= 5$ ($\tilde{\epsilon} = 0.4675$); (e) SU method with SNR $= 3$ ($\tilde{\epsilon} = 0.5078$); (f) SU method with SNR $= 5$ ($\tilde{\epsilon} = 0.4705$).

DISCUSSION

The numerical experiments presented in this research verify that the proposed f-p domain FWI with slant update strategy can reconstruct the velocity model very well with reducing the computational cost considerably. While the inversion results may be influenced by the ray parameter setting obviously. The ray parameter range should be determined to illuminate the subsurface layers with different dip angles. The ray parameters should be sampled properly according to the anti-aliasing rule of τ-p transform (Zhang et al., 2005)

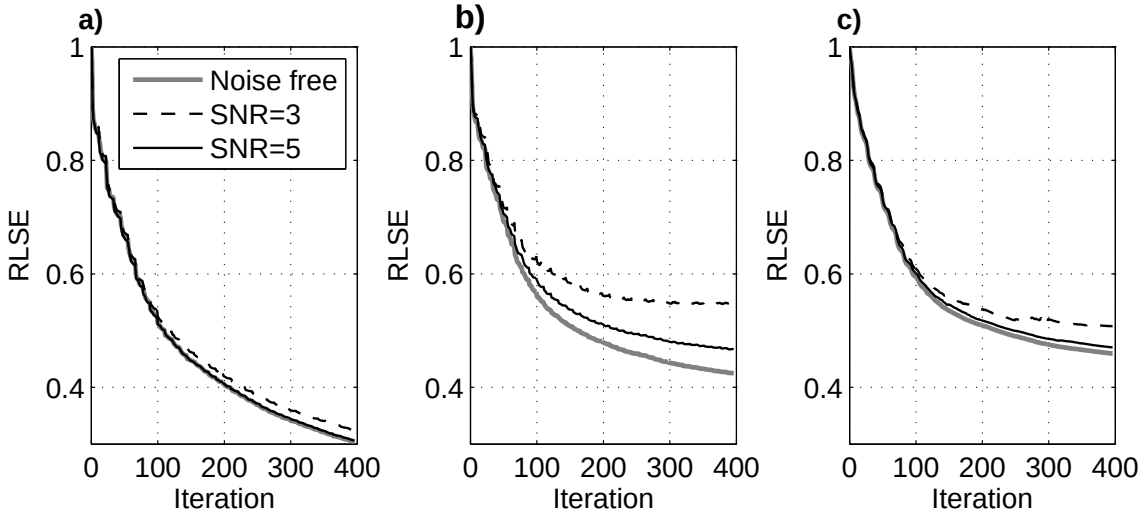

FIG. 10. RLSE vs. Iterations for partial overlap-frequency SP (a), TLPE ($N_p = 7$) (b) and SU (c) methods with noisy data. The black, red, and blue lines indicate the RLSE for noise-free data, noisy data with SNR $= 3$ and SNR $= 5$ respectively.

Table 4 Comparison of the sensitivities to random noise for different methods.

Methods	Iteration n	f (Hz)	Noise-free ϵ	$\tilde{\epsilon}$ with SNR $= 5$ ($\Delta\epsilon$)	$\tilde{\epsilon}$ with SNR $= 3$ ($\Delta\epsilon$)
SP	100	9-11	0.5116	0.5135 (0.37 %)	0.5251 (2.64 %)
	200	18-20	0.4048	0.4060 (0.30 %)	0.4190 (3.51 %)
	300	27-29	0.3418	0.3441 (0.67 %)	0.3597 (5.24 %)
TLPE	100	9-11	0.5607	0.5886 (4.96 %)	0.6229 (11.08 %)
	200	18-20	0.4789	0.5103 (6.55 %)	0.5613 (17.20 %)
	300	27-29	0.4433	0.4801 (8.31 %)	0.5479 (23.60 %)
SU	100	9-11	0.5944	0.6009 (1.08 %)	0.6086 (2.32 %)
	200	18-20	0.5090	0.5176 (1.66 %)	0.5378 (5.36 %)
	300	27-29	0.4753	0.4855 (2.09 %)	0.5201 (8.62 %)

and Radon transform (Vigh and Starr, 2008; Pan et al., 2015a). Examining the influences ray parameter setting needs further studies.

FWI is an ill-posed problem meaning that an infinite number of models matches the data. Regularization technique can alleviate the non-uniqueness of the ill-posed inverse problem (Menke, 1984). No regularization technique is used in this research. So, for further research, introducing regularization term is necessary for the proposed strategies. In this paper, we employ l-BFGS method for inversion compared to gradient-based methods. The propose strategies should be extended to more advanced optimization methods, such as Hessian-free optimization methods.

The numerical experiments are carried out for inversion from very low frequency (1 Hz), which mitigates the cycle-skipping problem. While in real data application, the low frequency information is always missed. Hence, testing the performance of the proposed

methods without low frequencies is necessary. In this research, we only discuss the sensitivity of the methods to random noise. The influences of other types of noise (e.g., surface related multiples) should also be taken into consideration. For marine-streamer survey, the inconsistent acquisition geometries between the observed data and modeled data can result in strong artifacts, which can be mitigated by correlation-based method (Choi and Alkhalifah, 2012).

CONCLUSION

In this paper, employing a linear-phase encoding technique, we develop an efficient frequency-ray parameter (f-p) domain FWI with slant update strategy for reducing the computation burden. The gradient constructed with single ray parameter is used to update the model at each iteration but the ray parameter varies as iteration proceeds. In addition to the common mono-frequency selection strategy for inversion, we also employ a partial overlap-frequency selection strategy for comparison. We notice that inverting multiple frequencies simultaneously is important to reduce the cross-talk artifacts and guarantee the quality of the inversion result for the proposed strategies. The numerical examples are illustrated to show that the slant update strategy can obtain comparable quality inversion results with reducing the computational cost significantly in comparison with traditional methods. We also show the faster convergence rate of l-BFGS optimization method compared to the gradient-based methods. Finally, we illustrate that the proposed slant update strategy is robust to random noise, especially when using low frequencies for inversion.

ACKNOWLEDGMENTS

The authors thank the sponsors of CREWES for continued support. This work was funded by CREWES industrial sponsors and NSERC (Natural Science and Engineering Research Council of Canada) through the grant CRDPJ 461179-13. Author 1 was also supported by a SEG/Chevron scholarship.

THE PHASE-ENCODED GAUSS-NEWTON HESSIAN

With Born approximation, the Gauss-Newton Hessian $\mathbf{H}_a$ (equation (25)) can be written with four Green's functions (Plessix and Mulder, 2004):

$$\mathbf{H}_a\left(\mathbf{x},\mathbf{x}'\right) \approx \sum_{\mathbf{x}_g}\sum_{\mathbf{x}_s}\sum_{\omega} \Re\left(\omega^4|f_s(\omega)|^2 G\left(\mathbf{x},\mathbf{x}_s,\omega\right) G\left(\mathbf{x}_g,\mathbf{x},\omega\right) G^*\left(\mathbf{x}',\mathbf{x}_s,\omega\right) G^*\left(\mathbf{x}_g,\mathbf{x}',\omega\right)\right). \tag{40}$$

As we can see, the source-side and receiver-side Green's functions need to be constructed to form the Gauss-Newton Hessian. Similarly, the linear phase-encoding technique can be employed to construct the Gauss-Newton Hessian, which gives the phase-encoded Gauss-

Newton Hessian $\tilde{\mathbf{H}}_a$:

$$\begin{aligned}\tilde{\mathbf{H}}_a\left(\mathbf{x},\mathbf{x}'\right) = &\sum_{\mathbf{x}_s}\sum_{\mathbf{p}_s}\sum_{\omega}\Re\left(\omega^4|f_s(\omega)|^2 G\left(\mathbf{x},\mathbf{x}_s,\omega\right)A_s\left(\omega\right)\mathbf{exp}\left(i\omega\mathbf{p}_s\cdot\left(\mathbf{x}_s-\hat{\mathbf{x}}_s\right)\right)\right)\\ &\times\sum_{\mathbf{x}'_s}\sum_{\mathbf{p}_s}\sum_{\omega}\Re\left(G^*\left(\mathbf{x}',\mathbf{x}'_s,\omega\right)A_s^*\left(\omega\right)\mathbf{exp}\left(i\omega\mathbf{p}_s\cdot\left(\hat{\mathbf{x}}_s-\mathbf{x}'_s\right)\right)\right)\\ &\times\sum_{\mathbf{x}_g}\sum_{\mathbf{p}_g}\sum_{\omega}\Re\left(G\left(\mathbf{x},\mathbf{x}_g,\omega\right)A_g\left(\omega\right)\mathbf{exp}\left(i\omega\mathbf{p}_g\cdot\left(\mathbf{x}_g-\hat{\mathbf{x}}_g\right)\right)\right)\\ &\times\sum_{\mathbf{x}'_g}\sum_{\mathbf{p}_g}\sum_{\omega}\Re\left(G^*\left(\mathbf{x}',\mathbf{x}'_g,\omega\right)A_g^*\left(\omega\right)\mathbf{exp}\left(i\omega\mathbf{p}_g\cdot\left(\hat{\mathbf{x}}_g-\mathbf{x}'_g\right)\right)\right),\end{aligned} \tag{41}$$

where $A_s(\omega)$ and $A_g(\omega)$ are the source-side and receiver-side weighting functions for Hessian construction, $\mathbf{p}_s$ and $\mathbf{p}_g$ are source-side and receiver-side ray parameter vectors, and for $\hat{\mathbf{x}}_s$ and $\hat{\mathbf{x}}_g$, we have:

$$\begin{cases}\hat{\mathbf{x}}_s=\mathbf{x}_s^0, p_s\geq 0\\ \hat{\mathbf{x}}_s=\mathbf{x}_s^{\text{end}}, p_s<0\end{cases}, \tag{42}$$

and

$$\begin{cases}\hat{\mathbf{x}}_g=\mathbf{x}_g^0, p_g\geq 0\\ \hat{\mathbf{x}}_g=\mathbf{x}_g^{\text{end}}, p_g<0\end{cases}, \tag{43}$$

where $\mathbf{x}_g^0$ and $\mathbf{x}_g^{\text{end}}$ indicate the positions of the intial reciever and right most receiver. Reorganizing equation (41) gives:

$$\begin{aligned}\tilde{\mathbf{H}}_a\left(\mathbf{x},\mathbf{x}'\right) = &\sum_{\mathbf{x}_s}\sum_{\mathbf{x}'_s}\sum_{\mathbf{p}_s}\sum_{\omega}\Re\left(\omega^4|f_s(\omega)|^2 G\left(\mathbf{x},\mathbf{x}_s,\omega\right)G^*\left(\mathbf{x}',\mathbf{x}'_s,\omega\right)A_s^2\left(\omega\right)\mathbf{exp}\left(i\omega\mathbf{p}_s\cdot\left(\mathbf{x}_s-\mathbf{x}'_s\right)\right)\right)\\ &\times\sum_{\mathbf{x}_g}\sum_{\mathbf{x}'_g}\sum_{\mathbf{p}_g}\sum_{\omega}\Re\left(G\left(\mathbf{x},\mathbf{x}_g,\omega\right)G^*\left(\mathbf{x}',\mathbf{x}'_g,\omega\right)A_g^2\left(\omega\right)\mathbf{exp}\left(i\omega\mathbf{p}_g\cdot\left(\mathbf{x}_g-\mathbf{x}'_g\right)\right)\right)\end{aligned} \tag{44}$$

If $\mathbf{x}_s=\mathbf{x}'_s$ and $\mathbf{x}_g=\mathbf{x}'_g$, the phase-encoded Hessian $\tilde{\mathbf{H}}_a$ is identical to the non-encoded Hessian $\mathbf{H}_a$. If $\mathbf{x}_s=\mathbf{x}'_s$ or $\mathbf{x}_g=\mathbf{x}'_g$, the phase-encoded Hessian $\tilde{\mathbf{H}}_a$ becomes the cross-talk term $\mathbf{H}_a^{\text{cross}}$. So, the phase-encoded Hessian $\tilde{\mathbf{H}}_a$ is equal to the summation of non-encoded Hessian $\mathbf{H}_a$ with the cross-talk term: $\tilde{\mathbf{H}}_a=\mathbf{H}_a+\mathbf{H}_a^{\text{cross}}$. To disperse this cross-talk term, stacking over sufficient ray parameters is required. The source-side encoding function $\psi_s\left(\mathbf{x}_s,\mathbf{x}'_s,\omega\right)$ and receiver-side encoding function $\psi_g\left(\mathbf{x}_g,\mathbf{x}'_g,\omega\right)$ can be expressed as:

$$\psi_s\left(\mathbf{x}_s,\mathbf{x}'_s,\omega\right)=\sum_{\mathbf{p}_s}A_s^2\left(\omega\right)\mathbf{exp}\left(i\omega\mathbf{p}_s\cdot\left(\mathbf{x}_s-\mathbf{x}'_s\right)\right). \tag{45}$$

$$\psi_g\left(\mathbf{x}_g,\mathbf{x}'_g,\omega\right)=\sum_{\mathbf{p}_g}A_g^2\left(\omega\right)\mathbf{exp}\left(i\omega\mathbf{p}_g\cdot\left(\mathbf{x}_g-\mathbf{x}'_g\right)\right). \tag{46}$$

Following equation (20), if the numbers of ray parameters in $\mathbf{p}_s$ and $\mathbf{p}_g$ are large enough, applying integration over the ray parameters gives:

$$\psi_s\left(\mathbf{x}_s,\mathbf{x}'_s,\omega\right)=\frac{A_s^2\left(\omega\right)}{\omega^2}\delta\left(\mathbf{x}_s-\mathbf{x}'_s\right). \tag{47}$$

$$\psi_g\left(\mathbf{x}_g,\mathbf{x}'_g,\omega\right)=\frac{A_g^2\left(\omega\right)}{\omega^2}\delta\left(\mathbf{x}_g-\mathbf{x}'_g\right). \tag{48}$$

Inserting equations (47) and (48) into equation (44) gives:

$$\begin{aligned}\tilde{\mathbf{H}}_a\left(\mathbf{x},\mathbf{x}'\right)=&\sum_{\mathbf{x}_s}\sum_{\mathbf{x}'_s}\sum_{\omega}\Re\left(\omega^4|f_s(\omega)|^2G\left(\mathbf{x},\mathbf{x}_s,\omega\right)G^*\left(\mathbf{x}',\mathbf{x}'_s,\omega\right)\frac{A_s^2\left(\omega\right)}{\omega^2}\delta\left(\mathbf{x}_s-\mathbf{x}'_s\right)\right)\\&\times\sum_{\mathbf{x}_g}\sum_{\mathbf{x}'_g}\sum_{\omega}\Re\left(G\left(\mathbf{x},\mathbf{x}_g,\omega\right)G^*\left(\mathbf{x}',\mathbf{x}'_g,\omega\right)\frac{A_g^2\left(\omega\right)}{\omega^2}\delta\left(\mathbf{x}_g-\mathbf{x}'_g\right)\right),\end{aligned} \tag{49}$$

Making $A_s^2\left(\omega\right)=A_g^2\left(\omega\right)=\omega^2$ and considering the sifting property of the delta function, the cross-talk artifacts in the phase-encoded Hessian can be totally reduced:

$$\begin{aligned}\tilde{\mathbf{H}}_a\left(\mathbf{x},\mathbf{x}'\right)&=\mathbf{H}_a\left(\mathbf{x},\mathbf{x}'\right)\\&=\sum_{\mathbf{x}_s}\sum_{\mathbf{x}_g}\sum_{\omega}\Re\left(\omega^4|f_s(\omega)|^2G\left(\mathbf{x},\mathbf{x}_s,\omega\right)G^*\left(\mathbf{x}',\mathbf{x}_s,\omega\right)G\left(\mathbf{x},\mathbf{x}_g,\omega\right)G^*\left(\mathbf{x}',\mathbf{x}_g,\omega\right)\right).\end{aligned} \tag{50}$$

THE PHASE-ENCODED DIAGONAL PSEUDO-HESSIAN PRECONDITIONER

The pseudo-Hessian $\mathbb{H}$ is constructed by replacing the Fréchet derivative wavefield with the virtual source $\tilde{f}_s\left(\omega\right)$ in the correlation process (Shin et al., 2001b). Taking partial derivative with respect to model parameter on both sides of equation (1) gives:

$$\left(\nabla^2+\mathbf{m}\left(\mathbf{x}\right)\omega^2\right)\frac{\partial\mathbf{u}\left(\mathbf{x},\mathbf{x}_s,\omega\right)}{\partial\mathbf{m}\left(\mathbf{x}\right)}=\tilde{f}_s\left(\omega\right), \tag{51}$$

where $\tilde{f}_s\left(\mathbf{x},\omega\right)=-\omega^2\mathbf{u}\left(\mathbf{x},\mathbf{x}_s,\omega\right)$ is the virtual source. Considering equation (2), the virtual source can be expressed with Green's function:

$$\tilde{f}_s\left(\mathbf{x},\omega\right)=-\omega^2 f_s\left(\omega\right)G\left(\mathbf{x},\mathbf{x}_s,\omega\right). \tag{52}$$

The pseudo-Hessian is obtained by correlating two virtual sources (Shin et al., 2001a; Plessix and Mulder, 2004):

$$\mathbb{H}\left(\mathbf{x},\mathbf{x}'\right)=\tilde{f}_s\left(\mathbf{x},\omega\right)\tilde{f}_s^*\left(\mathbf{x}',\omega\right)=\sum_{\mathbf{x}_s}\sum_{\mathbf{x}_g}\sum_{\omega}\Re\left(\omega^4\mid f_s\left(\omega\right)\mid^2G\left(\mathbf{x},\mathbf{x}_s,\omega\right)G^*\left(\mathbf{x}',\mathbf{x}_s,\omega\right)\right), \tag{53}$$

where when $\mathbf{x}=\mathbf{x}'$, we can obtain the diagonal pseudo-Hessian $\mathbb{H}_{\text{diag}}$, the auto-correlation of the two virtual sources:

$$\mathbb{H}_{\text{diag}}\left(\mathbf{x}\right)=\sum_{\mathbf{x}_s}\sum_{\mathbf{x}_g}\sum_{\omega}\Re\left(\omega^4\mid f_s\left(\omega\right)\mid^2G\left(\mathbf{x},\mathbf{x}_s,\omega\right)G^*\left(\mathbf{x},\mathbf{x}_s,\omega\right)\right). \tag{54}$$

Similarly, following the derivation process in Appendix A, the diagonal pseudo-Hessian can also be constructed with linear phase-encoding :

$$\begin{aligned}\tilde{\mathbb{H}}_{\text{diag}}\left(\mathbf{x}\right)=&\sum_{\mathbf{x}_s}\sum_{\mathbf{x}'_s}\sum_{\mathbf{x}_g}\sum_{\mathbf{p}_s}\sum_{\omega}\Re\left(\omega^4\mid f_s\left(\omega\right)\mid^2G\left(\mathbf{x},\mathbf{x}_s,\omega\right)G^*\left(\mathbf{x},\mathbf{x}_s,\omega\right)\right.\\&\left.A_s^2\left(\omega\right)\mathbf{exp}\left(i\omega\mathbf{p}_s\cdot\left(\mathbf{x}_s-\mathbf{x}'_s\right)\right)\right).\end{aligned} \tag{55}$$

This phase-encoded diagonal pseudo-Hessian is used as the preconditioner in truncated Gauss-Newton method (equation (30)). Furthermore, no additional computation cost is needed for constructing this preconditioner.

REFERENCES

Anagaw, A. Y., 2014, Full waveform inversion using simultaneous encoded sources based on first and second-order optimization methods: Ph.D. thesis, University of Alberta.

Anagaw, A. Y., and Sacchi, M. D., 2014, Comparison of multifrequency selection strategies for simultaneous-source full-waveform inversion: Geophysics, **79**, R165–R181.

Armijo, L., 1966, Minimization of functions having Lipschitz continuous first partial derivatives: Pacific J. Math, **16**, 1–3.

Ben-Hadj-Ali, H., Operto, S., and Virieux, J., 2011, An efficient frequency-domain full waveform inversion method using simultaneous encoded sources: Geophysics, **76**, WCC177–WCC188.

Boonyasiriwat, C., Valasek, P., Routh, P., Cao, W., Schuster, G. T., and Macy, B., 2009, An efficient multiscale method for time-domain waveform tomography: Geophysics, **74**, WCC59–WCC68.

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

Broyden, C. G., 1970, The convergence of a class of double-rank minimization algorithms: IMA Journal of Applied Mathematics, **6**, 222–231.

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

Byrd, R. H., Lu, P., and Nocedal, J., 1995, A limited memory algorithm for bound constrained optimization: SIAM Journal on Scientific and Statistical Computing, **16**, 1190–1208.

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

Choi, Y., and Alkhalifah, T., 2012, Application of multi-source waveform inversion to marine streamer data using the global correlation norm: Geophysical Prospecting, **60**, 748–758.

Davis, T. A., and Duff, I. S., 1997, An unsymmetric pattern multifrontal method for sparse lu factorization: SIAM Journal on Matrix Analysis and Applications, **18**, 140–158.

Engquist, B., and Majda, A., 1977, Absorbing boundary conditions for the numerical simulations of waves: Mathematical Computation, **31**, 629–651.

Fletcher, R., 1970, A new approach to variable metric algorithms: The Computer Journal, **13**, 317–322.

Fletcher, R., and Reeves, C. M., 1964, Function minimization by conjugate gradients: Comp. J., **7**, 149–154.

Gauthier, O., Virieux, J., and Tarantola, A., 1986, Two-dimensional nonlinear inversion of seismic waveforms: numerical results: Geophysics, **51**, 1387–1403.

Goldfarb, D., 1970, A family of variable-metric methods derived by variational means: Mathematics of Computation, 23–26.

Hu, W., Abubakar, A., Habashy, T. M., and Liu, J., 2011, Preconditioned non-linear conjugate gradient method for frequency domain full-waveform seismic inversion: Geophysical Prospecting, **59**, 477–491.

Jo, C. H., Shin, C., and Suh, J. H., 1996, An optimal 9-point, finite-difference, frequency-space, 2-d scalar wave extrapolator: Geophysics, **61**, 529–537.

Kim, Y., Cho, H., Min, D. J., and Shin, C., 2011, Comparison of frequency-selection strategies for 2D frequency-domain acoustic waveform inversion: Pure and Applied Geophysics, **168**, 1715–1727.

Krebs, J. R., Anderson, J. E., Henkley, D., Neelamani, R., Lee, S., Baumstein, A., and Lacasse, M., 2009, Fast full-wavefield seismic inversion using encoded sources: Geophysics, **74**, WCC177–WCC188.

Kwon, T., Seol, S. J., and Byun, J., 2015, Efficient full-waveform inversion with normalized plane-wave data: Geophysical Journal International, **201**, 53–60.

Lailly, P., 1983, The seismic inverse problem as a sequence of before stack migration: Conference on Inverse Scattering, Theory and Applications, SIAM, Expanded Abstracts, 206–220.

Liu, F., Hanson, D. W., Whitemore, N. D., Day, R. S., and Stolt, R. H., 2006, Toward a unified analysis for source plane-wave migration: Geophysics, **71**, S129–S139.

Ma, Y., and Hale, D., 2012, Quasi-Newton full-waveform inversion with a projected Hessian matrix: Geophysics, **77**, R207–R216.

Marfurt, K., 1984, Accuracy of finite-difference and finite-elements modeling of the scalar and elastic wave equation: Geophysics, **49**, 533–549.

Menke, W., 1984, Geophysical data analysis: Discrete inverse theory: Academic Press.

Métivier, L., Bretaudeau, F., Brossier, R., Virieux, J., and Operto, S., 2014, Full waveform inversion and the truncated Newton method: quantitative imaging of complex subsurface structures: Geophysical Prospecting, **62**, 1–23.

Métivier, L., Brossier, R., Virieux, J., and Operto, S., 2012, The truncated Newton method for full waveform inversion: SEG Technical Program Expanded Abstracts, 1–5.

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

Mora, P., 1987, Nonlinear two-dimensional elastic inversion of multioffset seismic data: Geophysics, **52**, 1211–1228.

Morton, S. A., and Ober, C. C., 1998, Faster shot-record depth migrations using phase encoding test: SEG Technical Program Expanded Abstracts, 1131–1134.

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

Nocedal, J., 1980, Updating quasi-Newton matrices with limited storage: Mathematics of Computation, **35**, 773–782.

Nocedal, J., and Wright, S. J., 2006, Numerical Optimization: Springer.

Pan, W., Innanen, K. A., and Margrave, G. F., 2014a, A comparison of different scaling methods for least-squares migration/inversion: EAGE Expanded Abstracts, We G103 14.

Pan, W., Innanen, K. A., Margrave, G. F., and Cao, D., 2015a, Efficient pseudo-Gauss-Newton full-waveform inversion in the τ-p domain: Geophysics, **80**, R225–R14.

Pan, W., Innanen, K. A., Margrave, G. F., Fhler, M. C., Fang, X., and Li, J., 2015b, Estimation of elastic constants in HTI media using Gauss-Newton and Full-Newton multi-parameter full-waveform inversion: SEG Technical Program Expanded Abstracts, 1177–1182.

Pan, W., Margrave, G. F., and Innanen, K. A., 2014b, Iterative modeling migration and inversion (IMMI): Combining full waveform inversion with standard inversion methodology: SEG Technical Program Expanded Abstracts, 938–943.

Pica, A., Diet, J. P., and Tarantola, A., 1990, Nonlinear inversion of seismic reflection data in a laterally invariant medium: Geophysics, **55**, 284–292.

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.

Plessix, R. E., and Mulder, W. A., 2004, Frequency-domain finite-difference amplitude-preserving migration: Geophysical Journal International, **157**, 975–987.

Pratt, R. G., and Chapman, C. H., 1992, Traveltime tomography in anisotropic media-II. application: Geophysical Journal International, **109**, 20–37.

Pratt, R. G., Shin, C., and Hicks, G. J., 1998, Gauss-Newton and full Newton methods in frequency-space seismic waveform inversion: Geophysical Journal International, **133**, 341–362.

Romero, L. A., Ghiglia, D. C., Ober, C. C., and Morton, S. A., 2000, Phase encoding of shot records in prestack migration: Geophysics, **65**, 426–436.

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

Shanno, D. F., 1970, Conditioning of quasi-Newton methods for function minimization: Mathematics of Computation, **24**, 647–656.

Shin, C., Jang, S., and Min, D., 2001a, Improved amplitude preservation for prestack depth migration by inverse scattering theory: Geophysical Prospecting, **49**, 592–606.

Shin, C., Yoon, K., Marfurt, K. J., Park, K., Yang, D., Lim, H. Y., Chung, S., and Shin, S., 2001b, Efficient calculation of a partial-derivative wavefield using reciprocity for seismic imaging and inversion: Geophysics, **66**, 1856–1863.

Sirgue, L., and Pratt, R. G., 2004, Efficient waveform inversion and imaging: A strategy for selecting temporal frequencies: Geophysics, **69**, 231–248.

Stoffa, P. L., Sen, M. K., Seifoullaev, R., Pestana, R., and Fokkema, J., 2006, Plane-wave depth migration: Geophysics, **71**, S261–S272.

Stolt, R. H., and Benson, A., 1986, Seismic migration: Theory and practice: Geophysical Press.

Tang, Y., 2009, Target-oriented wave-equation least-squares migration/inversion with phase-encoded Hessian: Geophysics, **74**, WCA95–WCA107.

Tao, Y., and Sen, M. K., 2013, Frequency-domain full waveform inversion with plane-wave data: Geophysics, **78**, R13–R23.

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

Valenciano, A., 2008, Imaging by wave-equation inversion: Ph.D. thesis, Stanford University.

van Leeuwen, T., Aravkin, A. Y., and Herrmann, F. J., 2011, Seismic waveform inversion by stochastic optimization: International Journal of Geophysics, 1–18.

Vigh, D., and Starr, E. W., 2008, 3D prestack plane-wave, full-waveform inversion: Geophysics, **73**, VE135–VE144.

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

Wang, B., Gao, F., Wheaton, D., and Dirks, V., 2006, Model-based decimation of input data for delayed-shot/plane-wave migration for the purpose of subsalt velocity model building: SEG Technical Program Expanded Abstracts, 2186–2190.

Wu, S., Wang, Y., Zheng, Y., and Chang, X., 2015, Limited-memory BFGS based least-squares pre-stack Kirchhoff depth migration: Geophysical Journal International, **202**, 738–747.

Zhang, Y., Sun, J., Notfors, C., Gray, S. H., Chernis, L., and Young, J., 2005, Delayed-shot 3D depth migration: Geophysics, **70**, E21–E28.