SIAM J. Sci. Comput.
Vol. 45, No. 3, pp. B314–B336

SEISMIC TOMOGRAPHY WITH RANDOM BATCH GRADIENT RECONSTRUCTION[*]

YIXIAO HU[†], LIHUI CHAI[‡], ZHONGYI HUANG[†], AND XU YANG[§]

Abstract. Seismic tomography solves high-dimensional optimization problems in the imaging of Earth's subsurface structures. In this paper, we propose using random batch methods to construct the gradient used for iterations in seismic tomography. Specifically, we use the frozen Gaussian approximation to compute seismic wave propagation and construct stochastic gradients by random batch methods. This method inherits the spirit of stochastic gradient descent methods for solving high-dimensional optimization problems. The proposed idea is general in the sense that it does not rely on the use of frozen Gaussian approximation, and one can replace it with any other efficient wave propagation solver, e.g., Gaussian beam methods and spectral element methods. We prove the convergence of the random batch method in the mean-square sense and show the numerical performance of the proposed method by two- and three-dimensional examples of wave-equation-based travel-time inversion and full-waveform inversion, respectively. As a by-product, we also prove the convergence of the accelerated full-waveform inversion using dynamic mini-batches and spectral element methods.

Key words. seismic tomography, random batch method, frozen Gaussian approximation, high-dimensional optimization, inverse problems

MSC codes. 86A22, 65M32, 65M99, 35B40

DOI. 10.1137/21M1452342

1. Introduction.

Seismic tomography can provide crucial information via computed images of Earth's subsurface structures at different scales and enhances our understanding of tectonics, volcanism, and geodynamics [1, 33, 31, 45]. Wave-equation-based seismic tomography iteratively solves nonlinear high-dimensional optimization problems for velocity models by computing seismograms and sensitivity kernels in complex models [39, 24, 23, 38]. Successful applications include imaging the velocity models of the southern California crust [36, 37], the European upper mantle [46], the North Atlantic region [32], and the Japan islands [34]. The performance of seismic tomography is restricted by how accurately and efficiently one can compute synthetic seismograms and sensitivity kernels [39], which are used to construct descent directions of velocity models for iterations. The computation of synthetic seismograms

[*] Submitted to the journal's Computational Methods in Science and Engineering section October 13, 2021; accepted for publication (in revised form) January 19, 2023; published electronically June 8, 2023.

https://doi.org/10.1137/21M1452342

Funding: The second author was partially supported by National Key R&D Program of China grants 2021YFA1003001 and 2020YFA0712503, NSFC projects 12271537 and 11901601, and the Fundamental Research Funds for the Central Universities, Sun Yat-sen University under grant 2021qntd21. The third author was partially supported by NSFC projects 1205104, 11871298, and 81930119. The fourth author was partially supported by NSF grants DMS-1818592 and DMS-2109116.

[†] Department of Mathematical Sciences, Tsinghua University, Beijing, 100084, China (yx-hu16@mails.tsinghua.edu.cn, zhongyih@tsinghua.edu.cn).

[‡] School of Mathematics, Sun Yat-sen University, Guangzhou, 510275, China (chailihui@mail.sysu.edu.cn).

[§] Department of Mathematics, University of California, Santa Barbara, CA 93106 USA (xuyang@math.ucsb.edu).

could be extremely expensive for large-scale and high-frequency three-dimensional (3D) simulations. Velocity models live in high-dimensional space, making it challenging to compute descent directions and find global minima; this is known as the curse of dimensionality. Recent research works have sought to overcome this challenge, e.g., by using a randomized optimizer to search the global minima [22], and using the Wasserstein metric to improve the convexity [30, 44].

Stochastic gradient descent (SGD), frequently used in training deep neural networks, has proved to be efficient in solving high-dimensional optimization problems. Note that a common gradient descent needs to accurately compute gradients in high dimensions at each iteration, which can be computationally prohibitive and leave the gradients stuck in bad local minima, while SGD is better at overcoming these issues as illustrated, for example, by the Adam method [21]. Motivated by the success of SGD, we propose solving the high-dimensional optimization problem in seismic tomography by constructing descent directions of velocity models using the random batch method (RBM) recently proposed for computing dynamics of interacting particles [17].

To use RBM in seismic tomography, possible choices for computing synthetic seismograms are numerical methods of particle type, e.g., generalized ray theory [13, 41], Kirchhoff migration [9, 20], Gaussian beam migration [15, 16, 27, 10, 8, 28], the Gaussian beam method [29, 2], and frozen Gaussian approximation (FGA) [25, 5, 6, 11, 4]. One may also use direct numerical methods (e.g., the spectral element method) to compute synthetic seismograms if the random batch is applied to source-receiver pairs [40]. Here, for the sake of convenience we use FGA to compute wave equations. FGA was originally used in quantum chemistry for the Schrödinger equation [12, 14], with systematic justifications given in [18, 19, 35]. Then the formula was generalized to linear hyperbolic systems [25, 26], with applications in seismic tomography [43, 22, 5, 6, 11]. FGA does not need to solve ray paths by shooting to reach the receivers, and it can provide accurate solutions in the presence of caustics and multipathing, with no requirement on tuning beam width parameters to achieve a good resolution [3, 15, 7, 29, 25].

In this paper, we focus on seismic tomography based on acoustic wave propagation (P wave). We shall study both wave-equation-based travel-time inversion (TTI) and full-waveform inversion (FWI). Specifically, we compute the wave equations by FGA and construct the sensitivity kernel by RBM, yielding the stochastic decent directions for the velocity model. Then the convergent iterations in TTI and FWI will produce the velocity model expected in seismic tomography. We analyze the convergence of the proposed method in the mean-square sense and show the accuracy by a two-dimensional (2D) Gaussian perturbation model, a 2D gradually changing background model, a 2D three-layered model, and a 3D Gaussian perturbation model.

The rest of the paper is organized as follows. In section 2, we introduce the model setup and the formulation of seismic tomography. In section 3, we systematically describe the construction of stochastic gradients by FGA and RBM, and then prove the convergence to the deterministic gradient descent in the mean-square sense. In addition, we generalize the idea of and provide convergence analysis for the accelerated FWI method [40], which performs the random batch on source-receiver pairs. We present its numerical performance with several examples in section 4, and we make conclusive remarks in section 5.

2. Seismic tomography.

In this section, we introduce the formulation of seismic tomography, where the propagation of seismic waves is modeled by the acoustic wave equations (wave),

$$
\begin{aligned}
&\rho(x)\,\partial_t^2 u - \Delta_x u = s(t,x), \qquad x\in\mathbb{R}^3,\ t>0,\\
&u(t=0,x)=0, \quad \partial_t u(t=0,x)=0.
\end{aligned}
\tag{1}
$$

Here $\rho(x)$ is the reference media density at location $x \in \mathbb{R}^3$, from which one can get the P-wave velocity $c = 1/\sqrt{\rho}$, Δ_x is the Laplace operator in x, and $s(t,x)$ is the source term. When an earthquake is modeled by a point source, one can choose $s(t,x) = f(t)\delta_d(x - x_s)$, with $f(t)$ as the source time function at x_s with compact support on $[0,\infty)$, and δ_d as the Dirac delta function. Note that the formulation here can be easily generalized to elastic wave propagation as in, e.g., [39, 11], and we focus on seismic tomography using the P-wave for the sake of simplicity.

Seismic tomography aims to solve an inverse problem which minimizes the misfit functional

$$
\mathrm{J}(\rho) := \frac{1}{2}\int_0^T \mathrm{d}t \int_\Omega \mathrm{d}x\, w(x)\left(\mathcal{A}[u_{\mathrm{obs}}(t,x)] - \mathcal{A}[u(t,x;\rho)]\right)^2,
\tag{2}
$$

where $[0,T]$ is a fixed time window, Ω is the region of interest, $w(x)$ represents the distribution of receivers (e.g., in a finite-receiver setup, $w(x) = \frac{1}{N_\mathrm{R}}\sum_{r=1}^{N_\mathrm{R}}\delta_d(x - x_r)$, where N_R is the number of receivers and x_r is the location of the rth receiver station), $u_{\mathrm{obs}}(t,x)$ is the observed signal, and $u(t,x)$ is the synthetic signal satisfying the forward propagating wave equation (1). We use $\mathcal{A}$ to denote an observation operator to extract useful information from the signals, e.g., in the FWI, $\mathcal{A}[u] = u$ means that all the information contained in a signal u is used; in the TTI, $\mathcal{A}[u]$ is the time spent by a signal u generated from a earthquake location x_s propagated to a seismic receiver x_r.

In order to solve the minimization problem, one needs to compute the Fréchet derivative of the misfit functional $\delta\mathrm{J}/\delta\rho$. Without loss of generality, we take FWI as an example and compute (cf. [6])

$$
\begin{aligned}
\delta\mathrm{J} &= -\int_0^T \mathrm{d}t \int_\Omega \mathrm{d}x\, w(x)\left(u_{\mathrm{obs}}(t,x) - u(t,x)\right)\delta u(t,x)\\
&= \int_0^T \mathrm{d}t \int_\Omega \mathrm{d}x \int_0^t \mathrm{d}\tau \int_\Omega \mathrm{d}y\, w(x)\,[u_{\mathrm{obs}} - u](t,x)\, G(t,x;\tau,y)\partial_t^2 u(\tau,y)\,\delta\rho(y)\\
&= \int_\Omega \mathrm{d}x \int_\Omega \mathrm{d}y \int_0^T \mathrm{d}\tau \int_0^{T-\tau} \mathrm{d}t\, w(x)\,[u_{\mathrm{obs}} - u](T-t,x)\, G(T-t,x;\tau,y)\,\partial_t^2 u(\tau,y)\,\delta\rho(y)\\
&= \int_\Omega \mathrm{d}y \int_0^T \mathrm{d}\tau \int_0^{T-\tau} \mathrm{d}t \int_\Omega \mathrm{d}x\, w(x)\,[u_{\mathrm{obs}} - u](T-t,x)\, G(T-\tau,y;t,x)\,\partial_t^2 u(\tau,y)\,\delta\rho(y)\\
&= \int \mathrm{d}y \int_0^T \mathrm{d}\tau\, \delta\rho(y)\, u^\dagger(T-\tau,y)\,\partial_t^2 u(\tau,y)\,,
\end{aligned}
$$

where the Green's function $G = G(t,x;\tau,y)$ solves

$$
\rho(x)\,\partial_t^2 G - \Delta_x G = \delta_d(t-\tau, x-y),
$$

and $u^\dagger$ solves the adjoint wave equations

$$
\begin{aligned}
&\rho(x)\,\partial_t^2 u^\dagger - \Delta_x u^\dagger = s^\dagger(t,x), \qquad x\in\mathbb{R}^3,\ t>0,\\
&u^\dagger(t=0,x)=0, \quad \partial_t u^\dagger(t=0,x)=0,
\end{aligned}
\tag{3}
$$

with the adjoint source function

$$
s^\dagger(t,x) = w(x)\,[u_{\mathrm{obs}} - u](T-t,x).
\tag{4}
$$

Define the sensitivity kernel

$$K(x;\rho) := \int_0^T \mathrm{d}t\, \rho(x)\, u^\dagger(T-t,x)\, \partial_t^2 u(t,x) = \int_0^T \mathrm{d}t\, \rho(x)\, \partial_t u^\dagger(T-t,x)\, \partial_t u(t,x); \tag{5}$$

then one gets

$$\delta \mathrm{J} = \int_\Omega \mathrm{d}x\, K(x;\rho)\, \delta \log \rho(x). \tag{6}$$

Remark 2.1. The above computation can also be performed by wave-equation-based TTI, yielding a similar formulations except that the adjoint source function becomes (cf. [6])

$$s^\dagger(t,x) = w(x) g(t) \partial_t u(t,x),$$

where $g(t)$ is a window function supported in $[0, T]$.

After computing the sensitivity kernel (5) and the Fréchet derivative of the misfit functional (6), one can apply optimization methods to find the minimizer of (2), producing the desired velocity model by the relation $c = 1/\sqrt{\rho}$. Classical methods include, but are not limited to, the gradient descent, conjugate gradient, Newtonian, and quasi-Newtonian methods. In this paper, we choose the gradient descent method for its simplicity, bringing convenience to the derivation and proofs. We remark that the idea of reconstructing the gradient using RBM can be used in essentially the same way as in other kinds of gradient-based optimization methods, at least for the purpose of numerical computing.

The gradient descent method can be formulated by

$$\frac{\mathrm{dX}}{\mathrm{d}s} = -\nabla_{\mathrm{X}} \mathrm{J}, \tag{7}$$

where $\mathrm{X} := \log \rho$ and $\nabla_{\mathrm{X}} \mathrm{J} := K(x;\rho)$.

Given $\rho > \rho_0 > 0$, the map $\rho \to \mathrm{X}$ is one-to-one. Therefore, we shall use X, Y, $\tilde{\mathrm{X}}$ to denote density (or velocity) models throughout this paper. Note that X is a function of spatial variable $x \in \mathbb{R}^3$ and the iteration index $s \in \mathbb{R}^+$. Let us define $|\mathrm{X}(s)|^q := \int_\Omega |\mathrm{X}(x,s)|^q \mathrm{d}x$ and $\|\mathrm{X}(s)\| := (\mathbb{E}|\mathrm{X}(s)|^2)^{1/2}$. For the remainder of the paper, we shall not write the dependence of X on the spatial variable x explicitly but use $\mathrm{X} = \mathrm{X}(s)$ to put more focus on the iteration procedure. We also write $K(\mathrm{X}) = K(x;\rho)$.

To make the inverse problem well-posed, a regularization term is added to the misfit functional (2), and the above gradient flow is modified by

$$\frac{\mathrm{dX}}{\mathrm{d}s} = -K(\mathrm{X}) - \nabla_{\mathrm{X}} V(\mathrm{X}), \tag{8}$$

where V is a given regularization potential that does not rely on solving the wave equations, and we assume that V is strongly convex in X so that $V(\mathrm{X}) - \frac{r}{2}\mathrm{X}^2$ is convex for some $r > 0$, and $\nabla_{\mathrm{X}} V$, $\nabla_{\mathrm{X}}^2 V$ have polynomial growth. We remark here that the assumptions on V are only for technical use in order to prove the convergence in the following section. In practice, these assumptions may be removed. In section 4 it will be seen that all of the numerical examples simply take $V \equiv 0$, and we can still get numerical convergence results.

3. Frozen Gaussian approximation and random batch method.

In this section, we systematically introduce the construction of the stochastic gradient by FGA and RBM. We first describe how to use FGA to construct the gradient, and then we use RBM to construct the stochastic gradient.

3.1. FGA-based gradient construction.

The sensitivity kernel K defined in (5) is a cross-correlation of the forward and adjoint wavefields. For convenience of latter discussion, we write $K = K(\mathrm{X})$ to indicate dependence on the velocity model (noting that both u and $u^\dagger$ depend on X since they are synthetic solutions of (1) and (3)).

FGA approximates the wavefields by (cf. [5])

$$u(t,x;\mathrm{X}) = \frac{1}{N}\sum_{j=1}^{N} G_j(t,x;\mathrm{X}) \quad \text{and} \quad u^\dagger(t,x;\mathrm{X}) = \frac{1}{N}\sum_{j=1}^{N} G_j^\dagger(t,x;\mathrm{X}), \tag{9}$$

where G_j and $G_j^\dagger$ are Gaussian functions in the form of, e.g.,

$$A\exp\left(\frac{\mathrm{i}}{\varepsilon}P\cdot(x-Q) - \frac{1}{2\varepsilon}|x-Q|^2\right),$$

where A, Q, and P are functions of (t,q,p) determined by a set of ordinary differential equations (ODEs),

$$\begin{cases} \dfrac{\mathrm{d}Q}{\mathrm{d}t} = \partial_P H, & Q(0,q,p) = q, \\ \dfrac{\mathrm{d}P}{\mathrm{d}t} = -\partial_Q H, & P(0,q,p) = p, \\ \dfrac{\mathrm{d}A}{\mathrm{d}t} = A\,\dfrac{\partial_P H\cdot\partial_Q H}{H} + \dfrac{A}{2}\,\mathrm{Tr}\left(Z^{-1}\dfrac{\mathrm{d}Z}{\mathrm{d}t}\right), & A(0,q,p) = A_0(q,p), \end{cases} \tag{10}$$

with $H(Q,P) = \pm c(Q)|P|$ and the shorthand notation $\partial_z = \partial_q - \mathrm{i}\partial_p$ and $Z = \partial_z(Q + \mathrm{i}P)$; see more details in [5, 6, 11]. In (9) we assume the same beam number N for both forward and adjoint simulations, and N does not change during the iteration procedure (8). For convenience of notation and latter discussion, we introduce

$$\dot{G}_j(t,x;\mathrm{X}) := \partial_t G_j(t,x;\mathrm{X}), \quad \text{and} \quad \dot{G}_j^\dagger(t,x;\mathrm{X}) := \rho(x)\,\partial_t G_j^\dagger(t,x;\mathrm{X}). \tag{11}$$

Then we approximate the kernel

$$K(\mathrm{X}) = \sum_{k=1}^{T/\tau} \tau\, u_k^\dagger\, u_k \tag{12}$$

where $u_k = \partial_t u(t_k,x)$ and $u_k^\dagger = \rho(x)\,\partial_t u^\dagger(T-t_k,x)$, $t_k = k\tau$, $k = 1,2,\dots,T/\tau$. The velocity model follows

$$\frac{\mathrm{dX}}{\mathrm{d}s} = -\nabla_{\mathrm{X}} V(\mathrm{X}) - K(\mathrm{X}) \tag{13}$$

at each iteration step $s \in [s_{m-1}, s_m)$.

3.2. Stochastic gradient by random batch method.

At each iteration step $s_m = mh$ and each time $t_k = k\tau$, we randomly choose index sets $\mathfrak{B}_{m,k}$, $\mathfrak{B}_{m,k}^\dagger \subset \{1,2,\dots,N\}$ such that, first, the number of indices $|\mathfrak{B}_{m,k}| = |\mathfrak{B}_{m,k}^\dagger| = p \ll N$; second, all $\mathfrak{B}_{m,k}$, $\mathfrak{B}^\dagger_{m,k}$ for $m = 0, 1, 2, \ldots$ and $k = 1, 2, \ldots, T/\tau$ are independent and identically distributed, and the probability $\mathbb{P}(j \in \mathfrak{B}_{m,k}) = \mathbb{P}(j \in \mathfrak{B}^\dagger_{m,k}) = p/N$ for $1 \le j \le N$. We call such $\mathfrak{B}_{m,k}$ and $\mathfrak{B}^\dagger_{m,k}$ *random batches*.

We first approximate the wavefields by RBM

$$\tilde{u}(t, x; \mathrm{X}) = \frac{1}{p} \sum_{j \in \mathfrak{B}_{m,k}} G_j(t, x; \mathrm{X}) \quad \text{and} \quad \tilde{u}^\dagger(t, x; \mathrm{X}) = \frac{1}{p} \sum_{j \in \mathfrak{B}^\dagger_{m,k}} G^\dagger_j(t, x; \mathrm{X}). \tag{14}$$

Then we define the stochastic gradient as

$$\tilde{K}(\mathrm{X}) = \sum_{k=1}^{T/\tau} \tau \, \tilde{u}^\dagger_k \, \tilde{u}_k, \tag{15}$$

where $\tilde{u}_k = \partial_t \tilde{u}(t_k, x)$ and $\tilde{u}^\dagger_k = \rho(x) \, \partial_t \tilde{u}^\dagger(T - t_k, x)$.

We call $\tilde{K}(\mathrm{X})$ the random batch kernel and update the velocity model by

$$\frac{\mathrm{d}\mathrm{X}}{\mathrm{d}s} = -\nabla_{\mathrm{X}} V(\mathrm{X}) - \tilde{K}(\mathrm{X}) \tag{16}$$

for each iteration step $s \in [s_{m-1}, s_m)$.

Remark 3.1. The main idea here is to use random batch summation to construct randomized wavefields and the corresponding sensitivity kernels. It does not rely on using FGA, and one can replace it by any other efficient wave propagation solver, e.g., Gaussian beam method or spectral element method.

Remark 3.2. As discussed in [5], the ODEs (10) can, embarrassingly, be parallelly computed, but the parallelization for computing the summation (9) is less efficient and technically involved. Therefore, the random batch summation (14) can significantly reduce the workload of reconstructing wavefields and further improve the efficiency of the FGA computation.

Remark 3.3. To get a well-resolved wavefield, the number of beams N is typically on the order of $\varepsilon^{-d/2}$, where ε is proportional to the wavelength, and $d = 2, 3$ is the dimensionality of the space; thus N is usually a large number when performing a high-frequency simulation where $\varepsilon \ll 1$. Recently, a random sampling method [42] was developed for the Schrödinger equation which can reduce the number of beams significantly by choosing beams via a preliminary distribution. This sampling idea is restricted to Gaussian or WKB initial data for which one can determine whether a beam is more or less "important." However, for wave equations with Dirac delta sources, the beams are almost uniformly distributed, and thus it is not yet clear how to find a good sampling strategy.

As in Algorithm 1, we present a brief pseudocode for implementation of the procedure of the proposed stochastic gradient descent by random batch method. We remark that if, in lines 16 and 22, we set both $\mathfrak{B}_{m,k}$ and $\mathfrak{B}^\dagger_{m,k}$ as the whole index set $\{1, 2, \ldots, N\}$, then the algorithm will recover the classical gradient descent method.

3.3. Preliminary results.

In this subsection, we prove a few lemmas as preparation for proving the convergence theorem in the next subsection.

First, let us state the Lipschitz continuity properties of K and $\tilde{K}$.

Proposition 3.1. *The kernels defined in* (12) *and* (15) *are Lipschitz continuous in* X.

Algorithm 1. Stochastic gradient descent by random batch method.

```
procedure MAIN LOOP
    $\mathrm{X} = \mathrm{X}_0$, $m = 0$
    Compute $\mathrm{J}(\mathrm{X})$
    while $m < M^*$ do ▷ $M^*$ is a given number indicating the maximum iteration steps
        $\tilde{K} \leftarrow$ RANDOM BATCH KERNEL ▷ Compute the gradient by random batch method
        $\mathrm{X} \leftarrow \mathrm{X} - \alpha(\tilde{K} + \nabla_{\mathrm{X}} V)$, $m \leftarrow m + 1$
        Compute $\mathrm{J}(\mathrm{X})$
        if $(\mathrm{J}(\mathrm{X}) < \mathrm{J}^*)$ exit ▷ If the misfit smaller is than a given threshold $J^*$, exit iteration loop
    end while
end procedure
procedure RANDOM BATCH KERNEL($m$)
    Given discrete source and receiver locations $x_s$ and $x_r$
    Initialize $G_j(t, x; \mathrm{X})$ at $t = 0$ for all $j = 1, \dots, N$
    for $k = 1 : T/\tau$ do ▷ TIME EVOLUTION FOR FORWARD SIMULATION
        $G_j(t, x; \mathrm{X}) \leftarrow G_j(t + \tau)$ for all $j = 1, \dots, N$ ▷ Evolve FGA ODEs for one time step
        Generate independent random batch $\mathfrak{B}_{m,k}$
        Compute $\tilde{u}(t, x; \mathrm{X})$ by(14); Compute $u(t, x_r; \mathrm{X})$ by (9)
    end for
    Initialize $G_j^{\dagger}(t, x; \mathrm{X})$ at $t = 0$ for all $j = 1, \dots, N$
    for $k = 1 : T/\tau$ do ▷ TIME EVOLUTION FOR ADJOINT SIMULATION
        $G_j^{\dagger}(t, x; \mathrm{X}) \leftarrow G_j^{\dagger}(t + \tau)$ for all $j = 1, \dots, N$ ▷ Evolve FGA ODEs for one time step
        Generate independent random batch $\mathfrak{B}_{m,k}^{\dagger}$
        Compute $\tilde{u}^{\dagger}(t, x; \mathrm{X})$ by equation (14)
    end for
    Compute $\tilde{K}$ by equation (15)
end procedure return $\tilde{K}$ and $u(\cdot, x_r; \mathrm{X})$
```

Proof. As defined in equations (9)–(12), the kernel K can be seen as a summation of $\dot{G}_j \dot{G}_l^{\dagger}$'s. Each G_j or $G_l^{\dagger}$ is a Gaussian function whose parameters are given by a set of ODEs (10). By the smooth dependence on the initial condition and parameters for solution of ODEs, one can deduce that $\dot{G}_j$ and $\dot{G}_l^{\dagger}$ are smooth in X, and thus K is Lipschitz continuous in X. Similarly, $\tilde{K}$ is Lipschitz continuous in X. ∎

Note that equation (16) can be rewritten as

$$\frac{\mathrm{d}\mathrm{X}}{\mathrm{d}s} = -\nabla_{\mathrm{X}} V(\mathrm{X}) - K(\mathrm{X}(s)) - \chi_m(\mathrm{X}(s)), \tag{17}$$

where

$$\chi_m(\mathrm{X}) := \tilde{K}(\mathrm{X}) - K(\mathrm{X}). \tag{18}$$

Thus to analyze the convergence of the RBM, the key is a precise estimate on χ_m, for which we have the following lemma.

Lemma 3.1. *Let* Y *be a velocity model, fixed and determined, and let* $K(\mathrm{Y})$ *and* $\tilde K(\mathrm{Y})$ *be defined as in* (12) *and* (15), *respectively. Then*

$$\mathbb{E}[\chi_m(\mathrm{Y})] = 0, \tag{19}$$

$$\mathbb{E}[\chi_m^2(\mathrm{Y})] = \left(\frac{1}{p} - \frac{1}{N}\right)\tau\Lambda, \tag{20}$$

where

$$\Lambda = \frac{\tau}{N-1}\sum_{k=1}^{T/\tau}\sum_{j=1}^{N}\left[\left(\dot G_j - \frac{1}{N}\sum_{l=1}^{N}\dot G_l\right)^2 \mathbb{E}(\tilde u_k^\dagger)^2 + u_k^2\left(\dot G_j^\dagger - \frac{1}{N}\sum_{l=1}^{N}\dot G_l^\dagger\right)^2\right]. \tag{21}$$

Proof. The expectation (19) is straightforward, and we only prove the variance equality (20). Noting that $\tilde u_k$, $\tilde u_j^\dagger$ for $k, j = 1, \dots, T/\tau$ are independent, we then have

(22)

$$\begin{aligned}\mathbb{E}\chi_m^2 = \tau^2\sum_{k=1}^{T/\tau}\mathbb{E}\left[(u_k u_k^\dagger - \tilde u_k \tilde u_k^\dagger)^2\right] &= \tau^2\sum_{k=1}^{T/\tau}\left(\mathbb{E}(\tilde u_k)^2\,\mathbb{E}(\tilde u_k^\dagger)^2 - u_k^2(u_k^\dagger)^2\right)\\ &= \tau^2\sum_{k=1}^{T/\tau}\Big(\big(\mathbb{E}(\tilde u_k)^2 - u_k^2\big)\,\mathbb{E}(\tilde u_k^\dagger)^2 + u_k^2\left(\mathbb{E}(\tilde u_k^\dagger)^2 - (u_k^\dagger)^2\right)\Big),\end{aligned}$$

where u_k, $\tilde u_k$, $u_k^\dagger$, and $\tilde u_k^\dagger$ take the form of

$$u_k = \frac{1}{N}\sum_{j=1}^{N}\dot G_j, \quad \tilde u_k = \frac{1}{p}\sum_{j\in\mathfrak{B}_{m,k}}\dot G_j, \quad u_k^\dagger = \frac{1}{N}\sum_{j=1}^{N}\dot G_j^\dagger, \quad \tilde u_k^\dagger = \frac{1}{p}\sum_{j\in\mathfrak{B}^\dagger_{m,k}}\dot G_j^\dagger.$$

Then one can compute

$$\begin{aligned}\mathbb{E}(\tilde u_k)^2 &= \frac{1}{p^2}\sum_{j=1}^{N}\dot G_j^2\,\mathbb{P}(j\in\mathfrak{B}_{m,k}) + \frac{1}{p^2}\sum_{j,l:j\neq l}\dot G_j\,\dot G_l\,\mathbb{P}\left(j\in\mathfrak{B}_{m,k}\text{ and } l\in\mathfrak{B}_{m,k}\right)\\ &= \frac{1}{pN}\sum_{j=1}^{N}\dot G_j^2 + \frac{p-1}{pN(N-1)}\sum_{j,l:j\neq l}\dot G_j\,\dot G_l,\end{aligned}$$

and thus

$$\begin{aligned}\mathbb{E}(\tilde u_k)^2 - u_k^2 &= \left(\frac{1}{pN} - \frac{1}{N^2}\right)\sum_{j=1}^{N}\dot G_j^2 + \left(\frac{p-1}{pN(N-1)} - \frac{1}{N^2}\right)\sum_{j,l:j\neq l}\dot G_j\,\dot G_l\\ &= \left(\frac{1}{p} - \frac{1}{N}\right)\left(\frac{1}{N}\sum_{j=1}^{N}\dot G_j^2 - \frac{1}{N(N-1)}\sum_{j,l:j\neq l}\dot G_j\,\dot G_l\right)\\ &= \left(\frac{1}{p} - \frac{1}{N}\right)\frac{1}{N-1}\sum_{j=1}^{N}\left(\dot G_j - \frac{1}{N}\sum_{l=1}^{N}\dot G_l\right)^2.\end{aligned}$$

The $\mathbb{E}(\tilde u_k^\dagger)^2 - (u_k^\dagger)^2$ can be compute in an analogous way, and then we can obtain (20). ∎

Let $\mathrm{X}(s)$ be a solution to (13) and $\tilde{\mathrm{X}}(s)$ be solution to (16). Define $Z(s) := \tilde{\mathrm{X}}(s) - \mathrm{X}(s)$, and let $\mathcal{F}_{m-1}$ be a σ-algebra generated by the random batch construction for $s \le s_{m-1}$. The following lemmas are devoted to the stability and truncation error analysis of the random batch method.

LEMMA 3.2. *One can have the following estimate for* X *and* $\tilde{\mathrm{X}}$*:*

$$\sup_{t>0}\left(|\mathrm{X}|^q + \mathbb{E}|\tilde{\mathrm{X}}|^q\right) \le C_q. \tag{23}$$

Proof.

$$\frac{\mathrm{d}|\mathrm{X}|^q}{\mathrm{d}s} = |\mathrm{X}|^{q-2}\mathrm{X}\cdot\frac{\mathrm{dX}}{\mathrm{d}s} = -|\mathrm{X}|^{q-2}\mathrm{X}\cdot(\nabla_{\mathrm{X}}V(\mathrm{X}) + K(\mathrm{X}))$$

$$\begin{aligned}\mathrm{X}\cdot\nabla_{\mathrm{X}}V(\mathrm{X}) &= (\mathrm{X}-0)\cdot(\nabla_{\mathrm{X}}V(\mathrm{X}) - \nabla_{\mathrm{X}}V(0)) + \mathrm{X}\cdot\nabla_{\mathrm{X}}V(0)\\ &= (\mathrm{X}-0)^2 : \nabla^2_{\mathrm{X}}V(\mathrm{X}^*) + \mathrm{X}\cdot\nabla_{\mathrm{X}}V(0),\end{aligned}$$

$$\frac{\mathrm{d}|\mathrm{X}|^q}{\mathrm{d}s} \le -qr|\mathrm{X}|^q + \|K\|_\infty|\mathrm{X}|^{q-1} \le -qr|\mathrm{X}|^q + \|K\|_\infty\left(\frac{q-1}{q}\nu|\mathrm{X}|^q + \frac{1}{q\nu^{q-1}}\right).$$

Thus $|\mathrm{X}|^q \le C_q$. Similarly, $\mathbb{E}|\tilde{\mathrm{X}}|^q \le C_q$. ∎

LEMMA 3.3. *For* $s \in [s_{m-1}, s)$*, it holds that*

$$\left\|\tilde{\mathrm{X}}(s) - \tilde{\mathrm{X}}(s_{m-1})\right\| \le Ch. \tag{24}$$

Proof. Direct computation shows that

$$\frac{\mathrm{d}}{\mathrm{d}s}\left\|\tilde{\mathrm{X}}(s) - \tilde{\mathrm{X}}(s_{m-1})\right\|^2 = -2\,\mathbb{E}\left[\left(\tilde{\mathrm{X}}(s) - \tilde{\mathrm{X}}(s_{m-1})\right)\left(\nabla_{\mathrm{X}}V\left(\tilde{\mathrm{X}}(s)\right) + \tilde{K}\left(\tilde{\mathrm{X}}(s)\right)\right)\right];$$

then by Hölder's inequality, one has

$$\frac{\mathrm{d}}{\mathrm{d}s}\left\|\tilde{\mathrm{X}}(s) - \tilde{\mathrm{X}}(s_{m-1})\right\|^2 \le C\left\|\tilde{\mathrm{X}}(s) - \tilde{\mathrm{X}}(s_{m-1})\right\|\left(\left\|\nabla_{\mathrm{X}}V\left(\tilde{\mathrm{X}}(s)\right)\right\| + \left\|\tilde{K}\left(\tilde{\mathrm{X}}(s)\right)\right\|\right).$$

Note that $\nabla_{\mathrm{X}}V \le C(1+|\mathrm{X}|^q)$ for some q, and thus $\|\nabla_{\mathrm{X}}V\|$ is bounded, and $\tilde{K}$ is a cross-correlation of two wavefields constructed from Gaussians, where the Gaussians are determined by a set of ODEs depending on the velocity model $\tilde{\mathrm{X}}$ smoothly, so Gaussians are bounded and so is $\|\tilde{K}\|$. Thus

$$\frac{\mathrm{d}}{\mathrm{d}s}\left\|\tilde{\mathrm{X}}(s) - \tilde{\mathrm{X}}(s_{m-1})\right\|^2 \le C\left\|\tilde{\mathrm{X}}(s) - \tilde{\mathrm{X}}(s_{m-1})\right\|,$$

and then the estimate (24) follows. ∎

LEMMA 3.4. *For* $s \in [s_{m-1}, s)$*,*

$$\|Z(s) - Z(s_{m-1})\| \le Ch \tag{25}$$

and

$$\mathbb{E}\left|(Z(s) - Z(s_{m-1}))\,\chi_m\left(\mathrm{X}(s)\right)\right| \le Ch\left[\left(\|Z(s)\| + \|Z(s)\|^2\right) + h\right] + \frac{h\tau}{p}\|\Lambda\|_\infty. \tag{26}$$

Proof.

$$\frac{\mathrm{d}Z}{\mathrm{d}s} = -\nabla_{\mathrm{X}} V(\tilde{\mathrm{X}}) + \nabla_{\mathrm{X}} V(\mathrm{X}) - \tilde{K}(\tilde{\mathrm{X}}) + K(\mathrm{X});$$

thus

$$\frac{1}{2}\frac{\mathrm{d}Z^2}{\mathrm{d}s} \le -(r-L)Z^2,$$

which implies that, for $s \in [s_{m-1}, s_m)$, one has

$$|Z(s)| \le |Z(s_{m-1})| + Ch \quad \text{and} \quad \|Z(s) - Z(s_{m-1})\| \le Ch$$
$$-\frac{\mathrm{d}Z}{\mathrm{d}s} = \nabla_{\mathrm{X}} V(\tilde{\mathrm{X}}) - \nabla_{\mathrm{X}} V(\mathrm{X}) + \tilde{K}(\tilde{\mathrm{X}}) - \tilde{K}(\mathrm{X}) + \chi_m(\mathrm{X}).$$

Since

$$\left|\nabla_{\mathrm{X}} V(\tilde{\mathrm{X}}) - \nabla_{\mathrm{X}} V(\mathrm{X})\right| \le \left|(\tilde{\mathrm{X}} - \mathrm{X}) \cdot \nabla^2_{\mathrm{X}} V(\mathrm{X}^*)\right|,$$

we have

$$\begin{aligned}
&\mathbb{E}\left|\left(\nabla_{\mathrm{X}} V(\tilde{\mathrm{X}}(s')) - \nabla_{\mathrm{X}} V(\mathrm{X}(s'))\right)\chi_m(\mathrm{X}(s))\right| \\
&\le \|\chi_m(\mathrm{X}(s))\|_\infty \left\|(\tilde{\mathrm{X}}(s') - \mathrm{X}(s'))\right\| \left(\mathbb{E}\left[|\tilde{\mathrm{X}}(s')|^{q_1} + |\mathrm{X}(s')|^{q_1}\right]^2\right)^{1/2} \le C\|Z(s')\|.
\end{aligned}$$

Therefore

$$\begin{aligned}
&\mathbb{E}\left|(Z(s) - Z(s_{m-1}))\,\chi_m\left(\mathrm{X}(s)\right)\right| \\
&\le \int_{s_{m-1}}^{s} \mathrm{d}s' \Big\{ C\left\|Z(s')\right\| + \mathbb{E}\left|\chi_m\left(\mathrm{X}(s')\right)\chi_m\left(\mathrm{X}(s)\right)\right| \\
&\quad + \mathbb{E}\left|\left(\tilde{K}(\tilde{\mathrm{X}}(s')) - \tilde{K}(\mathrm{X}(s'))\right)\chi_m\left(\mathrm{X}(s)\right)\right| \Big\} \\
&\le \int_{s_{m-1}}^{s} \mathrm{d}s' \Big\{ C\left\|Z(s')\right\| + \mathbb{E}\left[\chi_m\left(\mathrm{X}(s)\right)^2\right] + \frac{1}{2}\mathbb{E}\left[\chi_m\left(\mathrm{X}(s')\right)^2\right] \\
&\quad + \frac{1}{2}\mathbb{E}\left[\left(\tilde{K}(\tilde{\mathrm{X}}(s')) - \tilde{K}(\mathrm{X}(s'))\right)^2\right] \Big\}.
\end{aligned}$$

The second and third terms are controlled by Lemma 3.1 since X is independent of the random batch, and thus

$$\mathbb{E}\left[\chi_m\left(\mathrm{X}(s)\right)^2 + \chi_m\left(\mathrm{X}(s)\right)^2\right] \le C\left(\frac{1}{p} - \frac{1}{N}\right)\tau\|\Lambda\|_\infty.$$

The fourth term is controlled by using the Lipschitz continuity of $\tilde{K}$ as follows:

$$\mathbb{E}\left[\left(\tilde{K}(\tilde{\mathrm{X}}) - \tilde{K}(\mathrm{X})\right)^2\right] \le L^2\,\mathbb{E}\left[\left(\tilde{\mathrm{X}} - \mathrm{X}\right)^2\right] = L^2\|Z\|^2.$$

Then

$$\begin{aligned}
\mathbb{E}\left|(Z(s) - Z(s_{m-1}))\,\chi_m\left(\mathrm{X}(s)\right)\right| &\le C\left[\left(\|Z(s)\| + \|Z(s)\|^2\right)h + h^2\right] \\
&\quad + \left(\frac{1}{p} - \frac{1}{N}\right)h\tau\|\Lambda\|_\infty.
\end{aligned}$$

□

3.4. Main theorem. In this subsection, we present the main convergence theorem. As in Remark 3.1, the main idea of the proposed method is to use RBM for the construction of randomized wavefields and the corresponding sensitivity kernels, and one can replace FGA by any other efficient wave propagation solvers. Therefore, we shall only focus on the convergence of RBM to the deterministic gradient decent method, assuming that the chosen wave propagation solvers can provide convergent numerical results. We refer the reader to [26, 4] for the convergent results of the FGA solvers.

THEOREM 3.1. *As the iteration step* h *goes to zero,* $\tilde{\mathrm{X}}$ *converges to* X *in the mean-square sense. More precisely, we have the following estimate:*

$$\sup_{s\geq 0}\|Z(s)\| \leq C\sqrt{\frac{h\tau}{p}+Ch^2}. \tag{27}$$

Proof.

$$\begin{aligned}\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}s}\mathbb{E}Z^2 =& -\mathbb{E}\left[Z(s)\left(\nabla_{\mathrm{X}}V(\tilde{\mathrm{X}}(s)) - \nabla_{\mathrm{X}}V\left(\mathrm{X}(s)\right) + K(\tilde{\mathrm{X}}(s)) - K\left(\mathrm{X}(s)\right)\right)\right]\\ &-\mathbb{E}\left[Z(s)\,\chi_m\left(\tilde{\mathrm{X}}(s)\right)\right]\\ \leq& -(r-L)\mathbb{E}Z^2 - \mathbb{E}\left[Z(s)\,\chi_m\left(\tilde{\mathrm{X}}(s)\right)\right].\end{aligned}$$

Let

$$R(s) := \mathbb{E}\left[Z(s)\,\chi_m\left(\tilde{\mathrm{X}}(s)\right)\right],$$

$$\begin{aligned}R(s) =&\ \mathbb{E}\left[Z(s_{m-1})\,\chi_m\left(\tilde{\mathrm{X}}(s_{m-1})\right)\right]\\ &+\mathbb{E}\left[Z(s_{m-1})\left(\chi_m\left(\tilde{\mathrm{X}}(s)\right) - \chi_m\left(\tilde{\mathrm{X}}(s_{m-1})\right)\right)\right]\\ &+\mathbb{E}\left[\left(Z(s) - Z(s_{m-1})\right)\chi_m\left(\mathrm{X}(s)\right)\right]\\ &+\mathbb{E}\left[\left(Z(s) - Z(s_{m-1})\right)\left(\chi_m\left(\tilde{\mathrm{X}}(s)\right) - \chi_m\left(\mathrm{X}(s)\right)\right)\right]\\ =:&\ I_1 + I_2 + I_3 + I_4.\end{aligned}$$

For the first term,

$$\begin{aligned}I_1 &= \mathbb{E}\left[\mathbb{E}\left[Z(s_{m-1})\,\chi_m\left(\tilde{\mathrm{X}}(s_{m-1})\right)\Big|\,\mathcal{F}_{m-1}\right]\right]\\ &= \mathbb{E}\left[Z(s_{m-1})\,\mathbb{E}\left[\chi_m\left(\tilde{\mathrm{X}}(s_{m-1})\right)\Big|\,\mathcal{F}_{m-1}\right]\right] = 0.\end{aligned}$$

For the second term,

$$\begin{aligned}I_2 &= \mathbb{E}\left[Z(s_{m-1})\left(\chi_m\left(\tilde{\mathrm{X}}(s)\right) - \chi_m\left(\tilde{\mathrm{X}}(s_{m-1})\right)\right)\right]\\ &\leq C\left\|Z(s_{m-1})\right\|\left\|\chi_m\left(\tilde{\mathrm{X}}(s)\right) - \chi_m\left(\tilde{\mathrm{X}}(s_{m-1})\right)\right\|\\ &\leq 2LC\left\|Z(s_{m-1})\right\|\left\|\tilde{\mathrm{X}}(s) - \tilde{\mathrm{X}}(s_{m-1})\right\|\\ &\leq C\|Z(s)\|h + Ch^2,\end{aligned}$$

where, for the second inequality, we used the Lipschitz continuity of K and $\tilde{K}$. For the third term, by Lemma 3.4

$$I_3 \leq Ch\left[\left(\|Z(s)\| + \|Z(s)\|^2\right) + h\right] + \frac{h\tau}{p}\|\Lambda\|_\infty.$$

For the fourth term,

$$I_4 \leq \|Z(s) - Z(s_{m-1})\| \left\|\chi_m\left(\tilde{\mathrm{X}}(s)\right) - \chi_m\left(\mathrm{X}(s)\right)\right\| \leq C\|Z(s)\|h.$$

Hence,

$$R(s) \leq C\|Z(s)\|h + C\frac{h\tau}{p} + Ch^2,$$

and

$$\frac{\mathrm{d}}{\mathrm{d}s}\|Z\|^2 \leq -(r-L)\|Z\|^2 + Ch\|Z\| + C\frac{h\tau}{p} + Ch^2,$$

which implies

$$\sup_{s\geq 0}\|Z(s)\|^2 \leq C\frac{h\tau}{p} + Ch^2. \qquad \square$$

3.5. Random batch on source-receiver pairs. In this subsection, as a by-product and by essentially following the same proof strategies as in sections 3.3 and 3.4, we provide convergence results for the accelerated FWI method using dynamic mini-batches as proposed in [40]. This method uses spectral element methods to compute synthetic seismograms and applies random batches on the source-receiver pairs. For convenience, we briefly review the method. Assume there are N_R receiver stations and N_S earthquake events, one can rewrite the sensitivity kernel by

(28)
$$K = \frac{1}{N_\mathrm{R}N_\mathrm{S}}\sum_{r=1}^{N_\mathrm{R}}\sum_{s=1}^{N_\mathrm{S}} K_{rs}, \quad \text{where} \quad K_{rs} = \int_0^T \mathrm{d}t\,\rho(x)\,\partial_t u^\dagger(T-t,x;x_r)\,\partial_t u(t,x;x_s),$$

where one uses the spectral element method to solve $u(\cdot,\cdot;x_s)$ by the wave equation (1), with the source located at $x = x_s$, and $u^\dagger(\cdot,\cdot;x_r)$ by the adjoint wave equation (3), with the adjoint source function

$$s^\dagger(t,x) = [u_\mathrm{obs} - u](T-t,x)\,\delta_d(x-x_r).$$

To apply the RBM, in each iteration step m we choose a receiver index subset $\mathfrak{R}_m \subset \{1,2,\ldots,N_\mathrm{R}\}$ and a source index subset $\mathfrak{S}_m \subset \{1,2,\ldots,N_\mathrm{S}\}$ randomly and independently. The random batch kernel is then given by

$$\tilde{K} = \frac{1}{p_\mathrm{R}p_\mathrm{S}}\sum_{r\in\mathfrak{R}_m}\sum_{s\in\mathfrak{S}_m} K_{rs}. \tag{29}$$

Now one can use this kernel in the main loop of Algorithm 1 to update the velocity model X.

Our contribution here is to give a convergence result for the RBM in source-receiver pairs in analogy to Theorem 3.1. Since the strategy of the proof is essentially the same, we only state the following key lemma.

LEMMA 3.5. *Let* Y *be a velocity model, fixed and determined, and let* $\chi_m(\mathrm{Y}) := \tilde{K}(\mathrm{Y}) - K(\mathrm{Y})$. *Then it holds that* $\mathbb{E}_{\mathrm{S}}[\chi_m(\mathrm{Y})] = \mathbb{E}[\chi_m(\mathrm{Y})] = 0$ *and*

$$\mathbb{E}_{\mathrm{S}}[\chi_m^2(\mathrm{Y})] = \left(\frac{1}{p_{\mathrm{R}}} - \frac{1}{N_{\mathrm{R}}}\right)\Lambda_{\mathrm{R}}, \tag{30}$$

$$\mathbb{E}[\chi_m^2(\mathrm{Y})] = \left(\frac{1}{p_{\mathrm{S}}} - \frac{1}{N_{\mathrm{S}}}\right)\mathbb{E}\Lambda_{\mathrm{S}} + \mathbb{E}_{\mathrm{S}}[\chi_m^2(\mathrm{Y})], \tag{31}$$

where we have used the shorthand notation $\mathbb{E}_{\mathrm{S}}$ *for conditional expectation* $\mathbb{E}_{\mathrm{S}}[\,\cdot\,] := \mathbb{E}[\,\cdot\,|\,\mathfrak{S}_m = 1, \ldots, N_{\mathrm{obs}}]$, *and*

$$\Lambda_{\mathrm{R}} = \frac{1}{N_{\mathrm{R}}-1}\sum_{r=1}^{N_{\mathrm{R}}}\left(K_r - \frac{1}{N_{\mathrm{R}}}\sum_{l=1}^{N} K_l\right)^2, \quad \textit{with} \quad K_r = \frac{1}{N_{\mathrm{S}}}\sum_{s=1}^{N_{\mathrm{S}}} K_{rs}, \tag{32}$$

$$\Lambda_{\mathrm{S}} = \frac{1}{N_{\mathrm{S}}-1}\sum_{r=1}^{N_{\mathrm{S}}}\left(\tilde{K}_s - \frac{1}{N_{\mathrm{S}}}\sum_{l=1}^{N} \tilde{K}_l\right)^2, \quad \textit{with} \quad \tilde{K}_s = \frac{1}{p_{\mathrm{R}}}\sum_{r\in\mathfrak{R}_m} K_{rs}. \tag{33}$$

Proof. It is straightforward to show $\mathbb{E}_{\mathrm{S}}[\chi_m(\mathrm{Y})] = \mathbb{E}[\chi_m(\mathrm{Y})] = 0$. To show (30), we compute

$$\begin{aligned}\mathbb{E}_{\mathrm{S}}[\chi_m^2(\mathrm{Y})] &= \mathbb{E}\left[\left(\frac{1}{p_{\mathrm{R}}}\sum_{r\in\mathfrak{R}_m} K_r - \frac{1}{N_{\mathrm{R}}}\sum_{r=1}^{N_{\mathrm{R}}} K_r\right)^2\right] \\ &= \frac{1}{p_{\mathrm{R}}^2}\mathbb{E}\left(\sum_{r\in\mathfrak{R}_m} K_r\right)^2 - \frac{1}{N_{\mathrm{R}}^2}\left(\sum_{r=1}^{N_{\mathrm{R}}} K_r\right)^2.\end{aligned}$$

Note that

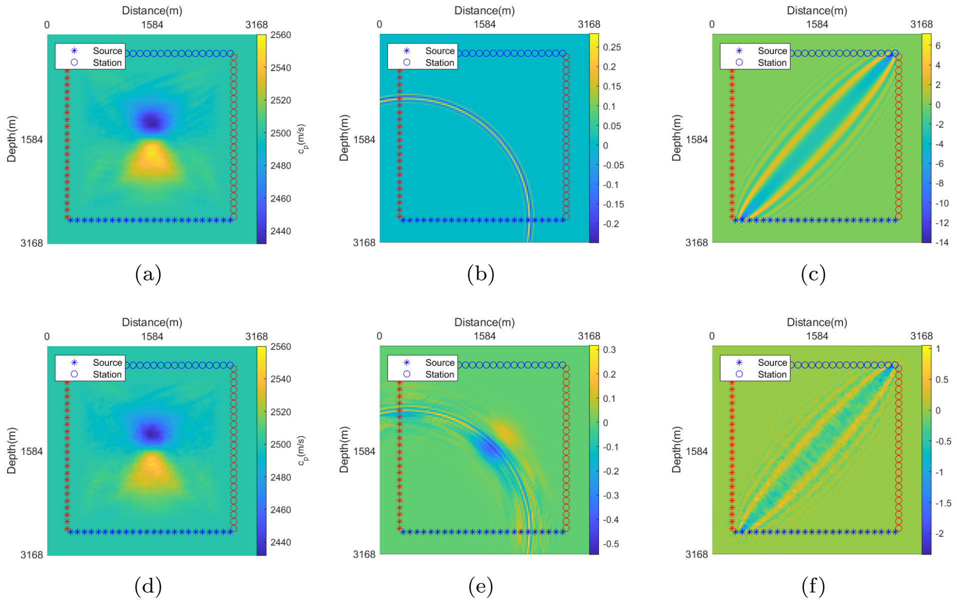
$$\begin{aligned}\mathbb{E}\left(\sum_{r\in\mathfrak{R}_m} K_r\right)^2 &= \sum_{r=1}^{N_{\mathrm{R}}} K_r^2\,\mathbb{P}(r\in\mathfrak{R}_m) + \sum_{r,l:r\neq l} K_r\,K_l\,\mathbb{P}\,(r\in\mathfrak{R}_m \text{ and } l\in\mathfrak{R}_m) \\ &= \frac{p_{\mathrm{R}}}{N_{\mathrm{R}}}\sum_{r=1}^{N_{\mathrm{R}}} K_r^2 + \frac{p_{\mathrm{R}}(p_{\mathrm{R}}-1)}{N_{\mathrm{R}}(N_{\mathrm{R}}-1)}\sum_{r,l:r\neq l} K_r\,K_l,\end{aligned}$$

and thus

$$\begin{aligned}\mathbb{E}_{\mathrm{S}}[\chi_m^2(\mathrm{Y})] &= \left(\frac{1}{p_{\mathrm{R}}N_{\mathrm{R}}} - \frac{1}{N_{\mathrm{R}}^2}\right)\sum_{r=1}^{N_{\mathrm{R}}} K_r^2 + \left(\frac{p_{\mathrm{R}}-1}{p_{\mathrm{R}}\,N_{\mathrm{R}}(N_{\mathrm{R}}-1)} - \frac{1}{N_{\mathrm{R}}^2}\right)\sum_{r,l:r\neq l} K_r\,K_l, \\ &= \left(\frac{1}{p_{\mathrm{R}}} - \frac{1}{N_{\mathrm{R}}}\right)\frac{1}{N_{\mathrm{R}}-1}\sum_{r=1}^{N_{\mathrm{R}}}\left(K_r - \frac{1}{N_{\mathrm{R}}}\sum_{l=1}^{N_{\mathrm{R}}} K_l\right)^2,\end{aligned}$$

yielding (30). Then (31) can be obtained by taking the expectation with respect to $\mathfrak{S}_m$ and $\mathfrak{R}_m$ separately. □

4. Numerical examples. In this section, we present some synthetic tomography tests using random batch gradient reconstruction, where the wave equations are solved by FGA and the sensitivity kernel are constructed using (15). Note that the random batch gradient reconstruction method we propose can be applied to any



(a) (b) (c)

(d) (e) (f)

Fig. 1. *Example* 4.1. 2*D FWI results with batch strategy* 1. (a) *and* (d) *plot the resulted velocities after five and three iterations using batch strategy* 1 *with sampling rate* 10% *and* 2.5%, *respectively. Comparisons of the same setup are also given for the wavefields* ((b) *and* (e)) *and kernels* ((c) *and* (f)).

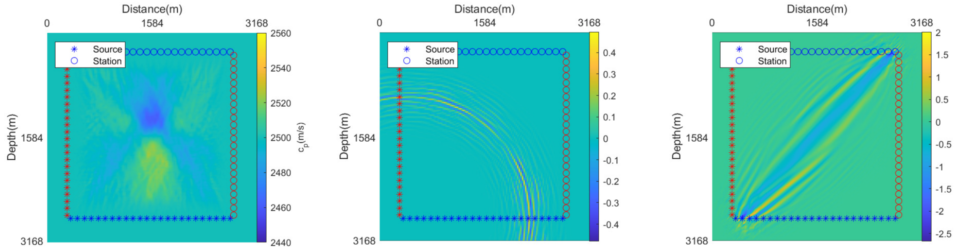
gradient-based iterative method, but the proof of the convergence may be more complicated than that in the previous section. So in the following subsections, we mainly use L-BFGS for the iterations and show the convergence numerically but leave the rigorous proof for further studies. We present one gradient descent example in section 4.5. We remark that all the computations are performed on a Dell T7920 workstation with dual Intel Xeon Gold 6130 Processor (16 Cores, 22 M Cache, 2.10 GHz) and compiled with GFORTRAN and MPICH.

4.1. Full-waveform inversion for a 2D model. In the first example, we present a test using FWI to image a 2D (in the $x-z$ plane) square region. As a proof of methodology, we set point receivers on the top and right of the square region and set point sources aligning on the bottom and left of the square region. The target velocity field is set as

$$
\begin{aligned}
c(x,z) = C_0 \Bigg(1 - \alpha \exp\left(-\frac{\beta}{L^2}\left((x-x_{c1})^2 + (z-z_c)^2\right)\right) \\
+\alpha \exp\left(-\frac{\beta}{L^2}\left((x-x_{c2})^2 + (z-z_c)^2\right)\right)\Bigg),
\end{aligned}
\tag{34}
$$

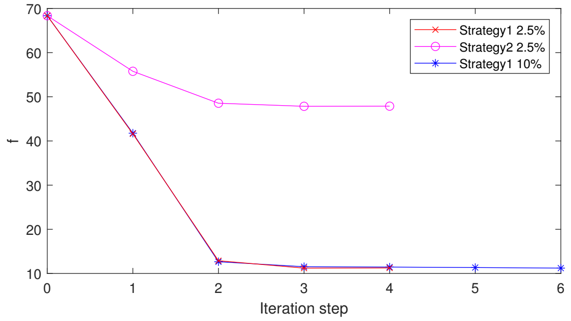
where $C_0 = 2500$ m/s, $x_{c1} = 1344$ m, $x_{c2} = 1824$ m, $z_c = L = 1584$ m, $\alpha = 0.03$, $\beta = 24.2$. See Figure 1(a) for a demonstration of the setup. FWI iteration starts with the background velocity, that is, $c_0 \equiv 2500\,\mathrm{m/s}$ homogeneously. In this example, the beam number $N = 32766$, and $\epsilon = L/256$.

We use FGA to simulate the forward and adjoint wave equations. To reconstruct the wavefields and kernels, we use the two following strategies to generate random batch:



(a) (b) (c)

Fig. 2. *Example* 4.1. *2D FWI results with batch Strategy* 2 *using the sampling rate* 2.5%. *After three iterations, we have* (a) *velocity,* (b) *wavefield,* (c) *kernel.*



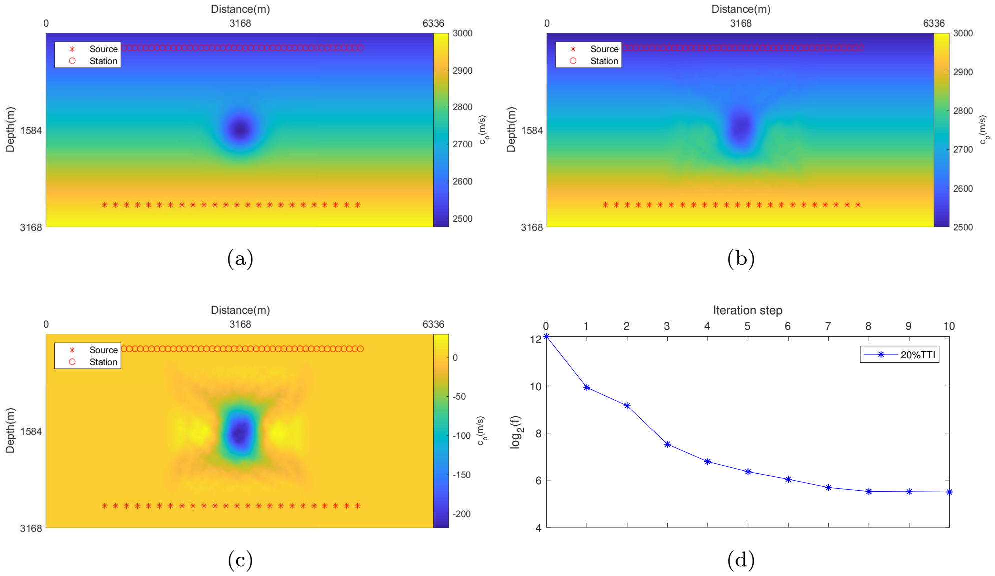
(a)

Fig. 3. *Example* 4.1. *2D FWI results: Decay of the misfit function.*

- *Strategy* 1. As we proposed in section 3.2, for each iteration step m and time evolution step k, we choose batches such that $\{\mathfrak{B}_{m,k}, \mathfrak{B}^{\dagger}_{m,k} : m \in \mathbb{N}, k = 0, 1, \ldots, T/\tau\}$ is independent.
- *Strategy* 2. For each iteration step m we choose two batches $\mathfrak{B}_m$ and $\mathfrak{B}^{\dagger}_m$ independently and set $\mathfrak{B}_{m,k} = \mathfrak{B}_m$, $\mathfrak{B}^{\dagger}_{m,k} = \mathfrak{B}^{\dagger}_m$ for all $k = 0, 1, \ldots, T/\tau$, that is, we lose the independence for time evolution steps.

In Figure 1(a) and(d), we plot the resulting velocities after four iteration steps using batch Strategy 1 with sampling rates $p/N = 10\%$ and $p/N = 10\%$, respectively, and one can see that the low-velocity region has already been captured (though there are blurs and artifacts). We also plot time-shots of the wavefields for both 10% and 2.5% reconstructions in Figure 1(b),(e), respectively, and the kernels for both 10% and 2.5% reconstructions in Figure 1(c),(f), respectively. For a comparison, we use batch Strategy 2 to generate batches and redo the test with the sampling rate 2.5%. In Figure 2, the inversion result is poor even though the wavefield and kernel look "okay." The decay of the misfit functional for these two different strategies is shown in Figure 3.

We can see from Figures 1(b),(e), and 2(b) that different strategies of batch generation in FGA capture similar wavefront shapes; this is because FGA as a ray-based asymptotic method gives correct ray-path information; the wavefields are smooth because they are reconstructed by complex-valued Gaussian functions. Figures 1(c),(f), and 2(c) also show that the banana-doughnut shapes of the kernels look similar to one another. But one can see a significant difference by looking at small-scale structures:



(a) (b) (c) (d)

FIG. 4. *Example* 4.2. *2D TTI model results with gradually changing background.* (a) *Target velocity field;* (b) *velocity field after* 10 *iteration steps;* (c) *velocity field after* 10 *iteration steps, subtracting the background velocity field* $c_{10} - c_0$*;* (d) *decay of the misfit function.*

apparently, there is roughness or a “noise-like” structure in Figure 1(f) generated by Strategy 1, while Figure 2(c) generated by Strategy 2 shows a smooth kernel. The reason for this difference lies is because Strategy 2 has no randomness in time evolution, and so it is smooth when integrating (or summing, for numerical purposes) over time to get the kernel (15), while Strategy 1 uses an independent random batch for each time evolution step and thus shows more randomness. We note here that it is the time-independence that helps Strategy 1 attain a better convergence than Strategy 2, which can be seen in the proof of Lemma 3.1 where the computation of variance (22) relies on the independence directly.

4.2. Travel-time inversion for a 2D gradually changing background model. In the previous subsection, the perturbation in the velocity field is small (3%). As observed in the literature (see, e.g., [6]), TTI has a much wider convergence zone than FWI, so when the perturbation is large, one can use TTI instead of FWI to get a convergent result. To further test the performance of the proposed method, we look at a region with gradually changing background velocity and aim to image a target of low-velocity perturbation using travel-time inversion. As shown in Figure 4(a), 48 stations are put near the surface, 24 sources are put deep inside the earth near the bottom of the target region of size 6336 m × 3168 m, and the background velocity field has a gradual change from 2500m/s at the top ground to 3000 m/s at the deep bottom, and the velocity field is given by

$$(35)\quad c(x,z) = \left(C_1\left(1-\frac{z}{2L}\right)+C_2\frac{z}{2L}\right)\left(1-\alpha\exp\left(-\frac{\beta}{L^2}\left((x-x_c)^2+(z-z_c)^2\right)\right)\right),$$

where $C_1 = 2500$ m/s, $C_2 = 3000$ m/s, $z_c = L = 1584$ m, $x_c = 3168$ m, $\alpha = 0.1$, $\beta = 24.2$. TTI iteration starts with the background velocity c_0, which is given by (35), with the same parameter values as specified above except that here $\alpha = 0$. In this numerical example, the beam number $N = 32768$, and $\epsilon = L/256$. We use FGA to simulate the forward and adjoint wave equations with sampling rate $p/N = 20\%$ to generate the random batch for wavefield and kernel reconstructions. In Figure 4(b), we plot the resulting velocity model c_{10} after 10 iteration steps, and one can see a good match with the target velocity model; this also can be seen in Figure 4(c) where $c_{10} - c_0$ is plotted; Figure 4(d) shows the decay of the misfit functional, and one can see the iteration has numerical convergence.

4.3. Travel-time inversion for a three-layered model. In this example, we apply TTI inversion with random batch gradient reconstruction in a cross-well setup, which is often used for high-resolution reservoir characterization in exploration geophysics. Two wells with 24 sources and 48 stations, respectively, are set on the left and right sides of a region of size 3168 m × 6436 m. The target velocity model is chosen to be three-layered in the form of

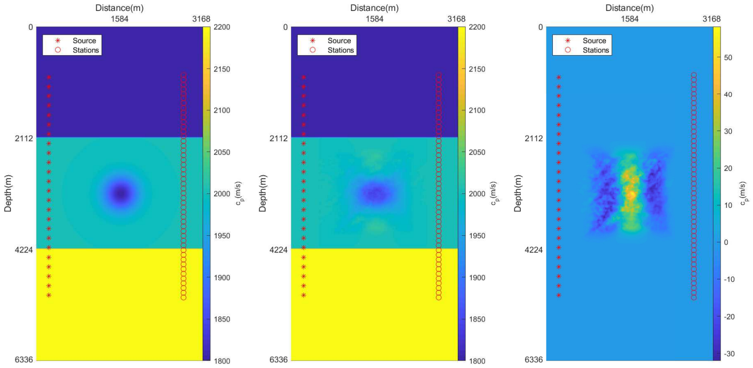
$$
\begin{aligned}
c(x,y,z) &= c(x,z) \\
&= \begin{cases} C_1, & \text{if } z_0 < z < z_1, \\ C_2\left(1-\alpha\exp\left(-\frac{\beta}{L^2}\left((x-x_c)^2+(z-z_c)^2\right)\right)\right) & \text{if } z_1 < z < z_2, \\ C_3, & \text{if } z > z_2, \end{cases}
\end{aligned} \tag{36}
$$

where the background velocities in three layers are $C_1 = 1800$ m/s, $C_2 = 2000$ m/s, $C_3 = 2200$ m/s, and the layer interfaces are located at $z_0 = 0$ m, $z_1 = 2112$ m, $z_2 = 4224$ m. A low-velocity region characterized by a Gaussian perturbation is located at the center of the second layer with $x_c = 1584$ m, $z_c = 3168$ m, and we choose $\alpha = 10\%$, $\beta = 24.2$, and $L = 3168$ m. See Figure 5(a) for an illustration of the velocity model and the source-receiver setup. In this numerical example, the beam number $N = 32766$, and $\epsilon = L/256$. Travel-time inversion iteration starts with a piecewise constant background velocity, which is given by (36), with the same parameter values specified above except that here $\alpha = 0$. We use FGA to simulate the forward and adjoint wave equations with sampling rate $p/N = 20\%$ to generate the random batch for wavefield and kernel reconstructions. In Figure 5(b), we plot the resulting velocity model c_9 after nine iteration steps, and one can see a good match with the target velocity model. Figure 5(c) shows the difference in resulting and target velocity models $c_9 - c$, and one can see that the residual is relatively small compared to the background and the Gaussian perturbation. Decay of the misfit during iteration is shown in Figure 6, and one can see that the iteration has numerical convergence.

4.4. Travel-time inversion for a 3D model. In this subsection, we present a test using TTI to image a 3D cube region. The target velocity field is assumed homogeneous in the y-direction and is set to be a 10% Gaussian perturbation of a homogeneous background with velocity 2500 m/s, i.e.,

$$
c(x,y,z) = c(x,z) = C_0\left(1-\alpha\exp\left(-\frac{\beta}{L^2}\left((x-x_c)^2+(z-z_c)^2\right)\right)\right), \tag{37}
$$

where $C_0 = 2500$ m/s, $x_c = z_c = L = 1584$ m, $\alpha = 0.1$, $\beta = 24.2$. Travel-time inversion iteration starts with a velocity field with 2500 m/s homogeneously. In this numerical example, the beam number $N = 65534$, and $\epsilon = L/64$. We use FGA to simulate the forward and adjoint wave equations in three dimensions. To reconstruct the wavefields and kernels, we use Strategy 1 with sampling rates $p/N = 5\%$, 10%, and 20% to generate the random batch for wavefield and kernel reconstructions.



(a) (b) (c)

Fig. 5. *Example* 4.3. 2*D TTI results for a three-layered model.* (a) *Target velocity;* (b) *velocity after nine iteration steps;* (c) *the difference between velocity after nine iteration steps and target velocity.*

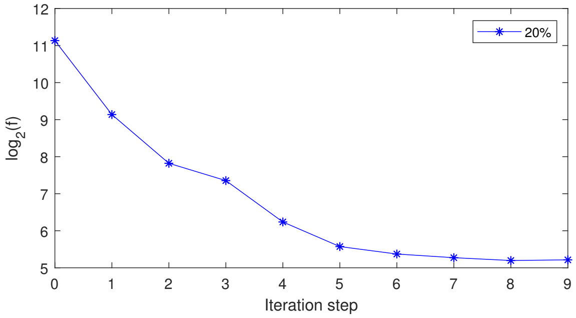
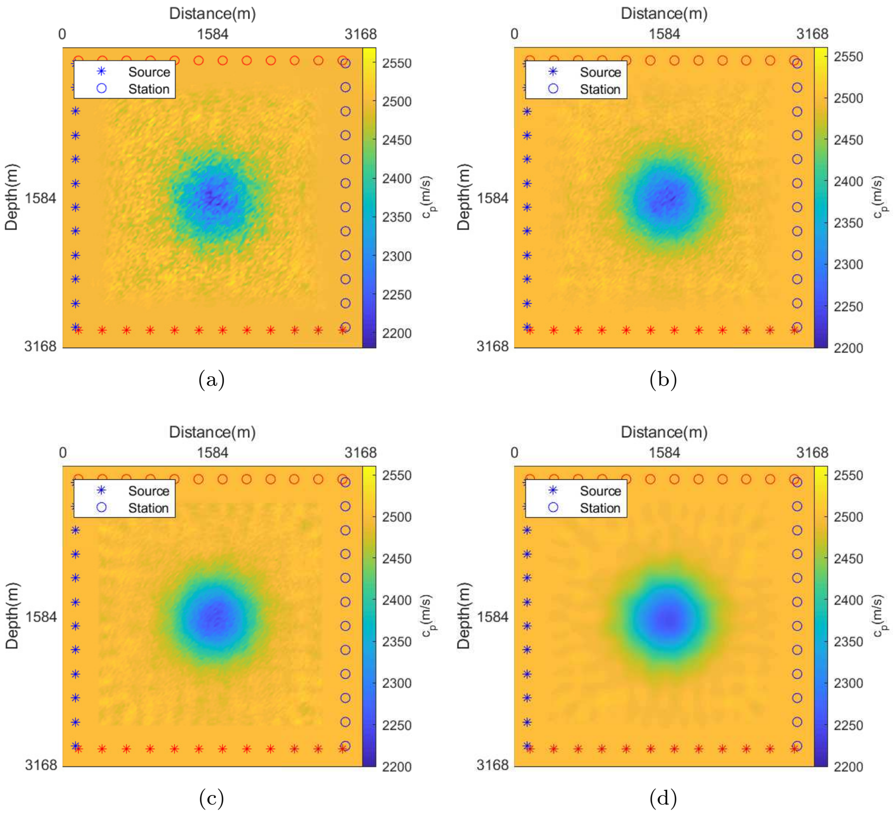


Fig. 6. *Example* 4.3. 2*D TTI results for a three-layered model: Decay of the misfit function.*

It can be seen from Figure 7 that even when 5% of the beams are used, the tomography can capture at least the location and shape of the low-velocity region and give a reasonable model. The larger the sampling rate, the better the resolution of the resulting image. This phenomenon can be further seen in Figure 8 as we look at the values of the misfit functional. One can observe a smaller misfit value for a larger sampling rate, which is consistent with Theorem 3.1 since larger p/N implies smaller variance in (20) for fixed N. On the other hand, for the first several iteration steps, the values of the misfit functions are almost the same for different sampling rates, which numerically indicates that the convergence rate of the iteration is not sensitive to the batch size.

4.5. Travel-time inversion for a 3D refraction model. In this subsection, we give an example using TTI to image a 3D cuboid region with a completely reflective bottom interface. The target velocity field is assumed homogeneous in the y-direction and is set to be a 10% Gaussian perturbation of a homogeneous background with velocity 2500 m/s, i.e.,

$$c(x,y,z) = c(x,z) = C_0\left(1 - \alpha \exp\left(-\frac{\beta}{L^2}\left((x-x_c)^2 + (z-z_c)^2\right)\right)\right), \tag{38}$$



(a) (b) (c) (d)

FIG. 7. *Example* 4.4. *3D TTI results.* (a) *Velocity model after nine iteration steps with sampling rate* 5%*;* (b) *velocity model after seven iteration steps with sampling rate* 10%*;* (c) *velocity model after seven iteration steps with sampling rate* 20%*;* (d) *velocity model after nine iteration steps with sampling rate* 100%*.*

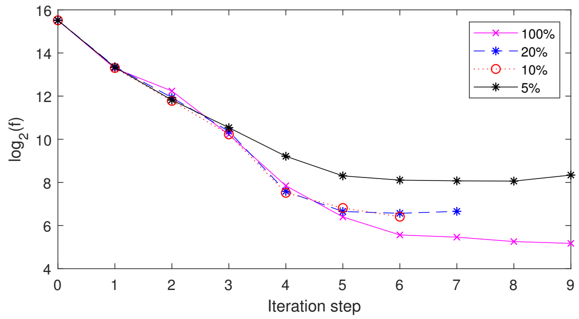


FIG. 8. *Example* 4.4. *3D TTI results: Decay of misfit function.*

where $C_0 = 2500$ m/s, $x_c = L = 1584$ m, $z_c = 891$ m, $\alpha = 0.1$, $\beta = 6.05$. Travel-time inversion iteration starts with a velocity field with 2500 m/s homogeneously. In this numerical example, the beam number $N = 65534$, and $\epsilon = L/64$. We use FGA to simulate the forward and adjoint wave equations in three dimensions. To reconstruct the wavefields and kernels, we use Strategy 1 with sampling rates $p/N = 5\%$, 20%, and 100% to generate the random batch for wavefield and kernel reconstructions.

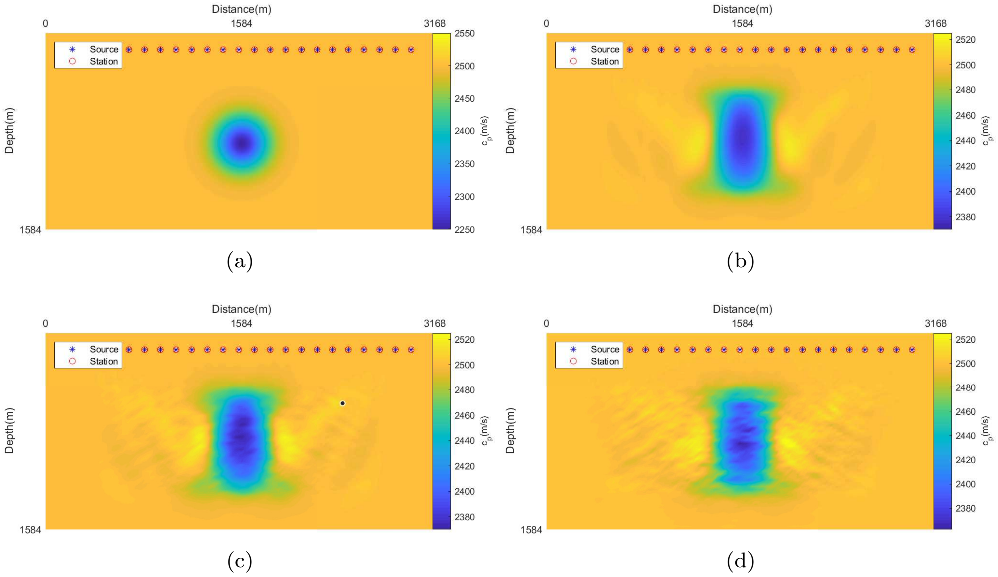


FIG. 9. *Example* 4.5. *3D TTI model results with reflection bottom.* (a) *Target velocity field;* (b) *velocity field after* 14 *iteration steps with sampling rate* 100%*;* (c) *velocity field after* 11 *iteration steps with sampling rate* 20%*;* (d) *velocity field after* 15 *iteration steps with sampling rate* 5%*.*

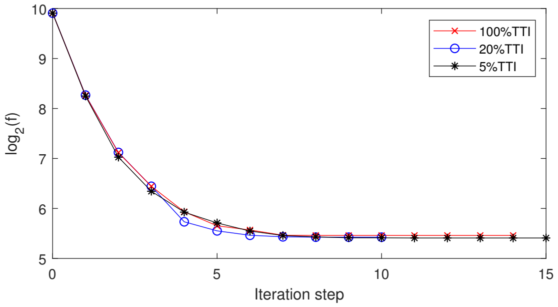


FIG. 10. *Example* 4.5. *3D TTI Model results with reflection bottom: Decay of misfit function.*

It can be seen from Figure 10 that, in this test, the tomography can capture at least the location and shape of the low-velocity region and give a reasonable model. The results with different sampling rates are nearly the same. This phenomenon can be further observed in Figure 10 as we look at the values of the misfit function.

We can also compare the computation times to show the effectiveness of the stochastic method in saving computations. As one can see from Table 1, for different sampling rates from 100% to 5%, the computation time spent for one iteration step varies from 30.08 hours to 4.18 hours, which indicates that it can save about 86.1% CPU time by using the random batch method.

5. Conclusion and discussion. In this paper, we propose a type of stochastic gradient descent method for seismic tomography. Specifically, we use the frozen Gaussian approximation (FGA) to compute seismic wave propagation, and then we construct stochastic gradients by random batch methods (RBMs). One can easily gen-

Table 1
Table of efficiency for different sampling rates.

Efficiency of random batch method		
Sampling rate	Computation time	Savings
100%	30.08h	0%
20%	7.65h	74.6%
10%	5.08h	83.1%
5%	4.18h	86.1%

eralize this idea by replacing FGA with any other efficient wave propagation solver, e.g., the Gaussian beam method. The convergence of the proposed method is proved in the mean-square sense, and we present four examples of both wave-equation-based travel-time inversion (TTI) and full-waveform inversion (FWI) to show the numerical performance. This method introduces the possibility of efficiently solving high-dimensional optimization problems in seismic tomography. We plan to apply it to the imaging of Earth's subsurface structures using realistic seismic signals.

REFERENCES

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

[2] G. Bao, J. Qian, L. Ying, and H. Zhang, *A convergent multiscale Gaussian-beam parametrix for the wave equation*, Commun. Partial Differential Equations, 38 (2013), pp. 92–134, https://doi.org/10.1080/03605302.2012.727130.

[3] V. Cerveny, M. M. Popov, and I. Psencik, *Computation of wave fields in inhomogeneous media—Gaussian beam approach*, Geophys. J. Roy. Astr. Soc., 70 (1982), pp. 109–128.

[4] L. Chai, J. C. Hateley, E. Lorin, and X. Yang, *On the convergence of frozen Gaussian approximation for linear non-strictly hyperbolic systems*, Commun. Math. Sci., 19 (2021), pp. 585–606.

[5] L. Chai, P. Tong, and X. Yang, *Frozen Gaussian approximation for* 3*-D seismic wave propagation*, Geophys. J. Int., 208 (2017), pp. 59–74.

[6] L. Chai, P. Tong, and X. Yang, *Frozen Gaussian approximation for* 3*-D seismic tomography*, Inverse Problems, 34 (2018), 055004.

[7] S. Fomel and N. Tanushev, *Time-domain seismic imaging using beams*, in SEG Technical Program Expanded Abstracts 2009, Society of Exploration Geophysicists, 2009, pp. 2747–2752.

[8] S. Gray and N. Bleistein, *True-amplitude Gaussian-beam migration*, Geophysics, 74 (2009), pp. S11–S23.

[9] S. H. Gray, *Efficient traveltime calculations for Kirchhoff migration*, Geophysics, 51 (1986), pp. 1685–1688.

[10] S. H. Gray, *Gaussian beam migration of common-shot records*, Geophysics, 70 (2005), pp. S71–S77.

[11] J. C. Hateley, L. Chai, P. Tong, and X. Yang, *Frozen Gaussian approximation for* 3*-D elastic wave equation and seismic tomography*, Geophys. J. Int., 216 (2019), pp. 1394–1412.

[12] E. J. Heller, *Frozen Gaussians: A very simple semiclassical approximation*, J. Chem. Phys., 75 (1981), pp. 2923–2931.

[13] D. Helmberger, *The crust-mantle transition in the Bering sea*, Bull. Seism. Soc. Amer., 58 (1968), pp. 179–214.

[14] M. F. Herman and E. Kluk, *A semiclassical justification for the use of non-spreading wavepackets in dynamics calculations*, Chem. Phys., 91 (1984), pp. 27–34.

[15] N. R. Hill, *Gaussian beam migration*, Geophysics, 55 (1990), pp. 1416–1428.

[16] N. R. Hill, *Prestack Gaussian-beam depth migration*, Geophysics, 66 (2001), pp. 1240–1250.

[17] S. Jin, L. Li, and J.-G. Liu, *Random batch methods (rbm) for interacting particle systems*, J. Comput. Phys., 400 (2020), 108877.

[18] K. Kay, *Integral expressions for the semi-classical time-dependent propagator*, J. Chem. Phys., 100 (1994), pp. 4377–4392.

[19] K. Kay, *The Herman-Kluk approximation: Derivation and semiclassical corrections*, Chem. Phys., 322 (2006), pp. 3–12.

[20] T. H. Keho and W. B. Beydoun, *Paraxial ray Kirchhoff migration*, Geophysics, 53 (1988), pp. 1540–1546.

[21] D. P. Kingma and J. Ba, *Adam: A Method for Stochastic Optimization*, preprint, https://arxiv.org/abs/1412.6980, 2014.

[22] J. Li, G. Lin, and X. Yang, *A frozen Gaussian approximation-based multi-level particle swarm optimization for seismic inversion*, J. Comput. Phys., 296 (2015), pp. 58–71.

[23] Q. Liu and Y. J. Gu, *Seismic imaging: From classical to adjoint tomography*, Tectonophysics, 566/567 (2012), pp. 31–66.

[24] Q. Liu and J. Tromp, *Finite-frequency sensitivity Kernels for global seismic wave propagation based upon adjoint methods*, Geophys. J. Int., 174 (2008), pp. 265–286.

[25] J. Lu and X. Yang, *Frozen Gaussian approximation for high frequency wave propagation*, Commun. Math. Sci., 9 (2011), pp. 663–683.

[26] J. Lu and X. Yang, *Convergence of frozen Gaussian approximation for high frequency wave propagation*, Comm. Pure Appl. Math., 65 (2012), pp. 759–789.

[27] R. L. Nowack, M. K. Sen, and P. L. Stoffa, *Gaussian beam migration for sparse common-shot and common-receiver data*, in SEG Technical Program Expanded Abstracts 2003, Society of Exploration Geophysicists, 2003, pp. 1114–1117.

[28] M. M. Popov, N. M. Semtchenok, P. M. Popov, and A. R. Verdel, *Depth migration by the Gaussian beam summation method*, Geophysics, 75 (2010), pp. S81–S93.

[29] J. Qian and L. Ying, *Fast multiscale gaussian wavepacket transforms and multiscale Gaussian beams for the wave equation*, Multiscale Model. Simul., 8 (2010), pp. 1803–1837, https://doi.org/10.1137/100787313.

[30] L. Qiu, J. Ramos-Martínez, A. Valenciano, Y. Yang, and B. Engquist, *Full-waveform inversion with an exponentially encoded optimal-transport norm*, in SEG Technical Program Expanded Abstracts 2017, Society of Exploration Geophysicists, 2017, pp. 1286–1290.

[31] N. Rawlinson, S. Pozgay, and S. Fishwick, *Seismic tomography: A window into deep earth*, Phys. Earth Planet. Inter., 178 (2010), pp. 101–135.

[32] F. Rickers, A. Fichtner, and J. Trampert, *The Iceland–Jan Mayen plume system and its impact on mantle dynamics in the North Atlantic region: Evidence from full-waveform inversion*, Earth Planet. Sci. Lett., 367 (2013), pp. 39–51.

[33] B. Romanowicz, *Seismic tomography of the earth's mantle*, Annu. Rev. Earth Planet. Sci., 19 (1991), pp. 77–99.

[34] S. Simute, H. Steptoe, L. Cobden, and A. Fichtner, *Full-waveform inversion of the Japanese Islands region*, J. Geophys. Res. Solid Earth, 121 (2016), pp. 3722–3741.

[35] T. Swart and V. Rousse, *A mathematical justification of the Herman-Kluk propagator*, Commun. Math. Phys., 286 (2009), pp. 725–750.

[36] C. Tape, Q. Liu, A. Maggi, and J. Tromp, *Adjoint tomography of the southern California crust*, Science, 325 (2009), pp. 988–992.

[37] C. Tape, Q. Liu, A. Maggi and J. Tromp, *Seismic tomography of the southern California crust based on spectral-element and adjoint methods*, Geophys. J. Int., 180 (2010), pp. 433–462.

[38] P. Tong, C.-W. Chen, D. Komatitsch, P. Basini, and Q. Liu, *High-resolution seismic array imaging based on an SEM-FK hybrid method*, Geophys. J. Int., 197 (2014), pp. 369–395.

[39] J. Tromp, C. Tape, and Q. Liu, *Seismic tomography, adjoint methods, time reversal and banana-doughnut kernels*, Geophys. J. Int., 160 (2005), pp. 195–216.

[40] D. P. van Herwaarden, C. Boehm, M. Afanasiev, S. Thrastarson, L. Krischer, J. Trampert, and A. Fichtner, *Accelerated full-waveform inversion using dynamic mini-batches*, Geophys. J. Int., 221 (2020), pp. 1427–1438.

[41] J. Vidale and D. Helmberger, *Elastic finite-difference modeling of the* 1971 *San Fernando, California earthquake*, Bull. Seism. Soc. Amer., 78 (1988), pp. 122–141.

[42] Y. Xie and Z. Zhou, *Frozen Gaussian Sampling: A Mesh-Free Monte Carlo Method for Approximating Semiclassical Schrödinger Equations*, https://arxiv.org/abs/2112.05405, 2021.

[43] X. Yang, J. Lu, and S. Fomel, *Seismic modeling using the frozen Gaussian approximation*, SEG Technical Program Expanded Abstracts 2013, Society of Exploration Geophysicists, 2013, pp. 4677–4682.

[44] Y. Yang, B. Engquist, J. Sun, and B. F. Hamfeldt, *Application of optimal transport and the quadratic Wasserstein metric to full-waveform inversion*, Geophysics, 83 (2018), pp. R43–R62.

[45] D. Zhao, *Tomography and dynamics of Western-Pacific subduction zones*, Monogr. Environ. Earth Planets, 1 (2012), pp. 1–70.

[46] H. Zhu, E. Bozdag, D. Peter, and J. Tromp, *Structure of the European upper mantle revealed by adjoint tomography*, Nat. Geosci., 5 (2012), pp. 493–498.