FAST ONLINE ORTHONORMAL DICTIONARY LEARNING FOR EFFICIENT FULL WAVEFORM INVERSION

Lingchen Zhu, Entao Liu, James H. McClellan

Center for Energy and Geo Processing (CeGP) at Georgia Tech and KFUPM
Georgia Institute of Technology, 75 5th St NW, Atlanta, GA 30308
{lczhu, eliu, jim.mcclellan}@gatech.edu

ABSTRACT

Full waveform inversion (FWI) delivers high-resolution images of a subsurface medium property by minimizing iteratively the misfit between observed and simulated seismic data, and is commonly used by the oil and gas industry for geophysical exploration. FWI is a challenging problem because seismic surveys cover ever larger areas of interest and collect massive volumes of data. The dimensionality of the problem and the heterogeneity of the medium both stress the need for faster algorithms, so sparse regularization techniques can be used to accelerate and improve imaging results.

In this paper, we propose a compressive sensing method for the FWI problem by exploiting the sparsity of geological model perturbations over learned dictionaries. Based on stochastic approximations, the dictionaries are updated iteratively to adapt changing models during FWI iterations. Meanwhile, the dictionaries are kept orthonormal in order to maintain the corresponding transform in a fast and compact manner so that these transforms do not introduce extra computational overhead to FWI. Establishing such a sparsity regularization on the model enables us to significantly reduce the workload by only collecting 0.625% of the field data without introducing subsampling artifacts. Hence, the computational burden of large-scale FWI problems can be greatly reduced.

Index Terms— sparse representation, dictionary learning, orthonormal basis, compressive sensing, full waveform inversion

1. INTRODUCTION

Dictionary learning has now become a promising technique for sparse signal representation and approximation. Compared to traditional transforms such as wavelet, curvelet, etc. with a predefined dictionary, dictionary learning based transforms are better able to adapt to nonintuitive signal regularities beyond piecewise smoothness. Numerical signal processing tasks such as denoising [1, 2] and inpainting [3] have benefited from the use of adaptive dictionaries that lead to more sparse representations of high dimensional signals and achieve state-of-the-art results. Such a technique has great potential on seismic imaging problems such as full waveform inversion (FWI) due to its continued demand on high dimensional seismic data.

Recent overcomplete dictionary learning algorithms such as K-SVD and its variants [4–6] train a dictionary $\mathbf{D}$ with two steps: (1) sparse coding and (2) dictionary update. The first step uses matching pursuit algorithms [7–9] to find the sparse coefficients of the input training samples over the current $\mathbf{D}$ by solving an ℓ_0-norm regularized minimization problem. The second step updates $\mathbf{D}$ by solving a gradient descent problem using the known sparse coefficients. However, the drawback of K-SVD algorithms is that they have to train atoms in $\mathbf{D}$ sequentially with high computational complexity. Such an issue can be bypassed by imposing orthonormality on $\mathbf{D}$ [10], which yields an orthonormal dictionary learning algorithm that trains all atoms in $\mathbf{D}$ at once.

All above algorithms are iterative batch procedures that require to access a fixed set of training set to learn $\mathbf{D}$. Though they have shown the ability to exploit sparsity from the data, they may not effectively handle dynamic training data changing over time such as geophysical models used in FWI. To address this issue, we propose an online approach that updates the orthonormal dictionary to adapt the currently obtained model. The model for the next iteration is sparsely represented by the updated dictionary and, therefore, such sparsity allows us to make the industrial-scale FWI feasible and much more efficient by significantly reducing its problem dimensionality based on the compressive sensing method.

The contributions in this paper are the following:

- We propose an iterative online algorithm for orthonormal dictionary learning by minimizing the expectation of the cost function when new training samples join.
- We implement the compressive sensing framework into large-scale FWI problems by admitting sparse representation of model perturbations over learned dictionaries and reducing the problem dimensionality with randomized encoding.
- As shown in Section 5, our method can significantly reduce the amount of data used in FWI and decrease the running time without introducing any visible artifacts.

2. ORTHONORMAL DICTIONARY LEARNING

Given a signal set $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_R] \in \mathbb{R}^{N\times R}$ in which each element represents a vectorized training sample, dictionary learning minimizes the following empirical cost function

$$e_R(\mathbf{Y}, \mathbf{D}) \triangleq \frac{1}{R}\sum_{i=1}^{R}\left(\|\mathbf{y}_i - \mathbf{D}\mathbf{x}_i\|_2^2 + \lambda\|\mathbf{x}_i\|_0\right) \quad (1)$$

where $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_R] \in \mathbb{R}^{L\times R}$ are the sparse coefficients of $\mathbf{Y}$ over the dictionary $\mathbf{D} \in \mathbb{R}^{N\times L}(N \le L)$, λ is a Lagrange multiplier and $\|\cdot\|_0$ is the ℓ_0-norm that counts the nonzero entries of a vector. Since there are two unknown variables $\mathbf{D}$ and $\mathbf{X}$, this problem can be solved by minimizing over one while keeping the other one fixed, as commonly done in K-SVD algorithms [4–6].

Orthonormal dictionary learning places a constraint on $\mathbf{D} \in \mathbb{R}^{N\times N}$ such that $\mathbf{D}^T\mathbf{D} = \mathbf{I}$ and minimizes the empirical cost function $e_R(\mathbf{Y}, \mathbf{D})$ in (1), whose matrix form is

$$\min_{\mathbf{D},\mathbf{X}} \frac{1}{R}\left(\|\mathbf{Y} - \mathbf{D}\mathbf{X}\|_F^2 + \lambda\|\mathbf{X}\|_0\right) \quad \text{s.t.} \quad \mathbf{D}^T\mathbf{D} = \mathbf{I} \quad (2)$$

where $\|\cdot\|_F$ is the Frobenius norm. To solve the minimization problem in (2), the first step is to find the sparsest representation of $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_R]$ over a fixed orthonormal dictionary $\mathbf{D}$. This first step would be formulated as

$$\hat{\mathbf{X}} = \underset{\mathbf{X}}{\operatorname{argmin}}\left(\|\mathbf{Y} - \mathbf{D}\mathbf{X}\|_F^2 + \lambda\|\mathbf{X}\|_0\right) \quad (3)$$

whose solution is straightforward by hard-thresholding the entries of $\mathbf{C} = \mathbf{D}^T\mathbf{Y}$ with threshold $\sqrt{\lambda}$ [10, 11] as

$$\hat{x}_{ij} = \begin{cases} c_{ij}, & |c_{ij}| \ge \sqrt{\lambda} \\ 0, & |c_{ij}| < \sqrt{\lambda}. \end{cases} \quad (4)$$

The second step is to optimize the orthonormal dictionary $\mathbf{D}$ by solving an orthogonal procrustes problem [12] that minimizes the reconstruction error given the present values of the sparse coefficients $\hat{\mathbf{X}} = [\hat{\mathbf{x}}_1, \hat{\mathbf{x}}_2, \ldots, \hat{\mathbf{x}}_R]$, i.e.,

$$\hat{\mathbf{D}} = \underset{\mathbf{D}}{\operatorname{argmin}} \|\mathbf{Y} - \mathbf{D}\hat{\mathbf{X}}\|_F^2 \quad \text{s.t.} \quad \mathbf{D}^T\mathbf{D} = \mathbf{I}. \quad (5)$$

It is proved in [10, 12] that if we define a matrix $\mathbf{P} = \hat{\mathbf{X}}\mathbf{Y}^T$ and and let $\mathbf{P} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ denote its singular value decomposition (SVD), then the orthonormal matrix $\hat{\mathbf{D}} = \mathbf{V}\mathbf{U}^T$ solves (5). The orthonormal dictionary $\mathbf{D}$ can thus be learned by alternating between the two steps (3) and (5) iteratively until the cost function $e_R(\mathbf{Y}, \mathbf{D})$ converges to a steady state.

When compared to the overcomplete dictionary learning method K-SVD, the computational complexity of orthonormal dictionary learning is significantly lower. For each learning iteration, we need only one matrix-vector multiplication to obtain the sparse coding and one SVD to update the entire dictionary $\mathbf{D}$ in (5). There is no need for complex iterative algorithms such as basis pursuit or matching pursuit that have been widely used in the K-SVD to update the dictionary atoms sequentially.

3. ONLINE ORTHONORMAL DICTIONARY LEARNING

For large-scale and dynamic dictionary learning problems, an online method based on stochastic approximation is attractive. In this case, [13] suggests minimizing the expectation of the cost function

$$\begin{aligned} e(\mathbf{D}) &\triangleq \mathbb{E}_{\mathbf{y}}\left[\|\mathbf{y} - \mathbf{D}\mathbf{x}\|_2^2 + \lambda\|\mathbf{x}\|_0\right] \\ &= \lim_{R\to\infty} e_R(\mathbf{Y}, \mathbf{D}) \quad \text{almost surely} \end{aligned} \quad (6)$$

instead of the empirical cost function $e_R(\mathbf{Y}, \mathbf{D})$. Then we can make a trade-off between the computational complexity and estimation error. Minimizing $e(\mathbf{D})$ relies on the (unknown) stochastic characteristics of the training samples, not the number of samples. Also, the online method takes the previous learning information into account so that it always keeps the representation sparse for dynamic data, which is a vital property for applying compressive sensing in the FWI problems.

Input: input data $\mathbf{y} \in \mathbb{R}^N$ that follows an unknown distribution $p(\mathbf{y})$, initial orthonormal dictionary $\mathbf{D}^{(0)} \in \mathbb{R}^{N\times N}$, Lagrange multiplier λ, number of update iterations T, mini-batch size R
Output: learned orthonormal dictionary $\mathbf{D}^{(T)}$, sparse representation matrix $\mathbf{X}^{(T)}$
Initialization : $\mathbf{P}^{(0)} = \mathbf{0}$
for $t = 1$ **to** T **do**
  Draw a mini-batch of data $\mathbf{Y}^{(t)} = [\mathbf{y}_1^{(t)}, \mathbf{y}_2^{(t)}, \ldots, \mathbf{y}_R^{(t)}]$;
  while $\|\mathbf{Y} - \mathbf{D}\mathbf{X}\|_F^2 + \lambda\|\mathbf{X}\|_0$ *not converged* **do**
    $\mathbf{C}^{(t)} = [\mathbf{D}^{(t-1)}]^T\mathbf{Y}^{(t)}$;
    $\hat{x}_{ij}^{(t)} = \begin{cases} c_{ij}^{(t)}, & |c_{ij}^{(t)}| \ge \sqrt{\lambda} \\ 0, & |c_{ij}^{(t)}| < \sqrt{\lambda} \end{cases}$;
    $\beta^{(t)} = \frac{tR+1-R}{tR+1}$;
    $\mathbf{P}^{(t)} = \beta^{(t)}\mathbf{P}^{(t-1)} + \mathbf{X}^{(t)}[\mathbf{Y}^{(t)}]^T$;
    SVD of $\mathbf{P}^{(t)} = \mathbf{U}^{(t)}\mathbf{\Sigma}^{(t)}[\mathbf{V}^{(t)}]^T$;
    $\mathbf{D}^{(t)} = \mathbf{V}^{(t)}[\mathbf{U}^{(t)}]^T$;
  end
end

Algorithm 1: Online Orthonormal Dictionary Learning

Algorithm 1 summarizes the online version of orthonormal dictionary learning method. In each iteration, we first draw R training samples $\mathbf{Y}^{(t)}$, which can be from a large dataset or from the current data snapshot. Then we carry out sparse coding over the dictionary $\mathbf{D}^{(t-1)}$ by hard-thresholding with $\sqrt{\lambda}$, and obtain the updated dictionary $\mathbf{D}^{(t)}$ aided by the SVD of $\mathbf{P}^{(t)}$. Essentially, the above two alternating steps keep reducing the value of the cost function

$$\hat{e}_t(\mathbf{D}) \triangleq \frac{1}{t}\sum_{i=1}^{t}\left(\|\mathbf{Y}^{(i)} - \mathbf{D}\mathbf{X}^{(i)}\|_F^2 + \lambda\|\mathbf{X}^{(i)}\|_0\right) \quad (7)$$

which aggregates all historical information computed during the previous learning iterations. One practical implementa-

tion is to rescale the older information so that newer updates $\mathbf{X}^{(t)}[\mathbf{Y}^{(t)}]^T$ can have more weight in $\mathbf{P}^{(t)}$, which is done by using a weighting factor $\beta^{(t)}$. It is proved in [13] that $\hat{e}_t(\mathbf{D}^{(t)})$ converges to $e(\mathbf{D}^{(t)})$ with probability one, so the online orthonormal dictionary learning converges to a stationary point.

4. A FULL WAVEFORM INVERSION FRAMEWORK REGULARIZED BY GEOLOGICAL SPARSITY

FWI uses two-way wave equations to recover velocity models from seismic survey data. The schematic workflow is shown in Figure 1. Without loss of generality, we only consider 2D acoustic waves in this paper (to conserve space).

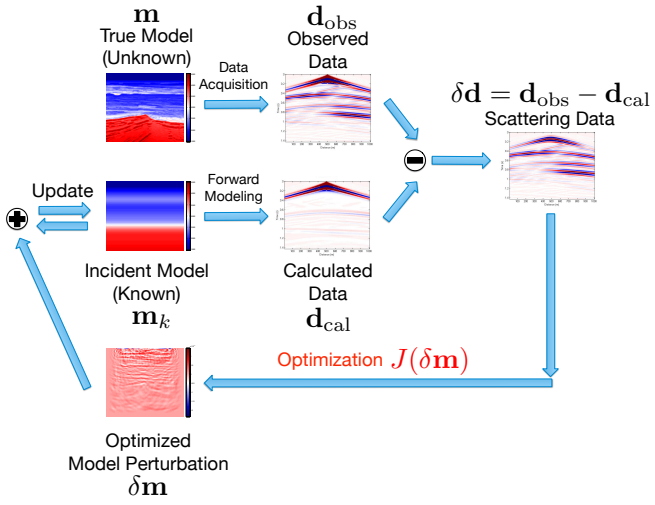


Fig. 1: Schematic FWI Workflow

In the frequency domain, forward modeling of an acoustic wavefield $p(\mathbf{x},\omega;\mathbf{x}_s)$ on a constant-density velocity model $m(\mathbf{x})$ of size $N_z \times N_x$ with a point source $f(\omega)\delta(\mathbf{x}-\mathbf{x}_s)$ at position $\mathbf{x}_s$, where $f(\omega)$ is usually a wavelet source, can be written as the following partial differential equation (PDE)

$$\left(-m(\mathbf{x})\omega^2 - \nabla^2\right) p(\mathbf{x},\omega;\mathbf{x}_s) = f(\omega)\delta(\mathbf{x}-\mathbf{x}_s) \quad (8)$$

Let $d_{\text{obs}}(\mathbf{x}_r,\omega;\mathbf{x}_s)$ denote the recorded seismic data collected at receivers located at $\mathbf{x}_r$ and $d_{\text{cal}}(\mathbf{x}_r,\omega;\mathbf{x}_s)$ denote the synthetic seismic data obtained by sampling the solution of PDE (8) at the same receiver positions. FWI aims to minimize the following nonlinear least squares misfit function [14]

$$J(\mathbf{m}) \triangleq \frac{1}{2}\|\mathbf{d}_{\text{obs}} - \mathbf{d}_{\text{cal}}\|_2^2 \quad (9)$$

where we stack the data of $d(\mathbf{x}_r,\omega;\mathbf{x}_s)$ for all N_r receivers, N_s sources and N_ω frequencies into a vector $\mathbf{d}$ and denote the velocity model as a vector $\mathbf{m}$.

FWI is essentially a local optimization problem, where we minimize (9) by approaching the true model iteratively

$$\mathbf{m}_{k+1} = \mathbf{m}_k + \delta\mathbf{m}. \quad (10)$$

With the Born approximation of acoustic scattering [15], Eq. (9) can be reformulated in the following linearized form

$$J(\delta\mathbf{m}) \triangleq \frac{1}{2}\|\delta\mathbf{d} - \mathcal{F}\delta\mathbf{m}\|_2^2 \quad (11)$$

where $\delta\mathbf{d} \triangleq \mathbf{d}_{\text{obs}} - \mathbf{d}_{\text{cal}}$ and $\mathcal{F} \triangleq \partial\mathbf{d}_{\text{cal}}/\partial\delta\mathbf{m}$ of size $N_\omega N_s N_r \times N_z N_x$ measures the sensitivity of the wave field with respect to the model perturbation $\delta\mathbf{m}$. Particularly, $\mathcal{F}$ involves products of monochromatic Green's functions obtained from the PDE (8) for all source and receiver pairs and all frequencies. Thus, for the FWI problem with N_ω frequencies, N_s shots and N_r receivers, minimizing $J(\delta\mathbf{m})$ in (11) typically requires solving $N_\omega(N_s+N_r)$ PDEs.

Initialization : $k=0$, $\mathbf{m}_0=\mathbf{m}_s$, $\Delta_0=\infty$
while $\Delta_k > \epsilon$ *and* $k \le k_{\max}$ **do**
1. Randomly draw N'_ω out of N_ω frequencies to form a set Ω' and N'_s out of N_s sources to form a set $\mathcal{S}'$;
2. Generate $w_\Omega(\omega)$ and $w_\mathcal{S}(\mathbf{x}_s)$;
3. Extract $d_{\text{obs}}(\mathbf{x}_r,\omega;\mathbf{x}_s)$ on frequencies $\omega\in\Omega'$ and shots $\mathbf{x}_s\in\mathcal{S}'$ and stack them as $\mathbf{d}_{\text{obs}}$;
4. Solve Eq. (8) for $\mathbf{W}\mathbf{d}_{\text{cal}}$ on $\omega\in\Omega'$ and $\mathbf{x}_s\in\mathcal{S}'$ with randomized sources $w_\Omega(\omega)w_\mathcal{S}(\mathbf{x}_s)f(\omega)\delta(\mathbf{x}-\mathbf{x}_s)$;
5. Denote $\mathbf{W}\delta\mathbf{d} = \mathbf{W}\mathbf{d}_{\text{obs}} - \mathbf{W}\mathbf{d}_{\text{cal}}$;
6. $\boldsymbol{\alpha}_k = \operatorname{argmin} \frac{1}{2}\|\mathbf{W}\delta\mathbf{d} - \mathbf{W}\mathcal{F}\mathcal{D}\boldsymbol{\alpha}\|_2^2$ s.t. $\|\boldsymbol{\alpha}\|_1 \le \tau$;
7. Inverse block-wise transform $\delta\mathbf{m}_k = \mathcal{D}_k\boldsymbol{\alpha}_k$;
8. Learn $\mathbf{D}^{(k+1)}$ by **Algorithm 1** from R blocks of $\delta\mathbf{m}_k$;
9. Update $\mathbf{m}_{k+1} = \mathbf{m}_k + \delta\mathbf{m}$;
10. $\Delta_k = \|\mathbf{m}_{k+1} - \mathbf{m}_k\|_2/\|\mathbf{m}_k\|_2$;
11. $k \leftarrow k+1$;
end

Algorithm 2: Sparsity-Promoting FWI

In order to exploit sparsity, the model perturbation should be $\delta\mathbf{m} = \mathcal{D}\boldsymbol{\alpha}$ where $\mathcal{D}$ is a block-wise transform that converts each block of $\delta\mathbf{m}$ into sparse coefficients for a dictionary $\mathbf{D}$. In order to reduce data dimensionality, an ℓ_1-norm constraint is then placed on the sparse coefficients such that $\|\boldsymbol{\alpha}\|_1 \le \tau$. As a result, we propose a sparsity-promoting FWI based on compressive sensing that minimizes the following misfit

$$J(\boldsymbol{\alpha}) \triangleq \frac{1}{2}\|\mathbf{W}\delta\mathbf{d} - \mathbf{W}\mathcal{F}\mathcal{D}\boldsymbol{\alpha}\|_2^2 \quad \text{s.t.} \quad \|\boldsymbol{\alpha}\|_1 \le \tau \quad (12)$$

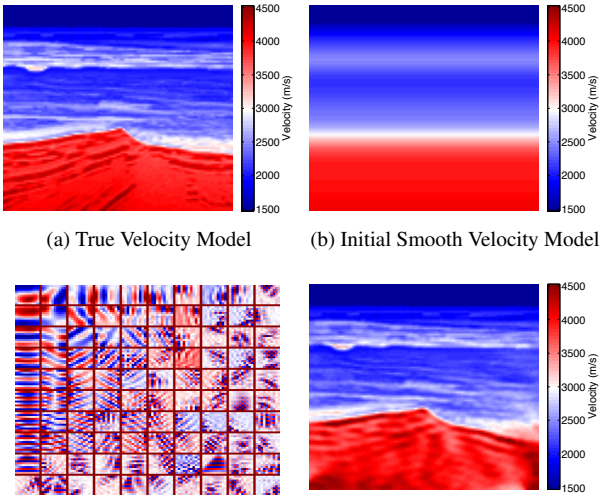
where $\mathbf{W} = (\mathbf{R}_{\Omega'}\text{diag}(\mathbf{w}_\Omega)) \otimes (\mathbf{R}_{\mathcal{S}'}\text{diag}(\mathbf{w}_\mathcal{S})) \otimes \mathbf{I}_{N_r} \in \mathbb{C}^{N'_\omega N'_s N_r \times N_\omega N_s N_r}$ is the random spatial-frequency subsampling matrix. Random vectors $\mathbf{w}_\mathcal{S}$ and $\mathbf{w}_\Omega$ randomize $f(\omega)\delta(\mathbf{x}-\mathbf{x}_s)$ on different shot positions $\mathbf{x}_s$ and frequencies ω. The restriction matrix $\mathbf{R}_{\Omega'}$ randomly selects N'_ω out of N_ω frequencies and $\mathbf{R}_{\mathcal{S}'}$ randomly selects N'_s out of N_s sources.

Algorithm 2 outlines the overall FWI optimization procedure which is initialized by a smooth model $\mathbf{m}_s$. For each iteration, we don't have to generate explicitly the whole matrix $\mathbf{W}$; instead, only two random vectors $\mathbf{w}_\mathcal{S}$ and $\mathbf{w}_\Omega$ are needed for random spatial-frequency modulations. We employ the limited-memory projected quasi-Newton (l-PQN) algorithm [16] to minimize $J(\boldsymbol{\alpha})$ because of its ability to project $\boldsymbol{\alpha}$

into an ℓ_1-norm ball of radius τ. Each optimized $\delta\mathbf{m}$ is divided into R blocks for online dictionary learning to update $\mathbf{D}$, which will be used in the next FWI iteration.

5. RESULTS

We test the proposed FWI method on two benchmark velocity models named "BG-Compass" and "Marmousi" whose true forms are shown in Figs. 2(a) and 3(a). Both models are scaled to $N_z \times N_x = 100 \times 100$ grids and cover a width of 17 km and a depth of 3.5 km. We deploy $N_r = 100$ receivers and generate $N_s = 100$ shots evenly spaced along the surface of the model to collect wave data for FWI. The shot source is a Ricker wavelet with $N_\omega = 256$ frequency components centered at 20 Hz. FWI starts from initial smooth models shown in Figs. 2(b) and 3(b). For every FWI iteration we pick data $\mathbf{d}_{\text{obs}}$ and $\mathbf{d}_{\text{cal}}$ from $N_s' = 10$ random shots and $N_\omega' = 16$ random frequencies. In practical FWI implementations, these $N_\omega' = 16$ random frequencies are equally chosen within 4 consecutive frequency bands between 2 Hz and 42 Hz to avoid local minima.



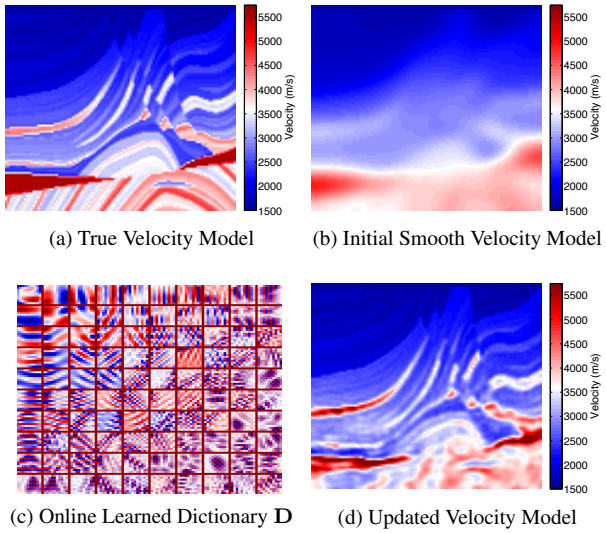
(a) True Velocity Model (b) Initial Smooth Velocity Model

(c) Online Learned Dictionary $\mathbf{D}$ (d) Updated Velocity Model

Fig. 2: FWI results for the "BG-Compass" model with velocity range of 1500 to 4500 m/s.

After we finish 20 FWI iterations on one frequency band, the more accurate model serves as the initial model for another 20 FWI iterations on the next higher frequency band. Thus, each evaluation of the compressed misfit $J(\boldsymbol{\alpha})$ is $(N_\omega N_s)/(N_\omega' N_s') = 160$ times cheaper than the evaluation of the full-data misfit $J(\delta\mathbf{m})$, implying that our reduced-data FWI could be roughly 160 times faster than the full-data FWI.

We compare the running time of one single FWI iteration on a Quad-core i7 desktop computer equipped with 16 GB RAM, and provide the times in Table 1 for one iteration. Multiple PDEs with different ω and $\mathbf{x}_s$ can be solved in parallel. Thus, only a few hours are needed to finish 80 iterations of



(a) True Velocity Model (b) Initial Smooth Velocity Model

(c) Online Learned Dictionary $\mathbf{D}$ (d) Updated Velocity Model

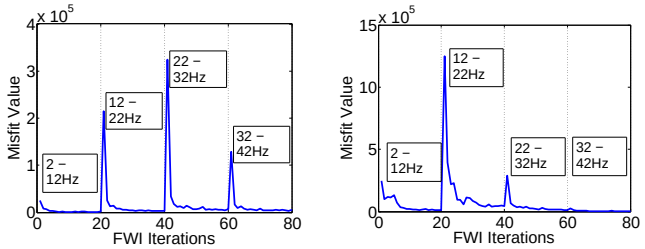
Fig. 3: FWI results for the "Marmousi" model with velocity range of 1500 to 5800 m/s.

FWI by using our reduced-data scheme, but the computation would be prohibitive if the full dataset were used.

Model	Reduced-data FWI	Full-data FWI
BG-Compass	290 s	41325 s
Marmousi	315 s	44312 s

Table 1: Comparison of running time for one FWI iteration

The orthonormal dictionary $\mathbf{D}$ is initialized as an $N \times N = 100 \times 100$ DCT matrix (for 10×10 blocks) to provide sufficient block-wise sparsity on $\delta\mathbf{m}$ at the first iteration and then updated by the optimized $\delta\mathbf{m}$ from each following iteration. After 80 iterations, the online updated orthonormal dictionaries are shown in Figs. 2(c) and 3(c). The updated FWI results are given in Figs. 2(d) and 3(d), which clearly shows the validity of this method on complex velocity models. Fig. 4 shows the misfit value versus the number of FWI iterations when using dictionary-based sparsity regularization. We can see that the proposed method has quick convergence within each frequency band.



(a) BG-Compass (b) Marmousi

Fig. 4: Misfit convergence versus iterations

6. REFERENCES

[1] M. Elad and M. Aharon, "Image denoising via sparse and redundant representations over learned dictionaries," *Image Processing, IEEE Transactions on*, vol. 15, no. 12, pp. 3736–3745, Dec 2006.

[2] L. Zhu, E. Liu, and J. H. McClellan, "Seismic data denoising through multiscale and sparsity-promoting dictionary learning," *GEOPHYSICS*, vol. 80, no. 6, pp. WD45–WD57, 2015.

[3] M. Filipovic, I. Kopriva, and A. Cichocki, "Inpainting color images in learned dictionary," in *Signal Processing Conference (EUSIPCO), 2012 Proceedings of the 20th European*, Aug 2012, pp. 66–70.

[4] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An Algorithm for Designing Overcomplete Dictionaries for Sparse Representation," *Signal Processing, IEEE Transactions on*, vol. 54, no. 11, pp. 4311–4322, Nov 2006.

[5] R. Rubinstein, M. Zibulevsky, and M. Elad, "Double sparsity: Learning sparse dictionaries for sparse signal approximation," *Signal Processing, IEEE Transactions on*, vol. 58, no. 3, pp. 1553–1564, March 2010.

[6] R. Rubinstein, T. Peleg, and M. Elad, "Analysis K-SVD: A dictionary-learning algorithm for the analysis sparse model," *Signal Processing, IEEE Transactions on*, vol. 61, no. 3, pp. 661–677, Feb 2013.

[7] S. Chen, S. A. Billings, and W. Luo, "Orthogonal least squares methods and their application to non-linear system identification," *International Journal of control*, vol. 50, no. 5, pp. 1873–1896, 1989.

[8] S. G. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *Signal Processing, IEEE Transactions on*, vol. 41, no. 12, pp. 3397–3415, Dec 1993.

[9] J. A. Tropp and A. C. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," *Information Theory, IEEE Transactions on*, vol. 53, no. 12, pp. 4655–4666, Dec 2007.

[10] O. G. Sezer, O. G. Guleryuz, and Y. Altunbasak, "Approximation and compression with sparse orthonormal transforms," *Image Processing, IEEE Transactions on*, vol. 24, no. 8, pp. 2328–2343, Aug 2015.

[11] T. Blumensath and M. E. Davies, "Iterative hard thresholding for compressed sensing," *Applied and Computational Harmonic Analysis*, vol. 27, no. 3, pp. 265 – 274, 2009.

[12] P. H. Schönemann, "A generalized solution of the orthogonal procrustes problem," *Psychometrika*, vol. 31, no. 1, pp. 1–10, 1966.

[13] O. Bousquet and L. Bottou, "The tradeoffs of large scale learning," in *Advances in Neural Information Processing Systems 20*, J.c. Platt, D. Koller, Y. Singer, and S. Roweis, Eds., pp. 161–168. MIT Press, Cambridge, MA, 2007.

[14] J. Virieux and S. Operto, "An overview of full-waveform inversion in exploration geophysics," *GEOPHYSICS*, vol. 74, no. 6, pp. WCC1–WCC26, 2009.

[15] M. Nieto-Vesperinas, *Scattering and diffraction in physical optics*, Wiley New York, 1991.

[16] M. W. Schmidt, E. Berg, M. P. Friedlander, and K. P. Murphy, "Optimizing costly functions with simple constraints: A limited-memory projected quasi-newton algorithm," in *International Conference on Artificial Intelligence and Statistics*, April 2009, pp. 456–463.