Randomized dimension reduction for full-waveform inversion

Felix J. Hermann

SLIM
University of British Columbia
Randomized dimension reduction for full-waveform inversion

Felix J. Hermann, Peyman Moghaddam, and Xiang Li
Impediments

Full-waveform inversion (FWI) is suffering from

- multimodality, i.e., a multitude of velocity models explain data
- local minima, i.e., requirement of an accurate initial model
- over- and underdeterminacy

Curse of dimensionality for $d>2$

- requires implicit (Helmholtz) solvers to address bandwidth
- # RHS's makes computation of gradients prohibitively expensive
Wish list

An inversion technology that

• is based on a time-harmonic PDE solver, which is easily parallelizable, and scalable to 3D

• does not require multiple iterations with all data

• removes the linearly increasing costs of implicit solvers for increasing numbers of frequencies & RHS’s

• allows for a dimensionality reduction commensurate the model’s complexity
Key technologies

Numerical linear algebra [Erlangga & Nabben,’08, Erlanga & FJH ’08-’09]
  • multi-level Krylov preconditioner for Helmholtz

Simultaneous sources [Beasley, ’98, Neelamani et. al., ’08]
  • supershots

Simultaneous sources [Krebs et.al., ’09, Operto et. al., ’09, FJH et.al., ’08-10’]

Stochastic optimization & machine learning [Bersekas, ’96]
  • stochastic gradient decent

Compressive sensing [Candès et.al, Donoho, ’06]
  • sparse recovery & randomized subsampling
Full-waveform inversion

**Single-source / frequency PDE-constrained optimization problem:**

\[
\min_{u \in U, m \in M} \frac{1}{2} \| p - Du \|_2^2 \quad \text{subject to} \quad H[m]u = q
\]

- \(p\) = Single-source and single-frequency data
- \(D\) = Detection operator
- \(u\) = Solution of the Helmholtz equation
- \(H\) = Discretized monochromatic Helmholtz system
- \(q\) = Unknown seismic source
- \(m\) = Unknown model, e.g. \(c^{-2}(x)\)
Unconstrained problem

For each separate source $q$ solve the unconstrained problem:

$$\min_{m \in \mathcal{M}} \frac{1}{2} \| p - \mathcal{F}[m, q] \|_2^2$$

with $q$ a single source function and

$$\mathcal{F}[m, q] = DH^{-1}[m]q$$

- Gradient updates involve 3 PDE solves
- Increased matrix bandwith of Helmholtz discretization makes scaling to 3D challenging for direct methods...
Preconditioner

\[ \lambda(H) \xrightarrow{M^-} \lambda(HM^{-1}) \xrightarrow{Q} \lambda(HM^{-1}Q) \]

- **shifts** the eigenvalues to **positive** half plane - solves **indefiniteness**
- **clusters** eigenvalues **near** one - solves **ill-conditioning**
Scaling

**Number of iteration**
- MKMG
- MG

**Number of matrix-vec multiplies**

![Graph showing frequency vs. number of iterations and matrix-vec multiplies.](image)

<table>
<thead>
<tr>
<th>Frequency, Hz</th>
<th>Unit memory</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>$751 \times 201$</td>
</tr>
<tr>
<td>15</td>
<td>$1501 \times 401$</td>
</tr>
<tr>
<td>20</td>
<td>$2001 \times 534$</td>
</tr>
</tbody>
</table>

**Graph Details**
- **X-axis**: Frequency, Hz
- **Y-axis**: Number of iteration and number of matrix-vec multiplies
Example: Marmousi

Hard Model
1500-4000 m/s

Smooth Model
Simulations

Real part of $u$, freq = 10 Hz, 9 grid/wavelength

Real part of $u$, freq = 10 Hz, 18 grid/wavelength
Convergence

- Hard model
- Smooth model
Observations

*Implicit* solver that

- converges for *high* frequencies
- *scales* & embarrassingly *parallelizable*
- same of order *complexity* as TDFD
- *trivial* implementation of *imaging* conditions

*But* it’s *complexity* grows *linearly* with the number of frequencies and sources...
**Full-waveform inversion**

Multiexperiment PDE-constrained optimization problem:

$$\min_{U \in \mathcal{U}, m \in \mathcal{M}} \frac{1}{2} \| P - DU \|_{2,2}^2 \quad \text{subject to} \quad H[m]U = Q$$

- **P** = Total multi-source and multi-frequency data volume
- **U** = Solution of the Helmholtz equation
- **H** = Discretized multi-frequency Helmholtz system
- **Q** = Unknown seismic sources
Simultaneous sources

Randomized superposition of sequential source functions creates a supershot

[Beasley, '98, Neelamani et. al., '08, Krebs et.al., '09, Operto & Virieux, '09-'10, FJH et.al., '08-10']
Simultaneous shot at 5 Hz

sequential source wavefield

simultaneous source wavefield

Notice the increased wavenumber contend
Increased wavenumber contend leads to improved image/gradient updates ...
Supershots

sub sampler

\[
\mathbf{RM} = \begin{bmatrix}
\mathbf{R}^\Sigma_1 \otimes \mathbf{I} \otimes \mathbf{R}^\Omega_1 \\
\vdots \\
\mathbf{R}^\Sigma_{n_s'} \otimes \mathbf{I} \otimes \mathbf{R}^\Omega_{n_s'}
\end{bmatrix}
\]

random phase encoder

\[
\left( \mathbf{F}_2^* \operatorname{diag} \left( e^{i\theta} \right) \otimes \mathbf{I} \right) \mathbf{F}_3,
\]

\( n_s' \ll n_s \)

The sparsifying transform:

Aside from proper CS sampling the recovery from simultaneous simulations depends on a sparsifying transform that compresses seismic data, is fast and reasonably incoherent with the CS sampling matrix. We accomplish this by defining the sparsity transform as the Kronecker product between the discrete curvelet transform and the discrete wavelet transform along the source-coordinate and the time coordinate, respectively:

\[
\mathbf{S} = \mathbf{C} \otimes \mathbf{W}
\]

with \( \mathbf{C} \) and \( \mathbf{W} \) the curvelet and wavelet transform matrices, respectively.
Dimensionality reduction

\[
\begin{cases}
Q = D^* \\
HU = Q \\
\end{cases}
\quad \text{single shots}
\quad \leftrightarrow
\quad \begin{cases}
Q = \mathbf{D}^* \\
HU = Q \\
\end{cases}
\quad \text{simul. shots}

\mathbf{P} := \mathbf{RMDU} = \mathbf{DU}

Reduced system

\[
\min_{\mathbf{U}, \mathbf{m}} \quad \frac{1}{2} \| \mathbf{P} - \mathbf{DU} \|_2^2 \quad \text{subject to} \quad \mathbf{H} [\mathbf{m}] \mathbf{U} = \mathbf{Q}
\]

[Neelamani et.al., '08, Krebs et. al., '09, FJH et. al., '08-'10]
Reduced gradient

Replace gradient updates for all sequential sources:

\[ m^{k+1} := m^k - \eta_k \sum_{i=1}^{N} \nabla f(m^k, q^i) \quad \text{with} \quad q^i := i^{th} \text{Col}(Q) \]

by

\[ m^{k+1} := m^k - \eta_k \sum_{i=1}^{N'} \nabla f(m^k, q^i) \quad \text{with} \quad q^i := (RM)_i Q \]

- \( N' \ll N \) number of multifrequency simultaneous experiments
- creates incoherent “Gaussian” simultaneous-source crosstalk
Stochastic optimization

Stochastic “batch” gradient decent:

\[
m^{k+1} := m^k - \eta_k \frac{1}{n} \sum_{i=1}^{n} \nabla f(m^k, q^i) \quad \text{with} \quad q^i := (RM)_{i} Q
\]

- for \( n \to \infty \), the updates become deterministic

- prohibitively expensive

[Robbins and Monro, 1951]

[Bersekas, ’96, Nemirovski, ’09]
Stochastic Average Approximation (SAA)

Approximate expectation with *ensemble* average

\[
E\{f(m, \hat{q})\} \approx \frac{1}{N} \sum_{i=1}^{N} \min_{m} f(m, q^i) \quad \text{with} \quad q^i := (RM)_i Q
\]

- for \( n \to \infty \) becomes equality
- well studied and known as Monte Carlo sampling
- slow but embarrassingly parallel

[Bersekas, ’96, Nemirovski, ’09]
Renewals [Krebs et.al, ’09]

Use different *supershots* for each (gradient) update:

\[ m^{k+1} := m^k - \eta_k \nabla f(m^k, Q^k) \quad \text{with} \quad Q^k := (RM)_k Q \]

- uses *different* random $RM$ for each *iteration*
- cheap but introduces more “noise” and does not converge...
Stochastic Approximation (SA)

Stochastic “online” gradient descent with mini batches:

\[
\begin{align*}
\mathbf{m}^{k+1} &= \mathbf{m}^k - \eta_k \nabla f(\mathbf{m}^k, Q^k) \quad \text{with} \quad Q^k := (RM)_k Q \\
\mathbf{m}^{k+1} &= \frac{1}{k+1} \left( \sum_{i=1}^{k} \mathbf{m}^i + \mathbf{m}^{k+1} \right)
\end{align*}
\]

- averages over the past iterations and converges
- reasonable well understood
- not understood for (quasi) Newton

Marmoussi model

original model

initial model
Full-waveform inversion

recovered model

L-BFGS

recovered
stochastic gradient
Speed up

Full scenario:
- 113 sequential shots with 50 frequencies
- 18 iterations of l-BFGS (90=5*18 Helmholtz solves)

Reduced scenario:
- 16 randomized simultaneous shots with 4 frequencies
- 40 iterations of SA (2.27=16*4/(113*50)*40*5 solves)

Speed up of 40 X or > week vs 8 h on 32 CPUs
Observations

SAA can be applied to *supercharge* FWI

But it

- requires care with line searches
- does not extend to Newton methods
- is relatively poorly understood mathematically

What does *Compressive Sensing* have to offer?
Compressive recovery

Consider “Newton” updates of the reduced system as sparse recovery problems

• feasible because of the reduced system size
• imposes transform-domain sparsity on the updates
• corresponds to sparse linearized inversions
Sparse linearized inversion

Invert adjoint of the Jacobian, i.e., linearized Born approximation

\[ \delta d^k = K[m^k, Q] \delta m^k \text{ with } \delta d^k = \text{vec}(P - \mathcal{F}[m^k, Q]) \]

with \( \delta m^k = S^H \tilde{x} \) obtained via sparse inversion

\[ \tilde{x} = \arg \min_x ||x||_1 \text{ subject to } b = Ax \]

where

\[ A := RMKS^H = KS^H \text{ and } b = \delta d^k := RM\delta d^k \]
Initial model

SLIM

Lateral (× 15 meters)

Depth (× 15 meters)

0 50 100 150 200 250

20 40 60 80 100 120

2000 2500 3000 3500 4000 4500

SLIM
Linearized sparse inversion

30 simultaneous shots 10 random frequencies

true reflectivity

sparse recovery with wavelets
Linearized sparse inversion

20 simultaneous shots 10 random frequencies

true reflectivity  sparse recovery with wavelets
Linearized sparse inversion

10 simultaneous shots 5 random frequencies

true reflectivity

sparse recovery with wavelets
## Linearized sparse inversion

<table>
<thead>
<tr>
<th>Subsample ratio</th>
<th>0.015</th>
<th>0.006</th>
<th>0.002</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n_f' / n_s'$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>17.44 (1.32)</td>
<td>11.66 (0.78)</td>
<td>6.83 (-0.14)</td>
</tr>
<tr>
<td>1</td>
<td>17.53 (1.59)</td>
<td>11.89 (1.05)</td>
<td>7.19 (0.15)</td>
</tr>
<tr>
<td>0.2</td>
<td>18.22 (1.68)</td>
<td>12.11 (1.32)</td>
<td>7.46 (0.27)</td>
</tr>
<tr>
<td>Speed up (×)</td>
<td>66</td>
<td>166</td>
<td>500</td>
</tr>
</tbody>
</table>

Errors for “migration” in parentheses
Observations

Reconstruct model updates
- from randomized subsamplings
- with correct amplitudes (like Newton updates)

Removed the “curse of dimensionality” by reducing the number of RHS’s

Recovery quality depends on degree of subsampling

What about stochastic optimization?
Reduced batch

10 X
Stochastic mini batch

66X
Stochastic average approximation

$66 \times$
Stochastic mini batch

Algorithm 1: FWI without renewal of CS experiments.

Result: Estimate for the model \( \tilde{m} \)

\[
\begin{align*}
\mathbf{m} &\leftarrow \mathbf{m}_0; \\
\mathbf{j} &\leftarrow 0; \\
\mathbf{Q} &\leftarrow (\mathbf{RM})\mathbf{Q}; \\
\text{while } &\| \mathbf{P} - \mathbf{F}[\mathbf{m}, \mathbf{Q}] \|_2^2 \geq \epsilon \\
&\left\{ \\
\mathbf{j} &\leftarrow \mathbf{j} + 1; \\
\mathbf{A} &\leftarrow \mathbf{K}[\mathbf{m}, \mathbf{Q}]\mathbf{S}^*; \\
\mathbf{\delta P} &\leftarrow \mathbf{P} - \mathbf{F}[\mathbf{m}, \mathbf{Q}]; \\
\mathbf{\delta x} &\leftarrow \arg \min_{\mathbf{\delta x}} \| \mathbf{\delta x} \|_1 \text{ s.t. } \| \mathbf{\delta P} - \mathbf{A}\mathbf{\delta x} \|_2 \leq \sigma; \\
\mathbf{m} &\leftarrow \mathbf{m} + \mathbf{S}^*\mathbf{\delta x}; \\
\end{align*}
\]

end
Algorithm 1: FWI with renewal of CS experiments.

Result: Estimate for the model \( \tilde{m} \)

\[
\begin{align*}
\text{m} &\leftarrow \text{m}_0; \quad \text{// initial model} \\
j &\leftarrow 0; \quad \text{// loop counter} \\
Q &\leftarrow (\text{RM})_iQ; \quad \text{// Draw random sim. shot} \\
\text{while } &\| \text{P} - \mathcal{F}[\text{m}, Q] \|_{2,2}^2 \geq \epsilon \text{ do} \\
&j := j + 1; \quad \text{// increase counter} \\
&A &\leftarrow K[\text{m}, Q]S^*; \quad \text{// Calculate Jacobian} \\
&\delta \text{P} &\leftarrow \text{P} - \mathcal{F}[\text{m}, Q]; \quad \text{// Calculate residual} \\
&\delta \tilde{x} &\leftarrow \arg \min_{\delta x} \| \delta x \|_{\ell_1} \quad \text{s.t. } \| \delta \text{P} - A\delta x \|_2 \leq \sigma; \quad \text{// Sparse recovery} \\
&m &\leftarrow \text{m} + S^* \delta \tilde{x}; \quad \text{// Compute model update} \\
&Q &\leftarrow (\text{RM})_iQ; \quad \text{// Draw random sim. shot} \\
\text{end}
\end{align*}
\]

SA mini batch
Update with sparse recoveries

10 iterations

with renewal (SNR = 20.8)

without renewal (SNR = 19.9)
Updates

1st update

9th update
Conclusions

Reconstruct “Newton-like” updates from randomized subsamplings

Remove the “curse of dimensionality”

Algorithms have parallel pathways

Results are encouraging but rigorous theory still lacking
Acknowledgments

This work was in part financially supported by the Natural Sciences and Engineering Research Council of Canada Discovery Grant (22R81254) and the Collaborative Research and Development Grant DNOISE (334810-05).

This research was carried out as part of the SINBAD II project with support from the following organizations: BG Group, BP, Petrobras, and WesternGeco.
Thank you

slim.eos.ubc.ca
Further reading

Compressive sensing
- Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information by Candes, 06.
- Compressed Sensing by D. Donoho, ’06

Simultaneous acquisition
- A new look at simultaneous sources by Beasley et. al., ’98.
- Changing the mindset in seismic data acquisition by Berkhout ’08.

Simultaneous simulations, imaging, and full-wave inversion:
- Phase encoding of shot records in prestack migration by Romero et. al., ’00.
- Efficient Seismic Forward Modeling using Simultaneous Random Sources and Sparsity by N. Neelamani et. al., ’08.
- Compressive simultaneous full-waveform simulation by FJH et. al., ’09.
- Fast full-wavefield seismic inversion using encoded sources by Krebs et. al., ’09
- Randomized dimensionality reduction for full-waveform inversion by FJH & X. Li, ’10

Stochastic optimization and machine learning:
- A Stochastic Approximation Method by Robbins and Monro, 1951
- Neuro-Dynamic Programming by Bersekas, ’96
- Robust stochastic approximation approach to stochastic programming by Nemirovski et. al., ’09
- Stochastic Approximation and Recursive Algorithms and Applications by Kushner and Lin
- Stochastic Approximation approach to Stochastic Programming by Nemirovski