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

SLIM
University of British Columbia

[Symes, '08]

SLIM

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 [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

SLIM

Full-waveform inversion

Single-source / frequency PDE-constrained optimization problem:

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

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

Unconstrained problem

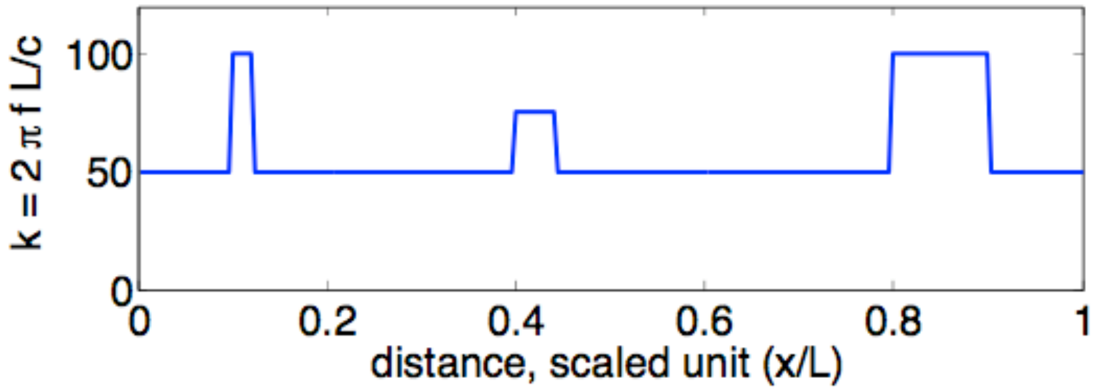
For each ***separate*** source **q** solve the **unconstrained problem:**

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

with **q** a single source function and

$$\mathcal{F}[\mathbf{m},\mathbf{q}] = \mathbf{D}\mathbf{H}^{-1}[\mathbf{m}]\mathbf{q}$$

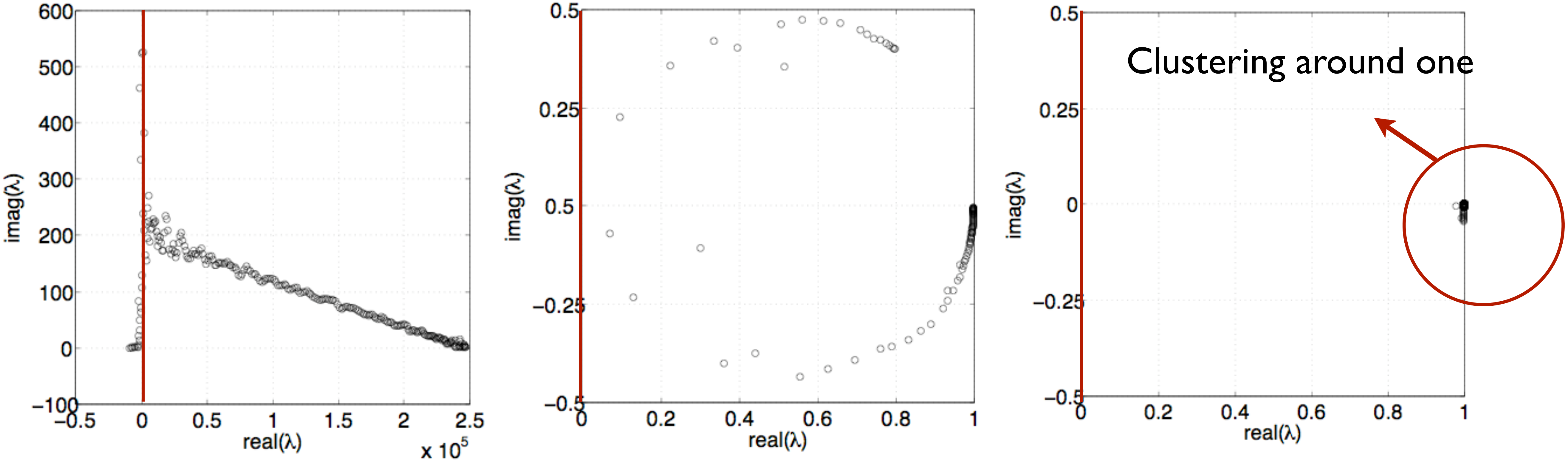
- Gradient updates involve 3 PDE solves
- Increased matrix bandwith of Helmholtz discretization makes *scaling* to 3D *challenging* for *direct* methods...



[Erlanga, Nabben, ’08]
[Erlanga and F.J.H, ‘08]

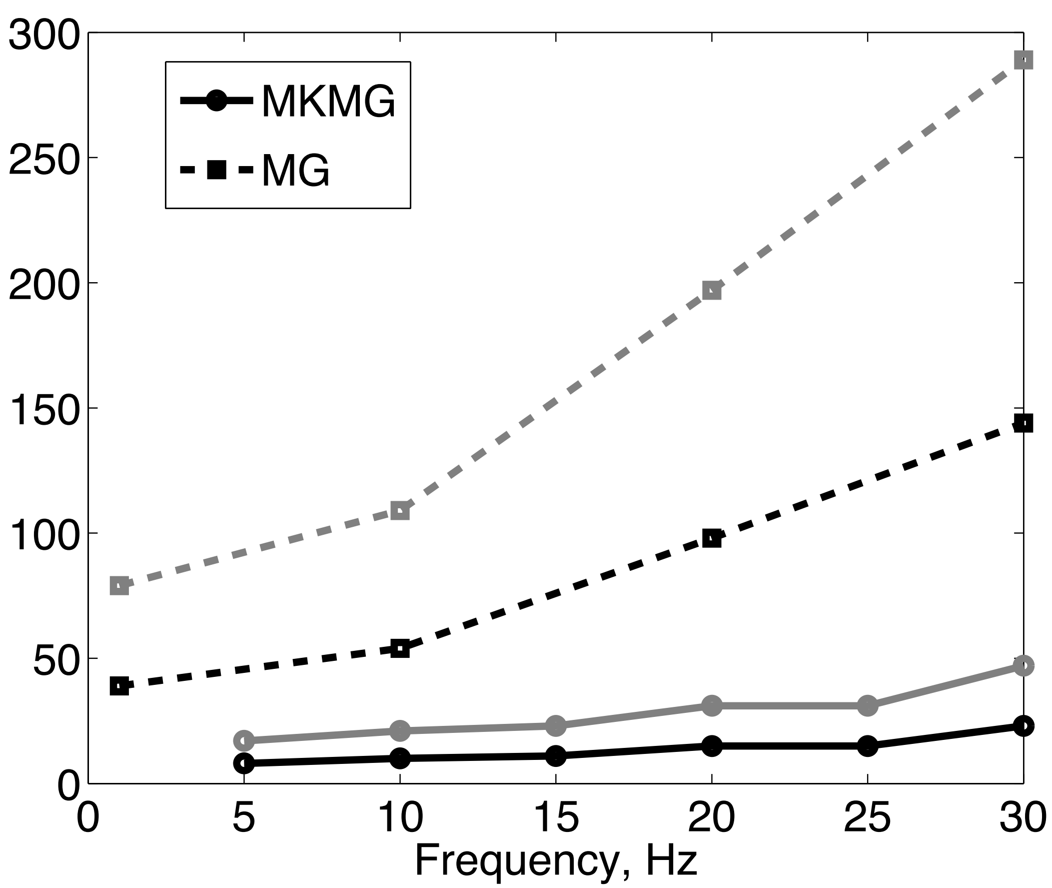
Preconditioner

$$\lambda(\mathcal{H}) \xrightarrow{\mathcal{M}^{-}} \lambda(\mathcal{H}\mathcal{M}^{-1}) \xrightarrow{\mathcal{Q}} \lambda(\mathcal{H}\mathcal{M}^{-1}\mathcal{Q})$$

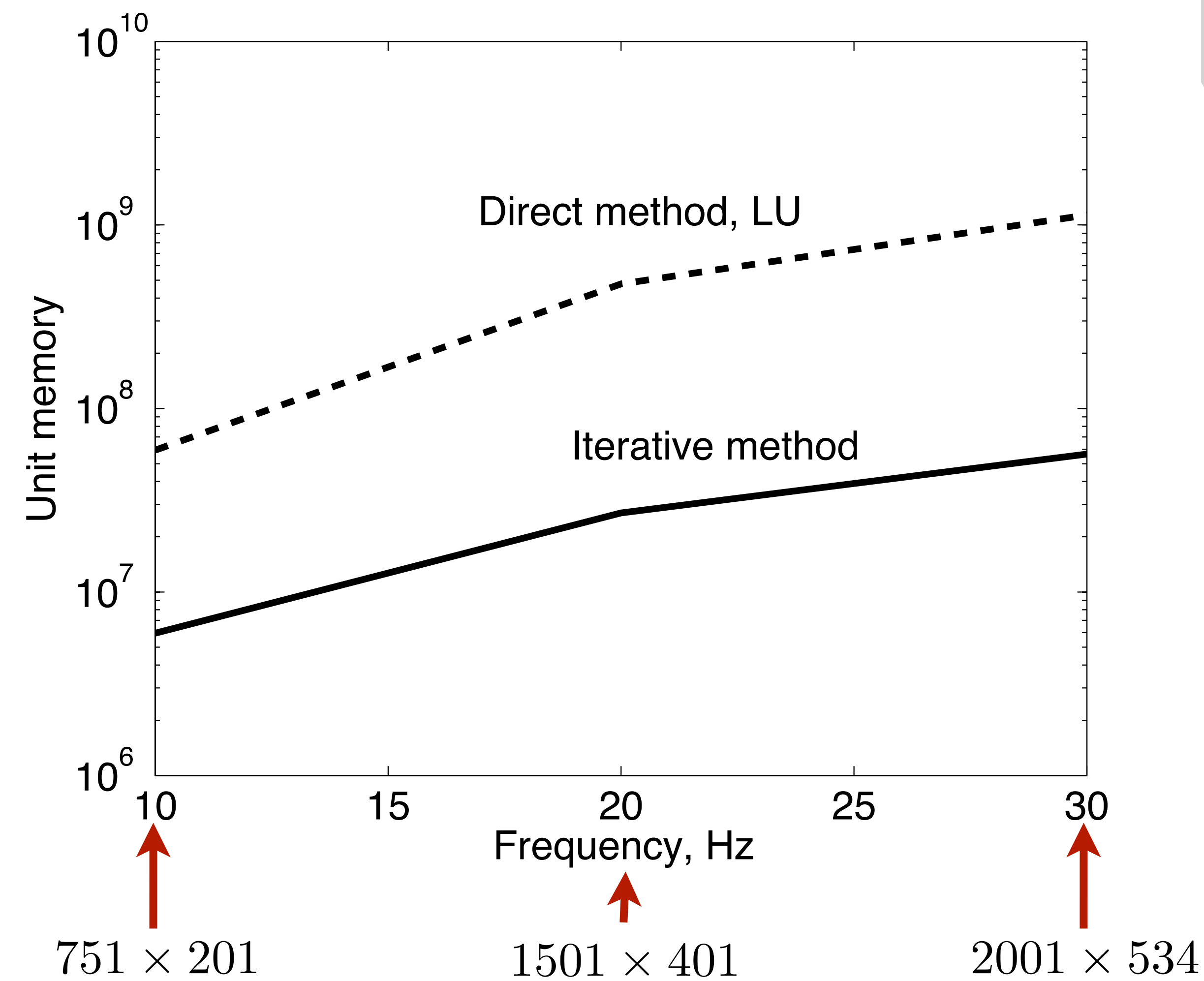


- *shifts* the eigenvalues to *positive* half plane - solves *indefiniteness*)
- *clusters* eigenvalues *near* one - solves *ill-conditioning*

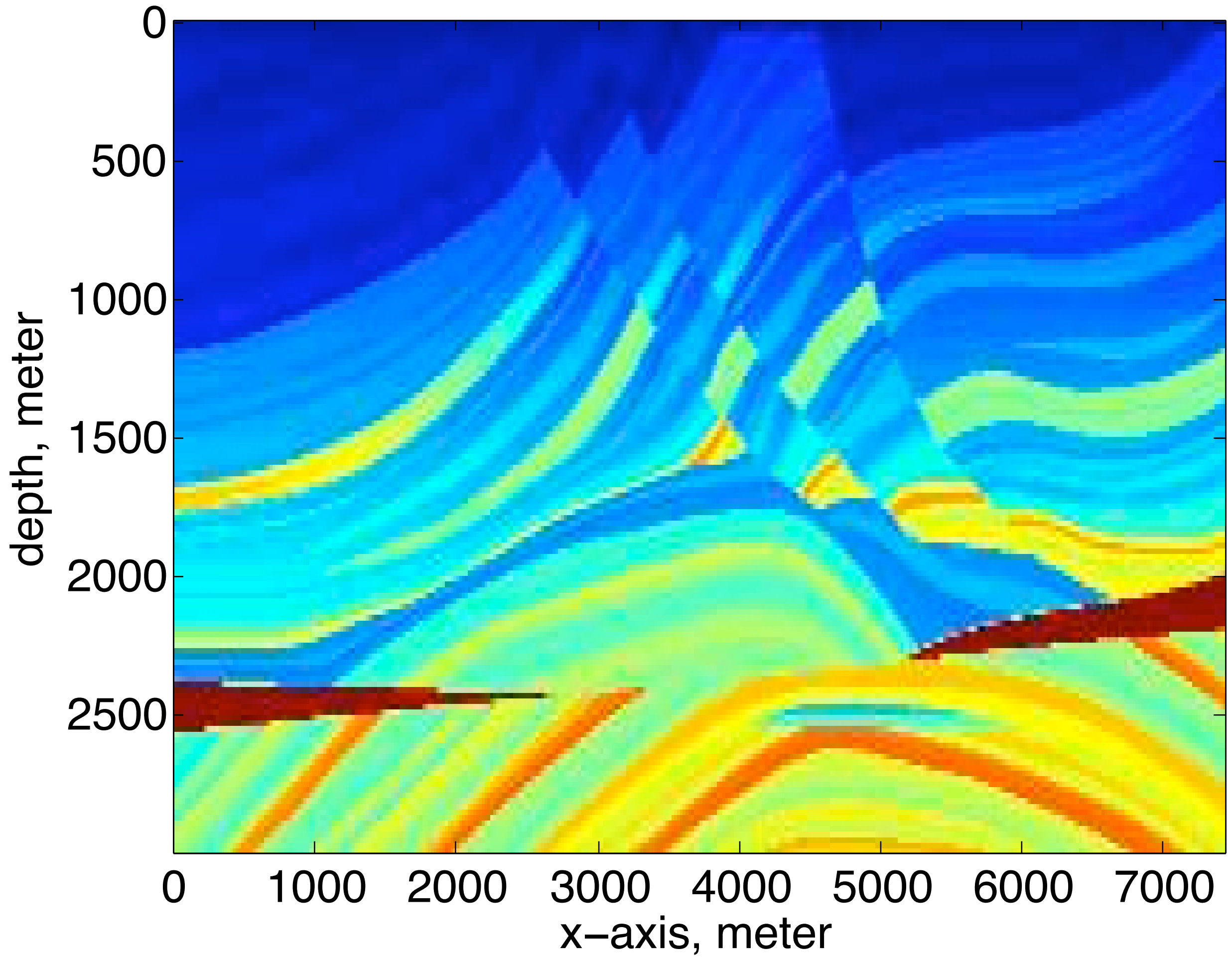
Scaling



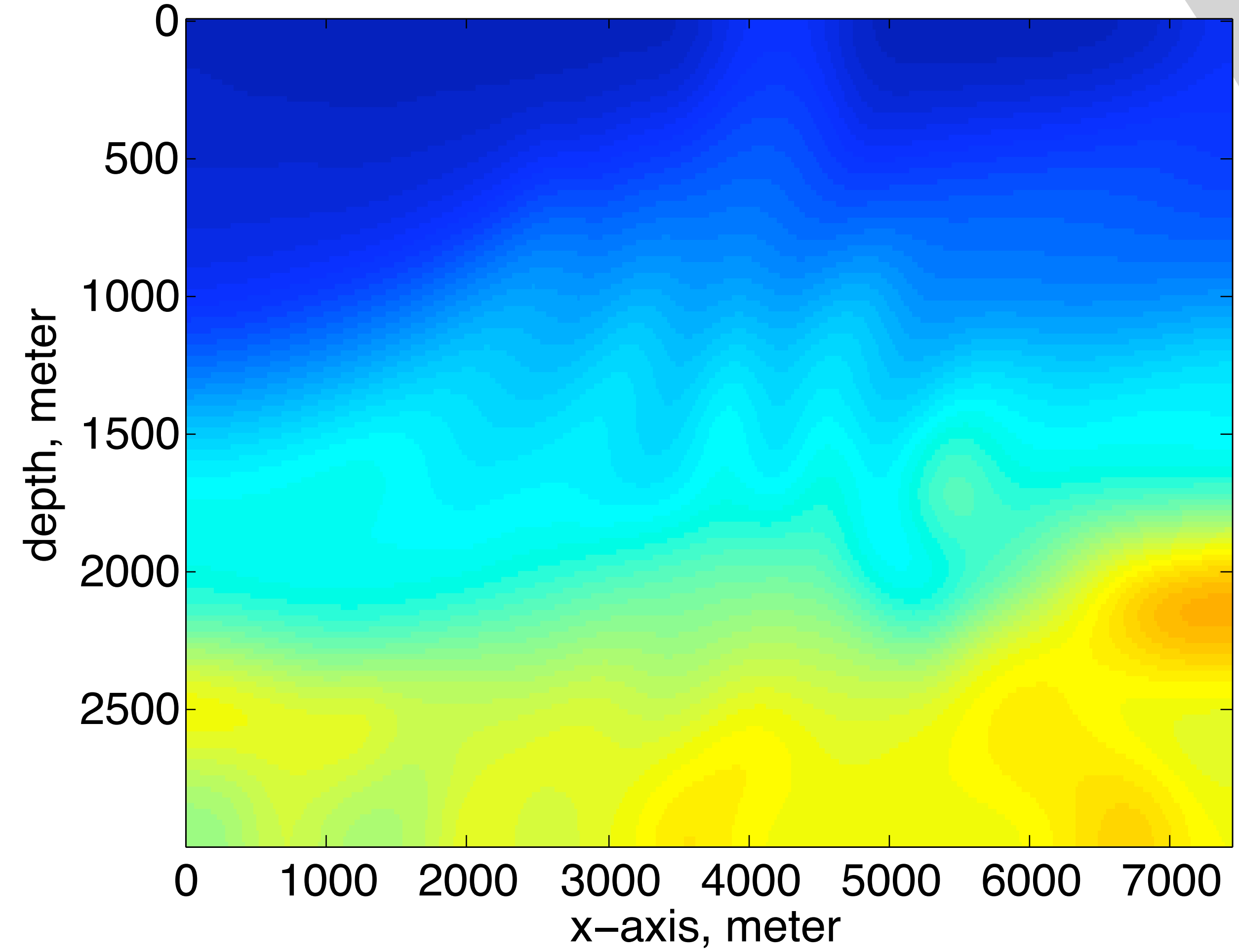
Number of iteration
Number of matrix-vec multiplies



Example: Marmousi

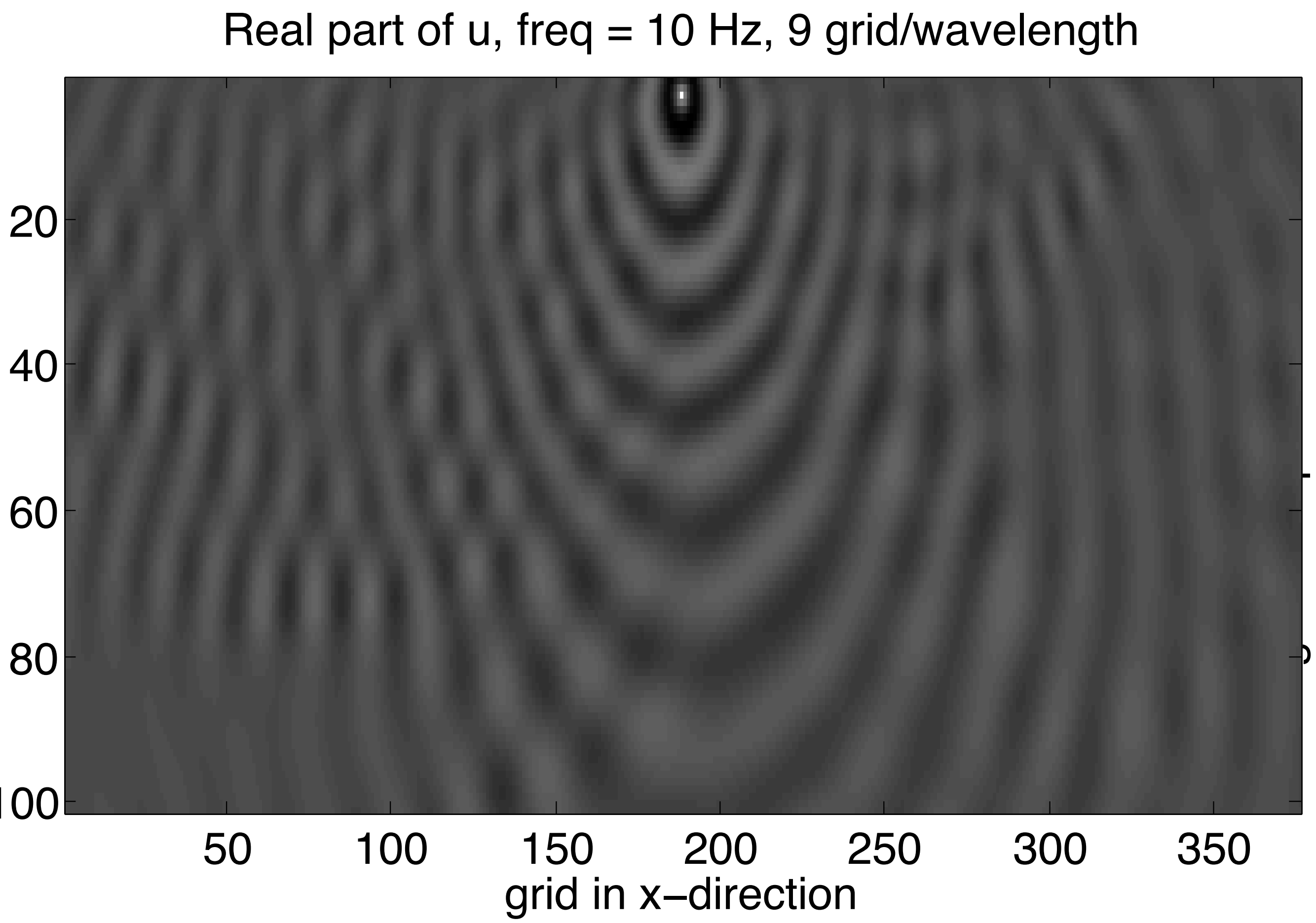


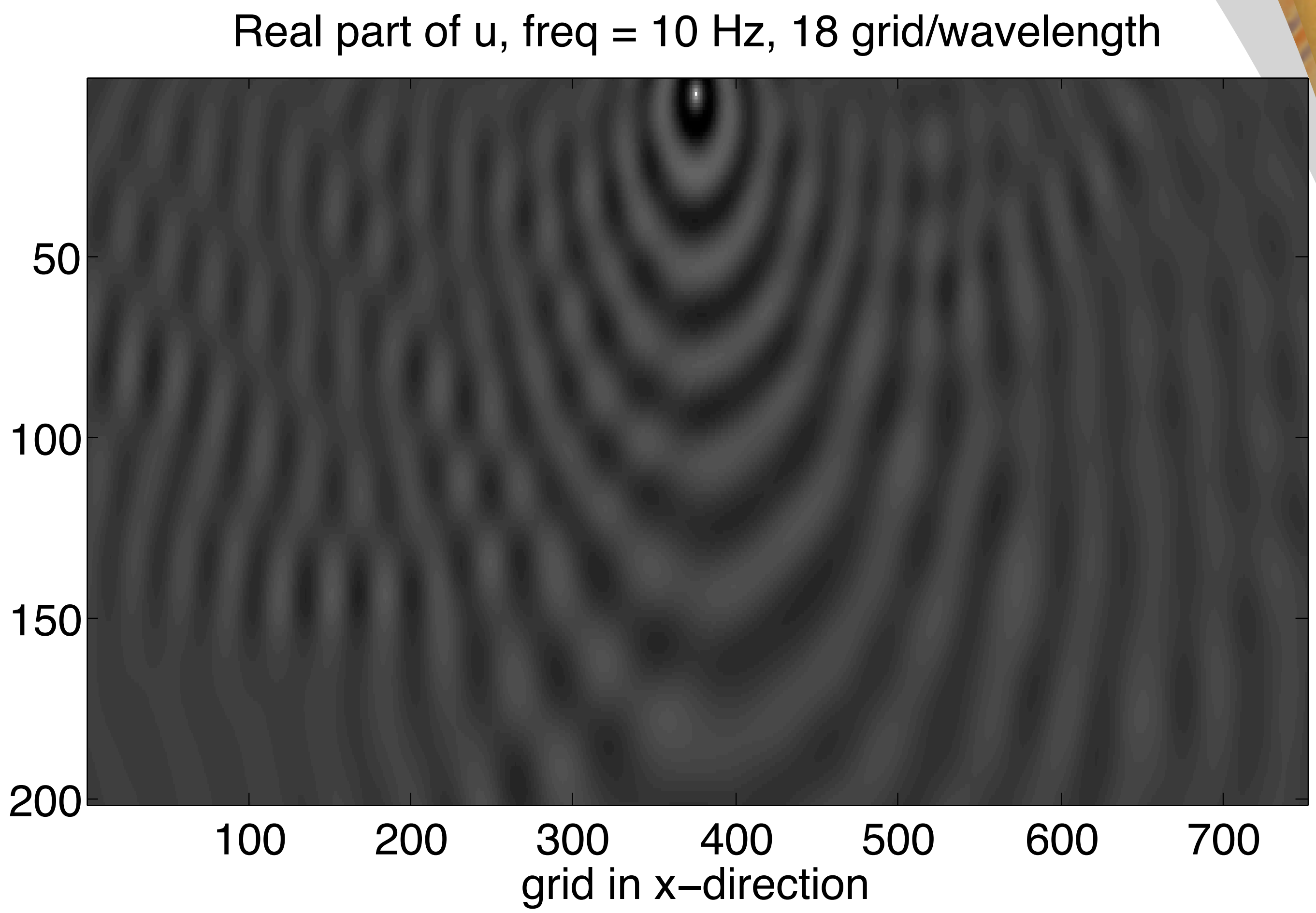
Hard Model
1500-4000 m/s



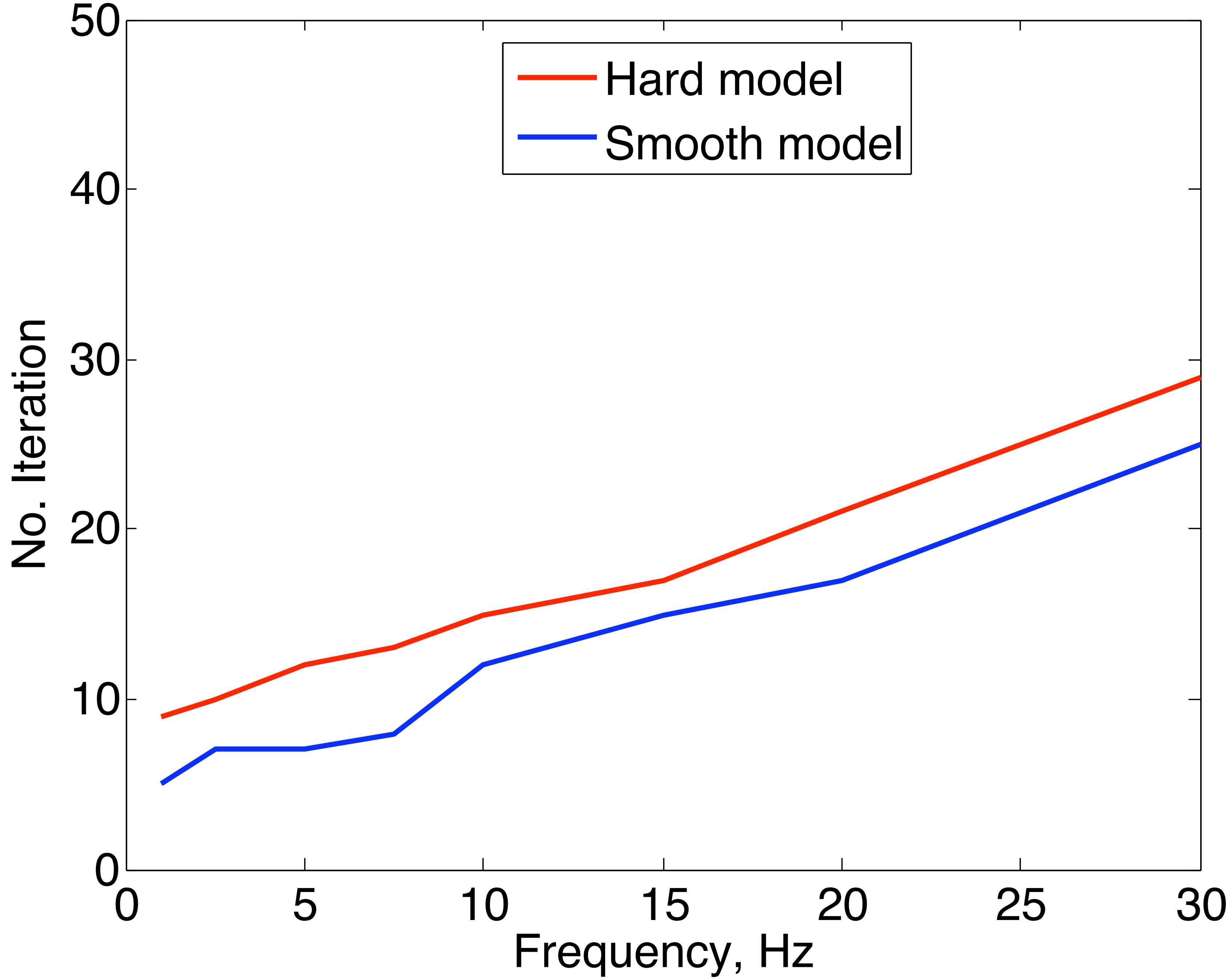
Smooth Model

Simulations





Convergence



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

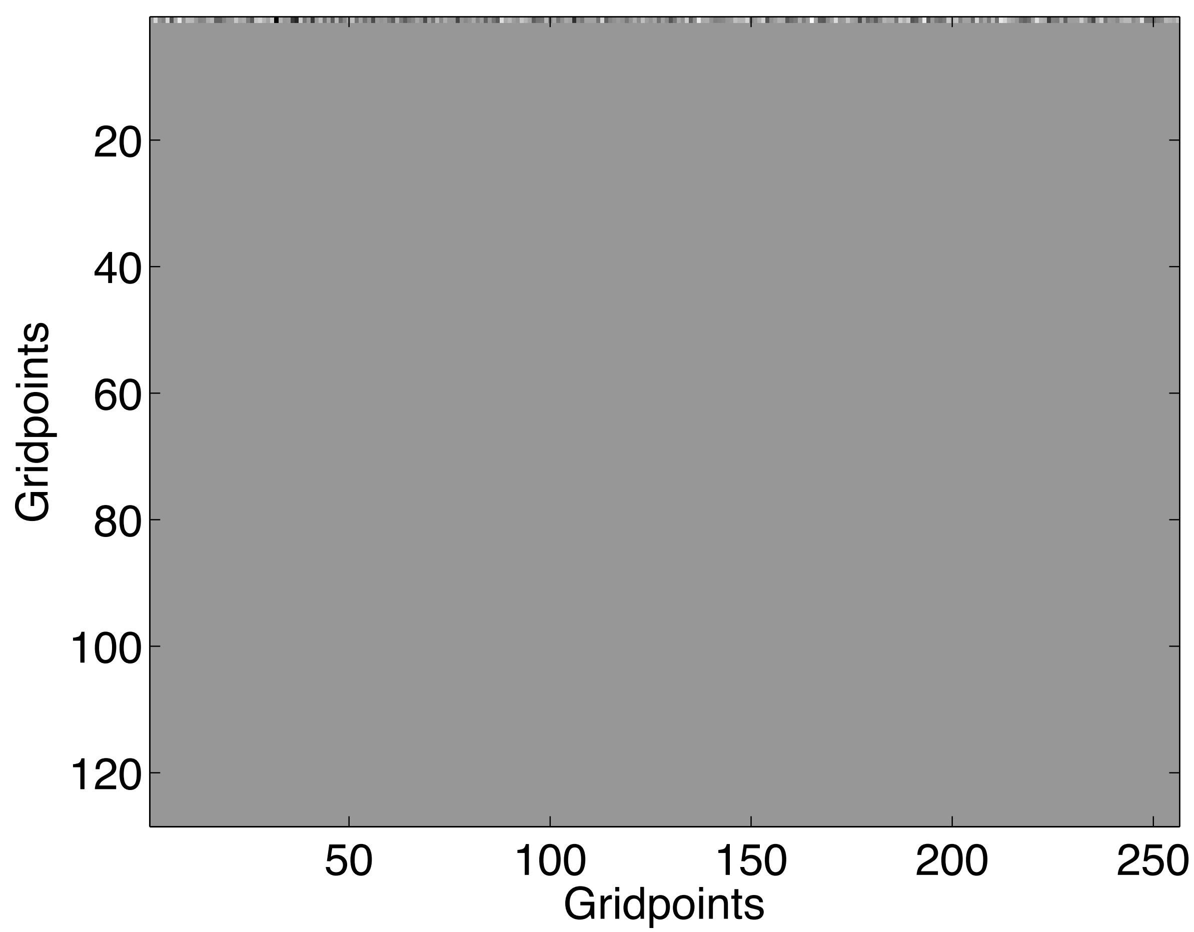
Multiexperiment PDE-constrained optimization problem:

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

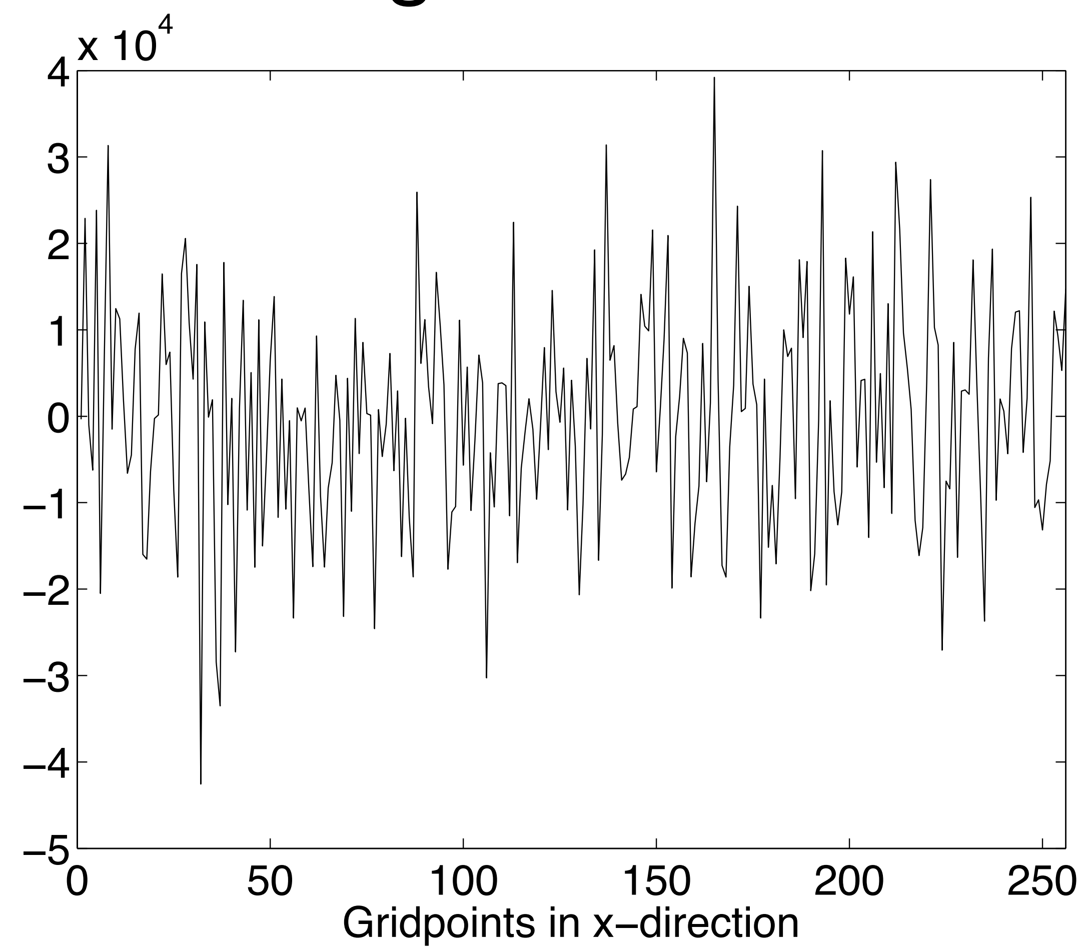
$$\begin{aligned}
\mathbf{P} &= \text{Total multi-source and multi-frequency data volume}\\
\mathbf{U} &= \text{Solution of the Helmholtz equation}\\
\mathbf{H} &= \text{Discretized multi-frequency Helmholtz system}\\
\mathbf{Q} &= \text{Unknown seismic sources}
\end{aligned}$$

Simultaneous sources

simultaneous source



Randomized amplitudes along the shot line

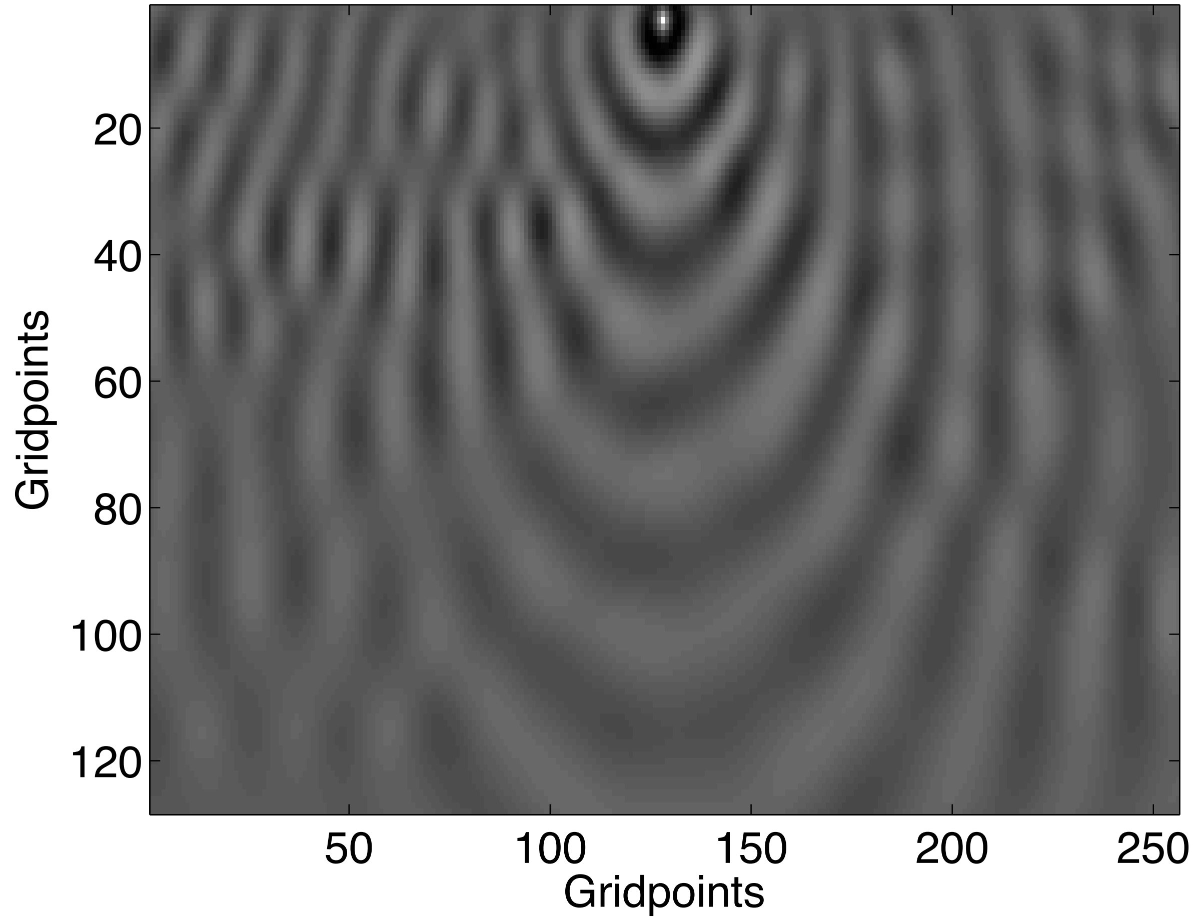


Randomized superposition of *sequential* source functions creates a *supershot*

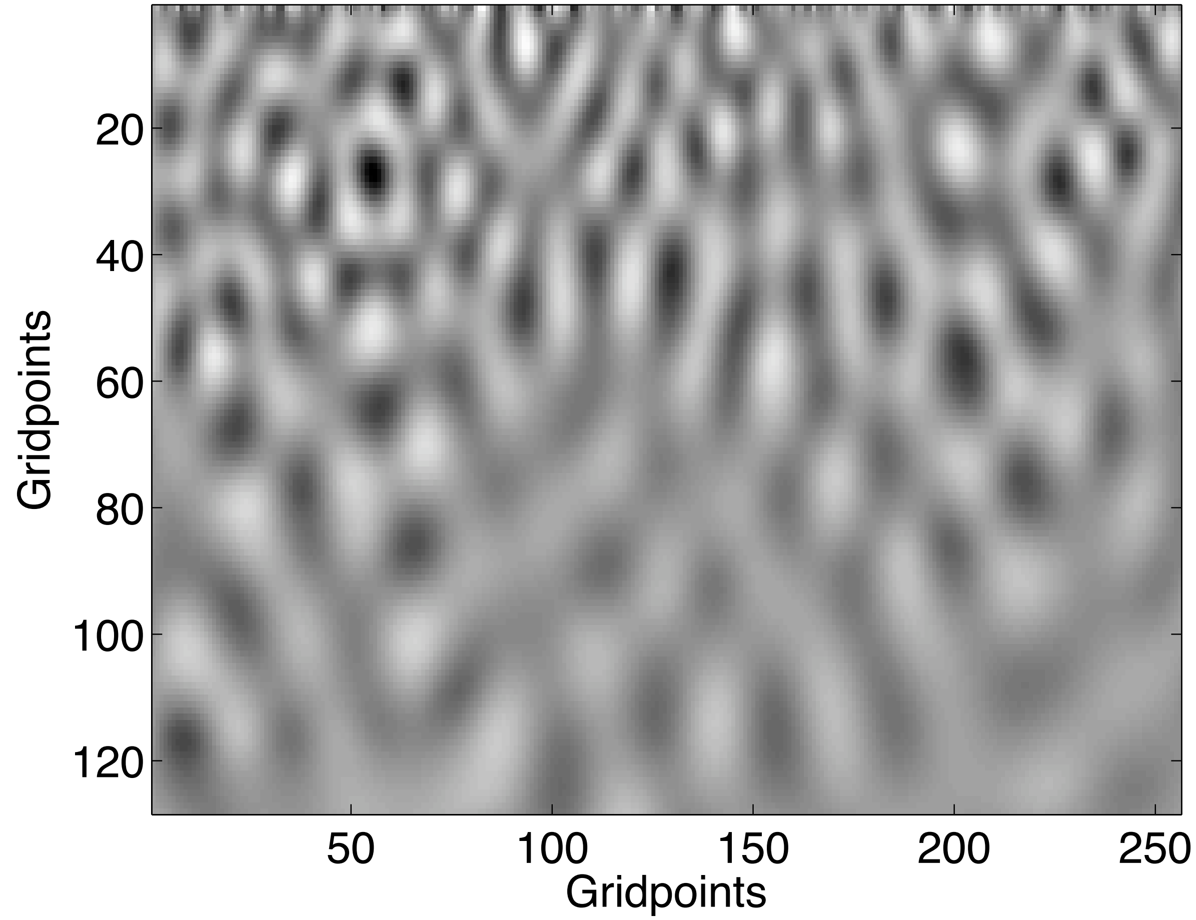
[Beasley, '98, Neelamani et. al,, '08, Krebs et.al., '09, Operto & Virieux, '09-'10, FJH et.al., '08-10']

Simultaneous shot at 5 Hz

sequential source wavefield



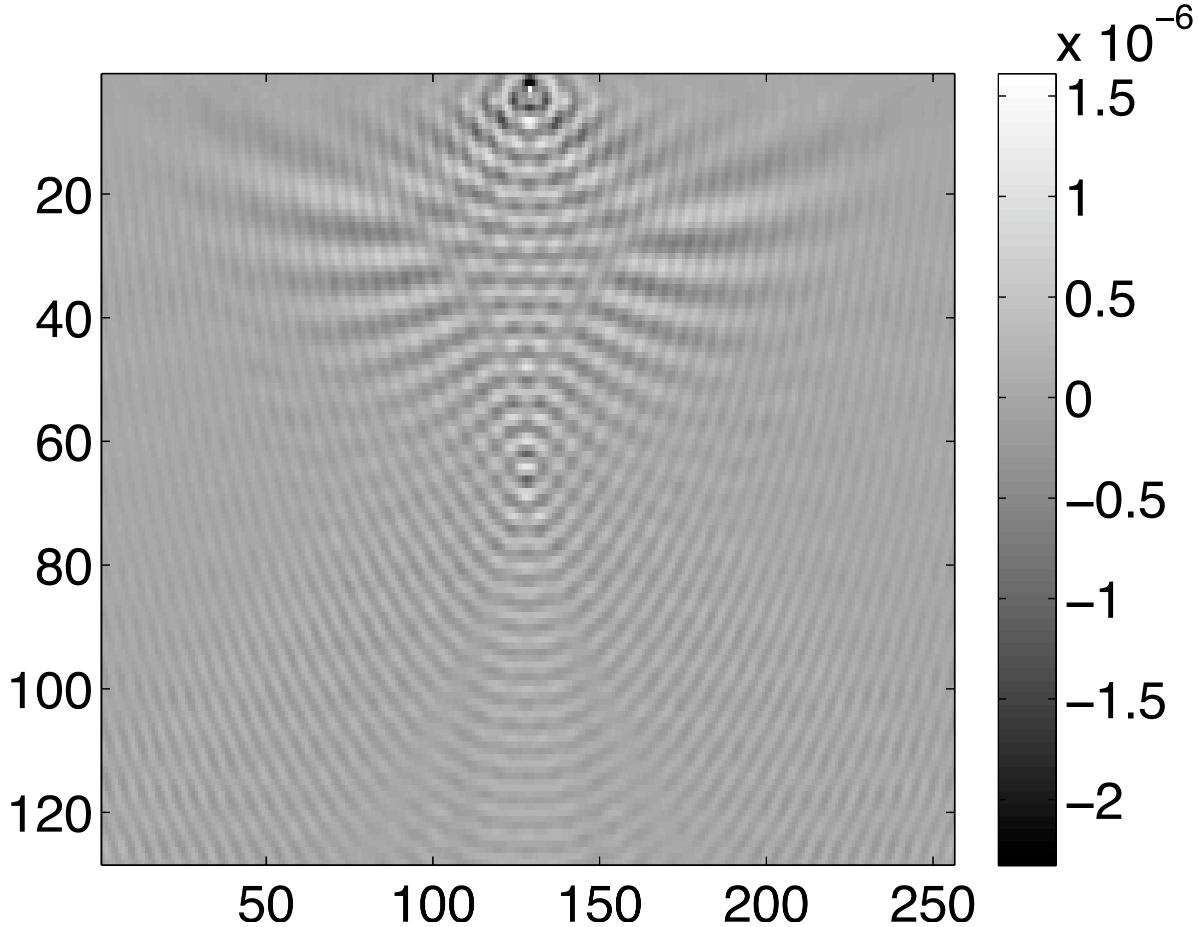
simultaneous source wavefield



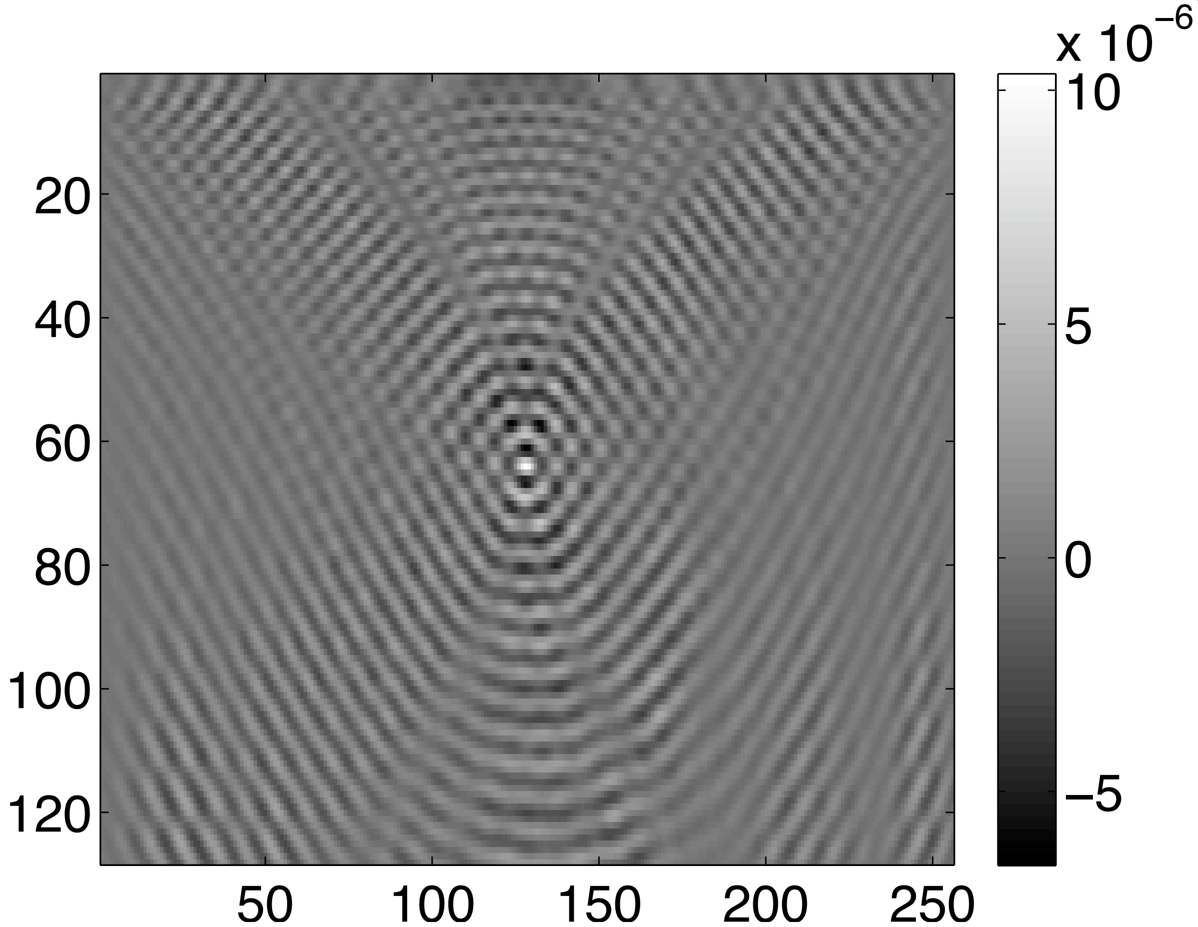
Notice the *increased wavenumber contend*

Image

sequential source image



simultaneous source image

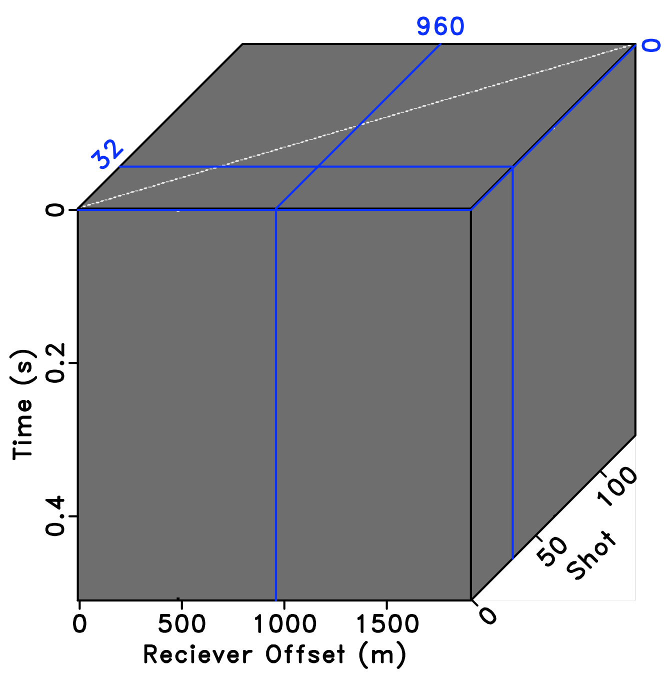


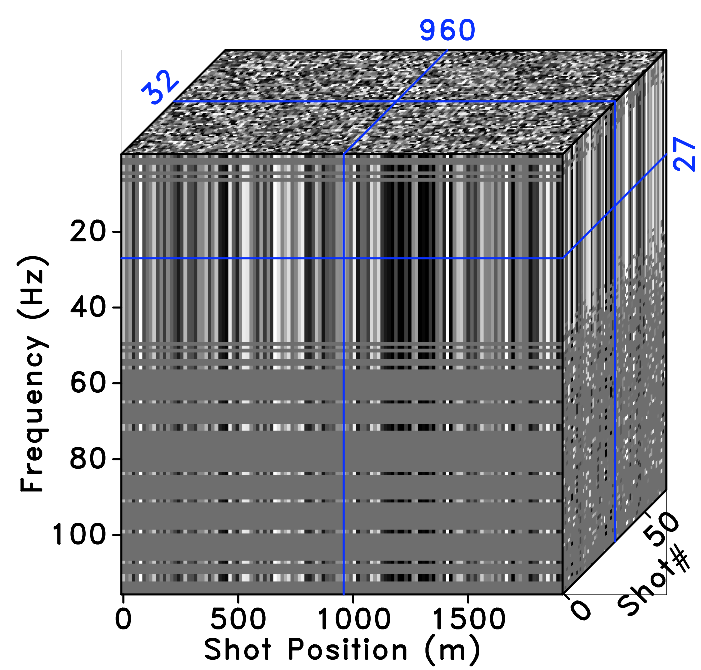
Increased wavenumber contend leads to improved image/ gradient updates ...

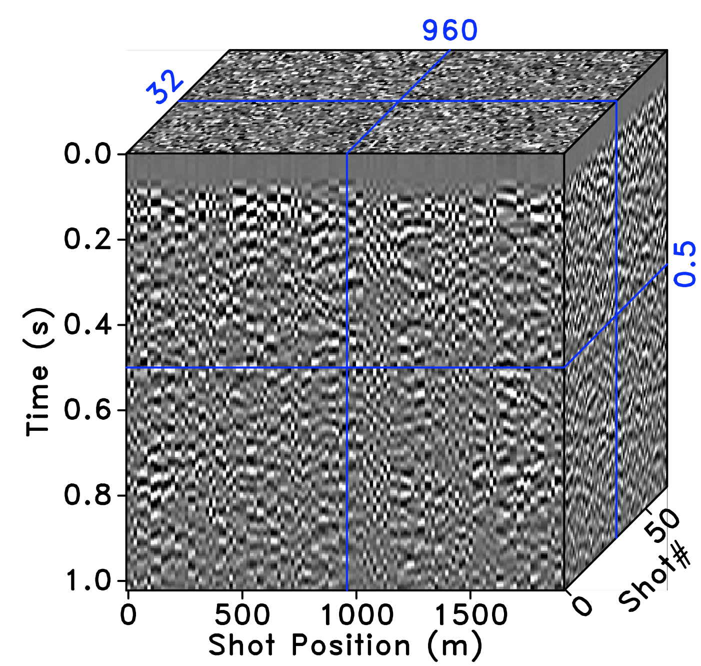
Supershots

$$\mathbf{RM} = \overbrace{\begin{bmatrix} \mathbf{R}_1^{\Sigma} \otimes \mathbf{I} \otimes \mathbf{R}_1^{\Omega} \\ \vdots \\ \mathbf{R}_{n_{s'}}^{\Sigma} \otimes \mathbf{I} \otimes \mathbf{R}_{n_{s'}}^{\Omega} \end{bmatrix}}^{\text{sub sampler}} \overbrace{\left(\mathbf{F}_2^* \operatorname{diag}\left(e^{\hat{i}\boldsymbol{\theta}}\right) \otimes \mathbf{I}\right) \mathbf{F}_3}^{\text{random phase encoder}}, \qquad n_s' \ll n_s$$

separated source







DQ → RM → DQ → $\mathbf{H}^{-1}$

SLIM

Dimensionality reduction

$$\begin{cases} \mathbf{Q} = \mathbf{D}^* & \underbrace{\mathbf{s}}_{\text{single shots}} \\ \mathbf{H}\mathbf{U} = \mathbf{Q} \end{cases} \Longleftrightarrow \begin{cases} \underline{\mathbf{Q}} = \underline{\mathbf{D}}^* & \underbrace{\underline{\mathrm{RMs}}}_{\text{simul. shots}} \\ \underline{\mathbf{HU}} = \underline{\mathbf{Q}} \end{cases}$$

$$\underline{\mathbf{P}} := \mathrm{RM}\mathbf{DU} = \underline{\mathbf{DU}}$$

Reduced system

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

[Neelamani et.al., ’08, Krebs et. al., ’09, FJH et. al., ’08-’10]

Reduced gradient

Replace gradient updates for *all sequential sources*:

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

by

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

- *N'*<<*N* number of *multifrequency simultaneous* experiments
- creates *incoherent* "Gaussian" simultaneous-source *crosstalk*

SLIM

Stochastic optimization

[Robbins and Monro, 1951]

Stochastic "batch" gradient decent :

$$\underline{\mathbf{m}}^{k+1} := \underline{\mathbf{m}}^{k} - \eta_k \frac{1}{n} \sum_{i=1}^{n} \nabla f(\underline{\mathbf{m}}^{k}, \underline{\mathbf{q}}^{i}) \quad \text{with} \quad \underline{\mathbf{q}}^{i} := (\mathbf{RM})_i \, \mathbf{Q}$$

- for $n \to \infty$, the updates become *deterministic*
- *prohibitively* expensive

[Bersekas, '96, Nemirovski, '09]

Stochastic Average Approximation (SAA)

Approximate expectation with *ensemble* average

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

- for $n \rightarrow \infty$ becomes equality
- well studied and known as Monte Carlo sampling
- slow but embarrassingly parallel

[Bersekas, ’96, Nemirovski, ’09]

SLIM

Renewals [Krebs et.al, '09]

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

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

- uses *different* random **RM** for each *iteration*
- cheap but introduces more "noise" and does not converge...

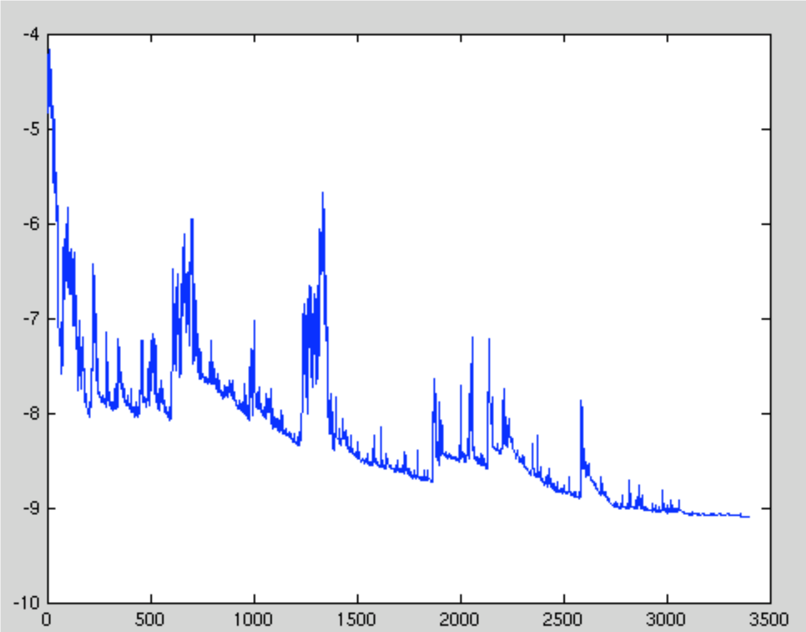
Stochastic [Bersekas, '96] Approximation (SA)

Stochastic "online" gradient descent with *mini batches* :

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

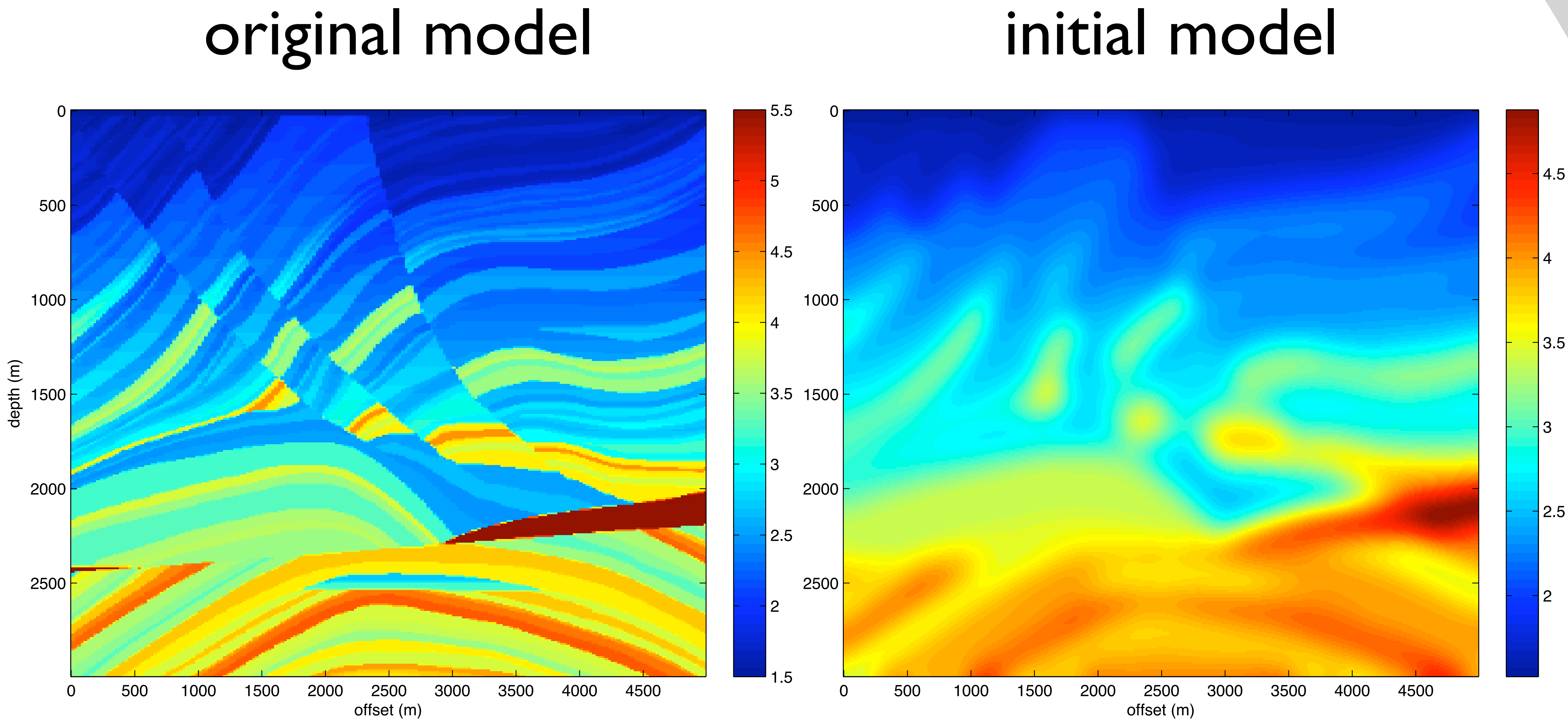
$$\mathbf{m}^{k+1} = \frac{1}{k+1}\left(\sum_{i=1}^{k} \mathbf{m}^i + \underline{\mathbf{m}}^{k+1}\right)$$

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



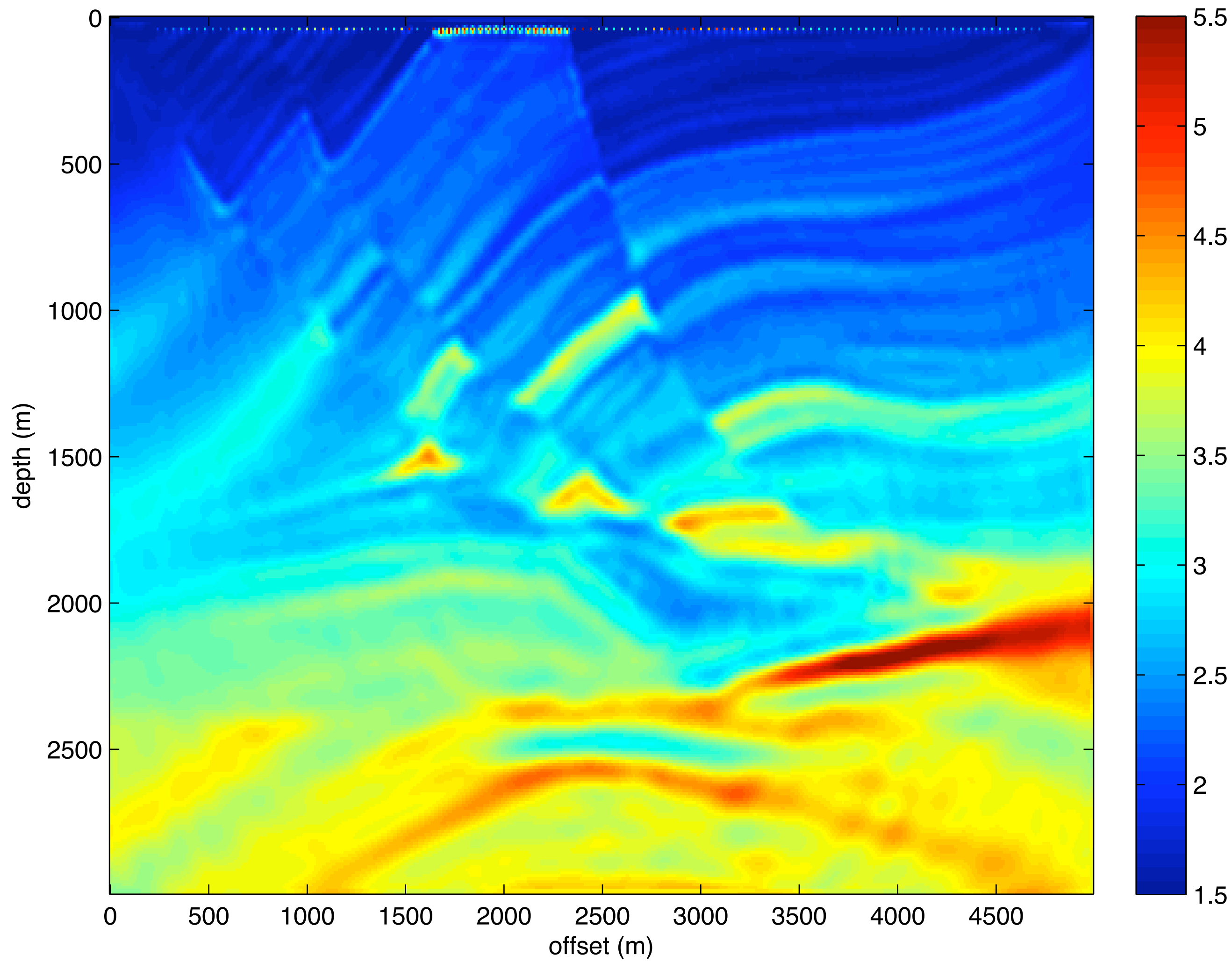
source: http://en.wikipedia.org/wiki/Stochastic_gradient_descent

Marmoussi model

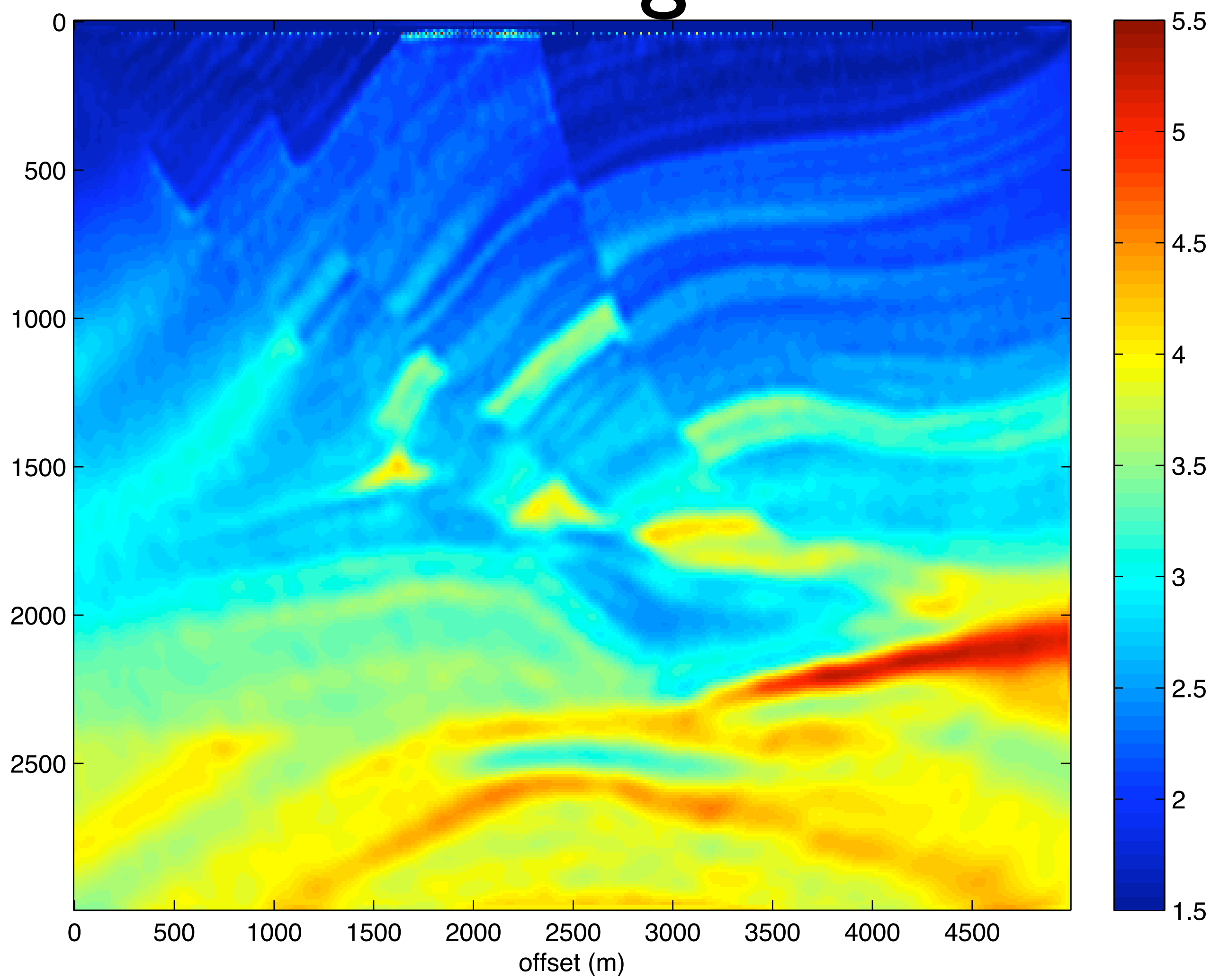


Full-waveform inversion

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

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

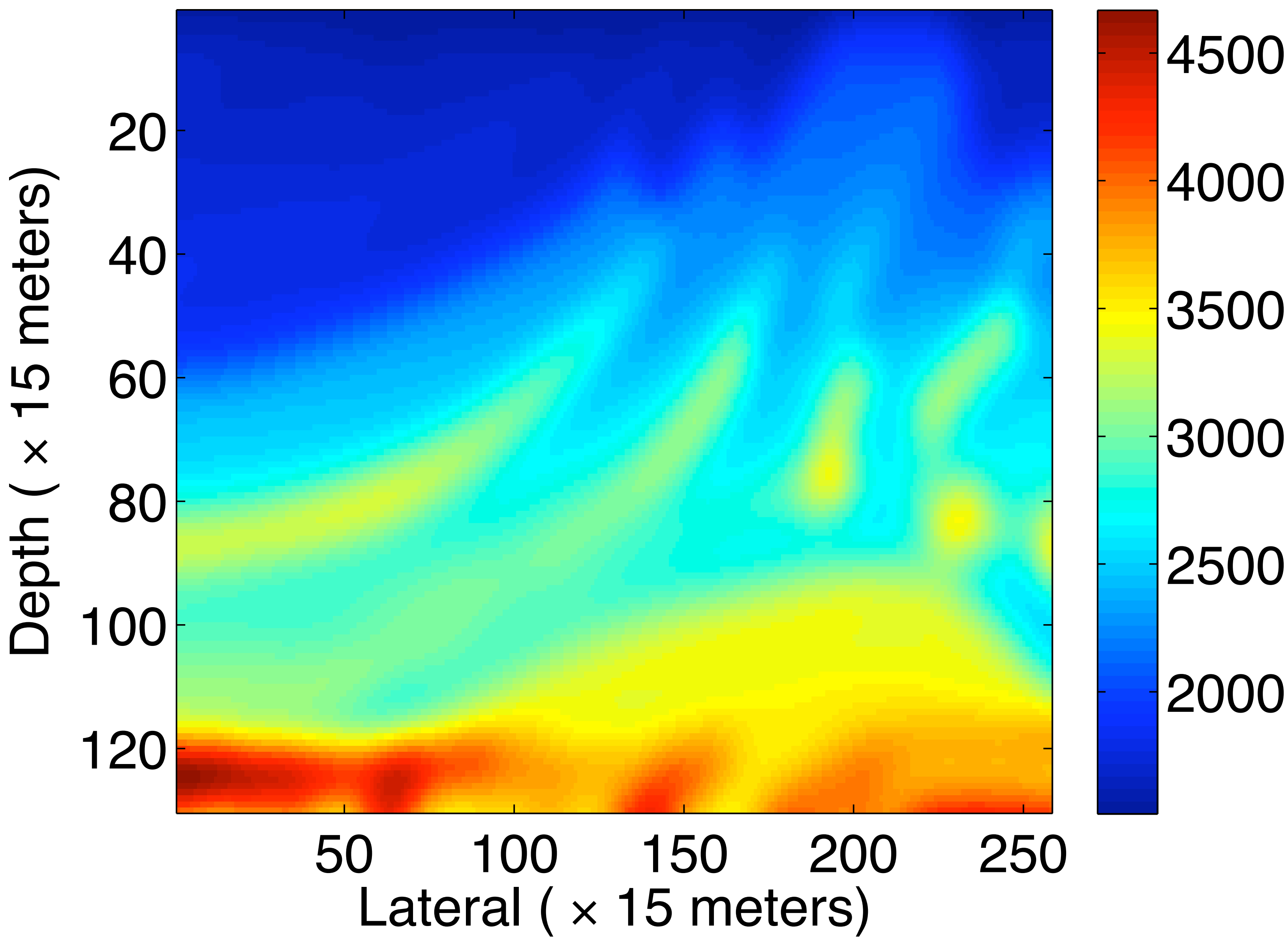
with $\delta\mathbf{m}^k = \mathbf{S}^H\tilde{\mathbf{x}}$ obtained via *sparse* inversion

$$\boxed{\tilde{\mathbf{x}} = \arg\min_{\mathbf{x}} ||\mathbf{x}||_1 \quad \text{subject to} \quad \mathbf{b} = \mathbf{A}\mathbf{x}}$$

where

$$\mathbf{A} := \mathbf{RMKS}^H = \underline{\mathbf{K}}\mathbf{S}^H \text{ and } \mathbf{b} = \delta\underline{\mathbf{d}}^k := \mathbf{RM}\boldsymbol{\delta}\mathbf{d}^k$$

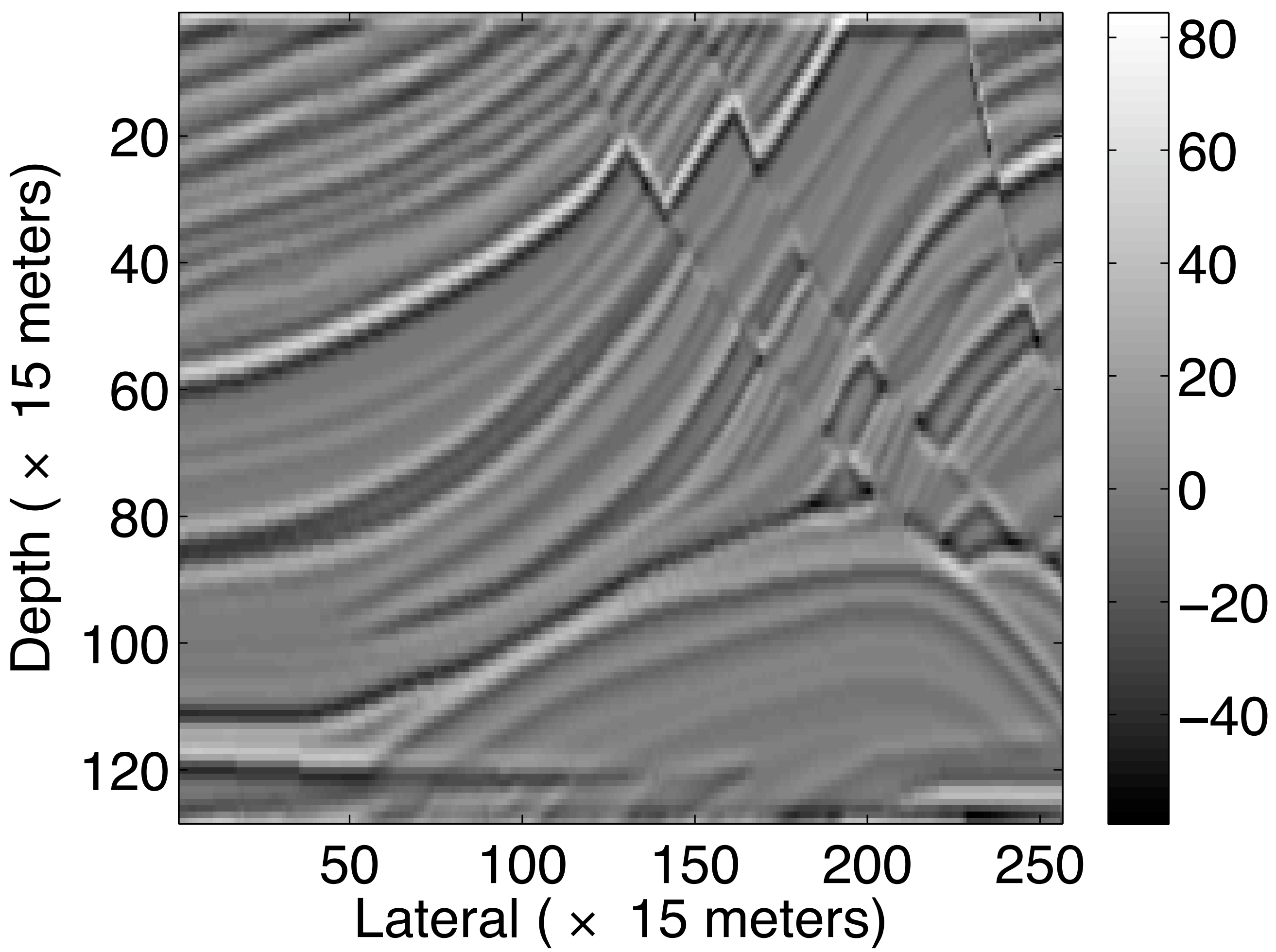
Initial model



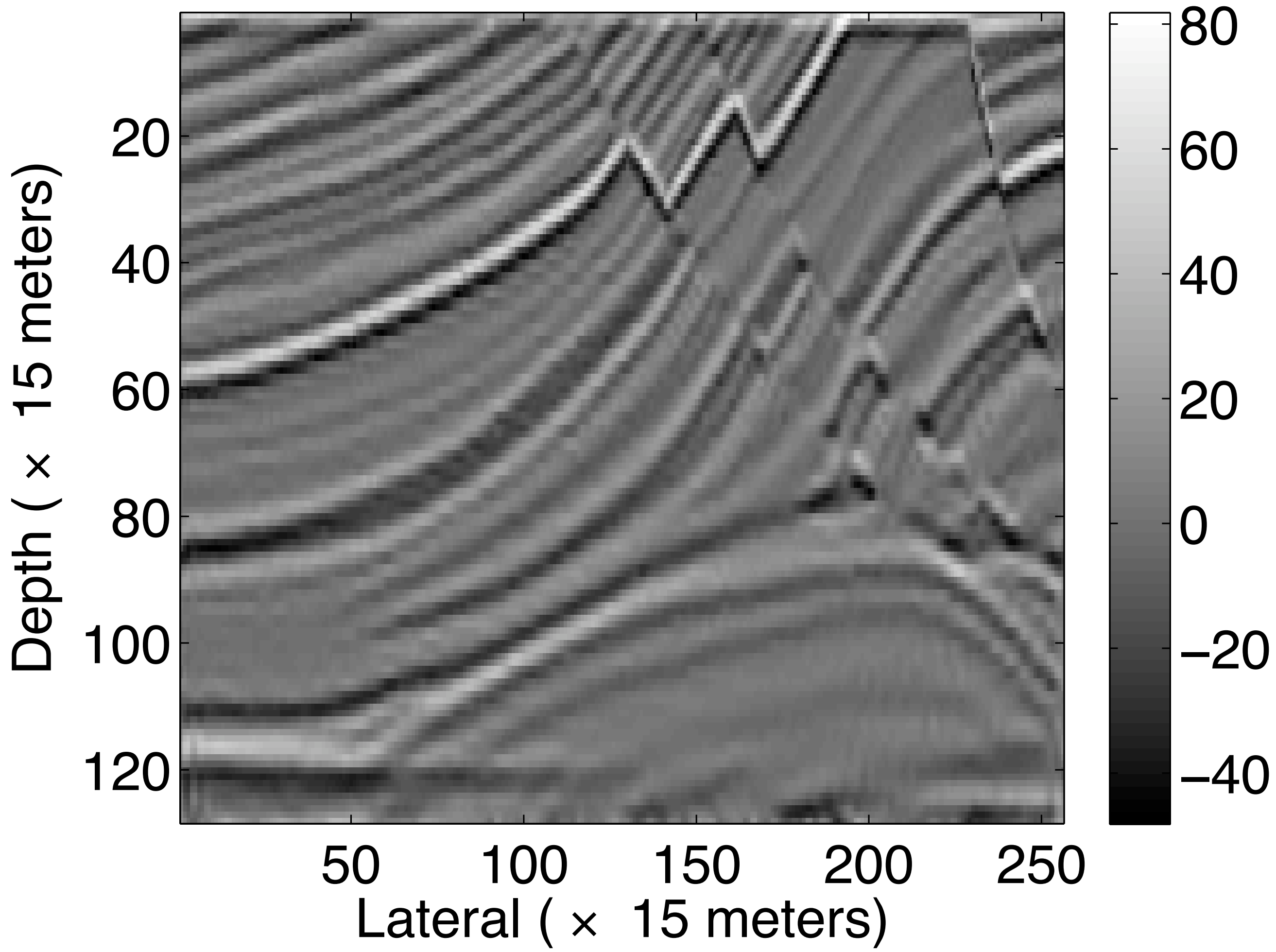
Linearized sparse inversion

30 simultaneous shots 10 random frequencies

***true* reflectivity**



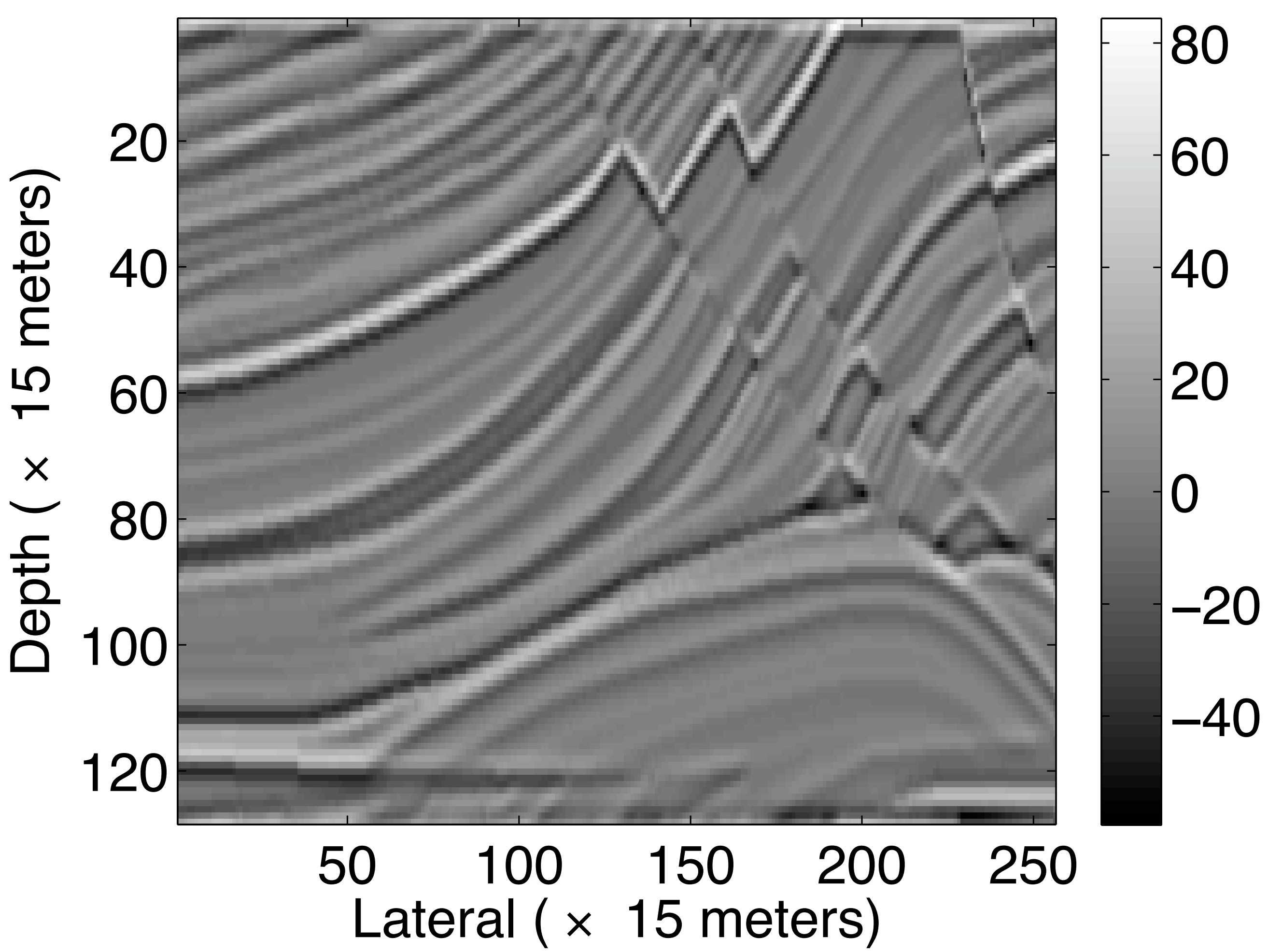
***sparse* recovery with wavelets**



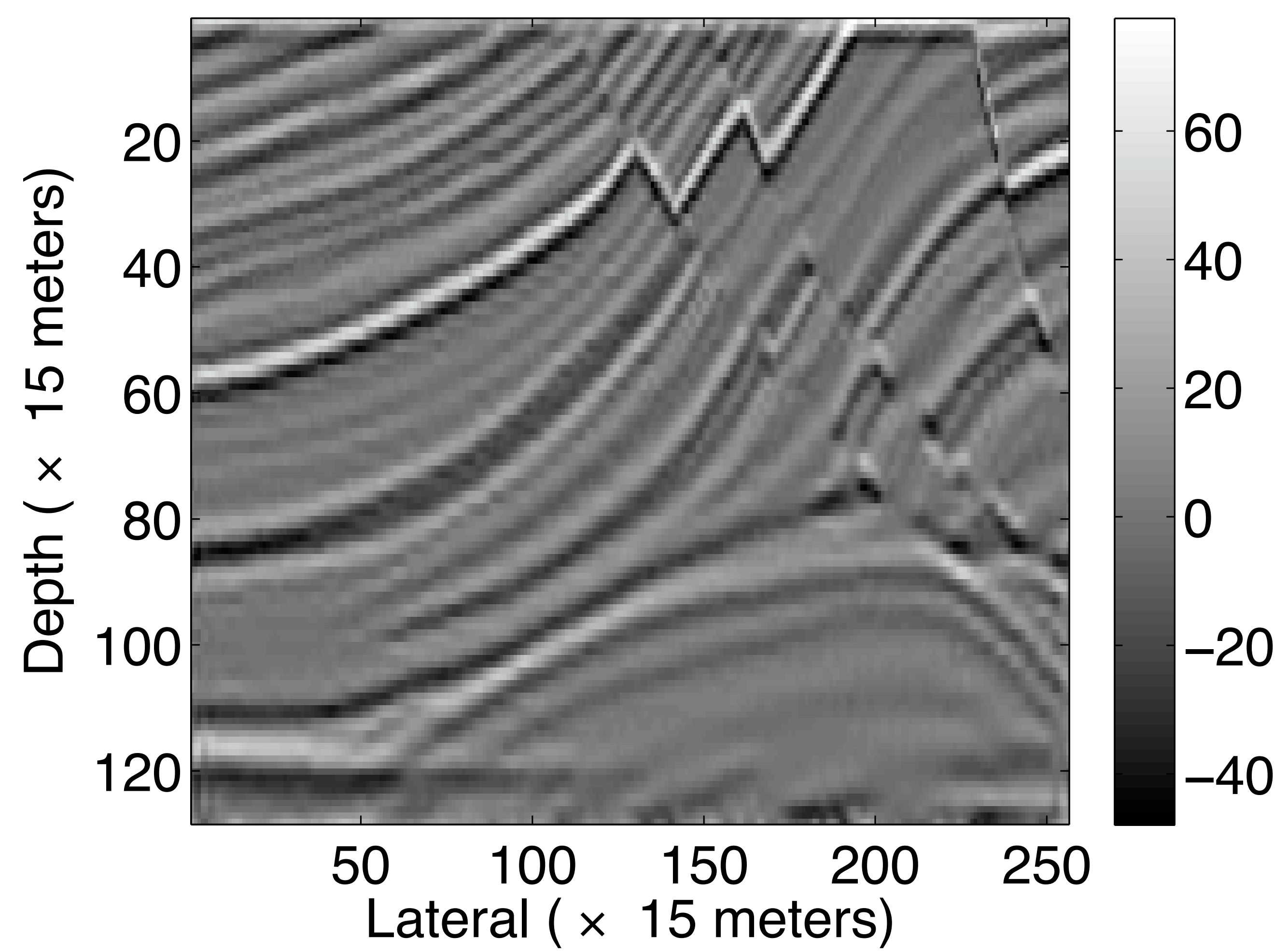
Linearized sparse inversion

20 simultaneous shots 10 random frequencies

***true* reflectivity**

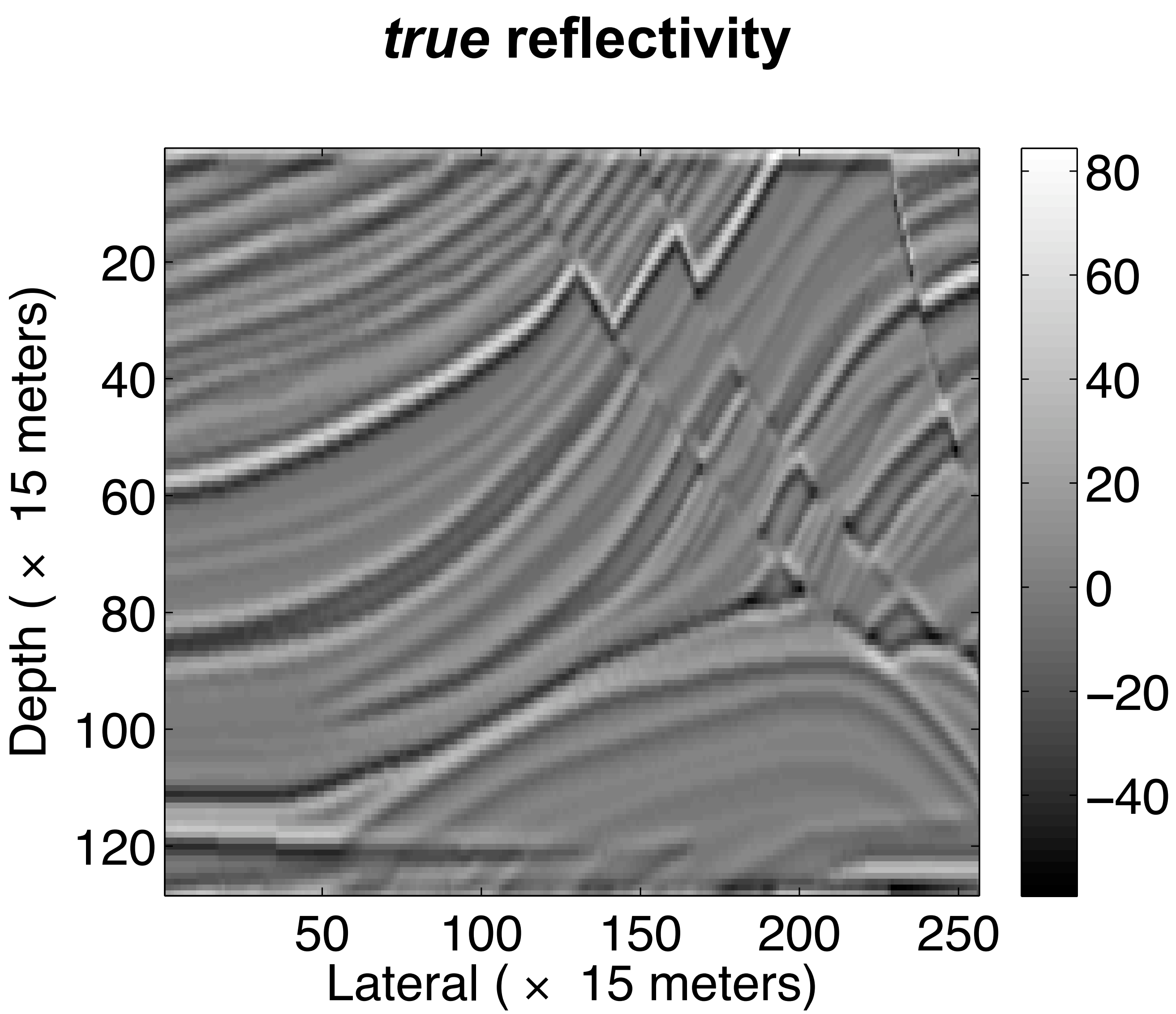


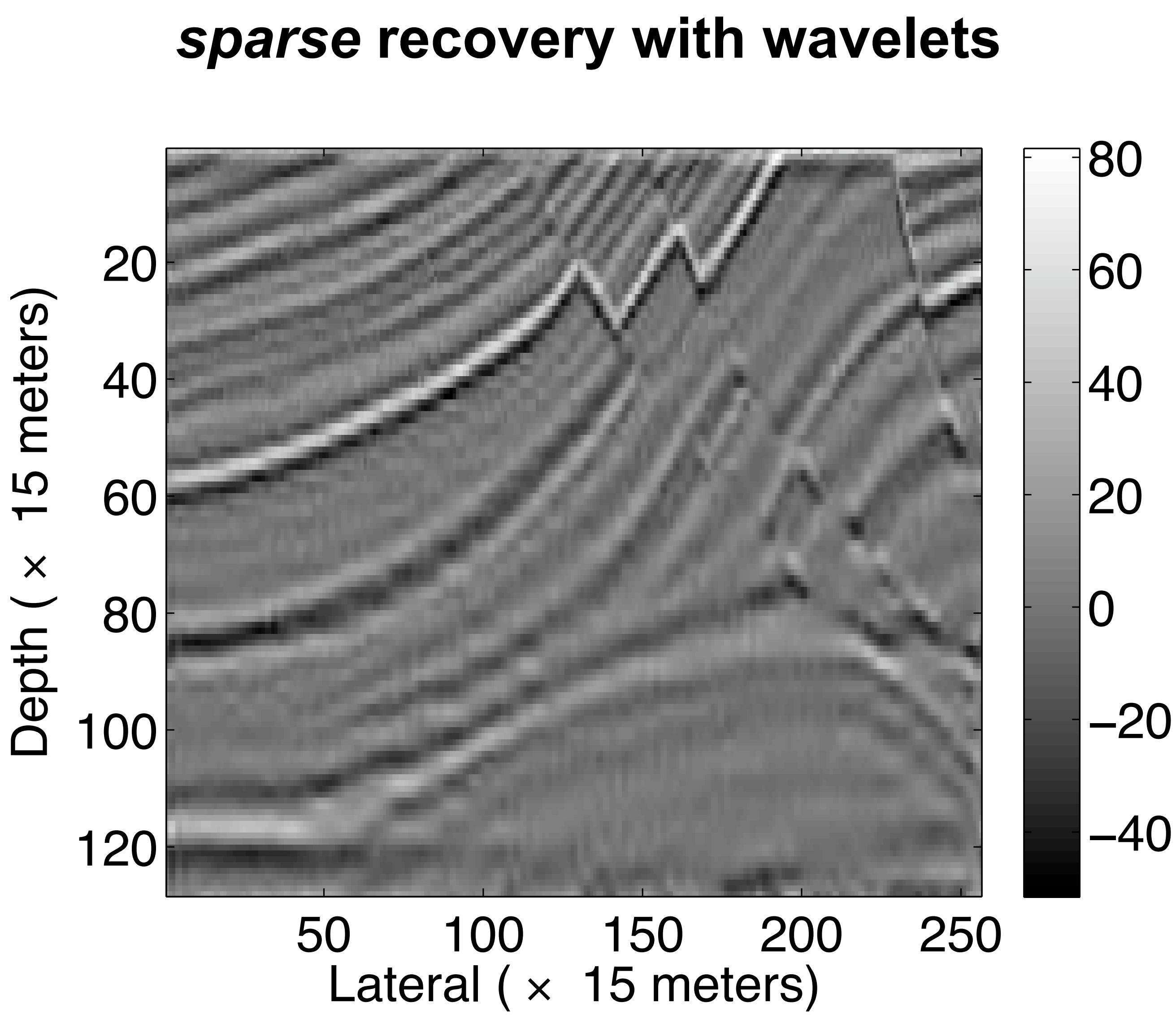
***sparse* recovery with wavelets**



Linearized sparse inversion

10 simultaneous shots 5 random frequencies





SLIM

Linearized sparse inversion

Subsample ratio	0.015	0.006	0.002
n'_f/n'_s	recovery error (dB)		
5	**17.44** (1.32)	**11.66** (0.78)	**6.83** (-0.14)
1	**17.53** (1.59)	**11.89** (1.05)	**7.19** (0.15)
0.2	**18.22** (1.68)	**12.11** (1.32)	**7.46** (0.27)
Speed up ($\times$)	66	166	500

errors for "migration" in parentheses

Observations

Reconstruct model updates

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

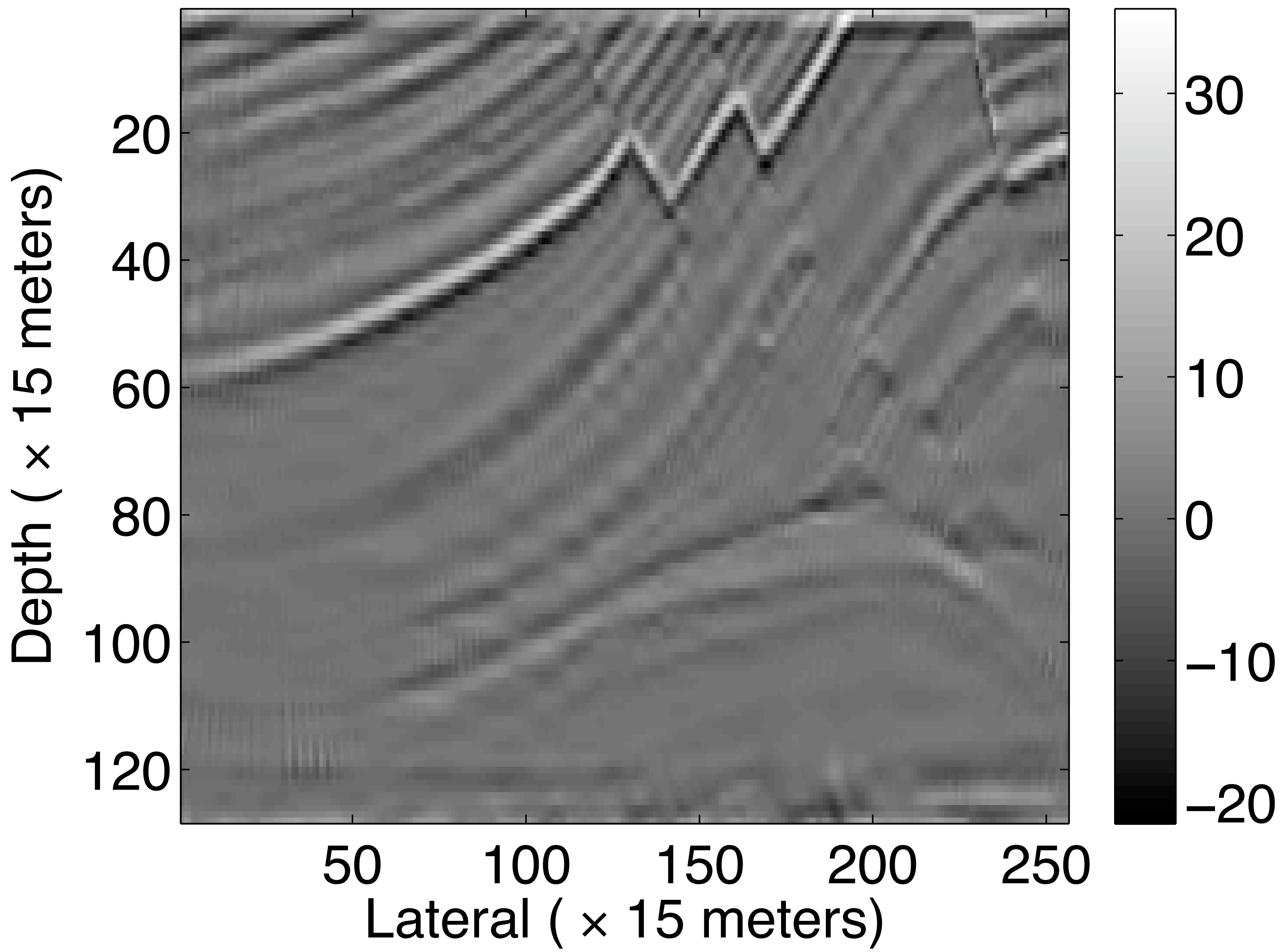
Removed the "curse of dimensionality" by reducing the number of RHS's

Recovery quality depends on *degree* of *subsampling*

What about *stochastic optimization*?

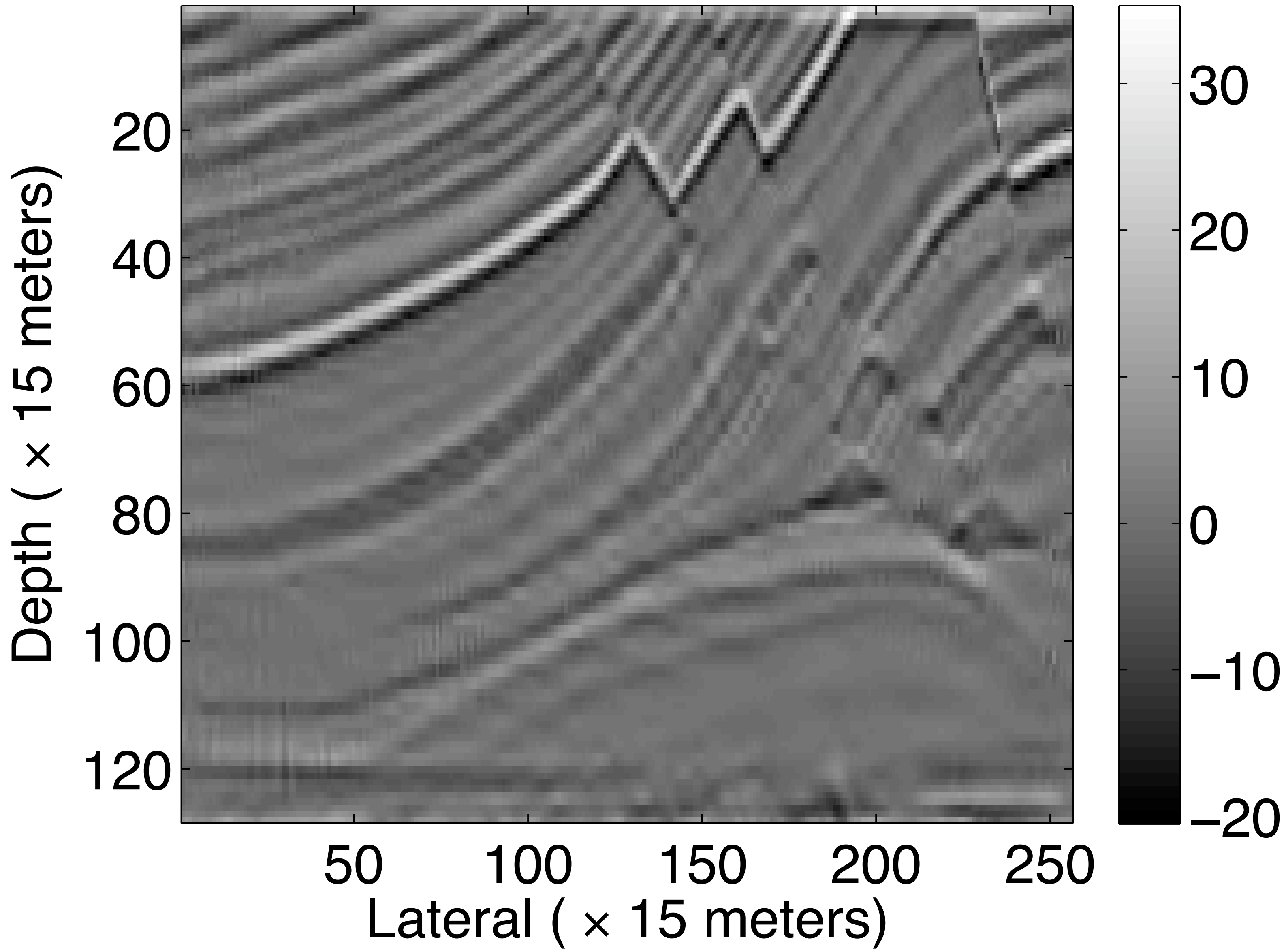
Reduced batch

10 X



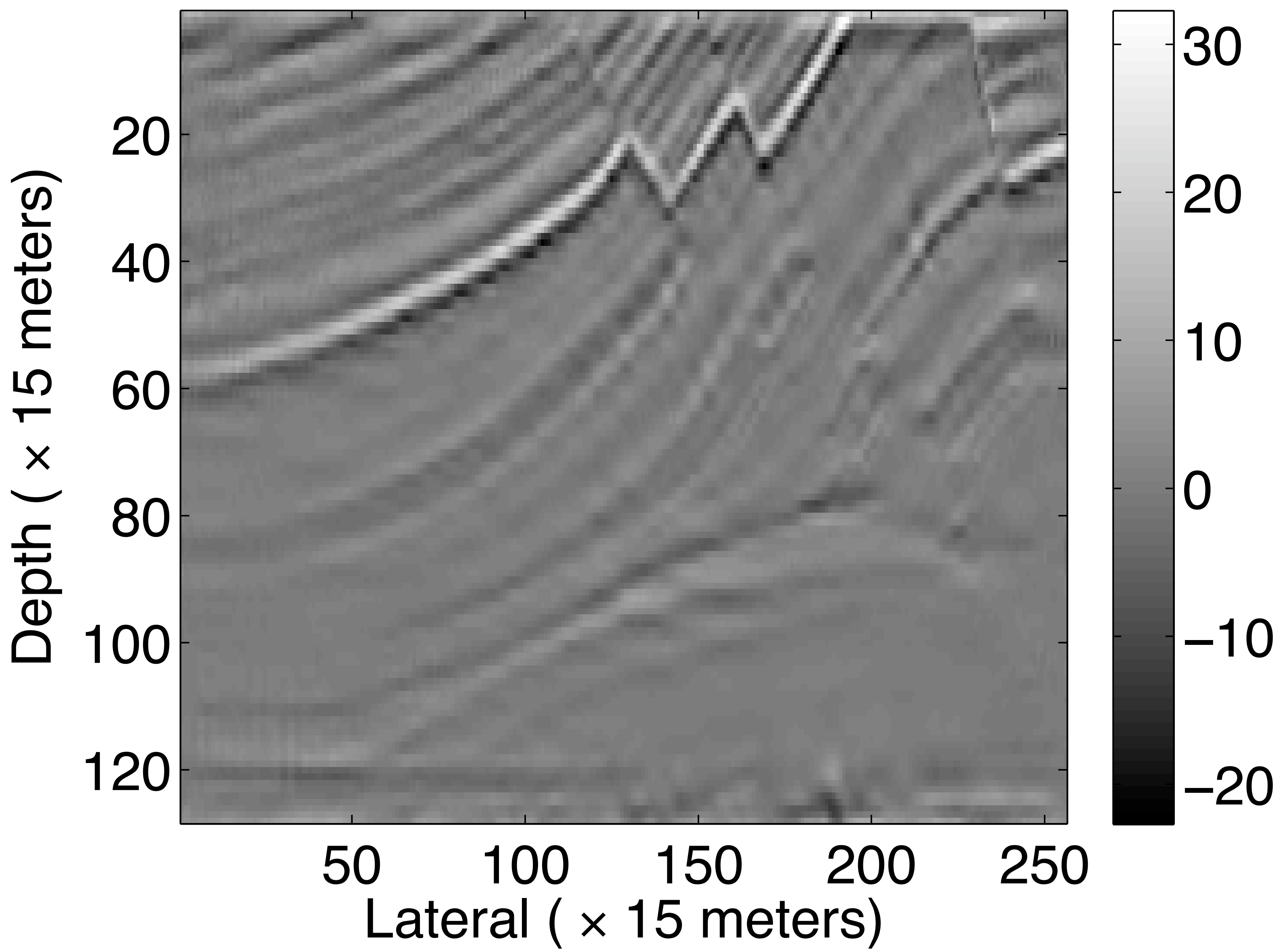
Stochastic mini batch

66 X



Stochastic average approximation

66 X



Stochastic mini batch

Algorithm 1: FWI without renewal of CS experiments.

Result: Estimate for the model $\widetilde{\mathbf{m}}$

$\mathbf{m} \longleftarrow \mathbf{m}_0$; `// initial model`

$j \longleftarrow 0$; `// loop counter`

$\underline{\mathbf{Q}} \longleftarrow (\mathbf{RM})\mathbf{Q}$; `// Draw random sim. shot`

while $\|\underline{\mathbf{P}} - \underline{\boldsymbol{\mathcal{F}}}[\mathbf{m}, \underline{\mathbf{Q}}]\|_{2,2}^2 \geq \epsilon$ **do**

  $j := j + 1$; `// increase counter`

  $\mathbf{A} \longleftarrow \underline{\mathbf{K}}[\mathbf{m}, \underline{\mathbf{Q}}]\mathbf{S}^*$; `// Calculate Jacobian`

  $\delta\underline{\mathbf{P}} = \underline{\mathbf{P}} - \underline{\boldsymbol{\mathcal{F}}}[\mathbf{m}, \underline{\mathbf{Q}}]$; `// Calculate residual`

  $\delta\widetilde{\mathbf{x}} = \arg\min_{\delta\mathbf{x}} \|\delta\mathbf{x}\|_{\ell_1} \quad \text{s.t.} \quad \|\delta\underline{\mathbf{P}} - \mathbf{A}\delta\mathbf{x}\|_2 \leq \sigma$; `// Sparse recovery`

  $\mathbf{m} \longleftarrow \mathbf{m} + \mathbf{S}^*\delta\widetilde{\mathbf{x}}$; `// Compute model update`

end

SLIM

SA mini batch

Algorithm 1: FWI with renewal of CS experiments.

Result: Estimate for the model $\widetilde{\mathbf{m}}$

$\mathbf{m} \longleftarrow \mathbf{m}_0$; // initial model

$j \longleftarrow 0$; // loop counter

$\underline{\mathbf{Q}} \longleftarrow (\mathbf{RM})_i\mathbf{Q}$; // Draw random sim. shot

while $\|\underline{\mathbf{P}} - \underline{\mathcal{F}}[\mathbf{m}, \underline{\mathbf{Q}}]\|_{2,2}^2 \geq \epsilon$ **do**

$j := j + 1$; // increase counter

$\mathbf{A} \longleftarrow \underline{\mathbf{K}}[\mathbf{m}, \underline{\mathbf{Q}}]\mathbf{S}^*$; // Calculate Jacobian

$\delta\underline{\mathbf{P}} = \underline{\mathbf{P}} - \underline{\mathcal{F}}[\mathbf{m}, \underline{\mathbf{Q}}]$; // Calculate residual

$\delta\widetilde{\mathbf{x}} = \arg\min_{\delta\mathbf{x}} \|\delta\mathbf{x}\|_{\ell_1} \quad \text{s.t.} \quad \|\delta\underline{\mathbf{P}} - \mathbf{A}\delta\mathbf{x}\|_2 \leq \sigma$; // Sparse recovery

$\mathbf{m} \longleftarrow \mathbf{m} + \mathbf{S}^*\delta\widetilde{\mathbf{x}}$; // Compute model update

$\underline{\mathbf{Q}} \longleftarrow (\mathbf{RM})_i\mathbf{Q}$; // Draw random sim. shot

end

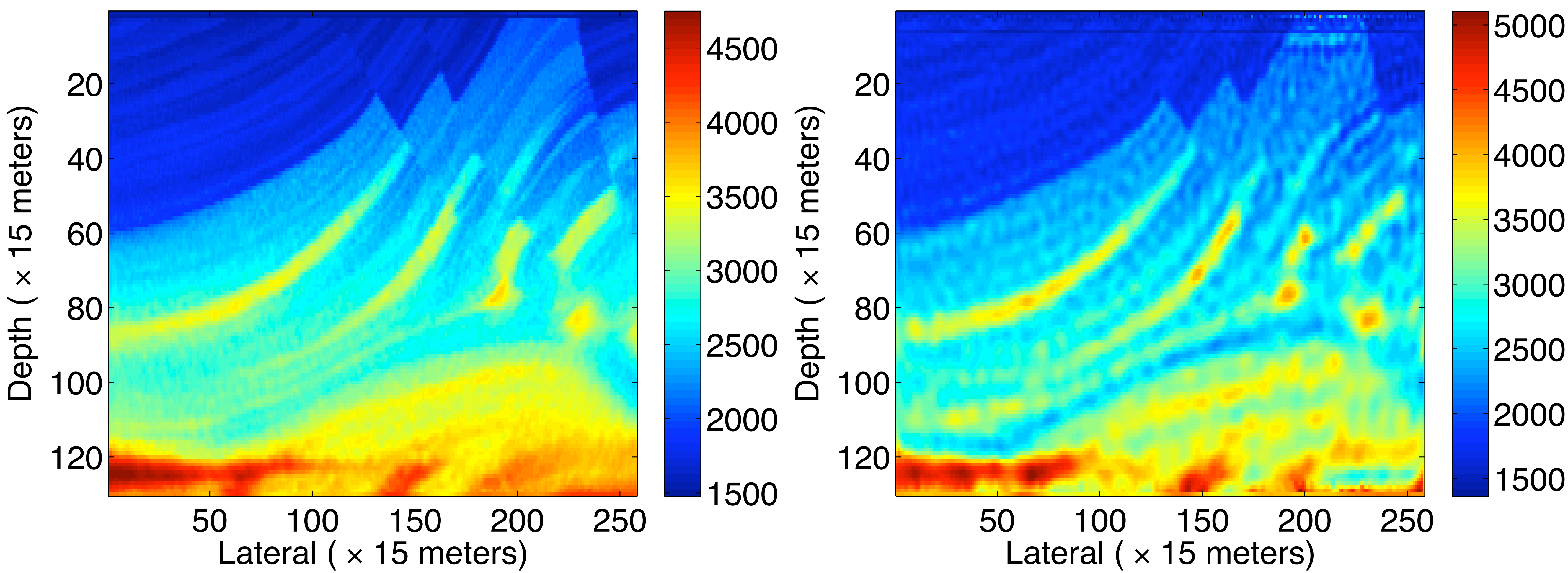
Update with sparse recoveries

SLIM

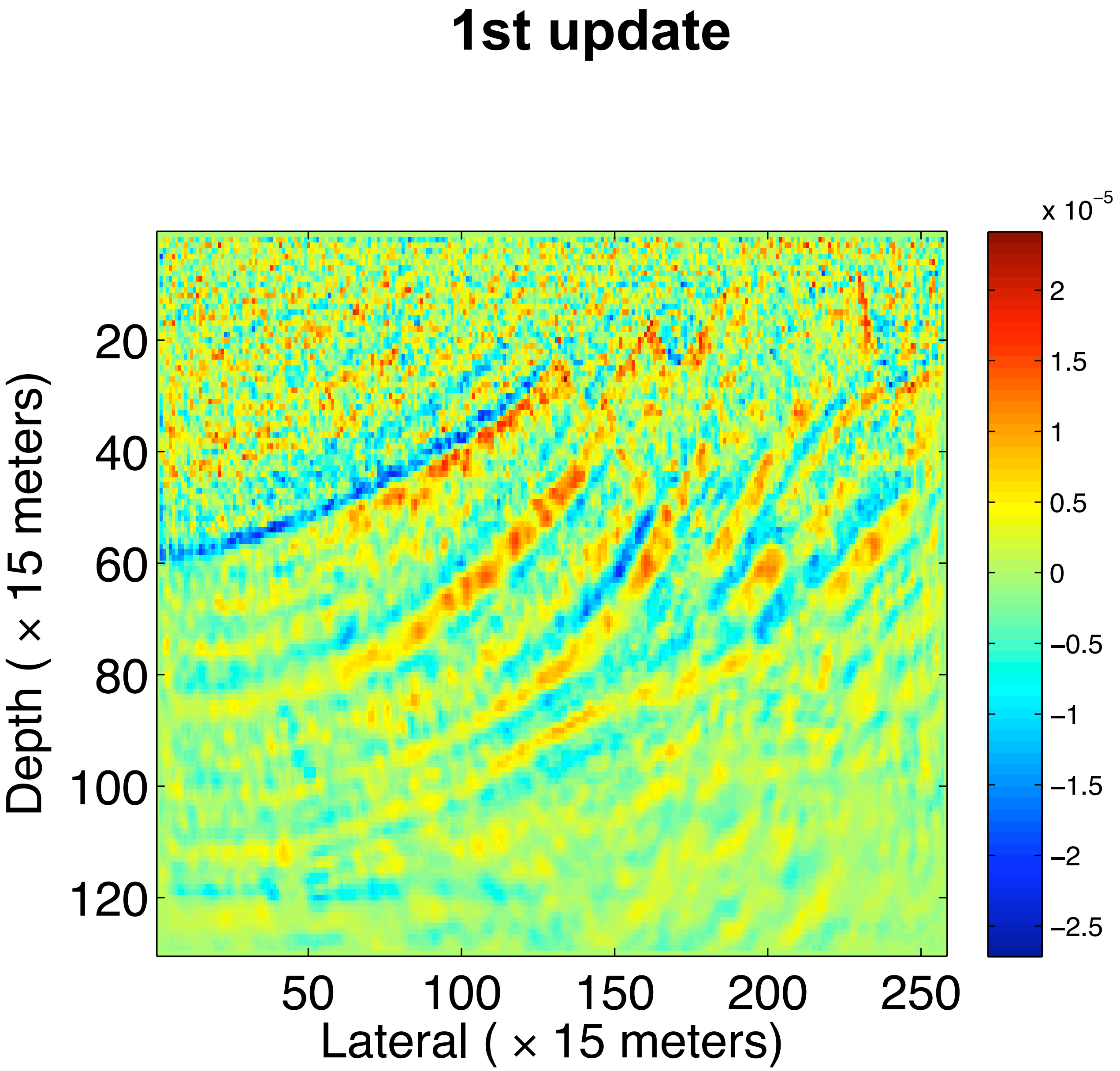
10 iterations

with renewal
(SNR = 20.8)

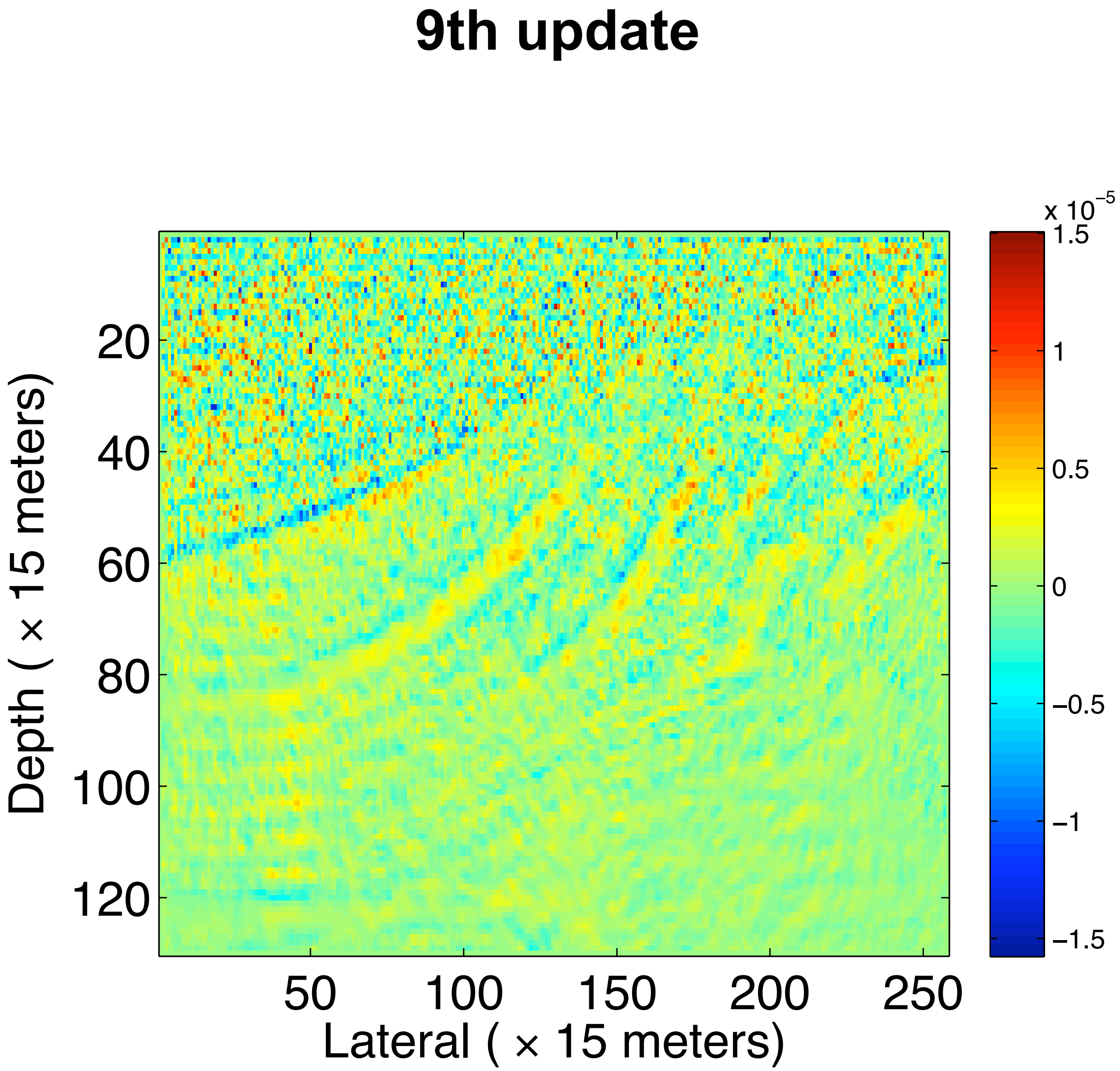
without renewal
(SNR = 19.9)



Updates



1st update
x 10^−5
2
1.5
1
0.5
0
−0.5
−1
−1.5
−2
−2.5
Depth (× 15 meters)
20
40
60
80
100
120
50
100
150
200
250
Lateral (× 15 meters)



9th update
x 10^−5
1.5
1
0.5
0
−0.5
−1
−1.5
Depth (× 15 meters)
20
40
60
80
100
120
50
100
150
200
250
Lateral (× 15 meters)

SLIM

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

SLIM

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:

- *Faster shot-record depth migrations using phase encoding* by Morton & Ober, '98.
- *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