Randomized full-waveform inversion: A dimensionality-reduction approach

Peyman Moghaddam & Felix J. Herrmann*

SLIM
University of British Columbia

Randomized full-waveform inversion: A dimensionality-reduction approach

Felix J. Hermann,
Peyman Moghaddam, and
Tristan van Leeuwen

SLIM
University of British Columbia

Motivation

Curse of dimensionality for d>2

- *Exponentially* increasing *data volumes*
- *Helmholtz* requires *implicit* solvers to address *bandwidth*
- Computational complexity grows *linearly* with # RHS's
- Makes *computation* of the misfit functional & 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 passes* over *all* data
- removes the *linearly* increasing costs of *implicit* solvers for increasing numbers of frequencies & RHS's

Key technologies

Simultaneous sources & phase encoding [Beasley, '98, Berkhout, '08] [Morton, '98, Romero, '00]

- supershots [Krebs et.al., '09, Operto et. al., '09, FJH et.al., '08-10']

Stochastic optimization & machine learning [Bertsekas, '96]

- stochastic gradient decent/stochastic approximation [Nemirovski, '09]

Compressive sensing [Candès et.al, Donoho, '06]

- *sparse* recovery & *randomized* subsampling

FWI formulation

Multiexperiment unconstrained optimization problem:

$$\min_{\mathbf{m}\in\mathcal{M}} \frac{1}{2}\|\mathbf{D}-\mathcal{F}[\mathbf{m};\mathbf{Q}]\|_{2,2}^2 \quad \text{with} \quad \mathcal{F}[\mathbf{m};\mathbf{Q}] := \mathbf{P}\mathbf{H}^{-1}[\mathbf{m}]\mathbf{Q}$$

- requires large number of PDE solves
- linear in the sources
- apply *randomized* dimensionality reduction

[Tarantola, 84; Pratt, '98; Plessix, '06]

SLIM

Reduced FWI formulation

Multiexperiment unconstrained optimization problem:

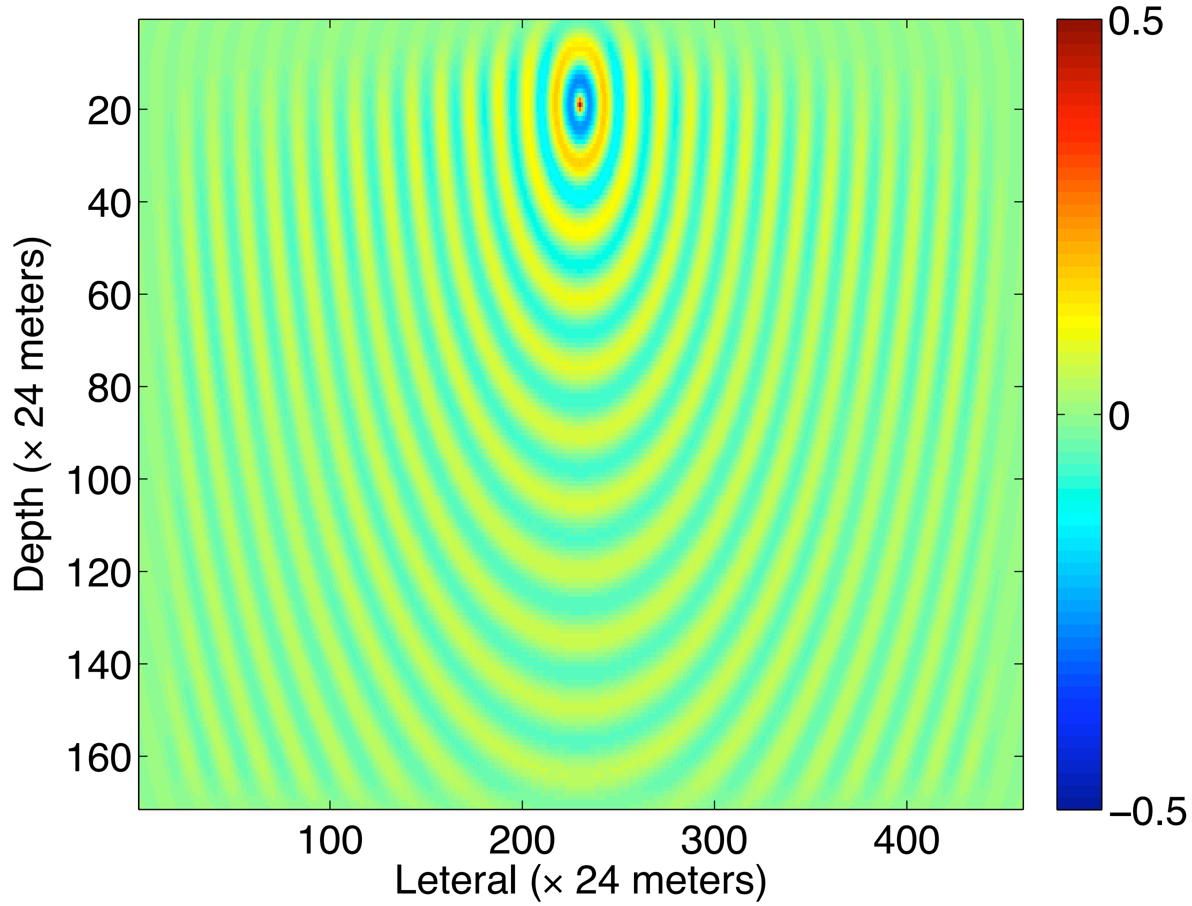
$$\min_{\mathbf{m}\in\mathcal{M}} \frac{1}{2}\|\underline{\mathbf{D}} - \mathcal{F}[\mathbf{m};\underline{\mathbf{Q}}]\|_{2,2}^2 \quad \text{with} \quad \mathcal{F}[\mathbf{m};\underline{\mathbf{Q}}] := \mathbf{P}\underline{\mathbf{H}}^{-1}\underline{\mathbf{Q}}$$

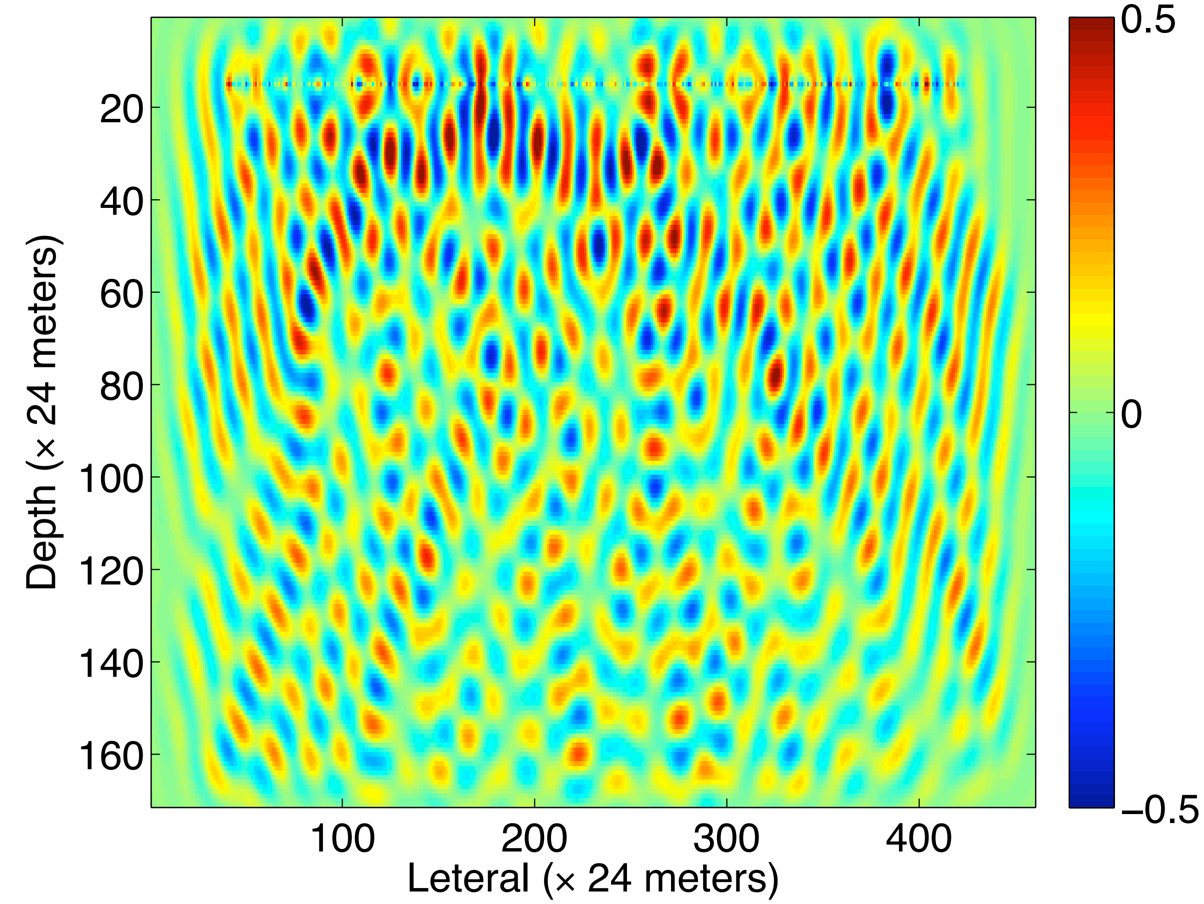
- requires *smaller* number of PDE solves
- explores *linearity* in the *sources* & *block-diagonal* structure of the *Helmholtz system*
- uses *randomized* frequency *selection* and *phase encoding*

Simultaneous shot at 5 Hz

Sequential-source wavefield

Simultaneous-source wavefield

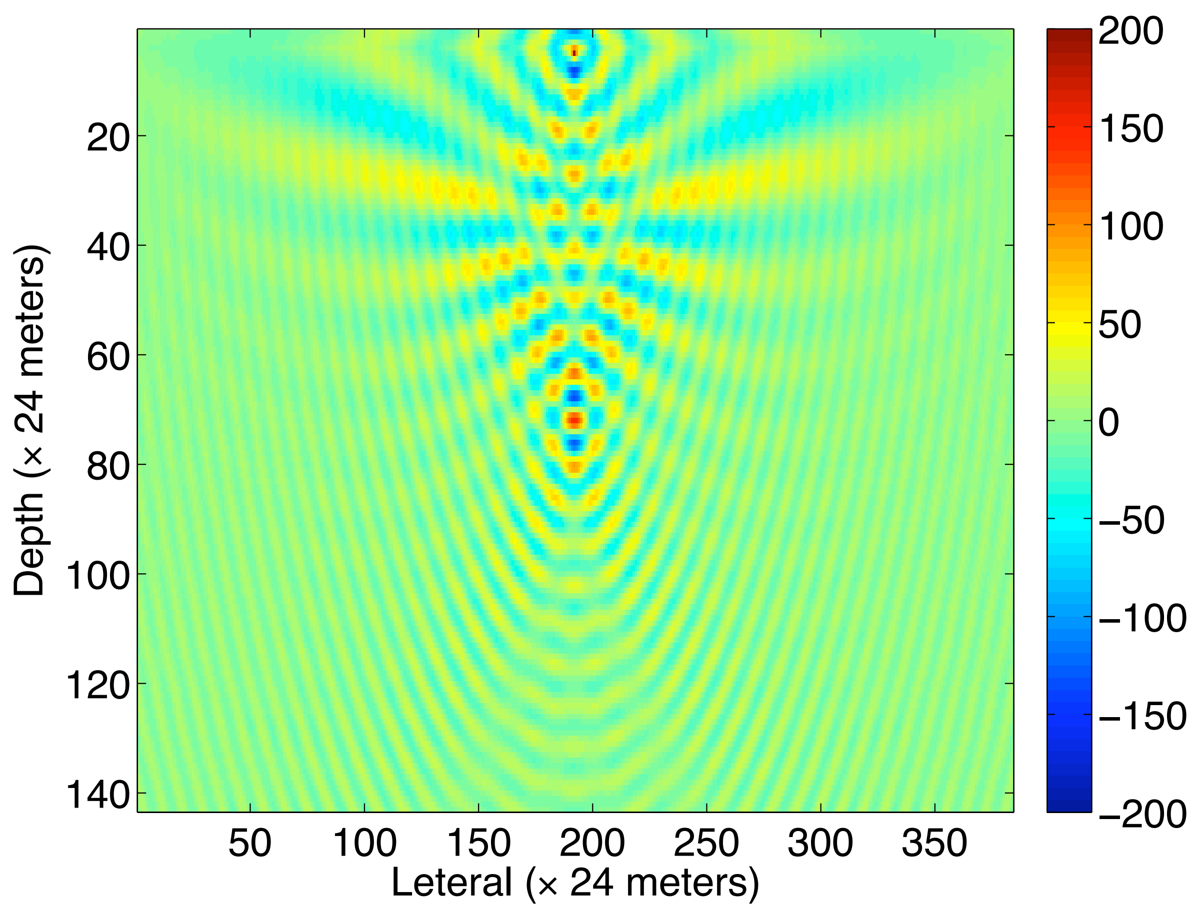




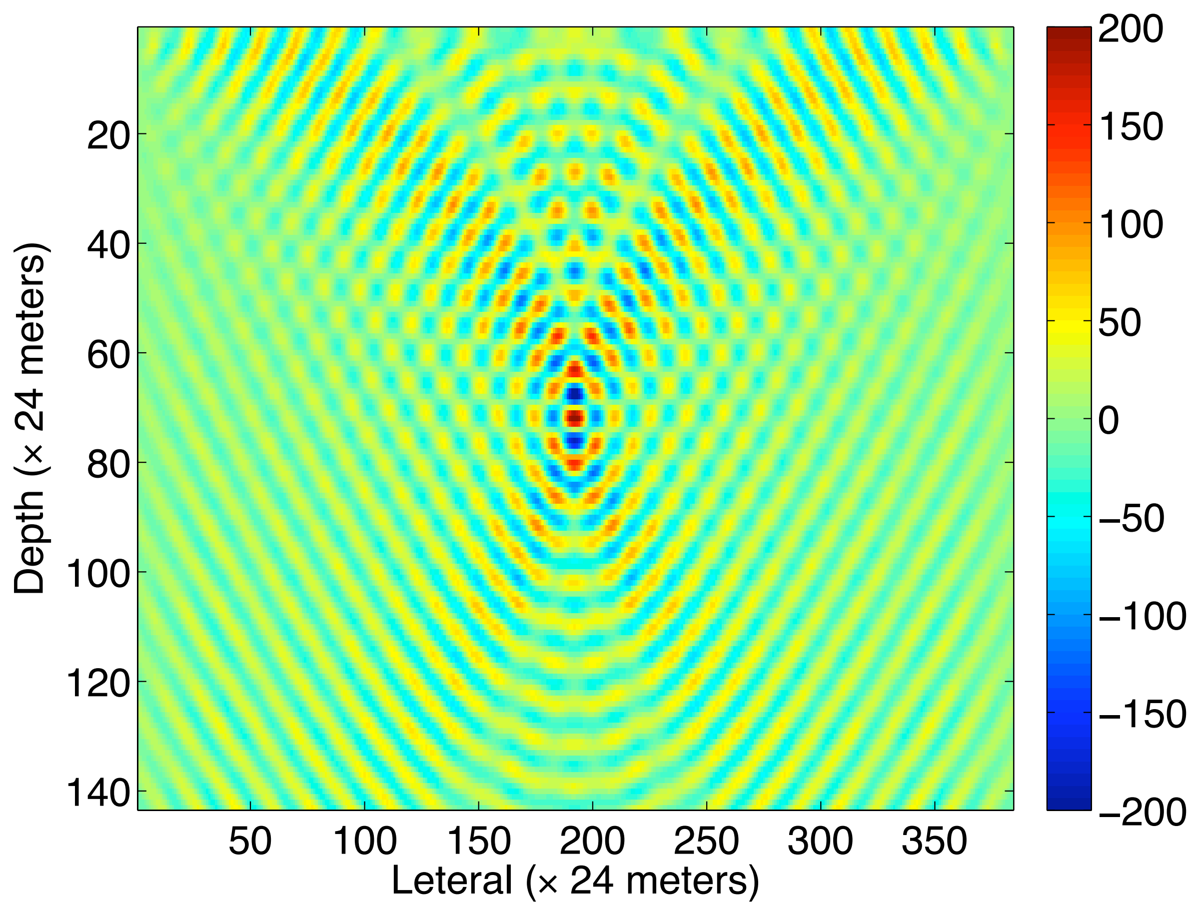
[Morton, '98, Romero, '00]

Image at 5Hz

Sequential-source image



Simultaneous-source image

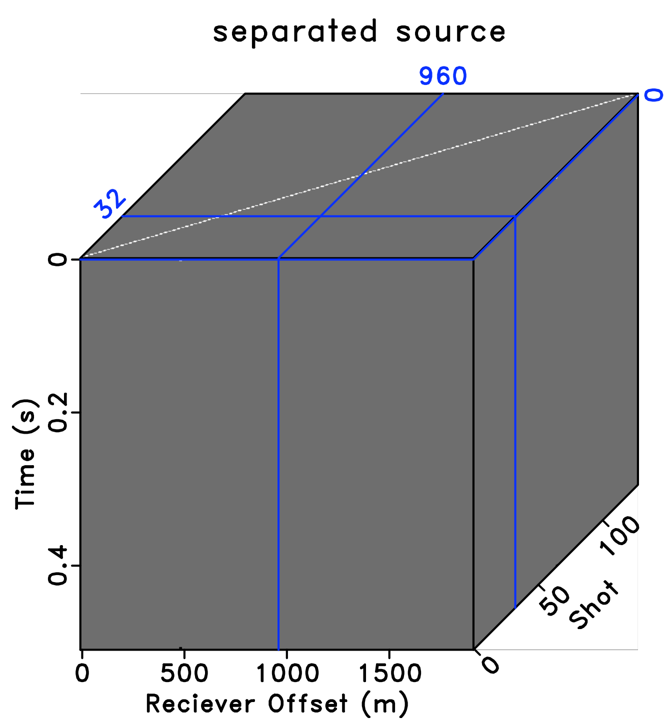


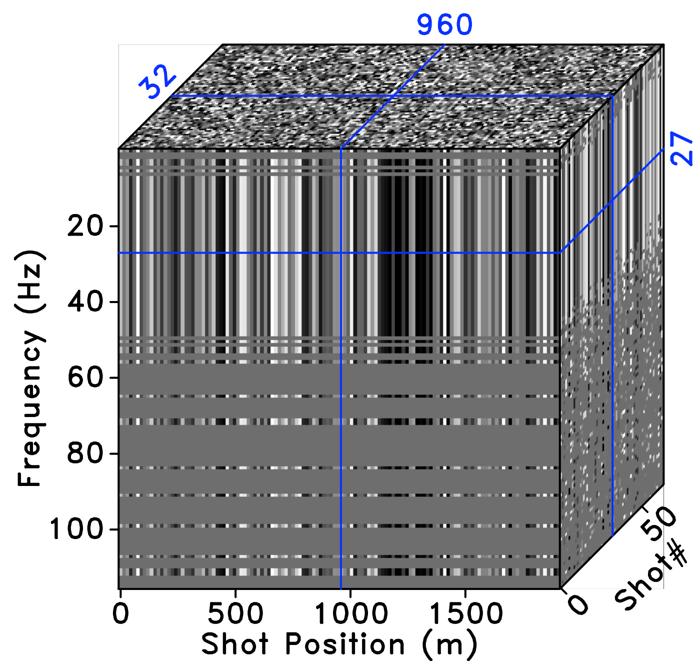
[Morton, ’98, Romero, ’00]

[FJH et. al. '08-'10]

Batch/mini experiment

adapted from FJH et. al. ,09





$$\mathbf{Q} \qquad \underline{\mathbf{Q}} = \mathbf{RMQ}$$

Collection of K *simultaneous-source* experiments with *batch* size $K \ll n_f \times n_s$

[FJH et. al. '08-'10]

Observations

Increased wavenumber *content* leads to *improved* image

Severe subsampling leads to *interferences* (source crosstalk and aliases)

Increasing the *number* of frequencies & simultaneous sources reduces *incoherent* interference *noise*

Is there something more we can say...

Interpretations

Consider *randomized* dimensionality reduction as instances of

- *stochastic optimization & machine learning*
  (today's talk)

- *compressive sensing* [FJH et. al, '08-'10]
  (tomorrow's talk by Xiang Li, 10:35 am, Room 405/406)

Stochastic optimization

Replace *deterministic*-optimization problem

$$\min_{\mathbf{m}\in\mathcal{M}} f(\mathbf{m}) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{2}\|\mathbf{d}_i - \mathcal{F}[\mathbf{m};\mathbf{q}_i]\|_2^2$$

with *sum* cycling over *different sources* & *corresponding shot records*
(columns of D & Q)

[Natterer, '01]

Stochastic *average* approximation

[Haber, Chung, and FJH, '10]

by *a stochastic-optimization* problem

$$\min_{\mathbf{m}\in\mathcal{M}} \mathbf{E}_{\mathbf{w}}\{f(\mathbf{m},\mathbf{w}) = \frac{1}{2}\|\mathbf{D}\mathbf{w} - \mathcal{F}[\mathbf{m};\mathbf{Q}\mathbf{w}]\|_2^2\}$$

$$\approx \frac{1}{K}\sum_{j=1}^{K}\frac{1}{2}\|\underline{\mathbf{d}}_j - \mathcal{F}[\mathbf{m};\underline{\mathbf{q}}_j]\|_2^2$$

with $\mathbf{w} \in N(0,1)$ and $\mathbf{E}_{\mathbf{w}}\{\mathbf{w}\mathbf{w}^H\} = \mathbf{I}$

and $\underline{\mathbf{d}}_j = \mathbf{D}\mathbf{w}_j,\ \underline{\mathbf{q}}_j = \mathbf{Q}\mathbf{w}_j$

FWI with phase encoding

Multiexperiment unconstrained optimization problem:

$$\min_{\mathbf{m}\in\mathcal{M}} \frac{1}{2}\|\underline{\mathbf{D}} - \mathcal{F}[\mathbf{m};\underline{\mathbf{Q}}]\|_{2,2}^2 \quad \text{with} \quad \mathcal{F}[\mathbf{m};\underline{\mathbf{Q}}] := \mathbf{P}\underline{\mathbf{H}}^{-1}\underline{\mathbf{Q}}$$

- requires *smaller* number of PDE solves
- exploits *linearity* in the *sources* & *block-diagonal* structure of the *Helmholtz system*
- uses *randomized* frequency *selection* and *phase encoding*

[Krebs et.al., '09, Operto et. al., '09 ; FJH et. al. '08-'10]

Stochastic *average* approximation

In the *limit* $K \to \infty$, *stochastic* & *deterministic* formulations are *identical*

We *gain* as long as $K \ll N$ ...

Since the error in *Monte-Carlo* methods decays only slowly $(\mathcal{O}(K^{-1/2}))$

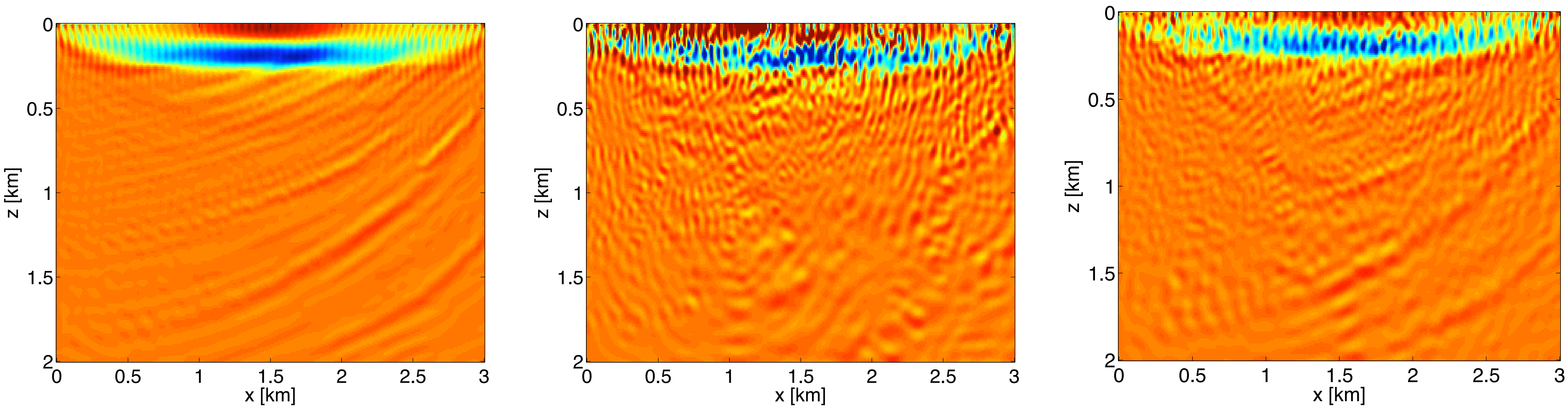
this approach may be problematic...

However, the location for the *minimum* of the *misfit* may be relatively *robust*...

Stylized example

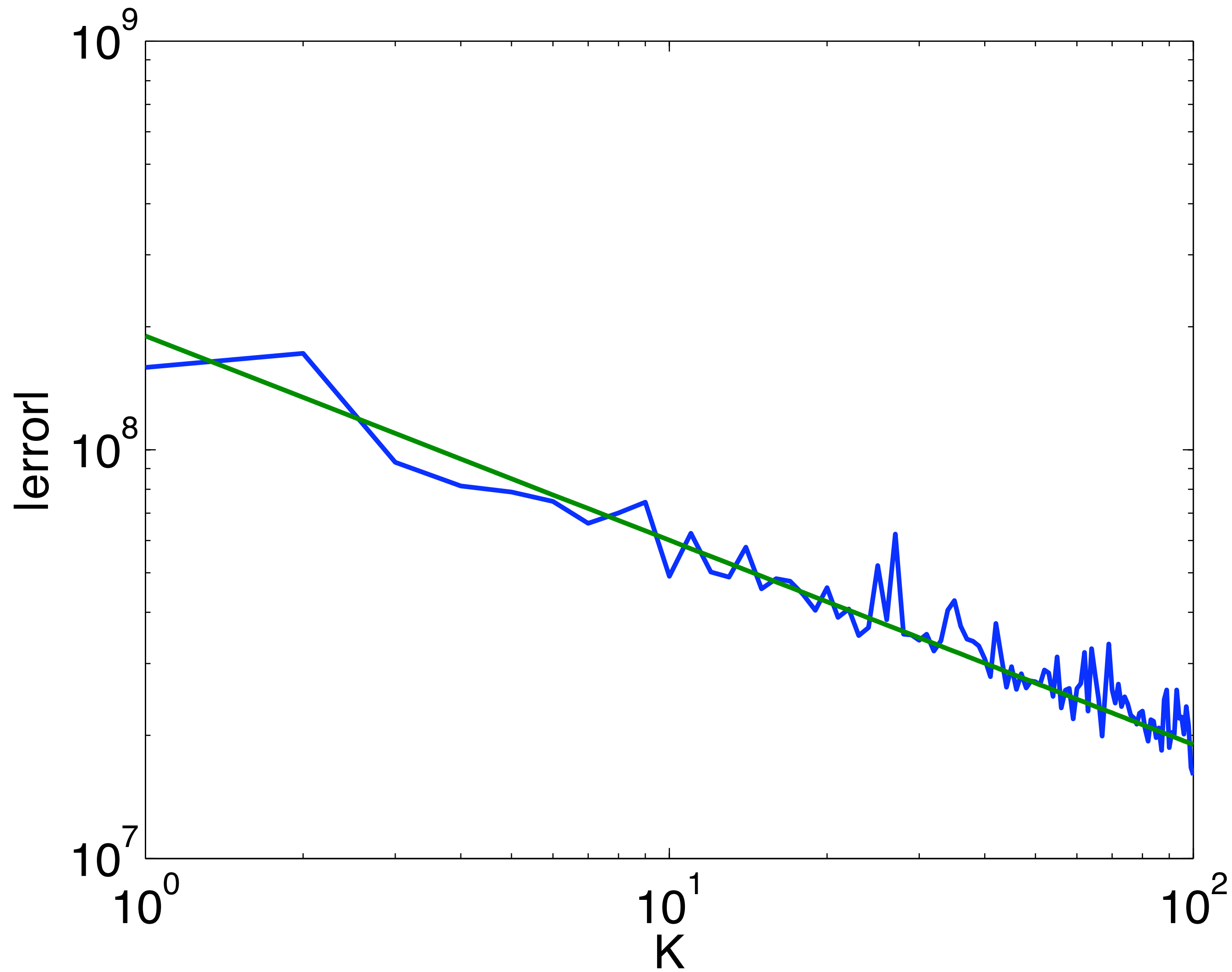
Search direction for batch size K:

$$\mathbf{g}_K \approx \frac{1}{K}\sum_{j=1}^{K} \nabla\mathcal{F}^*[\mathbf{m};\underline{\mathbf{q}}_j]\delta\underline{\mathbf{d}}_j$$



full K=1 K=5

Decay

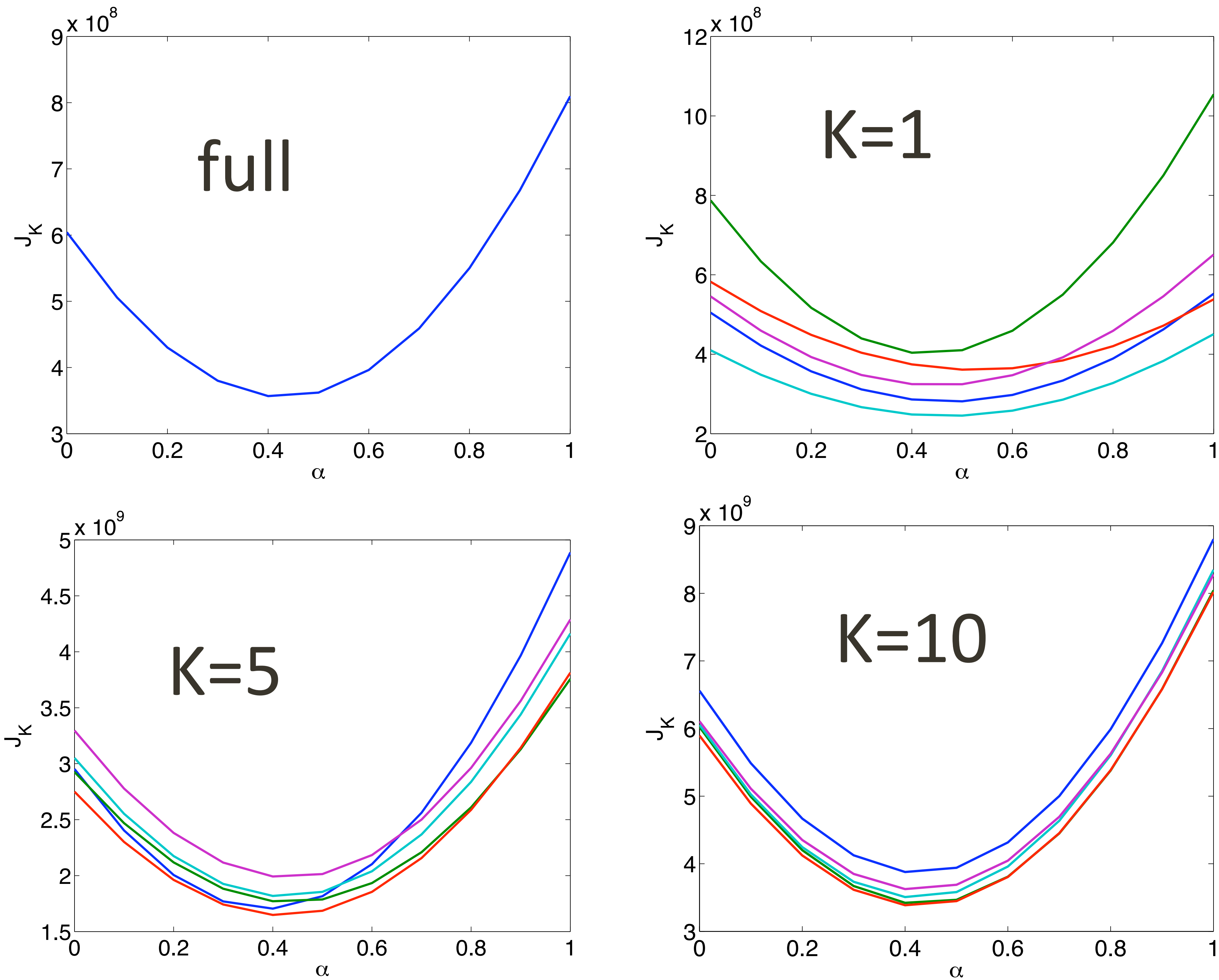


error between full and sampled gradient

Misfit functional

[adapted from Haber, Chung, and FJH, '10]

$$f_K(\mathbf{g}_K) = \frac{1}{K}\sum_{j=1}^{K}\frac{1}{2}\|\underline{\mathbf{d}}_j - \mathcal{F}[\mathbf{m}+\alpha\mathbf{g}_K;\underline{\mathbf{q}}_j]\|_2^2$$



Stochastic approximation [Bertsekas,' '96; Nemirovski, '09]

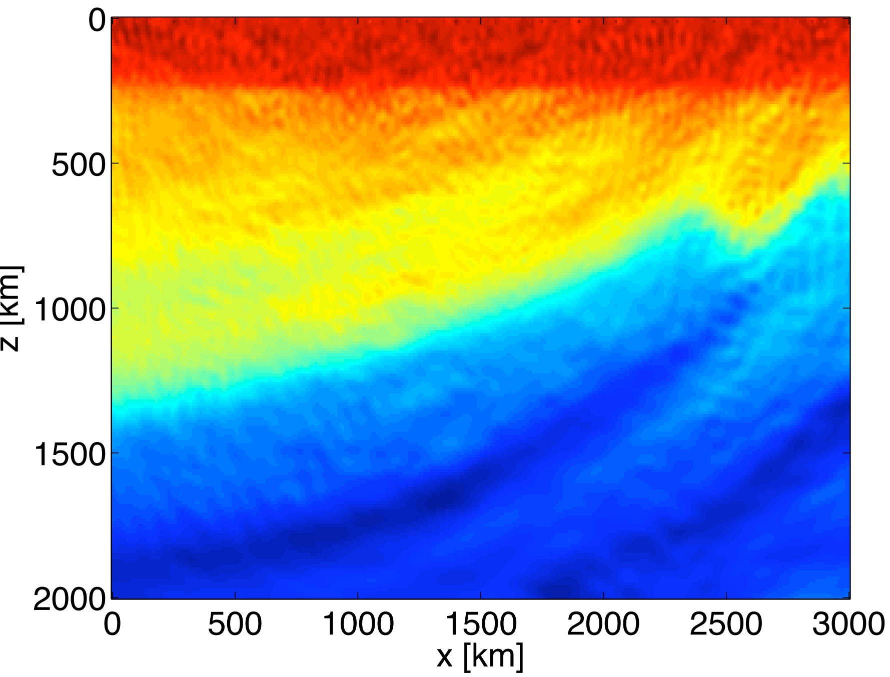
Use *different* simultaneous shots for each *subproblem*, i.e.,

$$\underline{\mathbf{Q}} \quad \mapsto \quad \underline{\mathbf{Q}}^{k}$$

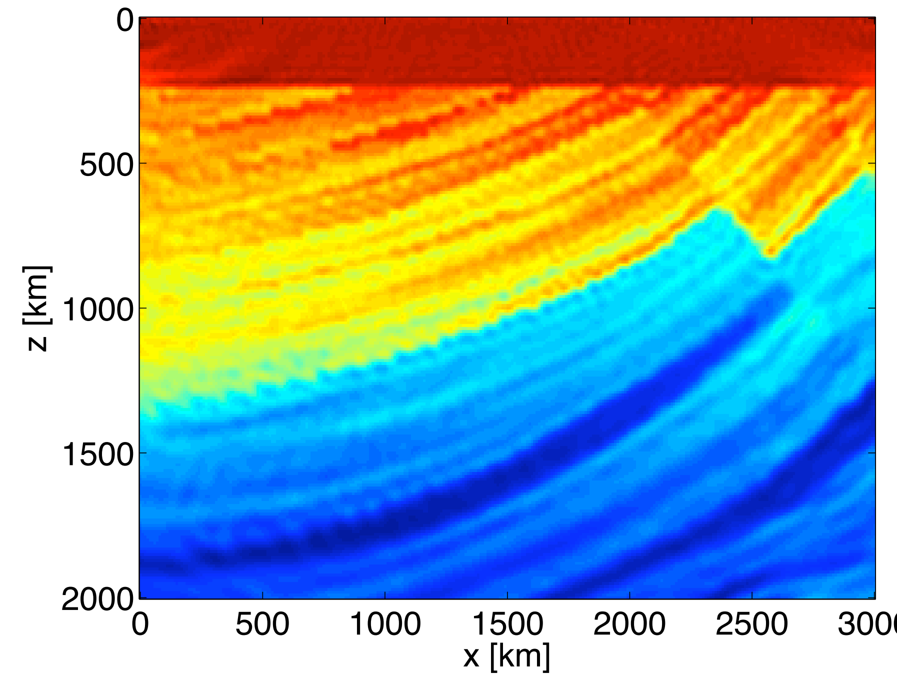
Requires *fewer* PDE solves for each GN *subproblem*...

- corresponds to *stochastic approximation* [Nemirovski, '09]
- related to Kaczmarz ('37) method applied by Natterer, '01
- *supersedes ad hoc* approach by Krebs *et.al.*, '09

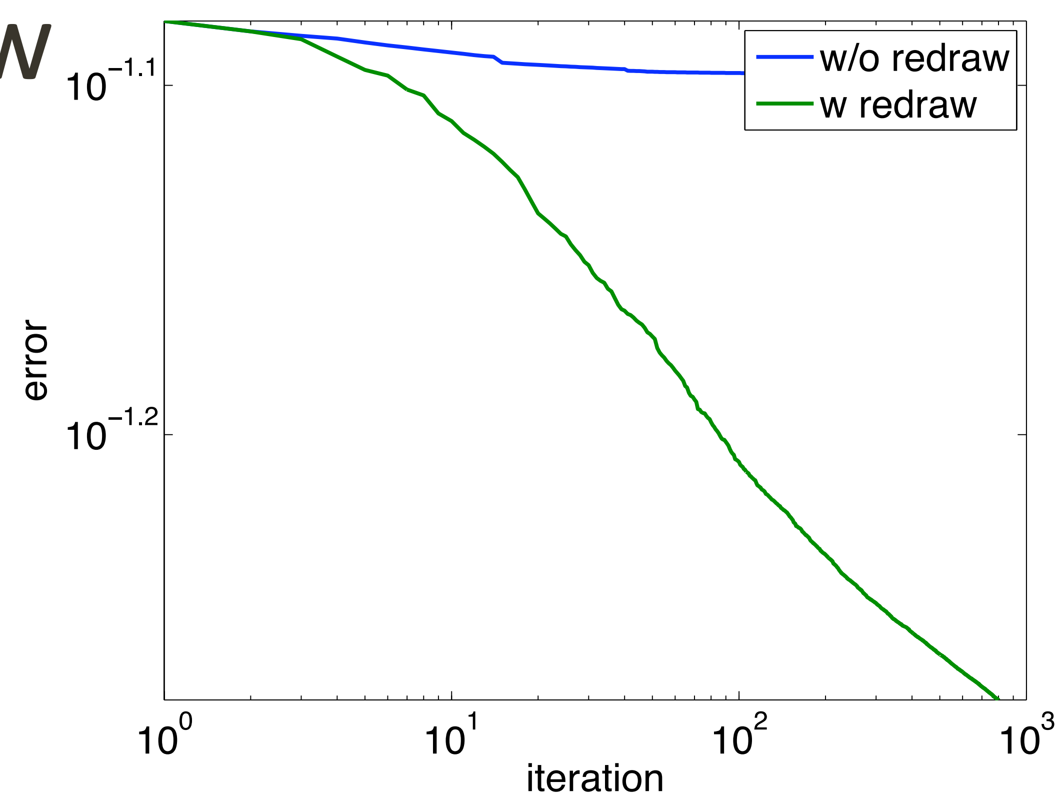
K=1 w and w/o redraw [noise-free case]



w/o redraw



w redraw



model error K=1, no averaging

Known issues

Renewals introduce *stochasticity* in the *gradients*

May lead to

- lack of convergence
- sensitivity to noise in data [Krebs, '09-'10]

Solutions

- increase the batch size
- average over the past model updates

Stochastic approximation

SLIM

Algorithm 1: Stochastic gradient descent

Result: Output estimate for the model $\mathbf{m}$

$\mathbf{m} \longleftarrow \mathbf{m}_0$; $k \longleftarrow 0$; `// initial model`

while not converged **do**

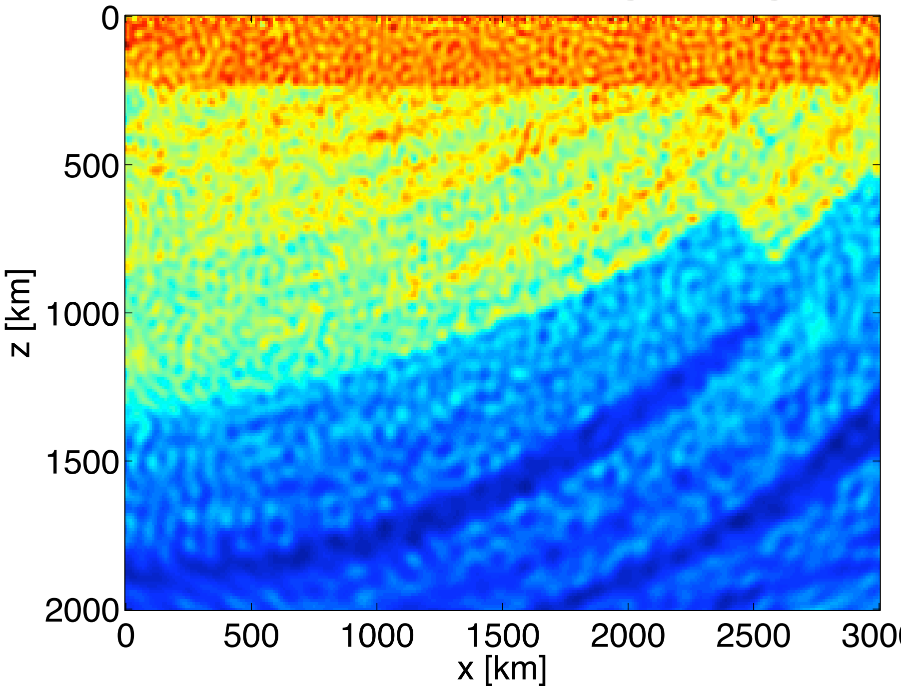
- $\{\underline{\mathbf{d}}^k, \underline{\mathbf{q}}^k\} \longleftarrow \{\mathbf{D}\mathbf{w}^k, \mathbf{Q}\mathbf{w}^k\}$ with $\mathbf{w}^k \in N(0,1)$; `// draw sim. exp.`
- $\mathbf{g}^k \longleftarrow \nabla\mathcal{F}^*[\mathbf{m}^{k-1}, \underline{\mathbf{q}}^k](\underline{\mathbf{d}}^k - \mathcal{F}[\mathbf{m}^{k-1}, \underline{\mathbf{q}}^k])$; `// gradient`
- $\underline{\mathbf{m}}^{k+1} \longleftarrow \mathbf{m}^k - \gamma^k \mathbf{g}^k$; `// update with linesearch`
- $\mathbf{m}^{k+1} = \frac{1}{k+1}\left(\sum_{i=1}^{k} \mathbf{m}^i + \underline{\mathbf{m}}^{k+1}\right)$; `// average`
- $k \longleftarrow k+1$;

end

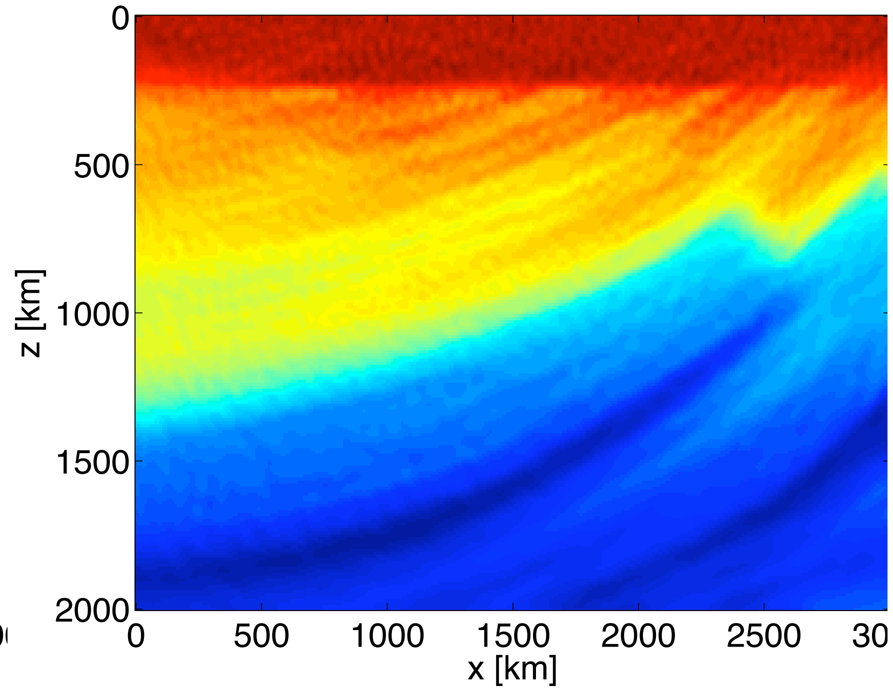
[Bertsekas, '96; Haber, Chung, and FJH, '10]

K=1
[noisy case]

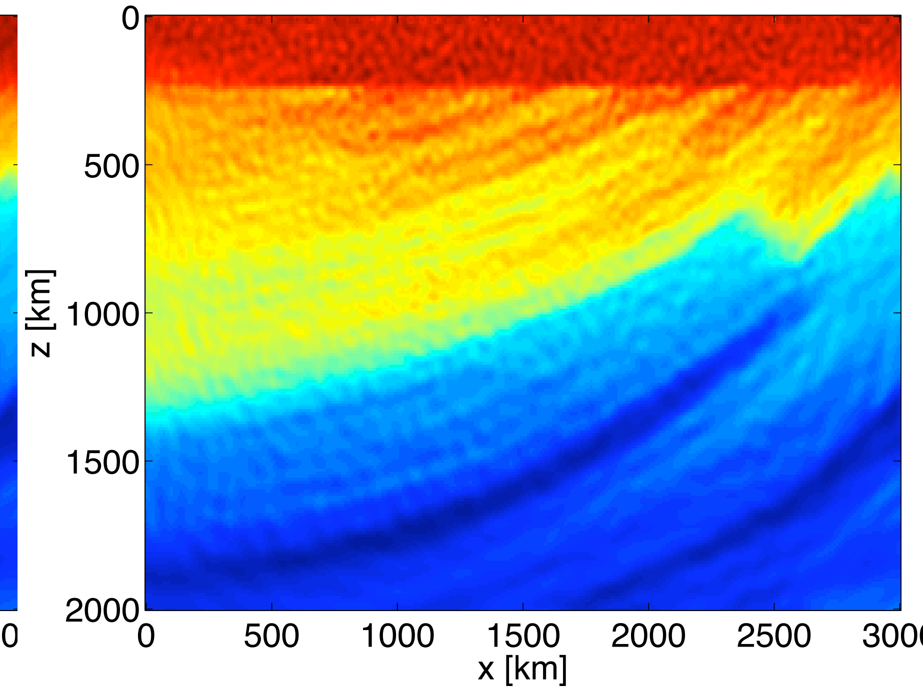
w/o averaging

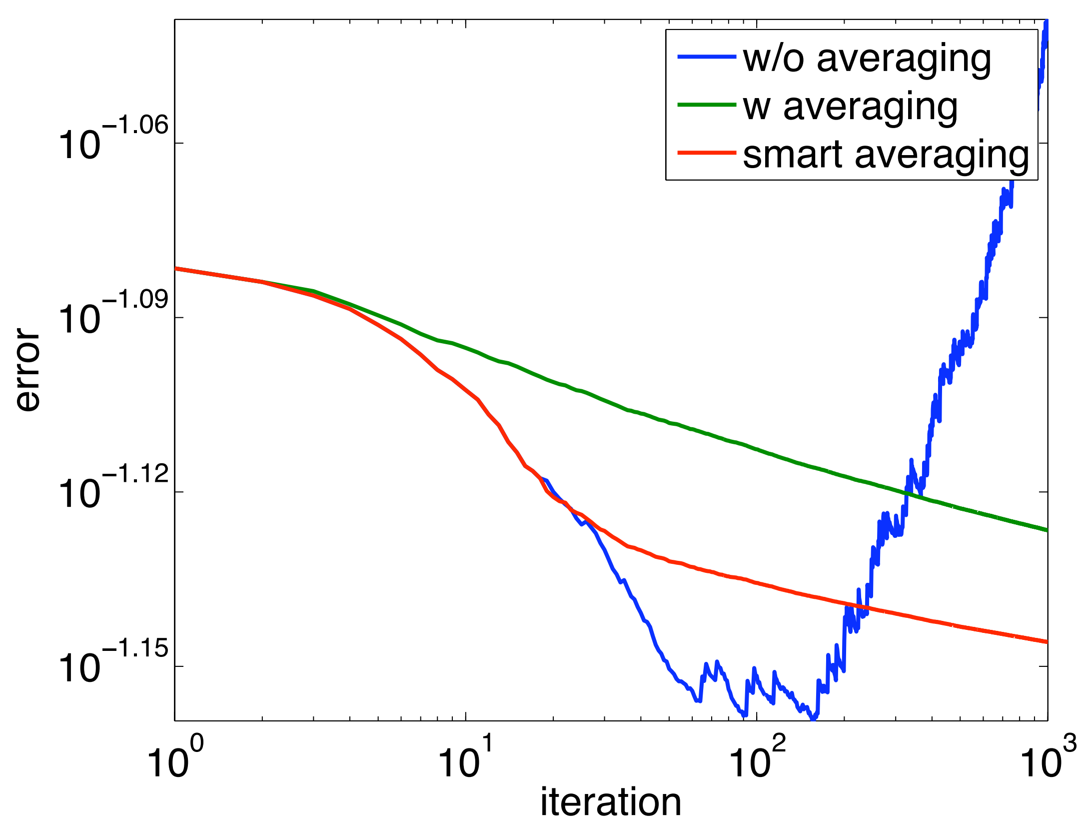


w averaging

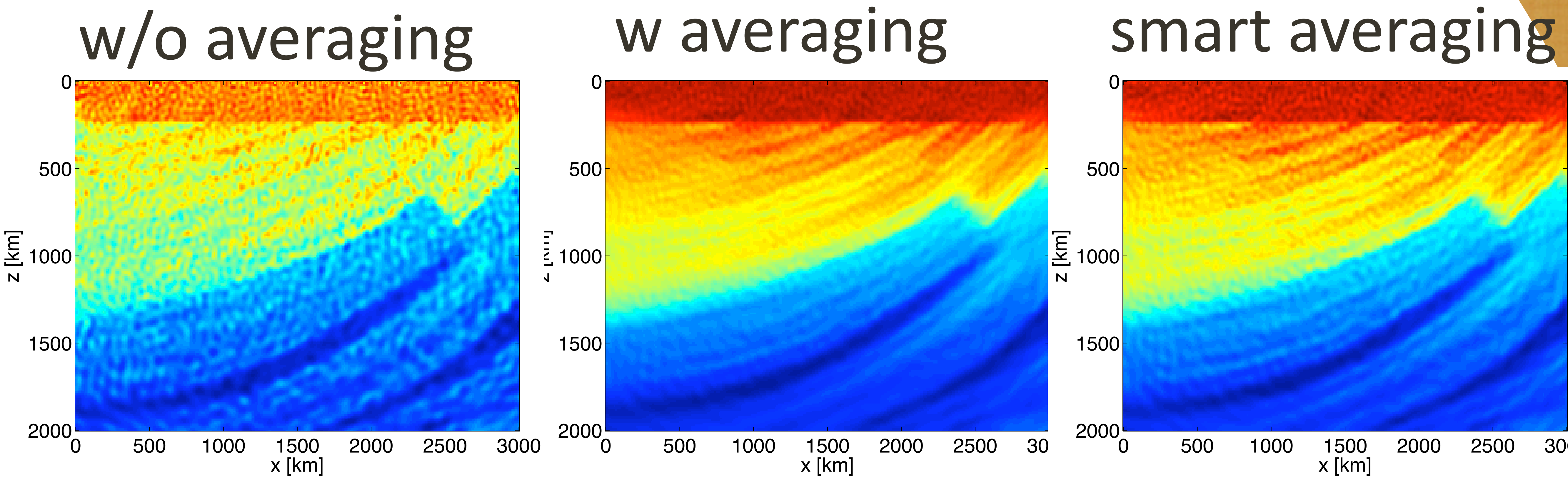


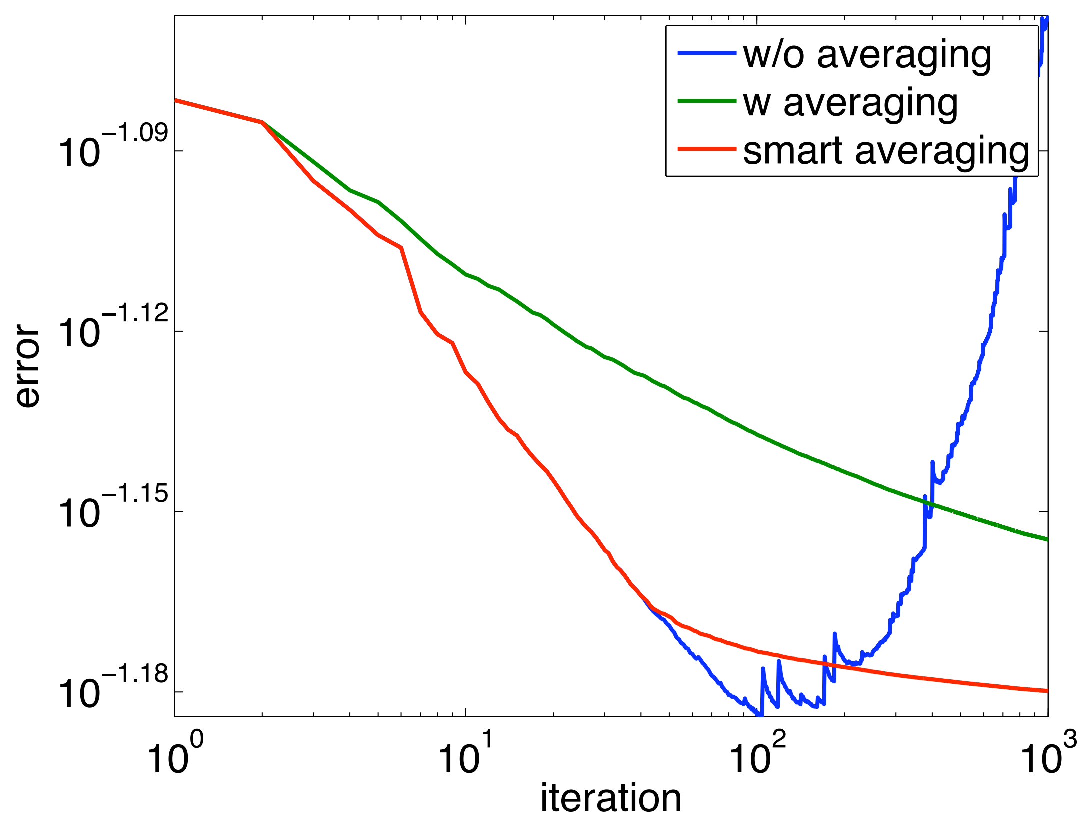
smart averaging





K=5 [noisy case]





Sources of noise

Noise contributions

- Noisy data
- *Interference* noise (source cross talk & aliases)
- *Inter gradient* noise (renewals)

can lead to a noise level that is too high

- leads to divergence

SLIM

Observations

Renewals improve convergence *significantly*

Averaging removes noise *instability* but is *detrimental* to the *convergence*

Smart averaging over *limited history* improves *convergence*

Increasing the *batch size* in *combination* with *smart averaging* leads to *superior convergence*

Alternative I

[*integrated* stochastic gradient descend]

Average the *gradients* instead, i.e.,

$$\mathbf{m}_{k+1} = \mathbf{m}_k - \eta_k \overline{\nabla \mathcal{F}(\mathbf{m}_k)}$$

with

$$\overline{\nabla \mathcal{F}(\mathbf{m}_k)} = \frac{\sum_{i=k-m}^{k} e^{\alpha[i-k-m]} \nabla \mathcal{F}(\mathbf{m}_i)}{\sum_{i=k-m}^{k} e^{\alpha[i-k-m]}}$$

over *last m iterations*.

SLIM

Case study I

1. Measure performance of l-BFGS with *renewals* as *function* of the *batch size K*
2. Compare l-BFGS on *complete* data with *integrated stochastic gradient descend* (iSGD)

Experimental setup

Marmousi model:

- 10 m grid spacing (3000 X 5000 m)
- 113 shots with 40m spacing and offsets 250-4749m
- 249 receivers with 20m spacing and offsets 20-4980m
- Ricker wavelet with central frequency of 10Hz
- 3.6s recording time with 0.009s sample interval

FWI setup

l-BFGS (reference):

- 50 frequencies between 5-33Hz
- 18 iterations

integrated Stochastic Gradient Descend

- *randomized simultaneous* shots
- *randomly* selected *frequencies* between 5-33Hz

SLIM

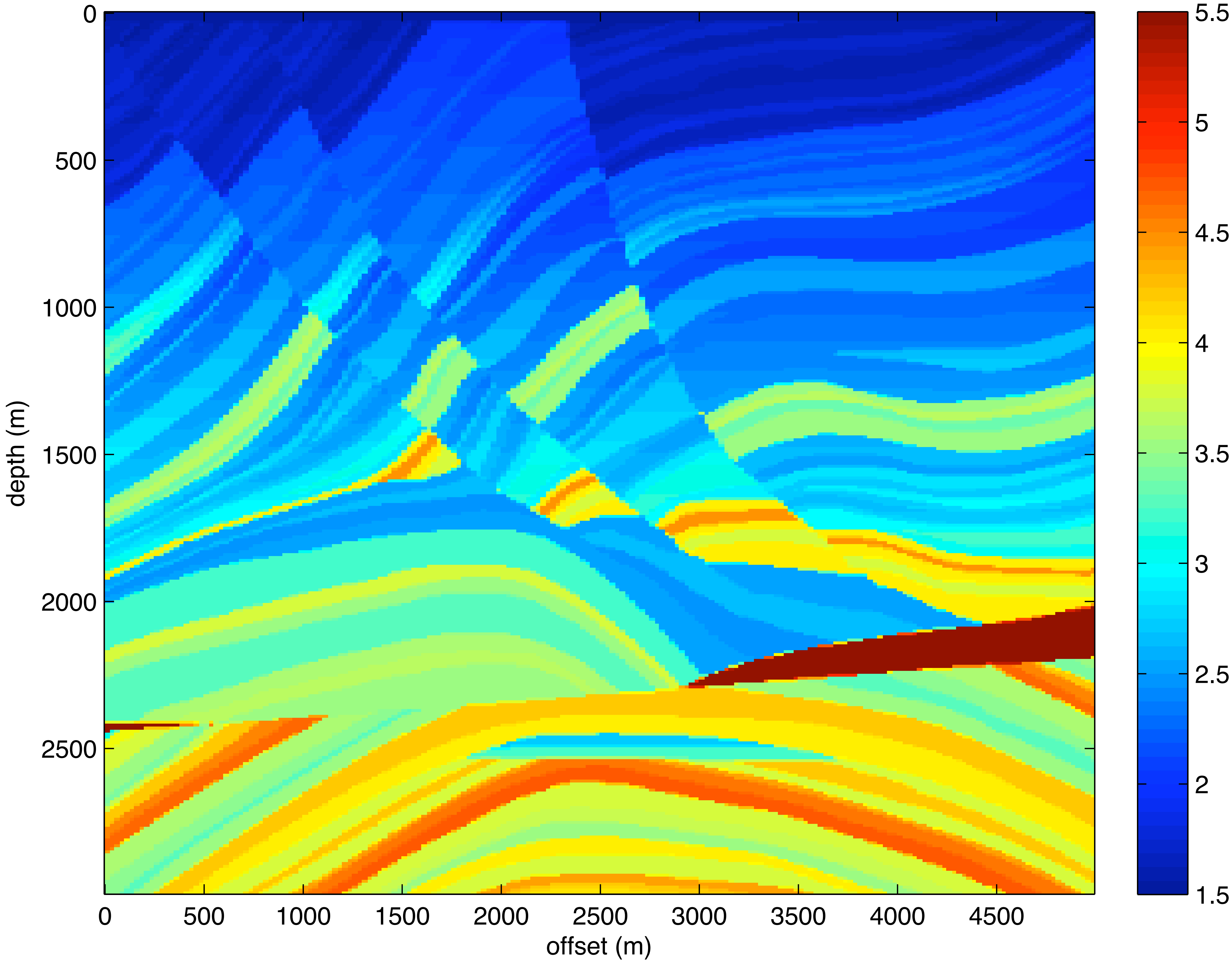
Performance
[l-BFGS w renewals]

Subsample ratio	0.0113	0.0028	0.0007
n'_f/n'_s	recovery error (dB)		
.25	**6.46**	**3.31**	**0.78**
1	**3.22**	**2.17**	**0.74**
4	**3.66**	**3.10**	**0.45**
Speed up ($\times$)	88	352	1410

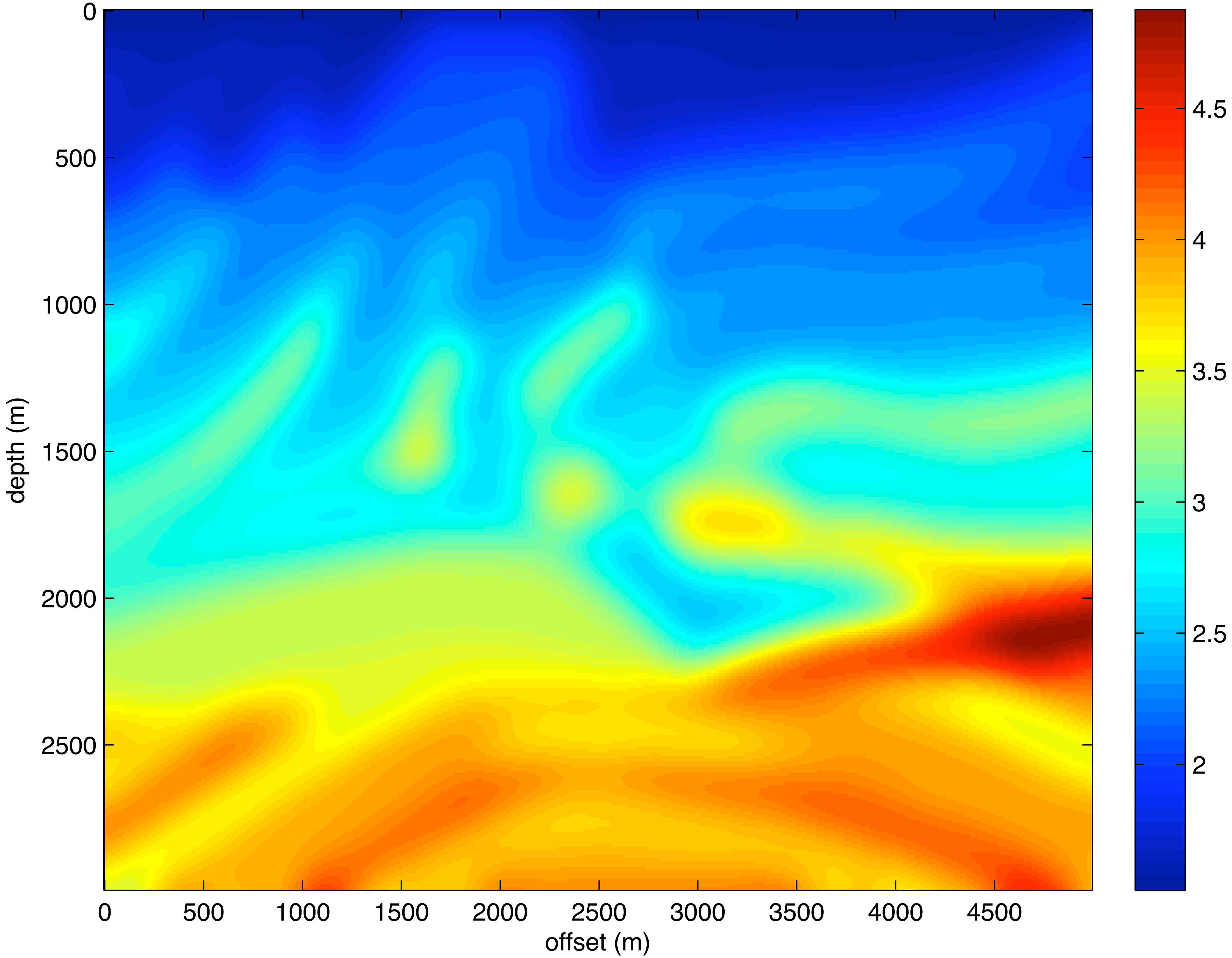
Exhaustive search

α	-.2	-.1	0	.1	.2	.3	.4
SNR(dB)	**1.7712**	**2.0776**	**2.0199**	**2.9072**	**5.2496**	**7.0717**	**7.2719**
α	.5	.6	.7	.8	.9	1	
SNR(dB)	**7.8315**	**6.5770**	**6.7162**	**7.4953**	**5.8569**	**5.9605**	

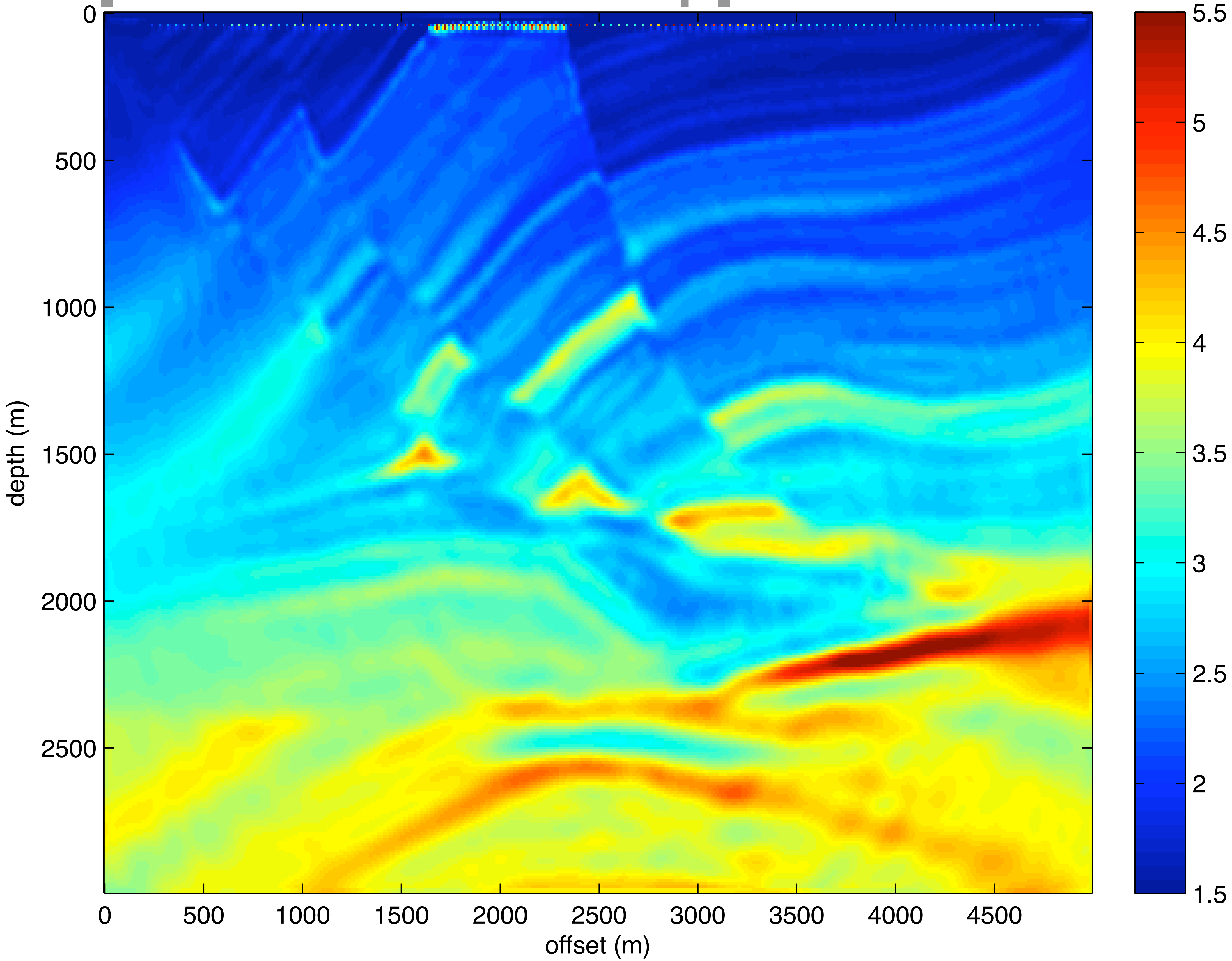
True model


0
500
1000
1500
2000
2500
depth (m)
0
500
1000
1500
2000
2500
3000
3500
4000
4500
offset (m)
5.5
5
4.5
4
3.5
3
2.5
2
1.5

Initial model

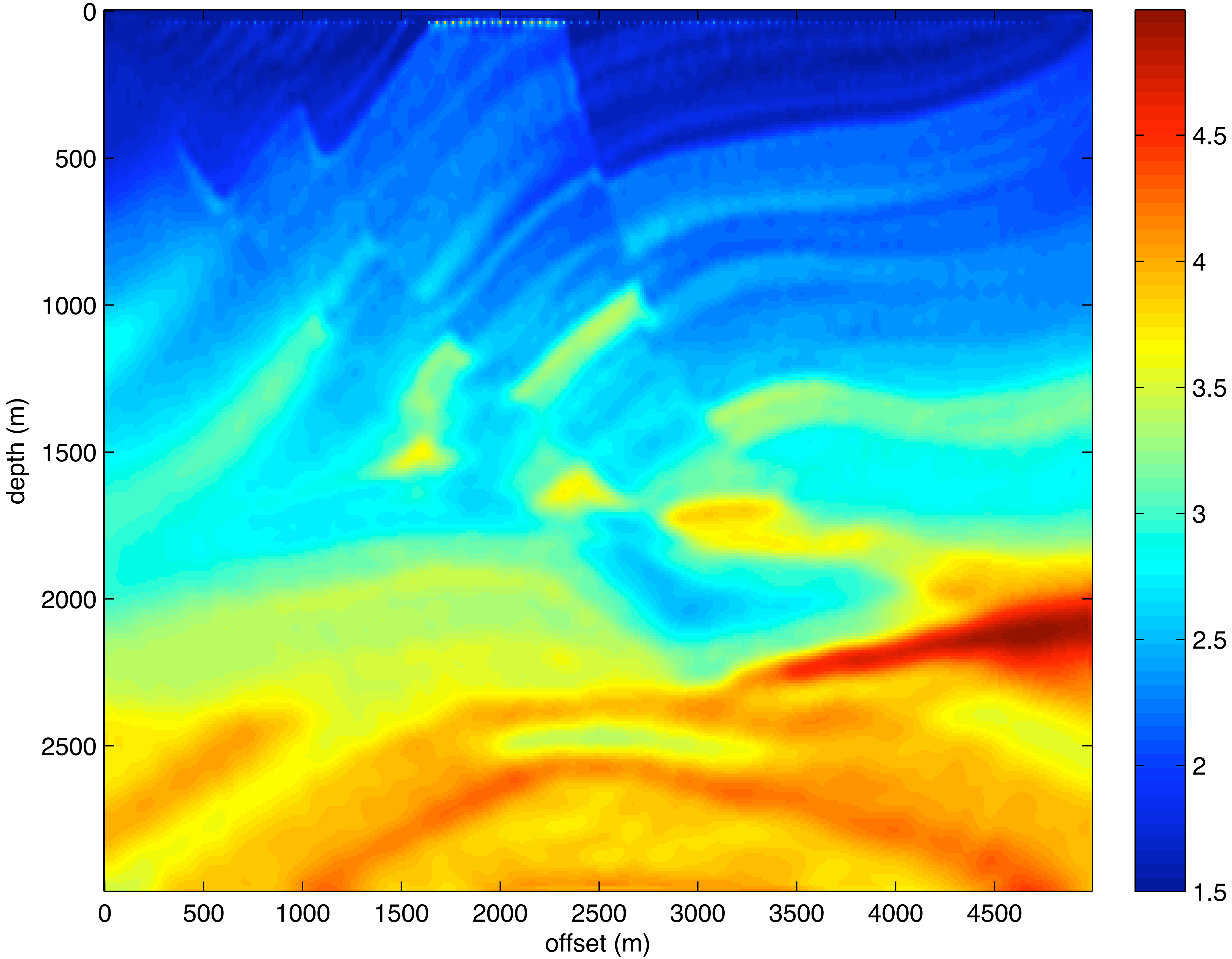


Reference
[all 113 shots & 50 freqs]



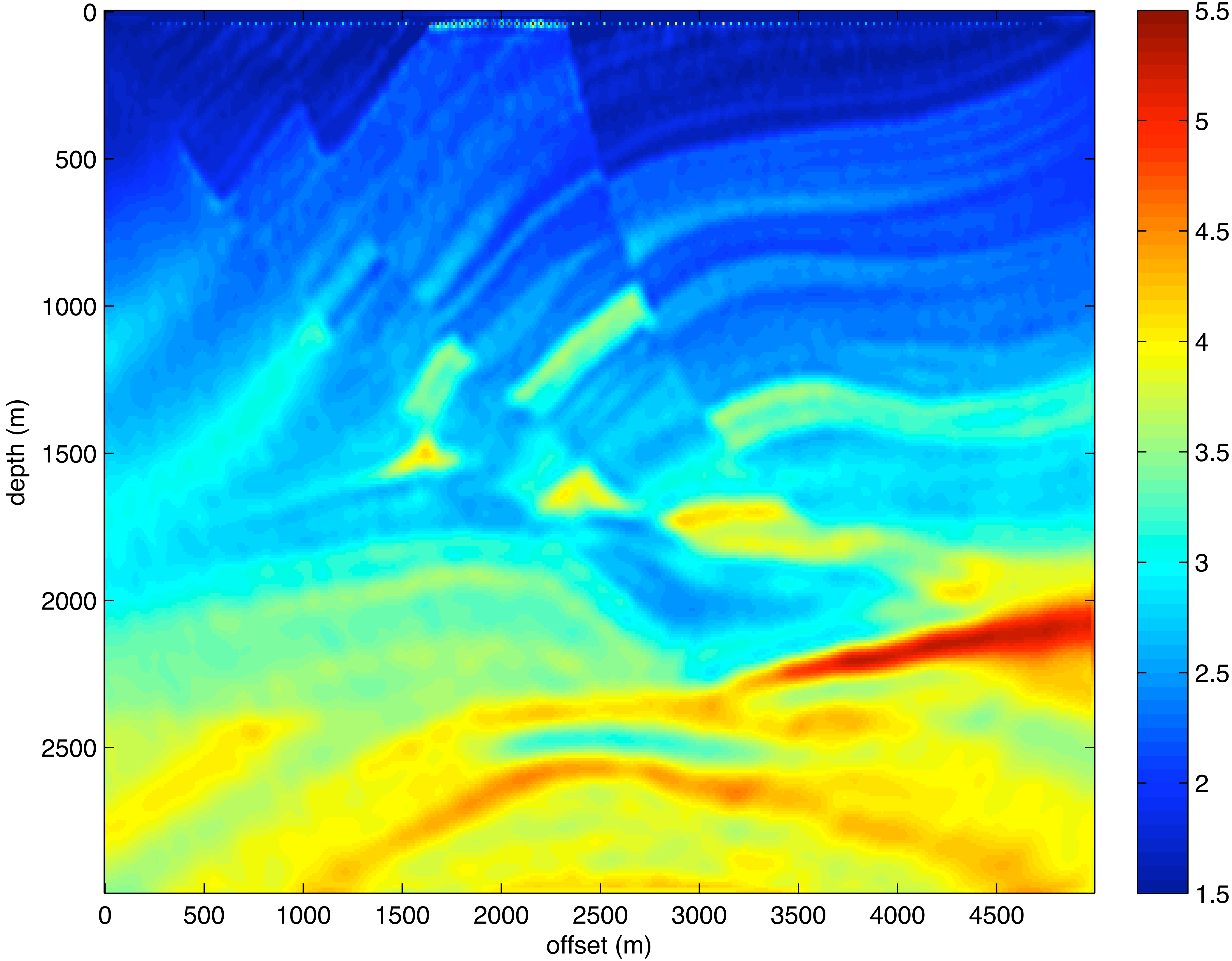
SGD

[16 sim shot & 4 freq: 40X, 4.65 dB]

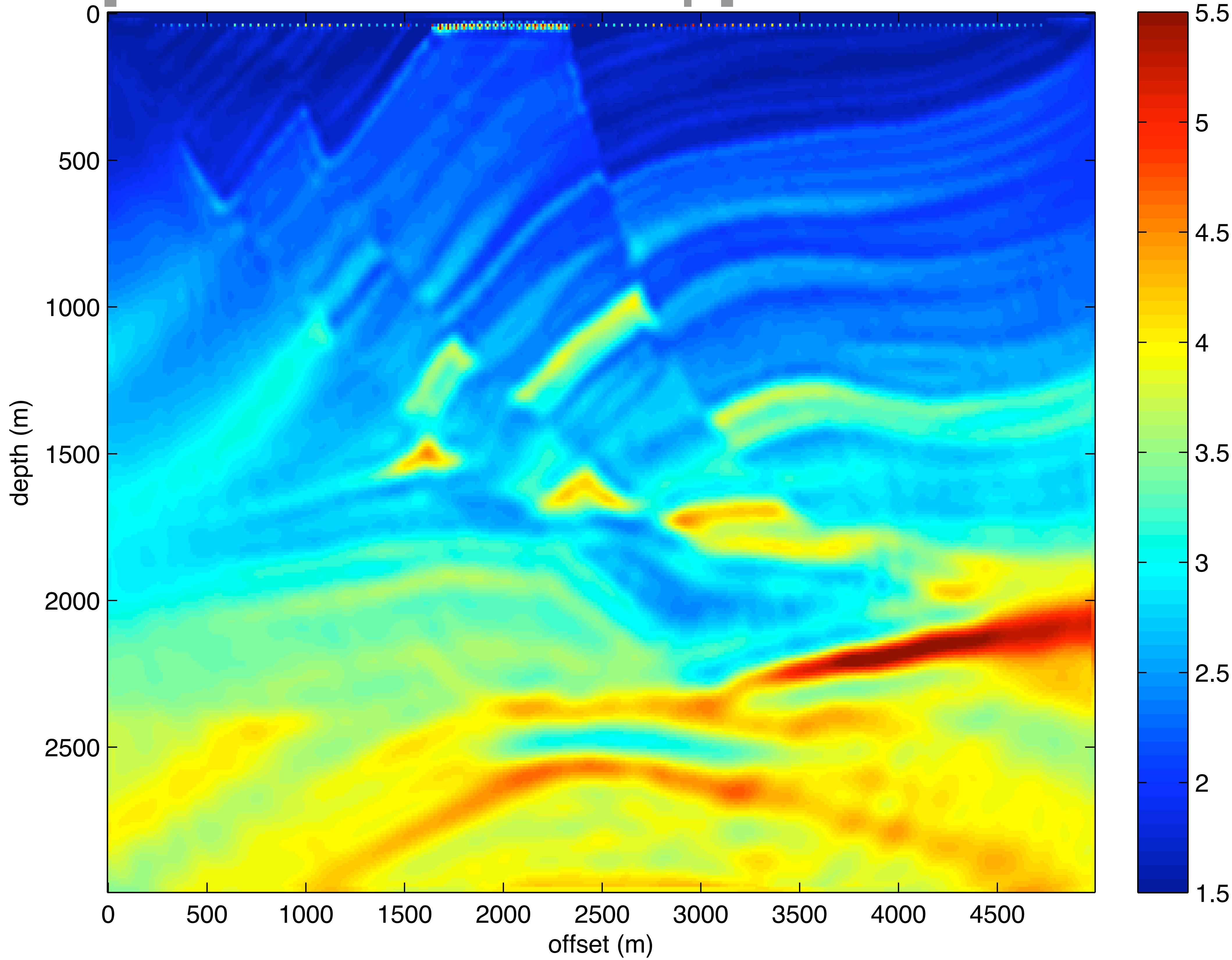


iSGD

[16 sim shot & 4 freq: 40X, 9.10 dB]



Reference
[all 113 shots & 50 freqs]



Observations

Averaging of *gradients* damps *stochasticity*

'Ad hoc' weighted averaging of *iSGD* leads to a *significant acceleration*

Consistent with *asymptotic* theory for *first-order SGD* [Bertsekas, '96]

Formulation is *amenable to incomplete acquisition* [Haber, Chung, and FJH, '10]

Results remain *noisy*, and lack *sharp edges*

Alternative II

Leverage findings from *sparse recovery* & *compressive sensing*

- consider each *phase-encoded* Gauss-Newton update as separate *compressive-sensing* experiment
- remove *interferences* by *curvelet-domain sparsity* promotion
- exploit properties of the Pareto curve

[Candes et al., ’06; Donoho, ’06]

[Demanet et. al. ’07; FJH & Li, ’08-’09]

Gauss-Newton

Algorithm 1: Vanilla Gauss Newton

Result: Output estimate for the model $\mathbf{m}$

$\mathbf{m} \longleftarrow \mathbf{m}_0$; $k \longleftarrow 0$; `// initial model`

while not converged **do**

  $\mathbf{p}^k \longleftarrow \arg\min_{\mathbf{p}} \frac{1}{2}\|\delta\mathbf{d} - \nabla\boldsymbol{\mathcal{F}}[\mathbf{m}^k;\mathbf{Q}]\mathbf{p}\|_2^2 + \lambda^k\|\mathbf{p}\|_2^2$; `// search dir.`

  $\mathbf{m}^{k+1} \longleftarrow \mathbf{m}^k + \gamma^k\mathbf{p}^k$; `// update with linesearch`

  $k \longleftarrow k+1$;

end

Compressive updates

Algorithm 1: Gauss Newton with sparse updates

Result: Output estimate for the model $\mathbf{m}$

$\mathbf{m} \longleftarrow \mathbf{m}_0$; $k \longleftarrow 0$; `// initial model`

while not converged **do**

$\mathbf{p}^k \longleftarrow \mathbf{S}^* \arg\min_{\mathbf{x}} \frac{1}{2}\|\delta\underline{\mathbf{d}}^k - \nabla\mathcal{F}[\mathbf{m}^k;\underline{\mathbf{Q}}^k]\mathbf{S}^*\mathbf{x}\|_2^2 \quad \text{s.t.} \quad \|\mathbf{x}\|_1 \leq \tau^k$

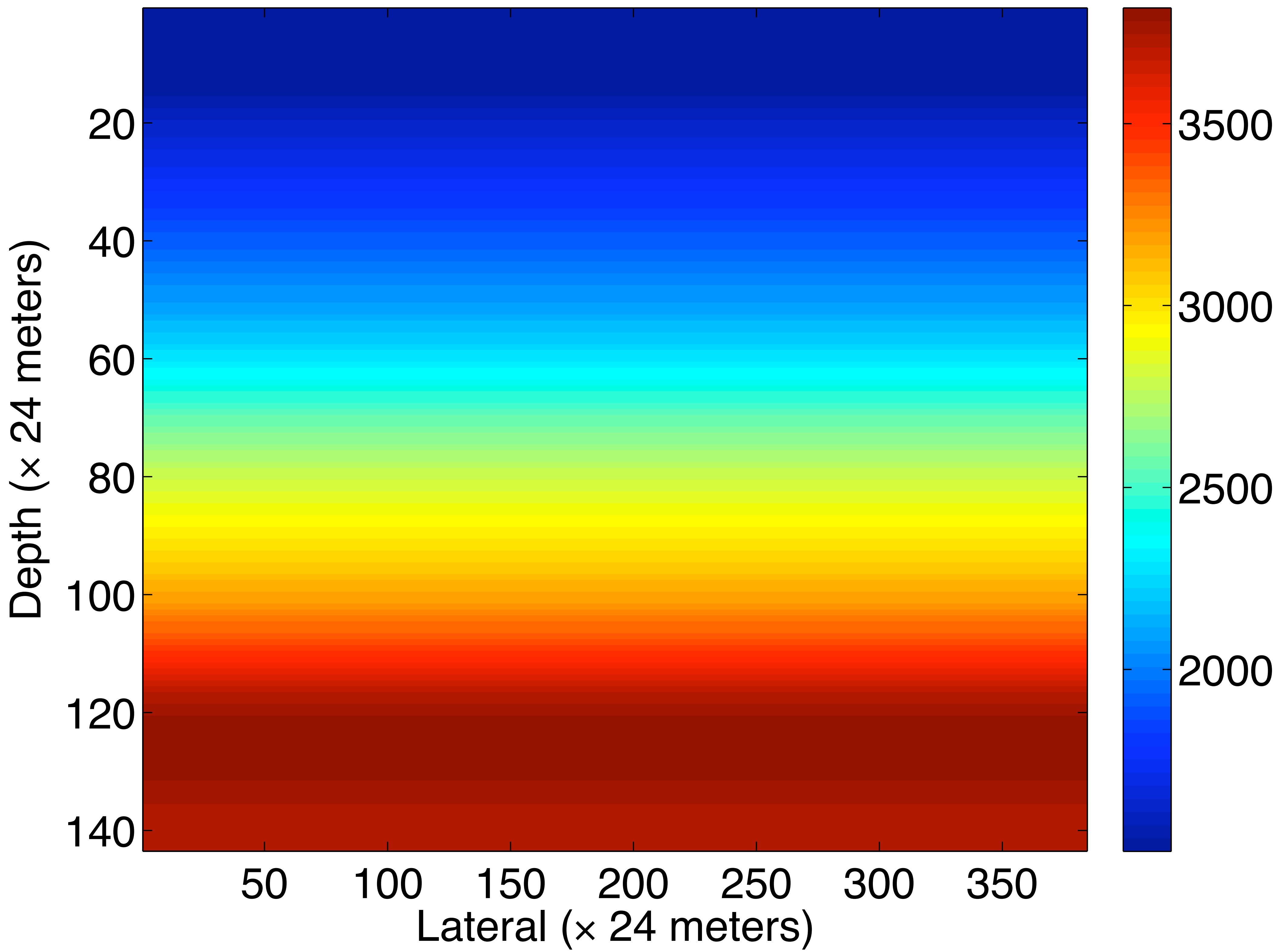
$\mathbf{m}^{k+1} \longleftarrow \mathbf{m}^k + \gamma^k \mathbf{p}^k$; `// update with linesearch`

$k \longleftarrow k+1$;

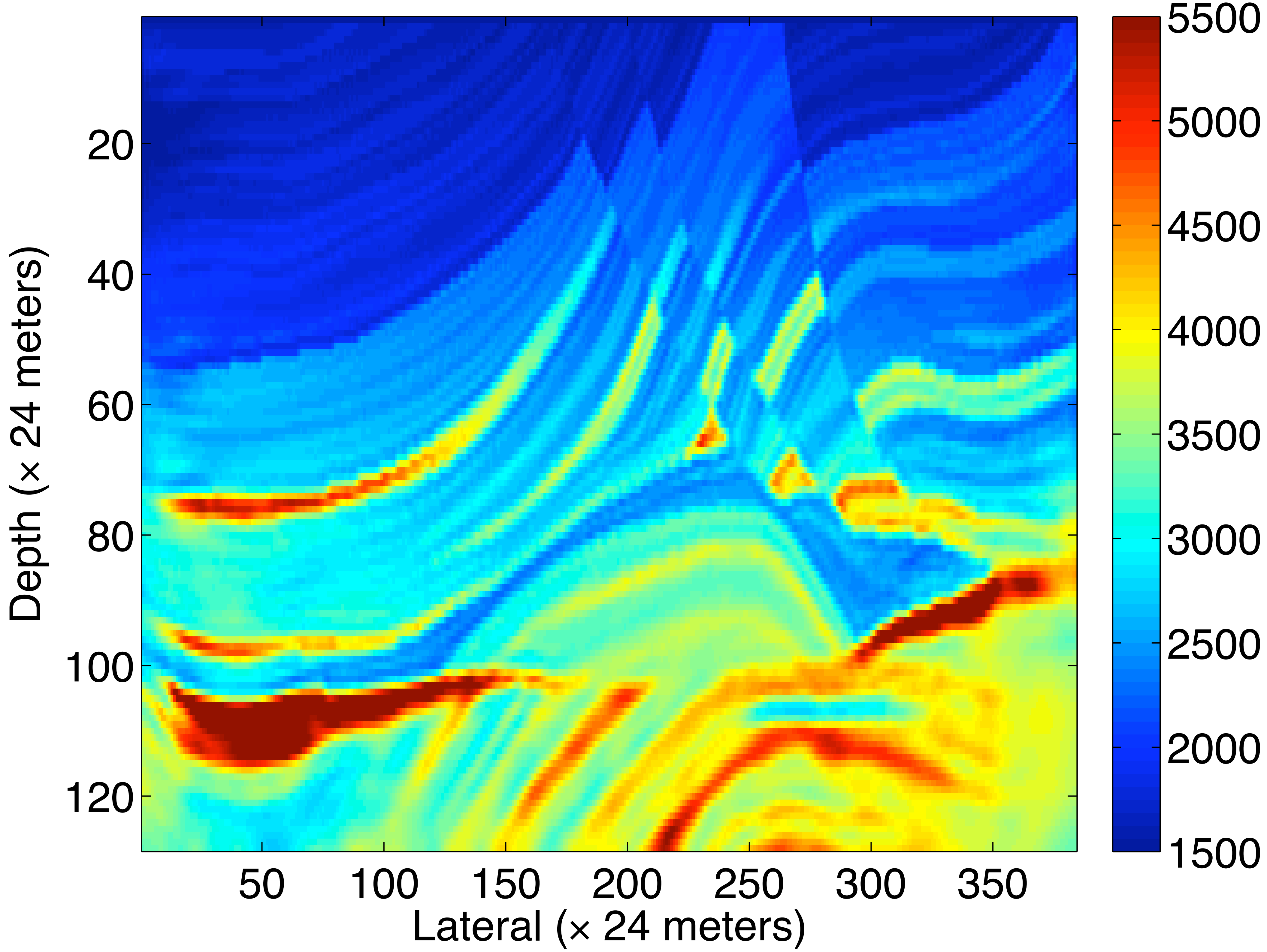
end

[Li & FJH, '10; van den Berg & Friedlander, '08]

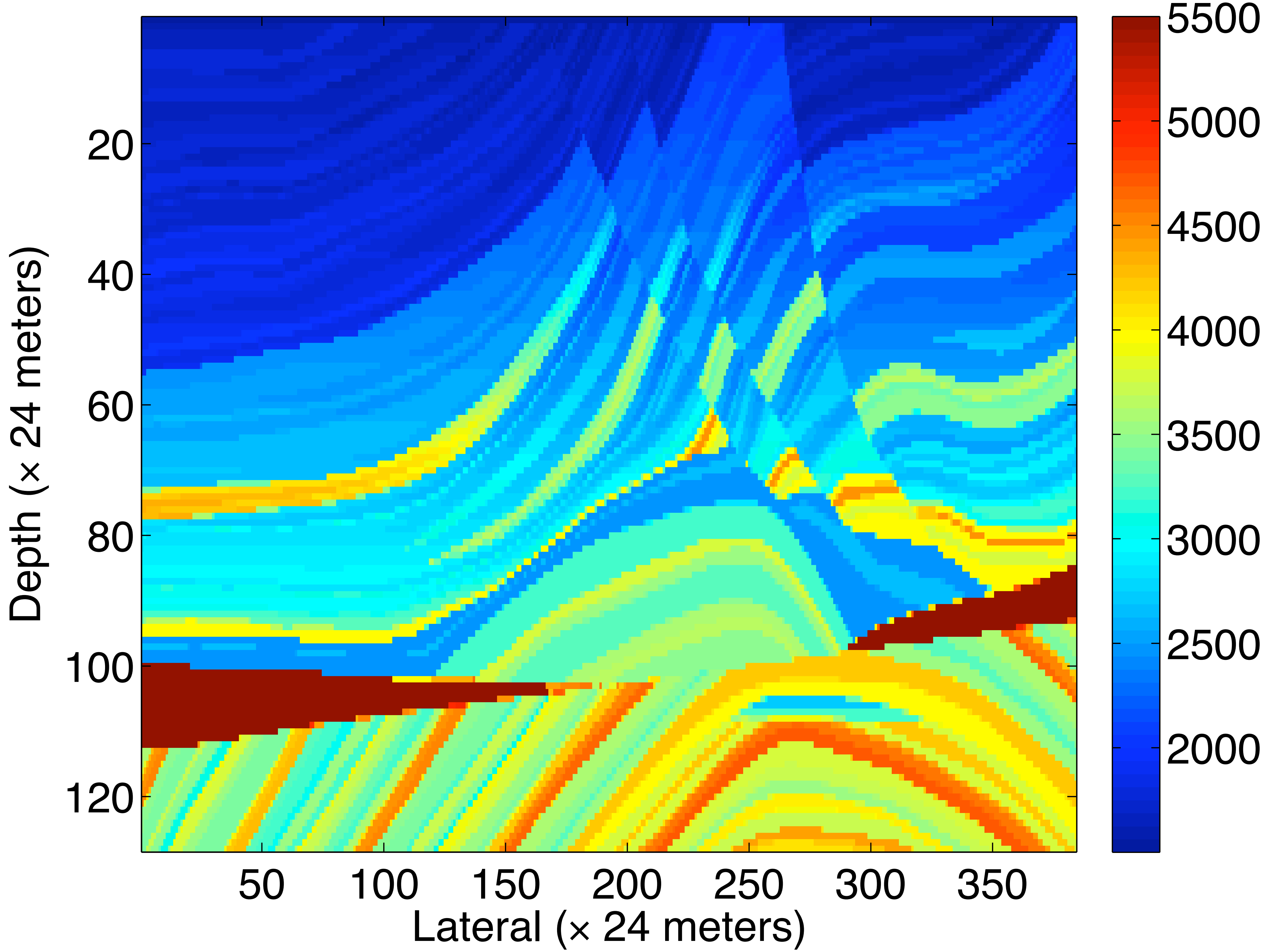
Initial model



Inverted model



True model



Depth (× 24 meters)
20
40
60
80
100
120
50
100
150
200
250
300
350
Lateral (× 24 meters)
5500
5000
4500
4000
3500
3000
2500
2000

SLIM

Conclusions

We *established phase-encoded FWI* with *renewals* as an instance *stochastic approximation*

- understand factors that contribute to *noise sensitivity*
- *factors* that *stabilize*

Identified shortcoming of *slow decay* for the *error* as *batch size increases*

Indications that *compressive sensing* supersedes the *stochastic approximation by sparse recovery of* dimensionality reduced subproblems

See tomorrow's talk by Xiang Li, 10:35 am, Room 405/406

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.
- *High-resolution wave-equation amplitude-variation-with-ray-parameter (AVP) imaging with sparseness constraints* by Wang & Sacchi, '07
- 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 Bertsekas, '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
- *An effective method for parameter estimation with PDE constraints with multiple right hand sides. by* Eldad Haber, Matthias Chung, and Felix J. Herrmann. '10

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 II (375142-08).

We also would like to thank the authors of CurveLab.

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

NSERC
CRSNG

Thank you

slim.eos.ubc.ca