Fast waveform inversion *without* source encoding

T. van Leeuwen

joint work with F. Herrmann, M. Friedlander, A. Aravkin, M. Schmidt

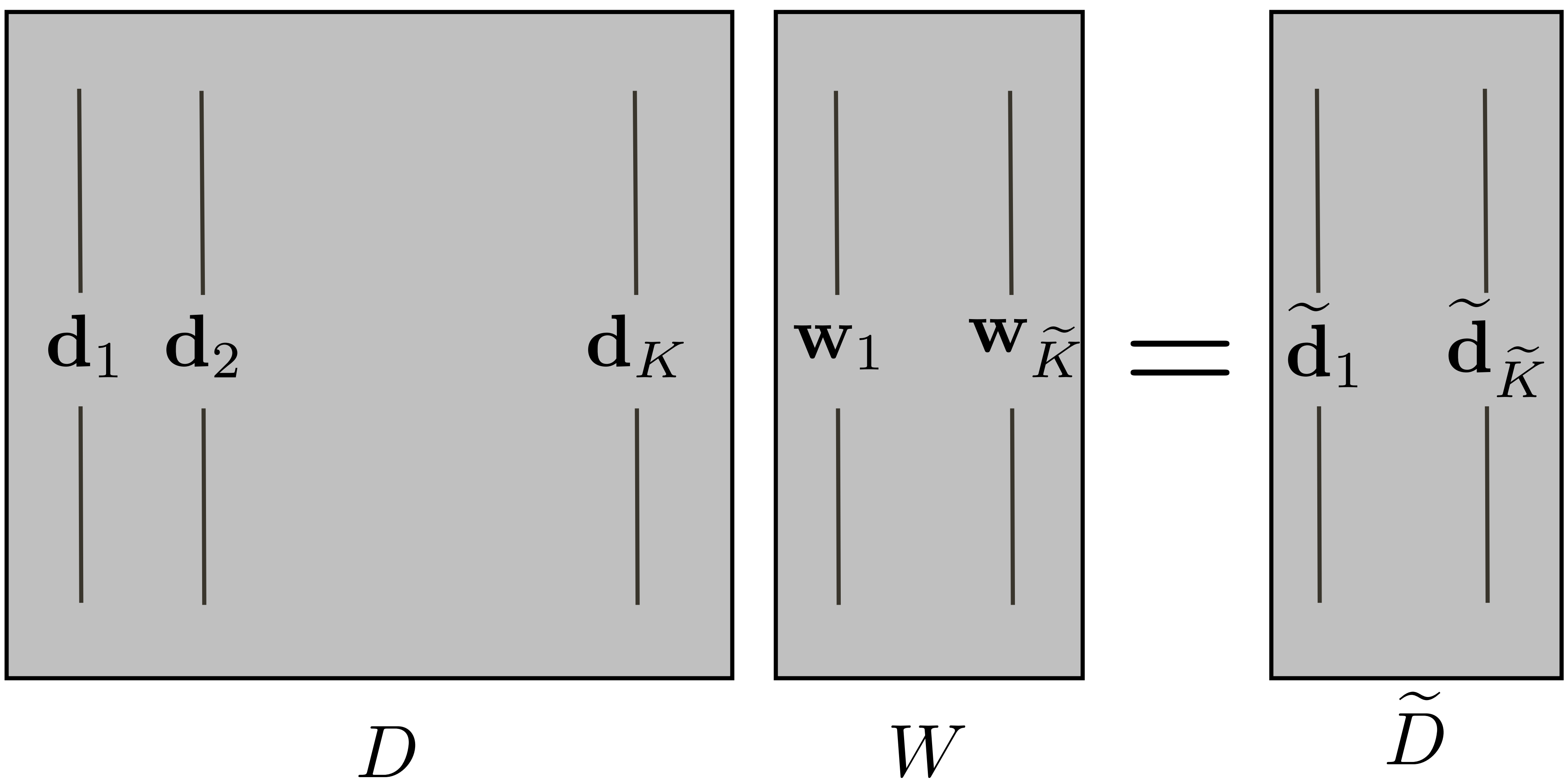
SLIM
University of British Columbia

Least-squares fitting of *multi-experiment* data that are linear in the source:

- *computational costs* are proportional to # of sources;
- can be reduced by *synthesizing* simultaneous sources from sequential data

[Beasley `98; Berkhout `08; Romero `00;]
[Ikelle `07; Neelamani `08;Herrmann `09]
[Krebs `09; Haber `10; TvL `10; Ben-Hadj-Ali `11]

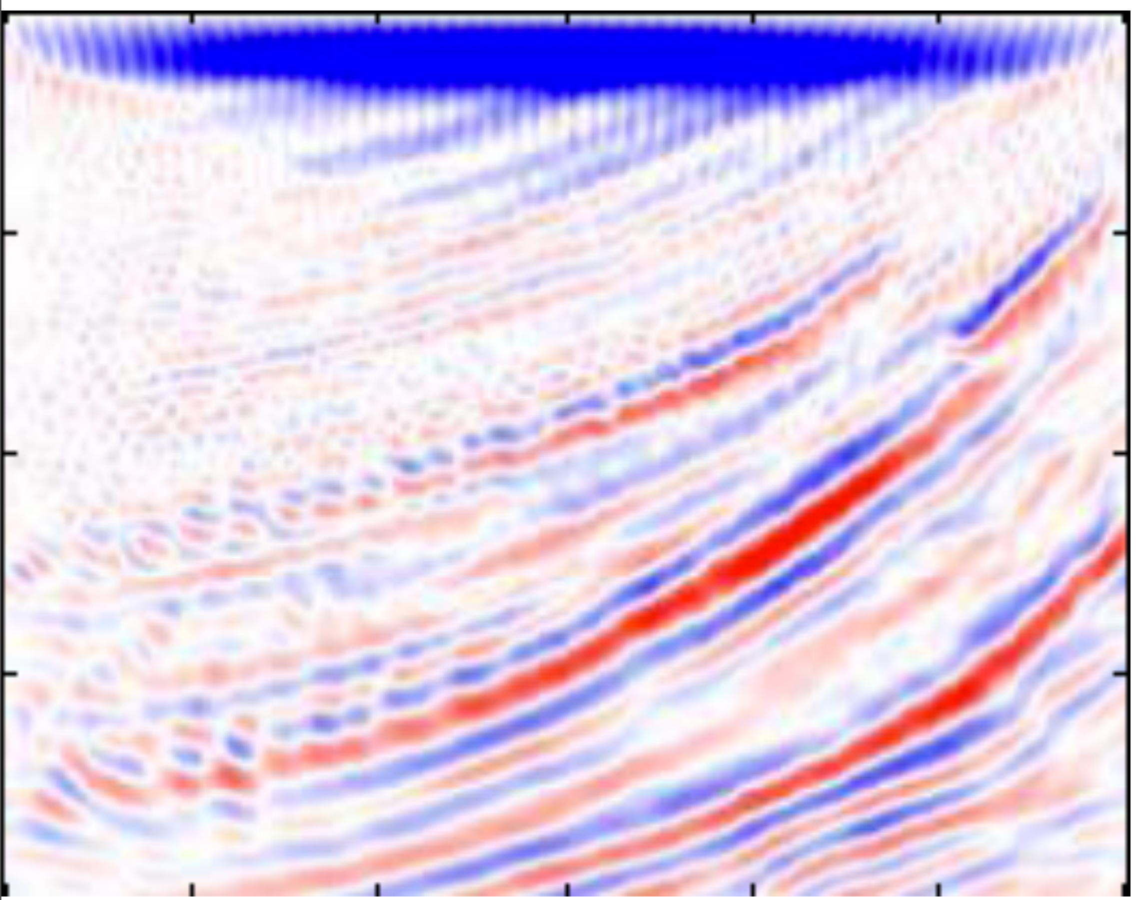
Replace data volume by `subsampled' volume



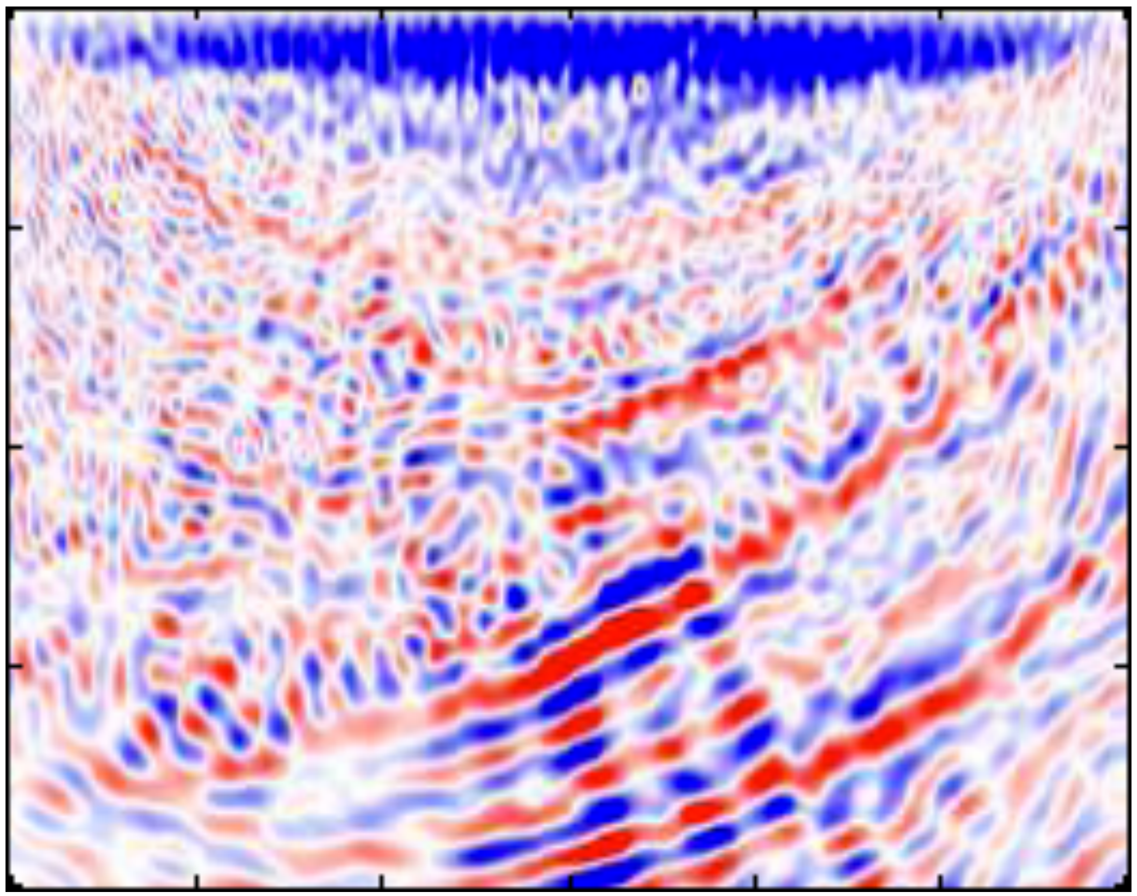
$$\widetilde{K} \ll K$$

What about `cross-talk'?

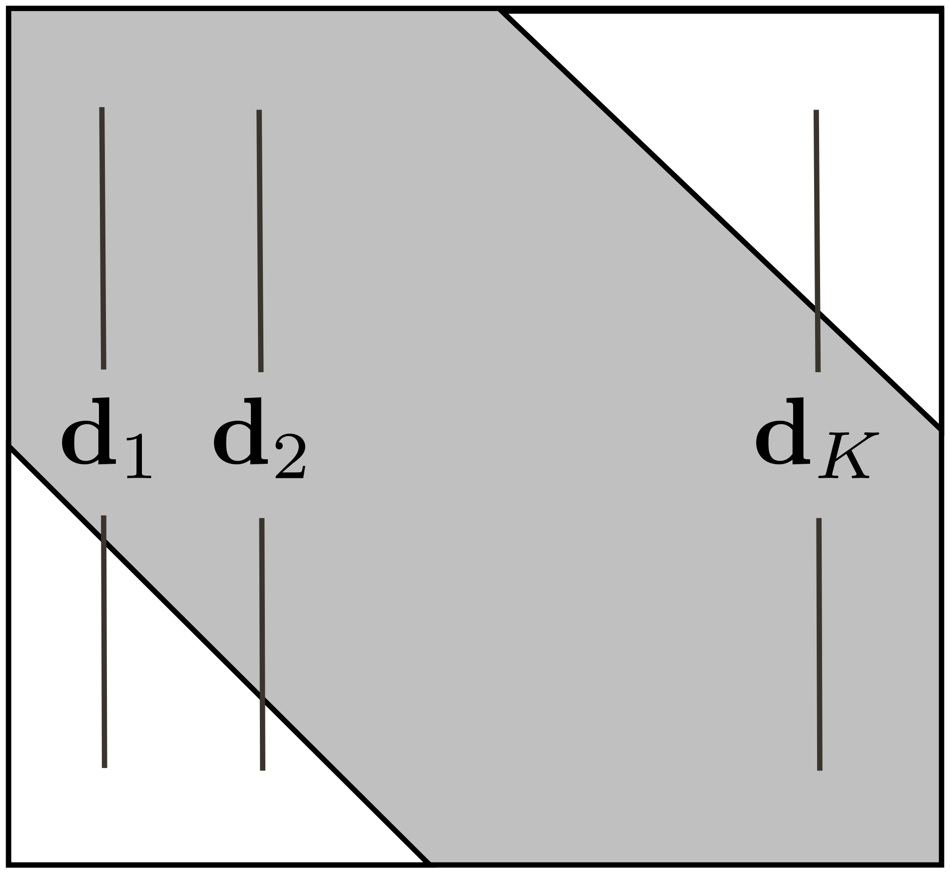
$$\sum_i \mathbf{u}_i \otimes \mathbf{v}_i$$



$$\left(\sum_i \mathbf{u}_i\right) \otimes \left(\sum_i \mathbf{v}_i\right)$$



What about `moving' receiver arrays?



Overview

- Approximating the misfit
- Optimization strategies
- Results
- Conclusions

SLIM

Approximating the misfit

Misfit is given by

$$\min_{\mathbf{m}} \Phi[\mathbf{m}] = \frac{1}{K}\sum_{i=1}^{K} \phi_i[\mathbf{m}]$$

$$\phi_i[\mathbf{m}] = ||\mathbf{d}_i - F[\mathbf{m}]\mathbf{q}_i||_2^2$$

Costs of evaluating the misfit are proportional to K.

Source encoding

Replace *sequential* sources by **one** *simultaneous* source: $\widetilde{\mathbf{q}} = \sum_j w_j \mathbf{q}_j$

$$\widetilde{\phi}[\mathbf{m}] = ||\widetilde{\mathbf{d}} - F[\mathbf{m}]\widetilde{\mathbf{q}}||_2^2$$

if $\mathsf{E}\{w_i w_j\} = \delta_{ij}$ we get $\mathsf{E}\left\{\widetilde{\phi}[\mathbf{m}]\right\} = \Phi[\mathbf{m}]$
and $\mathsf{E}\left\{\nabla\widetilde{\phi}[\mathbf{m}]\right\} = \nabla\Phi[\mathbf{m}]$

[Haber ’10; TVL ’10]

Trace estimation

Now replace *expectation* by *sample* average:

$$\widetilde{\Phi}[\mathbf{m}] = \frac{1}{\widetilde{K}} \sum_{i=1}^{\widetilde{K}} ||\widetilde{\mathbf{d}}_i - F[\mathbf{m}]\widetilde{\mathbf{q}}_i||_2^2$$

Can be seen as an instance of *trace* estimation:

$$||D - F[\mathbf{m}]Q||_F^2 \approx \frac{1}{\widetilde{K}} \sum_{i=1}^{\widetilde{K}} ||(D - F[\mathbf{m}]Q)\mathbf{w}_i||_2^2$$

Trace estimation

Some bounds for trace estimators in terms of (ϵ, δ)

$$\Pr(\text{error} \leq \epsilon) \geq 1 - \delta$$

estimator	variance	$\widetilde{K} \geq$
Gaussian	$2\|\|A\|\|_F^2$	$20\epsilon^{-2}\ln(2/\delta)$
random ± 1	$2\|\|A\|\|_F^2 - 2\sum_i a_{ii}^2$	$6\epsilon^{-2}\ln(2\mathrm{rank}(\mathrm{A})/\delta)$

[adapted from Avron '10]

Trace estimation

Error in the gradient:

$$||\nabla\Phi - \nabla\widetilde{\Phi}||_2^2 = \mathcal{O}(1/\widetilde{K})$$

with the *constants* dependent on the *random* weights.

[Hutchinson ‘89; Avron ’10]

Source encoding

Choice of *random* weights:

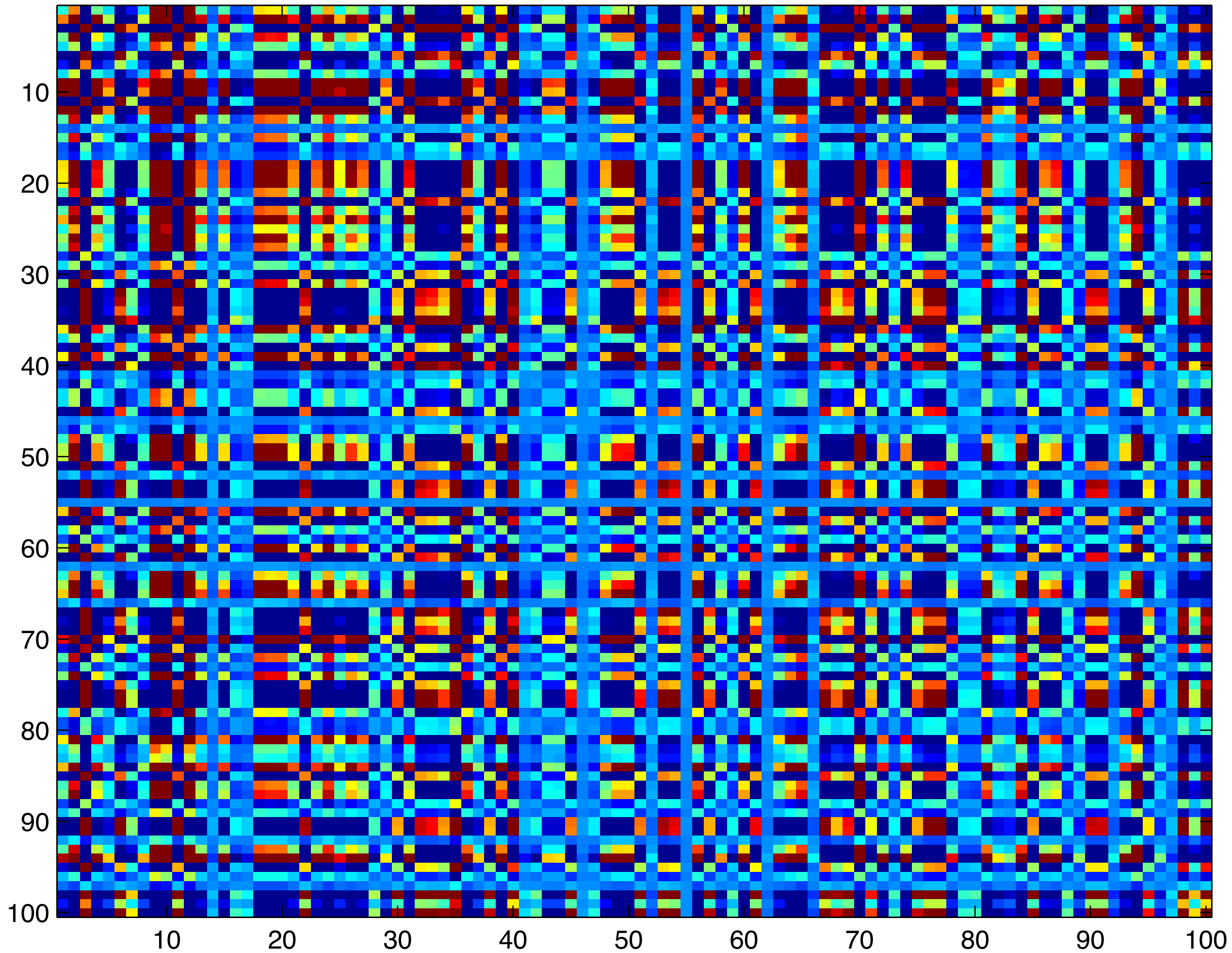
- Gaussian, ± 1, random phases: *efficient in sampling the whole matrix, but problematic for moving receiver array*

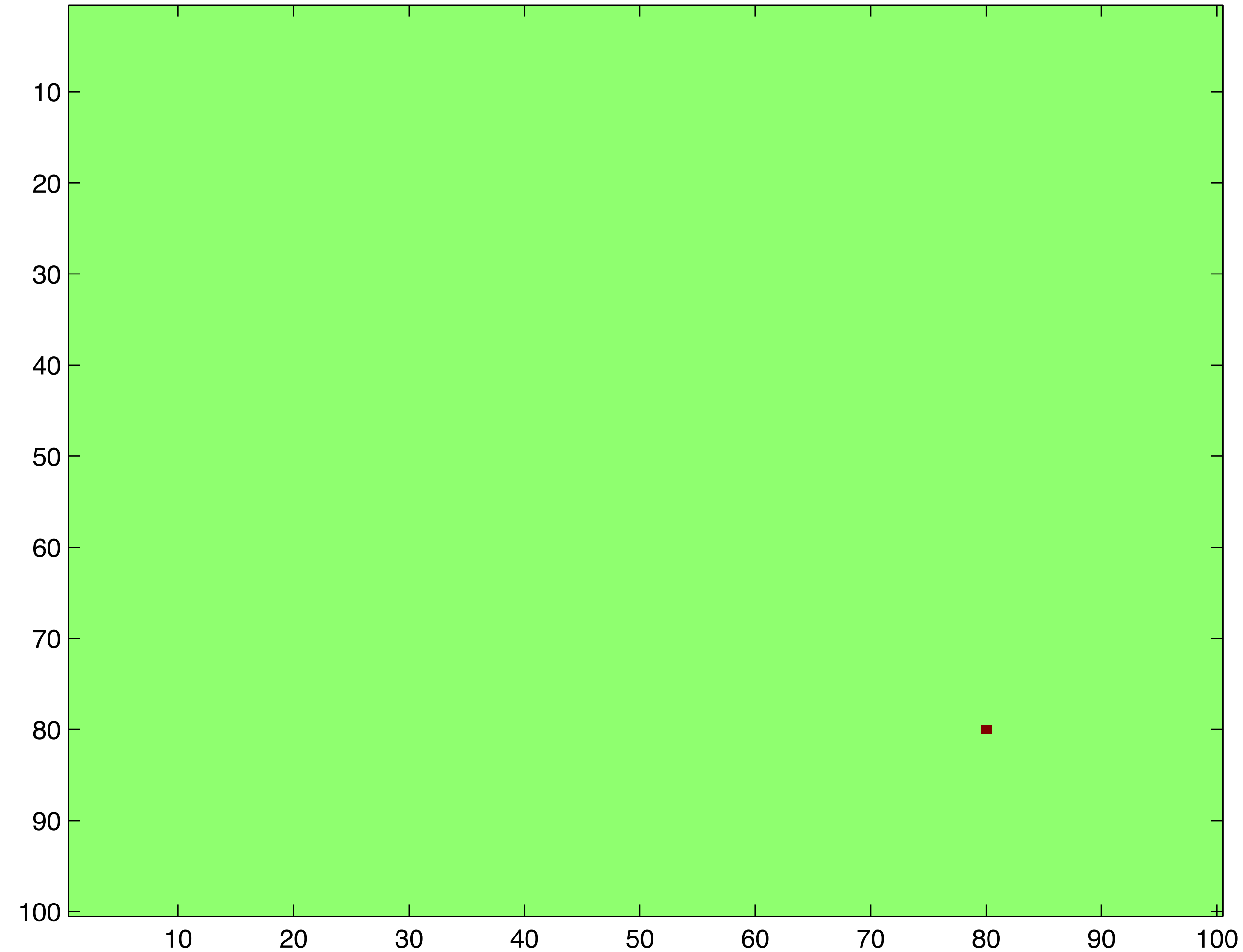
- Random unit vector: *less efficient sampling, but applicable for moving receiver array!*

Source encoding

$$\frac{1}{K}\sum_{i=0}^{K-1}\mathbf{w}_i\mathbf{w}_i^{\mathrm{T}} \approx I$$

K=1

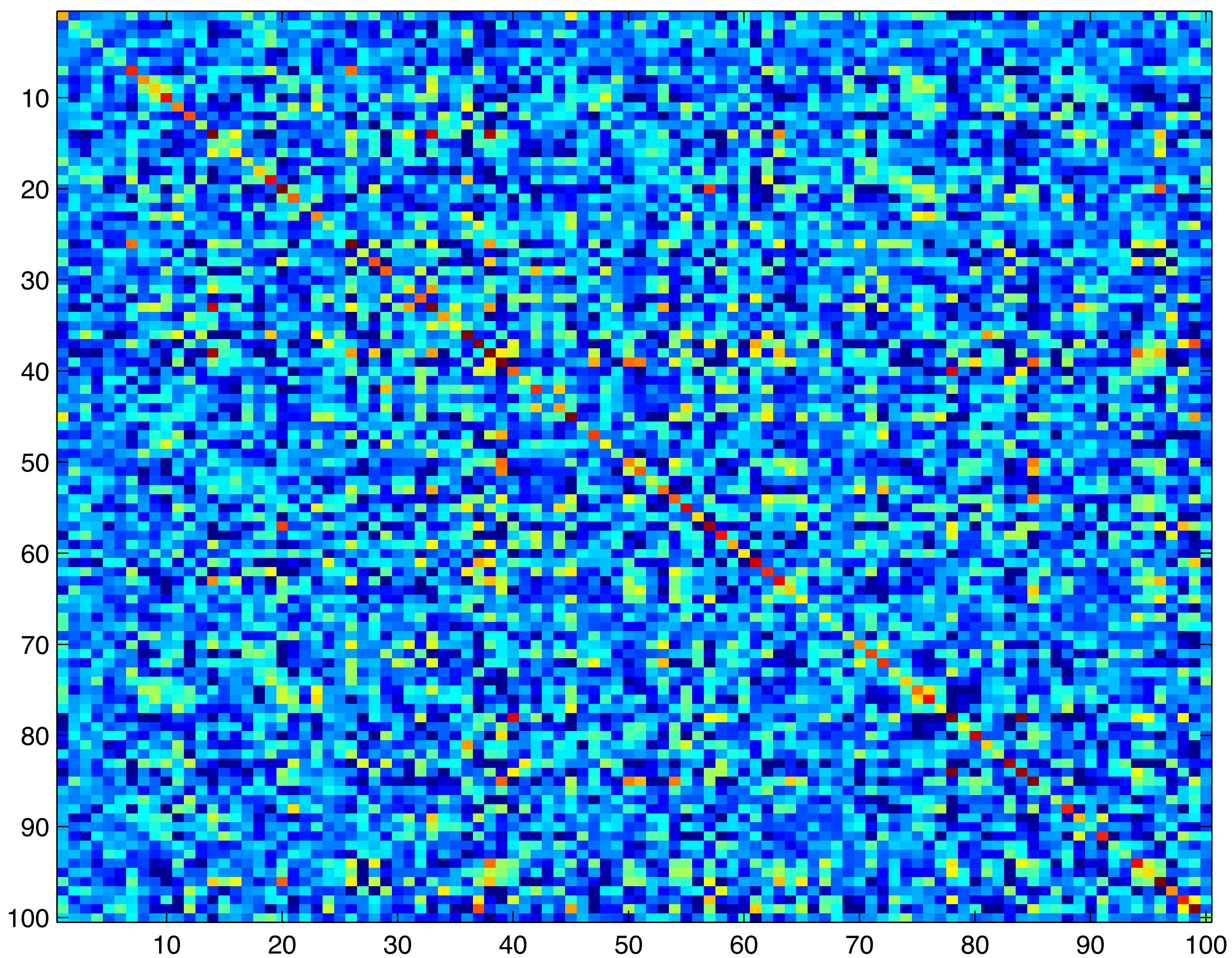


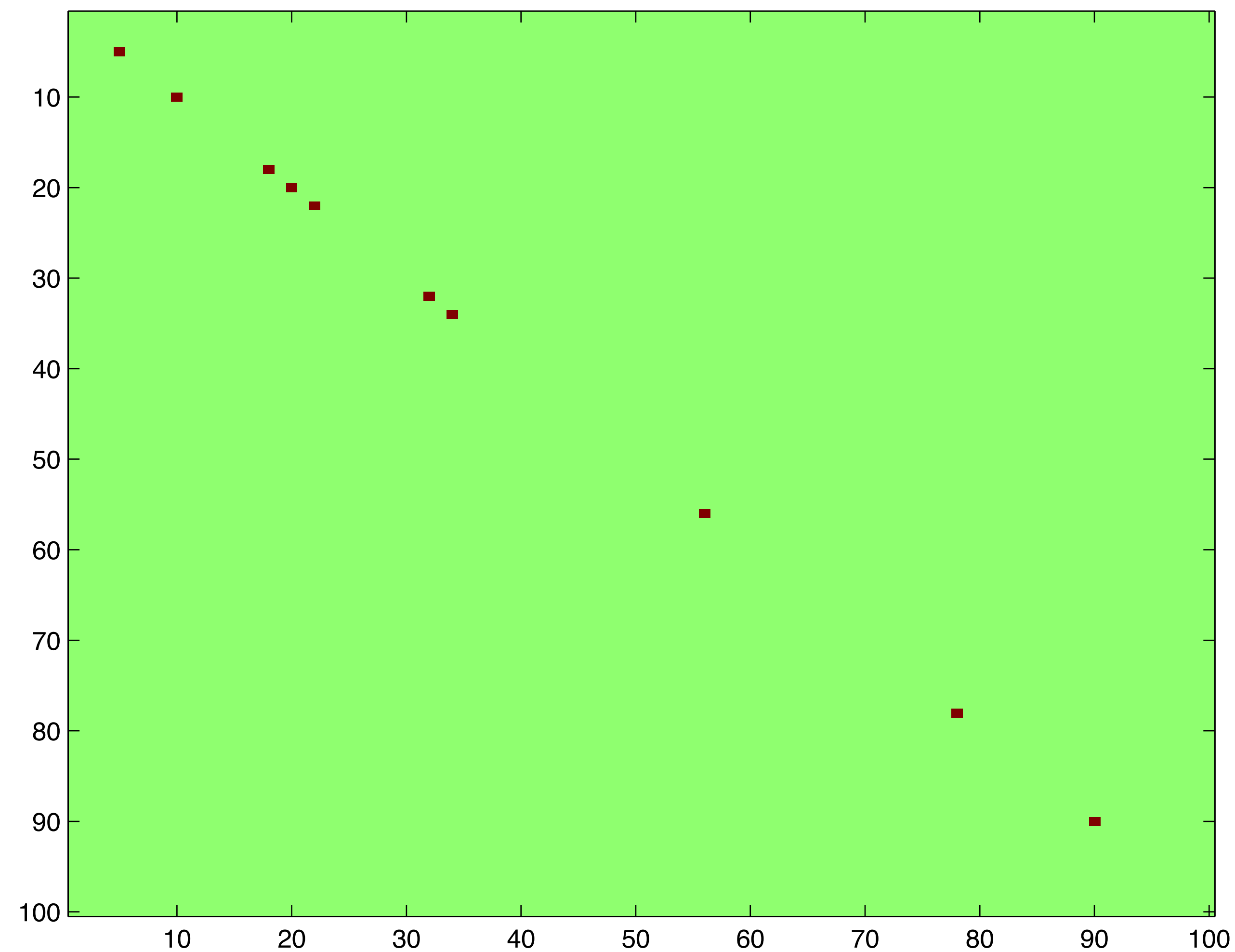


Source encoding

$$\frac{1}{K}\sum_{i=0}^{K-1}\mathbf{w}_i\mathbf{w}_i^{\mathrm{T}} \approx I$$

K=10

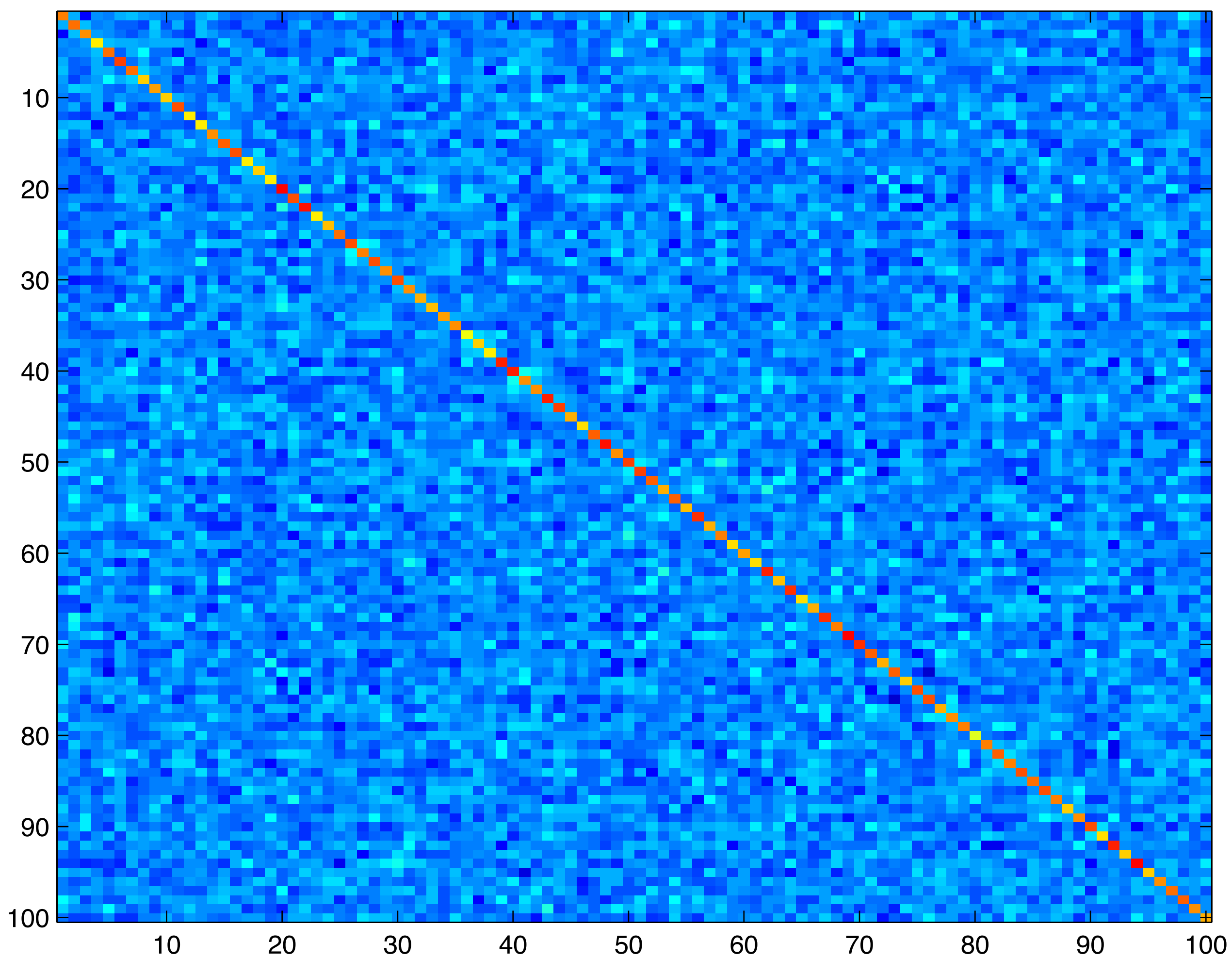


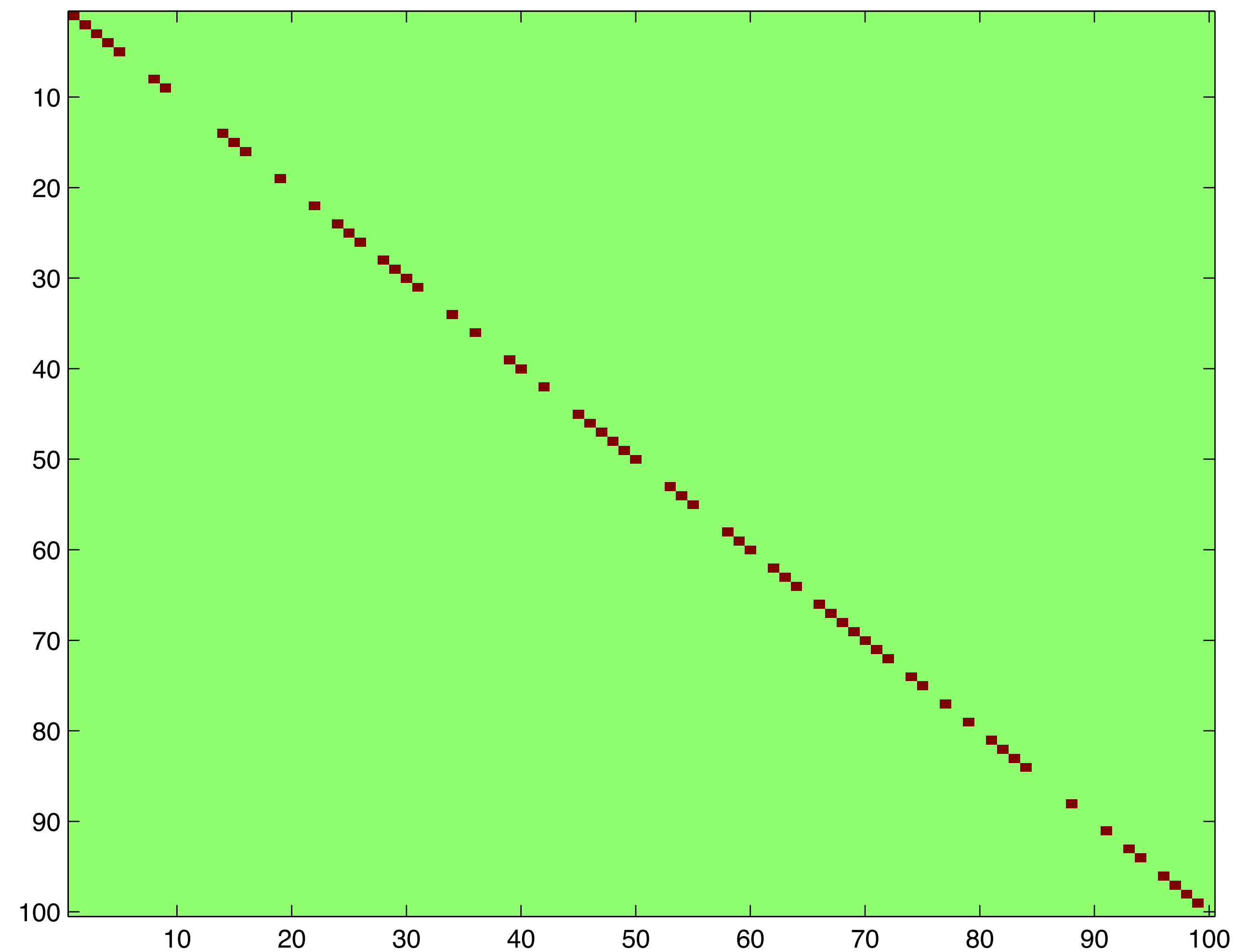


Source encoding

$$\frac{1}{K}\sum_{i=0}^{K-1}\mathbf{w}_i\mathbf{w}_i^{\mathrm{T}} \approx I$$

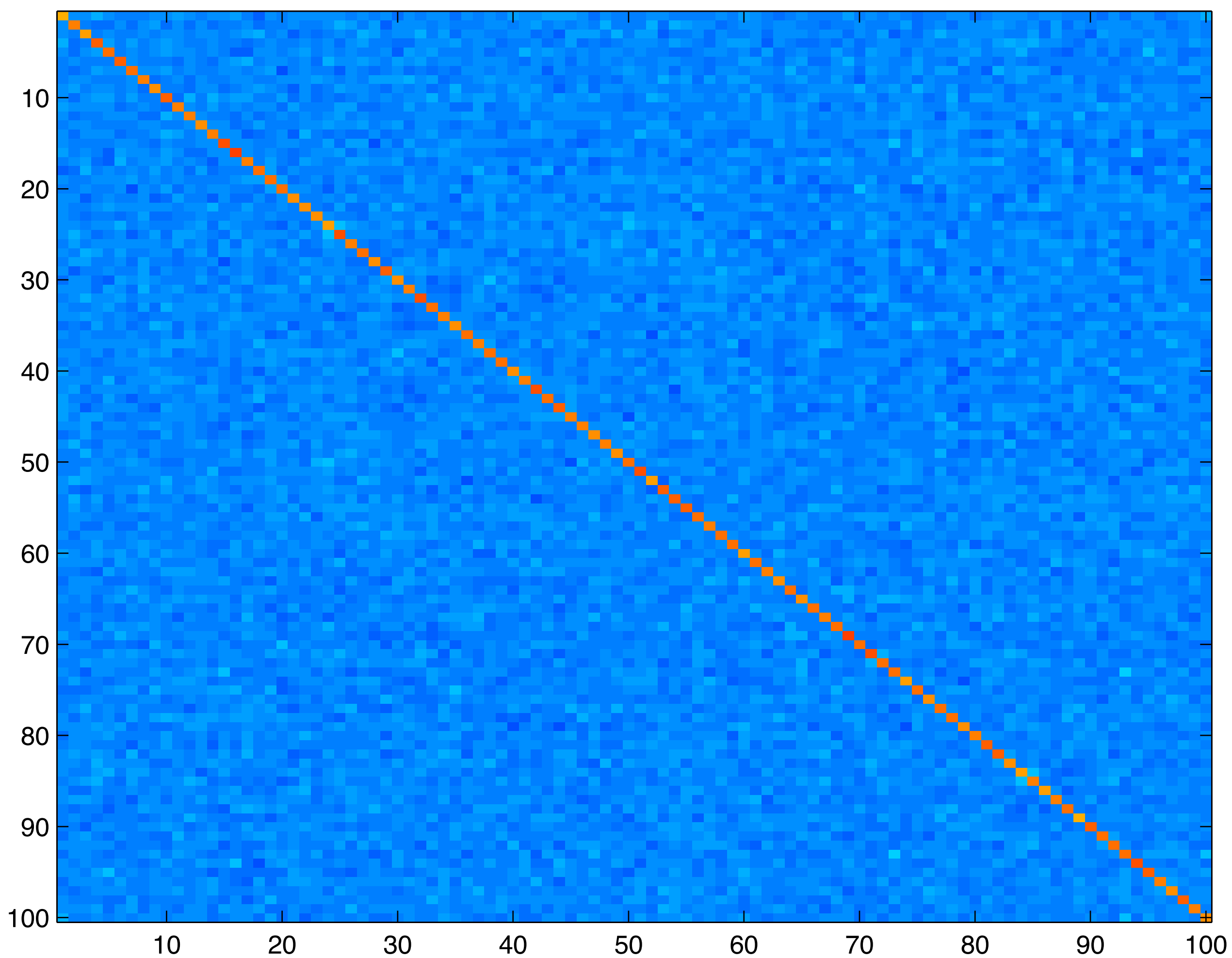
K=100

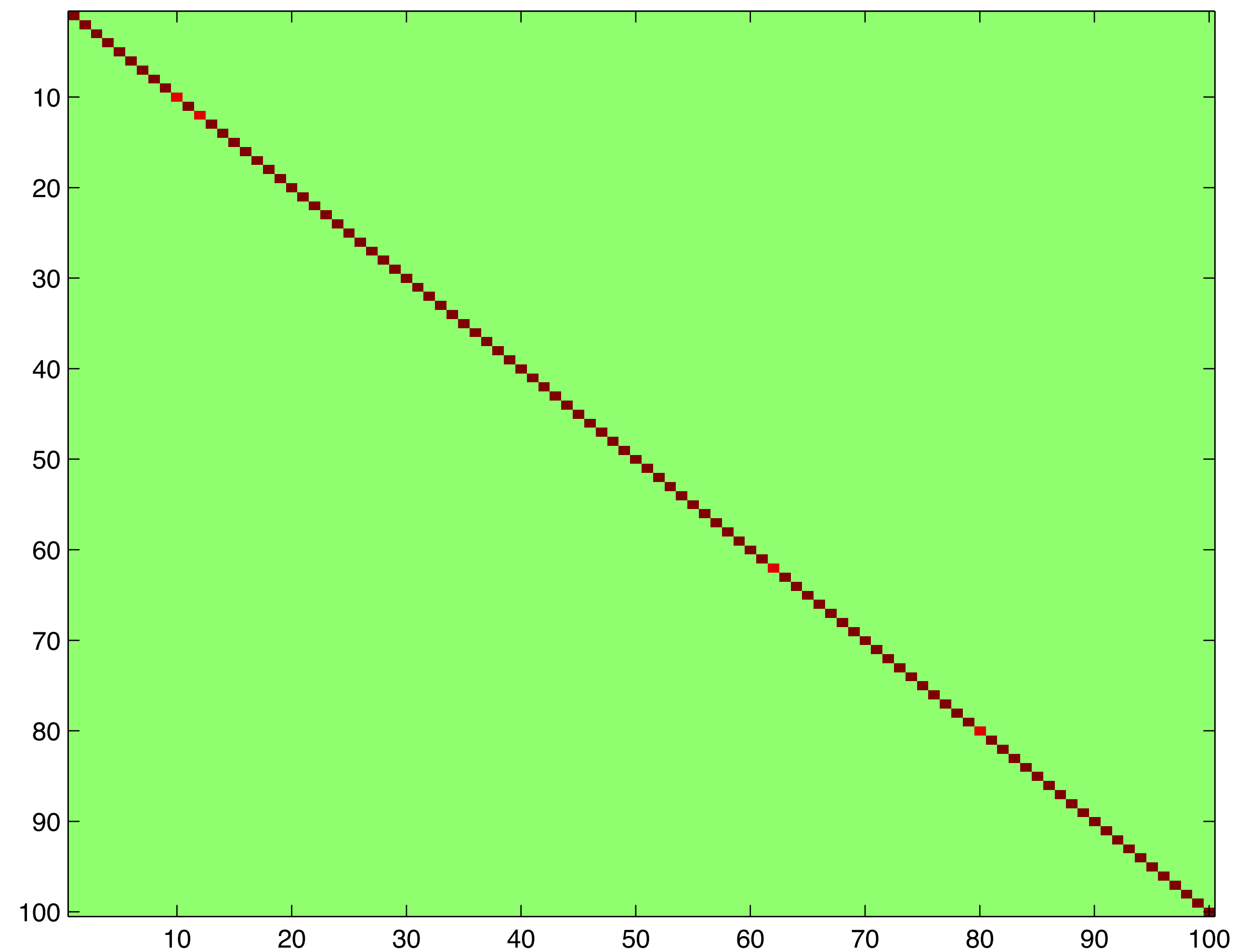




Source encoding

$$\frac{1}{K}\sum_{i=0}^{K-1}\mathbf{w}_i\mathbf{w}_i^{\mathrm{T}} \approx I \qquad K{=}1000$$

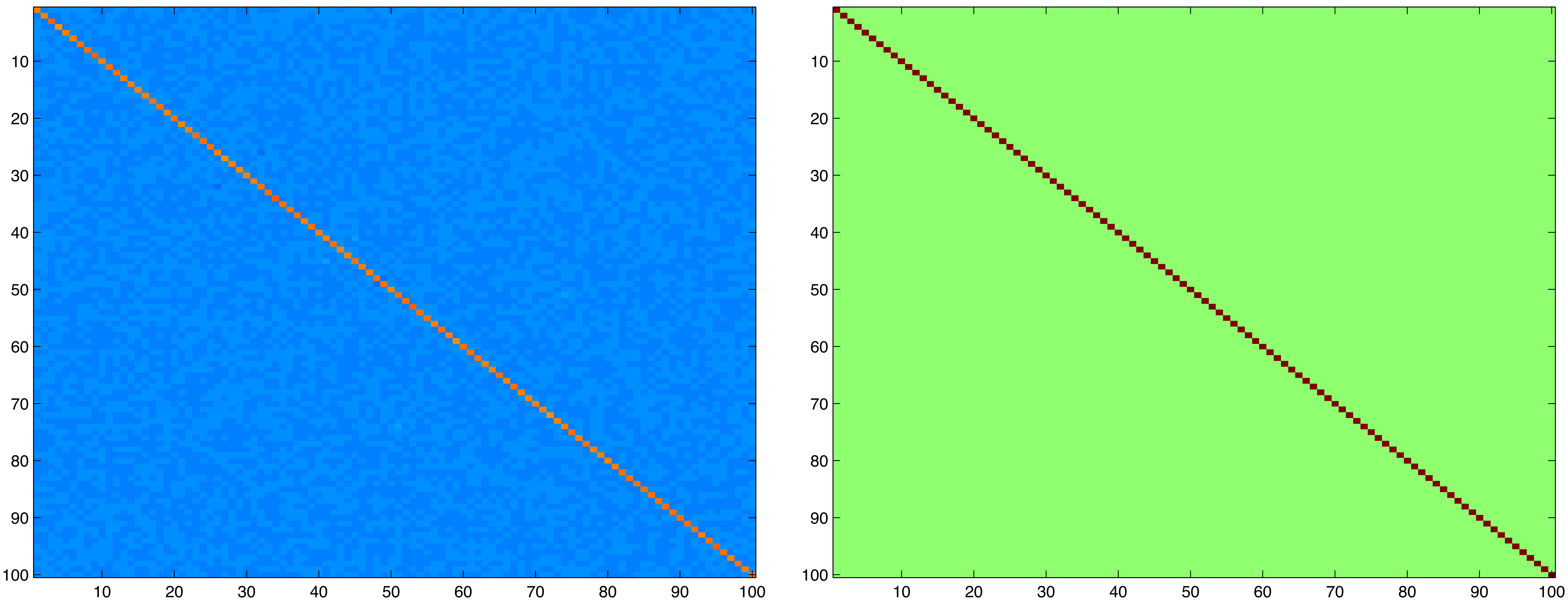




Source encoding

$$\frac{1}{K}\sum_{i=0}^{K-1}\mathbf{w}_i\mathbf{w}_i^{\mathrm{T}} \approx I$$

K=10000

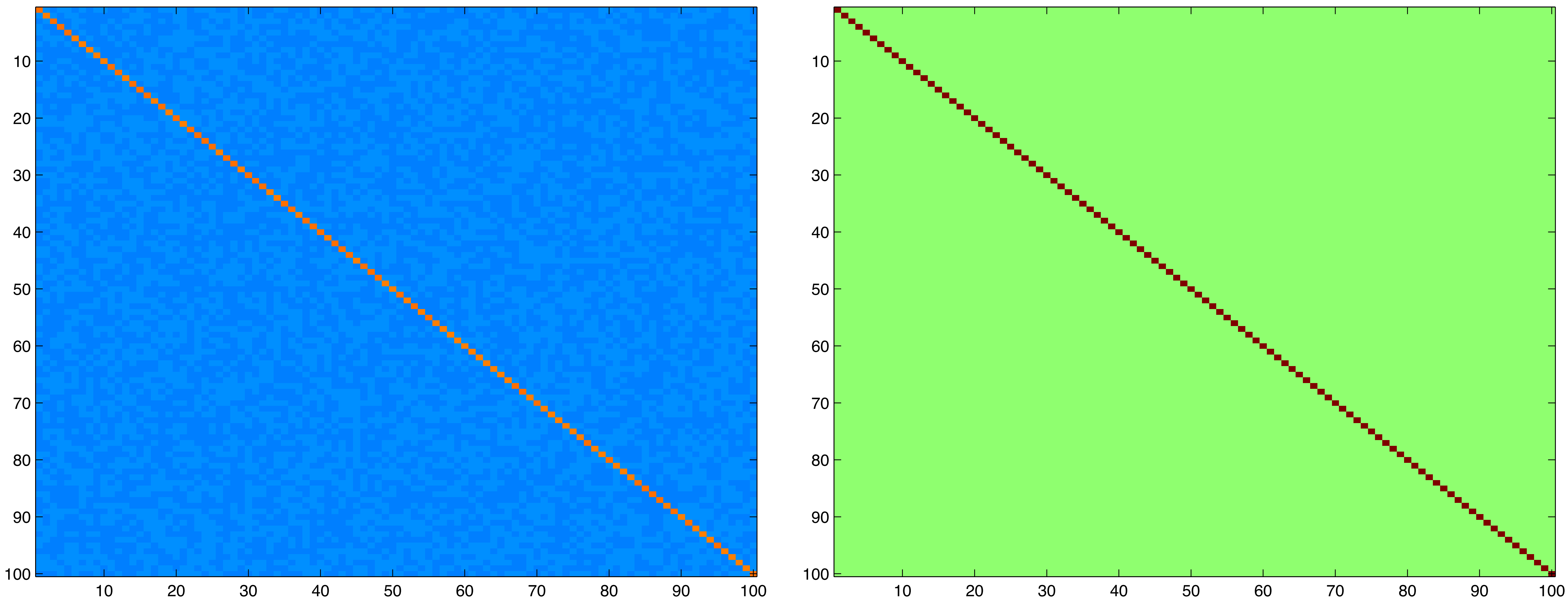


Tuesday, 6 December, 11

Source encoding

$$\frac{1}{K}\sum_{i=0}^{K-1}\mathbf{w}_i\mathbf{w}_i^{\mathrm{T}} \approx I$$

K=100000



Batching

Replace full misfit by sum over batch

$$\widetilde{\Phi}[\mathbf{m}] = \frac{1}{|B|}\sum_{i\in B}||\mathbf{d}_i - F[\mathbf{m}]\mathbf{q}_i||_2^2$$

Equivalent to choosing W to be *random* subset of the *identity* matrix

Batching

Write the error as

$$\mathbf{e} = \left(\frac{K - \widetilde{K}}{K\widetilde{K}}\right) \sum_{i \in B} \nabla \phi_i + \frac{1}{K} \sum_{i \notin B} \nabla \phi_i$$

The worst-case error:

$$||\mathbf{e}||_2^2 \leq 4 \left(\frac{K - \widetilde{K}}{K}\right)^2 \max_i ||\nabla \phi_i||_2^2$$

SLIM

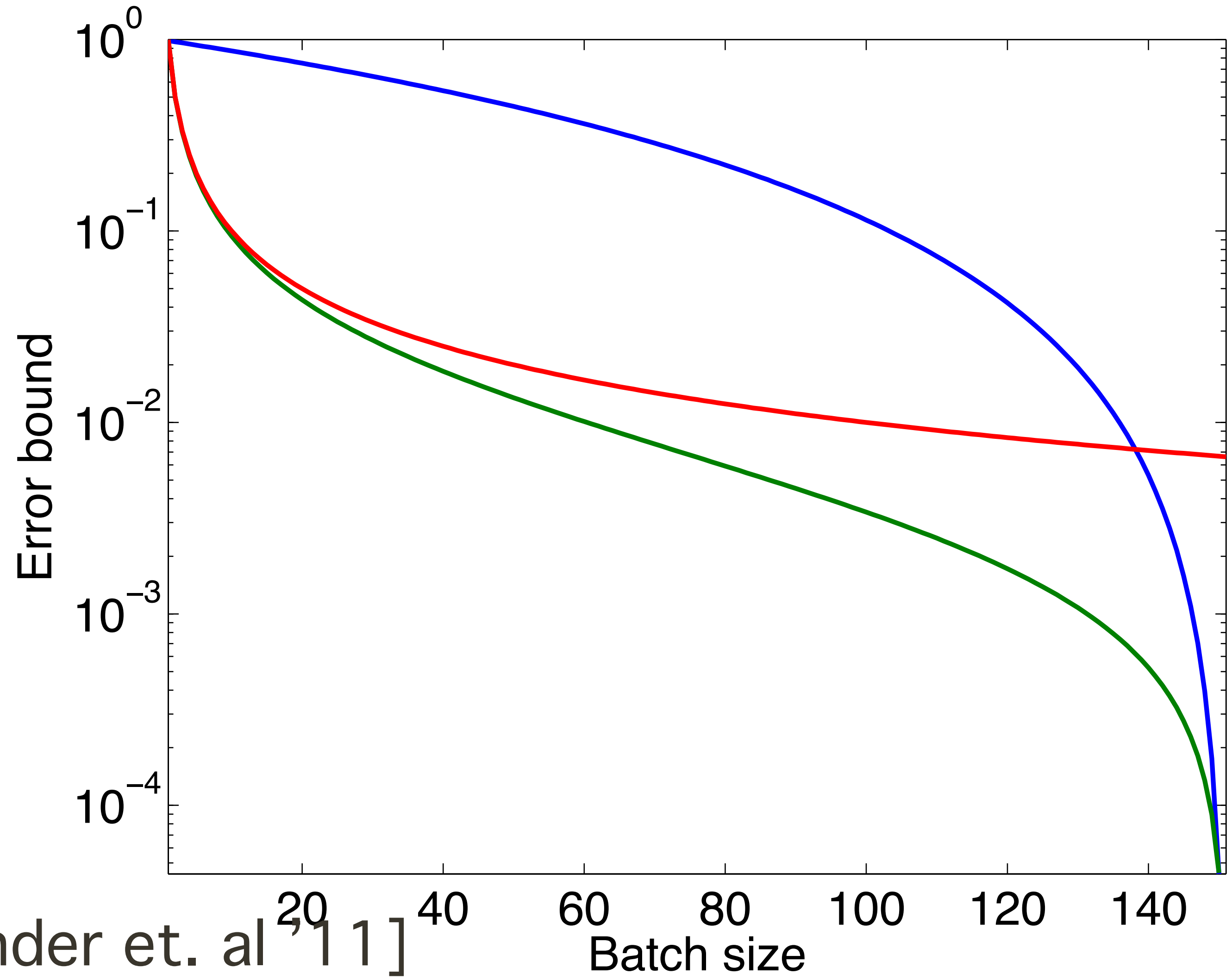
Batching

If we sample the sources randomly *without* replacement, we find:

$$\mathsf{E}(||\mathbf{e}||_2^2) \leq 2\left(\frac{K - \widetilde{K}}{\widetilde{K}(K-1)}\right) \max_i ||\nabla\phi_i - \nabla\Phi||_2^2$$

Batching strategy

Batching *strategy* controls theoretical *decay* of the *error*.



deterministic batching

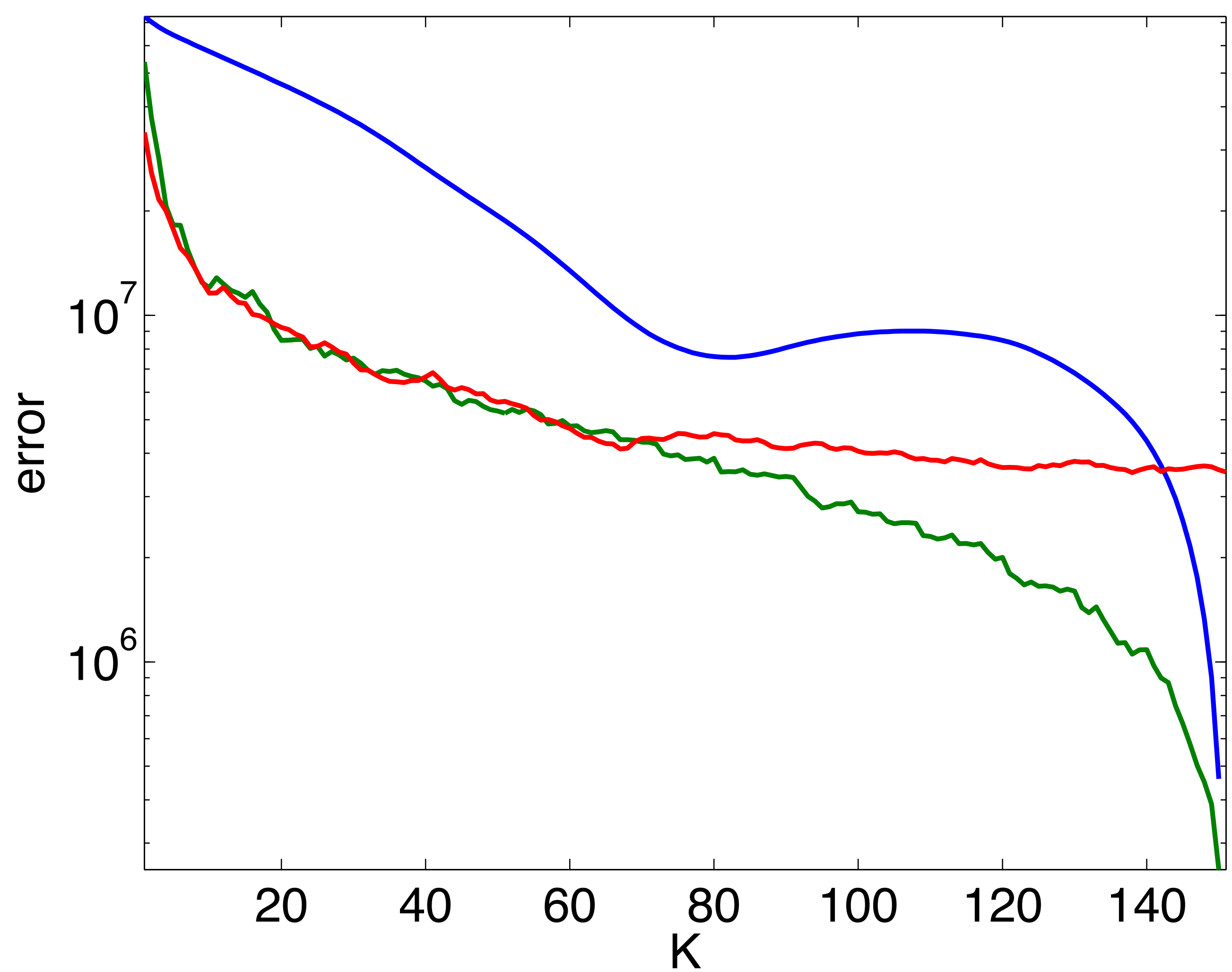
random batching

source encoding

[Friedlander et. al '11]

Batching strategy

Actual decay of the *error* in the *gradient*



deterministic batching

random batching

source encoding

Optimization with noisy gradients

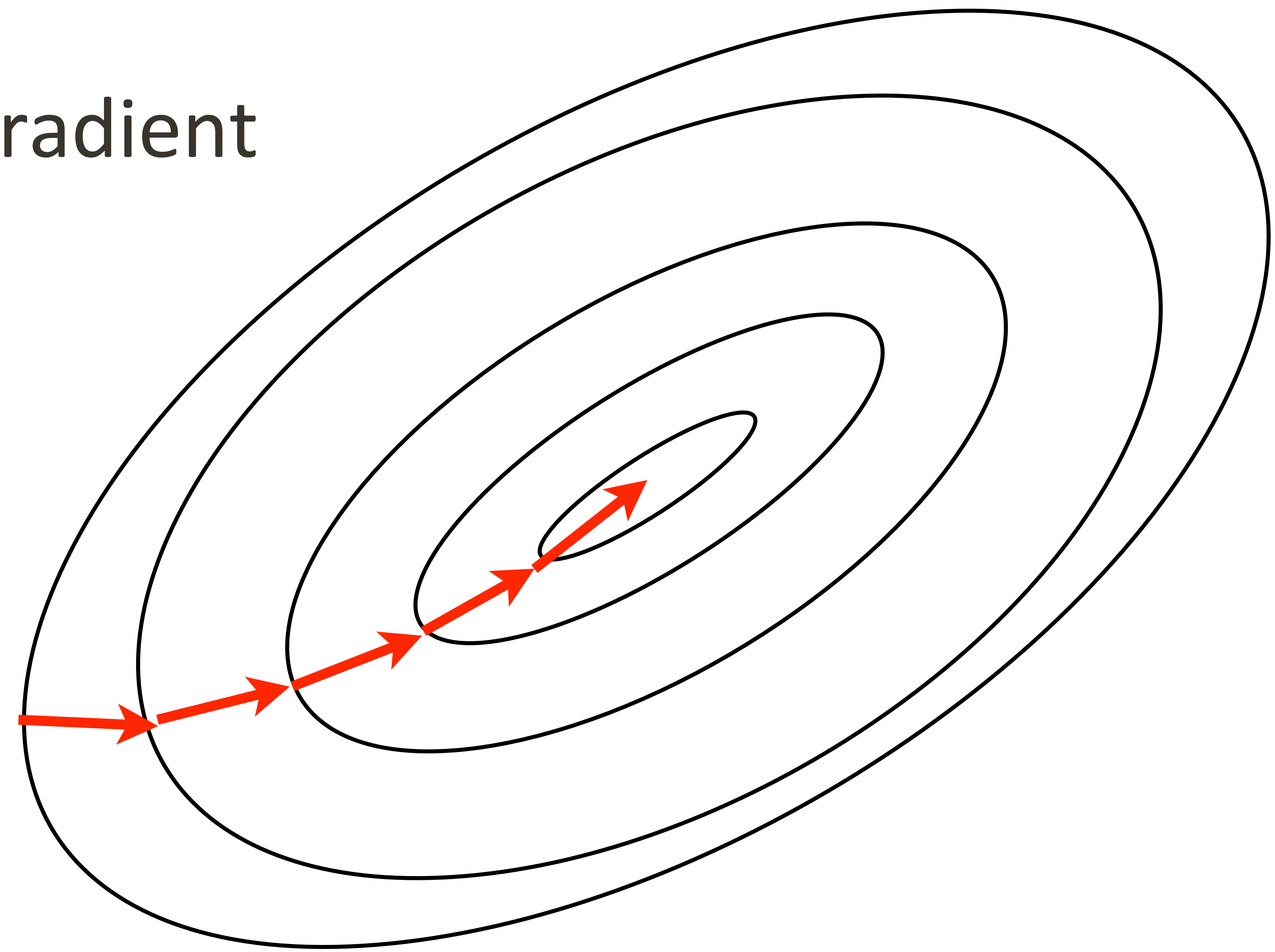
$$\mathbf{m}_{k+1} = \mathbf{m}_k + \gamma_k \mathbf{s}_k$$

$$\mathbf{s}_k \simeq \nabla\Phi[\mathbf{m}_k] + \mathbf{e}_k$$

Use either source-encoding or batching to control error

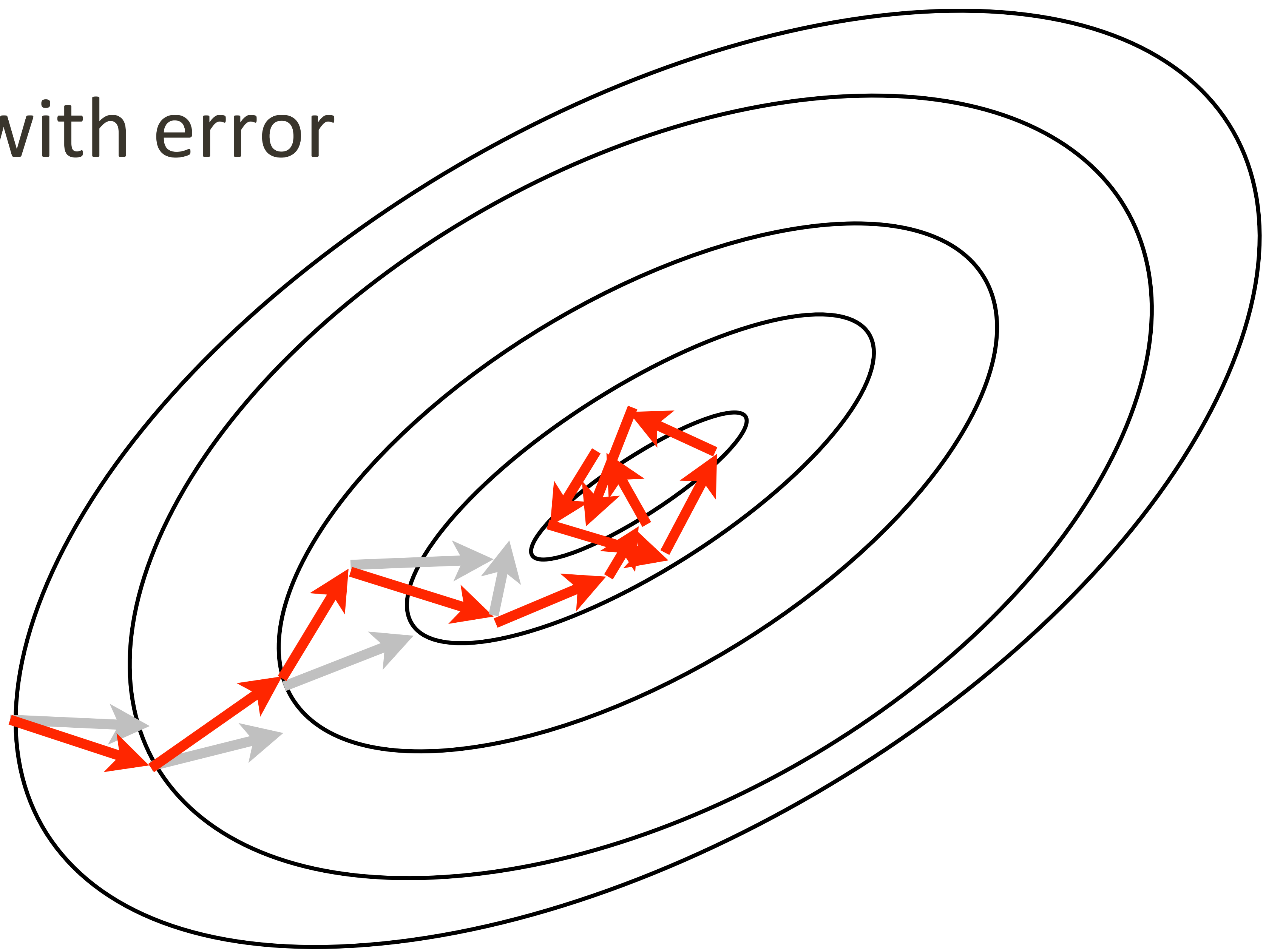
Optimization with noisy gradients

exact gradient



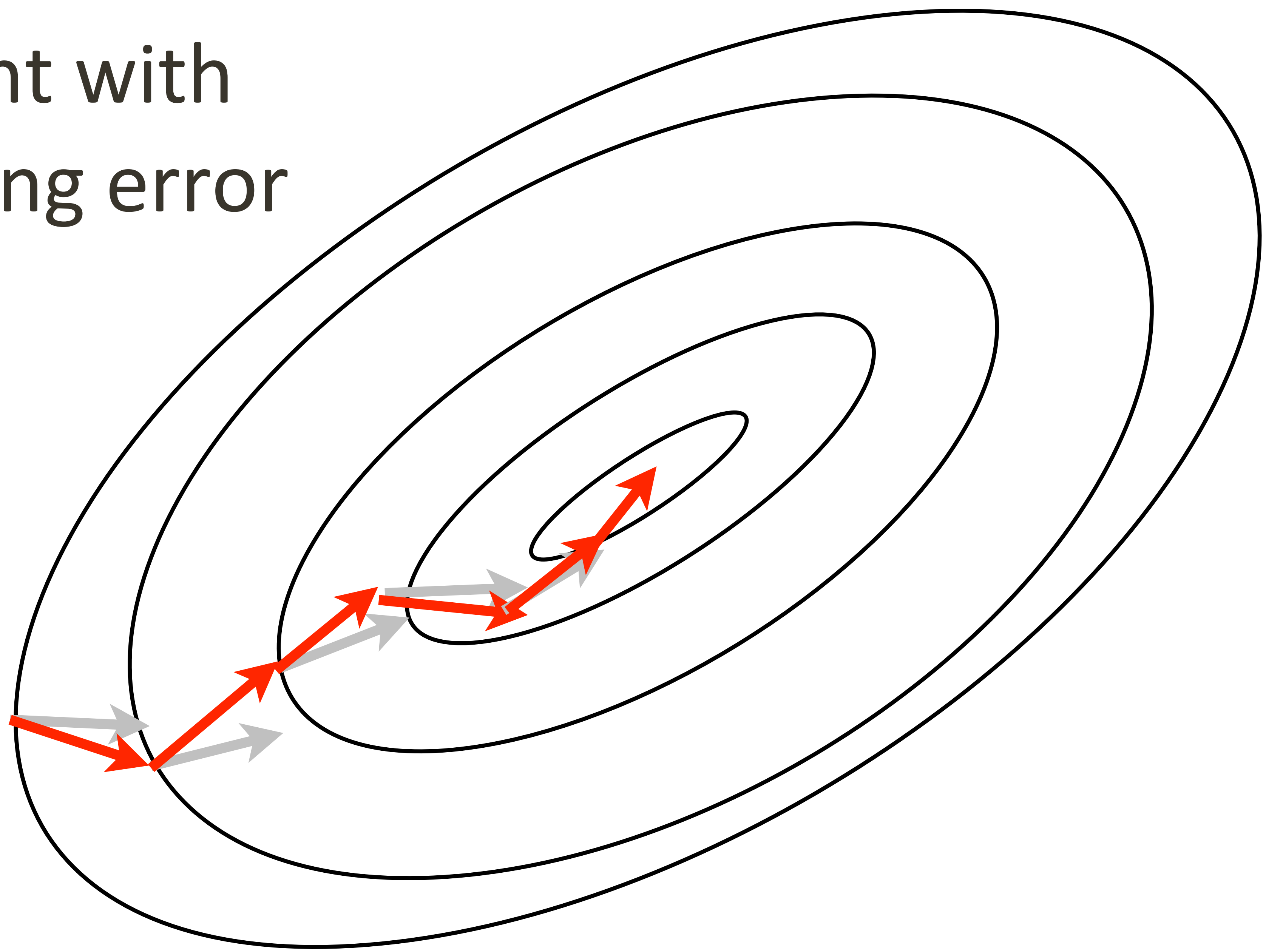
Optimization with noisy gradients

gradient with error



Optimization with noisy gradients

gradient with decreasing error



Deterministic optimization

- update: $\mathbf{s}_k = -H_k^{-1}\nabla\Phi[\mathbf{m}_k]$
- linesearch for γ_k
- cost per iteration: $\mathcal{O}(K)$
- convergence rate:

$$|\Phi[\mathbf{m}_*] - \Phi[\mathbf{m}_k]| = \mathcal{O}(c^k), \quad 0 < c \leq 1$$

Stochastic optimization

- update: $\mathbf{s}_k = -\left(\nabla\Phi[\mathbf{m}_k] + \mathbf{e}_k\right)$
- assumption: $\mathsf{E}\left\{\mathbf{s}_k\right\} = -\nabla\Phi[\mathbf{m}_k]$
- predetermined sequence $\gamma_k \downarrow 0$
- cost per iteration: $\mathcal{O}(1)$
- convergence rate:

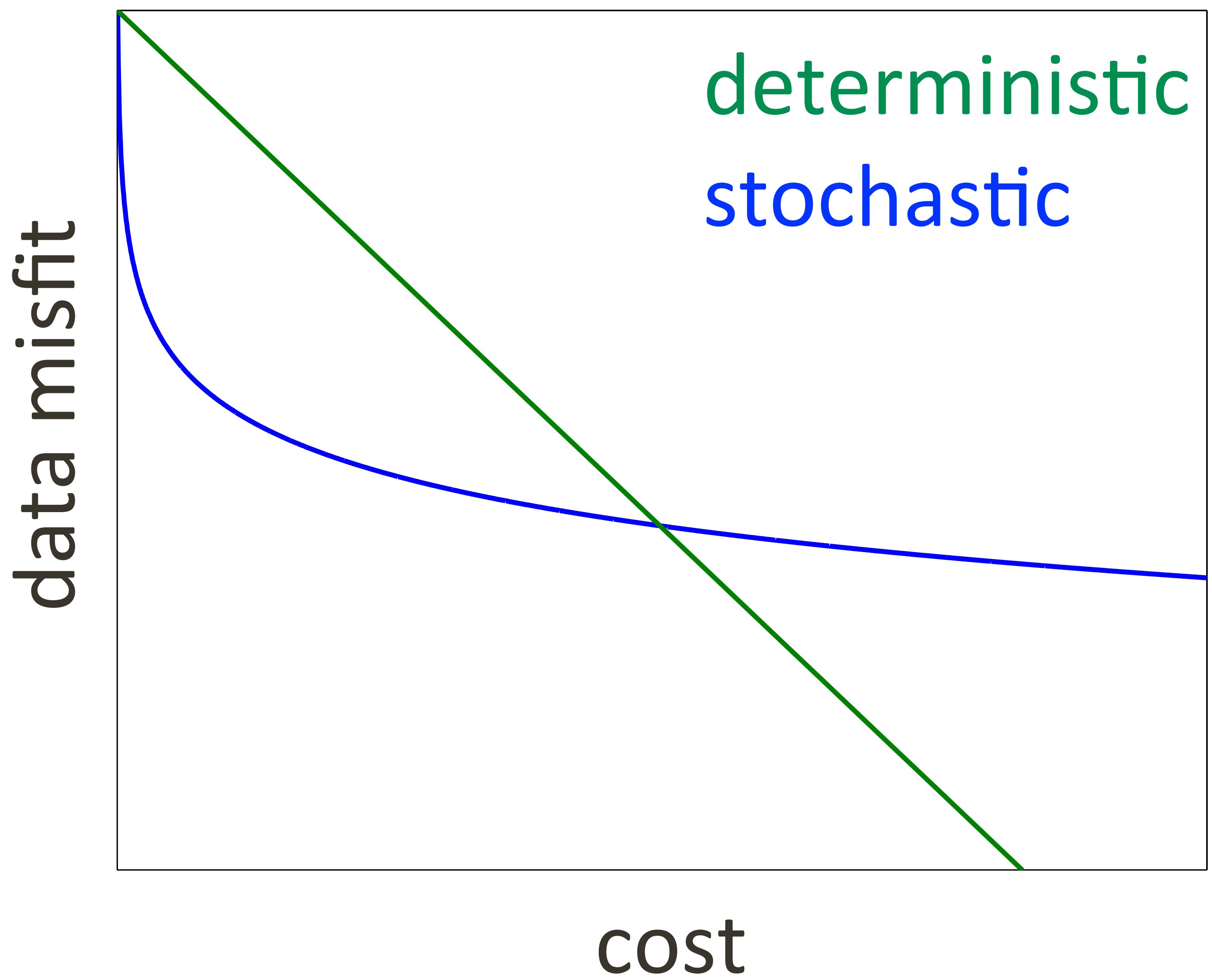
$$|\Phi[\mathbf{m}_*] - \Phi[\mathbf{m}_k]| = \mathcal{O}(1/k)$$

[Robbins et al. '50; Bertsekas et. al '96]

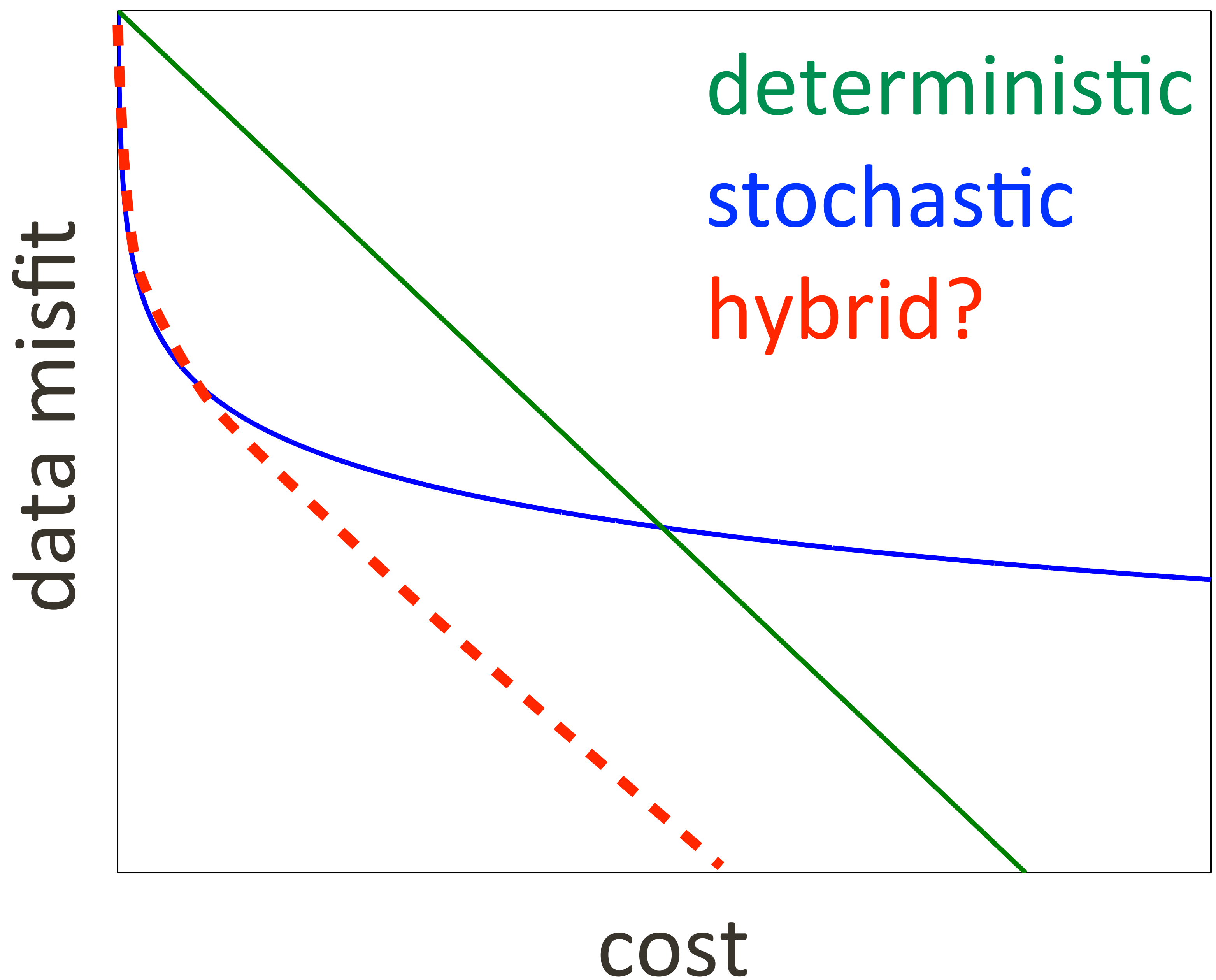
Stochastic optimization

- *cheap* iterations
- can be used with *any* encoding
- *no* theory for *Hessian* and *linesearch*
- *slow* convergence (relies on law of large numbers)

Stochastic vs. deterministic



Stochastic vs. deterministic



Hybrid method

- update: $\mathbf{s}_k = -H_k^{-1}\left(\nabla\Phi[\mathbf{m}_k] + \mathbf{e}_k\right)$
- assumption: $||\mathbf{e}_k||_2 \downarrow 0$
- cost per iteration: $\mathcal{O}(|B_k|)$
- convergence rate

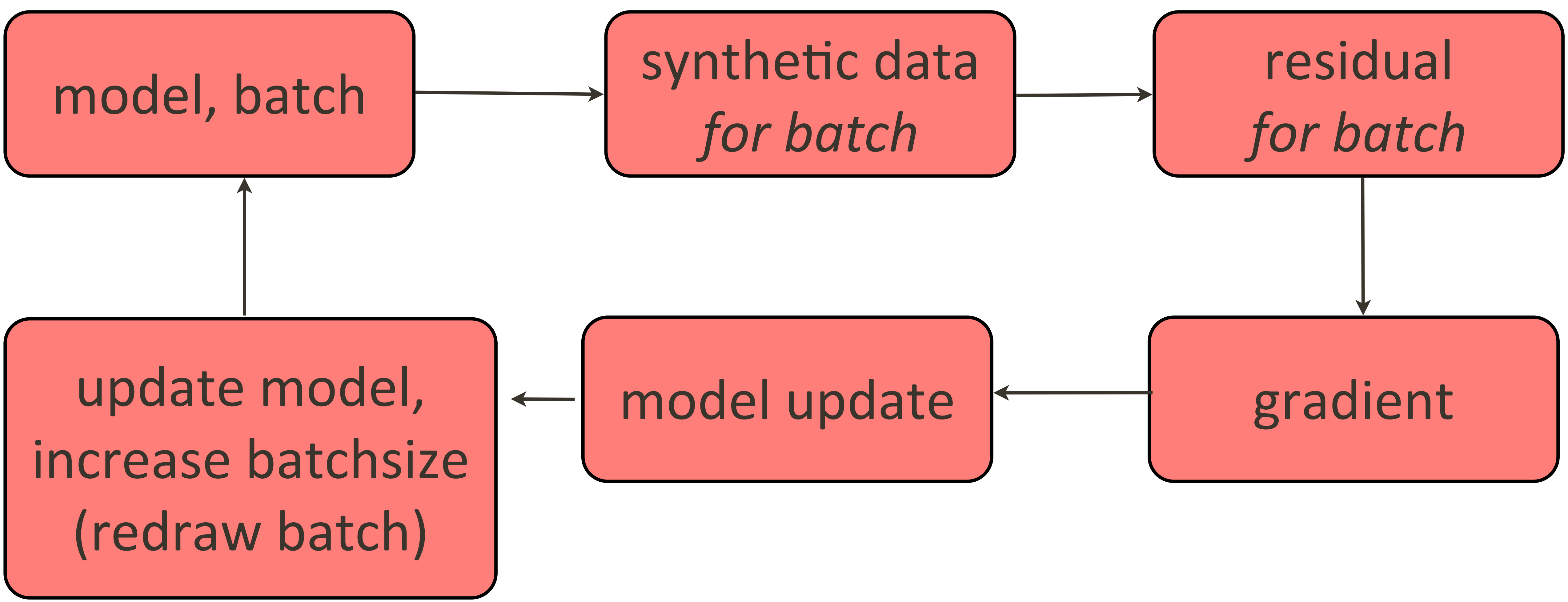
$$|\Phi[\mathbf{m}_*] - \Phi[\mathbf{m}_k]| = \mathcal{O}(c^k), \quad 0 < c < 1$$

[Friedlander et. al ’11]

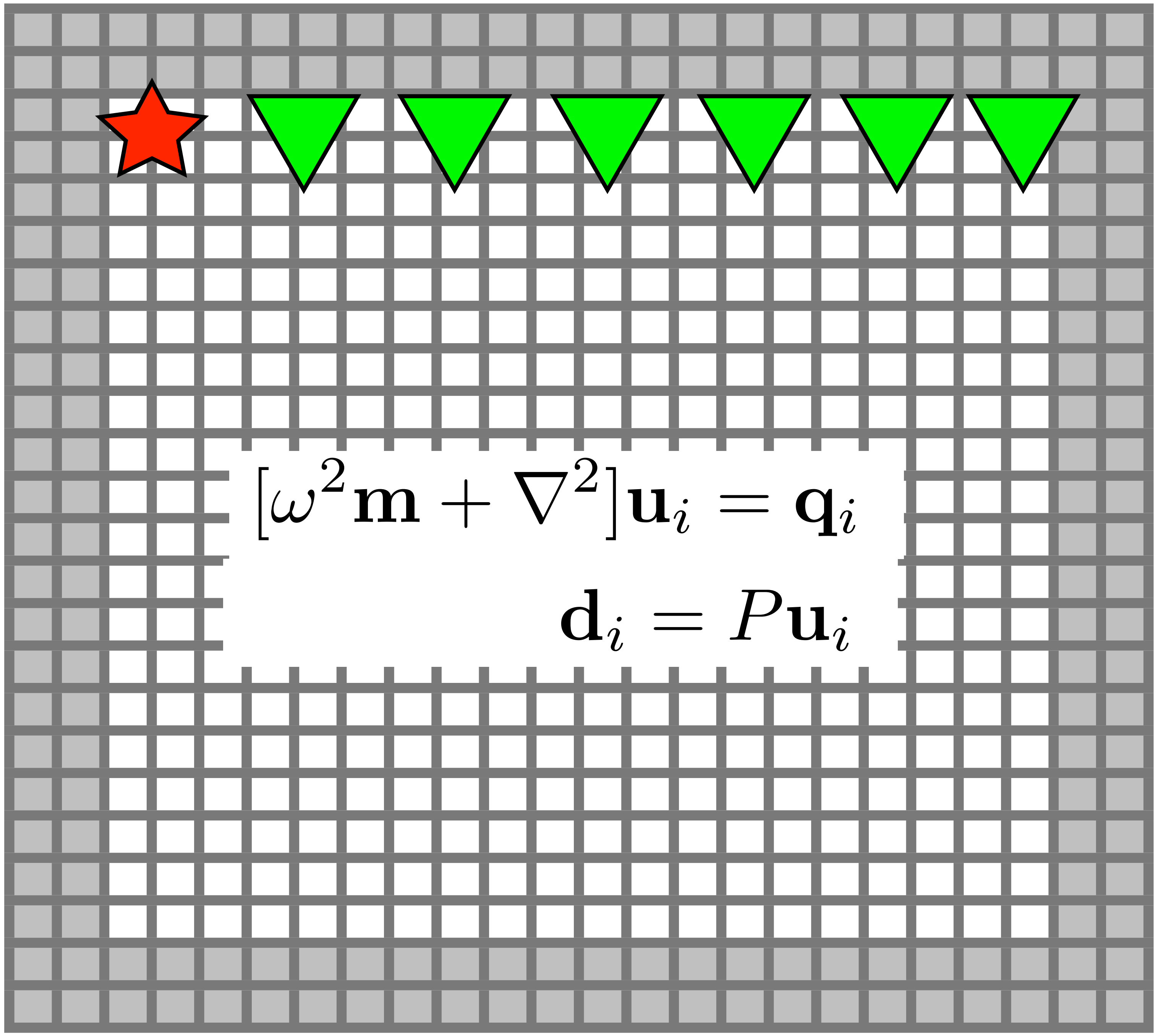
Hybrid method

- Batching allows us to bring down the error *fast* enough to ensure *linear* convergence
- Hessian and linesearch `allowed'

Hybrid method



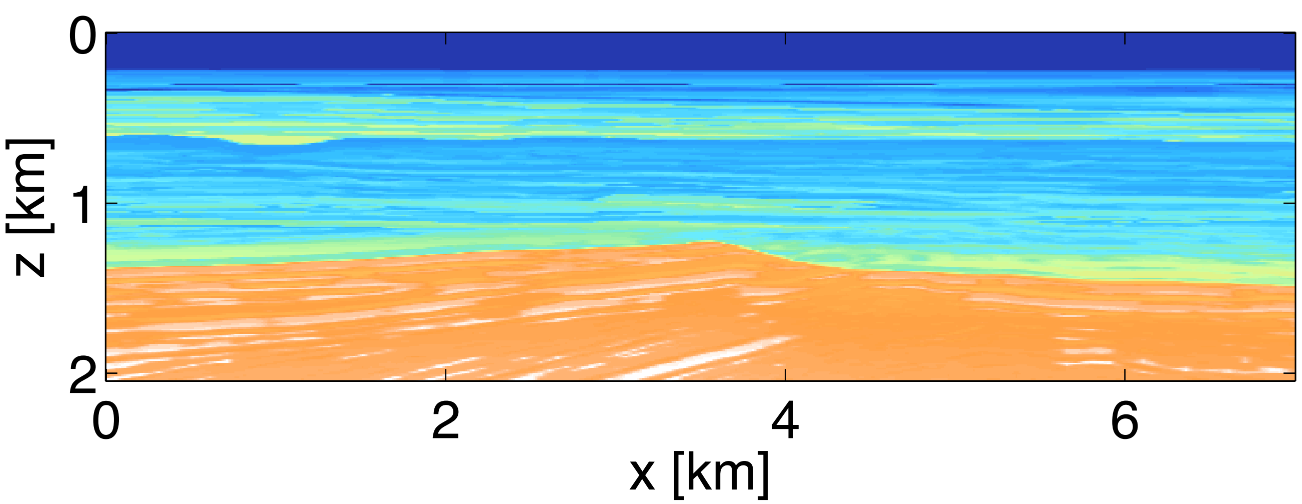
Results

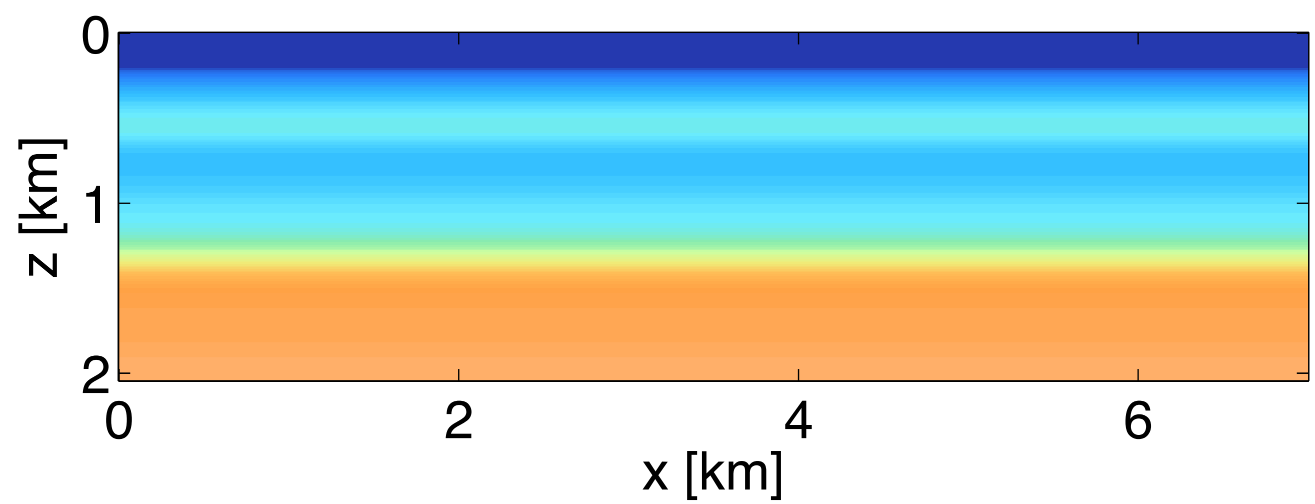


$$[\omega^2\mathbf{m} + \nabla^2]\mathbf{u}_i = \mathbf{q}_i$$

$$\mathbf{d}_i = P\mathbf{u}_i$$

Full waveform inversion



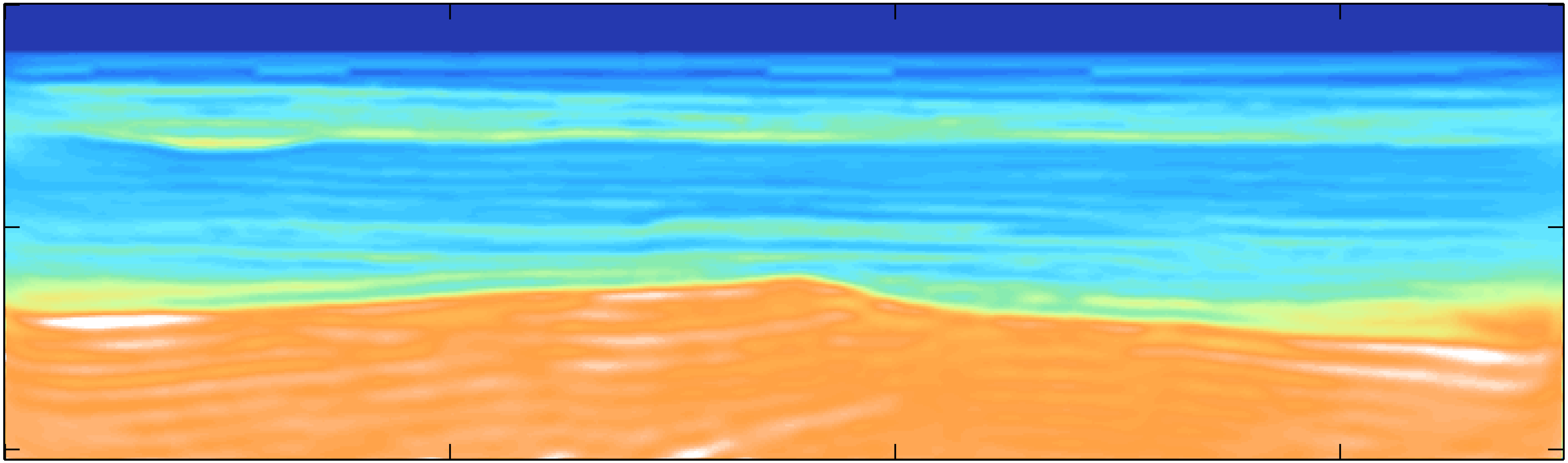


data for
141 sources, 281
receivers, 15 Hz Ricker

multi-scale frequency
domain inversion:
[2.5-20] Hz in 16 bands

[Bunks `95; Pratt `98]

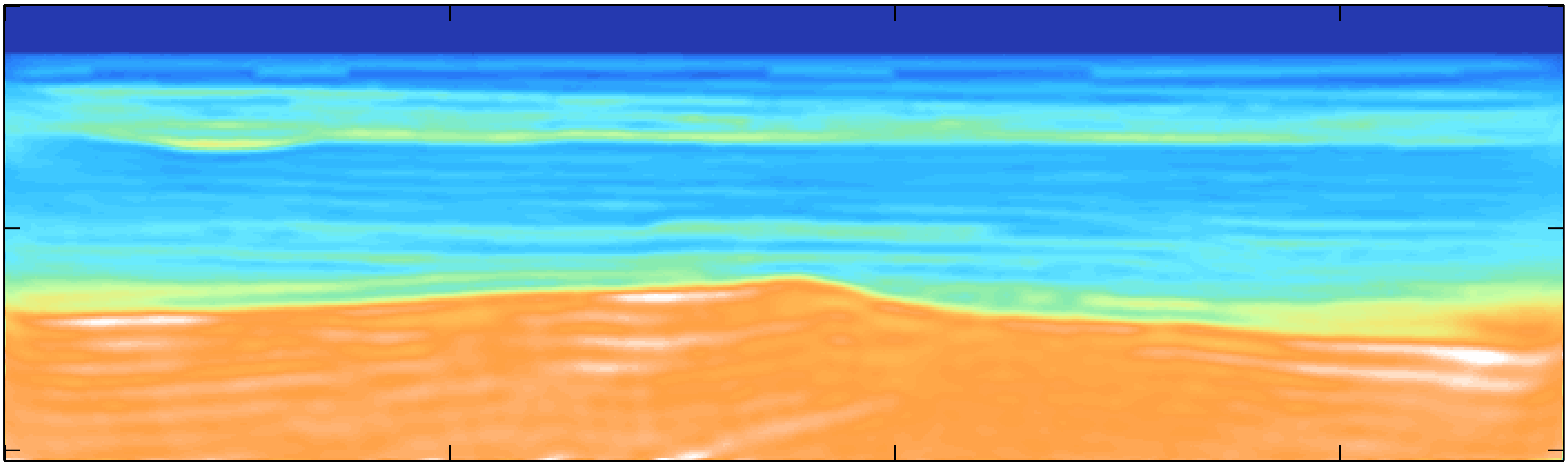
FWI



traditional L-BFGS

~15 full evaluations per frequency band

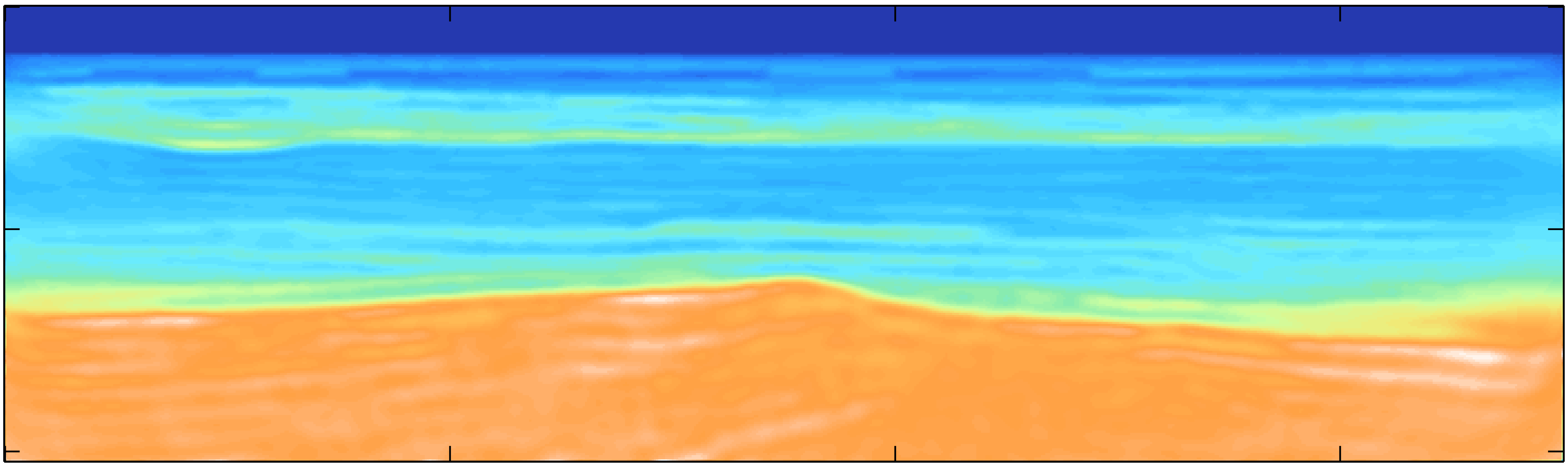
FWI



hybrid method

~1.5 full evaluations per frequency band

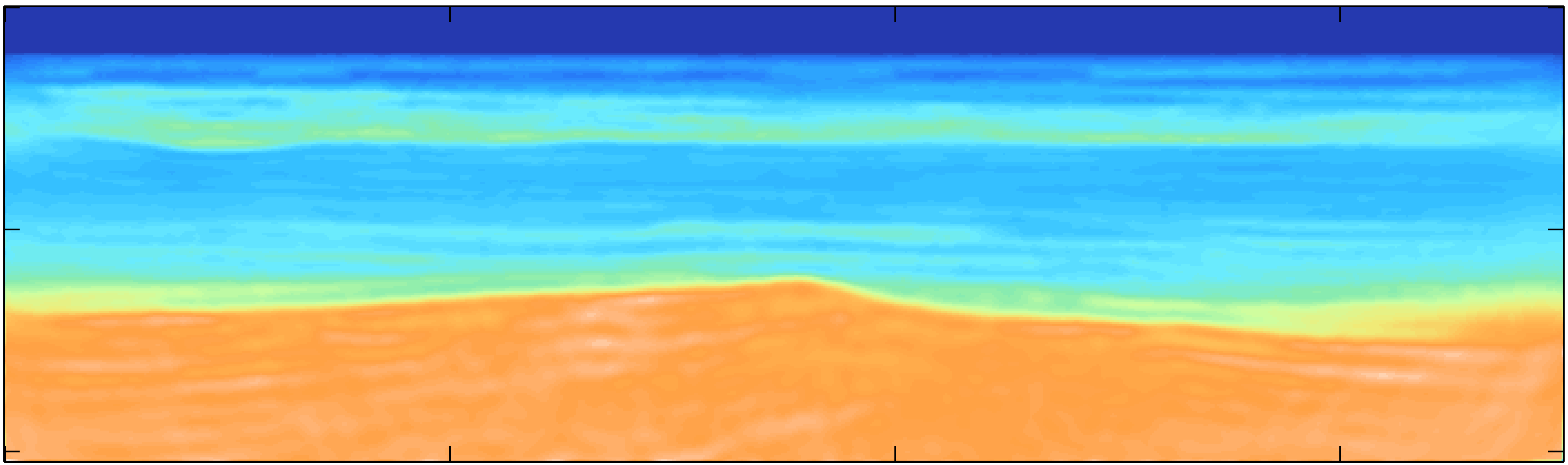
FWI



hybrid method

~.75 full evaluations per frequency band

FWI



hybrid method

~.5 full evaluations per frequency band

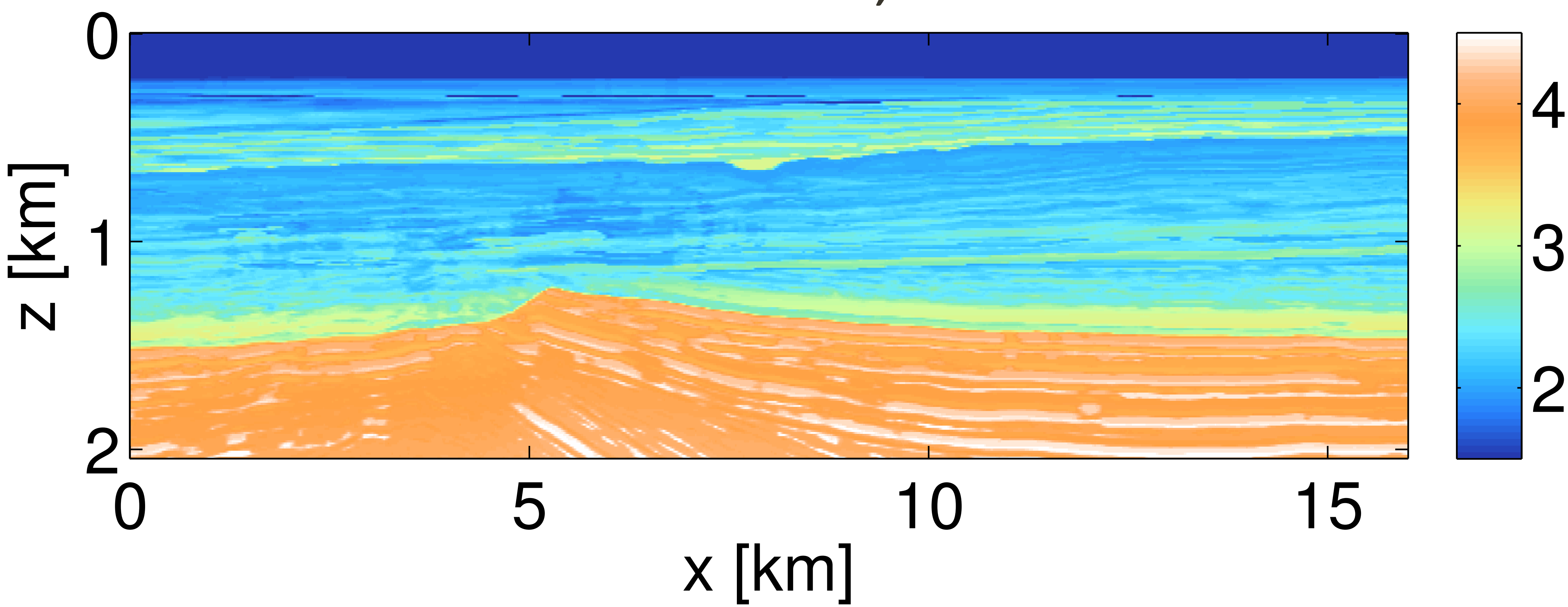
SLIM

FWI 2

time domain data

min offset 100m, max offset 3 km

320 sources at 50m, 15 Hz Ricker



FWI 2

Estimate source wavelet:

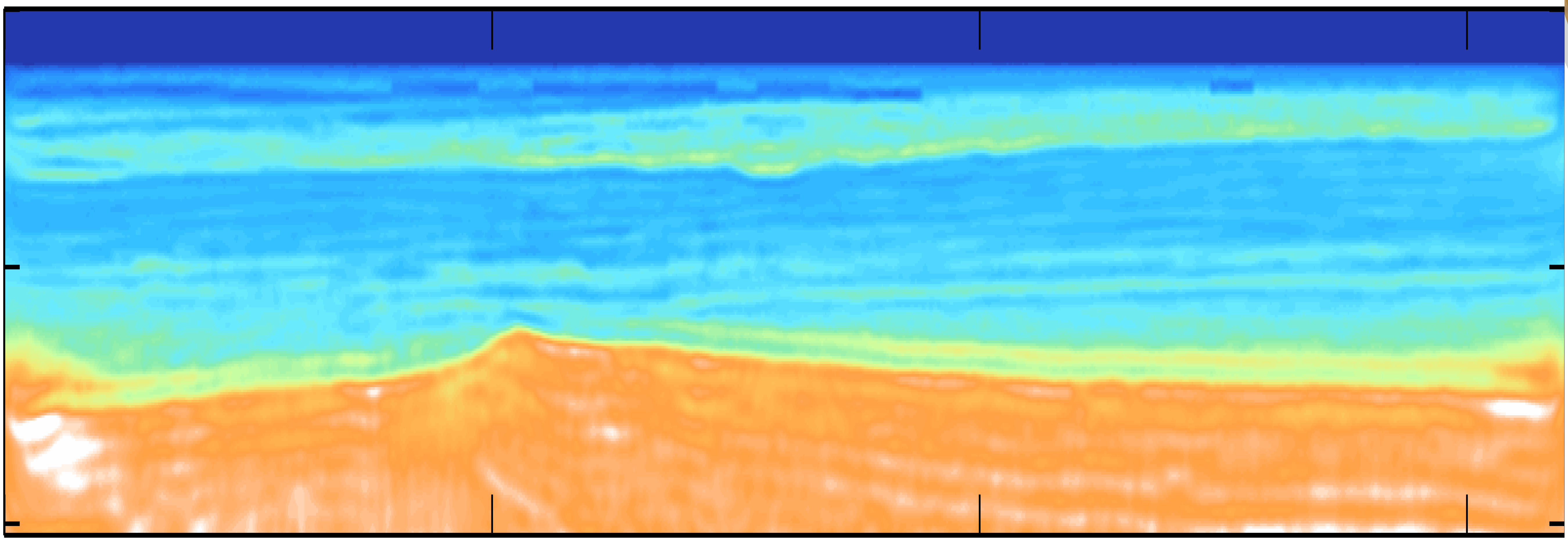
$$\Phi[\mathbf{m},\mathbf{a}] = \sum_i ||a_i F[\mathbf{m}]\mathbf{q}_i - \mathbf{d}_i||_2^2$$

LS solution for $\mathbf{a}$:

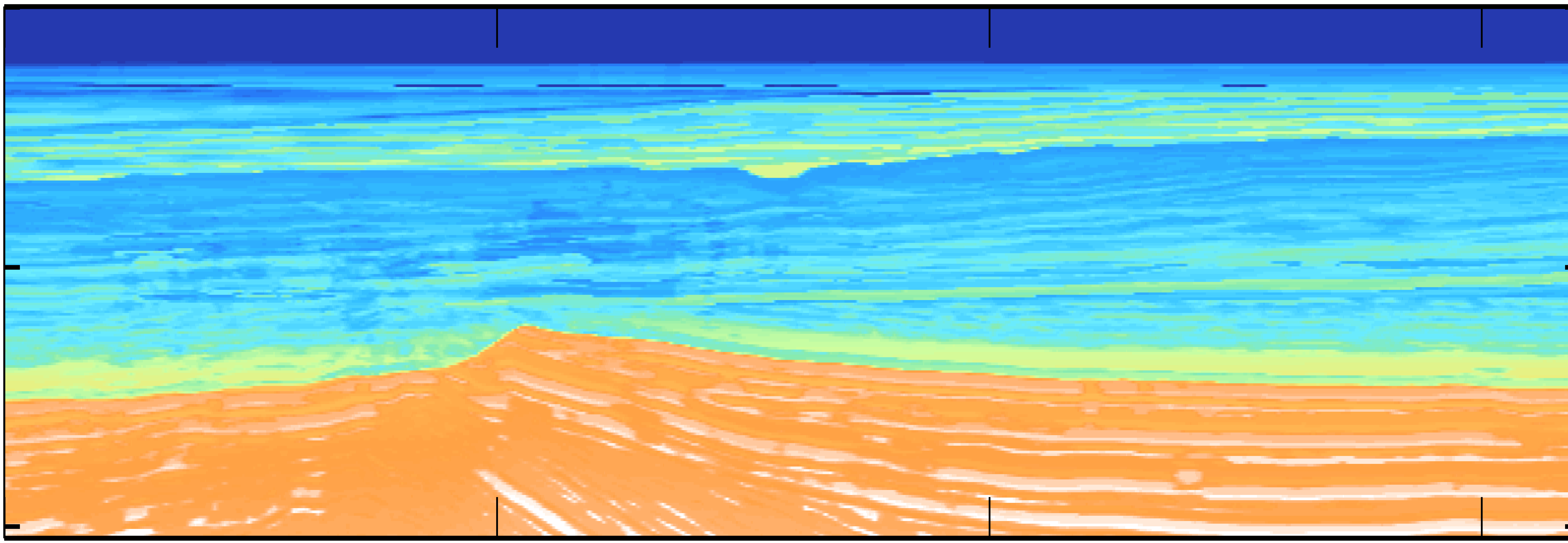
$$\hat{a}_i = \frac{(F[\mathbf{m}]\mathbf{q}_i)^H \mathbf{d}_i}{||\mathbf{d}_i||_2^2}$$

then:

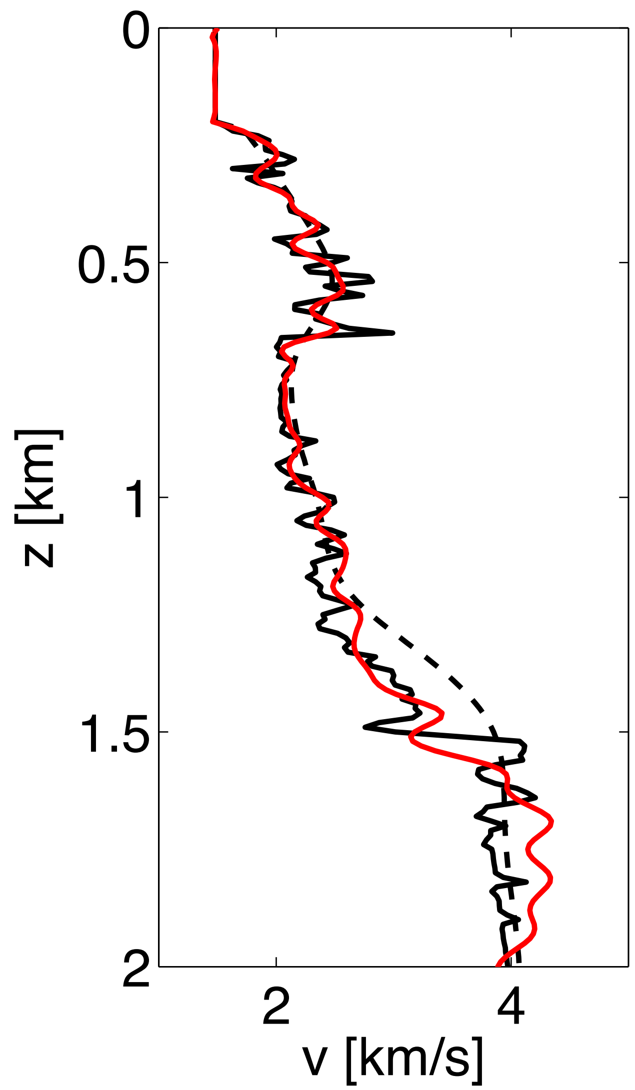
$$\nabla\Phi[\mathbf{m},\hat{\mathbf{a}}] = \sum_i \left(\frac{\partial \hat{a}_i F[\mathbf{m}]\mathbf{q}_i}{\partial \mathbf{m}}\right)^H (\hat{a}_i F[\mathbf{m}]\mathbf{q}_i - \mathbf{d}_i)$$



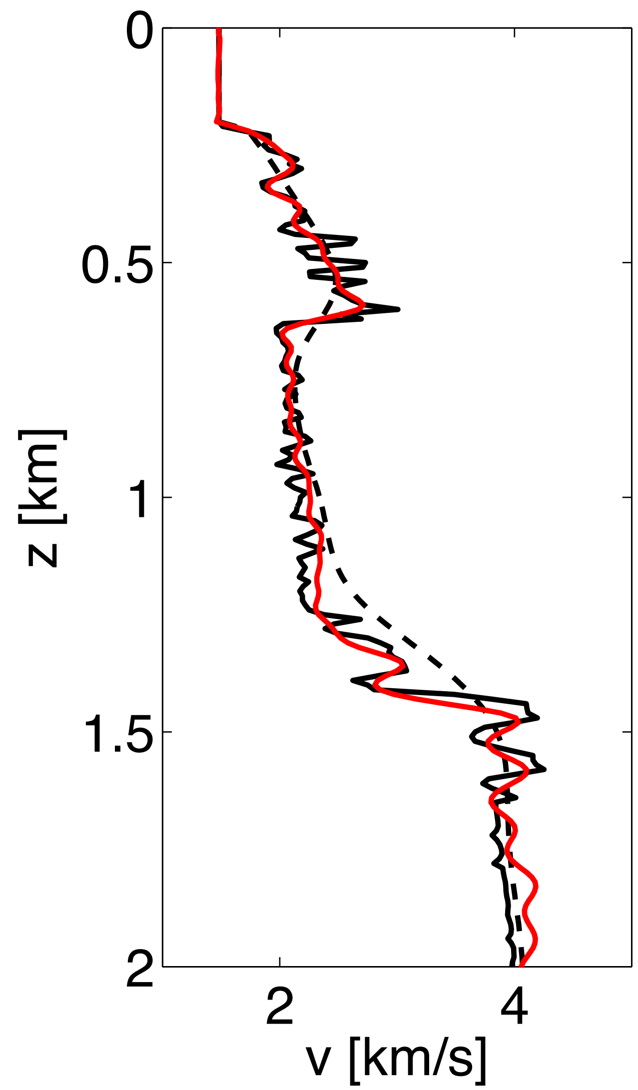
~2 iterations per freq. band



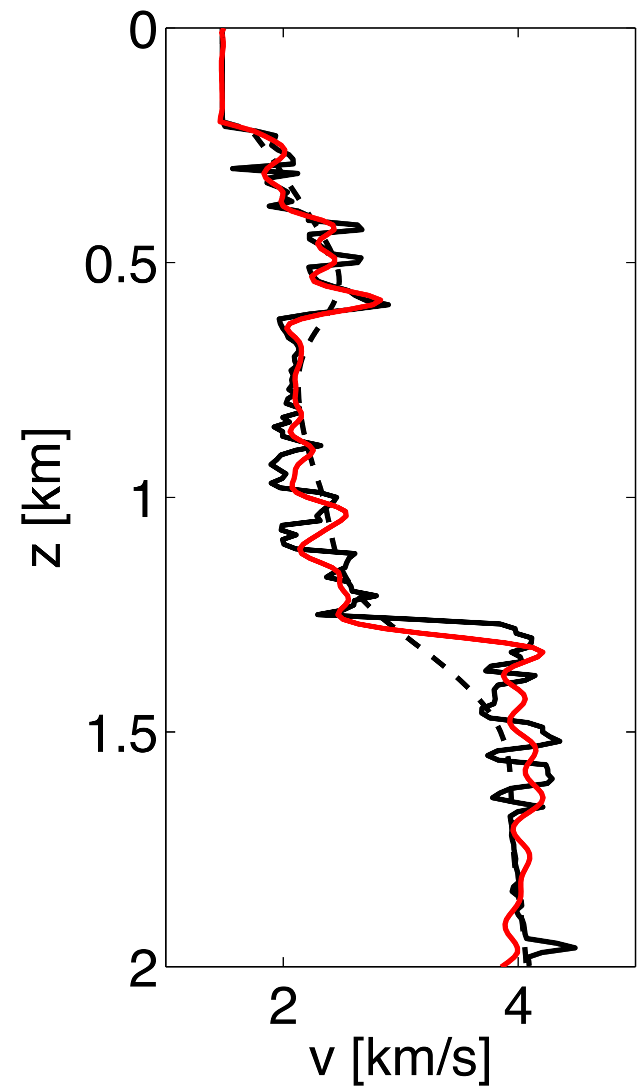
FWI 2



x=1000 m

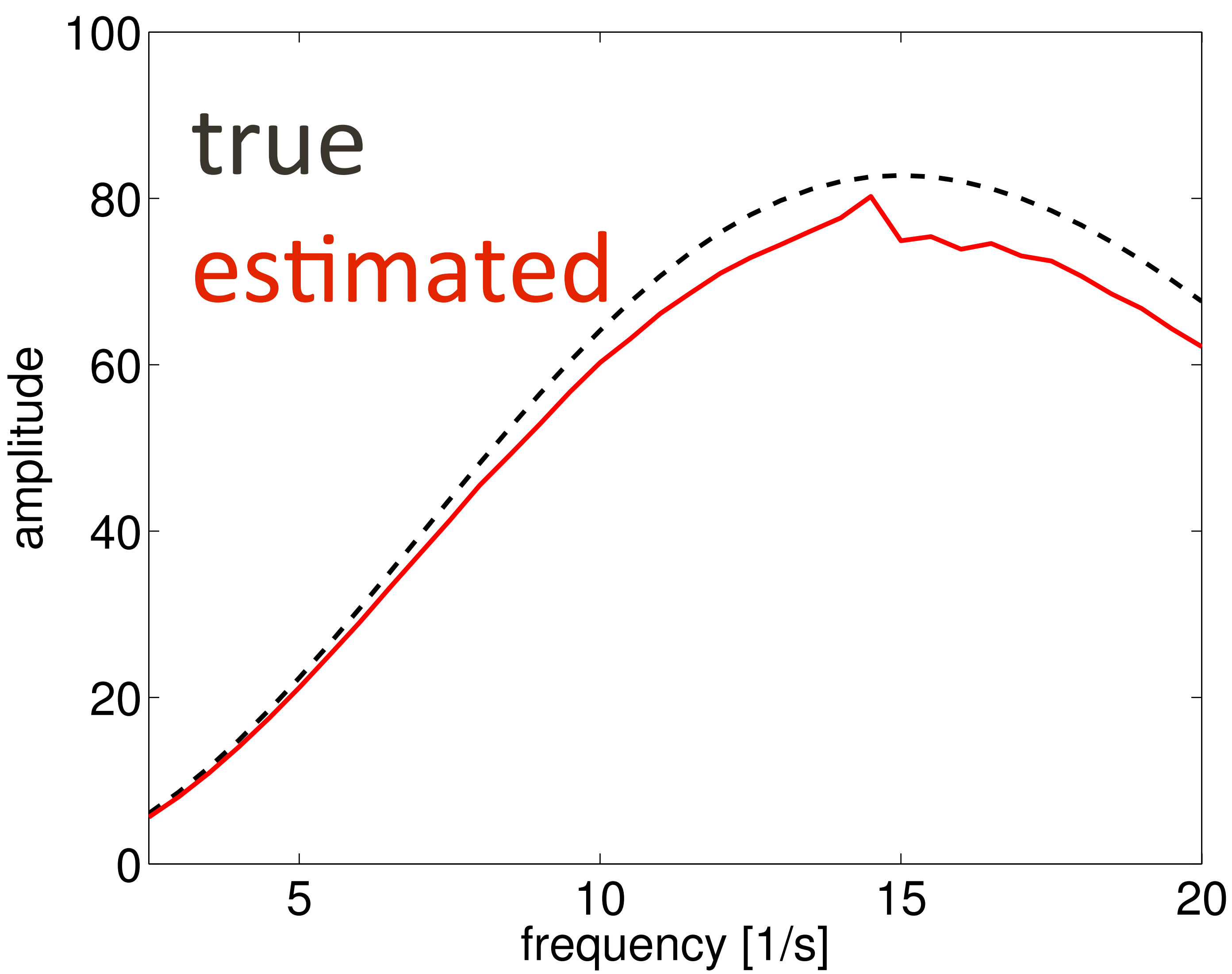


x=3500 m



x=6000 m

FWI 2


true
estimated
100
80
60
40
20
0
amplitude
5
10
15
20
frequency [1/s]

Conclusions

- `Source-encoding' can also be done with *unit* vectors: marine data
- Batching is a more *efficient* strategy to *approximate* the *full* misfit
- Batching can also be applied to plane-wave or eigenvector encoding

Conclusions

- Hybrid method gives both speed-up of *stochastic* method and *convergence* rate of deterministic method
- Applicable to *any* optimization problem of the form

$$\min_{\mathbf{m}} \Phi[\mathbf{m}] = \frac{1}{K}\sum_{i=0}^{K-1} \phi_i[\mathbf{m}]$$

(e.g., robust *FWI,* ray-based tomography)

SLIM

Acknowledgements

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). 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, BGP, PGS and WesternGeco.