Stochastic optimization and its application to seismic inversion

Zhilong Fang, and Tim Burgess

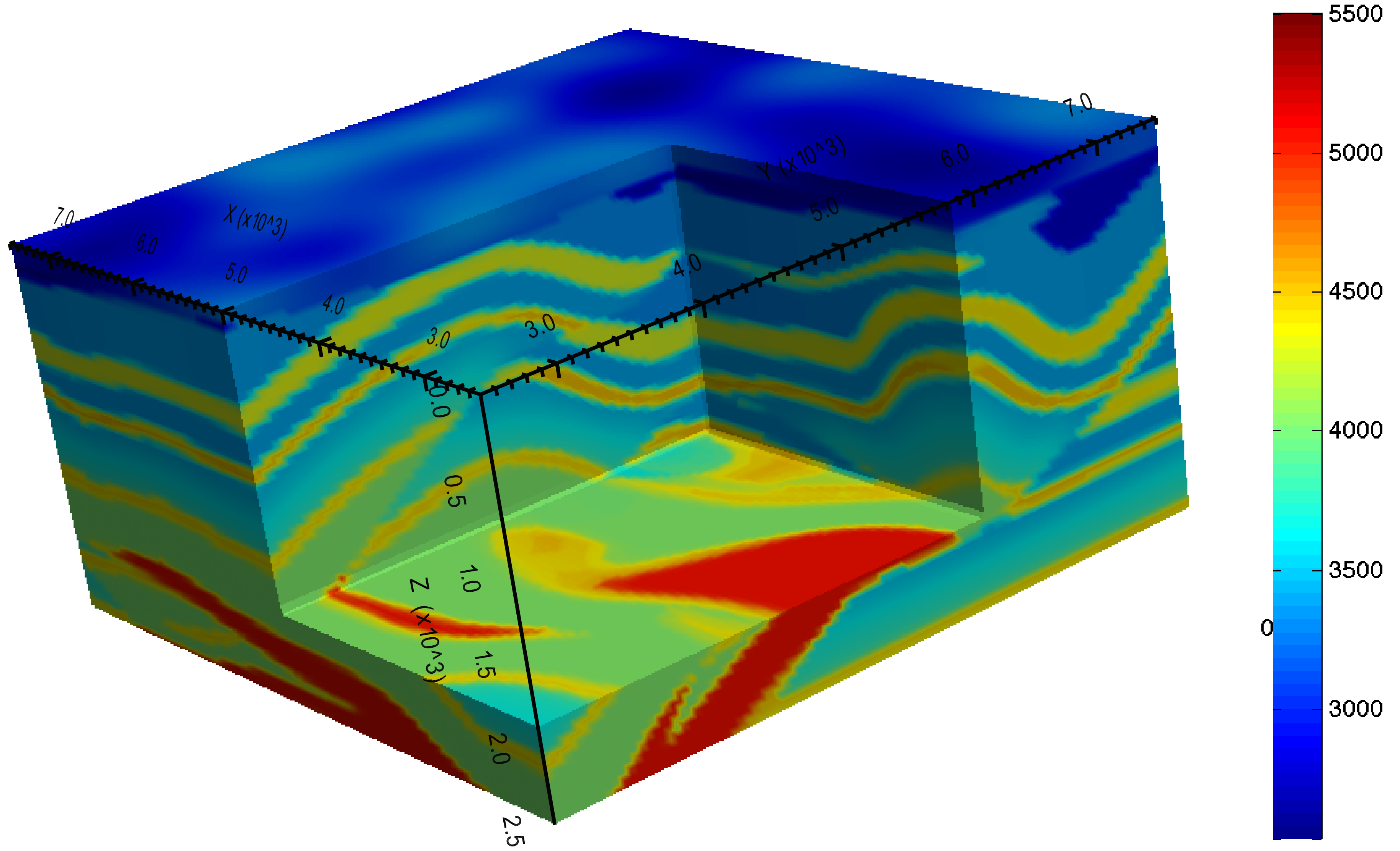
SLIM

University of British Columbia

Outline

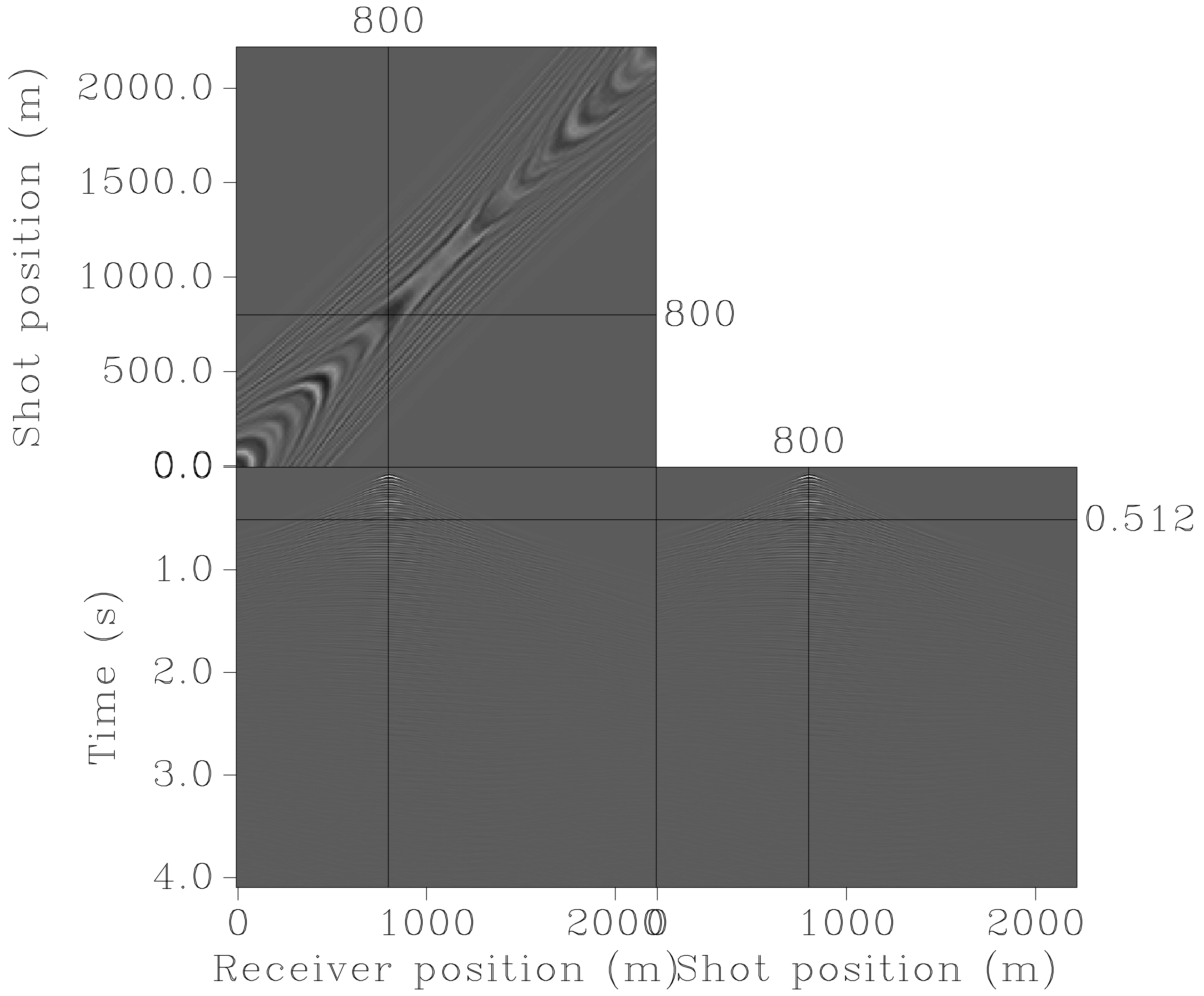
- Stochastic optimization
- Stochastic gradient method
- Stochastic average gradient method
- Stochastic gradient method with growing batch size
- Application on seismic inversion

Seismic inversion



Forward

Inversion



3D Model
Size: nx * ny * nz

5D Data
Size: nxsrc * nysrc * nxrec * nyrec * nt

Stochastic optimization

Data fitting problem:

$$\min_{\mathbf{m}} \varphi(\mathbf{m}) = \frac{1}{N}\sum_{i=1}^{N} f_i(\mathbf{m})$$

Full gradient (FG):

$$\min_{\mathbf{m}} G(\mathbf{m}) = \frac{1}{N}\sum_{i=1}^{N} g_i(\mathbf{m})$$

[Nesterov, 2004]

SLIM

Stochastic optimization

Full gradient method:

$$\mathbf{m}^{k+1} = \mathbf{m}^k - \alpha_k G(\mathbf{m}^k) = \mathbf{m}^k - \frac{\alpha_k}{N} \sum_{i=1}^{N} g_i(\mathbf{m}^k)$$

Linear convergence rate:

$$\varphi(\mathbf{m}^k) - \varphi(\mathbf{m}^*) = \mathcal{O}(\rho^k)$$

for some $\rho < 1$

[Nemirovski and Yudin, 1983]
[Nemirovski *et al.*, 2009]
[Agarwal *et al.*, 2012]

Stochastic optimization

Stochastic optimization

$$\min_{\mathbf{m}} \overline{\varphi}_k(\mathbf{m}) = \frac{1}{n_{\mathcal{I}_k}} \sum_{i \in \mathcal{I}_k} f_i(\mathbf{m})$$

Stochastic gradient (SG):

$$\overline{G}_k(\mathbf{m}) = \frac{1}{n_{\mathcal{I}_k}} \sum_{i \in \mathcal{I}_k} g_i(\mathbf{m})$$

[Nemirovski and Yudin, 1983]
[Nemirovski *et al.*, 2009]
[Agarwal *et al.*, 2012]

SLIM

Stochastic optimization

Stochastic gradient method:

$$\mathbf{m}^{k+1} = \mathbf{m}^k - \alpha_k \overline{G}_k(\mathbf{m}^k)$$

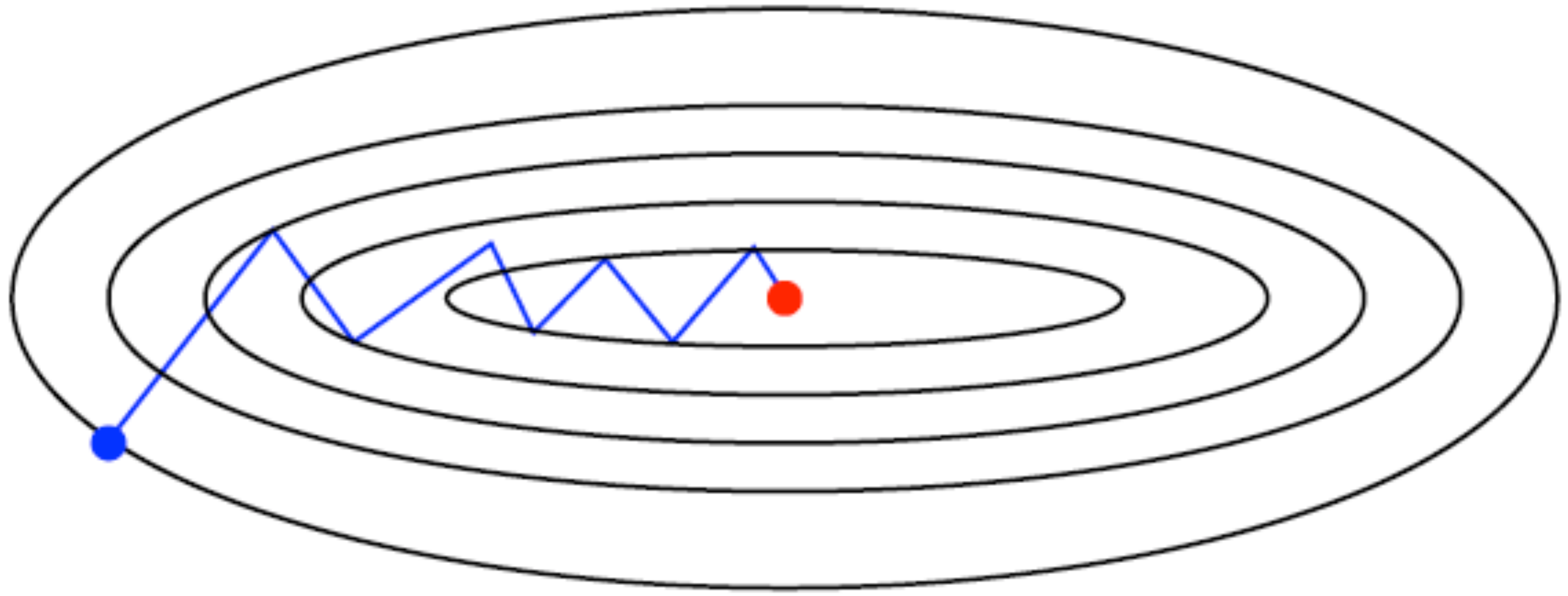
Sublinear convergence rate:

$$\mathbb{E}[\varphi(\mathbf{m}^k)] - \varphi(\mathbf{m}^*) = \mathcal{O}(1/k)$$

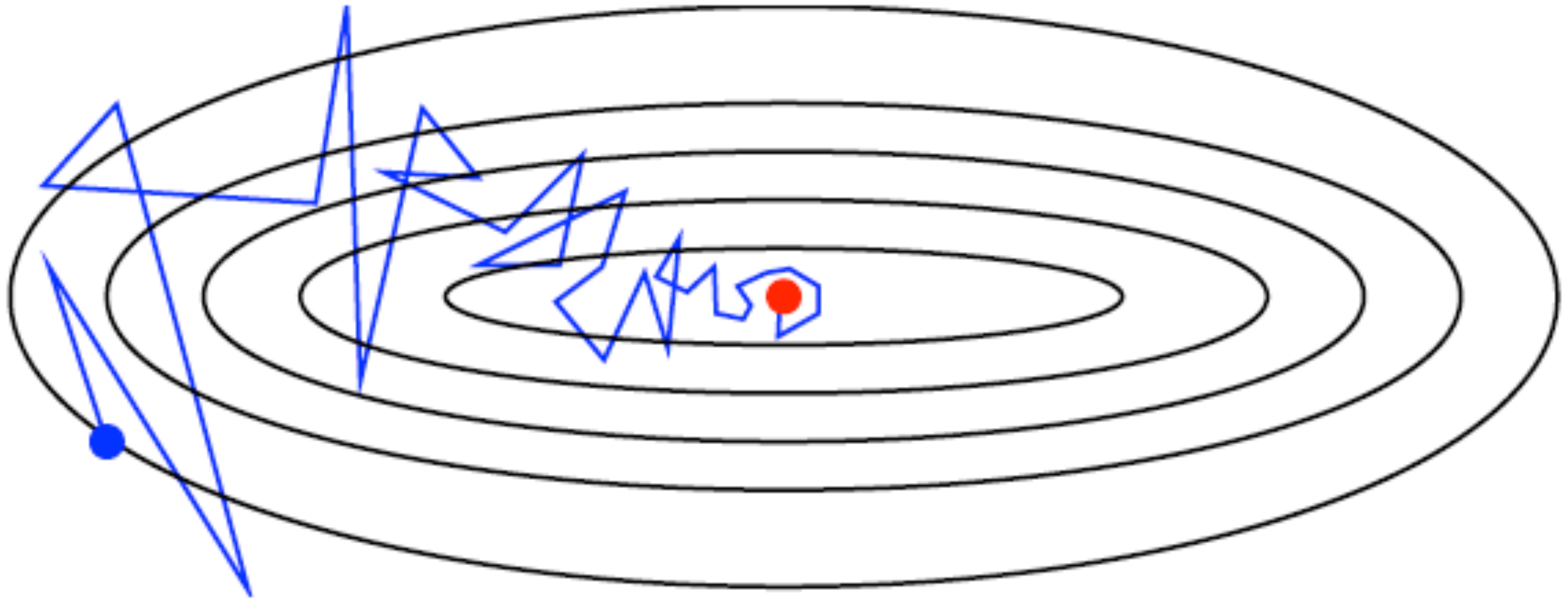
[Schmidt 2014]

FG vs SG

Full gradient method:



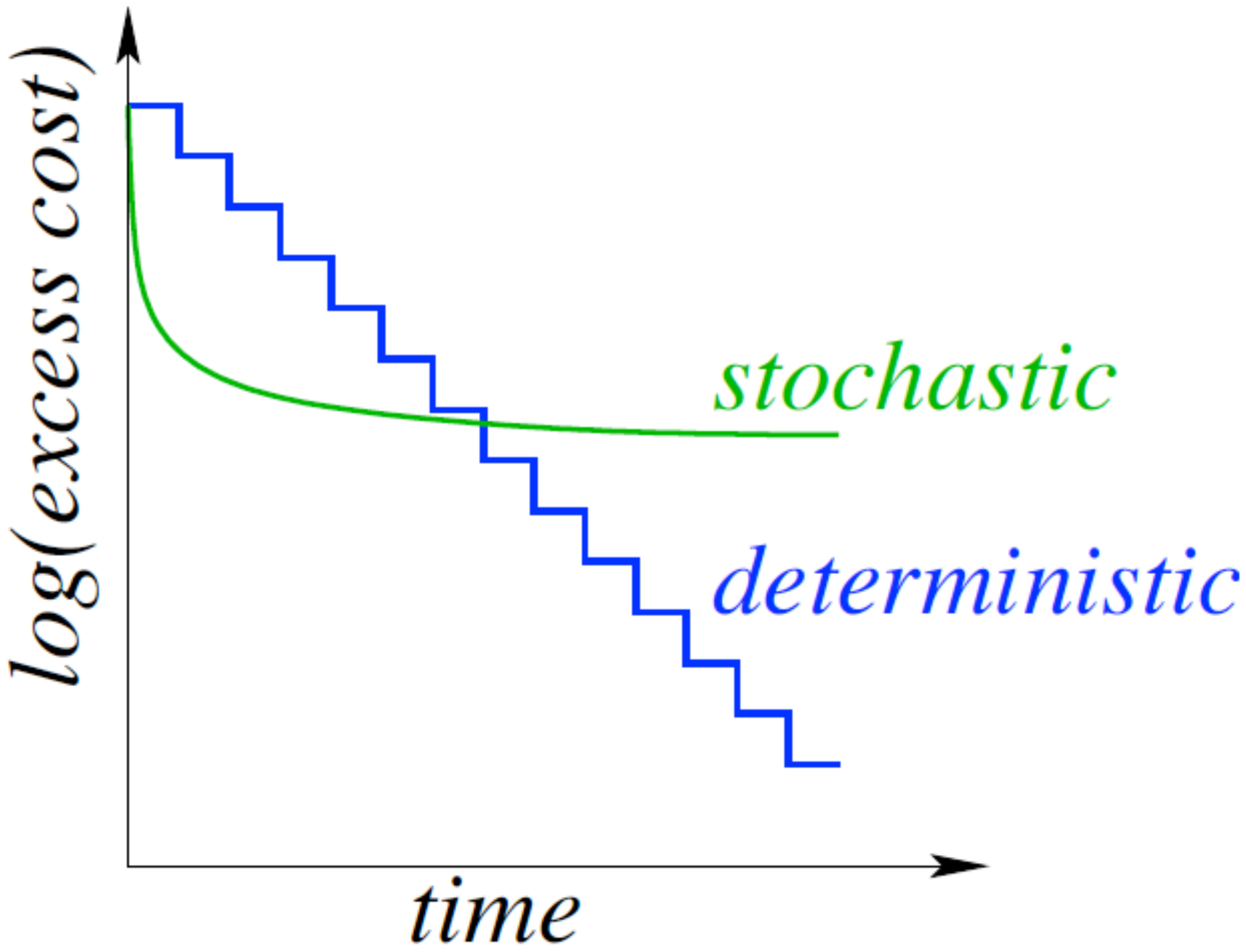
Stochastic gradient method:



FG vs SG

Convergence comparison:

- FG method has O(N) cost with $\mathcal{O}(\rho^k)$ rate;
- SG method has O(1) cost with O(1/t) rate;



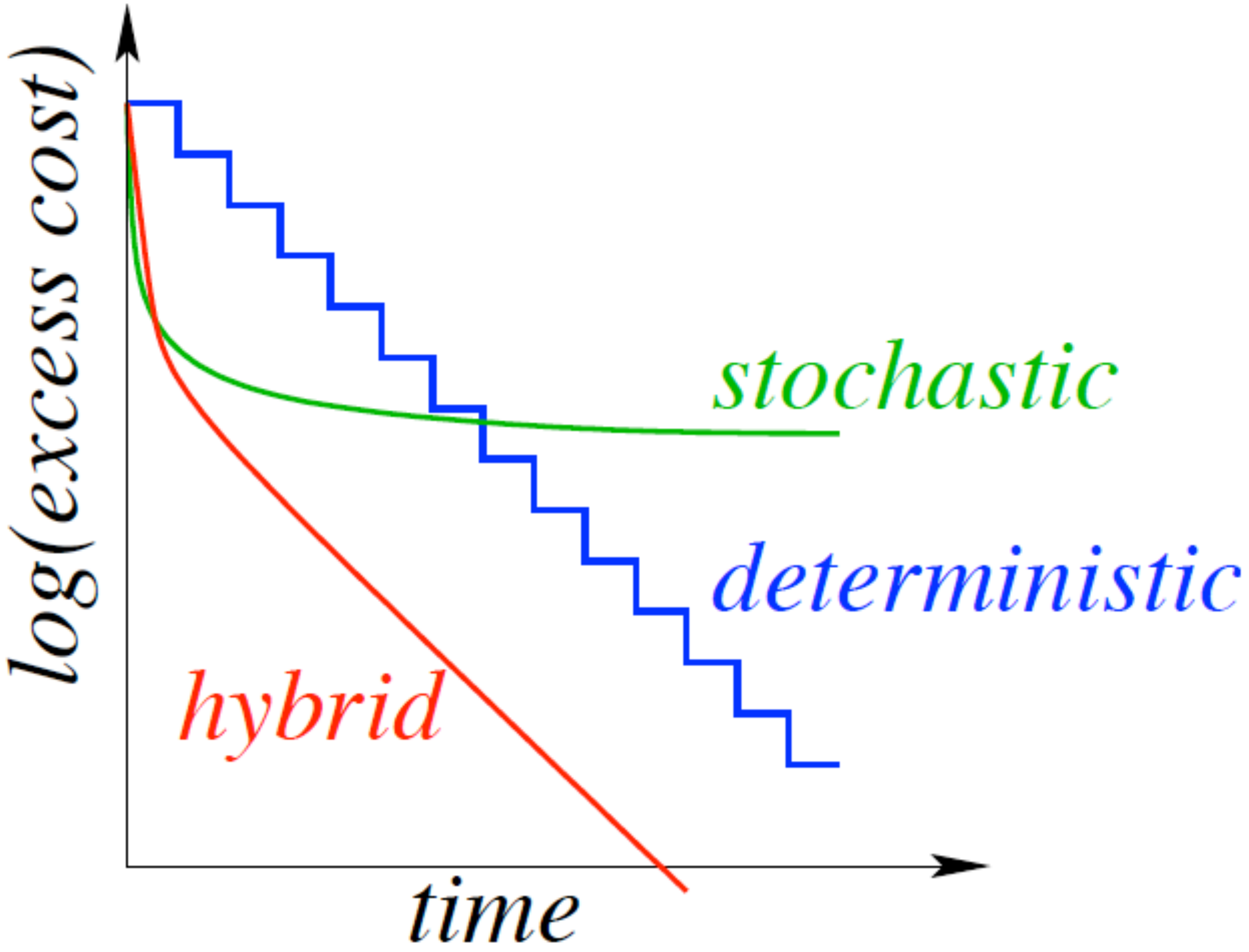
SLIM

FG vs SG

Convergence comparison:

- FG method has O(N) cost with $\mathcal{O}(\rho^k)$ rate;
- SG method has O(1) cost with O(1/t) rate;

Hybrid method:



Stochastic average gradient method

Stochastic average gradient method (SAG):

$$\mathbf{m}^{k+1} = \mathbf{m}^k - \frac{\alpha_k}{n} \sum_{i=1}^{N} y_i^k$$

where

$$y_i^k = \begin{cases} g_i(\mathbf{m}^k) & \text{if } i = i_k, \\ g_i^{k-1} & \text{otherwise} \end{cases}$$

SLIM

Stochastic average gradient method

Convergence rate:

Assume f_i is convex, f_i' is *L*- continuous, φ is μ-strongly convex, with $\alpha_k = \frac{1}{16L}$ the SAG iterations satisfy

$$\mathbb{E}[\varphi(\mathbf{m}^k) - \varphi(\mathbf{m}^*)] \leq (1 - \min\{\frac{\mu}{16L}, \frac{1}{8N}\})^k C$$

Stochastic average gradient method

Convergence rate comparison

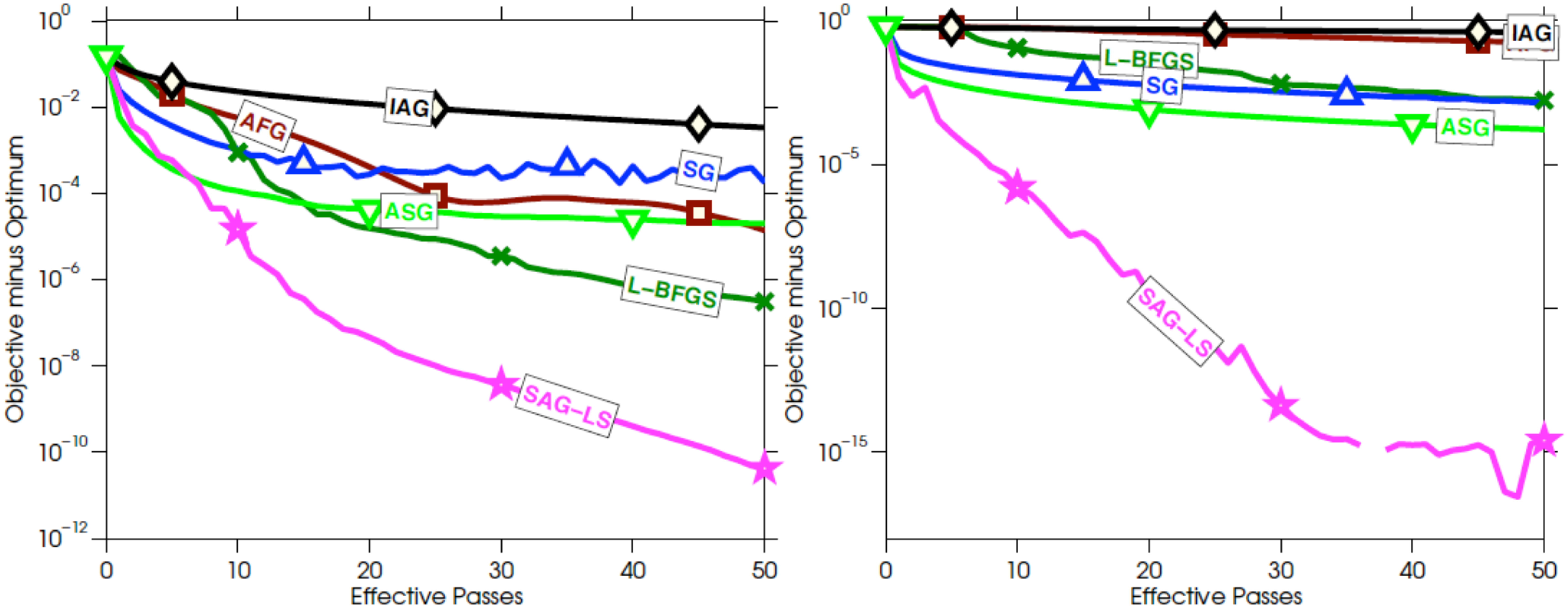
Number of f_i' evaluations to reach an accuracy of ϵ :

- SG – $\mathcal{O}(\frac{L}{\mu}(1/\epsilon))$;
- FG – $\mathcal{O}(N\frac{L}{\mu}log(1/\epsilon))$;
- SAG – $\mathcal{O}(\max\{N, \frac{L}{\mu}\}log(1/\epsilon))$

[Schmidt 2014]

Convergence comparison

quantum (n=50000,p=78) and rcv1(n=697641, p=47236)



[Friedlander and Schmidt 2012]

SLIM

Stochastic gradient method with growing batch size

Stochastic gradient:

$$\overline{G}_k(\mathbf{m}) = \frac{1}{n_{\mathcal{I}_k}} \sum_{i \in \mathcal{I}_k} g_i(\mathbf{m})$$

here, $n_{\mathcal{I}_k} \to N$ slowly.

Sampling strategies:

- Deterministic: pre-determined sample sequence
- Randomized: uniform sampling

SLIM

Stochastic gradient method with growing batch size

Convergence rate:

deterministic –

$$\varphi(\mathbf{m}^k) - \varphi(\mathbf{m}^*) = \mathcal{O}(\rho^k) + \mathcal{O}([\frac{N - n_{\mathcal{I}_k}}{N}]^2)$$

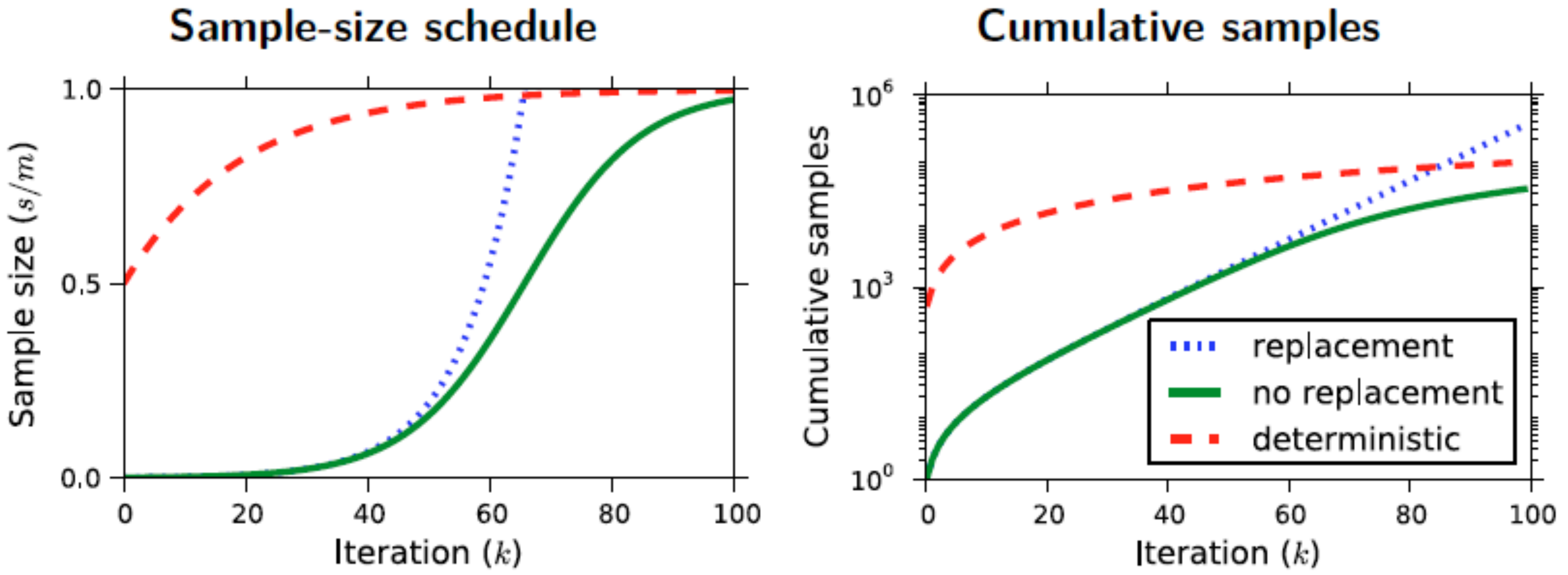
sampling w/o replacement –

$$\mathbb{E}[\varphi(\mathbf{m}^k)] - \varphi(\mathbf{m}^*) = \mathcal{O}(\rho^k) + \mathcal{O}(\frac{N - n_{\mathcal{I}_k}}{N} \cdot \frac{1}{n_{\mathcal{I}_k}})$$

sampling w/ replacement –

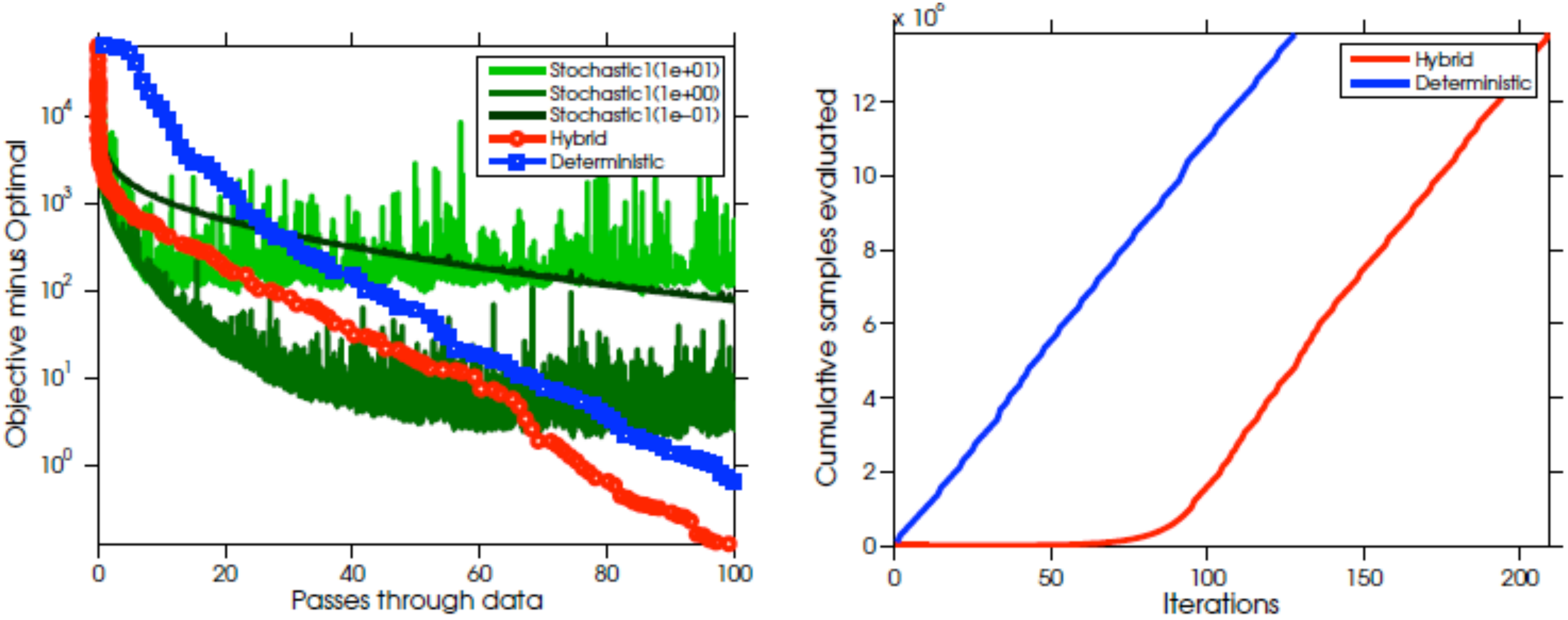
$$\mathbb{E}[\varphi(\mathbf{m}^k)] - \varphi(\mathbf{m}^*) = \mathcal{O}(\rho^k) + \mathcal{O}(\frac{1}{n_{\mathcal{I}_k}})$$

Stochastic gradient method with growing batch size

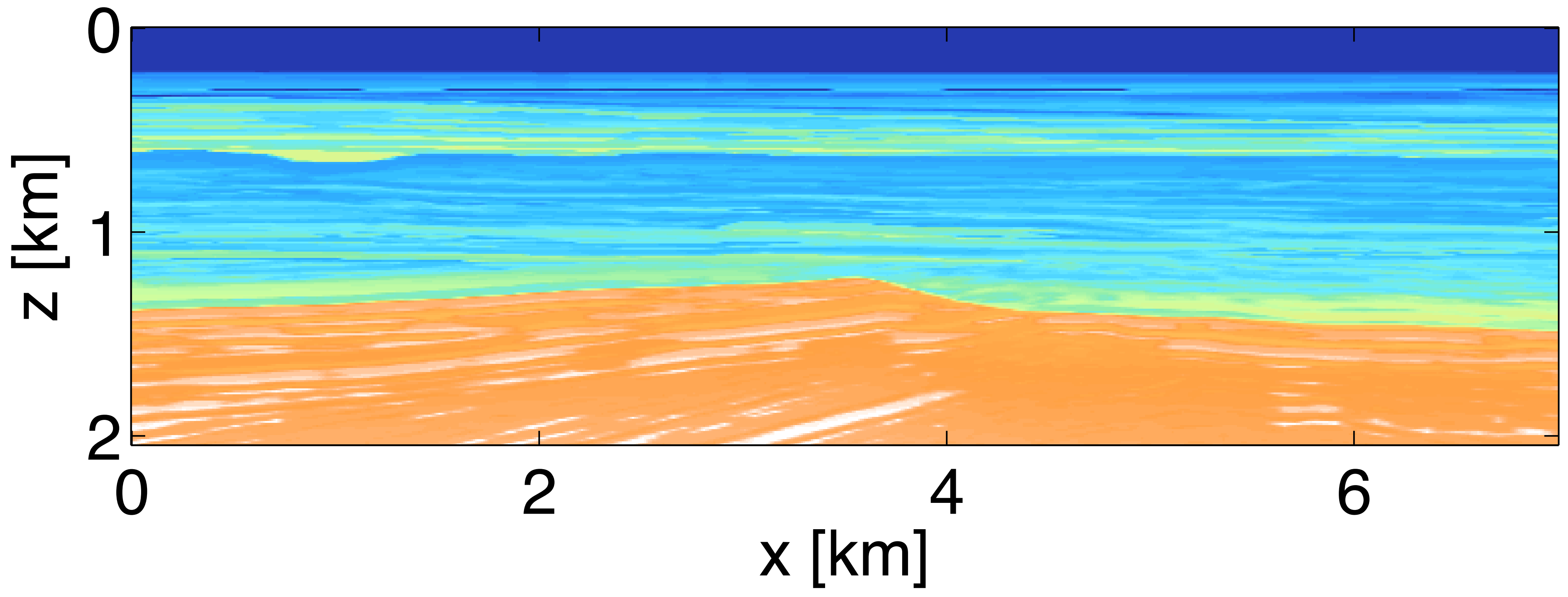


Stochastic gradient method with growing batch size

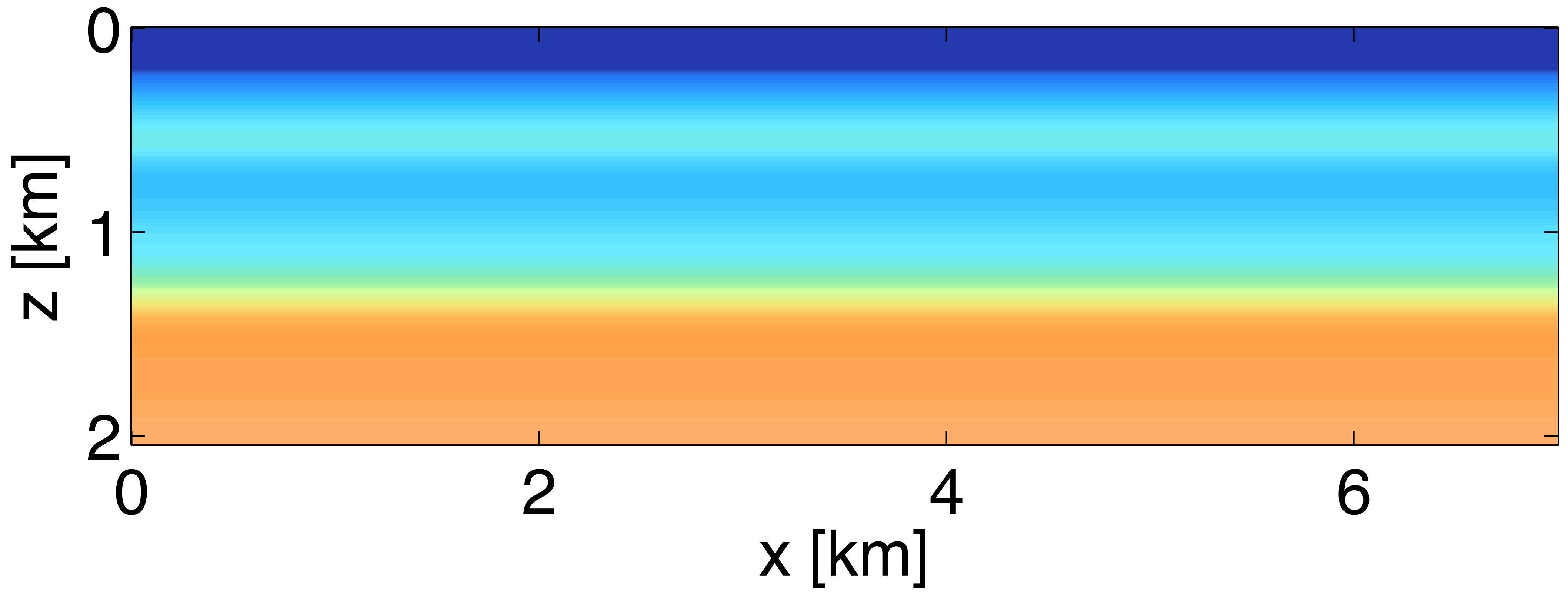
Binary logistic regression experiments:



Application to FWI



data for
141 sources, 281
receivers, 15 Hz Ricker

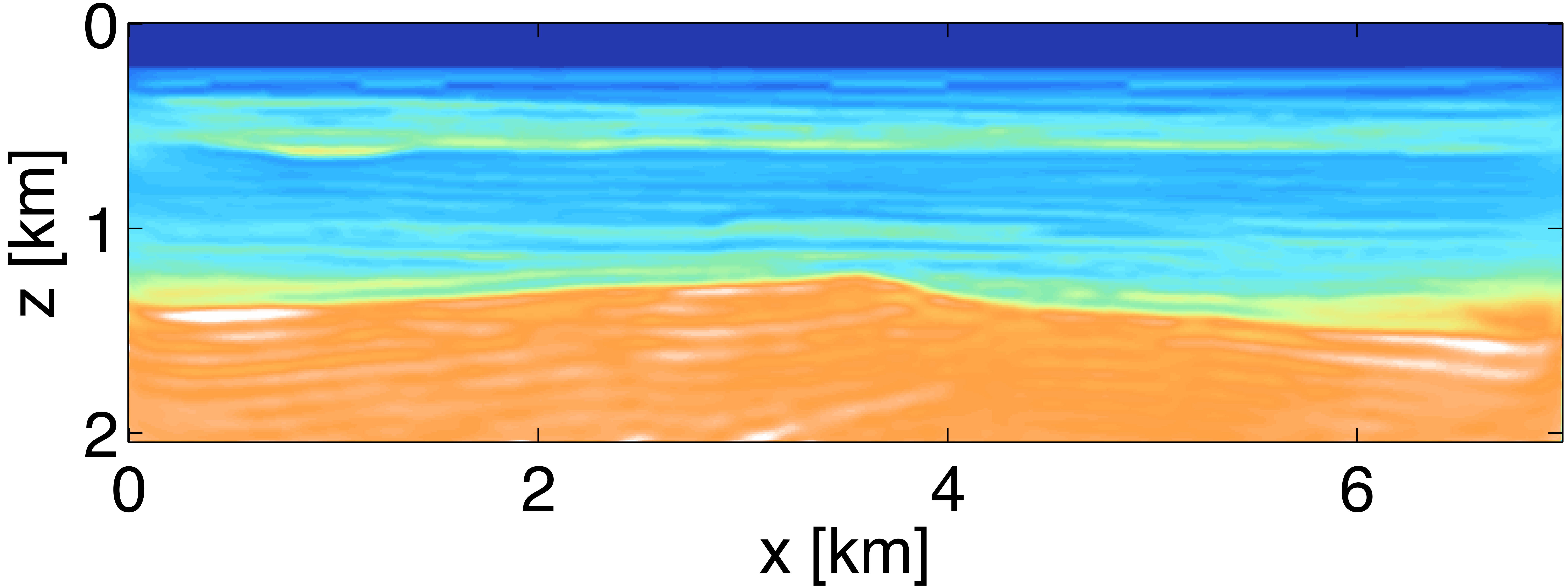


multi-scale frequency
domain inversion:
[2.5-20] Hz in 16 bands

[van Leewwen and FJH 2011]

SLIM

Application to FWI

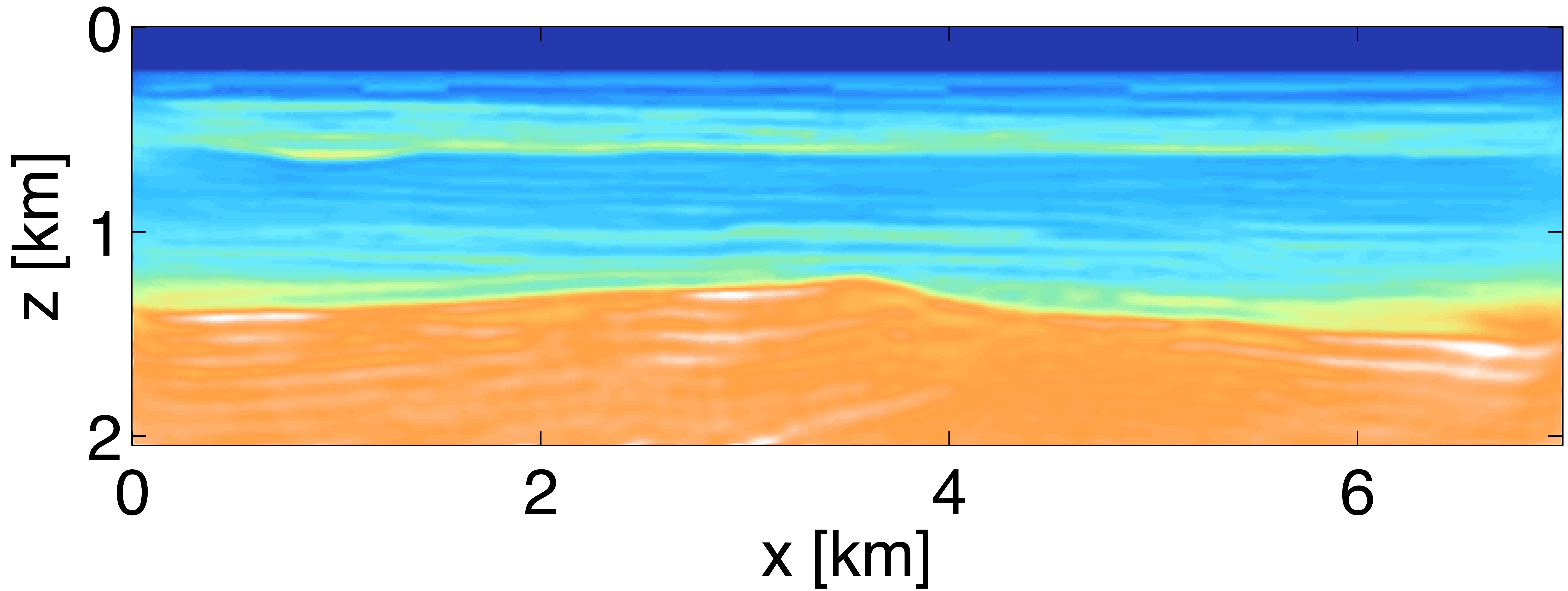


traditional L-BFGS

~15 full evaluations per frequency band

[van Leewwen and FJH 2011]

Application to FWI



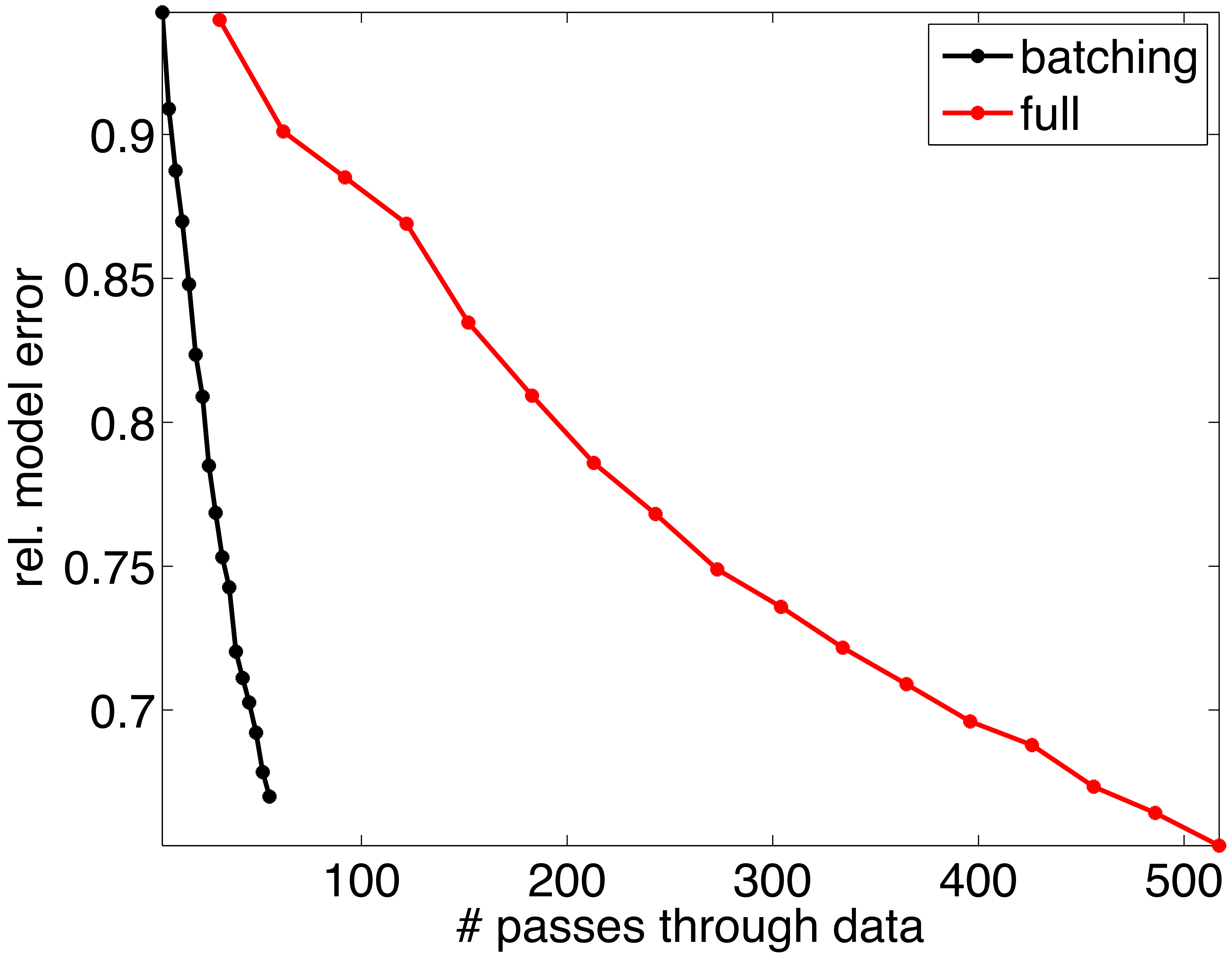
hybrid method

~1.5 full evaluations per frequency band

SLIM

Application to FWI

10 x speedup



Conclusion

- Hybrid method and SGA gives both speed-uo of stochastic method and convergence rate of deterministic method.
- Hybrid method can be applied to the seismic inversion and reduce the computational cost.