Randomized sampling: How confident are you?

Michael P. Friedlander
University of British Columbia

Sinbad Consortium Meeting
December 3–5, 2012

Collaborators: Mark Schmidt and Gabriel Goh

Model problem

$$\underset{x}{\text{minimize}} \quad f(x) \quad := \quad \sum_{i=1}^{m} f_i(x)$$

Examples:

- least-squares $\quad f_i(x) = (a_i^T x - b_i)^2 \qquad f(x) = \|Ax - b\|^2$
- log likelihood $\quad f_i(x) = -\log p(b_i; x)$

Model problem

$$\underset{x}{\text{minimize}} \quad f(x) \quad := \frac{1}{\boldsymbol{m}} \sum_{i=1}^{m} f_i(x)$$

Examples:

- least-squares $f_i(x) = (a_i^T x - b_i)^2$ $f(x) = \|Ax - b\|^2$
- log likelihood $f_i(x) = -\log p(b_i; x)$
- sample average $f_i(x) = f(x, \omega_i)$ $f(x) \approx \mathbf{E}_\omega[f(x, \omega)]$

Model problem

$$\underset{x}{\text{minimize}} \quad f(x) \quad := \frac{1}{m} \sum_{i=1}^{m} f_i(x)$$

Examples:

- least-squares $f_i(x) = (a_i^T x - b_i)^2$ $f(x) = \|Ax - b\|^2$
- log likelihood $f_i(x) = -\log p(b_i; x)$
- sample average $f_i(x) = f(x, \omega_i)$ $f(x) \approx \mathbf{E}_\omega[f(x, \omega)]$

Context:

- m large
- each $f_i(x)$ and $\nabla f_i(x)$ expensive to evaluate

Computing costs:

- count $f_i/\nabla f_i$ evals, not $f/\nabla f$ evals
- minimize passes through full data set $(f_1, \ldots, f_m)$

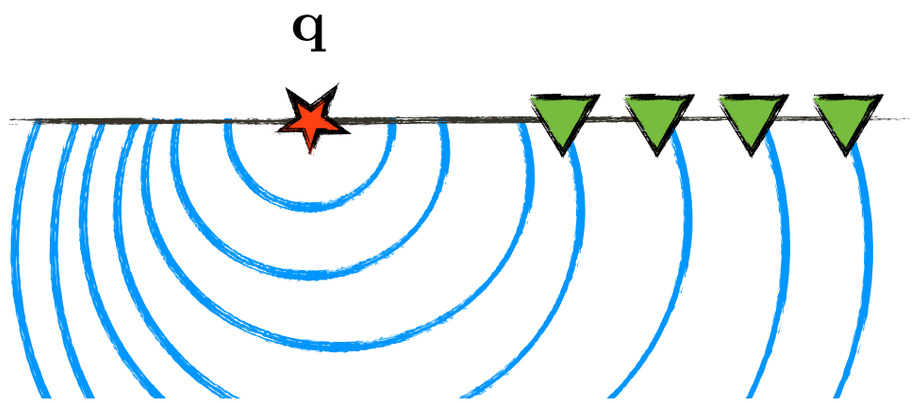
SEISMIC INVERSION

Full waveform inversion

Each of m **experiments** yields a vector of measurements:

sources: $q_1, \ldots, q_m$

measurements: $d_1, \ldots, d_m$



1 source, 1 frequency:

$$\underset{x,u}{\text{minimize}} \ \|d - Pu\|^2 \quad \text{subj to} \quad H_\omega(x)u = q$$

All sources, all frequencies: (eg, 1k sources, $\sim$ 10 freqs)

$$\underset{x}{\text{minimize}} \sum_{i}^{m} \sum_{\omega \in \Omega} \|d_i - PH_\omega(x)^{-1} q_i\|^2$$

Main cost is solution of Helmholtz equation for each (i, ω) pair:

$$H_\omega(x)u = q_i$$

Dimensionality reduction

0. Use all your data (no reduction):

$$f(x) = \sum_{i}^{m} \rho(r_i) \qquad \text{with} \qquad r_i(x) = d_i - F(x)q_i$$

Dimensionality reduction

0. Use all your data (no reduction):

$$f(x) = \sum_{i}^{m} \rho(r_i) \qquad \text{with} \qquad r_i(x) = d_i - F(x)q_i$$

1. Data mixing: form s weighted averages ($s \ll m$)

$$\bar{d}_j := \sum_{i}^{m} w_{ij} d_i \quad \text{and} \quad \bar{q}_j := \sum_{i}^{m} w_{ij} q_i, \qquad j = 1, \ldots, s$$

and optimize

$$\bar{f}(x) = \sum_{j}^{s} \rho(\bar{r}_j) \qquad \text{with} \qquad \bar{r}_j(x) = \bar{d}_j - F(x)\bar{q}_j$$

Dimensionality reduction

0. Use all your data (no reduction):

$$f(x) = \sum_{i}^{m} \rho(r_i) \qquad \text{with} \qquad r_i(x) = d_i - F(x)q_i$$

1. Data mixing: form s weighted averages ($s \ll m$)

$$\bar{d}_j := \sum_{i}^{m} w_{ij} d_i \quad \text{and} \quad \bar{q}_j := \sum_{i}^{m} w_{ij} q_i, \qquad j = 1, \ldots, s$$

and optimize

$$\bar{f}(x) = \sum_{j}^{s} \rho(\bar{r}_j) \qquad \text{with} \qquad \bar{r}_j(x) = \bar{d}_j - F(x)\bar{q}_j$$

2. Sample-average approximations: draw samples and optimize

$$\bar{f}(x) = \sum_{i \in \mathcal{S}} \rho(r_i) \quad \text{with} \quad \mathcal{S} = \{i_1, i_2, \ldots, i_s\}$$

Dimensionality reduction I: data mixing

Least-squares misfit is a **matrix trace**:

$$f(x) = \sum_i^m \|r_i\|^2 = \text{trace}(R^T R) \qquad \text{with} \qquad R = [\, r_1 \mid \cdots \mid r_m \,]$$

Mixing equivalent optimizing original problem over mixed data

$$\bar{f}(x) = \sum_j^s \|\bar{r}_j\|^2 = \text{trace}(\bar{R}^T \bar{R}) \qquad \text{with} \qquad \bar{R} = [\, \bar{r}_1 \mid \cdots \mid \bar{r}_s \,], \quad s \ll m$$

Expected objective property holds:

$$\underset{W}{\mathbf{E}}[\bar{f}(x)] = f(x) \qquad \underset{W}{\mathbf{E}}[\nabla \bar{f}(x)] = \nabla f(x)$$

Stochastic trace estimation: Haber, Chung, Herrmann ('12)

Careful: Expected objective property **only** holds for least-squares

Dimensionality reduction II: sample-average

Applies to generic inverse problem:

$$\min_x \; f(x) := \frac{1}{m}\sum_i^m \rho(r_i) \quad \text{with} \quad R(x) = [\; r_1 \mid \cdots \mid r_m \;]$$

Randomly sample subsets of the data, ie, for $s \ll m$,

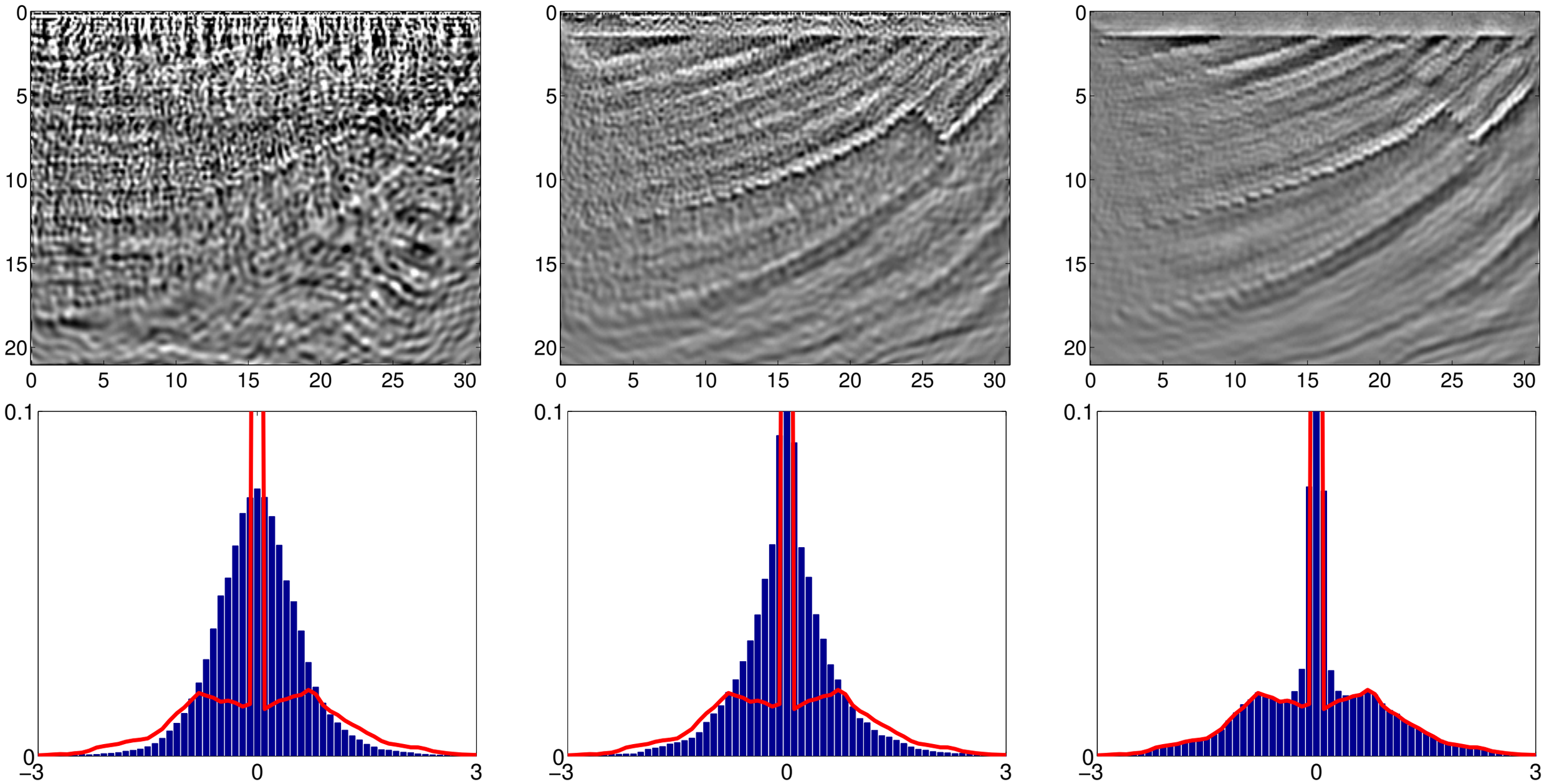
$$\bar{f}(x) := \frac{1}{s}\sum_{i=1}^{s} \rho(r_{(i)}) \quad \text{with} \quad \bar{R}(x) = [\; r_{(1)} \mid r_{(2)} \mid \cdots \mid r_{(s)}]$$

Generically yields desired **"expected objective"** property

$$\mathop{\mathbf{E}}_{W}[\bar{f}(x)] = f(x) \qquad \mathop{\mathbf{E}}_{W}[\nabla \bar{f}(x)] = \nabla f(x)$$

minimize $\rho(R(x))$ with $R(x) = D - F(x)Q$

Normal	Laplace	Student's-t
$\|R(x)\|_F^2$	$\sum_{ij} \|R_{ij}(x)\|_1$	$\sum_{ij} \log\left(k + R_{ij}(x)^2\right)$



MODEL PROBLEM

$$\underset{x}{\text{minimize}} \quad f(x) := \frac{1}{m} \sum_{i=1}^{m} f_i(x)$$

Complexity of steepest descent

Baseline Algorithm:

$$x_{k+1} \leftarrow x_k - \alpha_k \nabla f(x_k), \qquad \alpha_k \equiv 1/L$$

Assume:

convex f; Lipschitz ∇f with param L; $\pi_k := f(x_k) - f(x_*)$

Sublinear rate:

- $\pi_k = \mathcal{O}(1/k)$ (constant stepsize)
- $\pi_k = \mathcal{O}(1/k^2)$ (optimal rate with extrapolation)

Linear rate: additionally assume that f is strongly convex w/ param μ

- $\pi_k = \rho^k \pi_0$ where $\rho := 1 - \mu/L < 1$

Note: if f is twice differentiable, $\mu I \preceq \nabla^2 f(x) \preceq LI$

Incremental gradient methods

Algorithm: $x_{k+1} \leftarrow x_k - \alpha \nabla f_i(x_k), \ i \in \{1, \ldots, m\}$ **cyclic**

Assume: f strongly convex (μ) and Lipschitz ∇f (L)

Constant stepsize: $\alpha_k \equiv \bar{\alpha}$

- $\pi_k \leq \rho^k \pi_0 + \mathcal{O}(\boldsymbol{m^2 \bar{\alpha}})$ k **full cycles**

Decreasing stepsize: $\sum_k \alpha_k = \infty, \ \sum_k \alpha_k^2 < \infty$

- $\pi_k \leq \mathcal{O}(1/k)$ k **full cycles**

Incremental gradient methods

Algorithm: $x_{k+1} \leftarrow x_k - \alpha \nabla f_i(x_k),\ i \in \{1, \ldots, m\}$ **cyclic** , **randomized**

Assume: f strongly convex (μ) and Lipschitz ∇f (L)

Constant stepsize: $\alpha_k \equiv \bar{\alpha}$

- $\pi_k \leq \rho^k \pi_0 + \mathcal{O}(\boldsymbol{m^2 \bar{\alpha}})$ k **full cycles**
- $\mathbf{E}\pi_k \leq \rho^k \pi_0 + \mathcal{O}(\boldsymbol{m \bar{\alpha}})$ k **iterations**

Decreasing stepsize: $\sum_k \alpha_k = \infty,\ \sum_k \alpha_k^2 < \infty$

- $\pi_k \leq \mathcal{O}(1/k)$ k **full cycles**
- $\mathbf{E}\pi_k \leq \mathcal{O}(1/k)$ k **iterations**

Incremental gradient methods

Algorithm: $x_{k+1} \leftarrow x_k - \alpha \nabla f_i(x_k),\ i \in \{1, \ldots, m\}$ **cyclic** , **randomized**

Assume: f strongly convex (μ) and Lipschitz ∇f (L)

Constant stepsize: $\alpha_k \equiv \bar{\alpha}$

- $\pi_k \leq \rho^k \pi_0 + \mathcal{O}(\boldsymbol{m^2 \bar{\alpha}})$ k **full cycles**
- $\mathbf{E}\pi_k \leq \rho^k \pi_0 + \mathcal{O}(\boldsymbol{m \bar{\alpha}})$ k **iterations**

Decreasing stepsize: $\sum_k \alpha_k = \infty,\ \sum_k \alpha_k^2 < \infty$

- $\pi_k \leq \mathcal{O}(1/k)$ k **full cycles**
- $\mathbf{E}\pi_k \leq \mathcal{O}(1/k)$ k **iterations**

Many variations:

Luo/Tseng '94; Nedić/Bertsekas '00/'10; Blatt et al '08; Bottou '10

EXAMPLE

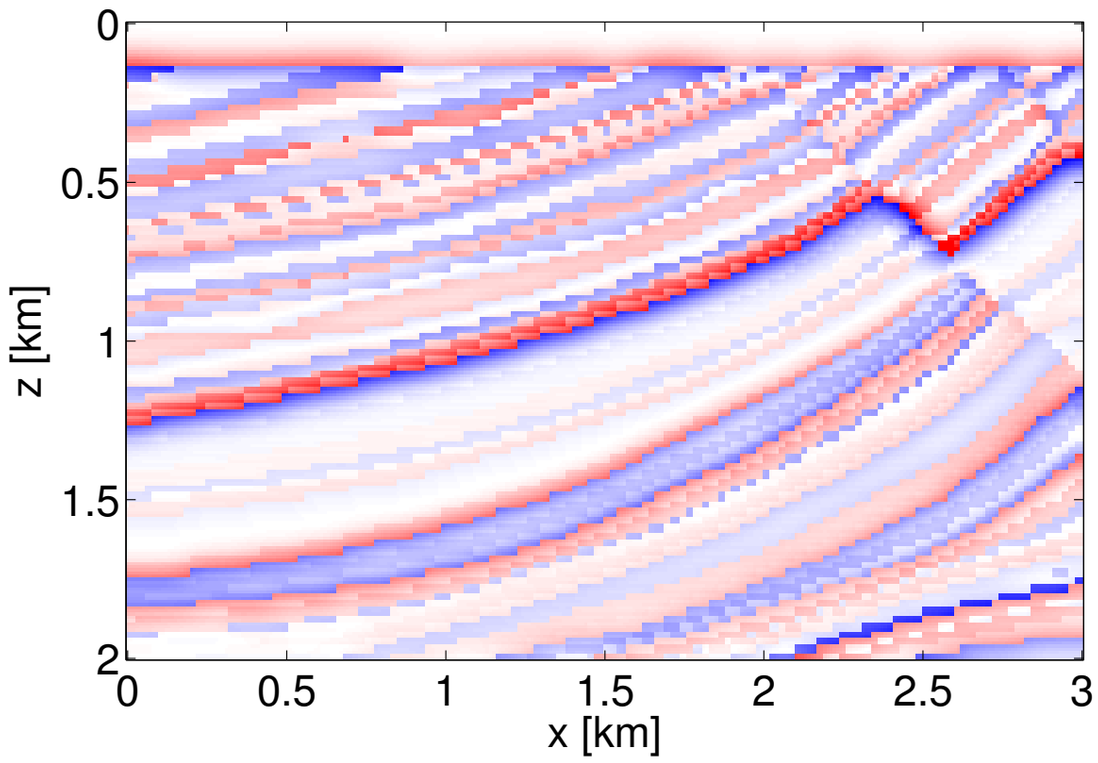
Seismic inversion

Recover image of geological structures via nonlinear least squares

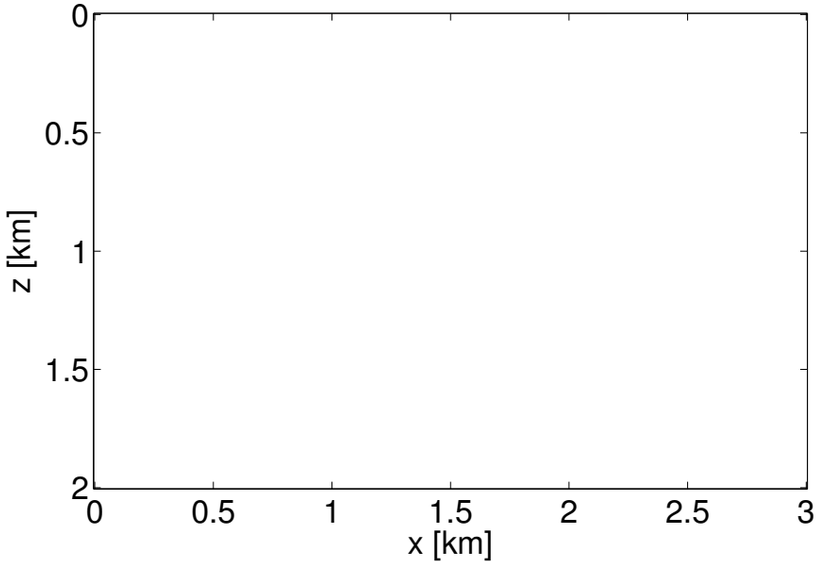
$$\underset{x}{\text{minimize}} \sum_{i}^{m} \sum_{\omega \in \Omega} \| d_i - P H_\omega(x)^{-1} q_i \|^2$$

Observations: Each of m "shots" is an experiment:

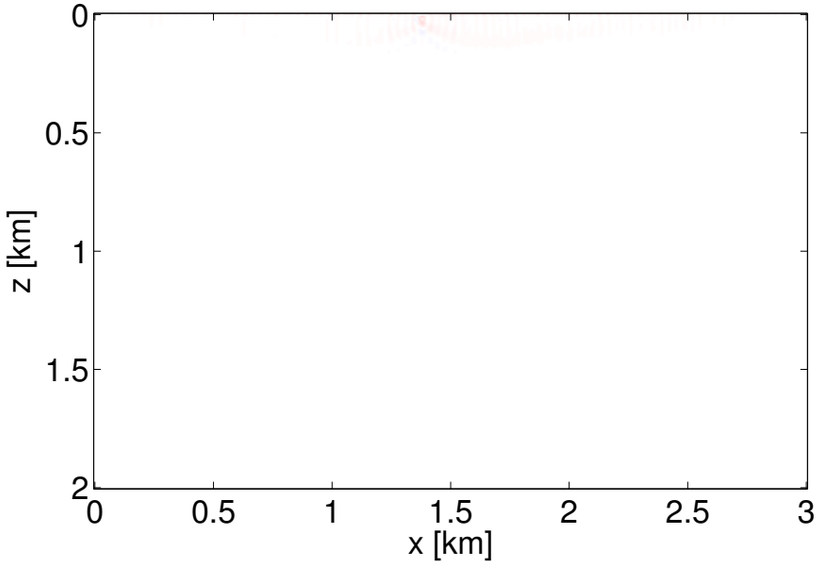
sources: $q_1, \ldots, q_m$, measurements: $d_1, \ldots, d_m$



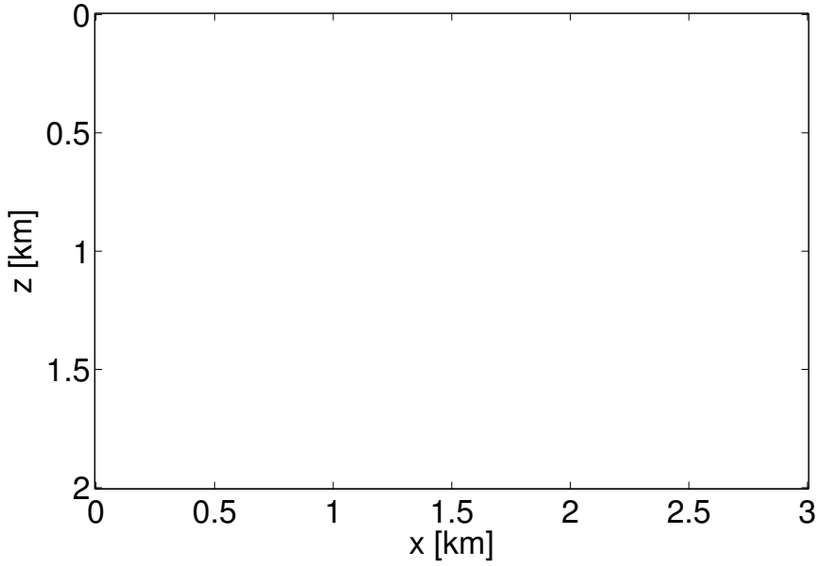
full gradient



incremental gradient

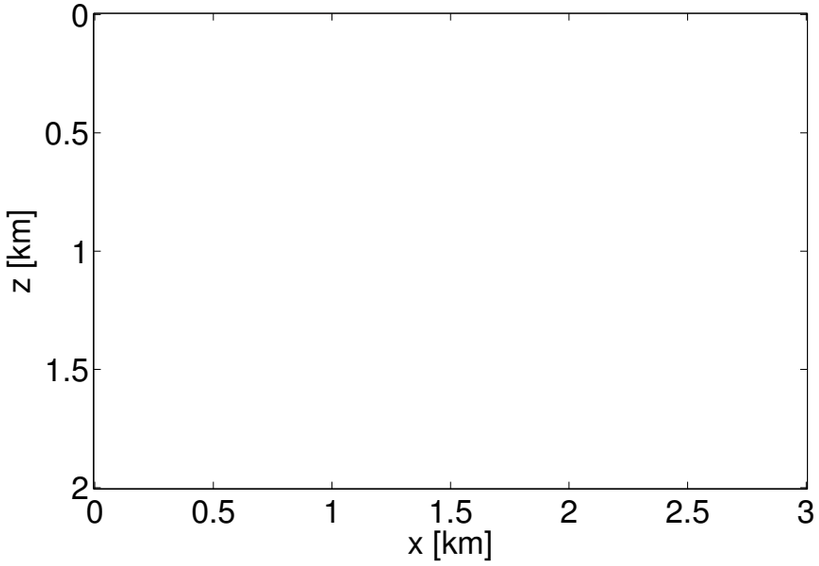


gradient sampling

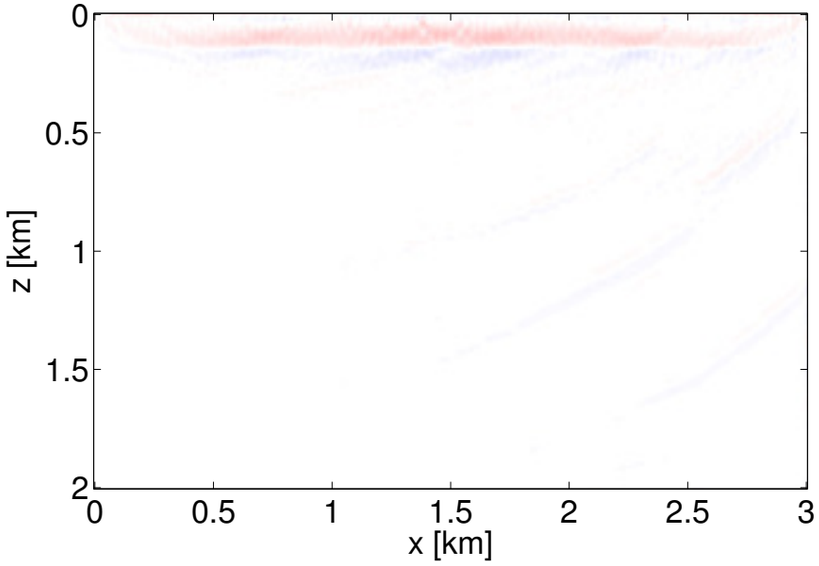


0.01 of 39 passes

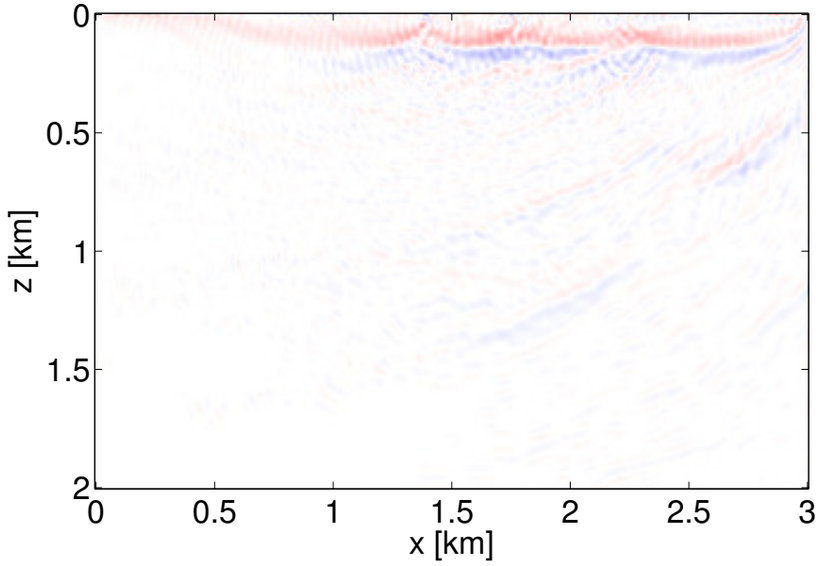
full gradient



incremental gradient

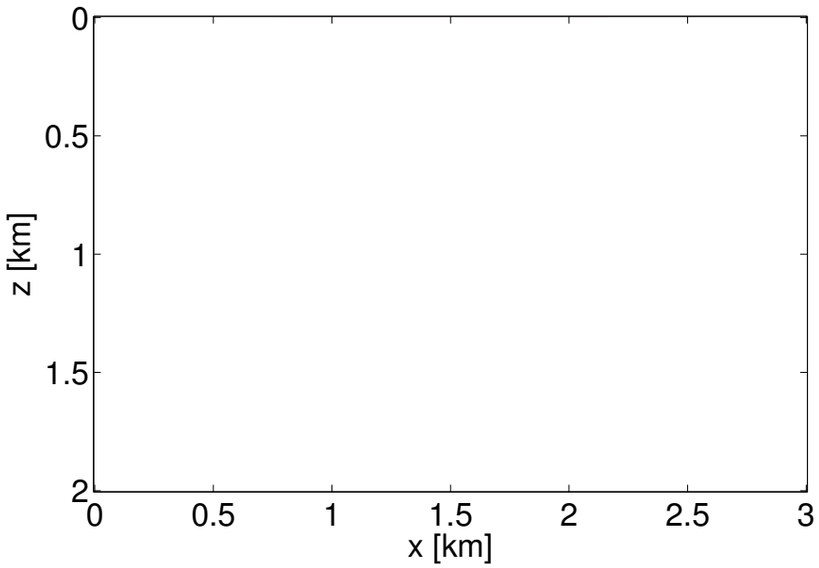


gradient sampling

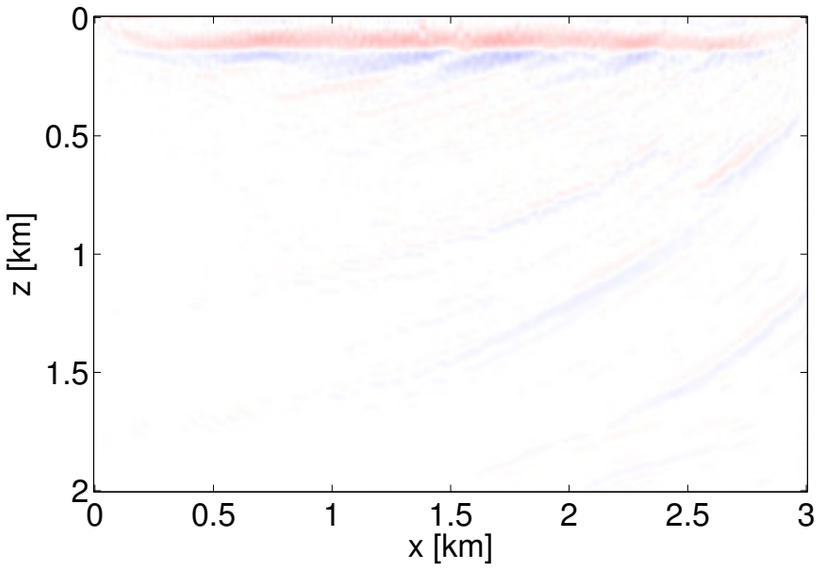


0.4 of 39 passes

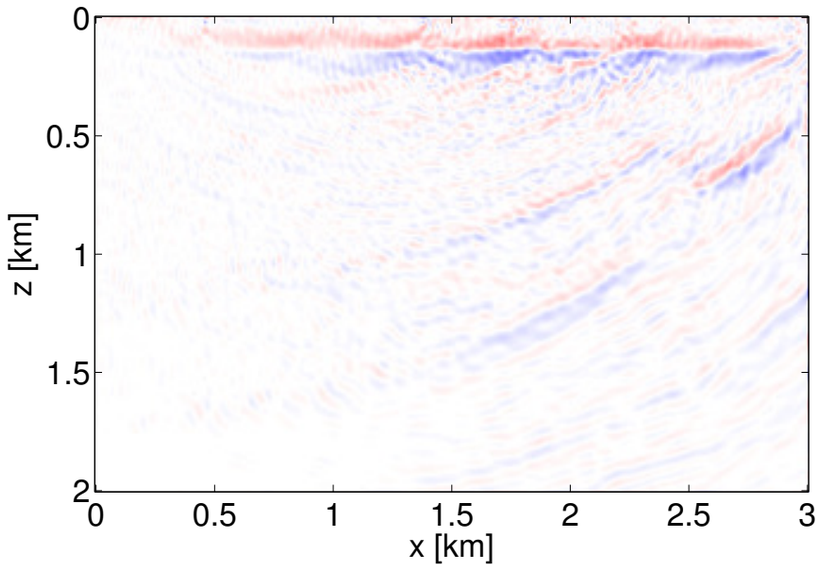
full gradient



incremental gradient

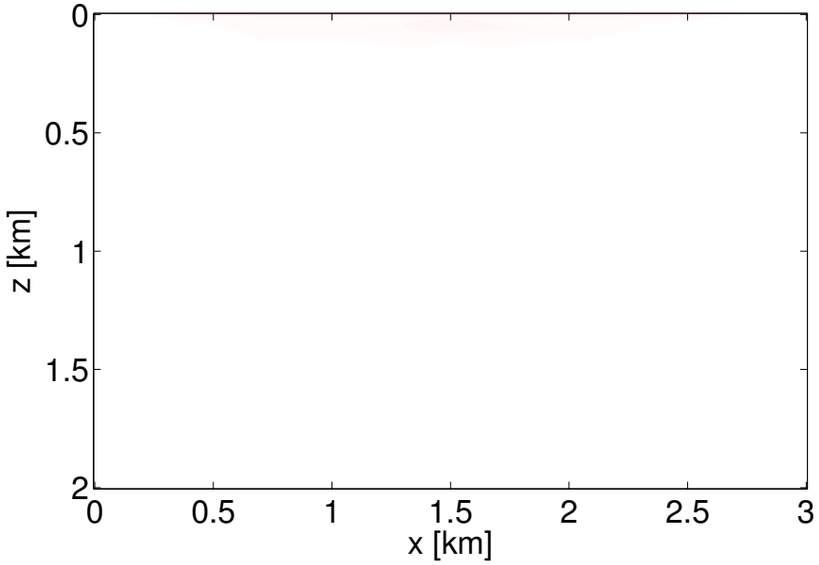


gradient sampling

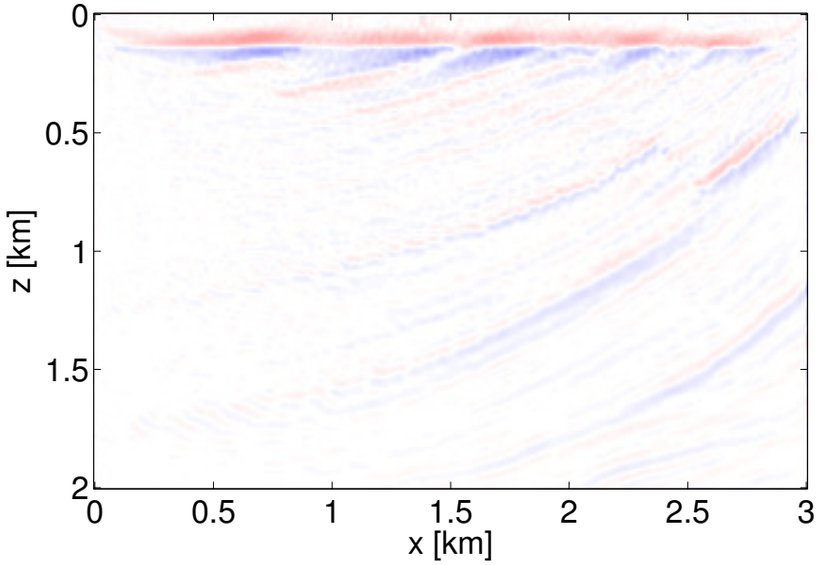


0.8 of 39 passes

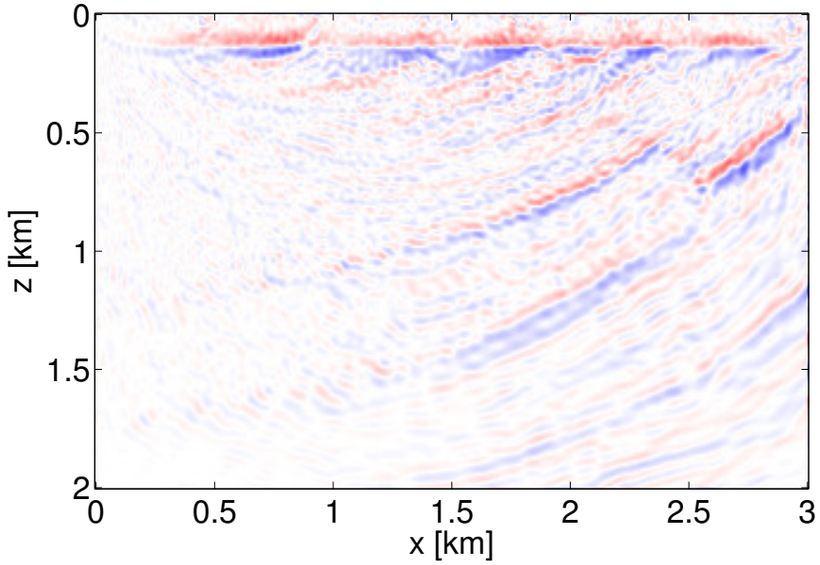
full gradient



incremental gradient

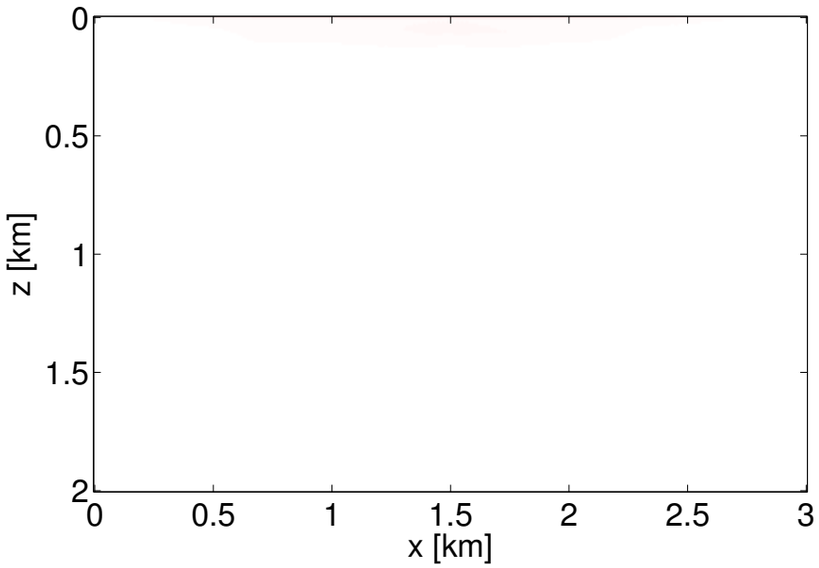


gradient sampling

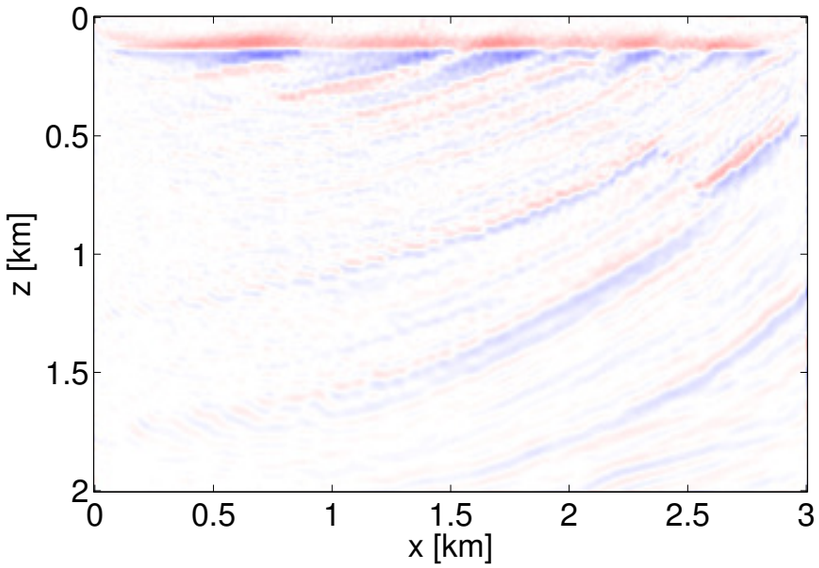


2 of 39 passes

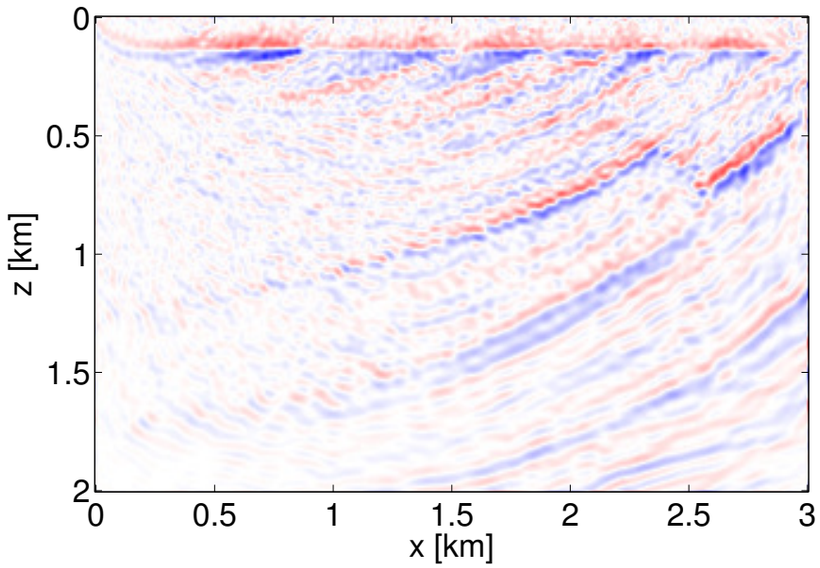
full gradient



incremental gradient

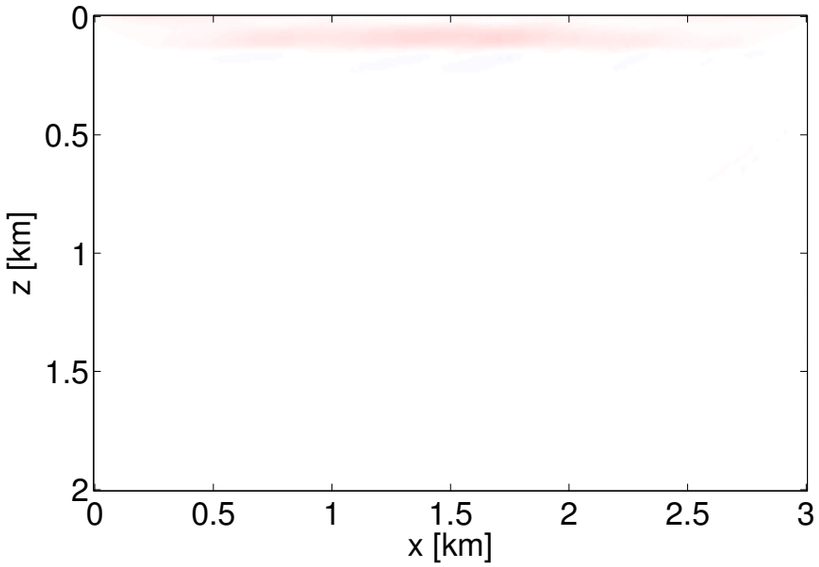


gradient sampling

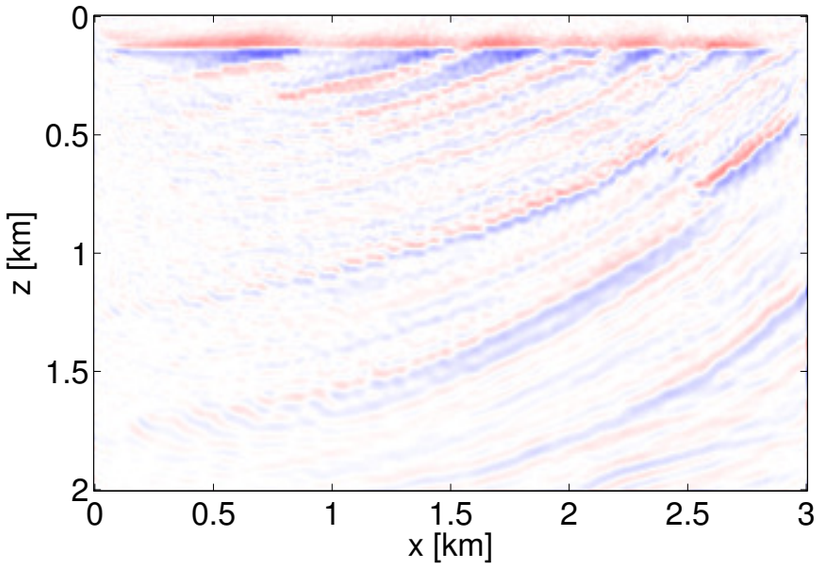


2.6 of 39 passes

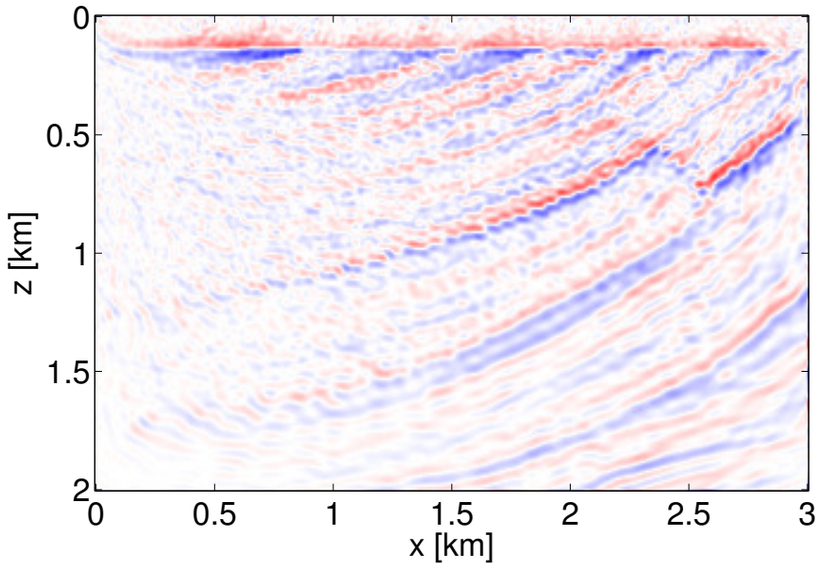
full gradient



incremental gradient



gradient sampling

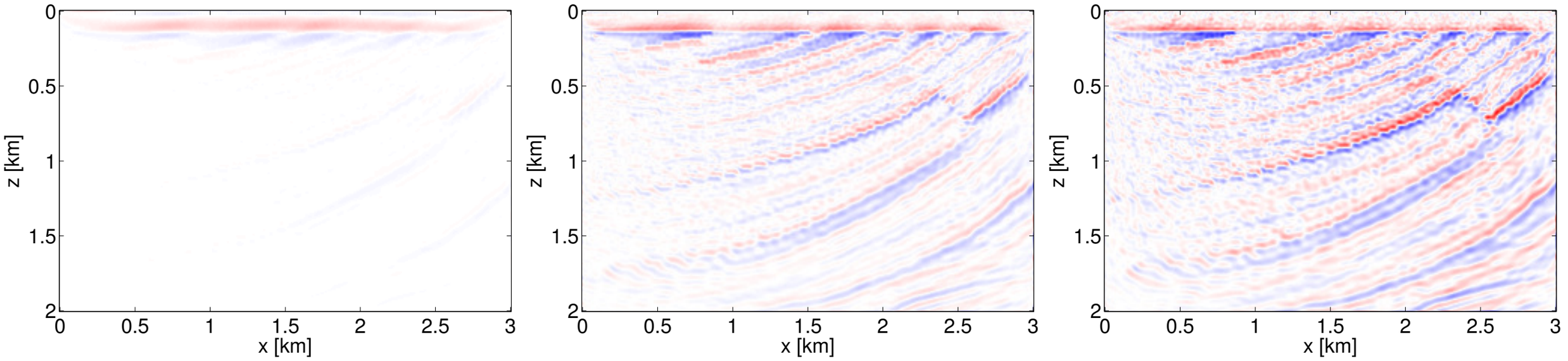


4 of 39 passes

full gradient

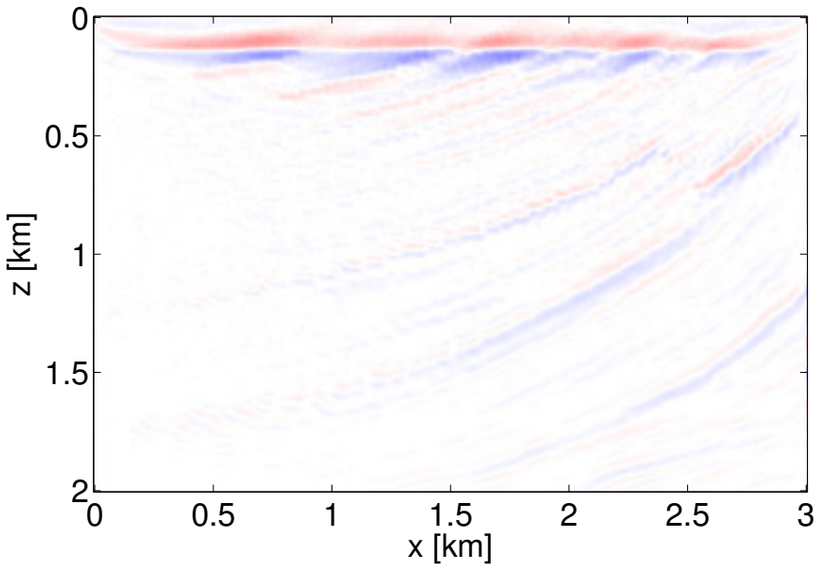
incremental gradient

gradient sampling

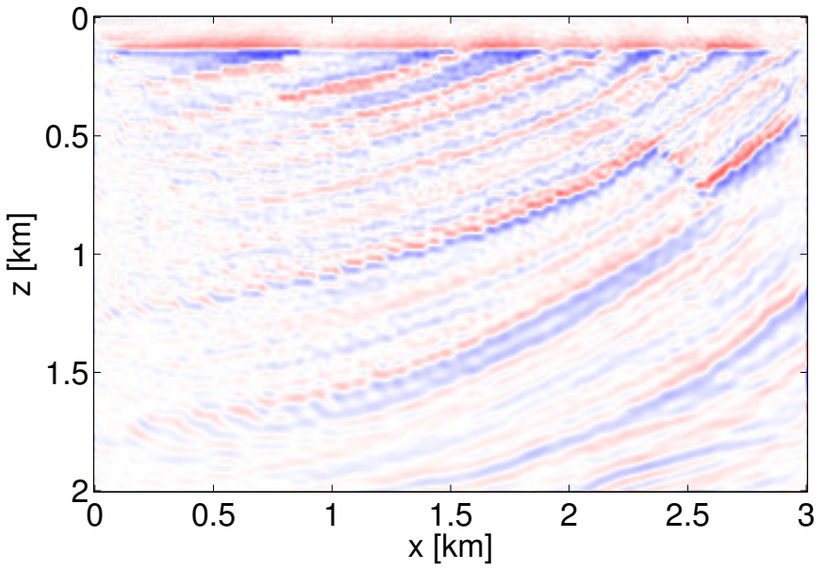


7 of 39 passes

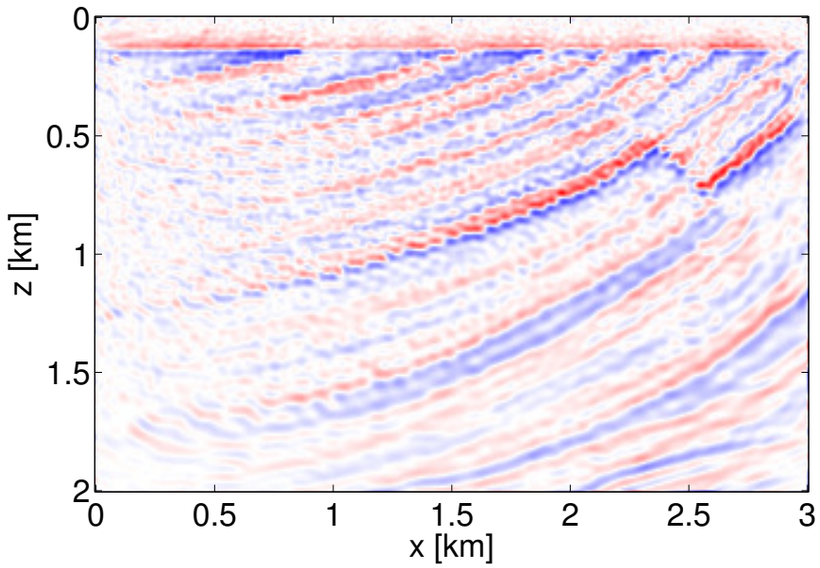
full gradient



incremental gradient

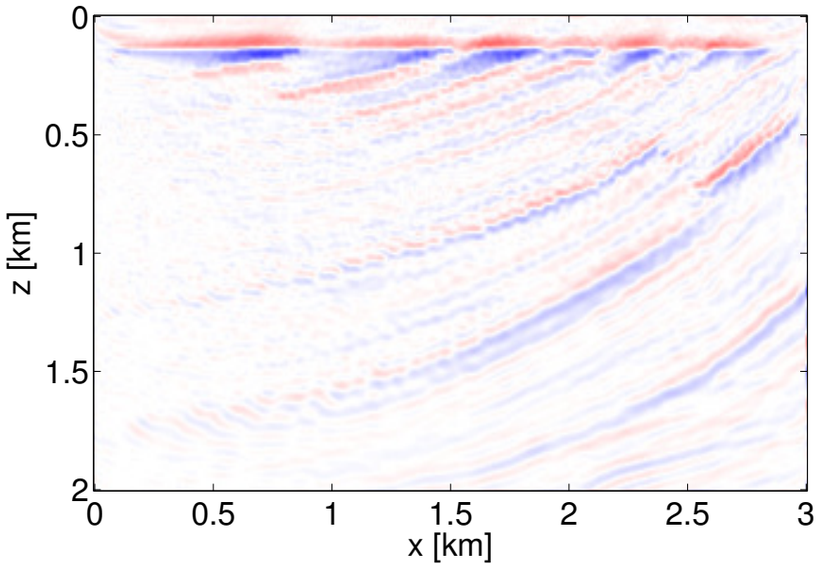


gradient sampling

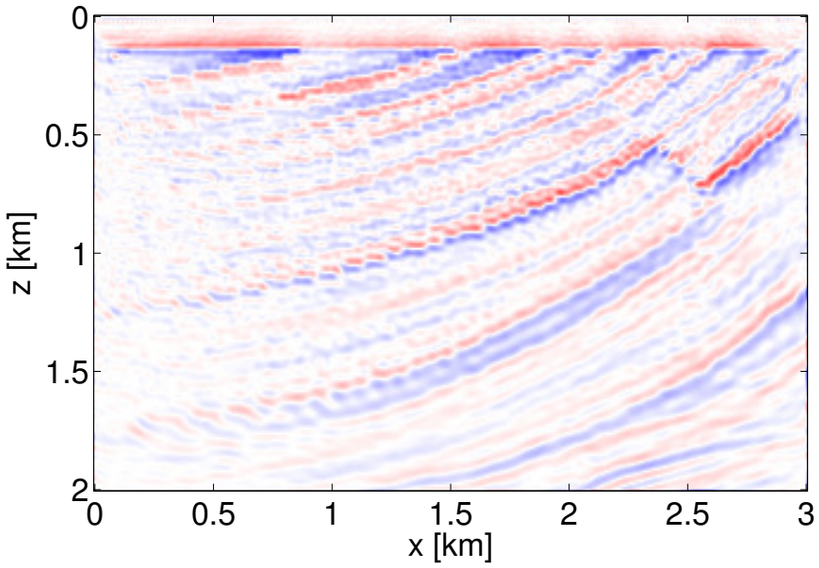


10 of 39 passes

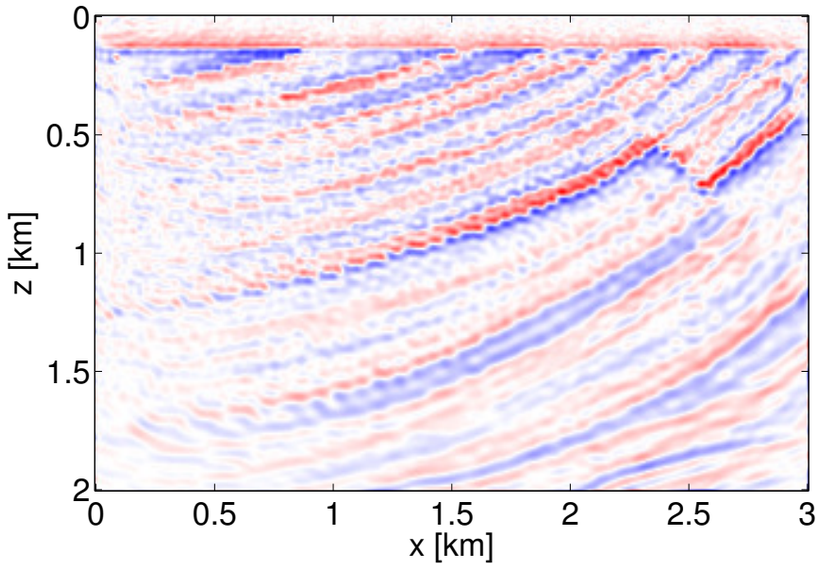
full gradient



incremental gradient

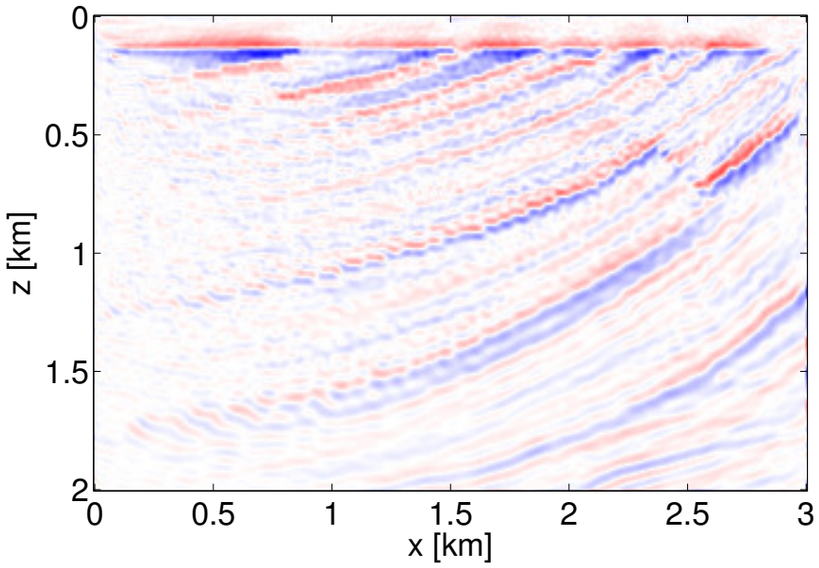


gradient sampling

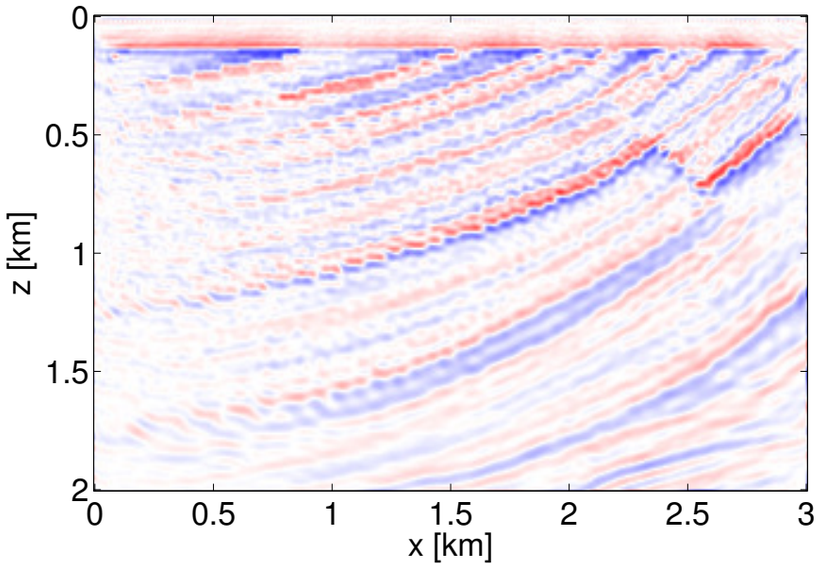


16 of 39 passes

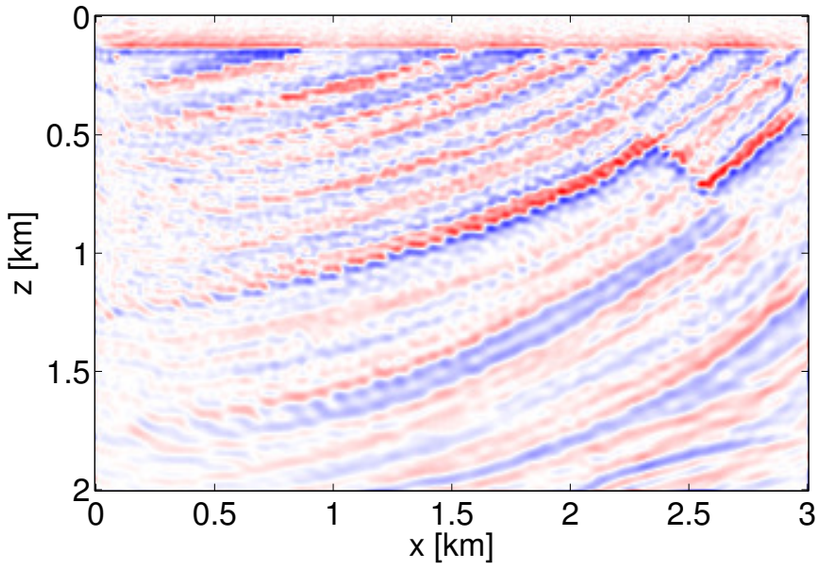
full gradient



incremental gradient

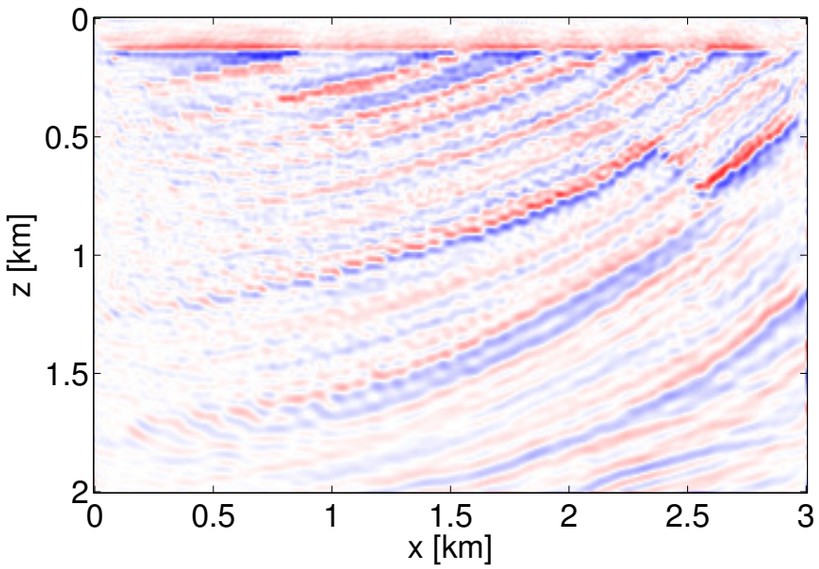


gradient sampling

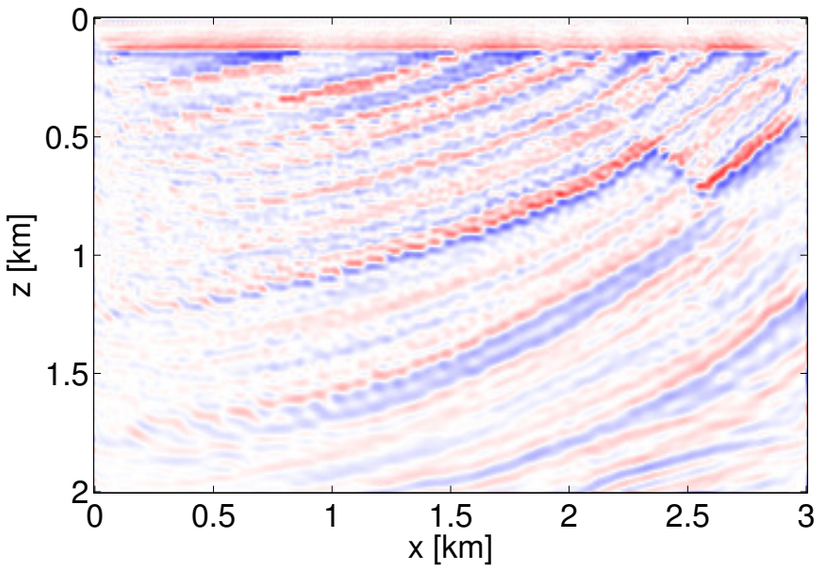


22 of 39 passes

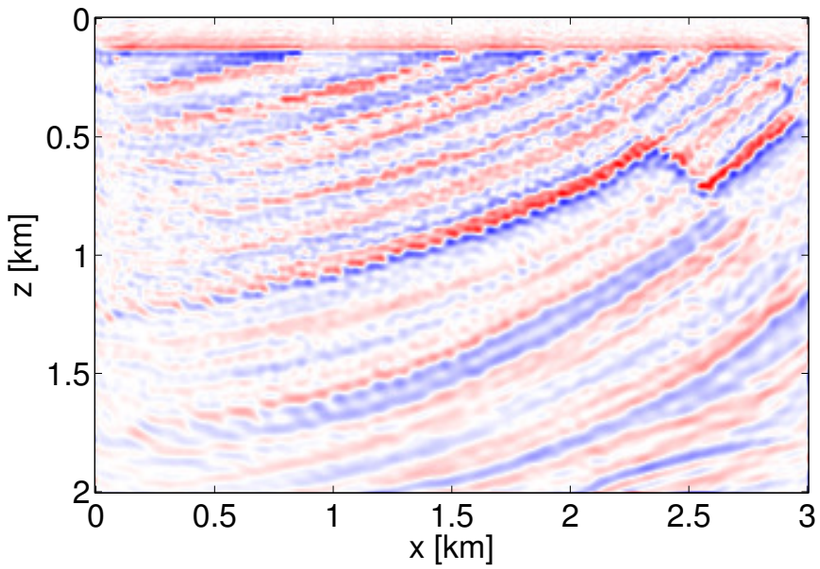
full gradient



incremental gradient

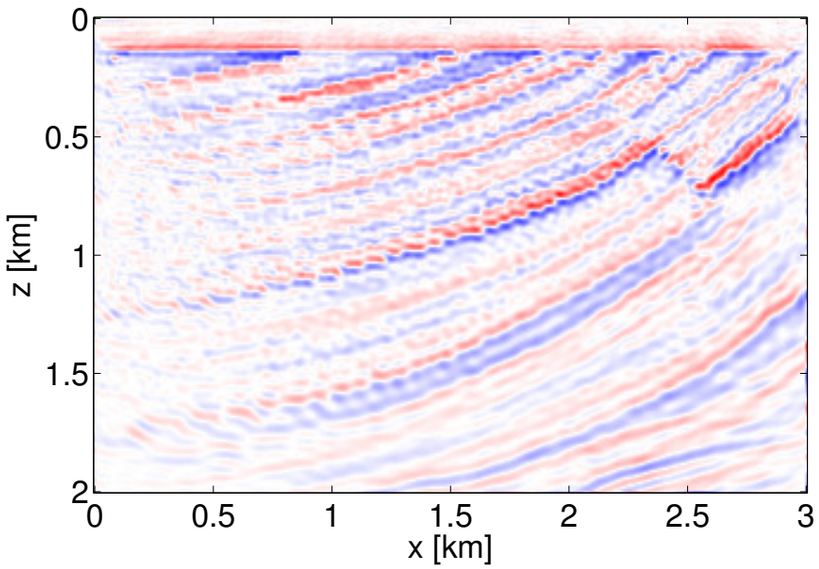


gradient sampling

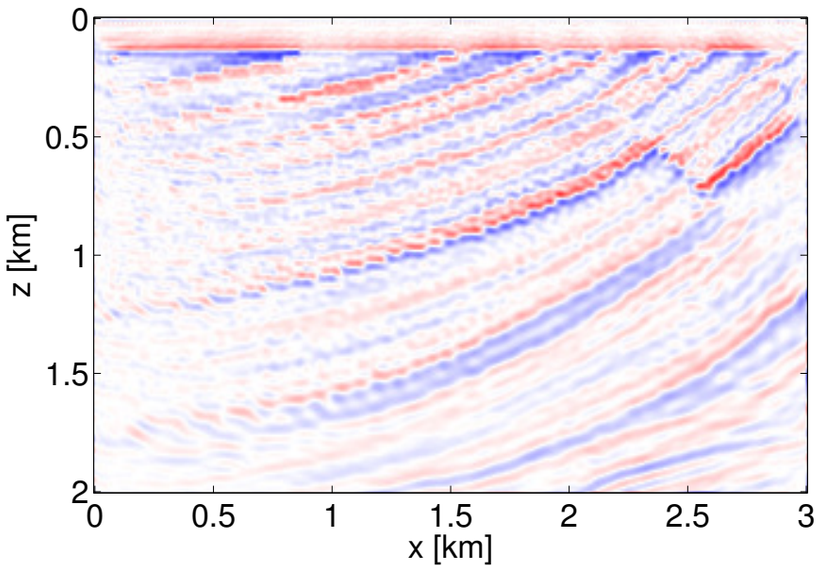


30 of 39 passes

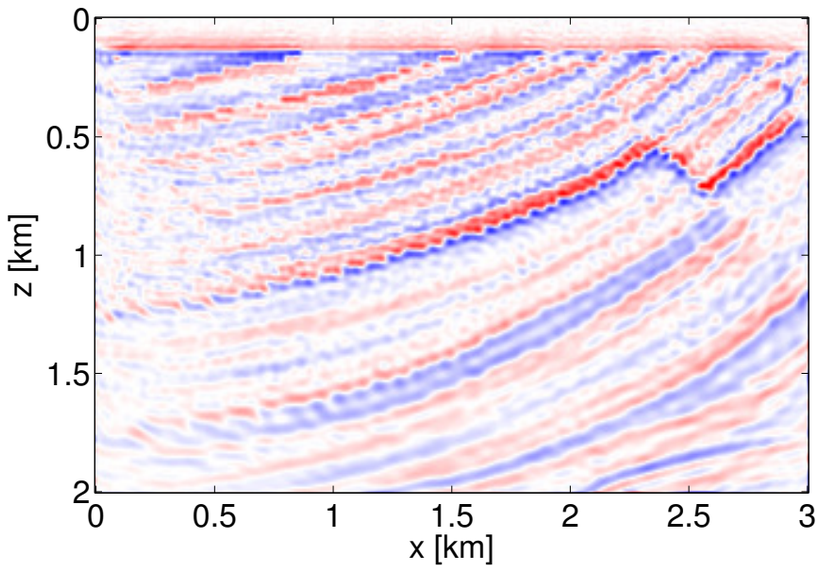
full gradient



incremental gradient



gradient sampling



39 of 39 passes

Sampling approach

Increasing batch: [Bertsekas/Tsitsiklis '96, Shapiro/H-do-M '00]

$$\mathcal{S}_k \subseteq \{1, \ldots, m\}, \qquad s_k \to m \quad \text{(slowly)}$$

Sample-average gradient:

$$g_k(x) := \frac{1}{s_k} \sum_{i \in \mathcal{S}_k} \nabla f_i(x)$$

Algorithm:

$$x_{k+1} \leftarrow x_k - \alpha_k d \quad \text{with} \quad H_k d = -g_k(x_k)$$

Goal: **non-asymptotic analysis** based on controlling gradient error

- $g_k(x) = \nabla f(x_k) + e_k$ where $\mathbf{E}[\|e_k\|^2] \leq \epsilon_k$
- How to control sample size s_k?

Gradient with generic errors

Prototype algo:

$$x_{k+1} \leftarrow x_k - \alpha g_k, \qquad g_k = \nabla f(x_k) + e_k, \qquad \alpha \text{ fixed}$$

Assumptions: Lipschitz gradient (L); strong convexity (μ)

Theorem (convergence rate): for all $k = 1, 2, \ldots$ [F. & Schmidt '12]

$$\mathbf{E}\,\pi_k \leq \rho^k \pi_0 + \mathcal{O}(\mathbf{E}\,\|e_k\|^2)$$

Examples:

linear: $\mathbf{E}\,\|e_k\|^2 = \mathcal{O}(\gamma^k) \implies \pi_k = \mathcal{O}(\max\{\gamma, \rho\}^k)$

sublinear: $\mathbf{E}\,\|e_k\|^2 = \mathcal{O}(1/k^2) \implies \pi_k = \mathcal{O}(1/k^2)$

persistent err: $\inf_k \mathbf{E}\,\|e_k\|^2 > 0 \implies \pi_k = \mathcal{O}(\mathbf{E}\,\|e_k\|^2)$

Generic errors: tail bounds

Convergence rate in **expectation**:

$$\mathbf{E}\,\pi_k \leq \rho^k \pi_0 + \mathcal{O}(\mathbf{E}\,\|e_k\|^2)$$

> "But I don't observe the mean!"
> — An observant critic

Answer: bound the **probability of deviation** from the mean

$$\Pr(\pi_k - \rho^k \pi_0 \geq \epsilon)$$

Theorem (tail bounds): for all $k = 1, 2, \ldots,$ [F. & Goh '12]

$$\Pr(\pi_k - \rho^k \pi_0 \geq \epsilon) \leq \inf_{\theta > 0} \left\{ \exp(-\theta\epsilon[1-\rho]) \sum_{i=0}^{k-1} \rho^{k-1-i} \gamma_i(\theta) \right\}$$

where $\gamma_i(\theta)$ is the moment generating function of $\|e_i\|^2$

Example

Steepest descent with independent Gaussian noise:

$$x_{k+1} = x_k - \frac{1}{L} g_k, \qquad g_k = \nabla f(x_k) + e_k \qquad e_k \sim N(0, \sigma^2 I)$$

Tail-bound simplifies to

$$\Pr(\pi_k - \rho^k \pi_0 \geq \epsilon) = \mathcal{O}\left[\exp\left(-L\frac{\epsilon}{\sigma}\right)\right] \quad \text{(for } \epsilon \text{ large enough)}$$

Exponential decrease in ϵ (deviation from steepest descent)

Growing the sample size

Prototype algo:

$$x_{k+1} \leftarrow x_k - \alpha g_k, \qquad g_k = \frac{1}{s_k} \sum_{i \in \mathcal{S}_k} \nabla f_i(x_k), \qquad \mathcal{S}_k \subseteq \{1, \dots, m\}$$

Sampling strategies

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

Theorem (convergence rates): for all $k = 1, 2, \dots$ [F. & Schmidt '12]

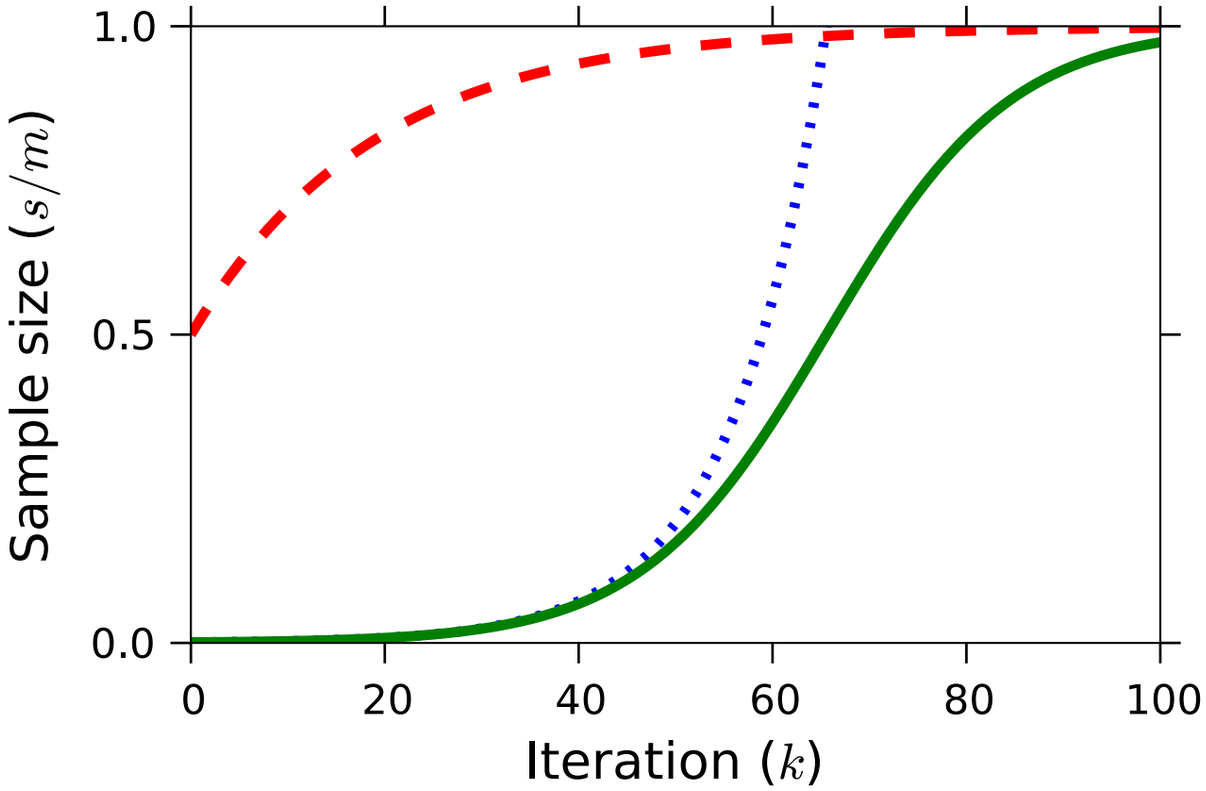
deterministic: $\pi_k = \mathcal{O}(\rho^k) + \mathcal{O}\left(\left[\frac{m - s_k}{m}\right]^2\right)$

sampling w/o replacement $\mathbf{E}\pi_k = \mathcal{O}(\rho^k) + \mathcal{O}\left(\frac{m - s_k}{m} \cdot \frac{1}{s_k}\right)$

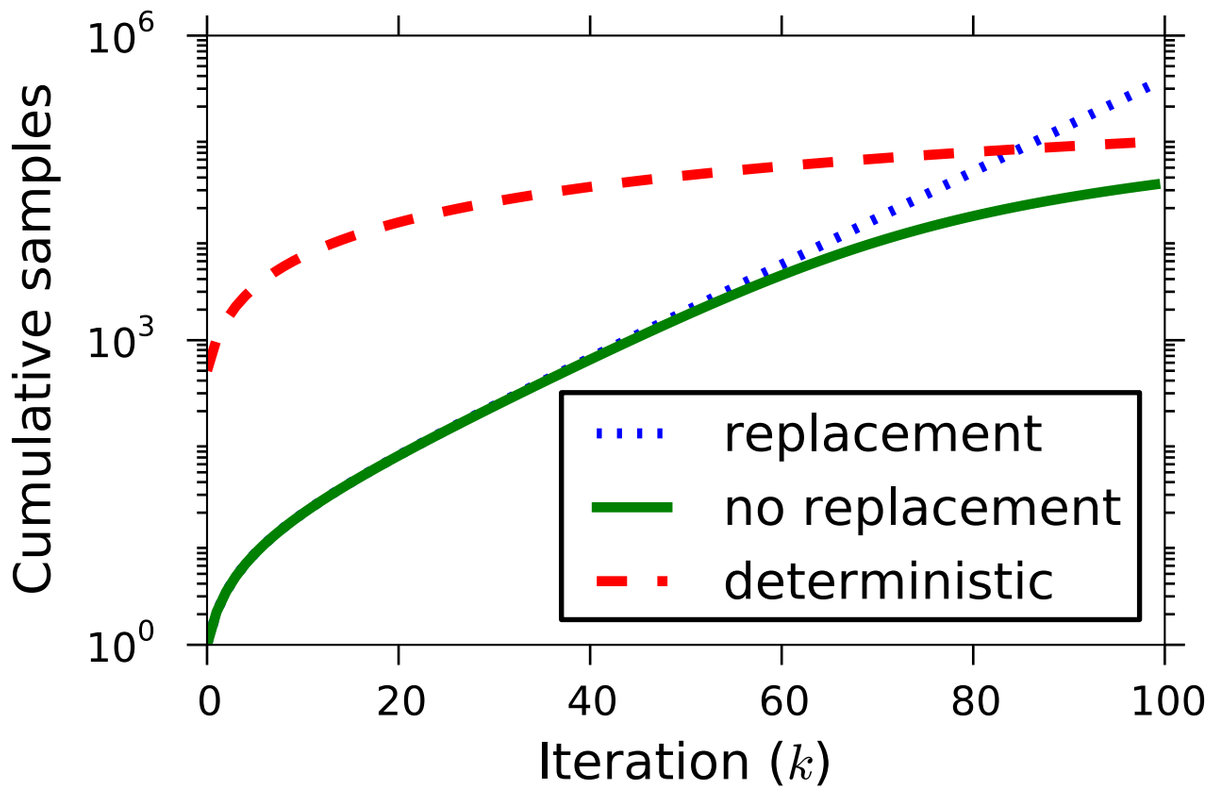
sampling w/ replacement $\mathbf{E}\pi_k = \mathcal{O}(\rho^k) + \mathcal{O}\left(\frac{1}{s_k}\right)$

Illustration

Sample-size schedule



Cumulative samples



Tail bounds and sample policies

Increase sample at geometric rate:

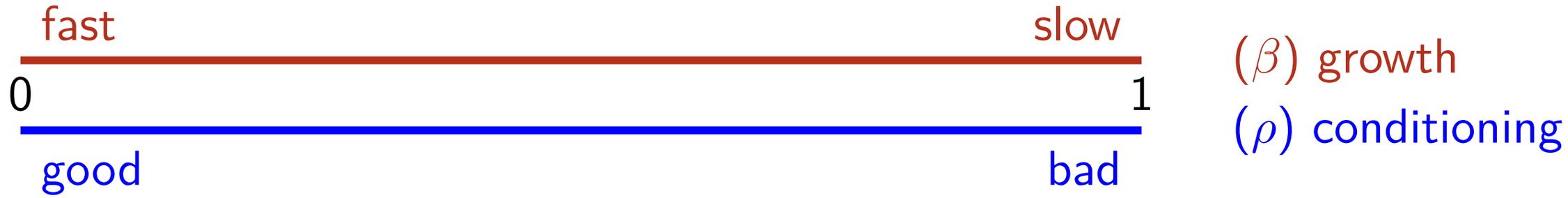
$$s_{k+1} = (1/\beta)s_k \quad \text{for} \quad \beta \in (0, 1]$$

Expectation bound:

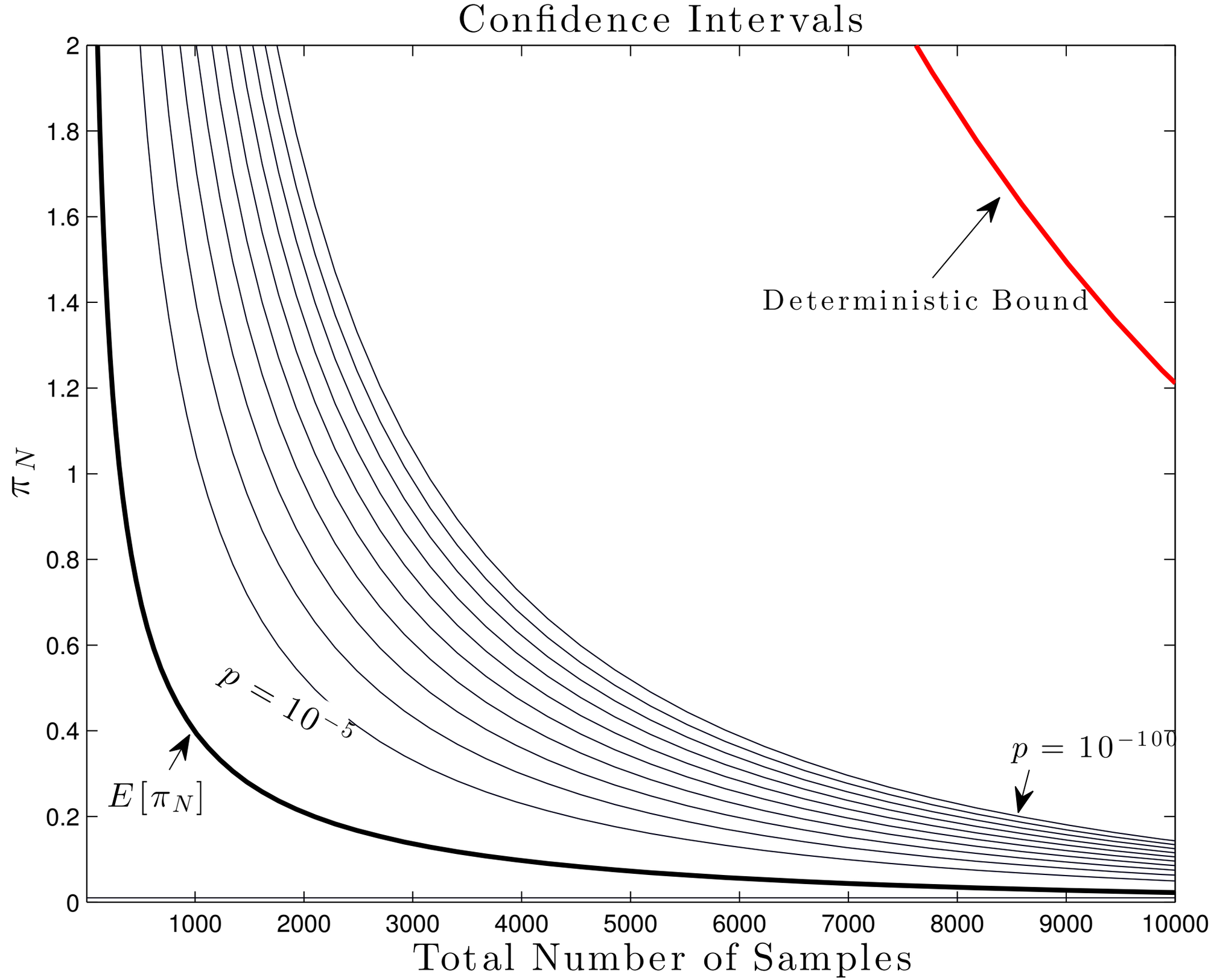
$$\mathbf{E}(\pi_k - \rho^k \pi_0) = \mathcal{O}(\max\{\beta, \rho\}^k)$$

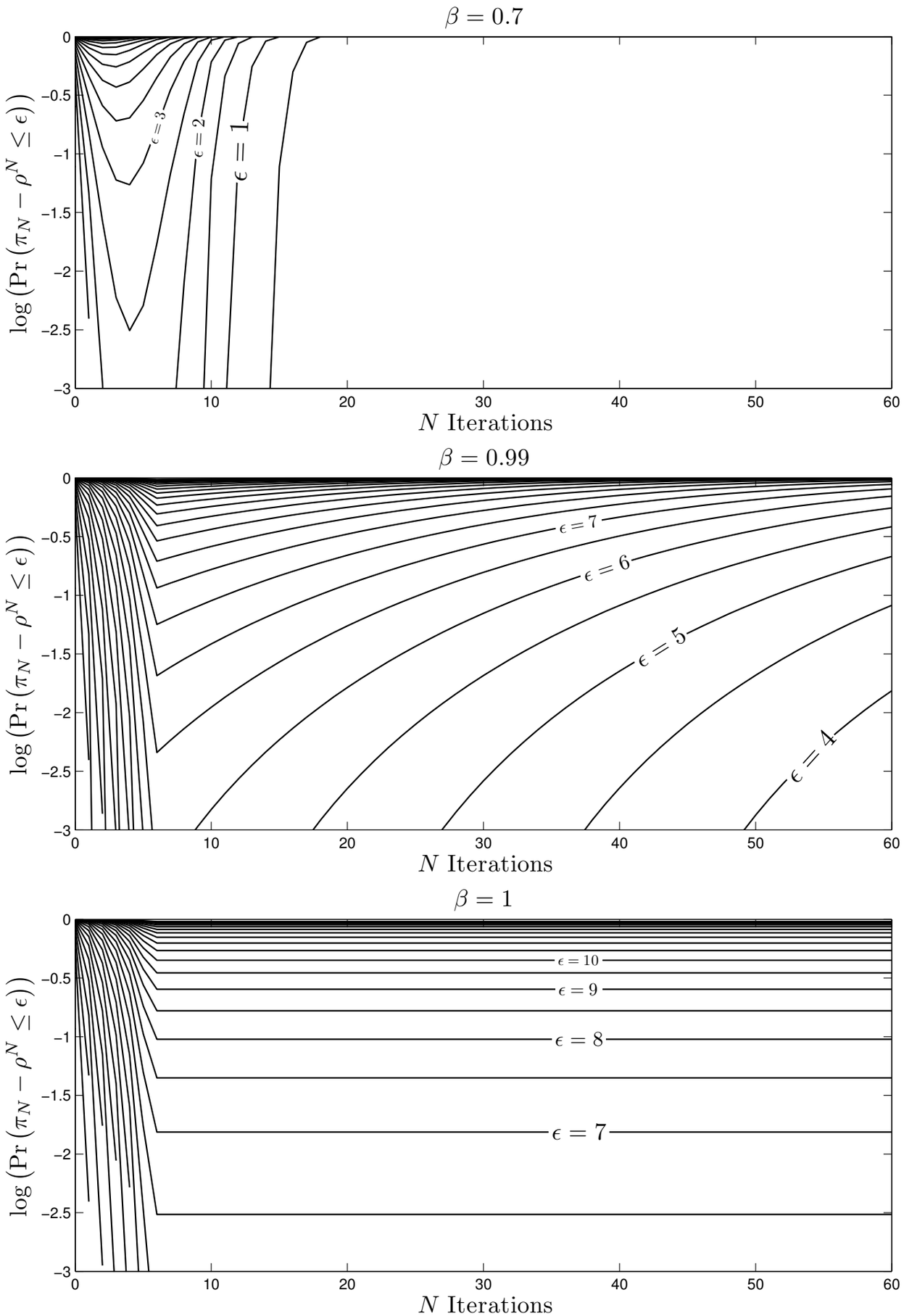
Tail bound:

$$\Pr(\pi_k - \rho^k \pi_0 \geq \epsilon) = \mathcal{O}\left(\exp\left[-\frac{\epsilon}{\max\{\beta, \rho\}^k}\right]\right)$$



Distance to solution vs. cumulative samples ($M = 500,\ \beta = \rho = .9$)




$\beta = 0.7$
$\epsilon = 3$
$\epsilon = 2$
$\epsilon = 1$
$\log\left(\Pr\left(\pi_N - \rho^N \leq \epsilon\right)\right)$
N Iterations
$\beta = 0.99$
$\epsilon = 7$
$\epsilon = 6$
$\epsilon = 5$
$\epsilon = 4$
$\log\left(\Pr\left(\pi_N - \rho^N \leq \epsilon\right)\right)$
N Iterations
$\beta = 1$
$\epsilon = 10$
$\epsilon = 9$
$\epsilon = 8$
$\epsilon = 7$
$\log\left(\Pr\left(\pi_N - \rho^N \leq \epsilon\right)\right)$
N Iterations

In practice

Algorithm:

$$x_{k+1} \leftarrow x_k - \alpha_k \boldsymbol{d}, \qquad \boldsymbol{B_k d} = -\boldsymbol{g_k}(x_k), \qquad \boldsymbol{g_k}(x) = \frac{1}{s_k} \sum_{i \in \mathcal{S}_k} \nabla f_i(x)$$

In practice

Algorithm:

$$x_{k+1} \leftarrow x_k - \alpha_k \boldsymbol{d}, \qquad \boldsymbol{B_k d} = -\boldsymbol{g_k}(x_k), \qquad \boldsymbol{g_k}(x) = \frac{1}{s_k} \sum_{i \in \mathcal{S}_k} \nabla f_i(x)$$

Hessian approximations $\boldsymbol{B_k}$:

- (limited-memory) quasi-Newton: [Schraudolph et al '07]

$$s_k := x_{k+1} - x_k, \qquad y_k := \boldsymbol{g_k}(x_{k+1}) - \boldsymbol{g_k}(x_k)$$

In practice

Algorithm:

$$x_{k+1} \leftarrow x_k - \alpha_k \boldsymbol{d}, \qquad \boldsymbol{B_k d} = -\boldsymbol{g_k}(x_k), \qquad \boldsymbol{g_k}(x) = \frac{1}{s_k} \sum_{i \in \mathcal{S}_k} \nabla f_i(x)$$

Hessian approximations $\boldsymbol{B_k}$:

- (limited-memory) quasi-Newton: [Schraudolph et al '07]

$$s_k := x_{k+1} - x_k, \qquad y_k := \boldsymbol{g_k}(x_{k+1}) - \boldsymbol{g_k}(x_k)$$

- sample-average Hessian: [Byrd et al '11]

$$\boldsymbol{B_k} := \frac{1}{h_k} \sum_{i \in \mathcal{H}_k} \nabla^2 f_i(x_k), \quad h_k \ll s_k$$

In practice

Algorithm:

$$x_{k+1} \leftarrow x_k - \alpha_k \boldsymbol{d}, \qquad \boldsymbol{B_k d} = -\boldsymbol{g_k}(x_k), \qquad \boldsymbol{g_k}(x) = \frac{1}{s_k} \sum_{i \in \mathcal{S}_k} \nabla f_i(x)$$

Hessian approximations $\boldsymbol{B_k}$:

- (limited-memory) quasi-Newton: [Schraudolph et al '07]

$$s_k := x_{k+1} - x_k, \qquad y_k := \boldsymbol{g_k}(x_{k+1}) - \boldsymbol{g_k}(x_k)$$

- sample-average Hessian: [Byrd et al '11]

$$\boldsymbol{B_k} := \frac{1}{h_k} \sum_{i \in \mathcal{H}_k} \nabla^2 f_i(x_k), \quad h_k \ll s_k$$

- Fisher information [for $f(x) = -\sum_i \log p_i(\omega_i; x)$]: [Osborne '92]

$$\boldsymbol{B_k} \approx \mathbf{E}_\omega[\nabla^2 f(x)] = \mathbf{E}_\omega[\nabla f(x) \nabla f(x)^T]$$

Outstanding questions

- How to keep $s_k \ll M$ for all k? (And keep up the pace.)
- When to stop? (Without observing the full gradient.)

Thanks!

Read:

- M. P. Friedlander and G. Goh. Probabilisting bounds for minimizing sums of functions. *In preparation*, December 2012.
- M. P. Friedlander and M. Schmidt. Hybrid deterministic-stochastic methods for data fitting. *SIAM J. Scientific Computing*, 34(3), April 2012.
- A. Aravkin, M. P. Friedlander, F. Herrmann, and T. van Leeuwen. Robust inversion, dimensionality reduction, and randomized sampling. *Mathematical Programming*, 134(1):101–125, 2012.

Email:

- `mpf@cs.ubc.ca`

Surf:

- `http://www.cs.ubc.ca/~mpf`