Robust inversion and randomized sampling

Michael P. Friedlander
University of British Columbia

International Symposium on Mathematical Programming
August 19–24, 2012

Collaborators:
Sasha Aravkin, Felix Herrmann, Tristan van Leeuwen, and Mark Schmidt

Model problem

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

Examples:

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

Context:

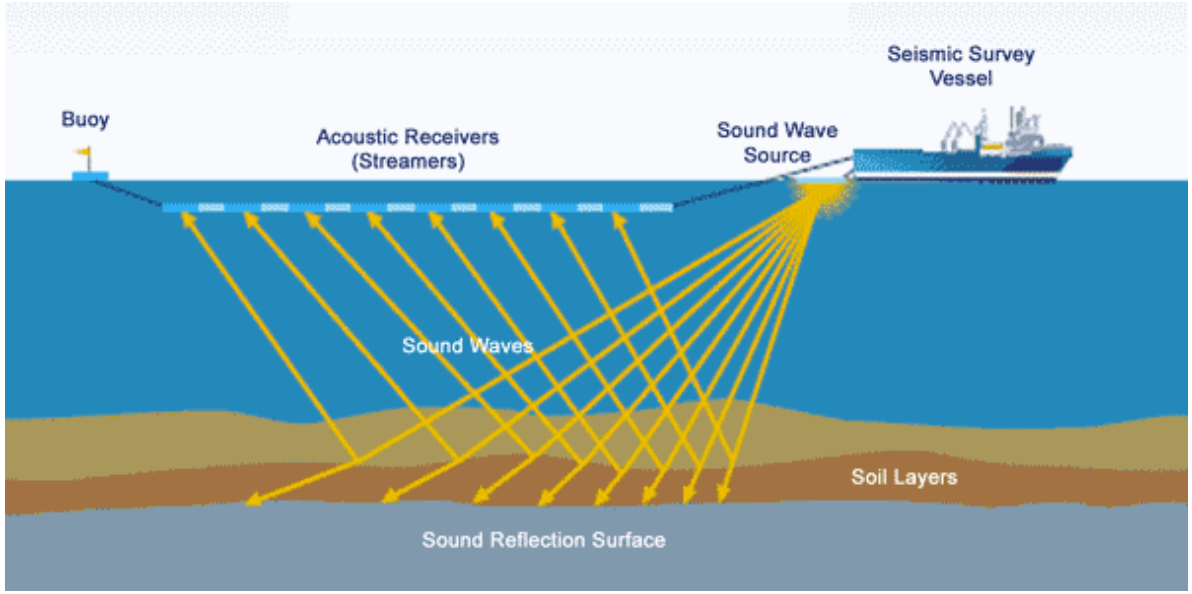
- m large
- each $f_i(x)$ and $\nabla f_i(x)$ expensive to evaluate

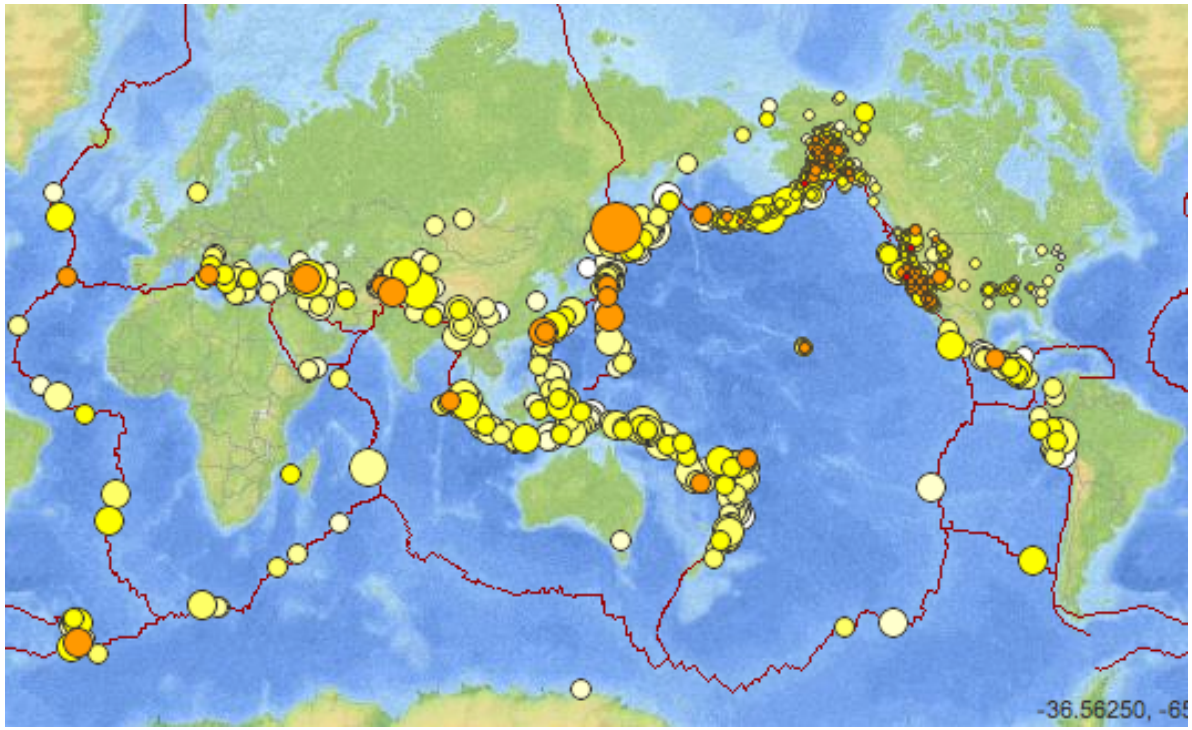
Computing costs:

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

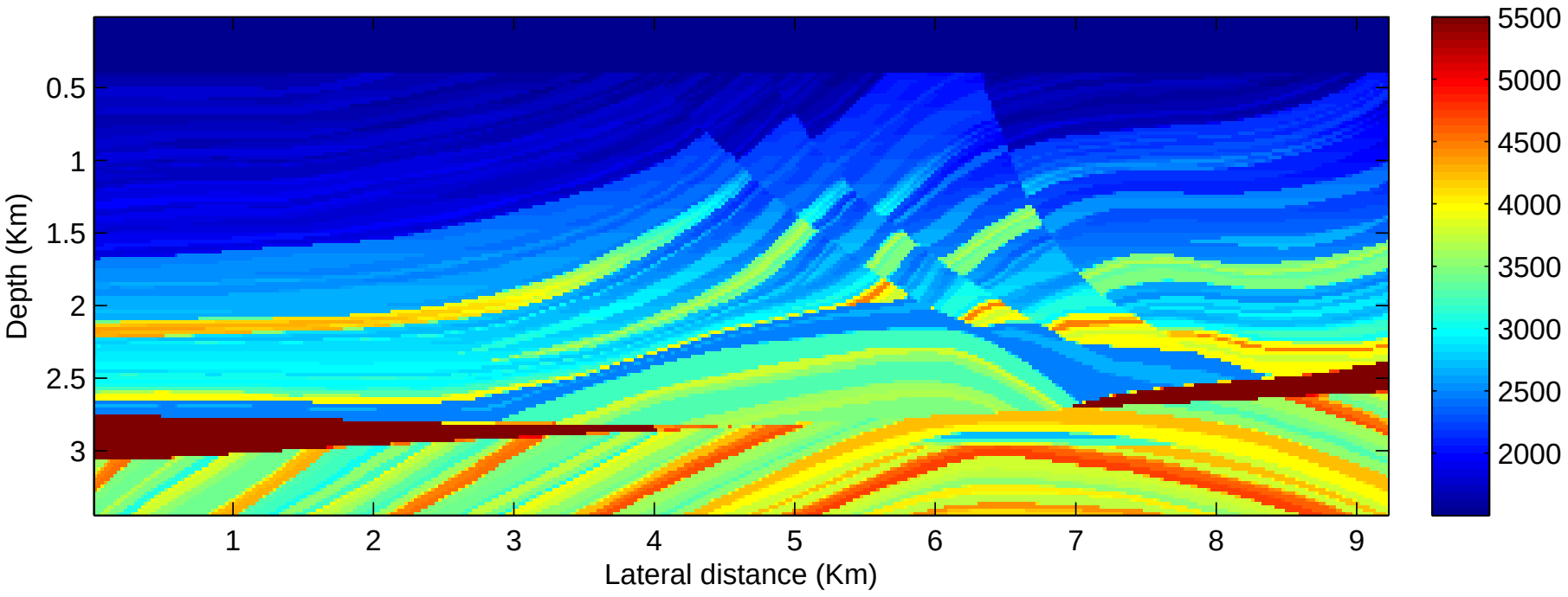
SEISMIC INVERSION

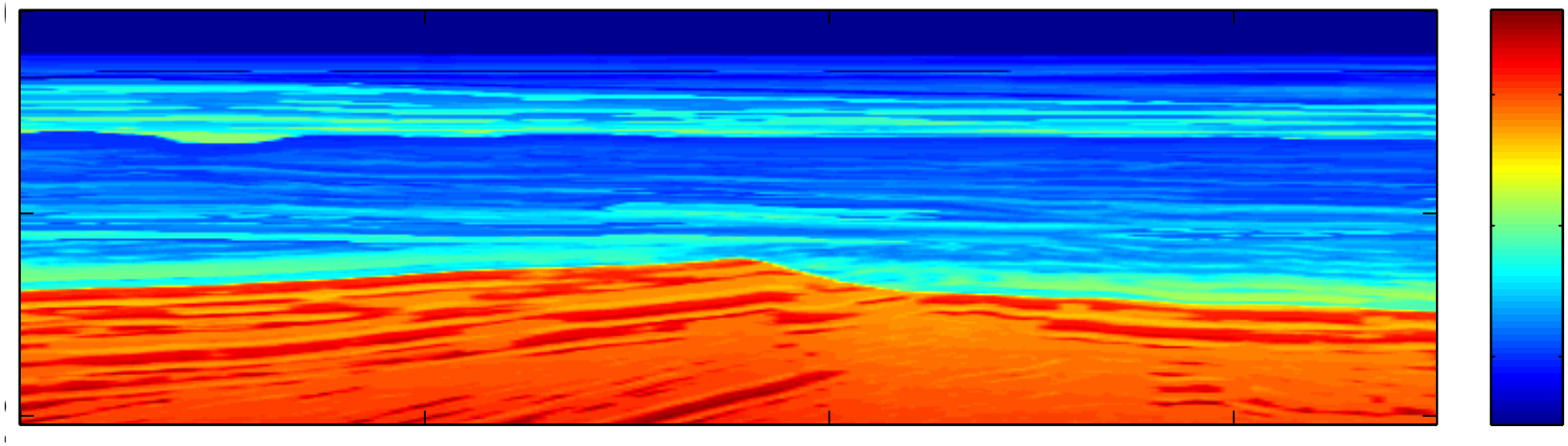
Reflection seismology





Full waveform inversion: velocity models



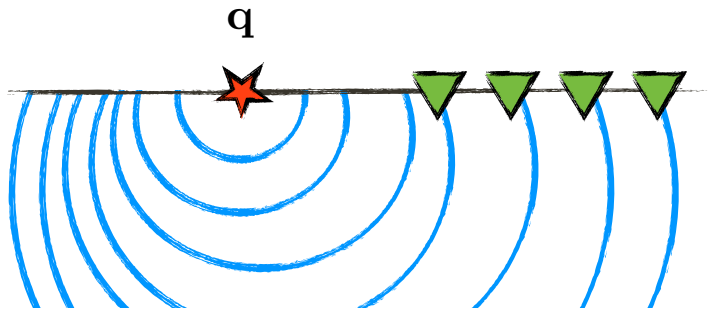


Full waveform inversion

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 and stochastic optimization

Use all your data:

$$f(x) = \sum_{i}^{m} \|r_i(x)\|^2 \qquad \text{with} \qquad r_i(x) = d_i - F(x)q_i$$

Dimensionality reduction: form s weighted averages ($s \ll m$)

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

Stochastic approximation of the misfit:

$$\bar{f}(x) = \sum_{j}^{s} \|\bar{r}_j(x)\|^2 \qquad \text{with} \qquad \bar{r}_j(x) = \bar{d}_j - F(x)\bar{q}_j$$

Stochastic optimization interpretation holds if $\mathbb{E}[WW^T] = I$:

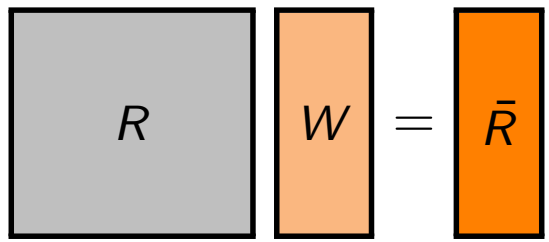
$$\mathbb{E}[\bar{f}(x)] = f(x) \quad \text{and} \quad \mathbb{E}[\nabla \bar{f}(x)] = \nabla f(x)$$

Stochastic trace estimation

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

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

Data mixing (dimensionality reduction) **equivalent** to



gives **sample-average** function

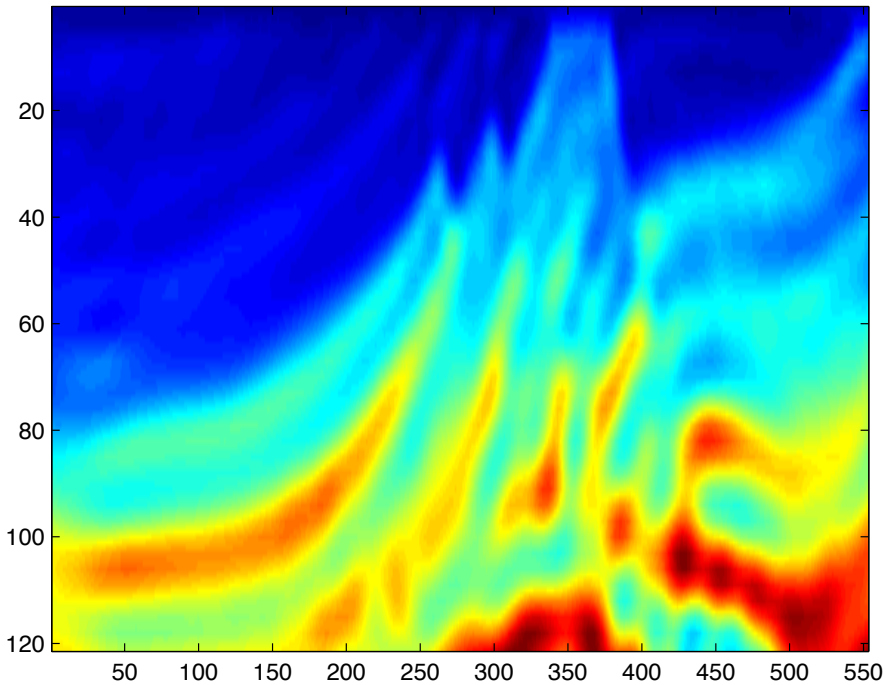
$$\bar{f}(x) = \sum_j^s \|\bar{r}_j\|^2 = \text{trace}(\bar{R}^T \bar{R}) \qquad \text{with} \qquad \bar{R} = [\, \bar{r}_1 \mid \cdots \mid \bar{r}_s \,], \quad s \ll m$$

Connected to **stochastic trace estimation.**

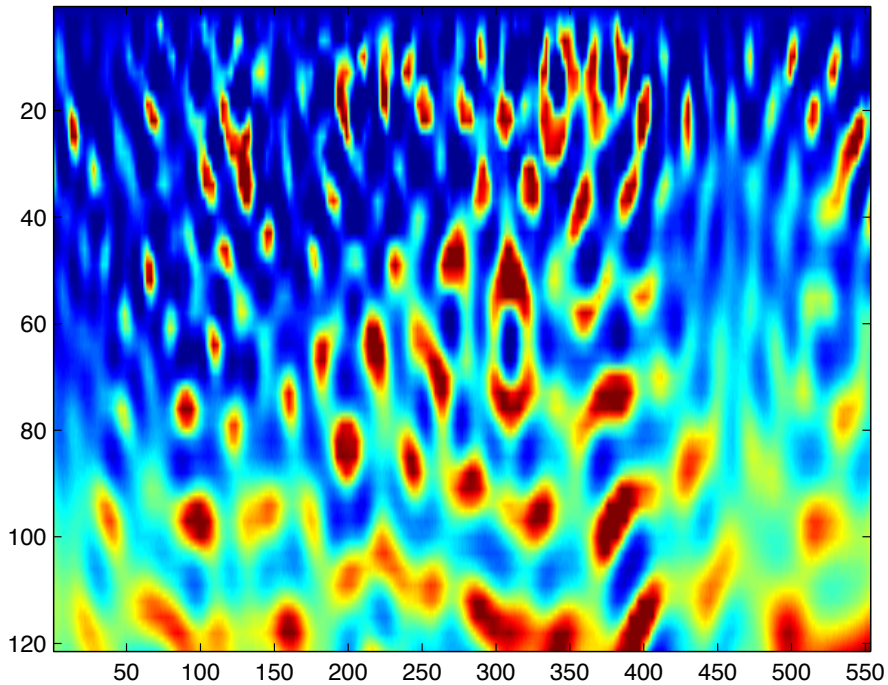
- Hutchinson ('90) — minimize var[trace($\bar{R}^T \bar{R}$)] with $\boldsymbol{W} \sim$ Rademacher
- Avron & Toledo ('11) — other optimal choices for $\boldsymbol{W}$
- Haber, Chung, Herrmann ('12) — connection to inverse problems

Nonlinear least-squares with corrupted data

good data



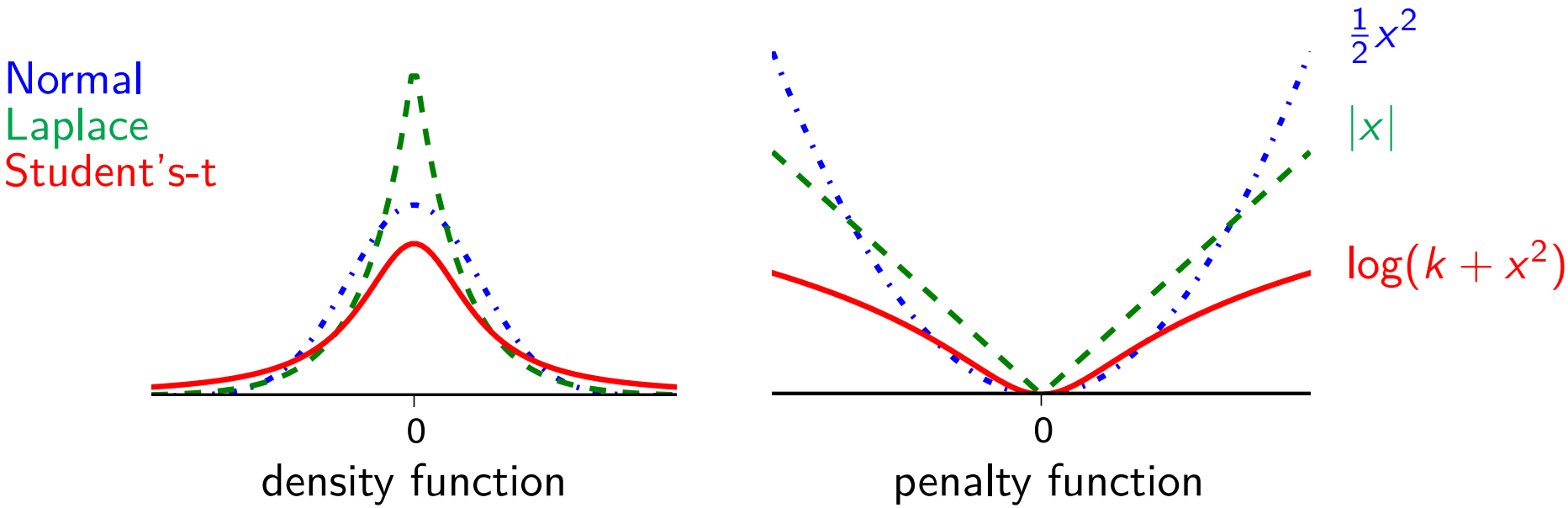
4% bad data



collaboration with Total

Robust misfit measures

$$d_i = F(x)q_i + \varepsilon$$



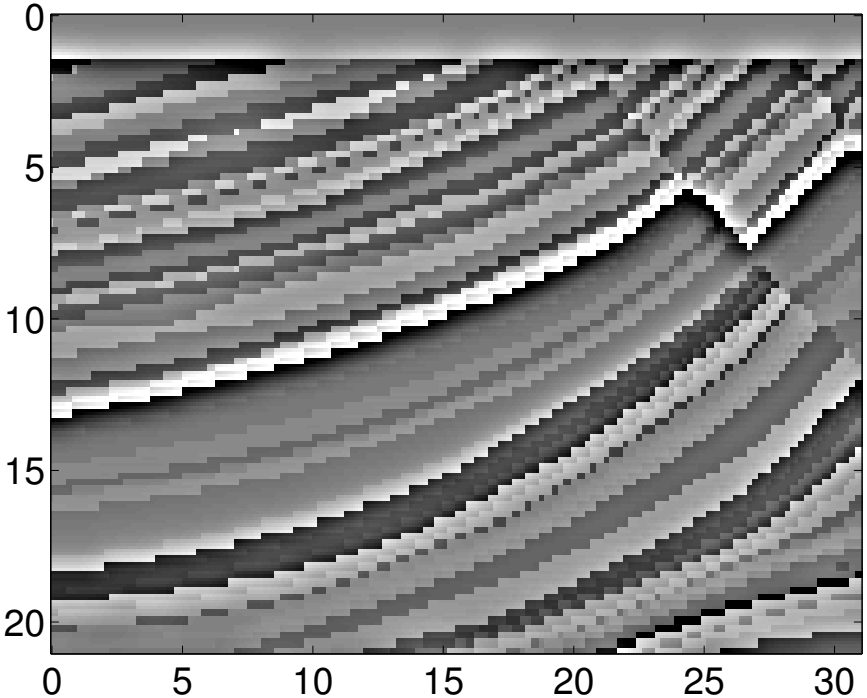
- **Least-squares** penalty assumes $\varepsilon \sim \mathcal{N}(0, 1)$
- **Theorem:** Log-concave densities—ie, those induced by convex penalties—are all **exponentially bounded**:

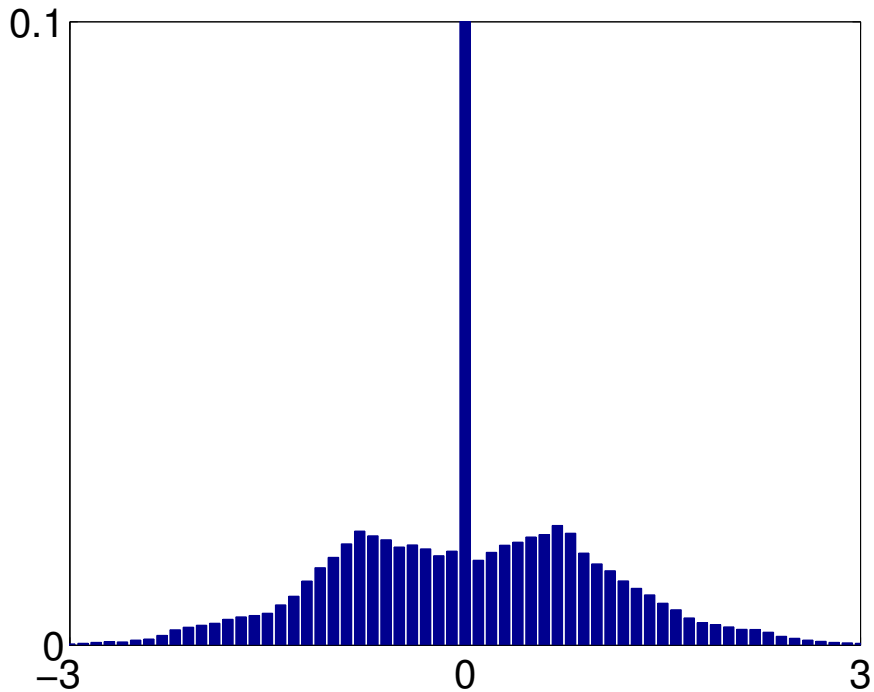
$$\textbf{prob}\big(|x| > t + \Delta t \quad \text{given} \quad |x| > t\big) = \mathcal{O}(e^{-\Delta t})$$

- **Heavy-tailed densities:** robust to wrong data & approximate models

Robust fullwaveform inversion results

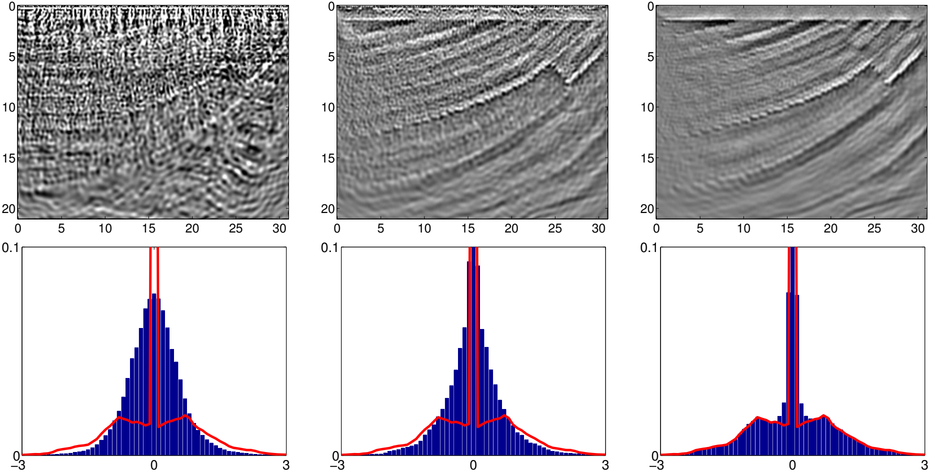
- Recover velocity on a 2D grid: 201×301
- 151 sources, 6 frequencies: 906 PDE solves per $f/\nabla f$-eval
- 50% corrupted data





$$\text{minimize } \rho(R(x)) \quad \text{with} \quad R(x) = D - F(x)Q$$

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



Sampling strategies for dimensionality reduction

Generic inverse problem (assuming iid observations):

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

Moment condition $E[WW^T] = I$ is **not generally sufficient** to guarantee

$$\begin{aligned} \mathbb{E}[\quad \bar{\boldsymbol{f}}(\boldsymbol{x})] &= \quad f(x) \\ \mathbb{E}[\nabla \bar{\boldsymbol{f}}(\boldsymbol{x})] &= \nabla f(x) \end{aligned} \qquad \text{with} \qquad \bar{\boldsymbol{R}}(\boldsymbol{x}) = R(x)W$$

Random subset selection without replacement, ie,

$$\bar{R}(x) = [\; r_{i(1)} \mid r_{i(2)} \mid \cdots \mid r_{i(s)} \;]$$

does yield desired **"expected objective"** property

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

Sublinear rate:

- $f(x_k) - f(x_*) = \mathcal{O}(1/k)$ (constant stepsize)
- $f(x_k) - f(x_*) = \mathcal{O}(1/k^2)$ (optimal rate with extrapolation)
  [Nesterov '83; Tseng '10]

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

- $f(x_k) - f(x_*) = \mathcal{O}([1 - \mu/L]^k)$

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** , **randomized**

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

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

- $\|x_k - x_*\|^2 \leq \mathcal{O}([1 - \mu/L]^k) + \mathcal{O}(\boldsymbol{m^2\bar{\alpha}})$ k **full cycles**
- $\mathbb{E}\|x_k - x_*\|^2 \leq \mathcal{O}([1 - \mu/L]^k) + \mathcal{O}(\boldsymbol{m\bar{\alpha}})$ k **iterations**

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

- $\|x_k - x_*\|^2 = \mathcal{O}(1/k)$ k **full cycles**
- $\mathbb{E}\|x_k - x_*\|^2 = \mathcal{O}(1/k)$ k **iterations**

Many variations:

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

EXAMPLES

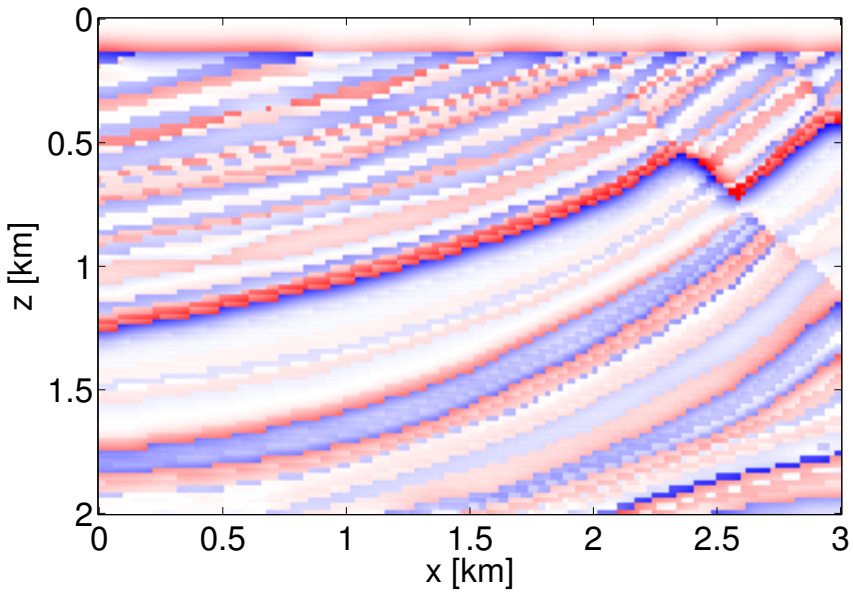
Seismic inversion

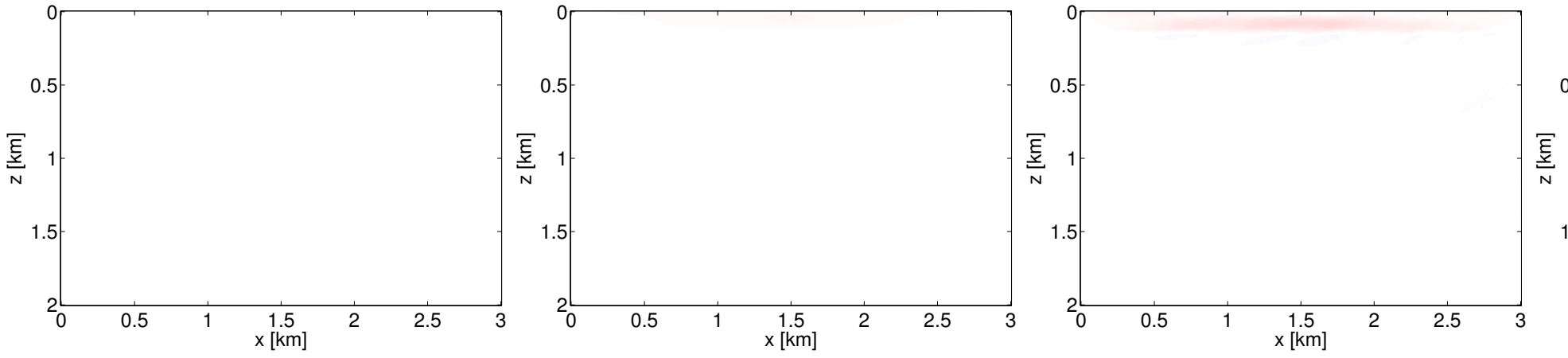
Recover image of geological structures via nonlinear least squares

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

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

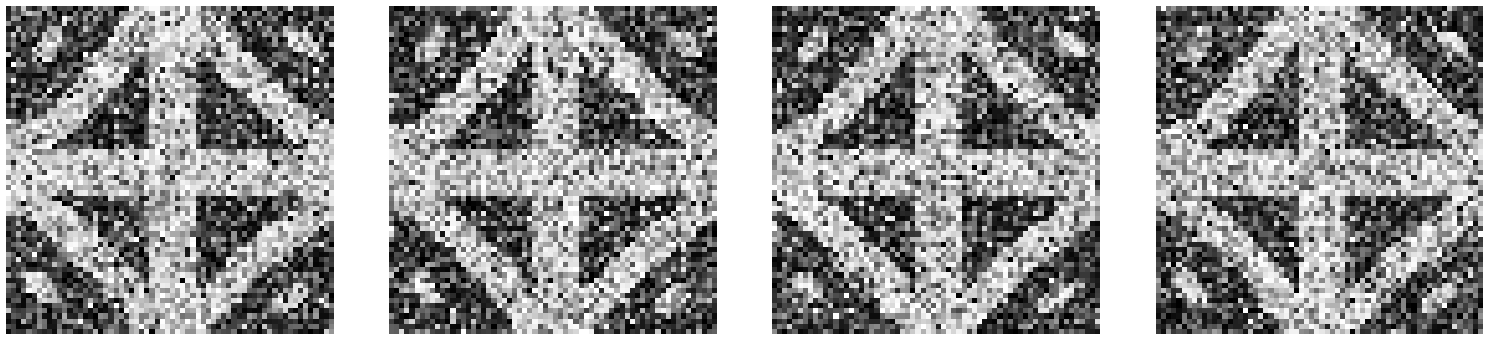
sources: $q_1, \ldots, q_m$, measurements: $d_1, \ldots, d_m$





0.010.40.822.6471016223039 of 39 passes

Image denoising

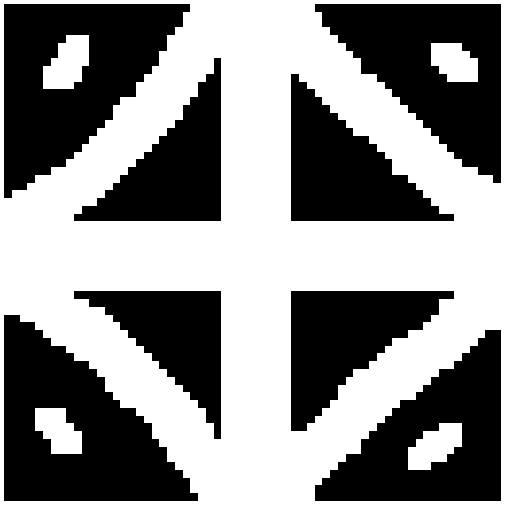


- Statistical denoising via conditional random fields
- Dataset of 50 synthetic 64×64 images [Kumar/Hebert '04]
- Generalization of logistic model to capture dependencies among labels

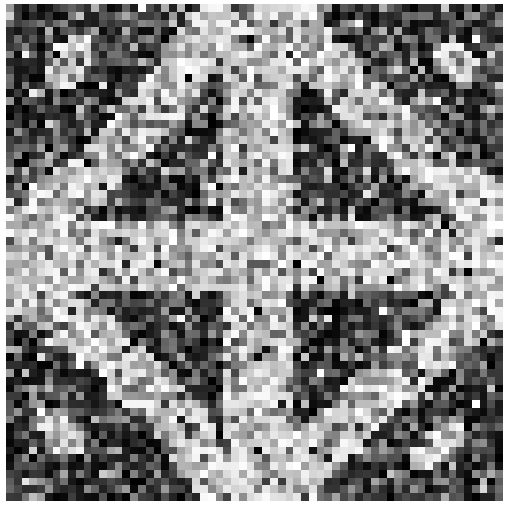
$$\underset{x}{\text{maximize}} \sum_{i=1}^{m} \log p(b_i; x)$$

- p is intractable and approximated

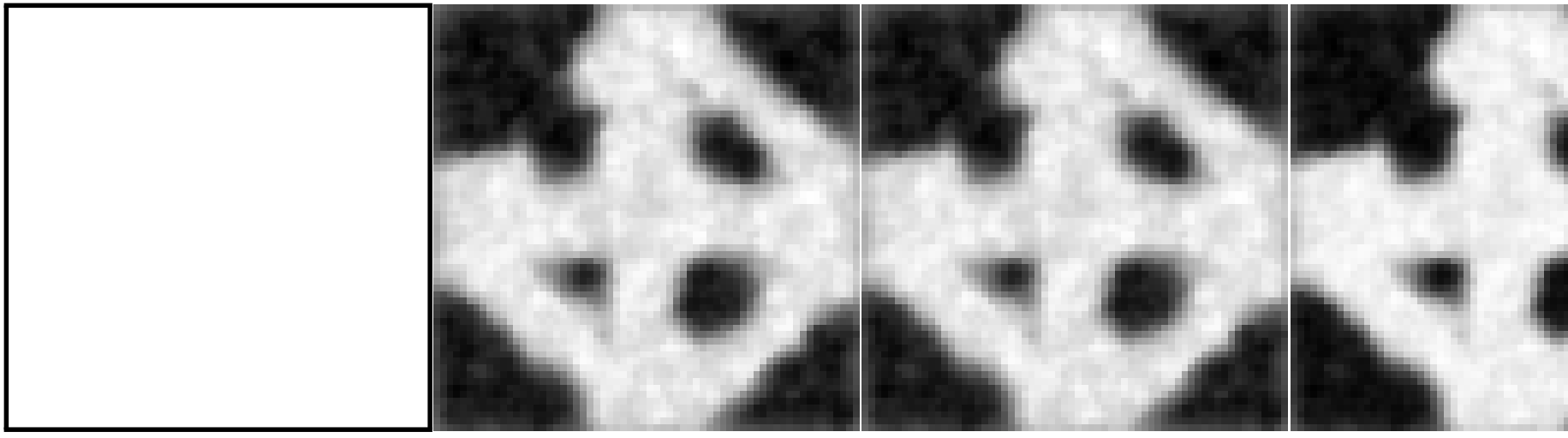
true image



noisy sample



full gradient



0.250.500.7512345 of 5 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

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

Gradient with generic errors

Prototype algo:

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

Assumptions:

- Lipschitz gradient (L); strong convexity (μ)
- $\|e_k\|^2 \leq \epsilon_k$

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

$$\|x_k - x_*\|^2 \leq \mathcal{O}([1 - \mu/L]^k) + \mathcal{O}(\epsilon_k)$$

Examples:

linear: $\epsilon_k = \mathcal{O}(\gamma^k) \implies f(x_k) - f_* = \mathcal{O}(\max\{\gamma,\ 1 - \mu/L\}^k)$

sublinear: $\epsilon_k = \mathcal{O}(1/k^2) \implies f(x_k) - f_* = \mathcal{O}(1/k^2)$

persistent err: $\inf_k \epsilon_k > 0 \implies f(x_k) - f_* = \mathcal{O}(\epsilon_k)$

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

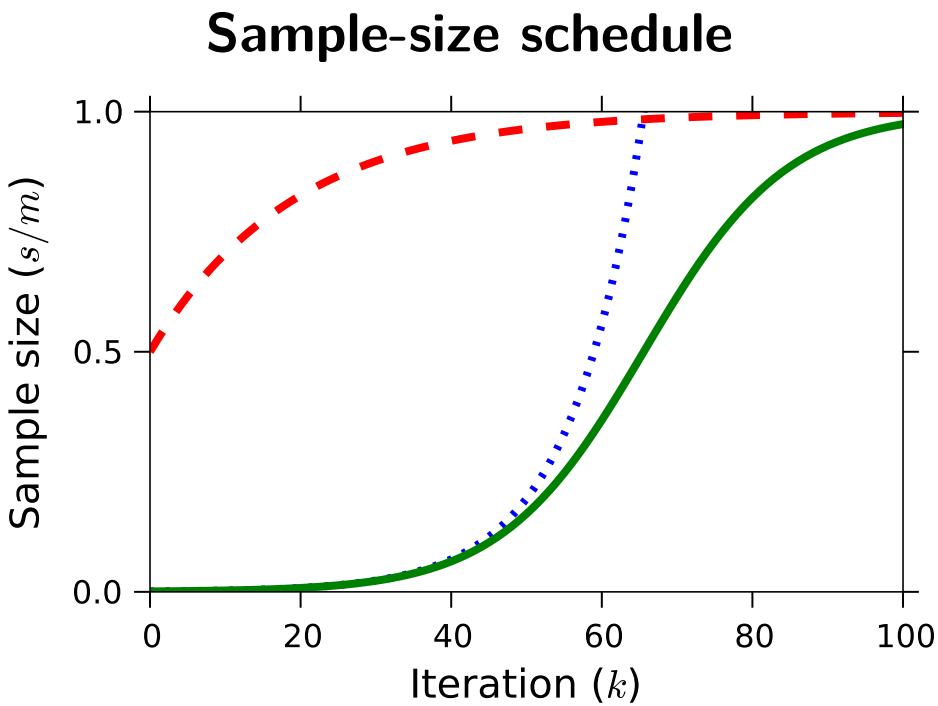
Convergence rates for all $k = 1, 2, \dots$ [F. & Schmidt '12]

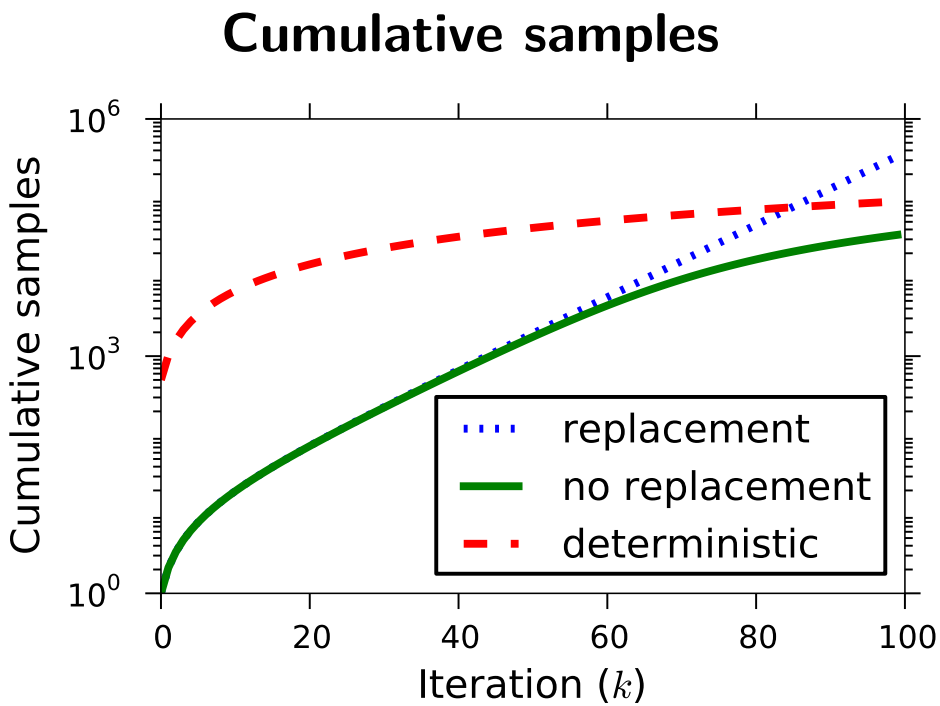
deterministic: $$\|x_k - x_*\|^2 = \mathcal{O}([1 - \mu/L]^k) + \mathcal{O}\left(\left[\frac{m - s_k}{m}\right]^2\right)$$

sampling w/o replacement $$\boldsymbol{E}\|x_k - x_*\|^2 = \mathcal{O}([1 - \mu/L]^k) + \mathcal{O}\left(\frac{m - s_k}{m} \cdot \frac{1}{s_k}\right)$$

sampling w/ replacement $$\boldsymbol{E}\|x_k - x_*\|^2 = \mathcal{O}([1 - \mu/L]^k) + \mathcal{O}\left(\frac{1}{s_k}\right)$$

Illustration





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]

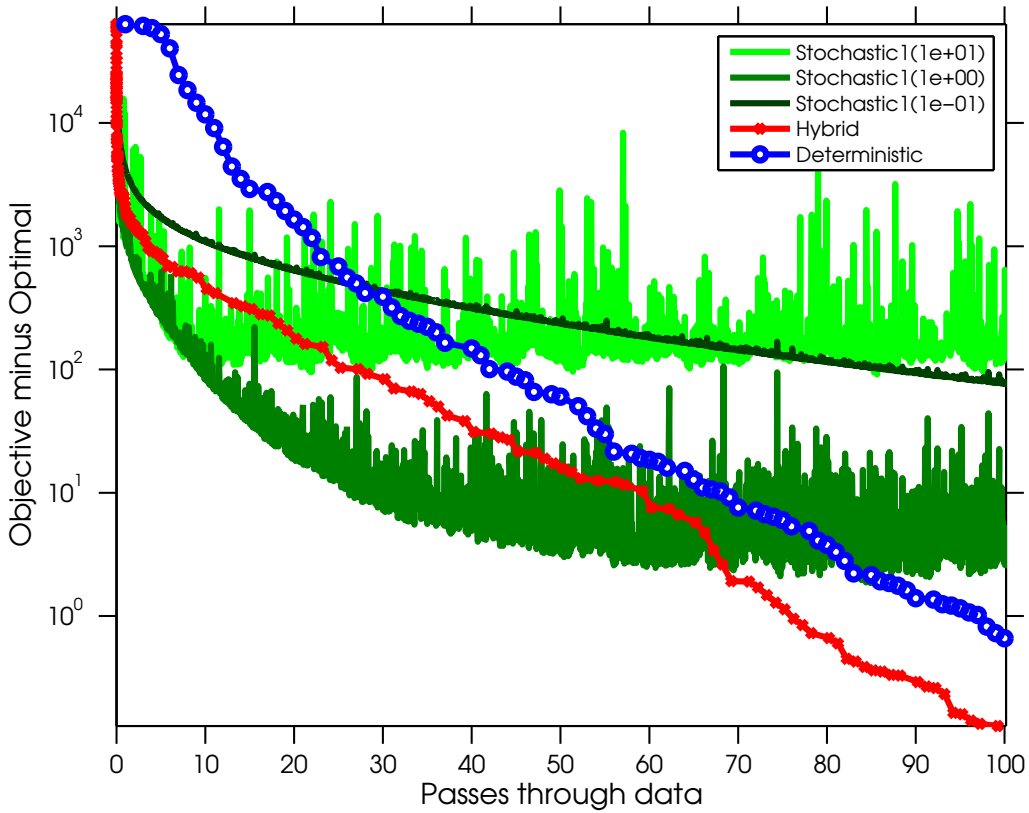
$$\boldsymbol{B_k} \approx \mathbb{E}_\omega[\nabla^2 f(x)] = \mathbb{E}_\omega[\nabla f(x) \nabla f(x)^T]$$

APPLICATIONS

Binary logistic regression

$$\max_x \sum_{i}^{m} \log p(b_i \mid a_i, x), \quad p(b_i \mid a_i, x) = \frac{1}{1 + \exp(-b_i a_i^T x)}, \quad b_i \in \{-1, 1\}$$

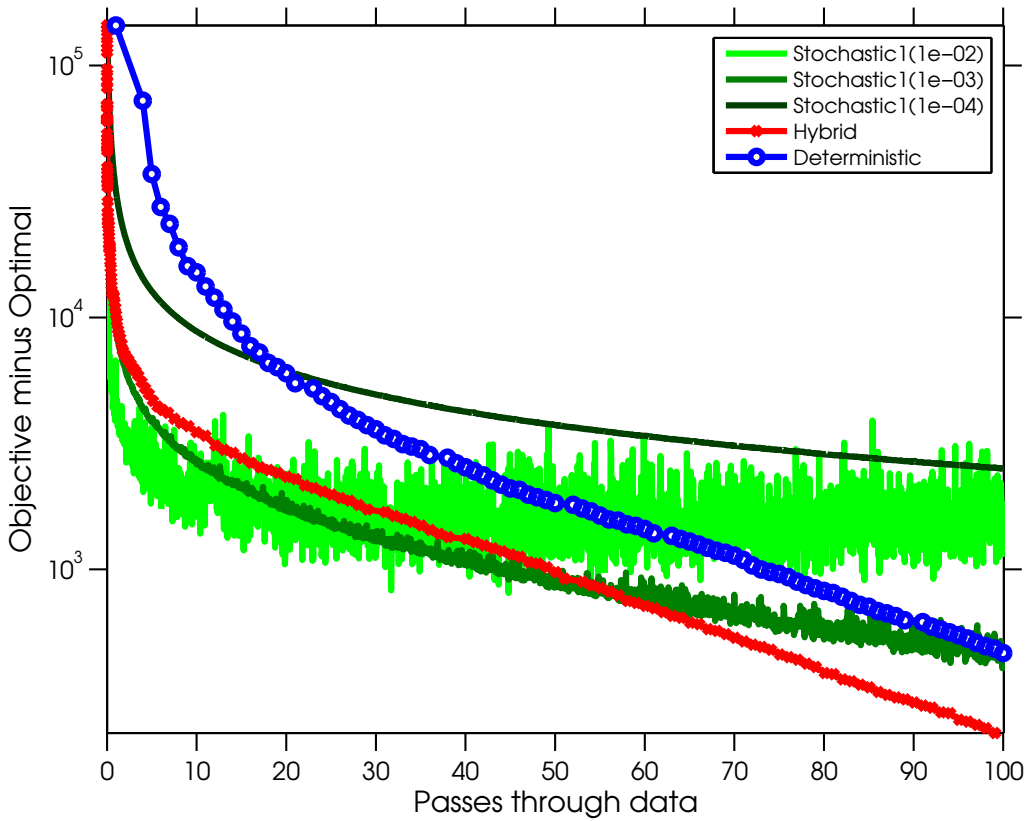
- Email spam classifier (Cormack and Lynam, 2005)
- TREC 2005 dataset: 92,189 email msgs from Enron investigation



Multinomial logistic regression

$$\max_x \sum_i^m \log p(b_i = j \mid a_i, \{x\}_{j \in \mathcal{C}}), \quad b_i \in \mathcal{C}$$

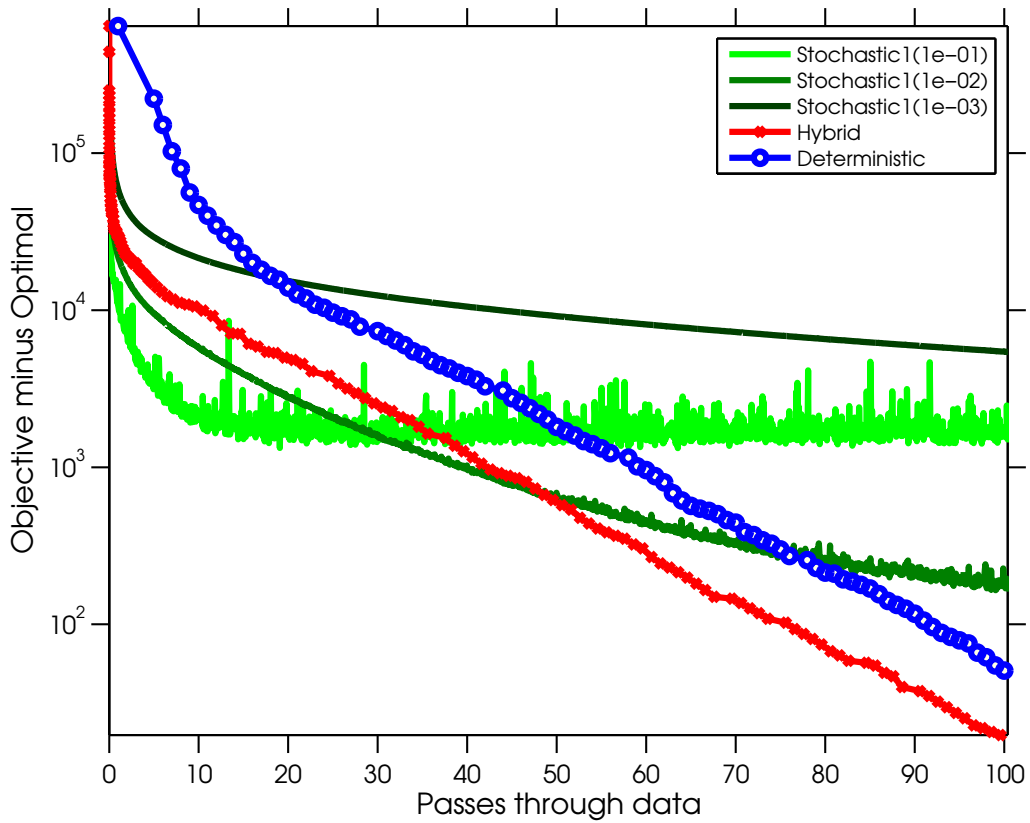
- Digit classification 
- MNIST dataset: 70,000 handwritten 28×28 images of digits



Chain-structured conditional random fields

$$\max_x \sum_i^m \log p(\{b_i^k = j_k\}_{k\in\Omega} \mid \{a_i^k\}_{k\in\Omega}, \{x_j\}_{j\in\mathcal{C}}), \quad b_i^k \in \mathcal{C},\ k \in \Omega$$

- noun-phrase chunking task from natural-language processing
- CoNLL-2000 Shared Task dataset: 211,727 words in 8,936 sentences

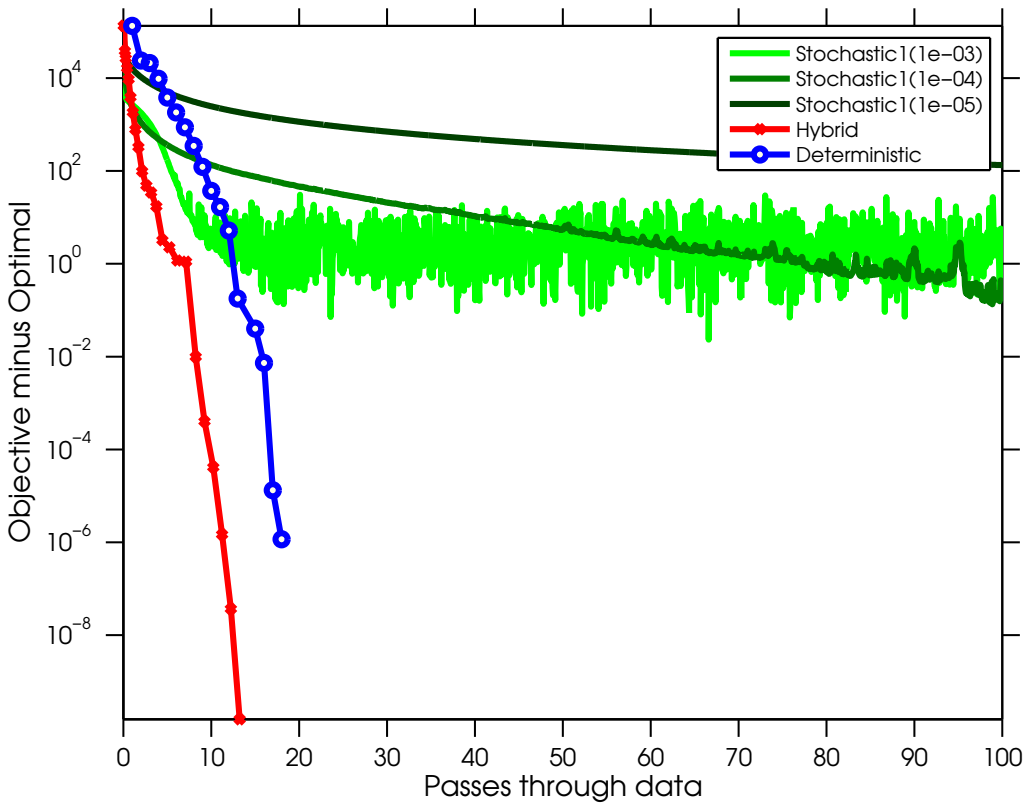


General conditional random fields

$$\min_x \sum_{i=1}^{m} p(b_i; x)$$

- Statistical denoising

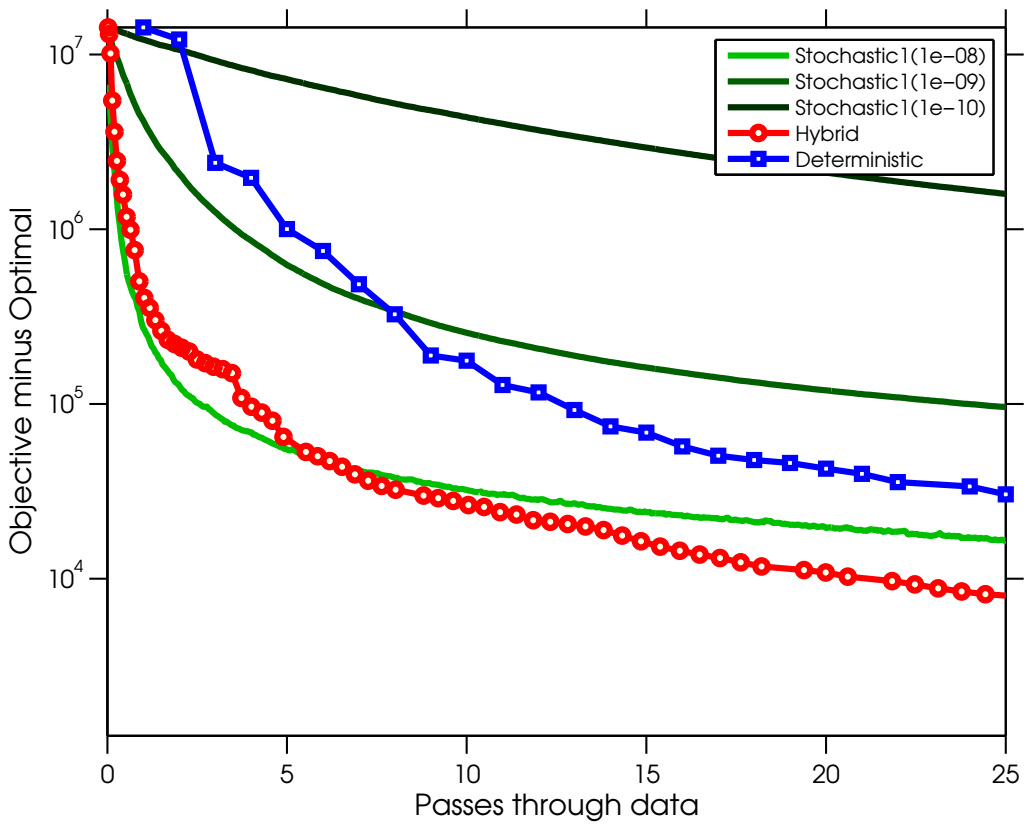
- Kumar/Hebert dataset of 50 synthetic 64×64 images



Seismic inversion

$$\min_x \sum_i^m \sum_{\omega \in \Omega} \|d_i - PH_\omega(x)^{-1} q_i\|^2$$

- Recover seismic image via nonlinear least squares
- Marmousi 2D acoustic model; 101 sources/receivers; 8 frequencies



No strong convexity, no linear convergence

Steepest descent

$$x_{k+1} \leftarrow x_k - \alpha \nabla f(x_k)$$

is sublinear:

$$f(x_k) - f(x_*) = \mathcal{O}(1/k)$$

Approximate steepest descent

$$x_{k+1} \leftarrow x_k - \alpha d_k \quad \text{with} \quad d_k = \nabla f(x_k) + e_k$$

Summable errors, ie, $\sum_k^\infty \|e_k\| < \infty$, gives sublinear iterate average:

$$f(\bar{x}_k) - f(x_*) = \mathcal{O}(1/k), \quad \bar{x}_k := \frac{1}{k} \sum_{i=1}^{k} x_i$$

References I

- H. Avron and S. Toledo. Randomized algorithms for estimating the trace of an implicit symmetric positive semi-definite matrix. *J. ACM*, 58:8:1–8:34, April 2011.
- D. Bertsekas. Incremental Gradient, Subgradient, and Proximal Methods for Convex Optimization: A Survey. In S. Sra, S. Nowozin, and S. J. Wright, editors, *Optimization for machine learning*, chapter 4, pages 85–115. MIT Press, 2012.
- D. Bertsekas and J. Tsitsiklis. *Neuro-dynamic programming*. Athena Scientific, 1996.
- D. Blatt, A. Hero, and H. Gauchman. A convergent incremental gradient method with a constant step size. *SIAM J. Optim.*, 18(1):29–51, 2008.
- L. Bottou. Large-scale machine learning with stochastic gradient descent. In Y. Lechevallier and G. Saporta, editors, *Proceedings of the 19th International Conference on Computational Statistics (COMPSTAT'2010)*, pages 177–187, Paris, France, August 2010. Springer.
- R. H. Byrd, G. Chin, J. Nocedal, and Y. Wu. Sample size selection in optimization methods for machine learning. *Math. Program.*, 134:127–155, 2012.
- R. H. Byrd, G. M. Chin, W. Neveitt, and J. Nocedal. On the use of stochastic hessian information in optimization methods for machine learning. *SIAM J. Optim.*, 21(3):977–995, 2011.
- L. Grippo. A class of unconstrained minimization methods for neural network training. *Optim. Methods Softw.*, 4(2):135–150, 1994.
- E. Haber, M. Chung, and F. J. Herrmann. An effective method for parameter estimation with PDE constraints with multiple right-hand sides. *SIAM J. Optim.*, 22(3):739–757, 2012.

References II

- M. Hutchinson. A stochastic estimator of the trace of the influence matrix for laplacian smoothing splines. *Comm. Statist. Simulation Comput.*, 19(2):433–450, 1990.
- A. Kleywegt, A. Shapiro, and T. Homem-de Mello. The sample average approximation method for stochastic discrete optimization. *SIAM J. Optim.*, 12(2):479–502, 2002.
- Z. Luo and P. Tseng. Error bounds and convergence analysis of feasible descent methods: A general approach. *Ann. Oper. Res.*, 46(1):157–178, 1993.
- A. Nedic and D. Bertsekas. Convergence rate of incremental subgradient algorithms. *Stochastic Optimization: Algorithms and Applications*, pages 263–304, 2000.
- Y. Nesterov. A method for unconstrained convex minimization problem with the rate of convergence $O(1/k^2)$. *Soviet Math. Dokl.*, 269, 1983.
- M. Osborne. Fisher's method of scoring. *Intern. Stat. Rev.*, pages 99–117, 1992.
- N. N. Schraudolph, J. Yu, and S. Günter. A Stochastic Quasi-Newton Method for Online Convex Optimization. In M. Meila and X. Shen, editors, *Proc. 11th Intl. Conf. Artificial Intelligence and Statistics (AIStats)*, volume 2 of *Workshop Conf. Proc.*, pages 436–443, San Juan, Puerto Rico, 2007. J. Machine Learning Res.
- P. Tseng. Approximation accuracy, gradient methods, and error bound for structured convex optimization. *Math. Program.*, 125:263–295, 2010.
- K. van den Doel and U. M. Ascher. Adaptive and stochastic algorithms for electrical impedance tomography and dc resistivity problems with piecewise constant solutions and many measurements. *SIAM J. Comput.*, 34(1), 2012.

Thanks!

Read:

- 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`