Robust inversion, data-fitting, and inexact gradient methods

Michael P. Friedlander
Computer Science
University of British Columbia

SINBAD Sponsor Meeting
December 5, 2011

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

Model problem

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

Examples:

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

Model problem

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

Examples:

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

Context:

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

Model problem

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

Examples:

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

Context:

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

Computing costs:

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

Model problem

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

Examples:

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

Context:

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

Computing costs:

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

FULL WAVEFORM INVERSION

Experiments: Each of m "shots" yields a vector of measurements:

$$\text{sources: } q_1, \ldots, q_m, \qquad \text{measurements: } d_1, \ldots, d_m$$

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, 10k 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

Generic inverse problem:

$$D = F(x)Q + \mathcal{E}$$

$$D = \left[d_1, \ldots, d_m\right]$$

$$Q = \left[q_1, \ldots, q_m\right]$$

Dimensionality reduction and stochastic optimization

Generic inverse problem:

$$D = F(x)Q + \mathcal{E} \qquad\qquad D = \left[d_1, \ldots, d_m\right], \quad Q = \left[q_1, \ldots, q_m\right]$$

Nonlinear least-squares formulation:

$$\min_x \; f(x) := \frac{1}{m}\|R(x)\|_F^2 \qquad \text{with} \qquad R(x) = D - F(x)Q$$

Dimensionality reduction and stochastic optimization

Generic inverse problem:

$$D = F(x)Q + \mathcal{E} \qquad\qquad D = \left[d_1, \ldots, d_m\right], \quad Q = \left[q_1, \ldots, q_m\right]$$

Nonlinear least-squares formulation:

$$\min_x \; f(x) := \frac{1}{m}\|R(x)\|_F^2 \qquad \text{with} \qquad R(x) = D - F(x)Q$$

Reduction via averaging: generate small number of avgs ($s \ll m$)

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

Dimensionally reduced misfit: $f_w(x) = \dfrac{1}{s}\|\widetilde{R}(x)\|_f^2$

Dimensionality reduction and stochastic optimization

Generic inverse problem:

$$D = F(x)Q + \mathcal{E} \qquad\qquad D = \left[d_1, \ldots, d_m\right] \qquad Q = \left[q_1, \ldots, q_m\right]$$

Nonlinear least-squares formulation:

$$\min_x \; f(x) := \frac{1}{m}\|R(x)\|_F^2 \qquad \text{with} \qquad R(x) = D - F(x)Q$$

Reduction via averaging: generate small number of avgs ($s \ll m$)

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

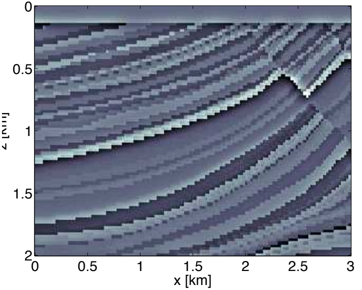
Dimensionally reduced misfit: $f_w(x) = \dfrac{1}{s}\|\widetilde{R}(x)\|_f^2$

Stochastic optimization interpretation: if weights w_{ij} are iid,

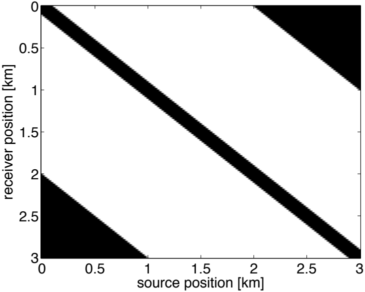
$$E[f_w(x)] = f(x) \quad \text{and} \quad E[\nabla f_w(x)] = \nabla f(x)$$

Nonlinear LS with missing data (partial Marmoussi)

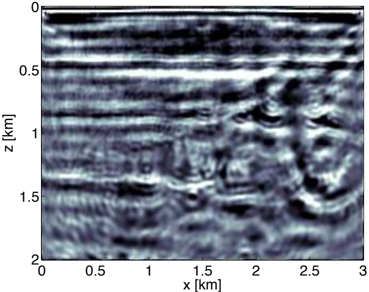
true reflectivity



marine acquisition mask



recovery via nonlinear least squares

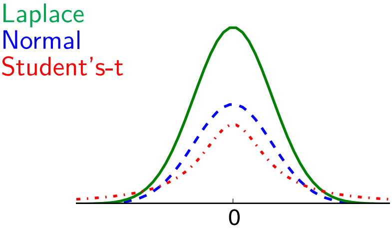


Robust misfit measures

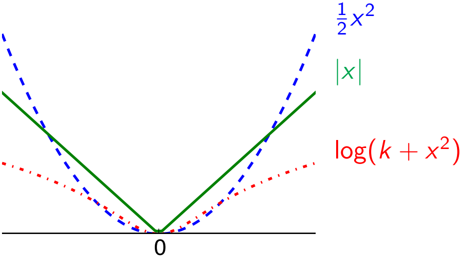
$$D = F(x)Q + \mathcal{E} \qquad \text{with} \qquad \mathcal{E} \sim \text{heavy-tailed dist'n}$$

Robust error model to capture

- missing data
- artifacts not captured by forward model



density function

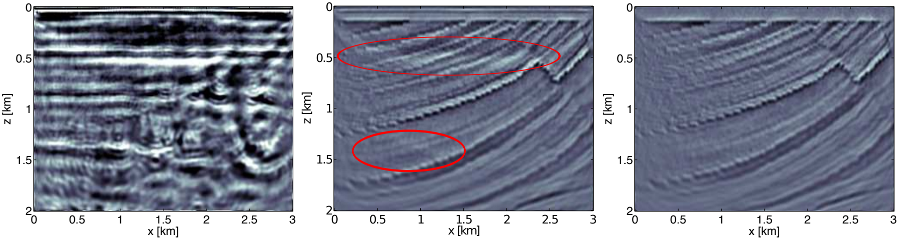


penalty function

Robust full-waveform inversion

minimize misfit $R(x) = D - F(x)Q$

$\|R(x)\|_F^2$ $\qquad$ $\sum_{ij} \|R_{ij}(x)\|_1$ $\qquad$ $\sum_{ij} \log\left(k + R_{ij}(x)^2\right)$



Normal $\qquad$ Laplace $\qquad$ Student's-t

Sampling strategies for dimensionality reduction

Generic inverse problem

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

Data averaging is generally **not** sufficient to guarantee that

$$E_w[f_w(x)] = f(x) \quad \text{and} \quad E_w[\nabla f_w(x)] = \nabla f(x)$$

However, **random subset selection**

$$\widetilde{R}(x) = \frac{1}{s}[\; r_{i(1)}, r_{i(2)}, \ldots, r_{i(s)} \;]$$

yields desired **“expected objective”** property for dimensionally reduced misfit f_w.

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: 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 '08]

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**

Constant stepsize: $\alpha_k \equiv 1/L$

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

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

- $\|x_k - x_*\|^2 = \mathcal{O}(1/k)$ $\quad$ 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**

Constant stepsize: $\alpha_k \equiv 1/L$, $\boldsymbol{\alpha_k \equiv m/L}$

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

EXAMPLES

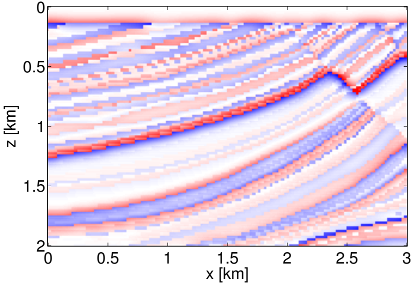
Seismic inversion

Recover image of geological structures via nonlinear least squares

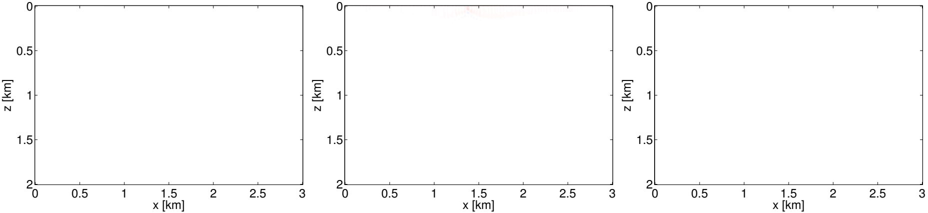
$$\underset{x}{\text{minimize}} \;\sum_{i}^{m} \sum_{\omega \in \Omega} \|d_i - PH_\omega(x)^{-1} q_i\|^2$$

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

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

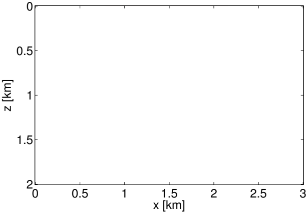


full gradient   incremental gradient   batched sampling

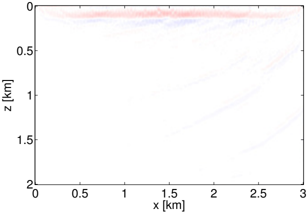


0.01 of 39 passes

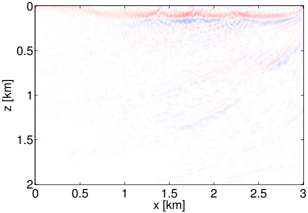
full gradient



incremental gradient

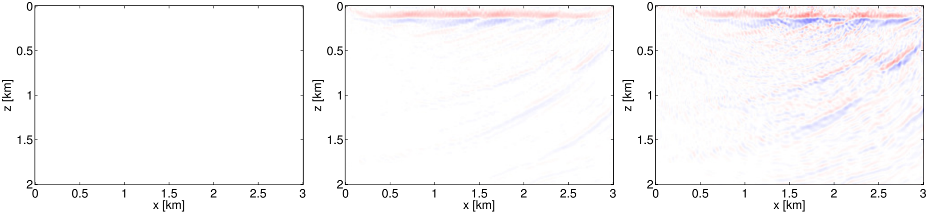


batched sampling



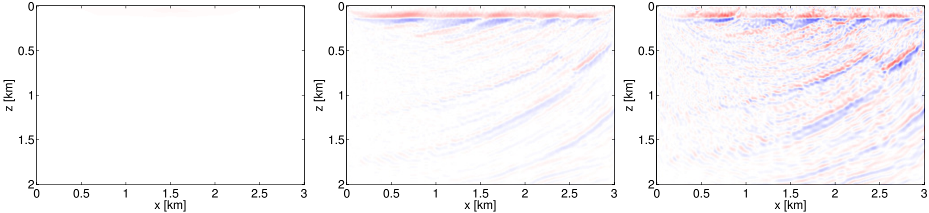
0.4 of 39 passes

full gradient | incremental gradient | batched sampling



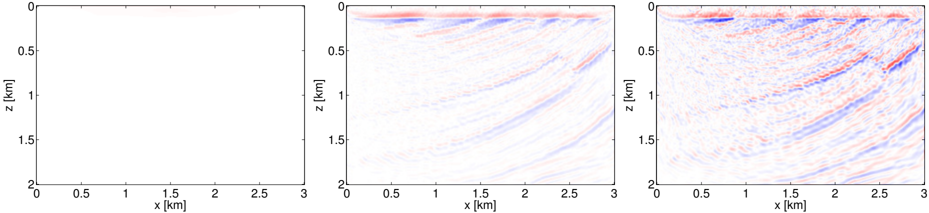
0.8 of 39 passes

full gradient | incremental gradient | batched sampling



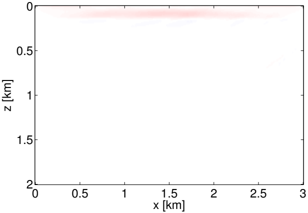
2 of 39 passes

full gradient | incremental gradient | batched sampling

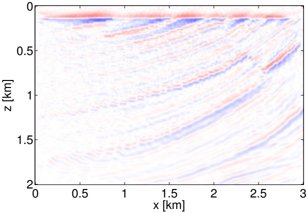


2.6 of 39 passes

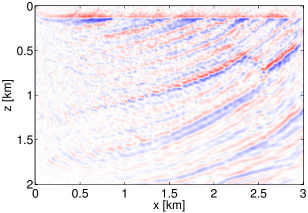
full gradient



incremental gradient

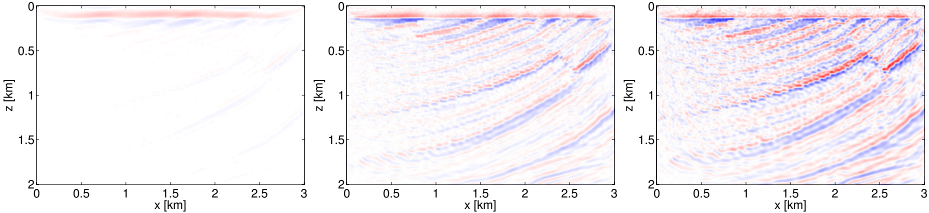


batched sampling



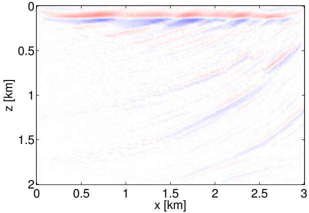
4 of 39 passes

full gradient | incremental gradient | batched sampling

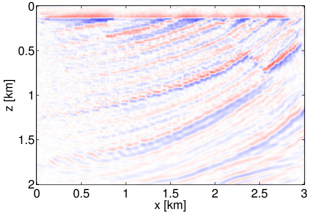


7 of 39 passes

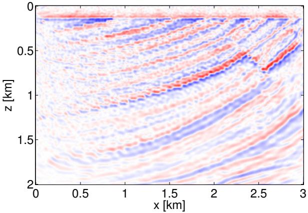
full gradient



incremental gradient

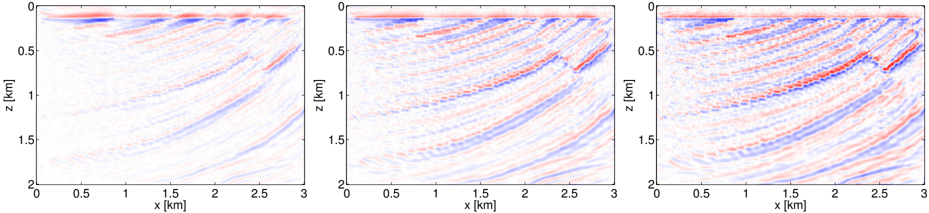


batched sampling



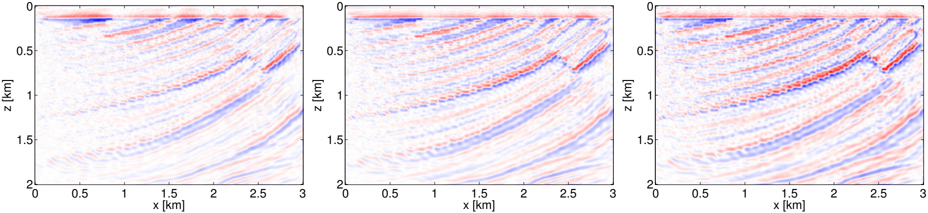
10 of 39 passes

full gradient | incremental gradient | batched sampling

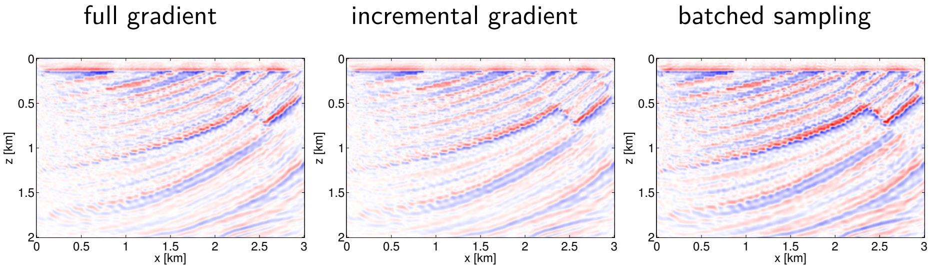


16 of 39 passes

full gradient | incremental gradient | batched sampling

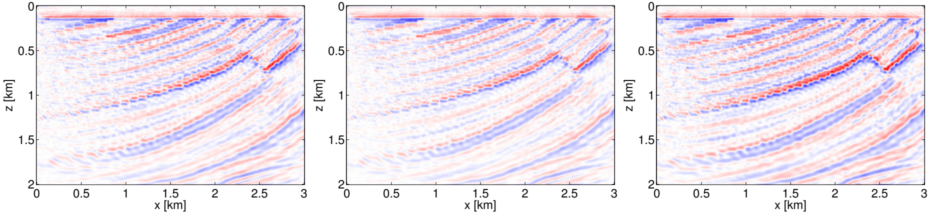


22 of 39 passes



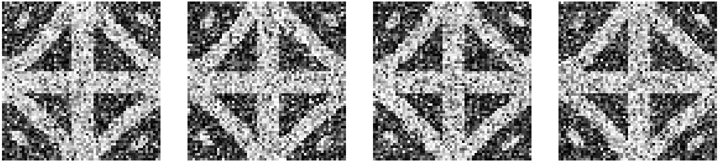
30 of 39 passes

full gradient | incremental gradient | batched sampling



39 of 39 passes

Image denoising

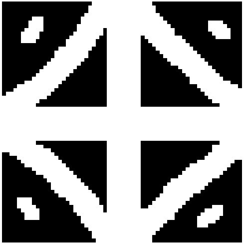


- Statistical denoising via conditional random fields
- Kumar/Hebert ('04) dataset of 50 synthetic 64×64 images
- Generalization of logistic model to capture dependencies among labels

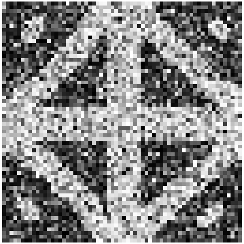
$$\min_x \sum_{i=1}^{m} p(b_i; x)$$

- p is intractable and approximated

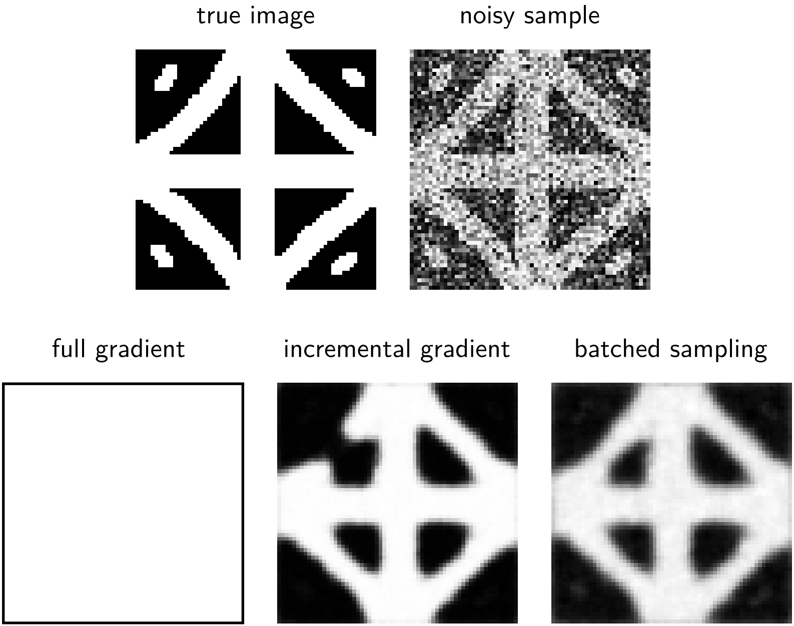
true image



noisy sample

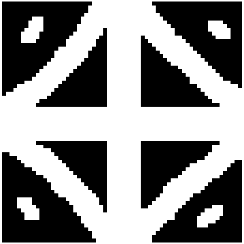


full gradient   incremental gradient   batched sampling

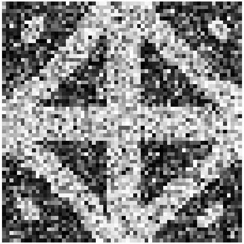


0.25 of 5 passes

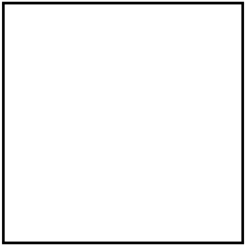
true image



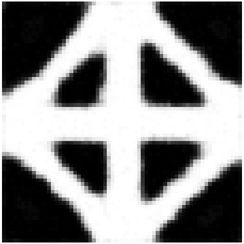
noisy sample



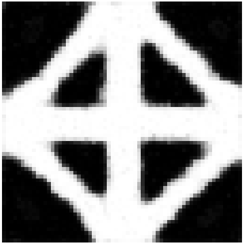
full gradient



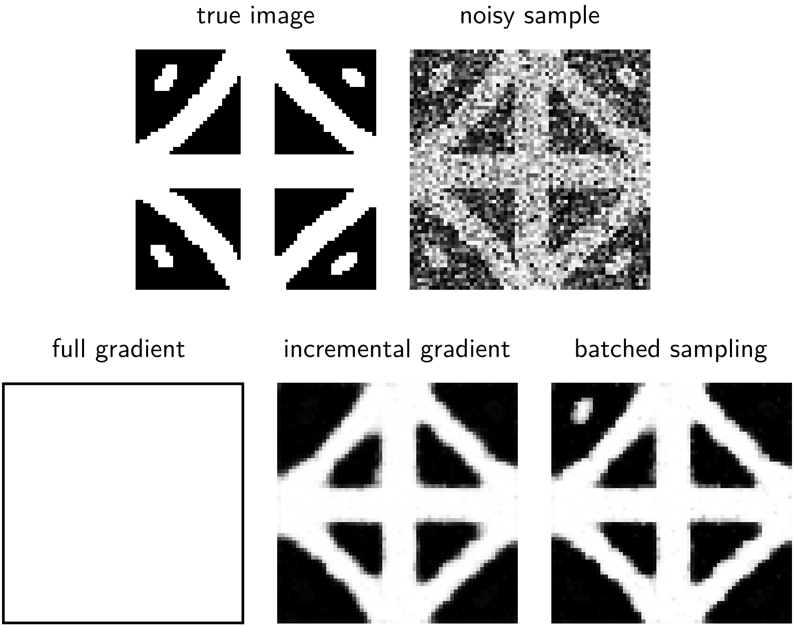
incremental gradient



batched sampling

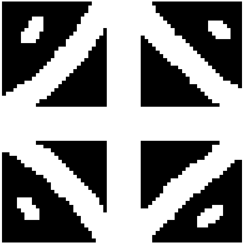


0.50 of 5 passes

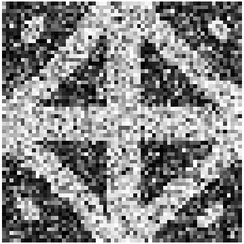


0.75 of 5 passes

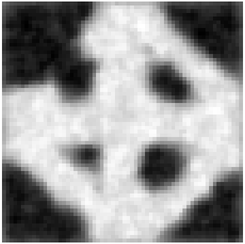
true image



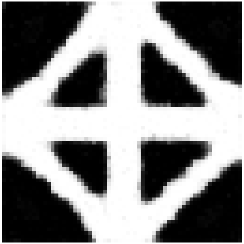
noisy sample



full gradient



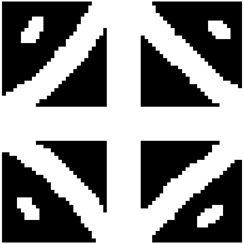
incremental gradient



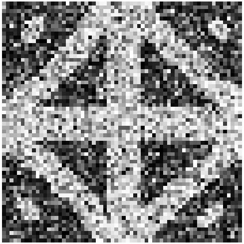
batched sampling

1 of 5 passes

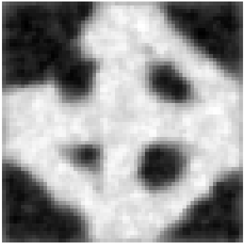
true image



noisy sample



full gradient

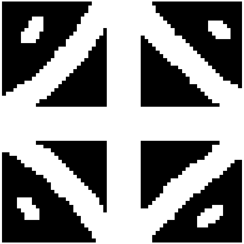


incremental gradient

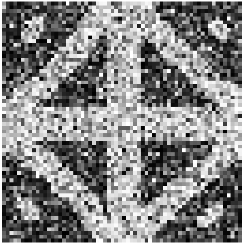
batched sampling

2 of 5 passes

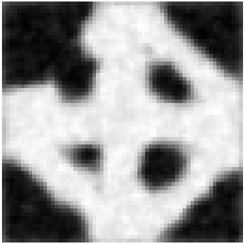
true image



noisy sample

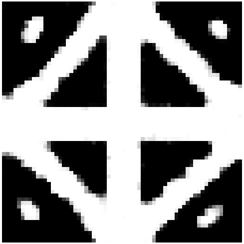


full gradient



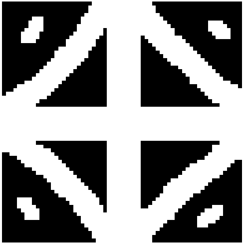
incremental gradient

batched sampling

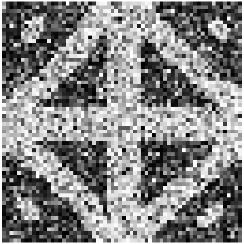


3 of 5 passes

true image



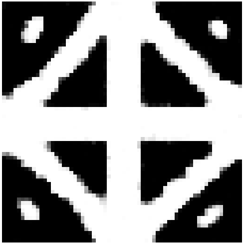
noisy sample



full gradient

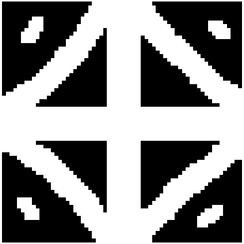
incremental gradient

batched sampling

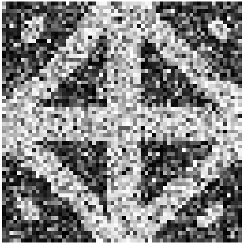


4 of 5 passes

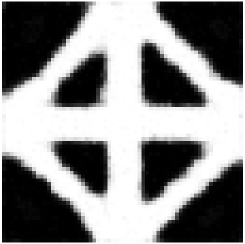
true image



noisy sample

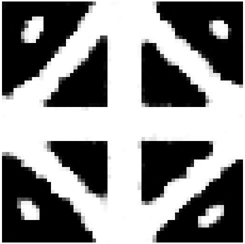


full gradient



incremental gradient

batched sampling



5 of 5 passes

ALGORITHM

Sampling approach

Increasing batch:

$$\mathcal{B}_k \subseteq \{1, \ldots, m\}, \quad |\mathcal{B}_k| \to m$$

Sample gradient:

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

Algorithm:

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

Analysis: Based on controlling gradient error e_k:

$$g_k(x) = \nabla f(x_k) + e, \qquad \|e_k\|^2 \le \epsilon_k \quad \text{or} \quad \mathbb{E}[\|e_k\|^2] \le \epsilon_k$$

Gradient with errors

Prototype algo: $x_{k+1} \leftarrow x_k - \alpha g_k, \quad g_k = \nabla f(x_k) + e_k, \quad \alpha \equiv 1/L$

Given

- Lipschitz gradient (L); strong convexity (μ)
- $\|e_k\|^2 \leq \epsilon_k$
- $\lim_{k\to\infty} \epsilon_{k+1}/\epsilon_k \leq 1$

Convergence for all $k = 1, 2, \ldots$

$$\|x_k - x_*\|^2 \leq (1 - \mu/L)^k [f(x_0) - f(x_*)] + \mathcal{O}(\epsilon_k)$$

Growing batch size

Prototype algo:

$$x_{k+1} \leftarrow x_k - \alpha g_k, \qquad g_k = \frac{1}{s} \sum_{i \in \mathcal{B}_k} \nabla f_i(x_k), \qquad \alpha \equiv 1/L$$

Batch strategy:

1. **Deterministic**: pre-determined batch sequence
2. **Randomized**: uniform sampling without replacement

Unsampled fraction of the population:

$$\rho_k := \frac{m - s}{m}$$

Convergence: for all $k = 1, 2, \ldots$

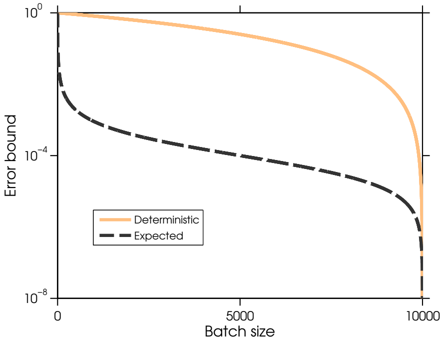
$$\begin{aligned} \|\mathbf{x_k} - \mathbf{x}_*\|^2 &= \mathcal{O}([1 - \mu/L]^k) + \mathcal{O}(\rho_k^2) \\ \boldsymbol{E}[\|\boldsymbol{x_k} - \boldsymbol{x}_*\|^2] &= \mathcal{O}([1 - \mu/L]^k) + \mathcal{O}(\rho_k/s) \end{aligned}$$

Randomization is key

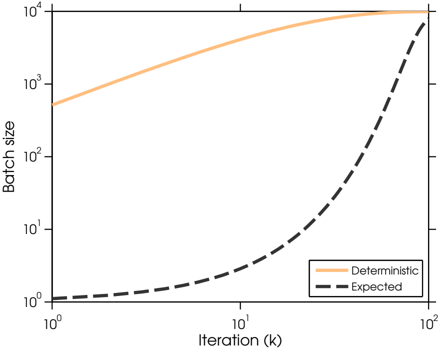
Sample average:

$$g_k(x) = \frac{1}{s} \sum_{i \in \mathcal{B}_k} \nabla f_i(x)$$

Gradient error



Batching schedule



Batching algorithm in practice

Sample approximation:

$$\bar{f}_k(x) = \frac{1}{s} \sum_{i \in \mathcal{B}_k} f_i(x), \qquad g_k(x) = \frac{1}{s} \sum_{i \in \mathcal{B}_k} \nabla f_i(x)$$

Algorithm:

$$x_{k+1} \leftarrow x_k - \alpha_k d_k, \qquad H_k d = -g_k(x_k)$$

Quasi-Newton Hessian using

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

Linesearch on sample function

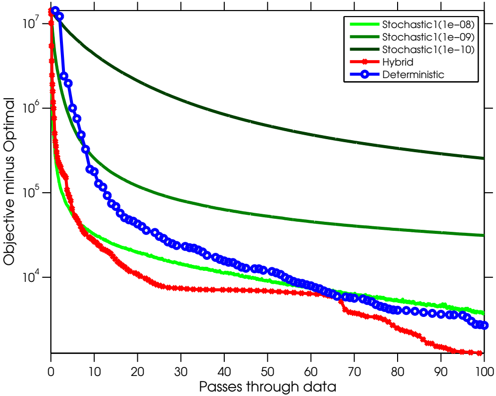
$$\bar{f}(x_k + \alpha d_k) < \bar{f}(x_k)$$

APPLICATIONS

Seismic inversion

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

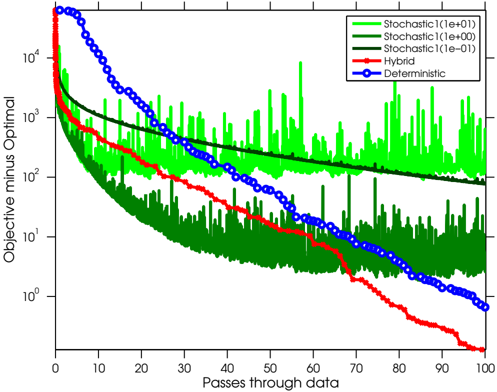
- Recover seismic image via nonlinear least squares
- Marmousi 2D acoustic model; 101 sources/receivers; 8 frequencies



Binary logistic regression

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

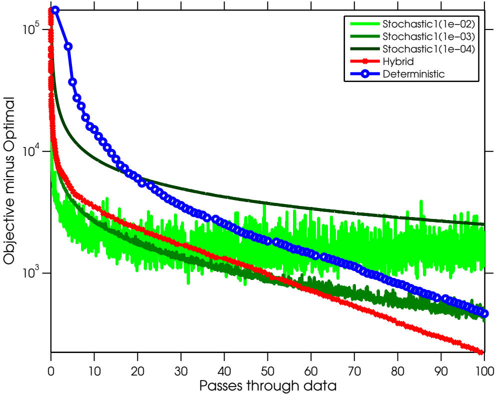
- Email spam classifier (Cormack and Lynam, 2005)
- TREC 2005 dataset: 92,189 email msgs from Enron investigation



Multinomial logistic regression

$$\min_{x} \sum_{i}^{m} -\log p(b_i = j \mid a_i, \{x\}_{j\in\mathcal{C}}), \quad b_i \in \mathcal{C}$$

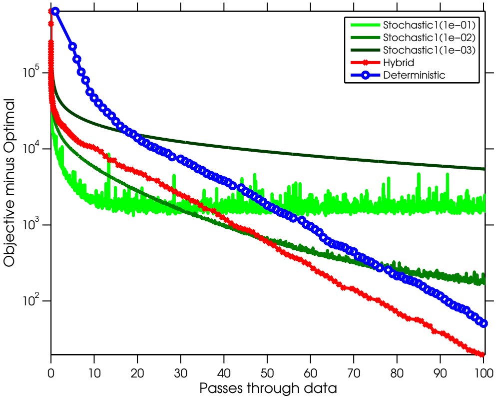
- Digit classification 
- MNIST dataset: 70,000 handwritten 28×28 images of digits



Chain-structured conditional random fields

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

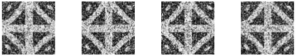
- Noun phrase chunking in natural-language processing
- CoNLL-2000 Shared Task dataset: 211,727 words in 8,936 sentences



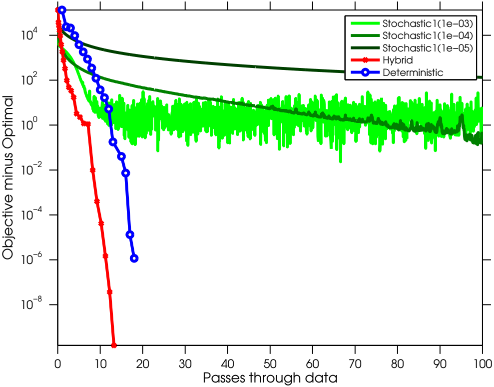
General conditional random fields

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

- Statistical denoising



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



Thanks!

Read:

- Herrmann, Friedlander, and Yılmaz, "Fighting the curse of dimensionality: compressive sensing in exploration seismology", August 2011
- Friedlander and Schmidt, "Hybrid deterministic-stochastic methods for data fitting", to appear in *SIAM J. Scientific Computing*, September 2011
- Aravkin, Friedlander, and van Leeuwen, "Robust inversion via semistochastic dimensionality reduction", October 2011
- Aravkin, Friedlander, Herrmann, and van Leeuwen, "Robust inversion, dimensionality reduction, and randomize sampling", November 2011

Email:

- mpf@cs.ubc.ca

Surf:

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