Image reconstruction for low-dose X-ray CT

Jeffrey A. Fessler & Donghwan Kim

EECS Dept., BME Dept., Dept. of Radiology
University of Michigan

web.eecs.umich.edu/~fessler

Gordon research conference
Accelerating the pace of system design and task-based evaluation

9 June 2014

Disclosure

- Research support from GE Healthcare
- Supported in part by NIH grants R01 HL-098686 and P01 CA-87634
- Equipment support from Intel Corporation

Credits

Current (CT) students / post-docs

- Jang Hwan Cho
- Donghwan Kim
- Jungkuk Kim
- Madison McGaffin
- Hung Nien
- Stephen Schmitt

GE collaborators

- Bruno De Man
- Jiang Hsieh
- Jean-Baptiste Thibault

CT collaborators

- Mitch Goodsitt, UM
- Ella Kazerooni, UM
- Neal Clinthorne, UM
- Paul Kinahan, UW
- Adam Alessio, UW

Former MS / undegraduate students

- Kevin Brown, Philips
- Meng Wu, Stanford
- ...

Former PhD students / post-docs (who did/do CT)

- Se Young Chun, UNIST
- Sathish Ramani, GE GRC
- Yong Long, UM-SJTU Joint Inst.
- Wonseok Huh, Bain & Company
- Hugo Shi, Continuum Analytics
- Joonki Noh, Emory
- Somesh Srivastava, GE HC
- Rongping Zeng, FDA
- Yingying, Zhang-O'Connor, RGM Advisors
- Matthew Jacobson, Xoran
- Sangtae Ahn, GE GRC
- Idris Elbakri, CancerCare / Univ. of Manitoba
- Saowapak Sotthivirat, NSTDA Thailand
- Web Stayman, JHU
- Feng Yu, Univ. Bristol
- Mehmet Yavuz, Qualcomm
- Hakan Erdoğan, Sabanci University

Statistical image reconstruction: a CT revolution

Thin-slice FBP

≈ 1974

ASIR
(denoised)

≈ 2008

Statistical
Reconstruction

≈ 2012

Why statistical/iterative methods for CT?

Benefits:

- Accurate physics models
  (reduced artifacts; improved quantification, spatial resolution, contrast)
- Nonstandard geometries
- Appropriate statistical models for measurements
  (reduced noise, hence reduced dose)
- Object constraints / priors

Disadvantages:

- Computation time (super computer)
- Must reconstruct entire FOV
- Complexity of models and software
- Algorithm nonlinearities
  - Difficult to analyze resolution/noise properties (*cf.* FBP)
  - Tuning parameters
  - Challenging to characterize performance / assess image quality

SIR for X-ray CT

Low-dose X-ray CT image reconstruction is a (constrained) optimization problem:

$$\hat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x} \succeq \boldsymbol{0}} \frac{1}{2} \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_W^2 + \beta R(\boldsymbol{x})$$

Ingredients:

- Sinogram data $\boldsymbol{y}$
- System matrix $\boldsymbol{A}$
- Statistical model (diagonal weighting matrix $\boldsymbol{W}$)
- Regularizer / log prior $\mathrm{R}(\boldsymbol{x})$
- Regularization parameter β
- Optimizer “arg min”

Regularization options for CT reconstruction

- Quadratic regularization: uselessly blurry
- Edge-preserving regularization (used clinically):

$$\mathsf{R}(\boldsymbol{x}) = \sum_{j=1}^{N} \sum_{k \in \mathcal{N}_j} \psi(x_j - x_k),$$

typically with strictly convex, non-quadratic potential functions ψ
- Total variation (akin to $\psi(t) = |t|$) to encourage “gradient sparsity”
- Extensions of TV
- Wavelet-based sparsity?
- Patch-based regularity:

$$\mathsf{R}(\boldsymbol{x}) = \sum_{j=1}^{N} \sum_{k \in \mathcal{N}_j} \psi(\boldsymbol{P}_j(\boldsymbol{x}) - \boldsymbol{P}_k(\boldsymbol{x}))$$

- Sparse representations in terms of patch dictionary
  - learned from training images (*e.g.*, high-dose CT scans)
  - learned adaptively from sinogram data

Relatively little work on task-based assessment of IQ for regularizer design in CT!

Why no consensus on best regularizer?

Non-quadratic regularizers lead to nonlinear estimators $\hat{\boldsymbol{x}}(\boldsymbol{y})$.

- Hard to analyze.
- Tedious to evaluate empirically - 3D helical X-ray CT

Need faster optimization algorithms:

- clinical X-ray CT
- regularization design
- task-based assessment investigations

Optimization problem:

$$\hat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x} \succeq \boldsymbol{0}} f(\boldsymbol{x}), \quad \underbrace{f(\boldsymbol{x}) = \frac{1}{2} \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_W^2 + \beta R(\boldsymbol{x})}_{\text{cost function}}$$

Challenges:

- large-scale
- non-quadratic
- constraints

Optimization problems in image reconstruction

(work of Donghwan Kim)

$$\hat{\boldsymbol{x}} \in \arg\min_{\boldsymbol{x}} f(\boldsymbol{x})$$

- Unconstrained
- Large-scale
  - Hessian too big to store
  - Even limited-memory Quasi-Newton is unattractive
- Cost function assumptions (throughout)
  - $f : \mathbb{R}^M \mapsto \mathbb{R}$
  - convex (need not be strictly convex)
  - non-empty set of global minimizers:

$$\hat{\boldsymbol{x}} \in \mathscr{X}^* = \left\{ \boldsymbol{x}_\star \in \mathbb{R}^M : f(\boldsymbol{x}_\star) \leq f(\boldsymbol{x}), \ \forall \boldsymbol{x} \in \mathbb{R}^M \right\}$$

  - smooth (differentiable with L-Lipschitz gradient)

$$\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{z})\|_2 \leq L \|\boldsymbol{x} - \boldsymbol{z}\|_2, \quad \forall \boldsymbol{x}, \boldsymbol{z} \in \mathbb{R}^M$$

Algorithms

Gradient descent (review)

Iteration with step size $1/L$ ensures monotonic descent of f:

$$\boldsymbol{x}_{n+1} = \boldsymbol{x}_n - \frac{1}{L} \nabla f(\boldsymbol{x}_n)$$

Classic $O(1/n)$ convergence rate of cost function descent:

$$\underbrace{f(\boldsymbol{x}_n) - f(\boldsymbol{x}_\star)}_{\text{inaccuracy}} \leq \frac{L \left\| \boldsymbol{x}_0 - \boldsymbol{x}_\star \right\|_2^2}{2n}.$$

$O(1/n)$ rate is undesirably slow.

Heavy ball method

Heavy ball iteration (Polyak, 1987):

$$\boldsymbol{x}_{n+1} = \boldsymbol{x}_n - \frac{\alpha}{L} \nabla f(\boldsymbol{x}_n) + \underbrace{\beta\,(\boldsymbol{x}_n - \boldsymbol{x}_{n-1})}_{\text{momentum!}} \qquad \text{(for implementation)}$$

$$= \boldsymbol{x}_n - \frac{1}{L} \sum_{k=0}^{n} \underbrace{\alpha \beta^{n-k}}_{\substack{\text{step-size} \\ \text{coefficients}}} \nabla f(\boldsymbol{x}_k) \qquad \text{(for analysis)}$$

- How to choose α and β?
- How to optimize step-size coefficients more generally?

General first-order method class

General "first-order" (FO) iteration:

$$\boldsymbol{x}_{n+1} = \boldsymbol{x}_n - \frac{1}{L} \sum_{k=0}^{n} h_{n+1,k} \nabla f(\boldsymbol{x}_k)$$

Primary goals:

- Analyze convergence rate of FO for any *given* set of step-size coefficients $H = \{h_{n,k} : n = 0, \ldots, N-1, k = 0, \ldots, n\}$
- Optimize set of step-size coefficients H.
  - Fast convergence
  - Efficient recursive implementation
  - Universal (design *prior* to iterating)
    Excludes CG, QN, BBGM, etc.

Nesterov's fast gradient method (FGM1)

Nesterov (1983) iteration: Initialize: $t_0 = 1$, $z_0 = x_0$

$$
\begin{aligned}
z_{n+1} &= x_n - \frac{1}{L}\nabla f(x_n) && \text{(usual GD update)}\\
t_{n+1} &= \frac{1}{2}\left(1 + \sqrt{1 + 4t_n^2}\right) && \text{(magic momentum factors)}\\
x_{n+1} &= z_{n+1} + \frac{t_n - 1}{t_{n+1}}(z_{n+1} - z_n) && \text{(update with momentum)} .
\end{aligned}
$$

Reverts to GD if $t_n = 1, \forall n$.

FGM1 is in class FO: $$x_{n+1} = x_n - \frac{1}{L}\sum_{k=0}^{n} h_{n+1,k}\nabla f(x_k)$$

$$
h_{n+1,k} = \begin{cases} \frac{t_n - 1}{t_{n+1}} h_{n,k}, & k = 0, \dots, n-2 \\ \frac{t_n - 1}{t_{n+1}} (h_{n,n-1} - 1), & k = n-1 \\ 1 + \frac{t_n - 1}{t_{n+1}}, & k = n. \end{cases}
$$

n	$h_{n,0}$	$h_{n,1}$	$\dots$			
0	1	0	0	0	0	0
1	0	1.25	0	0	0	0
2	0	0.10	1.40	0	0	0
3	0	0.05	0.20	1.50	0	0
4	0	0.03	0.11	0.29	1.57	0
5	0	0.02	0.07	0.18	0.36	1.62

Nesterov FGM1 optimal convergence rate

Shown by Nesterov to be $O(1/n^2)$ for "auxiliary" sequence:

$$f(\boldsymbol{z}_n) - f(\boldsymbol{x}_\star) \leq \frac{2L \|\boldsymbol{x}_0 - \boldsymbol{x}_\star\|_2^2}{(n+1)^2}.$$

Nesterov constructed a convex function f with L-Lipschitz gradient such that any first-order method achieves:

$$\frac{\frac{3}{32} L \|\boldsymbol{x}_0 - \boldsymbol{x}_\star\|_2^2}{(n+1)^2} \leq f(\boldsymbol{x}_n) - f(\boldsymbol{x}_\star).$$

- $O(1/n^2)$ rate of FGM1 is optimal.
- Potential acceleration by constant factor of > 20.

Overview

General first-order (FO) iteration:

$$\boldsymbol{x}_{n+1} = \boldsymbol{x}_n - \frac{1}{L} \sum_{k=0}^{n} h_{n+1,k} \nabla f(\boldsymbol{x}_k)$$

- Analyze (*i.e.*, bound) convergence rate as a function of
  - number of iterations N
  - Lipschitz constant L
  - step-size coefficients $H = \{h_{n+1,k}\}$
  - Distance to a solution: $R = \|\boldsymbol{x}_0 - \boldsymbol{x}_\star\|$
- Optimize step-size coefficients H by minimizing the bound

Ideal "universal" bound for first-order methods

For given

- number of iterations N
- Lipschitz constant L
- step-size coefficients $H = \{h_{n+1,k}\}$
- distance to a solution: $R = \|\boldsymbol{x}_0 - \boldsymbol{x}_\star\|$

Drori & Teboulle (2014) bound the worst-case convergence rate of FO algorithm:

$$B_1(H,R,L,N) \triangleq \max_{f \in \mathscr{F}_L} \max_{\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_N \in \mathbb{R}^M} \max_{\substack{\boldsymbol{x}_\star \in \mathscr{X}^*(f) \\ \|\boldsymbol{x}_0 - \boldsymbol{x}_\star\| \leq R}} f(\boldsymbol{x}_N) - f(\boldsymbol{x}_\star)$$

such that $\quad \boldsymbol{x}_{n+1} = \boldsymbol{x}_n - \frac{1}{L} \sum_{k=0}^{n} h_{n+1,k} \nabla f(\boldsymbol{x}_k), \quad n = 0, \ldots, N-1.$

Clearly for any FO method:

$$f(\boldsymbol{x}_N) - f(\boldsymbol{x}_\star) \leq B_1(H,R,L,N).$$

Towards practical bounds for first-order methods

For convex functions with L-Lipschitz gradients

$$\frac{1}{2L}\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{z})\|^2 \leq f(\boldsymbol{x}) - f(\boldsymbol{z}) - \langle \nabla f(\boldsymbol{z}), \boldsymbol{x} - \boldsymbol{z} \rangle, \quad \forall \boldsymbol{x}, \boldsymbol{z} \in \mathbb{R}^M.$$

Drori & Teboulle (2014) use this inequality to propose a "more tractable" bound:

$$B_2(H, R, L, N) \triangleq \max_{\boldsymbol{g}_0,\ldots,\boldsymbol{g}_N \in \mathbb{R}^M} \max_{\delta_0,\ldots,\delta_N \in \mathbb{R}} \max_{\boldsymbol{x}_0,\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N \in \mathbb{R}^M} \max_{\boldsymbol{x}_\star : \|\boldsymbol{x}_0 - \boldsymbol{x}_\star\| \leq R} LR\delta_N^2$$

such that $\quad \boldsymbol{x}_{n+1} = \boldsymbol{x}_n - \frac{1}{L}\sum_{k=0}^{n} h_{n+1,k} R\, \boldsymbol{g}_k, \quad n = 0, \ldots, N-1$

$$\frac{1}{2}\|\boldsymbol{g}_i - \boldsymbol{g}_j\|^2 \leq \delta_i - \delta_j - \frac{1}{R}\langle \boldsymbol{g}_j, \boldsymbol{x}_i - \boldsymbol{x}_j \rangle, \quad i, j = 0, \ldots, N.$$

Looser bound for any FO method:

$$f(\boldsymbol{x}_N) - f(\boldsymbol{x}_\star) \leq B_1(H, R, L, N) \leq B_2(H, R, L, N).$$

However, even B_2 is as of yet unsolved.

Numerical bounds for first-order methods

Drori & Teboulle (2014) further relax the bound

Leads eventually to a still simpler optimization problem
(but still with no known closed-form solution):

$$f(\boldsymbol{x}_N) - f(\boldsymbol{x}_\star) \leq B_1(H, R, L, N) \leq B_2(H, R, L, N) \leq B_3(H, R, L, N).$$

For given step-size coefficients H, and given number of iterations N,
they compute B_3 numerically, using a semi-definite program (SDP).

Optimizing step-size coefficients numerically

Drori & Teboulle (2014) also compute numerically the minimizer over H of their relaxed bound for given N using a semi-definite program (SDP):

$$H^* = \arg\min_H B_3(H, R, L, N).$$

Numerical solution for H^* for $N = 5$ iterations: [Fig. from Drori & Teboulle (2014)]

0. Input: $f \in C_L^{1,1}(\mathbb{R}^d)$, $x_0 \in \mathbb{R}^d$,
1. $x_1 = x_0 - \frac{1.6180}{L} f'(x_0)$,
2. $x_2 = x_1 - \frac{0.1741}{L} f'(x_0) - \frac{2.0194}{L} f'(x_1)$,
3. $x_3 = x_2 - \frac{0.0756}{L} f'(x_0) - \frac{0.4425}{L} f'(x_1) - \frac{2.2317}{L} f'(x_2)$,
4. $x_4 = x_3 - \frac{0.0401}{L} f'(x_0) - \frac{0.2350}{L} f'(x_1) - \frac{0.6541}{L} f'(x_2) - \frac{2.3656}{L} f'(x_3)$,
5. $x_5 = x_4 - \frac{0.0178}{L} f'(x_0) - \frac{0.1040}{L} f'(x_1) - \frac{0.2894}{L} f'(x_2) - \frac{0.6043}{L} f'(x_3) - \frac{2.0778}{L} f'(x_4)$.

Drawbacks

- Must choose N in advance
- Requires $O(N)$ memory for all gradient vectors $\{\nabla f(\boldsymbol{x}_n)\}_{n=1}^N$
- $O(N^2)$ computation for N iterations

Benefit: convergence bound (for specific N) $\approx 2\times$ lower than for Nesterov's FGM1.

New results

(paper submitted in May 2014)

(skipping long derivations...)

New analytical solution

- Analytical solution for optimized step-size coefficients (Donghwan Kim, 2014):

$$H^* : \quad h_{n+1,k} = \begin{cases} \frac{\theta_n - 1}{\theta_{n+1}} h_{n,k}, & k = 0, \ldots, n-2 \\ \frac{\theta_n - 1}{\theta_{n+1}} (h_{n,n-1} - 1), & k = n-1 \\ 1 + \frac{2\theta_n - 1}{\theta_{n+1}}, & k = n. \end{cases}$$

$$\theta_n = \begin{cases} 1, & n = 0 \\ \frac{1}{2}\left(1 + \sqrt{1 + 4\theta_{n-1}^2}\right), & n = 1, \ldots, N-1 \\ \frac{1}{2}\left(1 + \sqrt{1 + 8\theta_{n-1}^2}\right), & n = N. \end{cases}$$

- Analytical convergence bound for these optimized step-size coefficients:

$$f(\boldsymbol{x}_N) - f(\boldsymbol{x}_\star) \leq B_3(H^*, R, L, N) = \frac{1L \|\boldsymbol{x}_0 - \boldsymbol{x}_\star\|_2^2}{(N+1)(N+1+\sqrt{2})}.$$

Of course bound is $O(1/N^2)$, but constant is twice better than that of Nesterov.
No numerical SDP needed $\Longrightarrow$ feasible for large N.

Optimized gradient method (OGM1)

Donghwan Kim (2014) found efficient recursive iteration:

Initialize: $\theta_0 = 1$, $\boldsymbol{z}_0 = \boldsymbol{x}_0$

$$\boldsymbol{z}_{n+1} = \boldsymbol{x}_n - \frac{1}{L} \nabla f(\boldsymbol{x}_n) \qquad \text{(usual GD update)}$$

$$\theta_n = \begin{cases} \frac{1}{2}\left(1 + \sqrt{1 + 4\theta_{n-1}^2}\right), & n = 1, \ldots, N-1 \\ \frac{1}{2}\left(1 + \sqrt{1 + 8\theta_{n-1}^2}\right), & n = N \end{cases} \qquad \text{(momentum factors)}$$

$$\boldsymbol{x}_{n+1} = \boldsymbol{z}_{n+1} + \frac{\theta_n - 1}{\theta_{n+1}} (\boldsymbol{z}_{n+1} - \boldsymbol{z}_n) + \underbrace{\frac{\theta_n}{\theta_{n+1}} (\boldsymbol{z}_{n+1} - \boldsymbol{x}_n)}_{\text{new momentum}}.$$

Reverts to Nesterov's FGM1 if the new terms are removed.

- Very simple modification of existing Nesterov code
- No need to choose N in advance (or solve SDP);
  use favorite stopping rule then run one last "decreased momentum" step.
- Factor of 2 better upper bound than Nesterov's "optimal" FGM1.

(Proofs omitted.)

Further acceleration...

Combining ordered subsets (OS) with momentum

Optimization problems in image reconstruction (and machine learning) involve sums of many similar terms:

$$f(\boldsymbol{x}) = \sum_{m=1}^{M} f_m(\boldsymbol{x}) \,.$$

Approximate gradients using just one term at a time:

$$\nabla f(\boldsymbol{x}) \approx M \nabla f_m(\boldsymbol{x})$$

- Ordered subsets (OS) in tomography
- Incremental gradients in optimization / machine learning

Combining OS with momentum leads to dramatic acceleration!

OS + OGM1 method

Initialize: $\theta_0 = 1$, $\boldsymbol{z}_0 = \boldsymbol{x}_0$
For each iteration n
For each subset $m = 1, \ldots, M$

$$k = nM + m - 1$$

$$\boldsymbol{z}_{k+1} = \boldsymbol{x}_k - \frac{M}{L} \nabla f_m(\boldsymbol{x}_k) \qquad \text{(usual OS update)}$$

$$\theta_k = \frac{1}{2}\left(1 + \sqrt{1 + 4\theta_{k-1}^2}\right) \qquad \text{(momentum factors)}$$

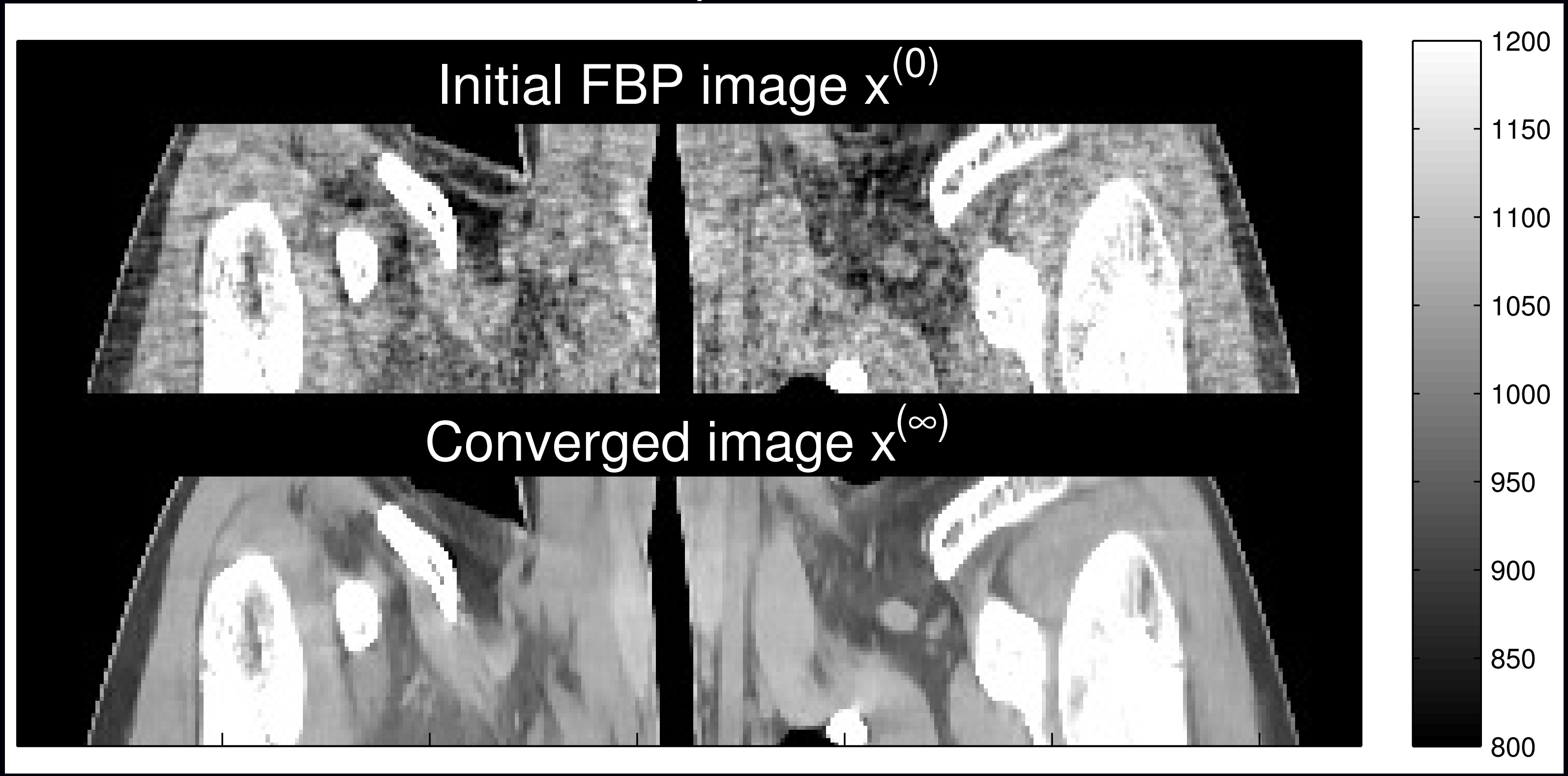
$$\boldsymbol{x}_{k+1} = \boldsymbol{z}_{k+1} + \frac{\theta_k - 1}{\theta_{k+1}}(\boldsymbol{z}_{k+1} - \boldsymbol{z}_k) + \underbrace{\frac{\theta_k}{\theta_{k+1}}(\boldsymbol{z}_{k+1} - \boldsymbol{x}_k)}_{\text{new momentum}}.$$

- Simple modification of existing OS code
- Roughly $O\left(1/(Mn)^2\right)$ decrease of cost function f in early iterations

New empirical results

Results: 3D X-ray CT patient scan

- 3D cone-beam helical CT scan with pitch 0.5

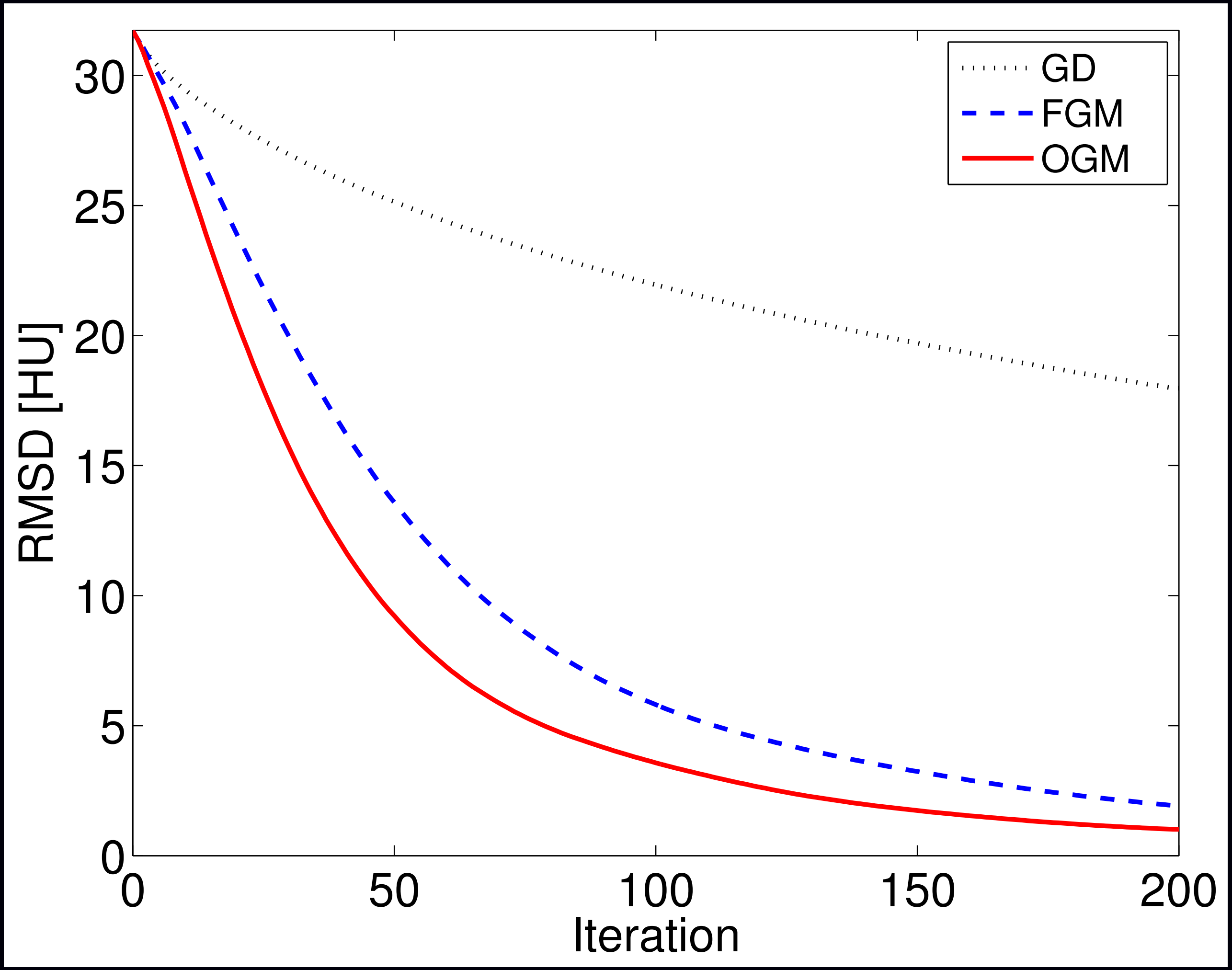


- Convergence rate in RMSD [HU], within ROI, versus iteration:

$$\mathrm{RMSD}_{\mathrm{ROI}}(\boldsymbol{x}_n) \triangleq \frac{||x_{\mathrm{ROI}}^{(n)} - \hat{x}_{\mathrm{ROI}}||_2}{\sqrt{N_{\mathrm{ROI}}}}.$$

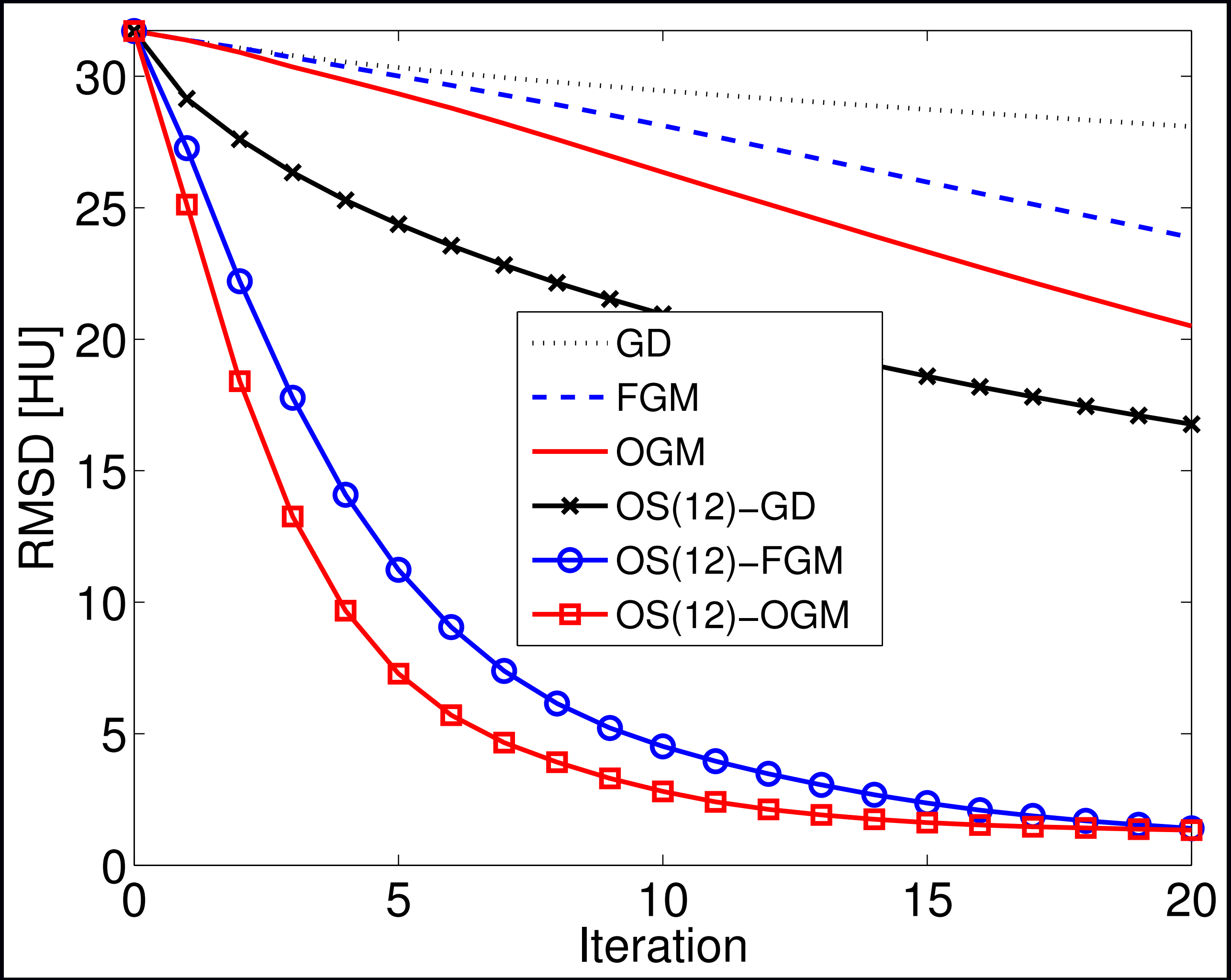
(Disclaimer: RMSD may not relate to task performance...)

Results: RMSD [HU] vs. iteration: without OS



- Convergence speed: GD $\ll$ FGM $<$ OGM
- OGM requires about $\frac{1}{\sqrt{2}}$-times fewer iterations than FGM to reach the same RMSD.

Results: RMSD [HU] vs. iteration: with OS



- $M = 12$ subsets in OS algorithm.
- Proposed OS-OGM converges faster than OS-FGM.
- Computation time per iteration of all algorithms are similar.

Summary

- New optimized first-order minimization algorithm
- Simple implementation akin to Nesterov's FGM
- Analytical converge rate bound
- Bound is $2\times$ better than Nesterov
- Combining with ordered subsets (OS) provides dramatic acceleration

Future work

- Optimization method
  - Constraints
  - Non-smooth cost functions, *e.g.*, ℓ_1
  - Tighter bounds
  - Strongly convex case
  - Asymptotic / local convergence rates
  - Incremental gradients / relaxation
  - Stochastic gradient descent
  - Adaptive restart
- Low-dose X-ray CT image reconstruction
  - Regularization design
  - Task-based IQ assessment

Bibliography

[1] Y. Drori and M. Teboulle. Performance of first-order methods for smooth convex minimization: A novel approach. Mathematical Programming, 145(1-2):451–82, June 2014.

[2] Y. Nesterov. A method for unconstrained convex minimization problem with the rate of convergence $O(1/k^2)$. Dokl. Akad. Nauk. USSR, 269(3):543–7, 1983.

[3] Y. Nesterov. Smooth minimization of non-smooth functions. Mathematical Programming, 103(1):127–52, May 2005.

[4] D. Kim and J. A. Fessler. Optimized first-order methods for smooth convex minimization. Mathematical Programming, 2015. Submitted.

[5] D. Kim, S. Ramani, and J. A. Fessler. Combining ordered subsets and momentum for accelerated X-ray CT image reconstruction. IEEE Trans. Med. Imag., 34(1):167–78, January 2015.

Not: Barzilai-Borwein gradient method

Barzilai & Borwein, 1988

$$\boldsymbol{g}^{(n)} \triangleq \nabla f(\boldsymbol{x}_n)$$

$$\alpha_n = \frac{\|\boldsymbol{x}_n - \boldsymbol{x}_{n-1}\|^2}{\langle \boldsymbol{x}_n - \boldsymbol{x}_{n-1}, \boldsymbol{g}^{(n)} - \boldsymbol{g}^{(n-1)} \rangle}$$

$$\boldsymbol{x}_{n+1} = \boldsymbol{x}_n - \alpha_n \nabla f(\boldsymbol{x}_n) .$$

Not in "first-order" class FO.
Neither are methods like

- ◦ steepest descent (with line search),
- ◦ conjugate gradient,
- ◦ quasi-Newton ...