Iterative Reconstruction in CT and MRI

Jeffrey A. Fessler
EECS Dept., BME Dept., Dept. of Radiology
University of Michigan
web.eecs.umich.edu/~fessler

Fully 3D Conference
02 June 2015
Iterative Reconstruction in CT and MRI

(and a bit of PET and SPECT)

Jeffrey A. Fessler

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

web.eecs.umich.edu/~fessler

Fully 3D Conference
02 June 2015
Disclosure

- Research support from GE Healthcare
- Supported in part by NIH grants P01 CA-87634, U01 EB018753
- Equipment support from Intel Corporation

Acknowledgment:
many collaborators and many students and post-docs
Outline

What
  CT
  MRI

Why
  Why CT iterative
  Why MRI iterative

How
  Optimization transfer
  Separable quadratic surrogates
  Momentum
  Ordered subsets

Parallelization
Outline

What
CT
MRI

Why
Why CT iterative
Why MRI iterative

How
Optimization transfer
Separable quadratic surrogates
Momentum
Ordered subsets

Parallelization
X-ray CT scans

CT image reconstruction problem:
Determine unknown attenuation map \( x \) given sinogram data \( y \) using system matrix \( A \).

cf. SPECT with orbiting gamma camera
MR image reconstruction problem:
Determine unknown magnetization image $x$ given k-space data $y$ using system matrix $A$

Defer motion for now...
Inverse problems

Unknown object $x$ → Imaging system → Data $y$ → Recon → Image $\hat{x}$

How to reconstruct object $x$ from data $y$?

Non-iterative methods:
- analytical / direct
  - Filtered back-projection (FBP) for CT (textbook: Radon transform)
  - Inverse FFT for MRI (textbook: FFT)
- idealized description of the system
  - geometry / sampling
  - disregards noise and simplifies physics
- typically fast

Iterative methods:
- model-based / statistical
- based on “reasonably accurate” models for physics and statistics
- usually much slower
Statistical image reconstruction: CT example

- A picture is worth 1000 words
- (and perhaps several 1000 seconds of computation?)

| Thin-slice FBP Seconds | ASIR (denoise) A bit longer | Statistical Much longer |

(Same sinogram, so all at same dose)
Outline

What
  CT
  MRI

Why
  Why CT iterative
  Why MRI iterative

How
  Optimization transfer
  Separable quadratic surrogates
  Momentum
  Ordered subsets

Parallelization
Why statistical/iterative methods for CT?

- **Accurate physics models**
  - X-ray spectrum, beam-hardening, scatter, ...
    → reduced artifacts? quantitative CT?
  - X-ray detector spatial response, focal spot size, ...
    → improved spatial resolution?
  - detector spectral response (e.g., photon-counting detectors)
    → improved contrast between distinct material types?

- **Nonstandard geometries**
  - transaxial truncation (wide patients)
  - long-object problem in helical CT
  - irregular sampling in “next-generation” geometries
  - coarse angular sampling in image-guidance applications
  - limited angular range (tomosynthesis)
  - “missing” data, e.g., bad pixels in flat-panel systems
Why iterative for CT ... continued

- Appropriate models of (data dependent) measurement statistics
  - weighting reduces influence of photon-starved rays (cf. FBP) \(\Rightarrow\) reducing image noise or X-ray dose

- **Object** constraints / priors
  - nonnegativity
  - object support
  - piecewise smoothness
  - object sparsity (e.g., angiography)
  - sparsity in some basis
  - motion models
  - dynamic models
  - ...

Constraints may help reduce image artifacts or noise or dose.

Similar motivations/benefits in PET and SPECT.
Disadvantages of iterative methods for CT?

- Computation time
- 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 IQ
Sub-mSv example

3D helical X-ray CT scan of abdomen/pelvis:
100 kVp, 25-38 mA, 0.4 second rotation, 0.625 mm slice, 0.6 mSv.
MBIR example: Chest CT

Helical chest CT study with dose = 0.09 mSv.
Typical CXR effective dose is about 0.06 mSv.

(Health Physics Soc.: http://www.hps.org/publicinformation/ate/q2372.html)

FBP

MBIR

Veo (MBIR) images courtesy of Jiang Hsieh, GE Healthcare
History: Statistical reconstruction for X-ray CT

- Iterative method for X-ray CT (Hounsfield, 1968)
- ART for tomography (Gordon, Bender, Herman, JTB, 1970)
- Roughness regularized LS for tomography (Kashyap & Mittal, 1975)
- Poisson likelihood (transmission) (Rockmore and Macovski, TNS, 1977)
- EM algorithm for Poisson transmission (Lange and Carson, JCAT, 1984)
- Iterative coordinate descent (ICD) (Sauer and Bouman, T-SP, 1993)
- Ordered-subsets algorithms (Manglos et al., PMB 1995)
  (Kamphuis & Beekman, T-MI, 1998)
  (Erdoğan & Fessler, PMB, 1999)
- Commercial OS for Philips BrightView SPECT-CT (2010)
- Commercial ICD for GE CT scanners (circa 2010)
- FDA 510(k) clearance of Veo (Sep. 2011)
- First Veo installation in USA (at UM) (Jan. 2012)

(* numerous omissions, including many denoising methods)
Optimization problem formulation: \( \hat{x} = \arg \min_{x \geq 0} \Psi(x) \)

\[
\Psi(x) \triangleq \frac{1}{2} \| y - Ax \|_W^2 + \beta \sum_{j=1}^{N} \sum_{k \in N_j} \psi(x_j - x_k)
\]

\( y \): measured data (sinogram)
\( A \): system matrix (physics / geometry)
\( W \): weighting matrix (statistics)
\( x \): unknown image (attenuation map)
\( \beta \): regularization parameter(s)
\( N_j \): neighborhood of \( j \)th voxel
\( \psi \): edge-preserving potential function
(piece-wise smoothness / gradient sparsity)
\[ \hat{x} = \arg \min_{x \geq 0} \psi(x), \quad \psi(x) \triangleq \frac{1}{2} \| y - Ax \|_W^2 + \sum_j \sum_k \beta_{j,k} \psi(x_j - x_k) \]

Apparent topics:
- regularization design / parameter selection \( \psi, \beta_{jk} \)
- statistical modeling \( W, \| \cdot \| \)
- system modeling \( A \)
- optimization algorithms (arg min)
- assessing IQ of \( \hat{x} \)

Other topics:
- system design
- motion
- spectral
- dose ...
Inverse FFT is fast (like FBP). Why change?

(Joint work with D. Noll, J. Nielsen, ...)

Recall rationale for CT/PET/SPECT:

- **physics** modeling
  - reduce artifacts
  - improve resolution
  - improve contrast
- **noise** modeling: (dose, variability)
- **sampling**: non-standard geometries
- **constraints** on object

Which of these matter for MRI?
Physics modeling (e.g., field inhomogeneity) $\implies$ reduced artifacts

Example: T2*-weighted imaging (Sutton et al., IEEE T-MI, 03)

\[
\hat{x} = \arg \min_x \frac{1}{2} \|y - Ax\|_2^2 + \beta R(x)
\]

System matrix $A$ depends on (measured) field map:

\[
a_{ij} = e^{-\omega_j t_i} e^{-i2\pi \vec{v}_i \cdot \vec{r}_j}
\]

No analytical inverse of $A$. cf. nonuniform attenuation correction in SPECT.
Joint estimation of field map $\omega$ and magnetization image $x$:

$$(\hat{x}, \hat{\omega}) = \arg\min_{x, \omega} \frac{1}{2} \| y - A(\omega)x \|_2^2 + \beta_1 R_1(x) + \beta_2 R_2(\omega)$$

Useful when field map drifts in dynamic imaging.

(Sutton et al., MRM 04) (Olafsson et al., T-MI 08)

$cf.$ joint estimation of attenuation map $\mu$ and activity image $\lambda$ in SPECT, PET and TOF-PET.

(Censor et al., T-NS 79) (Clinthorne et al., NSS 91) (Rezaei, Defrise, Nuyts, T-MI 14)
RF pulse design

\[ b \rightarrow \text{Bloch Eqn} \rightarrow m \]

Small-tip approximation: \( m \approx Ab \)

Iterative RF pulse design (with RF power regularization):

\[
\arg \min_b \| m - Ab \|_2^2 + \beta \| b \|_2^2
\]

Minimize using CG.

(Yip et al., MRM, Oct. 2005)

d. Non-iterative:
e. Iterative:
MRI why iterative: Noise

- MRI measurements: (complex) AWGN $\Rightarrow$ easy !?
MRI why iterative: Noise

- MRI measurements: (complex) AWGN $\implies$ easy !?
- Variance of image phase depends on image magnitude.
- Image phase useful in some applications, e.g., B1 mapping:

Unregularized vs regularized phase estimate. (Zhao et al., T-MI 14)
MRI why iterative: Sampling

- Reducing k-space sampling $\implies$ reduced scan time
- Especially compelling for dynamic imaging *(cf. CT and SPECT)*
- Popular “under-sampled” patterns: *(cf. sparse-view CT)*

Solution strategies
- Multiple receive coils
- Object model assumptions (*e.g.*, sparsity)
- Iterative reconstruction (“compressed sensing”)
Parallel MRI

Under-sampled Cartesian k-space: use multiple receive coils with individual spatial sensitivity patterns.  
(Pruessmann et al., MRM, 1999)

Compressed sensing parallel MRI $\equiv$ (random) under-sampling


cf. multiple-source CT (speed) or multi-camera SPECT (counts)
Regularized estimator:

\[ \hat{x} = \arg \min_x \frac{1}{2} \| y - FSx \|_2^2 + \beta \| Rx \|_p. \]

\( F \) is under-sampled DFT matrix (wide)

Features:
- coil sensitivity matrix \( S \) is block diagonal
- \( F'F \) is circulant (for Cartesian sampling)

Challenges:
- Data-fit Hessian \( S'F'FS \) is highly shift variant due to coil sensitivity maps
- Non-quadratic (edge-preserving) regularization \( \| \cdot \|_p \)
- Non-smooth regularization \( \| \cdot \|_1 \) (cf. sparse view CT)
- Complex quantities
- Large problem size (if 3D or dynamic or many coils)
2.5D parallel MR image reconstruction

Example of “compressed sensing” MRI reconstruction:

- Fully sampled body coil image of human brain (144 × 128)
- Poisson-disk-based k-space sampling, 16% sampling (acceleration 6.25)
- Square-root of sum-of-squares inverse FFT of zero-filled k-space data for 8 coils
- Regularized reconstruction \( x^{(\infty)} \) combined TV and \( \ell_1 \) norm of two-level undecimated Haar wavelets
- Difference image magnitude

(Sathish Ramani & JF, IEEE T-MI, Mar. 2011)
Summary of “What” and “Why”

- CT and MRI both involve inverse problems
- Some similarities in motivations and formulations
- Some similarities in computation challenges
- Some opportunities for cross-fertilization
- Caution: MRI reconstruction field is crowded!
Outline

What
CT
MRI

Why
Why CT iterative
Why MRI iterative

How
Optimization transfer
Separable quadratic surrogates
Momentum
Ordered subsets

Parallelization
SIR for CT: Optimization challenges

\[ \hat{x} = \arg \min_{x \geq 0} \psi(x), \quad \psi(x) \triangleq \frac{1}{2} \| y - Ax \|^2_W + \sum_{j=1}^N \sum_{k} \beta_{j,k} \psi(x_j - x_k) \]

Optimization challenges:

- large problem size: \( x \in \mathbb{R}^{512 \times 512 \times 600}, \ y \in \mathbb{R}^{888 \times 64 \times 7000} \)
- \( A \) is sparse but still too large to store; compute \( Ax \) on-the-fly
- \( W \) has enormous dynamic range \((1 \ \text{to} \ \exp(-9) \approx 1.2 \cdot 10^{-4})\)
- Gram matrix \( A'WA \) highly shift variant
- \( \Psi \) is non-quadratic but convex (and often smooth)
- nonnegativity constraint
- data size grows: dual-source CT, spectral CT, wide-cone CT, ...
- Moore’s law insufficient
Optimization transfer (Majorize-Minimize) methods: 1D

![Diagram showing optimization transfer]

\[ \phi^{(n)}(x) = \Psi(x^{(n)}) \]
\[ \phi^{(n)}(x) \geq \Psi(x) \]

\[ x^{(n+1)} = \arg\min_x \phi^{(n)}(x) \]

cf. ML-EM
Optimization transfer (Majorize-Minimize) methods: 2D
Outline

What
- CT
- MRI

Why
- Why CT iterative
- Why MRI iterative

How
- Optimization transfer
  - Separable quadratic surrogates
- Momentum
- Ordered subsets

Parallelization
Separable Quadratic Surrogates (SQS): Math

\[ L(x) = \frac{1}{2} \| y - Ax \|^2_W \]

\[ = L(x^{(n)}) + \nabla L(x^{(n)})(x - x^{(n)}) + \frac{1}{2} (x - x^{(n)})' A' WA (x - x^{(n)}) \]

non-separable

\[ \leq L(x^{(n)}) + \nabla L(x^{(n)})(x - x^{(n)}) + \frac{1}{2} (x - x^{(n)})' D (x - x^{(n)}) \]

separable

\[ \equiv \phi_L^{(n)}(x), \quad \text{a "SQS"}, \]

where \( A' WA \leq D = \text{diag}\{A' WA1\} \). (De Pierro, T-MI, Mar. 1995)

Proofs:

- Convexity of \( x^2 \)
- Geršgorin disk theorem
- Cauchy-Schwarz inequality
Separable Quadratic Surrogates (SQS): Pictures

- Find minimizer of $L(x)$: challenging
- Find minimizer of $\phi_{L}^{(n)}(x)$: easy (separate 1D problems)
WLS-SQS: Iteration

General optimization transfer (majorize-minimize) method:

\[ x^{(n+1)} = \arg \min_x \phi_{L}^{(n)}(x) \]

For SQS:

\[ \phi_{L}^{(n)}(x) = L(x^{(n)}) + \nabla L(x^{(n)})(x - x^{(n)}) + \frac{1}{2} (x - x^{(n)})' \quad D \quad (x - x^{(n)}) \]

\[ \nabla \phi_{L}^{(n)}(x) = \nabla L(x^{(n)}) + D \quad (x - x^{(n)}) \]

\[ 0 = \nabla \phi_{L}^{(n)}(x^{(n+1)}) = \nabla L(x^{(n)}) + D \quad (x^{(n+1)} - x^{(n)}) \]

\[ x^{(n+1)} = x^{(n)} - D^{-1} \nabla L(x^{(n)}) \]

“diagonally preconditioned gradient descent”

(Erdoğan & JF, PMB, 1999)
Ordinary gradient descent (GD) for WLS:

$$x^{(n+1)} = x^{(n)} - \alpha \nabla L(x^{(n)}) = x^{(n)} - \alpha A' W (Ax^{(n)} - y),$$

where textbook step size is reciprocal of Lipschitz constant:

$$\alpha = \frac{1}{\lambda_{\text{max}}(A'WA)}.$$

WLS-GD is equivalent to WLS-SQS with “isotropic” majorizer Hessian:

$$D = \lambda_{\text{max}}(A'WA) I.$$ 

Drawbacks:

- $$\lambda_{\text{max}}(A'WA)$$ usually impractical to compute (in CT)
- Usually slower convergence due to smaller step sizes
SQS versus GD: Pictures

Lipshitz Majorizer

SQS Majorizer
SQS versus GD: Pictures

Lipshitz Majorizer

SQS Majorizer
Outline

What
  CT
  MRI

Why
  Why CT iterative
  Why MRI iterative

How
  Optimization transfer
  Separable quadratic surrogates
  Momentum
  Ordered subsets

Parallelization
Classical gradient descent (GD)

Assumptions:
- \( \Psi \) is convex (need not be strictly convex)
- \( \Psi \) has non-empty set of global minimizers
  \[ \hat{x} \in \mathcal{X}^* = \{ x^{(*)} \in \mathbb{R}^N : \Psi(x^{(*)}) \leq \Psi(x), \ \forall x \in \mathbb{R}^N \} \]
- \( \Psi \) is smooth (differentiable with \( L \)-Lipschitz gradient)
  \[ \| \nabla \Psi(x) - \nabla \Psi(z) \|_2 \leq L \| x - z \|_2, \quad \forall x, z \in \mathbb{R}^N \]

GD with step size \( 1/L \) ensures monotonic descent of \( \Psi \):

\[ x^{(n+1)} = x^{(n)} - \frac{1}{L} \nabla \Psi(x^{(n)}) . \]

Drori & Teboulle (2014) derive tightest “inaccuracy” bound:

\[ \Psi(x^{(n)}) - \Psi(x^{(\ast)}) \leq \frac{L \| x^{(0)} - x^{(\ast)} \|_2^2}{4n + 2} . \]

For a Huber-like function \( \Psi \), GD achieves that (tight) bound. \( O(1/n) \) rate is undesirably slow.
Nesterov’s fast gradient method (FGM1)

Nesterov (1983) iteration: Initialize: \( t_0 = 1, \ z^{(0)} = x^{(0)} \)

\[
z^{(n+1)} = x^{(n)} - \frac{1}{L} \nabla \psi(x^{(n)})
\]  
(usual GD update)

\[
t_{n+1} = \frac{1}{2} \left( 1 + \sqrt{1 + 4t_n^2} \right)
\]  
(magic momentum factors)

\[
x^{(n+1)} = z^{(n+1)} + \frac{t_n - 1}{t_{n+1}} (z^{(n+1)} - z^{(n)})
\]  
(update with momentum)

- Reverts to GD if \( t_n = 1, \ \forall n. \)
- Comparable computation as GD
- Store one additional image-sized vector \( z^{(n)} \)
FGM1 properties

FGM1 shown by Nesterov to be $O(1/n^2)$ for “primary” sequence:

$$\psi(z^{(n)}) - \psi(x^{(*)}) \leq \frac{2L \|x^{(0)} - x^{(*)}\|^2}{(n + 1)^2}.$$

Nesterov constructed a function $\psi$ such that any first-order method achieves

$$\frac{3}{32} \frac{L \|x^{(0)} - x^{(*)}\|^2}{(n + 1)^2} \leq \psi(x^{(n)}) - \psi(x^{(*)}).$$

Thus $O(1/n^2)$ rate of FGM1 is optimal.

Donghwan Kim (2014) analyzed “secondary” sequence:

$$\psi(x^{(n)}) - \psi(x^{(*)}) \leq \frac{2L \|x^{(0)} - x^{(*)}\|^2}{(n + 2)^2}.$$
“Traditional” iterative soft thresholding algorithm (ISTA) uses (global) Lipschitz constant of data-fit term:

$$\nabla^2 \frac{1}{2} \| y - FS \|_2^2 = S' F' FS \leq S' S \leq \lambda_{\text{max}} I, \quad \lambda_{\text{max}} = \max_j [S'S]_{j,j}$$

$\lambda_{\text{max}}$ is maximum sum-of-squares value of sensitivity maps.

Augmented Lagrangian (AL) methods converge faster than ISTA, FISTA, MFISTA (Ramani & JF, T-MI, 2011)

BARISTA (B1-based, adaptive restart, ISTA) (Muckley, Noll, JF, T-MI, 2015)

For synthesis operator $x = Qz$ with $z$ sparse:

$$\nabla^2 \frac{1}{2} \| y - FSQ \|_2^2 = Q'S' F' FSQ \leq Q'S'SQ \leq D$$

for a suitable diagonal matrix $D$. (cf., SQS)

$D^{-1}$ becomes voxel-dependent step size, akin to SQS in CT
BARISTA convergence rates

“Compressed sensing” MRI reconstruction:
Total variation (TV) regularizer                Undecimated Haar Wavelets

Corresponding $D$ for each of the two cases:
BARISTA requires no algorithm parameter tuning, unlike AL.
Includes momentum with adaptive restart of O’Donoghue and Candès (2014).
FGM1 is in the general class of first-order methods:

\[ x^{(n+1)} = x^{(n)} - \frac{1}{L} \sum_{k=0}^{n} h_{n+1,k} \nabla \psi(x^{(k)}) \]

where the step-size factors \( \{h_{n,k}\} \) are

\[
\begin{bmatrix}
1 & 0 & 0 & 0 & 0 & 0 \\
0 & 1.25 & 0 & 0 & 0 & 0 \\
0 & 0.10 & 1.40 & 0 & 0 & 0 \\
0 & 0.05 & 0.20 & 1.50 & 0 & 0 \\
0 & 0.03 & 0.11 & 0.29 & 1.57 & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots 
\end{bmatrix}
\]

Use of previous gradients \( \iff \) “momentum”

Is this the optimal choice for \( \{h_{n,k}\} \)?

Can we improve on the constant 2 in worst-case convergence rate?

Drori & Teboulle (2014) numerically found 2× better \( \{h_{n,k}\} \)
Optimized gradient method (OGM1)

New approach by optimizing \( \{h_{n,k}\} \) analytically

Initialize: \( t_0 = 1, \ z^{(0)} = x^{(0)} \)  
(Donghwan Kim and JF; 2014, 2015)

\[
\begin{align*}
  z^{(n+1)} &= x^{(n)} - \frac{1}{\ell} \nabla \Psi(x^{(n)}) & \text{(usual GD update)} \\
  t_{n+1} &= \frac{1}{2} \left( 1 + \sqrt{1 + 4t_n^2} \right) & \text{(momentum factors)} \\
  x^{(n+1)} &= z^{(n+1)} + \frac{t_n - 1}{t_{n+1}} \left( z^{(n+1)} - z^{(n)} \right) + \frac{t_n}{t_{n+1}} \left( z^{(n+1)} - x^{(n)} \right) & \text{new momentum}
\end{align*}
\]

Smaller (worst-case) convergence bound than Nesterov by 2×:

\[
\Psi(z^{(n)}) - \Psi(x^{(\star)}) \leq \frac{1L \| x^{(0)} - x^{(\star)} \|^2}{(n + 1)^2}.
\]

Recently DK found a Huber-like function for which OGM1 achieves that upper bound (thus tight), inspired by numerical work of Taylor et al. (2015).
Example: Image restoration (?!)

True $x$

Blurry $y$

Restored $\hat{x}$

Rate

$\Psi(x^{(n)}) - \Psi(\hat{x})$ vs iteration $n$
Outline

What
CT
MRI

Why
Why CT iterative
Why MRI iterative

How
Optimization transfer
Separable quadratic surrogates
Momentum
Ordered subsets

Parallelization
Ordered subsets approximation

- Data decomposition (aka incremental gradients, cf. stochastic GD):

\[ \psi(x) = \sum_{m=1}^{M} \psi_m(x), \quad \psi_m(x) \triangleq \frac{1}{2} \| y_m - A_m x \|_{W_m}^2 + \frac{1}{M} R(x) \]

1/Mth of measurements

- Key idea. For \( x \) far from minimizer: \( \nabla \psi(x) \approx M \nabla \psi_m(x) \)

- SQS:

\[ x^{(n+1)} = x^{(n)} - D^{-1} \nabla \psi(x^{(n)}) \]

- OS-SQS:

for \( n = 0, 1, \ldots \) (iteration)

for \( m = 1, \ldots, M \) (subset)

\[ k = nM + m \text{ (subiteration)} \]

\[ x^{k+1} = x^k - D^{-1} M \nabla \psi_m(x^k) \]

less work

- Coil-wise in parallel MRI

(Muckley, Noll, JF, ISMRM 2014)
Ordered subsets version of OGM1

For more acceleration, combine OGM1 with ordered subsets (OS).

OS-OGM1:
Initialize: \( t_0 = 1, \ z^{(0)} = x^{(0)} \)
for \( n = 0, 1, \ldots \) (iteration)
  for \( m = 1, \ldots, M \) (subset)

\[
k = nM + m \quad \text{(subiteration)}
\]

\[
z^{k+1} = \left[ x^k - D^{-1}M\nabla \psi_m(x^k) \right]_+ \quad \text{(typical OS-SQS)}
\]

\[
t_{k+1} = \frac{1}{2} \left( 1 + \sqrt{1 + 4t_k^2} \right)
\]

\[
x^{k+1} = z^{k+1} + \frac{t_k - 1}{t_{k+1}} \left( z^{k+1} - z^k \right) + \frac{t_k}{t_{k+1}} \left( z^{k+1} - x^k \right)
\]
OS-OGM1 properties

- Approximate convergence rate for $\Psi$: $O\left(\frac{1}{n^2M^2}\right)$  
  (Donghwan Kim and JF; CT 2014)

- Same compute per iteration as other OS methods
  (One forward / backward projection and $M$ regularizer gradients per iteration)

- Same memory as OGM1 (two more images than OS-SQS)

- Guaranteed convergence for $M = 1$

- No convergence theory for $M > 1$
  - unstable for large $M$
  - small $M$ preferable for parallelization

- Now fast enough to show X-ray CT examples...
OS-OGM1 results: data

- 3D cone-beam helical X-ray CT scan
- pitch 0.5
- image $x$: $512 \times 512 \times 109$ with 70 cm FOV and 0.625 mm slices
- sinogram: $y$ 888 detectors $\times$ 32 rows $\times$ 7146 views
OS-OGM1 results: convergence rate

Root mean square difference (RMSD) between $x^{(n)}$ and $x^{(\infty)}$ over ROI (in HU), versus iteration.

(Compute times per iteration are very similar.)
OS-OGM1 results: images

At iteration $n = 10$ with $M = 12$ subsets.
OS divergence example

- one-pixel image
- three intersecting rays
- \( A = \begin{bmatrix} 1 \\ 1 \\ 4 \end{bmatrix} \)
- \( x = 2, \ y = Ax = \begin{bmatrix} 2 \\ 2 \\ 8 \end{bmatrix} \)
- condition number of \( A'A = 1 \)
- consistent system of eqns.
OS divergence example

OS-SQS-LS for $M = 3$ subsets:

$$x^{\text{new}} = x^{\text{old}} - D^{-1}3\nabla_m x^{\text{old}} = x^{\text{old}} - D^{-1}3A'(Ax^{\text{old}} - y)$$

$$D = \text{diag}\{A'A1\} = 1^2 + 1^2 + 4^2 = 18$$

After 3 updates:

$$x^{(n+1)} - x = \left(1 - \frac{3}{18}1^2\right) \left(1 - \frac{3}{18}1^2\right) \left(1 - \frac{3}{18}4^2\right) (x^{(n)} - x)$$

$$= -2(15/18)^3 (x^{(n)} - x) = -\frac{125}{108} (x^{(n)} - x)$$

Divergence of OS-SQS-LS is possible even in well-conditioned, consistent case.
Outline

What
  CT
  MRI

Why
  Why CT iterative
  Why MRI iterative

How
  Optimization transfer
  Separable quadratic surrogates
  Momentum
  Ordered subsets

Parallelization
Amazon Cloud version of OS-OGM

Distribute long object (320 useful slices) into (overlapping) slabs (128 slices each) across 5 separate clusters, each with 10 nodes having 16 cores.

Use MPI (message passing interface) for within-cluster communication:

- Forward Projection
- Broadcast Communication
- Back Projection
- Regularization
- Update
- Broadcast Communication
- Forward Projection

Rosen, Wu, Wenisch, JF (Fully 3D, 2013)

- Overlapping slabs is inefficient
- Communication time (within cluster, after every subset) is serious bottleneck
Conventional OS approach uses a voxel-wise SQS:

$$\psi(x) \leq \psi(x^{(n)}) + \nabla \psi(x^{(n)})(x - x^{(n)}) + \frac{1}{2}(x - x^{(n)})'D(x - x^{(n)})$$

$$= \psi(x^{(n)}) + \sum_{j=1}^{N} \frac{\partial}{\partial x_j} \psi(x^{(n)})(x_j - x_j^{(n)}) + \frac{1}{2} d_j \left(x_j - x_j^{(n)}\right)^2$$

Diagonal matrix $D$ majorizes the Hessian of $\psi$: $\nabla^2 \psi(x) \preceq D$.

Distributed computing alternative: slab-separable surrogate:

$$\psi(x) - \psi(x^{(n)}) \leq \sum_{b=1}^{B} \psi_b(x_b)$$

$$\psi_b(x_b) \triangleq \nabla x_b \psi(x^{(n)})(x_b - x_b^{(n)}) + \frac{1}{2} \left(x_b - x_b^{(n)}\right)'H_b \left(x_b - x_b^{(n)}\right)$$

Block diagonal matrix $H = \text{diag}\{H_1, \ldots, H_B\}$ majorizes $\nabla^2 \psi(x)$. 
\[
\psi_b(x_b) \triangleq \nabla_{x_b} \psi(x^{(n)})(x_b - x_b^{(n)}) + \frac{1}{2} \left( x_b - x_b^{(n)} \right)' H_b \left( x_b - x_b^{(n)} \right)
\]

\[
H_b \triangleq A_b' W \Lambda_b A_b, \quad \Lambda_b \triangleq \text{diag}\{A_1 \otimes A_b 1_b\}
\]

Updates parallelizable across blocks (slabs):

\[
x_b^{(n+1)} \triangleq \text{arg min}_{x_b \succeq 0} \psi_b(x_b).
\]

- Reduces communication.
- (Apply favorite optimization method within slab.)
- (Donghwan Kim and JF; Fully 3D, 2015) [Mo18]
1: Initialize $\tilde{x}^{(0)}$ by FBP, and compute $D$.
2: Distribute image $\tilde{x}^{(0)}$ and data $y$ into $B$ nodes.
3: for $n = 0, 1, \ldots$
4:  Minimize $\phi_{BSS}(x; \tilde{x}^{(n)})$ using $L$ sub-iterations of OS-SQS-mom.
   1) Initialize $x^{(0)} = z^{(0)}$ by $\tilde{x}^{(n)}$, and $t^{(0)} = 1$.
   2) for $l = 0, 1, \ldots, L - 1$
   3) $m = l \mod M$
   4) $t^{(l+1)} = \frac{1}{2} \left( 1 + \sqrt{1 + 4 \left[ t^{(l)} \right]^2} \right)$
   5) for $b = 1, \ldots, B$ simultaneously
   6) $g_{m,b}^{(l)} = M \nabla_b \phi_{BSS,m}(z^{(\frac{l}{M})}; z^{(0)})$ [subset gradient]
   7) $x_b^{(\frac{l+1}{M})} = \left[ z_b^{(\frac{l}{M})} - D_b^{-1} g_{m,b}^{(l)} \right]$ [OS-SQS update]
   8) $z_b^{(\frac{l+1}{M})} = x_b^{(\frac{l+1}{M})} + \frac{t^{(l)}-1}{t^{(l+1)}} \left( x_b^{(\frac{l+1}{M})} - x_b^{(\frac{l}{M})} \right)$ [momentum]
   9) end for
10) end for
11) $\tilde{x}^{(n+1)} = x^{(\frac{L}{M})}$
5: Communicate $\tilde{x}^{(n+1)}$.
6: end for
BSS OS-OGM: data

- $256 \times 256 \times 160$ XCAT phantom (Segars et al., 2008)
- Simulated helical CT, $444 \times 32 \times 492$
- $M = 12$ subsets, $B = 10$ blocks, $L = 5$ inner iterations
- Matlab emulation

FBP initializer $x^{(0)}$  Converged $x^{(\infty)}$
BSS OS-OGM: rates

- Outer loop interrupts momentum
  \[ \implies \text{BSS is slower per iteration than OS-OGM} \]
- Reduced communication reduces overall time
BSS OS-OGM: images

(a) $\mathbf{x}^{(10)}$ of OS-SQS-mom ($M = 12$)

(b) Difference between (a) and $\hat{\mathbf{x}}$

(c) $\mathbf{x}^{(20)}$ of BSS ($B = 10$, $M = 12$, $L/M = 5$)

(d) Difference between (c) and $\hat{\mathbf{x}}$

- Comparable images
- Algorithm designed for distributed computation
Duality approach for using GPU

- Data transfer between system RAM and GPU can be bottleneck
- “Hide” communication time by overlapping with computation

Algorithm synopsis: (Madison McGaffin and JF; Fully 3D, 2015) [Wed. AM]

- Write cost function $\Psi(\mathbf{x})$ in terms of dual variables $\mathbf{v}$ and $\mathbf{u}$ for data-fit and regularizer:

$$
\Psi(\mathbf{x}) = \sum_{i=1}^{M} h_i([A\mathbf{x}]_i) + \sum_{k} \psi([C\mathbf{x}]_k)
$$

$$
\mathbf{x}^{(n+1)} = \arg\min_{\mathbf{x}} \sup_{\mathbf{u}, \mathbf{v}} 
$$

$$
(A' \mathbf{u} + C' \mathbf{v})' \mathbf{x} - \sum_{i=1}^{M} h_i^*(u_i) - \sum_{k} \psi^*(v_k) + \frac{\mu}{2} \| \mathbf{x} - \mathbf{x}^{(n)} \|^2_2
$$

$h_i^*$ and $\psi^*$ denote convex conjugates of $h_i$ and $\psi$

- Alternate between updating
  - several projection view dual variables $\{u_i\}$
  - dual variables for one regularization direction $\{v_k\}$

- Using dual variables “decouples” regularizer and data terms
Duality-GPU: data

- 3D cone-beam helical X-ray CT scan
- pitch 0.5
- image $\mathbf{x}$: $512 \times 512 \times 109$ with 70 cm FOV and 0.625 mm slices
- sinogram: $\mathbf{y}$ 888 detectors $\times$ 32 rows $\times$ 7146 views
- OpenCL on aging NVIDIA GTX 480 GPU with 2.5 GB RAM
  - FBP initializer $\mathbf{x}^{(0)}$
  - Converged $\mathbf{x}^{(\infty)}$
Duality-GPU: timing results

- Algorithm designed specifically for GPU architecture characteristics
- Future work:
  - combine with BSS for multiple nodes?
Duality-GPU: image results

(a) Filtered backprojection

(b) Reference

(c) OS-OGM with 4 GPUs after 8 iterations (5.2 minutes)

(d) Proposed with 4 GPUs after 5 iterations (4.8 minutes)
Summary

- Model-based image reconstruction can
  - improve image quality for low-dose X-ray CT
  - enable faster MRI scans via under-sampling
- Much more: dynamic image reconstruction, motion compensation, ...
- Computation time remains a significant challenge
- Moore’s law will not solve the problem
- Algorithms designed for distributed computation are essential
  - Block-separable surrogates to reduce communication
    (Donghwan Kim and JF; Fully 3D, 2015) [Mo18]
  - Duality approach to overlap communication with computation
    Also provides a OS-like algorithm with convergence theory
    (Madison McGaffin and JF; Fully 3D, 2015) [Wed. AM]


