Convergence Stories of Algebraic Iterative Reconstruction

Per Christian Hansen

DTU Compute, Technical University of Denmark

Joint work with

Tommy Elfving – Linköping University

Touraj Nikazad – Iran Univ. Scence & Technology

Yiqiu Dong & Nicolai A. B. Riis – DTU Compute

Back in Hong Kong

Thank you so much for the invitation to present my recent work and giving me a chance to see Hong Kong once again!

Overview of my Talk

Algebraic iterative reconstruction methods (Kaczmarz, Cimmino, etc.) are successfully used in computed tomography:

- Very flexible – no assumptions about the CT scanning geometry.
- Easy to incorporate convex constraints (e.g., nonneg./box constraints).

Both statements about these methods are true:

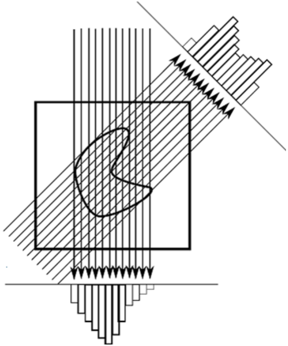
- We know a lot about the convergence – for exact data.
- We know so little about the convergence – for noisy data.

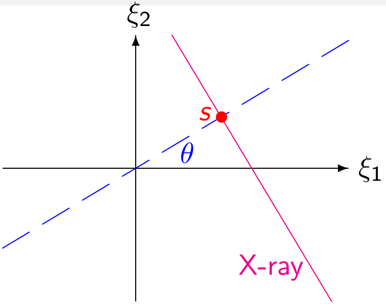
This talk tells the tale of how convergence theory is being established.

X-Ray Tomography and the Radon Transform

The Principle

Send X-rays through the object at different angles, and measure the attenuation.





$$f(\boldsymbol{\xi}) = \text{2D object/image}, \quad \boldsymbol{\xi} = \begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix}$$

$$g(\theta, s) = \mathcal{R}\, f = \text{Radon transform of } f$$

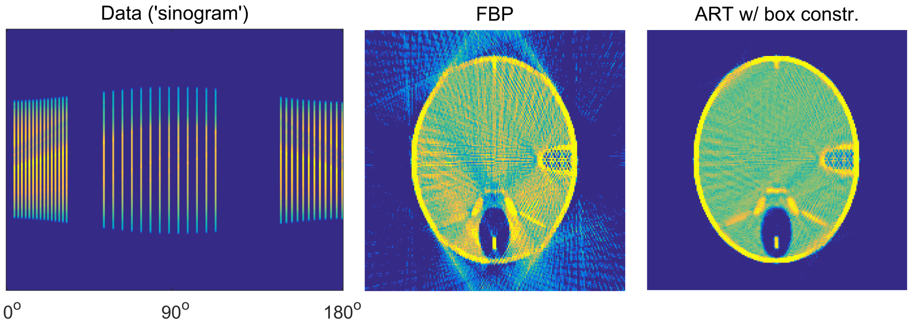
$$= \int_{-\infty}^{\infty} f\left(s \begin{bmatrix} \cos\theta \\ \sin\theta \end{bmatrix} + \tau \begin{bmatrix} -\sin\theta \\ \cos\theta \end{bmatrix} \right) d\tau$$

$$\mathcal{R}^{-1} = \text{Filtered Back Projection (FBP)}$$

Filtered Back Projection Versus Algebraic Reconstruction

- FBP: fast, low memory, good results with sufficiently many good data.
- But *artifacts* appear with noisy and/or limited data.
- Difficult to incorporate constraints (e.g., nonnegativity).
- Algebraic iterative reconstruction methods are more flexible and adaptive – but require more computational work.

Example with 3% noise and projection angles $15^\circ, 30^\circ, \ldots, 180^\circ$:



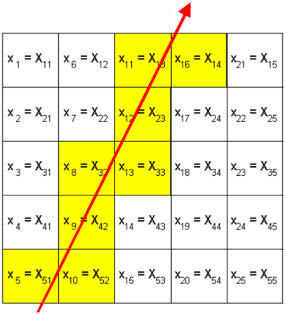
Setting Up the Algebraic Model

Damping of the ith X-ray through the domain is a line integral:

$$b_i = \int_{\text{ray}_i} f(\xi)\, d\ell, \qquad f(\xi) = \text{attenuation coef.}$$

Assume $f(\xi)$ is a constant x_j in pixel j, leading to:

$$b_i = \sum a_{ij}\, x_j, \qquad a_{ij} = \begin{cases} \text{length of ray } i \text{ in pixel } j \\ 0 \text{ otherwise.} \end{cases}$$



This leads to a linear system of equations with a LARGE and *s p a r s e* coefficient matrix:

$$\boxed{A\, x = b}$$

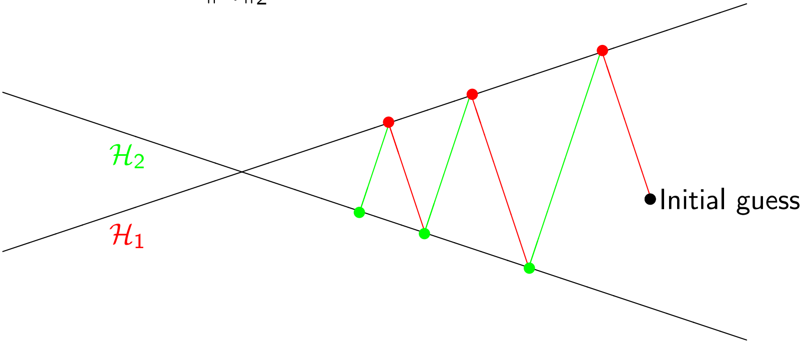
$A \sim$ measurement geometry,
$x \sim$ reconstruction,
$b \sim$ data.

ART: Algebraic Reconstruction Technique = Kaczmarz

Kaczmarz (1937): $x \leftarrow \mathcal{P}_i\, x =$ orthogonal projection on the hyperplane $\mathcal{H}_i$ defined by the ith row a_i^T of A and the corresp. element b_i of the rhs.

Repeat accessing the rows *sequentially*, e.g., in a cyclic fashion:

$$x \leftarrow \mathcal{P}_i\, x = x + \frac{b_i - a_i^T x}{\|a_i\|_2^2}\, a_i \ , \qquad i = 1, 2, \ldots, m, 1, 2, \ldots, m, 1, 2, \ldots$$



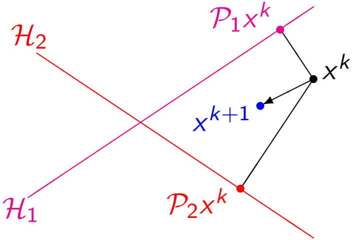
Can also access the rows in a randomized fashion.

From Sequential to Simultaneous Updates

Cimmino (1938): accesses all rows *simultaneously* and compute next iterate as the average of the all the projections of the previous iterates:

$$\begin{aligned} x^{k+1} &= \frac{1}{m}\sum_{i=1}^{m} \mathcal{P}_i x^k = \frac{1}{m}\sum_{i=1}^{m}\left(x^k + \frac{b_i - a_i^T x^k}{\|a_i\|_2^2}\, a_i\right) \\ &= x^k + \frac{1}{m}\sum_{i=1}^{m} \frac{b_i - a_i^T x^k}{\|a_i\|_2^2}\, a_i = x^k + A^T M \left(b - A\, x^k\right), \end{aligned}$$

where we introduced the diagonal matrix $M = \mathrm{diag}\big(m\|a_i\|_2^2\big)^{-1}$.



SIRT: Simultaneous Iterative Reconstruction Technique

A general of methods:

$$x^{k+1} = x^k + \omega\, D\, A^T M\, (b - A\, x^k), \qquad k = 0, 1, 2, \ldots$$

	D	M
Landweber	I	I
Projected gradient descent		
Cimmino	I	$\frac{1}{m}\,\mathrm{diag}\left(\frac{1}{\|a_i\|_2^2}\right)$
Landweber with row normalization		
CAV	I	$\mathrm{diag}\left(\frac{1}{\|a_i\|_S^2}\right)$
Component Averaging		$S = \mathrm{diag}(\mathtt{nnz}(\text{column } j))$
DROP	S^{-1}	$M = \mathrm{diag}\left(\frac{1}{\|a_i\|_2^2}\right)$
Diagonally relaxed orthogonal projection		
SART	$\mathrm{diag}(\text{row sums})^{-1}$	$M = \mathrm{diag}(\text{column sums})^{-1}$
Simultaneous algebraic reconstruction technique		

Asymptotic Convergence for Kaczmarz's Method

Galántai (2004); Strohmer and Vershynin (2009)

Assume that A is invertible and that all rows are scaled to unit 2-norm.

$$\left.\begin{aligned} \|x^k - \bar{x}\|_2^2 &\le \left(1 - \det(A)^2\right)^k \|x^0 - \bar{x}\|_2^2 \\ \mathcal{E}\left(\|x^k - \bar{x}\|_2^2\right) &\le \left(1 - \frac{1}{n\,\kappa^2}\right)^k \|x^0 - \bar{x}\|_2^2 \end{aligned}\right\} \quad k = 1, 2, \ldots\ ,$$

where $\mathcal{E}(\cdot)$ = expected value, $\bar{x} = A^{-1}b$ and $\kappa = \|A\|_2\,\|A^{-1}\|_2$.

This is linear convergence.

When κ is large we have

$$\left(1 - \frac{1}{n\,\kappa^2}\right)^k \approx 1 - \frac{k}{n\,\kappa^2}.$$

After $k = n$ updates, i.e., one "sweep," the reduction factor is $1 - 1/\kappa^2$.

Asymptotic Convergence for Cimmino (a SIRT Method)

Follows from Nesterov (2004)

Assume that A is invertible and scaled such that $\|A\|_2^2 = m$.

$$\|x^k - \bar{x}\|_2^2 \le \left(1 - \frac{2}{1+\kappa^2}\right)^k \|x^0 - \bar{x}\|_2^2$$

where $\bar{x} = A^{-1}b$ and $\kappa = \|A\|_2 \, \|A^{-1}\|_2$. Again: linear convergence.

When κ is large then we have the approximate upper bound

$$\|x^k - \bar{x}\|_2^2 \lesssim (1 - 2/\kappa^2)^k \, \|x^0 - \bar{x}\|_2^2,$$

showing that in each iteration the error is reduced by a factor $1 - 2/\kappa^2$.

Almost the same factor as in one "sweep" in Kaczmarz's method.

Real Problems Have Noisy Data

A standard topic of SIAM ALA conferences: solve $A\,x = b$.

Don't do this for inverse problems with noisy data!

The right-hand side b (the data) is a sum of noise-free data $\bar{b} = A\,\bar{x}$ from the ground-truth image $\bar{x}$ plus a noise component e:

$$b = A\,\bar{x} + e, \qquad \bar{x} = \text{ground truth}, \qquad e = \text{noise}.$$

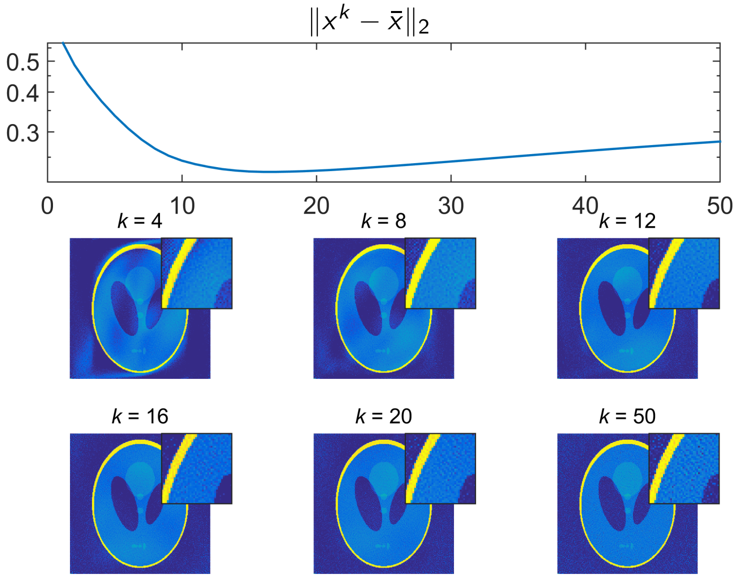
The naïve solution $x^{\text{naïve}} = A^{-1}b$ is undesired, because it has a large component coming from the noise in the data:

$$x^{\text{naïve}} = A^{-1}b = A^{-1}(A\,\bar{x} + e) = \bar{x} + A^{-1}e.$$

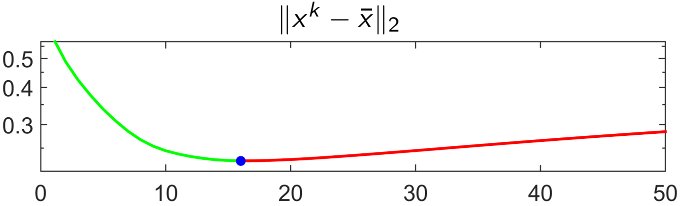
The component $A^{-1}e$ dominates over $\bar{x}$, because A is ill conditioned.

But something interesting happens during the iterations . . .

The Reconstruction Error for Kaczmarz's Method



Semi-Convergence



- In the initial iterations x^k approaches the unknown ground truth $\bar{x}$.
- During later iterations x^k converges to the undesired $x^{\text{naïve}} = A^{-1}b$.
- Stop the iterations when the convergence behavior changes.

Then we achieve a regularized solution: an approximation to the noise-free solution which is not too perturbed by the noise in the data.

- Today we explain *why* we have semi-convergence for noisy data.
- How to stop the iterations at the right time is a *different story*.

Analysis of Semi-Convergence for SIRT

Consider SIRT with $D = I$ and the SVD $\boxed{M^{\frac{1}{2}} A = \sum_{i=1}^n u_i \, \sigma_i \, v_i^T}$

Van der Sluis & Van der Vorst (1990)

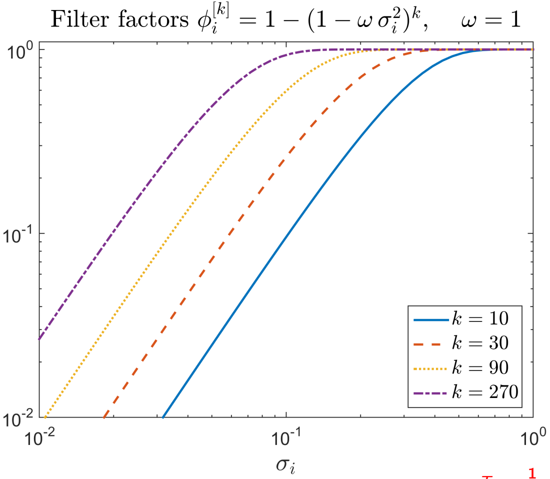
The iterate x^k is a filtered SVD solution:

$$x^k = \sum_{i=1}^n \varphi_i^{[k]} \, \frac{u_i^T (M^{\frac{1}{2}} b)}{\sigma_i} \, v_i, \qquad \varphi_i^{[k]} = 1 - \left(1 - \omega \, \sigma_i^2\right)^k .$$

Recall that we solve *noisy* systems $A\,x = b$ with $b = A\,\bar{x} + e$. Then:

$$x^k - \bar{x} = \underbrace{\sum_{i=1}^n \varphi_i^{[k]} \, \frac{u_i^T (M^{\frac{1}{2}} e)}{\sigma_i} \, v_i}_{\text{noise error}} - \underbrace{\sum_{i=1}^n (1 - \varphi_i^{[k]}) \, v_i^T \, \bar{x} \, v_i}_{\text{iteration error}} .$$

Fact: the iteration error decreases. Aim: show that noise error increases.



The filter factors *dampen* the "inverted noise" in $\varphi_i^{[k]} \, \dfrac{u_i^T (M^{\frac{1}{2}} e)}{\sigma_i}$.

Note: $\omega\,\sigma_i^2 \ll 1 \Rightarrow \varphi_i^{[k]} \approx k\,\omega\,\sigma_i^2$ showing that k and ω play the same role.

The Spectral Behavior of the Noise Error

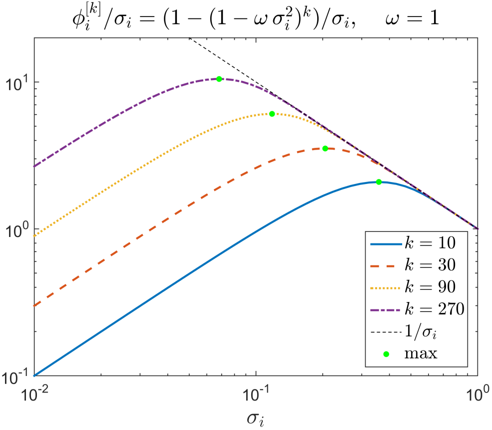
Recall: the noise error $=$

$\sum_{i=1}^{n} \frac{\varphi_i^{[k]}}{\sigma_i} u_i^T (M^{\frac{1}{2}} e) \, v_i$

and v_i is a *spectral basis:*

▷ large $\sigma_i \sim$ low-freq. v_i

▷ small $\sigma_i \sim$ high-freq. v_i



- Each curve has a maximum for $\sigma_i \approx 1.12/\sqrt{k\,\omega}$.
- As k increases, more noise is included and the SVD-spectrum changes.
- As k increases, the noise error gets dominated by higher frequencies.

Constrained Problems

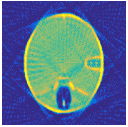
In many applications we can improved the reconstruction by including simple constraints:

$$\boxed{\min_x \|A\,x - b\|_2 \quad \text{s.t.} \quad x \in \mathcal{C}}$$

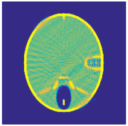
where $\mathcal{C}$ is a convex set, e.g.,

- $\mathcal{C} = \mathbb{R}^n$ – nonnegativity constraints.
- $\mathcal{C} = [0,1]^n$ – box constraints.

No constr.



Box constr.



Kaczmarz (ART) with projection:

$$x \leftarrow \mathcal{P}_{\mathcal{C}}\left(x + \omega\,\frac{b_i - a_i^T x}{\|a_i\|_2^2}\,a_i\right)\;, \qquad i = 1,2,3,\ldots$$

SIRT with projection:

$$x^{k+1} = \mathcal{P}_{\mathcal{C}}\left(x^k + \omega\,D\,A^T M\,(b - A\,x^k)\right)\,, \qquad k = 0,1,2,\ldots$$

Analysis of Semi-Convergence for Projected SIRT

For *constrained* problems we cannot perform an SVD analysis.

Let $\bar{x}$ be the solution to the noise-free problem:

$$\bar{x} = \text{argmin}_{x \in \mathcal{C}} \|A\,x - \bar{b}\|_M , \qquad \bar{b} = A\,\bar{x} = \text{pure data}$$

and let $\bar{x}^k$ denote the iterates when applying SIRT to $\bar{b}$. Then

$$\|x^k - \bar{x}\|_2 \leq \underbrace{\|x^k - \bar{x}^k\|_2}_{\text{noise error}} + \underbrace{\|\bar{x}^k - \bar{x}\|_2}_{\text{iteration error}} .$$

We already considered the decreasing iteration error:

$$\|\bar{x}^k - \bar{x}\|_2 \lesssim (1 - 2/\kappa^2)^k \, \|x^0 - \bar{x}\|_2^2 .$$

Now we must consider the noise error (which we expect to grow with k).

The Noise Error for Projected SIRT

Elfving, H, Nikazad (2012)

The noise error in Projected SIRT is bounded by

$$\| x^k - \bar{x}^k \|_2 \leq \frac{\sigma_1}{\sigma_n} \, \frac{1 - (1 - \omega \, \sigma_n^2)^k}{\sigma_n} \, \| M^{\frac{1}{2}} e \|_2 \ .$$

As long as $\omega \, \sigma_n^2 \ll 1$ we have $1 - (1 - \omega \, \sigma_n^2)^k \approx k \, \omega \, \sigma_n$ and thus

$$\boxed{\| x^k - \bar{x}^k \|_2 \lesssim \omega \, k \, \sigma_1 \| M^{\frac{1}{2}} e \|_2 \ ,}$$

showing again that k and ω play the same role in the error bound.

Sketch of Proof

By definition: $x^k - \bar{x}^k = \mathcal{P}_{\mathcal{C}}(\ldots x^{k-1} \ldots) - \mathcal{P}_{\mathcal{C}}(\ldots \bar{x}^{k-1} \ldots)$.

Using that $\mathcal{P}_{\mathcal{C}}$ is nonexpansive:

$$\begin{aligned} \varepsilon^k &= \|x^k - \bar{x}^k\|_2 = \|\mathcal{P}_{\mathcal{C}}(\ldots x^{k-1} \ldots) - \mathcal{P}_{\mathcal{C}}(\ldots \bar{x}^{k-1} \ldots)\|_2 \\ &\leq \|\ldots x^{k-1} \ldots - \ldots \bar{x}^{k-1} \ldots\|_2 \\ &= \|(I - \omega\, A^T M\, A)\,(x^{k-1} - \bar{x}^{k-1}) + \omega\, A^T M\, e\|_2 \\ &\leq \|I - \omega\, A^T M\, A\|_2\, \varepsilon^{k-1} + \omega\, \|A^T M\, e\|_2 \ . \end{aligned}$$

Induction & $x^0 = \bar{x}^0$ & $\|I - \omega\, A^T M\, A\|_2 \leq 1 - \omega\, \sigma_n^2$:

$$\begin{aligned} \varepsilon^k &\leq \omega\, \sigma_1\, \|M^{\frac{1}{2}} e\|_2 \textstyle\sum_{s=0}^{k-1} \|I - \omega\, A^T M\, A\|_2^s \\ &= \omega\, \sigma_1\, \|M^{\frac{1}{2}} e\|_2\, \frac{1 - \|I - \omega\, A^T M\, A\|_2^k}{1 - \|I - \omega\, A^T M\, A\|_2} \\ &= \omega\, \sigma_1\, \|M^{\frac{1}{2}} e\|_2\, \frac{1 - (1 - \omega\, \sigma_n^2)^k}{\omega\, \sigma_n^2} \ . \end{aligned}$$

Analysis of Semi-Convergence for ART – Setting the Stage

Elfving, Nikazad (2009)

An ART "sweep" can be written in a form that *resembles* SIRT:

$$x^{k+1} = x^k + \omega\, A^T \widehat{M}\, (b - A\, x^k)\,, \qquad \widehat{M} = (\Delta + \omega\, L)^{-1}\,.$$

where the **nonsymmetric** $\widehat{M}$ comes from the splitting:

$$A\, A^T = L + \Delta + L^T, \qquad \Delta = \mathrm{diag}(\|a_i\|_2^2)\,,$$

where L is strictly lower triangular.

Simple manipulations show that the noise error is given by

$$\begin{aligned} x^k - \bar{x}^k &= (I - \omega\, A^T \widehat{M}\, A)\, (x^{k-1} - \bar{x}^{k-1}) + \omega\, A^T \widehat{M}\, e \\ &= \omega \sum_{j=1}^{k-1} (I - \omega\, A^T \widehat{M}\, A)^j\, A^T \widehat{M}\, e\,. \end{aligned}$$

Analysis of Semi-Convergence for ART – Results

Elfving, H, Nikazad (2014)

Let $\delta = \|A^T \widehat{M}\, e\|_2$ and σ_r = smallest nonzero singular value of A.

We obtain a bound which resembles that of SIRT:

$$\boxed{\|x^k - \bar{x}^k\|_2 \leq \omega\, k\, \delta + O(\sigma_r^2)}$$

As long as $\omega\, \sigma_r^2 < 1$ we have:

$$\boxed{\|x^k - \bar{x}^k\|_2 \leq \frac{\sqrt{\omega}}{\sigma_r} \sqrt{k}\, \delta + O(\sigma_r^2)}$$

These results also hold for constrained problems, provided that

$$y \in \mathcal{R}(A^T) \quad \Rightarrow \quad \mathcal{P}_{\mathcal{C}}(y) \in \mathcal{R}(A^T)\ .$$

Proof: much too technical for this talk.

Numerical Results – Parallel-Beam X-Ray Tomography

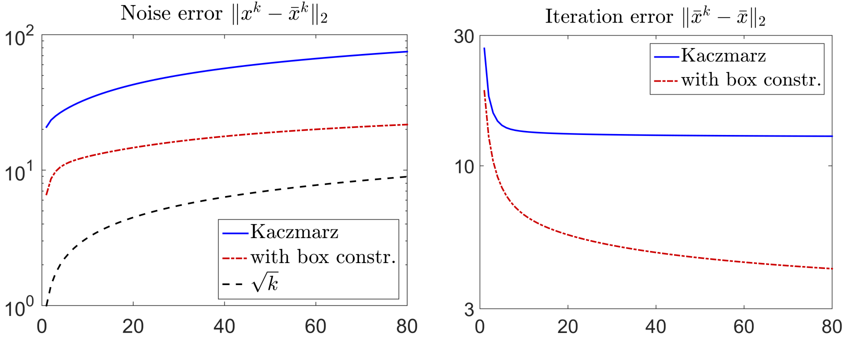
Test problem

- 200×200 phantom
- 60 projections at
- $3^\circ, 6^\circ, 9^\circ, \ldots, 180^\circ$
- $m = 15\,232$
- $n = 40\,000$

The upper bound. We estimate

$$\frac{\sqrt{\omega}}{\sigma_r}\,\delta \approx 10^7 .$$

Our bound $\frac{\sqrt{\omega}}{\sigma_r}\delta\sqrt{k}$ is a huge over-estimate; the factor $\sqrt{k}$ correctly *tracks* the noise error.



And Now: Matrix Perturbations

Motivations for studying how the convergence is influenced by matrix perturbations:

- Academic reason: "nice to have."
- Practical reason: "the backprojector is not the transpose."

Recall: the matrix A is not stored explicitly. Multiplications with A and A^T are computed "on the fly" using hardware accelerators.

But many software packages implement the backprojector in such a way that it is *not* the exact transposed of the forward projector A.

- Depends on the application and the traditions.
- Better utilization of hardware: multi-core processors, GPUs, etc.

> What is the influence of unmatched projector/backprojector pairs on the computed solutions and the convergence of the iterations?

Perturbation Theory for Unmatched Normal Equations

Let $\{A, A^T, \bar{b}\}$ be the unperturbed data, and consider the perturbations

$$\tilde{A} = A + E_A, \qquad \hat{A}^T = A^T + E_{A^T}, \qquad b = \bar{b} + e.$$

Also let $\bar{x}$ denote the unperturbed solution to $A^T A\,\bar{x} = A^T \bar{b}$.

Elfving, H (2018)

When we use the perturbed triple $\{\tilde{A}, \hat{A}^T, b\}$ then we aim at solving the *unmatched normal equations:*

$$\hat{A}^T \tilde{A}\,(\bar{x} + \delta x) = \hat{A}^T b.$$

Skipping higher-order terms, we obtain:

$$\boxed{\|\delta x\|_2 \leq \frac{1}{\sigma_r}\big(\|\mathcal{P}_{\mathcal{R}(A)}\, e\|_2 + \|E_A \bar{x}\|_2\big) + \frac{1}{\sigma_r^2}\|E_{A^T}(\bar{b} - A\,\bar{x})\|_2}$$

For inconsistent systems, the solution is more sensitive to E_{A^T} than E_A.

Convergence Analysis for Unmatched Pairs

To study many different cases we consider the generic **BA Iteration**

$$x^{k+1} = x^k + \omega\, B\,(b - A\,x^k)\ , \qquad \omega > 0\ .$$

- Any fixed point x^* satisfies $BA\,x^* = B\,b$.
- If $B\,A$ is invertible then $x^* = (BA)^{-1}B\,b$.
- If $\mathcal{N}(BA) = \mathcal{N}(A)$ and $b \in \mathcal{R}(A)$ then $A\,x^* = b$.

Shi, Wei, Zhang (2011); Elfving, H (2018)

The **BA Iteration** converges to a solution of $BA\,x = B\,b$ if and only if

$$0 < \omega < \frac{2\,\Re(\lambda_j)}{|\lambda_j|^2} \quad \text{and} \quad \Re(\lambda_j) > 0, \qquad \{\lambda_j\} = \mathtt{eig}(BA)\ .$$

Zeng & Gullberg (2000): similar analysis but ignoring complex λ_j.

Noise Error for BA Iteration

From the definition of the **BA Iteration** it follows that

$$x^k - \bar{x}^k = (I - \omega BA)\,(x^{k-1} - \bar{x}^{k-1}) + \omega\, B\, e\ ,$$

and hence by induction, and assuming $x^0 = \bar{x}^0 = 0$,

$$x^k - \bar{x}^k = \omega \sum_{j=0}^{k-1} (I - \omega BA)^j B\, e\ .$$

Elfving, H (2018)

Similar to ART and SIRT, with $b = A\,\bar{x} + e$ we have

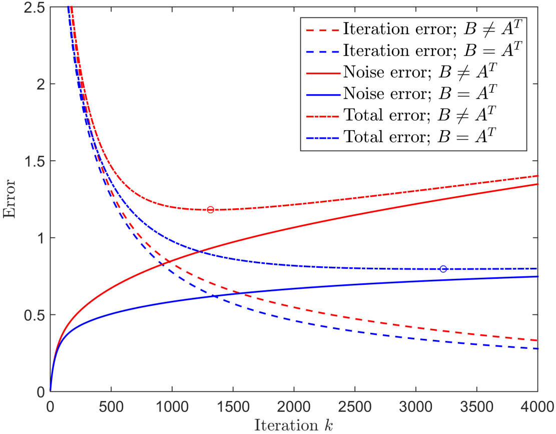
$$\boxed{\|x^k - \bar{x}^k\|_2 \ \leq\ (\omega\, c_{\mathrm{BA}} \|B\|_2)\, k\, \|e\|_2}$$

where we define the constant c_{BA} by: $\sup_j \|(I - \omega BA)^j\|_2 \leq c_{\mathrm{BA}}$.

Cimmino's method.

Test problem

- ▷ 64×64 phantom
- ▷ 180 projections at
- ▷ $1^\circ, 2^\circ, 3^\circ, \ldots, 180^\circ$
- ▷ $m = 16\,380$
- ▷ $n = 4\,096$



Iteration error: both versions converge to $\bar{x}$; the one with $B \neq A^T$ is slower.

Noise error: the one for $B \neq A^T$ increases faster.

Total error: semi-convergence, the iteration with $B \neq A^T$ reaches the min. error ○ 1.181 after 1314 iterations. This error is 48% larger than the min. error ○ 0.796 for the iterations with A^T, achieved after 3225 iterations.

Did We Prove Semi-Convergence?

Not really:

- we give an *upper* bound for the noise error;
- this bound increases with k,
- and it seems to track the actual noise error in numerical experiments.

Thus we have justified the observed behavior of

$$\text{total error} = \text{iteration error} + \text{noise error}.$$

But we also need a *lower* bound that increases with k:

- If the right-hand side error $e \in \mathcal{N}(A)$ then the lower bound is 0 (this is extremely unlikely).
- We (currently) don't know how to derive a nonzero increasing lower bound for the case $e \notin \mathcal{N}(A)$.

Conclusion

Interesting stuff not covered today:

- Convergence of column-action methods.
- Connections to first-order optimization methods.
- Pre-asymptotic convergence of ART; Jiao, Jin, Lu (2017).
- Choice of relaxation parameters; stopping rules.

What we did cover today:

- Review of the convergence for noise-free data (iteration error).
- Illustration of semi-convergence.
- New convergence results (upper bounds) for the noise error, for several important methods.
- New results for matrix/model errors $\rightarrow$ same behavior of the noise error $\rightarrow$ semi-convergence.