Randomized iterative methods for linear systems and inverting matrices

Robert Mansel Gower

Joint work with Peter Richtárik

University of Edinburgh

Cambridge, January 2016

RMG and Peter Richtárik
Randomized Iterative Methods for Linear Systems
SIAM. J. Matrix Anal. & Appl., 36(4), 1660–1690, 2015

RMG and Peter Richtárik
Stochastic Dual Ascent for Solving Linear Systems
Preprint arXiv:1512.06890, 2015

RMG and Peter Richtárik
Stochastic Iterative Matrix Inversion
In progress, 2016

Linear Systems

The Problem

$\in \mathbb{R}^n$

$$m\left\{\overbrace{A}^{n}\right. x = b \left.\right\} m$$

Assumption: The system is consistent (i.e., has a solution)

We can also think of this as *m* linear equations, where the i[th] equation looks as follows:

$$\sum_{j=1}^{n} A_{ij} x_j = b_i$$

$$A_{i:} x = b_i$$

The Problem

$$\langle x, y \rangle_B := x^T B y, \qquad \|x\|_B := \sqrt{\langle x, x \rangle_B}$$

B: Symmetric and positive definite

$$x^* := \arg\min ||x||_B^2 \quad \text{subject to} \quad Ax = b$$

Insight: As there are possibly multiple solutions, we compute the solution with the least B-norm.

Standard Randomized Methods

The return of old methods

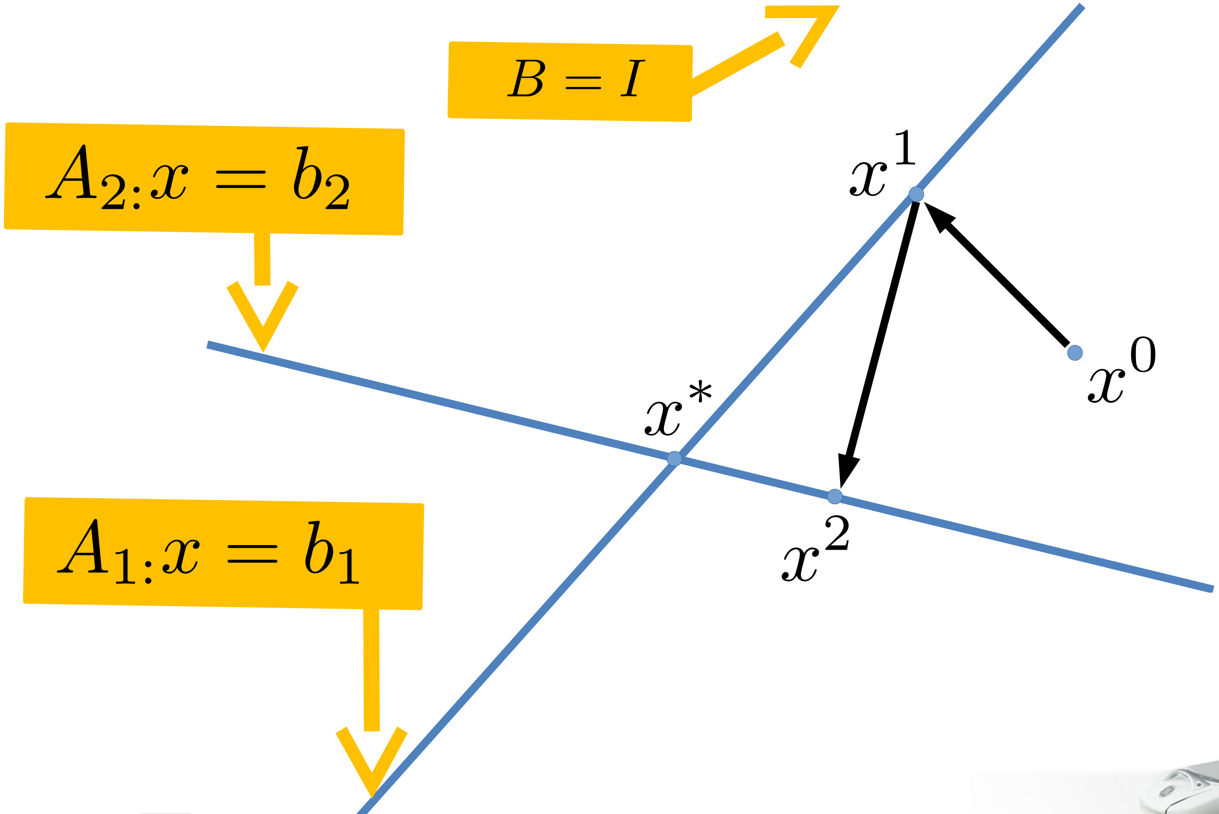
Old methods (Kaczmarz 1937, Guass-Seidel 1823) make a randomized return, why?

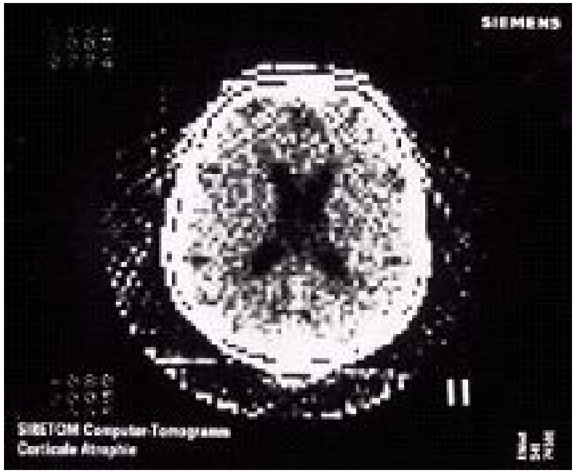
- Often suitable for Big Data problems (short recurrence, low iteration cost, low memory, block variants...etc)
- Easy to implement
- Easy to analyse, good complexity
- Often fits in parallel/distributed architecture

Randomized Kaczmarz

Karczmarz, M. S. (1937). **Angenäherte Auflösung von Systemen linearer Gleichungen**. *Bulletin International de l'Académie Polonaise Des Sciences et Des Lettres*, 35, 355–357.

$$x^{t+1} = \arg\min ||x - x^t||_2^2 \quad \text{subject to} \quad A_{i:}x = b_i$$





G.N. Hounsfield. Computerized transverse axial scanning (tomography): Part I. description of the system. British Journal Radiology. 1973

Framework for Randomized Methods

1. Relaxation Viewpoint "Sketch and Project"

$$\langle x, y \rangle_B := x^T B y, \qquad \|x\|_B := \sqrt{\langle x, x \rangle_B}$$

B: Symmetric and positive definite

$$x^{t+1} = \arg\min_{x \in \mathbb{R}^n} \|x - x^t\|_B^2$$

$$\text{subject to} \quad S^T A x = S^T b$$

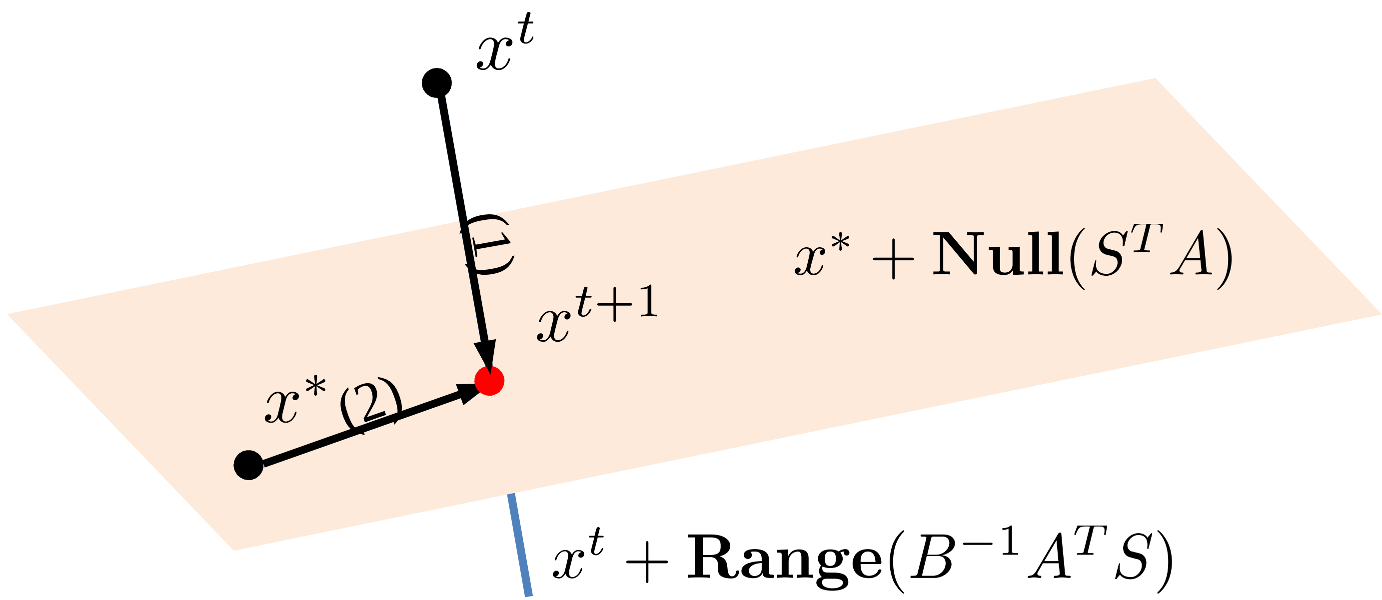
$$S^T \quad A \quad = \quad S^T A$$

S: random $m \times \tau$ matrix

2. Optimization Viewpoint
“Constrain and Approximate”

$$
\begin{aligned}
x^{t+1} \quad &= \quad \arg\min_{x \in \mathbb{R}^n} \|x - x^*\|_B^2 \\
\text{subject to} \quad & x = x^t + B^{-1}A^T S y \\
& y \quad \text{is free}
\end{aligned}
$$

3. Geometric Viewpoint "Random Intersect"



$$(1) \quad x^{t+1} = \arg\min ||x - x^t||_B^2 \quad \text{subject to} \quad S^T A x = S^T b$$

$$(2) \quad x^{t+1} = \arg\min ||x - x^*||_B^2 \quad \text{subject to} \quad x = x^t + B^{-1} A^T S y$$

$$\{x^{t+1}\} \quad = \quad \left(x^* + \mathbf{Null}(S^T A)\right) \quad \bigcap \quad \left(x^t + \mathbf{Range}(B^{-1} A^T S)\right)$$

4. Algebraic Viewpoint
“Random Linear Solve”

$$x^{t+1} \quad = \quad \text{solution in } x \text{ of the linear system}$$

$$S^T A x = S^T b$$

$$x = x^t + B^{-1} A^T S y$$

Unknown: *x*

Unknown: *y*

5. Algebraic Viewpoint "Random Update"

$$x^{t+1} = x^t - \overbrace{B^{-1}A^T S\underbrace{(S^T A B^{-1} A^T S)^\dagger}_{\text{Small } \tau \times \tau \text{ matrix}} S^T (Ax^t - b)}^{\text{Random Update Vector}}$$

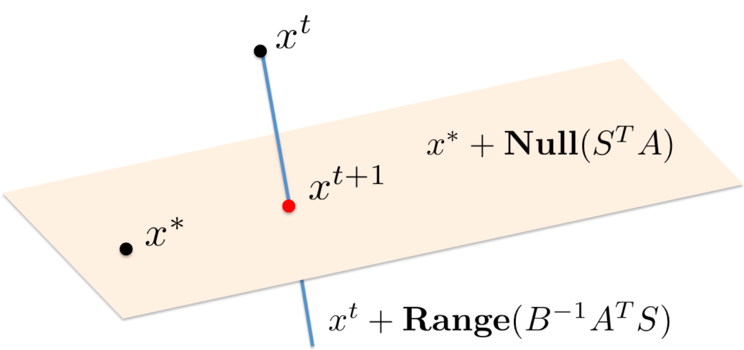
Moore-Penrose pseudo inverse

Fact: Every (not necessarily square) real matrix M has a real pseudo-inverse $M^\dagger$.

6. Analytic Viewpoint "Random Fixed Point"

$$H := S(S^T A B^{-1} A^T S)^\dagger S^T$$

$$x^{t+1} - x^* = \underbrace{(I - B^{-1} A^T H A)}_{\text{Random Iteration Matrix}}(x^t - x^*)$$



$B^{-1} A^T H A$ projects orthogonally onto $\mathbf{Range}(B^{-1} A^T S)$
$I - B^{-1} A^T H A$ projects orthogonally onto $\mathbf{Null}(S^T A)$

Theory

Complexity / Convergence

Theorem [GR'15]

$$\mathbf{E}[x^{t+1} - x^*] = (I - B^{-1}A^T\mathbf{E}[H]A)\mathbf{E}[x^t - x^*]$$

$$x^0 = 0 \quad \text{and} \quad \mathbf{E}[H] \succ 0 \quad \Rightarrow$$

1 $$||\mathbf{E}[x^t - x^*]||_B \leq \rho^t ||x^0 - x^*||_B$$

$$\rho := 1 - \lambda_{\min}^+(B^{-1/2}A^T\mathbf{E}[H]AB^{-1/2})$$

$\lambda_{\min}^+(A) :=$ smallest nonzero eigenvalue

2 $$\mathbf{E}[||x^t - x^*||_B^2] \leq \rho^t ||x^0 - x^*||_B^2$$

Proof of 1 for *A* full column rank

$$\mathbf{E}[x^{t+1} - x^*] = (I - B^{-1}A^T\mathbf{E}[H]A)\mathbf{E}[x^t - x^*]$$

Taking expectations conditioned on x^t, we get

$$\mathbf{E}[x^{t+1} - x^* \,|\, x^t] = (I - B^{-1}A^T\mathbf{E}[H]A)(x^t - x^*)$$

Taking expectation again gives

$$\begin{aligned}\mathbf{E}[x^{t+1} - x^*] &= \mathbf{E}[\mathbf{E}[x^{t+1} - x^* \,|\, x^t]] \\ &= \mathbf{E}[(I - B^{-1}A^T\mathbf{E}[H]A)(x^t - x^*)] \\ &= (I - B^{-1}A^T\mathbf{E}[H]A)\mathbf{E}[x^t - x^*]\end{aligned}$$

Applying norms to both sides we obtain

$$||\mathbf{E}[x^{t+1} - x^*]||_B \leq \underbrace{\boxed{||I - B^{-1}A^T\mathbf{E}[H]A||_B}}_{\rho} ||\mathbf{E}[x^t - x^*]||_B$$

Case study of $\mathbf{E}[H]$

$$H := \boxed{S}\boxed{(S^T A B^{-1} A^T S)^\dagger}\boxed{S^T}$$

Special Choice of Parameters

$$\mathbf{P}(S = e_i) = \frac{1}{m}$$

$$B = I$$

$$S = e^i$$

$$\mathbf{E}[H] = \frac{1}{m}\sum_{i=1}^{m} \frac{\boxed{e_i e_i^T}}{\boxed{||A_{i:}||_2^2}}$$

$$= \mathrm{diag}(||A_{i:}||_2^2)$$

No zero rows in A

$\mathbf{E}[H]$ is positive definite

Weak assumption

The rate: lower and upper bounds

Theorem [RG‘15]

$$\mathbf{E}[H] \succ 0$$

$$\Downarrow$$

$$0 \leq 1 - \frac{\mathbf{E}[\mathbf{Rank}(S^T A)]}{\mathbf{Rank}(A)} \leq \rho \leq 1$$

Insight: The method is a *contraction* (without any assumptions on *S* whatsoever). That is, things can not get worse.

Insight: The lower bound on the rate is better for *A* low rank and when the dimension of the search space in the “constrain and approximate” viewpoint grows.

Special Case: Randomized Kaczmarz Method

Randomized Kaczmarz: derivation and rate

General Method

$$x^{t+1} = x^t - \boxed{B^{-1}A^TS}\,\boxed{(S^TAB^{-1}A^TS)^{\dagger}}\,\boxed{S^T(Ax^t - b)}$$

Special Choice of Parameters

$$\mathbf{P}(S = e_i) = p_i$$

$$B = I$$

$$S = e_i$$

$$x^{t+1} = x^t - \frac{\boxed{A_{i:}x^t - b_i}}{\boxed{\|A_{i:}\|_2^2}}\boxed{(A_{i:})^T}$$

Complexity Rate. All rows of A are nonzero $\Rightarrow \mathbf{E}[H]$ is nonsingular

$$p_i = \frac{\|A_{i:}\|^2}{\|A\|_F^2}$$

$$\mathbf{E}\left[\|x^t - x^*\|_2^2\right] \leq \left(1 - \frac{\lambda_{\min}(A^TA)}{\|A\|_F^2}\right)^t \|x^0 - x^*\|_2^2$$

Special Case: Randomized Coordinate Descent

Randomized Coordinate Descent: derivation and rate

General Method

$$x^{t+1} = x^t - \boxed{B^{-1}A^T S}\,\boxed{(S^T A B^{-1} A^T S)^\dagger}\,\boxed{S^T(Ax^t - b)}$$

Special Choice of Parameters

positve definite

$$B = A$$

$$\mathbf{P}(S = e_i) = p_i$$

$$S = e_i$$

$$x^{t+1} = x^t - \frac{\boxed{(A_{i:})^T x^t - b_i}}{\boxed{A_{ii}}}\,\boxed{e^i}$$

Complexity Rate

$$p_i = \frac{A_{ii}}{\mathbf{Tr}(A)}$$

$$\mathbf{E}\left[\|x^t - x^*\|_A^2\right] \le \left(1 - \frac{\lambda_{\min}(A)}{\mathbf{Tr}(A)}\right)^t \|x^0 - x^*\|_A^2$$

Theory recovers known and new convergence results

Method	B	S	Convergence Rate ρ
Randomized CD Least square	$A^T A$	$P(S = e_i) = \frac{\|\|A_{:i}\|\|_2^2}{\|\|A\|\|_F^2}$	$1 - \frac{\lambda_{\min}(A^T A)}{\|\|A\|\|_F^2}^*$
Gaussian psd	A	$S \sim \mathcal{N}(0, I)$	$1 - \frac{2}{\pi}\frac{\lambda_{\min}(A^T A)}{\|\|A\|\|_F^2}$
Gaussian Kaczmarz	I	$S \sim \mathcal{N}(0, I)$	$1 - \frac{2}{\pi}\frac{\lambda_{\min}(A^T A)}{\|\|A\|\|_F^2}$

*Leventhal, D., & Lewis, A. S. (2010). **Randomized Methods for Linear Constraints: Convergence Rates and Conditioning**. Mathematics of Operations Research, 35(3), 641–654.

Convenient probability

Theorem [GR‘15]

$$\bar{S} := [S_1, \ldots, S_r] \quad \text{is nonsingular}$$

$$\mathbf{P}(S = S_i) = p_i = \frac{\mathbf{Tr}(S_i^T A B^{-1} A^T S_i)}{\mathbf{Tr}(\bar{S}^T A B^{-1} A^T \bar{S})}$$

$$\Downarrow$$

$$\rho = 1 - \frac{1}{\kappa^2(W^{1/2} A^T \bar{S})}$$

$$\kappa(W^{1/2} A^T \bar{S}) := ||(W^{1/2} A^T \bar{S})^{-1}||_2 ||W^{1/2} A^T \bar{S}||_F$$

Conclusion for linear systems

- **Unites** many randomized methods under a single framework
- **Improved convergence:** New lower bound, less assumptions, RK convergence without full rank assumption.
- **Design new methods:** S = Guassian, count-sketch, Walsh-Hadamard ...etc
- **Optimal Sampling:** We can choose a sampling that optimizes the convergence rate.

Inverting a Matrix

The Problem

$\in \mathbb{R}^{n\times n}$

Identity matrix

$$n\left\{\overbrace{A}^{n}\right. X = I$$

Assumption: The matrix *A* is nonsingular

Why iteratively invert a matrix?

- Needed to calculate a Schur complement or a projection operator
- Iterative methods are good when we can tolerate an error or have an initial guess $X^0 \approx A^{-1}$
- Staging for **randomized variable metric** methods and **randomized preconditioning**

Randomized Methods for Nonsymmetric Matrices

Equivalence to solving linear systems

$$\langle X, Y \rangle_{F(B)} = \mathbf{Tr}(X^T B Y B), \quad ||X||_{F(B)} = \sqrt{\langle X, X \rangle_{F(B)}}$$

B: Symmetric and positive definite

$$X^{t+1} = \arg \min_{X \in \mathbb{R}^{n \times n}} ||X - X^t||^2_{F(B)}$$
$$\text{subject to} \quad S^T A X = S^T$$

S: random $n \times \tau$ matrix

$$X^{t+1} = X^t - B^{-1} A^T S (S^T A B^{-1} A^T S)^{\dagger} S^T (A X^t - I)$$

This method is equivalent to the sketch and project method for solving linear systems, but applied simultanously to the *n* equations defined by *AX* = *I*

Randomized Methods for Symmetric Matrices

$$A = A^T$$

Sketch and Project

$$X^{t+1} = \arg \min_{X \in \mathbb{R}^{n \times n}} ||X - X^t||^2_{F(B)}$$

$$\text{subject to} \quad S^T AX = S^T, \quad X = X^T$$

Connection to quasi-Newton Methods: This is a randomized block extension of the quasi-Newton updates. In the quasi-Newton setting

$$S = \delta \in \mathbb{R}^n \quad \text{and} \quad \gamma := A\delta$$

and A is an *unknown* operator. However, we can sample its action $A\delta$ *and*

$$S^T AX = S^T \Leftrightarrow X\gamma = \delta$$

is known as the *secant equation*

Goldfarb, D. (1970). **A Family of Variable-Metric Methods Derived by Variational Means**. Mathematics of Computation, 24(109), 23.

Constrain and Approximate

$$X^{t+1} = \arg \min_{X \in \mathbb{R}^{n \times n}} ||X - A^{-1}||^2_{F(B)}$$

$$\text{subject to} \quad X = X^t + YS^T AB^{-1} + B^{-1} A^T S Y^T$$

$$Y \in \mathbb{R}^{n \times \tau} \text{ is free}$$

Duality: This is dual problem of the sketch and project viewpoint, new insight into quasi-Newton methods.

New viewpoint for BFGS

Sketch and project

$$X^{t+1} = \arg \min_{X \in \mathbb{R}^{n \times n}} ||X - X^t||^2_{F(A)}$$

$$\text{subject to} \quad X\gamma = \delta, \quad X = X^T$$

Constrain and approximate

$$X^{t+1} = \arg \min_{X \in \mathbb{R}^{n \times n}} ||AX - I||^2_F$$

$$\text{subject to} \quad X = X^t + y\delta^T + \delta y^T$$

$$y \in \mathbb{R}^n \text{ is free}$$

Duality: The BFGS minimizes a residual restricted to an affine space of symmetric matrices

$$H := S(S^T A B^{-1} A^T S)^{\dagger} S^T$$

Random Update

$$X^{t+1} = X^t - (X^t A - I) H A B^{-1} \\ + B^{-1} A H (A X^t - I) \left(A H A B^{-1} - I \right)$$

Low rank 3 X τ update

Random Fixed Point

$$X^{t+1} - A^{-1} = \\ (I - B^{-1} A^T H A)(X^t - A^{-1})(I - A H A^T B^{-1})$$

Complexity / Convergence

Theorem [GR'16]

$$||A||_B = ||B^{1/2}AB^{1/2}||_2$$

1. $$||\mathbf{E}[X^t - A^{-1}]||_B \leq \rho^t \, ||X^0 - A^{-1}||_B$$

2. $$\mathbf{E}[H] \succ 0 \quad \Rightarrow \quad 0 \leq \rho < 1$$

$$\Downarrow$$

$$\mathbf{E}[||X^t - A^{-1}||_{F(B)}^2] \leq \rho^t \, ||X^0 - A^{-1}||_{F(B)}^2$$

Special Case: Randomized Block BFGS

Randomized BFGS

$$X^{t+1} = \arg\min_{X \in \mathbb{R}^{n\times n}} ||X - X^t||_{F(A)}^2$$
$$\text{subject to} \quad S^T A X = S^T, \quad X = X^T$$

Special Choice of Parameters

positve definite $\Rightarrow B = A$

$\mathbf{P}(S = e_i) = p_i \Rightarrow S = e_i$

$$X^{t+1} = H + (I - HA)X^t(I - AH)$$

Complexity Rate. A is positive definite $\Rightarrow \mathbf{E}[H]$ is nonsingular

$$p_i = \frac{A_{ii}}{\mathbf{Tr}(A)}$$

$$\mathbf{E}[||AX^t - I||_F^2] \leq \left(1 - \frac{\lambda_{\min}(A)}{\mathbf{Tr}(A)}\right)^t ||AX^0 - I||_F^2$$

Randomized Block BFGS

$$X^{t+1} = \arg \min_{X \in \mathbb{R}^{n \times n}} ||X - X^t||^2_{F(A)}$$

$$\text{subject to} \quad S^T A X = S^T, \quad X = X^T$$

Special Choice of Parameters

positve definite

$$B = A$$

$$\mathbf{P}(S = S_i) = p_i$$

$$S = S_i$$

$$X^{t+1} = H + (I - HA)X^t(I - AH)$$

Complexity Rate If A is positive definite $\rightarrow$ $\mathbf{E}[H]$ is nonsingular

Idea: To minimize condition number, choose S so that $\bar{S}$ is an approximate inverse of $A^{1/2}$

$$p_i = \frac{\mathbf{Tr}(S_i^T A S_i)}{\mathbf{Tr}(\bar{S}^T A \bar{S})}$$

$$\mathbf{E}[||AX^t - I||_F^2] \leq \left(1 - \frac{1}{\boxed{\kappa^2(A^{1/2}\bar{S})}}\right) ||AX^0 - I||_F^2$$

BFGS with Randomized Self-Conditioning (RASC)

$$\mathbf{E}[||AX^t - I||_F] \leq \left(1 - \frac{1}{\boxed{\kappa^2(A^{1/2}\bar{S})}}\right)^t ||AX^0 - I||_F$$

$$X_k \to A^{-1} \qquad\qquad X_k^{1/2} \to A^{-1/2}$$

$$\text{Maintain and update } L_k = X_k^{1/2}*$$

Self conditioning sampling:

$\texttt{RASC\_cols:}$ $\quad S = L_k I_{:C}, \quad C \subset \{1, \ldots, n\}$ random set

$\texttt{RASC\_guass:}$ $\quad S \sim \mathcal{N}(0, X_k)$

*Gratton, S., Sartenaer, A., & Ilunga, J. T. (2011). **On a Class of Limited Memory Preconditioners for Large-Scale Nonlinear Least-Squares Problems**. SIAM Journal on Optimization, 21(3), 912–935.

Experiments

Current state of the art

Symmetric Newton-Schulz

$$X^{t+1} = 2X^t - X^t A X^t$$

Self-conditioning Minimal Residual (MR)

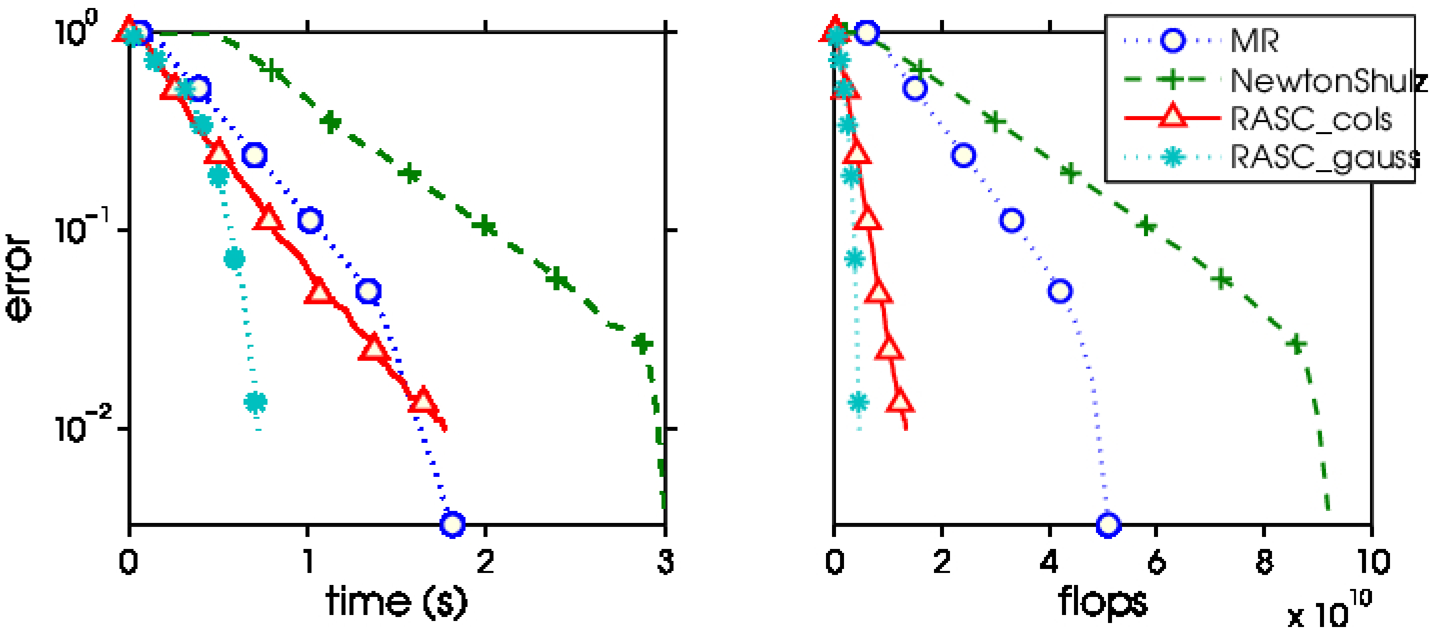
$$\min_{\alpha} ||AX - I||_F^2 \quad \text{subject to} \quad X = X^t + \alpha X^t R^t$$

$$\Rightarrow X^{t+1} = X^t - \frac{\mathbf{Tr}((R^t)^T A X^t R^t)}{\mathbf{Tr}(AX^t R^t)^T A X^t R^t} X^t R^t$$

where $R^t := I - AX^t$

Synthetic Problems

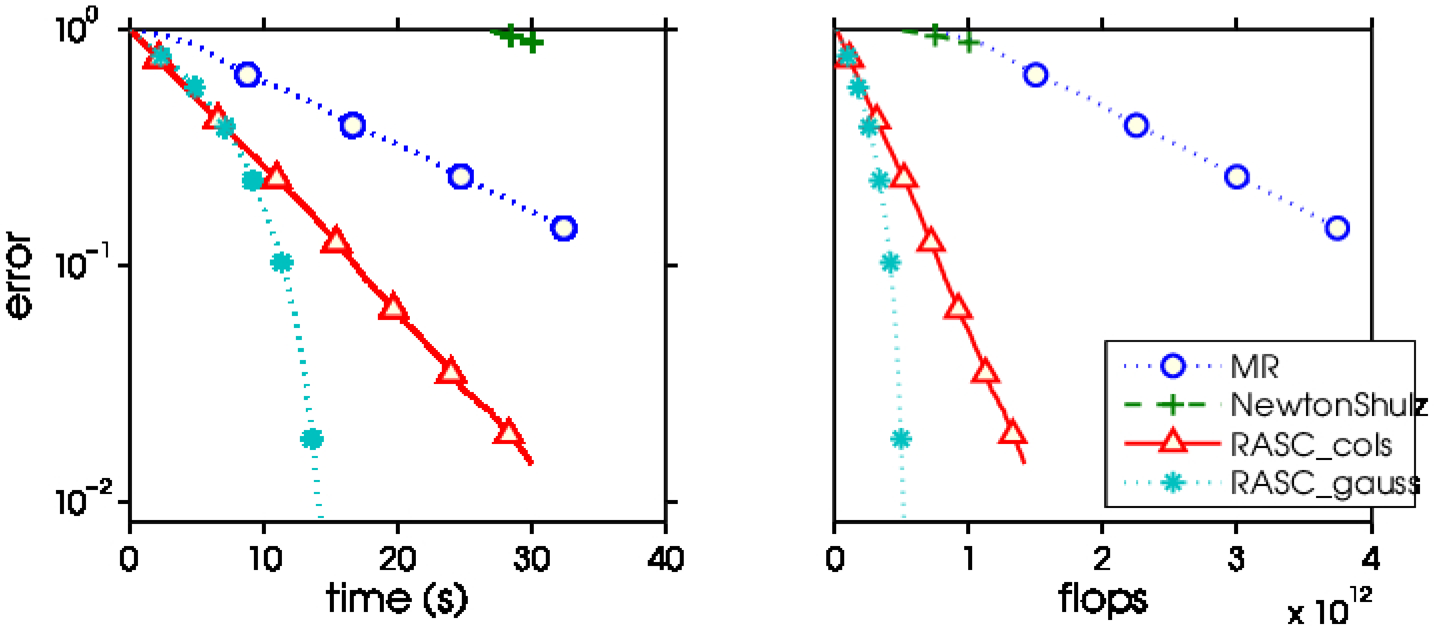
Synthetic data



($randn$, n = 1000)

Synthetic Problems

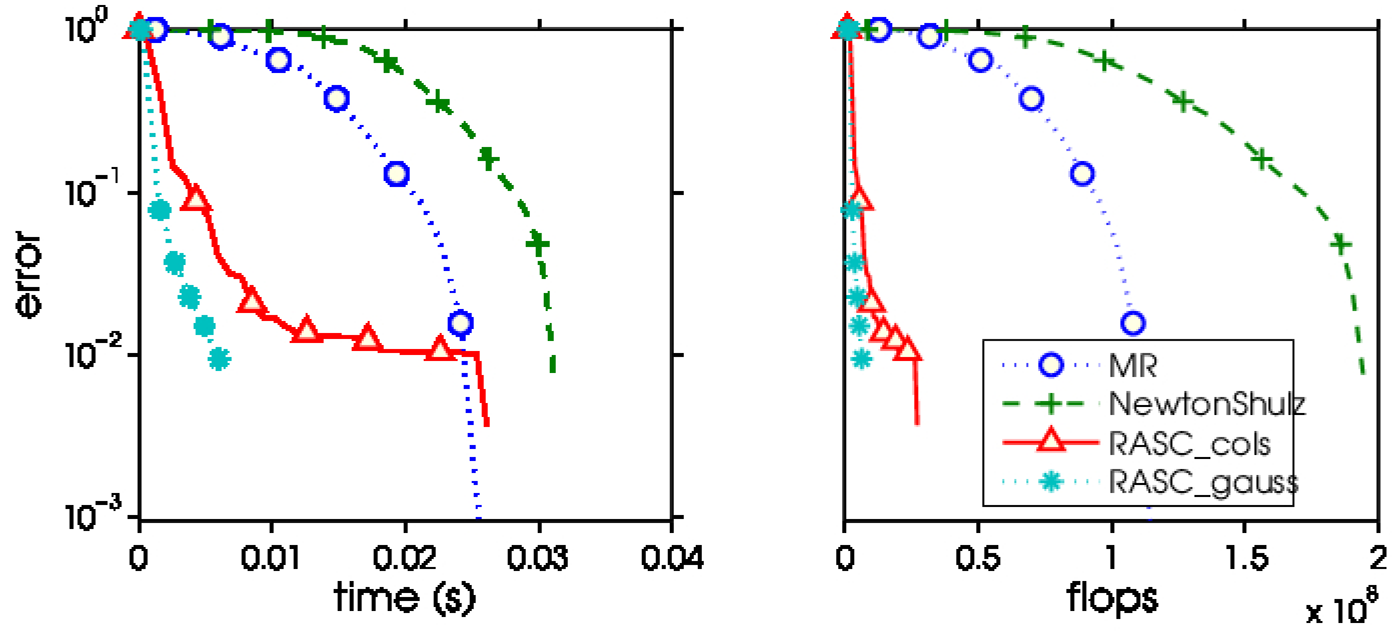
Synthetic data



(*randn*, n = 5000)

Ridge Regression Hessian

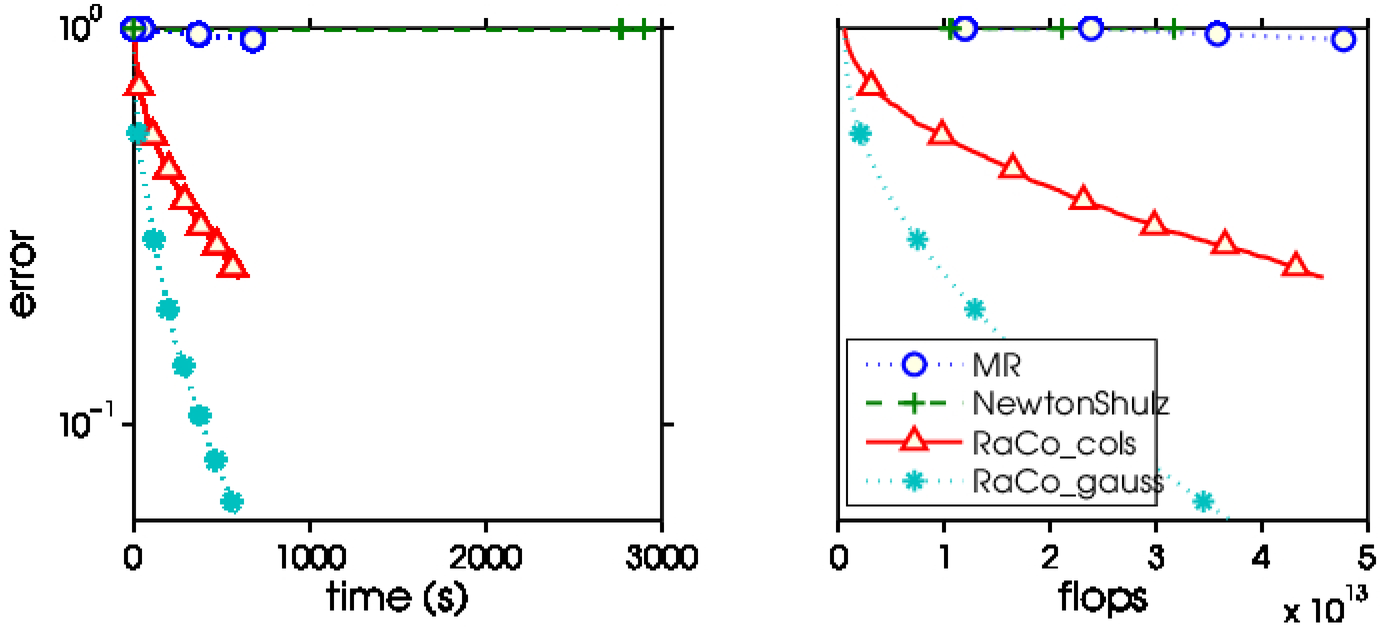
LIBSVM data



(aloi, n = 128)

Ridge Regression Hessian

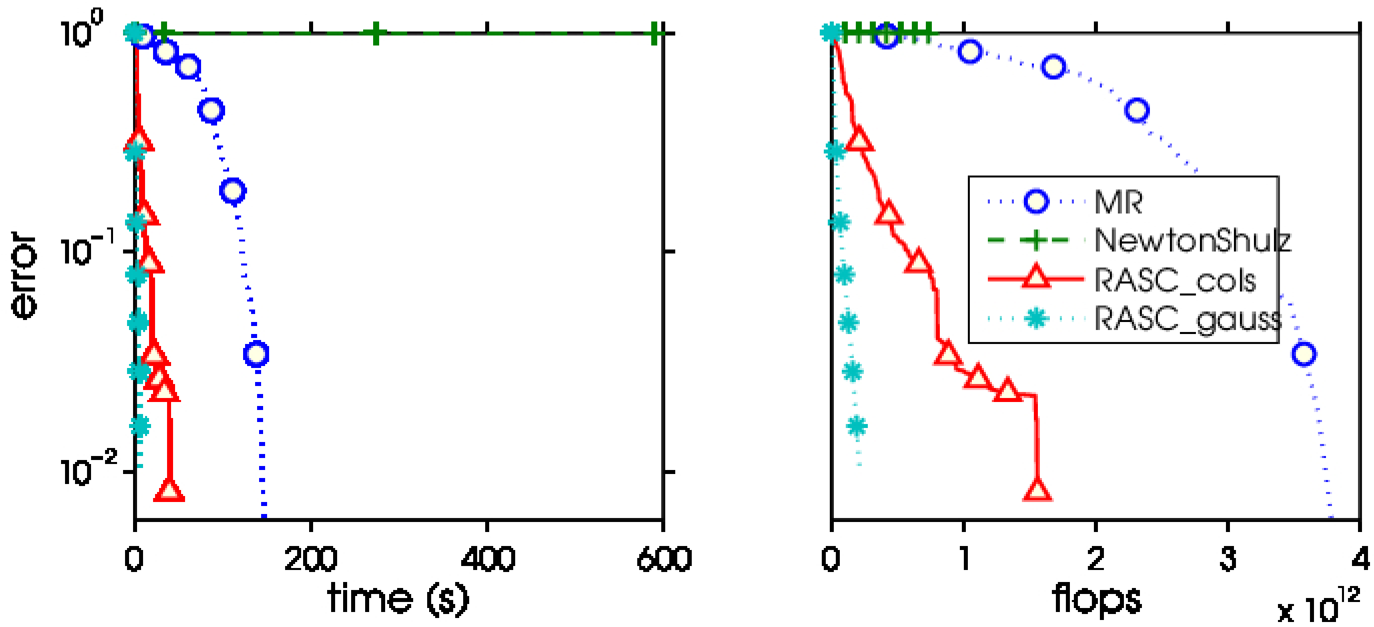
LIBSVM data



(aloi, n = 20,958)

Sparse Matrices from Engineering

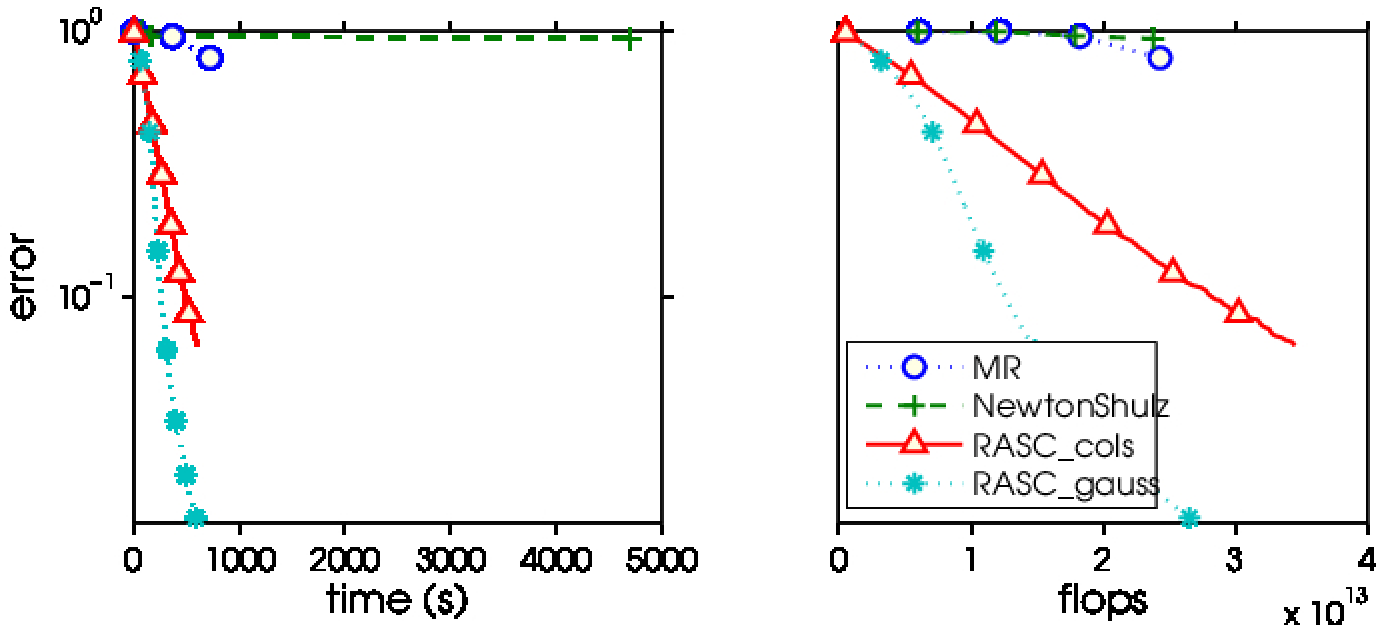
UF collection



(Nasa-nasa, n = 4,705)

Sparse Matrices from Engineering

UF collection



(ND-nd6k, n = 18,000)

Consequences and Future Work

Smooth minimization

$$\min_{x \in \mathbf{R}^n} f(x)$$

$$f \in C^2(\mathbf{R}^n)$$

$$\left(\frac{d}{dy} \nabla f(x + Sy) \right) |_{y=0} = \nabla^2 f(x) S$$

Cheap to calculate, costs
τ X function evaluations

Variable metric methods

$$
\begin{aligned}
&\textbf{Initialize } X^0 \in \mathbf{R}^{n\times n}\\
&\textbf{For } t = 0, 1, \ldots,\\
&1. \quad x^{t+1} = x^t + \alpha_t X^t \nabla f(x^t)\\
&2. \quad S = \mathrm{compute\_sample\_matrix}(X^t)\\
&3. \quad Y = \nabla^2 f(x^t) S\\
&4. \quad X^{t+1} = \arg\min_X ||X - X^t||^2_{\nabla^2 f(x^t)}\\
&\qquad\qquad \text{s.t.} \quad Y^T X = S^T,\ X = X^T
\end{aligned}
$$

Update metric with
RASC update

Preconditioning Sketched Newton

Sketch and project Newton system
$\nabla^2 f(x^t)x^t = -\nabla f(x^t)$

Initialize $X^0 \in \mathbf{R}^{n\times n}$

For $t = 0, 1, \ldots,$

1. $S = \text{compute\_sample\_matrix}(X^t)$
2. $Y = \nabla^2 f(x^t) S$
3. $x^{t+1} = \arg\min_x \|x - x^t\|^2_{\nabla^2 f(x^t)}$ s.t. $Y^T x = -S^T \nabla f(x^t)$
4. $X^{t+1} = \arg\min_X \|X - X^t\|^2_{\nabla^2 f(x^t)}$ s.t. $Y^T X = S^T,\ X = X^T$

Update metric with RASC update

Conclusion for Inverting Matrices

- **New randomized methods** capable of inverting large-scale matrices
- **Convergence rates** which can form the basis of convergence of preconditioning or variable metric methods.
- **Dual viewpoints** of classic quasi-Newton methods
- **Can be extended** to calculating pseudo-inverse

Thank you,
Questions?