Linearly Convergent Randomized Iterative Methods for Computing the Pseudoinverse

Robert Mansel Gower

Joint work with Peter Richtarik

ÉCOLE NORMALE SUPÉRIEURE

INVENTEURS DU MONDE NUMÉRIQUE

The 27th Biennial Numerical Analysis Conference, Strathclyde, June 2017

Sketch and project applications

Numerical Linear Algebra

- Linear systems
- Matrix inverse
- Pseudo inverse

Stochastic Optimization

- Stochastic Quasi-Newton methods
- Stochastic variance reduced gradients
- Stochastic Coordinate descent

Distributed Consensus

Three viewpoints of the Pseudoinverse
Three methods

Three Viewpoints

Given $A \in \mathbb{R}^{m \times n}$ compute an approx. $A^\dagger \in \mathbb{R}^{n \times m}$

$$A^\dagger \quad = \quad \arg \min_{X \in ^{n \times m}} ||X||_F^2$$

subject to

$(1)\; A^\top = A^\top A X$ or $(2)\; A^\top = X A A^\top$ or $(3)\; A X A = A$

Three Viewpoints

Given $A \in \mathbb{R}^{m\times n}$ compute an approx. $A^\dagger \in \mathbb{R}^{n\times m}$

$$A^\dagger \quad = \quad \arg \min_{X \in^{n\times m}} ||X||_F^2$$

subject to

$(1)\; A^\top = A^\top A X$ or $(2)\; A^\top = X A A^\top$ or $(3)\; A X A = A$

Design three methods based on approximate stochastic projections

Three Viewpoints

Given $A \in \mathbb{R}^{m\times n}$ compute an approx. $A^\dagger \in \mathbb{R}^{n\times m}$

$$A^\dagger \quad = \quad \arg\min_{X\in^{n\times m}} ||X||_F^2$$

subject to

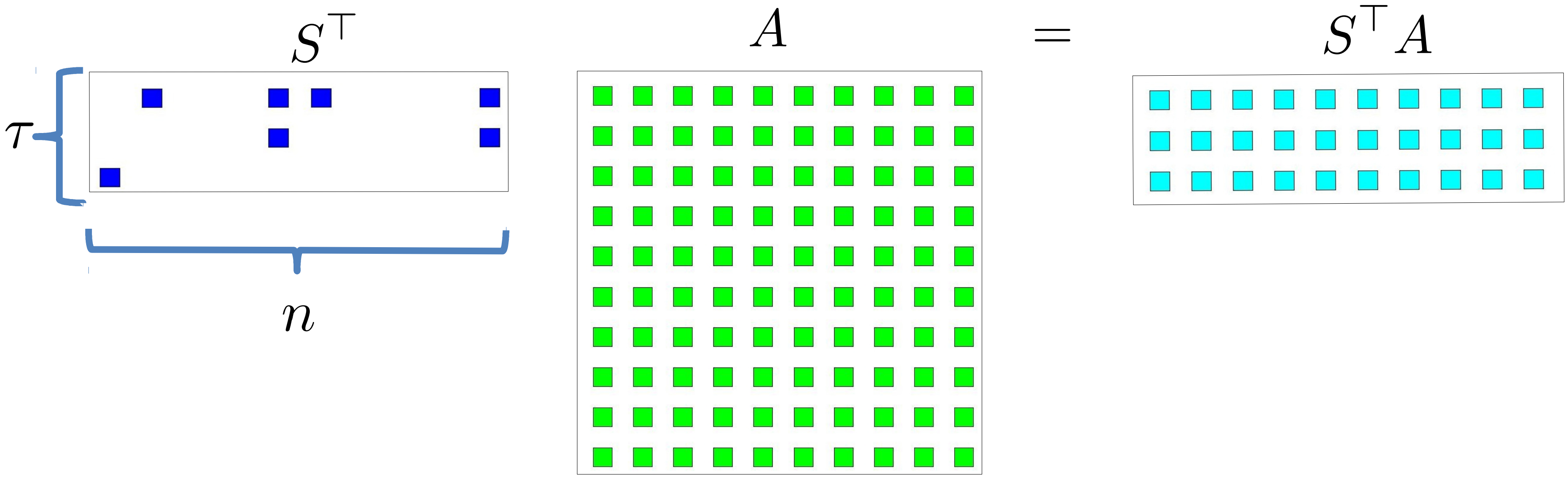
$(1)\ A^\top = A^\top A X$ or $(2)\ A^\top = XAA^\top$ or $(3)\ AXA = A$

Design three methods based on approximate stochastic projections

Use stochastic sketching to approximate the constraints

Sketching

Randomized Sketching



The Sketching Matrix

$$S \sim \mathcal{D} \text{ a distribution over matrices } S \in \mathbb{R}^{m \times \tau} \text{ and } \tau \ll m, n$$

W. B. Johnson and J. Lindenstrauss (1984). Contemporary Mathematics, 26, **Extensions of Lipschitz mappings into a Hilbert space.**

David P. Woodruff (2014), Foundations and Trends® in Theoretical Computer, **Sketching as a Tool for Numerical Linear Algebra.**

Sketching and Projecting

Method 1

Problem:

$$A^\dagger \quad = \quad \arg\min ||X||_F^2 \quad \text{subject to} \quad A^\top = A^\top A X$$

Sample $S \sim \mathcal{D}$

$$X_{t+1} \quad = \quad \arg\min_X ||X - X_t||_F^2 \quad \text{subject to} \quad S^\top A^\top = S^\top A^\top A X$$

Method 1

Problem:

$$A^\dagger = \arg\min ||X||_F^2 \quad \text{subject to} \quad A^\top = A^\top A X$$

Sample $S \sim \mathcal{D}$

$$X_{t+1} = \arg\min_X ||X - X_t||_F^2 \quad \text{subject to} \quad S^\top A^\top = S^\top A^\top A X$$

Or equivalently using duality

Method 1

Problem:

$$A^{\dagger} = \arg\min ||X||_F^2 \quad \text{subject to} \quad A^{\top} = A^{\top} A X$$

Sample $S \sim \mathcal{D}$

$$X_{t+1} = \arg\min_{X} ||X - X_t||_F^2 \quad \text{subject to} \quad S^{\top} A^{\top} = S^{\top} A^{\top} A X$$

Or equivalently using duality

$$X_{t+1} = \arg\min_{X,\Gamma} ||X - A^{\dagger}||_F^2 \quad \text{subject to} \quad X = X_t + A^{\top} A S \Gamma$$

Method 1

Problem:

$$A^\dagger = \arg\min ||X||_F^2 \quad \text{subject to} \quad A^\top = A^\top A X$$

Sample $S \sim \mathcal{D}$

$$X_{t+1} = \arg\min_X ||X - X_t||_F^2 \quad \text{subject to} \quad S^\top A^\top = S^\top A^\top A X$$

Or equivalently using duality

$$X_{t+1} = \arg\min_{X,\Gamma} ||X - A^\dagger||_F^2 \quad \text{subject to} \quad X = X_t + A^\top A S \Gamma$$

$$X_{t+1} = X_t - A^\top A S \underbrace{(S^\top (A^\top A)^2 S)^\dagger}_{\tau \times \tau} S^\top A^\top (A X_t - I)$$

Method 1

Problem:

$$A^\dagger = \arg\min ||X||_F^2 \quad \text{subject to} \quad A^\top = A^\top A X$$

$$\text{Sample } S \sim \mathcal{D}$$

$$X_{t+1} = \arg\min_X ||X - X_t||_F^2 \quad \text{subject to} \quad S^\top A^\top = S^\top A^\top A X$$

Or equivalently using duality

$$X_{t+1} = \arg\min_{X,\Gamma} ||X - A^\dagger||_F^2 \quad \text{subject to} \quad X = X_t + A^\top A S \Gamma$$

Use powerful direct solver

$$X_{t+1} = X_t - A^\top A S \underbrace{(S^\top (A^\top A)^2 S)}_{\tau \times \tau}{}^\dagger S^\top A^\top (A X_t - I)$$

Linear Convergence

Theorem [GR'16]

$$\text{Let } H_S := S(S^\top (A^\top A)^2 S)^\dagger S^\top \succeq 0.$$

Linear Convergence

Theorem [GR'16]

$$\text{Let } H_S := S(S^\top (A^\top A)^2 S)^\dagger S^\top \succeq 0.$$

If $X_0 \in \mathbf{Range}(A^\top A)$ and $\mathbf{E}[H_S] \succ 0$ then

Linear Convergence

Theorem [GR‘16]

$$\text{Let } H_S := S(S^\top (A^\top A)^2 S)^\dagger S^\top \succeq 0.$$

If $X_0 \in \mathbf{Range}(A^\top A)$ and $\mathbf{E}[H_S] \succ 0$ then

$$\mathbf{E}[||X_t - A^\dagger||_F^2] \leq \rho^t ||X_0 - A^\dagger||_F^2$$

where

$$\rho := 1 - \lambda_{\min}^+(A^T A \mathbf{E}[H_S] A^\top A)$$

Linear Convergence

Theorem [GR'16]

$$\text{Let } H_S := S(S^\top (A^\top A)^2 S)^\dagger S^\top \succeq 0.$$

If $X_0 \in \mathbf{Range}(A^\top A)$ and $\mathbf{E}[H_S] \succ 0$ then

$$\mathbf{E}[||X_t - A^\dagger||_F^2] \leq \rho^t ||X_0 - A^\dagger||_F^2$$

where

Smallest nonzero eigenvalue

$$\rho := 1 - \lambda_{\min}^{+}(A^T A \mathbf{E}[H_S] A^\top A)$$

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

$$H := S(S^T(A^\top A)^2 S)^\dagger S^T$$

RMG, P. Richtarik (2016). **Stochastic Dual Ascent for Solving Linear Systems**, arXiv:1512.06890

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

$$H := S(S^T(A^\top A)^2 S)^\dagger S^T$$

Special Choice of Parameters

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

RMG, P. Richtarik (2016). **Stochastic Dual Ascent for Solving Linear Systems**, arXiv:1512.06890

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

$$H := S(S^T(A^\top A)^2 S)^\dagger S^T$$

Special Choice of Parameters

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

$$\mathbf{E}[H_S] \quad = \frac{1}{m}\sum_{i=1}^m \frac{e_i e_i^T}{||A^\top A_{:i}||_2^2}$$
$$= \mathrm{diag}(||A^\top A_{:i}||_2^2)$$

RMG, P. Richtarik (2016). **Stochastic Dual Ascent for Solving Linear Systems**, arXiv:1512.06890

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

$$H := S\boxed{(S^T(A^\top A)^2 S)}^\dagger S^T$$

Special Choice of Parameters

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

$$\begin{aligned}\mathbf{E}[H_S] &= \frac{1}{m}\sum_{i=1}^{m} \frac{e_i e_i^T}{\boxed{||A^\top A_{:i}||_2^2}} \\ &= \mathrm{diag}(||A^\top A_{:i}||_2^2)\end{aligned}$$

RMG, P. Richtarik (2016). **Stochastic Dual Ascent for Solving Linear Systems**, arXiv:1512.06890

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

$$H := \boxed{S}\boxed{(S^T(A^\top A)^2 S)}^\dagger \boxed{S^T}$$

Special Choice of Parameters

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

$$\begin{aligned} \mathbf{E}[H_S] \quad &= \frac{1}{m}\sum_{i=1}^{m} \frac{\boxed{e_i e_i^T}}{\boxed{||A^\top A_{:i}||_2^2}} \\ &= \mathrm{diag}(||A^\top A_{:i}||_2^2) \end{aligned}$$

RMG, P. Richtarik (2016). **Stochastic Dual Ascent for Solving Linear Systems**, arXiv:1512.06890

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

$$H := S(S^T(A^\top A)^2 S)^\dagger S^T$$

Special Choice of Parameters

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

$$\mathbf{E}[H_S] = \frac{1}{m}\sum_{i=1}^{m} \frac{e_i e_i^T}{||A^\top A_{:i}||_2^2} = \mathrm{diag}(||A^\top A_{:i}||_2^2)$$

No zero columns in A → $\mathbf{E}[H]$ is positive definite

RMG, P. Richtarik (2016). **Stochastic Dual Ascent for Solving Linear Systems**, arXiv:1512.06890

Interpretable Convergence

Theorem [GR'16] If $S = S_i$ with probability

$$p_i = \frac{\mathbf{Tr}(S_i^\top (A^\top A)^2 S_i)}{\mathbf{Tr}(\bar{S}^T (A^\top A)^2 \bar{S})}, \quad \text{for } i = 1, \ldots, r$$

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

$$\rho := 1 - \lambda_{\min}^{+}(A^T A \mathbf{E}[H_S] A^\top A) \leq 1 - \frac{1}{\kappa^2(A^T A \bar{S})}$$

$$\kappa^2(A^\top A \bar{S}) := \frac{\mathbf{Tr}\left(\bar{S}^\top (A^\top A)^2 \bar{S}\right)}{\lambda_{\min}(\bar{S}^\top (A^\top A)^2 \bar{S})}$$

Interpretable Convergence

Theorem [GR'16] If $S = S_i$ with probability

$$p_i = \frac{\mathbf{Tr}(S_i^\top (A^\top A)^2 S_i)}{\mathbf{Tr}(\bar{S}^T (A^\top A)^2 \bar{S})}, \quad \text{for } i = 1, \ldots, r$$

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

$$\rho := 1 - \lambda_{\min}^{+}(A^T A \mathbf{E}[H_S] A^\top A) \leq 1 - \frac{1}{\kappa^2(A^T A \bar{S})}$$

where κ is the scaled condition number,

$$\kappa^2(A^\top A \bar{S}) := \frac{\mathbf{Tr}\left(\bar{S}^\top (A^\top A)^2 \bar{S}\right)}{\lambda_{\min}(\bar{S}^\top (A^\top A)^2 \bar{S})}$$

Interpretable Convergence

Theorem [GR'16] If $S = S_i$ with probability

$$p_i = \frac{\mathbf{Tr}(S_i^\top (A^\top A)^2 S_i)}{\mathbf{Tr}(\bar{S}^T (A^\top A)^2 \bar{S})}, \quad \text{for } i = 1, \ldots, r$$

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

$$\rho := 1 - \lambda_{\min}^{+}(A^T A \mathbf{E}[H_S] A^\top A) \leq 1 - \frac{1}{\kappa^2(A^T A \bar{S})}$$

where κ is the scaled condition number,

$\bar{S} = (A^\top A)^\dagger \approx X_t X_t^\top ?$

$$\kappa^2(A^\top A \bar{S}) := \frac{\mathbf{Tr}\left(\bar{S}^\top (A^\top A)^2 \bar{S}\right)}{\lambda_{\min}(\bar{S}^\top (A^\top A)^2 \bar{S})}$$

Adaptive Sketching

$$\mathbf{E}[||X_t - A^\dagger||_F^2] \leq \left(1 - \boxed{\frac{\lambda_{\min}(\bar{S}^T(A^\top A)^2\bar{S})}{\mathbf{Tr}(\bar{S}^T(A^\top A)^2\bar{S})}}\right)^t ||X_0 - A^\dagger||_F^2$$

To minimize condition number:

$$\text{If } \bar{S} = A^\dagger A^{\top\dagger} \text{ then } \frac{\lambda_{\min}(\bar{S}^T(A^\top A)^2\bar{S})}{\mathbf{Tr}(\bar{S}^T(A^\top A)^2\bar{S})} = \frac{\lambda_{\min}(I)}{\mathbf{Tr}(I)} = \frac{1}{n}$$

Adaptive Sketching

$$\mathbf{E}[||X_t - A^\dagger||_F^2] \leq \left(1 - \boxed{\frac{\lambda_{\min}(\bar{S}^T(A^\top A)^2\bar{S})}{\mathbf{Tr}(\bar{S}^T(A^\top A)^2\bar{S})}}\right)^t ||X_0 - A^\dagger||_F^2$$

To minimize condition number:

$$\text{If } \bar{S} = A^\dagger A^{\top\dagger} \text{ then } \frac{\lambda_{\min}(\bar{S}^T(A^\top A)^2\bar{S})}{\mathbf{Tr}(\bar{S}^T(A^\top A)^2\bar{S})} = \frac{\lambda_{\min}(I)}{\mathbf{Tr}(I)} = \frac{1}{n}$$

$$X_k \to A^\dagger \quad \Longrightarrow \quad X_t X_t^\top \to A^\dagger A^{\dagger\top}$$

Adaptive Sketching

$$\mathbf{E}[||X_t - A^\dagger||_F^2] \le \left(1 - \boxed{\frac{\lambda_{\min}(\bar{S}^T(A^\top A)^2\bar{S})}{\mathbf{Tr}(\bar{S}^T(A^\top A)^2\bar{S})}}\right)^t ||X_0 - A^\dagger||_F^2$$

To minimize condition number:

$$\text{If } \bar{S} = A^\dagger A^{\top\dagger} \text{ then } \frac{\lambda_{\min}(\bar{S}^T(A^\top A)^2\bar{S})}{\mathbf{Tr}(\bar{S}^T(A^\top A)^2\bar{S})} = \frac{\lambda_{\min}(I)}{\mathbf{Tr}(I)} = \frac{1}{n}$$

$$X_k \to A^\dagger \quad \Longrightarrow \quad X_t X_t^\top \to A^\dagger A^{\dagger\top}$$

$$S = I_C X_t X_t^\top, \quad C \subset \{1, \dots, n\}$$

Adaptive Sketching

$$\mathbf{E}[||X_t - A^\dagger||_F^2] \leq \left(1 - \boxed{\frac{\lambda_{\min}(\bar{S}^T(A^\top A)^2\bar{S})}{\mathbf{Tr}(\bar{S}^T(A^\top A)^2\bar{S})}}\right)^t ||X_0 - A^\dagger||_F^2$$

To minimize condition number:

$$\text{If } \bar{S} = A^\dagger A^{\top\dagger} \text{ then } \frac{\lambda_{\min}(\bar{S}^T(A^\top A)^2\bar{S})}{\mathbf{Tr}(\bar{S}^T(A^\top A)^2\bar{S})} = \frac{\lambda_{\min}(I)}{\mathbf{Tr}(I)} = \frac{1}{n}$$

$$X_k \to A^\dagger \quad \Longrightarrow \quad X_t X_t^\top \to A^\dagger A^{\dagger\top}$$

$$\cancel{S = I_C X_t X_t^\top} \quad C \subset \{1, \ldots, n\}$$

Adaptive Sketching

$$\mathbf{E}[||X_t - A^\dagger||_F^2] \leq \left(1 - \boxed{\frac{\lambda_{\min}(\bar{S}^T(A^\top A)^2 \bar{S})}{\mathbf{Tr}(\bar{S}^T(A^\top A)^2 \bar{S})}}\right)^t ||X_0 - A^\dagger||_F^2$$

To minimize condition number:

$$\text{If } \bar{S} = A^\dagger A^{\top\dagger} \text{ then } \frac{\lambda_{\min}(\bar{S}^T(A^\top A)^2 \bar{S})}{\mathbf{Tr}(\bar{S}^T(A^\top A)^2 \bar{S})} = \frac{\lambda_{\min}(I)}{\mathbf{Tr}(I)} = \frac{1}{n}$$

$$X_k \to A^\dagger \quad \Longrightarrow \quad X_t X_t^\top \to A^\dagger A^{\dagger\top}$$

Didn't work well in practice

$$S = I_C X_t X_t^\top \qquad C \subset \{1, \ldots, n\}$$

Choosing the Sketching

Sample $S \sim \mathcal{D}$

$$X_{t+1} = \arg\min_{X} ||X - X_t||_F^2 \quad \text{subject to} \quad S^\top A^\top = S^\top A^\top A X$$

Adaptive method

`SATAX-ada` $\quad \mathbf{Prob}[S = X_t I_C] = 1/\binom{|C|}{n}, \quad C \subset \{1, \ldots, n\}$

Uniform coordinates

`SATAX-uni` $\quad \mathbf{Prob}[S = I_C] = 1/\binom{|C|}{n}, \quad C \subset \{1, \ldots, n\}$

Numerics

Benchmark

Symmetric Newton-Schulz

$$X_{t+1} = 2X_t - X_t A X_t$$

$$X_0 = \frac{1}{2}\frac{A^\top}{||A||_F^2} \Rightarrow ||I - X_0 A|| < 1$$

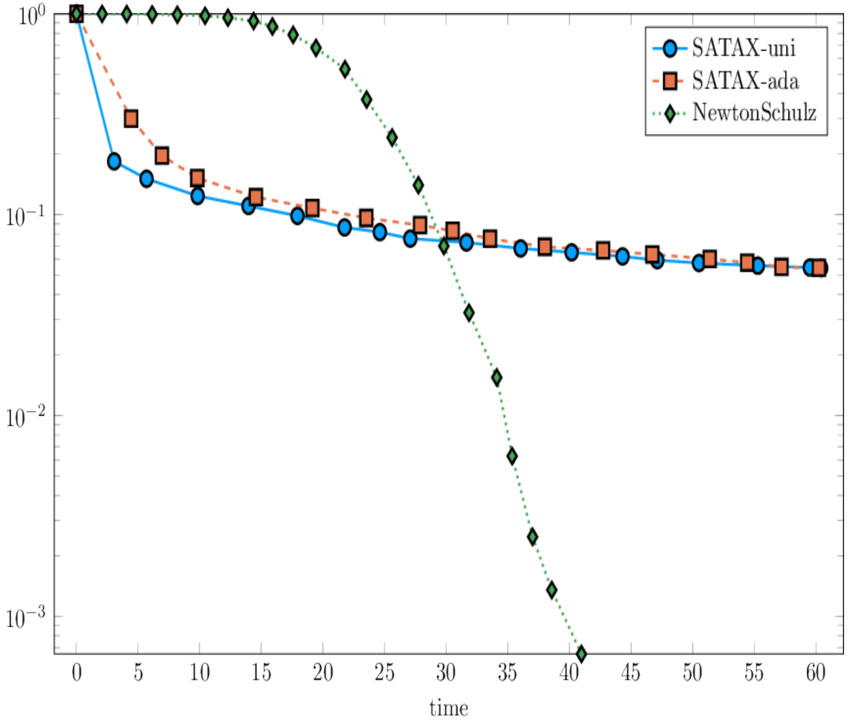
Guarantees convergence

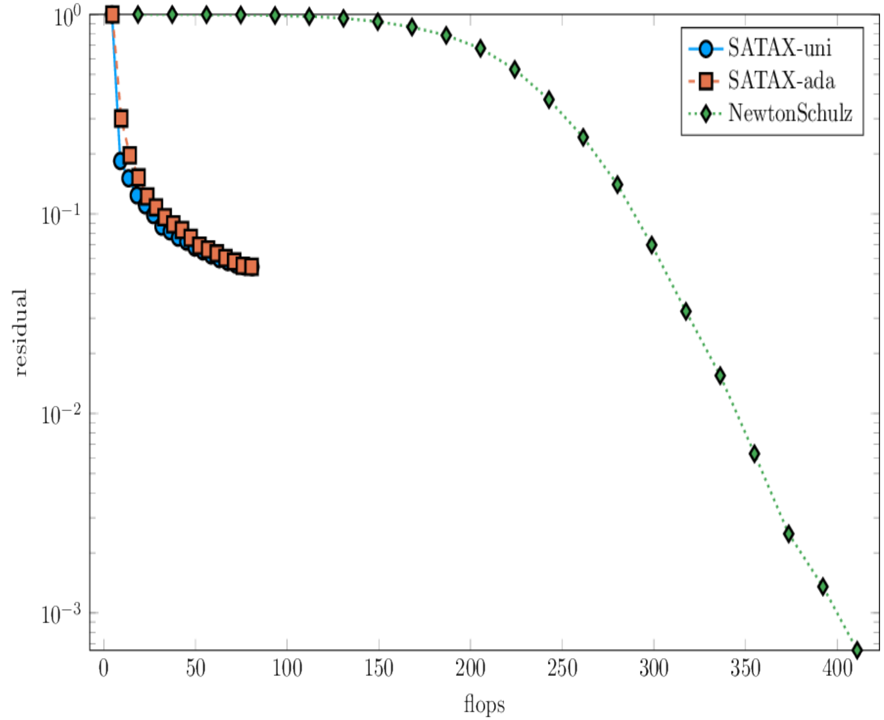
Residual

$$r_t = ||A - AX_tA||_F$$

Sparse Matrices from Engineering

UF collection



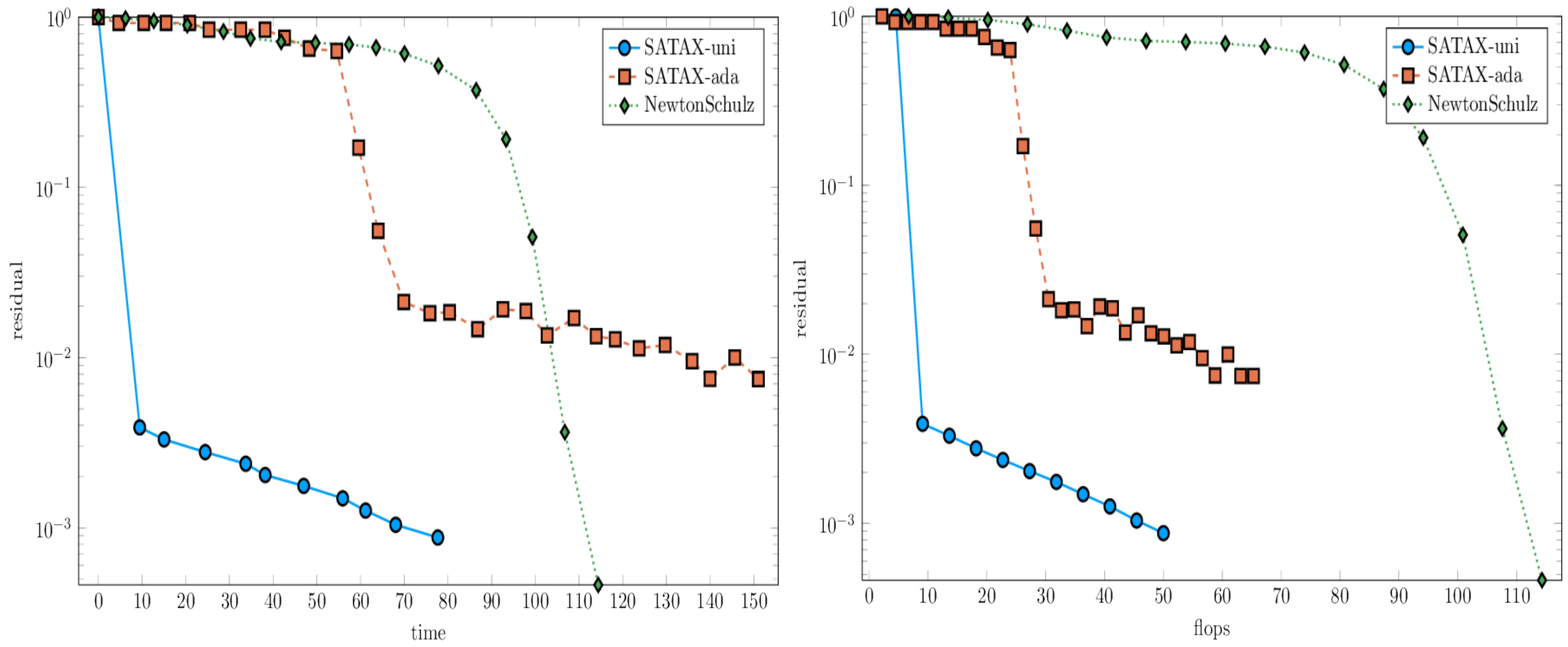


$$\tau = \lfloor \sqrt{m} \rfloor = 48$$

LPnetlib/lp ken 07 (m; n) = (2, 426; 3, 602).

Sparse Matrices from Engineering

UF collection

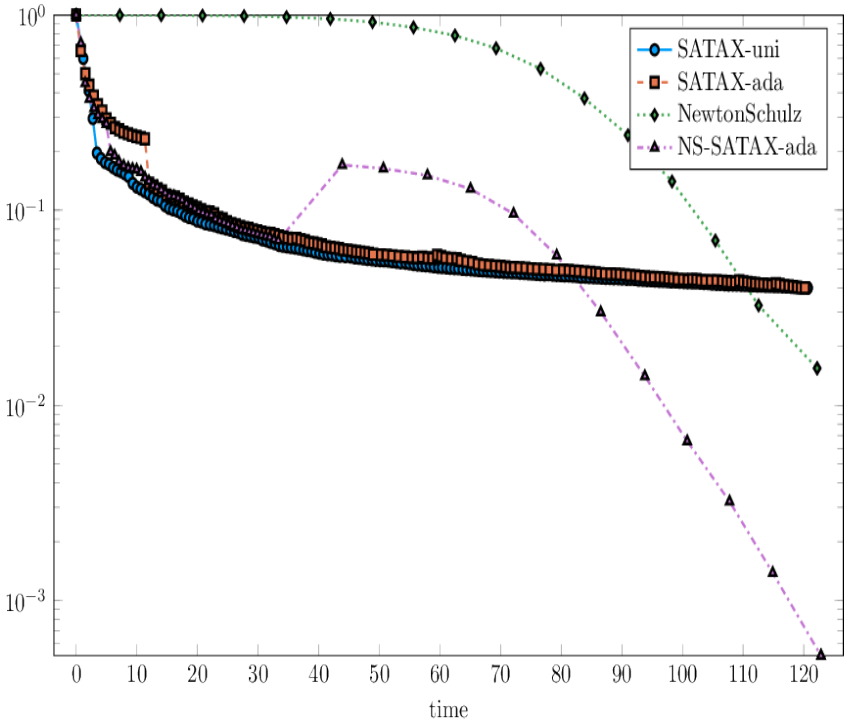


$$\tau = \lfloor \sqrt{m} \rfloor = 39$$

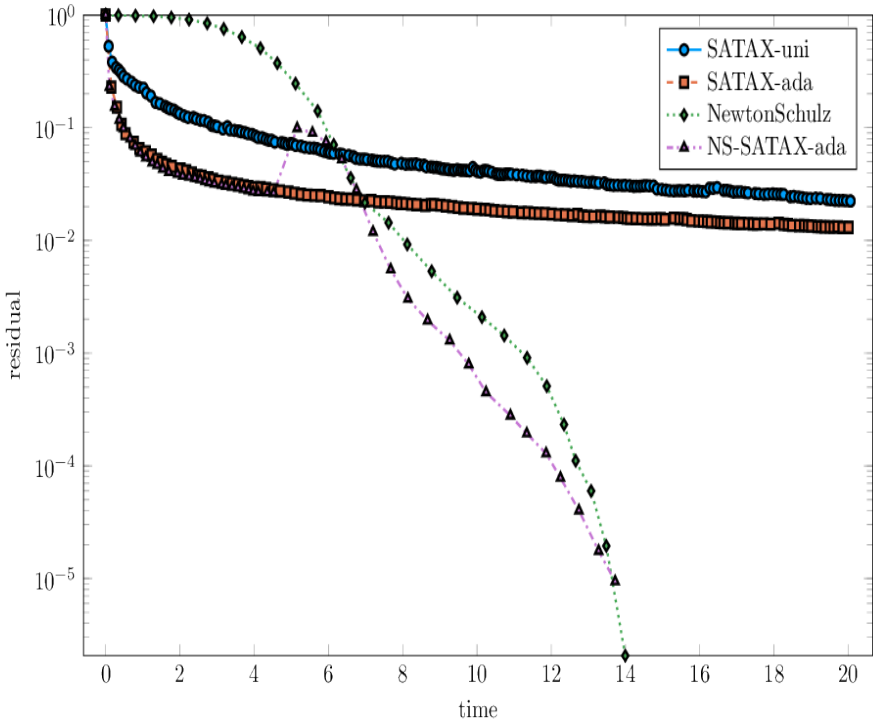
Meszaros/primagaz (m; n) = (1, 554; 10, 836)

Sparse Matrices from Engineering

UF collection



lp ken 07 (m; n) = (2, 426; 3, 602).



Maragal_3 (m; n) = (1,690; 860).

Symmetric Rank deficient Matrices $A = A^T$

The Symmetric Method

Problem:

$$\begin{aligned} A^{\dagger} \quad &= \quad \arg\min ||X||_F^2 \\ &\quad\ \text{subject to} \quad A = AXA \end{aligned}$$

The Symmetric Method

Problem:

$$A^\dagger \quad = \quad \arg\min ||X||_F^2 \quad \text{subject to} \quad A = AXA$$

Sample $S \sim \mathcal{D}$

$$X_{t+1} \quad = \quad \arg\min_{X} ||X - X_t||_F^2 \quad \text{subject to} \quad S^\top A S = S^\top A X A S$$

Or equivalently using duality

$$X_{t+1} \quad = \quad \arg\min_{X,\Gamma} ||X - A^\dagger||_F^2 \quad \text{subject to} \quad X = X_t + AS\Gamma S^\top A$$

The Symmetric Method

Problem:

$$A^\dagger = \arg\min ||X||_F^2 \quad \text{subject to} \quad A = AXA$$

Sample $S \sim \mathcal{D}$

$$X_{t+1} = \arg\min_X ||X - X_t||_F^2 \quad \text{subject to} \quad S^\top A S = S^\top A X A S$$

Or equivalently using duality

$$X_{t+1} = \arg\min_{X,\Gamma} ||X - A^\dagger||_F^2 \quad \text{subject to} \quad X = X_t + AS\Gamma S^\top A$$

$$X_{t+1} = X_t + AS(\underbrace{S^\top A^2 S}_{\tau \times \tau})^\dagger S^\top (A - AX_tA)S(S^\top A^2 S)^\dagger S^\top A$$

Symmetric iterates

Choosing the Sketching

Sample $S \sim \mathcal{D}$

$$\begin{aligned} X_{t+1} \quad &= \quad \arg\min_{X} ||X - X_t||_F^2 \\ &\quad \text{subject to} \quad S^\top A S = S^\top A X A S \end{aligned}$$

Adaptive method

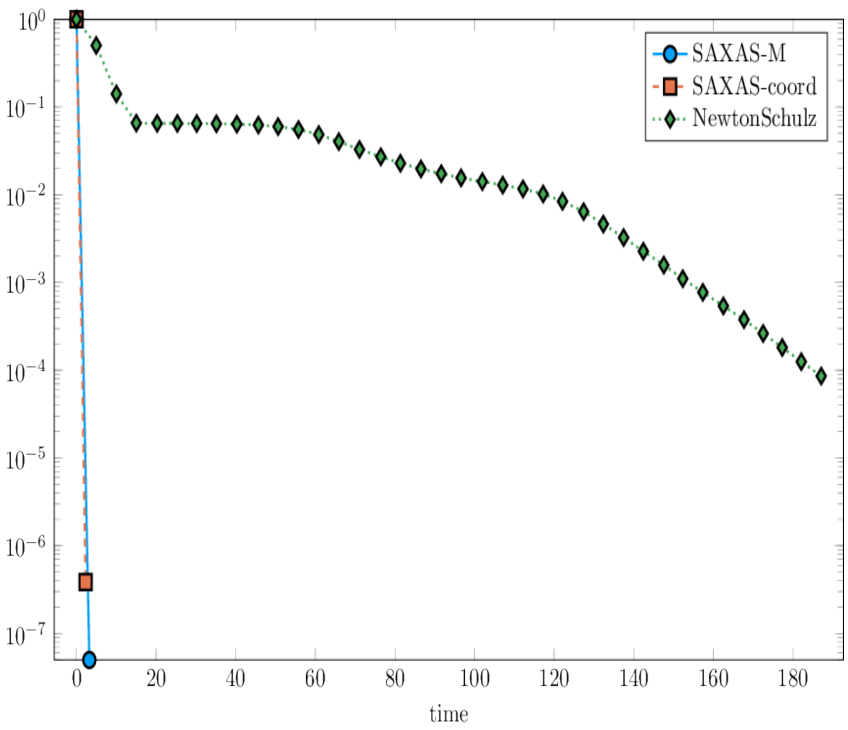
`SAXAS-ada` $\quad S = X_t I_C, \quad C \subset \{1, \ldots, n\}$

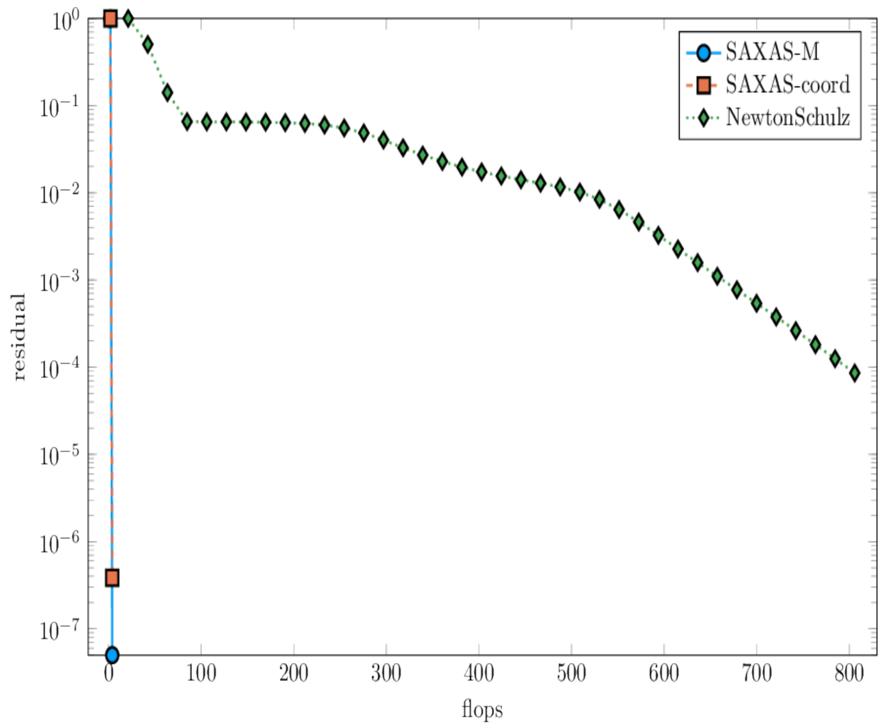
Uniform coordinates

`SAXAS-uni` $\quad S = I_C, \quad C \subset \{1, \ldots, n\}$

Hessian of linear least squares

LIBSVM data

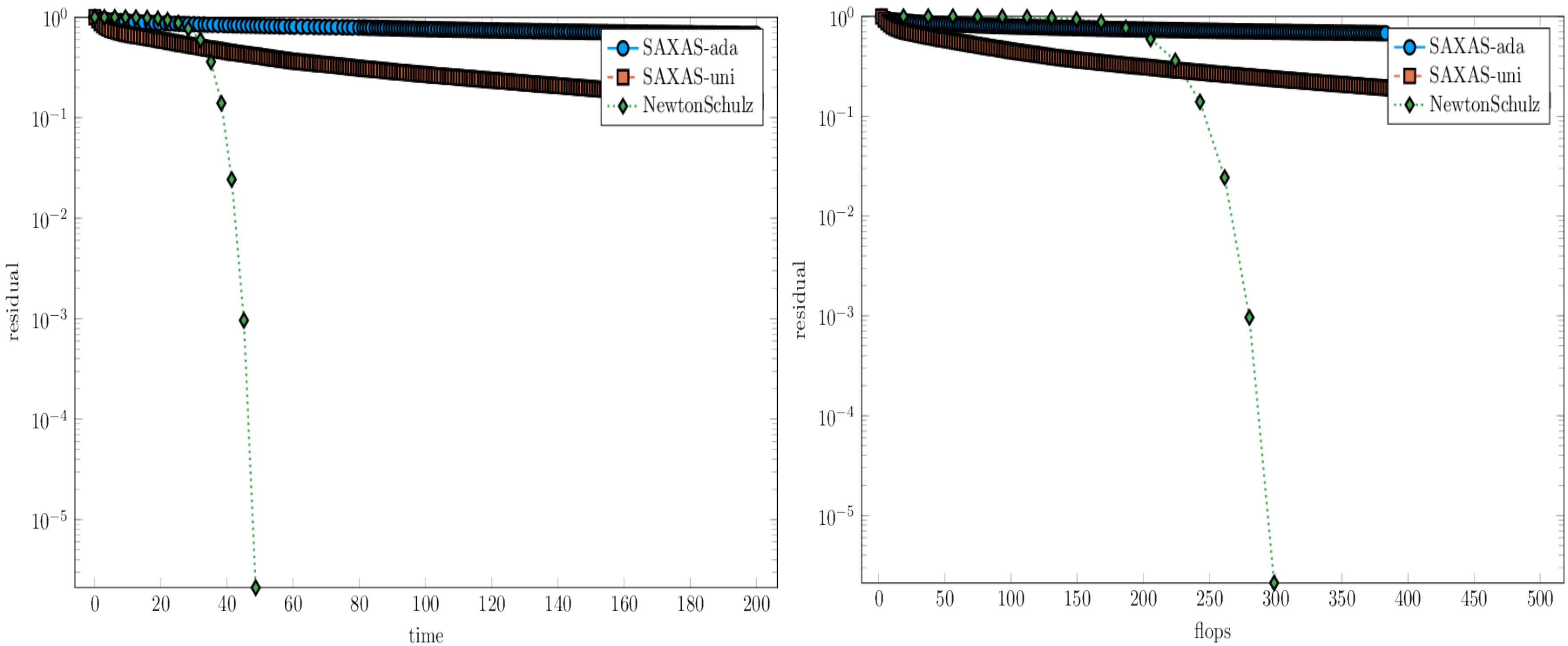




$$\tau = \lfloor \sqrt{m} \rfloor = 70$$

(gisette, n = 5,000)

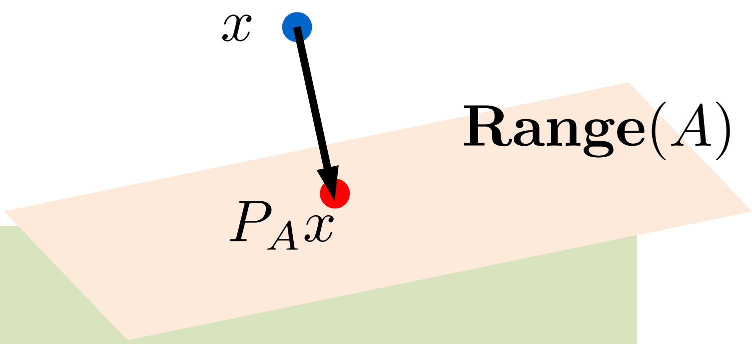
Low rank approx of Gaussian



$$\tau = \lfloor \sqrt{m} \rfloor = 70$$

(best rank 1000 approx to the matrix $G+G^T$
where G is a $5000X5000$ Gaussian matrix)

Related Problems



Range Space Projection

$$\begin{aligned} P_A \quad = \quad & \arg\min_P ||P||_F^2 \\ & \text{subject to} \quad PA = A, \quad P = P^\top, \quad P \succ 0 \end{aligned}$$

Sketch and Project

$$\begin{aligned} P_{t+1} \quad = \quad & \arg\min_P ||P - P_t||_F^2 \\ & \text{subject to} \quad PAS = AS, \quad P = P^\top, \quad P \succ 0 \end{aligned}$$

$$\mathbf{E}[||P_t - P_A||_F^2] \leq \left(1 - \frac{\lambda_{\min}(\bar{S}^T A^2 \bar{S})}{\mathbf{Tr}(\bar{S}^T A^2 \bar{S})}\right)^t ||P_0 - P_A||_F^2$$

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

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

RMG and Peter Richtárik
Randomized quasi-Newton updates are linearly convergent matrix inversion algorithms
Preprint arXiv:1602.01768, 2016

RMG and Peter Richtárik
Linearly Convergent Randomized Iterative Methods for Computing the Pseudoinverse
Preprint arXiv:1612.06255, 2016