Randomized iterative methods for linear systems and inverting matrices

Robert Gower
joint work with Peter Richtárik

Optimization and Big Data 2015, 7th of May, Edinburgh.

Gower, Robert M., Richtárik, Peter, April 2015.
Randomized Iterative Methods for Linear Systems (in progress)

Gower, Robert M., Richtárik, Peter, April 2015.
Randomized Iterative Methods for Inverting Matrices (in progress)

The Problem

The Problem

Solve a consistent linear system $Ax_* = b$, where $A \in \mathbb{R}^{m \times n}$, $m \geq n$.

$$\begin{bmatrix} \text{——} \; A_{1:} \; \text{——} \\ \text{——} \; A_{2:} \; \text{——} \\ \vdots \\ \vdots \\ \text{——} \; A_{m-1:} \; \text{——} \\ \text{——} \; A_{m:} \; \text{——} \end{bmatrix} \begin{bmatrix} x_*^1 \\ \vdots \\ x_*^n \end{bmatrix} = \begin{bmatrix} b^1 \\ b^2 \\ \vdots \\ \vdots \\ b^{m-1} \\ b^m \end{bmatrix}$$

Solve with an iterative method

$$x_{k+1} = \text{update\_formula}(A, x_k)$$

such that $x_k \to x_*$.

The Problem

The Problem

Solve a consistent linear system $Ax_* = b$, where $A \in \mathbb{R}^{m \times n}$, $m \geq n$.

$$\begin{bmatrix} \text{——} \; A_{1:} \; \text{——} \\ \text{——} \; A_{2:} \; \text{——} \\ \vdots \\ \vdots \\ \text{——} \; A_{m-1:} \; \text{——} \\ \text{——} \; A_{m:} \; \text{——} \end{bmatrix} \begin{bmatrix} x_*^1 \\ \vdots \\ x_*^n \end{bmatrix} = \begin{bmatrix} b^1 \\ b^2 \\ \vdots \\ \vdots \\ b^{m-1} \\ b^m \end{bmatrix}$$

Solve with an iterative method

$$x_{k+1} = \text{update\_formula}(A, x_k)$$

such that $x_k \to x_*$.

Notation: Let $\|x\|_B^2 := x^T B x$ for $B \succ 0$.

The Problem

Solve a consistent linear system $Ax_* = b$, where $A \in \mathbb{R}^{m \times n}$, $m \geq n$.

$$\begin{bmatrix} \text{——} \; A_{1:} \; \text{——} \\ \text{——} \; A_{2:} \; \text{——} \\ \vdots \\ \vdots \\ \text{——} \; A_{m-1:} \; \text{——} \\ \text{——} \; A_{m:} \; \text{——} \end{bmatrix} \begin{bmatrix} x_*^1 \\ \vdots \\ x_*^n \end{bmatrix} = \begin{bmatrix} b^1 \\ b^2 \\ \vdots \\ \vdots \\ b^{m-1} \\ b^m \end{bmatrix}$$

Solve with an iterative method

$$x_{k+1} = \text{update\_formula}(A, x_k)$$

such that $x_k \to x_*$.

Notation: Let $\|x\|_B^2 := x^T B x$ for $B \succ 0$.

Table of Contents

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 memory,...etc)
- Easy to implement
- Easy to analyse, good complexity
- Often fits in parallel architecture

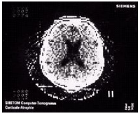
Kaczmarz method

Choose the ith row then iterate

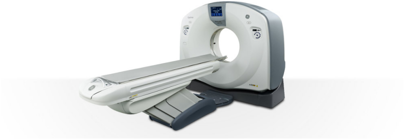
$$\boxed{x_{k+1} = \arg\min_{x \in \mathbb{R}^n} \|x - x_k\|_2^2 \quad \text{subject to} \quad A_{i:}x = b^i.}$$

$$x_{k+1} = x_k - \frac{A_{i:}x_k - b^i}{\|A_{i:}\|_2^2} A_{i:}^T$$

- Developed in 1937 Kaczmarz
- Implemented in the first CT scanner 1972



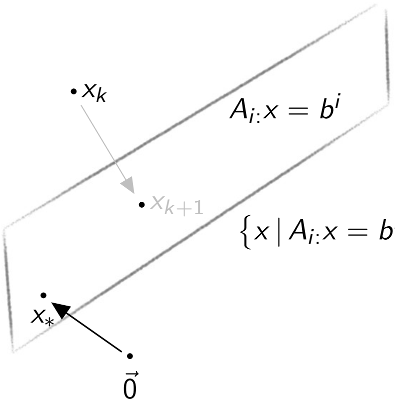
[1]



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

Kaczmarz Interpretation

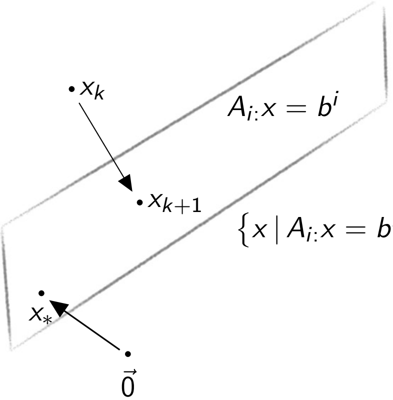
$$\boxed{x_{k+1} = \arg\min_{x\in\mathbb{R}^n} \|x - x_k\|_2^2 \quad \text{subject to} \quad A_{i:}x = b^i.}$$



$$\begin{aligned}\left\{x \,|\, A_{i:}x = b^i\right\} &= \{x \,|\, A_{i:}(x - x_*) = 0\}\\ &= x_* + \{x \,|\, A_{i:}x = 0\}\\ &= x_* + \mathbf{Null}\,(A_{i:})\end{aligned}$$

Kaczmarz Interpretation

$$\boxed{x_{k+1} = \arg\min_{x\in\mathbb{R}^n} \|x - x_k\|_2^2 \quad \text{subject to} \quad A_{i:}x = b^i.}$$



$$\begin{aligned}\left\{x \,|\, A_{i:}x = b^i\right\} &= \{x \,|\, A_{i:}(x - x_*) = 0\} \\ &= x_* + \{x \,|\, A_{i:}x = 0\} \\ &= x_* + \mathbf{Null}\,(A_{i:})\end{aligned}$$

How to choose i

$$\boxed{x_{k+1} = \arg\min_{x\in\mathbb{R}^n} \|x - x_k\|_2^2 \quad \text{subject to} \quad A_{i:}x = b^i.}$$

- Traditional Kaczmarz: Cycle $i = 1, 2, \ldots, m$. Slow in practice + difficult to interpret complexity
- Pick i with probability $p_i = 1/m$. Better in practice + difficult to interpret complexity
- Break-Through (Strohmer & Vershynin, 2009): pick i with probability $p_i = \|A_{i:}\|_2^2/\|A\|_F^2$.

$$\mathbf{E}\left[\|x_k - x_*\|_2^2\right] \leq \left(1 - \frac{\lambda_{\min}(A^TA)}{\|A\|_F^2}\right)^k \|x_0 - x_*\|_2^2.$$

$$\lambda_{\min}(A^TA)/\|A\|_F^2 = 1/\|A\|_F^2\|A^\dagger\|_2^2$$

How to choose i

$$\boxed{x_{k+1} = \arg\min_{x\in\mathbb{R}^n} \|x - x_k\|_2^2 \quad \text{subject to} \quad A_{i:}x = b^i.}$$

- Traditional Kaczmarz: Cycle $i = 1, 2, \ldots, m$. Slow in practice + difficult to interpret complexity
- Pick i with probability $p_i = 1/m$. Better in practice + difficult to interpret complexity
- Break-Through (Strohmer & Vershynin, 2009): pick i with probability $p_i = \|A_{i:}\|_2^2/\|A\|_F^2$.

$$\mathbf{E}\left[\|x_k - x_*\|_2^2\right] \leq \left(1 - \frac{\lambda_{\min}(A^TA)}{\|A\|_F^2}\right)^k \|x_0 - x_*\|_2^2.$$

$$\lambda_{\min}(A^TA)/\|A\|_F^2 = 1/\|A\|_F^2\|A^\dagger\|_2^2$$

Coordinate Descent (Gauss-Seidel)

Choose the ith coordinate then

$$\boxed{x_{k+1} = \arg\min_{x\in\mathbb{R}^n} \|Ax-b\|_2^2 \quad \text{subject to} \quad x = x_k + te_i, \quad t\in\mathbb{R}.}$$

$$x_{k+1} = x_k - \frac{(A_{:i})^T(Ax_k-b)}{\|A_{:i}\|_2^2} e_i$$

Note that $\|Ax-b\|_2^2 = \|A(x-x_*)\|_2^2 = \|x-x_*\|_{A^TA}^2$

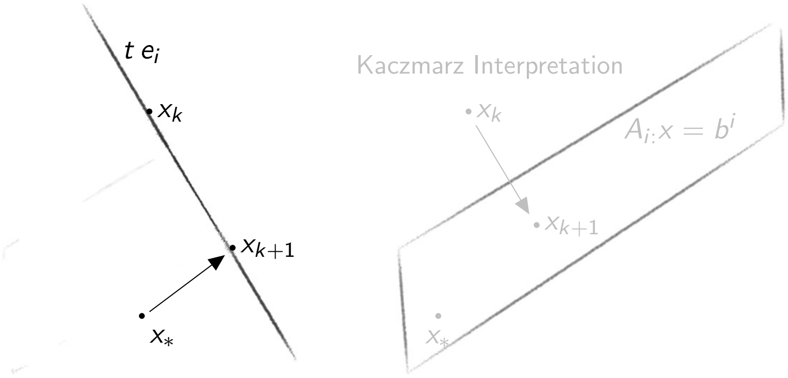
Convergence (Leventhal & Lewis, 2010)

Pick i with probability $p_i = \|A_{:i}\|_2^2/\|A\|_F^2$.

$$\mathbf{E}\left[\|x_k - x_*\|_{A^TA}^2\right] \leq \left(1 - \frac{\lambda_{\min}(A^TA)}{\|A\|_F^2}\right)^k \|x_0 - x_*\|_{A^TA}^2.$$

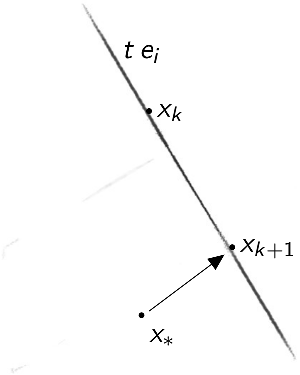
Coordinate Descent Interpretation

$$x_{k+1} = \arg\min_{x \in \mathbb{R}^n} \|x - x_*\|^2_{A^T A} \quad \text{subject to} \quad x = x_k + te_i, \quad t \in \mathbb{R}.$$



Coordinate Descent Interpretation

$$\boxed{x_{k+1} = \arg\min_{x\in\mathbb{R}^n} \|x - x_*\|^2_{A^TA} \quad \text{subject to} \quad x = x_k + te_i, \quad t \in \mathbb{R}.}$$



Kaczmarz Interpretation

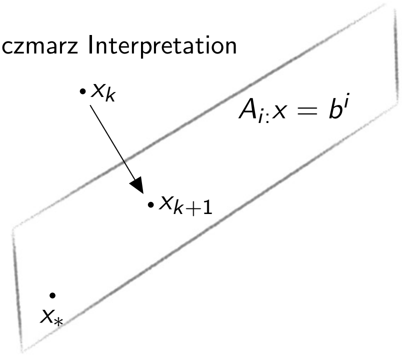


Table of Contents

Framework for designing randomized methods

Choose $B \succ 0 \in \mathbb{R}^{n\times n}$ and a random matrix S independently drawn at each iteration k. **Two** viewpoints of the **same** method.

$$\textbf{(I)} \quad x_{k+1} = \arg\min_{x\in\mathbb{R}^n} \|x - x_k\|_B^2 \quad \text{s. t.} \quad S^T A x = S^T b,$$

$$\textbf{(II)} \quad x_{k+1} = \arg\min_{x\in\mathbb{R}^n} \|x - x_*\|_B^2 \quad \text{s. t.} \quad x \in x_k + B^{-1}\textbf{Range}\left(A^T S\right)$$

(I) : Project x_k onto a randomly compacted system.

$$\boxed{S^T}\boxed{A}\boxed{x} = \boxed{S^T A}\boxed{x}$$

Kaczmarz fits nicely with $B = I$ and $S = e_i$.
Block Kaczmarz choose $B = I$ and $S = I_{:C}$ a subset of columns of identity.

Framework for designing randomized methods

Choose $B \succ 0 \in \mathbb{R}^{n\times n}$ and a random matrix S independently drawn at each iteration k. **Two** viewpoints of the **same** method.

$$\textbf{(I)} \quad x_{k+1} = \arg\min_{x\in\mathbb{R}^n} \|x - x_k\|_B^2 \quad \text{s. t.} \quad S^T A x = S^T b,$$

$$\textbf{(II)} \quad x_{k+1} = \arg\min_{x\in\mathbb{R}^n} \|x - x_*\|_B^2 \quad \text{s. t.} \quad x \in x_k + B^{-1}\textbf{Range}\left(A^T S\right)$$

(**I**) : Project x_k onto a randomly compacted system.

$$\boxed{S^T} \; \boxed{A} \; \boxed{x} = \boxed{S^T A} \; \boxed{x}$$

Kaczmarz fits nicely with $B = I$ and $S = e_i$.
Block Kaczmarz choose $B = I$ and $S = I_{:C}$ a subset of columns of identity.

Coordinate descent methods fit (II)

$$\boxed{\textbf{(II)} \quad x_{k+1} = \arg\min_{x \in \mathbb{R}^n} \|x - x_*\|_B^2 \quad \text{subject to} \quad x \in x_k + \textbf{Range}\left(B^{-1}A^T S\right)}$$

- Least-Squares Coord. Desc: With $B = A^T A$ and $S = Ae_i = A_{:i}$

$$x_{k+1} = \arg\min_{x \in \mathbb{R}^n} \|Ax - b\|_2^2 \quad \text{subject to} \quad x = x_k + t\, e_i, \quad t \in \mathbb{R}.$$

  Stochastic Newton (SDNA[2] Method 1) Let $S = AI_{:C} = A_{:C}$ subset of columns of A

$$x_{k+1} = \arg\min_{x \in \mathbb{R}^n} \|Ax - b\|_2^2 \quad \text{subject to} \quad x = x_k + t\, I_{:C}, \quad t \in \mathbb{R}^{|C|}.$$

- Positive Definite Coord. Desc: When $A \succ 0$, $B = A$ and $S = I_{:C}$ then

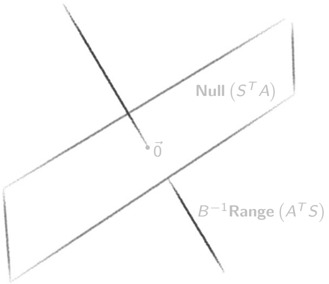
$$x_{k+1} = \arg\min_{x \in \mathbb{R}^n} \underbrace{\frac{1}{2} x^T A x - x^T b}_{= \|x - x_*\|_A^2} \quad \text{subject to} \quad x = x_k + t\, I_{:C}, \quad t \in \mathbb{R}^{|C|},$$

[2] Qu, Z., Richtárik, P., Takáč, M., & Fercoq, O. (2015). SDNA: Stochastic Dual Newton Ascent for Empirical Risk Minimization.

Let $\langle x, y\rangle = x^T B y$ in $\mathbb{R}^n$. $\left\{x \mid S^T A x = S^T b\right\} = x_* + \left\{x \mid S^T A x = 0\right\}$ and

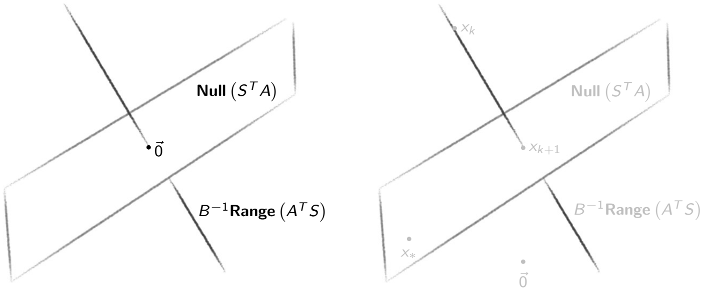
Let $\langle x, y \rangle = x^T By$ in $\mathbb{R}^n$. $\left\{x \mid S^T Ax = S^T b\right\} = x_* + \left\{x \mid S^T Ax = 0\right\}$ and

$$\textbf{Null}\left(S^T A\right) \oplus B^{-1}\textbf{Range}\left(A^T S\right) = \mathbb{R}^n.$$



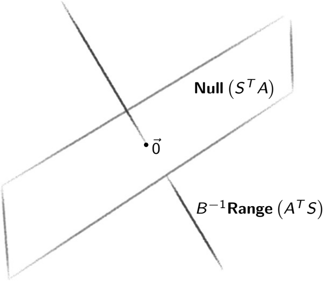
Let $\langle x, y\rangle = x^T B y$ in $\mathbb{R}^n$. $\left\{x \mid S^T A x = S^T b\right\} = x_* + \left\{x \mid S^T A x = 0\right\}$ and

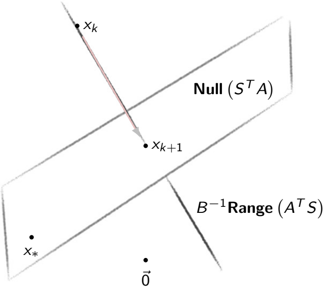
$$\textbf{Null}\left(S^T A\right) \oplus B^{-1}\textbf{Range}\left(A^T S\right) = \mathbb{R}^n.$$



Let $\langle x, y\rangle = x^T B y$ in $\mathbb{R}^n$. $\left\{x \mid S^T A x = S^T b\right\} = x_* + \left\{x \mid S^T A x = 0\right\}$ and

$$\mathbf{Null}\left(S^T A\right) \oplus B^{-1}\mathbf{Range}\left(A^T S\right) = \mathbb{R}^n.$$



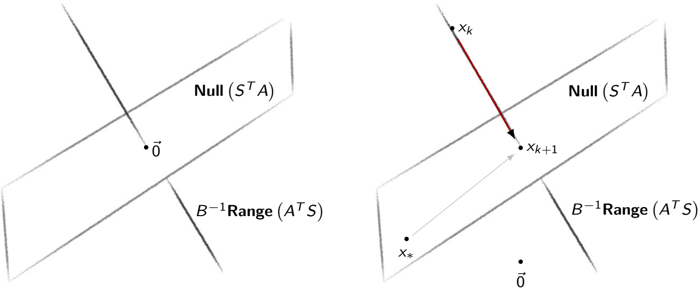


I Project x_k onto $x_* + \mathbf{Null}\left(S^T A\right)$

$$\boxed{x_{k+1} = \arg\min_{x\in\mathbb{R}^n} \|x - x_k\|_B^2 \quad \text{subject to} \quad S^T A x = S^T b}$$

Let $\langle x, y\rangle = x^T By$ in $\mathbb{R}^n$. $\left\{x \mid S^T Ax = S^T b\right\} = x_* + \left\{x \mid S^T Ax = 0\right\}$ and

$$\textbf{Null}\left(S^T A\right) \oplus B^{-1}\textbf{Range}\left(A^T S\right) = \mathbb{R}^n.$$



I Project x_k onto $x_* + \textbf{Null}\left(S^T A\right)$

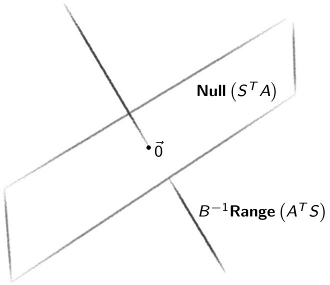
$$\boxed{x_{k+1} = \arg\min_{x\in\mathbb{R}^n} \|x - x_k\|_B^2 \quad \text{subject to} \quad S^T Ax = S^T b}$$

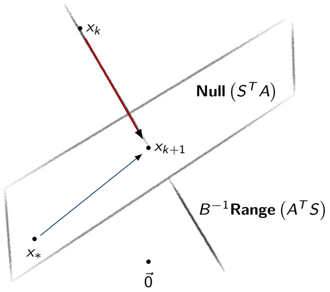
II Project x_* onto $x_k + B^{-1}\textbf{Range}\left(A^T S\right)$

$$\boxed{x_{k+1} = \arg\min_{x\in\mathbb{R}^n} \|x - x_*\|_B^2 \quad \text{subject to} \quad x \in x_k + B^{-1}\textbf{Range}\left(A^T S\right)}$$

Let $\langle x, y\rangle = x^T B y$ in $\mathbb{R}^n$. $\{x \mid S^T A x = S^T b\} = x_* + \{x \mid S^T A x = 0\}$ and

$$\textbf{Null}\left(S^T A\right) \oplus B^{-1}\textbf{Range}\left(A^T S\right) = \mathbb{R}^n.$$





I Project x_k onto $x_* + \textbf{Null}\left(S^T A\right)$

$$\boxed{x_{k+1} = \arg\min_{x\in\mathbb{R}^n} \|x - x_k\|_B^2 \quad \text{subject to} \quad S^T A x = S^T b}$$

II Project x_* onto $x_k + B^{-1}\textbf{Range}\left(A^T S\right)$

$$\boxed{x_{k+1} = \arg\min_{x\in\mathbb{R}^n} \|x - x_*\|_B^2 \quad \text{subject to} \quad x \in x_k + B^{-1}\textbf{Range}\left(A^T S\right)}$$

The Solution

Assuming $A^T S$ has full column rank $\Rightarrow$ closed form solution

II Project x_* onto $x_k + B^{-1}\textbf{Range}\left(A^T S\right)$

$$
\begin{aligned}
x_{k+1} &= x_k + \mathrm{proj}_{B^{-1}\textbf{Range}(A^T S)}\left(x_* - x_k\right) \\
&= x_k + B^{-1}A^T S(S^T A B^{-1} A^T S)^{-1} S^T A(x_k - x_*) \\
&= x_k + B^{-1}A^T S \underbrace{(S^T A B^{-1} A^T S)^{-1}}_{\text{Solve small system.}} S^T (A x_k - b)
\end{aligned}
$$

I Project x_k onto $x_* + \textbf{Null}\left(S^T A\right)$

$$
\begin{aligned}
x_{k+1} &= x_* + \mathrm{proj}_{\textbf{Null}(A^T S)}\left(x_k - x_*\right) \\
&= x_* + (I - B^{-1}\underbrace{A^T S(S^T A B^{-1} A^T S)^{-1} S^T A}_{Z})(x_k - x_*) \\
&= x_* + (I - B^{-1} Z)(x_k - x^*).
\end{aligned}
$$

The Solution

Assuming $A^T S$ has full column rank $\Rightarrow$ closed form solution

II Project x_* onto $x_k + B^{-1}\mathbf{Range}\left(A^T S\right)$

$$\begin{aligned} x_{k+1} &= x_k + \text{proj}_{B^{-1}\mathbf{Range}(A^T S)}\left(x_* - x_k\right) \\ &= x_k + B^{-1}A^T S(S^T A B^{-1} A^T S)^{-1} S^T A(x_k - x_*) \\ &= x_k + B^{-1}A^T S \underbrace{(S^T A B^{-1} A^T S)^{-1}}_{\text{Solve small system.}} S^T (A x_k - b) \end{aligned}$$

I Project x_k onto $x_* + \mathbf{Null}\left(S^T A\right)$

$$\begin{aligned} x_{k+1} &= x_* + \text{proj}_{\mathbf{Null}(A^T S)}\left(x_k - x_*\right) \\ &= x_* + (I - B^{-1}\underbrace{A^T S(S^T A B^{-1} A^T S)^{-1} S^T A}_{Z})(x_k - x_*) \\ &= x_* + (I - B^{-1} Z)(x_k - x^*). \end{aligned}$$

The Fixed point form

All randomness is in the range space projection $B^{-1}Z$

$$Z \stackrel{\text{def}}{=} A^T S (S^T A B^{-1} A^T S)^{-1} S^T A.$$

For analysis, fixed point form

$$x_{k+1} - x_* = (I - B^{-1} Z)(x_k - x_*).$$

The Fixed point form

All randomness is in the range space projection $B^{-1}Z$

$$Z \overset{\text{def}}{=} A^T S(S^T A B^{-1} A^T S)^{-1} S^T A.$$

For analysis, fixed point form

$$\mathbf{E}\left[x_{k+1} - x_* \,|\, x_k\right] = (I - B^{-1}\mathbf{E}\left[Z\right])(x_k - x_*).$$

$$\begin{aligned}
\mathbf{E}\left[\mathbf{E}\left[x_{k+1} - x_* \,|\, x_k\right]\right] &= \mathbf{E}\left[x_{k+1} - x_*\right] \\
&= \mathbf{E}\left[(I - B^{-1}\mathbf{E}\left[Z\right])(x_k - x_*)\right] \\
&= (I - B^{-1}\mathbf{E}\left[Z\right])\mathbf{E}\left[x_k - x_*\right].
\end{aligned}$$

The Fixed point form

All randomness is in the range space projection $B^{-1}Z$

$$Z \stackrel{\text{def}}{=} A^T S (S^T A B^{-1} A^T S)^{-1} S^T A.$$

For analysis, fixed point form

$$\begin{aligned}
\mathbf{E}\left[x_{k+1} - x_* \mid x_k\right] &= (I - B^{-1}\mathbf{E}\left[Z\right])(x_k - x_*).\\
\mathbf{E}\left[\mathbf{E}\left[x_{k+1} - x_* \mid x_k\right]\right] &= \mathbf{E}\left[x_{k+1} - x_*\right]\\
&= \mathbf{E}\left[(I - B^{-1}\mathbf{E}\left[Z\right])(x_k - x_*)\right]\\
&= (I - B^{-1}\mathbf{E}\left[Z\right])\mathbf{E}\left[x_k - x_*\right].
\end{aligned}$$

Convergence Theorems

$$\left\|\mathbf{E}\left[x_k\right] - x_*\right\| \leq \left(1 - \lambda_{\min}(B^{-1/2}\mathbf{E}\left[Z\right] B^{-1/2})\right)^k \|x_0 - x_*\|$$

and when $\mathbf{E}\left[Z\right]$ nonsingular

$$\mathbf{E}\left[\|x_k - x_*\|\right] \leq \left(1 - \lambda_{\min}(B^{-1/2}\mathbf{E}\left[Z\right] B^{-1/2})\right)^k \|x_0 - x_*\|$$

The Fixed point form

All randomness is in the range space projection $B^{-1}Z$

$$Z \stackrel{\text{def}}{=} A^T S(S^T A B^{-1} A^T S)^{-1} S^T A.$$

For analysis, fixed point form

$$\begin{aligned}
\mathbf{E}\left[x_{k+1} - x_* \,|\, x_k\right] &= (I - B^{-1}\mathbf{E}\left[Z\right])(x_k - x_*). \\
\mathbf{E}\left[\mathbf{E}\left[x_{k+1} - x_* \,|\, x_k\right]\right] &= \mathbf{E}\left[x_{k+1} - x_*\right] \\
&= \mathbf{E}\left[(I - B^{-1}\mathbf{E}\left[Z\right])(x_k - x_*)\right] \\
&= (I - B^{-1}\mathbf{E}\left[Z\right])\mathbf{E}\left[x_k - x_*\right].
\end{aligned}$$

Convergence Theorems

$$\|\mathbf{E}\left[x_k\right] - x_*\| \leq \left(1 - \lambda_{\min}(B^{-1/2}\mathbf{E}\left[Z\right] B^{-1/2})\right)^k \|x_0 - x_*\|$$

and when $\mathbf{E}\left[Z\right]$ nonsingular

$$\mathbf{E}\left[\|x_k - x_*\|\right] \leq \left(1 - \lambda_{\min}(B^{-1/2}\mathbf{E}\left[Z\right] B^{-1/2})\right)^k \|x_0 - x_*\|$$

Theorem (General S)

Let S be a random matrix such that $A^T S$ full column rank. Then for all $k \geq 0$,

$$\|\mathbf{E}[x_k] - x_*\| \leq \rho^k \|x_0 - x_*\|,$$

where

$$\rho = 1 - \lambda_{\min}(B^{-1/2}\mathbf{E}[Z]\,B^{-1/2}) \quad \text{and} \quad 0 \leq \rho \leq 1.$$

Proof.

Taking conditional expectation with respect to x_k, we get

$$\mathbf{E}[x_{k+1} - x_* \mid x_k] = (I - B^{-1}\mathbf{E}[Z])(x_k - x_*). \tag{1}$$

Taking full expectation, we get

$$\begin{aligned} \mathbf{E}[x_{k+1} - x_*] &= \mathbf{E}[\mathbf{E}[x_{k+1} - x_* \mid x_k]] \\ &\overset{(1)}{=} \mathbf{E}\left[(I - B^{-1}\mathbf{E}[Z])(x_k - x_*)\right] \\ &= (I - B^{-1}\mathbf{E}[Z])\mathbf{E}[x_k - x_*]. \end{aligned}$$

Now unroll the recurrence and apply the operator norm. As $B^{-1}Z$ is a projection, by Jensen's inequality with the convex functions $\lambda_{\max}$ and $-\lambda_{\min}$, we have

$$0 \leq \lambda_{\max}(B^{-1}\mathbf{E}[Z]) \leq \lambda_{\max}(B^{-1}Z) \leq 1. \qquad \square$$

Unifying previous methods & analysis

Theorem (Discrete random vector)

Let S be discrete r.v. such $S = s_i \in \mathbb{R}^n$ (for concreteness, think of $s_i = e_i$) with probability $p_i > 0$, for $i = 1, \ldots, m$, and let

$$\mathbf{S} = [s_1, \ldots, s_m] \,.$$

Then

$$x_{k+1} = x_k + \frac{s_i^T(Ax_k - b)}{s_i^T AB^{-1}A^T s_i} B^{-1}A^T s_i, \quad \textit{with prob } p_i.$$

If we choose

$$p_i = \frac{s_i^T AB^{-1}A^T s_i}{\|B^{-1/2}A^T\mathbf{S}\|_F^2}, \quad \textit{for } i = 1, \ldots, m,$$

then

$$\mathbf{E}\left[Z\right] = \frac{A^T\mathbf{S}\mathbf{S}^T A}{\|B^{-1/2}A^T\mathbf{S}\|_F^2} \quad \textit{and} \quad \rho = 1 - \frac{\lambda_{\min}\left(B^{-1/2}A^T\mathbf{S}\mathbf{S}^T AB^{-1/2}\right)}{\|B^{-1/2}A^T\mathbf{S}\|_F^2}.$$

Furthermore, if $\mathbf{S}^T A$ has full column rank then $\rho < 1$.

Proof.

$$\mathbf{E}[Z] = \sum_{i=1}^{m} A^T s_i (s_i^T A B^{-1} A^T s_i)^{-1} s_i^T A p_i$$
$$= \frac{1}{\|B^{-1/2} A^T \mathbf{S}\|_F^2} \sum_{i=1}^{m} A^T s_i s_i^T A$$
$$= \frac{1}{\|B^{-1/2} A^T \mathbf{S}\|_F^2} A^T \mathbf{S}\mathbf{S}^T A.$$

Thus the ρ is given by

$$\rho = 1 - \lambda_{\min}\left(B^{-1/2}\mathbf{E}[Z]\, B^{-1/2}\right) = 1 - \frac{\lambda_{\min}\left(B^{-1/2} A^T \mathbf{S}\mathbf{S}^T A B^{-1/2}\right)}{\|B^{-1/2} A^T \mathbf{S}\|_F^2}.$$

As $\mathbf{S}^T A$ has full column rank, $\mathbf{E}[Z]$ is positive definite and $\rho < 1$. $\square$

Unifying previous methods & analysis

$$p_i = \frac{s_i^T A B^{-1} A^T s_i}{\|B^{-1/2} A^T \mathbf{S}\|_F^2} \quad \text{with} \quad \rho = 1 - \frac{\lambda_{\min}\left(B^{-1/2} A^T \mathbf{S}\mathbf{S}^T A B^{-1/2}\right)}{\|B^{-1/2} A^T \mathbf{S}\|_F^2}.$$

Name	B	S	$\mathbf{S}$	p_i	$1-\rho$
Kaczmarz	I	e_i	I	$\|A_{i:}\|_2^2 / \|A\|_F^2$	$\lambda_{\min}(A^T A)/\|A\|_F^2$
CD $\|Ax-b\|_2^2$	$A^T A$	$A_{:i}$	A	$\|A_{:i}\|_2^2 / \|A\|_F^2$	$\lambda_{\min}(A^T A)/\|A\|_F^2$
CD $x^T Ax/2 - x^T b$	A	e_i	I	$A_{ii}/\mathbf{Tr}(A)$	$\lambda_{\min}(A)/\mathbf{Tr}(A)$

New possibilities suggested:

- Covers new cases, e.g., $S = \alpha_i e_i + \alpha_j e_j$

Unifying previous methods & analysis

$$p_i = \frac{s_i^T A B^{-1} A^T s_i}{\|B^{-1/2} A^T \mathbf{S}\|_F^2} \quad \text{with} \quad \rho = 1 - \frac{\lambda_{\min}\left(B^{-1/2} A^T \mathbf{S}\mathbf{S}^T A B^{-1/2}\right)}{\|B^{-1/2} A^T \mathbf{S}\|_F^2}.$$

Name	B	S	$\mathbf{S}$	p_i	$1-\rho$
Kaczmarz	I	e_i	I	$\|A_{i:}\|_2^2/\|A\|_F^2$	$\lambda_{\min}(A^T A)/\|A\|_F^2$
CD $\|Ax-b\|_2^2$	$A^T A$	$A_{:i}$	A	$\|A_{:i}\|_2^2/\|A\|_F^2$	$\lambda_{\min}(A^T A)/\|A\|_F^2$
CD $x^T Ax/2 - x^T b$	A	e_i	I	$A_{ii}/\mathbf{Tr}(A)$	$\lambda_{\min}(A)/\mathbf{Tr}(A)$

New possibilities suggested:

- Covers new cases, e.g., $S = \alpha_i e_i + \alpha_j e_j$
- For $B = I$, then ideally $\mathbf{S}^T \approx A^\dagger$ then $\rho \approx 1 - 1/n$. If we have a preconditioner $P \approx A^\dagger$ then $S =$ sample rows of P.

Unifying previous methods & analysis

$$p_i = \frac{s_i^T A B^{-1} A^T s_i}{\|B^{-1/2} A^T \mathbf{S}\|_F^2} \quad \text{with} \quad \rho = 1 - \frac{\lambda_{\min}\left(B^{-1/2} A^T \mathbf{S}\mathbf{S}^T A B^{-1/2}\right)}{\|B^{-1/2} A^T \mathbf{S}\|_F^2}.$$

Name	B	S	$\mathbf{S}$	p_i	$1-\rho$
Kaczmarz	I	e_i	I	$\|A_{i:}\|_2^2/\|A\|_F^2$	$\lambda_{\min}(A^T A)/\|A\|_F^2$
CD $\|Ax-b\|_2^2$	$A^T A$	$A_{:i}$	A	$\|A_{:i}\|_2^2/\|A\|_F^2$	$\lambda_{\min}(A^T A)/\|A\|_F^2$
CD $x^T Ax/2 - x^T b$	A	e_i	I	$A_{ii}/\mathbf{Tr}(A)$	$\lambda_{\min}(A)/\mathbf{Tr}(A)$

New possibilities suggested:

- Covers new cases, e.g., $S = \alpha_i e_i + \alpha_j e_j$
- For $B = I$, then ideally $\mathbf{S}^T \approx A^\dagger$ then $\rho \approx 1 - 1/n$. If we have a preconditioner $P \approx A^\dagger$ then $S =$ sample rows of P.

Guassian based sampling

Why not make S a continuous random matrix?
Sample $S = \xi \sim N(0, \Sigma)$ a normal random variable then

$$Z = A^T S(S^T A B^{-1} A^T S)^{-1} S^T A = \frac{A^T \xi \xi^T A}{\xi^T A B^{-1} A^T \xi}.$$

$$x_{k+1} = x_k - \frac{\xi^T (A x_k - b)}{\xi^T A B^{-1} A^T \xi} B^{-1} A^T \xi.$$

Iteration cost $O(\text{product } A^T \cdot \xi)$.
The convergence rate determined by

$$\rho = 1 - \lambda_{min}(B^{-1/2} \mathbf{E}[Z] B^{-1/2}) = 1 - \lambda_{\min}\left(\mathbf{E}\left[\frac{\bar{\xi}\bar{\xi}^T}{\bar{\xi}^T \bar{\xi}}\right]\right),$$

where $\bar{\xi} = B^{-1/2} A^T \xi \sim N(0, \Omega)$, and $\Omega = B^{-1/2} A \Sigma A^T B^{-1/2}$.

Guassian based sampling

Why not make S a continuous random matrix?
Sample $S = \xi \sim N(0, \Sigma)$ a normal random variable then

$$Z = A^T S(S^T AB^{-1}A^T S)^{-1} S^T A = \frac{A^T \xi\xi^T A}{\xi^T AB^{-1}A^T \xi}.$$

$$x_{k+1} = x_k - \frac{\xi^T (Ax_k - b)}{\xi^T AB^{-1}A^T \xi} B^{-1}A^T \xi.$$

Iteration cost $O(\text{product } A^T \cdot \xi)$.
The convergence rate determined by

$$\rho = 1 - \lambda_{min}(B^{-1/2}\mathbf{E}\left[Z\right] B^{-1/2}) = 1 - \lambda_{\min}\left(\mathbf{E}\left[\frac{\bar{\xi}\bar{\xi}^T}{\bar{\xi}^T \bar{\xi}}\right]\right),$$

where $\bar{\xi} = B^{-1/2}A^T \xi \sim N(0, \Omega)$, and $\Omega = B^{-1/2}A\Sigma A^T B^{-1/2}$.

New Gaussian Methods

Sample $S = \xi \sim N(0, \Sigma)$. Let $\eta \sim N(0, I)$.
Gauss. Kaczmarz $B = I$ and $\Sigma = I$

$$x_{k+1} = \arg\min_{x \in \mathbb{R}^n} \|x - x_k\|_2^2 \quad \text{subject to} \quad \eta^T (Ax - b) = 0.$$

Gauss Least-squares $B = A^T A$ and $\Sigma = AA^T$

$$x_{k+1} = \arg\min_{x \in \mathbb{R}^n} \|Ax - b\|_2^2 \quad \text{subject to} \quad x = x_k + t\,\eta, \quad t \in \mathbb{R}.$$

Gauss. Pos. Def. $B = A$ and $\Sigma = I$

$$x_{k+1} = \arg\min_{x \in \mathbb{R}^n} 1/2 x^T A x - x^T b \quad \text{subject to} \quad x = x_k + t\,\eta, \quad t \in \mathbb{R}.$$

Table of Contents

Dense Overdetermined Gaussian Matrix

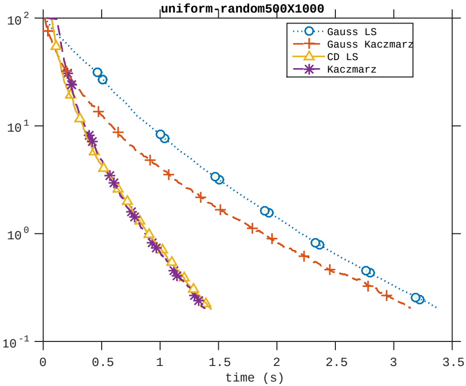


Figure : $m \times n = 500X1000$, A = `randn`(m, n)

Dense matrix $\Rightarrow$ High iteration cost of Gaussian methods $O(A \cdot \eta)$.

Sparse Square Gaussian Matrix

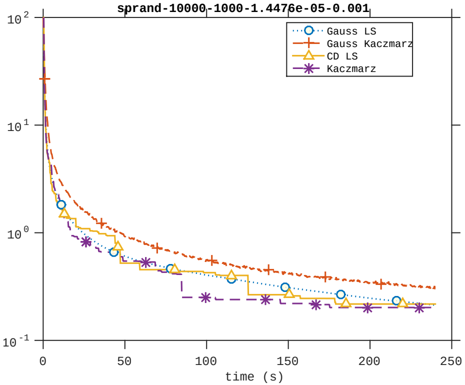
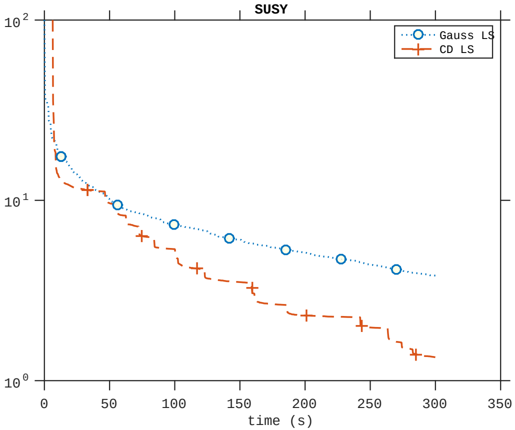


Figure : $m \times n = 10000 \times 1000$, density $= 1/\sqrt{m} = 1\%$; $\kappa = \sqrt{n}$; `A = sprandsym(n,density,rc)`

Sparse matrices $\Rightarrow$ Guass methods become competitive.

Regression SUSY

The SUSY[3] Classification problem



Solving least-squares regression $\min \|Ax - b\|_2^2$ with $m = 5 \cdot 10^6$ and $n = 18$

[3] Baldi, P., P. Sadowski, and D. Whiteson. Searching for Exotic Particles in High-energy Physics with Deep Learning. Nature Communications 5 (July 2, 2014)

Table of Contents

Why iteratively invert a matrix $A \in \mathbb{R}^{n\times n}$?

- Needed to calculate Schur complements, a orojection operator...etc
- Iterative is good when we can tolerate an error
- Iterative is good when we have an initial guess $X_0 \approx A^{-1}$.
- Staging for randomized variable metric methods and randomized Preconditioning.

New context: $A \in \mathbb{R}^{n\times n}$ non-singular.

Framework

- Assume we observe $S^T A$ where S is random.
- Given $X_k \approx A \in \mathbb{R}^{n\times n}$, we want to iteratively calculate

$$X_{k+1} = \text{update\_formula}(S^T A, X_k)$$

such that $X_{k+1} \to A^{-1}$.

$$\boxed{X_{k+1} = \arg\min_{X\in\mathbb{R}^{n\times n}} \|X - X_k\|^2_{\text{Frobenius}(B)} \quad \text{s.t.} \quad S^T AX = X,}$$

The solution

$$\begin{aligned} X_{k+1} &= X_k + \text{proj}_{B^{-1}\mathbf{Range}(A^T S)}(A^{-1} - X_k) \\ &= X_k + B^{-1}A^T S(S^T AB^{-1}A^T S)^{-1} S^T A(A^{-1} - X_k) \\ &= X_k + B^{-1}A^T S \underbrace{(S^T AB^{-1}A^T S)^{-1}}_{\text{Invert small matrix}} S^T(I - AX_k). \end{aligned}$$

What about the symmetric case $A^T = A$?

Framework

- Assume we observe $S^T A$ where S is random.
- Given $X_k \approx A \in \mathbb{R}^{n\times n}$, we want to iteratively calculate

$$X_{k+1} = \text{update\_formula}(S^T A, X_k)$$

such that $X_{k+1} \to A^{-1}$.

$$\boxed{X_{k+1} = \arg\min_{X\in\mathbb{R}^{n\times n}} \|X - X_k\|^2_{\text{Frobenius}(B)} \quad \text{s.t.} \quad S^T AX = X,}$$

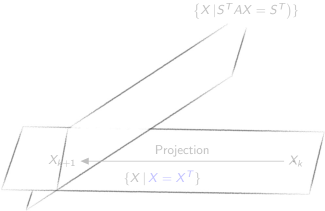
The solution

$$\begin{aligned} X_{k+1} &= X_k + \text{proj}_{B^{-1}\mathbf{Range}(A^T S)}(A^{-1} - X_k) \\ &= X_k + B^{-1}A^T S(S^T AB^{-1}A^T S)^{-1} S^T A(A^{-1} - X_k) \\ &= X_k + B^{-1}A^T S \underbrace{(S^T AB^{-1}A^T S)^{-1}}_{\text{Invert small matrix}} S^T (I - AX_k). \end{aligned}$$

What about the symmetric case $A^T = A$?

Symmetric matrices

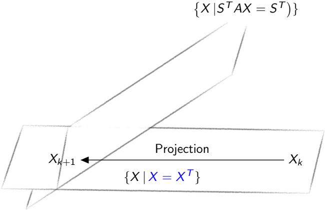
$$X_{k+1} = \arg\min_{X \in \mathbb{R}^{n\times n}} \|X - X_k\|^2_{\text{Frobenius}(B)}$$
$$\text{s.t.} \quad S^T A X = X$$
$$X = X^T$$



[4] Gower and Gondzio 2014

Symmetric matrices

$$X_{k+1} = \arg\min_{X \in \mathbb{R}^{n\times n}} \|X - X_k\|^2_{\text{Frobenius}(B)}$$
$$\text{s.t.} \quad S^T A X = X$$
$$X = X^T$$



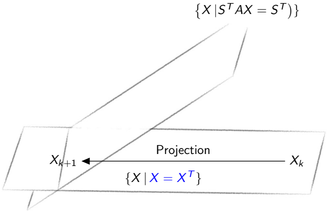
Solution:[4]

$$X_{k+1} = X_k + \text{proj}_{B^{-1}\mathbf{Range}(AS)}(X_k - A^{-1})\text{proj}_{B^{-1}\mathbf{Range}(AS)} - (X_k - A^{-1})\text{proj}_{B^{-1}\mathbf{Range}(AS)} - \text{proj}_{B^{-1}\mathbf{Range}(AS)}(X_k - A^{-1})$$

[4] Gower and Gondzio 2014

Symmetric matrices

$$X_{k+1} = \arg \min_{X \in \mathbb{R}^{n\times n}} \|X - X_k\|^2_{\text{Frobenius}(B)}$$
$$\text{s.t.} \quad S^T AX = X$$
$$X = X^T$$



Solution:[4]

$$X_{k+1} = X_k + \text{proj}_{B^{-1}\textbf{Range}(AS)}(X_k - A^{-1})\text{proj}_{B^{-1}\textbf{Range}(AS)} - (X_k - A^{-1})\text{proj}_{B^{-1}\textbf{Range}(AS)} - \text{proj}_{B^{-1}\textbf{Range}(AS)}(X_k - A^{-1})$$

[4] Gower and Gondzio 2014

Theorem (Convergence)

Let S be equal to a column of a full rank matrix $\mathbf{S} := [s_1, \ldots, s_n]$ with probability $\|B^{-1/2}As_i\|^2 / \|B^{-1/2}A\mathbf{S}\|_F^2$. Then from a given $X_0 \in \mathbb{R}^{n\times n}$, the iteration

$$\begin{aligned} X_{k+1} = X_k &+ proj_{B^{-1}\mathbf{Range}(AS)}(X_k - A^{-1})proj_{B^{-1}\mathbf{Range}(AS)} \\ &- (X_k - A^{-1})proj_{B^{-1}\mathbf{Range}(AS)} - proj_{B^{-1}\mathbf{Range}(AS)}(X_k - A^{-1}) \end{aligned}$$

converges with

$$\mathbf{E}\left[\|X_k - A^{-1}\|^2{}_{Frob(B)}\right] = \left(I - \frac{1}{\kappa_F^2(B^{-1/2}A\mathbf{S})}\right)^k \|X_0 - A^{-1}\|^2{}_{Frob(B)},$$

where $\kappa_F(B^{-1/2}A\mathbf{S}) = \|B^{-1/2}A\mathbf{S}\|_F \|\mathbf{S}^{-1}A^{-1}B^{1/2}\|_F$.

Self-preconditioning Method: This suggests that $\mathbf{S} \approx A^{-1}$. But $X_k \approx A^{-1}$ so try $S =$ sample columns of X_k.

Theorem (Convergence)

Let S be equal to a column of a full rank matrix $\mathbf{S} := [s_1, \ldots, s_n]$ with probability $\|B^{-1/2}As_i\|^2 / \|B^{-1/2}A\mathbf{S}\|_F^2$. Then from a given $X_0 \in \mathbb{R}^{n\times n}$, the iteration

$$X_{k+1} = X_k + proj_{B^{-1}\mathbf{Range}(AS)}(X_k - A^{-1})proj_{B^{-1}\mathbf{Range}(AS)} - (X_k - A^{-1})proj_{B^{-1}\mathbf{Range}(AS)} - proj_{B^{-1}\mathbf{Range}(AS)}(X_k - A^{-1})$$

converges with

$$\mathbf{E}\left[\|X_k - A^{-1}\|^2{}_{Frob(B)}\right] = \left(I - \frac{1}{\kappa_F^2(B^{-1/2}A\mathbf{S})}\right)^k \|X_0 - A^{-1}\|^2{}_{Frob(B)},$$

where $\kappa_F(B^{-1/2}A\mathbf{S}) = \|B^{-1/2}A\mathbf{S}\|_F \|\mathbf{S}^{-1}A^{-1}B^{1/2}\|_F$.

Self-preconditioning Method: This suggests that $\mathbf{S} \approx A^{-1}$. But $X_k \approx A^{-1}$ so try $S =$ sample columns of X_k.

Initial experiments A positive definite

Newton-Schulz: $X_0 = A^T/(0.99\|A^TA\|_2), \quad X_{k+1} = 2X_k - X_kAX_k.$

Self-preconditioning Method: $B = A,\ X_0 = I, \quad X_{k+1} = \mathrm{proj}_S + (I - \mathrm{proj}_S A)X_k(I - A\mathrm{proj}_S),$ where $S =$ sample columns of X_k.

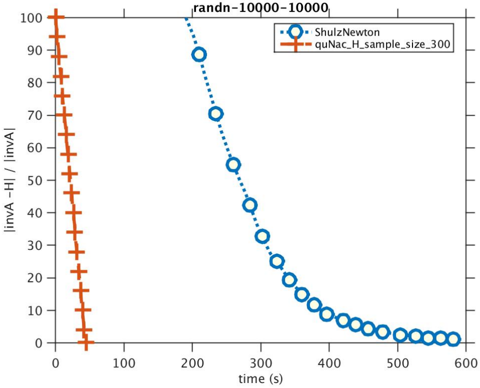


Figure : $n = 10'000$, with $nnz = 10^8$, A `= randn(`n,n`), A = (A')*A;`

Towards Randomized Preconditioning

Initialize $X_0 \in \mathbb{R}^{n\times n}$ and $x_0 \in \mathbb{R}^n$.
While (`stopping_criteria`)

$$S_k = \texttt{sample\_function}(A, X_k)$$
$$x_{k+1} = \arg\min_{x\in\mathbb{R}^n} \|x - x_k\|_B^2 \text{ s.t } S_k^T A x = S_k^T b$$
$$X_{k+1} = \arg\min_{X\in\mathbb{R}^{n\times n}} \|X - X_k\|_{\text{Frobenius}(B)} \text{ s.t } S_k^T A X = S_k^T, X = X^T.$$
$$k = k+1$$

end

What if $(A_k)_k$ is a slowly evolving sequence (like a Hessian matrix)?

Conclusion

- A natural framework for designing and analysing randomized iterative methods
- Analyse previous methods through one Theorem
- New Gaussian methods, with potential on sparse problems
- New randomized matrix inversion methods.
- Paving a path towards randomized preconditioning.

References

Gower, Robert M., Richtárik, Peter, April 2015.
Randomized Iterative Methods for Linear Systems (in progress)

Gower, Robert M., Richtárik, Peter, April 2015.
Randomized Iterative Methods for Inverting Matrices (in progress)

Qu, Z., Richtárik, P., Takáč, M., & Fercoq, O. (2015).
SDNA: Stochastic Dual Newton Ascent for Empirical Risk Minimization.

Gower, Robert M., Gondzio, Jacek (2014).
Action constrained quasi-Newton methods (in progress)