Randomized iterative methods for linear systems

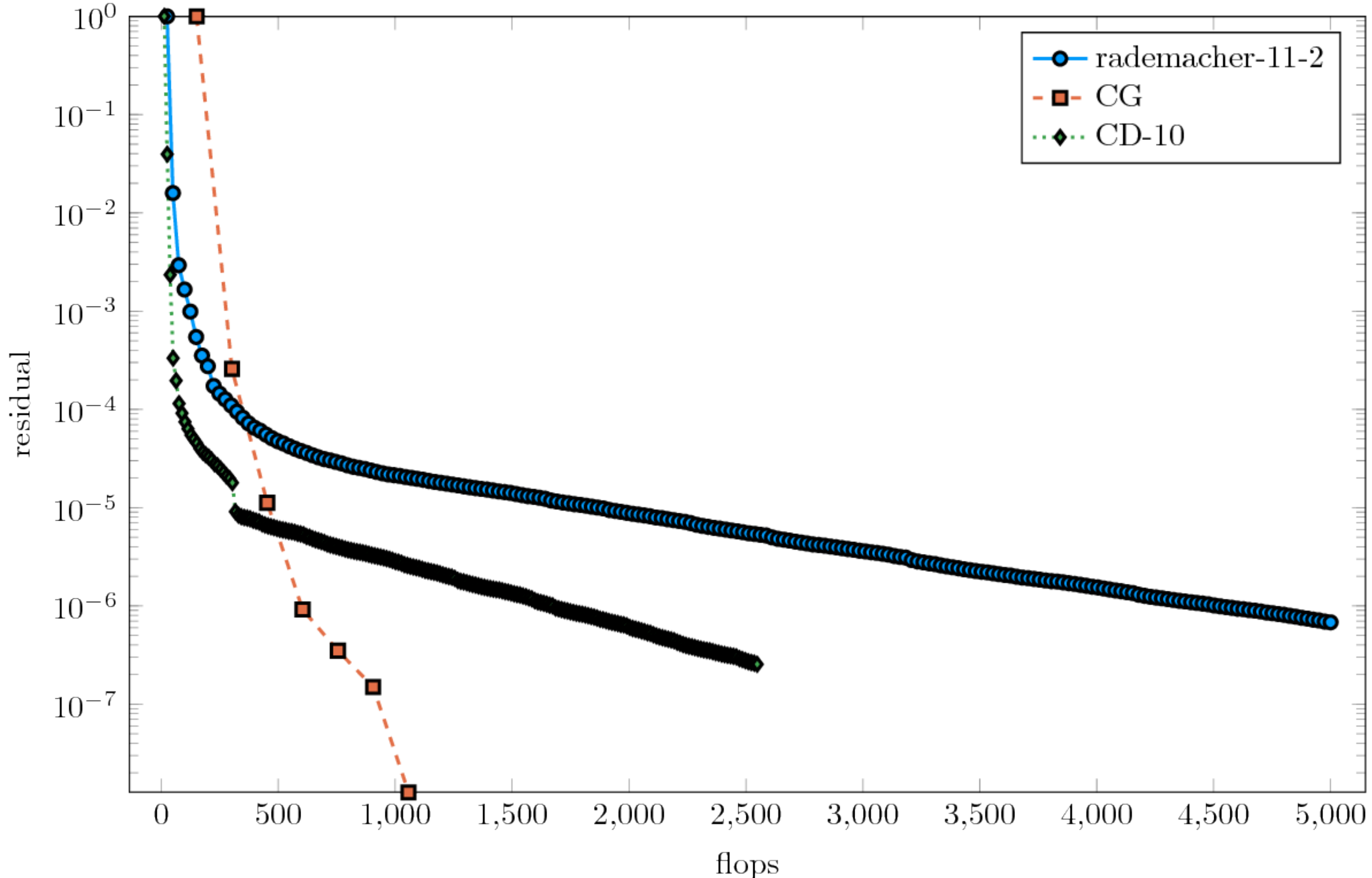
Robert Mansel Gower

IMA Leslie Fox Prize Meeting, Strathclyde, June 2017

Motivation

Large scale Kernel Ridge Regression

$$\min_x \frac{1}{2}||Ax - y||_2^2 + \frac{\lambda}{2}||x||_2^2$$



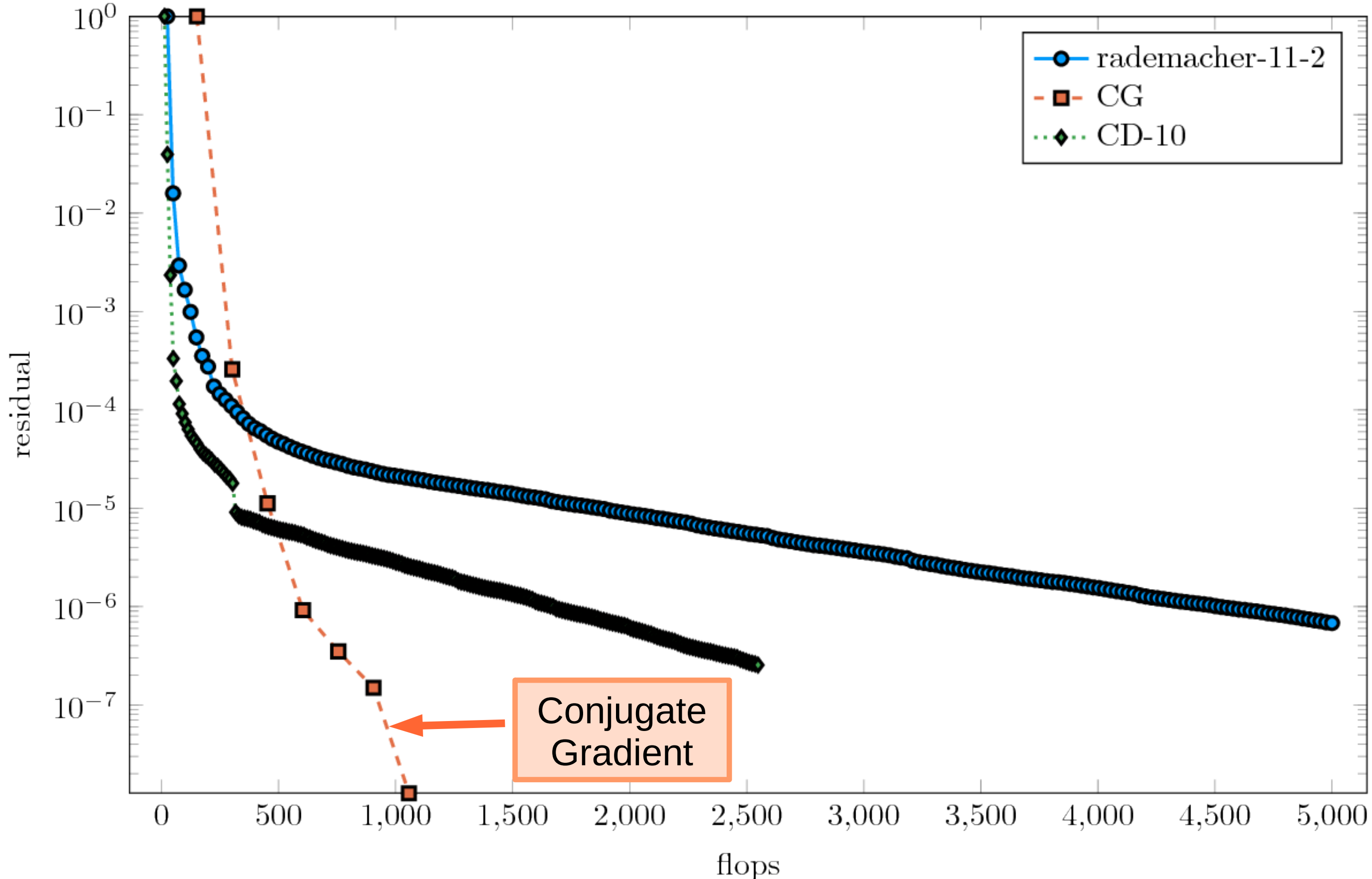
Problem: a9a

$A \in \mathbb{R}^{32,561 \times 123}$

Origin: LIBSVM

Large scale Kernel Ridge Regression

$$\min_x \frac{1}{2}||Ax - y||_2^2 + \frac{\lambda}{2}||x||_2^2$$



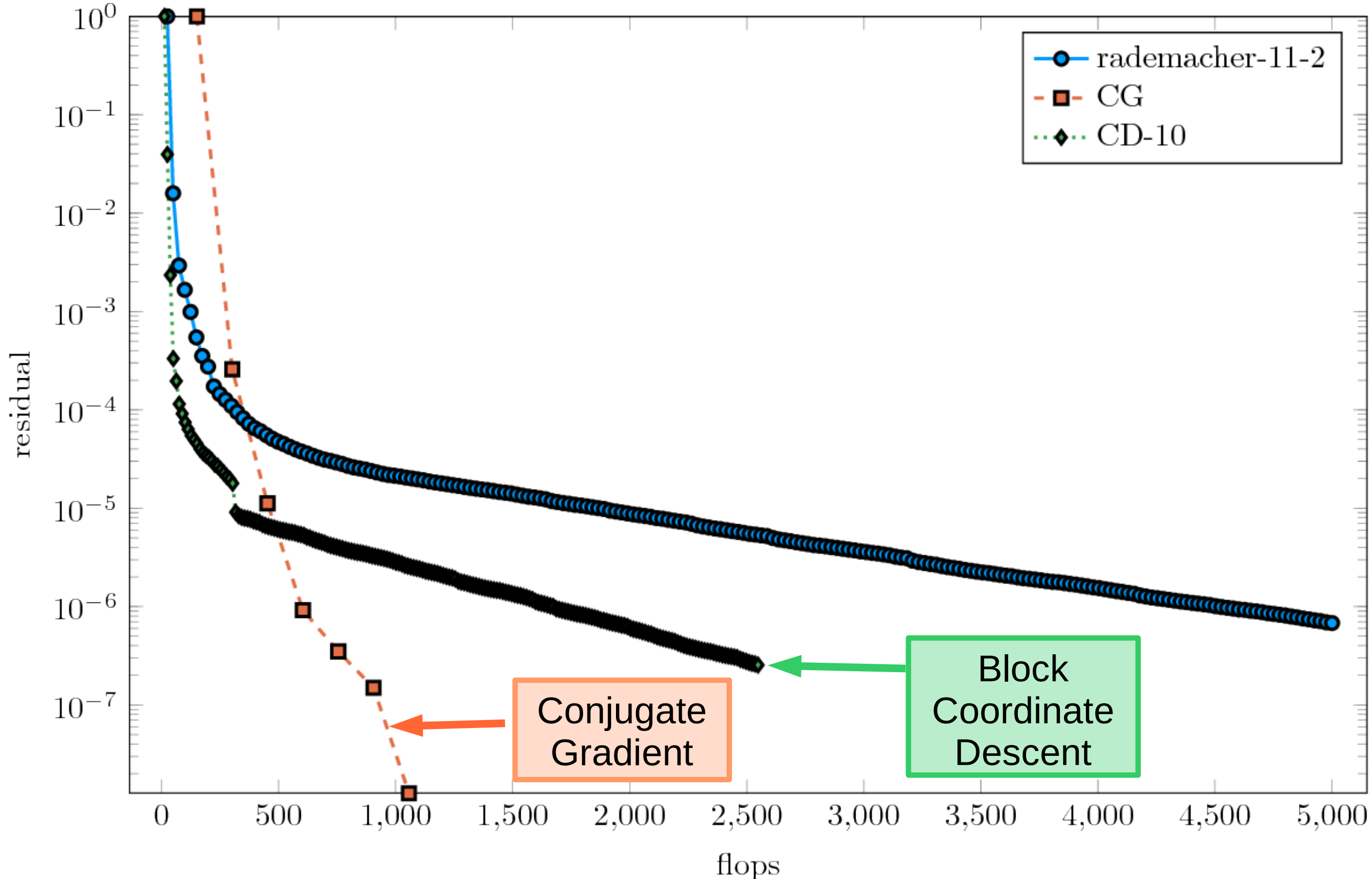
Problem: a9a

$A \in \mathbb{R}^{32{,}561 \times 123}$

Origin: LIBSVM

Large scale Kernel Ridge Regression

$$\min_x \frac{1}{2}||Ax - y||_2^2 + \frac{\lambda}{2}||x||_2^2$$



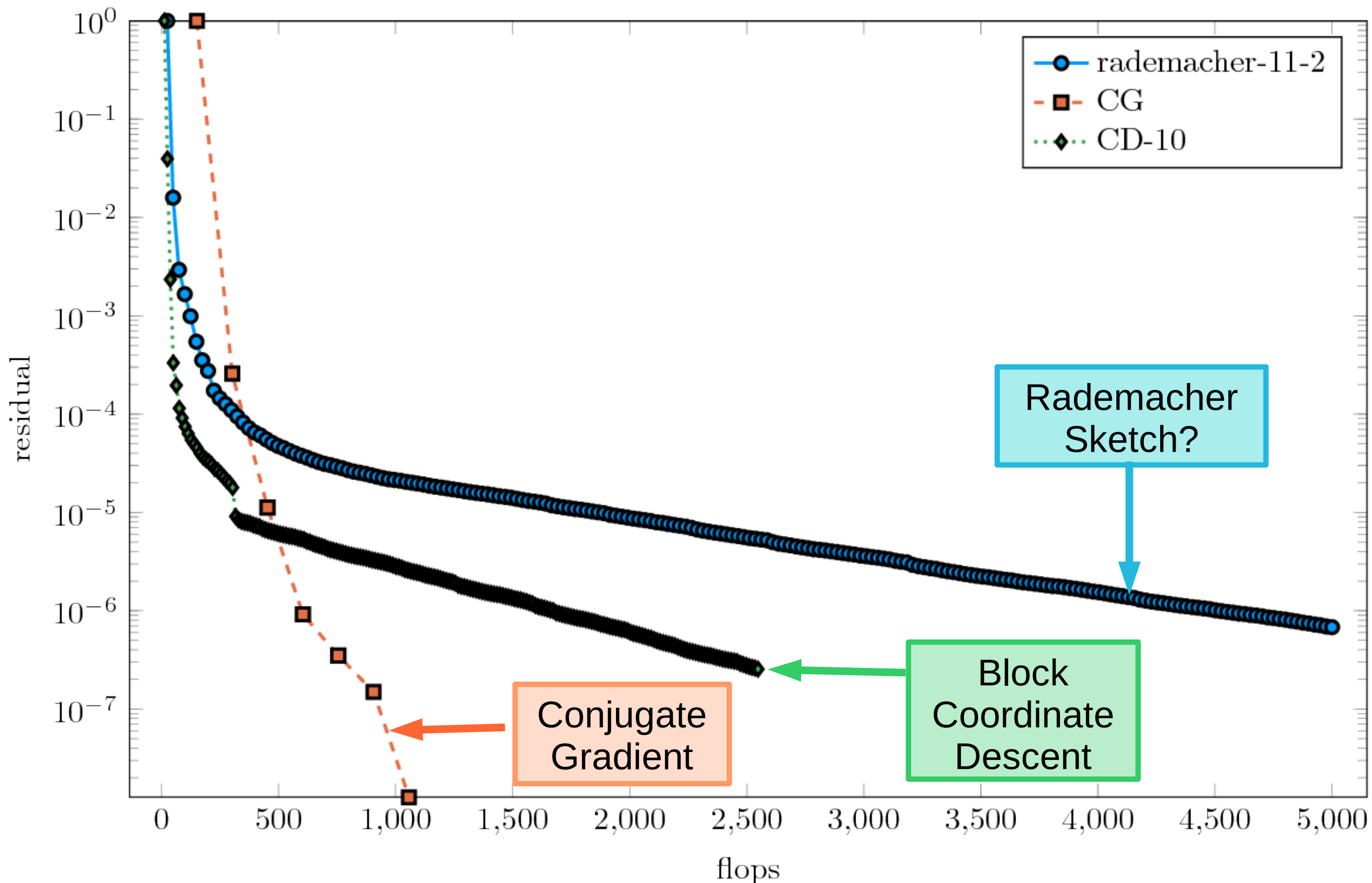
Problem: a9a

$A \in \mathbb{R}^{32,561 \times 123}$

Origin: LIBSVM

Large scale Kernel Ridge Regression

$$\min_x \frac{1}{2}||Ax - y||_2^2 + \frac{\lambda}{2}||x||_2^2$$



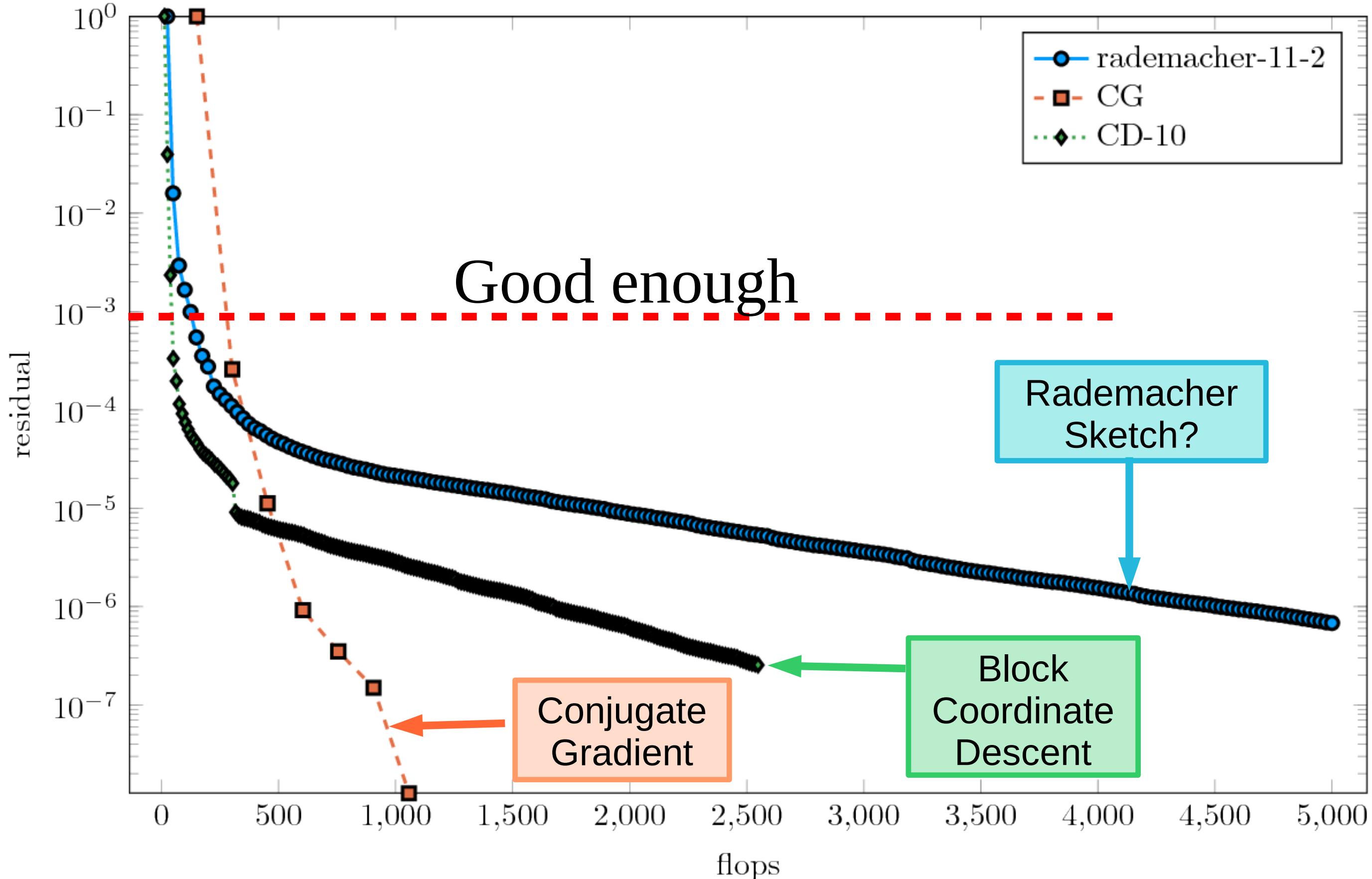
Problem: a9a

$A \in \mathbb{R}^{32,561 \times 123}$

Origin: LIBSVM

Large scale Kernel Ridge Regression

$$\min_x \frac{1}{2}||Ax - y||_2^2 + \frac{\lambda}{2}||x||_2^2$$



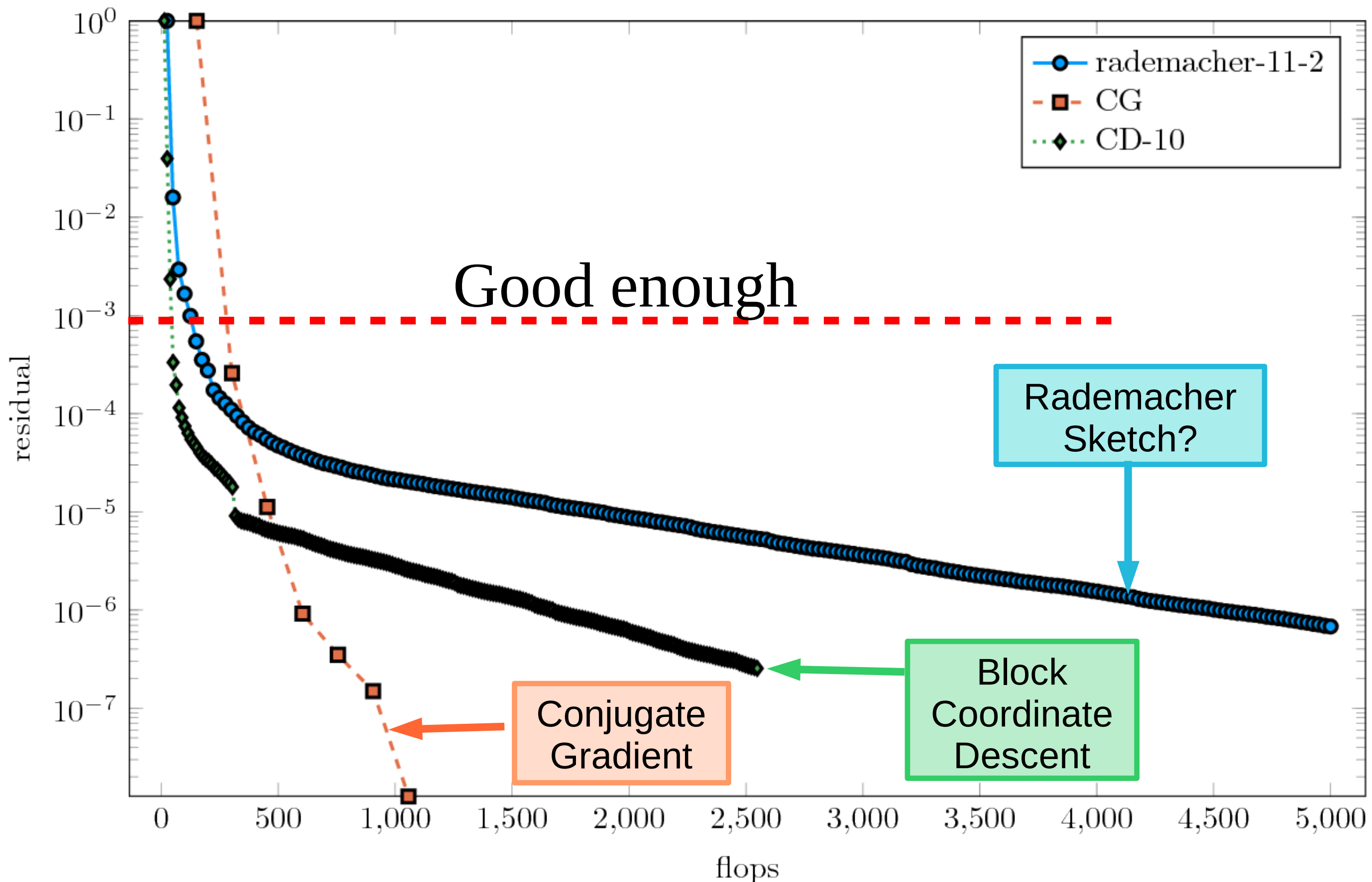
Problem: a9a

$A \in \mathbb{R}^{32,561 \times 123}$

Origin: LIBSVM

Large scale Kernel Ridge Regression

$$\min_{x} \frac{1}{2}||Ax - y||_2^2 + \frac{\lambda}{2}||x||_2^2$$

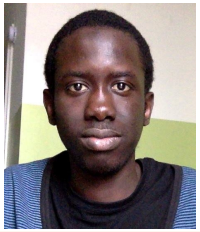


Problem: a9a

$A \in \mathbb{R}^{32,561 \times 123}$

Origin: LIBSVM

GitHub: BigRidge



Cheikh S. Toure

Linear Systems

The Problem

$$\underbrace{\overbrace{A}^{n}x = b}_{m \text{ rows}}$$

n

$\in \mathbb{R}^n$

$$m \left\{ \; Ax = b \; \right\} m$$

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

The Problem

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

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

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

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

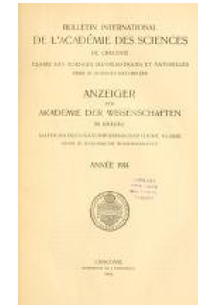
$$x^* := \arg\min ||x||_B^2 \quad \text{subject to} \quad Ax = b$$

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

Randomized Methods

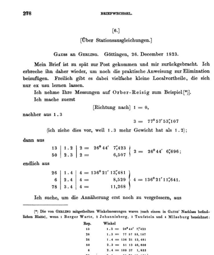
The return of old methods

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


BULLETIN INTERNATIONAL
DE L'ACADÉMIE DES SCIENCES

ANZEIGER

AKADEMIE DER WISSENSCHAFTEN


278 BRIEFWECHSEL.

[6.]

[Über Stationsausgleichungen.]

Gauss an Gerling. Göttingen, 26. December 1823.

Mein Brief ist zu spät zur Post gekommen und mir zurückgebracht. Ich erbreche ihn daher wieder, um noch die praktische Anweisung zur Elimination beizufügen. Freilich gibt es dabei vielfache kleine Localvortheile, die sich nur ex usu lernen lassen.

Ich nehme Ihre Messungen auf Orber-Reisig zum Beispiel[*]).

Ich mache zuerst

[Richtung nach] $1 = 0$,

nachher aus 1.3

$3 = 77^0 57' 53'',107$

(ich ziehe dies vor, weil 1.3 mehr Gewicht hat als 1.2);

dann aus

$$\left.\begin{array}{r|c|l} 13 & 1.2 & 2 = 26^0 44' \; 7'',423 \\ 50 & 2.3 & 2 = \qquad\quad 6,507 \end{array}\right\} 2 = 26^0 44' \; 6'',696;$$

endlich aus

$$\left.\begin{array}{r|c|l} 26 & 1.4 & 4 = 136^0 21' \; 13'',481 \\ 6 & 2.4 & 4 = \qquad\quad 8,529 \\ 78 & 3.4 & 4 = \qquad\quad 11,268 \end{array}\right\} 4 = 136^0 21' \; 11'',641.$$

Ich suche, um die Annäherung erst noch zu vergrössern, aus

[*] Die von Gerling mitgetheilten Winkelmessungen waren (nach einem in Gauss' Nachlass befindlichen Blatte), wenn 1 Berger Warte, 2 Johannisberg, 3 Taufstein und 4 Milseburg bezeichnet:

Rep.	Winkel
13	1.2 = 26° 44′ 7″,423
28	1.3 = 77 57 53,107
26	1.4 = 136 21 13,481
50	2.3 = 51 13 46,680
6	2.4 = 109 37 1,833
78	3.4 = 58 23 18,161.]

Suitable for large scale problems: short recurrence, low iteration cost and low memory

Easy to implement

Easy to analyse, good complexity

Often fits in parallel/distributed architecture

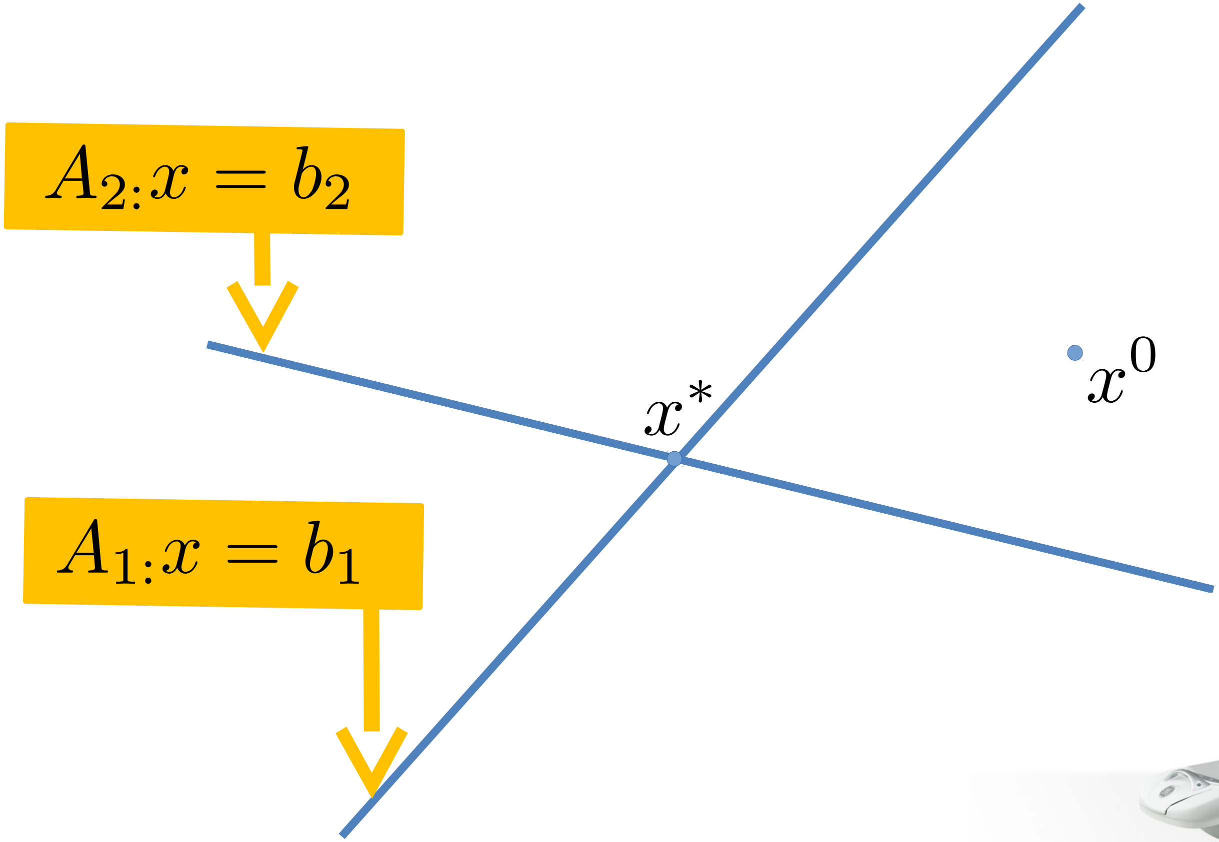
Stochasticity inherent in problem

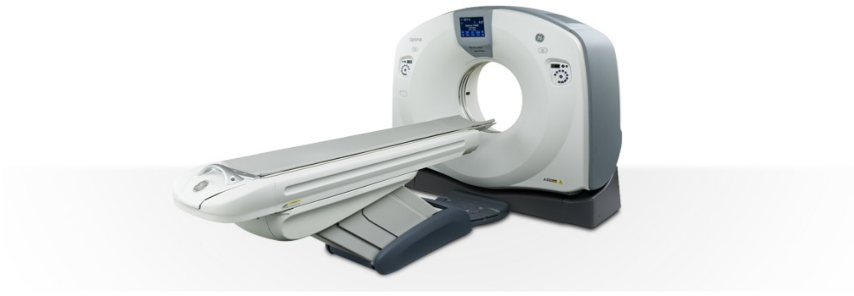
Old Methods

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.

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

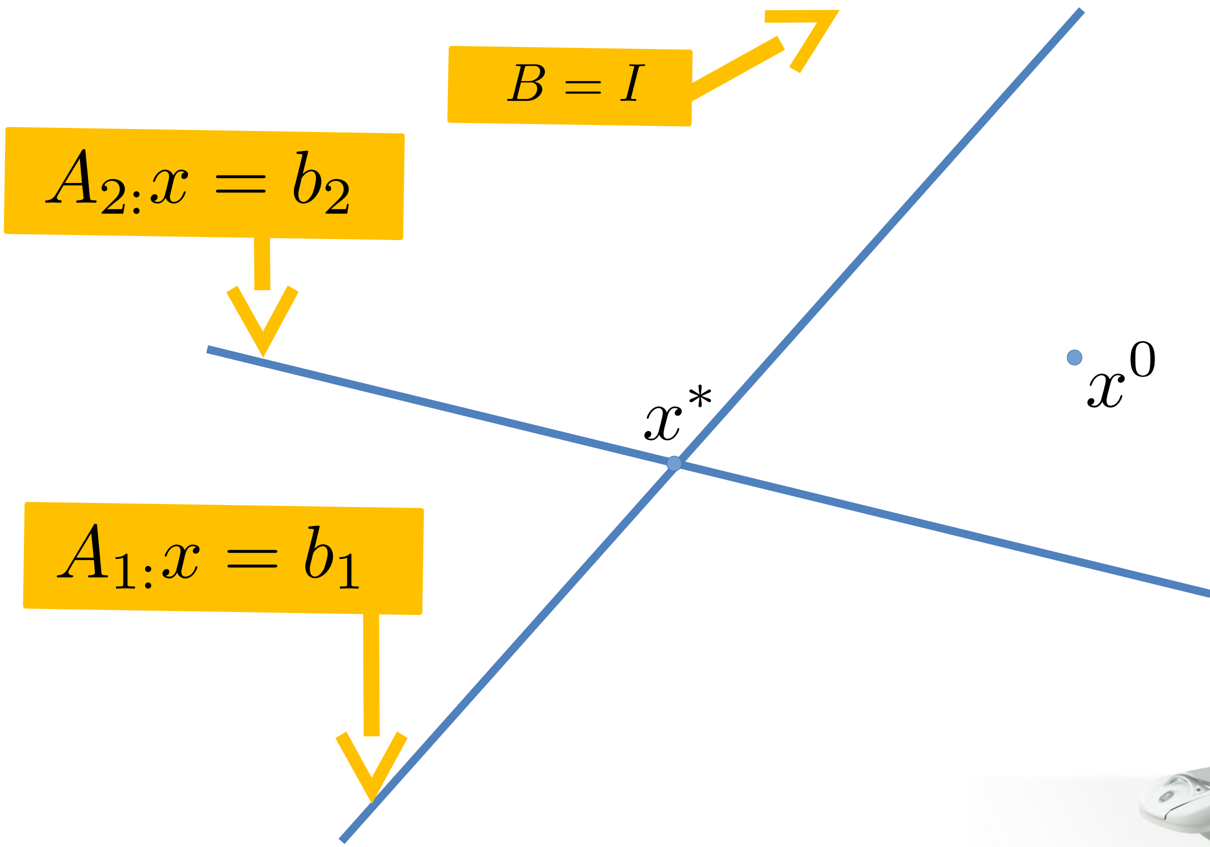


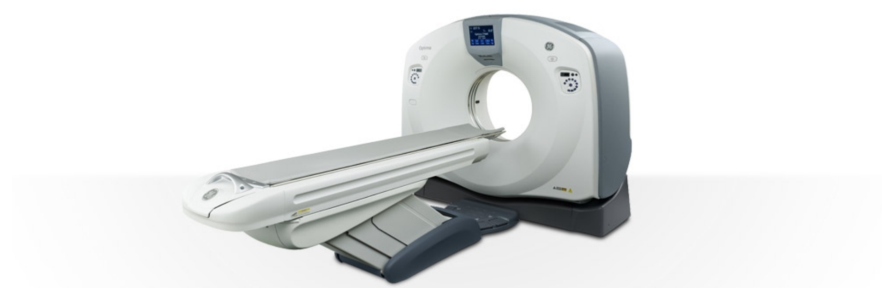


Randomized Kaczmarz

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

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

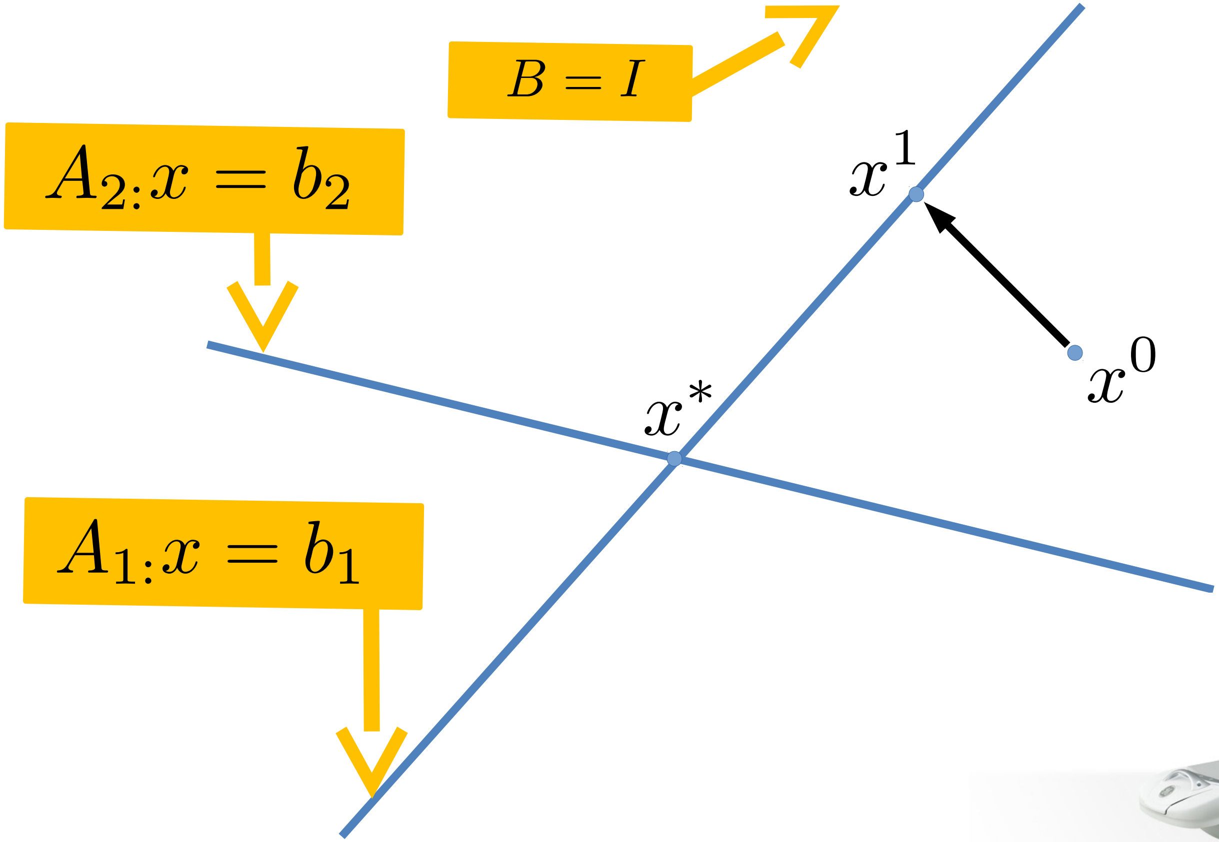


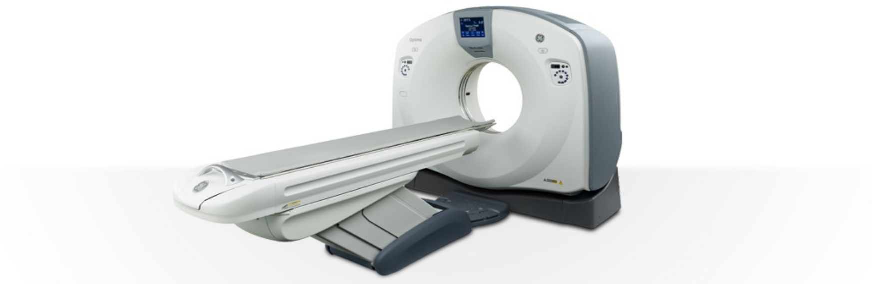


Randomized Kaczmarz

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

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

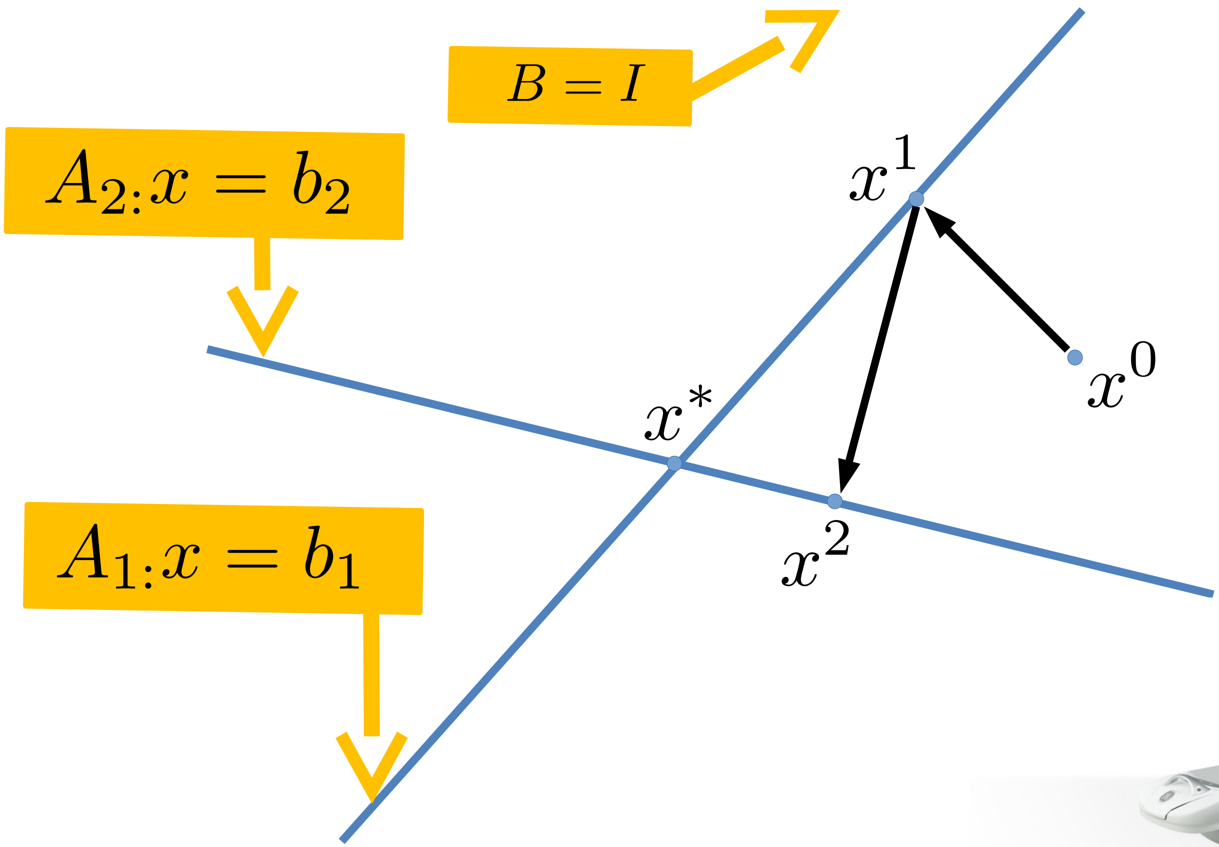


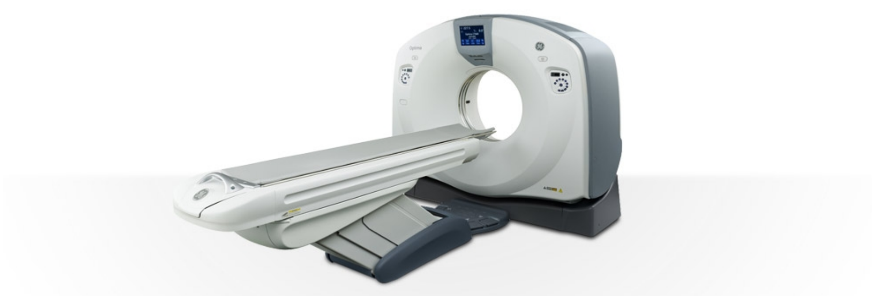


Randomized Kaczmarz

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

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

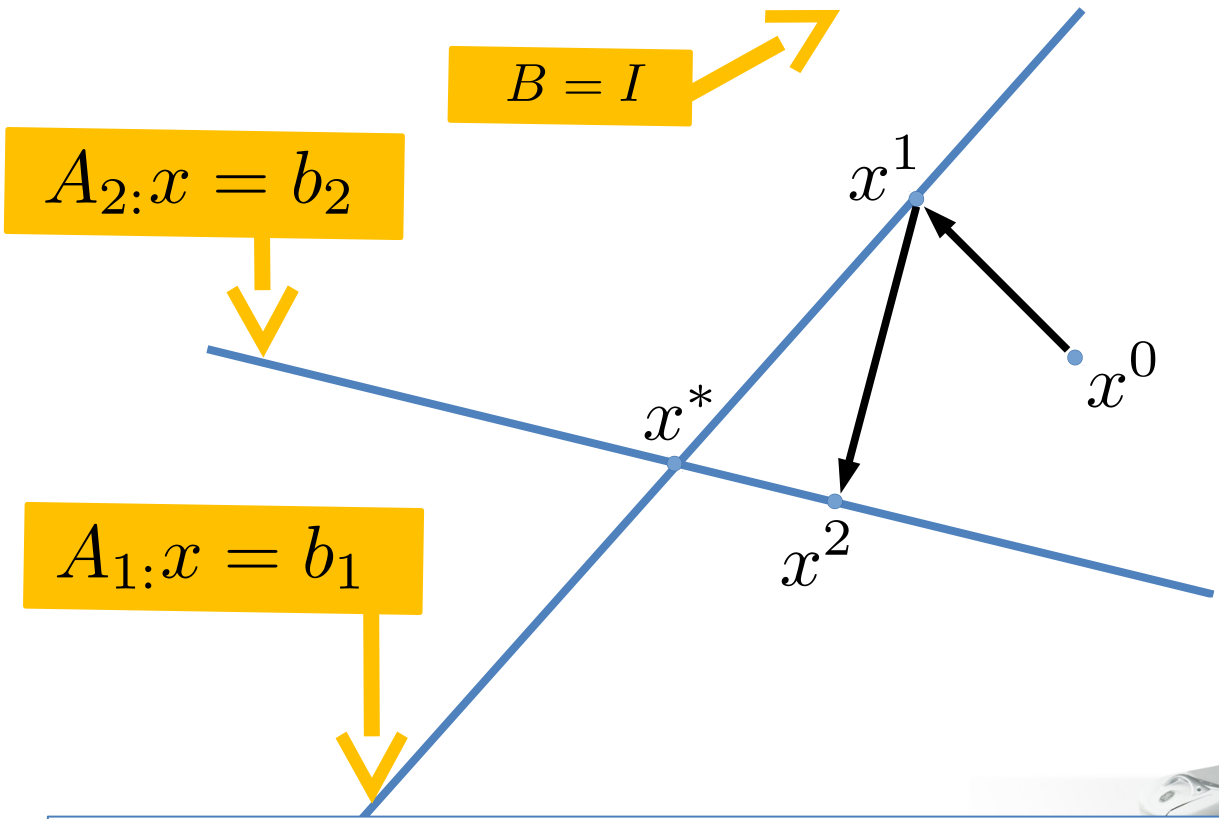


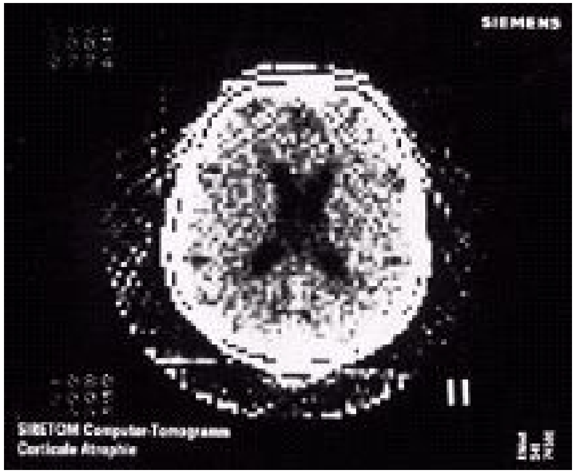


Randomized Kaczmarz

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

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



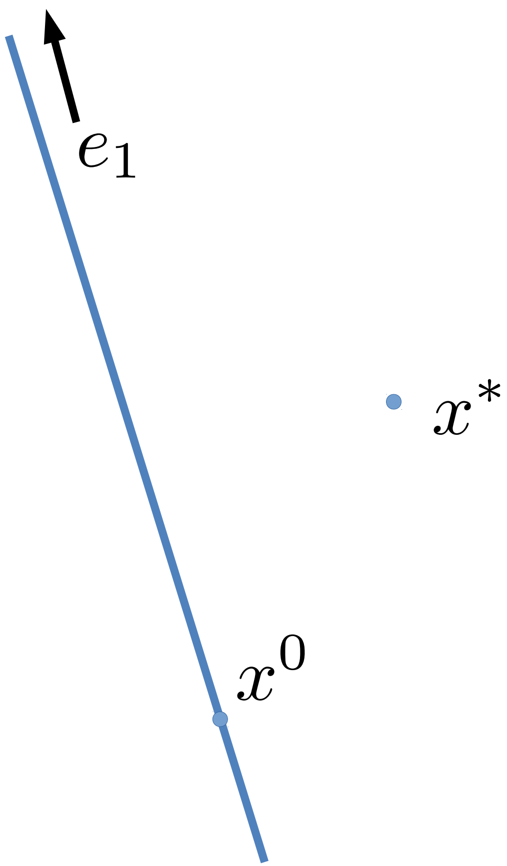


G.N. Hounsfield. Computerized transverse axial scanning (tomography): Part I. description of the system. British Journal Radiology. 1973

Randomized Coordinate Descent

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

$$x^{t+1} = \arg\min ||x - x^*||_A^2 \quad \text{subject to} \quad x = x^t + \alpha e_i$$

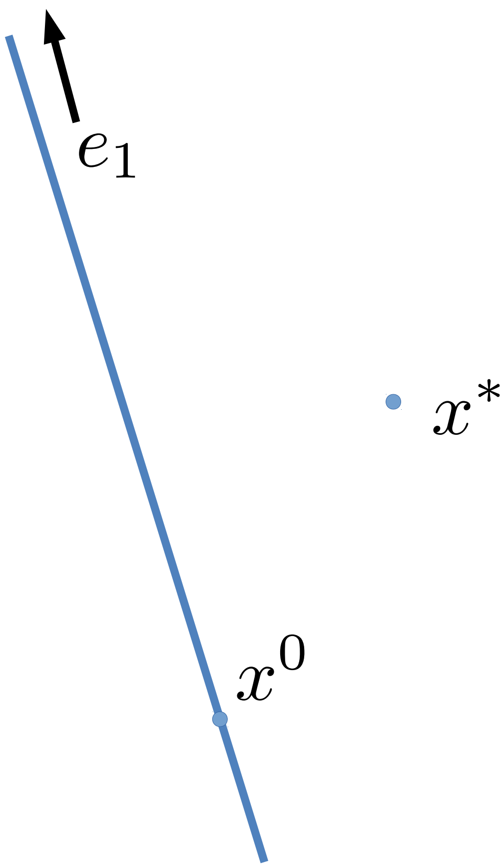


Randomized Coordinate Descent

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

$$x^{t+1} = \arg\min ||x - x^*||_A^2 \quad \text{subject to} \quad x = x^t + \alpha e_i$$

$B = A$

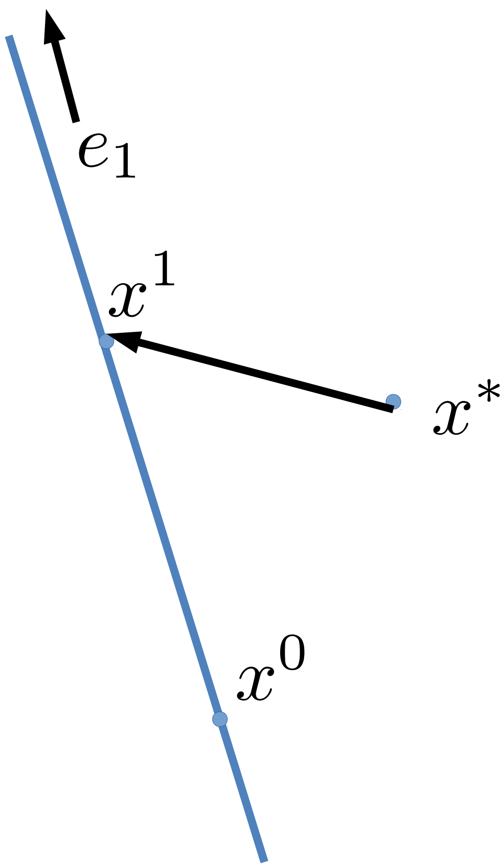


Randomized Coordinate Descent

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

$$x^{t+1} = \arg\min \|x - x^*\|_A^2 \quad \text{subject to} \quad x = x^t + \alpha e_i$$

$B = A$

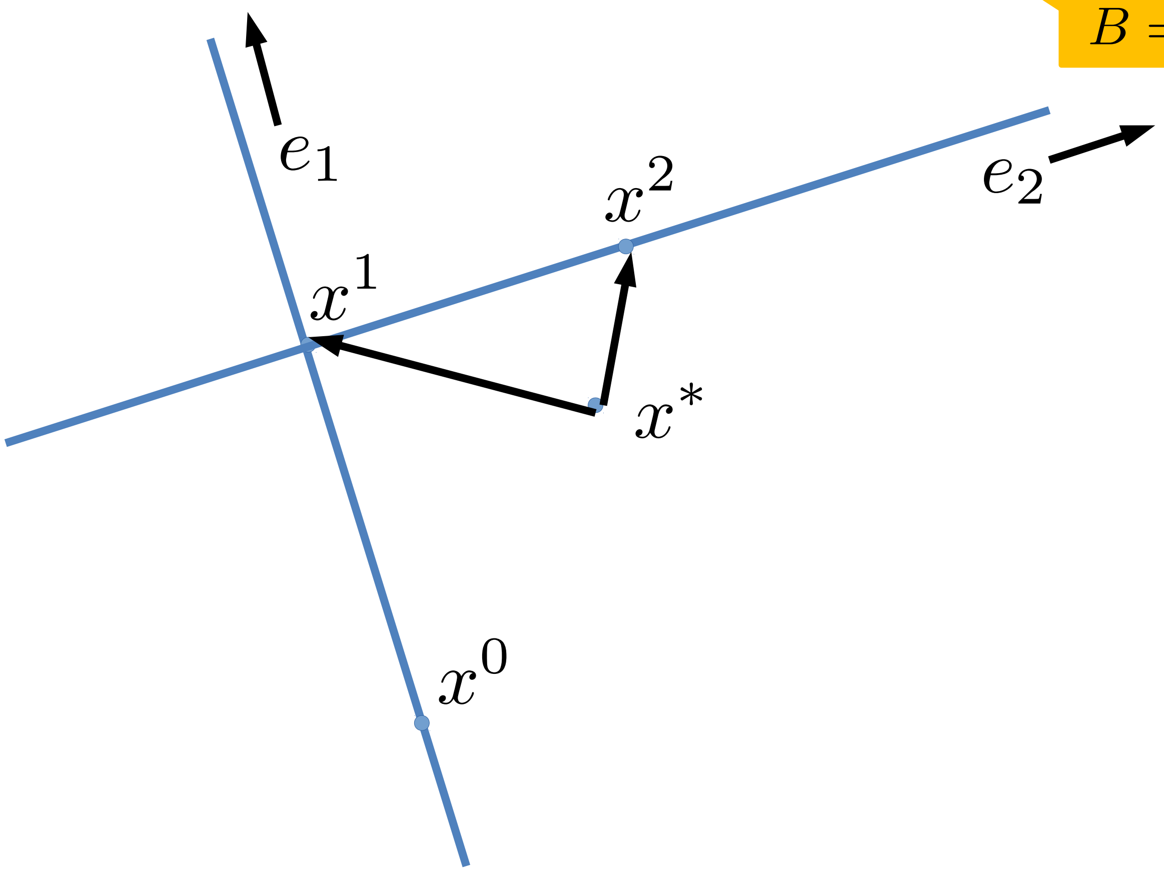


Randomized Coordinate Descent

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

$$x^{t+1} = \arg\min ||x - x^*||_A^2 \quad \text{subject to} \quad x = x^t + \alpha e_i$$

$B = A$

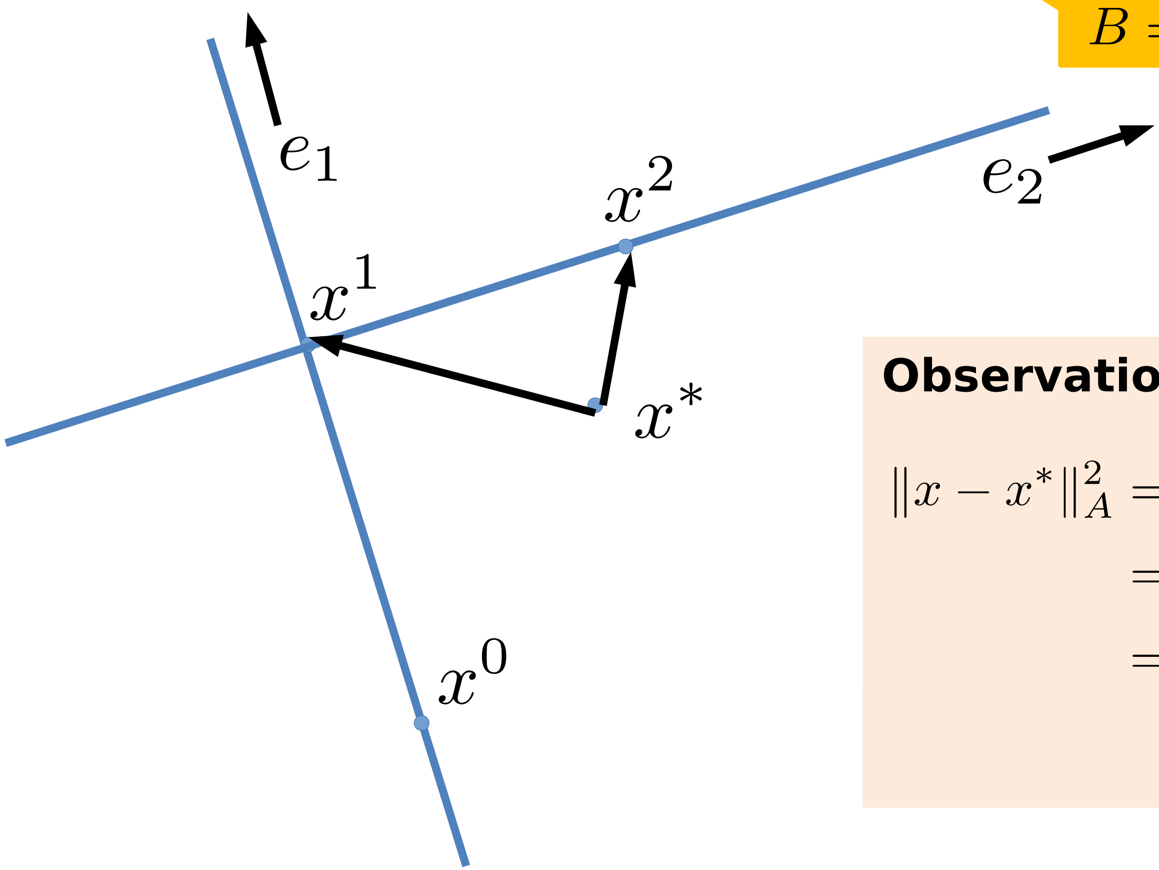


Randomized Coordinate Descent

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

$$x^{t+1} = \arg\min ||x - x^*||_A^2 \quad \text{subject to} \quad x = x^t + \alpha e_i$$

$B = A$



Observation:

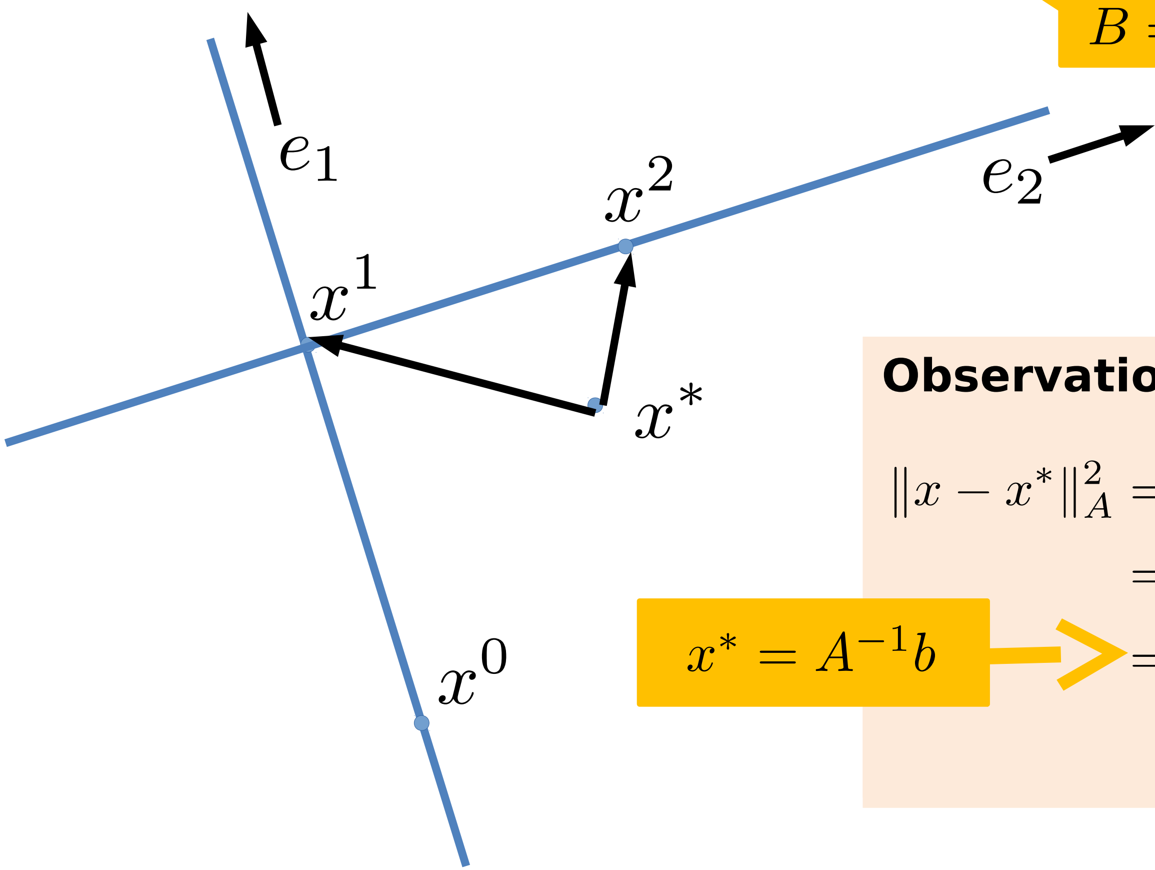
$$\begin{aligned} \|x - x^*\|_A^2 &= (x - x^*)^T A (x - x^*) \\ &= x^T A x - 2(x^*)^T A x + (x^*)^T A x^* \\ &= x^T A x - 2b^T x + b^T x^* \end{aligned}$$

Randomized Coordinate Descent

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

$$x^{t+1} = \arg\min ||x - x^*||_A^2 \quad \text{subject to} \quad x = x^t + \alpha e_i$$

$B = A$



$x^* = A^{-1}b$

Observation:

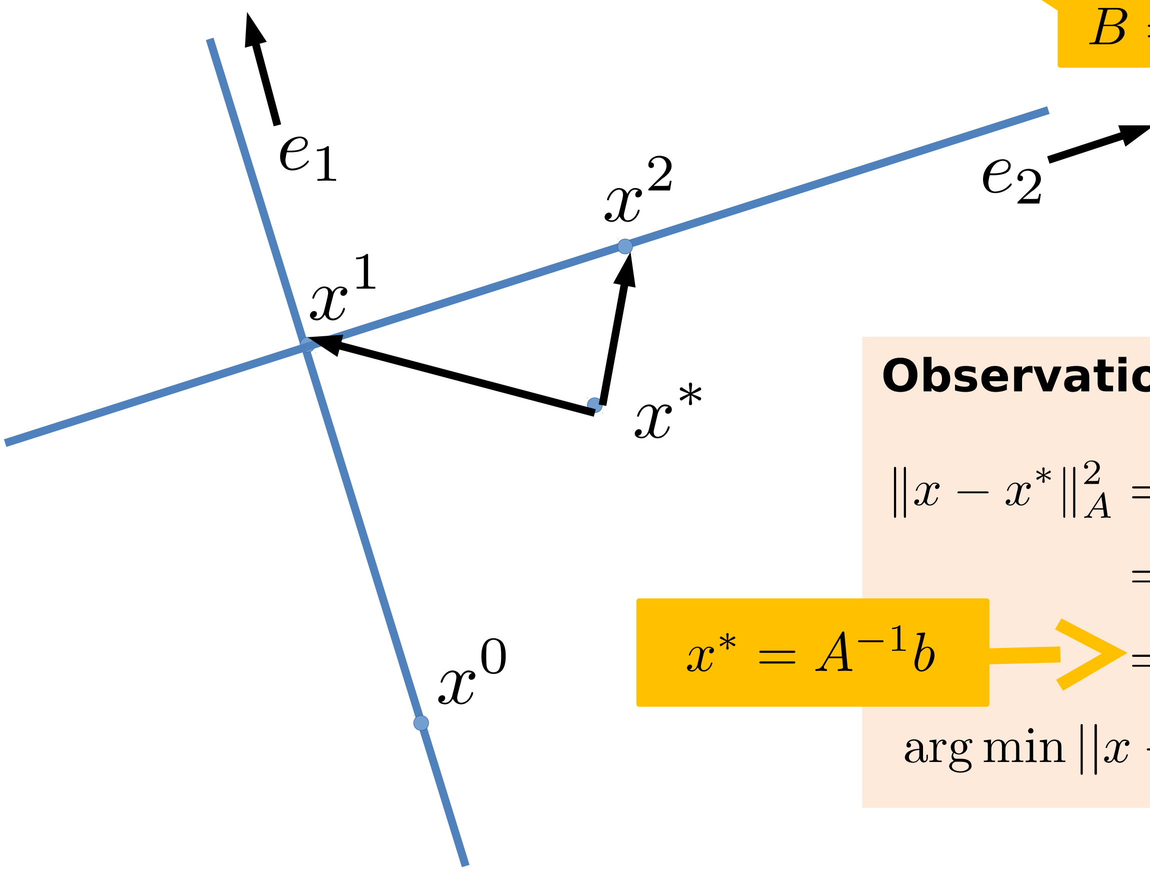
$$\begin{aligned}\|x - x^*\|_A^2 &= (x - x^*)^T A (x - x^*) \\ &= x^T A x - 2(x^*)^T A x + (x^*)^T A x^* \\ &= x^T A x - 2b^T x + b^T x^*\end{aligned}$$

Randomized Coordinate Descent

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

$$x^{t+1} = \arg\min ||x - x^*||_A^2 \quad \text{subject to} \quad x = x^t + \alpha e_i$$

$B = A$



$x^* = A^{-1}b$

Observation:

$$\begin{aligned}\|x - x^*\|_A^2 &= (x - x^*)^T A (x - x^*) \\ &= x^T A x - 2(x^*)^T A x + (x^*)^T A x^* \\ &= x^T A x - 2b^T x + b^T x^*\end{aligned}$$

$$\arg\min ||x - x^*||_A = \arg\min x^T A x - 2b^T x$$

Randomized Coordinate Descent

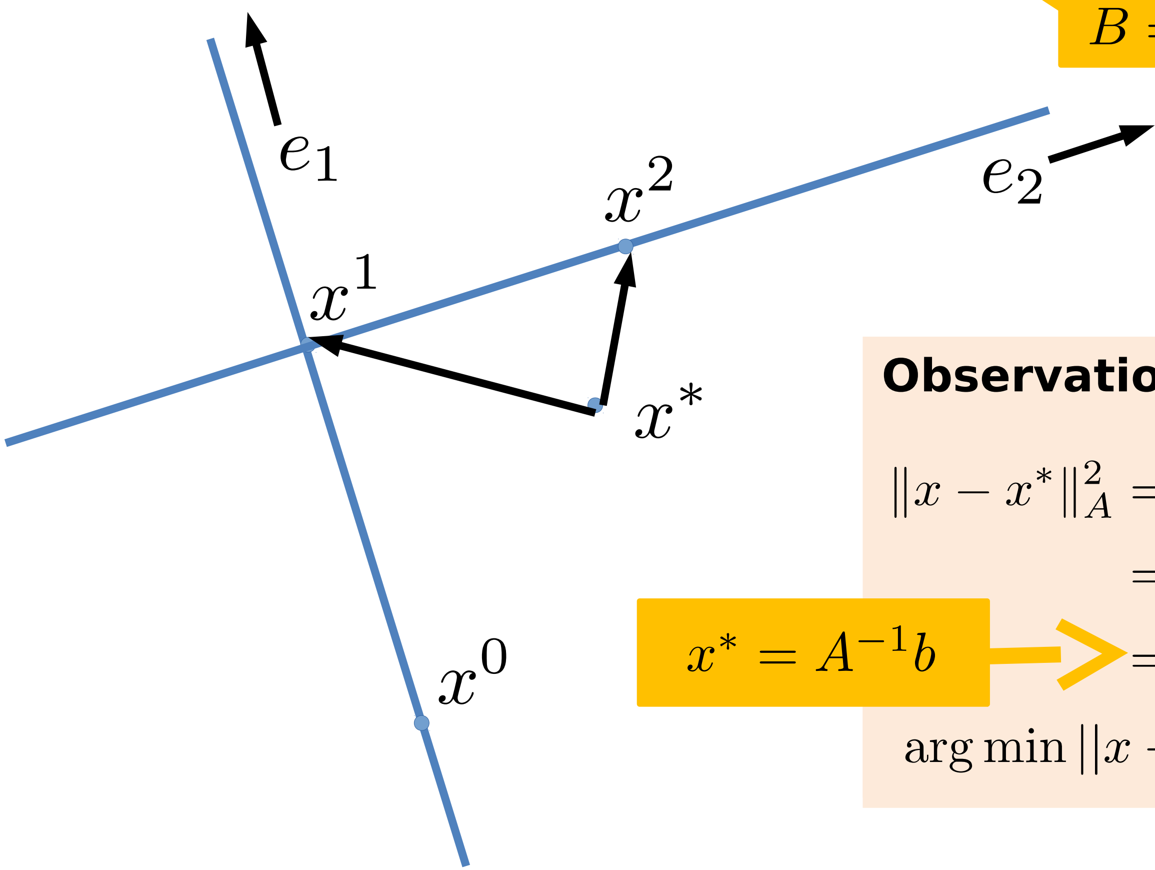
Leventhal, D., & Lewis, A. S. (2010). **Randomized Methods for Linear Constraints: Convergence Rates and Conditioning**. Mathematics of Operations Research, 35(3), 641–654.

$$x^{t+1} = \arg\min ||x - x^*||_A^2 \quad \text{subject to} \quad x = x^t + \alpha e_i$$

$B = A$

Block Coord. Descent

$$x = x^t + [e_{i_1} e_{i_2}] \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix}$$



Observation:

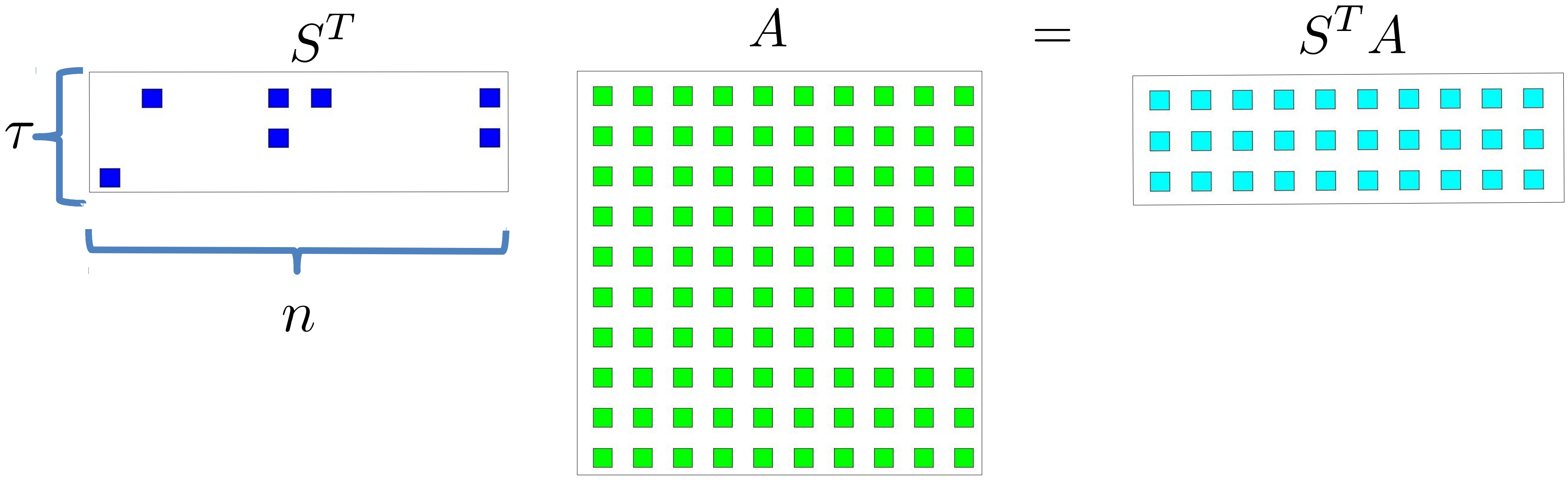
$$\begin{aligned} \|x - x^*\|_A^2 &= (x - x^*)^T A (x - x^*) \\ &= x^T A x - 2(x^*)^T A x + (x^*)^T A x^* \\ &= x^T A x - 2b^T x + b^T x^* \end{aligned}$$

$x^* = A^{-1} b$

$$\arg\min ||x - x^*||_A = \arg\min x^T A x - 2b^T x$$

Modern 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

1. Relaxation Viewpoint "Sketch and Project"

Sample $S \sim \mathcal{D}$

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

1. Relaxation Viewpoint "Sketch and Project"

Sample $S \sim \mathcal{D}$

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

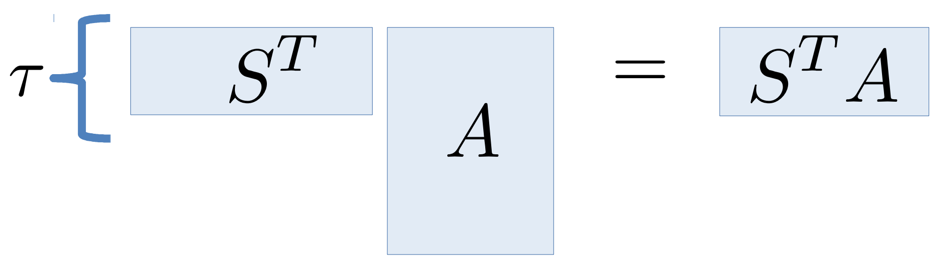
$$\text{subject to} \quad S^T A x = S^T b$$

1. Relaxation Viewpoint
"Sketch and Project"

Sample $S \sim \mathcal{D}$

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

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



2. Optimization Viewpoint "Constrain and Approximate"

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

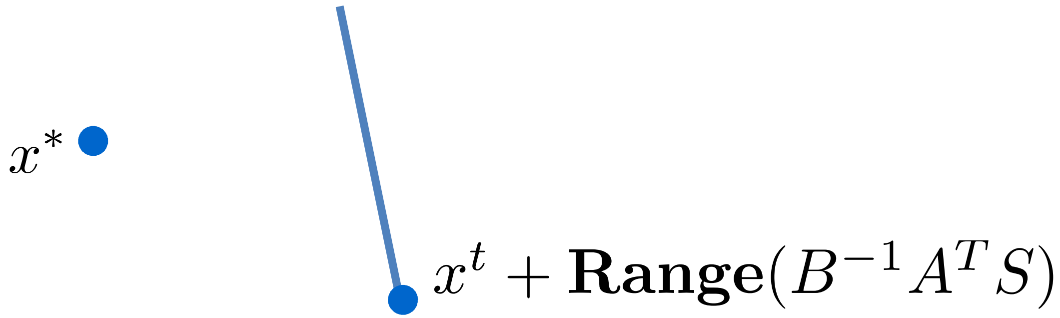
$$\begin{aligned} x^{t+1} \quad = \quad & \arg\min_{x \in \mathbb{R}^n} ||x - x^*||_B^2 \\ \text{subject to} \quad & x = x^t + B^{-1}A^T S y \\ & y \text{ is free} \end{aligned}$$

$x^t + \mathbf{Range}(B^{-1}A^T S)$

2. Optimization Viewpoint "Constrain and Approximate"

Sample $S \sim \mathcal{D}$

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



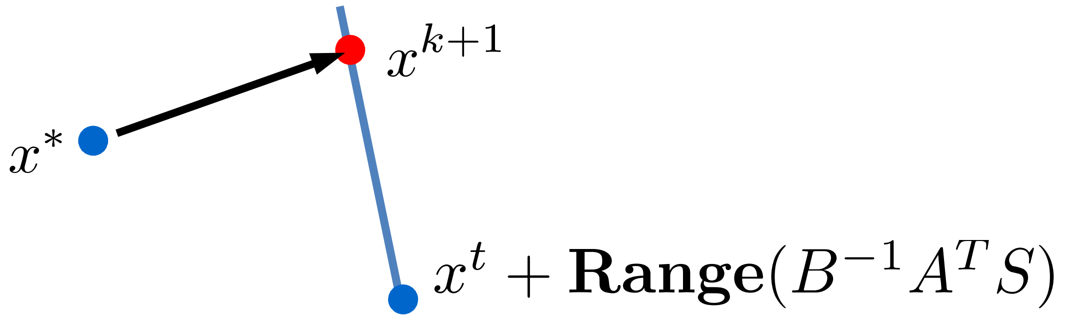
2. Optimization Viewpoint “Constrain and Approximate”

Sample $S \sim \mathcal{D}$

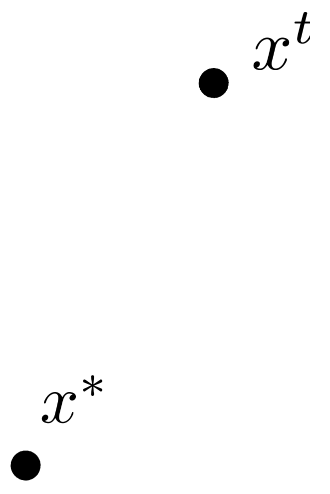
$$x^{t+1} \quad = \quad \arg\min_{x \in \mathbb{R}^n} ||x - x^*||_B^2$$

$$\text{subject to} \quad x = x^t + B^{-1} A^T S y$$

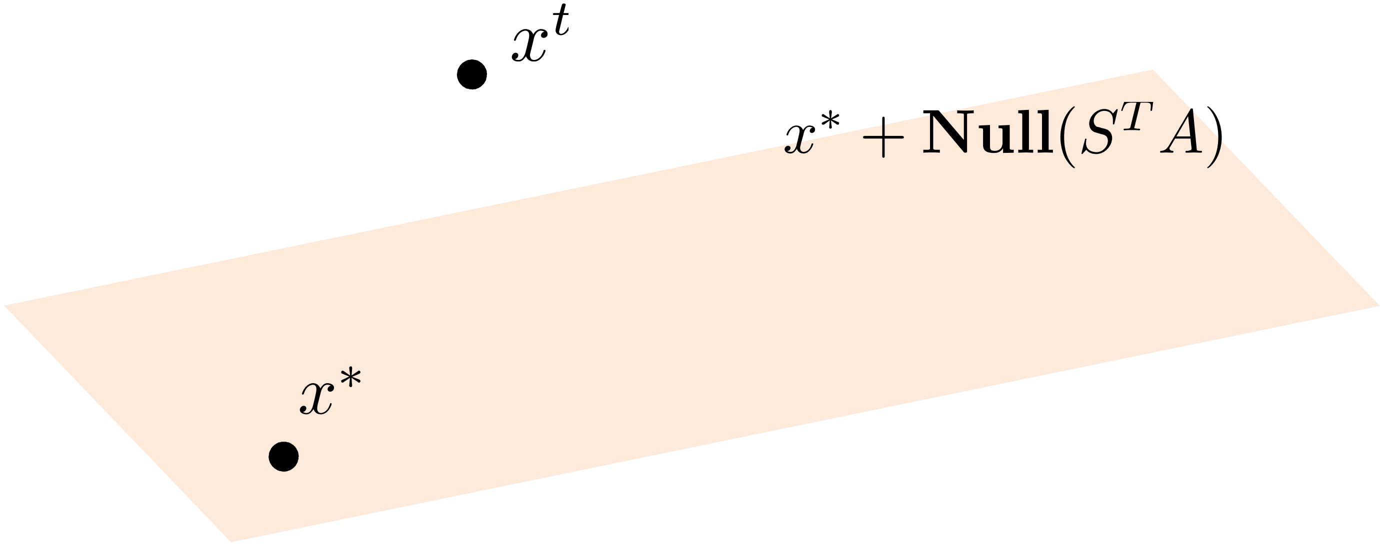
$$y \text{ is free}$$



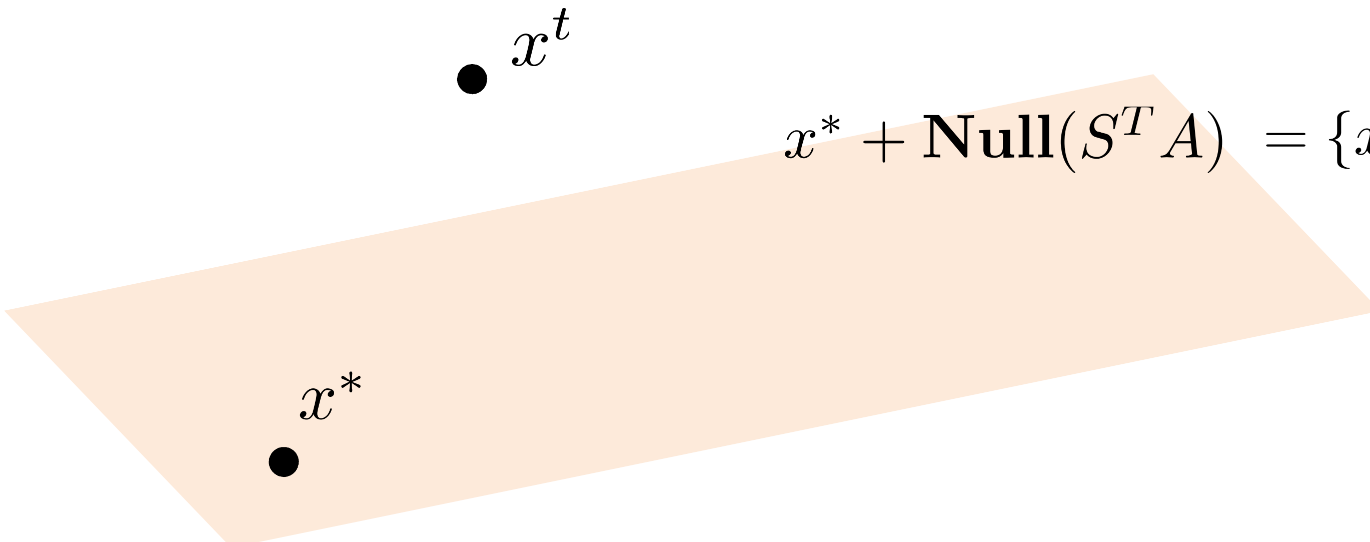
3. Geometric Viewpoint
“Random Intersect”



3. Geometric Viewpoint
“Random Intersect”

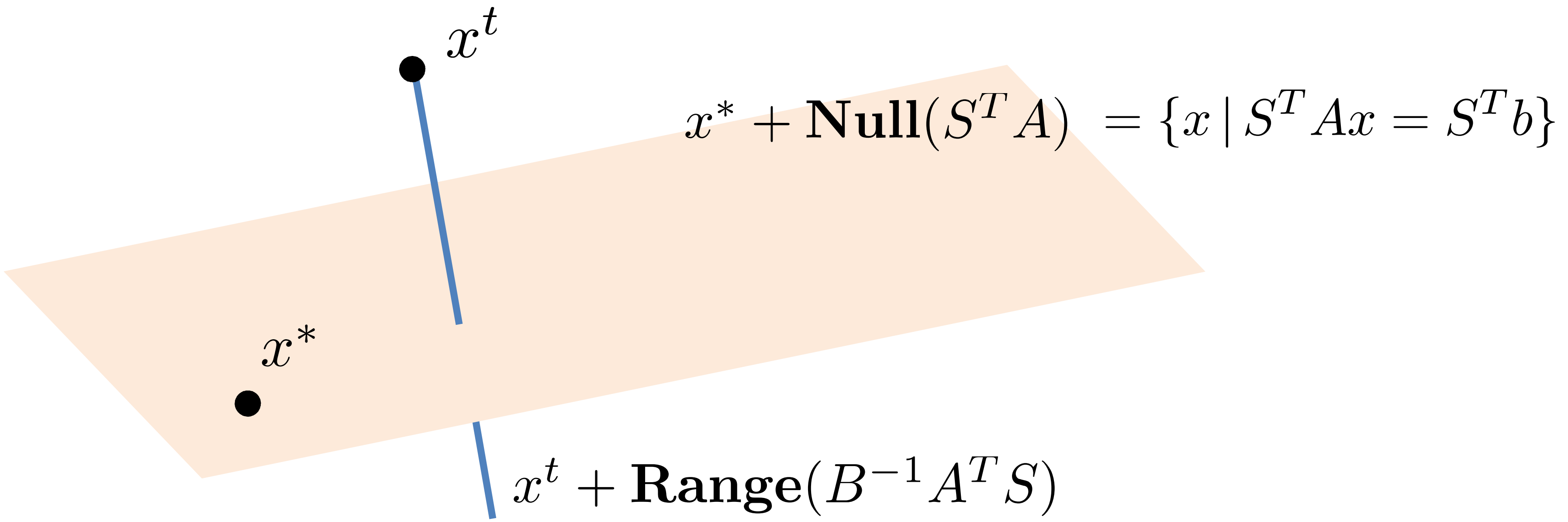


3. Geometric Viewpoint
“Random Intersect”

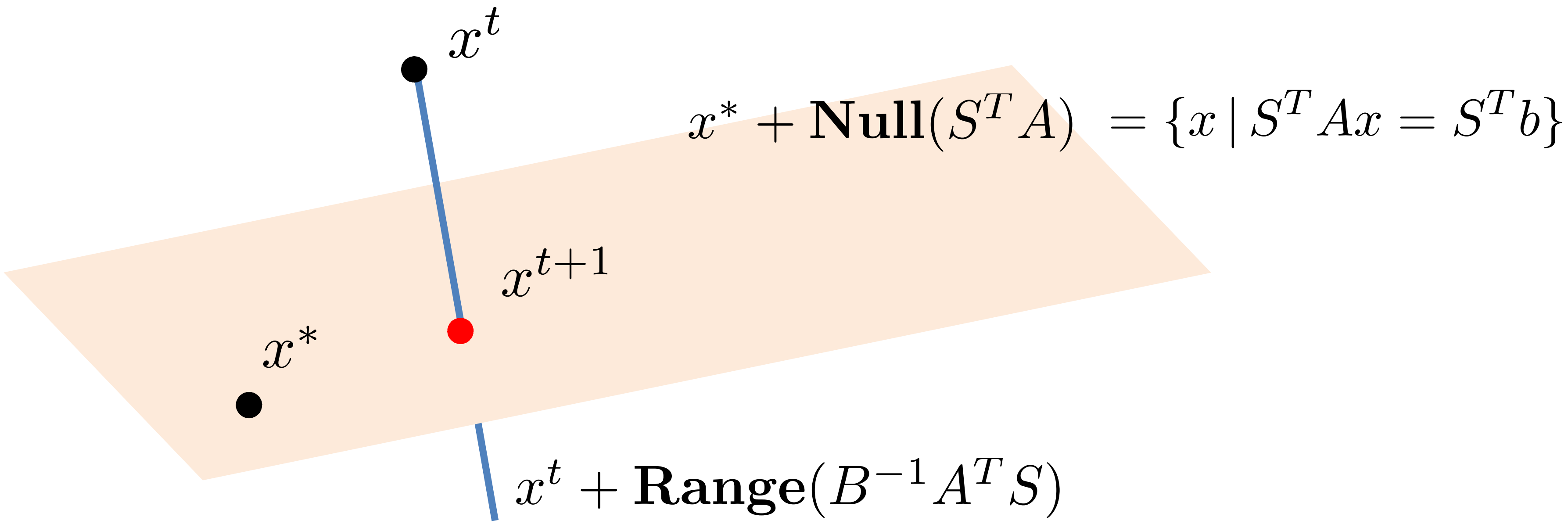


$$x^* + \mathbf{Null}(S^T A) \; = \{x \mid S^T A x = S^T b\}$$

3. Geometric Viewpoint
"Random Intersect"

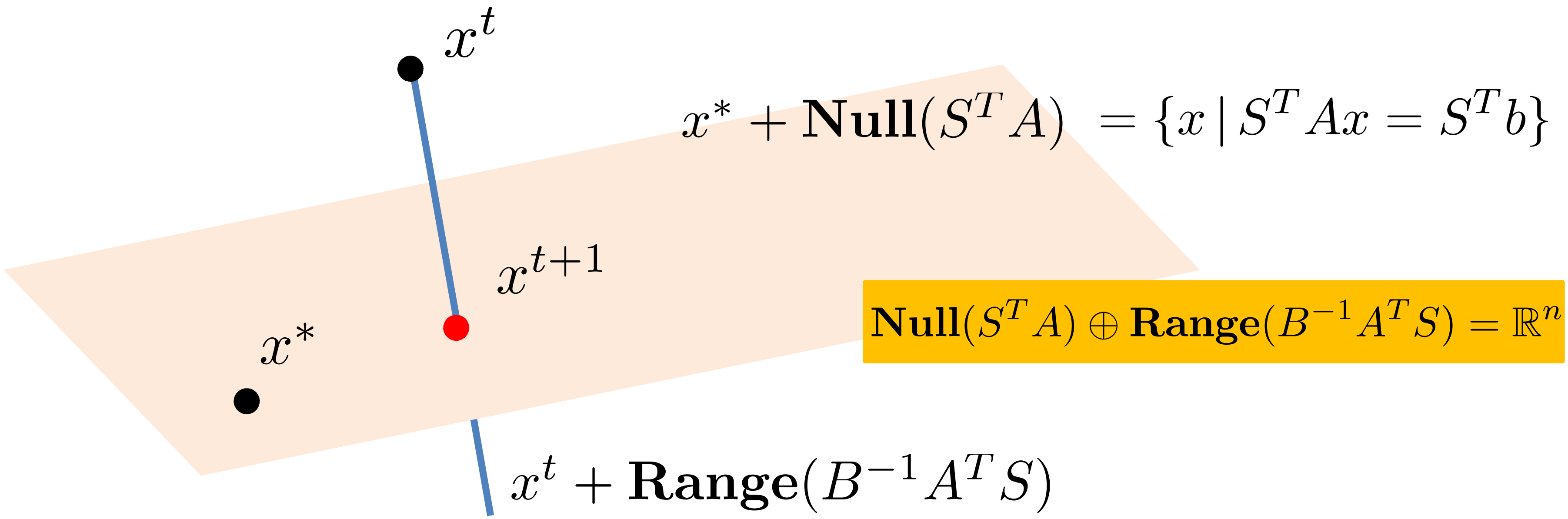


3. Geometric Viewpoint
"Random Intersect"



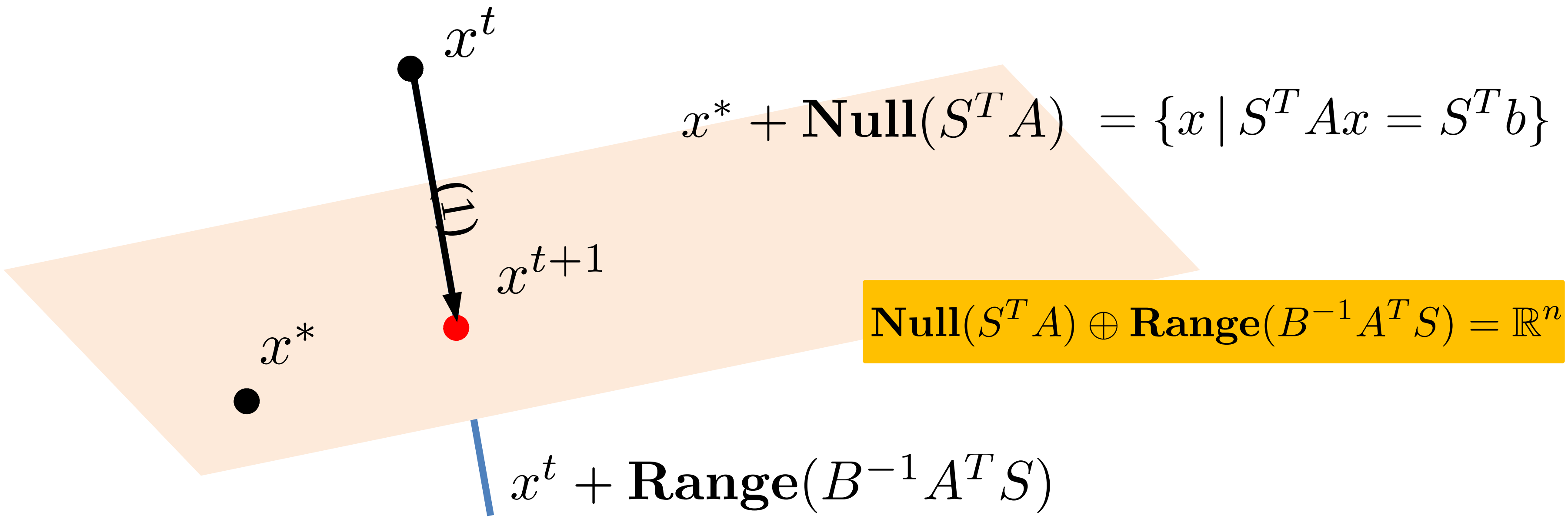
$$\{x^{t+1}\} = \left(x^* + \mathbf{Null}(S^T A)\right) \bigcap \left(x^t + \mathbf{Range}(B^{-1} A^T S)\right)$$

3. Geometric Viewpoint
“Random Intersect”



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

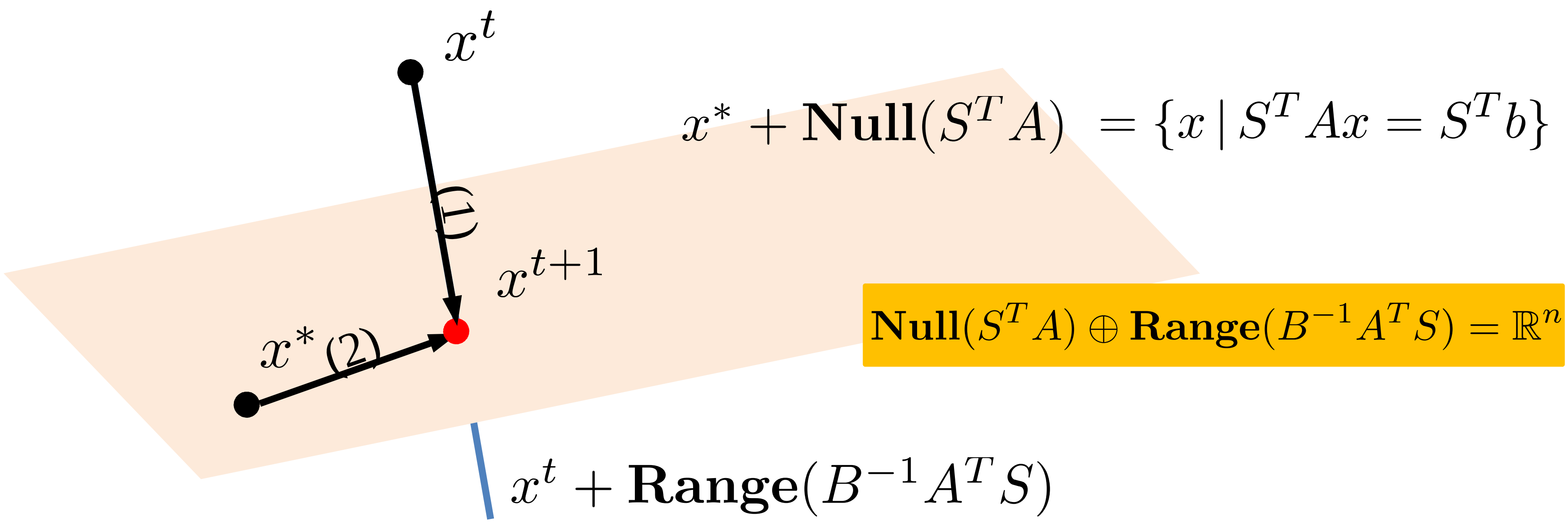
3. Geometric Viewpoint
"Random Intersect"



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

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

3. Geometric Viewpoint "Random Intersect"



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

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

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

4. Algebraic Viewpoint
"Random Update"

Random Update Vector

$$x^{t+1} = x^t - \underbrace{B^{-1}A^TS(S^TAB^{-1}A^TS)^{\dagger}S^T(Ax^t - b)}_{\text{Random Update Vector}}$$

4. Algebraic Viewpoint
“Random Update”

Random Update Vector

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

Moore-Penrose pseudo inverse

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

4. Algebraic Viewpoint “Random Update”

Random Update Vector

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

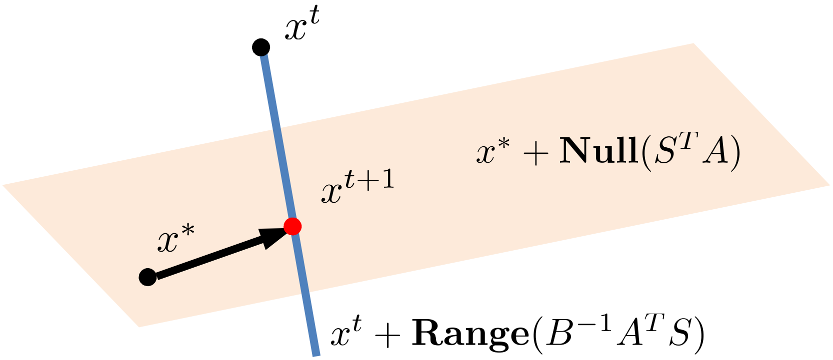
Small $\tau \times \tau$ matrix

Moore-Penrose pseudo inverse

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

5. Analytic Viewpoint "Random Fixed Point"

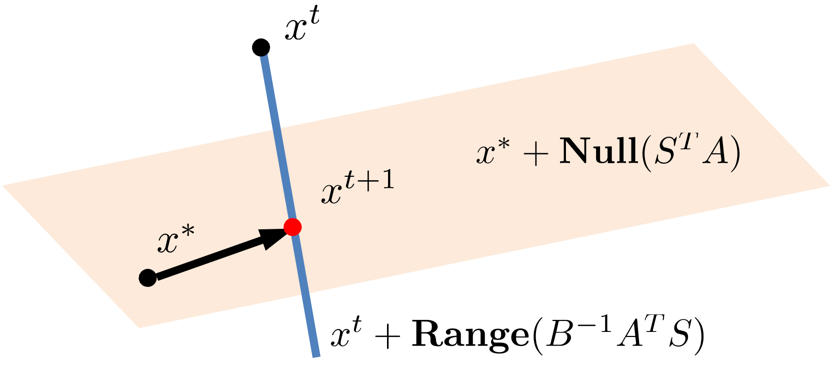
$$x^{t+1} - x^* = (I - B^{-1}A^T H A)(x^t - x^*)$$



5. Analytic Viewpoint
"Random Fixed Point"

$$H := S(S^TAB^{-1}A^TS)^\dagger S^T \in \mathbb{R}^{m\times m}$$

$$x^{t+1} - x^* = (I - B^{-1}A^THA)(x^t - x^*)$$

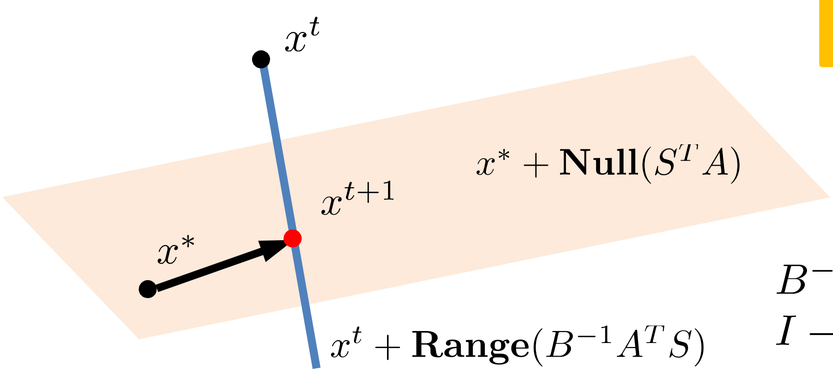


5. Analytic Viewpoint “Random Fixed Point”

$$H := S(S^T A B^{-1} A^T S)^{\dagger} S^T \in \mathbb{R}^{m \times m}$$

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

Random Iteration Matrix



$B^{-1} A^T H A$ 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]

If $x^0 \in \mathbf{Range}(A^T)$ and $\mathbf{E}[H] \succ 0$ then

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

where

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

Complexity / Convergence

Theorem [GR'15]

If $x^0 \in \mathbf{Range}(A^T)$ and $\mathbf{E}[H] \succ 0$ then

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

where

Smallest nonzero eigenvalue

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

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

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

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

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

Special Choice of Parameters

$$\mathbf{P}(S = e_i) = \frac{1}{m} \Rightarrow S = e^i$$

$$B = I$$

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

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

Special Choice of Parameters

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

$$B = I$$

$$S = e^i$$

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

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

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

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

Special Choice of Parameters

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

$$B = I$$

$$S = e^i$$

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

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

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

Special Choice of Parameters

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

$$B = I$$

$$S = e^i$$

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

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

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

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

Special Choice of Parameters

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

$$B = I$$

$$S = e^i$$

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

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

The rate: lower and upper bounds

Theorem [RG‘15]

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

⇓

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

The rate: lower and upper bounds

Theorem [RG'15]

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

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

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

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.

The rate: lower and upper bounds

Theorem [RG'15]

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

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

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

Insight: lower rank of A and great rank of $S^T A$ gives better lower bound. In other words, when the dimension of the search space in the "constrain and approximate" viewpoint grows.

Special Case: Randomized Kaczmarz Method

T. Strohmer and R. J. Vershynin, (2009). **A Randomized Kaczmarz Algorithm with Exponential Convergence** Journal of Fourier Analysis and Applications, 15:262

Randomized Kaczmarz: derivation and rate

General Method

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

Randomized Kaczmarz: derivation and rate

General Method

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

Special Choice of Parameters

$$B = I$$

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

Randomized Kaczmarz: derivation and rate

General Method

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

Special Choice of Parameters

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

$$B = I$$

$$S = e_i$$

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

Randomized Kaczmarz: derivation and rate

General Method

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

Special Choice of Parameters

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

$$B = I$$
$$S = e_i$$

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

Randomized Kaczmarz: derivation and rate

General Method

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

Special Choice of Parameters

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

$$B = I$$

$$S = e_i$$

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

Randomized Kaczmarz: derivation and rate

General Method

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

Special Choice of Parameters

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

$$B = I$$

$$S = e_i$$

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

Randomized Kaczmarz: derivation and rate

General Method

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

Special Choice of Parameters

$$B = I$$

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

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

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

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

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

Special Case: Randomized Coordinate Descent

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

Randomized Coordinate Descent: derivation and rate

General Method

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

$$B = A$$

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

Randomized Coordinate Descent: derivation and rate

General Method

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

Special Choice of Parameters

positive definite → $B = A$

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

Randomized Coordinate Descent: derivation and rate

General Method

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

Special Choice of Parameters

positive definite → $B = A$

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

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

Randomized Coordinate Descent: derivation and rate

General Method

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

Special Choice of Parameters

positive definite $\rightarrow$ $B = A$

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

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

Randomized Coordinate Descent: derivation and rate

General Method

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

Special Choice of Parameters

positive definite → $B = A$

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

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

Randomized Coordinate Descent: derivation and rate

General Method

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

Special Choice of Parameters

positive definite → $B = A$

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

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

Randomized Coordinate Descent: derivation and rate

General Method

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

Special Choice of Parameters

positive definite $\Rightarrow B = A$

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

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

Complexity Rate $A \succ 0 \Rightarrow \mathbf{E}[H] = \mathrm{diag}(A_{11}, \ldots, A_{nn}) \succ 0$

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

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

Theory recovers known and new convergence results

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

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

Designing New Methods

Optimal methods

Optimal choice

$$\max_{B,\mathcal{D}} \rho = \lambda_{\min}^{+}(B^{-1/2} A^T \mathbf{E}_{S\sim\mathcal{D}}[H] A B^{-1/2})$$

Optimal methods

Optimal choice

$$\max_{B,\mathcal{D}} \rho = \lambda_{\min}^{+}(B^{-1/2}A^T\mathbf{E}_{S\sim\mathcal{D}}[H]AB^{-1/2})$$

	$A \succ 0$	$\mathbf{Rank}(A) = n$	any A
B	A	A^TA	I

Optimal methods

Optimal choice

$$\max_{B,\mathcal{D}} \rho = \lambda_{\min}^{+}(B^{-1/2}A^T \mathbf{E}_{S\sim\mathcal{D}}[H] A B^{-1/2})$$

	$A \succ 0$	$\mathbf{Rank}(A) = n$	any A
B	A	$A^T A$	I

Optimal S

$$\mathbf{Range}(S) = \mathbf{Range}(A^{-T}B^{1/2})$$

Optimal methods

Optimal choice

$$\max_{B,\mathcal{D}} \rho = \lambda_{\min}^{+}(B^{-1/2} A^T \mathbf{E}_{S\sim\mathcal{D}}[H] A B^{-1/2})$$

	$A \succ 0$	$\mathbf{Rank}(A) = n$	any A
B	A	$A^T A$	I

Optimal S

$$\mathbf{Range}(S) = \mathbf{Range}(A^{-T} B^{1/2})$$

S with fixed range

$$\mathbf{Prob}[S = S_i] = p_i,$$

for $i = 1, \ldots, r$

Optimal methods

Optimal choice

$$\max_{B,\mathcal{D}} \rho = \lambda_{\min}^{+}(B^{-1/2}A^T\mathbf{E}_{S\sim\mathcal{D}}[H]AB^{-1/2})$$

	$A \succ 0$	$\mathbf{Rank}(A) = n$	any A
B	A	A^TA	I

Optimal S

$$\mathbf{Range}(S) = \mathbf{Range}(A^{-T}B^{1/2})$$

S with fixed range

$$\mathbf{Prob}[S = S_i] = p_i,$$

for $i = 1, \ldots, r$

Optimal p_i's

Optimal methods

Optimal choice

$$\max_{B,\mathcal{D}} \rho = \lambda^{+}_{\min}(B^{-1/2}A^T\mathbf{E}_{S\sim\mathcal{D}}[H]AB^{-1/2})$$

	$A \succ 0$	$\mathbf{Rank}(A) = n$	any A
B	A	A^TA	I

Optimal S

$$\mathbf{Range}(S) = \mathbf{Range}(A^{-T}B^{1/2})$$

S with fixed range

$$\mathbf{Prob}[S = S_i] = p_i,$$

for $i = 1, \ldots, r$

Optimal p_i's

$$\max_{t,p\in\triangle_r} \quad t$$

$$\text{sub. to} \quad \sum_{i=1}^{r} p_i V_i(V_i^T V_i)^{-1}V_i^T \succeq t \cdot I$$

$$V_i = B^{1/2}A^T S_i, \quad i = 1, \ldots, r$$

Difficult SDP

where $||x^* - x_s^*||_2 \leq (1+\epsilon)||x^*||$

Practical New Methods

One Shot Sketches

$$x_s^* = \arg_x \min ||S^T Ax - S^T b||_2$$

$$\text{where} \quad ||x^* - x_s^*||_2 \leq (1+\epsilon)||x^*||$$

N. Ailon and B. Chazelle (2006). **Approximate nearest neighbors and the fast Johnson-Lindenstrauss transform.** Mathematics of Operations Research, 35(3), 641–654.

where $||x^* - x_s^*||_2 \leq (1+\epsilon)||x^*||$

Practical New Methods

One Shot Sketches

$$x_s^* = \arg_x \min ||S^T A x - S^T b||_2$$

$$\text{where} \quad ||x^* - x_s^*||_2 \leq (1+\epsilon)||x^*||$$

S	Computing $S^T A$
Gaussian Matrix	$O(mn\tau)$
Subsampled Hadamard-Welsh	$O(mn\log(\tau))$
Countmin Sketch	$O(nnz(A))$

N. Ailon and B. Chazelle (2006). **Approximate nearest neighbors and the fast Johnson-Lindenstrauss transform.** Mathematics of Operations Research, 35(3), 641–654.

where $||x^* - x_s^*||_2 \leq (1+\epsilon)||x^*||$

Practical New Methods

One Shot Sketches

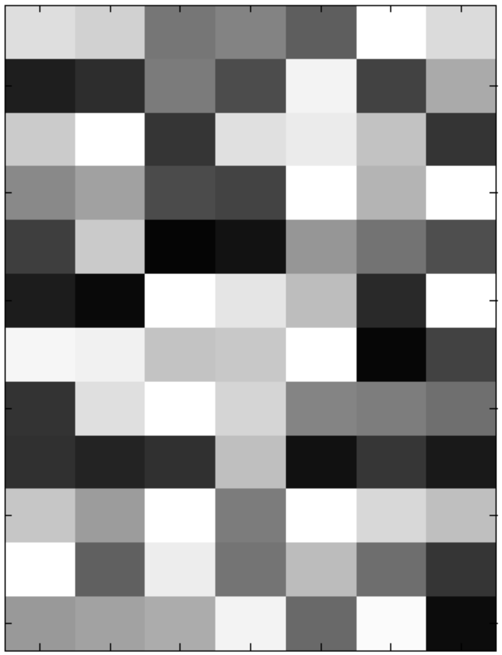
$$x_s^* = \arg_x \min ||S^T A x - S^T b||_2$$

where $$||x^* - x_s^*||_2 \leq (1+\epsilon)||x^*||$$

S	Computing $S^T A$
Gaussian Matrix	$O(mn\tau)$
Subsampled Hadamard-Welsh	$O(mn\log(\tau))$
Countmin Sketch	$O(nnz(A))$
Rademacher Sketch	$O(nnz(A))$

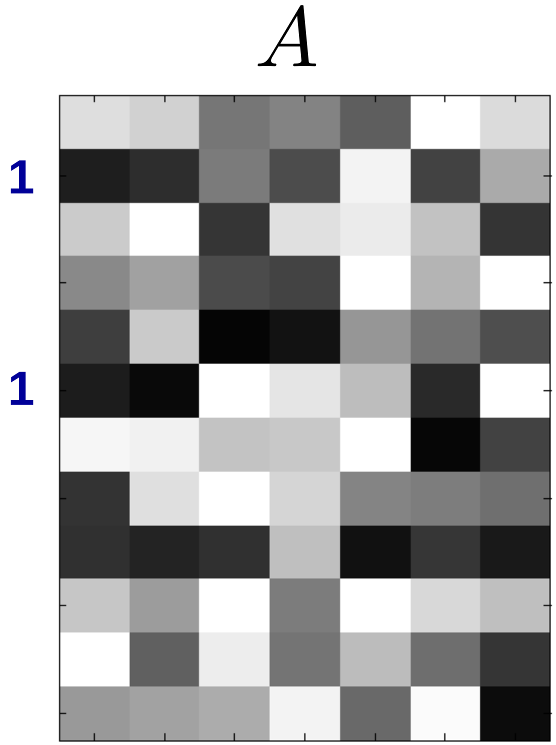
N. Ailon and B. Chazelle (2006). **Approximate nearest neighbors and the fast Johnson-Lindenstrauss transform.** Mathematics of Operations Research, 35(3), 641–654.

Sub-Rademacher Sketching



sketch size τ	$= 3$
density	$= 2$

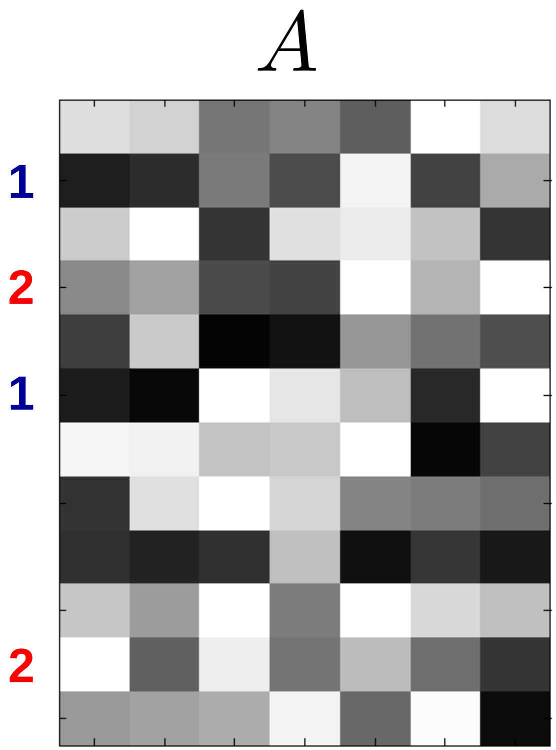
Sub-Rademacher Sketching



sketch size $\tau = 3$

density $= 2$

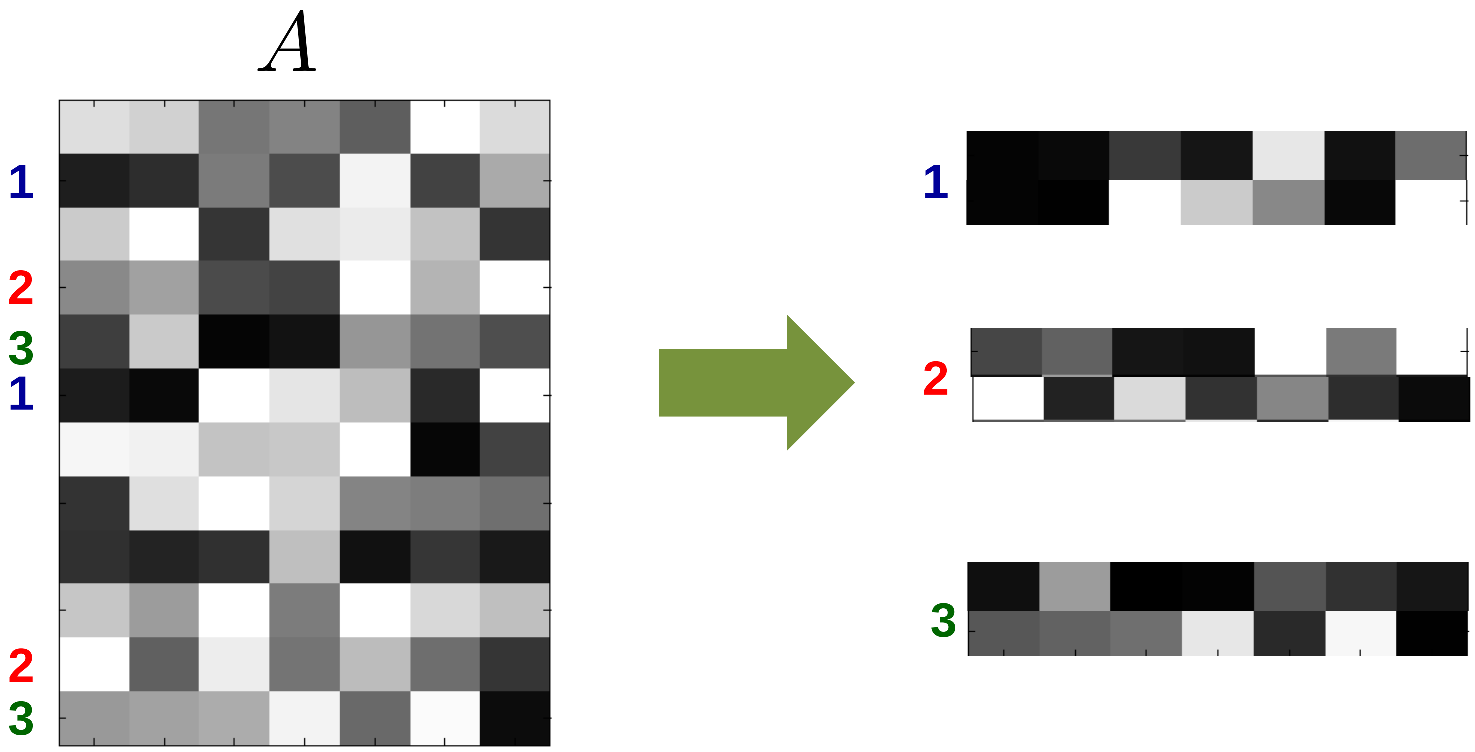
Sub-Rademacher Sketching



sketch size $\tau = 3$

density $= 2$

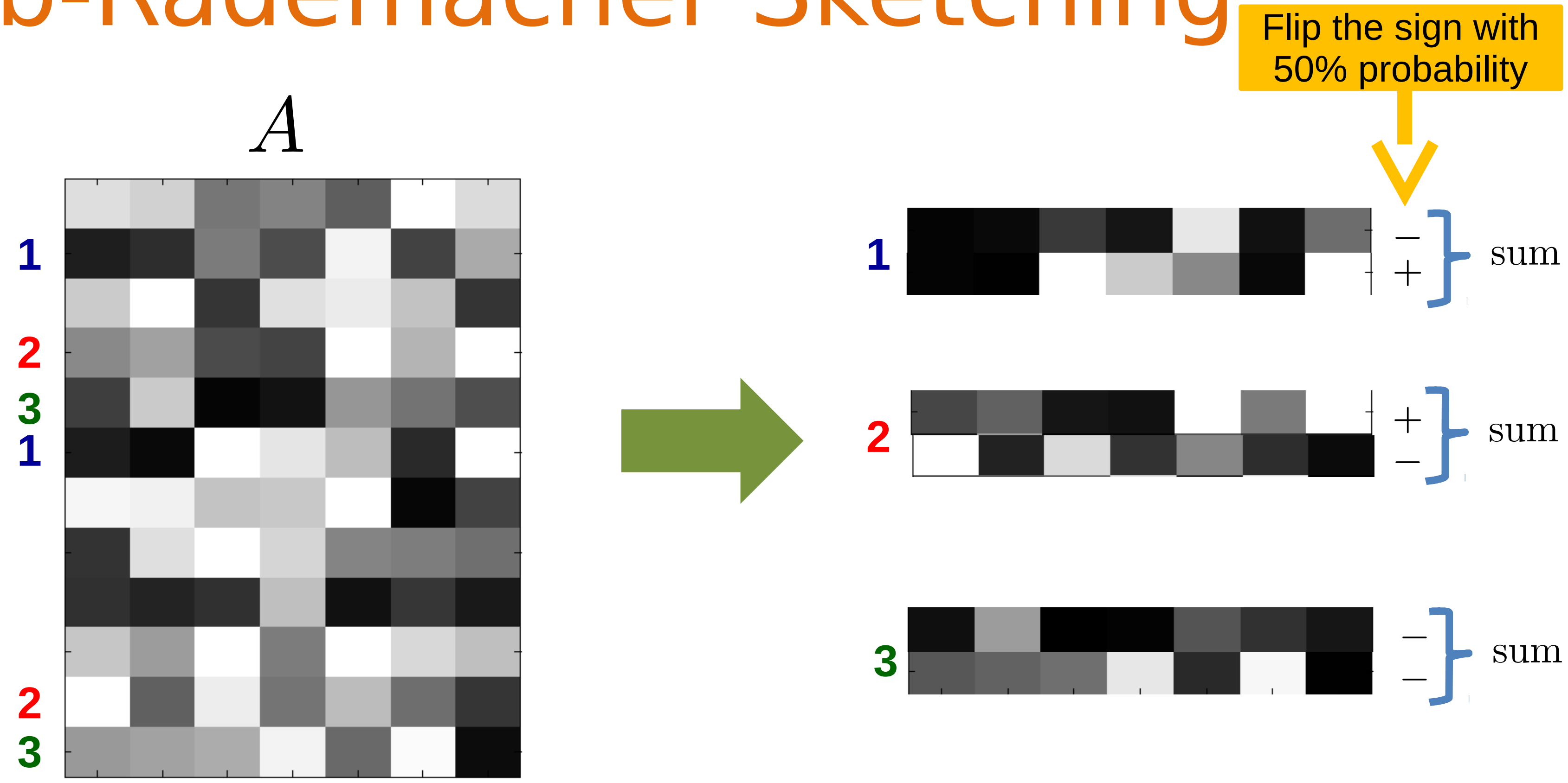
Sub-Rademacher Sketching



$$\text{sketch size } \tau = 3$$

$$\text{density } = 2$$

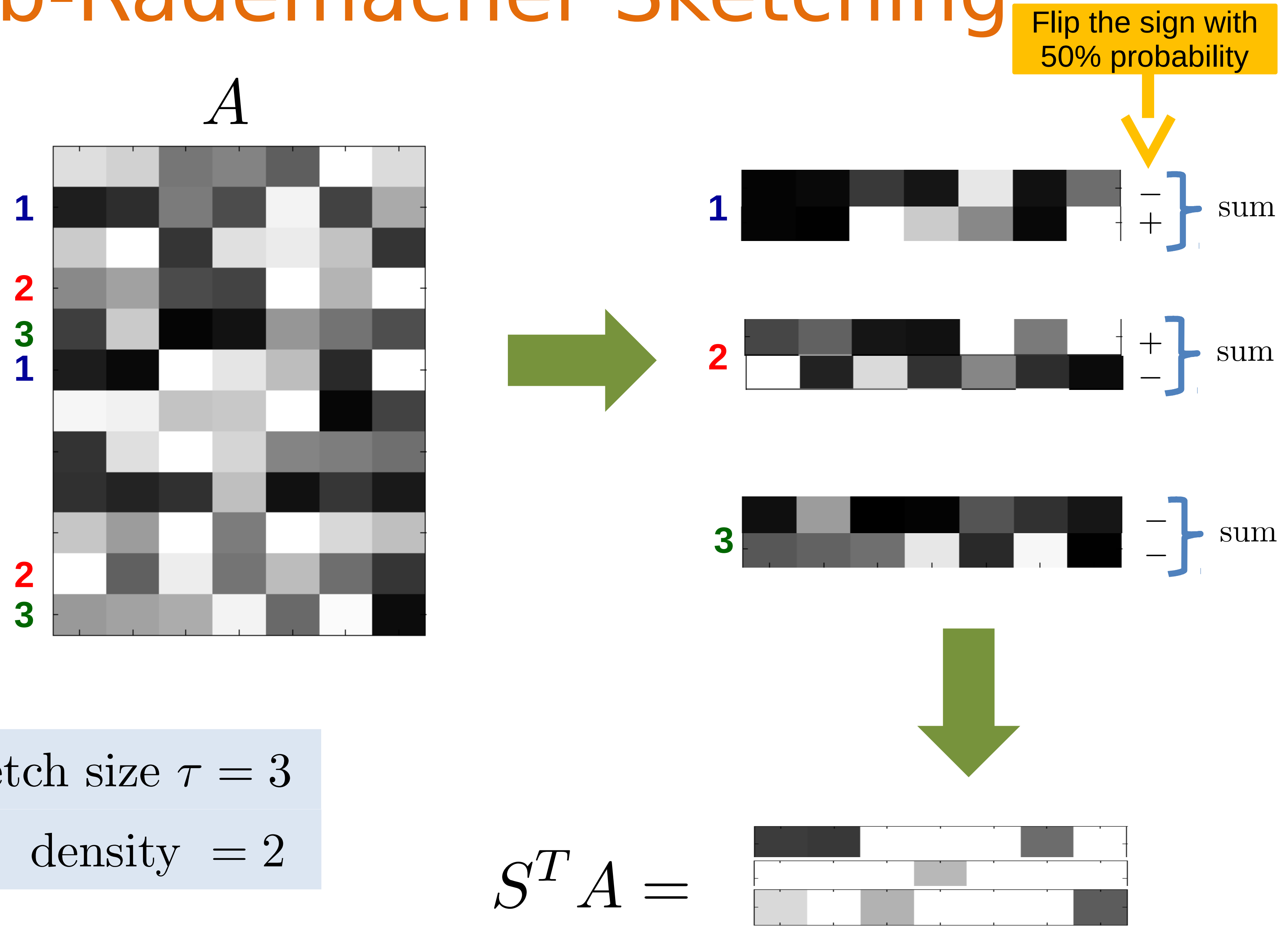
Sub-Rademacher Sketching



sketch size $\tau = 3$

density $= 2$

Sub-Rademacher Sketching



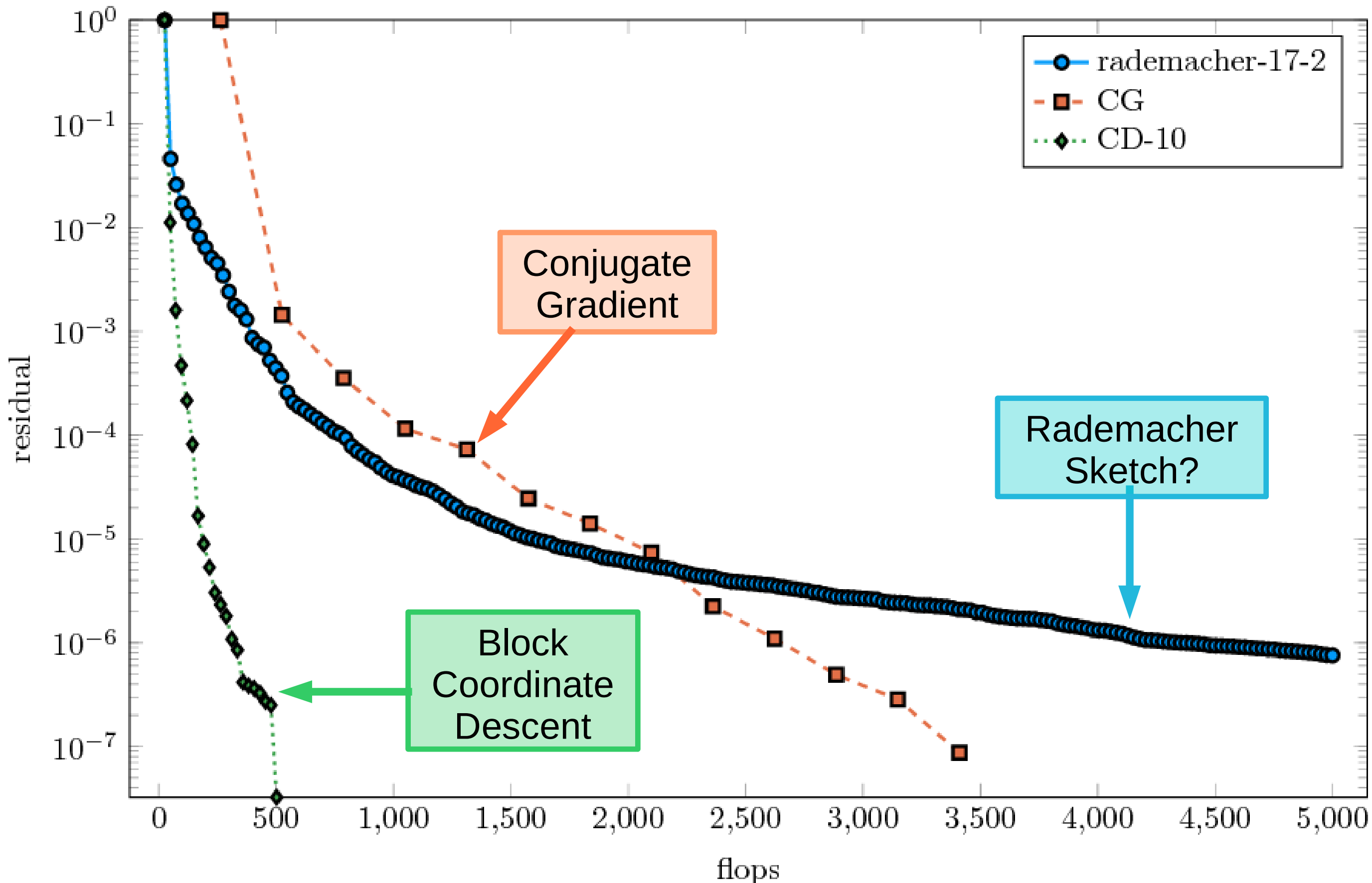
sketch size $\tau = 3$

density $= 2$

Experiments

Large scale Ridge Regression

$$\min_x \frac{1}{2}||Ax - y||_2^2 + \frac{\lambda}{2}||x||_2^2$$



Problem: w8a

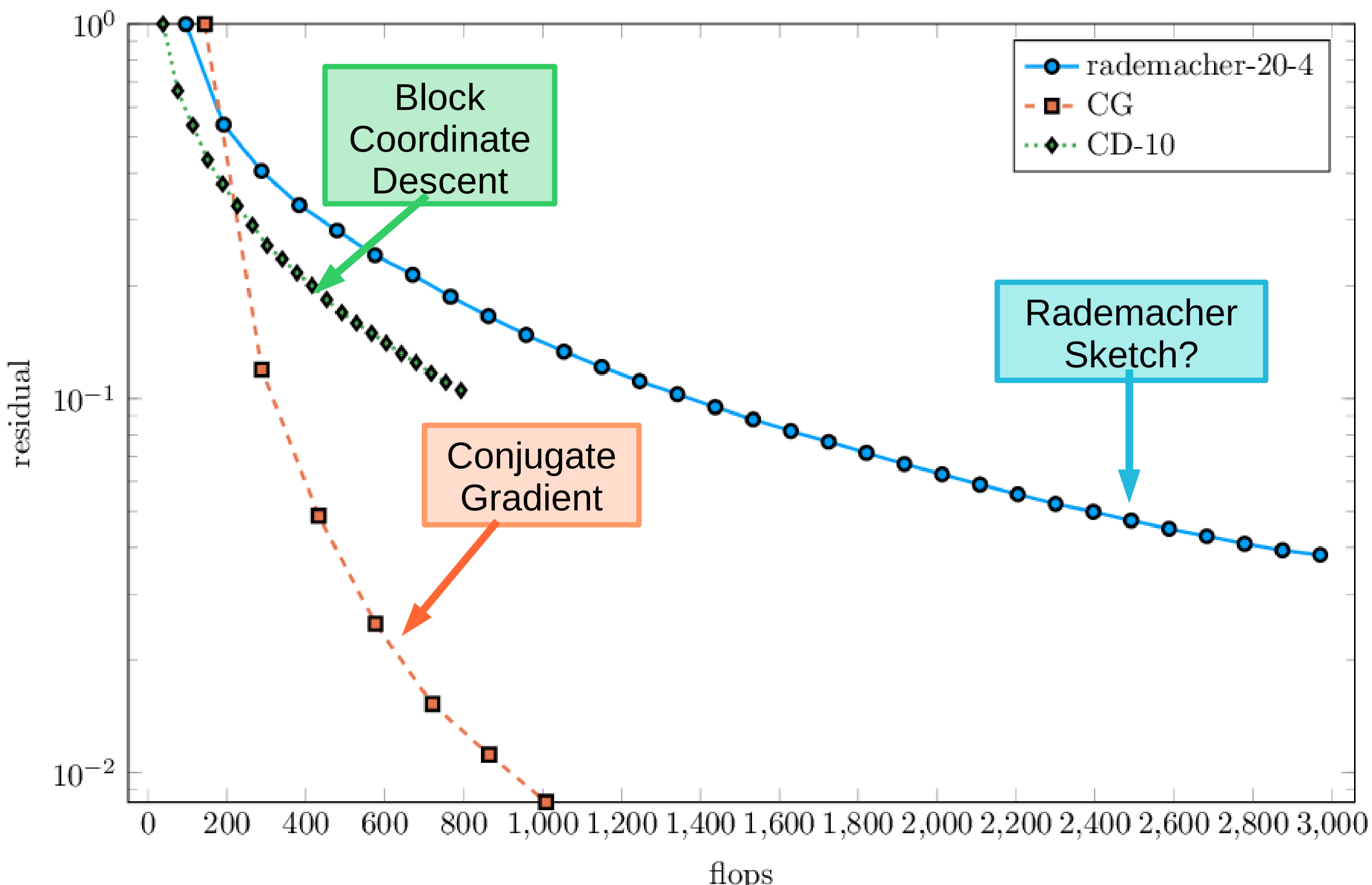
$A \in \mathbb{R}^{49\,749 \times 300}$

Origin: LIBSVM

GitHub: BigRidge

Large scale Ridge Regression

$$\min_x \frac{1}{2}||Ax - y||_2^2 + \frac{\lambda}{2}||x||_2^2$$



Problem: rcv1

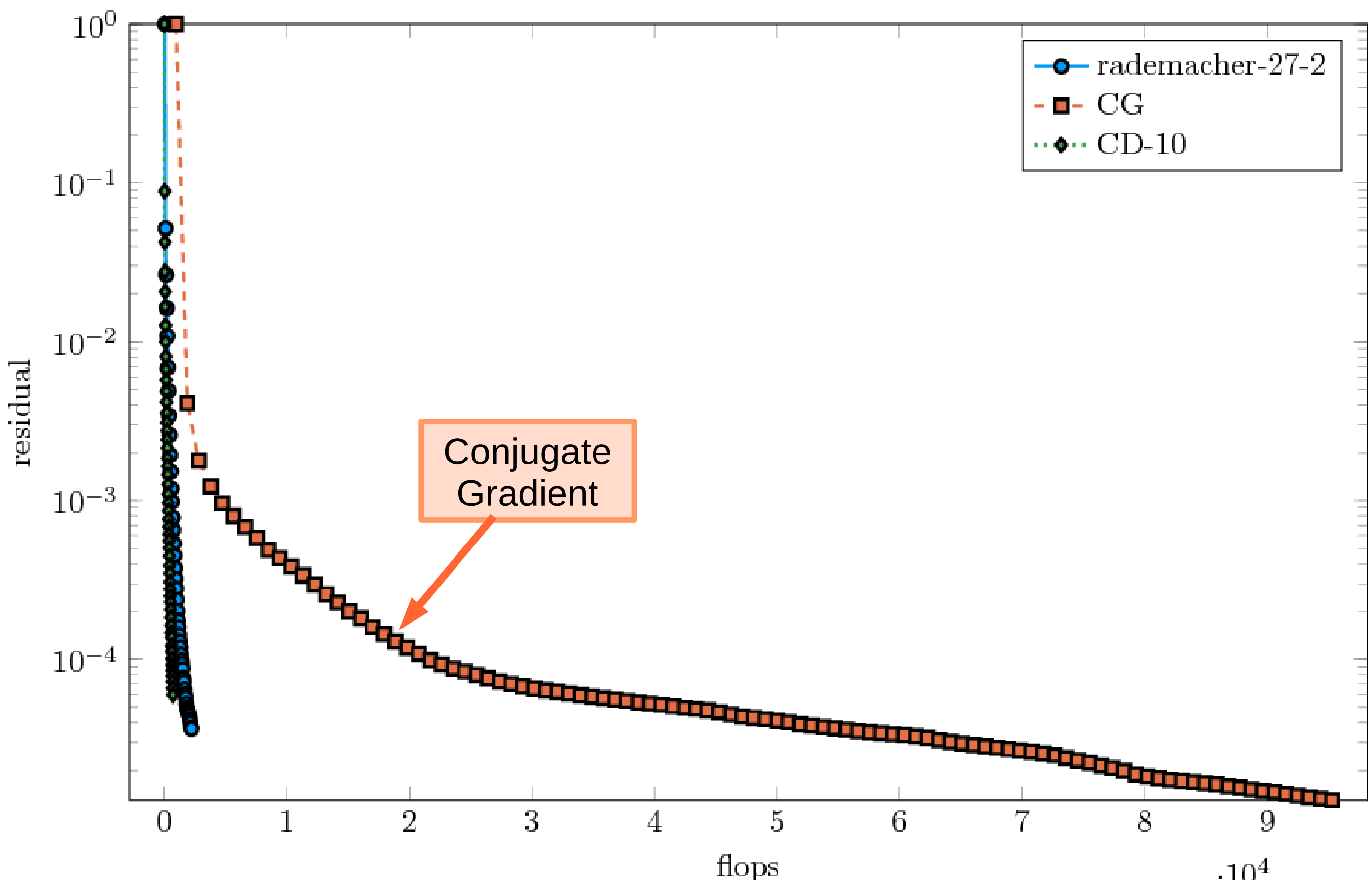
$A \in \mathbb{R}^{20\,242 \times 47\,236}$

Origin: LIBSVM

GitHub: BigRidge

Large scale Ridge Regression

$$\min_x \frac{1}{2}||Ax - y||_2^2 + \frac{\lambda}{2}||x||_2^2$$



Problem: mnist

$A \in \mathbb{R}^{60\,000 \times 780}$

Origin: LIBSVM

GitHub: BigRidge

Conclusions

Unites many randomized methods under a single framework

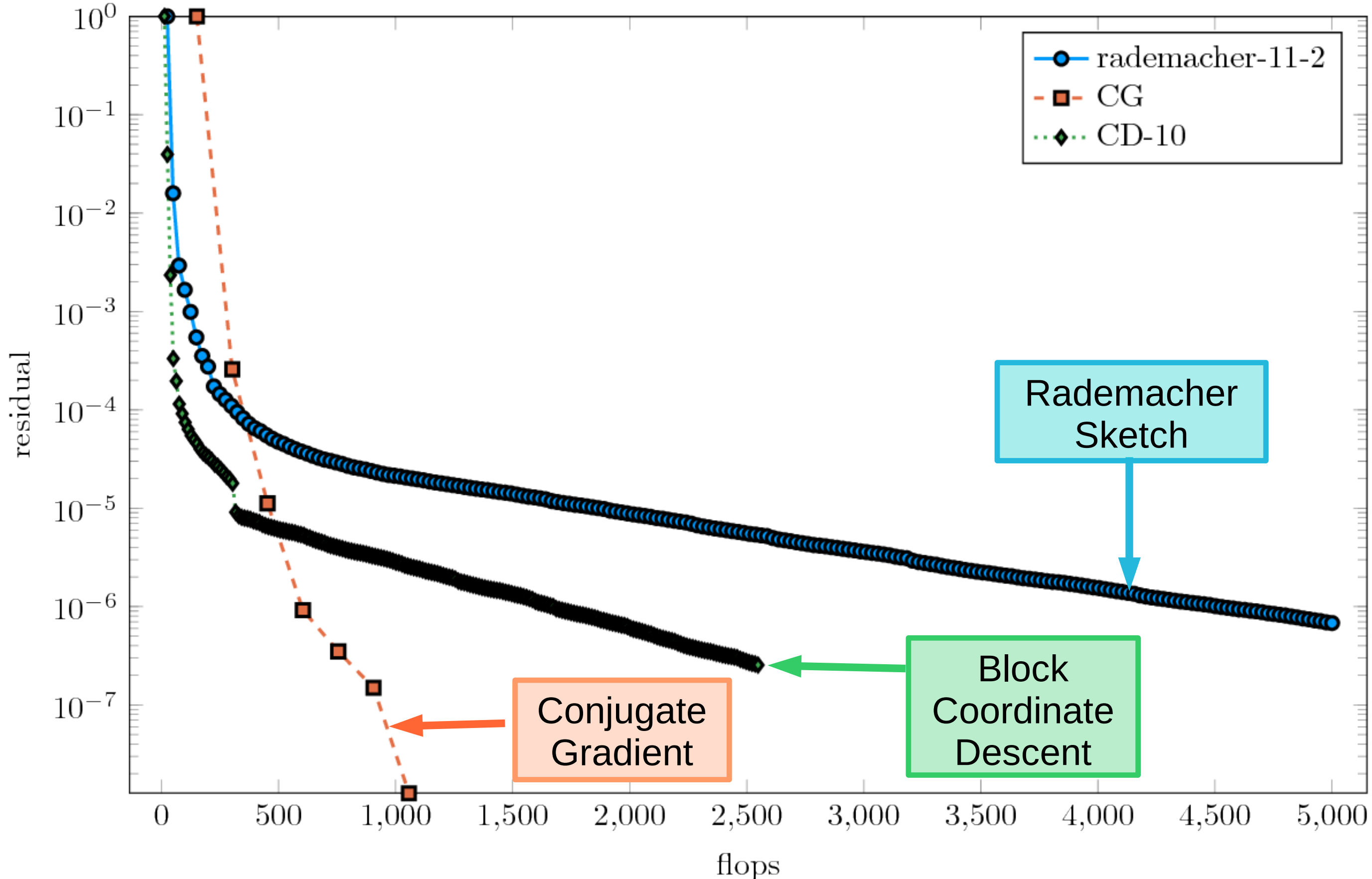
Improved convergence New lower bounds, less assumptions, tightest results.

Design new methods $S =$ Guassian, count-sketch, Walsh-Hadamard ...etc

Optimal Sampling We can choose a sampling that optimizes the convergence rate.

Large scale Ridge Regression

$$\min_x \frac{1}{2}||Ax - y||_2^2 + \frac{\lambda}{2}||x||_2^2$$



Problem: a9a

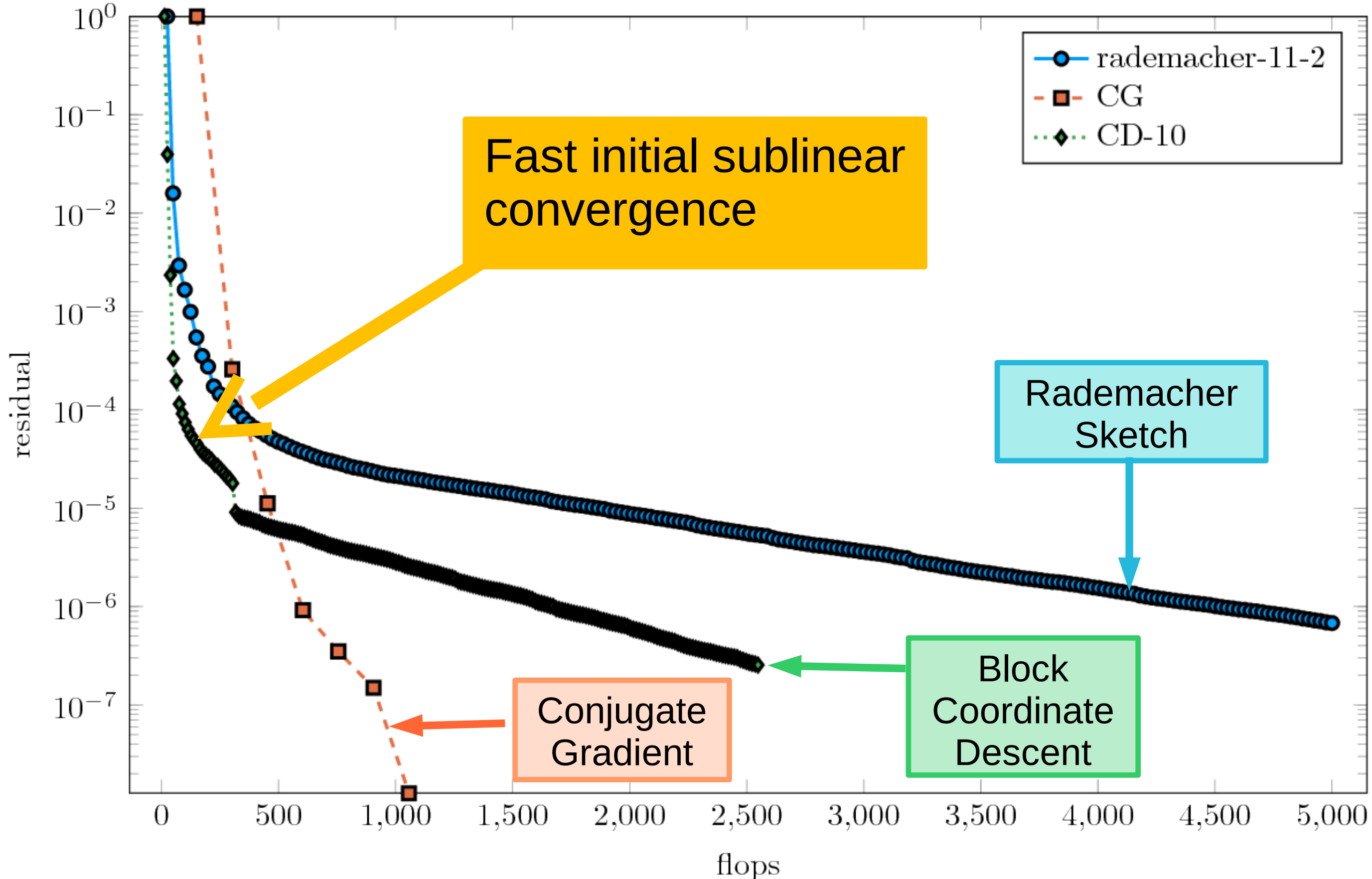
$A \in \mathbb{R}^{32,561 \times 123}$

Origin: LIBSVM

GitHub: BigRidge

Large scale Ridge Regression

$$\min_{x} \frac{1}{2}||Ax - y||_2^2 + \frac{\lambda}{2}||x||_2^2$$



Problem: a9a

$A \in \mathbb{R}^{32,561 \times 123}$

Origin: LIBSVM

GitHub: BigRidge

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

Thank you,
Questions?