



# Modern Optimization Methods for Big Data Problems

## MATH11146

### Lecture 1

# Randomized Iterative Methods for Linear Systems



Peter Richtárik



**Robert M Gower**  
(Edinburgh)



Robert M Gower and P.R.  
**Randomized Iterative Methods for Linear Systems**  
*SIAM Journal on Matrix Analysis and Applications* 36(4):  
1660-1690, 2015



Robert M Gower and P.R.  
**Stochastic Dual Ascent for Solving Linear Systems**  
*arXiv:1512.06890*, 2015

# The Problem

# The Problem

$$m \left[ \begin{array}{c} n \\ \text{---} \\ A\mathbf{x} = \mathbf{b} \end{array} \right] m$$

A blue brace above the matrix  $A$  indicates its width is  $n$ . A blue brace below the equation indicates its height is  $m$ . A yellow box containing the text  $\in \mathbb{R}^n$  has a yellow arrow pointing to the variable  $\mathbf{x}$ .

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

We can also think of this as  $m$  linear equations, where the  $i^{\text{th}}$  equation looks as follows:

$$\sum_{j=1}^n A_{ij}x_j = b_i$$
$$A_{i:\mathbf{x}} = b_i$$

# Minimizing Convex Quadratics

$$\min_{x \in \mathbb{R}^n} \left[ f(x) = \frac{1}{2} \|Ax - b\|^2 \right] \Rightarrow \nabla f(x) = 0 \Rightarrow A^T Ax = A^T b$$



This system is consistent

$$\min_{x \in \mathbb{R}^n} \left[ f(x) = \frac{1}{2} x^T Ax + b^T x + c \right] \Rightarrow \nabla f(x) = 0 \Rightarrow Ax = b$$



$A = \text{positive definite}$



This system is consistent

# The Solution (6 Ways to Skin a Cat)

TOP DEFINITION

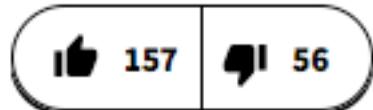


## skin the cat

Term refers to a task which has several ways by which it can be completed. Often used in the expression "there are many ways to skin the cat" or by using "skin this cat" in place of "skin the cat."

*My friends and I are going to start a business, but we don't even know where to begin because there are so many ways to skin the cat.*

by CRubio April 15, 2007



# 1. Relaxation Viewpoint

## “Sketch and Project”

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

$B$ : Symmetric and positive definite

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

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

**One Step Method:**  $S = m \times m$  invertible (with probability 1)

## 2. Optimization Viewpoint

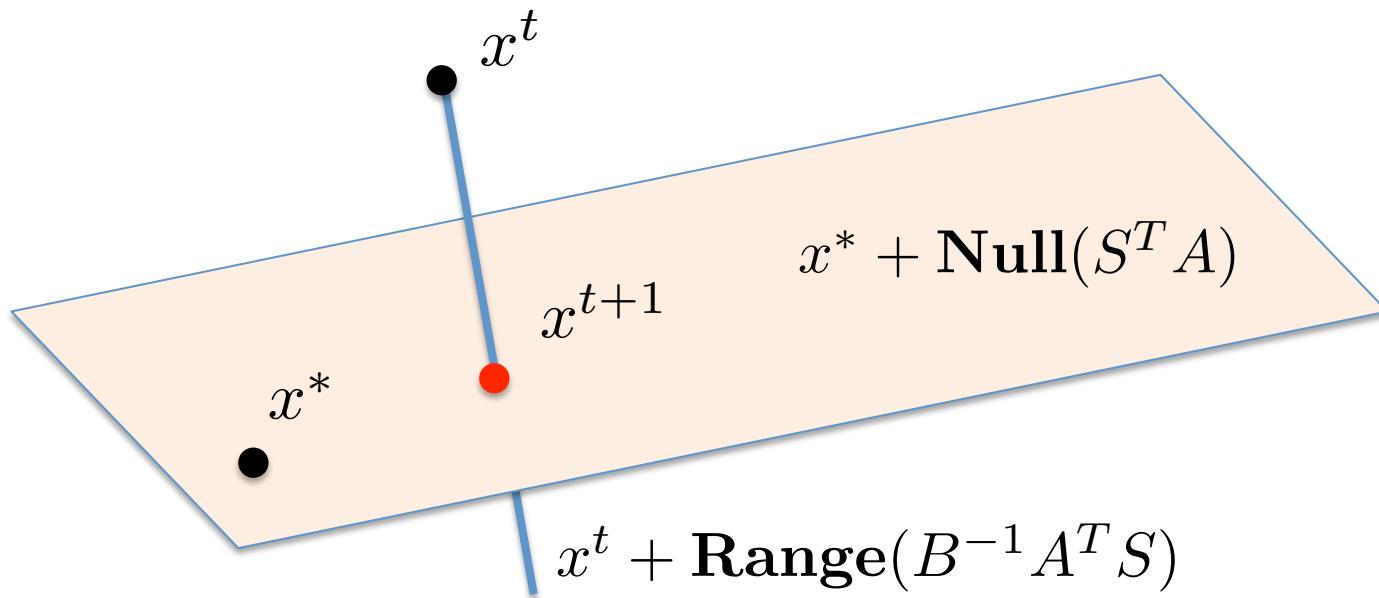
### “Constrain and Approximate”

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

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

$y$  is free

### 3. Geometric Viewpoint “Random Intersect”



**Lemma**  $\text{Null}(S^T A)$  and  $\text{Range}(B^{-1} A^T S)$  are  $B$ -orthogonal complements

*Proof*  $h \in \text{Null}(S^T A) \Rightarrow \langle B^{-1} A^T S y, h \rangle_B = (y^T S^T A B^{-1}) B h = y^T S^T A h = 0$

$$\{x^{t+1}\} = (x^* + \text{Null}(S^T A)) \cap (x^t + \text{Range}(B^{-1} A^T S))$$

## 4. Algebraic Viewpoint “Random Linear Solve”

$x^{t+1}$  = solution in  $x$  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”

Random Update Vector

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

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

Moore-Penrose  
pseudo-inverse

**Some properties:**

1.  $MM^\dagger M = M$
2.  $M^\dagger MM^\dagger = M^\dagger$
3.  $(M^\top M)^\dagger M^\top = M^\dagger$
4.  $(M^\top)^\dagger = (M^\dagger)^\top$
5.  $(MM^\top)^\dagger = (M^\dagger)^\top M^\dagger$

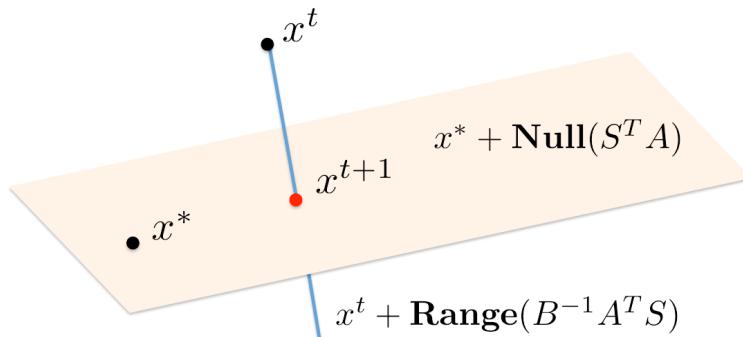
# 6. Analytic Viewpoint

## “Random Fixed Point”

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

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

Random Iteration Matrix



$$(B^{-1} Z)^2 = B^{-1} Z$$

$$(I - B^{-1} Z)^2 = I - B^{-1} Z$$

$B^{-1} Z$  projects orthogonally onto **Range**( $B^{-1} A^T S$ )  
 $I - B^{-1} Z$  projects orthogonally onto **Null**( $S^T A$ )

# Verifying that $B^{-1}Z$ is a Projection

$$\begin{aligned}(B^{-1}Z)^2 &= B^{-1}A^T S(S^T A B^{-1} A^T S)^\dagger S^T A B^{-1} A^T S(S^T A B^{-1} A^T S)^\dagger S^T A \\ &= B^{-1}A^T S(S^T A B^{-1} A^T S)^\dagger S^T A \\ &= B^{-1}Z\end{aligned}$$

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

$$M^\dagger M M M^\dagger = M^\dagger$$

Eigenvalues of  $B^{-1}Z$  are in  $\{0,1\}$

# Theory

# Complexity / Convergence

**Theorem [RG'15]** For every solution  $x^*$  of  $Ax = b$  we have

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

Moreover,

1

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

2

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



$$\rho := \|I - B^{-1}\mathbf{E}[Z]\|_B$$



$$\|M\|_B := \max_{\|x\|_B=1} \|Mx\|_B$$

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

# Proof of

1

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

Taking expectations conditioned on  $x^t$ , we get

$$\mathbf{E}[x^{t+1} - x^* \mid x^t] = (I - B^{-1}\mathbf{E}[Z])(x^t - x^*).$$

Taking expectation again gives

$$\begin{aligned}\mathbf{E}[x^{t+1} - x^*] &= \mathbf{E}[\mathbf{E}[x^{t+1} - x^* \mid x^t]] \\ &= \mathbf{E}[(I - B^{-1}\mathbf{E}[Z])(x^t - x^*)] \\ &= (I - B^{-1}\mathbf{E}[Z])\mathbf{E}[x^t - x^*].\end{aligned}$$

Applying the norms to both sides we obtain the estimate

$$\|\mathbf{E}[x^{t+1} - x^*]\|_B \leq \boxed{\|I - B^{-1}\mathbf{E}[Z]\|_B} \|\mathbf{E}[x^t - x^*]\|_B.$$

$\rho$

# The Rate: Lower and Upper Bounds

$$d := \text{Rank}(S^T A) = \dim(\text{Range}(B^{-1} A^T S)) = \text{Tr}(B^{-1} Z)$$

**Theorem [RG'15]**

$$0 \leq 1 - \frac{\mathbf{E}[d]}{n} \leq \rho \leq 1$$

**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 improves as the dimension of the search space in the “constrain and approximate” viewpoint grows.

# Proof

$$\begin{aligned}
 \rho &= \|I - B^{-1} \mathbf{E}[Z]\|_B \\
 \text{Direct calculation} \rightarrow &= \lambda_{\max}(I - B^{-1/2} \mathbf{E}[Z] B^{-1/2}) \\
 \|M\|_B := \max_{\|x\|_B=1} \|Mx\|_B &= 1 - \lambda_{\min}(B^{-1/2} \mathbf{E}[Z] B^{-1/2}) \\
 &= 1 - \lambda_{\min}(\mathbf{E}[B^{-1/2} Z B^{-1/2}]) \\
 \text{$XY$ and $YX$ have the same spectrum} \rightarrow &= 1 - \lambda_{\min}(\mathbf{E}[B^{-1} Z]) \\
 &\leq 1 - \frac{\text{Tr}(\mathbf{E}[B^{-1} Z])}{n} \\
 \text{Smallest eigenvalue is smaller than the average of all eigenvalues} \rightarrow &= 1 - \frac{\mathbf{E}[\text{Tr}(B^{-1} Z)]}{n}
 \end{aligned}$$

Upper bound

# The Rate: Sufficient Condition for Convergence

$$\rho = 1 - \lambda_{\min}(B^{-1}\mathbf{E}[Z])$$

## Lemma

If  $\mathbf{E}[Z]$  is invertible, then



- (i)  $\rho < 1$ ,
- (ii)  $A$  has full column rank, and
- (iii)  $x^*$  is unique

# Special Case: Randomized Kaczmarz Method

# Randomized Kaczmarz (RK) Method



M. S. Kaczmarz. **Angenäherte Auflösung von Systemen linearer Gleichungen**, *Bulletin International de l'Académie Polonaise des Sciences et des Lettres. Classe des Sciences Mathématiques et Naturelles. Série A, Sciences Mathématiques* 35, pp. 355–357, 1937

Kaczmarz method (1937)



T. Strohmer and R. Vershynin. **A Randomized Kaczmarz Algorithm with Exponential Convergence**. *Journal of Fourier Analysis and Applications* 15(2), pp. 262–278, 2009

Randomized Kaczmarz method (2009)

**RK arises as a special case for parameters  $B, S$  set as follows:**

$$B = I \quad S = e^i = (0, \dots, 0, 1, 0, \dots, 0) \text{ with probability } p_i$$

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

RK was analyzed for  $p_i = \frac{\|A_{i:}\|^2}{\|A\|_F^2}$



# RK: 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

$$\mathbf{P}(S = e^i) = p_i \rightarrow B = I \rightarrow S = e^i$$

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

## Complexity Rate

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

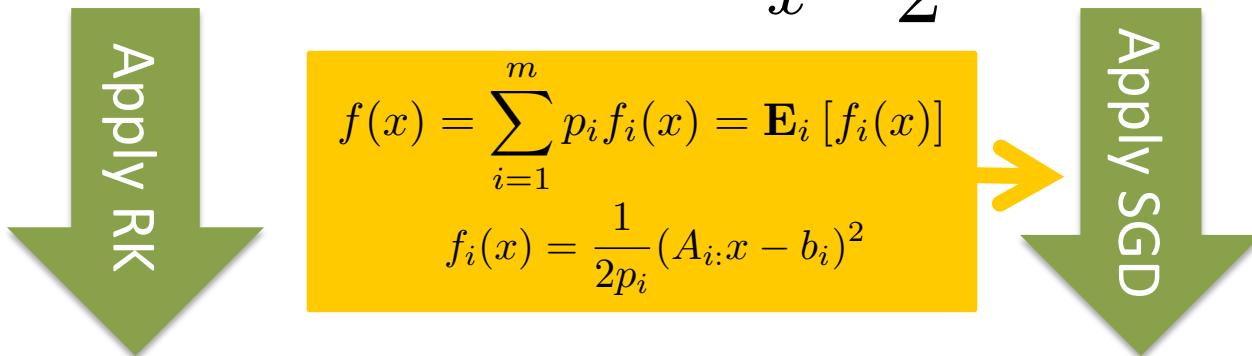
$$\mathbf{E} [\|x^t - x^*\|_2^2] \leq \left(1 - \frac{\lambda_{\min}(A^T A)}{\|A\|_F^2}\right)^t \|x^0 - x^*\|_2^2$$

# RK = SGD with a “smart” stepsize

$$Ax = b$$

vs

$$\min_x \frac{1}{2} \|Ax - b\|^2$$



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

$$\begin{aligned} x^{t+1} &= x^t - h^t \nabla f_i(x^t) \\ &= x^t - \frac{h^t}{p_i} (A_{i:}x^t - b_i) (A_{i:})^T \end{aligned}$$

RK is equivalent to applying SGD with a specific (smart!) constant stepsize!

$$x^{t+1} = \arg \min_{x \in \mathbb{R}^n} \|x - x^*\|_2^2 \quad \text{s.t.} \quad x = x^t + y (A_{i:})^T, \quad y \in \mathbb{R}$$

# RK: Further Reading



D. Needell. **Randomized Kaczmarz solver for noisy linear systems.** *BIT* 50 (2), pp. 395-403, 2010



D. Needell and J. Tropp. **Paved with good intentions: analysis of a randomized block Kaczmarz method.** *Linear Algebra and its Applications* 441, pp. 199-221, 2012



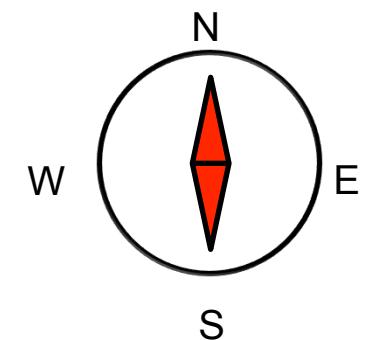
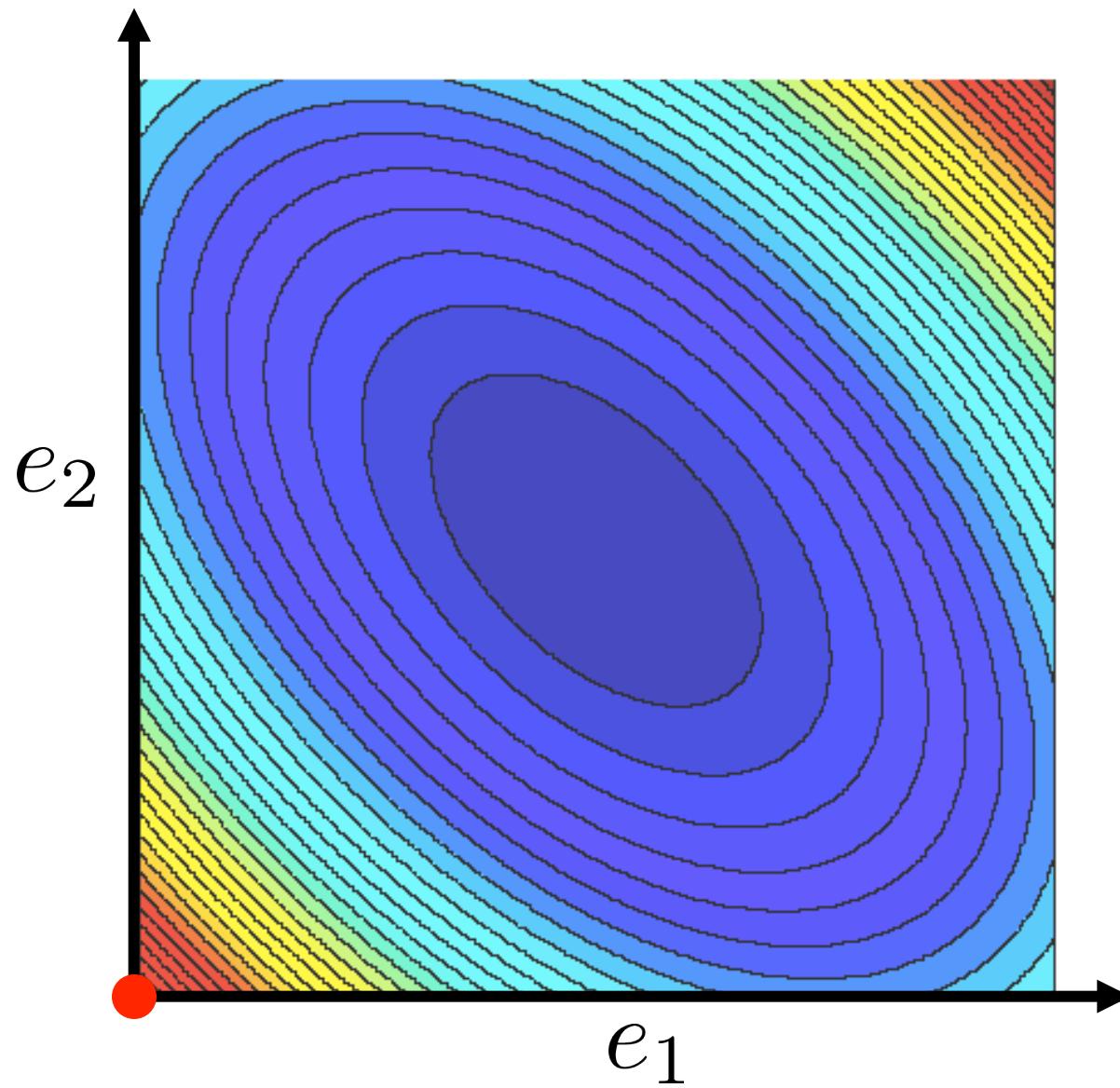
D. Needell, N. Srebro and R. Ward. **Stochastic gradient descent, weighted sampling and the randomized Kaczmarz algorithm.** *Mathematical Programming*, 2015 (arXiv:1310.5715)



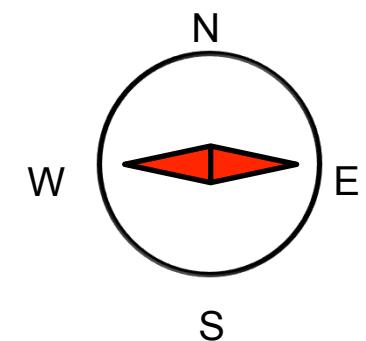
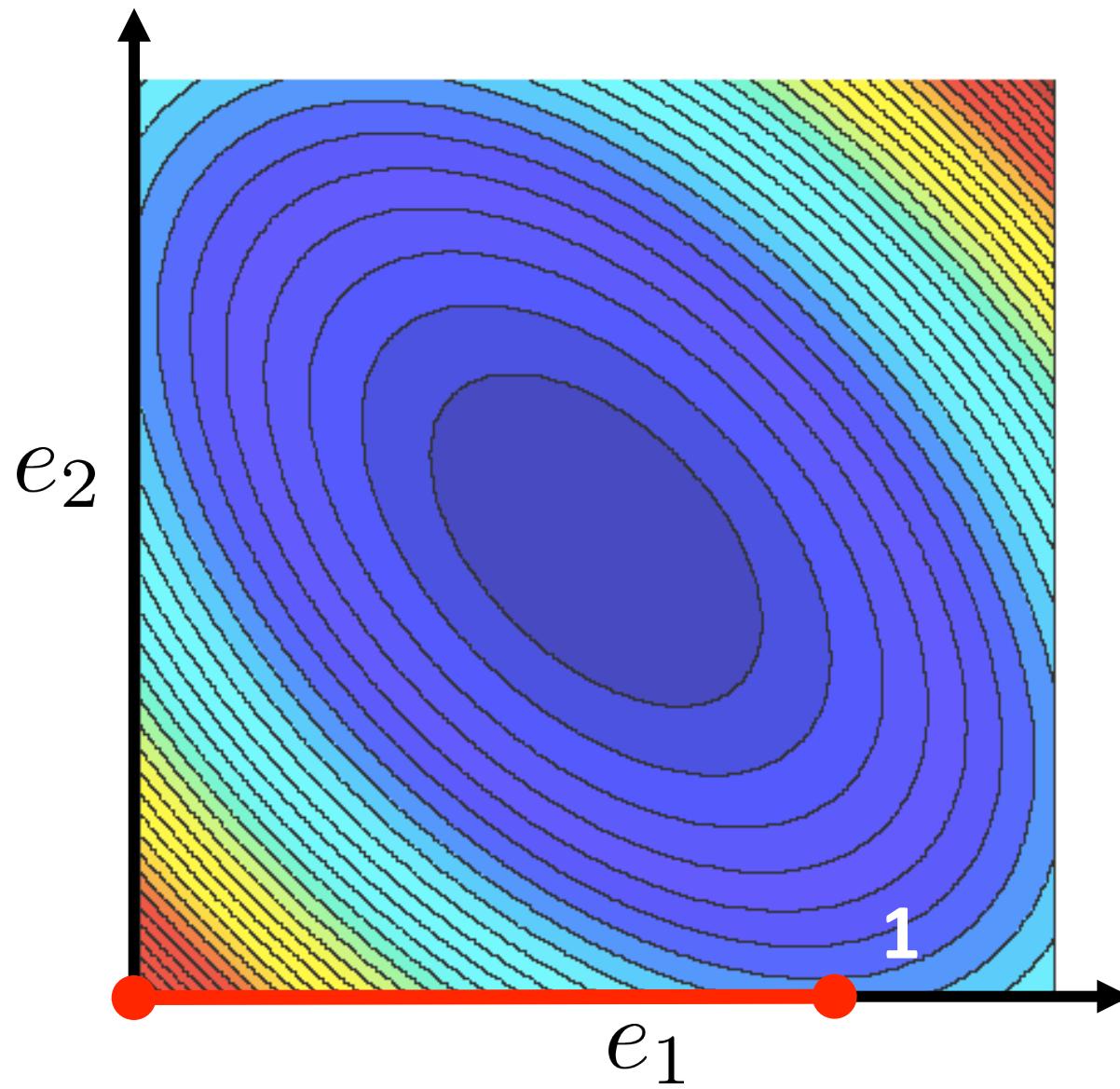
A. Ramdas. **Rows vs Columns for Linear Systems of Equations – Randomized Kaczmarz or Coordinate Descent?** *arXiv:1406.5295*, 2014

# Special Case: Randomized Coordinate Descent

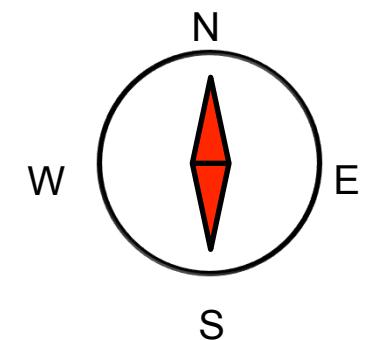
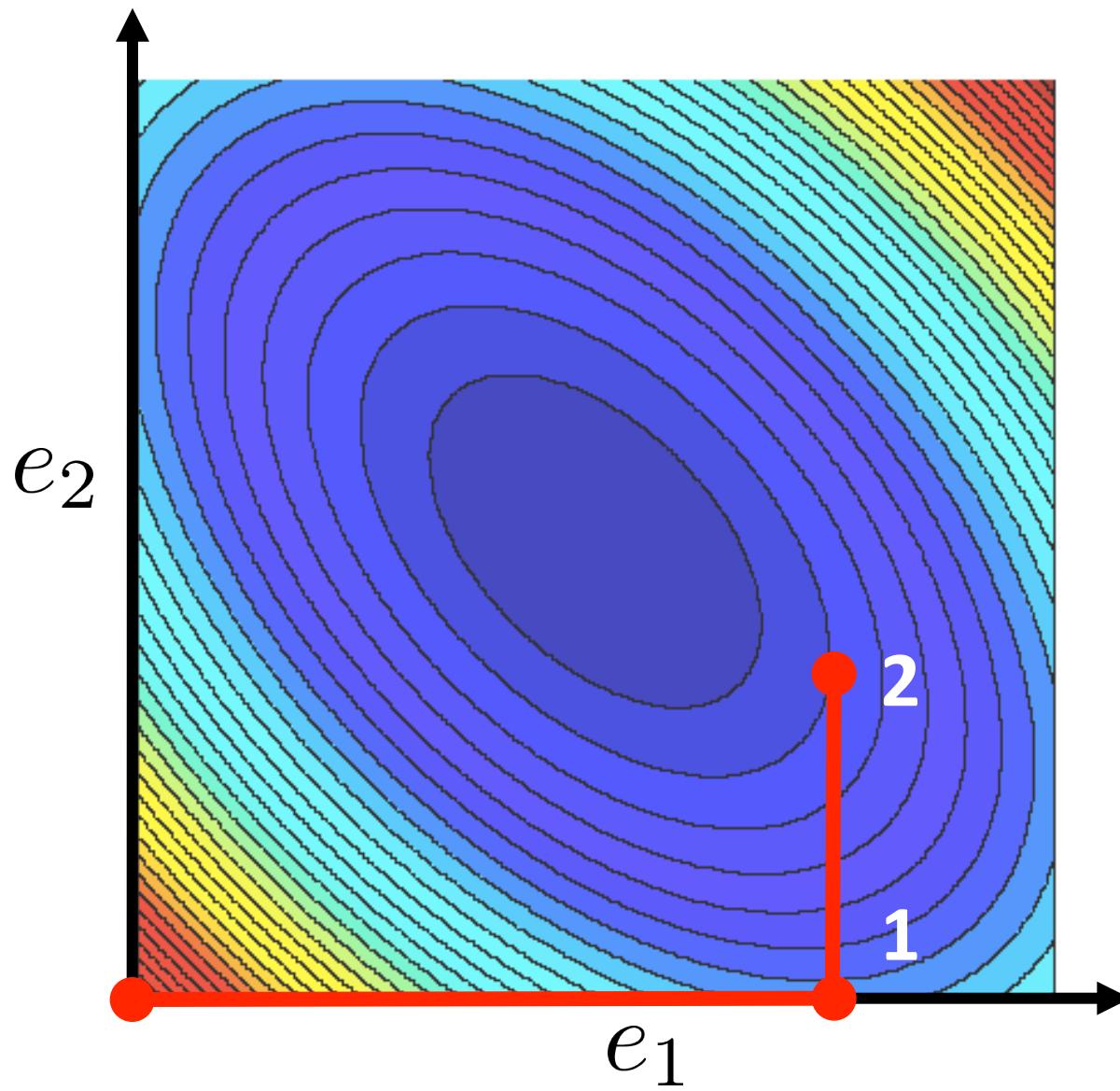
# Randomized Coordinate Descent in 2D



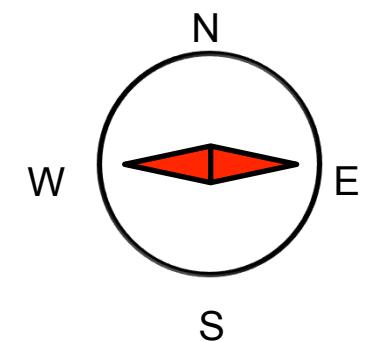
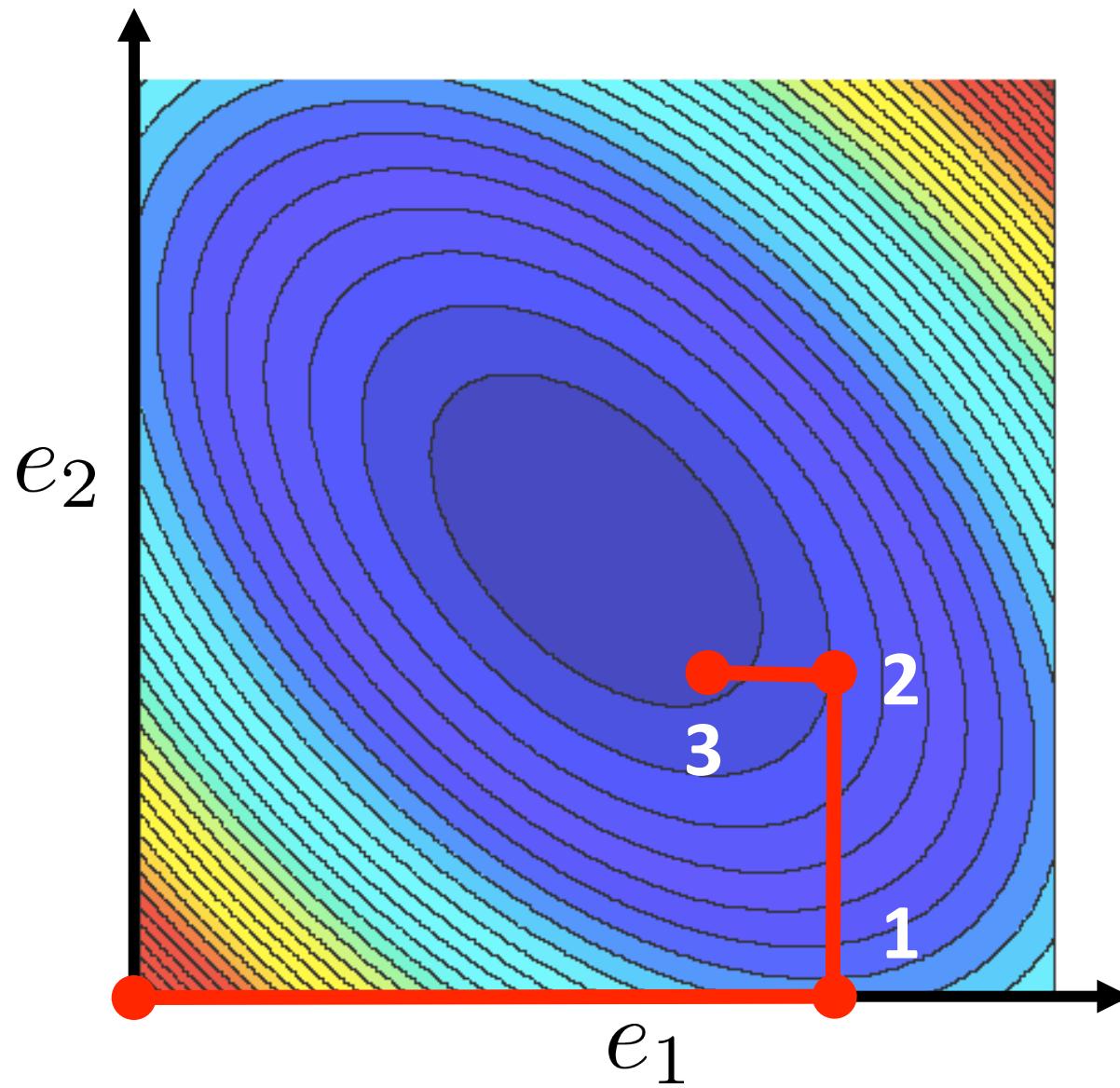
# Randomized Coordinate Descent in 2D



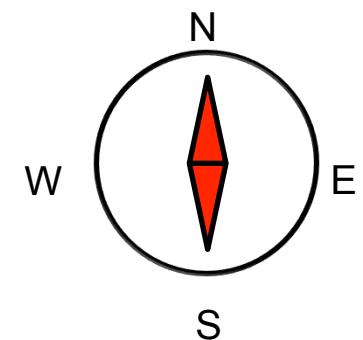
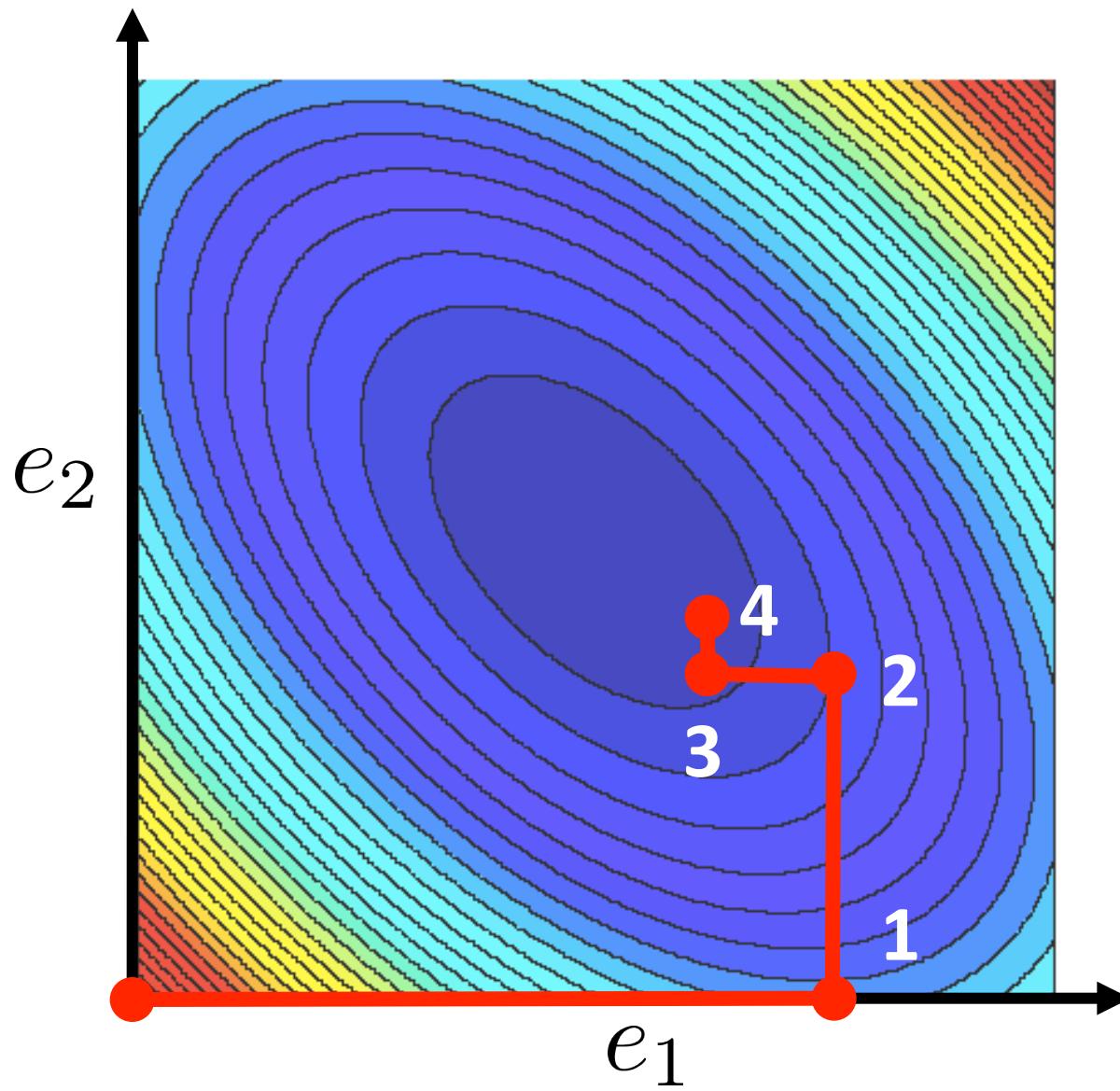
# Randomized Coordinate Descent in 2D



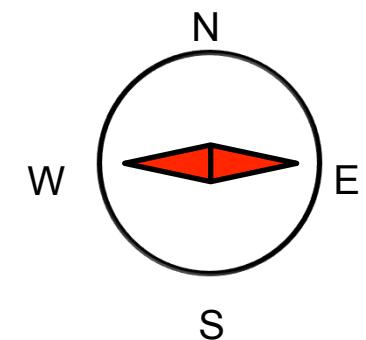
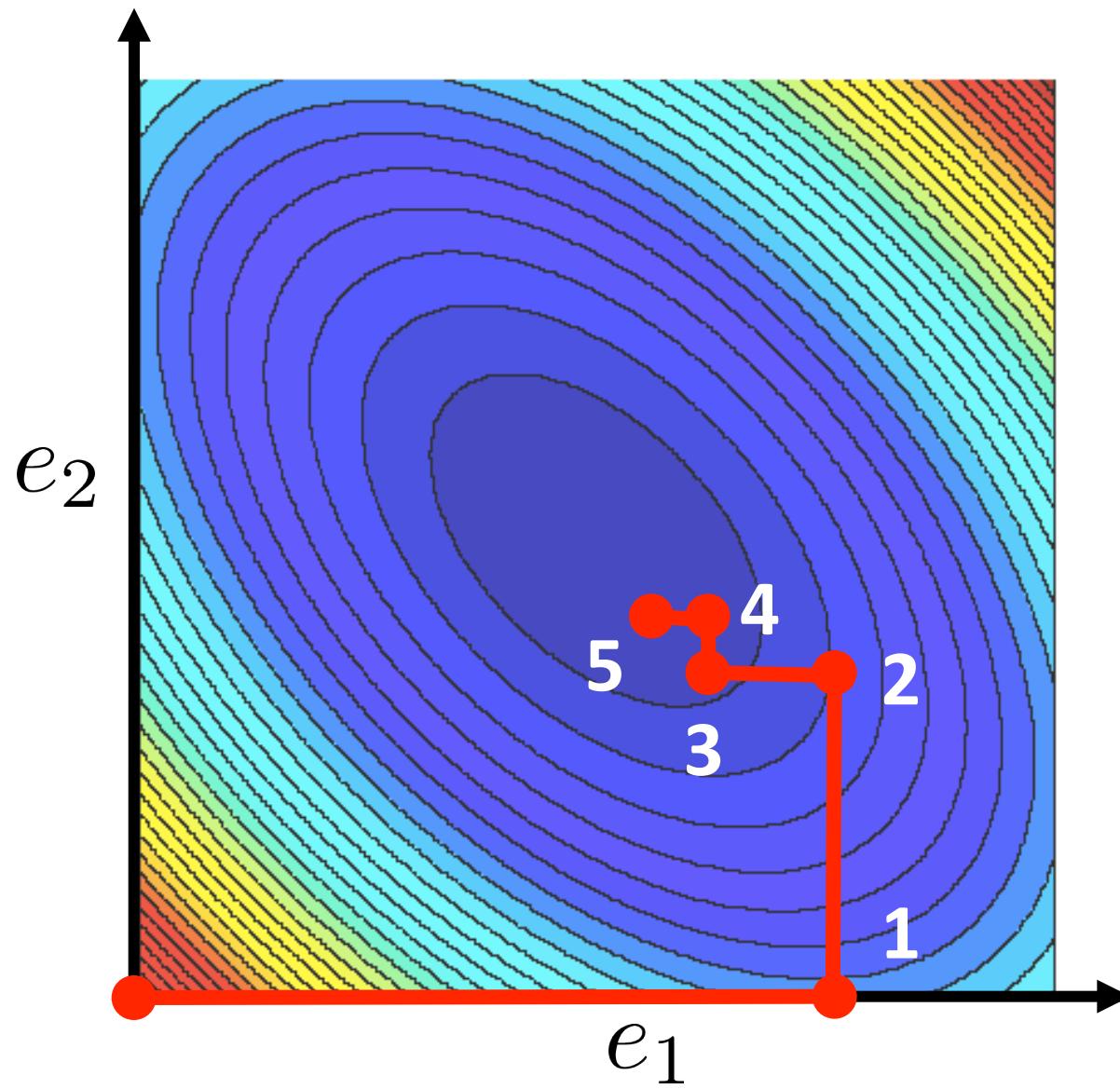
# Randomized Coordinate Descent in 2D



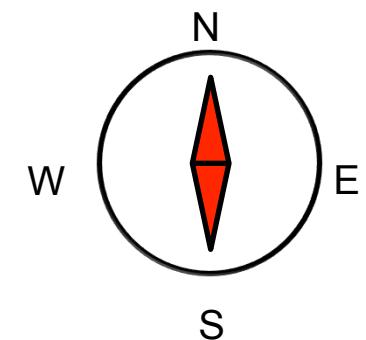
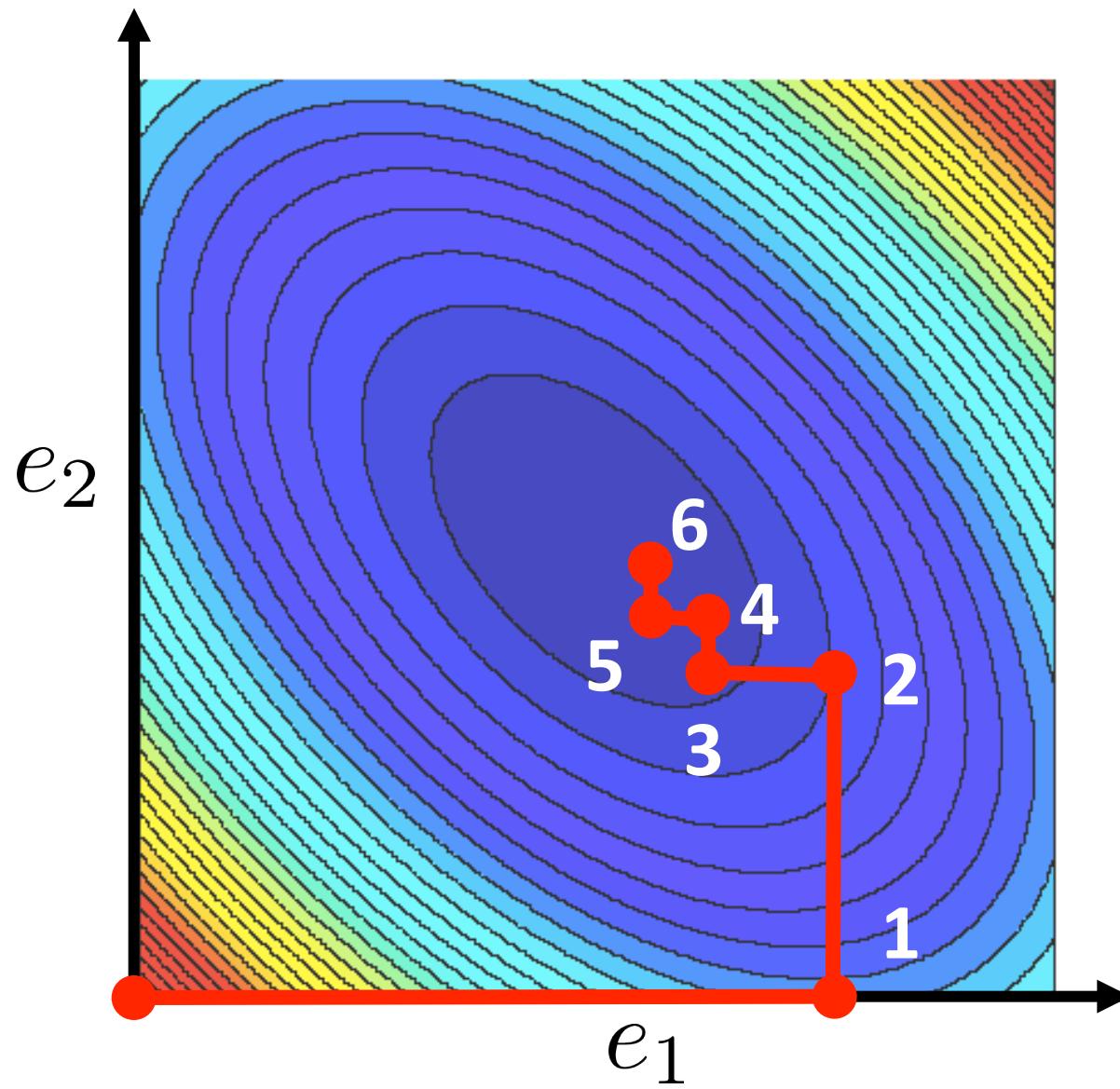
# Randomized Coordinate Descent in 2D



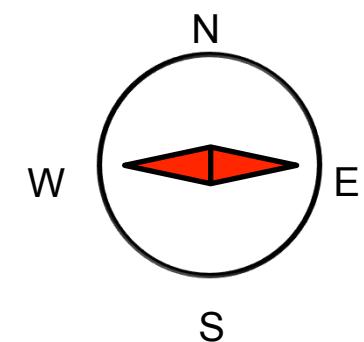
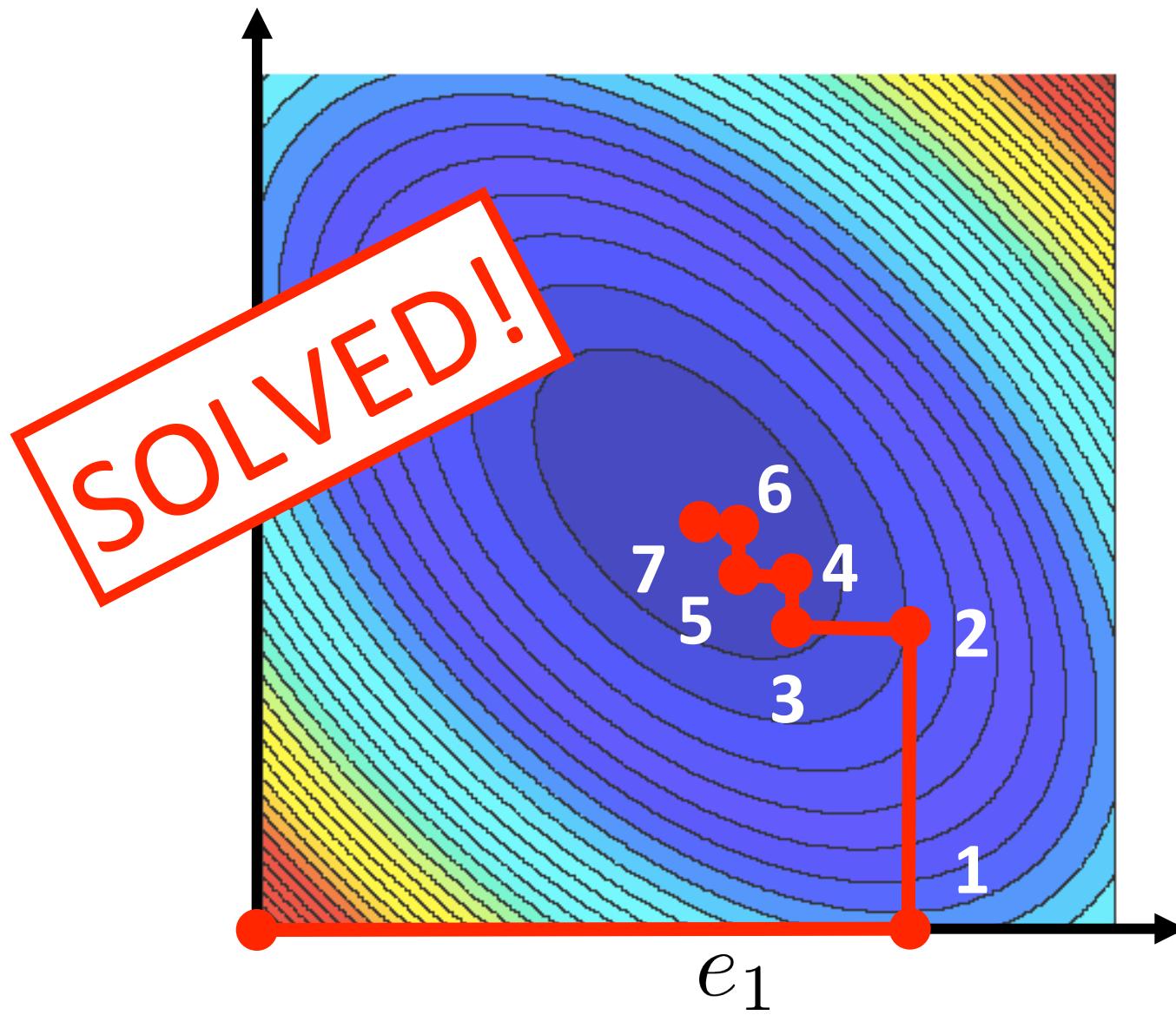
# Randomized Coordinate Descent in 2D



# Randomized Coordinate Descent in 2D



# Randomized Coordinate Descent in 2D



# Randomized Coordinate Descent (RCD)



A. S. Lewis and D. Leventhal. **Randomized methods for linear constraints: convergence rates and conditioning.** *Mathematics of OR* 35(3), 641-654, 2010 (arXiv:0806.3015)

RCD (2008)

$$\min_{x \in \mathbb{R}^n} [f(x) = \frac{1}{2}x^T Ax - b^T x]$$
$$x^* = A^{-1}b$$

Assume: Positive definite

**RCD arises as a special case for parameters  $B, S$  set as follows:**

$$B = A \quad S = e^i = (0, \dots, 0, 1, 0, \dots, 0) \text{ with probability } p_i$$

Recall: In RK we had  $B = I$

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

RCD was analyzed for  $p_i = \frac{A_{ii}}{\text{Tr}(A)}$

# RCD: 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

$$\begin{aligned} & B = A \\ \text{P}(S = e^i) = p_i \rightarrow & S = e^i \end{aligned}$$

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

## Complexity Rate

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

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

# RCD uses “Exact Line Search”

Recall Viewpoint 2 (“Constrain and Approximate”):

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

In RCD we have:  
 $B = A$     $S = e^i$

**Observation:**  $\|x - x^*\|_A^2 = (x - x^*)^T A(x - x^*)$

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

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

**Insight:**

$$\begin{aligned} x^{t+1} &= \arg \min_{x \in \mathbb{R}^n} f(x) \\ \text{subject to } &x = x^t + y e^i \\ &y \in \mathbb{R} \end{aligned}$$

RCD **exactly minimizes  $f$**  along a random coordinate direction!

# RCD: “Standard” Optimization Form



Yurii Nesterov. **Efficiency of coordinate descent methods on huge-scale optimization problems.** *SIAM J. on Optimization*, 22(2):341–362, 2012 (CORE Discussion Paper 2010/2)

Nesterov considered the problem:

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

Convex and smooth

Nesterov assumed that the following inequality holds for all  $x, h$  and  $i$ :

$$f(x + he^i) \leq f(x) + \nabla_i f(x)h + \frac{L_i}{2}h^2$$

Given a current iterate  $x$ , choosing  $h$  by minimizing the RHS gives:

**Nesterov’s RCD method:**

$$x^{t+1} = x^t - \frac{1}{L_i} \nabla_i f(x^t) e^i$$

$$f(x) = \frac{1}{2}x^T Ax - b^T x \Rightarrow \\ L_i = A_{ii} \quad \nabla_i f(x) = (A_{i:})^T x - b_i$$

We recover RCD as we have seen it:

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

# Special Case: Randomized Newton Method

# Randomized Newton (RN)



Z. Qu, PR, M. Takáč and O. Fercoq. **Stochastic Dual Newton Ascent for Empirical Risk Minimization.** *arXiv:1502.02268*, 2015

**SDNA**

$$\min_{x \in \mathbb{R}^n} [f(x) = \frac{1}{2}x^T Ax - b^T x]$$
$$x^* = A^{-1}b$$

Assume: Positive definite

RN arises as a special case for parameters  $B, S$  set as follows:

$$B = A \quad S = I_{:C} \text{ with probability } p_C$$

$$p_C \geq 0 \quad \forall C \subseteq \{1, \dots, n\} \quad \sum_{C \subseteq \{1, \dots, n\}} p_C = 1$$

RCD is special case with  $p_C = 0$  whenever  $|C| \neq 1$

# RN: Derivation

## General Method

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

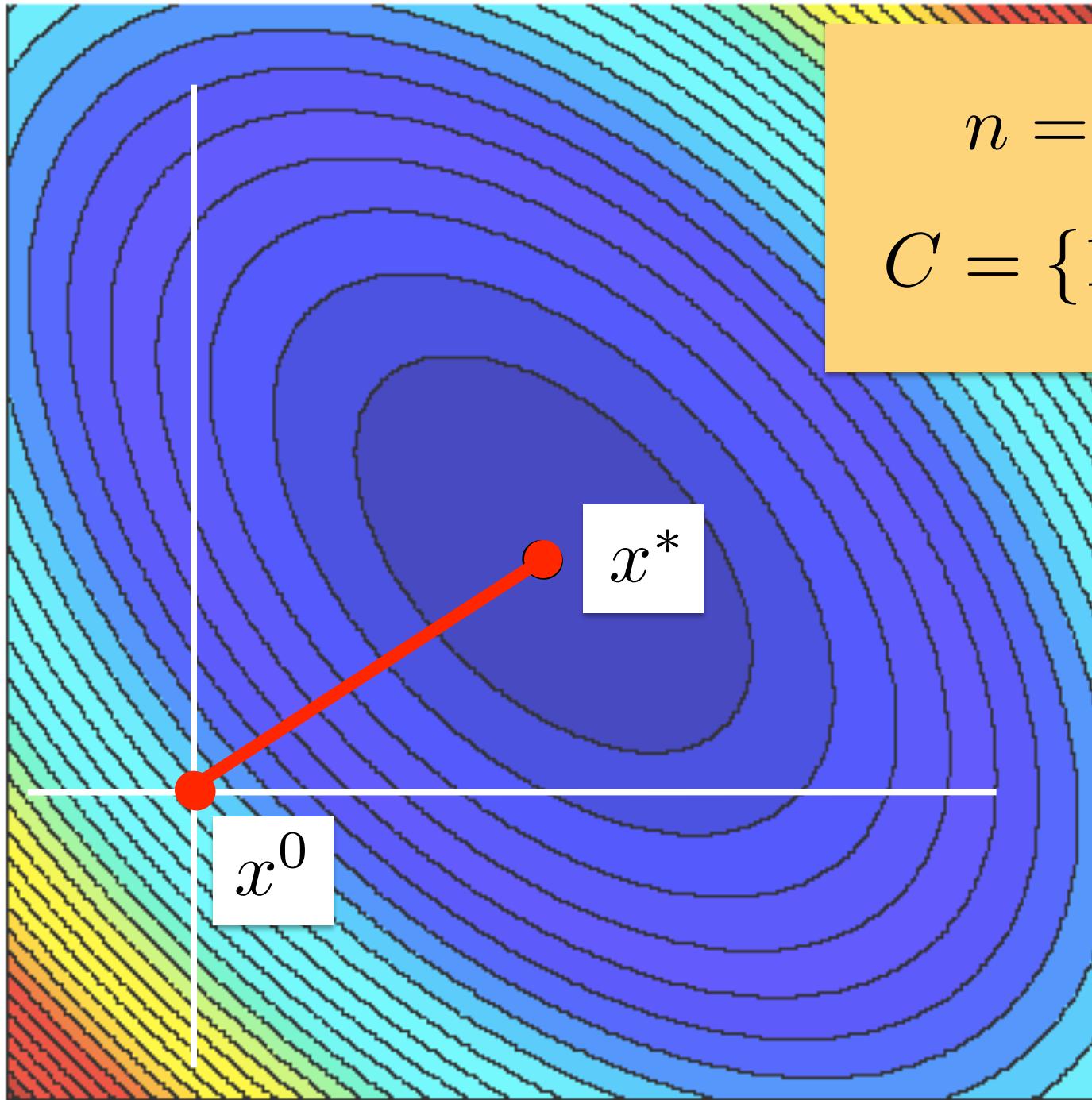
Special Choice of Parameters     $B = A$



$S = I_{:C}$  with probability  $p_C$

$$x^{t+1} = x^t - \boxed{I_{:C}} \boxed{((I_{:C})^T A I_{:C})^{-1}} \boxed{(I_{:C})^T (Ax^t - b)}$$

This method minimizes  $f$  exactly in a random subspace spanned by the coordinates belonging to  $C$



$$n = 2$$

$$C = \{1, 2\}$$

# Special Case: Gaussian Descent

# Gaussian Descent

## General Method

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

## Special Choice of Parameters

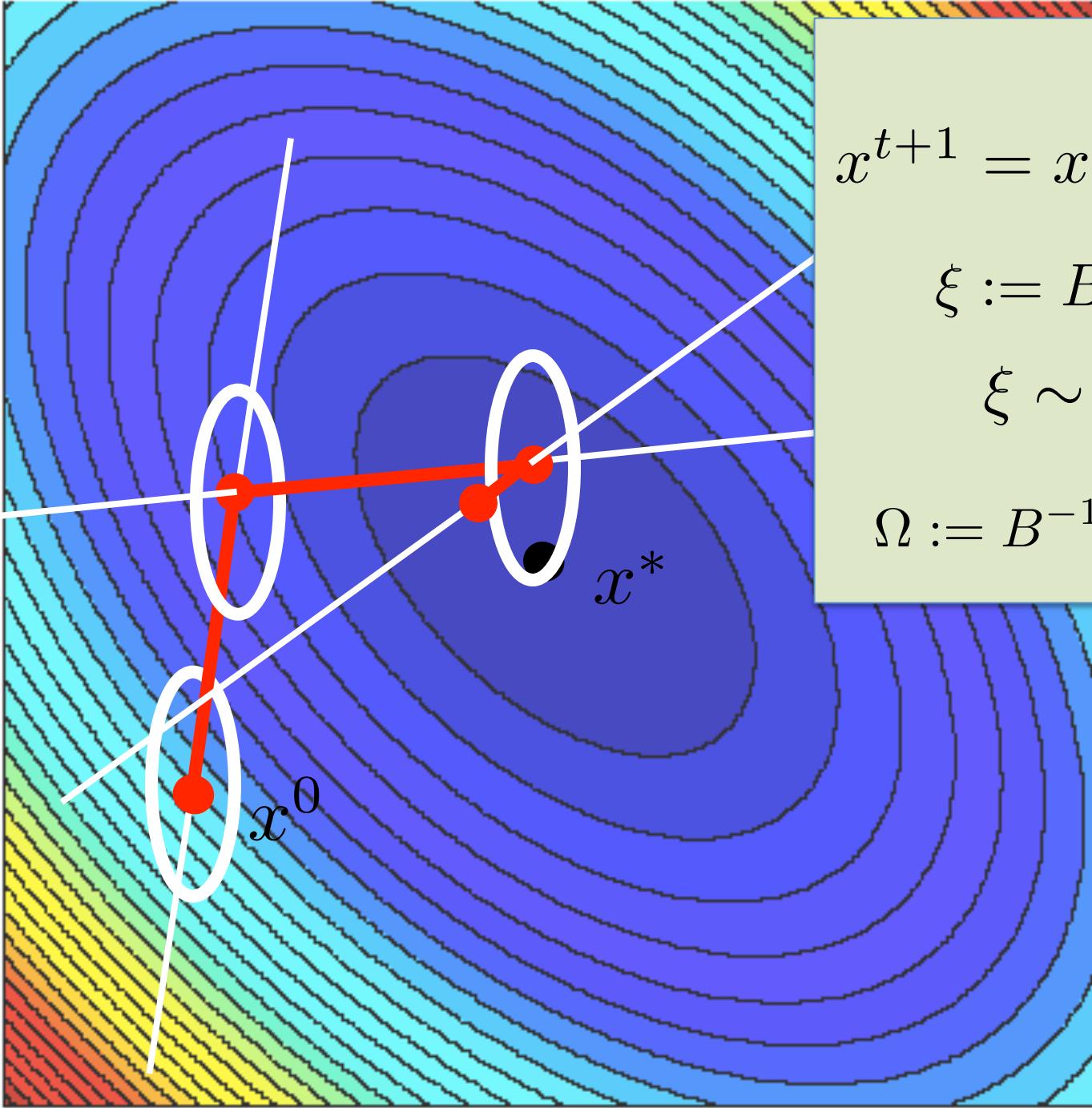
$$S \sim N(0, \Sigma) \quad \rightarrow$$

Positive definite covariance matrix

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

## Complexity Rate

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



A contour plot showing a function's value across a 2D space. The color bar on the left indicates values from red (high) to blue (low). The contours are concentric ellipses centered at a point labeled  $x^*$ . A red line with circular markers shows a path starting from  $x^0$  and moving towards  $x^*$ , illustrating the iterative steps of an optimization algorithm.

$$x^{t+1} = x^t - h^t B^{-1/2} \xi$$

$$\xi := B^{-1/2} A^T S$$

$$\xi \sim N(0, \Omega)$$

$$\Omega := B^{-1/2} A^T \Sigma A B^{-1/2}$$

# Gaussian Descent: The Rate

XY and YX  
have the  
same  
spectrum

$$\begin{aligned}\rho &= 1 - \lambda_{\min}(B^{-1} \mathbf{E}[Z]) \\ &= 1 - \lambda_{\min}\left(B^{-1/2} \mathbf{E}[Z] B^{-1/2}\right) \\ &= 1 - \lambda_{\min}\left(B^{-1/2} \mathbf{E}\left[A^T S (S^T A B^{-1} A^T S)^{\dagger} S^T A\right] B^{-1/2}\right) \\ &= 1 - \lambda_{\min}\left(\mathbf{E}\left[B^{-1/2} A^T S (S^T A B^{-1} A^T S)^{\dagger} S^T A B^{-1/2}\right]\right) \\ &= 1 - \lambda_{\min}\left(\mathbf{E}\left[\frac{\xi \xi^T}{\|\xi\|_2^2}\right]\right)\end{aligned}$$

$$\xi := B^{-1/2} A^T S$$

$$\xi \sim N(0, \Omega)$$

$$\Omega := B^{-1/2} A^T \Sigma A B^{-1/2}$$

# Gaussian Descent: The Rate

**Lemma [GR'15]**

$$\mathbf{E} \left[ \frac{\xi \xi^T}{\|\xi\|_2^2} \right] \succeq \frac{2}{\pi} \frac{\Omega}{\text{Tr}(\Omega)}$$

$$\rho \leq 1 - \frac{2}{\pi} \frac{\lambda_{\min}(\Omega)}{\text{Tr}(\Omega)}$$

This follows from the general lower bound  $1 - \frac{\mathbf{E}[d]}{n} \leq \rho$  since  $d = 1$

# Gaussian Descent: Further Reading



Yurii Nesterov. **Random gradient-free minimization of convex functions.** CORE Discussion Paper # 2011/1, 2011



S. U. Stich, C. L. Muller and G. Gartner. **Optimization of convex functions with random pursuit.** SIAM Journal on Optimization 23 (2), pp. 1284-1309, 2014



S. U. Stich. **Convex optimization with random pursuit.** PhD Thesis, ETH Zurich, 2014

# EXTRA MATERIAL

# Importance Sampling

# Importance Sampling

Assume that  $S$  is discrete:

$$S = S_i \quad \text{with probability} \quad p_i \quad (i = 1, \dots, r)$$

## Question

Consider  $S_1, \dots, S_r$  fixed. How to choose the probabilities  $p_1, \dots, p_r$  which optimize the convergence rate  $\rho = 1 - \lambda_{\min}(B^{-1}\mathbf{E}[Z])$  ?

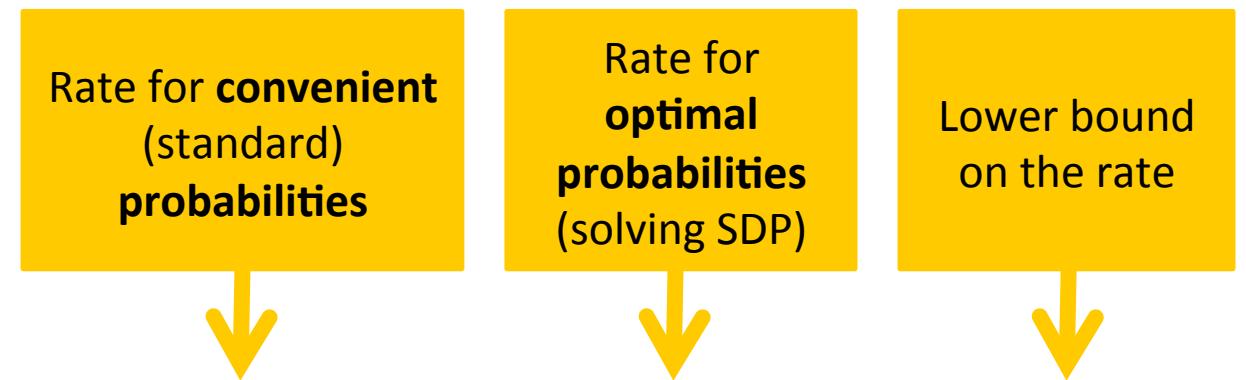
$$\max_p \left\{ \lambda_{\min}(B^{-1}\mathbf{E}[Z]) \quad \text{subject to} \quad \sum_{i=1}^r p_i = 1, \quad p \geq 0 \right\}$$

- Can be reformulated as an **SDP (Semidefinite Program)**
- Leads to different probabilities than those proposed for RK and RCD!

$$V_i = B^{-1/2} A^T S_i$$

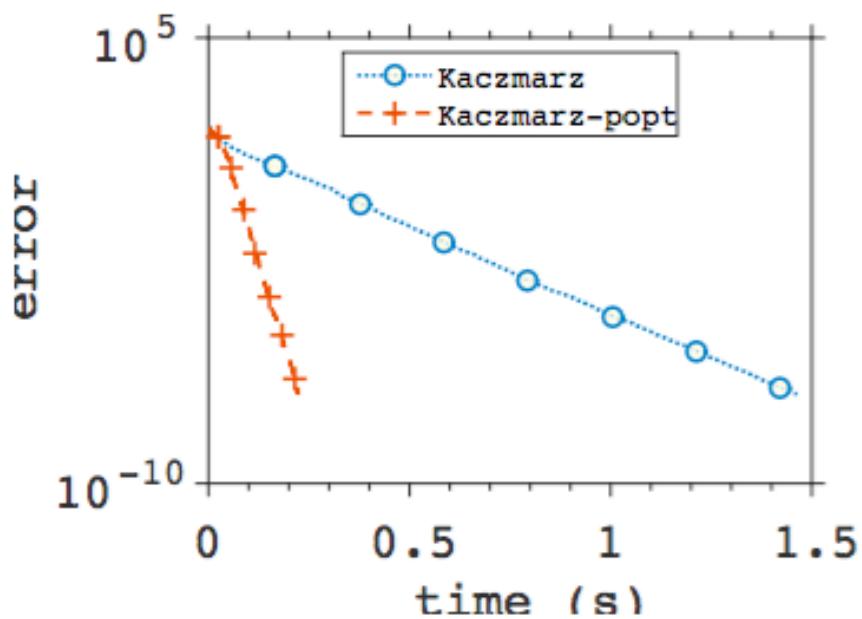
$$\begin{aligned} & \max_{p,t} && t \\ & \text{subject to} && \sum_{i=1}^r p_i (V_i(V_i^T V_i)^\dagger V_i^T) \succeq t \cdot I, \\ & && p \geq 0, \quad \sum_{i=1}^r p_i = 1 \end{aligned}$$

# RCD: Optimal Probabilities Can Lead to a Remarkable Improvement

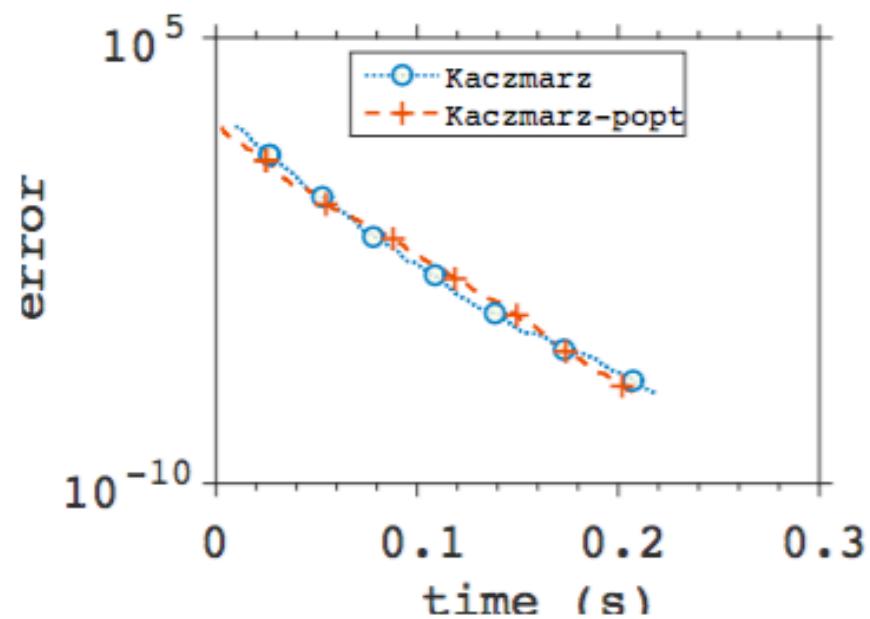


data set	$\rho_c$	$\rho^*$	$1 - 1/n$
rand(50,50)	$1 - 2 \cdot 10^{-6}$	$1 - 3.05 \cdot 10^{-6}$	$1 - 2 \cdot 10^{-2}$
mushrooms-ridge	$1 - 5.86 \cdot 10^{-6}$	$1 - 7.15 \cdot 10^{-6}$	$1 - 8.93 \cdot 10^{-3}$
aloi-ridge	$1 - 2.17 \cdot 10^{-7}$	$1 - 1.26 \cdot 10^{-4}$	$1 - 7.81 \cdot 10^{-3}$
liver-disorders-ridge	$1 - 5.16 \cdot 10^{-4}$	$1 - 8.25 \cdot 10^{-3}$	$1 - 1.67 \cdot 10^{-1}$
covtype.binary-ridge	$1 - 7.57 \cdot 10^{-14}$	$1 - 1.48 \cdot 10^{-6}$	$1 - 1.85 \cdot 10^{-2}$

# RK: Convenient vs Optimal

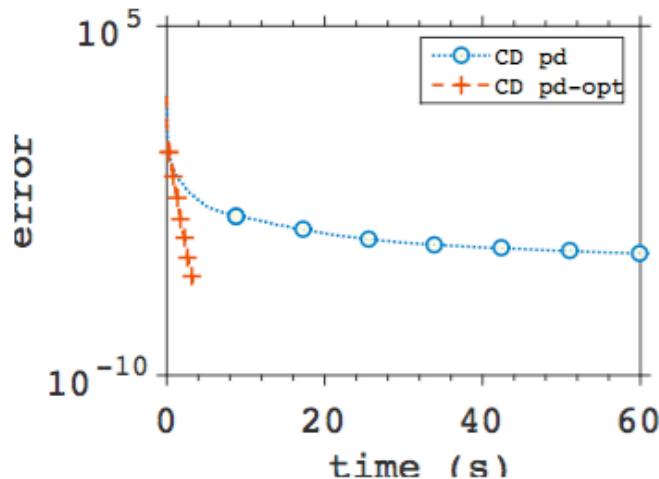


(a) `liver-disorders-popt-k`

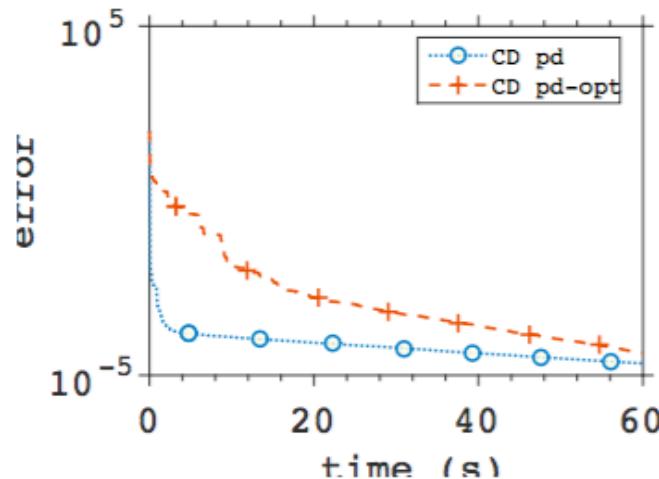


(b) `rand(500,100)`

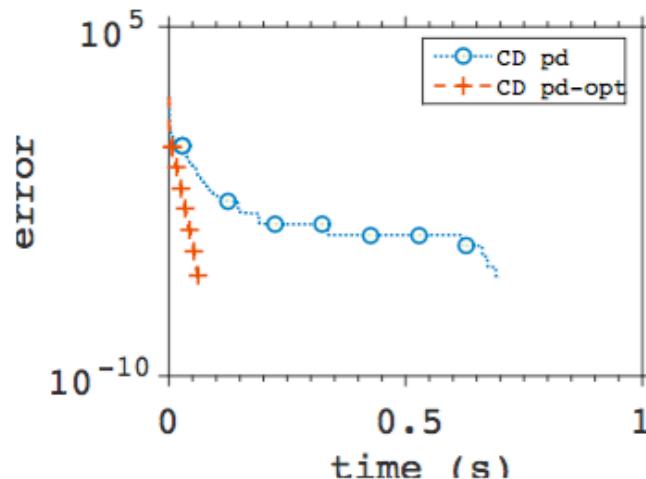
# RCD: Convenient vs Optimal



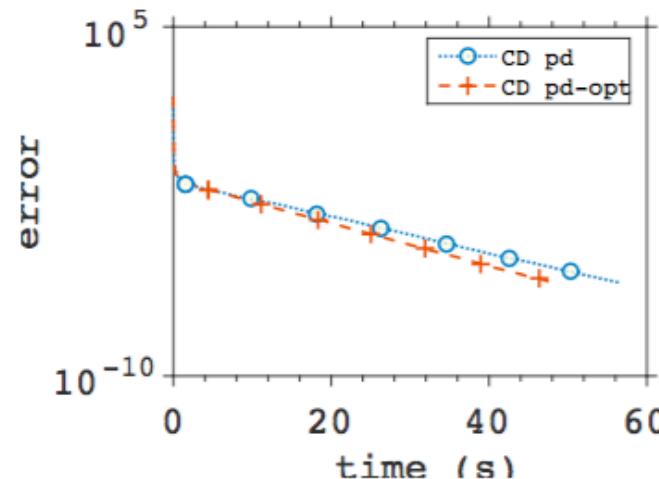
(a) aloi



(b) covtype.libsvm.binary



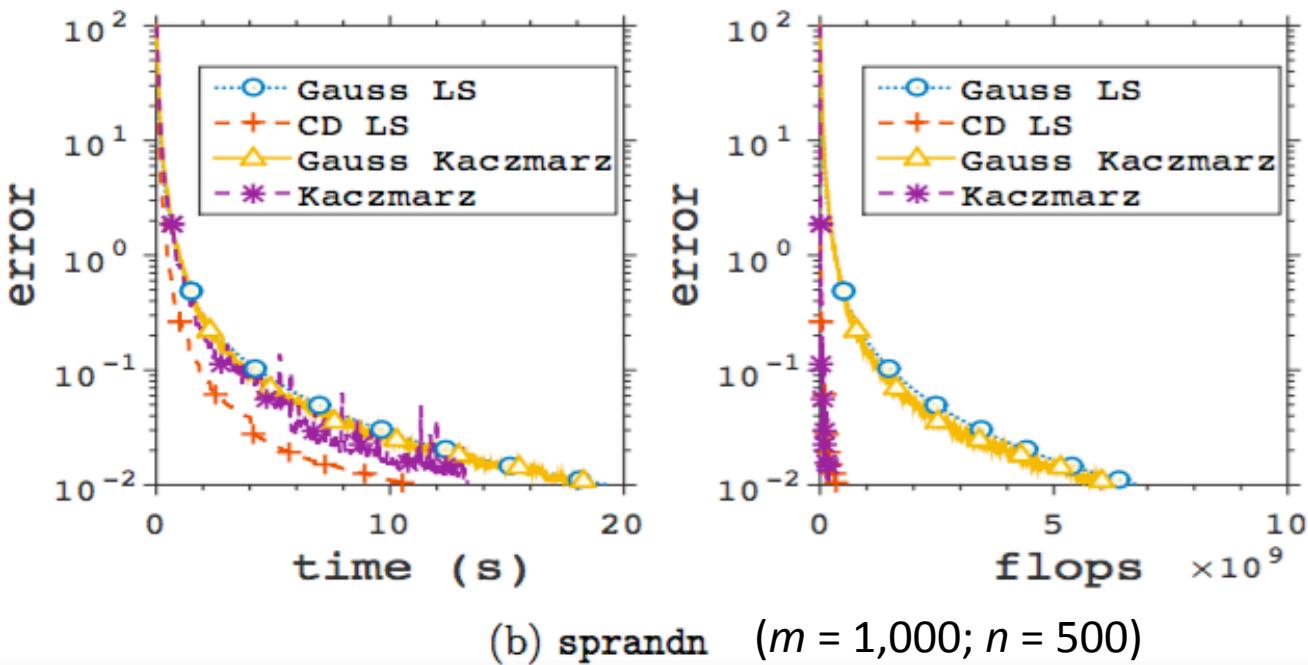
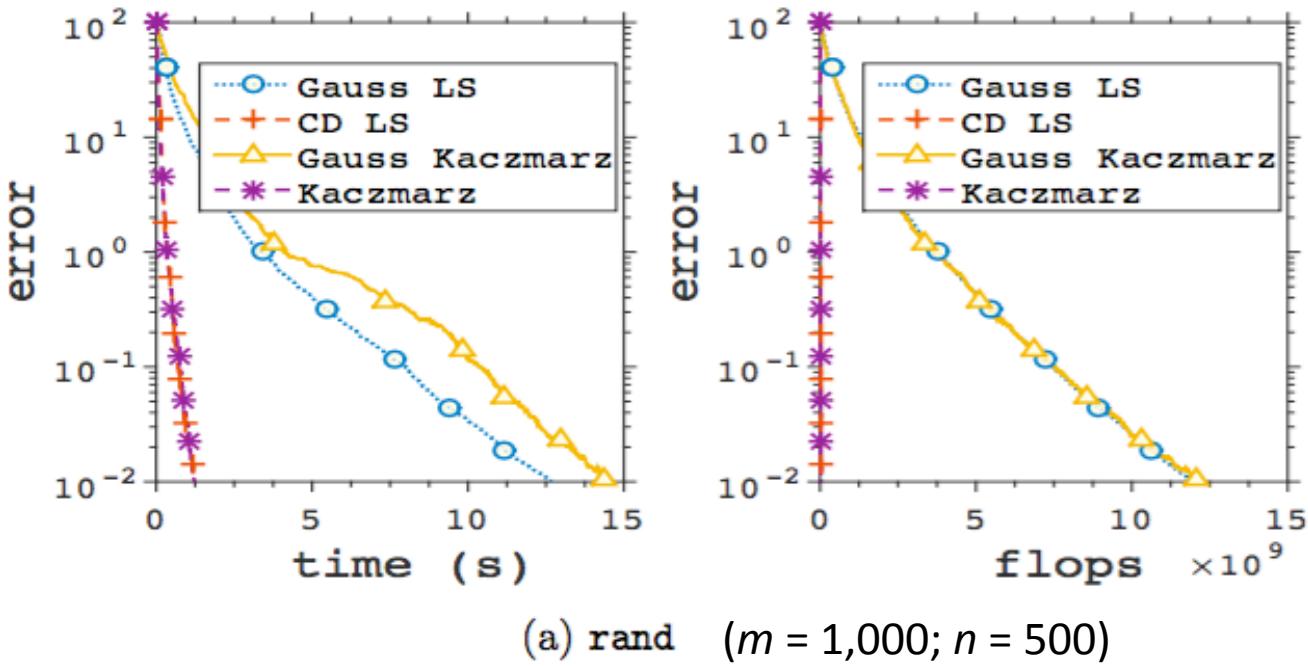
(c) liver-disorders-ridge



(d) mushrooms-ridge-opt

# Experiments

# Synthetic data



# Real data (Matrix Market)

