Randomized Iterative Methods for Linear Systems

Robert Mansel Gower & Peter Richtárik

University of Edinburgh

LMS Inverse Day on Large-scale and Nonlinear Inverse Problems

September 21, 2015

Robert M Gower
(Edinburgh)

Robert M Gower and P.R.
Randomized Iterative Methods for Linear Systems
arXiv:1506.03296, 2015

The Problem

The Problem

$$m \left\{ \overbrace{A}^{n} \overset{\in \mathbb{R}^n}{x} = b \right\} m$$

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

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

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

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

Minimizing Convex Quadratics

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

This system is consistent

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

A = 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

157 56

1. Relaxation Viewpoint "Sketch and Project"

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

B: Symmetric and positive definite

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

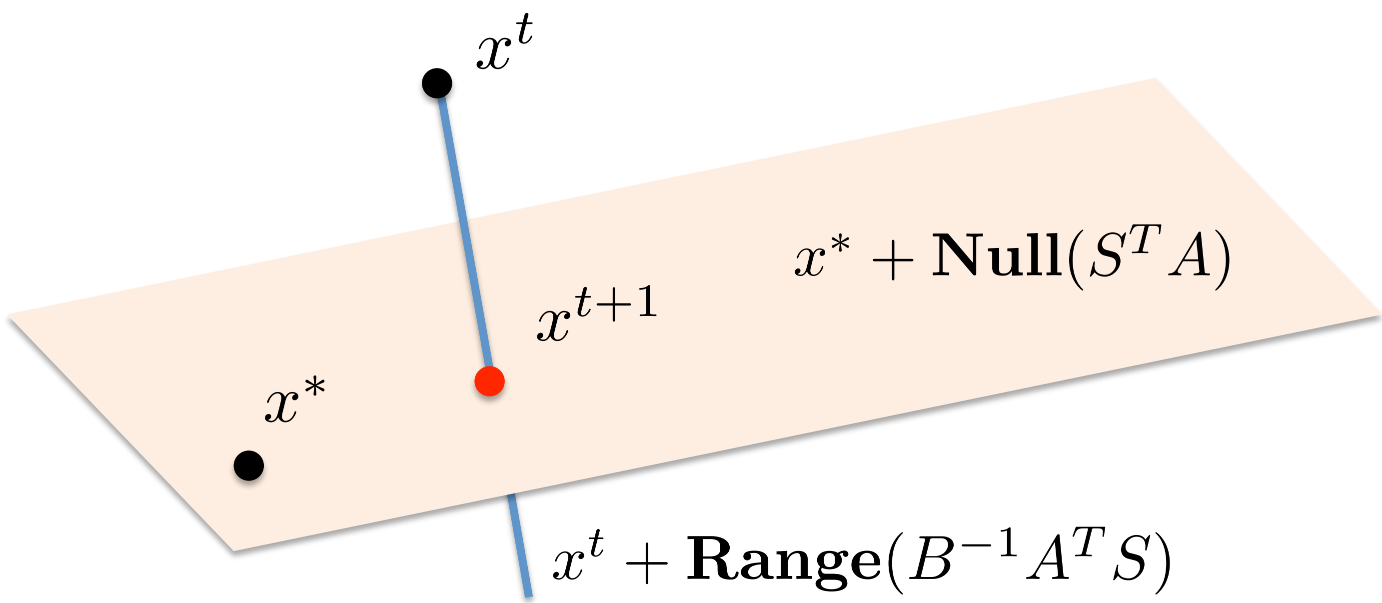
$$\text{subject to} \quad S^T A x = S^T b$$

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

2. Optimization Viewpoint "Constrain and Approximate"

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

3. Geometric Viewpoint "Random Intersect"



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

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

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

4. Algebraic Viewpoint
"Random Linear Solve"

$x^{t+1} \quad = \quad$ 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"

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

Moore-Penrose pseudo-inverse

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

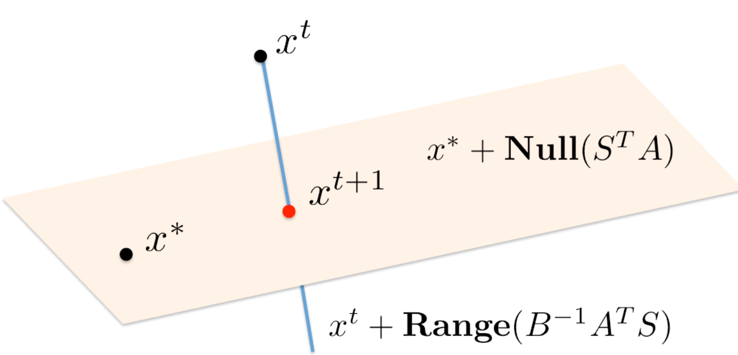
Some properties:

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"

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

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



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

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

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

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

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

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

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

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

Theory

Complexity / Convergence

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

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

Moreover,

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

2 $$\mathbf{E}[Z] \succ 0$$

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

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

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

Proof of 1

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

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

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

Applying the norms to both sides we obtain the estimate

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

The Rate: Lower and Upper Bounds

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

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

$$
\begin{aligned}
\rho &= \|I - B^{-1}\mathbf{E}[Z]\|_B \\
&= \lambda_{\max}(I - B^{-1/2}\mathbf{E}[Z]B^{-1/2}) \\
&= 1 - \lambda_{\min}(B^{-1/2}\mathbf{E}[Z]B^{-1/2}) \\
&= 1 - \lambda_{\min}(\mathbf{E}[B^{-1/2}ZB^{-1/2}]) \\
&= 1 - \lambda_{\min}(\mathbf{E}[B^{-1}Z]) \\
&\geq 1 - \frac{\mathbf{Tr}\left(\mathbf{E}\left[B^{-1}Z\right]\right)}{n} \\
&= 1 - \frac{\mathbf{E}\left[\mathbf{Tr}(B^{-1}Z)\right]}{n}
\end{aligned}
$$

Direct calculation

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

XY and *YX* have the same spectrum

Upper bound

Smallest eigenvalue is smaller then the average of all eigenvalues

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. **Angenaherte Auflosung 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*:

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

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

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

RK: Derivation and Rate

General Method

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

Special Choice of Parameters

$$B = I$$

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

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

Complexity Rate

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

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

RK = SGD with a "smart" stepsize

$$Ax = b \qquad \text{vs} \qquad \min_x \frac{1}{2}\|Ax - b\|^2$$

Apply RK

$$f(x) = \sum_{i=1}^{m} p_i f_i(x) = \mathbf{E}_i\left[f_i(x)\right]$$

$$f_i(x) = \frac{1}{2p_i}(A_{i:}x - b_i)^2$$

Apply SGD

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

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

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

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

RK: Further Reading

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

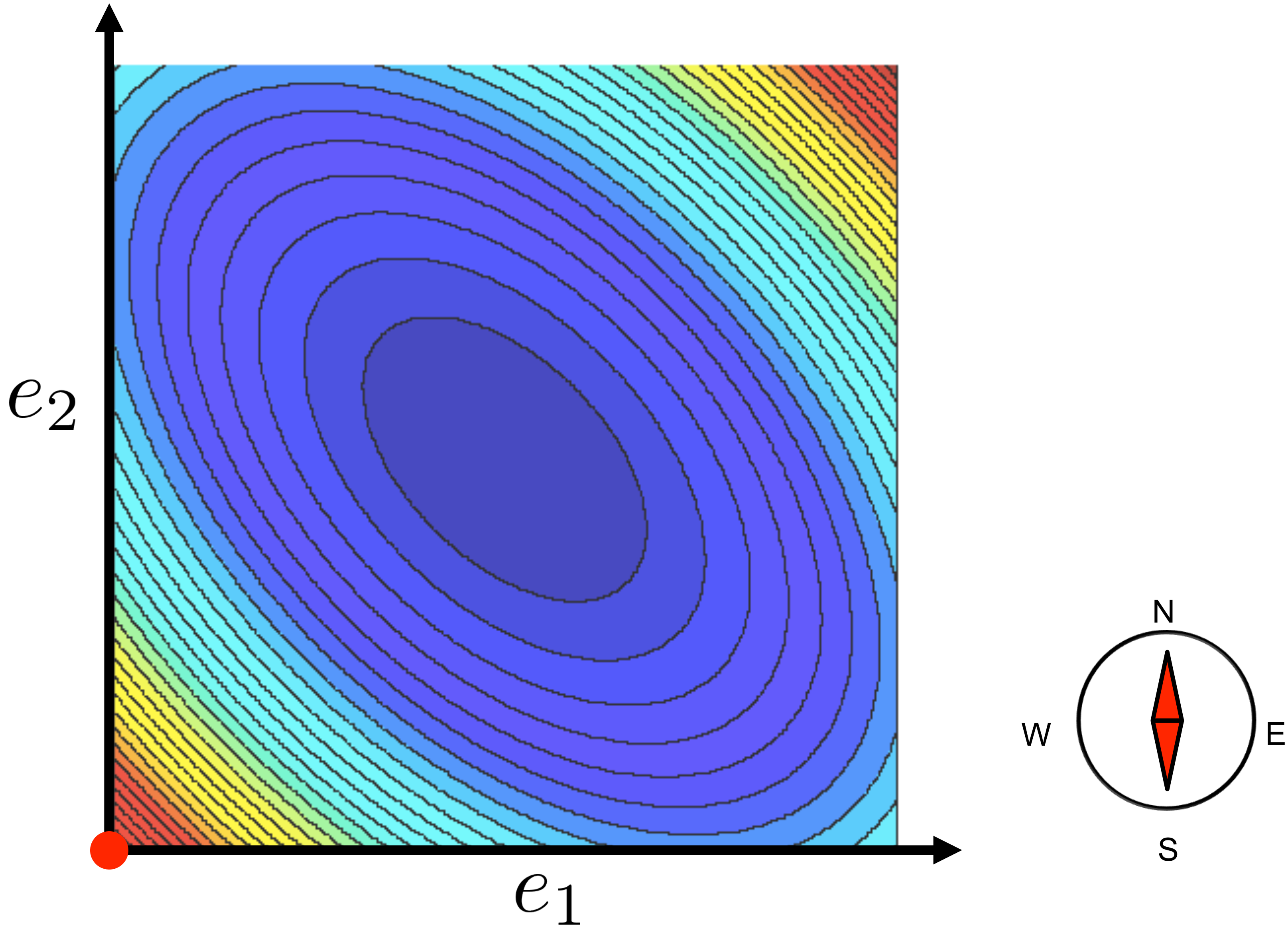
D. Needell and J. Tropp. **Paved with good intentions: analyzis of a randomized block Kaczmarz method.** *Linear Algebra and its Applications* 441, pp. 199-221, 2012

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

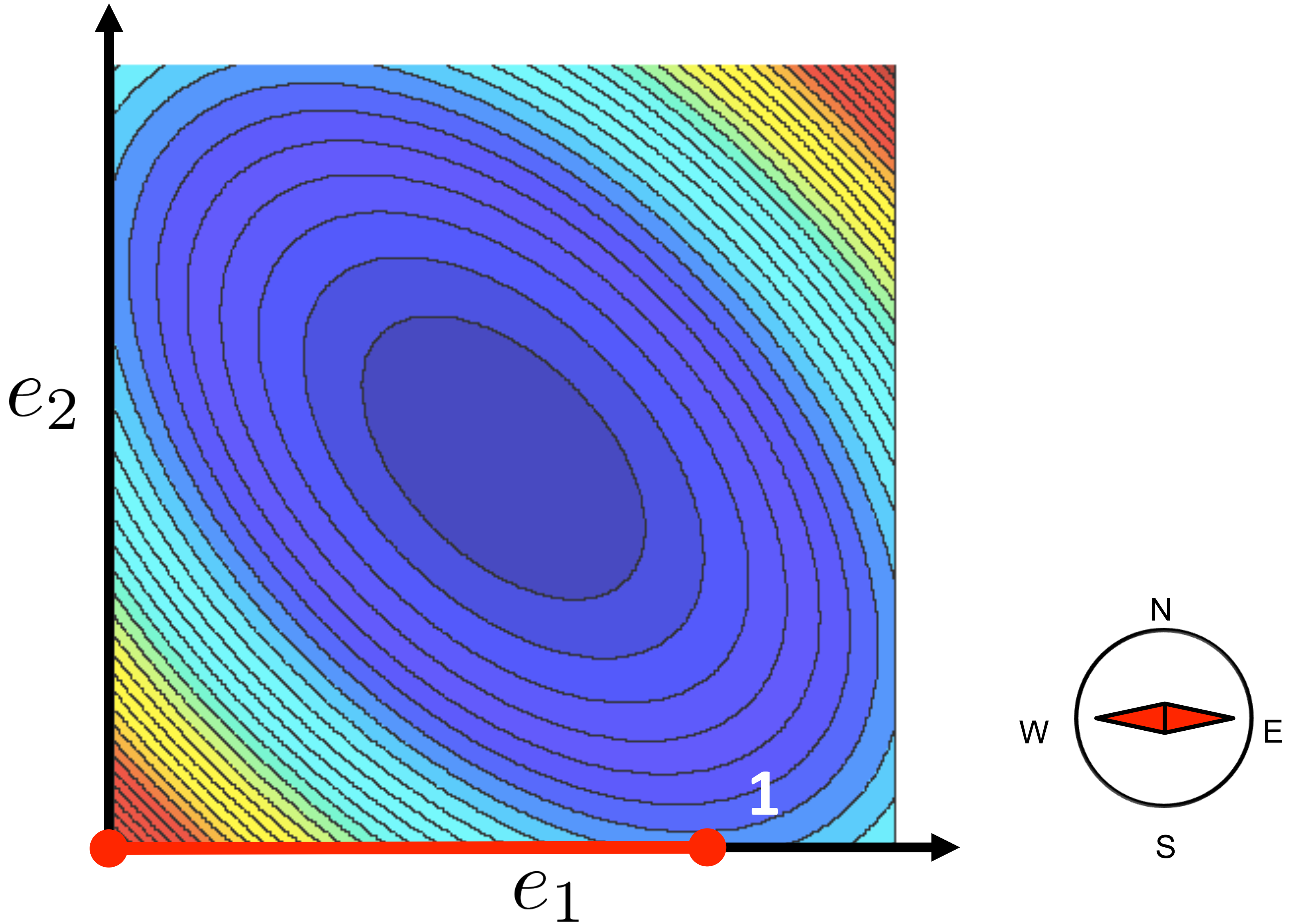
A. Ramdas. **Rows vs Columns for Linear Systems of Equations – Randomized Kaczmarz or Coordinate Descent?** *arXiv:1406.5295*, 2014

Special Case: Randomized Coordinate Descent

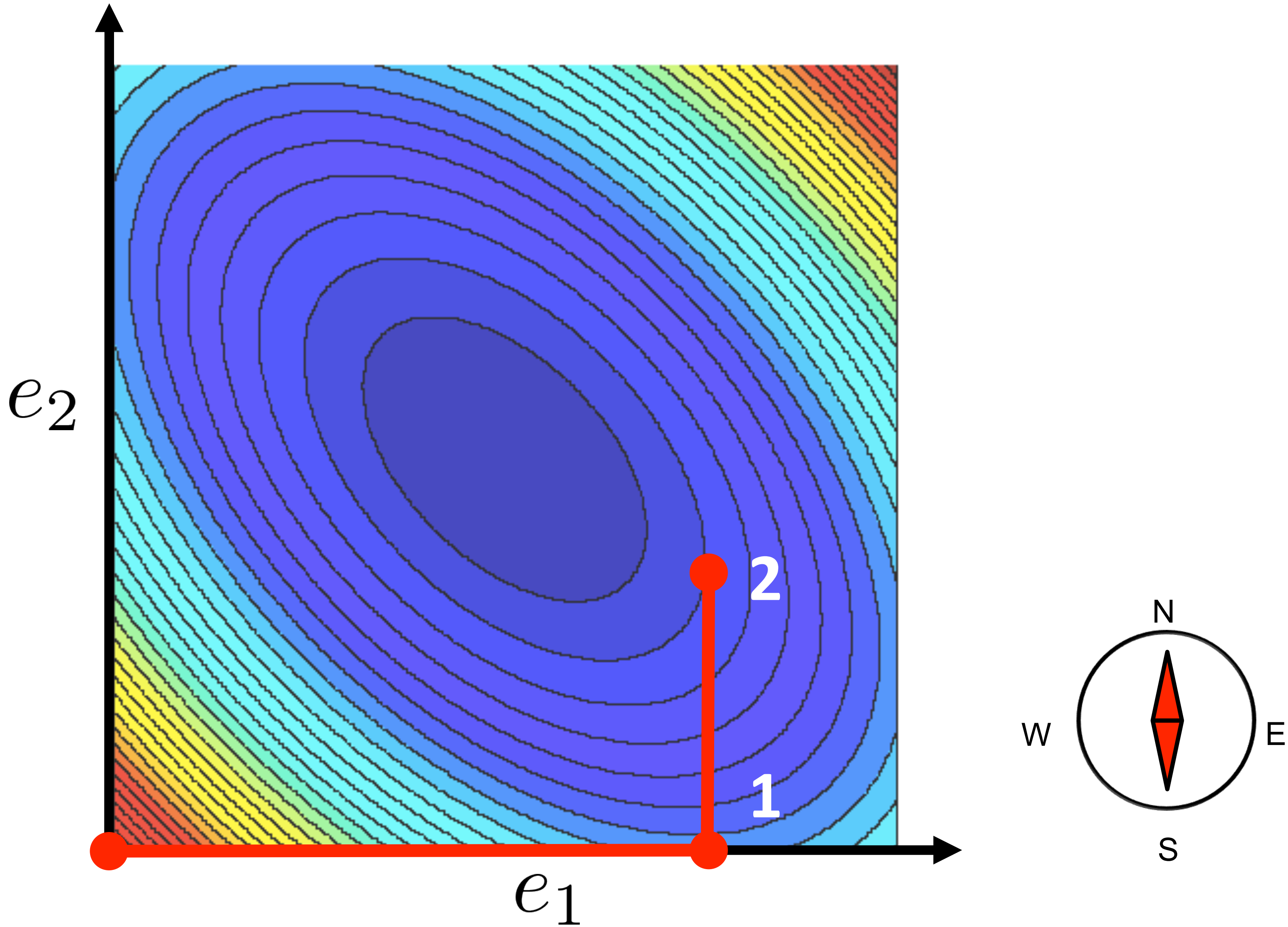
Randomized Coordinate Descent in 2D



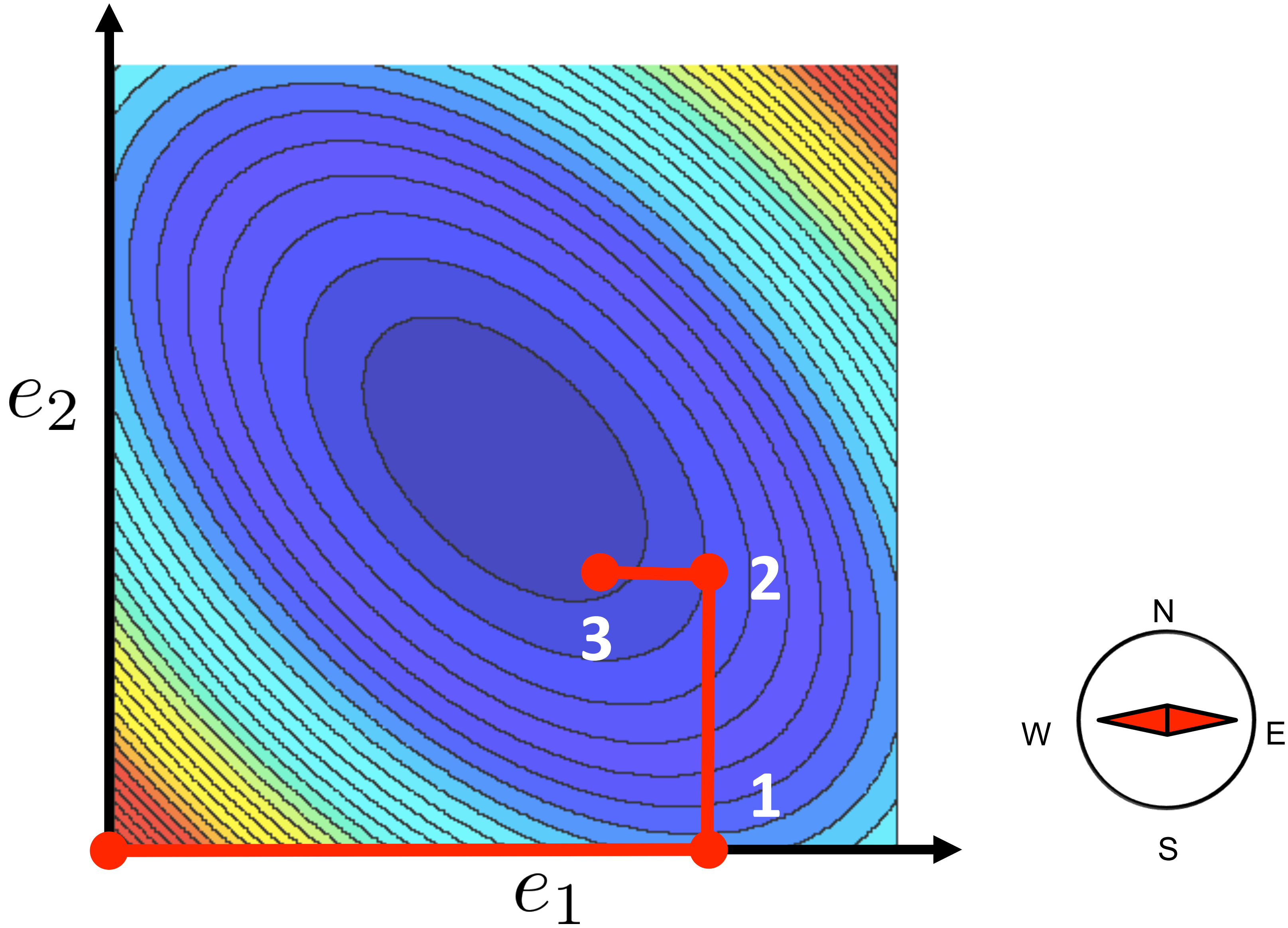
Randomized Coordinate Descent in 2D



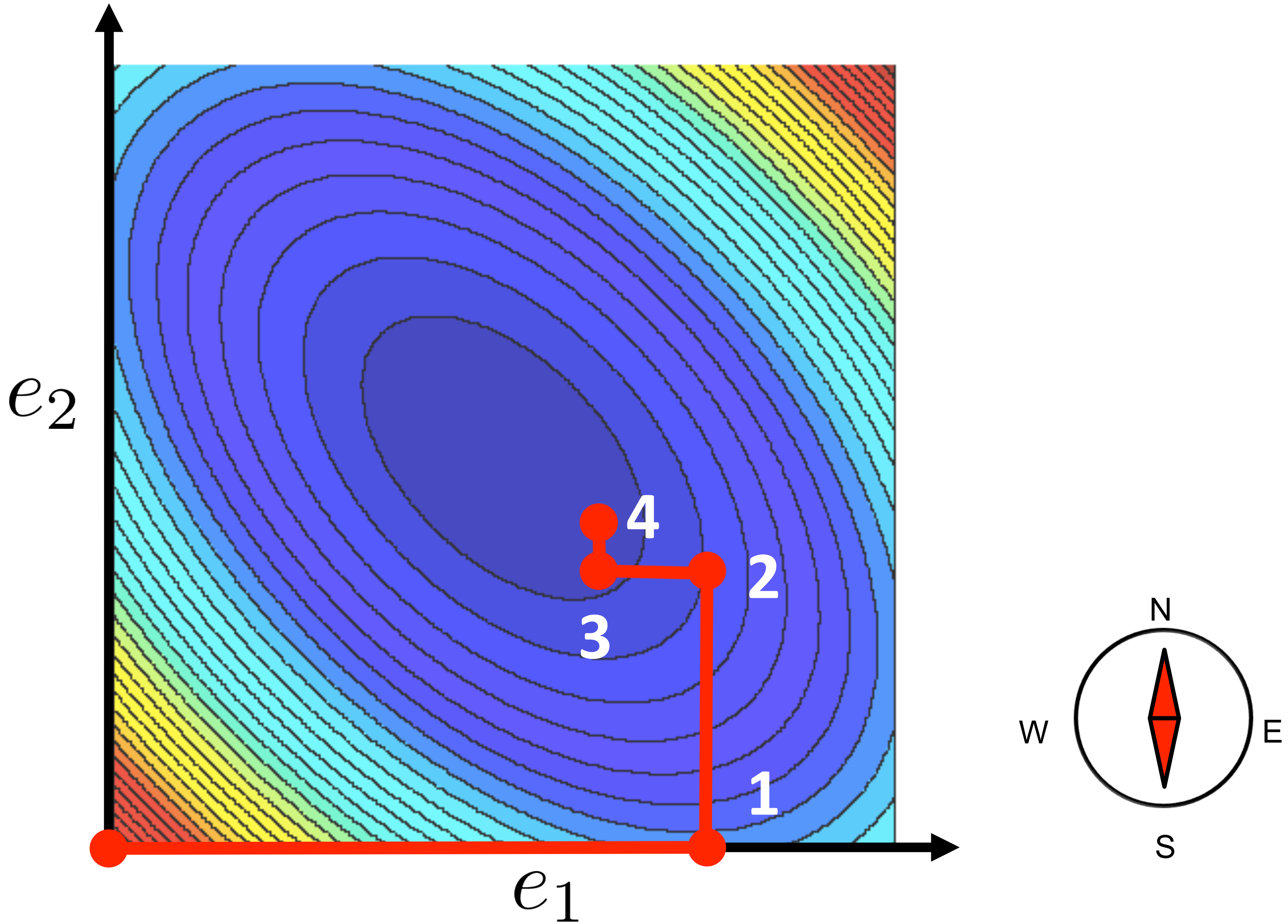
Randomized Coordinate Descent in 2D



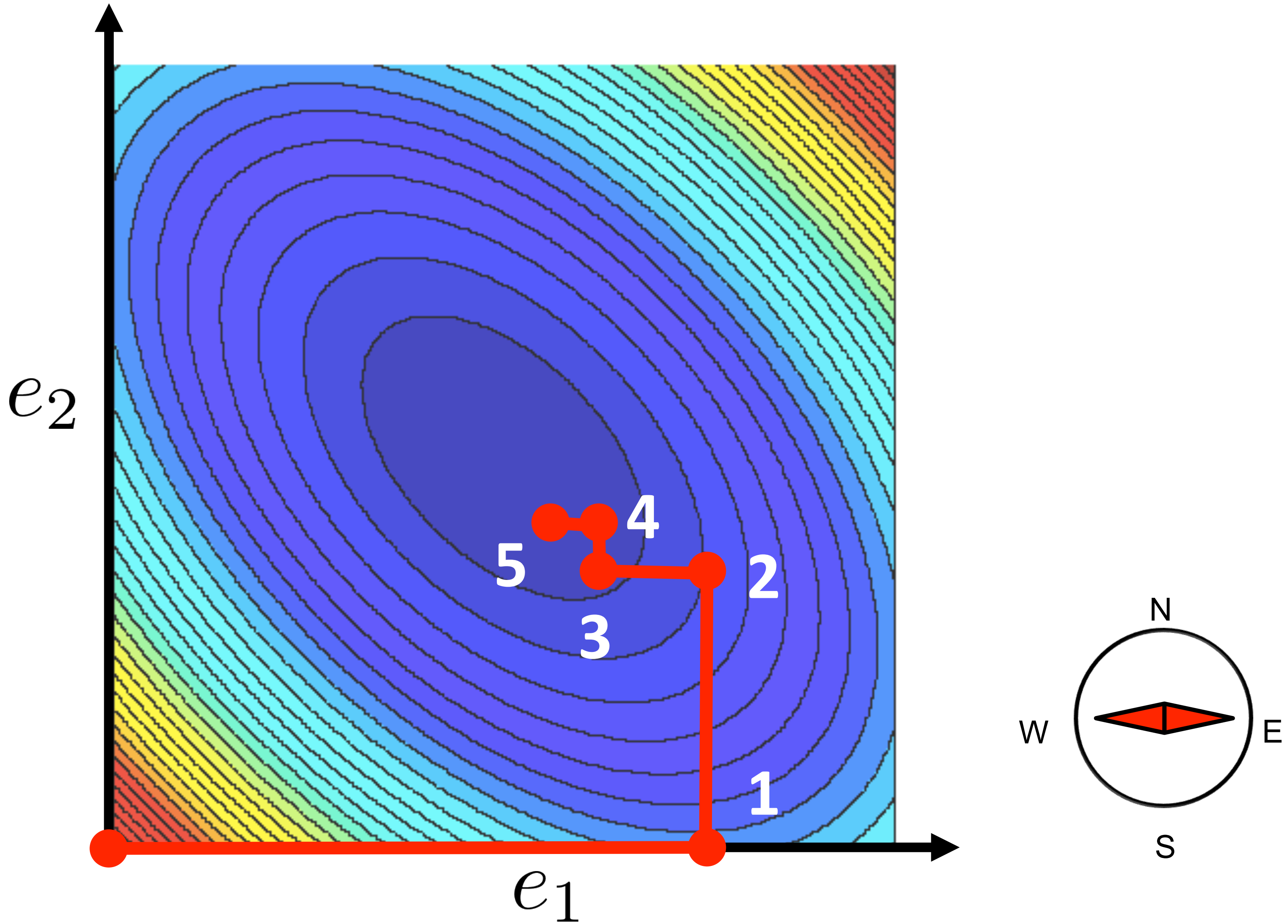
Randomized Coordinate Descent in 2D



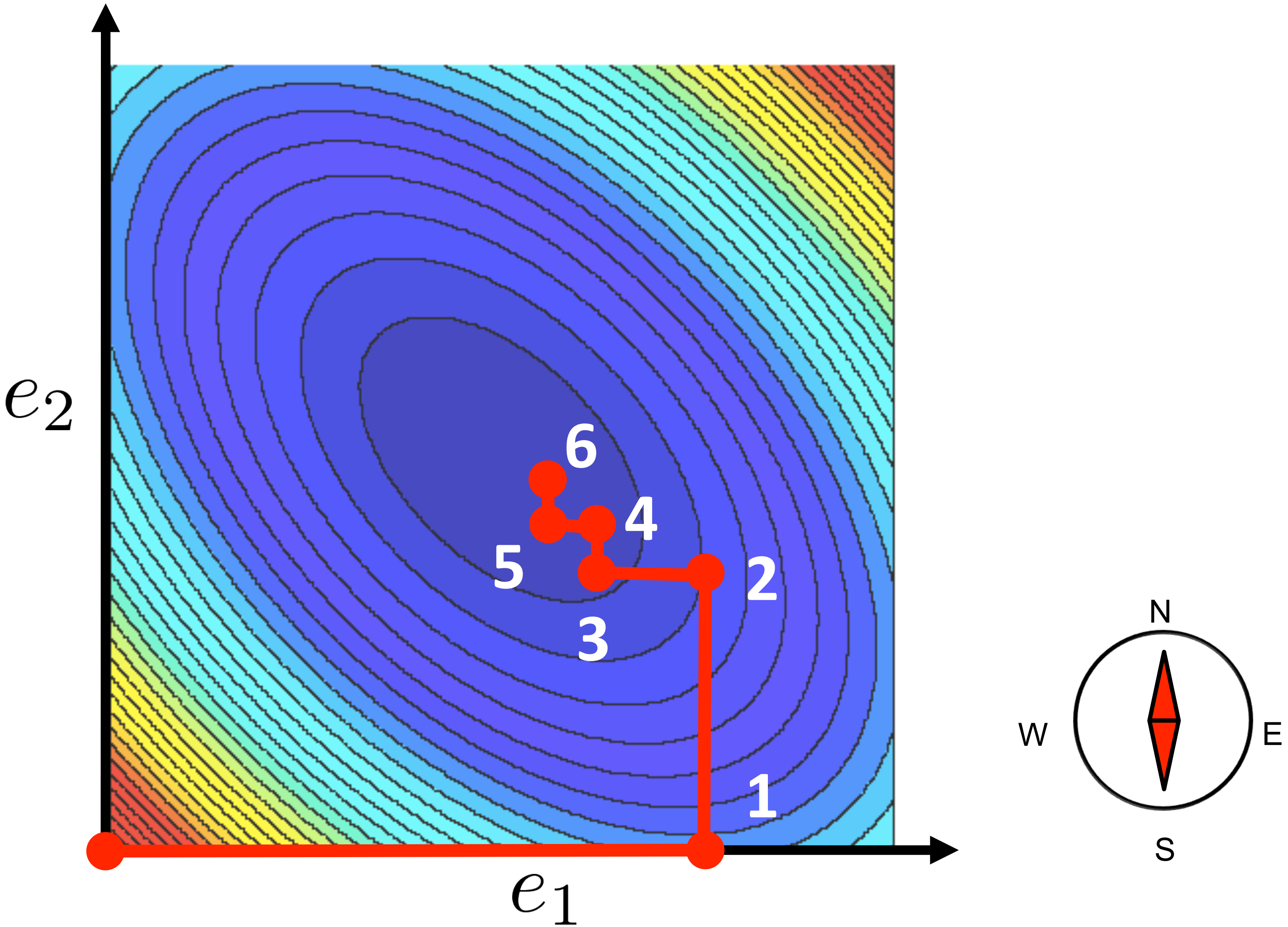
Randomized Coordinate Descent in 2D



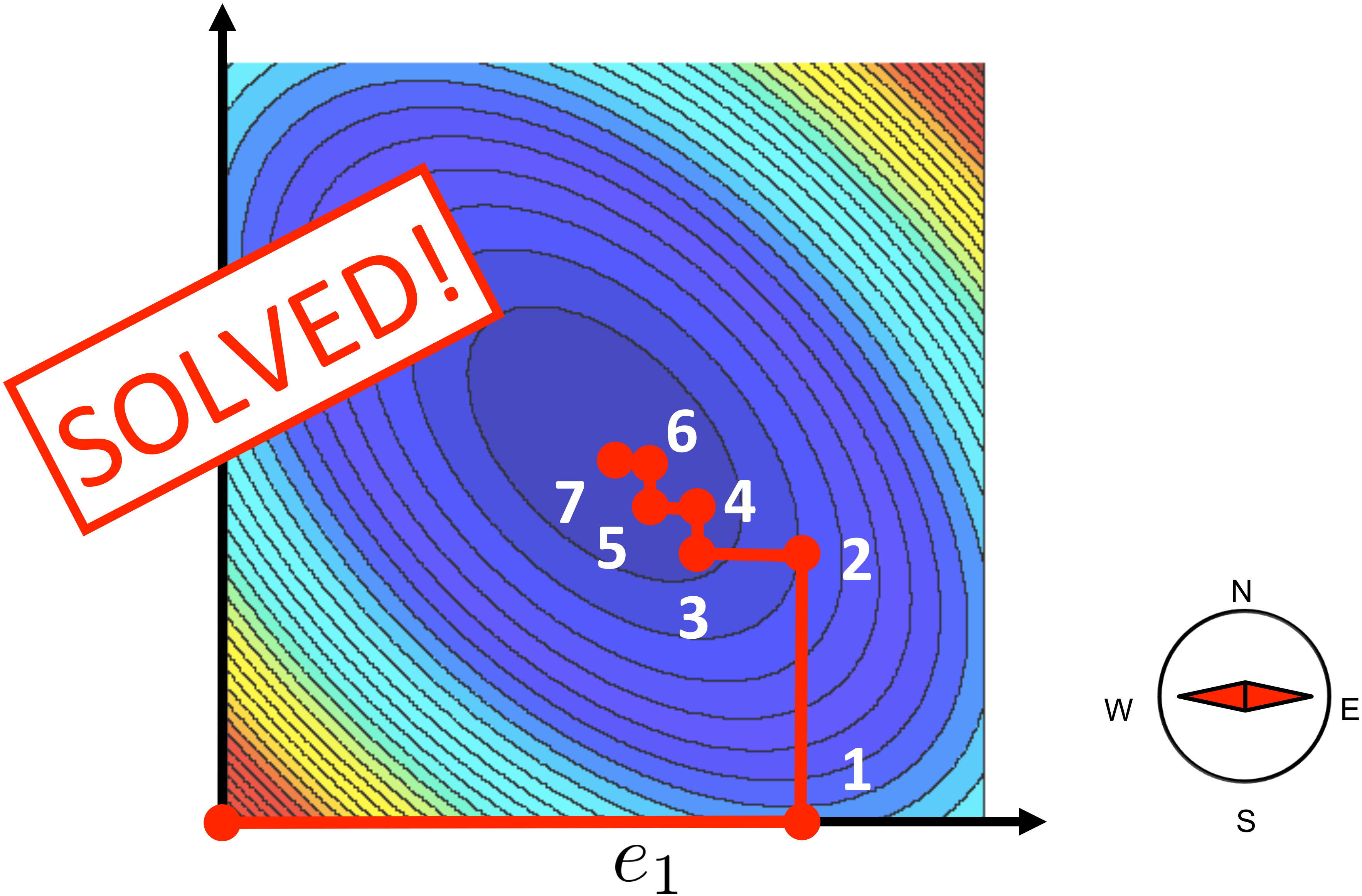
Randomized Coordinate Descent in 2D



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)

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

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

Assume: Positive definite

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

$B = A \qquad S = e^i = (0, \ldots, 0, 1, 0, \ldots, 0)$ 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 $p_i = \frac{A_{ii}}{\mathbf{Tr}(A)}$

RCD: Derivation and Rate

General Method

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

Special Choice of Parameters

$$B = A$$

$$\mathbf{P}(S = e^i) = p_i \quad S = e^i$$

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

Complexity Rate

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

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

RCD uses "Exact Line Search"

Recall Viewpoint 2 ("Constrain and Approximate"):

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

In RCD we have:

$B = A \quad S = e^i$

Observation:

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

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

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

Insight:

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*:

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

$$f(x) = \frac{1}{2} x^T A x - b^T x \quad \Rightarrow$$

$$L_i = A_{ii} \qquad \nabla_i f(x) = (A_{i:})^T x - b_i$$

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

Nesterov's RCD method:

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

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

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

Assume: Positive definite

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

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

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

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

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

RN: Derivation

General Method

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

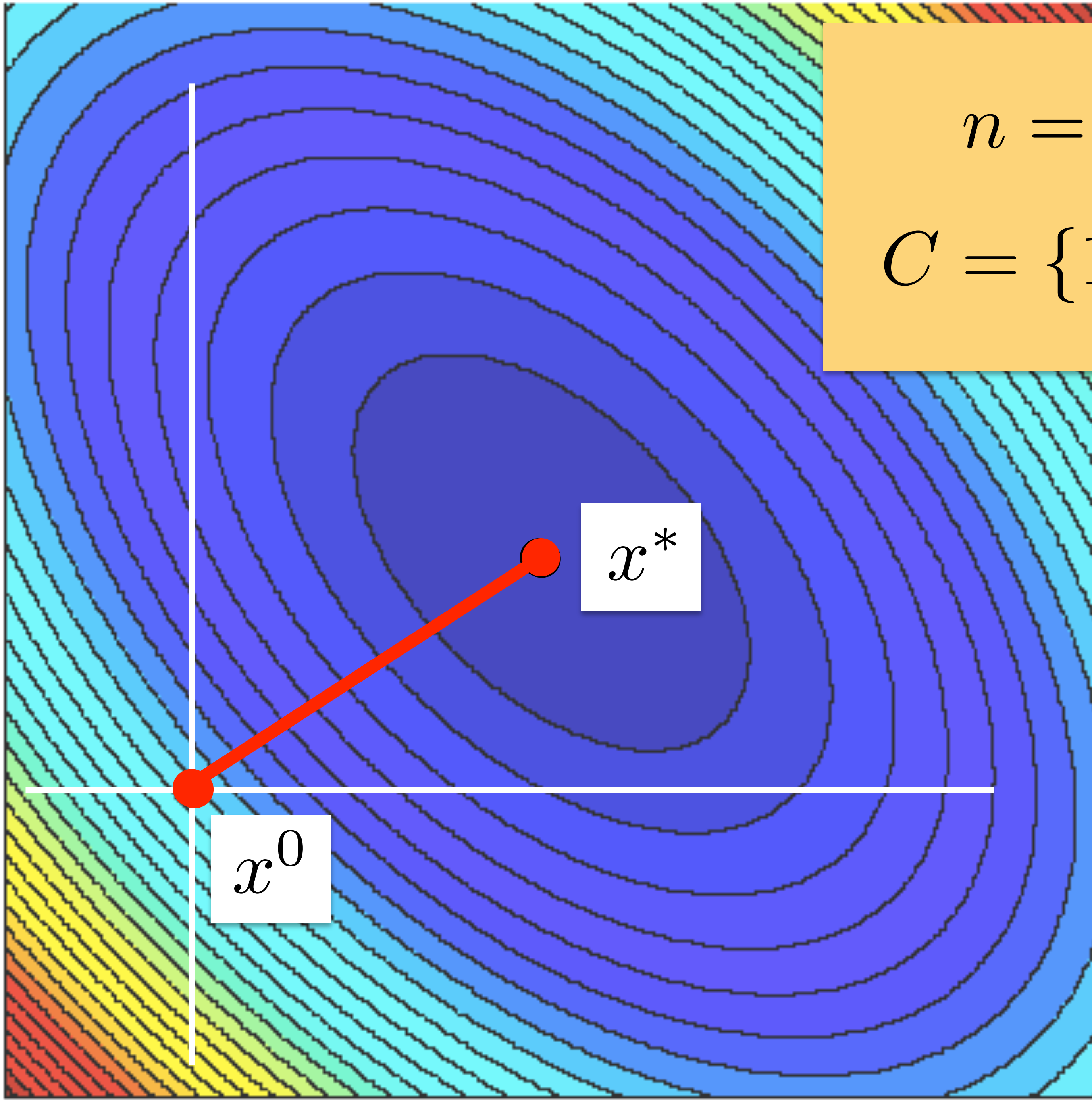
Special Choice of Parameters $B = A$

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

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

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

Complexity Rate Will talk about this more later in the "curvature" part



$n = 2$

$C = \{1, 2\}$

Special Case: Gaussian Descent

Gaussian Descent

General Method

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

Special Choice of Parameters

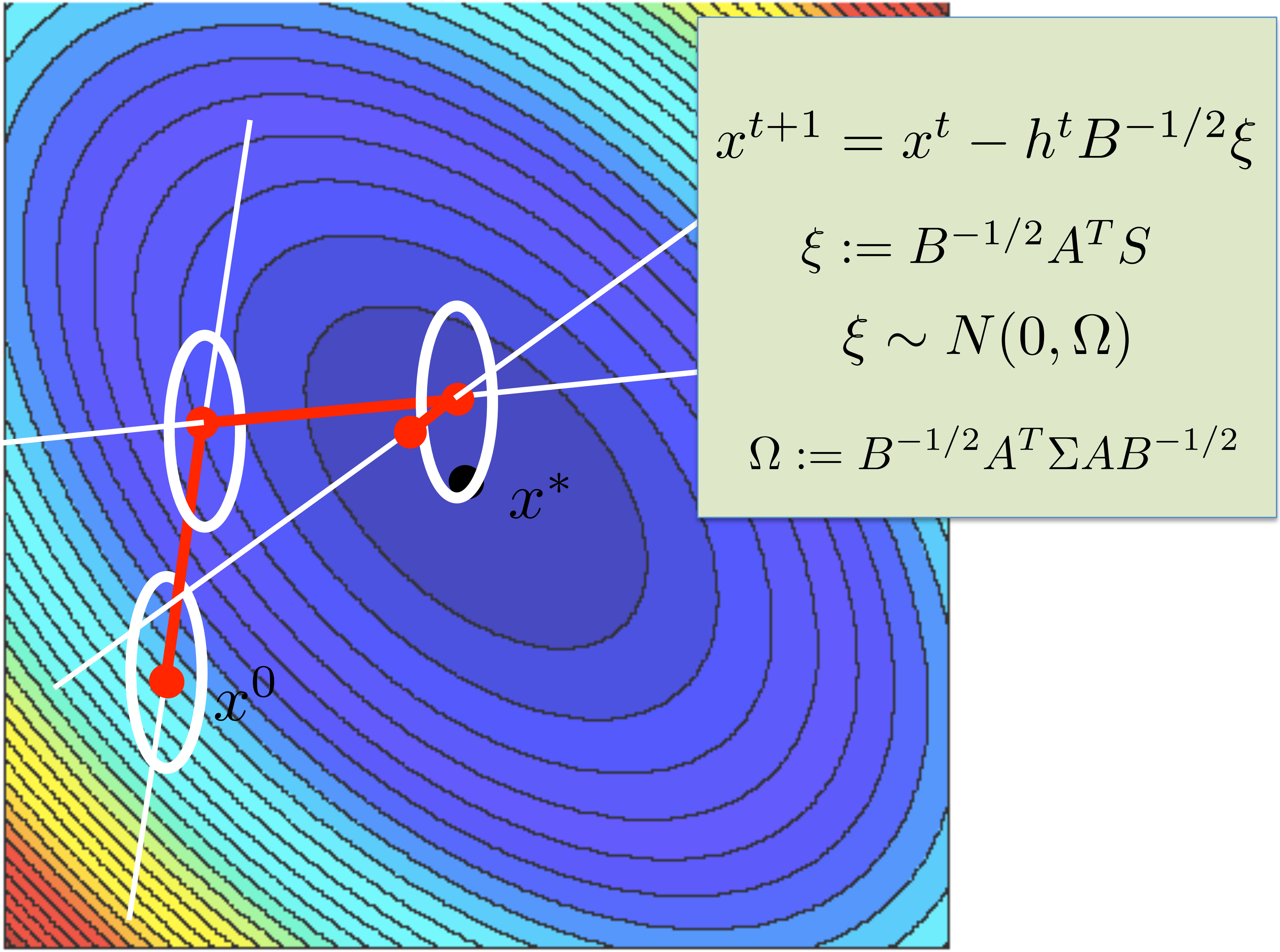
$$S \sim N(0, \Sigma)$$

Positive definite covariance matrix

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

Complexity Rate

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



$x^{t+1} = x^t - h^t B^{-1/2}\xi$
$\xi := B^{-1/2}A^T S$
$\xi \sim N(0,\Omega)$
$\Omega := B^{-1/2}A^T\Sigma A B^{-1/2}$
x^*
x^0

Gaussian Descent: The Rate

XY and *YX* have the same spectrum

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

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

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

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

Gaussian Descent: The Rate

Lemma [GR'15]

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

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

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

Gaussian Descent: Further Reading

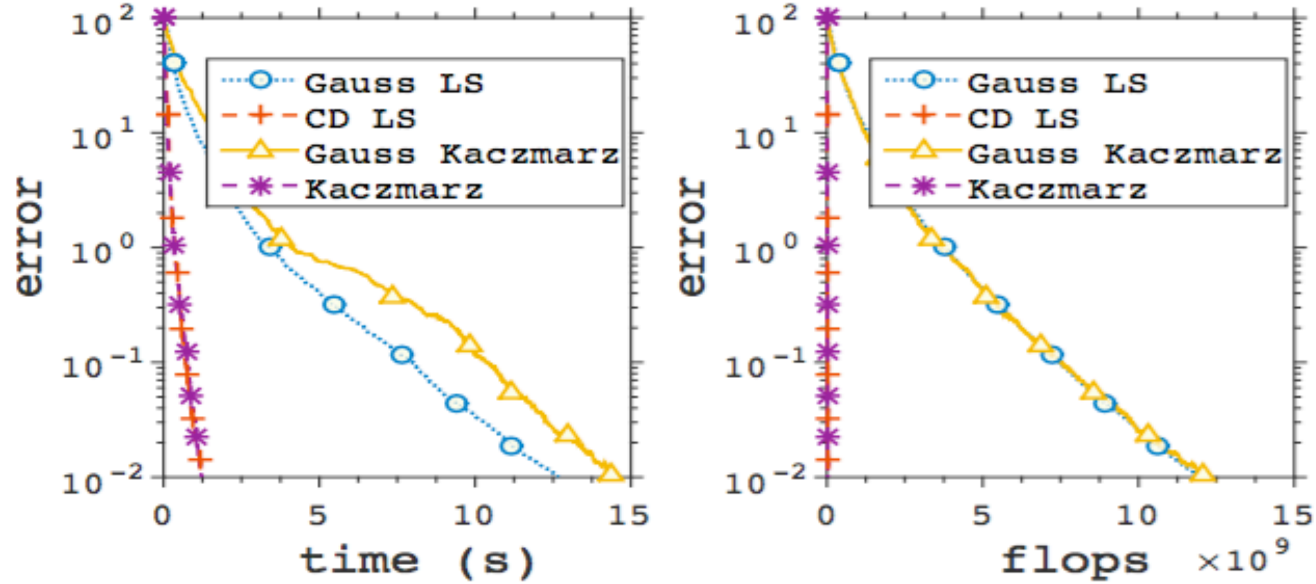
Yurii Nesterov. **Random gradient-free minimization of convex functions.** CORE Discussion Paper # 2011/1, 2011

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

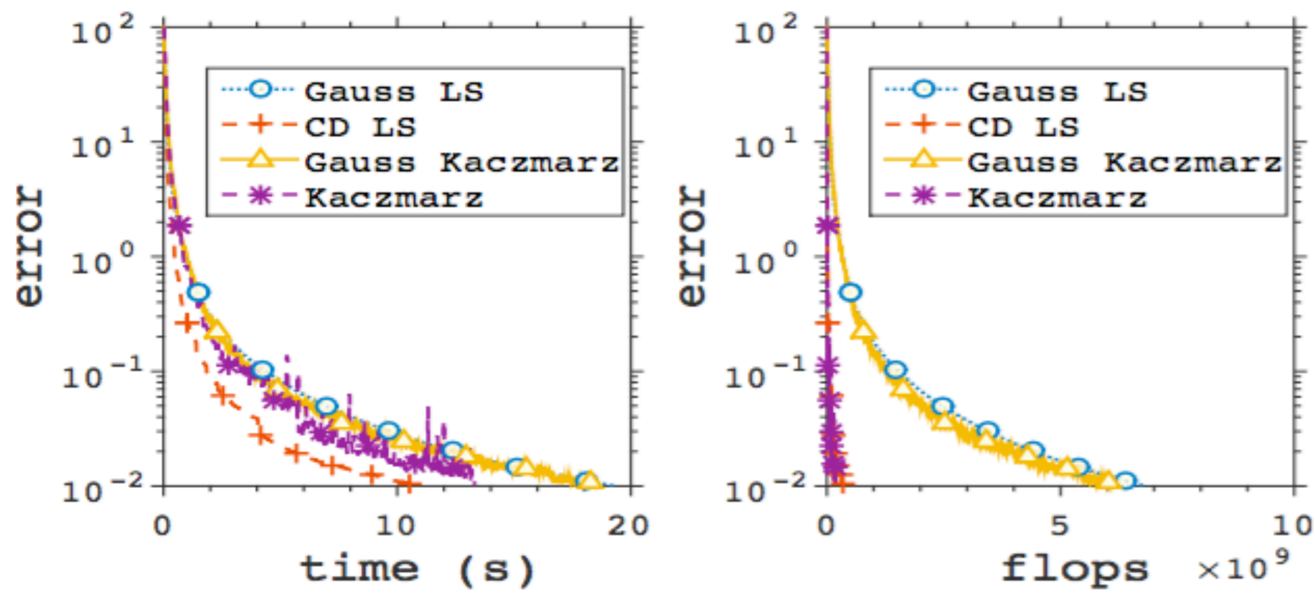
S. U. Stitch. **Convex optimization with random pursuit.** PhD Thesis, ETH Zurich, 2014

Experiments

Synthetic data

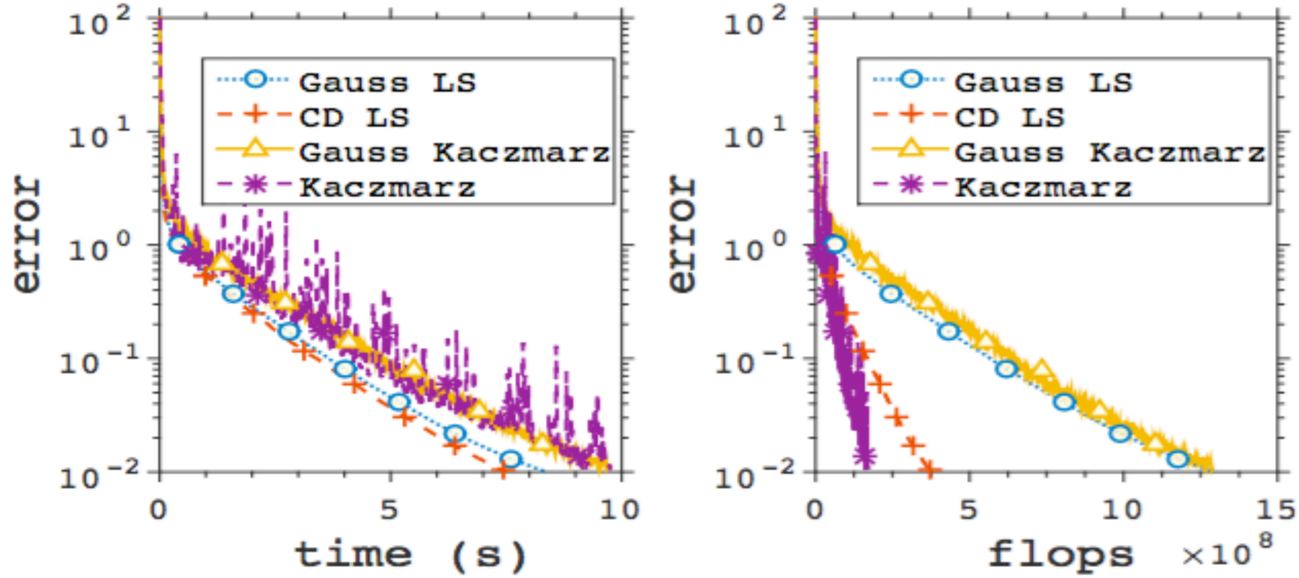


(a) **rand** (m = 1,000; n = 500)

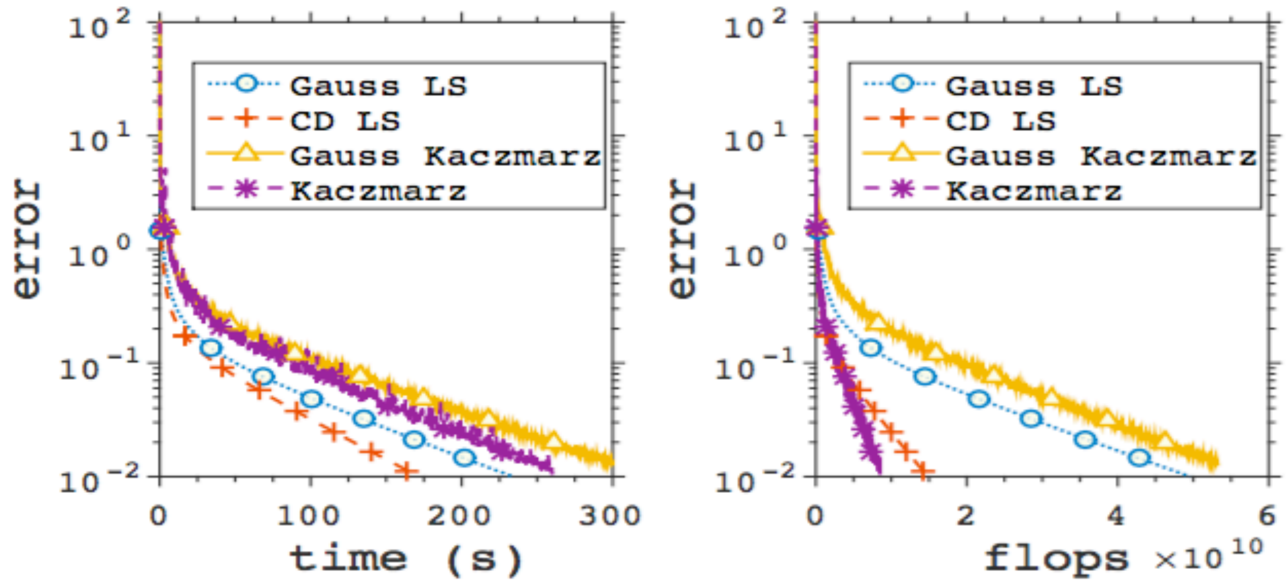


(b) **sprandn** (m = 1,000; n = 500)

Real data (Matrix Market)



(a) illc1033 (m = 1,850; n = 750)



(b) well1033 (m = 1,033; n = 320)

Importance Sampling

Importance Sampling

Assume that *S* is discrete:

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

Question

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

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

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

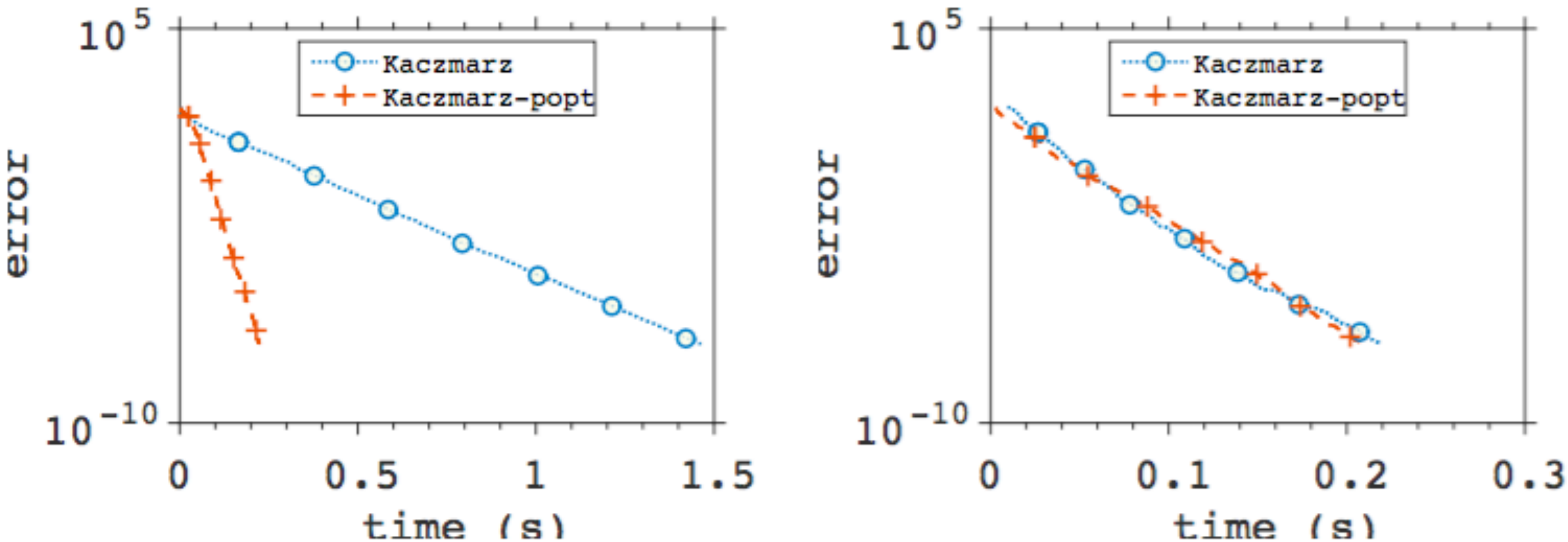
$$\begin{aligned} \max_{p,t} \quad & t \\ \text{subject to} \quad & \sum_{i=1}^r p_i \left(V_i (V_i^T V_i)^\dagger V_i^T \right) \succeq t \cdot I, \\ & p \geq 0, \quad \sum_{i=1}^r p_i = 1 \end{aligned}$$

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

RCD: Optimal Probabilities Can Lead to a Remarkable Improvement

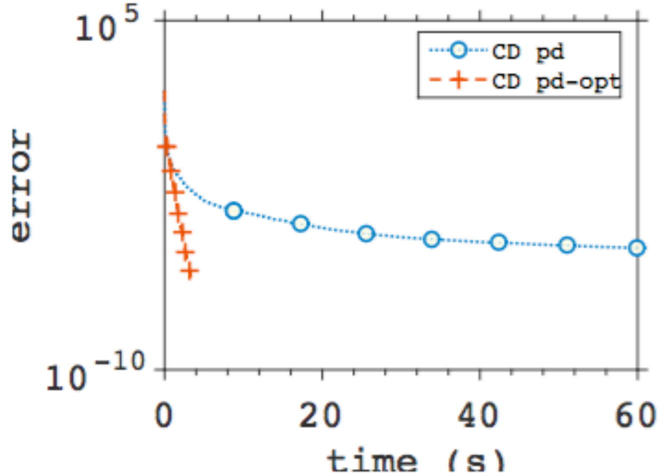
	Rate for **convenient** (standard) **probabilities**	Rate for **optimal probabilities** (solving SDP)	Lower bound on the rate
data set	ρ_c	ρ^*	$1 - 1/n$
rand(50,50)	$1 - 2 \cdot 10^{-6}$	$1 - 3.05 \cdot 10^{-6}$	$1 - 2.10^{-2}$
mushrooms-ridge	$1 - 5.86 \cdot 10^{-6}$	$1 - 7.15 \cdot 10^{-6}$	$1 - 8.93 \cdot 10^{-3}$
aloi-ridge	$1 - 2.17 \cdot \boxed{10^{-7}}$	$1 - 1.26 \cdot \boxed{10^{-4}}$	$1 - 7.81 \cdot 10^{-3}$
liver-disorders-ridge	$1 - 5.16 \cdot 10^{-4}$	$1 - 8.25 \cdot 10^{-3}$	$1 - 1.67 \cdot 10^{-1}$
covtype.binary-ridge	$1 - 7.57 \cdot \boxed{10^{-14}}$	$1 - 1.48 \cdot \boxed{10^{-6}}$	$1 - 1.85 \cdot 10^{-2}$

RK: Convenient vs Optimal

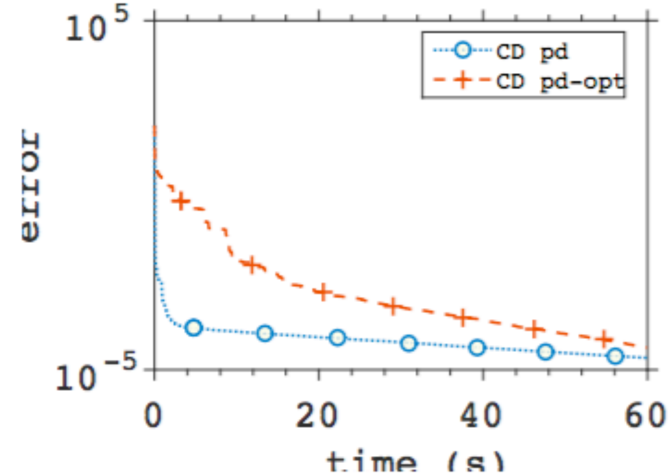


(a) liver-disorders-popt-k (b) rand(500,100)

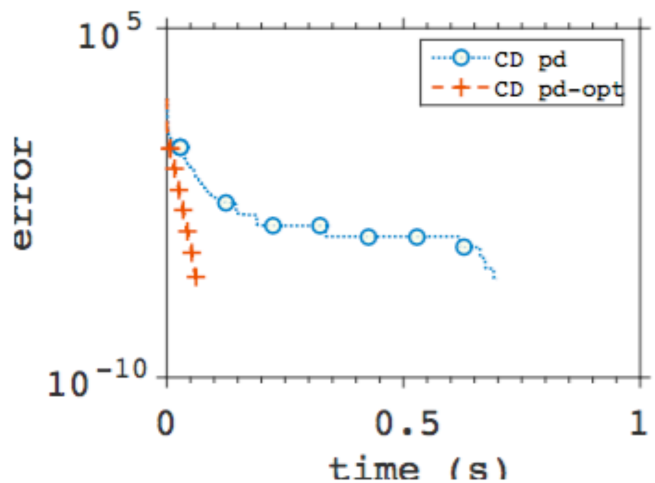
RCD: Convenient vs Optimal



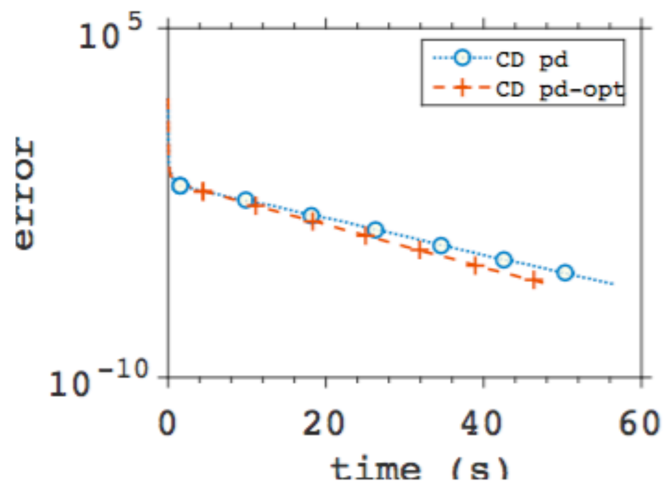
(a) aloi



(b) covtype.libsvm.binary



(c) liver-disorders-ridge



(d) mushrooms-ridge-opt

THE END