On Stochastic Algorithms in Linear Algebra, Optimization and Machine Learning

Peter Richtárik

Summer School: "Control, Information and Optimization"
Voronovo - June 12, 2018

A System of Linear Equations

m equations with n unknowns

$$A \in \mathbb{R}^{m \times n}, \quad x \in \mathbb{R}^{n}, \quad b \in \mathbb{R}^{m}$$

$$m\left\{ \overbrace{A}^{n} \right. x = b$$

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

Part I
Six Ways to Skin a Cat

[GR’15a]

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

1. Relaxation Viewpoint
"Sketch and Project"

$$\|x\|_B^2 = x^\top B x$$

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

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

S = identity matrix ➡ convergence in 1 step

$$\min_x \{\|x - x^0\| \; : \; Ax = 0\}$$

E.S. Coakley, V. Rokhlin and M. Tygert. **A Fast Randomized Algorithm for Orthogonal Projection**. *SIAM Journal on Scientific Computing* 33(2), pp. 849–868, 2011

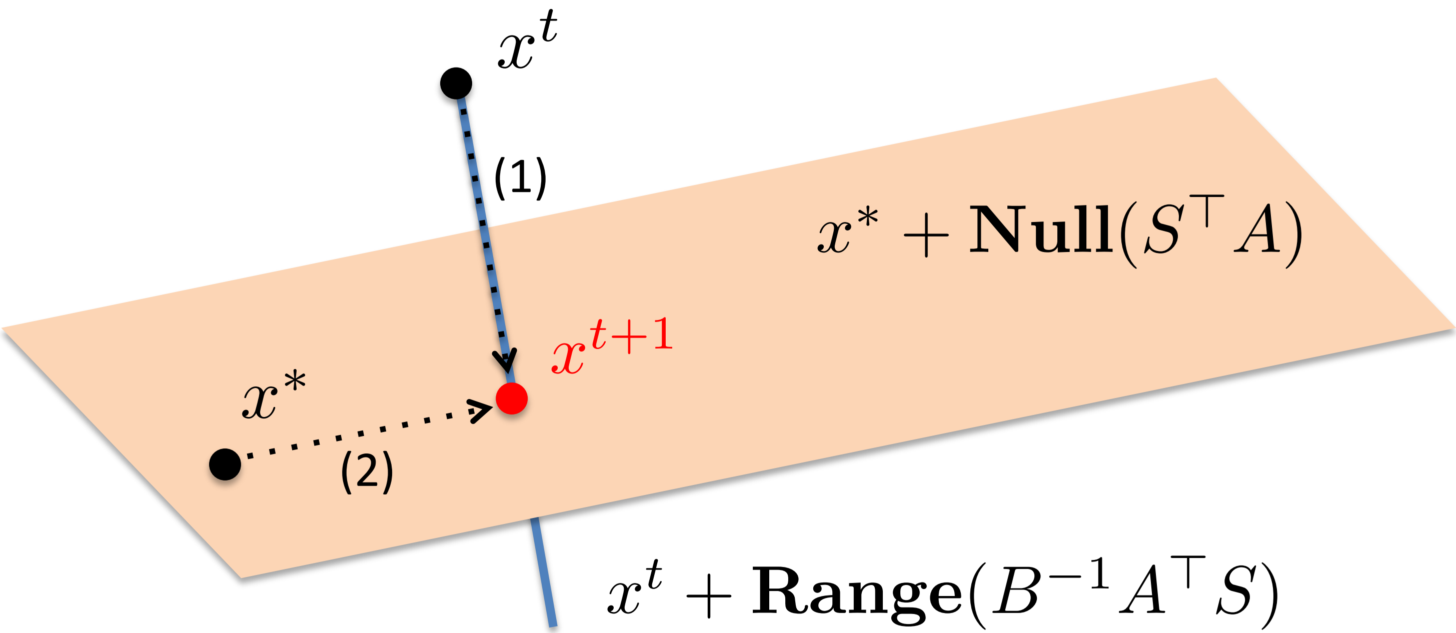
2. Approximation Viewpoint
“Constrain and Approximate”

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

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

$$\lambda \quad \text{is free}$$

3. Geometric Viewpoint
"Random Intersect"



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

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

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

4. Algebraic Viewpoint
"Random Linear Solve"

$$x^{t+1} \quad = \quad \text{solution in } x \text{ of the linear system}$$

$$S^\top A x = S^\top b$$

$$x = x^t + B^{-1} A^\top S \lambda$$

Unknown

Unknown

5. Algebraic Viewpoint
"Random Update"

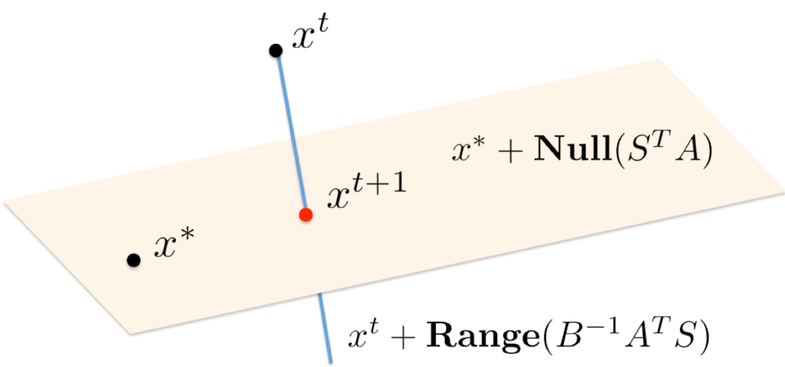
$$x^{t+1} = x^t - \overbrace{B^{-1}A^\top S(S^\top AB^{-1}A^\top S)^\dagger S^\top (Ax^t - b)}^{\text{Random Update Vector}}$$

Moore-Penrose
pseudo-inverse

6. Analytic Viewpoint
"Random Fixed Point"

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

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



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

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

Part II
Stochastic Reformulations

P.R. and Martin Takáč
Stochastic Reformulations of Linear Systems: Algorithms and Convergence Theory
arXiv:1706.01108, 2017

Stochastic Reformulations of Linear Systems

$n \times n$ pos def

distribution over $m \times q$ matrices

$$Ax = b \xrightarrow{B,\mathcal{D}}$$

1. Stochastic Optimization
2. Stochastic Linear System
3. Stochastic Fixed Point
4. Probabilistic Intersection

Example: $B =$ identity

$\mathcal{D} =$ uniform over $e_1, \ldots, e_m$ (unit basis vectors in $\mathbb{R}^m$)

Theorem

a) These 4 problems have the same solution sets

b) Necessary & sufficient conditions for the solution set to be equal to $\{x \;:\; Ax = b\}$

Reformulation 1: Stochastic Optimization

$$\text{Minimize } f(x) \overset{\text{def}}{=} \mathbf{E}_{S\sim\mathcal{D}}[f_S(x)]$$

$$f_S(x) = \frac{1}{2}\|x - \Pi^B_{\mathcal{L}_S}(x)\|_B^2 = \frac{1}{2}(Ax-b)^\top H(Ax-b)$$

$$\mathcal{L}_S = \{x \;:\; S^\top Ax = S^\top b\}$$

$$H = S(S^\top AB^{-1}A^\top S)^\dagger S^\top$$

Reformulation 2: Stochastic Linear System

Instead of $Ax = b$ we solve the preconditioned system:

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

$$\text{Solve } \underbrace{B^{-1}A^\top \mathbf{E}_{S\sim\mathcal{D}}[H]}_{\text{preconditioner}} Ax = B^{-1}A^\top \mathbf{E}_{S\sim\mathcal{D}}[H] b$$

Instead of $B^{-1}A^\top \mathbf{E}[H] A$ we have access to $B^{-1}A^\top H A$

Unbiased estimate of the preconditioner

Reformulation 3:
Stochastic Fixed Point Problem

$$\text{Solve } x = \mathbf{E}_{S \sim \mathcal{D}} \left[\Pi^B_{\mathcal{L}_S}(x) \right]$$

Projection in B-norm onto $\mathcal{L}_S = \{x \;:\; S^\top A x = S^\top b\}$

Reformulation 4:
Probabilistic Intersection Problem

$$\text{Find } x \in \mathbb{R}^n \text{ such that } \mathbf{P}(x \in \mathcal{L}_S) = 1$$

$$\mathcal{L}_S = \{x \ : \ S^\top A x = S^\top b\}$$

Sketched system

S discrete $\Rightarrow$ $\{x \ : \ \mathbf{P}(x \in \mathcal{L}_S) = 1\} = \bigcap_S \mathcal{L}_S$

Part III
Randomized Algorithms

Viewpoint 1:
Stochastic Optimization

Stochastic Gradient Descent

constant stepsize

$S \sim \mathcal{D}$

$$x^{t+1} = x^t - \omega \nabla f_S(x^t)$$

"stochastic gradient"

A key method in machine learning

Stochastic "Newton" Descent

Constant stepsize

"stochastic gradient"

$$x^{t+1} = x^t - \omega(\nabla^2 f_S)^{\dagger_B} \nabla f_S(x^t)$$

B- pseudoinverse of the "stochastic Hessian"

$S \sim \mathcal{D}$

Stochastic Proximal Point Method

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

$$x^{t+1} = \arg\min_{x \in \mathbb{R}^n} \left\{ f_S(x) + \frac{\omega - 1}{2\omega} \|x - x^t\|_B^2 \right\}$$

Viewpoint 3: Stochastic Fixed Point Method

Stochastic Fixed Point Method

Stochastic fixed point mapping

$$x^{t+1} = \omega \Pi^B_{\mathcal{L}_S}(x^t) + (1-\omega)x^t$$

Relaxation parameter

$S \sim \mathcal{D}$

Part IV
Complexity

Basic Method

Basic Method: Complexity

$$\mathbf{E}[U^\top B^{1/2}(x^t - x^*)] = (I - \omega\Lambda)^t U^\top B^{1/2}(x^0 - x^*)$$

stepsize / relaxation parameter

$$W = B^{-1/2}A^\top \mathbf{E}_{S\sim\mathcal{D}}[H]AB^{-1/2} = U\Lambda U^\top$$

$$H = S(S^\top AB^{-1}A^\top S)^\dagger S^\top$$

Basic Method: Complexity

Convergence of Expected Iterates

$$t \geq \frac{1}{\lambda_{\min}^{+}} \log\left(\frac{1}{\epsilon}\right) \quad \xrightarrow{\omega = 1} \quad \|\mathbf{E}[x^t - x^*]\|_B^2 \leq \epsilon$$

$$t \geq \frac{\lambda_{\max}}{\lambda_{\min}^{+}} \log\left(\frac{1}{\epsilon}\right) \quad \xrightarrow{\omega = 1/\lambda_{\max}} \quad \|\mathbf{E}[x^t - x^*]\|_B^2 \leq \epsilon$$

L2 Convergence

$$t \geq \frac{1}{\lambda_{\min}^{+}} \log\left(\frac{1}{\epsilon}\right) \quad \xrightarrow{\omega = 1} \quad \mathbf{E}\left[\|x^t - x^*\|_B^2\right] \leq \epsilon$$

Parallel Method

Parallel Method

"Run 1 step of the basic method from x^t
several times independently,
and average the results."

i.i.d.

$$x^{t+1} = \frac{1}{\tau}\sum_{i=1}^{\tau} \underbrace{\phi_\omega(x^t, S_i^t)}_{\text{One step of the basic method from } x^t}$$

Parallel Method: Complexity

L2 Convergence

$$\tau = 1$$

$$t \geq \frac{1}{\lambda_{\min}^{+}} \log\left(\frac{1}{\epsilon}\right)$$

or

$$\tau = +\infty$$

$$t \geq \frac{\lambda_{\max}}{\lambda_{\min}^{+}} \log\left(\frac{1}{\epsilon}\right)$$

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

Accelerated Method

Accelerated Method

Acceleration parameter (between 1 and 2)

$S^t, S^{t-1} \sim \mathcal{D}$ (independent)

$$x^{t+1} = \gamma \underbrace{\phi_\omega(x^t, S^t)}_{\text{One step of the basic method from } x^t} + (1-\gamma)\underbrace{\phi_\omega(x^{t-1}, S^{t-1})}_{\text{One step of the basic method from } x^{t-1}}$$

Accelerated Method: Complexity

Convergence of Iterates

$$t \geq \sqrt{\frac{\lambda_{\max}}{\lambda_{\min}^{+}}} \log\left(\frac{1}{\epsilon}\right) \Rightarrow \|\mathbf{E}[x^t - x^*]\|_B^2 \leq \epsilon$$

Basic Method depends on $\frac{\lambda_{\max}}{\lambda_{\min}^{+}}$!

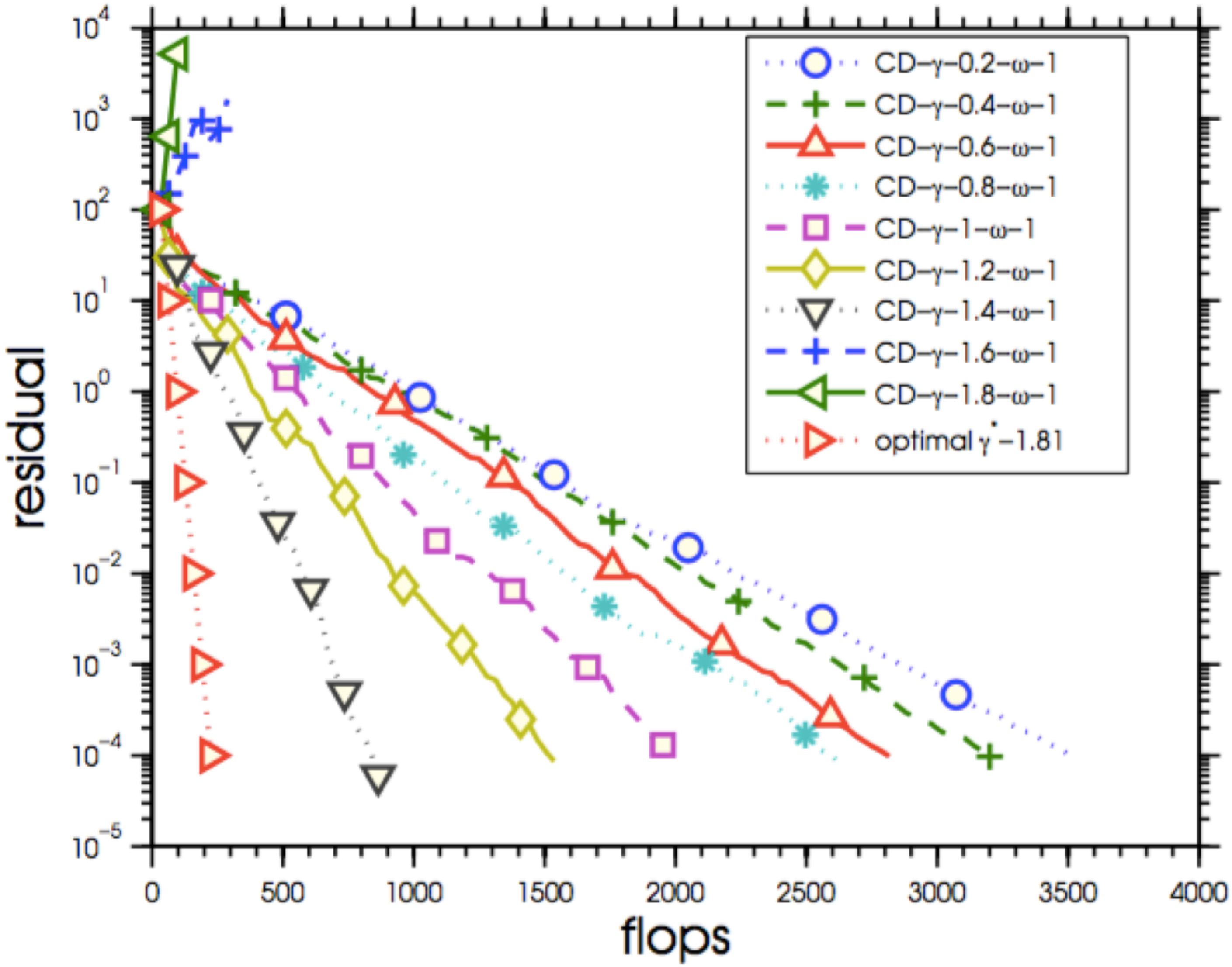
Detailed Complexity Results

Alg.	ω	τ	γ	Quantity	Rate	Complexity	Theorem
1	1	-	-	$\|\mathrm{E}\left[x_k - x_*\right]\|_{\mathbf{B}}^2$	$(1-\lambda_{\min}^+)^{2k}$	$1/\lambda_{\min}^+$	4.3, 4.4, 4.6
1	$1/\lambda_{\max}$	-	-	$\|\mathrm{E}\left[x_k - x_*\right]\|_{\mathbf{B}}^2$	$(1-1/\zeta)^{2k}$	ζ	4.3, 4.4, 4.6
1	$\frac{2}{\lambda_{\min}^+ + \lambda_{\max}}$	-	-	$\|\mathrm{E}\left[x_k - x_*\right]\|_{\mathbf{B}}^2$	$(1-2/(\zeta+1))^{2k}$	ζ	4.3, 4.4, 4.6
1	1	-	-	$\mathrm{E}\left[\|x_k - x_*\|_{\mathbf{B}}^2\right]$	$(1-\lambda_{\min}^+)^{k}$	$1/\lambda_{\min}^+$	4.8
1	1	-	-	$\mathrm{E}\left[f(x_k)\right]$	$(1-\lambda_{\min}^+)^{k}$	$1/\lambda_{\min}^+$	4.10
2	1	τ	-	$\mathrm{E}\left[\|x_k - x_*\|_{\mathbf{B}}^2\right]$	$\left(1-\lambda_{\min}^+\left(2-\xi(\tau)\right)\right)^k$		5.1
2	$1/\xi(\tau)$	τ	-	$\mathrm{E}\left[\|x_k - x_*\|_{\mathbf{B}}^2\right]$	$\left(1-\frac{\lambda_{\min}^+}{\xi(\tau)}\right)^k$	$\xi(\tau)/\lambda_{\min}^+$	5.1
2	$1/\lambda_{\max}$	∞	-	$\mathrm{E}\left[\|x_k - x_*\|_{\mathbf{B}}^2\right]$	$(1-1/\zeta)^{k}$	ζ	5.1
3	1	-	$\frac{2}{1+\sqrt{0.99\lambda_{\min}^+}}$	$\|\mathrm{E}\left[x_k - x_*\right]\|_{\mathbf{B}}^2$	$\left(1-\sqrt{0.99\lambda_{\min}^+}\right)^{2k}$	$\sqrt{1/\lambda_{\min}^+}$	5.3
3	$1/\lambda_{\max}$	-	$\frac{2}{1+\sqrt{0.99/\zeta}}$	$\|\mathrm{E}\left[x_k - x_*\right]\|_{\mathbf{B}}^2$	$\left(1-\sqrt{0.99/\zeta}\right)^{2k}$	$\sqrt{\zeta}$	5.3

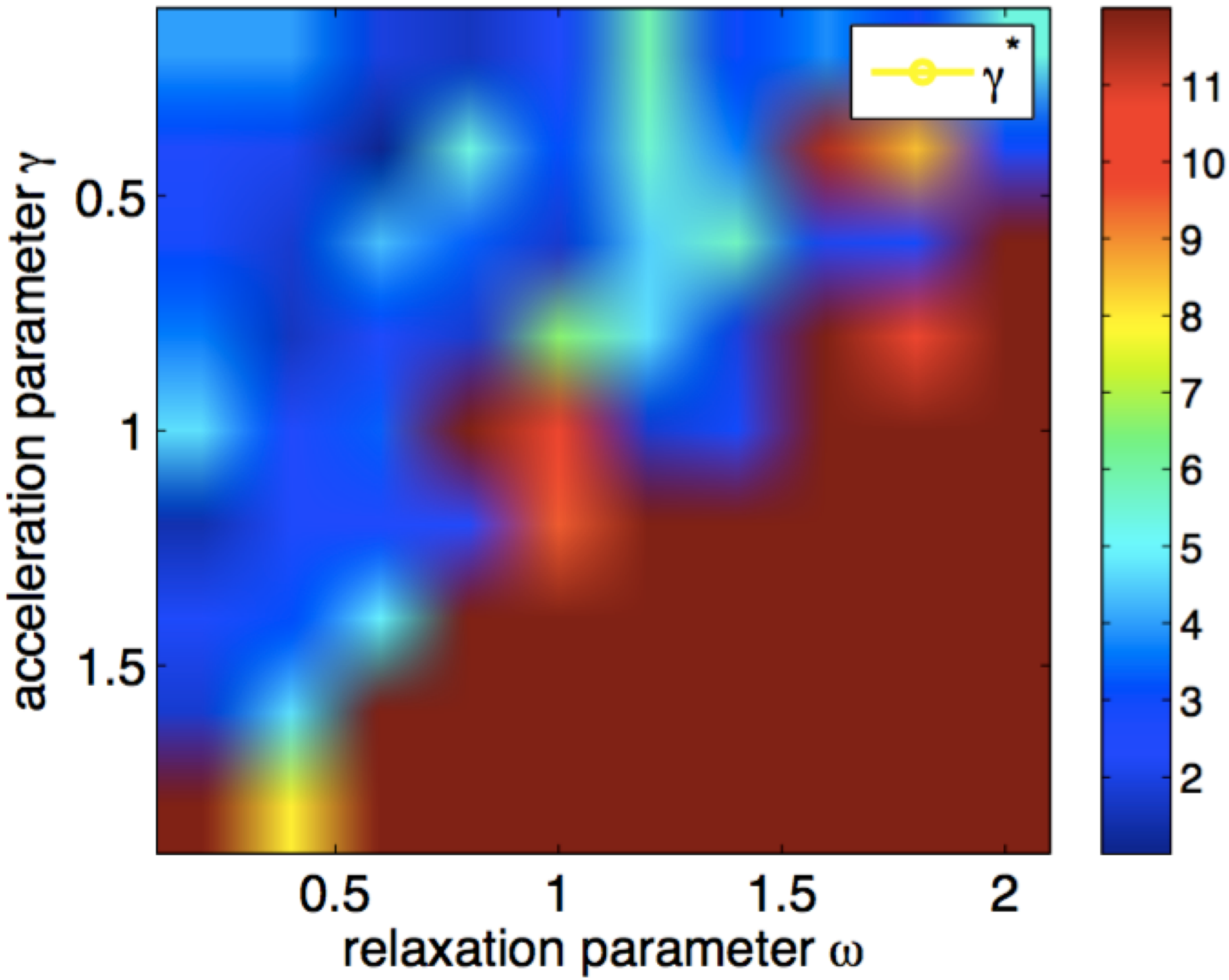
Table 1: Summary of the main complexity results. In all cases, $x_* = \Pi_{\mathcal{L}}^{\mathbf{B}}(x_0)$ (the projection of the starting point onto the solution space of the linear system). "Complexity" refers to the number of iterations needed to drive "Quantity" below some error tolerance $\epsilon > 0$ (we suppress a $\log(1/\epsilon)$ factor in all expressions in the "Complexity" column). In the table we use the following expressions: $\xi(\tau) = \frac{1}{\tau} + (1-\frac{1}{\tau})\lambda_{\max}$ and $\zeta = \lambda_{\max}/\lambda_{\min}^+$.

Part V
Experiments

Acceleration Accelerates



More Relaxation Requires More Acceleration



Part VI
Dual Viewpoint

[GR'15b]

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

Optimization Formulation

Primal Problem

$B \succ 0$

$$
\begin{aligned}
\text{minimize} \quad & P(x) := \tfrac{1}{2}\|x - c\|_B^2 \\
\text{subject to} \quad & Ax = b \\
& x \in \mathbb{R}^n
\end{aligned}
$$

$A \in \mathbb{R}^{m \times n}$

$\frac{1}{2}(x-c)^\top B(x-c)$

Dual Problem

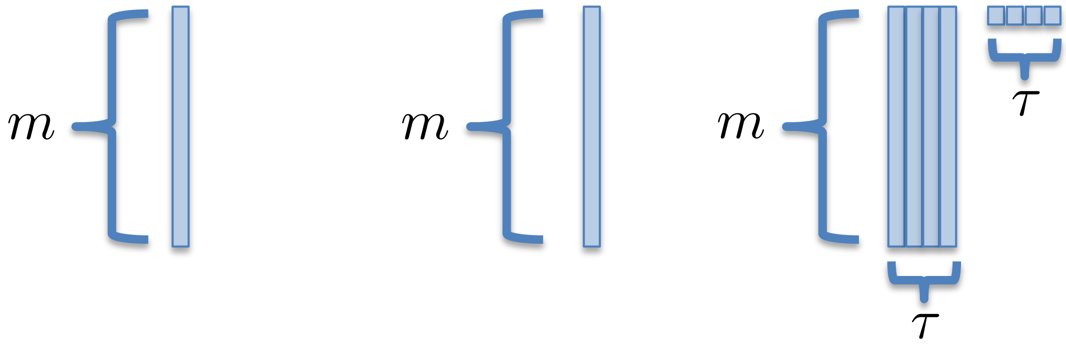
Unconstrained non-strongly concave quadratic maximization problem

$$
\begin{aligned}
\text{maximize} \quad & D(y) := (b - Ac)^\top y - \tfrac{1}{2}\|A^\top y\|_{B^{-1}}^2 \\
\text{subject to} \quad & y \in \mathbb{R}^m
\end{aligned}
$$

Stochastic Dual Ascent

A random $m \times \tau$ matrix drawn i.i.d. in each iteration $S \sim \mathcal{D}$

$$y^{t+1} = y^t + S\lambda^t$$



Moore-Penrose pseudo-inverse of a small $\tau \times \tau$ matrix

$$\lambda^t := \arg\min_{\lambda \in Q^t} \|\lambda\|_2$$

$$Q^t := \arg\max_\lambda D(y^t + S\lambda)$$

$$\lambda^t = \left(S^\top A B^{-1} A^\top S\right)^\dagger S^\top \left(b - A\left(c + B^{-1} A^\top y^t\right)\right)$$

Dual Correspondence Lemma

Lemma (GR'15b)

Affine mapping from $\mathbb{R}^m$ to $\mathbb{R}^n$

$$x(y) := c + B^{-1}A^\top y$$

(Any) dual optimal point

Primal optimal point

$$\underbrace{D(y^*) - D(y)}_{\substack{\text{Dual error} \\ \text{(in function values)}}} = \underbrace{\frac{1}{2}\|x(y) - x^*\|_B^2}_{\substack{\text{Primal error} \\ \text{(in distance)}}}$$

Primal Method = Linear Image of the Dual Method

$$x^t := x(y^t) = c + B^{-1}A^\top y^t$$

Corresponding primal iterates

Dual iterates produced by SDA

Convergence

Main Assumption

Assumption 2

The matrix

$$\mathbf{E}_{S\sim\mathcal{D}}\Big[\underbrace{S\left(S^\top A B^{-1} A^\top S\right)^\dagger S^\top}_{H}\Big]$$

is nonsingular

Complexity of SDA

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

$$U_0 = \frac{1}{2}\|x^0 - x^*\|_B^2$$

Theorem (GR'15b)

Primal iterates: $\mathbf{E}\left[\frac{1}{2}\|x^t - x^*\|_B^2\right] \leq \rho^t U_0$ GR'15a

Residual: $\mathbf{E}[\|Ax^t - b\|_B] \leq \rho^{t/2}\|A\|_B\sqrt{2 \times U_0}$

Dual error: $\mathbf{E}[OPT - D(y^t)] \leq \rho^t U_0$

Primal error: $\mathbf{E}[P(x^t) - OPT] \leq \rho^t U_0 + 2\rho^{t/2}\sqrt{OPT \times U_0}$

Duality gap: $\mathbf{E}[P(x^t) - D(y^t)] \leq 2\rho^t U_0 + 2\rho^{t/2}\sqrt{OPT \times U_0}$

The Rate: Lower and Upper Bounds

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

Theorem [RG'15ab]

$$0 \leq 1 - \frac{\mathbf{Rank}(S^\top A)}{\mathbf{Rank}(A)} \leq \rho < 1$$

Insight:

$\rho \leq 1$ always

$\rho < 1$ if Assumption 2 holds

Insight: The lower bound is good when:

i) the dimension of the search space in the "constrain and approximate" viewpoint is large,

ii) the rank of *A* is small

Part VII
Conclusion

Contributions

- 4 Equivalent stochastic reformulations of a linear system
  - Stochastic optimization
  - Stochastic fixed point problem
  - Stochastic linear system
  - Probabilistic intersection
- 3 Algorithms
  - Basic (SGD, stochastic Newton method, stochastic fixed point method, stochastic proximal point method, stochastic projection method, …)
  - Parallel
  - Accelerated
- Iteration complexity guarantees for various measures of success
  - Expected iterates (closed form)
  - L1 / L2 convergence
  - Convergence of *f*; ergodic …

Related Work

Basic method with unit stepsize and full rank *A*

- 2017 IMA Fox Prize (2^{nd} Prize) in Numerical Analysis
- Most downloaded SIMAX paper

★ Robert Mansel Gower and P.R.
Randomized Iterative Methods for Linear Systems
SIAM J. Matrix Analysis & Applications 36(4):1660-1690, 2015

Removal of full rank assumption + duality

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

Inverting matrices & connection to Quasi-Newton updates

Robert Mansel Gower and P.R.
Randomized Quasi-Newton Methods are Linearly Convergent Matrix Inversion Algorithms
arXiv:1602.01768, 2016

Computing the pseudoinverse

Robert Mansel Gower and P.R.
Linearly Convergent Randomized Iterative Methods for Computing the Pseudoinverse
arXiv:1612.06255, 2016

Application in machine learning

Robert Mansel Gower, Donald Goldfarb and P.R.
Stochastic Block BFGS: Squeezing More Curvature out of Data
ICML 2016

Related Work

Stochastic Reformulations

P.R. and Martin Takáč.
Stochastic Reformulations of Linear Systems: Algorithms and Convergence Theory
arXiv:1706.01108, 2017

+ Polyak Momentum

Nicolas Loizou and P.R.
Momentum and Stochastic Momentum for Stochastic Gradient, Newton, Proximal Point and Subspace Descent Methods
arXiv:1712.09677, 2017

Basic method with unit stepsize and full rank *A*

Dmitry Kovalev, Eduard Gorbunov, Elnur Gasanov and P.R.
Stochastic Spectral and Conjugate Descent Methods
arXiv:1802.03703, 2018

POSTER

1st acceleration of BFGS matrix update rules

Robert M. Gower, Filip Hanzely, P.R. and Sebastian Stich
Accelerated Stochastic Matrix Inversion: General Theory and Speeding up BFGS Rules for Faster Second-Order Optimization
arXiv:1802.04079, 2018

Convex Feasibility

Ion Necoara, Andrei Patrascu and P.R.
Randomized Projection Methods for Convex Feasibility Problems: Conditioning and Convergence Rates
arXiv:1801.04873, 2018

Extra Material:
Special Cases

Special Case 1: 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:

$B = I$ $\qquad S = e^i = (0, \ldots, 0, 1, 0, \ldots, 0)$ 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

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

Special Choice of Parameters

$$B = I$$

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

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

Complexity Rate

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

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

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!

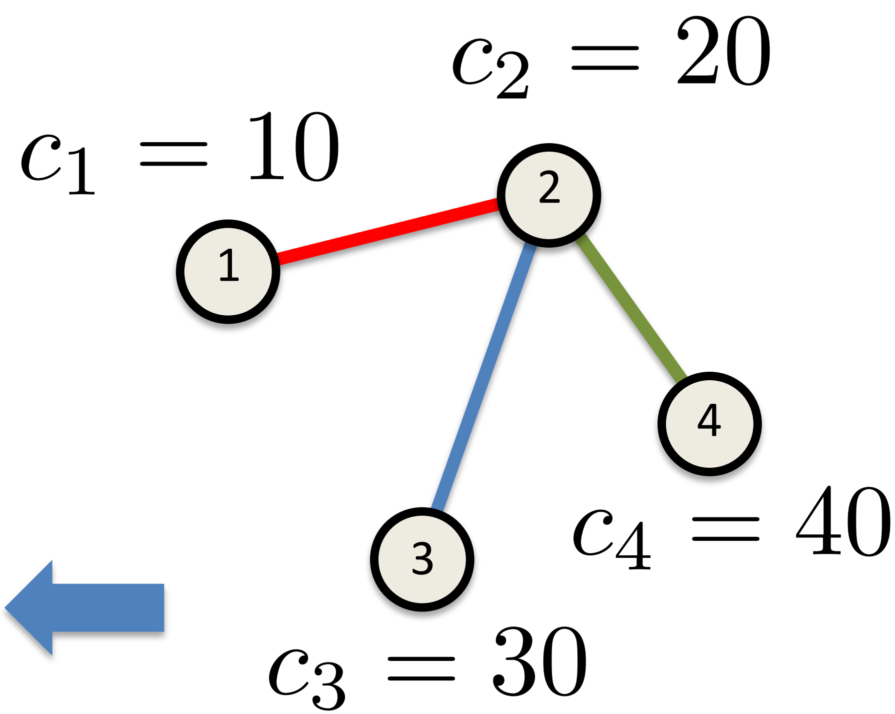
$$x^{t+1} = \arg\min_{x \in \mathbb{R}^n} \|x - x^*\|_2^2 \quad \text{s.t.} \quad x = x^t + \boxed{y}(A_{i:})^T, \quad y \in \mathbb{R}$$

Application: Average Consensus

$$\min_{x\in\mathbb{R}^4} \frac{1}{2}\|x-c\|_2^2$$

$$\text{subject to } Ax = 0$$

$$A = \begin{pmatrix} 1 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 1 & 0 & -1 \end{pmatrix}$$



Insight: Randomized Kaczmarz = Randomized Gossip

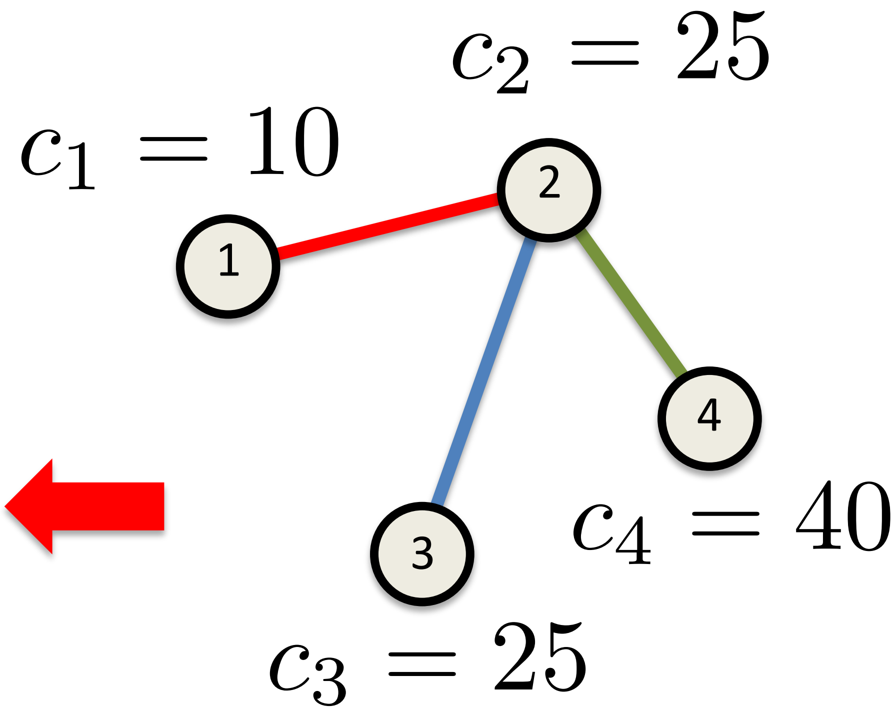
Now also have: dual interpretation, block variants, …

Application: Average Consensus

$$\min_{x \in \mathbb{R}^4} \frac{1}{2}\|x - c\|_2^2$$

$$\text{subject to } Ax = 0$$

$$A = \begin{pmatrix} 1 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 1 & 0 & -1 \end{pmatrix}$$



Insight: Randomized Kaczmarz = Randomized Gossip

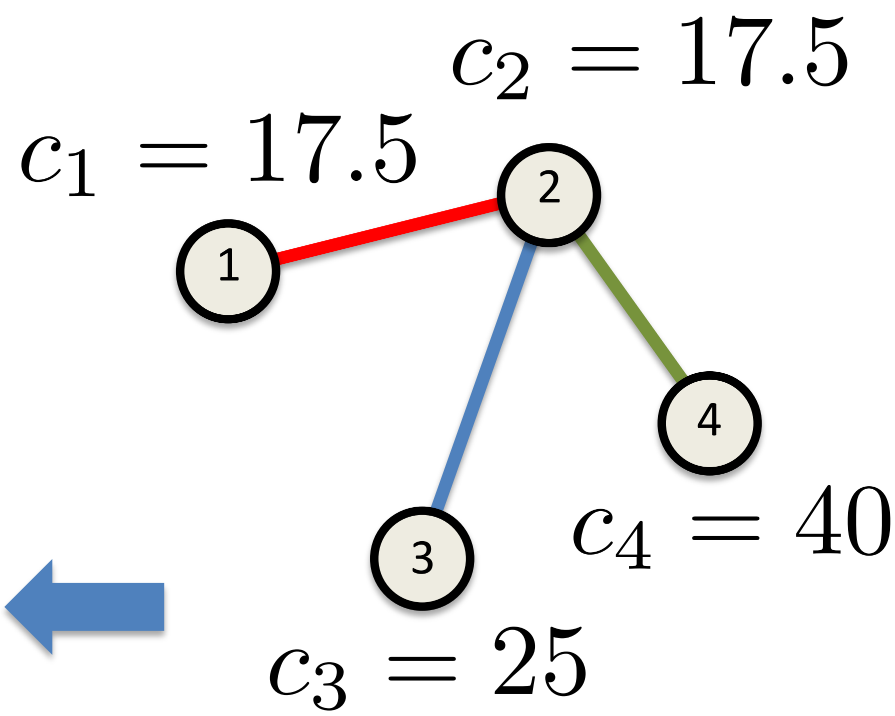
Now also have: dual interpretation, block variants, …

Application: Average Consensus

$$\min_{x\in\mathbb{R}^4} \frac{1}{2}\|x-c\|_2^2$$

$$\text{subject to } Ax = 0$$

$$A = \begin{pmatrix} 1 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 1 & 0 & -1 \end{pmatrix}$$



Insight: Randomized Kaczmarz = Randomized Gossip

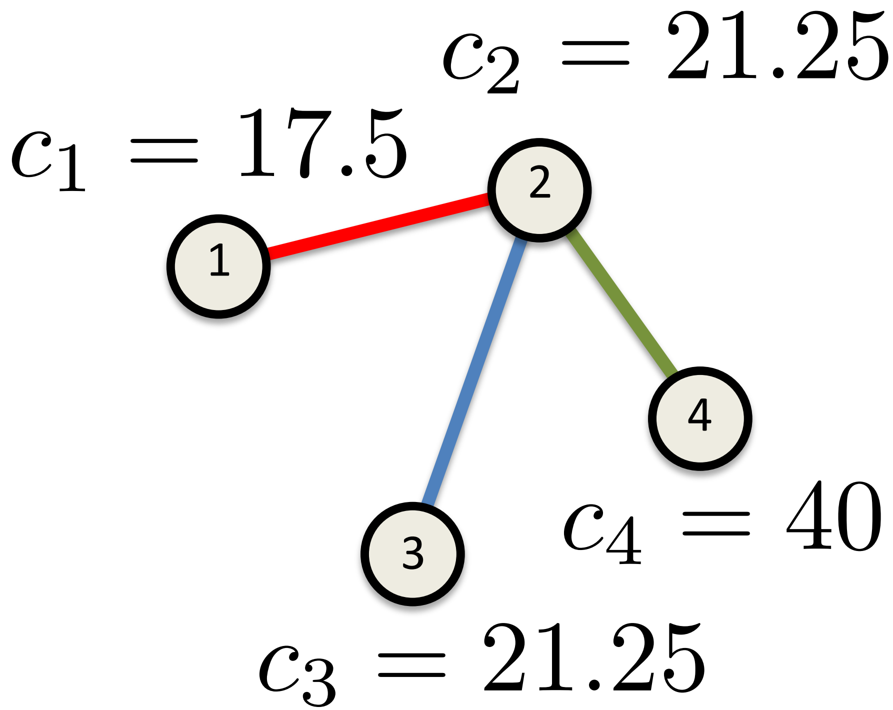
Now also have: dual interpretation, block variants, ...

Application: Average Consensus

$$\min_{x \in \mathbb{R}^4} \frac{1}{2}\|x - c\|_2^2$$

$$\text{subject to } Ax = 0$$

$$A = \begin{pmatrix} 1 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 1 & 0 & -1 \end{pmatrix}$$



Insight: Randomized Kaczmarz = Randomized Gossip

Now also have: dual interpretation, block variants, …

RK: Further Reading

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

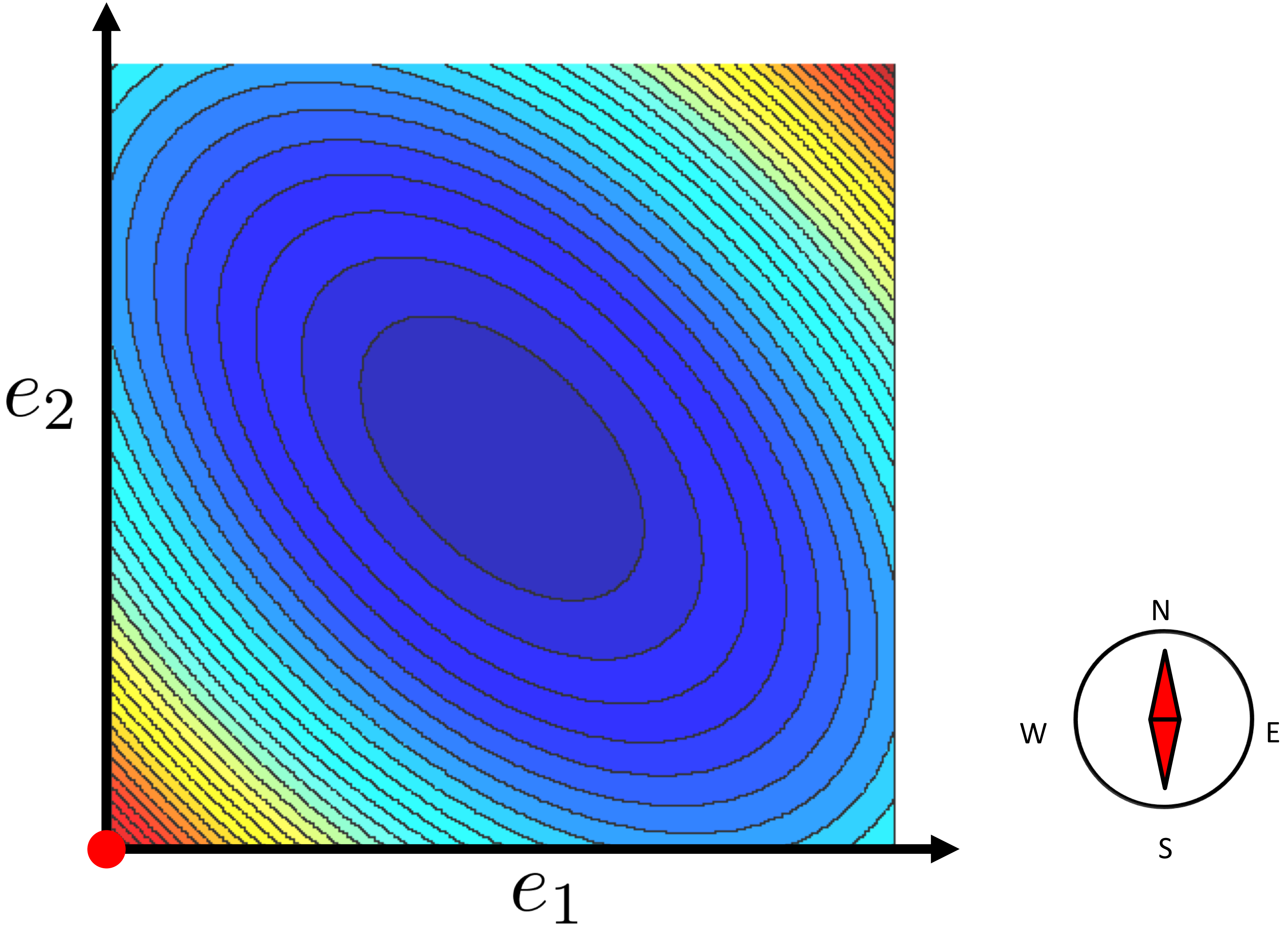
D. Needell and J. Tropp. **Paved with good intentions: analyzis of a randomized block Kaczmarz method.** *Linear Algebra and its Applications* 441, pp. 199-221, 2012

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

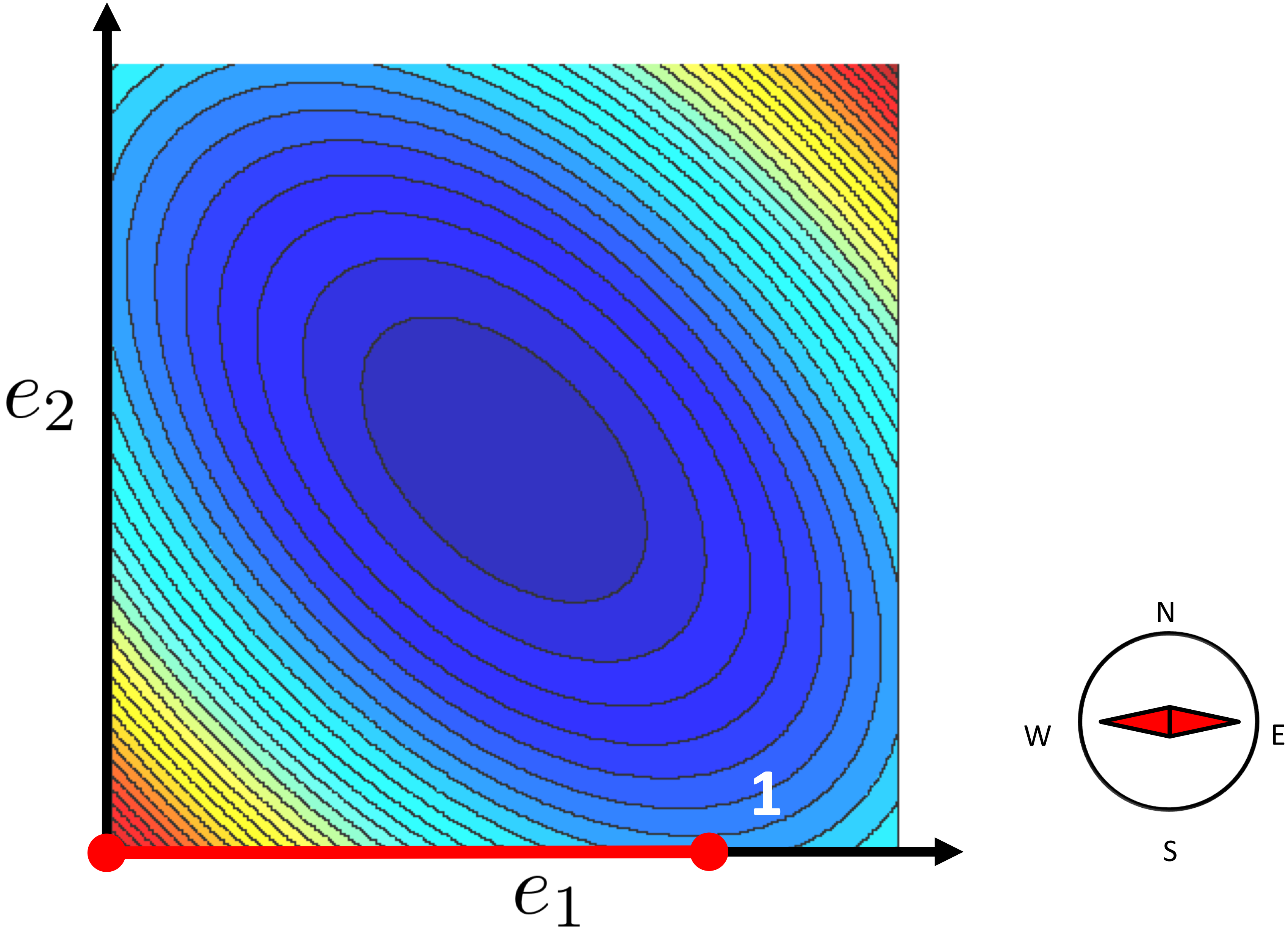
A. Ramdas. **Rows vs Columns for Linear Systems of Equations – Randomized Kaczmarz or Coordinate Descent?** *arXiv:1406.5295*, 2014

Special Case 2: Randomized Coordinate Descent

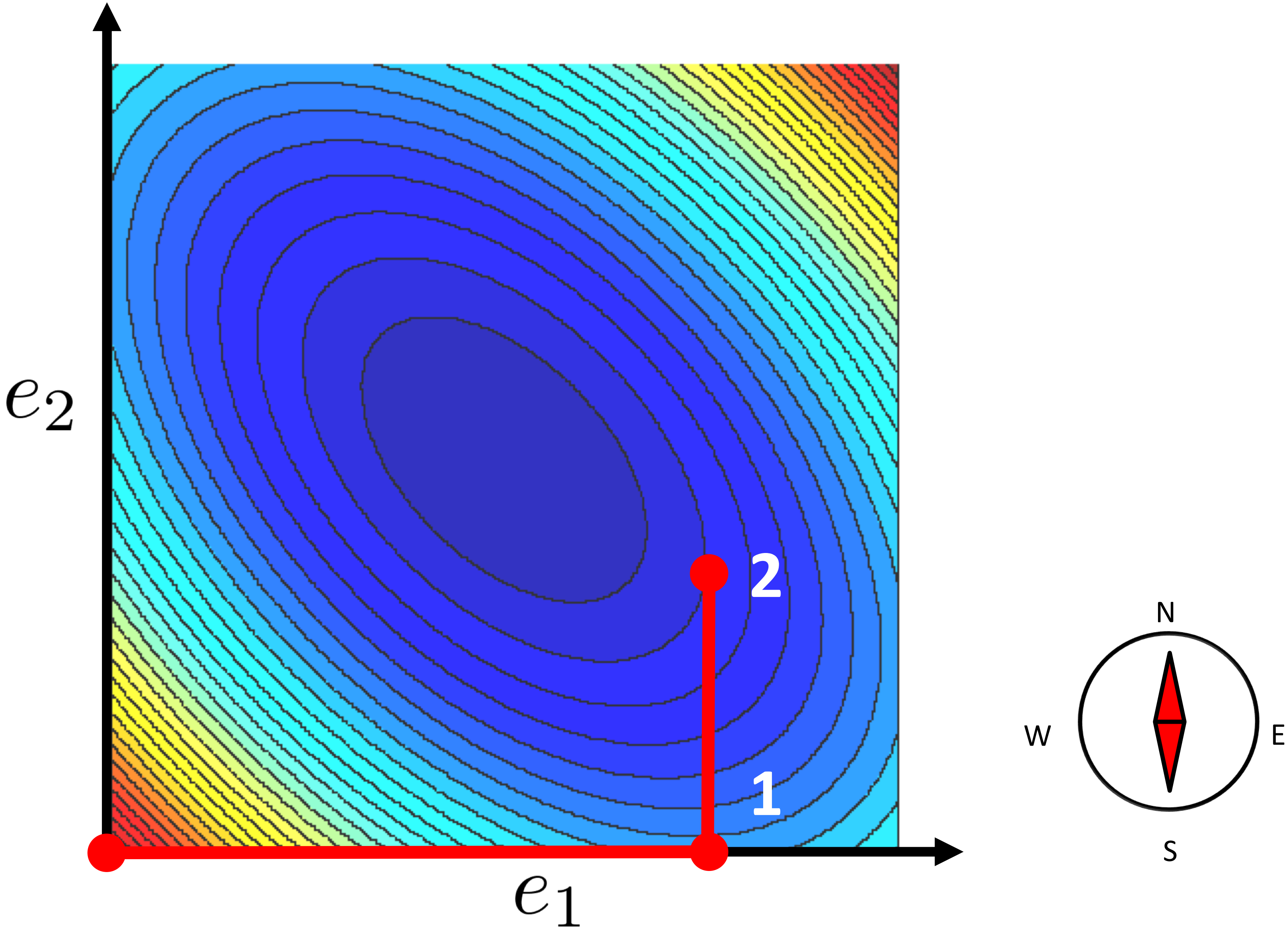
Randomized Coordinate Descent in 2D



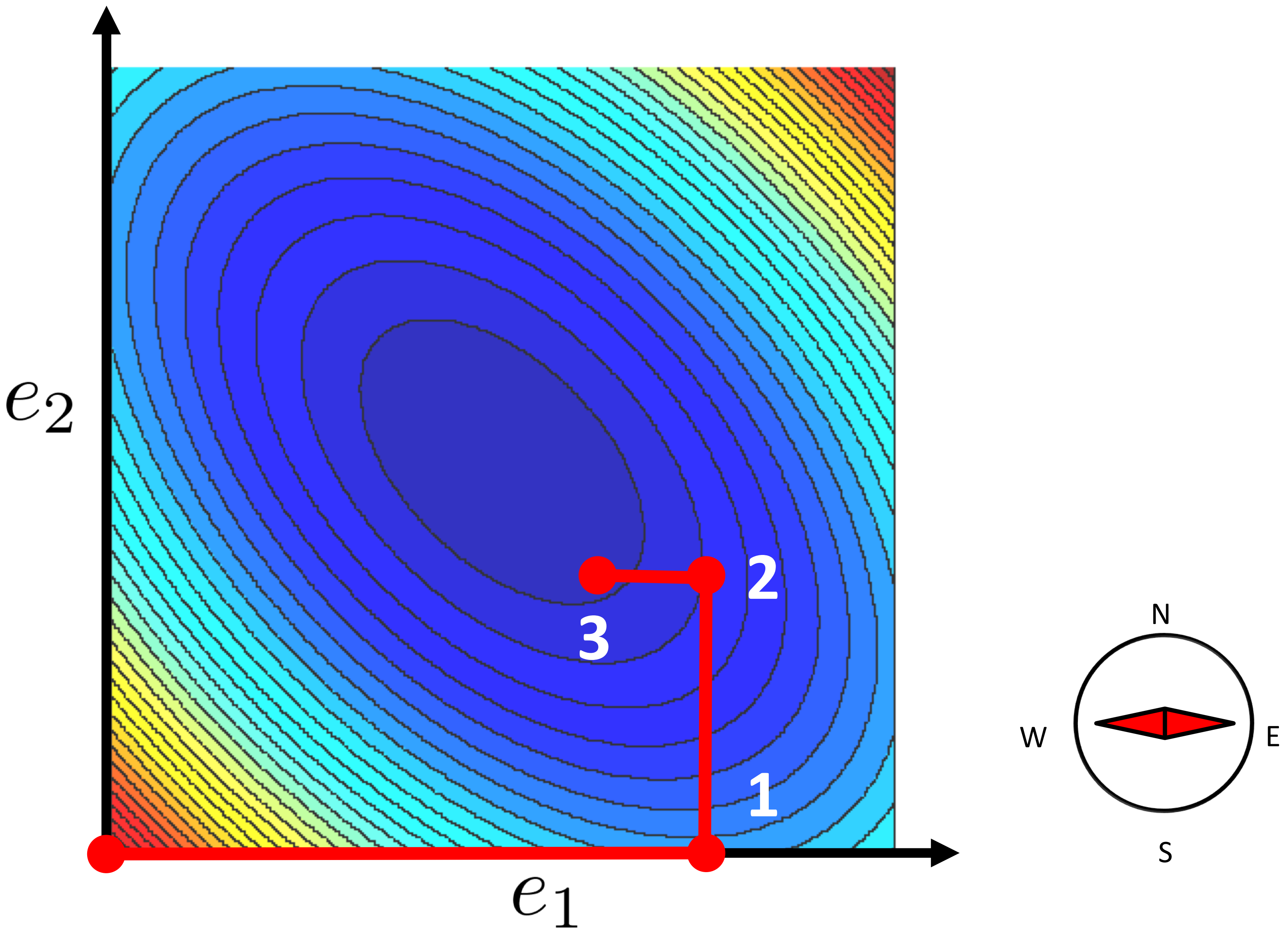
Randomized Coordinate Descent in 2D



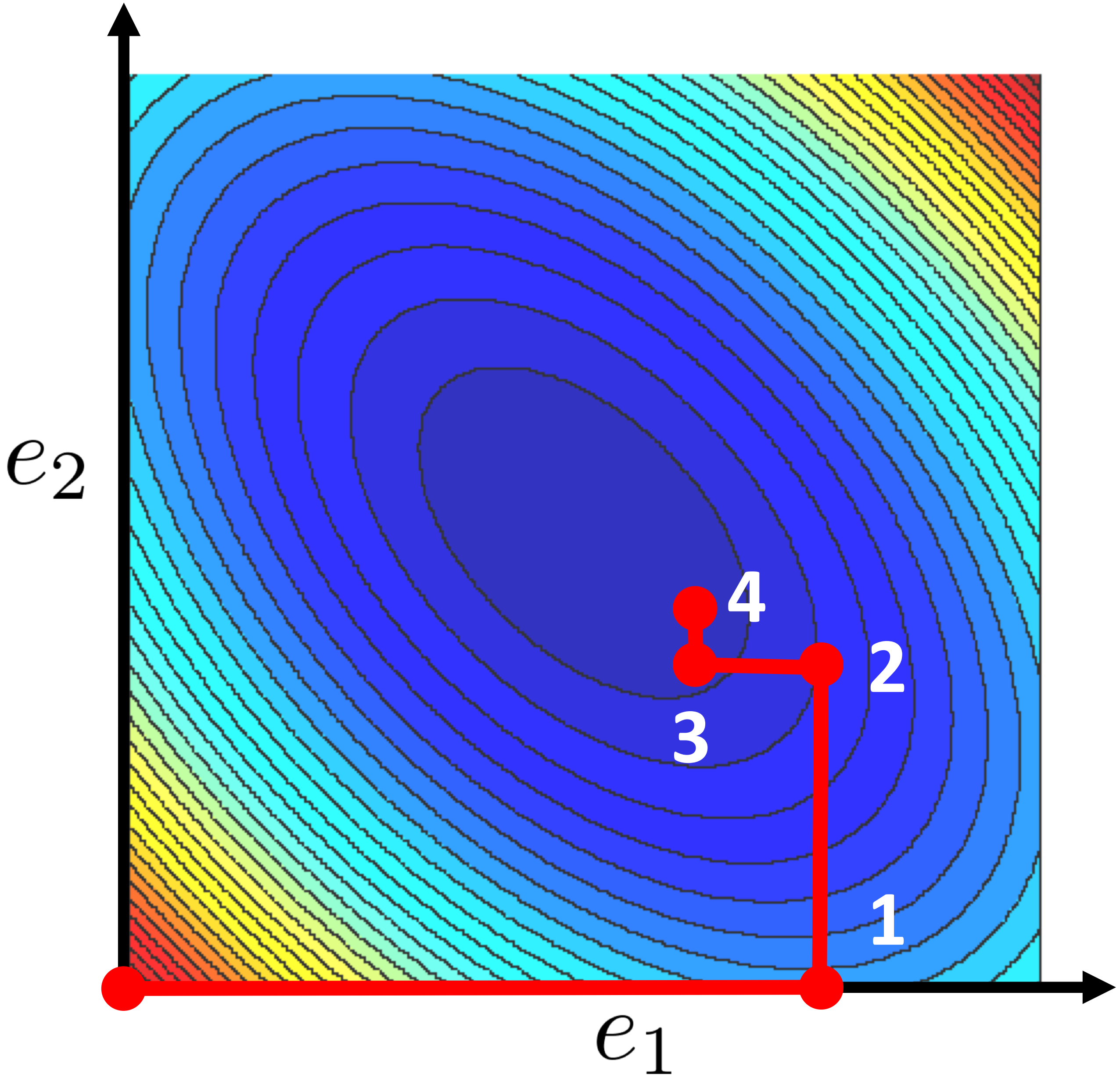
Randomized Coordinate Descent in 2D

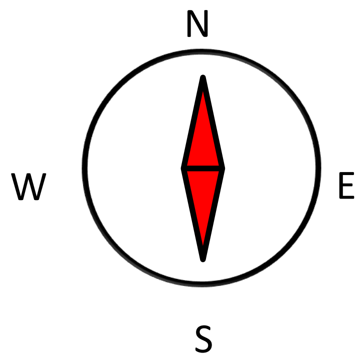


Randomized Coordinate Descent in 2D

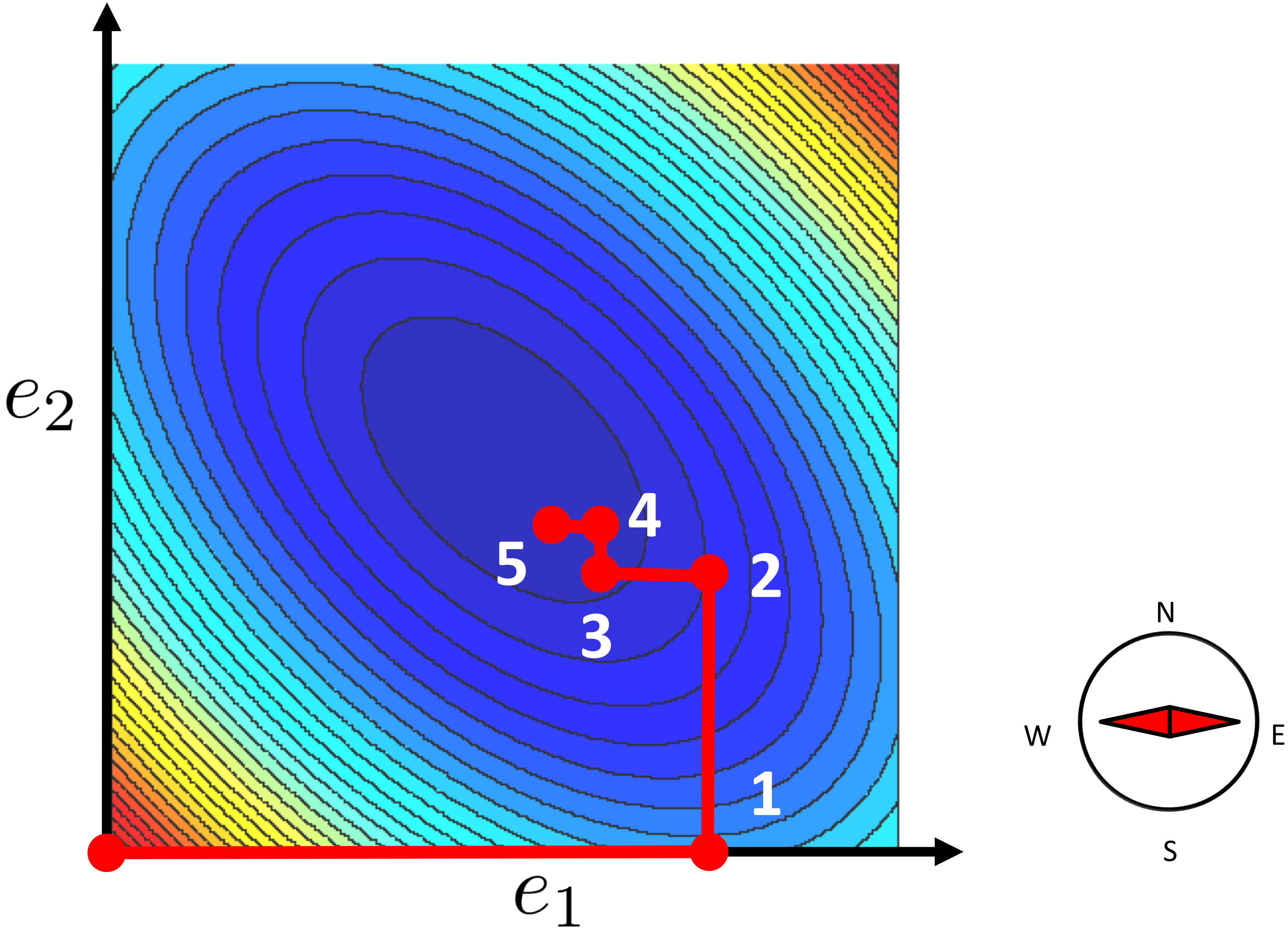


Randomized Coordinate Descent in 2D

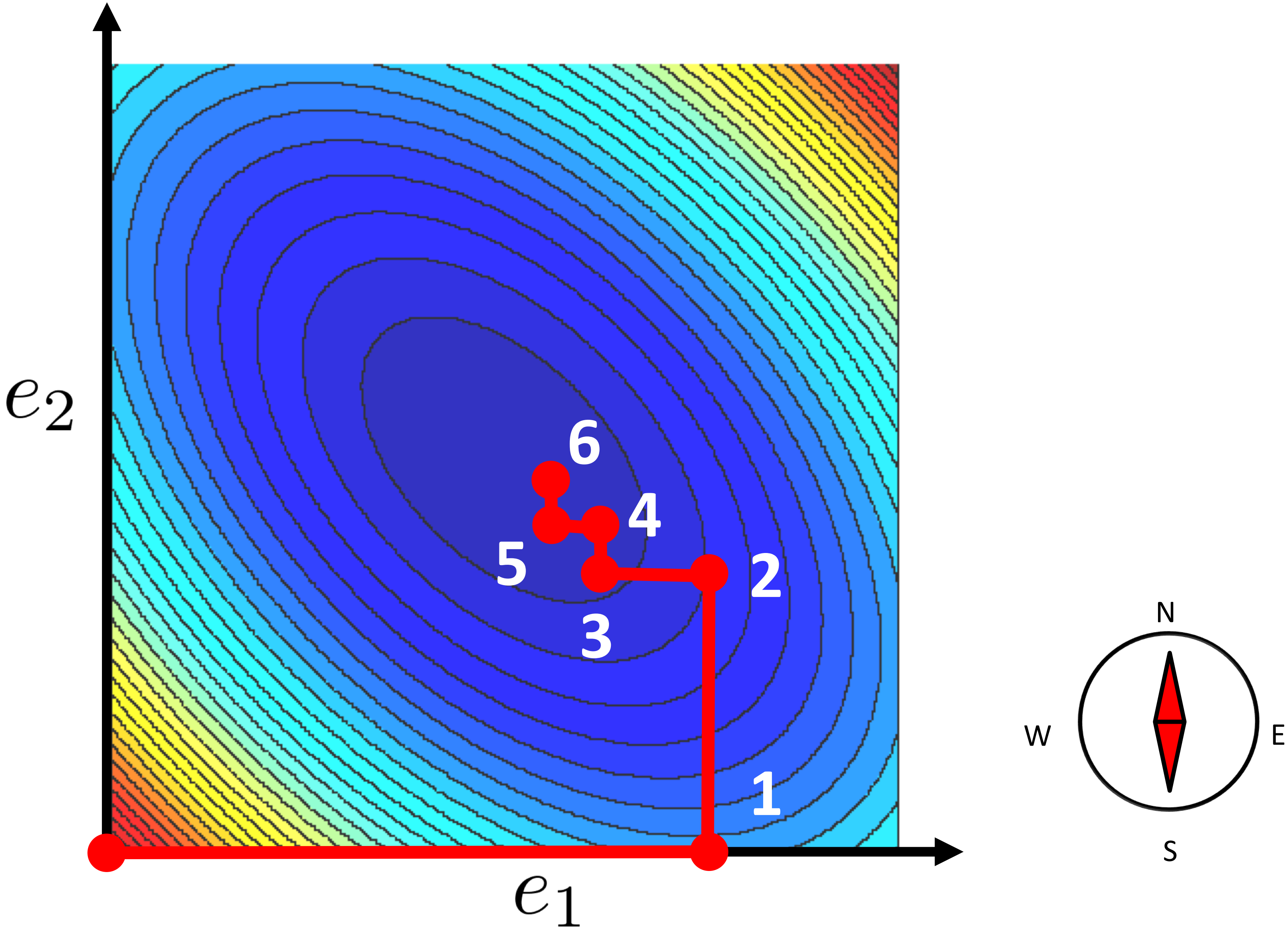




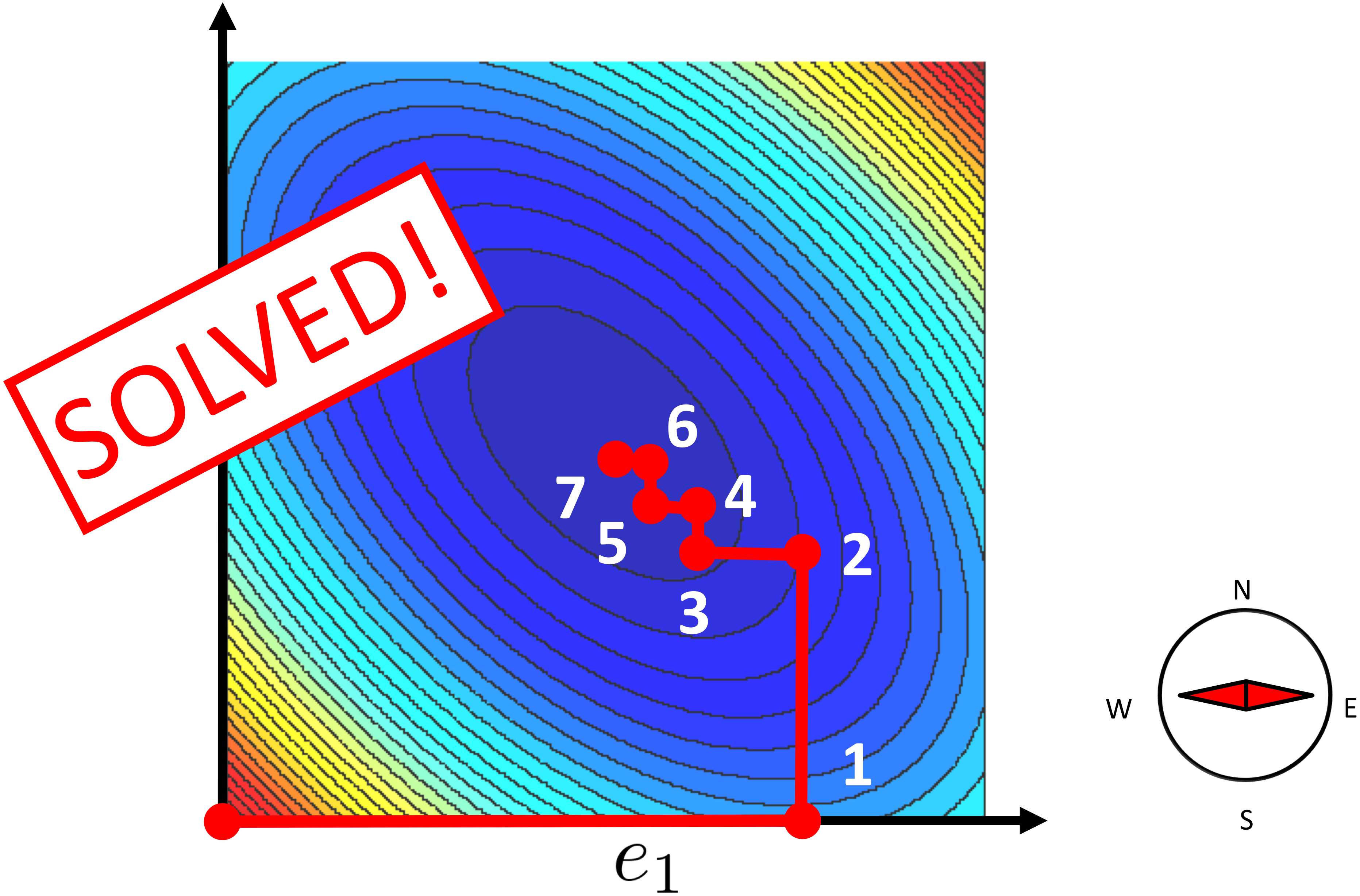
Randomized Coordinate Descent in 2D



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)

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

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

Assume: Positive definite

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

$B = A$ $\quad$ $S = e^i = (0,\dots,0,1,0,\dots,0)$ 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

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

Special Choice of Parameters

$$B = A$$

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

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

Complexity Rate

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

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

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:

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

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

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

We recover RCD as we have seen it:

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

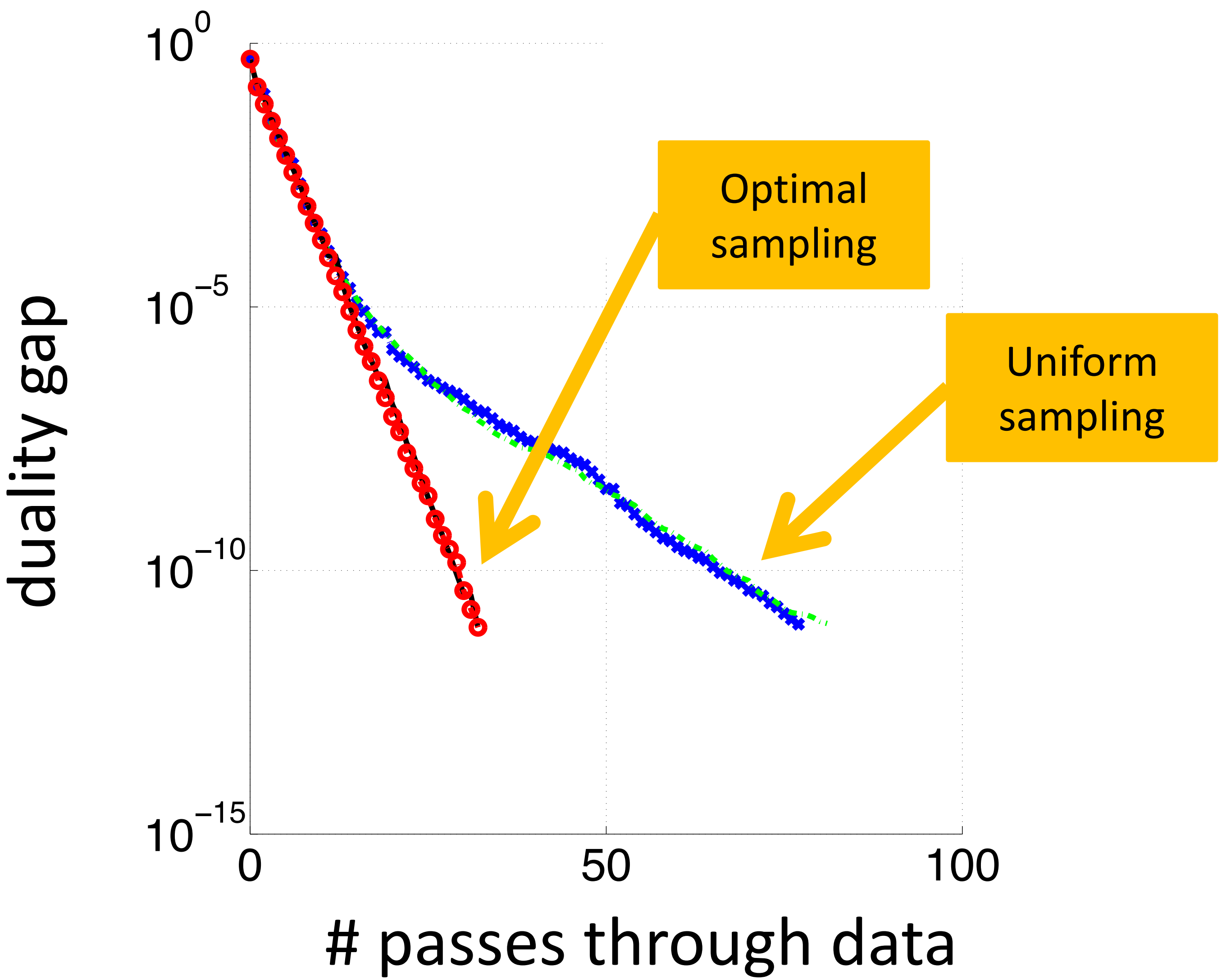
Experiment 1

Machine: laptop

Problem: logistic regression, n = 522,911, d = 54

Zheng Qu, P.R. and Tong Zhang. **Quartz: Randomized Dual Coordinate Ascent with Arbitrary Sampling.** In *Advances in Neural Information Processing Systems* 28, 2015

Logistic Regression: Laptop



$\mathrm{Data} = \texttt{cov1}, \quad n = 522{,}911, \quad \lambda = 10^{-6}$

Experiment 2

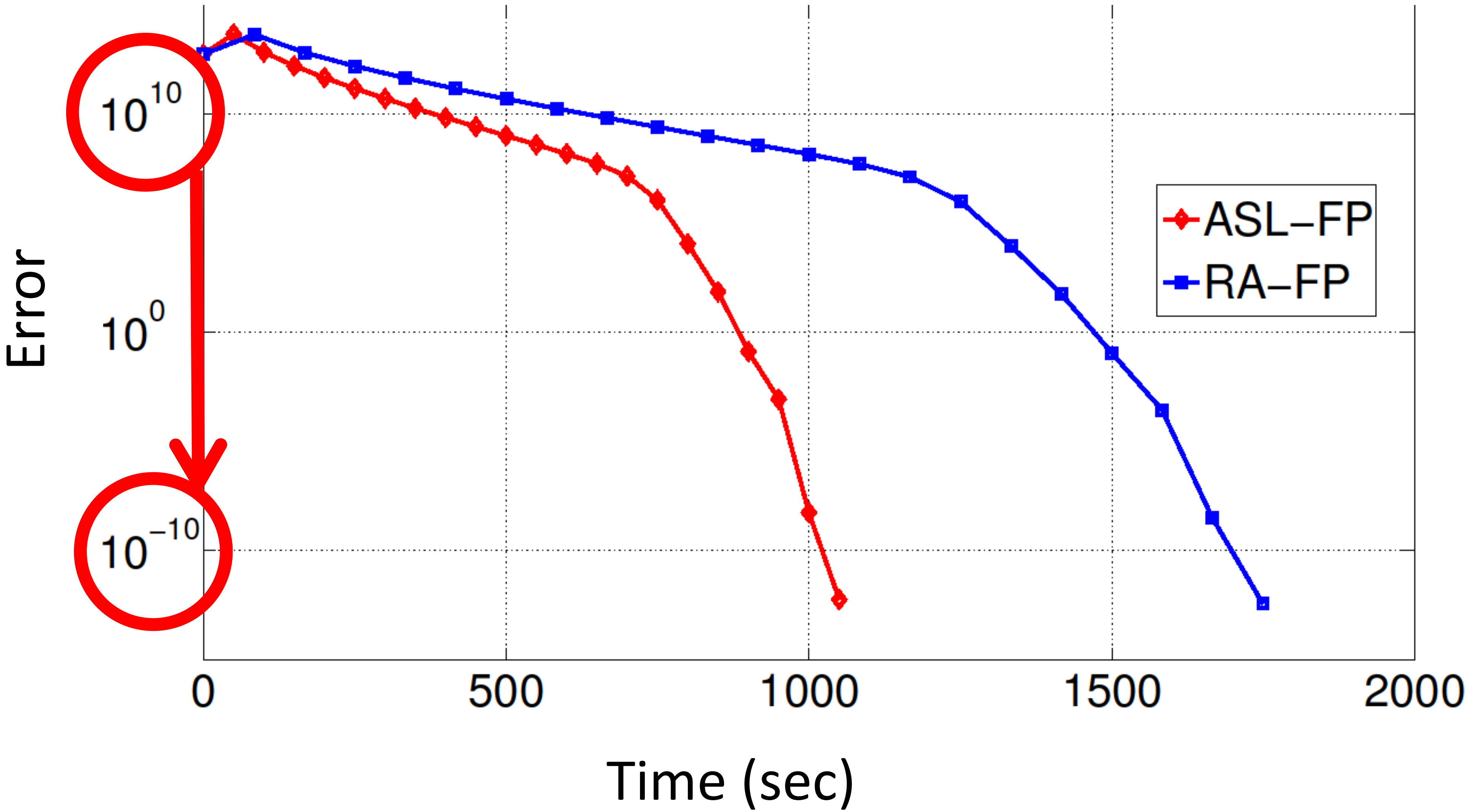
Machine: 128 nodes of Hector Supercomputer (4096 cores)

Problem: LASSO, n = 1 billion, d = 0.5 billion, 3 TB

HYDRA

P.R. and Martin Takáč. **Distributed coordinate descent for learning with big data.** *Journal of Machine Learning Research* 17, 2016 (*arXiv:1310.2059*, 2013)

LASSO: 3TB data + 128 nodes



Experiment 3

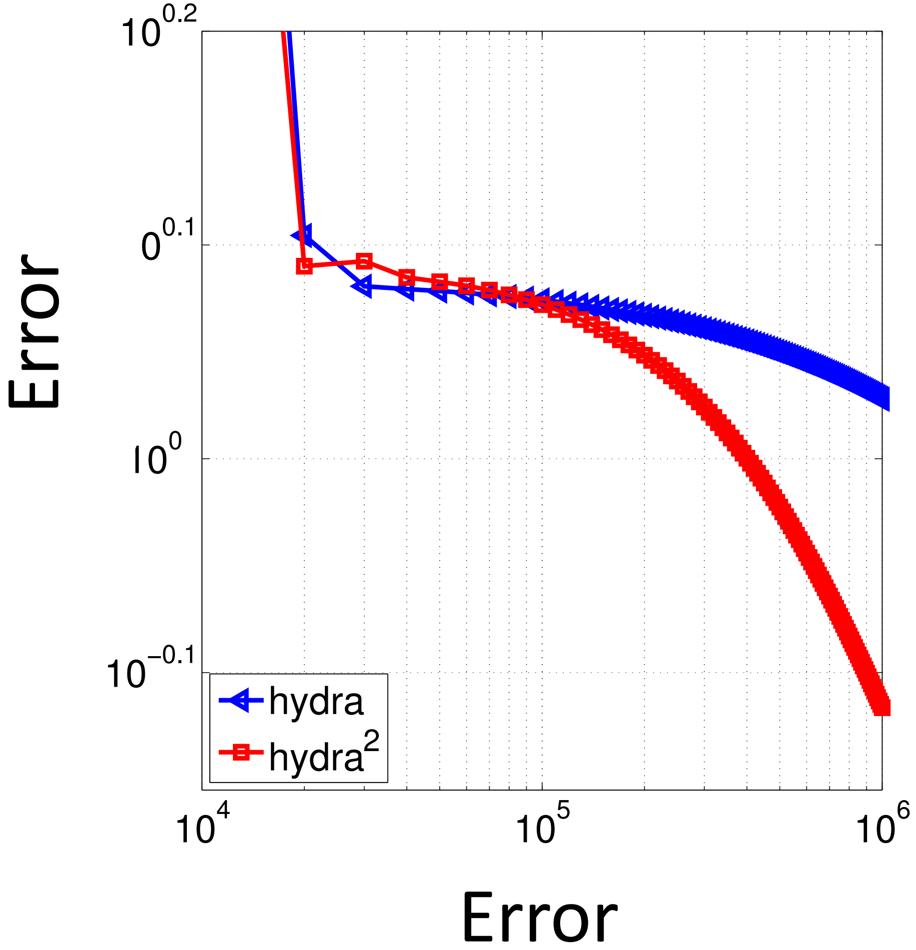
Machine: 128 nodes of Archer Supercomputer

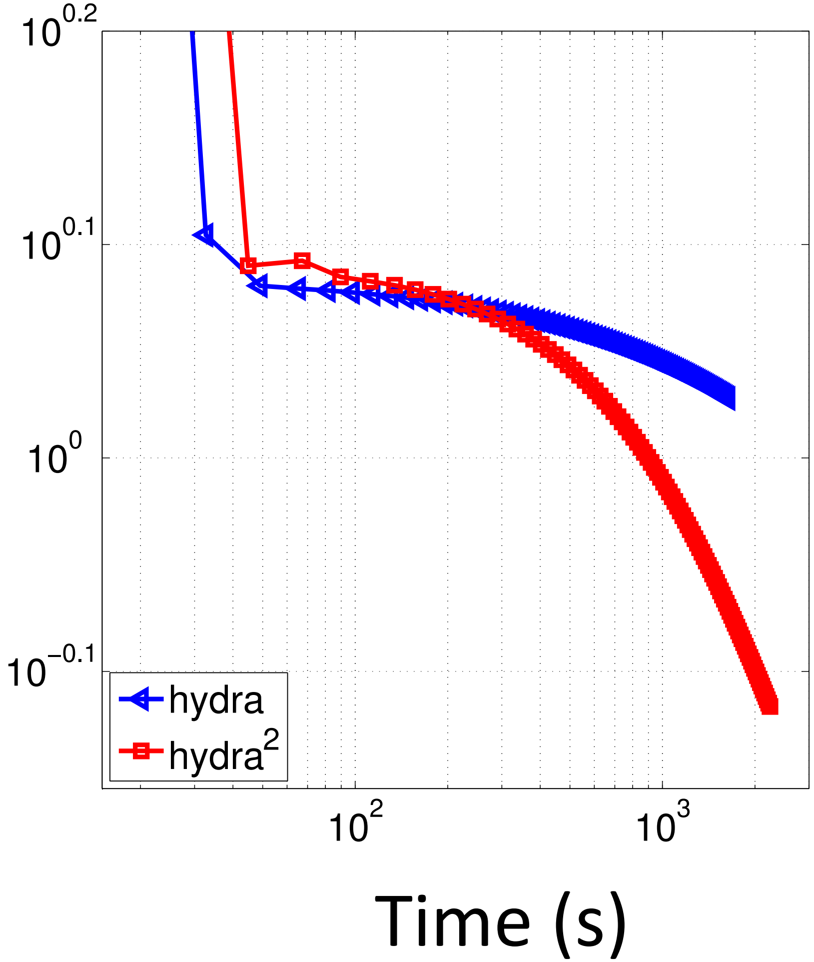
Problem: LASSO, n = 5 million, d = 50 billion, 5 TB
(60,000 nnz per row of A)

HYDRA^2

Olivier Fercoq, Zheng Qu, P.R. and Martin Takáč. **Fast distributed coordinate descent for minimizing non-strongly convex losses.** *In 2014 IEEE Int. Workshop on Machine Learning for Signal Proc*, 2014

LASSO: 5 TB data (d = 50 billion)
128 nodes





Special Case 3: Randomized Newton Method

Randomized Newton (RN)

Z. Qu, PR, M. Takáč and O. Fercoq. **Stochastic Dual Newton Ascent for Empirical Risk Minimization.** *ICML* 2016

SDNA

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

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

Assume: Positive definite

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

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

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

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

RN: Derivation

General Method

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

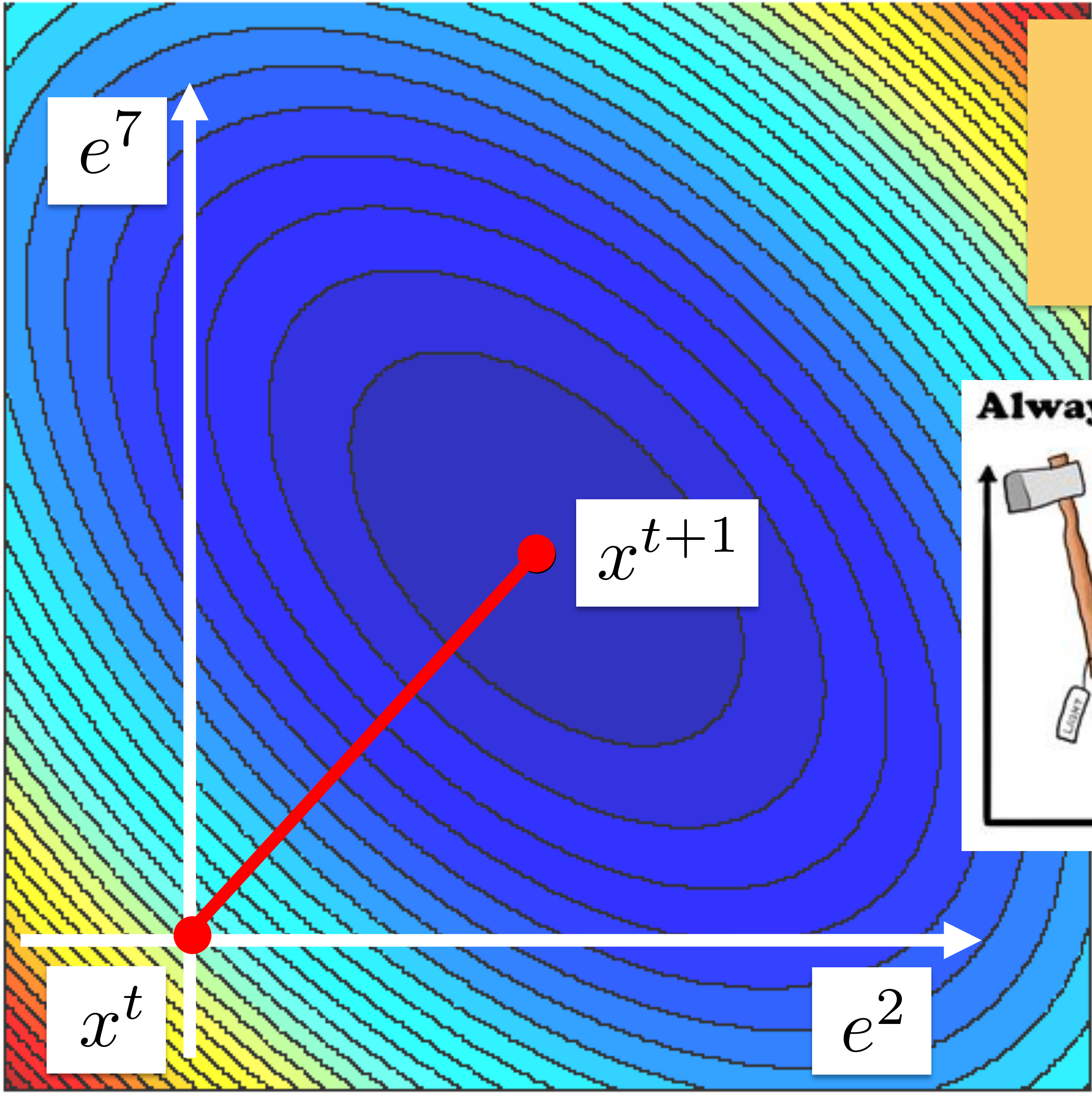
Special Choice of Parameters

$$B = A$$

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

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

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



$$C = \{2, 7\}$$
$$|C| = 2$$

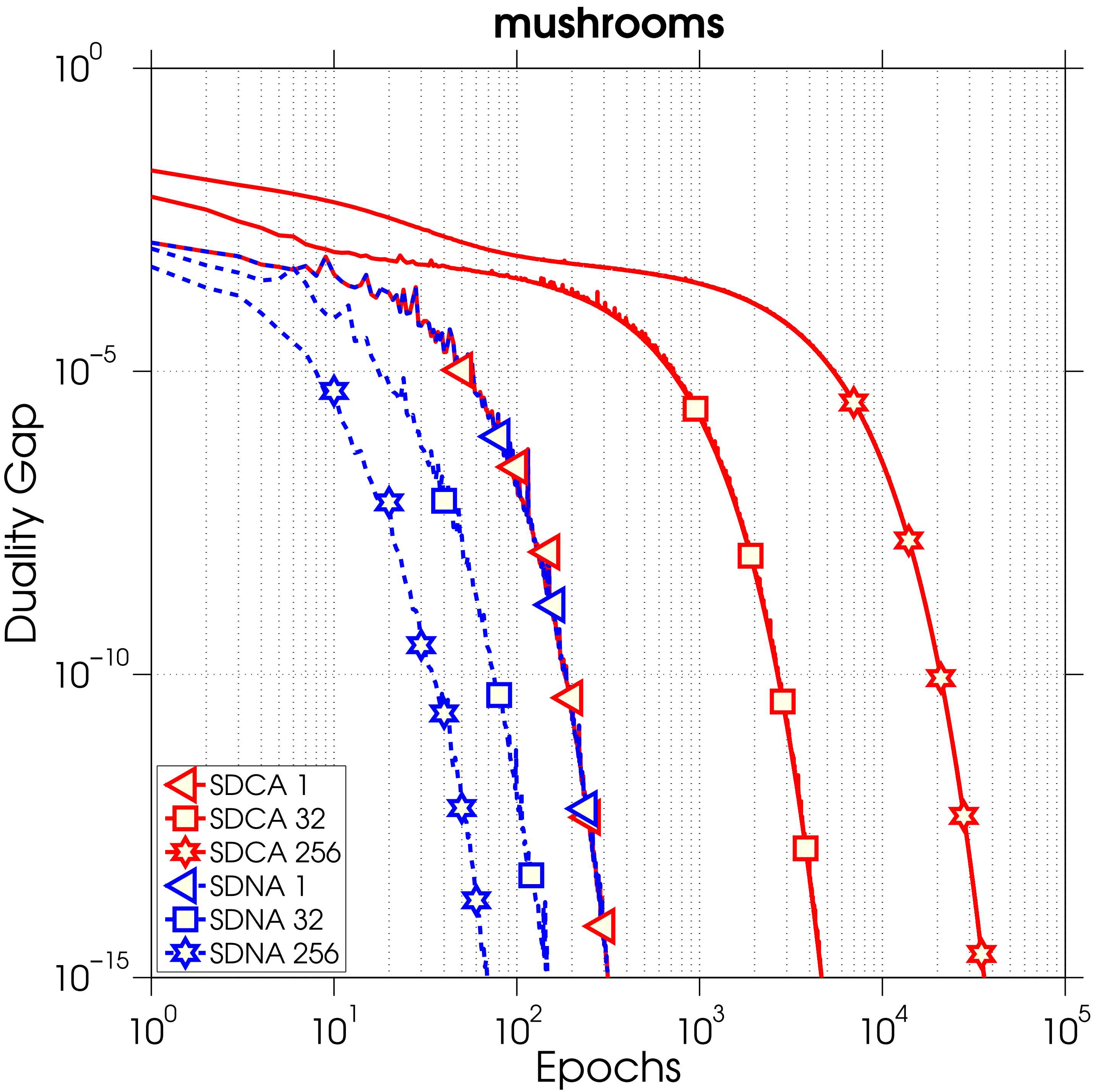
Experiment 4

Machine: laptop

Problem: Ridge Regression, n = 8124, d = 112

SDNA

Zheng Qu, P.R., Martin Takáč and Olivier Fercoq, **SDNA: Stochastic Dual Newton Ascent for Empirical Risk Minimization.** To appear in *ICML*, 2016


mushrooms
Duality Gap
Epochs
10^{0}
10^{-5}
10^{-10}
10^{-15}
10^{0}
10^{1}
10^{2}
10^{3}
10^{4}
10^{5}
SDCA 1
SDCA 32
SDCA 256
SDNA 1
SDNA 32
SDNA 256

Special Case 4: Gaussian Descent

Gaussian Descent

General Method

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

Special Choice of Parameters

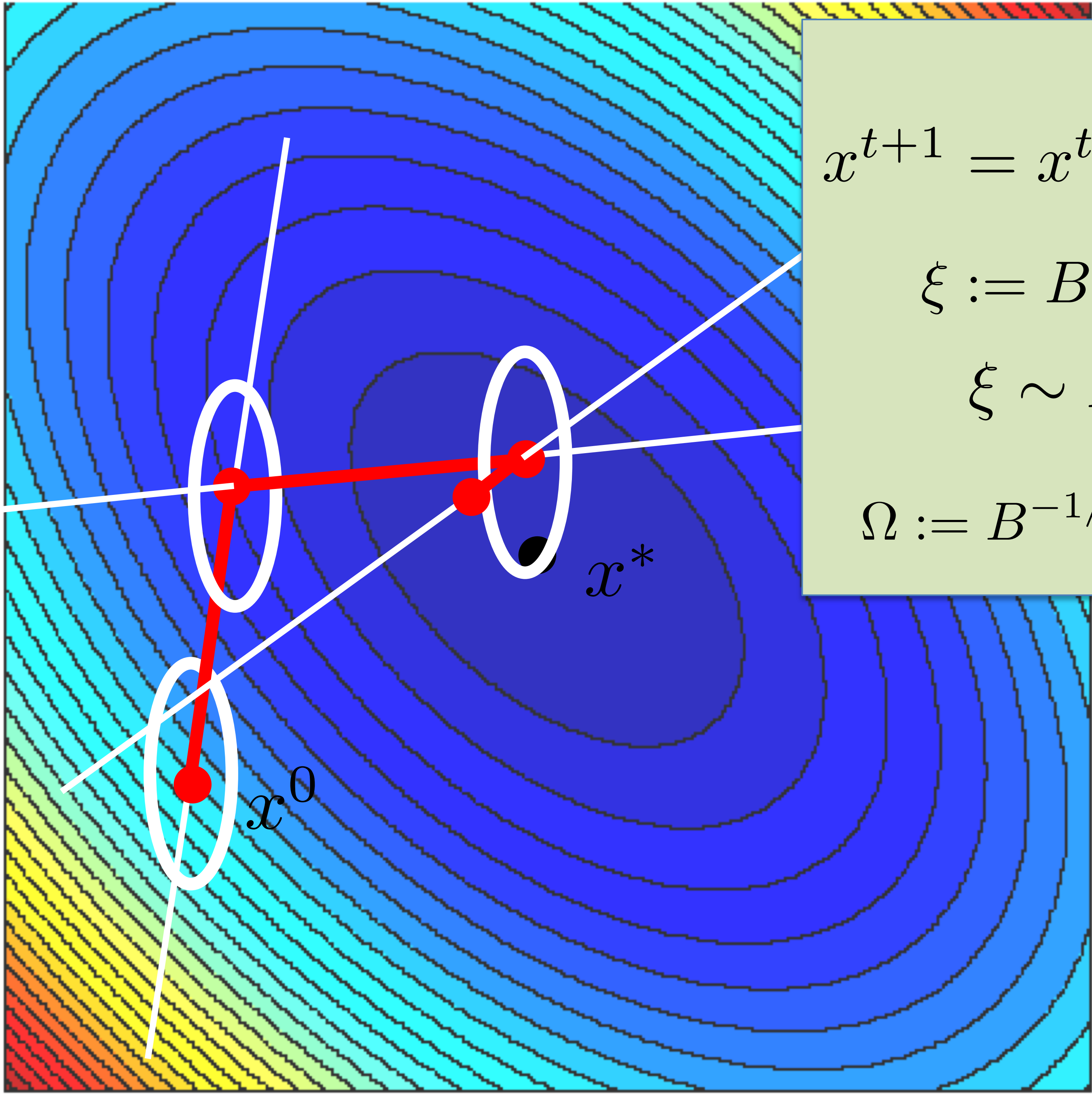
$$S \sim N(0, \Sigma)$$

Positive definite covariance matrix

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

Complexity Rate

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



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

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

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

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

Gaussian Descent: The Rate

Lemma [GR'15a]

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

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

This follows from the general lower

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

Extra Material: Importance Sampling

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

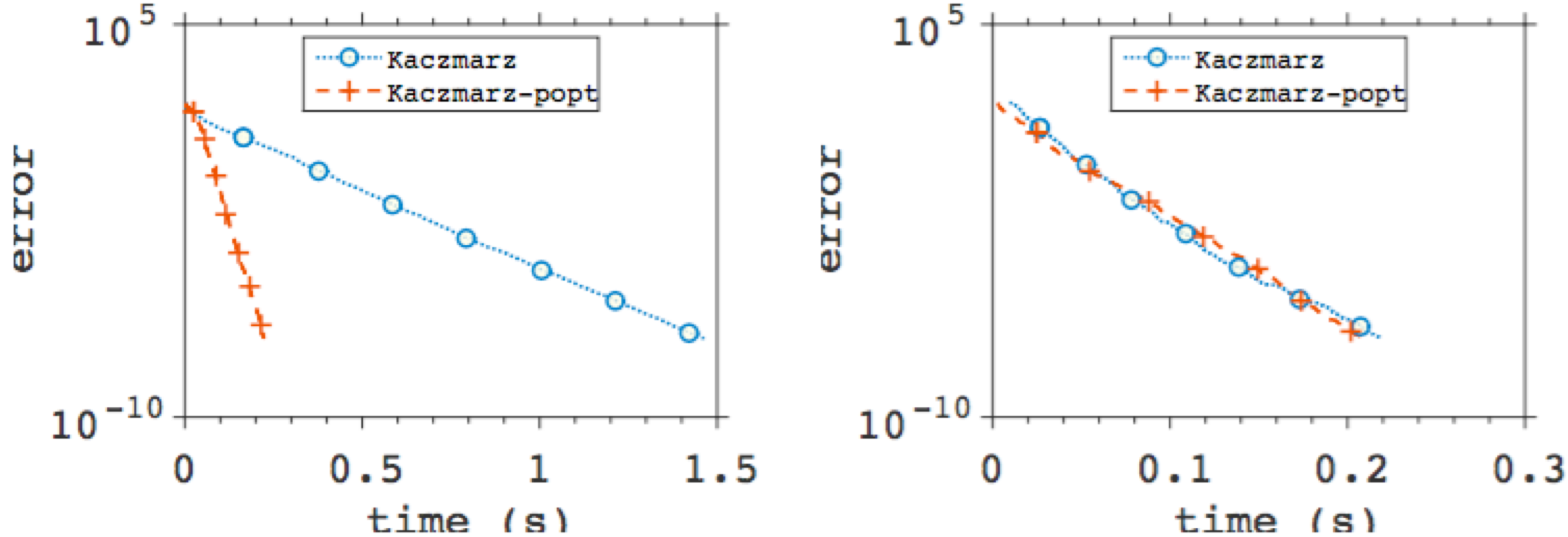
Rate for **convenient** (standard) **probabilities**

Rate for **optimal probabilities** (solving SDP)

Lower bound on the rate

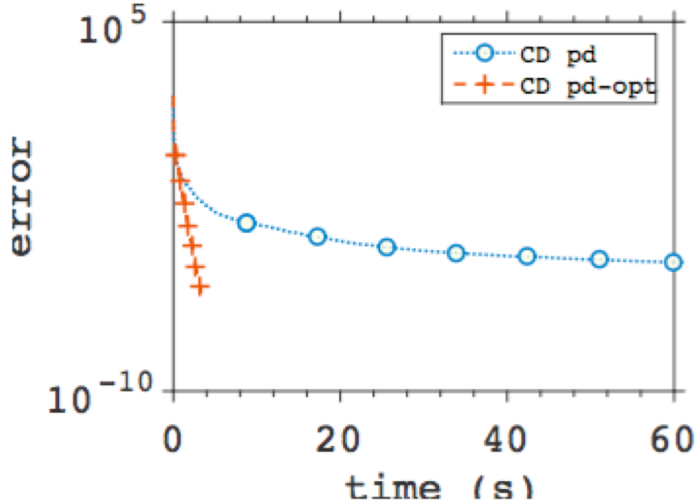
data set	ρ_c	ρ^*	$1 - 1/n$
rand(50,50)	$1 - 2 \cdot 10^{-6}$	$1 - 3.05 \cdot 10^{-6}$	$1 - 2.10^{-2}$
mushrooms-ridge	$1 - 5.86 \cdot 10^{-6}$	$1 - 7.15 \cdot 10^{-6}$	$1 - 8.93 \cdot 10^{-3}$
aloi-ridge	$1 - 2.17 \cdot \boxed{10^{-7}}$	$1 - 1.26 \cdot \boxed{10^{-4}}$	$1 - 7.81 \cdot 10^{-3}$
liver-disorders-ridge	$1 - 5.16 \cdot 10^{-4}$	$1 - 8.25 \cdot 10^{-3}$	$1 - 1.67 \cdot 10^{-1}$
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

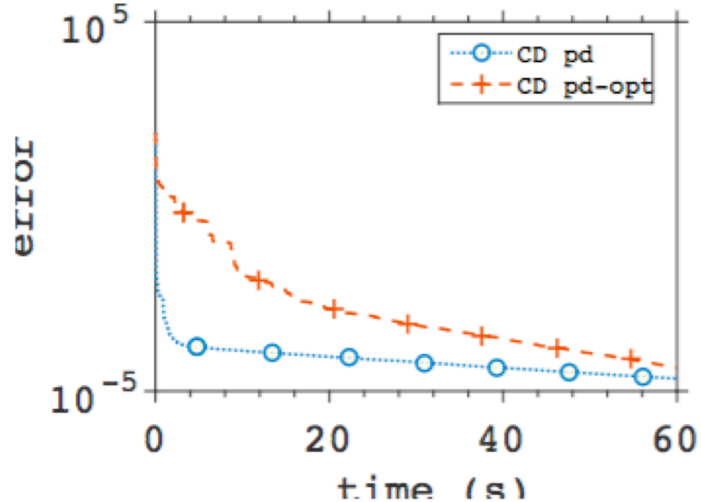


(a) `liver-disorders-popt-k` (b) `rand`(500,100)

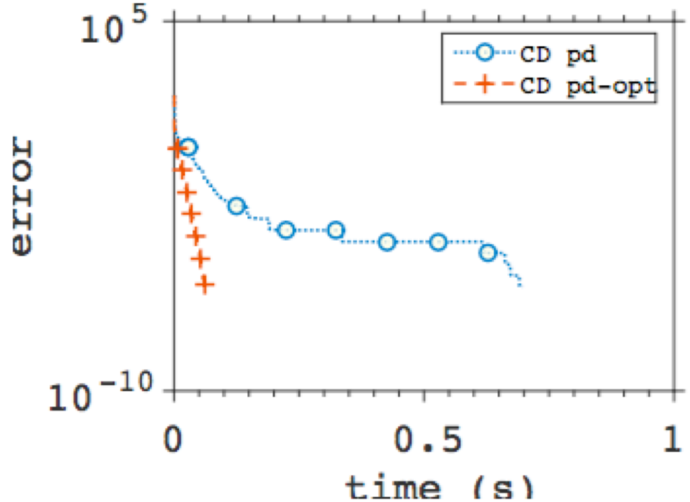
RCD: Convenient vs Optimal



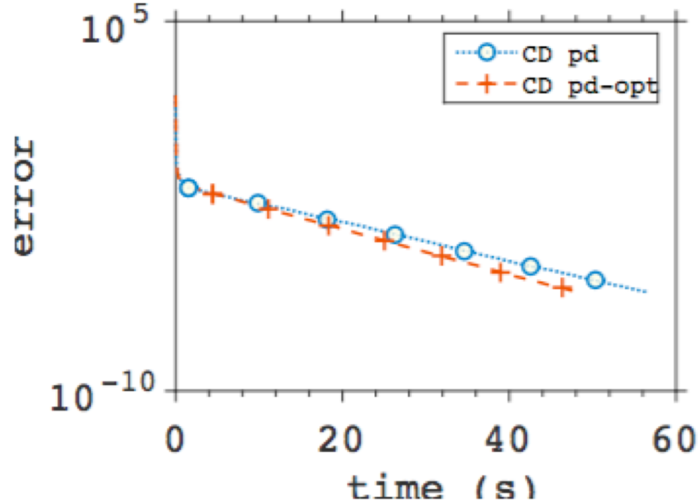
(a) **aloi**



(b) **covtype.libsvm.binary**



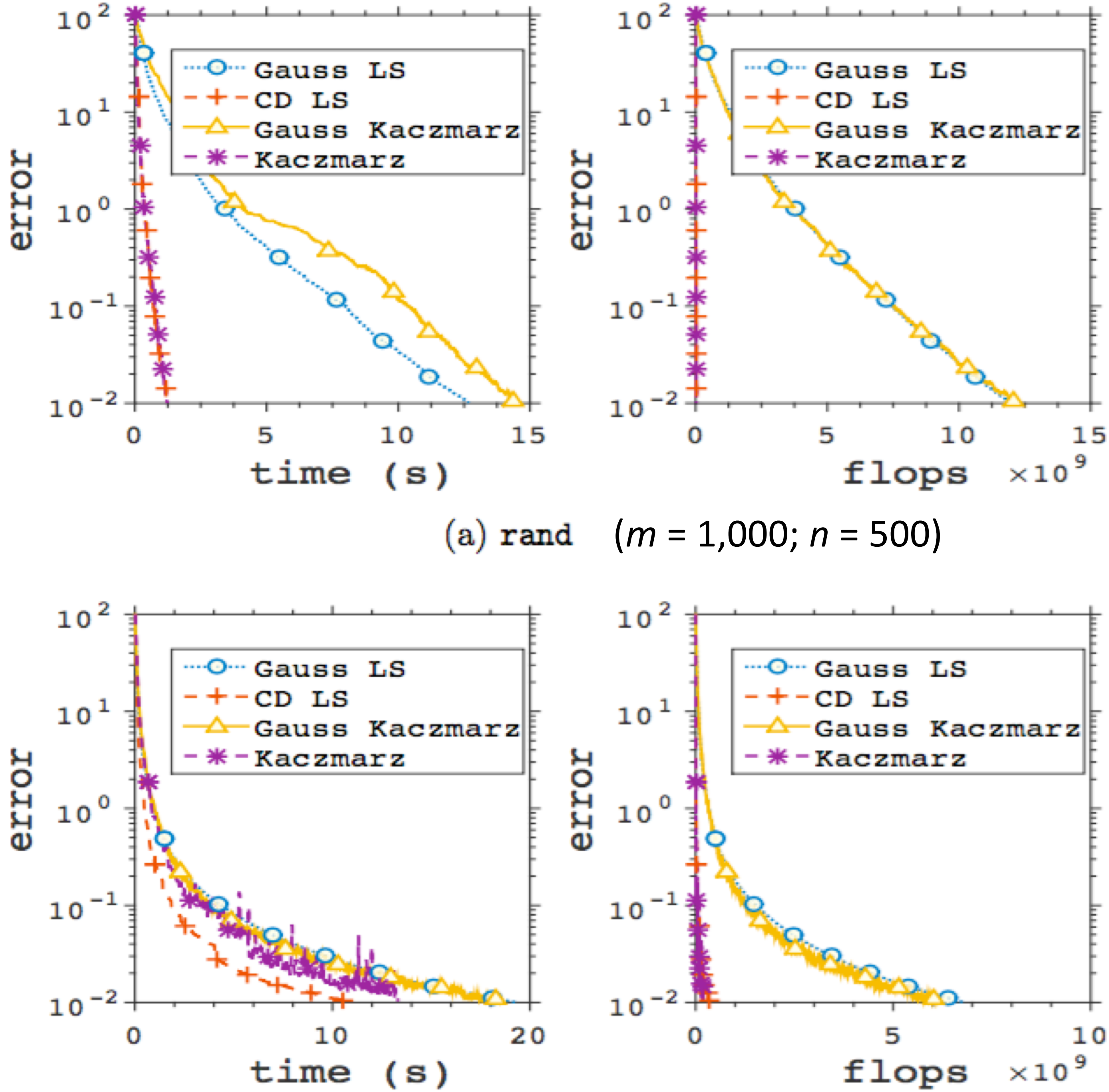
(c) **liver-disorders-ridge**



(d) **mushrooms-ridge-opt**

Experiments

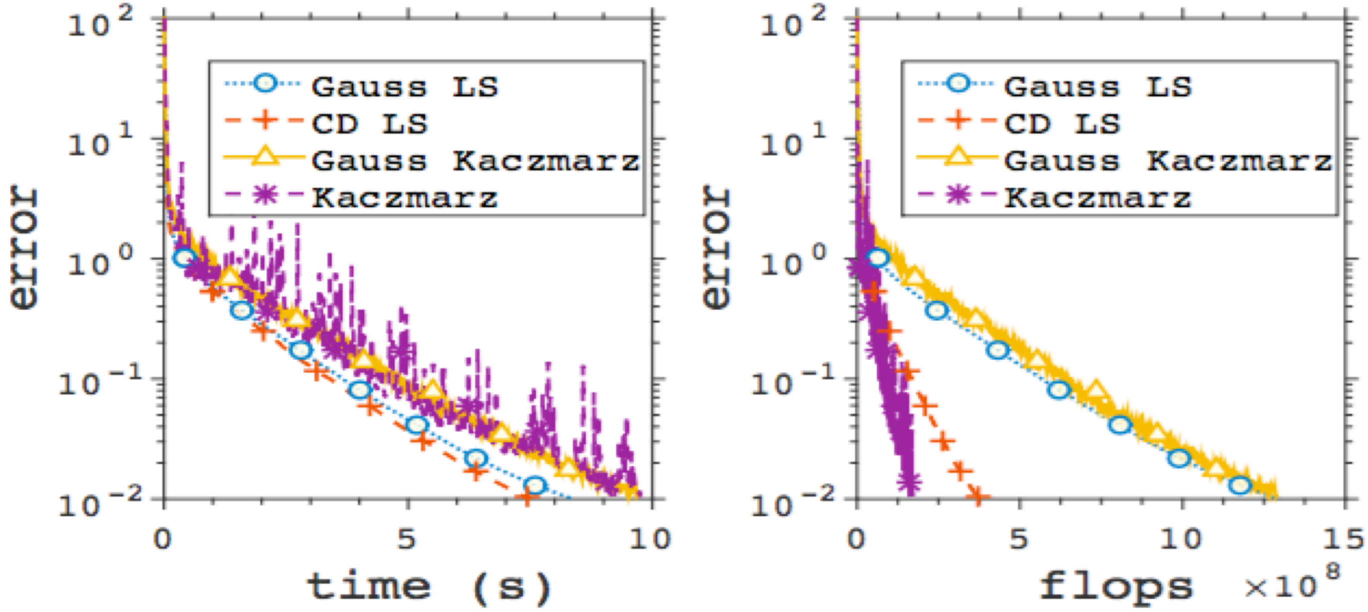
Synthetic data



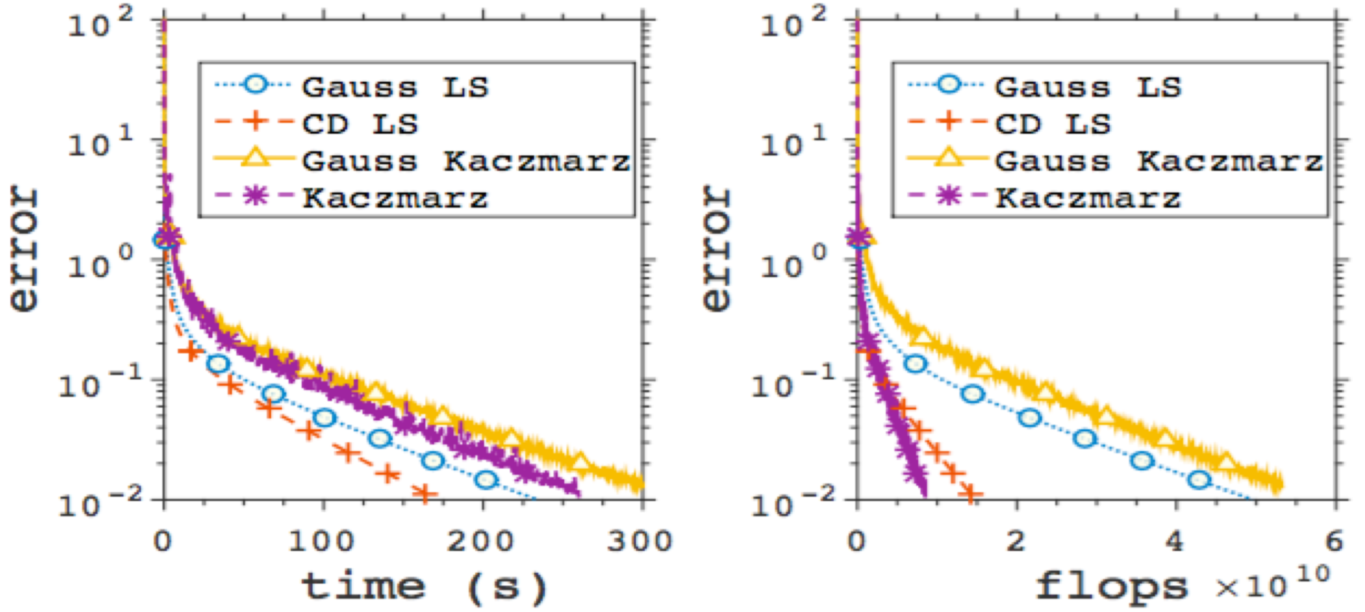
(a) rand (*m* = 1,000; *n* = 500)

(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)