EPSRC
Pioneering research and skills

Mathematics for Vast Digital Resources

London Mathematical Society
150 Years

NAIS Centre for Numerical Algorithms and Intelligent Software

amazon.com

Stochastic Dual Ascent

Linear Systems, Quasi-Newton Updates and Matrix Inversion

Peter Richtárik

(joint work with Robert M. Gower)

University of Edinburgh

Oberwolfach, March 8, 2016

Part I
Stochastic Dual Ascent for Linear Systems

Robert Mansel Gower (Edinburgh)

[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

[GR’15b]

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

The Problem

The Problem:
Solve a Linear System

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

Assumption 1

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

Optimization Formulation

Primal Problem

$B \succ 0$

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

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

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

Dual Problem

Unconstrained non-strongly concave quadratic maximization problem

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

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

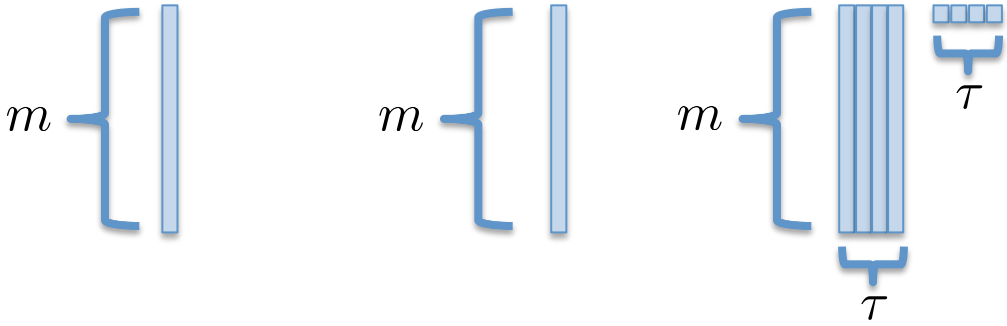
$$\underbrace{D(y^*) - D(y)}_{\text{Dual error (in function values)}} = \underbrace{\tfrac{1}{2}\|x(y) - x^*\|_B^2}_{\text{Primal error (in distance)}}$$

New Algorithm: Stochastic Dual Ascent (SDA)

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

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

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

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

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

Main Assumption

Assumption 2

The matrix

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

is nonsingular

Complexity of SDA

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

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

Theorem (GR'15b)

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

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

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

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

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

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:

$$\rho \leq 1 \text{ always}$$
$$\rho < 1 \text{ if Assumption 2 holds}$$

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

The Primal Iterates: 6 Equivalent Viewpoints

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

Corresponding primal iterates

Dual iterates produced by SDA

1. Relaxation Viewpoint “Sketch and Project”

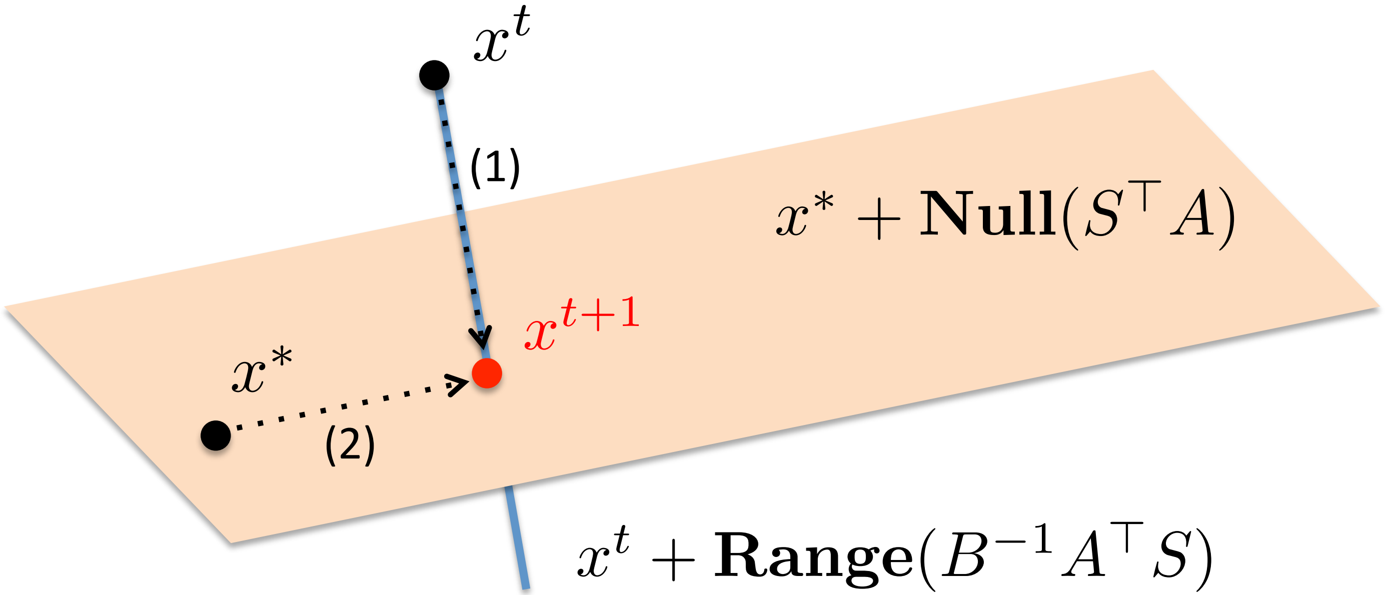
$$\begin{aligned} x^{t+1} \quad &= \quad \arg\min_{x\in\mathbb{R}^n} \|x - x^t\|_B^2 \\ \text{subject to} \quad &S^\top A x = S^\top b \end{aligned}$$

S = identity matrix → convergence in 1 step

2. Approximation Viewpoint “Constrain and Approximate”

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

3. Geometric Viewpoint "Random Intersect"



x^t
(1)
$x^* + \mathbf{Null}(S^\top A)$
x^{t+1}
x^*
(2)
$x^t + \mathbf{Range}(B^{-1}A^\top S)$

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

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

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

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”

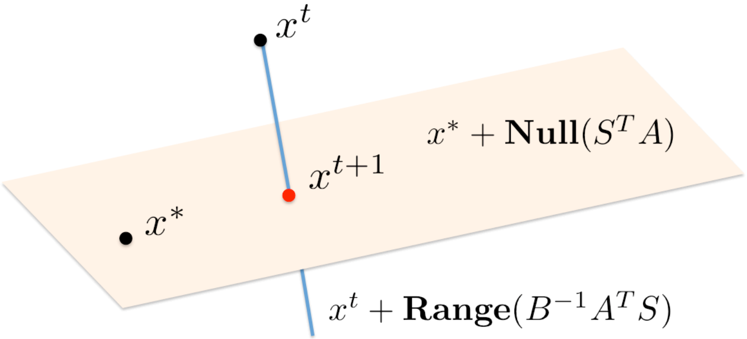
$$x^{t+1} = x^t - \overbrace{B^{-1}A^\top S(S^\top A B^{-1} A^\top S)^\dagger S^\top (Ax^t - b)}^{\text{Random Update Vector}}$$

Moore-Penrose pseudo-inverse

6. Analytic Viewpoint
"Random Fixed Point"

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

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


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

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

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

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*:

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

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

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

RK: Derivation and Rate

General Method

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

Special Choice of Parameters

$$B = I$$

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

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

Complexity Rate

$$p_i = \frac{\|A_{i:}\|^2}{\|A\|_F^2} \quad \Longrightarrow \quad \mathbf{E}\left[\|x^t - x^*\|_2^2\right] \le \left(1 - \frac{\lambda_{\min}\left(A^T A\right)}{\|A\|_F^2}\right)^t \|x^0 - x^*\|_2^2$$

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: 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

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

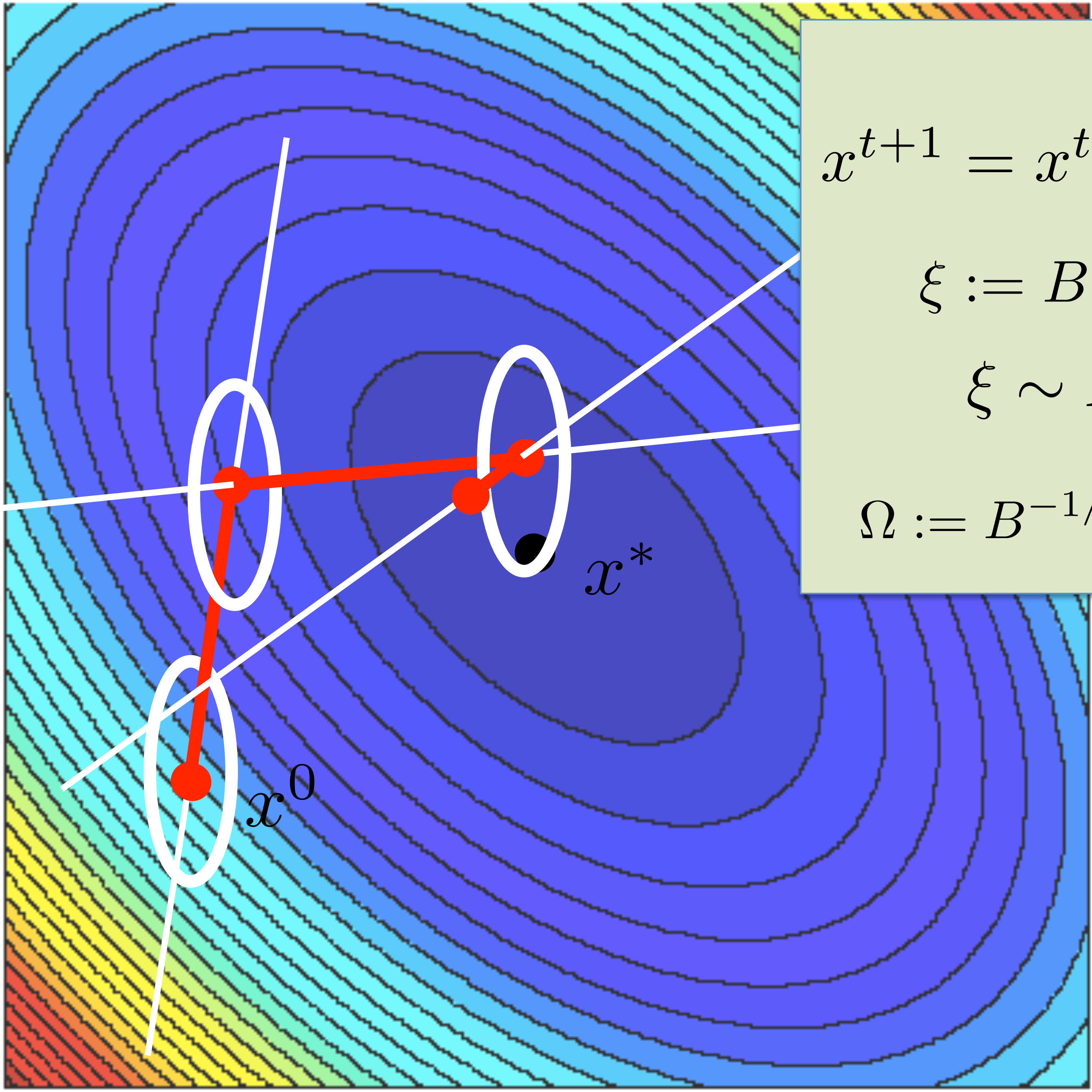
Positive definite covariance matrix

➡

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

Complexity Rate

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



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

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

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

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

Gaussian Descent: The Rate

Lemma [GR'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

Gaussian Descent: Further Reading

PDF Adobe

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

PDF Adobe

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

PDF Adobe

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

Special Case: Randomized Coordinate Descent

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

$$B = A$$

$\mathbf{P}(S=e^i)=p_i$ → $S = e^i$

→

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

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: “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*:

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

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

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

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

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

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

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

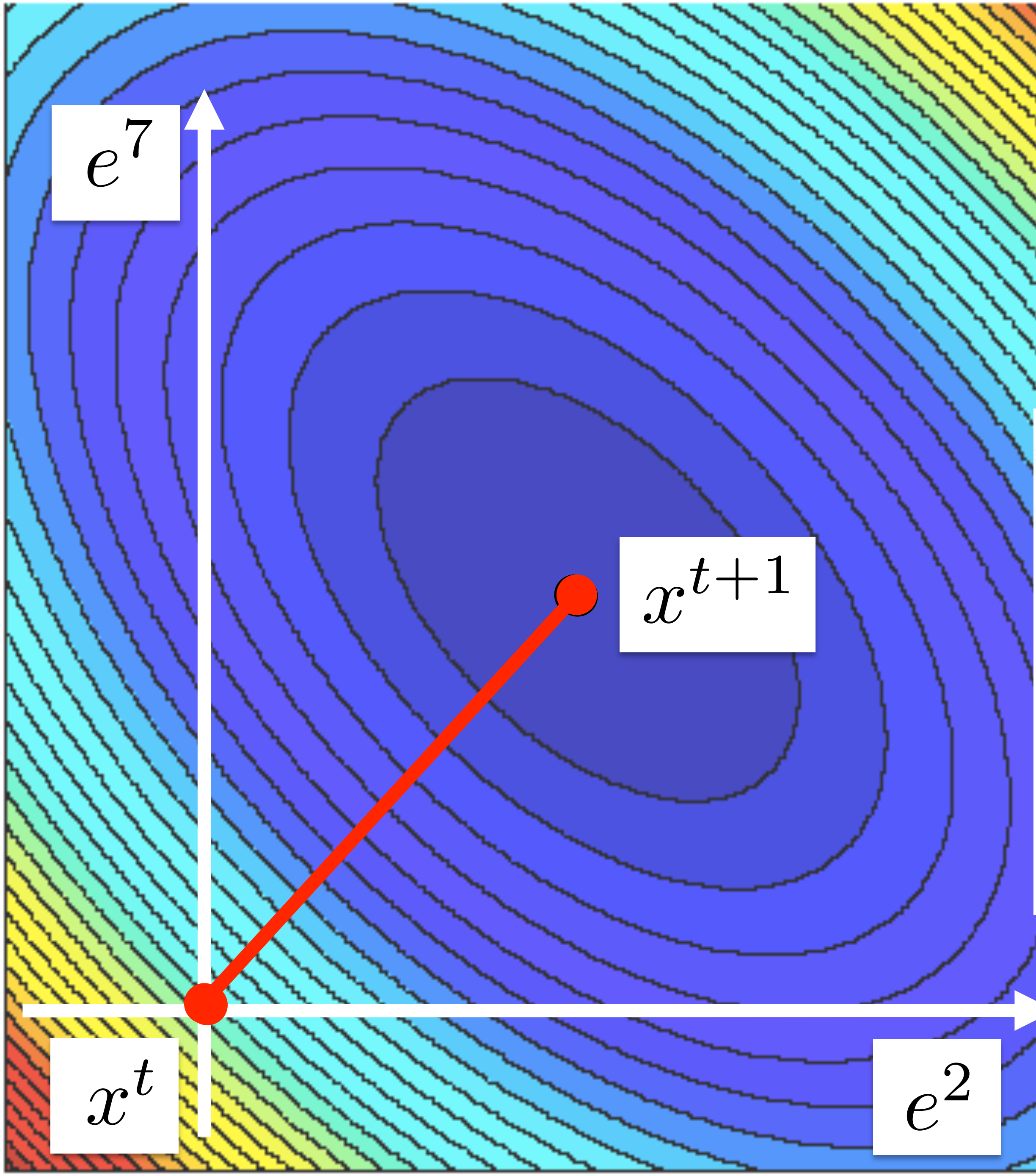
$$x^{t+1} = x^t - \boxed{B^{-1}A^TS}\boxed{(S^TAB^{-1}A^TS)^\dagger}\boxed{S^T(Ax^t - b)}$$

Special Choice of Parameters $B = A$

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

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

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

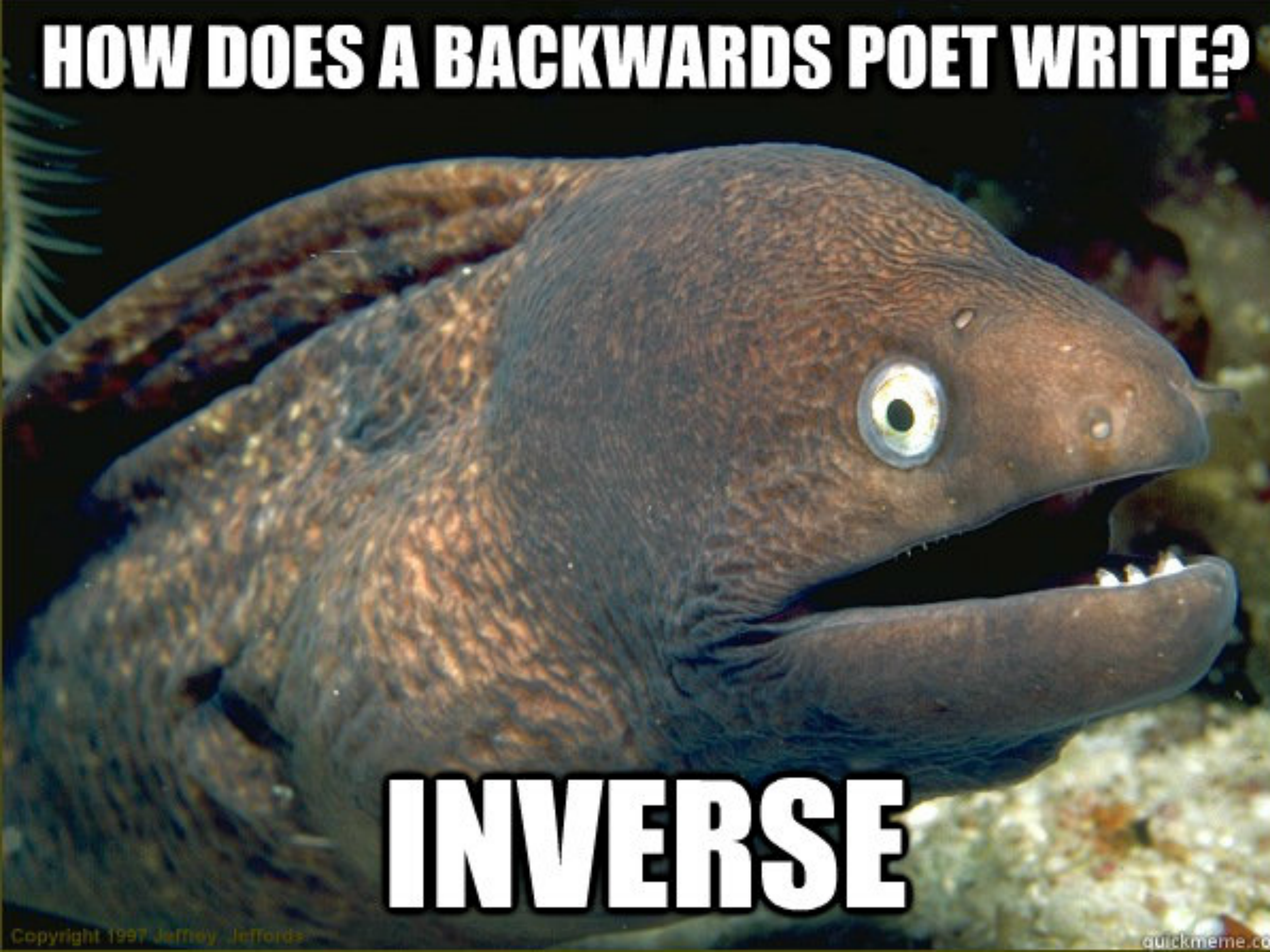


$$C = \{2, 7\}$$
$$|C| = 2$$

Always label your axes

Summary: Linear Systems

- SDA:
  - A new class of randomized optimization algorithms
  - Extremely versatile
    - Works for almost any random *S*
    - Get several existing algorithms in special cases (RK, RCD, RN, RBK)
    - Get many new algorithms in special cases
  - Linear convergence despite lack of strong concavity
  - RK in the primal = RCD in the dual
- Did not talk about:
  - Randomized gossip
  - Distributed variant
  - Optimal sampling via SDP
  - Experiments



Copyright 1997 Jeffrey Jeffords

Part II
Stochastic Dual Ascent for Matrix Inversion

Robert Mansel Gower (Edinburgh)

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

The Problem: Invert a Matrix

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

Identity matrix

$$n\left\{\overbrace{A}^{n}\right. X = I$$

Assumption 1 Matrix *A* is invertible

Inverting Symmetric Matrices

1. Sketch and Project

$$\|X\|_{F(B)} := \sqrt{\mathbf{Tr}(X^\top B X B)}$$

$$X^{t+1} = \arg \min_{X \in \mathbb{R}^{n \times n}} \|X - X^t\|^2_{F(B)}$$

$$\text{subject to} \quad S^\top A X = S^\top, \quad X = X^\top$$

- Quasi-Newton updates are of this form: *S* = deterministic column vector
- We get **randomized block** version of quasi-Newton updates!
- **Randomized quasi-Newton updates are linearly convergent matrix inversion methods**
- Interpretation**: Gaussian Inference** (Henning, 2015)

Donald Goldfarb. **A Family of Variable-Metric Methods Derived by Variational Means.** *Mathematics of Computation* 24(109), 1970

Gaussian Inference

Philipp Henning
Probabilistic Interpretation of Linear Solvers
SIAM Journal on Optimization 25(1):234-260, 2015

The new iterate X_{k+1} can be interpreted as

- the mean of a posterior distribution
- under a Gaussian prior with mean X_k and
- noiseless (and random) linear observation of A^{-1}

Randomized QN Updates

B	**Equation**	**Method**
I	$AX = I$	Powel-Symmetric-Broyden (PSB)
A^{-1}	$XA^{-1} = I$	Davidon-Fletcher-Powell (DFP)
A	$AX = I$	Broyden-Fletcher-Goldfarb-Shanno (BFGS)

- All these QN methods arise as **special cases of our framework**
- All are **linearly convergent,** with explicit convergence rates
- We also recover **non-symmetric updates** such as **Bad Broyden** and **Good Broyden**
- We get **block versions**
- We get randomized versions of **new QN updates**

2. Constrain and Approximate

$$X^{t+1} = \arg\min_{X\in\mathbb{R}^{n\times n}} \|X - A^{-1}\|^2_{F(B)}$$

$$\text{s.t.} \quad X = X^t + \Lambda S^\top A B^{-1} + B^{-1} A^\top S \Lambda^\top$$

$$\Lambda \in \mathbb{R}^{n\times\tau} \text{ is free}$$

New formulation even for standard QN methods

Randomized BFGS: $B = A,\ \tau = 1$

$$X^{t+1} = \arg\min_{X\in\mathbb{R}^{n\times n}} \|X - A^{-1}\|^2_{F(A)} = \boxed{\|AX - I\|^2_F}$$

$$\text{s.t.} \quad X = X^t + \boxed{\lambda S^\top + S\lambda^\top}$$

$$\lambda \in \mathbb{R}^n \text{ is free}$$

RBFGS performs "best" symmetric rank-2 update

4. Random Update

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

$$\begin{aligned} X^{t+1} = X^t - (X^tA - I)HAB^{-1} \\ + B^{-1}AH(AX^t - I)(AHAB^{-1} - I) \end{aligned}$$

6. Random Fixed Point

$$\begin{aligned} X^{t+1} - A^{-1} = \\ (I - B^{-1}A^\top HA)(X^t - A^{-1})(I - AHA^\top B^{-1}) \end{aligned}$$

Complexity / Convergence

Theorem [GR'16]

$$\|M\|_B := \|B^{1/2} M B^{1/2}\|_2$$

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

2 $$\mathbf{E}[H] \succ 0 \quad \Rightarrow \quad \rho < 1$$

$$\mathbf{E}\left[\|X^t - A^{-1}\|_{F(B)}^2\right] \leq \rho^t \|X^0 - A^{-1}\|_{F(B)}^2$$

Summary: Matrix Inversion

- Block version of QN updates
- New points of view (constrain and approximate, ...)
- New link between QN and approx. inverse preconditioning
- First time randomized QN updates are proposed
- First stochastic method for matrix inversion (with complexity bounds)?
- Linear convergence under weak assumptions
- Did not talk about:
  - Nonsymmetric variants
  - Theoretical bounds for discretely distributed S
  - Adaptive randomized BFGS
  - Limited memory and factored implementations
  - Experiments (Newton-Schultz; MinRes)
  - Use in empirical risk minimization (Gower, Goldfarb & R. '16)
  - Extension to the computation of a pseudoinverse

The End