

SDNA

Stochastic Dual Newton Ascent
for Empirical Risk Minimization



LEHIGH
UNIVERSITY®

Martin Takac

MOPTA: modeling and optimization: theory and applications
20th of July, 2015

Coauthors



Zheng Qu
(Edinburgh)



Peter Richtarik
(Edinburgh)



Olivier Fercoq
(Telecom ParisTech)



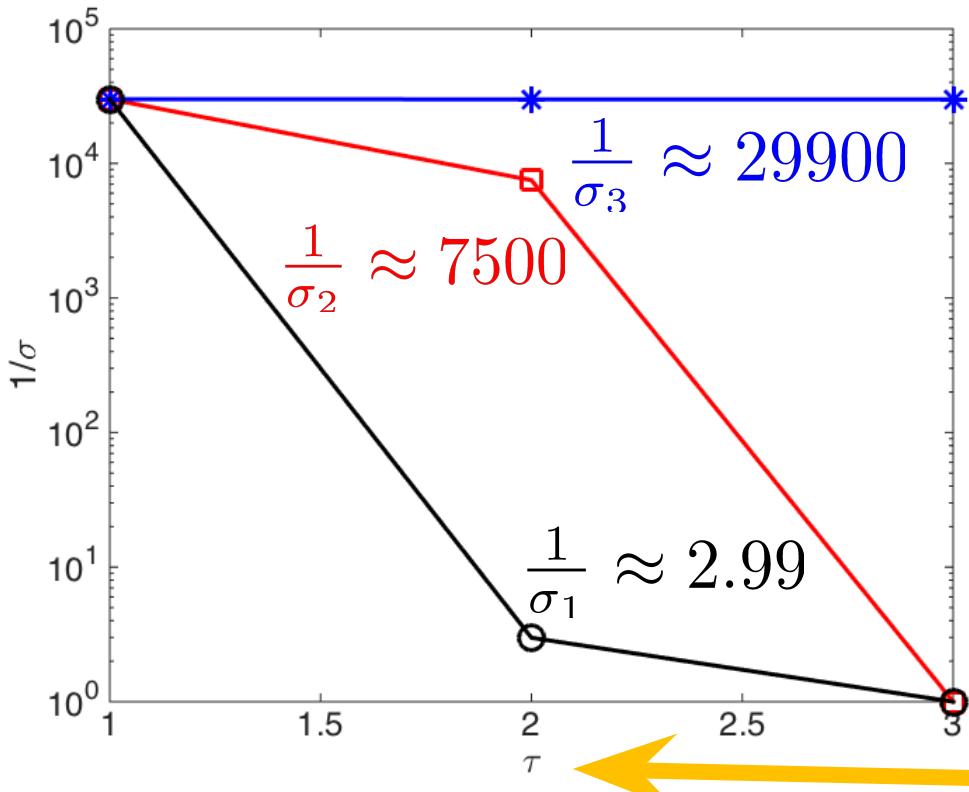
Zheng Qu, Peter Richtarik, M.T. and Olivier Fercoq
SDNA: Stochastic Dual Newton Ascent for empirical risk minimization
arXiv:1502.02268, 2015

Part A

Motivation

3D Quadratic

$$\min_{x \in \mathbb{R}^3} \left[f(x) = \frac{1}{2} x^T \mathbf{M} x + b^\top x + c \right]$$

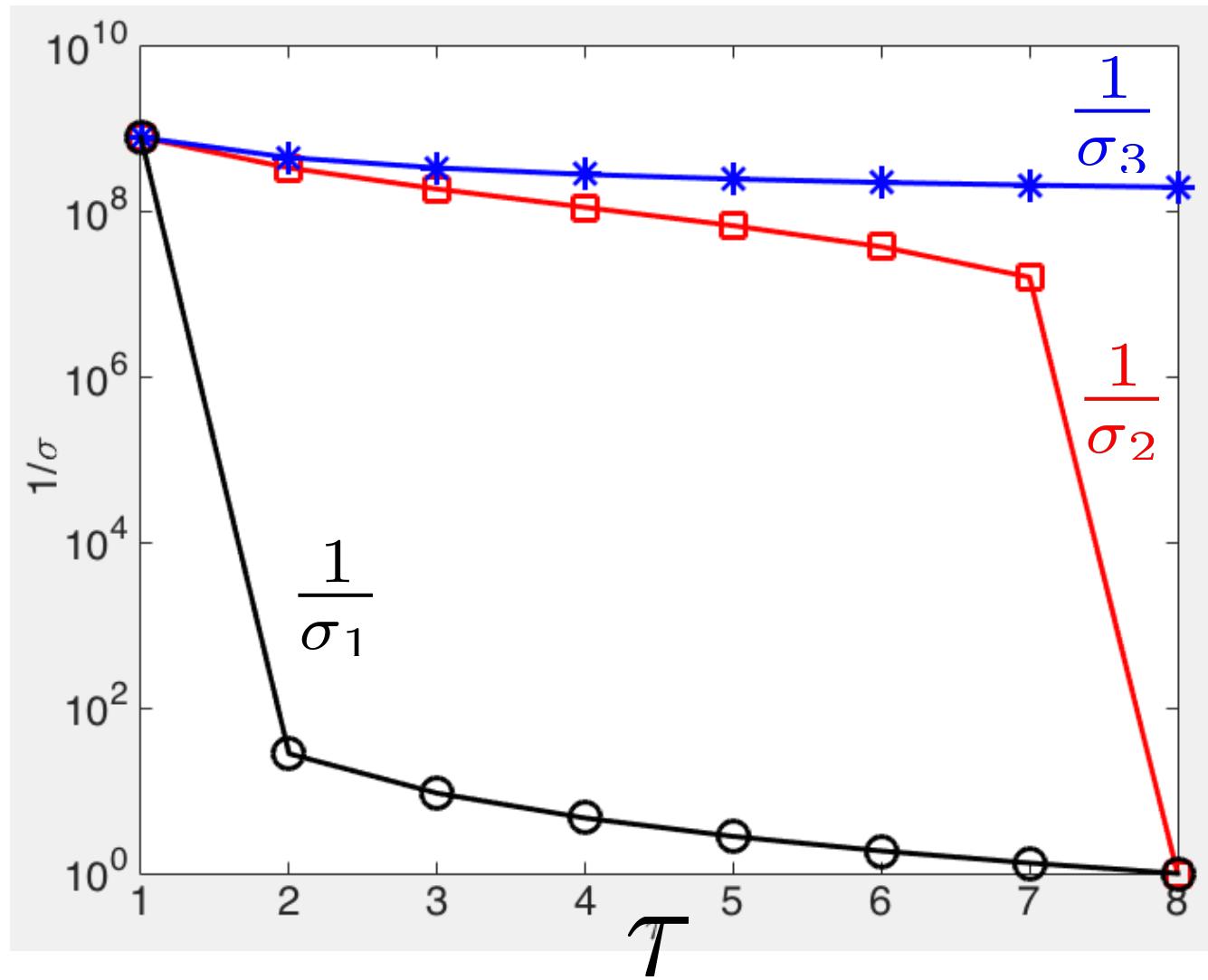


$$\mathbf{M} = \begin{pmatrix} 1.0000 & 0.9900 & 0.9999 \\ 0.9900 & 1.0000 & 0.9900 \\ 0.9999 & 0.9900 & 1.0000 \end{pmatrix}$$

condition number $\approx 3 \times 10^4$

$$\tau = \mathbb{E} [|S_k|]$$

8D Quadratic



Part B

Three Methods

The Problem & Assumptions

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

Strong convexity

Large dimension

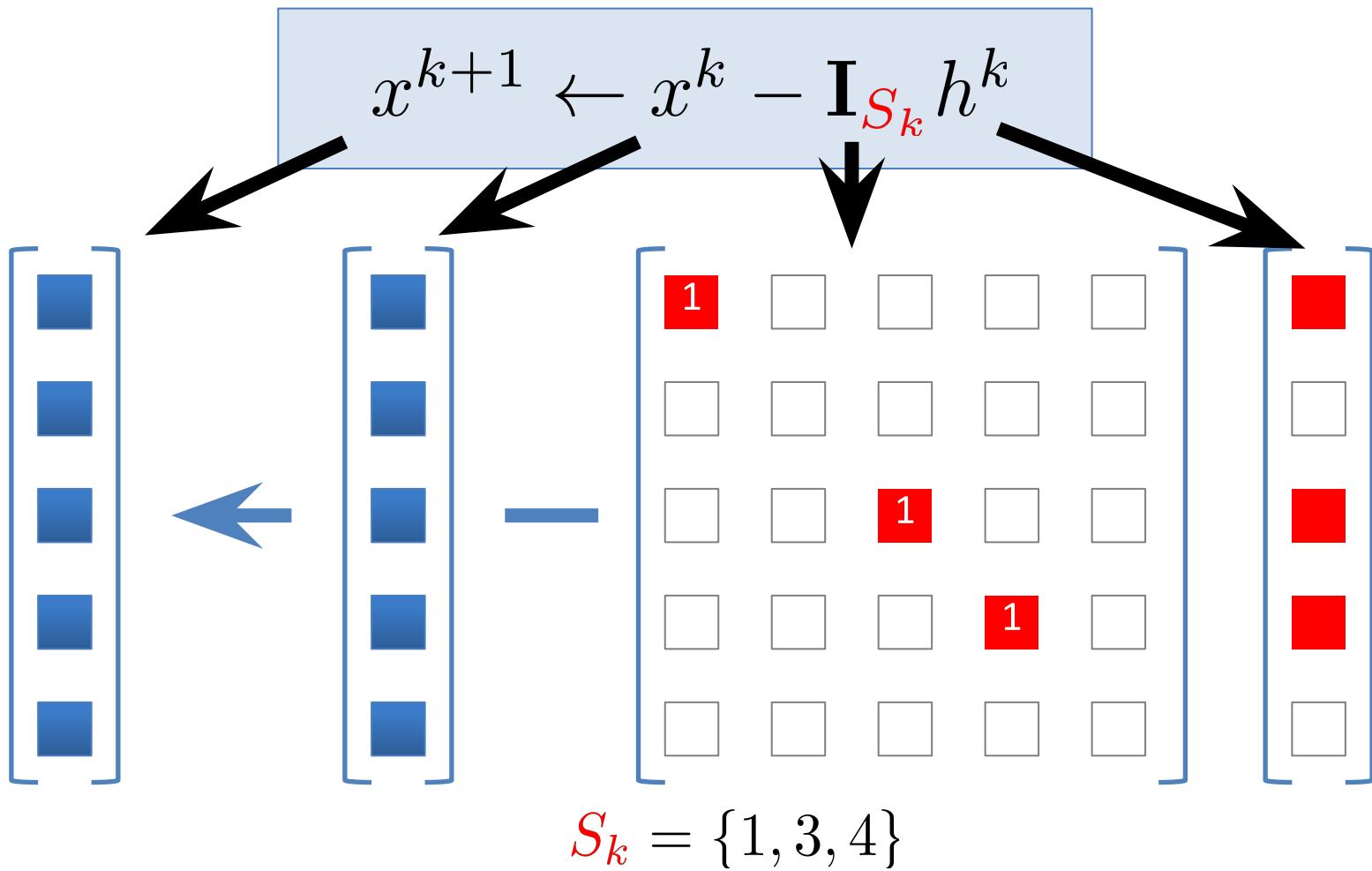
$$f(x) + (\nabla f(x))^\top h + \frac{1}{2} h^\top \mathbf{G} h \leq f(x + h)$$

Positive definite matrices

Smoothness

$$f(x + h) \leq f(x) + (\nabla f(x))^\top h + \frac{1}{2} h^\top \mathbf{M} h$$

Randomized Update



Method 3



Peter Richtarik and M.T.
On optimal probabilities in stochastic coordinate descent methods
In NIPS Workshop on Optimization for Machine Learning, 2013
Optimization Letters 2015 (arXiv:1310.3438)

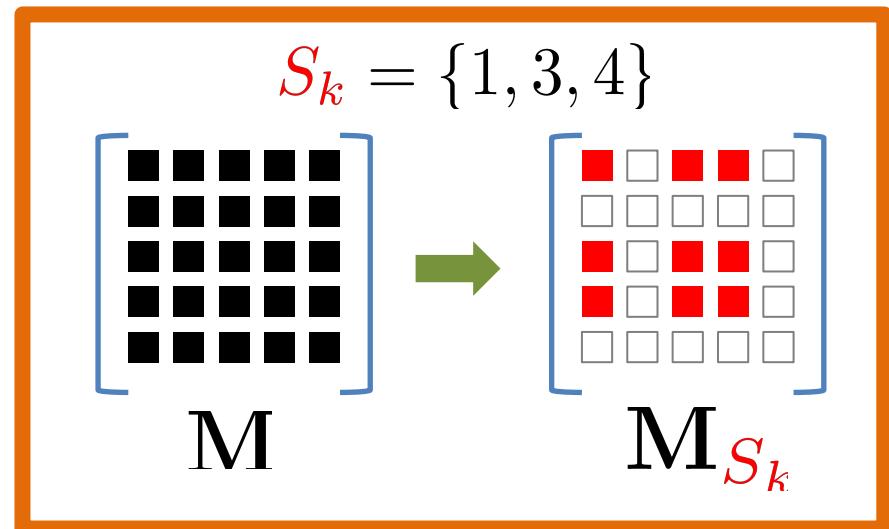
Key Inequality

$$f(x + h) \leq f(x) + (\nabla f(x))^\top h + \frac{1}{2} h^\top \mathbf{M} h$$



$$x \leftarrow x^k$$

$$h \leftarrow \mathbf{I}_{S_k} h = \sum_{i \in S_k} h_i e_i$$



$$f(x^k + \mathbf{I}_{S_k} h) \leq f(x^k) + (\nabla f(x^k))^\top (\mathbf{I}_{S_k} h) + \frac{1}{2} (\mathbf{I}_{S_k} h)^\top \mathbf{M}_{S_k} (\mathbf{I}_{S_k} h)$$



Method 3

$$f(x^k + \mathbf{I}_{S_k} h) \leq f(x^k) + (\mathbf{I}_{S_k} \nabla f(x^k))^{\top} h + \frac{1}{2} h^{\top} \mathbf{M}_{S_k} h$$

1. take expectations on both sides



$$p_i = \mathbb{P}(i \in S_k)$$

$$\mathbb{E}[f(x^k + \mathbf{I}_{S_k} h)] \leq f(x^k) + (\text{Diag}(p) \nabla f(x^k))^{\top} h + \frac{1}{2} h^{\top} \mathbb{E}[\mathbf{M}_{S_k}] h$$

2. diagonalize



$$\mathbb{E}[\mathbf{M}_{S_k}] \preceq \text{Diag}(p \circ v)$$

$$\mathbb{E}[f(x^k + \mathbf{I}_{S_k} h)] \leq f(x^k) + (\text{Diag}(p) \nabla f(x^k))^{\top} h + \frac{1}{2} h^{\top} \text{Diag}(p \circ v) h$$

3. minimize the RHS in h



$$x^{k+1} \leftarrow x^k - \mathbf{I}_{S_k} (\text{Diag}(v))^{-1} \nabla f(x^k)$$

Method 3

i.i.d. with arbitrary distribution

Choose a random set S_k of coordinates

For $i \in S_k$ do

$$x_i^{k+1} \leftarrow x_i^k - \frac{1}{v_i} (\nabla f(x^k))^{\top} e_i$$

For $i \notin S_k$ do

$$x_i^{k+1} \leftarrow x_i^k$$

Convergence

Theorem (RT'13)

$$\mathbb{E}[f(x^k) - f(x^*)] \leq (1 - \sigma_3)^k (f(x^0) - f(x^*))$$



$$\sigma_3 = \lambda_{\min} \left(\mathbf{G}^{1/2} \mathbf{Diag}(p \circ v^{-1}) \mathbf{G}^{1/2} \right)$$

Alternative formulation:

$$k \geq \frac{1}{\sigma_3} \log \left(\frac{f(x^0) - f(x^*)}{\epsilon} \right) \quad \Rightarrow \quad \mathbb{E}[f(x^k) - f(x^*)] \leq \epsilon$$

Uniform vs Optimal Sampling

Special case:

$$\mathbf{G} = \lambda \mathbf{I} \quad \Rightarrow \quad \frac{1}{\sigma_3} = \max_i \frac{v_i}{\lambda p_i}$$

$$\mathbb{P}(|S_k| = 1) = 1 \quad \Rightarrow \quad v_i = \mathbf{M}_{ii}$$

$$p_i = \frac{1}{n}$$



$$\frac{1}{\sigma_3} = \frac{n \max_i \mathbf{M}_{ii}}{\lambda}$$

$$p_i = \frac{\mathbf{M}_{ii}}{\sum_i \mathbf{M}_{ii}}$$



$$\frac{1}{\sigma_3} = \frac{\sum_{i=1}^n \mathbf{M}_{ii}}{\lambda}$$

Method 2

Method 2

$$f(x^k + \mathbf{I}_{S_k} h) \leq f(x^k) + (\mathbf{I}_{S_k} \nabla f(x^k))^{\top} h + \frac{1}{2} h^{\top} \mathbf{M}_{S_k} h$$



1. take expectations on both sides

$$\mathbb{E}[f(x^k + \mathbf{I}_{S_k} h)] \leq f(x^k) + (\text{Diag}(p) \nabla f(x^k))^{\top} h + \frac{1}{2} h^{\top} \mathbb{E}[\mathbf{M}_{S_k}] h$$



$$p_i = \mathbb{P}(i \in S_k)$$

2. minimize the RHS in h

$$x^{k+1} \leftarrow x^k - \mathbf{I}_{S_k} (\mathbb{E}[\mathbf{M}_{S_k}])^{-1} \text{Diag}(p) \nabla f(x^k)$$

Convergence of Method 2

Theorem (QRTF'15)

$$\mathbb{E}[f(x^k) - f(x^*)] \leq (1 - \sigma_2)^k (f(x^0) - f(x^*))$$

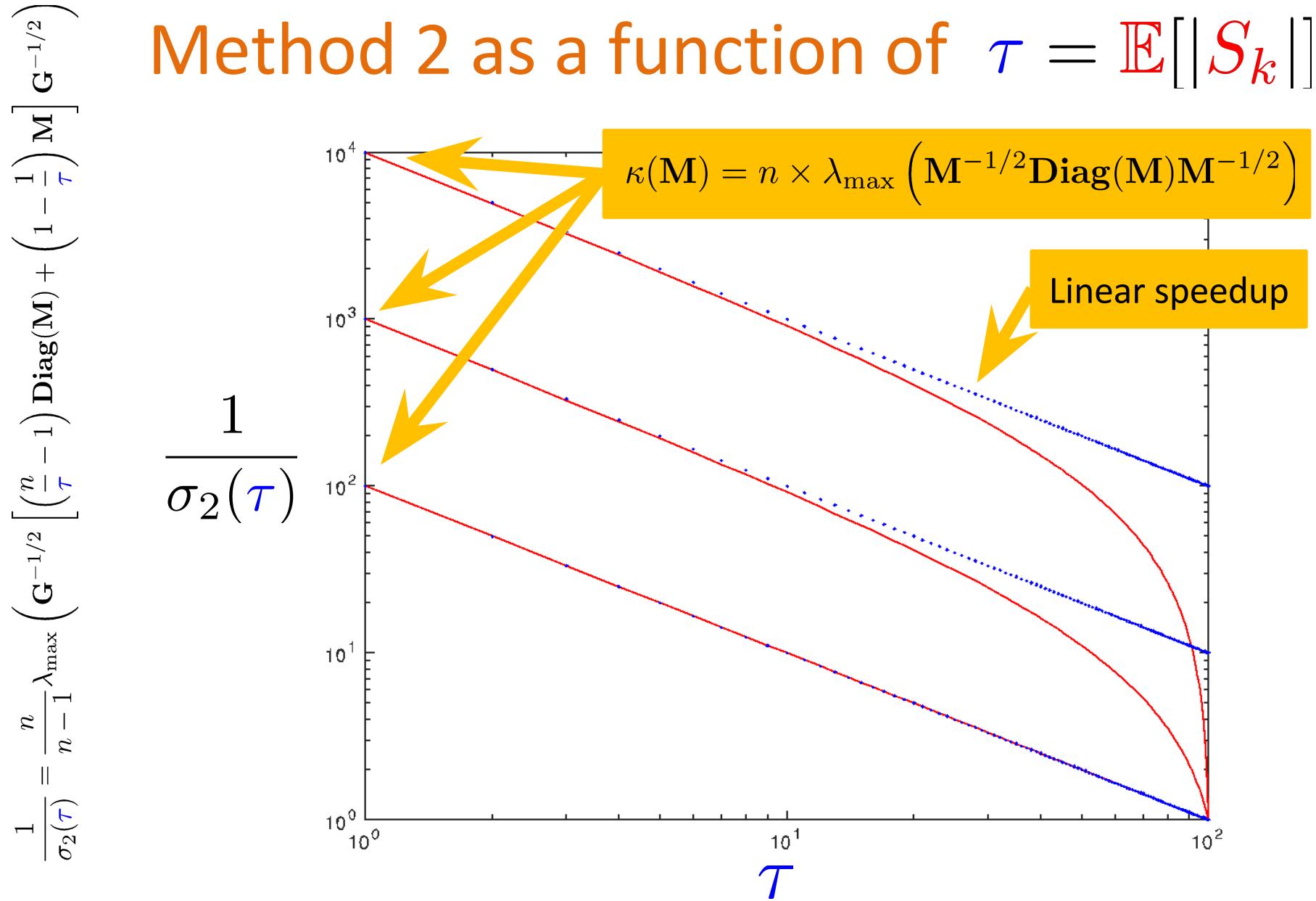


$$\sigma_2 = \lambda_{\min} \left(\mathbf{G}^{1/2} \mathbf{Diag}(p) (\mathbb{E} [\mathbf{M}_{S_k}])^{-1} \mathbf{Diag}(p) \mathbf{G}^{1/2} \right)$$

Alternative formulation:

$$k \geq \frac{1}{\sigma_2} \log \left(\frac{f(x^0) - f(x^*)}{\epsilon} \right) \Rightarrow \mathbb{E}[f(x^k) - f(x^*)] \leq \epsilon$$

Leading term in the complexity of Method 2 as a function of $\tau = \mathbb{E}[|S_k|]$



Method 1

Randomized Newton

Method

Method 1: Randomized Newton

$$f(x^k + \mathbf{I}_{S_k} h) \leq f(x^k) + (\mathbf{I}_{S_k} \nabla f(x^k))^{\top} h + \frac{1}{2} h^{\top} \mathbf{M}_{S_k} h$$



minimize the RHS in h

$$x^{k+1} \leftarrow x^k - (\mathbf{M}_{S_k})^{-1} \nabla f(x^k)$$

$$S_k = \{1, 3, 4\}$$

$$\mathbf{M}_{S_k} = \begin{bmatrix} \textcolor{red}{\square} & \textcolor{white}{\square} & \textcolor{red}{\square} & \textcolor{red}{\square} \\ \textcolor{white}{\square} & \textcolor{white}{\square} & \textcolor{white}{\square} & \textcolor{white}{\square} \\ \textcolor{red}{\square} & \textcolor{white}{\square} & \textcolor{red}{\square} & \textcolor{red}{\square} \\ \textcolor{red}{\square} & \textcolor{white}{\square} & \textcolor{red}{\square} & \textcolor{red}{\square} \\ \textcolor{white}{\square} & \textcolor{white}{\square} & \textcolor{white}{\square} & \textcolor{white}{\square} \end{bmatrix}$$

$$(\mathbf{M}_{S_k})^{-1} = \begin{bmatrix} \textcolor{green}{\square} & \textcolor{white}{\square} & \textcolor{green}{\square} & \textcolor{green}{\square} \\ \textcolor{white}{\square} & \textcolor{white}{\square} & \textcolor{white}{\square} & \textcolor{white}{\square} \\ \textcolor{green}{\square} & \textcolor{white}{\square} & \textcolor{green}{\square} & \textcolor{green}{\square} \\ \textcolor{green}{\square} & \textcolor{white}{\square} & \textcolor{green}{\square} & \textcolor{green}{\square} \\ \textcolor{white}{\square} & \textcolor{white}{\square} & \textcolor{white}{\square} & \textcolor{white}{\square} \end{bmatrix}$$

$$\mathbf{I}_{S_k} = \begin{bmatrix} 1 & \textcolor{white}{\square} & \textcolor{white}{\square} & \textcolor{white}{\square} & \textcolor{white}{\square} \\ \textcolor{white}{\square} & \textcolor{white}{\square} & \textcolor{white}{\square} & \textcolor{white}{\square} & \textcolor{white}{\square} \\ \textcolor{white}{\square} & \textcolor{white}{\square} & \textcolor{white}{\square} & \textcolor{white}{\square} & \textcolor{white}{\square} \\ \textcolor{white}{\square} & \textcolor{white}{\square} & \textcolor{white}{\square} & \textcolor{white}{\square} & \textcolor{white}{\square} \\ \textcolor{white}{\square} & \textcolor{white}{\square} & \textcolor{white}{\square} & \textcolor{white}{\square} & \textcolor{white}{\square} \end{bmatrix}$$

$$\mathbf{M}_{S_k}$$

$$(\mathbf{M}_{S_k})^{-1}$$

$$\mathbf{I}_{S_k}$$

Convergence of Method 1 (Randomized Newton Method)

Theorem (QRTF'15)

$$\mathbb{E}[f(x^k) - f(x^*)] \leq (1 - \sigma_1)^k (f(x^0) - f(x^*))$$


$$\sigma_1 = \lambda_{\min} \left(\mathbf{G}^{1/2} \mathbb{E} \left[(\mathbf{M}_{S_k})^{-1} \right] \mathbf{G}^{1/2} \right)$$

Alternative formulation:

$$k \geq \frac{1}{\sigma_1} \log \left(\frac{f(x^0) - f(x^*)}{\epsilon} \right) \Rightarrow \mathbb{E}[f(x^k) - f(x^*)] \leq \epsilon$$

Three Convergence Rates

3 Convergence Rates

Theorem (QRTF'15)

$$0 < \sigma_3 \leq \sigma_2 \leq \sigma_1 \leq 1$$

$$\sigma_1(1) = \sigma_2(1) = \sigma_3(1)$$

$$\sigma_1(n) = \sigma_2(n) = \frac{1}{\kappa_f}$$

$$\sigma_2(\tau) \geq \tau \sigma_2(1)$$

$$\sigma_3(\tau) \leq \tau \sigma_2(1)$$

$$\sigma_1(\tau) \leq \frac{\tau}{n \kappa_f}$$

$$\kappa_f = \lambda_{\max} \left(\mathbf{G}^{-1/2} \mathbf{M} \mathbf{G}^{-1/2} \right)$$

The 3 methods coincide if we update 1 coordinate at a time

Methods 2 and 3 coincide if we update all coordinates

Randomized Newton:
superlinear speedup

Randomized Coordinate Descent:
sublinear speedup

Lower bound on complexity

Part C

Empirical Risk

Minimization

Primal Problem

$$|\phi'_i(a) - \phi'_i(b)| \leq \frac{1}{\gamma} |a - b| \quad \forall a, b \in \mathbb{R}$$

$P = \text{Regularized Empirical Risk}$

$1/\gamma$ - smooth & convex
functions (“risk”)

positive
regularization
parameter

$$\min_{w \in \mathbb{R}^d} P(w) \equiv \frac{1}{n} \sum_{i=1}^n \phi_i(A_i^\top w) + \lambda g(w)$$

$w = \text{linear predictor}$

n data vectors
 (“examples”)

$d = \# \text{ features}$
(parameters)

1 - strongly convex function (“regularizer”)

$$g(w) \geq g(w') + \langle \nabla g(w'), w - w' \rangle + \frac{1}{2} \|w - w'\|^2, \quad w, w' \in \mathbb{R}^d$$

Dual Problem

n dual variables: as many as
examples in the primal

$\in \mathbb{R}^d$

$$\max_{\alpha \in \mathbb{R}^n} \left[D(\alpha) \equiv -\lambda g^* \left(\frac{1}{\lambda n} \sum_{i=1}^n A_i \alpha_i \right) - \frac{1}{n} \sum_{i=1}^n \phi_i^*(-\alpha_i) \right]$$

1 – smooth & convex

γ - strongly convex

$$g^*(w') = \max_{w \in \mathbb{R}^d} \{(w')^\top w - g(w)\}$$

$$\phi_i^*(a') = \max_{a \in \mathbb{R}^m} \{(a')^\top a - \phi_i(a)\}$$

SDNA

Initialization:

$$\alpha^0 \in \mathbb{R}^n \quad \bar{\alpha}^0 = \frac{1}{\lambda n} \mathbf{A} \alpha^0$$

Iterate:

Primal update: $w^k = \nabla g^*(\bar{\alpha}^k)$

Generate a random set S_k

Compute:

$$h^k = \arg \min_{h \in \mathbb{R}^n} ((\mathbf{A}^\top w^k)_{S_k})^\top h + \frac{1}{2} h^\top \mathbf{X}_{S_k} h + \sum_{i \in S_k} \phi_i^*(-\alpha_i^k - h_i)$$

Dual update: $\alpha^{k+1} \leftarrow \alpha^k + \sum_{i \in S_k} h_i^k e_i$

Maintain average: $\bar{\alpha}^{k+1} = \bar{\alpha}^k + \frac{1}{\lambda n} \sum_{i \in S_k} h_i^k A_i$

$$\mathbf{A} = [A_1, A_2, \dots, A_n] \in \mathbb{R}^{d \times n}$$

$$\mathbf{X} = \frac{1}{\lambda n} \mathbf{A}^\top \mathbf{A}$$

Convergence of SDNA

Theorem (QRTF'15)

Better rate than SDCA

Assume that S_k is uniform

$$\mathbb{E}[P(w^k) - D(\alpha^k)] \leq (1 - \sigma_1^{prox})^k \frac{D(\alpha^*) - D(\alpha^0)}{\theta(S_k)}$$

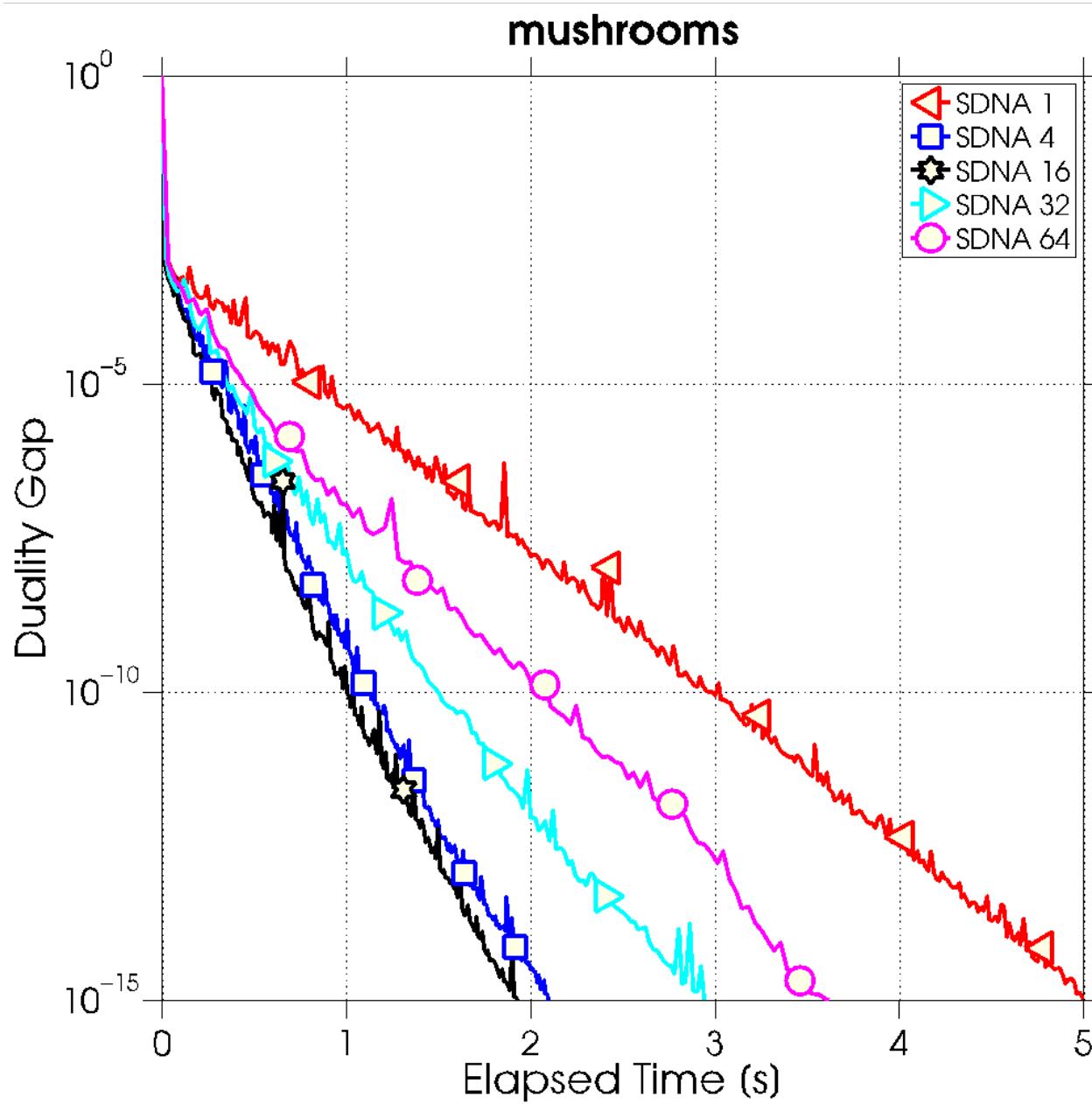
Expected duality gap
after k iterations

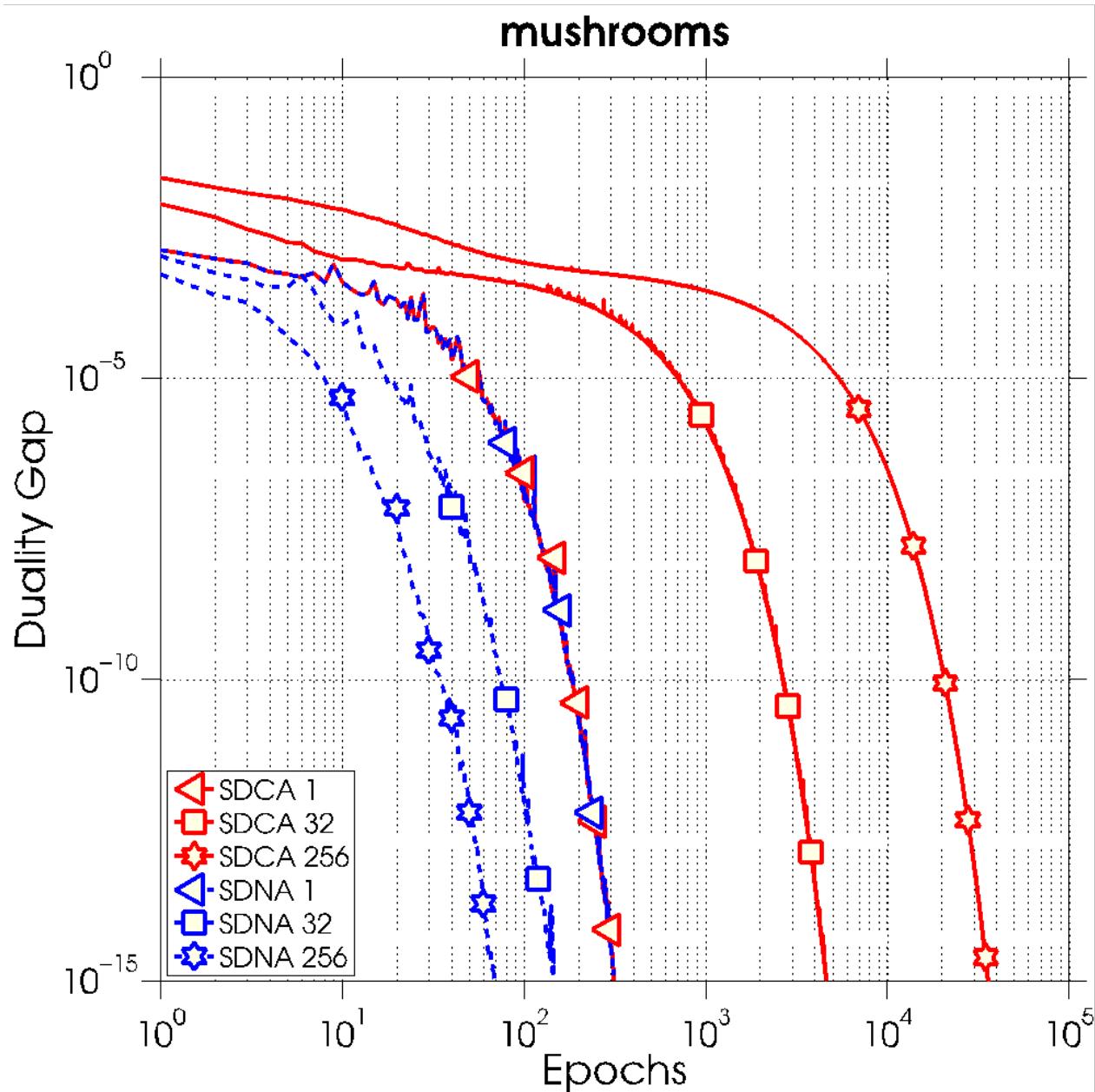
$$\sigma_1^{prox} = \frac{\tau}{n} \min\{1, s_1\}$$

$$\tau = \mathbb{E}[|S_k|] \quad s_1 = \lambda_{\min} \left[\left(\frac{1}{\tau \gamma \lambda} \mathbb{E}[(\mathbf{A}^\top \mathbf{A})_{S_k}] + \mathbf{I} \right)^{-1} \right]$$

Real Dataset: mushrooms

$d = 112$ $n = 8,124$

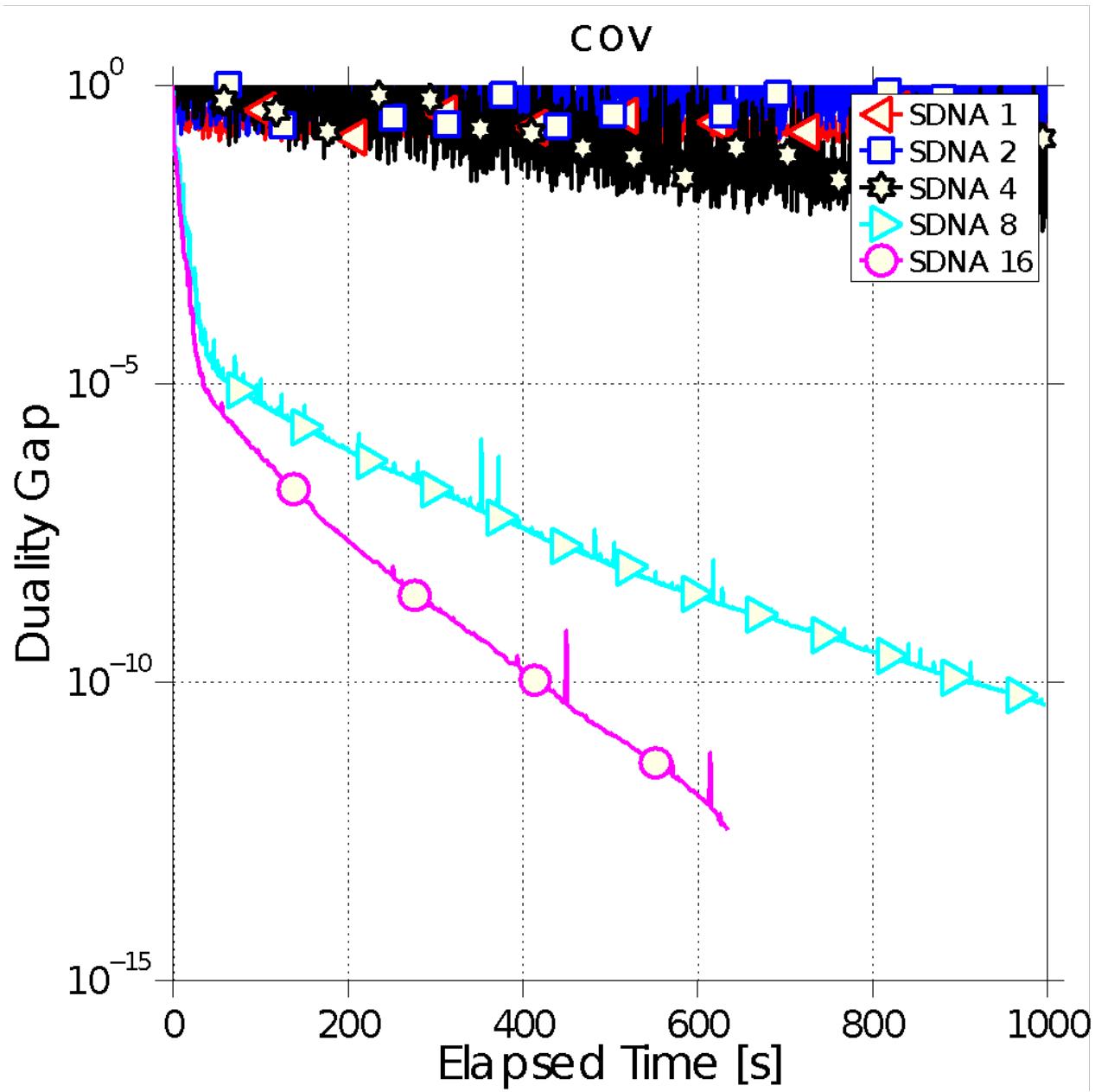




Real Dataset:

COV

$d = 54$ $n = 581,012$



Part D

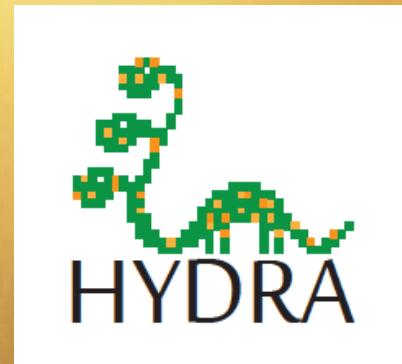
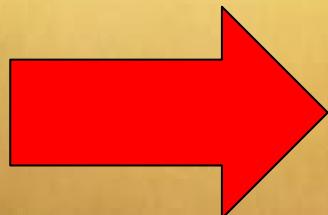
Conclusion

Summary

- Can combine **curvature & randomization** and get complexity rates
- Curvature is utilized by doing exact computations in small but **multidimensional subspaces**
- Randomized “Newton” (Method 1):
 - **Superlinear speedup** (always)
 - **Expensive iterations:** Needs to solve a “small” but potentially dense linear system in each step
- Randomized Coordinate Descent (Method 3):
 - **Sublinear speedup** (gets better with sparsity or good spectral properties)
 - **Cheap iterations:** Needs to solve a small diagonal linear system in each step
- Can apply to the **dual of ERM:** **SDNA**
 - Coincides with SDCA if minibatch size = 1
 - Improves on SDCA when minibatch size is “small enough”
 - New effect: # passes over data decreases as minibatch size increases
- Previous work: **Stochastic quasi-Newton** [Schraudolph, Yu, Gunter ’07] [Bordes, Bottou, Gallinari ’09] [Byrd, Hansen, Nocedal, Singer ’14] **Newton sketch** [Pilanci & Wainwright ’15]

THE END

Big Data Detour

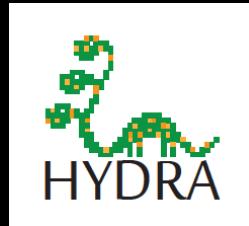


Experiment

Machine: 128 nodes of Hector Supercomputer (4096 cores)

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

Algorithm:



with $c = 512$



Peter Richtarik and M.T.

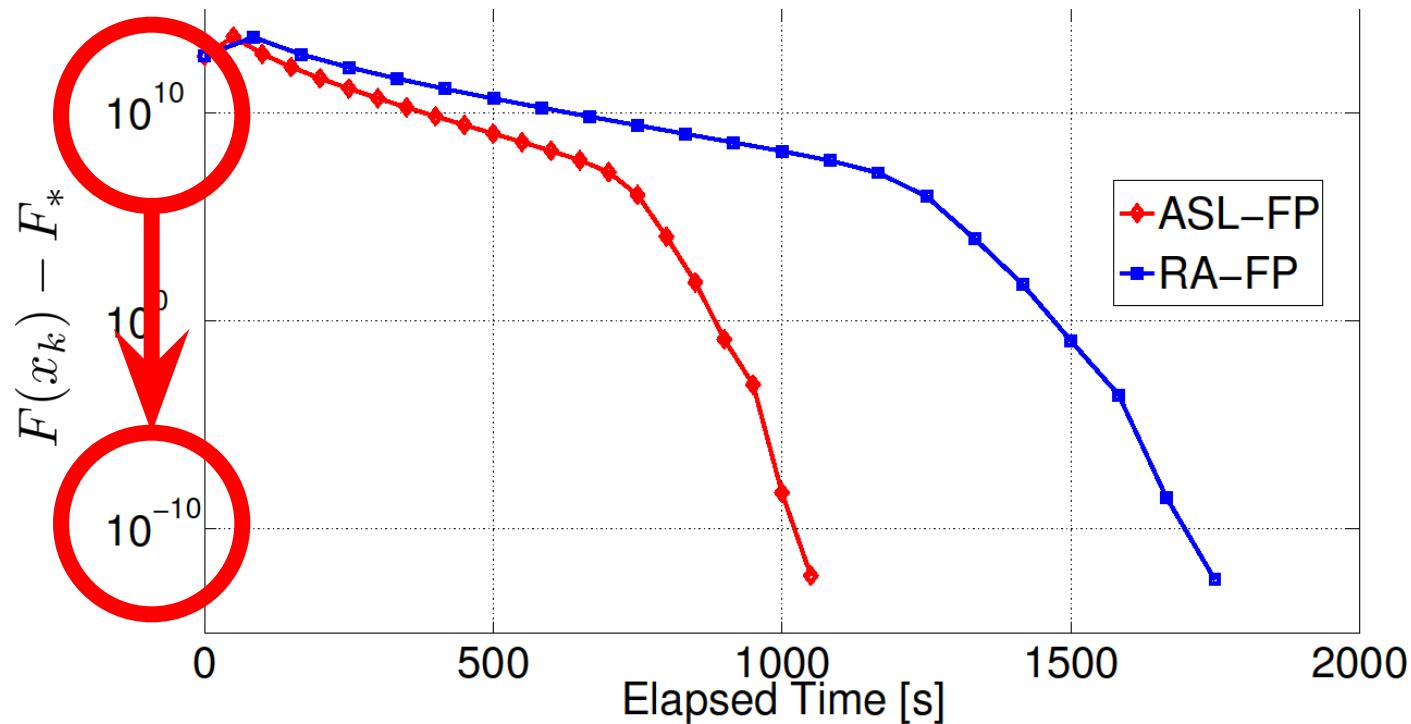
Distributed coordinate descent method for learning with big data

arXiv:1310.2059, 2013

LASSO with a 3TB data matrix

$$A = 10^9\text{-by-}0.5 \times 10^9$$

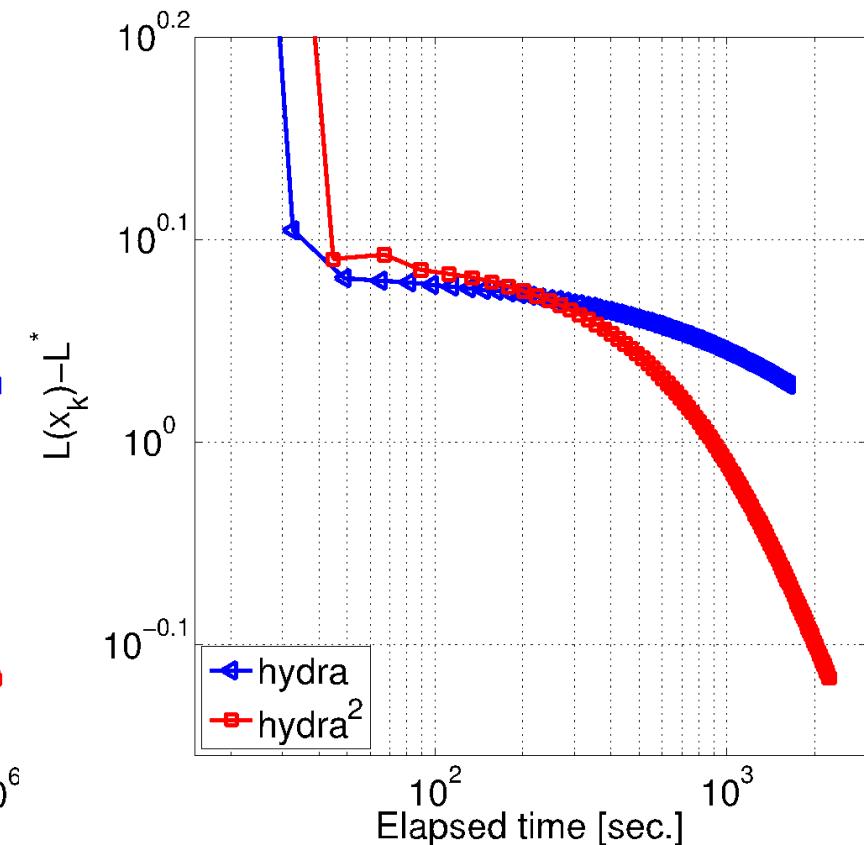
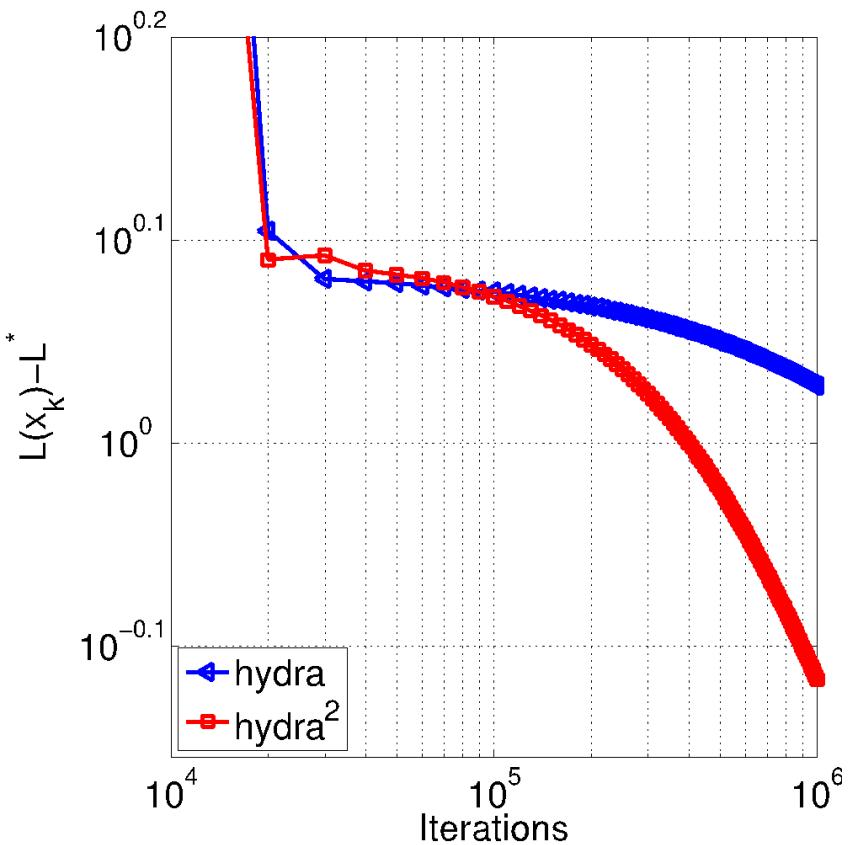
$\underbrace{}$
 $N = \# \text{ coordinates}$



128 Cray XE6 nodes with 4 MPI processes ($c = 512$)

Each node: 2 x 16-cores with 32GB RAM

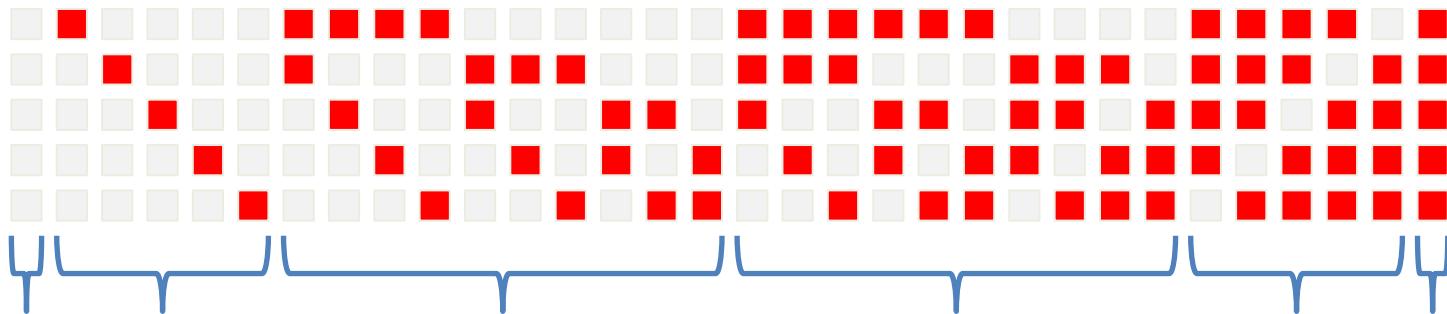
LASSO: 5TB data + 128 nodes



Arbitrary Sampling

$$\mathbb{P}(S_k = S) = ? \quad \forall S \subseteq \{1, 2, \dots, n\}$$

Example: $n = 5$



$|S| = 0$

$|S| = 1$

$|S| = 2$

$|S| = 3$ $|S| = 4$ $|S| = 5$