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

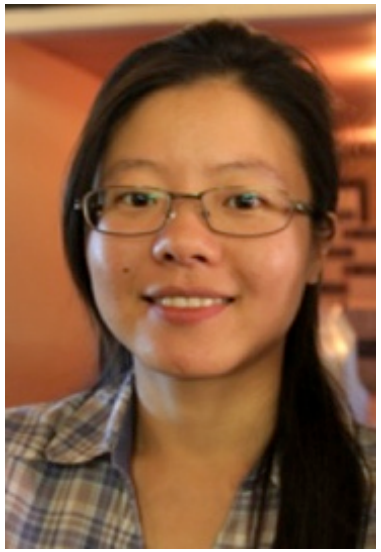
SDNA

Stochastic Dual Newton Ascent for Empirical Risk Minimization

Peter Richtárik

International Symposium on Mathematical Programming
Pittsburgh, July 2015

Coauthors



Zheng Qu
(Edinburgh)

Martin Takáč
(Lehigh)

Olivier Fercoq
(Telecom ParisTech)

SDNA: Stochastic Dual Newton Ascent
for Empirical Risk Minimization

Zheng Qu* Peter Richtárik† Martin Takáč‡ Olivier Fercoq§

February 8, 2015
(updated: May 2, 2015)

Abstract

We propose a new algorithm for minimizing regularized empirical loss: *Stochastic Dual Newton Ascent* (SDNA). One method is dual in nature: in each iteration we update a random subset of the dual variables. However, unlike existing methods such as stochastic dual coordinate ascent, SDNA is capable of utilizing all curvature information contained in the examples, which leads to striking improvements in both theory and practice – sometimes by orders of magnitude. In the special case when an L2-regularizer is used in the primal, the dual problem is a concave quadratic maximization problem plus a separable term. In this regime, SDNA in each step solves a proximal subproblem involving a random principal submatrix of the Hessian of the quadratic function; whence the name of the method. If, additionally, the loss is also quadratic, our method can be interpreted as a novel variant of the recently introduced Iterative Hessian Sketch.

Keywords: empirical risk minimization, randomized algorithms, iterative Hessian sketch, coordinate descent, convex optimization, duality.

*School of Mathematics, The University of Edinburgh, UK, (email: zheng.qu@ed.ac.uk).
†School of Mathematics, The University of Edinburgh, UK, (email: peter.richtarik@ed.ac.uk).
‡Industrial and Systems Engineering, Lehigh University, USA, (email: takac.mt@gmail.com).
§Signal and Image Processing department, Telecom ParisTech, France, (email: olivier.fercoq@telecom-paristech.fr)
Acknowledgments: The first two authors would like to acknowledge support from the EPSRC Grant EP/K02325X/1, Accelerated Coordinate Descent Methods for Big Data Optimization.

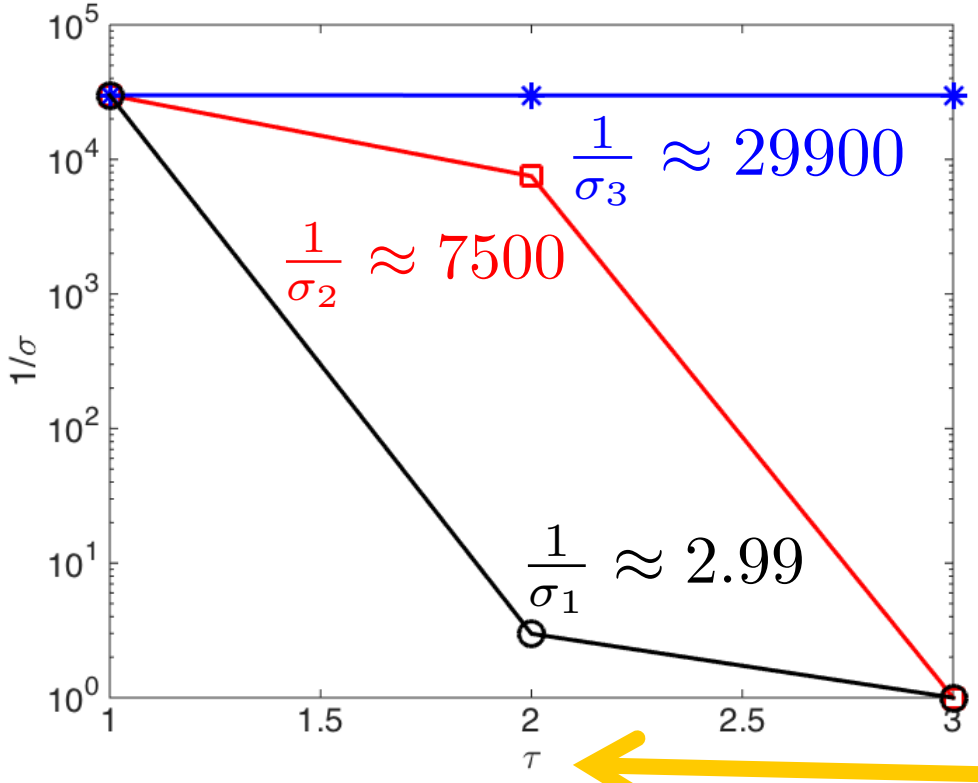
1

Zheng Qu, P.R., Martin Takáč and Olivier Fercoq
SDNA: Stochastic Dual Newton Ascent for empirical risk minimization
arXiv:1502.02268, 2015

Part A
Motivation

3D Quadratic

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

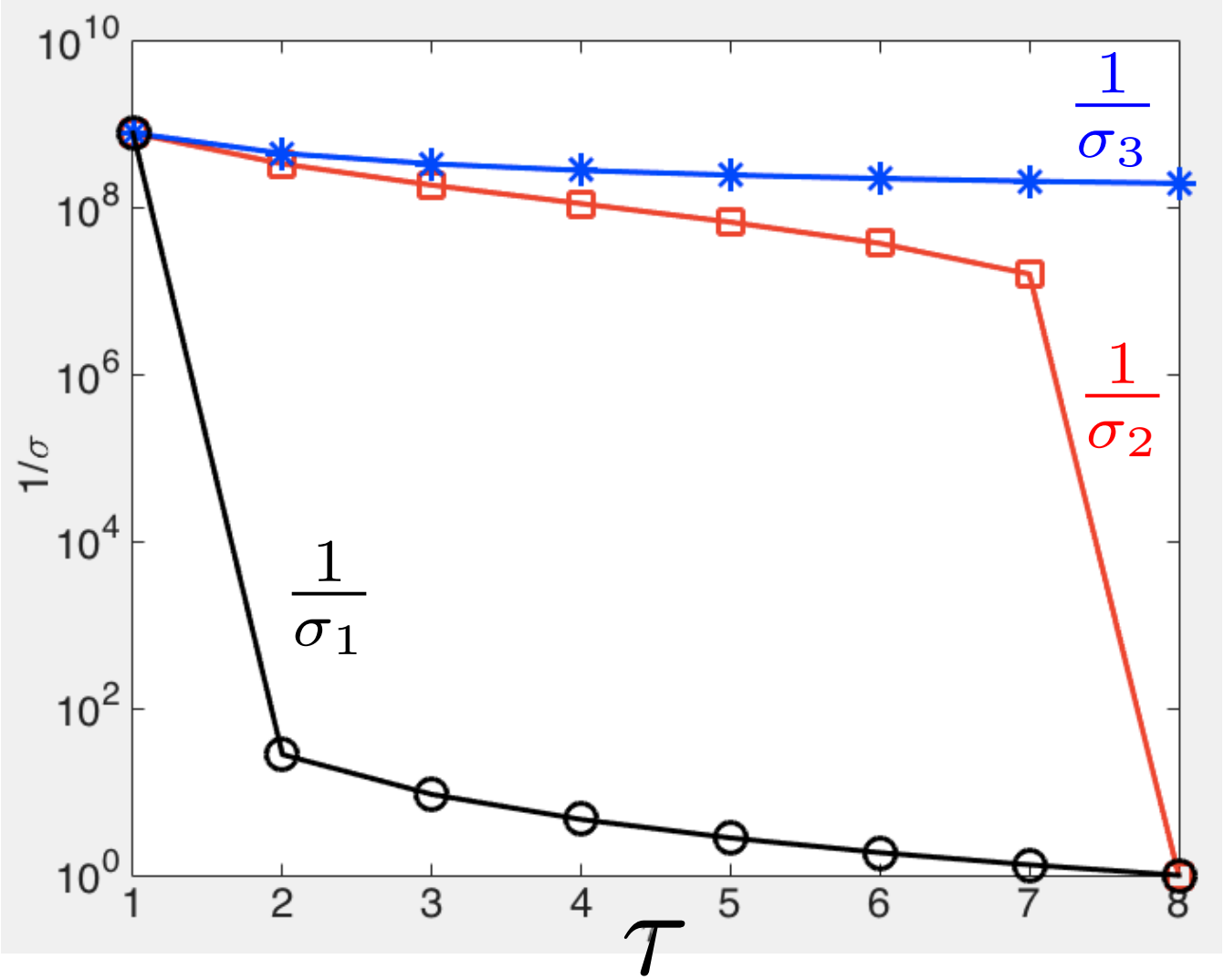


$$\mathbf{M} = \begin{pmatrix} 1.0000 & 0.9900 & 0.9999 \\ 0.9900 & 1.0000 & 0.9900 \\ 0.9999 & 0.9900 & 1.0000 \end{pmatrix}$$

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

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

8D Quadratic



Part B
Three Methods

The Problem & Assumptions

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

Large dimension

Strong convexity

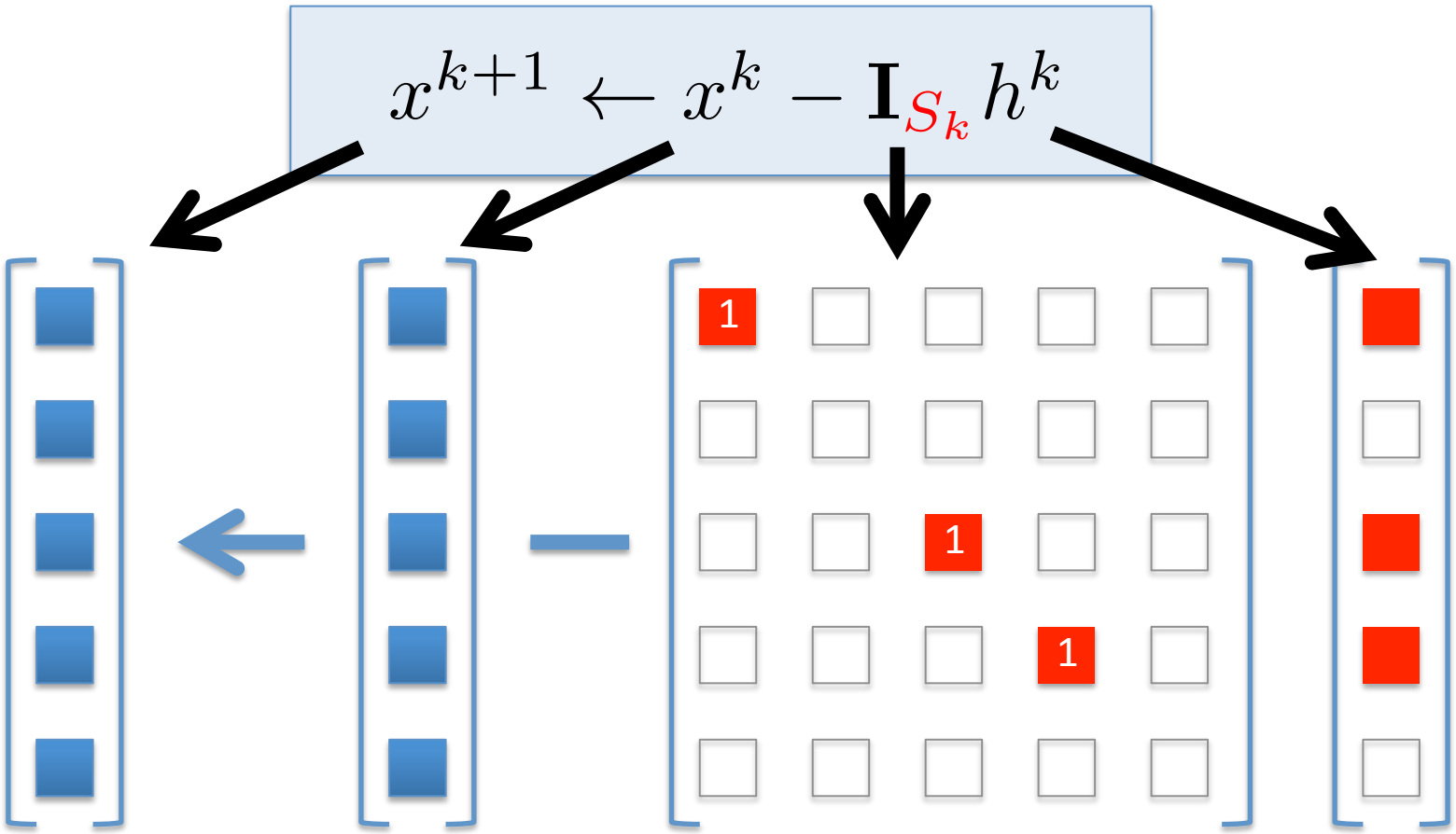
$$f(x) + (\nabla f(x))^\top h + \frac{1}{2} h^\top \mathbf{G} h \leq f(x+h)$$

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

P.R. and Martin Takáč

On optimal probabilities in stochastic coordinate descent methods

In NIPS Workshop on Optimization for Machine Learning, 2013

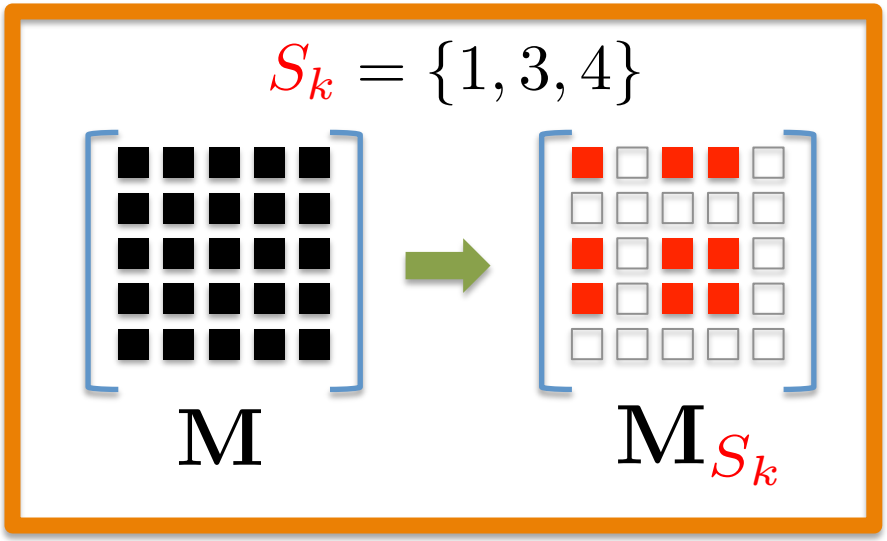
Optimization Letters 2015 (arXiv:1310.3438)

Key Inequality

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

$$x \leftarrow x^k$$

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



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

Method 3

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

1. take expectations on both sides

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

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

2. diagonalize

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

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

3. minimize the RHS in *h*

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

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:

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

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

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

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

Method 2

Method 2

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

1. take expectations on both sides

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

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

2. minimize the RHS in *h*

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

Convergence of Method 2

Theorem (QRTF'15)

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

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

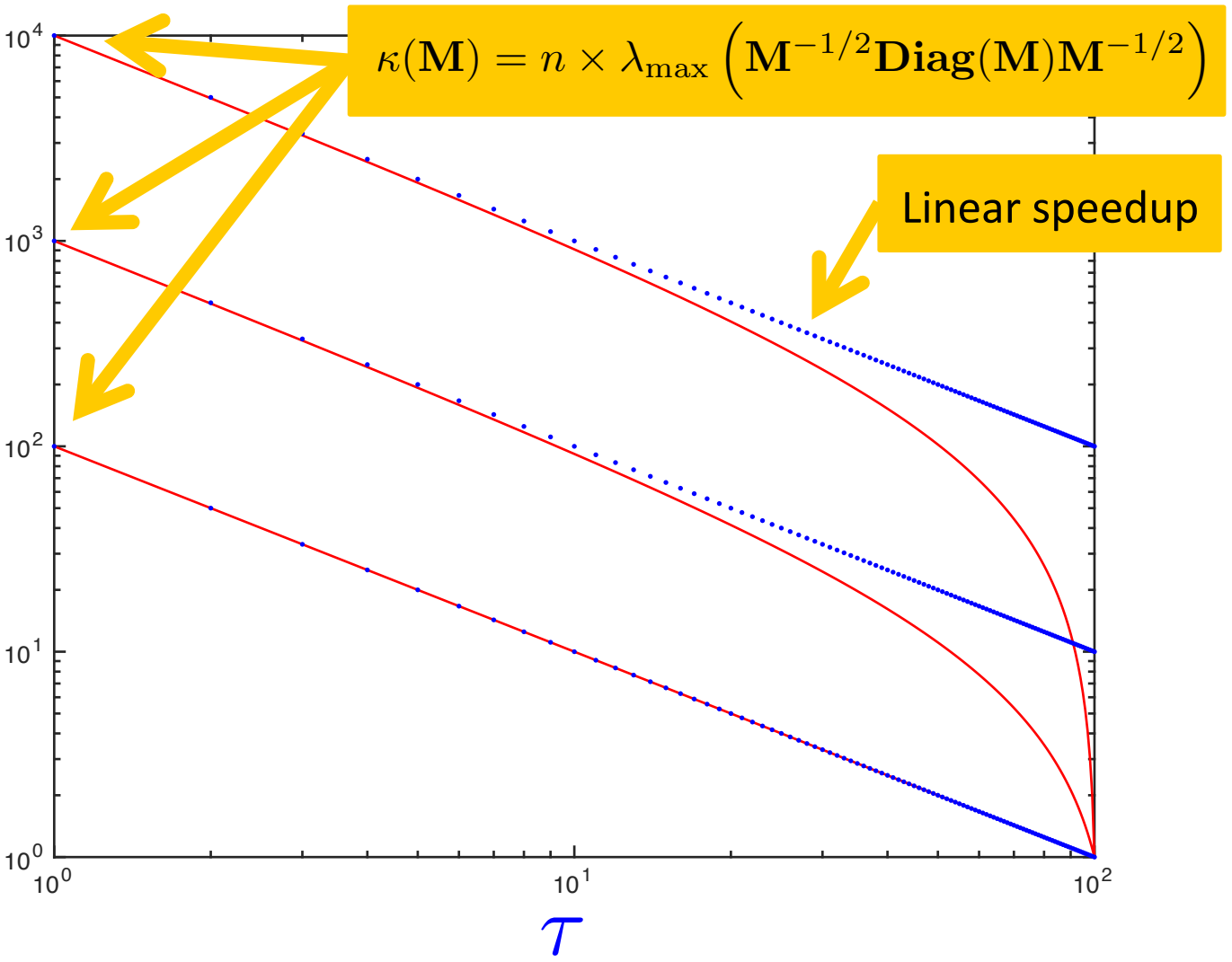
Alternative formulation:

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

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

$$\frac{1}{\sigma_2(\tau)} = \frac{n}{n-1}\lambda_{\max}\left(\mathbf{G}^{-1/2}\left[\left(\frac{n}{\tau}-1\right)\mathbf{Diag}(\mathbf{M}) + \left(1-\frac{1}{\tau}\right)\mathbf{M}\right]\mathbf{G}^{-1/2}\right)$$

$$\frac{1}{\sigma_2(\tau)}$$



Method 1
Randomized Newton Method

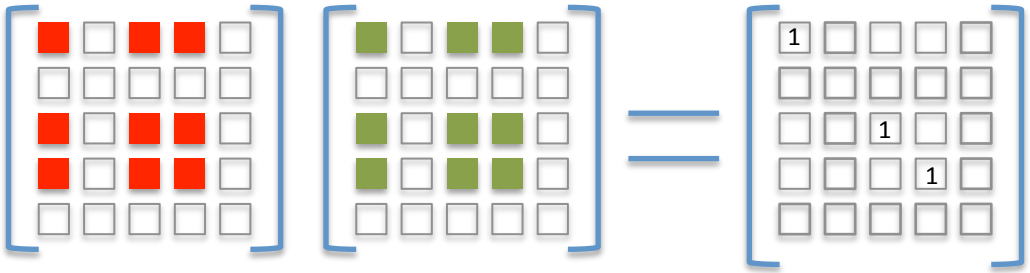
Method 1: Randomized Newton

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

minimize the RHS in *h*

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

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



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

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:

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

Three Convergence Rates

3 Convergence Rates

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

Theorem (QRTF'15)

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

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

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

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

Methods 2 and 3 coincide if we update all coordinates

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

Randomized Newton: **superlinear speedup**

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

Randomized Coordinate Descent: **sublinear speedup**

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

Lower bound on complexity

Part C
Empirical Risk Minimization

Primal Problem

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

P = Regularized Empirical Risk

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

positive regularization parameter

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

d = # features (parameters)

w = linear predictor

n data vectors ("examples")

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

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

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

SDNA

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

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

Iterate:

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

Generate a random set S_k

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

Compute:

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

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

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

Convergence of SDNA

Better rate than SDCA

Theorem (QRTF'15)

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

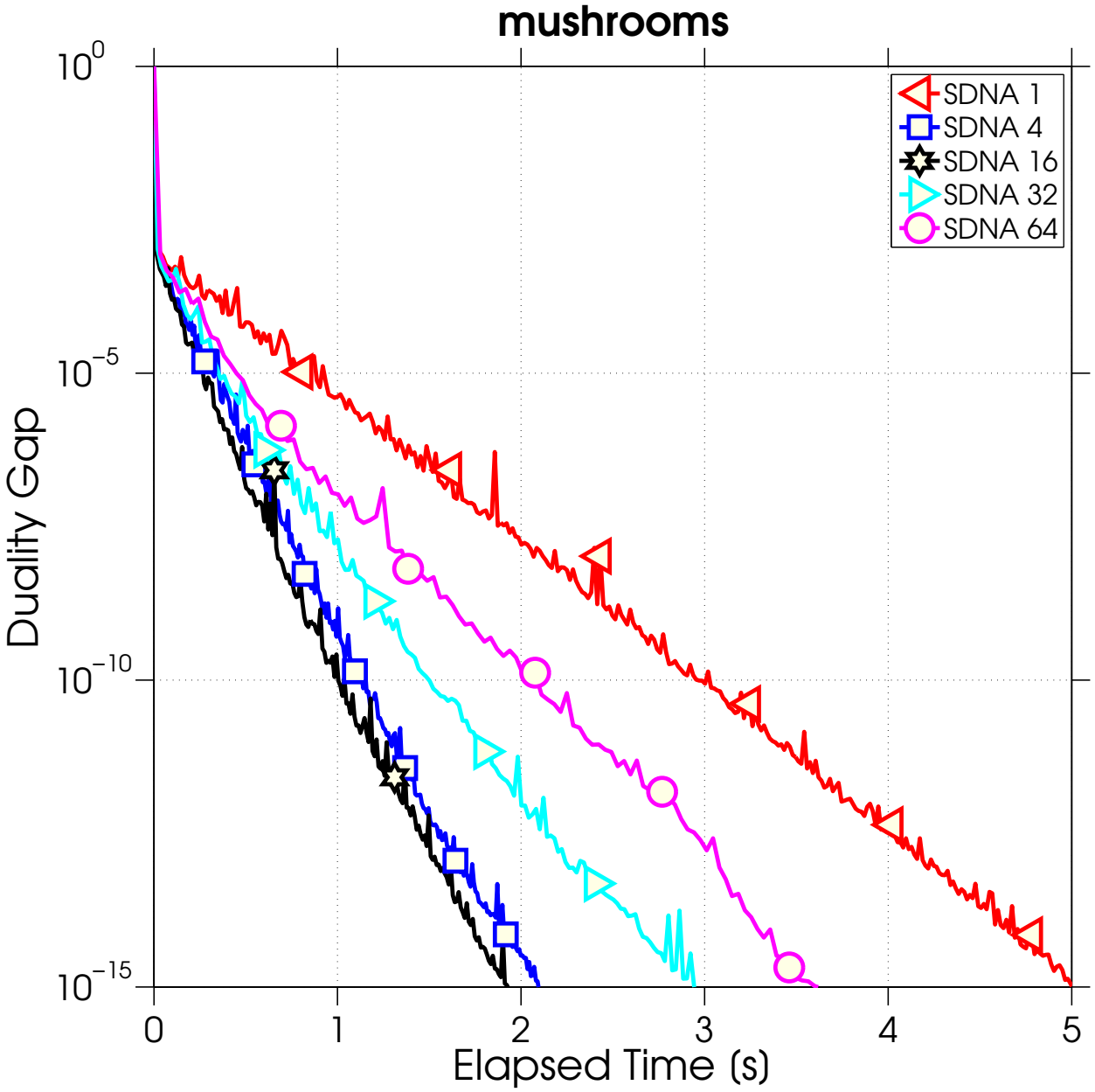
$$\sigma_1^{prox} = \frac{\tau}{n} \min\{1, s_1\}$$

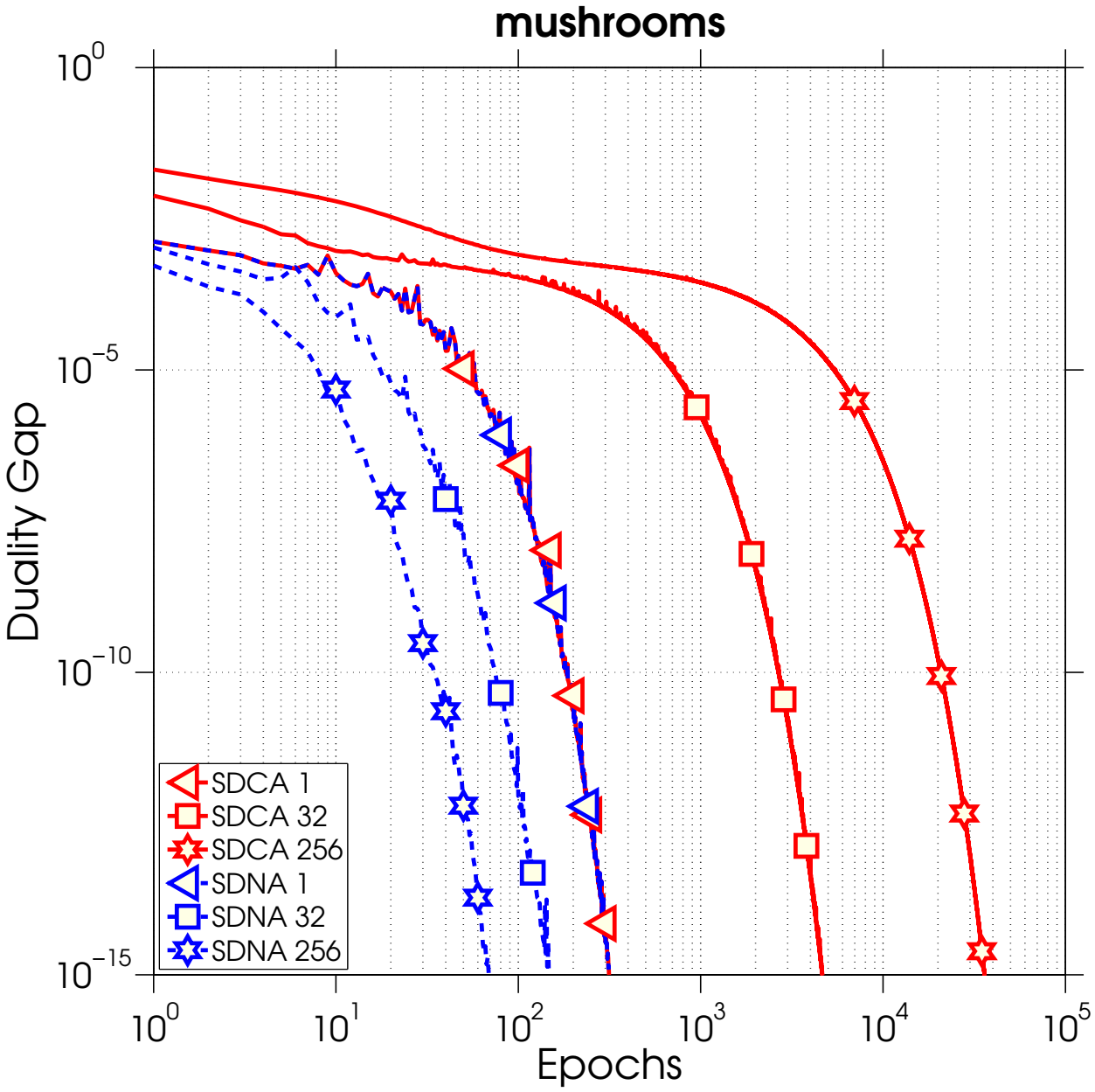
$$\tau = \mathbb{E}[|S_k|] \qquad s_1 = \lambda_{\min}\left[\left(\frac{1}{\tau\gamma\lambda}\mathbb{E}[(\mathbf{A}^\top\mathbf{A})_{S_k}] + \mathbf{I}\right)^{-1}\right]$$

Real Dataset:

`mushrooms`

$d = 112 \quad n = 8{,}124$

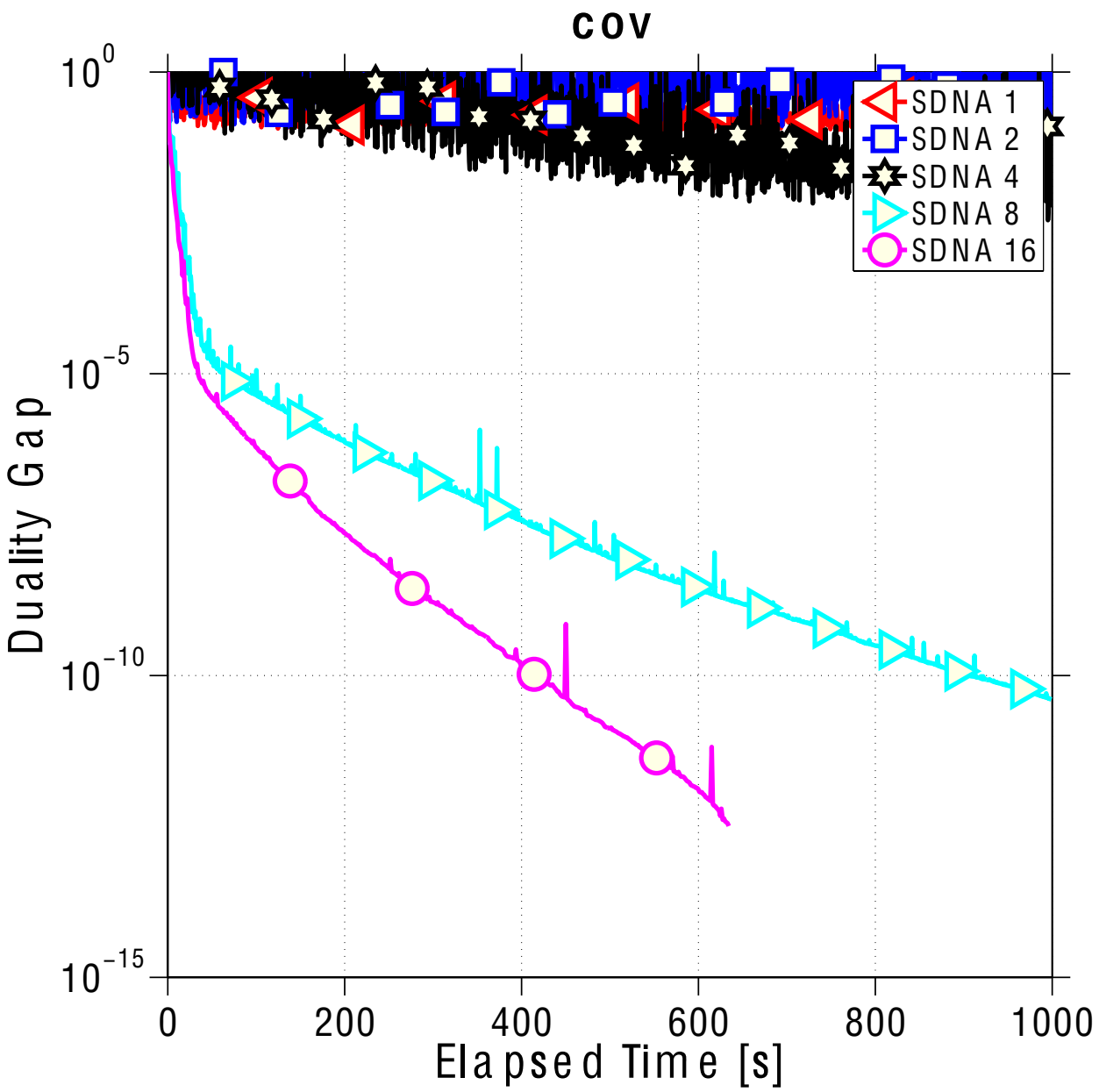

mushrooms
SDNA 1
SDNA 4
SDNA 16
SDNA 32
SDNA 64
Duality Gap
Elapsed Time (s)
10^0
10^-5
10^-10
10^-15
0
1
2
3
4
5


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

Real Dataset:

cov

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


cov
SDNA 1
SDNA 2
SDNA 4
SDNA 8
SDNA 16
Duality Gap
Elapsed Time [s]
10^{0}
10^{-5}
10^{-10}
10^{-15}
0
200
400
600
800
1000

Part D
Conclusion

Randomized Methods with Arbitrary Sampling

Method 3

FD13

SYNC

P.R. and Martin Takáč. **On optimal probabilities in stochastic coordinate descent methods.** *Optimization Letters, 2015 (arXiv:1310.3438)*

Zheng Qu, P.R. and Tong Zhang. **Randomized dual coordinate ascent with arbitrary sampling.** *arXiv:1411.5873*

α

Zheng Qu and P.R. **Coordinate descent with arbitrary sampling I: algorithms and complexity.** *arXiv:1412.8060*

Zheng Qu and P.R. **Coordinate descent with arbitrary sampling II: ESO.** *arXiv:1412.8063*

SDNA

Zheng Qu, P.R., Martin Takáč and Olivier Fercoq. **SDNA: Stochastic Dual Newton Ascent for empirical risk minimization.** *arXiv:1502.02268*

Dominik Csiba and P.R. **Primal method for ERM with flexible mini-batching schemes and non-convex losses.** ICML 2015 *(arXiv:1502.02268)*

Robert M. Gower and P.R. **Randomized iterative methods for linear systems.** *arXiv: 1502.02268*

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