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

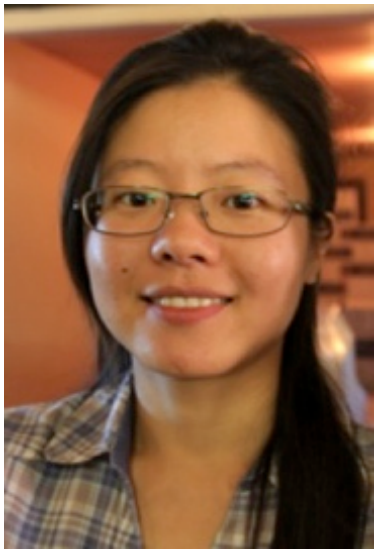
SDNA

Stochastic Dual Newton Ascent for Empirical Risk Minimization

Peter Richtárik

2016 International Workshop on Modern Optimization and Applications
Beijing, June 27-29, 2016

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). Our 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, in addition, the loss functions are 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.

Zheng Qu, P.R., Martin Takáč and Olivier Fercoq
SDNA: Stochastic Dual Newton Ascent for empirical risk minimization
ICML 2016 (arXiv:1502.02268)

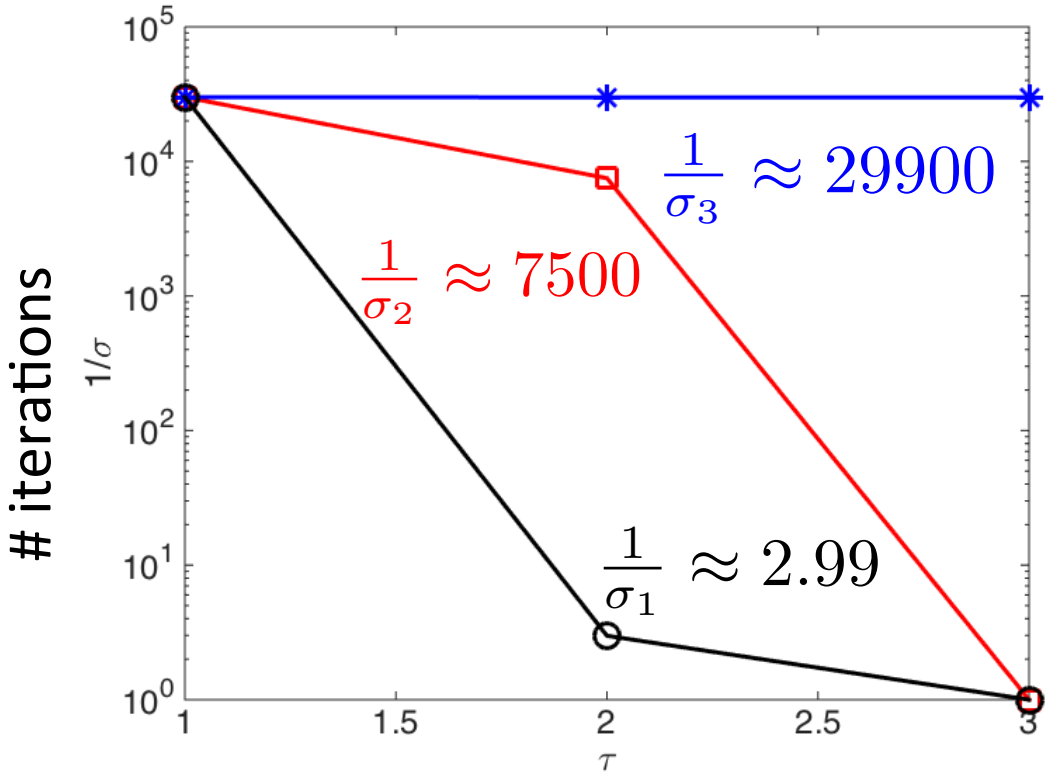
Part A
Motivation

Why Curvature Is Cute

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

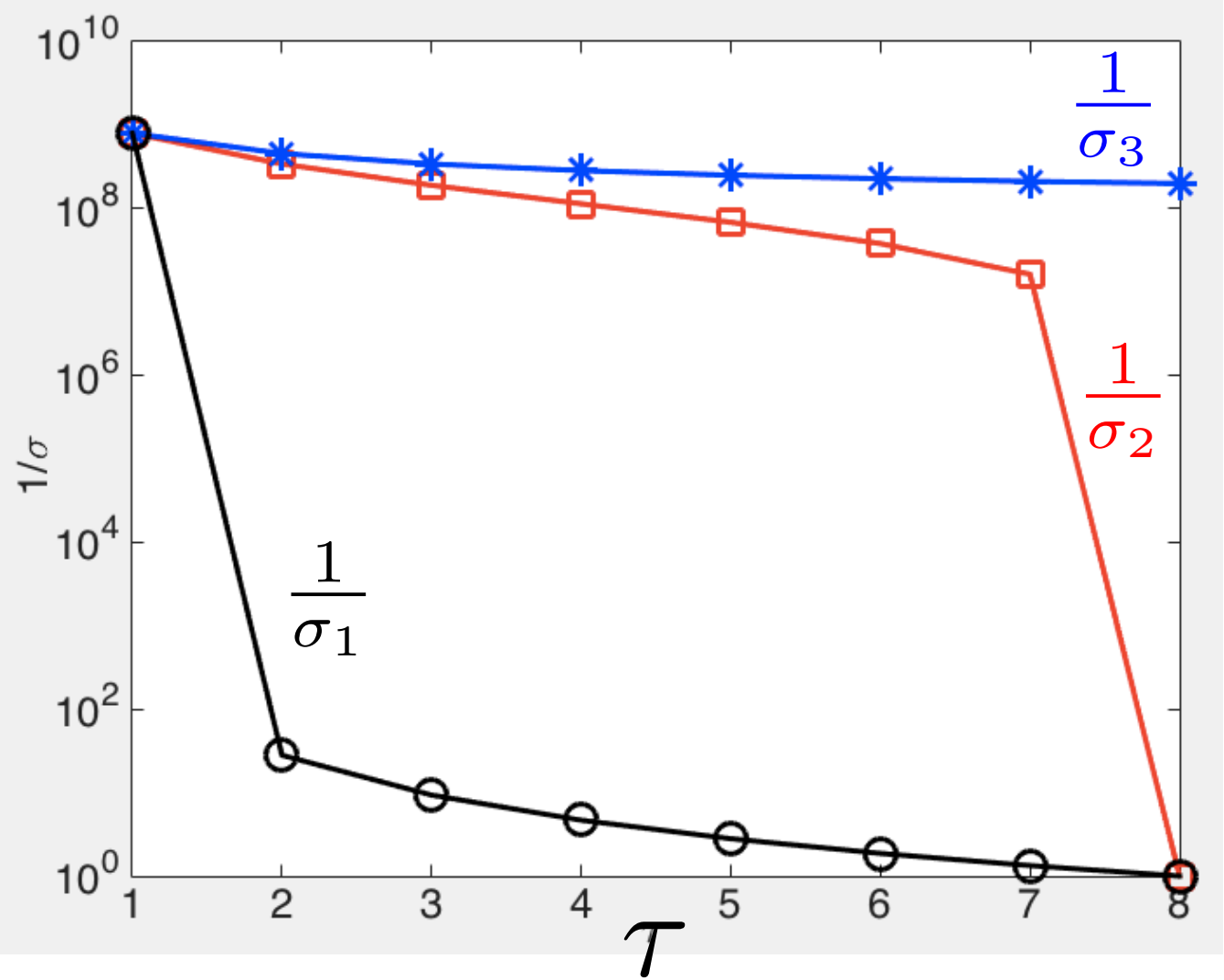
$$\mathbf{M} = \begin{pmatrix} 1.0000 & 0.9900 & 0.9999 \\ 0.9900 & 1.0000 & 0.9900 \\ 0.9999 & 0.9900 & 1.0000 \end{pmatrix}$$

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



- Phenomenon described in [Qu et al 15]
- Method 1: Two points of view: "Exact line search in higher dimensional subspaces" or "inversion of random submatrices of the Hessian"

8D Quadratic



Part B
Three Methods

The Problem & Assumptions

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

Large dimension

Strong convexity

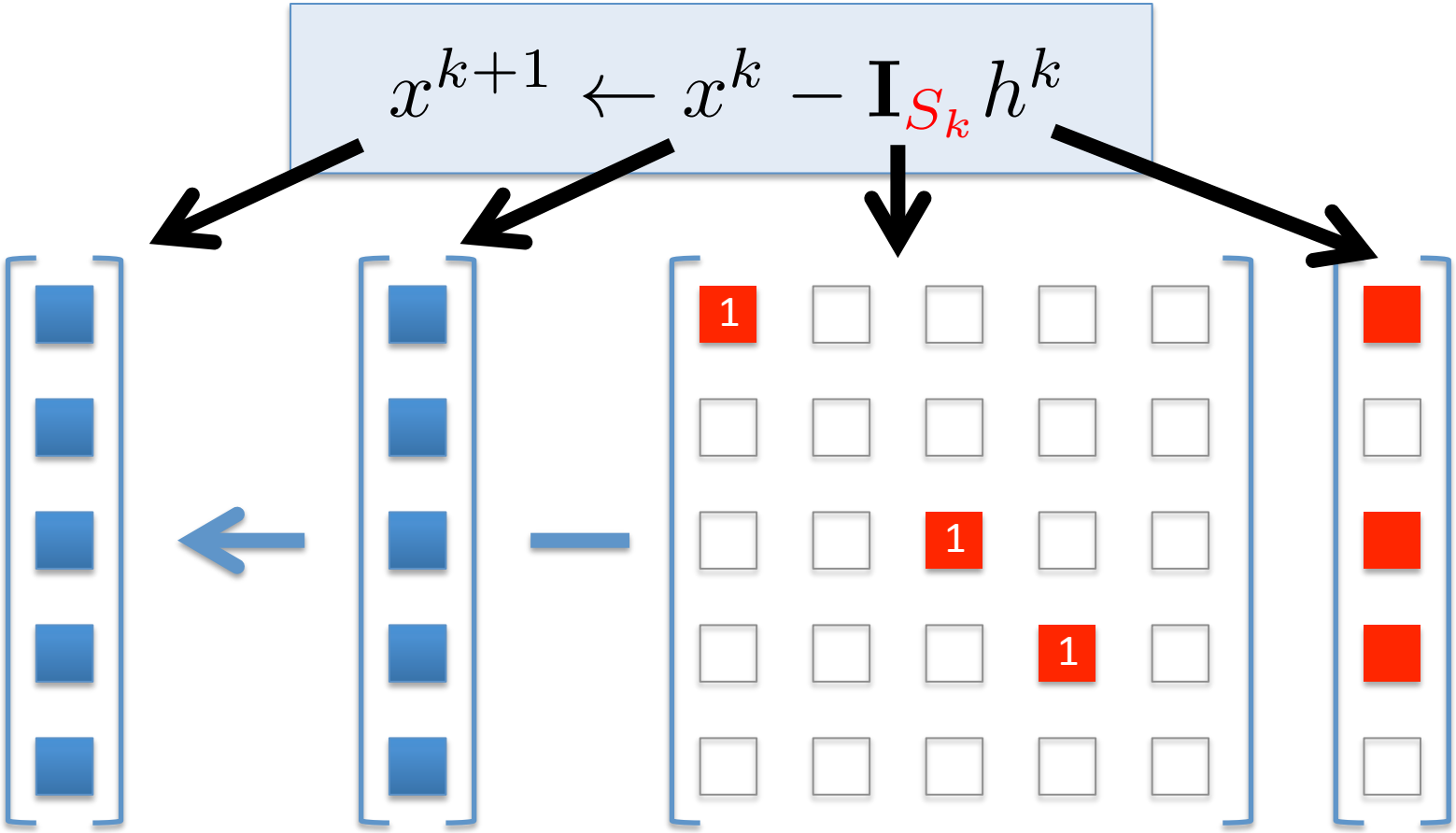
$$f(x) + (\nabla f(x))^\top h + \frac{1}{2} h^\top \mathbf{G} h \leq f(x+h)$$

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

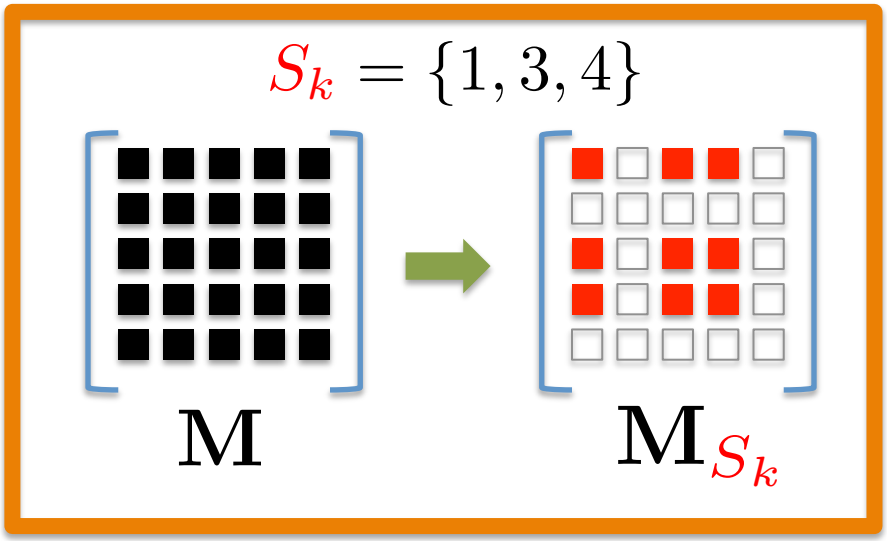
Optimization Letters 2015 (arXiv:1310.3438)

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*

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

Convergence of Method 2

Theorem (QRTF'15)

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

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

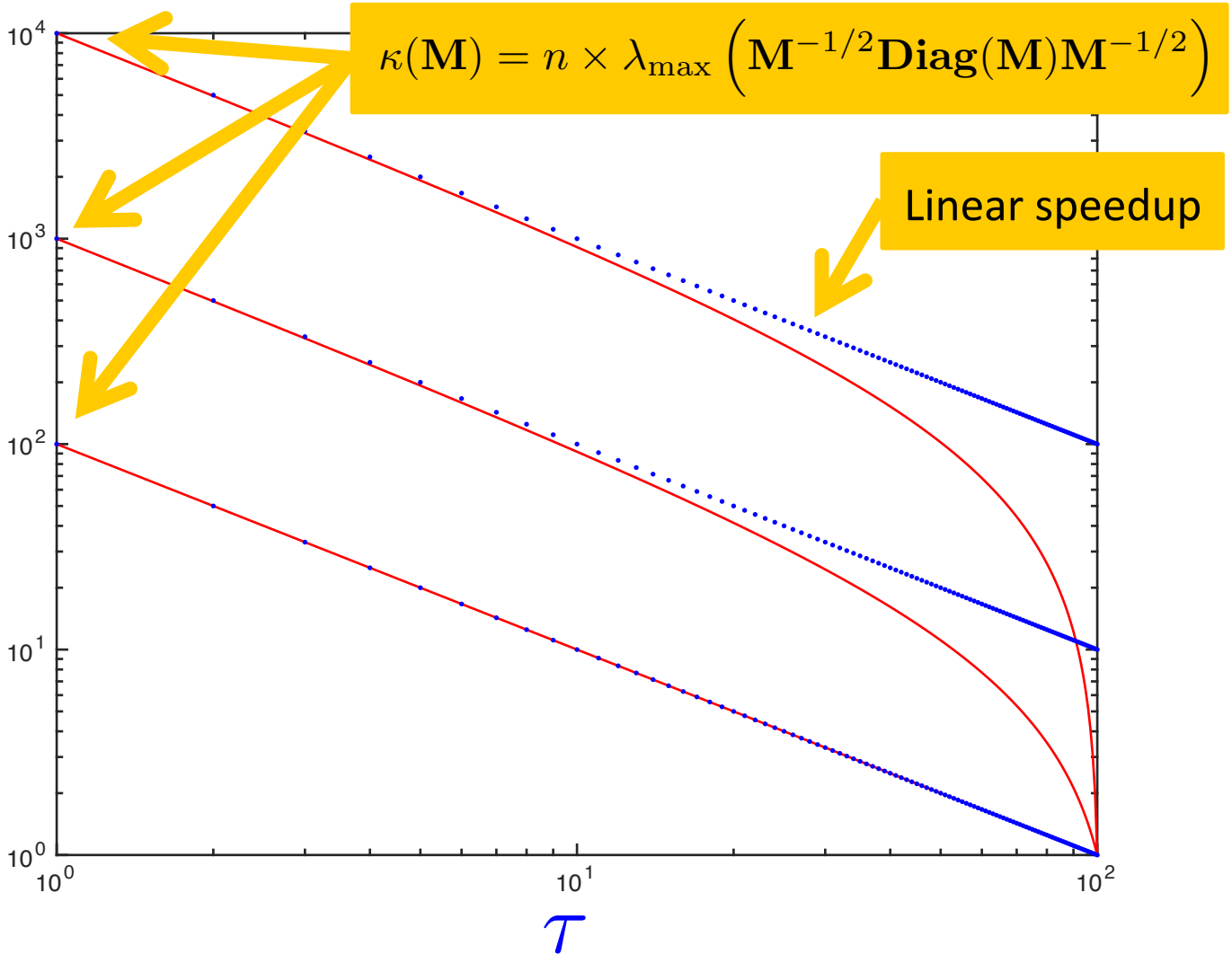
Alternative formulation:

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

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

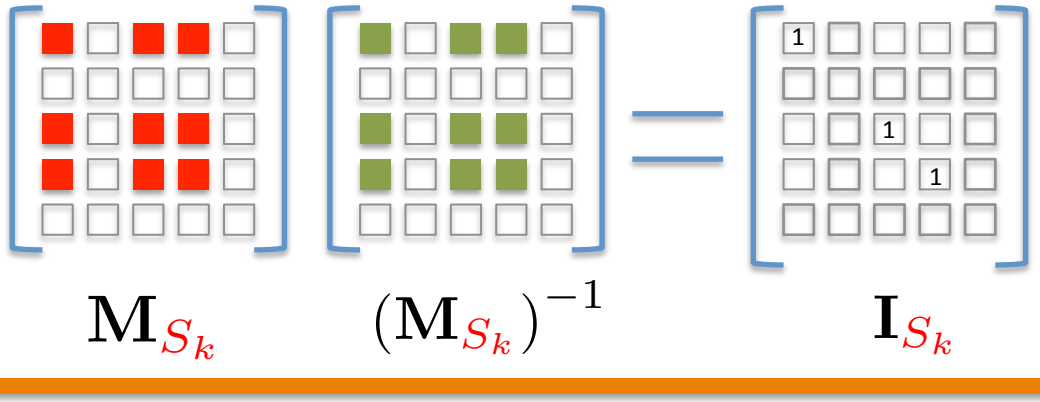
Method 1: Randomized Newton

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

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

Theorem [QRTF'15]

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

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

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

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

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

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

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

Methods 1 and 2 coincide if we update all coordinates

Randomized Newton: **superlinear speedup**

Randomized Coordinate Descent: **sublinear speedup**

Part C
Empirical Risk Minimization

Primal Problem

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

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

positive regularization parameter

P = Regularized Empirical Risk

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

w = linear predictor

n data vectors ("examples")

d = # 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

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

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

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

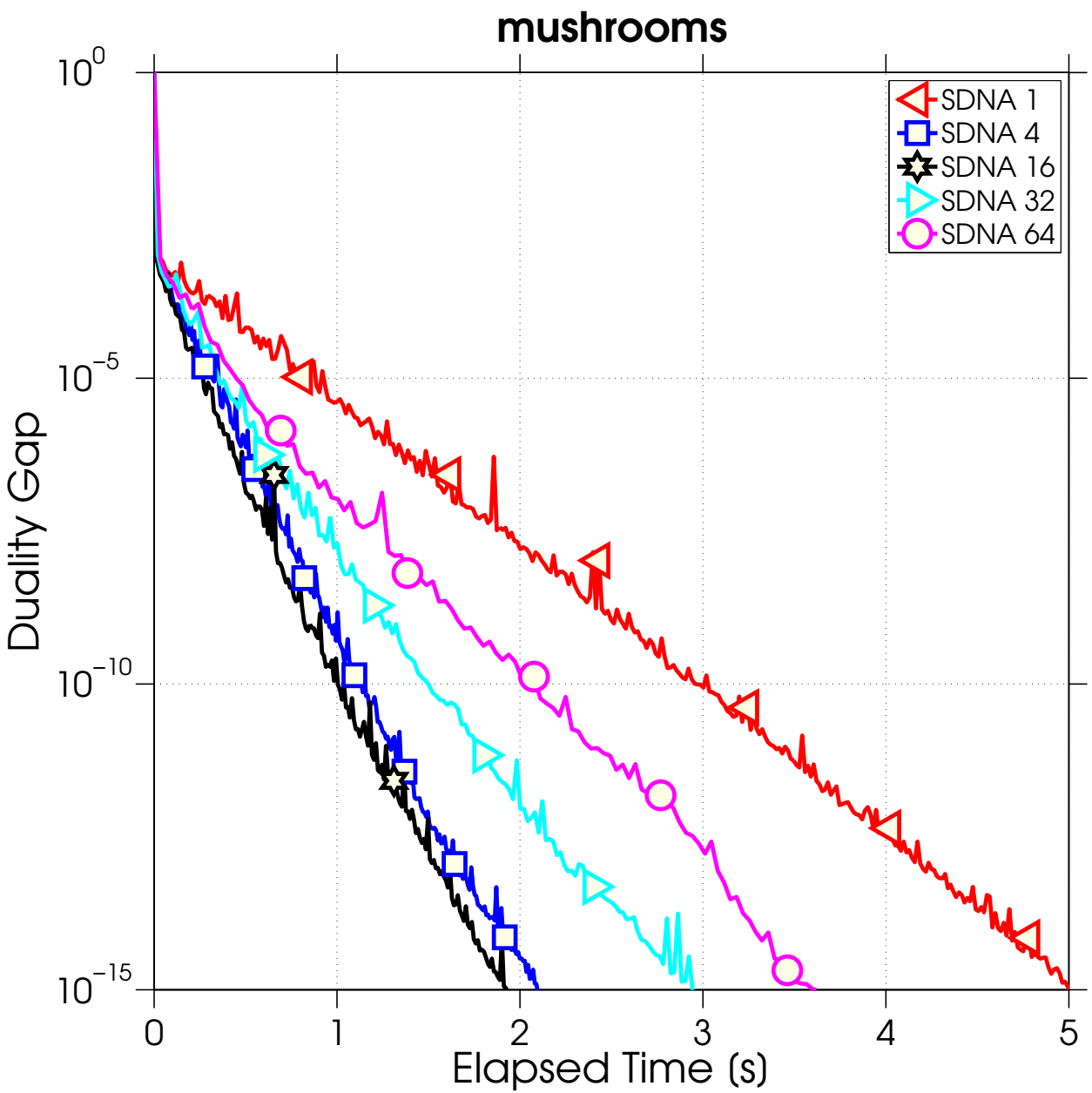
$$\sigma_1^{prox} = \frac{\tau}{n} \min\{1, s_1\}$$

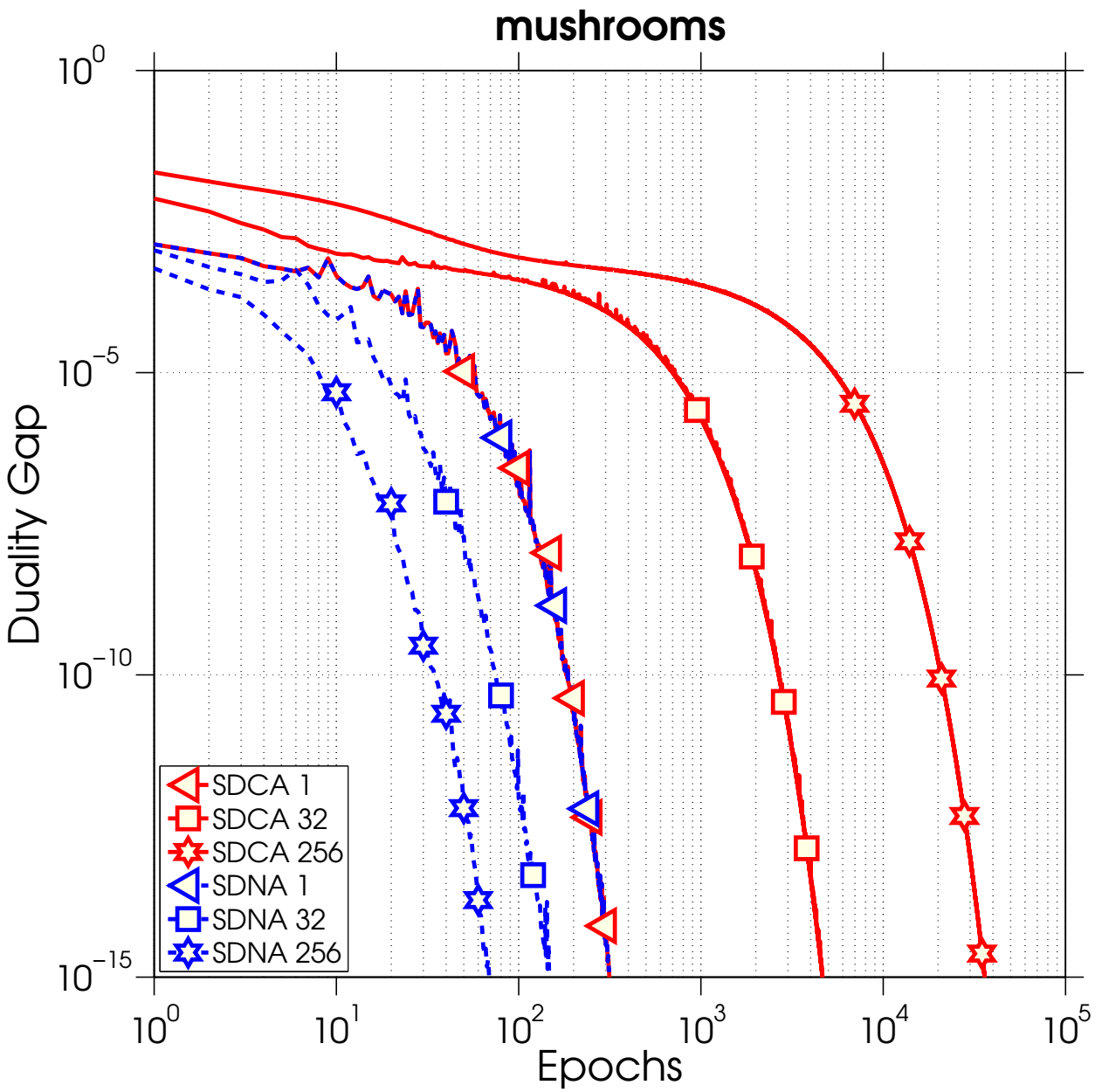
$$\tau = \mathbb{E}[|S_k|] \qquad s_1 = \lambda_{\min}\left[\left(\frac{1}{\tau\gamma\lambda}\mathbb{E}[(\mathbf{A}^\top\mathbf{A})_{S_k}] + \mathbf{I}\right)^{-1}\right]$$

Real Dataset:

`mushrooms`

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

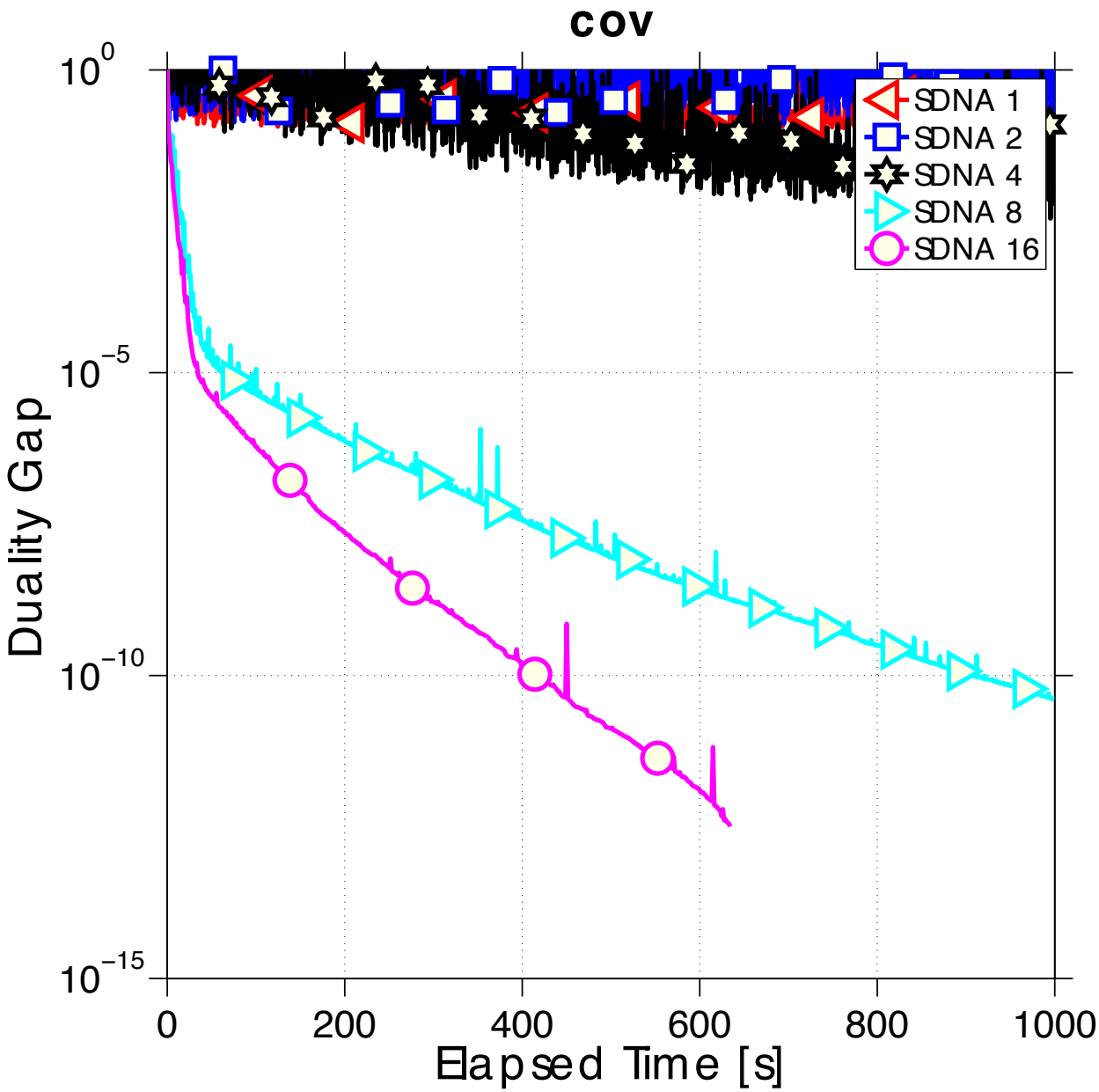

mushrooms
SDNA 1
SDNA 4
SDNA 16
SDNA 32
SDNA 64
Duality Gap
Elapsed Time (s)
10^0
10^-5
10^-10
10^-15
0
1
2
3
4
5


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

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


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

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]

Methods with Arbitrary Sampling

1st work on arbitrary (and optimal) sampling

P.R. and Martin Takáč

On optimal probabilities in stochastic coordinate descent methods

Optimization Letters, p 1-11, 2015 (arXiv:1310.3438)

Unification of randomized & deterministic methods, accelerated and non-accelerated methods via arbitrary sampling

Zheng Qu and P.R.

Coordinate descent with arbitrary sampling I: algorithms and complexity

Optimization Methods and Software, 2016 *(arXiv:1412.8060)*

Formulas for computing the ESO constants

Zheng Qu and P.R.

Coordinate descent with arbitrary sampling II: ESO

Optimization Methods and Software, 2016 *(arXiv:1412.8063)*

Empirical Risk Minimization & Arbitrary Sampling

1st work on arbitrary sampling for ERM

Zheng Qu, P.R. and Tong Zhang
Quartz: randomized dual coordinate ascent with arbitrary sampling
In *Neural Information Processing Systems*, 2015 *(arXiv:1411.5873)*

Randomization & curvature

Zheng Qu, P.R., Martin Takáč and Olivier Fercoq
SDNA: Stochastic Dual Newton Ascent for empirical risk minimization
ICML 2016 *(arXiv:1502.02268)*

SDNA

Similar to Quartz, but primal-only analysis

Dominik Csiba and P.R
Primal method for ERM with flexible mini-batching schemes and non-convex losses *(arXiv:1506.02227)*

Randomization, Linear Systems & ERM

Robert M. Gower and P.R
Randomized iterative methods for linear systems
SIAM J of Matrix Analysis and Applications 36(4), 1660-1690, 2015

Robert M. Gower and P.R
Stochastic dual ascent for solving linear systems *(arXiv:1512.06890)*

Robert M. Gower and P.R
Randomized quasi-Newton updates are linearly convergent matrix inversion algorithms *(arXiv:1602.01768)*

Robert M. Gower, Donald Goldfarb and P.R
Randomized block BFGS: squeezing more curvature out of data
ICML 2016

THE END

Proofs

Theorem 1 f is $\mathbf{G}$-strongly convex & $\mathbf{G} \succ 0$ $\quad S_k \overset{\text{i.i.d.}}{\sim} \hat{S}$

f is $\mathbf{M}$-smooth & $\mathbf{M} \succ 0$ $\quad \hat{S}$ is proper

Method m (for $m = 1, 2, 3$) converges linearly:

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

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

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

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

Definition of *p*

$$p = (p_1, \ldots, p_n) \in \mathbb{R}^n$$
$$p_i = \mathbb{P}(i \in \hat{S})$$

Definition of *v*

$$\mathbb{E}\left[\mathbf{M}_{\hat{S}}\right] \preceq \mathbf{D}(p)\mathbf{D}(v)$$

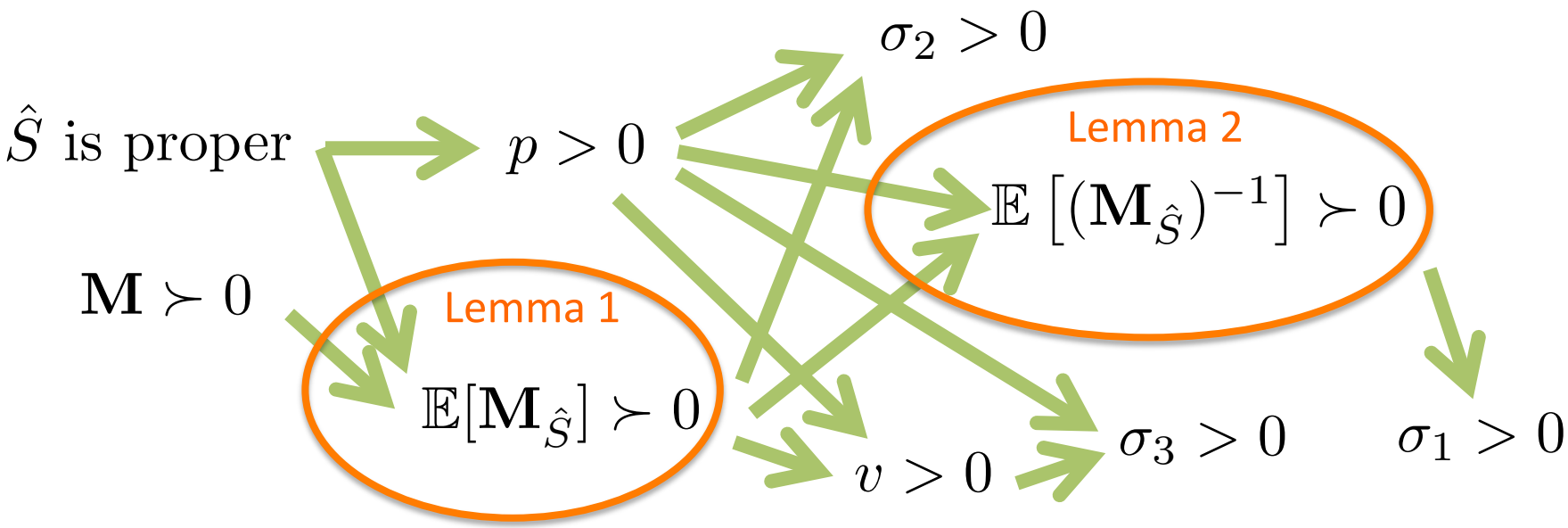
Sanity Check

Let us verify that the rates asserted by the theorem make sense (well defined & positive)

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

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

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



Lemma 1

Lemma 1

A

$$\mathbf{M} \succeq 0 \quad \& \quad \hat{S} \text{ is any sampling} \quad \Rightarrow \quad \mathbb{E}\left[\mathbf{M}_{\hat{S}}\right] \succeq 0$$

B

$$\mathbf{M} \succ 0 \quad \& \quad \hat{S} \text{ is a proper sampling} \quad \Rightarrow \quad \mathbb{E}\left[\mathbf{M}_{\hat{S}}\right] \succ 0$$

Proof of Lemma 1

A

The first claim follows from:

- $\mathbf{M}_S \succeq 0$ for all subsets S of $[n] = \{1, 2, \ldots, n\}$
- average of PSD matrices is a PSD matrix

B

Denote $supp\{x\} = \{i \in [n] : x_i \neq 0\}$. Since $\mathbf{M} \succ 0$, any principal submatrix of $\mathbf{M}$ is also positive definite. Hence, for any $x \in \mathbb{R}^n \backslash \{0\}$, $x^\top \mathbf{M}_S x = 0$ implies that $supp\{x\} \cap S = \emptyset$ for all $S \subseteq [n]$. If $x \in \mathbb{R}^n$ is such that

$$x^\top \mathbb{E}\left[\mathbf{M}_{\hat{S}}\right] x = \sum_{S \subseteq [n]} \mathbb{P}(\hat{S} = S) x^\top \mathbf{M}_S x = 0,$$

then $\mathbb{P}(supp\{x\} \cap \hat{S} = \emptyset) = 1$. Since $\hat{S}$ is proper, this only happens when $x = 0$. Therefore, $\mathbb{E}[\mathbf{M}_{\hat{S}}] \succ 0$.

Lemma 2

Lemma 2

$$\mathbf{M} \succ 0,\ \hat{S} \text{ is proper, and } \mathbb{P}(\hat{S} = \emptyset) = 0$$

A B

$$0 \prec \mathbf{D}(p)\left(\mathbb{E}\left[\mathbf{M}_{\hat{S}}\right]\right)^{-1}\mathbf{D}(p) \preceq \mathbb{E}\left[\left(\mathbf{M}_{\hat{S}}\right)^{-1}\right]$$

Proof of Lemma 2 A

Follows from

- Lemma 1, and
- the fact that for proper $\hat{S}$ we have $p > 0$ and hence $\mathbf{D}(p) \succ 0$.

Proof of Lemma 2 B

Fix $h \in \mathbb{R}^n$. For arbitrary $\emptyset \neq S \subseteq [n]$ and $y \in \mathbb{R}^n$ we have:

$$\begin{aligned}
\frac{1}{2} h^\top \left(\mathbf{M}_S\right)^{-1} h &= \frac{1}{2} h_S^\top \left(\mathbf{M}_S\right)^{-1} h_S \\
&= \max_{x \in \mathbb{R}^n} \langle x, h_S \rangle - \frac{1}{2} x^\top \mathbf{M}_S x \\
&\geq \langle y, h_S \rangle - \frac{1}{2} y^\top \mathbf{M}_S y.
\end{aligned}$$

Proof of Lemma 2 B

Substituting $S = \hat{S}$ and taking expectations, we obtain

$$\begin{aligned}\frac{1}{2}\mathbb{E}\left[h^\top \left(\mathbf{M}_{\hat{S}}\right)^{-1} h\right] &\geq \mathbb{E}\left[\langle y, h_{\hat{S}}\rangle - \frac{1}{2} y^\top \mathbf{M}_{\hat{S}} y\right] \\ &= y^\top \mathbf{D}(p) h - \frac{1}{2} y^\top \mathbb{E}\left[\mathbf{M}_{\hat{S}}\right] y.\end{aligned}$$

Proof of Lemma 2 B

Finally, maximizing in y gives:

$$\begin{aligned}\frac{1}{2}h^\top \mathbb{E}\left[\left(\mathbf{M}_{\hat{S}}\right)^{-1}\right]h &\geq \max_{y\in\mathbb{R}^n} y^\top \mathbf{D}(p)h - \frac{1}{2}y^\top \mathbb{E}\left[\mathbf{M}_{\hat{S}}\right]y \\ &= \frac{1}{2}h^\top \mathbf{D}(p)\left(\mathbb{E}\left[\mathbf{M}_{\hat{S}}\right]\right)^{-1}\mathbf{D}(p)h.\end{aligned}$$

Proof of Theorem 1: First Steps

- From $\mathbf{G}$-strong convexity of f (by minimizing both sides in h) we get:

$$f(x) - f(x^*) \leq \frac{1}{2}\langle \nabla f(x), \mathbf{G}^{-1}\nabla f(x)\rangle, \quad \forall x \in \mathbb{R}^n \qquad (*)$$

- From $\mathbf{M}$-smoothness of f we get:

$$f(x^k + \mathbf{I}_{S_k} h) \leq f(x^k) + \langle \nabla f(x^k), \mathbf{I}_{S_k} h\rangle + \frac{1}{2}\langle \mathbf{M}_{S_k} h, h\rangle, \quad \forall h \in \mathbb{R}^n \quad (**)$$

Proof of Theorem 1: Method 1

- Use (**) with $h \leftarrow h^k := -(\mathbf{M}_{S_k})^{-1}\nabla f(x^k)$:

$$f(x^{k+1}) \leq f(x^k) - \frac{1}{2}\langle \nabla f(x^k), (\mathbf{M}_{S_k})^{-1}\nabla f(x^k)\rangle$$

- Taking conditional expectations on both sides:

$$\begin{aligned}
\mathbb{E}[f(x^{k+1}) \mid x^k] - f(x^k) &\leq -\frac{1}{2}\langle \nabla f(x^k), \mathbb{E}[\left(\mathbf{M}_{\hat{S}}\right)^{-1}]\nabla f(x^k)\rangle \\
&\overset{\text{def of } \sigma_1}{\leq} -\frac{\sigma_1}{2}\langle \nabla f(x^k), \mathbf{G}^{-1}\nabla f(x^k)\rangle \\
&\overset{(*)}{\leq} -\sigma_1\left(f(x^k) - f(x^*)\right)
\end{aligned}$$

- Rearrange the inequality and take expectation to get:

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

Proof of Theorem 1: Method 2

- Let $\mathbf{D} = \mathbf{D}(p)$ and take expectations on both sides of (**):

$$\mathbb{E}[f(x^k + \mathbf{I}_{S_k} h) \mid x^k] \leq f(x^k) + \langle \mathbf{D}\nabla f(x^k), h\rangle + \frac{1}{2}\langle \mathbb{E}[\mathbf{M}_{S_k}]h, h\rangle$$

- Note that the choice $\tilde{h}^k := -(\mathbb{E}[\mathbf{M}_{\hat{S}}])^{-1}\mathbf{D}\nabla f(x^k)$ minimizes the RHS of the inequality in h. Since $h^k = \mathbf{I}_{S_k}\tilde{h}^k$,

$$\begin{aligned}
\mathbb{E}[f(x^{k+1}) \mid x^k] - f(x^k) &\leq -\frac{1}{2}\langle \nabla f(x^k), \mathbf{D}\left(\mathbb{E}[\mathbf{M}_{\hat{S}}]\right)^{-1}\mathbf{D}\nabla f(x^k)\rangle \\
&\overset{\text{def of } \sigma_2}{\leq} -\frac{\sigma_2}{2}\langle \nabla f(x^k), \mathbf{G}^{-1}\nabla f(x^k)\rangle \\
&\overset{(*)}{\leq} -\sigma_2\left(f(x^k) - f(x^*)\right)
\end{aligned}$$

- Rearrange the inequality and take expectation to get:

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

Proof of Theorem 1: Method 3

Same as for Method 2, except in the first inequality replace $\mathbb{E}[\mathbf{M}_{S_k}]$ by the upper bound:

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

Ordering Theorem

Theorem 2 $\sigma_3 \leq \sigma_2 \leq \sigma_1$

Proof:

$$\begin{aligned}
\mathbf{D}(p)\mathbf{D}(v^{-1}) &= \mathbf{D}(p)\mathbf{D}(p^{-1})\mathbf{D}(v^{-1})\mathbf{D}(p) \\
&\overset{\text{ESO}}{\preceq} \mathbf{D}(p)\left(\mathbb{E}\left[\mathbf{M}_{\hat{S}}\right]\right)^{-1}\mathbf{D}(p) \\
&\overset{\text{Lemma 2}}{\preceq} \mathbb{E}\left[\left(\mathbf{M}_{\hat{S}}\right)^{-1}\right]
\end{aligned}$$

$$\begin{aligned}
\sigma_1 &:= \lambda_{\min}\left(\mathbf{G}^{1/2}\mathbb{E}\left[\left(\mathbf{M}_{\hat{S}}\right)^{-1}\right]\mathbf{G}^{1/2}\right) \\
\sigma_2 &:= \lambda_{\min}\left(\mathbf{G}^{1/2}\mathbf{D}(p)\left(\mathbb{E}\left[\mathbf{M}_{\hat{S}}\right]\right)^{-1}\mathbf{D}(p)\mathbf{G}^{1/2}\right) \\
\sigma_3 &:= \lambda_{\min}\left(\mathbf{G}^{1/2}\mathbf{D}(p)\mathbf{D}(v^{-1})\mathbf{G}^{1/2}\right)
\end{aligned}$$