Training Machine Learning Models via Empirical Risk Minimization (Lecture 2)

Peter Richtárik

The 41st Woudschoten Conference - October 5-7, 2016

Part 1

Lecture 1 Condensed to 2 Slides

Lecture 1

- Empirical Risk Minimization
  - Primal Formulation (minimize the average of *n* convex functions of *d* variables)
  - Dual Formulation (maximize a concave function of *n* variables)
- 5 Basic Tools
  - Gradient Descent (GD)
  - Accelerated Gradient Descent (AGD)
  - Handling Nonsmoothness: Proximal Gradient Descent:
  - Randomized Decomposition
    - Stochastic Gradient Descent (SGD)
    - Randomized Coordinate Descent (RCD)
  - Parallelism / Minibatching

Summary of Complexity Results from Lecture 1

Method	# iterations	Cost of 1 iter.
Gradient Descent (GD)	$\frac{L}{\mu}\log(1/\epsilon)$	n
Accelerated Gradient Descent (AGD)	$\sqrt{\frac{L}{\mu}}\log(1/\epsilon)$	n
Proximal Gradient Descent (PGD)	$\frac{L}{\mu}\log(1/\epsilon)$	n + Prox Step
Stochastic Gradient Descent (SGD)	$\left(\frac{\max_i L_i}{\mu}+\frac{\sigma^2}{\mu^2\epsilon}\right)\log(1/\epsilon)$	1
Randomized Coordinate Descent (RCD)	$\frac{\max_i L_i}{\mu}\log(1/\epsilon)$	1

Part 2
Arbitrary Sampling
(A Unified Theory of Deterministic and Randomized Gradient-Type Methods)

P.R. and Martin Takáč
On optimal probabilities in stochastic coordinate descent methods
Optimization Letters 10(6), 1233-1243, 2016 *(arXiv:1310.3438)*

The Problem

The Problem

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

Smooth and λ-strongly convex

The Algorithm

i.i.d. with arbitrary distribution

$S_t \subseteq \{1, 2, \ldots, n\}$

Choose a random set S_t of coordinates

For $i \in S_t$ do

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

For $i \notin S_t$ do

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

Example $n = 3$

$$e_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \quad e_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$$

Complexity

Key Assumption

Parameters $v_1, \dots, v_n$ satisfy:

$$\mathbf{E}\left[f\left(x + \sum_{i \in \hat{S}_t} h_i e_i\right)\right] \leq f(x) + \sum_{i=1}^n p_i \nabla_i f(x) h_i + \sum_{i=1}^n p_i v_i h_i^2$$

Inequality must hold for all

$x, h \in \mathbb{R}^n$

$p_i = \mathbf{P}(i \in \hat{S}_t)$

Complexity Theorem

$$t \geq \left(\max_i \frac{v_i}{p_i \lambda} \right) \log \left(\frac{f(x^0) - f(x^*)}{\epsilon \rho} \right)$$

$p_i = \mathbf{P}(i \in S_t)$

strong convexity constant

$$\mathbf{P}\left(f(x^t) - f(x^*) \leq \epsilon \right) \geq 1 - \rho$$

Uniform vs Optimal Sampling

$$p_i = \frac{1}{n} \quad \Rightarrow \quad \max_i \frac{v_i}{p_i \lambda} = \frac{n \max_i v_i}{\lambda}$$

$$p_i = \frac{v_i}{\sum_i v_i} \quad \Rightarrow \quad \max_i \frac{v_i}{p_i \lambda} = \frac{\sum_i v_i}{\lambda}$$

How to Compute the Stepsize Parameters?

Zheng Qu and P.R.
Coordinate descent with arbitrary sampling I: algorithms and complexity
Optimization Methods and Software 31(5), 829-857, 2016
(arXiv:1412.8060)

Zheng Qu and P.R.
Coordinate descent with arbitrary sampling II: expected separable overapproximation
Optimization Methods and Software 31(5), 858-884, 2016

Part 3
Quartz

Zheng Qu, P.R. and Tong Zhang
Quartz: Randomized dual coordinate ascent with arbitrary sampling
In *Advances in Neural Information Processing Systems* 28, 2015

Empirical Risk Minimization

Statistical Nature of Data

Label

$$(A_i, y_i) \sim Distribution$$

Data
(e.g., image, text, measurements, ...)

$$A_i \in \mathbb{R}^{d \times m}$$
$$y_i \in \mathbb{R}^m$$

Prediction of Labels from Data

Find $w \in \mathbb{R}^d$

Linear predictor

Such that when (data, label) pair is drawn from the distribution

$$(A_i, y_i) \sim Distribution$$

Then

$$A_i^\top w \approx y_i$$

Predicted label

True label

Measure of Success

$$loss(a, b)$$

data

label

We want the expected loss (=risk) to be small:

$$\mathbf{E}\left[loss(A_i^\top w, y_i)\right]$$

$$(A_i, y_i) \sim Distribution$$

Finding a Linear Predictor via Empirical Risk Minimization (ERM)

Draw i.i.d. data (samples) from the distribution

$$(A_1, y_1), (A_2, y_2), \ldots, (A_n, y_n) \sim Distribution$$

Output predictor which minimizes the empirical risk:

$$\min_{w \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} loss(A_i^\top w, y_i)$$

ERM:
Primal & Dual Problems

Primal Problem

Loss: convex & $1/\gamma$ - smooth

$$\|\nabla\phi_i(t) - \nabla\phi_i(t')\| \leq \gamma^{-1} \cdot \|t - t'\|$$

Lipschitz constant

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

Regularizer: 1-strongly convex

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

Dual Problem

$$D(\alpha) \equiv -\lambda g^* \left(\underbrace{\frac{1}{\lambda n} \sum_{i=1}^{n} A_i \alpha_i}_{\in \mathbb{R}^d} \right) - \frac{1}{n} \sum_{i=1}^{n} \phi_i^* (\overbrace{-\alpha_i}^{\in \mathbb{R}^m})$$

1 – smooth
& convex

γ - strongly convex

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

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

$$\max_{\alpha = (\underbrace{\alpha_1}_{\in \mathbb{R}^m}, \ldots, \underbrace{\alpha_n}_{\in \mathbb{R}^m}) \in \mathbb{R}^N = \mathbb{R}^{nm}} D(\alpha)$$

Fenchel Duality

$$\bar{\alpha} = \frac{1}{\lambda n} \sum_{i=1}^{n} A_i \alpha_i$$

$$P(w) - D(\alpha) \;=\; \lambda\left(g(w) + g^*\left(\bar{\alpha}\right)\right) + \frac{1}{n}\sum_{i=1}^{n} \phi_i(A_i^\top w) + \phi_i^*(-\alpha_i) =$$

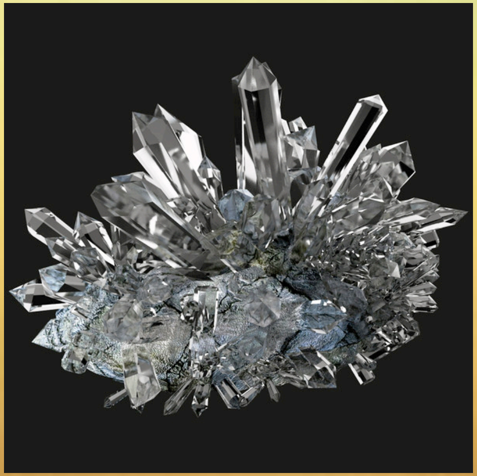
$$\underbrace{\lambda(g(w) + g^*\left(\bar{\alpha}\right) - \langle w, \bar{\alpha}\rangle)}_{\geq 0} + \frac{1}{n}\sum_{i=1}^{n} \underbrace{\phi_i(A_i^\top w) + \phi_i^*(-\alpha_i) + \left\langle A_i^\top w, \alpha_i \right\rangle}_{\geq 0}$$

Weak duality

Optimality conditions

$$w = \nabla g^*(\bar{\alpha}) \qquad\qquad \alpha_i = -\nabla \phi_i(A_i^\top w)$$

Quartz Algorithm



$$(\alpha^t, w^t) \quad \Rightarrow \quad (\alpha^{t+1}, w^{t+1})$$

Quartz: Bird's Eye View

$\theta = \min_i \frac{p_i \lambda\gamma n}{v_i + \lambda\gamma n}$

STEP 1: PRIMAL UPDATE

$$w^{t+1} \leftarrow (1-\theta)w^t + \theta\nabla g^*(\bar{\alpha}^t)$$

STEP 2: DUAL UPDATE

Choose a random set S_t of dual variables

$p_i = \mathbf{P}(i \in S_t)$

For $i \in S_t$ do

$$\alpha_i^{t+1} \leftarrow \left(1 - \frac{\theta}{p_i}\right)\alpha_i^t + \frac{\theta}{p_i}\left(-\nabla\phi_i(A_i^\top w^{t+1})\right)$$

Randomized Primal-Dual Methods

Algorithm	1-nice	1-optimal	τ-nice	arbitrary	additional speedup	direct p-d analysis	acceleration
SDCA	•						
mSDCA	•		•		•		
ASDCA	•		•				•
AccProx-SDCA	•						•
DisDCA	•		•				
Iprox-SDCA	•	•					
APCG	•						•
SPDC	•	•	•			•	•
Quartz	•	•	•	•	•	•	

SDCA:	SS Shwartz & T Zhang, 09/2012
mSDCA	M Takac, A Bijral, P R & N Srebro, 03/2013
ASDCA:	SS Shwartz & T Zhang, 05/2013
AccProx-SDCA:	SS Shwartz & T Zhang, 10/2013
DisDCA:	T Yang, 2013
Iprox-SDCA:	P Zhao & T Zhang, 01/2014
APCG:	Q Lin, Z Lu & L Xiao, 07/2014
SPDC:	Y Zhang & L Xiao, 09/2014
Quartz:	Z Qu, P R & T Zhang, 11/2014

Complexity

Assumption 3
(Expected Separable Overapproximation)

Parameters $v_1, \dots, v_n$ satisfy:

$$\mathbf{E} \left\| \sum_{i \in S_t} A_i \alpha_i \right\|^2 \leq \sum_{i=1}^{n} p_i v_i \|\alpha_i\|^2$$

inequality must hold for all

$\alpha_1, \dots, \alpha_n \in \mathbb{R}^m$

$p_i = \mathbf{P}(i \in S_t)$

Complexity Theorem (QRZ'14)

$$t \;\geq\; \max_i \left(\frac{1}{p_i} + \frac{v_i}{p_i \lambda \gamma n} \right) \log \left(\frac{P(w^0) - D(\alpha^0)}{\epsilon} \right)$$

$$\mathbf{E}\left[P(w^t) - D(\alpha^t) \right] \leq \epsilon$$

Part 4

Quartz: Special Cases

Special Case 1: Serial Sampling

Complexity

Optimal sampling $p_i = \frac{L_i}{\sum_j L_j}$	$n + \frac{\frac{1}{n}\sum_{i=1}^n L_i}{\lambda\gamma}$
Uniform sampling $p_i = \frac{1}{n}$	$n + \frac{\max_i L_i}{\lambda\gamma}$

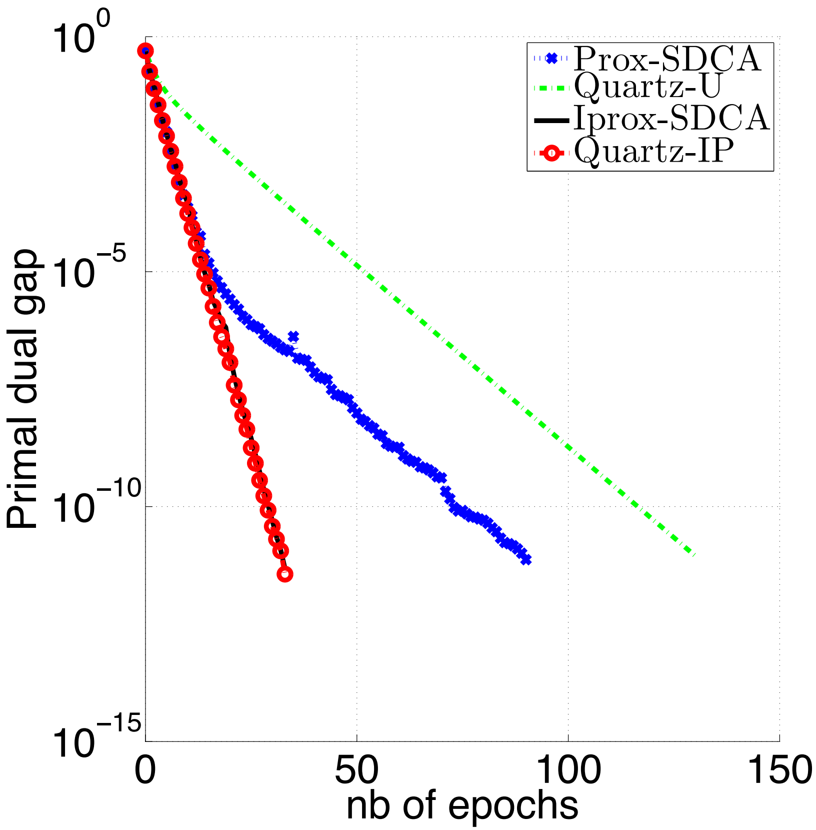
$$L_i \equiv \lambda_{\max}\left(A_i^\top A_i\right)$$

Data

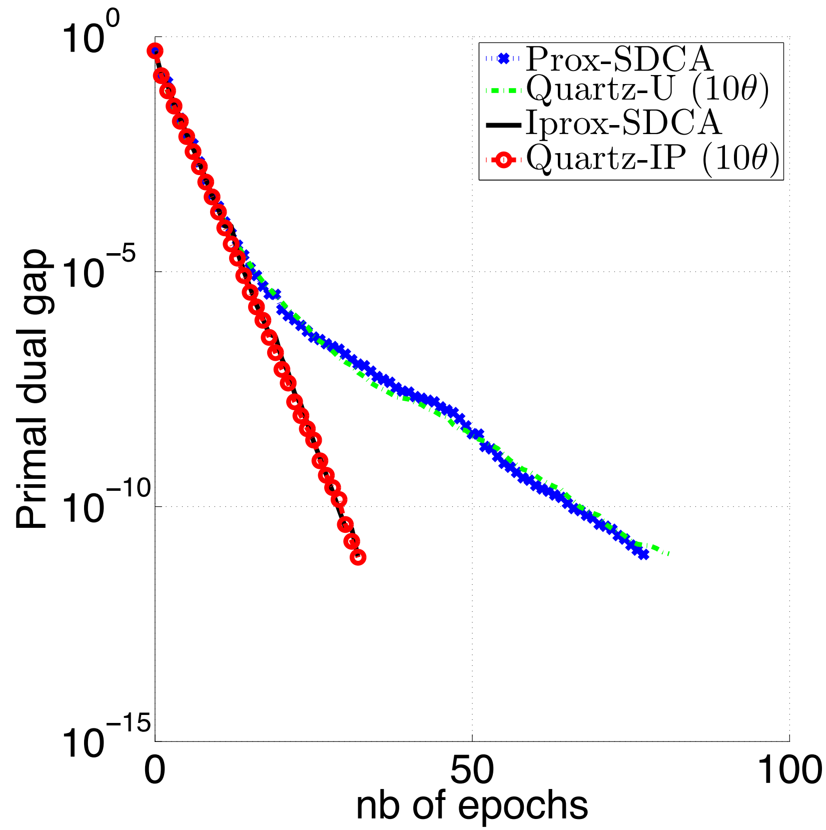
Dataset	**# Samples** n	**# features** d	**density** $nnz(A)/(nd)$
astro-ph	29,882	99,757	0.08%
CCAT	781,265	47,236	0.16%
cov1	522,911	54	22.22%
w8a	49,749	300	3.91%
ijcnn1	49,990	22	59.09%
webspam	350,000	254	33.52%

Experiment: Quartz vs SDCA, Uniform vs Optimal Sampling

Standard primal update



"Aggressive" primal update



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

Special Case 2: Minibatching & Sparsity

Data Sparsity

$$1 \leq \tilde{\omega} \leq n$$

"Fully sparse data"

A normalized measure of average sparsity of the data

"Fully dense data"

Complexity of Quartz

Fully sparse data ($\tilde{\omega} = 1$)	$\frac{n}{\tau} + \frac{\max_i L_i}{\lambda\gamma\tau}$
Fully dense data ($\tilde{\omega} = n$)	$\frac{n}{\tau} + \frac{\max_i L_i}{\lambda\gamma}$
Any data ($1 \leq \tilde{\omega} \leq n$)	$\frac{n}{\tau} + \frac{\left(1 + \frac{(\tilde{\omega}-1)(\tau-1)}{n-1}\right)\max_i L_i}{\lambda\gamma\tau}$

$$\equiv T(\tau)$$

Speedup

Assume the data is normalized: $L_i \equiv \lambda_{\max}(A_i^\top A_i) \leq 1$

Then:

$$T(\tau) = \frac{\left(1 + \frac{(\tilde{\omega}-1)(\tau-1)}{(n-1)(1+\lambda\gamma n)}\right)}{\tau} \times T(1)$$

Linear speedup up to a certain data-independent minibatch size:

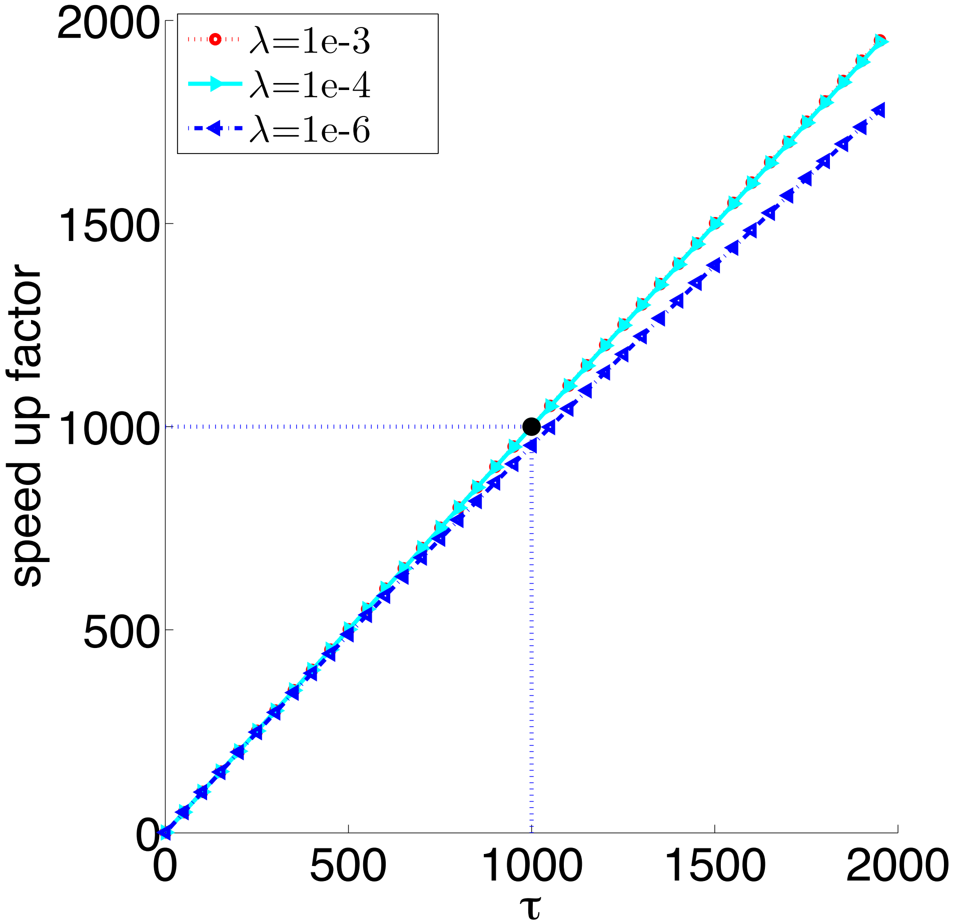
$$\tau \leq 2 + \lambda\gamma n \quad \Rightarrow \quad T(\tau) \leq \frac{2}{\tau} \times T(1)$$

Further data-dependent speedup, up to the extreme case:

$$\tilde{\omega} = \mathcal{O}(\lambda\gamma n) \quad \Rightarrow \quad T(\tau) = \mathcal{O}\left(\frac{T(1)}{\tau}\right)$$

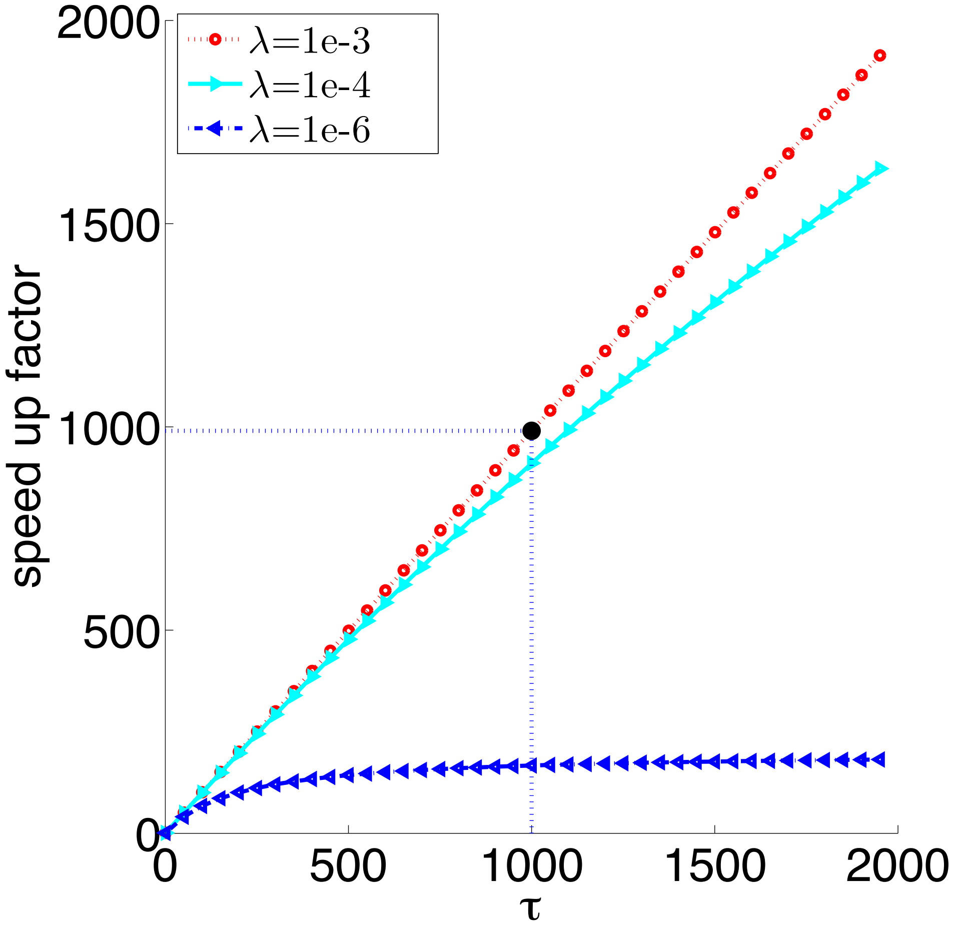
Speedup: sparse data

$$n = 10^6, \tilde{\omega} = 10^2, \gamma = 1$$



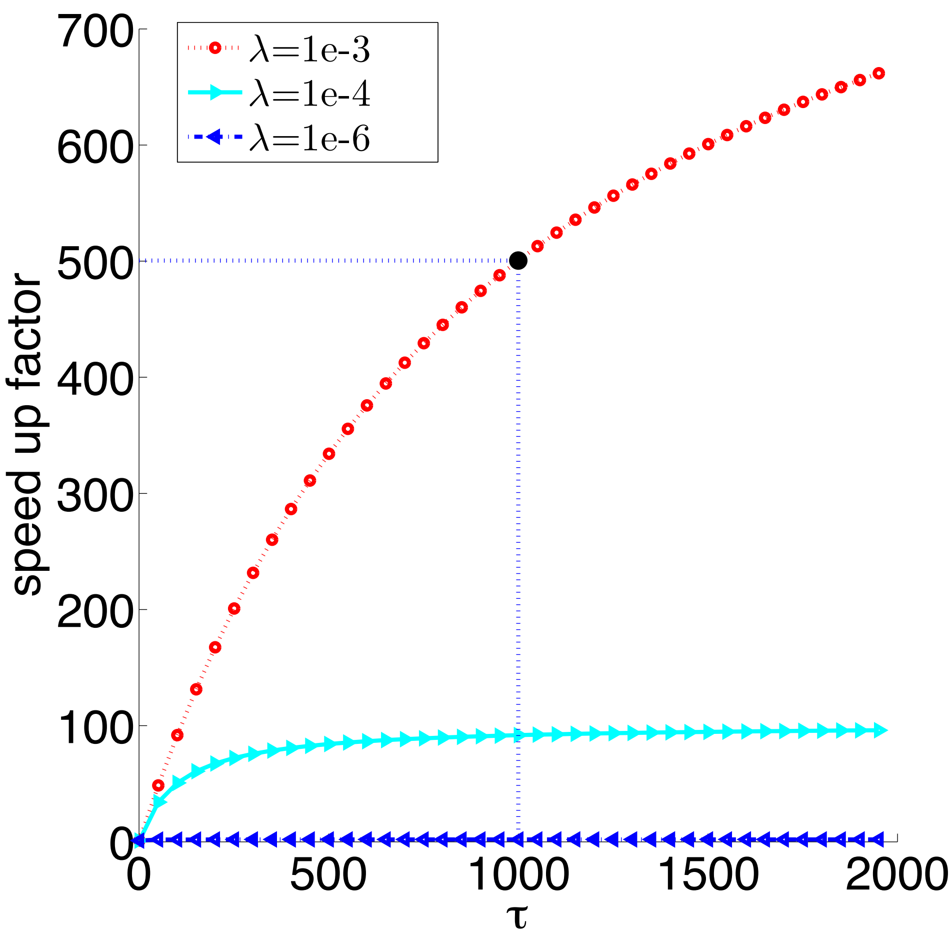
Speedup: denser data

$$n = 10^6,\ \tilde{\omega} = 10^4,\ \gamma = 1$$

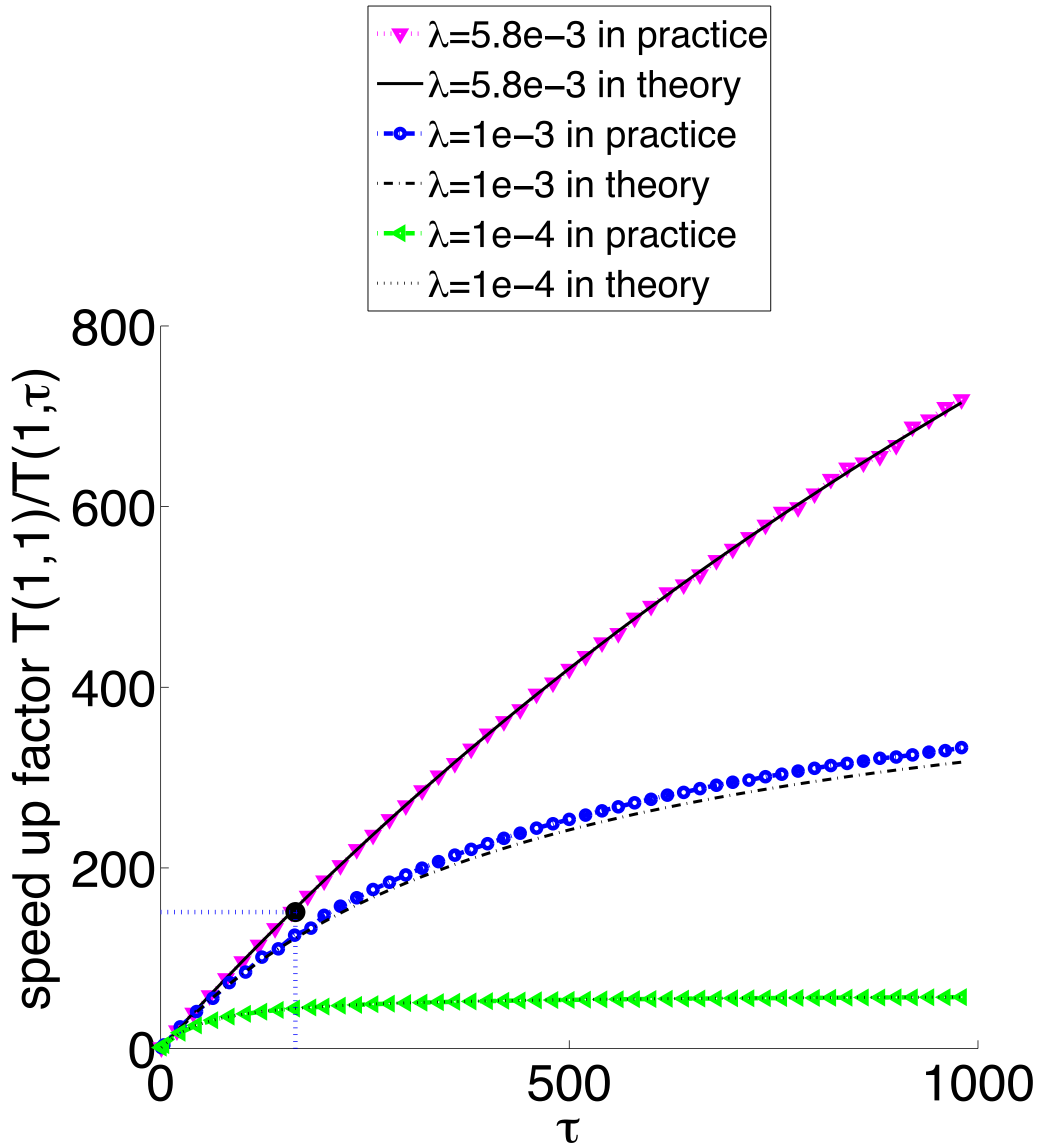


Speedup: fully dense data

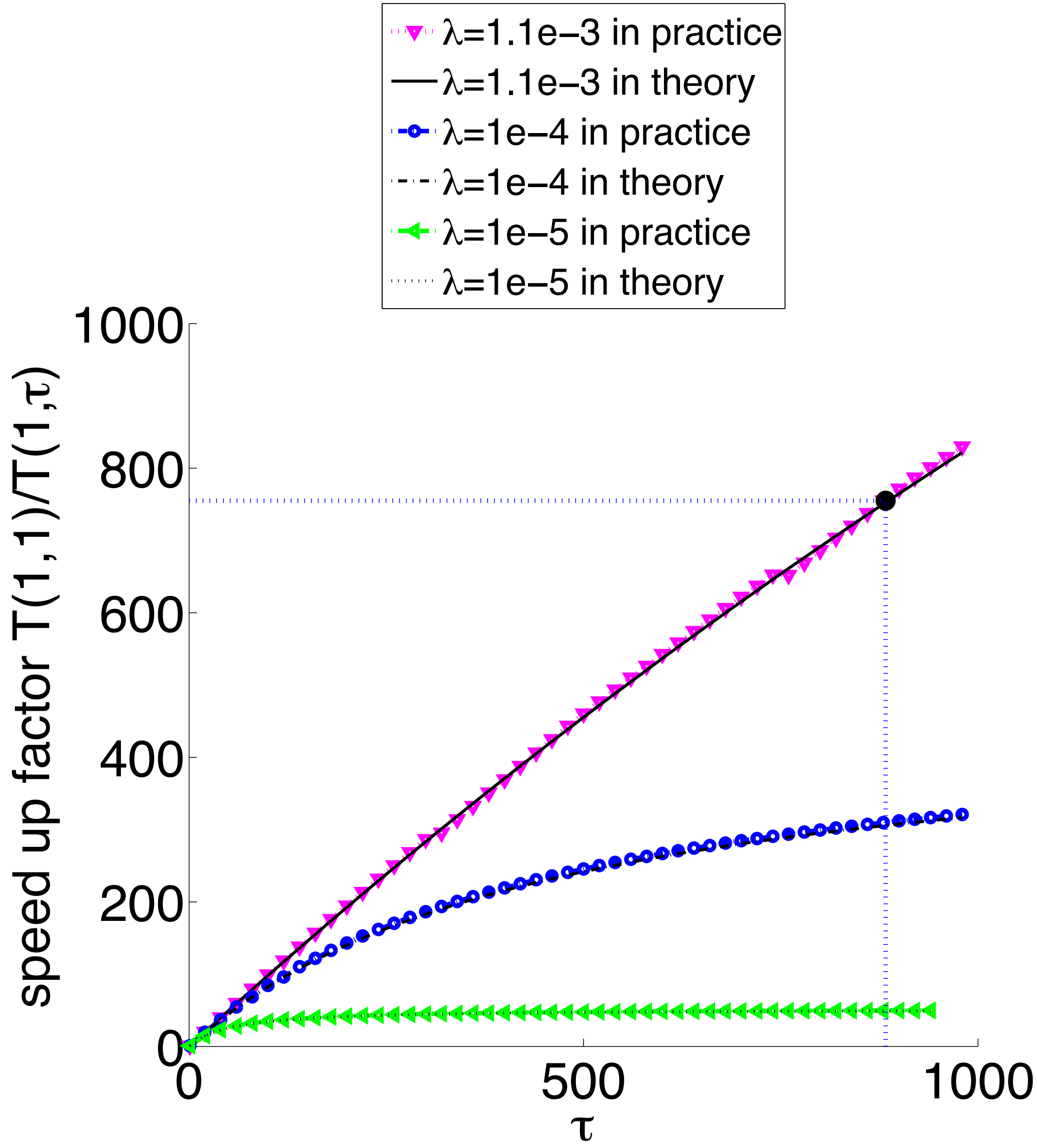
$$n = 10^6,\ \tilde{\omega} = 10^6,\ \gamma = 1$$



astro_ph: *n* = 29,882   density = 0.08%



CCAT: n = 781,265 density = 0.16%



Primal-Dual Methods with tau-nice Sampling

$L_i = 1$

Algorithm	Iteration complexity	g
SDCA SS-Shwartz & T Zhang '13	$n + \frac{1}{\lambda\gamma}$	$\frac{1}{2}\|\cdot\|^2$
ASDCA SS-Shwartz & T Zhang '13	$4 \times \max\left\{\frac{n}{\tau}, \sqrt{\frac{n}{\lambda\gamma\tau}}, \frac{1}{\lambda\gamma\tau}, \frac{n^{\frac{1}{3}}}{(\lambda\gamma\tau)^{\frac{2}{3}}}\right\}$	$\frac{1}{2}\|\cdot\|^2$
SPDC Y Zhang & L Xiao '14	$\frac{n}{\tau} + \sqrt{\frac{n}{\lambda\gamma\tau}}$	general
Quartz	$\frac{n}{\tau} + \left(1 + \frac{(\tilde{\omega} - 1)(\tau - 1)}{n - 1}\right)\frac{1}{\lambda\gamma\tau}$	general

For sufficiently sparse data, Quartz wins even when compared against accelerated methods

Algorithm	$\gamma\lambda n = \Theta(\frac{1}{\tau})$	$\gamma\lambda n = \Theta(1)$	$\gamma\lambda n = \Theta(\tau)$	$\gamma\lambda n = \Theta(\sqrt{n})$
	$\kappa = n\tau$	$\kappa = n$	$\kappa = n/\tau$	$\kappa = \sqrt{n}$
SDCA	$n\tau$	n	n	n
ASDCA	n	$\frac{n}{\sqrt{\tau}}$	$\frac{n}{\tau}$	$\frac{n}{\tau} + \frac{n^{3/4}}{\sqrt{\tau}}$
SPDC	n	$\frac{n}{\sqrt{\tau}}$	$\frac{n}{\tau}$	$\frac{n}{\tau} + \frac{n^{3/4}}{\sqrt{\tau}}$
Quartz	$n + \tilde{\omega}\tau$	$\frac{n}{\tau} + \tilde{\omega}$	$\frac{n}{\tau}$	$\frac{n}{\tau} + \frac{\tilde{\omega}}{\sqrt{n}}$

Accelerated

Special Case 3: Distributed Sampling

References

P.R. and Martin Takáč
Distributed coordinate descent for learning with big data
Journal of Machine Learning Research 17(75), 1-25, 2016
(arXiv:1310.2059)

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

Zheng Qu, P.R. and Tong Zhang
Quartz: Randomized dual coordinate ascent with arbitrary sampling
Neural Information Processing Systems 28, 865-873, 2015

Distributed Quartz: Perform the Dual Updates in a Distributed Manner

Quartz STEP 2: DUAL UPDATE

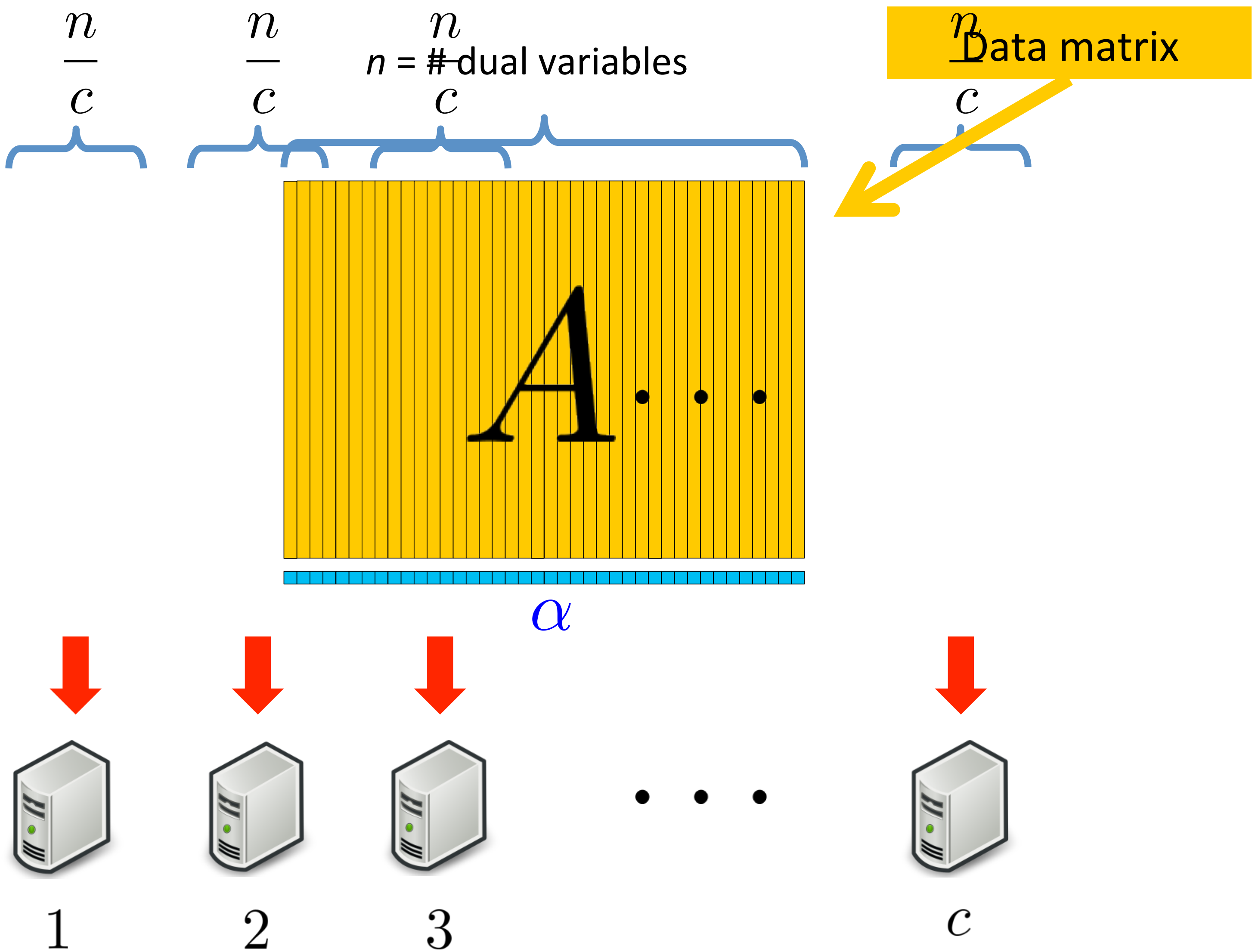
Choose a random set S_t of dual variables

For $i \in S_t$ do

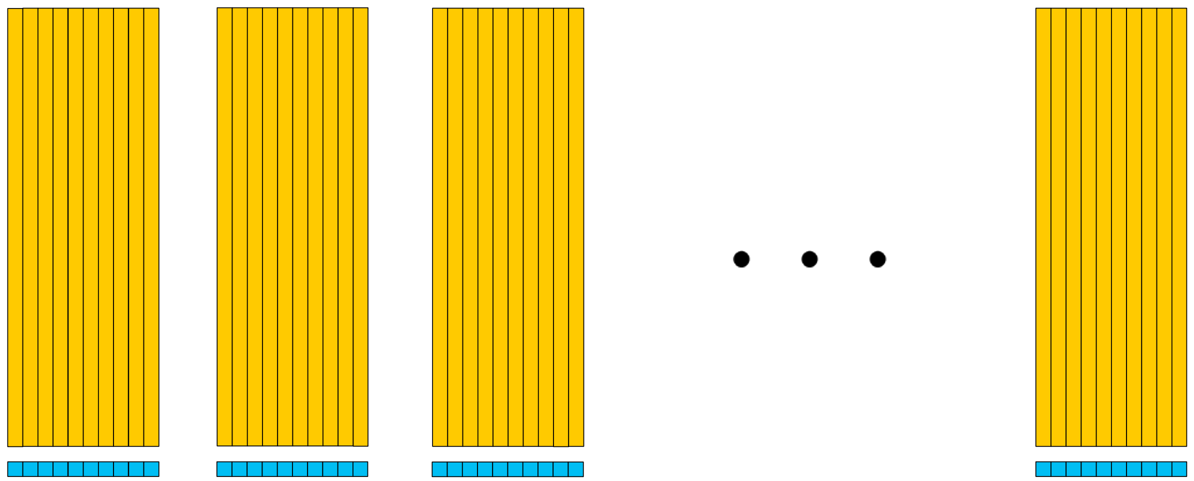
$$\alpha_i^{t+1} \leftarrow \left(1 - \frac{\theta}{p_i}\right)\alpha_i^t + \frac{\theta}{p_i}\left(-\nabla \phi_i(A_i^\top w^{t+1})\right)$$

Data required to compute the update

Distribution of Data



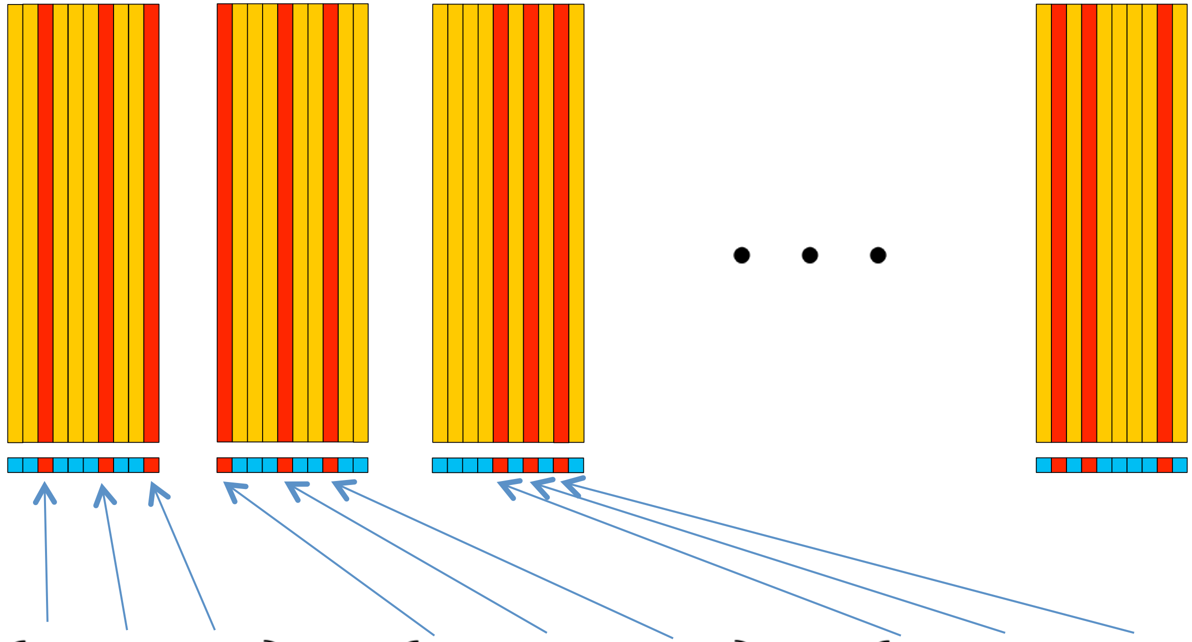
Distributed Sampling



Distributed Sampling

Each node independently picks τ dual variables from those it owns, uniformly at random

$\tau = 3$



$$\hat{S} = \{3, 7, 10\} \cup \{11, 15, 18\} \cup \{25, 27, 29\} \cup \cdots$$

Random set of dual variables

Also see: CoCoA+ [Ma, Smith, Jaggi et al 15]

Complexity of Distributed Quartz

Key: Get the right stepsize parameters *v* (so that the ESO inequality holds)

The leading term in the complexity bound then is:

$$\max_i \left(\frac{1}{p_i} + \frac{v_i}{p_i \lambda \gamma n} \right)$$

$$||$$

$$\frac{n}{c\tau} + \frac{\text{Something that looks complicated}}{\lambda \gamma c \tau}$$

$$||$$

$$\frac{n}{c\tau} + \max_i \frac{\lambda_{\max}\left(\sum_{j=1}^{d}\left(1 + \frac{(\tau-1)(\omega_j-1)}{\max\{n/c-1,1\}} + \left(\frac{\tau c}{n} - \frac{\tau-1}{\max\{n/c-1,1\}}\right)\frac{\omega'_j-1}{\omega'_j}\omega_j\right) A_{ji}^\top A_{ji}\right)}{\lambda\gamma c\tau}$$

Experiment

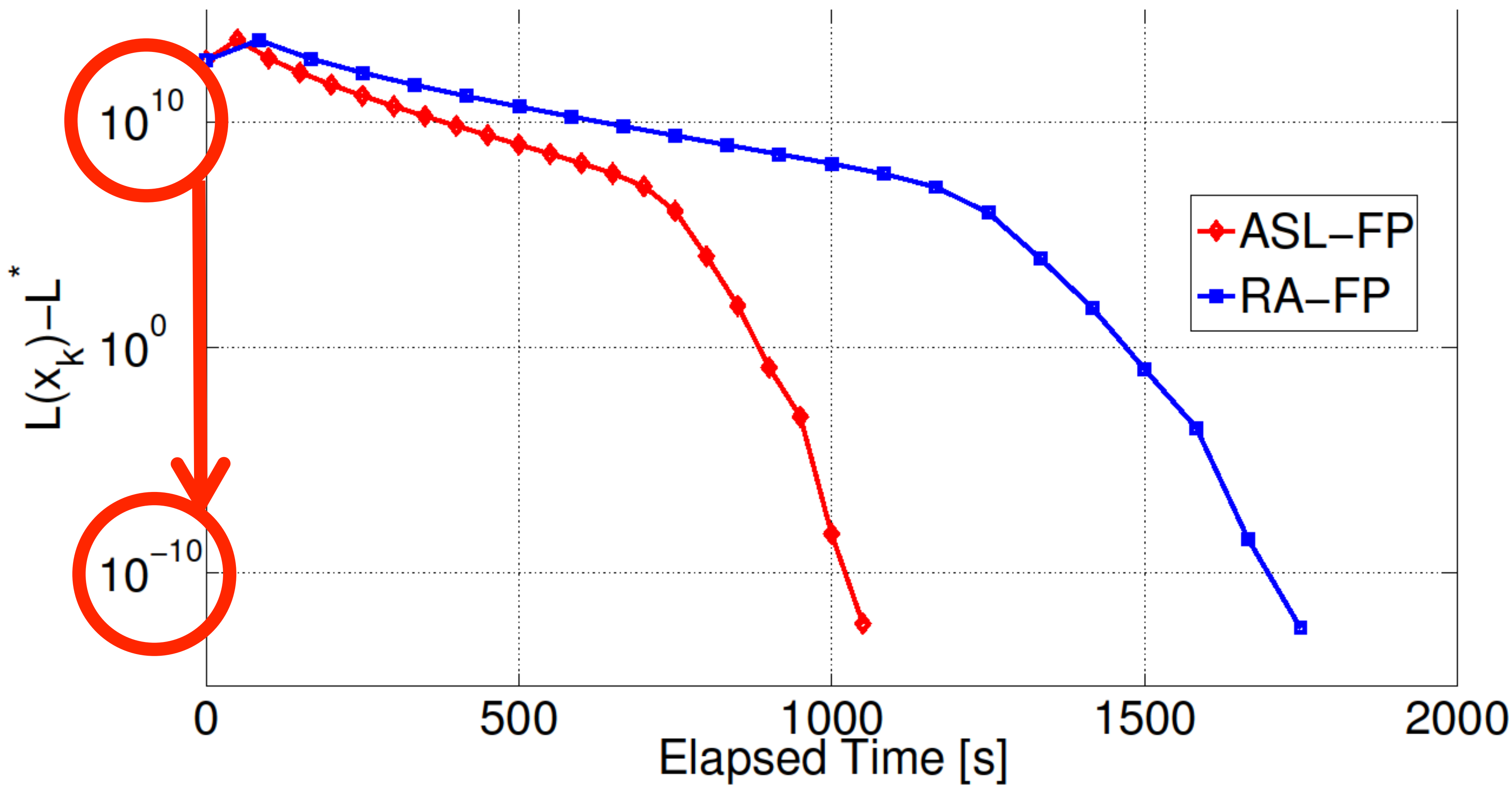
Machine: 128 nodes of Hector Supercomputer (4,096 cores)

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(75), 1-25, 2016
(arXiv:1310.2059)

LASSO: 3TB data + 128 nodes



Experiment (Acceleration)

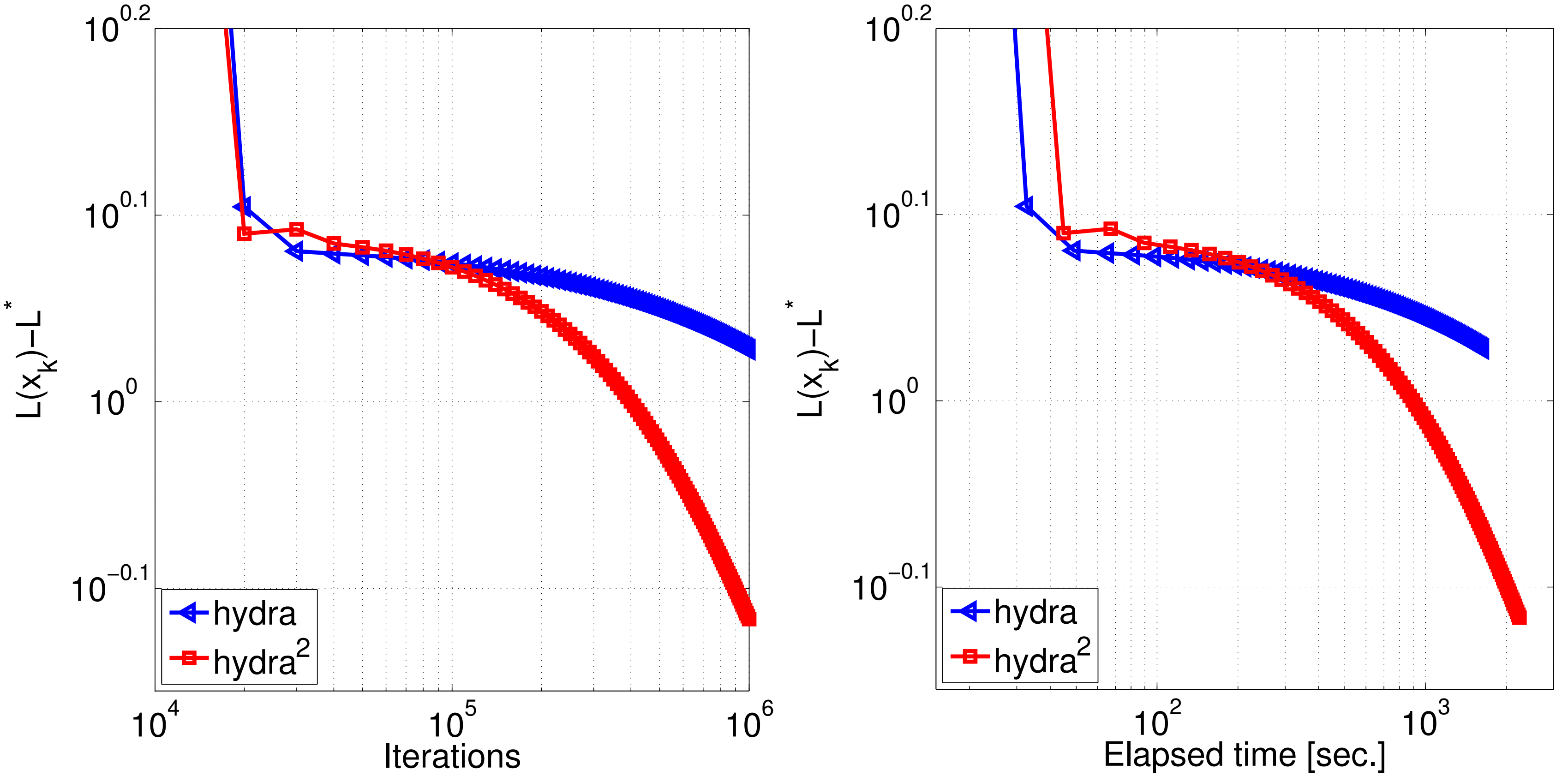
Machine: 128 nodes of Archer Supercomputer

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
IEEE Int. Workshop on Machine Learning for Signal Proc., 2014

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



THE END

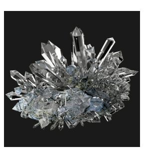
Coauthors

Martin Takáč
(Lehigh)

Zheng Qu
(Hong Kong)

Tong Zhang
(Baidu)

P.R. and Martin Takáč
On optimal probabilities in stochastic coordinate descent methods
Optimization Letters 10(6), 1233-1243, 2016 *(arXiv:1310.3438)*



Zheng Qu, P.R. and Tong Zhang
Quartz: Randomized dual coordinate ascent with arbitrary sampling
In *Advances in Neural Information Processing Systems* 28, 865-873, 2015 *(arXiv:1411.5873)*