Modern Optimization Methods for Big Data Problems
MATH11146

Lecture 8
Empirical Risk Minimization

Peter Richtárik

Machine Learning:

Training Linear Predictors via 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 ← $A_i^\top w$ → 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

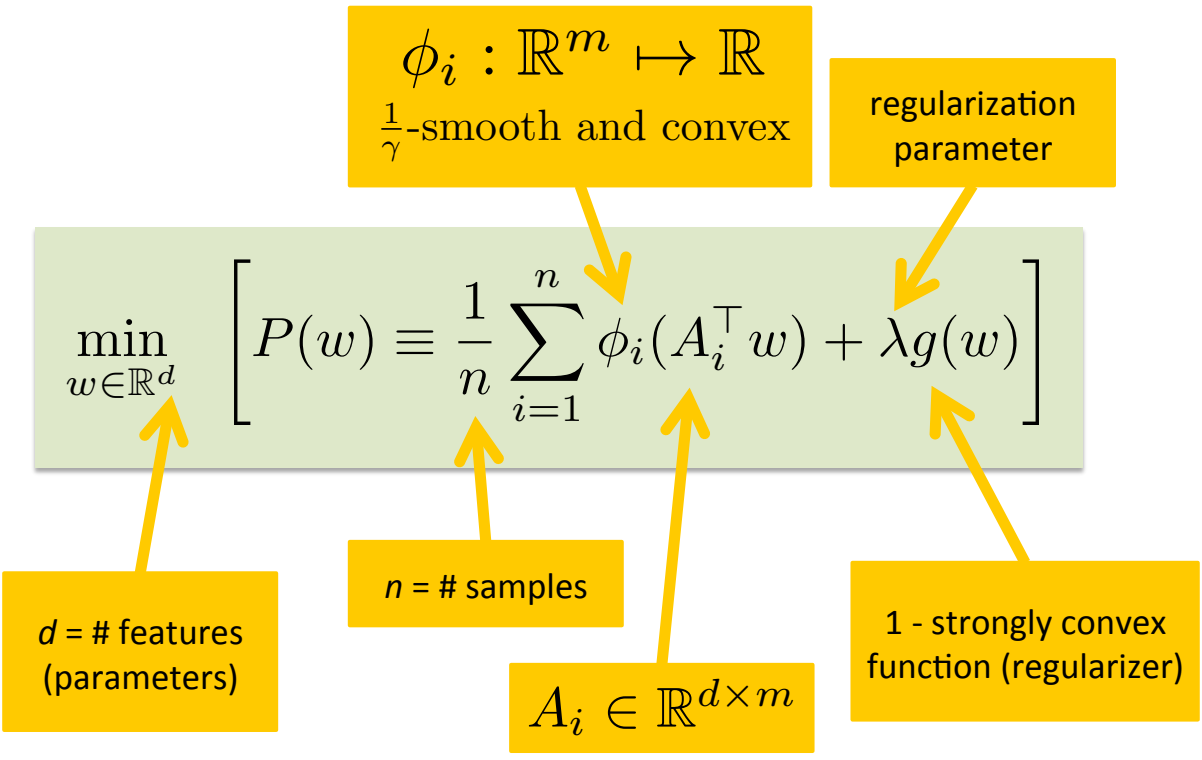
$$(A_1, y_1), (A_2, y_2), \ldots, (A_n, y_n) \sim Distribution$$

Output predictor which minimizes the empirical risk:

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

Primal and Dual Problems

Primal Problem: ERM



Is the difficulty in *n* or *d*?

- **Big n**
  - Work in the **primal**
  - Process one loss function (= one example) at a time
  - Type of methods: stochastic gradient descent (modern variants: SAG, SVRG, S2GD, mS2GD, SAGA, S2CD, MISO, FINITO, …)
- **Big d**
  - Work in the **primal**
  - Process one primal variable at a time
  - Type of methods: randomized coordinate descent (e.g., Hydra, Hydra2)
- **Big n**
  - Work in the **dual**
  - Process one dual variable (=one example) at a time
  - Type of methods: randomized coordinate descent (modern variants: RCDM, PCDM, Shotgun, SDCA, APPROX, Quartz, ALPHA, SDNA, SPDC, ASDCA, …)
  - E.g. SDCA = run coordinate descent on the dual problem

Dual Problem

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

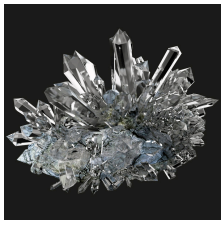
1 – smooth & convex

γ - strongly convex

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

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

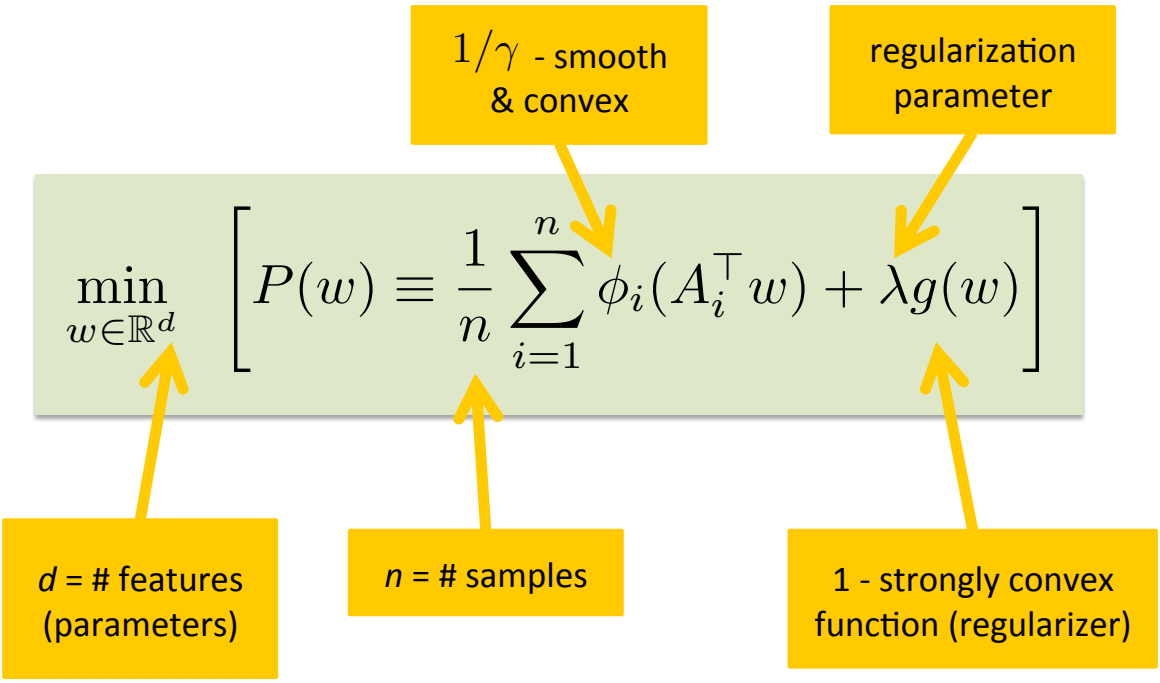
An Efficient Dual Method



Zheng Qu, P.R. and Tong Zhang
Randomized dual coordinate ascent with arbitrary sampling
In NIPS 2015 (arXiv:1411.5873)

Empirical Risk Minimization

Primal Problem: ERM



Assumption 1

The loss functions $\phi_i : \mathbb{R}^m \mapsto \mathbb{R}$ are $\frac{1}{\gamma}$-smooth:

$$\|\nabla \phi_i(a) - \nabla \phi_i(a')\| \leq \frac{1}{\gamma} \|a - a'\|, \quad a, a' \in \mathbb{R}^m$$

Lipschitz constant of the gradient of the function

Assumption 2

Regularizer is 1-strongly convex

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

subgradient

Dual Problem

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

$\in \mathbb{R}^m$

$\in \mathbb{R}^d$

1 – smooth & convex

γ - strongly convex

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

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

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

$\in \mathbb{R}^m \quad \in \mathbb{R}^m$

The Algorithm: Quartz

Fenchel Duality

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

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

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

≥ 0 ⟵ **Weak duality** ⟶ ≥ 0

Optimality conditions

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

The Algorithm

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

Quartz: Bird's Eye View

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

For $i \in S_t$ do

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

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

Algorithm 1 Quartz

Parameters: proper random sampling $\hat{S}$ and a positive vector $v \in \mathbb{R}^n$

Initialization: Choose $\alpha^0 \in \mathbb{R}^N$ and $w^0 \in \mathbb{R}^d$

Set $p_i = \mathbb{P}(i \in \hat{S})$, $\theta = \min_i \frac{p_i \lambda\gamma n}{v_i + \lambda\gamma n}$ and $\bar{\alpha}^0 = \frac{1}{\lambda n}\sum_{i=1}^n A_i \alpha_i^0$

Convex combination constant

for $t \geq 1$ **do**

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

$\alpha^t = \alpha^{t-1}$

Generate a random set $S_t \subseteq [n]$, following the distribution of $\hat{S}$

for $i \in S_t$ **do**

Calculate $\Delta\alpha_i^t$ using one of the following options:

Option I :

$\Delta\alpha_i^t = \arg\max_{\Delta \in \mathbb{R}^m} \left[-\phi_i^*(-(\alpha_i^{t-1} + \Delta)) - \nabla g^*(\bar{\alpha}^{t-1})^\top A_i \Delta - \frac{v_i \|\Delta\|^2}{2\lambda n} \right]$

Option II :

$\Delta\alpha_i^t = -\theta p_i^{-1} \alpha_i^{t-1} - \theta p_i^{-1} \nabla \phi_i(A_i^\top w^t)$

$\alpha_i^t = \alpha_i^{t-1} + \Delta\alpha_i^t$ **STEP 2**

end for

$\bar{\alpha}^t = \bar{\alpha}^{t-1} + (\lambda n)^{-1} \sum_{i \in S_t} A_i \Delta\alpha_i^t$

Just maintaining $\bar{\alpha}$

end for

Output: w^t, α^t

Some Other Stochastic Dual Methods for ERM

Randomized Dual Coordinate Ascent Methods for ERM

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 of Quartz

Assumption 3
(Expected Separable Overapproximation)

Parameters $v_1, \ldots, v_n$ 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 $\alpha_1, \ldots, \alpha_n \in \mathbb{R}^m$

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

Assumption 3 for *m* = 1

If *m* = 1 (labels are real numbers and not vectors), then the inequality from previous slide is equivalent to

$$\alpha^T \left(P \bullet A^T A\right) \alpha \leq \alpha^T \mathrm{Diag}\left(p \bullet v\right) \alpha$$

which is in turn equivalent to

$$P \bullet A^T A \preceq \mathrm{Diag}\left(p \bullet v\right)$$

- This is precisely the same inequality as in Theorem 10
- We have discussed at length how to compute *v* for which this holds

Complexity

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

Theorem [Qu, R & Zhang 14]

$$\mathbf{E}[P(w^t) - D(\alpha^t)] \leq (1-\theta)^t (P(w^0) - D(\alpha^0))$$

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

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

Example

Data: $n = 7 \times 10^5$

$\gamma = \frac{1}{4}$ $\quad v_i \equiv \lambda_{\max}(A_i^\top A_i) \leq 1$

Method: $|S_t| \equiv 1$ $\quad p_i = \frac{1}{n}$ $\quad \lambda = \frac{1}{n}$

$$(1-\theta)^n = 0.8187$$

$$(1-\theta)^{12n} = 0.0907 < \tfrac{1}{10}$$

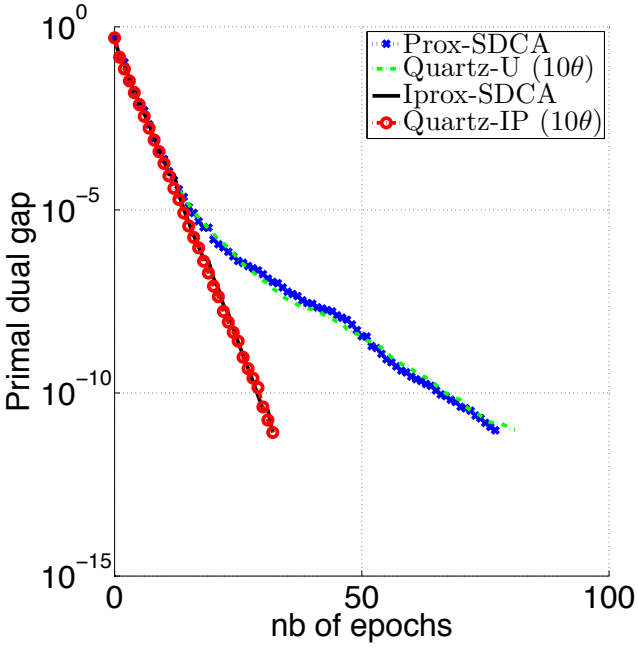
Updating One Dual Variable at a Time

Complexity of Quartz specialized to serial sampling

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

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

Experiment: Quartz vs SDCA, uniform vs optimal sampling



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

An Efficient Primal Method

S. Shalev-Shwartz
SDCA without Duality, *NIPS 2015* (arXiv:1502.06177)

Dominik Csiba and P.R.
Primal method for ERM with flexible mini-batching schemes and non-convex losses, *arXiv:1506.02227,* 2015

Empirical Risk Minimization

Primal Problem: ERM

$$\min_{w \in \mathbb{R}^d} \left[P(w) \equiv \frac{1}{n} \sum_{i=1}^{n} \phi_i(A_i^\top w) + \frac{\lambda}{2} \|w\|_2^2 \right]$$

$\phi_i : \mathbb{R}^m \mapsto \mathbb{R}$ is $\frac{1}{\gamma}$-smooth and convex

regularization parameter

d = # features (parameters)

n = # samples

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

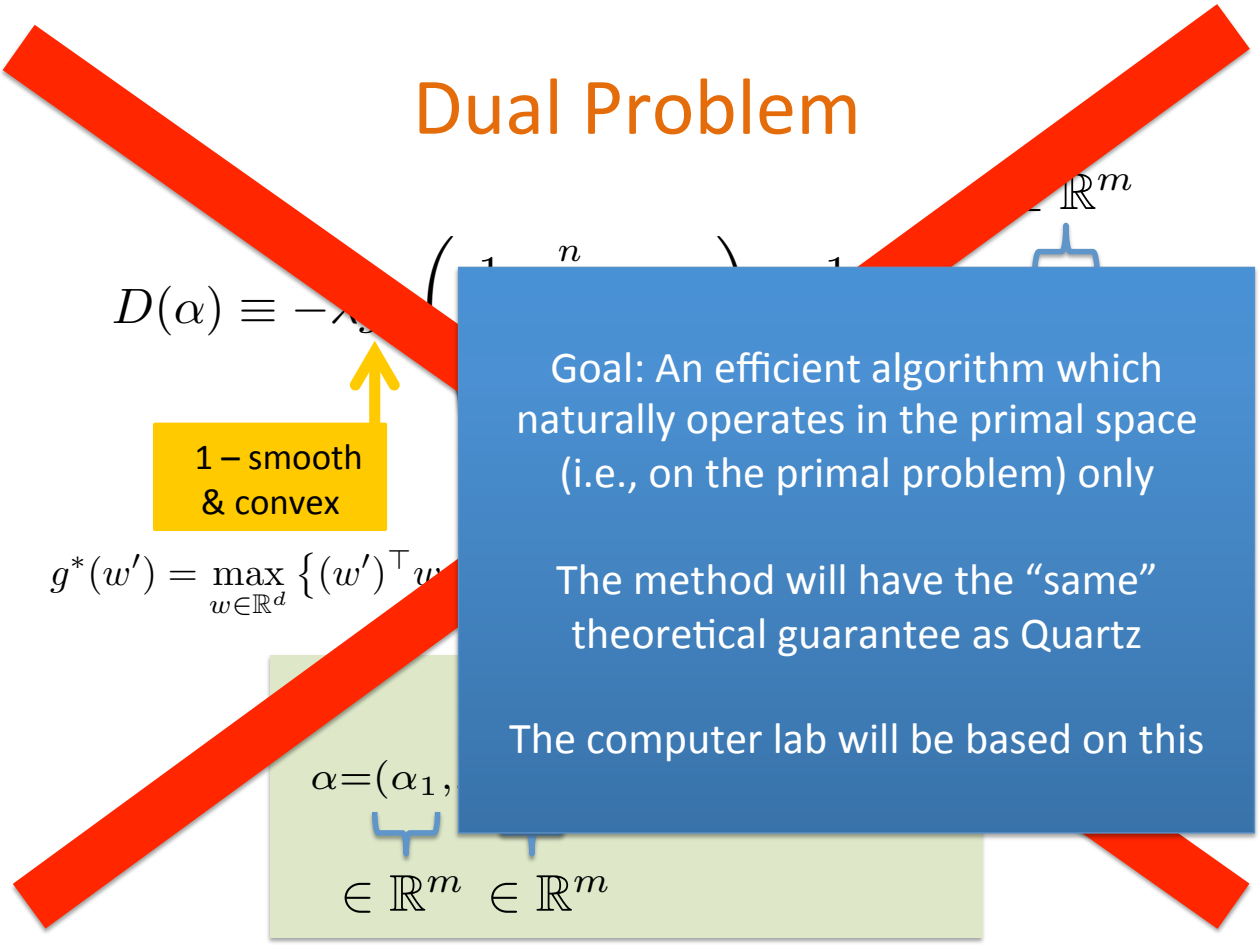
We had a general 1-strongly convex function g here before

Assumption

The loss functions $\phi_i : \mathbb{R}^m \mapsto \mathbb{R}$ are $\frac{1}{\gamma}$-smooth:

$$\|\nabla \phi_i(a) - \nabla \phi_i(a')\| \leq \frac{1}{\gamma} \|a - a'\|, \quad a, a' \in \mathbb{R}^m$$

Lipschitz constant of the gradient of the function



The Algorithm

Motivation I

w^* is optimal

$$0 = \nabla P(w^*) = \left(\frac{1}{n} \sum_{i=1}^{n} A_i \nabla \phi_i(A_i^\top w^*) \right) + \lambda w^*$$

$$w^* = \frac{1}{\lambda n} \sum_{i=1}^{n} A_i \alpha_i^*$$

$\alpha_i^* := -\nabla \phi_i(A_i^\top w^*)$

Motivation II

Algorithmic Ideas:

1. Simultaneously search for both w^* and $\alpha_1^*, \ldots, \alpha_n^*$
2. Try to do "something like"

$$\alpha_i^{t+1} \leftarrow -\nabla \phi_i(A_i^\top w^t)$$

3. Maintain the relationship

Does not quite work: too "greedy"

$$w^t = \frac{1}{\lambda n} \sum_{i=1}^{n} A_i \alpha_i^t$$

The Algorithm: dfSDCA

STEP 0: INITIALIZE

Choose $\alpha_1^0, \ldots, \alpha_n^0 \in \mathbb{R}^m$

Initialize the relationship

$$w^0 = \frac{1}{\lambda n} \sum_{i=1}^{n} A_i \alpha_i^0$$

STEP 1: "DUAL" UPDATE

Choose a random set S_t of "dual variables"

For $i \in S_t$ do

Controlling "greed" by taking a convex combination

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

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

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

STEP 2: PRIMAL UPDATE

$$w^{t+1} \leftarrow w^t - \sum_{i \in S_t} \frac{\theta}{n \lambda p_i} A_i \left(\nabla \phi_i(A_i^\top w^t) + \alpha_i^t\right)$$

This is just maintaining the relationship

Complexity

ESO Assumption (same as before!)

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 [Csiba & R '15]

A constant depending on $P, w^0, \alpha_i^0, w^*, \alpha_i^*$

$$t \geq \max_i \left(\frac{1}{p_i} + \frac{v_i}{p_i \lambda\gamma n}\right) \log\left(\frac{C}{\epsilon}\right)$$

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

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

Experiments

Some More Efficient Primal Methods for ERM: SAG, SVRG and S2GD

SAG: Stochastic Average Gradient

N. Le Roux, M. Schmidt, and F. Bach. **A stochastic gradient method with an exponential convergence rate for finite training sets**. *NIPS*, 2012

SVRG: Stochastic Variance Reduced Gradient

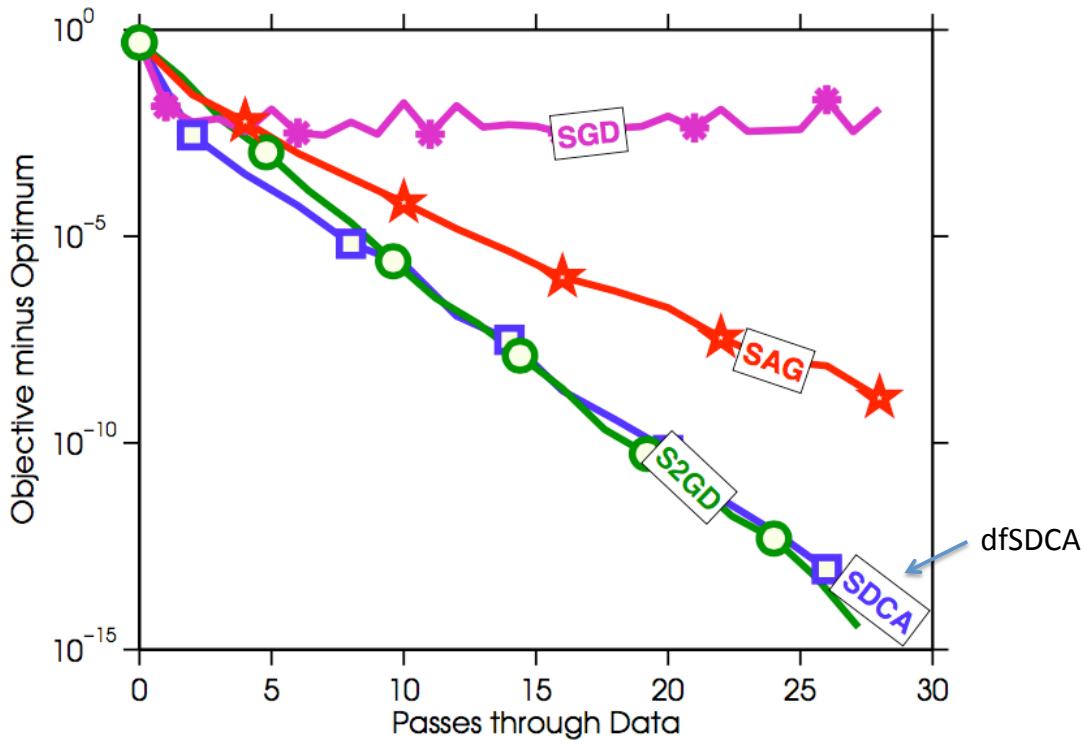
Rie Johnson and Tong Zhang. **Accelerating stochastic gradient descent using predictive variance reduction.** *NIPS*, 2013.

S2GD: Semi-Stochastic Gradient Descent

J. Konečný and P. R. **Semi-stochastic gradient descent methods.** *arXiv:1312.1666*, 2013

Modern Methods for ERM vs SGD

Dataset: rcv1 (n = 20,241 ; d = 47,232)



Behavior of dfSDCA for various λ

Dataset: rcv1 (n = 20,241 ; d = 47,232)