Randomized Coordinate Descent with Arbitrary Sampling: Algorithms and Complexity

Zheng Qu

University of Hong Kong

CAM, 23-26 Aug 2016
Hong Kong

based on joint work with Peter Richtarik and Dominique Cisba(University of Edinburgh)

Outline

- First-order methods for composite convex optimization
- Randomized coordinate descent method
- Adaptive sampling
- Expected separable overapproximation

Problem and Motivation

Problem Setup

$$\min_{x \in \mathbb{R}^n} [F(x) := f(x) + \psi(x)]$$

- $f : \mathbb{R}^n \to \mathbb{R}$ is convex and smooth:

$$f(x+h) \leq f(x) + \langle \nabla f(x), h \rangle + \frac{1}{2}\|Ah\|^2, \ \ \forall x, h \in \mathbb{R}^n$$

- $\psi : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is proper, convex, closed and separable:

$$\psi(x) \equiv \sum_{i=1}^{n} \psi^i(x^i)$$

Motivation: Empirical Risk Minimization

ERM:

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

- supervised learning/image processing...;
- train a linear predictor $w \in \mathbb{R}^d$;
- n training samples $A_1, \ldots, A_n \in \mathbb{R}^d$;

Motivation: Empirical Risk Minimization

ERM:

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

- supervised learning/image processing...;
- train a linear predictor $w \in \mathbb{R}^d$;
- n training samples $A_1, \ldots, A_n \in \mathbb{R}^d$;
- convex loss function $\phi_i : \mathbb{R} \to \mathbb{R}$;
    - ex.: Squared loss ($\phi_i(a) = \frac{1}{2}(a - b_i)^2$), Logistic loss ($\phi_i(a) = \log(1 + e^a)$), ...

Motivation: Empirical Risk Minimization

ERM:

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

- supervised learning/image processing...;
- train a linear predictor $w \in \mathbb{R}^d$;
- n training samples $A_1, \ldots, A_n \in \mathbb{R}^d$;
- convex loss function $\phi_i : \mathbb{R} \to \mathbb{R}$;
  - ex.: Squared loss ($\phi_i(a) = \frac{1}{2}(a - b_i)^2$), Logistic loss ($\phi_i(a) = \log(1 + e^a)$), ...
- convex regularizer $g : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$;
  - ex.: L_1 regularization ($g(w) = \|w\|_1$), L_2 regularization ($g(w) = \frac{1}{2}\|w\|_2^2$), ...

Primal Dual Formulation

- ERM:

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

- Dual problem of ERM:

$$\max_{\alpha\in\mathbb{R}^n} \quad D(\alpha) \stackrel{\text{def}}{=} \underbrace{-\lambda g^\star\left(\frac{1}{\lambda n}\sum_{i=1}^{n} A_i\alpha_i\right)}_{\substack{smooth \ \ if \\ g \ \ strongly \ \ convex}} \underbrace{-\frac{1}{n}\sum_{i=1}^{n}\phi_i^\star(-\alpha_i)}_{\substack{convex \\ and \ \ separable}}$$

Primal Dual Formulation

- ERM:

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

- Dual problem of ERM:

$$\max_{\alpha\in\mathbb{R}^n} \quad D(\alpha) \stackrel{\text{def}}{=} \underbrace{-\lambda g^\star\left(\frac{1}{\lambda n}\sum_{i=1}^{n} A_i\alpha_i\right)}_{\substack{\text{smooth if} \\ g \text{ strongly convex}}} \underbrace{-\frac{1}{n}\sum_{i=1}^{n}\phi_i^\star(-\alpha_i)}_{\substack{\text{convex} \\ \text{and separable}}}$$

- Optimality conditions:

$$\mathbf{OPT1}: \; w^* = \nabla g^\star\left(\frac{1}{\lambda n} A\alpha^*\right)$$

$$\mathbf{OPT2}: \; \alpha_i^* = -\nabla\phi_i\left(A_i^\top w^*\right), \quad \forall i = 1,\ldots,n.$$

First-order methods for non-strongly convex composite optimization

Problem Setup

$$\min_{x \in \mathbb{R}^n} [F(x) := f(x) + \psi(x)]$$

- $f : \mathbb{R}^n \to \mathbb{R}$ is convex and smooth:

$$f(x+h) \leq f(x) + \langle \nabla f(x), h \rangle + \frac{1}{2}\|Ah\|^2, \ \ \forall x, h \in \mathbb{R}^n$$

- $\psi : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is proper, convex, closed and separable:

$$\psi(x) \equiv \sum_{i=1}^{n} \psi^i(x_i)$$

Proximal Gradient

Parameters: vector $v \in \mathbb{R}^n_{++}$
Initialization: choose $x_0 \in \mathrm{dom}\,\psi$
for $k \geq 0$ **do**
  for $i \in [n]$ **do**
    $x^i_{k+1} = \arg\min_{x \in \mathbb{R}} \left\{ \langle \nabla_i f(x_k), x \rangle + \frac{v_i}{2} \|x - x^i_k\|^2_i + \psi^i(x) \right\}$
  end for
end for

Proximal Gradient

Parameters: vector $v \in \mathbb{R}^n_{++}$
Initialization: choose $x_0 \in \operatorname{dom}\psi$
for $k \geq 0$ **do**
  for $i \in [n]$ **do**
    $x^i_{k+1} = \arg\min_{x\in\mathbb{R}} \left\{ \langle \nabla_i f(x_k), x\rangle + \frac{v_i}{2}\|x - x^i_k\|^2_i + \psi^i(x) \right\}$
  end for
end for

- proximal operator of ψ^i is easily computable.

Proximal Gradient

Parameters: vector $v \in \mathbb{R}^n_{++}$
Initialization: choose $x_0 \in \operatorname{dom} \psi$
for $k \geq 0$ **do**
  for $i \in [n]$ **do**
    $x^i_{k+1} = \arg\min_{x \in \mathbb{R}} \left\{ \langle \nabla_i f(x_k), x \rangle + \frac{v_i}{2} \|x - x^i_k\|^2_i + \psi^i(x) \right\}$
  end for
end for

- proximal operator of ψ^i is easily computable.
- a.k.a. explicite-implicite/forward-backward method $\subset$ splitting algorithm [Lions & Mercier 79], [Eckstein & Bertsekas 89]

Accelerated Proximal Gradient

Parameters: vector $v \in \mathbb{R}^n_{++}$

Initialization: choose $x_0 \in \mathrm{dom}(\psi)$, set $z_0 = x_0$ and $\theta_0 = 1$

for $k \geq 0$ **do**

  for $i \in [n]$ **do**

    $z^i_{k+1} = \arg\min_{z \in \mathbb{R}} \left\{ \langle \nabla_i f((1-\theta_k)x_k + \theta_k z_k), z \rangle + \frac{\theta_k v_i}{2} \|z - z^i_k\|^2_i + \psi^i(z) \right\}$

  end for

  $x_{k+1} = (1-\theta_k)x_k + \theta_k z_{k+1}$

  $\theta_{k+1} = \frac{\sqrt{\theta_k^4 + 4\theta_k^2} - \theta_k^2}{2}$

end for

[Nesterov 83, 04], [Beck & Teboulle 08](FISTA), [Tseng 08],

Accelerated Proximal Gradient

Parameters: vector $v \in \mathbb{R}^n_{++}$
Initialization: choose $x_0 \in \mathrm{dom}(\psi)$, set $z_0 = x_0$ and $\theta_0 = 1$
for $k \geq 0$ **do**
  for $i \in [n]$ **do**
    $z^i_{k+1} = \arg\min_{z \in \mathbb{R}} \left\{ \langle \nabla_i f((1-\theta_k)x_k + \theta_k z_k), z \rangle + \frac{\theta_k v_i}{2} \|z - z^i_k\|^2_i + \psi^i(z) \right\}$
  end for
  $x_{k+1} = (1-\theta_k)x_k + \theta_k z_{k+1}$
  $\theta_{k+1} = \frac{\sqrt{\theta_k^4 + 4\theta_k^2} - \theta_k^2}{2}$
end for

[Nesterov 83, 04], [Beck & Teboulle 08](FISTA), [Tseng 08], [Su, Boyd & Candès 14], [Chambolle & Pock 15], [Chambolle & Dossal 15], [Attouch & Peypouquet 15]

Convergence Analysis

Theorem

If $v_i = L$ for any $i \in [n]$ with $L \geq \lambda_{\max}(A^\top A)$, then the iterates $\{x_k\}$ of the proximal gradient method satsify:

$$F(x_k) - F(x_*) \leq \frac{L\|x_0 - x_*\|^2}{2k}, \ \ \forall k \geq 1.$$

Theorem (Tseng 08)

If $v_i = L$ for any $i \in [n]$ with $L \geq \lambda_{\max}(A^\top A)$, then the iterates $\{x_k\}_k$ of the accelerated proximal gradient algorithm satisfy:

$$F(x_k) - F(x_*) \leq \frac{2L\|x_0 - x_*\|^2}{(k+1)^2}, \ \ \forall k \geq 1.$$

Randomized Coordinate Descent

Randomized coordinate descent

Parameters: vector $v \in \mathbb{R}^n_{++}$
Initialization: choose $x_0 \in \operatorname{dom} \psi$
for $k \geq 0$ **do**
    Generate random $i \in [n]$ uniformly
    $x_{k+1} \leftarrow x_k$
    $x^i_{k+1} = \arg\min_{x \in \mathbb{R}} \left\{ \langle \nabla_i f(x_k), x \rangle + \frac{v_i}{2} \| x - x^i_k \|^2_i + \psi^i(x) \right\}$
end for

Randomized Coordinate Descent

Randomized coordinate descent

Parameters: vector $v \in \mathbb{R}^n_{++}$
Initialization: choose $x_0 \in \operatorname{dom} \psi$
for $k \geq 0$ **do**
    Generate random $i \in [n]$ uniformly
    $x_{k+1} \leftarrow x_k$
    $x^i_{k+1} = \arg\min_{x \in \mathbb{R}} \left\{ \langle \nabla_i f(x_k), x \rangle + \frac{v_i}{2} \|x - x^i_k\|^2_i + \psi^i(x) \right\}$
end for

- $v = \mathrm{Diag}(A^\top A)$ [Nesterov 10], [Shalev-Shwartz & Tewari 11], [Richtarik & Takac 11]

Randomized Coordinate Descent

Randomized coordinate descent

1: **Parameters:** vector $v \in \mathbb{R}^n_{++}$
2: **Initialization:** choose $x_0 \in \operatorname{dom}\psi$
3: **for** $k \geq 0$ **do**
4: Generate random $i \in [n]$ uniformly
5: $x_{k+1} \leftarrow x_k$
6: $x^i_{k+1} = \arg\min_{x \in \mathbb{R}} \left\{ \langle \nabla_i f(x_k), x \rangle + \frac{v_i}{2} \|x - x^i_k\|^2_i + \psi^i(x) \right\}$
7: **end for**

- $v = \mathrm{Diag}(A^\top A)$ [Nesterov 10], [Shalev-Shwartz & Tewari 11], [Richtarik & Takac 11]
- Other variants [Wright 15]
  - Cyclic (Gauss-Seidel) [Canutescu & Dunbrack 03]
  - Greedy [Wu & Lange 08] [Nutini et. al 15]

Parallel Randomized Coordinate Descent

Parallel coordinate descent

Parameters: $\tau \in [n]$, vector $v \in \mathbb{R}^n_{++}$
Initialization: choose $x_0 \in \operatorname{dom} \psi$
for $k \geq 0$ **do**
    Generate a random subset $S_k \subset [n]$ of size τ uniformly
    $x_{k+1} \leftarrow x_k$
    for $i \in S_k$ **do**
        $x^i_{k+1} = \arg\min_{x \in \mathbb{R}} \left\{ \langle \nabla_i f(x_k), x \rangle + \frac{v_i}{2} \|x - x^i_k\|^2_i + \psi^i(x) \right\}$
    end for
end for

Parallel Randomized Coordinate Descent

Parallel coordinate descent

Parameters: $\tau \in [n]$, vector $v \in \mathbb{R}^n_{++}$
Initialization: choose $x_0 \in \operatorname{dom}\psi$
for $k \geq 0$ **do**
  Generate a random subset $S_k \subset [n]$ of size τ uniformly
  $x_{k+1} \leftarrow x_k$
  for $i \in S_k$ **do**
    $x^i_{k+1} = \arg\min_{x \in \mathbb{R}} \left\{ \langle \nabla_i f(x_k), x \rangle + \frac{v_i}{2} \|x - x^i_k\|^2_i + \psi^i(x) \right\}$
  end for
end for

$v = \left(1 + \frac{(\tau-1)(\omega-1)}{\max(n-1,1)}\right) \operatorname{Diag}(A^\top A)$ [Richtarik & Takac 13]

where ω is the maximal number of nonzero elements in each row of A.

Convergence Analysis

Theorem (Richtarik & Takac 13)

Define the level-set distance

$$\mathcal{R}_v(x_0, x_*) \stackrel{def}{=} \max_x \{\|x - x_*\|_v^2 : F(x) \leq F(x_0)\}.$$

Under the assumption

$$\mathcal{R}_v(x_0, x_*) < +\infty,$$

we have:

$$\mathbb{E}[F(x_k)] - F(x_*) \leq \frac{2n \max\{\mathcal{R}_v(x_0, x_*), F(x_0) - F(x_*)\}}{2n \max\{\mathcal{R}_v(x_0, x_*)/\left(F(x_k - F(x_*)\right), 1\} + \tau k}$$

Accelerated Parallel Proximal Coordinate Descent

Parameters: $\tau \in [n]$, vector $v \in \mathbb{R}^n_{++}$
Initialization: choose $x_0 \in \mathrm{dom}(\psi)$, set $z_0 = x_0$ and $\theta_0 = \tau/n$
for $k \geq 0$ **do**
    $y_k = (1-\theta_k)x_k + \theta_k z_k$
    Generate a random subset $S_k \subset [n]$ of size τ uniformly
    $z_{k+1} \leftarrow z_k$
    for $i \in S_k$ **do**
        $z^i_{k+1} = \arg\min_{z\in\mathbb{R}} \left\{ \langle \nabla_i f(y_k), z\rangle + \frac{\theta_k v_i n}{2\tau} \|z - z^i_k\|^2_i + \psi^i(z) \right\}$
    end for
    $x_{k+1} = y_k + \theta_k n/\tau \cdot (z_{k+1} - z_k)$
    $\theta_{k+1} = \frac{\sqrt{\theta_k^4 + 4\theta_k^2} - \theta_k^2}{2}$
end for

[Nesterov 10], [Lee & Sidford 13], [Fercoq & Richtarik 13]...

Convergence Analysis

Theorem (Fercoq & Richtarik 13)

Choose

$$v_i = \sum_{j=1}^{m} \left(1 + \frac{(\tau - 1)(\omega_j - 1)}{\max(n-1, 1)}\right) A_{ji}^2, \quad i = 1, 2, \ldots, n.$$

The iterates $\{x_k\}$ *of APPROX for all* $k \geq 1$ *satisfies:*

$$\mathbb{E}[F(x_k) - F(x_*)] \leq \frac{4\left[\left(1 - \frac{\tau}{n}\right)(F(x_0) - F(x_*)) + \frac{1}{2}\|x_0 - x_*\|_v^2\right]}{((k-1)\tau/n + 2)^2}.$$

Summary

Parallel Coordinate Descent (choose subset of size τ uniformly)

Summary

Parallel Coordinate Descent
(choose subset of size τ uniformly)

$\tau = 1$

Randomized Coordinate Descent

Summary

Parallel Coordinate Descent (choose subset of size τ uniformly)

$\tau = 1$

Randomized Coordinate Descent

$\tau = n$

Proximal Gradient

Summary

Parallel Coordinate Descent (choose subset of size τ uniformly)

$\tau = 1$

Randomized Coordinate Descent

$\tau = n$

Proximal Gradient

Accelerated Parallel Proximal Coordinate Descent (choose subset of size τ uniformly)

Summary

Parallel Coordinate Descent (choose subset of size τ uniformly)

$\tau = 1$

Randomized Coordinate Descent

$\tau = n$

Proximal Gradient

Accelerated Parallel Proximal Coordinate Descent (choose subset of size τ uniformly)

$\tau = n$

Accelerated Proximal Gradient

Summary

Parallel Coordinate Descent
(choose subset of size τ uniformly)

$\tau = 1$

Randomized Coordinate Descent

$\tau = n$

Proximal Gradient

Accelerated Parallel Proximal Coordinate Descent
(choose subset of size τ uniformly)

$\tau = n$

Accelerated Proximal Gradient

Randomized coordinate descent method with arbitrary sampling

Q. and Richtarik. Coordinate descent with arbitrary sampling I: algorithms and complexity, *Optimization methods and software, 2016*.

Sampling

- Sampling is a set-valued random variable:

$$\hat{S} \subset \{1, \dots, n\}$$

- Probability vector:

$$p_i = \mathbb{P}(i \in \hat{S}), \quad i \in \{1, \dots, n\}$$

- Proper sampling:

$$p_i = \mathbb{P}(i \in \hat{S}) > 0, \quad \forall i \in \{1, \dots, n\}$$

- Serial sampling:

$$\mathbb{P}(|\hat{S}| = 1) = 1$$

- Uniform sampling:

$$p_1 = \cdots = p_n = \frac{\mathbb{E}[|\hat{S}|]}{n}$$

Algorithm

Parameters: proper sampling $\hat{S}$ with probability vector $p = (p_1, \dots, p_n) \in [0,1]^n$, $v \in \mathbb{R}^n_{++}$, sequence $\{\theta_k\}_{k \geq 0} \subset (0,1]$
Initialization: choose $x_0 \in \operatorname{dom} \psi$ and set $z_0 = x_0$
for $k \geq 0$ **do**
  $y_k = (1 - \theta_k) x_k + \theta_k z_k$
  Generate a random set of blocks $S_k \sim \hat{S}$
  $z_{k+1} \leftarrow z_k$
  for $i \in S_k$ **do**
    $$z^i_{k+1} = \arg\min_{z \in \mathbb{R}} \left\{ \langle \nabla_i f(y_k), z \rangle + \frac{\theta_k v_i}{2 p_i} \| z - z^i_k \|_i^2 + \psi^i(z) \right\}$$
  end for
  $x_{k+1} = y_k + \theta_k p^{-1} \cdot (z_{k+1} - z_k)$
end for

Efficient Implementation

Parameters: proper sampling $\hat{S}$ with probability vector $p = (p_1, \ldots, p_n)$, $v \in \mathbb{R}^n_{++}$, sequence $\{\theta_k\}_{k \geq 0}$
Initialization: choose $x^0 \in \mathrm{dom}\, \psi$, set $z^0 = x^0$, $u^0 = 0$ and $\alpha_0 = 1$
for $k \geq 0$ **do**
  Generate a random set of coordinates $S_k \sim \hat{S}$
  $z^{k+1} \leftarrow z^k$, $u^{k+1} \leftarrow u^k$
  for $i \in S_k$ **do**
    $\Delta z_i^k = \arg\min_{t \in \mathbb{R}} \left\{ t \nabla_i f(\alpha_k u^k + z^k) + \frac{\theta_k v_i}{2 p_i} |t|^2 + \psi_i(z_i^k + t) \right\}$
    $z_i^{k+1} \leftarrow z_i^k + \Delta z_i^k$
    $u_i^{k+1} \leftarrow u_i^k - \alpha_k^{-1}(1 - \theta_k p_i^{-1}) \Delta z_i^k$
    $\alpha_{k+1} = (1 - \theta_{k+1}) \alpha_k$
  end for
end for
OUTPUT: $x^{k+1} = z^k + \alpha_k u^k + \theta_k p^{-1}(z^{k+1} - z^k)$

Convergence Analysis

Lemma

Let $\hat{S}$ be an arbitrary proper sampling and $v \in \mathbb{R}^n_{++}$ be such that

$$\mathbb{E}[f(x + h_{[\hat{S}]})] \leq f(x) + \langle \nabla f(x), h \rangle_p + \frac{1}{2}\|h\|^2_{v \circ p}, \quad \forall x, h \in \mathbb{R}^n.$$

Let $\{\theta_k\}_{k \geq 0}$ be arbitrary sequence of positive numbers in $(0, 1]$. Then for the sequence of iterates produced by the algorithm and all $k \geq 0$, the following recursion holds:

$$\begin{aligned}\mathbb{E}_k &\left[\hat{F}_{k+1} + \frac{\theta_k^2}{2}\|z^{k+1} - x^*\|^2_{v \circ p^{-2}}\right] \\ &\leq \left[\hat{F}_k + \frac{\theta_k^2}{2}\|z^k - x^*\|^2_{v \circ p^{-2}}\right] - \theta_k(\hat{F}_k - F^*) \ .\end{aligned}$$

Convergence Analysis

Lemma

Let $\hat{S}$ be an arbitrary proper sampling and $v \in \mathbb{R}^n_{++}$ be such that

$$\mathbb{E}[f(x + h_{[\hat{S}]})] \leq f(x) + \langle \nabla f(x), h \rangle_p + \frac{1}{2}\|h\|^2_{v \circ p}, \quad \forall x, h \in \mathbb{R}^n.$$

Let $\{\theta_k\}_{k \geq 0}$ be arbitrary sequence of positive numbers in $(0, 1]$. Then for the sequence of iterates produced by the algorithm and all $k \geq 0$, the following recursion holds:

$$\mathbb{E}_k\left[\hat{F}_{k+1} + \frac{\theta_k^2}{2}\|z^{k+1} - x^*\|^2_{v \circ p^{-2}}\right] \leq \left[\hat{F}_k + \frac{\theta_k^2}{2}\|z^k - x^*\|^2_{v \circ p^{-2}}\right] - \theta_k(\hat{F}_k - F^*) \ .$$

$\hat{F}_k \geq F(x_k)$ if $\psi \equiv 0$ or $\theta_k \leq \min p_i$

Convergence Results

$$(f, \hat{S}) \sim ESO(v) + \begin{cases} \psi \equiv 0 & \text{or} \\ \theta_k \leq \min p_i \end{cases}$$

- $\theta_k = \theta_0$

$$\mathbb{E}\left[F\left(\frac{x^k + \theta_0 \sum_{t=1}^{k-1} x^t}{1 + (k-1)\theta_0}\right)\right] - F^* \leq \frac{C}{(k-1)\theta_0 + 1}$$

- $\theta_{k+1} = \frac{\sqrt{\theta_k^4 + 4\theta_k^2} - \theta_k^2}{2}$

$$\mathbb{E}[F(x^k)] - F^* \leq \frac{4C}{((k-1)\theta_0 + 2)^2}$$

where

$$C = (1 - \theta_0)(F(x^0) - F^*) + \frac{\theta_0^2}{2}\|x^0 - x^*\|^2_{v \circ p^{-2}}$$

Corollaries-Parallel Coordinate Descent

Corollary

The iterates $\{x_k\}$ of Parallel Coordinate Descent satisfy:

$$\mathbb{E}[F(x_k)] - F(x_*) \leq \frac{n}{(k-1)\tau + n}\left[\left(1 - \frac{\tau}{n}\right)(F(x_0) - F(x_*)) + \frac{1}{2}\|x_0 - x_*\|_v^2\right]$$

Compare with

- [Richtarik & Takac 13]:

$$\max_x\{\|x - x_*\|_v^2 : F(x) \leq F(x_0)\} < +\infty$$

- [Lu & Xiao 14] ($\tau = 1$):

$$\mathbb{E}[F(x_k)] - F(x_*) \leq \frac{n}{n+k}\left[(F(x_0) - F(x_*)) + \frac{1}{2}\|x_0 - x_*\|_v^2\right]$$

Corollaries-Smooth Minimization

Corollary

If $\psi \equiv 0$, then the iterates $\{x_k\}$ of accelerated coordinate descent satisfy:

$$\mathbb{E}\left[f(x^k)\right] - f^* \le \frac{2\|x^0 - x^*\|^2_{v \circ p^{-2}}}{(k+1)^2}, \quad k \ge 1.$$

Corollaries-Smooth Minimization

Corollary

If $\psi \equiv 0$, then the iterates $\{x_k\}$ of accelerated coordinate descent satisfy:

$$\mathbb{E}\left[f(x^k)\right] - f^* \leq \frac{2\|x^0 - x^*\|^2_{v \circ p^{-2}}}{(k+1)^2}, \quad k \geq 1.$$

Define $L_i = A_i^\top A_i$ for $i = 1, \ldots, n$.

Corollary

If each step we update coordinate i with probability

$$p_i \sim \sqrt{L_i},$$

then $\mathbb{E}\left[f(x^k)\right] - f^* \leq \dfrac{2(\sum_i \sqrt{L_i})^2 \|x^0 - x^*\|^2}{(k+1)^2}, \quad k \geq 1.$

Corollaries-Smooth Minimization

Serial sampling $\hat{S}$, $v = L$:

$$\mathbb{E}\left[f(x^k)\right] - f^* \leq \frac{2\|x^0 - x^*\|^2_{L\circ p^{-2}}}{(k+1)^2}, \quad k \geq 1.$$

Corollaries-Smooth Minimization

Serial sampling $\hat{S}$, $v = L$:

$$\mathbb{E}\left[f(x^k)\right] - f^* \leq \frac{2\|x^0 - x^*\|^2_{L \circ p^{-2}}}{(k+1)^2}, \;\; k \geq 1.$$

The probability minimizing the right-hand side is:

$$p_i^* = \frac{(L_i \|x_i^* - x_i^0\|^2)^{\frac{1}{3}}}{\sum_{j=1}^{n} (L_j \|x_j^* - x_j^0\|^2)^{\frac{1}{3}}}, \;\; i = 1, \ldots, n.$$

Stochastic dual coordinate ascent with adaptive sampling

Cisba, Q. and Richtarik. Stochastic dual coordinate ascent with adaptive sampling, *International Conference on Machine Learning, 2015*.

Primal Dual Formulation

- ERM:

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

- Dual problem of ERM:

$$\max_{\alpha \in \mathbb{R}^n} \quad D(\alpha) \stackrel{\text{def}}{=} \underbrace{-\lambda g^\star \left(\frac{1}{\lambda n} \sum_{i=1}^{n} A_i \alpha_i \right)}_{smooth} - \underbrace{\frac{1}{n} \sum_{i=1}^{n} \phi_i^\star(-\alpha_i)}_{\substack{\gamma-strongly\ \ convex \\ and\ \ separable}}$$

Primal Dual Formulation

- ERM:

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

- Dual problem of ERM:

$$\max_{\alpha\in\mathbb{R}^n}\ D(\alpha)\stackrel{\text{def}}{=}\underbrace{-\lambda g^\star\left(\frac{1}{\lambda n}\sum_{i=1}^{n}A_i\alpha_i\right)}_{smooth}\underbrace{-\frac{1}{n}\sum_{i=1}^{n}\phi_i^\star(-\alpha_i)}_{\substack{\gamma-strongly\ \ convex\\ and\ \ separable}}$$

- Optimality conditions:

$$\textbf{OPT1}:\ w^*=\nabla g^\star\left(\frac{1}{\lambda n}A\alpha^*\right)$$

$$\textbf{OPT2}:\ \alpha_i^*=-\nabla\phi_i\left(A_i^\top w^*\right),\ \ \forall i=1,\ldots,n.$$

Stochastic Dual Coordinate Ascent

Primal solution

For $t \geq 0$:

1. $w^t = \nabla g^\star(\frac{1}{\lambda n} A\alpha^t)$

Dual solution

For $t \geq 0$:

1. $\alpha^{t+1} = \alpha^t$;
2. **Randomly pick** $i_t \in \{1, \ldots, n\}$**;**
3. Update $\alpha_{i_t}^{t+1}$:

$$\alpha_{i_t}^{t+1} = \arg\max_{\beta \in \mathbb{R}} \left\{ -\phi_{i_t}^\star(-\beta) - (A_{i_t}^\top w^t)\beta - \frac{\|A_{i_t}\|^2}{2\lambda n} |\beta - \alpha_{i_t}^t|^2 \right\}$$

Stochastic Dual Coordinate Ascent

Primal solution

For $t \geq 0$:

1. $w^t = \nabla g^\star(\frac{1}{\lambda n} A\alpha^t)$

Dual solution

For $t \geq 0$:

1. $\alpha^{t+1} = \alpha^t$;
2. **Randomly pick $i_t \in \{1, \ldots, n\}$ according to a fixed distribution p;**
3. Update $\alpha_{i_t}^{t+1}$:
$$\alpha_{i_t}^{t+1} = \arg\max_{\beta \in \mathbb{R}} \left\{ -\phi_{i_t}^\star(-\beta) - (A_{i_t}^\top w^t)\beta - \frac{\|A_{i_t}\|^2}{2\lambda n} |\beta - \alpha_{i_t}^t|^2 \right\}$$

Uniform and Importance Sampling

Uniform sampling (SDCA: [Shalev-Shwartz & Zhang 13],...)

$$p_i = \mathbf{Prob}(i_t = i) \sim \frac{1}{n},$$

Iteration complexity:

$$\tilde{O}\left(n + \frac{\max_i \|A_i\|^2}{\lambda\gamma}\right)$$

Uniform and Importance Sampling

Uniform sampling (SDCA: [Shalev-Shwartz & Zhang 13],...)

$$p_i = \mathbf{Prob}(i_t = i) \sim \frac{1}{n},$$

Iteration complexity:

$$\tilde{O}\left(n + \frac{\max_i \|A_i\|^2}{\lambda\gamma}\right)$$

Importance sampling (Iprox-SDCA: [Zhao & Zhang 15'],...)

$$p_i = \mathbf{Prob}(i_t = i) \sim \|A_i\|^2 + \lambda\gamma n,$$

Iteration complexity:

$$\tilde{O}\left(n + \frac{\frac{1}{n}\sum_{i=1}^n \|A_i\|^2}{\lambda\gamma}\right)$$

Adaptive Sampling

- Each dual variable has a natural measure of progress:

$$\kappa_i^t \stackrel{\text{def}}{=} \alpha_i^t + \nabla\phi_i(A_i^\top w^t), \quad i = 1, \ldots, n$$

called dual residue.

Adaptive Sampling

- Each dual variable has a natural measure of progress:

$$\kappa_i^t \stackrel{\text{def}}{=} \alpha_i^t + \nabla\phi_i(A_i^\top w^t), \quad i = 1, \ldots, n$$

  called dual residue.

- Optimality conditions:

$$\textbf{OPT1}: \; w^* = \nabla g^\star\left(\frac{1}{\lambda n} A\alpha^*\right)$$

$$\textbf{OPT2}: \; \alpha_i^* = -\nabla\phi_i\left(A_i^\top w^*\right), \; \forall i \in [n].$$

Adaptive Sampling

- Each dual variable has a natural measure of progress:

$$\kappa_i^t \stackrel{\text{def}}{=} \alpha_i^t + \nabla\phi_i(A_i^\top w^t), \quad i = 1, \dots, n$$

  called dual residue.

- Optimality conditions:

$$\begin{aligned} \textbf{OPT1}: \; w^* &= \nabla g^\star\left(\frac{1}{\lambda n} A\alpha^*\right) \\ \textbf{OPT2}: \; \alpha_i^* &= -\nabla\phi_i\left(A_i^\top w^*\right), \;\; \forall i \in [n]. \end{aligned}$$

- A sampling distribution p is coherent with κ^t if for all $i \in [n]$:

$$\kappa_i^t \neq 0 \;\; \Rightarrow p_i > 0.$$

Stochastic Dual Coordinate Ascent

Primal solution

For $t \geq 0$:

1. $w^t = \nabla g^\star(\frac{1}{\lambda n} A\alpha^t)$

Dual solution

For $t \geq 0$:

1. $\alpha^{t+1} = \alpha^t$;
2. **Randomly pick $i_t \in \{1, \ldots, n\}$ according to a fixed distribution p;**
3. Update $\alpha^{t+1}_{i_t}$:
$$\alpha^{t+1}_{i_t} = \arg\max_{\beta \in \mathbb{R}} \left\{ -\phi^\star_{i_t}(-\beta) - (A_{i_t}^\top w^t)\beta - \frac{\|A_{i_t}\|^2}{2\lambda n}|\beta - \alpha^t_{i_t}|^2 \right\}$$

Adaptive Stochastic Dual Coordinate Ascent

Primal solution

For $t \geq 0$:

1. $w^t = \nabla g^\star(\frac{1}{\lambda n} A\alpha^t)$

Dual solution

For $t \geq 0$:

1. $\alpha^{t+1} = \alpha^t$;
2. **Randomly pick $i_t \in \{1, \ldots, n\}$ according to a distribution p^t coherent with dual residue κ^t;**
3. Update $\alpha_{i_t}^{t+1}$:
$$\alpha_{i_t}^{t+1} = \arg\max_{\beta \in \mathbb{R}} \left\{ -\phi_{i_t}^\star(-\beta) - (A_{i_t}^\top w^t)\beta - \frac{\|A_{i_t}\|^2}{2\lambda n} |\beta - \alpha_{i_t}^t|^2 \right\}$$

Convergence Theorem

Theorem (AdaSDCA)

Consider AdaSDCA. If at each iteration $t \geq 0$,

$$\theta(\kappa^t, p^t) \overset{def}{=} \frac{n\lambda\gamma \sum_i |\kappa_i^t|^2}{\sum_{i:\kappa_i^t \neq 0} (p_i^t)^{-1}(\|A_i\|^2 + n\lambda\gamma)|\kappa_i^t|^2} \leq \min_{i:\kappa_i^t \neq 0} p_i^t,$$

then

$$\mathbb{E}[P(w^t) - D(\alpha^t)] \leq \frac{1}{\tilde{\theta}_t} \prod_{k=0}^{t} (1 - \tilde{\theta}_k) \left(D(\alpha^*) - D(\alpha^0)\right),$$

for all $t \geq 0$ where

$$\tilde{\theta}_t \overset{def}{=} \frac{\mathbb{E}[\theta(\kappa^t, p^t)(P(w^t) - D(\alpha^t))]}{\mathbb{E}[P(w^t) - D(\alpha^t)]}.$$

Optimal Adaptive Sampling Probability

$$
\begin{aligned}
p^*(\kappa^t) = \quad \operatorname{arg\,max} \quad & \theta(\kappa^t, p) \\
\text{s.t.} \quad & p \in \mathbb{R}^n_+, \quad \textstyle\sum_i p_i = 1 \\
& p \text{ is coherent with } \kappa^t \\
& \theta(\kappa^t, p) \le \min_{i:\kappa_i^t \neq 0} p_i
\end{aligned}
$$

Optimal Adaptive Sampling Probability

$$
\begin{aligned}
p^*(\kappa^t) = \quad \arg\max \quad & \theta(\kappa^t, p) \\
\text{s.t.} \quad & p \in \mathbb{R}^n_+, \ \ \textstyle\sum_i p_i = 1 \\
& p \text{ is coherent with } \kappa^t \\
& \theta(\kappa^t, p) \leq \min_{i:\kappa_i^t \neq 0} p_i
\end{aligned}
$$

Relaxation:

$$
\begin{aligned}
\tilde{p}^*(\kappa^t) = \quad \arg\max \quad & \theta(\kappa^t, p) \\
\text{s.t.} \quad & p \in \mathbb{R}^n_+, \ \ \textstyle\sum_{i=1}^n p_i = 1
\end{aligned}
$$

Optimal Adaptive Sampling Probability

$$
\begin{aligned}
p^*(\kappa^t) = \quad \arg\max \quad & \theta(\kappa^t, p) \\
\text{s.t.} \quad & p \in \mathbb{R}^n_+, \ \ \textstyle\sum_i p_i = 1 \\
& p \text{ is coherent with } \kappa^t \\
& \theta(\kappa^t, p) \le \min_{i:\kappa_i^t \neq 0} p_i
\end{aligned}
$$

Relaxation:

$$
\begin{aligned}
\tilde{p}^*(\kappa^t) = \quad \arg\max \quad & \frac{n\lambda\gamma \sum_i |\kappa_i^t|^2}{\sum_{i:\kappa_i^t \neq 0} (p_i)^{-1} |\kappa_i^t|^2 (\|A_i\|^2 + n\lambda\gamma)} \\
\text{s.t.} \quad & p \in \mathbb{R}^n_+, \ \ \textstyle\sum_{i=1}^n p_i = 1
\end{aligned}
$$

Optimal Adaptive Sampling Probability

$$\begin{aligned} p^*(\kappa^t) = \quad & \arg\max \quad \theta(\kappa^t, p) \\ & \text{s.t.} \qquad p \in \mathbb{R}^n_+, \ \ \textstyle\sum_i p_i = 1 \\ & \qquad\qquad p \text{ is coherent with } \kappa^t \\ & \qquad\qquad \theta(\kappa^t, p) \le \min_{i:\kappa_i^t \neq 0} p_i \end{aligned}$$

Relaxation:

$$\begin{aligned} \tilde{p}^*(\kappa^t) = \quad & \arg\max \quad \frac{n\lambda\gamma \sum_i |\kappa_i^t|^2}{\sum_{i:\kappa_i^t \neq 0} (p_i)^{-1} |\kappa_i^t|^2 (\|A_i\|^2 + n\lambda\gamma)} \\ & \text{s.t.} \qquad p \in \mathbb{R}^n_+, \ \ \textstyle\sum_{i=1}^n p_i = 1 \end{aligned}$$

$$(\tilde{p}^*(\kappa^t))_i \sim |\kappa_i^t| \sqrt{\|A_i\|^2 + n\lambda\gamma}, \ \ \forall i \in [n].$$

Exact Relaxation for Squared Loss

Theorem (AdaSDCA)

Consider AdaSDCA. If at each iteration $t \geq 0$,

$$\theta(\kappa^t, p^t) \stackrel{def}{=} \frac{n\lambda\gamma \sum_i |\kappa_i^t|^2}{\sum_{i:\kappa_i^t \neq 0} (p_i^t)^{-1} |\kappa_i^t|^2 (\|A_i\|^2 + n\lambda\gamma)} \leq \min_{i:\kappa_i^t \neq 0} p_i^t,$$

then

$$\mathbb{E}[P(w^t) - D(\alpha^t)] \leq \frac{1}{\tilde{\theta}_t} \prod_{k=0}^{t} (1 - \tilde{\theta}_k) \left(D(\alpha^*) - D(\alpha^0) \right),$$

for all $t \geq 0$ where

$$\tilde{\theta}_t \stackrel{def}{=} \frac{\mathbb{E}[\theta(\kappa^t, p^t)(P(w^t) - D(\alpha^t))]}{\mathbb{E}[P(w^t) - D(\alpha^t)]}.$$

Exact Relaxation for Squared Loss

Theorem (AdaSDCA for squared loss)

Consider AdaSDCA. If all the loss functions $\{\phi_i\}$ are squared loss functions, then

$$\mathbb{E}[P(w^t) - D(\alpha^t)] \le \frac{1}{\tilde{\theta}_t} \prod_{k=0}^{t} (1 - \tilde{\theta}_k) \left(D(\alpha^*) - D(\alpha^0)\right),$$

for all $t \ge 0$ where

$$\tilde{\theta}_t \stackrel{def}{=} \frac{\mathbb{E}[\theta(\kappa^t, p^t)(P(w^t) - D(\alpha^t))]}{\mathbb{E}[P(w^t) - D(\alpha^t)]}.$$

Exact Relaxation for Squared Loss

Theorem (AdaSDCA for squared loss)

Consider AdaSDCA. If all the loss functions $\{\phi_i\}$ are squared loss functions, then

$$\mathbb{E}[P(w^t) - D(\alpha^t)] \leq \frac{1}{\tilde{\theta}_t} \prod_{k=0}^{t} (1 - \tilde{\theta}_k) \left(D(\alpha^*) - D(\alpha^0) \right),$$

for all $t \geq 0$ where

$$\tilde{\theta}_t \overset{def}{=} \frac{\mathbb{E}[\theta(\kappa^t, p^t)(P(w^t) - D(\alpha^t))]}{\mathbb{E}[P(w^t) - D(\alpha^t)]}.$$

Optimal adaptive sampling probability is given by:

$$(\tilde{p}^*(\kappa^t))_i \sim |\kappa_i^t| \sqrt{\|A_i\|^2 + n\lambda\gamma}, \quad \forall i \in [n].$$

AdaSDCA

Dual solution

For $t \geq 1$:

1. Compute dual residue κ^t: $\kappa_i^t = \alpha_i^t + \nabla\phi_i(A_i^\top w^t)$
   Set $p_i^t \sim |\kappa_i^t|\sqrt{\|A_i\|^2 + n\lambda\gamma}$
2. Randomly pick $i_t \in \{1, \dots, n\}$ with probability proportional to p^t
3. Update $\alpha_{i_t}^t$
   $$\alpha_{i_t}^t = \arg\max_{\beta\in\mathbb{R}} \left\{ -\phi_{i_t}^\star(\beta) - (A_{i_t}^\top w^{t-1})\beta - \frac{\|A_{i_t}\|^2}{2\lambda n}|\beta - \alpha_{i_t}^{t-1}|^2 \right\}$$

Heuristic and Efficient Variant of AdaSDCA

AdaSDCA+:

Dual solution

For $t \geq 1$:

1. If $mod(t, n) = 0$, then

   Option I: Adaptive Sampling Probability

   Compute dual residue κ^t: $\kappa_i^t = \alpha_i^t + \nabla\phi_i(A_i^\top w^t)$

   Set $p_i^t \sim |\kappa_i^t|\sqrt{\|A_i\|^2 + n\lambda\gamma}$

   Option II: Importance Sampling Probability

   Set $p_i^t \sim \|A_i\|^2 + n\lambda\gamma$

2. Randomly pick $i_t \in \{1, \ldots, n\}$ according to p^t
3. Update $\alpha_{i_t}^t$
4. Update Probability: $p^{t+1} \sim (p_1^t, \ldots, p_{i_t}^t/m, \ldots p_n^t)$

Computational Cost per Epoch

Algorithm	cost of an epoch
SDCA	$O(\text{nnz})$
Iprox-SDCA	$O(\text{nnz} + n\log(n))$
AdaSDCA	$O(n \cdot \text{nnz})$
AdaSDCA+	$O(\text{nnz} + n\log(n))$

Table 1: One epoch computational cost of different algorithms

Numerical Experiments

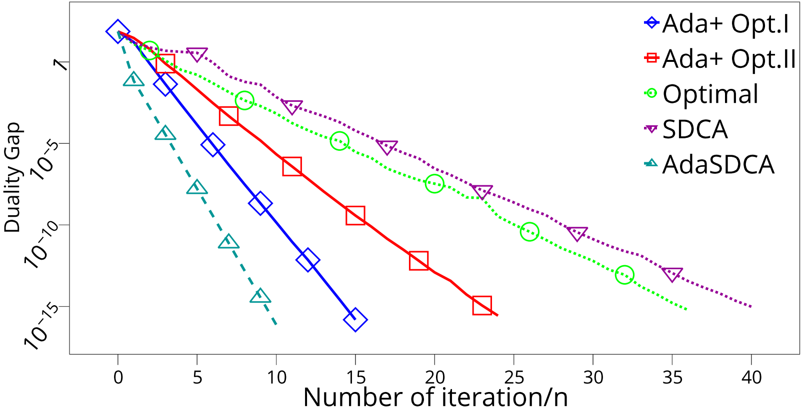

Figure 1: **w8a** dataset $d = 300$, $n = 49749$, Quadratic loss with L_2 regularizer, $\lambda = 1/n, \gamma = 1$.

Numerical Experiments

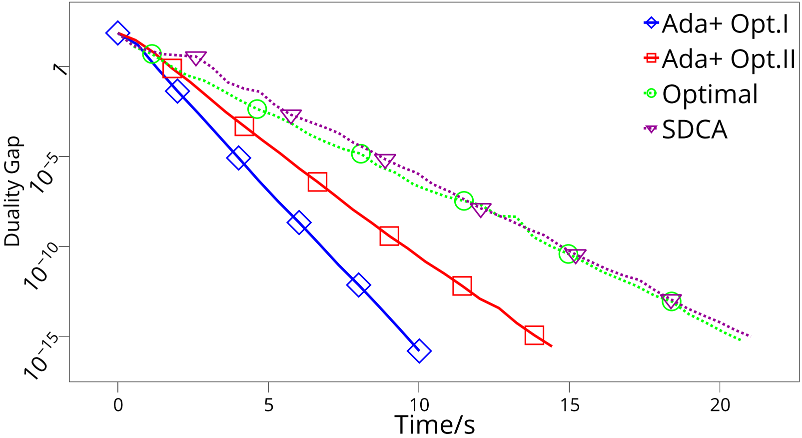

Figure 2: **w8a** dataset $d = 300$, $n = 49749$, Quadratic loss with L_2 regularizer, $\lambda = 1/n, \gamma = 1$.

Numerical Experiments

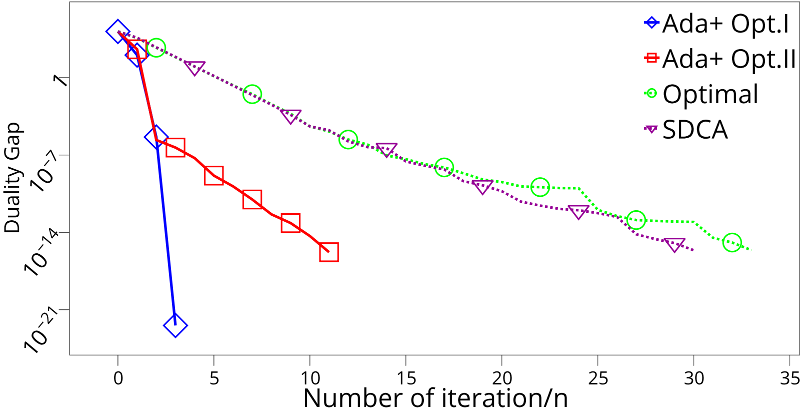

Figure 3: **cov1** dataset: $d = 54, n = 581,012$. Smooth Hinge loss with L_2 regularizer, $\lambda = 1/n, \gamma = 1$.

Numerical Experiments

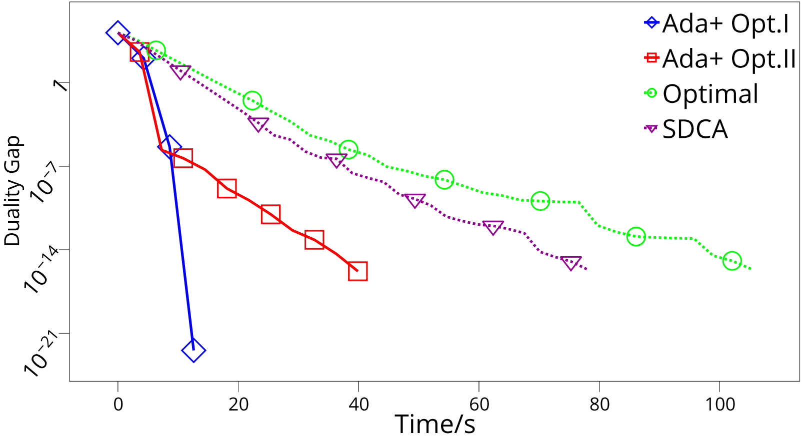

Figure 4: **cov1** dataset: $d = 54, n = 581,012$. Smooth Hinge loss with L_2 regularizer, $\lambda = 1/n, \gamma = 1$.

Numerical Experiments

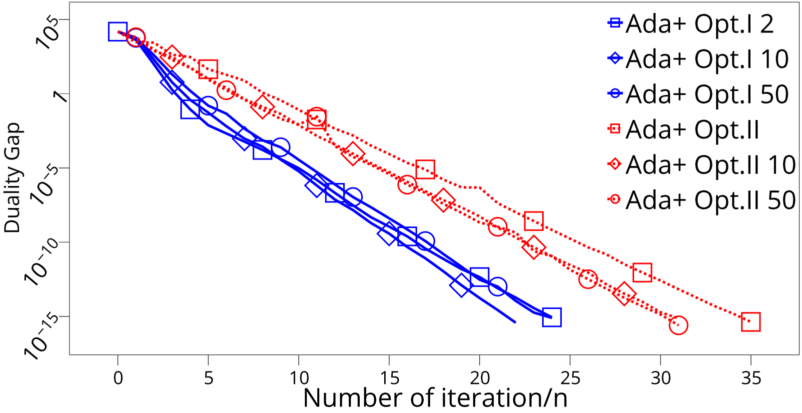

Figure 5: **cov1** dataset: $d = 54, n = 581,012$. Smooth Hinge loss with L_2 regularizer, $\lambda = 1/n, \gamma = 1$. comparison of different choices of the constant m.

More on ESO

Q. and Richtarik. Coordinate descent with arbitrary sampling II: expected separable overapproximation, *Optimization methods and software, 2016.*

ESO

- The function f admits an expected separable overapproximation (ESO) w.r.t. $\hat{S}$ and $v \in \mathbb{R}^n_+$, denoted as $(f, \hat{S}) \sim ESO(v)$, if

$$\mathbb{E}[f(x + h_{[\hat{S}]})] \leq f(x) + \langle \nabla f(x), h \rangle_p + \frac{1}{2}\|h\|^2_{v \circ p}, \quad \forall x, h \in \mathbb{R}^n.$$

ESO

- The function f admits an expected separable overapproximation (ESO) w.r.t. $\hat{S}$ and $v \in \mathbb{R}^n_+$, denoted as $(f, \hat{S}) \sim ESO(v)$, if

$$\mathbb{E}[f(x + h_{[\hat{S}]})] \leq f(x) + \langle \nabla f(x), h \rangle_p + \frac{1}{2} \|h\|^2_{v \circ p}, \quad \forall x, h \in \mathbb{R}^n.$$

- Recall the smoothness assumption:

$$f(x + h) \leq f(x) + \langle \nabla f(x), h \rangle + \frac{1}{2} \|Ah\|^2, \quad \forall x, h \in \mathbb{R}^n$$

ESO

- The function f admits an expected separable overapproximation (ESO) w.r.t. $\hat{S}$ and $v \in \mathbb{R}^n_+$, denoted as $(f, \hat{S}) \sim ESO(v)$, if

$$\mathbb{E}[f(x + h_{[\hat{S}]})] \leq f(x) + \langle \nabla f(x), h \rangle_p + \frac{1}{2}\|h\|^2_{v \circ p}, \quad \forall x, h \in \mathbb{R}^n.$$

- Recall the smoothness assumption:

$$f(x + h) \leq f(x) + \langle \nabla f(x), h \rangle + \frac{1}{2}\|Ah\|^2, \quad \forall x, h \in \mathbb{R}^n$$

- $(f, \hat{S}) \sim ESO(v)$ if

ESO

- The function f admits an expected separable overapproximation (ESO) w.r.t. $\hat{S}$ and $v \in \mathbb{R}^n_+$, denoted as $(f, \hat{S}) \sim ESO(v)$, if

$$\mathbb{E}[f(x + h_{[\hat{S}]})] \leq f(x) + \langle \nabla f(x), h \rangle_p + \frac{1}{2}\|h\|^2_{v \circ p}, \ \ \forall x, h \in \mathbb{R}^n.$$

- Recall the smoothness assumption:

$$f(x + h) \leq f(x) + \langle \nabla f(x), h \rangle + \frac{1}{2}\|Ah\|^2, \ \ \forall x, h \in \mathbb{R}^n$$

- $(f, \hat{S}) \sim ESO(v)$ if

$$\mathbb{E}[\|Ah_{[\hat{S}]}\|^2] \leq \|h\|^2_{v \circ p}, \ \ \forall h \in \mathbb{R}^n$$

ESO

- The function f admits an expected separable overapproximation (ESO) w.r.t. $\hat{S}$ and $v \in \mathbb{R}^n_+$, denoted as $(f, \hat{S}) \sim ESO(v)$, if

$$\mathbb{E}[f(x + h_{[\hat{S}]})] \leq f(x) + \langle \nabla f(x), h \rangle_p + \frac{1}{2}\|h\|^2_{v \circ p}, \quad \forall x, h \in \mathbb{R}^n.$$

- Recall the smoothness assumption:

$$f(x + h) \leq f(x) + \langle \nabla f(x), h \rangle + \frac{1}{2}\|Ah\|^2, \quad \forall x, h \in \mathbb{R}^n$$

- $(f, \hat{S}) \sim ESO(v)$ if

$$\mathbb{E}[\|Ah_{[\hat{S}]}\|^2] = h^\top \mathbb{E}[I_{\hat{S}}^\top A^\top A I_{[\hat{S}]}]h \leq \|h\|^2_{v \circ p}, \quad \forall h \in \mathbb{R}^n$$

Deriving Stepsize

Find $v \in \mathbb{R}^n_+$ scuh that

$$\mathbb{E}[I_{\hat{S}}^\top A^\top A I_{[\hat{S}]}] \preceq \mathrm{Diag}(v \circ p)$$

Deriving Stepsize

Find $v \in \mathbb{R}^n_+$ scuh that

$$\mathbb{E}[I_{\hat{S}}^\top A^\top A I_{[\hat{S}]}] \preceq \mathrm{Diag}(v \circ p)$$

- $\mathbb{E}[I_{\hat{S}}^\top A^\top A I_{[\hat{S}]}] = P \circ (A^\top A)$ where $P_{ij} = \mathbb{P}(i \in \hat{S}, j \in \hat{S})$

Deriving Stepsize

Find $v \in \mathbb{R}^n_+$ scuh that

$$\mathbb{E}[I_{\hat{S}}^\top A^\top A I_{[\hat{S}]}] \preceq \mathrm{Diag}(v \circ p)$$

- $\mathbb{E}[I_{\hat{S}}^\top A^\top A I_{[\hat{S}]}] = P \circ (A^\top A)$ where $P_{ij} = \mathbb{P}(i \in \hat{S}, j \in \hat{S})$
- Let $A = (A_1^\top, \ldots A_m^\top)^\top$, then

$$P \circ (A^\top A) = \sum_{j=1}^{m} P \circ (A_j^\top A_j)$$

Deriving Stepsize

Find $v \in \mathbb{R}^n_+$ scuh that

$$\mathbb{E}[I_{\hat{S}}^\top A^\top A I_{[\hat{S}]}] \preceq \mathrm{Diag}(v \circ p)$$

- $\mathbb{E}[I_{\hat{S}}^\top A^\top A I_{[\hat{S}]}] = P \circ (A^\top A)$ where $P_{ij} = \mathbb{P}(i \in \hat{S}, j \in \hat{S})$
- Let $A = (A_1^\top, \ldots A_m^\top)^\top$, then

$$P \circ (A^\top A) = \sum_{j=1}^{m} P \circ (A_j^\top A_j)$$

- Denote

$$J_j := \{i \in [n] \ : \ A_{ji} \neq 0\},$$

Deriving Stepsize

Find $v \in \mathbb{R}^n_+$ scuh that

$$\mathbb{E}[I_{\hat{S}}^\top A^\top A I_{[\hat{S}]}] \preceq \mathrm{Diag}(v \circ p)$$

- $\mathbb{E}[I_{\hat{S}}^\top A^\top A I_{[\hat{S}]}] = P \circ (A^\top A)$ where $P_{ij} = \mathbb{P}(i \in \hat{S}, j \in \hat{S})$
- Let $A = (A_1^\top, \ldots A_m^\top)^\top$, then

$$P \circ (A^\top A) = \sum_{j=1}^m P \circ (A_j^\top A_j)$$

- Denote

$$J_j := \{i \in [n] \ : \ A_{ji} \neq 0\},$$

then

$$P \circ (A^\top A) = \sum_{j=1}^m P \circ (A_j^\top A_j) = \sum_{j=1}^m P_{[J_j]} \circ (A_j^\top A_j)$$

Deriving Stepsize

Theorem (ESO with coupling between sampling and data)

Let $\hat{S}$ be an arbitrary sampling and $v = (v_1, \ldots, v_n)$ be defined by:

$$v_i = \sum_{j=1}^{m} \lambda'(J_j \cap \hat{S}) A_{ji}^2, \quad i = 1, 2, \ldots, n,$$

where

$$\lambda'(J \cap \hat{S}) := \max_{h \in \mathbb{R}^n} \{ h^\top P_{[J]} h \ : \ h^\top \mathrm{Diag}(\mathbb{P}_{[J]}) h \leq 1 \}.$$

Then $(f, \hat{S}) \sim ESO(v)$.

Deriving Stepsize

Tight bounds for:

- serial sampling $\lambda'(J \cap \hat{S}) = 1$;

Deriving Stepsize

Tight bounds for:

- serial sampling $\lambda'(J \cap \hat{S}) = 1$;
- uniform distribution over subsets of fixed size τ (aka τ-nice sampling) ([Richtarik & Takac 13])

$$\lambda'(J \cap \hat{S}) = 1 + \frac{(|J| - 1)(\tau - 1)}{\max(n-1, 1)} .$$

Deriving Stepsize

Tight bounds for:

- serial sampling $\lambda'(J \cap \hat{S}) = 1$;
- uniform distribution over subsets of fixed size τ (aka τ-nice sampling) ([Richtarik & Takac 13])

$$\lambda'(J \cap \hat{S}) = 1 + \frac{(|J| - 1)(\tau - 1)}{\max(n - 1, 1)} \quad .$$

- distributed sampling with datas equally partitionned on c processors, each of which draws independently a τ-nice sampling ([Fercoq, Q. , Richtarik & Takac 14])

$$\lambda'(J \cap \hat{S}) \leq \left(1 + \frac{1}{\tau - 1}\right)\left(1 + \frac{(|J| - 1)(\tau - 1)}{\max(n/c - 1, 1)}\right) \quad .$$

Conclusion

- Unified convergence analysis for Randomized coordinate descent method
  - Accelerated Randomized coordinate descent method
  - Arbitrary sampling
- Convergence condition (ESO)+Formulae for computing admissible stepsizes
- Adaptive sampling using duality gap