Smooth minimization of nonsmooth functions by parallel coordinate descent

Olivier Fercoq and Peter Richtárik

The University of Edinburgh

1$^{\mathrm{st}}$ May 2013

Minimization of functions with max structure

Minimize for $x \in \mathbb{R}^n$ the structured nonsmooth function f

$$f(x) = \max_{u \in Q_2 \subseteq \mathbb{R}^m} \langle Ax, u \rangle - \langle b, u \rangle$$

- $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$, Q_2 convex
- Also for composite functions $F(x) = f(x) + \psi(x)$
  $\psi : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$, convex, closed, separable

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

Setting

Norms:

- norm in $\mathbb{R}^n$: $\|\cdot\|_1$
- norm in $\mathbb{R}^m$: $\|\cdot\|_2$
- induced norm

$$\|A\|_{1,2} = \max_{x \in \mathbb{R}^n} \{\|Ax\|_2^* : \|x\|_1 = 1\} = \max_{u \in \mathbb{R}^m} \{\|A^T u\|_1^* : \|u\|_2 = 1\}$$

Prox function d_2: σ_2-strongly convex on Q_2 with minimizer u_0

- $d_2(u) \geq \sigma_2 \|u - u_0\|_2^2, \ \forall u \in Q_2$
- $u_0 = \arg\min_{u \in Q_2} d_2(u)$
- $D_2 = \max\{d_2(u) : u \in Q_2\}$

Nesterov's smoothing

For $\mu > 0$, smooth approximation f_μ of f:

$$f_\mu(x) = \max_{u \in Q_2 \subseteq \mathbb{R}^m} \{\langle Ax, u\rangle - \langle b, u\rangle - \mu d_2(u)\}$$

Theorem (Nesterov, 2005)

$$f_\mu(x) \leq f(x) \leq f_\mu(x) + \mu D_2, \quad \forall x \in \mathbb{R}^n$$

$$\|\nabla f_\mu(x+h) - \nabla f_\mu(x)\|_1^* \leq \frac{\|A\|_{1,2}^2}{\mu \sigma_2} \|h\|_1$$

Any ε-solution of f_μ is $(\varepsilon + \mu D_2)$-solution of f

Examples

- Sum of absolute values

$$f(x) = \sum_{j=1}^{m} |e_j^T Ax - b^j| = \max_{u \in [-1,1]^n} \langle Ax, u \rangle - \langle b, u \rangle$$

$$f_\mu(x) = \sum_{j=1}^{m} \|e_j^T A\|_1^* \psi_\mu \Big(\frac{|e_j^T Ax - b^j|}{\|e_j^T A\|_1^*} \Big)$$

$$\psi_\mu(t) = \begin{cases} \frac{t^2}{2\mu}, & 0 \le t \le \mu \\ t - \frac{\mu}{2}, & \mu \le t \end{cases}$$

- Maximum of linear functions: $\tilde{A} = \begin{bmatrix} A \\ -A \end{bmatrix}$, $\tilde{b} = \begin{bmatrix} b \\ -b \end{bmatrix}$

$$f(x) = \max_{1 \le j \le m} |e_j^T Ax - b^j| = \max_{u \in \Sigma_{2m}} \langle \tilde{A}x, u \rangle - \langle \tilde{b}, u \rangle$$

$$f_\mu(x) = \mu \log \left(\frac{1}{2m} \sum_{j=1}^{2m} \exp\Big(\frac{e_j^T \tilde{A}x - \tilde{b}^j}{\mu} \Big) \right)$$

Smoothness of f_μ

Proposition

If $\|x\|_1 = (\sum_{i=1}^n (x^i)^2)^{1/2}$, ∇f_μ is coordinate-wise Lipschitz with constants $L_i = \frac{(\|Ae_i\|_2^*)^2}{\mu\sigma_2}$:

$$\forall x \in \mathbb{R}^n, t \in \mathbb{R}, i \in \{1, \dots, n\},$$

$$|\nabla_i f_\mu(x) - \nabla_i f_\mu(x + te_i)| \leq \frac{(\|Ae_i\|_2^*)^2}{\mu\sigma_2} \|te_i\|_1$$

Fact

Let $t \in \mathbb{R}$ and $\|x\|_L = (\sum_{i=1}^n L_i (x^i)^2)^{1/2}$

$$f_\mu(x + te_i) \leq f_\mu(x) + \langle \nabla f_\mu(x), te_i \rangle + \frac{1}{2}\|te_i\|_L^2$$

Serial Coordinate Descent

At each iteration:

1. Choose at random a coordinate i
2. Compute update $t \in \mathbb{R}$ that minimizes the overapproximation of $f_\mu(x + te_i)$
3. Update variable $x = x + te_i$.

Remarks:

- Very cheap iterations
- Many iterations required

Parallel Coordinate Descent Method

[Richtárik, Takáč, 2012]

At each iteration:

1. Choose a random subset of variables $\hat{S}$ (sampling)
2. In parallel for $i \in \hat{S}$
   a. Compute update $h^i \in \mathbb{R}$
   b. Update variable $x^i = x^i + h^i$.

- More general overapproximation to calculate updates
- Theory for smooth partially separable functions

$$f(x) = \sum_{J \in \mathcal{J}} f_J(x)$$

f_J depends on variable i only if $i \in J$

$|J| \leq \omega$ for all $J \in \mathcal{J}$

Max-partially separable functions

Definition

f is max-partially separable of degree ω
if it can be written in the form

$$f(x) = \max_{u \in Q_2 \subseteq \mathbb{R}^m} \langle Ax, u \rangle - \langle b, u \rangle$$

where $Q_2 \subseteq \mathbb{R}^m$ is convex, $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$ and

$$\max_{1 \leq j \leq m} |\{i : A_{j,i} \neq 0\}| \leq \omega$$

Deterministic Separable Overapproximation

Theorem

For $v_1, \ldots, v_m > 0$, $1 \leq p \leq 2$, *define* $\|u\|_2 = (\sum_{j=1}^m v_j u_j^p)^{1/p}$
Let $\|x\|_1 = \|x\|_w = (\sum_{i=1}^n w_i x_i^2)^{1/2}$ *where*

$$w_i = (\|Ae_i\|_2^*)^2 = \begin{cases} (\sum_{j=1}^m v_j^{-1} |A_{j,i}|^q)^{2/q}, & 1 < p \leq 2, \frac{1}{p} + \frac{1}{q} = 1 \\ \max_{1 \leq j \leq m} v_j^{-2} A_{j,i}^2, & p = 1 \end{cases}$$

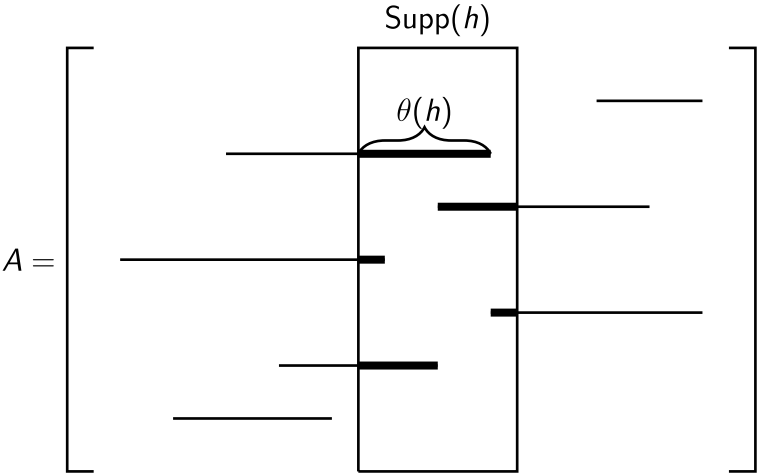
If f *is max-partially separable of degree* ω, *then:*

$$f_\mu(x+h) \leq f_\mu(x) + \langle \nabla f_\mu(x), h \rangle + \frac{\theta(h)}{2\mu\sigma_2} \|h\|_w{}^2$$

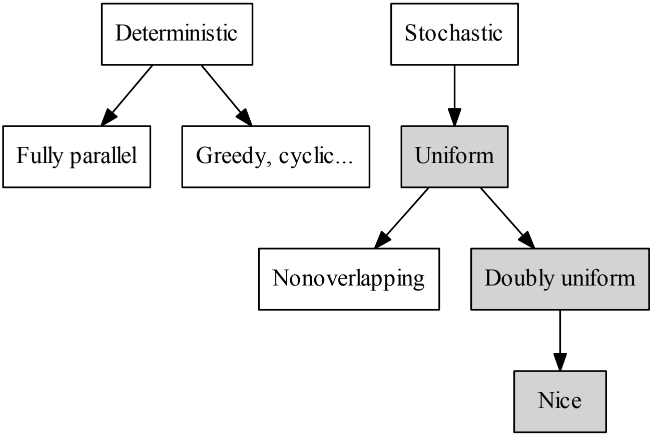
where

$$\theta(h) = \max_{1 \leq j \leq m} |\{i : A_{j,i} \neq 0 \text{ and } h_i \neq 0\}| \leq \min(\omega, \|h\|_0)$$

Illustration of $\theta(h)$



Samplings



Uniform sampling:
$\mathbf{P}(i \in \hat{S}) = constant$

Doubly uniform:
if $|S'| = |S''| = j$,
$\mathbf{P}(S') = \mathbf{P}(S'') = q_j$

τ-nice sampling:
$q_\tau = 1$.

Expected Separable Overapproximation (ESO)

Definition

φ admits a (β, w)-ESO with respect to sampling $\hat{S}$ if

$$\mathbf{E}[\varphi(x + h_{[\hat{S}]})] \leq \varphi(x) + \frac{\mathbf{E}[|\hat{S}|]}{n}\Big(\langle \nabla \varphi(x), h\rangle + \frac{\beta}{2}\|h\|_w^2\Big)$$

Theorem

If f is max-partially separable of degree $\boldsymbol{\omega}$ and $\hat{S}$ is a uniform sampling with $\mathbf{P}(|\hat{S}| = \boldsymbol{\tau}) = 1$, then

$$(f_\mu, \hat{S}) \sim \mathrm{ESO}\Big(\min(\boldsymbol{\omega}, \boldsymbol{\tau}), \frac{w}{\mu\sigma_2}\Big)$$

ESO with 1-norm

Theorem

Assume that:

- f *is max-partially separable of degree* ω
- $\|u\|_2 = \sum_{j=1}^m v_j |u_j|$ *a weighted 1-norm*
- $\hat{S}$ *is a* τ*-nice sampling*
- $\mu > 0$

Then

$$(f_\mu, \hat{S}) \sim \mathrm{ESO}\Big(\sum_{k=1}^{\min(\omega,\tau)} \min\Big(1, m \sum_{l=k}^{\min(\omega,\tau)} p_l\Big), \frac{w}{\mu\sigma_2}\Big)$$

where $\qquad p_l = \dfrac{\binom{\omega}{l}\binom{n-\omega}{\tau-l}}{\binom{n}{\tau}}$

ESO with 2-norm

Theorem

Assume that:

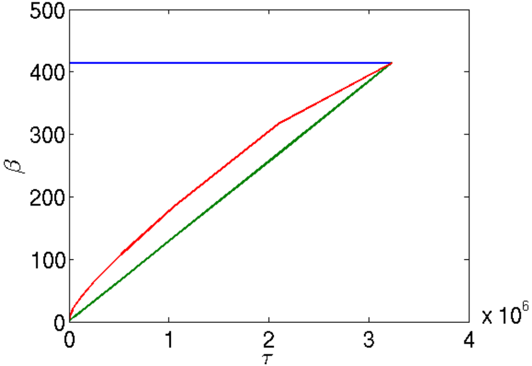
- f *is max-partially separable of degree* ω
- $\|\cdot\|_2$ *is a weighted 2-norm*
- $\hat{S}$ *is a* τ*-nice sampling*
- $\mu > 0$

Then

$$(f_\mu, \hat{S}) \sim \mathrm{ESO}\left(1 + \frac{(\omega - 1)(\tau - 1)}{\max(1, n-1)}, \frac{w}{\mu \sigma_2}\right)$$

Exactly the same formula as for partially separable functions!

Comparison of ESO's



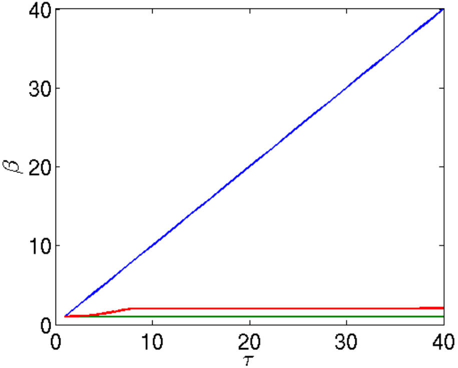
$m = 2,396,130$
$n = 3,231,961$
$\omega = 414$

β as a function of the number of processors

$\beta_1 = \sum_{k=1}^{\min(\omega,\tau)} \min\left(1, m \sum_{l=k}^{\min(\omega,\tau)} p_l\right)$

$\beta_{DSO} = \min(\omega, \tau)$ $\beta_2 = (1 + \frac{(\omega-1)(\tau-1)}{\max(1,n-1)})$

Zoom for small number of processors



$m = 2,396,130$
$n = 3,231,961$
$\omega = 414$

β as a function of the number of processors

— $\beta_1 = \sum_{k=1}^{\min(\omega,\tau)} \min\left(1, m \sum_{l=k}^{\min(\omega,\tau)} p_l\right)$

— $\beta_{DSO} = \min(\omega, \tau)$ — $\beta_2 = \left(1 + \frac{(\omega-1)(\tau-1)}{\max(1,n-1)}\right)$

Iteration complexity

Theorem

For $\mu = \varepsilon/(2D_2)$, *let* (x_k) *be the sequence generated by Parallel Coordinate Descent with sampling* $\hat{S}$ *applied to* $F_\mu = f_\mu + \psi$. *If* $(f_\mu, \hat{S}) \sim ESO(\beta, \frac{w}{\mu\sigma_2})$ *and*

$$k \geq \frac{8D_2\mathcal{R}_w^2(x_0)}{\sigma_2}\frac{n\beta}{\tau}\frac{1}{\varepsilon^2}\left(1 + \log\frac{1}{\rho}\right) + 2$$

Then $\mathbf{P}(F(x_k) - F^* \leq \varepsilon) \geq 1 - \rho$

Comparison of algorithms: Infinity norm

- $f(x) = \max_{1 \leq j \leq m} |(Ax)_j - b_j|$
- Dorothea dataset: $m = 800$, $n = 100,000$, $\omega = 6,061$
- $\varepsilon = 0.01$

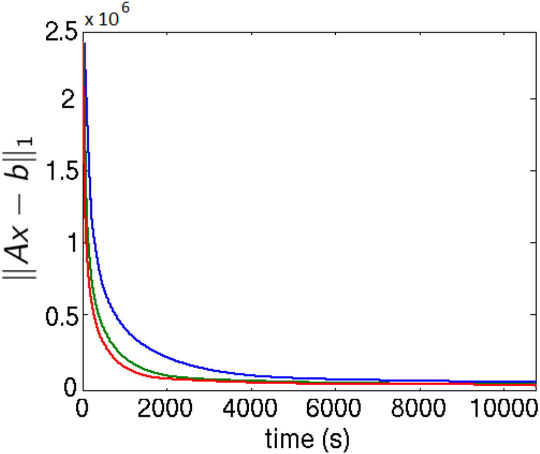
Algorithm	Comp time
GLPK simplex	681 s
Accelerated gradient[1]	10,000 s
Sparse subgradient[2] opt value *known*	6.4s
Sparse subgradient[2] opt value unknown	544 s
Smoothed PCDM[3], τ=4 cores	55 s
Smoothed PCDM[3], τ=16 cores	34 s

[1] Nesterov 2005, Smooth minimization of non-smooth functions
[2] Nesterov 2012, Subgradient Methods for Huge-Scale Optimization Problems
[3] Smoothed Parallel Coordinate Descent Method

Bigger Dataset

- URL reputation dataset
- Feature matrix $A \in \mathbb{R}^{m \times n}$
  $m = 2,396,130$, $n = 3,231,961$, $\omega = 414$
- $b \in \mathbb{R}^m$: spam or non spam page
- Two test cases
  1. Least absolute deviations: $f(x) = \sum_{j=1}^m |(Ax)_j - b_j|$
  2. Exponential loss: $f(x) = \sum_{j=1}^m \exp(b_j \sum_{i=1}^n A_{j,i} x_i)$
- 64-cores Intel(R) Xeon(R) @ 2.60GHz, 128GB RAM
- Asynchronous implementation

Minimization of 1-norm



Number of processors

$\tau = 1$

$\tau = 2$

$\tau = 4$

Minimization of $\|Ax - b\|_1$ by parallel coordinate descent

$m = 2,396,130,\ n = 3,231,961,\ \omega = 414,\ \varepsilon = 5.10^3,\ \mu = 2.10^{-4}$

Alternative smoothing

Comparison of two smoothings for least absolute deviations

	Smoothed PCDM	Mini-batch SDCA
Regularization	Dual space	Primal space
Parameter	$\mu = \frac{\varepsilon}{2D_2}$	$\lambda = \frac{\varepsilon}{2\|x^*\|^2}$
Parallelization speedup factor	$\frac{1}{\tau}(1 + \frac{(\omega-1)(\tau-1)}{n-1})$	$\frac{1}{\tau}(1 + \frac{(\sigma^2-1)(\tau-1)}{n-1})$ $\sigma^2(\frac{A}{\max_i \|Ae_i\|}) \leq \omega$
Complexity	$O(\|x^*\|^2/\varepsilon^2)$	$O(\|x^*\|^2/\varepsilon^2)$
If $\|x^*\|$ unknown	Cannot stop	Cannot start
Stopping criterion	Iteration complexity	Duality gap

Shalev-Schwartz and Zhang, 2012, Proximal Stochastic Dual Coordinate Ascent

Takáč, Bijral, Richtárik and Srebro, 2013, Mini-batch primal and dual methods for support vector machines

Parallel Adaboost

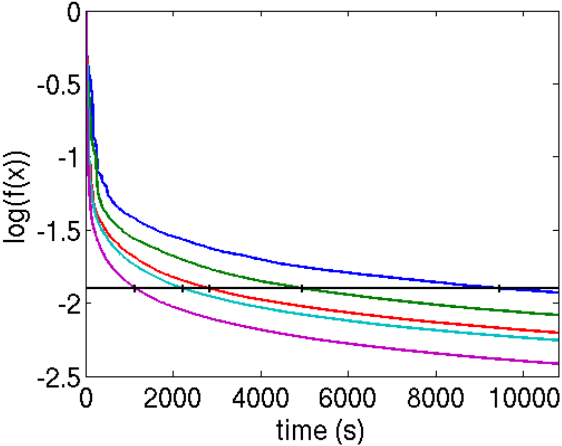
Adaboost = greedy serial coordinate descent for

$$f(x) = \sum_{j=1}^{m} \exp(b_j \sum_{i=1}^{n} A_{j,i} x_i)$$

Parallelization:

- Colins, Shapire and Singer 2002: $\beta = \omega$ (fully parallel)
- Palit and Reddy, 2012: $\beta = \tau$ (generalized greedy)
- here: $\varphi(x) = \max_{1 \leq j \leq m} b_j (Ax)_j$ is max-partially separable
  With $\mu = 1$, $\log \circ f = \varphi_\mu$
  $\beta = \min(\omega, \tau)$ for any sampling
  $\beta_1 < \min(\omega, \tau)$ for τ-nice sampling

Experiment on URL reputation dataset



Number of processors

- $\tau = 1$
- $\tau = 2$
- $\tau = 4$
- $\tau = 8$
- $\tau = 16$

Minimization of exponential loss by parallel coordinate descent

$m = 2,396,130$, $n = 3,231,961$, $\omega = 414$

Conclusion

- Fine study of coordinate-wise Lipschitz constants of max-partially separable functions
- Expected Separable Overapproximations giving the theoretical parallelization speedup
- Promising numerical experiments
- Commom framework for partially and max-partially separable functions?