Smooth Primal-Dual Coordinate Descent Algorithms for Nonsmooth Convex Optimization

A. Alacaoglu, V. Cevher, *O. Fercoq*, Q. Tran-Dinh

14 November 2017

THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL

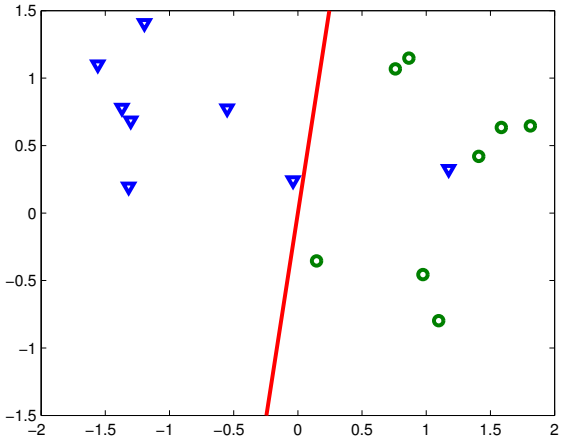
Programme PGMO
Pour l'optimisation et la recherche opérationnelle

Composite convex optimisation

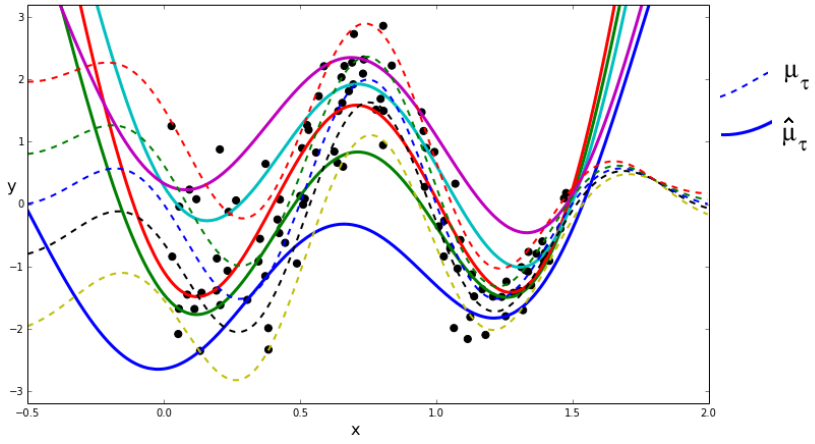
$$\min_{x \in \mathcal{X}} \quad f(x) + g(x) + h(Ax)$$

- $f : \mathcal{X} \to \mathbb{R}$ is a differentiable convex function with $L(\nabla_i f)$-Lipschitz partial derivatives
- $g : \mathcal{X} \to \mathbb{R} \cup \{+\infty\}$ is a l.s.c. convex function with a simple proximal operator
- $h : \mathcal{Y} \to \mathbb{R} \cup \{+\infty\}$ is a l.s.c. convex function with a simple proximal operator
- $A : \mathcal{X} \to \mathcal{Y}$ is a linear map

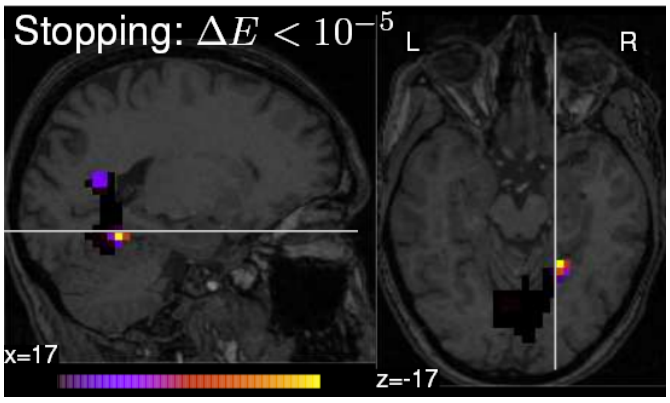
Examples



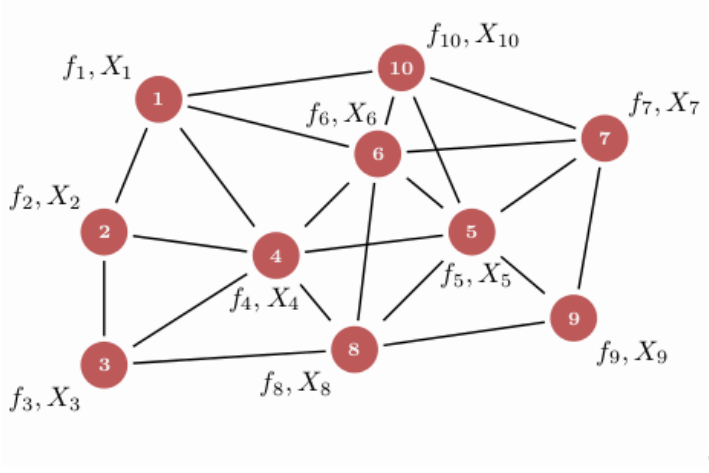
Support Vector Machines with intercept



Quantile regression



TV-regularized regression



©Mota

Distributed optimization

Primal-dual algorithms

- Equivalent saddle point formulation
  if $0 \in \operatorname{relint}(\operatorname{dom} h - A \operatorname{dom} g)$

$$\min_{x \in \mathcal{X}} \; f(x) + g(x) + h(Ax)$$

$$= \max_{y \in \mathcal{Y}} \min_{x \in \mathcal{X}} f(x) + g(x) + \langle Ax, y \rangle - h^\star(y)$$

  where $h^*(y) = \sup_{z \in \mathcal{Y}} \langle z, y \rangle - h(z)$ (Fenchel conjugate)

- Example: Vũ-Condat method

$$y_{k+1} = \operatorname{prox}_{\sigma h^\star}(y_k + Ax_k)$$
$$x_{k+1} = \operatorname{prox}_{\tau g}\left(x_k - \tau(\nabla f(x_k) + A^\top(2y_{k+1} - y_k))\right)$$

  Splits f, g, h and A.

Coordinate descent

- Updates 1 coordinate per iteration:
  - needs many iterations
  - efficient updates
  - allows longer step sizes
- Comparison gradient descent vs coordinate descent for $g = h = 0$ and $f(x) = \|Mx - b\|_2^2$

Algo	$x^{k+1} = x^k - \frac{1}{L(\nabla f)} \nabla f(x^k)$	$x_i^{k+1} = x_i^k - \frac{1}{L(\nabla_i f)} \nabla_i f(x^k)$
Parameter	$L(\nabla f) = \|M\|_2^2$	$L(\nabla_i f) = \|Me_i\|_2^2$
Complexity	$O(L(\nabla f)/k)$	$O(n \max_i L(\nabla_i f)/k)$
Cost 1 iter	$2\,\mathrm{nnz}(M)$	$2\,\mathrm{nnz}(Me_i)$
Total cost	$O(\mathrm{nnz}(M) L(\nabla f)/k)$	$O(\mathrm{nnz}(M) \max_i L(\nabla_i f)/k)$

Previous works

[Combettes, Pesquet, 2014]+[Bianchi, Hachem, Iutzeler, 2016]
no convergence rate, small step sizes

[Fercoq, Bianchi, 2015]
no convergence rate, longer step sizes

[Zhang, Xiao, 2017]
only strongly convex-concave lagrangians + all primal variables updated together

[Gao, Xu, Zhang, 2017] + [Chambolle, Ehrhardt, Richtárik, Schönlieb, 2017]
all dual variables updated together

Idea of this work

Combine features from:

- Smooth minimization of nonsmooth functions
  [Nesterov, 2005] + [Fercoq, Richtárik, 2013]
- Accelerated proximal coordinate descent
  [Fercoq, Richtárik, 2015]
- Accelerated smoothed gap reduction algorithm
  [Tran-Dinh, Fercoq, Cevher, 2017]

Smoothing nonsmooth functions

- We define the smoothed function [Nesterov 2005]

$$h_\beta(z;\dot{y}) = \max_y \langle z, y\rangle - h^*(y) - \frac{\beta}{2}\|y-\dot{y}\|_{\mathcal{Y}}^2$$

- h_β is differentiable wrt z and $\nabla_z h_\beta$ is $\frac{1}{\beta}$-Lipschitz
- $h_\beta(Ax^k,\dot{y}) = \langle \dot{y}, Ax^k - c\rangle + \frac{1}{2\beta}\|Ax^k - c\|_{\mathcal{Y},*}^2$

Fundamental theorem

$F = f + g$; $S_\beta(x, \dot{y}) = F(x) + h_\beta(Ax; \dot{y}) - F(x^\star) - h(Ax^\star)$

Theorem (Nesterov)

Suppose that $L(h) < +\infty$. *Then*
$F(x) + h(Ax) - F(x^\star) - h(Ax^\star) \leq S_\beta(x; \dot{y}) + 2\beta L(h)$

Fundamental theorem

$F = f + g$; $S_\beta(x, \dot{y}) = F(x) + h_\beta(Ax; \dot{y}) - F(x^\star) - h(Ax^\star)$

Theorem (Nesterov)

Suppose that $L(h) < +\infty$. *Then*
$F(x) + h(Ax) - F(x^\star) - h(Ax^\star) \leq S_\beta(x; \dot{y}) + 2\beta L(h)$

Theorem (Tran-Dinh, Fercoq, Cevher)

Consider $h = \delta_{\{c\}}$

$$\begin{cases} \|Ax - c\|_{\mathcal{Y},*} \leq \beta\Big[\|y^\star - \dot{y}\|_{\mathcal{Y}} + \big(\|y^\star - \dot{y}\|_{\mathcal{Y}}^2 + 2\beta^{-1} S_\beta(x; \dot{y})\big)^{1/2}\Big] \\ F(x) - F(x^\star) \geq -\|y^\star\|_{\mathcal{Y}} \|Ax - c\|_{\mathcal{Y},*} \\ F(x) - F(x^\star) \leq S_\beta(x, \dot{y}) + \|y^\star\|_{\mathcal{Y}} \|Ax - c\|_{\mathcal{Y},*} + \frac{\beta}{2}\|y^\star - \dot{y}\|_{\mathcal{Y}}^2 \end{cases}$$

If β and $S_\beta(x, \dot{y})$ are small, we have an approximate solution

SMooth, Accelerate, Randomize The Coordinate Descent (SMART-CD)

Choose $\beta_1 > 0$, $x^0 \in \mathcal{X}$ and probabilities $q \in \mathbb{R}^n$

Set $\tau_0 = \min_{1 \le i \le n} q_i$ and $\bar{x}^0 = \tilde{x}^0 = x^0$

For $k \in \{0, 1, \cdots, k_{\max}\}$

$\hat{x}^k = (1 - \tau_k)\bar{x}^k + \tau_k \tilde{x}^k$

Select $i_k \sim q$

$B_{i_k}^k = L(\nabla_{i_k}(f + h_{\beta_{k+1}} \circ A))$

$\tilde{x}_{i_k}^{k+1} = \text{prox}_{\frac{\tau_0}{\tau_k B_{i_k}^k} g_{i_k}} \left(\tilde{x}_{i_k}^k - \frac{\tau_0}{\tau_k B_{i_k}^k} \nabla_{i_k}(f + h_{\beta_{k+1}} \circ A)(\hat{x}^k)\right)$

$\bar{x}^{k+1} = \hat{x}^k + \frac{\tau_k}{\tau_0}(\tilde{x}^{k+1} - \tilde{x}^k)$

$\tau_{k+1} \in (0, 1)$ solution to $\tau^3 + \tau^2 + \tau_k^2 \tau - \tau_k^2 = 0$

$\beta_{k+2} = \frac{\beta_{k+1}}{1 + \tau_{k+1}}$

SMooth, Accelerate, Randomize The Coordinate Descent (SMART-CD)

Choose $\beta_1 > 0$, $x^0 \in \mathcal{X}$ and probabilities $q \in \mathbb{R}^n$

Set $\tau_0 = \min_{1 \le i \le n} q_i$ and $\bar{x}^0 = \tilde{x}^0 = x^0$

For $k \in \{0, 1, \cdots, k_{\max}\}$

$\hat{x}^k = (1 - \tau_k)\bar{x}^k + \tau_k \tilde{x}^k$

Select $i_k \sim q$

$B_{i_k}^k = L(\nabla_{i_k} f) + \frac{\|A_{i_k}\|^2}{\beta_{k+1}}$

$y_k^* = \text{prox}_{\beta_{k+1}^{-1} h^*}(\dot{y} + \beta_{k+1}^{-1} A\hat{x}^k)$

$\tilde{x}_{i_k}^{k+1} = \text{prox}_{\frac{\tau_0}{\tau_k B_{i_k}^k} g_{i_k}}\left(\tilde{x}_{i_k}^k - \frac{\tau_0}{\tau_k B_{i_k}^k}(\nabla_{i_k} f(\hat{x}^k) + A_{i_k}^\top y_k^*)\right)$

$\bar{x}^{k+1} = \hat{x}^k + \frac{\tau_k}{\tau_0}(\tilde{x}^{k+1} - \tilde{x}^k)$

$\tau_{k+1} \in (0, 1)$ solution to $\tau^3 + \tau^2 + \tau_k^2 \tau - \tau_k^2 = 0$

$\beta_{k+2} = \frac{\beta_{k+1}}{1 + \tau_{k+1}}$

Efficient implementation

Choose $\beta_1 > 0$, $x^0 \in \mathcal{X}$ and probabilities $q \in \mathbb{R}^n$

Set $\tau_0 = \min_{1 \le i \le n} q_i$ and $\bar{x}^0 = \tilde{x}^0 = x^0$

For $k \in \{0, 1, \cdots, k_{\max}\}$

Select $i_k \sim q$

$$B_{i_k}^k = L(\nabla_{i_k} f) + \frac{\|A_{i_k}\|^2}{\beta_{k+1}}$$

$$y_k^* = \operatorname{prox}_{\beta_{k+1}^{-1} h^*}(\dot{y} + \beta_{k+1}^{-1}(c_k A \check{x}^k + A\tilde{x}^k))$$

$$\tilde{x}_{i_k}^{k+1} = \operatorname{prox}_{\frac{\tau_0}{\tau_k B_{i_k}^k} g_{i_k}}\left(\tilde{x}_{i_k}^k - \frac{\tau_0}{\tau_k B_{i_k}^k}(\nabla_{i_k} f(c_k A \check{x}^k + A\tilde{x}^k) + A_{i_k}^\top y_k^*)\right)$$

$$\check{x}_{i_k}^{k+1} = \check{x}_{i_k}^k - \frac{1 - \tau_k/\tau_0}{c_k}(\tilde{x}_{i_k}^{k+1} - \tilde{x}_{i_k}^k)$$

$\tau_{k+1} \in (0, 1)$ solution to $\tau^3 + \tau^2 + \tau_k^2 \tau - \tau_k^2 = 0$

$$\beta_{k+2} = \frac{\beta_{k+1}}{1 + \tau_{k+1}}$$

Speed of convergence

$F = f + g$, $\tau_0 = \min_{1 \leq i \leq n} q_i \in O(1/n)$

Theorem

SMART-CD has a $O(n/k)$ convergence speed. If $h = \delta_{\{c\}}$,

$$\mathbb{E}(\|A\bar{x}^k - c\|_*) \leq \frac{\beta_1}{\tau_0(k-1)+1}\left[\|y^* - \dot{y}\| + \left(\|y^* - \dot{y}\|^2 + 4\beta_1^{-1}C^*\right)^{1/2}\right]$$

$$-\|y^\star\|\mathbb{E}(\|A\bar{x}^k - c\|_*) \leq \mathbb{E}(F(\bar{x}^k) - F(x^\star))$$

$$\mathbb{E}(F(\bar{x}^k) - F(x^\star)) \leq \frac{(1-\tau_0)C^*}{\tau_0(k-1)+1} + \frac{\beta_1\|y^\star - \dot{y}\|^2}{2(\tau_0(k-1)+1)} + \|y^*\|\mathbb{E}(\|A\bar{x}^k - c\|_*)$$

where $C^* = (1-\tau_0)(F_{\beta_0}(x^0) - F(x^\star)) + \sum_{i=1}^n \frac{\tau_0 B_i^0}{2q_i}\|x_i^\star - \tilde{x}_i^0\|_i^2$

Same kind of results for Lipschitzian h

Sketch of proof

- $\bar{x}^k$ is a convex combination of the $\tilde{x}^l$'s $(l \leq k)$
- Denote $\hat{F}^k_{\beta_k} = f(\bar{x}^k) + \sum_{l=0}^k \gamma^{k,l} g(\tilde{x}^l) + h_{\beta_k}(A\bar{x}^k; \dot{y})$
  If $\beta_{k+1}(1+\tau_k) - \beta_k \geq 0$, then

$$\mathbb{E}(\hat{F}^{k+1}_{\beta_{k+1}} - F^\star) + \mathbb{E}\Big[\sum_{i=1}^n \frac{\tau_k^2 B_i^k}{2q_i\tau_0}\|x_i^\star - \tilde{x}_i^{k+1}\|^2_{(i)}\Big]$$

$$\leq (1-\tau_k)\mathbb{E}(\hat{F}^k_{\beta_k} - F^\star) + \mathbb{E}\Big[\sum_{i=1}^n \frac{\tau_k^2 B_i^k}{2q_i\tau_0}\|x_i^\star - \tilde{x}_i^k\|^2_{(i)}\Big]$$

- Choose τ_k so that $\tau_k^2 B_i^k \leq (1-\tau_k)\tau_{k-1} B_i^{k-1}$
- Show that $\tau_k \in O(1/k)$, $\beta_k \in O(1/k)$ and $\frac{\tau_k^2}{\beta_k} \in O(1/k)$

Restart

Like accelerated gradient methods, one can restart SMART-CD

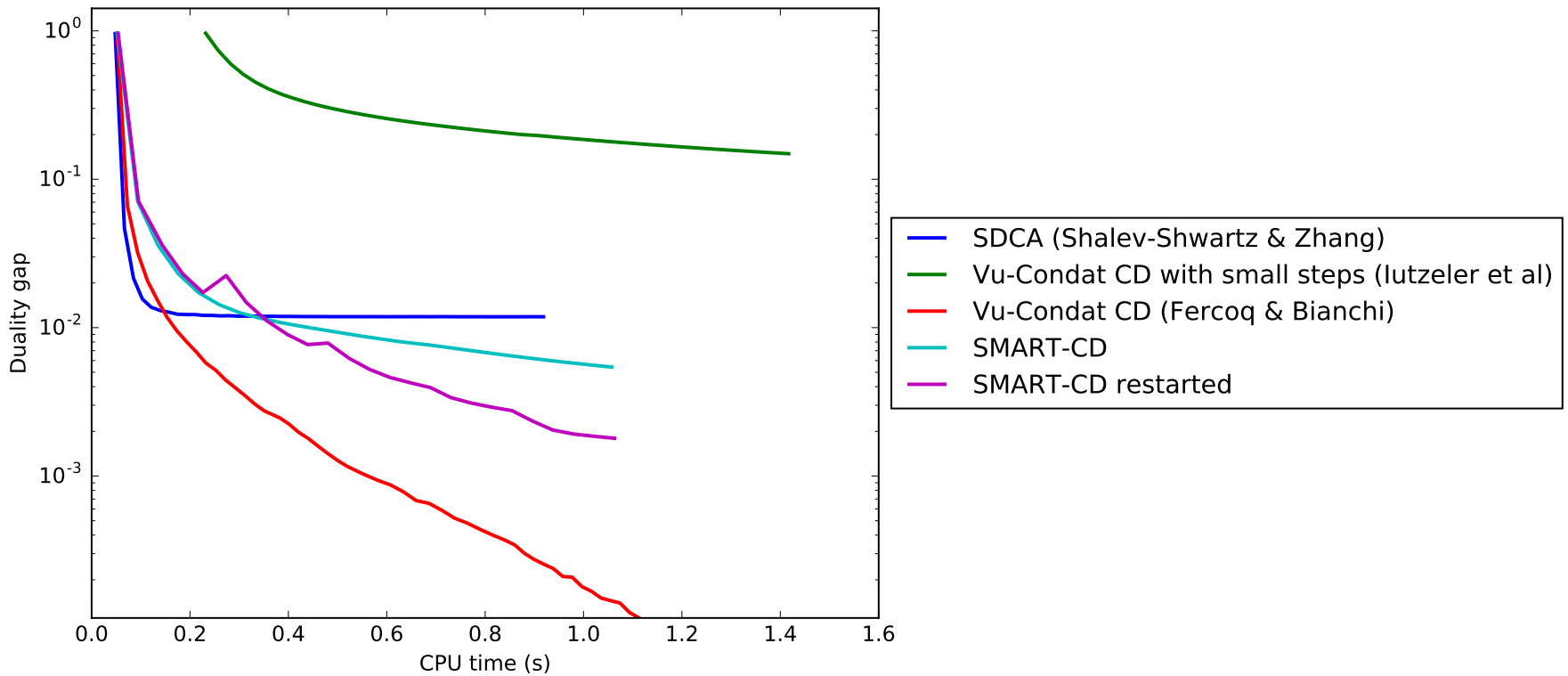
$$\begin{cases} \tilde{x}^k & \leftarrow \bar{x}^k, \\ \dot{y} & \leftarrow y^*_{\beta_{k+1}}(\bar{x}^k; \dot{y}), \\ \beta_{k+1} & \leftarrow \beta_1, \\ \tau_k & \leftarrow \tau_0, \end{cases}$$

$\rightarrow$ Analysis in progress

Experiment on Support Vector Machines

SVM problem on RCV1 dataset $m = 20,242$ and $n = 47,236$

$$\min_{\alpha \in \mathbb{R}^n} \frac{1}{2\lambda} \|X^\top D(y)\alpha\|^2 - e^\top \alpha + I_{[0,C]}(\alpha) + I_{y^\perp}(\alpha)$$



Distributed training for logistic regression 1/2

- Data distributed across M machines, $R(w) = \sum_{i=1}^{p} |w_i|$

$$\min_{w \in \mathbb{R}^{p+1}} \sum_{m=1}^{M} \sum_{j=1}^{n_m} \log(1 + \exp(-y_j x_j^\top w)) + R(w)$$

- Equivalent problem

$$\min_{\boldsymbol{w} \in \mathbb{R}^{(p+1)\times M}} \sum_{m=1}^{M} \left(\sum_{j=1}^{n_m} \log(1+\exp(-y_j x_j^\top w_m)) + \frac{1}{M} R(w_m) \right) + I_C(\boldsymbol{w})$$

where $C = \{\boldsymbol{w} \in \mathbb{R}^{(p+1)\times M} \ : \ w_1 = \ldots = w_M\}$

Distributed training for logistic regression 2/2

- master / slave implementation

For $k \in \mathbb{N}$

For $m \in \{1, \dots, M\}$

Select at random $i_m^k \in \{0, \dots, p\}$

$(z^*_{m,i_m^k})^k = \frac{1}{\beta_{k+1}}(\tilde{w}_{m,i_m^k} - (P_C(\tilde{\boldsymbol{w}}))_{i_m^k}) + \frac{c_k}{\beta_{k+1}}(\check{w}_{m,i_m^k} - (P_C(\check{\boldsymbol{w}}))_{i_m^k})$

$B^k_{m,i_m^k} = L(\nabla_{i_m^k} f_m) + \frac{1}{\beta_{k+1}}$

$\tilde{w}^{k+1}_{m,i_k} = \mathrm{prox}_{\frac{\tau_0}{\tau_k B^k_{i_m^k} M} R_{i_m^k}} \left(\tilde{w}^k_{m,i_m^k} - \frac{\tau_0}{\tau_k B^k_{m,i_m^k}}(\nabla_{i_m^k} f_m(\hat{w}^k_m) + (z^*_{m,i_m^k})^k)\right)$

$\check{w}^{k+1}_{m,i_k} = \check{w}^k_{m,i_k} - \frac{1-\tau_k/\tau_0}{c_k}(\tilde{w}^{k+1}_{m,i_k} - \tilde{w}^k_{m,i_k})$

Update $P_C(\tilde{\boldsymbol{w}}) = \frac{1}{M}\sum_{m=1}^M \tilde{w}_m$ and $P_C(\check{\boldsymbol{w}}) = \frac{1}{M}\sum_{m=1}^M \check{w}_m$

$\tau_{k+1} \in (0,1)$ solution to $\tau^3 + \tau^2 + \tau_k^2 \tau - \tau_k^2 = 0$

$\beta_{k+2} = \frac{\beta_{k+1}}{1+\tau_{k+1}}$

Distributed training for logistic regression 2/2

- master / slave implementation – communicated values

For $k \in \mathbb{N}$

For $m \in \{1, \dots, M\}$

Select at random $i_m^k \in \{0, \dots, p\}$

$(z^*_{m,i_m^k})^k = \frac{1}{\beta_{k+1}}(\tilde{w}_{m,i_m^k} - (P_C(\tilde{\boldsymbol{w}}))_{i_m^k}) + \frac{c_k}{\beta_{k+1}}(\check{w}_{m,i_m^k} - (P_C(\check{\boldsymbol{w}}))_{i_m^k})$

$B^k_{m,i_m^k} = L(\nabla_{i_m^k} f_m) + \frac{1}{\beta_{k+1}}$

$\tilde{w}^{k+1}_{m,i_k} = \mathrm{prox}_{\frac{\tau_0}{\tau_k B^k_{i_m^k} M} R_{i_m^k}} \left(\tilde{w}^k_{m,i_m^k} - \frac{\tau_0}{\tau_k B^k_{m,i_m^k}}(\nabla_{i_m^k} f_m(\hat{w}^k_m) + (z^*_{m,i_m^k})^k)\right)$

$\check{w}^{k+1}_{m,i_k} = \check{w}^k_{m,i_k} - \frac{1-\tau_k/\tau_0}{c_k}(\tilde{w}^{k+1}_{m,i_k} - \tilde{w}^k_{m,i_k})$

Update $P_C(\tilde{\boldsymbol{w}}) = \frac{1}{M}\sum_{m=1}^M \tilde{w}_m$ and $P_C(\check{\boldsymbol{w}}) = \frac{1}{M}\sum_{m=1}^M \check{w}_m$

$\tau_{k+1} \in (0,1)$ solution to $\tau^3 + \tau^2 + \tau_k^2 \tau - \tau_k^2 = 0$

$\beta_{k+2} = \frac{\beta_{k+1}}{1+\tau_{k+1}}$

Conclusion

Summary

- Designed a primal-dual coordinate descent method
- Rate in $O(n/k)$
- Good performance in practice

Future work

- Understand better the restart scheme
- Speed of convergence of Vu-Condat CD
- Smooth gap reduction schemes based on Frank-Wolfe and the stochastic gradient method