An Improved Convergence Analysis of Cyclic Block Coordinate Descent-type Methods for Strongly Convex Minimization

Xingguo Li*, Tuo Zhao◇, Raman Arora◇, Han Liu†, Mingyi Hong‡

*University of Minnesota ◇Johns Hopkins University †Princeton University ‡Iowa State University

Background

A class of strongly convex minimization problems:

$$\text{P1}: \quad \min_{x \in \mathbb{R}^d} \mathcal{L}(x) + \mathcal{R}(x),$$

- A partition of p blocks: $x = [x_1, ..., x_p]^\top$;
- $\mathcal{L}(x)$: differentiable and convex;
- $\mathcal{R}(x) = \sum_{j=1}^{p} \mathcal{R}_j(x_j)$: strongly convex and possibly nonsmooth for each $\mathcal{R}_j(\cdot)$.

Popular examples:

- Elastic-net Penalized Regression:

$$\min_x \frac{1}{2n}||b - Ax||^2 + \lambda_1 ||x||^2 + \lambda_2 ||x||_1;$$

- Ridge Penalized Logistic Regression:

$$\min_x \frac{1}{n} \sum_{i=1}^{n} \log(1 + \exp[-b_i x^\top \underbrace{A_{i*}}_{i\text{-th row of } \mathbf{A}}]) + \lambda ||x||^2;$$

- Support Vector Machines:

$$\min_x \frac{1}{2n} \sum_{i=1}^{n} \max\{1 - b_i(x^\top A_{i*} - y), 0\}^2 + \lambda ||x||^2.$$

Popular Solvers for model P1:

- Gradient decent methods;
- Alternating direction method of multipliers;
- Cyclic block coordinate descent methods.

Algorithms of Our Interests for model P1:

(1) **C**yclic **B**lock **C**oordinate **M**inimization;

(2) **C**yclic **B**lock **C**oordinate **G**radient **D**escent.

Algorithm

Define two auxiliary variables:

$$x_{1:(j-1)}^{(t+1)} = [x_1^{(t+1)\top}, \ldots, x_{j-1}^{(t+1)\top}]^\top,$$
$$x_{(j+1):p}^{(t)} = [x_{j+1}^{(t)\top}, \ldots, x_p^{(t)\top}]^\top.$$

At $t+1$-th iteration, we take

(1) CBCM: For all $j = 1, \ldots, p$,

$$x_j^{(t+1)} = \operatorname*{argmin}_{x_j} \mathcal{L}\left(x_{1:(j-1)}^{(t+1)}, x_j, x_{(j+1):p}^{(t)}\right) + \mathcal{R}_j(x_j);$$

(2) CBCGD: For all $j = 1, \ldots, p$,

$$x_j^{(t+1)} = \operatorname*{argmin}_{x_j} \nabla_j \mathcal{L}\left(x_{1:(j-1)}^{(t+1)}, x_j, x_{(j+1):p}^{(t)}\right)^\top (x_j - x_j^{(t)}) + \frac{\eta_j}{2}||x_j - x_j^{(t)}||^2 + \mathcal{R}_j(x_j),$$

$\eta_j > 0$: the step-size parameter for updating x_j.

Assumptions

Assumption 1. $\nabla \mathcal{L}(\cdot)$ is **Lipschitz continuous** and **blockwise Lipschitz continuous**, i.e., $\exists\ L$ and L_j's such that $\forall\ x, x' \in \mathbb{R}^d$ and $\forall\ j = 1, \ldots, p$, we have

$$||\nabla \mathcal{L}(x') - \nabla \mathcal{L}(x)|| \leq L||x - x'||,$$
$$||\nabla_j \mathcal{L}\left(x_{1:(j-1)}, x'_j, x_{(j+1):p}\right) - \nabla_j \mathcal{L}(x)|| \leq L_j ||x_j - x'_j||.$$

Assumption 2. $\mathcal{R}(\cdot)$ is **strongly convex** and **blockwise strongly convex**, i.e., $\exists\ \mu$ and μ_j's such that $\forall\ x, x' \in \mathbb{R}^d$ and $\forall\ j = 1, \ldots, p$, we have

$$\mathcal{R}(x) \geq \mathcal{R}(x') + (x - x')^\top \xi' + \frac{\mu}{2}||x - x'||^2,$$
$$\mathcal{R}_j(x_j) \geq \mathcal{R}_j(x'_j) + (x_j - x'_j)^\top \xi'_j + \frac{\mu_j}{2}||x_j - x'_j||^2.$$

Main Results

Table 1. Comparison between Our Results and Beck & Tetruashvili [1].

	Method	$\mathcal{L}(\cdot)$	$\mathcal{R}(\cdot)$	Improved Iteration Complexity	Beck & Tetruashvili [1]
[a]	CBCGD	Quadratic	Smooth	$\mathcal{O}\left(\mu^{-1} \log^2 p L^2 \log(1/\epsilon)\right)$	$\mathcal{O}\left(\mu^{-1} p L^2 \log(1/\epsilon)\right)$
[b]	CBCGD	Quadratic	Nonsmooth	$\mathcal{O}\left(\mu^{-1} \log^2 p L^2 \log(1/\epsilon)\right)$	N/A
[c]	CBCGD	General Convex	Smooth	$\mathcal{O}\left(\mu^{-1} p \cdot \min\{L^2, p\} \log(1/\epsilon)\right)$	$\mathcal{O}\left(\mu^{-1} p L^2 \log(1/\epsilon)\right)$
[d]	CBCGD	General Convex	Nonsmooth	$\mathcal{O}\left(\mu^{-1} p L^2 \log(1/\epsilon)\right)$	N/A
[e]	CBCM	Quadratic	Smooth	$\mathcal{O}\left(\mu^{-1} \log^2 p L^2 \log(1/\epsilon)\right)$	N/A
[f]	CBCM	Quadratic	Nonsmooth	$\mathcal{O}\left(\mu^{-1} \log^2 p L^2 \log(1/\epsilon)\right)$	N/A
[g]	CBCM	General Convex	Smooth	$\mathcal{O}\left(\mu^{-1} p L^2 \log(1/\epsilon)\right)$	N/A
[h]	CBCM	General Convex	Nonsmooth	$\mathcal{O}\left(\mu^{-1} p L^2 \log(1/\epsilon)\right)$	N/A

Our main contributions:

(1) **Develop** the iteration bounds of CBCM and CBCGD for different specifications on $\mathcal{L}(\cdot)$ and $\mathcal{R}(\cdot)$;

(2) **Significantly improve** the dependence on dimension p of the iteration bound of CBCGD for quadratic $\mathcal{L}(\cdot)$;

(3) **Improve** the iteration bound of CBCGD for smooth $\mathcal{R}(\cdot)$.

Proof Sketch for Quadratic $\mathcal{L}(\cdot)$ via CBCGD

Let $\mathcal{F}(x) = \mathcal{L}(x) + \mathcal{R}(x)$, $L_{\min}^{\mu} = \min_j L_j + \mu_j$, $x^* = \operatorname{argmin}_x \mathcal{F}(x)$, $d_{\max} = \max_j \{d_j : x_j \in \mathbb{R}^{d_j}\}$, ϵ be a pre-specified accuracy of the objective value:

(1) Characterize the successive descent after each CBCGD iteration:

$$\mathcal{F}(x^{(t)}) - \mathcal{F}(x^{(t+1)}) \geq \frac{L_{\min}^{\mu}}{2}||x^{(t)} - x^{(t+1)}||^2;$$

(2) Characterize the gap towards the optimal objective value after each CBCGD iteration:

$$\mathcal{F}(x^{(t+1)}) - \mathcal{F}(x^*) \leq \frac{L^2 \log^2(2p \cdot d_{\max})}{2\mu}||x^{(t+1)} - x^{(t)}||^2;$$

A new proof technique: Symmetrification.

(3) Combine (1) and (2): To guarantee $\mathcal{F}(x^{(t)}) - \mathcal{F}(x^*) \leq \epsilon$,

$$\#\text{ of iterations} \leq \left\lceil \frac{\mu L_{\min}^{\mu} + L^2 \log^2(2p \cdot d_{\max})}{\mu L_{\min}^{\mu}} \log\left(\frac{\mathcal{F}(x^{(0)}) - \mathcal{F}(x^*)}{\epsilon}\right) \right\rceil.$$

Tight up to a $\log^2 p$ factor.

Tightness of the Iteration Complexity for Quadratic $\mathcal{L}(\cdot)$

Consider: $\min_x \mathcal{H}(x) := ||Bx||^2$, $x^* = \operatorname{argmin} \mathcal{H}(x) = [0, 0, \ldots, 0]^\top$. Let

$$B = \begin{bmatrix} 3 & 1 & 0 & 0 & \ldots & 0 & 0 & 0 \\ 1 & 3 & 1 & 0 & \ldots & 0 & 0 & 0 \\ 0 & 1 & 3 & 1 & \ldots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \ldots & 0 & 3 & 1 \\ 0 & 0 & 0 & 0 & \ldots & 0 & 1 & 3 \end{bmatrix}, \quad L = \overbrace{\lambda_{\max}(2B^\top B)}^{\text{Constant}} \leq 50, \quad \mu = \underbrace{\lambda_{\min}(2B^\top B)}_{\text{Constant}} \geq 2$$

Choose $x^{(0)} = \left[1, \frac{9}{32}, \frac{7}{8}, 1, \ldots, 1, 1\right]^\top \implies x^{(1)} = \left[-\frac{1}{2}, -\frac{1}{2}, \ldots, -\frac{1}{2}, -\frac{3}{10}, -\frac{17}{40}\right]^\top$

$$\implies \mathcal{H}(x^{(1)}) - \mathcal{H}(x^*) \geq \frac{25}{4}(p - 3)$$

$$\mathcal{H}(x^{(0)}) - \mathcal{H}(x^*) \leq \frac{L}{2}||x^{(0)} - x^*||^2 \leq 25\left(p - 2 + \left(\frac{9}{32}\right)^2 + \left(\frac{7}{8}\right)^2\right) \leq 25(p - 1)$$

$$\implies \frac{\mathcal{H}(x^{(1)}) - \mathcal{H}(x^*)}{\mathcal{H}(x^{(0)}) - \mathcal{H}(x^*)} \geq \frac{25(p-3)}{100p} \geq \underbrace{\frac{11}{50}}_{\text{Lower Bound}}$$

Iteration Complexity is independent of p and **cannot be further improved** when $\lambda_{\max}$ and $\lambda_{\min}$ of Hessian **do not scale with** p.

References

[1] A. Beck and L. Tetruashvili. On the convergence of block coordinate descent type methods. *SIAM Journal on Optimization*, 23(4):2037–2060, 2013.

University of Minnesota | Johns Hopkins University | Princeton University | Iowa State University