Background
A class of strongly convex minimization problems:
\[ P_1: \min_{x \in \mathbb{R}^d} \ell(x) + R(x). \]
- A partition of \( p \) blocks: \( x = [x_1, \ldots, x_p]^T \);
- \( L(x) \): differentiable and convex;
- \( R(x) = \sum_{j=1}^p R_j(x_j) \): strongly convex and possibly nonsmooth for each \( R_j(\cdot) \).

Popular examples:
- Elastic-net Penalized Regression:
  \[ \min_{x \in \mathbb{R}^d} \frac{1}{2n} \sum_{i=1}^n \log(1 + \exp[-b_i^T A_i x]) + \lambda_1 |x| + \lambda_2 |x|^2; \]
- Ridge Penalized Logistic Regression:
  \[ \min_{x \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^n \log(1 + \exp[-b_i^T A_i x]) + \lambda |x|^2; \]
- Support Vector Machines:
  \[ \min_{x \in \mathbb{R}^d} \frac{1}{2n} \sum_{i=1}^n \max\{1 - b_i^T A_i x, 0\|^2 + \lambda |x|^2. \]

Popular solvers for model \( P_1 \):
- Gradient decent methods;
- Alternating direction method of multipliers;
- Cyclic block coordinate descent methods.

Algorithms of Our Interests for model \( P_1 \):
1. *Cyclic Block Coordinate Minimization;*
2. *Cyclic Block Coordinate Gradient Descent.*

Algorithm
Define two auxiliary variables:
\[ x_j^{(t+1)} = \left[ x_1^{(t+1)} \cdot \cdots \cdot x_{j-1}^{(t+1)} \cdot x_j + \mu L \right] \]
\[ x_j^{(t)} = \left[ x_1^{(t)} \cdot \cdots \cdot x_{j-1}^{(t)} \cdot x_j \right] \]

At \( t+1 \)-th iteration, we take
1. CBCM: For all \( j = 1, \ldots, p \),
   \[ x_j^{(t+1)} = \text{argmin}_{x_j} \ell\left(x_j^{(t+1)}\right) + R_j(x_j). \]
2. CBCGD: For all \( j = 1, \ldots, p \),
   \[ x_j^{(t+1)} = \text{argmin}_{x_j} \nabla \ell\left(x_j^{(t+1)}\right) + R_j(x_j). \]

Assumptions
Assumption 1. \( \nabla \ell(\cdot) \) 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 \( j = 1, \ldots, p \), we have
\[ \| \nabla \ell(x) - \nabla \ell(x') \| \leq L_\ell \| x - x' \|, \]
\[ \| \nabla \ell(x_j) - \nabla \ell(x_j') \| \leq L_j \| x_j - x_j' \|. \]

Assumption 2. \( \mathcal{R}(\cdot) \) is strongly convex and blockwise strongly convex, i.e., \( \exists \mu \) and \( \mu_j \) ‘s such that \( \forall x, x' \in \mathbb{R}^d \) and \( j = 1, \ldots, p \), we have
\[ \mathcal{R}(x) \geq \mathcal{R}(x') + (x - x')^T \ell'(x') + \frac{\mu}{2} \| x - x' \|^2, \]
\[ \mathcal{R}_j(x_j) \geq \mathcal{R}_j(x_j') + (x_j - x_j')^T \ell_j'(x_j') + \frac{\mu_j}{2} \| x_j - x_j' \|^2. \]

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

<table>
<thead>
<tr>
<th>Method</th>
<th>( \mathcal{L}(\cdot) )</th>
<th>( \mathcal{R}(\cdot) )</th>
<th>Improved Iteration Complexity</th>
<th>Beck &amp; Tetruashvili [1]</th>
</tr>
</thead>
<tbody>
<tr>
<td>[a] CBCGD</td>
<td>Quadratic</td>
<td>Smooth</td>
<td>( \mathcal{O}(\mu^{-1} \log^2 \lambda_2 L^2 \log(1/\epsilon)) )</td>
<td>( \mathcal{O}(\mu^{-1} \log^2 (1/\epsilon)) )</td>
</tr>
<tr>
<td>[b] CBCGD</td>
<td>Quadratic</td>
<td>Nonsmooth</td>
<td>( \mathcal{O}(\mu^{-1} \log^2 (1/\epsilon)) )</td>
<td>N/A</td>
</tr>
<tr>
<td>[c] CBCGD</td>
<td>General Convex</td>
<td>Smooth</td>
<td>( \mathcal{O}(\mu^{-1} p \cdot \min(L_j L_j') \log(1/\epsilon)) )</td>
<td>( \mathcal{O}(\mu^{-1} \log^2 (1/\epsilon)) )</td>
</tr>
<tr>
<td>[d] CBCGD</td>
<td>General Convex</td>
<td>Nonsmooth</td>
<td>( \mathcal{O}(\mu^{-1} p \log^2 (1/\epsilon)) )</td>
<td>N/A</td>
</tr>
<tr>
<td>[e] CBCM</td>
<td>Quadratic</td>
<td>Smooth</td>
<td>( \mathcal{O}(\mu^{-1} \log^2 L^2 \log(1/\epsilon)) )</td>
<td>N/A</td>
</tr>
<tr>
<td>[f] CBCM</td>
<td>Quadratic</td>
<td>Nonsmooth</td>
<td>( \mathcal{O}(\mu^{-1} \log^2 L^2 \log(1/\epsilon)) )</td>
<td>N/A</td>
</tr>
<tr>
<td>[g] CBCM</td>
<td>General Convex</td>
<td>Smooth</td>
<td>( \mathcal{O}(\mu^{-1} p \log^2 (1/\epsilon)) )</td>
<td>N/A</td>
</tr>
<tr>
<td>[h] CBCM</td>
<td>General Convex</td>
<td>Nonsmooth</td>
<td>( \mathcal{O}(\mu^{-1} p \log^2 (1/\epsilon)) )</td>
<td>N/A</td>
</tr>
</tbody>
</table>

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 \( F(x) = \mathcal{L}(x) + R(x) \), \( L_{\min} = \min_j L_j + \mu_j\), \( x^* = \arg\min_x F(x) \), \( d_{\max} = \max_j \{|d_j| : x_j \in \mathbb{R}^d\} \), \( \epsilon \) be a pre-specified accuracy of the objective value:

1. Characterize the successive descent after each CBCGD iteration:
   \[ F(x^{(t)}) - F(x^{(t+1)}) \leq \frac{\mu L}{2 \mu L_{\min}} \| x^{(t)} - x^{(t+1)} \|^2; \]
2. Characterize the gap towards the optimal value after each CBCGD iteration:
   \[ F(x^{(t)}) - F(x^*) \leq \frac{L_{\min}^2}{\mu^2} \log \left( \frac{2 \mu L_{\min}^2}{\mu^2} \| x^{(0)} - x^* \|^2 \right) \]

A new proof technique: Symmetrization

3. Combine (1) and (2): To guarantee \( F(x^{(t)}) - F(x^*) \leq \epsilon \),
   \[ \# \text{ of iterations} \leq \frac{\mu L_{\min}^2 + L_{\min} \log^2 (2 \mu d_{\max})}{\mu^2} \log \left( \frac{F(x^{(0)}) - F(x^*)}{\epsilon} \right). \]

Tightness of the Iteration Complexity for Quadratic \( \mathcal{L}(\cdot) \)
Consider: \( \mathcal{H}(x) := \| Bx \|^2, x^* = \arg\min_x \mathcal{H}(x) = [0, 0, \ldots, 0]^T \).

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

\[ \Rightarrow \mathcal{H}(x^{(1)}) - \mathcal{H}(x^*) \geq \frac{25}{4} (p - 3) \]

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

\[ \Rightarrow \mathcal{H}(x^{(1)}) - \mathcal{H}(x^*) \geq \frac{25 (p - 3)}{100} \geq \frac{11}{50} \]

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