Iteration Complexity Analysis of Block Coordinate Descent Method

Mingyi Hong

IMSE and ECE Department,
Iowa State University

INFORMS 2015
**Question**: What is the iteration complexity of the BCD (with deterministic update rules) for convex problems?

**Answer**: Scales sublinearly as $O(1/r)$ [H.-Wang-Razaviyayn-Luo 14]

1. Covers popular algorithms like BCPG, BCM, etc
2. Covers popular block selection rule like cyclic, Gauss-Southwell, Essentially cyclic
3. Does not require per-block strong convexity
The Main Content of the Talk

- **Question**: How does the rate depend on the problem dimension?
- **Answer**: Scales (almost) independently, linearly, etc, requires case-by-case study [Sun-H. 15]
Outline

1 Introduction of BCD
   - The Algorithm and Applications
   - The Prior Art

2 Analyzing the BCD-type Algorithm
   - The BCPG and its iteration complexity analysis
   - The BCM and its complexity analysis

3 Sharpening the Bounds on $K$
The Problem

- We consider the following problem (with $K$ block variables)

$$
\begin{align*}
\text{minimize} & \quad f(x) := g(x_1, \ldots, x_K) + \sum_{k=1}^{K} h_k(x_k) \\
\text{subject to} & \quad x_k \in X_k, \quad k = 1, \ldots, K
\end{align*}
$$

(P)

- $g(\cdot)$ smooth convex function; $h_k(\cdot)$ nonsmooth convex function;
- $x = (x_1^T, \ldots, x_K^T)^T \in \mathbb{R}^n$ is a partition of $x$; $X_k \subseteq \mathbb{R}^{n_k}$
Applications

• Lots of applications in practice

• One of the most well-known application is the LASSO problem

\[
\min_x \frac{1}{2} \|Ax - b\|^2 + \lambda \|x\|_1
\]

• Each scalar \(x_k \in \mathbb{R}\) is a block variable

\[
\min_x \frac{1}{2} \left\| \sum_{k=1}^{K} A_k x_k - b \right\|^2 + \lambda \sum_{k=1}^{K} |x_k|
\]
Rate maximization in uplink wireless communication network

$K$ users, a single base station (BS)

Each user has $n_t$ ($n_r$) transmit (receive) antennas

Let $C_k \in \mathbb{R}^{n_t \times n_t}$ denote user $k$’s transmit covariance matrix

$H_k \in \mathbb{R}^{n_r \times n_t}$ the channel matrix between user $k$ and the BS

Then the uplink channel capacity optimization problem is

$$\min_{\{C_k\}_{k=1}^K} - \log \det \left| \sum_{k=1}^K H_k C_k H_k^T + I_{n_r} \right|$$

s.t. $C_k \succeq 0$, $\text{Tr}[C_k] \leq P_k$, $k = 1, \cdots, K$

The celebrated iterative water-filling algorithm (IWFA) [Yu-Cioffi 04] is simply BCD with cyclic update rule
Let us assume that the gradient of $g(\cdot)$ is block-wise Lipschitz continuous

$$\|\nabla_k g([x_{-k}, x_k]) - \nabla_k g([x_{-k}, \hat{x}_k])\| \leq M_k \|x_k - \hat{x}_k\|, \quad \forall \ x \in X, \ \forall \ k$$

$$\|\nabla g(x) - \nabla g(z)\| \leq M \|x - z\|, \quad \forall \ x, z \in X$$

Let $M_{\text{min}} = \min M_k$, $M_{\text{max}} = \max M_k$
Consider the cyclic block coordinate minimization (BCM)

At iteration $r + 1$, block $k$ updated by

$$x_k^{r+1} \in \arg \min_{x_k \in X_k} g(x_1^{r+1}, \ldots, x_{k-1}^{r+1}, x_k, x_{k+1}^{r}, \ldots, x_K^{r}) + h_k(x_k)$$

Sweep over all blocks in cyclic order (a.k.a Gauss-Seidel rule)

Popular for solving modern large-scale problems
Variants: Block Selection Rule

- Lots of block selection rules
- Cyclic, randomized, parallel, greedy (Gauss-Southwell), randomly permutated
- Interested in analyzing deterministic rules
  - Provides worse case analysis
  - Sheds lights on the randomly permutated variants
Variants: Block Update Rule

- **Block coordinate proximal gradient (BCPG)**
  
  At iteration $r + 1$, block $k$ updated by
  
  $$
  x_{k}^{r+1} = \arg\min_{x_k \in X_k} u_k(x_k; x_1^{r+1}, \ldots, x_{k-1}^{r+1}, x_k^r, \ldots, x_K^r) + h_k(x_k)
  $$

- $u_k(x_k; y)$: a quadratic approximation of $g(\cdot)$ w.r.t. $x_k$
  
  $$
  u_k(x_k; y) = g(y) + \langle \nabla_k g(y), x_k - y_k \rangle + \frac{L_k}{2} \| x_k - y_k \|^2.
  $$

- $L_k$ is the penalty parameter; $1/L_k$ is the stepsize
Prior Art: BCPG

- A large literature on analyzing BCPG-type algorithm
- **Randomized BCPG**: blocks picked randomly
- Strongly convex problems, **linear rate** [Richtárik-Takáč 12]
- Convex problems with $\mathcal{O}(1/r)$ rate
  1. Smooth [Nesterov 12]
  2. Smooth + L1 penalization [Shalev-Shwartz-Tewari 11]
  3. General nonsmooth (P) [Richtárik-Takáč 12][Lu-Xiao 13]
Prior Art: BCPG (cont.)

- How about cyclic BCPG? Less is known
- For strongly convex problems, linear rate
- For general convex problems, $O(1/r)$ rate?
  1. LASSO, when satisfies the so-called “isotonicity assumption” (assumption on the data matrix) [Saha-Tewari 13]
  2. Smooth [Beck-Tetruashvili 13]
  3. General nonsmooth, i.e., (P)?
Prior Art: BCM

- How about cyclic BCM? Even less is known
- For strongly convex problems, linear rate
- For convex problem \( O(1/r) \) rate
  1. Smooth unconstrained problem with \( K = 2 \) (two-block variables) [Beck-Tetruashvili 13]
  2. Other cases?
- Other coordinate update rules (e.g., Gauss-Southwell)?
The Summary of Results

- A summary of existing results on sublinear rate for BCD-type
- \( NS = \text{NonSmooth}, \ S = \text{Smooth}, \ C = \text{Constrained}, \ U = \text{Unconstrained}, \ K = \text{K-block}, \ 2 = \text{2-Block} \)
- \( GS = \text{Gauss-Seidel}, \ GSo = \text{Gauss-Southwell}, \ EC = \text{Essentially-Cyclic} \)

**Table:** Summary of Prior Art

<table>
<thead>
<tr>
<th>Method</th>
<th>Problem</th>
<th>( O(1/r) ) Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>GS-BCPG</td>
<td>S-C-K</td>
<td>√</td>
</tr>
<tr>
<td>GS/GSo/EC-BCPG</td>
<td>NS-C-K</td>
<td>?</td>
</tr>
<tr>
<td>GS-BCM</td>
<td>S-U-2</td>
<td>√</td>
</tr>
<tr>
<td>GS/EC-BCM</td>
<td>NS-C-K</td>
<td>?</td>
</tr>
</tbody>
</table>
This Work

- This work shows the following results [H.-Wang-Razaviyayn-Luo 14]
- NS=NonSmooth, S=Smooth, C=Constrained, U=Unconstrained, K=K-block, 2=2-Block
- GS=Gauss-Seidel, GSo=Gauss-Southwell, EC=Essentially-Cyclic

Table: Summary of Prior Art + This Work

<table>
<thead>
<tr>
<th>Method</th>
<th>Problem</th>
<th>$O(1/r)$ Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>GS-BCPG</td>
<td>S-C-K</td>
<td>✓</td>
</tr>
<tr>
<td>GS/GSo/EC-BCPG</td>
<td>NS-C-K</td>
<td>✓</td>
</tr>
<tr>
<td>GS-BCM</td>
<td>S-U-2</td>
<td>✓</td>
</tr>
<tr>
<td>GS/EC-BCM</td>
<td>NS-C-K</td>
<td>✓</td>
</tr>
</tbody>
</table>
Outline

1. Introduction of BCD
   - The Algorithm and Applications
   - The Prior Art

2. Analyzing the BCD-type Algorithm
   - The BCPG and its iteration complexity analysis
   - The BCM and its complexity analysis

3. Sharpening the Bounds on $K$
Iteration Complexity Analysis for BCPG

Define $X^*$ as the optimal solution set, and let $x^* \in X^*$ be one of the optimal solutions.

Define the optimality gap as

$$\Delta^r := f(x^r) - f(x^*)$$

Main proof steps

1. **Sufficient Descent**: $\Delta^r - \Delta^{r+1}$ large enough
2. **Estimate Cost-To-Go**: $\Delta^{r+1}$ small enough
3. **Estimate The Rate**: Show $(\Delta^{r+1})^2 \leq c (\Delta^r - \Delta^{r+1})$
Bound for BCPG

- Define $R := \max_{x \in X} \max_{x^* \in X^*} \{ \|x - x^*\| : f(x) \leq f(x^1) \}$

- For BCPG, pick $L_k = M_k$ for all $k$, the bound final bound is

$$\Delta^r \leq 8 \frac{cKM}{\min_k M_k} \frac{MR^2}{r}$$

- The bound is in the same order as the one stated in [Beck-Tetruashvili 13, Theorem 6.1]

- In the worst case $M/M_{\min} = O(K)$, so the red part scales with $K^2$
Remark

- The analysis extends to other popular update rules
  1. Gauss-Southwell (i.e., greedy coordinate selection)
  2. Maximum Block Improvement (MBI) [Chen-Li-He-Zhang 13]
  3. Essentially cyclic
  4. Random permutation

- Same analysis for GS-BCM with per-block strongly convexity (BSC)

- **Key challenge.** Complexity without BSC?
Motivating Examples

- **Example 1.** Consider the group-LASSO problem

\[
\min_x \left\| \sum_{k=1}^{K} A_k x_k - b \right\|^2 + \lambda \sum_{k=1}^{K} \| x_k \|_2
\]

- $x_k$ subproblem (semi)closed-form solution; $A_k$’s can be rank-deficient

- **Example 2.** Consider a rate maximization problem in wireless networks ($K$ user, multiple antenna, etc)

\[
\min_{\{C_k\}_{k=1}^{K}} - \log \det \left| \sum_{k=1}^{K} H_k C_k H_k^T + I_{n_r} \right|, \quad \text{s.t.} \quad C_k \succeq 0, \ \text{Tr}[C_k] \leq P_k, \ \forall \ k
\]

- If $H_k$ is not full row rank, each $C_k$ subproblem is not strongly convex

Our previous results do not apply!
Iteration Complexity for BCM?

- **The Algorithm.** At iteration $r + 1$, update:

  $$
  x_k^{r+1} \in \min_{x_k \in X_k} g \left( x_1^{r+1}, \ldots, x_{k-1}^{r+1}, x_k, x_{k+1}^r, \ldots, x_K^r \right) + h_k(x_k)
  $$

- **Key challenges.**
  1. No BSC anymore
  2. Multiple optimal solutions
  3. The sufficient descent estimate is lost
Rate Analysis for GS-BCM (no BSC)

- **Key idea.** Using a different measure to gauge progress.

- **Key steps:** Still three-step approach

  - Sufficient descent
    \[
    \Delta^r - \Delta^{r+1} \geq \frac{1}{2M} \sum_{k=1}^{K} \| \nabla g(w_{k+1}^{r+1}) - \nabla g(w_{k+1}^{r+1}) \|^2.
    \]

  where
  \[
  w_{k+1}^{r+1} := [x_1^{r+1}, \ldots, x_{k-1}^{r+1}, x_k^r, x_{k+1}^r, \ldots, x_K^r].
  \]

  - Cost-to-go estimate
    \[
    (\Delta^{r+1})^2 \leq 2K^2R^2 \sum_{k=1}^{K} \| \nabla g(w_{k+1}^{r+1}) - \nabla g(w_{k+1}^{r+1}) \|^2, \forall x^* \in X^*.
    \]

  - Matching the previous two and obtain...
Rate Analysis for cyclic BSUM (no BSC)

Theorem

(H.-Wang-Razaviyayn-Luo 14) Let \( \{x^r\} \) be the sequence generated by the BCM algorithm with G-S rule. Then we have

\[
\Delta^r = f(x^r) - f^* \leq \frac{c_5}{\sigma_5 r}, \quad \forall \ r \geq 1,
\]

(2.1)

where the constants are given below

\[
c_5 = \max\{4\sigma_5 - 2, f(x^1) - f^*, 2\},
\]

\[
\sigma_5 = \frac{1}{2MK^2R^2},
\]

(2.2)
Outline

1. Introduction of BCD
   - The Algorithm and Applications
   - The Prior Art

2. Analyzing the BCD-type Algorithm
   - The BCPG and its iteration complexity analysis
   - The BCM and its complexity analysis

3. Sharpening the Bounds on $K$
What’s Missing?

- Why we care about iteration complexity?
  - Characterize how fast the algorithms progresses
  - Estimate practical performance for big data problems

- For large scale problems, the scaling with respect to $K$ matters!

- When $K = 10^6$, two algorithms with $\mathcal{O}(K/r)$ and $\mathcal{O}(K^2/r^2)$...
The State-of-the-Art

- The classical gradient method scales $O(M/r)$ (independent of $K$)

- In [Saha-Tewari 13], GS-BCPG for LASSO with “isotonicity assumption” scales $O(M/r)$

- The GS-BCPG for smooth problems [Beck-Tetruashvili 13] scales $O(K^2M/r)$ in the worst case

- The analysis in [H.-Wang-Razaviyayn-Luo 14] scales similarly

- Better bounds?
Sharpening the Bounds on $K$?

- All the rates scale \textit{quadratically} in $K$ (in worst case)

- Next we sharpen the bound for the following quadratic problem

$$\min f(x) := \frac{1}{2} \left\| \sum_{k=1}^{K} A_k x_k - b \right\|^2 + \sum_{k=1}^{K} h_k(x_k), \quad \text{s.t. } x_k \in X_k, \ \forall \ k \quad (Q)$$
Sharpening the Bounds on $K$?

- Consider the BCPG algorithm
- Remove a $K$ factor in the worst case
- Matches the complexity of gradient descent (almost independent of $K$) for some special cases
The Result

**Question:** How does the rate bound depend on $K$?

**Result 1:** BCPG+quadratic $g$. If $L_k = M_k$, then the rate scales

$$O \left( \log^2(K) \left( \frac{M_{\text{max}}}{M} + \frac{M}{M_{\text{min}}} \right) M R^2 \right)$$

**Result 2:** BCPG+quadratic $g$. If $L_k = M$, then the rate scales

$$O \left( \log^2(K) M R^2 \right)$$

**Result 3:** For problems with $\frac{M_{\text{max}}}{M_{\text{min}}} = O(1)$, the above rates are tight.
Rate Analysis for GS-BCPG

Theorem (Sun-H. 15)

The iteration complexity of using BCPG to solve (Q) is given below

1. Suppose the stepsizes are chosen as \( L_k = M \), \( \forall k \), then

\[
\Delta^{(r+1)} \leq 3 \max \left\{ \Delta^0, 4 \log^2 (2K) M \right\} \frac{R^2}{r+1}.
\]

2. Suppose the stepsizes are chosen according to:

\[
L_k = \lambda_{\text{max}}(A_k^T A_k) = M_k, \quad \forall k.
\]

Then we have

\[
\Delta^{(r+1)} \leq 3 \max \left\{ \Delta^0, 2 \log^2 (2K) \left( M_{\text{max}} + \frac{M^2}{M_{\text{min}}} \right) \right\} \frac{R^2}{r+1}
\]
We briefly discuss the tightness of the bound over $K$

**Question:** Can we improve the bounds further in the order of $K$?

Construct a simple quadratic problem

$$\min g(x) := \left\| \sum_{k=1}^{K} A_k x_k \right\|^2$$

with

$$A = \begin{bmatrix}
1 & 1 & 0 & 0 & \cdots & 0 & 0 & 0 \\
1 & 1 & 1 & 0 & \cdots & 0 & 0 & 0 \\
0 & 1 & 1 & 1 & \cdots & 0 & 0 & 0 \\
\vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\
0 & 0 & 0 & 0 & \cdots & 1 & 1 & 1 \\
0 & 0 & 0 & 0 & \cdots & 0 & 1 & 1
\end{bmatrix}.$$
We can show that after running a single pass of GS-BCD/BCPG

\[
\Delta^{(1)} \geq \frac{9(K-3)}{4(K-1)} \|x^{(0)} - x^*\|^2, \quad \forall \ K \geq 3.
\]

Specialized our bound to this problem predicts that the gap is at most

\[
\Delta^{(1)} \leq \left( \frac{M - M_{\text{min}}}{M_{\text{min}}} \frac{1}{2} + \frac{1}{4} \right) \|x^{(0)} - x^*\|^2 M \leq 36 \|x^{(0)} - x^*\|^2.
\]

As \( K \rightarrow \infty \), the previous two bounds match, up to a constant dimensionless factor \( 1/16 \)

Conclusion. The derived bound is tight for the case \( M/M_{\text{min}} = \mathcal{O}(1) \)
Numerical Comparison

- We compare the performance of GD and BCD over the constructed problem.

Figure: Comparison of the gradient method and GS-BCD method to solve the constructed. Left, $K = 100$. Right, $K = 1000$. 

Mingyi Hong (Iowa State University)
The proposed technique also applies to the following scenarios:

- Quadratic strongly convex problems (reduces a $K$ factor)
- General nonlinear convex smooth problems
## Comparisons

<table>
<thead>
<tr>
<th>Lip-constant 1/Stepsize</th>
<th>Diag. Hess. $M_i = M$</th>
<th>Full Hess. $M_i = \frac{M}{K}$</th>
<th>Full Hess. $M_i = \frac{M}{K}$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$L_i = M$</td>
<td>Large step $L_i = \frac{M}{K}$</td>
<td>Small step $L_i = M$</td>
</tr>
<tr>
<td><strong>GD</strong></td>
<td>$M/r$</td>
<td>N/A</td>
<td>$M/r$</td>
</tr>
<tr>
<td><strong>R BCGD</strong></td>
<td>$M/r$</td>
<td>$M/(Kr)$</td>
<td>$M/r$</td>
</tr>
<tr>
<td><strong>GS BCGD</strong></td>
<td>$KM/r$</td>
<td>$K^2M/r$</td>
<td>$KM/r$</td>
</tr>
<tr>
<td><strong>GS BCGD (QP)</strong></td>
<td>$\log^2(2K)M/r$</td>
<td>$\log^2(2K)KM/r$</td>
<td>$\log^2(2K)M/r$</td>
</tr>
</tbody>
</table>

**Table:** Comparison of Various Iteration Complexity Results
Conclusion

- We show the $O(1/r)$ sublinear rate of convergence for a few BCD-type methods.
- We also manage to reduce the dependency of the rates on the problem dimension.
- **Observation.** Conservative stepsize obtains better theoretical rate bound (but worse practical performance).
Future Work

- Still a gap in the rate bound

**Question.** Can the GS-BCD/BCPG matches the bound of GD for general convex $K$-block problems (i.e., independent of problem dimension)? If not, construct an example?

**Question.** Does random permutation help?
Thank You!
Reference


Reference


Reference


