Convergence of a Block Coordinate Descent Method for Nondifferentiable Minimization

Paul Tseng
Presenter: Lei Tang

Department of CSE
Arizona State University

Nov. 7th, 2008

Introduction

- Popular method for minimizing a real-valued continuously differentiable function f of n variables, subject to bound constraint, is (block) coordinate descent (BCD).
- In this work, coordinate descent actually refers to alternating optimization(AO). Each step find the exact minimizer.
- Popular for its efficiency, simplicity and scalability.
- Applied to large-scale SVM, Lasso etc.
- Unfortunately, the convergence of coordinate descent is not clear. Not like steepest descent method.
- In this work, it is shown that if the function satisfy some mild conditions, BCD converges to the stationary point.

Introduction

- Popular method for minimizing a real-valued continuously differentiable function f of n variables, subject to bound constraint, is (block) coordinate descent (BCD).
- In this work, coordinate descent actually refers to alternating optimization(AO). Each step find the exact minimizer.
- Popular for its efficiency, simplicity and scalability.
- Applied to large-scale SVM, Lasso etc.
- Unfortunately, the convergence of coordinate descent is not clear. Not like steepest descent method.
- In this work, it is shown that if the function satisfy some mild conditions, BCD converges to the stationary point.

Questions?

1. Does BCD Converge?
2. Does BCD Converge to the local minimizer?
3. When does BCD converge to the stationary point?
4. What's the convergence rate?

Existing works

- Convergence of coordinate descent method requires typically that f be strictly convex (or quasiconvex and hemivariate) differentiable
- the strict convexity is relaxed to pseudoconvexity, which allows f to have non-unique minimum along coordinate directions.
- If f is not differentiable, the coordinate descent method may get stuck at a nonstationary point even when f is convex.
- However, this method still works when the nondifferentiable part of f is seperable.

$$f(x_1, \cdots, x_N) = f_0(x_1, \cdots, x_N) + \sum_{k=1}^{N} f_k(x_k)$$

where f_k is non-differentiable, each x_k represents one block.

- This work shows that BCD converges to a stationary point if f_0 has certain smoothness property.

Existing works

- Convergence of coordinate descent method requires typically that f be strictly convex (or quasiconvex and hemivariate) differentiable
- the strict convexity is relaxed to pseudoconvexity, which allows f to have non-unique minimum along coordinate directions.
- If f is not differentiable, the coordinate descent method may get stuck at a nonstationary point even when f is convex.
- However, this method still works when the nondifferentiable part of f is seperable.

$$f(x_1, \cdots, x_N) = f_0(x_1, \cdots, x_N) + \sum_{k=1}^{N} f_k(x_k)$$

  where f_k is non-differentiable, each x_k represents one block.
- This work shows that BCD converges to a stationary point if f_0 has certain smoothness property.

An Example of Alternating Optimization

$$\begin{aligned}\phi_1(x,y,z) &= -xy - yz - zx + (x-1)_+^2 + (-x-1)_+^2 + \\ &\quad (y-1)_+^2 + (-y-1)_+^2 + (z-1)_+^2 + (-z-1)_+^2\end{aligned}$$

Note that the optimal x given fixed y and z is

$$x = sign(y+z)\left(1 + \frac{1}{2}|y+z|\right)$$

Suppose you start from $(-1-\epsilon, 1+\frac{1}{2}\epsilon, -1-\frac{1}{4}\epsilon)$:

$$(1+\frac{1}{8}\epsilon, 1+\frac{1}{2}\epsilon, -1-\frac{1}{4}\epsilon)$$
$$(1+\frac{1}{8}\epsilon, -1-\frac{1}{16}\epsilon, -1-\frac{1}{4}\epsilon)$$
$$(1+\frac{1}{8}\epsilon, -1-\frac{1}{16}\epsilon, 1+\frac{1}{32}\epsilon)$$

$$(-1-\frac{1}{64}\epsilon, -1-\frac{1}{16}\epsilon, 1+\frac{1}{32}\epsilon)$$
$$(-1-\frac{1}{64}\epsilon, 1+\frac{1}{128}\epsilon, 1+\frac{1}{32}\epsilon)$$
$$(-1-\frac{1}{64}\epsilon, 1+\frac{1}{128}\epsilon, -1-\frac{1}{256}\epsilon)$$

Cycle around 6 edges of the cube $(\pm1, \pm1, \pm1)$!!

An Example of Alternating Optimization

$$\phi_1(x, y, z) = -xy - yz - zx + (x-1)_+^2 + (-x-1)_+^2 + (y-1)_+^2 + (-y-1)_+^2 + (z-1)_+^2 + (-z-1)_+^2$$

Note that the optimal x given fixed y and z is

$$x = sign(y+z)\left(1 + \frac{1}{2}|y+z|\right)$$

Suppose you start from $(-1-\epsilon, 1+\frac{1}{2}\epsilon, -1-\frac{1}{4}\epsilon)$:

$$(1+\frac{1}{8}\epsilon, 1+\frac{1}{2}\epsilon, -1-\frac{1}{4}\epsilon)$$
$$(1+\frac{1}{8}\epsilon, -1-\frac{1}{16}\epsilon, -1-\frac{1}{4}\epsilon)$$
$$(1+\frac{1}{8}\epsilon, -1-\frac{1}{16}\epsilon, 1+\frac{1}{32}\epsilon)$$

$$(-1-\frac{1}{64}\epsilon, -1-\frac{1}{16}\epsilon, 1+\frac{1}{32}\epsilon)$$
$$(-1-\frac{1}{64}\epsilon, 1+\frac{1}{128}\epsilon, 1+\frac{1}{32}\epsilon)$$
$$(-1-\frac{1}{64}\epsilon, 1+\frac{1}{128}\epsilon, -1-\frac{1}{256}\epsilon)$$

Cycle around 6 edges of the cube $(\pm 1, \pm 1, \pm 1)$!!

An Example of Alternating Optimization

$$\phi_1(x, y, z) = -xy - yz - zx + (x-1)_+^2 + (-x-1)_+^2 + (y-1)_+^2 + (-y-1)_+^2 + (z-1)_+^2 + (-z-1)_+^2$$

Note that the optimal x given fixed y and z is

$$x = sign(y+z)\left(1 + \frac{1}{2}|y+z|\right)$$

Suppose you start from $(-1-\epsilon, 1+\frac{1}{2}\epsilon, -1-\frac{1}{4}\epsilon)$:

$$(1+\frac{1}{8}\epsilon, 1+\frac{1}{2}\epsilon, -1-\frac{1}{4}\epsilon) \qquad (-1-\frac{1}{64}\epsilon, -1-\frac{1}{16}\epsilon, 1+\frac{1}{32}\epsilon)$$

$$(1+\frac{1}{8}\epsilon, -1-\frac{1}{16}\epsilon, -1-\frac{1}{4}\epsilon) \qquad (-1-\frac{1}{64}\epsilon, 1+\frac{1}{128}\epsilon, 1+\frac{1}{32}\epsilon)$$

$$(1+\frac{1}{8}\epsilon, -1-\frac{1}{16}\epsilon, 1+\frac{1}{32}\epsilon) \qquad (-1-\frac{1}{64}\epsilon, 1+\frac{1}{128}\epsilon, -1-\frac{1}{256}\epsilon)$$

Cycle around 6 edges of the cube $(\pm 1, \pm 1, \pm 1)$!!

An Example of Alternating Optimization

$$\begin{aligned}\phi_1(x,y,z) &= -xy - yz - zx + (x-1)_+^2 + (-x-1)_+^2 + \\ &\quad (y-1)_+^2 + (-y-1)_+^2 + (z-1)_+^2 + (-z-1)_+^2\end{aligned}$$

Note that the optimal x given fixed y and z is

$$x = sign(y+z)\left(1 + \frac{1}{2}|y+z|\right)$$

Suppose you start from $(-1-\epsilon, 1+\frac{1}{2}\epsilon, -1-\frac{1}{4}\epsilon)$:

$$(1+\frac{1}{8}\epsilon, 1+\frac{1}{2}\epsilon, -1-\frac{1}{4}\epsilon)$$
$$(1+\frac{1}{8}\epsilon, -1-\frac{1}{16}\epsilon, -1-\frac{1}{4}\epsilon)$$
$$(1+\frac{1}{8}\epsilon, -1-\frac{1}{16}\epsilon, 1+\frac{1}{32}\epsilon)$$

$$(-1-\frac{1}{64}\epsilon, -1-\frac{1}{16}\epsilon, 1+\frac{1}{32}\epsilon)$$
$$(-1-\frac{1}{64}\epsilon, 1+\frac{1}{128}\epsilon, 1+\frac{1}{32}\epsilon)$$
$$(-1-\frac{1}{64}\epsilon, 1+\frac{1}{128}\epsilon, -1-\frac{1}{256}\epsilon)$$

Cycle around 6 edges of the cube $(\pm 1, \pm 1, \pm 1)$!!

An Example of Alternating Optimization

$$\begin{aligned}\phi_1(x,y,z) &= -xy - yz - zx + (x-1)_+^2 + (-x-1)_+^2 + \\ &\quad (y-1)_+^2 + (-y-1)_+^2 + (z-1)_+^2 + (-z-1)_+^2\end{aligned}$$

Note that the optimal x given fixed y and z is

$$x = sign(y+z)\left(1 + \frac{1}{2}|y+z|\right)$$

Suppose you start from $(-1-\epsilon, 1+\frac{1}{2}\epsilon, -1-\frac{1}{4}\epsilon)$:

$$(1+\frac{1}{8}\epsilon, 1+\frac{1}{2}\epsilon, -1-\frac{1}{4}\epsilon)$$
$$(1+\frac{1}{8}\epsilon, -1-\frac{1}{16}\epsilon, -1-\frac{1}{4}\epsilon)$$
$$(1+\frac{1}{8}\epsilon, -1-\frac{1}{16}\epsilon, 1+\frac{1}{32}\epsilon)$$

$$(-1-\frac{1}{64}\epsilon, -1-\frac{1}{16}\epsilon, 1+\frac{1}{32}\epsilon)$$
$$(-1-\frac{1}{64}\epsilon, 1+\frac{1}{128}\epsilon, 1+\frac{1}{32}\epsilon)$$
$$(-1-\frac{1}{64}\epsilon, 1+\frac{1}{128}\epsilon, -1-\frac{1}{256}\epsilon)$$

Cycle around 6 edges of the cube $(\pm 1, \pm 1, \pm 1)$!!

Some Examples

- The gradient in the example is not zero at any $(\pm 1, \pm 1, \pm 1)$.
- The example we show is unstable to perturbations.
- The example has non-smooth 2nd derivatives.
- More complicated examples could be constructed to show that even if the function is infinitely differentiable, stable cyclic behavior still occurs, whose gradient is bounded away from zero in the limiting path.
- Please see *On Search Directions for Minimization Algorithms*, Mathematical Programming, 1974.

Some Examples

- The gradient in the example is not zero at any $(\pm 1, \pm 1, \pm 1)$.
- The example we show is unstable to perturbations.
- The example has non-smooth 2nd derivatives.
- More complicated examples could be constructed to show that even if the function is infinitely differentiable, stable cyclic behavior still occurs, whose gradient is bounded away from zero in the limiting path.
- Please see *On Search Directions for Minimization Algorithms*, Mathematical Programming, 1974.

Some Examples

- The gradient in the example is not zero at any $(\pm 1, \pm 1, \pm 1)$.
- The example we show is unstable to perturbations.
- The example has non-smooth 2nd derivatives.
- More complicated examples could be constructed to show that even if the function is infinitely differentiable, stable cyclic behavior still occurs, whose gradient is bounded away from zero in the limiting path.
- Please see *On Search Directions for Minimization Algorithms*, Mathematical Programming, 1974.

Alternating Optimization Algorithm

AO-1 Let $\Psi_i \subseteq \Re^{p_i}$, for i = 1,…,t, and let $\Psi = \Psi_1 \times \ldots \times \Psi_t$. Partition $x \in \Re^s$ as $x = (X_1, X_2, \ldots, X_t)^T$, with $X_i \in \Re^{p_i}$ for i = 1,…,t, $\bigcup_{i=1}^{t} X_i = X;\ X_i \cap X_i = \varnothing\ \text{for}\ i \neq j;\ \text{and}\ s = \sum_{i=1}^{t} p_i$. Pick an initial iterate $x^{(0)} = (X_1^{(0)}, X_2^{(0)}, \ldots, X_t^{(0)})^T \in \Psi = \Psi_1 \times \ldots \times \Psi_t$, a vector norm $\|\cdot\|$, termination threshold ε, and iteration limit L. Set r = 0.

AO-2 For i = 1,…, t, compute the restricted minimizer

$$X_i^{(r+1)} = \underset{X_i \in \Psi_i \subset R^{p_i}}{\arg\min} \left\{ f(X_1^{(r+1)}, \ldots, X_{i-1}^{(r+1)}, X_i, X_{i+1}^{(r)}, \ldots, X_t^{(r)}) \right\} \qquad (1)$$

AO-3 If $\| x^{(r+1)} - x^{(r)} \| \leq \varepsilon$ or r > L, then quit; otherwise, set r = r+1 and go to AO-2.

Figure: Alternating Optimization Algorithm

EU Assumption

Before we go into the proof details, I would like to introduce some convergence properties of AO that might be useful.
Typically, we have this EU assumption:

Existence and Uniqueness (EU) Assumption. Let $\Psi_i \subseteq \Re^{p_i}$, i = 1,…,t; and let $\Psi = \Psi_1 \times \ldots \times \Psi_t$. Partition $x = (X_1,\ldots,X_t)^T$, $X_i \in \Re^{p_i}$, and let $g(X_i) = f(\cancel{X}_1, \ldots, \cancel{X}_{i-1}, X_i, \cancel{X}_{i+1}, \ldots, \cancel{X}_t)$, i = 1,…,t, If $x \in \Psi$, then

$$g(X_i) \text{ has a unique (global) minimizer for } X_i \in \Psi_i. \qquad \text{(EU)}$$

Global Convergence

Theorem 2 [10]. Suppose that (EU) holds for $f : \Re^s \mapsto \Re$. Let x = $(X_1,\ldots,X_t)^T$, and $\Psi = \Psi_1 \times \ldots \times \Psi_t$, where Ψ_i is a compact subset of $\Re^{p_i}$, i = 1,…,t. Let $\{x^{(r+1)} = T(x^{(r)})\}$ denote the AO iterate sequence begun at $x^{(0)} \in \Psi$, and denote the fixed points of T as $\Omega = \{x \in \Psi : x = T(x)\}$. Then:

(i) if $x^* \in \Omega$, then $x^* = (X_1^*,\ldots,X_t^*)^T$ satisfies, for i = 1,…,t,

$$X_i^* = \underset{X_i \in \Psi_i \subset R^{p_i}}{arg\,min} \left\{ f(X_1^*,\ldots,X_{i-1}^*,X_i,X_{i+1}^*,\ldots,X_t^*) \right\};$$

(ii) $f(x^{(r+1)}) \le f(x^{(r)})$, equality if and only if $x^{(r)} \in \Omega$;

(iii) either: (a) $\exists\, x^* \in \Omega$ and $r_o \in \Re$ so that $x^{(r)} = x^*$ for all $r \ge r_o$;

or (b) the limit of every convergence subsequence of $\{x^{(r)}\}$ is in Ω.

Indications

- Under certain conditions, all limit points of an AO sequence are either saddle points of a special type of minimizers.
- However, not all saddle point can be captured by AO. Only those which looks like a minimizer along the grouped coordinate (X_1, X_2, etc) can be captured.
- The potential for convergence to a saddle point is a "price" need to pay.
- What if strict convex functions? Converge to the global optimal q-linearly

Indications

- Under certain conditions, all limit points of an AO sequence are either saddle points of a special type of minimizers.
- However, not all saddle point can be captured by AO. Only those which looks like a minimizer along the grouped coordinate (X_1, X_2, etc) can be captured.
- The potential for convergence to a saddle point is a "price" need to pay.
- What if strict convex functions? Converge to the global optimal q-linearly

Theorem 3 [10]. Let x^* be a local minimizer of $f:\Re^s \mapsto \Re$ for which $\nabla^2 f(x^*)$ is positive definite, and let f be C^2 in a neighborhood $N(x^*,\delta)$. Let $0 < \varepsilon \leq \delta$ be chosen so that f is strictly convex on $N(x^*,\varepsilon)$. Finally, assume that if $y = (\bar{X}_1, \ldots, \bar{X}_{i-1}, Y_i, \bar{X}_{i+1}, \ldots, \bar{X}_t)^T \in N(x^*,\varepsilon)$, and Y_i^* locally minimizes $g_i(Y_i) = f(\bar{X}_1, \ldots, \bar{X}_{i-1}, Y_i, \bar{X}_{i+1}, \ldots, \bar{X}_t)$, then Y_i^* is also the unique global minimizer of g_i:

Then for any $x^{(0)} \in N(x^*,\varepsilon)$, the corresponding AO iterate sequence $\{x^{(r+1)} = T(x^{(r)})\} \to x^*$ q-linearly.

- The previous two results are making strong assumptions:
  - Each restricted minimization problem has a unique solution.
  - Strict convexity near the optimal.
- Here, study the functions with relaxed assumptions:
  - Minimize a nondifferentiable (nonconvex) function $f(x_1, \cdots, x_N)$ with certain separability and regularity properties.
  - Converge to a stationary point if f is
    - pseudoconvex in every pair of coordinate blocks from among $N-1$ coordinate blocks; or
    - f has at most one minimum in each of $N-2$ coordinate blocks
  - If f is quasiconvex and hemivariate in every coordinate block, the assumption could be relaxed further.

- The previous two results are making strong assumptions:
  - Each restricted minimization problem has a unique solution.
  - Strict convexity near the optimal.
- Here, study the functions with relaxed assumptions:
  - Minimize a nondifferentiable (nonconvex) function $f(x_1, \cdots, x_N)$ with certain separability and regularity properties.
  - Converge to a stationary point if f is
    - pseudoconvex in every pair of coordinate blocks from among $N-1$ coordinate blocks; or
    - f has at most one minimum in each of $N-2$ coordinate blocks
  - If f is quasiconvex and hemivariate in every coordinate block, the assumption could be relaxed further.

Preliminary

- Effective domain: $dom\ h = \{x \in R^m | h(x) < \infty\}$
- A function f is proper if $f \neq \infty$.
- A space is compact if it is closed and bounded.
- Lower Directional derivative:

$$h'(x; d) = lim \inf_{\lambda \to 0} \frac{h(x + \lambda d) - h(x)}{\lambda}$$

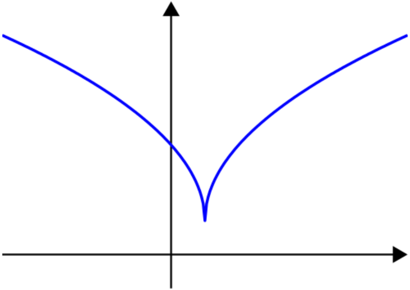
- Gateaux-Differentiable:

$$h'(x; d) = lim_{\lambda \to 0} \frac{h(x + \lambda d) - h(x)}{\lambda} = \frac{d}{d\lambda} h(x + \lambda d)|_{\lambda = 0}$$

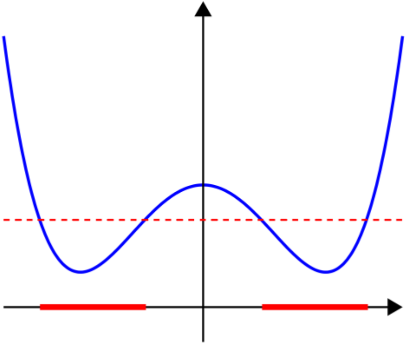
If the transformation $H(d) : d \to h'(x; d)$ is continuous and linear, then F is said to be Gateaux differentiable at u.
In other words,

$$h'(x; \alpha d) = \alpha h'(x; d);$$
$$h'(x; (d_1 + d_2)) = h'(x; d_1) + h'(x; d_2)$$

- *Quasiconvex*: a real-valued function defined on an interval or on a convex subset or a real vector space such that the inverse image of any set of the form $(-\infty, a)$ is a convex set.



Quasiconvex but not convex



Not Quasiconvex

$$h(\lambda x + (1-\lambda)y) \leq \max(h(x), h(y)) \quad \forall \lambda \in [0,1]$$

or $$h(x + \lambda d) \leq \max(h(x), h(x+d))$$

PseudoConvex

- Pseudoconvex: a function satisifying the following property:

$$h(x+d) \geq h(x), \quad \text{whenever x} \in \text{dom h and } h'(x;,d) \geq 0$$

- arctan(x) is pseudo convex, but not convex. Its derivative is

$$\frac{1}{1+x^2}$$

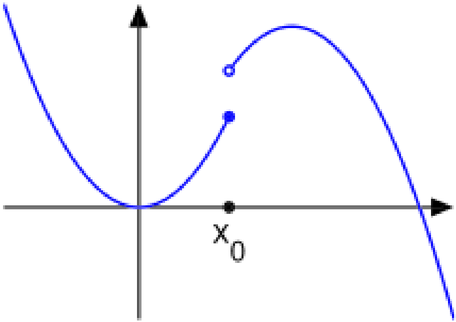
  which is always positive. But it's not convex function.

- hemivariate: h is not constant on any line segment belonging to dom h. Used to guarantee the unique minimizer for each restricted minimization problem.

Lower Semi-continous

- lower semi-continuous:

$$\lim_{x \longrightarrow x_0} \inf f(x) \geq f(x_0)$$



- A Lower Semi-Continuous Function indicates that the limit point x_0 (if in the effective domain), the function value f is always smaller than the limiting value of f.

Stationary Point & Regular Function

- z is a stationary point if

$$f'(z; d) \geq 0, \quad \forall d$$

- f is regular if $\forall d = (d_1, \cdots, d_N)$ which satisfy

$$f'(z; (0, \cdots, d_k, \cdots, 0)) \geq 0 \Longrightarrow f'(z; d) \geq 0$$

- coordinatewise minimum point:

$$f(z + (0, \cdots, d_k, \cdots, 0)) \geq f(z), \quad \forall d_k$$

- This is less strong than the following condition:

$$f'(z; d) = \sum_{k=1}^{N} f'(z; (0, \cdots, d_k, \cdots, 0)), \quad \textit{for all } d = (d_1, \cdots, d_N)$$

Stationary Point & Regular Function

- z is a stationary point if

$$f'(z;d) \geq 0, \quad \forall d$$

- f is regular if $\forall d = (d_1, \cdots, d_N)$ which satisfy

$$f'(z;(0,\cdots,d_k,\cdots,0)) \geq 0 \Longrightarrow f'(z;d) \geq 0$$

- coordinatewise minimum point:

$$f(z+(0,\cdots,d_k,\cdots,0)) \geq f(z), \quad \forall d_k$$

- This is less strong than the following condition:

$$f'(z;d) = \sum_{k=1}^{N} f'(z;(0,\cdots,d_k,\cdots,0)), \quad \textit{for all } d = (d_1,\cdots,d_N)$$

Stationary Point & Regular Function

- z is a stationary point if

$$f'(z;d) \geq 0, \quad \forall d$$

- f is regular if $\forall d = (d_1, \cdots, d_N)$ which satisfy

$$f'(z;(0,\cdots,d_k,\cdots,0)) \geq 0 \Longrightarrow f'(z;d) \geq 0$$

- coordinatewise minimum point:

$$f(z+(0,\cdots,d_k,\cdots,0)) \geq f(z), \quad \forall d_k$$

- This is less strong than the following condition:

$$f'(z;d) = \sum_{k=1}^{N} f'(z;(0,\cdots,d_k,\cdots,0)), \quad \textit{for all } d = (d_1,\cdots,d_N)$$

An example of Regular Function with no additive property

$$f(x_1, x_2) = \phi(x_1, x_2) + \phi(-x_1, x_2) + \phi(x_1, -x_2) + \phi(-x_1, -x_2)$$
$$where\ \phi(a, b) = max\{0, a + b - \sqrt{a^2 + b^2}\}$$

It's easy to verify that

$$f'(\mathbf{0}; (d_1, 0)) = 0, \quad f'(\mathbf{z}; (0, d_2)) = 0;$$
$$f'(\mathbf{0}; d) = |d_1| + |d_2| - \sqrt{d_1^2 + d_2^2} \neq f'(\mathbf{0}; (d_1, 0)) + f'(\mathbf{0}; (0, d_2))$$

Stationary Point = Coordinate-wise Minimum?

- z is a stationary point if

$$f'(z; d) \geq 0, \quad \forall d$$

- f is regular if $\forall d = (d_1, \cdots, d_N)$ which satisfy

$$f'(z; (0, \cdots, d_k, \cdots, 0)) \geq 0 \Longrightarrow f'(z; d) \geq 0$$

- coordinatewise minimum point:

$$f(z + (0, \cdots, d_k, \cdots, 0)) \geq f(z), \quad \forall d_k$$

- A coordinatewise minimum point z is a stationary point whenever f is regular at z.
- When is a function regular?

- z is a stationary point if

$$f'(z;d) \geq 0, \quad \forall d$$

- f is regular if $\forall d = (d_1, \cdots, d_N)$ which satisfy

$$f'(z;(0,\cdots,d_k,\cdots,0)) \geq 0 \Longrightarrow f'(z;d) \geq 0$$

- coordinatewise minimum point:

$$f(z + (0,\cdots,d_k,\cdots,0)) \geq f(z), \quad \forall d_k$$

- A coordinatewise minimum point z is a stationary point whenever f is regular at z.
- When is a function regular?

Smoothness Assumptions

$$f(x_1, \cdots, x_N) = f_0(x_1, \cdots, x_N) + \sum_{k=1}^{N} f_k(x_k)$$

A1 dom f_0 is open and f_0 is Gateaux-differentiable on dom f_0.

A2 f_0 is Gateaux-differentiable on $int(dom\ f_0)$ and for every $z \in \text{dom } f \cap$ bdry(dom f_0) , there exist

$$f(z + (0, \cdots, d_k, \cdots, 0)) < f(z)$$

Essentially the minimizer never occurs at the boundary point.

Lemma 3.1 Under A1, f is regular at each $z \in domf$; Under A2, f is regular at each coordinatewise minimum point z of f.

Smoothness Assumptions

$$f(x_1, \cdots, x_N) = f_0(x_1, \cdots, x_N) + \sum_{k=1}^{N} f_k(x_k)$$

A1 dom f_0 is open and f_0 is Gateaux-differentiable on dom f_0.

A2 f_0 is Gateaux-differentiable on $int(dom\ f_0)$ and for every $z \in \text{dom } f \cap \text{bdry}(\text{dom } f_0)$, there exist

$$f(z + (0, \cdots, d_k, \cdots, 0)) < f(z)$$

Essentially the minimizer never occurs at the boundary point.

Lemma 3.1 Under A1, f is regular at each $z \in domf$; Under A2, f is regular at each coordinatewise minimum point z of f.

Proof for Lemma 3.1

Lemma 3.1 Under A1, f is regular at each $z \in domf$; Under A2, f is regular at each coordinatewise minimum point z of f.

Under A1, if $z \in dom\ f \Longrightarrow z \in dom\ f_0$; Under A2, $z \in int(dom\ f_0)$ for any d such that $f'(z;(0,\cdots,d_k,\cdots,0)) \geq 0 \quad k = 1,\cdots N$

We need to prove $f'(z;d) \geq 0$.

$$\begin{aligned}
f'(z;d) &= \underbrace{<\nabla_0(z), d>}_{Gateaux-differentiable} + \liminf_{\lambda \downarrow 0} \sum_{k=1}^{N} [f_k(x_k + \lambda d_k) - f_k(x_k)]/\lambda \\
&\geq <\nabla_0(z), d> + \sum_{k=1}^{N} \liminf_{\lambda \downarrow 0} [f_k(x_k + \lambda d_k) - f_k(x_k)]/\lambda \qquad (1) \\
&= <\nabla f_0(z), d> + \sum_{k=1}^{N} f_k'(z_k; d_k) \qquad (2) \\
&= \sum_{k=1}^{N} f'(z;(0,\cdots,d_k,\cdots,0)) \geq 0 \qquad (3)
\end{aligned}$$

Proof for Lemma 3.1

Lemma 3.1 Under A1, f is regular at each $z \in dom f$; Under A2, f is regular at each coordinatewise minimum point z of f.

Under A1, if $z \in dom\ f \implies z \in dom\ f_0$; Under A2, $z \in int(dom\ f_0)$ for any d such that $f'(z; (0, \cdots, d_k, \cdots, 0)) \geq 0 \quad k = 1, \cdots N$

We need to prove $f'(z; d) \geq 0$.

$$
\begin{aligned}
f'(z; d) &= \underbrace{< \nabla_0(z), d >}_{Gateaux-differentiable} + \liminf_{\lambda \downarrow 0} \sum_{k=1}^{N} [f_k(x_k + \lambda d_k) - f_k(x_k)]/\lambda \\
&\geq < \nabla_0(z), d > + \sum_{k=1}^{N} \liminf_{\lambda \downarrow 0} [f_k(x_k + \lambda d_k) - f_k(x_k)]/\lambda \quad (1) \\
&= < \nabla f_0(z), d > + \sum_{k=1}^{N} f'_k(z_k; d_k) \quad (2) \\
&= \sum_{k=1}^{N} f'(z; (0, \cdots, d_k, \cdots, 0)) \geq 0 \quad (3)
\end{aligned}
$$

Comments of Regularity

- This work makes the assumption of A1 or A2.
- Under such assumptions, a coordinate-wise minimum is a stationary point.
- So the following convergence analysis just need to show that the algorithm converges to a coordinate-wise minimum point.
- A1 & A2 only care about the smoothness of f_0. Even if $f_1, \cdots, f_N$ are not smooth, the claim here is still valid.
- Need additional properties to guarantee the convergence.

Block Coordinate Descent Algorithm

BCD Method.

Initialization. Choose any $x^0 = (x_1^0, \ldots, x_N^0) \in \operatorname{dom} f$.

Iteration $r+1$, $r \geq 0$. Given $x^r = (x_1^r, \ldots, x_N^r) \in \operatorname{dom} f$, choose an index $s \in \{1, \ldots, N\}$ and compute a new iterate

$$x^{r+1} = (x_1^{r+1}, \ldots, x_N^{r+1}) \in \operatorname{dom} f$$

satisfying

$$x_s^{r+1} \in \arg\min_{x_s} f(x_1^r, \ldots, x_{s-1}^r, x_s, x_{s+1}^r, \ldots, x_N^r), \tag{2}$$

$$x_j^{r+1} = x_j^r, \qquad \forall j \neq s. \tag{3}$$

Cyclic Rule

Essentially Cyclic Rule. There exists a constant $T \geq N$ such that every index $s \in \{1, \ldots, N\}$ is chosen at least once between the rth iteration and the $(r + T - 1)$th iteration, for all r.

A well-known special case of this rule, for which $T = N$, is given below.

Cyclic Rule. Choose $s = k$ at iterations k, $k + N$, $k + 2N$, $\ldots$, for $k = 1, \ldots, N$.

Assuming f continuous, without using the Special Structure

Theorem 4.1 Assume the level set $X^0 = \{x : f(x) \leq f(x^0)\}$ is compact and that f is continuous on X^0. Then, the sequence generated by BCD is defined and bounded. Moreover,

(a) If $f(x_1, \ldots, x_N)$ is pseudoconvex in (x_k, x_i) for every $i, k \in \{1, \ldots, N\}$, and if f is regular at every $x \in X^0$, then every cluster point of $\{x^r\}$ is a stationary point of f.

(b) If $f(x_1, \ldots, x_N)$ is pseudoconvex in (x_k, x_i) for every $i, k \in \{1, \ldots, N-1\}$, if f is regular at every $x \in X^0$, and if the cyclic rule is used, then every cluster point of $\{x^r\}_{r \equiv (N-1) \bmod N}$ is a stationary point of f.

(c) If $f(x_1, \ldots, x_N)$ has at most one minimum in x_k for $k = 2, \ldots, N-1$, and if the cyclic rule is used, then every cluster point z of $\{x^r\}_{r \equiv (N-1) \bmod N}$ is a coordinatewise minimum point of f. In addition, if f is regular at z, then z is a stationary point of f.

Proof

- Goal: To show that the BCD algorithm converges to z such that

$$f(z + (0, \cdots, d_k, \cdots, 0)) \geq f(z); \quad \forall d_k, k = 1, \cdots, N$$

- The stationary point property is obtained if the function is regular.
- The key process is to show the following by induction: for $j = 1, \cdots, T-1$,

$$f(z^j) \leq f(z^j + (0, \cdots, d_k, \cdots, 0)), \quad \forall d_k, \forall k = s^1, \cdots, s^j.$$

$$
\begin{aligned}
& X^0 = \{x : f(x) \le f(x^0)\} \textit{ is compact} \\
\Rightarrow \quad & f(x^{r+1}) \le f(x^r) \textit{ and } x^{r+1} \in X^0 \textit{ for all } r = 0, 1, \cdots \\
\Rightarrow \quad & \{x^r\} \text{is bounded.} \\
\Rightarrow \quad & \text{Consider any subsequence } \{x^r\}_{r \in R}, \text{converging to } z, \\
& \text{where } R \subseteq \{0, 1, \cdots\}, \\
& \{x^{r-T+1+j}\}_{r \in R} \text{ is bounded.} \\
& \text{By passing } r \text{ to a subsubsequence, we have} \\
\Rightarrow \quad & \{x^{r-T+1+j}\}_{r \in R} \to z^j,\ j = 1, \cdots, T \\
\text{Note that} \quad & z^{T-1} = z; \\
\Rightarrow \quad & \underbrace{f(x^0) \ge lim_{r \to \infty} f(x^r) = f(z^1) = \cdots f(z^T)}_{\text{f decreasing monotonically, and f is continuous}}
\end{aligned}
$$

Assume that the index s chosen at iteration $r - T + 1 + j$, $j \in \{1, \cdots, T\}$, is the same for all $r \in R$ (denoted as s^j), then

$$\begin{aligned} f(x^{r-T+1+j}) \leq f(x^{r-T+1+j} + (0, \cdots, d_{s^j}, \cdots, 0)), \quad & \forall d^{s^j}, j = 1, \cdots, T \\ x_k^{r-T+1+j} = x_K^{r-T+j} \quad & \forall k \neq s^j, j = 2, \cdots, T \end{aligned}$$

Based on the continuity of f on X^0, we have

$$\begin{aligned} & f(z^j) \leq f(z^j + (0, \cdots, d_{s^j}, \cdots, 0)), \quad \forall d_{s^j}, j = 1, \cdots, T \\ & z_k^j = z_k^{j-1} \quad \forall k \neq s^j, j = 2, \cdots, T \\ \Rightarrow \quad & f(z^{j-1}) = f(z^j) \leq \underbrace{f(z^{j-1} + (0, \cdots, d_{s^j}, \cdots, 0))}_{z^j \text{ and } z^{j-1} \text{ only differ at index } s^j} \\ & \forall d_{s^j}, j = 2, \cdots, T \end{aligned}$$

The limit point z^{j-1} is also the directional minimizer for d_{s^j}.

We have

$$f(z^{j-1}) \leq f(z^{j-1} + (0, \cdots, d_{s^j}, \cdots, 0)), \quad j = 2, \cdots, T$$

(a). f is pseudoconvex in (x_k, x_i) for every i, k in $\{1, \cdots, N\}$

(b). f is pseudoconvex in (x_k, x_i) for every i, k in $\{1, \cdots, N-1\}$

$\Rightarrow$ if f is pseudoconvex in (x_k, x_i), $\forall i, k \in s^1 \cup \cdots, \cup s^{T-1}$

Claim for $j = 1, \cdots, T-1$,

$$f(z^j) \leq f(z^j + (0, \cdots, d_k, \cdots, 0)), \qquad \forall d_k, \forall k = s^1, \cdots, s^j. \tag{4}$$

Note that

$$f(z) = f(z^{T-1}) \leq f(z^{T-1} + (0, \cdots, d_{s^T}, \cdots, 0)$$

Then we have z is a coordinate-wise minimum.

We have

$$f(z^{j-1}) \leq f(z^{j-1} + (0, \cdots, d_{s^j}, \cdots, 0)), \quad j = 2, \cdots, T$$

(a). f is pseudoconvex in (x_k, x_i) for every i, k in $\{1, \cdots, N\}$

(b). f is pseudoconvex in (x_k, x_i) for every i, k in $\{1, \cdots, N-1\}$

$\Rightarrow$ if f is pseudoconvex in (x_k, x_i), $\forall i, k \in s^1 \cup \cdots, \cup s^{T-1}$

Claim for $j = 1, \cdots, T-1$,

$$f(z^j) \leq f(z^j + (0, \cdots, d_k, \cdots, 0)), \quad \forall d_k, \forall k = s^1, \cdots, s^j. \tag{5}$$

Proof by Induction

- $j = 1$, automatically satisfied by the minimization.
- Suppose (5) holds for $j = 1, \cdots, \ell - 1$ for $\ell \in \{2, \cdots, T-1\}$, we'll show (5) holds for ℓ.

$$
\begin{aligned}
& f(z^{j-1}) \le f(z^{j-1} + (0,\cdots,d_{s^j},\cdots,0)) \quad \forall d_{s^j}, j = 2,\cdots,T \\
\Rightarrow \quad & f(z^{\ell-1}) \le f(z^{\ell-1} + (0,\cdots,d_{s^\ell},\cdots,0)) \quad \forall d_{s^\ell} \\
\Rightarrow \quad & f'(z^{\ell-1};(0,\cdots,z^{\ell}_{s^\ell} - z^{\ell-1}_{s^\ell},\cdots,0)) \ge 0 \quad (\textit{pseudoconvexity})
\end{aligned}
$$

Based on Induction assumption, we have

$$
f'(z^{\ell-1};(0,\cdots,d_k,\cdots,0)) \ge 0, \forall d_k, k = s^1,\cdots,s^{\ell-1}
$$

$$
\Rightarrow \quad \underbrace{f'(z^{\ell-1};(0,\cdots,d_k,\cdots,0) + (0,\cdots,z^{\ell}_{s^\ell} - z^{\ell-1}_{s^\ell},\cdots,0)) \ge 0}_{\text{as f is regular}} \tag{6}
$$

$$
\begin{aligned}
\Rightarrow \quad & f(z^{\ell-1}) \le f(z^{\ell} + (0,\cdots,d_k,\cdots,0)) \quad (\text{f is pseudoconvex}) && (7) \\
\Rightarrow \quad & f(z^{\ell}) = f(z^{\ell-1}) \le f(z^{\ell} + (0,\cdots,d_k,\cdots,0)) \quad k = s^1,\cdots,s^{\ell-1} && (8) \\
As \quad & f(z^{j}) \le f(z^{j} + (0,\cdots,d_{s^j},\cdots,0)), \quad \forall d_{s^j}, j = 1,\cdots,T && (9) \\
\Rightarrow \quad & f(z^{\ell}) \le f(z^{\ell} + (0,\cdots,d_k,\cdots,0)) \quad k = s^1,\cdots,s^{\ell} && (10) \\
\Rightarrow \quad & \text{Claim holds for } \ell. && (11)
\end{aligned}
$$

$$
\begin{aligned}
& f(z^{j-1}) \le f(z^{j-1} + (0, \cdots, d_{s^j}, \cdots, 0)) \quad \forall d_{s^j}, j = 2, \cdots, T \\
\Rightarrow\ & f(z^{\ell-1}) \le f(z^{\ell-1} + (0, \cdots, d_{s^\ell}, \cdots, 0)) \quad \forall d_{s^\ell} \\
\Rightarrow\ & f'(z^{\ell-1}; (0, \cdots, z^{\ell}_{s^\ell} - z^{\ell-1}_{s^\ell}, \cdots, 0)) \ge 0 \quad (\textit{pseudoconvexity})
\end{aligned}
$$

Based on Induction assumption, we have

$$
f'(z^{\ell-1}; (0, \cdots, d_k, \cdots, 0)) \ge 0, \forall d_k, k = s^1, \cdots, s^{\ell-1}
$$

$$
\begin{aligned}
\Rightarrow\ & \underbrace{f'(z^{\ell-1}; (0, \cdots, d_k, \cdots, 0) + (0, \cdots, z^{\ell}_{s^\ell} - z^{\ell-1}_{s^\ell}, \cdots, 0)) \ge 0}_{\text{as f is regular}} && (6) \\
\Rightarrow\ & f(z^{\ell-1}) \le f(z^{\ell} + (0, \cdots, d_k, \cdots, 0)) \quad (\text{f is pseudoconvex}) && (7) \\
\Rightarrow\ & f(z^{\ell}) = f(z^{\ell-1}) \le f(z^{\ell} + (0, \cdots, d_k, \cdots, 0)) \quad k = s^1, \cdots, s^{\ell-1} && (8) \\
As\ & f(z^{j}) \le f(z^{j} + (0, \cdots, d_{s^j}, \cdots, 0)), \quad \forall d_{s^j}, j = 1, \cdots, T && (9) \\
\Rightarrow\ & f(z^{\ell}) \le f(z^{\ell} + (0, \cdots, d_k, \cdots, 0)) \quad k = s^1, \cdots, s^{\ell} && (10) \\
\Rightarrow\ & \text{Claim holds for } \ell. && (11)
\end{aligned}
$$

$$
\begin{aligned}
& f(z^{j-1}) \le f(z^{j-1} + (0, \cdots, d_{s^j}, \cdots, 0)) \quad \forall d_{s^j}, j = 2, \cdots, T \\
\Rightarrow \quad & f(z^{\ell-1}) \le f(z^{\ell-1} + (0, \cdots, d_{s^\ell}, \cdots, 0)) \quad \forall d_{s^\ell} \\
\Rightarrow \quad & f'(z^{\ell-1}; (0, \cdots, z^{\ell}_{s^\ell} - z^{\ell-1}_{s^\ell}, \cdots, 0)) \ge 0 \quad (\textit{pseudoconvexity})
\end{aligned}
$$

Based on Induction assumption, we have

$$
f'(z^{\ell-1}; (0, \cdots, d_k, \cdots, 0)) \ge 0, \forall d_k, k = s^1, \cdots, s^{\ell-1}
$$

$$
\begin{aligned}
\Rightarrow \quad & \underbrace{f'(z^{\ell-1}; (0, \cdots, d_k, \cdots, 0) + (0, \cdots, z^{\ell}_{s^\ell} - z^{\ell-1}_{s^\ell}, \cdots, 0)) \ge 0}_{\text{as f is regular}} && (6) \\
\Rightarrow \quad & f(z^{\ell-1}) \le f(z^{\ell} + (0, \cdots, d_k, \cdots, 0)) \quad (\text{f is pseudoconvex}) && (7) \\
\Rightarrow \quad & f(z^{\ell}) = f(z^{\ell-1}) \le f(z^{\ell} + (0, \cdots, d_k, \cdots, 0)) \quad k = s^1, \cdots, s^{\ell-1} && (8) \\
As \quad & f(z^{j}) \le f(z^{j} + (0, \cdots, d_{s^j}, \cdots, 0)), \quad \forall d_{s^j}, j = 1, \cdots, T && (9) \\
\Rightarrow \quad & f(z^{\ell}) \le f(z^{\ell} + (0, \cdots, d_k, \cdots, 0)) \quad k = s^1, \cdots, s^{\ell} && (10) \\
\Rightarrow \quad & \text{Claim holds for } \ell. && (11)
\end{aligned}
$$

Brief Summary

As $f(z^{j-1}) = f(z^j) \leq f(z^{j-1} + (0, \cdots, d_{s^j}, \cdots, 0)) \quad \forall d_{s^j}, j = 2, \cdots, T$,

$$f(z^{T-1}) \leq f(z^{T-1} + (0, \cdots, d_k, \cdots, 0)) \quad k = s^T$$

Combined with our induction proof, we have

$$f(z^{T-1}) \leq f(z^{T-1} + (0, \cdots, d_k, \cdots, 0)) \quad k = s^1, \cdots, s^T$$

Recall that $z^{T-1} = z$, hence z is coordinate-wise minimum.
As f is regular, z is also a stationary point.

Unique Minimizer at Each Step $\Longrightarrow$ unique limiting point?

(c). f has at most one minimum in x_k for $k = 2, \cdots, N-1$, and if the cycle rule is used. Then every cluster point z of $\{x^r\}_{r \equiv (N-1) mod N}$, is a coordinatewise minimum point of f. If f is regular at z, then it's also a stationary point.

Proof

$$\text{Define a function as } d_{s^j} \to f(z^j + (0, \cdots, d_{s^j}, \cdots, 0))$$

$$f(z^{j-1}) = f(z^j) \leq f(z^{j-1} + (0, \cdots, d_{s^j}, \cdots, 0)) \quad \forall d_{s^j}, j = 2, \cdots, T \tag{12}$$

attains its minimum at both 0 and $z_{s^j}^{j-1} - z_{s^j}^{j}$.

$\Longrightarrow z_{s^j}^{j-1} - z_{s^j}^{j} = 0 \quad$ (uniqueness of minimization function)

$\Longrightarrow z^{j-1} = z^j \Longrightarrow z^1 = z^2 = \cdots, z^{T-1} = z$

Plus, $f(z^{j-1}) = f(z^j) \leq f(z^{j-1} + (0, \cdots, d_{s^j}, \cdots, 0)) \quad \forall d_{s^j}, j = 2, \cdots, T$

Hence, z is the coordinate-wise minimizer.

Recap the Theorem

Assuming f continuous, without using the Special Structure

Theorem 4.1 Assume the level set $X^0 = \{x : f(x) \leq f(x^0)\}$ is compact and that f is continuous on X^0. Then, the sequence generated by BCD is defined and bounded. Moreover,

(a) If $f(x_1, \ldots, x_N)$ is pseudoconvex in (x_k, x_i) for every $i, k \in \{1, \ldots, N\}$, and if f is regular at every $x \in X^0$, then every cluster point of $\{x^r\}$ is a stationary point of f.

(b) If $f(x_1, \ldots, x_N)$ is pseudoconvex in (x_k, x_i) for every $i, k \in \{1, \ldots, N-1\}$, if f is regular at every $x \in X^0$, and if the cyclic rule is used, then every cluster point of $\{x^r\}_{r \equiv (N-1) \bmod N}$ is a stationary point of f.

(c) If $f(x_1, \ldots, x_N)$ has at most one minimum in x_k for $k = 2, \ldots, N-1$, and if the cyclic rule is used, then every cluster point z of $\{x^r\}_{r \equiv (N-1) \bmod N}$ is a coordinatewise minimum point of f. In addition, if f is regular at z, then z is a stationary point of f.

Summary & Comments

- if f is pseudoconvex, then f is pseudoconvex in (x_k, x_i) for all k, i
- if f is quasiconvex and hemivariate in x_k, then f has at most one minimum in x_k. Some papers refer it as strict quasiconvex.
- If f is continuous, and only 2-blocks are involved. Then it does not require unique minimizer to converge to a stationary point. (This result is used in the convergence proof of alternating least-square proof in NMF)
- The previous proof does not take advantage of the special structure and assume f to conbinuous on a bounded level set.
- Next we show that considering the special structure without requiring f to be smooth.

- if f is pseudoconvex, then f is pseudoconvex in (x_k, x_i) for all k, i
- if f is quasiconvex and hemivariate in x_k, then f has at most one minimum in x_k. Some papers refer it as strict quasiconvex.
- If f is continuous, and only 2-blocks are involved. Then it does not require unique minimizer to converge to a stationary point. (This result is used in the convergence proof of alternating least-square proof in NMF)
- The previous proof does not take advantage of the special structure and assume f to conbinuous on a bounded level set.
- Next we show that considering the special structure without requiring f to be smooth.

Sleepy? Shall we continue?

Assumptions

(B1) f_0 is continuous on dom f_0

(B2) for each $k \in \{1, \cdots, N\}$, and$(x_j)_{j \neq k}$, the function $x_k \to f(x_1, \cdots, x_N)$ is quasiconvex and hemivariate.

(B3) $f_0, f_1, \cdots, f_N$ is lower semi-continuous.

Meanwhile, f_0 satisfy the one of the following assumption:

(C1) dom f_0 is open and f_0 tends to ∞ at every boundary point of dom f_0

(C2) dom $f_0 = Y_1 \times \cdots \times Y_N$ for some $Y_k \subseteq R^{n_k}, k = 1, \cdots, N$

- C2 allows a finite value at boundary point.
- We'll show that Assumption B1-B3, together with either C1 or C2, ensure that every cluster poiint of the iterates generated by the BCD methods is a coordinate minimum point of f.

Proposition 5.1

Suppose that $f, f_0, \cdots, f_N$ satisfy B1-B3 and f_0 satisfy C1 or C2. Then, either $\{f(x^r)\} \downarrow -\infty$ or else every cluster point $z = (z_1, \cdots, z_N)$ is a coordinatewise minimum point of f.

Proof Strategy

$$
\begin{array}{ll}
\text{As} & f(x^0) < \infty, \text{ and } f(x^{r+1}) \leq f(x^r) \\
\Rightarrow & \{f(x^r)\} \downarrow -\infty \\
& \text{or } \{f(x^r)\} \text{ converges to some limit and } \{f(x^{r+1}) - f(x^r)\} \to 0 \\
& \text{Let } z \text{ be any cluster point of } \{x^r\} \\
\Rightarrow & f(z) \leq \lim_{r\to\infty} f(x^r) \leq \infty \quad (\text{as f is lower semi-continuous})
\end{array}
$$

- First, we show that for any convergent sequence $\{x^r\} \to z$, we have $\{x^{r+1}\} \to z$;
- We'll prove this by contradiction.
- Then, we prove z is a coordinate-wise minimum.

Claim of convergence for x^r

Claim: for any convergent subsequence $\{x^r\}_{r \in R} \to z$, we have

$$\{x^{r+1}\} \to z$$

Sketch of the Proof

- Proof by contradiction
- If $\{x^{r+1}\}$ converges to a different value z', then all the values between z and z' have

$$f(\lambda z + (1-\lambda)z') = f(z) = f(z')$$

  contradicting to the uniqueness of each minimization of coordinate block.

Claim of convergence for x^r

Claim: for any convergent subsequence $\{x^r\}_{r\in R} \to z$, we have

$$\{x^{r+1}\} \to z$$

Prove by Contradiction

Suppose the above is not true, then there exits an infinite subsequence $R' \subseteq R$ and a scalar $\epsilon > 0$ such that

$$||x^{r+1} - x^r|| \geq \epsilon, \quad \text{for all r} \in R'$$

So we can assume that there is some nonzero vector d such that

$$\{(x^{r+1} - x^r)/||x^{r+1} - x^r||\}_{r\in R'} \to d \quad (\text{not quite sure why?})$$

and the same coordinate block, say x_s is chosen at the $r+1$-th iteration. So

$$\{f_0(x^r) + f_s(x_s^r)\}_{r\in R'} \to \theta$$

Fix any $\lambda \in [0, \epsilon]$, Let $\hat{z} = z + \lambda d$, and for each $r \in R'$, let

$$\begin{aligned}
& \hat{x}^r = x^r + \lambda(x^{r+1} - x^r)/||x^{r+1} - x^r|| && (13) \\
\Rightarrow \quad & \{\hat{x}^r\}_{r \in R'} \rightarrow \hat{z} && (14)
\end{aligned}$$

$\hat{x}^r$ lies in the segment of x^{r+1} and x^r, thus

$$\begin{aligned}
& f(\hat{x}^r) \leq f(x^r) \quad \forall r \in R' \qquad (\text{f is quasiconvex}) && (15) \\
\Rightarrow \quad & f_0(\hat{x}^r) + f_s(\hat{x}_s^r) \leq f_0(x^r) + f_s(x_s^r) \rightarrow \theta && (16) \\
\Rightarrow \quad & \lim_{r \to \infty, r \in R'} \sup\{f_0 \hat{x}^r) + f_s(\hat{x}_s^r)\} \leq \theta && (17) \\
As \quad & \{f(x^{r+1}) - f(x^r)\} \rightarrow 0 && (18) \\
\Rightarrow \quad & \{f_0(x^{r+1}) + f_s(x_s^{r+1}) - f_0(x^r) - f_s(x_s^r)\}_{r \in R'} \rightarrow 0 && (19) \\
\Rightarrow \quad & \{f_0(x^{r+1}) + f_s(x_s^{r+1})\} \rightarrow \theta && (20) \\
Define \quad & \delta = f_0(\hat{z}) + f_s(\hat{z}_s) - \theta && (21) \\
Then \quad & \delta \leq 0, \text{actually } \delta = 0 && (22)
\end{aligned}$$

$$As \qquad \underbrace{\{(x_1^r, \cdots, x_{s-1}^r, \hat{z}_s, x_{s+1}^r, \cdots, x_N^r)\}}_{\hat{x}^r \ and \ x^r \text{only differ in } s\text{-th block}} \to \hat{z} \tag{23}$$

$$\lim_{r\to\infty, r\in R'} \sup\{f_0\hat{x}^r) + f_s(\hat{x}_s^r)\} \le \theta \tag{24}$$

if $\delta \neq 0$, then for r sufficiently large

$$f_0(x_1^r, \cdots, x_{s-1}^r, \hat{z}_s, x_{s+1}^r, \cdots, x_N^r) \le f_0(x^{r+1}) + f_s(x_s^{r+1}) + \delta/2 \tag{25}$$

$$f(x_1^r, \cdots, x_{s-1}^r, \hat{z}_s, x_{s+1}^r, \cdots, x_N^r) \le f(x^{r+1}) + \delta/2 \tag{26}$$

A contradiction to the fact that x^{r+1} is obtained from x^r by minimizing f with respect to the s-th coordinate block. Hence

$$\delta = 0 \ so \ f_0(\hat{z}) + f_s(\hat{z}_s) = \theta \tag{27}$$

$$f_0(z + \lambda d) + f_s(z_s + \lambda d_s) = \theta, \forall \lambda \in [0, \epsilon] \tag{28}$$

A contradiction to B2 that f is hemivariate in each block. Therefore,

$$\{x^{r+1}\}_{r\in R} \to z$$

$$\{x^{r+j}\}_{r\in R} \to z, \quad \forall j = 0, 1, \cdots, T \tag{29}$$

all converge to the same value, but the sequence could be different?

With (29) and Assumption C1 or C2,

$$f_0(z) + f_k(z_k) \leq f_0(z_1, \cdots, z_{k-1}, x_k, z_{k+1}, \cdots, z_N) + f_k(x_k)$$

$$f_0(x^{r+j}) + f_k(x_k^{r+j}) \leq f_0(x_1^{r+1}, \cdots, x_{k-1}^{r+j}, x_k^{r+j}, x_{k+1}^{r+j}, \cdots, x_N^{r+j}) + f_k(x_k) \forall x_k$$

Based on the continuity of f_0 and lower-semi-continuous property of of f_k, we can push the above inequality to the limit and obtain the solution.

Theorem 5.1

Suppose that $f, f_0, \cdots, f_N$ satisfy Assumptions B1-B3 and that f_0 satisfies Assumption C1 or C2. Also, assume that $\{x : f(x) \leq f(x^0)\}$ is bounded. Then the sequence $\{x^r\}$ generated by the BCD method using the essentially cyclic rule is defined, bounded, and every cluster point is a coordinate-wise minimum point of f.

(B1) f_0 is continuous on dom f_0

(B2) for each $k \in \{1, \cdots, N\}$, and$(x_j)_{j \neq k}$, the function $x_k \to f(x_1, \cdots, x_N)$ is quasiconvex and hemivariate.

(B3) $f_0, f_1, \cdots, f_N$ is lower semi-continuous.

(C1) dom f_0 is open and f_0 tends to ∞ at every boundary point of dom f_0

(C2) dom $f_0 = Y_1 \times \cdots \times Y_N$ for some $Y_k \subseteq R^{n_k}, k = 1, \cdots, N$

Questions

- Does BCD always converge on a compact subset?
- If BCD converges, are all the sequence converging to the same value?
- What if those assumption are not satisfied, could we make any conclusion?