Complexidade e convergência de algoritmos de descida coordenadas de alta ordem para minimização com restrições de caixa.

Vitaliano S. Amaral - UFPI

V Congresso Brasileiro de Jovens Pesquisadores em Matemática - V CBJME

UFMG

11 de Setembro de 2024

Sumário

Inicialmente consideramos o seguinte problema:

$$\text{Minimizar } f(x) \text{ sujeito a } x \in \Omega \subset \mathbb{R}^n.$$

Basicamente um método CD para resolver o problema acima consiste no seguinte:

Algoritmo 1.

Dados $k \leftarrow 0$, $x^0 \in \Omega$.
Repetir os seguintes passos:

Passo 1. *Dado* x^k *escolher* $I_k \subset \{1, 2, \cdots, n\}$;

Passo 2. *Determinar* $x^{k+1} \in \Omega$, *com* $x_i^{k+1} = x_i^k$ *para* $i \notin I_k$ *de modo que satisfaça alguma condição (pode ser decréscimo em* f*). Após isso voltar ao Passo 1.*

Por um bom tempo, esses métodos receberam pouca atenção por parte dos pesquisadores.

Dois dos motivos foram:

1. Baixo desempenho em muitos problemas da época.
2. Falta de desafios em termos de teoria de convergência.

A situação mudou drasticamente nas últimas décadas. Dentre vários motivos, destacamos:

1. O surgimento de problemas de grande porte em que os métodos CD se mostraram úteis para resolvê-los.
2. A necessidade de teorias esclarecedoras sobre os métodos CD.

Alguns dos trabalhos que contribuíram com essa retomada:

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

Foi estudado problema irrestrito onde a função objetivo possui gradiente Lipschitz. Considerando f convexa obtiveram bons resultados de convergência.

S. J. Wright. - *Coordinate descent algorithms.* Mathematical Programming, 151(1):3-34, 2015

Foi analisado abordagens tradicionais e avanços modernos na introdução e análise dos métodos CD. Wright deu uma atenção especial a problemas que surgem com frequência em aplicações de aprendizado de máquina.

Recentemente surgiram vários trabalhos sobre os métodos CD.
Por exemplo, os trabalhos:

- V. S. Amaral; R. Andreani; E. J. G. Birgin; D. S. Marcondes and J. M. Martínez. - *On complexity and convergence of high-order coordinate descent algorithms for smooth nonconvex box-constrained minimization.* Journal of Global Optimization (2022): 1-35
- E. G. Birgin, and J. M. Martínez. - *Block coordinate descent for smooth nonconvex constrained minimization.* Computational Optimization and Applications 83.1 (2022): 1-27.
- I. Necoara and F. Chorobura. - *Random Coordinate Descent Methods for Nonseparable Composite Optimization.* SIAM Journal on Optimization. 33, 2160-2190 (2023)

O restante desta apresentação é dedicado ao trabalho[1], onde foi considerado o seguinte problema:

$$\text{Minimize } \; f(x) \text{ subject to } x \in \Omega, \tag{1}$$

onde $\Omega \subset \mathbb{R}^n$ é dado por $\Omega = \{x \in \mathbb{R}^n \,|\, \mathbf{l} \leqslant x \leqslant \mathbf{u}\}$, e $\mathbf{l}, \mathbf{u} \in \mathbb{R}^n$ são tais que $\mathbf{l} < \mathbf{u}$.

[1] V. S. Amaral; R. Andreani; E. J. G. Birgin; D. S. Marcondes and J. M. Martínez. - *On complexity and convergence of high-order coordinate descent algorithms for smooth nonconvex box-constrained minimization.* Journal of Global Optimization (2022): 1-35

Consideramos:

- f com derivadas primeiras contínuas em Ω.
- $g_P(x) = P_\Omega(x - \nabla f(x)) - x$ para todo $x \in \Omega$, onde P_Ω é a projeção euclidiana em Ω.
- $g_{P,I}(x) \in \mathbb{R}^n$ definida por

$$[g_{P,I}(x)]_i = \begin{cases} [g_P(x)]_i, & \text{se } i \in I, \\ 0, & \text{se } i \notin I. \end{cases}$$

- $M_{\bar{x}}(\cdot)$ uma aproximação de f em torno de $\bar{x}$.

Introduzimos o seguinte método para resolver o Problema (1):

Algoritmo 2.

Suponha $p \in \{1, 2, 3, \ldots\}$, $\alpha > 0$, $\sigma_{min} > 0$, $\tau_2 \geqslant \tau_1 > 1$, $\theta > 0$ *e* $x_0 \in \Omega$ *sejam fornecidos. Inicialize* $k \leftarrow 0$ *e* $\sigma_0 \leftarrow 0$*.*

Passo 1. *Escolha um conjunto não vazio* $I_k \subseteq \{1, \ldots, n\}$*.*

Passo 2. *Calcule* $x^{trial} \in \Omega$, $x_i^{trial} = x_i^k$ *para todos* $i \notin I_k$ *tal que*

$$M_{x^k}(x^{trial}) + \sigma_k \|x^{trial} - x^k\|^{p+1} \leqslant M_{x^k}(x^k)$$

e

$$\|P_\Omega\left[x^{trial} - \nabla\left(M_{x^k}(x) + \sigma_k\|x - x^k\|^{p+1}\right)\Big|_{x = x^{trial}}\right] - x^{trial}\| \leqslant \theta \|x^{trial} - x^k\|^p.$$

Passo 3. *Se*

$$f(x^{trial}) \leqslant f(x^k) - \alpha \|x^{trial} - x^k\|^{p+1},$$

definir $x^{k+1} = x^{trial}$, $\sigma_{k+1} = \sigma_k$ *e voltar ao Passo 1. Caso contrário, atualize* $\sigma_k \leftarrow \max\{\sigma_{min}, \tau\sigma_k\}$ *com* $\tau \in [\tau_1, \tau_2]$ *e voltar ao Passo 2.*

Assumimos a seguinte Suposição:

Suposição 1.

Para todo $x \in \mathbb{R}^n$ *existe* $L > 0$ *tal que*

$$\|\nabla f(x) - \nabla M_{\bar{x}}(x)\| \leqslant L\|x - \bar{x}\|^p,$$

$$M_{\bar{x}}(\bar{x}) = f(\bar{x}) \quad e \quad f(x) \leqslant M_{\bar{x}}(x) + L\|x - \bar{x}\|^{p+1}.$$

Observação 1.

Para satisfazer a Suposição 1, $M_{\bar{x}}(\cdot)$ *pode ser:*

- *o polinômio de Taylor de ordem* p *de* f *ao redor de* $\bar{x}$ *se as derivadas de ordem* p *de* f *satisfazem condições de Lipschitz.*
- *a própria* f. *Neste caso,* p *pode ser arbitrariamente grande, mas apenas as primeiras derivadas de* f *precisam existir.*

Teorema 1.

Suponha que a Suposição 1 seja válida. Se $\sigma_k \geqslant L + \alpha$*, então o ponto* x^{k+1} *calculado pelo Algoritmo 2 está bem definido e satisfaz*

$$f(x^{k+1}) \leqslant f(x^k) - \alpha\|x^{k+1} - x^k\|^{p+1} \tag{2}$$

e

$$\|\nabla g_{P,I_k}(x^{k+1})\| \leqslant \big(L + \tau_2(L+\alpha)(p+1) + \theta\big)\|x^{k+1} - x^k\|^p. \tag{3}$$

As provas de (2) e (3) segue quase que direto das condições do Passo 2 e da Suposição 1.

Teorema 2.

Suponha que a Suposição 1 seja válida. Seja $\{x^k\}$ a sequência gerada pelo Algoritmo 2. Então,

$$\lim_{k\to\infty} \|x^{k+1} - x^k\| = 0, \tag{4}$$

$$\lim_{k\to\infty} \|\nabla g_{P,I_k}(x^{k+1})\| = 0, \tag{5}$$

e

$$\lim_{k\to\infty} \|\nabla g_{P,I_k}(x^k)\| = 0. \tag{6}$$

Resumo da prova:

- Prova de (4): Ω é compacto $\Longrightarrow$ f é limitada em Ω $\Longrightarrow \lim_{k\to\infty}(f(x^k) - f(x^{k+1})) = 0 \overset{(2)}{\Longrightarrow} \lim_{k\to\infty} \|x^{k+1} - x^k\| = 0.$
- Prova de (5): Segue direto de (3) e (4).
- Prova de (6): De Ω compacto e continuidade do gradiente, temos que $\|\nabla g_{P,I}(x)\|$ é uniformemente contínua. Daí, passando limite em

$$\|\nabla g_{P,I}(x^k)\| \leqslant \|\nabla g_{P,I}(x^k) - \nabla g_{P,I}(x^{k+1})\| + \|\nabla g_{P,I}(x^{k+1})\|,$$

usando (4) e (5) obtemos (6).

Observação 2.

Observe que os resultados anteriores não garantem

$$\lim_{k\to\infty} \|g_P(x^k)\| = 0. \tag{7}$$

Para provar (7) foi necessário a seguinte suposição.

Suposição 2.

Existe $\bar{m} < +\infty$ *tal que, para todo* $i \in \{1, \ldots, n\}$*:*

1. *Existe* $k \leqslant \bar{m}$ *tal que* $i \in I_k$*;*
2. *Para qualquer* $k \in \mathbb{N}$*, se* $i \in I_k$*, então existe* $m \leqslant \bar{m}$ *tal que* $i \in I_{k+m}$.

Note que a Suposição 2 nos permite escolher o bloco de coordenadas em cada iteração de diversas formas, basta que, a cada $\bar{m}$ iterações, todos os blocos sejam escolhidos pelo menos uma vez.

Teorema 3.

Suponha que as Suposições 1 e 2 sejam válidas. Seja $\{x^k\}$ a sequência gerada pelo Algoritmo 2. Então,

$$\lim_{k\to\infty} \|\nabla g_P(x^k)\| = 0. \tag{8}$$

Além disso, se $x^* \in \Omega$ for um ponto de acumulação de $\{x^k\}$, então temos que $\|\nabla g_P(x^*)\| = 0$.

Resumo da prova: Seja $i \in \{1, \dots, n\}$. Pela Suposição 2, existe um conjunto $K = \{k_1, k_2, k_3, \dots\}$ tal que $i \in I_{k_\ell}$ para todo $\ell = 1, 2, 3, \dots$. Dado $I \subseteq \{1, \dots, n\}$, de (6) temos,

$$\lim_{k\in K} [\nabla g_P(x^k)]_i = \lim_{k\in K} \nabla g_{P,I}(x^k)]_i = 0 \text{ para todo } i \in I. \tag{9}$$

Seja $j \in \{1, 2, \ldots\}$ arbitrário. De (9) e da continuidade uniforme de ∇g_P, temos

$$\begin{aligned}\lim_{k \in K} |[\nabla g_P(x^{k+j})]_i| &\leqslant \lim_{k \in K} \left|[\nabla g_P(x^{k+j})]_i - [\nabla g_P(x^k)]_i\right| \\ &+ \lim_{k \in K} |[\nabla g_P(x^k)]_i| = 0.\end{aligned}$$

Em particular, isto é válida para todo $j = 1, \ldots, \bar{m}$. Isso implica que

$$\lim_{k \to \infty} [\nabla g_P(x^k)]_i = 0.$$

Portanto

$$\lim_{k \to \infty} \|\nabla g_P(x^k)\| = 0.$$

Assim, temos que todo ponto de acumulação é ponto estacionário.

Dado uma tolerância $\epsilon > 0$, desejamos saber o esforço computacional máximo necessário para obter um iterado x^k na qual a função objetivo seja menor que um dado alvo ou $\|g_P(x^k)\|$ seja menor que ϵ.

Teorema 4.

Suponha que a Suoisção 1 seja válida. Seja $f_{target} < f(x^0)$ *e* $\epsilon > 0$ *dados. Então, a quantidade máxima de iterações* k *para obter*

(i) $f(x^{k+1}) \leqslant f_{target}$ *ou*

(ii) $|[g_P(x^{k+1})]_i| \leqslant \epsilon$ *para todo* $i \in I_k$ *é*

$$\frac{f(x^0) - f_{target}}{c} \epsilon^{-\frac{p+1}{p}}, \tag{10}$$

onde c *depende apenas de* α, τ_2, L, p, *e* θ.

Resumo da prova: Substituindo $\left|[g_P(x^{k+1})]_i\right| > \epsilon$ em (2) obtemos

$$f(x^{k+1}) \leqslant f(x^k) - c\epsilon^{\frac{p+1}{p}}.$$

Daí e de $f(x^{k+1}) > f_{\text{target}}$ obtemos que

$$k \leqslant \frac{f(x^0) - f_{\text{target}}}{c} \epsilon^{-\frac{p+1}{p}}.$$

Observação 3.

O resultado anterior se refere a obter $|[g_P(x^{k+1})]_i| \leqslant \epsilon$ *para todos* $i \in I_k$. *Para obter* $|[g_P(x^{k+1})]_i| \leqslant \epsilon$ *para todos* $i \notin I_k$ *precisamos que as iterações consecutivas estejam suficientemente próximas.*

Teorema 5.

Suponha que a Suposição 1 seja válida. Seja $f_{target} < f(x^0)$, $\epsilon > 0$, *e* $\delta > 0$ *dados. Então, a quantidade de iterações para obter*

(i) $f(x^{k+1}) \leqslant f_{target}$ *ou*

(ii) $\|x^{k+1} - x^k\| \leqslant \delta$ *e* $|[g_P(x^{k+1})]_i| \leqslant \epsilon$ *para todo* $i \in I_k$

é no máximo

$$\frac{f(x^0) - f_{target}}{c\,\epsilon^{\frac{p+1}{p}}} + \frac{f(x^0) - f_{target}}{\alpha\,\delta^{p+1}},$$

onde c *depende apenas de* α, τ_2, L, p, *e* θ.

Resumo de prova: A prova segue direto de (ii), de $f(x^{k+1}) \leqslant f(x^k) - \alpha\|x^{k+1} - x^k\|^{p+1}$ e do Teorema 4.

Agora dividimos as iterações ciclos, onde cada ciclo possui $\overline{m}$ iterações.

- os ciclos sucessivos começam em $x^0, x^{\overline{m}}, x^{2\overline{m}}, \ldots, x^{l\overline{m}}, \ldots$.
- As iterações $k = l\overline{m}, \ldots, l\overline{m} + \overline{m} - 1$ são internas ao ciclo l.

Com a noção de ciclo em mãos, reformulamos os Teoremas 4 e 5 da seguinte forma.

Teorema 6.

Suponha que as Suposições 1 e 2 sejam válidas. Seja $f_{target} < f(x^0)$ e $\epsilon > 0$ dados. Então, o número de ciclos l para obter uma iteração interna k tal que

(i) $f(x^{k+1}) \leqslant f_{target}$ *ou*

(ii) $|[g_P(x^{k+1})]_i| \leqslant \epsilon$ *para algum $i \in I_k$ é no máximo*

$$\frac{f(x^0) - f_{target}}{c} \epsilon^{-\frac{p+1}{p}}, \tag{11}$$

onde c depende apenas de α, τ_2, L, p, e θ.

Resumo da prova: A prova segue direto do Teorema 4.

Teorema 7.

Suponha que as Suposições 1 e 2 sejam válidas. Seja $f_{target} < f(x^0)$, $\epsilon > 0$, *e* $\delta > 0$ *dados. Então, o número de ciclos* l *para obter uma iteração* k *tal que*

(i) $f(x^{k+1}) \leqslant f_{target}$ *ou*

(ii) $\|x^{k+1} - x^k\| \leqslant \delta$ *e* $|[g_P(x^{k+1})]_i| \leqslant \epsilon$ *para algum* $i \in I_k$

é no máximo

$$\frac{f(x^0) - f_{target}}{c\,\epsilon^{\frac{p+1}{p}}} + \frac{f(x^0) - f_{target}}{\alpha\,\delta^{p+1}}, \tag{12}$$

onde c *depende apenas de* α, τ_2, L, p, *e* θ.

Resumo da prova: A prova segue diretamente dos Teoremas 4-6.

A suposição a seguir garante que pequenos incrementos causam pequenas diferenças nos gradientes projetados.

Suposição 3.

Existe $L_g > 0$ *tal que para todo* $i = 1, \dots, n$ *e* $x, z \in \Omega$,

$$|[g_P(x)]_i - [g_P(z)]_i| \leqslant L_g \|x - z\|.$$

A Suposição 3 é satisfeita se o gradiente de f satisfaz a condição de Lipschitz com constante L_g.

Agora somos capazes de estabelecer um limite para o número de ciclos em que todo o gradiente projetado é maior que uma tolerância dada.

Teorema 8.

Suponha que as Suposições 1-3 sejam válidas. Seja $f_{target} < f(x^0)$, $\epsilon > 0$, *e* $\delta > 0$ *dados. Então, existe um ciclo* l *com*

$$l \leqslant \frac{f(x^0) - f_{target}}{c\,\epsilon^{\frac{p+1}{p}}} + \frac{f(x^0) - f_{target}}{\alpha\,\delta^{p+1}} + 1 \tag{13}$$

tal que:

(i) *para alguma iteração* k *interna ao ciclo* l, *temos* $f(x^k) \leqslant f_{target}$, *ou*

(ii) *para todas as iterações* k *internas ao ciclo* l, *temos que*

$$|[g_P(x^{k+1})]_i| \leqslant \epsilon + \overline{m}L_g\delta \quad \textit{para todo } i = 1, \ldots, n.$$

Resumo da prova: Pelo Teorema 7, existe um ciclo l satisfazendo (13), tal que, para cada iteração k interna ao ciclo l,

$$f(x^{k+1}) \leqslant f_{\text{target}} \text{ ou}$$

$$\|x^{k+1} - x^k\| \leqslant \delta \quad \text{e} \quad |[g_P(x^{k+1})]_i| \leqslant \epsilon \quad \text{para todo } i \in I_k. \tag{14}$$

- Se existir uma iteração k interna a l tal que $f(x^{k+1}) \leqslant f_{\text{target}}$, ok.

Neste caso o Teorema é válido.

- Suponha (14) válido e seja $i \in \{1, \ldots, n\}$ arbitrário.
  Da Suposição 2 existe uma iteração k em l tal que $i \in I_k$, daí e de (14), temos $|[g_P(x^{k+1})]_i| \leqslant \epsilon$. Para qualquer iterado z em l, temos

$$
\begin{aligned}
|[g_P(z)]_i| &\leqslant |[g_P(x^{k+1})]_i| + |[g_P(z)]_i - [g_P(x^{k+1})]_i| \\
&\leqslant \epsilon + L_g \|z - x^{k+1}\| \\
&\leqslant \epsilon + L_g \sum_{j=0}^{\overline{m}-1} \|x^{l\overline{m}+j} - x^{l\overline{m}+j+1}\| \\
&\leqslant \epsilon + \overline{m} L_g \delta.
\end{aligned}
$$

Finalizando a prova.

Substituindo ϵ por $\epsilon/2$ e definindo $\delta = \epsilon/(2\overline{m}L_g)$, o Teorema 8 pode ser reescrito da seguinte forma.

Teorema 9.

Suponha que as suposições 1, 2 e 3 sejam válidas. Seja $f_{target} < f(x_0)$, $\epsilon > 0$, *e* $\delta > 0$ *dados. Então, existe um ciclo* l *não maior que*

$$\frac{f(x_0) - f_{target}}{2^{-\frac{p+1}{p}} c} \epsilon^{-\frac{p+1}{p}} + \frac{[f(x_0) - f_{target}]\overline{m}^{p+1}}{\alpha(2\bar{L})^{-(p+1)}} \epsilon^{-(p+1)} + 1, \tag{15}$$

onde c *depende apenas de* α, τ_2, L, p *e* θ, *tal que, na sua primeira iteração interna* k,

$$f(x^k) \leqslant f_{target} \text{ ou } |[g_P(x^{k+1})]_i| \leqslant \epsilon \quad \text{para todo } i = 1, \ldots, n.$$

No resultado anterior, não levamos em consideração o número de vezes em que o parâmetro de regularização é atualizado.

Observe que:

- Por definição, a sequência de σ_k é limitada inferiormente pela sequência $0, \tau_1^0\sigma_{\min}, \tau_1^1\sigma_{\min}, \tau_1^2\sigma_{\min}, \tau_1^3\sigma_{\min}, \ldots$.
- Sendo r o número máximo de vezes em que o parâmetro de regularização é atualizado, logo

$$\tau_1^{r-2}\sigma_{\min} \leqslant L + \alpha, \text{ pois na pior das hipóteses } \tau_1^{r-1}\sigma_{\min} \geqslant L + \alpha.$$

De onde segue que

$$r \leqslant \log_{\tau_1}\left(\frac{L+\alpha}{\sigma_{\min}}\right) + 2.$$

O número de ciclos para obter $f(x^k) \leqslant f_{\text{target}}$ ou $|[g_P(x^{k+1})]_i| \leqslant \epsilon$ para todo $i = 1, \ldots, n$ é no máximo

$$\frac{f(x_0) - f_{\text{target}}}{2^{-\frac{p+1}{p}} c} \epsilon^{-\frac{p+1}{p}} + \frac{[f(x_0) - f_{\text{target}}] \overline{m}^{p+1}}{\alpha(2\bar{L})^{-(p+1)}} \epsilon^{-(p+1)} + 1. \qquad (16)$$

Alguns pontos sobre (16):

1. p cresce $\implies$ a complexidade piora, o contrário dos métodos tradicionais.
2. $\overline{m}$ cresce $\implies$ o segundo termo de (15) cresce.
3. $\bar{m}$ crescer junto com n se o tamanho dos subproblemas permanecer limitado.
4. O tipo de escolha do bloco pode influenciar na complexidade.

Com, isso, os resultado de complexidade obtido é um argumento para descartar algoritmos CD de alta ordem?

OBRIGADO A TODOS PELA PACIÊNCIA!!