Uma Breve Introdução aos Avanços dos Métodos de Descida Coordenada por Blocos

Vitaliano S. Amaral – UFPI
vitalianoamaral@ufpi.edu.br
vitalianoamaral.github.io

VI Encontro de Matemática Aplicada e Computacional - ERMAC 2025

Novembro de 2025

Ideia geral de um algoritmo de descida

Considere o seguinte problema:

$$\text{Minimizar } f(x) \quad \text{sujeito a} \quad x \in \mathbb{R}^n \tag{1}$$

onde $f : \mathbb{R}^n \to \mathbb{R}$ é diferenciável.

Objetivo: encontrar $x^\star$ tal que $f(x^\star) \leqslant f(x)$ para todo $x \in \mathbb{R}^n$.

Partimos de um ponto inicial x^0 e, a cada iteração, escolhemos uma direção d^k e um tamanho de passo $t_k > 0$ e definimos

$$x^{k+1} = x^k + t_k d^k.$$

Uma escolha razoável é determinar d^k de modo que seja uma direção de descida para f, isto é, de forma que exista $t > 0$ tal que $f(x^k + t\, d^k) < f(x^k)$.

Ideia geral de um algoritmo de descida

- Se $\nabla f(\bar{x})^\top d < 0$, então d é uma direção de descida para f em $\bar{x}$.
- Uma escolha natural é $d = -\nabla f(\bar{x})$, pois quando $\nabla f(\bar{x}) \neq 0$ temos $\nabla f(\bar{x})^\top(-\nabla f(\bar{x})) = -\|\nabla f(\bar{x})\|^2 < 0$.

Um dos métodos clássicos de otimização é o método de Cauchy, ou método do gradiente, que busca minimizar uma função ao longo da direção de maior descida.

Método do Gradiente (forma básica)

Dado $x^0 \in \mathbb{R}^n$, para $k = 0, 1, 2, \ldots$:

Passo 1 Se $\nabla f(x^k) = 0$, pare. Caso contrário ir ao Passo 2.

Passo 2 Escolha uma direção $d^k = -\nabla f(x^k)$ e ir ao Passo 3.

Passo 3 Escolha um passo $t_k > 0$ e ir ao Passo 4.

Passo 4 Defina $x^{k+1} = x^k + t_k d^k$, atualize $k \leftarrow k + 1$ e voltar ao Passo 1.

Controle no tamanho do passo

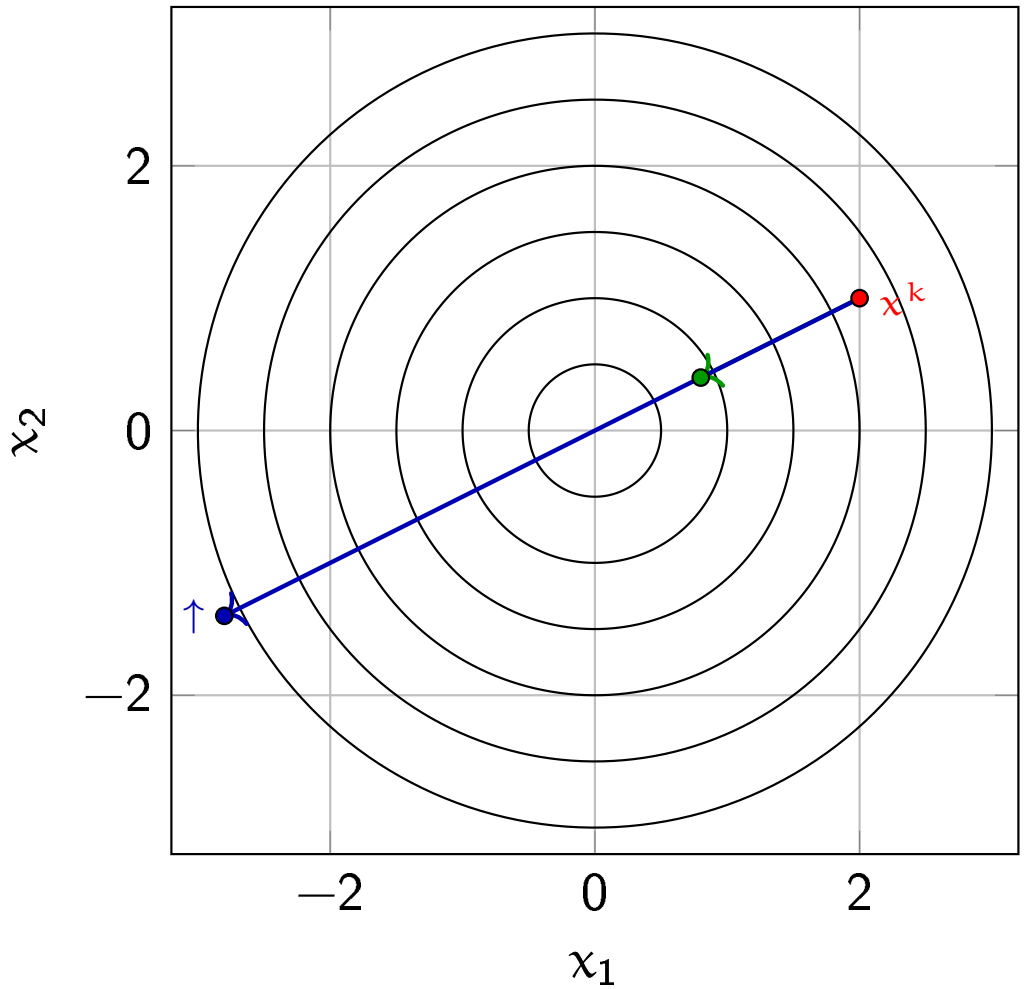


Figura: Curvas de nível de $f(x, y) = x^2 + y^2$. Um passo pequeno reduz f, enquanto um passo grande pode aumentar seu valor.

Critérios de parada usuais

Na prática, nem sempre o método encontra um ponto estacionário exato ou um ponto ótimo exato; por isso, estabelece-se um critério de parada para o método.

Critérios de parada comumente utilizados:

- $\|\nabla f(x^k)\| \leqslant \varepsilon$ (estacionariedade aproximada);
- $|f(x^k) - f(x^{k-1})| \leqslant \varepsilon_f$ (pequena variação no valor objetivo);
- $\|x^k - x^{k-1}\| \leqslant \varepsilon_x$ (pequena variação iterativa);
- $k \geqslant k_{max}$ (limite máximo de iterações).
- $f(x^k) \leqslant f_{target}$ (valor ótimo aproximado).

Métodos de Descida Coordenada por Blocos

Com o crescimento do número de problemas de grande porte, aumenta a necessidade de métodos iterativos que, na prática, apresentem esforços computacionais cada vez menores durante sua execução.

Nessa perspectiva, os métodos **BCD** (Block Coordinate Descent) têm se mostrado importantes por buscarem reduzir a dimensão do problema e, assim, melhorar consideravelmente o custo computacional.

A ideia principal dos métodos BCD é decompor um grande problema de otimização em uma sequência de problemas menores, resolvidos de forma alternada ou cíclica.

Considere o seguinte problema:

$$\text{Minimizar } f(x) \text{ sujeito a } x \in \Omega \subset \mathbb{R}^n. \tag{2}$$

Métodos de Descida Coordenada por Blocos BCD

Basicamente um método BCD para resolver o Problema (2) consiste em:

Método BCD

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. Atualizar apenas as coordenadas do bloco escolhido:

$$x_i^{k+1} = \begin{cases} x_i^k, & i \notin I_k, \\ x_i^{k+1} \text{ (valor que gera decréscimo em f,)} & i \in I_k. \end{cases}$$

Definir $k \leftarrow k + 1$ e voltar ao Passo 1.

Métodos de Descida Coordenada por Blocos BCD

- Os métodos BCD estão entre os primeiros de decomposição de variáveis, onde um dos primeiros algoritmos foi descrito para a minimização de funções quadráticas por Gauss e Seidel.
- Ideia central: decompor um grande problema em subproblemas menores.
- Por um bom tempo, esses métodos receberam pouca atenção por parte dos pesquisadores. Alguns motivos:
  - Baixo desempenho prático.
  - Poucos desafios teóricos.
- A situação mudou drasticamente nas últimas décadas. Dentre vários motivos, destacamos:
  - Aplicações em problemas de grande porte.
  - Desenvolvimento de novas teorias.

Um dos pioneiros na retomada do estudo dos métodos BCD foi o trabalho de Beck e Tetruashvili[1].

Problema considerado:

$$\min f(x) \quad \text{s.a.} \quad x \in \mathbb{R}^n.$$

- Os autores assumiram f convexa,
- Gradiente parcial $\nabla_{(i)} f$ Lipschitz.

Método do Gradiente Coordenado por Blocos - BCGD

Tome $x^0 \in \mathbb{R}^n$ e L_i $(i = 1, \ldots, p)$ constantes de Lipschitz do gradiente parcial $\nabla_{(i)} f$.

Passo 1. Tome $x^{k,0} = x^k$ e defina recursivamente:

$$x^{k,i} = x^{k,i-1} - \frac{1}{L_i} U_i \nabla_{(i)} f(x^{k,i-1}), \quad i = 1, \ldots, p$$

Passo 2. Defina $x^{k+1} = x^{k,p}$

[1] A. Beck, and L. Tetruashvili (2013). - *On the convergence of block coordinate descent type methods.* SIAM Journal on Optimization, 23(4):20372060.

Convergência do método BCGD

Assumindo f convexa e condições de Lipschitz nos gradientes parciais, os autores provaram que a sequência $\{f(x^k)\}$ satisfaz as seguintes condições:

- Para todo $k = 0, 1, 2, \ldots$

$$f(x^k) - f(x^{k+1}) \geqslant \frac{1}{4L_{max}(1 + pL^2/L_{min}^2)} \|\nabla f(x^k)\|^2$$

$$f(x^k) - f(x^*) \leqslant 4L^2\left(1 + pL^2/L_{min}^2\right) R^2(x^0) \frac{1}{k + 8/p}, \quad k = 0, 1, 2, \ldots\ldots,$$

onde L_{max} e L_{min} são as constantes de Lipschitz máxima e mínima dos p-blocos, respectivamente,

$$R(x_0) \equiv \max_{x \in \mathbb{R}^n} \max_{x^* \in X^*} \{\|x - x^*\| : f(x) \leqslant f(x_0)\}$$

e $f(x^*)$ é um valor ótimo de f.

Método BCGD no $\mathbb{R}^2$

Tome $x^0 \in \mathbb{R}^2$ e L_i $(i = 1, 2)$ constantes de Lipschitz de $\nabla_{(i)} f$.

Passo 1. Tome $x^{k,0} = x^k = (x_1^{k,0}, x_2^{k,0})$ e defina recursivamente:

$$x^{k,1} = x^{k,0} - \frac{1}{L_1} e_1 \frac{\partial f(x^{k,0})}{\partial x_1} = \left(x_1^{k,0} - \frac{1}{L_1} \frac{\partial f(x^{k,0})}{\partial x_1}, \, x_2^{k,0} \right)$$

$$x^{k,2} = x^{k,1} - \frac{1}{L_2} e_2 \frac{\partial f(x^{k,1})}{\partial x_2} = \left(x_1^{k,1}, \, x_2^{k,1} - \frac{1}{L_2} \frac{\partial f(x^{k,1})}{\partial x_2} \right)$$

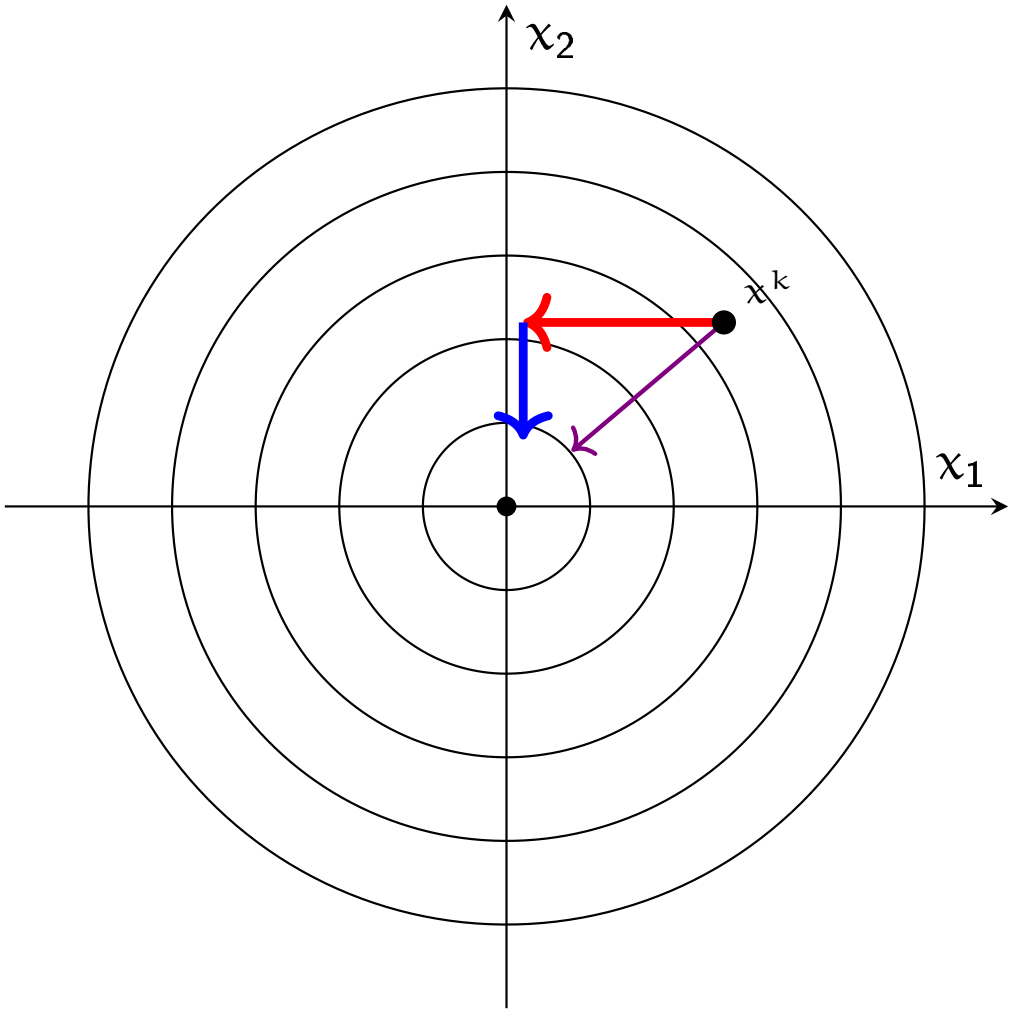
Passo 2. Defina

$$x^{k+1} = x^{k,2} = \left(x_1^{k,0} - \frac{1}{L_1} \frac{\partial f(x^{k,0})}{\partial x_1}, \, x_2^{k,1} - \frac{1}{L_2} \frac{\partial f(x^{k,1})}{\partial x_2} \right).$$

- Vale destacar que os subproblemas resolvidos no Passo 1 são resolvidos na reta real.

BCD vs. gradiente clássico

Curvas de nível de $f(x, y) = x^2 + y^2$



Teste do BCGD em Problemas de Mínimos Quadrados

Considere a função objetivo $f : \mathbb{R}^2 \to \mathbb{R}$, tal que:

$$f(x) = \frac{1}{2}\|Ax - b\|^2, \quad \text{onde } A = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}_{2\times 2} \text{ e } b = \begin{pmatrix} 1 \\ 1 \end{pmatrix}_{2\times 1}$$

Temos $\nabla f(x) = A^T(Ax - b)$ e $L_i = 9$. Consideramos $x^0 = \left(1, \frac{2}{3}\right)$.

	Bloco $i = 1$		Bloco $i = 2$		
It.	$U_1 \nabla_1 f(x^{k,0})$	$x^{k,1}$	$U_2 \nabla_2 f(x^{k,1})$	$x^{k,2} = x^{k+1}$	$f(x^{k+1})$
1	(4.6666, 0)	(0.4814, 0.6666)	(0, 2.2592)	(0.4814, 0.4156)	0.120578
2	(1.0699, 0)	(0.3625, 0.4156)	(0, 0.5285)	(0.3625, 0.3569)	0.006290
3	(0.2406, 0)	(0.3358, 0.3569)	(0, 0.1279)	(0.3625, 0.3569)	0.000329
4	(0.0500, 0)	(0.3303, 0.3426)	(0, 0.0346)	(0.3303, 0.3388)	0.000032
5	(0.0068, 0)	(0.3295, 0.3388)	(0, 0.0123)	(0.3295, 0.3374)	0.000016
6	(-0.0024, 0)	(0.3298, 0.3374)	(0, 0.0065)	(0.3298, 0.3367)	0.000012
7	(-0.0040, 0)	(0.3302, 0.3367)	(0, 0.0047)	(0.3302, 0.3362)	0.000009

Tabela: Resultados do Método BCGD em f para $k = 0, 1, 2, 3, 4, 5, 6, 7$.

Um outro trabalho importante na retomada dos estudos método BCD foi o trabalho de S. J. Wright[2].

- O artigo descreve os fundamentos da abordagem, variantes, extensões e propriedades de convergência, sobretudo no caso de funções convexas.
- Dá atenção especial a estruturas que surgem frequentemente em aprendizado de máquina, mostrando implementações eficientes de versões aceleradas.
- Além disso, apresenta variantes paralelas e discute suas propriedades de convergência em diferentes modelos de execução.

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

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)

Amaral, V. (2025). - *A partially derivative-free cyclic block coordinate descent method for nonseparable composite optimization.* Mathematical Modelling and Analysis, 30(3), 535-552.

No primeiro trabalho[3] foi considerado o seguinte problema:

$$\text{Minimize } f(x) \text{ s.a. } x \in \Omega, \tag{3}$$

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

- 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 $\bar{x}$.

---

[3] 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

Introduzimos o seguinte método para resolver o Problema (11):

Algoritmo 1.

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.*

Boa definição

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

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

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

Para satisfazer as condições anteriores, $M_{\bar{x}}(\cdot)$ pode ser Taylor de ordem p de f ao redor de $\bar{x}$ se as derivadas de ordem p de f satisfazem condições de Lipschitz ou a própria f.
Provamos que par $\sigma_k \geqslant L + \alpha$, o ponto x^{trial} satisfaz

$$f(x^{trial}) \leqslant f(x^k) - \alpha\|x^{trial} - x^k\|^{p+1} \tag{4}$$

e

$$\|\nabla g_{P,I_k}(x^{trial})\| \leqslant \big(L + \tau_2(L + \alpha)(p + 1) + \theta\big)\|x^{trial} - x^k\|^p, \tag{5}$$

Garantindo a boa definição do método.

Convergência

Provamos que a sequência $\{x^k\}$ gerada pelo Algoritmo 1 satisfaz as seguintes condições:

$$\lim_{k\to\infty} \|x^{k+1} - x^k\| = 0, \tag{6}$$

$$\lim_{k\to\infty} \|\nabla g_{P,I_k}(x^{k+1})\| = 0, \tag{7}$$

e

$$\lim_{k\to\infty} \|\nabla g_{P,I_k}(x^k)\| = 0. \tag{8}$$

Observe que os resultados anteriores não garantem

$$\lim_{k\to\infty} \|g_P(x^k)\| = 0. \tag{9}$$

Para provar (9) foi necessário uma suposição adicional em relação a escolha dos blocos.

Suposição 1.

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 1 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.

Com essa suposição adicional, mostramos que a sequência $\{x^k\}$ satisfaz,

$$\lim_{k \to \infty} \|\nabla g_P(x^k)\| = 0. \tag{10}$$

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

A complexidade do método

Dado as estimativas $f_{\text{target}} < f(x^0)$ e $\epsilon > 0$ dados. Provamos que o número de iterações k para obter

$$f(x^{k+1}) \leqslant f_{\text{target}} \text{ ou } |[g_P(x^{k+1})]_i| \leqslant \epsilon \text{ para todo } i \in I_k$$

é no máximo

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

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

Podemos observar que a estimativa anterior para o gradiente projetado se refere apenas às coordenadas escolhidas no bloco.

No entanto, obter $|[g_P(x^k)]_i| \leqslant \varepsilon$ para todos $i \notin I_k$ é mais difícil, pois para este propósito, é necessário que as iterações consecutivas estejam suficientemente próximas.

Para isso, foi assumido a seguinte suposição.

A complexidade do método

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

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

Agora, considerando a suposição adicional anterior, provamos que o número máximo de iterações necessário para obter $f(x^k) \leqslant f_{\text{target}}$ ou $|[g_P(x^k)]_i| \leqslant \varepsilon$ para todo $i = 1, \ldots, n$ é

$$\frac{f(x^0) - f_{\text{target}}}{c(\frac{1}{2})^{\frac{p+1}{p}}} \varepsilon^{-\frac{p+1}{p}} + \frac{f(x^0) - f_{\text{target}}}{\alpha(\frac{1}{2\bar{m}L_g})^{p+1}} \varepsilon^{-(p+1)} + 1.$$

Assim, a complexidade obtida é da ordem $O(\epsilon^{-(p+1)})$.

A ordem da complexidade obtida chama a atenção, pois piora quando aumentamos a ordem de regularização, ao contrário do que ocorre em métodos tradicionais.

Comentários

Alguns pontos sobre

$$\frac{f(x^0) - f_{\text{target}}}{c(\frac{1}{2})^{\frac{p+1}{p}}} \varepsilon^{-\frac{p+1}{p}} + \frac{f(x^0) - f_{\text{target}}}{\alpha(\frac{1}{2\bar{m}L_g})^{p+1}} \varepsilon^{-(p+1)} + 1.$$

1. p cresce $\Longrightarrow$ a complexidade piora, o contrário dos métodos tradicionais.
2. $\overline{m}$ cresce $\Longrightarrow$ o segundo termo da expressão acima 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.

No trabalho[4] foi considerado o seguinte problema:

$$\text{Minimizar } F(x) := f(x) + h(x) \text{ sujeito a } x \in \mathbb{R}^n, \tag{11}$$

onde f diferenciável e $h : \mathbb{R}^n \to \mathbb{R}$ é uma função convexa.

Quando h é a função indicadora de um conjunto convexo C, então o Problema (11) é equivalente a resolver o problema de minimizar $F(x) = f(x)$ restrito ao conjunto convexo C.

Muitos problemas de otimização em aplicações do mundo real são frequentemente de grande escala e alta dimensão, o que os torna desafiadores de resolver, especialmente quando há derivadas difíceis de calcular.

Neste trabalho foi proposto uma versão dos métodos BCD parcialmente livre de derivadas.

---

[4] V. S. Amaral. - *A partially derivative-free cyclic block coordinate descent method for nonseparable composite optimization.* Mathematical Modelling And Analysis, v. 30, p. 535-552, 2025.

Para o desenvolvimento do método proposto, foi considerado a função

$$\varphi_f : \mathbb{R}^n \times [0,1] \to \mathbb{R}^n \text{ onde } \lim_{\lambda \to 0} \varphi_f(x,\lambda) = \nabla f(x). \qquad (12)$$

A função em (12) pode ser definida como $\varphi_f = \nabla f$. No entanto, essa definição não é recomendada quando o cálculo do gradiente de f é computacionalmente proibitivo. Nesses casos, é preferível definir φ_f como uma aproximação do gradiente ∇f, que pode ser obtida com um custo computacional menor.

A função φ_f pode ser definido por uma das seguintes maneiras:

- $\varphi_f(x,\lambda) = \left[\frac{f(x+\lambda e_1) - f(x)}{\lambda}, \cdots, \frac{f(x+\lambda e_n) - f(x)}{\lambda}\right].$
- $\varphi_f(x,\lambda) = \left[\frac{f(x) - f(x-\lambda e_1)}{\lambda}, \cdots, \frac{f(x) - f(x-\lambda e_n)}{\lambda}\right].$
- $\varphi_f(x,\lambda) = \left[\frac{f(x+\lambda e_1) - f(x-\lambda e_1)}{2\lambda}, \cdots, \frac{f(x+\lambda e_n) - f(x-\lambda e_n)}{2\lambda}\right].$

Seja $x^0 \in \mathbb{R}^n$; para cada $i = 1, \ldots, q$, uma matriz simétrica semidefinida positiva $B_{(i)}(x^0) \in \mathbb{R}^{n_i \times n_i}$; tome $\alpha \in (0,1)$, $\epsilon \in (0,1)$, $\sigma_0 \geqslant 1$ e $F_{\text{target}} \in \mathbb{R}$. Inicialize $k \leftarrow 0$.

Etapa 1: Escolha $\lambda_k \in \left[0, \ \frac{\epsilon}{\sigma_k \sqrt{n}}\right]$, e considere $\varphi_f(x^k, \lambda_k)$.

Etapa 2: Ponha $x^{k,0} = x^k$ e, para cada $i = 1, \ldots, q$, calcule $x^{k,i} = x^{k,i-1} + U_i s^k_{(i)}$, onde $s^k_{(i)} \in \mathbb{R}^{n_i}$ é solução de

$$\min_{s \in \mathbb{R}^{n_i}} \ \langle U_i^T \varphi_f(x^{k,i-1}, \lambda_k), s\rangle + \frac{1}{2}\langle B_{(i)}(x^{k,i-1})s, s\rangle + h(x^{k,i-1} + U_i s) - h(x^{k,i-1}) + \frac{\sigma_k}{2}\|s\|^2.$$

Etapa 3: Se

$$\left\| \sum_{i=1}^{q} U_i s^k_{(i)} \right\|_\infty < \frac{\epsilon}{\sigma_k} \quad \text{ou} \quad F(x^{k,q}) \leqslant F_{\text{target}},$$

pare e declare $x^{k,q}$ como uma solução aceitável. Caso contrário, vá para a Etapa 4.

Etapa 4: Se

$$F(x^{k,q}) \leqslant F(x^k) - \frac{\alpha}{\sigma_k}\,\epsilon^2,$$

faça $k \leftarrow k+1$, defina $x^{k+1} = x^{k,q}$ e $\sigma_{k+1} = \sigma_k$, escolha $B_{(i)}(x^{k+1}) \in \mathbb{R}^{n_i \times n_i}$ simétrica s.d.p. e volte à Etapa 1. Caso contrário, defina $\sigma_k \leftarrow 2\sigma_k$ e volte à Etapa 1.

Boa definição

Para garantir a boa definição do método foi considerado que:

Suposição 2.

Existem $L, M \in (0, +\infty)$ *e* $\beta \in (0, 1]$ *tais que para todo* $x, y \in \mathbb{R}^n$

$$f(y) \leqslant f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|y - x\|^2 + \frac{M}{\beta + 1}\|y - x\|^{\beta+1} \tag{13}$$

e

$$\|\nabla f(x) - \varphi_f(x, \lambda)\| \leqslant \frac{\sqrt{n}L}{2}\lambda + \frac{\sqrt{n}M}{\beta + 1}\lambda^{\beta} \tag{14}$$

Se $f := g_1 + g_2$, onde ∇g_1 L-Lipschitz e ∇g_2 M-Hölder com expoente β então (13) e (14) são válida, pois

$$g_1(y) \leqslant g_1(x) + \langle \nabla g_1(x), y - x \rangle + \frac{L}{2}\|y - x\|^2 \ \forall x, y \in \mathbb{R}^n$$

e

$$g_2(y) \leqslant g_2(x) + \langle \nabla g_2(x), y - x \rangle + \frac{M}{\beta + 1}\|y - x\|^{\beta+1} \ \forall x, y \in \mathbb{R}^n.$$

Boa definição

Além das suposições anterior foi considerado que: existe $\overline{B} \geqslant 1$ tal que $\|B_{(i)}(x^{k,i-1})\| \leqslant \overline{B}$ para todos os $i \in \{1, \ldots, q\}$ e k.

Assim, provamos que se

$$\|s^k\|_\infty \geqslant \frac{\epsilon}{\sigma_k} \text{ e } \sigma_k \geqslant \left[\frac{2L + q\overline{B}}{(1-\alpha)} + \frac{2q(Mn^{\frac{1-\beta}{2}} + M)}{(\beta+1)(1-\alpha)}\epsilon^{\beta-1}\right]^{\frac{1}{\beta}}$$

então

$$F(x^{k,q}) \leqslant F(x^k) - \alpha\frac{\sigma_k}{2}\|s^k\|_\infty^2 \tag{15}$$

e

$$F(x^{k,q}) \leqslant F(x^k) - \frac{\alpha}{\sigma_k}\epsilon^2. \tag{16}$$

Garantindo assim a boa definição do método.

Complexidade

- Números de Iterações é limitado superiormente por:

$$\frac{c\left(F(x^0) - F_{target}\right)}{\alpha} \epsilon^{-\frac{\beta+1}{\beta}}.$$

- Número de avaliações de f e seus subdiferenciais é limitado superiormente por:

$$\frac{c\left(F(x^0) - F_{target}\right)}{\alpha} \epsilon^{-\frac{\beta+1}{\beta}} + \log_2\left(\frac{\sigma_{max}}{\sigma_{min}}\right).$$

com

$$\sigma_{max} = 2\epsilon^{\frac{\beta-1}{\beta}} \max\left\{\sigma_{min}, \left[\frac{2L + q\overline{B}}{(1-\alpha)} + \frac{2q(Mn^{\frac{1-\beta}{2}} + M)}{(\beta+1)(1-\alpha)}\right]^{\frac{1}{\beta}}\right\}.$$

Aplicações

O método BCDC-Dfree pode resolver problemas:

- **Caso diferenciável:** se $h \equiv 0$

$$\min_x f(x).$$

- **o problema do lasso**

$$\min_{x \in \mathbb{R}^n} \left\{ \frac{1}{2}\|Ax - b\|^2 + \lambda \|x\|_1 \right\}, \ \lambda > 0, \ A \in \mathbb{R}^{s \times n}, \tag{17}$$

considerando $h(x) = \lambda\|x\|_1$, $f(x) = \frac{1}{2}\|Ax - b\|^2$.

- **o problema**

$$\min \left\{ \|Ax - b\|^2 : 0 \leqslant x_i \leqslant 1, i = 1, 2, \ldots, n \right\} \tag{18}$$

onde $A \in \mathbb{R}^{s \times n}$ e $b \in \mathbb{R}^s$.
Basta minimizar $F(x) = f(x)$ com $f(x) = \delta_\Omega(x)$,
$\Omega = \{x \in \mathbb{R}^n : 0 \leqslant x_i \leqslant 1, i = 1, 2, \ldots, n\}$, $f(x) = \|Ax - b\|^2$.

- **Problema de mínimos quadrados penalizado por norma L_p $(1 < p < 2)$:**

$$\min F(x) = \frac{1}{2}\|A(x) - b\|_2^2 + \frac{\lambda}{p}\|\Phi(x)\|_p^p,$$

onde $A : \mathbb{R}^n \to \mathbb{R}^m$ é diferenciável (não necessariamente linear), $b \in \mathbb{R}^m$, e Φ é operador linear.

Detalhes em:


A. Bernigaud, S. Gratton and E. Simon, A non-linear conjugate gradient in dual space for L_p-norm regularized non-linear least squares with application in data assimilation, *Numerical Algorithms*, **95**, 471–497 (2024). https://doi.org/10.1007/s11075-023-01578-x

Testes Numéricos

Aplicamos nosso método proposto para resolver alguns problemas em que ambos foi considerados como um caso particular de

$$F(x) = \frac{1}{2}\|A(x) - b\|_2^2 + \frac{\lambda}{p}\|\Phi(x)\|_p^p, \tag{19}$$

com $1 < p < 2$, A diferenciável e Φ linear.

Nos exemplos a seguir, consideramos $B_{(i)} \equiv 0$, $n = 10$, $\alpha = 0,5$, $\epsilon = 5 \times 10^{-5}$, $F_{\text{target}} = 0,5$ e a função $\varphi_f : \mathbb{R}^{10} \times \mathbb{R} \to \mathbb{R}^{10}$ definida por

$$\varphi_f(x, \lambda) = \left[\frac{f(x + \lambda e_1) - f(x)}{\lambda}, \cdots, \frac{f(x + \lambda e_{10}) - f(x)}{\lambda}\right].$$

Exemplo Numérico: Influência de σ_0

A desigualdade em (16) mostra que valores muito grandes de σ_0 podem gerar pequenas reduções na função objetivo, demandando mais iterações para alcançar uma solução aceitável.

Consideramos o seguinte problema de mínimos quadrados lineares, penalizado por uma norma L_p, com $1 < p < 2$:

$$\min F(x) = \frac{1}{2}\|Ax - b\|_2^2 + \frac{\lambda}{p}\|\Phi(x)\|_p^p,$$

onde $A \in \mathbb{R}^{10\times 10}$ e $b \in \mathbb{R}^{10}$ são fixados.

No algoritmo utilizamos $f(x) = 0.5\|Ax - b\|^2 + 5 \times 10^{-5} \sum_{i=1}^{10} |x_i|^{3/2}$, $h(x) = 0$ e $x^0 = 0 \in \mathbb{R}^{10}$.

λ_k	N. Blocos	σ_0	N. Iterações	Approx. F*
$\frac{\epsilon}{\sigma_k\sqrt{10}}$	10	1	10	0.4731
	10	500	40	0.3021
	10	1000	83	0.3003

Tabela: Comportamento do BCDC-Dfree para diferentes σ_0.

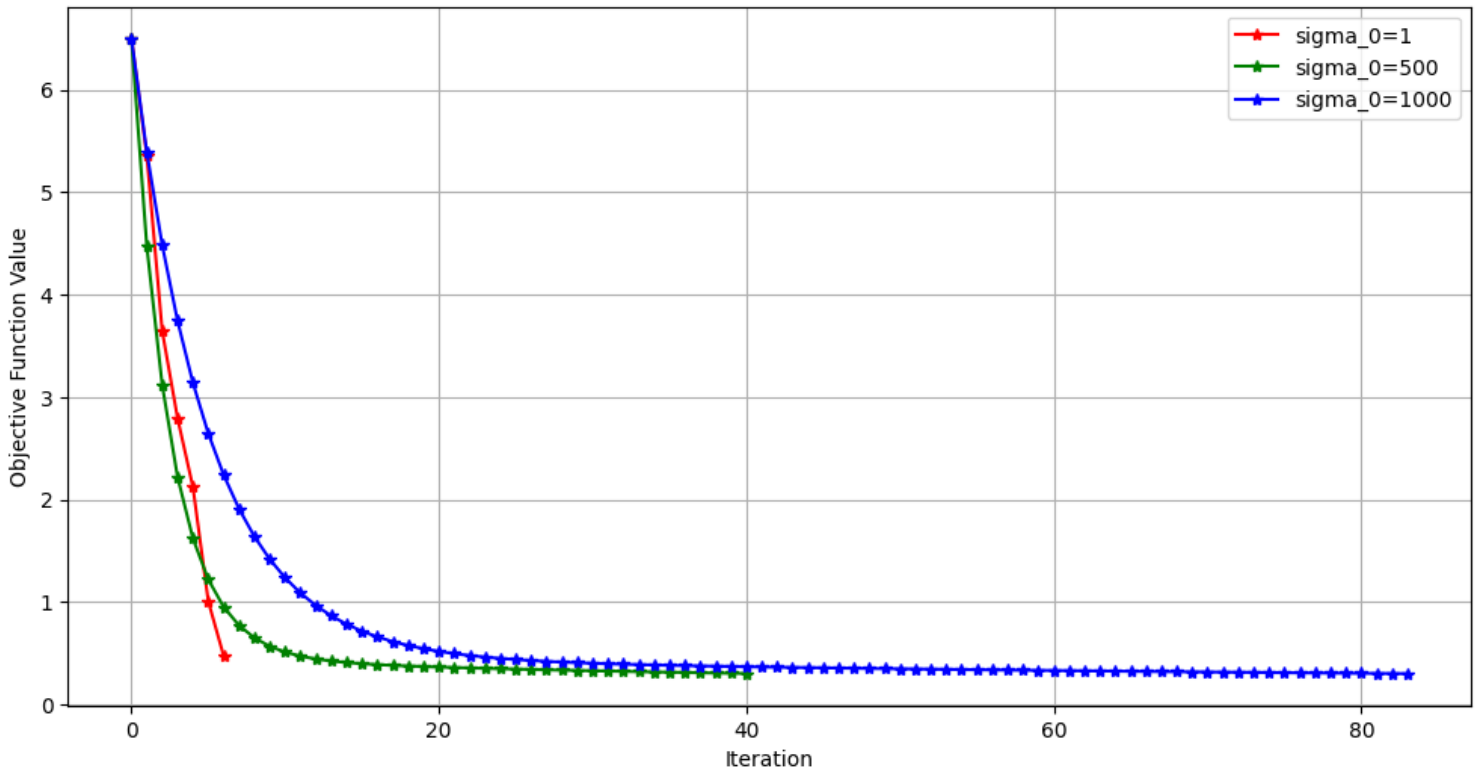


Figura: Evolução do BCDC-Dfree para diferentes valores de σ_0.

Exemplos Numéricos com A e b Aleatórios

Consideramos o Problema (11) com f dado em (19), onde $A \in \mathbb{R}^{10\times 10}$ e $b \in \mathbb{R}^{10}$ são gerados aleatoriamente, Φ é a função identidade e h é nulo. Foram avaliadas 100 instâncias distintas, todas com x^0 aleatório, comparando a abordagem com 10 blocos e com um único bloco.

Blocos	σ_0	Iterações Médias	Média-F
10	1	11.55	0.4329
1	1	17.56	0.4682

Tabela: BCDC-Dfree com A e b aleatórios.

Em seguida, σ_0 foi sorteado no intervalo $[1, 1000]$, mantendo os demais dados inalterados.

Blocos	σ_0	Iterações Médias	Média-F
10	rand()	805.45	0.4988
1	rand()	809.02	0.4988

Tabela: BCDC-Dfree com A, b e σ_0 aleatórios.

Exemplo com A Não Linear

Agora, consideramos o caso em que A em (19) é não linear, definido por

$$A(x) = (A_1(x), \ldots, A_n(x)), \quad A_i(x) = \frac{x_i}{1 + x_{i+1}^2}, \ i = 1, \ldots, n-1,$$

com $x_{n+1} = x_1$. Tomamos $\Phi(x) = Bx$, com $B \in \mathbb{R}^{m \times n}$ aleatório e b também aleatório. Foram avaliadas 100 instâncias, todas com x^0 aleatório e σ_0 sorteado em $[1, 100]$.

Blocos	σ_0	Iterações Médias	Média-F
10	rand()	648.96	0.5008
1	rand()	664.61	0.5008

Tabela: BCDC-Dfree para o caso com A não linear.

OBRIGADO!

vitalianoamaral@ufpi.edu.br

https://vitalianoamaral.github.io

vitalianoamaral.github.io

sbmac.org.br

CCN

Centro de Ciências da Natureza - UFPI

SBMAC