UM ESTUDO SOBRE MÉTODOS DE MÁXIMA DESCIDA E ACELERAÇÕES

Gearlisson dos Santos Mendonça

Orientador: Profº Dr. Douglas Soares Gonçalves

Universidade Federal de Santa Catarina, SC, Brasil

gearlisson.sm@grad.ufsc.br

I ENCONTRO DE PÓS-GRADUANDOS EM MATEMÁTICA NA UFSC

UNIVERSIDADE FEDERAL DE SANTA CATARINA

Palavras-Chave: Máxima Descida, Aceleração de Nesterov, Gradiente Espectral.

Os **métodos de direções de descida** são métodos iterativos para resolução de problemas de otimização irrestrita:

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

$f : \mathbb{R}^n \to \mathbb{R}$ é continuamente diferenciável;

$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|, \quad \forall x, y \in \mathbb{R}^n$;

$x^* \in \mathbb{R}^n$ é a solução do problema quando $f(x^*) \leq f(x), \forall x \in \mathbb{R}^n$.

Tais métodos, a partir de um ponto inicial $x_0 \in \mathbb{R}^n$ dado, geram uma sequência $\{x_k\}_{k=0}^{\infty} \subset \mathbb{R}^n$ dada por:

$$x_{k+1} = x_k + t_k d_k, \tag{1}$$

em que $d_k \in \mathbb{R}^n$ é direção de descida tal que $\nabla f(x_k)^T d_k < 0$ e $t_k > 0$ é o tamanho de passo, em geral escolhido de modo a garantir $f(x_{k+1}) < f(x_k)$.

Neste trabalho damos ênfase ao método de Cauchy [3], e variações recentes, como o método de Nesterov [1] e o método do gradiente espectral [2]. Para estes métodos estudamos tanto a complexidade de pior caso quanto o desempenho prático.

O **método de Cauchy** utiliza em (1) $d_k = -\nabla f(x_k)$ e o $t_k > 0$ pode ser calculado de diferentes maneiras. Por exemplo, $t_k = 1/L$, com $L > 0$ e uma outra escolha possível é dada pela busca linear exata: $t_k = \arg\min \{f(x_k - t\nabla f(x_k)) \mid t > 0\}$.

Se f é convexa, é possível mostrar que $\{x_k\}_k$ gerada por este método, com $t_k = 1/L$ satisfaz,

$$f(x_k) - f(x_*) \leq \frac{L\|x_0 - x_*\|^2}{2k} \tag{2}$$

Assim, dizemos que este método tem taxa de convergência $\mathcal{O}(1/k)$.

No **método de Nesterov**, a iteração é descrita:

$$\begin{cases} \text{Se } k \geq 1: \ \text{escolha } \theta_k \in (0,1) \ \text{tal que } \dfrac{(1-\theta_k)\, t_k}{\theta_k^2} \leq \dfrac{t_k - 1}{\theta_{k-1}^2}. \\ y = (1-\theta_k)\, x_k + \theta_k v_k \\ x_{k+1} = y - t_k \nabla f(y) \\ v_{k+1} = x_k + \dfrac{1}{\theta_k}(x_{k+1} - x_k) \end{cases}$$

Além das escolhas de t_k já discutidas para o método de Cauchy, para este método podemos considerar a busca linear inexata ou *backtracking* em que $t_k > 0$ deve satisfazer

$$f(x_{k+1}) \leq f(y) - \frac{t_k}{2}\|\nabla f(y)\|^2.$$

Em comparação com (2), para este método com $t_k = t \in (0, 1/L]$ e com $\theta_k = 2/(k+2)$ satisfaz (veja [1, Teorema 5.1]):

$$f(x_k) - f(x^*) \leq \frac{2}{t(k+1)^2}\|x_0 - x^*\|^2 \tag{3}$$

Dessa forma, esse método garante a taxa de convergência $\mathcal{O}(1/k^2)$ para métodos de primeira ordem.

Por sua vez, no **método do gradiente espectral**, introduzido por [2], em (1) é usada $d_k = -\lambda_k^{-1}\nabla f(x_k)$, em que o escalar

$$\lambda_k = \max\left\{\delta_{\min}, \min\left\{\delta_{\max}, \frac{(x_k - x_{k-1})^T(\nabla f(x_k) - \nabla f(x_{k-1}))}{(x_k - x_{k-1})^T(x_k - x_{k-1})}\right\}\right\}$$

e que $0 < \delta_{\min} < \delta_{\max} < \infty$ são salva-guardas.

O t_k é determinado usando busca linear não-monótona: Dados $\gamma \in (0,1)$, $M \in \mathbb{Z}_+^*$, $0 \leq m_k \leq \min\{m_{k-1}+1, M\}$ escolha $t_k > 0$ que satisfaça a condição:

$$f(x_k + t_k d_k) < \max_{0 \leq j \leq m_k} f(x_{k-j}) + t_k \gamma \nabla f(x_k)^T d_k.$$

Para este método, se f é convexa, com valor ótimo $f(x^*)$, então para $\varepsilon \in (0,1]$, o algoritmo da busca linear não-monótona leva (veja [4, Teorema 2])

$$\left[\left(\frac{(\beta\bar{t})^{-1}\|x_0 - x^*\|^2 + \rho^{-1}\delta_{\max}[(f(x_0) - f(x^*)) + (\sum_{k=0}^{+\infty} \upsilon_k)]}{2\delta_{\min}}\right)\varepsilon^{-1}\right]$$

iterações para garantir $f(x_k) - f(x^*) \leq \varepsilon$ em no máximo $\mathcal{O}(1/k)$.

Para avaliar o desempenho prático dos métodos, estes foram implementados em Matlab e consideramos experimentos envolvendo funções quadráticas com Hessiana simétrica definida positiva com dimensão n variando de 2 a 5000. Na implementação utilizamos como critério de parada $\|\nabla f(x_k)\| < 10^{-6}$ ou um número máximo de iterações $K = 100000$.

	Método de Cauchy		Método de Nesterov		Gradiente Espectral
n	$1/L$	Busca Exata	$1/L$	Backtracking	Não-monótona
2	13,4	9,2	15,6	2,4	4
5	41,4	34,4	35	20,8	19,4
10	85	64,6	62,2	33,6	29,2
50	379,8	337,8	213	187,4	71,6
100	826,6	716,4	321	359,2	133
500	4387	3318,6	1279,4	1047	349,4
1000	13459	6635,3	4791,6	2985,2	546
5000	33871,8	25783,5	18543	12752	1692

Tabela 1: Média das iterações com diferentes distribuição de autovalores.

Os resultados numéricos indicam que o método de gradiente espectral, em média, alcança o primeiro critério de parada em menos iterações que os demais. A Figura 1 ilustra os valores funcionais ao longo das iterações para um problema de dimensão $n = 100$. Neste gráfico observamos que todos os métodos foram capazes de reduzir o valor funcional até próximo do valor ótimo $f(x^*) = 0$. Ainda, observa-se que os métodos de Cauchy garantem um decréscimo monótono da função, em oposição aos demais.

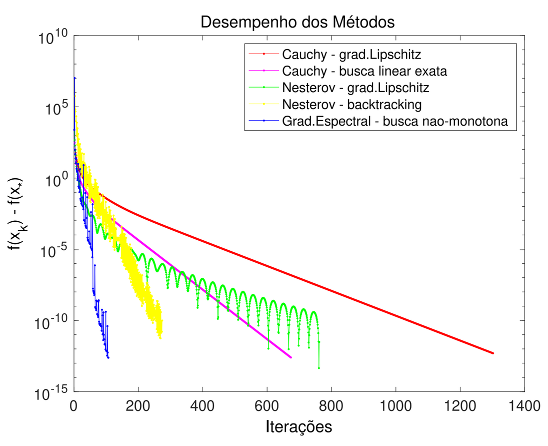

Figura 1: Desempenho dos métodos

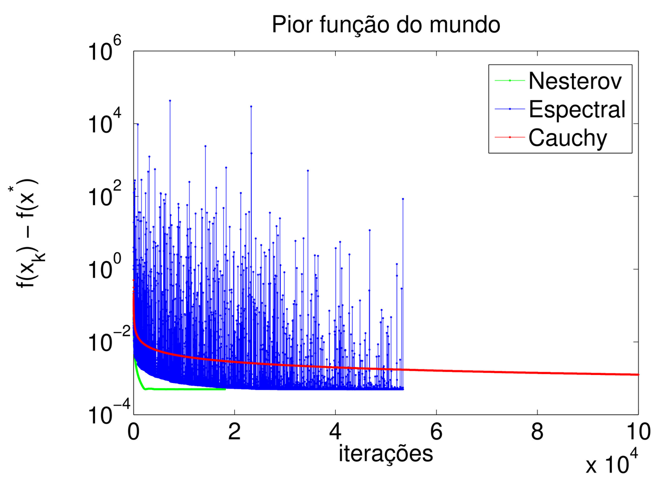

Figura 2: Pior Função do Mundo

A fim de evidenciar a performance de pior caso dos métodos, também fizemos testes com a "pior função do mundo" [1, Página 57] que é uma quadrática com Hessiana Tridiagonal definida da forma

$$f_q(x) = \frac{L}{4}\left\{\frac{1}{2}\left[x_1^2 + \sum_{i=1}^{q-1}(x_i - x_{i+1})^2 + x_q^2\right] - x_1\right\}, \ para\ q = 1, \ldots, n.$$

Neste teste, consideramos $n = 2001$, $q = 1000$ e $K = 10^5$. A Figura 2 ilustra a diferença $f(x_k) - f(x^*)$ ao longo das iterações (veja [1, Página 59]). Aqui fica claro o melhor desempenho do método de Nesterov.

Referências

[1] Y. Nesterov. *Introductory lectures on convex optimization: A basic course*, Springer Science & Business Media, 2003.

[2] J. Barzilai and J. M. Borwein.*Two-point step size gradient methods*. IMA Journal of Numerical analysis. v.8, p.141-148, 1988.

[3] A. A. Ribeiro and E. W. Karas. *Otimização Contínua: aspectos teóricos e computacionais*. Cengage Learning Editora, 2014.

[4] G. N. Grapiglia; E.W. Sachs. *On the worst-case evalution complexity of non-monotone line search algorithms*. Computacional Optimization and Applications, 68(3), 555–577, 2017.