Proceeding Series of the Brazilian Society of Computational and Applied Mathematics, v. 10, n. 1, 2023.

Trabalho apresentado no XLII CNMAC, Universidade Federal de Mato Grosso do Sul - Bonito - MS, 2023

Proceeding Series of the Brazilian Society of Computational and Applied Mathematics

Um estudo sobre métodos de máxima descida e acelerações

Gearlisson dos Santos Mendonça[1] Douglas Soares Gonçalves[2]
Departamento de Matemática, CFM, UFSC, Florianópolis, SC

Os **métodos de direções de descida** são métodos iterativos para resolução de problemas de otimização irrestrita: minimizar $f(x)$, sujeito a $x \in \mathbb{R}^n$, em que $f : \mathbb{R}^n \to \mathbb{R}$ é uma função continuamente diferenciável e com gradiente Lipschitz, de constante $L > 0$. 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 discutiremos sobre dois métodos nos quais a direção d_k é um múltiplo de $-\nabla f(x_k)$, a saber, o método de Cauchy [1] e o método do gradiente espectral [2]. Além disso, estudaremos também o método de Nesterov [1]. Para estes métodos investigamos tanto a complexidade de pior caso quanto o desempenho prático.

O **método de Cauchy**, também conhecido como método de máxima descida, utiliza em (1) $d_k = -\nabla f(x_k)$ e o tamanho de passo $t_k > 0$ pode ser calculado de diferentes maneiras. Por exemplo, se a constante de Lipschitz L é conhecida, podemos usar $t_k = 1/L$ e outra escolha possível é dada pela busca linear exata: $t_k = \arg\min \{f(x_k - t\nabla f(x_k)) \mid t > 0\}$. Para este método, é possível mostrar (veja [1, Teorema 1.2.4]) que quando f é também convexa, com minimizador global $x^* \in \mathbb{R}^n$ e usamos $t_k = 1/L$, vale para $k = 1, 2, \dots$

$$f(x_k) - f(x^*) \leq (L||x_0 - x^*||^2)/(2k). \tag{2}$$

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

No **método de Nesterov** (ou gradiente acelerado de Nesterov) ao invés de x_k, o gradiente da função é avaliado em um ponto y que é definido como uma extrapolação de x_{k-1} ao longo da direção $x_k - x_{k-1}$, ou seja, $y_k = x_k + \beta_k(x_k - x_{k-1})$, em que $\beta_k = \theta_k(\theta_{k-1}^{-1} - 1)$ e $\theta_k \in (0, 1)$ [1]. Além das escolhas de tamanho de passo 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

$$f(y_k - t_k\nabla(y_k)) \leq f(y_k) - (t_k/2)\|\nabla f(y_k)\|^2. \tag{3}$$

Em comparação com (2), o método de Nesterov com tamanho de passo constante $t_k = t \in (0, 1/L]$ e com $\theta_k = 2/(k + 2)$ satisfaz (veja [1, Teorema 5.1]):

$$f(x_k) - f(x^*) \leq (2||x_0 - x^*||^2)/(t(k + 1)^2). \tag{4}$$

Esse método garante a taxa de convergência ótima $\mathcal{O}(k^{-2})$ para métodos de primeira ordem.

Outro método de primeira ordem introduzido em [2] é o **método do gradiente espectral**. Para este método em (1) é usada $d_k = -\lambda_k^{-1}\nabla f(x_k)$, em que o escalar é conhecido como parâmetro espectral e é definido como

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

[1]gearlissonmendonca@gmail.com
[2]douglas.goncalves@ufsc.br

e $0 < \delta_{\min} < \delta_{\max} < \infty$ são salva-guardas para evitar que este fique muito grande ou muito próximo de zero. O λ_k recebe este nome pois a razão que aparece envolvendo a diferença dos gradientes e diferença dos iterados é um quociente de Rayleigh para uma "Hessiana média"[2]. O tamanho de passo 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 é possível provar que para $\varepsilon \in (0,1]$, e f convexa e com gradiente Lipschitz, temos que $f(x_k) - f(x^*) \leq \varepsilon$ em no máximo $\mathcal{O}(\varepsilon^{-1})$ iterações [3].

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 $K = 2000$. 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 $f(x^*) = 0$. 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

$$f_q(x) = \frac{L}{4}\left\{\frac{1}{2}\left[x_1^2 + \sum_{i=1}^{q-1}(x_i - x_{i+1})^2 + x_q^2\right] - x_1\right\}, \; para \; q \in \{1, \ldots, n\}. \tag{5}$$

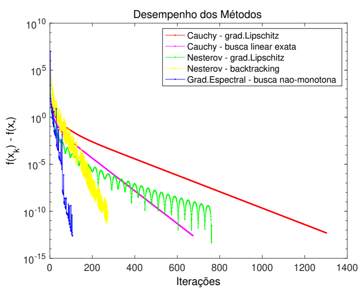

Figura 1: Desempenho dos Métodos

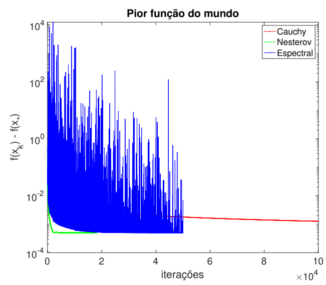

Figura 2: Pior Função do Mundo

Em (5), consideramos $n = 2001$, $q = 1000$ e $K = 10^5$ e no gráfico da Figura 2 fica evidente que existem funções quadráticas convexas para as quais métodos em que d_k é um múltiplo de $-\nabla f(x_k)$ não atingem a taxa de convergência de $\mathcal{O}(1/k^2)$, como no método de Nesterov.

Referências

[1] Y. Nesterov. **Introductory lectures on convex optimization: A basic course**. Vol. 87. Springer Science & Business Media, 2003.

[2] J. Barzilai e J. M. Borwein. "Two-point step size gradient methods". Em: **IMA Journal of Numerical Analysis** 8.1 (1988), pp. 141–148.

[3] G. N. Grapiglia e E. W. Sachs. "On the worst-case evaluation complexity of non-monotone line search algorithms". Em: **Computational Optimization and Applications** 68.3 (2017), pp. 555–577.

