Trabalho apresentado no XLI CNMAC, Unicamp - Campinas - SP, 2022.

Proceeding Series of the Brazilian Society of Computational and Applied Mathematics
Preprint

Métodos de primeira ordem acelerados e buscas adaptativas para minimização suave

Gabriel Grillo[1], Profa. Dra. Sandra Augusta Santos[2]
IMECC/Unicamp, Campinas, SP

Neste trabalho foi considerado o problema de minimizar $f(x)$, s.a $x \in \mathbb{R}^n$, em que $f : \mathbb{R}^n \to \mathbb{R}$ é continuamente diferenciável ($f \in \mathcal{C}^1$) e convexa. Adicionalmente, foram consideradas as hipóteses de f ser L-fortemente suave e μ-fortemente convexa. Um tratamento teórico desse problema com essas hipóteses pode ser visto, por exemplo, em [3, Capítulo 3] ou [6, Capítulo 2].

Esse problema é, em geral, resolvido utilizando-se métodos numéricos iterativos. Dois métodos muito conhecidos são o método do gradiente descendente [4, §3.1.2] e o método de Newton [4, §3.2.2]. Outra opção são os métodos de primeira ordem acelerados, que são métodos iterativos que utilizam apenas informação de gradiente, assim como o gradiente descendente, mas buscam uma melhor taxa de convergência, como o método de Newton.

Entre esses métodos está o método *heavy-ball* [8], que se caracteriza pela atualização

$$x^{k+1} = x^k - \alpha_k \nabla f(x^k) + \beta_k (x^k - x^{k-1}). \tag{1}$$

Em (1) vemos que o método *heavy-ball* se diferencia do método do gradiente descendente por conta do termo de inércia $\beta_k(x^k - x^{k-1})$, o qual tem o papel de defletir o gradiente e amenizar o zigue-zague da trajetória, o que acelera a convergência.

A escolha dos tamanhos de passo α_k e β_k pode seguir os indicados em [8, Teorema 9] quando f for L-fortemente suave e μ-fortemente convexa. Entretanto, quando L ou μ são desconhecidos ou não existem, ainda é possível utilizarmos (1), mas agora será necessário um esquema de buscas adaptativas para seleção dos passos.

Neste trabalho propomos uma maneira de selecionar α_k e β_k de forma adaptativa utilizando-se a regra de Armijo [4, §3.1.1] em cada uma das direções $(-\nabla f(x^k)$ e $(x^k - x^{k-1}))$ individualmente, com possíveis contrações, controladas por α_{contr} e β_{contr}, no caso de falha na regra.

Na estratégia proposta, os palpites iniciais de α_k e β_k são dilatações, controladas por α_{dil} e β_{dil}, dos respectivos passos selecionados na iteração anterior. Com essa dilatação podemos obter a não monotonicidade dos tamanhos de passo, e utilizando a memória dos passos da iteração anterior, podemos diminuir a quantidade de avaliações de f na busca.

Veja que uma exceção deve ser feita para quando a direção de inércia $(x^k - x^{k-1})$ não for de descida. Nesse caso, substituímos β_k por $\delta\beta_{k-1}$ em (1), com $0 \leq \delta \ll 1$, realizamos a atualização $\beta_k \leftarrow \beta_{k-1}$ e esse passo não é dilatado caso haja busca linear na próxima iteração.

Como problemas-teste, consideramos funções quadráticas da forma $f(x) = \sum_{i=1}^{n} d_i x_i^2$, em que os coeficientes d_i foram tais que $d_1 = \mu = 1$, $d_n = L = 1000$ e $d_i \leq d_{i+1}$. Além disso, os coeficientes foram divididos em 1 até 5 *clusters*, sendo possível controlar o quão concentrados ao redor do centro dos *clusters* os coeficientes estavam e se os valores extremos estariam isolados ou não. Ao total, consideramos 400 problemas distintos e para cada um deles foram gerados aleatoriamente 3 pontos iniciais x^0 tais que $\|x^0\|_\infty = 1$. Foi permitido um máximo de 2000 iterações e admitiu-se convergência quando $\|\nabla f(x^k)\|_2 \leq 1\text{e}{-6}$. Os hiperparâmetros da proposta adaptativa foram: $\alpha_0 = \beta_0 = 0.01$, $\alpha_{\text{contr}} = 0.5$, $\alpha_{\text{dil}} = 1.1$, $\beta_{\text{contr}} = 0.2$, $\beta_{\text{dil}} = 2$ e $\delta = 0.001$.

[1] gabriel-grillo@live.com
[2] sasantos@unicamp.br

Comparamos a proposta adaptativa do método *heavy-ball* (`HB adapt`) com outros métodos de primeira ordem acelerados, como o método de Nesterov de 1983 – `N83` [6, Seção 2.2], o método de Nesterov de 2007 [5] com recomeço adaptativo – `N07` [7] e o método de Gonzaga e Karas de 2013 – `GK13` [2]. Os hiperparâmetros de `HB adapt` foram escolhidos manualmente com uma pequena quantidade de problemas-teste quadráticos. A comparação foi feita com a utilização de *performance profile*, conforme proposto em [1]. Os resultados podem ser vistos na Figura 1.

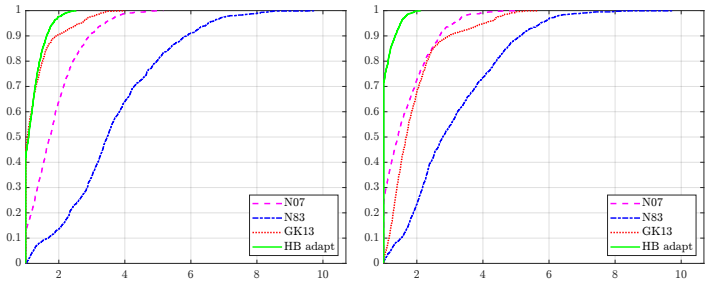

Figura 1: Perfis baseados em iterações (esq.) e tempo (dir.) para problemas com funções quadráticas

Com esses resultados, vemos que a proposta apresentada atingiu desempenho competitivo com métodos já conhecidos, e teve vantagem no conjunto de problemas-teste.

Agradecimentos

Este projeto foi desenvolvido com financiamento da FAPESP (nº de processo 2020/13946-3).

Referências

[1] E. D. Dolan e J. J. Moré. "Benchmarking optimization software with performance profiles". Em: **Mathematical Programming** 91.2 (2002), pp. 201–213.

[2] C. C. Gonzaga e E. W. Karas. "Fine tuning Nesterov's steepest descent algorithm for differentiable convex programming". Em: **Mathematical Programming** 138.1 (2013), pp. 141–166.

[3] A. Izmailov e M. Solodov. **Otimização, Volume 1: Condições de Otimalidade, Elementos de Análise Convexa e de Dualidade**. 3ª ed. Rio de Janeiro: IMPA, 2014.

[4] A. Izmailov e M. Solodov. **Otimização, Volume 2: Métodos Computacionais**. 3ª ed. Rio de Janeiro: IMPA, 2018.

[5] Y. Nesterov. **Gradient methods for minimizing composite objective function**. Discussion paper 76. Bélgica: CORE, UCL, 2007.

[6] Y. Nesterov. **Introductory Lectures on Convex Optimization: A Basic Course**. Vol. 87. Applied Optimization. New York: Springer Science & Business Media, 2003.

[7] B. O'Donoghue e E. Candès. "Adaptive restart for accelerated gradient schemes". Em: **Foundations of Computational Mathematics** 15.3 (2015), pp. 715–732.

[8] B. T. Polyak. "Some methods of speeding up the convergence of iteration methods". Em: **USSR Computational Mathematics and Mathematical Physics** 4.5 (1964), pp. 1–17.