MÉTODOS DE PRIMEIRA ORDEM ACELERADOS E BUSCAS ADAPTATIVAS PARA MINIMIZAÇÃO SUAVE[1]

Palavras-Chave: Gradiente acelerado, métodos adaptativos, testes computacionais.

Autores/as:
Gabriel Grillo (UNICAMP)
Profa. Dra. Sandra Augusta Santos (orientadora) (UNICAMP)

1 Introdução

O presente trabalho se propõe a estudar métodos numéricos para minimização de funções suaves utilizando-se apenas informações de primeira ordem. Em outras palavras, dada uma função $f : \mathbb{R}^n \to \mathbb{R}$, $f \in \mathcal{C}^1$, tal que possua minimizador global $x^* \in \mathbb{R}^n$, i.e. $f(x^*) \leq f(x)$, $\forall x \in \mathbb{R}^n$, objetiva-se estudar métodos numéricos iterativos que geram uma sequência $\{x_k\}_{k=0}^{\infty} \subset \mathbb{R}^n$ tal que $x_k \to x^*$ e x_{k+1} é obtido como combinação linear de x_i e $\nabla f(x_i)$, para $i \in \{0, 1, \ldots, k\}$.

Embora métodos de segunda ordem, como o de Newton, possuam convergência rápida, eles demandam o cálculo e o armazenamento da matriz Hessiana da função f, bem como a resolução de sistemas lineares a cada iteração. O tradicional método de Cauchy, por outro lado, apesar de baixo custo, tem convergência lenta.

Com o aumento da dimensão, os requerimentos de memória e os custos associados, cria-se a necessidade de estratégias aceleradoras. Na área de aprendizado de máquinas, por exemplo, o treinamento de um modelo preditivo é um problema de minimização, em que a dimensão corresponde à quantidade de dados disponíveis. Por balancearem características de eficiência e custo computacional, os métodos de primeira ordem acelerados vem ganhando destaque, e constituem o foco deste trabalho.

1.1 Método de Cauchy

Este é o método de primeira ordem mais simples, sendo o iterado x_{k+1} calculado apenas como um deslocamento do iterado x_k na direção de $-\nabla f(x_k)$ considerando um passo $\alpha_k > 0$. Assim, dado $x_0 \in \mathbb{R}^n$, a sequência $\{x_k\}_{k=0}^{\infty}$ é gerada pelo processo

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k),$$

em que o passo α_k pode ser pré-definido ou obtido a cada iteração com alguma estratégia de busca. A seguinte proposição apresenta o comportamento desse método quando o tamanho de passo é escolhido aproveitando-se informações da estrutura da função.

[1]Projeto financiado pela FAPESP (Número do Processo: 2020/13946-3).

Proposição 1 (Teorema 2.1.15 de Nesterov (2003) adaptado). *Seja $f : \mathbb{R}^n \to \mathbb{R}$ tal que $f \in \mathcal{C}^2$ é fortemente convexa com constante $\mu > 0$ e seu gradiente é Lipchitz contínuo com constante $L \geq \mu$. Então, f possui minimizador global $x^* \in \mathbb{R}^n$ e, dado $x_0 \in \mathbb{R}^n$ qualquer, o método de Cauchy com passo constante $\alpha_k = \frac{2}{L+\mu}$ gera sequência $\{x_k\}_{k=0}^{\infty}$ tal que*

$$\|x_k - x^*\|_2 \leq \left(\frac{\kappa - 1}{\kappa + 1}\right)^k \|x_0 - x^*\|_2, \text{ em que } \kappa \coloneqq \frac{L}{\mu}.$$

No caso de as informações de L e μ não estarem disponíveis ou as constantes não existirem, é possível determinar o tamanho de passo de uma iteração do método de Cauchy pela *regra de Armijo*. Essa regra estabelece um decréscimo suficiente da função a cada iteração, em que $\alpha_k > 0$ deve ser selecionado de forma que

$$f(x_k - \alpha_k \nabla f(x_k)) \leq f(x_k) - \sigma \alpha_k \nabla f(x_k)^T \nabla f(x_k), \tag{1.1}$$

em que $\sigma > 0$ é um hiperparâmetro de controle de exigência do decréscimo. Para mais detalhes, veja, por exemplo, o Capítulo 3 de Nocedal e Wright (2006).

1.2 Método *Heavy-ball*

O método da *bola pesada* é um dos primeiros métodos de primeira ordem acelerados conhecidos, e consiste em acrescentar à iteração do método de Cauchy um termo de inércia, com iterações dadas por

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k) + \beta_k (x_k - x_{k-1}), \tag{1.2}$$

em que $\alpha_k > 0$ e $\beta_k > 0$. A direção do método da bola pesada é, então, uma combinação linear entre a direção de Cauchy (gradiente) e o passo dado na última iteração (termo de inércia).

Nesse método temos os tamanhos de passo α_k e β_k, que podem ser pré-definidos ou selecionados a cada iteração, de forma semelhante à seleção de α_k no método de Cauchy. A seguinte proposição apresenta tamanhos de passos fixos que garantem a convergência da sequência $\{x_k\}_{k=0}^{\infty}$.

Proposição 2 (Item (3) do Teorema 9 de Polyak (1964) adaptado). *Seja $f : \mathbb{R}^n \to \mathbb{R}$ tal que $f \in \mathcal{C}^2$ é fortemente convexa com constante $\mu > 0$ e seu gradiente é Lipchitz contínuo com constante $L \geq \mu$. Então, f possui minimizador global $x^* \in \mathbb{R}^n$ e, dados $x_0, x_1 \in \mathbb{R}^n$ suficientemente próximos de x^*, o método da bola pesada com passos $\alpha_k = \left(\frac{2}{\sqrt{L}+\sqrt{\mu}}\right)^2$ e $\beta_k = \frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}$ gera sequência $\{x_k\}_{k=0}^{\infty}$ tal que*

$$\|x_k - x^*\|_2 \leq c_\delta \left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} + \delta\right)^k \sqrt{\|x_0 - x^*\|_2^2 + \|x_1 - x^*\|_2^2},$$

em que $\kappa \coloneqq \frac{L}{\mu}$, $\delta > 0$ é suficientemente pequeno para que $\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1} + \delta < 1$ e, uma vez fixado δ, c_δ é uma constante positiva.

Os passos propostos na Proposição 2 dependem da existência e conhecimento de constantes estruturais da função objetivo. Propomos nesse trabalho a determinação de α_k e β_k de maneira adaptativa. O tamanho de passo na direção do gradiente, α_k, deverá ser selecionado conforme (1.1). Já o tamanho de passo na direção de inércia possui a sua seleção dividida em dois casos. Se a direção de inércia for de descida ($\nabla f(x_k)^T (x_k - x_{k-1}) < 0$), então aplicamos a regra de Armijo para determinar β_k:

$$f(x_k + \beta_k (x_k - x_{k-1})) \leq f(x_k) + \sigma \beta_k \nabla f(x_k)^T (x_k - x_{k-1}). \tag{1.3}$$

Caso contrário, então $\beta_k \leftarrow \beta_{k-1}$, mas o passo na direção de inércia é amortecido pelo hiperparâmetro $\omega \ll 1$, i.e. trocamos o termo $\beta_k(x_k - x_{k-1})$ em (1.2) por $\omega\beta_k(x_k - x_{k-1})$.

O processo adaptativo se dá nas buscas lineares para garantir tanto (1.1) quanto (1.3). Na implementação, são empregados processos de *backtracking* para obtenção de α_k e de β_k, em que os valores são inicializados com α_{k-1} e β_{k-1}, respectivamente. Hiperparâmetros de contração $\alpha_{\text{contr}} < 1$ e $\beta_{\text{contr}} < 1$ são utilizados durante a busca. Após a obtenção de x_{k+1}, os tamanhos de passo α_k e de β_k são então dilatados por hiperparâmetros $\alpha_{\text{dil}} \geq 1$ e $\beta_{\text{dil}} \geq 1$, respectivamente, para serem usados como inicializações das buscas do passo seguinte.

Vale destacar que somente haverá busca para obtenção de β_k caso a direção de inércia seja de descida, de forma que β_{contr} e β_{dil} somente são aplicados nesse caso. No caso da direção de inércia não ser de descida, $\beta_k \leftarrow \beta_{k-1}$ sem que haja qualquer contração ou dilatação, e então β_k é utilizado como inicialização, se for o caso, para a busca de β_{k+1}.

A estratégia adotada busca tirar proveito da direção de inércia sempre que essa for de descida, e permite que os tamanhos de passo sejam maiores do que os passos utilizados na Proposição 2 por conta dos termos de dilatação, com a garantia de decréscimo suficiente. Na situação da direção de inércia não ser de descida, ainda é considerada essa direção, mesmo que com um amortecimento, para evitar um passo puramente do método de Cauchy.

1.3 Método de Nesterov de 1983

Este é um método da família de métodos ótimos, no sentido em que alcança a melhor taxa de convergência possível para métodos de primeira ordem (veja, por exemplo, Nesterov (2003, Seção 2.2)). O método é apresentado no contexto de funções convexas diferenciáveis com gradiente Lipchitz contínuo. A informação de convexidade forte da função objetivo, quando existir e estiver disponível, é usada pelo método. De maneira geral, o método precisa de $L > 0$, constante de Lipchitz do gradiente, e de $\mu \geq 0$, constante de convexidade forte da função objetivo, que no caso de $\mu = 0$ indica apenas a convexidade da função. Definindo $q \coloneqq \frac{\mu}{L}$ e dados $x_0 \in \mathbb{R}^n$, $y_0 = x_0$ e $\alpha_0 \in (0, 1)$, o processo iterativo é definido por:

$$\begin{cases} x_{k+1} = y_k - \dfrac{1}{L}\nabla f(y_k), \\ \alpha_{k+1} = \dfrac{q - \alpha_k^2 + \sqrt{(q-\alpha_k^2)^2 + 4\alpha_k^2}}{2}, \\ y_{k+1} = x_{k+1} + \dfrac{\alpha_k(1-\alpha_k)}{\alpha_k + \alpha_{k+1}^2}(x_{k+1} - x_k). \end{cases} \tag{1.4}$$

Esse é o esquema iterativo simplificado para o método, em que o passo de gradiente para obtenção de x_{k+1} possui tamanho de passo fixo. Dessa maneira é possível eliminar uma sequência numérica e uma sequência auxiliar no $\mathbb{R}^n$. Em Nesterov (2003, Subseção 2.2.1), é possível acompanhar a versão mais geral do método, assim como sua dedução. A seguinte proposição apresenta a taxa de convergência da sequência $\{x_k\}_{k=0}^{\infty}$ obtida via (1.4).

Proposição 3 (Teorema 2.2.3 de Nesterov (2003) adaptado)**.** *Seja $f : \mathbb{R}^n \to \mathbb{R}$ tal que $f \in \mathcal{C}^1$ é fortemente convexa com constante $\mu > 0$ e seu gradiente é Lipchitz contínuo com constante $L \geq \mu$. Então, f possui minimizador global $x^* \in \mathbb{R}^n$ e, dado $x_0 \in \mathbb{R}^n$ qualquer, o esquema iterativo* (1.4) *com α_0 dado pela maior raiz de $\alpha_0^2 + (1-q)\alpha_0 - 1 = 0$ gera sequência $\{x_k\}_{k=0}^{\infty}$ tal que*

$$\|x_k - x^*\|_2 \leq \sqrt{2\frac{L}{\mu}\min\left\{\left(1 - \sqrt{\frac{\mu}{L}}\right)^k, \frac{4}{(k+2)^2}\right\}}\|x_0 - x^*\|_2.$$

O método de Nesterov de 1983 apresenta uma das melhores taxas de convergência garantidas, visto que é um método ótimo. Entretanto, possui uma grande dependência das constantes estruturais μ e L da função objetivo, que nem sempre estão disponíveis. Além disso, o método como exposto usa as informações estruturais globais, mas para algumas funções pode ser interessante analisar essas informações localmente via estratégias adaptativas.

Para lidar com essas dificuldades, algumas propostas foram feitas for diversos autores. Em Nesterov (2007) é introduzido um esquema adaptativo para o ajuste da constante de Lipschtiz a cada iteração. Já em Gonzaga e Karas (2013), um método com uma diferente escolha para α_k, que independe de L, e uma estratégia adaptativa para seleção de μ a cada iteração são introduzidas. Por fim, em O'Donoghue e Candès (2015), uma estratégia de reinício do método de Nesterov de 1983 é apresentada para lidar com o não conhecimento de μ e L.

2 Experimentos computacionais

No presente trabalho, todos os métodos estudados foram implementados em linguagem `MATLAB`. Para análise dos resultados foi utilizada a técnica de *performance profile* de Dolan e Moré (2002).

Para os experimentos computacionais foram utilizadas funções quadráticas simples, i.e. $f : \mathbb{R}^n \to \mathbb{R}$ da forma $f(x) = \frac{1}{2}\sum_{i=1}^{n} d_i x_i^2$, em que $d \in \mathbb{R}^n$, $\mu = d_1 \leq d_2 \leq \cdots \leq d_{n-1} \leq d_n = L$. As constantes estruturais de f foram escolhidas como $\mu = 1$ e $L = 1000$, de forma que o número de condição de todos os problemas tratados é $\kappa = 1000$.

Para introduzir variabilidade aos problemas quadráticos, desenvolvemos uma sistemática para geração dos autovalores d. Contemplamos a possibilidade dos autovalores estarem concentrados em até 5 *clusters*, os autovalores extremos d_1 e/ou d_n estarem ou não isolados, além de um fator $t \in [0, 1]$ que determina o quanto os autovalores estão concentrados em cada *cluster*. A distribuição usada para determinar os autovalores em cada *cluster* foi a uniforme, e os *clusters* foram uniformemente espaçados no intervalo $[\mu, L]$.

Usando-se a sistemática apresentada, consideramos as dimensões $n \in \{10, 200, 500, 1000, 2000\}$, uma quantidade de *clusters* variando de 1 até 5, o isolamento ou não de d_1 e d_n e o fator $t \in \{0, 0.3, 0.6, 0.9\}$. Com isso, temos $5 \cdot 5 \cdot 2 \cdot 2 \cdot 4 = 400$ possibilidades para geração de problemas. Ademais, para cada problema gerado, três pontos iniciais x_0 foram gerados utilizando-se a distribuição uniforme no intervalo $[-1, 1]$, exigindo que $\|x_0\|_\infty = 1$. Em cada rodada, a convergência foi estabelecida quando $\|\nabla f(x_k)\|_2 < 1\text{e}{-6}$ e limitamos os métodos a 2000 iterações.

Quatro métodos foram utilizados para resolver os problemas, sendo eles: método de Nesterov de 2007 (`N07`) e de 1983 (`N83`), método de Gonzaga e Karas (`GK`) e o método da bola pesada adaptativo (`HB adapt`).

A implementação de `N07` levou em conta o esquema (1.4) com $\alpha_0 = 1$, a estratégia adaptativa de Nesterov (2007) com $\gamma_u = 1.5$ e $\gamma_d = 2$ (notação do autor) e de reinício (esquema com gradiente) de O'Donoghue e Candès (2015), em que não informamos μ e L. A implementação de `N83` usada foi exatamente a do esquema (1.4) com $\alpha_0 = 1$, logo informamos L e μ. A implementação de `GK` seguiu as sugestões originais em Gonzaga e Karas (2013), em que informamos apenas L e na determinação de θ_k usamos $\beta_{\text{dil}} = 2$ (Foi usado, como sugerido, $\beta_\mu = 1.02$ no contexto de seleção de μ_k) e, ao invés de um esquema de redução de intervalo, usamos *backtracking* com constante de contração igual a 0.5. Além disso, usamos tamanho de passo de gradiente fixo $\nu = \frac{1}{L}$. A implementação de `HB adapt` levou em conta as discussões da Subseção 1.2 com as escolhas $\alpha_0 = \beta_0 = 0.01$, $\alpha_{\text{contr}} = 0.5$, $\beta_{\text{contr}} = 0.2$,

$\alpha_{\text{dil}} = 1.1$, $\beta_{\text{dil}} = 2$, $\omega = 1\text{e}{-3}$, $\sigma = 1\text{e}{-4}$ e x_1 obtido como uma perturbação aleatória (distribuição normal) de x_0 com norma-2 igual a 0.1. As implementações aqui apresentadas foram escolhidas como tais pois se destacaram em rodadas preliminares comparando opções distintas em cada método.

Na Figura 1 vemos os perfis comparando os resultados dos métodos apresentados. Assim como observado em Gonzaga e Karas (2013) em termos das iterações, `GK` se destaca em relação aos métodos `N07` e `N83`. O método `HB adapt`, que não foi considerado em Gonzaga e Karas (2013), apresenta desempenho competitivo com o método `GK` nesse quesito. Já em relação ao tempo de execução, veja que a grande quantidade de buscas existentes no método `GK` resulta em um processamento mais lento, o que faz com que o método `N07` tenha melhor desempenho. Ademais, o método `HB adapt` supera o método `N07` e resolve pouco mais de 70% dos problemas mais rapidamente. De qualquer forma, destaca-se que todos os métodos são igualmente robustos no conjunto de problemas gerados, visto que todos conseguem alcançar a convergência com a tolerância estabelecida.

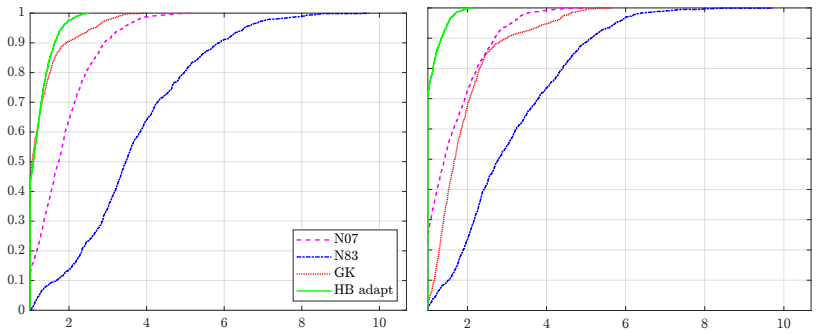

Figura 1: Perfis baseados em iterações (esq.) e tempo (dir.) para problemas com funções quadráticas

3 Conclusão

No desenvolvimento deste trabalho foi possível verificar que os métodos de primeira ordem acelerados são capazes de incrementar a eficiência do método de Cauchy sem que haja a necessidade da utilização de informação de segunda ordem, que é muitas vezes inadequada. Além disso, vimos que as estratégias adaptativas e de reinício incrementaram a eficiência desses métodos nos testes realizados.

Referências

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

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

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

— (2007). **Gradient methods for minimizing composite objective function**. Discussion paper 76. Bélgica: CORE, UCL.

Nocedal, J. e S. Wright (2006). **Numerical Optimization**. Springer Science & Business Media.

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

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