

UNIVERSIDADE FEDERAL DE SANTA CATARINA
CENTRO DE CIÊNCIAS FÍSICAS E MATEMÁTICAS
BACHARELADO EM MATEMÁTICA

Gearlisson dos Santos Mendonça

Um estudo sobre métodos de máxima descida e acelerações

Florianópolis - SC

2022

Gearlisson dos Santos Mendonça

Um estudo sobre métodos de máxima descida e acelerações

Trabalho de Conclusão de Curso de Graduação em Bacharelado em Matemática do Centro de Ciências Físicas e Matemáticas da Universidade Federal de Santa Catarina para a obtenção do título de Bacharel em Matemática.

Orientador: Prof. Douglas Soares Gonçalves

Florianópolis - SC

2022

Gearlisson dos Santos Mendonça

Um estudo sobre métodos de máxima descida e acelerações

Este Trabalho Conclusão de Curso foi julgado adequado para obtenção do Título de Bacharel em Matemática e aprovado em sua forma final pelo Curso de Bacharelado em Matemática.

Florianópolis - SC, 15 de dezembro de 2022.

Prof. Dr. Silvia Martini de Holanda
Coordenadora do Curso

Banca Examinadora:

Prof. Douglas Soares Gonçalves (Orientador)
Universidade Federal de Santa Catarina

Prof. Silvia Martini de Holanda
Universidade Federal de Santa Catarina

Prof. Leonardo Koller Sacht
Universidade Federal de Santa Catarina

Ficha de identificação da obra elaborada pelo autor,
através do Programa de Geração Automática da Biblioteca Universitária da UFSC.

Mendonça, Gearlisson dos Santos

Um estudo sobre métodos de máxima descida e acelerações
/ Gearlisson dos Santos Mendonça ; orientador, Douglas
Soares Gonçalves , 2022.

68 p.

Trabalho de Conclusão de Curso (graduação) -
Universidade Federal de Santa Catarina, Centro de Ciências
Físicas e Matemáticas, Graduação em Matemática e Computação
Científica, Florianópolis, 2022.

Inclui referências.

1. Matemática e Computação Científica. 2. Otimização
irrestrita. 3. Métodos iterativos de descida. 4. Pior
função do mundo. I. , Douglas Soares Gonçalves. II.
Universidade Federal de Santa Catarina. Graduação em
Matemática e Computação Científica. III. Título.

Dedico este trabalho à minha tia/avó, Maria de Jesus, que esteve presente em todos os momentos da minha vida, e todas as minhas conquistas. Infelizmente, hoje ela não se encontra mais entre nós, e algumas de suas últimas palavras foram para que chegasse onde cheguei.

Agradecimentos

No decorrer da minha graduação, inúmeras pessoas passaram por minha vida, deixando marcas e lições para toda ela, proporcionando-me alegrias, conhecimento e crescimento pessoal. Neste momento gostaria de agradecê-las, pois, de alguma forma, contribuíram para a conclusão desta etapa.

A Deus, causa primeira de todas as coisas, cuja infinita sabedoria e imensurável amor nos guia até Ele através dos séculos. Pelo dom da vida e por ter me dado todas as condições para que eu pudesse chegar ao almejado lugar em que cheguei.

Agradeço a minha mãe, Domingas, pela confiança inabalável na minha pessoa e pela sua coragem e resistência inspiradoras, aos meus irmãos: Dalberlanes e Gean, por sempre apoiarem nos meus objetivos. A minha vó Ribinha, que em qualquer oportunidade declara o orgulho de me ter como neto.

Quero expressar também meu mais sincero agradecimento ao professor Douglas Soares Gonçalves, pela sua paciência, tempo, apoio e dedicação ao me orientar na Iniciação Científica e no Trabalho de Conclusão de Curso (TCC). Sua confiança e capacidade para guiar minhas ideias foram de muito valor, não só no desenvolvimento do TCC, mas também na minha formação como profissional. Suas ideias, rigorosidade e críticas com finalidade de realizar um bom trabalho. Muito obrigado Professor.

Agradeço a minha amada Igreja Presbiteriana na Trindade que me acolheu desde o primeiro dia em Florianópolis. Em especial ao meu grupo de estudo “Família Santa Mônica” pelos momentos vividos juntos. Obrigado por dividir comigo as angústias e alegrias e por sempre estarem orando por mim. Foi bom poder contar com vocês!

Agradeço ao meu amigo Vinicius Cerutti, com o qual tive a

chance de estudar, sendo ele um dos grandes incentivadores da minha área de pesquisa. Mas também, ao Carlos Gabriel. Cada qual com sua humildade, paciência, esforço e sua disponibilidade em ajudar, foram inestimáveis na confecção deste trabalho.

A Cristiane e a Duda, minha segunda família por terem me ajudado sempre que precisei.

Agradeço aos professores do Departamento de Matemática da UFSC que contribuíram na minha formação.

Aos meus amigos construídos ao longo do curso, que estavam sempre procurando ajudar uns aos outros.

Por fim, mas não menos importante, eu deixo os agradecimentos aos meus grandes, velhos e maravilhosos amigos. Foram eles quem mais seguraram e, principalmente, suportaram as minhas reclamações, e ouviram minhas dificuldades, sem jamais deixar de incentivar e mostrar extrema confiança no meu potencial.

*“Quero implantar no mundo
Um pouco mais de Otimização.
Quero torná-lo mais profundo,
Para ouvir o meu coração
Quero minimizar as dores,
E restringir todo tipo de guerra.
Quero um mundo cheio de flores,
E felicidade em toda Terra.
Quero maximizar as alegrias,
Sem restrições quero igualdade.
Quero encher os meus dias,
De muita felicidade.”
(FONSECA, 2014)*

Resumo

Neste trabalho é apresentado um estudo sobre o método de máxima descida (método do gradiente ou método de Cauchy) e variações mais recentes, como o gradiente acelerado de Nesterov e gradiente espectral. Foi realizada uma revisão dos principais resultados teóricos destes métodos de primeira ordem, com foco na complexidade de iteração destes métodos. Além de estudar a complexidade de pior caso, também foi investigado o desempenho prático destes métodos através de experimentos numéricos com funções quadráticas com Hessiana simétrica positiva definida. Por fim, comentamos brevemente sobre a “pior função do mundo” para a qual a taxa de convergência ótima do método de Nesterov fica evidenciada.

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

Abstract

This work presents a study on the steepest descent method (also called gradient method or Cauchy method) and more recent variants, such as Nesterov's accelerated gradient and spectral gradient. A review of the main theoretical results for these methods is presented, focusing on their iteration complexity. Besides the worst case analysis, the practical behaviour of such methods was also investigated through numerical experiments on quadratics with symmetric positive definite Hessian. Finally, a brief discussion about "the worst function in the world" is given and some experiments carried out in order to highlight the optimal convergence rate of Nesterov's method.

Keywords: Maximum descent, Nesterov's accelerated gradient, Spectral gradient.

Lista de figuras

Figura 1 – Função convexa	30
Figura 2 – Aproximação linear de f	32
Figura 3 – Direção de descida	35
Figura 4 – Desempenho dos métodos	61
Figura 5 – Desempenho de pior caso	64

Lista de Algoritmos

1	Busca linear inexata com condição de Armijo	37
2	Método do gradiente	40
3	Gradiente Acelerado de Nesterov	44
4	Gradiente Espectral	48
5	Algoritmo de descida com busca não-monótona	49

Lista de tabelas

Tabela 1 – Média das iterações com autovalores av_1	59
Tabela 2 – Média das iterações com autovalores av_2	60
Tabela 3 – Média das iterações com autovalores av_3	60
Tabela 4 – Resultado obtido com a pior função do mundo	63

Sumário

1	INTRODUÇÃO	23
2	CONCEITOS PRELIMINARES	25
2.1	NOÇÕES DE ANÁLISE NO \mathbb{R}^n	25
2.2	NOÇÕES DE CONVEXIDADE	29
2.3	NOÇÕES DE OTIMIZAÇÃO CONTÍNUA	32
2.3.1	Busca Linear	34
2.3.1.1	Busca Exata	35
2.3.1.2	Busca inexata e condição de Armijo	36
2.3.1.3	Busca não-monótona	37
3	MÉTODOS DE DESCIDA	39
3.1	MÉTODO DE CAUCHY	39
3.2	MÉTODO DE NESTEROV	43
3.3	MÉTODO DO GRADIENTE ESPECTRAL	46
4	COMPLEXIDADE DE PIOR CASO DO MÉTODO DO GRADIENTE ESPECTRAL	49
5	EXPERIMENTOS NUMÉRICOS	57
5.1	DETALHES DE IMPLEMENTAÇÃO	57
5.2	FUNÇÕES QUADRÁTICAS	58
5.3	RESULTADOS DOS TESTES	58
5.4	A PIOR FUNÇÃO DO MUNDO DE NESTEROV	61
6	CONSIDERAÇÕES FINAIS E TRABALHOS FUTUROS	65
	REFERÊNCIAS	67

1 Introdução

A Otimização é uma área da matemática que trata do estudo de problemas nos quais se busca minimizar ou maximizar um funcional f sobre um determinado conjunto. Tais problemas possuem inúmeras aplicações em diversas áreas do conhecimento, como por exemplo, Engenharia, Medicina, Economia, Administração entre outras [13, 14].

Em geral, para tais problemas não há uma fórmula fechada para a solução. Assim, normalmente são empregados métodos de otimização que buscam gerar uma sequência que se aproxime da solução do problema.

Neste trabalho é apresentado um estudo sobre métodos de descida para resolução de problemas de otimização irrestrita: encontrar $x^* \in \mathbb{R}^n$ tal que $f(x^*) \leq f(x)$ para todo $x \in \mathbb{R}^n$. Mais especificamente, focaremos em funções convexas e com gradiente Lipschitz.

Como mencionado, para resolver este problema, normalmente são utilizados métodos iterativos que, a partir de um ponto inicial x_0 , geram uma sequência $\{x_k\}$ em \mathbb{R}^n tal que $f(x_{k+1}) < f(x_k)$.

O objetivo deste trabalho é fazer uma revisão dos principais resultados teóricos para métodos com iteração definida por

$$x_{k+1} = x_k + t_k d_k$$

em que a direção d_k é um vetor de \mathbb{R}^n , chamado direção de descida, e $t_k > 0$ é chamado tamanho de passo: ambos determinados usando apenas informação sobre $\nabla f(x)$ em pontos do segmento x_k, x_{k-1} . Tais métodos são conhecidos na literatura como métodos de primeira ordem. Essa classe de métodos tem re-ganhado a atenção de pesquisadores por ter uma iteração barata do ponto de vista computacional, ao contrário, por exemplo, do método de Newton o qual, além de necessitar de informações sobre derivadas segundas, ainda demanda a resolução de

um sistema linear a cada iteração, o que pode ser computacionalmente caro, sobretudo para problemas com um grande número de variáveis.

Daremos ênfase aos métodos de Cauchy [14] e variações recentes, como gradiente acelerado de Nesterov [11] e o método de gradiente espectral [1], e analisaremos tanto a complexidade de pior caso quanto o desempenho prático destes métodos.

O texto foi elaborado de modo que qualquer leitor com um conhecimento básico em Álgebra Linear e Cálculo possa acompanhar o desenvolvimento do trabalho.

O trabalho se organiza da seguinte maneira: no **Capítulo 2** listamos alguns resultados sobre Análise no \mathbb{R}^n , Convexidade e Otimização que serão utilizados praticamente durante todo o texto. O **Capítulo 3** apresenta uma revisão dos Métodos de Cauchy, de Nesterov e Gradiente Espectral, seus algoritmos e as taxas de convergências dos dois primeiros. Já no **Capítulo 4** estudaremos mais a fundo a convergência da sequência $\{x_k\}$ gerada pelo método do Gradiente Espectral com base em um trabalho recente da literatura. No **Capítulo 5** serão apresentados experimentos numéricos para avaliar o desempenho prático destes métodos em funções quadráticas definidas positivas com diferentes distribuições de autovalores. Além disso, a fim de evidenciar a performance de pior caso, também realizamos testes com a “pior função do mundo” (segundo Nesterov).

Por fim, o **Capítulo 6** traz as considerações finais e as direções para trabalhos futuros.

2 Conceitos Preliminares

Para uma melhor compreensão do conteúdo subsequente, neste capítulo iremos revisar alguns resultados e definições de Análise no \mathbb{R}^n , e noções de Convexidade e de Otimização Contínua que serão necessários ao desenvolvimento deste trabalho. As principais referências utilizadas neste capítulo foram [3], [4], [8] e [14].

2.1 Noções de Análise no \mathbb{R}^n

Nesta seção enunciaremos alguns resultados sobre sequências que serão úteis na análise da convergência dos métodos. Além disso, revisaremos os conceitos de gradiente e Hessiana de uma função de várias variáveis e um resultado importante que será usado posteriormente, sobre funções com gradiente Lipschitz. A principal referência desta seção foi [8].

Definição 2.1.1. *Uma sequência em \mathbb{R}^n é uma função*

$$\begin{aligned} x : \mathbb{N} &\rightarrow \mathbb{R}^n \\ k &\rightarrow x_k \end{aligned}$$

Observação 1. *A notação x_k é utilizada com frequência ao longo deste trabalho para denotar o k -ésimo elemento da sequência. Também pode-se escrever $(x_k)_{k \in \mathbb{N}}$, $(x_0, x_1, \dots, x_k, \dots)$ ou $\{x_k\}$.*

Exemplo 1. $x_k = (-1)^k$ define a sequência $(-1, 1, -1, 1, -1, 1, -1, \dots)$ em \mathbb{R} .

Definição 2.1.2. *A sequência $\{x_k\}$ em \mathbb{R}^n tem um limite $L \in \mathbb{R}^n$ se para qualquer $\varepsilon > 0$ existir um número $k_0 \in \mathbb{N}$, tal que se k for inteiro de tal modo que $k \geq k_0$, então $\|x_k - L\| < \varepsilon$ e escrevemos*

$$\lim_{k \rightarrow \infty} x_k = L$$

Exemplo 2. Vamos mostrar que a sequência $x_k = \frac{k}{2k+1}$ tem limite $\frac{1}{2}$.

Dado $\varepsilon > 0$. Temos que encontrar $k_0 \in \mathbb{N}$ tal que, para todo $k \geq k_0$, satisfaça $|x_k - \frac{1}{2}| < \varepsilon$.

Note que

$$|x_k - \frac{1}{2}| = \left| \frac{k}{2k+1} - \frac{1}{2} \right| = \left| \frac{-1}{2(2k+1)} \right| = \frac{1}{2(2k+1)}.$$

Como queremos encontrar um k_0 tal que $k \geq k_0$, então podemos escrever

$$2k + 1 \geq 2k_0 + 1 > 2k_0,$$

assim,

$$2(2k+1) > 2(2k_0) = 4k_0,$$

e portanto

$$\frac{1}{2(2k+1)} < \frac{1}{4k_0}$$

Disto, temos que se $\frac{1}{4k_0} < \varepsilon$ então obtemos $k_0 > \frac{1}{4\varepsilon}$.

Portanto,

$$\lim_{k \rightarrow \infty} x_k = \lim_{k \rightarrow \infty} \frac{k}{2k+1} = \frac{1}{2}$$

.

Definição 2.1.3. Se a sequência $\{x_k\}$ em \mathbb{R}^n tiver um limite L , dizemos que essa sequência é convergente, e $\{x_k\}$ converge para L . Caso contrário, essa sequência é dita divergente.

Definição 2.1.4. Uma sequência $\{x_k\}$ em \mathbb{R}^n é limitada, quando o conjunto formado pelos seus elementos é limitado, ou seja, quando existe um número real $M > 0$ tal que $\|x_k\| \leq M$, para todo $k \in \mathbb{N}$.

Teorema 2.1.1. Se a sequência $\{x_k\}$ em \mathbb{R}^n for convergente, então $\{x_k\}$ é limitada.

Demonstração. Seja $\lim_{k \rightarrow \infty} x_k = L$, temos que existe $k_0 \in \mathbb{N}$, tal que $k \geq k_0$ isso implica que $\|x_k - L\| < 1$. Logo,

$$\|x_k\| = \|(x_k - L) + L\| \leq \|x_k - L\| + \|L\| < 1 + \|L\|, \quad \forall k \geq k_0.$$

Fazendo-se $\mu = \max\{\|x_1\|, \dots, \|x_{k_0-1}\|, 1 + \|L\|\}$, tem-se $\|x_k\| \leq \mu$, $\forall k \in \mathbb{N}$.

Portanto, (x_k) é limitada. □

Definição 2.1.5. Considere $f : \mathbb{R}^n \rightarrow \mathbb{R}$ continuamente diferenciável. O vetor $\nabla f(x) \in \mathbb{R}^n$, que denotamos na forma matricial por

$$\nabla f(x) = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{bmatrix}$$

em que $\frac{\partial f}{\partial x_i}$ é a derivada parcial de f com respeito a x_i , avaliada em x é chamado vetor gradiente de f em x .

Definição 2.1.6. Se f for duas vezes continuamente diferenciável, a matriz $\nabla^2 f(x) \in \mathbb{R}^{n \times n}$ dada por

$$\nabla^2 f(x) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n x_1} & \frac{\partial^2 f}{\partial x_n x_2} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix}$$

é chamada Hessiana de f em x .

Exemplo 3. Seja $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ definida por $f(x) = x_1^2 + x_2^2$. Temos que,

$$\nabla f(x) = \begin{bmatrix} 2x_1 \\ 2x_2 \end{bmatrix} \quad e \quad \nabla^2 f(x) = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}.$$

Definição 2.1.7. Uma função $f : \mathbb{R}^n \rightarrow \mathbb{R}$ chama-se Lipschitz quando existe uma constante $L > 0$ tal que

$$|f(x) - f(y)| \leq L\|x - y\|$$

para todo $x, y \in \mathbb{R}^n$.

Exemplo 4. A função $f(x) = \|x\|$ é Lipschitz de constante $L = 1$ em \mathbb{R}^n .

Sejam $x, y \in \mathbb{R}^n$. Veja que, da desigualdade triangular temos

$$\|x\| = \|x - y + y\| \leq \|x - y\| + \|y\|$$

Assim,

$$\|x\| - \|y\| \leq \|x - y\| \quad (1)$$

De modo análogo,

$$\|y\| = \|y - x + x\| \leq \|x - y\| + \|x\|$$

Dessa forma,

$$\|y\| - \|x\| \leq \|x - y\| \quad (2)$$

Juntando as desigualdades (1) e (2), obtemos

$$|\|x\| - \|y\|| \leq \|x - y\|.$$

Proposição 2.1.2. Sejam $C \subseteq \mathbb{R}^n$ um aberto e $f : \mathbb{R}^n \rightarrow \mathbb{R}$ uma função diferenciável com $\nabla f : \mathbb{R}^n \rightarrow \mathbb{R}^n$ Lipschitz de constante $L > 0$, isto é, $\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$, para todo $x, y \in C$. Então

$$f(y) \leq f(x) + \nabla f(x)^T (y - x) + \frac{L}{2} \|y - x\|^2 \quad (3)$$

Demonstração. Pelo teorema do valor médio

$$\begin{aligned} f(y) - f(x) &= \int_0^1 \nabla f(x + t(y-x))^T (y-x) dt \\ &= \nabla f(x)^T (y-x) + \int_0^1 [(\nabla f(x + t(y-x)) - \nabla f(x))(y-x)] dt \end{aligned}$$

Assim, tomando normas, obtemos

$$\begin{aligned}
 \|f(y) - f(x) - \nabla f(x)^T(y-x)\| &\leq \int_0^1 \|(\nabla f(x+t(y-x)) - \nabla f(x))\| \|y-x\| dt \\
 &\leq \|y-x\| \int_0^1 Lt \|y-x\| dt \\
 &= L \|y-x\| \int_0^1 t dt \\
 &= \frac{L}{2} \|y-x\|^2
 \end{aligned}$$

em que na segunda desigualdade usamos o fato de $\nabla f(x)$ ser Lipschitz de constante L . \square

2.2 Noções de convexidade

Veremos agora algumas noções sobre conjuntos convexos e funções convexas que se mostrarão bastante úteis em capítulos posteriores. As principais referências utilizadas nesta seção foram [2], [9], [13] e [14].

Definição 2.2.1. *Um conjunto não-vazio $C \subseteq \mathbb{R}^n$ é dito convexo se para quaisquer dois pontos $x, y \in C$ tivermos*

$$\alpha x + (1 - \alpha)y \in C, \quad \forall \alpha \in [0, 1].$$

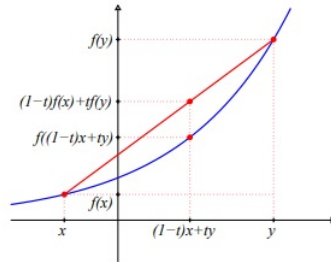
Definição 2.2.2. *Seja $C \subseteq \mathbb{R}^n$ um conjunto convexo. A função $f : \mathbb{R}^n \rightarrow \mathbb{R}$ é dita convexa em C quando*

$$f(\alpha x + (1 - \alpha)y) \leq \alpha f(x) + (1 - \alpha)f(y)$$

para todos $x, y \in C$ e $\alpha \in [0, 1]$.

Geometricamente podemos dizer que o gráfico de uma função convexa fica sempre por baixo da reta secante a dois pontos do gráfico, ao longo do segmento que une tais pontos. A Figura 1 ilustra uma função convexa.

Figura 1 – Função convexa



Fonte: Ribeiro e Karas [14].

Exemplo 5. A função $f(x) = |x|$ é convexa em \mathbb{R} .

De fato, tomamos $\alpha \in [0, 1]$ e para todo $x, y \in \mathbb{R}$, queremos mostrar que

$$f(\alpha x + (1 - \alpha)y) \leq \alpha f(x) + (1 - \alpha)f(y).$$

Veja que

$$f(\alpha x + (1 - \alpha)y) = |\alpha x + (1 - \alpha)y|.$$

Pela desigualdade triangular, obtemos

$$\begin{aligned} |\alpha x + (1 - \alpha)y| &\leq |\alpha x| + |(\alpha - 1)y| \\ &= \alpha|x| + (1 - \alpha)|y| \\ &= \alpha f(x) + (1 - \alpha)f(y). \end{aligned}$$

Portanto, $f(x) = |x|$ é convexa.

Dada uma função f diferenciável, a convexidade admite algumas caracterizações alternativas para determinar se f é convexa ou não.

Lema 2.2.1. Sejam $C \subseteq \mathbb{R}^n$ um conjunto convexo e $f : \mathbb{R}^n \rightarrow \mathbb{R}$ diferenciável. A função f é convexa em C se, e somente se,

$$f(y) \geq f(x) + (y - x)^T \nabla f(x), \quad \forall x, y \in C.$$

Demonstração. (\leftarrow) Sejam $x, z \in C$ e $\alpha \in [0, 1]$. Defina $y = \alpha x + (1 - \alpha)z$. Dessa forma,

$$f(x) \geq f(y) + (x - y)^T \nabla f(y) \quad (4)$$

$$f(z) \geq f(y) + (z - y)^T \nabla f(y) \quad (5)$$

Multiplicando α por (4) e $(1 - \alpha)$ por (5) obtemos,

$$\begin{aligned} \alpha f(x) &\geq \alpha f(y) + \alpha(x - y)^T \nabla f(y) \\ (1 - \alpha)f(z) &\geq (1 - \alpha)f(y) + (1 - \alpha)(z - y)^T \nabla f(y) \end{aligned}$$

Somando as duas desigualdades obtidas acima temos,

$$\begin{aligned} \alpha f(x) + (1 - \alpha)f(z) &\geq \alpha f(y) + \alpha(x - y)^T \nabla f(y) + (1 - \alpha)f(y) + (1 - \alpha)(z - y)^T \nabla f(y) \\ &= f(y)(\alpha + (1 - \alpha)) + [\alpha(x - y)^T + (1 - \alpha)(z - y)^T] \nabla f(y) \\ &= f(y) + (\alpha x + (1 - \alpha)z - y)^T \nabla f(y) \end{aligned}$$

Como $y = \alpha x + (1 - \alpha)z$ temos,

$$\begin{aligned} \alpha f(x) + (1 - \alpha)f(z) &\geq f(y) + (y - y)^T \nabla f(y) \\ &= f(y) \end{aligned}$$

Logo, f é convexa.

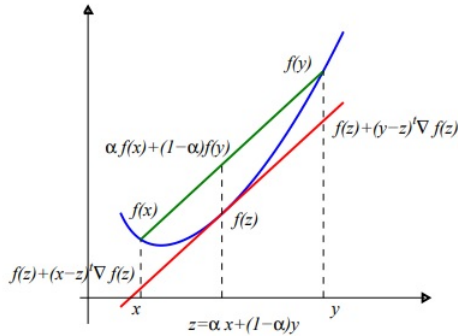
(\rightarrow) Suponha que f seja convexa. Para $x, y \in C$ quaisquer e $t \in [0, 1]$, definimos $d = y - x$. Assim, temos $x + td \in C$ e $f(x + td) = f((1 - t)x + ty) \leq (1 - t)f(x) + tf(y)$

Portanto,

$$\begin{aligned} f(y) - f(x) &\geq \lim_{t \rightarrow 0^+} \frac{(x + td) - f(x)}{t} = \nabla f(x)^T d = \nabla f(x)^T (y - x) \\ f(y) - f(x) &\geq \nabla f(x)^T (y - x) \\ f(y) &\geq f(x) + \nabla f(x)^T (y - x) \end{aligned}$$

□

Pelo Lema 2.2.1 temos que uma função convexa está sempre acima da sua aproximação linear. A Figura 2 ilustra o teorema

Figura 2 – Aproximação linear de f 

Fonte: Ribeiro e Karas [14].

2.3 Noções de otimização contínua

Nesta seção faremos uma revisão das principais noções de otimização contínua. As principais referências utilizadas nesta seção foram [4], [7], [9], [10], [13] e [14]

Vamos considerar o seguinte problema de otimização irrestrita:

$$\begin{aligned} \min \quad & f(x) \\ \text{sujeito a} \quad & x \in \mathbb{R}^n, \end{aligned} \tag{6}$$

em que $f : \mathbb{R}^n \rightarrow \mathbb{R}$ é uma função continuamente diferenciável e com gradiente Lipschitz, chamada *função objetivo*.

A solução do problema (6) é um vetor $x^* \in \mathbb{R}^n$ cujo o valor de $f(x^*)$ é o menor possível; chamamos x^* de *minimizador* de f em \mathbb{R}^n .

Definição 2.3.1. *Considere uma função $f : \mathbb{R}^n \rightarrow \mathbb{R}$ e $x^* \in \mathbb{R}^n$. Dizemos que x^* é um minimizador global de f em \mathbb{R}^n se $f(x^*) \leq f(x)$, para todo $x \in \mathbb{R}^n$. Por outro lado, Dizemos que x^* é um minimizador local de f em \mathbb{R}^n se existir $\varepsilon > 0$ tal que $f(x^*) \leq f(x)$, para todo $x \in \mathbb{R}^n$ e $\|x - x^*\| < \varepsilon$, ou seja, $x \in B_\varepsilon(x^*) \cap \mathbb{R}^n$.*

Abaixo veremos uma condição que deve ser satisfeita por minimizadores locais quando a função objetivo f é diferenciável.

Teorema 2.3.1. *Considere $f : \mathbb{R}^n \rightarrow \mathbb{R}$ continuamente diferenciável. Se x^* é minimizador local de f em \mathbb{R}^n , então $\nabla f(x^*) = 0$.*

Demonstração. Fixe um vetor não-nulo $d \in \mathbb{R}^n$ arbitrário e considere a função $\phi : \mathbb{R} \rightarrow \mathbb{R}$ definida por $\phi(t) = f(x^* + td)$. Como x^* é um minimizador local de f , resulta que $t = 0$ é minimizador local de ϕ e portanto $\phi'(0) = 0$. Pela regra da cadeia, obtemos

$$\phi'(t) = \nabla f(x^* + td)^T d.$$

Substituindo para $t = 0$, resulta $0 = \phi'(0) = \nabla f(x^*)^T d$. Como $d \in \mathbb{R}^n$ é arbitrário, então $\nabla f(x^*)$ é um vetor ortogonal a todos os vetores do espaço. Portanto $\nabla f(x^*) = 0$. \square

Exemplo 6. *Seja $f : \mathbb{R}^3 \rightarrow \mathbb{R}$ dada por $f(x) = \text{sen}(3x_1^2 + x_2^2) + \cos(x_1^2 - x_2^2) + 4x_3$. Verificaremos se f tem minimizadores em \mathbb{R}^3 . Note que $\nabla f(x) \neq 0$, para todo $x \in \mathbb{R}^3$, pois $\frac{\partial f}{\partial x_3}(x) = 4$. Portanto, pelo Teorema 2.3.1, não existe minimizador de f em \mathbb{R}^3 .*

Agora usando os resultados de convexidade da Seção 2.2, veremos que para f convexa, todo minimizador local é global.

Teorema 2.3.2. *Sejam $C \subseteq \mathbb{R}^n$ convexo e $f : C \rightarrow \mathbb{R}$ uma função convexa. Se $x^* \in C$ é minimizador local de f , então x^* é minimizador global de f .*

Demonstração. Suponha que x^* não seja um minimizador global de f . Então existe $x \in C$ tal que $f(x) < f(x^*)$. Para $\alpha \in [0, 1]$, consideremos $y = \alpha x + (1 - \alpha)x^*$. Pela convexidade de C temos $y \in C$. Agora, pela convexidade da f ,

$$f(y) \leq \alpha f(x) + (1 - \alpha)f(x^*) = f(x^*) + \alpha(f(x) - f(x^*)) < f(x^*).$$

Assim, para α suficientemente próximo de 0, y torna-se arbitrariamente próximo de x^* , mas $f(y) < f(x^*)$. Portanto, x^* não poderia ser um minimizador local. \square

A seguir, falaremos um pouco sobre o conceito de *direção de descida* no qual se baseiam os métodos estudados neste trabalho.

Definição 2.3.2. *Considere $f : \mathbb{R}^n \rightarrow \mathbb{R}$ e sejam $x \in \mathbb{R}^n$ e $d \in \mathbb{R}^n$. Se existir $\varepsilon > 0$ tal que $f(x + td) < f(x)$, com $t \in [0, \varepsilon]$ dizemos que d é uma direção de descida.*

Apresentamos abaixo uma condição suficiente para uma direção ser de descida.

Lema 2.3.3. *Seja f continuamente diferenciável. Se $\nabla f(x)^T d < 0$ então d é uma direção de descida para f a partir de x .*

Demonstração. Como $\nabla f(x)^T d = \lim_{t \rightarrow 0} \frac{f(x+td) - f(x)}{t}$ e por hipótese $\nabla f(x)^T d < 0$, então para todo $t > 0$ suficientemente pequeno, temos $f(x + td) < f(x)$. \square

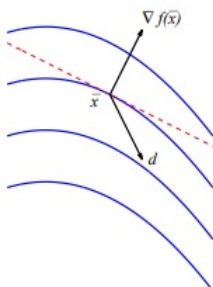
As direções que formam um ângulo maior que 90 graus com o gradiente são direções de descida [9], como ilustra a Figura 3.

Note que se $\nabla f(x) \neq 0$, então $d = -\nabla f(x)$ é direção de descida para f a partir de x . De fato, o que mostra pelo Lema 2.3.3 que $-\nabla f(x)$ é uma direção de descida.

2.3.1 Busca Linear

Uma vez fixada uma direção de descida d_k , uma pergunta natural é o quanto devemos “andar” nesta direção, isto é, qual o escalar $t_k > 0$ para que $x_{k+1} = x_k + t_k d_k$ seja uma boa atualização. O escalar t_k é chamado *tamanho de passo* e pode ser calculado observando o comportamento da função f ao longo da semirreta a partir de x_k na

Figura 3 – Direção de descida



Fonte: Ribeiro e Karas [14].

direção d_k , através de um processo que damos nome de *busca linear*. Descrevemos abaixo algumas estratégias para a busca linear.

2.3.1.1 Busca Exata

Considere uma função $f : \mathbb{R}^n \rightarrow \mathbb{R}$, um ponto $\bar{x} \in \mathbb{R}^n$ e uma direção de descida $d \in \mathbb{R}^n$. A busca exata é um método que consiste em minimizar unidimensionalmente f a partir do ponto \bar{x} ao longo da direção d , ou seja, devemos resolver o problema

$$\begin{aligned} \min \quad & f(\bar{x} + td) \\ \text{s.a} \quad & t > 0. \end{aligned} \tag{7}$$

Este problema é em geral difícil de resolver. Entretanto, dependendo do tipo de função considerada, é possível obter fórmulas fechadas para a solução de (7), como no caso das funções quadráticas com Hessiana simétrica positiva definida.

Seja f uma função quadrática da forma $f(x) = \frac{1}{2}x^T Ax - b^T x + c$ com A sendo a matriz simétrica positiva definida. Então, definindo

$\varphi(t) = f(\bar{x} + td)$, temos que, pela regra da cadeia,

$$\varphi'(t) = \nabla f(x + td)^T d$$

Note que se t é minimizador de φ , então $\varphi'(t) = 0$. Daí

$$(A(x + td) - b)^T d = 0$$

$$(Ax + tAd - b)^T d = 0$$

$$(Ax - b)^T d + t(Ad)^T d = 0$$

$$t = \frac{-(Ax - b)^T d}{d^T Ad}$$

Se $d = -\nabla f(x) = Ax - b$, obtemos

$$t = \frac{d^T d}{d^T Ad}.$$

2.3.1.2 Busca inexata e condição de Armijo

Na busca linear inexata, ao invés de resolver (7) exatamente, busca-se um tamanho de passo t que garanta um decréscimo suficiente no valor da função objetivo. Uma condição de decréscimo suficiente é dada pela condição de Armijo:

$$f(x + td) \leq f(x) + \rho t \nabla f(x)^T d.$$

em que $\rho \in (0, 1)$.

Se d é uma direção de descida para f a partir de x , então a existência de $t > 0$ satisfazendo a condição de Armijo é dada pelo seguinte teorema.

Teorema 2.3.4. [9, Teorema 6.1.3] *Sejam $x, d \in \mathbb{R}^n$ tais que $\nabla f(x)^T \neq 0$, $\nabla f(x)^T d < 0$ e $\rho \in (0, 1)$. Existe $\varepsilon = \varepsilon(\rho) > 0$ tal que*

$$f(x + td) \leq f(x) + \rho t \nabla f(x)^T d \tag{8}$$

para todo $t \in (0, \varepsilon]$.

Demonstração. Note que

$$\nabla f(x)^T d = \lim_{t \rightarrow 0} \frac{f(x + td) - f(x)}{t} \implies \lim_{t \rightarrow 0} \frac{f(x + td) - f(x)}{t \nabla f(x)^T d} = 1.$$

Assim, existe um $\varepsilon > 0$ tal que para todo $t \in (0, \varepsilon]$, temos

$$\frac{f(x + td) - f(x)}{t \nabla f(x)^T d} \geq \rho.$$

ou seja, para todo $t \in (0, \varepsilon]$,

$$f(x + td) \leq f(x) + \rho t \nabla f(x)^T d.$$

□

Esse resultado nos mostra que quando d satisfaz a condição $\nabla f(x)^T d < 0$, então a regra de Armijo está bem definida (é possível encontrar um $t > 0$ aceitável, após um número finito de reduções de um valor inicial $\bar{t} > 0$).

A seguir, apresentamos o algoritmo básico de busca linear associado a condição de Armijo.

Algoritmo 1: Busca linear inexata com condição de Armijo

Entrada: $x_k \in \mathbb{R}^n$, $d_k \in \mathbb{R}^n$ uma direção de descida,
 $\alpha, \beta \in (0, 1)$

Inicialização: Faça $t = 1$

Enquanto $f(x_k + td_k) > f(x_k) + \rho t \nabla f(x_k)^T d$

Faça: $t = \beta t$

2.3.1.3 Busca não-monótona

Na busca linear monótona, o tamanho de passo t_k é escolhido de modo que $f(x_{k+1}) < f(x_k)$. Por outro lado, na busca linear não-monótona é permitido algum crescimento no valor da função objetivo

de uma iteração para a seguinte. No contexto de algoritmos de busca linear, Grippo et al. [5] propuseram a regra de Armijo modificada para determinar o tamanho de passo: dados $\rho \in (0, 1)$, $M \in \mathbb{Z}_+^*$, escolha $t_k > 0$ que satisfaça a condição

$$f(x_k + t_k d_k) \leq \max_{0 \leq j \leq m_k} [f(x_{k-j})] + \rho t_k \nabla f(x_k)^T d_k \quad (9)$$

em que m_k é uma sequência de inteiros não-decrescente e limitada por M , $m_0 = 0$ e para $k > 0$, $0 \leq m_k \leq \min\{m_{k-1} + 1, M\}$.

Ao invés de impor um decréscimo de uma iteração para a seguinte, a condição (9) pede apenas uma melhoria em relação ao maior valor funcional das últimas M iterações.

3 Métodos de Descida

Os métodos de descida são métodos iterativos que a partir de um ponto inicial $x_0 \in \mathbb{R}^n$ dado, geram uma sequência $\{x_k\}_{k=0}^{\infty} \subset \mathbb{R}^n$ da seguinte maneira:

$$x_{k+1} = x_k + t_k d_k \quad (10)$$

em que $d_k \in \mathbb{R}^n$ é uma direção tal que $\nabla f(x_k)^T d_k < 0$, e que chamaremos de *direção de descida* (veja Lema 2.3.3) e $t_k > 0$ é o tamanho de passo que pode ser determinado por uma das estratégias descritas ao final do Capítulo 2.

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 e o método do gradiente espectral. Além disso, estudaremos também o método de Nesterov. Para estes métodos vamos investigar tanto a complexidade de pior caso quanto o desempenho prático.

3.1 Método de Cauchy

Considerado como um dos métodos mais antigos de minimização o Método da máxima descida, também conhecido como Método do Gradiente ou Método de Cauchy, foi desenvolvido por Augustin-Louis Cauchy em 1847. É um método de descida cuja direção corresponde ao oposto do vetor gradiente da função objetivo e o tamanho de passo é determinado de modo a minimizar a função ao longo dessa direção [4].

Normalmente, o termo “método do gradiente” é empregado para métodos que usam como direção de busca $d_k = -\nabla f(x_k)$ e cuja iteração é descrita por

$$x_{k+1} = x_k - t_k \nabla f(x_k) \quad (11)$$

Um protótipo de tais métodos é descrito no Algoritmo 2.

Algoritmo 2: Método do gradiente

Entrada: $x_0 \in \mathbb{R}^n$, $\varepsilon > 0$, $N \in \mathbb{Z}_+$

Inicialização: Faça $k = 0$

Enquanto $\|\nabla f(x_k)\| > \varepsilon$ e $k \leq N$, faça:

Passo 1: Defina $d_k = -\nabla f(x_k)$ e obtenha $t_k > 0$;

Passo 2: Atualize $x_{k+1} = x_k - t_k \nabla f(x_k)$, $k \leftarrow k + 1$.

Retorne: x_k

O algoritmo acima deixa em aberto o tamanho de passo t_k que podemos calcular de diferentes maneiras como vimos no Capítulo 2.

Para os experimentos que serão apresentados adiante neste trabalho, iremos considerar funções convexas e com gradiente Lipschitz. Mais especificamente, vamos considerar funções quadráticas positivas definidas pois minimizar esse tipo de função é um dos problemas mais fundamentais na área da otimização, fazendo também com que seja utilizado frequentemente como subproblema auxiliar em algoritmos para resolver problemas mais complicados [9]. Neste caso, daremos particular atenção a duas escolhas de tamanho de passo:

- Passo pré-definido: $\frac{1}{L}$, em que $L > 0$ é a constante de Lipschitz de $\nabla f(x)$;
- Busca exata: $t_k = \arg \min_t f(x_k - t \nabla f(x_k))$ e $t > 0$.

Embora o termo “método de Cauchy” (ou método de máxima descida) seja mais frequentemente utilizado para designar o método do gradiente com busca linear exata, com abuso de linguagem, neste trabalho iremos utilizar “método de Cauchy com passo pré-definido” e “método de Cauchy com busca exata” para nomear o Algoritmo 2 com estas respectivas escolhas de tamanho de passo.

Em relação ao tamanho de passo $t_k = 1/L$, os resultados a seguir levam à *complexidade de iteração* do método de Cauchy: qual estimativa de iterações são necessárias para obter uma “solução aproximada com precisão ε ”.

Proposição 3.1.1. [11, p. 29] *Seja $f : \mathbb{R}^n \rightarrow \mathbb{R}$ convexa e diferenciável em que o gradiente é Lipschitz de constante $L > 0$. Então, para todo $k \geq 0$*

$$f(x_k) - f(x_{k+1}) \geq \frac{1}{2L} \|\nabla f(x_k)\|^2.$$

Demonstração. Como f tem gradiente Lipschitz de contante $L > 0$, da Proposição 2.1.2 temos

$$\begin{aligned} f(x_{k+1}) &\leq f(x_k) + \nabla f(x_k)^T (x_{k+1} - x_k) + \frac{L}{2} \|x_{k+1} - x_k\|^2 \\ &= f(x_k) - \nabla f(x_k)^T \left(\frac{1}{L} \nabla f(x_k)\right) + \frac{L}{2} \left\| \frac{-1}{L} \nabla f(x_k) \right\|^2 \\ &= f(x_k) - \frac{1}{L} \|\nabla f(x_k)\|^2 + \frac{1}{2L} \|\nabla f(x_k)\|^2 \\ &= f(x_k) - \frac{1}{2L} \|\nabla f(x_k)\|^2 \end{aligned}$$

Portanto,

$$f(x_k) - f(x_{k+1}) \geq \frac{1}{2L} \|\nabla f(x_k)\|^2$$

□

Teorema 3.1.2. [11, p. 30] *Seja $f : \mathbb{R}^n \rightarrow \mathbb{R}$ uma função convexa e diferenciável com constante de Lipschitz $L > 0$ para o gradiente. Seja x^* um minimizador de f em \mathbb{R}^n . Então, para todo $k \geq 1$*

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

Demonstração. Como f tem gradiente Lipschitz de constante $L > 0$, da Proposição 2.1.2 temos

$$f(x_{k+1}) \leq f(x_k) + \nabla f(x_k)^T (x_{k+1} - x_k) + \frac{L}{2} \|x_{k+1} - x_k\|^2$$

Usando o Lema 2.2.1 com $y = x^*$ e $x = x_k$

$$f(x^*) \geq f(x_k) + \nabla f(x_k)^T(x^* - x_k)$$

Dessa forma,

$$f(x_{k+1}) \leq f(x^*) - \nabla f(x_k)^T(x^* - x_k) + \nabla f(x_k)^T(x_{k+1} - x_k) + \frac{L}{2} \|x_{k+1} - x_k\|^2$$

Como $x_{k+1} = x_k - \frac{1}{L} \nabla f(x_k)$, então

$$f(x_{k+1}) - f(x^*) \leq L(x_k - x_{k+1})^T(x^* - x_{k+1}) + \frac{L}{2} \|x_{k+1} - x_k\|^2$$

Após manipulações algébricas com produto interno, chegamos à

$$f(x_{k+1}) - f(x^*) \leq \frac{L}{2} \|x_{k+1} - x^*\|^2 - \frac{L}{2} \|x_k - x^*\|^2$$

Observe que se tomamos $k = 0$ temos

$$f(x_1) - f(x^*) \leq \frac{L}{2} \|x_1 - x^*\|^2 - \frac{L}{2} \|x_0 - x^*\|^2$$

Se $k = 1$ temos

$$f(x_2) - f(x^*) \leq \frac{L}{2} \|x_2 - x^*\|^2 - \frac{L}{2} \|x_1 - x^*\|^2$$

E se fizermos o procedimento até k , temos

$$f(x_k) - f(x^*) \leq \frac{L}{2} \|x_k - x^*\|^2 - \frac{L}{2} \|x_{k-1} - x^*\|^2$$

Somando as k inequações teremos

$$k[f(x_k) - f(x^*)] \leq \frac{L}{2} \|x_k - x^*\|^2 - \frac{L}{2} \|x_0 - x^*\|^2$$

Usando o fato de que $\{f(x_k)\}$ é decrescente (Proposição 3.1.1), obtemos

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

□

O resultado acima mostra que o método de Cauchy precisa de $\mathcal{O}(k^{-1})$ iterações para obter $f(x_k) - f(x^*) < \varepsilon$, quando f é convexa.

3.2 Método de Nesterov

Conhecido como gradiente acelerado de Nesterov, esse método possui a melhor taxa de convergência para métodos de primeira ordem [12]. O método foi descrito por Yuri Nesterov em um artigo de 1983.

A partir de um ponto inicial $x_0 \in \mathbb{R}^n$, $\theta_0 = 1$, $v_0 = x_0$, o processo iterativo é definido por:

$$\left\{ \begin{array}{l} \text{Se } k \geq 1 : \text{ escolha } \theta_k \in (0, 1) \text{ tal que } \frac{(1 - \theta_k)t_k}{\theta_k^2} \leq \frac{t_k - 1}{\theta_{k-1}^2} \\ y_k = (1 - \theta_k)x_k + \theta_k v_k \\ x_{k+1} = y_k - t_k \nabla f(y_k) \\ v_{k+1} = x_k + \frac{1}{\theta_k}(x_{k+1} - x_k) \end{array} \right. \quad (12)$$

Neste método, em vez de x_k , o gradiente é avaliado em um ponto y_k , que é a combinação convexa de x_k com uma extrapolação de x_{k-1} ao longo da direção $x_k - x_{k-1}$. O parâmetro $\theta_k \in (0, 1)$ controla tanto esta combinação convexa quanto o passo de extrapolação e, por questões de convergência que veremos mais a frente, deve satisfazer a condição da primeira linha de (12).

É possível descrever a iteração do método sem o termo v_k . Neste caso, y_k pode ser escrito da forma $y_k = x_k + \beta_k(x_k - x_{k-1})$ em que $\beta_k = \theta_k(\theta_{k-1}^{-1} - 1)$.

Assim como no método de Cauchy, o tamanho de passo t_k no método de Nesterov pode ser calculado de diferentes maneiras, entre elas destacamos:

- Passo pré-definido: $\frac{1}{L}$, com $L > 0$ constante de Lipschitz para $\nabla f(x)$;
- Busca linear inexata (ou *Backtracking*): Escolher $t_k > 0$ que

satisfaça a condição [3]:

$$f(y_k - t_k \nabla f(y_k)) \leq f(y_k) - \frac{t_k}{2} \|\nabla f(y_k)\|^2 \quad (13)$$

Uma possível escolha de θ_k , satisfazendo a condição imposta em (12), é dada por $\theta_k = 2/(k+2)$, que leva ao Algoritmo 3.

Algoritmo 3: Gradiente Acelerado de Nesterov

Entrada: $x_0 \in \mathbb{R}^n$, $\varepsilon > 0$, $N \in \mathbb{Z}_+$.

Inicialização: Faça $k = 0$, $\theta_0 = 1$, $v_0 = x_0$ e $y_0 = x_0$

Enquanto: $\|\nabla f(x_k)\| > \varepsilon$ e $k \leq N$

Passo 1: Faça $\theta_k = \frac{2}{k+2}$ e $y_k = (1 - \theta_k)x_k + \theta_k v_k$

Passo 2: Faça $t_k = 1/L$ (ou determine t_k satisfazendo (13))

Passo 3: Atualize $x_{k+1} = y_k - t_k \nabla f(y_k)$

Passo 4: Calcule $v_{k+1} = x_k + \frac{1}{\theta_k}(x_{k+1} - x_k)$, $k \leftarrow k + 1$.

Retorne: x_k

O resultado a seguir estabelece a taxa de convergência $\mathcal{O}(k^{-2})$ para o método de Nesterov com tamanho de passo pré-definido.

Teorema 3.2.1. *Seja f convexa com gradiente Lipschitz. A iteração do método de Nesterov com tamanho de passo constante $t_k = t \in \left(0, \frac{1}{L}\right]$ e com $\theta_k = \frac{2}{k+2}$ satisfaz, para todo $k \geq 1$*

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

Demonstração. Seja $x_{k+1} = y_k - t \nabla f(y_k)$. Como $0 < t < 1/L$, então segue do Proposição 2.1.2

$$f(x_{k+1}) \leq f(y_k) + \nabla f(y_k)^T (x_{k+1} - y_k) + \frac{1}{2t} \|x_{k+1} - y_k\|^2 \quad (14)$$

Como f é convexa, temos do Lema 2.2.1 que para qualquer $z \in \mathbb{R}^n$:

$$f(y_k) - f(z) \leq \nabla f(y_k)^T (y_k - z) \quad (15)$$

Combinando (14) e (15), e usando o fato que $\nabla f(y_k) = -\frac{1}{t}(x_{k+1} - y_k)$ obtemos,

$$f(x_{k+1}) \leq f(z) + \frac{1}{t}(x_{k+1} - y_k)^T(z - x_{k+1}) + \frac{1}{2t}\|x_{k+1} - y_k\|^2. \quad (16)$$

Iremos avaliar (16) para $z = x^*$ e $z = x_k$ e considerar a combinação convexa com pesos $\{\theta, 1 - \theta\}$. Com isso, teremos

$$\begin{aligned} \theta(f(x_{k+1}) - f(x^*)) + (1 - \theta)(f(x_{k+1}) - f(x_k)) &\leq \frac{1}{t}(x_{k+1} - y_k)^T(\theta x_{k+1} + (1 - \theta)x_k - x_{k+1}) + \\ &\quad + \frac{1}{2t}\|x_{k+1} - y_k\|^2 \end{aligned}$$

e simplificando,

$$\begin{aligned} f(x_{k+1}) - f(x^*) - (1 - \theta)(f(x_k) - f(x^*)) &\leq \frac{1}{2t}[\|\theta x_{k+1} + (1 - \theta)x_k - y_k\|^2 \\ &\quad - \|\theta x_{k+1} + (1 - \theta)x_k - x_{k+1}\|^2]. \end{aligned}$$

Sabemos, pelo processo iterativo, que $y_k = (1 - \theta)x_k + \theta v_k$. Logo

$$f(x_{k+1}) - f(x^*) - (1 - \theta)(f(x_k) - f(x^*)) \leq \frac{\theta^2}{2t}[\|x^* - v_k\|^2 - \|x^* - v_{k+1}\|^2]$$

Reorganizando para colocar as iterações $k + 1$ em um lado da desigualdade e as iterações k no outro lado, obtemos

$$\frac{t}{\theta^2_k}(f(x_{k+1}) - f(x^*)) + \frac{1}{2}\|x^* - v_{k+1}\|^2 \leq \frac{(1 - \theta_k)t}{\theta^2_k}(f(x_k) - f(x^*)) + \frac{1}{2}\|x^* - v_k\|^2$$

Agora, usando fato de que $\frac{(1 - \theta_k)t_k}{\theta^2_k} \leq \frac{t_{k-1}}{\theta^2_{k-1}}$, teremos

$$\frac{t}{\theta^2_k}(f(x_{k+1}) - f(x^*)) + \frac{1}{2}\|x^* - v_{k+1}\|^2 \leq \frac{t}{\theta^2_{k-1}}(f(x_k) - f(x^*)) + \frac{1}{2}\|x^* - v_k\|^2$$

A desigualdade acima nos diz que a quantidade de

$$V_k = \frac{t}{\theta^2_{k-1}}(f(x_k) - f(x^*)) + \frac{1}{2}\|x^* - v_k\|^2$$

é não-crescente com k . Assim temos $V_k \leq V_{k-1} \leq \dots \leq V_1$ que resulta em

$$\begin{aligned} \frac{t}{\theta_k^2} (f(x_{k+1}) - f(x^*)) + \frac{1}{2} \|x^* - v_{k+1}\|^2 &\leq \frac{t}{\theta_0^2} (f(x_k) - f(x^*)) + \frac{1}{2} \|x^* - v_k\|^2 \\ &\leq \frac{t}{\theta_{k-1}^2} (f(x_k) - f(x^*)) + \frac{1}{2} \|x^* - v_k\|^2 \\ &= \frac{1}{2} \|x^* - x_0\|^2. \end{aligned}$$

Portanto, $f(x_k) - f(x^*) \leq \frac{\theta_{k-1}^2}{2t} \|x_0 - x^*\|^2$, e com $\theta_{k-1} = \frac{2}{k+1}$ temos que $f(x_k) - f(x^*) \leq \frac{2}{(k+1)^2 t} \|x_0 - x^*\|^2$. \square

Perceba que através do teorema acima no método de Nesterov é possível que ocorra $f(x_{k+1}) > f(x_k)$, ao contrário do método de Cauchy. No entanto, a demonstração acima nos mostra que uma determinada combinação de $f(x_k) - f(x^*)$ e $\|x^* - v_k\|^2$ diminui ao longo das iterações.

Ainda, em comparação com o método de Cauchy, este último possui taxa de convergência $\mathcal{O}(k^{-1})$ enquanto o Teorema 3.2.1 mostra a taxa $\mathcal{O}(k^{-2})$ para o método de Nesterov. Assim, a medida que k aumenta, o limitante superior para a diferença $f(x_k) - f(x^*)$ irá para zero mais rapidamente no método de Nesterov que no método de Cauchy. É mencionado na literatura que este limitante do método de Nesterov é ótimo dentre métodos de primeira ordem. Mais evidência da taxa de convergência do método de Nesterov será obtida na Seção 5.4 do Capítulo 5.

3.3 Método do Gradiente Espectral

Introduzido por Barzilai e Borwein [1], este método considera como direção de busca $d_k = -\lambda_k^{-1} \nabla f(x_k)$, de modo que a iteração é dada por

$$x_{k+1} = x_k - t_k \lambda_k^{-1} \nabla f(x_k)$$

em que o escalar

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

é conhecido na literatura como parâmetro espectral 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”[1]. Mais ainda, essa razão é uma estimativa da curvatura de f ao longo da direção $x_k - x_{k-1}$.

Neste método é utilizada a busca linear não-monótona proposta por Grippo et al. [15] que foi discutida na Seção 2.3.1.3. De fato, Barzilai e Borwein observaram que essa nova escolha de tamanho de passo exige menos trabalho computacional e que também acelera a convergência do método do gradiente espectral para funções quadráticas convexas.

Para este método de gradiente espectral com busca não-monótona é possível provar que para $\varepsilon \in (0, 1]$, f convexa e com gradiente Lipschitz, temos que $f(x_k) - f(x^*) \leq \varepsilon$, em no máximo $\mathcal{O}(\varepsilon^{-1})$ iterações. Esta análise, baseada no trabalho [5], é apresentada em detalhes no próximo capítulo.

A seguir, apresentamos o algoritmo associado ao Método do Gradiente Espectral.

Algoritmo 4: Gradiente Espectral

Entrada: $x_0 \in \mathbb{R}^n$, $0 < \delta_{\min} < \delta_{\max}$, ε , $\bar{t}_0 > 0$, $\beta, \rho \in (0, 1)$,
 $1 \leq M$, $N \in \mathbb{Z}_+$

Inicialização: Faça $k = 0$, $\lambda_0 = 1$.

Passo 1: Se $\|\nabla f(x_k)\| \leq \varepsilon$ ou $k > N$, pare.

Passo 2: Se $k \geq 1$, defina $s_k = x_k - x_{k-1}$ e
 $g_k = \nabla f(x_k) - \nabla f(x_{k-1})$ e calcule

$$\lambda_k = \max \left(\delta_{\min}, \min \left(\delta_{\max}, \frac{s_k^T g_k}{s_k^T s_k} \right) \right)$$

Passo 3: defina $d_k = -\lambda_k^{-1} \nabla f(x_k)$, $t = \bar{t}_0$ e

$$f_{\max} = \max\{f(x_{k-j}) \mid 0 \leq j \leq \min(k, M)\}$$

Passo 4: Se: $f(x_k + td_k) < f_{\max} + \rho t \nabla f(x_k)^T d_k$, então
defina $t_k \leftarrow t$.

Do contrário: Calcule $t \leftarrow t\beta$ e retorne ao **Passo 4**

Passo 5 Atualize $x_{k+1} = x_k + t_k d_k$, $t_0 = t_k/\beta$, $k \leftarrow k + 1$ e
vá para o **Passo 1**.

Retorne: x_k

4 Complexidade de Pior caso do Método do Gradiente Espectral

Diferente das variações do método de Cauchy que estudamos até aqui, nas quais a busca linear usada garante um decréscimo monótono no valor da função objetivo, vimos que no método de gradiente espectral é utilizada uma busca não-monótona. Com isso, precisamos de uma teoria diferente para estabelecer a complexidade de iteração do gradiente espectral com busca não-monótona.

Neste capítulo faremos um estudo sobre a complexidade de pior caso de métodos de direções de descida que usam busca linear não-monótona para determinar o tamanho de passo. Todos os resultados aqui apresentados são dos trabalhos de Grapiglia e Sachs [5] e Sachs e Sachs [15]. Este estudo permitirá estabelecer a complexidade de iteração para o gradiente espectral, visto como um caso particular de um algoritmo mais geral.

Considere o protótipo de algoritmo listado no Algoritmo 5.

Algoritmo 5: Algoritmo de descida com busca não-monótona

Entrada: $x_0 \in \mathbb{R}^n$, $t_0 > 0$, $\beta, \rho \in (0, 1)$ e $\nu_0 > 0$

Inicialização: Faça $k = 0$.

Passo 1: Calcule uma d_k tal que $\nabla f(x_k)^T d_k < 0$ para f a partir de x_k

Passo 2: Encontre o menor inteiro $l_k \geq 0$ tal que

$$f(x_k + t_k \beta^{l_k} d_k) \leq f(x_k) + \rho t_k \beta^{l_k} \nabla f(x_k)^T d_k + \nu_k \quad (17)$$

Passo 3: Calcule $\nu_{k+1} \geq 0$, faça $x_{k+1} = x_k + t_k \beta^{l_k} d_k$, $t_{k+1} = t_k \beta^{l_{k+1}}$, $k \leftarrow k + 1$ e vá para o **Passo 1**.

A busca (17) que vamos considerar para provar a complexidade de iteração do Algoritmo 5 é mais geral do que a busca (9) no sentido que (9) é um caso particular de (17) para uma escolha adequada para

a sequência ν_k [5] e, por consequência, o Algoritmo 4 pode ser visto como um caso particular do Algoritmo 5.

O Algoritmo 5 proposto em [5], é um algoritmo de descida com busca não-monótona um pouco diferente da busca de Grippo et. al [6], e permite a não monotonia através da sequência de parâmetros positivos $\{\nu_k\}$.

Para analisarmos a complexidade do Algoritmo 5, inicialmente faremos uma estimativa do tamanho de passo t_k através do seguinte lema.

Lema 4.0.1. [5, **Lema 2**] *Seja $f : \mathbb{R}^n \rightarrow \mathbb{R}$ com gradiente Lipschitz de constante $L > 0$. Suponha que existam $0 < \delta_{\min} < \delta_{\max} < \infty$, tais que, para todo k ,*

$$\nabla f(x_k)^T d_k \leq -\delta_{\min} \|\nabla f(x_k)\|^2 \quad e \quad \|d_k\| \leq \delta_{\max} \|\nabla f(x_k)\|.$$

Então

$$t_k \geq \min \left\{ t_0, \frac{2(1-\rho)\delta_{\min}}{L\delta_{\max}^2} \right\} \equiv \bar{t}. \quad (18)$$

Demonstração. Faremos a demonstração desse lema por indução para l_k . Assuma a validade de (18) para um k arbitrário. Veja que $t_{k+1} = t_k \beta^{l_k - 1}$. Para $l_k = 0$, obtemos

$$t_{k+1} \geq t_k \beta^{-1} \geq t_k \geq \bar{t}.$$

Agora, para $l_k \geq 1$ por (17) e pela definição de t_{k+1} , obtemos

$$f(x_k + t_{k+1}d_k) - f(x_k) > \rho t_{k+1} \nabla f(x_k)^T d_k + \nu_k \quad (19)$$

Por outro lado, como f tem gradiente Lipschitz, temos

$$f(x_k + t_{k+1}d_k) - f(x_k) \leq t_{k+1} \nabla f(x_k)^T d_k + \frac{L t_{k+1}^2}{2} \|d_k\|^2$$

Disto e por (19), obtemos

$$\rho t_{k+1} \nabla f(x_k)^T d_k + \nu_k \leq t_{k+1} \nabla f(x_k)^T d_k + \frac{L t_{k+1}^2}{2} \|d_k\|^2$$

Dividindo a expressão acima por t_{k+1} e simplificando, temos

$$\begin{aligned} \rho \nabla f(x_k)^T d_k + \frac{v_k}{t_{k+1}} &\leq \nabla f(x_k)^T d_k + \frac{L t_{k+1}}{2} \|d_k\|^2 \\ \frac{v_k}{t_{k+1}} &\leq (1 - \rho) \nabla f(x_k)^T d_k + \frac{L t_{k+1}}{2} \|d_k\|^2 \\ \frac{v_k}{t_{k+1}} &\leq \frac{2 \nabla f(x_k)^T d_k (1 - \rho) + L t_{k+1} \|d_k\|^2}{2} \\ 2v_k &\leq t_{k+1} [2(1 - \rho) \nabla f(x_k)^T d_k + L t_{k+1} \|d_k\|^2] \\ \frac{2v_k}{L \|d_k\|^2 t_{k+1}} &\leq \frac{2(1 - \rho) \nabla f(x_k)^T d_k}{L \|d_k\|^2} + t_{k+1} \end{aligned}$$

Assim, podemos fazer uma estimativa de t_{k+1} da forma

$$\begin{aligned} t_{k+1} &\geq \frac{2(1 - \rho)}{L} \left(\frac{-\nabla f(x_k)^T d_k}{\|d_k\|^2} \right) + \frac{2v_k}{L \|d_k\|^2 t_{k+1}} \\ &\geq \frac{2(1 - \rho) \delta_{\min}}{L \delta_{\max}^2}. \end{aligned}$$

□

Com este resultado auxiliar, já podemos apresentar um resultado de complexidade para funções f com gradiente Lipschitz, não necessariamente convexas. Nestes casos, em vez de determinar o número máximo de iterações para obter $f(x_k) - f(x^*) < \varepsilon$, queremos garantir que $\|\nabla f(x_k)\| < \varepsilon$.

Teorema 4.0.2. [5, Teorema 1] *Seja $f : \mathbb{R}^n \rightarrow \mathbb{R}$ continuamente diferenciável e com gradiente Lipschitz, com $L > 0$. Suponha que existam $0 < \delta_{\min} < \delta_{\max} < \infty$ tais que, para todo k ,*

$$\nabla f(x_k)^T d_k \leq -\delta_{\min} \|\nabla f(x_k)\|^2 \text{ e } \|d_k\| \leq \delta_{\max} \|\nabla f(x_k)\|.$$

Além disso, assuma que existe $f_{\min} \in \mathbb{R}$ tal que $f(x) \geq f_{\min}, \forall x \in \mathbb{R}^n$ e a sequência $\{\nu_k\}$ é somável. Sob tais hipóteses, dado $\varepsilon \in (0, 1]$ o algoritmo leva no máximo $\mathcal{O}(\varepsilon^{-2})$ iterações para que possa garantir $\|\nabla f(x_k)\| \leq \varepsilon$.

Demonstração. Manipulando a expressão (17) do algoritmo, temos

$$\nu_k + f(x_k) - f(x_{k+1}) \geq \rho t_k \beta^{l_k} (-\nabla f(x_k))^T d_k$$

Usando as hipóteses sobre d_k , reescrevemos a expressão acima da forma

$$\nu_k + f(x_k) - f(x_{k+1}) \geq \rho t_{k+1} \beta \delta_{\min} \|\nabla f(x_k)\|^2. \quad (20)$$

Seja $j_1 < +\infty$ a primeira iteração a satisfazer $\|\nabla f(x_k)\| \leq \varepsilon$. Assim, $\|\nabla f(x_k)\| > \varepsilon$, para $k = 0, 1, \dots, j_1$. Logo, segue de (20) que

$$\nu_k + f(x_k) - f(x_{k+1}) \geq k_c \varepsilon^2$$

em que k_c é dado por

$$k_c = \min \left\{ \rho \beta t_0 \delta_{\min}, \frac{2\beta \rho (1 - \rho) \delta_{\min}^2}{L \delta_{\max}^2} \right\}$$

Disto, do fato de f ser limitada inferiormente e de ν_k ser somável, obtemos

$$\begin{aligned} (j_1 + 1)k_c \varepsilon^2 &= \sum_{k=0}^{j_1} k_c \varepsilon^2 \\ &\leq \sum_{k=0}^{j_1} (\nu_k + f(x_k) - f(x_{k+1})) \\ &= f(x_0) - f(x_{j_1+1}) + \sum_{k=0}^{j_1} \nu_k \\ &\leq f(x_0) - f_{\min} + \sum_{k=0}^{\infty} \nu_k. \end{aligned}$$

Portanto, segue que $j_1 + 1 \leq \left(\frac{f(x_0) - f_{\min} + \sum_{k=0}^{\infty} \nu_k}{k_c} \right) \varepsilon^{-2}$. \square

Agora veremos que se além das hipóteses acima f também for convexa, com valor ótimo $f(x^*)$, então o número máximo de iterações para obter $f(x_k) - f(x^*) < \varepsilon$ é $\mathcal{O}(\varepsilon^{-1})$.

Teorema 4.0.3. [5, Teorema 2] *Seja $f : \mathbb{R}^n \rightarrow \mathbb{R}$ convexa e continuamente diferenciável, com gradiente Lipschitz de constante $L > 0$. Assuma que existam $0 < \delta_{\min} < \delta_{\max} < \infty$, tais que, para todo k ,*

$$\nabla f(x_k)^T d_k \leq -\delta_{\min} \|\nabla f(x_k)\|^2 \text{ e } \|d_k\| \leq \delta_{\max} \|\nabla f(x_k)\|.$$

Seja $x^ \in \mathbb{R}^n$ o minimizador global de f . Então, para $\varepsilon \in (0, 1]$, o algoritmo leva*

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

iterações para garantir $f(x_k) - f(x^) \leq \varepsilon$.*

Demonstração. Lembrando de (20) que

$$v_k + f(x_k) - f(x_{k+1}) \geq \rho\beta\delta_{\min} t_{k+1} \|\nabla f(x_k)\|^2.$$

Assim, temos que

$$t_{k+1} \|\nabla f(x_k)\|^2 \leq \frac{f(x_k) - f(x_{k+1}) + v_k}{\rho\beta\delta_{\min}} \quad (21)$$

Suponha que $j_1 < +\infty$ seja o primeiro índice tal que $f(x_{j_1}) - f(x^*) \leq \varepsilon$. Logo, $f(x_k) - f(x^*) > \varepsilon$, para $k = 0, 1, \dots, j_1 - 1$.

Veja que, de (21) obtemos

$$\sum_{k=0}^{j_1-1} t_{k+1} \|\nabla f(x_k)\|^2 \leq \frac{f(x_0) - f(x^*) + (\sum_{k=0}^{\infty} v_k)}{\rho\beta\delta_{\min}} \quad (22)$$

Por outro lado,

$$\begin{aligned} \|x_{k+1} - x^*\|^2 &= \|(x_k - x^*) - t_{k+1} \beta \lambda_k \nabla f(x_k)\|^2 \\ &= \|x_k - x^*\|^2 - 2t_{k+1} \lambda_k \beta (x_k - x^*)^T \nabla f(x_k) + t_{k+1}^2 \beta^2 \lambda^2 \|\nabla f(x_k)\|^2 \\ &\leq \|x_k - x^*\|^2 - 2t_{k+1} \lambda_k \beta (f(x_k) - f(x^*))^T \nabla f(x_k) + t_{k+1}^2 \beta^2 \lambda^2 \|\nabla f(x_k)\|^2 \end{aligned}$$

em que a última desigualdade segue do fato de f ser convexa. Dessa forma,

$$f(x_k) - f(x^*) \leq \frac{1}{2\delta_{\min}\beta\bar{t}} (\|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2) + \frac{\delta_{\max}\beta}{2} t_{k+1} \|\nabla f(x_k)\|^2.$$

Fazendo a soma dessas desigualdades para $k = 0, 1, \dots, j_1 - 1$ e usando (22) conseguimos

$$\sum_{k=0}^{j_1-1} (f(x_k) - f(x^*)) \leq \frac{(\beta\bar{t})^{-1} \|x_0 - x^*\|^2 + \rho^{-1} \delta_{\max} [f(x_0) - f(x^*) + (\sum_{k=0}^{+\infty} v_k)]}{2\delta_{\min}}$$

Disto e pelo fato de $f(x_k) - f(x^*) > \varepsilon$, concluímos que

$$j_1 \varepsilon \leq \frac{(\beta\bar{t})^{-1} \|x_0 - x^*\|^2 + \rho^{-1} \delta_{\max} [f(x_0) - f(x^*) + (\sum_{k=0}^{+\infty} v_k)]}{2\delta_{\min}}$$

ou seja,

$$j_1 \leq \left(\frac{(\beta\bar{t})^{-1} \|x_0 - x^*\|^2 + \rho^{-1} \delta_{\max} [f(x_0) - f(x^*) + (\sum_{k=0}^{+\infty} v_k)]}{2\delta_{\min}} \right) \varepsilon^{-1}.$$

□

Note que as constantes que multiplicam ε^{-2} e ε^{-1} nos resultados acima dependem de $\sum_{k=0}^{\infty} v_k$. Tais constantes poderiam ser menores se um decréscimo no valor da função for imposto a cada iteração ($v_k = 0$). No entanto, uma vantagem de não exigir um comportamento monótono das iterações externas é que as iterações internas (backtracking) devem terminar em um estágio anterior [5].

O teorema seguinte nos dá uma estimativa para o número total de avaliações de funções após $k \geq 1$ iterações.

Teorema 4.0.4. *Considere as mesmas hipóteses do Teorema 4.0.3. Seja N_k o número total de avaliações da função após $k \geq 1$ iterações do algoritmo. Então*

$$N_k \leq 2(k+1) + (\log \beta)^{-1} [\log(\bar{t}) - \log(t_0)]. \quad (23)$$

Demonstração. Por definição de t_{k+1} no algoritmo e aplicando as propriedades de logaritmo, obtemos

$$\begin{aligned}
 t_{k+1} &= \beta^{l_k - 1} t_k \\
 \log(t_{k+1}) &= (l_k - 1) \log(\beta) + \log(t_k) \\
 (l_k - 1) \log \beta &= \log(t_{k+1}) - \log(t_k) \\
 l_k - 1 &= (\log \beta)^{-1} [\log(t_{k+1}) - \log(t_k)] \\
 l_k - 1 + 2 &= 2 + (\log \beta)^{-1} [\log(t_{k+1}) - \log(t_k)] \\
 l_k + 1 &= 2 + (\log \beta)^{-1} [\log(t_{k+1}) - \log(t_k)]
 \end{aligned}$$

Percebe-se que, a cada iteração do algoritmo $l_k + 1$ avaliações da função são necessárias. Assim, o número total de avaliações de função após $k \geq 1$ iterações satisfaz a relação

$$\begin{aligned}
 N_k &= \sum_{i=0}^k (l_i + 1) \\
 &\leq 2(k + 1) + (\log \beta)^{-1} [\log(t_{k+1}) - \log(t_0)] \\
 &\leq 2(k + 1) + (\log \beta)^{-1} [\log(\bar{t}) - \log(t_0)]
 \end{aligned}$$

□

Os resultados neste capítulo mostram que o Algoritmo 5 possui complexidade de iteração $\mathcal{O}(\varepsilon^{-1})$ quando f é convexa e com gradiente Lipschitz. Em outras palavras, uma estimativa do número de iterações necessárias para obter $f(x_k) - f(x^*) < \varepsilon$ é de no máximo $\mathcal{O}(k^{-1})$. Note que a complexidade de iteração para o método do gradiente espectral é a mesma complexidade no método de Cauchy.

Como o Algoritmo 4 é um caso particular do Algoritmo 5, segundo [5], podemos dizer que o método do gradiente espectral tem taxa de convergência de $\mathcal{O}(k^{-1})$.

5 Experimentos Numéricos

Para avaliar o desempenho prático dos métodos, neste capítulo, apresentaremos experimentos numéricos envolvendo funções quadráticas com Hessiana simétrica definida positiva com dimensão variando de 2 a 5000. Apresentaremos também a “pior função do mundo” a fim de evidenciar a performance de pior caso dos métodos.

5.1 Detalhes de implementação

Todos os algoritmos dos métodos apresentados foram implementados em linguagem Matlab R2013a, e todos os testes foram realizados em um Notebook Acer, 32GB RAM, processador Intel core i3, sistema operacional Windows 10.

Para todos os algoritmos, definimos como ponto inicial o vetor $x_0 = (1, 1, \dots, 1)^T \in \mathbb{R}^n$. Os critérios de parada utilizados foram:

- $\|\nabla f(x_k)\| < 10^{-6}$.
- número máximo de iterações 10^5 .

Se a execução for interrompida por ter atingido o número máximo de iterações, consideramos um insucesso.

Na implementação do Algoritmo 2, consideramos dois tamanhos de passo: busca linear exata (veja Seção 2.3.1.1)

$$t_k = \frac{\nabla f(x)^T \nabla f(x)}{\nabla f(x)^T A \nabla f(x)}$$

e tamanho de passo pré-definido $t_k = \frac{1}{L}$, em que $L > 0$ é a constante de Lipschitz do gradiente.

Na implementação do Algoritmo 3, definimos os parâmetros iniciais $\theta_0 = 1$, $v_0 = x_0$ e $y_0 = x_0$ e consideramos como estratégias de tamanho de passo $t_k = \frac{1}{L}$ e o procedimento de backtracking de modo

a satisfazer (13) com passo tentativo inicial $t_0 = 1$ e fator de redução de passo $\beta = 0,8$.

Por sua vez, na implementação do Algoritmo 4, utilizamos como parâmetros $\delta_{\min} = 10^{-10}$, $\delta_{\max} = 10^{10}$, $\lambda_0 = 1$. Definimos como parâmetro da busca linear não-monótona $M = 10$, valor de passo inicial $\bar{t}_0 = 1$, $\beta = 0,8$ e $\rho = 0,5$.

5.2 Funções Quadráticas

Para problemas testes, foram consideradas funções quadráticas com Hessiana simétrica positiva definida da forma $f(x) = \frac{1}{2}x^T Ax$.

Com o objetivo de gerar a matriz A simétrica positiva definida com autovalores positivos, geramos inicialmente uma matriz aleatória $M \in \mathbb{R}^{n \times n}$ com entradas aleatórias de uma distribuição normal padrão. A seguir, ortonormalizamos as colunas dessa matriz utilizando o processo de Gram-Schmidt e obtemos a matriz $P \in \mathbb{R}^{n \times n}$.

Definimos uma matriz diagonal $D \in \mathbb{R}^{n \times n}$ tal que $D_{ii} = \lambda_i$ em que $\lambda_i > 0, i = 1, 2, \dots, n$ serão os autovalores de A . Assim, obtemos

$$A = PDP^T.$$

Note que, $\nabla f(x) = Ax$. Para estimar a constante de Lipschitz utilizamos

$$\|\nabla f(x) - \nabla f(y)\| = \|Ax - Ay\| \leq \|A\| \|x - y\|,$$

em que $\|\cdot\|$ denota a norma - 2 e norma matricial induzida.

5.3 Resultados dos testes

A fim de estudar a performance dos algoritmos, para cada dimensão de problema teste consideramos três diferentes distribuições de autovalores que vamos descrever abaixo.

Na distribuição que chamaremos de av_1 os autovalores são da seguinte forma $\lambda_i = i$, para $i = 1, \dots, n$. Na distribuição av_2 , os autovalores são do tipo $\lambda_i = 1$, para $i = 1, \dots, n - 1$ e $\lambda_n = 2n - 3$ e para a distribuição av_3 os autovalores são da forma $\lambda_i = \gamma_i$, em que γ_i é um número aleatório segundo uma distribuição uniforme no intervalo $[0, 1]$.

Essa escolha de distribuições de autovalores se deu pelo fato de que métodos de tipo gradiente são fortemente afetados pelo número de condição da Hessiana da função quadrática. Neste caso, como A é simétrica, positiva definida, o número de condição é quociente entre o maior e o menor autovalor da matriz.

Nas tabelas de 1 a 3, apresentamos a média de iterações (sobre 5 exemplares) para cada variação dos métodos estudados.

n	Método Cauchy		Método Nesterov		Gradiente Espectral
	$1/L$	Busca Exata	$1/L$	Backtracking	Não-monótona
2	13,4	9,2	15,4	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	71276,8	25783,5	8543	6752	1692

Tabela 1 – Média das iterações com autovalores av_1 .

n	Método Cauchy		Método Nesterov		Gradiente Espectral
	$1/L$	Busca Exata	$1/L$	Backtracking	Não-monótona
2	1	1	1	6,2	4
5	78,6	28,8	42	33,8	9,4
10	246,4	26,4	105	29,6	9
50	1520,6	80,2	316	449,2	13,2
100	3167,2	784	495	895,4	12,8
500	16862	155	1611	257,6	10
1000	34478,2	72,2	2842	857,8	50
5000	49521,6	998,5	5103	1789,4	352

Tabela 2 – Média das iterações com autovalores av_2 .

n	Método Cauchy		Método Nesterov		Gradiente Espectral
	$1/L$	Busca Exata	$1/L$	Backtracking	Não-monótona
2	17,4	6,2	13	16,8	5
5	62,4	44,2	29,6	26	16,8
10	178,4	102,6	47	33,6	28,7
50	675,8	2241	88,4	78	49,1
100	15376	5081	143,4	381,6	156
500	6113,6	2298,2	203	647	98,5
1000	11218,8	7249,4	269,2	347	201
5000	62181,8	15875	399,8	284	187

Tabela 3 – Média das iterações com autovalores av_3 .

Os resultados numéricos indicam que o método de gradiente espectral, em média, alcança o primeiro critério de parada em menos iterações que os demais.

Considerando os resultados obtidos, a seguir, a Figura 4 ilustra o valor $f(x_k)$ em função do número de iterações para um problema com $n = 100$ e distribuição de autovalores av_1 .

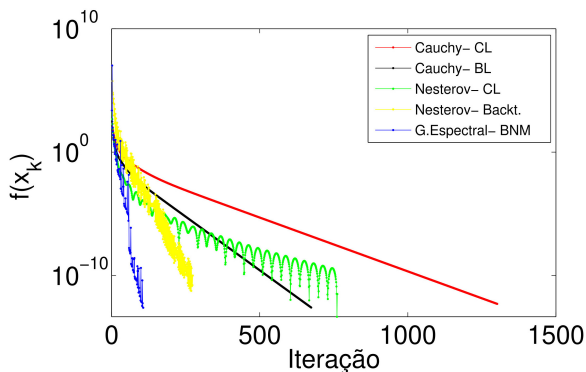


Figura 4 – Desempenho dos métodos

Neste gráfico observa-se 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 decrescimento monótono da função, em oposição aos métodos de Nesterov e gradiente espectral. É curioso notar que, apesar de possuir taxa de convergência ótima $\mathcal{O}(1/k^2)$, o método de Nesterov foi superado pelo gradiente espectral nos testes que realizamos.

5.4 A pior função do mundo de Nesterov

Fixando $L > 0$, considere a seguinte família de funções quadráticas $f_q : \mathbb{R}^n \rightarrow \mathbb{R}$ definidas por Nesterov (veja, [11, Página 59]),

Também podemos calcular explicitamente o valor ótimo

$$\begin{aligned}
 f_q(x_q^*) &= \frac{L}{4} \left\{ \frac{1}{2} \left[x_1^{*2} + \sum_{i=1}^{q-1} (x_i^* - x_{i+1}^*)^2 + x_q^{*2} \right] - x_1^* \right\} \\
 &= \frac{L}{4} \left(\frac{1}{2} x^{*T} A_q x^* - e_1^T x^* \right) \\
 &= \frac{L}{4} \left(-\frac{1}{2} e_1^T x^* \right) \\
 &= -\frac{L}{8} e_1^T x^* \\
 &= -\frac{L}{8} \left(1 - \frac{1}{q+1} \right)
 \end{aligned}$$

Em [11, Teorema 2.1.7, p.61]), Nesterov mostra que para $1 \leq q \leq \frac{1}{2}(n-1)$, $x_0 = 0 \in \mathbb{R}^n$, para métodos de primeira ordem da forma

$$x_k \in x_0 + \text{span}\{\nabla f(x_0), \dots, \nabla f(x_{k-1})\},$$

a função $f(x) = f_{2q+1}(x)$ é tal que

$$\begin{aligned}
 f(x_k) - f(x^*) &\geq \frac{3}{32} L \|x_0 - x^*\|^2 \frac{1}{(k+1)^2}, \quad (25) \\
 \|x_k - x^*\|^2 &\geq \frac{1}{8} \|x_0 - x^*\|^2
 \end{aligned}$$

para $k \leq q$.

Para os experimentos numéricos da função (24) consideramos $L = 4$ e os critérios de parada foram os mesmos utilizados na Seção 5.1.

Para avaliar os métodos, consideramos dimensão $n = 2001$ e $q = 1000$ e o ponto inicial foi $x_0 = 0 \in \mathbb{R}^n$.

Métodos	t_k	k	$\ \nabla f(x_k)\ $
Cauchy	$1/L$	100000	1,58e-04
Nesterov	$1/L$	18110	9,99e-07
Grad. Espectral	BNM	54058	9,99e-10

Tabela 4 – Resultado obtido com a pior função do mundo

O gráfico da Figura 5 ilustra o desempenho dos métodos quando testamos na função (24).

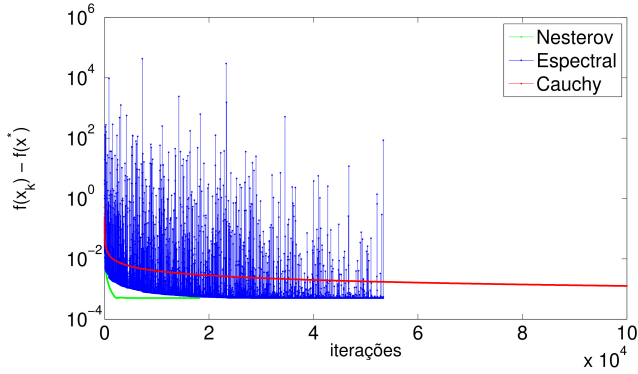


Figura 5 – Desempenho de pior caso

Assim, existem funções quadráticas convexas, como a pior função do mundo, para as quais métodos de descida com direção de busca sendo um múltiplo de $-\nabla f(x_k)$ não atingem a taxa de convergência ótima de $\mathcal{O}(1/k^2)$, como no método de Nesterov.

6 Considerações finais e trabalhos futuros

Neste trabalho foi apresentado um estudo sobre métodos de primeira ordem, que a cada iteração usam apenas informação de $\nabla f(x)$ em pontos do segmento x_{k-1}, x_k e uma análise de complexidade de iterações de tais métodos. Foram considerados os métodos de Cauchy, Nesterov e gradiente espectral, com diferentes esquemas de busca linear e seu desempenho prático investigado.

Para isso realizamos experimentos numéricos considerando problemas irrestritos em funções quadráticas com Hessiana simétrica positiva definida. De acordo com os resultados numéricos, na média, o método de gradiente espectral com busca não-monótona mostrou um desempenho melhor que os demais.

Por outro lado, nos testes feitos a “pior função do mundo”, tanto o método de Cauchy quanto o gradiente espectral tiveram seu comportamento ditado pela desigualdade (25) e o melhor desempenho do método de Nesterov neste caso ficou evidente.

Como trabalhos futuros, gostaríamos de estender o estudo numérico realizado para funções convexas mais gerais e considerar alguma aplicação em problemas reais, por exemplo, problemas de ajustes de curvas.

REFERÊNCIAS

- [1] J. Barzilai e J. M. Borwein. “Two-point step size gradient methods”. Em: *IMA Journal of Numerical Analysis* 8.1 (1988), pp. 141–148.
- [2] D. Bertsekas, A. Nedic e A. Ozdaglar. *Convex analysis and optimization*. Vol. 1. Athena Scientific, 2003.
- [3] H. Fawzi. “Topics in Convex Optimisation (L16)”. Em: *Mathematical Tripos Part III Guide to Courses* (2017), p. 71.
- [4] B. A. Ferraz. *Métodos computacionais de otimização*. Universidade Estadual Paulista (Unesp), 2017.
- [5] 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.
- [6] L. Grippo, F. Lampariello e S. Lucidi. “A nonmonotone line search technique for Newton’s method”. Em: *SIAM journal on Numerical Analysis* 23.4 (1986), pp. 707–716.
- [7] A. Izmailov e M. Solodov. *Otimização, volume 2: métodos computacionais*. IMPA, 2007.
- [8] E. L. Lima. *Análise real*. Vol. 1. Impa Rio de Janeiro, 2004.

-
- [9] J. M. Martínez e S.A. Santos. *Métodos computacionais de otimização*. XX Colóquio Brasileiro de Matemática, SBM, 1995.
- [10] M. W. Mendonça e D. S. Gonçalves. *MTM 3531 – Programação Linear*. 2020. Disp. em: <https://www.overleaf.com/read/kycmpxvpcznr>.
- [11] Y. Nesterov. *Introductory lectures on convex optimization: A basic course*. Vol. 87. Springer Science & Business Media, 2003.
- [12] Y. Nesterov. *Lectures on convex optimization*. Vol. 137. Springer, 2018.
- [13] J. Nocedal e S.J. Wright. *Numerical optimization*. Springer, 1999.
- [14] A. A. Ribeiro e E.W. Karas. *Otimização Contínua: aspectos teóricos e computacionais*. Cengage Learning Editora, 2014.
- [15] E. W Sachs e S. M Sachs. “Nonmonotone line searches for optimization algorithms”. Em: *Control and Cybernetics* 40.4 (2011), pp. 1059–1075.