Universidade Federal do Rio de Janeiro
Instituto de Matemática

Anderson de Oliveira Calixto

Métodos de Otimização Aplicados à Estatística

Rio de Janeiro
2020

Anderson de Oliveira Calixto

Métodos de Otimização Aplicados à Estatística

Dissertação de Mestrado submetida ao Programa de Pós-Graduação em Estatística do Instituto de Matemática da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do grau de Mestre em Estatística.

Área de Concentração: Estatística

Orientadores: Ralph dos Santos Silva
Heudson Tosta Mirandola

Rio de Janeiro

2020

Métodos de Otimização Aplicados à Estatística

Anderson de Oliveira Calixto

Orientadores: Ralph dos Santos Silva

Heudson Tosta Mirandola

Dissertação de Mestrado submetida ao Programa de Pós-Graduação em Estatística do Instituto de Matemática da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do grau de Mestre em Estatística.

Ralph dos Santos Silva
IM-UFRJ

Heudson Tosta Mirandola
IM-UFRJ

Marina Silva Paez
IM-UFRJ

Bernardo Freitas Paulo da Costa
IM-UFRJ

Guilherme Ost de Aguiar
IM-UFRJ

Rio de Janeiro

2020

CIP - Catalogação na Publicação

C154m

Calixto, Anderson de Oliveira
Métodos de Otimização Aplicados à Estatística / Anderson de Oliveira Calixto. -- Rio de Janeiro, 2020.
102 f.

Orientador: Ralph dos Santos Silva.
Coorientador: Heudson Tosta Mirandola.
Dissertação (mestrado) - Universidade Federal do Rio de Janeiro, Instituto de Matemática, Programa de Pós-Graduação em Estatística, 2020.

1. Algoritmos. 2. Estimação de Parâmetros. 3. Modelos Lineares Generalizados. I. Silva, Ralph dos Santos, orient. II. Mirandola, Heudson Tosta, coorient. III. Título.

Elaborado pelo Sistema de Geração Automática da UFRJ com os dados fornecidos pelo(a) autor(a), sob a responsabilidade de Miguel Romeu Amorim Neto - CRB-7/6283.

Dedico esse trabalho à minha mãe Margarida, que sempre esteve ao meu lado em todos os momentos de minha caminhada. Seu apoio e carinho constantes foram fundamentais para que eu caminhasse até aqui; à minha irmã Aline, que por intermédio do carinho de sua amizade, trouxe mais leveza aos desafios da vida acadêmica e à minha grande companheira Janaína, por sempre incentivar os meus sonhos ao longo de todos esses anos, sem o seu companherismo, cumplicidade e amor não teria chegado até aqui.

Agradecimentos

À Fundação CAPES, pelo apoio financeiro para a realização deste trabalho. Aos meus orientadores, professores Ralph dos Santos Silva e Heudson Tosta Mirandola, pelos ensinamentos, apoio e dedicação na construção de cada etapa desta dissertação e, principalmente, pela generosidade e amizade; contem sempre comigo. Aos professores do DME-IM - Departamento de Métodos Estatísticos do Instituto de Matemática, pela contribuição dada através das disciplinas cursadas ao longo da pós-graduação. À toda equipe do DME-IM que proporcionou as condições necessárias para a realização de cada fase do meu mestrado. Por fim, agradeço a todos aqueles que acreditaram e contribuíram direta ou indiretamente para a realização desta dissertação, minha eterna gratidão!

Lições tardias

“ Não devemos aprender a esperar.
Devemos, sim,
esquecer as coisas esperadas.
Ainda que nos digam:
“espere-me, à tal hora, em tal jardim”,
o jardim nos deve bastar.
Que a chegada daquilo
que nos fez esperar
seja algo normal naquele mundo,
como a morte de uma borboleta
ou a fuga de um lagarto nas pedras.
Se nada chega,
se ninguém aparece,
não notaremos a sua falta”

Alberto da Cunha Melo

Resumo

Nesta dissertação, considera-se a interseção entre as áreas dos métodos de otimização e da estimação de parâmetros em modelos estatísticos. Trabalha-se essa interseção em duas vias. Na primeira, realiza-se um estudo de caso na área dos métodos de otimização, por meio dos algoritmos do gradiente descendente, gradiente acelerado, gradiente acelerado de alta ordem e Newton-Raphson e são estudadas as taxas de convergência teóricas das sequências numéricas geradas por esses algoritmos. Para a segunda via, realiza-se um estudo de caso na área dos modelos estatísticos, fazendo uso dos modelos lineares generalizados e é analisado o processo de estimação de parâmetros nessa classe de modelos via logaritmo da função de verossimilhança. Por fim, é concretizada a interseção entre essas áreas, implementando os estimadores de máxima verossimilhança para o modelo da regressão Logística. Por intermédio da análise empírico-estatístico dos métodos de otimização em estudo, chega-se à conclusão de que o método do gradiente acelerado de alta ordem, quando faz o uso da informação de uma derivada da função objetivo, tem uma performance empírica competitiva em relação aos demais métodos de primeira ordem, quando considerado a taxa de convergência empírica e o tempo de execução, apesar de teoricamente ter uma taxa de convergência inferior ao método de Newton-Raphson, abrindo a possibilidade de trabalhos futuros para a investigação analítica desse fenômeno.

Palavras-chave: Algoritmos, Estimação de Parâmetros, Modelos Lineares Generalizados.

Abstract

In this thesis, we study an intersection between the areas of optimization methods and the statistical models. We work in this intersection in two ways. First, we take as a case study, in the field of optimization methods, the gradient descent, accelerated gradient descent, high-order accelerated gradient and Newton-Raphson and we study the theoretical convergence rates of the sequences generated by these methods. Second, we take as a case study, in the field of statistical models, the generalized linear models and studies their estimation process via logarithm of the likelihood function. Finally, to consolidate the intersection between these areas, we implemented the maximum likelihood estimators for the logistic regression model. Through the empirical statistical analysis of the optimization methods under study, we concluded that the high-order accelerated gradient method, when it uses information from one derivative of the objective function, has an empirical performance superior to the other methods of first order when we look to the empirical rates of convergence and the run times, despite theoretically having a convergence rate lower than the Newton-Raphson method. Thus opening the possibility of future work to an analytical investigation of this phenomenon.

Keywords: Algorithms, Parameter Estimation, Generalized Linear Models.

Lista de Figuras

3.1 Comportamento da sequência gerada pelo método de Newton-Raphson . . 27
3.2 Comportamento da sequência gerada pelo método do gradiente descendente 35
3.3 Comportamento da sequência gerada pelo método do gradiente acelerado. . 42

4.1 Comportamento da sequência gerada pelo método do gradiente acelerado de alta ordem 69

Lista de Tabelas

5.1 Comparação entre as ordens de convergência dos métodos do gradiente descendente, gradiente acelerado, gradiente acelerado de alta ordem para uma derivada e Newton-Raphson. 72

5.2 Dimensão das amostras para cada experimento. 74

5.3 Estatísticas descritivas sobre o número de iterações até a convergência dos algoritmos considerando o modelo de regressão Logística para 3 tamanhos de amostras, $n \in \{250, 1000, 5000\}$ e para 3 tamanhos do vetor paramétrico, $p \in \{25, 50, 70\}$. Os resultados são baseados em 100 réplicas de Monte Carlo para cada combinação. 79

5.4 Estatísticas descritivas sobre os tempos de execução até a convergência dos algoritmos considerando o modelo de regressão Logístico para 3 tamanhos de amostras, $n \in \{250, 1000, 5000\}$ e para 3 tamanhos do vetor paramétrico, $p \in \{25, 50, 70\}$. Os resultados são baseados em 100 réplicas de Monte Carlo para cada combinação. 83

Sumário

1. Introdução 14

1.1 Objetivo 15

2. Modelos Lineares Generalizados 17

2.1 Definição e propriedades 17

2.2 Funções de ligação canônicas 19

2.3 Informação de Fisher 21

2.4 Modelo de regressão logística 22

3. Métodos de Otimização 24

3.1 Newton-Raphson 26

3.1.1 Construção 26

3.1.2 Taxa de convergência 28

3.1.3 Escore de Fisher 29

3.2 Gradiente descendente 30

3.2.1 Taxa de convergência 35

3.2.2 EDO de primeira ordem associada 38

3.2.3 Taxa de convergência das soluções da EDO associada 39

3.3 Gradiente acelerado 41

3.3.1 Taxa de convergência 42

3.3.2 EDO de segunda ordem associada 47

3.3.3 Taxa de convergência das soluções da EDO associada 49

3.4 Gradiente acelerado de alta ordem 50

3.4.1 EDO de segunda ordem associada 51

3.4.2 Taxa de convergência das soluções da EDO associada 53
3.4.3 Taxa de convergência 55

4. Gradiente acelerado de alta ordem aplicado aos MLGs 62
4.1 Gradiente acelerado de alta ordem em forma explícita para MLGs 66

5. Comparação entre os métodos de otimização 70
5.1 Comparação matemática entre as taxas de convergência 72
5.2 Análise empírica de algoritmos 73
5.2.1 Modelo de regressão Logística 76
5.2.1.1 Comparação empírica entre as taxas de convergência . . . 76
5.2.1.2 Comparação empírica entre os tempos de execução 80

6. Conclusão 85

Referências 87

Capítulo 1

Introdução

Os métodos de otimização (Izmailov e Solodov, 2012; Baumaister e Leitão, 2014) são amplamente usados nos mais variados campos da ciência. Na Estatística, os métodos de otimização são principalmente empregados na estimação de parâmetros em modelos que não apresentam solução analítica explícita. Por exemplo, o método de Newton-Raphson é amplamente aplicado no contexto da estimação por máxima verossimilhança - em particular nos modelos lineares generalizados (McCullagh e Nelder, 1989).

Os métodos de otimização que se baseiam na derivada primeira (vetor gradiente) da função objetivo são denominados de métodos de primeira ordem. Já os métodos que utilizam as derivadas primeira e segunda (matriz hessiana) são chamados de métodos de segunda ordem, e assim por diante.

Um método de primeira ordem muito conhecido é o gradiente descendente. Esse método tem como uma de suas características a taxa de convergência da ordem de $\mathcal{O}(1/k)$, em que k representa o número de iterações (ver Gower, 2019, e referências). Ao longo das últimas décadas vários estudos têm sido realizados para melhorar as taxas de convergência dos métodos de primeira e segunda ordens. Por exemplo, Nesterov (1983) mostrou que é possível obter uma variação do método do gradiente descendente com taxa de convergência da ordem de $\mathcal{O}(1/k^2)$. Esse método ficou conhecido como *método do gradiente acelerado* ou *método de Nesterov*. Além disso, Nesterov (2004) mostrou que a taxa ótima para métodos que utilizam somente a derivada primeira é de $\mathcal{O}(1/k^2)$.

Em busca de uma melhor compreensão do método de Nesterov, Su et al. (2016) investigaram o gradiente acelerado via teoria das equações diferenciais ordinárias e, a partir disso, evidenciaram vários aspectos interessantes sobre o comportamento da sequência gerada por esse método. Outrossim, Wibisono et al. (2016) propuseram uma formulação

variacional, que toma como inspiração a técnica de aceleração que surge com o método de Nesterov, que possibilitou o desenvolvimento de um modelo geral de aceleração para uma determinada classe de algoritmos numéricos conhecidos como *método do gradiente de alta ordem*. Essa técnica de aceleração, chamada de *método do gradiente acelerado de alta ordem*, pode ser particularizada para qualquer ordem (primeira, segunda etc). A parte empírica desta dissertação foca principalmente no caso de primeira ordem.

Sob condições gerais, as taxas de convergência (teóricas) dos métodos de primeira ordem são, em geral, bem diferentes das taxas de convergência dos métodos de segunda ordem. Por exemplo, o gradiente descendente e o gradiente acelerado têm taxas de convergência menores (mais lentas) do que o método de Newton-Raphson. Porém, esse último utiliza mais informações via derivada segunda (inversa da matriz hessiana), o que tem um custo computacional elevado (pelo menos para problemas considerados "grandes"). Esta dissertação se concentra no estudo de alguns métodos de otimização de primeira ordem – gradiente descendente, gradiente acelerado e gradiente acelerado de alta ordem para uma derivada – e de segunda ordem – método de Newton-Raphson – do ponto de vista teórico e prático. Vale ressaltar que a parte teórica do método do gradiente acelerado de alta ordem é abordada de maneira geral, mas na parte empírica dos estudos de Monte Carlo se faz o uso somente do caso de primeira orden.

1.1 Objetivo

O objetivo desta dissertação de mestrado é estudar, do ponto de vista teórico e prático, as taxas de convergência dos seguintes métodos de otimização:

- gradiente descendente (ver Nesterov, 2004);
- gradiente acelerado (Nesterov, 1983); e
- gradiente acelerado de alta ordem (Wibisono et al., 2016).

O último método é abordado de forma geral na parte teórica. Contudo, as aplicações utilizam somente o caso particular de primeira ordem. Portanto, outra meta desta dissertação também é comparar esses métodos de primeira ordem (que não utilizam a derivada segunda) dos pontos de vistas teórico e empírico. Ademais, para efeitos de comparação, apresenta-se também do ponto de vista teórico e empírico as taxas de convergência do

clássico método de otimização de segunda ordem: método de Newton-Raphson (ver Izmailov e Solodov, 2012).

As provas teóricas das taxas de convergência são baseadas nas sequências que os algoritmos produzem. Além disso, os três primeiros algoritmos têm relações diretas com determinadas equações diferenciais ordinárias (EDO) e essas são utilizadas na interpretação do sentido do processo de aceleração. Por isso, as taxas de convergências das curvas solução dessas EDOs também são demonstradas.

Ainda, um outro objetivo desta dissertação é a comparação empírica desses algoritmos quando aplicados à estimação por máxima verossimilhança em modelos lineares generalizados, em particular no modelo de regressão logístico.

Esta dissertação de mestrado está organizada da seguinte forma. O capítulo 2 apresenta uma revisão dos modelos lineares generalizados, suas propriedades e um caso particular – modelo de regressão logístico – que serve de base para as comparações empíricas dos métodos de otimização. O capítulo 3 aborda os métodos de otimização começando com a dedução do método de Newton-Raphson (e Escore de Fisher) e sua respectiva taxa de convergência. Em seguida, estuda-se o clássico algoritmo do gradiente descendente e sua taxa de convergência, o método do gradiente acelerado e sua taxa de convergência, e finalmente o gradiente acelerado de alta ordem e sua taxa de convergência. Esses métodos são explorados do ponto de vista discreto (sequência gerada pelo algoritmo) e por meio das equações diferenciais ordinárias associadas a eles. No capítulo 4 se aplica o método do gradiente acelerado de alta ordem para o caso de uma derivada aos modelos lineares generalizados de modo a se obter para esses modelos a forma explícita desse algoritmo. No capítulo 5 se faz um estudo de Monte Carlo para a comparação dos diversos métodos de otimização apresentados e esses são aplicados à estimação por máxima verossimilhança no modelo de regressão logístico. O capítulo 6 resume os resultados obtidos e as possíveis limitações do estudo - principalmente referente aos resultados empíricos. Aborda-se também algumas direções para trabalhos futuros.

Capítulo 2

Modelos Lineares Generalizados

Os modelos lineares normais durante muito tempo constituíram a base para a modelagem estatística de diversos fenômenos nas mais diversas áreas do conhecimento. Entretanto, para vários desses fenômenos não é possível assumir que a variável resposta cujo comportamento os quantifica seja normalmente distribuída e, portanto, é necessário realizar alguma transformação nos dados observados a fim de se obter uma aproximação normal deles.

Infelizmente, isso leva a uma importante perda na capacidade de interpretação por parte do estatístico sobre o comportamento dos fenômenos em estudo. Algumas propostas de extensão dos modelos lineares normais foram apresentadas ao longo dos anos. Nelder e Wedderburn (1972) introduziram a classe dos modelos lineares generalizados (MLGs). Essa classe de modelos amplia as possibilidades de distribuição da variável resposta para todos os membros da família exponencial de distribuições. Essa classe de modelos também flexibiliza a relação funcional entre a média da variável resposta e o preditor linear.

Este capítulo apresenta a definição dos MLGs e algumas de suas propriedades. Ademais, aborda-se o modelo de regressão logístico.

2.1 Definição e propriedades

Como afirmado na introdução deste capítulo, os MLGs permitem que a distribuição da variável resposta possa pertencer à família exponencial de distribuições. Essa classe de distribuições de probabilidade contém as distribuições clássicas como, por exemplo, a normal, exponencial, Bernoulli, binomial, gama, dentre outras. A definição formal da família exponencial no caso unidimensional é dada a seguir.

Definição 1 (Família Exponencial)**.** *Seja Y uma variável aleatória. A distribuição de Y pertence à* família exponencial *de distribuições se a sua função de densidade (ou de probabilidade) é escrita na seguinte forma*

$$p(y|\theta,\phi) = \exp\{\phi[y\theta - b(\theta)] + c(y,\phi)\}. \tag{2.1}$$

Na equação (2.1), temos os seguintes elementos:

1 $\phi > 0$ é o parâmetro de dispersão.

2 θ é o parâmetro canônico que indexa a distribuição.

3 $b : \mathbb{R} \to \mathbb{R}$ é uma função estritamente convexa e diferenciável chamada de *função de partição.*

4 $c : \mathbb{R} \times \mathbb{R}_+ \to \mathbb{R}$ é uma função da amostra e da dispersão.

A função $b : \mathbb{R} \to \mathbb{R}$ desempenha um papel importante na classe dos MLGs e também em diversos outros modelos estatísticos que tomam como base a família exponencial de distribuições. A derivada da função b é um difeomorfismo, isto é, uma função diferenciável cuja a inversa também é diferenciável. Através do difeomorfismo b' podemos realizar mudanças de coordenadas entre o espaço dos parâmetros canônicos do modelo - o espaço dos $\theta \in \mathbb{R}$ - e o espaço dos parâmetros naturais - o espaço dos $\mu \in \mathbb{R}$.

A família exponencial tem diversas propriedades interessantes e algumas delas serão utilizadas ao longo desta dissertação. Para mais informações sobre os demais aspectos teóricos relacionados a família exponencial, ver a obra (Brown, 1986).

Definição 2 (Modelos Lineares Generalizados)**.** *Considere $(Y_i)_{0\leq i\leq n}$ uma sequência de variáveis aleatórias independentes, cada uma com função de probabilidade (ou densidade) pertencente à família exponencial de distribuições. Seja X uma matriz $n \times p$ com posto p e $\boldsymbol{x}_i^T = (x_{i1}, ..., x_{ip})$ a i-ésima linha da matriz X. Por fim, seja $\beta^T = (\beta_1, ..., \beta_p)$ um vetor de parâmetros do modelo e $G : \mathbb{R} \to \mathbb{R}$ um difeomorfismo chamado de* função de ligação*. O* modelo linear generalizado *é definido por:*

$$p(y_i|\theta_i,\phi) = \exp\{\phi\{y_i\theta_i - b(\theta_i)\} + c(y_i,\phi)\}, \tag{2.2a}$$

$$G(\mu_i) = \boldsymbol{x}_i^T\beta, \tag{2.2b}$$

$$b'(\theta_i) = \mu_i. \tag{2.2c}$$

O objetivo, ao longo das próximas seções, é estimar o vetor de parâmetros β. Por isso, só os aspectos relacionados com os precedimentos necessários para a estimação desse vetor serão apresentados nesta dissertação. Maiores informações e detalhes sobre os MLGs podem ser encontradas em McCullagh e Nelder (1989) e Dobson (2002). Além disso, o parâmetro de dispersão $\phi > 0$ é tomado como constante nesta e nas próximas seções. O vetor de parâmetros β, em conjunto com a matriz do modelo e a função de ligação, estruturam o vetor de médias $\mu^T = (\mu_1, ..., \mu_n)$ através da seguinte relação funcional

$$\mu_i = G^{-1}(\mathbf{x}_i^T\beta), \quad i = 1, 2, \ldots, n.$$

A função de ligação desempenha um papel de destaque na estimação de parâmetros na classe dos MLGs. Isso se deve ao fato de que, dependendo da classe de funções de ligação escolhida, o logaritmo da função de verossimilhança passa a ter propriedades como, por exemplo, ser estritamente côncavo. Essa propriedade garante que o estimador de máxima verossimilhança obtido seja único, quando esse existir (Paula, 2013). A classe de funções de ligação que assegura que o logaritmo da função de verossimilhança seja côncava é a classe das ligações canônicas. Na próxima subseção, abordaremos os aspectos mais relevantes dessa classe de funções para o processo de estimação do parâmetro β.

2.2 Funções de ligação canônicas

Seja $\phi > 0$ conhecido. O logaritmo da função de verossimilhança do MLG definido em (2.2) é dado por

$$\mathcal{L}(\beta|X, y) = \sum_{i=1}^{n} \phi\big\{y_i\theta_i - b(\theta_i)\big\} + \sum_{i=1}^{n} c(y_i, \phi). \tag{2.3}$$

Supondo que o parâmetro canônico é igual ao preditor linear, $\theta_i = \mathbf{x}_i^T\beta$, a equação (2.3) pode ser reescrita como

$$\mathcal{L}(\beta|X, y) = \sum_{i=1}^{n} \phi\big\{y_i\mathbf{x}_i^T\beta - b(\mathbf{x}_i^T\beta)\big\} + \sum_{i=1}^{n} c(y_i, \phi). \tag{2.4}$$

Definindo a estatística $S^T := \phi\sum_{i=1}^{n} y_i\mathbf{x}_i^T$, a equação (2.4) pode ser expressa por

$$\mathcal{L}(\beta|X, y) = S^T\beta - \phi\sum_{i=1}^{n} b(\mathbf{x}_i^T\beta) + \sum_{i=1}^{n} c(y_i, \phi). \tag{2.5}$$

Então, pelo teorema da fatoração de Neyman–Pearson, a estatística S^T é suficiente para o vetor paramétrico β. As funções de ligação associadas a tais estatísticas são chamadas

de *ligações canônicas*. Por um lado, como a função $b : \mathbb{R} \to \mathbb{R}$ é um difeomorfismo, existe $(b')^{-1} : \mathbb{R} \to \mathbb{R}$ tal que,

$$(b')^{-1}(\mu_i) = \theta_i \quad i = 1, 2, \ldots, n.$$

Por outro lado, das equações (2.2b) e (2.2c), segue-se que

$$G \circ b'(\theta_i) = \mathbf{x}_i^T \beta.$$

Com isso, concluímos que, se G é uma ligação canônica, então

$$G(\mu_i) = (b')^{-1}(\mu_i),$$

pois nesse caso

$$\theta_i = (b')^{-1} \circ b'(\theta_i) = G \circ b'(\theta_i) = \mathbf{x}_i^T \beta.$$

No lema a seguir, formaliza-se a afirmação de que: sob a hipótese da função de ligação ser canônica, o logaritmo da função de verossimilhança (2.3) é estritamente côncavo.

Lema 3. *Seja $\mathcal{L} : \mathbb{R}^p \to \mathbb{R}$ o logaritmo da função de verossimilhança para o modelo linear generalizado (2.3). Se a função de ligação $G : \mathbb{R} \to \mathbb{R}$ é uma ligação canônica, então o logaritmo da função de verossimilhança é estritamente côncavo.*

Demonstração. A função $\mathcal{L} : \mathbb{R}^p \to \mathbb{R}$, para o modelo (2.2) (sob a hipótese da função de ligação ser canônica e $\phi > 0$ ser conhecido) é dada em (2.5). Como $b : \mathbb{R} \to \mathbb{R}$ é uma função estritamente convexa e a aplicação $\beta \mapsto \mathbf{x}_i^T \beta$ é linear, temos que, a aplicação $\beta \mapsto b(\mathbf{x}_i^T \beta)$ é estritamente convexa. A soma de funções estritamente convexas é uma função estritamente convexa. Logo, a aplicação $\beta \mapsto \phi \sum_{i=1}^n b(\mathbf{x}_i^T \beta)$ é estritamente convexa.

Considere $\beta_a, \beta_b \in \mathbb{R}^p$ e $\alpha \in (0, 1)$. Inicialmente, temos que

$$\mathcal{L}(\alpha\beta_a + (1-\alpha)\beta_b | X, y) = S^T[\alpha\beta_a + (1-\alpha)\beta_b] - \phi \sum_{i=1}^n b(\mathbf{x}_i^T[\alpha\beta_a + (1-\alpha)\beta_b]) + \sum_{i=1}^n c(y_i, \phi).$$

Agora, dado que $\beta \mapsto S^T\beta$ é uma aplicação linear

$$S^T[\alpha\beta_a + (1-\alpha)\beta_b] = \alpha S^T\beta_a + (1-\alpha)S^T\beta_b.$$

Como a aplicação $\beta \mapsto -\phi \sum_{i=1}^n b(\mathbf{x}_i^T \beta)$ é estritamente côncava, segue que

$$-\phi \sum_{i=1}^n b(\mathbf{x}_i^T[\alpha\beta_a + (1-\alpha)\beta_b]) > \alpha \left[-\phi \sum_{i=1}^n b(\mathbf{x}_i^T\beta_a)\right] + (1-\alpha)\left[-\phi \sum_{i=1}^n b(\mathbf{x}_i^T\beta_b)\right],$$

e reescrevendo $\sum_{i=1}^{n} c(y_i, \phi)$ como uma combinação convexa,

$$\sum_{i=1}^{n} c(y_i, \phi) = \alpha \left[\sum_{i=1}^{n} c(y_i, \phi) \right] + (1 - \alpha) \left[\sum_{i=1}^{n} c(y_i, \phi) \right],$$

obtemos que

$$\begin{aligned} \mathcal{L}(\alpha\beta_a + (1-\alpha)\beta_b | X, y) > \alpha & \left[S^T \beta_a - \phi \sum_{i=1}^{n} b(\mathbf{x}_i^T \beta_a) + \sum_{i=1}^{n} c(y_i, \phi) \right] \\ & + (1-\alpha) \left[S^T \beta_b - \phi \sum_{i=1}^{n} b(\mathbf{x}_i^T \beta_b) + \sum_{i=1}^{n} c(y_i, \phi) \right] \\ & = \alpha \mathcal{L}(\beta_a | X, y) + (1-\alpha) \mathcal{L}(\beta_b | X, y). \end{aligned}$$

Portanto, $\mathcal{L}$ é uma função estritamente côncava. □

2.3 *Informação de Fisher*

Em Estatística é comum se pensar em termos da quantidade de informação contida na amostra com a qual se está trabalhando. Seja $\mathcal{D}$ o conjunto que representa uma amostra disponível para um modelo estatístico $p(\cdot|\theta)$, em que θ é o parâmetro que indexa o modelo. Assim, o logaritmo da função de verossimilhança do parâmetro θ em relação a amostra $\mathcal{D}$ é dada por

$$\mathcal{L}(\theta|\mathcal{D}) := \log(p(\mathcal{D}|\theta)). \tag{2.6}$$

Se a função (2.6) for duas vezes diferenciável, então define-se a *matriz de informação de Fisher* por

$$\mathcal{I}_{\mathcal{D}}(\theta) := \mathbb{E}[-\nabla^2 \mathcal{L}(\theta|\mathcal{D})]. \tag{2.7}$$

A matriz de informação de Fisher nos diz sobre o quanto de informação uma amostra $\mathcal{D}$ contém sobre o parâmetro desconhecido θ. No caso dos MLGs, o conjunto $\mathcal{D}$ se reduz a $y^T = (y_1, \ldots, y_n)$ – toda análise é, em geral, feita condicionalmente ao conhecimento da matriz X – e o logaritmo da função de verossimilhança é dado em (2.5) para as funções de ligação canônicas. Explicitamente, para os MLGs, obtemos

$$\mathcal{I}(\beta) = X^T \text{diag}(b''(\mathbf{x}_1^T\beta), ..., b''(\mathbf{x}_n^T\beta))X,$$

em que diag$(\cdot)$ é a matriz diagonal dada pelo vetor $(b''(\mathbf{x}_1^T\beta), ..., b''(\mathbf{x}_n^T\beta))$. A matriz de informação de Fisher tem papel fundamental no processo de estimação do parâmetro β, principalmente, na construção de métodos numéricos para essa finalidade.

2.4 Modelo de regressão logística

Dentro da classe dos MLGs, o modelo de regressão logístico possui uma posição de destaque e tem sido amplamente empregado ao longo dos últimos anos. Isso se deve ao fato, desse modelo ser a base para a construção de muitos classificadores em problemas de aprendizagem de máquina (*machine learning*). Assim, esses modelos servem para tratar problemas de classificação supervisionado. Dado a matriz modelo X, desejamos saber se um determinado objeto pertence a uma classe pré-estabelecida ou não. Isso significa que o vetor de variáveis respostas deve assumir valores binários que, em geral, toma-se como 0 ou 1. A distribuição de probabilidade utilizada para isto é a *Bernoulli* com parâmetro π que representa a probabilidade de ocorrência do fenômeno aleatório de interesse.

Como visto na seção §2.1, assume-se que o parâmetro π da distribuição Bernoulli tem relação com os dados da matriz modelo por meio da função de ligação G e do vetor de parâmetros β cujo valor deve ser estimado. Na definição abaixo, apresentamos a formulação do modelo de regressão logístico pressupondo que a função de ligação é canônica.

Definição 4 (Modelo de Regressão Logístico)**.** *Considere* $(Y_i)_{0\leq i\leq n}$ *uma sequência de variáveis aleatórias independentes cada uma com função de probabilidade dada por*

$$p(y_i|\theta_i) = \exp\{y_i\theta_i - \log(1+\exp\{\theta_i\})\}.$$

Sejam X *uma matriz* $n \times p$ *com posto* p, $\boldsymbol{x}_i^T = (x_{i1}, ..., x_{ip})$ *a* i*-ésima linha da matriz* X *e* $\beta^T = (\beta_1, ..., \beta_p)$ *o vetor de parâmetros do modelo. O modelo de regressão logístico é definido por*

$$\begin{aligned} p(y_i|\theta_i) &= \exp\{y_i\theta_i - \log(1+\exp\{\theta_i\})\}, \\ \theta_i &= \log\left(\frac{\pi_i}{1-\pi_i}\right) = \boldsymbol{x}_i^T\beta, \\ \pi_i &= \frac{\exp\{\theta_i\}}{1+\exp\{\theta_i\}}. \end{aligned}$$

Note-se que, existem, no modelo de regressão logístico, algumas características interessantes do ponto de vista da estimação do parâmetro β. A primeira delas é que, o parâmetro de dispersão do modelo ϕ é constante e igual a 1, isso garante que a suposição que foi feita no início desta seção ($\phi > 0$ conhecido) seja satisfeita automaticamente. Também, temos que, a função $c(y_i, \phi)$ é constante e igual a zero, o que simplifica a função de verossimilhança

cujo o logaritmo fica dado por

$$\mathcal{L}(\beta|X,y) = S^T\beta - \sum_{i=1}^{n} \log(1+\exp\{\mathbf{x}_i^T\beta\}), \quad \text{com} \quad S^T = \sum\nolimits_{i=1}^{n} y_i\mathbf{x}_i^T. \tag{2.8}$$

Além disso, como a função de partição $b : \mathbb{R} \to \mathbb{R}_+$ possui uma forma simples

$$b(\theta_i) = \log(1+\exp\{\theta_i\}),$$

segue-se que a matriz de informação de Fisher associada ao modelo pode ser facilmente calculada e assume a seguinte forma

$$\mathcal{I}(\beta) = X^T \text{diag}\left[\frac{\exp\{\mathbf{x}_1^T\beta\}}{(1+\exp\{\mathbf{x}_1^T\beta\})^2}, ..., \frac{\exp\{\mathbf{x}_n^T\beta\}}{(1+\exp\{\mathbf{x}_n^T\beta\})^2}\right] X.$$

Com base no modelo de regressão logístico e nas propriedades expostas acima, é realizado no capítulo 5 um estudo estatístico dos métodos de otimização discutidos no capítulo 3 que passamos a discutir a seguir.

Capítulo 3

Métodos de Otimização

Em diversas áreas, – por exemplo, – na Estatística ou nas Engenharias – existem problemas que podem ser resolvidos recorrendo a um processo de otimização. Considera-se que o fenômeno ou problema em questão seja definido ou modelado por uma função – por exemplo, $f : \mathbb{R}^p \to \mathbb{R}$ – cuja dinâmica do fenômeno ou solução do problema seja obtida quando minimizarmos (ou maximizarmos) a função f em um subconjunto ou no próprio $\mathbb{R}^p$. Em outras palavras, considere $\mathcal{A} \subseteq \mathbb{R}^p$ um conjunto cujos elementos representam os candidatos a solução de um determinado problema e que a solução é obtida através de um processo de minimização da função f sobre o conjunto $\mathcal{A}$. Matematicamente, pode-se modelar esse problema do seguinte modo:

$$\beta_* = \arg\min_{\beta \in \mathcal{A}} f(\beta), \tag{3.1}$$

em que o elemento $\beta_* \in \mathcal{A}$ é a solução do problema de minimização. O conjunto $\mathcal{A}$ é chamado de conjunto dos pontos viáveis para o problema, ou seja, é o conjunto em que se busca os elementos que tornam a equação (3.1) verdadeira. A função f é chamada de função objetivo.

Nos diversos casos encontrados na prática, não é possível obter a solução β_* de modo analítico. Portanto, é necessário a utilização de algum procedimento numérico para a obtenção de β_* - sempre de forma aproximada. A resolução do problema (3.1) via métodos numéricos é obtida através da construção de uma sequência $(\beta_k)_{k \geqslant 0}$ tal que

$$\lim_{k \to \infty} f(\beta_k) = f(\beta_*).$$

Nesse contexto, um aspecto teórico fundamental é a taxa de convergência da sequência $(f(\beta_k))_{k \geqslant 0}$ – uma medida de rapidez – que é representada por

$$f(\beta_k) - f(\beta_*) \leqslant \mathcal{O}(H(k)),$$

em que H é uma função da iteração k.

Nas próximas seções, são apresentados procedimentos numéricos que podem ser utilizados na resolução de alguns casos particulares da equação (3.1). A seção §3.1 traz o conhecido método de Newton-Raphson que é frequentemente encontrado em pacotes computacionais para a estimação de parâmetros em modelos estatísticos. Nas seções §3.2, §3.3 e §3.4 são abordados os métodos do gradiente descendente, do gradiente acelerado e do gradiente acelerado de alta ordem, respectivamente.

Na seção §3.2 dissertamos sobre o método do gradiente descendente e apresentamos uma EDO de primeira ordem que pode ser associada a esse método de modo natural. Demonstramos então, que as taxas de convergência do método do gradiente descendente é $\mathcal{O}(1/\epsilon k)$, em que $\epsilon > 0$, e que da EDO de primeira ordem associada é $\mathcal{O}(1/t)$. Observamos que as taxas de convergência podem ser associadas por intermédio da identificação $t = \epsilon k$. Isso permite a interpretação do método do gradiente descendente como um método de discretização para essa EDO que preserva a taxa de convergência da curva $t \mapsto f(\beta(t))$ e de sua discretização $k \mapsto f(\beta_k)$, em que $t \mapsto \beta(t)$ é a curva solução da EDO de primeira ordem associada ao método do gradiente descendente.

Em seguida, na seção §3.3 é introduzido uma sequência auxiliar no método do gradiente descendente, a qual permite que a taxa de convergência passe de $\mathcal{O}(1/\epsilon k)$ para $\mathcal{O}(1/\epsilon k^2)$ sem a necessidade de se utilizar mais informações sobre a função objetivo f, como por exemplo, a segunda derivada. Esse novo método é conhecido como método do gradiente acelerado ou método de Nesterov. Ademais, mostra-se que é possível associar ao método de Nesterov uma EDO de segunda ordem. Depois, ao se calcular a taxa de convergência da curva $t \mapsto f(\beta(t))$, em que $t \mapsto \beta(t)$ é agora a curva solução da EDO de segunda ordem associada ao método de Nesterov, obtém-se a estimativa $\mathcal{O}(1/t^2)$. Ao se identificar $t = \sqrt{\epsilon}k$, nota-se, mais uma vez, que o método em questão pode ser interpretado como um método de discretização que garante a compatibilidade entre as taxas de convergência a tempo contínuo e discreto.

Por fim, na seção §3.4 se aborda o método do gradiente acelerado de alta ordem que pode ser entendido, de modo simplificado, como uma generalização do processo de aceleração desenvolvido por Nesterov. Nesta parte a interpretação desse processo, como uma técnica de discretização, para uma determinada classe de EDOs, a qual preserva a taxa de convergência é consolidada. Isto é possível através do cálculo variacional em que uma

classe de EDOs de segunda ordem é deduzida e um método geral de discretização é obtido de modo que as taxas de convergência dessa classe de EDOs e do método geral de discretização são sempre compatíveis.

Na tentativa de tornar a exposição dos métodos mais didática e facilitar as comparações das sequências geradas pelos métodos, este capítulo considera apenas a seguinte função objetivo

$$f(\beta_1, \beta_2) := 2.10^{-2}\beta_1^2 + 5.10^{-3}\beta_2^2, \tag{3.2}$$

com domínio no conjunto $[-2\,;\,2]\times[-2\,;\,2]$. A escolha desta função ocorre pela simplicidade das suas curvas de nível, por ser uma função convexa e duas vezes diferenciável e por possibilitar a visualização do comportamento das trajetórias produzidas pelas sequências geradas por cada método. Já a escolha dos coeficientes tem como única função possibilitar uma melhor visualização gráfica.

3.1 Newton-Raphson

Nos mais variados campos das engenharias e da Estatística inúmeros fenômenos só podem ser corretamente modelados por equações não lineares. A compreensão do fenômeno em estudo se dá, em certos casos, pela obtenção de um elemento que seja uma raiz da equação que modela o problema. Em diversos casos, esse elemento não pode ser obtido explicitamente e recorre-se a métodos numéricos para a obtenção de um valor aproximado. Se a equação não linear que modela o problema pudesse ser transformada em uma função diferenciável que atendesse a um certo conjunto de critérios, então o método de Newton-Raphson poderia ser utilizado para a obtenção da solução.

3.1.1 Construção

Considere que um determinado problema ou fenômeno pode ser representado por uma função $g : \mathbb{R}^p \to \mathbb{R}^p$ diferenciável. Seja $\beta_* \in \mathbb{R}^p$ um elemento que representa a solução do problema

$$g(\beta_*) = 0. \tag{3.3}$$

Considere também que, não seja possível obter a solução explícita em (3.3). Supondo que $\beta_0 \in \mathbb{R}^p$ seja uma primeira aproximação para o valor de β_*, segue-se, pela expansão de

Taylor de primeira ordem, que

$$0 = g(\beta_*) = g(\beta_0 + h) = g(\beta_0) + \langle \nabla g(\beta_0), h \rangle + o(||h||),$$

em que $h := \beta_* - \beta_0$. Agora, assuma que a inversa da matriz jacobiana ∇g exista no ponto β_0. Logo, podemos aproximar o valor de h por

$$h \approx -[\nabla g(\beta_0)]^{-1} g(\beta_0).$$

A partir daí, define-se uma nova estimativa para o valor de β_* como

$$\beta_* \approx \beta_1 := \beta_0 - [\nabla g(\beta_0)]^{-1} g(\beta_0),$$

em que β_1 é uma segunda aproximação para β_*. Ao se repetir esse procedimento diversas vezes, constrói-se o seguinte processo iterativo:

$$\beta_{k+1} = \beta_k - [\nabla g(\beta_k)]^{-1} g(\beta_k). \tag{3.4}$$

O processo iterativo (3.4) é conhecido como método de Newton-Raphson.

A figura 3.1 exemplifica o caminho da sequência gerada pelo método de Newton-Raphson. Note que a sequência segue uma linha reta em direção ao ponto de mínimo. Uma possível explicação que justifica esse comportamento é que, ao fazer o uso implícito de informações sobre a curvatura da superfície, via o inverso da matriz hessiana da função objetivo f, o método de Newton-Raphson constrói uma trajetória, em certo sentido, ótima até o ponto de mínimo.

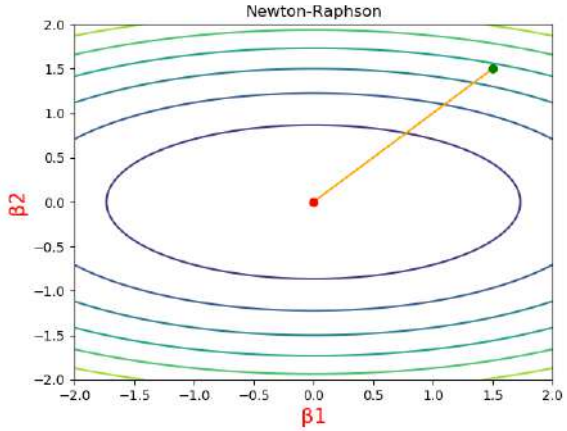

Figura 3.1: Comportamento da sequência gerada pelo método de Newton-Raphson para a função objetivo (3.2). O ponto verde é o valor em que é inicializado o método, $\beta_0 = (1.5\,;\,1.5)$ e o ponto vermelho é o valor estimado pelo método, $\beta_* = (0\,;\,0)$ para o ponto de mínimo.

Esse método funciona bem em muitas situações, mas pode apresentar desempenho ruim em alguns casos. Por exemplo, no contexto de otimização, se a função objetivo não for aproximadamente quadrática ou se a estimativa corrente se encontra distante do ponto ótimo, pode haver problemas de convergência da sequência para o ponto de mínimo (C.Frery e Cribari-Neto, 2011).

3.1.2 Taxa de convergência

O método de Newton-Raphson utiliza a informação da derivada da função g. Contudo, algumas restrições devem ser aplicadas como, por exemplo, a função g não pode se afastar muito de ser linear, pois o método de Newton-Raphson usa aproximações lineares para determinar a raiz da equação definida pela função g. Um modo de medir a não linearidade de uma função é por intermédio da constante de Lipschitz. Como a derivada de uma função mede a inclinação da "reta tangente"(plano tangente) ao gráfico da função em um determinado ponto, é natural exigir, para todo $\beta_a, \beta_b \in B(\beta_*, r)$ com $r > 0$ (ou seja, numa vizinhança de β_*), que

$$||\nabla g(\beta_a) - \nabla g(\beta_b)|| \leqslant L||\beta_a - \beta_b||, \tag{3.5}$$

em que $L > 0$ é a constante de Lipschitz. Temos que, quanto menor for a constante L, mais próximo estará a inclinação das "retas tangentes"(planos tangentes) ao gráfico da função g em dois pontos quaisquer, o que pode ser interpretado como o quanto a função g se aproxima de ser linear.

Uma segunda condição é que a derivada da função g seja invertível no ponto β_*. Além disso, pedimos que exista uma constante $\rho > 0$ tal que

$$||[\nabla g(\beta_*)]^{-1}|| \leqslant \frac{1}{\rho}. \tag{3.6}$$

Por fim, o ponto inicial β_0 não pode estar muito longe da solução β_*, isto é, existe $0 < \eta < r < 1$ tal que

$$||\beta_0 - \beta_*|| < \eta. \tag{3.7}$$

Teorema 5 (Bierlaire (2018) teorema 7.13)**.** *Considere $g \in C^1(\mathbb{R}^p; \mathbb{R}^p)$ e assuma que as condições (3.5), (3.6) e (3.7) são satisfeitas. Então, a sequência $(\beta_k)_{k\geqslant 0}$ gerada pela método de Newton-Raphson (equação (3.4)) satisfaz a seguinte desigualdade*

$$||\beta_{k+1} - \beta_*|| \leqslant \frac{L}{\rho}||\beta_k - \beta_*||^2. \tag{3.8}$$

Note que, recursivamente, a desigualdade (3.8) implica a seguinte desigualdade

$$||\beta_k - \beta_*|| \leqslant \left(\frac{L}{\rho}\right)^{2^k-1} ||\beta_0 - \beta_*||^{2^k} \quad \forall k \geqslant 0. \tag{3.9}$$

Como $g \in C^1(\mathbb{R}^p; \mathbb{R}^p)$ segue-se que no compacto $\overline{B(\beta_*, \eta)}$ essa função é Lipschitz contínua. Assim, existe uma constante $\tilde{L} > 0$ tal que

$$||g(\beta_a) - g(\beta_b)|| \leqslant \tilde{L}||\beta_a - \beta_b|| \quad \forall \beta_a, \beta_b \in \overline{B(\beta_*, \eta)}. \tag{3.10}$$

Combinando as desigualdades (3.9) e (3.10) obtemos

$$||g(\beta_k) - g(\beta_*)|| \leqslant \frac{\rho\tilde{L}}{L}\left(\frac{L\eta}{\rho}\right)^{2^k} \quad \beta_k \in \overline{B(\beta_*, \eta)}.$$

O que implica

$$||g(\beta_{k+1}) - g(\beta_*)|| \leqslant \mathcal{O}\left(\left(\frac{L\eta}{\rho}\right)^{2^k}\right) \quad \beta_k \in \overline{B(\beta_*, \eta)}. \tag{3.11}$$

3.1.3 Escore de Fisher

Em modelagem Estatística, quando há a necessidade de recorrer a um método numérico para estimar os parâmetros de um determinado modelo, é comum empregar uma variação do método de Newton-Raphson. Isso se deve ao fato do procedimento de estimação ser via o logaritmo da função de verossimilhança. Nos MLGs, dado o logaritmo da função de verossimilhança (2.5), a função g é definida por

$$g(\beta) := -\nabla\mathcal{L}(\beta|X, y).$$

Se β_* for a estimativa de máxima verossimilhança do parâmetro β, então necessariamente

$$g(\beta_*) = 0.$$

A derivada da função g é o negativo da matriz hessiana do logaritmo da função de verossimilhança, isto é,

$$\nabla g(\beta) = -\nabla^2\mathcal{L}(\beta|X, y).$$

Considere, para todo $\beta \in \mathbb{R}^p$, que $\nabla g(\beta)$ seja inversível. Logo, o método de Newton-Raphson para os MLGs é dado por

$$\beta_{k+1} = \beta_k - \left[\nabla^2\mathcal{L}(\beta_k|X, y)\right]^{-1}\nabla\mathcal{L}(\beta_k|X, y). \tag{3.12}$$

Como discutido na seção §2.3, em Estatística é comum utilizar toda a informação disponível na amostra, de modo a obter a melhor estimativa possível para os parâmetros do modelo em estudo. Além disso, naquela seção, a matriz de informação de Fisher foi apresentada como um conceito que auxilia a mensurar a quantidade de informação disponível em uma determinada amostra. Assim, a combinação da matriz de informação de Fisher (2.7) com o método de Newton-Raphson (3.4) recebe o nome de *método de Escore de Fisher*. No caso dos MLGs, a matriz de informação de Fisher coincide com a derivada segunda do logaritmo da função de verossimilhança, isto é

$$\mathcal{I}(\beta) = X^T \text{diag}(b''(\mathbf{x}_1^T\beta), ..., b''(\mathbf{x}_n^T\beta))X = \nabla^2\mathcal{L}(\beta|X, y)$$

e, portanto, podemos reescrever (3.12) do seguinte modo:

$$\beta_{k+1} = \beta_k - [\mathcal{I}(\beta_k)]^{-1}\nabla\mathcal{L}(\beta_k|X, y).$$

O método do Escore de Fisher desempenha um papel importante não só na classe dos MLGs, mas também em muitas outras classes de modelos e teorias estatísticas, como, por exemplo, a teoria assintótica de estimadores de máxima verossimilhança.

3.2 Gradiente descendente

O método do gradiente está inserido na classe dos métodos de descida. Os métodos de descida têm como base a ideia natural de construir uma sequência de pontos tal que a função objetivo seja ao longo desses pontos decrescente. Em outras palavras, dada uma aproximação β_k do ponto β_* que minimiza o problema, busca-se encontrar uma nova aproximação β_{k+1} tal que

$$f(\beta_{k+1}) < f(\beta_k).$$

Esse procedimento pode ser realizado de várias maneiras. Uma delas é tomar uma direção $d_k \in \mathbb{R}^p$ a partir do ponto β_k na qual a função objetivo f decresce, pelo menos para passos curtos, e calcular um comprimento de passo $\alpha_k > 0$ tal que

$$f(\beta_k + \alpha_k d_k) < f(\beta_k).$$

Desse modo, pode-se definir uma sequência $(\beta_k)_{k\geqslant 0}$ dada por

$$\beta_{k+1} = \beta_k + \alpha_k d_k, \quad \forall k \in \mathbb{N}.$$

Definição 6 (Direção de Descida)**.** *O vetor $d \in \mathbb{R}^p$ é uma* direção de descida *da função $f : \mathbb{R}^p \to \mathbb{R}$ no ponto $\beta \in \mathbb{R}^p$, se existe $\eta > 0$ tal que*

$$f(\beta + td) < f(\beta), \quad \forall t \in (0, \eta]. \tag{3.13}$$

Denotamos por $\mathcal{D}_f(\beta)$ o conjunto de todas as direções de descida da função f no ponto $\beta \in \mathbb{R}^p$. Como exemplo, podemos mostrar que o vetor $d := -\nabla f(\beta)$ é uma direção de descida. De fato, se $f : \mathbb{R}^p \to \mathbb{R}$ é uma função diferenciável tal que $\nabla f(\beta) \neq 0$ para todo $\beta \in \mathbb{R}^p$, segue-se, pela expansão de Taylor de primeira ordem, que

$$f(\beta + td) = f(\beta) + t\langle \nabla f(\beta), d\rangle + o(t)$$

com

$$\lim_{t \longmapsto 0} \frac{o(t)}{t} = 0.$$

A expressão acima diz que, para todo $\epsilon > 0$, exite $\eta(\epsilon) > 0$ tal que

$$|t| < \eta(\epsilon) \Rightarrow \left| \frac{o(t)}{t} \right| < \epsilon,$$

Tomando $\epsilon = ||\nabla f(\beta)||^2$, existe $\eta(\epsilon) > 0$ tal que,

$$|t| < \eta(\epsilon) \Rightarrow \frac{o(t)}{t} < ||\nabla f(\beta)||^2.$$

Assim, obtemos que

$$f(\beta + td) = f(\beta) - t||\nabla f(\beta)||^2 + o(t) < f(\beta) - t||\nabla f(\beta)||^2 + t||\nabla f(\beta)||^2 = f(\beta)$$

Logo, $d = -\nabla f(\beta) \in \mathcal{D}_f(\beta)$.

O próximo lema oferece um modo prático de caracterizar as direções de descida. A sua demonstração pode ser encontrada em Izmailov e Solodov (2012).

Lema 7. *Seja $f : \mathbb{R}^p \to \mathbb{R}$ uma função diferenciável no ponto $\beta \in \mathbb{R}^p$. As seguintes afirmações valem:*

1 Para todo $d \in \mathcal{D}_f(\beta)$, temos $\langle \nabla f(\beta), d\rangle \leqslant 0$;

2 Se $d \in \mathbb{R}^p$ satisfaz $\langle \nabla f(\beta), d\rangle < 0$, então $d \in \mathcal{D}_f(\beta)$.

A estratégia para calcular o comprimento de passo em um método de descida é um procedimento fundamental. Dentre as várias possibilidades de estratégias a serem adotadas,

como, por exemplo, a *regra de Wolfe* (Wolfe, 1969) e a *busca linear*, será apresentado nesse trabalho uma das estratégias mais simples e amplamente adotada em aplicações práticas, conhecida como *regra de Armijo* (Armijo, 1966). Para isso, faz-se necessário alguns resultados preliminares.

Teorema 8. *(Fórmula de Newton-Leibnitz - ver Izmailov e Solodov (2012)) Seja* $f \in C^1(\mathbb{R}^p)$. *Então, para todo* $\beta_a, \beta_b \in \mathbb{R}^p$, *vale que*

$$f(\beta_a + \beta_b) = f(\beta_a) + \int_0^1 \langle \nabla f(\beta_a + t\beta_b), \beta_b \rangle dt.$$

Lema 9. *Seja* $f \in C^1(\mathbb{R}^p)$ *e com derivada Lipschitz contínua de modulo* $L > 0$. *Então,*

$$|f(\beta + d) - f(\beta) - \langle \nabla f(\beta), d \rangle| \leqslant \frac{L}{2}||d||^2,$$

para todo $\beta, d \in \mathbb{R}^p$.

Demonstração. Pela fórmula de Newton-Leibnitz, obtemos

$$\begin{aligned}
|f(\beta + d) - f(\beta) - \langle \nabla f(\beta), d \rangle| &= \left| \int_0^1 \langle \nabla f(\beta + td), d \rangle dt - \langle \nabla f(\beta), d \rangle \right| \\
&= \left| \int_0^1 \langle \nabla f(\beta + td), d \rangle dt - \int_0^1 \langle \nabla f(\beta), d \rangle dt \right| \\
&= \left| \int_0^1 \langle \nabla f(\beta + td) - \nabla f(\beta), d \rangle dt \right| \\
&\leqslant \int_0^1 \left| \langle \nabla f(\beta + td) - \nabla f(\beta), d \rangle \right| dt
\end{aligned}$$

Por outro lado, pela desigualdade de Cauchy-Schwarz, temos que

$$\left| \langle \nabla f(\beta + td) - \nabla f(\beta), d \rangle \right| \leqslant ||\nabla f(\beta + td) - \nabla f(\beta)||.||d||.$$

Isso implica que

$$|f(\beta + d) - f(\beta) - \langle \nabla f(\beta), d \rangle| \leqslant \int_0^1 ||\nabla f(\beta + td) - \nabla f(\beta)||.||d|| dt.$$

Como o ∇f é Lipschitz contínuo em $\mathbb{R}^p$ com módulo $L > 0$, obtemos

$$||\nabla f(\beta + td) - \nabla f(\beta)|| \leqslant L||td||.$$

Isso implica que

$$|f(\beta + d) - f(\beta) - \langle \nabla f(\beta), d \rangle| \leqslant \int_0^1 L||td||.||d|| dt = L||d^2|| \int_0^1 tdt = \frac{L}{2}||d||^2.$$

□

Passamos então a apresentação da regra de Armijo. Considere que a função objetivo f seja diferenciável no ponto β_k. Fixados os parâmetros $\hat{\alpha} > 0$, $\sigma, \theta \in (0,1)$, toma-se $\alpha := \hat{\alpha}$ e realizamos os seguintes passos:

1 Verifica-se a desigualdade

$$f(\beta_k + \alpha d_k) \leqslant f(\beta_k) + \sigma\alpha\langle\nabla f(\beta_k), d_k\rangle. \tag{3.14}$$

2 Se a desigualdade (3.14) não é satisfeita, tomasse $\alpha = \theta\alpha$ e retornasse ao passo anterior. Caso contrário, aceitasse $\alpha_k = \alpha$ como valor do comprimento de passo.

Em outras palavras, o termo α_k é o maior número entre os números da forma $(\hat{\alpha}\theta^i)_{i\geqslant 0}$ que satisfaz a desigualdade (3.14). $\alpha\langle\nabla f(\beta_k), d_k\rangle$ representa o valor da redução na função objetivo f previsto pela sua aproximação linear para o passo α na direção de d_k. Se faz necessário demonstrar alguns resultados sobre a regra de Armijo, como, por exemplo, que ela está bem definida.

Lema 10. *Seja $f : \mathbb{R}^p \to \mathbb{R}$ uma função diferenciável no ponto $\beta_k \in \mathbb{R}^p$. Assuma que $d_k \in \mathbb{R}^p$ satisfaz a condição,*

$$\langle\nabla f(\beta_k), d_k\rangle < 0. \tag{3.15}$$

Então, a desigualdade (3.14) é satisfeita para todo $\alpha > 0$ suficientemente pequeno. Em outras palavras, a regra de Armijo está bem definida e termina com um valor $\alpha_k > 0$.

Demonstração. Pela expansão de Taylor de primeira ordem, obtemos

$$\begin{aligned} f(\beta_k + \alpha d_k) &= f(\beta_k) + \langle\nabla f(\beta_k), \alpha d_k\rangle + o(\alpha) \\ &= f(\beta_k) + \sigma\alpha\langle\nabla f(\beta_k), d_k\rangle + (1-\sigma)\alpha\langle\nabla f(\beta_k), d_k\rangle + o(\alpha) \\ &= f(\beta_k) + \sigma\alpha\langle\nabla f(\beta_k), d_k\rangle + \alpha\left((1-\sigma)\langle\nabla f(\beta_k), d_k\rangle + \frac{o(\alpha)}{\alpha}\right). \end{aligned}$$

Tomando $\epsilon := -\frac{(1-\sigma)}{2}\langle\nabla f(\beta_k), d_k\rangle$, existe $\delta(\epsilon) > 0$ tal que

$$|\alpha| < \delta(\epsilon) \Rightarrow \left|\frac{o(\alpha)}{\alpha}\right| < \epsilon.$$

Logo, para $|\alpha| < \delta(\epsilon)$, segue que

$$(1-\sigma)\langle\nabla f(\beta_k), d_k\rangle + \frac{o(\alpha)}{\alpha} \leqslant \frac{(1-\sigma)}{2}\langle\nabla f(\beta_k), d_k\rangle < 0.$$

Isso implica que

$$f(\beta_k + \alpha d_k) \leqslant f(\beta_k) + \sigma\alpha\langle\nabla f(\beta_k), d_k\rangle.$$

□

O lema a seguir será fundamental na demonstração da taxa de convergência do método do gradiente descendente, pois possibilitará que um mesmo tamanho de passo seja usado para todas as iterações.

Lema 11. *Seja $f \in C^1(\mathbb{R}^p)$ e com derivada Lipschitz contínua de módulo $L > 0$. Se $\beta_k, d_k \in \mathbb{R}^p$ satisfazem a condição (3.15). Então, a desigualdade (3.14) é válida para todo $\alpha \in (0, \bar{\alpha}_k]$, em que*

$$\bar{\alpha}_k := -\frac{2(1-\sigma)\langle\nabla f(\beta_k), d_k\rangle}{L||d_k||^2}. \tag{3.16}$$

Demonstração. Para todo $\alpha \in (0, \bar{\alpha}_k]$, segue de (3.16) que

$$\alpha < \bar{\alpha}_k \Rightarrow \frac{L\alpha}{2}||d_k||^2 < (\sigma - 1)\langle\nabla f(\beta_k), d_k\rangle.$$

Pelo lema 9, obtemos

$$\begin{aligned} f(\beta_k + \alpha d_k) - f(\beta_k) &\leqslant \langle\nabla f(\beta_k), \alpha d_k\rangle + \frac{L\alpha^2}{2}||d_k||^2 \\ &\leqslant \alpha\left(\langle\nabla f(\beta_k), d_k\rangle + \frac{L\alpha}{2}||d_k||^2\right). \end{aligned}$$

Isso implica que,

$$f(\beta_k + \alpha d_k) - f(\beta_k) \leqslant \alpha\left(\langle\nabla f(\beta_k), d_k\rangle + (\sigma - 1)\langle\nabla f(\beta_k), d_k\rangle\right) \leqslant \sigma\alpha\langle\nabla f(\beta_k), d_k\rangle.$$

Ou seja,

$$f(\beta_k + \alpha d_k) \leqslant f(\beta_k) + \sigma\alpha\langle\nabla f(\beta_k), d_k\rangle.$$

□

Uma consequência do lema 11 é que, ao tomar $d_k = -\nabla f(\beta_k)$, obtemos

$$\bar{\alpha}_k = -\frac{2(1-\sigma)\nabla f(\beta_k)^T d_k}{L||d_k||^2} = \frac{2(1-\sigma)}{L}.$$

Portanto, $\bar{\alpha}_k$ não depende de k e pelo lema 11, a desigualdade (3.14) vale para todo $\alpha \in (0, \bar{\alpha}]$, em que

$$\bar{\alpha} := \frac{2(1-\sigma)}{L} > 0.$$

Depois desses desenvolvimentos, podemos apresentar o método do gradiente descendente. Formalmente, considere $f \in C^1(\mathbb{R}^p)$ com derivada Lipschitz contínua de modulo $L > 0$. Seja $(\beta_k)_{k\geqslant 0}$ uma sequência definida por

$$\beta_{k+1} = \beta_k - \epsilon \nabla f(\beta_k), \tag{3.17}$$

em que $0 < \epsilon < 1/L$ é o comprimento de passo. Note que, pelo lema 11, tomando $\sigma = 1/2$, ocorre que $\bar{\alpha} = 1/L$ e também que, para todo $\epsilon \in (0, 1/L]$, a regra de Armijo sempre vale.

A figura 3.2 exibe o comportamento da sequência gerada pelo método do gradiente descendente. A trajetória gerada é ortogonal às curvas de nível da função objetivo (3.2), o que faz surgir o característico movimento em zigue-zague. Em conformidade com a justificativa dada para o comportamento da sequência do método de Newton-Raphson, em que o uso implícito de informação sobre a curvatura da função objetivo (via inversa da matriz hessiana) garantia que a trajetória fosse, em certo sentido, "ótima", temos que, no caso do método do gradiente descendente isso não é mais possível, pois essa informação está ausente nesse método.

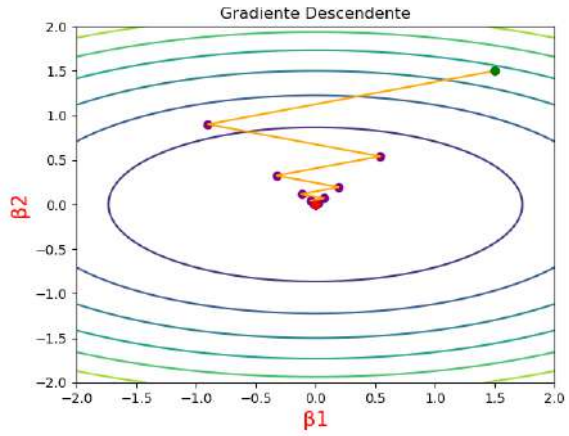

Figura 3.2: Comportamento da sequência gerada pelo método do gradiente descendente para a função objetivo (3.2). O ponto verde é o valor em que é inicializado o método, $\beta_0 = (1.5\,,\, 1.5)$ e o ponto vermelho é o valor estimado pelo método, $\beta_* = (0.0\,,\, 0.0)$ para o ponto de mínimo.

3.2.1 Taxa de convergência

O método do gradiente descendente possui diversas características interessantes tanto ao nível teórico quanto de aplicação. A característica que se mostra mais relevante para o estudo desenvolvido nessa subseção é a taxa de convergência do método.

Lema 12 (Taxa de convergência do método do gradiente descendente)**.** *Seja $f \in C^1(\mathbb{R}^p)$ uma função convexa com derivada Lipschitz contínua de modulo $L > 0$. A taxa de convergência do método do gradiente descendente é da ordem de*

$$f(\beta_k) - f(\beta_*) \leqslant \mathcal{O}\left(\frac{1}{\epsilon k}\right),$$

em que β_* é o ponto de mínimo para f.

Demonstração. Pela expansão de Taylor de primeira ordem com resto de Lagrange, obtemos

$$\begin{aligned} f(\beta_k) &= f(\beta_{k-1} - \epsilon \nabla f(\beta_{k-1})) \\ &\leqslant f(\beta_{k-1}) - \langle \nabla f(\beta_{k-1}), \epsilon \nabla f(\beta_{k-1}) \rangle + \frac{\epsilon^2 L}{2} ||\nabla f(\beta_{k-1})||^2 \\ &\leqslant f(\beta_{k-1}) - \epsilon ||\nabla f(\beta_{k-1})||^2 + \frac{\epsilon^2 L}{2} ||\nabla f(\beta_{k-1})||^2 \\ &\leqslant f(\beta_{k-1}) + \left(\frac{\epsilon^2 L}{2} - \epsilon\right) ||\nabla f(\beta_{k-1})||^2. \end{aligned}$$

Ou seja

$$f(\beta_k) \leqslant f(\beta_{k-1}) + \left(\frac{\epsilon^2 L}{2} - \epsilon\right) ||\nabla f(\beta_{k-1})||^2. \tag{3.18}$$

Por outro lado, seja $(\mathcal{E}_k)_{k \geqslant 1}$ uma sequência definida por

$$\mathcal{E}_k := \epsilon k (f(\beta_k) - f(\beta_*)) + \frac{1}{2} ||\beta_k - \beta_*||^2. \tag{3.19}$$

Pela equações (3.17) e (3.18), segue que

$$\begin{aligned} \mathcal{E}_k &\leqslant \epsilon k (f(\beta_{k-1}) - f(\beta_*)) + \frac{1}{2} ||\beta_* - \beta_{k-1} + \epsilon \nabla f(\beta_{k-1})||^2 + \left(\frac{\epsilon^2 L}{2} - \epsilon\right) ||\nabla f(\beta_{k-1})||^2 \\ &\leqslant \epsilon k (f(\beta_{k-1}) - f(\beta_*)) + \frac{1}{2} ||\beta_* - \beta_{k-1}||^2 + \epsilon \langle \beta_* - \beta_{k-1}, \nabla f(\beta_{k-1}) \rangle + \frac{\epsilon^2}{2} ||\nabla f(\beta_{k-1})||^2 \\ &+ \left(\frac{\epsilon^2 L}{2} - \epsilon\right) ||\nabla f(\beta_{k-1})||^2 \\ &\leqslant \epsilon k (f(\beta_{k-1}) - f(\beta_*)) + \frac{1}{2} ||\beta_* - \beta_{k-1}||^2 + \epsilon \langle \beta_* - \beta_{k-1}, \nabla f(\beta_{k-1}) \rangle \\ &+ \left[\frac{\epsilon^2}{2} + \left(\frac{\epsilon^2 L}{2} - \epsilon\right)\right] ||\nabla f(\beta_{k-1})||^2 \\ &\leqslant \epsilon k (f(\beta_{k-1}) - f(\beta_*)) + \frac{1}{2} ||\beta_* - \beta_{k-1}||^2 + \epsilon \langle \beta_* - \beta_{k-1}, \nabla f(\beta_{k-1}) \rangle \\ &+ \left[\frac{\epsilon^2 (1 + L)}{2} - \epsilon\right] ||\nabla f(\beta_{k-1})||^2. \end{aligned}$$

Agora, tomando $\epsilon > 0$ como

$$\epsilon < \frac{2}{(1+L)},$$

obtemos que

$$\frac{\epsilon^2(1+L)}{2} - \epsilon < 0.$$

Isso implica que

$$\mathcal{E}_k \leqslant \epsilon k(f(\beta_{k-1}) - f(\beta_*)) + \frac{1}{2}||\beta_* - \beta_{k-1}||^2 + \epsilon\langle\beta_* - \beta_{k-1}, \nabla f(\beta_{k-1})\rangle.$$

Somando e subtraindo $\epsilon(f(\beta_{k-1}) - f(\beta_*))$ do lado direito da desigualdade acima, obtemos

$$\begin{aligned}\mathcal{E}_k &\leqslant \epsilon k(f(\beta_{k-1}) - f(\beta_*)) + \frac{1}{2}||\beta_* - \beta_{k-1}||^2 + \epsilon(f(\beta_{k-1}) - f(\beta_*)) - \epsilon(f(\beta_{k-1}) - f(\beta_*)) \\ &\quad + \epsilon\langle\beta_* - \beta_{k-1}, \nabla f(\beta_{k-1})\rangle \\ &\leqslant \epsilon(k-1)(f(\beta_{k-1}) - f(\beta_*)) + \frac{1}{2}||\beta_* - \beta_{k-1}||^2 + \epsilon(f(\beta_{k-1}) - f(\beta_*)) \\ &\quad + \epsilon\langle\beta_* - \beta_{k-1}, \nabla f(\beta_{k-1})\rangle.\end{aligned}$$

Pela equação (3.19) temos, da desigualdade acima, que

$$\begin{aligned}\mathcal{E}_k &\leqslant \mathcal{E}_{k-1} + \epsilon(f(\beta_{k-1}) - f(\beta_*)) + \epsilon\langle\beta_* - \beta_{k-1}, \nabla f(\beta_{k-1})\rangle \\ &\leqslant \mathcal{E}_{k-1} + \epsilon\big[(f(\beta_{k-1}) - f(\beta_*)) + \langle\beta_* - \beta_{k-1}, \nabla f(\beta_{k-1})\rangle\big].\end{aligned}$$

Da convexidade de f sabemos que

$$f(\beta_*) \geqslant f(\beta_{k-1}) + \langle\beta_* - \beta_{k-1}, \nabla f(\beta_{k-1})\rangle.$$

Isso implica que

$$\mathcal{E}_k \leqslant \mathcal{E}_{k-1}.$$

Então, a sequência $(\mathcal{E}_k)_{k\geqslant 1}$ é monótona não crescente. Assim, obtemos

$$\epsilon k(f(\beta_k) - f(\beta_*)) \leqslant \mathcal{E}_k \leqslant \mathcal{E}_0 = \frac{1}{2}||\beta_0 - \beta_*||^2.$$

O que implica

$$f(\beta_k) - f(\beta_*) \leqslant \mathcal{O}\left(\frac{1}{\epsilon k}\right).$$

□

3.2.2 EDO de primeira ordem associada

Em muitos casos, é mais fácil a análise de um determinado procedimento no contexto contínuo do que no discreto pela maior disponibilidade de ferramentas analíticas desenvolvidas para o campo contínuo. Uma dessas ferramentas é a teoria das equações diferenciais ordinárias (EDO) que é uma teoria matemática bem estabelecida com ramificações importantes, como, por exemplo, a teoria dos sistemas dinâmicos. Além disso, há uma parte considerável das ciências da natureza que modelam os seus fenômenos de estudo via essa teoria com especial destaque para física na área da mecânica clássica. Para algumas classes de métodos numéricos é possível associar, de modo natural, uma EDO de uma determinada ordem que possibilita estudar algumas propriedades da sequência gerada pelo método através de sua contraparte contínua. Esse é o caso do método de gradiente descendente que pode ser associado a uma EDO de primeira ordem.

Considere $\beta \in C^1(\mathbb{R}_+; \mathbb{R}^p)$ e $\nabla f : \mathbb{R}^p \to \mathbb{R}^p$ o campo de vetores Lipschitz contínuo dado pelo vetor gradiente da função convexa $f : \mathbb{R}^p \to \mathbb{R}$. Então, temos o seguinte problema de valor inicial [1]:

$$\dot{\beta}(t) = -\nabla f(\beta(t)), \tag{3.20a}$$

$$\beta(0) = \beta_0. \tag{3.20b}$$

A EDO de primeira ordem (3.20) se relaciona com o método do gradiente descendente por meio de sua discretização pelo método das *diferenças finitas*. Como a curva $\beta : \mathbb{R}_+ \to \mathbb{R}^p$ é diferenciável, temos pela expansão de Taylor de primeira ordem, que

$$\beta(t + \epsilon) = \beta(t) + \dot{\beta}(t)\epsilon + \mathcal{O}(\epsilon^2), \tag{3.21}$$

Passando $\beta(t)$ para o lado esquerdo da igualdade (3.21) e dividindo a expressão por $\epsilon > 0$, obtemos

$$\frac{\beta(t + \epsilon) - \beta(t)}{\epsilon} = \dot{\beta}(t) + \mathcal{O}(\epsilon). \tag{3.22}$$

Fazendo a identificação $t = \epsilon k$ e $\beta(\epsilon k) =: \beta_k$, segue que

$$\frac{\beta_{k+1} - \beta_k}{\epsilon} \approx \dot{\beta}(\epsilon k). \tag{3.23}$$

[1] As notações $\dot{\beta}$ e $\ddot{\beta}$ têm origem na Física e representam a primeira e a segunda derivada da curva β, respectivamente.

As etapas apresentadas nas equações (3.22) e (3.23) são conhecidas como *método de Euler progressivo*. Esse é um método de discretização de EDOs muito comum e utilizado na prática. Agora, substituindo a equação (3.23) na equação (3.20), obtemos

$$\frac{\beta_{k+1} - \beta_k}{\epsilon} = -\nabla f(\beta_k) \quad \Rightarrow \quad \beta_{k+1} = \beta_k - \epsilon \nabla f(\beta_k),$$

Da definição da EDO de primeira ordem (3.20), temos que, a curva definida por $\beta(t) := \beta_*$ (em que β_* é ponto de mínimo da função objetivo f) é solução para o problema de valor inicial (3.20) quando se considera $\beta_0 = \beta_*$. Ademais, esta é chamada de solução estacionária ou estado estacionário. Como o campo de vetores, dado pelo vetor gradiente da função f é ortogonal as curvas de nível dessa função, temos que as curvas solução do problema de valor inicial (3.20) são ortogonais as curvas de nível da função objetivo e, portanto, a sequência gerada pelo método do gradiente descendente terá um comportamento próximo a este. Quanto a acurácia do método de Euler, pode-se apresentar o seguinte resultado.

Teorema 13 (Burden e Faires (2015))**.** *Seja $F \in C^1(\mathbb{R}^p; \mathbb{R}^p)$ um campo de vetores com derivada Lipschitz contínua de módulo $L > 0$. Considere $M > 0$ uma constante tal que $||\ddot{\beta}(t)|| < M$, em que $\beta : \mathbb{R}_+ \to \mathbb{R}^p$ é a curva solução do problema de valor inicial (3.20). Por fim, seja $(\beta_k)_{k \geqslant 0}$ a sequência gerada pelo método de Euler progressivo para algum tamanho de passo $\epsilon > 0$. Então,*

$$||\beta(\epsilon k) - \beta_k|| \leqslant \frac{\epsilon M}{2L}\big(\exp\{\epsilon k L\} - 1\big).$$

No caso do método do gradiente descendente $F(x) := -\nabla f(x)$ e com ferramentas mais sofisticadas da teoria das EDOs, outras propriedades podem ser inferidas sobre o comportamento da curva solução e, consequentemente, sobre o comportamento da sequência gerada pelo método do gradiente descendente.

3.2.3 Taxa de convergência das soluções da EDO associada

A maioria das EDOs não apresentam solução analítica, o que torna necessário a utilização de algum método numérico para a obtenção de soluções aproximadas. Nesse caso, é necessário verificar a compatibilidade entre a taxa de convergência da curva solução da EDO e da sequência gerada pelo método que a discretiza, para ter uma estimativa de quão custoso é a obtenção da aproximação numérica da solução real. O lema 14 trata da taxa de convergência para as curvas solução da EDO de primeira ordem (3.20).

Lema 14 (Taxa de convergência das curvas solução da EDO (3.20) associada ao método do gradiente descendente)**.** *Considere $f \in C^1(\mathbb{R}^p)$ uma função convexa e $\beta : \mathbb{R}_+ \to \mathbb{R}^p$ a curva solução da EDO (3.20). Então, a taxa de convergência da curva solução desta EDO é da ordem de*

$$f(\beta(t)) - f(\beta_*) \leqslant \mathcal{O}\left(\frac{1}{t}\right),$$

em que β_* é o ponto de mínimo para f.

Demonstração. Considere a função $\mathcal{E} : \mathbb{R}_+ \to \mathbb{R}_+$ definida por

$$\mathcal{E}(t) := t(f(\beta(t)) - f(\beta_*)) + \frac{1}{2}||\beta(t) - \beta_*||^2.$$

Derivando $\mathcal{E}$, obtemos que

$$\dot{\mathcal{E}}(t) = (f(\beta(t)) - f(\beta_*)) + t\langle \nabla f(\beta(t)), \dot{\beta}(t)\rangle + \langle \dot{\beta}(t), \beta(t) - \beta_*\rangle.$$

Como a curva β é solução da EDO (3.20), temos que

$$\dot{\mathcal{E}}(t) = (f(\beta(t)) - f(\beta_*)) - \langle \nabla f(\beta(t)), \beta(t) - \beta_*\rangle - t||\nabla f(\beta(t))||^2.$$

Além disso, dado que f é uma função convexa e diferenciável,

$$0 \geqslant f(\beta(t)) - f(\beta_*) - \langle \nabla f(\beta(t)), \beta(t) - \beta_*\rangle.$$

Como $-t||\nabla f(\beta(t))||^2$ é negativo para todo $t > 0$, obtemos $\dot{\mathcal{E}}(t) \leqslant 0$. Portanto, a função $\mathcal{E}$ é decrescente, o que implica

$$t(f(\beta(t)) - f(\beta_*)) \leqslant \mathcal{E}(t) \leqslant \mathcal{E}(0) = \frac{1}{2}||\beta_0 - \beta_*||^2.$$

Concluímos que

$$f(\beta(t)) - f(\beta_*) \leqslant \mathcal{O}\left(\frac{1}{t}\right).$$

□

Na seção §3.3, discutimos como o método do gradiente descendente pode ser acelerado dando origem ao *método de Nesterov* e a uma EDO (de segunda ordem) associada a ele.

3.3 Gradiente acelerado

O método do gradiente acelerado, ou de Nesterov, tem como base o gradiente descendente. Esse procedimento, utilizando a mesma informação disponível para o gradiente descendente, a saber, a primeira derivada da função objetivo, consegue obter uma taxa de convergência da ordem de $\mathcal{O}(1/\epsilon k^2)$. Essa aceleração ocorre por uma leve modificação do gradiente descendente (a introdução de uma sequência auxiliar) que altera o comportamento do método, fazendo com que a aceleração apareça na parte final da convergência.

Formalmente, considere $f \in C^1(\mathbb{R}^p)$ uma função convexa com derivada Lipschitz contínua de modulo $L > 0$. Agora, considere as sequências $(\beta_k)_{k\geqslant 1}$ e $(\zeta_k)_{k\geqslant 1}$ definidas por:

$$\beta_k := \zeta_{k-1} - \epsilon \nabla f(\zeta_{k-1}), \tag{3.24a}$$

$$\zeta_k := \beta_k + \frac{k-1}{k+2}(\beta_k - \beta_{k-1}), \tag{3.24b}$$

em que $0 < \epsilon < 1/L$ é o tamanho do passo.

O método de Nesterov atinge o limite inferior ótimo para os métodos que utilizam apenas a informação da primeira derivada da função objetivo. Esse fato surpreendente foi provado pelo próprio Nesterov e sua demonstração completa pode ser encontrada em Nesterov (2004). Nas próximas subseções vamos compreender melhor como a introdução da sequência $(\zeta_k)_{k\geqslant 0}$ em conjunto com o termo $\frac{k-1}{k+2}(\beta_k - \beta_{k-1})$ permitem que a sequência $(\beta_k)_{k\geqslant 1}$ seja acelerada uma ordem acima, isto é, de $\mathcal{O}(1/\epsilon k)$ para $\mathcal{O}(1/\epsilon k^2)$.

A figura 3.3, exibe o comportamento da sequência gerada pelo método do gradiente acelerado. Note como, a partir da introdução da sequência auxiliar, o comportamento do método do gradiente descendente se modificou consideravelmente e agora a trajetória da sequência não apresenta mais um movimento em zigue-zague.

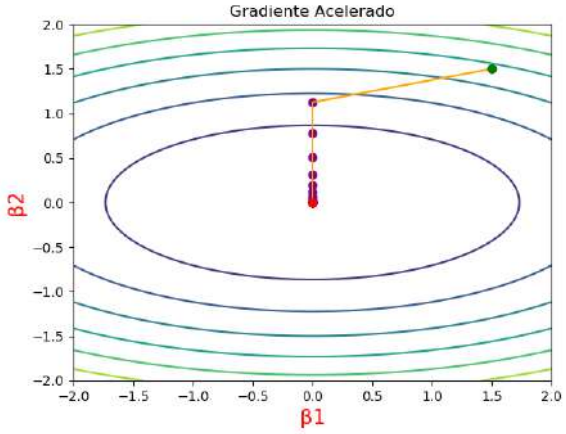

Figura 3.3: Comportamento da sequência gerada pelo método do gradiente acelerado para a função objetivo (3.2). O ponto verde é o valor em que é inicializado o método $\beta_0 = (1.5\,;\,1.5)$ e o ponto vermelho é o valor estimado pelo método $\beta_* = (0\,;\,0)$ para o ponto de mínimo.

3.3.1 Taxa de convergência

O método do gradiente acelerado tem um conjunto de características interessantes e uma das mais significativas para o campo da otimização é a sua taxa de convergência. Para estudar essa característica são necessários alguns resultados técnicos preliminares que são expostos abaixo em forma de lemas.

Lema 15. *Seja $f \in C^1(\mathbb{R}^p)$ uma função convexa com derivada Lipschitz contínua de modulo $L > 0$. Fixado $\epsilon \in (0, 1/L]$, temos, para todo $\beta_a, \beta_b \in \mathbb{R}^p$, que vale a seguinte desigualdade*

$$f(\beta_a - \epsilon \nabla f(\beta_a)) \leqslant f(\beta_b) + \langle \nabla f(\beta_a), \beta_a - \beta_b \rangle - \frac{\epsilon}{2} ||\nabla f(\beta_a)||^2. \tag{3.25}$$

Demonstração. Como f é uma função convexa e diferenciável, temos que

$$f(\beta_b) \geqslant f(\beta_a) + \langle \nabla f(\beta_a), \beta_b - \beta_a \rangle.$$

Isso implica

$$f(\beta_a) \leqslant f(\beta_b) + \langle \nabla f(\beta_a), \beta_a - \beta_b \rangle. \tag{3.26}$$

Agora, novamente pela diferenciabilidade de f, segue que

$$f(\beta_a - \epsilon \nabla f(\beta_a)) = f(\beta_a) - \epsilon ||\nabla f(\beta_a)||^2 + E_{\beta_a}(\epsilon),$$

em que $E_{\beta_a} : \mathbb{R}_+ \to \mathbb{R}_+$ é uma função tal que,

$$\lim_{\epsilon \to 0} \frac{E_{\beta_a}(\epsilon)}{\epsilon} = 0.$$

Dado $\eta := \frac{1}{2}||\nabla f(\beta_a)||^2$, existe $\delta(\eta) > 0$ tal que, se $0 < \epsilon < \delta(\eta)$, então

$$\frac{E_{\beta_a}(\epsilon)}{\epsilon} < \eta.$$

Isso implica

$$\begin{aligned} f(\beta_a - \epsilon\nabla f(\beta_a)) &\leqslant f(\beta_a) - \epsilon||\nabla f(\beta_a)||^2 + \frac{\epsilon}{2}||\nabla f(\beta_a)||^2 \\ &= f(\beta_a) - \frac{\epsilon}{2}||\nabla f(\beta_a)||^2. \end{aligned}$$

Combinando a desigualdade (3.26) com a apresentada acima, obtemos que

$$f(\beta_a - \epsilon\nabla f(\beta_a)) \leqslant f(\beta_b) + \langle \nabla f(\beta_a), \beta_a - \beta_b\rangle - \frac{\epsilon}{2}||\nabla f(\beta_a)||^2.$$

□

Lema 16. *Seja $f \in C^1(\mathbb{R}^p)$ uma função convexa com derivada Lipschitz contínua de modulo $L > 0$. Dado $\epsilon \in (0, 1/L]$, considere as sequências $(\zeta_k)_{k\geqslant 0}$ e $(\beta_k)_{k\geqslant 0}$ definidas na equação (3.24). Por fim, seja $(U_k)_{k\geqslant 0}$ uma sequência definida por*

$$U_k := \frac{k+2}{2}\zeta_k - \frac{k}{2}\beta_k. \tag{3.27}$$

Então, temos que

$$U_k = U_{k-1} - \epsilon\frac{k+1}{2}\nabla f(\zeta_{k-1}). \tag{3.28}$$

Demonstração. Inicialmente, note que

$$U_{k-1} - \epsilon\frac{k+1}{2}\nabla f(\zeta_{k-1}) = \frac{k+1}{2}\zeta_{k-1} - \frac{k-1}{2}\beta_{k-1} - \epsilon\frac{k+1}{2}\nabla f(\zeta_{k-1}).$$

Agora, dado que

$$\beta_k + \epsilon\nabla f(\zeta_{k-1}) = \zeta_{k-1}.$$

Podemos escrever

$$\begin{aligned} &U_{k-1} - \epsilon\frac{k+1}{2}\nabla f(\zeta_{k-1}) \\ &= \frac{k+1}{2}\beta_k + \frac{k+1}{2}\epsilon\nabla f(\zeta_{k-1}) - \frac{k-1}{2}\beta_{k-1} - \frac{k+1}{2}\epsilon\nabla f(\zeta_{k-1}) \\ &= \frac{k+1}{2}\beta_k - \frac{k-1}{2}\beta_{k-1}. \end{aligned}$$

Por outro lado,

$$\begin{aligned}\zeta_k &= \beta_k + \frac{k-1}{k+2}(\beta_k - \beta_{k-1}) \\ &= \beta_k + \frac{k-1}{k+2}\beta_k - \frac{k-1}{k+2}\beta_{k-1} \\ &= \left(1 + \frac{k-1}{k+2}\right)\beta_k - \frac{k-1}{k+2}\beta_{k-1} \\ &= \frac{2k+1}{k+2}\beta_k - \frac{k-1}{k+2}\beta_{k-1}.\end{aligned}$$

Multiplicando a igualdade acima por $(k+2)/2$, segue que

$$\frac{k+2}{2}\zeta_k - \frac{2k+1}{2}\beta_k = -\frac{k-1}{2}\beta_{k-1}.$$

Portanto,

$$\begin{aligned}&U_{k-1} - \epsilon\frac{k+1}{2}\nabla f(\zeta_{k-1}) \\ &= \frac{k+1}{2}\beta_k + \frac{k+2}{2}\zeta_k - \frac{2k+1}{2}\beta_k \\ &= \frac{k+2}{2}\zeta_k - \frac{k}{2}\beta_k = U_k.\end{aligned}$$

□

Com os dois resultados acima, é demonstrado no teorema abaixo que a taxa de convergência do método de Nesterov tem ordem $\mathcal{O}(1/\epsilon k^2)$.

Teorema 17 (Taxa de convergência do método de Nesterov)**.** *Seja $f \in C^1(\mathbb{R}^p)$ uma função convexa e com derivada Lipschitz contínua de modulo $L > 0$. Dado $\epsilon \in (0, 1/L]$, considere $(\beta_k)_{k\geqslant 0}$ a sequência gerada pelo método de Nesterov (3.24). Então, temos que*

$$f(\beta_k) - f(\beta_*) \leqslant \mathcal{O}\left(\frac{1}{\epsilon(k+1)^2}\right),$$

em que β_* é o ponto de mínimo para f.

Demonstração. Considere $(\mathcal{E}_k)_{k\geqslant 0}$ uma sequência definida por,

$$\mathcal{E}_k := \epsilon(k+1)^2(f(\beta_k) - f(\beta_*)) + 2||U_k - \beta_*||^2,$$

em que $(U_k)_{k\geqslant 0}$ é a sequência (3.27). Substituindo na equação (3.25) os valores $\beta_a = \zeta_{k-1}$ e $\beta_b = \beta_{k-1}$, obtemos

$$f(\zeta_{k-1} - \epsilon\nabla f(\zeta_{k-1})) \leqslant f(\beta_{k-1}) + \langle \nabla f(\zeta_{k-1}), \zeta_{k-1} - \beta_{k-1}\rangle - \frac{\epsilon}{2}||\nabla f(\zeta_{k-1})||^2. \tag{3.29}$$

Repetindo o processo acima para os valores $\beta_a = \zeta_{k-1}$ e $\beta_b = \beta_*$, segue que

$$f(\zeta_{k-1} - \epsilon\nabla f(\zeta_{k-1})) \leqslant f(\beta_*) + \langle \nabla f(\zeta_{k-1}), \zeta_{k-1} - \beta_* \rangle - \frac{\epsilon}{2}||\nabla f(\zeta_{k-1})||^2. \tag{3.30}$$

Multiplicando a equação (3.29) por $(k-1)/(k+1)$, a equação (3.30) por $2/(k+1)$ e depois somando ambas, obtemos que

$$\begin{aligned} f(\zeta_{k-1} - \epsilon\nabla f(\zeta_{k-1})) \leqslant{}& \frac{k-1}{k+1} f(\beta_{k-1}) + \frac{2}{k+1} f(\beta_*) \\ &+ \left\langle \nabla f(\zeta_{k-1}), \left[\frac{k-1}{k+1}\big(\zeta_{k-1} - \beta_{k-1}\big) + \frac{2}{k+1}\big(\zeta_{k-1} - \beta_*\big) \right] \right\rangle - \frac{\epsilon}{2}||\nabla f(\zeta_{k-1})||^2. \end{aligned}$$

Lembrando que $\beta_k = \zeta_{k-1} - \epsilon\nabla f(\zeta_{k-1})$, podemos reescrever a desigualdade acima como

$$\begin{aligned} f(\beta_k) \leqslant{}& \frac{k-1}{k+1} f(\beta_{k-1}) + \frac{2}{k+1} f(\beta_*) \\ &+ \left\langle \nabla f(\zeta_{k-1}), \left[\frac{k-1}{k+1}\big(\zeta_{k-1} - \beta_{k-1}\big) + \frac{2}{k+1}\big(\zeta_{k-1} - \beta_*\big) \right] \right\rangle - \frac{\epsilon}{2}||\nabla f(\zeta_{k-1})||^2. \end{aligned}$$

Dado que $U_{k-1} = \frac{k+1}{2}\zeta_{k-1} - \frac{k-1}{2}\beta_{k-1}$, obtemos

$$f(\beta_k) \leqslant \frac{k-1}{k+1} f(\beta_{k-1}) + \frac{2}{k+1} f(\beta_*) + \frac{2}{k+1}\langle \nabla f(\zeta_{k-1}), U_{k-1} - \beta_* \rangle - \frac{\epsilon}{2}||\nabla f(\zeta_{k-1})||^2.$$

Pela igualdade (3.28), temos que

$$f(\beta_k) \leqslant \frac{k-1}{k+1} f(\beta_{k-1}) + \frac{2}{k+1} f(\beta_*) + \frac{4}{\epsilon(k+1)^2}\langle U_{k-1} - U_k, U_{k-1} - \beta_* \rangle - \frac{\epsilon}{2}||\nabla f(\zeta_{k-1})||^2.$$

Agora, dado que

$$||\nabla f(\zeta_{k-1})||^2 = \frac{4}{\epsilon^2(k+1)^2}||U_{k-1} - U_k||^2.$$

Obtemos

$$\begin{aligned} f(\beta_k) \leqslant{}& \frac{k-1}{k+1} f(\beta_{k-1}) + \frac{2}{k+1} f(\beta_*) + \frac{4}{\epsilon(k+1)^2}\langle U_{k-1} - U_k, U_{k-1} - \beta_* \rangle \\ &- \frac{2}{\epsilon(k+1)^2}||U_{k-1} - U_k||^2. \end{aligned}$$

Por outro lado,

$$||U_{k-1} - \beta_* + \beta_* - U_k||^2 = ||U_{k-1} - \beta_*||^2 + 2\langle U_{k-1} - \beta_*, \beta_* - U_k \rangle + ||U_k - \beta_*||^2$$

e portanto,

$$\begin{aligned} f(\beta_k) \leqslant{}& \frac{k-1}{k+1} f(\beta_{k-1}) + \frac{2}{k+1} f(\beta_*) + \frac{4}{\epsilon(k+1)^2}\langle U_{k-1} - U_k, U_{k-1} - \beta_* \rangle \\ &- \frac{2}{\epsilon(k+1)^2}||U_{k-1} - \beta_*||^2 - \frac{4}{\epsilon(k+1)^2}\langle U_{k-1} - \beta_*, \beta_* - U_k \rangle - \frac{2}{\epsilon(k+1)^2}||U_k - \beta_*||^2. \end{aligned}$$

Além disso, temos que

$$\begin{aligned}&\langle U_{k-1} - \beta_* + \beta_* - U_k, U_{k-1} - \beta_* \rangle \\ &= \langle U_{k-1} - \beta_*, U_{k-1} - \beta_* \rangle + \langle \beta_* - U_k, U_{k-1} - \beta_* \rangle \\ &= ||U_{k-1} - \beta_*||^2 + \langle \beta_* - U_k, U_{k-1} - \beta_* \rangle.\end{aligned}$$

Isso implica que

$$\begin{aligned}f(\beta_k) \leqslant \frac{k-1}{k+1} f(\beta_{k-1}) + \frac{2}{k+1} f(\beta_*) + \frac{4}{\epsilon(k+1)^2} ||U_{k-1} - \beta_*||^2 + \frac{4}{\epsilon(k+1)^2} \langle \beta_* - U_k, U_{k-1} - \beta_* \rangle \\ - \frac{2}{\epsilon(k+1)^2} ||U_{k-1} - \beta_*||^2 - \frac{4}{\epsilon(k+1)^2} \langle U_{k-1} - \beta_*, \beta_* - U_k \rangle - \frac{2}{\epsilon(k+1)^2} ||U_k - \beta_*||^2.\end{aligned}$$

Portanto,

$$f(\beta_k) \leqslant \frac{k-1}{k+1} f(\beta_{k-1}) + \frac{2}{k+1} f(\beta_*) + \frac{2}{\epsilon(k+1)^2} ||U_{k-1} - \beta_*||^2 - \frac{2}{\epsilon(k+1)^2} ||U_k - \beta_*||^2.$$

Multiplicando ambos os lados da expressão acima por $\epsilon(k+1)^2$, obtemos

$$\epsilon(k+1)^2 f(\beta_k) \leqslant \epsilon(k+1)(k-1) f(\beta_{k-1}) + 2\epsilon(k+1) f(\beta_*) + 2||U_{k-1} - \beta_*||^2 - 2||U_k - \beta_*||^2.$$

Dessa desigualdade, segue-se que

$$\begin{aligned}\epsilon(k+1)^2 f(\beta_k) \leqslant \epsilon k^2 f(\beta_{k-1}) - \epsilon f(\beta_{k-1}) + 2\epsilon(k+1) f(\beta_*) \\ + 2||U_{k-1} - \beta_*||^2 - 2||U_k - \beta_*||^2.\end{aligned}$$

Como

$$\epsilon(k+1)^2 = \epsilon k^2 + 2\epsilon k + \epsilon \Rightarrow \epsilon(k+1)^2 - \epsilon k^2 - \epsilon = 2\epsilon k$$

Obtemos

$$\begin{aligned}\epsilon(k+1)^2 f(\beta_k) \leqslant \epsilon k^2 f(\beta_{k-1}) - \epsilon f(\beta_{k-1}) + \epsilon(k+1)^2 f(\beta_*) \\ - \epsilon k^2 f(\beta_*) - \epsilon f(\beta_*) + 2\epsilon f(\beta_*) + 2||U_{k-1} - \beta_*||^2 - 2||U_k - \beta_*||^2.\end{aligned}$$

O que implica

$$\begin{aligned}&\epsilon(k+1)^2 (f(\beta_k) - f(\beta_*)) + 2||U_k - \beta_*||^2 \\ &\leqslant \epsilon k^2 (f(\beta_{k-1}) - f(\beta_*)) + 2||U_{k-1} - \beta_*||^2 + \epsilon(f(\beta_*) - f(\beta_{k-1})).\end{aligned}$$

Portanto,

$$\mathcal{E}_k \leqslant \mathcal{E}_{k-1} + \epsilon(f(\beta_*) - f(\beta_{k-1})) \Rightarrow \mathcal{E}_k + \epsilon(f(\beta_{k-1}) - f(\beta_*)) \leqslant \mathcal{E}_{k-1}.$$

Trocando o índice k por i e somando de $i = 1$ até $i = k$, obtemos

$$\mathcal{E}_k + \sum_{i=1}^{k} \epsilon(f(\beta_{i-1}) - f(\beta_*)) \leqslant \mathcal{E}_0 = \epsilon(f(\beta_0) - f(\beta_*)) + 2||\beta_0 - \beta_*||^2.$$

Isso implica que

$$\mathcal{E}_k + \sum_{i=2}^{k} \epsilon(f(\beta_{i-1}) - f(\beta_*)) \leqslant 2||\beta_0 - \beta_*||^2.$$

Como $\sum_{i=2}^{k} \epsilon(f(\beta_{i-1}) - f(\beta_*)) > 0$, obtemos que

$$\epsilon(k+1)^2(f(\beta_k) - f(\beta_*)) \leqslant \mathcal{E}_k \leqslant 2||\beta_0 - \beta_*||^2.$$

O que implica

$$f(\beta_k) - f(\beta_*) \leqslant \mathcal{O}\left(\frac{1}{\epsilon(k+1)^2}\right).$$

□

3.3.2 EDO de segunda ordem associada

Nesta subseção, realizamos uma série de procedimentos análogos aos que foram efetuados na subseção 3.2.2. Mostraremos como uma EDO de segunda ordem emerge de modo natural do método de Nesterov. Nessa equação, destacamos dois componentes: o primeiro, é o termo de aceleração $\ddot{\beta}$, responsável pela alteração da velocidade da curva solução ao longo do tempo. O segundo, é o termo $\frac{3}{t}\dot{\beta}$, o qual pode ser entendido como um fator de resistência do meio ao longo da trajetória.

Considere $f \in C^1(\mathbb{R}^p)$ uma função convexa com derivada Lipschitz contínua de módulo $L > 0$. Vamos definir o seguinte problema de valor inicial:

$$\ddot{\beta}(t) + \frac{3}{t}\dot{\beta}(t) + \nabla f(\beta(t)) = 0 \quad \forall t \in \mathbb{R}_+, \tag{3.31a}$$

$$\beta(0) = \beta_0, \tag{3.31b}$$

$$\dot{\beta}(0) = 0. \tag{3.31c}$$

Observando a EDO (3.31) em conjunto com a figura 3.3, podemos empreender uma análise qualitativa simples que comece a esclarecer o comportamento da trajetória do método de Nesterov. Olhando para o segundo termo da EDO (3.31), temos, para valores pequenos de t, um valor elevado e interpretando a curva solução da EDO (3.31) como a trajetória de um corpo em um meio que lhe oferece resistência ao movimento, o corpo progride lentamente sem oscilações por esse meio.

Todavia, para valores elevados de t, o termo de resistência é aproximadamente zero e nesse meio de baixa resistência o termo de aceleração se torna dominante, surgindo oscilações próximas ao ponto no qual o campo de forças ∇f se anula. Para mostrar que a EDO (3.31) pode ser deduzida das sequências dadas na equação (3.24) é necessário a seguintes definições:

$$\mathcal{F}_L := \left\{ f \in C^1(\mathbb{R}^p) : ||\nabla f(\beta_a) - \nabla f(\beta_b)|| \leqslant L||\beta_a - \beta_b||; \forall \beta_a, \beta_b \in \mathbb{R}^p \right\},$$
$$\mathcal{F}_\infty := \bigcup_{L>0} \mathcal{F}_L.$$

Teorema 18 (Existência e unicidade de soluções para a EDO (3.31) - ver Su et al. (2016))**.** *Para toda $f \in \mathcal{F}_\infty$ e todo $\beta_0 \in \mathbb{R}^p$, a EDO (3.31), com condições iniciais $\beta(0) = \beta_0$ e $\dot{\beta}(0) = 0$, tem uma única solução global $\beta \in C^2(\mathbb{R}_+; \mathbb{R}^p) \cap C^1([0, +\infty); \mathbb{R}^p)$.*

Teorema 19 (Convergência do método de Nesterov (3.24) para as curvas solução da EDO (3.31) - ver Su et al. (2016))**.** *Para toda $f \in \mathcal{F}_\infty$, quando $\epsilon \to 0$, o método do gradiente acelerado (3.24) converge para a EDO (3.31) no seguinte sentido: fixado $T > 0$,*

$$\lim_{\epsilon \to 0} \max_{0 \leqslant k \leqslant \frac{T}{\sqrt{\epsilon}}} ||\beta_k - \beta(k\sqrt{\epsilon})|| = 0.$$

Pelo resultado do teorema 19, podemos estudar o comportamento da sequência gerada na equação (3.24) através da curva solução da EDO (3.31), o que justifica a análise qualitativa que foi feita anteriormente. Como uma segunda ilustração, podemos fazer o seguinte estudo sobre o comportamento assintótico inicial da curva solução da EDO (3.31).

Considere que o $\lim_{t \to 0} \ddot{\beta}(t) = \ddot{\beta}(0)$ existe. Pelo teorema do valor médio, existe $\xi \in (0, t)$ tal que, $\dot{\beta}(t) - \dot{\beta}(0) = \ddot{\beta}(\xi)t$. Como $\dot{\beta}(0) = 0$, podemos escrever $\dot{\beta}(t)/t = \ddot{\beta}(\xi)$ e pela EDO (3.24), obtemos

$$\ddot{\beta}(t) + 3\ddot{\beta}(\xi) + \nabla f(\beta(t)) = 0.$$

Tomando o limite $t \to 0$ na expressão acima, temos $\ddot{\beta}(0) = -\nabla f(\beta_0)/4$. Agora, considere um expansão de Taylor de segunda ordem para a curva solução da EDO (3.31) na vizinhança de 0. Então, temos que

$$\beta(t) = \beta(0) + \dot{\beta}(0)t + \ddot{\beta}(0)\frac{t^2}{2} + o(t^2).$$

Substituindo os valores de $\dot{\beta}(0)$ e $\ddot{\beta}(0)$ na expressão acima

$$\beta(t) = -\frac{\nabla f(\beta_0)t^2}{8} + \beta_0 + o(t^2). \tag{3.32}$$

Então, para valores pequenos de t, podemos estudar o comportamento das curvas solução da EDO (3.31) pela expressão (3.32). Outro contexto no qual é possível obter uma forma explícita para a curva β é quando a função f é quadrática. Nesse caso, a EDO (3.31) torna-se a EDO de Bessel de ordem 1 e podemos obter a curva β em função das funções de Bessel do primeiro tipo com ordem 1. Um estudo nessa direção é desenvolvido em Su et al. (2016).

3.3.3 *Taxa de convergência das soluções da EDO associada*

Nesta subseção, demonstramos que a taxa de convergência das curvas solução da EDO (3.31) é da ordem de $\mathcal{O}(1/t^2)$. Esse resultado, em conjunto com, os resultados alcançados nas subseções anteriores, possibilita a construção de uma interpretação que comece a dar significado ao procedimento desenvolvido por Nesterov. Uma discussão mais detalhada é realizada após a demonstração do lema abaixo.

Lema 20. *Para toda $f \in \mathcal{F}_\infty$, seja $\beta : \mathbb{R}_+ \to \mathbb{R}^p$ a única solução global da EDO (3.31) com condições iniciais $\beta(0) = \beta_0$ e $\dot{\beta}(0) = 0$. Então, para todo $t > 0$*

$$f(\beta(t)) - f(\beta_*) \leqslant \mathcal{O}\left(\frac{1}{t^2}\right).$$

Demonstração. Considere $\mathcal{E} : \mathbb{R}_+ \to \mathbb{R}_+$ uma função diferenciável definida por

$$\mathcal{E}(t) := t^2(f(\beta(t)) - f(\beta_*)) + 2||\beta(t) + \frac{t}{2}\dot{\beta}(t) - \beta_*||^2.$$

Derivando $\mathcal{E}$, obtemos

$$\dot{\mathcal{E}}(t) = 2t(f(\beta(t)) - f(\beta_*)) + t^2\langle\nabla f(\beta(t)), \dot{\beta}(t)\rangle + 4\left\langle \beta(t) + \frac{t}{2}\dot{\beta}(t) - \beta_*, \frac{3}{2}\dot{\beta}(t) + \frac{t}{2}\ddot{\beta}(t)\right\rangle.$$

Como a curva β é solução da EDO (3.31), temos que

$$\dot{\mathcal{E}}(t) = 2t(f(\beta(t)) - f(\beta_*)) - 2t\langle\nabla f(\beta(t)), \beta(t) - \beta_*\rangle.$$

Dado que f é uma função convexa e diferenciável, obtemos

$$0 \geqslant f(\beta(t)) - f(\beta_*) - \langle\nabla f(\beta(t)), \beta(t) - \beta_*\rangle.$$

Isso implica que

$$\dot{\mathcal{E}}(t) \leqslant 0.$$

Portanto, $\mathcal{E}$ é decrescente e segue-se que

$$t^2(f(\beta(t)) - f(\beta_*)) \leqslant \mathcal{E}(t) \leqslant \mathcal{E}(0) = 2||\beta_0 - \beta_*||^2.$$

O que implica

$$f(\beta(t)) - f(\beta_*) \leqslant \mathcal{O}\left(\frac{1}{t^2}\right).$$

□

Com os resultados acima, fazendo a discretização $t = \sqrt{\epsilon}k$ no domínio da EDO (3.31), obtemos, via o método do gradiente acelerado (3.24), uma discretização que preserva a taxa de convergência dessa EDO. A técnica de Nesterov pode ser aplicada a outros algoritmos clássicos, permitindo que sejam acelerados uma ordem acima. Isso leva a considerar a possibilidade de desenvolver um método geral de aceleração que permite, já de partida, a construção de algoritmos mais eficientes para muitos problemas, como, por exemplo, a estimação de parâmetros em modelos lineares generalizados.

3.4 *Gradiente acelerado de alta ordem*

Nas seções anteriores se começou com o método numérico de otimização e, a partir dele, foi obtido uma dinâmica, representada pela curva solução de uma EDO, no espaço em que se desejava minimizar a função objetivo. Aqui, fazemos o caminho inverso. Primeiro é munido o espaço $\mathbb{R}^p$, em que a função objetivo f está definida, com uma "métrica fraca", isto é, uma forma de medir que não possui todas as propriedades de uma métrica, como, por exemplo, simetria e a desigualdade triangular. Essa métrica fraca é chamada de *divergência de Bregman*.

Em seguida, definimos um funcional lagrangiano no espaço de fase $\mathbb{R}_+ \times \mathbb{R}^p \times \mathbb{R}^p$, em que se toma como energia potencial a função objetivo f e como energia cinética a divergência de Bregman. Pelo cálculo variacional é obtido uma dinâmica através da equação de *Euler-Lagrange*. As curvas solução da EDO obtida pela equação de Euler-Lagrange generaliza os métodos acelerados no domínio contínuo pela reparametrização de uma curva base gerada pelas condições iniciais com as quais se deseja inicializar o método.

Por fim, discretizamos essa EDO com o método de Euler progressivo e regressivo e é introduzido uma sequência auxiliar que possui algumas propriedades. Essa sequência é construída através de um operador que, sob certas condições de regularidade, generaliza

os métodos de descida. É provado, então, que as taxas de convergência das curvas solução da EDO obtida pelo método de Euler-Lagrange é compatível com a taxa de convergência da sequência gerada pela discretização desenvolvida acima.

3.4.1 EDO de segunda ordem associada

Como afirmado na introdução desta seção, a divergência de Bregman pode ser interpretada como uma métrica fraca em um espaço munido com uma estrutura gerada por uma função convexa e diferenciável. Como será visto na seção de aplicações, essa "métrica" apresenta vantagens no contexto da análise de dados em relação a uma métrica Riemanniana, pois permite um ajuste melhor a estrutura geométrica do espaço paramétrico.

Definição 21 (Divergência de Bregman)**.** *Considere $h \in C^1(\mathbb{R}^p)$ uma função convexa. A aplicação $D_h : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ definida por*

$$D_h(\beta_a : \beta_b) := h(\beta_a) - h(\beta_b) - \langle \nabla h(\beta_b), \beta_a - \beta_b \rangle, \tag{3.33}$$

para todo $\beta_a, \beta_b \in \mathbb{R}^p$ é chamada de divergência de Bregman.

Uma propriedade imediata que decorre da convexidade e diferenciabilidade da função h é que para todo $\beta_b, \beta_a \in \mathbb{R}^p$, temos que $D_h(\beta_a : \beta_b) \geqslant 0$. Outra propriedade importante é que a divergência de Bregman é uma função convexa na primeira variável.

Seja $f \in C^1(\mathbb{R}^p)$ uma função convexa. Desejamos minimizar essa função sobre o espaço $\mathbb{R}^p$. Considerando $\mathbb{R}^p$ o espaço de "configurações", interpreta-se f como a função "energia potencial" nos pontos desse espaço e ao muní-lo com a divergência de Bregman (21), $(\mathbb{R}^p, D_h)$, interpreta-se D_h como a "energia cinética". Com essas duas funções definimos o funcional $\mathcal{L}_B : \mathbb{R}_+ \times \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ no espaço de fase $\mathbb{R}_+ \times \mathbb{R}^p \times \mathbb{R}^p$ do seguinte modo.

Definição 22 (Lagrangiana de Bregman)**.** *Considere o espaço $(\mathbb{R}^p, D_h)$ e seja $f \in C^1(\mathbb{R}^p)$ uma função convexa. Considere $\alpha, \gamma, \eta : \mathbb{R}_+ \to \mathbb{R}$ funções diferenciáveis dadas a priori. A lagrangiana de Bregman é definida por*

$$\mathcal{L}_B(t, \beta, \dot{\beta}) := e^{\alpha(t)+\gamma(t)}(D_h(\beta + e^{-\alpha(t)}\dot{\beta} : \beta) - e^{-\eta(t)} f(\beta)). \tag{3.34}$$

As funções α, γ e η são tratadas como parâmetros a serem introduzidos no sistema e agem como a resistência do meio para a dinâmica definida por essa lagrangiana. Além

disso, vamos considerar que essas funções satisfazem as seguintes relações:

$$\dot{\eta}(t) \leqslant e^{\alpha(t)}, \tag{3.35a}$$

$$\dot{\gamma}(t) = e^{\alpha(t)}. \tag{3.35b}$$

Essas propriedades permitirão que algumas equações que são apresentadas a seguir sejam simplificadas possibilitando demonstrações mais simples.

A lagrangiana de Bregman define um sistema no espaço $(\mathbb{R}^p, D_h)$ e a dinâmica desse sistema é regida pelo *princípio da mínima ação*, isto é, que a dinâmica do sistema é obtida quando se minimiza a integral da lagrangiana sobre o espaço de trajetórias - curvas - desse sistema. Toma-se, então, $C^2(\mathbb{R}_+, \mathbb{R}^p)$ como o espaço das curvas que dão as trajetórias do sistema e assim, o *funcional de ação* $\Phi : C^2(\mathbb{R}_+, \mathbb{R}^p) \to \mathbb{R}$ é definido por

$$\Phi(\beta) := \int_{\mathbb{R}_+} \mathcal{L}_B(t, \beta(t), \dot{\beta}(t))dt, \tag{3.36}$$

Do cálculo variacional, a condição necessária para que uma curva em $C^2(\mathbb{R}_+, \mathbb{R}^p)$ seja minimizante para (3.36) é que essa curva satisfaça a equação de Euler-Lagrange

$$\frac{d}{dt}\frac{\partial \mathcal{L}_B}{\partial \dot{\beta}}(t, \beta, \dot{\beta}) = \frac{\partial \mathcal{L}_B}{\partial \beta}(t, \beta, \dot{\beta}). \tag{3.37}$$

Desenvolvendo a lagrangiana de Bregman (3.34) em relação a equação de Euler-Lagrange (3.37), obtemos a seguinte EDO de segunda ordem

$$\begin{aligned}
&\ddot{\beta}(t) + \big(e^{\alpha(t)} - \dot{\alpha}(t)\big)\dot{\beta}(t) \\
&\quad + e^{2\alpha(t)+\eta(t)}\big[\nabla^2 h(\beta(t) + e^{-\alpha(t)}\dot{\beta}(t))\big]^{-1}\nabla f(\beta(t)) \\
&\quad + e^{\alpha(t)}\big(\dot{\gamma}(t) - e^{\alpha(t)}\big)\big[\nabla^2 h(\beta(t) + e^{-\alpha(t)}\dot{\beta}(t))\big]^{-1}\nabla h(\beta(t) + e^{-\alpha(t)}\dot{\beta}(t)) \\
&\quad - e^{\alpha(t)}\big(\dot{\gamma}(t) - e^{\alpha(t)}\big)\big[\nabla^2 h(\beta(t) + e^{-\alpha(t)}\dot{\beta}(t))\big]^{-1}\nabla h(\beta(t)) = 0.
\end{aligned} \tag{3.38}$$

Como podemos notar, a EDO (3.38) tem uma estrutura complexa e difícil de trabalhar. Por isso, será imposto sobre as funções α, γ e η as condições (3.35) e como consequência, obtemos a EDO

$$\ddot{\beta}(t) + \big(e^{\alpha(t)} - \dot{\alpha}(t)\big)\dot{\beta}(t) + e^{2\alpha(t)+\eta(t)}\big[\nabla^2 h(\beta(t) + e^{-\alpha(t)}\dot{\beta}(t))\big]^{-1}\nabla f(\beta(t)) = 0 \tag{3.39}$$

em que a matriz hessiana $\nabla^2 h$ é inversível.

Antes de prosseguir as análises, é necessário um aprofundamento maior sobre a lagrangiana de Bregman. Essa possui diversas propriedades, e a mais notável, é que essa função é

fechada em relação a dilatações temporais. Isso significa que, se $\beta : \mathbb{R}_+ \to \mathbb{R}^p$ é solução da equação de Euler-Lagrange (3.39), então a sua reparametrização $\beta_R : \mathbb{R}_+ \to \mathbb{R}^p$, definida de modo a percorrer a mesma trajetória em um tempo distinto, também é solução de uma equação de Euler-Lagrange com parâmetros modificados. Formalmente, seja $\tau : \mathbb{R}_+ \to \mathbb{R}_+$ uma função crescente e diferenciável. Dado uma curva $\beta : \mathbb{R}_+ \to \mathbb{R}^p$, considere a curva $\beta_R : \mathbb{R}_+ \to \mathbb{R}^p$ definida por

$$\beta_R(t) := \beta(\tau(t)). \tag{3.40}$$

Isto é, a curva β_R é obtida da curva original β por uma contração ou dilatação temporal.

Teorema 23 (Dilatação temporal de soluções - ver Wibisono et al. (2016) teorema 2.2)**.** *Considere que a curva $\beta : \mathbb{R}_+ \to \mathbb{R}^p$ satisfaz a equação de Euler-Lagrange (3.38) para a lagrangiana de Bregman com parâmetros α, γ e η. Então, a curva (3.40) satisfaz a equação de Euler-Lagrange para a lagrangiana de Bregman com os parâmetros:*

$$\bar{\alpha}(t) := \alpha(\tau(t)) + \log(\dot{\tau}(t)), \tag{3.41a}$$

$$\bar{\eta}(t) := \eta(\tau(t)), \tag{3.41b}$$

$$\bar{\gamma}(t) := \gamma(\tau(t)). \tag{3.41c}$$

Além disso, $\bar{\alpha}$, $\bar{\eta}$ e $\bar{\gamma}$ satisfazem as condições (3.35) se, e somente se, α, η e γ também satisfazem.

O resultado estabelecido no teorema 23 afirma que, a família completa de métodos acelerados em tempo contínuo corresponde a uma única curva no espaço $\mathbb{R}^p$ que é percorrida com diferentes velocidades.

3.4.2 Taxa de convergência das soluções da EDO associada

Assim como no caso dos métodos do gradiente descendente e gradiente acelerado nesta subseção, é estabelecido a taxa de convergência para as curvas solução da EDO (3.39).

Teorema 24 (Taxa de convergência das curvas solução da EDO (3.39))**.** *Seja $f \in C^1(\mathbb{R}^p)$ uma função convexa. A taxa de convergência das soluções da EDO (3.39) é da ordem de*

$$f(\beta(t)) - f(\beta_*) \leqslant \mathcal{O}\big(e^{-\eta(t)}\big).$$

Demonstração. Considere a função $\mathcal{E} : \mathbb{R}_+ \to \mathbb{R}_+$ definida por

$$\mathcal{E}(t) := D_h(\beta_* : \beta(t) + e^{-\alpha(t)}\dot{\beta}(t)) + e^{\eta(t)}\big(f(\beta(t)) - f(\beta_*)\big). \tag{3.42}$$

Derivando (3.42), obtemos que

$$\dot{\mathcal{E}}(t) = \left\langle -\frac{d}{dt}\nabla h(\beta(t) + e^{-\alpha(t)}\dot{\beta}(t)), \beta_* - \beta(t) - e^{-\alpha(t)}\dot{\beta}(t) \right\rangle + \dot{\eta}(t)e^{\eta(t)}\big(f(\beta(t)) - f(\beta_*)\big) + e^{\eta(t)}\langle \nabla f(\beta(t)), \dot{\beta}(t)\rangle.$$

Se a curva $\beta : \mathbb{R}_+ \to \mathbb{R}^p$ satisfaz a equação de Euler-Lagrange (3.39), então, a expressão acima se simplifica para

$$\dot{\mathcal{E}}(t) = -e^{\alpha(t)+\eta(t)} D_f(\beta_* : \beta(t)) + \big(\dot{\eta}(t) - e^{\alpha(t)}\big)e^{\eta(t)}\big(f(\beta(t)) - f(\beta_*)\big),$$

em que $D_f(\beta_* : \beta(t)) = f(\beta_*) - f(\beta(t)) - \langle \nabla f(\beta(t)), \beta_* - \beta(t)\rangle$ é a divergência de Bregman em relação a função f. Como f é uma função convexa e diferenciável, temos que $D_f(\beta_* : \beta(t)) \geqslant 0$ e, portanto, o primeiro termo em $\dot{\mathcal{E}}$ é negativo. Dadas as propriedades (3.35), o segundo termo em $\dot{\mathcal{E}}$ também é negativo. Então, $\dot{\mathcal{E}}(t) \leqslant 0$. Logo, $\mathcal{E}$ é uma função decrescente e obtemos, para todo $t \geqslant 0$, que

$$\begin{aligned} & e^{\eta(t)}\big(f(\beta(t)) - f(\beta_*)\big) \leqslant \mathcal{E}(t) \leqslant \mathcal{E}(0) \\ \Rightarrow\ & f(\beta(t)) - f(\beta_*) \leqslant \mathcal{O}\big(e^{-\eta(t)}\big). \end{aligned}$$

□

Note que esse resultado afirma que é possível obtermos taxas de convergência exponenciais. Entretanto, nessa dissertação será abordada apenas a convergência polinomial, pois no caso exponencial, o procedimento de discretização da EDO (3.39) que garante a compatibilidade das taxas ainda não é completamente compreendido. Uma discussão mais elaborada sobre as complexidades envolvidas nesse processo pode ser encontrada em Wibisono et al. (2016).

Por intermédio da escolha das funções α, γ e η, podemos selecionar subfamílias de lagrangianas de Bregman. O caso em que temos interesse é o da subfamília polinomial. Para isso, dado $C > 0$ uma constante real e $m \geqslant 2$ tal que $f \in C^{m-1}(\mathbb{R}^p)$, definimos as funções:

$$\alpha(t) := \log(m) - \log(t), \tag{3.43a}$$

$$\eta(t) := m\log(t) + \log(C), \tag{3.43b}$$

$$\gamma(t) := m\log(t). \tag{3.43c}$$

As funções (3.43) satisfazem as condições apresentadas em (3.35) com uma igualdade na primeira condição. Combinando as funções (3.43) com a EDO (3.39), obtemos

$$\ddot{\beta}(t) + \frac{m+1}{t}\dot{\beta}(t) + Cm^2t^{m-2}\left[\nabla^2 h\left(\beta(t) + \frac{t}{m}\dot{\beta}(t)\right)\right]^{-1}\nabla f(\beta(t)) = 0. \tag{3.44}$$

Pelo teorema 24, segue-se que as curvas solução da EDO (3.44) têm taxa de convergência da ordem de $\mathcal{O}(1/t^m)$. Temos, como consequência direta do teorema 23 que, por exemplo, dado uma curva base, com $m = 2$, todas as demais curvas podem ser obtidas pela dilatação temporal $\tau(t) := t^{\frac{m}{2}}$. Além disso, quando a matriz $\nabla^2 h$ é a identidade, obtemos a EDO (3.31).

3.4.3 Taxa de convergência

O processo de discretização de uma EDO sempre constitui um desafio, principalmente quando desejamos que a discretização obtida preserve certas propriedades do caso contínuo. Começamos o processo de discretização da EDO (3.44) pela transformação desta em uma sistema de EDOs de primeira ordem. Para isso, consideramos a variável auxiliar (curva auxiliar) $Z : \mathbb{R}_+ \to \mathbb{R}^p$ definida por

$$Z(t) := \beta(t) + \frac{t}{m}\dot{\beta}(t).$$

Com base nessa variável, obtemos o sistema de EDOs de primeira ordem:

$$Z(t) = \beta(t) + \frac{t}{m}\dot{\beta}(t), \tag{3.45a}$$

$$\frac{d}{dt}\nabla h(Z(t)) = -Cmt^{m-1}\nabla f(\beta(t)). \tag{3.45b}$$

Passamos, então, a discretização dos domínios das curvas Z e β. Dado $\delta > 0$, definimos a sequência $t_k := \delta k$ e com base nela, as sequências $(\beta_k)_{k\geqslant 1}$ e $(Z_k)_{k\geqslant 1}$ do seguinte modo:

$$\beta_k := \beta(t_k) \quad \text{e} \quad Z_k := Z(t_k).$$

Para as derivadas de β e Z, aplicamos o método de Euler progressivo

$$\dot{\beta}(t_k) \approx \frac{\beta(t_k+\delta) - \beta(t_k)}{\delta} = \frac{\beta_{k+1} - \beta_k}{\delta},$$
$$\dot{Z}(t_k) \approx \frac{Z(t_k+\delta) - Z(t_k)}{\delta} = \frac{Z_{k+1} - Z_k}{\delta}.$$

Aplicando essas duas discretizações na equação (3.45a), segue-se que

$$\beta_{k+1} = \frac{m}{k}Z_k + \frac{k-m}{k}\beta_k. \tag{3.47}$$

Analogamente, aplicando o método de Euler regressivo na equação (3.45b), obtemos que

$$\frac{\nabla h(Z_k) - \nabla h(Z_{k-1})}{\delta} = -Cm(\delta k)^{m-1}\nabla f(\beta_k),$$

e a expressão acima pode ser reescrita como um problema de otimização

$$Z_k = \arg\min_Z \{Cmk^{m-1}\langle \nabla f(\beta_k), Z\rangle + (1/\epsilon)D_h(Z : Z_{k-1})\}, \tag{3.48}$$

com $\epsilon = \delta^m$. A princípio, as sequências nas equações (3.47) e (3.48) definem um algoritmo que implementa a dinâmica da EDO (3.45) em tempo discreto. Contudo, não é possível estabelecer uma taxa de convergência para esse algoritmo e de fato, empiricamente, encontramos que esse algoritmo é instável. Para transformar o algoritmo definido pelas sequências (3.47) e (3.48) em um algoritmo que, realmente, reproduza a dinâmica (3.45) em tempo discreto é necessária a introdução de uma sequência auxiliar que possui algumas propriedades. Considere a sequência $(\zeta_k)_{k\geqslant 1}$ (a ser definida mais a frente) e substitua nas sequências (3.47) e (3.48) o termo β_k por ζ_k obtendo assim o algoritmo abaixo

$$\beta_{k+1} = \frac{m}{(k+m)}Z_k + \frac{k}{(k+m)}\zeta_k, \tag{3.49a}$$

$$Z_k = \arg\min_Z \left\{Cmk^{(m-1)}\langle \nabla f(\zeta_k), Z\rangle + (1/\epsilon)D_h(Z : Z_{k-1})\right\}. \tag{3.49b}$$

em que $k^{(m-1)} := k(k+1)\dots(k+m-2)$ é o fatorial crescente. Afirmamos que uma condição suficiente para que o algoritmo (3.49) tenha uma taxa de convergência de $\mathcal{O}(1/\epsilon k^m)$ é que a sequência $(\zeta_k)_{k\geqslant 1}$ satisfaça a desigualdade

$$\langle \nabla f(\zeta_k), \beta_k - \zeta_k\rangle \geqslant M\epsilon^{\frac{1}{m-1}}||\nabla f(\zeta_k)||^{\frac{m}{m-1}}, \tag{3.50}$$

para alguma constante $M > 0$. Note que algumas modificações foram feitas nas sequências definidas por (3.47) e (3.48) em relação a sequência (3.49). Houve uma substituição do peso m/k por $m/(k+m)$. Essa mudança é somente para simplificar os cálculos, pois não altera o comportamento assintótico da sequência, dado que $m/k = \mathcal{O}(m/(k+m))$ quando $k \to \infty$. Analogamente houve a substituição de k^{m-1} pelo fatorial crescente $k^{(m-1)}$ em que também temos $k^{(m-1)} = \mathcal{O}\big(k^{m-1}\big)$ quando $k \to \infty$.

Para o próximo resultado, necessitamos da hipótese de que a função h, que gera a divergência de Bregman sobre o espaço $\mathbb{R}^p$, é *1-uniformemente convexa de ordem* $m \geqslant 2$. Por isso, esse conceito é relembrado na definição a seguir.

Definição 25. *Seja $h \in C^1(\mathbb{R}^p)$ uma função convexa. Então, h é uma função* σ-uniformemente convexa de ordem $m \geqslant 2$ *se para todo $\beta_a, \beta_b \in \mathbb{R}^p$ ocorre que*

$$D_h(\beta_a : \beta_b) \geqslant \frac{\sigma}{m}||\beta_a - \beta_b||^m.$$

Lema 26 (Ver Wibisono et al. (2016) lema A.1)**.** *Considere $f \in C^1(\mathbb{R}^p)$ uma função convexa e $h \in C^1(\mathbb{R}^p)$ uma função 1-uniformemente convexa de ordem $m \geqslant 2$. Dado $\epsilon > 0$, seja $(\zeta_k)_{k\geqslant 0}$ uma sequência que satisfaz a desigualdade (3.50) e considere $(\psi_k)_{k\geqslant 0}$ uma sequência de funções definidas por*

$$\psi_k(\beta) := Cm\sum_{i=0}^{k} i^{(m-1)}\big[f(\zeta_i) + \langle \nabla f(\zeta_i), \beta - \zeta_i\rangle\big] + (1/\epsilon)D_h(\beta : \beta_0), \tag{3.51}$$

em que $C > 0$ é uma constante. Então, para todo $k \geqslant 0$, segue-se que

$$\psi_k(Z_k) \geqslant Ck^{(m)}f(\zeta_k), \tag{3.52}$$

em que $(Z_k)_{k\geqslant 0}$ é a sequência gerada pelo algoritmo (3.49).

Teorema 27. *Suponha que a função h é 1-uniformemente convexa de ordem $m \geqslant 2$ e que a sequência $(\zeta_k)_{k\geqslant 0}$ satisfaz a desigualdade (3.50) para todo $k \geqslant 0$. Então, o algoritmo (3.49) com constante $C \leqslant M^{m-1}/m^m$ e condições iniciais $\beta_0 = Z_0 \in \mathbb{R}^p$ tem taxa de convergência da ordem de*

$$f(\zeta_k) - f(\beta_*) \leqslant \frac{D_h(\beta_* : \beta_0)}{C\epsilon k^{(m)}} = \mathcal{O}\left(\frac{1}{\epsilon k^m}\right). \tag{3.53}$$

Demonstração. Como f é convexa, a função ψ_k pode ser limitada superiormente por

$$\psi_k(\beta) \leqslant Cm\sum_{i=0}^{k} i^{(m-1)}f(\beta) + \frac{1}{\epsilon}D_h(\beta : \beta_0) = Ck^{(m)}f(\beta) + \frac{1}{\epsilon}D_h(\beta : \beta_0).$$

Essa desigualdade vale para todo $\beta \in \mathbb{R}^p$ e, em particular, para o minimizador β_* de f. Combinando a cota superior acima com o resultado do lema 26 e relembrando que Z_k é o minimizador de ψ_k, obtemos

$$Ck^{(m)}f(\zeta_k) \leqslant \psi_k(Z_k) \leqslant \psi_k(\beta_*) \leqslant Ck^{(m)}f(\beta_*) + \frac{1}{\epsilon}D_h(\beta_* : \beta_0).$$

Dividindo ambos os lados da desigualdade acima por $k^{(m)}$, segue-se que

$$f(\zeta_k) - f(\beta_*) \leqslant \frac{D_h(\beta_* : \beta_0)}{C\epsilon k^{(m)}} = \mathcal{O}\left(\frac{1}{\epsilon k^m}\right).$$

□

Observe que tomando $\epsilon = \delta^m$ obtemos que a taxa de convergência em tempo discreto, $\mathcal{O}(1/\epsilon k^m)$, torna-se compatível com a taxa de convergência em tempo contínuo, $\mathcal{O}(1/t^m)$ da EDO (3.44). Além disso, o resultado do teorema 27 não requer nenhuma hipótese adicional sobre a função f além da capacidade de produzir a sequência $(\zeta_k)_{k\geqslant 0}$. Nos próximos parágrafos, mostramos que é possível definir um operador que produz uma sequência que satisfaz a desigualdade (3.50).

Definição 28. *Seja $f \in C^{m-1}(\mathbb{R}^p)$ com $m \geqslant 2$. Então, o polinômio de Taylor de f em torno de $\beta \in \mathbb{R}^p$ de ordem $m-1$ é dada por*

$$f_{m-1}(\rho : \beta) := \sum_{i=0}^{m-1} \frac{1}{i!} \nabla^{(i)} f(\beta)(\rho - \beta)^i. \tag{3.54}$$

Com base na fórmula (3.54), definimos um operador de $\mathbb{R}^p$ em $\mathbb{R}^p$ do seguinte modo.

Definição 29. *Dados $N > 0$ e $\epsilon > 0$, definimos o operador $G_{m,\epsilon,N} : \mathbb{R}^p \to \mathbb{R}^p$ por*

$$G_{m,\epsilon,N}(\beta) := \arg\min_{\rho} \left\{ f_{m-1}(\rho : \beta) + (N/\epsilon m)||\rho - \beta||^m \right\}. \tag{3.55}$$

As definições a seguir serão necessárias para a demonstração que o operador (3.55) satisfaz a desigualdade (3.50).

Definição 30. *Seja $\mathcal{T}_m(\mathbb{R}^p; \mathbb{R}^p)$ o espaço das formas m lineares. Definimos em $\mathcal{T}_m(\mathbb{R}^p; \mathbb{R}^p)$ a seguinte norma*

$$||A||_{\mathcal{T}_m} := \sup_{||u_1||=\cdots=||u_m||=1} ||A(u_1, ..., u_m)||, \tag{3.56}$$

em que $A \in \mathcal{T}_m(\mathbb{R}^p; \mathbb{R}^p)$. Quando $u_1 = \cdots = u_m = u$, usa-se a notação $A(u)^m$ para indicar que o vetor u é distribuído nas m entradas de A, isto é, $A(u, ..., u)$.

Definição 31. *Seja $f \in C^{m-1}(\mathbb{R}^p)$. Dizemos que f é uma função L-*suave *de ordem $m-1$ se a derivada de f de ordem $m-1$ é Lipschitz contínua com constante de Lipschitz $L > 0$, isto é, para todo $\beta_a, \beta_b \in \mathbb{R}^p$, temos que*

$$||\nabla^{(m-1)} f(\beta_a) - \nabla^{(m-1)} f(\beta_b)||_{\mathcal{T}_m} \leqslant L||\beta_a - \beta_b||.$$

Com as definições acima, podemos passar para a demonstração do resultado principal.

Lema 32. *Considere $\beta \in \mathbb{R}^p$ e $\zeta := G_{m,\epsilon,N}(\beta)$ com $N > 1$. Se f é $L := \frac{(m-1)!}{\epsilon}$-suave de ordem $m-1$, então segue que*

$$\langle \nabla f(\zeta), \beta - \zeta \rangle \geqslant \frac{(N^2-1)^{\frac{m-2}{2m-2}}}{2N} \epsilon^{\frac{1}{m-2}} ||\nabla f(\zeta)||^{\frac{m}{m-1}}. \tag{3.57}$$

Demonstração. Seja $\Phi^f_{N,m,\epsilon} : \mathbb{R}^p \to \mathbb{R}$ uma função definida por

$$\Phi^f_{N,m,\epsilon}(\rho|\beta) := f_{m-1}(\rho : \beta) + (N/\epsilon m)||\rho - \beta||^m.$$

Podemos reescrever o operador (3.55) como

$$G_{m,\epsilon,N}(\beta) := \arg\min_{\rho} \Phi^f_{N,m,\epsilon}(\rho|\beta).$$

Dado que ζ é solução para o problema de otimização acima, a seguinte condição abaixo é satisfeita

$$\nabla\Phi^f_{N,m,\epsilon}(\zeta|\beta) = \sum_{i=1}^{m-1} \frac{1}{(i-1)!}\nabla^{(i)}f(\beta)(\zeta-\beta)^{i-1} + (N/\epsilon)||\zeta-\beta||^{m-2}(\zeta-\beta) = 0. \quad (3.58)$$

A expansão de Taylor de ordem $m-1$ para ∇f é dada por

$$\nabla f(\zeta) = \sum_{i=0}^{m-1} \frac{1}{i!}\nabla^{(i+1)}f(\beta)(\zeta-\beta)^i. \quad (3.59)$$

Fazendo a mudança de variável $i = j-1$ na equação (3.59), obtemos que

$$\nabla f(\zeta) = \sum_{j=1}^{m} \frac{1}{(j-1)!}\nabla^{(j)}f(\beta)(\zeta-\beta)^{j-1}.$$

Como $\nabla^{(m-1)}f$ é $(m-1)!/\epsilon$-suave, temos a seguinte desigualdade

$$\begin{aligned}
&\left|\left|\nabla f(\zeta) - \sum_{j=1}^{m-1} \frac{1}{(j-1)!}\nabla^{(j)}f(\beta)(\zeta-\beta)^{j-1}\right|\right| \\
&= \left|\left|\frac{1}{(m-1)!}\nabla^{(m)}f(\beta)(\zeta-\beta)^{m-1}\right|\right| \leqslant \left|\left|1(m-1)!\nabla^{(m)}f(\beta)\right|\right|_{\mathcal{T}_m} ||\zeta-\beta||^{m-1} \\
&= \frac{1}{(m-1)!}||\nabla^{(m)}f(\beta)||_{\mathcal{T}_m}||\zeta-\beta||^{m-1} \leqslant \frac{1}{(m-1)!}\frac{(m-1)!}{\epsilon}||\zeta-\beta||^{m-1} \\
&= \frac{1}{\epsilon}||\zeta-\beta||^{m-1}.
\end{aligned}$$

Ou seja,

$$\left|\left|\nabla f(\zeta) - \sum_{j=1}^{m-1} \frac{1}{(j-1)!}\nabla^{(j)}f(\beta)(\zeta-\beta)^{j-1}\right|\right| \leqslant \frac{1}{\epsilon}||\zeta-\beta||^{m-1}. \quad (3.61)$$

Substituindo (3.58) em (3.61) e tomando $r = ||\zeta-\beta||$, obtemos

$$\left|\left|\nabla f(\zeta) + \frac{Nr^{m-2}}{\epsilon}(\zeta-\beta)\right|\right| \leqslant \frac{r^{m-1}}{\epsilon}. \quad (3.62)$$

Elevando a norma ao quadrado e expandindo, segue-se que

$$\left|\left|\nabla f(\zeta) + \frac{Nr^{m-2}}{\epsilon}(\zeta-\beta)\right|\right|^2 = ||\nabla f(\zeta)||^2 - \frac{2Nr^{m-2}}{\epsilon}\langle\nabla f(\zeta), \beta-\zeta\rangle + \left(\frac{Nr^{m-2}}{\epsilon}\right)^2 ||\zeta-\beta||^2.$$

Combinando a expressão acima com a desigualdade 3.62, temos que

$$||\nabla f(\zeta)||^2 - \frac{2Nr^{m-2}}{\epsilon}\langle \nabla f(\zeta), \beta - \zeta\rangle + \left(\frac{Nr^{m-2}}{\epsilon}\right)^2 r^2 \leqslant \left(\frac{r^{m-1}}{\epsilon}\right)^2.$$

Rearranjando os termos obtemos a desigualdade

$$\langle \nabla f(\zeta), \beta - \zeta\rangle \geqslant \frac{\epsilon}{2Nr^{m-2}}||\nabla f(\zeta)||^2 + \frac{(N^2-1)r^m}{2N\epsilon}. \tag{3.63}$$

Note que, para $m = 2$, o primeiro termo em (3.63) já implica na desigualdade (3.57). Assumindo que $m \geqslant 3$, temos que o lado direito da equação (3.63) é da forma $\varphi(r) = A/r^{m-2} + Br^m$. A função φ está definida em $(0, +\infty)$ e é uma função convexa, pois $1/r^{m-2}$ e r^m são funções convexas nesse domínio e a combinação linear de funções convexas é uma função convexa. Se r^* é ponto de mínimo para φ, então temos que

$$\nabla\varphi(r^*) = 0,$$

e seque-se que

$$\begin{aligned} 0 = \nabla\varphi(r^*) &= -(m-2)A(r^*)^{-(m-1)} + Bm(r^*)^{m-1} \\ &\Rightarrow (m-2)A(r^*)^{-(m-1)} = Bm(r^*)^{m-1} \\ &\Rightarrow r^* = \left\{\frac{(m-2)}{m}\frac{A}{B}\right\}^{\frac{1}{2m-2}}. \end{aligned}$$

Então,

$$\begin{aligned} \varphi(r^*) &= A/(r^*)^{m-2} + B(r^*)^m \\ &= A^{\frac{m}{2m-2}}B^{\frac{m-2}{2m-2}}\left[\left(\frac{m}{m-2}\right)^{\frac{m-2}{2m-2}} + \left(\frac{m-2}{m}\right)^{\frac{m}{m-2}}\right] \\ &\geqslant A^{\frac{m}{2m-2}}B^{\frac{m-2}{2m-2}}. \end{aligned}$$

Substituindo os valores $A = (\epsilon/2N)||\nabla f(\zeta)||^2$ e $B = (1/2N\epsilon)(N^2-1)$ na desigualdade (3.63) obtemos

$$\begin{aligned} \langle \nabla f(\zeta), \beta - \zeta\rangle &\geqslant \left(\frac{\epsilon}{2N}||\nabla f(\zeta)||^2\right)^{\frac{m}{2m-2}}\left(\frac{1}{2N\epsilon}(N^2-1)\right)^{\frac{m-2}{2m-2}} \\ &= \frac{(N^2-1)^{\frac{m-2}{2m-2}}}{2N}\epsilon^{\frac{1}{m-1}}||\nabla f(\zeta)||^{\frac{m}{m-1}}. \end{aligned}$$

□

Com o resultado do lema 32, temos que o operador (3.55) produz uma sequência que satisfaz a desigualdade (3.50) e, portanto, completando o algoritmo (3.49) com essa sequência, obtemos um novo algoritmo que implementa a dinâmica, a tempo discreto, da EDO (3.44). Explicitamente, temos que

$$\begin{aligned}\beta_{k+1} &= \frac{m}{(k+m)}Z_k + \frac{k}{(k+m)}\zeta_k,\\ \zeta_k &= G_{m,\epsilon,N}(\beta_k),\\ Z_k &= \arg\min_{Z}\left\{Cmk^{(m-1)}\langle\nabla f(\zeta_k), Z\rangle + (1/\epsilon)D_h(Z : Z_{k-1})\right\}.\end{aligned}$$

Pelos teoremas 27 e 32 segue o seguinte corolário.

Corolário 33. *Suponha que f é $\frac{(m-1)!}{\epsilon}$-suave de ordem $m-1$, e que h é 1-uniformemente convexa de ordem m. Então, o algoritmo (3.65) com constantes $N > 1$ e $C \leqslant \frac{(N^2-1)^{\frac{m-2}{2}}}{(2N)^{m-1}m^m}$ e condições iniciais $\beta_0 = Z_0 \in \mathbb{R}^p$ tem taxa de convergência da ordem de $\mathcal{O}(1/\epsilon k^m)$.*

Observação: a implementação do exemplo da função (3.2) será deixado para o próximo capítulo em que é obtido o algoritmo (3.65) de forma explícita para o caso $m = 2$.

Capítulo 4

Gradiente acelerado de alta ordem aplicado aos MLGs

Para o processo de estimação de parâmetros nos MLGs via método do gradiente acelerado de alta ordem é necessário munir o espaço de parâmetros $\mathbb{R}^p$ com uma " medida" que permita obter informações sobre a estrutura do espaço paramétrico. Considerando que os MLGs tem por base a família exponencial de distribuições um caminho intuitivo, é olhar se é possível obter a função que desempenhará o papel de uma "métrica" como consequência das propriedades da família exponencial que articula os modelos. Para isso, considere a família exponencial em uma forma distinta da apresentada na definição 1,

$$p(y|\theta) = \exp\{y\theta - \psi(\theta)\}. \tag{4.1}$$

A expressão para a família exponencial em (4.1) pode ser obtida a partir da expressão dada na definição 1 pela aplicação de uma transformação na medida de probabilidade. Assim como na definição 1, temos na equação (4.1) que y é a variável resposta, θ o parâmetro canônico que indexa a família e $\psi : \mathbb{R} \to \mathbb{R}$ uma função convexa e diferenciável tal que $\psi' : \mathbb{R} \to \mathbb{R}$ é um difeomorfismo.

Associada a função de participação ψ, temos a função $\psi^* : \mathbb{R} \to \mathbb{R}$ dada pela *transformada de Legendre* que, por definição, é dada por

$$\psi^*(\mu) = \sup_{\theta \in \mathbb{R}} \{\mu\theta - \psi(\theta)\}. \tag{4.2}$$

Derivando a expressão dentro do supremo e igualando a zero (condição necessária de otimalidade para o cálculo do supremo), obtemos que

$$\psi'(\theta) = \mu.$$

Como a função ψ' é um difeomorfismo, a inversa de ψ' existe e, portanto, podemos escrever

$$\theta(\mu) := (\psi')^{-1}(\mu).$$

Substituindo a expressão acima na equação (4.2), segue-se que

$$\psi^*(\mu) = \mu\theta(\mu) - \psi(\theta(\mu)).$$

Derivando essa expressão, obtemos que

$$(\psi^*)'(\mu) = \theta(\mu) + \mu\theta'(\mu) - \psi'(\theta(\mu))\theta'(\mu) = \theta(\mu) + \mu\theta'(\mu) - \mu\theta'(\mu) = \theta(\mu).$$

Com esses resultados, podemos fazer o seguinte cálculo:

$$\begin{aligned}
y\theta - \psi(\theta) &= y\theta - \psi(\theta) - \mu\theta + \mu\theta \\
&= (\mu\theta - \psi(\theta)) + \theta(y - \mu) \\
&= \psi^*(\mu) + (\psi^*)'(\mu)(y - \mu) \\
&= \psi^*(\mu) + (\psi^*)'(\mu)(y - \mu) - \psi^*(y) + \psi^*(y) \\
&= -\big(\psi^*(y) - \psi^*(\mu) - (\psi^*)'(\mu)(y - \mu)\big) + \psi^*(y) \\
&= -D_{\psi^*}(y : \mu) + \psi^*(y),
\end{aligned}$$

Assim, podemos reescrever a função de probabilidade (ou densidade) na equação (4.1) como

$$p(y|\theta) = \exp\{-D_{\psi^*}(y : \mu(\theta)) + \psi^*(y)\}. \tag{4.3}$$

A variável dual μ é o parâmetro natural do modelo e a função $D_{\psi^*} : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$ é a divergência de Bregman dada na definição 21. Como discutido na subseção 3.4.1 a divergência de Bregman possui propriedades que a tornam, em certo sentido, semelhante a uma métrica. Note que o termo $D_{\psi^*}(y : \mu(\theta))$ "mede" a discrepância entre o parâmetro μ, que é a média do modelo, pois $\psi(\theta) = \mu = E[Y]$, e os dados y.

Portanto, a divergência de Bregman é uma forma natural de estruturar o espaço paramétrico, pois emerge da própria família exponencial usada para modelar os dados. Existe uma relação importante entre as divergências de Bregman associadas as funções ψ e ψ^*. Essa relação é dada por

$$D_{\psi^*}(y : \mu) = D_{\psi}(\theta : \tilde{y}).$$

Para provar essa relação inicialmente note que

$$\begin{aligned} D_{\psi^*}(y:\mu) &= \psi^*(y) - \psi^*(\mu) - (\psi^*)'(\mu)(y-\mu) \\ &= \psi^*(y) - \psi^*(\mu) - \theta(y-\mu) \\ &= \psi^*(y) - \psi^*(\mu) - \theta y + \theta\mu \\ &= (\theta\mu - \psi^*(\mu)) + \psi^*(y) - \theta y \\ &= \psi(\theta) + \psi^*(y) - \theta y. \end{aligned}$$

Por outro lado, utilizando o difeomorfismo $(\psi)'$ para fazer a transformação nos dados $(\psi)'(\tilde{y}) = y$, obtemos que

$$\begin{aligned} D_{\psi}(\theta:\tilde{y}) &= \psi(\theta) - \psi(\tilde{y}) - (\psi)'(\tilde{y})(\theta-\tilde{y}) \\ &= \psi(\theta) - \psi(\tilde{y}) - y(\theta-\tilde{y}) \\ &= \psi(\theta) - \psi(\tilde{y}) - y\theta + y\tilde{y} \\ &= (y\tilde{y} - \psi(\tilde{y})) + \psi(\theta) - \theta y \\ &= \psi^*(y) + \psi(\theta) - \theta y. \end{aligned}$$

Combinando esses dois desenvolvimentos segue-se que

$$D_{\psi^*}(y:\mu) = \psi(\theta) + \psi^*(y) - \theta y = \psi^*(y) + \psi(\theta) - \theta y = D_{\psi}(\theta:\tilde{y}).$$

Pelo resultado acima, podemos reescrever a distribuição de probabilidade (ou densidade) dada na equação (4.3) do seguinte modo

$$p(y|\theta) = \exp\{-D_{\psi}(\theta:(\psi')^{-1}(y)) + \psi^*(y)\}. \tag{4.4}$$

No método do gradiente acelerado de alta ordem, a divergência de Bregman aparece através da função convexa e diferenciável ψ. Por isso é necessário determinar a forma dessa função para os MLGs. Com a forma geral da família exponencial dada na equação (4.1) os MLGs assumem a seguinte forma

$$\begin{aligned} p(y_i|\theta_i) &= \exp\{y_i\theta_i - \psi(\theta_i)\}, \\ G(\mu_i) &= \mathbf{x}_i^T\beta, \\ \psi'(\theta_i) &= E[Y_i] = \mu_i. \end{aligned} \tag{4.5}$$

Nos próximos parágrafos será considerado apenas funções de ligação canônicas. A função de distribuição conjunta para as variáveis resposta do modelo (4.5) é dada por

$$p(y|\beta) := \exp\{T(y)^T\beta - \Psi(\beta)\}, \tag{4.6}$$

em que

$$T(y) := \sum_{i=1}^{n} y_i \mathbf{x}_i \quad \text{e} \quad \Psi(\beta) := \sum_{i=1}^{n} \psi(\mathbf{x}_i^T \beta).$$

Pelo que foi desenvolvido nos parágrafos anteriores, é possível associar uma divergência de Bregman ao modelo (4.6) que passa a ter a seguinte forma

$$p(y|\beta) = \exp\{-D_\Psi(\beta : (\Psi')^{-1}(T(y))) + \Psi^*(T(y))\}. \tag{4.7}$$

Para o método do gradiente acelerado de alta ordem, é necessário apenas que exista uma função convexa $h \in C^2(\mathbb{R}^p)$ que induza uma divergência de Bregman no espaço paramétrico. Também é necessário que a matriz hessiana da função h seja inversível. De fato, na formulação final do algoritmo apenas aparece a inversa da matriz hessiana da função h e por isso o estudo da matriz hessiana da função Ψ pode oferecer um caminho razoável para a definição da função h. A matriz hessiana da função Ψ é dada por

$$\nabla^2 \Psi(\beta) = X^T \text{diag}(\psi''(\mathbf{x}_1^T \beta), ..., \psi''(\mathbf{x}_n^T \beta)) X \tag{4.8}$$

em que diag$(\cdot)$ é a matriz diagonal cuja a diagonal é formada pelos elementos do vetor $(\psi''(\mathbf{x}_1^T \beta), ..., \psi''(\mathbf{x}_n^T \beta))$ e X é uma matriz $n \times p$ com posto p. Note que a matriz na equação (4.8) é a informação de Fisher para o MLG sob a forma geral da família exponencial.

Para evitar a necessidade de inverter a matriz hessiana da função h a cada iteração, o que aumentaria a complexidade computacional do método, busca-se uma função h que tenha matriz hessiana constante, mas que mantenha alguma relação com a estrutura da matriz $\nabla^2 \Psi(\beta)$. Como a matriz (4.8) é positiva semidefinida, segue-se para todo $\beta \in \mathbb{R}^p$ que $\nabla^2 \Psi(\beta) \in \mathbf{S}_+^p$, em que $\mathbf{S}_+^p$ é o conjunto das matrizes positivas semidefinidas de dimensão $p \times p$. Em $\mathbf{S}_+^p$ é possível definir uma relação de ordem parcial $\succeq$ do seguinte modo: dadas as matrizes $A \in \mathbf{S}_+^p$ e $B \in \mathbf{S}_+^p$, temos que $A \succeq B$ se, e somente se, para todo $u \in \mathbb{R}^p$ ocorre que $u^T(A - B)u \geqslant 0$. Busca-se, então, uma função h tal que $\nabla^2 h(\beta) \succeq \nabla^2 \Psi(\beta)$ com $\nabla^2 h(\beta)$ constante.

O caso de interesse nesta dissertação é quando a função de distribuição é uma Bernoulli. Nesse caso, a equação (4.8) fica dada por

$$\nabla^2 \Psi(\beta) = X^T \text{diag}\left(\frac{\exp\{\mathbf{x}_1^T \beta\}}{(1 + \exp\{\mathbf{x}_1^T \beta\})^2}, ..., \frac{\exp\{\mathbf{x}_n^T \beta\}}{(1 + \exp\{\mathbf{x}_n^T \beta\})^2} \right) X.$$

Como para todo $i \in \{1, ..., n\}$

$$0 < \frac{\exp\{\mathbf{x}_i^T \beta\}}{(1 + \exp\{\mathbf{x}_i^T \beta\})^2} < \frac{1}{4}, \tag{4.9}$$

obtemos que

$$\frac{1}{4}X^TX \succeq \nabla^2\Psi(\beta).$$

De fato, para todo $u \in \mathbb{R}^p$ segue-se que

$$\begin{aligned} u^T\left(\frac{1}{4}X^TX - \nabla^2\Psi(\beta)\right)u &= u^T\left(\frac{1}{4}X^TI_nX - \nabla^2\Psi(\beta)\right)u \\ &= u^T\left(\frac{1}{4}X^TI_nX - X^T\text{diag}\left(\frac{\exp\{\mathbf{x}_1^T\beta\}}{(1+\exp\{\mathbf{x}_1^T\beta\})^2}, ..., \frac{\exp\{\mathbf{x}_n^T\beta\}}{(1+\exp\{\mathbf{x}_n^T\beta\})^2}\right)X\right)u \\ &= u^TX^T\left(\frac{1}{4}I_n - \text{diag}\left(\frac{\exp\{\mathbf{x}_1^T\beta\}}{(1+\exp\{\mathbf{x}_1^T\beta\})^2}, ..., \frac{\exp\{\mathbf{x}_n^T\beta\}}{(1+\exp\{\mathbf{x}_n^T\beta\})^2}\right)\right)Xu. \end{aligned}$$

Tomando $Xu = v$, a desigualdade (4.9) implica que

$$u^T\left(\frac{1}{4}X^TX - \nabla^2\Psi(\beta)\right)u = v\left(\frac{1}{4}I_n - \text{diag}\left(\frac{\exp\{\mathbf{x}_1^T\beta\}}{(1+\exp\{\mathbf{x}_1^T\beta\})^2}, ..., \frac{\exp\{\mathbf{x}_n^T\beta\}}{(1+\exp\{\mathbf{x}_n^T\beta\})^2}\right)\right)v \geqslant 0,$$

de onde segue o resultado.

Uma observação a ser feita é que pela desigualdade (4.9) a matriz $\nabla^2\Psi(\beta)$ é positiva definida e, portanto, inversível. Esse fato pode ser traduzido em termos de desigualdades entre matrizes como $\nabla^2\Psi(\beta) \succ 0$ e pelo resultado acima, obtemos que

$$\frac{1}{4}X^TX \succeq \nabla^2\Psi(\beta) \succ 0.$$

Com os resultados já demonstrados a função h mais simples que pode ser escolhida para munir o espaço paramétrico com uma estrutura de forma a atender as condições estabelecidas para a matriz hessiana é

$$h(\beta) := \frac{1}{8}\beta^TX^TX\beta. \tag{4.10}$$

4.1 Gradiente acelerado de alta ordem em forma explícita para MLGs

Nesta subseção mostramos que o método do gradiente acelerado de alta ordem pode ser obtido explicitamente para os MLGs. Isso é possível por dois motivos. O primeiro vem do fato da derivada da função h ser invertível e, portanto, o procedimento de minimização que define a sequência $(Z_k)_{k\geqslant 0}$ pode ser calculado de forma explícita. O segundo vem do fato de que para $m = 2$ o procedimento de minimização que define a sequência $(\zeta_k)_{k\geqslant 1}$ também pode ser resolvido explicitamente e consequentemente o algoritmo dado na equação (3.65), como um todo, pode ser obtido explicitamente.

Tomando $m = 2$, o algoritmo (3.65) assume a seguinte forma

$$\beta_{k+2} = \frac{2}{(k+3)} Z_{k+1} + \frac{(k+1)}{(k+3)} \zeta_{k+1}, \tag{4.11a}$$

$$\zeta_{k+1} = \arg\min_{\zeta} \left\{ f(\beta_{k+1}) + \langle \nabla f(\beta_{k+1}), \zeta - \beta_{k+1} \rangle + N/(2\epsilon) ||\zeta - \beta_{k+1}||^2 \right\}, \tag{4.11b}$$

$$Z_{k+1} = \arg\min_{Z} \left\{ 2C(k+1) \langle \nabla f(\zeta_{k+1}), Z \rangle + (1/\epsilon) D_h(Z : Z_k) \right\}. \tag{4.11c}$$

Primeiro, calculamos a sequência $(Z_k)_{k \geqslant 0}$. Considere $(\mathcal{G}_k)_{k \geqslant 0}$ uma sequência de funções definidas por

$$\mathcal{G}_k(Z) := 2C(k+1) \langle \nabla f(\zeta_{k+1}), Z \rangle + (1/\epsilon) D_h(Z : Z_k).$$

Então, a sequência (4.11c) pode ser reescrita como

$$Z_{k+1} = \arg\min_{Z} \mathcal{G}_k(Z).$$

A condição necessária para que Z_{k+1} seja mínimo para $\mathcal{G}_k$ é que o vetor gradiente de $\mathcal{G}_k$ seja nulo em Z_{k+1}. Então

$$0 = \nabla \mathcal{G}_k(Z_{k+1}) = 2C(k+1) \nabla f(\zeta_{k+1}) + (1/\epsilon)(\nabla h(Z_{k+1}) - \nabla h(Z_k))$$

e isso implica que

$$\nabla h(Z_{k+1}) = \nabla h(Z_k) - 2C\epsilon(k+1) \nabla f(\zeta_{k+1}).$$

Como a derivada da função h é investível, pois $X^T X$ é invertível, segue-se que

$$\begin{aligned} \frac{1}{4} X^T X Z_{k+1} &= \frac{1}{4} X^T X Z_k - 2C\epsilon(k+1) \nabla f(\zeta_{k+1}) \Rightarrow \\ Z_{k+1} &= Z_k - 8C\epsilon(k+1)(X^T X)^{-1} \nabla f(\zeta_{k+1}). \end{aligned}$$

Passamos, então, para a sequência auxiliar $(\zeta_k)_{k \geqslant 1}$. Considere $(\mathcal{Q}_k)_{k \geqslant 1}$ uma sequência de funções definidas por

$$\mathcal{Q}_k(\zeta) := f(\beta_{k+1}) + \langle \nabla f(\beta_{k+1}), \zeta - \beta_{k+1} \rangle + N/(2\epsilon) ||\zeta - \beta_{k+1}||^2.$$

Então, a sequência (4.11b) pode ser reescrita como

$$\zeta_{k+1} = \arg\min_{\zeta} \mathcal{Q}_k(\zeta).$$

Assim como antes, a condição necessária para que ζ_{k+1} seja mínimo para $\mathcal{Q}_k$ é que o vetor gradiente de $\mathcal{Q}_k$ seja nulo em ζ_{k+1}. Então

$$0 = \nabla \mathcal{Q}_k(\zeta_{k+1}) = \nabla f(\beta_{k+1}) + (N/\epsilon)(\zeta_{k+1} - \beta_{k+1}),$$

e isso implica que

$$\zeta_{k+1} = \beta_{k+1} - (\epsilon/N)\nabla f(\beta_{k+1}).$$

Concluímos que o algoritmo do gradiente acelerado de alta ordem para $m = 2$ na forma explícita é dado por

$$\begin{aligned}
\beta_{k+2} &= \frac{2}{(k+3)} Z_{k+1} + \frac{(k+1)}{(k+3)} \zeta_{k+1}, \\
\zeta_{k+1} &= \beta_{k+1} - \frac{\epsilon}{N} \nabla f(\beta_{k+1}), \\
Z_{k+1} &= Z_k - 8C\epsilon(k+1)(X^T X)^{-1} \nabla f(\zeta_{k+1}).
\end{aligned} \tag{4.12}$$

Uma das hipóteses do Corolário 33 é que a constante $C > 0$ que aparece em (4.12) deve satisfazer a seguinte desigualdade

$$C \leqslant \frac{(N^2-1)^{\frac{m-2}{2}}}{(2N)^{m-1} m^m}$$

em que $N > 1$. Como estamos tomando $m = 2$, a desigualdade acima fica dada por

$$C \leqslant \frac{1}{8N}.$$

Tomando $N = 2$ obtemos que $C \leqslant 1/16$ e podemos reescrever o algoritmo (4.12) como

$$\begin{aligned}
\beta_{k+2} &= \frac{2}{(k+3)} Z_{k+1} + \frac{(k+1)}{(k+3)} \zeta_{k+1}, \\
\zeta_{k+1} &= \beta_{k+1} - \frac{\epsilon}{N} \nabla f(\beta_{k+1}), \\
Z_{k+1} &= Z_k - \frac{\epsilon(k+1)}{2} (X^T X)^{-1} \nabla f(\zeta_{k+1}).
\end{aligned} \tag{4.13}$$

Na figura 4.1 é apresentado o comportamento da sequência gerada pelo método do gradiente acelerado de alta ordem aplicado a função objetivo (3.2). Note a semelhança com o gráfico da sequência gerada pelo método de Newton-Raphson. Um dos fatores que pode justificar esse comportamento é o fato que, assim como no caso do Newton-Raphson, o método do gradiente acelerado de alta ordem também faz o uso implícito de informações sobre a curvatura da superfície via inversa da matriz hessiana da função h que, nesse caso, foi tomada como a própria função objetivo f.

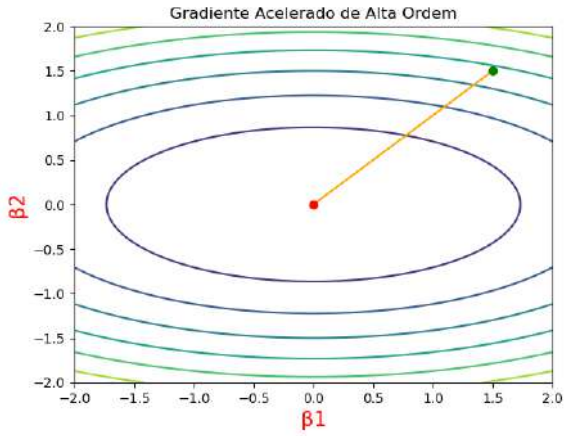

Figura 4.1: Comportamento da sequência gerada pelo método do gradiente acelerado de alta ordem para a função objetivo (3.2). O ponto verde é o valor em que é inicializado o método $\beta_0 = (1,5\,;\,1,5)$ e o ponto vermelho é o valor estimado pelo método $\beta_* = (0,0\,;\,0,0)$ para o ponto de mínimo.

Capítulo 5

Comparação entre os métodos de otimização

Neste capítulo se realiza uma série de análises a fim de comparar os métodos de Newton-Raphson, gradiente descendente, gradiente acelerado e gradiente acelerado de alta ordem. Como o estudo se dá sobre os algoritmos que implementam esses métodos, entra-se no domínio de uma área da ciência da computação conhecida como *análise de algoritmos*. Esse ramo do conhecimento costuma ser dividido em duas linhas de análise distintas. A primeira é conhecida como *análise matemática de algoritmos* e tem como objeto de estudo os aspectos formais da implementação de um determinado algoritmo, como, por exemplo, a ordem de complexidade desse algoritmo quando se olha para o tamanho da entrada de dados. Em geral, nesse caso, busca-se obter uma função, $H : \mathbb{N} \to \mathbb{N}$ tal que

$$H(n) := \text{Número de operações significativas que são efetuadas no algoritmo,}$$

em que $n \in \mathbb{N}$ é o tamanho da entrada de dados. O número de operações significativas é, em geral, para algoritmos numéricos, o número de operações de multiplicação e de comparações realizadas por esse algoritmo até que uma resposta seja obtida. Quando se efetua essas contagens e se soma, em geral, se obtêm um polinômio e como é conhecido, o termo de maior grau em um polinômio domina os demais termos quando a entrada (variável independente) tende ao infinito. Esse tipo de análise matemática em algoritmos recebe o nome de análise assintótica e se adota a notação grande $\mathcal{O}$("termo dominante") para indicar esse termo.

A segunda forma de análise de um algoritmo é do ponto de vista empírico e recebe o nome de *análise empírica de algoritmos*. Nesse caso, busca-se entender o algoritmo em estudo concretamente aplicado em uma linguagem de programação. Como em qualquer estudo empírico, a análise dos experimentos se dá por meio da Estatística e, nesse contexto, faz-se necessário estabelecer métricas e padronizações. Em relação as métricas, tem-se, por

exemplo, o tempo de execução de cada instância do algoritmo e as taxas de convergência empíricas relacionadas a elas. Quanto às padronizações, tem-se, por exemplo, a escolha de como as diferentes amostras para a entrada de dados no algoritmo serão geradas, quantas vezes será necessário rodar os experimentos para cada algoritmo e quais as técnicas de análise dentro da Estatística podem ser utilizadas para a avaliação dos resultados obtidos.

No contexto desta dissertação, opta-se por dar mais ênfase a via da análise empírica de algoritmos, apesar de na subseção 5.1 ser realizada uma discussão sobre os estudos desenvolvidos no capítulo 3 das taxas de convergência dos métodos de Newton-Raphson, gradiente descendente, gradiente acelerado e gradiente acelerado de alta ordem.

A análise empírica pode se mostrar muito reveladora em relação ao comportamento dos algoritmos. Uma das informações que podem ser reveladas por esse tipo de estudo é que a taxa de convergência empírica de um algoritmo para alguns tipos de dados pode ser competitiva em comparação com outros algoritmos que teoricamente tem uma taxa de convergência superior.

Na subseção 5.2, realiza-se a análise empírica dos métodos em estudo. Como discutido nos parágrafos acima, sobre a necessidade de se estabelecer métricas, adota-se duas. A primeira é a análise das taxas de convergência empírica dos métodos. A relevância desse estudo consiste em verificar concretamente se as taxas de convergência teóricas são encontradas na prática ou se há discrepâncias. Esses estudos são conduzidos na subseção 5.2.1.1 para o caso do modelo de regressão Logística.

A segunda métrica adotada é o tempo de execução de cada instância implementada. A relevância desta análise consiste em verificar se o tempo que um determinado algoritmo leva para executar uma instrução não é por demais elevado. Caso isso se verifique a eficiência do algoritmo estará comprometida mesmo se esse possuir uma boa taxa de convergência. Essas análises são conduzidas na subseção 5.2.1.2 e, mais uma vez, toma-se como estudo de caso o modelo de regressão Logística. O aspecto da padronização se realiza sobre a escolha das entradas de dados que serão utilizadas para o estudo das relações entre o tamanho e tipo de entrada e os tempos de execução e as taxas de convergência dos métodos em estudo.

5.1 Comparação matemática entre as taxas de convergência

Com base nos resultados obtidos no capítulo 3 e no capítulo 4, é realizado um estudo de comparação entre as taxas de convergência dos métodos apresentados nessa dissertação. Aqui, começamos a justificar a escolha de termos exposto as taxas de convergência ao longo do capítulo 3 com uma medida assintótica unificada, a saber, a notação $\mathcal{O}$ grande, pois isso possibilita comparar as ordens de convergência obtidas. Para que a exposição tenha um caráter estrutural e progressivo, fazemos esta introdução expondo os resultados na tabela abaixo

Métodos	Ordens de Convergência	Derivadas de f
Gradiente descendente	$\leqslant \mathcal{O}(1/k)$	∇f
Gradiente acelerado	$= \mathcal{O}(1/k^2)$	∇f
Gradiente acelerado de alta ordem	$\leqslant \mathcal{O}(1/k^2)$	∇f
Newton-Raphson	$\leqslant \mathcal{O}((L\eta/\rho)^{2^k})$	∇f e $\nabla^2 f$

Tabela 5.1 - Comparação entre as ordens de convergência dos métodos do gradiente descendente, gradiente acelerado, gradiente acelerado de alta ordem para uma derivada e Newton-Raphson.

É necessário alguns comentários sobre a taxa de convergência do método do gradiente acelerado de alta ordem apresentada na linha três da tabela 5.1. Para que as taxas de $\leqslant \mathcal{O}(1/k^2)$ seja obtida, o método também faz o uso de uma informação extra, a saber, a derivada segunda da função h que estrutura o espaço paramétrico em que se dá a convergência. A tabela 5.1 também possibilita observar o processo de aceleração com mais clareza, pelo menos em dois aspectos: do método do gradiente descendente para o gradiente acelerado, a reestruturação do método via introdução de uma sequência auxiliar permite um melhor aproveitamento da diferenciabilidade da função objetivo sem a necessidade de informações sobre a própria estrutura do espaço em que o processo de otimização está ocorrendo.

Quando se passa ao método do gradiente acelerado de alta ordem, a dependência estrutural se torna explícita na construção do próprio método, justamente na introdução da divergência de Bregman para a definição da lagrangiana de Bregman, pois esse é o ponto de partida para a dedução do método. No início do Capítulo 4 a função h, que define a divergência de Bregman, foi obtida de modo a representar parte da informação geométrica que emerge dos modelos lineares generalizados via família exponencial de distribuições.

Isso leva a considerar que a informação advinda da estrutura geométrica do espaço paramétrico via divergência de Bregman desempenha, de fato, um importante papel para o estabelecimento da taxa de convergência do método do gradiente acelerado de alta ordem.

No método de Newton-Raphson, isso ocorre de modo implícito, em sua variação, o método de Escore de Fisher. Pois, pode-se associar a matriz de informação de Fisher a uma métrica no espaço paramétrico que traz informação sobre a sua estrutura e, portanto, pode-se considerar que esse método também faz uso de informações estruturais para atingir sua elevada taxa de convergência.

5.2 Análise empírica de algoritmos

Nesta seção é realizado o principal trabalho empírico dessa dissertação em que é estudado a relação entre o tamanho da entrada de dados com as taxas de convergência empíricas e também o tempo de execução dos algoritmos, para isso é necessário delimitar de modo preciso qual é a forma da entrada de dados que os algoritmos que implementam os métodos em estudo demandam. Como essas implementações tomam como base a classe dos modelos lineares generalizados a entrada de dados é dada pelo par (X, y), em que X é a matriz modelo com dimensão $n \times p$ e y o vetor de variáveis resposta de dimensão n.

As variáveis explicativas $\{x_{ij}\}_{i=1,j=1}^{n,p}$ foram geradas de normais independentes com média 0 e variância 1 enquanto as variáveis $\{y_i\}_{i=1}^{n}$ são geradas via modelo de regressão Logística, ou seja, são variáveis Bernoulli. O algoritmo retorna uma estimativa $\hat{\beta}$ de dimensão p. A dimensão mais significativa para a análise do desempenho dos métodos em questão é a dimensão p relacionada ao vetor de parâmetros do modelo. Isso se deve ao fato de que, quanto maior é a dimensão do vetor paramétrico maior será o espaço em que o algoritmo deve buscar a estimativa $\hat{\beta}$. Em relação a dimensão n, tem-se que, quanto maior o seu valor maior será o volume de informação disponível para o processo de estimação, melhorando assim, a precisão do resultado. Em contrapartida as operações de multiplicação e inversão matriciais efetuadas pelo algoritmo tornam-se mais trabalhosas para valores de n muito altos.

Dado isso, tem-se que a escolha das dimensões da amostra se constitui um ponto crucial na análise e assim sendo o aumento das dimensões da amostra deve ser feito progressivamente ao longo dos experimentos de modo que as reações dos algoritmos a essas mudanças possam ser avaliadas. São considerados três valores para p, n e realiza-se as seguintes

combinações expostas na Tabela 5.2,

(p,n)	250	1000	5000
25	(25,250)	(25,1000)	(25,5000)
50	*	(50,1000)	(50,5000)
70	*	(70,1000)	(70,5000)

Tabela 5.2 - Dimensão das amostras para cada experimento.

em que cada par (p, n) dá a dimensão das amostras para a entrada de dados a serem utilizadas nos experimentos. Marca-se com $*$ as combinações que não são testadas, a saber, os pares $(50, 250)$ e $(70, 250)$. Isso se deve ao fato que 250 é uma amostra muito pequena para estimar vetores paramétricos de dimensão 50 e 70, o que compromete a qualidade das estimativas e consequentemente a sua utilidade para os testes de performance dos algoritmos. Os valores 25, 50 e 70 são pequenos quando pensados em termos dos campos de estudo como Aprendizagem de Máquina, mas para os fenômenos dentro do campo da Estatística em que os modelos lineares generalizados são utilizados esses valores são adequados.

Outro ponto que se faz necessário estabelecer é o número de vezes que cada experimento será replicado. A experiência durante os trabalhos empíricos mostrou que o valor de 100 réplicas é produtivo para se obter as informações necessárias para o tipo de análise empírica que é empreendida aqui.

Algumas palavras também são necessárias sobre os seguintes aspectos: número limite de iterações estabelecido para os algoritmos por execução, método de inicialização dos algoritmos adotados e a linguagem de programação escolhida para a implementação deles. Foi considerado o número de iterações limite igual a 2000. Esse número foi adotado pois em experimentos preliminares se constatou que, em muitos casos, o número mínimo de iterações que os métodos de primeira ordem demandavam até a convergência eram superiores a 1000 e, dessa maneira, considerar mais 1000 iterações como limite superior ainda está dentro dos critérios de razoabilidade que foram adotados.

O método de inicialização dos algoritmos que foi utilizado toma por base a versão de mínimos quadrados reponderados do método de Newton-Raphson. Essa versão é definida

por

$$\beta^{(m+1)} = \left(X^T W^{(m)} X\right)^{-1} X^T W^{(m)} z^{(m)}, \tag{5.1}$$

em que $z = \eta + W^{-\frac{1}{2}} V^{-\frac{1}{2}}(y - \mu)$. A variável z realiza a função de variável dependente modificada, enquanto W é uma matriz de pesos que muda a cada passo do processo iterativo. A matriz V é uma matriz diagonal das funções de variância dos modelos considerados. Normalmente, o método na equação (5.1) é inicializado com $\eta = G(y)$, em que η é o preditor linear, isto é, $\eta = (\mathbf{x}_1^T\beta, ..., \mathbf{x}_n^T\beta)$. Maiores informações sobre essa forma do método de Newton-Raphson pode ser encontrada em Paula (2013). Como se utiliza a equação (5.1) unicamente como inicializador, considera-se apenas o primeiro valor calculado para beta que nesse caso é dado por

$$\beta_0 = \left(X^T W X\right)^{-1} X^T W \eta.$$

Empiricamente foi verificado que esse procedimento de inicialização tem uma boa eficiência, pois gera um valor inicial na vizinhança do valor verdadeiro do parâmetro que foi utilizado para gerar as amostra de teste.

Outro ponto em que se faz necessário alguns esclarecimentos é em relação ao tamanho de passo ϵ utilizado nos métodos do gradiente descendente, gradiente acelerado e gradiente acelerado de alta ordem. Como poderá ser visto na tabela 5.3, os tamanhos dos passos variam de um método para o outro e estão em potências de 10. Isso se deve ao seguinte fato: os tamanhos dos passos foram escolhidos de modo a garantir a "melhor performance" dos métodos em questão e os valores que garantiam essa melhor performance foram determinados durante os procedimentos experimentais. Essa estratégia foi adotada em função da seguinte compreensão: os métodos devem ser julgados dentro de um contexto que permite a eles entregarem os melhores resultados que são capazes.

Por fim, tem-se a linguagem de programação que foi adotada para a implementação dos método e modelos em estudo nessa dissertação. A linguagem *Julia*[1] tem se tornado nos últimos anos uma referência quando se trata de linguagens de alta performance para computação científica (Sengupta, 2019). Essa linguagem consegue combinar a simplicidade sintática das linguagens *Python*[2] e *MATLAB*[3] com o alto desempenho em processamento de linguagens como *C* e *Fortran*. Apesar de ser recente, a linguagem Julia possui uma

[1] https://julialang.org/
[2] https://www.python.org/
[3] https://www.mathworks.com/products/matlab.html

ampla gama de pacotes para as mais variadas tarefas, como, por exemplo, os pacotes *Distributions* (que implementa distribuições clássicas de probabilidade) e *Statistics* (que implementa funções estatísticas) que foram amplamente usados para a codificação dos modelos em que se trabalha nessa subseção.

5.2.1 Modelo de regressão Logística

Nesta subseção é realizado, com base nos parâmetros estabelecidos na subseção anterior e na introdução desse capítulo, a análise empírica dos métodos do gradiente descendente, gradiente acelerado, gradiente acelerado de alta ordem para $m = 2$ e Newton-Raphson para a estimação do parâmetro β do preditor linear do modelo de regressão Logística com o objetivo de estudar as taxas de convergência empírica desses métodos.

Os resultados dos testes empíricos foram organizados na tabela 5.3 e estruturados com os seguintes campos: métodos, que podem assumir os valores: gradiente descendente (GD), gradiente acelerado (GA), gradiente acelerado de alta ordem para uma derivada (GAAO1) e Newton-Raphson (NR); tamanho do passo, assume como valores potências de 10; dimensão do vetor paramétrico, que assume os seguintes valores: 25, 50 e 70; estatísticas do número de iterações, em que se adota as seguintes estatísticas: mínimo, máximo, mediana, moda, média e desvio padrão. Por fim, tem-se o número de derivadas que podem assumir os valores 1 e 2. Além disso, divide-se a tabela em três seções para os tamanhos de amostras que foram utilizados: 250, 1000 e 5000.

5.2.1.1 Comparação empírica entre as taxas de convergência

Nesta subseção são analisados os dados resultantes dos testes empíricos realizados com os métodos de otimização em estudo aplicados ao modelo de regressão Logístico que foi apresentado na subseção 2.4.

Na tabela 5.3 se observa, para a dimensão 25 do vetor paramétrico e tamanho da amostra 250, entre os métodos de primeira ordem, que o método do gradiente acelerado de alta ordem apresenta a melhor performance em relação ao número de iterações até a convergência e, vale notar que esse ganho ocorre segundo todas as estatísticas. Em sequência se tem o método do gradiente acelerado que apresenta um aumento no número de iterações em relação ao gradiente acelerado de alta ordem. Essa diferença chega a 215 iterações para a estatística do máximo. Por fim, para essa classe de métodos, observa-se o

gradiente descendente com a performance mais fraca entre os três, principalmente quando se olha para as estatísticas do máximo e da média do número de iterações.

Nessa primeira análise, tendo por base a dimensão 25 do vetor paramétrico, tamanho de amostra 250 e as estatísticas adotadas, conclui-se que o método do gradiente acelerado de alta ordem para uma derivada da função objetivo apresenta desempenho superior a todos os demais métodos de primeira ordem em relação a taxa de convergência empírica.

Na sequência, para a amostra de tamanho 1000, é analisado em conjunto os comportamentos de todos os métodos durante as transições de dimensão do vetor paramétrico. Para a dimensão 25 nota-se um comportamento, para todos os métodos, semelhante ao caso anterior, tamanho de amostra 250, com a diferença que, como o tamanho da amostra aumentou, se tem uma maior disponibilidade de informação o que se traduz em uma redução no valor de todas as estatísticas e, em particular, um menor desvio padrão.

Para a dimensão 50, continuando com o mesmo tamanho de amostra, observa-se um comportamento consideravelmente distinto para os métodos de primeira ordem. Agora esses métodos, para a estatística do número máximo de iterações, não há convergência, pois atingem o limite superior de iterações permitidas nos experimentos que, como é conhecido, é igual a 2000. A estatística da moda para os métodos do gradiente descendente e gradiente acelerado também tem como valor 2000. Isso significa que, na maioria dos casos, entre as 100 repetições, esses métodos não conseguiram convergir. Uma das causas para isso é o aumento da dimensão do vetor paramétrico, o que torna a busca da estimativa mais difícil para o mesmo tamanho da amostra. Quanto ao método do gradiente acelerado de alta ordem tem-se que a moda é 1012 o que, em oposição aos dois casos anteriores, significa que, na maioria dos casos, esse método convergiu. Também se nota que a média e o desvio padrão do gradiente acelerado de alta ordem são menores em relação aos métodos do gradiente descendente e gradiente acelerado.

Para a dimensão 70 do vetor paramétrico e mesmo tamanho de amostra, observa-se uma piora das estatísticas para todos os métodos de primeira ordem com praticamente todos eles não atingindo a convergência em boa parte das réplicas. Mesmo nesse cenário, o método do gradiente acelerado de alta ordem apresenta o melhor desempenho entre esses métodos, principalmente, quando se olha para as estatísticas da média e do desvio padrão.

Por fim, tem-se a análise para o tamanho de amostra 5000. Com dimensão 25 para o vetor paramétrico, os métodos de primeira ordem tem o melhor desempenho dentre os três

tamanhos de amostra analisados até aqui. Isso decorre do fato que, para essa dimensão do vetor paramétrico, esse é o maior volume de dados disponível entre os três utilizados.

Para a dimensão do vetor paramétrico igual a 50, tem-se agora um comportamento distinto em relação a essa mesma dimensão para o tamanho de amostra 1000, exceto para o gradiente descendente, em que as estatísticas do número máximo de iterações e da moda atingem o limite superior de iterações permitidas para o algoritmo, que como é conhecido, tem o valor 2000. Para os outros dois casos dentro da classe de métodos de primeira ordem, observa-se que a convergência ocorre para um número relativamente baixo de iterações, para essa dimensão do vetor paramétrico, quando comparado aos resultados anteriores com amostras menores. Nessa classe de métodos se destaca, assim como nos casos anteriores, o método do gradiente acelerado de alta ordem com o melhor desempenho, como mostrado pelas estatísticas adotadas.

Por fim, para a dimensão do vetor paramétrico igual a 70, tem-se que os métodos de primeira ordem convergem, segundo todas as estatísticas. Dentre esses métodos o método do gradiente acelerado de alta ordem tem o melhor desempenho, com algumas estatísticas, como, por exemplo a média, mostrando uma diferença entre esse método e os demais métodos de primeira ordem que chega a 200 iterações.

Métodos	Tamanho do Passo (ϵ)	Dimensão do vetor (p)	Estatística - Número de Passos						Nº de Derivadas
			Mín	Máx	Mediana	Moda	Média	Desv.Pad.	
n = 250									
GD	$10^{-1.5}$	25	133	1300	368	199	404,1	208,97	1
GA	$10^{-1.73}$	25	115	1116	320	115	345,4	180,98	1
GAAO1	$10^{-1.2}$	25	95	900	258	110	283,4	144,38	1
NR	1	25	6	9	7	7	7,3	0,61	2
n = 1000									
GD	$10^{-2.17}$	25	149	353	220	229	224,8	42,36	1
GA	$10^{-2.4}$	25	124	301	177	210	181,8	39,92	1
GAAO1	$10^{-1.6}$	25	51	121	75	68	77,3	14,59	1
NR	1	25	6	7	7	7	6,8	0,37	2
GD	$10^{-1.8}$	50	639	2000	1171	2000	1283,9	407,98	1
GA	$10^{-1.98}$	50	508	2000	1016	2000	1113,9	427,24	1
GAAO1	$10^{-1.6}$	50	504	2000	901	1112	1013,2	359,39	1
NR	1	50	8	10	9	9	9,1	0,37	2
GD	$10^{-1.6}$	70	1216	2000	2000	2000	1987,5	81,87	1
GA	10^{-2}	70	892	2000	2000	2000	1865,0	236,91	1
GAAO1	$10^{-1.4}$	70	546	2000	1506	2000	1506,4	438,53	1
NR	1	70	9	13	10	10	10,1	0,72	2
n = 5000									
GD	$10^{-2.5}$	25	70	143	83	82	86,6	12,78	1
GA	$10^{-2.8}$	25	60	89	74	67	74,9	6,93	1
GAAO1	$10^{-2.2}$	25	39	78	46	47	47,8	6,10	1
NR	1	25	6	7	7	7	6,8	0,42	2
GD	$10^{-2.2}$	50	294	2000	2000	2000	1680,2	593,56	1
GA	$10^{-2.6}$	50	297	491	413	415	411,1	42,61	1
GAAO1	10^{-2}	50	160	272	225	217	225	23,62	1
NR	1	50	8	9	8	8	8,3	0,47	2
GD	$10^{-2.3}$	70	423	831	603	554	603,5	81,60	1
GA	$10^{-2.56}$	70	395	767	559	514	559,5	74,47	1
GAAO1	10^{-2}	70	235	456	334	346	333,3	44,21	1
NR	1	70	8	9	9	9	8,9	0,14	2

Tabela 5.3 - Estatísticas descritivas sobre o número de iterações até a convergência dos algoritmos considerando o modelo de regressão Logística para 3 tamanhos de amostras, $n \in \{250, 1000, 5000\}$ e para 3 tamanhos do vetor paramétrico, $p \in \{25, 50, 70\}$. Os resultados são baseados em 100 réplicas de Monte Carlo para cada combinação.

Com base em toda a análise que se procedeu para os métodos de otimização em estudo aplicados ao modelo de regressão Logística, conclui-se que, para todos os cenários testados, o método do gradiente acelerado de alta ordem apresenta, em relação a taxa de convergência empírica, um desempenho superior na classe dos métodos de primeira ordem.

Na próxima subseção é analisado os tempos de execução dos métodos de otimização

em estudo nessa dissertação. Essa análise é complementar ao estudo das taxas de convergência empíricas, pois ela revela se um algoritmo que executa uma tarefa em poucos passos também demanda pouco tempo para executar cada passo. Caso isso não se verifique a eficiência atribuída a um algoritmo, em função de ter uma taxa de convergência empírica reduzida, está comprometida, pois outros algoritmos que, apesar de demandarem mais passos até obterem a solução, o fazem em passos com tempo reduzido e terão globalmente uma eficiência maior.

5.2.1.2 Comparação empírica entre os tempos de execução

Nesta subseção é realizada a análise empírica para os tempos de execução dos métodos de otimização. Para que a consistência das análises fosse mantida, as mesmas estatísticas adotadas para o caso das taxas de convergência empírica foram utilizadas para a avaliação dos tempos de execução. Antes que os resultados obtidos sejam expostos se faz necessário comentar alguns aspectos técnicos que influenciam nas estatísticas quando se estuda os tempos de execução em qualquer algoritmo.

O primeiro fator é a habilidade do programador de ser capaz de implementar os algoritmos da forma mais eficiente. Isso depende da experiência do programador no desenvolvimento de algoritmos e do domínio dos recursos da linguagem escolhida para a implementação desses algoritmos. Na linguagem de programação Julia, por exemplo, se tem uma ampla gama de ferramentas disponíveis na própria linguagem para avaliar a performance dos algoritmos implementados nela. Duas dessas ferramentas são os macros *@time* e *@elapsed* que são utilizados para avaliar o tempo de execução de algoritmos. O macro *@time* também pode ser usado para avaliar o espaço de memória consumido pelo algoritmo. Ambos os macros retornam os tempos de execução em segundos com a diferença que o macro *@elapsed* suprime o valor de retorno do algoritmo avaliado enquanto que o macro *@time* retorna a solução obtida pelo algoritmo em conjunto com os tempos de execução e espaço de memória consumidos.

O segundo fator são as configurações das componentes físicas de um computador utilizado para os testes dos algoritmos. Esse aspecto tem forte influência sobre os tempos de execução, por exemplo, uma máquina que tenha grande capacidade de processamento paralelo permite uma redução considerável nos tempos de execução. Também se tem que a arquitetura do processador e o modo como os processos e as *Threads* são executados

influenciam na velocidade com que os algoritmos são capazes de executar a cada uma de suas instruções.

É comum na análise dos tempos de execução de um algoritmo não se usar configurações muito avançadas das componentes físicas de um computador, como por exemplo, grande capacidade de processamento paralelo. Isso ocorre pois a eficiência do algoritmo e de sua implementação está justamente em demandar o menor volume de recursos possíveis das componentes físicas de um computador. No próximo parágrafo se inicia a análise dos resultados obtidos para os tempos de execução.

Na tabela 5.4 se tem, para a dimensão 25 do vetor paramétrico e tamanho de amostra 250, que os métodos de primeira ordem apresentam tempos de execução próximos entre si segundo todas as estatísticas consideradas. Uma causa que pode ser considerada para esse comportamento é que, para dimensões baixas do vetor paramétrico e tamanho da amostra reduzido, o esforço computacional não difere significativamente nessa classe de métodos.

Passando para o tamanho de amostra 1000 é analisado em conjunto as três dimensões do vetor paramétrico com especial atenção ao comportamento dos métodos nas fases de transição entre as dimensões. Para a dimensão 25 do vetor paramétrico se tem, para os métodos de primeira ordem, um aumento em todos os tempos de execução. Um fator que influencia esse comportamento é o aumento das dimensões da matriz modelo o que faz com que as operações envolvendo essa matriz se tornem computacionalmente mais custosas. Mas mesmo com esse aumento no valor dos tempos de execução é possível observar que o método do gradiente acelerado de alta ordem tem uma redução perceptível nesses tempos em relação aos métodos do gradiente descendente e gradiente acelerado principalmente quando se observa as estatísticas da média e do desvio padrão.

Seguindo para a dimensão 50 do vetor paramétrico os resultados apresentados para os métodos de primeira ordem corroboram o que foi constatado em relação as taxas de convergência empíricas em que se observa a não convergência para os métodos do gradiente descendente e gradiente acelerado, em particular, quando considerado as estatísticas do máximo do número de iterações. As estatística do máximo do tempo de execução para esses métodos mostram valores acima de $1,0000$ segundos, o que evidencia a não convergência desses métodos quando comparado ao método do gradiente acelerado de alta ordem. Esse último apresenta estatísticas em linha com as apresentadas para a dimensão 25 do vetor paramétrico, excetuando, como esperado, que os valores são maiores em função da elevação

dos custos computacionais para as operações com uma matriz modelo de dimensão maior.

Por fim, conclui-se a análise para o tamanho de amostra 1000 observando que para a dimensão 70 do vetor paramétrico os métodos de primeira ordem apresentam valores elevados em todas as estatísticas. Deve ser destacado que os valores dessas estatísticas para os métodos do gradiente descendente e gradiente acelerado são inferiores aos apresentados pelo método do gradiente acelerado de alta ordem. Um dos fatores que justifica esse resultado é que esse método executa mais operações por iteração do que os métodos do gradiente descendente e gradiente acelerado. Somando-se a isso o aumento da dimensão do vetor paramétrico se tem como resultado um maior custo computacional e consequentemente um aumento nos valores das estatísticas dos tempos de execução.

Para o tamanho da amostra 5000 se tem que, na dimensão 25 do vetor paramétrico, os métodos de primeira ordem têm, segundo todas as estatísticas, tempos de execução com valores relativamente baixos. Pode-se destacar que as estatísticas da média e do desvio padrão apresentam valores consideravelmente próximos entre esses métodos.

Para a dimensão 50 se tem, para os métodos de primeira ordem, que as estatísticas assumem valores muito elevados, e em particular, para o método do gradiente descendente, as estatísticas do máximo, mediana e média apresentam os valores mais elevados. Os métodos do gradiente acelerado e gradiente acelerado de alta ordem, apesar de também apresentarem valores elevados para todas as estatísticas, esses são consideravelmente menores do que os apresentados pelo método do gradiente descendente.

Por fim, para a dimensão 70 do vetor paramétrico se observa, para os métodos de primeira ordem, valores elevados para as estatísticas dos métodos do gradiente descendente e gradiente acelerado com os tempos de execução para esses dois métodos sendo relativamente próximos. Nesse caso se destaca o método do gradiente acelerado de alta ordem que apresentar os menores valores em todas as estatísticas quando comparado aos demais métodos de primeira ordem.

Métodos	Tamanho do Passo (ϵ)	Dimensão do vetor (p)	Estatística - Tempo de Execução (segundos)						Nº de Derivadas
			Mín	Máx	Mediana	Moda	Média	Desv.Pad.	
n = 250									
GD	$10^{-1.5}$	25	0,0124	0,1252	0,0354	0,0285	0,0386	0,0200	1
GA	$10^{-1.73}$	25	0,0115	0,1792	0,032	0,0247	0,0377	0,0245	1
GAAO1	$10^{-1.2}$	25	0,012	0,1161	0,033	0,0313	0,0370	0,0190	1
NR	1	25	0,0021	0,0250	0,0078	0,0031	0,0086	0,0052	2
n = 1000									
GD	$10^{-2.17}$	25	0,0869	0,1218	0,1406	0,1218	0,1428	0,0288	1
GA	$10^{-2.4}$	25	0,0800	0,4798	0,1250	0,1011	0,1309	0,0438	1
GAAO1	$10^{-1.6}$	25	0,0450	0,1200	0,071	0,0609	0,0734	0,0150	1
NR	1	25	0,0070	0,0213	0,0119	0,0086	0,0121	0,0028	2
GD	$10^{-1.8}$	50	0,4683	1,9360	0,8300	1,8500	0,9320	0,3298	1
GA	$10^{-1.98}$	50	0,3625	1,8650	0,7700	1,7740	0,8621	0,3567	1
GAAO1	$10^{-1.6}$	50	0,1149	0,2290	0,1348	0,1193	0,1348	0,0075	1
NR	1	50	0,0126	0,0777	0,0247	0,0300	0,0256	0,0106	2
GD	$10^{-1.6}$	70	0,8737	1,8810	1,5416	1,7711	1,5480	0,1022	1
GA	10^{-2}	70	0,8096	2,5900	1,5503	2,325	1,5435	0,2937	1
GAAO1	$10^{-1.4}$	70	0,7295	3,5344	2,0642	2,0771	2,0730	0,6877	1
NR	1	70	0,0398	0,0870	0,0565	0,0531	0,0557	0,0096	2
n = 5000									
GD	$10^{-2.5}$	25	0,2706	0,6227	0,3054	0,6227	0,3234	0,0562	1
GA	$10^{-2.8}$	25	0,2665	0,7844	0,5064	0,7844	0,3173	0,0598	1
GAAO1	$10^{-2.2}$	25	0,2323	0,5983	0,3093	0,2871	0,3314	0,0697	1
NR	1	25	0,0250	0,0563	0,0341	0,0387	0,0348	0,0056	2
GD	$10^{-2.2}$	50	1,2762	9,3408	6,0890	1,4433	5,2871	1,7445	1
GA	$10^{-2.6}$	50	1,2209	2,0913	1,5995	1,7781	1,6166	0,1840	1
GAAO1	10^{-2}	50	1,0535	1,7645	1,2185	1,2326	1,2510	0,1157	1
NR	1	50	0,0250	0,0565	0,0341	0,0387	0,0348	0,0056	2
GD	$10^{-2.3}$	70	1,7874	3,0959	2,3943	3,0177	2,3830	0,2689	1
GA	$10^{-2.56}$	70	1,8322	3,2180	2,4304	2,5740	2,4594	0,2875	1
GAAO1	10^{-2}	70	0,8198	0,3741	0,7054	0,4285	0,6683	0,1157	1
NR	1	70	0,1812	0,2827	0,2021	0,2827	0,2048	0,0135	2

Tabela 5.4 - Estatísticas descritivas sobre os tempos de execução até a convergência dos algoritmos considerando o modelo de regressão Logístico para 3 tamanhos de amostras, $n \in \{250, 1000, 5000\}$ e para 3 tamanhos do vetor paramétrico, $p \in \{25, 50, 70\}$. Os resultados são baseados em 100 réplicas de Monte Carlo para cada combinação.

Com base nas análises empreendidas nos parágrafos anteriores se tem que os métodos de primeira ordem demandam um tempo maior até a convergência que os métodos de segunda ordem. Na classe dos métodos de primeira ordem todos apresentam, em geral, tempos de execução relativamente próximos entre si para cada tamanho de amostra e cada dimensão do vetor paramétrico tendo o método do gradiente acelerado de alta ordem uma

pequena vantagem em relação aos demais.

Capítulo 6

Conclusão

No estudo desenvolvido ao longo dessa dissertação, a interseção entre o campo da Otimização, através dos métodos do gradiente descendente, gradiente acelerado, gradiente acelerado de alta ordem e Newton-Raphson, com o campo da Estatística, através dos modelos lineares generalizados foi explorada. As taxas de convergência teóricas dos métodos de otimização foram demonstradas e a sua conexão com a teoria das equações diferenciais ordinárias foi articulada de modo que os métodos de otimização em estudo foram interpretados como técnicas de discretização que possibilitam que as taxas de convergência das curvas solução da EDO associada sejam compatíveis com as taxas de convergência das sequências que discretizam essas curvas.

Esse estudo teórico mostrou que o método de Newton-Raphson apresenta uma taxa de convergência exponencial e os demais métodos apresentam taxas de convergência polinomiais com o método do gradiente acelerado de alta ordem com informação de uma derivada da função objetivo tendo entre os métodos polinomiais a melhor taxa de convergência, $\leq \mathcal{O}(1/k^2)$. Dessa compreensão global foi articulado a aplicação desses métodos para a estimação do parâmetro do preditor linear nos modelos lineares generalizados, em particular, no modelo de regressão Logística.

Esses procedimentos foram implementados na linguagem de programação Julia e em uma análise empírica foi constatado que o método do gradiente acelerado de alta ordem apresenta uma taxa de convergência empírica superior aos demais métodos de primeira ordem e um tempo de execução, que na maioria dos casos, é competitivo com os tempos de execução do demais métodos de primeira ordem. Isso abre a possibilidade de investigações analíticas futuras sobre esse método que tem, nas condições exigidas pelo estudo aqui realizado, uma taxa de convergência superior o que pode resultar em um método mais

eficiente para a estimação de parâmetros não só nos modelos lineares generalizados como em outras classes de modelos estatísticos. Para que esse objetivo futuro seja realizado se faz necessário investigar o método do gradiente acelerado de alta ordem em pelo menos duas direções. A primeira, no campo teórico, é a possibilidade de se obter uma versão exponencial para o caso de uma derivada da função objetivo. Nesse caso se faz necessário um estudo para saber se para atingir essa taxa exponencial deve se impor mais restrições ao negativo do logaritmo da função de verossimilhança além de ser convexo, como por exemplo, ser fortemente convexo. Em caso afirmativo, o campo estatístico da análise assintótica pode se revelar uma ferramenta crucial ao estudar as propriedades assintóticas na estimação de parâmetros via teoria da verossimilhança.

Para a segunda linha de análise é necessário um estudo de Monte Carlo sobre o tamanho de passo a ser adotado para esse método. No trabalho aqui realizado o tamanho dos passos utilizados para cada dimensão do vetor paramétrico e cada tamanho de amostra decorreu das experiências acumuladas no processo de desenvolvimento e teste dos algoritmos. Portanto, uma investigação mais profunda e sistemática se torna fundamental para que os avanços nessas pesquisas futuras possam ser disponibilizados para a comunidade acadêmica na forma de novos pacotes implementados nas linguagens mais utilizadas no campo da Estatística e da Otimização como, por exemplo, as linguagens Python, R[1] e Julia.

[1] https://www.r-project.org/

Referências Bibliográficas

Armijo L., Minimization of functions having Lipschitz continuous first partial derivatives, Pacific Journal of Mathematics, 1966, vol. 16, p. 1

Baumaister J., Leitão A., Introdução à Teoria de Controle e Programação Dinâmica 1 edn. IMPA Rio de Janeiro, Brasil, 2014

Bierlaire M., Optimization: Principles and Algorithms 2 edn. EPFL Press Lausanne, Switzerland, 2018

Brown L. D., Fundamentals of Statistical Exponential Families. Institute of Mathematical Statistics, 1986

Burden R., Faires D., Numerical Analysis 10 edn. Cengage Learning, 2015

C.Frery A., Cribari-Neto F., Elementos de Estatística Computacional Usando Plataformas de Software Livre/Gratuito 2 edn. IMPA Rio de Janeiro, Brasil, 2011

Dobson A. J., An Introduction to Generalized Linear Models 2 edn. Chapman & Hall/CRC Boca Raton, EUA, 2002

Gower R. M., Convergence theorems for gradient descent, 2019

Izmailov A., Solodov M., Otimização - volume 2: Métodos Computacionais 2 edn. IMPA Rio de Janeiro, Brasil, 2012

McCullagh P., Nelder J. A., Generalized Linear Models 2 edn. Chapman & Hall/CRC Boca Raton, EUA, 1989

Nelder J. A., Wedderburn R. W. M., Generalized linear models, Journal of the Royal Statistical Society, Series A, 1972, vol. 135, p. 370

Nesterov Y., A method of solving a convex programming problem with convergence rate $O(1/k^2)$, Soviet Mathematics Doklady, 1983, vol. 27, p. 372

Nesterov Y., Introductory Lectures on Convex Optimization: A Basic Course 1 edn. vol. 87, Springer, 2004

Paula G. A., Modelos de Regressão com Apoio Computacional 2 edn. IME-USP São Paulo, Brasil, 2013

Sengupta A., Julia High Performance 2 edn. Packt Publishing, 2019

Su W., Boyd S., Candès E. J., A differential equation for modeling Nesterov's accelerated gradient method: Theory and insights, Journal of Machine Learning Research, 2016, vol. 17, p. 1

Wibisono A., Wilson A. C., Jordan M. I., A variational perspective on accelerated methods in optimization, Proceedings of the National Academy of Sciences, 2016, vol. 113, p. E7351

Wolfe P., Convergence conditions for ascent methods, SIAM Review, 1969, vol. 11, p. 226