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

Calixto, Anderson de Oliveira

Métodos de Otimização Aplicados à Estatística/Anderson de Oliveira Calixto - Rio de Janeiro: UFRJ/IM, 2020.

vi, 111f.: il.; 31cm.

Orientadores: Ralph dos Santos Silva e Heudson Tosta Mirandola

Dissertação (mestrado) - UFRJ/IM/ Programa de Pós-graduação em Estatística, 2020.

Referências Bibliográficas:f.83-87.

1. Algoritmos 2. Estimação de Parâmetros. 3. Modelos Lineares Generalizados. I. Silva, Ralph S. e Mirandola, Heudson T. II. Universidade Federal do Rio de Janeiro, Instituto de Matemática. III. Título.

*Dedico esse trabalho à minha mãe 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.*

# Agradecimentos

Agradeço à minha mãe Margarida, por todos os ensinamentos, força e encorajamento. Minha eterna gratidão e admiração. À Deus que me deu força, equilíbrio e sabedoria até a finalização deste estudo. Sem essa luz tudo seria mais difícil. À minha irmã Aline, pelo carinho de sua amizade, por todo incentivo e por dividir as alegrias e dificuldades da universidade comigo. Aos meus orientadores, professores Ralph dos Santos Silva e Heudson Tosta Mirandola, pelo apoio, paciência e dedicação na construção de cada etapa deste trabalho e, principalmente, pela generosidade e amizade. Contem sempre comigo. Aos professores do DME-IM - Departamento de Matemática e Estatística do Instituto de Matemática, pelos ensinamentos 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, se realiza um estudo de caso na área dos métodos de otimização, através dos algoritmos do gradiente descendente, gradiente acelerado, gradiente acelerado de alta ordem e Newton-Raphson e é estudado as taxas de convergência teóricas das sequências numéricas geradas por esses algoritmos. Para a segunda via, se realiza um estudo de caso na área dos modelos estatísticos, através 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. Através da análise empírico-estatístico dos métodos de otimização em estudo, chega-se a conclusão que, o método do gradiente acelerado de alta ordem, quando faz o uso da informação das duas derivadas da função objetivo, tem uma performance empírica competitiva em relação aos demais métodos 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 dissertation, we study an intersection between the areas of optimization methods and the statistical models. We work in this intersection in two ways: In the first one, 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. For the second way, we take as a case study, in the area of statistical models, the generalized linear models and studies their estimation process via the likelihood function logarithm. 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 the two derivatives of the objective function, has an empirical performance superior to the other methods 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 and thus opening the possibility of future work to an analytical investigation of this phenomenon.

**Keywords:** Algorithms, Parameter Estimation, Generalized Linear Models

# Lista de Figuras

# Lista de Tabelas

# Sumário

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 baseiam-se 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 de $\mathcal{O}(1/k)$ (ver Gower, 2019, e referências). Ao longo das ultimas 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 de $\mathcal{O}(1/k^2)$. Esse método ficou conhecido como *método do gradiente acelerado* ou *método de Nesterov*. Além disso, Nesterov também mostrou que a taxa ótima para métodos que utilizam somente a derivada primeira é de $\mathcal{O}(1/k^2)$ (ver Nesterov, 2004).

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 de primeira ordem. 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 acelerado de alta ordem*. Esse algoritmo pode ser particularizado para qualquer ordem (primeira, segunda, etc). Não obstante, a parte empírica desta dissertação foca principalmente nos casos de primeira e segunda ordem.

O método de Newton-Raphson, quando utilizado no contexto da otimização, é um método de segunda ordem que sob determinadas condições tem uma taxa de convergência elevada. Por exemplo: se $g : \mathbb{R}^p \longrightarrow \mathbb{R}^p$ é uma função duas vezes continuamente diferenciável tal que a sua matriz Jacobiana seja não nula e inversível; $\beta_* \in \mathbb{R}^p$ seja um ponto tal que $g(\beta_*) = 0$ e $\{\beta_k\}_{k\geqslant 0}$ seja a sequência gerada pelo método de Newton-Raphson, então a sequência $\{g(\beta_k)\}_{k\geqslant 0}$ tem taxa de convergência da ordem de

$$||g(\beta_k) - g(\beta^*)|| \leqslant \mathcal{O}\big(\eta^{2^k}\big),$$

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

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 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 concentra-se 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- gradiente acelerado de alta ordem para duas derivadas e 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 faz-se o uso somente dos casos de primeira e segunda ordens.

## *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 e segunda ordem. Portanto, outra meta desta dissertação também é comparar esses métodos de primeira ordem (que não utilizam derivadas segundas) do 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: 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 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. Então, apresenta-se provas teóricas de duas formas.

Ainda, um outro objetivo dessa 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 a dedução do método de Newton-Raphon (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 da sequência do algoritmo e através das equações diferenciais ordinárias associadas a eles. No Capítulo 4 aplica-se o método do gradiente acelerado de alta ordem para os casos de uma e duas derivadas aos modelos lineares generalizados de modo a obter-se para esses modelos a forma explícita desse algoritmo. No

Capítulo 5 faz-se um estudo de Monte Carlo para a comparação dos diversos algoritmos de otimização apresentados e estes 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, em vários desses fenômenos não é possível assumir que a variável resposta seja normalmente distribuída, e portanto, faz-se necessário realizar alguma transformação nos dados observados afim de obter-se uma aproximação normal dos mesmos.

Infelizmente, isso leva a uma importante perda na capacidade de interpretação do analista em relação aos 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. 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 modelos lineares generalizados e algumas de suas propriedades. Ademais, aborda-se o modelo de regressão logístico.

## *2.1 Definição e propriedades*

Os modelos lineares generalizados permitem que a distribuição da variável resposta possa pertencer a família exponencial de distribuições. Essa classe contém distribuições clássicas como, por exemplo, a normal, exponencial, Bernoulli, binomial, gama, dentre outras. A definição formal é dada a seguir.

**Definição 1.** *Considere $Y_1, ..., Y_n$ variáveis aleatórias independentes cada uma com função*

*de densidade ou de probabilidade dada por,*

$$p(y_i|\theta_i, \phi) = \exp\{\phi[y_i\theta_i - b(\theta_i)] + c(y_i, \phi)\},$$

*em que* $b : \mathbb{R} \longrightarrow \mathbb{R}$ *é uma função convexa e diferenciável tal que* $b' : \mathbb{R} \longrightarrow \mathbb{R}$ *é um difeomorfismo. Além disso,* $\phi > 0$ *é o parâmetro de dispersão,* $\theta_i$ *é o parâmetro canônico da distribuição e* $c : \mathbb{R} \times \mathbb{R}_+ \longrightarrow \mathbb{R}$ *é uma função da amostra e da dispersão.*

*Seja* $X$ *uma matriz* $n \times p$ *com posto* $p$ *e* $x_i^T = (x_{i1}, ..., x_{ip})$ *a* $i$*-ésima linha da matriz* $X$*. Por fim, seja* $\beta^T = (\beta_1, ..., \beta_p)$ *um vetor de parâmetros do modelo e* $G : \mathbb{R} \longrightarrow \mathbb{R}$ *um difeomorfismo chamado de função de ligação. O modelo linear generalizado é definido por:*

$$\begin{aligned} p(y_i|\theta_i, \phi) &= \exp\{\phi\{y_i\theta_i - b(\theta_i)\} + c(y_i, \phi)\}, \\ G(\mu_i) &= x_i^T\beta, \\ b'(\theta_i) &= E(Y_i) = \mu_i. \end{aligned} \tag{2.1}$$

A função $b : \mathbb{R} \longrightarrow \mathbb{R}$ é conhecida como função de partição e desempenha um papel importante na classe dos modelos lineares generalizados e também em diversas outras teorias estatísticas que tomam como base a família exponencial de distribuições. O difeomorfismo $b' : \mathbb{R} \longrightarrow \mathbb{R}$ possibilita uma mudança de coordenadas entre o espaço dos parâmetros canônicos do modelo na forma da família exponencial - o espaço dos $\theta_i \in \mathbb{R}$ - e o espaço dos parâmetros naturais - o espaço dos $\mu_i \in \mathbb{R}$. Maiores informações e detalhes sobre os modelos lineares generalizados podem ser encontrados em McCullagh e Nelder (1989) e Dobson (2002).

O objetivo a longo das próximas seções é estimar o vetor de parâmetros $\beta$ que em conjunto com a matriz do modelo e a função de ligação, estrutura o vetor de médias $\mu^T = (\mu_1, ..., \mu_n)$ através da seguinte relação funcional:

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

Considera-se que o parâmetro de dispersão $\phi > 0$ seja conhecido, pois o objetivo é estimar o vetor paramétrico $\beta$. Esse procedimento é realizado pelo método de Newton-Raphson em diversos programas de análise de dados como, por exemplo, o `R` (R Core Team, 2019).

A função de ligação, $G : \mathbb{R} \longrightarrow \mathbb{R}$, desempenha um papel de destaque na estimação de parâmetros na classe dos modelos lineares generalizados. Isso deve-se 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 côncava. Essa propriedade garante que o estimador de máxima verossimilhança obtido seja único - quando esse existir. Uma classe de funções de ligação importante é das ligações canônicas que garantem que logaritmo da função de verossimilhança seja côncava.

## 2.2 Funções de ligação canônicas

Seja $\phi > 0$ conhecido. Então, o logaritmo da função de verossimilhança do modelo linear generalizado definido na Equação (2.1) é dada por

$$\mathcal{L}(\beta|X,y) = \sum_{i=1}^{n} \phi\big\{y_i\theta_i - b(\theta_i)\big\} + \sum_{i=1}^{n} c(y_i,\phi). \tag{2.2}$$

Supondo que o parâmetro canônico é igual ao preditor linear, $\theta_i = x_i^T\beta$, tem-se que a Equação (2.2) pode ser reescrita como

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

Definindo a estatística $S^T = \phi\sum_{i=1}^{n} y_i x_i^T$, a Equação (2.3) pode ser expressa por

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

Então, pelo teorema da fatoração a estatística $S^T$ é suficiente para o vetor $\beta$. As funções de ligação associadas a tais estatísticas são chamadas de ligações canônicas. O lema a seguir formaliza a afirmação que sob a hipótese da função de ligação ser canônica o logaritmo da função de verossimilhança dada pela Equação (2.2) é côncava.

**Lema 2.** *Seja $\mathcal{L}:\mathbb{R}^p \longrightarrow \mathbb{R}$ o logaritmo da função de verossimilhança para o modelo linear generalizado dado pela Equação (2.2). Se a função de ligação $G:\mathbb{R}\longrightarrow\mathbb{R}$ é uma ligação canônica, então o logaritmo da função de verossimilhança é côncava.*

*Demonstração.* O logaritmo da função de verossimilhança $\mathcal{L}:\mathbb{R}^p\longrightarrow\mathbb{R}$ para o modelo dado pela Equação (2.1), sob a hipótese da função de ligação ser canônica e $\phi > 0$ ser conhecido, é dada pela Equação (2.4). Como $b:\mathbb{R}\longrightarrow\mathbb{R}$ é uma função convexa e a aplicação $\beta \longmapsto x_i^T\beta$ é linear, tem-se que a aplicação $\beta\longmapsto b(x_i^T\beta)$ é convexa. A soma de funções convexas é uma função convexa. Logo, a aplicação $\beta\longmapsto\phi\sum_{i=1}^{n} b(x_i^T\beta)$ é convexa.

Considere $\hat{\beta} \in \mathbb{R}^p$, $\bar{\beta} \in \mathbb{R}^p$ e $\alpha \in (0,1)$. Tem-se que

$$\mathcal{L}(\alpha\hat{\beta} + (1-\alpha)\bar{\beta}|X,y) = S^T[\alpha\hat{\beta} + (1-\alpha)\bar{\beta}] - \phi\sum_{i=1}^{n} b(x_i^T[\alpha\hat{\beta} + (1-\alpha)\bar{\beta}]) + \sum_{i=1}^{n} c(y_i,\phi).$$

Agora, dado que $\beta \longmapsto S^T\beta$ é uma aplicação linear, segue-se que

$$S^T[\alpha\hat{\beta} + (1-\alpha)\bar{\beta}] = \alpha S^T\hat{\beta} + (1-\alpha)S^T\bar{\beta}.$$

Como a aplicação $\beta \longmapsto -\phi\sum_{i=1}^{n} b(x_i^T\beta)$ é côncava, tem-se que

$$-\phi\sum_{i=1}^{n} b(x_i^T[\alpha\hat{\beta} + (1-\alpha)\bar{\beta}]) \geqslant \alpha\left[-\phi\sum_{i=1}^{n} b(x_i^T\hat{\beta})\right] + (1-\alpha)\left[-\phi\sum_{i=1}^{n} b(x_i^T\bar{\beta})\right],$$

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

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

obtem-se que

$$\begin{aligned}
\mathcal{L}(\alpha\hat{\beta} + (1-\alpha)\bar{\beta}|X,y) &\geqslant \alpha\left[S^T\hat{\beta} - \phi\sum_{i=1}^{n} b(x_i^T\hat{\beta}) + \sum_{i=1}^{n} c(y_i,\phi)\right] \\
&+ (1-\alpha)\left[S^T\bar{\beta} - \phi\sum_{i=1}^{n} b(x_i^T\bar{\beta}) + \sum_{i=1}^{n} c(y_i,\phi)\right] \\
&= \alpha\mathcal{L}(\hat{\beta}|X,y) + (1-\alpha)\mathcal{L}(\bar{\beta}|X,y).
\end{aligned}$$

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

A seguir discute-se o conceito de informação de Fisher e como essa se aplica aos modelos lineares generalizados.

## 2.3 Informação de Fisher

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

$$\mathcal{L}(\theta|\mathcal{D}) = \log(p(\mathcal{D}|\theta)). \tag{2.5}$$

Se a função definida pela Equação (2.5) for duas vezes diferenciável, então defini-se a *matriz de informação de Fisher* por

$$\mathcal{I}_{\mathcal{D}}(\theta) = E_{\mathcal{D}}\left(-\nabla^2 \mathcal{L}(\theta|\mathcal{D})\right), \tag{2.6}$$

em que $E_{\mathcal{D}}(.)$ é o operador esperança para o modelo $p(.|\theta)$. A matriz de informação de Fisher nos diz sobre o quanto de informação uma amostra $\mathcal{D}$ contém sobre o parâmetro desconhecido $\theta$. No caso dos modelos lineares generalizados o conjunto $\mathcal{D}$ reduz-se 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 o dado pela Equação (2.4) para as funções de ligação canônicas.

A seguir, revisa-se um caso particular de modelo linear generalizado, a saber, o modelo de regressão logístico.

## 2.4 Modelo de regressão logística

Dentro da classe dos modelos lineares generalizados, o modelo de regressão logístico possui uma posição de destaque e tem sido amplamente empregado ao longo dos últimos anos. Isso deve-se 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 $X$, deseja-se 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 $\pi$ que representa a probabilidade de ocorrência de um fenômeno aleatório.

Como visto na Seção 2.1, assume-se que o parâmetro $\pi$ da distribuição Bernoulli tem relação com os dados da matriz modelo através da função de ligação $G$ - um difeomorfismo - e um vetor $\beta$ cujo valor deve ser estimado. Na definição abaixo, apresenta-se a formulação do modelo de regressão logístico pressupondo que a função de ligação $G : \mathbb{R} \longrightarrow \mathbb{R}$ é canônica.

**Definição 3.** *Considere $Y_1, ..., Y_n$ variáveis aleatórias independentes cada uma com função de probabilidade dada por*

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

*Seja $X$ uma matriz $n \times p$ com posto $p$ e $x_i^T = (x_{i1}, ..., x_{ip})$ a $i$-ésima linha da matriz $X$. Por fim, seja $\beta^T = (\beta_1, ..., \beta_p)$ um vetor de parâmetros do modelo e $G : \mathbb{R} \longrightarrow \mathbb{R}$ a função de ligação canônica. O modelo de regressão logístico é definido por*

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

A função de partição $b : \mathbb{R} \longrightarrow (0, +\infty)$ para o modelo de regressão logístico assume a forma simples da composição de uma função exponencial com a função logarítmica no espaço dos parâmetros canônicos, isto é,

$$b(\theta_i) = \log(1 + \exp\{\theta_i\}), \tag{2.7}$$

e a sua derivada, $b' : \mathbb{R} \longrightarrow (0, 1)$, que é um difeomorfismo, assume a forma,

$$b'(\theta_i) = \frac{\exp\{\theta_i\}}{1+\exp\{\theta_i\}}. \tag{2.8}$$

Além disso, assume-se ligação canônica no modelo de regressão logístico e, portanto, a função de ligação $G : \mathbb{R} \longrightarrow \mathbb{R}$ toma a forma do difeomorfismo inverso $(b')^{-1} : (0, 1) \longrightarrow \mathbb{R}$ que é dado por,

$$G(\pi_i) = \left(b'\right)^{-1}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right). \tag{2.9}$$

Com base nas Equações (2.7), (2.8) e (2.9), o logaritmo da função de verossimilhança dada pela Equação (2.4) para o modelo de regressão logístico pode ser reescrita por

$$\mathcal{L}_{ML}(\beta|X, y) = S^T\beta - \sum_{i=1}^{n} \log(1 + \exp\{x_i^T\beta\}), \quad \text{com} \quad S^T = \sum\nolimits_{i=1}^{n} y_i x_i^T. \tag{2.10}$$

Note que para o modelo de regressão logístico o parâmetro de dispersão é constante, $\phi = 1$, e que o logaritmo da função de verossimilhança assume uma forma simples. Com base nesse modelo é realizado no Capítulo 5 uma análise estatística de métodos dos otimização a serem discutidos no Capítulo 3 a seguir.

Capítulo 3

# Métodos de Otimização

Em diversas áreas - por exemplo, na Estatística ou nas Engenharias - existem problemas que podem ser resolvidos através de um processo de otimização. Considera-se que o fenômeno ou problema em questão seja definido ou modelado por uma função - por exemplo, $f : \mathbb{R}^p \longrightarrow \mathbb{R}$ - cuja dinâmica do fenômeno ou solução do problema seja obtida quando minimizamos (ou maximizamos) a função $f$ em um subconjunto ou no próprio $\mathbb{R}^p$. Para isso, 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 das soluções viáveis para o problema, ou seja, é o conjunto em que busca-se 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 $\beta_*$ de modo analítico. Portanto, faz-se necessário a utilização de algum procedimento numérico para a obtenção de $\beta_*$ - sempre de forma aproximada. A resolução do problema dado pela Equação (3.1) via métodos numéricos é obtida através da construção de uma sequência $\{\beta_k\}_{k \geqslant 0}$ tal que

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

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

$$f(\beta_k) - f(\beta_*) \leqslant \mathcal{O}\big(F(k|\epsilon)\big),$$

em que $F$ é uma função da iteração $k$ e do tamanho de passo $\epsilon > 0$. Como os métodos numéricos devem ser implementado em alguma linguagem de programação, é possível fazer comparações empíricas baseadas na linguagem escolhida. 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 se expõe sobre o método do gradiente descendente e mostra-se que existe um equação diferencial ordinária (EDO) de primeira ordem que pode ser associada esse método de modo natural. Demonstra-se, então, as taxas de convergência do método do gradiente descendente, $\mathcal{O}\big(1/\epsilon k\big)$, e da EDO de primeira ordem associada a ele, $\mathcal{O}\big(1/t\big)$. Observa-se que as taxas de convergência podem ser associadas através 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 $t \longmapsto f(\beta(t))$, em que $t \longmapsto \beta(t)$ é a curva solução da EDO de primeira ordem associada ao método do gradiente descendente, e de sua discretização $k \longmapsto f(\beta_k)$.

Em seguida, na Seção 3.3 introduz-se uma sequência auxiliar no método do gradiente descendente. Isso permite que a taxa de convergência passe de $\mathcal{O}\big(1/\epsilon k\big)$ para $\mathcal{O}\big(1/\epsilon k^2\big)$ 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 $t \longmapsto f(\beta(t))$, em que $t \longmapsto \beta(t)$ é agora a curva solução da EDO de segunda ordem associada ao método de Nesterov, obtem-se a estimativa $\mathcal{O}\big(1/t^2\big)$. Ao se identificar $t = \sqrt{\epsilon}k$, nota-se mais uma vez que o método em questão poder 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 aborda-se 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, que 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 Su et al. (2016):

$$f(\beta_1, \beta_2) = 2 \times 10^{-2}\beta_1^2 + 5 \times 10^{-3}\beta_2^2. \tag{3.2}$$

Além disso, toma-se o domínio da Equação (3.2) como $[-2\,;\,2] \times [-2\,;\,2]$.

## 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 puder ser transformada em uma função diferenciável que atende a um certo conjunto de propriedades, então o método de Newton-Raphson pode ser utilizado para a obtenção da solução.

### 3.1.1 Construção

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

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

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

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

em que $\beta_* - \beta_0 = h$. Agora, suponha que a inversa da matriz Jacobiana, $\nabla g$, exista no ponto $\beta_0$. Logo, pode-se aproximar o valor de $h$ por

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

A partir daí, defini-se uma nova estimativa para o valor de $\beta_*$ como

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

em que $\beta_1$ é uma segunda aproximação para $\beta_*$ mais precisa que $\beta_0$. Ao repetir-se esse procedimento diversas vezes, constrói-se o seguinte processo iterativo:

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

O processo iterativo dado pela Equação (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 o caminho da sequência segue uma linha reta em direção ao ponto de mínimo. Uma possível explicação que pode justificar 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. Essa convergência ocorre em apenas 2 passos.

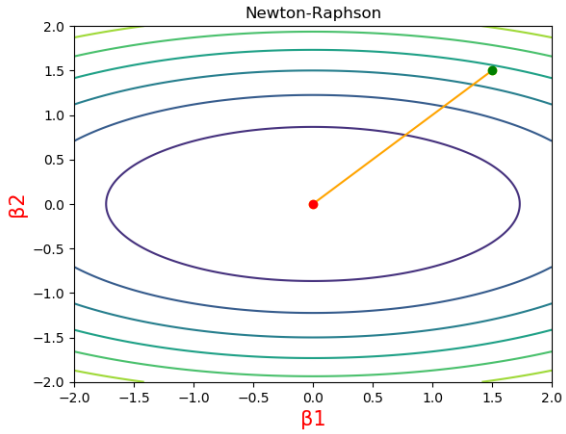


*Figura 3.1:* Comportamento da sequência gerada pelo método de Newton-Raphson para a função objetivo dada pela Equação (3.2). O ponto vermelho é o valor em que é inicializado o método, $\beta_0 = (1,5\,;\,1,5)$ e o ponto verde é o valor estimado pelo método, $\beta_{min} = (0,0\,;\,0,0)$ para o ponto 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 encontra-se distante do ponto ótimo, pode haver problemas de convergência da sequência para o ponto de mínimo. Como mencionado anteriormente, um aspecto que deve ser considerado para a adoção de um método numérico é sua taxa de convergência.

### *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 derivada não pode ser nula em uma vizinhança do ponto que é a solução da equação $g(\beta) = 0$. Dada a informação e as restrições, o método de Newton-Raphson é capaz de gerar uma sequência $\{\beta_k\}_{k\geqslant 0}$ tal que a sequência $\{g(\beta_k)\}_{k\geqslant 0}$ tem uma ordem de convergência consideravelmente elevada. Para mostrar esse resultado, precisa-se definir algumas normas sobre os espaços das derivadas de ordem um e dois da função $g$.

**Definição 4.** *Seja $\mathcal{B}\left(\mathbb{R}^p;\mathbb{R}^p\right)$ o espaço vetorial das aplicações lineares. Defini-se nesse espaço a norma $||.||_{\mathcal{B}} : \mathcal{B}\left(\mathbb{R}^p;\mathbb{R}^p\right) \longrightarrow \mathbb{R}$ por,*

$$||A||_{\mathcal{B}} := \sup_{||v||=1} \{||Av||\},$$

*em que $A \in \mathcal{B}\left(\mathbb{R}^p;\mathbb{R}^p\right)$.*

**Definição 5.** *Seja $\mathcal{B}_2\left(\mathbb{R}^p;\mathbb{R}^p\right)$ o espaço vetorial das aplicações bilineares. Defini-se nesse espaço a norma $||.||_{\mathcal{B}_2} : \mathcal{B}_2\left(\mathbb{R}^p;\mathbb{R}^p\right) \longrightarrow \mathbb{R}$ por,*

$$||A||_{\mathcal{B}_2} := \sup_{||u||=||v||=1} \{||u^T Av||\},$$

*em que $A \in \mathcal{B}_2\left(\mathbb{R}^p;\mathbb{R}^p\right)$.*

Com as Definições 4 e 5, segue-se para a demonstração do principal resultado dessa seção.

**Lema 6.** *Sejam $g \in C^2(\mathbb{R}^p;\mathbb{R}^p)$ e $\beta_* \in \mathbb{R}^p$ tal que $g(\beta_*) = 0$. Considere $\epsilon > 0$ tal que para todo $\beta \in B[\beta_*;\epsilon]$ a matriz jacobiana, $\nabla g(\beta)$, seja não nula e que a inversa da matriz jacobiana, $\left[\nabla g(\beta)\right]^{-1}$, exista e seja contínua. Então, dada a sequência $\{\beta_k\}_{k\geqslant 0}$ gerada pelo algoritmo definido na Equação (3.4) com $\{\beta_k\}_{k\geqslant 0} \subset B[\beta_*;\epsilon]$, tem-se que a taxa de convergência da sequência $\{g(\beta_k)\}_{k\geqslant 0}$ é exponencial no seguinte sentido:*

$$||g(\beta_k) - g(\beta_*)|| \leqslant \mathcal{O}\left(\eta^{2^k}\right), \quad com \quad \eta \in (0,1).$$

*Demonstração.* Dada a sequência $\{\beta_k\}_{k\geqslant 0} \subset B[\beta_*;\epsilon]$ gerada pelo método de Newton-Raphson, tem-se pela expansão de Taylor de primeira ordem com resto de Lagrange na vizinhança de $\beta_{k-1}$ que

$$g(\beta_k) = g(\beta_{k-1}) + \nabla g(\beta_{k-1})\big(\beta_k - \beta_{k-1}\big) + \frac{1}{2}\big(\beta_k - \beta_{k-1}\big)^T \nabla^2 g(\xi_k)\big(\beta_k - \beta_{k-1}\big),$$

em que $\xi_k$ é um ponto pertencente ao seguimento de reta que tem como extremidades os pontos $\beta_{k-1}$ e $\beta_k$. Pela Equação (3.4), segue-se que

$$\begin{aligned}
\beta_k &= \beta_{k-1} - \big[\nabla g(\beta_{k-1})\big]^{-1} g(\beta_{k-1}) \quad \Rightarrow \\
\beta_k - \beta_{k-1} &= -\big[\nabla g(\beta_{k-1})\big]^{-1} g(\beta_{k-1}) \quad \Rightarrow \\
\nabla g(\beta_{k-1})\big(\beta_k - \beta_{k-1}\big) &= -g(\beta_{k-1}) \quad \Rightarrow \\
g(\beta_{k-1}) + \nabla g(\beta_{k-1})\big(\beta_k - \beta_{k-1}\big) &= 0.
\end{aligned}$$

Portanto,

$$g(\beta_k) = \frac{1}{2}\big(\beta_k - \beta_{k-1}\big)^T \nabla^2 g(\xi_k)\big(\beta_k - \beta_{k-1}\big).$$

Aplicando-se a norma da Definição 5, obtem-se que

$$||g(\beta_k)|| \leqslant \frac{1}{2}||\nabla^2 g(\xi_k)||_{\mathcal{B}_2}||\beta_k - \beta_{k-1}||^2, \tag{3.5}$$

e como $\nabla^2 g(\xi)$ é contínua, tem-se pela compacidade de $B[\beta_*;\epsilon]$, que existe $M_1 > 0$ tal que

$$M_1 = \sup_{\xi\in B[\beta_*;\epsilon]} \{(1/2)||\nabla^2 g(\xi)||_{\mathcal{B}_2}\}. \tag{3.6}$$

Ao substituir o resultado da Equação (3.6) na Equação (3.5), tem-se que

$$||g(\beta_k)|| \leqslant M_1||\beta_k - \beta_{k-1}||^2.$$

Novamente pela Equação (3.4), segue-se que

$$||g(\beta_k)|| \leqslant M_1||\big[\nabla g(\beta_{k-1})\big]^{-1} g(\beta_{k-1})||^2.$$

Aplicando-se a norma dada pela Definição 4, obtem-se que

$$||g(\beta_k)|| \leqslant M_1||\big[\nabla g(\beta_{k-1})\big]^{-1}||^2_{\mathcal{B}}||g(\beta_{k-1})||^2. \tag{3.7}$$

Como, por hipótese, a aplicação $\big[\nabla g(\xi)\big]^{-1}$ em $B[\beta_*;\epsilon]$ é contínua, segue-se que existe $M_2 > 0$ tal que

$$M_2 = \sup_{\xi\in B[\beta_*;\epsilon]} \{||\big[\nabla g(\xi)\big]^{-1}||^2_{\mathcal{B}}\}. \tag{3.8}$$

Ao substituir o resultado da Equação (3.8) na Equação (3.7), obtem-se

$$||g(\beta_k)|| \leqslant M_1 M_2 ||g(\beta_{k-1})||^2. \tag{3.9}$$

Defina $C_1 = M_1 M_2$. Então, pode-se reescrever a Equação (3.9) como

$$||g(\beta_k)|| \leqslant C_1 ||g(\beta_{k-1})||^2. \tag{3.10}$$

Agora, considere a seguinte afirmação: a desigualdade definida a seguir é verdadeira,

$$||g(\beta_k)|| \leqslant C_1^{2^k-1} ||g(\beta_0)||^{2^k} \quad \text{(é verdadeira)}. \tag{3.11}$$

A prova é feita por indução sobre $k \geqslant 1$. Para $k = 1$, segue-se que

$$||g(\beta_1)|| \leqslant C_1 ||g(\beta_0)||^2$$

é verdadeira de acordo com a Equação (3.10). Agora, suponha que a Equação (3.11) seja verdadeira. Multiplicando-se ambos os lados da Equação (3.11) por $\sqrt{C_1}$, tem-se que

$$\sqrt{C_1}||g(\beta_k)|| \leqslant \sqrt{C_1} C_1^{2^k-1} ||g(\beta_0)||^{2^k},$$

e tomando-se a potência quadrática em ambos os lados da desigualdade resulta que

$$\begin{aligned} C_1||g(\beta_k)||^2 \leqslant C_1\big(C_1^{2^k-1}||g(\beta_0)||^{2^k}\big)^2 = C_1^{2^{k+1}-1}||g(\beta_0)||^{2^{k+1}} \quad &\Rightarrow \\ ||g(\beta_{k+1})|| \leqslant C_1^{2^{k+1}-1}||g(\beta_0)||^{2^{k+1}} \qquad &\text{(pela Equação (3.10))}. \end{aligned}$$

Portanto, segue-se por indução que a desigualdade dada pela Equação (3.11) é verdadeira para todo $k \geqslant 1$. Note que a desigualdade dada pela Equação (3.11) pode ser reescrita do seguinte modo:

$$||g(\beta_k)|| \leqslant C_1^{-1}\big(C_1||g(\beta_0)||\big)^{2^k}. \tag{3.12}$$

Agora, como $g$ é uma função contínua, tem-se que dado $0 < \eta < 1$ existe $\delta > 0$ tal que

$$||\beta - \beta_*|| < \delta \quad \Rightarrow \quad ||g(\beta) - g(\beta_*)|| \leqslant \frac{\eta}{C_1}.$$

Logo, para $\beta_0 \in B[\beta_*; \delta] \cap B[\beta_*; \epsilon]$, tem-se que

$$C_1||g(\beta_0)|| < \eta < 1 \quad \Rightarrow \quad \big(C_1||g(\beta_0)||\big)^{2^k} < \eta^{2^k} \quad \Rightarrow \quad C_1^{-1}\big(C_1||g(\beta_0)||\big)^{2^k} < C_1^{-1}\eta^{2^k},$$

para todo $k \geqslant 1$. Portanto, pela Equação (3.12), segue-se que

$$||g(\beta_k)|| \leqslant C_1^{-1}\eta^{2^k} \quad \Rightarrow \quad ||g(\beta_k) - g(\beta_*)|| \leqslant \mathcal{O}\big(\eta^{2^k}\big).$$

□

### *3.1.3 Escore de Fisher*

Em modelagem Estatística, quando há a necessidade de recorrer-se a um método numérico para estimar os parâmetros de um determinado modelo, é comum empregar-se uma variação do método de Newton-Raphson. Isso deve-se ao fato do procedimento de estimação ser via o logaritmo da função de verossimilhança. Nos modelos lineares generalizados, dado o logaritmo da função de verossimilhança dada pela Equação (2.4), a função objetivo $g$ é definida por

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

Segue-se que se $\beta_*$ for a estimativa de máxima verossimilhança do parâmetro $\beta$, 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 que $\nabla g$ seja inversível. Logo, tem-se que o método de Newton-Raphson para os modelos lineares generalizados é dado por

$$\beta_k = \beta_{k-1} - \left[\nabla g(\beta_{k-1})\right]^{-1} g(\beta_{k-1}).$$

Como discutido na Seção 2.3, em Estatística é comum utilizar toda a informação disponível na amostra de modo a se 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 dada pela Equação (2.6) com o método de Newton-Raphson apresentado na Equação (3.4) recebe o nome de *método de Escore de Fisher* e é definido por

$$\beta_k = \beta_{k-1} - \left[\mathcal{I}(\beta_{k-1})\right]^{-1} g(\beta_{k-1}).$$

Esse método desempenha um papel importante na classe do modelos lineares generalizado porque permite a estimação dos parâmetros do preditor linear do modelo com uma boa eficiência.

## 3.2 Gradiente descendente

Do cálculo em várias variáveis, sabe-se que o vetor gradiente de uma função de $p$ variáveis reais aponta para a direção de maior crescimento dessa função, e portanto, o negativo do vetor gradiente aponta para a direção de decrescimento da função. O método do gradiente descendente utiliza esse segundo fato para construir uma sequência de valores tal que ao longo da trajetória formada, por essa sequência, a função em questão seja decrescente. Formalmente, considere $f : \mathbb{R}^p \longrightarrow \mathbb{R}$ uma função diferenciável e $\nabla f : \mathbb{R}^p \longrightarrow \mathbb{R}^p$ seu vetor gradiente. Defíne-se a sequência $\{\beta_k\}_{k\geqslant 0}$ por

$$\beta_k = \beta_{k-1} - \epsilon_k \nabla f(\beta_{k-1}), \tag{3.13}$$

em que $\epsilon_k > 0$ é escolhido a cada iteração de tal modo que ao longo da sequência $\{\beta_k\}_{k\geqslant 0}$ tenha-se

$$f(\beta_k) < f(\beta_{k-1}). \tag{3.14}$$

A escolha do valor de $\epsilon_k > 0$ de modo que a desigualdade dada na Equação (3.14) seja satisfeita é parte fundamental na aplicação desse método, pois se o valor de $\epsilon_k$ for muito pequeno o método torna-se lento e se o valor de $\epsilon_k$ for muito grande ele pode divergir. Um modo de se obter o valor ótimo para $\epsilon_k$ é através do procedimento conhecido como busca linear. Considere a sequência de funções $\phi_k : \mathbb{R}_+ \longrightarrow \mathbb{R}$ definidas por

$$\phi_k(\epsilon) = f(\beta_{k-1} - \epsilon \nabla f(\beta_{k-1}))$$

e, então, defina $\epsilon_k$ pelo seguinte procedimento de minimização:

$$\epsilon_k = \arg\min_{\epsilon \in \mathbb{R}_+} \{\phi_k(\epsilon)\}. \tag{3.15}$$

Note que se o procedimento da Equação (3.15) não puder ser realizado de modo a se obter uma forma fechada para $\epsilon_k$, então os custos computacionais podem ser elevados caso a minimização dada pela Equação (3.15) seja executada a cada iteração. Na literatura sobre métodos de otimização algumas alternativas ao procedimento dado pela Equação (3.15) são abordadas, como, por exemplo, os métodos de *Armijo* (Armijo, 1966) e *Wolfe* (Wolfe, 1969). Contudo, este não é o objetivo desta dissertação de mestrado. Para um tratamento mais profundo sobre a determinação do passo $\epsilon_k$ veja, por exemplo, Izmailov e Solodov (2012) e referências.

A Figura 3.2 exibe o comportamento da sequência gerada pelo método do gradiente descendente. A trajetória gerada pelo método segue ortogonal as curvas de nível da função objetivo dada pela Equação (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 da função objetivo garantia que a trajetória fosse, em certo sentido ótima, tem-se que no caso do método do gradiente descendente isso não é mais possível, pois essa informação está ausente nesse método. A convergência observada na Figura 3.2 ocorreu em 25 passos.

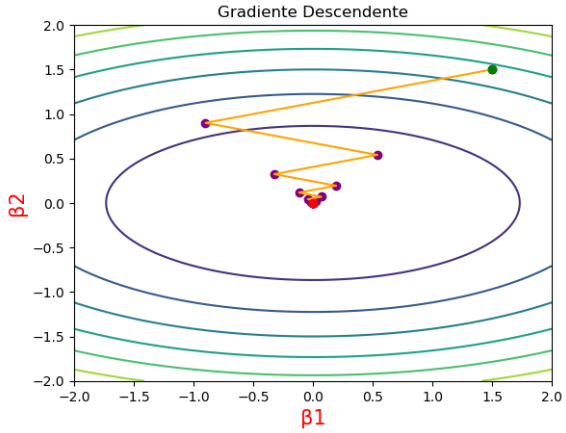


*Figura 3.2:* Comportamento da sequência gerada pelo método do gradiente descendente para a função objetivo dada pela Equação (3.2). O ponto vermelho é o valor em que é inicializado o método, $\beta_0 = (1,5\,;\,1,5)$ e o ponto verde é o valor estimado pelo método, $\beta_{min} = (0,0\,;\,0,0)$ para o ponto de mínimo.

### *3.2.1 Taxa de convergência*

O método do gradiente descendente tem diversas características interessantes tanto ao nível teórico quanto de aplicação. A característica que mostra-se mais relevante para o estudo desenvolvido nessa subseção é a taxa de convergência do método. Começa-se o estudo dessa propriedade pelo Lema 7 abaixo.

**Lema 7.** *Seja* $f : \mathbb{R}^p \longrightarrow \mathbb{R}$ *uma função convexa e diferenciável. Considere* $\epsilon > 0$ *um tamanho de passo escolhido de tal modo que a desigualdade dada pela Equação* (3.14) *seja satisfeita. Então, 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 $\beta_*$ é o ponto de mínimo para $f$.*

*Demonstração.* Seja $\{\mathcal{E}_k\}_{k\geqslant 1}$ uma sequência definida por

$$\mathcal{E}_k = \epsilon k(f(\beta_k) - f(\beta_*)) + \frac{1}{2}||\beta_k - \beta_*||^2. \tag{3.16}$$

Pela Equação (3.13) - método do gradiente descendente - e pelo fato de $f$ ser diferenciável, segue-se que

$$f(\beta_k) = f(\beta_{k-1}) - \epsilon||\nabla f(\beta_{k-1})||^2 + o(\epsilon). \tag{3.17}$$

Combinado-se as Equações (3.16) e (3.17), obtem-se

$$\mathcal{E}_k = \epsilon k(f(\beta_{k-1}) - f(\beta_*)) - \epsilon^2 k||\nabla f(\beta_{k-1})||^2 + o(\epsilon^2) + \frac{1}{2}||\beta_k - \beta_*||^2. \tag{3.18}$$

Novamente pela Equação (3.13), tem-se que

$$\frac{1}{2}||\beta_k - \beta_*||^2 = \frac{1}{2}||\beta_{k-1} - \beta_*||^2 - \epsilon[\nabla f(\beta_{k-1})]^T\big(\beta_{k-1} - \beta_*\big) + \frac{\epsilon^2}{2}||\nabla f(\beta_{k-1})||^2. \tag{3.19}$$

Combinado-se as Equações (3.18) e (3.19), obtem-se

$$\begin{aligned}\mathcal{E}_k &= \epsilon k(f(\beta_{k-1}) - f(\beta_*)) + \frac{1}{2}||\beta_{k-1} - \beta_*||^2 \\ &+ \big(\frac{1}{2} - k\big)\epsilon^2||\nabla f(\beta_{k-1})||^2 + o(\epsilon^2) - \epsilon\nabla f(\beta_{k-1})^T\big(\beta_{k-1} - \beta_*\big).\end{aligned} \tag{3.20}$$

Somando e subtraindo $\epsilon(f(\beta_{k-1}) - f(\beta_*))$ na Equação (3.20), tem-se que

$$\begin{aligned}\mathcal{E}_k &= \epsilon(k-1)(f(\beta_{k-1}) - f(\beta_*)) + \frac{1}{2}||\beta_{k-1} - \beta_*||^2 \\ &+\big(\frac{1}{2} - k\big)\epsilon^2||\nabla f(\beta_{k-1})||^2 + o(\epsilon^2) + \epsilon[f(\beta_{k-1}) - f(\beta_*) - \nabla f(\beta_{k-1})^T\big(\beta_{k-1} - \beta_*\big)].\end{aligned}$$

Como $f$ é uma função convexa e diferenciável, segue-se que

$$f(\beta_{k-1}) - f(\beta_*) - \nabla f(\beta_{k-1})^T\big(\beta_{k-1} - \beta_*\big) \leqslant 0,$$

e dado que $\big(\frac{1}{2} - k\big)\epsilon^2||\nabla f(\beta_{k-1})||^2$ é negativo, obtem-se

$$\mathcal{E}_k \leqslant \epsilon(k-1)(f(\beta_{k-1}) - f(\beta_*)) + \frac{1}{2}||\beta_{k-1} - \beta_*||^2 + o(\epsilon^2) = \mathcal{E}_{k-1} + o(\epsilon^2).$$

E concluí-se que

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

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

□

### *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 a teoria dos sistemas dinâmicos. Além disso, tem-se que uma parte considerável das ciências da natureza 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 $\beta \in C^1([0,+\infty);\mathbb{R}^p)$ e $\nabla f : \mathbb{R}^p \longrightarrow \mathbb{R}^p$ o campo de vetores Lipschitz contínuo dado pelo vetor gradiente da função convexa $f : \mathbb{R}^p \longrightarrow \mathbb{R}$. Então, tem-se o seguinte problema de valor inicial:

$$\begin{aligned} \dot{\beta}(t) &= -\nabla f(\beta(t)), \\ \beta(0) &= \beta_0. \end{aligned} \tag{3.21}$$

A EDO de primeira ordem dada pela Equação (3.21) relaciona-se com o método do gradiente descendente através de sua discretização pelo método das diferenças finitas. Como a curva $\beta : [0,+\infty) \longrightarrow \mathbb{R}^p$ é diferenciável, tem-se pela expansão de Taylor de primeira ordem que

$$\beta(t+\epsilon) = \beta(t) + \dot{\beta}(t)\epsilon + \mathcal{O}(\epsilon^2). \tag{3.22}$$

Passando-se $\beta(t)$ para o lado esquerdo da igualdade da Equação (3.22) e dividindo a expressão por $\epsilon$, obtem-se,

$$\frac{\beta(t+\epsilon) - \beta(t)}{\epsilon} = \dot{\beta}(t) + \mathcal{O}(\epsilon). \tag{3.23}$$

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

$$\frac{\beta_{k+1} - \beta_k}{\epsilon} = \dot{\beta}(t) + \mathcal{O}(\epsilon). \tag{3.24}$$

As etapas apresentadas nas Equações (3.23) e (3.24) são conhecidas como método de Euler progressivo. Esse é um método de discretização de equações diferenciais muito comum e

utilizado na prática. Agora, substituindo-se a Equação (3.24) na Equação (3.21), obtem-se que

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

que por sua vez recai no método do gradiente descendente. Da definição da EDO de primeira ordem dada pela Equação (3.21), tem-se que a curva definida por $\beta(t) = \beta_*$ - em que $\beta_*$ é ponto de mínimo da função objetivo $f$ - é solução para o problema de valor inicial dado pela Equação (3.21) quando considera-se $\beta_0 = \beta_*$. 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, tem-se que as curvas solução do problema de valor inicial na Equação (3.21) são ortogonais as curvas de nível da função objetivo e, portanto, a sequência gerada pelo método do gradiente descendente que discretiza essa curva terá um comportamento próximo a este. Com ferramentas mais sofisticadas da teoria das equações diferenciais ordinárias outras propriedades podem ser inferidas sobre o comportamento da curva solução e, consequentemente, sobre o comportamento da sequência que a discretiza.

### *3.2.3 Taxa de convergência das soluções da EDO associada*

A maioria das equações diferenciais ordinárias não apresentam solução analítica o que torna necessário a utilização de algum método numérico para a obtenção de soluções aproximadas. Nesse contexto faz-se necessário verificar a compatibilidade entre a taxa de convergência da curva solução da EDO e da sequência gerada pelo método que a discretiza para se poder ter uma estimativa de quão custoso é a obtenção da aproximação numérica da solução real. O Lema 8 trata da taxa de convergência para as curvas solução da EDO de primeira ordem dada pela Equação (3.21).

**Lema 8.** *Seja $\beta : [0, +\infty) \longrightarrow \mathbb{R}^p$ uma curva solução para a Equação (3.21). Então, a taxa de convergência da curva solução desta EDO é dada por*

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

*em que $\beta_*$ é o ponto de mínimo para $f$.*

*Demonstração.* Considere a função $\mathcal{E} : [0, +\infty) \longrightarrow \mathbb{R}$ definida por

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

Derivando $\mathcal{E}$, obtem-se que

$$\dot{\mathcal{E}}(t) = (f(\beta(t)) - f(\beta_*)) + t\nabla f(\beta(t))^T\dot{\beta}(t) + \dot{\beta}(t)^T\big(\beta(t) - \beta_*\big).$$

Como a curva $\beta(t)$ é solução para a Equação (3.21), tem-se que

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

Além disso, dado que $f$ é uma função convexa e diferenciável, segue-se que

$$0 \geqslant f(\beta(t)) - f(\beta_*) - \nabla f(\beta(t))^T\big(\beta(t) - \beta_*\big).$$

Como $-t||\nabla f(\beta(t))||^2$ é negativo para todo $t > 0$, obtem-se que

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

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

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

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

□

Na Seção 3.3, a seguir, discute-se como o método do gradiente descendente pode ser acelerado dando origem ao método de Nesterov e como esse método também tem uma EDO (de segunda ordem) associada a ele.

## 3.3 *Gradiente acelerado*

O método do gradiente acelerado ou método de Nesterov tem como base o método do gradiente descendente e apesar de não ser um método de descida, isto é, a desigualdade (3.14) não é satisfeita pela sequência, $\{\beta_k\}_{k\geqslant 0}$, gerada pela método, esse procedimento, utilizando a mesma informação disponível para o método do gradiente descendente, a saber, a primeira derivada da função objetivo, consegue obter uma taxa de convergência da ordem de $\mathcal{O}\big(1/\epsilon k^2\big)$. Essa aceleração ocorre por uma leve modificação do método 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 $f : \mathbb{R}^p \longrightarrow \mathbb{R}$ uma função diferenciável e $\nabla f : \mathbb{R}^p \longrightarrow \mathbb{R}^p$ seu vetor gradiente. Defíne-se as sequências $\{\beta_k\}_{k\geqslant 0}$ e $\{\zeta_k\}_{k\geqslant 0}$ por:

$$\begin{aligned} \beta_{k+1} &= \zeta_k - \epsilon \nabla f(\zeta_k), \\ \zeta_{k+1} &= \beta_{k+1} + \tfrac{k-1}{k+2}\big(\beta_{k+1} - \beta_k\big). \end{aligned} \tag{3.25}$$

em que $\epsilon > 0$ 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 buscaremos compreender melhor como a introdução da sequência $\{\zeta_k\}_{k\geqslant 0}$ com o seu termo de momento $\beta_{k+1} - \beta_k$ e de resistência do meio $\frac{k-1}{k+2}$ permitem que a sequência $\{\beta_k\}_{k\geqslant 0}$ seja acelerada em uma ordem acima, isto é, de $\mathcal{O}\big(1/\epsilon k\big)$ para $\mathcal{O}\big(1/\epsilon k^2\big)$. Os termos *momento* e *resistência do meio* tem sua origem na Física e serão utilizados para facilitar a compreensão do comportamento das sequências dada na Equação (3.25) através de analogias com fenômenos físicos como a dinâmica de partículas.

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 modificou-se consideravelmente. Agora a trajetória da sequência não apresenta mais um movimento em zigue-zague e em seu lugar surge um movimento em duas partes: Na primeira, a sequência se desloca em um passo até um ponto médio. Na segunda, segue em linha reta até o ponto de mínimo. A convergência observada na Figura 3.3 ocorreu em 21 passos.

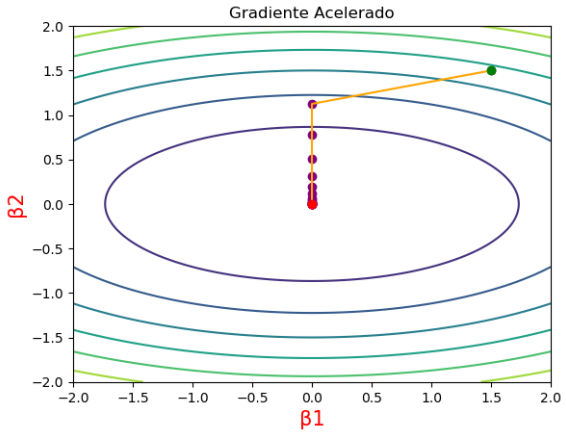


*Figura 3.3:* Comportamento da sequência gerada pelo método do gradiente acelerado para a função objetivo dada pela Equação (3.2). O ponto vermelho é o valor em que é inicializado o método, $\beta_0 = (1,5\,;\,1,5)$ e o ponto verde é o valor estimado pelo método, $\beta_{min} = (0,0\,;\,0,0)$ para o 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 faz-se necessário alguns resultados técnicos preliminares que são expostos abaixo em forma de lemas.

**Lema 9.** *Sejam $\epsilon > 0$ e $f : \mathbb{R}^p \longrightarrow \mathbb{R}$ uma função convexa é diferenciável. Para todo $\bar{\beta}, \hat{\beta} \in \mathbb{R}^p$ tem-se que vale a seguinte desigualdade,*

$$f(\hat{\beta} - \epsilon \nabla f(\hat{\beta})) \leqslant f(\bar{\beta}) + \nabla f(\hat{\beta})^T(\hat{\beta} - \bar{\beta}) - \frac{\epsilon}{2}||\nabla f(\hat{\beta})||^2 \tag{3.26}$$

*Demonstração.* Como $f$ é uma função convexa e diferenciável tem-se que,

$$f(\bar{\beta}) \geqslant f(\hat{\beta}) + \nabla f(\hat{\beta})^T(\bar{\beta} - \hat{\beta}).$$

E isso implica que,

$$f(\hat{\beta}) \leqslant f(\bar{\beta}) + \nabla f(\hat{\beta})^T(\hat{\beta} - \bar{\beta}). \tag{3.27}$$

Dado que $f$ é diferenciável segue-se que,

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

em que $E_{\hat{\beta}} : \mathbb{R}_+ \longrightarrow \mathbb{R}_+$ é uma função tal que,

$$\lim_{\epsilon \longrightarrow 0} \frac{E_{\hat{\beta}}(\epsilon)}{\epsilon} = 0.$$

Dado $\eta = \frac{1}{2}||\nabla f(\hat{\beta})||^2$ tem-se que existe $\delta(\eta) > 0$ tal que, se $\delta(\eta) > \epsilon > 0$, então

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

E segue-se que,

$$f(\hat{\beta} - \epsilon \nabla f(\hat{\beta})) \leqslant f(\hat{\beta}) - \epsilon||\nabla f(\hat{\beta})||^2 + \frac{\epsilon}{2}||\nabla f(\hat{\beta})||^2 = f(\hat{\beta}) - \frac{\epsilon}{2}||\nabla f(\hat{\beta})||^2.$$

Combinando a desigualdade dada na Equação (3.27) com a desigualdade apresentada acima se obtêm que,

$$f(\hat{\beta} - \epsilon \nabla f(\hat{\beta})) \leqslant f(\bar{\beta}) + \nabla f(\hat{\beta})^T(\hat{\beta} - \bar{\beta}) - \frac{\epsilon}{2}||\nabla f(\hat{\beta})||^2.$$

□

**Lema 10.** *Considere $f : \mathbb{R}^p \longrightarrow \mathbb{R}$ uma função convexa e diferenciável. Dados $\epsilon > 0$ e as sequências $\{\zeta_k\}_{k\geqslant 0}$ e $\{\beta_k\}_{k\geqslant 0}$ definidas na Equação (3.25). Seja $\{U_k\}_{k\geqslant 0}$ uma sequência definida por,*

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

*Então tem-se que,*

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

*Demonstração.* Inicialmente tem-se que,

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

Como,

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

Segue-se que,

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

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 $(k+2)/2$ tem-se que,

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

Portanto segue-se que,

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

□

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

**Teorema 11.** *Considere $f : \mathbb{R}^p \longrightarrow \mathbb{R}$ uma função convexa e diferenciável. Seja $\{\beta_k\}_{k\geqslant 0}$ a sequência dada na Equação (3.25). Então tem-se que,*

$$f(\beta_k) - f(\beta_*) \leqslant \mathcal{O}\left(\frac{1}{\epsilon(k+1)^2}\right), \tag{3.29}$$

*em que $\beta_*$ é o ponto de mínimo para $f$.*

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

$$\mathcal{E}_k = \epsilon(k+1)^2(f(\beta_k) - f(\beta_*)) + 2||U_k - \beta_*||^2, \tag{3.30}$$

em que $U_k = \frac{k+2}{2}\zeta_k - \frac{k}{2}\beta_k$. Substituindo na Equação (3.26) $\hat{\beta} = \zeta_{k-1}$ e $\bar{\beta} = \beta_{k-1}$ se obtêm,

$$f(\zeta_{k-1} - \epsilon\nabla f(\zeta_{k-1})) \leqslant f(\beta_{k-1}) + \nabla f(\zeta_{k-1})^T(\zeta_{k-1} - \beta_{k-1}) - \frac{\epsilon}{2}||\nabla f(\zeta_{k-1})||^2. \tag{3.31}$$

Repetindo o processo acima para os valores, $\hat{\beta} = \zeta_{k-1}$ e $\bar{\beta} = \beta_*$ segue-se que,

$$f(\zeta_{k-1} - \epsilon\nabla f(\zeta_{k-1})) \leqslant f(\beta_*) + \nabla f(\zeta_{k-1})^T(\zeta_{k-1} - \beta_*) - \frac{\epsilon}{2}||\nabla f(\zeta_{k-1})||^2. \tag{3.32}$$

Multiplicando a Equação (3.31) por $(k-1)/(k+1)$ e a Equação (3.32) por $2/(k+1)$ e depois somando ambas segue-se que,

$$f(\zeta_{k-1} - \epsilon\nabla f(\zeta_{k-1})) \leqslant \frac{k-1}{k+1}f(\beta_{k-1}) + \frac{2}{k+1}f(\beta_*) +$$

$$\nabla f(\zeta_{k-1})^T\left[\frac{k-1}{k+1}\big(\zeta_{k-1} - \beta_{k-1}\big) + \frac{2}{k+1}\big(\zeta_{k-1} - \beta_*\big)\right] - \frac{\epsilon}{2}||\nabla f(\zeta_{k-1})||^2.$$

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

$$f(\beta_k) \leqslant \frac{k-1}{k+1}f(\beta_{k-1}) + \frac{2}{k+1}f(\beta_*) +$$

$$\nabla f(\zeta_{k-1})^T\left[\frac{k-1}{k+1}\big(\zeta_{k-1} - \beta_{k-1}\big) + \frac{2}{k+1}\big(\zeta_{k-1} - \beta_*\big)\right] - \frac{\epsilon}{2}||\nabla f(\zeta_{k-1})||^2$$

e dado que $U_{k-1} = \frac{k+1}{2}\zeta_{k-1} - \frac{k-1}{2}\beta_{k-1}$ tem-se que,

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

Pela igualdade dada na Equação (3.28) tem-se que,

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

E dado que,

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

Segue-se que,

$$f(\beta_k) \leqslant \frac{k-1}{k+1} f(\beta_{k-1}) + \frac{2}{k+1} f(\beta_*) + \frac{4}{\epsilon(k+1)^2} \big(U_{k-1} - U_k\big)^T \big(U_{k-1} - \beta_*\big)$$

$$- \frac{2}{\epsilon(k+1)^2} ||U_{k-1} - U_k||^2.$$

Mas tem-se que,

$$||U_{k-1} - \beta_* + \beta_* - U_k||^2 = ||U_{k-1} - \beta_*||^2 + 2\big(U_{k-1} - \beta_*\big)^T \big(\beta_* - U_k\big) + ||U_k - \beta_*||^2$$

e portanto,

$$f(\beta_k) \leqslant \frac{k-1}{k+1} f(\beta_{k-1}) + \frac{2}{k+1} f(\beta_*) + \frac{4}{\epsilon(k+1)^2} \big(U_{k-1} - U_k\big)^T \big(U_{k-1} - \beta_*\big)$$

$$- \frac{2}{\epsilon(k+1)^2} ||U_{k-1} - \beta_*||^2 - \frac{4}{\epsilon(k+1)^2} \big(U_{k-1} - \beta_*\big)^T \big(\beta_* - U_k\big) - \frac{2}{\epsilon(k+1)^2} ||U_k - \beta_*||^2.$$

Além disso tem-se que,

$$\big(U_{k-1} - \beta_* + \beta_* - U_k\big)^T \big(U_{k-1} - \beta_*\big) = \big(U_{k-1} - \beta_*\big)^T \big(U_{k-1} - \beta_*\big) + \big(\beta_* - U_k\big)^T (U_{k-1} - \beta_*)$$

$$= ||U_{k-1} - \beta_*||^2 + \big(\beta_* - U_k\big)^T (U_{k-1} - \beta_*).$$

E segue-se que,

$$f(\beta_k) \leqslant \frac{k-1}{k+1} f(\beta_{k-1}) + \frac{2}{k+1} f(\beta_*) + \frac{4}{\epsilon(k+1)^2} ||U_{k-1} - \beta_*||^2 + \frac{4}{\epsilon(k+1)^2} \big(\beta_* - U_k\big)^T (U_{k-1} - \beta_*)$$

$$- \frac{2}{\epsilon(k+1)^2} ||U_{k-1} - \beta_*||^2 - \frac{4}{\epsilon(k+1)^2} \big(U_{k-1} - \beta_*\big)^T \big(\beta_* - U_k\big) - \frac{2}{\epsilon(k+1)^2} ||U_k - \beta_*||^2.$$

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 $\epsilon(k+1)^2$ se obtêm,

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

E segue-se que,

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

$$+ 2||U_{k-1} - \beta_*||^2 - 2||U_k - \beta_*||^2.$$

Como,

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

Tem-se que,

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

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

E segue-se que,

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

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 $i = k$ se obtêm que,

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

E isso implica que,

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

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

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

E conclui-se que,

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

□

### *3.3.2 EDO de segunda ordem associada*

Nesta subseção realiza-se um série de procedimentos análogos aos que foram efetuados na Subseção 3.2.2 principalmente na utilização de referências a física, em especial a mecânica clássica. Será mostrado como uma equação diferencial ordinária de segunda ordem emerge de modo natural do método de Nesterov. Nessa equação pode-se destacar dois componentes: O primeiro é o termo de aceleração $\ddot{\beta}$ responsável pela alteração da

velocidade da curva ao longo do tempo, o segundo é o termo $\frac{3}{t}\dot{\beta}$ que pode ser entendido como um fator de resistência do meio ao longo da trajetória.

Considere $\beta \in C^2([0,+\infty);\mathbb{R}^p)$ e $\nabla f : \mathbb{R}^p \longrightarrow \mathbb{R}^p$ um campo de vetores Lipschitz contínuo dado pelo vetor gradiente da função convexa e diferenciável $f : \mathbb{R}^p \longrightarrow \mathbb{R}$. Então tem-se o seguinte problema de valor inicial:

$$\begin{aligned} \ddot{\beta}(t) + \tfrac{3}{t}\dot{\beta}(t) + \nabla f(\beta(t)) = 0, \\ \beta(0) = \beta_0, \\ \dot{\beta}(0) = 0. \end{aligned} \tag{3.33}$$

Observando a EDO dada na Equação (3.33) em conjunto com a Figura 3.3 pode-se empreender uma análise qualitativa simples que comece a esclarecer o comportamento da trajetória do método de Nesterov. Olhando para o segundo termo da EDO dada na Equação (3.33) tem-se o termo de resistência do meio, esse termo assume, para valores pequenos de $t$, um valor elevado e interpretando a curva solução da EDO dada na Equação (3.33) como a trajetória de uma partícula em um meio viscoso, tem-se que a partícula progride lentamente sem oscilações por esse meio.

Agora para valores elevados de $t$ tem-se que o termo de resistência é aproximadamente zero e nesse meio de baixa resistência o termo de aceleração torna-se dominante surgindo oscilações próximas ao ponto onde o campo de forças $\nabla f$ se anula. Para mostrar que a EDO dada na Equação (3.33) pode ser deduzida das sequências dadas na Equação (3.25) necessita-se da seguinte definição.

**Definição 12.** *O conjunto $\mathcal{F}_L$, com $L > 0$, é o conjunto das funções convexas e diferenciáveis com a seguinte propriedade,*

$$\mathcal{F}_L := \{f : \mathbb{R}^p \longrightarrow \mathbb{R} : \quad ||\nabla f(\hat{\beta}) - \nabla f(\bar{\beta})|| \leqslant L||\hat{\beta} - \bar{\beta}||; \quad \forall \hat{\beta}, \bar{\beta} \in \mathbb{R}^p\}.$$

Também faz-se necessário os seguintes teoremas cujas as demonstração serão deixadas para as referências pois suas apresentações não são necessárias para a compreensão dos resultados dessa dissertação.

**Teorema 13.** *Para toda função $f \in \mathcal{F}_\infty := \cup_{L>0}\mathcal{F}_L$ e todo $\beta_0 \in \mathbb{R}^p$, a EDO dada na Equação (3.33) com condições iniciais $\beta(0) = \beta_0$ e $\dot{\beta}(0) = 0$ tem uma única solução global $\beta \in C^2((0,+\infty);\mathbb{R}^p) \cap C^1([0,+\infty);\mathbb{R}^p)$.*

*Demonstração.* Ver (Su et al., 2016) página 28. □

**Teorema 14.** *Para toda função $f \in \mathcal{F}_\infty := \cup_{L>0}\mathcal{F}_L$, quando $\epsilon \longrightarrow 0$ tem-se que o método do gradiente acelerado converge para a EDO dada na Equação (3.33) no seguinte sentido. Fixado $T > 0$, tem-se que,*

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

*Demonstração.* Ver (Su et al., 2016) página 32. □

Com bases nas informações dadas na Definição 12 e nos Teoremas 13 e 14 segue-se a dedução da EDO dada na Equação (3.33) a partir do método do gradiente acelerado.

**Lema 15.** *Seja $f \in \mathcal{F}_\infty$ e $\beta_0 \in \mathbb{R}^p$. A sequência gerada pela método do gradiente acelerado dado na Equação (3.25) converge para a curva solução da equação diferencial ordinária de segunda ordem dada na Equação (3.33) na sentido apresentado no Teorema 14.*

*Demonstração.* Primeiramente note que combinando as duas equações do sistema dado na Equação (3.25) se obtêm,

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

Passando $\beta_k$ para o primeiro membro e dividindo tudo por $\sqrt{\epsilon}$ segue-se que,

$$\frac{\beta_{k+1} - \beta_k}{\sqrt{\epsilon}} = \frac{k-1}{k+2}\left(\frac{\beta_k - \beta_{k-1}}{\sqrt{\epsilon}}\right) - \sqrt{\epsilon}\nabla f(\zeta_k). \tag{3.34}$$

Agora, dada a solução $\beta : (0, +\infty) \longrightarrow \mathbb{R}^p$ da EDO dada na Equação (3.33) tem-se pelo Teorema 14 que para $\epsilon > 0$ pequeno,

$$\beta_k \approx \beta(k\sqrt{\epsilon}).$$

Fixando $t > 0$ tal que $k = \frac{t}{\sqrt{\epsilon}}$ segue-se que $\beta(t) \approx \beta_k$, assim como $\beta(t+\sqrt{\epsilon}) \approx \beta_{k+1}$. Tem-se pelo Teorema 13 que a curva $\beta : (0, +\infty) \longrightarrow \mathbb{R}^p$ é de classe $C^2$ e pode-se expandi-la com um polinômio de Taylor de segunda ordem ao redor de $t$ de duas formas:

$$\beta(t + \sqrt{\epsilon}) = \beta(t) + \dot{\beta}(t)\sqrt{\epsilon} + \frac{1}{2}\ddot{\beta}(t)\epsilon + o(\sqrt{\epsilon}),$$

$$\beta(t - \sqrt{\epsilon}) = \beta(t) - \dot{\beta}(t)\sqrt{\epsilon} + \frac{1}{2}\ddot{\beta}(t)\epsilon + o(\sqrt{\epsilon}).$$

Passando $\beta(t)$ para o primeiro membro e dividindo tudo por $\sqrt{\epsilon}$ se obtêm,

$$\frac{\beta(t + \sqrt{\epsilon}) - \beta(t)}{\sqrt{\epsilon}} = \dot{\beta}(t) + \frac{1}{2}\ddot{\beta}(t)\sqrt{\epsilon} + o(\sqrt{\epsilon}),$$

$$\frac{\beta(t)-\beta(t-\sqrt{\epsilon})}{\sqrt{\epsilon}} = \dot{\beta}(t) - \frac{1}{2}\ddot{\beta}(t)\sqrt{\epsilon} + o(\sqrt{\epsilon}).$$

Como,

$$||(\beta_{k+1}-\beta_k) - (\beta(t+\sqrt{\epsilon}) - \beta(t))|| = ||(\beta_{k+1}-\beta_k) - (\beta(k\sqrt{\epsilon}+\sqrt{\epsilon}) - \beta(k\sqrt{\epsilon}))|| =$$

$$= ||(\beta_{k+1}-\beta_k) - (\beta((k+1)\sqrt{\epsilon}) - \beta(k\sqrt{\epsilon}))|| = ||(\beta_{k+1} - \beta((k+1)\sqrt{\epsilon})) - (\beta_k - \beta(k\sqrt{\epsilon}))||$$

$$\leqslant ||(\beta_{k+1} - \beta((k+1)\sqrt{\epsilon}))|| + ||(\beta_k - \beta(k\sqrt{\epsilon}))||,$$

tem-se que,

$$\max_{0\leqslant k\leqslant \frac{t}{\sqrt{\epsilon}}} ||(\beta_{k+1} - \beta((k+1)\sqrt{\epsilon})) - (\beta_k - \beta(k\sqrt{\epsilon}))|| \leqslant \max_{0\leqslant k\leqslant \frac{t}{\sqrt{\epsilon}}} ||(\beta_{k+1} - \beta((k+1)\sqrt{\epsilon}))||$$

$$+ \max_{0\leqslant k\leqslant \frac{t}{\sqrt{\epsilon}}} ||(\beta_k - \beta(k\sqrt{\epsilon}))||.$$

E segue-se pelo Teorema 14, fazendo $\epsilon \longrightarrow 0$ que,

$$\lim_{\epsilon\longrightarrow 0} \max_{0\leqslant k\leqslant \frac{t}{\sqrt{\epsilon}}} ||(\beta_{k+1}-\beta_k) - (\beta((k+1)\sqrt{\epsilon}) - \beta(k\sqrt{\epsilon}))|| = 0,$$

portanto pode-se considerar para $\epsilon$ pequeno que,

$$\frac{\beta_{k+1}-\beta_k}{\sqrt{\epsilon}} \approx \frac{\beta(t+\sqrt{\epsilon})-\beta(t)}{\sqrt{\epsilon}} \tag{3.35}$$

e pelo mesmo argumento também tem-se que,

$$\frac{\beta_k-\beta_{k-1}}{\sqrt{\epsilon}} \approx \frac{\beta(t)-\beta(t-\sqrt{\epsilon})}{\sqrt{\epsilon}}. \tag{3.36}$$

Agora, como $f \in \mathcal{F}_L$, se obtêm que,

$$||\nabla f(\beta_k) - \nabla f(\beta(k\sqrt{\epsilon}))|| \leqslant L||\beta_k - \beta(k\sqrt{\epsilon})||,$$

e segue-se que,

$$\max_{0\leqslant k\leqslant \frac{t}{\sqrt{\epsilon}}} ||\nabla f(\beta_k) - \nabla f(\beta(k\sqrt{\epsilon}))|| \leqslant L \max_{0\leqslant k\leqslant \frac{t}{\sqrt{\epsilon}}} ||\beta_k - \beta(k\sqrt{\epsilon})||.$$

Passando o limite,

$$\lim_{\epsilon\longrightarrow 0} \max_{0\leqslant k\leqslant \frac{t}{\sqrt{\epsilon}}} ||\nabla f(\beta_k) - \nabla f(\beta(k\sqrt{\epsilon}))|| \leqslant L \lim_{\epsilon\longrightarrow 0} \max_{0\leqslant k\leqslant \frac{t}{\sqrt{\epsilon}}} ||\beta_k - \beta(k\sqrt{\epsilon})|| = 0,$$

e pode-se considerar, para $\epsilon$ pequeno que,

$$\nabla f(\beta_k) \approx \nabla f(\beta(t)).$$

Pela segunda equação do método do gradiente acelerado tem-se que,

$$\nabla f(\zeta_k) = \nabla f(\beta_k + \frac{k-1}{k+2}(\beta_k - \beta_{k-1})),$$

e considerando uma expansão de Taylor de primeira ordem para $\beta : (0, +\infty) \longrightarrow \mathbb{R}^p$ segue-se que,

$$\nabla f(\beta_k + \frac{k-1}{k+2}(\beta_k - \beta_{k-1})) \approx \nabla f(\beta(t) + \frac{k-1}{k+2}(\beta(t) - \beta(t - \sqrt{\epsilon}))) =$$

$$\nabla f(\beta(t) + \frac{k-1}{k+2}(-\dot{\beta}(t)\sqrt{\epsilon} + o(\sqrt{\epsilon}))) = \nabla f(\beta(t) - \frac{k-1}{k+2}\dot{\beta}(t)\sqrt{\epsilon} + o(\sqrt{\epsilon})).$$

Fazendo $\epsilon \longrightarrow 0$ tem-se que,

$$\nabla f(\zeta_k) = \lim_{\epsilon \longrightarrow 0} \nabla f(\beta_k + \frac{k-1}{k+2}(\beta_k - \beta_{k-1})) \approx \lim_{\epsilon \longrightarrow 0} \nabla f(\beta(t) - \frac{k-1}{k+2}\dot{\beta}(t)\sqrt{\epsilon} + o(\sqrt{\epsilon})) = \nabla f(\beta(t)).$$

Por fim, multiplicando por $\sqrt{\epsilon}$, segue-se que,

$$\sqrt{\epsilon}\nabla f(\zeta_k) \approx \sqrt{\epsilon}\nabla f(\beta(t)). \tag{3.37}$$

Substituindo na Equação dada em (3.34) as expressões dadas nas Equações (3.35), (3.36) e (3.37) se obtêm,

$$\frac{\beta_{k+1} - \beta_k}{\sqrt{\epsilon}} = \frac{k-1}{k+2}\left(\frac{\beta_k - \beta_{k-1}}{\sqrt{\epsilon}}\right) - \sqrt{\epsilon}\nabla f(\zeta_k) \approx$$

$$\approx \frac{\beta(t + \sqrt{\epsilon}) - \beta(t)}{\sqrt{\epsilon}} = \frac{k-1}{k+2}\left(\frac{\beta(t) - \beta(t - \sqrt{\epsilon})}{\sqrt{\epsilon}}\right) - \sqrt{\epsilon}\nabla f(\beta(t)).$$

Como tem-se que,

$$\frac{k-1}{k+2} = \frac{k+2-2-1}{k+2} = \frac{k+2-3}{k+2} = 1 - \frac{3}{k+2},$$

e lembrando que $k = \frac{t}{\sqrt{\epsilon}}$ segue-se que,

$$1 - \frac{3}{k+2} = 1 - \frac{3}{\frac{t}{\sqrt{\epsilon}} + 2} = 1 - \frac{3\sqrt{\epsilon}}{t + 2\sqrt{\epsilon}}.$$

Pode-se escrever então,

$$\frac{\beta_{k+1} - \beta_k}{\sqrt{\epsilon}} = \frac{k-1}{k+2}\left(\frac{\beta_k - \beta_{k-1}}{\sqrt{\epsilon}}\right) - \sqrt{\epsilon}\nabla f(\zeta_k) \approx$$

$$\approx \frac{\beta(t + \sqrt{\epsilon}) - \beta(t)}{\sqrt{\epsilon}} = \left(1 - \frac{3\sqrt{\epsilon}}{t + 2\sqrt{\epsilon}}\right)\left(\frac{\beta(t) - \beta(t - \sqrt{\epsilon})}{\sqrt{\epsilon}}\right) - \sqrt{\epsilon}\nabla f(\beta(t)).$$

Com base nas expansões de Taylor de segunda ordem que foram feitas anteriormente tem-se que,

$$\frac{\beta_{k+1}-\beta_k}{\sqrt{\epsilon}} = \frac{k-1}{k+2}\left(\frac{\beta_k-\beta_{k-1}}{\sqrt{\epsilon}}\right) - \sqrt{\epsilon}\nabla f(\zeta_k) \approx$$

$$\approx \left(\dot{\beta}(t) + \frac{1}{2}\ddot{\beta}(t)\sqrt{\epsilon} + o(\sqrt{\epsilon})\right) =$$

$$\left(1 - \frac{3\sqrt{\epsilon}}{t+2\sqrt{\epsilon}}\right)\left(\dot{\beta}(t) - \frac{1}{2}\ddot{\beta}(t)\sqrt{\epsilon} + o(\sqrt{\epsilon})\right) - \sqrt{\epsilon}\nabla f(\beta(t)).$$

Fazendo algumas manipulações algébricas se obtêm,

$$\left(\dot{\beta}(t) + \frac{1}{2}\ddot{\beta}(t)\sqrt{\epsilon} + o(\sqrt{\epsilon})\right) =$$

$$\left(\dot{\beta}(t) - \frac{1}{2}\ddot{\beta}(t)\sqrt{\epsilon} + o(\sqrt{\epsilon})\right) - \frac{3\sqrt{\epsilon}}{t+2\sqrt{\epsilon}}\left(\dot{\beta}(t) - \frac{1}{2}\ddot{\beta}(t)\sqrt{\epsilon} + o(\sqrt{\epsilon})\right) - \sqrt{\epsilon}\nabla f(\beta(t))$$

$$\Rightarrow \ddot{\beta}(t)\sqrt{\epsilon} = -\frac{3\sqrt{\epsilon}}{t+2\sqrt{\epsilon}}\left(\dot{\beta}(t) - \frac{1}{2}\ddot{\beta}(t)\sqrt{\epsilon} + o(\sqrt{\epsilon})\right) - \sqrt{\epsilon}\nabla f(\beta(t))$$

$$\Rightarrow \ddot{\beta}(t)\sqrt{\epsilon} = -\frac{3\sqrt{\epsilon}}{t+2\sqrt{\epsilon}}\dot{\beta}(t) + \frac{3\sqrt{\epsilon}}{t+2\sqrt{\epsilon}}\frac{1}{2}\ddot{\beta}(t)\sqrt{\epsilon} + \frac{3\sqrt{\epsilon}}{t+2\sqrt{\epsilon}}o(\sqrt{\epsilon}) - \sqrt{\epsilon}\nabla f(\beta(t)).$$

Dividindo todos os membros por $\sqrt{\epsilon}$ segue-se que,

$$\ddot{\beta}(t) = -\frac{3}{t+2\sqrt{\epsilon}}\dot{\beta}(t) + \frac{3}{t+2\sqrt{\epsilon}}\frac{1}{2}\ddot{\beta}(t)\sqrt{\epsilon} + \frac{3}{t+2\sqrt{\epsilon}}o(\sqrt{\epsilon}) - \nabla f(\beta(t)).$$

Fazendo $\epsilon \longrightarrow 0$ tem-se que,

$$\ddot{\beta}(t) = -\frac{3}{t}\dot{\beta}(t) - \nabla f(\beta(t)).$$

E conclui-se que,

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

□

Pelo resultado do Lema 15 pode-se estudar o comportamento da sequência gerada na Equação dada em (3.25) através da curva solução da EDO dada na Equação (3.33) o que justifica a primeira análise qualitativa simplificada que foi feita anteriormente. Como uma segunda ilustração disso, pode-se fazer o seguinte estudo sobre o comportamento assintótico inicial da curva solução da EDO dada na Equação (3.33).

Considere que o limite $\lim_{t\longrightarrow 0}\ddot{\beta}(t)$ existe. Pelo teorema do valor médio tem-se que existe $\xi \in (0,t)$ tal que, $\dot{\beta}(t)-\dot{\beta}(0)=\ddot{\beta}(\xi)t$. Como $\dot{\beta}(0)=0$ pode-se escrever $\dot{\beta}(t)/t=\ddot{\beta}(\xi)$ e pela EDO dada na Equação (3.25) segue-se que,

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

Tomando o limite $t\longrightarrow 0$ na expressão acima se obtêm $\ddot{\beta}(0)=-\nabla f(\beta_0)/4$. Considere agora um expansão de Taylor de segunda ordem para a curva solução da EDO dada na Equação (3.33) na vizinhança de 0, segue-se que,

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

e substituindo os valores de $\dot{\beta}(0)$ e $\ddot{\beta}(0)$ na expressão acima se obtêm,

$$\beta(t)=-\frac{\nabla f(\beta_0)t^2}{8}+\beta_0+o(t^2), \tag{3.38}$$

então para valores pequenos de $t$ pode-se estudar o comportamento das curvas solução da EDO dada na Equação (3.33) pela expressão dada na Equação (3.38). Outro contexto onde é possível obter uma forma explícita para a curva solução $\beta(t)$ é quando a função $f$ é quadrática e nesse caso a equação diferencial ordinária dada na Equação (3.33) torna-se a equação diferencial ordinária de Bessel de ordem 1 e pode-se obter a curva solução $\beta(t)$ 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 é apresentado a demonstração que a taxa de convergência das curvas solução da EDO dada na Equação (3.33) é da ordem de $\mathcal{O}\big(1/t^2\big)$. Esse resultado em conjunto com os resultados alcançados nas subseções anteriores possibilitam a construção de uma interpretação que comece dar significado ao procedimento desenvolvido por Nesterov. Uma discussão mais detalhada é realizado após a demonstração do lema abaixo.

**Lema 16.** *Para todo $f\in\mathcal{F}_\infty$, seja $\beta:[0,+\infty)\longrightarrow\mathbb{R}^p$ a única solução global para a EDO dada na Equação (3.33) com condições iniciais $\beta(0)=\beta_0$ e $\dot{\beta}(0)=0$. Então para todo $t>0$ tem-se que,*

$$f(\beta(t))-f(\beta_*)\leqslant\mathcal{O}\left(\frac{1}{t^2}\right). \tag{3.39}$$

*Demonstração.* Considere $\mathcal{E} : [0, +\infty) \longrightarrow \mathbb{R}$ uma função diferenciável definida por,

$$\mathcal{E}(t) = t^2(f(\beta(t)) - f(\beta_*)) + 2||\beta(t) + \frac{t}{2}\dot{\beta}(t) - \beta_*||^2. \tag{3.40}$$

Derivando $\mathcal{E}$ se obtêm que,

$$\dot{\mathcal{E}}(t) = 2t(f(\beta(t))-f(\beta_*))+t^2\nabla f(\beta(t))^T\dot{\beta}(t)+4\big(\beta(t)+\frac{t}{2}\dot{\beta}(t)-\beta_*\big)^T\big(\frac{3}{2}\dot{\beta}(t)+\frac{t}{2}\ddot{\beta}(t)\big). \tag{3.41}$$

Como a curva $\beta(t)$ é solução para a EDO dada na Equação (3.33) e $\frac{3}{2}\dot{\beta}(t) + \frac{t}{2}\ddot{\beta}(t) = -\frac{t}{2}\nabla f(\beta(t))$ tem-se que,

$$\dot{\mathcal{E}}(t) = 2t(f(\beta(t)) - f(\beta_*)) - 2t\nabla f(\beta(t))^T\big(\beta(t) - \beta_*\big). \tag{3.42}$$

Dado que $f$ é uma função convexa e diferenciável segue-se que,

$$0 \geqslant f(\beta(t)) - f(\beta_*) - \nabla f(\beta(t))^T\big(\beta(t) - \beta_*\big).$$

E se obtêm que,

$$\dot{\mathcal{E}}(t) \leqslant 0. \tag{3.43}$$

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

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

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

□

Com os resultados acima tem-se que, fazendo a discretização $t = \sqrt{\epsilon}k$ no domínio da equação diferencial ordinária de segunda ordem dada na Equação (3.33), se obtêm, via o método do gradiente acelerado dado na Equação (3.25), uma discretização que preserva a taxa de convergência dessa EDO. A técnica de Nesterov pode ser aplicada a muitos outros algoritmos clássicos, permitindo que esses 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 muito problemas, como, por exemplo, a estimação de parâmetros em modelos lineares generalizados.

## 3.4 Gradiente acelerado de alta ordem

Nesta subseção são apresentados os resultados mais importantes do ponto de vista teórico para o que é desenvolvido nas próximas seções. Para que a apresentação dos resultados seja a mais natural possível busca-se construir o método do gradiente acelerado de alta ordem de modo sistemático e progressivo.

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

Em seguida, define-se um funcional lagrangiano no espaço de fase $\mathbb{R}^p \times \mathbb{R}^p \times [0, +\infty)$, onde toma-se 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 deseja-se inicializar o método.

Por fim, discretiza-se essa EDO com o método de Euler progressivo e regressivo e é introduzido uma sequência auxiliar que possui algumas propriedades, esse 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.

Esse procedimento é comumente utilizado nas ciências da natureza. Primeiro observa-se um fenômeno repetidas vezes até que identifique-se algum padrão, no caso aqui trabalhado, a relação que mostra a compatibilidade entre as taxas de convergência do método de Nesterov e da EDO associada a ele. Esse processo parte da multiplicidade dos fenômenos em direção a alguns primeiros princípios que explique a multiplicidade do que foi observado e depois a partir desses primeiros princípios, no caso um grande primeiro princípio, a saber, a lagrangiana de Bregman, tenta-se chegar novamente a multiplicidade dos fenômenos

observados, mas dessa vez dando a eles uma interpretação que possa esclarecê-los. Esse é o caso, por exemplo, da mecânica lagrangiana em que os fenômenos físicos são deduzidos a partir do princípio da mínima ação.

### *3.4.1 EDO de segunda ordem associada*

Como afirmado na introdução desta seção, a divergência de Bregman pode ser interpretada como uma medida fraca em um espaço munido com uma estrutura gerada por uma função convexa e diferenciável. Como será visto na seção de aplicações, essa medida apresenta vantagens no contexto da análise de dados em relação a uma métrica Riemanniana, pois não impõe uma estrutura de simetria no espaço onde esta definida. Formalmente tem-se a seguinte definição,

**Definição 17.** *Considere $h : \mathbb{R}^p \longrightarrow \mathbb{R}$ uma função convexa e diferenciável. A função $D_h : \mathbb{R}^p \times \mathbb{R}^p \longrightarrow \mathbb{R}$ definida por,*

$$D_h(\hat{\beta} : \bar{\beta}) = h(\hat{\beta}) - h(\bar{\beta}) - \nabla h(\bar{\beta})^T(\hat{\beta} - \bar{\beta}), \tag{3.44}$$

*para todo $\hat{\beta}, \bar{\beta} \in \mathbb{R}^p$ é chamada de divergência de Bregman.*

Uma propriedade imediata que decorre da convexidade e diferenciabilidade da função $h$ é que para todo $\bar{\beta}, \hat{\beta} \in \mathbb{R}^p$ tem-se que $D_h(\hat{\beta} : \bar{\beta}) \geq 0$, outra propriedade importante é que a divergência de Bregman é uma função convexa na primeira variável.

Para tornar a apresentação dos resultados mais intuitiva é adotado algumas terminologias oriundas da mecânica lagrangiana. Seja $f : \mathbb{R}^p \longrightarrow \mathbb{R}$ uma função convexa e diferenciável. Deseja-se 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 definida na subseção anterior, $(\mathbb{R}^p, D_h)$, interpreta-se $D_h$ como a energia cinética. Com essas duas funções define-se o funcional $\mathcal{L}_B : \mathbb{R}^p \times \mathbb{R}^p \times [0, +\infty) \longrightarrow \mathbb{R}$ no espaço de fase $\mathbb{R}^p \times \mathbb{R}^p \times [0, +\infty)$ do seguinte modo.

**Definição 18.** *Considere o espaço $(\mathbb{R}^p, D_h)$ e seja $f : \mathbb{R}^p \longrightarrow \mathbb{R}$ uma função convexa e diferenciável. Considere $\alpha, \gamma, \eta : [0, +\infty) \longrightarrow \mathbb{R}^p$ funções diferenciáveis dadas a priori. A lagrangiana de Bregman é definida por,*

$$\mathcal{L}_B(t, \beta, \dot{\beta}) = e^{\alpha(t)+\gamma(t)}\big(D_h(\beta + e^{-\alpha(t)}\dot{\beta} : \beta) - e^{-\eta(t)}f(\beta)\big). \tag{3.45}$$

As funções $\alpha, \gamma$ e $\eta$ são tratadas como parâmetros a serem introduzidos no sistema e agem como a resistência do meio para a dinâmica definida por essa lagrangiana. Considera-se que essas funções satisfazem as seguintes relações:

$$\begin{aligned} \dot{\eta}(t) &\leqslant e^{\alpha(t)}, \\ \dot{\gamma}(t) &= e^{\alpha(t)}. \end{aligned} \tag{3.46}$$

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

Recorrendo uma vez mais a conceitos da mecânica lagrangiana considere que a lagrangiana de Bregman define um sistema de partículas no espaço $(\mathbb{R}^p, D_h)$ e que a dinâmica desse sistema é regida pelo princípio da mínima ação, isto é, que a dinâmica do sistema é obtida quando minimiza-se a integral da lagrangiana sobre o espaço de trajetórias - curvas - desse sistema. Define-se então $\mathcal{S} := \{\beta \in C^2([0, +\infty); \mathbb{R}^p)\}$ como o espaço das curvas de $[0, +\infty)$ em $\mathbb{R}^p$ de classe $C^2$ que dão as trajetórias do sistema de partículas, segue-se que o funcional $\Phi : \mathcal{S} \longrightarrow \mathbb{R}$ dado por,

$$\Phi(\beta) := \int_{[0,+\infty)} \mathcal{L}_B(t, \beta(t), \dot{\beta}(t))dt, \tag{3.47}$$

ao ser minimizado sobre o espaço de curvas $\mathcal{S}$ retorna a dinâmica do sistema. Do cálculo variacional tem-se que uma condição necessária para que uma determinada curva em $\mathcal{S}$ seja minimizante para a Equação (3.47) é que essa curva satisfaça a equação de Euler-Lagrange,

$$\frac{d}{dt}\frac{\partial \mathcal{L}_B}{\partial \dot{\beta}}(t, \beta, \dot{\beta}) = \frac{\partial \mathcal{L}_B}{\partial \beta}(t, \beta, \dot{\beta}). \tag{3.48}$$

Desenvolvendo a lagrangiana de Bregman dada na Equação (3.45) em relação a equação de Euler-Lagrange dada na Equação (3.48) se obtêm a seguinte equação diferencial ordinária de segunda ordem,

$$\begin{aligned} &\ddot{\beta}(t) + \left(e^{\alpha(t)} - \dot{\alpha}(t)\right)\dot{\beta}(t) \\ &+e^{2\alpha(t)+\eta(t)}\left[\nabla^2 h(\beta(t) + e^{-\alpha(t)}\dot{\beta}(t))\right]^{-1}\nabla f(\beta(t)) \\ &+e^{\alpha(t)}\left(\dot{\gamma}(t) - e^{\alpha(t)}\right)\left[\nabla^2 h(\beta(t) + e^{-\alpha(t)}\dot{\beta}(t))\right]^{-1}\nabla h(\beta(t) + e^{-\alpha(t)}\dot{\beta}(t)) \\ &-e^{\alpha(t)}\left(\dot{\gamma}(t) - e^{\alpha(t)}\right)\left[\nabla^2 h(\beta(t) + e^{-\alpha(t)}\dot{\beta}(t))\right]^{-1}\nabla h(\beta(t)) = 0. \end{aligned} \tag{3.49}$$

Como pode-se notar, a EDO acima tem uma estrutura consideravelmente complexa e difícil de trabalhar. Por isso, será imposto sobre as funções $\alpha$, $\gamma$ e $\eta$ as condições expostas nas Equações (3.46) e como consequência se obtêm a seguinte EDO,

$$\ddot{\beta}(t) + \left(e^{\alpha(t)} - \dot{\alpha}(t)\right)\dot{\beta}(t) + e^{2\alpha(t)+\eta(t)}\left[\nabla^2 h(\beta(t) + e^{-\alpha(t)}\dot{\beta}(t))\right]^{-1}\nabla f(\beta(t)) = 0. \tag{3.50}$$

Na Equação (3.50) está se assumindo que a matriz hessiana $\nabla^2 h$ é inversível, entretanto pode-se reescrever a Equação (3.50) de modo que ela seja apenas diferenciável,

$$\frac{d}{dt}\nabla h(\beta(t) + e^{-\alpha(t)}\dot{\beta}(t)) = -e^{\alpha(t)+\eta(t)}\nabla f(\beta(t)). \tag{3.51}$$

Antes de prosseguir as análises faz-se necessário um aprofundamento um pouco maior sobre a lagrangiana de Bregman. Ela tem diversas propriedades, a mais notável, é que essa função é fechada em relação a dilatações temporais. Isso significa que se $\beta : [0, +\infty) \longrightarrow \mathbb{R}^p$ é solução da equação de Euler-Lagrange apresentada na Equação (3.51), então a sua reparametrização $\beta_R : [0, +\infty) \longrightarrow \mathbb{R}^p$, 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 $\tau : [0, +\infty) \longrightarrow [0, +\infty)$ uma função crescente e diferenciável. Dado uma curva $\beta : [0, +\infty) \longrightarrow \mathbb{R}^p$, considere a curva $\beta_R : [0, +\infty) \longrightarrow \mathbb{R}^p$ definida por,

$$\beta_R(t) = \beta(\tau(t)). \tag{3.52}$$

Isto é, a curva $\beta_R$ é obtida da curva original $\beta$ por uma contração ou dilatação temporal. No próximo teorema é provado um resultado fundamental para compreensão desta subseção.

**Teorema 19.** *Se a curva $\beta : [0, +\infty) \longrightarrow \mathbb{R}^p$ satisfaz a equação de Euler-Lagrange na Equação (3.49) para a lagrangiana de Bregman com parâmetros $\alpha$, $\gamma$ e $\eta$. Então a curva dada na Equação (3.52) satisfaz a equação de Euler-Lagrange para a lagrangiana de Bregman com os parâmetros,*

$$\begin{aligned} \bar{\alpha}(t) &= \alpha(\tau(t)) + \log(\dot{\tau}(t)), \\ \bar{\eta}(t) &= \eta(\tau(t)), \\ \bar{\gamma}(t) &= \gamma(\tau(t)), \end{aligned} \tag{3.53}$$

*além disso $\bar{\alpha}$, $\bar{\eta}$ e $\bar{\gamma}$ satisfazem as condições dadas nas Equações (3.46) se, e somente se $\alpha$, $\eta$ e $\gamma$ também satisfazem.*

*Demonstração.* A velocidade e a aceleração da curva reparametrizada $\beta_R(t) = \beta(\tau(t))$ são dadas por,

$$\begin{aligned} \dot{\beta_R}(t) &= \dot{\tau}(t)\dot{\beta}(\tau(t)), \\ \ddot{\beta_R}(t) &= \ddot{\tau}(t)\dot{\beta}(\tau(t)) + \dot{\tau}(t)^2\ddot{\beta}(\tau(t)), \end{aligned}$$

e invertendo essas relações se obtêm que,

$$\begin{aligned} \dot{\beta}(\tau(t)) &= \tfrac{1}{\dot{\tau}(t)}\dot{\beta_R}(t), \\ \ddot{\beta}(\tau(t)) &= \tfrac{1}{\dot{\tau}(t)^2}\ddot{\beta_R}(t) - \tfrac{\ddot{\tau}(t)}{\dot{\tau}(t)^3}\dot{\beta_R}(t). \end{aligned} \tag{3.54}$$

Por hipótese a curva original $\beta(t)$ satisfaz a equação de Euler-Lagrange na Equação (3.49) para a lagrangiana de Bregman com parâmetros $\alpha$, $\gamma$ e $\eta$, então para o tempo $\tau(t)$ tem-se que,

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

Usando as relações definidas nas Equações (3.54), multiplicando por $\dot{\tau}(t)^2$ e simplificando os termos se obtêm que,

$$\begin{gathered}\ddot{\beta}_R(t) + \big(\dot{\tau}(t)e^{\alpha(\tau(t))} - \dot{\tau}(t)\dot{\alpha}(\tau(t)) - \tfrac{\ddot{\tau}(t)}{\dot{\tau}(t)}\big)\dot{\beta}_R(t) \\ +\dot{\tau}(t)^2 e^{2\alpha(\tau(t))+\eta(\tau(t))}\big[\nabla^2 h(\beta_R(t) + \tfrac{e^{-\alpha(\tau(t))}}{\dot{\tau}(t)}\dot{\beta}_R(t)\big]^{-1}\nabla f(\beta_R(t)) \\ +\dot{\tau}(t)^2 e^{\alpha(\tau(t))}\big(\dot{\gamma}(\tau(t)) - e^{\alpha(\tau(t))}\big)\big[\nabla^2 h(\beta_R(t) + \tfrac{e^{-\alpha(\tau(t))}}{\dot{\tau}(t)}\dot{\beta}_R(t)\big]^{-1}\nabla h(\beta_R(t) + \tfrac{e^{-\alpha(\tau(t))}}{\dot{\tau}(t)}\dot{\beta}_R(t)) \\ -\dot{\tau}(t)^2 e^{\alpha(\tau(t))}\big(\dot{\gamma}(\tau(t)) - e^{\alpha(\tau(t))}\big)\big[\nabla^2 h(\beta_R(t) + \tfrac{e^{-\alpha(\tau(t))}}{\dot{\tau}(t)}\dot{\beta}_R(t)\big]^{-1}\nabla h(\beta_R(t)) = 0.\end{gathered}$$

Agora, com a definição dos parâmetros modificados dados nas Equações (3.53) pode-se reescrever a equação acima como,

$$\begin{gathered}\ddot{\beta}_R(t) + \big(e^{\bar{\alpha}(t)} - \dot{\bar{\alpha}}(t)\big)\dot{\beta}_R(t) \\ +e^{2\bar{\alpha}(t)+\bar{\eta}(t)}\big[\nabla^2 h(\beta_R(t) + e^{-\bar{\alpha}(t)}\dot{\beta}_R(t))\big]^{-1}\nabla f(\beta_R(t)) \\ +e^{\bar{\alpha}(t)}\big(\dot{\bar{\gamma}}(t) - e^{\bar{\alpha}(t)}\big)\big[\nabla^2 h(\beta_R(t) + e^{-\bar{\alpha}(t)}\dot{\beta}_R(t))\big]^{-1}\nabla h(\beta_R(t) + e^{-\bar{\alpha}(t)}\dot{\beta}_R(t)) \\ -e^{\bar{\alpha}(t)}\big(\dot{\bar{\gamma}}(t) - e^{\bar{\alpha}(t)}\big)\big[\nabla^2 h(\beta_R(t) + e^{-\bar{\alpha}(t)}\dot{\beta}_R(t))\big]^{-1}\nabla h(\beta_R(t)) = 0,\end{gathered} \tag{3.55}$$

que é a equação de Euler-Lagrange para a lagrangiana de Bregman com os parâmetros dados nas Equações (3.53). Além disso, considerando que $\alpha$, $\eta$ e $\gamma$ satisfazem as condições dadas nas Equações (3.46), então segue-se que,

$$\dot{\bar{\eta}}(t) = \frac{d}{dt}\eta(\tau(t)) = \dot{\tau}(t)\dot{\eta}(\tau(t)) \leqslant \dot{\tau}(t)e^{\alpha(\tau(t))} = e^{\alpha(\tau(t))+\log(\dot{\tau}(t))} = e^{\bar{\alpha}(t)},$$

$$\dot{\bar{\gamma}}(t) = \frac{d}{dt}\gamma(\tau(t)) = \dot{\tau}(t)\dot{\gamma}(\tau(t)) = \dot{\tau}(t)e^{\alpha(\tau(t))} = e^{\alpha(\tau(t))+\log(\dot{\tau}(t))} = e^{\bar{\alpha}(t)},$$

o que significa que os parâmetros dados nas Equações (3.53) também satisfazem as condições dadas nas Equações (3.46). Para recíproca basta considerar $\tau^{-1}(t)$ no lugar de $\tau(t)$. □

O resultado estabelecido no Teorema 19 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 na Equação (3.51). Note que esse resultado afirma que é possível se obter taxas de convergência exponenciais, entretanto nessa dissertação será abordada apenas a convergência polinomial, pois no caso exponencial, o procedimento de discretização da EDO na Equação (3.51) 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).

**Teorema 20.** *A taxa de convergência das soluções da Equação (3.51) é da ordem de,*

$$f(\beta(t)) - f(\beta_*) \leqslant \mathcal{O}\big(e^{-\eta(t)}\big). \tag{3.56}$$

*Demonstração.* Considere a função $\mathcal{E} : [0, +\infty) \longrightarrow \mathbb{R}$ definida por,

$$\mathcal{E}(t) = D_h(\beta_* : \beta(t) + e^{-\alpha(t)}\dot{\beta}(t)) + e^{\eta(t)}\big(f(\beta(t)) - f(\beta_*)\big). \tag{3.57}$$

Derivando a Equação (3.57) se obtêm que,

$$\dot{\mathcal{E}}(t) = -\frac{d}{dt}\nabla h(\beta(t) + e^{-\alpha(t)}\dot{\beta}(t))^T\big(\beta_* - \beta(t) - e^{-\alpha(t)}\dot{\beta}(t)\big) + \dot{\eta}(t)e^{\eta(t)}\big(f(\beta(t)) - f(\beta_*)\big)$$

$$+e^{\eta(t)}\nabla f(\beta(t))^T\dot{\beta}(t).$$

Se a curva $\beta : [0, +\infty) \longrightarrow \mathbb{R}^p$ satisfaz a equação de Euler-Lagrange na Equação (3.51). Então a expressão acima simplifica-se 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 $D_f(\beta_* : \beta(t)) = f(\beta_*) - f(\beta(t)) - \nabla f(\beta(t))^T\big(\beta_* - \beta(t)\big)$ é a divergência de Bregman em relação a função $f$. Como $f$ é uma função convexa e diferenciável, tem-se que $D_f(\beta_* : \beta(t)) \geqslant 0$ e portanto o primeiro termo em $\dot{\mathcal{E}}$ é negativo. Dadas as propriedades apresentadas nas Equações (3.46) segue-se que 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 se obtêm para todo $t \geqslant 0$ que,

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

□

Através da escolha das funções $\alpha$, $\gamma$ e $\eta$ pode-se selecionar subfamílias de lagrangianas de Bregman. O caso em que tem-se interesse é o da subfamília polinomial, para isso, dado $C > 0$ constante e $m \geqslant 2$ tal que $f \in C^{m-1}\big(\mathbb{R}^p; \mathbb{R}\big)$ tem-se que:

$$\begin{aligned} \alpha(t) &= \log(m) - \log(t), \\ \eta(t) &= m\log(t) + \log(C), \\ \gamma(t) &= m\log(t). \end{aligned} \tag{3.58}$$

As funções nas Equações (3.58) satisfazem as condições apresentadas nas Equações (3.46) com uma igualdade na primeira condição. Dada as funções nas Equações (3.58) tem-se que a EDO de segunda ordem resultante da EDO na Equação (3.50) é dada por,

$$\ddot{\beta}(t) + \frac{m+1}{t}\dot{\beta}(t) + Cm^2t^{m-2}\big[\nabla^2 h\big(\beta(t) + \frac{t}{m}\dot{\beta}(t)\big)\big]^{-1}\nabla f(\beta(t)) = 0. \tag{3.59}$$

Pelo Teorema 20 segue-se que as curvas solução da Equação (3.59) tem taxa de convergência da ordem de $\mathcal{O}\big(1/t^m\big)$. Tem-se como consequência direta do Teorema 19 que, por exemplo, dado uma curva base, no caso $m = 2$, todas as demais curvas podem ser obtidas pela dilatação temporal $\tau(t) = t^{\frac{m}{2}}$ e também que, quando a matriz $\nabla^2 h$ é identidade se obtêm a EDO na Equação (3.33). Na próxima subseção é apresentado o processo de discretização da EDO na Equação (3.59) que garante a compatibilidade das taxas de convergência entre o caso contínuo e discreto.

### *3.4.3 Taxa de convergência*

O processo de discretização de uma EDO sempre constitui-se um desafio, principalmente quando deseja-se que a discretização obtida preserve certas propriedades do caso contínuo. Começa-se o processo de discretização da EDO na Equação (3.59) pela transformação desta em uma sistema de EDOs de primeira ordem. Para isso, considere-se a variável auxiliar - curva auxiliar - $Z : [0, +\infty) \longrightarrow \mathbb{R}^p$ definida por,

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

Com base nessa variável se obtêm o sistema de EDOs de primeira ordem dado abaixo:

$$\begin{aligned} Z(t) &= \beta(t) + (t/m)\dot{\beta}(t), \\ \tfrac{d}{dt}\nabla h(Z(t)) &= -Cmt^{m-1}\nabla f(\beta(t)). \end{aligned} \tag{3.60}$$

Passa-se então a discretização dos domínios das curvas $Z(t)$ e $\beta(t)$. Dado $\delta > 0$ define-se a sequência $t_k = \delta k$ e com base nela se define as sequências $\{\beta_k\}_{k\geqslant 0}$ e $\{Z_k\}_{k\geqslant 0}$ do seguinte modo:

$$\beta_k = \beta(t_k),$$

$$Z_k = Z(t_k).$$

Para as derivadas de $\beta(t)$ e $Z(t)$ aplica-se o método de Euler progressivo e se obtêm,

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

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

Aplicando essas duas discretizações a primeira equação em (3.60) segue-se que,

$$\beta_{k+1} = \frac{m}{k}Z_k + \frac{k-m}{k}\beta_k. \tag{3.61}$$

Analogamente aplicando o método de Euler regressivo a segunda equação em (3.60) se obtêm 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,

$$Z_k = \arg\min_Z \{Cmk^{m-1}\nabla f(\beta_k)^T Z + (1/\epsilon)D_h(Z : Z_{k-1})\}, \tag{3.62}$$

com $\epsilon = \delta^m$. A princípio as sequências nas Equações (3.61) e (3.62) definem um algoritmo que implementa a dinâmica da EDO na Equação (3.60) em tempo discreto. Contudo não é possível estabelecer uma taxa de convergência para esse algoritmo e de fato, empiricamente, encontra-se que esse algoritmo é instável. Para transformar o algoritmo definido pelas sequências nas Equações (3.61) e (3.62) em um algoritmo que, de fato, reproduza a dinâmica na Equação (3.60) em tempo discreto faz-se necessária a introdução de uma sequência auxiliar que possui algumas propriedades. Começa-se definindo uma sequência $\{\zeta_k\}_{k\geqslant 0}$ e substitui-se nas sequências nas Equações (3.61) e (3.62) o termo $\beta_k$ por $\zeta_k$ obtendo assim o algoritmo abaixo,

$$\begin{aligned} \beta_{k+1} &= [(m/(k+m)]Z_k + [k/(k+m)]\zeta_k, \\ Z_k &= \arg\min_Z \Big\{Cmk^{(m-1)}\nabla f(\zeta_k)^T Z + (1/\epsilon)D_h(Z : Z_{k-1})\Big\}. \end{aligned} \tag{3.63}$$

em que $k^{(m-1)} = k(k+1)\dots(k+m-2)$ é o fatorial crescente. Afirma-se que uma condição suficiente para que o algoritmo na Equação (3.63) tenha uma taxa de convergência de $\mathcal{O}\big(1/\epsilon k^m\big)$ é que a sequência $\{\zeta_k\}_{k\geqslant 0}$ satisfaça a desigualdade,

$$\nabla f(\zeta_k)^T\big(\beta_k - \zeta_k\big) \geqslant M\epsilon^{\frac{1}{m-1}}||\nabla f(\zeta_k)||^{\frac{m}{m-1}}, \tag{3.64}$$

para alguma constante $M > 0$. Note-se que algumas modificações foram feitas das sequências definidas nas Equações (3.61) e (3.62) para a sequência definida na Equação (3.63). Houve uma substituição do peso $m/k$ por $m/(k+m)$, essa mudança é somente para simplificar os cálculos que se seguirão, pois não alteram o comportamento assintótico da sequência, dado que $m/k = \mathcal{O}\big(m/(k+m)\big)$ quando $k \longrightarrow \infty$. Analogamente houve a substituição de $k^{m-1}$ pelo fatorial crescente $k^{(m-1)}$ em que também tem-se que $k^{(m-1)} = \mathcal{O}\big(k^{m-1}\big)$ quando $k \longrightarrow \infty$.

Para o próximo resultado necessita-se 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 21.** *Seja* $h : \mathbb{R}^p \longrightarrow \mathbb{R}$ *uma função convexa e diferenciável.* $h$ *é uma função* $\sigma$*-uniformemente convexa de ordem* $m \geqslant 2$ *se para todo* $\hat{\beta}, \bar{\beta} \in \mathbb{R}^p$ *ocorre que,*

$$D_h(\hat{\beta} : \bar{\beta}) \geqslant \frac{\sigma}{m}||\hat{\beta} - \bar{\beta}||^m. \tag{3.65}$$

Também será necessário o seguinte resultado auxiliar expresso no lema abaixo.

**Lema 22.** *Considere* $f : \mathbb{R}^p \longrightarrow \mathbb{R}$ *uma função convexa e diferenciável e* $h : \mathbb{R}^p \longrightarrow \mathbb{R}$ *uma função 1-uniformemente convexa de ordem* $m \geqslant 2$ *e diferenciável. Dado* $\epsilon > 0$. *Seja* $\{\zeta_k\}_{k\geqslant 0}$ *uma sequência que satisfaz a desigualdade na Equação (3.64) e considere* $\{\psi_k\}_{k\geqslant 0}$ *uma sequência de funções* $\psi_k : \mathbb{R}^p \longrightarrow \mathbb{R}$ *definidas tais que,*

$$\psi_k(\beta) = Cm\sum_{i=0}^{k} i^{(m-1)}\big[f(\zeta_i) + \nabla f(\zeta_i)^T\big(\beta - \zeta_i\big)\big] + (1/\epsilon)D_h(\beta : \beta_0), \tag{3.66}$$

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

$$\psi_k(Z_k) \geqslant Ck^{(m)}f(\zeta_k), \tag{3.67}$$

*em que* $\{Z_k\}_{k\geqslant 0}$ *é a sequência gerada pelo algoritmo na Equação (3.63).*

Para dar suporte a algumas passagens na demonstração do Lema 22 segue-se abaixo as seguintes observações: As funções $\psi_k$ para todo $k \geqslant 0$ são funções convexas. De fato, as aplicações $\beta \longmapsto f(\zeta_i) + \nabla f(\zeta_i)^T\big(\beta - \zeta_i\big)$ são convexas para todo $i \in \{0, ..., k\}$, pois elas são funções afins na variável $\beta$ e dado que a multiplicação de uma função convexa por uma contante positiva e a soma de funções convexas tem como resultado uma função convexa, segue-se que a aplicação,

$$\beta \longmapsto Cm \sum_{i=0}^{k} i^{(m-1)} \big[f(\zeta_i) + \nabla f(\zeta_i)^T\big(\beta - \zeta_i\big)\big],$$

é uma função conexa. Por fim tem-se que a aplicação $\beta \longmapsto \frac{1}{\epsilon} D_h(\beta : \beta_0)$ é convexa. Isso decorre do fato da divergência de Bregman ser convexa na primeira variável. Somando essas duas aplicações expostas acima tem-se que as funções $\psi_k$ são funções convexas para todo $k \geqslant 0$.

A segunda observação é que a segunda sequência do algoritmo na Equação (3.63) tem condição de otimalidade expressa por,

$$\nabla h(Z_k) = \nabla h(Z_{k-1}) - \epsilon Cm k^{(m-1)} \nabla f(\zeta_k).$$

Fazendo a mudança de variável de $k$ para $i$ e somando os termos de $i = 1$ até $i = k$ se obtêm pela propriedade das somas telescópicas que,

$$\nabla h(Z_k) - \nabla h(Z_0) = \sum_{i=1}^{k} \nabla h(Z_i) - \nabla h(Z_{i-1}) = -\epsilon Cm \sum_{i=0}^{k} i^{(m-1)} \nabla f(\zeta_i) \Rightarrow$$

$$\nabla h(Z_k) = \nabla h(Z_0) - \epsilon Cm \sum_{i=0}^{k} i^{(m-1)} \nabla f(\zeta_i).$$

Como $Z_0 = \beta_0$ tem-se da sequência de funções definidas na Equação (3.66) que,

$$\nabla \psi_k(Z_k) = Cm \sum_{i=0}^{k} i^{(m-1)} \nabla f(\zeta_i) + (1/\epsilon)\big(\nabla h(Z_k) - \nabla h(\beta_0)\big) = 0$$

e dado que $\psi_k$ é uma função convexa para todo $k \geqslant 0$ segue-se que $Z_k$ é, de fato, o minimizador de $\psi_k$ e pode-se escrever que,

$$Z_k = \arg\min_{Z} \big\{\psi_k(Z)\big\}. \tag{3.68}$$

Por fim a última observação será apresentada na forma de lema abaixo.

**Lema 23.** *Se $h$ é uma função 1-uniformemente convexa de ordem $m \geqslant 2$, então as funções $\psi_k$ são, para todo $k \geqslant 0$, funções $1/\epsilon$-uniformemente convexas.*

*Demonstração.* Seja $Q : \mathbb{R}^p \longrightarrow \mathbb{R}$ uma função definida por,

$$Q(\beta) = \frac{1}{\epsilon} D_h(\beta : \beta_0) = \frac{1}{\epsilon}(h(\beta) - h(\beta_0) - \nabla h(\beta_0)^T(\beta - \beta_0)). \tag{3.69}$$

Derivando a função dada na Equação (3.69) se obtêm,

$$\nabla Q(\beta) = \frac{1}{\epsilon}(\nabla h(\beta) - \nabla h(\beta_0)). \tag{3.70}$$

Como função $Q$ é convexa e diferenciável pode-se definir sua divergência de Bregman,

$$D_Q(\beta : \beta_0) = Q(\beta) - Q(\beta_0) - \nabla Q(\beta_0)^T(\beta - \beta_0).$$

Das definições dadas nas Equações (3.69) e (3.70) segue-se que,

$$D_Q(\beta : \beta_0) = Q(\beta) = \frac{1}{\epsilon} D_h(\beta : \beta_0).$$

Como a função $h$ é 1-uniformemente convexa de ordem $m \geqslant 2$, isto é,

$$D_h(\beta : \beta_0) \geqslant \frac{1}{m}||\beta - \beta_0||^m.$$

Tem-se que,

$$D_Q(\beta : \beta_0) = \frac{1}{\epsilon} D_h(\beta : \beta_0) \geqslant \frac{1/\epsilon}{m}||\beta - \beta_0||^m.$$

Portanto a função $Q$ é $1/\epsilon$-uniformemente convexa de ordem $m \geqslant 2$. Passa-se então para as funções $\psi_k$. Os vetores gradientes das funções $\psi_k$ para todo $k \geqslant 0$ são dados por,

$$\nabla \psi_k(\beta) = Cm \sum_{i=0}^{k} i^{(m-1)} \nabla f(\zeta_i) + \frac{1}{\epsilon}(\nabla h(\beta) - \nabla h(\beta_0)). \tag{3.71}$$

A divergência de Bregman para as funções $\psi_k$ são dadas por,

$$D_{\psi_k}(\beta : \beta_0) = \psi_k(\beta) - \psi_k(\beta_0) - \nabla \psi_k(\beta_0)^T(\beta - \beta_0) \tag{3.72}$$

Tem-se das Equações (3.66) e (3.71) que,

$$D_{\psi_k}(\beta : \beta_0) = \left( Cm \sum_{i=0}^{k} i^{(m-1)} \left[ f(\zeta_i) + \nabla f(\zeta_i)^T (\beta - \zeta_i) \right] + Q(\beta) \right)$$

$$- \left( Cm \sum_{i=0}^{k} i^{(m-1)} \left[ f(\zeta_i) + \nabla f(\zeta_i)^T (\beta_0 - \zeta_i) \right] \right) - \left( \left[ Cm \sum_{i=0}^{k} i^{(m-1)} \nabla f(\zeta_i) \right]^T (\beta - \beta_0) \right) \Rightarrow$$

$$D_{\psi_k}(\beta : \beta_0) = Q(\beta) + Cm \sum_{i=0}^{k} i^{(m-1)} \nabla f(\zeta_i)^T (\beta - \zeta_i) - Cm \sum_{i=0}^{k} i^{(m-1)} \nabla f(\zeta_i)^T (\beta_0 - \zeta_i)$$

$$-Cm\sum_{i=0}^{k} i^{(m-1)}\nabla f(\zeta_i)^T(\beta-\beta_0) = Q(\beta) + Cm\sum_{i=0}^{k} i^{(m-1)}\nabla f(\zeta_i)^T(\beta-\beta_0)$$

$$-Cm\sum_{i=0}^{k} i^{(m-1)}\nabla f(\zeta_i)^T(\beta-\beta_0) = Q(\beta).$$

E segue-se que,

$$D_{\psi_k}(\beta:\beta_0) = Q(\beta) = D_Q(\beta:\beta_0) \geqslant \frac{1/\epsilon}{m}||\beta-\beta_0||^m.$$

E portanto as funções $\psi_k$ para todo $k \geq 0$ são $1/\epsilon$-uniformemente convexas de ordem $m \geqslant 2$. □

Com base nas observações apresentadas acimas passa-se a demonstração do Lema (22).

*Demonstração.* Procede-se por indução sobre $k \geqslant 0$. Para $k = 0$ tem-se que a afirmação dada na Equação (3.67) é verdadeira pois tanto o lado direito quanto o lado esquerdo da Equação (3.67) são iguais a zero para esse valor de $k$. Considerando-se que a desigualdade dada na Equação (3.67) é verdadeira para um valor de $k$ fixo, será mostrado que essa Equação (3.67) também é valida para $k+1$. Dado que $Z_k$ é minimizante para $\psi_k$ e que $\nabla\psi_k(Z_k) = 0$, tem-se que,

$$D_{\psi_k}(\beta:Z_k) = \psi_k(\beta) - \psi_k(Z_k) - \nabla\psi_k(Z_k)^T(\beta-Z_k) = \psi_k(\beta) - \psi_k(Z_k).$$

Do Lema 23 segue-se que as função $\psi_k$ são $1/\epsilon$-uniformemente convexas de ordem $m \geqslant 2$ e portanto, para todo $\beta \in \mathbb{R}^p$ segue-se que,

$$\psi_k(\beta) = \psi_k(Z_k) + D_{\psi_k}(\beta:Z_k) \geqslant \psi_k(Z_k) + \frac{1}{\epsilon m}||\beta-Z_k||^m.$$

Pela convexidade e diferenciabilidade da função $f$ tem-se que,

$$f(\zeta_k) \geq f(\zeta_{k+1}) + \nabla f(\zeta_{k+1})^T\big(\zeta_k - \zeta_{k+1}\big).$$

Aplicando a hipótese de indução dada na Equação (3.67) e usando a desigualdade acima se obtêm que,

$$\psi_k(\beta) \geqslant Ck^{(m)}\big[f(\zeta_{k+1}) + \nabla f(\zeta_{k+1})^T\big(\zeta_k - \zeta_{k+1}\big)\big] + \frac{1}{\epsilon m}||\beta-Z_k||^m.$$

Somando $Cm(k+1)^{(m-1)}\big[f(\zeta_{k+1}) + \nabla f(\zeta_{k+1})^T\big(\beta - \zeta_{k+1}\big)\big]$ a ambos os lados da equação acima segue-se que,

$$Cm(k+1)^{(m-1)}\big[f(\zeta_{k+1}) + \nabla f(\zeta_{k+1})^T\big(\beta - \zeta_{k+1}\big)\big] + \psi_k(\beta) \geqslant$$

$$Cm(k+1)^{(m-1)}\big[f(\zeta_{k+1})+\nabla f(\zeta_{k+1})^T\big(\beta-\zeta_{k+1}\big)\big]$$
$$+Ck^{(m)}\big[f(\zeta_{k+1})+\nabla f(\zeta_{k+1})^T\big(\zeta_k-\zeta_{k+1}\big)\big]+\frac{1}{\epsilon m}||\beta-Z_k||^m.$$

Como $k^{(m)}=k(k+1)^{(m-1)}$ tem-se que,

$$Cm(k+1)^{(m-1)}\big[f(\zeta_{k+1})+\nabla f(\zeta_{k+1})^T\big(\beta-\zeta_{k+1}\big)\big]+Cm\sum_{i=0}^{k}i^{(m-1)}\big[f(\zeta_i)+\nabla f(\zeta_i)^T\big(\beta-\zeta_i\big)\big]$$
$$+(1/\epsilon)D_h(\beta:\beta_0)\geqslant Cm(k+1)^{(m-1)}\big[f(\zeta_{k+1})+\nabla f(\zeta_{k+1})^T\big(\beta-\zeta_{k+1}\big)\big]+$$
$$Ck(k+1)^{(m-1)}\big[f(\zeta_{k+1})+\nabla f(\zeta_{k+1})^T\big(\zeta_k-\zeta_{k+1}\big)\big]+\frac{1}{\epsilon m}||\beta-Z_k||^m.$$

E isso implica que,

$$Cm\sum_{i=0}^{k+1}i^{(m-1)}\big[f(\zeta_i)+\nabla f(\zeta_i)^T\big(\beta-\zeta_i\big)\big]+(1/\epsilon)D_h(\beta:\beta_0)\geqslant$$
$$Ck(k+1)^{(m-1)}\big[f(\zeta_{k+1})+\nabla f(\zeta_{k+1})^T\big(\zeta_k-\zeta_{k+1}\big)\big]+$$
$$Cm(k+1)^{(m-1)}\big[f(\zeta_{k+1})+\nabla f(\zeta_{k+1})^T\big(\beta-\zeta_{k+1}\big)\big]+\frac{1}{\epsilon m}||\beta-Z_k||^m.$$

Portanto segue-se que

$$\psi_{k+1}(\beta)\geqslant C(k+1)^{(m-1)}\big\{k\big[f(\zeta_{k+1})+\nabla f(\zeta_{k+1})^T\big(\zeta_k-\zeta_{k+1}\big)\big]+m\big[f(\zeta_{k+1})+\nabla f(\zeta_{k+1})^T\big(\beta-\zeta_{k+1}\big)\big]\big\}$$
$$+\frac{1}{\epsilon m}||\beta-Z_k||^m\Rightarrow$$
$$\psi_{k+1}(\beta)\geqslant C(k+1)^{(m-1)}\big\{(k+m)f(\zeta_{k+1})+\nabla f(\zeta_{k+1})^T\big[m(\beta-\zeta_{k+1})+k(\zeta_k-\zeta_{k+1})\big]\big\}$$
$$+\frac{1}{\epsilon m}||\beta-Z_k||^m\Rightarrow$$
$$\psi_{k+1}(\beta)\geqslant C(k+1)^{(m-1)}\big\{(m+k)f(\zeta_{k+1})+\nabla f(\zeta_{k+1})^T\big[m\beta+k\zeta_k-(m+k)\zeta_{k+1}\big]\big\}$$
$$+\frac{1}{\epsilon m}||\beta-Z_k||^m.$$

Da primeira sequência dada na Equação (3.63) tem-se que,

$$\beta_{k+1}=[(m/(k+m)]Z_k+[k/(k+m)]\zeta_k\Rightarrow$$
$$(m+k)\beta_{k+1}-mZ_k=k\zeta_k$$

De onde segue-se que,

$$\psi_{k+1}(\beta)\geqslant C(k+1)^{(m-1)}\big\{(m+k)f(\zeta_{k+1})+\nabla f(\zeta_{k+1})^T\big[m\beta+(m+k)\beta_{k+1}-mZ_k-(m+k)\zeta_{k+1}\big]\big\}$$
$$+\frac{1}{\epsilon m}||\beta-Z_k||^m\Rightarrow$$

$$\psi_{k+1}(\beta) \geqslant C(k+1)^{(m-1)}\big\{(m+k)f(\zeta_{k+1}) + \nabla f(\zeta_{k+1})^T\big[m(\beta - Z_k) + (m+k)(\beta_{k+1} - \zeta_{k+1})\big]\big\}$$
$$+\frac{1}{\epsilon m}||\beta - Z_k||^m \Rightarrow$$
$$\psi_{k+1}(\beta) \geqslant C(k+1)^{(m-1)}\big\{(m+k)f(\zeta_{k+1}) + \nabla f(\zeta_{k+1})^T\big[(m+k)\frac{m}{m+k}(\beta - Z_k) + (m+k)(\beta_{k+1} - \zeta_{k+1})\big]\big\}$$
$$+\frac{1}{\epsilon m}||\beta - Z_k||^m \Rightarrow$$
$$\psi_{k+1}(\beta) \geqslant C(m+k)(k+1)^{(m-1)}\big\{f(\zeta_{k+1}) + \nabla f(\zeta_{k+1})^T\big[\beta_{k+1} - \zeta_{k+1} + \frac{m}{m+k}(\beta - Z_k)\big]\big\}$$
$$+\frac{1}{\epsilon m}||\beta - Z_k||^m.$$

Como tem-se que $(k+1)^{(m)} = (m+k)(k+1)^{(m-1)}$ e definindo $\tau_k = m/(m+k)$ segue-se,

$$\psi_{k+1}(\beta) \geqslant C(k+1)^{(m)}\big[f(\zeta_{k+1}) + \nabla f(\zeta_{k+1})^T\big(\beta_{k+1} - \zeta_{k+1} + \tau_k(\beta - Z_k)\big)\big] + \frac{1}{\epsilon m}||\beta - Z_k||^m.$$

Note que o primeiro termo da expressão acima é a desigualdade desejada para $k+1$, então é necessário mostrar que os termos remanescentes são não negativos. Isso será feito em duas partes: Primeiro aplica-se a desigualdade dada na Equação (3.64) ao termo $\nabla f(\zeta_{k+1})^T\big(\beta_{k+1} - \zeta_{k+1}\big)$ de onde se obtêm que,

$$\psi_{k+1}(\beta) \geqslant C(k+1)^{(m)}f(\zeta_{k+1}) + C(k+1)^{(m)}M\epsilon^{\frac{1}{m-1}}||\nabla f(\zeta_{k+1})||^{\frac{m}{m-1}}$$
$$+Cm(k+1)^{(m-1)}\nabla f(\zeta_{k+1})^T\big(\beta - Z_k\big) + \frac{1}{\epsilon m}||\beta - Z_k||^m.$$

Agora aplicando a desigualdade de *Fenchel-Young*,

$$s^T u + \frac{1}{m}||u||^m \geqslant -\frac{m-1}{m}||s||^{\frac{m}{m-1}}, \tag{3.73}$$

com as escolhas $u = \epsilon^{-\frac{1}{m}}\big(\beta - Z_k\big)$ e $s = \epsilon^{\frac{1}{m}}Cm(k+1)^{(m-1)}\nabla f(\zeta_{k+1})$ segue-se que,

$$\psi_{k+1}(\beta) \geqslant C(k+1)^{(m)}\left[f(\zeta_{k+1}) + \left(M - \frac{m-1}{m}m^{\frac{m}{m-1}}C^{\frac{1}{m-1}}\frac{\big\{(k+1)^{(m-1)}\big\}^{\frac{m}{m-1}}}{(k+1)^{(m)}}\right)\epsilon^{\frac{1}{m-1}}||\nabla f(\zeta_{k+1})||^{\frac{m}{m-1}}\right].$$

Note que $\big\{(k+1)^{(m-1)}\big\}^{\frac{m}{m-1}} \leqslant (k+1)^{(m)}$. Então da hipótese $C \leqslant M^{(m-1)}/m^m$ tem-se que o segundo termo dentro do parêntesis é não negativo e conclui-se que,

$$\psi_{k+1}(\beta) \geqslant C(k+1)^{(m)}f(\zeta_{k+1}).$$

Dado que $\beta \in \mathbb{R}^p$ é arbitrário, a desigualdade acima também vale para $Z_{k+1}$ e portanto,

$$\psi_{k+1}(Z_{k+1}) \geq C(k+1)^{(m)}f(\zeta_{k+1}),$$

concluindo a indução. □

**Teorema 24.** *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 dada na Equação (3.64) para todo $k \geqslant 0$. Então o algoritmo dado na Equação (3.63) 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,*

$$f(\zeta_k) - f(\beta_*) \leqslant \frac{D_h(\beta_* : \beta_0)}{C\epsilon k^{(m)}} = \mathcal{O}\left(\frac{1}{\epsilon k^m}\right). \tag{3.74}$$

*Demonstração.* Como $f$ é convexa a função $\psi_k$ 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 $\beta_*$ de $f$. Combinando o limitante superior acima com o resultado do Lema 22 e relembrando que $Z_k$ é o minimizador de $\psi_k$ se obtêm que,

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

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 $\epsilon = \delta^m$ se obtêm que a taxa de convergência em tempo discreto, $\mathcal{O}\big(1/\epsilon k^m\big)$, torna-se compatível com a taxa de convergência em tempo contínuo, $\mathcal{O}\big(1/t^m\big)$ da EDO dada na Equação (3.59). Além disso, o resultado do Teorema 24 não requer nenhuma hipótese adicional sobre a função $f$ além da capacidade de produzir a sequência $\{\zeta_k\}_{k\geqslant 0}$. Nos próximos parágrafos mostra-se que é possível definir um operador que produz uma sequência que satisfaz a desigualdade dada na Equação (3.64). Começa-se com algumas definições que são necessárias para a prova.

**Definição 25.** *Seja $f \in C^{m-1}(\mathbb{R}^p; \mathbb{R})$ 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,*

$$f_{m-1}(\rho : \beta) = \sum_{i=0}^{m-1} \frac{1}{i!} \nabla^{(i)} f(\beta)(\rho - \beta)^i. \tag{3.75}$$

Com base na fórmula apresentada na Equação (3.75) define-se um operador de $\mathbb{R}^p$ em $\mathbb{R}^p$ do seguinte modo.

**Definição 26.** *Dados $N > 0$ e $\epsilon > 0$ define-se o operador $G_{m,\epsilon,N} : \mathbb{R}^p \longrightarrow \mathbb{R}^p$ por,*

$$G_{m,\epsilon,N}(\beta) = \arg\min_{\rho} \Big\{ f_{m-1}(\rho : \beta) + (N/\epsilon m)||\rho - \beta||^m \Big\}. \tag{3.76}$$

As definições a seguir serão necessárias para a demonstração que o operador na Equação (3.76) satisfaz a desigualdade na Equação (3.64). Nessas definições o espaço $\mathcal{B}_m^f(\mathbb{R}^p; \mathbb{R}^p)$ é o espaço das formas $m$ lineares dadas pela $m$-ésima derivada da função objetivo $f$.

**Definição 27.** *Sejam $f \in C^m(\mathbb{R}^p; \mathbb{R})$ e $\mathcal{B}_m^f(\mathbb{R}^p; \mathbb{R}^p)$ o espaço das formas $m$ lineares definidas pela $m$-ésima derivada de $f$. Dado $\beta \in \mathbb{R}^p$. Definimos em $\mathcal{B}_m^f$ a seguinte norma,*

$$||\nabla^{(m)} f(\beta)||_{\mathcal{B}_m^f} := \sup_{||\zeta - \beta|| = 1} \big\{ ||\nabla^{(m)} f(\beta)(\zeta - \beta)^m|| \big\}. \tag{3.77}$$

**Definição 28.** *Seja $f \in C^{m-1}(\mathbb{R}^p; \mathbb{R})$. $f$ é uma função $L$-suave de ordem $m-1$ se a $(m-1)$-ésima derivada de $f$ é $L$-Lipschitz, isto é, para todo $\bar{\beta}, \hat{\beta} \in \mathbb{R}^p$ tem-se que,*

$$||\nabla^{(m-1)} f(\hat{\beta}) - \nabla^{(m-1)} f(\bar{\beta})||_{\mathcal{B}_{m-1}^f} \leq L||\hat{\beta} - \bar{\beta}||. \tag{3.78}$$

Com as definições acima pode-se passar para a demonstração do resultado principal.

**Lema 29.** *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-se que,*

$$\nabla f(\zeta)^T \big(\beta - \zeta\big) \geqslant \frac{(N^2 - 1)^{\frac{m-2}{2m-2}}}{2N} \epsilon^{\frac{1}{m-2}} ||\nabla f(\zeta)||^{\frac{m}{m-1}}. \tag{3.79}$$

*Demonstração.* Seja $\Phi_{N,m,\epsilon}^f : \mathbb{R}^p \longrightarrow \mathbb{R}$ uma função definida por,

$$\Phi_{N,m,\epsilon}^f(\rho|\beta) = f_{m-1}(\rho : \beta) + (N/\epsilon m)||\rho - \beta||^m. \tag{3.80}$$

Pode-se reescrever o operador na Equação (3.76) como,

$$G_{m,\epsilon,N}(\beta) = \arg\min_{\rho} \Big\{ \Phi_{N,m,\epsilon}^f(\rho|\beta) \Big\}.$$

Dado que $\zeta$ é solução para o problema de otimização acima segue-se que a seguinte condição abaixo é satisfeita,

$$\nabla \Phi_{N,m,\epsilon}^f(\zeta|\beta) = \sum_{i=1}^{m-1} \frac{1}{(i-1)!} \nabla^{(i)} f(\beta)(\zeta - \beta)^{i-1} + (N/\epsilon)||\zeta - \beta||^{m-2}(\zeta - \beta) = 0. \tag{3.81}$$

A expansão de Taylor de ordem $m-1$ para $\nabla f$ é dada por,

$$\nabla f(\zeta) = \sum_{i=0}^{m-1} \frac{1}{i!} \nabla^{(i+1)} f(\beta)(\zeta - \beta)^i. \tag{3.82}$$

Fazendo a mudança de variável $i = j - 1$ na Equação (3.82) segue-se que,

$$\nabla f(\zeta) = \sum_{j=1}^{m} \frac{1}{(j-1)!} \nabla^{(j)} f(\beta)(\zeta - \beta)^{j-1}. \tag{3.83}$$

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

$$\begin{aligned} ||\nabla f(\zeta) - \textstyle\sum_{j=1}^{m-1} \frac{1}{(j-1)!} \nabla^{(j)} f(\beta)(\zeta - \beta)^{j-1}|| &= ||\tfrac{1}{(m-1)!} \nabla^{(m)} f(\beta)(\zeta - \beta)^{m-1}|| \\ \leqslant ||\tfrac{1}{(m-1)!} \nabla^{(m)} f(\beta)||_{\mathcal{B}_m^f} ||\zeta - \beta||^{m-1} &= \tfrac{1}{(m-1)!} ||\nabla^{(m)} f(\beta)||_{\mathcal{B}_m^f} ||\zeta - \beta||^{m-1} \\ \leqslant \tfrac{1}{(m-1)!} \left(\tfrac{(m-1)!}{\epsilon}\right) ||\zeta - \beta||^{m-1} &= \tfrac{1}{\epsilon} ||\zeta - \beta||^{m-1}. \end{aligned} \tag{3.84}$$

Substituindo a Equação (3.81) na Equação (3.84) e tomando $r = ||\zeta - \beta||$ se obtêm que,

$$||\nabla f(\zeta) + \frac{N r^{m-2}}{\epsilon} (\zeta - \beta)|| \leqslant \frac{r^{m-1}}{\epsilon}. \tag{3.85}$$

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

$$||\nabla f(\zeta) + \frac{N r^{m-2}}{\epsilon} (\zeta - \beta)||^2 = ||\nabla f(\zeta)||^2 - \frac{2N r^{m-2}}{\epsilon} \nabla f(\zeta)^T (\beta - \zeta) + \left(\frac{2N r^{m-2}}{\epsilon}\right)^2 ||\zeta - \beta||^2.$$

Combinando a expressão acima com a desigualdade na Equação (3.85) segue-se que,

$$||\nabla f(\zeta)||^2 - \frac{2N r^{m-2}}{\epsilon} \nabla f(\zeta)^T (\beta - \zeta) + \left(\frac{2N r^{m-2}}{\epsilon}\right)^2 r^2 \leqslant \left(\frac{r^{m-1}}{\epsilon}\right)^2$$

Rearranjando os termos se obtêm a desigualdade,

$$\nabla f(\zeta)^T (\beta - \zeta) \geqslant \frac{\epsilon}{2N r^{m-2}} ||\nabla f(\zeta)||^2 + \frac{(N^2 - 1) r^m}{2N\epsilon}. \tag{3.86}$$

Note que, para $m = 2$, o primeiro termo na Equação (3.86) já implica na desigualdade na Equação (3.79). Assumindo que $m \geqslant 3$ tem-se que o lado direito da Equação (3.86) é da forma $\varphi(r) = A/r^{m-2} + Br^m$. A função $\varphi$ 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 $\varphi$, então tem-se 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 tem-se que,

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

Substituindo os valores $A = (\epsilon/2N)||\nabla f(\zeta)||^2$ e $B = (1/2N\epsilon)(N^2-1)$ na desigualdade dada na Equação (3.86) se obtêm,

$$\nabla f(\zeta)^T(\beta - \zeta) \geqslant \left( \frac{\epsilon}{2N} ||\nabla f(\zeta)||^2 \right)^{\frac{m}{2m-2}} \left( \frac{1}{2N\epsilon}(N^2-1) \right)^{\frac{m-2}{2m-2}} = \frac{(N^2-1)^{\frac{m-2}{2m-2}}}{2N} \epsilon^{\frac{1}{m-1}} ||\nabla f(\zeta)||^{\frac{m}{m-1}}.$$

□

Com o resultado do Lema 29 tem-se que o operador na Equação (3.76) produz uma sequência que satisfaz a desigualdade dada na Equação (3.64) e portanto, completando o algoritmo na Equação (3.63) com essa sequência, se obtêm um novo algoritmo que implementa a dinâmica, a tempo discreto, da EDO de segunda ordem para a lagrangiana de Bregman polinomial na Equação (3.59). Explicitamente tem-se que:

$$\begin{aligned} \beta_{k+1} &= \big[m/(k+m)\big] Z_k + \big[k/(k+m)\big] \zeta_k, \\ \zeta_k &= G_{m,\epsilon,N}(\beta_k), \\ Z_k &= \arg\min_Z \Big\{ Cmk^{(m-1)} \nabla f(\zeta_k)^T Z + (1/\epsilon) D_h(Z : Z_{k-1}) \Big\}. \end{aligned} \tag{3.87}$$

Pelos Teoremas 24 e 29 segue o seguinte corolário.

**Corolário 30.** *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 na Equação (3.87) 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 $\mathcal{O}\big(1/\epsilon k^m\big)$.*

**Observação.:** A implementação do exemplo da função na Equação (3.2) será deixado para o próximo capítulo em que é obtido o algoritmo na Equação (3.87) de forma explicita para os casos $m = 2$ e $m = 3$.

Capítulo 4

# Gradiente acelerado de alta ordem aplicado aos MLGs

Neste capítulo é implementado o algoritmo do gradiente acelerado de alta ordem dado na Equação (3.87) para estimação de parâmetros nos modelos lineares generalizados dados na Equação (2.1). Como visto no capítulo anterior, Subseção 3.4, isso é feito seguindo algumas etapas: Primeiro, uma função $h : \mathbb{R}^p \longrightarrow \mathbb{R}$ convexa e diferenciável é definida no espaço paramétrico de modo a representar alguma informação sobre a estrutura presente nesse espaço. Com base nessa função defíne-se a divergência de Bregman sobre esse espaço para obter, aquilo que foi chamado nesse texto, uma medida fraca, isto é, uma medida que não satisfaz todas as propriedades de uma métrica.

Também uma função objetivo $f : \mathbb{R}^p \longrightarrow \mathbb{R}$ convexa e diferenciável é definida sobre o espaço paramétrico. Para os modelos lineares generalizados essa função é o negativo do logaritmo da função de verossimilhança pois, como mostrado no Lema 2, para funções de ligação canônicas ela é côncava e portanto o seu negativo é uma função convexa.

O método de Newton-Raphson para os modelos lineares generalizados exige que o logaritmo da função de verossimilhança seja duas vezes diferenciável e que a matriz hessiana seja inversível. Para que o método do gradiente acelerado de alta ordem possa ser competitivo, deve-se exigir que ele utilize no máximo as mesmas informações que o método de Newton-Raphson, logo é tomado $m = 2$ e $m = 3$, respectivamente na Equação dada em (3.58), e com isso, se obtêm taxas de convergência menores ou iguais à $\mathcal{O}\big(1/\epsilon k^2\big)$ e $\mathcal{O}\big(1/\epsilon k^3\big)$, respectivamente. Defíne-se a função $h$ de tal modo que o seu vetor gradiente seja inversível, e assim, o algoritmo dada na Equação (3.87) será obtido explicitamente e portanto os procedimentos de minimização para as sequencias $\{\zeta_k\}_{k\geqslant 0}$ e $\{Z_k\}_{k\geqslant 0}$ não serão necessários.

## 4.1 Estrutura geométrica para o espaço paramétrico

Nesta subseção constroem-se uma função $h : \mathbb{R}^p \longrightarrow \mathbb{R}$ convexa e diferenciável que, como descrito na introdução desse capítulo, possa trazer alguma informação sobre a estrutura do espaço paramétrico e portanto sobre o parâmetro $\beta$. A informação que existe disponível sobre o parâmetro $\beta$ está armazenada no par $(X, y)$ em que, como dado na Definição 1, $X$ é a matriz do modelo com vetor linha $x_i^T = (x_1, ..., x_n)$ e $y^T = (y_1, ..., y_n)$ é o vetor de variáveis resposta. Essa informação é usada para dotar o espaço $\mathbb{R}^p$ de uma métrica.

Pelo o que foi discutido na Subseção 2.3 uma escolha natural é a matriz de informação de Fisher, que mede a quantidade de informação que uma amostra $(X, y)$ carrega sobre o parâmetro $\beta$. Olhando-se para a matriz de informação de Fisher no caso da distribuição normal tem-se que está assume a seguinte forma,

$$\mathcal{I}(\beta) = \phi X^T X, \tag{4.1}$$

em que, mais uma vez, considera-se o parâmetro $\phi$ conhecido. Como para valores grandes do tamanho da amostra a distribuição de probabilidade Bernoulli pode ser aproximada com boa precisão pela distribuição normal, é considerado que a matriz de informação de Fisher dada na Equação (4.1) traz, para tamanhos grandes da amostra, uma informação aproximada a informação que a matriz de informação de Fisher associada ao modelo de regressão logístico traz. A matriz dada na Equação (4.1) pode ser usada para definir a função $h : \mathbb{R}^p \longrightarrow \mathbb{R}$ da seguinte forma: Dado $\beta \in \mathbb{R}^p$ defíne-se,

$$h(u) = \frac{1}{2} u^T X^T X u, \tag{4.2}$$

para todo $u \in \mathbb{R}^p$. Note que a função $h$ é convexa e diferenciável. Se a matriz dada na Equação (4.1) for positiva definida e lembrando que para a distribuição Bernoulli $\phi = 1$, defíne-se o produto interno $\langle ., . \rangle : \mathbb{R}^p \times \mathbb{R}^p \longrightarrow \mathbb{R}$ por,

$$\langle u, v \rangle = u^T \nabla^2 h(\xi) v = u^T X^T X v, \tag{4.3}$$

em que $\xi$ é um ponto qualquer em $\mathbb{R}^p$. Como base nesse produto interno pode-se definir uma norma $||.|| : \mathbb{R}^p \longrightarrow \mathbb{R}$ tal que,

$$||u|| = \sqrt{\langle u, u \rangle}. \tag{4.4}$$

e por sua vez, com base na norma dada na Equação (4.4), defíne-se uma métrica no espaço paramétrico $\mathbb{R}^p$. Considere $d : \mathbb{R}^p \times \mathbb{R}^p \longrightarrow \mathbb{R}$ uma métrica definida por,

$$d(u, v) = \frac{1}{2}||u - v||^2. \tag{4.5}$$

Desse modo é obtido o seguinte espaço métrico $(\mathbb{R}^p, d)$ para os parâmetros do modelo. Quando trabalha-se com dados estatísticos, não é razoável assumir, por exemplo, que esses dados são simétricos. Quando o espaço paramétrico $\mathbb{R}^p$ foi munido com a métrica $d$ ficou implícito que uma determinada simetria estava presente no espaço dos parâmetros. Por isso, olha-se o comportamento local da métrica $d$ afim de relaxar a condição de simetria. Quando isso for feito será perdida a métrica, mas introduz-se uma nova medida para os dados estatísticos que é bem mais abrangente.

Dados $\epsilon > 0$ e $v \in \mathbb{R}^p$, toma-se $u \in \mathbb{R}^p$ tal que $||u - v|| < \epsilon$. Então tem-se que,

$$d(u, v) = \frac{1}{2}||u - v||^2 = \frac{1}{2}(u - v)^T \nabla^2 h(v)(u - v) =$$

$$\nabla h(v)^T (u - v) + \frac{1}{2}(u - v)^T \nabla^2 h(v)(u - v) - \nabla h(v)^T (u - v) =$$

$$h(u) - h(v) + o(||u - v||^2) - \nabla h(v)^T (u - v) =$$

$$h(u) - h(v) - \nabla h(v)^T (u - v) + o(||u - v||^2) =$$

$$D_h(u : v) + o(||u - v||^2),$$

em que $D_h : \mathbb{R}^p \times \mathbb{R}^p \longrightarrow \mathbb{R}$ é a divergência de Bregman, definida na Equação (3.44) da função $h$. O resultado acima diz que, localmente a métrica $d$ comporta-se como uma divergência de Bregman. Uma das propriedades da divergência de Bregman é que esta não é simétrica e portanto, mais adequada para dar estrutura ao espaço paramétrico. Então considera-se a seguinte estrutura $(\mathbb{R}^p, D_h)$ para o espaço paramétrico do vetor de parâmetros $\beta$ do modelo linear generalizado dado na Definição 1.

## 4.2 *Gradiente acelerado de alta ordem em forma explicita para MLGs*

Nesta subseção mostra-se que o método do gradiente acelerado de alta ordem pode ser obtido explicitamente para os modelos lineares generalizados. Isso é possível por dois motivos: O primeiro vem do fato do gradiente 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

explicita. O segundo vem do fato de que para $m = 2$ e $m = 3$ o procedimento de minimização que define a sequência $\{\zeta_k\}_{k\geqslant 0}$ também pode ser resolvida explicitamente e consequentemente o algoritmo dado na Equação (3.87), como um todo, pode ser obtido explicitamente. Tomando $m = 2$ tem-se que o algoritmo definido na Equação (3.87) assume a seguinte forma,

$$\begin{aligned} \beta_{k+1} &= [2/(k+2)]Z_k + [k/(k+2)]\zeta_k, \\ \zeta_k &= \arg\min_{\zeta}\Big\{f(\beta) + \nabla f(\beta)^T(\zeta - \beta_k) + (N/(2\epsilon))||\zeta - \beta_k||^2\Big\}, \\ Z_k &= \arg\min_{Z}\Big\{2Ck\nabla f(\zeta_k)^T Z + (1/\epsilon)D_h(Z : Z_{k-1})\Big\}. \end{aligned} \tag{4.6}$$

Começa-se calculando a sequência $\{Z_k\}_{k\geqslant 1}$. Considere $\{q_k\}_{k\geq 1}$ uma sequência de funções definidas por,

$$q_k(Z) = 2Ck\nabla f(\zeta_k)^T Z + (1/\epsilon)D_h(Z : Z_{k-1}).$$

Então a sequência $\{Z_k\}_{k\geq 1}$ pode ser reescrita como,

$$Z_k = \arg\min_{Z}\Big\{q_k(Z)\Big\}.$$

A condição necessária para que $Z_k$ seja mínimo para $q_k$ é que o vetor gradiente de $q_k$ seja nulo em $Z_k$. Então tem-se que,

$$0 = \nabla q_k(Z_k) = 2Ck\nabla f(\zeta_k) + (1/\epsilon)\big(\nabla h(Z_k) - \nabla h(Z_{k-1})\big).$$

E portanto,

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

Como o vetor gradiente da função $h$ é investível, pois $X^TX$ é invertível, segue-se da equação acima que,

$$X^TXZ_k = X^TXZ_{k-1} - 2C\epsilon k\nabla f(\zeta_k) \Rightarrow Z_k = Z_{k-1} - 2C\epsilon k\big(X^TX\big)^{-1}\nabla f(\zeta_k).$$

Passa-se agora para a sequência auxiliar $\{\zeta_k\}_{k\geq 0}$. Considere $\{g_k\}_{k\geq 0}$ uma sequência de funções definidas por,

$$g_k(\zeta) = f(\beta_k) + \nabla f(\beta_k)^T(\zeta - \beta_k) + (N/(2\epsilon))||\zeta - \beta_k||^2.$$

Então a sequência $\{\zeta_k\}_{k\geq 0}$ pode ser reescrita como,

$$\zeta_k := \arg\min_{\zeta}\{g_k(\zeta)\}.$$

A condição necessária para que $\zeta_k$ seja mínimo para $g_k$ é que o vetor gradiente de $g_k$ seja nulo em $\zeta_k$. Então tem-se que,

$$0 = \nabla g_k(\zeta_k) = \nabla f(\beta_k) + (N/\epsilon)\big(\zeta_k - \beta_k\big).$$

E portanto segue que,

$$\zeta_k = \beta_k - (\epsilon/N)\nabla f(\beta_k).$$

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

$$\begin{aligned} \zeta_k &= \beta_k - (\epsilon/N)\nabla f(\beta_k), \\ Z_k &= Z_{k-1} - 2C\epsilon k\big(X^TX\big)^{-1}\nabla f(\zeta_k), \\ \beta_{k+1} &= [2/(k+2)]Z_k + [k/(k+2)]\zeta_k. \end{aligned} \tag{4.7}$$

Uma das hipóteses do Corolário 30 é que a constante $C > 0$ que aparece na Equação (4.7) 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 está se tomando $m = 2$ tem-se que a desigualdade acima fica dada por,

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

Tomando $N = 2$ se obtêm que $C \leqslant 1/16$ e pode-se reescrever o algoritmo dado na Equação (4.7) como,

$$\begin{aligned} \zeta_k &= \beta_k - (\epsilon/2)\nabla f(\beta_k), \\ Z_k &= Z_{k-1} - (\epsilon k/8)\big(X^TX\big)^{-1}\nabla f(\zeta_k), \\ \beta_{k+1} &= [2/(k+2)]Z_k + [k/(k+2)]\zeta_k. \end{aligned} \tag{4.8}$$

Agora para o caso $m = 3$ tem-se que o algoritmo definido na Equação (3.87) assume a seguinte forma,

$$\begin{aligned} \beta_{k+1} &= [3/(k+3)]Z_k + [k/(k+3)]\zeta_k, \\ \zeta_k &= \arg\min_\zeta \Big\{ f(\beta) + \nabla f(\beta)^T(\zeta - \beta_k) + (1/2)(\zeta - \beta_k)^T\nabla^2 f(\beta_k)(\zeta - \beta_k) \\ &\qquad +(N/(3\epsilon))||\zeta - \beta_k||^3 \Big\}, \\ Z_k &= \arg\min_Z \Big\{ 3Ck(k+1)\nabla f(\zeta_k)^TZ + (1/\epsilon)D_h(Z : Z_{k-1}) \Big\}. \end{aligned} \tag{4.9}$$

Como antes começa-se calculando a sequência $\{Z_k\}_{k\geq 1}$. Considere agora que $\{q_k\}_{k\geq 1}$ é uma sequência de funções definidas por,

$$q_k(Z) = 3Ck(k+1)\nabla f(\zeta_k)^TZ + (1/\epsilon)D_h(Z : Z_{k-1}).$$

Então a sequência $\{Z_k\}_{k\geqslant 1}$ pode ser reescrita como,

$$Z_k = \arg\min_{Z} \Big\{ q_k(Z) \Big\}.$$

Como é conhecido, a condição necessária para que $Z_k$ seja mínimo para $q_k$ é que o vetor gradiente de $q_k$ seja nulo em $Z_k$. Então tem-se que,

$$0 = \nabla q_k(Z_k) = 3Ck(k+1)\nabla f(\zeta_k) + (1/\epsilon)\big(\nabla h(Z_k) - \nabla h(Z_{k-1})\big).$$

E portanto,

$$\nabla h(Z_k) = \nabla h(Z_{k-1}) - 3C\epsilon k(k+1)\nabla f(\zeta_k).$$

O vetor gradiente da função $h$ é investível, pois $X^TX$ é invertível, e segue-se da equação acima que,

$$X^TXZ_k = X^TXZ_{k-1} - 3C\epsilon k(k+1)\nabla f(\zeta_k) \Rightarrow Z_k = Z_{k-1} - 3C\epsilon k(k+1)\big(X^TX\big)^{-1}\nabla f(\zeta_k).$$

Passa-se para a sequência auxiliar $\{\zeta_k\}_{k\geqslant 0}$. Considere agora que $\{g_k\}_{k\geqslant 0}$ é uma sequência de funções definidas por,

$$g_k(\zeta) = f(\beta_k) + \nabla f(\beta_k)^T(\zeta - \beta_k) + (1/2)(\zeta - \beta_k)^T\nabla^2 f(\beta_k)(\zeta - \beta_k) + (N/(3\epsilon))||\zeta - \beta_k||^3.$$

Dado $\delta > 0$ toma-se $\eta(\delta) = \big(3\epsilon\delta/N\big)^{\frac{1}{3}}$ e tem-se para todo $\zeta \in B(\beta_k, \eta(\delta))$ que,

$$g_k(\zeta) = f(\beta_k) + \nabla f(\beta_k)^T(\zeta - \beta_k) + (1/2)(\zeta - \beta_k)^T\nabla^2 f(\beta_k)(\zeta - \beta_k) + \delta.$$

Então a sequência $\{\zeta_k\}_{k\geq 0}$ pode ser reescrita da seguinte forma,

$$\zeta_k := \arg\min_{\zeta \in B(\beta_k, \eta(\delta))} \{g_k(\zeta)\}.$$

A condição necessária para que $\zeta_k$ seja mínimo para $g_k$ é que o vetor gradiente de $g_k$ seja nulo em $\zeta_k$. Então tem-se que,

$$0 = \nabla g_k(\zeta_k) = \nabla f(\beta_k) + \nabla^2 f(\beta_k)\big(\zeta_k - \beta_k\big).$$

Portanto assumindo que a matriz hessiana $\nabla^2 f$ é invertível segue-se que,

$$\zeta_k = \beta_k - \big[\nabla^2 f(\beta_k)\big]^{-1}\nabla f(\beta_k).$$

Logo o algoritmo do gradiente acelerado de alta ardem para $m = 3$ na forma explícita é dado por:

$$\begin{aligned} \zeta_k &= \beta_k - \big[\nabla^2 f(\beta_k)\big]^{-1}\nabla f(\beta_k), \\ Z_k &= Z_{k-1} - 3C\epsilon k(k+1)\big(X^TX\big)^{-1}\nabla f(\zeta_k), \\ \beta_{k+1} &= [3/(k+3)]Z_k + [k/(k+3)]\zeta_k. \end{aligned} \tag{4.10}$$

Como antes, a constante $C > 0$ na Equação (4.10) deve satisfazer a desigualdade,

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

em que $N > 1$. Como está se tomando $m = 3$ tem-se que a desigualdade acima fica dada por,

$$C \leqslant \frac{1}{108}\left[\frac{(N^2-1)^{\frac{1}{2}}}{N^2}\right].$$

Note que, para todo $N > 1$, tem-se que,

$$\frac{(N^2-1)^{\frac{1}{2}}}{N^2} < 1.$$

Considerando a seguinte equação do quarto grau,

$$cN^4 - N^2 + 1 = 0, \tag{4.11}$$

tem-se, fazendo a mudança de variável $M = N^2$, que

$$cM^2 - M + 1 = 0. \tag{4.12}$$

O discriminante da equação do segundo grau dada em (4.12) é da forma, $\Delta = 1 - 4c$ e para que o valor de $M$ seja único e real toma-se $c = 1/4$, onde se obtêm $M = 2$ e portanto $N = \sqrt{2}$. Dado que,

$$\sqrt{c} = \frac{(N^2-1)^{\frac{1}{2}}}{N^2}.$$

Segue-se, tomando para $N$ o valor de $\sqrt{2}$, a seguinte desigualdade para $C > 0$,

$$C \leqslant \frac{1}{216}$$

e pode-se reescrever o algoritmo dado na Equação (4.10) como,

$$\begin{aligned} \zeta_k &= \beta_k - \left[\nabla^2 f(\beta_k)\right]^{-1}\nabla f(\beta_k), \\ Z_k &= Z_{k-1} - (1/72)\epsilon k(k+1)\left(X^TX\right)^{-1}\nabla f(\zeta_k), \\ \beta_{k+1} &= [3/(k+3)]Z_k + [k/(k+3)]\zeta_k. \end{aligned} \tag{4.13}$$

Na Figura 4.1 é apresentado o comportamento da sequência gerada pelo método do gradiente acelerado de alta ordem aplicado a função objetivo dada na Equação (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 uso implícito de informações sobre a curvatura da superfície via inverso da matriz hessiana da função objetivo. Também tem-se o adicional que uma outra fração da informação é advinda da função $h$ que, nesse caso, foi tomada como a própria função objetivo $f$. A convergência observada ocorreu em 2 passos.

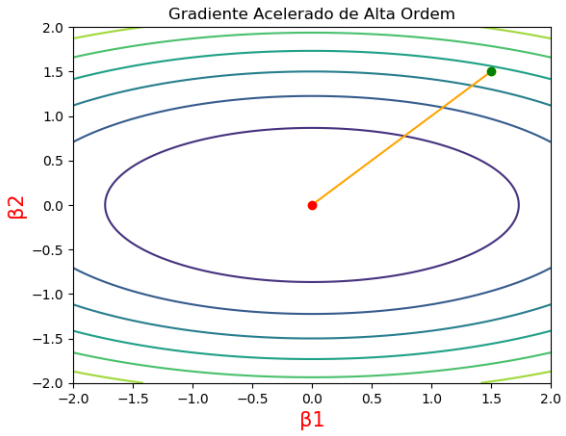


*Figura 4.1:* Comportamento da sequência gerada pelo método do gradiente acelerado de alta ordem para a função objetivo dada pela Equação (3.2). O ponto vermelho é o valor em que é inicializado o método, $\beta_0 = (1,5\,;\,1,5)$ e o ponto verde é o valor estimado pelo método, $\beta_{min} = (0,0\,;\,0,0)$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

Capítulo 5

# Comparação entre os métodos de otimização

Neste capítulo realiza-se uma série de análises afim 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, entramos 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 olha-se para o tamanho da entrada de dados. Em geral, nesse caso, buscá-se obter uma função, $F : \mathbb{N} \longrightarrow \mathbb{N}$ tal que,

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

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 efetua-se essas contagens e soma-se, em geral, se obtêm um polinômio e como é conhecido, dos cursos de cálculo da graduação, 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 adota-se 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 estatística 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, se faz necessário estabelecer métricas e padronizações. Em relação as métricas se tem, por

exemplo, o tempo de execução de cada instância do algoritmo e as taxas de convergência empíricas relacionadas a elas. Quanto as padronizações se tem, 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 empírico-estatística de algoritmos, apesar de na subseção 5.1 ser realizada uma discussão sobre os estudos desenvolvidos no capitulo 3 das taxas de convergência dos métodos de Newton-Raphson, gradiente descendente, gradiente acelerado e gradiente acelerado de alta ordem.

A análise empírico-estatística 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 tipo 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 empírico-estatística 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 Logístico.

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. Pois 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 Logístico. 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 começa-se 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 carácter estrutural e progressivo faz-se esta introdução expondo os resultados na tabela abaixo,

| métodos | Ordens de Convergência | Derivadas de $f$ |
|---|---|---|
| Gradiente descendente | $\leqslant \mathcal{O}(1/k)$ | $\nabla f$ |
| Gradiente acelerado | $= \mathcal{O}(1/k^2)$ | $\nabla f$ |
| Gradiente acelerado de alta ordem | $\leqslant \mathcal{O}(1/k^2)$ | $\nabla f$ |
| Gradiente acelerado de alta ordem | $\leqslant \mathcal{O}(1/k^3)$ | $\nabla f$ e $\nabla^2 f$ |
| Newton-Raphson | $\leqslant \mathcal{O}(\eta^{2^k})$ | $\nabla f$ 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 e duas derivadas 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 $\leqslant \mathcal{O}(1/k^2)$ e $\leqslant \mathcal{O}(1/k^3)$ sejam obtidas o método também faz o uso de uma informação extra, a saber, a derivada primeira da função $h$ que estrutura o espaço paramétrico em que se dá a convergência como discutimos na subseção 4.1. 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 explicita 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 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. Na subseção 3.4.1, em que construímos a divergência de Bregman, não foi exigida da função $h$ nada além de ser diferenciável e convexa de modo a poder definir uma divergência com ela. Isso leva a considerar que a informação advinda da estruturação do espaço via divergência de Bregman, apesar de sua função na construção do método, não é a fonte principal de informação para o estabelecimento da taxa de convergência do método do gradiente acelerado de alta ordem e, de fato, a aceleração ocorre mais em função da estrutura do próprio método, assim como no gradiente acelerado, do que por fatores externos, como a estrutura do espaço onde ocorre o processo de otimização.

## *5.2 Análise estatística 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 necessária 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 $n$. O algoritmo retorna um estimador $\hat{\beta}$ 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 dever ao fato de que, quanto maior é dimensão do vetor paramétrico maior será o espaço em que o algoritmo deve buscar os estimador $\hat{\beta}$. Em relação a dimensão $n$ tem-se que, quanto maior o seu valor maior será o volume de informação disponível para a o processo de estimação, melhorando assim, a precisão do resultado, em contra partida as operações de multiplicação e inversão matriciais efetuadas pelo algoritmo se tornam mais trabalhosas para valores de $n$ muito altos.

Dada isso, tem-se que a escolha das dimensões da amostra se constitui um ponto crucial na análise e portanto 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 abaixo,

| **(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 $(p, n)$ dar 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. 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 replicas é produtivo para se obter as informações necessárias para o tipo de análise estatística que 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 adotado e a linguagem de programação escolhida para a implementação deles. Foi considerado o número de iterações limite igual a 2000. Esse número, elevado, 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 portanto 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 (5.1) é inicializado com $\eta = G(y)$, em que $\eta$ é o preditor linear, isto é, $\eta = (x_1^T\beta, ..., x_n^T\beta)$. Maiores informações sobre essa forma do método de Newton-Raphson pode ser encontrada em (Paula, 2013). Como se utiliza (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.

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* tem-se tornado nos últimos anos uma referência quando se trata de linguagens de alta performance para computação cientifica (Sengupta, 2019). Essa linguagem consegue combinar a simplicidade sintática das linguagens *Python* e *MATLAB* com o alto desempenho em processamento de linguagens como *C* e *Fortran*. Apesar de ser uma linguagem recente, a linguagem Julia possui uma 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 Logístico*

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

Os resultados dos testes empíricos foram organizados em uma tabela e estruturados com os seguintes campos: métodos, que pode assumir os valores: gradiente descendente (GD), gradiente acelerado (GA), gradiente acelerado de alta ordem (GAAO) 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 adota-se as seguintes estatísticas: mínimo, máximo, mediana, moda, média e desvio padrão e por fim tem-se o número de derivadas que pode assumir os valores 1 e 2. Além disso se dividi a tabela em três seções para os tamanhos de amostra que foram utilizados: 250, 1000 e 5000.

#### *5.2.1.1 Comparação estatística 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 observa-se, 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 à estatística do máximo. Por fim, para essa classe de métodos, se observa 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 .

Para a classe dos métodos de segunda ordem, tem-se que o método do gradiente acelerado de alta ordem, quando faz uso da segunda derivada tem um ganho de desempenho realmente notável. Esse ganho de desempenho pode ser observado para todas as estatísticas, como por exemplo, para a média que sai de 110 iterações até a convergência quando só uma derivada esta disponível para 6 iterações até a convergência quando as duas derivadas estão disponíveis. Esse desempenho é superior ao método de Newton-Raphson segundo todas as estatísticas, em particular para a estatística do máximo do número de iterações até a convergência em que esse diferença é de 2 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 duas derivadas da função objetivo apresenta desempenho superior a todos os demais métodos 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 durantes as transições de dimensão do vetor paramétrico.

Para dimensão 25 nota-se um comportamento, para os 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.

Na classe dos métodos de segunda ordem se destaca novamente o gradiente acelerado de alta ordem que converge em 5 iterações segundo todas as estatísticas e isso corresponde a uma diferença no número de iterações em relação ao método de Newton-Raphson de 2 iterações.

Para a dimensão 50 e 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 2000 e a estatística da moda para os métodos do gradiente descendente e gradiente acelerado também têm 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 do estimador mais difícil para o mesmo tamanho da amostra. Quanto ao método do gradiente acelerado de alta ordem se tem que a moda é 1012 o que, em oposição aos dois casos anteriores, significa que, na maioria dos casos esse método convergiu. Também nota-se que a média e o desvio padrão do gradiente acelerado de alta ordem são menores em relação aos método do gradiente descendente e gradiente acelerado.

Com relação os métodos de segunda ordem, o método do gradiente acelerado de alta ordem continua a apresentar a melhor performance segundo todas as estatísticas. Note o aumento do número de iterações para todas as estatísticas que, como antes, tem como causa o aumento da dimensão do vetor paramétrico a ser estimado. As diferenças entre o número de iterações dos métodos de primeira ordem e dos métodos de segunda ordem continua acentuada para essa dimensão.

Para a dimensão 70 do vetor paramétrico e mesmo tamanho de amostra se observa 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 de 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.

Quanto aos métodos de segunda ordem se tem que o padrão observado nos casos anteriores permanece, com o método do gradiente acelerado de alta ordem mantendo a consistência em relação aos suas estatísticas.

Por fim, se tem 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.

Os métodos de segunda ordem também apresentam uma performance melhor para esse volume de informação e novamente com destaque para o método do gradiente acelerado de alta ordem em que todas as estatísticas marcam 5 iterações até a convergência, mesmo resultado que o obtido para dimensão 1000. Esse resultado é, segundo todas as estatísticas, superior aos resultados do método de Newton-Raphson.

Para a dimensão do vetor paramétrico igual a 50 se tem agora uma 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. Na classe dos métodos de segunda ordem o método do gradiente acelerado de alta ordem também demonstra, em linha com os casos anteriores, um desempenho superior em relação ao método de Newton-Raphson.

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. Para os métodos de segunda ordem o gradiente acelerado de alta ordem consegue atingir uma performance de 7 iterações até a convergência segundo todas as estatísticas com uma diferença de 2 iterações em relação ao método de Newton-Raphson.

| Métodos | Tamanho do Passo ($\epsilon$) | 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 |
| GAAO | $10^{-1.2}$ | 25 | 95 | 900 | 258 | 110 | 283,4 | 144,38 | 1 |
| GAAO | $10^{-1.1}$ | 25 | 5 | 7 | 6 | 6 | 5,7 | 0,54 | 2 |
| 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 |
| GAAO | $10^{-1.6}$ | 25 | 51 | 121 | 75 | 68 | 77,3 | 14,59 | 1 |
| GAAO | $10^{-1.6}$ | 25 | 5 | 5 | 5 | 5 | 5 | 0 | 2 |
| 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 |
| GAAO | $10^{-1.6}$ | 50 | 504 | 2000 | 901 | 1112 | 1013,2 | 359,39 | 1 |
| GAAO | $10^{-1.6}$ | 50 | 7 | 8 | 7 | 7 | 7,3 | 0,45 | 2 |
| 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 |
| GAAO | $10^{-1.4}$ | 70 | 546 | 2000 | 1506 | 2000 | 1506,4 | 438,53 | 1 |
| GAAO | $10^{-1.4}$ | 70 | 7 | 11 | 8 | 8 | 8,3 | 0,66 | 2 |
| 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 |
| GAAO | $10^{-2.2}$ | 25 | 39 | 78 | 46 | 47 | 47,8 | 6,10 | 1 |
| GAAO | $10^{-1}$ | 25 | 5 | 5 | 5 | 5 | 5 | 0 | 2 |
| 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 |
| GAAO | $10^{-2}$ | 50 | 160 | 272 | 225 | 217 | 225 | 23,62 | 1 |
| GAAO | $10^{-1.9}$ | 50 | 6 | 7 | 7 | 7 | 6,8 | 0,41 | 2 |
| 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 |
| GAAO | $10^{-2}$ | 70 | 235 | 456 | 334 | 346 | 333,3 | 44,21 | 1 |
| GAAO | $10^{-1,9}$ | 70 | 7 | 7 | 7 | 7 | 7 | 0 | 2 |
| NR | 1 | 70 | 8 | 9 | 9 | 9 | 8,9 | 0,14 | 2 |

*Tabela 5.3* - Estatísticas descritivas sobre os números de iterações 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 em toda a análise que se procedeu para os métodos de otimização em estudo aplicados ao modelo de regressão Logístico, se conclui 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 tanto na classe dos métodos de primeira ordem quanto na classe dos métodos de segunda ordem e para o caso em que esse método usa a segunda derivada da função objetivo se tem que esse método apresenta o melhor desempenho globalmente.

Na próxima subseção é analisado os tempos de execução dos métodos de otimização em estudo nessa dissertação. Essa análise e complementar ao estudo das taxas de convergência empíricas, pois ela revela se um algoritmo que executa um 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, esta comprometida, pois outros algoritmos que, apesar de demandarem mais passos até obterem a solução, o fazem em passos com tempo reduzido terão globalmente uma eficiência maior.

#### *5.2.1.2 Comparação estatística entre os tempos de execução*

Nesta subseção é realizada a análise empírico-estatística 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 recurso 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 memoria 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 do hardware utilizado para os testes dos algoritmos. Esse aspecto tem forte influência sobre os tempos de execução, por exemplo, uma maquina que tenha grande capacidade de processamento paralelo permite uma redução considerável nos tempo de execução. Também se tem que a arquitetura do processador e o modo como os processos e as *Threads* são executado influenciam na velocidade com que os algoritmos são capazes de executar as cada uma de suas instruções.

E comum na análise dos tempos de execução de um algoritmo não se usar configurações de hardware muito avançadas, como por exemplo, grande capacidade de processamento paralelo. Isso ocorre pois a eficiência do algoritmo e de sua implementação esta justamente em demandar o menor volume de recursos de hardware possível. No próximo paragrafo se inicia a análise dos resultados obtidos para os tempos de execução.

Na tabela 5.4 tem-se, 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.

Quanto aos métodos de segundo ordem, representados pelos métodos de Newton-Raphson e gradiente acelerado de alta ordem, é possível observar uma pequena diferença nos tempos de execução em favor do gradiente acelerado de alta ordem. Essa diferença é mais significativa quando se observa as estatísticas do máximo, mediana, média e desvio padrão.

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 paramétricos se tem, para os métodos de primeira ordem, um aumento em todos os tempos de execução. Um fator que influência 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.

Em relação aos métodos de segunda ordem se observa valores muito próximos para os tempos de execução segundo todas as estatísticas. Uma causa para esse comportamento é que o procedimento mais custoso computacionalmente para esses métodos é a inversão da matriza hessiana da função objetivo, o que acaba por dominar o tempo de execução desses algoritmos.

Seguindo para a dimensão 50 do vetor paramétrico os resultados apresentados para os métodos de primeira ordem corrobora o que foi constado em relação as taxas de convergência empíricas em que se observa a não convergência para o 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 apresentam 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.

Para os métodos de segunda ordem tem-se que os valores das estatísticas para os métodos do gradiente acelerado de alta ordem e Newton-Raphson estão próximas, com os tempos de execução para o gradiente acelerado de alta ordem estando um pouco acima dos valores apresentados pelo método de Newton-Raphson.

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 e 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.

Quanto aos métodos de segunda ordem tem-se que as estatísticas dos tempos de execução continuam a apresentar valores próximos entre si. Também nota-se que houve um

aumento no valor dessas estatísticas em relação ao demais casos tratados para o tamanho de amostra 1000.

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.

Em relação aos métodos de segunda ordem ambos apresentam valores próximos de tempo de execução, segundo todas as estatísticas, com os valores para o método do gradiente acelerado de alta ordem sendo ligeiramente inferiores as valores do método de Newton-Raphson.

Para a dimensão 50 tem-se, 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.

Para os métodos de segunda ordem se tem que o método de Newton-Raphson apresentam valores inferiores para os tempos de execução do que o método do gradiente acelerado de alta ordem, em particular para a estatística do máximo, onde se encontra a maior diferença. Um dos fatores para esse comportamento é que, o método do gradiente acelerado de alta ordem, executa mais operações por iteração do que o método de Newton-Raphson.

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 métodos do gradiente acelerado de alta ordem que apresentar os menores valores em todas as estatísticas quando comparado aos demais métodos de primeira ordem.

Quanto as métodos de segunda ordem para essa dimensão do vetor paramétrico se tem que ambos apresentam valores relativamente baixos para os tempos de execução, segundo todas as estatísticas, com esses tempos de execução sendo consideravelmente próximos entre os dois métodos.

| Métodos | Tamanho do Passo ($\epsilon$) | 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 |
| GAAO | $10^{-1.2}$ | 25 | 0,012 | 0,1161 | 0,033 | 0,0313 | 0,0370 | 0,0190 | 1 |
| GAAO | $10^{-1.1}$ | 25 | 0,0018 | 0,0153 | 0,0040 | 0,0030 | 0,0044 | 0,0022 | 2 |
| 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 |
| GAAO | $10^{-1.6}$ | 25 | 0,0450 | 0,1200 | 0,071 | 0,0609 | 0,0734 | 0,0150 | 1 |
| GAAO | $10^{-1.6}$ | 25 | 0,0057 | 0,0154 | 0,0072 | 0,0070 | 0,0076 | 0,0014 | 2 |
| 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 |
| GAAO | $10^{-1.6}$ | 50 | 0,1149 | 0,2290 | 0,1348 | 0,1193 | 0,1348 | 0,0075 | 1 |
| GAAO | $10^{-1.6}$ | 50 | 0,0240 | 0,0546 | 0,0322 | 0,0315 | 0,0344 | 0,0074 | 2 |
| 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 |
| GAAO | $10^{-1.4}$ | 70 | 0,7295 | 3,5344 | 2,0642 | 2,0771 | 2,0730 | 0,6877 | 1 |
| GAAO | $10^{-1.4}$ | 70 | 0,0371 | 0,1404 | 0,0571 | 0,0466 | 0,0564 | 0,0118 | 2 |
| 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 |
| GAAO | $10^{-2.2}$ | 25 | 0,2323 | 0,5983 | 0,3093 | 0,2871 | 0,3314 | 0,0697 | 1 |
| GAAO | $10^{-1}$ | 25 | 0,0236 | 0,0343 | 0,0275 | 0,0256 | 0,0278 | 0,0022 | 2 |
| 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 |
| GAAO | $10^{-2}$ | 50 | 1,0535 | 1,7645 | 1,2185 | 1,2326 | 1,2510 | 0,1157 | 1 |
| GAAO | $10^{-1.9}$ | 50 | 0,08420 | 0,1163 | 0,0990 | 0,0970 | 0,0990 | 0,0063 | 2 |
| 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 |
| GAAO | $10^{-2}$ | 70 | 0,8198 | 0,3741 | 0,7054 | 0,4285 | 0,6683 | 0,1157 | 1 |
| GAAO | $10^{-1,9}$ | 70 | 0,1363 | 0,2365 | 0,1454 | 0,1571 | 0,1477 | 0,0116 | 2 |
| 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 de execução maior 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 um pequena vantagem em relação aos demais. Na classe dos métodos de segunda ordem o método do gradiente acelerado de alta ordem também apresenta, na média, uma pequena vantagem em relação o método de Newton-Raphson para boa parte dos casos estudados e para a maioria das estatísticas utilizadas.

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 diferencias 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 polinomial com o método do gradiente acelerado de alta ordem com informação das duas derivadas da função objetivo tendo entre os métodos polinomiais a melhor taxas de convergência, $\mathcal{O}\big(1/k^3\big)$. 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 Logístico.

Esses procedimentos foram implementados na linguagem de programação Julia e em uma análise empírico-estatística foi constatado que o método do gradiente acelerado de alta ordem, fazendo o uso das duas derivadas da função objetivo, apresenta uma taxa de convergência empírica superior aos demais métodos e um tempo de execução, que na maioria dos casos, e competitivo com os tempos de execução do método de Newton-Raphson. Isso abre a possibilidade de investigações analíticas futuras sobre esse método que pode ter, nas condições exigidas pelo estudo aqui realizado, uma taxa de convergência superior

ao método de Newton-Raphson o que pode resultar em um método mais eficiente para a estimação de parâmetros não só nos modelos lineares generalizados com em outras classes de modelos estatísticos em que for possível obter as duas derivadas do logaritmo da função de verossimilhança associada ao modelo. Para que esse objetivo futuro seja realizado se faz necessário a investigar o método do gradiente acelerado de alta ordem em pelo menos duas direções: A primeira, no campo teórico, e a possibilidade de se obter um versão exponencial para os casos de uma e duas derivadas 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 a sistemática se torna fundamental para que os avanços nessas pesquisas futuras possam ser disponibilizadas para a comunidade acadêmica na forma de novos pacotes implementados nas linguagem mais utilizadas no campo da Estatística e da Otimização como, por exemplo, as linguagens `Python`, `R` e `Julia`.

# 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

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

R Core Team, 2019 R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing Vienna, Austria

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