

RAFAEL MARTIN GONÇALEZ

Funções de Bregman e Métodos das direções alternadas para multiplicadores



UNIVERSIDADE FEDERAL DE UBERLÂNDIA
FACULDADE DE MATEMÁTICA
2015

RAFAEL MARTIN GONÇALEZ

Funções de Bregman e Métodos das direções alternadas para multiplicadores

Dissertação apresentada ao Programa de Pós-Graduação em Matemática da Universidade Federal de Uberlândia, como parte dos requisitos para obtenção do título de **MESTRE EM MATEMÁTICA**.

Área de Concentração: Matemática.
Linha de Pesquisa: Matemática Aplicada.

Orientadora: Profa. Dra. Celia Aparecida Zorzo Barcelos

UBERLÂNDIA - MG
2015

Dados Internacionais de Catalogação na Publicação (CIP)
Sistema de Bibliotecas da UFU, MG, Brasil.

M381f Martin González, Rafael.
2015 Funções de Bregman e Métodos das direções alternadas para
multiplicadores / Rafael Martin González. - 2015.
97 f. : il.

Orientadora: Celia Aparecida Zorzo Barcelos.
Dissertação (mestrado) - Universidade Federal de Uberlândia,
Programa de Pós-Graduação em Matemática.
Inclui bibliografia.

1. Matemática - Teses. 2. Otimização matemática - Teses. I.
Barcelos, Celia Aparecida Zorzo. II. Universidade Federal de
Uberlândia, Programa de Pós-Graduação em Matemática. III. Título.

CDU: 51



UNIVERSIDADE FEDERAL DE UBERLÂNDIA
FACULDADE DE MATEMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM MATEMÁTICA
Av. João Naves de Ávila, 2121, Bloco 1F, Sala 1F 152
Campus Santa Mônica, Uberlândia - MG, CEP 38400-902

ALUNO(A): Rafael Martin Gonzalez

NÚMERO DE MATRÍCULA: 11312MAT008.

ÁREA DE CONCENTRAÇÃO: Matemática.

LINHA DE PESQUISA: Matemática Aplicada.

PÓS-GRADUAÇÃO EM MATEMÁTICA: Nível Mestrado.

TÍTULO DA DISSERTAÇÃO: Funções de Bregman e Métodos das direções alternadas para multiplicadores

ORIENTADORA: Profa. Dra. Celia Aparecida Zorzo Barcelos.

Esta dissertação foi **APROVADA** em reunião pública realizada na Sala Multiuso da Faculdade de Matemática, Bloco 1F, Campus Santa Mônica, em 22 de Junho de 2015, às 14h 30min, pela seguinte Banca Examinadora:

NOME

ASSINATURA

Prof(a). Dr(a). Celia Aparecida Zorzo Barcelos
UFU - Universidade Federal de Uberlândia

Celia A. Z. Barcelos

Prof. Dr. Geraldo Nunes Silva
UNESP - Universidade Estadual Paulista "Júlio de Mesquita Filho"

GNS

Prof. Dr. César Guilherme de Almeida
UFU - Universidade Federal de Uberlândia

C. G. Almeida

Uberlândia-MG, 22 de Junho de 2015.

Dedicatória

Dedico este trabalho a minha filha Beatriz que é o sorriso que alegra meus dias, a minha esposa Mirian pelo imenso amor e dedicação que foram proporcionados ao longo desses anos, aos meus pais Márcia e Claudemir (*in memoriam*) e a minha irmã Carolina.

Agradecimentos

Aos meus pais, Márcia e Claudemir (*in memorian*), por respeitarem as minhas escolhas. Obrigado por me ensinarem a viver.

A minha irmã Carolina, pelo amor de irmã.

A Mirian pelo amor, dedicação, compreensão, respeito e por fazer minha vida mais feliz.

A minha filha Beatriz, por me ensinar o que é amor incondicional.

A minha orientadora Celia Ap. Zorzo Barcelos, pela orientação durante esses anos, pela paciência, pela confiança e pelo conhecimento que me foi passado. Obrigado!

A banca de pesquisadores Prof. César Guilherme de Almeida e Prof. Geraldo Nunes Silva, pelas dicas e correções deste trabalho.

Aos meus avós e aos meus tios e tias que também participaram, de um modo ou de outro, desta caminhada.

A minhas tias Mariza e Idê por sempre me incentivarem a estudar.

Aos meus sogros Márcia e Miguel, por me tomarem como um filho.

Aos meus cunhados Celso, Mariana e Murilo.

Ao meu sobrinho Fellipe, por me ensinar a ser um tio que mima e dá bronca.

A todos os professores e amigos do programa de pós graduação em Matemática, pelo crescimento, pela aprendizagem e bons momentos que me proporcionaram.

A FAPEMIG pelo auxílio financeiro.

A Deus por tudo e por todos.

GONÇALEZ, R. M. *Funções de Bregman e Métodos das direções alternadas para multiplicadores* 2015. 97 p. Dissertação de Mestrado, Universidade Federal de Uberlândia, Uberlândia-MG.

Resumo

Neste trabalho apresentamos alguns métodos ótimos para otimização de uma função convexa diferenciável sujeito ou não a restrições. Apresentamos as condições dos métodos de Nesterov (1983) em [20], Auslender e Teboulle (2006) em [2], Nesterov (2013) em [24], Tseng (2008) em [30] e Rossetto (2012) em [29]. E demonstramos com rigor os resultados do trabalho *An accelerated linearized alternating direction method of multipliers* de Ouyang, Chen, Lan e Pasiliao (2015) em [26].

Palavras-chave: Bregman, gradiente Lipschitz contínuo, Nesterov, Métodos ótimos

GONÇALEZ, R. M. *Bregman functions and methods of alternating directions for multipliers* 2015. 97 p. M. Sc. Dissertation, Federal University of Uberlândia, Uberlândia-MG.

Abstract

In this work we deal with optimal methods for optimizing a differentiable convex function with and without constraints. We presented the methods given by the authors: Nesterov (1983) in [20], Auslender e Teboulle (2006) in [2], Nesterov (2013) in [24], Tseng (2008) in [30] and Rossetto (2012) in [29], also, we demonstrated rigorously results of the paper "An accelerated linearized alternating direction method of multipliers" given by Ouyang, Chen, Lan and Pasiliao (2015) in [26].

Keywords: Bregman, gradient Lipschitz continuous, Nesterov, optimal methods

Sumário

Resumo	vii
Abstract	viii
Introdução	1
1 Conceitos Básicos	4
1.1 Definições e resultados gerais	4
1.2 Condições de Otimalidade	9
1.3 Distâncias de Bregman	11
2 Método de Nesterov para problemas sem restrições	17
2.1 O método de Nesterov	17
2.2 Construindo a sequência $\{\phi_k(\cdot)\}$	19
2.3 Escolhendo as sequências que garantem otimalidade e o Método de Nesterov . .	21
2.4 A otimalidade do método	23
3 Modificações do algoritmo de Nesterov	25
3.1 A modificação proposta em Nesterov (2013)	25
3.2 As modificações propostas por Auslender e Teboulle (2006) e Rossetto (2012) . .	27
3.2.1 Construção da sequência $\{\phi_k(\cdot)\}$ usando distâncias de Bregman coercivas na fronteira	27
3.2.2 Generalização para projeções não interiores	30
3.3 As modificações propostas por Tseng (2008)	34
3.3.1 Introdução	34
3.3.2 Propriedades das funções de distância de Bregman	35
3.3.3 Método de otimização convexa de ordem $\mathcal{O}\left(\sqrt{\frac{L}{\epsilon}}\right)$	36
4 Método das direções alternadas acelerado e linearizado de multiplicadores	41
4.1 Condições do problema	42
4.1.1 Notações	43
4.1.2 Lagrangeano aumentado e método de direção alternada de multiplicado- res	44
4.2 Um framework acelerado ADMM	45

4.3	Principais Estimativas	48
4.4	Considerações sobre os resultados de convergência na resolução de problemas UCO em domínios limitados	56
4.5	Resultados de convergência na resolução de problemas AECCO	71
4.6	Um esquema de retrocesso	80

Introdução

Neste trabalho apresentamos alguns métodos ótimos para otimização de uma função convexa diferenciável sujeita ou não a restrições.

Começamos apresentando as condições do método de Nesterov que foi o precursor desses esquemas de resolução em 1983 em [20], onde se considera o problema de otimização convexa irrestrita $\min_{x \in \mathbb{R}^n} f(x)$ onde $f : \mathbb{R}^n \rightarrow \mathbb{R}$ é convexa continuamente diferenciável em \mathbb{R}^n , e o gradiente da função objetivo é Lipschitz contínuo, com constante de Lipschitz $L > 0$.

Na sequência apresentamos algumas modificações deste método de otimização convexa, porém com restrições. Nesterov em [24] estende suas ideias e trabalha com o seguinte problema $\min_{x \in \mathbb{R}^n} F(x)$ onde $F(x) = f(x) + G(x)$ e $G(\cdot)$ é a função indicadora do conjunto C onde C é um conjunto não vazio, convexo e fechado em \mathbb{R}^n . Auslender e Teboulle (2006) em [2] inspirados em Nesterov [20] desenvolvem um algoritmo para resolver o seguinte problema $\min_{x \in C} f(x)$ onde C é um conjunto não vazio, convexo e fechado em \mathbb{R}^n , $f : \mathbb{R}^n \rightarrow \mathbb{R}$ é convexa e diferenciável em C e o gradiente da função objetivo é Lipschitz contínuo, com constante de Lipschitz $L > 0$. Para isso usam uma sequência de funções $\{\phi_k(\cdot)\}$ que guardam informações sobre o conjunto viável através de funções distâncias conhecidas como distâncias de Bregman. Rossetto em [29] estende o método de Auslender e Teboulle para permitir o uso da distância euclidiana ao quadrado além de estimar a constante de Lipschitz para o gradiente da função objetivo para o problema tratado em seu trabalho. Tseng em [30] apresenta um tratamento unificado para os métodos de Nesterov usando distância de Bregman.

Finalizamos demonstrando rigorosamente os resultados do trabalho *An accelerated linearized alternating direction method of multipliers* de Ouyang, Chen, Lan e Pasiliao (2015) em [26] onde é apresentado um novo framework, que chamamos de AADMM, para acceleration of linearized alternating direction method of multipliers (ADMM) - aceleração do método direção alternada de multiplicadores. A ideia básica do AADMM é incorporar um sistema de aceleração para ADMM com múltiplos passos linearizados. Foi demonstrado que, para resolver uma classe de otimização convexa composta com restrições lineares, a taxa de convergência de AADMM é melhor do que a de ADMM, em termos de sua dependência da constante de Lipschitz do componente suave. Além disso, AADMM é capaz de lidar com a situação em que a região viável é ilimitada, desde que sejam satisfeitas certas condições.

O texto foi estruturado da seguinte maneira:

- Capítulo 1: apresentamos os conceitos básicos para a compreensão deste trabalho, alguns resultados de otimização convexa, além das definições de função de Bregman e distância de

Bregman, apresentando vários exemplos e propriedades. Em seguida definimos funções de Bregman coercivas na fronteira de sua zona, tal propriedade impede que uma projeção definida por esta distância seja um ponto da fronteira, sendo assim, as funções de Bregman coercivas na fronteira funcionam como barreiras impedindo que a iteração resulte num ponto fora do conjunto viável;

- Capítulo 2: descrevemos o método de Nesterov para um problema de otimização convexa irrestrita do tipo $\min_{x \in \mathbb{R}^n} f(x)$ onde $f : \mathbb{R}^n \rightarrow \mathbb{R}$. Este método foi o precursor desses esquemas de resolução que visam solucionar um problema de otimização convexa irrestrito ou restrito. Para isso Nesterov propõe a construção de uma sequência de funções $\{\phi_k(\cdot)\}$ que aproximam a função objetivo f e estabelece critérios para que a escolha de tais funções $\{\phi_k(\cdot)\}$ garantam que o algoritmo do método proposto seja ótimo;

- Capítulo 3: descrevemos algumas modificações do algoritmo de Nesterov para problemas sem restrição. As diferenças e as particularidades dessas modificações se concentram nas restrições impostas, dentre elas as feitas por Nesterov [24] que estende as ideias do capítulo anterior, trabalhando agora com restrições, para resolver o problema $\min_{x \in \mathbb{R}^n} F(x)$ onde $F(x) = f(x) + G(x)$ e $G(\cdot)$ é a função indicadora do conjunto C onde C é um conjunto não vazio, convexo e fechado em \mathbb{R}^n . As diferenças para o método restrito estão no fato das iterações serem obtidas através de projeções no conjunto viável, uma diferença na linearização feita para aproximar a função $f(\cdot)$, além de ser possível estimar o valor do gradiente da função $f(\cdot)$ sem perder a otimalidade do método. Auslender e Teboulle em [2] visam resolver o problema $\min_{x \in C} f(x)$ onde C é um conjunto não vazio, convexo e fechado em \mathbb{R}^n usando funções $\{\phi_k(\cdot)\}$ construídas de modo a guardar informações do conjunto viável por meio de distâncias de Bregman e para isso é necessário que essas distâncias de Bregman sejam coercivas na fronteira de sua zona, porém essa construção impede que a distância Eucliana seja usada. Rossetto [29] propõe o acréscimo de algumas hipóteses para que a distância Eucliana possa ser usada, aproveitando, assim, suas inúmeras propriedades. Além disso é exposto em Rossetto [29] uma tentativa de aumentar o decréscimo da função objetivo e de estimar o valor da constante de Lipschitz do gradiente da função objetivo. E, por fim, as modificações propostas por Tseng [30] o qual apresenta um tratamento unificado para os métodos de Nesterov usando funções de Bregman para otimizar o problema $\min_x f^P(x)$, em que $f^P(x) = f(x) + P(x)$;

- Capítulo 4: neste capítulo exibimos as ideias propostas no trabalho *An accelerated linearized alternating direction method of multipliers*. Após o trabalho de Nesterov (2005) [22], muitos outros trabalhos visaram tornar métodos de primeira ordem para otimização não diferenciável mais eficientes como, por exemplo, o trabalho do próprio Nesterov (2005) [23] e os trabalhos de Auslender e Teboulle [2], Tseng [30] e Lan, Lu e Monteiro [14], alguns destes detalhados no capítulo 3. Entretanto, de acordo com Ouyang et al. [26], o fato de o conjunto viável Y ser limitado é crítico para a análise do esquema de suavização de Nesterov. Seguindo o esquema de Nesterov em [22] vários estudos sobre os problemas AECCO e UCO têm ocorrido. Foi demonstrado que melhores resultados de aceleração podem ser obtidos considerando mais algumas hipóteses para os problemas AECCO - $\min_{x \in X, w \in W} G(x) + F(w)$, onde $Bw - Kx = b$

e $\text{UCO} \min_{x \in X} f(x) := G(x) + F(Kx)$. Tais problemas têm sido muito usados em aplicações em aprendizado de máquina e processamento de imagem. Na maioria das aplicações, $G(\cdot)$ é conhecido como o termo fidelidade e $F(\cdot)$ é o termo de regularização. Os problemas do tipo AECCO e UCO podem ser reformulados como problemas de ponto de sela utilizando o método dos multiplicadores de Lagrange e, a partir dessa reformulação, Ouyang, Chen, Lan e Pasiliao propõem algoritmos, proposições, teoremas e corolários nos quais focamos nossos esforços em demonstrá-los com cuidado.

Rafael Martin Gonzalez
Uberlândia-MG, 22 de Junho de 2015.

Capítulo 1

Conceitos Básicos

Neste capítulo apresentaremos os pré-requisitos necessários para a compreensão deste trabalho, assim como, algumas definições, teoremas e notações que serão utilizadas.

1.1 Definições e resultados gerais

Definição 1.1 *Um conjunto não vazio E é um **espaço vetorial** sobre \mathbb{K} , onde o símbolo \mathbb{K} denotará o corpo \mathbb{R} dos números reais ou o corpo \mathbb{C} dos números complexos, se em seus elementos, denominados **vetores**, estiverem definidas as duas seguintes operações:*

(A) *A cada par u, v de vetores de E corresponde um vetor $u + v \in E$, chamado de **soma de u e v** , de modo que:*

(A1) $u + v = v + u, \forall u, v \in E$ (propriedade comutativa).

(A2) $(u + v) + w = u + (v + w), \forall u, v, w \in E$ (propriedade associativa).

(A3) *exista em E um vetor, denominado **vetor nulo** e denotado por 0 , tal que $0 + v = v, \forall v \in E$.*

(A4) *a cada $v \in E$, exista um vetor em E , denotado por $-v$, tal que $v + (-v) = 0$.*

(M) *A cada par $\alpha \in \mathbb{K}$ e $v \in E$, corresponde um vetor $\alpha.v \in E$, denominado **produto por escalar de α por v** de modo que:*

(M1) $(\alpha\beta).v = \alpha(\beta.v), \forall \alpha, \beta \in \mathbb{K} \text{ e } v \in E$ (propriedade associativa).

(M2) $1.v = v, \forall v \in E$ (onde 1 é o elemento identidade de \mathbb{K}).

Além disso, vamos impor que as operações dadas em (A) e (M) se distribuam, isto é, que valham as seguintes propriedades:

(D1) $\alpha.(u + v) = \alpha.u + \alpha.v, \forall \alpha \in \mathbb{K} \text{ e } \forall u, v \in E$.

(D2) $(\alpha + \beta).v = \alpha.v + \beta.v, \forall \alpha, \beta \in \mathbb{K} \text{ e } \forall v \in E$.

Definição 1.2 *Seja E um espaço vetorial sobre \mathbb{K} , onde o símbolo \mathbb{K} denotará o corpo \mathbb{R} dos números reais ou o corpo \mathbb{C} dos números complexos. Uma função*

$$\|\cdot\| : E \longrightarrow \mathbb{R}$$

é uma **norma** se as seguintes propriedades forem satisfeitas:

1. $\|x\| \geq 0$ para todo $x \in E$ e $\|x\| = 0 \Leftrightarrow x = 0$.
2. $\|ax\| = |a| \cdot \|x\|$ para todo escalar a e todo $x \in E$.
3. $\|x + y\| \leq \|x\| + \|y\|$ para quaisquer $x, y \in E$.

Denotamos abaixo as normas usuais dos espaços \mathbb{K}^n , $n \in \mathbb{N}$:

$$\begin{aligned} \|(a_1, \dots, a_n)\|_1 &= |a_1| + \dots + |a_n|, \\ \|(a_1, \dots, a_n)\|_2 &= (|a_1|^2 + \dots + |a_n|^2)^{\frac{1}{2}}, \text{ e} \\ \|(a_1, \dots, a_n)\|_\infty &= \max \{|a_1| + \dots + |a_n|\}. \end{aligned}$$

Um espaço vetorial munido de uma norma será chamado de **espaço vetorial normado**, ou simplesmente **espaço normado**.

Definição 1.3 Uma **sequência** de números reais é uma função $x : \mathbb{N} \rightarrow \mathbb{R}$, que associa a cada número natural n um número real x_n , chamado o **n -ésimo termo** da sequência. Se a cada natural n fizermos corresponder uma função f_n , definida em A (isto é, $f_n : A \rightarrow \mathbb{R}$), então (f_n) , será dita **sequência de funções**.

Definição 1.4 Um **operador linear contínuo** do espaço normado E no espaço normado F , ambos sobre o mesmo corpo \mathbb{K} , é uma função $T : E \rightarrow F$, que é linear, isto é:

- $T(x + y) = T(x) + T(y)$ para quaisquer $x, y \in E$ e
- $T(ax) = aT(x)$ para todo $a \in \mathbb{K}$ e qualquer $x \in E$

e contínua, isto é, para todos $x_0 \in E$ e $\epsilon > 0$ existe $\delta > 0$ tal que $\|T(x) - T(x_0)\| < \epsilon$ sempre que $x \in E$ e $\|x - x_0\| < \delta$.

O conjunto de todos os operadores lineares contínuos de E em F será denotado por $\mathcal{L}(E, F)$. Quando F é o corpo de escalares, escrevemos E' no lugar de $\mathcal{L}(E, \mathbb{K})$, e chamamos esse espaço de **dual topológico de E** , ou simplesmente **dual de E** , e dizemos que seus elementos são **funcionais lineares contínuos**.

Dizemos, ainda, que T é um **operador linear limitado** se existe um $M > 0$ tal que:

$$\|T(x)\| \leq M\|x\|, \quad \forall x \in E.$$

Definição 1.5 Seja E um espaço vetorial sobre um corpo $\mathbb{K} = \mathbb{R}$ ou \mathbb{C} . Um **produto interno** em E é uma aplicação

$$\langle \cdot, \cdot \rangle : E \times E \longrightarrow \mathbb{K}, \quad (x, y) \longrightarrow \langle x, y \rangle,$$

tal que para quaisquer $u, v, w \in E$ e $\lambda \in \mathbb{K}$:

1. $\langle u + v, w \rangle = \langle u, w \rangle + \langle v, w \rangle$, (*Distributividade ou Linearidade*)
2. $\langle \lambda u, v \rangle = \lambda \langle u, v \rangle$, (*Homogeneidade ou Associatividade*)
3. $\langle v, v \rangle \geq 0$ e igual a zero se, e somente se, $v = 0$.
4. $\langle u, v \rangle = \overline{\langle v, u \rangle}$ (*Simetria hermitiana*)

Sendo que \bar{z} representa o conjugado complexo de $z \in \mathbb{C}$.

O par $(E, \langle \cdot, \cdot \rangle)$ é chamado de **espaço com produto interno**. Neste caso dizemos que a função

$$\| \cdot \| : E \longrightarrow \mathbb{R}, \|x\| = \sqrt{\langle x, x \rangle},$$

é a **norma induzida pelo produto interno** $\langle \cdot, \cdot \rangle$.

Definição 1.6 Um espaço com produto interno que é completo na norma induzida pelo produto interno é chamado de **espaço de Hilbert**.

Definição 1.7 *Argumento do mínimo* de uma função representa o valor (ou valores) do argumento x no conjunto S que minimiza (ou minimizam) a função objetivo f , denota-se por:

$$\arg \min_{x \in S} f(x)$$

Definição 1.8 Diz-se que um **conjunto** $A \subset \mathbb{R}^n$ é **aberto** em \mathbb{R}^n se para todo $x \in A$, existe $B(x, r)$ tal que $B(x, r) \subset A$.

Definição 1.9 Seja $A \subset \mathbb{R}^n$. Um ponto $x \in \mathbb{R}^n$ é dito **ponto da fronteira** ou do bordo de A se toda vizinhança de x intersecta A e $\mathbb{R}^n - A$. Denotamos o conjunto dos pontos da fronteira do conjunto A por ∂A .

Definição 1.10 Um subconjunto: $A \subset \mathbb{R}^n$ diz-se **convexo** se para quaisquer $x \in A$, $y \in A$ e $t \in [0, 1]$, tem-se que $tx + (1 - t)y \in A$. O ponto $tx + (1 - t)y$, onde $t \in [0, 1]$, se chama **combinação convexa** de x e y (com parâmetro t).

Isso quer dizer que um conjunto é convexo quando, para quaisquer dois pontos deste conjunto, o segmento que une esses pontos está inteiramente contido nele.

Definição 1.11 Seja $A \subset \mathbb{R}^n$ um conjunto convexo. Uma função $f : A \rightarrow \mathbb{R}^n$ chama-se **convexa** quando, para quaisquer pontos $x, y \in A$ e $t \in [0, 1]$, tem-se:

$$f(tx + (1 - t)y) \leq tf(x) + (1 - t)f(y).$$

A função é dita **estritamente convexa** quando a desigualdade acima é estrita. Se, para quaisquer pontos x e y e $t \in [0, 1]$, valer a desigualdade

$$f(tx + (1 - t)y) \leq tf(x) + (1 - t)f(y) - \frac{1}{2}\mu t(1 - t)\|x - y\|^2.$$

dizemos que $f(\cdot)$ é **fortemente convexa** em A com parâmetro de convexidade $\mu > 0$.

Definição 1.12 Seja $U \subseteq \mathbb{R}^m$ aberto, dizemos que a **função** $f : U \rightarrow \mathbb{R}^n$ é **diferenciável** em $a \in U$ quando existe uma transformação linear $T : \mathbb{R}^m \rightarrow \mathbb{R}^n$ tal que, para todo $v \in \mathbb{R}^m$ com $a + v \in U$ temos

$$f(a + v) = f(a) + T(v) + r(v) \quad \text{com} \quad \lim_{v \rightarrow 0} \frac{r(v)}{\|v\|} = 0$$

A diferenciabilidade de f no ponto a significa que podemos obter uma boa aproximação linear para f numa vizinhança de a . Essa boa aproximação de $f(a + v)$ por $f(a) + T(v)$ numa vizinhança de a é expressa pela condição $\lim_{v \rightarrow 0} \frac{r(v)}{\|v\|} = 0$. Pondo $\rho(v) = \frac{r(v)}{\|v\|}$ se $v \neq 0$ e $\rho(0) = 0$ podemos exprimir a diferenciabilidade de f no ponto a por:

$$f(a + v) = f(a) + T(v) + \rho(v)\|v\| \quad \text{com} \quad \lim_{v \rightarrow 0} \rho(v) = 0.$$

Diz-se que f é **continuamente diferenciável** ou de classe C^1 se f for diferenciável e, além disso, a sua derivada for contínua.

Definição 1.13 Sejam $U \subseteq \mathbb{R}^n$ aberto, $a \in U$ e $f : U \rightarrow \mathbb{R}$ diferenciável. O **gradiente de f em a** é o vetor

$$\nabla f(a) = \left(\frac{\partial f}{\partial x_1}(a), \dots, \frac{\partial f}{\partial x_n}(a) \right)$$

também pode-se denotar o vetor gradiente de f em a por $\text{grad}f(a)$.

Nas próximas definições, consideramos uma extensão da função $f(\cdot)$ que pode assumir valores $-\infty$ ou ∞ .

Definição 1.14 Chamamos de **conjunto de nível** de $f(\cdot)$ o conjunto $\mathcal{L}_f(r) = \{x \in \mathbb{R}^n | f(x) \leq r\}$.

Definição 1.15 Uma função $f : S \rightarrow [-\infty, \infty]$ é **semicontínua inferiormente** (lsc) em um ponto $\bar{x} \in S$ se, para toda sequência $\{x^k\} \subset S$ que converge para \bar{x} é possível verificar

$$f(\bar{x}) \leq \liminf_{k \rightarrow \infty} f(x^k).$$

Definição 1.16 Seja $A \subset \mathbb{R}^n$. Uma função $f : S \rightarrow [-\infty, \infty]$ é **própria** em S se, para pelo menos um $x \in A$, acontecer $f(x) < \infty$ e $f(x) > -\infty$ para todo $x \in S$.

Definição 1.17 Seja $S \subset \mathbb{R}^n$ um subconjunto aberto e uma função $f : S \rightarrow \mathbb{R}$ continuamente diferenciável. Dizemos que o **gradiente de $f(\cdot)$** é **Lipschitz contínuo** em S , se existir um escalar $L > 0$ tal que:

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

para todo $x, y \in S$.

Proposição 1.1 *Se φ é estritamente convexa e diferenciável em C então:*

$$\varphi(x) - \varphi(y) - \langle \nabla \varphi(y), x - y \rangle \geq 0.$$

Definição 1.18 *Seja $C \subset \mathbb{R}^n$ um subconjunto convexo e uma função $f : C \rightarrow \mathbb{R}$. O **subdiferencial**, representado por $\partial f(x)$, de $f(\cdot)$ no ponto $x \in \text{dom} f$ é definido pelo seguinte conjunto:*

$$\partial f(x) = \{g \in \mathbb{R}^n | f(y) \geq f(x) + \langle g, y - x \rangle \text{ para todo } y \in C\}.$$

Um vetor $g \in \partial f(x)$ é chamado de **subgradiente** de f em x

O Lema a seguir nos mostra como separar o produto interno $\langle \cdot, \cdot \rangle$ e transformá-lo em norma $\|\cdot\|^2$.

Lema 1.1 *Seja H um espaço de Hilbert, $x, y, z \in H$ e $\lambda \in \mathbb{R}$, a seguinte igualdade é válida:*

$$\langle \lambda(x - y), y - z \rangle = \frac{\lambda}{2} [\|x - z\|^2 - \|y - z\|^2 - \|x - y\|^2].$$

Teorema 1.1 *Seja f uma função convexa e própria. Para que o $\text{dom} f$ seja limitado é suficiente e necessário que: Se f assume valores finitos então existe um número real $\alpha \geq 0$ tal que:*

$$|f(x) - f(y)| \leq \alpha |x - y|, \quad \forall x, \quad \forall y.$$

Dem: Corolário 13.3.3 página 116 da referência [28].

Teorema 1.2 *Para qualquer função convexa e própria f e para qualquer vetor x , as seguintes condições para um vetor x^* são equivalentes:*

- (a) $x^* \in \partial f(x)$;
- (b) $\langle z, x^* \rangle - f(z)$ atinge o supremo em z para $z = x$;
- (c) $f(x) + f^*(x^*) \leq \langle x, x^* \rangle$;
- (d) $f(x) + f^*(x^*) = \langle x, x^* \rangle$.

Se $\bar{f}(x) = f(x)$, as seguintes condições são adicionadas à lista de equivalências:

- (a') $x^* \in \partial f^*(x^*)$;
- (b') $\langle x, z^* \rangle - f^*(z^*)$ atinge o supremo em z^* para $z^* = x^*$;
- (a'') $x^* \in \partial \bar{f}(x)$.

em que $f^*(x^*) = \sup \{ \langle \sum_{i=1}^m \lambda_i a_i, x^* \rangle - \sum_{i=1}^m \lambda_i \alpha_i \}$ e \bar{f} denota o fecho de f .

Dem: Teorema 23.5 página 218 da referência [28].

Teorema 1.3 (Decomposição de Moreau) *Seja K um cone convexo e fechado. Para $x, x_1, x_2 \in \mathbb{R}^n$, as seguintes propriedades são equivalentes:*

(a) $x = x_1 + x_2$ com $x_1 \in K$, $x_2 \in K^0$ e $\langle x_1, x_2 \rangle = 0$.

(b) $x_1 = P_K(x)$ e $x_2 = P_{K^0}(x)$

Onde $P_K : \mathbb{R}^n \rightarrow \mathbb{R}$ é chamada de projeção definida por, $P_K(y) = \operatorname{argmin} \{\|y - x\| / x \in K\}$ e $K^0 = \{y \in \mathbb{R} : \langle y, x \rangle \leq 0, \forall x \in K\}$

Dem: Lema 1.7 página 17 da referência [25].

Método dos Multiplicadores de Lagrange. Considere uma função de n variáveis $f(x_1, x_2, \dots, x_n)$ e m funções de restrição $g_1(x_1, x_2, \dots, x_n) \dots g_m(x_1, x_2, \dots, x_n)$. Sejam estas funções deriváveis em primeira ordem com derivadas contínuas e que para qualquer ponto do domínio existe algum i para o qual $\nabla g_i(x) \neq 0$, se f tiver um extremo relativo dentro de suas restrições, este ponto ocorre em um ponto $P(x_1^*, x_2^*, \dots, x_n^*)$ tal que P pertença a uma superfície de restrição de f na qual a seguinte condição seja satisfeita:

$$\nabla f(x_1^*, x_2^*, \dots, x_n^*) = \sum_{i=1}^m \lambda_i \nabla g_i(x_1^*, x_2^*, \dots, x_n^*)$$

$\lambda = (\lambda_1, \lambda_2, \dots, \lambda_m)$ são os multiplicadores de Lagrange.

1.2 Condições de Otimalidade

O objetivo desta seção é apresentar as condições para que um ponto seja uma solução de um problema de otimização convexa.

Para isso consideremos o problema de otimização convexa sem restrições:

$$\min_{x \in \mathbb{R}^n} f(x) \quad (1.1)$$

onde $f : \mathbb{R}^n \rightarrow \mathbb{R}$ é convexa e diferenciável em \mathbb{R}^n .

Para que um ponto x^* seja solução de um determinado problema de otimização é necessário que o gradiente da função neste ponto seja nulo. Considerando a hipótese de convexidade, esta condição torna-se também suficiente, conforme o teorema a seguir.

Teorema 1.4 *Seja $f : \mathbb{R}^n \rightarrow \mathbb{R}$ convexa e diferenciável em \mathbb{R}^n . Se o gradiente de $f(\cdot)$ em x^* for nulo, isto é, $\nabla f(x^*) = 0$, então o ponto x^* é uma solução global do problema 1.1.*

Dem: Teorema 1.18 página 10 da referência [29].

Quando $f(\cdot)$ não for diferenciável podemos enunciar um teorema análogo, apenas trocando o gradiente pelo subgradiente.

Consideremos agora o problema de otimização de uma função convexa em um conjunto convexo:

$$\min_{x \in C} f(x) \quad (1.2)$$

em que C é um conjunto não vazio, convexo e fechado em \mathbb{R}^n e $f : \mathbb{R}^n \rightarrow \mathbb{R}$ é convexa e diferenciável em C .

Definição 1.19 Dizemos que $x^* \in [-\infty, \infty)$ definido por

$$\inf_{x \in C} f(x)$$

é o **valor ótimo** do problema 1.2.

As condições para que um ponto seja solução ótima do problema 1.2 serão dadas no próximo teorema.

Teorema 1.5 Seja C um conjunto convexo e fechado do \mathbb{R}^n e $f : C \rightarrow \mathbb{R}$ uma função convexa e diferenciável em C . Um ponto x^* é uma solução do problema 1.2 se, e somente se, para todo $x \in C$,

$$\langle \nabla f(x^*), x - x^* \rangle \geq 0. \quad (1.3)$$

Dem: Teorema 1.20 página 11 da referência [29].

Definição 1.20 Seja f^* o **valor ótimo** de um problema de otimização 1.2. Dado $\epsilon > 0$, resolver este problema com precisão $\epsilon > 0$, significa encontrar uma solução aproximada $\bar{x} \in C$ de forma que $f(\bar{x}) - f^* < \epsilon$.

Definição 1.21 Considerando uma função positiva $g : \mathbb{R}^n \rightarrow \mathbb{R}$. Dizemos que a **cota de complexidade inferior** de uma classe \mathcal{F} de funções de uma classe de métodos de otimização \mathcal{M} é $\Omega(g(\epsilon))$ quando para todo M da classe de métodos \mathcal{M} , existe uma função $\bar{f} \in \mathcal{F}$ para qual o método gastará $\Omega(g(\epsilon))$ iterações com precisão ϵ . Isto quer dizer que, existe $N_1 > 0$ que depende das classes \mathcal{M} e \mathcal{F} tal que se $\{x^k\}$ é a sequência gerada pelo método M de \mathcal{M} e se $k \leq N_1 g(\epsilon)$, então:

$$\bar{f}(x^k) - \bar{f}^* \geq \epsilon$$

A complexidade inferior está relacionada com as funções de \mathcal{F} que são difíceis de minimizar pelos métodos de \mathcal{M} .

Por outro lado, consideremos uma função positiva $h : \mathbb{R}^n \rightarrow \mathbb{R}$. Fixado um método M , a **cota de complexidade superior** deste método é $O(h(\epsilon))$ quando ele minimiza, com precisão ϵ , toda função da classe \mathcal{F} em $O(h(\epsilon))$ iterações. Isto quer dizer que, existe $N_2 > 0$ que depende do método M tal que se $\{x^k\}$ é a sequência gerada pelo método M para minimizar $f \in \mathcal{F}$ e se $k \geq N_2 h(\epsilon)$, então:

$$f(x^k) - f^* \leq \epsilon$$

Definição 1.22 Um método M de uma classe de métodos \mathcal{M} é **ótimo** para uma classe \mathcal{F} de funções quando sua cota de complexidade superior é proporcional a cota de complexidade inferior da classe de funções \mathcal{F} e da classe de métodos \mathcal{M} .

Nesterov mostrou, em [21], que existe uma função \bar{f} na classe das funções convexas tal que para todo método de primeira ordem,

$$\bar{f}(x^k) - \bar{f}^* \geq \frac{3L\|x^0 - x^*\|^2}{32(k+1)^2}.$$

Isto quer dizer que:

$$\bar{f}(x^k) - \bar{f}^* = \Omega\left(\frac{1}{k^2}\right).$$

De onde concluí-se que a cota de complexidade inferior para a classe de funções convexas e os métodos de primeira ordem é $\Omega\left(\frac{1}{\sqrt{\epsilon}}\right)$. Dessa forma, para um método de primeira ordem ser considerado ótimo, para a classe de funções convexas, ele deve satisfazer a cota de complexidade superior $O\left(\frac{1}{\sqrt{\epsilon}}\right)$. Porém, para isto, basta que:

$$f(x^k) - f^* = O\left(\frac{1}{k^2}\right)$$

Em resumo, pode-se dizer que um método de primeira ordem é ótimo para a classe de funções convexas quando determina uma solução, com precisão ϵ , em $\frac{1}{\sqrt{\epsilon}}$ iterações (ROSSETTO, 2012) [29].

1.3 Distâncias de Bregman

Nesta seção apresentamos a definição de uma distância de Bregman, alguns exemplos e algumas de suas propriedades.

Definição 1.23 *Seja Ω um conjunto convexo, aberto e não vazio do \mathbb{R}^n . Seja $\varphi : \bar{\Omega} \rightarrow \mathbb{R}$ uma função própria, semicontínua inferiormente, estritamente convexa, contínua em $\bar{\Omega}$ e diferenciável em Ω . Defina:*

$$D_\varphi(x, y) = \varphi(x) - \varphi(y) - \langle \nabla \varphi(y), x - y \rangle (> 0 \text{ para } x \neq y).$$

A função $\varphi(\cdot)$ é dita uma função de Bregman e $D_\varphi(\cdot, \cdot)$ a distância de Bregman induzida por $\varphi(\cdot)$ se valem:

B1.

- Para todo $x \in \bar{\Omega}$, $D_\varphi(x, \cdot)$ tem conjuntos de nível limitado em Ω ;
- Para todo $y \in \Omega$, $D_\varphi(\cdot, y)$ tem conjuntos de nível limitado em $\bar{\Omega}$;

B2. Para todo $y \in \bar{\Omega}$, para todo $\{y^k\} \subset \Omega$ com $\lim_{k \rightarrow \infty} y^k = y$, tem-se $\lim_{k \rightarrow \infty} D_\varphi(y, y^k) = 0$;

B3. Se $\{y^k\}$ é uma sequência limitada em Ω e $y \in \bar{\Omega}$, tal que $\lim_{k \rightarrow \infty} D_\varphi(y, y^k) = 0$, então $\lim_{k \rightarrow \infty} y^k = y$.

O conjunto Ω é chamado **zona de $\varphi(\cdot)$** .

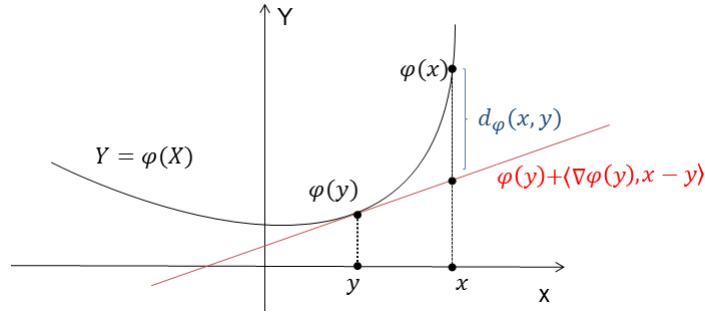


Figura 1.1: Distância de Bregman de x em relação a y .

Exemplos de distância de Bregman

Exemplo 1.1 A distância euclidiana é talvez o mais simples e mais amplamente utilizado exemplo de distância de Bregman. A função $\varphi(x) = \langle x, x \rangle$ é estritamente convexa, diferenciável em \mathbb{R}^d e

$$\begin{aligned} D_\varphi(x, y) &= \langle x, x \rangle - \langle y, y \rangle - \langle x - y, \nabla \varphi(y) \rangle \\ &= \langle x, x \rangle - \langle y, y \rangle - \langle x - y, 2y \rangle \\ &= \langle x - y, x - y \rangle = \|x - y\|^2. \end{aligned}$$

Exemplo 1.2 Uma outra distância de Bregman amplamente utilizada é a KL-divergência.

Se $p = (p_1, p_2, \dots, p_d)$; $p_j > 0$, $1 \leq j \leq d$, $d \in \mathbb{N}$ é de modo que $\sum_{j=1}^d p_j = 1$, então

$\varphi(p) = \sum_{j=1}^d p_j \log_2 p_j$ é uma função convexa. A distância de Bregman correspondente é:

$$\begin{aligned} D_\varphi(p, q) &= \sum_{j=1}^d p_j \log_2 p_j - \sum_{j=1}^d q_j \log_2 q_j - \langle p - q, \nabla \varphi(q) \rangle \\ &= \sum_{j=1}^d p_j \log_2 p_j - \sum_{j=1}^d q_j \log_2 q_j - \sum_{j=1}^d (p_j - q_j)(\log_2 q_j + \log_2 e) \\ &= \sum_{j=1}^d p_j \log_2 \left(\frac{p_j}{q_j} \right) - \log_2 e \sum_{j=1}^d (p_j - q_j) \end{aligned}$$

Chamada de KL-divergência entre as duas distribuições quando $\sum_{j=1}^d p_j = \sum_{j=1}^d q_j = 1$

Exemplo 1.3 Outras funções convexas e suas respectivas distâncias de Bregman correspondentes:

1. Temos que $\varphi(x) = x^2$ é uma função convexa e diferenciável em \mathbb{R} . A distância de Bregman correspondente é:

$$\begin{aligned}
D_\varphi(x, y) &= x^2 - y^2 - \langle x - y, \nabla \varphi(y) \rangle \\
&= x^2 - y^2 - \langle x - y, 2y \rangle \\
&= x^2 - y^2 - (2xy - 2y^2) \\
&= x^2 - 2xy + y^2 \\
&= (x - y)^2.
\end{aligned}$$

Esta distância de Bregman é chamada Perda Quadrática.

2. Temos que $\varphi(x) = x \log x$ é uma função convexa e diferenciável em \mathbb{R}_+ . A distância de Bregman correspondente é:

$$\begin{aligned}
D_\varphi(x, y) &= x \log x - y \log y - \langle x - y, \nabla \varphi(y) \rangle \\
&= x \log x - y \log y - \langle x - y, \log y + 1 \rangle \\
&= x \log x - y \log y - (y \log y + y - x \log x - x) \\
&= y(\log y - \log x) - (y - x) \\
&= y \log \left(\frac{y}{x} \right) - (y - x).
\end{aligned}$$

3. Temos que $\varphi(x) = x \log x + (1 - x) \log(1 - x)$ é uma função convexa e diferenciável em $[0, 1]$. A distância de Bregman correspondente é:

$$\begin{aligned}
D_\varphi(x, y) &= x \log x + (1 - x) \log(1 - x) - y \log y + (1 - y) \log(1 - y) - \langle x - y, \nabla \varphi(y) \rangle \\
&= x \log x + (1 - x) \log(1 - x) - y \log y + (1 - y) \log(1 - y) - \langle x - y, \log y - \log(1 - y) \rangle \\
&= x \log x + (1 - x) \log(1 - x) - y \log y + (1 - y) \log(1 - y) \\
&\quad - (x \log y - x \log(1 - y) - y \log y + y \log(1 - y)) \\
&= x \log x - x \log y + (1 - x) \log(1 - x) - (1 - x) \log(1 - y) \\
&= x \log \left(\frac{x}{y} \right) - (1 - x) \log \left(\frac{1 - x}{1 - y} \right).
\end{aligned}$$

Esta distância de Bregman é chamada Perda Logística.

4. Temos que $\varphi(x) = -\log x$ é uma função convexa e diferenciável em \mathbb{R}_{++} . A distância de Bregman correspondente é:

$$\begin{aligned}
D_\varphi(x, y) &= -\log x - \log y - \langle x - y, \nabla \varphi(y) \rangle \\
&= -\log x + \log y - \left\langle x - y, -\frac{1}{y} \right\rangle \\
&= -\log x + \log y - \left(-\frac{x}{y} + 1 \right) \\
&= -\log \left(\frac{x}{y} \right) + \frac{x}{y} - 1.
\end{aligned}$$

Esta distância de Bregman é chamada Distância de Itakura-Saito.

5. Temos que $\varphi(x) = e^x$ é uma função convexa e diferenciável em \mathbb{R} . A distância de Bregman correspondente é:

$$\begin{aligned} D_\varphi(x, y) &= e^x - e^y - \langle x - y, \nabla \varphi(y) \rangle \\ &= e^x - e^y - \langle x - y, e^y \rangle \\ &= e^x - e^y - (xe^y - ye^y) \\ &= e^x - e^y - (x - y)e^y. \end{aligned}$$

6. Se x é de modo que $\sum_{j=1}^d x_j = 1$, então $\varphi(x) = \sum_{j=1}^d x_j \log x_j$ é uma função convexa. A distância de Bregman correspondente é:

$$\begin{aligned} D_\varphi(x, y) &= \sum_{j=1}^d x_j \log x_j - \sum_{j=1}^d y_j \log y_j - \langle x - y, \nabla \varphi(y) \rangle \\ &= \sum_{j=1}^d x_j \log x_j - \sum_{j=1}^d y_j \log y_j - \left\langle x - y, \sum_{j=1}^d (\log y_j + 1) \right\rangle \\ &= \sum_{j=1}^d x_j \log x_j - \sum_{j=1}^d y_j \log y_j - \sum_{j=1}^d (x_j \log y_j + x_j - y_j \log y_j - y_j) \\ &= \sum_{j=1}^d x_j \log \left(\frac{x_j}{y_j} \right) - \sum_{j=1}^d (x_j - y_j) \end{aligned} \tag{1.4}$$

$$\tag{1.5}$$

Esta é a Generalizada I-divergência entre as duas distribuições quando $\sum_{j=1}^d x_j = \sum_{j=1}^d y_j = 1$

Propriedades das distâncias de Bregman

As seguintes propriedades das distâncias de Bregman foram retiradas do trabalho de Banerjee e Meguru [3].

Propriedade 1.1 Não Negatividade. $D_\varphi(x, y) \geq 0, \forall x \in \overline{\Omega}, y \in \Omega$, é igual a 0, se, e somente se $x = y$.

Propriedade 1.2 Convexidade. D_φ é uma função convexa no primeiro argumento, mas não necessariamente convexa no segundo argumento.

Distância euclidiana ao quadrado e KL-divergência são exemplos de distância de Bregman que são convexas em ambos os seus argumentos, mas a distância de Bregman correspondente a uma função φ estritamente convexa $\varphi(x) = x^3$, definido em R_+ , determinado por $D_\varphi(x, y) = x^3 - y^3 - 3(x - y)y^2$ é um exemplo de divergência que não é convexa em y .

Propriedade 1.3 Linearidade. A Distância de Bregman é um operador linear, ou seja, $\forall x \in \overline{\Omega}, y \in \Omega$,

$$\begin{aligned} D_{\varphi_1 + \varphi_2}(x, y) &= D_{\varphi_1}(x, y) + D_{\varphi_2}(x, y) \\ D_{c\varphi}(x, y) &= cD_{\varphi}(x, y) \text{ (para } c \geq 0 \text{)}. \end{aligned}$$

Propriedade 1.4 Classes de equivalência. As distâncias de Bregman de funções que diferem apenas nos termos afins são idênticas, ou seja, se $\varphi(x) = \varphi_0(x) + \langle b, x \rangle + c$ em que $b \in \mathbb{R}^n$ e $c \in \mathbb{R}$, então $D_{\varphi}(x, y) = D_{\varphi_0}(x, y)$, $\forall x \in \overline{\Omega}, y \in \Omega$. Assim, o conjunto de todas as funções diferenciáveis e estritamente convexas em um conjunto convexo S pode ser dividida em classes de equivalência da forma:

$$[\varphi_0] = \{ \varphi | D_{\varphi}(x, y) = D_{\varphi_0}(x, y), \quad \forall x \in \overline{\Omega}, y \in \Omega \}.$$

Propriedade 1.5 Separação Linear. O lugar geométrico de todos os pontos $x \in \overline{\Omega}$ que estão equidistantes entre dois pontos fixos $\mu_1, \mu_2 \in \Omega$ em termos de uma distância de Bregman é um hiperplano, ou seja, as partições induzidas por distâncias de Bregman tem separadores lineares dados por:

$$\begin{aligned} D_{\varphi}(x, \mu_1) &= D_{\varphi}(x, \mu_2) \\ \varphi(x) - \varphi(\mu_1) - \langle x - \mu_1, \nabla \varphi(\mu_1) \rangle &= \varphi(x) - \varphi(\mu_2) - \langle x - \mu_2, \nabla \varphi(\mu_2) \rangle \\ \langle x, \nabla \varphi(\mu_2) - \nabla \varphi(\mu_1) \rangle &= (\varphi(\mu_1) - \varphi(\mu_2)) - (\langle \mu_1, \nabla \varphi(\mu_1) \rangle - \langle \mu_2, \nabla \varphi(\mu_2) \rangle). \end{aligned}$$

Propriedade 1.6 Generalização do teorema de Pitágoras (Identidade dos três pontos). Para qualquer $x \in \overline{\Omega}$ e $y, z \in \Omega$, a sequência de três pontos $\{x, y, z\}$ detém a propriedade:

$$D_{\varphi}(x, y) = D_{\varphi}(x, z) + D_{\varphi}(z, y) + \langle \nabla_1 D_{\varphi}(z, y), x - z \rangle. \quad (1.6)$$

Dem: Realmente, da definição temos que:

$$D_{\varphi}(x, y) = \varphi(x) - \varphi(y) - \langle \nabla \varphi(y), x - y \rangle$$

Somando e subtraindo termos convenientes, reescrevemos a equação acima como

$$\begin{aligned} D_{\varphi}(x, y) &= \varphi(x) - \varphi(y) - \langle \nabla \varphi(y), x - y \rangle + (\varphi(z) - \varphi(z)) \\ &\quad + \langle \nabla \varphi(z), x - z \rangle - \langle \nabla \varphi(z), x - z \rangle - \langle \nabla \varphi(y), z - y \rangle + \langle \nabla \varphi(y), z - y \rangle. \end{aligned}$$

Reorganizando os termos, temos:

$$\begin{aligned} D_\varphi(x, y) &= \varphi(x) - \varphi(z) - \langle \nabla \varphi(z), x - z \rangle \\ &\quad + \varphi(z) - \varphi(y) - \langle \nabla \varphi(y), z - y \rangle \\ &\quad - \langle \nabla \varphi(y), x - y \rangle + \langle \nabla \varphi(z), x - z \rangle + \langle \nabla \varphi(y), z - y \rangle. \end{aligned}$$

Usando a linearidade do produto interno, obtemos:

$$D_\varphi(x, y) = D_\varphi(x, y) + D_\varphi(x, y) + \langle \nabla \varphi(z), x - z \rangle + \langle \nabla \varphi(y), -x + z \rangle$$

De onde segue a equação (1.6).

No capítulo 3 apresentaremos o método de Auslender e Teboulle onde a distância de Bregman utilizada deixa implícito qual será o conjunto viável. Para isso, o fecho da zona Ω da distância de Bregman deve ser igual ao conjunto viável, ou seja $\overline{\Omega} = C$ e, além disso, é necessário que a distância de Bregman seja coerciva na fronteira de Ω (veja a definição a seguir).

Definição 1.24 *Seja $\{y^i\}_{i \in \mathbb{N}} \in \Omega$ uma sequência de elementos com $y^i \rightarrow y$ e y na fronteira de $\overline{\Omega}$, se para todo $x \in \Omega$ vale:*

$$\langle \nabla \varphi(y^i), x - y^i \rangle \rightarrow -\infty.$$

Então, dizemos que a função de Bregman $\varphi(\cdot)$ é **coerciva na fronteira** de Ω .

Uma observação elencada por Rossetto [29] é que o fato de uma distância de Bregman ser coerciva na fronteira de sua zona impede que uma projeção definida por esta distância seja um ponto da fronteira. Iusem [13] mostra que isto se deve ao fato de que na fronteira, o subdiferencial de uma distância de Bregman coerciva é vazio.

Abaixo um exemplo de distância de Bregman com a propriedade de coercividade.

Exemplo 1.4 *O exemplo já comentado em 1.3 é baseado em $\varphi(x) = \sum_{j=1}^d x_j \log x_j$, onde $D_\varphi(x, y)$ é dado por (1.4). Onde a zona é o interior de \mathbb{R}_+^n (veja a Definição 1.23).*

Mais detalhes sobre estas funções podem ser encontrados em Auslender e Teboulle [2].

A norma euclidiana ao quadrado é um exemplo de distância de Bregman que satisfaz a identidade dos três pontos, entretanto sua zona é o espaço \mathbb{R}^n . Sendo assim, usar essa distância no método de Auslender e Teboulle [2] só faz sentido se quisermos resolver um problema irrestrito (ROSSETTO, 2012) [29].

Capítulo 2

Método de Nesterov para problemas sem restrições

O objetivo deste capítulo é apresentar um método ótimo para problemas sem restrições proposto por Nesterov ([21], [29]). Para isso vamos considerar o problema de otimização convexa irrestrita:

$$\min_{x \in \mathbb{R}^n} f(x) \quad (2.1)$$

onde $f : \mathbb{R}^n \rightarrow \mathbb{R}$ é convexa e continuamente diferenciável em \mathbb{R}^n , e o gradiente ∇f da função objetivo é Lipschitz contínuo, com constante de Lipschitz $L > 0$, isto é, para todo $x, y \in \mathbb{R}^n$

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\| \quad (2.2)$$

2.1 O método de Nesterov

Levando em consideração que um método de primeira ordem ótimo para a classe de funções convexas deve satisfazer:

$$f(x^k) - f^* = \mathcal{O}\left(\frac{1}{k^2}\right)$$

em que f^* denota o valor ótimo do problema e $\{x^k\}$ é a sequência gerada pelo método para minimizar o problema. Nesterov propôs a construção de uma sequência de funções $\{\phi_k(\cdot)\}$ que aproximam a função objetivo $f(\cdot)$ possibilitando escrever a desigualdade:

$$f(x^k) - f^* \leq \lambda_k(\phi_0(x^*) - f(x^*)), \quad (2.3)$$

onde $\lambda_k \geq 0$ e $\phi_0(\cdot)$ é a primeira função da sequência $\{\phi_k(\cdot)\}$. Assim, se $\lambda_k = \mathcal{O}(\frac{1}{k^2})$, então, o método será ótimo (ROSSETTO, 2012) [29].

Se a sequência de funções $\phi_k(\cdot)$, que aproximam a função $f(\cdot)$, for construída de forma que em cada iteração $k \geq 0$ a seguinte razão seja satisfeita para $\alpha_k \in (0, 1)$ e para todo $x \in \mathbb{R}^n$,

$$\phi_{k+1}(x) - f(x) \leq (1 - \alpha_k)(\phi_k(x) - f(x)), \quad (2.4)$$

então a desigualdade (2.3) é válida. E a partir de (2.4), obtemos:

$$\phi_k(x) - f(x) \leq (1 - \alpha_{k-1})(1 - \alpha_{k-2}) \cdots (1 - \alpha_0)(\phi_0(x) - f(x)). \quad (2.5)$$

Devemos determinar, ainda, a sequência $\{x^k\}$ de modo que, para todo $k \geq 0$, ϕ_k^* que é o valor mínimo de $\phi_k(\cdot)$, seja maior ou igual ao valor de $f(x^k)$.

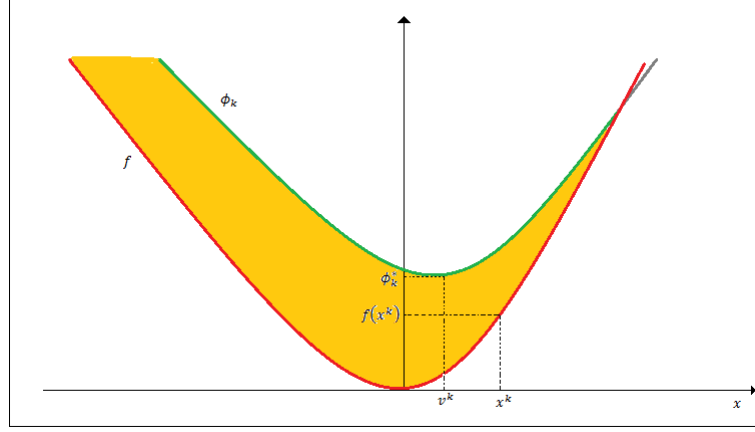


Figura 2.1: $\phi_k^* \geq f(x^k)$

Se a sequência $\phi_k(\cdot)$ satisfizer a desigualdade (2.5) e de $\phi_k^* \geq f(x^k)$, para todo $k \geq 0$, segue de forma imediata a desigualdade (2.3) com:

$$\lambda_k := \prod_{i=0}^{k-1} (1 - \alpha_i) \quad (2.6)$$

A seguir, definimos o conceito de **função simples** para darmos continuidade à construção da sequência $\phi_k(\cdot)$.

Definição 2.1 *Seja $\phi_k : \mathbb{R}^n \rightarrow \mathbb{R}$ uma função fortemente convexa. Dizemos que $\phi_k(\cdot)$ é **função simples** se é da forma:*

$$\phi_k(x) = \phi_k^* + \frac{\gamma_k}{2} D(x, v^k),$$

onde $\phi_k^* \in \mathbb{R}$ é o valor ótimo de $\phi_k(\cdot)$, $v^k \in \mathbb{R}^n$ é o minimizador irrestrito de $\phi_k(\cdot)$, $\gamma_k \in \mathbb{R}$ é o parâmetro de convexidade forte e $D(x, v^k)$ é uma função que representa uma distância em relação a v^k . Se $D(x, v^k) = \|x - v^k\|^2$ dizemos que $\phi(\cdot)$ é uma **função quadrática convexa simples**.

Vejam, a seguir, que as funções da sequência $\{\phi_k(\cdot)\}$ também são funções quadráticas convexas simples. A figura nos mostra uma sequência de funções aproximadoras, ou seja, uma sequência de funções quadráticas convexas simples se aproximando da função $f(\cdot)$.

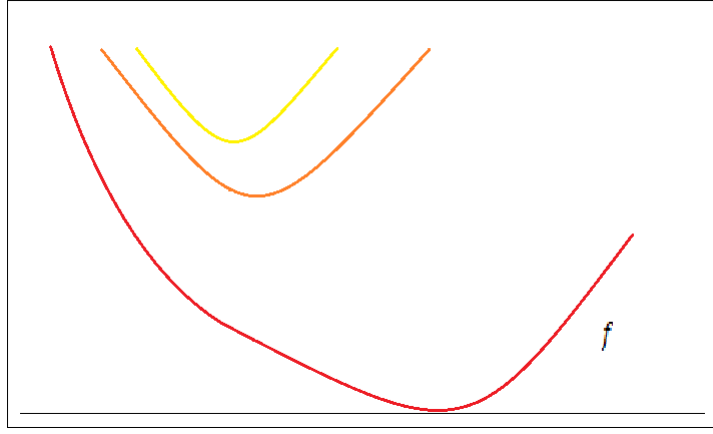


Figura 2.2: Sequência de funções aproximadoras

Nesterov, afim de justificar a construção de tal sequência de funções, satisfazendo (2.3) e (2.4), introduz o conceito de sequência de estimativas.

Definição 2.2 Um par de sequências $\{\phi_k(\cdot)\}$ e $\{\lambda_k\}$, $\lambda_k \geq 0$ é chamado uma **sequência de estimativas** da função $f(\cdot)$ se

$$\lambda_k \rightarrow 0$$

e para qualquer $x \in \mathbb{R}^n$ e todo $k \geq 0$ temos:

$$\phi_k(x) - f(x) \leq \lambda_k(\phi_0(x) - f(x)) \quad (2.7)$$

Podemos observar que, se as funções da sequência satisfazem a desigualdade (2.4): $\phi_{k+1}(x) - f(x) \leq (1 - \alpha_k)(\phi_k(x) - f(x))$, então a sequência $\{\phi_k(\cdot)\}$ satisfaz a desigualdade (2.7) com $\lambda_k := \prod_{i=0}^{k-1} (1 - \alpha_i)$, ou seja, λ_k dado pela equação (2.6).

Lema 2.1 Seja uma função $f(\cdot)$ que admite sequência de estimativa e x^* uma solução do problema (2.1). Se uma sequência $\{x^k\}$ satisfaz

$$f(x^k) \leq \phi_k^* \equiv \min\{\phi_k(x) \mid x \in \mathbb{R}^n\} \quad (2.8)$$

então,

$$f(x^k) - f(x^*) \leq \lambda_k(\phi_0(x^*) - f(x^*)) \rightarrow 0 \quad (2.9)$$

Este lema nos mostra que se a função $f(\cdot)$ admite sequência de estimativas, então pode-se medir a taxa de convergência de $f(x^k) \rightarrow f(x^*)$ analisando-se apenas a velocidade com que $\lambda_k \rightarrow 0$.

2.2 Construindo a sequência $\{\phi_k(\cdot)\}$

Nesta seção apresentamos como Nesterov construiu a sequência $\{\phi_k(\cdot)\}$ de modo a satisfazer a desigualdade (2.4).

Definição 2.3 *Sejam $y^k \in \mathbb{R}^n$, $\alpha_k \in (0, 1)$, $\phi_0(\cdot)$ uma função arbitrária em \mathbb{R}^n e $f(\cdot)$ uma função convexa do \mathbb{R}^n . Para todo $x \in \mathbb{R}^n$ e todo $k \geq 0$, $\phi_k(\cdot)$ é definida por:*

$$\phi_{k+1}(x) = (1 - \alpha_k)\phi_k(x) + \alpha_k l(x, y^k) \quad (2.10)$$

$$l(x, y^k) = f(y^k) + \langle \nabla f(y^k), x - y^k \rangle + \frac{\mu}{2} \|x - y^k\|^2 \quad (2.11)$$

Se $f(\cdot)$ é fortemente convexa definimos μ como um limitante inferior do parâmetro de convexidade forte desta função, senão, definimos $\mu = 0$.

Ilustramos na figura abaixo o modo como Nesterov obteve cada uma das funções da sequência $\{\phi_k(\cdot)\}$ por meio de combinações, respeitando a Definição 2.3

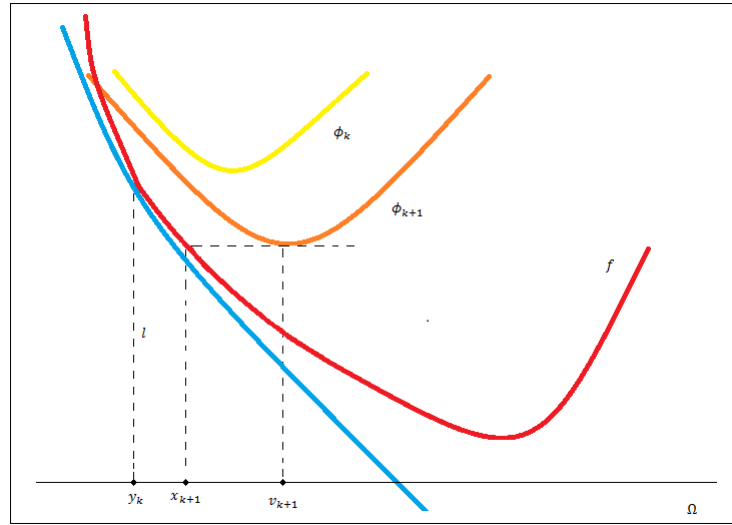


Figura 2.3: Construção de $\{\phi_k(\cdot)\}$

Nos próximos lemas vemos que a desigualdade (2.4) é satisfeita pela sequência de funções definidas acima. Veremos que se as condições da desigualdade (2.4) forem satisfeitas para uma função quadrática simples $\phi_0(\cdot)$, então todas as funções da sequência $\{\phi_k(\cdot)\}$ serão funções quadráticas simples. É apresentado, ainda, o minimizador e o valor ótimo ϕ_{k+1}^* de $\phi_{k+1}(\cdot)$.

Lema 2.2 *Consideremos a Definição 2.3. Então,*

$$\phi_{k+1}(x) - f(x) \leq (1 - \alpha_k)(\phi_k(x) - f(x)) \quad (2.12)$$

Lema 2.3 *Seja $\phi_0^* \in \mathbb{R}$ e $\gamma_0 > 0$, se*

$$\phi_0(x) = \phi_0^* + \frac{\gamma_0}{2} \|x - v^0\|^2,$$

então, as funções da sequência $\{\phi_k(\cdot)\}$ definidas pelas equações (2.10) e (2.11) admitem, para todo $k \geq 0$, a forma:

$$\phi_k(x) = \phi_k^* + \frac{\gamma_k}{2} \|x - v^k\|^2, \quad (2.13)$$

onde as sequência γ_k , v^k e ϕ_k^* são definidas por:

$$\gamma_{k+1} = (1 - \alpha_k)\gamma_k + \alpha_k\mu, \quad (2.14)$$

$$v^{k+1} = \frac{1}{\gamma_{k+1}} \left((1 - \alpha_k)\gamma_k v^k + \alpha_k\mu y^k - \alpha_k \nabla f(y^k) \right), \quad (2.15)$$

$$\begin{aligned} \phi_{k+1}^* &= (1 - \alpha_k)\phi_k^* + \alpha_k f(y^k) - \frac{\alpha_k^2}{2\gamma_{k+1}} \|\nabla f(y^k)\|^2 \\ &\quad + \frac{(1 - \alpha_k)\gamma_k \alpha_k}{\gamma_{k+1}} \left(\frac{\mu}{2} \|y^k - v^k\|^2 \langle \nabla f(y^k), v^k - y^k \rangle \right). \end{aligned} \quad (2.16)$$

e como na Definição 2.3, se $f(\cdot)$ é fortemente convexa definimos μ como um limitante inferior do parâmetro de convexidade forte desta função, senão, definimos $\mu = 0$.

2.3 Escolhendo as sequências que garantem otimalidade e o Método de Nesterov

Nesta seção apresentamos como determinar x^k , y^k e α_k de modo que $\phi_k^* \geq f(x^k)$ e sintetizamos os resultados, afim de obter o algoritmo intitulado de método de Nesterov.

De acordo com Rossetto [29], neste método uma iteração qualquer se inicia com $x^k, v^k \in \mathbb{R}^n$ e $\gamma_k \in \mathbb{R}$. Então, tomamos α_k como sendo a maior raiz positiva da equação $L\alpha_k^2 = (1 - \alpha_k)\gamma_k + \alpha_k\mu$ (L é a constante de Lipschitz dada em (2.2)) e escolhemos um ponto intermediário y^k , como podemos ver na figura abaixo, a partir de x^k na direção de x^k para v^k .

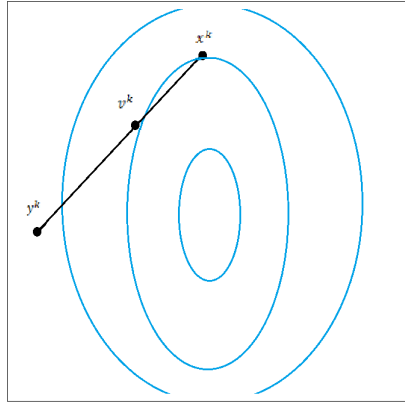


Figura 2.4: y^k a partir de x^k na direção $(v^k - y^k)$

Devemos, agora, determinar o ponto x^{k+1} .

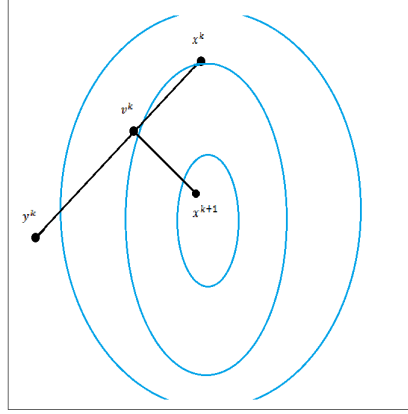


Figura 2.5: x^{k+1} na direção $-\nabla f(y^k)$

E, então, atualizar o valor de γ_{k+1} e determinar o ponto v^{k+1} .

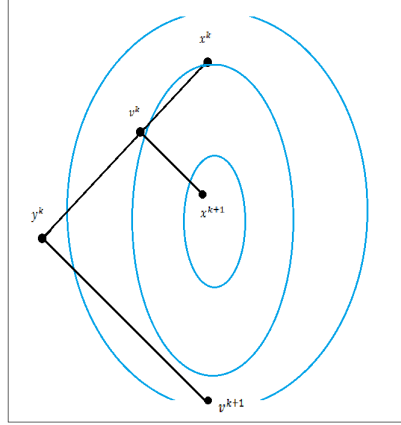


Figura 2.6: v^{k+1} na direção $-\nabla f(y^k)$

Nesterov demonstrou que com as escolhas acima é garantido que $\phi_k^* \geq f(x^k)$. Esses resultados compõem o próximo teorema.

Teorema 2.1 *Seja $L > 0$. Suponha que a iteração $k \geq 0$ se inicia com $\gamma_k > 0, v^k, x^k \in \mathbb{R}^n$ tal que $\phi_k^* \geq f(x^k)$ e $\alpha_k \in (0, 1)$ sendo a maior raiz da equação $L\alpha_k^2 = (1 - \alpha_k)\gamma_k + \alpha_k\mu$. Se as seguintes escolhas são feitas*

$$y^k = x^k + \theta_k(v^k - x^k),$$

$$\text{com } \theta_k = \frac{\alpha_k \gamma_k}{\gamma_k + \alpha_k \mu}$$

$$x^{k+1} = y^k - \frac{1}{L} \nabla f(y^k),$$

$\gamma_{k+1} = (1 - \alpha_k)\gamma_k + \alpha_k\mu$. Então, a iteração termina com:

$$\phi_{k+1}^* \geq f(x^{k+1})$$

A partir dos resultados apresentados neste capítulo, podemos sintetizar o método de Nesterov, no seguinte algoritmo, que nos é apresentado em Rossetto ([29], p.29) :

Algoritmo 2.1 *Método de Nesterov*

- 1: Dados: $x^0 \in \mathbb{R}^n$, $\gamma_0 > \mu$, $L > 0$,
- 2: Faça $k = 0$, $v^0 = x^0$
- 3: Repita
- 4: $d^k = v^k - x^k$
- 5: Calcule a maior raiz da equação $L\alpha_k^2 = (1 - \alpha_k)\gamma_k + \alpha_k\mu$
- 6: Faça $\theta_k = \frac{\alpha_k\gamma_k}{\gamma_k + \alpha_k\mu}$
- 7: $y^k = x^k + \theta_k d^k$
- 8: $x^{k+1} = y^k - \frac{1}{L} \nabla f(y^k)$
- 9: $\gamma_{k+1} = (1 - \alpha_k)\gamma_k + \alpha_k\mu$
- 10: $v^{k+1} = \frac{1}{\gamma_{k+1}}((1 - \alpha_k)\gamma_k v^k + \alpha_k\mu_k y^k - \alpha_k \nabla f(y^k))$
- 11: $k = k + 1$

2.4 A otimalidade do método

Nesta seção enunciaremos os resultados apresentados por Nesterov que garantem a otimalidade do método. Nos Lemas 2.4 e 2.5 encontramos os resultados necessários para provar a otimalidade do método dada no Teorema 2.2 donde obtemos que:

$$f(x^k) - f(x^*) = \mathcal{O}\left(\frac{1}{k^2}\right)$$

Isso quer dizer que com precisão ϵ em um número proporcional a $\frac{1}{\sqrt{\epsilon}}$ iterações, pode-se estimar uma solução ótima x^* . Isso garante que o método seja ótimo.

Lema 2.4 *Se $\lambda_0 = 1$ e no Algoritmo 2.1 $\gamma_0 \leq \mu$, então*

$$\lambda_k \leq \min \left\{ \left(1 - \sqrt{\frac{\mu}{L}}\right)^k, \frac{4L}{(2\sqrt{L} + k\sqrt{\gamma_0})^2} \right\}.$$

Lema 2.5 *Sejam $\phi_0^* = f(x^0)$, $\lambda_0 = 1$ e $\lambda_k = \prod_{i=0}^{k-1} (1 - \alpha_i)$. O Algoritmo 2.1 gera uma sequência $\{x^k\}$ tal que*

$$f(x^k) - f(x^*) \leq \lambda_k \left[f(x^0) - f(x^*) + \frac{\gamma_0}{2} \|x^0 - x^*\|^2 \right].$$

Teorema 2.2 *O Algoritmo 2.1 gera uma sequência $\{x^k\}$ tal que:*

$$f(x^k) - f(x^*) \leq \min \left\{ \left(1 - \sqrt{\frac{\mu}{L}} \right)^k, \frac{4L}{(2\sqrt{L} + k\sqrt{\gamma_0})^2} \right\} \left(\frac{(L + \gamma_0)}{2} \|x^0 - x^*\|^2 \right).$$

Capítulo 3

Modificações do algoritmo de Nesterov

Neste capítulo apresentamos quatro modificações do método de Nesterov sem restrição, apresentado no capítulo anterior, suas diferenças e particularidades se concentram nas restrições impostas.

Para isso, iniciamos considerando o seguinte problema de otimização convexa restrita

$$\min_{x \in C} f(x) \quad (3.1)$$

onde C é um conjunto não vazio, convexo e fechado em \mathbb{R}^n , $f : \mathbb{R}^n \rightarrow \mathbb{R}$ é convexa e diferenciável em C e o gradiente da função objetivo é Lipschitz contínuo em C com constante de Lipschitz $L > 0$.

Apresentaremos, portanto, o método ótimo desenvolvido por Nesterov [24] para resolver este problema. Em seguida, apresentaremos o método ótimo de Auslender e Teboulle [2] que será usado como base para um novo método ótimo desenvolvido por Rossetto [29] e, por fim, o método ótimo de Tseng [30].

3.1 A modificação proposta em Nesterov (2013)

Nesta seção apresentaremos como Nesterov estendeu as ideias apresentadas no capítulo 2, trabalhando, agora, com restrições. Para tanto, consideramos o seguinte problema:

$$\min_{x \in \mathbb{R}^n} F(x) \quad (3.2)$$

onde $F(x) = f(x) + G(x)$ e $G(\cdot)$ é a função indicadora do conjunto C , isto é:

$$G(x) = \begin{cases} 0, & \text{se } x \in C \\ \infty, & \text{caso contrário} \end{cases}$$

As diferenças elencadas por Rossetto ([29], pp.33-34), entre o método irrestrito para o método restrito são:

- os pontos v^k e x^k são obtidos através de projeções no conjunto viável;
- a aproximação $l(\cdot, \cdot)$ definida em (2.11) é feita apenas no ponto x^{k+1} e aproxima apenas a

função $f(\cdot)$;

- é possível calcular uma estimativa para o valor da constante de Lipschitz que pode ser usada no método sem perder a otimalidade.

Para isso, Nesterov introduz o conceito de mapeamento gradiente, que é o argumento que minimiza uma combinação entre uma aproximação convexa de $f(\cdot)$ em um ponto $y \in \mathbb{R}^n$ e a função indicadora do conjunto C (ROSSETTO, 2012) [29], ou seja:

Definição 3.1 *Sejam $x, y \in \mathbb{R}^n$ e $L^\alpha > 0$. Utilizando a Definição 1.7, define-se **mapeamento gradiente** por:*

$$T_{L^\alpha}(y) = \arg \min_{x \in \mathbb{R}^n} m_{L^\alpha}(y; x)$$

onde $m_{L^\alpha}(y; x) = f(x) + \langle \nabla f(y), x - y \rangle + \frac{L^\alpha}{2} \|x - y\|^2 + G(x)$.

A partir dos estudos feitos por Rossetto [29] podemos afirmar que duas observações podem ser feitas a partir dessa definição. A primeira é que se o conjunto viável for todo o espaço \mathbb{R}^n então $T_{L^\alpha}(y)$ é um ponto a partir de y na direção oposta ao gradiente de $f(\cdot)$. A segunda é que no caso da função $G(\cdot)$ ser a função indicadora de um conjunto convexo e fechado de \mathbb{R}^n , $T_{L^\alpha}(y)$ é uma projeção ortogonal do ponto descrito acima no conjunto viável.

Com essa definição, as observações acima e o subgradiente de $F(\cdot)$

$$F'(z) = \nabla f(z) + \zeta_G(z),$$

onde $\zeta_G(z) = L^\alpha (z - T_L(z)) - \nabla f(z) \in \partial G(z)$, Nesterov prova o Lema 3.1 que fornece um critério para estimar o valor da constante de Lipschitz L .

Lema 3.1 *Se $L^\alpha > L$, então:*

$$\langle F'(T_{L^\alpha}(y)), y - T_{L^\alpha}(y) \rangle \geq \frac{1}{L^\alpha} \|F'(T_{L^\alpha}(y))\|^2.$$

Em Nesterov [24] encontramos a demonstração deste lema supracitado, bem como, uma descrição mais detalhada dos resultados que compõe o Método de Nesterov, que pode ser sintetizado no Algoritmo 3.1.

Algoritmo 3.1 *Método de Nesterov para problemas com restrição*

1: Dados: $x^0 \in \mathbb{R}^n$, $A_0 = 0$, $L_0 > 0$,

2: Faça $k = 0$, $v^0 = x^0$

3: Repita

4: $L = L_k$

5: $d^k = v^k - x^k$

- 6: Calcule a maior raiz da equação $La^2 = 2(A_k + a)$
- 7: $\theta = \frac{a}{A_k + a}$
- 8: Faça $y = x^k + \theta d^k$ e calcule $T_L(y)$
- 9: Se $\langle F'(T_L(y)), y - T_L(y) \rangle < \frac{1}{L} \|F'(T_L(y))\|^2$, então $L = 2L$
- 10: Até que $\langle F'(T_L(y)), y - T_L(y) \rangle \geq \frac{1}{L} \|F'(T_L(y))\|^2$
- 11: Defina
- 12: $y^k = y, a_{k+1} = a$
- 13: $x^{k+1} = T_L(y_k), L_{k+1} = \frac{L}{2}, A_{k+1} = A_k + a_{k+1}$
- 14: $v^{k+1} = \arg \min_x \left\{ l_k(x, x^{k+1}) + A_k G(x) + \frac{1}{2} \|x - x_0\|^2 \right\}$
- 15: $k = k + 1$
- 16: Volte ao passo 3.

Nesterov [24] mostrou que este método é ótimo para a classe de funções convexas e que o número máximo N_k de avaliações do gradiente, após k -iterações é $4(k+1) + \log_2 \frac{L}{L_0}$.

3.2 As modificações propostas por Auslender e Teboulle (2006) e Rossetto (2012)

Auslender e Teboulle [2], usando uma sequência de funções $\{\phi_k(\cdot)\}$ construídas de modo a guardar informações sobre o conjunto viável por meio de uma distância não convencional, chamada de distância de Bregman, e inspirados em Nesterov, desenvolveram um algoritmo para resolver o Problema 3.1 (ROSSETTO, 2012) [29]. De acordo com Rossetto [29] *essas distâncias de Bregman estão definidas no interior do conjunto viável e funcionam como barreiras, impedindo que a sequência de iteradas geradas pelo algoritmo saiam desse conjunto. Dessa forma, os iterados estão sempre no interior do conjunto viável* (p.35). Diante dessas considerações, nosso objetivo, nesta seção, é mostrar como foi possível estender esse método usando distâncias de Bregman, mas agora definidas num conjunto maior que o viável. Assim definidas, as distâncias de Bregman, contemplam, de acordo com Rossetto [29], o caso da distância euclidiana ao quadrado (que não é válido para o método de Auslender e Teboulle) e permitem que o iterado atinja a borda do conjunto viável. Análogo ao que foi feito por Nesterov, Rossetto [29] mostra que é possível estabelecer um critério para determinar, aproximadamente, o valor da constante de Lipschitz de forma adaptativa.

3.2.1 Construção da sequência $\{\phi_k(\cdot)\}$ usando distâncias de Bregman coercivas na fronteira

Sejam $\gamma_0 > 0$, $x^0 \in \text{int}(C)$, $\alpha_k \in (0, 1)$, $y^k \in \text{int}(C)$, $\varphi(\cdot)$ uma função de Bregman coerciva na fronteira de C (Definição 1.24). Para todo $k \geq 0$ e para qualquer $x \in C$, a sequência $\{\phi_k(\cdot)\}$ é definida recursivamente da seguinte forma:

$$\phi_0(x) = f(x^0) + \gamma_0 D_\varphi(x, x^0), \quad (3.3)$$

$$\phi_{k+1}(x) = (1 - \alpha_k)\phi_k(x) + \alpha_k l_k(x, y^k), \quad (3.4)$$

$$l_k(x, y^k) = f(y^k) + \langle \nabla f(y^k), x - y^k \rangle. \quad (3.5)$$

Essa construção da sequência proposta por Auslender e Teboulle [2] se difere da construção proposta por Nesterov apenas no que se refere ao fato da função $\varphi(\cdot)$ ser coerciva na fronteira de sua zona. Isso quer dizer que o restante da construção é idêntico a feita por Nesterov para o problema sem restrições quando tomamos $\varphi(x) = \frac{\|x\|^2}{2}$. Para Rossetto [29] a diferença fundamental deste método para o descrito no capítulo anterior é que naquele método a aproximação linear é feita em relação ao ponto x^{k+1} e o minimizador de $\phi_{k+1}(\cdot)$ depende de x^{k+1} .

A partir de agora enunciaremos os resultados propostos por Rossetto [29] e omitiremos as demonstrações por não serem o foco principal de nosso trabalho.

Lema 3.2 *A sequência $\{\phi_k(\cdot)\}$ definida pelas equações 3.10, (3.4) e (3.5) satisfaz para todo $x \in C$*

$$\phi_{k+1}(x) - f(x) \leq (1 - \alpha_k)(\phi_k - f(x)). \quad (3.6)$$

Lema 3.3 *Sejam $\gamma_0 > 0$, $v^0 = x^0 \in \text{int}(C)$ e $\phi_0^* = f(x^0)$. Então, para toda sequência $\{y^k\} \subset C$, as funções $\phi_k(\cdot)$ definidas por (3.3), (3.4) e (3.5) têm a forma:*

$$\phi_k(x) = \phi_k^* + \gamma_k D_\varphi(x, v^k), \quad (3.7)$$

onde, $v^k = \text{argmin} \{\phi_k(x) | x \in C\}$, $\phi_k^* = \phi_k(v^k)$, $\gamma_{k+1} = (1 - \alpha_k)\gamma_k$ e D_φ é uma distância de Bregman coerciva.

Lema 3.4 *Se v^{k+1} é o minimizador de $\phi_{k+1}(\cdot)$ no conjunto C , então v^{k+1} satisfaz:*

$$v^{k+1} = \text{argmin} \left\{ \left\langle x, \frac{\alpha_k}{\gamma_{k+1}} \nabla f(y^k) \right\rangle + D_\varphi(x, v^k) \right\}. \quad (3.8)$$

Teorema 3.1 *Seja $L > 0$ e suponha que $D_\varphi(\cdot, \cdot)$ seja fortemente convexa com parâmetro de convexidade $\sigma > 0$. Suponha também, que a iteração $k \geq 0$ se inicia com $\alpha_k \in (0, 1)$ definido como a maior raiz da equação $L\alpha_k^2 = \sigma(1 - \alpha_k)\gamma_k$, $v^k \in \text{int}(C)$, $x^k \in C$ tal que $\phi_k^* \geq f(x^k)$. Se as seguintes escolhas são feitas:*

$$y^k = x^k + \alpha_k(v^k - x^k), \quad (3.9)$$

$$x^{k+1} = x^k + \alpha_k(v^{k+1} - x^k), \quad (3.10)$$

v^{k+1} é dado pelo Lema 3.4 e $\gamma_{k+1} = (1 - \alpha_k)\gamma_k$, então o processo iterativo termina quando se obtém:

$$\phi_{k+1}^* \geq f(x^{k+1}). \quad (3.11)$$

A partir dos resultados apresentados nesta seção, podemos sintetizar o método de Auslender e Teboulle, no seguinte algoritmo:

Algoritmo 3.2 *Método de Auslender e Teboulle*

- 1: Dados: $x^0 \in \text{int}(C)$, $L > 0$, $\gamma_0 > 0$, $\sigma > 0$
- 2: Faça $k = 0$, $v^0 = x^0$
- 3: Repita
- 4: Calcule a maior raiz da equação $L\alpha_k^2 = \sigma(1 - \alpha_k)\gamma_k$
- 5: $\gamma_{k+1} = (1 - \alpha_k)\gamma_k$
- 6: $y^k = x^k + \alpha_k(v^k - x^k)$
- 7: $v^{k+1} = \arg \min_x \left\{ \left\langle x, \frac{\alpha_k}{\gamma_{k+1}} \nabla f(y^k) \right\rangle + D_\varphi(x, v^k) \right\}$
- 8: $x^{k+1} = x^k + \alpha_k(v^{k+1} - x^k)$
- 9: $k = k + 1$
- 10: Volte ao passo 3.

Agora enunciaremos os resultados que garantem a otimalidade do método.

Lema 3.5 *Sejam $\gamma_k > 0$ e $\alpha_k \in (0, 1)$ definidos por $L\alpha_k^2 = \sigma(1 - \alpha_k)\gamma_k$ com $\gamma_{k+1} = (1 - \alpha_k)\gamma_k$. Defina $\lambda_k := \prod_{i=0}^{k-1} (1 - \alpha_i)$, então:*

$$\lambda_k \leq \frac{4L}{\left(2\sqrt{L} + k\sqrt{\sigma\gamma_0}\right)^2} \quad (3.12)$$

Teorema 3.2 *Sejam $\{x^k\}$, $\{y^k\}$ e $\{\alpha^k\}$ as sequências geradas pelo método de Auslender e Teboulle e seja x^* a solução ótima do Problema 3.1. Então, para qualquer $k \geq 0$, temos:*

$$f(x^k) - f(x^*) \leq \frac{4L}{\sigma\gamma_0 k^2} C(x^*, x^0) = \mathcal{O}\left(\frac{1}{k^2}\right), \quad (3.13)$$

onde $C(x^*, x^0) = f(x^0) + \gamma_0 D_\varphi(x^*, x^0) - f(x^*)$. Essa sequência $\{x^k\}$ é minimizante, isto é, $f(x^k) \rightarrow f(x^*)$ e é viável.

Este teorema garante que o método é ótimo.

3.2.2 Generalização para projeções não interiores

Nesta seção descreveremos o método proposto por Rossetto [29], que estende o método proposto por Auslender e Teboulle [2], permitindo o uso da distância euclidiana ao quadrado como distância de Bregman em métodos restritos, que não era possível pois, a zona da distância euclidiana é todo o espaço \mathbb{R}^n , embora esta distância seja uma distância de Bregman, ela só poderia ser usada para minimizar todo o \mathbb{R}^n .

A diferença fundamental do método proposto por Rossetto [29] com o método de Auslender e Teboulle [2] é que neste são definidas distâncias de Bregman com zonas que podem conter propriamente o interior do conjunto viável. Como feito por Auslender e Teboulle [2], Rossetto ([29], p.48) tem a intenção de construir *uma sequência de funções* $\{\phi_k(\cdot)\}$ *de forma que cada função, que é obtida combinando com a função anterior (que é simples) com uma aproximação inferior, continue sendo uma função simples*. Para isso ela fará uso de uma sequência de funções auxiliares $\{\hat{\phi}_k(\cdot)\}$. Antes de tratarmos sobre essas construções enunciaremos algumas hipóteses que devem ser consideradas:

Hipótese 3.1 *Consideremos $\varphi(\cdot)$ uma função de Bregman como na Definição 1.23 e sobre ela supõe-se ainda as seguintes hipóteses:*

1. *o interior do conjunto viável C está contido na zona de $\varphi(\cdot)$, isto é, $\text{int}(C) \subset \Omega$;*
2. *$\varphi(\cdot)$ é σ - fortemente convexa em Ω (σ é o parâmetro de convexidade de acordo com a Definição 1.11);*
3. *o subdiferencial de $\varphi(\cdot)$ na fronteira de $\bar{\Omega}$ é vazio*

Definição 3.2 *Sejam $y^k \in \mathbb{R}^n$, $v^k \in C$, $\gamma_k > 0$, $\phi_k^* \in \mathbb{R}$ e $\alpha_k \in (0, 1)$. Onde:*

$$\begin{aligned}\phi_k(x) &= \phi_k^* + \gamma_k D_\varphi(x, v^k) \\ l(x, y^k) &= f(y^k) + \langle \nabla f(y^k), x - y^k \rangle\end{aligned}$$

definimos:

$$\hat{\phi}_{k+1}(x) = (1 - \alpha_k)\phi_k(x) + \alpha_k l(x, y^k) \quad (3.14)$$

Definição 3.3 *Sejam \bar{v} , v^+ , $\hat{\phi}_{k+1}^*$ e γ_+ definidos por:*

$$\bar{v} = \operatorname{argmin} \left\{ \hat{\phi}_{k+1}(x) \mid x \in \mathbb{R}^n \right\}, \quad (3.15)$$

$$v^+ = \operatorname{argmin} \left\{ D_\varphi(x, \bar{v}) \mid x \in C \right\}, \quad (3.16)$$

$$\hat{\phi}_{k+1}^* = \hat{\phi}_{k+1}(v^+), \quad (3.17)$$

$$\gamma_+ = (1 - \alpha_k)\gamma_k. \quad (3.18)$$

Pode-se provar que \bar{v} e v^+ estão bem definidas considerando que $\varphi(\cdot)$ é fortemente convexa, para maiores detalhes consulte Rossetto ([29], p.49)

Lema 3.6 *Consideremos a Definição 3.2 com $\varphi(\cdot)$ satisfazendo as Hipóteses 3.1, e os pontos descritos na Definição 3.3. Então, para todo $x \in C$, temos:*

$$\widehat{\phi}_{k+1}(x) \geq \widehat{\phi}_{k+1}^* + \gamma_+ D_\varphi(x, v^+). \quad (3.19)$$

Este lema mostra que as funções da sequência $\{\widehat{\phi}_k(\cdot)\}$ não são funções simples, porém qualquer função desta sequência sempre majora uma função simples. Sendo assim, precisamos buscar uma sequência de funções simples majorada pela sequência $\{\widehat{\phi}_k(\cdot)\}$. Escolhendo a função $\phi_{k+1}(\cdot)$ como função majorada por $\widehat{\phi}_k(\cdot)$, temos a formalização desse fato na seguinte definição:

Definição 3.4 *Sejam γ_{k+1} , v^{k+1} , ϕ_{k+1}^* e $\phi_{k+1}(x)$ definidos por:*

$$\gamma_{k+1} = \gamma_+ \quad (3.20)$$

$$v^{k+1} = v^+ \quad (3.21)$$

$$\phi_{k+1}^* = \widehat{\phi}_{k+1}^* \quad (3.22)$$

$$\phi_{k+1}(x) = \phi_{k+1}^* + \gamma_{k+1} D_\varphi(x, v^{k+1}). \quad (3.23)$$

Daí para todo $x \in C$ temos $\widehat{\phi}_{k+1}(x) \geq \phi_{k+1}(x)$. E, então, o método de Rossetto [29] trabalhará com duas sequências de funções e $\{\widehat{\phi}_{k+1}(\cdot)\}$ funcionará como uma sequência auxiliar para a construção da sequência $\{\phi_k(\cdot)\}$.

Lema 3.7 *As sequências $\{\widehat{\phi}_k(\cdot)\}$ e $\{\phi_k(\cdot)\}$ definidas acima, satisfazem para todo $x \in C$,*

$$\widehat{\phi}_{k+1}(x) - f(x) \leq (1 - \alpha_k)(\phi_k(x) - f(x)). \quad (3.24)$$

Lema 3.8 *Se $\gamma_0 = 1$, x^* é a solução ótima do Problema 3.1 e a sequência $\{\phi_k(\cdot)\}$ dada na Definição 3.4 satisfaz $\phi_{k+1}^* \geq f(x^{k+1})$ para todo $k \geq 0$. Então, pode-se garantir que:*

$$f(x^{k+1}) - f(x^*) \leq \lambda_{k+1}(\phi_0(x^*) - f(x^*)). \quad (3.25)$$

Com estes resultados Rossetto [29] já poderia enunciar um novo método, idêntico ao de Auslender e Teboulle [2], exceto pela troca da distância de Bregman cuja a zona contém o conjunto viável, de maneira que a distância euclidiana ao quadrado possa ser usada. No entanto, inspirada em Karas e Gonzaga [11], Rossetto [29] expõe uma tentativa de aumentar o decréscimo da função objetivo e estima um valor para a constante de Lipschitz do gradiente da função objetivo.

Proposição 3.1 *Se $L^\alpha \geq L$, então, para qualquer $x, y \in C$ temos:*

$$f(y) + \langle \nabla f(y), x - y \rangle + \frac{L^\alpha}{2} \|x - y\|^2 \geq f(x). \quad (3.26)$$

Rossetto [29] afirma que qualquer valor $L^\alpha \geq L$ garante que podemos construir uma aproximação superior da função $f(\cdot)$, como descrito na proposição acima, e sempre que a aproximação seja superior é possível demonstrar um resultado equivalente ao Teorema 3.1. O teorema que segue garante que, com as sugestões de Rossetto [29], ainda é possível garantir que $\phi_{k+1}^* \geq f(x^{k+1})$.

Teorema 3.3 *Suponha que $D_\varphi(\cdot, \cdot)$ seja fortemente convexa com parâmetro de convexidade $\sigma > 0$ e que a iteração $k \geq 0$ começa com $\gamma_k > 0$, $v^k, x^k \in \mathbb{R}^n$ tal que $\phi_k^* \geq f(x^k)$ e $\alpha_k \in (0, 1)$ sendo a maior raiz da equação $L^\alpha \alpha_k^2 = \sigma(1 - \alpha_k)\gamma_k$. Se $z = x_k + \beta(v^k - x^k)$ está no segmento $[x^k, v^k]$ e é tal que $f(z) \leq f(x^k)$ e se:*

$$y^k = x^k + \theta(v^k - x^k), \text{ com } \theta = \beta + \alpha_k(1 - \beta), \quad (3.27)$$

v^{k+1} dado pelo Lema 3.4. Se ainda,

$$\hat{x} = y^k + \alpha_k(v^{k+1} - v^k), \quad (3.28)$$

e $\gamma_{k+1} = (1 - \alpha_k)\gamma_k$ e $L^\alpha > 0$ satisfaz a desigualdade (3.26), então, a iteração termina com

$$\phi_{k+1}^* \geq f(x^{k+1}). \quad (3.29)$$

Abaixo, apresentamos uma descrição completa do método de Rossetto [29]:

Algoritmo 3.3 *Método com projeção não interior*

- 1: Dados: $x^0 \in C$, L_0 , $\gamma_0 = 1$, $\sigma > 0$ e $\alpha \in (0, 1)$
- 2: Faça $k = 0$, $v^0 = x^0$
- 3: Repita
- 4: Faça $L = L_k$, $\gamma = \gamma_k$
- 5: Calcule a maior raiz da equação $L^\alpha \alpha^2 = \sigma(1 - \alpha)\gamma$
- 6: $\gamma_+ = (1 - \alpha)\gamma$
- 7: Determine $z = x^k + \beta(v^k - x^k)$ com $\beta \in [0, 1]$, tal que $f(z) \leq f(x^k)$
- 8: Defina $\theta = \beta + \alpha(1 - \beta)$ e $y = x^k + \theta(v^k - x^k)$
- 9: $v^+ = \operatorname{argmin} \{D_\varphi(x, \bar{v}) | x \in C\}$
- 10: $\hat{x} = x^k + \alpha(v^+ - x^k)$
- 11: Se $f(y) + \langle \nabla f(y), \hat{x} - y \rangle + \frac{L^\alpha}{2} \|\hat{x} - y\|^2 < f(\hat{x})$, faça $L = 2L$
- 12: Até que $f(y) + \langle \nabla f(y), \hat{x} - y \rangle + \frac{L^\alpha}{2} \|\hat{x} - y\|^2 \geq f(\hat{x})$.
- 13: Defina:

- 14: $y^k = y, \gamma_{k+1} = \gamma_+, v^{k+1} = v^+, L_{k+1} = \frac{L}{2}$
 15: *Escolha, se possível, x^{k+1} tal que $f(x^{k+1}) \leq f(\hat{x})$*
 16: $k = k + 1$
 17: *Volte ao passo 3.*

Ao apresentar este método, Rossetto ([29], pp.59-60) afirma que:

1. Se L^α não for aceito pelo algoritmo (isso ocorre se a desigualdade no passo 15 não ocorrer), sabemos que $L^\alpha < L$ e como o próximo valor testado pelo algoritmo é sempre o dobro do valor anterior podemos concluir que L^α é sempre menor ou igual a $2L$;
2. Em cada iteração do laço interno do algoritmo o gradiente é avaliado apenas uma vez e são feitas duas avaliações da função;
3. Usando esses dois fatos, Rossetto apresentou um limitante para o número de avaliações do gradiente e da função.

O número de avaliações do gradiente (NG_k) é $NG_k \leq 2(K + 1) + \log_2 \frac{L}{L_0}$, e o número de avaliações da função (NF_k) é $NF_k \leq 4(K + 1) + \log_2 \frac{L}{L_0}$.

Isso garante que o laço interno será feito em número finito e constante de vezes.

O Lema 3.9 e o Teorema 3.4 garantem a otimalidade do método de Rossetto [29], são equivalentes ao Lema 3.5 e ao Teorema 3.2, respectivamente. Para o método de Auslender e Teboulle eles diferem apenas pelo fato de que a proposta de Rossetto usa um valor aproximado para a constante de Lipschitz com a garantia que este valor nunca é maior que duas vezes o valor verdadeiro de L , enquanto que no método de Auslender e Teboulle usamos o valor verdadeiro de L .

Lema 3.9 *Sejam $\gamma_k \geq 0$ e $\alpha_k \in (0, 1)$ definidos por $L^\alpha \alpha_k^2 = \sigma(1 - \alpha_k)\gamma_k$ e $\gamma_{k+1} = (1 - \alpha_k)\gamma_k$. Como $\lambda_k := \prod_{i=0}^{k-1} (1 - \alpha_i)$, então:*

$$\lambda_k \leq \frac{8L}{\left(2\sqrt{2L} + k\sqrt{\sigma\gamma_0}\right)^2}. \quad (3.30)$$

Rossetto [29] mostra, no próximo teorema, que o algoritmo proposto encontra a solução ótima com precisão ϵ com um número de iterações proporcional a $\frac{1}{\sqrt{\epsilon}}$, garantindo, assim sua otimalidade.

Teorema 3.4 *Sejam $\gamma_0 = 1$ e $\{x^k\}, \{y^k\}$ e $\{\alpha^k\}$ as sequências geradas pelo Algoritmo 3.3 e seja x^* a solução ótima do Problema 3.1. Então, para qualquer $k \geq 0$ temos:*

$$f(x^k) - f(x^*) \leq \frac{8L}{\sigma\gamma_0 k^2} C(x^*, x^0) = \mathcal{O}\left(\frac{1}{k^2}\right), \quad (3.31)$$

onde $C(x^*, x^0) = f(x^0) + D_\varphi(x^*, x^0) - f(x^*)$.

3.3 As modificações propostas por Tseng (2008)

Paul Tseng, em 2008, em seu paper intitulado "*On accelerated proximal gradient methods for convex-concave optimization*" [30] - Métodos de aceleração gradiente proximais para otimização concava-convexa, apresentou um tratamento unificado para os métodos de Nesterov usando funções de Bregman, que descreveremos no decorrer desta seção.

3.3.1 Introdução

Seja E um espaço vetorial real dotado de uma norma $\|\cdot\|$. Seja E^* o espaço vetorial de funcionais lineares contínuos em E , dotado de uma norma dual $\|x^*\|_* = \sup_{\|x\| \leq 1} \langle x^*, x \rangle$, onde $\langle x^*, x \rangle$ denota o valor de $x^* \in E^*$ e $x \in E$. Consideremos o problema de otimização convexa não diferenciável

$$\min_x f^P(x), \quad \text{em que } f^P(x) = f(x) + P(x) \quad (3.32)$$

onde $P : E \rightarrow (-\infty, \infty]$ e $f : E \rightarrow (-\infty, \infty]$ são próprias, semicontínuas inferiormente (lsc) e convexas. Para este problema assumimos que $\text{dom}P$ é fechado, que f é diferenciável no conjunto aberto contendo $\text{dom}P$ e ∇f é Lipschitz contínua em $\text{dom}P$, isto é:

$$\|\nabla f(x) - \nabla f(y)\|_* \leq L\|x - y\| \quad (3.33)$$

para algum $L > 0$ (TSENG, 2008)[30].

Note que, diferentemente do que foi tratado até agora, Tseng [30] usa a norma dual para gradiente da função. Este autor nos apresenta alguns exemplos para o Problema 3.32 que serão especificados a seguir.

Para qualquer $y \in \text{dom}P$ considere a aproximação de f^P substituindo f pela sua aproximação linear de y , dada da seguinte maneira:

$$l_f(x, y) := f(y) + \langle \nabla f(y), x - y \rangle + P(x) \quad (3.34)$$

Exemplo 3.1 Um caso bastante conhecido é a otimização convexa restrita suave, onde P é a função de indicador para um conjunto não vazio, fechado e convexo $X \subseteq E$, isto é:

$$P(x) = \begin{cases} 0, & \text{se } x \in X \\ \infty, & \text{caso contrário} \end{cases} \quad (3.35)$$

Este exemplo se resume ao caso tratado por Nesterov, apresentado na seção 3.1.

Exemplo 3.2 Um segundo caso, é quando P é a norma $\|\cdot\|_1$. Esta escolha para P tem sido usada em várias aplicações esparsas como em [10], [31] e [6].

Exemplo 3.3 O terceiro exemplo é o grupo para o qual P é uma soma ponderada de $\|\cdot\|_2$, isto é:

$$P(x) = c_1 \|x^1\|_2 + \cdots + c_N \|x^N\|_2 \quad (3.36)$$

onde x^1, \dots, x^N denotam subvetores de x , disjuntos, ou seja, não têm componentes em comum e $c_j > 0 \forall j$.

Fundamentado nos trabalhos de Polyak [27]; Fukushima e Mine [9]; Nemirovski e Yudin [18]; Nesterov [20]; Nesterov ([22] e [23]); Auslender e Teboulle [1]; d'Aspremont [8]; Lu [15]; Lu, Monteiro e Yuan [16]; Beck e Teboulle [4], Tseng [30] propõe uma estrutura unificada e uma análise mais simples dos métodos de ordem $\mathcal{O}\left(\sqrt{\frac{L}{\epsilon}}\right)$, estendendo a solução do problema (3.32). Como um subproduto, Tseng [30] derivou, novas variantes e refinamentos dos métodos de ordem $\mathcal{O}\left(\sqrt{\frac{L}{\epsilon}}\right)$ que usam uma ou duas projeções por iteração (isso pode ser verificado nos Corolários 3.1, 3.2 e 3.3 da Seção 3.3.3). O ponto principal da análise feita por Tseng [30] são as duas propriedades de distância de Bregman descritas na próxima seção e a reinterpretação dos métodos contidos nos trabalhos supracitados. Além disso, Tseng [30] faz uma comparação dos métodos acima referidos por meio de exemplos numéricos e discute extensões dos resultados obtidos (para aprofundar o estudo dessa comparação sugerimos consultar [30] seções 6 e 7, visto que este aprofundamento foge do escopo do nosso trabalho e, por isso, não o faremos).

3.3.2 Propriedades das funções de distância de Bregman

Listaremos agora, fundamentados em Tseng [30], duas propriedades de distância utilizadas para provar os resultados apresentados nesta seção.

Propriedade 3.1 *Para qualquer função semicontínua inferiormente, convexa e própria $\psi : E \rightarrow (-\infty, \infty]$, se:*

$$z_+ = \arg \min_x \{\psi(x) + D(x, z_+)\}$$

e φ é diferenciável em z , então:

$$\psi(x) + D(x, z) \geq \psi(z) + D(z_+, z) + D(x, z_+) \quad \forall x \in \text{dom} P \quad (3.37)$$

Isso decorre da condição de otimização para $z_+ : \psi(x) + \langle \nabla_x D(z_+, z), x - z_+ \rangle \geq \psi(z_+) \quad \forall z$ e $\nabla_x D(z_+, z) = \nabla \varphi(z_+) - \nabla \varphi(z)$. De fato, reorganizando os termos, temos:

$$\psi(x) - \langle \nabla \varphi(z), x - z \rangle \geq \psi(z_+) - \langle \nabla \varphi(z), z_+ - z \rangle - \langle \nabla \varphi(z_+), x - z_+ \rangle,$$

e somando $\varphi(x) - \varphi(z)$ em ambos os lados concluímos (3.37).

Propriedade 3.2 *Para qualquer função lsc convexa própria $\psi : E \rightarrow (-\infty, \infty]$, se:*

$$z_+ = \arg \min_x \{\psi(x) + \varphi(x)\},$$

e φ é diferenciável em z , então:

$$\psi(x) + \varphi(x) \geq \psi(z) + \varphi(z) + D(x, z) \quad \forall x \in \text{dom} P.$$

Isso decorre da condição de otimização para z : $\psi(x) + \langle \nabla \varphi(z), x - z_+ \rangle \geq \psi(z) \quad \forall z$.

A partir de agora enunciaremos os algoritmos e os resultados propostos por Tseng [30].

3.3.3 Método de otimização convexa de ordem $\mathcal{O}\left(\sqrt{\frac{L}{\epsilon}}\right)$

Nesta seção, apresentamos uma estrutura unificada para o segundo método de Nesterov e suas variantes para a solução de (3.32).

Algoritmo 3.4 Tome $\theta_0 \in (0, 1]$, $x^0, z^0 \in \text{dom}P$ e $k \leftarrow 0$.

Passo 1. Escolha um conjunto fechado, convexo e não vazio $X_k \subseteq E$ de forma que $X_k \cap \text{dom}(P) \neq \emptyset$. Seja:

$$y^k = (1 - \theta_k)x^k + \theta_k z^k \quad (3.38)$$

$$z^{k+1} = \arg \min_{x \in X_k} \{l_f(x, y^k) + \theta_k LD(x, z^k)\} \quad (3.39)$$

$$\hat{x}^{k+1} = (1 - \theta_k)x^k + \theta_k z^{k+1} \quad (3.40)$$

onde $l_f(\cdot, \cdot)$ é dado por (3.34) *Passo 2.* x_{k+1} deve ser tomado de modo que:

$$l_f(x^{k+1}, y^k) + \frac{L}{2} \|x^{k+1} - y^k\|^2 \leq l_f(\hat{x}^{k+1}, y^k) + \frac{L}{2} \|\hat{x}^{k+1} - y^k\|^2 \quad (3.41)$$

Passo 3. Tome $\theta_{k+1} \in (0, 1]$ satisfazendo:

$$\frac{1 - \theta_{k+1}}{\theta_{k+1}^2} \leq \frac{1}{\theta_k^2}, \quad (3.42)$$

$k \leftarrow k + 1$ e retorne ao passo 1.

De acordo com Tseng [30], o Algoritmo 3.4 produz uma sequência x^{k+1} que deve obedecer às condições citadas no passo 2. Além disso, o conjunto X_k deve ser escolhido de modo que contenha uma solução de (3.32). A escolha mais simples é $X_k = E$, mas pode ser desejável utilizar um conjunto X_k menor, para acelerar a convergência. Uma sugestão é escolher X_k da seguinte maneira:

$$X_k = \{x / \sum_{i \in I_{k,j}} a_{k,i} (l_f(x, w^{k,i}) - f^P(w^{k,i})) \leq \epsilon, j = 1, \dots, n_k\} \quad (3.43)$$

com $w_{k,i} \in \text{dom}P$. Por exemplo: se $w^{k,i} \in \{x^0, x^1, \dots, x^k, z^0, \dots, z^k\}$, $\alpha_{k,i} \geq 0$, $\sum_{i \in I_{k,j}} a_{k,i} = 1$, $I_{k,j} \subseteq \{1, 2, \dots\}$ e $n_k \geq 0$, então X_k contém todos pontos de ϵ -mínimo de f^P , onde um ponto ϵ -mínimo de f^P é definido como $f^P(x) \leq \inf f^P + \epsilon$ para $\epsilon \geq 0$.

Tseng [30] afirma, ainda, que uma escolha para x^{k+1} é:

$$x^{k+1} = \arg \min_x \{l_f(x, y^k) + \frac{L}{2} \|x - y^k\|^2\}. \quad (3.44)$$

E que, no caso do Exemplo 3.35 com $X_k = E$ e tomando x^{k+1} como em (3.44), o Algoritmo 3.4 se reduz ao caso estudado por Lan, Lu e Monteiro [14] que é um variante do método de Nesterov.

Uma outra escolha para x^{k+1} é

$$x^{k+1} = \hat{x}^{k+1}.$$

E neste caso, o Exemplo 3.35 com o Algoritmo 3.4 e $X_k = E$, esta escolha de x^{k+1} se reduz à extensão de Auslender e Teboulle do método de Nesterov, onde o termo quadrático é substituído por uma função de Bregman.

Para simplificação de notação denotamos:

$$\triangle_f(x, y) := f^P(x) - l_f(x, y) = f(x) - f(y) - \langle \nabla f(y), x - y \rangle \quad \forall x, y \in \text{dom} P.$$

Proposição 3.2 *Seja $\{(x^k, y^k, z^k, \theta_k, X_k)\}$ gerado pelo Algoritmo 3.4. Para todo $k = 0, 1, \dots$ e para todo $x \in X_k \cap \text{dom} P$, se $f^P(x) \leq f^P(x^{k+1})$ ou $\theta_{k+1} = \frac{\sqrt{\theta_k^4 + 4\theta_k^2 - \theta_k^2}}{2}$, então:*

$$\frac{1 - \theta_{k+1}}{\theta_{k+1}^2} (f^P(x^{k+1}) - f^P(x^k)) + LD(x, z^{k+1}) \leq \frac{1 - \theta_k}{\theta_k^2} (f^P(x^k) - f^P(x)) + LD(x, z^k) - \frac{\triangle_f(x, y^k)}{\theta_k} \quad (3.45)$$

Corolário 3.1 *Seja $\{(x^k, y^k, z^k, \theta_k, X_k)\}$ gerado pelo Algoritmo 3.4 para $\theta_0 = 1$.*

(a) *Fixar qualquer $\epsilon > 0$. Suponha $\theta_k \leq \frac{2}{k+2}$ e que X_k é dado pela equação (3.43) para todo k . Em seguida, para todo $x \in \text{dom} P$ com $f^P(x) \leq \inf f^P + \epsilon$, temos:*

$$\min_{i=0,1,\dots,k+1} f^P(x^i) \leq f^P + \epsilon, \quad \text{quando,} \quad k \geq \sqrt{\frac{4LD(x, z^0)}{\epsilon}} - 2$$

(b) *Suponha E de dimensão finita, $\text{dom} P$ é limitado, $f(x) = \max_{u \in U} \phi(x, u)$, $\theta_{k+1} = \frac{\sqrt{\theta_k^4 + 4\theta_k^2 - \theta_k^2}}{2}$ e $X_k = E$. Então:*

$$0 \leq f^P(x^{k+1}) - q^P(\hat{u}^k) \leq \theta_k^2 L \max_{x \in \text{dom} P} D(x, z^0), \quad k = 0, 1, \dots,$$

onde $q^P(u) := \min_x \{\phi(x, u) + P(x)\}$, $u^k = \arg \max_u \phi(y^k, u)$, $(\hat{u}^{-1} = 0)$ e

$$\hat{u}^k = \frac{\left(\sum_{i=0}^k \frac{u^i}{\theta_i}\right)}{\left(\sum_{i=0}^k \frac{1}{\theta_i}\right)} = (1 - \theta_k) \hat{u}^{k-1} + \theta_k u^k.$$

No restante dessa seção assumiremos que f é diferenciável em E e que E é um espaço de Hilbert, isto é, $E^* = E$, $\|\cdot\| = \sqrt{\langle \cdot, \cdot \rangle}$. O seguinte algoritmo é uma extensão do primeiro método de Nesterov, da mesma ordem do Algoritmo 3.4 (TSENG, 2008) [30].

Algoritmo 3.5 *Tome $\theta_0 = \theta_{-1} \in (0, 1]$, $x^0 = x^{-1} \in \text{dom} P$ e $k = 0$.*

*Passo 1. Escolher um conjunto fechado, convexo e não vazio $X_k \subseteq E$ tal que $X_k \cap \text{dom}P \neq \emptyset$.
Sejam:*

$$y^k = x^k + \theta_k (\theta_{k-1}^{-1} - 1)(x^k - x^{k-1}) \quad (3.46)$$

$$x^{k+1} = \arg \min_{x \in X_k} \left\{ l_f(x, y^k) + \frac{L}{2} \|x - y^k\|^2 \right\} \quad (3.47)$$

Passo 2. Escolher $\theta_{k+1} \in (0, 1]$ satisfazendo (3.42). $k \leftarrow k + 1$ retorne ao passo 1.

De acordo com Tseng [30], o conjunto X_k e a constante θ_k podem ser escolhidos como no Algoritmo 3.4. Uma boa escolha é $\theta_k = \frac{2}{k+2}$, pois a convergência é mais rápida. Pode acontecer de y^k estar fora do $\text{dom}P$ e, portanto, precisamos que f seja diferenciável fora do $\text{dom}P$.

Uma prova simples da complexidade do Algoritmo 3.5 é dada abaixo, usando propriedades de norma e a propriedade (3.1).

Proposição 3.3 *Suponha f diferenciável em E e $E = E^*$, $\|\cdot\| = \sqrt{\langle \cdot, \cdot \rangle}$. Tomando $(x^k, y^k, \theta_k, X_k)$ gerados pelo Algoritmo 3.1. Para cada $k = 0, 1, \dots$ e para todo $x \in X_k \cap \text{dom}P$, se $f^p(x) \leq f^p(x^{k+1})$ ou $\theta_{k+1} = \frac{\sqrt{\theta_k^4 + 4\theta_k^2} - \theta_k^2}{2}$ ocorre, então, que a equação (3.45) é válida com $D(x, z) = \frac{1}{2} \|x - z\|^2$ e $z^k = x^{k-1} + \theta_{k-1}^{-1}(x^k - x^{k-1})$.*

Corolário 3.2 *Seja f diferenciável em E e $E = E^*$, $\|\cdot\| = \sqrt{\langle \cdot, \cdot \rangle}$. Tomando $\{(x^k, y^k, \theta_k, X_k)\}$ gerados pelo Algoritmo 3.5 e $\theta_0 = 1$. Então, a afirmação do Corolário 3.1 se mantém com $z^0 = x^0$.*

O método de Nesterov de ordem $\mathcal{O}\left(\sqrt{\frac{L}{\epsilon}}\right)$, tratado em [22], é notavelmente diferente dos métodos de Auslender e Teboulle em [2], ou ainda do método proposto por Lan, Lu e Monteiro em [14] e dos próprios métodos de Nesterov ([20], [21] e [19]), na medida em que utilizam, em cada iteração, duas projeções, além de uma soma ponderada de gradientes. A partir de agora apresentaremos uma análise, ou seja, "um framework" do método de Nesterov para estender a solução de (3.32), feita por Tseng [30]. Essa análise fornece a possibilidade de com apenas uma projeções por iteração obter a mesma complexidade do método de Nesterov e, além disso, apresenta uma estreita conexão entre os métodos de ordem $\mathcal{O}\left(\sqrt{\frac{L}{\epsilon}}\right)$ (TSENG, 2008) [30].

Algoritmo 3.6 *Tome $0 < \theta_0 \leq 1$, $v_0 \geq \theta_0$, $x^0 \in \text{dom}P$. Tome $z^0 = \arg \min_{x \in \text{dom}P} \varphi(x)$, $X_{-1} = E$ e $k = 0$.*

Passo 1. Considere um conjunto fechado, convexo e não vazio $X_k \subseteq X_{k+1}$ tal que $X_k \cap \text{dom}P \neq \emptyset$. Tome :

$$\begin{aligned} y^k &= (1 - \theta_k)x^k + \theta_k z^k \\ \psi_{k+1}(x) &:= \sum_{i=0}^k \frac{l_f(x; y^i)}{v_i}, \quad \forall x \\ z^{k+1} &= \arg \min_{x \in X_k} \{ \psi_{k+1}(x) + L\varphi(x) \} \end{aligned}$$

e, tome ainda, x^k dado por (3.40), (3.41) no Algoritmo 3.4.

Passo 2. Considere $0 < \theta_{k+1} \leq 1$, $v_{k+1} \geq \theta_{k+1}$ satisfazendo

$$\frac{1 - \theta_{k+1}}{\theta_{k+1}v_{k+1}} = \frac{1}{\theta_k v_k} \quad (3.48)$$

$k \leftarrow k + 1$ e ir para 1.

Podemos tomar $X_k = E$ ou X_k dado por (3.43), uma vez que $X_k \subseteq X_{k-1}$. O método de Nesterov (2005) dado em [22], corresponde ao algoritmo acima com $X_k = E$, x^{k+1} dado por (3.44), e

$$\theta_k = \frac{2}{k+2}, \quad v_k = \frac{2}{k+1} \quad \text{que satisfazem (3.48)} \quad (3.49)$$

Nesterov [24], propôs um método para resolver o problema (3.32), similar ao Algoritmo 3.6, onde x^{k+1} é dado pela equação (3.44), $D(x, y) = \frac{1}{2}\|x - y\|^2$ e $X_k = E$. Em Nesterov [23] é apresentada uma modificação do método de Fukushima e Mine [9] que substitui:

$$\hat{z}^{k+1} = \arg \min_{x \in X_k} \{l_f(x, y^k) + \theta_k LD(x, z^k)\} \quad (3.50)$$

$$x^{k+1} = (1 - \theta_k)x^k + \theta_k \hat{z}^{k+1} \quad (3.51)$$

Tseng [30], observou que esta modificação pode ser vista como uma mistura dos Algoritmos 3.4 e 3.6. A proposição abaixo apresenta uma demonstração da complexidade do Algoritmo 3.6 que tem ordem $\mathcal{O}\left(\sqrt{\frac{L}{\epsilon}}\right)$.

Proposição 3.4 *Seja $\{(x^k, y^k, z^k, \theta_k, v_k, X_k)\}$ gerado pelo Algoritmo 3.6 ou essa modificação em que (3.40), (3.41) são substituídos por (3.50), (3.51) e $\psi_0 = \delta_{\text{dom}P}$. Então, para cada $k = 0, 1, \dots$, temos:*

$$\frac{1 - \theta_{k+1}}{\theta_{k+1}v_{k+1}} f^P(x^{k+1}) - (\psi_{k+1} + Lh)(z^{k+1}) \leq \frac{1 - \theta_k}{\theta_k v_k} f^P(x^k) - (\psi_k + L\varphi)(z^k).$$

Corolário 3.3 *Seja $\{(x^k, y^k, z^k, \theta_k, v_k, X_k)\}$, gerado pelo Algoritmo 3.6 ou pela modificação descrita na Proposição 3.4, com $\theta_0 = 1$:*

(a) Para cada $k \geq 0$ e cada $x \in X_k \cap \text{dom}P$, temos:

$$\frac{1}{\theta_k v_k} (f^P(x^{k+1}) - f^P(x^k)) + \sum_{i=0}^k \frac{\Delta_f(x; y^i)}{v^i} \leq L(\varphi(x) - \varphi(z^0)).$$

(b) Fixar qualquer $\epsilon > 0$. Suponha $v_k \leq \frac{2}{k+1}$ e que X_k é dado por (3.43) para todo k . Então, para qualquer $x \in \text{dom}P$ com $f^P(x) \leq \inf f^P + \epsilon$, temos:

$$f^P(x^{k+1}) \leq f^P(x) + \epsilon \quad \text{sempre que} \quad k \geq \sqrt{\frac{4L(\varphi(x) - \varphi(z^0))}{\epsilon}} - 1$$

Esses são os resultados da complexidade dos algoritmos propostos por Tseng [30]. As demonstrações foram omitidas pois o foco principal do nosso trabalho são as demonstrações do trabalho *An accelerated linearized alternating direction method of multipliers*, de Ouyang, Chen, Lan e Pasiliao [26].

Capítulo 4

Método das direções alternadas acelerado e linearizado de multiplicadores

Neste capítulo é estudado com detalhes os resultados apresentados por Ouyang, Chen, Lan e Pasiliao [26], que apresentam um novo framework para aceleração do método linear das direções alternadas para multiplicadores denominado AADMM - *acceleration of linearized alternating direction method of multipliers*, para resolver problemas compostos e irrestritos de otimização denominados por UCO e problemas compostos cuja restrição é uma equação afim, aqui denominados AECCO. A ideia básica proposta pelos autores é a incorporação de uma técnica de passo múltiplo na versão linearizada ADMM para aceleração. Com isso, consegue-se obter uma técnica AADMM que apresenta uma ordem de convergência superior à versão linearizada ADMM em termos da dependência da constante de Lipschitz da componente suave quando usados para resolver uma classe de problemas de otimização composta convexa com restrições lineares. Mostram ainda que o método proposto, além de acelerar a convergência pode ser usado em situações onde a região viável não é limitada, desde que o problema ponto de sela correspondente tenha solução.

O método de Nesterov, quando aplicado a problemas de UCO, tem taxa de convergência ótima cuja ordem é:

$$\mathcal{O}\left(\frac{L_G D_{x^*}^2}{N^2} + \frac{\|K\| D_{x^*} D_Y}{N}\right) \quad (4.1)$$

onde Y é o espaço dual limitado para problemas UCO, L_G descrita em (4.4), K é um operador linear limitado e N a quantidade de iterações.

Após o trabalho de Nesterov (2005) [22], muitos outros trabalhos visaram tornar métodos de primeira ordem para otimização não diferenciável mais eficientes como, por exemplo, o trabalho do próprio Nesterov [23] e os trabalhos de Auslender e Teboulle [2], Tseng [30] e Lan, Lu e Monteiro [14]. Embora a taxa de (4.1) também seja de ordem $\mathcal{O}\left(\frac{1}{N}\right)$ o que o diferencia é o fato de permitir constante de Lipschitz tão grande quanto se queira, sem que esta afete a taxa de

convergência (a menos de um fator constante). Entretanto, de acordo com Ouyang et al. [26], o fato de o conjunto viável Y ser limitado é crítico para a análise do esquema de suavização de Nesterov. Seguindo o esquema de Nesterov em [22] vários estudos sobre os problemas AECCO e UCO têm ocorrido, foi demonstrado que melhores resultados de aceleração podem ser obtidos considerando mais algumas hipóteses para os problemas AECCO e UCO. Essas hipóteses e resultados são encontradas em Ouyang et al. [26].

4.1 Condições do problema

Suponha que $\mathcal{W}, \mathcal{X}, \mathcal{Y}$ sejam espaços vetoriais munidos de produto interno $\langle \cdot, \cdot \rangle$, norma $\| \cdot \|$ e norma dual $\| \cdot \|_*$. O problema de interesse é o que chamaremos de AECCO (*affine equality constrained composite optimization*):

$$\min_{x \in X, w \in \mathcal{W}} G(x) + F(w), \quad \text{onde } Bw - Kx = b \quad (4.2)$$

em que $X \subseteq \mathcal{X}$ é um conjunto fechado e convexo, $G(\cdot) : X \rightarrow \mathbb{R}$ e $F(\cdot) : \mathcal{W} \rightarrow \mathbb{R}$ são funções convexas, semicontínuas inferiormente e assumem valores finitos. Ainda $K : X \rightarrow \mathcal{Y}$, $B : \mathcal{W} \rightarrow \mathcal{Y}$ são operadores lineares limitados.

De acordo com Ouyang et al. [26], assumimos que $F(\cdot)$ é solucionável no sentido de que o problema de otimização

$$\min_{w \in \mathcal{W}} \frac{\eta}{2} \|w - c\|^2 + F(w), \quad \text{onde } c \in \mathcal{W}, \eta \in \mathbb{R} \quad (4.3)$$

pode ser resolvido de maneira eficiente e o termo não solucionável é usado no sentido oposto. Assumimos que $G(\cdot)$ é não solucionável, continuamente diferenciável e que existe uma constante $L_G > 0$ tal que:

$$G(x_2) - G(x_1) - \langle \nabla G(x_1), x_2 - x_1 \rangle \leq \frac{L_G}{2} \|x_2 - x_1\|^2, \quad \forall x_1, x_2 \in X. \quad (4.4)$$

Um caso especial de problema AECCO é quando $B = I$ e $b = 0$. Nessas condições, o problema (4.2) é equivalente ao seguinte problema de otimização composta irrestrito, que Ouyang et al. [26] denominam de UCO (*unconstrained composite optimization* - otimização composta irrestrita):

$$\min_{x \in X} f(x) := G(x) + F(Kx) \quad (4.5)$$

Os problemas AECCO e UCO têm sido muito usados em aplicações em aprendizado de máquina (aprendizagem automática) e processamento de imagem. Na maioria das aplicações, $G(\cdot)$ é conhecido como o termo fidelidade e $F(\cdot)$ é o termo de regularização.

Salientamos aqui que os problemas do tipo AECCO e UCO podem ser reformulados como problemas de ponto de sela utilizando o método dos multiplicadores de Lagrange. O problema AECCO é equivalente ao seguinte problema de ponto de sela:

$$\min_{x \in X, w \in \mathcal{W}} \max_{y \in \mathcal{Y}} G(x) + F(w) - \langle y, Bw - Kx - b \rangle \quad (4.6)$$

A partir dessa equivalência, Ouyang, Chen, Lan e Pasiliao propõem algoritmos, proposições, teoremas e corolários nos quais focamos nossos esforços em demonstrá-los com cuidado.

4.1.1 Notações

Assumimos que existe uma solução ótima (w^*, x^*) do problema (4.2) e que existe $y^* \in \mathcal{Y}$ tal que $z^* := (w^*, x^*, y^*) \in \mathcal{Z}$ é um ponto de sela de (4.6), onde $\mathcal{Z} = \mathcal{W} \times \mathcal{X} \times \mathcal{Y}$ e se o conjunto $Y \subseteq \mathcal{Y}$ é dado usamos $\mathcal{Z} = \mathcal{W} \times \mathcal{X} \times Y$. Considere $f^* := G(x^*) + F(w^*)$ o valor objetivo ótimo do Problema 4.2. Nos problemas UCO (4.5), que são os casos especiais de AECCO (4.2), também usaremos f^* para denotar o valor ótimo para $G(x^*) + F(Kx^*)$. Considerando que tanto o valor da função objetivo bem como a viabilidade da restrição interferem na solução do Problema 4.2, a definição de solução aproximada é dada por:

Definição 4.1 *Um par $(w, x) \in \mathcal{W} \times X$ é chamado de (ϵ, δ) -solução do problema (4.2) se:*

$$G(x) + F(w) - f^* \leq \epsilon \quad e \quad \|Bw - Kx - b\| \leq \delta.$$

Neste caso, dizemos que (w, x) tem resíduo primal ϵ e resíduo de viabilidade δ . Em particular, se (w, x) é uma $(\epsilon, 0)$ -solução, então dizemos que é uma ϵ -solução.

O resíduo de viabilidade δ na definição acima mensura a violação da restrição de igualdade, bem como o resíduo primal ϵ mensura a diferença (gap) entre o valor objetivo $G(x) + F(w)$ na solução aproximada e o valor ótimo f^* .

As seguintes notações serão utilizadas para simplificar:

$$\begin{aligned} D_{w^*, B} &:= \|B(w_1 - w^*)\| \\ D_{x^*, K} &:= \|K(x_1 - x^*)\| \\ D_{x^*} &:= \|x_1 - x^*\| \\ D_{y^*} &:= \|y_1 - y^*\| \\ D_{X, K} &:= \sup_{x_1, x_2 \in \mathcal{X}} \|Kx_1 - Kx_2\| \\ D_S &:= \sup_{s_1, s_2 \in \mathcal{S}} \|s_1 - s_2\| \quad \text{para qualquer conjunto compacto } S. \end{aligned} \quad (4.7)$$

Além disso, utilizamos $x_{[t]}$ para denotar a sequência $\{x_i\}_{i=1}^t$, onde x'_i s podem ser números reais ou pontos em espaços vetoriais. Também representaremos algumas operações na notação de sequências. Suponha que $\mathcal{V}_1, \mathcal{V}_2$ sejam espaços vetoriais, $v_{[t+1]} \subset \mathcal{V}_1$ é qualquer sequência em \mathcal{V}_1 e $A : \mathcal{V}_1 \rightarrow \mathcal{V}_2$ é qualquer operação. Usamos $Av_{[t+1]}$ para designar a sequência $\{Av_i\}_{i=1}^{t+1}$. Se $\eta_{[t]}, \tau_{[t]} \subset \mathbb{R}$ são quaisquer duas sequências a valores reais e $L \in \mathbb{R}$ é qualquer número real, então $\eta_{[t]} - L\tau_{[t]}$ denota $\{\eta_i - L\tau_i\}_{i=1}^t$ ainda, denotamos por $\eta_{[t]}^{-1}$ a sequência $\{\eta_i^{-1}\}_{i=1}^t$ para qualquer sequência $\eta_{[t]}$ diferente de zero (OUYANG et al., 2015) [26].

4.1.2 Lagrangeano aumentado e método de direção alternada de multiplicadores

Neste trabalho estudamos o problema AECCO via formulação lagrangeana, visto em (4.6):

$$\min_{x \in X, w \in \mathcal{W}} \max_{y \in \mathcal{Y}} G(x) + F(w) - \langle y, Bw - Kx - b \rangle + \frac{\rho}{2} \|Bw - Kx - b\|^2 \quad (4.8)$$

em que ρ é um parâmetro de penalidade.

Uma análise de (4.8) a fim de resolver o problema (4.2) é o método lagrangeano aumentado (ALM) dado por Hestenes [12] que é uma variação do algoritmo ADMM que, por sua vez, é um método para resolver o problema (4.8) de forma alternada que minimiza x e w e atualiza o coeficiente de Lagrange y . No que se segue, será apresentada uma revisão do ADMM e algumas variantes. O método de direção alternada de multiplicadores ADMM, para obter a solução de (4.2), é dado pelo seguinte algoritmo:

Algoritmo 4.1 *Passo 1. Tome $x_1 \in \mathcal{X}$, $w_1 \in \mathcal{W}$ e $y_1 \in \mathcal{Y}$ e para $t = 1, \dots, n - 1$ faça*

$$x_{t+1} = \arg \min_{x \in X} G(x) - \langle y_t, Bw_t - Kx - b \rangle + \frac{\rho}{2} \|Bw_t - Kx - b\|^2 \quad (4.9)$$

$$w_{t+1} = \arg \min_{w \in \mathcal{W}} F(w) - \langle y_t, Bw - Kx_{t+1} - b \rangle + \frac{\rho}{2} \|Bw - Kx_{t+1} - b\|^2 \quad (4.10)$$

$$y_{t+1} = y_t - \rho(Bw_{t+1} - Kx_{t+1} - b) \quad (4.11)$$

De acordo com Ouyang et al. [26], se G não for solucionável, uma linearização de ADMM, chamada de L-ADMM, gera um esquema onde x_{t+1} em (4.9) é substituído por:

$$x_{t+1} = \arg \min_{x \in X} \langle \nabla G(x_t), x \rangle + \langle y_t, Kx \rangle + \frac{\rho}{2} \|Bw_t - Kx - b\|^2 + \frac{\eta}{2} \|x - x_t\|^2. \quad (4.12)$$

Também podemos linearizar o termo $\|Bw - Kx_t - b\|^2$, e gerar x_{t+1} como:

$$x_{t+1} = \arg \min_{x \in X} G(x) + \langle y_t, Kx \rangle - \rho \langle Bw_t - Kx_t - b, Kx \rangle + \frac{\eta}{2} \|x - x_t\|^2. \quad (4.13)$$

Esta variação é chamada por Ouyang et al. [26], de ADMM pré-condicionada denotada por P-ADMM. Se $G(x)$ e $\|Bw - Kx_{t+1} - b\|^2$ forem linearizados o método é denominado LP-ADMM (*linearized preconditioned ADMM*) e a equação (4.9) é substituída por:

$$x_{t+1} = \arg \min_{x \in X} \langle \nabla G(x_t), x \rangle + \langle y_t, Kx \rangle - \rho \langle Bw_t - Kx_t - b, Kx \rangle + \frac{\eta}{2} \|x - x_t\|^2 \quad (4.14)$$

Existem vários trabalhos sobre a análise da convergência e aplicações de ADMM, L-ADMM e P-ADMM, dentre elas podemos destacar os trabalhos de Chambolle e Pock [5]; Chen, Hager, Huang, Phan, Ye e Yin [7]; Ye, Chen, Lin, e Huang [32] e Monteiro e Svaiter [17].

4.2 Um framework acelerado ADMM

Ouyang, et al. [26] propõem um framework para acelerar os métodos ADMM, ou seja, AADMM para resolver (4.2) e (4.5). O Algoritmo 4.2 apresenta a descrição deste framework. Em AADMM, χ é uma constante binária na equação (4.16), que vale 0 ou 1. O sobrescrito "ag" significa agregado e "md" significa "médio". O ponto x_t^{md} e os pontos agregados $w_{t+1}^{ag}, x_{t+1}^{ag}$ e y_{t+1}^{ag} são somas ponderadas de todas as iterações $\{x_i\}_{i=1}^t, \{w_i\}_{i=1}^{t+1}, \{x_i\}_{i=1}^{t+1}$ e $\{y_i\}_{i=1}^{t+1}$, respectivamente. Se os $\alpha_t \equiv 1$, então $x_t^{md} = x_t$ e os agregados são exatamente as iterações atuais w_{t+1}, x_{t+1} e y_{t+1} . Neste caso, se $\chi = 0$ e $\theta_t = \tau_t = \rho_t \equiv \rho$, então, AADMM torna-se ADMM, e se, além disso, G é solucionável, AADMM torna-se P-ADMM. Chamamos de aceleração para $\chi = 0$ o L-ADMM acelerado (AL-ADMM) e $\chi = 1$ o AL-ADMM acelerado (ALP-ADMM) (OUYANG et al., 2015) [26].

Algoritmo 4.2 *Framework ADMM acelerado (AADMM)*

1. Escolha $x_1 \in X$ e $w_1 \in \mathcal{W}$ tal que $Bw_1 = Kx_1 + b$.

2. Defina $x_1^{ag} = x_1, w_1^{ag} = w_1$ e $y_1^{ag} = y_1 = 0$.

Para $t = 1, \dots, N - 1$ faça

$$x_t^{md} = (1 - \alpha_t)x_t^{ag} + \alpha_t x_t \quad (4.15)$$

$$x_{t+1} = \arg \min_{x \in X} \langle \nabla G(x_t^{md}), x \rangle - \chi \theta_t \langle Bw_t - Kx_t - b, Kx \rangle + \frac{(1 - \chi)\theta_t}{2} \|Bw_t - Kx - b\|^2 + \langle y_t, Kx \rangle + \frac{\eta_t}{2} \|x - x_t\|^2 \quad (4.16)$$

$$x_{t+1}^{ag} = (1 - \alpha_t)x_t^{ag} + \alpha_t x_{t+1} \quad (4.17)$$

$$w_{t+1} = \arg \min_{w \in \mathcal{W}} F(w) - \langle y_t, Bw \rangle + \frac{\tau_t}{2} \|Bw - Kx_{t+1} - b\|^2 \quad (4.18)$$

$$w_{t+1}^{ag} = (1 - \alpha_t)w_t^{ag} + \alpha_t w_{t+1} \quad (4.19)$$

$$y_{t+1} = y_t - \rho_t (Bw_{t+1} - Kx_{t+1} - b) \quad (4.20)$$

$$y_{t+1}^{ag} = (1 - \alpha_t)y_t^{ag} + \alpha_t y_{t+1} \quad (4.21)$$

a saída do Algoritmo 4.2 é $z_N^{ag} = (w_N^{ag}, x_N^{ag})$.

Na formulação de ponto de sela usamos as funções gap, definidas abaixo:

Definição 4.2 Funções Gap: Para quaisquer $\hat{z} = (\hat{w}, \hat{x}, \hat{y}) \in \mathcal{Z}$ e $z = (w, x, y) \in \mathcal{Z}$ definimos:

$$Q(\hat{w}, \hat{x}, \hat{y}, w, x, y) := [G(x) + F(w) - \langle \hat{y}, Bw - Kx - b \rangle] - [G(\hat{x}) + F(\hat{w}) - \langle y, B\hat{w} - K\hat{x} - b \rangle] \quad (4.22)$$

denominaremos $Q(\hat{z}, z) = Q(\hat{w}, \hat{x}, \hat{y}, w, x, y), Q(\hat{z}, w, x, y)$ ou $Q(\hat{w}, \hat{x}, \hat{y}, z)$ para o mesmo significado.

Verificamos que $Q(z^*, z) \geq 0$ e $Q(z, z^*) \leq 0$ para todo $z \in Z$, onde z^* é um ponto de sela de (4.6).

De acordo com Ouyang et al. [26] para conjuntos compactos $W \subset \mathcal{W}$, $X \subset \mathcal{X}$, $Y \subset \mathcal{Y}$ a função gap dualidade

$$\sup_{\hat{w} \in W, \hat{x} \in X, \hat{y} \in Y} Q(\hat{w}, \hat{x}, \hat{y}, w, x, y) \quad (4.23)$$

mensura a precisão de uma solução aproximada (w, x, y) do problema ponto de sela

$$\min_{x \in X, w \in W} \max_{y \in Y} G(x) + F(w) - \langle y, Bw - Kx - b \rangle$$

Entretanto, o problema de interesse (4.2) tem a formulação de ponto de sela (4.6), em que o conjunto viável (W, X, Y) pode ser ilimitado. Um critério de finalização baseado em perturbação é empregado por Monteiro e Suaiter [17] para resolver as desigualdades variacionais e problemas ponto de sela. Ouyang et al. [26] modificam este critério e propõem uma versão modificada da função gap de (4.23). Apresentando a seguinte definição:

$$g_Y(v, z) := \sup_{\hat{y} \in Y} \{Q(w^*, x^*, \hat{y}; z) + \langle v, \hat{y} \rangle\} \quad (4.24)$$

para qualquer conjunto fechado $Y \subset \mathcal{Y}$, para qualquer $z \in Z$ e $v \in Y$. Pode-se, ainda, denotar:

$$\hat{g}_Y(z) := g_Y(0, z) = \sup_{\hat{y} \in Y} Q(w^*, x^*, \hat{y}; z) \quad (4.25)$$

Se $Y = \mathcal{Y}$, é omitido o subscrito Y e simplesmente usa-se a notação $g(v, z)$ e $\hat{g}(z)$.

As duas próximas proposições, descrevem a relação entre as funções gap (4.24)-(4.25) e as soluções aproximadas para os problemas (4.2) e (4.5).

Proposição 4.1 *Para qualquer $Y \subseteq \mathcal{Y}$, se $g_Y(Bw - Kx - b, z) \leq \epsilon < \infty$ e $\|Bw - Kx - b\| \leq \delta$ onde $z = (w, x, y) \in Z$, então (w, x) é uma (ϵ, δ) -solução de (4.2). Em particular no caso $Y = \mathcal{Y}$ para qualquer v de modo que $g(v, z) \leq \epsilon < \infty$ e $\|v\| \leq \delta$ temos $v = Bw - Kx - b$.*

Dem: Por (4.24), temos:

$$g_Y(v, z) = \sup_{\hat{y} \in Y} Q(w^*, x^*, \hat{y}; z) + \langle v, \hat{y} \rangle$$

substituindo $Q(w^*, x^*, \hat{y}; z)$ de acordo com (4.22), temos:

$$g_Y(v, z) = \sup_{\hat{y} \in Y} [G(x) + F(w) - \langle \hat{y}, Bw - Kx - b \rangle] - [G(x^*) + F(w^*) - \langle y, Bw^* - Kx^* - b \rangle] + \langle v, \hat{y} \rangle$$

o fato de w^* e x^* serem soluções ótimas de (4.2), torna nulo o termo $\langle y, Bw^* - Kx^* - b \rangle$ devido a restrição $Bw^* - Kx^* - b = 0$. Assim:

$$g_Y(v, z) = \sup_{\hat{y} \in Y} [G(x) + F(w) - \langle \hat{y}, Bw - Kx - b \rangle] - [G(x^*) + F(w^*)] + \langle v, \hat{y} \rangle.$$

Usando propriedades do sup e substituindo $G(x^*) + F(w^*)$ por f^* , temos:

$$g_Y(v, z) = G(x) + F(w) - f^* - \sup_{\hat{y} \in Y} \langle \hat{y}, Bw - Kx - b \rangle + \langle v, \hat{y} \rangle$$

usando a associatividade, simetria e a linearidade do produto interno temos:

$$g_Y(v, z) = G(x) + F(w) - f^* + \sup_{\hat{y} \in Y} \langle -\hat{y}, Bw - Kx - b - v \rangle$$

Usamos acima o fato de $\langle v, \hat{y} \rangle = \langle -v, -\hat{y} \rangle = \langle -\hat{y}, -v \rangle$.

Daí vemos que, se $g_Y(Bw - Kx - b, z) = G(x) + F(w) - f^* \leq \epsilon$ (pois se $v = Bw - Kx - b$ temos $\sup_{\hat{y} \in Y} \langle -\hat{y}, Bw - Kx - b - v \rangle = \sup_{\hat{y} \in Y} \langle -\hat{y}, 0 \rangle$) e $\|Bw - Kx - b\| \leq \delta$ então (w, z) por definição é uma (ϵ, δ) -solução de (4.2). Além disso, se $Y = \mathcal{Y}$ pode-se ver que $g(v, z) = \infty$ se $v \neq Bw - Kx - b$ (pois daí teríamos $\sup_{\hat{y} \in Y} \langle -\hat{y}, Bw - Kx - b - v \rangle = \infty$), consequentemente, se $g(v, z) < \infty \implies v = Bw - Kx - b$.

Proposição 4.2 Assuma que B é um operador injetor de modo que $B\mathcal{W} = \mathcal{Y}$ e $F(\cdot)$ é Lipschitz contínua, então o conjunto $Y := (B^*)^{-1} \text{dom} F^*$ é limitado. Além disso, se $\bar{g}_y(z) \leq \epsilon$, então o par (\hat{w}, x) é uma ϵ -solução de (4.2) onde $\hat{w} = (B^*)^{-1}(Kx + b)$.

Dem: Podemos ver que \hat{w} está bem definido desde que $B\mathcal{W} = \mathcal{Y}$. Além disso, usando o fato que $F(\cdot)$ assume valores finitos, pelo Teorema 1.1, temos que $\text{dom} F^*$ é limitado, portanto, Y é limitado.

Além disso, como $B\hat{w} - Kx - b = 0$, temos:

$$\bar{g}_y(z) = g_y(0, z) = \sup_{\hat{y} \in Y} Q(w^*, x^*, \hat{y}; z)$$

substituindo $Q(w^*, x^*, \hat{y}; z)$ de acordo com (4.22), temos:

$$\bar{g}_y(z) = \sup_{\hat{y} \in Y} [G(x) + F(w) - \langle \hat{y}, Bw - Kx - b \rangle] - [G(x^*) + F(w^*)]$$

Usando propriedades do sup, associatividade do produto interno, substituindo $G(x^*) + F(w^*)$ por f^* e $-Kx - b$ por $B\hat{w}$, temos:

$$\bar{g}_y(z) = G(x) + F(w) - f^* + \sup_{\hat{y} \in Y} \langle -\hat{y}, Bw - B\hat{w} \rangle$$

Usando propriedades do sup, a propriedade dos operadores adjuntos e somando e subtraindo $F(\hat{w})$, temos:

$$\bar{g}_y(z) = G(x) + F(\hat{w}) - f^* + \sup_{\hat{y} \in Y} [F(w) - F(\hat{w}) - \langle B^*\hat{y}, w - \hat{w} \rangle]$$

Usamos acima o fato de $\langle \hat{y}, Bw - B\hat{w} \rangle = \langle B^*\hat{y}, w - \hat{w} \rangle$, onde B^* é o operador adjunto de B .

Se $B^*Y \cap \partial F(\hat{w}) \neq \emptyset$ então, a partir da convexidade de $F(\cdot)$, temos:

$$\bar{g}_y(z) \geq G(x) + F(\hat{w}) - f^*,$$

assim, (\hat{w}, x) é uma ϵ -solução de (4.2). Para terminar a demonstração basta mostrar que $B^*Y \cap \partial F(\hat{w}) \neq \emptyset$. De fato, observando que:

$$\sup_{w' \in B^*Y} \langle \hat{w}, w' \rangle - F^*(w') = \sup_{w' \in \text{dom} F^*} \langle \hat{w}, w' \rangle - F^*(w') = \sup_{w' \in W} \langle \hat{w}, w' \rangle - F^*(w')$$

e utilizando o fato de \mathcal{Y} ser um conjunto fechado, pode-se concluir que $B^*\hat{y} \in B^*Y$ de tal modo que $B^*\hat{y}$ atinge o supremo da função $\langle \hat{w}, w' \rangle - F^*(w')$, com respeito a w' . Pelo Teorema 1.2, Seção 1.1 temos $B^*\hat{y} \in \partial F(\hat{w})$ e, portanto, $B^* \cap \partial F(\hat{w}) \neq \emptyset$. ■

4.3 Principais Estimativas

Para provar a taxa de convergência para AADMM são usadas algumas estimativas. Na sequência deste texto apresentamos as principais delas.

Lema 4.1 *Seja*

$$\Gamma_t = \begin{cases} \Gamma_1 = 1, & \text{se } \alpha_t = 1 \\ (1 - \alpha_t)\Gamma_{t-1}, & \text{se } t > 1 \end{cases} \quad (4.26)$$

para todos $y \in \mathcal{Y}$, as iterações $\{z_t^{ag}\}_{t \geq 1} := \{(w_t^{ag}, x_t^{ag}, y_t^{ag})\}_{t \geq 1}$ do Algoritmo 4.2 satisfazem:

$$\begin{aligned} & \frac{1}{\Gamma_t} Q(w^*, x^*, y; z_{t+1}^{ag}) - \sum_{i=2}^t \left(\frac{1 - \alpha_i}{\Gamma_i} - \frac{1}{\Gamma_{i-1}} \right) Q(w^*, x^*, y; z_i^{ag}) \\ & \leq \mathcal{B}_t(x^*, x_{[t+1]}, \eta_{[t]}) + \mathcal{B}_t(y, y_{[t+1]}, \rho_{[t]}^{-1}) + \mathcal{B}_t(Bw^*, Bw_{[t+1]}, \theta_{[t]}) - \chi \mathcal{B}_t(Kx^*, Kx_{[t+1]}, \theta_{[t]}) \\ & - \sum_{i=1}^t \frac{\alpha_i(\tau_i - \theta_i)}{2\Gamma_i} \|Bw_{i+1} - Kx^* - b\|^2 + \sum_{i=1}^t \frac{\alpha_i(\tau_i - \theta_i)}{2\Gamma_i} \|B(x_{i+1} - x^*)\|^2 \\ & - \sum_{i=1}^t \frac{\alpha_i(\tau_i - \rho_i)}{2\Gamma_i \rho_i^2} \|y_i - y_{i+1}\|^2 - \sum_{i=1}^t \frac{\alpha_i}{2\Gamma_i} (\eta_i - L_G \alpha_i - \chi \theta_i \|K\|^2) \|x_i - x_{i+1}\|^2. \end{aligned} \quad (4.27)$$

em que L_G é dado em (4.4) e o termo $\mathcal{B}_t(\cdot, \cdot, \cdot)$ é definido da seguinte forma: Para qualquer ponto v e qualquer sequência $v_{[t+1]}$ em qualquer espaço vetorial \mathcal{V} , e qualquer sequência de números reais $\gamma_{[t]}$, temos:

$$\mathcal{B}_t(v, v_{[t+1]}, \gamma_{[t]}) := \sum_{i=1}^t \frac{\alpha_i}{2\Gamma_i} \gamma_i (\|v_i - v\|^2 - \|v_{i+1} - v\|^2). \quad (4.28)$$

Dem: Para começar, provaremos uma propriedade importante da função $Q(\cdot, \cdot)$ no Algoritmo 4.2. Pela convexidade de $G(\cdot)$, temos:

$$G(x_{t+1}^{ag}) \leq G(x_t^{md}) + \langle \nabla G(x_t^{md}), x_{t+1}^{ag} - x_t^{md} \rangle + \frac{L_G}{2} \|x_{t+1}^{ag} - x_t^{md}\|^2. \quad (4.29)$$

Além disso, por meio das equações (4.15), (4.16), temos $x_{t+1}^{ag} - x_t^{md} = \alpha_t(x_{t+1} - x_t)$. Usando esta observação, a desigualdade (4.29) e a convexidade de $G(\cdot)$, temos:

$$G(x_{t+1}^{ag}) \leq G(x_t^{md}) + \langle \nabla G(x_t^{md}), x_{t+1}^{ag} - x_t^{md} \rangle + \frac{L_G}{2} \|\alpha_t(x_{t+1} - x_t)\|^2 \quad (4.30)$$

$$= G(x_t^{md}) + \langle \nabla G(x_t^{md}), x_{t+1}^{ag} - x_t^{md} \rangle + \frac{L_G}{2} \alpha_t^2 \|x_{t+1} - x_t\|^2 \quad (4.31)$$

$$= G(x_t^{md}) + \langle \nabla G(x_t^{md}), (1 - \alpha_t)x_t^{ag} + \alpha_t x_{t+1} - x_t^{md} \rangle + \frac{L_G}{2} \alpha_t^2 \|x_{t+1} - x_t\|^2 \quad (4.32)$$

$$= G(x_t^{md}) + \langle \nabla G(x_t^{md}), (1 - \alpha_t)x_t^{ag} + \alpha_t x_{t+1} - x_t^{md} + \alpha_t x_t^{md} - \alpha_t x_t^{md} \rangle + A \quad (4.33)$$

$$= G(x_t^{md}) + \langle \nabla G(x_t^{md}), (1 - \alpha_t)x_t^{ag} - (1 - \alpha_t)x_t^{md} + \alpha_t(x_{t+1} - x_t^{md}) \rangle + A \quad (4.34)$$

$$= G(x_t^{md}) + \alpha_t G(x_t^{md}) - \alpha_t G(x_t^{md}) + \langle \nabla G(x_t^{md}), (1 - \alpha_t)(x_t^{ag} - x_t^{md}) \rangle + \langle \nabla G(x_t^{md}), \alpha_t(x_{t+1} - x_t^{md}) \rangle + A \quad (4.35)$$

$$= G(x_t^{md}) + \alpha_t G(x_t^{md}) - \alpha_t G(x_t^{md}) + (1 - \alpha_t) \langle \nabla G(x_t^{md}), x_t^{ag} - x_t^{md} \rangle + \alpha_t \langle \nabla G(x_t^{md}), x_{t+1} - x_t^{md} \rangle + A \quad (4.36)$$

$$= G(x_t^{md}) + \alpha_t G(x_t^{md}) - \alpha_t G(x_t^{md}) + (1 - \alpha_t) \langle \nabla G(x_t^{md}), x_t^{ag} - x_t^{md} \rangle + \alpha_t \langle \nabla G(x_t^{md}), x_{t+1} - x + x - x_t^{md} \rangle + A \quad (4.37)$$

$$= (1 - \alpha_t) [G(x_t^{md}) + \langle \nabla G(x_t^{md}), x_t^{ag} - x_t^{md} \rangle] + \alpha_t [G(x_t^{md}) + \langle \nabla G(x_t^{md}), x - x_t^{md} \rangle] + \alpha_t \langle \nabla G(x_t^{md}), x_{t+1} - x \rangle + A \quad (4.38)$$

Usamos em (4.30) uma propriedade da norma $\|\cdot\|$ para obter (4.31), substituímos x_{t+1}^{ag} , de acordo com (4.17) para obter (4.32), somamos e subtraímos o termo $\alpha_t x_t^{md}$ além de chamar o termo $\frac{L_G}{2} \alpha_t^2 \|x_{t+1} - x_t\|^2$ de A para simplificar, obtendo (4.33), agrupando os termos dentro do produto interno, obtemos (4.34), somando e subtraindo o termo $\alpha_t G(x_t^{md})$, usando a linearidade do produto interno, resulta em (4.35), pela associatividade do produto interno, temos (4.36), somando e subtraindo o termo x , obtemos (4.37) e finalmente usando a linearidade do produto interno e agrupando os termos, resulta em (4.38).

Por (4.29), temos:

$$G(x_{t+1}^{ag}) \leq (1 - \alpha_t)G(x_t^{ag}) + \alpha_t G(x) + \alpha_t \langle \nabla G(x_t^{md}), x_{t+1} - x \rangle + A \quad (4.39)$$

Temos, então, por (4.22) :

$$\begin{aligned} Q(z, z_{t+1}^{ag}) - (1 - \alpha_t)Q(z, z_t^{ag}) &= [G(x_{t+1}^{ag}) + F(w_{t+1}^{ag}) - \langle y, Bw_{t+1}^{ag} - Kx_{t+1}^{ag} - b \rangle] \\ &\quad - [G(x) + F(w) - \langle y_t^{ag}, Bw - Kx - b \rangle] \\ &\quad - (1 - \alpha_t) [G(x_t^{ag}) + F(w_t^{ag}) - \langle y, Bw_t^{ag} - Kx_t^{ag} - b \rangle] \\ &\quad + (1 - \alpha_t) [G(x) + F(w) - \langle y_t^{ag}, Bw - Kx - b \rangle] \end{aligned}$$

Reorganizando os termos e simplificando o que for possível, temos:

$$\begin{aligned}
Q(z, z_{t+1}^{ag}) - (1 - \alpha_t)Q(z, z_t^{ag}) &= [G(x_{t+1}^{ag}) - (1 - \alpha_t)G(x_t^{ag}) - \alpha_t G(x)] \\
&\quad + [F(w_{t+1}^{ag}) - (1 - \alpha_t)F(w_t^{ag}) - \alpha_t F(x)] \\
&\quad - \langle y, Bw_{t+1}^{ag} - Kx_{t+1}^{ag} - b \rangle + (1 - \alpha_t)\langle y, Bw_t^{ag} - Kx_t^{ag} - b \rangle \\
&\quad + \langle y_{t+1}^{ag}, Bw - Kx - b \rangle - (1 - \alpha_t)\langle y_t^{ag}, Bw - Kx - b \rangle
\end{aligned}$$

Usando a linearidade e a associatividade do produto interno, temos:

$$\begin{aligned}
Q(z, z_{t+1}^{ag}) - (1 - \alpha_t)Q(z, z_t^{ag}) &= [G(x_{t+1}^{ag}) - (1 - \alpha_t)G(x_t^{ag}) - \alpha_t G(x)] \\
&\quad + [F(w_{t+1}^{ag}) - (1 - \alpha_t)F(w_t^{ag}) - \alpha_t F(x)] \\
&\quad + \langle y, B[-w_{t+1}^{ag} + (1 - \alpha_t)w_t^{ag}] - K[-x_{t+1}^{ag} + (1 - \alpha_t)x_t^{ag}] - b(-\alpha_t) \rangle \\
&\quad + \langle y_{t+1}^{ag} - (1 - \alpha_t)y_t^{ag}, Bw - Kx - b \rangle
\end{aligned}$$

Usando as igualdades em (4.17), (4.19) e (4.21), temos:

$$\begin{aligned}
Q(z, z_{t+1}^{ag}) - (1 - \alpha_t)Q(z, z_t^{ag}) &= [G(x_{t+1}^{ag}) - (1 - \alpha_t)G(x_t^{ag}) - \alpha_t G(x)] \\
&\quad + [F(w_{t+1}^{ag}) - (1 - \alpha_t)F(w_t^{ag}) - \alpha_t F(x)] \\
&\quad + \langle y, B(-\alpha_t w_{t+1}) - K(-\alpha_t x_{t+1}) - b(-\alpha_t) \rangle \\
&\quad + \langle \alpha_t y_{t+1}, Bw - Kx - b \rangle
\end{aligned}$$

Usando a associatividade do produto interno, temos:

$$\begin{aligned}
Q(z, z_{t+1}^{ag}) - (1 - \alpha_t)Q(z, z_t^{ag}) &= [G(x_{t+1}^{ag}) - (1 - \alpha_t)G(x_t^{ag}) - \alpha_t G(x)] \\
&\quad + [F(w_{t+1}^{ag}) - (1 - \alpha_t)F(w_t^{ag}) - \alpha_t F(x)] \\
&\quad - \alpha_t \langle y, Bw_{t+1} - Kx_{t+1} - b \rangle \\
&\quad + \alpha_t \langle y_{t+1}, Bw - Kx - b \rangle
\end{aligned} \tag{4.40}$$

De (4.39), e do fato de $F(\cdot)$ ser convexa, temos:

$$\begin{aligned}
G(x_{t+1}^{ag}) - (1 - \alpha_t)G(x_t^{ag}) - \alpha_t G(x) &\leq \alpha_t \langle \nabla G(x_t^{md}), x_{t+1} - x \rangle + A \\
F(\alpha_t w_{t+1} - \alpha_t w) = F(w_{t+1}^{ag} - (1 - \alpha_t)w_t^{ag} - \alpha_t w) &\geq F(w_{t+1}^{ag}) - (1 - \alpha_t)F(w_t^{ag}) - \alpha_t F(w)
\end{aligned}$$

daí usando em (4.40) e as duas desigualdades acima, segue a desigualdade:

$$\begin{aligned}
Q(z, z_{t+1}^{ag}) - (1 - \alpha_t)Q(z, z_t^{ag}) &\leq \alpha_t \left\{ \langle \nabla G(x_t^{md}), x_{t+1} - x \rangle + [F(w_{t+1}) - F(w)] + \frac{L_G}{2} \alpha_t^2 \|x_{t+1} - x_t\|^2 \right. \\
&\quad \left. - \langle y, Bw_{t+1} - Kx_{t+1} - b \rangle + \langle y_{t+1}, Bw - Kx - b \rangle \right\}.
\end{aligned} \tag{4.41}$$

Agora, vamos analisar as condições de otimalidade em (4.16) e (4.18), linearizando as equações, temos para todo $x \in X$ e $w \in \mathcal{W}$:

$$\langle \nabla G(x_t^{md}) + \eta_t(x_{t+1} - x_t), x_{t+1} - x \rangle - \langle \theta_t(Bw_t - K\hat{x}_t - b) - y_t, K(x_{t+1} - x) \rangle \leq 0$$

e

$$F(w_{t+1}) - F(w) + \langle \tau_t(Bw_{t+1} - Kx_{t+1} - b) - y_t, B(w_{t+1} - w) \rangle \leq 0$$

em que $\hat{x}_t := \chi x_t + (1 - \chi)x_{t+1}$.

Observando a partir de (4.20), concluímos:

$$Bw_{t+1} - Kx_{t+1} - b = \frac{(y_t - y_{t+1})}{\rho_t}.$$

e daí resulta:

$$Bw_t - K\hat{x}_t - b = \frac{(y_t - y_{t+1})}{\rho_t} - K(\hat{x}_t - x_{t+1} + B(w_t - w_{t+1})).$$

Assim, as condições de otimalidade tornam-se:

$$\begin{aligned} & \langle \nabla G(x_t^{md}) + \eta_t(x_{t+1} - x_t), x_{t+1} - x \rangle \\ & - \theta_t \left\langle \left(\frac{(y_t - y_{t+1})}{\rho_t} \right) - K(\hat{x}_t - x_{t+1}) + B(w_t - w_{t+1}) - y_t, K(x_{t+1} - x) \right\rangle \leq 0 \end{aligned}$$

e

$$F(w_{t+1}) - F(w) + \left\langle \tau_t \left[\left(\frac{(y_t - y_{t+1})}{\rho_t} \right) - y_t \right], B(w_{t+1} - w) \right\rangle \leq 0$$

Usando a associatividade e a linearidade do produto interno, além de manipulações algébricas, obtemos:

$$\begin{aligned} & \langle \nabla G(x_t^{md}) + \eta_t(x_{t+1} - x_t), x_{t+1} - x \rangle + \left\langle \left(\frac{\theta_t}{\rho_t} - 1 \right) (y_t - y_{t+1}) - y_{t+1}, -K(x_{t+1} - x) \right\rangle \\ & + \theta_t \langle K(\hat{x}_t - x_{t+1}), K(x_{t+1} - x) \rangle + \theta_t \langle B(w_t - w_{t+1}), K(x_{t+1} - x) \rangle \leq 0 \end{aligned}$$

e

$$F(w_{t+1}) - F(w) + \left\langle \left(\frac{\tau_t}{\rho_t} - 1 \right) (y_t - y_{t+1}) - y_{t+1}, B(w_{t+1} - w) \right\rangle \leq 0$$

Somando as duas condições de otimalidade acima e usando novamente a associatividade e a

linearidade do produto interno, resulta em:

$$\begin{aligned} & \langle \nabla G(x_t^{md}), x_{t+1} - x \rangle + \langle \eta_t(x_{t+1} - x_t), x_{t+1} - x \rangle + \left\langle \left(\frac{\theta_t}{\rho_t} - 1 \right) (y_t - y_{t+1}), -K(x_{t+1} - x) \right\rangle \\ & + \langle -y_{t+1}, -K(x_{t+1} - x) \rangle + \theta_t \langle K(\hat{x}_t - x_{t+1}), K(x_{t+1} - x) \rangle + \theta_t \langle B(w_t - w_{t+1}), -K(x_{t+1} - x) \rangle \\ & + F(w_{t+1}) - F(w) + \left\langle \left(\frac{\tau_t}{\rho_t} - 1 \right) (y_t - y_{t+1}), B(w_{t+1} - w) \right\rangle + \langle -y_{t+1}, B(w_{t+1} - w) \rangle \leq 0 \end{aligned}$$

Para simplificar somaremos $\langle -y_{t+1}, -K(x_{t+1} - x) \rangle$ com $\langle -y_{t+1}, B(w_{t+1} - w) \rangle$. Usando a associatividade e a linearidade do produto interno, temos:

$$\begin{aligned} \langle -y_{t+1}, B(w_{t+1} - w) - K(x_{t+1} - x) \rangle &= \langle -y_{t+1}, Bw_{t+1} - Bw - Kx_{t+1} + Kx + b - b \rangle \\ &= \langle -y_{t+1}, Bw_{t+1} - Kx_{t+1} - b - (Bw - Kx - b) \rangle \\ &= \langle -y_{t+1}, Bw_{t+1} - Kx_{t+1} - b - (Bw - Kx - b) \rangle \\ &= \langle -y_{t+1}, Bw_{t+1} - Kx_{t+1} - b \rangle - \langle -y_{t+1}, Bw - Kx - b \rangle \\ &= \langle -y_{t+1} + y - y, Bw_{t+1} - Kx_{t+1} - b \rangle - \langle -y_{t+1}, Bw - Kx - b \rangle \\ &= \langle -y, Bw_{t+1} - Kx_{t+1} - b \rangle + \langle -y_{t+1} + y, Bw_{t+1} - Kx_{t+1} - b \rangle \\ &\quad - \langle -y_{t+1}, Bw - Kx - b \rangle \end{aligned}$$

Agora, reorganizando os termos e usando a associatividade do produto interno, finalmente, obtemos:

$$\begin{aligned} & \langle \nabla G(x_t^{md}), x_{t+1} - x \rangle + F(w_{t+1}) - F(w) - \langle y, Bw_{t+1} - Kx_{t+1} - b \rangle + \langle y_{t+1}, Bw - Kx - b \rangle \\ & \leq \langle \eta_t(x_t - x_{t+1}), x_{t+1} - x \rangle + \langle y_{t+1} - y, Bw_{t+1} - Kx_{t+1} - b \rangle \\ & \quad - \left\langle \left(\frac{\theta_t}{\rho_t} - 1 \right) (y_t - y_{t+1}), -K(x_{t+1} - x) \right\rangle - \left\langle \left(\frac{\tau_t}{\rho_t} - 1 \right) (y_t - y_{t+1}), B(w_{t+1} - w) \right\rangle \\ & \quad + \theta_t \langle K(x_{t+1} - \hat{x}_t), K(x_{t+1} - x) \rangle + \theta_t \langle B(w_{t+1} - w_t), -K(x_{t+1} - x) \rangle \end{aligned} \quad (4.42)$$

Faremos três observações sobre o lado direito de (4.42). Em primeiro lugar por (4.20), vale a seguinte igualdade:

$$\begin{aligned} -\langle \eta_t(x_t - x_{t+1}), x_{t+1} - x \rangle + \langle y_{t+1} - y, Bw_{t+1} - Kx_{t+1} - b \rangle &= -\langle \eta_t(x_t - x_{t+1}), x_{t+1} - x \rangle \\ &\quad + \left\langle y_{t+1} - y, \frac{(y_t - y_{t+1})}{\rho_t} \right\rangle \end{aligned}$$

Aplicando o Lema 1.1, temos:

$$\begin{aligned} -\langle \eta_t(x_t - x_{t+1}), x_{t+1} - x \rangle + \left\langle y_{t+1} - y, \frac{(y_t - y_{t+1})}{\rho_t} \right\rangle &= \frac{\eta_t}{2} (\|x_t - x\|^2 - \|x_{t+1} - x\|^2) - \frac{\eta_t}{2} \|x_t - x_{t+1}\|^2 \\ &\quad + \frac{1}{2\rho_t} (\|y_t - y\|^2 - \|y_{t+1} - y\|^2 - \|y_t - y_{t+1}\|^2) \end{aligned} \quad (4.43)$$

Em segundo lugar, também por (4.20), vale a seguinte igualdade:

$$B(w_{t+1} - w) = \frac{1}{\rho_t} (y_t - y_{t+1} + (Kx_{t+1} - Kx) - (Bw - Kx - b)) \quad (4.44)$$

e usando propriedades de produto interno temos:

$$\begin{aligned} & - \left\langle \left(\frac{\theta_t}{\rho_t} - 1 \right) (y_t - y_{t+1}), -K(x_{t+1} - x) \right\rangle - \left\langle \left(\frac{\tau_t}{\rho_t} - 1 \right) (y_t - y_{t+1}), \frac{1}{\rho_t} (y_t - y_{t+1}) + (Kx_{t+1} - Kx) \right\rangle \\ & = \left(\frac{\theta_t}{\rho_t} - 1 \right) \langle y_t - y_{t+1}, K(x_{t+1} - x) \rangle \\ & \quad - \left(\frac{\tau_t}{\rho_t} - 1 \right) \left[\left\langle (y_t - y_{t+1}), \frac{(y_t - y_{t+1})}{\rho_t} \right\rangle + \langle (y_t - y_{t+1}), (Kx_{t+1} - Kx) \rangle \right] \\ & = \left(\frac{\theta_t}{\rho_t} - 1 - \left(\frac{\tau_t}{\rho_t} - 1 \right) \right) \langle y_t - y_{t+1}, K(x_{t+1} - x) \rangle - \left(\frac{\tau_t}{\rho_t^2} - \frac{1}{\rho_t} \right) \langle (y_t - y_{t+1}), (y_t - y_{t+1}) \rangle \\ & = \left(\frac{\theta_t - \tau_t}{\rho_t} \right) \langle y_t - y_{t+1}, K(x_{t+1} - x) \rangle - \left(\frac{\tau_t - \rho_t}{\rho_t^2} \right) \|y_t - y_{t+1}\|^2 \\ & = \left(\frac{\tau_t - \theta_t}{2} \right) \left[\frac{1}{\rho_t^2} \|y_t - y_{t+1}\|^2 - \|K(x_{t+1} - x)\|^2 - \left\| \frac{1}{\rho_t} (y_t - y_{t+1}) + (K(x_{t+1} - x)) \right\|^2 \right] \\ & \quad - \left(\frac{\tau_t - \rho_t}{\rho_t^2} \right) \|y_t - y_{t+1}\|^2 \end{aligned} \quad (4.45)$$

$$\begin{aligned} & = \left(\frac{\tau_t - \theta_t}{2} \right) \left[\frac{1}{\rho_t^2} \|y_t - y_{t+1}\|^2 - \|K(x_{t+1} - x)\|^2 - \|Bw_{t+1} - Kx - b\|^2 \right] \\ & \quad - \left(\frac{\tau_t - \rho_t}{\rho_t^2} \right) \|y_t - y_{t+1}\|^2. \end{aligned} \quad (4.46)$$

Pelo Lema (1.1), obtemos (4.45) e por (4.44), resulta em (4.46).

E em terceiro lugar, temos:

$$\theta_t \langle K(x_{t+1} - \hat{x}_t), K(x_{t+1} - x) \rangle + \theta_t \langle B(w_{t+1} - w_t), -K(x_{t+1} - x) \rangle = A_1$$

Substituindo \hat{x}_t , obtemos:

$$A_1 = \theta_t \langle K(x_{t+1} - \chi x_t - x_{t+1} + \chi x_{t+1}), K(x_{t+1} - x) \rangle + \theta_t \langle B(w_{t+1} - w_t), -K(x_{t+1} - x) \rangle$$

Simplificando, colocando χ em evidência e usando a associatividade do produto interno, resulta em:

$$A_1 = -\chi \theta_t \langle K(x_{t+1} - x_t), K(x_{t+1} - x) \rangle + \theta_t \langle B(w_{t+1} - w_t), -K(x_{t+1} - x) \rangle$$

E pelo Lema 1.1:

$$\begin{aligned} A_1 & = -\frac{\chi \theta_t}{2} (\|K(x_t - x)\|^2 - \|K(x_{t+1} - x)\|^2 - \|K(x_t - x_{t+1})\|^2) \\ & \quad + \frac{\theta_t}{2} (\|Bw_t - Kx - b\|^2 - \|Bw_{t+1} - Kx - b\|^2 + \|Bw_{t+1} - Kx_{t+1} - b\|^2 - \|Bw_t - Kx_{t+1} - b\|^2) \\ & \leq -\frac{\chi \theta_t}{2} (\|K(x_t - x)\|^2 - \|K(x_{t+1} - x)\|^2) + \frac{\chi \theta_t \|K\|^2}{2} \|x_t - x_{t+1}\|^2 \\ & \quad + \frac{\theta_t}{2} (\|Bw_t - Kx - b\|^2 - \|Bw_{t+1} - Kx - b\|^2) + \frac{\theta_t}{2\rho^2} \|y_t - y_{t+1}\|^2 - \frac{\theta_t}{2} \|Bw_t - Kx_{t+1} - b\|^2 \end{aligned} \quad (4.47)$$

Em que a última desigualdade resulta do fato de:

$$\chi \|K(x_t - x_{t+1})\| \leq \chi \|K\| \|x_t - x_{t+1}\| \quad (4.48)$$

Aplicando (4.42), (4.43), (4.44), (4.46) e (4.47) em (4.41), temos:

$$\begin{aligned} & \frac{1}{\Gamma_t} Q(z, z_{t+1}^{ag}) - \frac{(1 - \alpha_t)}{\Gamma_t} Q(z, z_t^{ag}) \\ \leq & \frac{\alpha_t}{\Gamma_t} \left\{ \frac{\eta_t}{2} (\|x_t - x\|^2 - \|x_{t+1} - x\|^2) + \frac{1}{2\rho_t} (\|y_t - y\|^2 - \|y_{t+1} - y\|^2 - \frac{\tau_t - \rho_t}{2\rho_t^2} \|y_t - y_{t+1}\|^2 \right. \\ & + \frac{\theta_t}{2} \|Bw_t - Kx - b\|^2 - + \frac{\tau_t}{2} \|Bw_t - Kx - b\|^2 - \frac{\chi\theta_t}{2} (\|K(x_t - x)\|^2 - \|K(x_{t+1} - x)\|^2) \\ & + \left\langle \left(\frac{\tau_t}{\rho_t} - 1 \right) (y_t - y_{t+1}), Bw - Kx - b \right\rangle + \frac{(\tau_t - \theta_t)}{2} \|K(x_{t+1} - x)\|^2 - \frac{\theta_t}{2} \|Bw_t - Kx_{t+1} - b\|^2 \\ & \left. - \frac{1}{2} (\eta_t - L_G \alpha_t - \chi \theta_t \|K\|^2) \|x_t - x_{t+1}\|^2 \right\} \end{aligned} \quad (4.49)$$

Tomando $w = w^*$ e $x = x^*$ na equação anterior, observando a partir de (4.26) que $\Gamma_{t-1} = \frac{1-\alpha_t}{\Gamma_t}$ e aplicando a desigualdade acima, indutivamente concluímos (4.27). ■

Observamos que há duas grandes consequências do Lema 4.1.

1) Se $\alpha_t \equiv 1$ para todo t , então, o lado esquerdo de (4.27) torna-se $\frac{1}{\Gamma_1} \sum_{i=2}^{t+1} Q(z; z_i^{ag})$.

De fato, por (4.27), obtemos:

$$\frac{1}{\Gamma_t} Q(z; z_{t+1}^{ag}) - \sum_{i=2}^t \left(\frac{1 - \alpha_i}{\Gamma_i} - \frac{1}{\Gamma_{i-1}} \right) Q(z; z_i^{ag})$$

tomando $\alpha_t \equiv 1$ temos $\Gamma_t = \Gamma_1$ para todo t , então:

$$\frac{1}{\Gamma_1} Q(z; z_{t+1}^{ag}) + \sum_{i=2}^t \frac{1}{\Gamma_1} Q(z; z_i^{ag}) = \frac{1}{\Gamma_1} \left(Q(z; z_{t+1}^{ag}) + \sum_{i=2}^t Q(z; z_i^{ag}) \right) = \frac{1}{\Gamma_1} \sum_{i=2}^{t+1} Q(z; z_i^{ag}).$$

2) Por outro lado, se $\alpha_t \in [0, 1)$ o lado esquerdo de (4.27) torna-se $\frac{Q(z; z_{t+1}^{ag})}{\Gamma_t}$.

De fato, por (4.27), obtemos:

$$\frac{1}{\Gamma_t} Q(z; z_{t+1}^{ag}) - \sum_{i=2}^t \left(\frac{1 - \alpha_i}{\Gamma_i} - \frac{1}{\Gamma_{i-1}} \right) Q(z; z_i^{ag})$$

Tomando $\alpha_t \in [0, 1)$, temos por (4.26), $\Gamma_t = (1 - \alpha_t) \Gamma_{t-1}$ para todo t , então, note que $\sum_{i=2}^t \left(\frac{1 - \alpha_i}{\Gamma_i} - \frac{1}{\Gamma_{i-1}} \right) Q(z; z_i^{ag}) = 0$. De fato, basta substituir Γ_i e segue:

$$\sum_{i=2}^t \left(\frac{1 - \alpha_i}{(1 - \alpha_t) \Gamma_{i-1}} - \frac{1}{\Gamma_{i-1}} \right) Q(z; z_i^{ag}) = \sum_{i=2}^t \left(\frac{1}{\Gamma_{i-1}} - \frac{1}{\Gamma_{i-1}} \right) Q(z; z_i^{ag}) = 0.$$

Atribui-se a esta diferença a principal razão pela qual se pode acelerar a taxa de convergência de AADMM em termos de L_G .

O lema que apresentamos a seguir fornece possíveis limitações de $\mathcal{B}_t(\cdot, \cdot, \cdot)$ no Lema 4.1.

Lema 4.2 *Suponha que \mathcal{V} é um espaço vetorial qualquer e $V \subset \mathcal{V}$ é um conjunto convexo qualquer. Para cada $v \in V$, $v_{[t+1]} \subset \mathcal{V}$ e $\gamma_{[t]} \subset \mathbb{R}$, temos:*

(a) *Se a sequência $\left\{\frac{\alpha_i}{\Gamma_i}\gamma_i\right\}$ é decrescente, então:*

$$\mathcal{B}_t(v, v_{[t+1]}, \gamma_{[t]}) \leq \frac{\alpha_1}{2\Gamma_1}\gamma_1\|v_1 - v\|^2 - \frac{\alpha_t}{2\Gamma_t}\gamma_t\|v_{t+1} - v\|^2 \quad (4.50)$$

(b) *Se a sequência $\left\{\frac{\alpha_i}{\Gamma_i}\gamma_i\right\}$ é crescente, V é limitado e $v_{[t+1]} \subset V$, então:*

$$\mathcal{B}_t(v, v_{[t+1]}, \gamma_{[t]}) \leq \frac{\alpha_t}{2\Gamma_t}\gamma_tD_V^2 - \frac{\alpha_t}{2\Gamma_t}\gamma_t\|v_{t+1} - v\|^2 \quad (4.51)$$

Dem: Temos por (4.28):

$$\mathcal{B}_t(v, v_{[t+1]}, \gamma_{[t]}) := \sum_{i=1}^t \frac{\alpha_i}{2\Gamma_i}\gamma_i (\|v_i - v\|^2 - \|v_{i+1} - v\|^2).$$

Desenvolvendo o somatório, temos:

$$= \frac{\alpha_1}{2\Gamma_1}\gamma_1 (\|v_1 - v\|^2 - \|v_2 - v\|^2) + \frac{\alpha_2}{2\Gamma_2}\gamma_2 (\|v_2 - v\|^2 - \|v_3 - v\|^2) + \dots + \frac{\alpha_t}{2\Gamma_t}\gamma_t (\|v_t - v\|^2 - \|v_{t+1} - v\|^2)$$

Agrupando os termos, obtemos:

$$= \frac{\alpha_1}{2\Gamma_1}\gamma_1\|v_1 - v\|^2 + \left(\frac{\alpha_2}{2\Gamma_2}\gamma_2 - \frac{\alpha_1}{2\Gamma_1}\gamma_1\right)\|v_2 - v\|^2 + \dots + \left(\frac{\alpha_t}{2\Gamma_t}\gamma_t - \frac{\alpha_{t-1}}{2\Gamma_{t-1}}\gamma_{t-1}\right)\|v_t - v\|^2 - \frac{\alpha_t}{2\Gamma_t}\gamma_t\|v_{t+1} - v\|^2$$

Podemos reescrever a equação acima da seguinte maneira:

$$= \frac{\alpha_1}{2\Gamma_1}\gamma_1\|v_1 - v\|^2 - \sum_{i=1}^{t-1} \left(\frac{\alpha_i}{2\Gamma_i}\gamma_i - \frac{\alpha_{i+1}}{2\Gamma_{i+1}}\gamma_{i+1}\right)\|v_{i+1} - v\|^2 - \frac{\alpha_t}{2\Gamma_t}\gamma_t\|v_{t+1} - v\|^2$$

$\left\{\frac{\alpha_i}{\Gamma_i}\gamma_i\right\}$ é decrescente, então (4.50) é válida, pois:

$$- \sum_{i=1}^{t-1} \left(\frac{\alpha_i}{2\Gamma_i}\gamma_i - \frac{\alpha_{i+1}}{2\Gamma_{i+1}}\gamma_{i+1}\right)\|v_{i+1} - v\|^2 \longrightarrow a, \quad a \leq 0 \quad \text{ou diverge para } -\infty.$$

Se a sequência $\left\{\frac{\alpha_i}{\Gamma_i}\gamma_i\right\}$ é crescente, V é limitado e $v_{[t+1]} \subset V$. A partir da equação acima, temos:

$$\mathcal{B}_t(v, v_{[t+1]}, \gamma_{[t]}) = \frac{\alpha_1}{2\Gamma_1}\gamma_1\|v_1 - v\|^2 - \sum_{i=1}^{t-1} \left(\frac{\alpha_i}{2\Gamma_i}\gamma_i - \frac{\alpha_{i+1}}{2\Gamma_{i+1}}\gamma_{i+1}\right)\|v_{i+1} - v\|^2 - \frac{\alpha_t}{2\Gamma_t}\gamma_t\|v_{t+1} - v\|^2$$

Por (4.7), obtemos:

$$\mathcal{B}_t(v, v_{[t+1]}, \gamma_{[t]}) \leq \frac{\alpha_1}{2\Gamma_1} \gamma_1 D_V^2 - \sum_{i=1}^{t-1} \left(\frac{\alpha_i}{2\Gamma_i} \gamma_i - \frac{\alpha_{i+1}}{2\Gamma_{i+1}} \gamma_{i+1} \right) D_V^2 - \frac{\alpha_t}{2\Gamma_t} \gamma_t \|v_{t+1} - v\|^2$$

Desenvolvendo o somatório, resulta em:

$$\mathcal{B}_t(v, v_{[t+1]}, \gamma_{[t]}) \leq \frac{\alpha_t}{2\Gamma_t} \gamma_t D_V^2 - \frac{\alpha_t}{2\Gamma_t} \gamma_t \|v_{t+1} - v\|^2$$

e, portanto, (4.51) é válida. ■

4.4 Considerações sobre os resultados de convergência na resolução de problemas UCO em domínios limitados

Veremos, agora, problemas UCO com conjuntos viáveis limitados e, para isso, tomemos que X e $Y := \text{dom} F^*$ são compactos, B é o operador identidade ($B = I$) e $b = 0$. A condição do conjunto viável Y ser limitado é equivalente a $F(\cdot)$ ser Lipschitz contínua (Teorema 1.1).

O Teorema 4.1 generaliza as propriedades de convergência de algoritmos ADMM e possibilita uma visão unificada das propriedades de convergência de todos os algoritmos ADMM.

Teorema 4.1 *Em AADMM se os parâmetros são definidos da seguinte maneira $\alpha_t \equiv 1$, $\theta_t \equiv \tau_t \equiv \rho_t \equiv \rho$ e $\eta_t \equiv L_G + \chi\rho\|K\|^2$, então:*

$$G(x^{t+1}) + F(Kx^{t+1}) - f^* \leq \frac{L_G}{2t} D_X^2 + \frac{\chi\rho}{2t} \|K\|^2 D_X^2 + \frac{(1-\chi)\rho}{2t} D_{X,K}^2 + \frac{D_Y^2}{2\rho t}, \quad (4.52)$$

em que $x^{t+1} := \frac{1}{t} \sum_{i=2}^{t+1} x_i$. E, se ρ é dado por:

$$\rho = \frac{D_Y}{\chi\|K\|D_X + (1-\chi)D_{X,K}}, \quad (4.53)$$

temos:

$$G(x^{t+1}) + F(\hat{w}^{t+1}) - f^* \leq \frac{L_G}{2t} D_X^2 + \frac{\chi\|K\|D_X D_Y + (1-\chi)D_{X,K} D_Y}{t}. \quad (4.54)$$

Dem: Se $\alpha_t \equiv 1$, por (4.17), (4.19) e (4.21), temos $x_t^{ag} = x_t$, $w_t^{ag} = w_t$ e $y_t^{ag} = y_t$, respectivamente, e concluímos que $\Gamma_t \equiv 1$ satisfaz (4.26). Agora aplicando as definições dos parâmetros para o lado direito de (4.27), temos:

$$(I) : \mathcal{B}_t(x^*, x_{[t+1]}, \eta_{[t]}) \stackrel{(4.28)}{=} \sum_{i=1}^t \frac{\alpha_i}{2\Gamma_i} \eta_i (\|x_i - x^*\|^2 - \|x_{i+1} - x^*\|^2).$$

Substituindo $\alpha_t \equiv 1$, $\Gamma_t \equiv 1$ e desenvolvendo o somatório, obtemos:

$$= \frac{\eta}{2} \|x_1 - x^*\|^2 - \frac{\eta}{2} \|x_2 - x^*\|^2 + \frac{\eta}{2} \|x_2 - x^*\|^2 - \frac{\eta}{2} \|x_3 - x^*\|^2 + \cdots + \frac{\eta}{2} \|x_t - x^*\|^2 - \frac{\eta}{2} \|x_{t+1} - x^*\|^2.$$

Simplificando, temos:

$$\mathcal{B}_t(x^*, x_{[t+1]}, \eta_{[t]}) = \frac{\eta}{2} (\|x_1 - x^*\|^2 - \|x_{t+1} - x^*\|^2).$$

Por (4.7), resulta em:

$$\mathcal{B}_t(x^*, x_{[t+1]}, \eta_{[t]}) \leq \frac{\eta}{2} (D_{x^*}^2 - \|x_{t+1} - x^*\|^2).$$

Substituindo η , obtemos:

$$\begin{aligned} \mathcal{B}_t(x^*, x_{[t+1]}, \eta_{[t]}) &= \frac{L_G + \chi\rho\|K\|^2}{2} (D_{x^*}^2 - \|x_{t+1} - x^*\|^2) \\ &= \frac{L_G}{2} D_{x^*}^2 + \frac{\chi\rho\|K\|^2}{2} D_{x^*}^2 - \frac{L_G}{2} \|x_{t+1} - x^*\|^2 - \frac{\chi\rho\|K\|^2}{2} \|x_{t+1} - x^*\|^2, \end{aligned}$$

e como $-\frac{L_G}{2} \|x_{t+1} - x^*\|^2 \leq 0$, concluímos que:

$$\mathcal{B}_t(x^*, x_{[t+1]}, \eta_{[t]}) \leq \frac{L_G}{2} D_{x^*}^2 + \frac{\chi\rho\|K\|^2}{2} D_{x^*}^2 - \frac{\chi\rho\|K\|^2}{2} \|x_{t+1} - x^*\|^2.$$

$$(II) : \mathcal{B}_t(w^*, w_{[t+1]}, \theta_{[t]}) \stackrel{(4.28)}{=} \sum_{i=1}^t \frac{\alpha_i}{2\Gamma_i} \theta_i (\|w_i - w^*\|^2 - \|w_{i+1} - w^*\|^2).$$

Substituindo $\theta_{[t]}$ por ρ e de maneira totalmente análoga ao que foi feito em I, obtemos:

$$\mathcal{B}_t(w^*, w_{[t+1]}, \theta_{[t]}) = \frac{\rho}{2} (\|w_1 - w^*\|^2 - \|w_{t+1} - w^*\|^2).$$

Por (4.7), temos:

$$\mathcal{B}_t(w^*, w_{[t+1]}, \theta_{[t]}) \leq \frac{\rho}{2} (D_{w^*}^2 - \|w_{t+1} - w^*\|^2).$$

Como $-\frac{\rho}{2} \|w_{t+1} - w^*\|^2 \leq 0$, concluímos que:

$$\mathcal{B}_t(w^*, w_{[t+1]}, \theta_{[t]}) \leq \frac{\rho}{2} D_{w^*}^2 = \frac{\rho}{2} D_{x^*, K}^2.^1$$

¹Pela hipótese do Algoritmo 4.2, temos $Bw_1 = Kw_1 - b$ e, como considerado na introdução desta seção, sabemos que $B = I$ e $b = 0$.

$$(III) : -\chi \mathcal{B}_t(Kx^*, Kx_{[t+1]}, \theta_{[t]}) \stackrel{(4.28)}{=} \sum_{i=1}^t \frac{\alpha_i}{2\Gamma_i} \theta_i (\|Kx_i - Kx^*\|^2 - \|Kx_{i+1} - Kx^*\|^2).$$

Substituindo $\theta_{[t]}$ por ρ e de maneira totalmente análoga ao que foi feito em I, obtemos:

$$-\chi \mathcal{B}_t(Kx^*, Kx_{[t+1]}, \theta_{[t]}) = -\frac{\chi\rho}{2} (\|Kx_1 - Kx^*\|^2 - \|Kx_{t+1} - Kx^*\|^2).$$

Por (4.7), temos:

$$\begin{aligned} -\chi \mathcal{B}_t(Kx^*, Kx_{[t+1]}, \theta_{[t]}) &\leq -\frac{\chi\rho}{2} (D_{x^*, K}^2 - \|Kx_{t+1} - Kx^*\|^2) \\ &= -\frac{\chi\rho}{2} D_{x^*, K}^2 + \frac{\chi\rho}{2} \|Kx_{t+1} - Kx^*\|^2 \\ &= -\frac{\chi\rho}{2} D_{x^*, K}^2 + \frac{\chi\rho}{2} \|K\|^2 \|x_{t+1} - x^*\|^2. \end{aligned}$$

e

$$(IV) : \mathcal{B}_t(y, y_{[t+1]}, \rho_{[t]}^{-1}) \stackrel{(4.28)}{=} \sum_{i=1}^t \frac{\alpha_i}{2\Gamma_i} \rho_i (\|y_i - y\|^2 - \|y_{i+1} - y\|^2).$$

Substituindo $\rho_{[t]}$ por ρ e de maneira análoga ao que foi feito em I, obtemos:

$$\mathcal{B}_t(y, y_{[t+1]}, \rho_{[t]}^{-1}) = \frac{1}{2\rho} (\|y_1 - y\|^2 - \|y_{t+1} - y\|^2).$$

Por (4.7), temos:

$$\mathcal{B}_t(y, y_{[t+1]}, \rho_{[t]}^{-1}) \leq \frac{1}{2\rho} (D_y^2 - \|Kx_{t+1} - Kx^*\|^2) = \frac{1}{2\rho} D_y^2 - \frac{1}{2\rho} \|y_{t+1} - y\|^2.$$

Como $-\frac{1}{2\rho} \|y_{t+1} - y\|^2 \leq 0$, resulta em:

$$\mathcal{B}_t(y, y_{[t+1]}, \rho_{[t]}^{-1}) \leq \frac{1}{2\rho} D_y^2, \quad \forall y \in Y.$$

Note que para todo $y \in Y$, pela convexidade de $Q(w^*, x^*, y; \cdot)$, sabemos que:

$$Q(w^*, x^*, y; z^{t+1}) = Q(w^*, x^*, y; \frac{1}{t} \sum_{i=2}^{t+1} z_i) \leq \frac{1}{t} \sum_{i=2}^{t+1} Q(w^*, x^*, y; z_i). \quad (4.55)$$

Temos também que:

$$\frac{1}{\Gamma_t} Q(w^*, x^*, y^*; z_{t+1}^{ag}) - \sum_{i=2}^t \left(\frac{1 - \alpha_i}{\Gamma_i} - \frac{1}{\Gamma_{i-1}} \right) Q(w^*, x^*, y^*; z_i^{ag}).$$

Substituindo $\Gamma_t \equiv 1$ e $\alpha_t \equiv 1$, obtemos:

$$\begin{aligned}
\frac{1}{\Gamma_t} Q(w^*, x^*, y^*; z_{t+1}^{ag}) - \sum_{i=2}^t \left(\frac{1 - \alpha_i}{\Gamma_i} - \frac{1}{\Gamma_{i-1}} \right) Q(w^*, x^*, y^*; z_i^{ag}) &= Q(w^*, x^*, y^*; z_{t+1}) + \sum_{i=2}^t Q(w^*, x^*, y^*; z_i^{ag}) \\
&= \sum_{i=2}^{t+1} Q(w^*, x^*, y^*; z_i^{ag}),
\end{aligned}$$

daí pelo Lema 4.1 da Seção 4.3 temos:

$$\sum_{i=2}^{t+1} Q(w^*, x^*, y^*; z_i^{ag}) \leq \mathcal{B}_t(x^*, x_{[t+1]}, \eta_{[t]}) + \mathcal{B}_t(y, y_{[t+1]}, \rho_{[t]}^{-1}) + \mathcal{B}_t(Bw^*, Bw_{[t+1]}, \theta_{[t]}) - \chi \mathcal{B}_t(Kx^*, Kx_{[t+1]}, \theta_{[t]}).$$

Os demais termos que ficam à direita de (4.27) tornam-se nulos pelas definições dos parâmetros. Substituindo I, II, III e IV acima, temos:

$$\sum_{i=2}^{t+1} Q(w^*, x^*, y^*; z_i^{ag}) \leq \frac{L_G}{2} D_{x^*}^2 + \frac{\chi \rho}{2} \|K\|^2 D_{x^*}^2 + \frac{(1 - \chi) \rho}{2} D_{x^*, K}^2 + \frac{D_Y^2}{2\rho}. \quad (4.56)$$

Temos ainda:

$$\sup_{\hat{y} \in Y} Q(w^*, x^*, \hat{y}; z) \stackrel{(4.25)}{=} \hat{g}_y(z) \leq^{Prop. 4.2} G(x^*) + F(\hat{w}) - f^*.$$

Assim, aplicando (4.55) à desigualdade acima, obtemos:

$$G(x^{t+1}) + F(\hat{w}^{t+1}) - f^* \leq Q(\hat{w}^{t+1}, x^{t+1}, y; z_{t+1}) \stackrel{(4.55)}{\leq} \frac{1}{t} \sum_{i=2}^{t+1} Q(w^*, x^*, y; z_i).$$

E, por (4.56), concluímos que:

$$G(x^{t+1}) + F(\hat{w}^{t+1}) - f^* \leq \frac{1}{t} \left(\frac{L_G}{2} D_{x^*}^2 + \frac{\chi \rho}{2} \|K\|^2 D_{x^*}^2 + \frac{(1 - \chi) \rho}{2} D_{x^*, K}^2 + \frac{D_Y^2}{2\rho} \right).$$

e (4.52) fica provado.

Agora tomando ρ de acordo com (4.53) e substituindo em (4.52), temos:

$$\begin{aligned}
G(x^{t+1}) + F(\hat{w}^{t+1}) - f^* &\leq \frac{L_G}{2t} D_X^2 + \frac{\chi}{2t} \frac{D_Y}{\chi \|K\| D_X + (1 - \chi) D_{X,K}} \|K\|^2 D_X^2 \\
&\quad + \frac{(1 - \chi)}{2t} \frac{D_Y}{\chi \|K\| D_X + (1 - \chi) D_{X,K}} D_{X,K}^2 \\
&\quad + \frac{D_Y^2}{2t} \frac{\chi \|K\| D_X + (1 - \chi) D_{X,K}}{D_Y} \\
&= \frac{L_G}{2t} D_X^2 + \frac{\chi}{2t} \frac{D_Y \|K\|^2 D_X^2}{\|K\| D_X} + \frac{(1 - \chi)}{2t} \frac{D_Y D_{X,K}^2}{D_{X,K}} \\
&\quad + \frac{D_Y}{2t} \chi \|K\| D_X + \frac{D_Y}{2t} (1 - \chi) D_{X,K},
\end{aligned}$$

pois $\chi \neq 0$ se $(1 - \chi) = 0$. Simplificando as expressões, obtemos:

$$G(x^{t+1}) + F(\hat{w}^{t+1}) - f^* \leq \frac{L_G}{2t} D_X^2 + \frac{\chi}{2t} D_Y \|K\| D_X + \frac{(1 - \chi)}{2t} D_Y D_{X,K} + \frac{D_Y}{2t} \chi \|K\| D_X + \frac{D_Y}{2t} (1 - \chi) D_{X,K}.$$

Agrupando os termos resulta em:

$$G(x^{t+1}) + F(\hat{w}^{t+1}) - f^* \leq \frac{L_G D_X^2}{2t} + \frac{\chi \|K\| D_X D_Y + (1 - \chi) D_{X,K} D_Y}{t}.$$

Provando, assim, a desigualdade (4.54). ■

De acordo com Ouyang et al. [26] embora AADMM unifique todos os algoritmos ADMM, o que o diferencia é a ponderação da sequência $\{\alpha_t\}_{t \geq 1}$ (em vez de $\alpha_t \equiv 1$) que é o que acelera a velocidade de convergência no que diz respeito à sua dependência de L_G . Isso pode ser verificado no Teorema 4.2

Teorema 4.2 *Em AADMM, se os parâmetros são ajustados para*

$$\alpha_t = \frac{2}{t+1}, \tau_t = \rho_t \equiv \rho, \theta_t = \frac{(t-1)\rho}{t}, \eta_t = \frac{2L_G + \chi \rho t \|K\|^2}{t}, \quad (4.57)$$

então:

$$G(x_{t+1}^{ag}) + F(Kx_{t+1}^{ag}) - f^* \leq \frac{2L_G D_X^2}{t(t+1)} + \frac{1}{(t+1)} \left[\chi \rho \|K\|^2 D_X^2 + (1 - \chi) \rho D_{X,K}^2 + \frac{D_Y^2}{\rho} \right]. \quad (4.58)$$

Em particular, se ρ é dada por (4.53), temos:

$$G(x_{t+1}^{ag}) + F(Kx_{t+1}^{ag}) - f^* \leq \frac{2L_G D_X^2}{t(t+1)} + \frac{2}{(t+1)} [\chi \|K\| D_X D_Y + (1 - \chi) D_{X,K} D_Y]. \quad (4.59)$$

Dem: Sabemos que $\alpha_t = \frac{2}{t+1}$ e $\Gamma_t = \frac{2}{t(t+1)}$ satisfazem (4.26). De fato, provaremos por indução sobre t .

Para $t = 1$ temos: $\alpha_1 = \frac{2}{1+1}$ e $\Gamma_1 = \frac{2}{1(1+1)} = 1$.

Supondo válido para $t - 1$, ou seja, $\alpha_{t-1} = \frac{2}{t}$ e $\Gamma_{t-1} = \frac{2}{(t-1)t}$.

E mostremos que é válido para t .

Por (4.26), temos:

$$\Gamma_t = (1 - \alpha_t) \Gamma_{t-1}$$

substituindo α_t e Γ_{t-1} , temos:

$$\Gamma_t = \left(1 - \frac{2}{t+1}\right) \frac{2}{t(t+1)} = \frac{t-1}{t+1} \frac{2}{(t-1)t} = \frac{2}{t(t+1)}.$$

Mais ainda $\frac{\alpha_t}{\Gamma_t} = \frac{\frac{2}{t+1}}{\frac{2}{t(t+1)}} = t$.

Agora, pelos parâmetros definidos em (4.57) e pela definição de $\mathcal{B}_t(\cdot, \cdot, \cdot)$ em (4.28), podemos obter os seguintes resultados:

$$(I) : \eta_t - L_G \alpha_t - \chi \theta_t \|K\|^2 \geq 0.$$

De fato, substituindo η_t , α_t e θ_t , temos:

$$\eta_t - L_G \alpha_t - \chi \theta_t \|K\|^2 = \frac{2L_G + \chi \rho t \|K\|^2}{t} - \frac{2L_G}{t+1} - \chi \frac{(t-1)\rho}{t} \|K\|^2.$$

Reorganizando os termos, temos:

$$\eta_t - L_G \alpha_t - \chi \theta_t \|K\|^2 = \frac{2L_G}{t} - \frac{2L_G}{t+1} + \frac{\chi \rho t \|K\|^2}{t} - \frac{\chi(t-1)\rho}{t} \|K\|^2 = \frac{2L_G}{t(t+1)} + \frac{\chi \rho \|K\|^2}{t} \geq 0,$$

pois todos os termos são positivos.

$$(II) : \tau_t \geq \theta_t.$$

Por definição:

$$\tau_t = \rho \geq \frac{(t-1)}{t} = \theta_t.$$

$$(III) : A_2 = \mathcal{B}_t(w^*, w_{[t+1]}, \theta_t) - \sum_{i=1}^t \frac{\alpha_i(\tau_i - \theta_i)}{2\Gamma_i} \|w_{i+1} - w^*\|^2 = -\frac{\rho t}{2} \|w_{t+1} - w^*\|^2 \leq 0.$$

De fato, pela definição de $\mathcal{B}_t(\cdot, \cdot, \cdot)$ em (4.28), temos:

$$A_2 = \sum_{i=1}^t \frac{\alpha_i \theta_i}{2\Gamma_i} (\|w_i - w^*\|^2 - \|w_{i+1} - w^*\|^2) - \sum_{i=1}^t \frac{\alpha_i(\tau_i - \theta_i)}{2\Gamma_i} \|w_{i+1} - w^*\|^2.$$

Substituindo θ_i , α_i , Γ_i e τ_i , temos:

$$A_2 = \sum_{i=1}^t \frac{i}{2} \frac{(i-1)\rho}{i} (\|w_i - w^*\|^2 - \|w_{i+1} - w^*\|^2) - \sum_{i=1}^t \frac{i}{2} \left[\rho - \frac{(i-1)\rho}{i} \right] \|w_{i+1} - w^*\|^2.$$

Simplificando os termos, temos:

$$A_2 = \sum_{i=1}^t \frac{(i-1)\rho}{2} (\|w_i - w^*\|^2 - \|w_{i+1} - w^*\|^2) - \sum_{i=1}^t \frac{\rho}{2} \|w_{i+1} - w^*\|^2.$$

E desenvolvendo os somatórios, chegamos em:

$$\begin{aligned}
A_2 &= 0 - \frac{\rho}{2}\|w_2 - w^*\|^2 + \frac{\rho}{2}\|w_2 - w^*\|^2 - \frac{\rho}{2}\|w_3 - w^*\|^2 - \frac{\rho}{2}\|w_3 - w^*\|^2 + \dots + \frac{(t-1)\rho}{2}\|w_t - w^*\|^2 \\
&\quad - \frac{(t-1)\rho}{2}\|w_t - w^*\|^2 - \frac{t\rho}{2}\|w_t - w^*\|^2 = -\frac{\rho t}{2}\|w_{t+1} - w^*\|^2 \leq 0,
\end{aligned}$$

pois $\rho > 0$ e $t > 0$.

$$\begin{aligned}
(IV) : A_3 &= -\chi \mathcal{B}_t(Kx^*, Kx_{[t+1]}, \theta_t) + \sum_{i=1}^t \frac{\alpha_i(\tau_i - \theta_i)}{2\Gamma_i} \|Kx_{i+1} - Kx^*\|^2 \\
&= \frac{\chi \rho t}{2} \|Kx_{t+1} - Kx^*\|^2 + \frac{(1-\chi)\rho}{2} \sum_{i=1}^t \|Kx_{i+1} - Kx^*\|^2 \leq \frac{\chi \rho t}{2} \|K\|^2 \|x_{t+1} - x^*\|^2 + \frac{(1-\chi)\rho t}{2} D_{X,K}^2.
\end{aligned}$$

De fato, pela definição de $\mathcal{B}_t(\cdot, \cdot, \cdot)$ em (4.28), temos:

$$A_3 = -\chi \left[\sum_{i=1}^t \frac{\alpha_i \theta_i}{2\Gamma_i} (\|Kx_i - x^*\|^2 - \|Kx_{i+1} - x^*\|^2) \right] + \sum_{i=1}^t \frac{\alpha_i(\tau_i - \theta_i)}{2\Gamma_i} \|Kx_{i+1} - Kx^*\|^2.$$

Substituindo θ_i , α_i , Γ_i e τ_i , temos:

$$A_3 = -\chi \left[\sum_{i=1}^t \frac{i(i-1)\rho}{2} (\|Kx_i - Kx^*\|^2 - \|Kx_{i+1} - Kx^*\|^2) \right] + \sum_{i=1}^t \frac{i}{2} \left[\rho - \frac{(i-1)\rho}{i} \right] \|Kx_{i+1} - Kx^*\|^2.$$

Simplificando os termos, temos:

$$A_3 = -\chi \left[\sum_{i=1}^t \frac{(i-1)\rho}{2} (\|Kx_i - Kx^*\|^2 - \|Kx_{i+1} - Kx^*\|^2) \right] + \sum_{i=1}^t \frac{\rho}{2} \|Kx_{i+1} - Kx^*\|^2.$$

E desenvolvendo os somatórios, chegamos em:

$$\begin{aligned}
A_3 &= 0 + \frac{\rho}{2}\|Kx_2 - Kx^*\|^2 - \chi \frac{\rho}{2}\|Kx_2 - Kx^*\|^2 + \chi \frac{\rho}{2}\|Kx_3 - Kx^*\|^2 + \frac{\rho}{2}\|Kx_3 - Kx^*\|^2 + \dots + \\
&\quad + \chi \frac{(t-1)\rho}{2} (\|Kx_t - Kx^*\|^2 - \|Kx_{t+1} - Kx^*\|^2) + \frac{\rho}{2}\|Kx_{t+1} - Kx^*\|^2 \\
&= \frac{\chi \rho t}{2} \|Kx_{t+1} - Kx^*\|^2 + \frac{(1-\chi)\rho}{2} \sum_{i=1}^t \|Kx_{i+1} - Kx^*\|^2.
\end{aligned}$$

Cada termo do somatório na equação acima é substituído por $D_{X,K}^2$ gerando a desigualdade:

$$A_3 \leq \frac{\chi \rho t}{2} \|K\|^2 \|x_{t+1} - x^*\|^2 + \frac{(1-\chi)\rho t}{2} D_{X,K}^2.$$

$$(V) : \mathcal{B}_t(y, y_{[t+1]}, \rho_t^{-1}) \leq \frac{t}{2\rho} D_Y^2 \quad \forall y \in Y.$$

Temos, por (4.18), (4.20) e pelo Teorema 1.3, que $y_{[t+1]} \subset Y$ (A projeção não permite que um ponto da sequência fique fora do conjunto Y). Usando o fato de que $\frac{\alpha_t}{\Gamma_t \rho_t} = \frac{t}{\rho}$ e aplicando (4.51) do lema 4.2, temos:

$$\mathcal{B}_t(y, y_{[t+1]}, \rho_t^{-1}) \stackrel{(4.51)}{\leq} \frac{\alpha_t}{2\Gamma_t} \rho_t^{-1} D_Y^2 - \frac{\alpha_t}{2\Gamma_t} \rho_t^{-1} \|y_{t+1} - y\|^2.$$

Como $-\frac{\alpha_t}{2\Gamma_t} \rho_t^{-1} \|y_{t+1} - y\|^2 \leq 0$, temos:

$$\mathcal{B}_t(y, y_{[t+1]}, \rho_t^{-1}) \leq \frac{\alpha_t}{2\Gamma_t} \rho_t^{-1} D_Y^2.$$

Substituindo $\frac{\alpha_t}{\Gamma_t \rho_t}$ por $\frac{t}{\rho}$, resulta em:

$$\mathcal{B}_t(y, y_{[t+1]}, \rho_t^{-1}) \leq \frac{t}{2\rho} D_Y^2.$$

$$(VI) : \mathcal{B}_t(x^*, x_{[t+1]}, \eta_t) \leq L_G D_X^2 + \frac{\chi \rho t}{2} \|K\|^2 D_X^2 - \frac{\chi \rho t}{2} \|K\|^2 \|x_{t+1} - x^*\|^2.$$

Observando que $\frac{\alpha_t \eta_t}{2\Gamma_t} = \frac{2}{t+1} \frac{2L_G + \chi \rho t \|K\|^2}{t} \frac{t(t+1)}{2} = 2L_G + \chi \rho t \|K\|^2$, por (4.51) no Lema 4.2, temos:

$$\mathcal{B}_t(x^*, x_{[t+1]}, \eta_t) \stackrel{(4.51)}{\leq} \frac{\alpha_t}{2\Gamma_t} \eta_t D_X^2 - \frac{\alpha_t}{2\Gamma_t} \eta_t \|x_{t+1} - x^*\|^2.$$

Substituindo $\frac{\alpha_t \eta_t}{2\Gamma_t}$, obtemos:

$$\mathcal{B}_t(x^*, x_{[t+1]}, \eta_t) \leq \frac{2L_G + \chi \rho t \|K\|^2}{2} D_X^2 - \frac{2L_G + \chi \rho t \|K\|^2}{2} \|x_{t+1} - x^*\|^2.$$

Reorganizando os termos, obtemos:

$$\mathcal{B}_t(x^*, x_{[t+1]}, \eta_t) \leq L_G D_X^2 + \frac{\chi \rho t \|K\|^2}{2} D_X^2 - 2L_G \|x_{t+1} - x^*\|^2 - \frac{\chi \rho t \|K\|^2}{2} \|x_{t+1} - x^*\|^2.$$

Como $-2L_G \|x_{t+1} - x^*\|^2 \leq 0$, segue a desigualdade:

$$\mathcal{B}_t(x^*, x_{[t+1]}, \eta_t) \leq L_G D_X^2 + \frac{\chi \rho t \|K\|^2}{2} D_X^2 - \|x^*\|^2 - \frac{\chi \rho t \|K\|^2}{2} \|x_{t+1} - x^*\|^2.$$

Por (4.27), sabemos que:

$$\frac{1}{\Gamma_t} Q(w^*, x^*, y^*; z_{t+1}^{ag}) - \sum_{i=2}^t \left(\frac{1 - \alpha_i}{\Gamma_i} - \frac{1}{\Gamma_{i-1}} \right) Q(w^*, x^*, y^*; z_i^{ag})$$

$$\begin{aligned}
&\leq \mathcal{B}_t(x^*, x_{[t+1]}, \eta_{[t]}) + \mathcal{B}_t(y, y_{[t+1]}, \rho_{[t]}^{-1}) + \mathcal{B}_t(Bw^*, Bw_{[t+1]}, \theta_{[t]}) - \chi \mathcal{B}_t(Kx^*, Kx_{[t+1]}, \theta_{[t]}) \\
&\quad - \sum_{i=1}^t \frac{\alpha_i(\tau_i - \theta_i)}{2\Gamma_i} \|Bw_{i+1} - Kx^* - b\|^2 + \sum_{i=1}^t \frac{\alpha_i(\tau_i - \theta_i)}{2\Gamma_i} \|B(x_{i+1} - x^*)\|^2 \\
&\quad - \sum_{i=1}^t \frac{\alpha_i(\tau_i - \rho_i)}{2\Gamma_i \rho_i} \|y_i - y_{i+1}\|^2 - \sum_{i=1}^t \frac{\alpha_i}{2\Gamma_i} (\eta_i - L_G \alpha_i - \chi \theta_i \|K\|^2) \|x_i - x_{i+1}\|^2.
\end{aligned}$$

Substituindo os parâmetros determinados em (4.57), concluímos que:

$$-\sum_{i=2}^t \left(\frac{1 - \alpha_i}{\Gamma_i} - \frac{1}{\Gamma_{i-1}} \right) Q(w^*, x^*, y^*; z_i^{ag}) = 0,$$

pois temos:

$$\begin{aligned}
\left(\frac{1 - \frac{2}{i+1}}{\frac{2}{i(i-1)}} - \frac{1}{\frac{2}{(i-1)i}} \right) &= \left(\frac{(1-i)(i+1)i}{(i+1)2} - \frac{(i-1)i}{2} \right) = 0. \\
-\sum_{i=1}^t \frac{\alpha_i(\tau_i - \rho_i)}{2\Gamma_i \rho_i} \|y_i - y_{i+1}\|^2 &= 0,
\end{aligned}$$

pois $\tau_t = \rho_t$.

E, finalmente:

$$-\sum_{i=1}^t \frac{\alpha_i(\tau_i - \theta_i)}{2\Gamma_i} \|Bw_{i+1} - Kx^* - b\|^2 + \sum_{i=1}^t \frac{\alpha_i(\tau_i - \theta_i)}{2\Gamma_i} \|B(x_{i+1} - x^*)\|^2 = 0.$$

Já que $\|Bw_{i+1} - Kx^* - b\|^2 = \|w_{i+1} - Kx^*\|^2$ (B é o operador identidade e $b = 0$). Além disso, pelo Algoritmo 4.2, $Bw_{i+1} - Kx_{i+1} - b = 0 \implies w_{i+1} - Kx_{i+1} = 0 \implies w_{i+1} = Kx_{i+1}$. Deste modo, $\|w_{i+1} - Kx^*\|^2 = \|Kx_{i+1} - Kx^*\|^2 = \|K(x_{i+1} - x^*)\|^2$ (o que torna os termos idênticos).

Logo, podemos escrever:

$$\begin{aligned}
\frac{1}{\Gamma_t} Q(w^*, x^*, y^*; z_{t+1}^{ag}) &\leq L_G D_X^2 + \frac{\chi \rho t}{2} \|K\|^2 D_X^2 - \frac{\chi \rho t}{2} \|K\|^2 \|x_{t+1} - x^*\|^2 + \frac{t}{2\rho} D_Y^2 \\
&\quad - \frac{\rho t}{2} \|w_{t+1} - w^*\|^2 + \frac{\chi \rho t}{2} \|K\|^2 \|x_{t+1} - x^*\|^2 + \frac{(1-\chi)\rho t}{2} D_{X,K}^2 \\
&\quad - \sum_{i=1}^t \frac{\alpha_i}{2\Gamma_i} (\eta_i - L_G \alpha_i - \chi \theta_i \|K\|^2) \|x_i - x_{i+1}\|^2.
\end{aligned}$$

Por I , temos: $-\sum_{i=1}^t \frac{\alpha_i}{2\Gamma_i} (\eta_i - L_G \alpha_i - \chi \theta_i \|K\|^2) \|x_i - x_{i+1}\|^2 \leq 0$. Além disso, $-\frac{\rho t}{2} \|w_{t+1} - w^*\|^2 \leq 0$. Simplificando e desprezando esses termos negativos, obtemos:

$$\frac{1}{\Gamma_t} Q(w^*, x^*, y^*; z_{t+1}^{ag}) \leq L_G D_X^2 + \frac{\chi \rho t}{2} \|K\|^2 D_X^2 + \frac{(1-\chi)\rho t}{2} D_{X,K}^2 + \frac{t}{2\rho} D_Y^2.$$

Usando os parâmetros $\alpha_t = \frac{2}{t+1}$ e $\Gamma_t = \frac{2}{t(t+1)}$, pela proposição 4.2, temos: $\hat{w} = (B^*)^{-1}(Kx + b)$. Como B é o operador identidade e $b = 0$, temos $\hat{w} = Kx$. Daí:

$$F(w_{t+1}^{ag}) = F(Kx_{t+1}^{ag}). \quad (4.60)$$

e, consequentemente:

$$\begin{aligned} G(x_{t+1}^{ag}) + F(Kx_{t+1}^{ag}) - f^* &\leq \frac{1}{\Gamma_t} \sup_{y \in Y} Q(w^*, x^*, y; z) \\ &\leq \frac{2}{t(t+1)} \sup_{y \in Y} \left[L_G D_X^2 + \frac{\chi \rho t}{2} \|K\|^2 D_X^2 + \frac{(1-\chi)\rho t}{2} D_{X,K}^2 + \frac{t}{2\rho} D_Y^2 \right]. \end{aligned}$$

Como D_Y^2 já é definido como $\sup_{y \in Y}$, temos:

$$G(x_{t+1}^{ag}) + F(Kx_{t+1}^{ag}) - f^* \leq \frac{2L_G D_X^2}{t(t+1)} + \frac{1}{(t+1)} \left[\chi \rho \|K\|^2 D_X^2 + (1-\chi)\rho D_{X,K}^2 + \frac{D_Y^2}{\rho} \right], \quad \forall y \in Y.$$

provando (4.58).

E se ρ é dado por (4.53), a partir de (4.58), substituindo ρ , temos:

$$\left[\frac{\chi D_Y \|K\|^2 D_X^2}{\chi \|K\| D_X + (1-\chi) D_{X,K}} + \frac{(1-\chi) D_Y D_{X,K}^2}{\chi \|K\| D_X + (1-\chi) D_{X,K}} + \frac{D_Y^2 \chi \|K\| D_X + (1-\chi) D_{X,K}}{D_Y} \right],$$

pois $\chi \neq 0$ se $(1-\chi) = 0$.

Simplificando as expressões, obtemos:

$$\left[\chi \rho \|K\|^2 D_X^2 + (1-\chi)\rho D_{X,K}^2 + \frac{D_Y^2}{\rho} \right] = [\chi D_Y \|K\| D_X + (1-\chi) D_{X,K} D_Y + D_Y (\chi \|K\| D_X + (1-\chi) D_{X,K})].$$

Agrupando os termos, temos:

$$\left[\chi \rho \|K\|^2 D_X^2 + (1-\chi)\rho D_{X,K}^2 + \frac{D_Y^2}{\rho} \right] = 2[\chi D_Y \|K\| D_X + (1-\chi) D_{X,K} D_Y].$$

Logo:

$$G(x_{t+1}^{ag}) + F(Kx_{t+1}^{ag}) - f^* \leq \frac{2L_G D_X^2}{t(t+1)} + \frac{2}{(t+1)} [\chi \|K\| D_X D_Y + (1-\chi) D_{X,K} D_Y].$$

provando (4.59)

■

Podemos fazer várias observações sobre os algoritmos AADMM a partir dos Teoremas 4.1 e 4.2. Podemos afirmar que o Teorema 4.2 fornece um exemplo de como escolher os parâmetros em AL-ADMM e ALP-ADMM que levam a uma melhor convergência com relação a dependência de L_G de L-ADMM e LP-ADMM, respectivamente. Notamos, ainda que AL-ADMM e ALP-ADMM permitem L_G tão grande quanto se queira, sem afetar a taxa de convergência (a menos de um fator constante). Outra consideração importante é que enquanto o Teorema 4.1 descreve apenas a convergência dos algoritmos ADMM, o Teorema 4.2 descreve a convergência de sequências agregadas $\{z_{t+1}^{ag}\}_{t \geq 1}$, que são exatamente as saídas dos esquemas acelerados. Finalmente, em métodos ADMM temos $\tau_t = \rho_t = \theta_t$, enquanto no Teorema 4.2 só temos $\tau_t = \rho_t$, embora $\theta_t \rightarrow \rho_t$ quando $t \rightarrow \infty$. Na verdade é possível escolher o conjunto de parâmetros iguais, tal como será descrito no próximo teorema (OUYANG et al., 2015) [26].

Teorema 4.3 *Em AADMM, se o número total de iterações N é escolhido, e os parâmetros são ajustados para:*

$$\alpha_t = \frac{2}{t+1}, \theta_t = \tau_t = \rho_t = \frac{\rho N}{t}, \eta_t = \frac{2L_G + \chi \rho N \|K\|^2}{t}$$

em que ρ é dado por (4.53), temos:

$$G(x_N^{ag}) + F(Kx_N^{ag}) - f^* \leq \frac{2L_G D_X^2}{N(N-1)} + \frac{2}{N-1} [\chi \rho \|K\| D_X D_Y + (1-\chi) \rho D_{X,K} D_Y]. \quad (4.61)$$

Dem: Sabemos que $\alpha_t = \frac{2}{t+1}$, $\Gamma_t = \frac{2}{t(t+1)}$ satisfazem (4.26). E, ainda que, $\frac{\alpha_t}{\Gamma_t} = \frac{\frac{2}{t+1}}{\frac{2}{t(t+1)}} = t$. Agora, pelos parâmetros definidos no Teorema 4.3 e pela definição de $\mathcal{B}_t(\cdot, \cdot, \cdot)$ em (4.28), podemos obter os seguintes resultados:

$$(I) : \eta_t - L_G \alpha_t - \chi \theta_t \|K\|^2 \geq 0.$$

De fato, substituindo η_t , α_t e θ_t , temos:

$$\eta_t - L_G \alpha_t - \chi \theta_t \|K\|^2 = \frac{2L_G + \chi \rho N \|K\|^2}{t} - \frac{2L_G}{t+1} - \chi \frac{(t-1)\rho N}{t} \|K\|^2.$$

Reorganizando os termos, temos:

$$\eta_t - L_G \alpha_t - \chi \theta_t \|K\|^2 = \frac{2L_G}{t} - \frac{2L_G}{t+1} + \frac{\chi \rho N \|K\|^2}{t} - \frac{\chi \rho N}{t} \|K\|^2 = \frac{2L_G}{t(t+1)} \geq 0,$$

pois todos os termos são positivos.

$$(II) : \mathcal{B}_t(x^*, x_{[t+1]}, \eta_t) \leq \frac{2L_G + \chi \rho N \|K\|^2}{2} (D_{x^*}^2 - \|x_{t+1} - x^*\|^2).$$

De fato, pela definição de $\mathcal{B}_t(\cdot, \cdot, \cdot)$ em (4.28), temos:

$$\mathcal{B}_t(x^*, x_{[t+1]}, \eta_t) = \sum_{i=1}^t \frac{\alpha_i \eta_i}{2\Gamma_i} (\|x_i - x^*\|^2 - \|x_{i+1} - x^*\|^2).$$

Substituindo η_i , α_i e Γ_i , temos:

$$\mathcal{B}_t(x^*, x_{[t+1]}, \eta_t) = \sum_{i=1}^t \frac{i}{2} \left(\frac{2L_G + \chi \rho N \|K\|^2}{i} \right) (\|x_i - x^*\|^2 - \|x_{i+1} - x^*\|^2).$$

Simplificando os termos, obtemos:

$$\mathcal{B}_t(x^*, x_{[t+1]}, \eta_t) = \frac{2L_G + \chi \rho N \|K\|^2}{2} \sum_{i=1}^t (\|x_i - x^*\|^2 - \|x_{i+1} - x^*\|^2).$$

E desenvolvendo os somatórios, temos:

$$\mathcal{B}_t(x^*, x_{[t+1]}, \eta_t) = \frac{2L_G + \chi \rho N \|K\|^2}{2} (\|x_1 - x^*\|^2 - \|x_2 - x^*\|^2 + \|x_2 - x^*\|^2 - \dots + \|x_t - x^*\|^2 - \|x_{t+1} - x^*\|^2).$$

Simplificando, obtemos:

$$\mathcal{B}_t(x^*, x_{[t+1]}, \eta_t) = \frac{2L_G + \chi \rho N \|K\|^2}{2} (\|x_1 - x^*\|^2 - \|x_{t+1} - x^*\|^2).$$

e por (4.7), temos:

$$\mathcal{B}_t(x^*, x_{[t+1]}, \eta_t) \leq \frac{2L_G + \chi \rho N \|K\|^2}{2} (D_{x^*}^2 - \|x_{t+1} - x^*\|^2).$$

$$(III) : \mathcal{B}_t(w^*, w_{[t+1]}, \theta_t) \leq \frac{\rho N D_{x^*, K}^2}{2}.$$

De fato, pela definição de $\mathcal{B}_t(\cdot, \cdot, \cdot)$ em (4.28), temos:

$$\mathcal{B}_t(w^*, w_{[t+1]}, \theta_t) = \sum_{i=1}^t \frac{\alpha_i \theta_i}{2\Gamma_i} (\|w_i - w^*\|^2 - \|w_{i+1} - w^*\|^2).$$

Substituindo θ_i , α_i e Γ_i , temos:

$$\mathcal{B}_t(w^*, w_{[t+1]}, \theta_t) = \sum_{i=1}^t \frac{i}{2} \frac{\rho N}{i} (\|w_i - w^*\|^2 - \|w_{i+1} - w^*\|^2).$$

Simplificando os termos, obtemos:

$$\mathcal{B}_t(w^*, w_{[t+1]}, \theta_t) = \frac{\rho N}{2} \sum_{i=1}^t (\|w_i - w^*\|^2 - \|w_{i+1} - w^*\|^2).$$

E desenvolvendo os somatórios, temos:

$$\begin{aligned}
\mathcal{B}_t(w^*, w_{[t+1]}, \theta_t) &= \frac{\rho N}{2} (\|w_1 - w^*\|^2 - \|w_2 - w^*\|^2 + \|w_2 - w^*\|^2 - \|w_3 - w^*\|^2 + \cdots + \|w_t - w^*\|^2 - \|w_{t+1} - w^*\|^2) \\
&= \frac{\rho N}{2} (\|w_1 - w^*\|^2 - \|w_{t+1} - w^*\|^2) \\
&\leq \frac{\rho N}{2} \|w_1 - w^*\|^2 = \frac{\rho N}{2} \|Kx_1 - Kx^*\|^2.
\end{aligned}$$

E, finalmente, por (4.7), temos:

$$\mathcal{B}_t(w^*, w_{[t+1]}, \theta_t) \leq \frac{\rho N D_{x^*, K}^2}{2}.$$

$$(IV) : -\chi \mathcal{B}_t(Kx^*, Kx_{[t+1]}, \theta_t) \leq -\frac{\chi \rho N}{2} D_{x^*, K}^2 + \frac{\chi \rho N}{2} \|K\|^2 \|x_{t+1} - x^*\|^2.$$

De fato, pela definição de $\mathcal{B}_t(\cdot, \cdot, \cdot)$ em (4.28), temos:

$$-\chi \mathcal{B}_t(Kx^*, Kx_{[t+1]}, \theta_t) = -\chi \left[\sum_{i=1}^t \frac{\alpha_i \theta_i}{2\Gamma_i} (\|Kx_i - x^*\|^2 - \|Kx_{i+1} - x^*\|^2) \right].$$

Substituindo θ_i , α_i e Γ_i , temos:

$$-\chi \mathcal{B}_t(Kx^*, Kx_{[t+1]}, \theta_t) = -\chi \left[\sum_{i=1}^t \frac{i}{2} \frac{\rho N}{i} (\|Kx_i - Kx^*\|^2 - \|Kx_{i+1} - Kx^*\|^2) \right].$$

Simplificando os termos, temos:

$$-\chi \mathcal{B}_t(Kx^*, Kx_{[t+1]}, \theta_t) = -\chi \frac{\rho N}{2} \sum_{i=1}^t (\|Kx_i - Kx^*\|^2 - \|Kx_{i+1} - Kx^*\|^2).$$

E desenvolvendo os somatórios, temos:

$$-\chi \mathcal{B}_t(Kx^*, Kx_{[t+1]}, \theta_t) = -\chi \frac{\rho N}{2} (\|Kx_1 - Kx^*\|^2 - \|Kx_2 - Kx^*\|^2 + \|Kx_2 - Kx^*\|^2 - \cdots + \|Kx_t - Kx^*\|^2 - \|Kx_{t+1} - Kx^*\|^2).$$

Simplificando, temos:

$$-\chi \mathcal{B}_t(Kx^*, Kx_{[t+1]}, \theta_t) = -\chi \frac{\rho N}{2} (\|Kx_1 - Kx^*\|^2 - \|Kx_{t+1} - Kx^*\|^2) = -\chi \frac{\rho N}{2} \|Kx_1 - Kx^*\|^2 + \chi \frac{\rho N}{2} \|Kx_{t+1} - Kx^*\|^2,$$

e, por (4.7), temos:

$$-\chi \mathcal{B}_t(Kx^*, Kx_{[t+1]}, \theta_t) \leq -\frac{\chi \rho N}{2} D_{x^*, K}^2 + \frac{\chi \rho N}{2} \|K\|^2 \|x_{t+1} - x^*\|^2.$$

$$(V) : \mathcal{B}_t(y, y_{[t+1]}, \rho_t^{-1}) \leq \frac{N}{2\rho} D_Y^2 \quad \forall y \in Y.$$

Observando que $\frac{\alpha_t}{\Gamma_t \rho_t} = \frac{t}{\frac{\rho N}{t}} = \frac{t^2}{\rho N}$ por (4.51), do Lema 4.2, temos:

$$\mathcal{B}_t(y, y_{[t+1]}, \rho_t^{-1}) \stackrel{(4.51)}{\leq} \frac{\alpha_t}{2\Gamma_t} \rho_t^{-1} D_Y^2 - \frac{\alpha_t}{2\Gamma_t} \rho_t^{-1} \|y_{t+1} - y\|^2.$$

Como $-\frac{\alpha_t}{2\Gamma_t} \rho_t^{-1} \|y_{t+1} - y\|^2 \leq 0$, temos:

$$\mathcal{B}_t(y, y_{[t+1]}, \rho_t^{-1}) \leq \frac{\alpha_t}{2\Gamma_t} \rho_t^{-1} D_Y^2.$$

Substituindo $\frac{\alpha_t}{\Gamma_t \rho_t}$ por $\frac{t^2}{\rho N}$, temos:

$$\mathcal{B}_t(y, y_{[t+1]}, \rho_t^{-1}) = \frac{t^2}{2\rho N} D_Y^2 \leq \frac{N^2}{2\rho N} D_Y^2 = \frac{N}{2\rho} D_Y^2, \quad \forall t \leq N.$$

Por (4.27), temos:

$$\begin{aligned} & \frac{1}{\Gamma_t} Q(w^*, x^*, y^*; z_{t+1}^{ag}) - \sum_{i=2}^t \left(\frac{1 - \alpha_i}{\Gamma_i} - \frac{1}{\Gamma_{i-1}} \right) Q(w^*, x^*, y^*; z_i^{ag}) \\ & \leq \mathcal{B}_t(x^*, x_{[t+1]}, \eta_{[t]}) + \mathcal{B}_t(y, y_{[t+1]}, \rho_{[t]}^{-1}) + \mathcal{B}_t(Bw^*, Bw_{[t+1]}, \theta_{[t]}) - \chi \mathcal{B}_t(Kx^*, Kx_{[t+1]}, \theta_{[t]}) \\ & \quad - \sum_{i=1}^t \frac{\alpha_i(\tau_i - \theta_i)}{2\Gamma_i} \|Bw_{i+1} - Kx^* - b\|^2 + \sum_{i=1}^t \frac{\alpha_i(\tau_i - \theta_i)}{2\Gamma_i} \|B(x_{i+1} - x^*)\|^2 \\ & \quad - \sum_{i=1}^t \frac{\alpha_i(\tau_i - \rho_i)}{2\Gamma_i \rho_i} \|y_i - y_{i+1}\|^2 - \sum_{i=1}^t \frac{\alpha_i}{2\Gamma_i} (\eta_i - L_G \alpha_i - \chi \theta_i \|K\|^2) \|x_i - x_{i+1}\|^2. \end{aligned}$$

Substituindo os parâmetros determinados, temos:

$$- \sum_{i=2}^t \left(\frac{1 - \alpha_i}{\Gamma_i} - \frac{1}{\Gamma_{i-1}} \right) Q(w^*, x^*, y^*; z_i^{ag}) = 0,$$

pois temos:

$$\left(\frac{1 - \frac{2}{i+1}}{\frac{2}{i(i-1)}} - \frac{1}{\frac{2}{(i-1)i}} \right) = \left(\frac{(1-i)(i+1)i}{(i+1)2} - \frac{(i-1)i}{2} \right) = 0.$$

O fato de $\tau_t = \theta_t = \rho_t$, resulta em:

$$\begin{aligned} & - \sum_{i=1}^t \frac{\alpha_i(\tau_i - \theta_i)}{2\Gamma_i} \|Bw_{i+1} - Kx^* - b\|^2 = 0, \\ & \sum_{i=1}^t \frac{\alpha_i(\tau_i - \theta_i)}{2\Gamma_i} \|B(x_{i+1} - x^*)\|^2 = 0, \\ & - \sum_{i=1}^t \frac{\alpha_i(\tau_i - \rho_i)}{2\Gamma_i \rho_i} \|y_i - y_{i+1}\|^2 = 0. \end{aligned}$$

Logo, temos:

$$\begin{aligned} \frac{1}{\Gamma_t} Q(w^*, x^*, y; z_{t+1}^{ag}) &\leq L_G D_{x^*}^2 + \frac{\chi \rho N}{2} \|K\|^2 D_{x^*}^2 - L_G \|x_{t+1} - x^*\|^2 - \frac{\chi \rho N}{2} \|K\|^2 \|x_{t+1} - x^*\|^2 \\ &\quad + \frac{N}{2\rho} D_Y^2 + \frac{\rho N D_{x^*,K}^2}{2} - \frac{\chi \rho N}{2} D_{x^*,K}^2 + \frac{\chi \rho N}{2} \|K\|^2 \|x_{t+1} - x^*\|^2 \\ &\quad - \sum_{i=1}^t \frac{\alpha_i}{2\Gamma_i} (\eta_i - L_G \alpha_i - \chi \theta_i \|K\|^2) \|x_i - x_{i+1}\|^2. \end{aligned}$$

Por I , temos: $-\sum_{i=1}^t \frac{\alpha_i}{2\Gamma_i} (\eta_i - L_G \alpha_i - \chi \theta_i \|K\|^2) \|x_i - x_{i+1}\|^2 \leq 0$. Além disso, $-\frac{L_G}{2} \|x_{t+1} - x^*\|^2 \leq 0$. Simplificando e desprezando esses termos negativos, obtemos:

$$\frac{1}{\Gamma_t} Q(w^*, x^*, y; z_{t+1}^{ag}) \leq L_G D_{x^*}^2 + \frac{\chi \rho N}{2} \|K\|^2 D_{x^*}^2 + \frac{N}{2\rho} D_Y^2 + \frac{(1-\chi)\rho N D_{x^*,K}^2}{2}.$$

e, por (4.7), temos:

$$\frac{1}{\Gamma_t} Q(w^*, x^*, y; z_{t+1}^{ag}) \leq L_G D_X^2 + \frac{\chi \rho N}{2} \|K\|^2 D_X^2 + \frac{N}{2\rho} D_Y^2 + \frac{(1-\chi)\rho N D_{X,K}^2}{2}.$$

Pela Proposição 4.2, temos: $\hat{w} = (B^*)^{-1}(Kx + b)$. Como B é o operador identidade e $b = 0$, temos $\hat{w} = Kx$. Daí:

$$F(w_{t+1}^{ag}) = F(Kx_{t+1}^{ag}).$$

e, consequentemente:

$$G(x_{t+1}^{ag}) + F(Kx_{t+1}^{ag}) - f^* \leq \frac{1}{\Gamma_t} \sup_{y \in Y} Q(w^*, x^*, y; z_{t+1}^{ag}).$$

Pela desigualdade acima e tomando $t = N - 1$, temos:

$$G(x_N^{ag}) + F(Kx_N^{ag}) - f^* \leq \frac{1}{\Gamma_t} \left[L_G D_X^2 + \frac{\chi \rho N}{2} \|K\|^2 D_X^2 + \frac{N}{2\rho} D_Y^2 + \frac{(1-\chi)\rho N D_{X,K}^2}{2} \right].$$

Substituindo ρ de acordo com (4.53), e de maneira totalmente análoga ao que foi feito no Teorema 4.2, para mostrar (4.59), temos:

$$G(x_N^{ag}) + F(Kx_N^{ag}) - f^* \leq \frac{2L_G D_X^2}{N(N-1)} + \frac{2}{N-1} [\chi \rho \|K\| D_X D_Y + (1-\chi) \rho D_{X,K} D_Y].$$

■

As taxas de convergência de AADMM são obtidas a partir do Teorema 4.2 e podem ser sintetizadas e comparadas conforme a tabela abaixo:

Tabela 4.1 : Taxas de convergência de instâncias AADMM para resolver problemas UCO com conjunto viável limitado

	Não pré-condicionada ($\chi = 0$)	pré-condicionada ($\chi = 1$)
ADMM	$\mathcal{O}\left(\frac{D_{X,K}D_Y}{t}\right)$	$\mathcal{O}\left(\frac{\ K\ D_XD_Y}{t}\right)$
L-ADMM	$\mathcal{O}\left(\frac{L_GD_X^2}{t} + \frac{D_{X,K}D_Y}{t}\right)$	$\mathcal{O}\left(\frac{L_GD_X^2}{t} + \frac{\ K\ D_XD_Y}{t}\right)$
Acelerado (AADMM)	$\mathcal{O}\left(\frac{L_GD_X^2}{t^2} + \frac{D_{X,K}D_Y}{t}\right)$	$\mathcal{O}\left(\frac{L_GD_X^2}{t^2} + \frac{\ K\ D_XD_Y}{t}\right)$

4.5 Resultados de convergência na resolução de problemas AECCO

Estudaremos, agora, a taxa de convergência de AADMM para resolver problemas gerais AECCO sem limitações pressupostas para X ou Y , considerando tanto termos de resíduos viabilidade quanto de resíduos primal. Iniciaremos nosso estudo com a análise de convergência de algoritmos ADMM num caso especial de AADMM, em que $\alpha_t \equiv 1$, $\theta_t = \tau_t = \rho_t = \rho$.

Teorema 4.4 *Em AADMM se os parâmetros são definidos da seguinte maneira $\alpha_t \equiv 1$, $\theta_t = \tau_t = \rho_t = \rho$ e $\eta_t \equiv \eta \geq L_G + \chi\rho\|K\|^2$, então, temos:*

$$G(x^{t+1}) + F(w^{t+1}) - f^* \leq \frac{1}{2t} (\eta D_{x^*}^2 + \rho(1 - \chi) D_{x^*,K}^2) \quad (4.62)$$

e,

$$\|Bw^{t+1} - Kx^{t+1} - b\|^2 \leq \frac{2}{t^2} \left(\frac{2D_{y^*}^2}{\rho^2} + \frac{\eta D_{x^*}^2}{\rho} + (1 - \chi) D_{x^*,K}^2 \right) \quad (4.63)$$

em que $x^{t+1} := \frac{1}{t} \sum_{i=2}^{t+1} x_i$ e $w^{t+1} := \frac{1}{t} \sum_{i=2}^{t+1} w_i$. E, se $\rho = 1$ e $\eta = L_G + \chi\|K\|^2$, temos :

$$G(x^{t+1}) + F(w^{t+1}) - f^* \leq \frac{1}{2t} (L_G D_{x^*}^2 + \chi\|K\|^2 D_{x^*}^2 + (1 - \chi) D_{x^*,K}^2) \quad (4.64)$$

e,

$$\|Bw^{t+1} - Kx^{t+1} - b\|^2 \leq \frac{2\sqrt{L_G}}{t} D_{x^*} + \frac{\chi\sqrt{2}\|K\|D_{x^*}}{t} + \frac{(1 - \chi)\sqrt{2}D_{x^*,K}}{t} + \frac{2D_{y^*}}{t}. \quad (4.65)$$

Dem: Semelhante à demonstração do Teorema 4.1, temos:

$$\begin{aligned} Q(w^*, x^*, y; z_{t+1}) &\leq \frac{1}{\Gamma_t} \sum_{i=2}^{t+1} Q(w^*, x^*, y; z_i) \\ &\leq \frac{1}{2t} \left(L_G D_{x^*}^2 + \chi\rho\|K\|^2 D_{x^*}^2 + (1 - \chi)\rho D_{x^*,K}^2 + \frac{1}{\rho} (\|y_1 - y\|^2 - \|y_{t+1} - y\|^2) \right) \\ &= \frac{1}{2t} (L_G D_{x^*}^2 + \chi\rho\|K\|^2 D_{x^*}^2 + (1 - \chi)\rho D_{x^*,K}^2) + \frac{1}{2t\rho} (\|y_1 - y\|^2 - \|y_{t+1} - y\|^2). \end{aligned} \quad (4.66)$$

Iremos, agora, obter uma majoração para o termo $\|y_1 - y\|^2 - \|y_{t+1} - y\|^2$. Sabemos que:

$$\|y_1 - y\|^2 - \|y_{t+1} - y\|^2 = \langle y_1 - y, y_1 - y \rangle - \langle y_{t+1} - y, y_{t+1} - y \rangle.$$

Usando a linearidade e a associatividade do produto interno, temos:

$$\|y_1 - y\|^2 - \|y_{t+1} - y\|^2 = \langle y_1 - y, y_1 \rangle - \langle y_1 - y, y \rangle - \langle y_{t+1} - y, y_{t+1} \rangle + \langle y_{t+1} - y, y \rangle.$$

Novamente usando a linearidade e a associatividade do produto interno, temos:

$$\|y_1 - y\|^2 - \|y_{t+1} - y\|^2 = \langle y - y_1 + y_{t+1} - y, y \rangle + \langle y_1, y_1 \rangle - \langle y, y_1 \rangle - \langle y_{t+1}, y_{t+1} \rangle + \langle y, y_{t+1} \rangle.$$

Como estamos num espaço de Hilbert, temos:

$$\|y_1 - y\|^2 - \|y_{t+1} - y\|^2 = \langle y_{t+1} - y_1, y \rangle + \|y_1\|^2 - \langle y, y_1 \rangle - \|y_{t+1}\|^2 + \langle y, y_{t+1} \rangle.$$

Usando a linearidade e a associatividade do produto interno, temos:

$$\|y_1 - y\|^2 - \|y_{t+1} - y\|^2 = 2\langle y_{t+1} - y_1, y \rangle + \|y_1\|^2 - \|y_{t+1}\|^2.$$

O fato de $\{y_t\}_{t \geq 1}$ ser uma sequência crescente implica na seguinte desigualdade:

$$\|y_1 - y\|^2 - \|y_{t+1} - y\|^2 \leq 2\langle y_{t+1} - y_1, y \rangle. \quad (4.67)$$

assim, substituindo esta majoração em (4.66), temos:

$$\frac{1}{\Gamma_t} \sum_{i=2}^{t+1} Q(w^*, x^*, y; z_i) \leq \frac{1}{2t} \left(L_G D_{x^*}^2 + \chi \rho \|K\|^2 D_{x^*}^2 + (1 - \chi) \rho D_{x^*, K}^2 \right) - \left\langle \frac{1}{t\rho} (y_{t+1} - y_1, y) \right\rangle, \quad (4.68)$$

onde $z^{t+1} = \sum_{i=2}^{t+1} z_i$. Observando que $Q(z^*, z^{t+1}) \geq 0$, por (4.66), temos:

$$0 \leq \frac{1}{2t} \left(L_G D_{x^*}^2 + \chi \rho \|K\|^2 D_{x^*}^2 + (1 - \chi) \rho D_{x^*, K}^2 \right) + \frac{1}{2t\rho} (\|y_1 - y^*\|^2 - \|y_{t+1} - y^*\|^2).$$

Multiplicando por $2t\rho$ e somando $\frac{1}{\rho} \|y_{t+1} - y^*\|^2$ em ambos os lados da desigualdade, temos:

$$\|y_{t+1} - y^*\|^2 \leq \rho L_G D_{x^*}^2 + \chi \rho^2 \|K\|^2 D_{x^*}^2 + (1 - \chi) \rho^2 D_{x^*, K}^2 \|y_1 - y^*\|^2.$$

Por (4.7), temos:

$$\|y_{t+1} - y^*\|^2 \leq \rho L_G D_{x^*}^2 + \chi \rho^2 \|K\|^2 D_{x^*}^2 + (1 - \chi) \rho^2 D_{x^*, K}^2 + D_y^{2*}.$$

Se tomarmos $v_{t+1} = \frac{(y_1 - y_{t+1})}{\rho t}$, então, teremos:

$$\|v_{t+1}\|^2 = \left\| \frac{(y_1 - y_{t+1})}{\rho t} \right\|^2 = \frac{1}{\rho^2 t^2} \|(y_1 - y^* + y^* - y_{t+1})\|^2.$$

Pela desigualdade triangular:

$$\|v_{t+1}\|^2 \leq \frac{1}{\rho^2 t^2} (\|y_1 - y^*\|^2 + \|y^* - y_{t+1}\|^2).$$

Substituindo a majoração de $\|y_{t+1} - y^*\|^2$, que temos acima, e majorando o termo $\|y_1 - y^*\|^2$ de acordo com (4.7), temos:

$$\|v_{t+1}\|^2 \leq \frac{2}{\rho^2 t^2} (\rho L_G D_{x^*}^2 + \chi \rho^2 \|K\|^2 D_{x^*}^2 + (1 - \chi) \rho^2 D_{x^*,K}^2 + 2D_y^{2*}).$$

Além disso, por (4.68), temos:

$$g(v_{t+1}; z^{t+1}) =_{\hat{y} \in Y}^{(4.24)} \sup Q(w^*, x^*, v_{t+1}; z^{t+1}) + \langle v_{t+1}, y \rangle \leq \frac{1}{2t} (L_G D_{x^*}^2 + \chi \rho \|K\|^2 D_{x^*}^2 + (1 - \chi) \rho D_{x^*,K}^2).$$

Usando as duas inequações acima para a Proposição 4.1 e substituindo η , temos imediatamente:

$$\begin{aligned} G(x^{t+1}) + F(w^{t+1}) - f^* = g(v_{t+1}; z^{t+1}) &\leq \frac{1}{2t} ([L_G + \chi \rho \|K\|^2] D_{x^*}^2 + (1 - \chi) \rho D_{x^*,K}^2) \\ &= \frac{1}{2t} (\eta D_{x^*}^2 + \rho(1 - \chi) D_{x^*,K}^2). \end{aligned}$$

Provando assim (4.62). E, ainda, temos:

$$\begin{aligned} \|Bw^{t+1} - Kx^{t+1} - b\|^2 = \|v_{t+1}\|^2 &\leq \frac{2}{\rho^2 t^2} (\rho L_G D_{x^*}^2 + \chi \rho^2 \|K\|^2 D_{x^*}^2 + (1 - \chi) \rho^2 D_{x^*,K}^2 + 2D_y^{2*}) \\ &= \frac{2}{t^2} \left(\frac{2D_y^{2*}}{\rho^2} + \frac{L_G + \chi \rho \|K\|^2 D_{x^*}^2}{\rho} + (1 - \chi) D_{x^*,K}^2 \right). \end{aligned}$$

Substituindo η , finalmente, temos:

$$\|Bw^{t+1} - Kx^{t+1} - b\|^2 \leq \frac{2}{t^2} \left(\frac{2D_y^{2*}}{\rho^2} + \frac{\eta D_{x^*}^2}{\rho} + (1 - \chi) D_{x^*,K}^2 \right).$$

provando (4.63).

Para provar (4.64) basta substituir $\rho = 1$ e $\eta \geq L_G + \chi \rho \|K\|^2$ e para provar (4.65), além desta substituição, é necessário extrair a raiz quadrada em ambos os lados da desigualdade.

■

De acordo com Ouyang et al. [26] do Teorema 4.4, podemos concluir que o algoritmo ADMM, tem taxa de convergência de ambos os resíduos de viabilidade e primal de ordem $\mathcal{O}(\frac{1}{t})$. Esses

resultados nos permitem observar que um valor maior de ρ aumenta o valor do lado direito de (4.62), mas diminui o lado direito de (4.63). Essa situação implica que uma seleção ótima de ρ deve ser determinada considerando os resíduos de viabilidade e primal juntos. Por questão de simplicidade os autores definem $\rho = 1$.

No próximo teorema, mostramos que existe uma sequência de ponderações $\{\alpha_t\}_{t \geq 1}$ que melhora a taxa de convergência do Algoritmo 4.2, em termos de sua dependência de L_G .

Teorema 4.5 *Em AADMM se o número de iterações é definido como N e os parâmetros são ajustados para $\alpha_t = \frac{2}{t+1}$, $\theta_t = \tau_t = \frac{N}{t}$, $\rho_t = \frac{t}{N}$ e $\eta_t = \frac{2L_G + \chi N \|K\|^2}{t}$, então, temos:*

$$G(x_N^{ag}) + F(w_N^{ag}) - f^* \leq \frac{2L_G D_{x^*}^2}{N(N-1)} + \frac{1}{(N-1)} (\chi \|K\|^2 D_{x^*}^2 + (1-\chi) D_{x^*,K}^2) \quad (4.69)$$

e

$$\|Bw_N^{ag} - Kx_N^{ag} - b\| \leq \frac{4\sqrt{L_G} D_{x^*}}{(N-1)\sqrt{N}} + \frac{2\sqrt{2}\chi \|K\| D_{x^*}}{N-1} + \frac{(1-\chi)2\sqrt{2} D_{x^*,K}}{N-1} + \frac{4D_{y^*}}{N-1}. \quad (4.70)$$

Dem: Sabemos que $\alpha_t = \frac{2}{t+1}$, $\Gamma_t = \frac{2}{t(t+1)}$ satisfazem (4.26). E, ainda, que $\frac{\alpha_t}{\Gamma_t} = \frac{\frac{2}{t+1}}{\frac{2}{t(t+1)}} = t$

Agora pelos parâmetros escolhidos no enunciado e pela definição de $\mathcal{B}_t(\cdot, \cdot, \cdot)$, em (4.28), podemos obter os seguintes resultados:

$$(I) : \eta_t - L_G \alpha_t - \chi \theta_t \|K\|^2 \geq 0.$$

De fato, substituindo η_t , α_t e θ_t , temos:

$$\eta_t - L_G \alpha_t - \chi \theta_t \|K\|^2 = \frac{2L_G + \chi \rho N \|K\|^2}{t} - \frac{2L_G}{t+1} - \chi \frac{\rho N}{t} \|K\|^2.$$

Reorganizando os termos, temos:

$$\eta_t - L_G \alpha_t - \chi \theta_t \|K\|^2 = \frac{2L_G}{t} - \frac{2L_G}{t+1} + \frac{\chi \rho N \|K\|^2}{t} - \frac{\chi \rho N}{t} \|K\|^2 = \frac{2L_G}{t(t+1)} \geq 0.$$

pois todos os termos são positivos.

$$(II) : \mathcal{B}_t(x^*, x_{[t+1]}, \eta_t) \leq \frac{2L_G + \chi \rho N \|K\|^2}{2} (D_{x^*}^2 - \|x_{t+1} - x^*\|^2).$$

De fato, pela definição de $\mathcal{B}_t(\cdot, \cdot, \cdot)$ em (4.28), temos:

$$\mathcal{B}_t(x^*, x_{[t+1]}, \eta_t) = \sum_{i=1}^t \frac{\alpha_i \eta_i}{2\Gamma_i} (\|x_i - x^*\|^2 - \|x_{i+1} - x^*\|^2).$$

Substituindo η_i , α_i e Γ_i , temos:

$$\mathcal{B}_t(x^*, x_{[t+1]}, \eta_t) = \sum_{i=1}^t \frac{i}{2} \left(\frac{2L_G + \chi\rho N \|K\|^2}{i} \right) (\|x_i - x^*\|^2 - \|x_{i+1} - x^*\|^2).$$

Simplificando os termos, temos:

$$\mathcal{B}_t(x^*, x_{[t+1]}, \eta_t) = \frac{2L_G + \chi\rho N \|K\|^2}{2} \sum_{i=1}^t (\|x_i - x^*\|^2 - \|x_{i+1} - x^*\|^2).$$

E os somatórios, temos:

$$\mathcal{B}_t(x^*, x_{[t+1]}, \eta_t) = \frac{2L_G + \chi\rho N \|K\|^2}{2} (\|x_1 - x^*\|^2 - \|x_2 - x^*\|^2 + \|x_2 - x^*\|^2 - \dots + \|x_t - x^*\|^2 - \|x_{t+1} - x^*\|^2).$$

Simplificando, temos:

$$\mathcal{B}_t(x^*, x_{[t+1]}, \eta_t) = \frac{2L_G + \chi\rho N \|K\|^2}{2} (\|x_1 - x^*\|^2 - \|x_{t+1} - x^*\|^2).$$

e, por (4.7), temos:

$$\mathcal{B}_t(x^*, x_{[t+1]}, \eta_t) \leq \frac{2L_G + \chi\rho N \|K\|^2}{2} (D_{x^*}^2 - \|x_{t+1} - x^*\|^2).$$

$$(III) : \mathcal{B}_t(w^*, w_{[t+1]}, \theta_t) \leq \frac{ND_{x^*, K}^2}{2}.$$

De fato, pela definição de $\mathcal{B}_t(\cdot, \cdot, \cdot)$, em (4.28), temos:

$$\mathcal{B}_t(w^*, w_{[t+1]}, \theta_t) = \sum_{i=1}^t \frac{\alpha_i \theta_i}{2\Gamma_i} (\|w_i - w^*\|^2 - \|w_{i+1} - w^*\|^2).$$

Substituindo θ_i , α_i e Γ_i , temos:

$$\mathcal{B}_t(w^*, w_{[t+1]}, \theta_t) = \sum_{i=1}^t \frac{i}{2} \frac{N}{i} (\|w_i - w^*\|^2 - \|w_{i+1} - w^*\|^2).$$

Simplificando os termos, temos:

$$\mathcal{B}_t(w^*, w_{[t+1]}, \theta_t) = \frac{N}{2} \sum_{i=1}^t (\|w_i - w^*\|^2 - \|w_{i+1} - w^*\|^2).$$

Desenvolvendo os somatórios, temos:

$$\mathcal{B}_t(w^*, w_{[t+1]}, \theta_t) = \frac{N}{2} (\|w_1 - w^*\|^2 - \|w_2 - w^*\|^2 + \|w_2 - w^*\|^2 - \|w_3 - w^*\|^2 + \dots + \|w_t - w^*\|^2 - \|w_{t+1} - w^*\|^2).$$

Simplificando, temos:

$$\begin{aligned}\mathcal{B}_t(w^*, w_{[t+1]}, \theta_t) &= \frac{N}{2} (\|w_1 - w^*\|^2 - \|w_{t+1} - w^*\|^2) \\ &\leq \frac{N}{2} \|w_1 - w^*\|^2 \stackrel{(1)}{=} \frac{N}{2} \|Kx_1 - Kx^*\|^2.\end{aligned}$$

E, finalmente, por (4.7), temos:

$$\mathcal{B}_t(w^*, w_{[t+1]}, \theta_t) \leq \frac{ND_{x^*, K}^2}{2}.$$

$$(IV) : -\chi \mathcal{B}_t(Kx^*, Kx_{[t+1]}, \theta_t) \leq -\frac{\chi N}{2} (D_{x^*, K}^2 + \|K\|^2 \|x_{t+1} - x^*\|^2).$$

De fato, pela definição de $\mathcal{B}_t(\cdot, \cdot, \cdot)$, em (4.28), temos:

$$-\chi \mathcal{B}_t(Kx^*, Kx_{[t+1]}, \theta_t) = -\chi \mathcal{B}_t(Kx^*, Kx_{[t+1]}, \theta_t) = -\chi \left[\sum_{i=1}^t \frac{\alpha_i \theta_i}{2\Gamma_i} (\|Kx_i - Kx^*\|^2 - \|Kx_{i+1} - Kx^*\|^2) \right].$$

Substituindo θ_i , α_i e Γ_i , temos:

$$-\chi \mathcal{B}_t(Kx^*, Kx_{[t+1]}, \theta_t) = -\chi \left[\sum_{i=1}^t \frac{i}{2} \frac{N}{i} (\|Kx_i - Kx^*\|^2 - \|Kx_{i+1} - Kx^*\|^2) \right].$$

Simplificando os termos, temos:

$$-\chi \mathcal{B}_t(Kx^*, Kx_{[t+1]}, \theta_t) = -\chi \frac{N}{2} \sum_{i=1}^t (\|Kx_i - Kx^*\|^2 - \|Kx_{i+1} - Kx^*\|^2).$$

E desenvolvendo os somatórios, temos:

$$-\chi \mathcal{B}_t(Kx^*, Kx_{[t+1]}, \theta_t) = -\chi \frac{N}{2} (\|Kx_1 - Kx^*\|^2 - \|Kx_2 - Kx^*\|^2 + \dots + \|Kx_t - Kx^*\|^2 - \|Kx_{t+1} - Kx^*\|^2).$$

Simplificando, temos:

$$-\chi \mathcal{B}_t(Kx^*, Kx_{[t+1]}, \theta_t) = -\chi \frac{N}{2} (\|Kx_1 - Kx^*\|^2 - \|Kx_{t+1} - Kx^*\|^2).$$

e, por (4.7), temos:

$$-\chi \mathcal{B}_t(Kx^*, Kx_{[t+1]}, \theta_t) \leq -\frac{\chi N}{2} (D_{x^*, K}^2 \|K\|^2 \|x_{t+1} - x^*\|^2).$$

$$(V) : \mathcal{B}_t(y, y_{[t+1]}, \rho_t^{-1}) \leq \frac{N}{2} (\|y_1 - y\|^2 - \|y_{t+1} - y\|^2), \quad \forall y \in Y.$$

De fato, pela definição de $\mathcal{B}_t(\cdot, \cdot, \cdot)$, em (4.28), temos:

$$\mathcal{B}_t(y, y_{[t+1]}, \rho_t^{-1}) = \sum_{i=1}^t \frac{\alpha_i \rho_t^{-1}}{2\Gamma_i} (\|y_i - y\|^2 - \|y_{i+1} - y\|^2).$$

Substituindo ρ_t^{-1} , α_i e Γ_i , temos:

$$\mathcal{B}_t(y, y_{[t+1]}, \rho_t^{-1}) = \sum_{i=1}^t \frac{i}{2} \frac{N}{i} (\|y_i - y\|^2 - \|y_{i+1} - y\|^2).$$

Simplificando os termos, temos:

$$\mathcal{B}_t(y, y_{[t+1]}, \rho_t^{-1}) = \frac{N}{2} \sum_{i=1}^t (\|y_i - y\|^2 - \|y_{i+1} - y\|^2).$$

E desenvolvendo os somatórios, temos:

$$\mathcal{B}_t(y, y_{[t+1]}, \rho_t^{-1}) = \frac{N}{2} (\|y_1 - y\|^2 - \|y_2 - y\|^2 + \|y_2 - y\|^2 - \|y_3 - y\|^2 + \cdots + \|y_t - y\|^2 - \|y_{t+1} - y\|^2).$$

$$\mathcal{B}_t(y, y_{[t+1]}, \rho_t^{-1}) = \frac{N}{2} (\|y_1 - y\|^2 - \|y_{t+1} - y\|^2), \quad \forall y \in Y.$$

$$(IV) : - \sum_{i=1}^t \frac{\alpha_i(\tau_i - \rho_i)}{2\Gamma_i \rho_i} \|y_i - y_{i+1}\|^2 \leq 0.$$

Segue de $\tau_t = \frac{N}{t} \geq \frac{t}{N} = \rho_t$ para todo $t \geq N$.

Por (4.27), temos:

$$\begin{aligned} & \frac{1}{\Gamma_t} Q(w^*, x^*, y^*; z_{t+1}^{ag}) - \sum_{i=2}^t \left(\frac{1 - \alpha_i}{\Gamma_i} - \frac{1}{\Gamma_{i-1}} \right) Q(w^*, x^*, y^*; z_i^{ag}) \\ & \leq \mathcal{B}_t(x^*, x_{[t+1]}, \eta_{[t]}) + \mathcal{B}_t(y, y_{[t+1]}, \rho_{[t]}^{-1}) + \mathcal{B}_t(Bw^*, Bw_{[t+1]}, \theta_{[t]}) - \chi \mathcal{B}_t(Kx^*, Kx_{[t+1]}, \theta_{[t]}) \\ & \quad - \sum_{i=1}^t \frac{\alpha_i(\tau_i - \theta_i)}{2\Gamma_i} \|Bw_{i+1} - Kx^* - b\|^2 + \sum_{i=1}^t \frac{\alpha_i(\tau_i - \theta_i)}{2\Gamma_i} \|B(x_{i+1} - x^*)\|^2 \\ & \quad - \sum_{i=1}^t \frac{\alpha_i(\tau_i - \rho_i)}{2\Gamma_i \rho_i} \|y_i - y_{i+1}\|^2 - \sum_{i=1}^t \frac{\alpha_i}{2\Gamma_i} (\eta_i - L_G \alpha_i - \chi \theta_i \|K\|^2) \|x_i - x_{i+1}\|^2. \end{aligned}$$

Substituindo os parâmetros determinados, temos:

$$- \sum_{i=2}^t \left(\frac{1 - \alpha_i}{\Gamma_i} - \frac{1}{\Gamma_{i-1}} \right) Q(w^*, x^*, y^*; z_i^{ag}) = 0,$$

pois temos:

$$\left(\frac{1 - \frac{2}{i+1}}{\frac{2}{i(i-1)}} - \frac{1}{\frac{2}{(i-1)i}} \right) = \left(\frac{(1-i)(i+1)i}{(i+1)2} - \frac{(i-1)i}{2} \right) = 0$$

O fato de $\tau_t = \theta_t$, resulta em:

$$\begin{aligned} - \sum_{i=1}^t \frac{\alpha_i(\tau_i - \theta_i)}{2\Gamma_i} \|Bw_{i+1} - Kx^* - b\|^2 &= 0, \\ \sum_{i=1}^t \frac{\alpha_i(\tau_i - \theta_i)}{2\Gamma_i} \|B(x_{i+1} - x^*)\|^2 &= 0. \end{aligned}$$

Logo,

$$\begin{aligned} \frac{1}{\Gamma_t} Q(w^*, x^*, y; z_{t+1}^{ag}) &\leq \frac{2L_G + \chi N \|K\|^2}{2} (D_{x^*}^2 - \|x_{t+1} - x^*\|^2) + \frac{N}{2} (\|y_1 - y\|^2 - \|y_{t+1} - y\|^2) \\ &+ \frac{N}{2} D_{x^*,K}^2 - \frac{\chi N}{2} (D_{x^*,K}^2 - \|K\|^2 \|x_{t+1} - x^*\|^2) - \sum_{i=1}^t \frac{\alpha_i(\tau_i - \rho_i)}{2\Gamma_i \rho_i} \|y_i - y_{i+1}\|^2 \\ &- \sum_{i=1}^t \frac{\alpha_i}{2\Gamma_i} (\eta_i - L_G \alpha_i - \chi \theta_i \|K\|^2) \|x_i - x_{i+1}\|^2. \end{aligned}$$

Por I, temos: $-\sum_{i=1}^t \frac{\alpha_i}{2\Gamma_i} (\eta_i - L_G \alpha_i - \chi \theta_i \|K\|^2) \|x_i - x_{i+1}\|^2 \leq 0$; por IV temos $-\sum_{i=1}^t \frac{\alpha_i(\tau_i - \rho_i)}{2\Gamma_i \rho_i} \|y_i - y_{i+1}\|^2 \leq 0$. Além disso, $-\frac{L_G}{2} \|x_{t+1} - x^*\|^2 \leq 0$. Simplificando e desprezando esses termos negativos, temos:

$$\frac{1}{\Gamma_t} Q(w^*, x^*, y; z_{t+1}^{ag}) \leq L_G D_{x^*}^2 + \frac{\chi N \|K\|^2}{2} D_{x^*}^2 + \frac{(1 - \chi) N D_{x^*,K}^2}{2} + \frac{N}{2} (\|y_1 - y\|^2 - \|y_{t+1} - y\|^2)$$

Duas consequências podem ser deduzidas da estimativa acima. Em primeiro lugar, uma vez que $Q(z^*, z_{t+1}^{ag}) \geq 0$, temos:

$$0 \leq L_G D_{x^*}^2 + \frac{\chi N \|K\|^2}{2} D_{x^*}^2 + \frac{(1 - \chi) N D_{x^*,K}^2}{2} + \frac{N}{2} (\|y_1 - y\|^2 - \|y_{t+1} - y\|^2).$$

Multiplicando por $\frac{2}{N}$ e somando $\|y_{t+1} - y^*\|^2$ em ambos os lados da desigualdade, temos:

$$\|y_{t+1} - y^*\|^2 \leq \frac{2L_G D_{x^*}^2}{N} + \chi \|K\|^2 D_{x^*}^2 + (1 - \chi) D_{x^*,K}^2 + \|y_1 - y^*\|^2.$$

Por (4.7), temos:

$$\|y_{t+1} - y^*\|^2 \leq \frac{2L_G D_{x^*}^2}{N} + \chi \|K\|^2 D_{x^*}^2 + (1 - \chi) D_{x^*,K}^2 + D_{y^*}^2$$

e

$$\|y_1 - y_{t+1}\|^2 = \|y_1 - y^* + y^* - y_{t+1}\|^2.$$

Pela desigualdade triangular:

$$\|y_{t+1} - y^*\|^2 \leq \|y_1 - y^*\|^2 + \|y_{t+1} - y^*\|^2 \leq 2 (\|y_1 - y^*\|^2 + \|y_{t+1} - y^*\|^2) = 2\|y_1 - y^*\|^2 + 2\|y_{t+1} - y^*\|^2.$$

Substituindo a majoração de $\|y_{t+1} - y^*\|^2$ que temos acima e majorando o termo $\|y_1 - y^*\|^2$ de acordo com (4.7), temos:

$$\|y_1 - y_{t+1}\|^2 \leq \frac{4L_G D_{x^*}^2}{N} + 2\chi\|K\|^2 D_{x^*}^2 + 2(1-\chi)D_{x^*,K}^2 + 4D_{y^*}^2.$$

Em segundo lugar, uma vez que:

$$\|y_1 - y\|^2 - \|y_{t+1} - y\|^2 = \|y_1\|^2 - \|y_{t+1}\|^2 - 2\langle y_1 - y_{t+1}, y \rangle \stackrel{(4.67)}{\leq} -2\langle y_1 - y_{t+1}, y \rangle.$$

Substituindo essa majoração acima, temos:

$$\frac{1}{\Gamma_t} Q(w^*, x^*, y; z_{t+1}^{ag}) \leq L_G D_{x^*}^2 + \frac{\chi N \|K\|^2}{2} D_{x^*}^2 + \frac{(1-\chi)N D_{x^*,K}^2}{2} - \frac{N}{2} 2\langle y_1 - y_{t+1}, y \rangle.$$

Resultando em:

$$\frac{1}{\Gamma_t} Q(w^*, x^*, y; z_{t+1}^{ag}) + N\langle y_1 - y_{t+1}, y \rangle \leq L_G D_{x^*}^2 + \frac{\chi N \|K\|^2}{2} D_{x^*}^2 + \frac{(1-\chi)N D_{x^*,K}^2}{2}, \quad \forall y \in Y.$$

Tomando $t = N - 1$, substituindo Γ_t e definindo $v_n = \frac{2(y_1 - y_{t+1})}{N-1}$, temos:

$$\frac{(N-1)N}{2} Q(w^*, x^*, y; z_{t+1}^{ag}) + N\langle y_1 - y_{t+1}, y \rangle \leq L_G D_{x^*}^2 + \frac{\chi N \|K\|^2}{2} D_{x^*}^2 + \frac{(1-\chi)N D_{x^*,K}^2}{2}.$$

Multiplicando por $\frac{2}{(N-1)N}$, e usando a associatividade do produto interno, temos:

$$Q(w^*, x^*, y; z_{t+1}^{ag}) + \left\langle \frac{2(y_1 - y_{t+1})}{N-1}, y \right\rangle \leq \frac{2}{(N-1)N} L_G D_{x^*}^2 + \frac{1}{(N-1)} [\chi \|K\|^2 D_{x^*}^2 + (1-\chi)D_{x^*,K}^2], \quad \forall y \in Y.$$

Substituindo $\frac{2(y_1 - y_{t+1})}{N-1}$ por v_n , temos:

$$Q(w^*, x^*, y; z_{t+1}^{ag}) + \langle v_n, y \rangle \leq \frac{2}{(N-1)N} L_G D_{x^*}^2 + \frac{1}{(N-1)} [\chi \|K\|^2 D_{x^*}^2 + (1-\chi)D_{x^*,K}^2], \quad \forall y \in Y.$$

Pela Proposição 4.2, temos:

$$G(x_N^{ag}) + F(Kx_N^{ag}) - f^* \leq \frac{2}{(N-1)N} L_G D_{x^*}^2 + \frac{1}{(N-1)} [\chi \|K\|^2 D_{x^*}^2 + (1-\chi)D_{x^*,K}^2].$$

Provando (4.69).

E, mais ainda:

$$\left\| \frac{N-1}{2} \frac{2}{N-1} (y_1 - y_{t+1}) \right\|^2 = \|y_1 - y_{t+1}\|^2 \leq \frac{4L_G D_{x^*}^2}{N} + 2\chi\|K\|^2 D_{x^*}^2 + 2(1-\chi)D_{x^*,K}^2 + 4D_{y^*}^2.$$

Substituindo v_n e extraíndo a raíz quadrada em ambos os lados da desigualdade (é possível pois todos os termos são não negativos):

$$\left\| \frac{N-1}{2} v_n \right\| \leq \frac{2\sqrt{L_G} D_{x^*}}{\sqrt{N}} + \sqrt{2}\chi\|K\| D_{x^*} + \sqrt{2}(1-\chi)D_{x^*,K} + 2D_{y^*}.$$

Usando uma propriedade de norma, temos:

$$\left| \frac{N-1}{2} \right| \|v_n\| \leq \frac{2\sqrt{L_G}D_{x^*}}{\sqrt{N}} + \sqrt{2}\chi\|K\|D_{x^*} + \sqrt{2}(1-\chi)D_{x^*,K} + 2D_{y^*}^2.$$

Multiplicando ambos os lados da desigualdade por $\frac{2}{N-1}$, resulta em:

$$\|v_n\| \leq \frac{4\sqrt{L_G}D_{x^*}}{(N-1)\sqrt{N}} + \frac{2\sqrt{2}\chi\|K\|D_{x^*}}{N-1} + \frac{(1-\chi)2\sqrt{2}D_{x^*,K}}{N-1} + \frac{4D_{y^*}^2}{N-1}.$$

Provando assim (4.70), pois pelo Algoritmo 4.2 temos $v_n = Bw_N^{ag} - Kx_N^{ag} - b$.

■

Ao compararmos (4.64) e (4.65) com (4.69) e (4.70), respectivamente, percebemos que as taxas de convergência com relação aos resíduos de viabilidade e primal de AL-ADMM são melhores que as taxas de L-ADMM e LP-ADMM, conforme verificamos nas tabelas que seguem:

Tabela 4.2 Taxas de convergência do resíduo primal de casos AADMM para resolver problemas AECCO

	Não pré-condicionada ($\chi = 0$)	pré-condicionada ($\chi = 1$)
ADMM	$\mathcal{O}\left(\frac{D_{x^*,K}^2}{N}\right)$	$\mathcal{O}\left(\frac{\ K\ ^2 D_{x^*}^2}{N}\right)$
L-ADMM	$\mathcal{O}\left(\frac{L_G D_{x^*}^2 + D_{x^*,K}^2}{N}\right)$	$\mathcal{O}\left(\frac{L_G D_{x^*}^2 + \ K\ ^2 D_{x^*}^2}{N}\right)$
Acelerado (AADMM)	$\mathcal{O}\left(\frac{L_G D_{x^*}^2}{N^2} + \frac{D_{x^*,K}^2}{N}\right)$	$\mathcal{O}\left(\frac{L_G D_{x^*}^2}{N^2} + \frac{\ K\ ^2 D_{x^*}^2}{N}\right)$

Tabela 4.3 Taxas de convergência do resíduo de viabilidade de casos AADMM para resolver problemas AECCO

	Não pré-condicionada ($\chi = 0$)	pré-condicionada ($\chi = 1$)
ADMM	$\mathcal{O}\left(\frac{D_{x^*,K} + D_{y^*}}{N}\right)$	$\mathcal{O}\left(\frac{\ K\ D_{x^*} + D_{y^*}}{N}\right)$
L-ADMM	$\mathcal{O}\left(\frac{\sqrt{L_G}D_{x^*} + D_{x^*,K} + D_{y^*}}{N}\right)$	$\mathcal{O}\left(\frac{\sqrt{L_G}D_{x^*} + \ K\ D_{x^*} + D_{y^*}}{N}\right)$
Acelerado (AADMM)	$\mathcal{O}\left(\frac{\sqrt{L_G}D_{x^*}}{N^{\frac{3}{2}}} + \frac{D_{x^*,K} + D_{y^*}}{N}\right)$	$\mathcal{O}\left(\frac{\sqrt{L_G}D_{x^*}}{N^{\frac{3}{2}}} + \frac{\ K\ D_{x^*} + D_{y^*}}{N}\right)$

4.6 Um esquema de retrocesso

Nesta seção discutiremos sobre a taxa de convergência do Algoritmo 4.2, pressupondo que L_G e $\|K\|$ sejam dados. Na prática, pode ser necessário o uso de uma técnica de retrocesso para estimar ambas as constantes e aqui é proposto uma técnica de retrocesso para AL-ADMM e ALP-ADMM (OUYANG et al., 2015) [26]. A partir do Lema 4.1, se L_G e $\|K\|$ em (4.29) e (4.48) são substituídos por L_t e M_t , respectivamente, teremos:

$$G(x_{t+1}^{ag}) \leq G(x_t^{md}) + \langle \nabla G(x_t^{md}), x_{t+1}^{ag} - x_t^{md} \rangle + \frac{L_t}{2} \|x_{t+1}^{ag} - x_t^{md}\|^2. \quad (4.71)$$

e

$$\chi\|K(x_t - x_{t+1})\| \leq \chi M_t \|x_t - x_{t+1}\|. \quad (4.72)$$

O Lema 4.1 ainda se mantém e para provarmos os Teoremas 4.1, 4.2, 4.3, 4.4 e 4.5, além do Lema 4.1 precisamos da monotonicidade das seguintes sequências :

$$\frac{\alpha_{[t]}\eta_{[t]}}{\Gamma_{[t]}}, \frac{\alpha_{[t]}\tau_{[t]}}{\Gamma_{[t]}}, \frac{\alpha_{[t]}\theta_{[t]}}{\Gamma_{[t]}}, \frac{\alpha_{[t]}}{\Gamma_{[t]}\rho_{[t]}}$$

bem como

$$\eta_t - L_t \alpha_t - \chi \theta_t M_t^2 \geq 0. \quad (4.73)$$

No Lema 4.2 também utilizamos a monotonicidade dessas sequências para mostrar a majoração dos fatores a direita de $\mathcal{B}(\cdot, \cdot, \cdot)$, em (4.27). Essas observações nos permitem usar as seguintes escolhas de parâmetros:

$$\theta_t = \tau_t = \frac{v_t \alpha_t}{\Gamma_t}, \rho_t = \frac{\alpha_t}{\rho_t \Gamma_t}, \eta_t = L_t \alpha_t + \chi \theta_t M_t^2,$$

em que assumimos v_t e M_t monótonas. Verificamos que a monotonicidade de $\frac{\alpha_t \eta_t}{\Gamma_t}$ depende de $\left\{ \frac{L_t \alpha_t^2}{\Gamma_t} \right\}_{t \geq 1}$ que se torna trivial se definirmos $L_t \alpha_t^2 = \Gamma_t$. Considerando os fatores a direita de (4.27), é necessário que $\tau_t \geq \rho_t$, isto é, $v_t \geq \frac{\alpha_t}{\Gamma_t}$. O procedimento a seguir resume essas considerações:

Procedimento 4.1 *Procedimento de retrocesso para AL-ADMM e ALP-ADMM na iteração t .*

0: *Procedimento de retrocesso* $(L_{t-1}, M_{t-1}, \Gamma_{t-1}, v_{t-1}, x_t, x_t^{ag}, L_{min})$

1: $L_t \leftarrow \max \left\{ L_{min}, \frac{L_{t-1}}{2} \right\}$, $M_t = M_{t-1}$ e $v_t = v_{t-1}$.

2: *Estimar* $\alpha_t \in [0, 1]$ *resolvendo a equação quadrática*

$$L_t \alpha_t^2 = \Gamma_{t-1} (1 - \alpha_t) \quad (4.74)$$

e definir $\Gamma_t \leftarrow \Gamma_{t-1} (1 - \alpha_t)$, $v_t = \max \left\{ v_{t-1}, \frac{\alpha_t}{\Gamma_t} \right\}$

3: *Escolher os parâmetros como*

$$\theta_t = \tau_t = \frac{\rho v_t \Gamma_t}{\alpha_t}, \rho_t = \frac{\rho \alpha_t}{\Gamma_t v_t}, \eta_t = \frac{\Gamma_t}{\alpha_t} + \chi \theta_t M_t^2 \quad (4.75)$$

e calcular as iterações (4.15), (4.16) e (4.17) do Algoritmo 4.2.

4: *Se* $G(x_{t+1}^{ag}) - G(x_t^{md}) - \langle \nabla G(x_t^{md}), x_{t+1}^{ag} - x_t^{md} \rangle > \frac{L_t}{2} \|x_{t+1}^{ag} - x_t^{md}\|^2$, *então:*

5: *Definir* $L_t \leftarrow 2L_t$ *e voltar ao passo 2. (retrocesso de L_G)*

6: *Mas se* $\chi \|K(x_t - x_{t+1})\| > \chi M_t \|x_t - x_{t+1}\|$, *então:*

7: *Definir* $M_t \leftarrow 2M_t$ *e voltar ao passo 2. (retrocesso de $\|K\|$)*

8: *O retrocesso acaba aqui.*

9: *Retornando* $L_t, M_t, \Gamma_t, v_t, x_{t+1}, x_{t+1}^{ag}, \tau_t, \rho_t, \alpha_t$

10: *Final do procedimento*

Sobre este procedimento, Ouyang et al. [26] fazem algumas observações:

1) As etapas de 2 a 8 são etapas de retrocesso, que terminam apenas quando as etapas 4 e 6 são ambas satisfeitas. Em cada tentativa para o retrocesso, os passos 4 e 6 só são realizados um número finito de vezes, e os valores retornados L_t e M_t satisfazem $L_{min} \leq L_t \leq 2L_G$ e $M_t \leq 2\|K\|$.

2) Enquanto $M_t \geq M_{t-1}$ e $v_t \geq v_{t-1}$, o valor de L_t no passo 9 não é necessariamente maior que

L_{t-1} .

3) O multiplicador para aumentar ou diminuir L_t e M_t é 2, que pode ser substituído por qualquer número maior que 1.

O algoritmo a seguir apresenta esse esquema de retrocesso.

Algoritmo 4.3 AADMM com retrocesso.

1. Escolha $x_1 \in X$ e $w_1 \in W$ tal que $Bw_1 = Kx_1 - b$, $L_0 \geq L_{min} > 0$ e $M_0, v_0, \rho > 0$. Faça $x_1^{ag} \leftarrow x_1$, $w_1^{ag} \leftarrow w_1$, $y_1^{ag} \leftarrow y_1 = 0$, $\Gamma_0 \leftarrow L_0$, $t \leftarrow 1$.

2. Para todo $t = 1, \dots, N - 1$ faça

$$(L_t, M_t, \Gamma_t, v_t, x_{t+1}, x_{t+1}^{ag}, \tau_t, \rho_t, \alpha_t) \leftarrow (L_{t-1}, M_{t-1}, \Gamma_{t-1}, v_{t-1}, x_t, x_t^{ag}, L_{min}) \quad (4.76)$$

calcular as iterações (4.18), (4.19), (4.20) e (4.21).

Consideramos problemas UCO com conjuntos viáveis X e Y limitados. O Teorema 4.7 resume as propriedades de convergência do Algoritmo 4.3 para resolver problemas UCO limitados.

Teorema 4.6 Se colocarmos $v_0 = -\infty$ e aplicarmos o Algoritmo 4.3 para o Problema UCO 4.5 sob hipóteses de (4.4), então:

$$G(x_{t+1}^{ag}) + F(Kx_{t+1}^{ag}) - f^* \leq \frac{4L_G D_X^2}{t^2} + \frac{4L_G}{L_{min}(t-1)} \left[6\chi \rho \max\{4M_0^2, \|K\|^2\} D_X^2 + (1-\chi)\rho D_{X,K}^2 + \frac{D_Y^2}{\rho} \right] \quad (4.77)$$

Em particular se $\rho = \frac{D_Y}{\sqrt{6\chi \max\{2M_0\|K\|\} D_X + (1-\chi) D_{X,K}}}$, então:

$$G(x_{t+1}^{ag}) + F(Kx_{t+1}^{ag}) - f^* \leq \frac{4L_G D_X^2}{t^2} + \frac{4L_G}{L_{min}(t-1)} \left[2\sqrt{6}\chi \max\{2M_0\|K\|\} D_X D_Y + 2(1-\chi) D_{X,K} D_Y \right]. \quad (4.78)$$

Dem: Como discutido no procedimento, temos:

$$L_{min} \leq L_t \leq 2L_G; 0 \leq M_t \leq 2\|K\| \quad (4.79)$$

Agora podemos estimar os limites de α_t e Γ_t . Por (4.26), temos $\frac{1}{\Gamma_t} = \frac{1}{\Gamma_{t-1}} + \frac{\alpha_t}{\Gamma_t}$. De fato, de $\Gamma_t = (1 - \alpha_t)\Gamma_{t-1}$, temos:

$$\frac{(1 - \alpha_t)}{\Gamma_t} = \frac{1}{\Gamma_{t-1}} \Rightarrow \frac{1}{\Gamma_t} - \frac{\alpha_t}{\Gamma_t} = \frac{1}{\Gamma_{t-1}} \Rightarrow \frac{1}{\Gamma_t} = \frac{1}{\Gamma_{t-1}} + \frac{\alpha_t}{\Gamma_t}.$$

Portanto, temos:

$$\sqrt{\frac{1}{\Gamma_t}} - \sqrt{\frac{1}{\Gamma_{t-1}}} = \left(\sqrt{\frac{1}{\Gamma_t}} - \sqrt{\frac{1}{\Gamma_{t-1}}} \right) \cdot \frac{\left(\sqrt{\frac{1}{\Gamma_t}} + \sqrt{\frac{1}{\Gamma_{t-1}}} \right)}{\left(\sqrt{\frac{1}{\Gamma_t}} + \sqrt{\frac{1}{\Gamma_{t-1}}} \right)} = \frac{\frac{1}{\Gamma_t} - \frac{1}{\Gamma_{t-1}}}{\sqrt{\frac{1}{\Gamma_t}} + \sqrt{\frac{1}{\Gamma_{t-1}}}} = \frac{\frac{\alpha_t}{\Gamma_t}}{\sqrt{\frac{1}{\Gamma_t}} + \sqrt{\frac{1}{\Gamma_{t-1}}}}.$$

Por (4.79), temos $L_{min} \leq L_t \leq 2L_G \Rightarrow \frac{1}{2L_G} \leq \frac{1}{L_t} \leq \frac{1}{L_{min}}(1)$.

Por (4.74), temos $L_t \alpha_t^2 = \Gamma_{t-1}(1 - \alpha_t) \Rightarrow L_t = \frac{\Gamma_{t-1}(1 - \alpha_t)}{\alpha_t^2}(2)$.

Por (4.26), temos $\Gamma_t = (1 - \alpha_t)\Gamma_{t-1}(3)$. Daí substituindo $\Gamma_{t-1} = \frac{\Gamma_t}{(1 - \alpha_t)}$ em (2), e substituindo L_t em (1), temos:

$$\frac{1}{2L_G} \leq \frac{\alpha_t^2}{\Gamma_t} \leq \frac{1}{L_{min}}. \quad (4.80)$$

Assim, temos:

$$\sqrt{\frac{1}{\Gamma_t}} - \sqrt{\frac{1}{\Gamma_{t-1}}} = \frac{\frac{\alpha_t}{\Gamma_t}}{\sqrt{\frac{1}{\Gamma_t}} + \sqrt{\frac{1}{\Gamma_{t-1}}}} \geq_{(\Gamma_t \geq \Gamma_{t-1})} \frac{\frac{\alpha_t}{\Gamma_t}}{2\sqrt{\frac{1}{\Gamma_t}}} = \frac{\alpha_t}{2\sqrt{\Gamma_t}} \geq_{(4.80)} \frac{1}{2\sqrt{2L_G}}.$$

Por outro lado, temos:

$$\sqrt{\frac{1}{\Gamma_t}} - \sqrt{\frac{1}{\Gamma_{t-1}}} = \frac{\alpha_t}{\Gamma_t} \frac{1}{\sqrt{\frac{1}{\Gamma_t}} + \sqrt{\frac{1}{\Gamma_{t-1}}}} \leq_{(4.80)} \frac{\sqrt{\frac{1}{\Gamma_t} \frac{1}{L_{min}}}}{\sqrt{\frac{1}{\Gamma_t}} + \sqrt{\frac{1}{\Gamma_{t-1}}}} \leq \frac{\sqrt{\frac{1}{\Gamma_t} \frac{1}{L_{min}}}}{\sqrt{\frac{1}{\Gamma_t}}} = \sqrt{\frac{1}{L_{min}}}.$$

Portanto, por um processo de indução, de acordo com Ouyang et al. [26], concluímos que:

$$\frac{t}{2\sqrt{2L_G}} \leq \sqrt{\frac{1}{\Gamma_t}} \leq t \frac{1}{\sqrt{L_{min}}} + \frac{1}{\sqrt{L_0}} \leq \frac{t+1}{\sqrt{L_{min}}}.$$

Reorganizando essas desigualdades, obtemos:

$$\frac{t}{2\sqrt{2L_G}} \leq \frac{1}{\sqrt{\Gamma_t}} \leq \frac{t+1}{\sqrt{L_{min}}}.$$

Elevando ao quadrado e invertendo as frações, temos:

$$\frac{L_{min}}{(t+1)^2} \leq \Gamma_t \leq \frac{8L_G}{t^2}. \quad (4.81)$$

Agora vamos examinar os fatores a direita de (4.27).

Sem perda de generalidade, assumiremos que $2M_0 \leq \|K\|$ (pois se, $2M_0 > \|K\|$, então $M_t = 2M_0$).

Desde que $v_{[t]}$ e $M_{[t]}$ sejam monotonicamente crescentes, por (4.75), temos:

$$\theta_t = \tau_t = \frac{\rho v_t \Gamma_t}{\alpha_t}, \rho_t = \frac{\rho \alpha_t}{\Gamma_t v_t}, \eta_t = \frac{\Gamma_t}{\alpha_t} + \chi \theta_t M_t^2.$$

$$(I) : \mathcal{B}_t(x^*, x_{[t+1]}, \eta_{[t]}) \leq \frac{1}{2} D_X^2 + \chi v_t \rho \|K\|^2 D_X^2.$$

Como estamos nas condições de (4.51) no Lema 4.2, temos:

$$\mathcal{B}_t(x^*, x_{[t+1]}, \eta_{[t]}) \leq \frac{\alpha_t \eta_t}{2\Gamma_t} (D_X^2 - \|x_{t+1} - x\|^2).$$

Substituindo η_t , temos:

$$\mathcal{B}_t(x^*, x_{[t+1]}, \eta_{[t]}) \leq \frac{\alpha_t}{2\Gamma_t} \left(\frac{\Gamma_t}{\alpha_t} + \frac{\chi \rho v_t \Gamma_t M_t^2}{\alpha_t} \right) (D_X^2 - \|x_{t+1} - x\|^2).$$

Simplificando e desprezando o termo negativo $-\|x_{t+1} - x\|^2$, temos:

$$\mathcal{B}_t(x^*, x_{[t+1]}, \eta_{[t]}) \leq \frac{1 + \chi \rho v_t M_t^2}{2} D_X^2 \leq \frac{1 + \chi \rho v_t \|K\|^2}{2} D_X^2 \leq \frac{1}{2} D_X^2 + \chi v_t \rho \|K\|^2 D_X^2.$$

$$(II) : \mathcal{B}_t(w^*, w_{[t+1]}, \theta_{[t]}) \stackrel{(4.28)}{=} \sum_{i=1}^t \frac{\alpha_i}{2\Gamma_i} \theta_i (\|w_i - w^*\|^2 - \|w_{i+1} - w^*\|^2).$$

Substituindo $\theta_{[t]}$ por $\frac{\rho v_t \Gamma_t}{\alpha_t}$ e desenvolvendo o somatório, temos:

$$\mathcal{B}_t(w^*, w_{[t+1]}, \theta_{[t]}) = \frac{\rho v_t}{2} \|w_1 - w^*\|^2 - \frac{\rho v_t}{2} \|w_2 - w^*\|^2 + \frac{\rho v_t}{2} \|w_2 - w^*\|^2 - \frac{\rho v_t}{2} \|w_3 - w^*\|^2 + \dots + \frac{\rho v_t}{2} \|w_t - w^*\|^2 - \frac{\rho v_t}{2} \|w_{t+1} - w^*\|^2.$$

Simplificando, temos:

$$\mathcal{B}_t(w^*, w_{[t+1]}, \theta_{[t]}) = \frac{\rho v_t}{2} (\|w_1 - w^*\|^2 - \|w_{t+1} - w^*\|^2).$$

Como $-\|w_{t+1} - w^*\|^2 \leq 0$, temos:

$$\mathcal{B}_t(w^*, w_{[t+1]}, \theta_{[t]}) \frac{\rho v_t}{2} \|w_1 - w^*\|^2 = \frac{\rho v_t}{2} \|Kx_1 - Kx^*\|^2$$

Essa igualdade é dada pela hipótese do Algoritmo 4.2, onde temos $Bw_1 = Kx_1 - b$.

Por (4.7), temos:

$$\mathcal{B}_t(w^*, w_{[t+1]}, \theta_{[t]}) \leq \frac{\rho v_t}{2} D_{X,K}^2.$$

$$(III) : \mathcal{B}_t(y, y_{[t+1]}, \rho_{[t]}^{-1}) \stackrel{(4.28)}{=} \sum_{i=1}^t \frac{\alpha_i}{2\Gamma_i} \rho_i (\|y_i - y\|^2 - \|y_{i+1} - y\|^2).$$

Substituindo $\rho_{[t]}$ por $\frac{\rho \alpha_t}{\Gamma_t v_t}$, usando o fato de $v_{t \geq 1}$ ser monotonicamente crescente e de maneira análoga ao que foi feito em II, temos:

$$\mathcal{B}_t(y, y_{[t+1]}, \rho_{[t]}^{-1}) \leq \frac{v_t}{2\rho} (\|y_1 - y\|^2 - \|y_{t+1} - y\|^2).$$

Por (4.7), temos:

$$\mathcal{B}_t(y, y_{[t+1]}, \rho_{[t]}^{-1}) \leq \frac{v_t}{2\rho} (D_y^2 - \|Kx_{t+1} - Kx^*\|^2) = \frac{v_t}{2\rho} D_y^2 - \frac{v_t}{2\rho} \|y_{t+1} - y\|^2.$$

Como $-\frac{v_t}{2\rho} \|y_{t+1} - y\|^2 \leq 0$, temos:

$$\mathcal{B}_t(y, y_{[t+1]}, \rho_{[t]}^{-1}) \leq \frac{v_t}{2\rho} D_y^2, \quad \forall y \in Y.$$

$$(IV) : -\chi \mathcal{B}_t(Kx^*, Kx_{[t+1]}, \theta_{[t]}) \leq \frac{\chi v_t \rho \|K\|^2}{2} D_X^2.$$

Como estamos nas condições de (4.50) no lema 4.2, temos:

Se $\left\{ \frac{\alpha_i \theta_i}{\Gamma_i} \right\}$ é decrescente, então:

$$-\chi \mathcal{B}_t(Kx^*, Kx_{[t+1]}, \theta_{[t]}) \leq -\chi \left[\frac{\alpha_1 \theta_1}{2\Gamma_1} \|Kx_1 - Kx^*\|^2 - \frac{\alpha_t \theta_t}{2\Gamma_t} \|Kx_{t+1} - Kx^*\|^2 \right].$$

Substituindo θ_i por $\frac{\rho v_i \Gamma_i}{\alpha_i}$ e simplificando a expressão, temos:

$$-\chi \mathcal{B}_t(Kx^*, Kx_{[t+1]}, \theta_{[t]}) \leq -\chi \frac{\rho v_1}{2} \|Kx_1 - Kx^*\|^2 + \chi \frac{\rho v_t}{2} \|Kx_{t+1} - Kx^*\|^2.$$

Como $-\chi \frac{\rho v_1}{2} \|Kx_1 - Kx^*\|^2 \leq 0$, temos:

$$-\chi \mathcal{B}_t(Kx^*, Kx_{[t+1]}, \theta_{[t]}) \leq \chi \frac{\rho v_t}{2} \|Kx_{t+1} - Kx^*\|^2 \leq \chi \frac{\rho v_t}{2} \|K\|^2 \|x_{t+1} - x^*\|^2$$

e por (4.7), temos:

$$-\chi \mathcal{B}_t(Kx^*, Kx_{[t+1]}, \theta_{[t]}) \leq \chi \frac{\rho v_t}{2} \|K\|^2 D_X^2.$$

Por (4.27), temos:

$$\begin{aligned} & \frac{1}{\Gamma_t} Q(w^*, x^*, y^*; z_{t+1}^{ag}) - \sum_{i=2}^t \left(\frac{1 - \alpha_i}{\Gamma_i} - \frac{1}{\Gamma_{i-1}} \right) Q(w^*, x^*, y^*; z_i^{ag}) \\ & \leq \mathcal{B}_t(x^*, x_{[t+1]}, \eta_{[t]}) + \mathcal{B}_t(y, y_{[t+1]}, \rho_{[t]}^{-1}) + \mathcal{B}_t(Bw^*, Bw_{[t+1]}, \theta_{[t]}) - \chi \mathcal{B}_t(Kx^*, Kx_{[t+1]}, \theta_{[t]}) \\ & \quad - \sum_{i=1}^t \frac{\alpha_i (\tau_i - \theta_i)}{2\Gamma_i} \|Bw_{i+1} - Kx^* - b\|^2 + \sum_{i=1}^t \frac{\alpha_i (\tau_i - \theta_i)}{2\Gamma_i} \|B(x_{i+1} - x^*)\|^2 \\ & \quad - \sum_{i=1}^t \frac{\alpha_i (\tau_i - \rho_i)}{2\Gamma_i \rho_i} \|y_i - y_{i+1}\|^2 - \sum_{i=1}^t \frac{\alpha_i}{2\Gamma_i} (\eta_i - L_G \alpha_i - \chi \theta_i \|K\|^2) \|x_i - x_{i+1}\|^2. \end{aligned}$$

Substituindo os parâmetros determinados por (4.75), temos:

$$-\sum_{i=2}^t \left(\frac{1 - \alpha_i}{\Gamma_i} - \frac{1}{\Gamma_{i-1}} \right) Q(w^*, x^*, y^*; z_i^{ag}) = 0,$$

pois:

$$\left(\frac{1 - \frac{2}{i+1}}{\frac{2}{i(i-1)}} - \frac{1}{\frac{2}{(i-1)i}} \right) = \left(\frac{(1-i)(i+1)i}{(i+1)2} - \frac{(i-1)i}{2} \right) = 0,$$

(V): O fato de $\tau_t = \theta_t$ resulta em:

$$- \sum_{i=1}^t \frac{\alpha_i(\tau_i - \theta_i)}{2\Gamma_i} \|Bw_{i+1} - Kx^* - b\|^2 = 0,$$

$$\sum_{i=1}^t \frac{\alpha_i(\tau_i - \theta_i)}{2\Gamma_i} \|B(x_{i+1} - x^*)\|^2 = 0.$$

(VI): Como pressuposto antes do procedimento $\tau_t \geq \rho_t$, temos:

$$- \sum_{i=1}^t \frac{\alpha_i(\tau_i - \rho_i)}{2\Gamma_i \rho_i} \|y_i - y_{i+1}\|^2 \leq 0.$$

(VII): Pela equação (4.73), temos:

$$- \sum_{i=1}^t \frac{\alpha_i}{2\Gamma_i} (\eta_i - L_G \alpha_i - \chi \theta_i \|K\|^2) \|x_i - x_{i+1}\|^2 \leq 0.$$

substituindo I, II, III, IV, V e desprezando os termos negativos de acordo com VI e VII, temos:

$$\frac{1}{\Gamma_t} Q(w^*, x^*, y; z_i^{ag}) \leq \frac{1}{2} D_X^2 + \frac{1}{2} \chi \rho v_t M_t^2 D_X^2 + \frac{v_t}{2\rho} D_Y^2 + \frac{\rho v_t}{2} D_{X,K}^2 + \frac{\chi v_t \rho}{2} \|K\|^2 D_X^2.$$

Como $M_t \leq 2\|K\|$, temos $M_t^2 \leq 4\|K\|^2$, logo:

$$\begin{aligned} \frac{1}{\Gamma_t} Q(w^*, x^*, y; z_i^{ag}) &\leq \frac{1}{2} D_X^2 + \chi \rho v_t 2\|K\|^2 D_X^2 + \frac{v_t}{2\rho} D_Y^2 + \frac{\rho v_t}{2} D_{X,K}^2 + \frac{\chi v_t \rho}{2} \|K\|^2 D_X^2 \\ &= \frac{1}{2} D_X^2 + \frac{5}{2} \chi \rho v_t \|K\|^2 D_X^2 + \frac{v_t}{2\rho} D_Y^2 + \frac{\rho v_t}{2} D_{X,K}^2 + \frac{\chi \rho v_t}{2} D_{X,K}^2 - \frac{\chi \rho v_t}{2} D_{X,K}^2. \end{aligned}$$

Agrupando os termos e usando (4.7), temos:

$$\frac{1}{\Gamma_t} Q(w^*, x^*, y; z_i^{ag}) \leq \frac{1}{2} D_X^2 + 3\chi \rho v_t \|K\|^2 D_X^2 + \frac{v_t}{2\rho} D_Y^2 + \frac{(1-\chi)\rho v_t}{2} D_{X,K}^2. \quad (4.82)$$

Por (4.80) e (4.81), temos $\frac{\alpha_t}{\Gamma_t} \leq \frac{t+1}{L_{min}}$.

De fato, por (4.80), temos $\frac{\alpha_t^2}{\Gamma_t} \leq \frac{1}{L_{min}}$ e, por (4.81), temos $\frac{L_{min}}{(t+1)^2} \leq \frac{\Gamma_t}{t^2} \leq \Gamma_t \Rightarrow \frac{L_{min}}{(t+1)^2} \leq \Gamma_t$. Multiplicando $\frac{\alpha_t^2}{\Gamma_t}$ por $\frac{L_{min}}{(t+1)^2}$ e $\frac{1}{L_{min}}$ por Γ_t a desigualdade se mantém pois todos os termos são positivos e resulta em $\frac{L_{min}\alpha_t^2}{(t+1)^2\Gamma_t} \leq \frac{\Gamma_t}{L_{min}}$. Como todos os termos são positivos podemos reorganizar de modo a obter $\frac{\alpha_t^2}{\Gamma_t^2} \leq \frac{(t+1)^2}{L_{min}^2}$. Basta extrair a raiz quadrada em ambos os lados da desigualdade e fica provado o resultado.

Por (4.74), temos:

$$v_t \leq \max_{i=1, \dots, t} \frac{\alpha_i}{\Gamma_i} \leq \frac{t+1}{L_{min}}. \quad (4.83)$$

Usando as duas desigualdades acima e (4.81), temos:

$$\widehat{g}_y(z_{t+1}^{ag}) := \sup_{\widehat{y} \in Y} Q(w^*, x^*, \widehat{y}; z_{t+1}^{ag}) = \Gamma_t \frac{1}{\Gamma_t} \sup_{\widehat{y} \in Y} Q(w^*, x^*, \widehat{y}; z_{t+1}^{ag}).$$

Por (4.82), temos:

$$\frac{1}{\Gamma_t} \sup_{\widehat{y} \in Y} Q(w^*, x^*, \widehat{y}; z_{t+1}^{ag}) \leq \Gamma_t \sup_{\widehat{y} \in Y} \left\{ \frac{1}{2} D_X^2 + 3\chi \rho v_t \|K\|^2 D_X^2 + \frac{v_t}{2\rho} D_Y^2 + \frac{(1-\chi)\rho v_t}{2} D_{X,K}^2 \right\}.$$

Como nenhum dos termos depende de \widehat{y} , temos:

$$\frac{1}{\Gamma_t} \sup_{\widehat{y} \in Y} Q(w^*, x^*, \widehat{y}; z_{t+1}^{ag}) \leq \Gamma_t \left[\frac{1}{2} D_X^2 + 3\chi \rho v_t \|K\|^2 D_X^2 + \frac{v_t}{2\rho} D_Y^2 + \frac{(1-\chi)\rho v_t}{2} D_{X,K}^2 \right].$$

Por (4.81), temos:

$$\frac{1}{\Gamma_t} \sup_{\widehat{y} \in Y} Q(w^*, x^*, \widehat{y}; z_{t+1}^{ag}) \leq \frac{8L_G}{t^2} \left[\frac{1}{2} D_X^2 + 3\chi \rho v_t \|K\|^2 D_X^2 + \frac{v_t}{2\rho} D_Y^2 + \frac{(1-\chi)\rho v_t}{2} D_{X,K}^2 \right].$$

Simplificando os termos, temos:

$$\frac{1}{\Gamma_t} \sup_{\widehat{y} \in Y} Q(w^*, x^*, \widehat{y}; z_{t+1}^{ag}) \leq \frac{4L_G D_X^2}{t^2} + \frac{24\chi L_G \rho v_t \|K\|^2 D_X^2}{t^2} + \frac{4L_G v_t D_Y^2}{\rho t^2} + \frac{4L_G (1-\chi) \rho v_t D_{X,K}^2}{t^2}.$$

Como $v_t \leq \frac{t+1}{L_{min}}$, temos:

$$\frac{1}{\Gamma_t} \sup_{\widehat{y} \in Y} Q(w^*, x^*, \widehat{y}; z_{t+1}^{ag}) \leq \frac{4L_G D_X^2}{t^2} + \frac{24\chi L_G \rho (t+1) \|K\|^2 D_X^2}{t^2 L_{min}} + \frac{4L_G (t+1) D_Y^2}{\rho t^2 L_{min}} + \frac{4L_G (1-\chi) \rho (t+1) D_{X,K}^2}{t^2 L_{min}}.$$

Como $t \geq 0$, então $(t^2 - 1) \leq t^2$ e disso temos: $\frac{1}{t^2} \leq \frac{1}{(t^2-1)}$. Daí:

$$\frac{1}{\Gamma_t} \sup_{\widehat{y} \in Y} Q(w^*, x^*, \widehat{y}; z_{t+1}^{ag}) \leq \frac{4L_G D_X^2}{t^2} + \frac{24\chi L_G \rho (t+1) \|K\|^2 D_X^2}{(t^2 - 1) L_{min}} + \frac{4L_G (t+1) D_Y^2}{\rho (t^2 - 1) L_{min}} + \frac{4L_G (1-\chi) \rho (t+1) D_{X,K}^2}{(t^2 - 1) L_{min}}.$$

Simplificando, temos:

$$\frac{1}{\Gamma_t} \sup_{\widehat{y} \in Y} Q(w^*, x^*, \widehat{y}; z_{t+1}^{ag}) \leq \frac{4L_G D_X^2}{t^2} + \frac{24\chi L_G \rho \|K\|^2 D_X^2}{(t-1) L_{min}} + \frac{4L_G D_Y^2}{\rho (t-1) L_{min}} + \frac{4L_G (1-\chi) \rho D_{X,K}^2}{(t-1) L_{min}}.$$

Pela Proposição 4.2, temos: $\widehat{w} = (B^*)^{-1}(Kx + b)$. Como B é o operador identidade e $b = 0$ temos $\widehat{w} = Kx$, daí:

$$F(w_{t+1}^{ag}) = F(Kx_{t+1}^{ag}) \quad (4.84)$$

e, consequentemente:

$$G(x_{t+1}^{ag}) + F(Kx_{t+1}^{ag}) - f^* \leq \frac{1}{\Gamma_t} \sup_{y \in Y} Q(w^*, x^*, y; z_{t+1}^{ag}).$$

Logo, temos:

$$G(x_{t+1}^{ag}) + F(Kx_{t+1}^{ag}) - f^* \leq \frac{4L_G D_X^2}{t^2} + \frac{24\chi L_G \rho \|K\|^2 D_X^2}{(t-1)L_{min}} + \frac{4L_G D_Y^2}{\rho(t-1)L_{min}} + \frac{4L_G(1-\chi)\rho D_{X,K}^2}{(t-1)L_{min}}.$$

Colocando $\frac{4L_G}{L_{min}(t-1)}$ em evidência, temos:

$$G(x_{t+1}^{ag}) + F(Kx_{t+1}^{ag}) - f^* \leq \frac{4L_G D_X^2}{t^2} + \frac{4L_G}{L_{min}(t-1)} \left[6\chi \rho \|K\|^2 D_X^2 + \frac{D_Y^2}{\rho} + (1-\chi)\rho D_{X,K}^2 \right].$$

Como $2M_0 \leq \|K\|$, podemos reescrever da seguinte maneira:

$$G(x_{t+1}^{ag}) + F(Kx_{t+1}^{ag}) - f^* \leq \frac{4L_G D_X^2}{t^2} + \frac{4L_G}{L_{min}(t-1)} \left[6\chi \rho \max\{4M_0^2, \|K\|^2\} D_X^2 + (1-\chi)\rho D_{X,K}^2 + \frac{D_Y^2}{\rho} \right]$$

e fica provado (4.77).

Agora, substituindo $\rho = \frac{D_Y}{\sqrt{6}\chi \max\{2M_0\|K\|\} D_X + (1-\chi)D_{X,K}}$, temos:

$$\begin{aligned} G(x_{t+1}^{ag}) + F(Kx_{t+1}^{ag}) - f^* &\leq \frac{4L_G D_X^2}{t^2} + \frac{4L_G}{L_{min}(t-1)} \frac{6\chi D_Y \max\{4M_0^2, \|K\|^2\} D_X^2}{\sqrt{6}\chi \max\{2M_0\|K\|\} D_X + (1-\chi)D_{X,K}} \\ &\quad + \frac{4L_G}{L_{min}(t-1)} \frac{(1-\chi)D_Y D_{X,K}^2}{\sqrt{6}\chi \max\{2M_0\|K\|\} D_X + (1-\chi)D_{X,K}} \\ &\quad + \frac{4L_G}{L_{min}(t-1)} \frac{D_Y^2 (\sqrt{6}\chi \max\{2M_0\|K\|\} D_X + (1-\chi)D_{X,K})}{D_Y}. \end{aligned}$$

Usando o fato de que $\chi \neq 0$ se $(1-\chi) = 0$ e simplificando as expressões, temos:

$$G(x_{t+1}^{ag}) + F(Kx_{t+1}^{ag}) - f^* \leq \frac{4L_G D_X^2}{t^2} + \frac{4L_G}{L_{min}(t-1)} \left[2\sqrt{6}\chi \max\{2M_0\|K\|\} D_X D_Y + 2(1-\chi)D_{X,K} D_Y \right].$$

e fica provado (4.78). ■

De acordo com Ouyang et al. [26], para problemas AECCO quando X e Y são limitados, também podemos aplicar o Algoritmo 4.3 com $\chi = 0$, contanto que o número de iterações N seja dado. O Teorema 2.2 abaixo descreve as propriedades de convergência de AL-ADMM com

retrocesso para a solução de problemas AECCO gerais.

Teorema 4.7 *Se escolhermos $\chi = 0$ o número de iterações é definido como N e os parâmetros são ajustados para $\rho = 1$ e $v_0 = \frac{N}{L_{min}}$ no Algoritmo 4.3 então, temos:*

$$G(x_N^{ag}) + F(w_N^{t+1}) - f^* \leq \frac{4L_G D_{x^*}^2}{(N-1)^2} + \frac{4L_G D_{x^*}^2}{L_{min}(N-1)} \quad (4.85)$$

e

$$\|Bw_N^{ag} - K_N^{ag} - b\| \leq \frac{16\sqrt{L_G} D_{x^*}}{(N-1)^{\frac{3}{2}} \sqrt{L_{min}}} + \frac{16\sqrt{2}\sqrt{L_G} D_{x^*,K}}{(N-1)\sqrt{L_{min}}} + \frac{32L_G D_{y^*}}{(N-1)L_{min}}. \quad (4.86)$$

Dem: *Em vista do passo 2 no procedimento, a equação (4.83), temos: $v_t = \frac{N}{L_{min}}$.*

De fato,

$v_t \leq \max_{i=1, \dots, t} \frac{\alpha_i}{\Gamma_i} \leq \frac{t+1}{L_{min}}$, $v_t = \max \left\{ v_{t-1}, \frac{\alpha_t}{\Gamma_t} \right\}$ e $v_0 = \frac{N}{L_{min}}$. *Basta definir $N := t+1$ que resulta em: $v_t = \frac{N}{L_{min}}$.*

Agora, pelos parâmetros definidos em (4.75) e pela definição de $\mathcal{B}_t(\cdot, \cdot, \cdot)$ em (4.28), podemos obter os seguintes resultados:

$$(I) : \mathcal{B}_t(x^*, x_{[t+1]}, \eta_t) \leq \frac{1}{2} D_{x^*}^2$$

De fato, pela definição de $\mathcal{B}_t(\cdot, \cdot, \cdot)$, em (4.28), temos:

$$\mathcal{B}_t(x^*, x_{[t+1]}, \eta_t) = \sum_{i=1}^t \frac{\alpha_i \eta_i}{2\Gamma_i} (\|x_i - x^*\|^2 - \|x_{i+1} - x^*\|^2)$$

Substituindo η_i , temos:

$$\mathcal{B}_t(x^*, x_{[t+1]}, \eta_t) = \sum_{i=1}^t \frac{\alpha_i}{2\Gamma_i} \left(\frac{\Gamma_i}{\alpha_i} + \chi \theta_t M_t^2 i \right) (\|x_i - x^*\|^2 - \|x_{i+1} - x^*\|^2).$$

Por hipótese $\chi = 0$, daí:

$$\mathcal{B}_t(x^*, x_{[t+1]}, \eta_t) = \sum_{i=1}^t \frac{\alpha_i}{2\Gamma_i} \left(\frac{\Gamma_i}{\alpha_i} \right) (\|x_i - x^*\|^2 - \|x_{i+1} - x^*\|^2).$$

Simplificando os termos, temos:

$$\mathcal{B}_t(x^*, x_{[t+1]}, \eta_t) = \sum_{i=1}^t \frac{1}{2} (\|x_i - x^*\|^2 - \|x_{i+1} - x^*\|^2).$$

Desenvolvendo os somatórios, temos:

$$\mathcal{B}_t(x^*, x_{[t+1]}, \eta_t) = \frac{1}{2} (\|x_1 - x^*\|^2 - \|x_2 - x^*\|^2 + \|x_2 - x^*\|^2 - \dots + \|x_t - x^*\|^2 - \|x_{t+1} - x^*\|^2).$$

Simplificando, temos:

$$\mathcal{B}_t(x^*, x_{[t+1]}, \eta_t) = \frac{1}{2} (\|x_1 - x^*\|^2 - \|x_{t+1} - x^*\|^2).$$

Por (4.7), temos:

$$\mathcal{B}_t(x^*, x_{[t+1]}, \eta_t) \leq \frac{1}{2} (D_{x^*}^2 - \|x_{t+1} - x^*\|^2).$$

Como $-\frac{1}{2}\|x_{t+1} - x^*\|^2 \leq 0$, temos:

$$\mathcal{B}_t(x^*, x_{[t+1]}, \eta_t) \leq \frac{1}{2} D_{x^*}^2.$$

$$(II) : \mathcal{B}_t(w^*, w_{[t+1]}, \theta_t) \leq \frac{ND_{x^*, K}^2}{2L_{min}}.$$

De fato, pela definição de $\mathcal{B}_t(\cdot, \cdot, \cdot)$, em (4.28), temos:

$$\mathcal{B}_t(Bw^*, Bw_{[t+1]}, \theta_t) = \sum_{i=1}^t \frac{\alpha_i \theta_i}{2\Gamma_i} (\|Bw_i - Bw^*\|^2 - \|Bw_{i+1} - Bw^*\|^2)$$

Substituindo θ_i , temos:

$$\mathcal{B}_t(w^*, w_{[t+1]}, \theta_t) = \sum_{i=1}^t \frac{\alpha_i}{2\Gamma_i} \frac{\rho v_i \Gamma_i}{\alpha_i} (\|Bw_i - Bw^*\|^2 - \|Bw_{i+1} - Bw^*\|^2).$$

Substituindo $\rho = 1$, $v_i = \frac{N}{L_{min}}$ e simplificando, temos:

$$\mathcal{B}_t(w^*, w_{[t+1]}, \theta_t) = \sum_{i=1}^t \frac{N}{2L_{min}} (\|Bw_i - Bw^*\|^2 - \|Bw_{i+1} - Bw^*\|^2).$$

Desenvolvendo os somatórios, temos:

$$\begin{aligned} \mathcal{B}_t(w^*, w_{[t+1]}, \theta_t) &= \frac{N}{2L_{min}} (\|Bw_1 - Bw^*\|^2 - \|Bw_2 - Bw^*\|^2 + \dots + \|Bw_t - Bw^*\|^2 - \|Bw_{t+1} - Bw^*\|^2) \\ &= \frac{N}{2L_{min}} (\|Bw_1 - Bw^*\|^2 - \|Bw_{t+1} - Bw^*\|^2) \\ &\leq \frac{N}{2L_{min}} \|Bw_1 - Bw^*\|^2 = \frac{N}{2L_{min}} \|Kx_1 - b - Kx^* + b\|^2 = \frac{N}{2L_{min}} \|Kx_1 - Kx^*\|^2. \end{aligned}$$

A última igualdade é válida pela hipótese do Algoritmo 4.2, onde temos: $Bw_1 = Kx_1 - b$. E, finalmente, por (4.7), temos:

$$\mathcal{B}_t(w^*, w_{[t+1]}, \theta_t) \leq \frac{ND_{x^*, K}^2}{2L_{min}}.$$

$$(III) : \mathcal{B}_t(y, y_{[t+1]}, \rho_t^{-1}) \leq \frac{N}{2L_{min}} (\|y_1 - y\|^2 - \|y_{t+1} - y\|^2), \quad \forall y \in Y.$$

De fato, pela definição de $\mathcal{B}_t(\cdot, \cdot, \cdot)$, em (4.28), temos:

$$\mathcal{B}_t(y, y_{[t+1]}, \rho_t^{-1}) = \sum_{i=1}^t \frac{\alpha_i \rho_t^{-1}}{2\Gamma_i} (\|y_i - y\|^2 - \|y_{i+1} - y\|^2).$$

Substituindo $\rho_t^{-1} = 1$, temos:

$$\mathcal{B}_t(y, y_{[t+1]}, \rho_t^{-1}) = \sum_{i=1}^t \frac{\alpha_i}{2\Gamma_i} (\|y_i - y\|^2 - \|y_{i+1} - y\|^2).$$

Como $\max_{i=1, \dots, t} \frac{\alpha_i}{\Gamma_i} \leq \frac{t+1}{L_{\min}} = \frac{N}{L_{\min}}$, temos:

$$\mathcal{B}_t(y, y_{[t+1]}, \rho_t^{-1}) \leq \sum_{i=1}^t \frac{N}{2L_{\min}} (\|y_i - y\|^2 - \|y_{i+1} - y\|^2).$$

Desenvolvendo os somatórios, temos:

$$\mathcal{B}_t(y, y_{[t+1]}, \rho_t^{-1}) \leq \frac{N}{2L_{\min}} (\|y_1 - y\|^2 - \|y_2 - y\|^2 + \|y_2 - y\|^2 - \|y_3 - y\|^2 + \dots + \|y_t - y\|^2 - \|y_{t+1} - y\|^2).$$

Simplificando, obtemos:

$$\mathcal{B}_t(y, y_{[t+1]}, \rho_t^{-1}) \leq \frac{N}{2L_{\min}} (\|y_1 - y\|^2 - \|y_{t+1} - y\|^2), \quad \forall y \in Y.$$

$$(IV) : - \sum_{i=1}^t \frac{\alpha_i(\tau_i - \rho_i)}{2\Gamma_i \rho_i} \|y_i - y_{i+1}\|^2 \leq 0.$$

Segue de $\tau_t \geq \rho_t$.

Por (4.27), temos:

$$\begin{aligned} & \frac{1}{\Gamma_t} Q(w^*, x^*, y^*; z_{t+1}^{ag}) - \sum_{i=2}^t \left(\frac{1 - \alpha_i}{\Gamma_i} - \frac{1}{\Gamma_{i-1}} \right) Q(w^*, x^*, y^*; z_i^{ag}) \\ & \leq \mathcal{B}_t(x^*, x_{[t+1]}, \eta_{[t]}) + \mathcal{B}_t(y, y_{[t+1]}, \rho_{[t]}^{-1}) + \mathcal{B}_t(Bw^*, Bw_{[t+1]}, \theta_{[t]}) - \chi \mathcal{B}_t(Kx^*, Kx_{[t+1]}, \theta_{[t]}) \\ & \quad - \sum_{i=1}^t \frac{\alpha_i(\tau_i - \theta_i)}{2\Gamma_i} \|Bw_{i+1} - Kx^* - b\|^2 + \sum_{i=1}^t \frac{\alpha_i(\tau_i - \theta_i)}{2\Gamma_i} \|B(x_{i+1} - x^*)\|^2 \\ & \quad - \sum_{i=1}^t \frac{\alpha_i(\tau_i - \rho_i)}{2\Gamma_i \rho_i} \|y_i - y_{i+1}\|^2 - \sum_{i=1}^t \frac{\alpha_i}{2\Gamma_i} (\eta_i - L_G \alpha_i - \chi \theta_i \|K\|^2) \|x_i - x_{i+1}\|^2. \end{aligned}$$

substituindo os parâmetros determinados, temos:

$$- \sum_{i=2}^t \left(\frac{1 - \alpha_i}{\Gamma_i} - \frac{1}{\Gamma_{i-1}} \right) Q(w^*, x^*, y^*; z_i^{ag}) = 0,$$

pois:

$$\left(\frac{1 - \frac{2}{i+1}}{\frac{2}{i(i-1)}} - \frac{1}{\frac{2}{(i-1)i}} \right) = \left(\frac{(1-i)(i+1)i}{(i+1)2} - \frac{(i-1)i}{2} \right) = 0.$$

,

O fato de $\tau_t = \theta_t$, resulta em:

$$\begin{aligned} - \sum_{i=1}^t \frac{\alpha_i(\tau_i - \theta_i)}{2\Gamma_i} \|Bw_{i+1} - Kx^* - b\|^2 &= 0, \\ \sum_{i=1}^t \frac{\alpha_i(\tau_i - \theta_i)}{2\Gamma_i} \|B(x_{i+1} - x^*)\|^2 &= 0. \end{aligned}$$

O fato de $\chi = 0$, resulta em:

$$\chi \mathcal{B}_t(Kx^*, Kx_{[t+1]}, \theta_{[t]}) = 0$$

e, por (4.73), temos:

$$- \sum_{i=1}^t \frac{\alpha_i}{2\Gamma_i} (\eta_i - L_G \alpha_i - \chi \theta_i \|K\|^2) \|x_i - x_{i+1}\|^2 \leq 0.$$

Logo:

$$\frac{1}{\Gamma_t} Q(w^*, x^*, y; z_{t+1}^{ag}) \leq \frac{1}{2} D_{x^*}^2 + \frac{N}{2L_{min}} D_{x^*,K}^2 + \frac{N}{2L_{min}} (\|y_1 - y\|^2 - \|y_{t+1} - y\|^2) \quad \forall y \in Y.$$

De maneira similar à demonstração do Teorema 4.5, temos:

$$\|y_{t+1} - y^*\|^2 \leq \frac{L_{min} D_{x^*}^2}{N} + D_{x^*,K}^2 + D_y^{2*}$$

e

$$\|y_1 - y^{t+1}\|^2 \leq \frac{2L_{min} D_{x^*}^2}{N} + 2D_{x^*,K}^2 + 4D_y^{2*}. \quad (4.87)$$

Resultando em:

$$\frac{1}{\Gamma_t} Q(w^*, x^*, y; z_{t+1}^{ag}) + \frac{N}{L_{min}} \langle y_1 - y_{t+1}, y \rangle \leq \frac{1}{2} D_{x^*}^2 + \frac{N}{2L_{min}} D_{x^*,K}^2, \quad \forall y \in Y.$$

Multiplicando ambos os lados da desigualdade por Γ_t , temos:

$$Q(w^*, x^*, y; z_{t+1}^{ag}) + \frac{N\Gamma_t}{L_{min}} \langle y_1 - y_{t+1}, y \rangle \leq \frac{\Gamma_t}{2} D_{x^*}^2 + \frac{\Gamma_t N}{2L_{min}} D_{x^*,K}^2, \quad \forall y \in Y.$$

Definindo $v_{t+1} = \frac{\Gamma_t N (y_1 - y_{t+1})}{L_{min}}$, $t = N - 1$ e aplicando (4.81), temos:

$$Q(w^*, x^*, y; z_{t+1}^{ag}) + \langle v_N, y \rangle \leq \frac{8L_G}{2(N-1)^2} D_{x^*}^2 + \frac{4L_G N}{(N-1)^2 L_{min}} D_{x^*,K}^2.$$

Pela Proposição 4.2, temos:

$$\begin{aligned}
G(x_N^{ag}) + F(Kx_N^{ag}) - f^* &\leq \frac{8L_G}{2(N-1)^2} D_{x^*}^2 + \frac{4L_G N}{(N-1)^2 L_{min}} D_{x^*,K}^2 \\
&\leq \frac{4L_G}{(N-1)^2} D_{x^*}^2 + \frac{4L_G N}{(N-1)N L_{min}} D_{x^*,K}^2 \\
&= \frac{4L_G}{(N-1)^2} D_{x^*}^2 + \frac{4L_G}{(N-1)L_{min}} D_{x^*,K}^2.
\end{aligned}$$

Provando (4.85).

Note que,

$$\|v_N\|^2 = \|v_{t+1}\|^2 = \left\| \frac{\Gamma_t N (y_1 - y_{t+1})}{L_{min}} \right\|^2 = \frac{\Gamma_t^2 N^2}{L_{min}^2} \|y_1 - y_{t+1}\|^2.$$

Pela majoração (4.87), obtemos:

$$\|v_N\|^2 \leq \frac{\Gamma_t^2 N^2}{L_{min}^2} \left(\frac{2L_{min} D_{x^*}^2}{N} + 2D_{x^*,K}^2 + 4D_y^{2*} \right).$$

Por (4.81), temos:

$$\|v_N\|^2 \leq \frac{64L_G^2 N^2}{(N-1)^4 L_{min}^2} \left(\frac{2L_{min} D_{x^*}^2}{N} + 2D_{x^*,K}^2 + 4D_y^{2*} \right).$$

Simplificando, temos:

$$\|v_N\|^2 \leq \frac{128L_G^2 N D_{x^*}^2}{(N-1)^4 L_{min}} + \frac{128L_G^2 N^2 D_{x^*,K}^2}{(N-1)^4 L_{min}^2} + \frac{256L_G^2 N^2 D_y^{2*}}{(N-1)^4 L_{min}^2}.$$

Multiplicando por 2 o lado direito da desigualdade e usando o fato de que $L_{min} \leq 2L_G$, temos:

$$\|v_N\|^2 \leq \frac{256L_G^2 N D_{x^*}^2}{(N-1)^4 L_{min}} + \frac{512L_G N^2 D_{x^*,K}^2}{(N-1)^4 L_{min}} + \frac{512L_G^2 N^2 D_y^{2*}}{(N-1)^4 L_{min}^2}.$$

Multiplicando por 2 o termo $\frac{512L_G^2 N^2 D_y^{2*}}{(N-1)^4 L_{min}^2}$, temos:

$$\begin{aligned}
\|v_N\|^2 &\leq \frac{256L_G^2 N D_{x^*}^2}{(N-1)^4 L_{min}} + \frac{512L_G N^2 D_{x^*,K}^2}{(N-1)^4 L_{min}} + \frac{1024L_G^2 N^2 D_y^{2*}}{(N-1)^4 L_{min}^2} \\
&\leq \frac{256L_G^2 N D_{x^*}^2}{(N-1)^3 N L_{min}} + \frac{512L_G N^2 D_{x^*,K}^2}{(N-1)^2 N^2 L_{min}} + \frac{1024L_G^2 N^2 D_y^{2*}}{(N-1)^2 N^2 L_{min}^2}.
\end{aligned}$$

Simplificando, temos:

$$\|v_N\|^2 \leq \frac{256L_G^2 D_{x^*}^2}{(N-1)^3 L_{min}} + \frac{512L_G D_{x^*,K}^2}{(N-1)^2 L_{min}} + \frac{1024L_G^2 D_y^{2*}}{(N-1)^2 L_{min}^2}.$$

Extraindo a raiz quadrada em ambos os lados da desigualdade, temos:

$$\|v_N\| \leq \frac{16\sqrt{L_G}D_{x^*}}{(N-1)^{\frac{3}{2}}\sqrt{L_{min}}} + \frac{16\sqrt{2}\sqrt{L_G}D_{x^*,K}}{(N-1)\sqrt{L_{min}}} + \frac{32L_GD_{y^*}}{(N-1)L_{min}}.$$

Provando assim (4.86), pois pelo Algoritmo 4.2, temos: $v_n = Bw_N^{ag} - K_N^{ag} - b$.

■

Referências Bibliográficas

- [1] AUSLENDER, A.; TEBOULLE, M. Interior projection-like methods for monotone variational inequalities. *Mathematical Programming*, vol. 104, Issue 1, pp. 39-68, September 2005.
- [2] AUSLENDER, A.; TEBOULLE, M. Interior gradient and proximal methods for convex and conic optimization. *SIAM Journal on Optimization*, vol. 16, Issue 3, pp. 697-725, July 2006.
- [3] BANERJEE, A.; MEGURU, S.; DHILLON, I. S.; GHOSH, J. Clustering with Bregman Divergence. *Journal of Machine Learning Research*, vol. 6, pp. 1705-1749, October 2005.
- [4] BECK, A.; TEBOULLE, M. A fast iterative shrinkage-threshold algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, vol. 2, Issue 1, pp. 183-202, March 2009.
- [5] CHAMBOLLE A.; POCK, T. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of Mathematical Imaging and Vision*, vol. 40, Issue 1, pp. 120-145, May 2011.
- [6] CHEN, S. S.; DONOHO, D. L.; SAUNDERS, M. A. Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, vol. 20, Issue 1, pp. 33-61, July 2006.
- [7] CHEN, Y.; HAGER, W.; HUANG, F.; PHAN, D.; YE, X.; YIN, W. Fast algorithms for image reconstruction with application to partially parallel MR imaging. *SIAM Journal on Imaging Sciences*, vol. 5, Issue 1, pp. 90-118, January 2012.
- [8] d'ASPREMONT, A.; BANERJEE, O.; GHAOU, L. E.; ZHANG, Y. First-order methods for sparse covariance selection. *SIAM Journal on Matrix Analysis and Applications*, vol. 30, Issue 1, pp. 56-66, January 2008.
- [9] FUKUSHIMA, M.; MINE, H. A generalized proximal point algorithm for certain non-convex minimization problems. *International Journal of Systems Science*, vol. 12, Issue 8, pp. 989-1000, 1981.
- [10] FRIEDLANDER, M. P.; TSENG, P. Exact Regularization of convex grams. *SIAM Journal on Optimization*, vol. 18, Issue 4, pp. 1326-1350, November 2007.
- [11] GONZAGA, C. C.; KARAS, E. W. Optimal Steepest descent algorithms for unconstrained convex problems: fine tuning nesterov's method. *Mathematical Programming*, vol. 138, Issue 1-2, pp. 141-166, April 2013.

- [12] HESTENES, M. R. Multiplier and gradient methods. *Journal of Optimization Theory and Applications*, vol. 4, Issue 5, pp. 303-320, November 1969.
- [13] IUSEM, A. *Método de ponto proximal em otimização*, Instituto de Matemática Pura e Aplicada - CNPq. Livro do 20º Colóquio Brasileiro de Matemática, 1995.
- [14] LAN, G.; LU, Z.; MONTEIRO, R. D. C. Primal-dual first-order methods with $O\left(\frac{1}{\epsilon}\right)$ iteration-complexity for cone programming. *Mathematical Programming*, vol. 126, Issue 1, pp.1-29, January 2011.
- [15] LU, Z. Smooth optimization approach for sparse covariance selection. *SIAM Journal on Optimization Archive*, vol. 19, Issue 4, pp. 1807-1827, December 2008.
- [16] LU, Z.; MONTEIRO, R. D. C.; YUAN, M. Convex optimization methods for dimension reduction and coefficient estimation in multivariate linear regression. *Mathematical Programming*, vol.131, Issue 1-2, pp.163-194, February 2012.
- [17] MONTEIRO, R. D C.; SVAITER, B. F. Iteration-complexity of block-decomposition algorithms and the alternating direction method of multipliers. *SIAM Journal on Optimization*, vol. 23, Issue 1, pp. 475-507, March 2013.
- [18] NEMIROVSKI, A.; YUDIM, D. Problem Complexity and Method Efficiency in Optimization. *SIAM Review*, vol.27, Issue 2, pp. 264-265, 1985.
- [19] NESTEROV, Y. On an approach to the construction of optimal methods of minimization of smooth convex functions. *Ekonomika i Mateaticheskie, Metody* 24, pp. 509-517, 1988
- [20] NESTEROV, Y. A method for unconstrained convex minimization problem with rate of convergence $o(\frac{1}{k^2})$. *Doklady AN SSSR 269 (translated as Soviet Math Docl)*, pp. 543-547, 1983.
- [21] NESTEROV, Y. *Introductory Lectures on Convex Optimization. A basic course*, Kluwer Academic Publishers, Boston,(2004).
- [22] NESTEROV, Y. Smooth minimization of nonsmooth functions. *Mathematical Programming*, vol. 103, Issue 1, pp. 127-152, May 2005.
- [23] NESTEROV, Y. Excessive gap technique in nonsmooth convex minimization. *SIAM Journal on Optimization*, vol. 16, Issue 1, July 2006.
- [24] NESTEROV, Y. Gradient method for minimizing composite objective function. *Mathematical Programming*, vol.140, Issue 1, pp. 125-161, August 2013.
- [25] OLIVEIRA, P. A. *Convergência Finita Algoritmo do Ponto Proximal para Desigualdade Variacional*. 2006, 47 p. Dissertação de Mestrado, Universidade Federal de Goiás, 2006.

- [26] OUYANG, Y.; CHEN, Y.; LAN, G.; PASILIAO JR, E. An accelerated linearized alternating direction method of multipliers. *SIAM Journal on Imaging Sciences*, vol. 8, Issue 1 , pp. 644-681, March 2015.
- [27] POLYAK, B. T. *Introduction to Optimization*, Optimization Software, New York, 1987.
- [28] ROCKAFELLAR, R. T. *Convex analysis*. Princeton University Press (Princeton, NJ), 1970.
- [29] ROSSETTO, D. R. *Tópicos em métodos ótimos para otimização convexa*. 2012, 90p. Tese de Doutorado, Universidade de São Paulo, 2012.
- [30] TSENG, P. On accelerated proximal gradient methods for convex-concave optimization. submitted to *SIAM Journal on Optimization*, 2008.
- [31] TSENG, P.; YUN, S. A block - coordinate gradient descent method for linearly constrained nonsmooth separable optimizatin. *Journal of Optimization Theory and Applications* vol. 140, Issue 3, pp 513-535, March 2009.
- [32] YE, X.; CHEN, Y.; HUANG, F. Computational acceleration for MR image reconstruction in partially parallel imaging. *Medical Imaging, IEEE Transactions on*, vol. 30, Issue 5, pp. 1055-1063, May 2011.
- [33] WANG, Y.; YANG, J.; YIN, W.; ZHANG, Y. A new alternating minimization algorithm for total variation image reconstruction. *SIAM Journal on Imaging Sciences*, vol. 1, Issue 3, pp. 248-272, August 2008.