UNIVERSIDADE ESTADUAL DE CAMPINAS
INSTITUTO DE MATEMÁTICA, ESTATÍSTICA E COMPUTAÇÃO CIENTÍFICA
DEPARTAMENTO DE MATEMÁTICA APLICADA

GABRIEL PASSOS

Investigação computacional do método ADMM inexato

Campinas
08/01/2021

GABRIEL PASSOS

Investigação computacional do método ADMM inexato*

Monografia apresentada ao Instituto de Matemática, Estatística e Computação Científica da Universidade Estadual de Campinas como parte dos requisitos para obtenção de créditos na disciplina Projeto Supervisionado, sob a orientação do(a) Prof. Dra. Sandra Augusta Santos.

*Este trabalho foi financiado pela FAPESP, processo 2019/15992-5.

Resumo

O método ADMM *(Alternating Direction Method of Multipliers)* é um algoritmo utilizado para abordar problemas convexos separáveis com restrições lineares. Ele divide o problema restrito original em uma sequência de subproblemas irrestritos simples que aproveitam da estrutura separável do problema original. Contudo, resolver os subproblemas de maneira exata como é exigido pelo método nem sempre é possível. Uma saída para essa dificuldade é tratar os subproblemas de forma aproximada, algo que demanda o desenvolvimento de critérios inexatos que garantam a convergência do método. Neste trabalho, são apresentados alguns critérios inexatos que podem ser utilizados como alternativa no método ADMM. Em especial, o método ADMM parcialmente inexato baseado no critério do HPE foi aplicado ao problema LASSO e seu desempenho foi comparado com o do método exato.

Abstract

The ADMM *(Alternating Direction Method of Multipliers)* method is an algorithm used to address separable convex problems with linear constraints. It divides the original restricted problem into a sequence of simple unrestricted subproblems that take advantage of the separable structure of the original problem. However, solving the subproblems exactly as required by the method is not always possible. One way to overcome this difficulty is to treat the subproblems in an approximate way, which demands the development of inexact criteria that guarantee the convergence of the method. In this work, some inexact criteria that can be used as an alternative in the ADMM method are presented. In particular, the partially inexact ADMM method based on the HPE criterion is applied to the LASSO problem and its performance is compared with the performance of the exact method.

Sumário

1 Critérios inexatos

Uma das técnicas mais usuais para abordar problemas que envolvem restrições consiste em dividir o problema original em uma série de problemas menores e mais simples. Esses subproblemas são resolvidos de forma sequencial, gerando uma sequência de termos que converge, ou possui subsequência convergente, para uma das soluções desejadas.

Quando esse tipo de abordagem é utilizada, o primeiro resultado de convergência que se obtém vem, em geral, associado à resolução exata de cada subproblema. Se existe uma fórmula analítica para essa solução, pode ser impossível aplicá-la por fatores como limitação de memória computacional. Nesse caso, deve-se abordar cada subproblema de uma outra maneira, como através de um método iterativo por exemplo. Caso a expressão analítica não exista, não há escolha, um método iterativo deve ser aplicado.

A maior desvantagem em usar métodos iterativos para solucionar cada subproblema de forma exata é o alto custo computacional para se atingir alta precisão na solução. Contudo, foi descoberto que alcançar uma solução precisa em todas as iterações não é algo obrigatório para se obter uma solução do problema original e assim, uma série de critérios inexatos foram desenvolvidos, como por exemplo em [6, 17, 11, 7, 18].

Discutiremos na Seção 1.1 um critério inexato para métodos baseados nas funções Lagrangianas aumentadas apresentado e testado em [7]. Esse critério foi utilizado nos subproblemas do método ADMM em [20, 19]. Na Seção 1.2 discutiremos um critério inexato usado no método do ponto proximal para operadores monótonos maximais. Esse critério foi apresentado pela primeira vez em [17] e utilizado nos subproblemas do método ADMM em [18].

1.1 Critério inexato para Lagrangianas Aumentadas

Considere o seguinte problema de otimização

$$\begin{aligned} \min \quad & f(x) \\ s.a. \quad & Ax = b, \end{aligned} \tag{1}$$

em que $f : \mathbb{R}^n \to \mathbb{R}$ é uma função convexa, $A \in \mathbb{R}^{l\times n}$ e $b \in \mathbb{R}^l$. Define-se a função Lagrangiana aumentada do problema (1), em termos de um parâmetro $\lambda \geq 0$, como

$$\mathcal{L}_\lambda(x, p) := f(x) + \langle p, Ax - b\rangle + \frac{\lambda}{2}\|Ax - b\|^2.$$

Um dos métodos clássicos que pode ser utilizado para abordar o problema (1) e que é inspirado na ideia de solucionar uma sequência de problemas menores é o método dos multiplicadores. O método dos multiplicadores trabalha com a minimização sequencial da função Lagrangiana aumentada e gera sequências (x^k), (p^k) que obedecem à seguinte recursividade:

$$x^{k+1} \in \arg\min_{x\in\mathbb{R}^n} \mathcal{L}_{\lambda_k}(x, p^k),$$

$$p^{k+1} := p^k + \lambda_k(Ax^{k+1} - b).$$

Como o conjunto viável do problema (1) é descrito por restrições lineares, trata-se de um conjunto qualificado. Consequentemente, a função Lagrangiana aumentada atua como uma penalização exata, no sentido que o parâmetro de penalização λ_k não precisa crescer ilimitadamente para se conquistar a viabilidade da sequencia gerada. Além disso, obtém-se os seguintes resultados:

- Toda subsequência convergente de (x^k) tem como ponto limite uma solução do problema (1).
- (p^k) converge para um multiplicador de Lagrange do problema (1).

Determinar um minimizador de $\mathcal{L}_{\lambda_k}(x, p^k)$ nem sempre é uma tarefa simples. Se não existe uma fórmula fechada, ou se é impossível de avaliar a fórmula existente devido a fatores externos, os subproblemas devem ser solucionados por métodos iterativos. Contudo, obter uma solução de forma iterativa com alta precisão pode ser uma operação computacionalmente custosa. Podemos contornar essa dificuldade estabelecendo um critério inexato que, se respeitado, garante a convergência para uma solução do problema original, mesmo com uma solução aproximada de cada subproblema.

Supondo que a função f do problema (1) seja diferenciável, uma técnica muito comum utilizada como critério de parada do método iterativo que resolve os subproblemas

é considerar x^{k+1} satisfazendo a condição

$$\|\nabla_x \mathcal{L}_{\lambda_k}(x^{k+1}, p^k)\| \leq \varepsilon_k,$$

para algum $\varepsilon_k \geq 0$ adequado. A ideia em utilizar esse critério é garantir alguma proximidade de x^{k+1} da solução real do subproblema, que é uma solução do sistema não-linear $\nabla_x \mathcal{L}_{\lambda_k}(x, p^k) = 0$. Note que esse critério é relativamente simples de ser testado pois só depende da avaliação do gradiente $\nabla_x \mathcal{L}_{\lambda_k}(x, p^k)$.

Um outro critério inexato para abordar métodos que se baseiam na minimização sequencial da função Lagrangiana aumentada foi desenvolvido recentemente em [7]. Utilizando esse critério no método dos multiplicadores e supondo que a função f do problema (1) seja diferenciável, o método gera sequências (x^k), (g^k), (p^k) e (w^k) que satisfazem para todo k as condições:

$$g^{k+1} = \nabla_x \mathcal{L}_{\lambda_k}(x^{k+1}, p^k),$$

$$\frac{2}{\lambda_k}|\langle w^k - x^{k+1}, g^{k+1}\rangle| + \|g^{k+1}\|^2 \leq \sigma \|Ax^{k+1} - b\|^2,$$

$$p^{k+1} := p^k + \lambda_k(Ax^{k+1} - b),$$

$$w^{k+1} := w^k - \lambda_k g^{k+1},$$

em que $\sigma \in [0, 1)$.

Veja que esse critério também requer somente a avaliação do gradiente $\nabla_x \mathcal{L}_{\lambda_k}(x, p^k)$, além da atualização da sequência (w^k).

1.2 Critério inexato para operadores

Um operador $\mathcal{T} : \mathbb{R}^n \to 2^{\mathbb{R}^n}$ é um mapa que associa a cada elemento $x \in \mathbb{R}^n$, um subconjunto $\mathcal{T}(x) \subset \mathbb{R}^n$. Dizemos que $\mathcal{T}$ é monótono quando

$$\langle x' - x, y' - y\rangle \geq 0, \quad \forall\, x, x' \in \mathbb{R}^n, \forall y \in \mathcal{T}(x), \forall y' \in \mathcal{T}(x').$$

Essas e outras definições, além de resultados importantes para este trabalho, podem ser encontrados no Apêndice B. Para um aprofundamento na teoria dos operadores monótonos, veja [1].

A vantagem dessa ferramenta é que muitos problemas podem ser tratados como o de determinar um zero de um operador apropriado, isto é, determinar x tal que

$$0 \in \mathcal{T}(x).$$

Esse tipo de interpretação é vantajosa pois permite a extensão de métodos já conhecidos para problemas mais gerais, como por exemplo o método do ponto proximal. Além disso, o poderoso ferramental fornecido por essa teoria é fundamental para demonstrar a convergência de diversos métodos.

Um operador de importância central para resolver o problema descrito anteriormente é o operador resolvente. Dado um operador $\mathcal{T}$, o resolvente de $\mathcal{T}$ é um operador na forma

$$\mathcal{J}_{\lambda\mathcal{T}}(x) := (\mathcal{I} + \lambda\mathcal{T})^{-1}(x),$$

em que λ é um escalar positivo.

O operador $\mathcal{J}_{\lambda\mathcal{T}}$ possui características que podem ser exploradas de diversas maneiras para determinar um zero de $\mathcal{T}$. Uma delas é a caracterização dos zeros de um operador monótono maximal $\mathcal{T}$ pelos pontos fixos de $\mathcal{J}_{\lambda\mathcal{T}}$. É possível demonstrar que para um operador monótono maximal $\mathcal{T}$, vale a equivalência (Teorema B.1):

$$0 \in \mathcal{T}(x) \Leftrightarrow x = \mathcal{J}_{\lambda\mathcal{T}}(x).$$

Observação 1.1. Note que foi utilizada a igualdade $x = \mathcal{J}_{\lambda\mathcal{T}}(x)$ ao invés da pertinência $x \in \mathcal{J}_{\lambda\mathcal{T}}(x)$ na equivalência anterior. Isso é permitido pois como $\mathcal{T}$ é monótono maximal, a Proposição B.1 garante que $\mathcal{J}_{\lambda\mathcal{T}}(x)$ é formado por exatamente um elemento. Assim, para facilitar a notação e evidenciar esse fato, utiliza-se do sinal de igualdade.

Um dos primeiros métodos desenvolvidos para determinar zeros de operadores monótonos maximais, chamado Método do Ponto Proximal, surgiu da tentativa de explorar essa propriedade de ponto fixo [10, 14, 15], e é constituído pelo seguinte processo

iterativo:

$$x^{k+1} = \mathcal{J}_{\lambda\mathcal{T}}(x^k). \tag{2}$$

Como $\mathcal{T}$ é monótono maximal, a Proposição B.1 garante que o resolvente seja um operador não-expansivo firme, que é uma propriedade suficiente para garantir a convergência da sequência (x^k) gerada pelo processo iterativo (2) a um ponto fixo de $\mathcal{J}_{\lambda\mathcal{T}}$, e portanto, a um zero de $\mathcal{T}$.

Embora o método do ponto proximal seja bem simples de compreender, ainda temos um obstáculo semelhante ao que encontramos anteriormente: pode ser muito difícil calcular $\mathcal{J}_{\lambda\mathcal{T}}(x)$. Todavia, é possível desenvolver métodos que aproveitem a estrutura do operador $\mathcal{T}$ de maneira eficiente, ou ainda, abordar o cálculo do resolvente de forma aproximada, como foi feito anteriormente com o critério inexato das Lagrangianas aumentadas.

Uma maneira de aproveitar da estrutura do operador $\mathcal{T}$ se dá quando este pode ser escrito na forma $\mathcal{T} = \mathcal{F} + \mathcal{G}$. Essa estrutura pode ser aproveitada por métodos baseados em separação, como *Douglas-Rachford* ou *Peaceman-Rachford*. Essas classes de métodos utilizam dos resolventes $\mathcal{J}_{\lambda\mathcal{F}}$ e $\mathcal{J}_{\lambda\mathcal{G}}$ ao invés do resolvente $\mathcal{J}_{\lambda\mathcal{T}}$. Um tratamento completo desses métodos é apresentado em [5].

Uma segunda alternativa para simplificar o método do ponto proximal é trabalhar com resolventes inexatos, utilizando por exemplo o ε-*enlargement* de $\mathcal{T}$, originando os métodos de ponto proximal inexatos. Algumas dessas estratégias para o caso particular com $\mathcal{T} = \partial f$ são descritas nos trabalhos [8, 16], em que $f : \mathbb{R}^n \to (-\infty, +\infty]$ é uma função convexa, própria e fechada[†]. Um método de ponto proximal inexato que considera um operador $\mathcal{T}$ monótono maximal genérico, denominado de *Hybrid Proximal Extragradient* (HPE) é apresentado em [17] e descrito brevemente a seguir.

1.3 O método HPE

O método HPE *(Hybrid Proximal Extragradient)* é um método de ponto proximal inexato apresentado pela primeira vez em [17]. Este método foi desenvolvido para determinar um zero de um operador monótono maximal $\mathcal{T}$ e sua principal vantagem está

[†]Para definição de função convexa, própria e fechada, veja Apêndice A

na maneira em que o operador resolvente é aproximado.

Como descrito na Seção 1.2, o método do ponto proximal aplicado em um operador $\mathcal{T}$, definido pelo processo apresentado em (2), depende do seu resolvente $\mathcal{J}_{\lambda\mathcal{T}}$, cujo cálculo pode ser computacionalmente caro em uma implementação prática do método. O HPE utiliza de uma propriedade de representação do resolvente $\mathcal{J}_{\lambda\mathcal{T}}(x)$ que é dependente somente do ponto x no qual o resolvente é avaliado e de um elemento único $y \in \mathcal{T}(x)$ na tentativa de relaxar o método do ponto proximal e utilizar um critério inexato ao invés de calcular o resolvente de forma precisa.

A propriedade de representação do resolvente nos diz que

$$y = \mathcal{J}_{\lambda\mathcal{T}}(x) \Leftrightarrow \exists v \in \mathcal{T}(y) \quad \text{tal que} \quad x = y + \lambda v. \tag{3}$$

Essa propriedade pode ser demonstrada facilmente utilizando a definição do operador resolvente e é apresentada em [17, Equação (1)]. A equivalência em (3) nos sugere que para aliviar o cálculo do resolvente, podemos relaxar as condições

$$v \in \mathcal{T}(y),$$

$$\lambda v + y - x = 0.$$

Uma das maneiras de realizar isso é considerar y satisfazendo, para um $\varepsilon \geq 0$ adequado, as relações

$$v \in \mathcal{T}^{\varepsilon}(y), \tag{4}$$

$$\|\lambda v + y - x\|^2 + 2\lambda\varepsilon \leq \sigma^2\|y - x\|^2, \quad \sigma \in [0, 1]. \tag{5}$$

Aqui, $\mathcal{T}^{\varepsilon}(x)$ denota o ε-*enlargement* do operador $\mathcal{T}$, apresentado na Definição B.10. Detalhes da dedução dessa relaxação podem ser encontrados em [17].

O método HPE aproveita dessa estratégia de relaxação e gera sequências que seguem a recursividade

$$v^k \in \mathcal{T}^{\varepsilon_k}(y^k),$$

$$\|\lambda v^k + y^k - x^k\|^2 + 2\lambda\varepsilon_k \leq \sigma^2\|y^k - x^k\|^2, \quad \sigma \in [0, 1].$$

$$x^{k+1} = x^k - \lambda v^k.$$

A demonstração da convergência desse método pode ser encontrada em [17, Teorema 3.1].

Note que o termo x^{k+1} não foi definido como y^k, que é a aproximação do resolvente pelo critério inexato apresentado, mas sim como $x^k - \lambda v^k$. Como a direção v^k deve ser calculada para verificar a condição, os autores optaram por aproveitar dela e então utilizaram uma abordagem semelhante à encontrada no método extragradiente [9]. O método extragradiente para minimizar uma função convexa diferenciável é dado por

$$x^{k+1} = x^k - \alpha_k \nabla f(x^{k+1/2}),$$

em que $x^{k+1/2}$ é um ponto auxiliar determinado, em geral, por

$$x^{k+1/2} = x^k - \alpha_k \nabla f(x^k).$$

Nesse caso, y^k pode ser interpretado como o ponto auxiliar $x^{k+1/2}$ e v^k como a direção do gradiente $\nabla f(x^{k+1/2})$ utilizado como no método extragradiente.

2 O método ADMM

Considere o seguinte problema de otimização

$$\begin{aligned} \min_{x,w} \quad & f(x) + g(w) \\ s.a \quad & Ax + Bw = c, \end{aligned} \tag{P}$$

em que $f : \mathbb{R}^m \to (-\infty, +\infty]$ e $g : \mathbb{R}^n \to (-\infty, +\infty]$ são funções convexas, próprias e fechadas, $A \in \mathbb{R}^{l\times m}$, $B \in \mathbb{R}^{l\times n}$ e $c \in \mathbb{R}^l$. A função Lagrangiana aumentada do problema (P) com parâmetro $\lambda > 0$ é dada por

$$\mathcal{L}_\lambda(x, w, p) := f(x) + g(w) + \langle p, Ax + Bw - c\rangle + \frac{\lambda}{2}\|Ax + Bw - c\|^2.$$

Definiremos por $\overline{\tau}$ o valor objetivo ótimo, isto é,

$$\overline{\tau} = \inf_{\substack{x\in\mathbb{R}^m \\ w\in\mathbb{R}^n}} \{f(x) + g(w) \mid Ax + Bw - c\}.$$

Será útil neste trabalho considerar uma segunda forma da função Lagrangiana aumentada, denominada de forma escalada:

$$\mathcal{L}_\lambda(x, w, p) = f(x) + g(w) + \frac{\lambda}{2}\|Ax + Bw - c + \frac{1}{\lambda}p\|^2 - \frac{1}{2\lambda}\|p\|^2. \tag{6}$$

Interpretando o problema (P) como um caso particular do problema (1), vemos que é possível utilizar o método dos multiplicadores para tentar resolvê-lo, o que nos leva ao seguinte processo iterativo:

$$(x^{k+1}, w^{k+1}) \in \operatorname*{arg\,min}_{\substack{x\in\mathbb{R}^m \\ w\in\mathbb{R}^n}} \mathcal{L}_{\lambda_k}(x, w, p^k),$$

$$p^{k+1} := p^k + \lambda_k(Ax^{k+1} + Bw^{k+1} - c).$$

Observando a estrutura do problema (P) e o método dos multiplicadores aplicado a ele, nota-se que a natureza separável deste problema não está sendo explorada. Um dos métodos apropriados para tratar problemas na forma de (P) é o ADMM *(Alternating*

Direction Method of Multipliers), que tenta manter as propriedades de convergência do método dos multiplicadores e ainda aproveitar da separabilidade do problema em questão [3].

Considere dados $x^0 \in \mathbb{R}^m$, $w^0 \in \mathbb{R}^n$, $p^0 \in \mathbb{R}^l$ e o escalar $\lambda > 0$. A proposta do ADMM é gerar, a partir de x^0, w^0 e p^0, sequências (x^k), (w^k) e (p^k) que seguem um processo iterativo dado por

$$\begin{aligned} x^{k+1} &\in \arg\min_{x \in \mathbb{R}^m} \mathcal{L}_\lambda(x, w^k, p^k), \\ w^{k+1} &\in \arg\min_{w \in \mathbb{R}^n} \mathcal{L}_\lambda(x^{k+1}, w, p^k), \\ p^{k+1} &:= p^k + \lambda(Ax^{k+1} + Bw^{k+1} - c). \end{aligned} \tag{7}$$

Mostra-se que as sequências (x^k), (w^k) e (p^k) geradas por esse processo satisfazem (cf. [3]):

- $Ax^k + Bw^k - c \to 0$.
- $f(x^k) + g(w^k) \to \overline{\tau}$.
- (p^k) converge para um multiplicador de Lagrange do problema (P).

Perceba que a diferença fundamental do ADMM em relação ao método dos multiplicadores é que, no ADMM, a função Lagrangiana aumentada é minimizada primeiramente na variável x, e em seguida na variável w, além do parâmetro λ ser considerado fixo.

Como as funções $\mathcal{L}_\lambda(x, w^k, p^k)$ e $\mathcal{L}_\lambda(x^{k+1}, w, p^k)$ podem possuir mais de um minimizador, não é possível considerar diretamente a igualdade na definição de x^{k+1} e w^{k+1} dada em (7). Por isso, considera-se a pertinência de x^{k+1} e w^{k+1} aos seus respectivos conjuntos de minimizadores.

Será necessário caracterizar as soluções do problema (7). Essa caracterização será baseada baseada nas condições de *Kuhn-Tucker* para o problema (P) e no desenvolvimento do método ADMM feito em [5].

Definição 2.1 (Tripla *Kuhn-Tucker*)**.** Dizemos que $(\overline{x}, \overline{w}, \overline{p}) \in \mathbb{R}^m \times \mathbb{R}^n \times \mathbb{R}^l$ é uma tripla

Kuhn-Tucker para o problema (P) se $(\overline{x}, \overline{w})$ satisfaz a igualdade $A\overline{x} + B\overline{w} = c$ e ainda

$$0 \in \partial f(\overline{x}) + A^t\overline{p},$$
$$0 \in \partial g(\overline{w}) + B^t\overline{p}.$$

Mostra-se que se $(\overline{x}, \overline{w}, \overline{p}) \in \mathbb{R}^m \times \mathbb{R}^n \times \mathbb{R}^l$ é uma tripla *Kuhn-Tucker* para o problema (P), então o par $(\overline{x}, \overline{w})$ é uma solução do problema (P) e $\overline{p}$ é um multiplicador de Lagrange.

2.1 O método ADMM generalizado

O método ADMM apresentado em (7) ainda sofre da mesma dificuldade comentada nas seções anteriores, a de que cada subproblema deve ser solucionado de maneira exata. Isso novamente restringe o método a ser utilizado somente em problemas que permitem fórmulas fechadas aplicáveis para calcular x^{k+1} e w^{k+1} em (7). Apresentaremos a seguir uma generalização do método ADMM que permitirá que os subproblemas sejam resolvidos de forma aproximada.

Para estabelecer a generalização do método ADMM, é conveniente modificar a forma dos subproblemas em (7). Expandindo os produtos internos e eliminando os termos constantes no cálculo dos minimizadores, obtemos

$$\arg\min_{x\in\mathbb{R}^m} \mathcal{L}_\lambda(x, w^k, p^k) = \arg\min_{x\in\mathbb{R}^m} \left\{ f(x) + \langle p^k, Ax\rangle + \frac{\lambda}{2}\|Ax + Bw^k - c\|^2 \right\}, \tag{8}$$

$$\arg\min_{w\in\mathbb{R}^n} \mathcal{L}_\lambda(x^{k+1}, w, p^k) = \arg\min_{w\in\mathbb{R}^n} \left\{ g(w) + \langle p^k, Bw\rangle + \frac{\lambda}{2}\|Ax^{k+1} + Bw - c\|^2 \right\} \tag{9}$$

A ideia fundamental do ADMM generalizado é gerar sequências (x^k), (w^k) e (p^k) tais que para cada $k = 0, 1, 2, ...$, x^{k+1} e w^{k+1} estejam suficientemente próximos das soluções exatas $\tilde{x}^k$ e $\tilde{w}^k$, respectivamente. Será considerado na versão generalizada que

$$\tilde{x}^k \in \arg\min_{x\in\mathbb{R}^m} \left\{ f(x) + \langle p^k, Ax\rangle + \frac{\lambda}{2}\|Ax + Bw^k - c\|^2 \right\},$$

$$\tilde{w}^k \in \arg\min_{w\in\mathbb{R}^n} \left\{ g(w) + \langle p^k, Bw\rangle + \frac{\lambda}{2}\|\rho_k Ax^{k+1} - (1-\rho_k)(Bw^k - c) + Bw - c\|^2 \right\}.$$

Observe que a definição para $\tilde{w}^k$ foi modificada em relação ao apresentado em (9). O que ocorreu foi a troca do termo Ax^{k+1} na penalização por

$$\rho_k Ax^{k+1} - (1-\rho_k)(Bw^k - c)$$

A introdução do parâmetro de relaxamento ρ_k da forma como foi apresentada é induzida quando analisamos o ADMM como um caso particular do método de *Douglas-Rachford* generalizado apresentado em [5]. A adição desse parâmetro permite acelerar, para uma escolha adequada, a convergência do método, em especial nos casos de sobre-relaxação com $\rho_k > 1$ para todo k. Para adequar a atualização dos termos p^k ao parâmetro de relaxamento, define-se

$$p^{k+1} := p^k + \lambda(\rho_k Ax^{k+1} - (1-\rho_k)(Bw^k - c) + Bw^{k+1} - c).$$

Teorema 2.1. *Considere o problema* (P) *e sejam dados* $x^0 \in \mathbb{R}^m$, $w^0 \in \mathbb{R}^n$, $p^0 \in \mathbb{R}^l$ *e* $\lambda \in \mathbb{R}$ *com* $\lambda > 0$. *Considere também as sequências* (μ_k), (ν_k) *e* (ρ_k) *de números reais tais que:*

- $\mu_k \geq 0$ *para todo* k *e* $\sum \mu_k < +\infty$.
- $\nu_k \geq 0$ *para todo* k *e* $\sum \nu_k < +\infty$.
- $\rho_k \in (0,2)$ *para todo* k *e* $0 < \liminf \rho_k \leq \limsup \rho_k < 2$.

Tome sequências (x^k), (w^k) *e* (p^k) *tais que para todo* k*:*

- $\|x^{k+1} - \tilde{x}^k\| \leq \mu_k$.
- $\|w^{k+1} - \tilde{w}^k\| \leq \nu_k$.
- $p^{k+1} = p^k + \lambda(\rho_k Ax^{k+1} - (1-\rho_k)(Bw^k - c) + Bw^{k+1} - c)$,

em que

$$\tilde{x}^k \in \underset{x\in\mathbb{R}^m}{\arg\min} \left\{ f(x) + \langle p^k, Ax\rangle + \frac{\lambda}{2}\|Ax + Bw^k - c\|^2 \right\},$$

$$\tilde{w}^k \in \underset{w\in\mathbb{R}^n}{\arg\min} \left\{ g(w) + \langle p^k, Bw\rangle + \frac{\lambda}{2}\|\rho_k Ax^{k+1} - (1-\rho_k)(Bw^k - c) + Bw - c\|^2 \right\}.$$

Então, se (P) *possui uma tripla Kuhn-Tucker, valem as afirmações*

1) p^k *converge para um multiplicador de Lagrange do problema* (P).

2) $Ax^k + Bw^k - c \to 0$.

3) $f(\tilde{x}^k) + g(\tilde{w}^k) \to \overline{\tau}$.

4) Toda subsequência convergente (x^{k_j}, w^{k_j}) *tem como ponto limite uma solução do problema* (P).

Demonstração. Veja [13, Teorema 3.3.1]. ■

Serão apresentados a seguir alguns corolários do Teorema 2.1. Tais resultados mostram algumas propriedades que as sequências (x^k), (w^k) e (p^k) passam a apresentar ao assumirmos algumas hipóteses adicionais. O primeiro resultado caracteriza a convergência de $f(x^k) + g(w^k)$ ao valor ótimo. O Teorema 2.1 estabelece que $f(\tilde{x}^k) + g(\tilde{w}^k)$ converge para $\overline{\tau}$, entretanto, não é possível garantir o mesmo para $f(x^k) + g(w^k)$ sem hipóteses adicionais.

Corolário 2.1. *Se* f *e* g *são funções reais, então,* $f(x^k) + g(w^k) \to \overline{\tau}$.

Demonstração. Veja [13, Corolário 3.3.1]. ■

Note que com a existência de uma tripla *Kuhn-Tucker* no Teorema 2.1, a sequência (p^k) sempre converge. Não é possível garantir o mesmo comportamento nas sequências (x^k) e (w^k), que podem não convergir, mas sim ter subsequências convergentes para uma solução de (P). Contudo, uma das maneiras de garantir a convergência das sequências (x^k) e (w^k) é supor que as matrizes A e B possuem posto coluna completo.

Corolário 2.2. *Suponha que as matrizes* A *e* B *possuam posto coluna completo. Dessa forma, se o problema* (P) *admite uma tripla Kuhn-Tucker, então, as sequências* (x^k) *e* (w^k) *convergem para uma solução do problema* (P). *Caso contrário, ao menos uma das sequências* (p^k) *ou* $(c - Bw^k)$ *é ilimitada.*

Demonstração. Veja [13, Corolário 3.3.2]. ■

Uma das dificuldades em utilizar os resultados do Teorema 2.1 em uma implementação prática do método ADMM inexato é verificar quando o critério de cada subproblema é satisfeito, isto é, quando são atendidas as condições

$$\|x^{k+1} - \tilde{x}^k\| \leq \mu_k, \quad \forall k = 0, 1, 2, ...,$$

$$\|w^{k+1} - \tilde{w}^k\| \leq \nu_k, \quad \forall k = 0, 1, 2, ...$$

Note que esses critérios são totalmente dependentes das soluções exatas $\tilde{x}^k$ e $\tilde{w}^k$. Como essas soluções são em geral desconhecidas, para utilizar desses critérios há a necessidade de adicionar hipóteses no problema (P) que garantem que estimativas das distâncias da solução ótima de cada subproblema possam ser feitas de alguma forma, mesmo desconhecendo tais soluções. Algumas dessas hipóteses são apresentadas em [5].

2.2 O critério HPE aplicado ao método ADMM

Os critérios inexatos apresentados no Teorema 2.1 são dependentes das soluções exatas. Isso dificulta a verificação desses critérios, inviabilizando a implementação prática do método para uma grande classe de problemas. Será apresentada nessa seção uma segunda versão do método ADMM, baseada em [18], que utiliza o critério inexato do método HPE apresentado na Seção 1.3. A desvantagem dessa abordagem é que ela demanda que um dos subproblemas seja resolvido de forma exata, o que torna o método menos geral no ponto de vista teórico em relação ao método estabelecido pelo Teorema 2.1.

A ideia do autor em [18] é tratar o subproblema na variável w em (7) como o de encontrar um zero de um operador apropriado. Essa perspectiva permite a aplicação do critério inexato utilizado no método HPE ao método ADMM. Como esse método é parcialmente inexato, um dos subproblemas deve ser escolhido para ser trabalhado de forma exata. Como em [18], será escolhido aqui o primeiro subproblema, na variável x, a fim de reduzir a propagação de erros.

Perceba que solucionar

$$w^{k+1} \in \underset{w \in \mathbb{R}^n}{\arg\min} \mathcal{L}_\lambda(x^{k+1}, w, p^k)$$

é equivalente a determinar w^{k+1} tal que

$$0 \in \partial_w \mathcal{L}_\lambda(x^{k+1}, w^{k+1}, p^k).$$

Portanto, trataremos o subproblema como o de encontrar um zero de

$$\mathcal{T}(w) = \partial_w \mathcal{L}_\lambda(\tilde{x}, w, \tilde{p}),$$

para $\tilde{x}$ e $\tilde{p}$ fixos.

As sequências geradas pelo método HPE aplicado ao operador $\mathcal{T}$ para um $\sigma \in (0, 1)$ seguem as regras

$$\begin{aligned} v^k &\in \mathcal{T}^{\varepsilon_k}(w^k) \\ \|\lambda v^k + w^k - \hat{w}^k\|^2 + 2\lambda\varepsilon_k &\leq \sigma^2 \|w^k - \hat{w}^k\|^2, \\ \hat{w}^{k+1} &= \hat{w}^k - \lambda v^k. \end{aligned} \tag{10}$$

Note que o tamanho de passo escolhido para o HPE é igual ao parâmetro de penalização λ da função Lagrangiana aumentada.

Lembre da Seção 1.3 que em (10), w^k é a aproximação do resolvente do operador $\mathcal{T}$. Já $\hat{w}^k$ é uma sequência auxiliar que aproveita da direção v^k calculada para aplicar um passo que segue a ideia do método extragradinte. Em [18], a sequência de soluções do segundo subproblema para o método ADMM é a sequência (w^k). Ela que será considerada para a atualização dos multiplicadores e para os subproblemas na variável x.

Analisaremos a condição

$$v^k \in \mathcal{T}^{\varepsilon_k}(w^k)$$

para o operador $\mathcal{T}$ escolhido. Observe que como a função g é convexa, própria e fechada, vale que

$$\begin{aligned} \mathcal{T}(w) &= \partial_w \mathcal{L}_\lambda(\tilde{x}, w, \tilde{p}) \\ &= \partial g(w) + B^t\tilde{p} + \lambda B^t(A\tilde{x} + Bw - c). \end{aligned} \tag{11}$$

Mostra-se facilmente que para uma função $h : \mathbb{R}^n \to (-\infty, +\infty]$ convexa e própria vale,

para todo $\varepsilon \geq 0$, a continência

$$\partial_\varepsilon h(x) \subset (\partial h)^\varepsilon(x), \quad \forall x \in \mathbb{R}^n, \tag{12}$$

isto é, o ε-subdiferencial de h é um subconjunto do ε-*enlargement* do subdiferencial de h [11, Proposição 2.3]. Combinando as relações (11) e (12), mostra-se que para todo $\varepsilon \geq 0$ vale a implicação

$$v \in \partial_\varepsilon g(w) \Rightarrow v + B^t\tilde{p} + \lambda B^t(A\tilde{x} + Bw - c) \in \mathcal{T}^\varepsilon(w).$$

Isso nos permite considerar, ao invés de (10), as regras

$$v^k \in \partial_{\varepsilon_k} g(w^k),$$

$$\|\lambda[v^k + B^t\tilde{p} + \lambda B^t(A\tilde{x} + Bw^k - c)] + w^k - \hat{w}^k\|^2 + 2\lambda\varepsilon_k \leq \sigma^2\|w^k - \hat{w}^k\|^2,$$

$$\hat{w}^{k+1} = \hat{w}^k - \lambda[v^k + B^t\tilde{p} + \lambda B^t(A\tilde{x} + Bw^k - c)].$$

A estratégia sugerida em [18] consiste em alternar entre a minimização na variável x, aplicar uma única vez o critério HPE na variável w e em seguida, atualizar o multiplicador. Essa abordagem é apresentada no Algoritmo 1. A análise de convergência desse algoritmo é apresentada em [18, Teorema 3.4].

2.3 O problema LASSO

Considere o problema LASSO *(Least Absolute Shrinkage and Selection Operator)*, dado por

$$\min_{x\in\mathbb{R}^n} \quad \frac{1}{2}\|Nx - b\|^2 + \alpha\|x\|_1, \tag{13}$$

em que $N \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$ e $\alpha > 0$. A resolução deste problema fornece a solução de mínimos quadrados com o maior número de zeros possível para o sistema

$$Nx = b.$$

Algoritmo 1: ADMM parcialmente inexato (HPE).

Entrada: $\lambda > 0$, $\sigma \in (0,1)$, $x^0 \in \mathbb{R}^m$, $w^0 \in \mathbb{R}^m$ e $p^0 \in \mathbb{R}^l$

$\hat{w}^0 = w^0$;

para *k=0,1,2,...* **faça**

- Determine x^{k+1} tal que

$$x^{k+1} \in \arg\min_{x \in \mathbb{R}^m} \left\{ f(x) + \langle p^k, Ax \rangle + \frac{\lambda}{2}\|Ax + Bw^k - c\|^2 \right\};$$

- Determine w^{k+1}, v^{k+1} e ε^{k+1} tais que:

$$v^{k+1} \in \partial_{\varepsilon_{k+1}} g(w^{k+1}),$$

$$\|\lambda[v^{k+1} + B^t p^k + \lambda B^t(Ax^{k+1} + Bw^{k+1} - c)] + w^{k+1} - \hat{w}^k\|^2 + \\ + 2\lambda\varepsilon_k \leq \sigma^2\|w^{k+1} - \hat{w}^k\|^2$$

- Faça

$$p^{k+1} = p^k + \lambda(Ax^{k+1} + Bw^{k+1} - c)$$
$$\hat{w}^{k+1} = \hat{w}^k - \lambda(v^{k+1} + B^t p^{k+1})$$

fim

Resolver problemas desse tipo tem grande importância em aplicações como aprendizado de máquina e reconstrução de imagem.

Para aplicar o método ADMM ao problema (13), precisamos primeiramente colocá-lo na forma do problema (P). Uma das maneiras de fazer isso é considerar o problema restrito:

$$\begin{aligned} \min_{x,w} \quad & \alpha\|x\|_1 + \frac{1}{2}\|Nw - b\|^2 \\ s.a \quad & x - w = 0. \end{aligned} \tag{14}$$

Fazendo as associações com o problema (P), temos:

- $f(x) = \alpha\|x\|_1$.
- $g(w) = \frac{1}{2}\|Nw - b\|^2$.
- $A = I$.
- $B = -I$.
- $c = 0$.

Analisaremos nesta seção a aplicação do método ADMM ao problema LASSO, tanto o método exato, quanto a sua versão parcialmente inexata. Aqui, $\text{prox}_{\alpha f}(x)$ denota o operador proximal de uma função $f : \mathbb{R}^n \to (-\infty, +\infty]$ avaliado no ponto $x \in \mathbb{R}^n$ em termos de um parâmetro $\alpha > 0$. Tal operador é definido por

$$\text{prox}_{\alpha f}(x) := \underset{y \in \mathbb{R}^n}{\arg\min} \left\{ f(y) + \frac{1}{2\alpha}\|y - x\|^2 \right\}.$$

Uma série de propriedades do operador proximal são apresentadas em [2, Capítulo 6].

2.3.1 Método exato

Nesta seção abordaremos o problema LASSO utilizando o método exato descrito em (7). Para este método, o subproblema na variável x consiste em resolver:

$$\begin{aligned} x^{k+1} &\in \underset{x \in \mathbb{R}^n}{\arg\min}\, \mathcal{L}_\lambda(x, w^k, p^k) \qquad \text{(forma escalada (6))} \\ &= \underset{x \in \mathbb{R}^n}{\arg\min} \left\{ \alpha\|x\|_1 + \frac{\lambda}{2}\|x - w^k + \frac{1}{\lambda}p^k\|^2 - \frac{1}{2\lambda}\|p^k\|^2 \right\} \\ &= \underset{x \in \mathbb{R}^n}{\arg\min} \left\{ \|x\|_1 + \frac{1}{2\frac{\alpha}{\lambda}}\|x - (w^k - \frac{1}{\lambda}p^k)\|^2) \right\} \\ &= \text{prox}_{\frac{\alpha}{\lambda}\|\cdot\|_1} \left(w^k - \frac{1}{\lambda}p^k \right). \end{aligned} \tag{15}$$

Define-se o operador de limiarização suave *(Soft-Thresholding)* $\mathcal{S}_\alpha : \mathbb{R} \to \mathbb{R}$ por

$$\mathcal{S}_\alpha(x) := \begin{cases} x + \alpha, \text{ se } x < -\alpha, \\ 0, \text{ se } -\alpha \leq x \leq \alpha, \\ x - \alpha, \text{ se } x > \alpha. \end{cases}$$

Quando $x \in \mathbb{R}^n$, denotaremos por $\mathcal{S}_\alpha(x)$ a aplicação coordenada a coordenada desse mesmo operador em x. Mostra-se em [2, Exemplo 6.8] que

$$\text{prox}_{\alpha\|\cdot\|_1}(x) = \mathcal{S}_\alpha(x).$$

Portanto, em (15), vemos que x^{k+1} é facilmente determinado por

$$x^{k+1} = \mathcal{S}_{\frac{\alpha}{\lambda}}\left(w^k - \frac{1}{\lambda}p^k\right)$$

Repare que não existe na relação anterior a pertinência x^{k+1} a um conjunto, que deveria ser herdada de (15). É possível usar a igualdade pois o conjunto dos minimizadores de $\mathcal{L}_\lambda(x, w^k, p^k)$ é formado por um único elemento, que pode ser calculado por meio do operador $\mathcal{S}_\alpha$.

A solução exata do segundo subproblema pode ser obtida com a minimização de uma quadrática. De fato, partindo de (7), abrindo a definição das normas e reagrupando alguns termos, temos

$$\begin{aligned} w^{k+1} &\in \arg\min_{w\in\mathbb{R}^n} \mathcal{L}_\lambda(x^{k+1}, w, p^k) \qquad \text{(forma escalada (6))} \\ &= \arg\min_{w\in\mathbb{R}^n} \left\{\frac{1}{2}\|Nw - b\|^2 + \frac{\lambda}{2}\|x^{k+1} - w + \frac{1}{\lambda}p^k\|^2 - \frac{1}{2\lambda}\|p^k\|^2\right\} \\ &= \arg\min_{w\in\mathbb{R}^n} \left\{\frac{1}{2}w^t(N^tN + \lambda I)w - (N^tb + \lambda x^{k+1} + p^k)^t w\right\} \end{aligned} \tag{16}$$

Dessa forma, w^{k+1} pode ser determinado com a resolução do sistema

$$(N^tN + \lambda I)w^{k+1} = N^tb + \lambda x^{k+1} + p^k.$$

Observe que a matriz $N^tN + \lambda I$ é simétrica e definida positiva, logo, o sistema linear que define w^{k+1} possui solução única. A simetria e a positividade de $N^tN + \lambda I$ são propriedades que podem ser bem exploradas em uma implementação prática do método. Como a matriz do sistema linear é constante em todas as iterações do método ADMM, é possível aplicar a fatoração de *Cholesky* uma única vez a $N^tN + \lambda I$ e então, reaproveitar essa fatoração para resolver os sistemas no decorrer das iterações do método.

Finalizada a discussão sobre as soluções dos subproblemas do ADMM para o caso do problema LASSO, tem-se que o método gera sequências (x^k), (w^k) e (p^k) a partir de pontos iniciais x^0, w^0 e p^0 dados tais que sejam válidas para todo $k = 0, 1, 2, ...$ as

igualdades

$$\begin{aligned} x^{k+1} &= \mathcal{S}_{\frac{\alpha}{\lambda}}\left(w^k - \frac{1}{\lambda}p^k\right), \\ w^{k+1} &= (N^tN + \lambda I)^{-1}(N^tb + \lambda x^{k+1} + p^k), \\ p^{k+1} &= p^k + \lambda(x^{k+1} - w^{k+1}). \end{aligned} \tag{17}$$

2.3.2 Método parcialmente inexato

Analisaremos a seguir a aplicação no problema LASSO do método ADMM parcialmente inexato com o critério HPE, apresentado na Seção 2.2. Repare que o primeiro subproblema é resolvido de forma exata no Algoritmo 1, portanto, manteremos a escolha

$$x^{k+1} = \mathcal{S}_{\frac{\alpha}{\lambda}}\left(w^k - \frac{1}{\lambda}p^k\right),$$

que foi feita anteriormente em (17) para a abordagem exata. Dessa forma, devemos trabalhar para determinar de uma maneira simples elementos w^{k+1}, v^{k+1} e ε_{k+1} que satisfazem as condições do critério inexato do Algoritmo 1, isto é, que satisfazem

$$v^{k+1} \in \partial_{\varepsilon_{k+1}} g(w^{k+1}),$$

$$\|\lambda[v^{k+1} - p^k - \lambda(x^{k+1} - w^{k+1})] + w^{k+1} - \hat{w}^k\|^2 + 2\lambda\varepsilon_k \leq \sigma^2\|w^{k+1} - \hat{w}^k\|^2.$$

Para simplificar a escrita, consideraremos x^{k+1}, $\hat{w}^k$ e p^k fixos respectivamente em $\tilde{x}$, $\hat{w}$ e $\tilde{p}$ e determinaremos w, v e ε tais que

$$v \in \partial_\varepsilon g(w), \tag{18}$$

$$\|\lambda[v - \tilde{p} - \lambda(\tilde{x} - w)] + w - \hat{w}\|^2 + 2\lambda\varepsilon \leq \sigma^2\|w - \hat{w}\|^2. \tag{19}$$

Observe que para o problema (14) temos que a função g é convexa, real e diferenciável. Isso implica que o subdiferencial $\partial g(x)$ é formado somente pelo gradiente de g, isto é

$$\partial g(w) = \{\nabla g(w)\}, \quad \forall w \in \mathbb{R}^n.$$

A demonstração dessa afirmação pode ser encontrada em [2, Teorema 3.33]. Além disso,

$\nabla g(w)$ pode ser facilmente calculado em todo $w \in \mathbb{R}^n$. Tal propriedade do subgradiente nos permite escolher

$$\begin{aligned} v &= \nabla g(w) \\ &= N^t(Nw - b). \end{aligned}$$

Como um elemento do subdiferencial pode ser facilmente calculado, não há a necessidade de trabalhar com o ε-subdiferencial. Dessa forma, buscaremos atender (18) e (19) com $\varepsilon = 0$ e $v = N^t(Nw-b)$, ou seja, tentaremos determinar um w que satisfaça a desigualdade

$$\|\lambda[N^t(Nw - b) - \tilde{p} - \lambda(\tilde{x} - w)] + w - \hat{w}\|^2 \leq \sigma^2 \|w - \hat{w}\|^2,$$

que é equivalente a

$$\lambda \left\| (N^t N + \lambda I)w - (N^t b + \lambda \tilde{x} + \tilde{p}) + \frac{1}{\lambda}(w - \hat{w}) \right\| \leq \sigma \|w - \hat{w}\|, \tag{20}$$

em que I denota a matriz identidade de dimensão n.

Suponha que w possa se escrito na forma

$$w = \hat{w} + u$$

para algum $u \in \mathbb{R}^n$ adequado. Substituindo essa expressão para w em (20), obtemos

$$\lambda \left\| \left(N^t N + \frac{\lambda^2 + 1}{\lambda} I \right) u + (N^t N + \lambda I)\hat{w} - (N^t b + \lambda \tilde{x} + \tilde{p}) \right\| \leq \sigma \|u\|. \tag{21}$$

Definindo

$$\begin{aligned} \mathcal{G}(w) &:= \nabla_w \mathcal{L}_\lambda(\tilde{x}, w, \tilde{p}) \\ &= N^t(Nw - b) - \tilde{p} + \lambda(\tilde{x} - w) \\ &= (N^t N + \lambda I)w - (N^t b + \lambda \tilde{x} + \tilde{p}), \end{aligned}$$

reescrevemos a desigualdade (21) como

$$\left\| \left(N^t N + \frac{\lambda^2 + 1}{\lambda} I \right) u + \mathcal{G}(\hat{w}) \right\| \leq \frac{\sigma}{\lambda} \|u\|. \tag{22}$$

Portanto, $w = \hat{w} + u$ respeita o critério (20) se, e somente se vale a condição (22) sobre u.

Mostraremos agora que existe um elemento u satisfazendo (22), e tal elemento pode ser facilmente determinado com a resolução de um sistema que pode ser solucionado de forma aproximada. Isso nos permitira encontrar um elemento w que respeita a condição (20). Basearemos essa discussão na resolução do sistema

$$\left(N^t N + \frac{\lambda^2 + 1}{\lambda} I \right) u = -\mathcal{G}(\hat{w}) \tag{23}$$

Note que a matriz do sistema linear (23) é simétrica e definida positiva, logo, vale que

- A solução $\hat{u}$ do sistema (23) existe e é única.
- A solução do sistema (23) é $\hat{u} = 0$ se, e somente se $\mathcal{G}(\hat{w}) = 0$.

No contexto em que a condição (22) é aplicada, é bem razoável assumir que $\mathcal{G}(\hat{w}) \neq 0$. Observe que resolver o subproblema diferenciável na variável w de forma exata é equivalente a determinar w tal que

$$\begin{aligned} 0 &= \nabla_w \mathcal{L}_\lambda(\tilde{x}, w, \tilde{p}) \\ &= \mathcal{G}(w). \end{aligned}$$

Dessa forma, se $\mathcal{G}(\hat{w}) = 0$, então $\hat{w}$ é a solução exata do subproblema e não há a necessidade de aplicar um critério inexato pois a solução já é conhecida. Assumiremos portanto que $\mathcal{G}(\hat{w}) \neq 0$.

Seja $\hat{u}$ a solução do sistema (23). Como por hipótese $\mathcal{G}(\hat{w}) \neq 0$, vale que $\hat{u} \neq 0$, o que por sua vez implica $\|\hat{u}\| > 0$. Por definição de $\hat{u}$,

$$\left\| \left(N^t N + \frac{\lambda^2 + 1}{\lambda} I \right) \hat{u} + \mathcal{G}(\hat{w}) \right\| = 0.$$

Isso mostra que $\hat{u}$ satisfaz a condição (22) e portanto,

$$w = \hat{w} + \hat{u}$$

satisfaz a condição (20).

Em um primeiro momento, parece não existir vantagem em aplicar o critério inexato que acabou de ser apresentado, visto que há a necessidade de se resolver um sistema linear semelhante ao sistema usado para determinar a solução exata w^{k+1} em (17). Entretanto, veremos que resolver esse sistema de forma exata não é a única maneira de atender a condição do método, como é em (17). Além disso, o critério para aceitar uma solução aproximada do sistema (23) é simples de ser verificado.

Seja (u^k) uma sequência tal que $u^k \to \hat{u}$, onde $\hat{u}$ é solução de (23). Relembramos que como estamos assumindo que $\mathcal{G}(\hat{w}) \neq 0$, então $\|\hat{u}\| > 0$, isso implica que

$$\lim_{k\to\infty} \|u^k\| > 0.$$

Como u^k converge para $\hat{u}$, vale que

$$\lim_{k\to\infty} \left\| \left(N^t N + \frac{\lambda^2 + 1}{\lambda} I \right) u^k + \mathcal{G}(\hat{w}) \right\| = 0.$$

Logo, deve existir um inteiro K tal que

$$k > K \Rightarrow \left\| \left(N^t N + \frac{\lambda^2 + 1}{\lambda} I \right) u^k + \mathcal{G}(\hat{w}) \right\| \leq \frac{\sigma}{\lambda} \|u^k\|.$$

Isso demonstra que a estratégia adequada é resolver o sistema de forma iterativa, até encontrar um termo u^k satisfazendo a condição (22), e então, tomar $u = u^k$ e definir

$$w = \hat{w} + u. \tag{24}$$

Com isso, w satisfaz a condição (20).

Comentaremos agora sobre o algoritmo que pode ser usado para determinar o u desejado. Como a matriz do sistema (23) é simétrica e definida positiva, é possível explorar dessa estrutura e aplicar o método de gradientes conjugados, que possui convergência

garantida para a solução do sistema linear em um número finito de iterações. A condição de parada para o método de gradientes conjugado deve ser atender a desigualdade (22). O método parcialmente inexato é apresentado no Algoritmo 2.

Algoritmo 2: ADMM parcial inexato LASSO.

Entrada: $\lambda > 0$, $\sigma \in (0,1)$, $x^0 \in \mathbb{R}^m$, $w^0 \in \mathbb{R}^m$ e $p^0 \in \mathbb{R}^l$

$\hat{w}^0 = w^0$;

para *k=0,1,2,...* **faça**

- Faça

$$x^{k+1} = \mathcal{S}_{\frac{\alpha}{\lambda}}\left(w^k - \frac{1}{\lambda}p^k\right);$$

- Calcule u^k aplicando o método de gradientes conjugados ao sistema

$$\left(N^tN + \frac{\lambda^2+1}{\lambda}I\right)u = -\mathcal{G}(\hat{w}^k),$$

com condição de parada (23);

- Faça

$$w^{k+1} = \hat{w}^k + u^k,$$
$$p^{k+1} = p^k + \lambda(x^{k+1} - w^{k+1}),$$
$$\hat{w}^{k+1} = \hat{w}^k - \lambda(N^t(Nw^{k+1} - b) - p^{k+1});$$

fim

Como o Algoritmo 2 é um caso particular do apresentado no Algoritmo 1, cuja convergência é demonstrada em [18, Teorema 3.4], valem os mesmos resultados.

3 Experimentos Computacionais

Apresentaremos aqui os experimentos computacionais de algumas variantes do método ADMM aplicados ao problema LASSO, apresentado na Seção 2.3. Serão abordadas duas variantes do método ADMM puro definido em (17). A primeira variante aborda o sistema que define w^{k+1} em (17) de maneira direta através da fatoração de *Cholesky* da matriz que define $N^tN + \lambda I$. Note que se λ é constante, a fatoração pode ser calculada uma única vez e então reutilizada ao longo das iterações. A segunda variante aborda o mesmo sistema, mas de forma iterativa com o método de gradientes conjugados. O erro na resolução do sistema pelo método iterativo deve ser pequeno o suficiente para mantê-lo equivalente ao método direto em termos do número de iterações. Será testado também o método inexato parcial com o critério do método HPE, definido no Algoritmo 2.

No Teorema 2.1, onde foi apresentado o resultado mais geral sobre a convergência do método ADMM inexato, foi introduzido um parâmetro de relaxamento ρ_k. Os subproblemas e a atualização do multiplicador a serem considerados nesse caso são

$$x^{k+1} \in \operatorname*{arg\,min}_{x\in\mathbb{R}^m} \left\{ f(x) + \langle p^k, Ax\rangle + \frac{\lambda}{2}\|Ax + Bw^k - c\|^2 \right\},$$

$$w^{k+1} \in \operatorname*{arg\,min}_{w\in\mathbb{R}^n} \left\{ g(w) + \langle p^k, Bw\rangle + \frac{\lambda}{2}\|\rho_k Ax^{k+1} - (1-\rho_k)(Bw^k - c) + Bw - c\|^2 \right\}.$$

$$p^{k+1} := p^k + \lambda(\rho_k Ax^{k+1} - (1-\rho_k)(Bw^k - c) + Bw^{k+1} - c).$$

Testaremos nessa seção os efeitos desse parâmetro quando ρ_k é escolhido constante em $\rho \in (0, 2)$ para todo k. A análise de desempenho para esse caso será feita no método ADMM exato e no método inexato parcial. A convergência do método exato com parâmetro de relaxamento constante é garantida pelo Teorema 2.1. Entretanto, ainda não existe um estudo de convergência que considere os efeitos desse parâmetro no método parcialmente inexato apresentado em [18]. Note que no caso do método relaxado, quando $\rho_k = 1$ para todo k, ele se converte no método ADMM original, definido em (7).

Os métodos serão testados em um conjunto de dados com 12 problemas, dos quais 10 são conjuntos de dados densos e os demais esparsos [4]. A Tabela 1 apresenta as

dimensões da cada problema para cada instância dos dados.

Problema	**Dimensão**
brain	42×5597
colon	62×2000
leukemia	72×3571
lymphoma	62×4026
prostate	102×6033
srbct	63×2308
Ball64	1638×4096
Logo64	1638×4096
Mug32	410×1024
Mug128	4770×16384
finance1000	30465×216842
PEMS	267×138672

Tabela 1: Dados utilizados e suas dimensões. Os problemas *finance1000* e *PEMS* possuem matrizes esparsas.

Os problemas *finance1000* e *PEMS* não podem ser solucionados pelo método ADMM exato direto. Isso ocorre pois não é possível armazenar na memória as matrizes $N^tN + \lambda I$ ou sua fatoração de Cholesky. Esses problemas só podem ser abordados pelo método exato iterativo, que usa o método de gradientes conjugados para resolver o sistema, e pelo método parcialmente inexato. Os métodos foram implementados em MATLAB versão R2018, e executados em um computador com sistema operacional Windows 10, processador Intel Core i7-10750H e 8GB de memória RAM.

Todos os testes tiveram como ponto inicial em x^0, w^0 e p^0 a origem e o parâmetro α associado ao problema LASSO foi considerado como

$$\alpha = \|N^t b\|_\infty.$$

Essa escolha para α é comum nas aplicações do problema LASSO e é sugerida em [4].

Analisaremos a seguir os critérios de parada adequados para cada uma das variantes do método ADMM. Nesse estudo, será possível observar o efeito do parâmetro de penalização λ na evolução do método. Essa análise permitirá o desenvolvimento de uma estratégia para adaptar o parâmetro λ ao longo das iterações do método. Tal estratégia também será testada.

3.1 Critérios de parada

Os critérios de parada escolhidos são deduzidos a partir das condições de *Kuhn-Tucker* para o problema (P), apresentadas na Definição 2.1. Eles são inspirados pelo critério de parada do método ADMM em sua verão mais pura, apresentado em [3]. Tais critérios foram escolhidos por serem fáceis de avaliar e independentes de custo computacional adicional no caso do problema LASSO. Desenvolveremos aqui os critérios para o caso geral do problema (P), e então, apresentaremos para o caso particular do problema LASSO.

Uma das três condições de *Kuhn-Tucker* para o problema (P) envolve a viabilidade do problema (P). Dessa forma, definiremos a sequências de resíduos primais (r^k) por

$$r^{k+1} := Ax^{k+1} + Bw^{k+1} - c.$$

A presença do resíduo primal será comum no critério de parada de qualquer um dos métodos testados aqui.

3.1.1 Critério de parada para o método exato

Analisaremos aqui o critério de parada para o método exato apresentado em (7). Note que pela definição de w^{k+1}, vale que

$$\begin{aligned} 0 &\in \partial g(w^{k+1}) + B^t \left[p^k + \lambda \left(Ax^{k+1} + Bw^{k+1} - c\right)\right] \\ &= \partial g(w^{k+1}) + B^t p^{k+1}. \end{aligned} \tag{25}$$

Dessa forma, a condição sobre a função g é sempre válida. Note que o mesmo não ocorre com a função f. Repare que tem-se a pertinência

$$\begin{aligned} 0 &\in \partial f(x^{k+1}) + A^t \left[p^k + \lambda \left(Ax^{k+1} + Bw^k - c\right)\right] \\ &= \partial f(x^{k+1}) + A^t \left[p^k + \lambda \left(Ax^{k+1} + Bw^{k+1} - c\right)\right] - \lambda A^t B \left(w^{k+1} - w^k\right) \\ &= \partial f(x^{k+1}) + A^t p^{k+1} - \lambda A^t B \left(w^{k+1} - w^k\right), \end{aligned} \tag{26}$$

em que a primeira igualdade foi obtida somando e subtraindo $\lambda A^t B w^{k+1}$. A segunda igualdade é obtida aplicando a definição de p^{k+1} dada em (7).

Repare que a relação (26) implica que

$$\lambda A^t B \left(w^{k+1} - w^k\right) \in \partial f(x^{k+1}) + A^t p^{k+1}, \tag{27}$$

o que nos leva a definir o resíduo dual s_f^k como

$$s_f^{k+1} := \lambda A^t B \left(w^{k+1} - w^k\right).$$

Como queremos satisfazer as condições da Definição 2.1, paramos o ADMM quando

$$\|r^k\| \leq \varepsilon$$
$$\|s_f^k\| \leq \varepsilon,$$

para algum $\varepsilon > 0$ adequado.

Para o caso do problema LASSO, os resíduos são:

$$r^{k+1} = x^{k+1} - w^{k+1},$$

$$s_f^{k+1} = \lambda(w^k - w^{k+1}).$$

Analisaremos de maneira breve o efeito do parâmetro de penalização λ na variação dos resíduos r^k e s_f^k. Pela definição dos subproblemas em (7), vemos que quanto maior o parâmetro λ, maior é a importância da viabilidade na resolução de cada subproblema, o que tende a fazer com que r^k decaia mais rapidamente. Contudo, um valor alto de λ é prejudicial para o resíduo s_f^k, que tende a crescer quando λ cresce. Esse comportamento indica que alterar o valor de λ em função da variação dos resíduos pode ser uma boa estratégia. Essa estratégia é apresentada em [3], onde podem ser encontradas referências para os trabalhos que aplicaram esta técnica. Ao variar o parâmetro de penalização, considera-se $\lambda_k > 0$ nas iterações em (7). Uma das escolhas possíveis é (cf. [3]):

$$\lambda_{k+1} := \begin{cases} 2\lambda_k, & \text{se } \|r^k\| > 10\|s_f^k\| \\ \lambda_k/2, & \text{se } \|s_f^k\| > 10\|r^k\| \\ \lambda_k, & \text{caso contrário.} \end{cases} \tag{28}$$

As relações entre os resíduos que implicam na atualização do valor do parâmetro λ em (28) são exemplificadas na Figura 1.

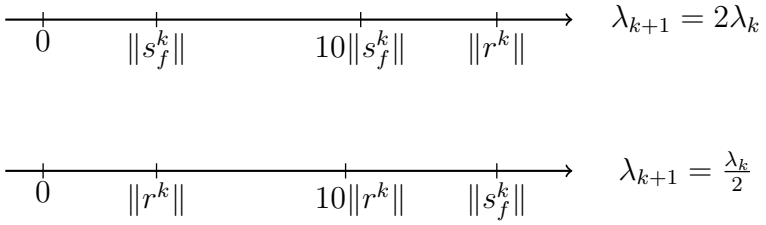

Figura 1: Relações entre os resíduos primal e dual que implicam atualização no valor de λ.

Critério de parada ao considerar o parâmetro de relaxamento ρ_k

O critério para o método ADMM exato com relaxamento é semelhante ao critério do método em sua versão pura e ele é deduzido da mesma forma. Veja que como foi feito em (25), podemos fazer considerando ρ_k e manter a pertinência

$$0 \in \partial g(w^{k+1}) + B^t p^{k+1}.$$

Em relação à condição sobre a função f, utilizando o mesmo tipo de manipulação feita em (26), somando e subtraindo termos convenientes a fim de expressar a condição em termos de p^{k+1}, obtemos

$$\lambda A^t \left[B(w^{k+1} - w^k) - (1-\rho_k)(Ax^{k+1} + Bw^k - c)\right] \in \partial f(x^{k+1}) + A^t p^{k+1}. \tag{29}$$

Dessa forma, para o método relaxado define-se o resíduo dual por

$$s_f^{k+1} := \lambda A^t \left[B(w^{k+1} - w^k) - (1-\rho_k)(Ax^{k+1} + Bw^k - c)\right].$$

Como anteriormente, o critério de parada para esta versão do método será

$$\|r^k\| \leq \varepsilon,$$
$$\|s_f^k\| \leq \varepsilon,$$

para $\varepsilon > 0$ dado.

Para o caso do problema LASSO, segue que

$$r^{k+1} = x^{k+1} - w^{k+1},$$

$$s_f^{k+1} := \lambda \left[w^k - w^{k+1} - (1 - \rho_k)(x^{k+1} - w^k) \right].$$

3.1.2 Critério de parada para o método parcialmente inexato

Para o método parcialmente inexato, com inexatidão presente na resolução do segundo subproblema, vale a pertinência (27) para o método sem relaxamento e (29) quando o relaxamento é considerado. A validade dessas pertinências decorre do fato de que o subproblema na variável x permanece sendo resolvido de maneira exata. Todavia, não é possível garantir a validade de (25), já que o subproblema na variável w está sendo solucionado de maneira aproximada. Isto implica que não existe um resíduo somente na condição de *Kuhn-Tucker* sobre função f, mas que existe um também na condição sobre a função g.

A ideia em (27) e (29) é determinar de maneira fácil, sem custo computacional adicional, um elemento de

$$\partial f(x^{k+1}) + A^t p^{k+1},$$

e medir o desvio da condição de otimalidade através da norma desse elemento. Nota-se aqui a importância de garantir que a norma desse elemento deve ter 0 como limite. O processo será semelhante para trabalharmos com a condição na função g, isto é, determinaremos u^{k+1} tal que

$$u^{k+1} \in \partial g(w^{k+1}) + B^t p^{k+1}.$$

Como comentado, não é possível considerar um elemento u^{k+1} qualquer de $\partial g(w^{k+1}) +$

$B^t p^{k+1}$ pois dessa forma não há como garantir que

$$\|u^k\| \to 0.$$

Escolheremos então o elemento de $\partial g(w^{k+1}) + B^t p^{k+1}$ que possui a menor norma, ou seja

$$u^{k+1} \in \underset{u \in \mathbb{R}^n}{\arg\min} \left\{ \|u\| \mid u \in \partial g(w^{k+1}) + B^t p^{k+1} \right\}, \tag{30}$$

e definiremos

$$s_g^{k+1} = u^{k+1}.$$

Determinar o elemento u^{k+1} adequado é uma tarefa simples em alguns casos, como por exemplo quando g é diferenciável.

As definições para r^k e s_f^k permanecem idênticas às definições apresentadas para o método exato. Seguem as definições para os resíduos:

- Sem parâmetro de relaxamento:

$$r^{k+1} := Ax^{k+1} + Bw^{k+1} - c,$$

$$s_f^{k+1} := \lambda A^t B \left(w^{k+1} - w^k \right),$$

$$s_g^{k+1} = u^{k+1},$$

- Com parâmetro de relaxamento:

$$r^{k+1} := Ax^{k+1} + Bw^{k+1} - c,$$

$$s_f^{k+1} := \lambda A^t \left[B(w^{k+1} - w^k) - (1 - \rho_k)(Ax^{k+1} + Bw^k - c) \right],$$

$$s_g^{k+1} = u^{k+1}.$$

Para ambos, u^{k+1} é definido como em (30).

O critério de parada será

$$\|r^k\| < \varepsilon,$$
$$\|s_f^k\| < \varepsilon,$$
$$\|s_g^k\| < \varepsilon.$$

Para o problema LASSO tem-se, sem considerar relaxamento, que

$$r^{k+1} := x^{k+1} - w^{k+1},$$

$$s_f^{k+1} := \lambda \left(w^k - w^{k+1}\right),$$

$$s_g^{k+1} = N^t(Nw^{k+1} - b) - p^{k+1}.$$

Se o parâmetro ρ_k está envolvido, então

$$s_f^{k+1} := \lambda \left(w^k - w^{k+1} - (1-\rho_k)(x^{k+1} - w^k)\right).$$

Para todos os métodos que serão testados aqui, será considerado

$$\varepsilon = 1 \times 10^{-4}.$$

3.2 Gradientes Conjugados

Como comentado nos capítulos anteriores, alguns sistemas lineares foram abordados pelo método dos gradientes conjugados. A teoria envolvendo este método iterativo para sistemas lineares pode ser encontrada em [12, Capítulo 5.1]. O método implementado é o apresentado em [12, Algoritmo 5.2].

No método exato, o sistema a ser resolvido é

$$(N^tN + \lambda I)w = N^tb + \lambda x^{k+1} + p^k.$$

Considerando w^j os termos gerados pelo método dos gradientes conjugados, o critério de

parada é

$$\|(N^tN + \lambda I)w^j - (N^tb + \lambda x^{k+1} + p^k)\| \leq \xi. \tag{31}$$

para algum $\xi > 0$ apropriado. Considera-se neste trabalho $\xi = 1 \times 10^{-8}$

Para o método inexato parcial baseado na condição do HPE, o sistema a ser resolvido é o da Equação (23) e a condição de parada é a Desigualdade (22).

Como o método de gradientes conjugados é utilizado dentro do método ADMM, usou-se como estimativa inicial de solução do sistema na iteração $k+1$ do ADMM a solução obtida do sistema anterior, isto é, a solução obtida pelo método de gradientes conjugados no sistema da iteração k.

3.3 Parâmetros

Nesta seção será realizada uma análise sobre o efeito da variação dos parâmetros σ e ρ_k, quando ρ_k é constante e igual a $\rho \in (0, 2)$, nos métodos apresentados. Uma análise criteriosa do parâmetro λ é complexa pois seu valor ideal é altamente dependente dos dados do problema a ser resolvido. Por outro lado, valores adequados de σ e ρ podem ser determinados de forma simples e a correlação desses parâmetros com o problema em questão é baixa.

Com essa discussão em mente, serão analisados o efeito dos parâmetros σ e ρ em 4 problemas dentre os 12 apresentados na Tabela 1, sendo eles: *brain*, *srbct*, *Mug32* e *Logo64*. Baseado nesse estudo, serão definidos os valores adequados desses parâmetros e para a comparação final dos métodos, uma escolha adequada para o valor de λ será feita para cada problema.

3.3.1 Método exato

O método ADMM exato em sua versão original possui como único parâmetro o de penalização. Esse parâmetro, como comentado anteriormente, é dependente do problema que está sendo resolvido, assim, a sua escolha será tratada separadamente.

Nesta seção será tratada primeiramente a equivalência entre o ADMM exato direto, onde o sistema que define w^{k+1} na Equação (17) é resolvido utilizando a fatoração de *Cholesky* da matriz $N^tN + \lambda I$, e o ADMM exato iterativo, onde o mesmo sistema é

resolvido pelo método de gradientes conjugados (GC). O critério de parada para o método de gradientes conjugados neste caso é dado na Equação (31), com

$$\xi = 1 \times 10^{-8}.$$

A Figura 2 representa o total de iterações para cada um dos métodos. Os testes foram feitos para $\lambda \in \{1, 2, 3, 4, 5\}$ e $\rho \in \{0.6, 1, 1.4, 1.8\}$. O parâmetro $\rho = 1$ foi um dos testados por converter o método relaxado no método sem relaxamento. Os marcadores $\circ$, acompanhado do gráfico de linhas, representam o total de iterações do método exato direto, enquanto os marcadores $\times$ representam o total de iterações do método iterativo.

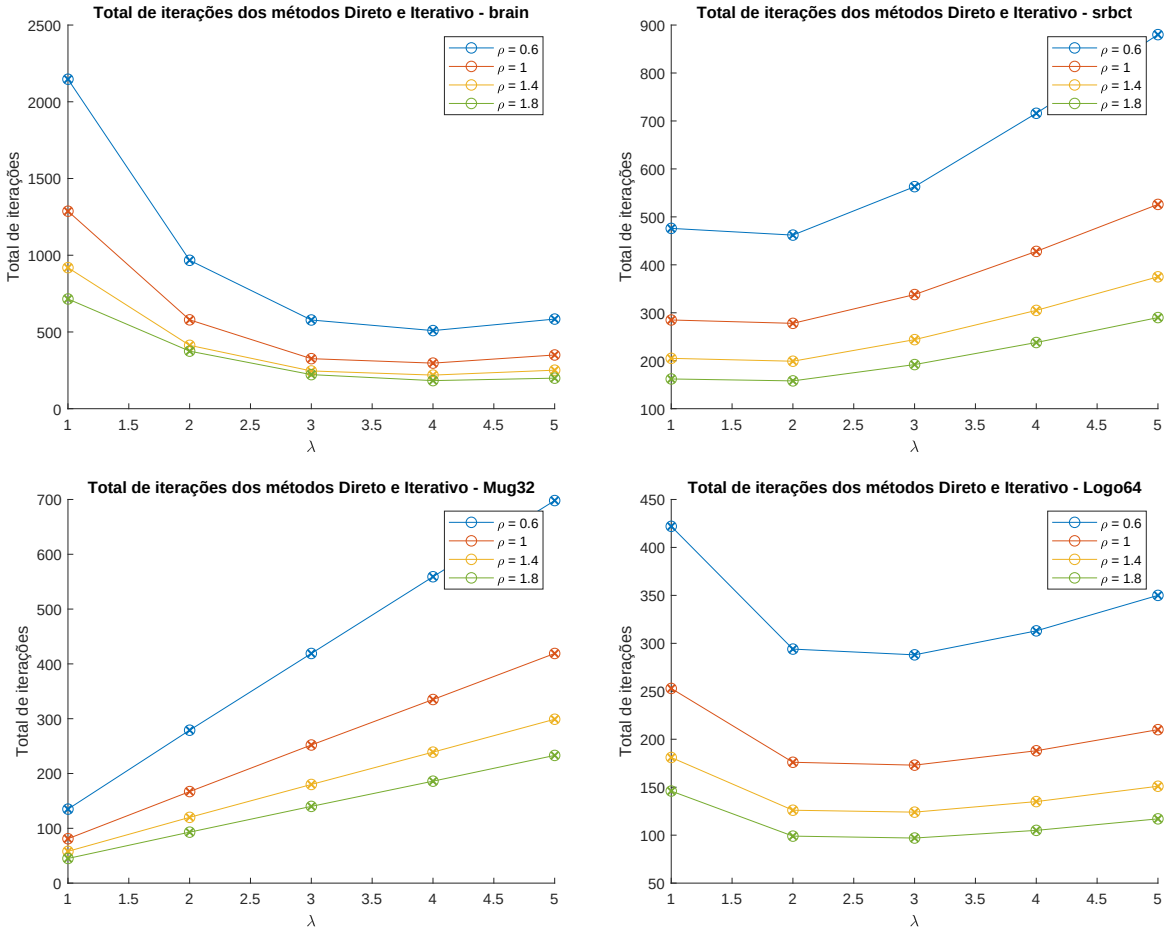

Figura 2: Total de iterações do método ADMM em suas variantes direta e iterativa em função do parâmetro λ. Cada cor representa uma escolha para o parâmetro ρ. Os métodos foram testados para $\lambda \in \{1, 2, 3, 4, 5\}$ e $\rho \in \{0.6, 1, 1.4, 1.8\}$. Os marcadores $\circ$ representam o desempenho da variante direta, enquanto os marcadores $\times$ representam o da variante iterativa.

.

Como é possível perceber, para todos os problemas e todas as combinações de parâmetros λ e ρ testadas, o total de iterações de ambos os métodos é idêntica. Contudo, o método em sua variante iterativa teve uma execução mais rápida na maioria dos problemas, além dele poder ser implementado de forma a lidar com problemas de grande porte, o que não é possível para o método direto, portanto, consideraremos somente a variante iterativa do método ADMM exato para os próximos testes.

Parâmetro ρ

Testaremos aqui os efeitos de trabalhar com o parâmetro de relaxamento ρ_k constante em $\rho \in (0, 2)$. Como mencionado em [3], espera-se que a sobre-relaxação com $\rho > 1$ tenha efeitos positivos para o método, acelerando a sua convergência. A Figura 3 apresenta o total de iterações do método ADMM exato em função da variação do parâmetro ρ para diversos valores do parâmetro de penalização λ. Na Figura 3, foi considerado ρ variando de 0.1 até 1.9 em intervalos de comprimento 0.3.

Nota-se claramente que a sobre-relaxação com $\rho \leq 1.9$ de fato funcionou bem para os problemas testados. Repare também que, pelo comportamento dos gráficos na Figura 3, a escolha de um valor ideal para ρ no intervalo testado não apresenta dependência do parâmetro λ, isto é, podemos afirmar com certeza que independentemente do valor de λ, $\rho = 1.9$ aprimora o desempenho do método.

Com base no sucesso da sobre-relaxação com $\rho \in [0.1, 1.9]$ apresentado na Figura 3, espera-se que valores mais próximos de 2 para ρ possam reduzir um pouco mais as iterações do método. Contudo, isso não ocorre. Quando ρ se aproxima de 2, o total de iterações do método ADMM exato tende a subir novamente. Esse fenômeno é apresentado na Figura 4 para o problema *Logo64*. Os demais problemas testados apresentaram o mesmo tipo de comportamento.

Como $\rho = 1.9$ teve o melhor desempenho nos problemas testados com o método ADMM exato, as versões relaxadas deste método serão executadas com

$$\rho_k = 1.9, \quad \forall k = 0, 1, 2, ...$$

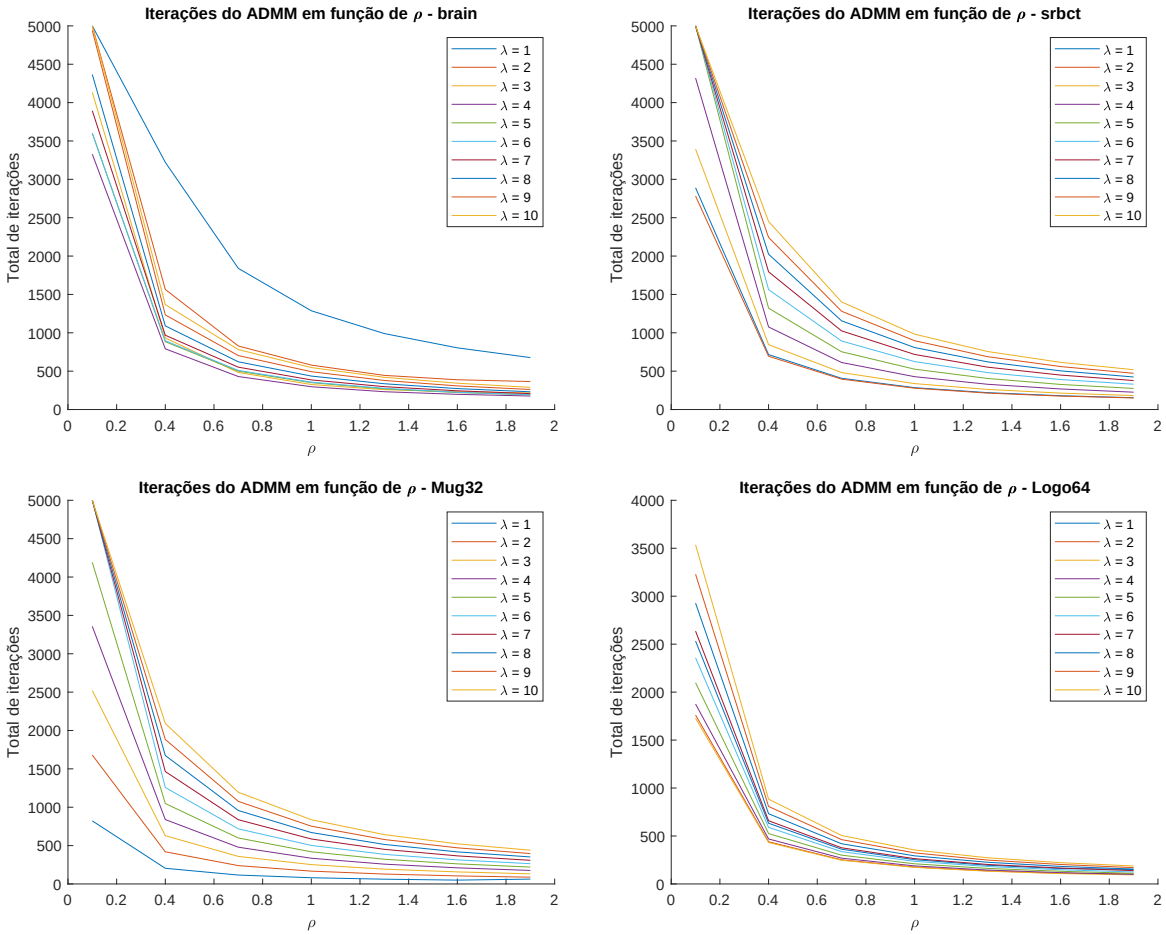

Figura 3: Total de iterações do método ADMM exato aplicado aos quatro problemas teste para ρ entre 0.1 e 1.9 em intervalos de comprimento 0.3. O parâmetro λ varia de 1 até 10.

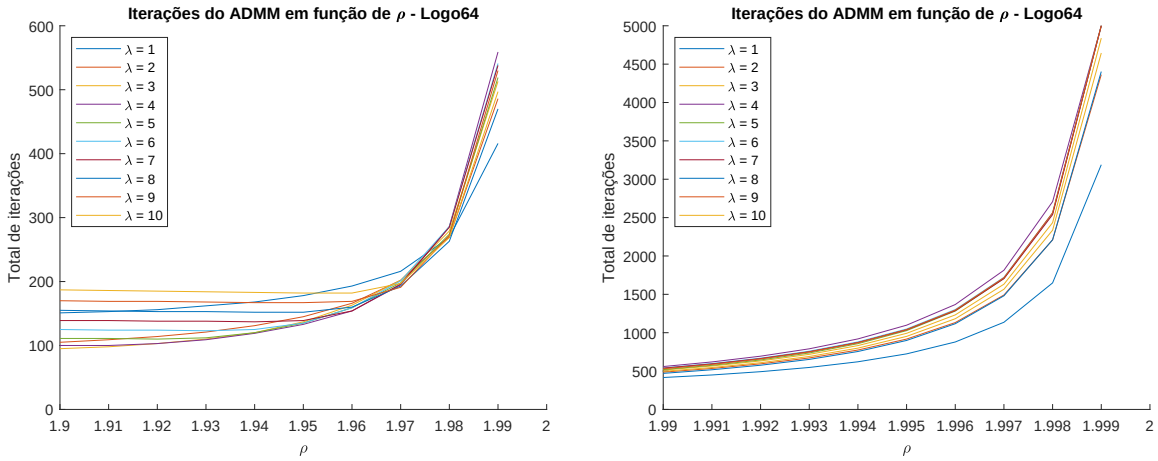

Figura 4: Efeito do parâmetro ρ no total de iterações do método ADMM exato aplicado ao problema *Logo64* quando $\rho \to 2$.

3.3.2 Método parcialmente inexato

O método ADMM parcialmente inexato possui originalmente os parâmetros λ e σ. Como já comentado, o parâmetro de penalização possui dependência dos dados do problema e a sua configuração será realizada separadamente. Embora sem apoio teórico, os efeitos da adição do parâmetro de relaxamento ρ_k constante serão analisados.

Parâmetro σ

O parâmetro $\sigma \in (0, 1)$ do método parcial inexato apresentado no Algoritmo 2 tem como único papel aliviar ou restringir a condição (22), controlando o quanto podemos desviar da solução do sistema (23). Se σ se aproxima de zero, exige-se mais precisão ao resolver o sistema. Caso contrário, uma solução mais relaxada para o sistema (23) é aceitável.

A Tabela 2 apresenta o total de iterações do método ADMM inexato, bem como o total de iterações internas necessárias para aproximar a solução do sistema (23) usando o critério de parada (22) para o método de gradientes conjugados. Para os resultados apresentados na Tabela 2 foi utilizado $\lambda = 1$ e não foi considerado o parâmetro de relaxamento.

	brain		***srbct***		***Mug32***		***Logo64***	
σ	**ADMM**	**GC**	**ADMM**	**GC**	**ADMM**	**GC**	**ADMM**	**GC**
0.1	1245	12721	339	3658	137	273	273	535
0.2	1245	11307	339	3127	137	205	272	502
0.3	1245	10825	339	2868	137	185	272	488
0.4	1245	10453	339	2758	139	175	272	478
0.5	1245	10270	346	2746	140	171	272	467
0.6	1245	10139	350	2728	140	167	274	463
0.7	1245	9981	354	2740	140	164	274	457
0.8	1245	9933	351	2689	145	168	274	449
0.9	1245	9886	358	2728	143	164	273	428

Tabela 2: Total de iterações do método ADMM e do método de gradientes conjugados (GC) necessárias para resolver cada problema teste com $\lambda = 1$, sem parâmetro de relaxamento.

Para $\lambda = 1$, nota-se que a variação do parâmetro σ não tem influência significativa no total de iterações do método ADMM para a maior parte dos problemas. Contudo, aumentar o valor de σ para próximo de 1 reduz o total de iterações do método

de gradientes conjugados ao resolver os sistemas lineares envolvidos. A Tabela 3 apresenta a média de iterações do método de gradientes conjugados por iteração do método ADMM para os quatro problemas selecionados. O mesmo padrão observado nas Tabelas 2 e 3 ocorre ao utilizar outros valores para o parâmetro de penalização λ e também ao considerar o parâmetro de relaxamento ρ.

σ	***brain***	***srbct***	***Mug32***	***Logo64***
0.3	8.69	8.46	1.35	1.79
0.6	8.14	7.79	1.19	1.69
0.9	7.94	7.62	1.15	1.57
0.93	7.93	7.64	1.16	1.57
0.96	7.92	7.67	1.15	1.56
0.99	7.89	7.65	1.14	1.56
0.993	7.91	7.67	1.15	1.57
0.996	7.93	7.63	1.15	1.57
0.999	7.94	7.63	1.14	1.57

Tabela 3: Média de iterações do método de gradientes conjugados com $\lambda = 1$.

Tendo em vista os dados das Tabelas 2 e 3, escolheremos para o método inexato baseado no HPE o valor

$$\sigma = 0.9.$$

Parâmetro ρ

Em [18], não existe a garantia de convergência para uma solução do problema (P) do ADMM parcialmente inexato com parâmetro de relaxamento ρ, introduzido ao ADMM no Capítulo 2.1. Contudo, testaremos seus efeitos quando aplicado a este método no problema LASSO. Consideraremos novamente, $\rho_k = \rho$ constante, com $\rho \in (0, 2)$.

Como já observado anteriormente para o método exato, bem como comentado em [3], espera-se que a sobre-relaxação acelere a convergência do método. Esse fenômeno também ocorre com o método parcialmente inexato e está representado na Figura 5, que apresenta o total de iterações do método ADMM em função da variação do parâmetro ρ para cada um dos problemas teste e para alguns valores do parâmetro λ.

Embora a sobre-relaxação reduza o total de iterações do método ADMM, é possível notar um ligeiro aumento na média de iterações de gradientes conjugados. Todavia, este aumento não é significativo para o desempenho global do método, já que o

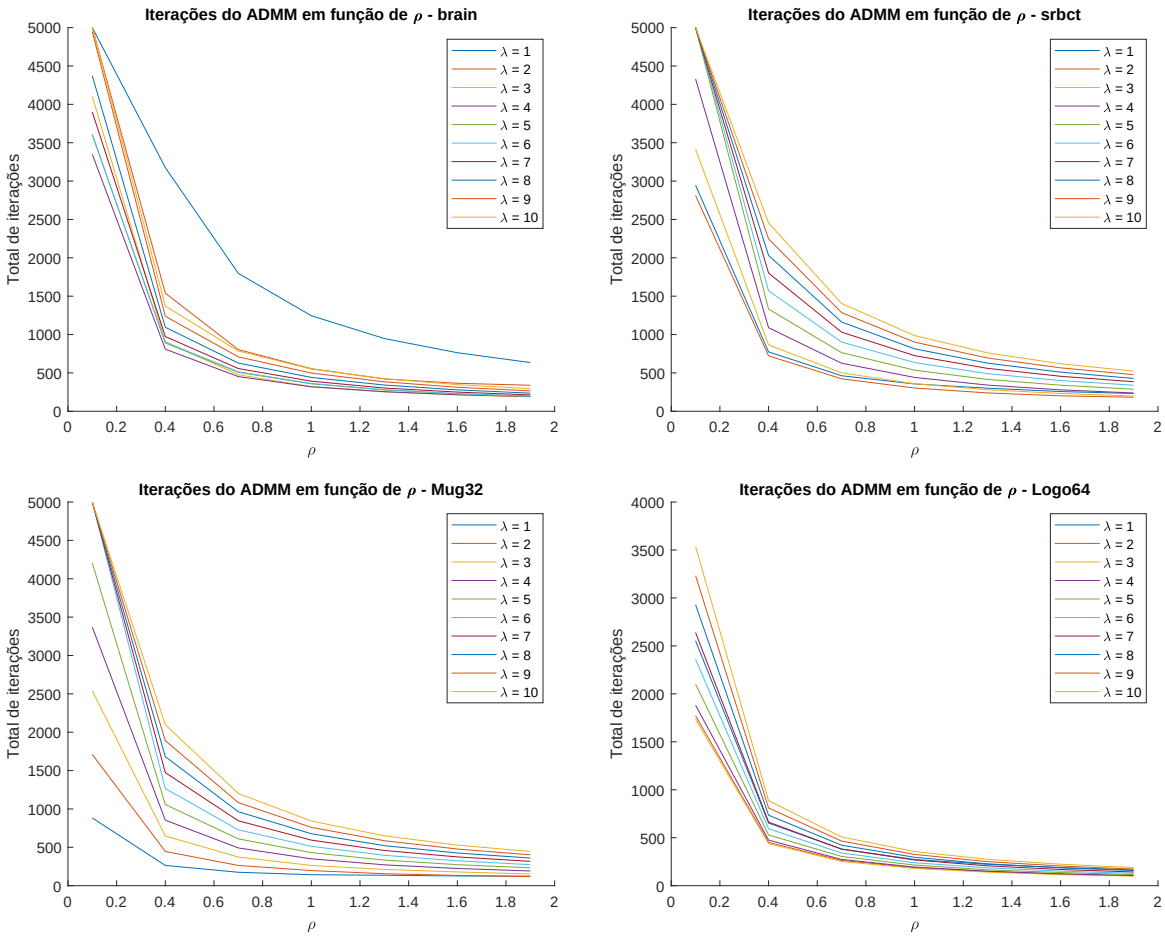

Figura 5: Total de iterações do método ADMM aplicado aos quatro problemas teste para ρ entre 0.1 e 1.9 em intervalos de comprimento 0.3. Os parâmetros λ variam de 1 até 10.

número de iterações do método ADMM diminui de forma significativa, inclusive reduzindo o total de iterações do método de gradientes conjugados. Isso está representado na Figura 6 para o problema *Mug32*. Para os demais problemas, o comportamento é idêntico. A variação da norma dos resíduos r^k, s_f^k e s_g^k, definidos na Seção 3.1.2, é apresentada na Figura 7.

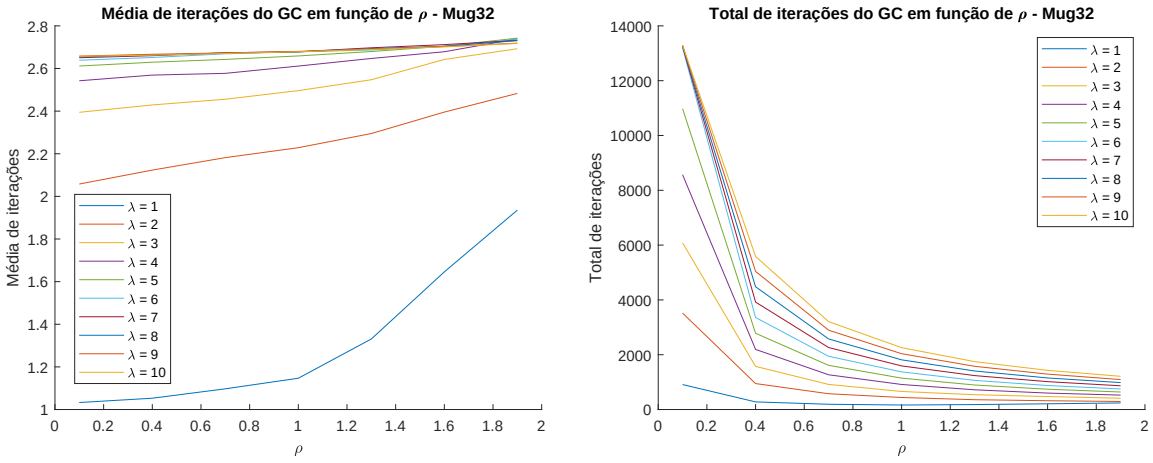

Figura 6: Média e total de iterações do método de gradientes conjugados em função do parâmetro ρ no problema *Mug32* para λ inteiro no intervalo $[1, 10]$.

Com indícios da melhora do desempenho quando $\rho \to 2$, foram realizados testes para alguns valores desse parâmetro no intervalo $[1.9, 1.999]$. Para $\rho \in [1.9, 1.99]$, foi observada uma leve queda no número de iterações do método ADMM inexato nos 4 problemas. Para $\rho \in [1.99, 1.999]$ não existe melhora significativa no desempenho do método em nenhum dos problemas em relação a $\rho \in [1.9, 1.99]$. Para o método ADMM parcial inexato aplicado ao problema LASSO, escolheremos portanto

$$\rho_k = 1.999, \quad \forall k = 0, 1, 2, ...$$

3.4 Comparações

Para as comparações finais, os métodos foram executados para diversos valores do parâmetro λ e o valor considerado como melhor foi aquele que permitiu o menor número possível de iterações do método. Com exceção dos problemas *Mug128*, *finance1000* e *PEMS*, os valores de λ testados vão de 0.5 até 10 em intervalos de comprimento 0.5. Para os problemas *Mug128* e *finance1000*, foi considerado λ entre 4 e 8 em intervalos

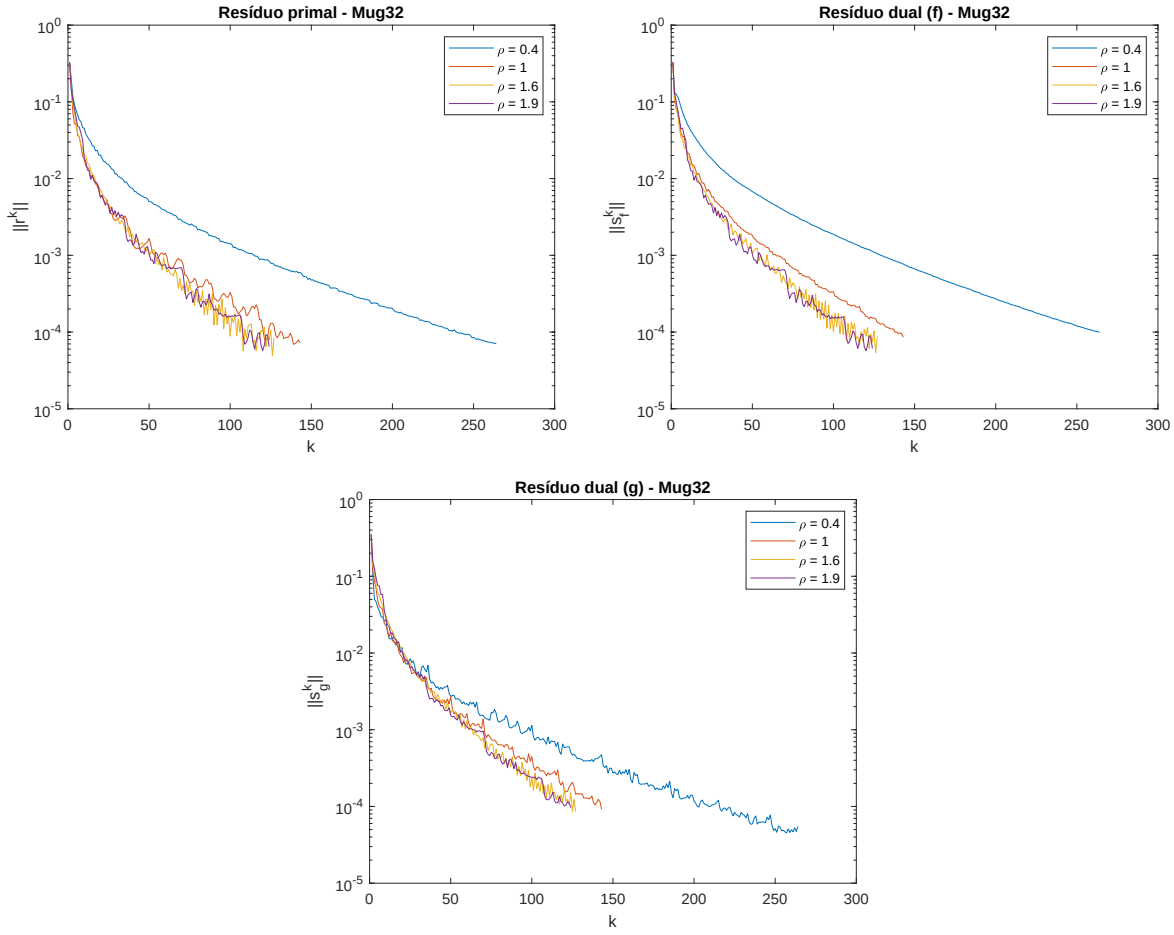

Figura 7: Variação da norma dos resíduos r^k, s_f^k e s_g^k ao longo das iterações do método ADMM inexato parcial aplicado ao problema *Mug32* para $\rho \in \{0.4, 1, 1.6, 1.9\}$ e $\lambda = 1$

.

de comprimento 0.5. Para o problema *PEMS*, foi considerado λ entre 5 e 7, também em intervalos de comprimento 0.5. Os testes foram feitos para os métodos com e sem o parâmetro de relaxamento ρ.

A Tabela 4 apresenta o total de iterações de cada método e entre parênteses como um par ordenado, o valor λ_{best} correspondente, seguido do total de iterações internas do método de gradientes conjugados, com os critérios de parada adequados a cada método.

	Sem relaxamento		**Com relaxamento**	
Problema	**Exato**	**Parc. Inexato**	**Exato**	**Parc. Inexato**
brain	282 (3.5, 6125)	285 (3.5, 3511)	162 (3.5, 3586)	164 (3.5, 2208)
colon	255 (1.5, 7157)	290 (1.5, 2764)	142 (1.5, 4109)	172 (1.5, 1887)
leukemia	347 (2, 8706)	361 (4.5, 4411)	190 (2, 4898)	193 (4, 2468)
lymphoma	376 (1.5, 11187)	494 (1.5, 6150)	221 (1.5, 6766)	323 (1.5, 4591)
prostate	321 (4.5, 8353)	328 (4, 4153)	174 (5, 4560)	171 (5, 2202)
srbct	259 (1.5, 8501)	300 (1.5, 2938)	149 (1.5, 5013)	178 (2, 2194)
Ball64	166 (2.5, 959)	175 (2.5, 461)	93 (3, 688)	93 (3, 281)
Logo64	171 (2.5, 983)	180 (2.5, 470)	95 (3, 691)	97 (3, 289)
Mug32	47 (0.5, 906)	143 (1, 164)	59 (0.5, 926)	103 (1.5, 246)
Mug128	285 (6, 1326)	287 (6.5, 924)	158 (6, 891)	159 (6.5, 524)
finance1000	114 (5.5, 6245)	114 (5.5, 2692)	96 (8, 4219)	102 (7.5, 2424)
PEMS	1558 (6.5, 67939)	1569 (6.5, 46933)	824 (6.5, 37484)	794 (6.5, 25042)

Tabela 4: Comparação entre os métodos trabalhados.

Uma propriedade notável do método ADMM inexato parcial é a sua capacidade de manter o total de iterações muito próximo do valor total apresentado pelo método exato. No problema *finance1000*, o total de iterações do método ADMM sem relaxamento é igual nas versões exata e inexata, mesmo com a versão inexata apresentando um total de iterações internas menor que a metade do total apresentado pela versão exata. O mesmo

ocorre para o *Ball64* ao considerar relaxamento. Esse comportamento é evidenciado na Figura 8, onde é possível observar que a média de iterações do ADMM exato e parcialmente inexato, exibido no gráfico à esquerda, são muito próximas considerando ou não o relaxamento. Mas a média de iterações internas de gradientes conjugados, exibida no gráfico à direita, é muito menor para o método parcialmente inexato.

Repare que para todos os testes, o método inexato atingiu uma quantidade muito menor de iterações internas. Como a etapa de maior custo computacional presente nos métodos é determinar a solução do sistema por gradientes conjugados, o método inexato tem toda a vantagem.

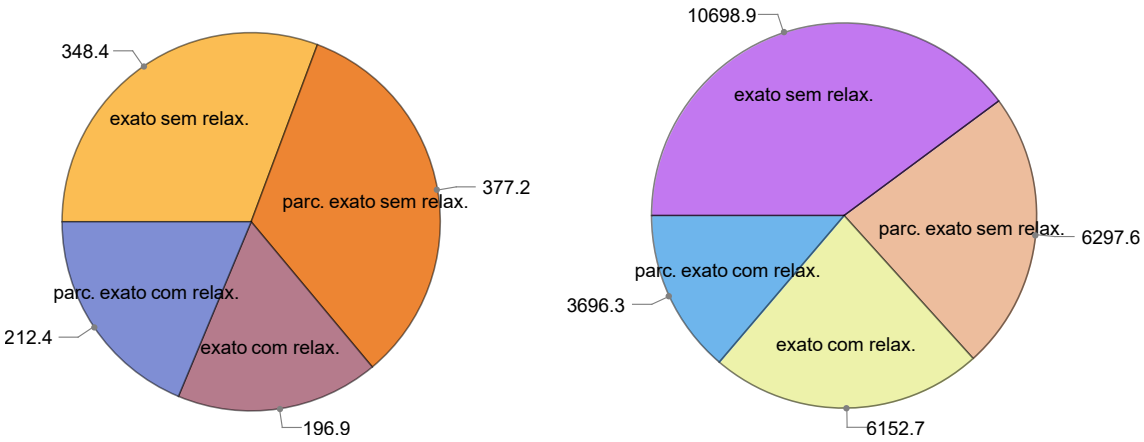

Figura 8: Médias das iterações efetuadas em cada método (esq.) e das iterações internas de gradientes conjugados (dir.) para os resultados apresentados na Tabela 4.

3.4.1 Parâmetro de penalização variável

Nessa seção, os métodos foram testados novamente, mas dessa vez com o parâmetro de penalização λ variável dado por (28), como sugerido em [3, Seção 3.4.1]. Para o método parcialmente inexato, como existe o resíduo dual associado à função g, ele também deve ser levado em conta na atualização de λ. Para tal, foi considerado no

método inexato

$$\lambda_{k+1} := \begin{cases} 2\lambda_k, & \text{se } \|r^k\| > 10\max\{\|s_f^k\|, \|s_g^k\|\} \\ \lambda_k/2, & \text{se } \max\{\|s_f^k\|, \|s_g^k\|\} > 10\|r^k\| \\ \lambda_k, & \text{caso contrário.} \end{cases}$$

A Tabela 5 apresenta o total de iterações do método ADMM, seguido do total de iterações internas do método de gradientes conjugados entre parênteses. Todos os testes foram realizados com $\lambda_0 = 1$ e os demais parâmetros idênticos aos utilizados nas seções anteriores.

	Sem relaxamento		**Com relaxamento**	
Problema	**Exato**	**Parc. Inexato**	**Exato**	**Parc. Inexato**
brain	299 (6319)	323 (3989)	677 (15168)	601 (6503)
colon	282 (7244)	356 (2651)	168 (5324)	216 (2072)
leukemia	348 (8724)	369 (4160)	279 (7721)	315 (3362)
lymphoma	518 (13680)	610 (8543)	293 (9184)	363 (4717)
prostate	547 (17089)	540 (6182)	601 (21251)	555 (5553)
srbct	279 (8530)	307 (3424)	154 (5641)	244 (2431)
Ball64	77 (732)	91 (236)	149 (768)	153 (322)
Logo64	68 (672)	100 (263)	151 (810)	159 (336)
Mug32	81 (1181)	143 (164)	63 (884)	128 (258)
Mug128	87 (693)	94 (278)	485 (1115)	482 (1044)
finance1000	327 (27250)	352 (6976)	315 (31569)	299 (5909)
PEMS	1186 (53520)	1210 (36729)	3444 (183716)	3264 (83784)

Tabela 5: Comparação entre os métodos com parâmetro de penalização varável.

Um comportamento que se manteve em relação aos dados da Tabela 4 é a redução do número total de iterações internas de gradientes conjugados em todos os

problemas no método parcialmente inexato. Note que isso é feito sem aumentar drasticamente o total de iterações do método ADMM. Isso é mostrado nos gráficos da Figura 9, que apresentam uma comparação entre as médias de iterações dos algoritmos testados.

Em relação ao desempenho do método sem relaxamento, considerar o parâmetro λ em sua variante adaptativa melhorou o desempenho dos métodos exato e parcialmente inexato nos problemas *Ball64*, *Logo64*, *Mug128* e *PEMS*. O aumento no total de iterações dos métodos com λ variável nos demais problemas parece prejudicial quando comparado ao desempenho considerando λ_{best}, apresentado na Tabela 4. Contudo, deve existir um pré-processamento do método para determinar o valor λ_{best} que funciona bem para cada problema, o que é algo que demanda tempo. Assim, considerar a variante adaptativa parece uma alternativa quando aproximar λ_{best} não é possível.

Considerar o parâmetro de relaxamento junto do parâmetro de penalização variável gera um comportamento mais instável no total de iterações do ADMM. Observe que com λ_{best} fixo, nota-se na Tabela 4 que adicionar o relaxamento reduziu o total de iterações em todos os casos o total de iterações. Contudo, isso não é verdade para todos os problemas na Tabela 5. Em alguns, houve a redução esperada, como no problema *colon*. Mas em outros, houve um aumento, como no problema *brain*.

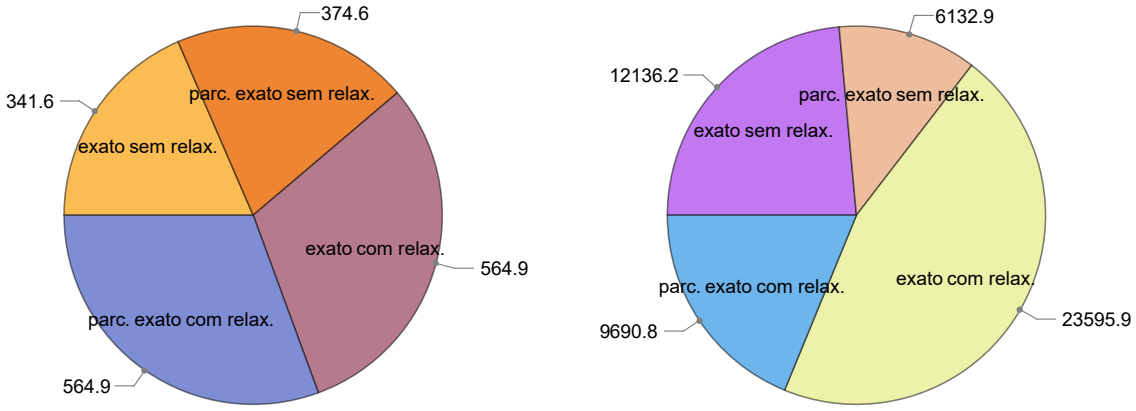

Figura 9: Médias das iterações efetuadas em cada método (esq.) e das iterações internas de gradientes conjugados (dir.) para os resultados apresentados na Tabela 5.

4 Considerações Finais

O método ADMM parcialmente inexato baseado na condição do HPE, apresentado no Algoritmo 1, teve um desempenho excelente em relação ao método ADMM em sua versão exata, quando aplicados ao problema LASSO. Para os problemas testados, a variante inexata foi capaz de manter o número total de iterações externas do método muito próximo do total de iterações do método exato, e ainda assim, manter o total de iterações internas de gradientes conjugados muito menor. Essa característica da variante inexata é vantajosa pois, como a etapa mais custosa dos métodos testados é a resolução dos sistemas lineares via gradientes conjugados, o esforço computacional exigido por ela é reduzido drasticamente.

Embora o ADMM parcialmente inexato seja dependente de um parâmetro $\sigma \in (0, 1)$ que deve ser configurado, nota-se na Seção 3 que sua variação não tem impacto significativo no total de iterações do método ADMM, mas sim no total de iterações internas de gradientes conjugados. Como indicado, valores mais próximos de 1 reduzem o esforço computacional ao determinar a solução dos sistemas, portanto, são mais vantajosos. Dessa forma, começar com valores altos para σ é recomendado.

A adição do parâmetro de relaxamento ρ_k constante em $\rho \in (1, 2)$, quando λ é mantido fixo, foi benéfica para todos os problemas abordados e para as duas variantes testadas. Em todos os casos, ocorreu a redução do total de iterações do ADMM e por consequência, a redução das iterações do método de gradientes conjugados. Ressalta-se que mesmo sem uma demonstração de convergência da versão relaxada do ADMM parcialmente inexato em [18], a adição do parâmetro de relaxamento funcionou bem, mantendo a convergência do método.

A escolha do parâmetro de penalização λ fixo influencia muito no desempenho do algoritmo e ainda não existe uma fórmula que o defina de maneira a garantir o bom funcionamento do método, nem mesmo para o caso do problema LASSO, algo que dificulta a utilização do ADMM. Contudo, a abordagem de considerar λ variável ao longo das iterações, com a atualização definida por (28), garantiu desempenho semelhante ao do método com λ fixo em λ_{best}. Logo, um bom desempenho é possível sem a necessidade de estimar λ_{best}. Porém, com parâmetro de penalização adaptativo, a adição do parâmetro

de relaxamento deve ser feita com cautela, pois a mesma pode causar uma queda no desempenho dos métodos.

Planos futuros para o projeto envolvem tratar os dois subproblemas do método ADMM de forma inexata utilizando o critério HPE e avaliar se o desempenho permanece satisfatório. Além disso, ainda está pendente a análise de convergência ao considerar a presença do parâmetro de relaxamento na versão inexata do ADMM apresentada no Algoritmo 1.

A Convexidade

Definição A.1 (Conjunto Convexo)**.** Um conjunto $\mathcal{C} \subset \mathbb{R}^n$ é dito convexo se

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

Para manter a consistência dos resultados, é conveniente adotar a convenção de que o conjunto vazio é um conjunto convexo.

Resumimos abaixo uma pequena parcela da teoria que envolve as funções convexas reais estendidas (funções que podem assumir os valores $+\infty$ e $-\infty$).

Definição A.2 (Epígrafo)**.** Sejam $\mathcal{C} \subset \mathbb{R}^n$ um conjunto convexo não vazio e $f : \mathcal{C} \to [-\infty, +\infty]$. Definimos o epígrafo de f como o conjunto

$$\text{epi}(f) := \{(x, \xi) \in \mathbb{R}^n \times \mathbb{R} \mid x \in \mathcal{C},\ \xi \in \mathbb{R},\ f(x) \leq \xi\}.$$

A seguir, apresentamos uma definição de função convexa aplicável às funções reais estendidas.

Definição A.3 (Função Convexa)**.** Sejam $\mathcal{C} \subset \mathbb{R}^n$ um conjunto não vazio e $f : \mathcal{C} \to [-\infty, +\infty]$. Dizemos que f é uma função convexa se $\text{epi}(f)$ é um conjunto convexo.

Os casos degenerados de funções convexas reais estendidas para as quais f é idêntica a $+\infty$ ou $-\infty$ em todo seu o domínio são excluídos por meio da seguinte definição.

Definição A.4 (Função Própria)**.** Sejam $\mathcal{C} \subset \mathbb{R}^n$ um conjunto não vazio e $f : \mathcal{C} \to [-\infty, +\infty]$. Dizemos que f é própria se $f(x) > -\infty$ para todo $x \in \mathcal{C}$ e ainda existe algum $x \in \mathcal{C}$ tal que $f(x) < +\infty$.

Agora, definiremos no contexto de análise convexa, o que vem a ser uma função fechada.

Definição A.5 (Função Fechada)**.** Sejam $\mathcal{C} \subset \mathbb{R}^n$ um conjunto não vazio e $f : \mathcal{C} \to [-\infty, +\infty]$. Dizemos que f é uma função fechada se $\text{epi}(f)$ é um conjunto fechado.

Os conceitos de subgradiente e subdiferencial são apresentados a seguir.

Definição A.6 (Subgradiente)**.** Seja $f : \mathbb{R}^n \to (-\infty, +\infty]$ uma função própria. Dizemos que $u \in \mathbb{R}^n$ é um subgradiente de f em x se

$$f(y) \geq f(x) + \langle u, y - x \rangle, \quad \forall\, y \in \mathbb{R}^n.$$

Definição A.7 (Subdiferencial)**.** Seja $f : \mathbb{R}^n \to (-\infty, +\infty]$ uma função própria. O subdiferencial de f em um ponto $x \in \mathrm{dom}(f)$, denotado por $\partial f(x)$, é o conjunto de todos os subgradientes de f em x, isto é,

$$\partial f(x) := \{u \in \mathbb{R}^n \mid f(y) \geq f(x) + \langle u, y - x \rangle, \quad \forall\, y \in \mathbb{R}^n\}.$$

Um outro conceito muito utilizado em algoritmos que possuem critérios inexatos é o de ε-subdiferencial.

Definição A.8 (ε-subdiferencial)**.** Seja $f : \mathbb{R}^n \to (-\infty, +\infty]$ uma função própria e $\varepsilon \geq 0$. O ε-subdiferencial de f em um ponto $x \in \mathrm{dom}(f)$ é definido pelo conjunto

$$\partial_\varepsilon f(x) := \{u \in \mathbb{R}^n \mid f(y) \geq f(x) + \langle u, y - x \rangle - \varepsilon, \quad \forall\, y \in \mathbb{R}^n\}.$$

O ε-subdiferencial pode ser entendido como uma ampliação do subdiferencial. De fato, é possível notar que $\partial f(x) \subseteq \partial_\varepsilon f(x)$ para todo $\varepsilon \geq 0$ e assim, a noção do conjunto ε-subdiferencial é útil quando não é possível calcular um subgradiente com precisão.

B Operadores Monótonos

Serão exibidos neste apêndice os resultados e definições necessários para o desenvolvimento deste trabalho. Para um tratamento completo sobre a teoria dos operadores monótonos, veja [1].

Definição B.1 (Operador)**.** Um operador $\mathcal{T} : \mathbb{R}^n \to 2^{\mathbb{R}^n}$ é um mapa que associa a cada elemento $x \in \mathbb{R}^n$, um subconjunto $\mathcal{T}(x) \subset \mathbb{R}^n$.

Denotaremos por $G_\mathcal{T}$ o gráfico do operador $\mathcal{T}$, isto é,

$$G_\mathcal{T} := \{(x, y) \in \mathbb{R}^n \times \mathbb{R}^n \mid x \in \mathbb{R}^n,\ y \in \mathcal{T}(x)\}.$$

Definição B.2 (Produto por Escalar)**.** Seja $\mathcal{T} : \mathbb{R}^n \to 2^{\mathbb{R}^n}$ um operador e $\lambda \in \mathbb{R}$ um escalar. Definimos o operador $\lambda\mathcal{T}$ por

$$\lambda\mathcal{T}(x) := \{\lambda y \mid y \in \mathcal{T}(x)\}, \quad \forall\, x \in \mathbb{R}^n.$$

Definição B.3 (Soma)**.** Sejam $\mathcal{F} : \mathbb{R}^n \to 2^{\mathbb{R}^n}$ e $\mathcal{G} : \mathbb{R}^n \to 2^{\mathbb{R}^n}$ operadores. Definimos o operador $\mathcal{F} + \mathcal{G}$ por

$$(\mathcal{F} + \mathcal{G})(x) := \{y + z \mid y \in \mathcal{F}(x),\ z \in \mathcal{G}(x)\}, \quad \forall\, x \in \mathbb{R}^n.$$

Definição B.4 (Inversão)**.** Seja $\mathcal{T} : \mathbb{R}^n \to 2^{\mathbb{R}^n}$ um operador. Definimos o operador inverso $\mathcal{T}^{-1}$ por

$$\mathcal{T}^{-1}(y) := \{x \in \mathbb{R}^n \mid y \in \mathcal{T}(x)\}, \quad \forall\, y \in \mathbb{R}^n.$$

Definição B.5 (Operador Monótono)**.** Seja $\mathcal{T} : \mathbb{R}^n \to 2^{\mathbb{R}^n}$ um operador. Dizemos que $\mathcal{T}$ é monótono se

$$\langle x' - x, y' - y\rangle \geq 0, \quad \forall\, (x, y),\ (x', y') \in G_{\mathcal{T}}.$$

Definição B.6 (Operador Monótono Maximal)**.** Seja $\mathcal{T} : \mathbb{R}^n \to 2^{\mathbb{R}^n}$ um operador monótono. Dizemos que $\mathcal{T}$ é monótono maximal se para todo operador monótono $\mathcal{T}' : \mathbb{R}^n \to 2^{\mathbb{R}^n}$ satisfazendo $\mathcal{T}(x) \subset \mathcal{T}'(x)$ para todo $x \in \mathbb{R}^n$, vale a igualdade $\mathcal{T}' = \mathcal{T}$.

Definição B.7 (Não-expansividade)**.** Um operador $\mathcal{T} : \mathbb{R}^n \to 2^{\mathbb{R}^n}$ é não-expansivo se

$$\|y' - y\| \leq \|x' - x\|, \quad \forall\, (x, y),\ (x', y') \in G_{\mathcal{T}}.$$

Definição B.8 (Não-expansividade Firme)**.** Um operador $\mathcal{T} : \mathbb{R}^n \to 2^{\mathbb{R}^n}$ é não-expansivo firme se

$$\|y' - y\|^2 \leq \langle x' - x, y' - y\rangle, \quad \forall\, (x, y),\ (x', y') \in G_{\mathcal{T}}.$$

Note que a não-expansividade ou a não-expansividade firme de um operador $\mathcal{T}$ implicam que $\mathcal{T}(x)$ é formado por no máximo um elemento. De fato, fazendo $x' = x$ nas Definições B.7 e B.8, obtemos $y' = y$.

Definição B.9 (Resolvente de $\mathcal{T}$)**.** Sejam $\mathcal{T} : \mathbb{R}^n \to 2^{\mathbb{R}^n}$ um operador e $\lambda > 0$ um escalar.

Definimos o operador resolvente de $\mathcal{T}$ por

$$\mathcal{J}_{\lambda\mathcal{T}}(x) := (\mathcal{I} + \lambda\mathcal{T})^{-1}(x).$$

Proposição B.1. *Sejam $\mathcal{T} : \mathbb{R}^n \to 2^{\mathbb{R}^n}$ um operador e $\lambda > 0$ um escalar. Então, $\mathcal{T}$ é monótono se, e somente se, o resolvente $\mathcal{J}_{\lambda\mathcal{T}}$ é um operador não-expansivo firme. Ainda, $\mathcal{T}$ é monótono maximal se, e somente se, $\mathcal{J}_{\lambda\mathcal{T}}$ é não-expansivo firme e possui domínio completo, isto é, $\mathcal{J}_{\lambda\mathcal{T}}(x)$ é possui ao menos um elemento para todo $x \in \mathbb{R}^n$.*

Demonstração. Veja [6, Teorema 2]. ■

Corolário B.1 (Lema de Representação)**.** *Sejam $\mathcal{T} : \mathbb{R}^n \to 2^{\mathbb{R}^n}$ um operador monótono e $\lambda > 0$ um escalar. Então, para cada $z \in \mathbb{R}^n$ existem no máximo um $x \in \mathbb{R}^n$ e um $y \in \mathcal{T}(x)$ tal que $z = x + \lambda y$. Se $\mathcal{T}$ é monótono maximal, então, existem únicos $x \in \mathbb{R}^n$ e $y \in \mathcal{T}(x)$ satisfazendo $z = x + \lambda y$.*

Demonstração. Veja [6, Corolário 2.3]. ■

Teorema B.1. *Sejam $\mathcal{T} : \mathbb{R}^n \to 2^{\mathbb{R}^n}$ um operador monótono maximal e $\lambda > 0$ um escalar. Então, $0 \in \mathcal{T}(x)$ se, e somente se, $x = \mathcal{J}_{\lambda\mathcal{T}}(x)$.*

Demonstração. Veja [6, Lema 2]. ■

Definimos a seguir o conceito de ε-*enlargement* de um operador

Definição B.10. Sejam $\mathcal{T} : \mathbb{R}^n \to 2^{\mathbb{R}^n}$ um operador e $\varepsilon \geq 0$. Define-se o ε-*enlargement* de $\mathcal{T}$ pelo conjunto

$$T^{\varepsilon}(x) = \{y \in \mathbb{R}^n \mid \langle x' - x, y' - y\rangle \geq -\varepsilon \quad \forall x' \in \mathbb{R}^n, y' \in \mathcal{T}(x')\}.$$

Assim como o ε-subdiferencial, o ε-*enlargement* de um operador é um conjunto que nos oferece flexibilidade ao avaliar operadores, permitindo algum grau de inexatidão, como feito em [16] e em trabalhos anteriores, de Burachik, Iusem e Svaiter, apontados pelos autores. Algumas relações relevantes entre $\mathcal{T}$ e T^{ε} são estabelecidas em [11]. São apresentadas também as relações entre $\partial_{\varepsilon} f$ e $(\partial f)^{\varepsilon}$, isto é, entre o ε-subdiferencial e o ε-*enlargement* do subdiferencial de uma função f.

Referências

[1] H. H. Bauschke e P. L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. New York: Springer, 2011.

[2] A. Beck. *First-Order Methods in Optimization*. Philadelphia: SIAM, 2017.

[3] S. Boyd, N. Parikh, E. Chu, B. Peleato e J. Eckstein. "Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers". Em: *Foundations and Trends in Machine Learning* 3.1 (2011), pp. 1–122.

[4] J. Eckstein. "Deriving solution value bounds from the ADMM". Em: *Optmization Letters* 14 (2020), pp. 1289–1303.

[5] J. Eckstein. "Splitting Methods for Monotone Operators with Applications to Parallel Optimization". Tese de dout. Massachusetts: MIT, 1989.

[6] J. Eckstein e D. P. Bertsekas. "On the Douglas—Rachford splitting method and the proximal point algorithm for maximal monotone operators". Em: *Mathematical Programming* 55.1-3 (1992), pp. 293–318.

[7] J. Eckstein e P. J. S. Silva. "A practical relative error criterion for augmented Lagrangians". Em: *Mathematical Programming* 141.1-2 (2013), pp. 319–348.

[8] C. Humes Jr. e P .J. S. Silva. "Inexact proximal point algorithms and descent methods in optimization". Em: *Optimization and Engineering* 6 (2005), pp. 257–271.

[9] G.M. Korpelevich. "The extragradient method for finding saddle points and other problems". Em: *Matecon* 12 (1976), pp. 747–756.

[10] B. Martinet. "Régularisation d'inéquations variationnelles par approximations successives. Rev. Française Informat". Em: *Recherche Opérationnelle* 4 (1970), pp. 154–158.

[11] R. D. C. Monteiro e B. F. Svaiter. "Iteration-complexity of block-decomposition algorithms and the alternating direction method of multipliers". Em: *SIAM Journal on Optimization* 23.1 (2013), pp. 475–507.

[12] J. Nocedal e S. Wright. *Numerical optimization*. 2ª ed. New York: Springer Science & Business Media, 2006.

[13] G. Passos. "Convergência do algoritmo ADMM e aplicações a problemas estruturados". Monografia disponível em `https://www.ime.unicamp.br/~mac/db/2020-1S-172351.pdf`. Jun. de 2020.

[14] R. T. Rockafellar. "Augmented Lagrangians and applications of the proximal point algorithm in convex programming". Em: *Mathematics of Operations Research* 1.2 (1976), pp. 97–116.

[15] R. T. Rockafellar. "Monotone operators and the proximal point algorithm". Em: *SIAM Journal on Control and Optimization* 14.5 (1976), pp. 877–898.

[16] S. Salzo e S. Villa. "Inexact and accelerated proximal point algorithms". Em: *Journal of Convex analysis* 19.4 (2012), pp. 1167–1192.

[17] M. V. Solodov e B. F. Svaiter. "A hybrid approximate extragradient–proximal point algorithm using the enlargement of a maximal monotone operator". Em: *Set-Valued Analysis* 7.4 (1999), pp. 323–345.

[18] B. F. Svaiter. "A partially inexact ADMM with o(1/n) asymptotic convergence rate, #(1/m) complexity, and immediate relative error tolerance". Em: *Optimization* 0.0 (2020), pp. 1–20.

[19] J. Xie. "On inexact ADMMs with relative error criteria". Em: *Computational Optimization and Applications* 71.3 (2018), pp. 743–765.

[20] J. Xie, A. Liao e X. Yang. "An inexact alternating direction method of multipliers with relative error criteria". Em: *Optimization Letters* 11.3 (2017), pp. 583–596.