UNIVERSIDADE FEDERAL DO PARANÁ

TATIANE CAZARIN DA SILVA

ALGORITMOS PRIMAIS-DUAIS DE PONTO FIXO APLICADOS AO PROBLEMA *RIDGE REGRESSION*

CURITIBA
2016

TATIANE CAZARIN DA SILVA

ALGORITMOS PRIMAIS-DUAIS DE PONTO FIXO APLICADOS AO PROBLEMA *RIDGE REGRESSION*

Tese apresentada ao Programa de Pós-Graduação em Métodos Numéricos em Engenharia, Área de Concentração em Programação Matemática, dos Setores de Tecnologia e de Ciências Exatas da Universidade Federal do Paraná, como requisito parcial à obtenção do título de Doutor em Métodos Numéricos em Engenharia.

Orientador:

Prof. Dr. Ademir Alves Ribeiro

Coorientadora:

Profa. Dra. Gislaine Aparecida Periçaro

CURITIBA

2016

TERMO DE APROVAÇÃO

Os membros da Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação em MÉTODOS NUMÉRICOS EM ENGENHARIA da Universidade Federal do Paraná foram convocados para realizar a arguição da Tese de Doutorado de TATIANE CAZARIN DA SILVA, intitulada: *"Algoritmos Primais-Duais de Ponto Fixo Aplicados ao Problema Ridge Regression*", após terem inquirido a aluna e realizado a avaliação do trabalho, são de parecer pela sua APROVAÇÃO.

Curitiba, 08 de Julho de 2016.

Prof ADEMIR ALVES RIBEIRO
Presidente da Banca Examinadora (UFPR)

Prof MAEL SACHINE
Avaliador Externo (UFPR)

Prof PAULO DOMINGOS CONEJO
Avaliador Externo (UNIOESTE)

Prof ROBERTO ANDREANI
Avaliador Externo (UNICAMP)

Prof RODRIGO GARCIA EUSTAQUIO
Avaliador Externo (UTFPR)

Prof DIANE RIZZOTTO ROSSETTO
Avaliador Externo (UTFPR)

Dedico ao meu pai Luiz.

AGRADECIMENTOS

Começo dizendo que nossa realização sempre depende da ajuda de muitas pessoas. Minha gratidão.

A Deus, acima de tudo, pelo dom da vida e por estar sempre comigo, conduzindo a oportunidades, momentos, caminhos e planos de vida.

Ao meu pai, Luiz, pelo imenso amor, carinho, apoio, estímulo e orações. Com certeza, com seu exemplo eu aprendi os melhores valores da vida. À toda minha família, base da vida, pela força e apoio de onde quer que estejam.

Ao meu orientador, Ademir, pelos ensinamentos, aprendizado, acompanhamento, dedicação, paciência, cobrança e exemplo de profissionalismo. À minha coorientadora, Gislaine, pelo estímulo, disponibilidade, dedicação, ensinamentos e acompanhamento durante todo o doutorado. Com certeza esse foi um trabalho em equipe, da qual me sinto muito honrada em ter feito parte e que me fez gostar de aprender a cada dia mais. Obrigada pelas discussões, dúvidas e questionamentos, pessoalmente ou virtualmente, que com toda certeza me fizeram crescer e amadurecer como profissional e pessoa.

Ao Leonardo Moreto Elias pela disponibilidade, supervisão e orientação em um dos assuntos da tese.

À Gislaine, minha amiga-irmã, pelo incentivo, acima de tudo, em cada um dos vários momentos difíceis em que me escutava, ajudava ou sentava comigo para estudar, sempre dizendo que um dia tudo daria certo.

Aos meus amigos, Adriano, Juliano, Lilian, Gislaine, Solange e Vanessa pela amizade, conselhos, viagens, enfim, simplesmente por todos os momentos. E, também, a todos os amigos que não faziam parte desta rotina.

À Dona Aparecida e a Solange, pela recepção, acolhida, convívio e amizade, especialmente neste último ano.

Aos professores do PPGMNE pelos ensinamentos e experiência transmitidos.

À todos os colegas e amigos que fiz por meio do PPGMNE, que proporcionaram a divisão de muitos momentos de angústia, alegria e amizade, especialmente, aqueles com quem pude ter o prazer de conviver mais.

À UTFPR, campus Campo Mourão, e ao DAMAT que proporcionaram o afastamento durante uma fase do doutorado.

Aos membros da banca, professores Diane, Mael, Paulo, Roberto e Rodrigo, pelas sugestões e contribuições no trabalho.

À todos que de alguma forma contribuíram durante este período.

Busque nos seus objetivos, seus sonhos e suas realizações e por mais que seja difícil não desista jamais.

RESUMO

Neste trabalho propomos algoritmos para resolver uma formulação primal-dual geral de ponto fixo aplicada ao problema de *Ridge Regression*. Estudamos a formulação primal para problemas de quadrados mínimos regularizado, em especial na norma L_2, nomeados *Ridge Regression* e descrevemos a dualidade convexa para essa classe de problemas. Nossa estratégia foi considerar as formulações primal e dual conjuntamente, e minimizar o *gap* de dualidade entre elas. Estabelecemos o algoritmo de ponto fixo primal-dual, nomeado SRP e uma reformulação para esse método, contribuição principal da tese, a qual mostrou-se mais eficaz e robusta, designada por método acc-SRP, ou versão acelerada do método SRP. O estudo teórico dos algoritmos foi feito por meio da análise de propriedades espectrais das matrizes de iteração associadas. Provamos a convergência linear dos algoritmos e apresentamos alguns exemplos numéricos comparando duas variantes para cada algoritmo proposto. Mostramos também que o nosso melhor método, acc-SRP, possui excelente desempenho numérico na resolução de problemas muito mal-condicionados quando comparado ao Método de Gradientes Conjugados, o que o torna computacionalmente mais atraente.

Palavras-chave: Métodos primais-duais, *Ridge Regression*, ponto fixo, dualidade, métodos acelerados.

ABSTRACT

In this work we propose algorithms for solving a fixed-point general primal-dual formulation applied to the Ridge Regression problem. We study the primal formulation for regularized least squares problems, especially L_2-norm, named Ridge Regression and then describe convex duality for that class of problems. Our strategy was to consider together primal and dual formulations and minimize the duality gap between them. We established the primal-dual fixed point algorithm, named SRP and a reformulation for this method, the main contribution of the thesis, which was more efficient and robust, called acc-SRP method or accelerated version of the SRP method. The theoretical study of the algorithms was done through the analysis of the spectral properties of the associated iteration matrices. We proved the linear convergence of algorithms and some numerical examples comparing two variants for each algorithm proposed were presented. We also showed that our best method, acc-SRP, has excellent numerical performance for solving very ill-conditioned problems, when compared to the conjugate gradient method, which makes it computationally more attractive.

Key-words: Primal-dual methods, ridge regression, fixed point, duality, accelerated methods.

Lista de Figuras

Lista de Tabelas

Sumário

Introdução

Ridge Regression é um método popular de regularização que foi introduzido por Hoerl e Kennard em 1970 [35] e pode ser visto como uma aplicação da regularização Tikhonov [3, 4, 34, 57]. Estatisticamente, tem como finalidade a obtenção de melhores resultados para a análise de regressão múltipla quando comparado à regressão de quadrados mínimos usual, nos casos em que há presença de multicolinearidade nas variáveis explicativas. Isso ocorre devido ao fato de que o método fornece estimativas para o vetor de parâmetros com um menor comprimento que aquelas obtidas pelo método de quadrados mínimos. Neste sentido, o objetivo da *Ridge Regression* é reduzir o erro padrão dos coeficientes de regressão por meio da imposição de uma penalidade, na norma L_2, sob os coeficientes [14, 31, 56, 58].

Vários trabalhos têm sido destinados a resolver problemas de *Ridge Regression* ou mesmo problemas com formulação geral, para os quais podem ser vistos como casos particulares. Alguns destes trabalhos consideram a formulação dual do problema, propondo algoritmos estocásticos ou determinísticos. Por exemplo, problemas de predição, tais como regressão linear e classificação, foram introduzidos em [62] para um formato geral para modelos lineares de predição regularizada, sob o qual foi derivada uma representação dual. Uma outra versão dual para o problema *Ridge Regression* foi proposta em [51], permitindo realizar a regressão não linear por meio da construção de uma função de regressão linear em dimensões maiores. Outras abordagens, ainda, podem ser encontradas em [32, 43, 50, 53, 61].

Nesta conjuntura de dualidade, existem algumas vantagens em trabalhar não apenas com a versão dual mas com o par de problemas primal-dual, inicialmente proposto por Dantzig, Ford e Fulkerson [12] para resolução de problemas lineares. O método primal-dual é uma ferramenta padrão no projeto de algoritmos para problemas de otimização combinatória, os quais podem ser modificados a fim de proporcionar bons algoritmos de aproximação para uma grande variedade de problemas de complexidade não polinomial (*NP-hard*) [19].

No contexto de programação linear, Zhu [63] apresenta os resultados de uma pesquisa sobre vários algoritmos com estrutura unificada primal-dual, na qual essa versão simultânea é empregada para melhorar os resultados, quando comparados aos problemas primal e dual, separadamente. Zhu destaca que tais algoritmos alcançam uma aceleração significativa em problemas de grande escala, atuando assim como uma

metodologia computacional promissora para muitos outros problemas. É realizada uma unificação entre as abordagens primal-dual, disponíveis na literatura, em um algoritmo computacional comum, para o qual a análise de complexidade foi feita com base em uma função exponencial.

Agora, envolvendo também conceitos de dualidade convexa mais gerais, a metodologia primal-dual ganha grande destaque em programação não-linear, contexto esse em que se situa o problema *Ridge Regression*. As vantagens em se trabalhar com o par de problemas primal-dual são apontadas por Komodakis e Pesquet [40], embasados em recentes avanços em análise convexa, otimização discreta, processamento paralelo e otimização não suave com ênfase nas questões de esparsidade. Além disso, Komodakis e Pesquet propõem os princípios dessa abordagem a fim de mostrar os benefícios de algoritmos primais-duais, tanto para resolver problemas de otimização convexa em grande escala como os discretos, apresentando uma série de métodos de otimização primal-dual utilizados para a resolução de problemas de sinal e processamento de imagem.

O problema de minimizar a média de um grande número de funções convexas suaves penalizada com um regularizador fortemente convexo, o que também é um problema do tipo *Ridge Regression*, é verificado em [50]. Os autores propõem um algoritmo, nomeado Quartz, que resolve simultaneamente os problemas primal e dual. O método consiste em, a cada iteração, selecionar e atualizar um subconjunto aleatório das variáveis duais, sem qualquer suposição sobre a distribuição de probabilidade para tais subconjuntos. Dessa forma, o diferencial do método consiste em ser o primeiro a fazer uma análise sob uma amostragem arbitrária. As atualizações duais são usadas para a atualização da variável primal, e o processo é repetido. Os experimentos numéricos foram realizados para o problema linear de *Support Vector Machine*, com norma L_2 regularizada.

Chambolle e Pock [7] apresentam o estudo de um algoritmo primal-dual de primeira ordem aplicado em problemas de otimização convexa onde o objetivo é calcular o movimento aparente em sequências de imagens. Os autores mostram que o algoritmo converge, com taxa ótima de $O\left(\frac{1}{N}\right)$ sob o *gap* de dualidade e, particularmente, que pode ser modificado, com uma nova taxa de convergência de $O\left(\frac{1}{N^2}\right)$ para quando os problemas primal e dual são uniformemente convexos. Para essa modificação acelerada, em problemas suaves, garantem que a convergência é linear, da ordem de $O\left(\frac{1}{e^N}\right)$. Outras aplicações da metodologia primal-dual podem ser vistas em [2, 19, 23].

Com base neste contexto de dualidade e regressão, neste trabalho, propomos algoritmos primais-duais de ponto fixo aplicados a uma versão de problemas *Ridge Regression*, que também são abordados, por exemplo em [50]. De acordo com [40], os métodos primais-duais têm sido empregados principalmente em problemas de otimização convexa com dualidade forte, obtendo sucesso quando aplicado a vários tipos de funções não-lineares, que surgem em diversas áreas de aplicação, tais como processamento de imagem, aprendizado de máquina, problemas inversos, entre outros.

Assim, em relação ao comentário anterior, somos encorajados a acreditar que esta é uma estratégia interessante para ser aplicada no problema *Ridge Regression*, visto que o problema primal-dual estabelecido é quadrático e fortemente convexo. Definimos a versão do problema primal, *Ridge Regression*, a ser tratado no trabalho e sob ele estabelecemos o problema dual, utilizando conceitos de dualidade convexa, em particular dualidade de Fenchel. Mostramos que o par de problemas satisfaz as condições de dualidade e, definida a condição de otimalidade, estabelecemos o algoritmo SRP assim como a sua prova de convergência linear, por meio de uma modificação que recai no formato ponto fixo. Baseados nas características de possível melhoria do algoritmo, quando utilizado o par primal-dual, propomos também uma versão acelerada para o método, nomeada acc-SRP, sendo este o principal resultado desta tese.

Experimentos numéricos foram executados em Matlab, a fim de verificar a eficiência e robustez dos algoritmos propostos. Os resultados indicam que a metodologia proposta é competitiva com os métodos de regularização clássicos existentes pois, teoricamente, possui garantias de convergência estabelecidas e, numericamente, apresenta excelente desempenho quando comparado à uma metodologia clássica na resolução de problemas muito mal-condicionados.

Este trabalho encontra-se estruturado conforme segue. No Capítulo 1 apresentamos uma discussão sobre os métodos de regularização existentes e sua aplicação em uma formulação do problema *Ridge Regression* no contexto estatístico. No Capítulo 2 revisamos alguns conceitos básicos de dualidade convexa e, em particular, construímos o par de problemas primal-dual, utilizado na pesquisa. Os algoritmos propostos para resolver o problema, assim como a análise de convergência teórica, estão apresentados no Capítulo 3. No Capítulo 4 descrevemos alguns detalhes referentes à implementação dos algoritmos SRP e acc-SRP, bem como os resultados numéricos obtidos a partir dessa implementação. Por fim, apresentamos a conclusão do trabalho.

Capítulo 1

Problemas mal-postos e regularização

Neste capítulo discutimos alguns conceitos fundamentais no campo da resolução de problemas mal-postos, a fim de justificar a teoria de regularização. Neste sentido, a idéia é destacar alguns dos principais métodos de regularização existentes, disponíveis na literatura, assim como suas vantagens e desvantagens. Tal discussão é estruturada a seguir com base em [3, 26, 27, 28, 44].

1.1 Problemas mal-postos

O conceito de problemas mal-postos surgiu em 1923 devido a Hadamard [24], que define problema mal-posto como aquele que não admite solução única ou quando pequenas perturbações arbitrárias nos dados afetam consideravelmente a solução. Dessa forma, um problema mal-posto é aquele que não satisfaz pelo menos uma dessas características, que dizem respeito à existência, unicidade e estabilidade da solução.

Neste sentido, considere o seguinte problema

$$\min_{x\in\mathbb{R}^n} \|Ax-b\|, \quad A\in\mathbb{R}^{m\times n} \quad \text{e} \quad b\in\mathbb{R}^m. \tag{1.1}$$

Dizemos que (1.1) é um problema mal-posto se (a) os valores singulares de A decrescem gradualmente até zero e (b) a razão entre o maior e o menor valores singulares não nulos é grande. Podemos perceber tais características, também, em sistemas de equações lineares e problemas lineares de quadrados mínimos resultantes da discretização de sistemas mal-postos, por exemplo as equações integrais de Fredholm, tendo aplicações nas mais diversas áreas [3, 8, 10, 11, 22, 27, 46].

A principal dificuldade em se trabalhar com os problemas mal-postos, por exemplo (1.1), ocorre quando a matriz A possui um conjunto de valores singulares muito próximos a zero. Por isso, é necessário incorporar mais informações sobre a solução desejada, a fim de estabilizar o problema e destacar uma solução útil e estável. Este é o propósito dos métodos de regularização, discutidos a seguir.

1.2 Métodos de regularização

A obtenção de soluções exatas para problemas mal-postos tem sido de grande interesse em diversas áreas do conhecimento, já que a solução proposta por quadrados mínimos pode ser contaminada por ruídos. Neste sentido, a teoria de regularização consiste na determinação de uma solução para problemas mal-postos que se aproxime da solução exata e não seja afetada por ruídos. Para isso, intuitivamente, a análise e solução de um problema mal-posto é feita via solução de um problema associado que é bem-posto [3, 26, 27, 28, 42].

Os métodos de regularização são classificados em métodos de projeção, métodos de penalidade e métodos híbridos [3, 26, 42, 55]. Dentre os métodos de projeção podemos citar a Decomposição em Valores Singulares Truncada (TSVD) [28], a Decomposição em Valores Singulares Truncada Generalizada (TGSVD) [26], o Método de Mínimos Resíduos Generalizado (GMRES) [6] e o Método de Regressão por Quadrados Mínimos (LSQR) [47, 48], matematicamente idêntico ao Método dos Gradientes Conjugados [3]. Dentre os métodos de penalidade podemos destacar a regularização Tikhonov, o método da variação total e o método L1 [55, 57, 59]. Os métodos híbridos consistem na combinação entre um método de penalidade e um método de projeção, a fim de contornar a dificuldade de escolha do critério de parada, que surge quando empregados separadamente. Entre os métodos híbridos, destacamos a bidiagonalização de Golub-Kahan, também nomeado algoritmo Lanczos, o algoritmo Arnoldi, entre outros [18, 41].

Os métodos de projeção caracterizam-se por considerar, logo nas primeiras iterações, informações relevantes para o problema. No entanto, se as iterações persistirem, as novas componentes passam a ser contaminadas por erros nos dados, possibilitando a desestabilização da solução, já que a solução do problema é a mesma definida pelo método dos quadrados mínimos usual [3]. Dessa forma, a dificuldade de aplicação dos métodos de projeção é a determinação da iteração de parada.

Dentre os métodos de projeção, um que tem atraído grande atenção é o método LSQR, por conter características de regularização e propriedade de semi-convergência [25, 28, 30, 36, 38, 54]. A semi-convergência garante que à medida que as iterações evoluem as soluções iteradas se aproximam da solução exata, porém a partir de dado momento as mesmas passam a se distanciar e a estabilidade do algoritmo não é alcançada [3]. Isso ocorre pelo fato de que o ruído afeta progressivamente o subespaço solução do problema após uma quantidade ótima de iterações, consequentemente, deteriorando a solução. Uma representação para tal comportamento pode ser visualizada na Figura 1.1.

Neste contexto de regularização, um dos métodos empregados com maior frequência na resolução numérica de problemas discretos mal-postos é a regularização Tikhonov, a qual descrevemos brevemente a seguir, baseados em [17, 34, 57].

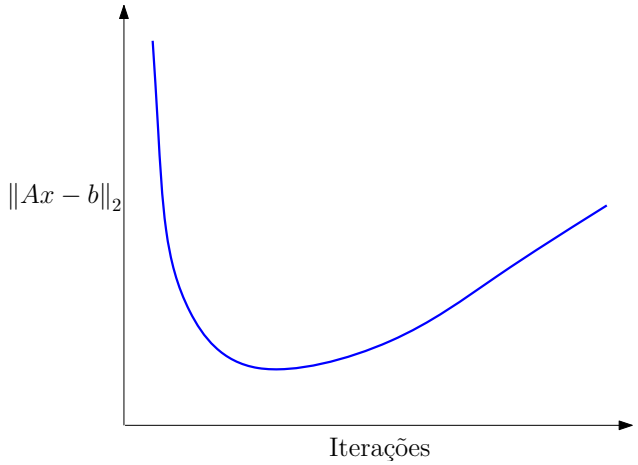


Figura 1.1: Norma residual para um crescente número de iterações

1.2.1 Regularização Tikhonov

A regularização Tikhonov substitui o problema de quadrados mínimos

$$x = \arg\min_{x \in \mathbb{R}^n} \|b - Ax\|_2^2 \tag{1.2}$$

onde $A \in \mathbb{R}^{m \times n}$ e $b \in \mathbb{R}^m$, por

$$x_\lambda = \arg\min_{x \in \mathbb{R}^n} \left\{ \|b - Ax\|_2^2 + \lambda^2 \|L(x - x_0)\|_2^2 \right\} \tag{1.3}$$

em que $\lambda > 0$ é o parâmetro de regularização, o vetor $x_0 \in \mathbb{R}^n$ é uma aproximação inicial para a solução e $L \in \mathbb{R}^{q \times n}$ uma matriz dada. Quando $L = I$ temos a formulação de Tikhonov padrão, caso contrário, Tikhonov no formato geral. A primeira proposta de resolução para o problema (1.3) é devida a Golub [20]. A ideia consiste em tratar (1.3) como um problema de quadrados mínimos da forma

$$x_\lambda = \arg\min_{x \in \mathbb{R}^n} \left\| \begin{pmatrix} A \\ \lambda L \end{pmatrix} x - \begin{pmatrix} b \\ \lambda L x_0 \end{pmatrix} \right\|_2^2. \tag{1.4}$$

No entanto, a formulação padrão da regularização Tikhonov substitui o problema usual de quadrados mínimos por

$$x_\lambda = \arg\min_{x \in \mathbb{R}^n} \left\{ \|b - Ax\|_2^2 + \lambda^2 \|x\|_2^2 \right\}, \tag{1.5}$$

ou seja, considerando $L = I$ e $x_0 = 0$.

A aproximação da solução x_λ em relação à solução exata do problema depende da escolha do parâmetro λ. Destacam-se na literatura diversos métodos para escolha de tal parâmetro, entre os quais podemos citar a Curva-L [28, 29], a Validação Cruzada Generalizada (GCV) [21], o Princípio da Discrepância [45], o Algoritmo de Ponto-Fixo de Bazán [4], entre outros. O escopo desta tese não diz respeito à escolha do parâmetro ótimo de regularização, mas à resolução do problema para uma variedade de parâmetros

definidos aleatoriamente. Sendo assim, maiores detalhes sobre os métodos de escolha do parâmetro de penalidade podem ser encontrados nos trabalhos anteriormente citados.

A formulação de Tikhonov, no formato padrão, pode ser observada como um problema na Estatística, especificamente, em regressão múltipla. *Ridge Regression* é considerada uma aplicação de Regularização Tikhonov no contexto estatístico, embora tenham sido desenvolvidas independentemente [1]. Sendo assim, destinamos a seção a seguir para a formulação do problema *Ridge Regression*, onde, por conveniência, optamos por empregar a notação estatística padrão.

1.3 Definição do problema *Ridge Regression*

A análise de regressão é uma técnica estatística que visa investigar e modelar a relação entre variáveis, sendo uma das ferramentas mais empregadas na análise de dados. Um dos objetivos desta técnica é a estimação de parâmetros desconhecidos do modelo. Geralmente, o interesse é avaliar a relação de uma variável de interesse Y em relação a p variáveis independentes $X_{ij}, i = 1, \ldots, n\ , j = 1, \ldots, p$.

Sendo assim, um possível modelo de regressão múltipla, no formato matricial, pode ser expresso como

$$Y = X\beta + \epsilon, \tag{1.6}$$

onde $Y \in \mathbb{R}^n$ é a variável dependente ou resposta, $X \in \mathbb{R}^{n\times p}$, assumindo posto completo, representa as variáveis independentes ou regressoras, $\beta \in \mathbb{R}^p \backslash \{0\}$ os coeficientes de regressão estimados e ϵ o erro residual.

A estimativa para os coeficientes β pode ser feita, por exemplo, pelo Método dos Quadrados Mínimos (M.Q.M.) que consiste em minimizar a soma dos quadrados dos resíduos. Definindo $f(\beta) = \|Y - X\beta\|_2^2$, temos então que o problema de minimizar o quadrado da norma residual retorna como minimizador

$$\widehat{\beta} = \arg\min f(\beta). \tag{1.7}$$

Observe que f é uma função convexa, uma vez que $\nabla^2 f(\beta) = 2X^T X$ é semi-definida positiva. Como um resultado clássico em otimização, segue que qualquer minimizador local de f é global. Assim, pela condição necessária de otimalidade de primeira ordem, se $\widehat{\beta}$ for um minimizador local de f, então $\widehat{\beta}$ é solução do sistema $\nabla f(\beta) = 0$. Portanto, se $X^T X$ for não singular, $\widehat{\beta}$ será dado por

$$\widehat{\beta} = (X^T X)^{-1} X^T Y, \tag{1.8}$$

que representa o estimador de quadrados mínimos usual. Pode-se provar que $\widehat{\beta}$, dado em (1.8), é um estimador não-tendencioso, ou seja, $\mathbb{E}\left(\widehat{\beta}\right) = \beta$. No entanto, note

que se X^TX é singular, não existem estimadores únicos para o problema. Além disso, mal-condicionamento, número de predições excedendo o número de observações, posto incompleto e multicolinearidade afetam diretamente essa solução.

Para contornar esse problema são aplicados métodos de regularização, dentre os quais o mais conhecido é o Método de *Ridge Regression*. Trata-se de uma metodologia estatística aplicada na análise de dados de regressão múltipla que sofrem de multicolinearidade ou não ortogonalidade. Neste caso, as estimativas de parâmetros com base no método dos quadrados mínimos usual têm uma probabilidade alta de ser insatisfatória, pois a variabilidade nos coeficientes é alta. Isso ocorre pelo fato de algumas variáveis explicativas serem combinações lineares de outras e não há estimadores de quadrados mínimos únicos para os parâmetros, pois a matriz X^TX é singular.

Neste sentido, *Ridge Regression* é um método de estimativa com base na adição de pequenas quantidades positivas para a diagonal da matriz X^TX, o que leva à obtenção de estimativas tendenciosas, porém com menor erro quadrático médio.

Considerando o problema de quadrados mínimos usual, dado em (1.7), temos que o problema *Ridge Regression* é definido como

$$\min \quad f(\beta) + \lambda \|\beta\|_2^2 . \tag{1.9}$$

A existência de uma solução para (1.9) deve-se ao mesmo argumento apresentado para o problema (1.7). Para simplificação, denotaremos a função objetivo por $g(\beta) = f(\beta) + \lambda \|\beta\|_2^2$. Assim, temos que $\nabla^2 g(\beta) = 2X^TX + 2\lambda I$ é definida positiva, o que implica na existência de um único minimizador para g. Então, o estimador de *Ridge Regression*, solução do sistema $\nabla g(\beta) = 0$, é dado por

$$\widehat{\beta}_{Ridge} = (X^TX + \lambda I)^{-1} X^TY. \tag{1.10}$$

Note que, se $\alpha_1, \ldots, \alpha_p$ com $\alpha_1 \geq \ldots \geq \alpha_p$, representam os autovalores de X^TX e $v_1, \ldots, v_p$ os autovetores associados, respectivamente, os autovalores de $(X^TX + \lambda I)^{-1}$ serão, exatamente, $(\alpha_i + \lambda)^{-1}, i = 1, \ldots, p$. Se X^TX for singular ou quase singular, com autovalor mínimo α_p, então o menor autovalor de $(X^TX + \lambda I)$ será $(\alpha_p + \lambda)$ e, esta matriz não estará tão próxima da singularidade.

O estimador de *Ridge Regression* pode ser obtido por meio da resolução do problema de otimização irrestrito, dado em (1.9). No entanto, esse problema de minimização pode, também, ser reformulado através do seguinte problema de otimização restrito

$$\begin{aligned} \min \quad & \|Y - X\beta\|_2^2 \\ s.a \quad & \|\beta\|_2^2 \leq c \end{aligned} \tag{1.11}$$

para algum $c > 0$. A existência de uma solução para o problema (1.11) é garantida

pelo fato de se tratar de uma função contínua e um conjunto viável compacto. Uma representação para a solução deste problema, para o caso em que $p = 2$ é ilustrada na Figura 1.2.

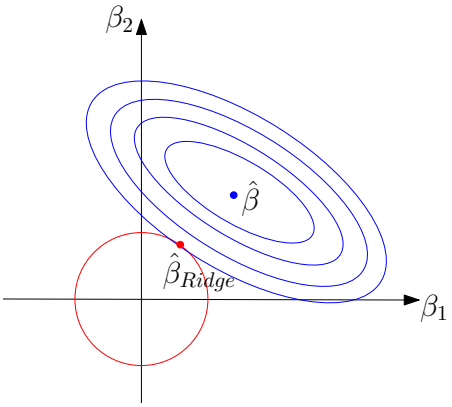


Figura 1.2: Solução do problema *Ridge Regression* para um exemplo com $p = 2$

Mostraremos que os problemas (1.9) e (1.11) são equivalentes. Para isso, considere o estimador de *Ridge Regression* para o problema (1.9), definido em (1.10). Pela condição necessária de otimalidade de primeira ordem, temos que $\nabla g(\widehat{\beta}_{Ridge}) = 0$, ou seja,

$$\nabla f(\widehat{\beta}_{Ridge}) + 2\lambda\widehat{\beta}_{Ridge} = 0. \tag{1.12}$$

Por outro lado, o problema de otimização restrito (1.11) pode ser resolvido pelas condições de Karush-Kuhn-Tucker (KKT). Mostraremos que, para uma adequada escolha de c, o estimador $\widehat{\beta}_{Ridge}$, dado em (1.10), também é solução de (1.11). De fato, como $\widehat{\beta}_{Ridge} \neq 0$, a condição de qualificação de independência linear (LICQ) é satisfeita em $\widehat{\beta}_{Ridge}$. Além disso, usando (1.12), temos que existe $\nu^* \geq 0$ tal que as condições de KKT são satisfeitas:

$$\begin{cases} -\nabla f(\widehat{\beta}_{Ridge}) = 2\nu^*\widehat{\beta}_{Ridge}, \\ \nu^*\left(\left\|\widehat{\beta}_{Ridge}\right\|_2^2 - c\right) = 0, \\ \left\|\widehat{\beta}_{Ridge}\right\|_2^2 \leq c. \end{cases} \tag{1.13}$$

De fato, tome $\nu^* = \lambda$ e $c = \left\|\widehat{\beta}_{Ridge}\right\|_2^2$. Assim, os problemas (1.9) e (1.11) têm a mesma solução, sendo esta destacada em (1.10).

Contornada a singularidade envolta ao problema e verificada a equivalência entre os problemas (1.9) e (1.11), estabeleceremos agora um comparativo entre o estimador de quadrados mínimos usual e o estimador de *Ridge Regression*, de acordo com [9, 35, 60]. Para isso, consideramos inicialmente informações referentes ao valor esperado para os coeficientes. Essa relação pode ser observada realizando a decomposição espectral da matriz X, como $X = UDV^T$, em que $U \in \mathbb{R}^{n\times n}$ e $V \in \mathbb{R}^{p\times p}$ são ortogonais, $D \in \mathbb{R}^{n\times p}$ e, usando (1.8) e (1.10), podemos escrever que o estimador de quadrados mínimos usual é dado por

$$\widehat{\beta} = VD^{-2}D^TU^TY, \tag{1.14}$$

e o estimador de *Ridge Regression*

$$\widehat{\beta}_{Ridge} = V(D^2 + \lambda I)^{-1}D^TU^TY, \tag{1.15}$$

para o qual há um encurtamento tanto dos estimadores como dos valores singulares. Além disso, note que, quando $\lambda = 0$, o estimador $\widehat{\beta}_{Ridge}$ coincide com $\widehat{\beta}$.

Com base no valor esperado para os estimadores, o método dos quadrados mínimos retorna a melhor solução não-tendenciosa para o modelo, ou seja, $\mathbb{E}\left(\hat{\beta}\right) = \beta$. Por outro lado, $\hat{\beta}_{Ridge}$ é um estimador tendencioso para β quando $\lambda \neq 0$, uma vez que

$$\mathbb{E}\left(\hat{\beta}_{Ridge}\right) = X^TX\left(\lambda I + X^TX\right)^{-1}\beta.$$

Observando o valor esperado para os estimadores de *Ridge Regression*, justificamos o fato do método encurtar os coeficientes obtidos pelo método de quadrados mínimos, aproximando-os de zero. De fato, de acordo com [60] temos que

$$\lim_{\lambda\to\infty} \mathbb{E}\left(\hat{\beta}_{Ridge}\right) = \lim_{\lambda\to\infty} X^TX\left(\lambda I + X^TX\right)^{-1}\beta = 0.$$

Analisando, agora, a variância do estimador de quadrados mínimos temos que

$$\mathbb{V}\left(\hat{\beta}\right) = \sigma^2(X^TX)^{-1}, \tag{1.16}$$

enquanto que a variância do estimador de *Ridge Regression* é dada por

$$\mathbb{V}\left(\hat{\beta}_{Ridge}\right) = \sigma^2\left(X^TX + \lambda I\right)^{-1}X^TX\left(\left(X^TX + \lambda I\right)^{-1}\right)^T. \tag{1.17}$$

Além disso, fazendo

$$\lim_{\lambda\to\infty} \mathbb{V}\left(\hat{\beta}_{Ridge}\right) = \lim_{\lambda\to\infty} \sigma^2\left(X^TX + \lambda I\right)^{-1}X^TX\left(\left(X^TX + \lambda I\right)^{-1}\right)^T = 0,$$

podemos concluir que a variância de $\widehat{\beta}_{Ridge}$ é uma função decrescente de λ, o que justifica a utilização do Método de *Ridge Regression* já que apresenta uma variância inferior à variância do Método dos Quadrados Mínimos usual para um certo valor de λ. Além disso, $\widehat{\beta}^T_{Ridge}\widehat{\beta}_{Ridge} < \widehat{\beta}^T\widehat{\beta}$ para todo λ positivo e $\widehat{\beta}^T_{Ridge}\widehat{\beta}_{Ridge}$ tende para zero conforme λ cresce.

Agora, após apresentarmos o problema *Ridge Regression* no formato mais simples, é importante destacar que sua característica principal de formulação se trata da minimização de uma função convexa suave sujeita a uma penalidade sob uma função fortemente convexa. Essa característica é o que possibilita a generalização do problema

a outras variações do tipo *Ridge Regression*. Por sua vez, é essa extensão que permite classificar o problema utilizado nesta pesquisa como um problema do tipo *Ridge Regression*. Para definir o problema proposto para estudo será necessário, também, a construção do par de problemas primal-dual, no contexto de otimização convexa. Devido a isso, o próximo capítulo destina-se a elencar alguns conceitos fundamentais no campo de dualidade convexa, a fim de estabelecer a relação entre os problemas considerados.

Capítulo 2

Um caso de Dualidade

Neste capítulo, nosso objetivo é apresentar um esquema de dualidade que será ferramenta para a construção do problema abordado nesta tese. Para o estudo dessa ideia, faremos uso de dois conceitos da literatura: conjugada de Fenchel e dualidade Lagrangiana. Iniciaremos com uma breve revisão destes tópicos.

2.1 A conjugada de Fenchel

A função conjugada foi introduzida por Fenchel em [16] e possui aplicações nas mais diversas áreas da Ciência, tais como Otimização, Equações Diferenciais e Economia. Apresentamos a seguir uma breve revisão sobre algumas propriedades que são de interesse no desenvolvimento desta tese e podem ser encontradas em [5, 15, 39, 49].

Vamos inicialmente dar uma motivação geométrica para a noção de conjugação. Dados $u \in \mathbb{R}^n$ e $b \in \mathbb{R}$, definimos o hiperplano em $\mathbb{R}^{n+1}$ por

$$H = \left\{(x, \lambda) \in \mathbb{R}^{n+1} \mid u^T x - \lambda = b\right\}. \tag{2.1}$$

Dada uma função $f : \mathbb{R}^n \to \mathbb{R}$ convexa, consideramos o problema de determinar $b \in \mathbb{R}$ tal que o hiperplano H, que pode ser descrito por $\lambda = u^T x - b$, seja tangente ao gráfico de f, ou seja, o hiperplano H seja suporte do epigrafo[1] de f. Isso é o que podemos observar na Figura 2.1, em que para efeitos de ilustração o eixo horizontal corresponde ao $\mathbb{R}^n$ e o eixo vertical corresponde a $\mathbb{R}$.

Nesse sentido, note que se considerarmos $-b$ como a distância vertical entre o gráfico da função f e o hiperplano dado por $\lambda = u^T x$, ou seja, se tomarmos

$$-b = \inf_{x \in \mathbb{R}^n} \left\{f(x) - u^T x\right\},$$

[1] Seja $f : \mathbb{R}^n \to \mathbb{R}$, definimos o epigrafo de f como $epi(f) = \{(x, \mu) \mid x \in \mathbb{R}^n, \mu \in \mathbb{R}, f(x) \leq \mu\} \subseteq \mathbb{R}^{n+1}$.

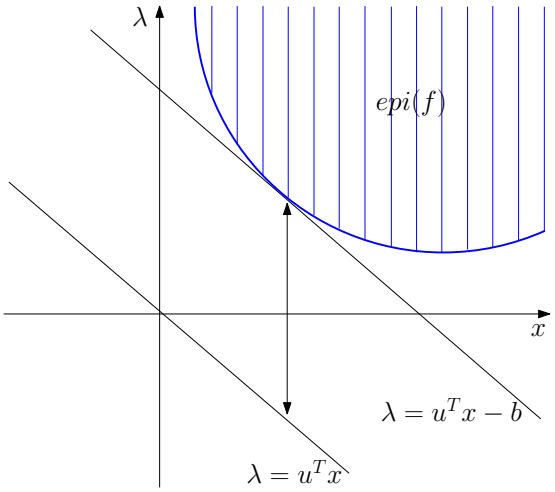


Figura 2.1: Hiperplano H tangente ao gráfico de f

ou equivalentemente

$$b = \sup_{x \in \mathbb{R}^n} \left\{ u^T x - f(x) \right\},$$

o hiperplano H será suporte para $epi(f)$ no ponto em que tangencia o gráfico de f. Considerando essa introdução e motivação geométrica, definimos então funções conjugadas.

Definição 2.1 *Dada uma função $f : \mathbb{R}^l \to \mathbb{R}$, a função conjugada de f, $f^* : \mathbb{R}^l \to \mathbb{R} \cup \{+\infty\}$ é definida por*

$$f^*(u) = \sup_{x \in \mathbb{R}^l} \{ u^T x - f(x) \}. \tag{2.2}$$

Essa função é conhecida como Conjugada de Fenchel ou Conjugada Clássica. Fazendo uma associação à motivação de conjugação apresentada, para uma função f convexa, temos que o hiperplano

$$H = \left\{ (x, \lambda) \mid u^T x - \lambda = f^*(u),\ u \in \mathbb{R}^n \right\}, \tag{2.3}$$

suporta o epigrafo de f no ponto em que tangencia o gráfico de f.

A fim de ilustrar a definição de função conjugada, considere o seguinte exemplo.

Exemplo 2.1 *Seja a função $f : \mathbb{R} \to \mathbb{R}$ dada por $f(x) = e^x$. Aplicando a Definição 2.1 temos que a conjugada de f é dada por*

$$f^*(u) = \sup_{x \in \mathbb{R}} \{ ux - e^x \},$$

para todo $u \in \mathbb{R}$. Para o caso em que $u < 0$, temos que a função definida como $h(x) = ux - e^x$ cresce infinitamente à medida que x decresce. Logo, temos que $f^*(u) = +\infty$. Considerando $u = 0$, temos que $f^*(u) = \sup_{x \in \mathbb{R}} \{-e^x\} = 0$. Finalmente, para o caso em

que $u > 0$, devemos analisar o comportamento da função $h(x) = ux - e^x$. Note que, neste caso, h possui maximizador global igual a $\bar{x} = \ln u$. Logo, a conjugada é definida como $f^*(u) = \sup_{x\in\mathbb{R}} \{ux - e^x\} = \max_{x\in\mathbb{R}} \{ux - e^x\} = u \ln u - u$. *Sendo assim, temos que*

$$f^*(u) = \begin{cases} +\infty, \ se\ u < 0 \\ 0, \ se\ u = 0 \\ u \ln u - u, \ se\ u > 0. \end{cases}$$

Uma ilustração para a reta suporte de f, no caso $u > 0$, pode ser observada na Figura 2.2. Note que para $u < 0$ nenhuma reta na forma $\lambda = ux - b$ é suporte para epi(f).

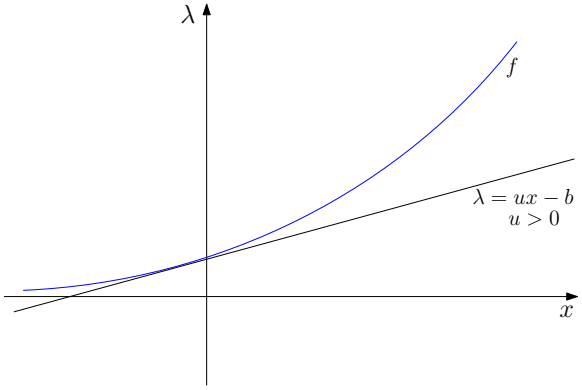


Figura 2.2: Ilustração para a reta λ suporte de f

Finalizada a discussão do exemplo, apresentamos a partir de agora algumas características das funções conjugadas. Conforme veremos no próximo resultado, se f é fortemente convexa[2], então $f^*(u)$ é finita para todo $u \in \mathbb{R}^l$.

Proposição 2.2 *Se $f : \mathbb{R}^l \to \mathbb{R}$ é uma função fortemente convexa, então $f^*(u) < +\infty$.*

Demonstração. Considere $u \in \mathbb{R}^l$ fixado. Como f é fortemente convexa, existe uma função quadrática q cuja Hessiana é positiva definida[3], tal que

$$f(x) \geq q(x), \forall x \in \mathbb{R}^l.$$

Assim, podemos escrever que

$$x^T u - f(x) \leq x^T u - q(x),$$

e usando o fato de que a função $x^T u - q(x)$ é quadrática, cuja Hessiana é negativa definida, então ela admite um maximizador $\bar{x}$, e concluímos assim que $f^*(u) \leq \bar{x}^T u - q(\bar{x}) < +\infty$. □

[2]Dado $\alpha > 0$, dizemos que a função $g : \mathbb{R}^l \to \mathbb{R}$ é $\alpha-$fortemente convexa quando para todo $x, y \in \mathbb{R}^l$ e $t \in [0, 1]$ temos que $g((1-t)x + ty) \leq (1-t)g(x) + tg(y) - \frac{\alpha}{2}(1-t)t \|x - y\|^2$.

[3]Seja $f : \mathbb{R}^l \to \mathbb{R}$ uma função $\alpha-$fortemente convexa. Podemos mostrar que $f(y) \geq f(x) + v(y - x) + \frac{\alpha}{2} \|x - y\|^2$ para todo $v \in \partial f(x)$. Ver Teorema 6.1.2 em [33].

O teorema a seguir apresenta algumas das propriedades que envolvem a função conjugada.

Teorema 2.3 *Sejam as funções* $f,\ g : \mathbb{R}^l \to \mathbb{R}$ *e* $f^*,\ g^* : \mathbb{R}^l \to \mathbb{R}$ *suas conjugadas, respectivamente. Então*

(i) $f(x) + f^*(u) \geq u^T x$, *para todo* $x, u \in \mathbb{R}^l$

(ii) $\inf\limits_{x \in \mathbb{R}^l} f(x) = -f^*(0)$

(iii) Se $f(x) \leq g(x)$ *para todo* $x \in \mathbb{R}^l$ *então* $f^*(u) \geq g^*(u)$ *para todo* $u \in \mathbb{R}^l$.

Demonstração. (i) Consequência da Definição 2.1.

(ii) Note que, usando a Definição 2.1, podemos escrever

$$-f^*(0) = -\sup_{x \in \mathbb{R}^l} \left\{0^T x - f(x)\right\} = \inf_{x \in \mathbb{R}^l} f(x).$$

(iii) A hipótese $f(x) \leq g(x)$ para todo $x \in \mathbb{R}^l$ nos garante que $u^T x - f(x) \geq u^T x - g(x)$.

Portanto,

$$f^*(u) = \sup_{x \in \mathbb{R}^l} \left\{u^T x - f(x)\right\} \geq \sup_{x \in \mathbb{R}^l} \left\{u^T x - g(x)\right\} = g^*(u),$$

para todo $u \in \mathbb{R}^l$, o que conclui a prova. □

O resultado a seguir estabelece uma condição para que na propriedade (i) do Teorema 2.3 possamos garantir a igualdade.

Proposição 2.4 *Seja* $f : \mathbb{R}^l \to \mathbb{R}$ *fortemente convexa e diferenciável. Considere* $x_0, u_0 \in \mathbb{R}^l$. *Então,* $f^*(u_0) + f(x_0) = x_0^T u_0$ *se, e somente se,* $u_0 = \nabla f(x_0)$ *e* $x_0 = \nabla f^*(u_0)$.

Demonstração. Primeiramente defina a função $h(x) = x^T u_0 - f(x)$ e suponha que $f^*(u_0) + f(x_0) = x_0^T u_0$. Pela definição de conjugada de Fenchel, podemos escrever que

$$f^*(u_0) \geq x^T u_0 - f(x), \tag{2.4}$$

para todo $x \in \mathbb{R}^l$. Temos que $h(x_0) = x_0^T u_0 - f(x_0) = f^*(u_0) \geq h(x)$, para todo $x \in \mathbb{R}^l$, ou seja, x_0 é maximizador de h. Portanto, $0 = \nabla h(x_0) = u_0 - \nabla f(x_0)$. De forma análoga, considerando a função $u \mapsto x_0^T u - f^*(u)$, podemos estabelecer que $x_0 = \nabla f^*(u_0)$.

Para provar a recíproca, suponha que $u_0 = \nabla f(x_0)$. Portanto, $\nabla h(x_0) = 0$. Isso implica que x_0 é um ponto crítico da função côncava h, logo maximizador global. Assim, temos que

$$h(x_0) \geq x^T u_0 - f(x),$$

para todo $x \in \mathbb{R}^l$. Em particular, temos que

$$h(x_0) \geq \sup_{x \in \mathbb{R}^l} \left\{ x^T u_0 - f(x) \right\} = f^*(u_0).$$

Por outro lado, $f^*(u_0) \geq x_0^T u_0 - f(x_0) = h(x_0)$. Logo, $f^*(u_0) + f(x_0) = x_0^T u_0$. □

Note que esse teorema está condicionado à existência de $x_0, u_0 \in \mathbb{R}^l$ tais que permitam a igualdade $f^*(u_0) + f(x_0) = x_0^T u_0$. A validade desse teorema é estabelecida com base no seguinte resultado.

Proposição 2.5 *Sejam $q : \mathbb{R}^l \to \mathbb{R}$ uma função quadrática côncava e $\xi : \mathbb{R}^l \to \mathbb{R}$ uma função contínua que satisfaz*

$$\xi(x) \leq q(x).$$

Então, existe x^* tal que $\xi(x) \leq \xi(x^*)$ para todo $x \in \mathbb{R}^l$.

Demonstração. Sabemos que ξ é limitada superiormente por uma função quadrática côncava, logo existe $M \in \mathbb{R}$ tal que $M = \sup_{x \in \mathbb{R}^l} \{\xi(x)\}$, ou seja, podemos afirmar que existe uma sequência $(x^k) \subset \mathbb{R}^l$ tal que

$$\xi(x^k) \to M. \tag{2.5}$$

Afirmamos que (x^k) é limitada. De fato, suponha por absurdo que a sequência (x^k) é ilimitada. Então, existe uma subsequência de (x^k), digamos $(x^k)_{k \in \mathbb{N}'}$ tal que $\left\|x^k\right\| \overset{\mathbb{N}'}{\to} +\infty$. Portanto, temos que $q(x^k) \overset{\mathbb{N}'}{\to} -\infty$, e como $\xi(x) \leq q(x)$, concluímos que $\xi(x^k) \overset{\mathbb{N}'}{\to} -\infty$, o que é uma contradição tendo em vista (2.5). Sendo assim, como (x^k) é limitada, existe uma subsequência $(x^k)_{k \in \mathbb{N}''}$ tal que $x^k \overset{\mathbb{N}''}{\to} x^*$, logo

$$\xi(x^k) \overset{\mathbb{N}''}{\to} \xi(x^*).$$

Assim, usando (2.5) temos que $M = \xi(x^*)$, o que implica que a função ξ admite um máximo. □

Podemos empregar esse resultado para validar as hipóteses da Proposição 2.4. Para isso, note que pelo fato da função f ser fortemente convexa, a função definida por $h(x) = u_0^T x - f(x)$ é limitada superiormente por uma função quadrática côncava q. Portanto, pela Proposição 2.5, a função h admite um maximizador. Logo, existe $x_0 \in \mathbb{R}^l$ tais que $f^*(u_0) = u_0^T x_0 - f(x_0)$.

Estabelecidos os conceitos fundamentais de conjugação, apresentamos a seguir conceitos básicos em dualidade Lagrangiana. Tal estudo é o que motiva a construção do par de problemas considerado nesta tese.

2.2 Dualidade Lagrangiana

No contexto de otimização, a teoria de dualidade baseia-se em associar a um problema de minimização de funções (*primal*), um outro problema, chamado *dual*, que sob certas condições é equivalente ao primal e pode ser de resolução mais fácil. Além disso, quando o problema primal é de natureza convexa, as relações de dualidade estabelecidas são consideradas mais fortes. Discutimos a seguir alguns conceitos de dualidade Lagrangiana que podem ser encontrados em [5, 39, 49].

Para apresentar o esquema de dualidade a ser empregado nesta tese, vamos considerar o problema primal como

$$\begin{array}{rl} \min & f(x) \\ s.a & x \in D, \end{array} \tag{2.6}$$

em que $D = \{x \in \Omega \mid h(x) = 0, g(x) \leq 0\}$, $\Omega \subset \mathbb{R}^n$, $f : \Omega \to \mathbb{R}$, $h : \Omega \to \mathbb{R}^l$ e $g : \Omega \to \mathbb{R}^m$. Nosso primeiro passo é reformulá-lo por meio da sua função Lagrangiana. Para isso, recordamos que a Lagrangiana associada a esse problema é dada por

$$\begin{array}{c} L : \Omega \times \mathbb{R}^l \times \mathbb{R}^m \to \mathbb{R} \\ L(x, \lambda, \mu) = f(x) + \lambda^T h(x) + \mu^T g(x), \end{array} \tag{2.7}$$

e que uma relação entre (2.6) e (2.7) pode ser expressa no teorema a seguir, conforme [39].

Teorema 2.6 *Dado o problema primal (2.6) e L a sua Lagrangiana associada, definida em (2.7), temos que*

$$\sup_{(\lambda,\, \mu) \in \mathbb{R}^l \times \mathbb{R}^m_+} L(x, \lambda, \mu) = \begin{cases} f(x), \ \textit{se } x \in D, \\ +\infty, \ \textit{se } x \in \Omega \backslash D. \end{cases}$$

Demonstração. Inicialmente, note que se $x \in D$ para todo $(\lambda, \mu) \in \mathbb{R}^l \times \mathbb{R}^m_+$ temos que $L(x, \lambda, \mu) \leq f(x)$, pois $\lambda^T h(x) = 0$ e $\mu^T g(x) \leq 0$. Além disso, temos que $L(x, \lambda, 0) = f(x)$. Portanto, para este caso temos que

$$\sup_{(\lambda,\mu) \in \mathbb{R}^l \times \mathbb{R}^m_+} L(x, \lambda, \mu) = f(x).$$

Considere, agora, $x \in \Omega \backslash D$. Neste caso, existe $j \in \{1, \ldots, l\}$ tal que $h_j(x) \neq 0$ e/ou existe $j \in \{1, \ldots, m\}$ tal que $g_j(x) > 0$. Supondo $h_j(x) \neq 0$ para algum $j \in \{1, \ldots, l\}$, definimos

$$\lambda_i = \begin{cases} t h_i(x), \text{ se } i = j, t > 0 \\ 0, \text{ se } i \in \{1, \ldots, l\} \backslash \{j\}. \end{cases}$$

Tomando $\mu_i = 0,\ i = 1, \ldots, m$, obtemos que

$$L(x, \lambda, \mu) = f(x) + t(h_j(x))^2 \to +\infty \text{ quando } t \to +\infty.$$

Por outro lado, suponha agora que $g_j(x) > 0$ para algum $j \in \{1, \ldots, m\}$. Considerando $\lambda_i = 0,\ i = 1, \ldots, l$, e

$$\mu_i = \begin{cases} t, \text{ se } i = j, t > 0 \\ 0, \text{ se } i \in \{1, \ldots, m\} \backslash \{j\}, \end{cases}$$

obtemos que

$$L(x, \lambda, \mu) = f(x) + tg_j(x) \to +\infty \text{ quando } t \to +\infty,$$

o que conclui a prova. □

Baseado neste teorema o problema primal (2.6) pode ser reformulado como

$$\begin{aligned} \min \quad & \left\{ \sup_{(\lambda,\mu) \in \mathbb{R}^l \times \mathbb{R}^m_+} L(x, \lambda, \mu) \right\} \\ s.a \quad & x \in D, \end{aligned} \tag{2.8}$$

onde $D = \left\{ x \in \Omega \mid \sup_{(\lambda,\mu) \in \mathbb{R}^l \times \mathbb{R}^m_+} L(x, \lambda, \mu) < +\infty \right\}$. No entanto, uma pergunta natural que surge é se a ordem das operações de minimizar e maximizar influencia na resolução do problema (2.8), ou seja, será que é possível optar pela ordem mais conveniente? Essa questão é o que motiva a construir o problema dual associado, como

$$\begin{aligned} \max \quad & \left\{ \inf_{x \in \Omega} L(x, \lambda, \mu) \right\} \\ s.a \quad & (\lambda, \mu) \in \Delta, \end{aligned} \tag{2.9}$$

em que $\Delta = \left\{ (\lambda, \mu) \in \mathbb{R}^l \times \mathbb{R}^m_+ \mid \inf_{x \in D} L(x, \lambda, \mu) > -\infty \right\}$. Definindo, então, a função dual

$$\varphi : \Delta \to \mathbb{R}, \quad \varphi(\lambda, \mu) = \inf_{x \in \Omega} L(x, \lambda, \mu),$$

o problema dual pode ser reescrito como

$$\begin{aligned} \max \quad & \varphi(\lambda, \mu) \\ s.a \quad & (\lambda, \mu) \in \Delta. \end{aligned} \tag{2.10}$$

Essa estrutura de análise nos permite estabelecer a relação entre as formulações primal e dual, dadas em (2.6) e (2.10), respectivamente, embasados pela dualidade convexa. Neste contexto, levantamos alguns resultados fundamentais associando tais problemas, que serão enunciados a seguir e encontram-se demonstrados em [39].

Inicialmente, veremos que o problema dual considerado consiste na maximiza-

ção de uma função côncava num conjunto convexo, ou seja, toda solução local é global e o conjunto de soluções é convexo. Esse resultado é garantido pelo teorema a seguir.

Teorema 2.2.1 *Para qualquer problema primal do tipo (2.6), o conjunto viável Δ do problema dual (2.10) é convexo e a função dual $\varphi : \Delta \to \mathbb{R}$ é côncava.*

Demonstração. Veja o Teorema 5.2.1 em [39]. □

A título de ilustração apresentamos uma interpretação geométrica para o problema dual, proposta em [52]. Para isso, considere o problema primal na forma

$$\begin{aligned} \min \quad & f(x) \\ s.a \quad & g_1(x) \leq 0 \\ & g_2(x) \leq 0 \\ & x \in X \subseteq \mathbb{R}, \end{aligned}$$

e o seu dual como sendo

$$\begin{aligned} \max_{\mu \geq 0} \quad & \left\{ \min_{x \in X} L(x, \mu) \right\} \\ s.a \quad & \inf_{x \in X} L(x, \mu) > -\infty, \end{aligned}$$

em que $L(x, \mu) = f(x) + \mu_1 g_1(x) + \mu_2 g_2(x)$. Considere o conjunto $I \subseteq \mathbb{R}^3$ a imagem de X sob as três funções, f, g_1 e g_2, ou seja, $I = \{(z_1, z_2, z_3) \in \mathbb{R}^3 \mid z_1 = g_1(x),\ z_2 = g_2(x)$ e $z_3 = f(x)$ para algum $x \in X\}$. Sobre este conjunto, a formulação equivalente do dual é dada por

$$\begin{aligned} \max_{\mu \geq 0} \quad & \left\{ \min_{z \in I} \tilde{L}(z, \mu) \right\} \\ s.a \quad & \inf_{z \in I} \tilde{L}(z, \mu) > -\infty, \end{aligned}$$

onde $\tilde{L}(z, \mu) = z_3 + \mu_1 z_1 + \mu_2 z_2$. Para valores fixos dos multiplicadores $\bar{\mu}_1 \geq 0$ e $\bar{\mu}_2 \geq 0$, temos que qualquer curva de nível da função $\tilde{L}(z, \bar{\mu})$ corresponde a um plano no $\mathbb{R}^3$.

Neste contexto, podemos interpretar geometricamente as soluções dos problemas primal e dual, conforme ilustrado na Figura 2.3. O ponto de interseção de I com o eixo z_3, denotado por P^*, é a imagem da solução ótima, x^*, do problema primal. A função dual $\varphi(\mu)$ no ponto $(\bar{\mu_1}, \bar{\mu_2})$ corresponde à determinação do plano mais baixo e paralelo a $\tilde{L}(z, \mu)$ que intercepta o conjunto I. Isto corresponde ao hiperplano suporte π tangente ao conjunto I no ponto $\bar{P}$ representado. O ponto P^* corresponde aos valores de $\bar{\mu}_1$ e $\bar{\mu}_2$ que maximizam z_3. Portanto, o problema dual consiste na determinação dos valores de $\bar{\mu}_1$ e $\bar{\mu}_2$ que definem a inclinação do hiperplano suporte do conjunto I, tal que a interseção seja com a maior cota, representada por $\bar{z_3}$.

Nem sempre podemos garantir a existência de uma equivalência entre as formulações primal e dual, porém, há uma rica relação entre suas funções objetivo, estabelecida no seguinte teorema.

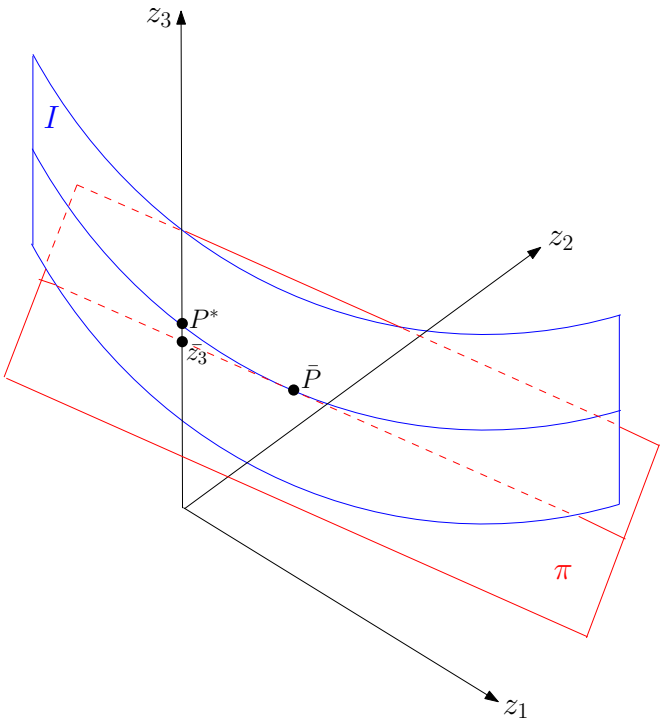


Figura 2.3: Representação geométrica do problema dual

Teorema 2.2.2 *(Teorema da Dualidade fraca) Para qualquer par de problemas primal e dual, tem-se que*

$$\varphi(\lambda, \mu) \leq f(x), \forall x \in D, \forall (\lambda, \mu) \in \Delta.$$

Em particular

$$\bar{w} = \sup_{(\lambda,\mu)\in\Delta} \varphi(\lambda, \mu) \leq \inf_{x\in D} f(x) = \bar{v}.$$

Demonstração. Veja o Teorema 5.2.2 em [39]. □

Caso as soluções ótimas sejam distintas, $\bar{w} < \bar{v}$, temos que há uma *brecha de dualidade*, ou *gap* de dualidade. No entanto, o próximo teorema nos fornece condições que garantem a não existência do *gap* de dualidade, ou seja, quando os valores ótimos dos problemas primal e dual coincidem.

Teorema 2.2.3 *(Teorema da Dualidade Forte) Sejam $\Omega = \mathbb{R}^n$, $f : \mathbb{R}^n \to \mathbb{R}$ e $g : \mathbb{R}^n \to \mathbb{R}^m$ funções convexas e $h : \mathbb{R}^n \to \mathbb{R}^l$ uma função afim. Suponha que o problema primal (2.6) satisfaça a condição de Slater:*

$$\exists\, \widehat{x} \in \mathbb{R}^n \text{ tal que } h(\widehat{x}) = 0 \text{ e } g_i(\widehat{x}) < 0, i = 1, \ldots, m.$$

Se o valor ótimo de (2.6) é finito, ou seja, $\bar{v} > -\infty$ (por exemplo, quando o problema primal tem uma solução), então o problema dual (2.9) possui uma solução e não há gap de dualidade.

Demonstração. Veja o Teorema 5.2.3 em [39]. □

Deste modo, vemos que quando as hipóteses do teorema anterior forem satisfeitas, temos liberdade de escolha do problema a ser resolvido (primal ou dual). Pelo Teorema 2.2.2, temos garantia de que a função dual avaliada em um ponto viável do dual sempre fornece uma cota inferior para o valor ótimo do problema primal.

Nosso foco agora é aplicar esta teoria para uma classe particular de problemas. Temos interesse nessa aplicação, pois se trata do nosso objeto de trabalho e a dualidade envolvida nos permitirá o desenvolvimento dos algoritmos propostos na tese.

2.2.1 Dualidade aplicada a uma classe de problemas

Considere $A_1, \ldots, A_n \in \mathbb{R}^{d\times m}$, $\lambda > 0$, $w \in \mathbb{R}^d$, $N = nm$, $\beta \in \mathbb{R}^N$ e a função

$$P(w) = \frac{1}{n}\sum_{i=1}^{n} \phi_i(A_i^T w) + \lambda g(w), \tag{2.11}$$

onde $\phi_i : \mathbb{R}^m \to \mathbb{R}$ e $g : \mathbb{R}^d \to \mathbb{R}$ são funções fortemente convexas. Vamos aplicar a teoria de dualidade sob a consideração de que o problema primal é dado por

$$\min_{w\in\mathbb{R}^d} P(w), \tag{2.12}$$

e que pode ser escrito como

$$\begin{array}{ll} \min & \left\{\dfrac{1}{n}\displaystyle\sum_{i=1}^{n} \phi_i(z_i) + \lambda g(w)\right\} \\ s.a & z_i - A_i^T w = 0,\ i = 1, \ldots, n. \end{array} \tag{2.13}$$

Pelo esquema da seção anterior, a função dual deste problema é dada por

$$\varphi(\beta) = \inf_{w\in\mathbb{R}^d,\, z_i\in\mathbb{R}^m} L(w, z_1, \ldots, z_n, \beta_1, \ldots, \beta_n),$$

em que $L(w, z_1, \ldots, z_n, \beta_1, \ldots, \beta_n) = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} \phi_i(z_i) + \lambda g(w) + \sum_{i=1}^{n} \beta_i^T \left(z_i - A_i^T w\right)$. Deste modo, temos que

$$\begin{aligned} \varphi(\beta) = & \sum_{i=1}^{n} \inf_{z_i\in\mathbb{R}^m} \left\{\beta_i^T z_i + \frac{1}{n}\phi_i(z_i)\right\} + \inf_{w\in\mathbb{R}^d} \left\{\left(-\sum_{i=1}^{n} A_i\beta_i\right)^T w + \lambda g(w)\right\} \\ = & -\frac{1}{n}\sum_{i=1}^{n} \sup_{z_i\in\mathbb{R}^m} \left\{-n\beta_i^T z_i - \phi_i(z_i)\right\} - \lambda \sup_{w\in\mathbb{R}^d} \left\{\left(\sum_{i=1}^{n} \frac{A_i\beta_i}{\lambda}\right)^T w - g(w)\right\} \\ = & -\frac{1}{n}\sum_{i=1}^{n} \phi_i^*(-n\beta_i) - \lambda g^*\left(\frac{1}{\lambda}\sum_{i=1}^{n} A_i\beta_i\right), \end{aligned} \tag{2.14}$$

em que ϕ_i^* e g^* são as conjugadas de ϕ_i e g, respectivamente, obtidas de acordo com a

Definição 2.1. Além disso, fazendo uma mudança de variáveis da forma $\alpha_i = n\beta_i$, com $i = 1, \ldots, n$, podemos escrever que

$$\varphi(\beta) \stackrel{\text{def}}{=} D(\alpha) = -\frac{1}{n}\sum_{i=1}^{n}\phi_i^*(-\alpha_i) - \lambda g^*\left(\frac{1}{\lambda n}\sum_{i=1}^{n}A_i\alpha_i\right). \tag{2.15}$$

Como $D(\alpha) > -\infty$ para todo α, o problema dual de (2.12) é dado por

$$\max_{\alpha\in\mathbb{R}^N} D(\alpha). \tag{2.16}$$

No capítulo a seguir estabelecemos o problema a ser tratado, assim como os métodos propostos para a sua resolução. Mostraremos que os problemas primal e dual considerados são casos particulares dos problemas (2.12) e (2.16), respectivamente. Veremos que, para estes casos, os Teoremas 2.2.2 e 2.2.3 serão válidos.

Capítulo 3

Métodos Propostos

Neste capítulo apresentaremos a formulação primal-dual do problema *Ridge Regression*, foco da pesquisa. Discutiremos a sistematização dos métodos propostos neste trabalho para a resolução do problema, os quais nomeamos algoritmos SRP e acc-SRP. Tais algoritmos serão estabelecidos no formato ponto fixo e, a partir disso, discutiremos as propriedades da análise de convergência teórica. Destacamos que a relevância teórica do trabalho está embasada no tratamento do problema na versão primal-dual. Por simplificação de notação, denotamos $\|\cdot\|_2$ por $\|\cdot\|$ e empregamos uma notação conveniente para os parâmetros.

3.1 O problema

Inicialmente, considere matrizes $A_1, \ldots, A_n \in \mathbb{R}^{d\times m}$, vetores $y_1, \ldots, y_n \in \mathbb{R}^m$ e um parâmetro de regularização $\lambda > 0$. Denotamos

$$A = (A_1 \ \cdots \ A_n) \in \mathbb{R}^{d\times N} \quad \text{e} \quad y = (y_1^T \ \cdots \ y_n^T)^T \in \mathbb{R}^N, \tag{3.1}$$

onde $N = nm$. Neste trabalho, consideramos o problema de minimização regularizado

$$\min_{w\in\mathbb{R}^d} P(w) \stackrel{\text{def}}{=} \frac{1}{2n}\|A^T w - y\|^2 + \frac{\lambda}{2}\|w\|^2 \tag{3.2}$$

e associamos o problema dual

$$\max_{\alpha\in\mathbb{R}^N} D(\alpha) \stackrel{\text{def}}{=} -\frac{1}{2n^2\lambda}\|A\alpha\|^2 - \frac{1}{2n}\|\alpha\|^2 + \frac{1}{n}\alpha^T y. \tag{3.3}$$

Mostraremos que os problemas (3.2) e (3.3) são casos particulares de (2.12) e (2.16), respectivamente, e conforme discutido no Capítulo 2, formam um par de problemas primal-dual. Note que devido à sua natureza o problema (3.2) pode ser tratado como um problema do tipo *Ridge Regression*, conforme destacado na Seção 1.3, por ser a minimização de uma espécie de média de funções convexas suaves penalizada

sob uma função fortemente convexa.

Sendo assim, nosso objetivo é analisar o par de problemas primal-dual por meio de uma mistura das variáveis primais e duais, a fim de englobarmos as vantagens em se trabalhar com tais variáveis, conjuntamente. Para isso, definindo $x = \begin{pmatrix} w \\ \alpha \end{pmatrix} \in \mathbb{R}^{d+N}$, nosso problema primal-dual tem natureza fortemente convexa e quadrática, consistindo na minimização do *gap* de dualidade, expresso por

$$\min_{x \in \mathbb{R}^{d+N}} f(x) = P(w) - D(\alpha), \tag{3.4}$$

em que P e D são definidas em (3.2) e (3.3), respectivamente. Considerando o problema no formato geral, dado em (3.4), apresentamos a seguir algumas discussões acerca das características de otimalidade e dualidade sob o mesmo e a relação de equivalência entre as formulações primal e dual, no contexto de dualidade convexa.

3.1.1 Condições de otimalidade e relações de dualidade

A princípio, note que para o problema dado em (3.4), temos que

$$\nabla f(x) = \begin{pmatrix} \frac{1}{n} A \left(A^T w - y \right) + \lambda w \\ \frac{1}{n^2 \lambda} A^T A \alpha + \frac{1}{n} \alpha - \frac{1}{n} y \end{pmatrix}, \tag{3.5}$$

e, a condição de otimalidade, $\nabla f(x) = 0$, pode ser reescrita como

$$\begin{pmatrix} \frac{1}{n} A A^T + \lambda I & 0 \\ 0 & \frac{1}{n^2 \lambda} A^T A + \frac{1}{n} I \end{pmatrix} \begin{pmatrix} w \\ \alpha \end{pmatrix} = \frac{1}{n} \begin{pmatrix} Ay \\ y \end{pmatrix}. \tag{3.6}$$

Neste formato matricial, com $\lambda, n > 0$, podemos verificar que a matriz dos coeficientes do sistema é definida positiva, o que implica na existência e unicidade da solução do sistema, a qual representamos por $x^* = \begin{pmatrix} w^* \\ \alpha^* \end{pmatrix} \in \mathbb{R}^{d+N}$.

Para discutir as relações que envolvem os problemas (3.2) e (3.3), buscamos reescrever as funções $P(w)$ e $D(\alpha)$, convenientemente, usando os conceitos de dualidade convexa apresentados no Capítulo 2.

Neste contexto, com base nas características da matriz $A \in \mathbb{R}^{d \times N}$ e dos vetores $y \in \mathbb{R}^N$ e $w \in \mathbb{R}^d$, podemos escrever que

$$A^T w - y = \begin{bmatrix} A_1^T w - y_1 \\ A_2^T w - y_2 \\ \vdots \\ A_n^T w - y_n \end{bmatrix},$$

o que nos leva a

$$\left\|A^T w - y\right\|^2 = \sum_{i=1}^{n} \left\|A_i^T w - y_i\right\|^2.$$

Assim, usando este fato, podemos reescrever P como

$$P(w) = \frac{1}{n}\sum_{i=1}^{n} \phi_i(A_i^T w) + \lambda g(w),$$

conforme definido em (2.11), em que

$$\phi_i(z) = \frac{1}{2}\|z - y_i\|^2 \quad \text{e} \quad g(w) = \frac{1}{2}\|w\|^2. \tag{3.7}$$

Similarmente, com o intuito de reescrevermos a função D, devemos encontrar as conjugadas das funções ϕ_i e g. Usando a função ϕ_i, definida em (3.7), e a definição de Conjugada de Fenchel (Definição 2.1), podemos escrever que

$$\phi_i^*(s) = \sup_{z\in\mathbb{R}^m} \left\{ s^T z - \frac{1}{2}\|z - y_i\|^2 \right\}. \tag{3.8}$$

Como a função $h(z) \stackrel{\text{def}}{=} s^T z - \frac{1}{2}\|z - y_i\|^2$ é quadrática côncava, o ponto crítico $\bar{z} = s + y_i$ é maximizador, e assim, concluímos que

$$\phi_i^*(s) = h(\bar{z}) = \frac{1}{2}\|s\|^2 + s^T y_i. \tag{3.9}$$

Analogamente, podemos verificar que a conjugada da função g, definida em (3.7), é dada por

$$g^*(u) = \frac{1}{2}\|u\|^2 \tag{3.10}$$

e, assim, podemos reescrever $D(\alpha)$, definida em (3.3), como

$$\begin{aligned}
D(\alpha) = & -\frac{1}{2n^2\lambda}\|A\alpha\|^2 - \frac{1}{2n}\|\alpha\|^2 + \frac{1}{n}\alpha^T y \\
= & -\frac{1}{2n^2\lambda}\left\|\sum_{i=1}^{n} A_i\alpha_i\right\|^2 - \frac{1}{2n}\sum_{i=1}^{n}\|\alpha_i\|^2 - \frac{1}{n}\sum_{i=1}^{n}(-\alpha_i)^T y_i \\
= & -\frac{\lambda}{2}\left\|\frac{1}{n\lambda}\sum_{i=1}^{n} A_i\alpha_i\right\|^2 - \frac{1}{2n}\sum_{i=1}^{n}\|\alpha_i\|^2 - \frac{1}{n}\sum_{i=1}^{n}(-\alpha_i)^T y_i \\
= & -\lambda g^*\left(\frac{1}{n\lambda}\sum_{i=1}^{n} A_i\alpha_i\right) - \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{2}\|-\alpha_i\|^2 + (-\alpha_i)^T y_i\right) \\
= & -\lambda g^*\left(\frac{1}{n\lambda}\sum_{i=1}^{n} A_i\alpha_i\right) - \frac{1}{n}\sum_{i=1}^{n}\phi_i^*(-\alpha_i)
\end{aligned} \tag{3.11}$$

conforme função estabelecida em (2.15). Assim, estabelecemos a equivalência entre o par de problemas (3.2) e (3.3) com (2.12) e (2.16), respectivamente, comprovando a relação primal-dual entre as formulações.

Se definirmos

$$\bar{\alpha} = \frac{1}{n\lambda}\sum_{i=1}^{n} A_i\alpha_i, \tag{3.12}$$

o *gap* de dualidade entre os problemas (3.2) e (3.3) pode ser reescrito como

$$P(w)-D(\alpha) = \lambda\big(g(w)+g^*(\bar{\alpha})-w^T\bar{\alpha}\big)+\frac{1}{n}\sum_{i=1}^{n}\Big(\phi_i(A_i^Tw)+\phi_i^*(-\alpha_i)+\alpha_i^TA_i^Tw\Big). \tag{3.13}$$

Além disso, note que a partir da definição de função conjugada podemos escrever que

$$g^*(\bar{\alpha}) \geq w^T\bar{\alpha} - g(w). \tag{3.14}$$

Analogamente,

$$\phi_i^*(-\alpha_i) \geq -\alpha_i^TA_i^Tw - \phi_i(A_i^Tw). \tag{3.15}$$

Assim, usando (3.14) e (3.15) em (3.13), podemos então concluir que

$$P(w) - D(\alpha) \geq 0,$$

ou seja, comprovamos a dualidade fraca, dada pelo Teorema 2.2.2.

Podemos, também, verificar que a dualidade forte, Teorema 2.2.3, é equivalente a

$$w = \nabla g^*(\bar{\alpha}) \quad \text{e} \quad \alpha_i = -\nabla\phi_i(A_i^Tw). \tag{3.16}$$

Para comprovar isso, basta aplicar a Proposição 2.4 para as funções g e ϕ_i. Em virtude de (3.7), (3.10) e (3.12) podemos reescrever (3.16) como

$$w = \bar{\alpha} = \frac{1}{n\lambda}A\alpha \ \text{ e } \ \alpha = y - A^Tw,$$

que é exatamente uma forma diferente para reescrever a solução do sistema (3.6).

Baseado neste contexto de dualidade convexa, mostraremos a seguir novas metodologias aplicadas ao problema primal-dual de *Ridge Regression*, definido em (3.4). Nosso diferencial consiste em trabalhar com a formulação primal-dual recaindo em um problema no formato ponto fixo, sob o qual analisamos a convergência linear. Apresentamos a seguir o método SRP, e uma reformulação acelerada, nomeada método acc-SRP, propostos neste trabalho.

3.2 O método SRP

O método SRP foi estabelecido a partir da condição de otimalidade do problema (3.4) no formato matricial, dado em (3.6). Multiplicando a equação (3.6) por $\frac{\theta}{\lambda}$, para $\theta, \lambda > 0$, e adicionando o vetor $\begin{pmatrix} w \\ \alpha \end{pmatrix}$ em ambos os membros da equação, recaímos no sistema

$$\begin{cases} w = w - \theta w - \dfrac{\theta}{n\lambda} AA^T w + \dfrac{\theta}{n\lambda} Ay \\ \alpha = \alpha - \dfrac{\theta}{n\lambda}\alpha - \dfrac{\theta}{(n\lambda)^2} A^T A\alpha + \dfrac{\theta}{n\lambda} y \end{cases},$$

que pode ser reescrito no formato

$$x = G(\theta)x + \theta b, \tag{3.17}$$

em que $x = \begin{pmatrix} w \\ \alpha \end{pmatrix}$, $b = \dfrac{1}{n\lambda}\begin{pmatrix} Ay \\ y \end{pmatrix}$ e

$$G(\theta) = \begin{pmatrix} (1-\theta)\, I - \dfrac{\theta}{n\lambda} AA^T & 0 \\ 0 & \left(1 - \dfrac{\theta}{n\lambda}\right) I - \dfrac{\theta}{(n\lambda)^2} A^T A \end{pmatrix}. \tag{3.18}$$

Usando a relação (3.17) podemos estabelecer o algoritmo a seguir.

Algoritmo 3.1 Método SRP

DADOS: $x^0 \in \mathbb{R}^{d+N}$; $\theta > 0$;
INICIALIZAÇÃO: $k = 0$
REPITA
$x^{k+1} = G(\theta)x^k + \theta b$
$k = k + 1$

Uma vez que a solução x^* do problema satisfaz a relação (3.17), podemos utilizar as relações $x^{k+1} = G(\theta)x^k + \theta b$ e $x^* = G(\theta)x^* + \theta b$, iterativamente, e verificar que

$$\begin{aligned} x^1 - x^* &= G(\theta)(x^0 - x^*) \\ x^2 - x^* &= G(\theta)(x^1 - x^*) = G(\theta)^2(x^0 - x^*) \\ &\vdots \\ x^k - x^* &= G(\theta)(x^{k-1} - x^*) = G(\theta)^k(x^0 - x^*), \end{aligned} \tag{3.19}$$

o que nos permite escrever que

$$\left\|x^k - x^*\right\| \leq \left\|G(\theta)^k\right\| \left\|x^0 - x^*\right\|. \tag{3.20}$$

Sendo assim, nosso objetivo é encontrar um limite para o espectro da matriz $G(\theta)$, a fim de garantir a convergência do algoritmo tendo como base (3.20). Neste contexto, definimos o raio espectral de uma matriz $G \in \mathbb{R}^{q\times q}$ por

$$\rho(G) = \max\{|\zeta| \mid \zeta \text{ é um autovalor de } G\}.$$

Uma relação que leva em consideração o espectro de uma matriz é dada por meio do seguinte resultado.

Teorema 3.1 *Seja $G \in \mathbb{R}^{q\times q}$. Então, $\lim_{k\to\infty} G^k = 0$ se, e somente se, $\rho(G) < 1$.*

Demonstração. Veja o Teorema 5.6.12 em [37]. □

Matrizes $G \in \mathbb{R}^{q\times q}$ tais que $\lim_{k\to\infty} G^k = 0$ são chamadas convergentes e por isso são de extrema importância na análise da convergência de processos iterativos. Com base em tais resultados podemos então afirmar que a convergência do Algoritmo 3.1 dependerá do espectro da matriz $G(\theta)$. Mais precisamente, há garantia de convergência se o raio espectral de $G(\theta)$ é menor do que 1, pois, neste caso, temos que $G(\theta)^k \to 0$ e, usando (3.20)

$$\left\|x^k - x^*\right\| \to 0.$$

Além disso, a convergência é linear tendo em vista (3.19).

Utilizando, então, o Teorema 3.1 realizamos a seguir um estudo sobre as garantias de convergência do Algoritmo SRP, baseado nos autovalores da matriz $G(\theta)$.

3.2.1 Análise de convergência do Método SRP

Os autovalores da matriz $G(\theta)$, definida em (3.18), podem ser calculados a partir da análise da sua estrutura especial. Para isso definimos, a partir dos blocos diagonais, duas matrizes

$$G_1(\theta) = (1-\theta)\, I - \frac{\theta}{n\lambda} AA^T \quad \text{e} \quad G_2(\theta) = \left(1 - \frac{\theta}{n\lambda}\right) I - \frac{\theta}{(n\lambda)^2} A^T A, \tag{3.21}$$

e os autovalores da matriz $G(\theta)$ passam a ser exatamente a união dos autovalores de $G_1(\theta)$ e $G_2(\theta)$, calculados separadamente. Além disso, pelas características de tais matrizes, seus autovalores podem ser obtidos por meio dos autovalores das matrizes AA^T e A^TA, pois os parâmetros θ, λ, e n são constantes. Assumimos aqui uma hipótese que ocorre em muitas aplicações práticas no contexto de regressão que é $d < n \leq N$.

Para estabelecer tais resultados, considere a decomposição em valores singulares da matriz A, que nos permite escrever que

$$A = U\Sigma V^T \tag{3.22}$$

em que $U \in \mathbb{R}^{d\times d}$ e $V \in \mathbb{R}^{N\times N}$ são matrizes ortogonais e

$$\Sigma = \begin{pmatrix} \widetilde{\Sigma} & 0 \\ 0 & 0 \end{pmatrix}, \tag{3.23}$$

com $\widetilde{\Sigma} = \text{diag}(\sigma_1, \ldots, \sigma_p)$, onde $\sigma_1 \geq \cdots \geq \sigma_p > 0$ representam os valores singulares da matriz A. Usando o fato de que os valores singulares de A correspondem às raízes quadradas dos autovalores não nulos das matrizes AA^T ou A^TA e, substituindo (3.22) em (3.21), concluímos que se posto$(A) < d$ então os autovalores de $G(\theta)$ são

$$\left\{1 - \theta - \frac{\theta\sigma_j^2}{n\lambda}\right\} \cup \left\{1 - \frac{\theta}{n\lambda} - \frac{\theta\sigma_j^2}{(n\lambda)^2}\right\} \cup \left\{1 - \theta, 1 - \frac{\theta}{n\lambda}\right\},\ j = 1, \ldots, p. \tag{3.24}$$

Por outro lado, se posto$(A) = d$, então a matriz AA^T não possui autovalores nulos. Logo, os autovalores de $G(\theta)$ são dados por

$$\left\{1 - \theta - \frac{\theta\sigma_j^2}{n\lambda}\right\} \cup \left\{1 - \frac{\theta}{n\lambda} - \frac{\theta\sigma_j^2}{(n\lambda)^2}\right\} \cup \left\{1 - \frac{\theta}{n\lambda}\right\},\ j = 1, \ldots, p. \tag{3.25}$$

Note que a existência do autovalor $\left\{1 - \frac{\theta}{n\lambda}\right\}$ não é afetada pelo posto de A pois, pela definição de A, a matriz A^TA sempre possuirá autovalores nulos. Assim, obtidos os autovalores de $G(\theta)$, podemos estabelecer o intervalo de variação do parâmetro θ para o qual o raio espectral de $G(\theta)$ seja menor que 1 e, consequentemente, garantir a convergência linear do algoritmo. Esse resultado é estabelecido no teorema a seguir.

Teorema 3.2 *Seja $x^0 \in \mathbb{R}^{d+N}$ um ponto inicial arbitrário e considere a sequência $(x^k)_{k\in\mathbb{N}}$ gerada pelo Algoritmo 3.1 com $n\lambda \geq 1$ e $\theta \in \left(0, \frac{2n\lambda}{n\lambda + \sigma_1^2}\right)$. Então, a sequência (x^k) converge linearmente para a solução do problema (3.4) com taxa de convergência*

$$\rho(\theta) = \max\left\{\left|1 - \theta - \frac{\theta\sigma_1^2}{n\lambda}\right|, \left|1 - \frac{\theta}{n\lambda}\right|\right\}.$$

Além disso, se escolhermos $\theta^* = \frac{2n\lambda}{n\lambda + \sigma_1^2 + 1}$, a taxa de convergência será ótima e dada por

$$\rho^* = \frac{\sigma_1^2 + n\lambda - 1}{\sigma_1^2 + n\lambda + 1}.$$

Demonstração. Com base nos autovalores da matriz $G(\theta)$ considere, sem perda de generalidade, funções modulares na variável θ, definidas por

$$y_1 = |1 - c\theta|,\ \ y_2 = \left|1 - \frac{c}{n\lambda}\theta\right|,\ \ y_3 = |1 - \theta| \quad \text{e} \quad y_4 = \left|1 - \frac{1}{n\lambda}\theta\right|,$$

em que $c = 1 + \dfrac{\sigma_1^2}{n\lambda}$, onde σ_1 representa o maior valor singular de A. Os zeros dessas funções são

$$\theta_1 = \frac{1}{c}, \quad \theta_2 = \frac{n\lambda}{c}, \quad \theta_3 = 1 \quad \text{e} \quad \theta_4 = n\lambda,$$

respectivamente. Uma representação gráfica para tais funções pode ser observada na Figura 3.1.

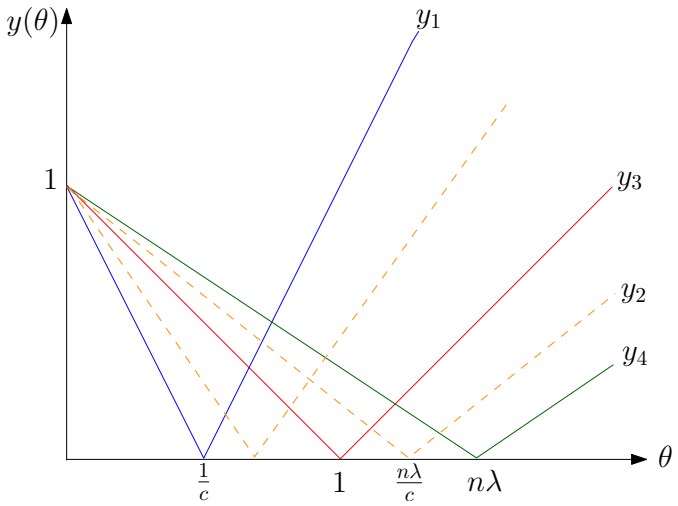


Figura 3.1: Representação das funções $y_i(\theta), i = 1, 2, 3, 4$ para o caso em que $n\lambda \geq 1$.

Note que $\theta_1 \leq \theta_3 \leq \theta_4$ e $\theta_1 \leq \theta_2 \leq \theta_4$. Além disso, o zero da função y_2 dependerá do valor de σ_1. Como o raio espectral da matriz $G(\theta)$ é dado por

$$\max_{1 \leq i \leq 4} \{y_i(\theta)\},$$

é fácil ver que y_2 e y_3 não interferem no cálculo do raio espectral. Sendo assim, sem perda de generalidade, para o caso $n\lambda \geq 1$, não há interferência do posto de A na análise, e o raio espectral de $G(\theta)$ passa a ser definido por

$$\rho(\theta) = \max \{y_1(\theta), y_4(\theta)\},$$

sendo explicitado graficamente na Figura 3.2.

Para obter a convergência linear, note que

$$\left\|x^{k+1} - x^*\right\| \leq \|G(\theta)\| \left\|x^k - x^*\right\| = \rho(\theta) \left\|x^k - x^*\right\|,$$

e a última igualdade é válida pois $G(\theta)$ é simétrica. Além disso, temos que $\rho(\theta) < 1$, se e somente se, $c\theta - 1 < 1$, o que implica que $\theta \in \left(0, \dfrac{2n\lambda}{n\lambda + \sigma_1^2}\right)$, comprovando a primeira afirmação.

Para finalizar a prova, note que pelo comportamento da função $\rho(\theta)$, o seu minimizador é calculado por meio da igualdade $1 - \dfrac{\theta}{n\lambda} = c\theta - 1$, resultando em $\theta^* =$

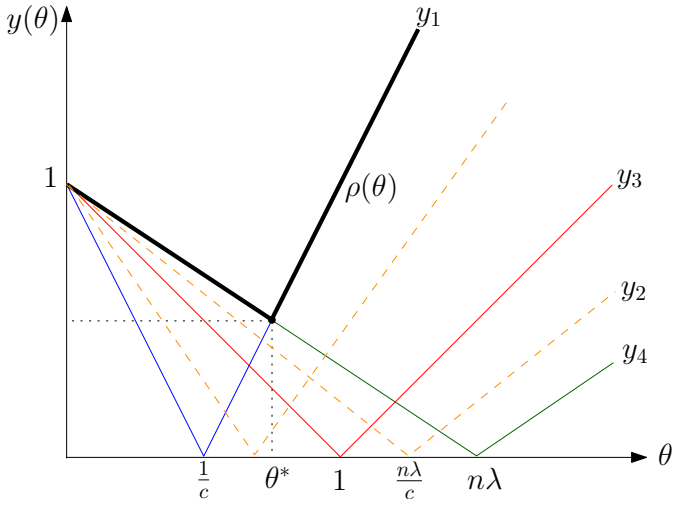


Figura 3.2: Representação da função ρ para o caso em que $n\lambda \geq 1$.

$\frac{2n\lambda}{n\lambda + \sigma_1^2 + 1}$. Além disso, usando o valor ótimo, θ^*, obtemos a taxa de convergência ótima $\rho^* = \frac{\sigma_1^2 + n\lambda - 1}{\sigma_1^2 + n\lambda + 1}$, completando a prova. □

Analogamente, podemos descrever o resultado da convergência para o caso em que $n\lambda < 1$. No entanto, agora a análise será influenciada pelo posto de A, pois a função y_3 também poderá definir a função $\rho(\theta)$. Uma representação gráfica para as funções y_i, com $i = 1, 2, 3, 4$, para o caso em que posto$(A) < d$, e para a função $\rho(\theta)$ é dada na Figura 3.3.

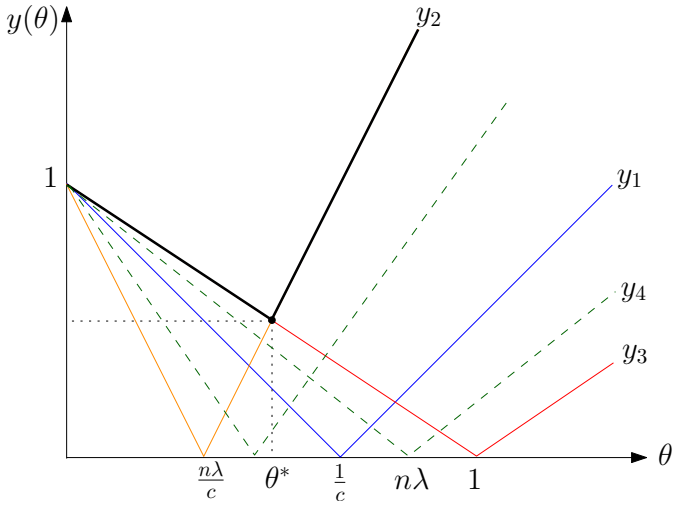
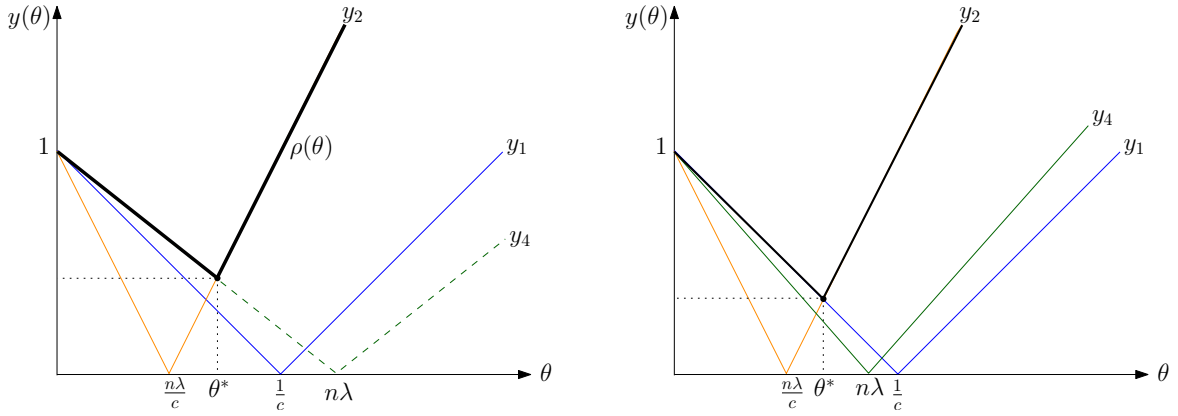


Figura 3.3: Representação das funções $y_i(\theta), i = 1, 2, 3, 4$ para $n\lambda < 1$ e posto$(A) < d$.

Para o caso em que posto$(A) = d$, não há existência do autovalor nulo para AA^T, logo, $1 - \theta$ não é autovalor de $G(\theta)$. A função $\rho(\theta)$ para este caso, passa a ser dividida em dois casos: $\frac{1}{c} \leq n\lambda < 1$ e $n\lambda < \frac{1}{c}$, conforme podemos observar na Figura 3.4.

Analogamente ao que foi feito no Teorema 3.2, podemos estender os resultados



(a) Caso $\frac{1}{c} \leq n\lambda < 1$ (b) Caso $n\lambda < \frac{1}{c}$

Figura 3.4: Representação das funções $y_i(\theta)$ para $n\lambda < 1$ e posto$(A) = d$

para os demais casos, resultando em três novos teoremas. Por simplificação, optamos apenas por explicitar os resultados. Sendo assim, resumimos as características espectrais para os casos em que posto$(A) < d$ e os apresentamos por meio da Tabela 3.1.

posto$(A) < d$	Intervalo	$\rho(\theta)$	ρ^*	θ^*
$n\lambda \geq 1$	$\left(0, \frac{2n\lambda}{n\lambda+\sigma_1^2}\right)$	$\max\left\{\left\|1-\theta-\frac{\theta\sigma_1^2}{n\lambda}\right\|, \left\|1-\frac{\theta}{n\lambda}\right\|\right\}$	$\frac{\sigma_1^2+n\lambda-1}{\sigma_1^2+n\lambda+1}$	$\frac{2n\lambda}{n\lambda+\sigma_1^2+1}$
$n\lambda < 1$	$\left(0, \frac{2(n\lambda)^2}{n\lambda+\sigma_1^2}\right)$	$\max\left\{\left\|1-\frac{\theta}{n\lambda}-\frac{\theta\sigma_1^2}{(n\lambda)^2}\right\|, \|1-\theta\|\right\}$	$\frac{\sigma_1^2+n\lambda-(n\lambda)^2}{\sigma_1^2+n\lambda+(n\lambda)^2}$	$\frac{2(n\lambda)^2}{\sigma_1^2+n\lambda+(n\lambda)^2}$

Tabela 3.1: Propriedades de convergência do método SRP quando posto$(A) < d$

Na Tabela 3.2 apresentamos também as características espectrais para os casos em que posto$(A) = d$.

posto$(A) = d$	Intervalo	$\rho(\theta)$	ρ^*	θ^*
$n\lambda \geq 1$	$\left(0, \frac{2n\lambda}{n\lambda+\sigma_1^2}\right)$	$\max\left\{\left\|1-\theta-\frac{\theta\sigma_1^2}{n\lambda}\right\|, \left\|1-\frac{\theta}{n\lambda}\right\|\right\}$	$\frac{\sigma_1^2+n\lambda-1}{\sigma_1^2+n\lambda+1}$	$\frac{2n\lambda}{n\lambda+\sigma_1^2+1}$
$\frac{1}{c} \leq n\lambda < 1$	$\left(0, \frac{2(n\lambda)^2}{n\lambda+\sigma_1^2}\right)$	$\max\left\{\left\|1-\frac{\theta}{n\lambda}-\frac{\theta\sigma_1^2}{(n\lambda)^2}\right\|, \left\|1-\frac{\theta}{n\lambda}\right\|\right\}$	$\frac{\sigma_1^2}{\sigma_1^2+2n\lambda}$	$\frac{2(n\lambda)^2}{\sigma_1^2+2n\lambda}$
$n\lambda < \frac{1}{c}$	$\left(0, \frac{2(n\lambda)^2}{n\lambda+\sigma_1^2}\right)$	$\max\left\{\left\|1-\frac{\theta}{n\lambda}-\frac{\theta\sigma_1^2}{(n\lambda)^2}\right\|, \left\|1-\theta\left(1+\frac{\sigma_1^2}{n\lambda}\right)\right\|\right\}$	$\frac{1-n\lambda}{1+n\lambda}$	$\frac{2(n\lambda)^2}{(\sigma_1^2+n\lambda)(n\lambda+1)}$

Tabela 3.2: Propriedades de convergência do método SRP quando posto$(A) = d$

Completamos, assim, a análise de convergência do Algoritmo 3.1 aplicado na resolução do problema definido em (3.4). Neste mesmo contexto, propomos uma versão acelerada para esse método, nomeada Método acc-SRP, discutida a seguir.

3.3 O método acc-SRP

Partindo do fato de que a solução ótima do nosso problema primal-dual, dado em (3.4), satisfaz a relação (3.17), a idéia aqui consiste em considerar um sistema aumentado e equivalente dado por

$$\begin{cases} x = z \\ x = G(\theta)x + \theta b \end{cases}. \tag{3.26}$$

Considerando um parâmetro não nulo $\gamma \in \mathbb{R}$, multiplicando a primeira equação por $1-\gamma$ e a segunda por γ, podemos escrever que

$$\begin{cases} x = (1-\gamma)z + \gamma(G(\theta)x + \theta b) \\ z = G(\theta)x + \theta b \end{cases}, \tag{3.27}$$

o que nos sugere o seguinte algoritmo.

Algoritmo 3.2 Método acc-SRP

DADOS: $x^0, z^0 \in \mathbb{R}^{d+N}; \theta, \gamma > 0;$
INICIALIZAÇÃO: $k = 0$
REPITA
$x^{k+1} = (1-\gamma)z^k + \gamma(G(\theta)x^k + \theta b)$
$z^{k+1} = G(\theta)x^k + \theta b$
$k = k + 1$

Note que se as sequências (x^k) e (z^k) convergem, digamos $x^k \to \bar{x}$ e $z^k \to \bar{z}$, então $\bar{x} = \bar{z} = \begin{pmatrix} w^* \\ \alpha^* \end{pmatrix}$. De fato, temos que $x^{k+1} \to \bar{x}$ e $z^{k+1} \to \bar{z}$. Por outro lado,

$$x^{k+1} = (1-\gamma)z^k + \gamma\left(G(\theta)x^k + \theta b\right) \to (1-\gamma)\bar{z} + \gamma\left(G(\theta)\bar{x} + \theta b\right)$$

e

$$z^{k+1} = G(\theta)x^k + \theta b \to G(\theta)\bar{x} + \theta b.$$

Portanto, $(1-\gamma)\bar{z} + \gamma\left(G(\theta)\bar{x} + \theta b\right) = \bar{x}$ e $G(\theta)\bar{x} + \theta b = \bar{z}$, donde segue a afirmação. Mostraremos a seguir que de fato as sequências (x^k) e (z^k) são convergentes. Além disso, numericamente, veremos que esta reformulação conduz a uma vantagem computacional significativa, uma vez que podemos alcançar uma convergência mais rápida em função

da escolha do parâmetro γ. Analisaremos, a seguir, a convergência téorica do algoritmo proposto.

3.3.1 Análise de convergência do Método acc-SRP

Nesta seção realizamos um estudo referente à convergência do Algoritmo 3.2. O objetivo é, para certos valores de γ, encontrar todos os valores do parâmetro θ para os quais há garantia de convergência para o algoritmo, assim como definirmos a taxa de convergência.

Para isso, note inicialmente que o Algoritmo 3.2 também pode ser visto no formato ponto fixo, uma vez que podemos escrever cada iteração do algoritmo na forma compacta

$$v^{k+1} = H_\gamma(\theta)v^k + r, \tag{3.28}$$

em que $v = \begin{pmatrix} x \\ z \end{pmatrix}$, $r = \begin{pmatrix} \gamma\theta b \\ \theta b \end{pmatrix}$ e

$$H_\gamma(\theta) = \begin{pmatrix} \gamma G(\theta) & (1-\gamma)I \\ G(\theta) & 0 \end{pmatrix}. \tag{3.29}$$

Assim, semelhante ao que foi discutido no Método SRP, em (3.19), podemos escrever que

$$v^k - v^* = H_\gamma(\theta)(v^{k-1} - v^*), \tag{3.30}$$

em que $v^* = \begin{pmatrix} x^* \\ x^* \end{pmatrix}$. Dessa forma, a convergência do Algoritmo 3.2 dependerá das propriedades espectrais da matriz de iteração $H_\gamma(\theta)$. Por uma questão de simplificação de notação denotaremos, a partir de agora, $H_\gamma(\theta)$ por H e $G(\theta)$, $G_1(\theta)$ e $G_2(\theta)$ por G, G_1 e G_2, respectivamente.

Assim, substituindo a matriz G, definida em (3.18), em (3.29) podemos escrever que

$$H = \begin{pmatrix} \gamma G_1 & 0 & (1-\gamma)I & 0 \\ 0 & \gamma G_2 & 0 & (1-\gamma)I \\ G_1 & 0 & 0 & 0 \\ 0 & G_2 & 0 & 0 \end{pmatrix}, \tag{3.31}$$

em que G_1 e G_2 foram definidas em (3.21). Para a análise dos autovalores da matriz H, nos lemas a seguir, definimos $q = p + 1$, $\sigma_q = 0$ e

$$\mu_{1j} = 1 + \frac{\sigma_j^2}{n\lambda}, \quad \mu_{2j} = \frac{1}{n\lambda}\left(1 + \frac{\sigma_j^2}{n\lambda}\right), \quad j = 1, \ldots, q. \tag{3.32}$$

Além disso, definimos as funções $\delta_{ij} : [0, \infty) \to \mathbb{R}, i = 1, 2,\ j = 1, \ldots, q$ por

$$\delta_{ij}(\theta) = \gamma^2(1 - \mu_{ij}\theta)^2 + 4(1 - \gamma)(1 - \mu_{ij}\theta), \tag{3.33}$$

e consideramos o seguinte resultado.

Proposição 3.3 *Seja $Q_j \in \mathbb{R}^{l\times l}, j = 1, 2, 3, 4$ matrizes diagonais cujos componentes das diagonais são $\alpha, \beta, \chi, \kappa, \in \mathbb{R}^l$, respectivamente. Então*

$$\det \begin{pmatrix} Q_1 & Q_2 \\ Q_3 & Q_4 \end{pmatrix} = \prod_{j=1}^{l} (\alpha_j\kappa_j - \beta_j\chi_j)\,.$$

Demonstração. Note que para $l = 1$, $Q_j \in \mathbb{R}, j = 1, 2, 3, 4$. Assim, por definição, temos que

$$\det \begin{pmatrix} \alpha & \beta \\ \chi & \kappa \end{pmatrix} = \alpha\kappa - \beta\chi. \tag{3.34}$$

Supondo agora o resultado válido para l, considere $Q_j \in \mathbb{R}^{\{l+1\}\times\{l+1\}}$ e a matriz

$$R = \begin{pmatrix} \alpha_1 & 0 & \cdots & 0 & 0 & \beta_1 & 0 & \cdots & 0 & 0 \\ 0 & \alpha_2 & \cdots & 0 & 0 & 0 & \beta_2 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \alpha_l & 0 & 0 & 0 & \cdots & \beta_l & 0 \\ 0 & 0 & \cdots & 0 & \alpha_{l+1} & 0 & 0 & \cdots & 0 & \beta_{l+1} \\ \chi_1 & 0 & \cdots & 0 & 0 & \kappa_1 & 0 & \cdots & 0 & 0 \\ 0 & \chi_2 & \cdots & 0 & 0 & 0 & \kappa_2 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \chi_l & 0 & 0 & 0 & \cdots & \kappa_l & 0 \\ 0 & 0 & \cdots & 0 & \chi_{l+1} & 0 & 0 & \cdots & 0 & \kappa_{l+1} \end{pmatrix}.$$

Note que permutando algumas linhas podemos escrever que

$$R^{'} = \begin{pmatrix} \alpha_1 & 0 & \cdots & 0 & 0 & \beta_1 & 0 & \cdots & 0 & 0 \\ 0 & \alpha_2 & \cdots & 0 & 0 & 0 & \beta_2 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \alpha_l & 0 & 0 & 0 & \cdots & \beta_l & 0 \\ \chi_1 & 0 & \cdots & 0 & 0 & \kappa_1 & 0 & \cdots & 0 & 0 \\ 0 & \chi_2 & \cdots & 0 & 0 & 0 & \kappa_2 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \chi_l & 0 & 0 & 0 & \cdots & \kappa_l & 0 \\ 0 & 0 & \cdots & 0 & \alpha_{l+1} & 0 & 0 & \cdots & 0 & \beta_{l+1} \\ 0 & 0 & \cdots & 0 & \chi_{l+1} & 0 & 0 & \cdots & 0 & \kappa_{l+1} \end{pmatrix},$$

e, novamente, permutando colunas, temos que a matriz resultante é

$$R^{''} = \begin{pmatrix} \alpha_1 & 0 & \cdots & 0 & \beta_1 & 0 & \cdots & 0 & 0 & 0 \\ 0 & \alpha_2 & \cdots & 0 & 0 & \beta_2 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \alpha_l & 0 & 0 & \cdots & \beta_l & 0 & 0 \\ \chi_1 & 0 & \cdots & 0 & \kappa_1 & 0 & \cdots & 0 & 0 & 0 \\ 0 & \chi_2 & \cdots & 0 & 0 & \kappa_2 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \chi_l & 0 & 0 & \cdots & \kappa_l & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & \alpha_{l+1} & \beta_{l+1} \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & \chi_{l+1} & \kappa_{l+1} \end{pmatrix}.$$

Como o número total de permutações é par, o determinante da matriz resultante é igual ao determinante da matriz original. Portanto,

$$\det(R) = \prod_{j=1}^{l} (\alpha_j \kappa_j - \beta_j \chi_j) (\alpha_{l+1}\kappa_{l+1} - \beta_{l+1}\chi_{l+1}) = \prod_{j=1}^{l+1} (\alpha_j \kappa_j - \beta_j \chi_j).$$

□

Podemos agora determinar os autovalores da matriz H.

Lema 3.4 *Os autovalores da matriz H, definida em (3.31), são dados por*

$$\left\{ \frac{\gamma(1 - \mu_{ij}\theta) \pm \sqrt{\delta_{ij}(\theta)}}{2} \ , \ i = 1, 2, \ j = 1, \ldots, q \right\}. \tag{3.35}$$

Demonstração. Realizando algumas permutações nas linhas e colunas da matriz $tI - H$, podemos observar que o polinômio característico da matriz H pode ser construído como

$$
\begin{aligned}
p(t) = & \det \begin{pmatrix} tI-\gamma G_1 & 0 & (\gamma-1)I & 0 \\ 0 & tI-\gamma G_2 & 0 & (\gamma-1)I \\ -G_1 & 0 & tI & 0 \\ 0 & -G_2 & 0 & tI \end{pmatrix} \\
= & (-1)^{dN} \det \begin{pmatrix} tI-\gamma G_1 & 0 & (\gamma-1)I & 0 \\ -G_1 & 0 & tI & 0 \\ 0 & tI-\gamma G_2 & 0 & (\gamma-I)\,I \\ 0 & -G_2 & 0 & tI \end{pmatrix} \\
= & (-1)^{dN}(-1)^{dN} \det \begin{pmatrix} tI-\gamma G_1 & (\gamma-1)I & 0 & 0 \\ -G_1 & tI & 0 & 0 \\ 0 & 0 & tI-\gamma G_2 & (\gamma-1)I \\ 0 & 0 & -G_2 & tI \end{pmatrix} \\
= & \det(H_1)\det(H_2),
\end{aligned}
\tag{3.36}
$$

em que $H_1 = \begin{pmatrix} tI-\gamma G_1 & (\gamma-1)I \\ -G_1 & tI \end{pmatrix}$ e $H_2 = \begin{pmatrix} tI-\gamma G_2 & (\gamma-1)I \\ -G_2 & tI \end{pmatrix}$.

Veremos agora como efetuar o cálculo dos determinantes das matrizes H_1 e H_2. Inicialmente, considerando a matriz H_1 e substituindo G_1, conforme definido em (3.21), obtemos que

$$
H_1 = \begin{pmatrix} \Big(t-\gamma\,(1-\theta)\Big)I + \dfrac{\gamma\theta}{n\lambda}AA^T & (\gamma-1)\,I \\ (\theta-1)\,I + \dfrac{\theta}{n\lambda}AA^T & tI \end{pmatrix}.
$$

Definindo as variáveis auxiliares reais

$$
a = t-\gamma(1-\theta), \quad b = \frac{\gamma\theta}{n\lambda}, \quad c = \gamma-1, \quad r = \theta-1, \quad s = \frac{\theta}{n\lambda}, \tag{3.37}
$$

e usando a decomposição em valores singulares de A, dada em (3.22), podemos escrever que

$$
H_1 = \begin{pmatrix} U & 0 \\ 0 & U \end{pmatrix} \begin{pmatrix} aI+b\Sigma\Sigma^T & cI \\ rI+s\Sigma\Sigma^T & tI \end{pmatrix} \begin{pmatrix} U^T & 0 \\ 0 & U^T \end{pmatrix}.
$$

Pelo fato de $U \in \mathbb{R}^{d\times d}$ ser ortogonal e usando propriedades de determinantes, obtemos

$$\begin{aligned}
\det(H_1) = &\ \det\begin{pmatrix} U & 0 \\ 0 & U \end{pmatrix} \det\begin{pmatrix} aI + b\Sigma\Sigma^T & cI \\ rI + s\Sigma\Sigma^T & tI \end{pmatrix} \det\begin{pmatrix} U^T & 0 \\ 0 & U^T \end{pmatrix} \\
= &\ \det\begin{pmatrix} U & 0 \\ 0 & U \end{pmatrix} \det\begin{pmatrix} U^T & 0 \\ 0 & U^T \end{pmatrix} \det\begin{pmatrix} aI + b\Sigma\Sigma^T & cI \\ rI + s\Sigma\Sigma^T & tI \end{pmatrix} \\
= &\ \det\left(\begin{pmatrix} U & 0 \\ 0 & U \end{pmatrix}\begin{pmatrix} U^T & 0 \\ 0 & U^T \end{pmatrix}\right) \det\begin{pmatrix} aI + b\Sigma\Sigma^T & cI \\ rI + s\Sigma\Sigma^T & tI \end{pmatrix} \\
= &\ \det\begin{pmatrix} aI + b\Sigma\Sigma^T & cI \\ rI + s\Sigma\Sigma^T & tI \end{pmatrix}.
\end{aligned} \tag{3.38}$$

Além disso, usando (3.23), temos que

$$\det(H_1) = \det\begin{pmatrix} aI + b\Sigma\Sigma^T & cI \\ rI + s\Sigma\Sigma^T & tI \end{pmatrix} = \det\begin{pmatrix} aI + b\tilde{\Sigma}^2 & 0 & cI & 0 \\ 0 & aI & 0 & cI \\ rI + s\tilde{\Sigma}^2 & 0 & tI & 0 \\ 0 & rI & 0 & tI \end{pmatrix}. \tag{3.39}$$

Permutando, novamente, linhas e colunas podemos escrever que

$$\det(H_1) = \det\begin{pmatrix} aI + b\tilde{\Sigma}^2 & 0 & cI & 0 \\ 0 & aI & 0 & cI \\ rI + s\tilde{\Sigma}^2 & 0 & tI & 0 \\ 0 & rI & 0 & tI \end{pmatrix} = \det\begin{pmatrix} aI + b\tilde{\Sigma}^2 & cI & 0 & 0 \\ rI + s\tilde{\Sigma}^2 & tI & 0 & 0 \\ 0 & 0 & aI & cI \\ 0 & 0 & rI & tI \end{pmatrix}. \tag{3.40}$$

Note que a matriz resultante é diagonal em blocos. Considerando a Proposição 3.3, o cálculo do determinante da matriz H_1 resulta em

$$\det(H_1) = (at - cr)^{d-p} \prod_{j=1}^{p} \left(at + b\sigma_j^2 t - cr - cs\sigma_j^2\right)$$

e, substituindo as variáveis definidas em (3.37), temos que

$$\begin{aligned}
\det(H_1) = &\ \left(t^2 - \gamma(1-\theta)t - (\theta-1)(\gamma-1)\right)^{d-p} \times \\
&\times \prod_{j=1}^{p} \left[t^2 - \gamma\left(1 - \theta - \frac{\theta\sigma_j^2}{n\lambda}\right)t - (1-\gamma)\left(1 - \theta - \frac{\theta\sigma_j^2}{n\lambda}\right)\right].
\end{aligned}$$

Analogamente, podemos realizar o cálculo do determinante da matriz H_2, re-

sultando em

$$\begin{aligned}\det(H_2) = & \left(t^2 - \gamma\left(1 - \frac{\theta}{n\lambda}\right)t - \left(\frac{\theta}{n\lambda} - 1\right)(\gamma - 1)\right)^{N-p} \times \\ & \times \prod_{j=1}^{p}\left[t^2 - \gamma\left(1 - \frac{\theta}{n\lambda} - \frac{\theta\sigma_j^2}{(n\lambda)^2}\right)t - (1-\gamma)\left(1 - \frac{\theta}{n\lambda} - \frac{\theta\sigma_j^2}{(n\lambda)^2}\right)\right].\end{aligned}$$

Logo, usando (3.36), o polinômio característico da matriz H é dado por

$$\begin{aligned}p(t) = & \left(t^2 - \gamma(1-\theta)t - (\theta-1)(\gamma-1)\right)^{d-p} \times \\ & \times \left(t^2 - \gamma\left(1 - \frac{\theta}{n\lambda}\right)t - \left(\frac{\theta}{n\lambda} - 1\right)(\gamma - 1)\right)^{N-p} \times \\ & \times \prod_{j=1}^{p}\left[t^2 - \gamma\left(1 - \theta - \frac{\theta\sigma_j^2}{n\lambda}\right)t - (1-\gamma)\left(1 - \theta - \frac{\theta\sigma_j^2}{n\lambda}\right)\right] \times \\ & \times \prod_{j=1}^{p}\left[t^2 - \gamma\left(1 - \frac{\theta}{n\lambda} - \frac{\theta\sigma_j^2}{(n\lambda)^2}\right)t - (1-\gamma)\left(1 - \frac{\theta}{n\lambda} - \frac{\theta\sigma_j^2}{(n\lambda)^2}\right)\right],\end{aligned}$$

e o resultado segue pelas definições (3.32) e (3.33). □

A partir de agora, denotaremos os autovalores dados em (3.35) por

$$\lambda_{ij}^{-}(\theta) = \frac{\gamma(1-\mu_{ij}\theta) - \sqrt{\delta_{ij}(\theta)}}{2} \quad \text{e} \quad \lambda_{ij}^{+}(\theta) = \frac{\gamma(1-\mu_{ij}\theta) + \sqrt{\delta_{ij}(\theta)}}{2} \tag{3.41}$$

para $i = 1, 2,\ j = 1, \ldots, q$ e assumiremos que $\gamma \in (1, 2)$.

A fim de identificar possíveis padrões de comportamento, realizamos a análise gráfica dos autovalores da matriz H como funções de θ no plano complexo, conforme ilustra a Figura 3.5. Note que, os autovalores são números reais ou complexos, de acordo com o sinal da função δ_{ij}. Além disso, podemos observar que uma parte da trajetória é descrita por um segmento de reta, seguida por uma circunferência que passa pela origem do plano complexo e, em seguida, por um segmento de reta ou por uma semi-reta, dependendo da direção. Ainda, para certos valores de θ, os autovalores tendem a $-\infty$ por um lado, e pelo outro, se aproximam de um certo valor real no intervalo $(0, 1)$.

Nosso objetivo, então, é sistematizar esse comportamento, a fim de auxiliar a

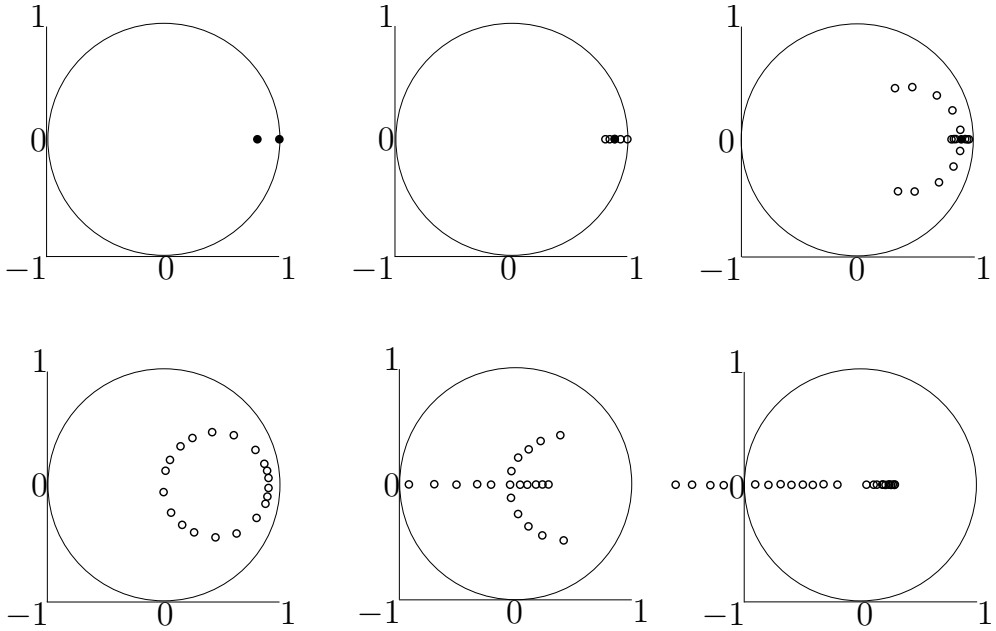


Figura 3.5: Representação do comportamento dos autovalores como funções de θ

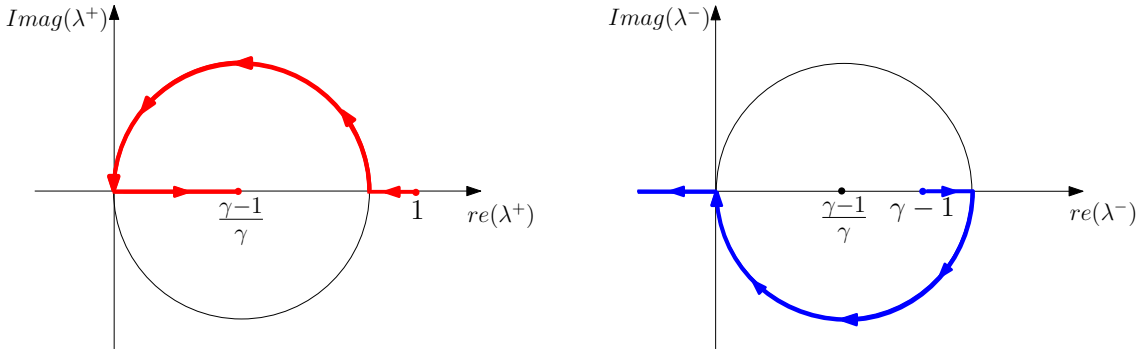
análise de convergência do método acc-SRP. Usando o fato de que

$$\begin{aligned}
\lim_{\theta\to\infty} \lambda_{ij}^{+}(\theta) &= \lim_{\theta\to\infty} \frac{\sqrt{\gamma^2(\mu_{ij}\theta-1)^2+4(\gamma-1)(\mu_{ij}\theta-1)}-\gamma(\mu_{ij}\theta-1)}{2} \\
&= \lim_{\theta\to\infty} \frac{2(\gamma-1)(\mu_{ij}\theta-1)}{\sqrt{\gamma^2(\mu_{ij}\theta-1)^2+4(\gamma-1)(\mu_{ij}\theta-1)}+\gamma(\mu_{ij}\theta-1)} \\
&= \lim_{\theta\to\infty} \frac{2(\gamma-1)}{\sqrt{\gamma^2+4\dfrac{(\gamma-1)}{(\mu_{ij}\theta-1)}}+\gamma} \\
&= \frac{\gamma-1}{\gamma},
\end{aligned} \tag{3.42}$$

podemos dizer que os autovalores $\lambda_{ij}^{+}(\theta)$ começam a trajetória no ponto $(1,0)$ e tendem para $\left(\frac{\gamma-1}{\gamma},0\right)$. Mais precisamente, podemos observar que, para cada par (i,j) fixado, para θ variando de 0 até $+\infty$, os autovalores $\lambda_{ij}^{+}(\theta)$ percorrem o caminho descrito pela Figura 3.6(a).

Por outro lado, os autovalores $\lambda_{ij}^{-}(\theta)$ iniciam a trajetória no ponto $(\gamma-1,0)$ indo em direção a $-\infty$, conforme destacado na Figura 3.6(b). Quando os autovalores são complexos, eles percorrem a circunferência de raio $\frac{\gamma-1}{\gamma}$ centrada no ponto $\left(\frac{\gamma-1}{\gamma},0\right)$. Além disso, observamos que para i e θ fixados, os autovalores obedecem a uma fila liderada por $\lambda_{i1}^{-}(\theta)$, seguido por $\lambda_{i2}^{-}(\theta)$, e assim sucessivamente. O mesmo ocorre com $\lambda_{ij}^{+}(\theta)$.

A partir de agora, buscamos formalizar tais afirmações por meio de resultados teóricos. Para isso, observe que os zeros das funções quadráticas δ_{ij}, definidas em



(a) λ_{ij}^+, $i = 1, 2$, $j = 1, \ldots, q$ (b) λ_{ij}^-, $i = 1, 2$, $j = 1, \ldots, q$

Figura 3.6: A trajetória descrita pelos autovalores

(3.33), são dados por

$$\theta_{ij}^- = \left(\frac{\gamma-2}{\gamma}\right)^2 \frac{1}{\mu_{ij}} \quad \text{e} \quad \theta_{ij}^+ = \frac{1}{\mu_{ij}}, \tag{3.43}$$

com $\theta_{ij}^- < \theta_{ij}^+$. Note que os autovalores, $\lambda_{ij}^-(\theta)$ ou $\lambda_{ij}^+(\theta)$, $i = 1, 2$, $j = 1, \ldots, q$, iniciam o caminho sendo números reais, em seguida tornam-se complexos e novamente, tornam-se reais. O resultado a seguir estabelece os valores de θ para os quais ocorrem tais variações numéricas.

Lema 3.5 *Considere os autovalores $\lambda_{ij}^-(\theta)$ e $\lambda_{ij}^+(\theta)$, definidos em (3.41), para i e j fixados. Então, para $\gamma \in (1, 2)$*

(i) Se $\theta \in [0, \theta_{ij}^-]$ os autovalores são números reais e

$$\gamma - 1 \leq \lambda_{ij}^-(\theta) \leq 2\left(\frac{\gamma-1}{\gamma}\right) \leq \lambda_{ij}^+(\theta) \leq 1;$$

(ii) Se $\theta \in (\theta_{ij}^-, \theta_{ij}^+)$ os autovalores são números complexos e

$$\left|\lambda_{ij}^-(\theta) - \frac{\gamma-1}{\gamma}\right| = \left|\lambda_{ij}^+(\theta) - \frac{\gamma-1}{\gamma}\right| = \frac{\gamma-1}{\gamma};$$

(iii) Se $\theta \geq \theta_{ij}^+$ os autovalores voltam a ser números reais e

$$\lambda_{ij}^-(\theta) \leq -\lambda_{ij}^+(\theta) \leq 0 \leq \lambda_{ij}^+(\theta) \leq \frac{\gamma-1}{\gamma}.$$

Demonstração. (i) Inicialmente, note que $\delta_{ij}(\theta) \geq 0$ para todo $\theta \in [0, \theta_{ij}^-]$. Além disso, $\theta \leq \left(\frac{\gamma-2}{\gamma}\right)^2 \frac{1}{\mu_{ij}}$ implica que

$$1 - \mu_{ij}\theta \geq 1 - \left(\frac{\gamma-2}{\gamma}\right)^2 = \frac{4(\gamma-1)}{\gamma^2} > \frac{2(\gamma-1)}{\gamma},$$

pois $\gamma \in (1,2)$. Portanto, $\gamma(1-\mu_{ij}\theta) - 2(\gamma-1) \geq 0$. Usando a definição de γ e $\delta_{ij}(\theta)$, podemos escrever que

$$\begin{aligned} 4(\gamma-1)^2 &\geq 0 \\ \delta_{ij}(\theta) + 4(\gamma-1)^2 &\geq \delta_{ij}(\theta) \\ \Big(\gamma(1-\mu_{ij}\theta) - 2(\gamma-1)\Big)^2 &\geq \delta_{ij}(\theta). \end{aligned}$$

Como $\delta_{ij}(\theta) \geq 0$, concluímos que

$$\lambda_{ij}^{-}(\theta) - (\gamma-1) = \frac{1}{2}\left(\gamma(1-\mu_{ij}\theta) - 2\,(\gamma-1) - \sqrt{\delta_{ij}(\theta)}\right) \geq 0.$$

Para a segunda desigualdade, note que

$$\lambda_{ij}^{-}(\theta) - 2\left(\frac{\gamma-1}{\gamma}\right) = \frac{1}{2\gamma}\left(\gamma^2(1-\mu_{ij}(\theta) + 4\,(1-\gamma) - \sqrt{\gamma^2\delta_{ij}(\theta)}\right).$$

Além disso,

$$\gamma^2\delta_{ij}(\theta) = \gamma^2(1-\mu_{ij}\theta)\Big(\gamma^2(1-\mu_{ij}\theta) + 4(1-\gamma)\Big) \geq 0$$

implica que $\gamma^2(1-\mu_{ij}\theta) + 4(1-\gamma) \geq 0$, pois $(1-\mu_{ij}\theta) \geq 0$. E, dessa forma, podemos escrever que

$$\begin{aligned} 4(1-\gamma) &\leq 0 \\ \gamma^2(1-\mu_{ij}\theta) + 4(1-\gamma) &\leq \gamma^2(1-\mu_{ij}\theta). \end{aligned}$$

Usando o fato de que $\Big(\gamma^2(1-\mu_{ij}\theta) + 4(1-\gamma)\Big)^2 - \gamma^2\delta_{ij}(\theta) \leq 0$, obtemos então que $\lambda_{ij}^{-}(\theta) \leq 2\left(\frac{\gamma-1}{\gamma}\right)$. Analogamente, podemos verificar que $\lambda_{ij}^{+}(\theta) - 2\left(\frac{\gamma-1}{\gamma}\right) \geq 0$.

Para concluir este caso, observe que

$$\lambda_{ij}^{+}(\theta) - 1 = \frac{1}{2}\left(\gamma(1-\mu_{ij}\theta) - 2 + \sqrt{\delta_{ij}(\theta)}\right),$$

e que $\gamma(1-\mu_{ij}\theta) \leq \gamma < 2$. Uma vez que $\Big(2-\gamma(1-\mu_{ij}\theta)\Big)^2 \geq \delta_{ij}(\theta)$, obtemos $\lambda_{ij}^{+}(\theta) - 1 \leq 0$.

(ii) Neste caso, temos $\delta_{ij}(\theta) < 0$. Assim, os números complexos $\lambda_{ij}^{-}(\theta)$ e $\lambda_{ij}^{+}(\theta)$ cumprem a seguinte igualdade

$$\left|\lambda_{ij}^{-}(\theta) - c\right|^2 = \left|\lambda_{ij}^{+}(\theta) - c\right|^2 = \left(\frac{\gamma(1-\mu_{ij}(\theta))}{2} - c\right)^2 - \frac{\delta_{ij}(\theta)}{4} = c^2,$$

em que $c = \dfrac{\gamma-1}{\gamma}$.

(iii) Note que, neste caso, temos $(1-\mu_{ij}\theta) \leq 0$. Logo, podemos escrever que

$$\lambda_{ij}^{-}(\theta) + \lambda_{ij}^{+}(\theta) = \frac{\gamma(1-\mu_{ij}\theta)}{2} + \frac{\gamma(1-\mu_{ij}\theta)}{2} \leq 0,$$

o que resulta em $\lambda_{ij}^{-}(\theta) \leq -\lambda_{ij}^{+}(\theta)$.

A segunda e terceira desigualdades seguem do fato que $\gamma(1-\mu_{ij}\theta) \leq 0$ e $\delta_{ij}(\theta) \geq \Big(\gamma(1-\mu_{ij}\theta)\Big)^2$. A fim de provar a desigualdade restante, note que

$$\lambda_{ij}^{+}(\theta) - \frac{\gamma-1}{\gamma} = \frac{1}{2\gamma}\left(\gamma^2(1-\mu_{ij}\theta) - 2\left(\gamma-1\right) + \sqrt{\gamma^2\delta_{ij}(\theta)}\right).$$

Usando o fato de que $\gamma \in (1,2)$ podemos escrever que

$$\begin{aligned}
4(1-\gamma)^2 &\geq 0 \\
4\gamma^2(1-\mu_{ij}\theta)(1-\gamma) + 4(1-\gamma)^2 &\geq 4\gamma^2(1-\mu_{ij}\theta)(1-\gamma) \\
\left(\gamma^2(1-\mu_{ij}\theta) + 2(1-\gamma)\right)^2 &\geq \gamma^2(1-\mu_{ij}\theta)\left(\gamma^2(1-\mu_{ij}\theta) + 4(1-\gamma)\right) \\
\left(-\gamma^2(1-\mu_{ij}\theta) - 2(1-\gamma)\right)^2 &\geq \gamma^2(1-\mu_{ij}\theta)\left(\gamma^2(1-\mu_{ij}\theta) + 4(1-\gamma)\right).
\end{aligned}$$

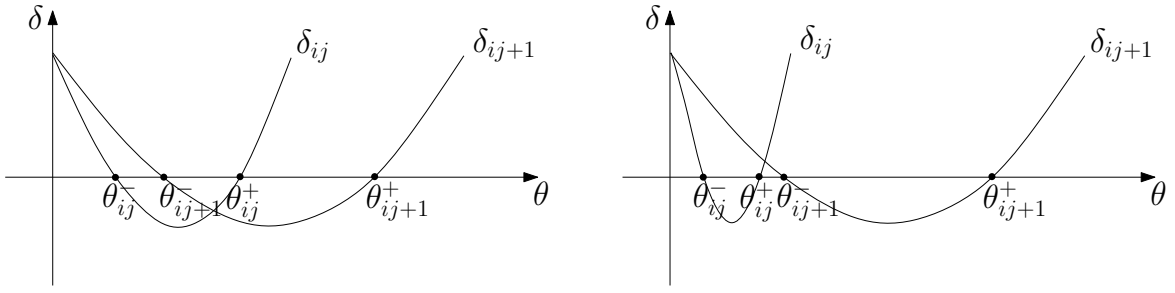
Uma vez que $\gamma^2\delta_{ij}(\theta) \leq \Big(-\gamma^2(1-\mu_{ij}(\theta)) + 2(\gamma-1)\Big)^2$ e $-\gamma^2(1-\mu_{ij}\theta) + 2(\gamma-1) \geq 0$, pois $(1-\mu_{ij}\theta) \leq 0$ para todo $\theta \geq \theta_{ij}^{+}$ e $\gamma \geq 1$, concluímos que $\lambda_{ij}^{+}(\theta) - \dfrac{\gamma-1}{\gamma} \leq 0$, completando a prova. □

Definida a trajetória descrita pelos autovalores, vamos estabelecer que os mesmos obedecem a uma fila para i e θ fixados. Para isso, observe na Figura 3.6 que para os dois casos, (a) e (b), o caminho percorrido pelos autovalores é dividido essencialmente em três partes, de acordo com os valores de θ. Na primeira e terceira parte, os autovalores estão sobre o eixo de números reais do plano complexo, e na segunda porção do caminho eles estão sobre uma semicircunferência com centro em $\left(\frac{\gamma-1}{\gamma}, 0\right)$ e raio $\left(\frac{\gamma-1}{\gamma}\right)$, seguindo a orientação das flexas. Esta orientação é fundamental para estabelecer o resultado a seguir.

Lema 3.6 *Considere o caminho dos autovalores descrito como função de θ, representado na Figura 3.6(a), começando no ponto $(1,0)$ e terminando em $\left(\frac{\gamma-1}{\gamma}, 0\right)$. O autovalor $\lambda_{ij}^{+}(\theta)$ inicia o percurso, sendo seguido por $\lambda_{ij+1}^{+}(\theta)$. Por outro lado, para o caminho representado na Figura 3.6(b), que é inicializado no ponto $(\gamma-1, 0)$, com direção a $-\infty$ sob o eixo real, temos que o autovalor $\lambda_{ij}^{-}(\theta)$ inicia o trajeto, seguido por $\lambda_{ij+1}^{-}(\theta)$.*

Demonstração. Note que os zeros das funções δ_{ij} e $\delta_{ij+1}, j = 1, \ldots, p$, dados em (3.43), cumprem $\theta_{ij}^{-} \leq \theta_{ij+1}^{-} \leq \theta_{ij}^{+} \leq \theta_{ij+1}^{+}$ ou $\theta_{ij}^{-} \leq \theta_{ij}^{+} \leq \theta_{ij+1}^{-} \leq \theta_{ij+1}^{+}$, conforme representado

na Figura 3.7. A análise destes dois casos é similar, por isso, detalhamos apenas o primeiro caso.



(a) $\theta_{ij}^- \leq \theta_{ij+1}^- \leq \theta_{ij}^+ \leq \theta_{ij+1}^+$ (b) $\theta_{ij}^- \leq \theta_{ij}^+ \leq \theta_{ij+1}^- \leq \theta_{ij+1}^+$

Figura 3.7: Representação da ordenação para as possibilidades de zeros das funções δ_{ij} e δ_{ij+1}

Inicialmente, vamos mostrar que o autovalor $\lambda_{ij}^+(\theta)$ precede $\lambda_{ij+1}^+(\theta)$. Para isso, também faremos a análise para cinco partições do domínio da função δ_{ij}, dadas a seguir.

(i) Considere $\theta \in \left(0, \theta_{ij}^-\right]$ com i, j fixados, onde $i = 1, 2, j = 1, \ldots, p$. Usando a definição da função δ_{ij} e o fato de que $\theta_{ij}^- \leq \theta_{ij+1}^-$, podemos concluir que $0 \leq \delta_{ij}(\theta) \leq \delta_{ij+1}(\theta)$, o que implica que $\lambda_{ij}^+(\theta) \leq \lambda_{ij+1}^+(\theta)$, pois $(1 - \mu_{ij}) < (1 - \mu_{ij+1})$.

(ii) Para $\theta \in \left(\theta_{ij}^-, \theta_{ij+1}^-\right]$, temos que $\delta_{ij}(\theta) < 0$ e $\delta_{ij+1}(\theta) \geq 0$. Assim, $\lambda_{ij}^+(\theta)$ já pertence à segunda porção do caminho, ou seja, à semicircunferência, enquanto $\lambda_{ij+1}^+(\theta)$ ainda permanece na primeira porção, como ilustramos na Figura 3.6.

(iii) Neste caso, temos que $\theta \in \left(\theta_{ij+1}^-, \theta_{ij}^+\right)$. Usando a definição de δ_{ij}, temos que $\delta_{ij}(\theta) < 0$ e $\delta_{ij+1}(\theta) < 0$. Assim, pelo Lema 3.5, temos que

$$re\left(\lambda_{ij}^+(\theta)\right) = \gamma\left(1 - \mu_{ij}\theta\right) \leq \gamma\left(1 - \mu_{ij+1}\theta\right) = re\left(\lambda_{ij+1}^+(\theta)\right),$$

o que implica que $\lambda_{ij}^+(\theta)$ precede $\lambda_{ij+1}^+(\theta)$.

(iv) Para $\theta \in \left[\theta_{ij}^+, \theta_{ij+1}^+\right)$, temos $\delta_{ij}(\theta) \geq 0$ e $\delta_{ij+1}(\theta) < 0$. Portanto, $\lambda_{ij}^+(\theta)$ já está na terceira parte do caminho, enquanto $\lambda_{ij+1}^+(\theta)$ ainda permanece na segunda parte, como observado da Figura 3.6.

(v) Considerando $\theta \geq \theta_{ij+1}^+$ sabemos que $\delta_{ij}(\theta) \geq 0$. Para provar que os autovalores cumprem $\lambda_{ij}^+(\theta) \geq \lambda_{ij+1}^+(\theta)$ basta relembrar que

$$\lambda_{ij}^+(\theta) = \frac{\gamma(1 - \mu_{ij}\theta) + \sqrt{\gamma^2\left(1 - \mu_{ij}\theta\right)^2 + 4(1-\gamma)(1 - \mu_{ij}\theta)}}{2},$$

$\mu_{ij} \geq \mu_{ij+1}$ e que a função $\phi : [1, +\infty) \to \mathbb{R}$ definida por

$$\phi(s) = \frac{\gamma(1-s) + \sqrt{\gamma^2 (1-s)^2 + 4(1-\gamma)(1-s)}}{2}$$

é crescente. Para comprovar, considere sua derivada

$$\phi'(s) = \frac{-\gamma\sqrt{\gamma^2 (1-s)^2 + 4(1-\gamma)(1-s)} - \gamma^2(1-s) - 2(1-\gamma)}{2\sqrt{\gamma^2 (1-s)^2 + 4(1-\gamma)(1-s)}}. \tag{3.44}$$

e o fato de que $\gamma^2 (1-s)^2 + 2(1-\gamma) \leq 0$. Isso nos leva a escrever

$$\begin{aligned} 4(\gamma-1)^2 &\geq 0 \\ 4\gamma^2(1-s)(1-\gamma) + 4(\gamma-1)^2 + \gamma^4(1-s)^2 &\geq 4\gamma^2(1-s)(1-\gamma) + \gamma^4(1-s)^2 \\ (\gamma^2(1-s) + 2(1-\gamma))^2 &\geq \gamma^2\left(\gamma^2(1-s)^2 + 4(1-\gamma)(1-s)\right) \\ -\gamma^2(1-s) - 2(1-\gamma) &\geq \gamma\sqrt{\gamma^2(1-s)^2 + 4(1-\gamma)(1-s)}, \end{aligned}$$

e nos permite concluir que $\phi'(s) \geq 0$.

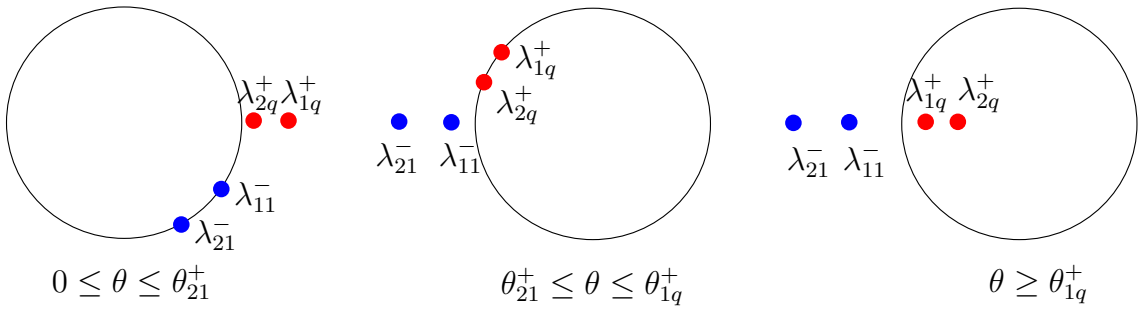
Analogamente, podemos provar que os autovalores $\lambda^-_{ij}(\theta)$ lideram a fila, seguidos por $\lambda^-_{ij+1}(\theta)$, o que conclui a prova. □

Podemos concluir do Lema 3.6 que para um i fixado, o autovalor que lidera a fila está sempre associado ao autovalor que possui o maior valor para o parâmetro μ_{ij}, e assim, sucessivamente. Esse resultado pode ser estendido para j e θ fixados, mas variando i. Pela definição de μ_{1j} e μ_{2j}, temos que se $n\lambda < 1$, então $\mu_{2j} > \mu_{1j}$, $\theta^-_{2j} < \theta^-_{1j}$ e $\theta^+_{2j} < \theta^+_{1j}$. Assim, $\lambda^+_{2j}(\theta)$ lidera a fila sendo seguido por $\lambda^+_{1j}(\theta)$, e $\lambda^-_{2j}(\theta)$ lidera a fila, sendo seguido por $\lambda^-_{1j}(\theta)$. Para o caso $n\lambda \geq 1$, temos $\mu_{2j} \leq \mu_{1j}$, $\theta^-_{1j} < \theta^-_{2j}$ e $\theta^+_{1j} < \theta^+_{2j}$. Logo, $\lambda^+_{1j}(\theta)$ precede $\lambda^+_{2j}(\theta)$, e $\lambda^-_{1j}(\theta)$ precede $\lambda^-_{2j}(\theta)$.

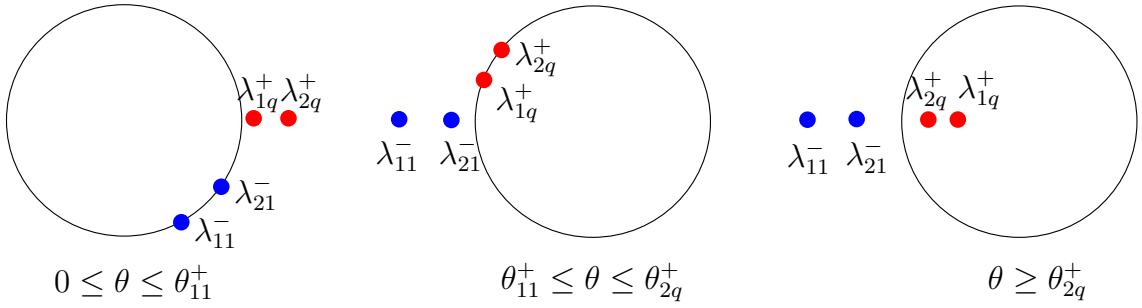
Essa observação sugere uma representação gráfica apenas entre os autovalores $\lambda^-_{i1}(\theta)$ e $\lambda^+_{iq}(\theta), i = 1, 2$, que são responsáveis pelo raio espectral, já que denotam os autovalores que lideram ou findam a fila, respectivamente. Na Figura 3.8 podemos ver a representação desses autovalores, especificamente, ao longo da trajetória.

Note que, na segunda parte do caminho, os autovalores $\lambda^+_{11}(\theta)$, $\lambda^+_{12}(\theta)$ e $\lambda^-_{11}(\theta)$, $\lambda^-_{12}(\theta)$, são simétricos, respectivamente, em relação ao eixo real, assim como $\lambda^+_{21}(\theta)$, $\lambda^+_{22}(\theta)$ e $\lambda^-_{21}(\theta)$, $\lambda^-_{22}(\theta)$. Portanto, possuem a mesma distância até a origem. Por esse motivo, apenas por simplificação, representamos no gráfico apenas os autovalores $\lambda^-_{11}(\theta)$ e $\lambda^-_{21}(\theta)$.

Agora, baseados nos Lemas 3.5 e 3.6, podemos estabelecer o raio espectral da matriz $H_\gamma(\theta)$, dada em (3.29), por meio da função $\rho : [0, +\infty) \to \mathbb{R}$ dada, para o caso $n\lambda < 1$, por



(a) $n\lambda < 1$



(b) $n\lambda \geq 1$

Figura 3.8: Representação dos autovalores que lideram e findam a fila

$$\rho(\theta) = \begin{cases} \left|\lambda_{1q}^+(\theta)\right|, \text{ se } 0 \leq \theta \leq \theta_{21}^+ \\ \max\left\{\left|\lambda_{1q}^+(\theta)\right|, -\lambda_{21}^-(\theta)\right\}, \text{ se } \theta_{21}^+ \leq \theta \leq \theta_{1q}^+ \\ -\lambda_{21}^-(\theta), \text{ se } \theta \geq \theta_{1q}^+ \end{cases} \tag{3.45}$$

e, para o caso $n\lambda \geq 1$ por

$$\rho(\theta) = \begin{cases} \left|\lambda_{2q}^+(\theta)\right|, \text{ se } 0 \leq \theta \leq \theta_{11}^+ \\ \max\left\{\left|\lambda_{2q}^+(\theta)\right|, -\lambda_{11}^-(\theta)\right\}, \text{ se } \theta_{11}^+ \leq \theta \leq \theta_{2q}^+ \\ -\lambda_{11}^-(\theta), \text{ se } \theta \geq \theta_{2q}^+. \end{cases} \tag{3.46}$$

Embora a função ρ esteja definida para dois casos, separadamente, seu comportamento gráfico, Figura 3.9, é mantido para os casos $n\lambda < 1$ ou $n\lambda \geq 1$.

Uma vez determinada a função ρ, podemos estabelecer o resultado que trata da convergência do Algoritmo 3.2.

Teorema 3.7 *Seja $v^0 \in \mathbb{R}^{2(d+N)}$ um ponto inicial arbitrário e considere a sequência $(v^k) = \begin{pmatrix} x^k \\ z^k \end{pmatrix}$ gerada pelo Algoritmo 3.2 com $\theta \in \left(0, \dfrac{2\gamma}{2\gamma - 1}\right) \dfrac{1}{\mu_{i1}}$, onde $i = 1$ se $n\lambda \geq 1$ ou $i = 2$ se $n\lambda < 1$. Então, a sequência (x^k) converge linearmente para*

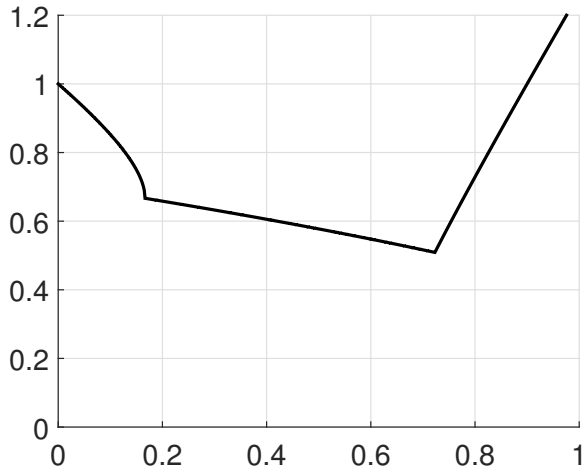


Figura 3.9: Gráfico da função ρ

a solução única do problema (3.4) com uma taxa de convergência assintótica $\rho(\theta)$, segundo definição (3.45) e (3.46), e o valor ótimo para o parâmetro θ é dado pela solução da equação

$$\left|\lambda_{1q}^{+}(\theta)\right| = -\lambda_{21}^{-}(\theta), \quad se \quad n\lambda < 1 \tag{3.47}$$

ou

$$\left|\lambda_{2q}^{+}(\theta)\right| = -\lambda_{11}^{-}(\theta), \quad se \quad n\lambda \geq 1. \tag{3.48}$$

Demonstração. Inicialmente, note que dado $\epsilon > 0$, existe uma norma matricial $\|\cdot\|$ tal que $\|H_\gamma(\theta)\| \leq \rho(\theta) + \epsilon$ (Ver Lema 5.6.10 em [37]). Assim, usando (3.30) obtemos

$$\left\|v^k - v^*\right\| \leq \|H_\gamma(\theta)\| \left\|v^{k-1} - v^*\right\| \leq (\rho(\theta) + \epsilon) \left\|v^{k-1} - v^*\right\|.$$

Portanto, a convergência assintótica segue de $\rho(\theta) < 1$ o que equivale a

$$\theta < \left(\frac{2\gamma}{2\gamma - 1}\right) \frac{1}{\mu_{i1}}.$$

De fato, de acordo com o Lema 3.6 o primeiro autovalor a deixar a região compreendida pelo círculo complexo será $\lambda_{11}^{-}(\theta)$ ou $\lambda_{21}^{-}(\theta)$. Assim, para verificar a primeira vez em que $\rho(\theta) = 1$, para $\theta \neq 0$, basta calcularmos $-\lambda_{i1}^{-}(\theta) = 1$, cuja solução é

$$\tilde{\theta}_i = \left(\frac{2\gamma}{2\gamma - 1}\right) \frac{1}{\mu_{i1}},$$

pois $\delta_{i1}(\theta) \geq 0$ para todo $\theta \geq \dfrac{1}{\mu_{i1}}$. Note que o valor mínimo de $\tilde{\theta}_i$ está associado com $\max\limits_{i=1,2}\{\mu_{i1}\}$. Usando a definição de μ_{ij}, dada em (3.32), temos que

$$\max_{i=1,2} \mu_{i1} = \begin{cases} \mu_{11}, & \text{se} \quad n\lambda \geq 1 \\ \mu_{21}, & \text{se} \quad n\lambda < 1 \end{cases}.$$

Finalmente, a solução ótima θ^* para o problema

$$\min_{\theta} \rho(\theta)$$

pode ser obtida diretamente da análise da função ρ, dada por (3.45) ou (3.46), como resultado das igualdades

$$\left|\lambda_{1q}^{+}(\theta)\right| = -\lambda_{21}^{-}(\theta)$$

ou

$$\left|\lambda_{2q}^{+}(\theta)\right| = -\lambda_{11}^{-}(\theta),$$

para os casos em que $n\lambda < 1$ e $n\lambda \geq 1$, respectivamente. □

Concluímos assim o estudo referente à convergência linear do método acc-SRP. No próximo capítulo, apresentaremos os detalhes da implementação dos algoritmos, assim como os resultados numéricos obtidos da aplicação dos algoritmos aqui propostos.

Capítulo 4

Resultados Numéricos

Apresentamos neste capítulo alguns resultados de testes numéricos realizados considerando um conjunto de problemas obtidos a partir da geração de matrizes $A \in \mathbb{R}^{d\times N}$ em Matlab, com variações nas dimensões e também no parâmetro λ, a fim de analisar o desempenho dos algoritmos propostos. Iniciamos com uma breve discussão envolvendo alguns detalhes da implementação, que foi feita em Matlab 8.4.0.150421 (R2014b) para Windows.

4.1 Algoritmo SRP

O algoritmo SRP, Algoritmo 3.1, consiste em, a partir de um ponto inicial, gerar uma sequência $\left(x^k\right)$ que converge para um ponto estacionário, solução do problema (3.4). O limite da sequência é um ponto estacionário para o problema (3.4), o qual, dada natureza do problema, é, de fato, seu minimizador global. Destacamos a seguir, o critério utilizado para escolha do parâmetro θ.

4.1.1 Escolha do parâmetro θ

Note que o parâmetro θ, Algoritmo 3.1, possui apenas a restrição $\theta > 0$. Sendo assim, para efeitos numéricos resolvemos utilizar duas variantes para a escolha deste parâmetro, a fim de padronizar os valores. Adotamos o valor constante $\dfrac{1}{L}$, onde L representa o maior autovalor da Hessiana de f e o valor ótimo θ^*. O primeiro valor é obtido por meio da função objetivo, enquanto o valor ótimo θ^* é o minimizador do raio espectral, destacado nas Tabelas 3.1 e 3.2.

4.2 Algoritmo acc-SRP

O algoritmo acc-SRP, Algoritmo 3.2, objetiva gerar duas sequências $\left(x^k\right)$ e $\left(z^k\right)$, ambas convergindo para o mesmo ponto, solução do problema (3.4). No entanto,

a diferença em relação ao método SRP consiste em ser uma reformulação acelerada. Mesmo atuando em uma dimensão maior, esse algoritmo tem a característica de ter o mesmo custo numérico por iteração da versão não acelerada. A seguir, destacamos os critérios para escolha dos parâmetros γ e θ, empregados no Algoritmo 3.2.

4.2.1 Escolha do parâmetro γ

O algoritmo acc-SRP foi estabelecido para os valores de $\gamma > 0$. No entanto, para a análise de convergência teórica do método foi necessário fixar $\gamma \in (1, 2)$. Logo, a convergência do algoritmo é garantida para qualquer valor neste intervalo, além disso há uma considerável melhoria no desempenho do método à medida que γ cresce. Neste sentido, para verificar a eficiência numérica do algoritmo acc-SRP consideramos, nos testes, valores de γ próximos a 2.

4.2.2 Escolha do parâmetro θ

Adotamos as mesmas escolhas para o parâmetro θ. Note que o valor constante $\frac{1}{L}$, onde L representa o maior autovalor da Hessiana de f segue da definição de f e valor ótimo θ^* é aquele que satisfaz as igualdades

$$\left|\lambda_{1q}^{+}(\theta)\right| = -\lambda_{21}^{-}(\theta),$$

se $n\lambda < 1$, ou

$$\left|\lambda_{2q}^{+}(\theta)\right| = -\lambda_{11}^{-}(\theta)$$

para o caso $n\lambda \geq 1$, conforme definição do raio espectral, dado em (3.45) e (3.46), respectivamente.

No entanto, tais igualdades não podem ser resolvidas explicitamente. Devido a esse fato, resolvemos aplicar um método numérico intervalar para encontrar a raiz das equações $\left|\lambda_{1q}^{+}(\theta)\right| + \lambda_{21}^{-}(\theta) = 0$ e $\left|\lambda_{2q}^{+}(\theta)\right| + \lambda_{11}^{-}(\theta) = 0$. Escolhemos o método da bissecção aplicado aos intervalos $\left[\theta_{21}^{+}, \theta_{1q}^{+}\right]$ ou $\left[\theta_{11}^{+}, \theta_{2q}^{+}\right]$, respectivamente, usando como critério de parada a precisão de 10^{-4}.

4.3 Resultados Numéricos

Analisamos a eficiência e robustez dos métodos SRP e acc-SRP com duas variações para cada algoritmo, que se referem à escolha do parâmetro θ. A fim de comparar os algoritmos, consideramos o número de iterações e o tempo computacional como medidas de desempenho. Os testes foram executados em um processador Intel(R) Core(TM) i5-3337U CPU, 1.8 GHz e 8.0 GB de memória RAM.

4.3.1 Análise dos resultados

Os algoritmos SRP e acc-SRP foram aplicados com duas variações para a escolha do parâmetro θ, conforme discutido anteriormente. Com o intuito de facilitar a comparação entre o desempenho dos algoritmos em relação ao número de iterações e o tempo computacional, optamos por usar os gráficos de desempenho propostos em [13], os quais constituem uma ferramenta para avaliação e comparação de um conjunto S de algoritmos aplicados a P problemas.

Para introduzir o formato de comparação, considere como medida de desempenho o tempo computacional, por exemplo, e n_p e n_s o número de problemas e o número de algoritmos, respectivamente. Para cada problema p e cada algoritmo s, foi definido $t_{p,s}$ como o tempo computacional requerido pelo algoritmo s para resolver o problema p e utilizado como medida comparativa o índice de desempenho, definido como

$$r_{p,s} = \frac{t_{p,s}}{\min\{t_{p,s} : s \in \mathsf{S}\}}.$$

Considere um parâmetro $r_M \geq r_{p,s}$ para todo p, s escolhidos. Assim, definimos $r_{p,s} = r_M$ quando o algoritmo s não resolve o problema p. A fim de obter uma visão global para o desempenho de cada algoritmo aplicado a um conjunto de problemas, foi definido

$$\rho_s(\tau) = \frac{1}{n_p} card\{p \in \mathsf{P} : r_{p,s} \leq \tau\},$$

como a probabilidade do índice $r_{p,s}$ estar dentro de um fator τ do melhor índice possível. Assim, $\rho_s(1)$ é a proporção de problemas que o algoritmo s resolve no menor tempo.

De forma geral, considerando uma medida de desempenho arbitrária, $\rho_s(\tau)$ representa a porcentagem de problemas que o algoritmo s resolve em τ vezes o valor da medida de desempenho do melhor algoritmo. Nos gráficos de desempenho, os valores do fator τ são indicados no eixo das abscissas, enquanto que no eixo das ordenadas são representados os valores das respectivas probabilidades $\rho_s(\tau)$.

A fim de verificar o comportamento dos algoritmos propostos para resolver o problema (3.4), fixamos os valores $n \in \{150,\ 200, 250, 300, 350\}$, $d \in \{10,\ 20, 30, 40, 50\}$ e $m = 1$. Os elementos da matriz A foram gerados aleatoriamente, de forma que a matriz A^TA seja singular, e conforme discutido na Seção 1.3, o método de regularização seja aplicado. Consideramos $\gamma = 1.95$ e como critérios de parada $f(x^k) \leq 10^{-6}$, já que o valor ótimo do problema (3.4) é $f(x^*) = 0$ e f é sempre não-negativa, e $k = 30000$ (número máximo de iterações).

Para o parâmetro de regularização usamos $\lambda \in (0, 1)$, pois para $\lambda \geq 1$ o desempenho numérico foi mantido. A análise numérica foi realizada, especificamente, para duas instâncias. A primeira consideramos $0.1 \leq \lambda < 1$ e a segunda $0 < \lambda \leq 0.1$, e os resultados encontram-se discutidos a seguir.

Considere a primeira instância com 100 problemas resultantes da variação

dos parâmetros n, d e m nos conjuntos já definidos e $\lambda \in \{0.6;\ 0.4;\ 0.2;\ 0.1\}$. Para tais problemas, observamos que utilizando como medida de desempenho o número de iterações, as duas variantes aplicadas aos algoritmos SRP e acc-SRP mostraram-se igualmente robustas, no sentido de que obtiveram sucesso na resolução dos 100 problemas. Os algoritmos acc-SRP com $\theta = \theta^*$ e acc-SRP com $\theta = \dfrac{1}{L}$ resolveram todos os problemas com o menor número de iterações, enquanto os algoritmos SRP com $\theta = \theta^*$ e SRP com $\theta = \dfrac{1}{L}$ resolveram 47% e 1%, respectivamente, conforme pode ser observado na Figura 4.1.

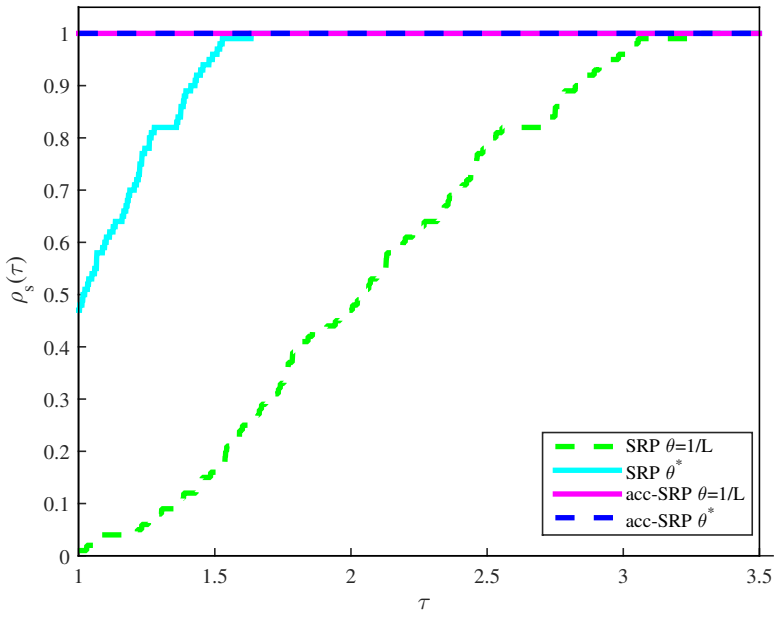


Figura 4.1: Gráfico de desempenho para o número de iterações para a instância 1

O desempenho dos algoritmos em relação ao tempo computacional pode ser observado na Figura 4.2, e para uma melhor visualização destacamos também o gráfico em uma maior escala considerando $0 \leq \tau \leq 20.5$. O algoritmo acc-SRP com $\theta = \theta^*$ resolveu 86% dos problemas com o menor tempo computacional e o algoritmo acc-SRP com $\theta = \dfrac{1}{L}$, 14%. Para resolver todos os problemas, os algoritmos SRP com $\theta = \theta^*$ e SRP com $\theta = \dfrac{1}{L}$ exigiram no máximo 65 e 198 vezes, respectivamente, o tempo exigido pelo melhor algoritmo. Sendo assim, temos evidências numéricas que comprovam a eficiência da versão acelerada.

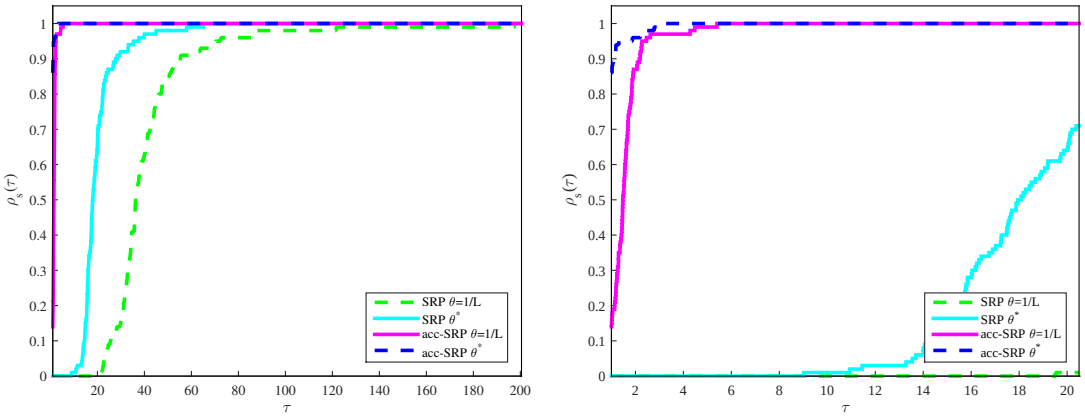


Figura 4.2: Gráfico de desempenho para o tempo computacional para a instância 1

Agora, realizamos os testes numéricos para a segunda instância de problemas, considerando $\lambda \in \{10^{-1};\ 10^{-2};\ 10^{-3};\ 10^{-4}\}$. Os algoritmos acelerados se mostraram mais robustos, resolvendo pelo menos 92% dos problemas contra a porcentagem máxima de 64% resolvidos pelos métodos não acelerados. Além disso, verificamos que o algoritmo acc-SRP com $\theta = \theta^*$ é o mais eficiente, resolvendo 87% dos problemas com o menor número de iterações, conforme pode ser visto na Figura 4.3.

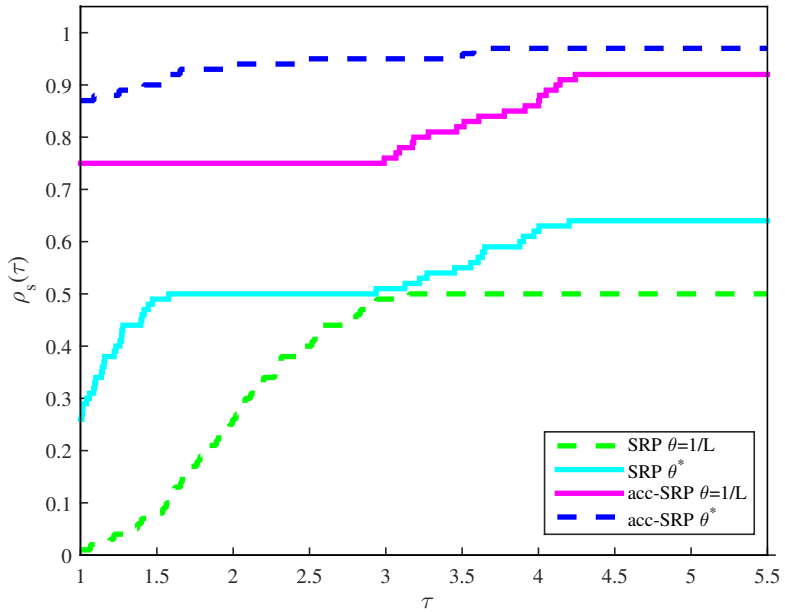


Figura 4.3: Gráfico de desempenho para o número de iterações para a instância 2

O algoritmo acc-SRP com $\theta = \dfrac{1}{L}$ resolveu 75% dos problemas com o menor número de iterações, enquanto os algoritmos SRP com $\theta = \theta^*$ e SRP com $\theta = \dfrac{1}{L}$, resolveram 26% e 1%, respectivamente.

Já na Figura 4.4 apresentamos o gráfico de desempenho para os quatro algoritmos referente ao tempo computacional, assim como um destaque em maior escala para

$0 \leq \tau \leq 20.5$. Note a eficiência dos algoritmos acelerados, uma vez que o algoritmo acc-SRP com $\theta = \theta^*$ resolveu 79% dos problemas com o menor tempo computacional, enquanto que para o algoritmo acc-SRP com $\theta = \dfrac{1}{L}$, a porcentagem foi de 21%. Para resolver 50% dos problemas o tempo computacional exigido pelo algoritmo SRP com $\theta = \theta^*$ é no máximo 22 vezes, aproximadamente, maior que o tempo requerido pelo melhor algoritmo.

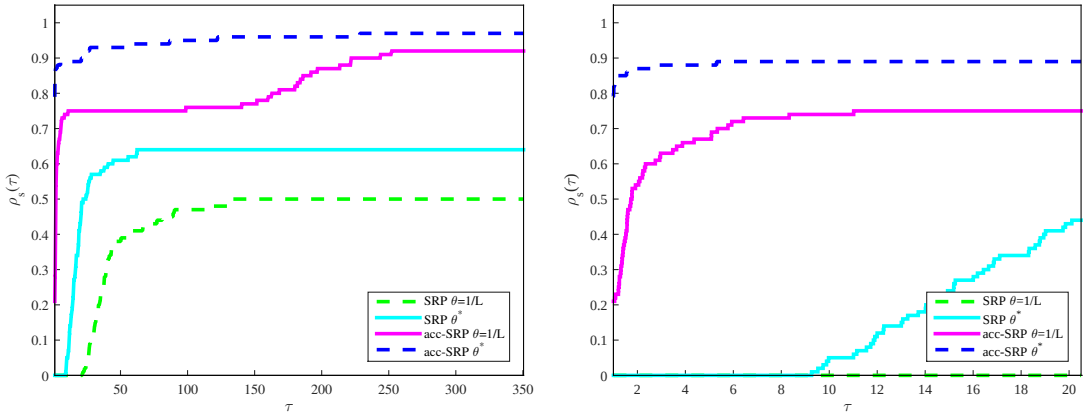


Figura 4.4: Gráfico de desempenho para o tempo computacional para a instância 2

Tendo em vista tais resultados, comprovamos a eficiência em relação ao número de iterações e tempo computacional dos algoritmos acelerados, principalmente do método acc-SRP com $\theta = \theta^*$, justificando a referência de reformulação acelerada do algoritmo SRP. Observamos que tais algoritmos foram mais robustos, resolvendo até 100% dos problemas quando foi considerada a primeira instância, ou seja, valores maiores de λ.

Após constatarmos a eficiência e robustez dos algoritmos propostos, por meio dos resultados numéricos, nosso objetivo foi realizar a comparação com um método clássico. Tal discussão encontra-se estruturada a seguir.

4.3.2 Comparação com Gradientes Conjugados

Para realizar uma análise comparativa do desempenho numérico dos algoritmos propostos utilizamos o Método de Gradientes Conjugados (GC), considerando os mesmos critérios de parada mencionados anteriormente ($f(x^k) \leq 10^{-6}$ e $k = 30000$). A escolha deste método deu-se por ser muito eficiente para minimização de funções quadráticas, de fácil implementação e matematicamente idêntico ao método LSQR. Como verificamos que o Método acc-SRP com $\theta = \theta^*$ mostrou-se numericamente superior em relação aos demais algoritmos propostos, realizamos a comparação numérica entre tal algoritmo e o Método de Gradientes Conjugados, aplicados ao problema (3.4), utilizando o número de iterações e o tempo computacional como medidas de desempenho.

Os testes foram realizados utilizando os mesmos parâmetros discutidos na Seção 4.3.1. No entanto, como era de se esperar, para problemas em que a Hessiana da função f é pouco mal-condicionada, o Método dos Gradientes Conjugados mostrou desempenho superior, uma vez que se trata de uma função quadrática. No entanto, à medida que o número de condição dessa matriz aumenta, o desempenho desse método cai, chegando a gastar mais que $d+N$ iterações, enquanto que o Método acc-SRP apresenta um comportamento superior. Sendo assim, a fim de ilustrar essa característica do Método acc-SRP, realizamos testes considerando matrizes muito mal-condicionadas, com número de condição da ordem de 10^{10}.

Comparamos inicialmente os algoritmos para os parâmetros n, m, d, γ fixados e $\lambda \in \{0.6; 0.4; 0.2; 0.1\}$, resultando em uma nova instância. Observamos que o algoritmo acc-SRP resolveu 100% dos problemas propostos, mostrando-se superior ao Método de Gradientes Conjugados, que resolveu 98% dos problemas. Em relação ao número de iterações, Figura 4.5, o algoritmo acc-SRP resolveu 97% dos problemas com o menor número de iterações, enquanto que a porcentagem atingida pelo Método de Gradientes Conjugados foi de 3%.

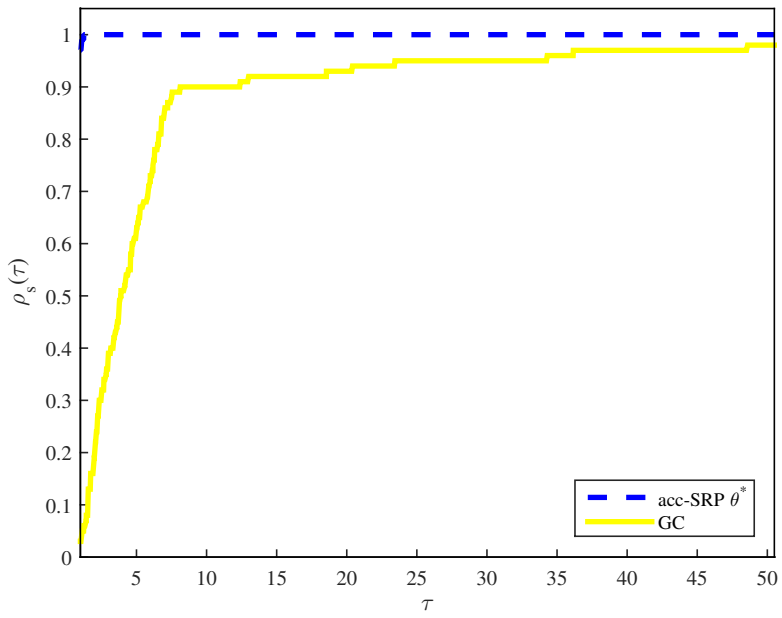


Figura 4.5: Gráfico de desempenho para o número de iterações para a instância 3

Na Figura 4.6 apresentamos o gráfico de desempenho para os algoritmos considerando como medida de desempenho o tempo computacional. Note que o algoritmo acc-SRP com $\theta = \theta^*$ resolveu 54% dos problemas com o menor tempo computacional, enquanto o Método de Gradientes Conjugados resolveu 46% dos problemas.

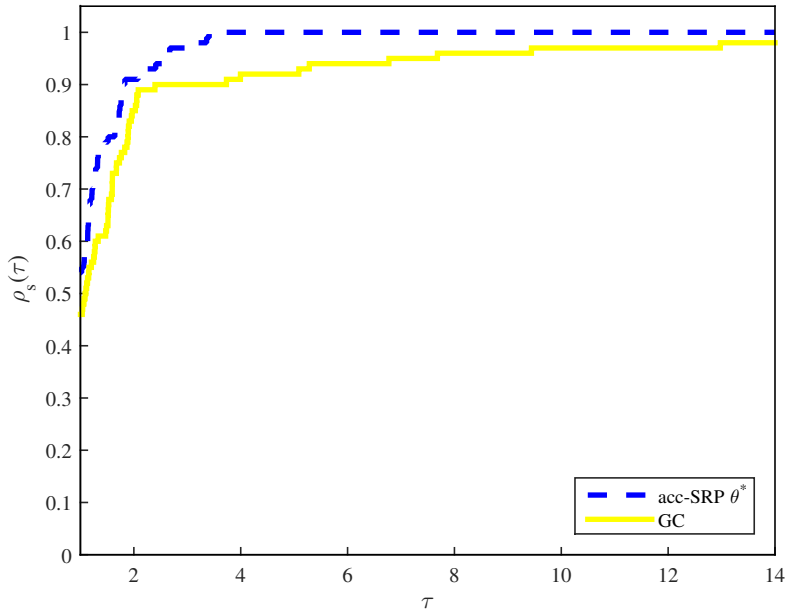


Figura 4.6: Gráfico de desempenho para o tempo computacional para a instância 3

Consideramos agora $\lambda \in \{10^{-1};\ 10^{-2};\ 10^{-3};\ 10^{-4}\}$ e os demais parâmetros anteriormente mencionados, originando a quarta instância de problemas. Para tal instância, o algoritmo acc-SRP com $\theta = \theta^*$ resolveu todos os problemas propostos, enquanto o Método de Gradientes Conjugados, 71%. O gráfico de desempenho para os algoritmos em relação ao número de iterações encontra-se na Figura 4.7. Note que o algoritmo acc-SRP com $\theta = \theta^*$ resolveu 95% dos problemas usando o menor número de iterações, enquanto o Método de Gradientes Conjugados a porcentagem foi de 5%.

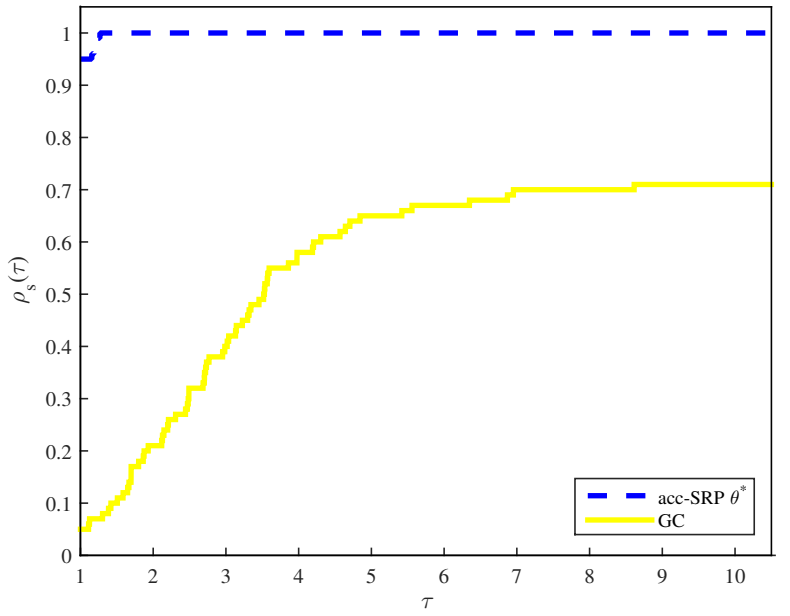


Figura 4.7: Gráfico de desempenho para o número de iterações para a instância 4

Na Figura 4.8 apresentamos um comparativo entre os algoritmos em relação ao tempo computacional, com um destaque para $0 \leq \tau \leq 2.5$. Note que o algoritmo acc-SRP com $\theta = \theta^*$ também apresentou um melhor desempenho, resolvendo 52% dos problemas com o menor tempo computacional, contra 48% resolvido pelo Método de Gradientes Conjugados.

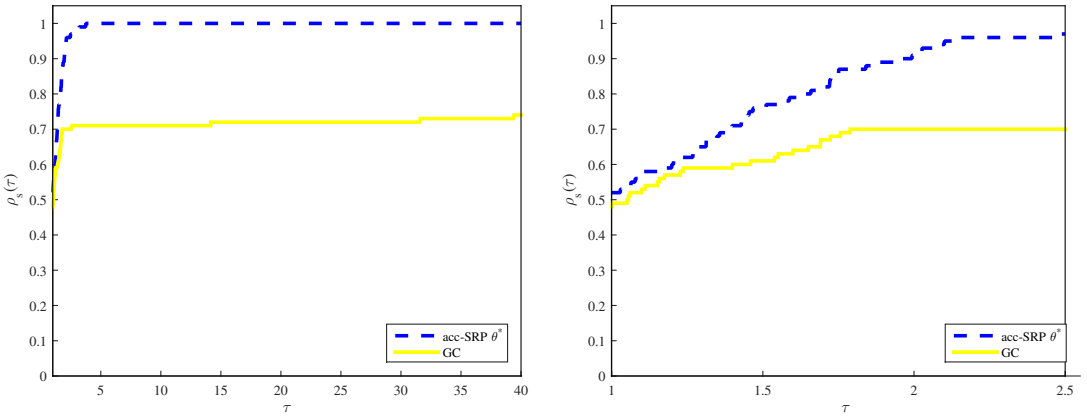
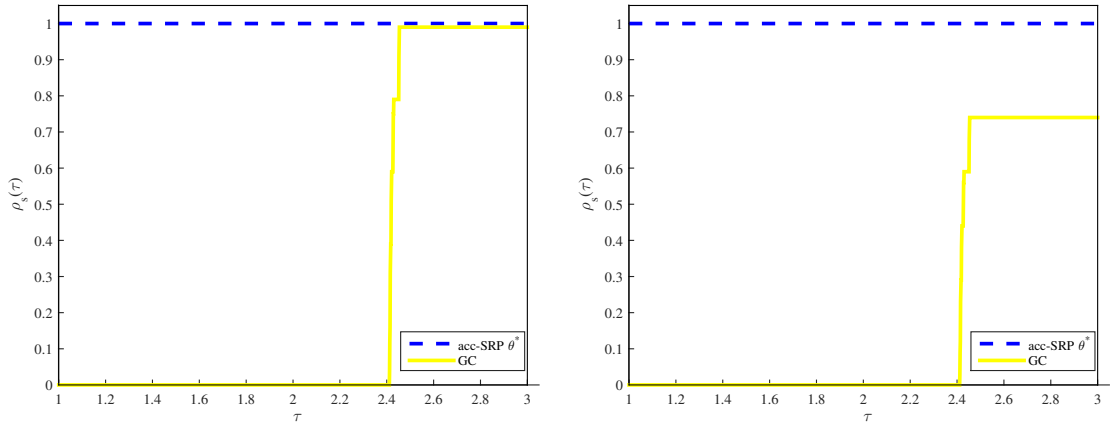


Figura 4.8: Gráfico de desempenho para o tempo computacional para a instância 4

Fazendo uma análise numérica comparativa do algoritmo acc-SRP e do Método de Gradientes Conjugados, para as instâncias 3 e 4, observamos que o custo computacional considerando produtos matriciais, por iteração, é menor para a metodologia proposta nesta tese, conforme pode ser observado na Figura 4.9. Isso significa que as iterações do algoritmo acc-SRP são mais baratas que as iterações do Método de Gradientes Conjugados sob esse critério. No entanto, os demais cálculos que caracterizam cada iteração do algoritmo acc-SRP reduzem a diferença entre o desempenho destes algoritmos em relação ao tempo computacional. Dessa forma, para matrizes muito mal-condicionadas verificamos que o algoritmo acc-SRP se mostrou superior, numericamente, ao Método de Gradientes Conjugados.



(a) Instância 3 (b) Instância 4

Figura 4.9: Gráfico de desempenho para o custo computacional por iteração

4.3.3 Conclusões dos resultados numéricos

Observamos que os algoritmos acc-SRP, com $\theta = \frac{1}{L}$ e $\theta = \theta^*$, destacaram-se em relação às medidas de desempenho consideradas. Verificamos, em especial, que o algoritmo acc-SRP, com $\theta = \theta^*$, apresentou melhor desempenho geral nos testes quando considerados o tempo computacional e o número de avaliações de função.

Comparando os algoritmos propostos com o Método de Gradientes Conjugados, aplicados ao problema (3.4), verificamos que o algoritmo acelerado, acc-SRP com $\theta = \theta^*$, mostrou-se numericamente superior à metodologia clássica quando consideradas instâncias de problemas muito mal-condicionados. Os testes apresentados foram realizados para $0 < \lambda < 1$. No entanto, para valores superiores para o parâmetro λ o Método de Gradientes Conjugados se mostrou melhor em relação ao algoritmo acc-SRP com $\theta = \theta^*$ quando utilizado o tempo computacional como medida de comparação de desempenho. Por outro lado, quando analisado o número de iterações, o algoritmo acc-SRP com $\theta = \theta^*$ ainda mostrou-se superior, numericamente. Esse comportamento pode ser explicado pelo fato de que à medida que o parâmetro λ aumenta, o condicionamento da Hessiana da função objetivo do problema (3.4) melhora, favorecendo assim o desempenho do Método de Gradientes Conjugados.

Dessa forma, acreditamos que a nossa metodologia é promissora no campo de regularização pelo fato de fazer uma abordagem teórica relevante, utilizando principalmente informações das variáveis primais e duais e estabelecer teoricamente a convergência linear, o que diferencia a nossa abordagem dos métodos específicos para resolução do problema regularizado. Ressaltamos que essa abordagem é indicada para ser aplicada a problemas de regularização cujo valor do parâmetro λ esteja definido no intervalo $(0, 1)$, o que na prática é o que fornecem os métodos de escolha de tal parâmetro, conforme verificado em [35].

Conclusão

Neste trabalho, discutimos e estabelecemos novos algoritmos de ponto fixo aplicados ao problema primal-dual de *Ridge Regression*. A característica comum aos algoritmos propostos é que as formulações primal e dual foram tratadas simultaneamente e a análise de convergência foi estabelecida usando propriedades espectrais das matrizes de iteração.

Propomos o algoritmo SRP e uma versão acelerada, acc-SRP, para tratar o problema. Os resultados numéricos mostraram que o desempenho da reformulação acelerada, Método acc-SRP, foi superior tanto em relação ao número de avaliações de função quanto ao tempo computacional, quando comparado à versão não-acelerada, Método SRP. Além disso, outra característica da reformulação acelerada é que embora aplicada a uma versão aumentada, o custo computacional por iteração é equivalente ao método SRP, pelo fato de estabelecer uma sequência com convergência mais rápida.

Fazendo um levantamento dos métodos específicos aplicados ao problema penalizado, em especial os métodos de projeção, verificamos que um inconveniente é a característica da semi-convergência. Embora, numericamente, para tais métodos a convergência seja rápida quanto ao número de iterações, a mesma está condicionada a uma escolha acertível do critério de parada, a fim de minimizar a interferência do resíduo na qualidade da solução. Nesse sentido, nossa estratégia foi, a partir do problema primal de *Ridge Regression* (3.2), abordar a formulação dual (3.3) e tratar o problema na versão primal-dual. Com isso, estabelecendo as relações de dualidade, a partir do conhecimento da solução ótima do problema reformulado, contornamos a dificuldade de escolha do critério de parada e, por consequência, da qualidade de convergência associada aos métodos de projeção.

Além disso, verificamos que numericamente o Método de Gradientes Conjugados perde em eficiência e robustez quando aplicado a problemas muito mal-condicionados. Nesse sentido, constatamos que o algoritmo acc-SRP com $\theta = \theta^*$ mostrou-se numericamente superior quando aplicado a tais problemas e com um custo computacional mais baixo, por iteração.

Neste contexto, com base no excelente desempenho do Método acc-SRP aplicado ao problema (3.4), deixamos como sugestão para trabalhos futuros o desevolvimento de métodos para a resolução de sistemas mal-condicionados gerais, a fim de generalizar a metodologia estabelecida nesta tese.

Referências Bibliográficas

[1] B. Anders. Ridge regression and inverse problems. Relatório técnico, 2001. Stock. Univ., Sweden.

[2] E. C. Baptista, E. A. Belati e G. R. M. Costa. Um método primal-dual aplicado na resolução do problema de fluxo de potência ótimo. *Pesquisa Operacional*, 91(2):215–226, 2004.

[3] F. S. V. Bazán e L. S. Borges. Métodos para problemas inversos de grande porte. In *Notas em Matemática Aplicada*, volume 39 of *Títulos publicados para o XXXII CNMAC*. Sociedade Brasileira de Matemática Aplicada e Computacional, São Paulo, 2009.

[4] F.S.V. Bazán e J.B. Francisco. Improved fixed-point algorithm for determining the Tikhonov regularization parameter. *Inverse Problems*, 25(4), 2009.

[5] D. P. Bertsekas, A. Nedic e A.E. Ozdaglar. *Convex Analysis and Optimization*. Athena Scientific, United States, 2003.

[6] D. Calvetti, B. Lewis e L. Reichel. On the regularizing properties of the gmres method. *Numerische Mathematik*, 91(4):605–625, 2002.

[7] A. Chambolle e T. Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *J Math Imaging Vis*, 40:120–145, 2011.

[8] D. Colton e R. Kress. *Integral Equation Methods for Scattering Theory*. John Wiley, New York, 1983.

[9] G. M. Cordeiro e G. A. Paula. *Modelos de Regressão para análise de dados univariados*. Instituto de Matemática Pura e Aplicada, Rio de Janeiro, 1989.

[10] I. J. D. Craig e J. C. Brown. *Inverse Problems in Astronomy*. Adam Hilger, Bristol, 1986.

[11] J. J. M. Cuppen. *A Numerical Solution of the Inverse Problem of Electrocardiography*. Tese, University of Amsterdam, 1983.

[12] G. B. Dantzig, L. R. Ford e D. R. Fulkerson. A primal-dual algorithm for linear programs. *Princeton University Press, Princeton*, pág. 171–181, 1956.

[13] E. D. Dolan e J. J. Moré. Benchmarking optimization software with performance profiles. *Mathematical Programming*, 91:201–213, 2002.

[14] M. El-Dereny e N. I. Rashwan. Solving multicollinearity problem using Ridge Regression Models. *Int. J. Contemp. Math. Sciences*, 6:585–600, 2011.

[15] L. M. Elias. Uma base teórica para conjugação de funções semicontínuas inferiormente. Dissertação, UFPR, Curitiba, 2013.

[16] W. Fenchel. On conjugate convex functions. *Canad. J. Math*, 1:73–77, 1949.

[17] A. Fuhry e L. Reichel. A new Tikhonov regularization method. *Numer. Algor.*, 59:433–445, 2012.

[18] S. Gazzola e J. G. Nagy. Generalized Arnoldi-Tikhonov Method for sparse reconstruction. *SIAM J. Sci. Comput.*, 36(2):225–247, 2014.

[19] M. X. Goemans e D. P. Williamson. *The primal-dual method for approximation algorithms and its application to network design problems*. PWS Publishing Co. Boston, USA, 1996.

[20] G. H. Golub. Numerical methods for solving linear least squares problems. *Numerische Mathematik*, 7:206–216, 1965.

[21] G. H. Golub, M. T. Heath e G. Wahba. Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21:215–223, 1979.

[22] C. W. Groetsch. *Inverse Problems in the Mathematical Sciences*. Vieweg Verlag, Wiesbaden, 19893.

[23] A. Gupta, R. Krishnaswamy e K. Pruhs. Online primal-dual for non-linear optimization with applications to speed scaling. Em Springer-Verlag Berlin Heidelberg, editor, *Approximation and Online Algorithms*, volume 7846 of *Lecture Notes in Computer Science*, pág. 173–186. Springer Berlin Heidelberg, Slovenia, 2012.

[24] J. Hadamard. *Lectures on Cauchy's Problem in Linear Partial Differential Equations*. Yale University Press, New Haven, 1923.

[25] M. Hanke. On lanczos based methods for the regularization of discrete ill-posed problems. *BIT Numer. Math.*, 41:1008–1018, 2001.

[26] P. C. Hansen. Regularization, GSVD and Truncated GSVD. *BIT Numerical Mathematics*, 29(3):491–504, 1989.

[27] P. C. Hansen. Regularization tools: A matlab package for analysis and solution of discrete ill-posed problems. *Numerical Algorithms*, 6(1):1–35, 1994.

[28] P. C. Hansen. *Rank-deficient and discrete ill-posed problems*. SIAM, Philadelphia, PA, 1998.

[29] P. C. Hansen. The l-curve and its use in the numerical treatment of inverse problems. In *Computational Inverse Problems in Electrocardiology*, volume 39 of *Títulos publicados para o XXXII CNMAC*. P. Johnston - WIT Press, Southampton, 2001.

[30] P. C. Hansen. *Discrete Inverse Problems: Insight and Algorithms*. SIAM, Philadelphia, PA, 2010.

[31] T. Hastie, R. Tibshirani e J. Friedman. *The Elements of Statistical Learning Data Mining, Inference, and Prediction*. Springer Series in Statistics, 2001.

[32] D. M. Hawkins e X. Yinb. A faster algorithm for Ridge Regression of reduced rank data. *Computational Statistics & Data Analysis*, 40(2):253–262, 2002.

[33] J-B. Hiriart-Urruty e C. Lemaréchal. *Convex Analysis and Minimization Algorithms I*. Springer-Verlag, New York, 1993.

[34] M. E. Hochstenbach e L. Reichel. An iterative method for Tikhonov Regularization with a general linear regularization operator. *Journal of Integral Equations and Applications*, 22(3):465–482, 2010.

[35] A. E. Hoerl e R. W. Kennard. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

[36] B. Hofmann, B. Kaltenbacher, C. Pöschl e O. Scherzer. A convergence rates result for Tikhonov regularization in Banach spaces with non-smooth operators. *Inverse Problems*, 23:987Ű1010, 2007.

[37] R. A. Horn e C. R. Johnson. *Matrix Analysis*. Cambridge University Press, Australia, 1985.

[38] Y. Huang e Z. Jia. Some results on the regularization of lsqr for large-scale discrete ill-posed problems. Relatório técnico, 2016. arXiv:1503.01864v3 [math.NA].

[39] A. F Izmailov e M. V. Solodov. *Otimização, Volume1: Condições de otimalidade, Elementos de análise convexa e de Dualidade*. IMPA, Rio de Janeiro, 2005.

[40] N. Komodakis e J. C. Pesquet. Playing with duality: An overview of recent primal-dual approaches for solving large-scale optimization problems. Relatório técnico, 2014. IEEE Signal Processing Magazine.

[41] C. Lanczos. An iteration method for the solution of the eigenvalue problem of linear differential and integral operators. *J. Res. Nat. Bureau Standards*, 45:255–282, 1950.

[42] J. Liu, L. Lin, W. Zhang e G. Li. A novel combined regularization algorithm of total variation and Tikhonov regularization for open electrical impedance tomography. *Physiol. Meas.*, 34(7):823–838, 2013.

[43] D. W. Marquardt e R. D. Snee. Ridge regression in practice. *The American Statistician*, 29(1):3–20, 1975.

[44] M. Martin Fuhry e L. Reichel. A new tikhonov regularization method. *Numer. Algor.*, 59, 2012.

[45] V. A. Morozov. On the solution of functional equations by the method of regularization. *Soviet Math. Dokl.*, 7:414–417, 1966.

[46] F. Natterer. *The Mathematics of Computerized Tomography*. John Wiley, New York, 1986.

[47] C. C. Paige e M. A. Saunders. Lsqr: An algorithm for sparse linear equations and sparse least squares. *ACM Trans. Math. Softw*, 8(1):43–71, 1982.

[48] C. C. Paige e M. A. Saunders. Lsqr: Sparse linear equations and least squares problems. *ACM Trans. Math. Softw*, 8(2):195–209, 1982.

[49] M. R. Pinheiro. Conjugação e dualidade em programação convexa. Relatório técnico, 1984. Universidade Nova de Lisboa.

[50] Z. Qu, P. Richtárik e T. Zhang. Randomized dual coordinate ascent with arbitrary sampling. Relatório técnico, 2014. math.OC.

[51] C. Saunders, A Gammerman e V. Vovk. Ridge regression learning algorithm in dual variables. In *Proceedings of the 15th International Conference on Machine Learning*, pág. 515–521. Morgan Kaufmann Publishers Inc., 1998.

[52] A. R. Secchi. Otimização de processos. Relatório técnico, 2015. UFRJ - COPPE, Programa de Engenharia Química.

[53] S. Shalev-Shwartz e T. Zhang. Accelerated proximal stochastic dual coordinate ascent for regularized loss minimization. *Math. Program., Ser. A*, pág. 1–41, 2014.

[54] F. Tabak. *Robust Algorithms for Discrete Tomography*. Tese, Delft University of Technology - Faculty of Electrical Engineering, Mathematics and Computer Science Delft Institute of Applied Mathematics, 2012.

[55] J. N. Tehrani, A. McEwan, C. Jin e A. van Schaik. L1 regularization method in electrical impedance tomography by using the l1-curve (pareto frontier curve). *Applied Mathematical Modelling*, 36, 2012.

[56] R. Tibshirani. Regression shrinkage and selection via the lasso. *J. R. Statist. Soc. B*, 58:267–288, 1994.

[57] A. N. Tikhonov. On the solution of ill-posed problems and the method of regularization. *Dokl. Akad. Nauk*, 151:501–504, 1963.

[58] H. D. Vinod. A survey of Ridge Regression and related techniques for improvements over ordinary least squares. *The Review of Economics and Statistics*, 60(1):121–131, 1978.

[59] E. de Vito, V. Umanità e S. Villa. A consistent algorithm to solve Lasso, elastic-net and Tikhonov regularization. *Journal of Complexity*, 27:188–200, 2011.

[60] W. van Wieringen. Lecture notes on Ridge Regression. Relatório técnico, 2015. arXiv:1509.09169v1.

[61] M. Xiangrui e H. Chen. Accelerating Nesterov's method for strongly convex functions with lipschitz gradient. Relatório técnico, 2011. math.OC.

[62] T. Zhang. On the dual formulation of regularized linear systems with convex risks. *Machine Learning*, 46(1):91–129, 2002.

[63] Q. Zhu. Primal-dual combinatorial algorithms. Relatório técnico, 2009.