UNIVERSIDADE FEDERAL DE SANTA CATARINA
CENTRO DE CIÊNCIAS FÍSICAS E MATEMÁTICAS
PROGRAMA DE PÓS-GRADUAÇÃO EM MATEMÁTICA PURA E APLICADA

Everton Boos

Avanços em técnicas iterativas para problemas inversos lineares e não lineares com aplicação na reconstrução de condutividade térmica

Florianópolis
2022

Everton Boos

Avanços em técnicas iterativas para problemas inversos lineares e não lineares com aplicação na reconstrução de condutividade térmica

Tese submetida ao Programa de Pós-Graduação em Matemática Pura e Aplicada da Universidade Federal de Santa Catarina para a obtenção do título de doutor em Matemática Pura e Aplicada, com área de concentração em Matemática Aplicada.
Orientador: Prof. Fermín S. V. Bazán, Dr.

Florianópolis
2022

Ficha de identificação da obra elaborada pelo autor,
através do Programa de Geração Automática da Biblioteca Universitária da UFSC.

Boos, Everton
Avanços em técnicas iterativas para problemas inversos lineares e não lineares com aplicação na reconstrução de condutividade térmica / Everton Boos ; orientador, Fermín Sinforiano Viloche Bazán, 2022.
157 p.

Tese (doutorado) - Universidade Federal de Santa Catarina, Centro de Ciências Físicas e Matemáticas, Programa de Pós-Graduação em Matemática Pura e Aplicada, Florianópolis, 2022.

Inclui referências.

1. Matemática Pura e Aplicada. 2. Problemas inversos. 3. Estimativas de erro. 4. Método de Levenberg-Marquardt. 5. Condutividade térmica. I. Bazán, Fermín Sinforiano Viloche. II. Universidade Federal de Santa Catarina. Programa de Pós-Graduação em Matemática Pura e Aplicada. III. Título.

Everton Boos

Avanços em técnicas iterativas para problemas inversos lineares e não lineares com aplicação na reconstrução de condutividade térmica

O presente trabalho em nível de doutorado foi avaliado e aprovado por banca examinadora composta pelos seguintes membros:

Prof. Saulo Pomponet Oliveira, Dr.
Universidade Federal do Paraná (UFPR)

Prof. Hugo José Lara Urdaneta, Dr.
Universidade Federal de Santa Catarina (UFSC – Blumenau)

Prof. Juliano de Bem Francisco, Dr.
Universidade Federal de Santa Catarina (UFSC)

Prof. Fermín S. V. Bazán, Dr.
Universidade Federal de Santa Catarina (UFSC - Orientador)

Certificamos que esta é a **versão original e final** do trabalho de conclusão que foi julgado adequado para obtenção do título de doutor em Matemática Pura e Aplicada, com área de concentração em Matemática Aplicada.

Documento assinado digitalmente
Daniel Goncalves
Data: 30/06/2022 19:06:41-0300
CPF: 017.022.459-77
Verifique as assinaturas em https://v.ufsc.br

Prof. Daniel Gonçalves, Dr.
Coordenador do Programa de
Pós-Graduação

Documento assinado digitalmente
Fermin Sinforiano Viloche Bazan
Data: 30/06/2022 23:15:09-0300
CPF: 137.612.718-08
Verifique as assinaturas em https://v.ufsc.br

Prof. Fermín S. V. Bazán, Dr.
Orientador

Florianópolis, 2022.

Este trabalho é dedicado aos sonhos
delirantes de um adolescente incauto.

AGRADECIMENTOS

Aos meus pais, Ana Carla e Laercio, pelo apoio incondicional sem o qual eu nada teria sido ou atingido. Serei sempre grato pela presença e força, especialmente nas ocasiões que mais precisei.

Ao Prof. Fermín, pela orientação e paciência ao longo desses anos. Além do aprendizado, me ensinou que mesmo nas horas mais sombrias é preciso manter a cabeça erguida e seguir lutando.

Àqueles que, ao cumprirem o papel da docência, formaram quem eu sou hoje. Em especial, aos membros da banca examinadora, pela leitura, comentários e valiosas contribuições à versão final deste trabalho.

À parceria entre CAPES e FAPESC[1], pelo apoio financeiro durante todo o doutorado, sem o qual este trabalho não seria possível. Que possam continuar e expandir a sua contribuição para que mais jovens pesquisadores e pesquisadoras tenham meios de produzir avanços concretos à tecnologia, inovação e análise crítica nacionais.

À UFSC, pela oportunidade de crescimento profissional e intelectual proporcionada por esta instituição de excelência a nível nacional e internacional. Em particular aos servidores públicos e terceirizados interessados em buscar soluções com celeridade e eficiência.

Ao empenho, para quando a motivação acaba. Talento ou genialidade no dia a dia são ofuscadas pela perseverança e pela disciplina. Atingir objetivos significa reconhecer as próprias limitações e aprender a contorná-las, um passo de cada vez.

Por fim, preciso dizer: este foi difícil. O doutorado por si só afeta as pessoas ao nosso redor, demandando compreensão e apoio, e escrever uma tese, mais ainda, é um trabalho árduo, isolado, muitas vezes enlouquecedor. Nada disso se faz ou se supera sozinho. Agradeço a todos que mesmo sem se darem conta, possam ter mudado completamente os meus dias para melhor apenas com poucas palavras ou algum alento de humanidade à uma mente cansada. Saibam que fizeram e fazem muita diferença. Então, aos amigos trazidos pelo caminho que trilhei, um agradecimento aos que se fizeram presentes, trazendo luz a um quarto escuro.

[1] Bolsista com suporte integral CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) e FAPESC (Fundo de Amparo à Pesquisa e Inovação do Estado de Santa Catarina), vinculado ao Projeto n° 88887.178114/2018-00.

RESUMO

Propomos avanços a métodos iterativos para problemas inversos lineares e não lineares, visando expandir a teoria existente e aplicar em situações práticas. Obter soluções estáveis para os chamados problemas discretos mal postos através de técnicas iterativas demanda o uso de critérios de parada especializados. Neste sentido, para um método proposto recentemente por Bazán e Boos, se as iterações são finalizadas através do princípio da discrepância de Morozov, existe uma estimativa ao erro relativo entre a solução exata e a aproximação gerada pelo método. Assumindo uma condição de fonte do tipo Hölder, mostramos que a estimativa pode ser significantemente melhorada, em dois resultados distintos. Uma terceira estimativa é obtida, sem condições adicionais, da ordem do ruído nos dados. Todas fornecem estreitamento ao resultado original que motivou o estudo. No caso não linear, resolver problemas de mínimos quadrados através do método de Levenberg-Marquardt (LMM) é constância na literatura, pela rápida convergência e estabilidade numérica. Diferentes versões propostas ao longo dos anos consideram modificações à matriz de *scaling*, tradicionalmente a matriz identidade, com a característica de serem não singulares. Estudamos o uso de LMM com matriz de *scaling* singular, motivados por bons resultados de técnicas similares em problemas lineares através, por exemplo, da regularização de Tikhonov. Verificamos que, sob hipóteses razoáveis, pontos de acumulação da sequência gerada por LMM com *scaling* singular são pontos estacionários do problema original. Com uso do critério de Armijo para escolha do passo e uma condição de *error bound*, mostramos que esta convergência ocorre com taxa quadrática localmente, tal qual a versão clássica de LMM. Em aplicações numéricas, estudamos um problema de condução de calor em 2D modelado por uma equação diferencial parcial com condições de fronteira mistas e condição inicial. Conhecer a condutividade térmica de um material é assunto de importância em processos industriais e na ciência, um tópico ativo na pesquisa das últimas décadas. Fornecemos uma forma de discretizar o modelo original através do método pseudo-espectral de Chebyshev nas variáveis espaciais, pelas suas boas capacidades de aproximação com baixo custo numérico, e o método de Crank-Nicolson na variável temporal. O problema inverso de aproximar a condutividade térmica a partir de dados capturados de temperatura é então elaborado como um problema de mínimos quadrados não linear, que faz uso ativo da discretização pregressa. A minimização é feita através de LMM com matrizes de *scaling* singular escolhidas para representarem operadores de derivação discretos de primeira e segunda ordens, com a intenção de introduzir suavidade nos iterados construídos. Para amenizar o efeito de imprecisões nos dados de temperatura fornecidos, o princípio da discrepância é utilizado como critério de parada. Resultados numéricos sintéticos ilustram as capacidades da técnica proposta, com reconstruções de qualidade a um baixo custo operacional, mesmo em situações com restrições de medição. Um exemplo com dados provenientes de um experimento físico com fonte de calor móvel é conduzido, evidenciando a robustez do método proposto, bem como sua aptidão para aplicação em problemas mal postos não lineares reais.

Palavras-chave: Problemas inversos. Métodos iterativos. Estimativas de erro. Semi-convergência. Método de Levenberg-Marquardt. Matriz de *scaling*. Condutividade térmica. Método pseudo-espectral de Chebyshev.

ABSTRACT

We propose advancements to iterative methods for solving linear and nonlinear inverse problems seeking to expand the existing theory and to apply in practical situations. Obtaining stable solutions for the so-called discrete ill-posed problems through iterative techniques demands the use of specialized stopping rules. In this sense, for a method proposed recently by Bazán and Boos, if the iterations are terminated using the Morozov discrepancy principle, there is an estimate of the relative error between the exact solution and the approximation generated by the method. Assuming a Hölder-type source condition, we show that the estimate can be significantly improved, in two distinct results. A third estimate is obtained, without additional conditions, of the order of the noise in the data. All provide enhancement to the original result that motivated the study. In the nonlinear case, solving least squares problems using the Levenberg-Marquardt method (LMM) is a constant in the literature, due to fast convergence and numerical stability. Different versions proposed over the years consider modifications to the scaling matrix, traditionally the identity matrix, with the characteristic of being non-singular. We studied the use of LMM with singular scaling matrix, motivated by good results of similar techniques in linear problems through, for example, Tikhonov regularization. We verified that, under reasonable assumptions, accumulation points of the sequence generated by LMM with singular scaling are stationary points of the original problem. Furthermore, using Armijo criterion to choose the step and an error bound condition, we show that this convergence occurs with a quadratic rate locally, as the classical version of LMM. In numerical applications, we study a 2D heat conduction problem modeled by a partial differential equation with mixed boundary conditions and initial condition. Knowing the thermal conductivity of a material is a matter of importance in industrial processes and in science, an active research topic in the last decades. We provide a way to discretize the original model through the Chebyshev pseudospectral method in the spatial variables, for its good approximation capabilities at low numerical cost, and Crank-Nicolson method in the time-dependent variable. The inverse problem of approximating the thermal conductivity from captured temperature data is then elaborated as a nonlinear least squares problem that makes active use of the provided discretization. The minimization is done through LMM with singular scaling matrices chosen to represent first and second order discrete derivative operators, with the intention of introducing smoothness in the constructed iterates. To mitigate the effect of inaccuracies in the provided temperature data, the discrepancy principle is used as stopping criterion. Synthetic numerical results illustrate the capabilities of the proposed technique, with high-quality reconstructions at a low operational cost, even in situations with restricted measurements. An example with data from a physical experiment with a mobile heat source is conducted, evidencing the robustness of the proposed method, as well as its capability for application in real nonlinear ill-posed problems.

Keywords: Inverse problems. Iterative methods. Error estimates. Semi-convergence. Levenberg-Marquardt method. Scaling matrix. Thermal conductivity. Chebyshev pseudospectral method.

LISTA DE ABREVIATURAS E SIGLAS

ART	Técnicas de Reconstrução Algébrica (*Algebraic Reconstruction Techniques*)
CAV	*Component Averaging*
CGLS	Método dos Gradientes Conjugados para mínimos quadrados (*Conjugate Gradient method for Least Squares*)
CN	Método de Crank-Nicolson
CPM	Método pseudo-espectral de Chebyshev
DP	Princípio da discrepância (*discrepancy principle*)
DROP	*Diagonally Relaxed Orthogonal Projections*
EDO	Equação diferencial ordinária
EDP	Equação diferencial parcial
FEM	Método de elementos finitos
GMRES	Método dos Resíduos Mínimos Generalizados (*Generalized Minimal Residual*)
GSVD	Decomposição em valores singulares generalizada (*generalized singular value decomposition*)
LMM	Método de Levenberg-Marquardt
LMMSS	Método de Levenberg-Marquardt com *scaling* singular
LSQR	Método *Least Squares* QR
MI	Número máximo de iterações
MINRES	Método dos Resíduos Mínimos (*Minimal Residual*)
MR-II	Variação de MINRES
MS1, MS2	Exemplos de cenários de medida restrita (*measument scenarios*)
NL	Nível de ruído (*noise level*)
PVI	Problema de valor inicial
RE	Erro relativo

RRGMRES	Variação de GMRES
RTRE	Erro restrito de reconstrução da temperatura (*restricted temperature reconstruction error*)
SIRT	Técnicas de Reconstrução Iterativa Simultânea (*Simultaneous Iterative Reconstruction Techniques*)
SOR	Sobre-Relaxação Sucessiva (*Successive Over-Relaxation*)
SVD	Decomposição em valores singulares (*singular value decomposition*)
SVE	Expansão em valores singulares (*singular value expansion*)
TGSVD	*Truncated* GSVD
TRE	Erro de reconstrução da temperatura (*temperature reconstruction error*)
TSVD	*Truncated* SVD

LISTA DE SÍMBOLOS

$\mathbb{N}$	Conjunto dos números naturais.
$\mathbb{R}$	Corpo dos números reais.
$\mathbb{R}^{m\times n}$	Espaço das matrizes de m linhas e n colunas com coeficientes em $\mathbb{R}$.
$\mathbb{R}^n$	Espaço de vetores (coluna) com coeficientes em $\mathbb{R}$. Para $x \in \mathbb{R}^n$, o representamos por $x = (x_1, x_2, \ldots, x_n)$, em que cada entrada x_i pertence a $\mathbb{R}$. Observamos também que $\mathbb{R}^n \cong \mathbb{R}^{n\times 1}$.
A^T	Transposta de A, isto é, se $A = (a_{ij}) \in \mathbb{R}^{m\times n}$, então $A^T = (b_{ij}) \in \mathbb{R}^{n\times m}$é tal que $b_{ij} := a_{ji}$.
$\langle x, y\rangle$	Produto interno canônico em $\mathbb{R}^n$, isto é, $\langle x, y\rangle = \sum_{i=1}^n x_i y_i \equiv x^T y$.
$\|x\|_2$	Norma vetorial canônica em $\mathbb{R}^n$ dada por $\|x\|_2 = \sqrt{\langle x, x\rangle} \equiv \sqrt{x^T x} = \sqrt{x_1^2 + \cdots + x_n^2}$, para $x \in \mathbb{R}^n$.
$I,\ I_n$	Matriz identidade de ordem n.
$\mathbf{0}$	Matriz ou vetor composto unicamente por zeros.
$\mathrm{diag}(v)$	Para $v = (v_1, v_2, \ldots, v_r) \in \mathbb{R}^r$ é vetor genérico, $\mathrm{diag}(v)$ é matriz (quadrada ou não, a depender do contexto) em que os primeiros r elementos da diagonal principal são, ordenadamente, $v_1, v_2, \ldots, v_r$ e os demais são nulos (tanto na diagonal quanto fora dela).
$e_k,\ \mathbf{e}_k$	k-ésimo vetor da base canônica, isto é, e_k é vetor composto por zeros, em que a k-ésima entrada é igual a 1.
$\mathcal{N}(A)$	Núcleo do operador A.
$\mathcal{R}(A)$	Imagem do operador A.
$\mathrm{span}\{\ldots\}$	Espaço gerado por uma seleção de vetores, isto é, conjunto de todas as combinações lineares dos vetores fornecidos.
A^{-1}	Matriz inversa de A, quando existente.
$A^\dagger$	Matriz pseudo-inversa de Moore-Penrose de A.
$\sigma_i,\ \sigma_i(A)$	i-ésimo valor singular de A.
$\gamma_i,\ \gamma_{i,k}$	i-ésimo valor singular generalizado de um par de matrizes $(A,\ L)$ ou $(A_k,\ L_k)$.

$\|A\|_2$	2-norma matricial (também chamada de norma espectral), dada por $\|A\|_2 = \sup_{x \neq \mathbf{0}} \frac{\|Ax\|_2}{\|x\|_2} \equiv \sqrt{\lambda_{max}(A^T A)}$, em que $\lambda_{max}(A^T A)$ corresponde ao maior autovalor de $A^T A$.
$\|A\|_F$	Norma de Frobenius de $A \in \mathbb{R}^{m \times n}$, dada por $\|A\|_F^2 = \sum_{i=1}^{m} \sum_{j=1}^{n} \|a_{ij}\|^2$.
$\rho(A)$	Raio espectral da matriz quadrada A, dado pelo máximo do valor absoluto dos autovalores de A.
$\kappa(A)$	Número de condição de A em relação à 2-norma matricial, i.e., $\kappa(A) = \|A\|_2 \|A^\dagger\|_2$.
posto(A)	Posto da matriz A, que corresponde, por exemplo, à quantidade de valores singulares não nulos de A.
$B(x, \rho)$	Bola (fechada) em torno do ponto $x \in \mathbb{R}^n$ e raio $\rho > 0$, mais especificamente, $B(x, \delta) = \{y \in \mathbb{R}^n \mid \|y - x\|_2 \leq \rho\}$.
$\nabla f(x)$	Vetor gradiente de $f : \mathbb{R}^n \longrightarrow \mathbb{R}$, isto é, $\nabla f(x) \in \mathbb{R}^n$ é vetor cuja i-ésima entrada é dada por $\frac{\partial f}{\partial x_i}(x)$, $i = 1, \ldots, n$.
$J(x)$	Matriz Jacobiana de $F : \mathbb{R}^n \to \mathbb{R}^m$, com $F(x) = (f_1(x), \ldots, f_m(x))^T$, isto é, $J(x) \in \mathbb{R}^{m \times n}$ é matriz cuja i-ésima linha é dada por $\nabla f_i(x)^T$, $i = 1, \ldots, m$.
$\{x_k\}_{k \in \mathcal{K}}$	Sequência de números/vetores/matrizes sujeitos ao subíndice $k \in \mathcal{K} \subseteq \mathbb{N}$, usualmente omitido quando $\mathcal{K} = \mathbb{N}$.
$\mathcal{K} \subseteq_\infty \mathbb{N}$	Denota que $\mathcal{K}$ é um subconjunto infinito de $\mathbb{N}$.
dist(x, D)	Distância entre o ponto $x \in \mathbb{R}^n$ e o conjunto $D \subset \mathbb{R}^n$, dada por dist$(x, D) = \inf\{\|x - y\|_2 \mid y \in D\}$.
δ	Notação para $\|e\|_2$, para problemas $Ax = b$ com perturbação nos dados de entrada da forma $\tilde{b} = b + e$.
$\mathcal{O}(\cdot)$	Para f, g apropriadas, $f(x) = \mathcal{O}(g(x))$ se, e somente se, existe $M > 0$ tal que $\|f(x)\|_2 \leq M \|g(x)\|_2$.

SUMÁRIO

1 INTRODUÇÃO

Problemas inversos são parte integrante da rotina diária de cientistas, matemáticos e engenheiros, com avanço expressivo nas últimas décadas [46], em parte impulsionado pelo crescimento do poder computacional no período. Uma enormidade de fenômenos pode ser entendida através de uma relação de causa e efeito, no sentido de que um estímulo aplicado a um sistema gera alguma resposta como fruto. Exemplos são inúmeros, partindo de problemas próximos às áreas de ciências exatas como sensoriamento remoto, aprendizagem de máquina, geofísica, astronomia, imagens médicas, identificação de parâmetros, mas multidisciplinar com acústica, ótica, biologia matemática, inferência estatística, identificação biométrica, ensaios não destrutivos, para citar alguns [46, 64, 79]. Matematicamente, os problemas são em geral modelados na estrutura

$$\mathsf{F}(\mathsf{x}) = \mathsf{y}, \tag{1.1}$$

para $\mathsf{F} : \mathsf{X} \longrightarrow \mathsf{Y}$ um operador (representando o "sistema"), possivelmente não linear, entre os conjuntos X e Y contendo, respectivamente, as "causas" x e os "efeitos" y. Determinar y conhecendo x é o que se chama de *problema direto*. Assim, a definição de *problema inverso* surge naturalmente quando conhecemos a resposta de um sistema e gostaríamos de saber quais foram os estímulos que a causaram. Em outras palavras, de posse de F e y, o problema inverso consiste da busca por elementos x que satisfaçam a igualdade $\mathsf{F}(\mathsf{x}) = \mathsf{y}$ (a inversão do operador, se existir).

Este trabalho tem por interesse central estudar técnicas iterativas para problemas inversos envolvendo operadores F entre espaços vetoriais reais de dimensão finita, lineares e não lineares. Especialmente, considerando situações em que o operador é sensível a pequenas modificações nos dados de entrada y, conhecidos como *problemas mal postos* [46, 65, 75], para os quais a construção de soluções estáveis costuma ser desafiadora. Nestes casos, frequentes em aplicações práticas, recebemos dados de entrada $\tilde{\mathsf{y}}$ (os dados observados de algum experimento físico, por exemplo), tais que $\tilde{\mathsf{y}} \approx \mathsf{y}$, porém os elementos que satisfazem $\mathsf{F}(\tilde{\mathsf{x}}) = \tilde{\mathsf{y}}$ pouco ou nada representam as soluções desejadas ao problema original (1.1).

Neste contexto, no caso linear, entendemos (1.1) como o conhecido sistema de equações lineares através da notação

$$Ax = \tilde{b}, \tag{1.2}$$

com $A : \mathbb{R}^n \longrightarrow \mathbb{R}^m$ severamente mal condicionada e $\tilde{b} = b + e$ representando um vetor de dados recebidos contaminados por um ruído desconhecido e, advindo de restrições nas medições, arredondamentos, falhas de equipamentos, etc. Esse problema foi classificado por Hansen [64, 65] como *problema discreto mal posto* e se apresenta em muitas frentes, uma usual sendo a discretização de equações integrais de Fredholm de primeira espécie [24, 46, 79]. Em geral, estamos interessados em solucionar o problema de mínimos quadrados associado

a (1.2),

$$\min_{x\in\mathbb{R}^n} \|Ax - \tilde{b}\|_2,$$

para o qual a solução livre de ruído $x^* = A^\dagger b$, com $A^\dagger$ a matriz pseudo-inversa de A [17, 92, 105], pode ser calculada por diversos métodos numéricos, o uso de técnicas de regularização sendo imperativo para amenizar o efeito do ruído. O método proposto por Tikhonov [115] é um exemplo, embora o destaque às técnicas iterativas seja natural com o crescimento da quantidade de variáveis envolvidas e também limitações relacionadas à avaliação da matriz A. Nestes, a solução é aproximada através de uma sequência de vetores $\{\tilde{x}^{(k)}\}$ que se caracterizam por, nas iterações iniciais, estarem próximos de x^* e irem deteriorando com o crescimento de k devido à maior influência do vetor e, no fenômeno chamado de *semi-convergência* [96]. O índice k atua, portanto, como parâmetro de regularização [75, 78, 115], e sua escolha para produzir boas aproximações de x^* é, por si só, não trivial. Entre exemplos de critérios especializados nesta escolha, citamos o princípio da discrepância (DP) [95], que se baseia no conhecimento da norma do ruído e, e métodos heurísticos [78] como a regra do produto mínimo (MPR) [27], critério da curva-L [63] e a validação cruzada generalizada (GCV) [54].

Entre a gama de métodos iterativos existentes, como LSQR [103], GMRES [110], TSVD/TGSVD [62, 64], método de Jacobi, Gauss-Seidel e SOR [59] e método de Kaczmarz [57], entre tantos outros [22, 30, 33–35, 53, 61, 80, 101], propomos em Bazán e Boos [11] o chamado de *método de Newton*, baseado nas iterações matriciais de Schultz [17, 111]. A ideia é construir iterados da forma $\tilde{x}^{(k)} = X_k\tilde{b}$, com $\{X_k\} \to A^\dagger$ proveniente do método de Schultz, as quais aliadas com critérios de parada apropriados produzem soluções de qualidade ao problema de interesse $Ax = b$ [11, 24]. Entre outras propriedades, verificamos que a convergência é quadrática e que a técnica é competitiva, especialmente se estratégias de projeção são utilizadas. Além disso, se DP é empregado como critério de parada às iterações e $b \in \mathcal{R}(A)$, provamos [11] que o erro relativo entre a solução exata e a capturada por DP, $\tilde{x}^{(k(\delta))}$, é da ordem de $\delta^{1/2}$, isto é,

$$\frac{\|x^* - \tilde{x}^{(k(\delta))}\|_2}{\|x^*\|_2} = \mathcal{O}\left(\delta^{1/2}\right), \tag{1.3}$$

em que $\delta := \|e\|_2$. Ou seja, é um resultado que garante a proximidade de $\tilde{x}^{(k(\delta))}$ de x^* se o ruído vai a zero e, além disso, fornece informações da qualidade da aproximação no pior caso. A pergunta natural quanto à possíveis refinamentos da estimativa acima caminham para a ideia de aumentar o expoente que acompanha δ para próximo ou igual a 1, de modo a obtermos estimativas diretamente proporcionais ao nível de ruído.

Parte deste trabalho é, então, dedicada a obter melhoramentos para (1.3). No contexto de dimensão infinita, é comum a suposição das chamadas *condições de fonte* [70], que basicamente fornecem informações quanto à suavidade de x^*, na busca por estimativas mais acuradas. Um exemplo é a *condição de fonte do tipo Hölder*, que assume

$x^* \in \mathcal{R}[(A^TA)^\mu]$, para $\mu > 0$, de modo que valores singulares pequenos de A (portanto seus vetores singulares associados) tem menor influência sobre x^*. Isto significa que as oscilações características destes vetores tendem a ser suprimidas, permitindo maior suavidade ao vetor [46, 64, 106]. A investigação proposta aqui busca provar que, sob a condição de fonte Hölder, o erro relativo em (1.3) é da ordem de $\delta^{\frac{2\mu}{2\mu+1}}$. Uma estimativa de mesma ordem é obtida, com a diferença de conter uma constante associada diferente e dependente dos chamados *coeficientes de Fourier* [65], que pode estreitar significantemente o resultado. Finalmente, obtemos uma última estimativa, sem dependência de condição de fonte, que mostra que o erro relativo pode ser da ordem de δ diretamente.

Todos estes partem de trabalhar com a desigualdade

$$\|x^* - \tilde{x}^{(k(\delta))}\|_2 \leq \|x^* - x^{(k)}\|_2 + \|x^{(k)} - \tilde{x}^{(k(\delta))}\|_2,$$

em que $x^{(k)}$ corresponde ao iterado gerado sem ruído nos dados, utilizando de propriedades da solução e do método, bem como a decomposição em valores singulares (SVD) [56], uma ferramenta essencial e muito poderosa à análise. Um ponto importante a ser observado diz respeito às estimativas no cenário discreto com condição de fonte. No contexto de dimensão infinita, as condições de fonte se relacionam diretamente com a suavidade da solução exata [46, 79, 96]. Em verdade, a análise segue válida para dimensão finita, embora a "suavidade" tenha um significado ligeiramente diferente, relacionado particularmente ao decaimento dos valores singulares, como comentado acima; veja [121, p. 8], [74].

Quanto aos problemas inversos não lineares, estamos interessados em lidar com problemas da forma

$$G(x) = y, \tag{1.4}$$

para $G : \mathbb{R}^n \longrightarrow \mathbb{R}^m$ continuamente diferenciável, um cenário desafiador pela estrutura de G ser em geral desconhecida. Assim como para problemas lineares, é comum substituirmos (1.4) pelo seu problema de mínimos quadrados associado, isto é,

$$\min_{x \in \mathbb{R}^n} \frac{1}{2}\|F(x)\|_2^2, \tag{1.5}$$

em que $F(x) := G(x) - y$, o que torna a resolução de uma equação não linear em um problema de otimização. Para a exposição deste trabalho, considere que o conjunto $\{x \in \mathbb{R}^n \mid F(x) = \mathbf{0}\}$ é não vazio. Como é sabido, buscar por mínimos globais para problemas de minimização irrestrita é uma tarefa, quando muito, numericamente indesejada, que pode inclusive ser impossível [42, 89, 97]. A busca então costuma se focar em pontos de mínimos locais, ou seja, que minimizam a função objetivo em uma certa vizinhança. Do ponto de vista teórico, garantir convergência de métodos para pontos de mínimo é, também, complexo: demanda informações de segunda ordem da função, que talvez nem estejam disponíveis. Na prática, a teoria para métodos iterativos de minimização irrestrita procura então garantir que pontos limite da sequência gerada pelo método sejam *estacionários*, ou seja, que o gradiente se anule em tal ponto. Esta é uma condição necessária ao ponto

ser mínimo local [19], geralmente "suficiente" no sentido de que, se o algoritmo tem por objetivo reduzir o valor da função em cada iteração, então é comum que os pontos obtidos sejam de fato mínimos locais [89].

Uma das técnicas mais importantes para (1.5) é o chamado *método de Levenberg-Marquardt* (LMM) [82, 87], que pode ser visto como uma variação do método de Gauss-Newton [77, 97] capaz de lidar com o mal condicionamento e posto incompleto das matrizes Jacobianas de F. A partir de um ponto $x_0 \in \mathbb{R}^n$ escolhido, LMM é definido pelas iterações

$$\begin{aligned}(J_k^T J_k + \lambda_k I) d_k^{LM} &= -J_k^T F_k \quad \text{e} \\ x_{k+1} &= x_k + \alpha_k d_k^{LM}, \quad k \geq 0,\end{aligned} \tag{1.6}$$

em que J_k é a Jacobiana de F em x_k, $F_k := F(x_k)$, $\lambda_k > 0$ é chamado de *parâmetro de* damping e $\alpha_k > 0$ é o *tamanho do passo*. Exemplos de formas de escolha destes parâmetros podem ser encontradas em [16, 19, 42, 122]. É sabido que LMM possui convergência de segunda ordem nas proximidades do minimizador [14, 18, 41, 82, 87, 93, 122], mesmo com hipóteses modestas como a de *error bound* [15, 49, 69, 76, 108], que exige menos que a condição clássica de posto completo da Jacobiana na solução.

O termo $\lambda_k I$ em (1.6) garante que a matriz $(J_k^T J_k + \lambda_k I)$ seja inversível independentemente da estrutura de J_k, e a matriz identidade é neste caso chamada de *matriz de* scaling [93]. No entanto, esta não é a única possibilidade, sendo que a literatura especializada é rica em casos de troca de I por outras matrizes [41, 48, 93, 98, 99, 112, 117, 124], a depender do cenário em estudo. Nestes, a matriz de *scaling* considerada é sempre não singular, de onde as demonstrações de convergência saem, por exemplo, visualizando LMM com um método de região de confiança [16, 48]. Por outro lado, existem exemplos em problemas lineares através da regularização de Tikhonov, por exemplo, em que a introdução de certas matrizes singulares junto ao processo iterativo podem produzir melhorias significativas às aproximações [64, 115]. De fato, a forma geral da regularização de Tikhonov [65] para sistemas lineares como em (1.2) pode ser vista na forma

$$(A^T A + \lambda^2 L^T L) x_{L,\lambda} = A^T \tilde{b}, \tag{1.7}$$

para λ parâmetro de regularização apropriado e $L \in \mathbb{R}^{p \times n}$, $p \leq n$, de posto completo, com a hipótese de $\mathcal{N}(A) \cap \mathcal{N}(L) = \{\mathbf{0}\}$ a fim de garantir unicidade da solução de (1.7). Pela natureza de muitos problemas discretos mal postos virem de discretizações de equações integrais, é comum considerar que as soluções procuradas sejam contínuas ou até diferenciáveis. Desta forma, uma escolha usual de L é na forma de matrizes retangulares representando versões discretas de operadores de derivação de primeira e segunda ordens, cuja atuação em (1.7) busca introduzir a suavidade de L nas aproximações obtidas (esperada da continuidade e/ou diferenciabilidade das mesmas). Os resultados são promissores [13, 62, 64, 68, 115] e as similaridades entre (1.7) e a direção d_k^{LM} em (1.6) sugerem uma construção similar para LMM, especialmente para problemas cuja suavidade é também esperada, como soluções de modelos de EDPs [8, 28, 73, 90, 100].

Uma parte deste trabalho se foca em formular uma versão de LMM com matriz de *scaling* singular com direção de descida calculada na forma

$$(J_k^T J_k + \lambda_k L^T L)d_k = -J_k^T F_k, \quad k \geq 0,$$

com L como acima, visando fornecer condições para a convergência do algoritmo e situações práticas em que o mesmo é vantajoso. Com a condição de $\mathcal{N}(J_k) \cap \mathcal{N}(L) = \{\mathbf{0}\}$ e escolha do parâmetro $\lambda_k = \|F_k\|_2^2$ [122], conseguimos mostrar que pontos de acumulação da sequência de iterados x_k produzidos são pontos estacionários para (1.5). Parte desta análise se utiliza da teoria de direções *gradient-related* e critérios de determinação de α_k como busca linear (exata e limitada) e critério de Armijo [19]. Mais ainda, de forma similar ao feito no trabalho de Yamashita e Fukushima [122], verificamos que o algoritmo converge localmente com taxa quadrática, provido que o tamanho do passo é determinado pelo critério de Armijo e F fornece um limitante de erro (*error bound*) nas proximidades do conjunto solução.

Exemplos em que esta variação de LMM pode ser de valia vem da reconstrução de condutividade térmica, tópico da parte final deste trabalho. Ali, descrevemos um problema de transporte de calor em duas dimensões modelado através de uma equação diferencial com condições de fronteira mistas e condição inicial [26, 86]. Ter informações térmicas de um material é de suma importância em diversas áreas técnicas, pelos óbvios usos na indústria, pesquisa e desenvolvimento [3, 5, 71, 85, 100]. Assim, existem inúmeros trabalhos na literatura que abordam este problema de diferentes formas, como em [2, 32, 86, 90, 104], apenas para citar alguns. Aqui, buscamos discretizar o modelo utilizando o método pseudo-espectral de Chebyshev (CPM) [31, 58], cujas capacidades de convergência chegam a ser exponenciais a depender da suavidade da solução, além de possibilitar implementações numéricas de baixo custo [8, 73]. CPM fornece então uma discretização às variáveis espaciais do problema, transformando o modelo contínuo em semidiscreto, dependente do tempo. Esta última variável é enfim discretizada através do método de Crank-Nicolson (CN) [39], de modo a obtermos uma forma em dimensão finita de aproximar valores de temperatura a partir dos demais dados (inclusive a condutividade), isto é, um *solver* ao problema direto. Este método foi escolhido por suas qualidades de convergência e estabilidade absoluta, ambas numericamente desejáveis.

Para o problema inverso de reconstruir valores de condutividade a partir de dados de temperatura (possivelmente contaminados com ruído proveniente de experimentos físicos ou mesmo imprecisão), um problema de mínimos quadrados não linear é formulado. Este se baseia intrinsecamente na solução do problema direto e é resolvido através do método de Levenberg-Marquardt com *scaling* singular. Resultados numéricos para exemplos sintéticos buscam ilustrar as capacidades da técnica em diferentes cenários, em que os operadores de derivação discretos são utilizados em conjunto das iterações. Testamos situações variadas de ruído e disponibilidade de dados, algo que pode ser comum em situações práticas, com resultados de baixo custo operacional e condizentes com o ruído introduzido. Comparativos entre a proposta de LMM deste trabalho, a versão clássica de LMM ($L = I$) e um método

de região de confiança reflexiva através da rotina *lsqnonlin* do Matlab mostram as boas funcionalidades da técnica discutida, capaz de lidar com o ruído e a natureza mal posta dos problemas na construção de soluções estáveis. Ao final, exibimos um exemplo baseado em dados experimentais conduzido por Luchesi e Coelho [85], com informações fornecidas em apenas 12 sensores em uma peça de aço AISI 4340. O problema é modelado com uma fonte de calor móvel dado que provém de um procedimento de fresamento de face [5]. Os resultados mostram que a estrutura proposta é aplicável em problemas reais, permitindo que boas reconstruções de condutividade sejam obtidas com baixo custo numérico, com as matrizes de *scaling* singulares cumprindo papel essencial nas aproximações geradas.

Das contribuições desta tese, destacamos: as novas estimativas de erro obtidas para o método de Newton, que visam estreitar e melhor medir a qualidade das aproximações obtidas; formulação, proposta e demonstração de convergência à variante de LMM que utiliza matrizes de *scaling* singulares, com aplicações promissoras; discretização e solução numérica de um problema de condução de calor via CPM, CN e LMM (com operadores discretos de derivação), uma estratégia eficiente e aplicável em cenários diversos, tanto com limitações nos dados fornecidos quanto perturbações nas informações de entrada. Os dois primeiros temas estão sendo organizados em forma de artigo científico visando submissão, enquanto que o material de condução de calor foi publicado em Boos, Luchesi e Bazán [26] e uma sequência está em fase de submissão para revisão em Boos, Bazán e Luchesi [25].

1.1 OBJETIVOS

- *Objetivo geral:* Obter avanços em métodos iterativos para problemas inversos lineares e não lineares, com foco em um problema de identificação de parâmetros em transferência de calor.

- *Objetivos específicos:*

  - Enunciar conceitos, resultados e contextualização necessários à compreensão do texto;
  - Encontrar novas estimativas de erro para o método de Newton [11] dependentes do ruído nos dados;
  - Propor uma versão do método de Levenberg-Marquardt que utiliza matrizes de *scaling* singulares;
  - Analisar a convergência do método de Levenberg-Marquardt proposto neste trabalho;
  - Exibir uma discretização via técnica pseudo-espectral de Chebyshev de um modelo que descreve a transferência de calor em um sólido bidimensional;

– Construir aproximações à condutividade térmica utilizando a discretização obtida e calculando soluções através de um problema inverso associado.

1.2 ORGANIZAÇÃO DO TRABALHO

Esta tese está organizada em capítulos da seguinte forma, além da presente introdução:

- *Capítulo 2:* apresenta preliminares relativos à solução de problemas inversos lineares e não lineares, necessários ao tratamento numérico de tais problemas encontrados em aplicações correntes. No caso linear, comentamos sobre soluções, técnicas existentes e formas de lidar com ruído nos dados de entrada. Fornecemos maneiras de introduzir informações a priori em problemas lineares através, por exemplo, da regularização de Tikhonov [115] e da GSVD [62, 64]. Estes exemplos servem de motivação para uma versão não linear que envolve técnicas similares, desenvolvida adiante. Para o caso não linear, introduzimos conceitos de Otimização pertinentes, bem como um apanhado geral de métodos conhecidos como o método de Newton, Gauss-Newton e Levenberg-Marquardt, de importância no seguir do trabalho.

- *Capítulo 3:* destinado a problemas inversos lineares através do método de Newton proposto por Bazán e Boos [11] e novas estimativas de erro para o mesmo. Iniciamos apresentando propriedades básicas do método, bem como uma estimativa de erro para quando o princípio da discrepância (DP) de Morozov [95] é utilizado como critério de parada, já presentes no artigo supracitado. Em seguida, elaboramos as ferramentas necessárias para a obtenção de duas estimativas, tanto absolutas quanto relativas, agora dependentes da chamada condição de fonte do tipo Hölder [46, 70]. Por fim, elaboramos uma última estimativa, sem condição adicional. Estes três avanços buscam exibir cotas de erro mais finas que a presente em [11].

- *Capítulo 4:* retornamos aos problemas não lineares com objetivo de mostrar a convergência do método de Levenberg-Marquardt com *scaling* singular. Motivamos o uso de matrizes singulares junto a LMM em exemplos simples, analisando em conjunto os motivos desta técnica poder ser vantajosa. Em seguida, elaboramos as ferramentas e resultados necessários para mostrar que as iterações de LMM com *scaling* singular convergem. Mais precisamente, verificamos que pontos de acumulação da sequência gerada pelo método são pontos críticos do problema de mínimos quadrados não linear. A parte final do capítulo se foca em provar que tal convergência ocorre com taxa quadrática sob a condição de *error bound* [14, 69], em uma análise relacionada à conduzida por Yamashita e Fukushima [122].

- *Capítulo 5:* orientado à recuperação de condutividade térmica em um problema de transferência de calor, com aplicação em um problema industrial [25, 26]. A

partir de um modelo de equações diferenciais parciais, buscamos construir uma discretização deste utilizando da técnica pseudo-espectral de Chebyshev [31, 58] e do método de Crank-Nicolson [39]. Em seguida, utilizamos da discretização para descrever o problema inverso associado à reconstrução da condutividade através da minimização de um funcional de mínimos quadrados não linear. Para obter aproximações, aplicamos o método de Levenberg-Marquardt como desenvolvido no capítulo anterior em conjunto com o princípio da discrepância [95] como forma de lidar com o ruído nos dados. Como forma de exemplificar a efetividade da técnica, exibimos resultados numéricos ao final, em especial um problema baseado em dados experimentais captados em [85].

Em seguida, finalizamos o trabalho com conclusões e referências. Após, apresentamos dois apêndices, para complementar a argumentação de alguns pontos do texto. O primeiro, no tema de transformação do problema de Tikhonov da forma geral para a forma padrão, completa parte do visto no Capítulo 2. O seguinte, que aborda limites superiores e inferiores, acrescenta informações a parte do Capítulo 4.

2 PRELIMINARES

Este trabalho atua essencialmente em duas frentes, tratando de problemas inversos lineares e não lineares através, especialmente, da formulação de mínimos quadrados associada. Em última instância, é claro que o primeiro é um caso particular do segundo, porém existe toda uma miríade de técnicas específicas desenvolvidas para cada caso, com seus usos a depender do contexto. Neste capítulo familiarizaremos o leitor com conceitos e motivações envolvendo o desenvolvimento posterior do texto, com foco especial em técnicas iterativas.

2.1 ESTRATÉGIAS PARA PROBLEMAS INVERSOS LINEARES

Considere o sistema de equações lineares da forma

$$Ax = b,$$

com $A \in \mathbb{R}^{m\times n}$ e $b \in \mathbb{R}^m$, para $m, n \in \mathbb{N}$. Se $m = n$ e A é não singular, então a solução do problema é simplesmente dada por $x = A^{-1}b$, cujo cálculo usualmente evita a construção de A^{-1} pelo alto custo computacional, especialmente para matrizes de ordem mais alta. Formas mais eficientes de atacar este problema envolvem fatorar a matriz A em matrizes mais simples e então resolver $Ax = b$ através de sistemas lineares que se utilizem da estrutura destas fatorações. Métodos baseadas nesta ideia são chamados de *métodos diretos*, para os quais alguns exemplos de fatoração envolvem:

- *Fatoração LU*: escrever $A = LU$, se existe, com L matriz triangular inferior e U triangular superior;
- *Fatoração Cholesky*: escrever $A = GG^T$, se existe, com G triangular inferior;
- *Fatoração QR*: escrever $A = QR$, com Q ortogonal e R triangular superior.

Uma outra fatoração importante e que utilizaremos neste trabalho é a chamada *decomposição em valores singulares* (SVD), sumarizada no teorema seguinte.

Teorema 2.1 (SVD, [56]). *Seja $A \in \mathbb{R}^{m\times n}$ tal que* $\text{posto}(A) = r \leq \min\{m, n\}$. *Então, existem matrizes $U = [u_1, \ldots, u_m] \in \mathbb{R}^{m\times m}$ e $V = [v_1, \ldots, v_n] \in \mathbb{R}^{n\times n}$ ortogonais tais que*

$$A = U\Sigma V^T, \quad \textit{com} \quad \Sigma = \begin{bmatrix} C & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \in \mathbb{R}^{m\times n},$$

em que $C = \text{diag}(\sigma_1, \ldots, \sigma_r) \in \mathbb{R}^{r\times r}$ e os números σ_i (chamados de valores singulares*) são ordenados de modo não crescente, ou seja, $\sigma_1 \geq \cdots \geq \sigma_r > 0$.*

Observação 2.1. Ainda na SVD, os vetores u_i, $i = 1, \ldots, m$, e v_j, $j = 1, \ldots, n$, são chamados de *vetores singulares à esquerda* e *à direita* de A, respectivamente. Neste sentido,

vale comentar que U e V não são unicamente determinados, variando de acordo com o algoritmo utilizado na construção da SVD, embora os valores singulares sejam únicos. De fato, $\sigma_i = \sqrt{\lambda_i(A^TA)}$, em que $\lambda_i(A^TA)$ corresponde ao i-ésimo autovalor não nulo de A^TA. Estas e outras propriedades podem ser encontradas em Golub e Van Loan [56, Section 2.4] e Allaire e Kaber [4, Section 2.7], por exemplo.

Assim como as fatorações mencionadas acima, a SVD conduz também à solução do sistema linear na forma de um método direto, especialmente se tratando do caso de A não singular. De fato, para este temos $r = m = n$ e, pela ortogonalidade de U e V, temos que $x = V\Sigma^{-1}U^Tb$, em que Σ^{-1} é facilmente calculada dada sua estrutura diagonal.

No caso geral de $A \in \mathbb{R}^{m\times n}$, introduzimos então o *problema de mínimos quadrados linear*,

$$\min_{x\in\mathbb{R}^n} \|Ax - b\|_2, \tag{2.1}$$

que procura entre os vetores que compõem a imagem de A aquele que possui a menor distância (Euclidiana) em relação à b. Para obtermos o conjunto solução de (2.1), utilizamos a chamada *matriz pseudo-inversa de Moore-Penrose* [92,105], denotada por $A^\dagger$, que estende o conceito de inversa para matrizes quaisquer. Uma forma de construí-la é através da SVD de A, de modo que

$$A^\dagger = V \begin{bmatrix} C^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}_{n\times m} U^T,$$

podendo também ser procedente de outras decomposições ou ainda métodos iterativos, tema que tocaremos adiante. Assim, as soluções do problema de mínimos quadrados (2.1) são da forma $\{A^\dagger b + w \mid w \in \mathcal{N}(A)\}$, em que a de maior interesse e que nos focaremos em obter é

$$x^* := A^\dagger b, \tag{2.2}$$

da qual as outras soluções podem ser construídas com adição de um vetor do núcleo do operador em questão. Vale comentar que x^* corresponde à solução de norma mínima entre todas os vetores que resolvem o problema de mínimos quadrados linear (2.1), a única com tal propriedade, uma consequência geométrica. Portanto, x^* traz uma noção de existência e unicidade à busca de soluções para (2.1), independente da estrutura de A. Estas e outras informações relacionadas podem ser encontradas em praticamente qualquer livro de Álgebra Linear como [4,72,91], mas também citamos Boos [23,24] como materiais adicionais.

Em termos práticos, com o crescimento do número de variáveis e dificuldade de acesso à informação (não raro, a matriz A pode ser dada apenas através de produtos matriz-vetor, sem ser explicitamente fornecida), computar x^* através de (2.2) pode acarretar em alto custo computacional e ser, em alguns casos, impraticável. Isto pois, equiparado ao esforço de calcular matrizes inversas, construir $A^\dagger$ (quando possível) costuma ser evitado. Para tanto, técnicas alternativas foram desenvolvidas com o passar das décadas, com

destaque para os *métodos iterativos*, que se baseiam no desenvolvimento de uma sequência de vetores $\{x^{(k)}\}$ convergente ao ponto de interesse, no caso, x^*. Como exemplos de tais métodos, podemos citar, entre outros: CGLS/LSQR [22, 103], MINRES/MR-II [61, 101], GMRES/RRGMRES [30, 110], TSVD/TGSVD [64], método de Jacobi/Gauss-Seidel/SOR [59], métodos da classe SIRT (Landweber/Cimmino/CAV/DROP) [33–35, 53, 80] e método de Kaczmarz (classe ART) [57]. Suas características particulares e vantagens/desvantagens de acordo com o cenário estão fora do escopo deste trabalho, embora seja importante mencionar a vasta presença de tais métodos na literatura, uma evidência do interesse teórico e prático envolvido.

Além destes, enfatizamos a técnica iterativa proposta por Bazán e Boos [11] e que se baseia nas chamadas iterações de Schultz [111] para a pseudo-inversa, isto é, uma sequência de matrizes posteriormente utilizada na construção de vetores que aproximam x^*. A estrutura do método e propriedades básicas serão melhor abordadas no capítulo seguinte, inteiramente focado nesta técnica.

2.1.1 Ruído nos dados e critérios de parada

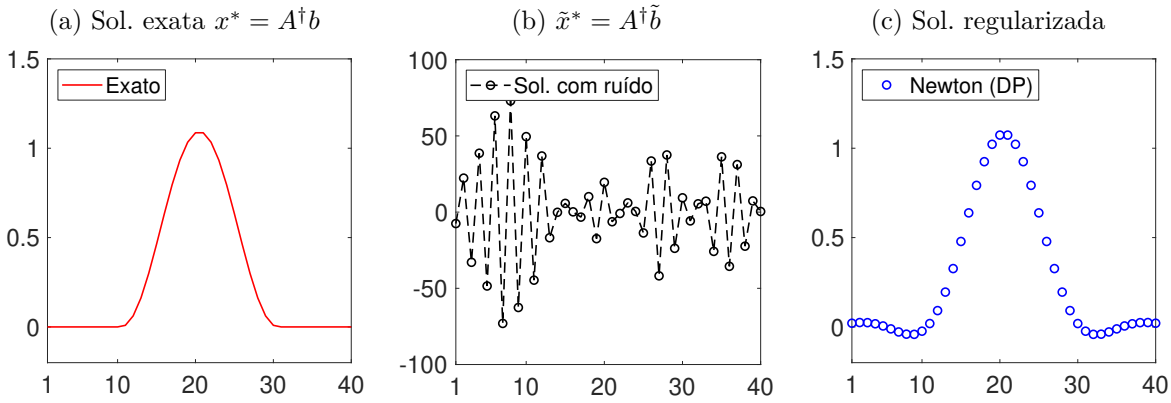
Um ponto importante a ser considerado diz respeito a resolver sistemas da forma

$$Ax = \tilde{b}, \quad \text{em que} \quad \tilde{b} = b + e, \quad \|e\|_2 =: \delta, \tag{2.3}$$

representa dados de entrada com ruído a partir dos dados exatos b (considerados indisponíveis), com a matriz A severamente mal condicionada. Problemas desta forma são conhecidos como *problemas discretos mal postos* [65] e surgem naturalmente na prática, como em processamento de sinais, restauração de imagens e tomografia computadorizada, para citar alguns [64, 96]. O ruído e é proveniente, por exemplo, de imprecisão numérica ou erros de medição, especialmente quando os dados são obtidas experimentalmente. Neste cenário, a solução do sistema com perturbação $Ax = \tilde{b}$, isto é, $\tilde{x}^* = A^\dagger \tilde{b}$, estará profundamente contaminada pelo ruído de modo a não apresentar proximidade com o vetor procurado $x^* = A^\dagger b$, solução do sistema com dados de entrada exatos [27, 64, 65]. Em outras palavras, para problemas discretos mal postos, a influência da amplificação do ruído representada pelo termo $A^\dagger e$ não pode ser ignorada [64, 65].

Para tentar mitigar o efeito de e, alguma estratégia de regularização se faz necessária para computarmos aproximações estáveis de x^*. No que tange métodos iterativos, alguns possuem a chamada propriedade de semi-convergência [96], que se caracteriza pela proximidade dos iterados $\{\tilde{x}^{(k)}\}$ de x^* nas nos passos iniciais para então o ruído gradativamente "deteriorar" as aproximações com o crescimento de k e finalmente convergir para $\tilde{x}^*$. Neste caso, escolher o índice k da iteração é fundamental para reduzir a influência de e e melhor aproximar x^*, na chamada *regularização iterativa* [27, 63, 78]. Em [11], verificamos que o método de Newton (3.2) apresenta a propriedade de semi-convergência e, mais ainda, captura informações relacionadas aos maiores valores singulares de A primeiro, importante

Figura 2.1 – Problema phillips [67], dados exatos e com ruído de 1%, $n = 40$. À direita, solução regularizada obtida por DP em conjunto com o método de Newton.



Fonte – o autor, 2022.

uma vez que a solução estável x^* é dominada pelos mesmos [64, 65]. O uso de critérios de parada eficientes e específicos é, então, essencial para tentarmos obter boas soluções, como é o caso de regras como a validação cruzada generalizada (GCV) [54], a regra do produto mínimo (MPR) [27] e o uso da curva-L [63], para citar alguns.

Outro critério de parada com ampla adoção e lastro teórico é o *princípio da discrepância* (DP) de Morozov [95], que sugere parar o processo iterativo no primeiro k tal que

$$\|\tilde{r}^{(k)}\|_2 \leq \tau\delta \leq \|\tilde{r}^{(k-1)}\|_2, \tag{2.4}$$

em que $\tilde{r}^{(k)} = \tilde{b} - A\tilde{x}^{(k)}$ e $\tau \gtrapprox 1$. Intuitivamente, DP deduz que as iterações devem terminar quando o resíduo encontrado seja da ordem do ruído introduzido aos dados. Veja que, como não temos acesso ao sistema livre de perturbações, considerar que as soluções capturadas dependam do tamanho do ruído é razoável. Na Figura 2.1 temos um exemplo que ilustra tanto o efeito do ruído em um problema mal posto como a ação de DP em conjunto com o método de Newton. Veja que mesmo sendo um problema pequeno (com matriz de ordem 40), a solução com ruído $\tilde{x}^*$ atinge valores de alta ordem e nada representam a informação procurada. Por outro lado, DP consegue capturar um iterado que compensa a influência do ruído com o que é dado do problema e fornece uma aproximação bastante coerente com a desejada. Vale mencionar que DP é utilizado também em problemas não lineares, uma vez que a natureza mal posta não é exclusiva de problemas lineares, como veremos adiante no trabalho. Outros comentários no efeito do ruído no caso linear podem ser encontrados em [24].

2.2 INCLUSÃO DE INFORMAÇÕES A PRIORI EM PROBLEMAS LINEARES

Em muitos casos, quando na tentativa de resolver sistemas da forma $Ax = b$, é possível que tenhamos algumas informações adicionais do problema que não estão contidas

na matriz A ou no vetor b. Esse tipo de informação inicial geralmente tem por característica introduzir algum tipo de regularização e/ou suavidade nas soluções aproximadas, podendo melhorar significativamente os resultados encontrados. Um exemplo vem da discretização de equações integrais, em que se espera que a solução seja contínua ou ainda diferenciável, algo não incluso naturalmente na discretização mas que pode contribuir na qualidade das soluções. Para o problema de minimização

$$\min_{x\in\mathbb{R}^n} \|Ax - b\|_2, \tag{2.5}$$

com $A \in \mathbb{R}^{m\times n}$ e $b \in \mathbb{R}^n$, considere possuímos informações adicionais presentes em uma matriz $L \in \mathbb{R}^{p\times n}$, $p \leq n$. Para os problemas que nos propomos nesta seção, L toma a forma de

$$L_1(m) = \begin{bmatrix} -1 & 1 & & \\ & \ddots & \ddots & \\ & & -1 & 1 \end{bmatrix} \in \mathbb{R}^{(m-1)\times m} \text{ ou } L_2(m) = \begin{bmatrix} 1 & -2 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & -2 & 1 \end{bmatrix} \in \mathbb{R}^{(m-2)\times m}, \tag{2.6}$$

embora nada impeça que tenhamos informações incluídas de outras formas. As matrizes L_1 e L_2 são versões discretas de operadores de derivação de primeira e segunda ordens, respectivamente, amplamente utilizadas, por exemplo, em problemas de reconstrução de imagens [13]. Daqui para frente assuma que $\mathcal{N}(A) \cap \mathcal{N}(L) = \{\mathbf{0}\}$ e que $L \in \mathbb{R}^{p\times n}$ possui posto completo, isto é, posto$(L) = p$.

Observação 2.2. Embora possam parecer restritivas, as duas hipóteses acima são bastante naturais em aplicações. De fato, em geral temos L como sendo alguma matriz retangular, o que impede considerações sobre sua inversibilidade; assim, a condição de posto completo para L é o máximo que se pode pedir. Por outro lado, se L não tem posto completo, é possível utilizar a SVD para construí-la alternativamente satisfazendo tal restrição, como veremos no último capítulo. Agora, a condição $\mathcal{N}(A) \cap \mathcal{N}(L) = \{\mathbf{0}\}$ é necessária para garantirmos unicidade à solução de sistemas lineares envolvendo o par de matrizes (A, L). Tendo em vista o caso em que A é proveniente de um problema mal posto (o que resulta em vetores singulares com bastante oscilação [64, 65]), em geral pensamos que L esteja introduzindo suavidade ao problema, de alguma maneira. Portanto, é pouco provável que um vetor de $\mathcal{N}(L)$ (supostamente suave) também seja um vetor de $\mathcal{N}(A)$.

Nas subseções seguintes, elaboraremos formas de incluir L na busca por soluções de (2.5) fazendo uso da regularização de Tikhonov e da GSVD, concluindo com exemplos ilustrativos.

2.2.1 Inclusão via problema de Tikhonov

Para o problema (2.5), é bastante conhecida a estratégia de regularização introduzida originalmente por Tikhonov [115] na forma

$$x_\lambda = \underset{x\in\mathbb{R}^n}{\operatorname{argmin}} \left\{ \|Ax - b\|_2^2 + \lambda^2 \|x - x_0\|_2^2 \right\}, \tag{2.7}$$

em que λ é o *parâmetro de regularização* e, em geral, x_0 é uma aproximação inicial da solução de (2.5) (caso não esteja disponível, tomamos $x_0 = \mathbf{0}$). A grande dificuldade se encontra em determinar λ de modo que a solução x_λ de (2.7) aproxime relativamente bem a solução do problema original (2.5). Esta técnica tem especial funcionalidade, como se sabe [64, 115], quando os dados de entrada contidos no vetor b possuem perturbações, embora possamos assumir dados exatos por hora. Num escopo mais geral, a regularização de Tikhonov toma a forma

$$x_{L,\lambda} = \underset{x\in\mathbb{R}^n}{\operatorname{argmin}} \left\{ \|Ax - b\|_2^2 + \lambda^2 \|L(x - x_0)\|_2^2 \right\}, \tag{2.8}$$

com L atuando diretamente no processo de minimização. Por notação, caso $L = I$, dizemos que (2.8) é um *problema de Tikhonov na forma padrão*. Para os outros casos, isto é, para $L \neq I$, dizemos que (2.8) é um *problema de Tikhonov na forma geral*. Agora, sabemos que (2.8) é equivalente a resolver

$$(A^T A + \lambda^2 L^T L) x_{L,\lambda} = A^T b - L^T L x_0, \tag{2.9}$$

que tem solução única por conta da hipótese $\mathcal{N}(A) \cap \mathcal{N}(L) = \{\mathbf{0}\}$. Além disso, veja que L atua conjuntamente a A na construção de $x_{L,\lambda}$, de modo que estamos ativamente incluindo componentes de L na solução.

Podemos também utilizar do problema de Tikhonov em outros métodos, como algoritmos iterativos, da forma que veremos a seguir. Lidar com a regularização de Tikhonov na forma geral dado na equação (2.8) pode levar a algumas de dificuldades, sem mencionar que já existem algoritmos eficientes para o problema na forma padrão (veja [12] para alguns exemplos). Portanto, surge a pergunta natural: é possível efetuar alguma transformação da forma geral para a forma padrão? Isto é, estamos buscando $\bar{A}$, $\bar{b}$ e $\bar{x}_0$ tais que (2.8) possa ser transformado em

$$\bar{x}_\lambda = \underset{\bar{x}\in\mathbb{R}^p}{\operatorname{argmin}} \left\{ \|\bar{A}\bar{x} - \bar{b}\|_2^2 + \lambda^2 \|\bar{x} - \bar{x}_0\|_2^2 \right\}. \tag{2.10}$$

Além disso, também queremos algum tipo de "regra de retorno", que permita relacionar a solução do problema transformado com a solução da sua forma geral. Em um exemplo didático, se L for não singular (i.e., $p = n$), temos:

$$\begin{aligned} & \|Ax - b\|_2^2 + \lambda^2 \|L(x - x_0)\|_2^2 \\ \Leftrightarrow \quad & \|(AL^{-1})Lx - b\|_2^2 + \lambda^2 \|Lx - Lx_0\|_2^2. \end{aligned} \tag{2.11}$$

Portanto, basta tomar $\bar{A} = AL^{-1}$, $\bar{b} = b$ e $\bar{x}_0 = Lx_0$ e temos a transformação para a forma padrão. Note que, para voltar ao problema original, basta resolver $Lx = \bar{x}$, o que implica em $x_{L,\lambda} = L^{-1}\bar{x}_\lambda$ (já que L é não singular), com $\bar{x}_\lambda$ solução do problema transformado. Para o caso de L geral, a transformação da forma geral para a forma padrão é sempre possível, embora um tanto mais complicada e, por isso, o leitor interessado pode encontrar no Apêndice A. Neste momento, a parte crucial de informação é que, ao efetuarmos a transformação, temos $\bar{A} = AL_A^\dagger$, em que $L_A^\dagger$ é chamada de *pseudo-inversa oblíqua (de A)* e definida por

$$L_A^\dagger = \left(I - \left(A\left(I - L^\dagger L\right)\right)^\dagger A\right) L^\dagger.$$

Novamente, para melhor entendimento com respeito à transformação e seu uso em métodos numéricos, recomendamos a leitura do Apêndice A. Veja que $\bar{A}$ possui influência tanto de A quanto de L, tornando factível a ideia de utilizar somente as informações do problema transformado em outros métodos como forma que incluir L na solução de (2.5).

Após a transformação, a estratégia sugere aplicarmos no problema transformado algum algoritmo conveniente para o problema na forma padrão, encontrando o parâmetro de regularização λ e, consequentemente, calculando $\bar{x}_\lambda$. Porém, para evitarmos o cálculo de λ ao mesmo tempo abrindo espaço para o uso de outras técnicas, consideramos o problema de minimização

$$\bar{x}^* = \underset{\bar{x}\in\mathbb{R}^p}{\operatorname{argmin}} \|\bar{A}\bar{x} - \bar{b}\|_2, \tag{2.12}$$

que é, em essência, apenas (2.10) com $\lambda = 0$. Pensando em técnicas regularizadoras, as soluções que obtemos através do problema de Tikhonov (2.7) tendem a ser similares às calculadas por métodos iterativos diretamente em (2.5) com algum critério de parada apropriado. Um comentário similar, portanto, ocorre entre os problemas (2.10) e (2.12). Desta forma, ao computar soluções estáveis para (2.12) e aplicando a transformação reversa nelas, esperamos estar, de fato, computando boas aproximações para $x^* = A^\dagger b$ e com influência de L.

Tendo em mente este cenário, podemos então aplicar essencialmente qualquer método (direto ou iterativo, como já mencionados na Seção 2.1) no sistema $\bar{A}\bar{x} = \bar{b}$ de modo a aproximar uma solução estável para $\bar{x}^* = \bar{A}^\dagger\bar{b}$; em seguida, aplicamos a transformação reversa nesta solução computada produzindo, portanto, alguma aproximação para x^*.

Observação 2.3. Existe ainda uma terceira forma de utilizar o problema de Tikhonov transformado, que é através da chamada estratégia *smoothing norm* (SN), proposta por Hansen e Jensen [68]. Falamos sobre tal abordagem com mais detalhe no Apêndice A.

2.2.2 Inclusão via GSVD

Uma outra forma de incluir informações a priori no sistema linear é através da GSVD (*decomposição em valores singulares generalizada*), que consiste de uma SVD conjunta para um par de matrizes (A, L). Esta decomposição em geral introduz dados que

uma das matrizes contém na decomposição da outra, o que torna a noção interessante no contexto dessa seção. Com efeito, são geradas duas decomposições SVD: uma para A e outra para L, com a propriedade de que A e L compartilham os mesmos vetores singulares à direita. Por precisão histórica, a GSVD foi introduzida inicialmente por Van Loan [119] e posteriormente desenvolvida/expandida em outros trabalhos [102, 120]. A versão que fornecemos aqui, por conveniência à abordagem, vem de Hansen [64]. Para uma demonstração e propriedades, além das referências acima, citamos também Björck [21].

Teorema 2.2 (GSVD, [64, p. 22]). *Considere o par matricial (A, L), com $A \in \mathbb{R}^{m\times n}$, $L \in \mathbb{R}^{p\times n}$, $m \geq n \geq p$, $\text{posto}(L) = p$ e $\mathcal{N}(A) \cap \mathcal{N}(L) = \{\mathbf{0}\}$. Então existem matrizes $U \in \mathbb{R}^{m\times n}$ e $V \in \mathbb{R}^{p\times p}$ com colunas ortonormais (i.e., $U^TU = I_n$ e $V^TV = I_p$) e $X \in \mathbb{R}^{n\times n}$ não singular tais que*

$$A = U \begin{bmatrix} \Sigma & \mathbf{0} \\ \mathbf{0} & I_{n-p} \end{bmatrix} X^{-1} \qquad e \qquad L = V \begin{bmatrix} M & \mathbf{0} \end{bmatrix} X^{-1}, \tag{2.13}$$

em que Σ e M são diagonais:

$$\Sigma = \text{diag}(\sigma_1, \ldots, \sigma_p) \in \mathbb{R}^{p\times p} \quad e \quad M = \text{diag}(\mu_1, \ldots, \mu_p) \in \mathbb{R}^{p\times p}. \tag{2.14}$$

Além disso, os elementos da diagonal de Σ e M são não negativos e ordenados de forma que

$$0 \leq \sigma_1 \leq \cdots \leq \sigma_p \leq 1 \quad e \quad 1 \geq \mu_1 \geq \cdots \geq \mu_p > 0 \tag{2.15}$$

e normalizados através da relação $\sigma_i^2 + \mu_i^2 = 1$, para $i = 1, \ldots, p$. Por notação, os quocientes

$$\gamma_i = \frac{\sigma_i}{\mu_i}, \quad i = 1, \ldots, p,$$

são chamados de valores singulares generalizados do par (A, L).

Para utilizarmos a GSVD, daremos um passo atrás. Quando de posse da SVD de A no formato do Teorema 2.1, um método iterativo que surge diretamente constrói iterados da forma

$$x^{(k)} = \sum_{i=1}^{k} \frac{u_i^T b}{\sigma_i} v_i,$$

para $k \leq r = \text{posto}(A)$, chamado de TSVD (*Truncated SVD*) [64]. Estes iterados convergem para x^*, uma vez que

$$x^* = A^\dagger b = \sum_{i=1}^{r} \frac{u_i^T b}{\sigma_i} v_i,$$

facilmente extraída da SVD. No caso de problemas com ruído no vetor de dados b, como vimos, k atua como parâmetro de regularização, "filtrando" as componentes de alta oscilação associadas a valores singulares pequenos. No caso livre de perturbações, esperamos que

$x^{(k)} \to x^*$ conforme $k \to r$. De forma análoga, a GSVD (Teorema 2.2) motiva a TGSVD (*Truncated GSVD*):

$$x^{(L,k)} = \sum_{i=p-k+1}^{p} \frac{u_i^T b}{\sigma_i} x_i + \sum_{i=p+1}^{n} (u_i^T b) x_i, \quad 1 \leq k \leq p.$$

Para maiores informações, veja [62, 64]. Em essência, este algoritmo é o método TSVD aplicado à decomposição SVD de A apresentada em (2.13), filtrando os menores σ_i's, com k atuando como parâmetro de regularização no caso com ruído. Como a GSVD consiste de uma decomposição conjunta do par (A, L), é natural pensar que teremos influência de L na solução computada por este método.

2.2.3 Exemplos numéricos

Para ilustrar na prática o efeito de incluir informações adicionais na forma discutida nesta seção, utilizaremos problemas provenientes do pacote *Regularization Tools*, de Hansen [67], mais especificamente deriv2 (exemplos 1 e 2) e baart. Os três correspondem à discretizações de equações integrais de Fredholm de primeira espécie, que usualmente conduzem a problemas mal postos [24, 65]. Como são provenientes de equações integrais, é razoável considerar que tais soluções sejam uma ou mesmo duas vezes diferenciáveis (suaves, portanto), justificando a introdução das matrizes L_1 e L_2 de (2.6) junto à reconstrução.

Quanto aos métodos, optamos por efetuar comparativos entre as soluções geradas por TSVD, TGSVD e Newton [11], em que aplicamos o método de Newton em dois cenários: em (2.5) e em (2.12). Veja que este último inclui a estratégia de Tikhonov, e é denotada por "Newton (Tik.)" nos resultados a seguir, que podem ser vistos na Figura 2.2 e na Tabela 2.1. São todos métodos iterativos para os quais escolhemos o princípio da discrepância (DP) como critério de parada, dado que incluímos 1% de ruído nos dados de entrada.

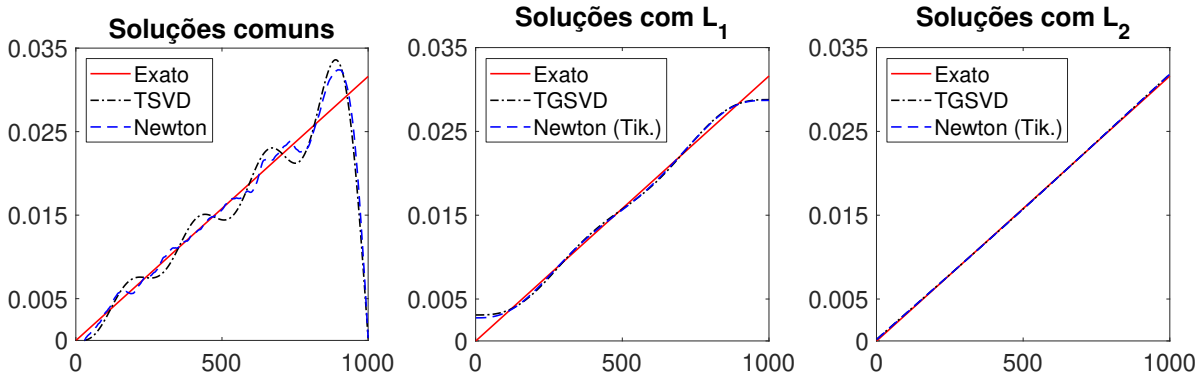
Visualmente, a Figura 2.2 exibe com clareza as diferenças entre incluir informação adicional ou não, especialmente para os exemplos de deriv2. Nestes, observe que o que chamamos de "soluções comuns", isto é, obtidas sem inclusão de dados a priori, possuem oscilações e erros de alta ordem nas fronteiras. Em contrapartida, incluir L_1 e L_2 tende a estabilizar os resultados obtidos através de suavidade e, no caso do operador de segunda ordem, essencialmente coincidir com a solução exata. No caso de baart, tecemos comentários similares, embora fique claro que L_2 produz as melhores soluções. Em termos dos erros relativos, a Tabela 2.1 complementa o visto nos gráficos: para deriv2, erros inicialmente da ordem de 25% reduzem para algo em torno de 4% e 0.5% com a aplicação de L_1 e L_2, respectivamente. Efeito similar ocorre com baart.

Portanto, estes exemplos ilustram que podemos ter ganhos efetivos e significativos na reconstrução de soluções para problemas lineares quando consideramos a inclusão

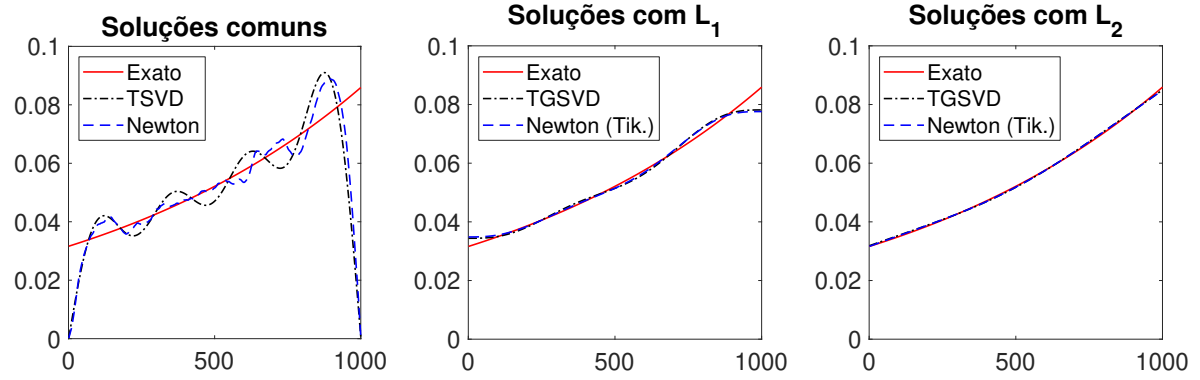
de informações adicionais no processo. Na próxima seção, discutiremos um pouco sobre problemas inversos não lineares, comentando sobre inclusões similares, com matrizes não singulares, feitas em métodos iterativos para tais problemas. Em conjunto, o assunto abordado nesta seção e na seguinte retorna para discussão aprofundada no Capítulo 4.

Figura 2.2 – Resultados da inclusão de informação adicional através das matrizes L_1 e L_2 (colunas do centro e direita) contra soluções usuais (coluna da esquerda), para diferentes problemas, dimensão 1000, 1% de ruído nos dados de entrada.

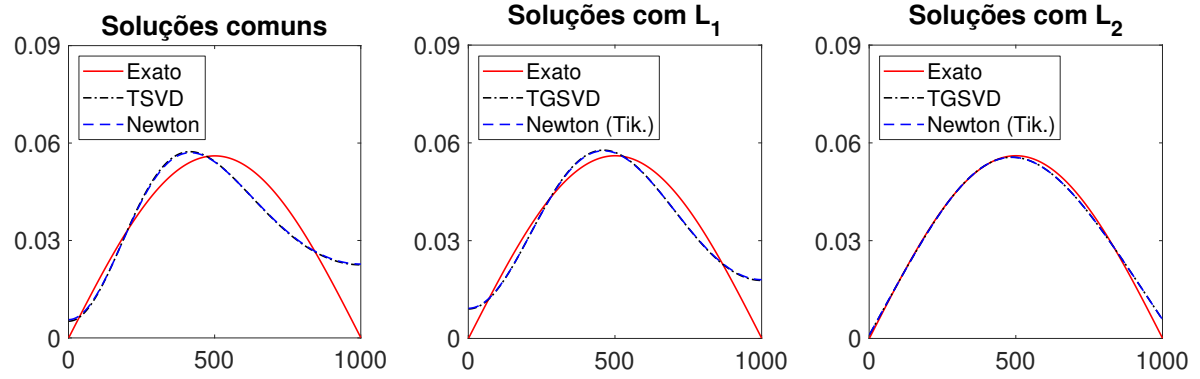
(a) deriv2 (exemplo 1).



(b) deriv2 (exemplo 2).



(c) baart.



Fonte – o autor, 2022.

Tabela 2.1 – Erros relativos entre a solução exata e a obtida com diferentes métodos e regularizadores, referente aos resultados da Figura 2.2. Problemas com dimensão 1000, ruído de 1% nos dados de entrada.

Problema	TSVD	Newton	**Efeito de L_1**		**Efeito de L_2**	
			TGSVD	Newton (Tik.)	TGSVD	Newton (Tik.)
deriv2 (ex. 1)	0.2675	0.2407	0.0458	0.0421	0.0053	0.0051
deriv2 (ex. 2)	0.2605	0.2310	0.0298	0.0317	0.0057	0.0052
baart	0.1670	0.1663	0.1177	0.1185	0.0371	0.0373

Fonte – o autor, 2022.

2.3 PRINCÍPIOS DE OTIMIZAÇÃO IRRESTRITA

Quando se trata do problema de minimização irrestrita de uma função $f : \mathbb{R}^n \longrightarrow \mathbb{R}$, isto é,

$$\min_{x \in \mathbb{R}^n} f(x), \tag{2.16}$$

idealmente gostaríamos de encontrar pontos que são soluções globais de tal problema, ou seja, pontos para os quais o valor de f seja o menor possível em todo o espaço. Mais precisamente, dizemos que x^* é um *ponto de mínimo global* se

$$f(x^*) \leq f(x), \quad \forall x \in \mathbb{R}^n,$$

o que costuma ser inviável em aplicações práticas pela dificuldade inerente em verificar a afirmação globalmente (salvo exceções como, por exemplo, quando a estrutura da função é bem compreendida e fornece informações assertivas quanto aos pontos de mínimo). Desta forma, métodos de Otimização frequentemente buscam por pontos que ao menos na sua vizinhança próxima são soluções de (2.16), isto é, pontos que satisfazem

$$f(x^*) \leq f(x), \quad \forall x \in B(x^*, \varepsilon), \tag{2.17}$$

para algum $\varepsilon > 0$, de modo que x^* é então denominado *ponto de mínimo local*. Ainda nesta caracterização, x^* é chamado *ponto de mínimo local estrito* se vale a desigualdade estrita em (2.17), para todo $x \in B(x^*, \varepsilon) \backslash \{x^*\}$.

No cenário de minimização irrestrita presente no problema (2.16), podemos então perguntar quais condições um ponto $x^* \in \mathbb{R}^n$ deve satisfazer para que possa ser entendido como mínimo local. O exemplo que motiva o que apresentaremos nas próximas linhas vem do Cálculo em uma variável real, ou seja, para $f : \mathbb{R} \longrightarrow \mathbb{R}$ diferenciável. Neste caso, como é sabido, se um ponto $x^* \in \mathbb{R}$ é mínimo local de f, então $f'(x^*) = 0$ e $f''(x^*) \geq 0$. Por outro lado, para que um ponto $x^* \in \mathbb{R}$ seja mínimo local, é suficiente que $f'(x^*) = 0$ e $f''(x^*) > 0$. No caso de funções de várias variáveis a valores reais, a análise segue de forma semelhante. Sem informações adicionais quanto à estrutura de f, as duas proposições seguintes apresentam as chamadas *condições necessárias e suficientes de otimalidade de*

segunda ordem, cujas demonstrações podem ser encontradas em Bertsekas [19, Propositions 1.1.1 e 1.1.3], ou também em [42, 89].

Proposição 2.1 (Cond. necessárias de otimalidade de segunda ordem). *Seja* $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ *continuamente diferenciável em um aberto* D*. Se* $x^* \in D$ *é um mínimo local de* f*, então*

$$\nabla f(x^*) = \mathbf{0}.$$

Mais ainda, se f *é duas vezes continuamente diferenciável em* D*, então* $\nabla^2 f(x^*)$ *é semi-definida positiva.*

Proposição 2.2 (Cond. suficientes de otimalidade de segunda ordem). *Seja* $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ *duas vezes continuamente diferenciável em um aberto* D*. Seja* $x^* \in D$ *tal que*

$$\nabla f(x^*) = \mathbf{0}$$

e $\nabla^2 f(x^*)$ *é definida positiva. Então,* x^* *é mínimo local estrito para* f*.*

É natural, portanto, pensarmos em procurar pontos que satisfaçam tais condições para produzirmos candidatos a minimizadores. Porém, na prática, este procedimento pode apresentar grande dificuldade, uma vez que mesmo a resolução do sistema de equações não linear $\nabla f(x) = \mathbf{0}$ é por si só uma tarefa não trivial, frequentemente comparável a resolver (2.16) diretamente [19]. Além disso, informações das derivadas de segunda ordem podem não estar disponíveis ou serem impossíveis de calcular, por motivos relacionados à formulação da função em questão. O que costuma ser feito, então, na tentativa de contornar custos numéricos e produzir técnicas viáveis, é utilizar de métodos iterativos que eventualmente convirjam para algum ponto que satisfaça, ao menos, a condição necessária de primeira ordem $\nabla f(x) = \mathbf{0}$. Por notação, $x^* \in \mathbb{R}^n$ que satisfaz

$$\nabla f(x^*) = \mathbf{0}$$

é chamado de *ponto estacionário* (ou *ponto crítico*). Como observam Martínez e Santos [89], é comum que pontos estacionários sejam de fato mínimos locais para f, especialmente se o método utilizado busca reduzir de forma efetiva o valor funcional de f em cada iteração. Desta forma, sem exigir mais informações da função, pedir que um ponto limite produzido pelo método seja estacionário é, do ponto de vista prático, o melhor que se pode esperar.

No que diz respeito à métodos numéricos para o problema de minimização sem restrições de funções suaves como proposto em (2.16), a literatura é rica em técnicas. Em geral, consistem de processos que partem de um iterado $x_0 \in \mathbb{R}^n$ e gradativamente o atualizam produzindo uma sequência $\{x_k\}$ que, idealmente, convirja a algum ponto x^* candidato a minimizador de $f(x)$. Ou, menos ainda, que a sequência $\{x_k\}$ tenha x^* como um de seus pontos de acumulação. A forma de produzir a atualização de x_k para x_{k+1} se baseia em informações de f (e possivelmente suas derivadas) em x_k e até nos pontos

anteriores gerados pela sequência, em alguns casos. Neste sentido, Nocedal e Wright [97] consideram que existem essencialmente duas classes fundamentais de métodos para efetuar esta atualização:

- *Busca linear:* a partir de x_k, o algoritmo decide por alguma direção p_k na qual deve procurar um novo iterado x_{k+1} que reduza o valor funcional de f. O termo *busca linear* (em inglês *line search*) é apropriado pois, após decidir a direção de busca p_k, precisamos resolver um problema secundário em uma variável para que x_{k+1} seja, aproximadamente, solução de

  $$\min_{\alpha>0} f(x_k + \alpha p_k).$$

  Claro que na prática se evita a solução exata do problema acima, que pode apresentar custo numericamente elevado, de modo que α é usualmente escolhido também de forma algorítmica, como veremos em parte no Capítulo 4. Então, um novo ponto x_{k+1} é escolhido e o processo é repetido.

- *Região de confiança:* a partir de x_k, a função objetivo f é substituída por uma outra função m_k que aproxime f localmente, a qual é então minimizada para produzir x_{k+1}. Um caso comum (mas não o único) é utilizar m_k como um modelo de segunda ordem de f, uma quadrática, que costuma ser mais fácil de manipular que f. Assim, o novo iterado é uma solução aproximada de

  $$\min_{p} m_k(x_k + p), \tag{2.18}$$

  de modo que $x_k + p$ esteja contido em um conjunto chamado de *região de confiança* (em inglês *trust region*). Este conjunto visa assegurar que não nos afastemos muito de x_k, uma vez que m_k pode perder significativamente suas capacidades de aproximação com distanciamento de x_k. Em geral, se considera a região de confiança como sendo uma bola em torno de x_k de raio $\Delta > 0$.

Algumas referências adicionais nestas técnicas incluem, por exemplo, [19, 37, 38, 89]. No escopo deste trabalho, despenderemos maiores esforços na tratativa da primeira estratégia, muito embora algoritmos de região de confiança tenham também surgimento na narrativa, justificando a menção aqui.

2.3.1 O problema de mínimos quadrados não linear

O maior interesse deste trabalho recai ao caso em que $f(x)$ em (2.16) corresponde à uma soma de quadrados. Mais especificamente, quando escrevemos

$$f(x) = \frac{1}{2}\|F(x)\|_2^2, \quad F : \mathbb{R}^n \longrightarrow \mathbb{R}^m, \tag{2.19}$$

a equação (2.16) se torna o que conhecemos como *problema de mínimos quadrados não linear*. Veja que, se $F(x) = Ax - b$, para A matriz e b vetor apropriados, recaímos no caso

comentado na Seção 2.1. Em geral, porém, esperamos que F seja uma função não linear cujo desconhecimento da estrutura torna a abordagem sensivelmente mais desafiadora que o caso linear. Por simplicidade na abordagem, assuma que F é continuamente diferenciável e, por notação, $F(x)$ e sua matriz Jacobiana $J(x) \in \mathbb{R}^{m\times n}$ são da forma:

$$F(x) = \begin{bmatrix} f_1(x) \\ \vdots \\ f_m(x) \end{bmatrix} \quad \text{e} \quad J(x) = \begin{bmatrix} \nabla f_1(x)^T \\ \vdots \\ \nabla f_m(x)^T \end{bmatrix},$$

com $f_i : \mathbb{R}^n \longrightarrow \mathbb{R}$, para $i = 1, \ldots, m$, e os gradientes são

$$\nabla f_i(x) = \left[\frac{\partial f_i}{\partial x_1}(x), \frac{\partial f_i}{\partial x_2}(x), \ldots, \frac{\partial f_i}{\partial x_n}(x)\right]^T \in \mathbb{R}^n, \quad i = 1, \ldots, m.$$

Do ponto de vista de aplicações, minimizar uma soma de quadrados como em (2.19) possui tantas aplicações quanto sua versão linear, ambos costumeiramente surgindo como coadjuvantes na resolução de problemas maiores. Um exemplo em que isto ocorre frequentemente diz respeito à identificação de parâmetros: um modelo composto por equações diferenciais é tratado por alguma técnica de discretização que eventualmente recai em um sistema de equações não linear. Alguns destes exemplos, muitos deles recentes e que retomaremos no Capítulo 5, podem ser encontrados em [8,25,26,28,32,73,86,90,104]. Estes casos podem ser vistos como casos particulares do problema de ajuste de modelos (em inglês *data fitting*) [19,22,77], que em linhas gerais busca encontrar parâmetros para que algum modelo escolhido melhor represente os dados disponíveis. Matematicamente, dispomos de dados na forma (t_i, y_i), para $i = 1, \ldots, m$, representando valores funcionais y_i encontrados em estágios de tempo t_i, por exemplo, e de um modelo $g(t; x)$ que depende de um vetor de parâmetros x, desconhecido. O problema então consiste de buscarmos por x de modo a minimizar os resíduos

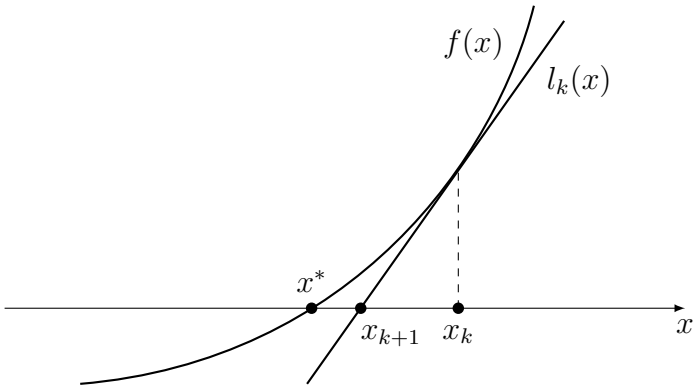
$$r_i(t_i) = y_i - g(t_i; x), \quad i = 1, \ldots, m.$$

A função g pode representar, por exemplo, um polinômio em t com x representando os coeficientes que acompanham as potências de t. Outras aplicações podem tomar g como uma distribuição de probabilidades, função exponencial, logarítmica, etc, ou combinações destas, a depender do entendimento do problema [97]. Problemas de ponto fixo podem também ser englobados nesta técnica, isto é, quando buscamos $x \in \mathbb{R}^n$ tal que $x = G(x)$, para G função de dimensões apropriadas, para o qual basta definir $F(x) := x - G(x)$.

Além destes, é importante lembrar da relevância de tal problema no contexto de Otimização, uma vez que obter candidatos a minimizador para (2.16) necessariamente demanda que tenhamos pontos satisfazendo $\nabla f(x) = \mathbf{0}$, um sistema de equações não linear. Este é um caso em que se busca resolver problemas da forma

$$F(x) = \mathbf{0}, \tag{2.20}$$

Figura 2.3 – Método de Newton para encontrar zeros de funções, caso unidimensional.



Fonte – o autor, 2022.

com $F : \mathbb{R}^n \longrightarrow \mathbb{R}^n$ continuamente diferenciável. Para a versão unidimensional ($n = 1$), é bastante conhecida a estratégia que toma $x_0 \in \mathbb{R}$ e atualiza

$$x_{k+1} = x_k - \frac{F(x_k)}{F'(x_k)}, \quad k \geq 0, \tag{2.21}$$

sempre que $F'(x_k) \neq 0$, que se conhece como *método de Newton-Raphson* [55,89]. A ideia consiste de, em cada iteração, produzir uma aproximação linear de F em torno de x_k utilizando a derivada $F'(x_k)$ e então construir o próximo iterado. De fato, é sabido que, nas proximidades de x_k,

$$F(x) \approx l_k(x) := F(x_k) + F'(x_k)(x - x_k),$$

de modo que x_{k+1} surge naturalmente como solução de $l_k(x) = 0$, sugerindo que ao encontrar um ponto que anule l_k estamos nos aproximando de x^* tal que $F(x^*) = 0$. A interpretação geométrica desta técnica é exibida na Figura 2.3, reforçando o interesse em resolver um problema de maior dificuldade através de uma sucessão de problemas mais simples. Neste caso em específico, encontrar um zero para F pode ser complexo, porém fazer o mesmo para funções lineares é algo facilmente alcançado.

Para o caso geral em $\mathbb{R}^n$, o então chamado somente de *método de Newton* [19,89,97] é construído analogamente. Pelo teorema de Taylor, podemos afirmar que

$$F(x) \approx l_k(x) := F_k + J_k(x - x_k), \tag{2.22}$$

em que $F_k := F(x_k)$ e $J_k := J(x_k)$. Igualando l_k a zero e assumindo que J_k é não singular, temos a atualização:

$$x_{k+1} = x_k - J_k^{-1} F_k, \quad k \geq 0.$$

É comum que este método seja escrito na forma

$$J_k d_k^N = -F_k \quad \text{e} \tag{2.23}$$

$$x_{k+1} = x_k + d_k^N, \quad \forall k \geq 0, \tag{2.24}$$

para $x_0 \in \mathbb{R}^n$ (chute inicial), explicitando que a direção d_k^N é obtida através da solução de um sistema linear e não pela construção da inversa J_k^{-1}, numericamente indesejável. Esta abordagem, simples e eficiente, permite diversas modificações que envolvem redução de custo operacional e/ou ganhos relacionados ao cálculo da matriz Jacobiana, dando origem a variantes como Newton inexato, Newton discreto, métodos Quasi-Newton, entre outros [89, 97].

A extensão direta desta técnica para o problema de mínimos quadrados não linear, isto é, com $f(x)$ como em (2.19), é apresentada pelo chamado *método de Gauss-Newton* [19, 77, 97], cuja direção d_k^{GN} é calculada através do sistema

$$(J_k^T J_k) d_k^{GN} = -J_k^T F_k, \quad k \geq 0,$$

e atualização de iterado dada por $x_{k+1} = x_k + d_k^{GN}$. Este método funciona sempre que J_k tenha posto completo de modo a d_k^{GN} estar bem definida. Neste caso, pela teoria de sistemas lineares,

$$d_k^{GN} = \underset{d \in \mathbb{R}^n}{\operatorname{argmin}} \frac{1}{2} \|J_k d + F_k\|_2^2$$

e, de fato, pelo posto completo de J_k, temos $d_k^{GN} = -J_k^{\dagger} F_k$. A forma de obter o método de Gauss-Newton consiste em aplicar o já conhecido método de Newton no sistema $\nabla f(x) = \mathbf{0}$, que neste caso consiste em $J(x)^T F(x) = \mathbf{0}$. Então, descartamos os termos de segunda ordem presentes na matriz Hessiana $\nabla^2 f(x)$, dada por

$$\nabla^2 f(x) = J(x)^T J(x) + S(x), \quad S(x) = \sum_{i=1}^{m} f_i(x) \nabla^2 f_i(x),$$

que seriam muito custosos para calcular em cada iteração. Além disso, é considerado que $S(x) \approx \mathbf{0}$ nas proximidades de um minimizador para f, ao menos no caso consistente [22, 77]. Como mencionam Nocedal e Wright [97, p. 254], na prática muitos problemas de mínimos quadrados não lineares possuem resíduo pequeno na solução, o que justifica ignorar $S(x)$ e, ainda, ter boas propriedades de convergência.

Para lidar com possível mal condicionamento e posto incompleto (perda de não singularidade de $J_k^T J_k$) nas Jacobianas, uma variação ao método de Gauss-Newton foi proposta por Levenberg [82] e Marquardt [87] e é ainda largamente aplicada. A diferença essencial consiste em calcular a direção d_k^{LM} através de

$$(J_k^T J_k + \lambda_k I) d_k^{LM} = -J_k^T F_k, \quad k \geq 0, \tag{2.25}$$

em que $\lambda_k > 0$ são escalares escolhidos em cada iteração e I é a matriz identidade de ordem n. Assim, com a atualização

$$x_{k+1} = x_k + d_k^{LM}, \quad k \geq 0,$$

temos a versão mais simples do chamado *método de Levenberg-Marquardt* (LMM), o qual abordaremos com mais detalhe no Capítulo 4. Um ponto interessante a observar é que

pela construção, LMM pode ser visto como um método de região de confiança [94], pois d_k^{LM} é solução de

$$\begin{array}{ll} \min & \frac{1}{2}\|J_k d + F_k\|_2^2 \\ \text{s.a} & \|d\|_2 \leq \Delta_k \end{array},$$

para algum $\Delta_k > 0$. Novamente, temos que a aproximação de primeira ordem obtida através do teorema de Taylor como em (2.22) é utilizada para modelar localmente o formato de F e caracterizando a estrutura da técnica de região de confiança como em (2.18).

Além da escolha de λ_k, diversos trabalhos ao longo das décadas propuseram substituir a matriz identidade em (2.25) por outros operadores que possam apresentar ainda mais vantagens na construção das direções de LMM. Uma opção comum substitui I por matrizes $\Omega_k \in \mathbb{R}^{n\times n}$ não singulares. Um exemplo emblemático na literatura considera tomar Ω_k como matriz diagonal [50]. Neste contexto, Moré [93] propôs que os elementos da diagonal contivessem informações da matriz Jacobiana J_k a fim de auxiliar na convergência e melhorar o condicionamento de (2.25). Outras formas de definir Ω_k podem ser encontradas em [41, 48, 98, 99, 112, 117, 124], com variações de acordo com o contexto do problema. Quanto à análise de convergência, é possível verificar que a sequência gerada por LMM com Ω_k possui pontos de acumulação que são estacionários [18, 41, 42, 93]. Mais ainda, tal convergência tem taxa quadrática localmente, assim como o método de Gauss-Newton. Estes resultados são válidos sempre que Ω_k é matriz não singular, de onde o próprio paralelo com o algoritmo de região de confiança conduz às demonstrações de convergência [16, 48].

Agora, vale ressaltar as similaridades entre a construção de d_k^{LM} em (2.25) e o cálculo do iterado de Tikhonov na forma geral, equação (2.9), especialmente se $L^TL = I$ ou $L^TL = \Omega_k$ e $x_0 = \mathbf{0}$. Desta forma, pensando nas possíveis vantagens apresentadas por introduzir L singular em sistemas lineares como comentado na última seção, surge o questionamento sobre fazer o mesmo com LMM. Seria possível reproduzir vantagens na convergência como as presenciadas para problemas lineares? Em quais cenários? É possível garantirmos convergência? Ou seja, quais hipóteses e estrutura necessitamos para garantir que iterações LMM com matrizes singulares sejam bem sucedidas e quais as possíveis vantagens numéricas da técnica? Este é o tema que norteia uma parte deste trabalho, como veremos no Capítulo 4 e também em aplicações no Capítulo 5. Em ambos, buscamos exibir casos em que a singularidade é aliada na construção de soluções de qualidade, em parte motivados pelos aspectos numéricos presentes em outros trabalhos, e.g. [8, 25, 26, 73].

3 MÉTODO ITERATIVO PARA PROBLEMAS LINEARES E ESTIMATIVAS DE ERRO

No artigo Bazán e Boos [11], propomos um método iterativo para a resolução de problemas lineares da forma $Ax = b$, em que $A \in \mathbb{R}^{m\times n}$ e $b \in \mathbb{R}^m$, para m e n naturais. No caso, estamos interessados em obter a solução de norma mínima do problema de mínimos quadrados associado,

$$\min_{x\in\mathbb{R}^n} \|Ax - b\|_2,$$

que existe e é única [17], denotada aqui por x^*. Mais ainda, tal solução tem forma definida por $x^* = A^\dagger b$, em que $A^\dagger$ representa a matriz pseudo-inversa de Moore-Penrose [92, 105] de A. O método se baseia nas iterações matriciais de Schultz [111], ou seja,

$$X_{k+1} = X_k(2I - AX_k), \quad k \geq 0, \tag{3.1}$$

de modo que, se $X_0 = \beta A^T$, $0 < \beta < \frac{2}{\rho(AA^T)}$, com $\rho(\cdot)$ denotando o *raio espectral de* $(\cdot)$ [56], segue a convergência $X_k \to A^\dagger$. Portanto, se construirmos a sequência vetorial

$$x^{(k)} = X_k b, \quad k \geq 0, \tag{3.2}$$

com $x^{(0)} = \beta A^T b$, temos que $x^{(k)} \to x^*$. A sequência vetorial assim gerada será denominada por *método de Newton* justamente pelo paralelo que possui com a busca por zeros de funções através do método homônimo. De fato, para $a \neq 0$ escalar real, calcular a^{-1} (o recíproco de a) é equivalente a buscar a raiz da função $f(x) = a - 1/x$. Se o fizermos utilizando o método (2.21), obtemos as iterações

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)} = x_k(2 - ax_k) \to \frac{1}{a},$$

para x_0 apropriado [24]. Generalizando para $f(X) = A - X^\dagger$, para $A \in \mathbb{R}^{m\times n}$ qualquer, produzimos exatamente (3.1) [17, 111].

Entre as propriedades interessantes obtidas em [11], provamos que $\{x^{(k)}\}$ de (3.2) herda a taxa de convergência quadrática verificada para o método matricial de Schultz. Além disso, verificamos que, se

$$0 < \beta < \frac{1}{\rho(AA^T)} = \frac{1}{\sigma_1^2},$$

então o método de Newton captura informações dos recíprocos dos valores singulares mantendo a sua ordem original [24, Teorema 3.7]. Esta informação é relevante pois evidencia a prioridade do método em incluir inicialmente elementos de baixa frequência da SVD de A, em geral responsáveis por conter componentes estáveis de x^* e, portanto, altamente desejados na construção de aproximações em problemas mal postos [27, 64, 65]. Vale também mencionar o resultado a seguir, que relaciona propriedades de monotonia na norma dos iterados e dos resíduos, a saber:

Teorema 3.1 (Bazán e Boos [11]). *Sejam $x^{(k)}$ sequência de vetores gerada pelo método de Newton e $r^{(k)} = b - Ax^{(k)}$ sequência dos resíduos correspondentes. Então, as seguintes propriedades são válidas:*

(i) $\|x^{(k+1)}\|_2 \geq \|x^{(k)}\|_2$, *para todo* $k \geq 1$.

(ii) $\|r^{(k+1)}\|_2 \leq \|r^{(k)}\|_2$, *para todo* $k \geq 0$.

Este teorema garante que algumas práticas de regularização heurísticas possam ser utilizadas como critério de parada junto ao método de Newton para mitigar o efeito no ruído nos dados de entrada. Exemplos de tais técnicas incluem, entre outras, a regra do produto mínimo (MPR) [27] e o uso da curva-L [63].

Observação 3.1. É interessante observar que, além dos aspectos teóricos, os resultados numéricos obtidos são atrativos, equiparando o que outros métodos bem estabelecidos na literatura produzem, como o método LSQR [103]. Pela brevidade da exposição, o leitor interessado encontra estas e outras informações complementares em Bazán e Boos [11]. Mais ainda, recomendamos a leitura de Boos [24] para uma explanação mais demorada em detalhes gerais, desde teoria, implementação e custo operacional, a técnicas de aceleração via projeção em subespaços.

Agora, caso consideremos o sistema com ruído nos dados, $Ax = \tilde{b}$, com $\tilde{b} = b + e$ e $\|e\|_2 =: \delta$, então sabemos que a sequência $\tilde{x}^{(k)} := X_k\tilde{b}$, $k \geq 0$, em geral necessita da chamada regularização iterativa para aproximar x^*. Neste sentido, é válido perguntar se há como estimar a qualidade das soluções obtidas com o método de Newton em conjunto com critérios de parada. Neste sentido, o teorema a seguir apresenta uma estimativa para o erro relativo entre a solução exata x^* e a capturada pelo método de Newton aliado ao princípio da discrepância (DP). Esta aproximação é aqui e no decorrer do texto denotada por $\tilde{x}^{(k(\delta))}$, com $k(\delta)$ representando o parâmetro de regularização.

Teorema 3.2 (Bazán e Boos [11]). *Suponha que $b \in \mathcal{R}(A)$ e denote o iterado gerado pelo princípio da discrepância por $\tilde{x}^{(k(\delta))}$. Então, é válida a estimativa*

$$\frac{\|x^* - \tilde{x}^{(k(\delta))}\|_2}{\|x^*\|_2} \leq C_{1/2}\,\kappa(A)\frac{\delta^{1/2}}{\|b\|_2^{1/2}} = \mathcal{O}\left(\delta^{1/2}\right), \tag{3.3}$$

com $C_{1/2} = \tau(1+\tau)^{1/2}/(\tau - 1)$.

Lembramos que $\kappa(A) = \|A\|_2\|A^\dagger\|_2$ denota o *número de condição de A* (com respeito à 2-norma matricial). Em palavras, este resultado fornece uma garantia teórica de que quanto menor o ruído imposto aos dados de entrada, mais próxima estará a solução gerada pelo método, $\tilde{x}^{(k(\delta))}$, da procurada x^*.

Além do aspecto da convergência quando $\delta \to 0$, este resultado fala da qualidade das aproximações obtidas. Neste sentido, vale se perguntar quão boa é uma estimativa

$\mathcal{O}(\delta^{1/2})$ e se seria possível torná-la mais fina. De fato, o melhor que podemos esperar é encontrar uma cota próxima ou exatamente de mesma ordem que δ, significando que o erro no pior caso é da ordem do ruído introduzido. Para visualizar o efeito das cotas, considere um exemplo em que $\delta = 0.01$, correspondendo a 1% de ruído. Assim, estimativas $\mathcal{O}(\delta^{1/2})$ e $\mathcal{O}(\delta)$ garantem, respectivamente, a menos de constantes, erros relativos no pior caso da ordem de 0.1 (10%) e 0.01 (1%), a segunda portanto significantemente mais confiante com relação à solução capturada. Obviamente, a influência das constantes que acompanham cada cota não pode ser ignorada, podendo conduzir a estimativas grosseiras mesmo com o crescimento do expoente que acompanha δ. Levando em conta estes apontamentos, abordamos nas subseções seguintes novas estimativas que obtivemos em estudos recentes, analisando também as constantes relacionadas.

3.1 FERRAMENTAS BÁSICAS E CONTEXTO

Parte essencial do material que exibiremos a seguir se baseia na decomposição SVD de A, do Teorema 2.1, que conduz à escrita da solução x^* e dos iterados $x^{(k)}$ (consequentemente $\widetilde{x}^{(k)}$ também) através de somatórios [11]:

$$x^* = \sum_{i=1}^{r} \frac{u_i^T b}{\sigma_i} v_i \quad \text{e} \quad x^{(k)} = \sum_{i=1}^{r} f_i^{(k)} \frac{u_i^T b}{\sigma_i} v_i, \tag{3.4}$$

para u_i e v_i o i-ésimo vetor-coluna de U e V, respectivamente, e

$$f_i^{(k)} = 1 - (1 - \beta\sigma_i^2)^{2^k}, \quad i = 1, 2, \ldots, r, \quad k \geq 0, \tag{3.5}$$

representando os chamados *fatores de filtro* [64,65] do método de Newton (3.2). A constante β é escolhida previamente satisfazendo $0 < \beta < 1/\sigma_1^2$ para garantirmos convergência da sequência $\{x^{(k)}\}$ e também boas propriedades do método [11]. Desta forma, vale que $|1 - \beta\sigma_i^2| < 1$ e, mais ainda,

$$0 < f_i^{(k)} < 1, \quad i = 1, 2, \ldots, r, \quad k \geq 0, \tag{3.6}$$

o que implica em $|1 - f_i^{(k)}| < 1$. Além disso, em todas as estimativas que exibiremos, estamos considerando o método de Newton (3.2) com parada dada pelo princípio da discrepância, isto é, para o primeiro $k = k(\delta)$ que satisfaz (2.4), e denotamos portanto este iterado por $\tilde{x}^{(k(\delta))}$.

Utilizando da estrutura acima, vamos mergulhar nos detalhes do trabalho desenvolvido e nas novas cotas de erro encontradas. Essencialmente, as estimativas são obtidas a partir da desigualdade triangular

$$\|x^* - \tilde{x}^{(k(\delta))}\|_2 \leq \|x^* - x^{(k)}\|_2 + \|x^{(k)} - \tilde{x}^{(k(\delta))}\|_2,$$

em que cada termo do lado direito é então estimado separadamente. As somas (3.4) permitem melhor tratamento de tais expressões, incluindo a criação de relações semelhantes

para os resíduos $r^{(k)}$ e $\tilde{r}^{(k)}$ que, em conjunto com as hipóteses impostas para cada estimativa e algebrismos, conduzem às cotas desejadas.

Por ordem de aparição no texto, apresentamos três estimativas: as duas primeiras relacionadas à utilização da hipótese adicional de condição de fonte do tipo Hölder e a última que tem por base as mesmas hipóteses da cota presente em [11]. Abordamos cada uma nas subseções seguintes e, após, apresentamos alguns comentários acerca dos resultados.

3.1.1 Condições de fonte

De forma sucinta, as chamadas *condições de fonte* consistem de informações prévias a respeito da solução exata x^* que, em algum sentido, consideram aspectos de suavidade da mesma. No ambiente de dimensão infinita, é possível provar [46, Proposition 3.11] que o erro entre x^* e suas aproximações obtidas através de técnicas de regularização para o problema (2.3) pode convergir para zero tão lentamente quanto se queira com $\delta \to 0$. Melhores taxas de convergência são então obtidas quando consideramos a solução exata restrita a subconjuntos do domínio do operador. Neste sentido, a *condição de fonte do tipo Hölder* sugere que $x^* \in \mathcal{R}[(A^TA)^\mu]$, para $\mu > 0$, de modo que $x^* = (A^TA)^\mu w$, para algum $w \in \mathbb{R}^n$. Observe que, pela SVD de A, é fácil verificar que o i-ésimo valor singular de $(A^TA)^\mu$ é da forma $\sigma_i^{2\mu}$. Como é sabido da literatura [64], vetores singulares associados a valores singulares pequenos são altamente oscilatórios. Assim, escrever $x^* = (A^TA)^\mu w$ induz à conclusão de que a expansão de x^*, isto é,

$$x^* = \sum_{i=1}^{r} \sigma_i^{2\mu}(v_i^T w)v_i,$$

é dominada por vetores singulares correspondentes a σ_i's maiores (especialmente com o crescimento de μ), que variam de forma suave. Situações que envolvem esta condição surgem naturalmente em aplicações como condução de calor, gradiometria de satélite e teoria de espalhamento, para citar alguns [70], motivando seu uso e foco nesta parte do trabalho. Como complemento, um outro exemplo é a chamada *condição de fonte logarítmica*, que assume algo da forma $x^* = \log^{-p}(A^TA)^{-1}v$, para $p > 0$ e v apropriado. Quando os problemas envolvem matrizes severamente mal condicionadas, em que os valores singulares tendem a zero exponencialmente, caracterizar x^* desta forma pode ser interessante. Novamente, a ideia gira em torno de esperar que σ_i's pequenos tenham pouca influência na solução exata, suprimindo a efeito de vetores singulares de alta frequência e, consequentemente, conduzindo a uma solução mais suave. Informações adicionais neste contexto de condições de fonte e decaimento podem ser encontradas, por exemplo, em [46, 64, 65, 70, 96, 106].

3.2 ESTIMATIVAS COM CONDIÇÃO DE FONTE DO TIPO HÖLDER

Para demonstrarmos o teorema central desta subseção, precisamos inicialmente de um resultado que conduz a uma cota superior para o quociente envolvendo os fatores de filtro e os valores singulares de A. Termos da forma $f_i^{(k)}/\sigma_i$ surgem no tratamento da estimativa por estarem presentes na formação dos iterados $x^{(k)}$ e precisam ser cotados superiormente em algumas situações. Como este desenvolvimento é paralelo e não se relaciona somente à presente estimativa, optamos por apresentar a investigação separadamente. Assim, considere a função

$$\Psi^{(k)}(\sigma,\lambda)=\frac{1-(1-\lambda\sigma^2)^{2^k}}{\sigma},\quad \text{restrita a}\quad 0<\lambda<\frac{1}{\sigma_1^2}\quad\text{e}\quad 0<\sigma<\frac{1}{\sqrt{\lambda}},\tag{3.7}$$

para a qual buscaremos pontos de máximo, que são caracterizados na proposição a seguir. Este resultado é baseado em estudo similar efetuado em [45] no contexto de uma análise de semi-convergência, que adaptamos para o caso em estudo.

Proposição 3.1. *Assuma (3.7) e seja λ fixo. Então, para todo $k\geq 1$, existe um ponto $\bar{\sigma}^{(k)}\in(0,1/\sqrt{\lambda})$ tal que*

$$\bar{\sigma}^{(k)}=\operatorname*{argmax}_{0<\sigma<1/\sqrt{\lambda}}\Psi^{(k)}(\sigma,\lambda).$$

Mais ainda, $\bar{\sigma}^{(k)}$ é único e dado por

$$\bar{\sigma}^{(k)}=\sqrt{\frac{1-\zeta^{(k)}}{\lambda}},\tag{3.8}$$

em que $\zeta^{(k)}$ é a única raiz em $(0,1)$ do polinômio

$$g_k(y)=(2^{k+1}-1)y^{2^k-1}-(y^{2^k-2}+\cdots+y+1).$$

Demonstração. Denotemos por $\Psi^{(k)'}$ a derivada de $\Psi^{(k)}$ com respeito a σ. Assim,

$$\Psi^{(k)'}(\sigma,\lambda)=\frac{2^{k+1}\lambda\sigma^2(1-\lambda\sigma^2)^{2^k-1}-(1-(1-\lambda\sigma^2)^{2^k})}{\sigma^2},$$

resultando em

$$\begin{aligned}\frac{1}{\lambda}\Psi^{(k)'}(\sigma,\lambda)&=2^{k+1}(1-\lambda\sigma^2)^{2^k-1}-\frac{1-(1-\lambda\sigma^2)^{2^k}}{\lambda\sigma^2}\\&=2^{k+1}(1-\lambda\sigma^2)^{2^k-1}-\frac{1-(1-\lambda\sigma^2)^{2^k}}{1-(1-\lambda\sigma^2)}\\&=2^{k+1}y^{2^k-1}-\frac{1-y^{2^k}}{1-y}\\&=(2^{k+1}-1)y^{2^k-1}-(y^{2^k-2}+\cdots+y+1)=:g_k(y),\end{aligned}$$

em que utilizamos da notação $y=1-\lambda\sigma^2$. Veja que g_k é contínua e satisfaz $g_k(0)=-1$ e $g_k(1)=2^k$, de modo que existe pelo menos um ponto em $\zeta^{(k)}\in(0,1)$ tal que $g_k(\zeta^{(k)})=0$. Para simplificar a visualização, daqui em diante considere $z:=\zeta^{(k)}$.

Observe que, como $y = 1 - \lambda\sigma^2$, então $\bar{\sigma}^{(k)} := \sqrt{\frac{1-z}{\lambda}}$ é um ponto crítico para $\Psi^{(k)}$ e, mais ainda, $\bar{\sigma}^{(k)} \in (0, 1/\sqrt{\lambda})$, por $z \in (0,1)$. Veremos agora que z é único. Como z é raiz de g_k, podemos escrever $g_k(y) = (y-z)Q_k(y)$, para Q_k polinômio apropriado. De fato, com algum algebrismo e fazendo uso de que $g_k(z) = 0$, podemos ver que

$$\begin{aligned}\frac{g_k(y)}{y-z} &= (2^{k+1}-1)y^{2^k-2} \\ &+ ((2^{k+1}-1)z - 1)y^{2^k-3} \\ &+ ((2^{k+1}-1)z^2 - z - 1)y^{2^k-4} + \dots \\ &+ ((2^{k+1}-1)z^{2^k-2} - z^{2^k-3} - \cdots - z - 1) =: Q_k(y).\end{aligned}$$

Disto, segue que $Q_k(0) = g_k(0)/(-z) = 1/z > 0$. Agora, provaremos que $Q_k(y)$ é função crescente em $0 < y < 1$, implicando portanto que $g_k(y) = (y-z)Q_k(y)$ se anula apenas em $y = z$. Tome $0 < t < 1$ e $\alpha > 0$ tal que $t + \alpha < 1$. Então

$$\begin{aligned}Q_k(t+\alpha) - Q_k(t) &= (2^{k+1}-1)\left((t+\alpha)^{2^k-2} - t^{2^k-2}\right) \\ &+ ((2^{k+1}-1)z - 1)\left((t+\alpha)^{2^k-3} - t^{2^k-3}\right) \\ &+ ((2^{k+1}-1)z^2 - z - 1)\left((t+\alpha)^{2^k-4} - t^{2^k-4}\right) + \dots \\ &+ ((2^{k+1}-1)z^{2^k-3} - z^{2^k-4} - \cdots - z - 1)\left((t+\alpha) - t\right), \qquad (3.9)\end{aligned}$$

e é claro que temos $((t+\alpha)^\nu - t^\nu) > 0$, para $\nu = 1, 2, \dots, (2^k - 2)$. Por outro lado, como $0 = g(z) = (2^{k+1}-1)z^{2^k-1} - (z^{2^k-2} + z^{2^k-3} + \cdots + z + 1)$, obtemos as relações:

$$\begin{aligned}z^{2^k-2}((2^{k+1}-1)z - 1) &= (z^{2^k-3} + \cdots + z + 1) \\ z^{2^k-3}((2^{k+1}-1)z^2 - z - 1) &= (z^{2^k-4} + \cdots + z + 1) \\ &\vdots \\ z^2((2^{k+1}-1)z^{2^k-3} - z^{2^k-4} - \cdots - z - 1) &= (z+1).\end{aligned}$$

Daí, como $z > 0$ implica que expoentes de z são positivos, segue que o lado direito das igualdades acima também é positivo, do que concluímos

$$\begin{aligned}((2^{k+1}-1)z - 1) &= (z^{2^k-3} + \cdots + z + 1)/z^{2^k-2} > 0 \\ ((2^{k+1}-1)z^2 - z - 1) &= (z^{2^k-4} + \cdots + z + 1)/z^{2^k-3} > 0 \\ &\vdots \\ ((2^{k+1}-1)z^{2^k-3} - z^{2^k-4} - \cdots - z - 1) &= (z+1)/z^2 > 0.\end{aligned}$$

Retornando para (3.9) com os apontamentos acima, asseguramos que $Q_k(t+\alpha) - Q_k(t) > 0$, o que torna Q_k crescente em $(0,1)$ e, de $Q_k(0) > 0$, garante unicidade de z. Mais do que isso, por consequência, $\bar{\sigma}^{(k)}$ também é único. Falta verificarmos que $\bar{\sigma}^{(k)}$ é ponto de máximo. Para isto, de $g_k(y) = (y-z)Q_k(y)$ e $Q_k(y) > 0$, temos

$$g_k(y) > 0, \text{ para } y > z, \text{ e } g_k(y) < 0, \text{ para } y < z.$$

Desta forma, podemos analisar o comportamento de $\Psi^{(k)\prime}$:

$$\sigma < \bar{\sigma}^{(k)} \Rightarrow \sigma < \sqrt{\frac{1-z}{\lambda}} \Rightarrow y = 1 - \lambda\sigma^2 > z \Rightarrow \Psi^{(k)\prime}(\sigma, \lambda) = \lambda g_k(y) > 0 \quad \text{e}$$

$$\sigma > \bar{\sigma}^{(k)} \Rightarrow \sigma > \sqrt{\frac{1-z}{\lambda}} \Rightarrow y = 1 - \lambda\sigma^2 < z \Rightarrow \Psi^{(k)\prime}(\sigma, \lambda) = \lambda g_k(y) < 0, \qquad \square$$

caracterizando que, de fato, $\bar{\sigma}^{(k)}$ é ponto de máximo para $\Psi^{(k)}$, como queríamos demonstrar.

Tendo em mãos este resultado e observando que os valores singulares σ_i de A satisfazem as restrições impostas a σ em (3.7), podemos aplicar (3.8) e extrair um limitante superior para $\Psi^{(k)}$ quando fixamos $\lambda \in (0, 1/\sigma_1^2)$, que assume a forma

$$\max_{1\leq i\leq r} \Psi^{(k)}(\sigma_i, \lambda) \leq \Psi^{(k)}(\bar{\sigma}^{(k)}, \lambda) = \frac{1 - (1 - \lambda(\bar{\sigma}^{(k)})^2)^{2^k}}{\bar{\sigma}^{(k)}} = \sqrt{\lambda}\,\frac{1 - (\zeta^{(k)})^{2^k}}{\sqrt{1 - \zeta^{(k)}}}.$$

Particularizando ao caso do método de Newton, temos que $\lambda = \beta$, escolhido tal que $0 < \beta < \frac{1}{\sigma_1^2}$, e portanto

$$\max_{1\leq i\leq r} \left(\frac{f_i^{(k)}}{\sigma_i}\right) = \max_{1\leq i\leq r} \left(\frac{1 - (1 - \beta\sigma_i^2)^{2^k}}{\sigma_i}\right) \leq \sqrt{\beta}\,\frac{1 - (\zeta^{(k)})^{2^k}}{\sqrt{1 - \zeta^{(k)}}} \leq \frac{1}{\sigma_1}\xi^{(k)}, \quad k \geq 1, \quad (3.10)$$

pois $\beta < \frac{1}{\sigma_1^2}$ e adotamos a notação

$$\xi^{(k)} = \frac{1 - (\zeta^{(k)})^{2^k}}{\sqrt{1 - \zeta^{(k)}}}, \quad k \geq 1.$$

A pergunta pertinente ao momento diz respeito aos valores que $\xi^{(k)}$ pode assumir, uma vez que podem ser calculados previamente devido à independência de $g_k(y)$ do problema em questão. Lembramos que o apresentado aqui busca limitantes para os termos $f_i^{(k)}/\sigma_i$, cuja estimativa ingênua é dada por

$$\max_{1\leq i\leq r} \left(\frac{1 - (1 - \beta\sigma_i^2)^{2^k}}{\sigma_i}\right) \leq \max_{1\leq i\leq r} \left(\frac{1}{\sigma_i}\right) = \frac{1}{\sigma_r}, \quad (3.11)$$

que costuma ser de alta ordem para problemas discretos mal postos pelo decrescimento vertiginoso dos valores singulares para zero [64,65]. Na Tabela 3.1 discriminamos algumas instâncias de $\zeta^{(k)}$ e também $\xi^{(k)}$, mostrando o crescimento modesto deste último ao longo das iterações. Por exemplo, caso o método pare em $k = 15$, temos que $\xi^{(k)} = 115.5227 = \mathcal{O}(10^2)$; cinco iterações adiante, $k = 20$ implica em $\xi^{(k)} = 653.4890$, ainda tal que $\xi^{(k)} = \mathcal{O}(10^2)$, sendo que resultados numéricos em [11] sugerem parada do algoritmo em até menos iterações devido à convergência quadrática. Ressaltamos portanto que, especialmente no caso de matrizes com número de condição elevado, é esperado que mesmo com a dependência em k seja válido

$$\frac{1}{\sigma_1}\xi^{(k)} << \frac{1}{\sigma_r}, \quad \text{ou, ainda,} \quad \xi^{(k)} << \kappa(A),$$

tornando a estimativa (3.10) mais acurada que a relação (3.11). De fato, não é raro obtermos $\kappa(A)$ da ordem de 10^{10}, 10^{15} ou ainda maiores [11, Table 2], evidenciando o interesse na análise proposta aqui.

Tabela 3.1 – Alguns exemplos do valor de $\zeta^{(k)}$ e $\xi^{(k)}$ para diferentes iterações k do método de Newton. Raízes $\zeta^{(k)}$ calculadas numericamente via Matlab com precisão de 10^{-6}.

k	$\zeta^{(k)}$	$\xi^{(k)}$	k	$\zeta^{(k)}$	$\xi^{(k)}$	k	$\zeta^{(k)}$	$\xi^{(k)}$	k	$\zeta^{(k)}$	$\xi^{(k)}$
1	0.3333	1.0887	6	0.9803	5.1306	11	0.9994	28.8848	16	0.9999	163.3730
2	0.6719	1.3900	7	0.9902	7.2379	12	0.9997	40.8462	17	0.9999	231.0437
3	0.8392	1.8803	8	0.9951	10.2233	13	0.9998	57.7630	18	0.9999	326.7448
4	0.9205	2.6043	9	0.9975	14.4491	14	0.9999	81.6877	19	0.9999	462.0867
5	0.9605	3.6460	10	0.9988	20.4278	15	0.9999	115.5227	20	0.9999	653.4890

Fonte – o autor, 2022.

3.2.1 Estimativa com condição de fonte Hölder

Agora, de posse do disposto acima, vamos ao teorema que caracteriza o erro absoluto cometido pelo método de Newton quando assumimos condição de fonte Hölder na solução exata. Após, utilizamos esta cota para exibir uma estimativa para o erro relativo.

Teorema 3.3. *Assuma que $b \in \mathcal{R}(A)$ e $x^* \in \mathcal{R}[(A^TA)^\mu]$, para $\mu > 0$. Então, é válido que*

$$\|x^* - \tilde{x}^{(k(\delta))}\|_2 \leq C_H \|w\|_2^{\frac{1}{2\mu+1}} \delta^{\frac{2\mu}{2\mu+1}} = \mathcal{O}\left(\delta^{\frac{2\mu}{2\mu+1}}\right),$$

com $C_H = (\tau+1)^{\frac{2\mu}{2\mu+1}} + \xi^{(k(\delta))}(\tau-1)^{\frac{-1}{2\mu+1}}$ e $w \in \mathbb{R}^n$ tal que $x^* = (A^TA)^\mu w$.

Demonstração. Pela desigualdade triangular,

$$\|x^* - \tilde{x}^{(k(\delta))}\|_2 \leq \|x^* - x^{(k)}\|_2 + \|x^{(k)} - \tilde{x}^{(k(\delta))}\|_2, \tag{3.12}$$

que vamos estimar separadamente. Inicialmente, de $x^* \in \mathcal{R}[(A^TA)^\mu]$, $\exists w \in \mathbb{R}^n$ tal que $x^* = (A^TA)^\mu w$. Assim, de $b \in \mathcal{R}(A)$, segue que

$$b = Ax^* = A(A^TA)^\mu w.$$

Pela decomposição SVD de A, temos que $A = U\Sigma V^T$ e então

$$b = U\Sigma V^T(V\Sigma^T\Sigma V^T)^\mu w = U\Sigma V^T V \begin{bmatrix} C & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}_{n\times n}^{2\mu} V^T w = U \begin{bmatrix} C^{2\mu+1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}_{m\times n} V^T w.$$

Logo,

$$U^T b = \begin{bmatrix} C^{2\mu+1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}_{m\times n} V^T w \quad \Rightarrow \quad \|w\|_2^2 = \sum_{i=1}^{r} \frac{(u_i^T b)^2}{\sigma_i^{2(2\mu+1)}}. \tag{3.13}$$

Observe que a primeira parte da cota em (3.12) satisfaz

$$\|x^* - x^{(k)}\|_2^2 = \left\| \sum_{i=1}^{r} \frac{u_i^T b}{\sigma_i} v_i - \sum_{i=1}^{r} f_i^{(k)} \frac{u_i^T b}{\sigma_i} v_i \right\|_2^2 = \sum_{i=1}^{r} (1 - f_i^{(k)})^2 \frac{(u_i^T b)^2}{\sigma_i^2} = \sum_{i=1}^{r} (1 - f_i^{(k)})^2 \frac{a_i^2}{\sigma_i^2},$$

definindo $a_i := u_i^T b$, para todo i. Agora, lembrando da desigualdade de Hölder, isto é,

$$\sum \alpha_i\beta_i \leq \left(\sum \alpha_i^p\right)^{\frac{1}{p}} \left(\sum \beta_i^q\right)^{\frac{1}{q}}, \quad \text{em que} \quad \frac{1}{p}+\frac{1}{q}=1 \text{ e } \alpha_i,\beta_i>0,$$

tomamos $p = 2\mu+1$ e $q = \frac{2\mu+1}{2\mu}$ e separamos o produto:

$$(1-f_i^{(k)})^2\frac{a_i^2}{\sigma_i^2} = \left((1-f_i^{(k)})^{\frac{2}{2\mu+1}}\frac{1}{\sigma_i^2}a_i^{\frac{2}{2\mu+1}}\right)\left((1-f_i^{(k)})^{\frac{4\mu}{2\mu+1}}a_i^{\frac{4\mu}{2\mu+1}}\right),$$

ou seja, neste caso,

$$\alpha_i = (1-f_i^{(k)})^{\frac{2}{2\mu+1}}\frac{1}{\sigma_i^2}a_i^{\frac{2}{2\mu+1}} \quad \text{e} \quad \beta_i = (1-f_i^{(k)})^{\frac{4\mu}{2\mu+1}}a_i^{\frac{4\mu}{2\mu+1}}.$$

Desta forma, temos que

$$\sum_{i=1}^{r}(1-f_i^{(k)})^2\frac{a_i^2}{\sigma_i^2} \leq \left(\sum_{i=1}^{r}(1-f_i^{(k)})^2\frac{a_i^2}{\sigma_i^{2(2\mu+1)}}\right)^{\frac{1}{2\mu+1}}\left(\sum_{i=1}^{r}(1-f_i^{(k)})^2a_i^2\right)^{\frac{2\mu}{2\mu+1}}. \tag{3.14}$$

De (3.13) e como $(1-f_i^{(k)})^2 \leq 1$ (veja (3.6)), segue diretamente que o primeiro fator pode ser limitado por

$$\left(\sum_{i=1}^{r}(1-f_i^{(k)})^2\frac{a_i^2}{\sigma_i^{2(2\mu+1)}}\right)^{\frac{1}{2\mu+1}} \leq \|w\|_2^{\frac{2}{2\mu+1}}. \tag{3.15}$$

Por outro lado, através da SVD de A, temos:

$$\begin{aligned}
r^{(k)} &= b - Ax^{(k)} \\
&= b - U\Sigma V^T\left(\sum_{i=1}^{r} f_i^{(k)}\frac{u_i^Tb}{\sigma_i}v_i\right) \\
&= b - U\Sigma\left[f_1^{(k)}\frac{u_1^Tb}{\sigma_1}, f_2^{(k)}\frac{u_2^Tb}{\sigma_2}, \ldots, f_r^{(k)}\frac{u_r^Tb}{\sigma_r}, 0, \ldots, 0\right]_{1\times n}^T \\
&= U\left(U^Tb - \left[f_1^{(k)}u_1^Tb, f_2^{(k)}u_2^Tb, \ldots, f_r^{(k)}u_r^Tb, 0, \ldots, 0\right]_{1\times m}^T\right),
\end{aligned}$$

de modo que

$$\|r^{(k)}\|_2^2 = \sum_{i=1}^{r}(1-f_i^{(k)})^2(u_i^Tb)^2 + \sum_{i=r+1}^{m}(u_i^Tb)^2. \tag{3.16}$$

Disto e lembrando que $b \in \mathcal{R}(A)$ implica que $u_i^Tb = 0$ para todo $i = r+1, \ldots, m$, temos

$$\sum_{i=1}^{r}(1-f_i^{(k)})^2a_i^2 = \sum_{i=1}^{r}(1-f_i^{(k)})^2a_i^2 + \sum_{i=r+1}^{m}a_i^2 = \|r^{(k)}\|_2^2. \tag{3.17}$$

Mas, referenciando o desenvolvimento prévio a (3.16), segue que

$$\begin{aligned}
r^{(k)} &= U\left(U^Tb - \left[f_1^{(k)}u_1^Tb, \ldots, f_r^{(k)}u_r^Tb, 0, \ldots, 0\right]_{1\times m}^T\right) \\
&= U\left(U^T(\tilde{b}-e) - \left[f_1^{(k)}u_1^T(\tilde{b}-e), \ldots, f_r^{(k)}u_r^T(\tilde{b}-e), 0, \ldots, 0\right]_{1\times m}^T\right) \\
&= \tilde{r}^{(k)} - U\left(U^Te - \left[f_1^{(k)}u_1^Te, \ldots, f_r^{(k)}u_r^Te, 0, \ldots, 0\right]_{1\times m}^T\right),
\end{aligned} \tag{3.18}$$

o que implica, através da desigualdade triangular, que

$$\begin{aligned}\|r^{(k)}\|_2 &\leq \|\tilde{r}^{(k)}\|_2 + \left(\sum_{i=1}^{r}(1-f_i^{(k)})^2(u_i^T e)^2 + \sum_{i=r+1}^{m}(u_i^T e)^2\right)^{1/2} \\ &\leq \|\tilde{r}^{(k)}\|_2 + \left(\sum_{i=1}^{m}(u_i^T e)^2\right)^{1/2} \\ &= \|\tilde{r}^{(k)}\|_2 + \|e\|_2. \end{aligned} \tag{3.19}$$

Tomando $k = k(\delta)$, segue portanto da primeira desigualdade em (2.4) que

$$\|r^{(k)}\|_2 \leq \tau\delta + \delta = (1+\tau)\delta. \tag{3.20}$$

Aplicando isto em (3.17) e juntando com (3.15) em (3.14), temos um limitante para o primeiro termo:

$$\|x^* - x^{(k)}\|_2 \leq (\tau+1)^{\frac{2\mu}{2\mu+1}}\|w\|_2^{\frac{1}{2\mu+1}}\delta^{\frac{2\mu}{2\mu+1}}. \tag{3.21}$$

Agora, para a segunda parte da cota (3.12), isto é, $\|\tilde{x}^{(k(\delta))} - x^{(k)}\|_2$, veja que

$$\begin{aligned}\|x^{(k)} - \tilde{x}^{(k(\delta))}\|_2^2 &= \sum_{i=1}^{r}\left(f_i^{(k)}\frac{u_i^T(b-\tilde{b})}{\sigma_i}\right)^2 \\ &\leq \max_{1\leq i\leq r}\left(\frac{1-(1-\beta\sigma_i^2)^{2^k}}{\sigma_i}\right)^2 \sum_{i=1}^{r}(u_i^T e)^2 \\ &\leq \frac{1}{\sigma_1^2}(\xi^{(k)})^2\|e\|_2^2, \end{aligned} \tag{3.22}$$

tomando $k = k(\delta)$ e utilizando (3.10) no último passo. Veja que temos aqui a inclusão direta do resultado da Proposição 3.1, importante para mantermos maior controle nas constantes da estimativa, como abordamos abaixo nos comentários após este teorema. Da segunda desigualdade em (2.4), temos $\tau\|e\|_2 \leq \|\tilde{r}^{(k(\delta)-1)}\|_2$, mas, por desenvolvimento análogo a (3.18), que culminou em (3.19), é fácil ver que $\|\tilde{r}^{(k)}\|_2 \leq \|r^{(k)}\|_2 + \|e\|_2$. Portanto, juntando ambas, temos

$$\tau\delta \leq \|\tilde{r}^{(k(\delta)-1)}\|_2 \leq \|r^{(k(\delta)-1)}\|_2 + \|e\|_2 = \|r^{(k(\delta)-1)}\|_2 + \delta \quad \Rightarrow \quad (\tau-1)\delta \leq \|r^{(k(\delta)-1)}\|_2.$$

Logo,

$$\begin{aligned}(\tau-1)^2\delta^2 \leq \|r^{(k(\delta)-1)}\|_2^2 &= \sum_{i=1}^{r}(1-f_i^{(k(\delta)-1)})^2(u_i^T b)^2 + \sum_{i=r+1}^{m}(u_i^T b)^2 \\ &\leq \sum_{i=1}^{r}(u_i^T b)^2 \qquad (\text{pois } b\in\mathcal{R}(A) \text{ e } |1-f_i^{(k)}|<1) \\ &\leq \sigma_1^{2(2\mu+1)}\sum_{i=1}^{r}\frac{(u_i^T b)^2}{\sigma_i^{2(2\mu+1)}} \\ &= \sigma_1^{2(2\mu+1)}\|w\|_2^2. \end{aligned}$$

Concluímos, portanto, que

$$\frac{1}{\sigma_1^2} \leq (\tau-1)^{\frac{-2}{2\mu+1}} \delta^{\frac{-2}{2\mu+1}} \|w\|_2^{\frac{2}{2\mu+1}}. \tag{3.23}$$

Introduzindo (3.23) em (3.22),

$$\begin{aligned}\|x^{(k)} - \tilde{x}^{(k(\delta))}\|_2^2 &\leq (\xi^{(k(\delta))})^2 (\tau-1)^{\frac{-2}{2\mu+1}} \delta^2 \delta^{\frac{-2}{2\mu+1}} \|w\|_2^{\frac{2}{2\mu+1}} \\ &= (\xi^{(k(\delta))})^2 (\tau-1)^{\frac{-2}{2\mu+1}} \|w\|_2^{\frac{2}{2\mu+1}} \delta^{\frac{4\mu}{2\mu+1}},\end{aligned}$$

De modo que o segundo termo em (3.12) se torna

$$\|x^{(k)} - \tilde{x}^{(k(\delta))}\|_2 \leq \xi^{(k(\delta))} (\tau-1)^{\frac{-1}{2\mu+1}} \|w\|_2^{\frac{1}{2\mu+1}} \delta^{\frac{2\mu}{2\mu+1}}. \tag{3.24}$$

Finalmente, de (3.21) e (3.24), segue que

$$\|x^* - \tilde{x}^{(k(\delta))}\|_2 \leq C_H \|w\|_2^{\frac{1}{2\mu+1}} \delta^{\frac{2\mu}{2\mu+1}}, \tag{3.25}$$

em que $C_H = (\tau+1)^{\frac{2\mu}{2\mu+1}} + \xi^{(k(\delta))}(\tau-1)^{\frac{-1}{2\mu+1}}$. Deste modo, obtemos

$$\|x^* - \tilde{x}^{(k(\delta))}\|_2 = \mathcal{O}\left(\delta^{\frac{2\mu}{2\mu+1}}\right). \qquad \square$$

Observação 3.2. Sem utilizarmos da Proposição 3.1 na demonstração acima, a constante associada à estimativa seria, ao final, de ordem mais alta. De fato, a abordagem inicial sugere o emprego da estimativa ingênua (3.11) em (3.22), que conduz a $C_H = (\tau+1)^{\frac{2\mu}{2\mu+1}} + \kappa(A)(\tau-1)^{\frac{-1}{2\mu+1}}$. Ou seja, trocaríamos o termo $\xi^{(k(\delta))}$ por $\kappa(A)$, indesejável na busca por estimativas mais finas.

De posse deste teorema, podemos rapidamente obter uma estimativa de erro relativo, como apresentada abaixo. Um comentário que tecemos diz respeito à constante presente em tal estimativa. Veja que, sem a Proposição 3.1, teríamos aqui $C_H\kappa(A) \approx \kappa(A)^2$, o que elevaria significativamente a cota (3.26), especialmente em problemas mal postos.

Corolário 3.1. *Nas mesmas condições do Teorema 3.3, vale a estimativa relativa*

$$\frac{\|x^* - \tilde{x}^{(k(\delta))}\|_2}{\|x^*\|_2} \leq C_H \kappa(A) \left(\frac{\delta}{\|b\|_2}\right)^{\frac{2\mu}{2\mu+1}} = \mathcal{O}\left(\delta^{\frac{2\mu}{2\mu+1}}\right), \tag{3.26}$$

em que $C_H = (\tau+1)^{\frac{2\mu}{2\mu+1}} + \xi^{(k(\delta))}(\tau-1)^{\frac{-1}{2\mu+1}}$.

Demonstração. Basta lembrar que

$$\|x^*\|_2 \geq \frac{\|b\|_2}{\sigma_1} \quad \text{e} \quad \|w\|_2^{\frac{1}{2\mu+1}} \leq \frac{1}{\sigma_r} \|b\|_2^{\frac{1}{2\mu+1}},$$

e utilizar o Teorema 3.3. $\square$

3.2.2 Estimativa de erro relativo dependente dos coeficientes de Fourier

A estimativa do teorema acima talvez possa ser melhorada no sentido de apresentar uma constante associada menor. O procedimento acompanha as equações apresentadas na demonstração do Teorema 3.3 e essencialmente buscamos efetuar a troca de w pelos coeficientes de Fourier [65] que o representam, isto é, as quantidades $u_i^T b$. Novamente, iniciamos com a estimativa para o erro absoluto e, em seguida, concluímos considerando o caso relativo.

Teorema 3.4 (Est. abs. com coef. de Fourier). *Assuma que $b \in \mathcal{R}(A)$ e, para $\mu > 0$, $x^* \in \mathcal{R}[(A^T A)^\mu]$. Então, é válido que*

$$\|x^* - \tilde{x}^{(k(\delta))}\|_2 \leq C_H \left(\sum_{i=1}^{r} (1 - f_i^{(k(\delta)-1)})^2 \frac{(u_i^T b)^2}{\sigma_i^{2(2\mu+1)}} \right)^{\frac{1}{2(2\mu+1)}} \delta^{\frac{2\mu}{2\mu+1}} = \mathcal{O}\left(\delta^{\frac{2\mu}{2\mu+1}}\right),$$

com $C_H = (\tau+1)^{\frac{2\mu}{2\mu+1}} + \xi^{(k(\delta))}(\tau-1)^{\frac{-1}{2\mu+1}}$ e $w \in \mathbb{R}^n$ tal que $x^* = (A^T A)^\mu w$.

Demonstração. Assumindo as mesmas condições de antes, iniciamos com

$$\|x^* - \tilde{x}^{(k(\delta))}\|_2 \leq \|x^* - x^{(k)}\|_2 + \|x^{(k)} - \tilde{x}^{(k(\delta))}\|_2, \tag{3.27}$$

novamente lidando com cada termo da cota separadamente. Da demonstração do teorema anterior, mais especificamente de (3.14) e (3.20), segue que

$$\begin{aligned} \|x^* - x^{(k)}\|_2^2 &\leq \left(\sum_{i=1}^{r} (1 - f_i^{(k)})^2 \frac{(u_i^T b)^2}{\sigma_i^{2(2\mu+1)}} \right)^{\frac{1}{2\mu+1}} \left(\sum_{i=1}^{r} (1 - f_i^{(k)})^2 (u_i^T b)^2 \right)^{\frac{2\mu}{2\mu+1}} \\ &\leq \left(\sum_{i=1}^{r} (1 - f_i^{(k)})^2 \frac{(u_i^T b)^2}{\sigma_i^{2(2\mu+1)}} \right)^{\frac{1}{2\mu+1}} (1+\tau)^{\frac{4\mu}{2\mu+1}} \delta^{\frac{4\mu}{2\mu+1}}, \end{aligned}$$

resultando em

$$\|x^* - x^{(k)}\|_2^2 \leq \left(\sum_{i=1}^{r} (1 - f_i^{(k)})^2 \frac{(u_i^T b)^2}{\sigma_i^{2(2\mu+1)}} \right)^{\frac{1}{2(2\mu+1)}} (1+\tau)^{\frac{2\mu}{2\mu+1}} \delta^{\frac{2\mu}{2\mu+1}}. \tag{3.28}$$

Por outro lado, para a segunda parte da estimativa, de (3.22) e do raciocínio seguinte:

$$\|x^{(k)} - \tilde{x}^{(k(\delta))}\|_2^2 \leq \frac{1}{\sigma_1^2} (\xi^{(k(\delta))})^2 \|e\|_2^2 \qquad \text{e} \qquad (\tau - 1) \leq \|r^{(k(\delta)-1)}\|_2. \tag{3.29}$$

Assim,

$$\begin{aligned}(\tau-1)^2\delta^2 &\leq \|r^{(k(\delta)-1)}\|_2^2\\ &= \sum_{i=1}^{r}(1-f_i^{(k(\delta)-1)})^2(u_i^Tb)^2+\sum_{i=r+1}^{m}(u_i^Tb)^2\\ &= \sum_{i=1}^{r}(1-f_i^{(k(\delta)-1)})^2(u_i^Tb)^2\\ &\leq \sigma_1^{2(2\mu+1)}\sum_{i=1}^{r}(1-f_i^{(k(\delta)-1)})^2\frac{(u_i^Tb)^2}{\sigma_i^{2(2\mu+1)}}.\end{aligned}$$

Logo,

$$\frac{1}{\sigma_1^2}\leq\left(\sum_{i=1}^{r}(1-f_i^{(k(\delta)-1)})^2\frac{(u_i^Tb)^2}{\sigma_i^{2(2\mu+1)}}\right)^{\frac{1}{2\mu+1}}(\tau-1)^{\frac{-2}{2\mu+1}}\delta^{\frac{-2}{2\mu+1}},$$

que introduzida na primeira desigualdade em (3.29) conduz a

$$\|x^{(k)}-\tilde{x}^{(k(\delta))}\|_2\leq\left(\sum_{i=1}^{r}(1-f_i^{(k(\delta)-1)})^2\frac{(u_i^Tb)^2}{\sigma_i^{2(2\mu+1)}}\right)^{\frac{1}{2(2\mu+1)}}\xi^{(k(\delta))}(\tau-1)^{\frac{-1}{2\mu+1}}\delta^{\frac{2\mu}{2\mu+1}}. \tag{3.30}$$

Desta forma, tomando $k=k(\delta)$ e usando as estimativas obtidas em (3.28) e (3.30), temos

$$\|x^*-\tilde{x}^{(k(\delta))}\|_2\leq C_H\left(\sum_{i=1}^{r}(1-f_i^{(k(\delta)-1)})^2\frac{(u_i^Tb)^2}{\sigma_i^{2(2\mu+1)}}\right)^{\frac{1}{2(2\mu+1)}}\delta^{\frac{2\mu}{2\mu+1}}, \tag{3.31}$$

com $C_H=(\tau+1)^{\frac{2\mu}{2\mu+1}}+\xi^{(k(\delta))}(\tau-1)^{\frac{-1}{2\mu+1}}$ (a mesma obtida no Teorema 3.3). □

A conclusão imediata que sai deste resultado é que esta estimativa absoluta é mais fina que (3.25), uma vez que temos

$$\left(\sum_{i=1}^{r}(1-f_i^{(k)})^2\frac{|u_i^Tb|^2}{\sigma_i^{2(2\mu+1)}}\right)^{\frac{1}{2\mu+1}}\leq\|w\|_2^{\frac{2}{2\mu+1}}.$$

O teorema busca explicitar os termos envolvidos na construção do vetor w, o que permite melhorar significativamente o resultado obtido. Agora, como fizemos na seção anterior, pensando em tomar erros relativos a partir de (3.31), temos a conclusão no corolário abaixo.

Corolário 3.2. *Nas mesmas condições do Teorema 3.4, vale a estimativa relativa*

$$\frac{\|x^*-\tilde{x}^{(k(\delta))}\|_2}{\|x^*\|_2}\leq C_H\,\sigma_1\left(\frac{1}{\|b\|_2^2}\sum_{i=1}^{r}(1-f_i^{(k(\delta)-1)})^2\frac{|u_i^Tb|^2}{\sigma_i^{2(2\mu+1)}}\right)^{\frac{1}{2(2\mu+1)}}\left(\frac{\delta}{\|b\|_2}\right)^{\frac{2\mu}{2\mu+1}}=\mathcal{O}\left(\delta^{\frac{2\mu}{2\mu+1}}\right). \tag{3.32}$$

Demonstração. Observe que $\|b\|_2 = \|Ax^*\|_2 \leq \sigma_1\|x^*\|_2$, de modo que (3.31) conduz ao resultado:

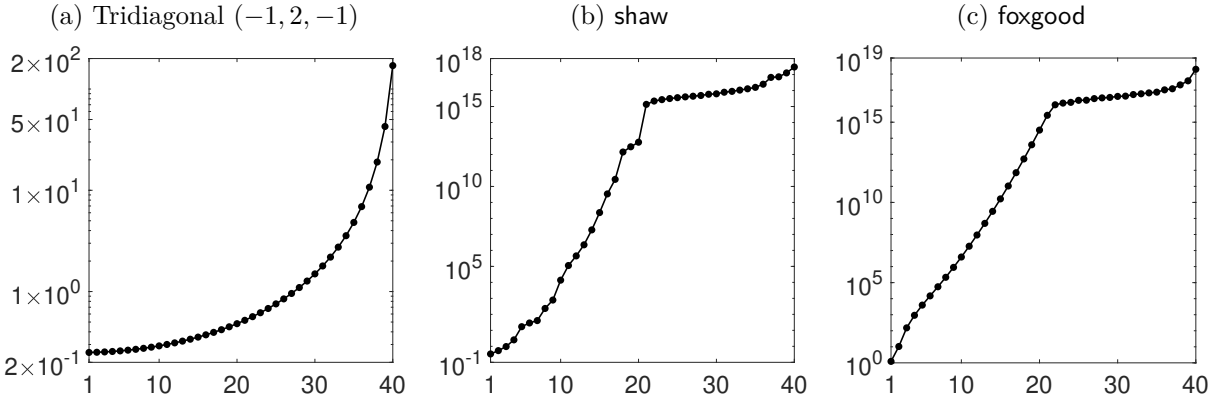
$$\begin{aligned}\frac{\|x^* - \tilde{x}^{(k(\delta))}\|_2}{\|x^*\|_2} &\leq \frac{C_H}{\|x^*\|_2}\left(\sum_{i=1}^{r}(1-f_i^{(k(\delta)-1)})^2\frac{|u_i^Tb|^2}{\sigma_i^{2(2\mu+1)}}\right)^{\frac{1}{2(2\mu+1)}}\delta^{\frac{2\mu}{2\mu+1}}\\ &\leq \frac{C_H\sigma_1}{\|b\|_2}\left(\sum_{i=1}^{r}(1-f_i^{(k(\delta)-1)})^2\frac{|u_i^Tb|^2}{\sigma_i^{2(2\mu+1)}}\right)^{\frac{1}{2(2\mu+1)}}\delta^{\frac{2\mu}{2\mu+1}}\\ &= C_H\sigma_1\left(\frac{1}{\|b\|_2^2}\sum_{i=1}^{r}(1-f_i^{(k(\delta)-1)})^2\frac{|u_i^Tb|^2}{\sigma_i^{2(2\mu+1)}}\right)^{\frac{1}{2(2\mu+1)}}\left(\frac{\delta}{\|b\|_2}\right)^{\frac{2\mu}{2\mu+1}}. \qquad \square\end{aligned}$$

Observe que esta estimativa difere de (3.26) pela não presença do termo $\kappa(A)$. A diferença central diz respeito à "troca" de

$$\frac{1}{\sigma_r} \quad \text{por} \quad \left(\frac{1}{\|b\|_2^2}\sum_{i=1}^{r}(1-f_i^{(k(\delta)-1)})^2\frac{|u_i^Tb|^2}{\sigma_i^{2(2\mu+1)}}\right)^{\frac{1}{2(2\mu+1)}}$$

na nova estimativa (3.32), o que traz benefícios concretos devido à alta ordem que $1/\sigma_r$ pode atingir em problemas mal postos. Na Figura 3.1 temos alguns exemplos que ilustram o crescimento das quantidades $1/\sigma_i$ em matrizes de dimensão $n = 40$. Ali, tridiagonal $(-1, 2, -1)$ corresponde à matriz em banda com 2 na diagonal principal e -1 na primeira subdiagonal superior e inferior e surge naturalmente como resultado da aplicação do método de diferenças finitas centrais no operador diferencial de segunda ordem. Esta matriz é considerada mal condicionada, embora $\kappa(A) \approx 10^3$, bastante inferior a shaw e foxgood, com números de condição da ordem de 8×10^{17} e 10^{18}, respectivamente. Estes últimos fazem parte dos problemas severamente mal condicionados, cujo decrescimento dos valores singulares ocorre de forma vertiginosa, levando ao comportamento visto na figura. Mesmo assim, operar com todos os recíprocos $1/\sigma_i$ em (3.31) e (3.32) ao invés de

Figura 3.1 – Recíprocos dos valores singulares para alguns problemas discretos mal postos, $n = 40$.



Fonte – o autor, 2022.

tomar apenas $1/\sigma_r$ tende a reduzir o valor das constantes, especialmente pelos primeiros recíprocos serem, em ordem, significantemente inferiores aos finais.

3.3 ESTIMATIVA SEM CONDIÇÃO DE FONTE ADICIONAL

Finalizamos esta seção com o próximo teorema, que é um avanço direto do apresentado em [11], uma vez que se utiliza das mesmas hipóteses. A diferença essencial entre este teorema e os imediatamente anteriores diz respeito à não utilização de condições de fonte na solução exata.

Teorema 3.5. *Assuma que $b \in \mathcal{R}(A)$. Então, é válido que*

$$\frac{\|x^* - \tilde{x}^{(k(\delta))}\|_2}{\|x^*\|_2} \leq C_\tau \kappa(A) \frac{\delta}{\|b\|_2} = \mathcal{O}(\delta), \tag{3.33}$$

com $C_\tau = 2 + \tau$.

Demonstração. Estimamos a quantidade procurada através da desigualdade triangular, resultando em

$$\|x^* - \tilde{x}^{(k(\delta))}\|_2 \leq \|x^* - x^{(k)}\|_2 + \|x^{(k)} - \tilde{x}^{(k(\delta))}\|_2, \tag{3.34}$$

de modo que cada parte da cota será tratada separadamente. Por simplicidade na abordagem, iniciemos com a segunda parte em (3.34). Assim, veja que, de (3.6),

$$\begin{aligned} \|x^{(k)} - \tilde{x}^{(k(\delta))}\|_2^2 &= \sum_{i=1}^{r} \frac{(f_i^{(k)})^2}{\sigma_i^2} (u_i^T(b - \tilde{b}))^2 \\ &\leq \frac{1}{\sigma_r^2} \sum_{i=1}^{r} (u_i^T(b - \tilde{b}))^2 \\ &\leq \frac{1}{\sigma_r^2} \|b - \tilde{b}\|_2^2. \end{aligned} \tag{3.35}$$

Lembrando que $\tilde{b} = b + e$, segue que

$$\|x^{(k)} - \tilde{x}^{(k(\delta))}\|_2 \leq \frac{1}{\sigma_r} \|e\|_2 = \frac{1}{\sigma_r} \delta. \tag{3.36}$$

A primeira parte da cota em (3.34) demanda mais cautela. Como $b = \tilde{b} - e$, temos

$$\begin{aligned} x^* - x^{(k)} &= \sum_{i=1}^{r} \frac{u_i^T b}{\sigma_i} v_i - \sum_{i=1}^{r} f_i^{(k)} \frac{u_i^T b}{\sigma_i} v_i \\ &= \sum_{i=1}^{r} (1 - f_i^{(k)}) \frac{u_i^T \tilde{b}}{\sigma_i} v_i - \sum_{i=1}^{r} (1 - f_i^{(k)}) \frac{u_i^T e}{\sigma_i} v_i \\ &= w_1 - w_2, \end{aligned}$$

em que

$$w_1 = \sum_{i=1}^{r} (1 - f_i^{(k)}) \frac{u_i^T \tilde{b}}{\sigma_i} v_i \quad \text{e} \quad w_2 = \sum_{i=1}^{r} (1 - f_i^{(k)}) \frac{u_i^T e}{\sigma_i} v_i.$$

Logo, através da desigualdade triangular,

$$\|x^* - x^{(k)}\|_2^2 \leq (\|w_1\|_2 + \|w_2\|_2)^2. \tag{3.37}$$

Assim, estimamos $\|w_1\|_2$ e $\|w_2\|_2$ separadamente. Para a primeira, da equação (3.16), temos

$$\begin{aligned}
\|w_1\|_2^2 &= \sum_{i=1}^{r}(1 - f_i^{(k)})^2 \frac{(u_i^T \tilde{b})^2}{\sigma_i^2} \\
&\leq \frac{1}{\sigma_r^2}\sum_{i=1}^{r}(1 - f_i^{(k)})^2 (u_i^T \tilde{b})^2 \\
&\leq \frac{1}{\sigma_r^2}\left(\sum_{i=1}^{r}(1 - f_i^{(k)})^2 (u_i^T \tilde{b})^2 + \sum_{i=r+1}^{m}(u_i^T \tilde{b})^2\right) \\
&= \frac{1}{\sigma_r^2}\|\tilde{r}^{(k)}\|_2^2 \\
&\leq \frac{1}{\sigma_r^2}\tau^2\delta^2,
\end{aligned}$$

em que a última desigualdade vem de DP (2.4). Portanto,

$$\|w_1\|_2 \leq \frac{1}{\sigma_r}\tau\delta. \tag{3.38}$$

Similarmente, lembrando que (3.6) implica que $(1 - f_i^{(k)})^2 < 1$, temos

$$\|w_2\|_2^2 = \sum_{i=1}^{r}(1 - f_i^{(k)})^2 \frac{(u_i^T e)^2}{\sigma_i^2} \leq \frac{1}{\sigma_r^2}\sum_{i=1}^{r}(u_i^T e)^2 \leq \frac{1}{\sigma_r^2}\|e\|_2^2 = \frac{1}{\sigma_r^2}\delta^2, \tag{3.39}$$

de modo que

$$\|w_2\|_2 \leq \frac{1}{\sigma_r}\delta. \tag{3.40}$$

Finalmente, aplicando (3.38) e (3.40) em (3.37), segue que

$$\|x^* - x^{(k)}\|_2 \leq \frac{1}{\sigma_r}(1 + \tau)\delta. \tag{3.41}$$

Assim, combinando as equações (3.36) e (3.41) novamente em (3.34), conseguimos

$$\|x^* - \tilde{x}^{(k(\delta))}\|_2 \leq \frac{1}{\sigma_r}(2 + \tau)\delta.$$

Lembrando que $\|b\|_2 \leq \sigma_1\|x^*\|_2$, a estimativa acima se torna

$$\frac{\|x^* - \tilde{x}^{(k(\delta))}\|_2}{\|x^*\|_2} \leq \frac{\sigma_1}{\sigma_r}(2 + \tau)\frac{\delta}{\|b\|_2} = C_\tau \kappa(A)\frac{\delta}{\|b\|_2},$$

em que $C_\tau = 2 + \tau$. □

O procedimento elaborado na Proposição 3.1 e adiante conduz a melhorias na estimativa proposta no Teorema 3.5. De fato, veja que (3.35) e (3.39) se tornam, respectivamente,

$$\|x^{(k)} - \tilde{x}^{(k(\delta))}\|_2 \leq \frac{1}{\sigma_1}\xi^{(k(\delta))}\delta \quad \text{e} \quad \|w_2\|_2 \leq \frac{1}{\sigma_1}\xi^{(k(\delta))}\delta.$$

Isto, juntamente com (3.37) e (3.38), implica em

$$\begin{aligned}\|x^* - \tilde{x}^{(k(\delta))}\|_2 &\leq \|x^* - x^{(k)}\|_2 + \|x^{(k)} - \tilde{x}^{(k(\delta))}\|_2 \\ &\leq \|w_1\|_2 + \|w_2\|_2 + \|x^{(k)} - \tilde{x}^{(k(\delta))}\|_2 \\ &\leq \frac{1}{\sigma_r}\tau\delta + 2\frac{1}{\sigma_1}\xi^{(k(\delta))}\delta \\ &= \frac{1}{\sigma_r}\left(\tau + 2\frac{1}{\kappa(A)}\xi^{(k(\delta))}\right)\delta.\end{aligned}$$

Portanto, usando $\|b\|_2 \leq \sigma_1\|x^*\|_2$, concluímos com a estimativa relativa, que toma a forma

$$\frac{\|x^* - \tilde{x}^{(k(\delta))}\|_2}{\|x^*\|_2} \leq \widetilde{C}_\tau\,\kappa(A)\frac{\delta}{\|b\|_2}, \tag{3.42}$$

em que

$$\widetilde{C}_\tau = \left(\tau + 2\frac{1}{\kappa(A)}\xi^{(k(\delta))}\right).$$

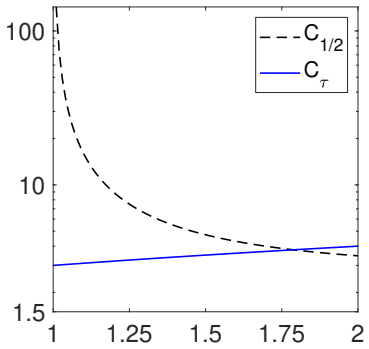
Uma observação importante a ser feita aqui é que, caso $\kappa(A) >> 1$ (como é comum em problemas mal postos), temos que $\widetilde{C}_\tau \approx \tau < (2+\tau) = C_\tau$, esta última constante oriunda do Teorema 3.5. Por outro lado, é claro que se $\kappa(A)$ é pequeno e dependendo do número de iterações até a parada do algoritmo (que faz crescer o valor de $\xi^{(k(\delta))}$ como visto na Tabela 3.1), podemos ter que $\widetilde{C}_\tau > C_\tau$, embora, provavelmente, raro.

3.4 COMENTÁRIOS E IMPLICAÇÕES

As estimativas encontradas neste capítulo avançam a qualidade das aproximações calculadas através do método de Newton com parada via princípio da discrepância. Em primeira instância, podemos mencionar o resultado que melhora diretamente a estimativa em Bazán e Boos [11], obtida no Teorema 3.5 e na análise subsequente. Ali, temos um resultado que leva o erro relativo à ordem do ruído nos dados, significativamente superior à $\mathcal{O}(\delta^{1/2})$. Mais ainda, na constante que acompanha (3.33), veja que $C_\tau = 2 + \tau$ é comparável e inferior a $C_{1/2}$ para diversos valores de τ, como é o caso, por exemplo, de $\tau \in (1, 1.7)$, como pode ser visto na Figura 3.2. Este efeito pode também ser esperado para o melhoramento em (3.42) com a constante $\widetilde{C}_\tau$, especialmente se pensamos em problemas mal postos em que $\widetilde{C}_\tau \approx \tau < C_\tau$. Portanto, o desenvolvimento na Seção 3.3 melhora tanto em ordem quanto nas constantes associadas ao resultado previamente obtido.

Já para as estimativas com condição de fonte Hölder, a análise necessita de maior cautela, especialmente por conta do cenário em dimensão infinita que motiva, na literatura,

Figura 3.2 – Comparação entre as constantes $C_{1/2}$ e C_τ, em função de τ.



Fonte – o autor, 2022.

o surgimento desse tipo de condição. De fato, pedir que $x^* \in \mathcal{R}([A^TA]^\mu)$ não é exatamente algo trivial de visualizar em geral, como veremos abaixo, após breve contextualização. Considere $K : \mathcal{X} \longrightarrow \mathcal{Y}$ um operador linear compacto entre os espaços de Hilbert $\mathcal{X}$ e $\mathcal{Y}$ com produtos internos $\langle\cdot,\cdot\rangle_\mathcal{X}$ e $\langle\cdot,\cdot\rangle_\mathcal{Y}$, respectivamente. Também, seja $K^* : \mathcal{Y} \longrightarrow \mathcal{X}$ o adjunto de K, definido através de $\langle Kx, y\rangle_\mathcal{Y} = \langle x, K^*y\rangle_\mathcal{X}$, para todo $x \in \mathcal{X}$ e $y \in \mathcal{Y}$. Assim, dizemos que $(\sigma_n; v_n, u_n)$ é *sistema singular associado a* K [46] se:

- $\{\sigma_n^2\}_{n\in\mathbb{N}}$ são os autovalores não nulos do operador K^*K (e também de KK^*), escritos em ordem não crescente.
- $\{v_n\}_{n\in\mathbb{N}} \subseteq \mathcal{X}$ são o sistema ortonormal completo de autovetores de K^*K correspondentes ao autovalores σ_n^2, que geram o espaço $\overline{\mathcal{R}(K^*)}$.
- $\{u_n\}_{n\in\mathbb{N}} \subseteq \mathcal{Y}$ são definidos por $u_n := \frac{Kv_n}{\|Kv_n\|_\mathcal{Y}}$ e formam um sistema ortonormal completo de autovetores de KK^* correspondentes ao autovalores σ_n^2, que geram o espaço $\overline{\mathcal{R}(K)}$.
- Valem as relações: $Kv_n = \sigma_n u_n$, $K^*u_n = \sigma_n v_n$, para todo $n \in \mathbb{N}$, e

$$Kx = \sum_{i=1}^{\infty} \sigma_n \langle x, v_n\rangle_\mathcal{X} u_n, \quad x \in \mathcal{X}, \quad \text{e} \quad K^*y = \sum_{i=1}^{\infty} \sigma_n \langle y, u_n\rangle_\mathcal{X} v_n, \quad y \in \mathcal{Y}, \quad (3.43)$$

  com estas séries convergindo em $\mathcal{X}$ e $\mathcal{Y}$, respectivamente.

A equação (3.43) apresenta o que é conhecido como *expansão em valores singulares* (SVE) [46, 64] e faz o paralelo no contexto de dimensão infinita com o seu análogo proporcionado pela SVD: para $A \in \mathbb{R}^{m\times n}$, sua SVD garante que

$$Ax = \sum_{i=1}^{r} \sigma_i (v_i^T x) u_i, \quad x \in \mathbb{R}^n, \quad \text{e} \quad A^T y = \sum_{i=1}^{r} \sigma_i (u_i^T y) v_i, \quad y \in \mathbb{R}^m,$$

com $(\cdot)^T$ tomando o papel do operador adjunto. Mais ainda, a "construção" do sistema singular associado a K feita acima é uma das formas de mostrar a existência da SVD: a

partir do conjunto de autovalores não nulos σ_i^2 e autovetores v_i de A^TA, formamos $u_i = \frac{Av_i}{\|Av_i\|_2}$, $i = 1, \ldots, r$. Os demais vetores componentes de U e V surgem como completamento de $\{u_1, \ldots, u_r\}$ e $\{v_1, \ldots, v_r\}$ à bases de $\mathbb{R}^m$ e $\mathbb{R}^n$, respectivamente.

Uma condição necessária e suficiente para que $K^\dagger y \in \mathcal{R}([K^*K]^\mu)$ é apresentada na proposição seguinte, como ilustração da condição de fonte Hölder para operadores com imagem aberta. Aqui, $K^\dagger$ representa a *inversa generalizada de Moore-Penrose* em sua versão geral, em paralelo com a matriz pseudo-inversa.

Proposição 3.2 (Engl, Hanke e Neubauer [46]). *Seja $K : \mathcal{X} \longrightarrow \mathcal{Y}$ um operador linear compacto com sistema singular $(\sigma_n; v_n, u_n)$. Então, para $\mu > 0$ e $y \in \mathcal{D}(K^\dagger)$,*

$$K^\dagger y \in \mathcal{R}([K^*K]^\mu)$$

se, e somente se,

$$\sum_{n=1}^{\infty} \frac{|\langle y, u_n\rangle|^2}{\sigma_n^{4\mu+2}} < \infty. \tag{3.44}$$

Vale comentar que $y \in \mathcal{D}(K^\dagger)$ se, e somente se,

$$\sum_{n=1}^{\infty} \frac{|\langle y, u_n\rangle|^2}{\sigma_n^2} < \infty$$

e, neste caso,

$$K^\dagger y = \sum_{n=1}^{\infty} \frac{\langle y, u_n\rangle}{\sigma_n} v_n,$$

estrutura análoga a que usamos, por exemplo, para caracterizar $x^* = A^\dagger b$ em (3.4). Assim, para $K^\dagger y$ satisfazer a condição de Hölder para algum $\mu > 0$, é preciso que (3.44) seja válida, o que pode não ser trivial.

Lembramos que, apesar da notória semelhança entre (3.26), (3.32), (3.33) e (3.42), o desenvolvimento de tais resultados é importante pelas constantes atreladas a cada estimativa, que são, como visto, diferentes. Observe de todas dependem do problema em questão (por exemplo, com o termo $\kappa(A)$ nas constantes), algumas mais intrinsecamente, como é o caso de (3.32) com os coeficientes $u_i^T b$. Esta, de mesma ordem que (3.26), apresenta uma constante associada inferior, que torna a cota mais fina. Especialmente para problemas mal postos em que se espera um decaimento expressivo dos valores singulares, ou seja, em que $1/\sigma_r$ é conduzido a quantidades de alta ordem, esta constante contribui no refinamento das cotas. Estas particularidades permitem analisar e possivelmente estreitar o erro no pior caso em alguma delas, consequentemente fornecendo maior garantia quanto à qualidade das aproximações geradas. Em resumo, obtivemos três novas estimativas de erro para o método de Newton em uma abordagem que, aparentemente, é inédita na literatura. Cada cota, com suas particularidades a parte, infere que as soluções obtidas apresentam erro no pior caso de ordem maior que o resultado que motivou o estudo, ou seja, cotas melhores neste sentido, e mais palpáveis às aplicações.

4 MÉTODO DE LEVENBERG-MARQUARDT COM *SCALING* SINGULAR

Considere o problema de mínimos quadrados não linear

$$\min_{x\in\mathbb{R}^n} \phi(x), \quad \text{com} \quad \phi(x) := \frac{1}{2}\|F(x)\|_2^2, \tag{4.1}$$

em que $F : \mathbb{R}^n \longrightarrow \mathbb{R}^m$ é continuamente diferenciável, e assuma que o problema possui soluções de resíduo nulo, i.e., que o conjunto

$$X^* := \{x \in \mathbb{R}^n \mid F(x) = \mathbf{0}\}$$

é não vazio. O objetivo central deste capítulo consiste em analisar propriedades de convergência da variante do método de Levenberg-Marquardt (LMM) apresentada a seguir. Dado um ponto inicial $x_0 \in \mathbb{R}^n$, definimos as iterações

$$(J_k^T J_k + \lambda_k L^T L)d_k = -J_k^T F_k \quad \text{e} \tag{4.2}$$

$$x_{k+1} = x_k + \alpha_k d_k, \quad \forall k \geq 0, \tag{4.3}$$

em que $F_k := F(x_k)$, $J_k := J(x_k)$ é a matriz Jacobiana de F em x_k, $\{\lambda_k\}$ é uma sequência de escalares positivos (conhecidos como *parâmetros de* damping ou *parâmetros de Levenberg-Marquardt*) e $\{\alpha_k\}$ correspondem ao tamanho do passo em cada iteração. Os termos λ_k e α_k serão melhor abordados em momento apropriado. Quanto à matriz L, fornecemos a seguinte estrutura:

Hipótese 4.1. *A matriz $L \in \mathbb{R}^{p\times n}$ é tal que* posto$(L) = p$*, com $m \geq n \geq p$. Adicionalmente, assumimos que*

$$\mathcal{N}(J_k) \cap \mathcal{N}(L) = \{\mathbf{0}\}, \quad k \geq 0. \tag{4.4}$$

Uma aplicação direta desta hipótese se encontra em garantir uma única solução ao sistema linear (4.2). De fato, veja que $(J_k^T J_k + \lambda_k L^T L) = H_k^T H_k$, em que

$$H_k = \begin{bmatrix} J_k \\ \sqrt{\lambda_k} L \end{bmatrix}.$$

Assim, como $\mathbb{R}^n = \mathcal{N}(L) \oplus \mathcal{N}(L)^\perp$, para $u \in \mathbb{R}^n \backslash \{\mathbf{0}\}$ qualquer, podemos escrever $u = v + w$, com $v \in \mathcal{N}(L)$ e $w \in \mathcal{N}(L)^\perp$. Daí, temos duas situações: para $v \in \mathcal{N}(L)$, então $J_k v \neq \mathbf{0}$, por (4.4); para $w \in \mathcal{N}(L)^\perp$, então $Lw \neq \mathbf{0}$. Em ambos os casos, conseguimos que $H_k u \neq \mathbf{0}$ ou, equivalentemente, $\mathcal{N}(H_k) = \{\mathbf{0}\}$. Segue que $u^T(J_k^T J_k + \lambda_k L^T L)u = u^T H_k^T H_k u > 0$, lembrando que $\lambda_k > 0$. Concluímos portanto que a matriz inversa de $(J_k^T J_k + \lambda_k L^T L)$ existe e, em particular, é também simétrica e definida positiva. Logo, d_k calculada através de (4.2) está bem definida.

Também assumiremos, em uma parte do texto, uma maneira de expressar a hipótese acima de forma mais abrangente, no que é conhecida como *condição de completude* (*completeness condition*, veja Morozov [95, p. 34] e Engl, Hanke e Neubauer [46, p. 197]) descrita como abaixo.

Hipótese 4.2. *Para $\bar{x} \in \mathbb{R}^n$ e $\rho > 0$ apropriados, existe $\gamma > 0$ tal que*

$$\|J(x)x\|_2^2 + \|Lx\|_2^2 \geq \gamma \|x\|_2^2, \quad \forall x \in B(\bar{x}, \rho), \tag{4.5}$$

em que $B(\bar{x}, \rho)$ denota uma bola fechada em $\mathbb{R}^n$ centrada em $\bar{x}$ e raio ρ.

No método LMM clássico (e em outros considerados na literatura), as direções d_k são obtidas de um sistema linear análogo a (4.2) em que, em lugar de $L^T L$ usa-se uma matriz Ω_k (chamada usualmente *matriz de* scaling) não singular. A teoria de convergência e taxas para esse caso é bem conhecida, em geral ressaltando as boas qualidades de convergência quadrática local de LMM [14,18,41,42,82,87,93,122]. Neste trabalho, porém, devido à estrutura de L, o produto $L^T L$ é uma matriz singular, proibindo a aplicação direta dos resultados já conhecidos da teoria clássica para LMM através, por exemplo, da convergência utilizando o método de região de confiança. Tendo em vista esta característica e para referência ao longo do texto, chamaremos de *método de Levenberg-Marquardt com* scaling *singular* (LMMSS) às iterações (4.2)-(4.3).

Além das hipóteses acima, durante todo o trabalho assumiremos a seguinte escolha do parâmetro de *damping* λ_k.

Hipótese 4.3. *A escolha do parâmetro de* damping *é dada por $\lambda_k = \lambda(x_k) := \|F(x_k)\|_2^2$, para todo $k \geq 0$.*

Esta determinação de λ_k é inspirada pelo trabalho de Yamashita e Fukushima [122], que tem forte influência na análise local que empreenderemos nas seções seguintes. É claro que, para LMM clássico, λ_k pode ser tomado de outras formas, a depender do autor e do problema em estudo. Em geral se entende que λ_k deve ser grande de modo a permitir que a função objetivo seja reduzida mais eficientemente no início das iterações mas, ao nos aproximarmos da convergência, este fator deveria ir a zero para que a taxa de convergência quadrática seja recuperada. Isso se dá pois, se $\lambda_k \to 0$, d_k tende à direção de Gauss-Newton (caso as matrizes Jacobianas tenham posto completo), que converge quadraticamente localmente, uma propriedade que queremos preservar [87]. Alguns exemplos de escolhas para λ_k com comentários adicionais o leitor encontra, por exemplo, em Benatti [16, Seção 3.1].

Para melhor apresentarmos as motivações que levaram ao desenvolvimento aqui proposto, nas linhas seguintes tratamos de estudar alguns exemplos simples, os quais também elucidam pontos fortes e fracos desta abordagem. Considere o problema de minimização (4.1) em que $F : \mathbb{R}^2 \longrightarrow \mathbb{R}^3$ é dada por

$$F(x, y) = (x^2, y^2, x + y),$$

de modo que

$$\phi(x, y) = \frac{1}{2}\|F(x, y)\|_2^2 = \frac{1}{2}(x^4 + y^4 + (x + y)^2), \tag{4.6}$$

que possui a origem como única solução, ou seja, o conjunto $X^* = \{x^*\}$, para $x^* = (0, 0)$. No âmbito da Otimização, x^* é neste caso dito ser um minimizador global estrito de ϕ, pois $\phi(x, y) > \phi(x^*)$, para todo $(x, y) \in \mathbb{R}^2 \backslash \{x^*\}$, uma afirmação facilmente verificável pela definição de ϕ. Além disso, a matriz Jacobiana $J(x, y)$ associada a F é

$$J(x, y) = \begin{bmatrix} 2x & 0 \\ 0 & 2y \\ 1 & 1 \end{bmatrix}$$

e, por precisão no texto, tomamos α_k em (4.3) escolhido através do critério de Armijo (o qual não terá papel central neste exemplo e será melhor abordado à frente).

Conduzimos testes comparativos para este problema, sumarizados na Figura 4.1 e na Tabela 4.1, entre a performance da versão clássica do LMM, isto é, com matriz de *scaling* dada por $L^T L = I_2$, e da versão proposta para LMMSS, considerando

$$L = L_1(1) = \begin{bmatrix} -1 & 1 \end{bmatrix} \in \mathbb{R}^{1\times 2},$$

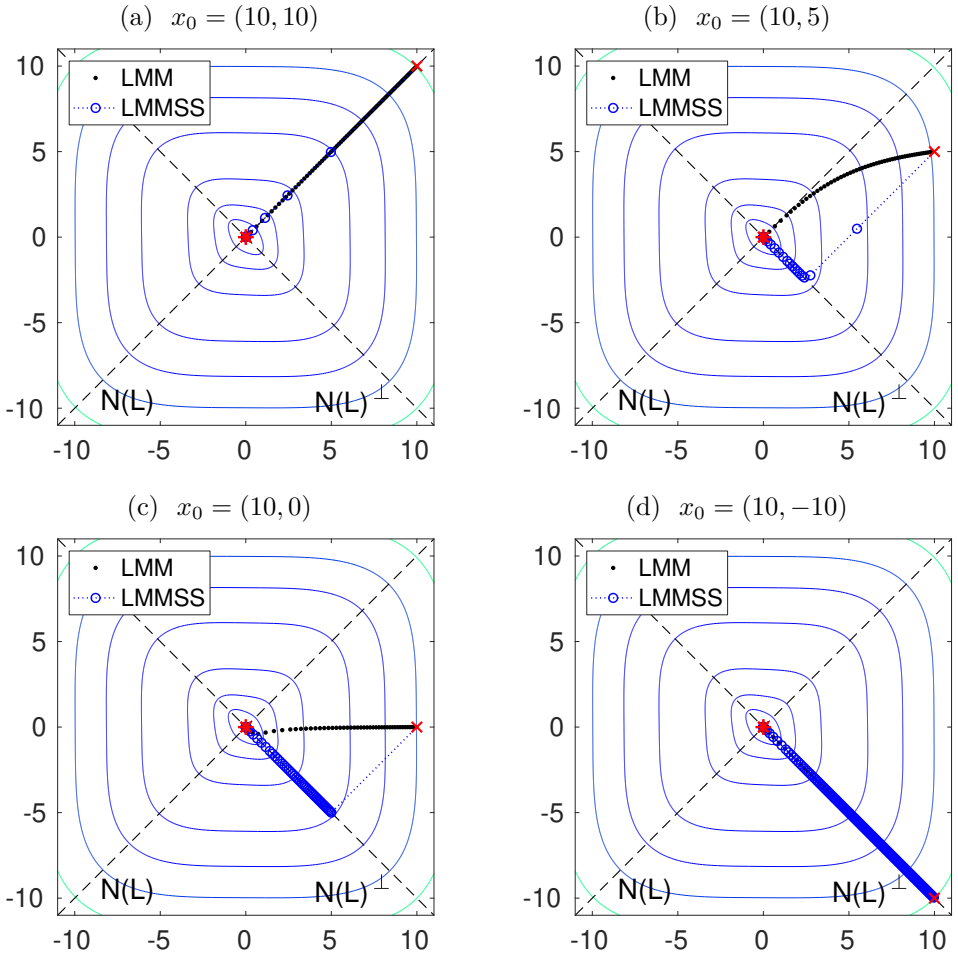
que consiste da versão discreta do operador de derivação de primeira ordem em 2 dimensões (falaremos mais destes operadores à frente). Ambas versões consideram λ_k dado pela Hipótese 4.3. Observe que

$$L^T L = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix},$$

claramente uma matriz singular. Mais além, $\mathcal{N}(L) = \text{span}\{(1, 1)^T\}$, que coincide com o gráfico da função identidade de $\mathbb{R}$ em $\mathbb{R}$ e, é fácil verificar, $\mathcal{N}(J(x, y)) \cap \mathcal{N}(L) = \{\mathbf{0}\}$ para todo $(x, y) \in \mathbb{R}^2$. Adicionalmente, $\mathcal{N}(L)^\perp = \text{span}\{(1, -1)^T\}$, constatação importante na explanação que segue. De fato, observe que os iterados gerados por LMMSS "caminham" com passos grandes o máximo possível em $\mathcal{N}(L)$ para finalmente utilizar direções em $\mathcal{N}(L)^\perp$, como facilmente visualizado na Figura 4.1, especialmente nas Figuras 4.1b e 4.1c. Já as outras duas possuem iterados totalmente em $\mathcal{N}(L)$ e $\mathcal{N}(L)^\perp$, representados pelas Figuras 4.1a e 4.1d, respectivamente, todas influenciadas pela escolha do chute inicial. Além do aspecto visual, as informações da Tabela 4.1 reforçam o ganho significativo no número de iterações em quase todos os casos, o que levanta o questionamento acerca dos motivos de tal comportamento. Ou seja, como justificar as iterações terem interesse em "cortar caminho" através do núcleo de L?

Ressaltamos também que o custo médio por iteração entre os métodos é essencialmente o mesmo, pela Tabela 4.1, de modo que o ganho em número de iterações mostra clara vantagem no custo total do algoritmo. Por exemplo, o tempo médio total para $x_0 = (10, 10)$ é de 1.0772×10^{-3} s e 8.1552×10^{-5} s para LMM e LMMSS, resp., justamente por necessitarmos de menos iterações com LMMSS. Ou seja, a troca da matriz de *scaling* parece não influenciar no custo em cada iteração, embora contribua no ganho geral a depender do iterado inicial.

Figura 4.1 – Comparativo entre iterados de LMM clássico e LMMSS. Nos gráficos, utilizamos a notação ($\times$) para diferentes pontos x_0, asterisco ($*$) para $x^* = (0, 0)$ e curvas de nível de $\phi(x, y)$ descrita em (4.6).



Fonte – o autor, 2022.

Tabela 4.1 – Comparativo entre LMM clássico e LMMSS, tanto em quantidade de iterações quanto em tempo médio por iteração (em segundos) após 500 resoluções, para diferentes chutes iniciais x_0 (veja Figura 4.1).

Método	**Chute inicial**	$(10, 10)$	$(10, 5)$	$(10, 0)$	$(10, -10)$
LMM clássico	Iterações	85	64	56	**92**
	Tempo por iteração ($\times 10^{-5}$)	1.2674	1.1976	1.1625	1.1968
LMMSS	Iterações	**7**	**27**	**55**	164
	Tempo por iteração ($\times 10^{-5}$)	1.1650	1.0753	1.1718	1.2009

Fonte – o autor, 2022.

Agora, o comportamento diferenciado de LMMSS do que faz LMM clássico tem a ver diretamente com a escolha das direções d_k. Em verdade, é sabido que d_k calculada

através de (4.2) é equivalente a resolver o problema de minimização

$$\min_{d\in\mathbb{R}^n}\{\|J_kd+F_k\|_2^2+\lambda_k\|Ld\|_2^2\}, \tag{4.7}$$

de modo que d_k é construída em um equilíbrio entre $\|J_kd+F_k\|_2^2$ e $\lambda_k\|Ld\|_2^2$. Nas iterações iniciais, com x_k "longe" de alguma solução x^*, é esperado que $\lambda_k=\|F_k\|_2^2$ assuma valores de ordem mais alta, naturalmente induzindo que $\|Ld\|_2$ seja pequeno a ponto de controlar o termo $\lambda_k\|Ld\|_2^2$ na minimização. Logo, quanto mais próximo d está de pertencer a $\mathcal{N}(L)$, tão menor fica $\|Ld\|_2$, a ponto de tornar efetivamente irrelevante $\lambda_k\|Ld\|_2^2$ na minimização. Com o andar das iterações e o valor de λ_k reduzindo, a influência de $\|J_kd+F_k\|_2^2$ se torna maior, introduzindo mais informações do problema original e conduzindo o método a algum minimizador para $\phi(x)$.

Mais ainda, os passos grandes dados através do núcleo de L se justificam justamente por esta ser uma direção degenerada, aqui pensando no método de região de confiança. Neste caso, é possível verificar que a direção calculada por (4.7) é dada equivalentemente pela resolução do problema [94]

$$\begin{array}{ll}\min & \frac{1}{2}\|J_kd+F_k\|_2^2\\ \text{s.a} & \|Ld\|_2\leq\Delta_k\end{array},$$

para algum limitante $\Delta_k>0$, conhecido como *raio da região de confiança*. A técnica vem justamente de perceber que $\|Ld\|_2\leq\Delta_k$ gera uma região dentro da qual procuramos d com segurança de minimização da função objetivo: *região de confiança* [37,38]. Se L é não singular, $\|L(\cdot)\|_2$ gera uma norma em $\mathbb{R}^n$ e, além disso, $\|Ld\|_2\leq\Delta_k$ corresponde a um elipsoide com extremos a depender dos valores singulares de L. Mais precisamente, quanto maior o valor singular, mais "achatado" estará o elipsoide na direção do vetor singular associado; analogamente, quanto menor o valor singular, mais alongada estará a região. Por este motivo dizemos que as direções através do núcleo de L são degeneradas para o caso da matriz de *scaling* ser singular: na direção dos vetores associados a $\mathcal{N}(L)$, d pode ser tão grande quanto se queira pois $\|Ld\|_2=0\leq\Delta_k$, ou seja, o "elipsoide" é ilimitado nestas direções. Isto permite que LMMSS tenha liberdade em tomar passos grandes neste subespaço sem que a restrição da região de confiança apareça, trazendo as vantagens comentadas acima e exibidas na Figura 4.1. Por outro lado, L retangular nos impede de utilizar a análise de convergência já existente para LMM através do método de região de confiança, basicamente por L possuir valores singulares nulos. O leitor interessado pode encontrar informações detalhadas e didáticas nesse tipo de análise, por exemplo, em Benatti [16, Seção 3.2] e Gardenghi e Santos [52], ou [38,93].

Retornemos ao exemplo da Figura 4.1b. O método inicia em $x_0=(10,5)$, para o qual $\lambda_0=1.0850\times10^4$ e

$$d_0=\begin{bmatrix}-4.5274\\-4.5095\end{bmatrix},$$

que mesmo não estando em $\mathcal{N}(L)$, está próximo a ponto de que $\|Ld_0\|_2 = 3.2295 \times 10^{-4}$, controlando λ_0 no segundo termo de (4.7). Veja que $d_0 \notin \mathcal{N}(L)$ apenas pela influência de $\|J_0 d + F_0\|_2^2$, uma vez que sem este termo é mais vantajoso à minimização que $\|Ld_0\|_2 = 0$. A iteração seguinte tem um comportamento similar, tomando $d_1 = [-2.7250, -2.7230]^T$ e conduzindo então a $x_2 = (2.7476, -2.2324)$, ponto próximo de $\mathcal{N}(L)^\perp$ a tal circunstância que, se d_2 estiver no núcleo de L ou próximo, teremos aumento no valor de ϕ (observe as curvas de nível da função), o que é indesejado. O método, portanto, gradualmente introduz termos de $\mathcal{N}(L)^\perp$ nas direções seguintes calculadas até estar virtualmente apenas neste espaço vetorial e sem a forte influência do núcleo de L presente anteriormente. De fato, LMMSS opta por tomar direções tais que

$$d_2 = \begin{bmatrix} -0.5032 \\ -0.1851 \end{bmatrix}, \quad d_3 = \begin{bmatrix} -0.2006 \\ 0.1566 \end{bmatrix} \quad \text{e} \quad d_4 = \begin{bmatrix} -0.1870 \\ 0.1837 \end{bmatrix},$$

em norma menores que d_0 e d_1 pela opção de tomada da direção não degenerada, limitada agora pela região de confiança. Os iterados seguintes são então essencialmente construídos em $\mathcal{N}(L)^\perp$ até atingir convergência nas proximidades da origem $(0, 0)$.

Observe que este exemplo é tal que o ponto procurando, $x^* = (0, 0)$, pertence à interseção entre $\mathcal{N}(L)$ e $\mathcal{N}(L)^\perp$, o que poderia contribuir para as propriedades de convergência. Considere, então, o exemplo seguinte, que toma $F : \mathbb{R}^2 \longrightarrow \mathbb{R}^2$ descrita por

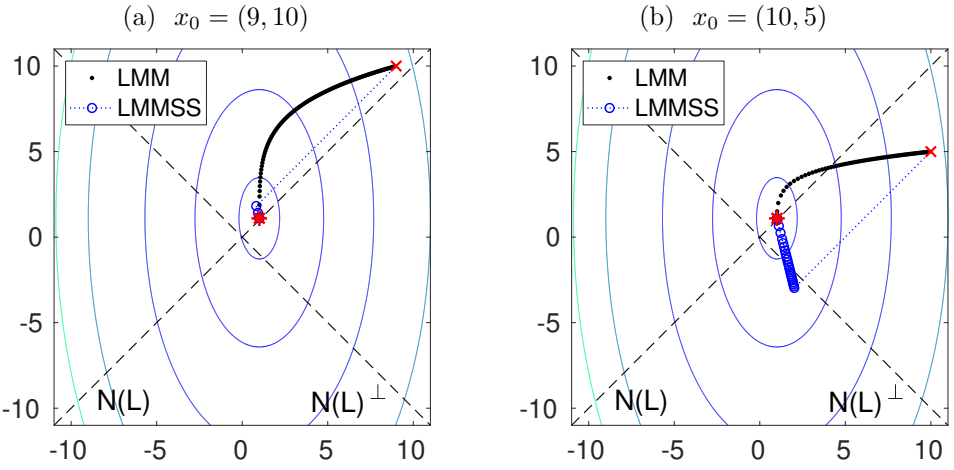
$$F(x, y) = \left(x - 1, \frac{1}{2}(y - 1.1)\right),$$

que conduz ao problema (4.1) com

$$\phi(x, y) = \frac{1}{2}\|F(x, y)\|_2^2 = \frac{1}{2}\left((x - 1)^2 + \frac{1}{4}(y - 1.1)^2\right). \tag{4.8}$$

A função ϕ representa, portanto, um paraboloide (elíptico) com vértice em $(1, 1.1)$, que é também o mínimo global (estrito) de (4.1) para este exemplo, isto é, $x^* = (1, 1.1)$. Aplicamos novamente LMM clássico e LMMSS neste problema, por efeitos de comparação, com os mesmos parâmetros do exemplo anterior. É fácil verificar que as condições para aplicação de LMMSS são satisfeitas, o que deixamos ao leitor interessado. Na Figura 4.2 exibimos alguns resultados obtidos para diferentes pontos iniciais, evidenciando um comportamento similar ao presenciado na Figura 4.1: iterações LMMSS com passos grandes visualmente paralelos a $\mathcal{N}(L)$ até a eventual mudança de direção seguindo a redução da função objetivo ϕ e posteriormente a convergência. Mesmo que x^* não seja um ponto em $\mathcal{N}(L)$ como no exemplo anterior, ainda verificamos ganho significativo em número de iterações efetuadas por LMMSS, como evidencia a Tabela 4.2. O que se infere é que, novamente, se os pontos iniciais x_0 estão próximos ao núcleo de L, então os iterados imediatamente seguintes tendem a preservar esta propriedade, por d_k dar preferência a estar em $\mathcal{N}(L)$ (ao menos inicialmente). Neste sentido, ao sabermos que x^* pertence às

Figura 4.2 – Comparativo entre iterados de LMM clássico e LMMSS. Nos gráficos, utilizamos a notação $(\times)$ para diferentes pontos x_0, asterisco $(*)$ para $x^* = (1, 1.1)$ e curvas de nível de $\phi(x, y)$ descrita em (4.8).



Fonte – o autor, 2022.

Tabela 4.2 – Comparativo entre número de iterações de LMM clássico e LMMSS, para diferentes chutes iniciais x_0 (veja Figura 4.2).

Chute inicial x_0	$(9, 10)$	$(10, 5)$
LMM clássico	111	77
LMMSS	**5**	**25**

Fonte – o autor, 2022.

proximidades de $\mathcal{N}(L)$, como acontece no exemplo corrente, é razoável considerar que LMMSS possa se aproximar rapidamente de x^* por caminhar através do núcleo de L. Quanto mais longe x_0 ou x^* estão de $\mathcal{N}(L)$, pior tende a ser o desempenho do método, como exemplificado pelas Figuras 4.1c e 4.1d (especialmente), mas também na Figura 4.2b. A aparente vantagem na aplicação de LMMSS reside, portanto, em combinar informações do ponto inicial com a solução procurada visando utilizar $\mathcal{N}(L)$ como caminho para rapidamente se aproximar de x^*, ganhando em iterações se comparado com LMM clássico.

Em resumo, LMMSS busca priorizar direções que mantenham d_k próxima ao núcleo de L, assim permitindo que passos maiores sejam dados por estarmos tomando uma direção degenerada. Eventualmente, tais direções se tornam indesejáveis à minimização, podendo conduzir a aumento de valor funcional, por exemplo. Então, os iterados invariavelmente tomam outras direções, incluindo mais termos de $\mathcal{N}(L)^\perp$, até que a convergência seja atingida. A pergunta natural que surge é: existem problemas para os quais é desejável que direções em $\mathcal{N}(L)$ sejam priorizadas? Um caso comum, por exemplo, é termos informações de suavidade da solução x^* procurada, tais como a mesma ser a versão discreta de uma função uma ou duas vezes diferenciável. Neste cenário, esperar que x^* possa ser escrita (mesmo que não totalmente) como combinação linear de vetores suaves é uma hipótese

razoável. Então, se tomarmos matrizes L que contenham em seu núcleo vetores suaves é uma forma natural de privilegiar buscas por iterados x_k de LMMSS com a mesma propriedade, especialmente se o chute inicial for também deste formato, imitando o comportamento visto na Figura 4.1a. Matrizes L que são interessantes em tais aplicações e que utilizaremos em exemplos são as versões discretas de operadores de derivação de primeira e segunda ordens, que buscam por soluções introduzindo a suavidade presente nas bases para o núcleo destas matrizes.

Este capítulo está dividido essencialmente em três partes e tem por interesse fundamental validar propriedades relacionadas à convergência das iterações definidas para LMMSS. Iniciaremos, na primeira seção, fornecendo alguns conceitos e ferramentas essenciais à compreensão do problema. Em seguida, analisaremos a convergência de LMMSS para os chamados pontos estacionários de $\phi(x)$, dada uma escolha apropriada dos parâmetros λ_k e α_k. Na última seção, verificaremos que tal convergência é de ordem quadrática localmente (como é sabido ocorrer na versão clássica de LMM) através da condição de *error bound* baseada no trabalho de Yamashita e Fukushima [122].

4.1 FERRAMENTAS PRELIMINARES

Nesta seção, buscamos estruturar as principais hipóteses e ferramentas necessárias à análise que empreenderemos. As condições propostas nas Hipóteses 4.1 e 4.3 fornecem a área de trabalho inicial com a qual buscaremos analisar propriedades de convergência. Desta forma, o objetivo aqui é desenvolver os utensílios de Álgebra Linear Matricial e Cálculo que faremos uso, visando obter conclusões e propriedades do problema a partir das hipóteses fornecidas acima.

Como o problema apresenta um par de matrizes, mais especificamente J_k e L, uma decomposição matricial que considere ambas se torna uma ferramenta analítica poderosa. De fato, utilizando da GSVD como formulada no Teorema 2.2, para cada $x \in \mathbb{R}^n$ fixo, podemos escrever o par $(J(x), L)$ como

$$J(x) = U(x) \begin{bmatrix} \Sigma(x) & \mathbf{0} \\ \mathbf{0} & I_{n-p} \end{bmatrix} X(x)^{-1} \quad \text{e} \quad L = V(x) \begin{bmatrix} M(x) & \mathbf{0} \end{bmatrix} X(x)^{-1},$$

com $\Sigma(x)$ e $M(x)$ possuindo

$$0 \leq \sigma_1(x) \leq \cdots \leq \sigma_p(x) \leq 1 \quad \text{e} \quad 1 \geq \mu_1(x) \geq \cdots \geq \mu_p(x) > 0$$

como elementos de suas diagonais respectivas. No caso em que x assume valores em uma sequência $\{x_k\}$, ou seja, $x = x_k$, para cada $k \geq 0$, é comum que simplifiquemos a notação como feito em (4.2), de modo que a GSVD do par $(J(x_k), L) = (J_k, L)$ assume a forma

$$J_k = U_k \begin{bmatrix} \Sigma_k & \mathbf{0} \\ \mathbf{0} & I_{n-p} \end{bmatrix} X_k^{-1} \quad \text{e} \quad L = V_k \begin{bmatrix} M_k & \mathbf{0} \end{bmatrix} X_k^{-1},$$

em que, como acima,

$$(\Sigma_k)_{ii} := \sigma_{i,k} \quad \text{e} \quad (M_k)_{ii} := \mu_{i,k}, \quad i = 1, \ldots, p.$$

Para referência futura, com o devido abuso de notação, podemos utilizar as mesmas definições acima para quando $x = y_k$ ou $x = z_k$, a depender do contexto.

Uma implicação de escrever J_k e L através da GSVD é, por exemplo, caracterizarmos

$$J_k^T J_k + \lambda_k L^T L = X_k^{-T} \begin{bmatrix} \Sigma_k^2 + \lambda_k M_k^2 & \mathbf{0} \\ \mathbf{0} & I_{n-p} \end{bmatrix} X_k^{-1}, \quad \forall k \in \mathbb{N}, \tag{4.9}$$

que é uma forma unificada de trabalhar com ambas as matrizes e que vamos utilizar nas argumentações consequentes. Em particular, disto segue que d_k construída via (4.2) pode ser expressa por

$$d_k = -X_k \begin{bmatrix} (\Sigma_k^2 + \lambda_k M_k^2)^{-1} & \mathbf{0} \\ \mathbf{0} & I_{n-p} \end{bmatrix} X_k^T \left(X_K^{-T} \begin{bmatrix} \Sigma_k & \mathbf{0} \\ \mathbf{0} & I_{n-p} \end{bmatrix} U_k^T \right) F_k, \quad \forall k \in \mathbb{N},$$

que se reduz a

$$d_k = -X_k \begin{bmatrix} \Gamma_k & \mathbf{0} \\ \mathbf{0} & I_{n-p} \end{bmatrix} U_k^T F_k, \quad \text{em que} \quad \Gamma_k := (\Sigma_k^2 + \lambda_k M_k^2)^{-1} \Sigma_k. \tag{4.10}$$

Daí, como as colunas de U_k são ortonormais, segue então que

$$\|d_k\|_2 \leq \|X_k\|_2 \max\{\|\Gamma_k\|_2, 1\} \|F_k\|_2, \quad \forall k \geq 0. \tag{4.11}$$

uma relação de suma importância que usaremos adiante. De fato, por (4.11) aparecer frequentemente e seus termos envolverem X_k e Γ_k, estimativas para as normas de ambas se tornam imperativas. Para melhor explanar cada uma, separamos o texto a seguir em duas subseções, abordando separadamente resultados auxiliares necessários à produção de limitantes às matrizes X_k e Γ_k, que aparecem na seção seguinte.

4.1.1 Limitando $\|X_k\|_2$

O seguinte resultado, de Hansen [62], exibe limitantes para a matriz X e sua inversa, se baseando nas propriedades da GSVD.

Teorema 4.1 (Hansen [62, Theorem 2.3]). *Se denotamos*

$$Z = \begin{bmatrix} A \\ L \end{bmatrix},$$

então

$$\|X^{-1}\|_2 = \|Z\|_2 \leq \|A\|_2 + \|L\|_2$$

e

$$\|X\|_2 = \|Z^\dagger\|_2 \leq \nu_p^{-1}, \quad com \quad \nu_p = \begin{cases} \|L^{-1}\|_2, & p = n \\ \max\{\|L^\dagger\|_2, \inf(AP_{\mathcal{N}(L)})^{-1}\}, & p < n \end{cases},$$

em que $P_{\mathcal{N}(L)}$ corresponde à matriz de projeção em $\mathcal{N}(L)$ e $\inf(AP_{\mathcal{N}(L)})$ denota o menor valor singular não nulo de $AP_{\mathcal{N}(L)}$.

Demonstração. Segue diretamente da construção da GSVD (veja [21, Theorem 22.2]) que $\sigma_i(X^{-1}) = \sigma_i(Z)$, para todo i, isto é, X^{-1} e Z possuem os mesmos valores singulares. Então, o limitante para $\|X^{-1}\|_2$ segue trivialmente da definição de Z. Quanto à limitação para $\|X\|_2$, as desigualdades de entrelaçamento para valores singulares [21, Theorem 3.5] implicam diretamente que

$$\sigma_i(Z) \geq \sigma_i\left(\begin{bmatrix} \mathbf{0} \\ L \end{bmatrix}\right) = \begin{cases} \sigma_i(L), & i = 1, \ldots, p \\ 0, & i = p+1, \ldots, n \end{cases}.$$

Para obtermos limitantes não nulos para $i = p+1, \ldots, n$, podemos considerar $\begin{bmatrix} A \\ \mathbf{0} \end{bmatrix}$ uma perturbação de $\begin{bmatrix} \mathbf{0} \\ L \end{bmatrix}$. Neste caso, segue de [114, Eq. (4)] que

$$\sigma_i(Z) \geq \inf\left(\begin{bmatrix} A \\ \mathbf{0} \end{bmatrix} P_{\mathcal{N}\left(\begin{bmatrix} \mathbf{0} \\ L \end{bmatrix}\right)}\right) = \inf(AP_{\mathcal{N}(L)}), \quad i = 1, \ldots, n,$$

em que $\inf(\cdot)$ definido como acima. Juntando as relações obtidas, um limitante inferior para $\|X\|_2^{-1} = \sigma_n(Z)$ é $\sigma_n(L)$ se $p = n$, e $\min\{\sigma_p(L), \inf(AP_{\mathcal{N}(L)})\}$ se $p < n$, o que completa o teorema. □

No primeiro momento, a conclusão que utilizaremos diz respeito à caracterização da norma de X, isto é,

$$\|X\|_2 = \left\| \begin{pmatrix} A \\ L \end{pmatrix}^\dagger \right\|_2, \tag{4.12}$$

de forma que apresentar cotas para $\|X_k\|_2$ nos induz a fornecer relações de dependência entre as variações das matrizes J_k e L. Para tanto, visto que a relação acima envolve a matriz pseudo-inversa, faremos uso de um resultado clássico da Álgebra Linear conhecido como Lema de Banach. Para a demonstração, utilizaremos a seguinte propriedade matricial, de Fan:

Proposição 4.1 (Fan [47, Theorem 2]). *Se $G,\ K,\ E \in \mathbb{R}^{m\times n}$ são tais que $G = K+E$, então, para $i,\ j \geq 0$,*

$$\sigma_{i+j+1}(G) \leq \sigma_{i+1}(K) + \sigma_{j+1}(E), \tag{4.13}$$

em que, lembramos, $\sigma_i(\cdot)$ corresponde ao i-ésimo valor singular de $(\cdot)$.

Este resultado vale sempre que $i + j + 1 \leq \text{posto}(G)$, ou seja, para que $\sigma_{i+j+1}(G)$ faça sentido. Com isto em mãos, podemos provar o lema seguinte.

Lema 4.1 (Lema de Banach). *Sejam* $A,\ E \in \mathbb{R}^{m\times n}$, $m \geq n$, *com* $\text{posto}(A) = n$. *Se* $\|A^\dagger\|_2\|E\|_2 < 1$, *então*

$$\|(A+E)^\dagger\|_2 \leq \frac{\|A^\dagger\|_2}{1 - \|A^\dagger\|_2\|E\|_2}. \tag{4.14}$$

Demonstração. Tome $G = -A$, $K = -(A+E)$. Assim, para $i+1 = n = \text{posto}(A)$ e $j = 0$, temos de (4.13) que

$$\begin{aligned} & \sigma_n(-A) \leq \sigma_n(-(A+E)) + \sigma_1(E) \\ \Leftrightarrow\quad & \sigma_n(A) \leq \sigma_n(A+E) + \|E\|_2 \\ \Leftrightarrow\quad & \sigma_n(A+E) \geq \sigma_n(A) - \|E\|_2, \end{aligned} \tag{4.15}$$

em que utilizamos a igualdade entre os valores singulares de uma matriz e de sua oposta. Agora, observe que da teoria de perturbação dos valores singulares [56, Corollary 8.6.2],

$$|\sigma_\ell(A+E) - \sigma_\ell(A)| \leq \|E\|_2, \quad \ell = 1, ..., n.$$

Desta forma, pela hipótese de que $\|E\|_2 < 1/\|A^\dagger\|_2 = \sigma_n(A)$, concluímos que $0 < \sigma_n(A+E) < 2\sigma_n(A)$, ou seja, $\text{posto}(A+E) = n$. Daí, segue que

$$\sigma_n(A+E) = \frac{1}{\|(A+E)^\dagger\|_2}$$

e, de (4.15), temos o resultado:

$$\frac{1}{\|(A+E)^\dagger\|_2} \geq \frac{1}{\|A^\dagger\|_2} - \|E\|_2 = \frac{1 - \|A^\dagger\|_2\|E\|_2}{\|A^\dagger\|_2} \quad \Leftrightarrow \quad \|(A+E)^\dagger\|_2 \leq \frac{\|A^\dagger\|_2}{1 - \|A^\dagger\|_2\|E\|_2}.$$

□

Uma versão alternativa deste resultado, que considera $\text{posto}(A) = \text{posto}(A+E) < n$, pode ser encontrada no livro de Lawson e Hanson [81, p. 43]. Vale mencionar que, na sua forma original, o Lema de Banach fornece limitantes para a inversa de matrizes com perturbações baseada na inversa original e no nível da perturbação, veja por exemplo [42, Theorem 3.1.4]. O enunciado que apresentamos acima é apenas uma parte do resultado original, voltado aos interesses deste trabalho, no contexto da matriz pseudo-inversa, diferentemente da versão mais conhecida.

4.1.2 Limitando $\|\Gamma_k\|_2$

Retornando à equação (4.10), temos a definição de

$$\Gamma_k = (\Sigma_k^2 + \lambda_k M_k^2)^{-1}\Sigma_k, \quad \forall k \geq 0.$$

Agora, se denotamos os valores singulares generalizados $\gamma_{i,k} = \sigma_{i,k}/\mu_{i,k}$, então é possível extrair relações como

$$\sigma_{i,k}^2 = \frac{\gamma_{i,k}^2}{\gamma_{i,k}^2 + 1} \quad \text{e} \quad \mu_{i,k}^2 = \frac{1}{\gamma_{i,k}^2 + 1},$$

que o leitor pode obter diretamente de $\sigma_{i,k}^2 + \mu_{i,k}^2 = 1$ e da definição de $\gamma_{i,k}$ (veja [62, p. 494], para maiores detalhes). Desta forma, podemos escrever Γ_k como

$$\Gamma_k = \operatorname{diag}\left(\frac{\gamma_{1,k}}{(\gamma_{1,k}^2 + \lambda_k)} \sqrt{1 + \gamma_{1,k}^2}, \ldots, \frac{\gamma_{p,k}}{(\gamma_{p,k}^2 + \lambda_k)} \sqrt{1 + \gamma_{p,k}^2} \right) \in \mathbb{R}^{p \times p}, \tag{4.16}$$

uma vez que

$$(\Gamma_k)_{ii} = \frac{\sigma_{i,k}}{\sigma_{i,k}^2 + \lambda_k \mu_{i,k}^2} = \frac{\gamma_{i,k}}{\gamma_{i,k}^2 + \lambda_k} \sqrt{\gamma_{i,k}^2 + 1}, \quad i = 1, \ldots, p.$$

Portanto, para produzir cotas para $\|\Gamma_k\|_2$ precisamos analisar o comportamento da relação acima com respeito a variações nos valores singulares generalizados e em λ_k, como faremos com o auxílio do seguinte lema.

Lema 4.2. *Considere a função*

$$\psi(\gamma, \lambda) = \frac{\gamma \sqrt{1 + \gamma^2}}{\gamma^2 + \lambda}, \quad \gamma \geq 0, \ \lambda > 0.$$

Então os seguintes itens são válidos:

(a) Para cada $\lambda \in (0, 1/2)$, a função $\psi(\gamma, \lambda)$ possui um único máximo em γ atingido no ponto $\gamma_{\max} = \sqrt{-\frac{\lambda}{2\lambda - 1}}$ com o valor

$$\max_{\gamma > 0} \psi(\gamma, \lambda) = \psi(\gamma_{\max}, \lambda) = \frac{1}{2\sqrt{\lambda - \lambda^2}}.$$

(b) Para $\lambda \geq 1/2$, a função $\psi(\gamma, \lambda)$ é não decrescente e limitada superiormente em γ. Mais especificamente,

$$\psi(\gamma, \lambda) \leq 1, \quad \forall \gamma > 0.$$

Demonstração. Iniciamos observando que, para qualquer λ, segue que

$$\lim_{\gamma \to \infty} \psi(\gamma, \lambda) = 1. \tag{4.17}$$

Agora, veja que, dado $\lambda > 0$,

$$\frac{\partial \psi}{\partial \gamma}(\gamma, \lambda) = \frac{(2\lambda - 1)\gamma^2 + \lambda}{\sqrt{1 + \gamma^2}(\gamma^2 + \lambda)^2}, \tag{4.18}$$

da qual extrairemos as relações mencionadas no enunciado. De fato:

(a) Seja $\lambda \in (0, 1/2)$. Como candidatos a ponto de máximo devem satisfazer $\frac{\partial \psi}{\partial \gamma}(\gamma, \lambda) = 0$, por (4.18), temos que

$$(2\lambda - 1)\gamma^2 + \lambda = 0 \quad \Rightarrow \quad \gamma_{\max} = \sqrt{-\frac{\lambda}{2\lambda - 1}},$$

por estarmos considerando apenas $\gamma \geq 0$. Observe que $\gamma_{\max} \in \mathbb{R}$ pois $2\lambda - 1 < 0$, de modo que o termo sob a raiz é positivo. Mais ainda, pelo comportamento de $(2\lambda-1)\gamma^2+\lambda$ como função de γ, é fácil ver que $\frac{\partial \psi}{\partial \gamma}(\gamma, \lambda) > 0$ se $\gamma < \gamma_{\max}$ e $\frac{\partial \psi}{\partial \gamma}(\gamma, \lambda) < 0$ se $\gamma > \gamma_{\max}$. Desta forma, concluímos que ψ é crescente para $0 \leq \gamma < \gamma_{\max}$ e em seguida decresce, caracterizando portanto $\gamma_{\max}$ como ponto de máximo. Para concluir, por substituição direta vemos que

$$\psi(\gamma_{\max}, \lambda) = \frac{1}{2\sqrt{\lambda - \lambda^2}}.$$

Ademais, comentamos que $\psi(\gamma_{\max}, \lambda) > 1$, o que é superior aos valores nos extremos para $\gamma \in [0, \infty]$, pois $\lim_{\gamma \to 0} \psi(\gamma, \lambda) = 0$ e $\lim_{\gamma \to \infty} \psi(\gamma, \lambda) = 1$ (por (4.17)).

(b) Seja $\lambda \geq 1/2$. Por (4.18), segue então que $\frac{\partial \psi}{\partial \gamma}(\gamma, \lambda) > 0$ para todo γ, uma vez que $(2\lambda - 1)\gamma^2 + \lambda > 0$. Portanto, ψ é crescente. Mais ainda, é claro que

$$\lim_{\gamma \to 0} \psi(\gamma, \lambda) = 0,$$

que juntamente com (4.17) prova a afirmação de que $\psi(\gamma, \lambda) \leq 1$, para todo $\gamma \geq 0$.

Com isto, concluímos a demonstração. □

Com este lema e o resultado da subseção anterior em mãos, podemos prosseguir com a análise de convergência. De fato, na próxima seção iremos mostrar que as iterações definidas por LMMSS através de (4.2)-(4.3) convergem para um ponto limite que possui propriedades particulares interessantes à Otimização.

4.2 CONVERGÊNCIA PARA PONTO ESTACIONÁRIO

Tendo em vista estas informações, o objetivo desta seção consiste em provar a convergência das iterações LMMSS para pontos estacionários com uso de regras para escolha do passo como, por exemplo, o critério de Armijo. Lembrando o desenvolvimento elaborado no Capítulo 2, pontos estacionários são tais que o gradiente de $\phi(x)$ se anula e são, em geral, o melhor que conseguimos garantir sem hipóteses adicionais [19, 89, 97]. Esta etapa da análise não busca estudar taxas de convergências, apenas garantir que as iterações (4.2)-(4.3) convergem; ordem de convergência é assunto da próxima seção.

O primeiro resultado que apresentamos garante que d_k é uma *direção de descida* para $\phi(x)$ a partir de x_k, ou seja, que sempre podemos encontrar $\alpha > 0$ tal que

$$\phi(x_k + \alpha d_k) < \phi(x_k).$$

Equivalentemente, d_k é de descida se $\nabla\phi(x_k)^T d_k < 0$ [19]. Em geral entende-se que esta é uma caracterização local, significando que ao menos nas proximidades de x_k, ao "caminharmos" através de d_k, vamos reduzir o valor funcional de ϕ. Este resultado consiste de uma importante constatação para a argumentos utilizados em seguida e, além disso, caracteriza LMMSS como um *método de descida* [89], isto é, um método descrito através da seguinte regra de atualização: dado um ponto inicial $w_0 \in \mathbb{R}^n$, construímos

$$w_{k+1} = w_k + \beta_k p_k, \quad k \geq 0, \tag{4.19}$$

para $\beta_k > 0$ correspondendo ao tamanho do passo e $p_k \in \mathbb{R}^n$ direção de descida. No caso de LMMSS, temos que $w_k \equiv x_k$, $\beta_k \equiv \alpha_k$ e $p_k \equiv d_k$, como explícito na equação (4.3). Para maiores informações sobre métodos e direções de descida, o leitor é direcionado a [19, 42]. Dando prosseguimento ao trabalho, fornecemos na proposição abaixo uma condição necessária e suficiente para x_k (iterado de LMMSS) ser ponto estacionário para $\phi(x)$. Mais ainda, intuitivamente, o resultado garante que ou o método está em um candidato a minimizador local ou ϕ ainda pode ter seu valor reduzido através da direção (de descida, portanto) d_k.

Proposição 4.2. *Assuma a Hipótese 4.1 e seja d_k a única solução de (4.2). Então, para cada $k \geq 0$, d_k satisfaz*

$$\nabla\phi(x_k)^T d_k \leq 0.$$

Além disso, $\nabla\phi(x_k)^T d_k = 0$ se, e somente se, x_k é um ponto estacionário para ϕ, ou seja, $\nabla\phi(x_k) = \mathbf{0}$.

Demonstração. Note que $\nabla\phi(x) = J(x)^T F(x)$, para todo $x \in \mathbb{R}^n$. Se $\lambda_k = 0$, estamos em um minimizador para ϕ, e não há mais nada a fazer. Considere, então, que $\lambda_k > 0$. Assim, com d_k dada por (4.2), temos que

$$\nabla\phi(x_k)^T d_k = -(J_k^T F_k)^T (J_k^T J_k + \lambda_k L^T L)^{-1} (J_k^T F_k). \tag{4.20}$$

Como $(J_k^T J_k + \lambda_k L^T L)$ é matriz definida positiva, podemos concluir que $\nabla\phi(x_k)^T d_k \leq 0$. Para a segunda parte, se $\nabla\phi(x_k)^T d_k = 0$, de (4.20) e da positividade de $(J_k^T J_k + \lambda_k L^T L)$, segue que $J_k^T F_k = \mathbf{0}$, i.e., x_k é ponto estacionário para ϕ. Reciprocamente, se $J_k^T F_k = \nabla\phi(x_k) = \mathbf{0}$, de (4.20) é claro que $\nabla\phi(x_k)^T d_k = 0$. □

Nas próximas linhas, desenvolvemos as ferramentas necessárias para garantir convergência das iterações de LMMSS para pontos estacionários. Para tanto, utilizaremos de uma definição e alguns resultados intermediários, exibidos a seguir. De forma simplificada, precisamos que as direções d_k de LMMSS satisfaçam uma propriedade específica, que verificaremos, e que implicam na convergência do método para pontos estacionários desde que sejam feitas escolhas sensatas para o tamanho do passo α_k em (4.3).

Inicialmente, observe que d_k pode ser entendida como uma função do ponto x_k, pela construção

$$d_k = -(J_k^T J_k + \lambda_k L^T L)^{-1} J_k^T F_k,$$

justamente por J_k, F_k e λ_k dependerem diretamente do iterado x_k em questão e L é fixa. Mais ainda, dado $x \in \mathbb{R}^n$, podemos definir

$$d := d(x) = -(J(x)^T J(x) + \lambda(x) L^T L)^{-1} J(x)^T F(x),$$

que corresponde à função que computa a direção d de LMMSS a partir de um ponto x qualquer através da equação (4.2). Agora, com esta ideia em mente, considere vetores (que queremos visualizar como direções) $p \in \mathbb{R}^n$ que são determinados unicamente para cada ponto de $\mathbb{R}^n$, isto é, temos uma função que associa cada ponto $y \in \mathbb{R}^n$ a uma única direção $p = p(y)$:

$$\begin{array}{rccc} p: & \mathbb{R}^n & \longrightarrow & \mathbb{R}^n \\ & y & \longmapsto & p(y) \end{array}.$$

São vetores desta forma para os quais utilizamos a definição seguinte:

Definição 4.1 (*Gradient-related*, Bertsekas [19, Eq. (1.13)]). *Seja $\{y_k\}$ uma sequência em $\mathbb{R}^n$ e $\{p_k\}$ a sequência de direções correspondente, com $p_k = p(y_k)$. Assim, $\{p_k\}$ é gradient-related a $\{y_k\}$ se, para qualquer subsequência $\{y_k\}_{k\in\mathcal{K}}$ de $\{y_k\}$, com $\mathcal{K} \subsetneq \mathbb{N}$, que converge a um ponto **não** estacionário, a subsequência correspondente $\{p_k\}_{k\in\mathcal{K}}$ é limitada e satisfaz*

$$\limsup_{k\to\infty,\ k\in\mathcal{K}} \nabla\phi(y_k)^T p_k < 0. \tag{4.21}$$

Observação 4.1. Para complementar detalhes no conceito de limites superiores (lim sup), recomendamos a leitura do Apêndice B, o qual contém uma breve explanação e exemplos que não caberiam a este momento do trabalho.

Dizer que uma sequência $\{p_k\}$ é *gradient-related* a $\{y_k\}$ implica basicamente que, com $\{\nabla\phi(y_k)\}_{k\in\mathcal{K}}$ convergindo a um vetor não nulo, então p_k e $\nabla\phi(y_k)$ não ficam ortogonais, o que poderia "travar" o andamento das iterações [19, 89]. Mais ainda, em conjunto com a limitação de p_k, esta propriedade faz com que as direções não sejam pequenas a ponto de não termos mudança efetiva nos pontos construídos em cada iteração e também que não sejam "grandes demais", relativamente ao tamanho de $\nabla\phi(y_k)$. Uma relação similar pode ser encontrada em Martínez e Santos [89] e Birgin e Martínez [20], que demandam a existência de constantes $\beta > 0$ e $\theta \in (0, 1)$ tais que

$$\|p_k\|_2 \geq \beta\|\nabla\phi(y_k)\|_2 \quad \text{e} \quad \nabla\phi(y_k)^T p_k \leq -\theta\|p_k\|_2\|\nabla\phi(y_k)\|_2, \quad k \geq 0. \tag{4.22}$$

Note que, em linhas gerais, estas condições buscam os objetivos semelhantes a *gradient-related*: evitar ortogonalidade entre o gradiente da função objetivo e as direções e manter

o tamanho da direção proporcional a $\nabla\phi(y_k)$. Inclusive, a segunda condição em (4.22) é conhecida como *condição do ângulo*, pois $\nabla\phi(y_k)^T p_k < 0$ implica que o ângulo entre $\nabla\phi(y_k)$ e p_k é maior que $\pi/2$, ou seja, não são ortogonais. Sob condições apropriadas para a escolha do tamanho do passo, as condições (4.22) garantem a convergência de métodos de descida na forma de (4.19); veja [89, Teorema 6.1.6] ou [20, Theorem 8.1] para maiores detalhes.

De forma similar, se um conjunto de direções é *gradient-related*, podemos verificar que a mesma converge para um ponto estacionário dado que o passo é escolhido adequadamente. Exemplos de aplicação desta técnica podem ser encontrados em [1, Chapter 4] e [113]. Desta forma, os próximos passos consistem em buscarmos mostrar que d_k gerada por (4.2) é *gradient-related*. Para tanto, iniciamos mostrando a proposição abaixo, que faz uso do lema de Banach para encontrar limitantes superiores à $\|X_k\|_2$, que usaremos logo em seguida.

Proposição 4.3. *Assuma a Hipótese 4.1. Seja $\{z_k\}$ uma sequência em $\mathbb{R}^n$ tal que $z_k \to z_\infty$. Então, existem uma constante $c_X > 0$ e um índice $k_0 \in \mathbb{N}$ de modo que*

$$\|X_k\|_2 \leq c_X, \quad \forall k \geq k_0, \tag{4.23}$$

em que $X_k = X(z_k)$.

Demonstração. Para simplificar a escrita, considere a notação

$$Z_k = \begin{pmatrix} J_k \\ L \end{pmatrix} \quad \text{e} \quad Z_\infty = \begin{pmatrix} J_\infty \\ L \end{pmatrix},$$

para $J_k = J(z_k)$ e $J_\infty = J(z_\infty)$. Como $z_k \to z_\infty$ e J é contínua (pois assumimos F continuamente diferenciável), temos que

$$\|Z_k - Z_\infty\|_2 = \left\| \begin{pmatrix} J_k \\ L \end{pmatrix} - \begin{pmatrix} J_\infty \\ L \end{pmatrix} \right\|_2 = \|J(z_k) - J(z_\infty)\|_2 \to 0,$$

implicando em $Z_k \to Z_\infty$. Daí, pela convergência de $\{Z_k\}$ para Z_∞, podemos afirmar que existe um índice k_0 tal que

$$\|Z_k - Z_\infty\|_2 \leq \frac{1}{2}\frac{1}{\|Z_\infty^\dagger\|_2}, \quad \forall k \geq k_0,$$

sempre possível pela convergência de $\{Z_k\}$ para Z_∞. Disto, segue que

$$\|Z_\infty^\dagger\|_2 \|Z_k - Z_\infty\|_2 \leq \frac{1}{2} < 1, \quad \forall k \geq k_0, \tag{4.24}$$

que se assemelha à formulação do Lema 4.1 caso tomemos $A := Z_\infty$ e $E = E_k := Z_k - Z_\infty$, $k \geq 0$. Mais ainda, da mesma forma que provamos que a direção d_k de (4.2) está bem definida, podemos verificar que $\text{posto}(Z_k) = n$ através da hipótese $\mathcal{N}(J_k) \cap \mathcal{N}(L) =$

$\{\mathbf{0}\}$, $k \geq 0$, ou mesmo da GSVD de (J_k, L). Assim, através do Lema 4.1, a equação (4.14) se torna

$$\|Z_k^\dagger\|_2 \leq \frac{\|Z_\infty^\dagger\|_2}{1 - \|Z_\infty^\dagger\|_2\|Z_k - Z_\infty\|_2} \leq \frac{\|Z_\infty^\dagger\|_2}{1 - \frac{1}{2}} = 2\|Z_\infty^\dagger\|_2, \quad \forall k \geq k_0.$$

Por (4.12), segue que $\|X_k\|_2 = \|Z_k^\dagger\|_2$ e, portanto, um limitante superior para a norma de X_k dado por

$$\|X_k\|_2 \leq 2\|Z_\infty^\dagger\|_2, \quad \forall k \geq k_0. \tag{4.25}$$

Basta tomar $c_X = 2\|Z_\infty^\dagger\|_2$ para concluir a demonstração. □

Observação 4.2. A cota apresentada na demonstração acima pode ser modificada ligeiramente, a depender da necessidade. No caso, considere $c > 1$ constante qualquer e tome $k_0 \in \mathbb{N}$ tal que (4.24) se torne

$$\|Z_\infty^\dagger\|_2\|Z_k - Z_\infty\|_2 \leq \frac{c-1}{c} < 1, \quad \forall k \geq k_0,$$

possível pela convergência de $\{Z_k\}$ para Z_∞. Desta forma, seguindo o raciocínio da demonstração através da aplicação do Lema 4.1, conseguimos que

$$\|X_k\|_2 \leq c\|Z_\infty^\dagger\|_2, \quad \forall k \geq k_0.$$

Veja que o resultado apresentado em (4.25) consiste de um caso particular da equação acima para quando $c = 2$ que, para os propósitos aos quais a proposição foi construída e que serão clarificados adiante, é suficiente. Porém, a informação extraída da desigualdade acima diz que se $c \approx 1$, então

$$\|X_k\|_2 \approx \|Z_\infty^\dagger\|_2 = \left\| \begin{pmatrix} J_\infty \\ L \end{pmatrix}^\dagger \right\|_2, \quad \forall k \geq k_0,$$

naturalmente levantando questionamento acerca da ordem de $\|X_k\|_2$ vista sua relação com J_∞ e L. Neste sentido, o Teorema 4.1 explicita que, em verdade, a relação entre $\|X\|_2$ e o par (A, L) é mais íntima:

$$\|X\|_2 = \left\| \begin{pmatrix} A \\ L \end{pmatrix}^\dagger \right\|_2 \leq \nu_p^{-1}, \quad \text{com} \quad \nu_p = \begin{cases} \|L^{-1}\|_2, & p = n, \\ \max\{\|L^\dagger\|_2, \inf(AP_{\mathcal{N}(L)})^{-1}\}, & p < n, \end{cases}$$

em que $P_{\mathcal{N}(L)}$ corresponde à matriz de projeção em $\mathcal{N}(L)$ e $\inf(AP_{\mathcal{N}(L)})$ denota o menor valor singular não nulo de $AP_{\mathcal{N}(L)}$. Daqui é possível afirmar [62, 64] que se ν_p não for grande e as matrizes são escaladas tais que $\|A\|_2 \approx \|L\|_2$, então se garante que X é bem condicionada. Mais ainda, não existe um vetor unitário z tal que $\|Az\|_2$ e $\|Lz\|_2$ são simultaneamente pequenos, intuitivamente afirmando que os valores singulares da matriz com blocos A e L não devem estar próximos da origem. Quando tratamos com problemas

mal postos, a relação entre estes valores singulares e a decomposição GSVD é ainda mais evidente e reveladora. Caso os vetores singulares de A possuam oscilações (como é comum em problemas mal postos), então esta propriedade é também existente nos vetores que compõem as colunas de X. Disto segue que se L é razoavelmente bem condicionada e escalada tal que $\|A\|_2 \approx \|L\|_2$, então o termo $\inf(AP_{\mathcal{N}(L)})^{-1}$ é de fato menor que $\|L^\dagger\|_2$. Ou seja, na prática, para problemas mal postos tais que L seja matriz bem condicionada, é usual que

$$\|X\|_2 \leq \|L^\dagger\|_2,$$

produzindo um limitante "pequeno" para $\|X\|_2$. Por outro lado, Hansen também mostra que

$$\|X^{-1}\|_2 = \left\| \begin{pmatrix} A \\ L \end{pmatrix} \right\|_2,$$

de modo que, obviamente, $\|X^{-1}\|_2 \leq \|A\|_2 + \|L\|_2$ e, portanto, o número de condição de X é

$$\kappa(X) = \|X\|_2\|X^{-1}\|_2 \leq (\|A\|_2 + \|L\|_2)\|L^\dagger\|_2,$$

levando à conclusão que X é, frequentemente, tão bem condicionada quanto L. Para provas detalhadas e outros comentários pertinentes ao tema, direcionamos o leitor para [62, 64].

Agora, com o uso da proposição anterior e de outros resultados preliminares, verificaremos que as direções de LMMSS, isto é,

$$d_k = -(J_k^T J_k + \lambda_k L^T L)^{-1} J_k^T F_k, \quad k \geq 0, \tag{4.26}$$

são *gradient-related*. Em palavras, a demonstração busca provar que se não estamos em um ponto estacionário no processo iterativo, então d_k oferece uma direção para decrescimento da função objetivo a partir do ponto em questão e tais direções são limitadas em norma. Este é um resultado central que permite, em seguida, mostrarmos a convergência das iterações propostas para LMMSS através de (4.2)-(4.3).

Proposição 4.4. *Assuma as Hipóteses 4.1 e 4.3. Assim, d_k gerada através de (4.2) é* gradient-related.

Demonstração. Seja $\{y_k\}$ sequência em $\mathbb{R}^n$ com $\{y_k\}_{k\in\mathcal{K}}$, $\mathcal{K} \subseteq \mathbb{N}$, uma subsequência qualquer que converge a um ponto y_∞ não estacionário para ϕ, isto é, $\nabla\phi(y_\infty) \neq \mathbf{0}$. Segundo a Definição 4.1, precisamos verificar que $\{d_k\}_{k\in\mathcal{K}}$, $d_k = d(y_k)$, é limitada e satisfaz (4.21). Para tanto, considere daqui em diante que $k \in \mathcal{K}$. Por (4.9) e a definição de d_k, segue que

$$\begin{aligned}
\nabla\phi(y_k)^T d_k &= -\nabla\phi(y_k)^T (J_k^T J_k + \lambda_k L^T L)^{-1} J_k^T F_k \\
&= -(X_k^T \nabla\phi(y_k))^T \begin{bmatrix} (\Sigma_k^2 + \lambda_k M_k^2)^{-1} & \mathbf{0} \\ \mathbf{0} & I_{n-p} \end{bmatrix} (X_k^T \nabla\phi(y_k)) \\
&\leq -\min\{\|(\Sigma_k^2 + \lambda_k M_k^2)^{-1}\|_2, 1\}\|X_k^T \nabla\phi(y_k)\|_2^2,
\end{aligned}$$

a segunda igualdade vindo de $\nabla\phi(y_k) = J_k^T F_k$. Pela convergência de y_k a y_∞ e da continuidade tanto das funções envolvidas quanto da GSVD, segue então que

$$\begin{aligned}\limsup_{k\to\infty,\ k\in\mathcal{K}} \nabla\phi(y_k)^T d_k &\leq \lim_{k\to\infty,\ k\in\mathcal{K}} \{-\min\{\|(\Sigma_k^2 + \lambda_k M_k^2)^{-1}\|_2, 1\}\|X_k^T\nabla\phi(y_k)\|_2^2\} \\ &= -\min\{\|(\Sigma_\infty^2 + \lambda_\infty M_\infty^2)^{-1}\|_2, 1\}\|X_\infty^T\nabla\phi(y_\infty)\|_2^2,\end{aligned}$$

pois, como y_∞ é não estacionário, temos $\lambda_\infty \neq 0$. Além disso, $X_\infty\nabla\phi(y_\infty) \neq \mathbf{0}$, já que X_∞ é não singular e $\nabla\phi(y_\infty) \neq \mathbf{0}$. Desta forma, concluímos da desigualdade acima que vale (4.21). Falta provar que a subsequência $\{d_k\}_{k\in\mathcal{K}}$ é limitada. Relembremos que, da equação (4.11), conseguimos

$$\|d_k\|_2 \leq \|X_k\|_2 \max\{\|\Gamma_k\|_2, 1\}\|F_k\|_2, \quad \forall k \geq 0, \tag{4.27}$$

com a notação $X_k = X(y_k)$, $F_k = F(y_k)$ e

$$\Gamma_k = \Gamma(y_k) = (\Sigma(y_k)^2 + \lambda(y_k)M(y_k)^2)^{-1}\Sigma(y_k),$$

definida paralelamente a (4.10). Os próximos argumentos visam construir um limitante superior para (4.27). Como $\{y_k\}_{k\in\mathcal{K}} \to y_\infty$, que é não estacionário, pela Proposição 4.3 existem $c_X > 0$ e $k_0 \geq 0$ tais que

$$\|X_k\|_2 \leq c_X, \quad \forall k \geq k_0. \tag{4.28}$$

Analisaremos agora o comportamento do termo $\max\{\|\Gamma_k\|_2, 1\}\|F_k\|_2$. Do Lema 4.2, segue que

$$\|\Gamma_k\|_2 \leq \frac{1}{2\sqrt{\lambda_k - \lambda_k^2}}, \quad \text{se} \quad 0 < \lambda_k < \frac{1}{2}, \quad \text{e} \quad \|\Gamma_k\|_2 \leq 1, \quad \text{se} \quad \lambda_k \geq \frac{1}{2}, \tag{4.29}$$

cuja dependência do valor de λ_k será desenvolvida nos dois casos a seguir. Observe que

(a) Se $0 < \lambda_k < 1/2$, é fácil ver que, do exposto acima,

$$\max\{\|\Gamma_k\|_2, 1\} \leq \frac{1}{2\sqrt{\lambda_k - \lambda_k^2}}.$$

Da Hipótese 4.3, temos que $\lambda_k = \|F_k\|_2^2$ e então

$$\begin{aligned}\max\{\|\Gamma_k\|_2, 1\}\|F_k\|_2 &\leq \frac{1}{2\sqrt{\lambda_k - \lambda_k^2}}\|F_k\|_2 \\ &= \frac{1}{2\sqrt{\|F_k\|_2^2(1 - \|F_k\|_2^2)}}\|F_k\|_2 \\ &= \frac{1}{2\sqrt{1 - \|F_k\|_2^2}} \\ &= \frac{1}{2\sqrt{1 - \lambda_k}} \\ &\leq \frac{\sqrt{2}}{2}.\end{aligned}$$

(b) Se $\lambda_k \geq 1/2$, então $\max\{\|\Gamma_k\|_2, 1\} \leq 1$. Por outro lado, considere $\varepsilon_0 > 0$ tal que $y_k \in B(y_\infty, \varepsilon_0)$, para todo $k \geq k_0$. Veja que ε_0 sempre existe devido à convergência de y_k para y_∞, com $k \in \mathcal{K}$. Sendo F contínua, do teorema de Weierstrass, é limitada no compacto $B(y_\infty, \varepsilon_0)$, isto é, podemos afirmar que $\|F_k\|_2 \leq M_F$, para todo $k \geq k_0$, para alguma constante $M_F > 0$. Assim,

$$\max\{\|\Gamma_k\|_2, 1\}\|F_k\|_2 \leq \|F_k\|_2 \leq M_F, \quad \forall k \geq k_0.$$

Portanto, os itens acima implicam que

$$\max\{\|\Gamma_k\|_2, 1\}\|F_k\|_2 \leq \max\left\{\frac{\sqrt{2}}{2}, M_F\right\}, \quad \forall k \geq k_0.$$

Disto e (4.28) aplicadas na cota (4.27) permitem que

$$\|d_k\|_2 \leq c_X \max\left\{\frac{\sqrt{2}}{2}, M_F\right\} =: M_1, \quad \forall k \geq k_0. \tag{4.30}$$

Agora, se $k < k_0$ (número finito de índices), é claro que

$$\|d_k\|_2 \leq \max_{k<k_0} \|d_k\|_2 =: M_2. \tag{4.31}$$

Combinando (4.30) e (4.31) e denotando $\hat{M} = \max\{M_1, M_2\}$, segue que $\|d_k\|_2 \leq \hat{M}$, para todo $k \in \mathcal{K}$, como desejado. □

Disto, podemos extrair um resultado de convergência, que é uma aplicação direta da proposição acima. Para tanto, considere iterações como definidas em (4.3), isto é,

$$x_{k+1} = x_k + \alpha_k d_k,$$

com d_k calculado através de (4.2) e a escolha do parâmetro α_k dada por uma das seguintes regras:

- *Busca linear exata:* tome α_k que minimiza a função objetivo na direção d_k, i.e.,

$$\phi(x_k + \alpha_k d_k) = \min_{\alpha>0} \phi(x_k + \alpha d_k).$$

- *Busca linear exata limitada:* estritamente relacionada à regra acima, esta pode ser mais fácil de aplicar em alguns contextos práticos. Dado $s > 0$, o parâmetro α_k é escolhido para efetuar a maior redução da função objetivo possível condicionada ao passo no intervalo $(0, s]$, i.e.,

$$\phi(x_k + \alpha_k d_k) = \min_{\alpha\in(0,s]} \phi(x_k + \alpha d_k).$$

  Ambas técnicas costumam ser aplicadas com alguma rotina auxiliar para a busca de mínimos em uma dimensão (*line search*).

- *Critério de Armijo:* dados escalares $s > 0$ e $0 < \nu, \eta < 1$, o passo é escolhido através do menor inteiro não negativo m que satisfaz

$$\phi(x_k + \eta^m s d_k) - \phi(x_k) \leq \nu \eta^m s \nabla\phi(x_k)^T d_k,$$

de modo que $\alpha_k = \eta^m s$. Esta regra possui uso extenso na prática, como é bem sabido. Neste trabalho (e também comum em outras aplicações), utilizamos $s = 1$.

Assim, é possível garantir a convergência de pontos limite da sequência $\{x_k\}$ para pontos estacionários de ϕ. Para tanto, referenciamos o resultado a seguir, apresentado em [19, Proposition 1.2.1], que exibimos também a demonstração por completude do texto.

Proposição 4.5 (Bertsekas [19, Proposition 1.2.1]). *Seja $\{x_k\}$ a sequência gerada por um método de descida da forma $x_{k+1} = x_k + \alpha_k d_k$, e assuma que $\{d_k\}$ é* gradient-related *e α_k é escolhido pela busca linear exata, ou busca linear exata limitada, ou critério de Armijo. Então cada ponto limite de $\{x_k\}$ é um ponto estacionário para ϕ.*

Demonstração. Considere o critério de Armijo e, para tentarmos obter uma contradição, assuma que x^* é um ponto limite de $\{x_k\}$ com $\nabla\phi(x^*) \neq \mathbf{0}$. Observe que, como $\{\phi(x_k)\}$ é monotonicamente não crescente, tal sequência converge para um valor limite ou diverge para $-\infty$. Como ϕ é contínua, $\phi(x^*)$ é um ponto limite de $\{\phi(x_k)\}$, de modo que toda a sequência $\{\phi(x_k)\}$ converge para $\phi(x^*)$. Portanto,

$$\phi(x_k) - \phi(x_{k+1}) \to 0.$$

Pela definição do critério de Armijo, temos

$$\phi(x_k) - \phi(x_{k+1}) \geq -\nu\alpha_k \nabla\phi(x_k)^T d_k. \tag{4.32}$$

Logo, $\alpha_k \nabla\phi(x^*)^T d_k \to 0$. Seja $\{x_k\}_{k\in\mathcal{K}}$ a subsequência convergindo para x^*. Como $\{d_k\}$ é *gradient-related*, temos que

$$\limsup_{k\to\infty,\ k\in\mathcal{K}} \nabla\phi(x_k)^T d_k < 0,$$

de modo que

$$\{\alpha_k\}_{k\in\mathcal{K}} \to 0.$$

Assim, pelo critério de Armijo, deve existir um índice $k_0 \geq 0$ tal que

$$\phi(x_k) - \phi(x_k + (\alpha_k/\eta)d_k) < -\nu(\alpha_k/\eta)\nabla\phi(x_k)^T d_k, \quad \forall k \in \mathcal{K},\ k \geq k_0, \tag{4.33}$$

ou seja, o passo inicial s deve ser reduzido ao menos uma vez para todo $k \in \mathcal{K},\ k \geq k_0$. Denote

$$p_k = \frac{d_k}{\|d_k\|_2}, \quad \bar{\alpha}_k = \frac{\alpha_k \|d_k\|_2}{\eta}.$$

Como $\{d_k\}$ é sequência *gradient-related*, $\{\|d_k\|_2\}_{k\in\mathcal{K}}$ é limitada, e segue que

$$\{\bar{\alpha}_k\}_{k\in\mathcal{K}} \to 0.$$

Como $\|p_k\|_2 = 1$ para todo $k \in \mathcal{K}$ (portanto limitada), existe uma subsequência $\{p_k\}_{k\in\bar{\mathcal{K}}}$, com $\bar{\mathcal{K}} \subseteq \mathcal{K}$, tal que

$$\{p_k\}_{k\in\bar{\mathcal{K}}} \to \bar{p},$$

em que $\bar{p}$ é um vetor com $\|\bar{p}\|_2 = 1$. De (4.33), temos que

$$\frac{\phi(x_k) - \phi(x_k + \bar{\alpha}_k p_k)}{\bar{\alpha}_k} < -\nu\nabla\phi(x_k)^T p_k, \quad \forall k \in \bar{\mathcal{K}},\ k \geq k_0. \tag{4.34}$$

Pelo teorema do valor médio, esta relação se torna

$$-\nabla\phi(x_k + \tilde{\alpha}_k p_k)^T p_k < -\nu\nabla\phi(x_k)^T p_k, \quad \forall k \in \bar{\mathcal{K}},\ k \geq k_0,$$

em que $\tilde{\alpha}_k$ é um escalar no intervalo $[0, \bar{\alpha}_k]$. Tomando limite para $k \in \bar{\mathcal{K}}$ na inequação acima, obtemos

$$-\nabla\phi(x^*)^T\bar{p} \leq -\nu\nabla\phi(x^*)^T\bar{p}$$

ou, ainda,

$$0 \leq (1-\nu)\nabla\phi(x^*)^T\bar{p}.$$

Como $\nu < 1$, segue que

$$0 \leq \nabla\phi(x^*)^T\bar{p}. \tag{4.35}$$

Por outro lado, temos

$$\nabla\phi(x_k)^T p_k = \frac{\nabla\phi(x_k)^T d_k}{\|d_k\|_2}.$$

Novamente tomando limite para $k \in \bar{\mathcal{K}}$, $k \to \infty$,

$$\nabla\phi(x^*)^T\bar{p} \leq \frac{\limsup_{k\to\infty,\ k\in\bar{\mathcal{K}}} \nabla\phi(x_k)^T d_k}{\limsup_{k\to\infty,\ k\in\bar{\mathcal{K}}} \|d_k\|_2} < 0,$$

o que contradiz (4.35). Isto prova o resultado para o caso do critério de Armijo.

Considere agora a busca linear exata, e seja $\{x_k\}_{k\in\mathcal{K}}$ convergente para x^* com $\nabla\phi(x^*) \neq \mathbf{0}$. Novamente temos que $\{\phi(x_k)\}$ decresce monotonicamente para $\phi(x^*)$. Denote por $\tilde{x}_{k+1}$ o ponto gerado a partir de x_k pelo critério de Armijo, e seja $\tilde{\alpha}_k$ o tamanho do passo correspondente. Temos então

$$\phi(x_k) - \phi(x_{k+1}) \geq \phi(x_k) - \phi(\tilde{x}_{k+1}) \geq -\nu\tilde{\alpha}_k\nabla\phi(x_k)^T d_k.$$

Repetindo os argumentos feitos acima seguindo a equação (4.32), substituindo α_k por $\tilde{\alpha}_k$, podemos obter uma contradição novamente. Em particular, temos que

$$\{\tilde{\alpha}_k\}_{k\in\mathcal{K}} \to 0,$$

e pela definição do critério de Armijo, temos que para algum índice $k_0 \geq 0$,

$$\phi(x_k) - \phi(x_k + (\tilde{\alpha}_k/\eta)d_k) < -\nu(\tilde{\alpha}_k/\eta)\nabla\phi(x_k)^T d_k, \quad \forall k \in \mathcal{K},\ k \geq k_0,$$

como em (4.33). Prosseguindo como antes, obtemos (4.34) e (4.35) (com $\bar{\alpha}_k = \tilde{\alpha}_k \|d_k\|_2/\eta$), e uma contradição com (4.35).

O argumento utilizado acima estabelece que qualquer regra para escolha de tamanho de passo que forneça uma maior redução que o critério de Armijo da função objetivo em cada iteração herda suas propriedades de convergência. Isto também prova a proposição para o caso da busca linear exata limitada. □

4.3 CONVERGÊNCIA COM TAXA QUADRÁTICA

Garantida a convergência global de LMMSS para pontos estacionários como visto acima, esta seção se foca em mostrar que localmente a taxa de tal convergência é quadrática. Em outras palavras, queremos verificar que se x_k está "suficientemente próximo" de algum ponto $x^* \in X^*$, então existe alguma constante, digamos c, positiva, tal que

$$\|x_{k+1} - x^*\|_2 \leq c\|x_k - x^*\|_2^2, \tag{4.36}$$

para todo k a partir de algum certo índice. Abordaremos cada parte da análise com os detalhes devidos a seguir, sendo que a estratégia apresentada aqui é fortemente baseada no desenvolvimento proposto por Yamashita e Fukushima [122], em que é feito uso da condição de *error bound* na demonstração da taxa de convergência. Em termos de organização, principiamos apresentando alguns conceitos e hipóteses consideradas, com foco na demonstração de dois teoremas centrais: o primeiro trata da convergência local quadrática propriamente dita, no contexto de (4.36); o segundo une tal resultado com a convergência global, afirmando que se pontos limites de $\{x_k\}$ são mínimos locais, então a partir de algum índice k, x_k se aproxima quadraticamente da solução. Ambos necessitam de resultados auxiliares exibidos na forma de lemas.

Iniciamos definindo a função

$$\theta_k(d) = \|J_k d + F_k\|_2^2 + \lambda_k \|Ld\|_2^2, \quad k \geq 0, \tag{4.37}$$

e note que d_k dada por (4.26) corresponde exatamente ao minimizador de $\theta_k(d)$ para $d \in \mathbb{R}^n$. De fato, podemos reescrever $\theta_k(d)$ como

$$\theta_k(d) = \left\| \begin{pmatrix} J_k d + F_k \\ \sqrt{\lambda_k} L d \end{pmatrix} \right\|_2^2 = \left\| \begin{pmatrix} J_k \\ \sqrt{\lambda_k} L \end{pmatrix} d - \begin{pmatrix} -F_k \\ \mathbf{0} \end{pmatrix} \right\|_2^2,$$

de modo que o minimizador d_k de $\theta_k(d)$ é dado pela solução das equações normais associadas ao sistema linear

$$\begin{pmatrix} J_k \\ \sqrt{\lambda_k} L \end{pmatrix} d = - \begin{pmatrix} F_k \\ \mathbf{0} \end{pmatrix},$$

que coincide exatamente com a equação (4.2). Daí concluímos que

$$d_k = \operatorname*{argmin}_{d \in \mathbb{R}^n} \theta_k(d).$$

Agora, além das condições consideradas pelo problema anteriormente, presentes nas Hipóteses 4.1 e 4.3, pedimos também as seguintes relações:

Hipótese 4.4. *Para algum $x^* \in X^*$, existem constantes $\delta \in (0, 1/2)$ e $c_1 \in (0, \infty)$ tais que*

$$\|J(y)(x-y) - (F(x) - F(y))\|_2 \leq c_1 \|x-y\|_2^2, \tag{4.38}$$

para todo $x, y \in B(x^*, 2\delta)$.

Vale pontuar que se F é continuamente diferenciável e J é Lipschitz contínua, a desigualdade (4.38) já é válida. De fato, J ser Lipschitz implica na existência de uma constante $M_J > 0$ tal que

$$\|J(x) - J(y)\|_2 \leq M_J \|x-y\|_2, \quad x, y \in B(x^*, 2\delta).$$

Neste caso, fazendo uso da aproximação de Taylor de primeira ordem para x e y, possível por F ser continuamente diferenciável, segue que

$$F(x) - F(y) = \int_0^1 J(x + t(y-x))(x-y)dt.$$

Portanto,

$$\begin{aligned}
\|J(y)(x-y) - (F(x) - F(y))\|_2 &= \left\| \int_0^1 \left[J(y) - J(x + t(y-x)) \right] (x-y) dt \right\|_2 \\
&\leq \int_0^1 \|J(y) - J(x + t(y-x))\|_2 \, \|x-y\|_2 dt \\
&\leq \int_0^1 M_J \|(1-t)(y-x)\|_2 \|x-y\|_2 dt \\
&= M_J \|x-y\|_2^2 \int_0^1 (1-t) dt \\
&= \frac{M_J}{2} \|x-y\|_2^2, \quad x, y \in B(x^*, 2\delta).
\end{aligned}$$

A relação (4.38) é comumente chamada de *erro da aproximação linear* e, para maiores detalhes na demonstração acima, veja Dennis e Schnabel [42, Theorem 4.1.12]. Adicionalmente, (4.38) implica a existência de uma constante $c_F > 0$ tal que

$$\|F(x) - F(y)\|_2 \leq c_F \|x-y\|_2, \quad \forall x, y \in B(x^*, 2\delta). \tag{4.39}$$

Para tanto, observe que por F ser continuamente diferenciável e $B(x^*, 2\delta)$ ser um conjunto compacto, existe uma constante $\widehat{M}_J > 0$ tal que $\|J(x)\|_2 \leq \widehat{M}_J$, para todo $x \in B(x^*, 2\delta)$, pelo teorema de Weierstrass. Assim, disto e usando (4.38), temos que

$$\begin{aligned}
\|F(x) - F(y)\|_2 &\leq \|F(x) - F(y) \pm J(y)(x-y)\|_2 \\
&\leq \|J(y)(x-y) - (F(x) - F(y))\|_2 + \|J(y)\|_2 \|x-y\|_2 \\
&\leq c_1 \|x-y\|_2^2 + \widehat{M}_J \|x-y\|_2 \\
&\leq c_F \|x-y\|_2, \quad \forall x, y \in B(x^*, 2\delta),
\end{aligned}$$

para $c_F := c_1 + \widehat{M}_J$ e considerando, claro, que os pontos estão suficientemente próximos a ponto de $\|x-y\|_2^2 \leq \|x-y\|_2$, o que é natural dada a caracterização local dos resultados aqui procurados.

Hipótese 4.5 (*Error bound*). *Para algum $x^* \in X^*$ da hipótese acima, $\|F(x)\|_2$ fornece um limitante de erro local em $B(x^*, 2\delta)$ para o sistema $F(x) = \mathbf{0}$, i.e., existe uma constante $c_2 \in (0, \infty)$ tal que*

$$c_2 \text{dist}(x, X^*) \leq \|F(x)\|, \quad \forall x \in B(x^*, 2\delta). \tag{4.40}$$

Condições *error bound* vem sendo utilizadas na literatura em muitos trabalhos, especialmente nos últimos 20 anos, apesar de que os primeiros artigos no assunto sejam ainda dos anos 1950 [14, 49, 69, 76, 108, 122]. Em geral, se tratam de condições de regularidade que permitem o tratamento/análise de soluções não isoladas (mas não somente) [14, 15]. Costumam também ser atrativas por exigirem menos que algumas das hipóteses de regularidade usuais, como a posto completo (ou ainda não singularidade) da matriz Jacobiana na solução. Para LMM clássico, por exemplo, esta é uma condição comum em demonstrações da convergência, como feito em [42, 93]. Como citam Behling, Gonçalves e Santos [14], se x^* é um minimizador local isolado de (4.1), ou seja, $\nabla\phi(x^*) = \mathbf{0}$ e $\nabla^2\phi(x^*)$ (Hessiana de ϕ em x^*) é definida positiva, então vale (4.40). De fato, utilizando a aproximação de Taylor até segunda ordem, nas proximidades de x^* temos

$$\phi(x) = \phi(x^*) + \nabla\phi(x^*)(x - x^*) + (x - x^*)^T \nabla^2\phi(\hat{x})(x - x^*),$$

com $\hat{x}$ no segmento de reta entre x e x^*. Como $\nabla^2\phi(x^*)$ é definida positiva, teremos uma vizinhança de x^* em que $\nabla^2\phi(\hat{x})$ é também definida positiva. Disto e de x^* ser solução do problema, segue que

$$\frac{1}{2}\|F(x)\|_2^2 = \phi(x) = (x - x^*)^T \nabla^2\phi(\hat{x})(x - x^*) \geq \sigma_n(\nabla^2\phi(\hat{x}))\|x - x^*\|_2^2.$$

Portanto, basta tomar $c_2 := 1/\sqrt{2\sigma_n(\nabla^2\phi(\hat{x}))}$ para a condição (4.40) ser válida. De forma similar, se $J(x^*)$ tem posto completo então a condição de *error bound* é verificada [14]. O contrário não é verdade em geral. Por exemplo, se $F(x) = Ax - b$ (função linear), inclusive para o caso de resíduo não nulo, podemos garantir a Hipótese 4.5 sem exigir informações do posto de A [14, 69, 76, 108]. Portanto, ao utilizarmos a condição de *error bound*, estamos permitindo análise de uma variedade maior de problemas do que as hipóteses clássicas assumem.

4.3.1 Análise local

De posse das ferramentas acima, podemos iniciar a análise de convergência local para LMMSS. Em seguida, apresentamos um lema chave, responsável por estender para

LMMSS um resultado equivalente apresentado por Yamashita e Fukushima [122]. Por rigorosidade na notação, para todo k, $\bar{x}_k$ denota um vetor em X^* tal que

$$\|x_k - \bar{x}_k\|_2 = \text{dist}(x_k, X^*).$$

Além disso, assumimos que x^* é como nas Hipóteses 4.4 e 4.5 e, adicionalmente, são válidas durante toda a seção as Hipóteses 4.1-4.3, permitindo que os coeficientes λ_k estejam bem determinados e que tenhamos propriedades acerca das matrizes J_k e L. Finalmente, é importante ressaltar que esta subseção se foca em verificar convergência local de iterações LMMSS sem escolha de passo, isto é, calculamos d_k como em (4.2) e atualizamos

$$x_{k+1} = x_k + d_k, \quad k \geq 0. \tag{4.41}$$

A questão de escolha de passo, que faremos através do critério de Armijo, retorna na subseção seguinte, em que faremos uma análise global da técnica. Note que a escolha de passo por Armijo garante que pontos limite da sequência gerada por LMMSS são estacionários. Agora, se tal ponto limite for um minimizador do problema, a ideia é verificar que os passos são sempre de tamanho 1 a partir de um certo índice k, permitindo portanto a aliança entre a análise local e a global (com Armijo).

Lema 4.3. *Suponha que as Hipóteses 4.4-4.5 são válidas e que a matriz L foi escalada de modo que $\|L\|_2 = 1$. Se $x_k \in B(x^*, \delta)$, então a solução d_k de (4.2) satisfaz*

$$\|d_k\|_2 \leq c_3 \text{dist}(x_k, X^*), \tag{4.42}$$

e

$$\|J_k d_k + F_k\|_2 \leq c_4 \text{dist}(x_k, X^*)^2, \tag{4.43}$$

em que $c_3 = \dfrac{1}{c_2\sqrt{\gamma}}\sqrt{c_1^2 + c_2^2(1 + c_F^2)}$ *e* $c_4 = \sqrt{c_1^2 + c_F^2}$.

Demonstração. Como $x_k \in B(x^*, \delta)$, segue que

$$\|\bar{x}_k - x^*\|_2 \leq \|\bar{x}_k - x_k\|_2 + \|x_k - x^*\|_2 \leq \|x_k - x^*\|_2 + \|x_k - x^*\|_2 \leq 2\delta,$$

implicando que $\bar{x}_k \in B(x^*, 2\delta)$.

Agora, usando a Hipótese 4.2 com $x = d_k$ temos

$$\gamma\|d_k\|_2^2 \leq \|J_k d_k\|_2^2 + \|L d_k\|_2^2. \tag{4.44}$$

Estimaremos cada termo do lado direito separadamente. Para o primeiro, observe que

$$\|J_k(J_k^T J_k + \lambda_k L^T L)^{-1} J_k^T\|_2 \leq 1,$$

a qual segue imediatamente do uso da GSVD do par (J_k, L). Com efeito,

$$J_k(J_k^T J_k + \lambda_k L^T L)^{-1} J_k^T = U_k \begin{bmatrix} \Sigma_k(\Sigma_k^2 + \lambda_k M_k^2)^{-1}\Sigma_k & \mathbf{0} \\ \mathbf{0} & I_{n-p} \end{bmatrix} U_k^T,$$

de modo que, pelas colunas de U_k serem unitárias,

$$\|J_k(J_k^TJ_k+\lambda_kL^TL)^{-1}J_k^T\|_2 \leq \max_{1\leq i\leq p}\left\{\frac{\sigma_{i,k}^2}{\sigma_{i,k}^2+\lambda_k\mu_{i,k}^2},1\right\}\leq 1,$$

pois $\lambda_k\mu_{i,k}^2>0$, para todo k. Disto e usando a expressão de d_k em (4.26), segue que

$$\|J_kd_k\|_2^2=\|J_k(J_k^TJ_k+\lambda_kL^TL)^{-1}J^TF_k\|_2^2\leq\|F_k\|_2^2\leq c_F^2\|\bar{x}_k-x_k\|_2^2, \tag{4.45}$$

a última desigualdade vindo de (4.39) e do fato que $F(\bar{x}_k)=\mathbf{0}$. Para estimar o segundo termo em (4.44), usamos o fato de que d_k é minimizador do funcional $\theta_k(d)$. De fato, como $\lambda_k\|Ld_k\|\leq\theta_k(d_k)\leq\theta_k(d)$, para todo $d\in\mathbb{R}^n$, temos

$$\lambda_k\|Ld_k\|_2^2\leq\theta_k(d_k)\leq\theta_k(\bar{x}_k-x_k).$$

Mas, usando a definição de θ_k e a Hipótese 4.4 temos

$$\begin{aligned}\theta_k(\bar{x}_k-x_k)&=\|J_k(x_k)(\bar{x}_k-x_k)+F_k\|_2^2+\lambda_k\|L(\bar{x}_k-x_k)\|_2^2\\&\leq c_1^2\|\bar{x}_k-x_k\|_2^4+\lambda_k\|\bar{x}_k-x_k\|_2^2,\end{aligned}$$

lembrando que $\|L\|_2=1$. Com base nesta desigualdade e no disposto acima, segue que

$$\|Ld_k\|_2^2\leq\frac{c_1^2}{\lambda_k}\|\bar{x}_k-x_k\|_2^4+\|\bar{x}_k-x_k\|_2^2\leq\frac{c_1^2+c_2^2}{c_2^2}\|\bar{x}_k-x_k\|_2^2, \tag{4.46}$$

em que utilizamos a Hipótese 4.3 e a condição de *error bound* para produzir $\lambda_k=\|F_k\|_2^2\geq c_2^2\|\bar{x}_k-x_k\|_2^2$. Substituindo (4.45) e (4.46) em (4.44) segue que

$$\|d_k\|_2\leq c_3\|\bar{x}_k-x_k\|_2,$$

com $c_3=\dfrac{1}{c_2\sqrt{\gamma}}\sqrt{c_1^2+c_2^2(1+c_F^2)}$.

Para a segunda parte do lema, por d_k minimizar $\theta_k(d)$ e por $F(\bar{x}_k)=\mathbf{0}$, temos

$$\begin{aligned}\|J_kd_k+F_k\|_2^2&\leq\theta_k(d_k)\\&\leq\theta_k(\bar{x}_k-x_k)\\&=\|J_k(\bar{x}_k-x_k)-(F(\bar{x}_k)-F_k)\|_2^2+\lambda_k\|L(\bar{x}_k-x_k)\|_2^2\\&\leq c_1^2\|\bar{x}_k-x_k\|_2^4+\lambda_k\|\bar{x}_k-x_k\|_2^2,\end{aligned}$$

a última desigualdade fazendo uso da Hipótese 4.4. Como $\lambda_k=\|F_k\|_2^2=\|F_k-F(\bar{x}_k)\|_2^2\leq c_F^2\|x_k-\bar{x}_k\|_2^2$, concluímos que

$$\|J_kd_k+F_k\|_2^2\leq(c_1^2+c_F^2)\|x_k-\bar{x}_k\|_2^4.$$

Definindo $c_4=\sqrt{c_1^2+c_F^2}$, finalizamos a demonstração com

$$\|J_kd_k+F_k\|_2\leq c_4\|x_k-\bar{x}_k\|_2^2=c_4\mathrm{dist}(x_k,X^*)^2.$$

□

Este resultado é de suma importância, sendo necessário nas demonstrações dos dois próximos lemas, que por sua vez baseiam o primeiro teorema da seção. Agora, dando continuidade à análise, no próximo enunciado mostramos que $\text{dist}(x_k, X^*)$ converge quadraticamente para zero contanto que os iterados $\{x_k\}$ estejam suficientemente próximos de x^*.

Lema 4.4 (Yamashita e Fukushima [122]). *Se $x_k, x_{k-1} \in B(x^*, \delta)$, então vale*

$$\text{dist}(x_k, X^*) \leq c_5 \text{dist}(x_{k-1}, X^*)^2,$$

em que $c_5 = (c_1 c_3^2 + c_4)/c_2$.

Demonstração. Como $x_k, x_{k-1} \in B(x^*, \delta)$ e $x_k = x_{k-1} + d_{k-1}$, segue da Hipótese 4.5 que

$$\begin{aligned} c_2 \text{dist}(x_k, X^*) &= c_2 \text{dist}(x_{k-1} + d_{k-1}, X^*) \\ &\leq \|F(x_{k-1} + d_{k-1})\|_2 \\ &= \|F(x_{k-1} + d_{k-1}) \pm J_{k-1} d_{k-1} \pm F(x_{k-1})\|_2 \\ &\leq \|J_{k-1} d_{k-1} - (F(x_{k-1} + d_{k-1}) - F(x_{k-1}))\|_2 + \|J_{k-1} d_{k-1} + F(x_{k-1})\|_2 \\ &\leq c_1 \|d_{k-1}\|_2^2 + \|J_{k-1} d_{k-1} + F(x_{k-1})\|_2, \end{aligned}$$

a última linha um resultado de aplicar a Hipótese 4.4 para $x := x_{k-1} + d_{k-1}$ e $y := x_{k-1}$. Agora, utilizando as desigualdades obtidas no Lema 4.3, temos

$$c_2 \text{dist}(x_k, X^*) \leq c_1 c_3^2 \text{dist}(x_{k-1}, X^*)^2 + c_4 \text{dist}(x_{k-1}, X^*)^2.$$

Basta então definir $c_5 = (c_1 c_3^2 + c_4)/2$ para concluir. □

Agora, devemos garantir que se o ponto inicial x_0 está suficientemente próximo de x^*, então todos os pontos x_k da sequência se mantém a uma certa distância de x^* também. Este é o foco do próximo resultado.

Lema 4.5 (Yamashita e Fukushima [122]). *Se $x_0 \in B(x^*, r)$, então temos que $x_k \in B(x^*, \delta)$ para todo k, em que*

$$r := \min \left\{ \frac{\delta}{2 + 4c_3}, \frac{1}{2c_5} \right\}.$$

Demonstração. Mostremos que, para cada k dado, se $x^\ell \in B(x^*, \delta)$, $\ell = 1, \ldots, k$, então $x_{k+1} \in B(x^*, \delta)$. Então, como $x_0 \in B(x^*, r) \subseteq B(x^*, \delta)$, isto mostra o lema. Consideremos dois casos: $k = 0$ e $k \geq 1$.

(i) Quando $k = 0$, segue do Lema 4.3 e de $x_0 \in B(x^*, r)$ que

$$\begin{aligned}\|x_1 - x^*\|_2 &= \|x_0 + d_0 - x^*\|_2 \\ &\leq \|x_0 - x^*\|_2 + \|d_0\|_2 \\ &\leq r + c_3 \text{dist}(x_0, X^*) \\ &\leq r + c_3 \|x_0 - x^*\|_2 \\ &\leq (1 + c_3)r. \end{aligned} \tag{4.47}$$

Como $(1 + c_3)r \leq \delta$ (pela hipótese sobre r), temos que $x_1 \in B(x^*, \delta)$.

(ii) Agora consideremos o caso $k \geq 1$. Então, como $x^\ell \in B(x^*, \delta)$, para $\ell = 0, 1, \ldots, k$, obtemos a partir do Lema 4.4 que, para $\ell = 1, \ldots, k$,

$$\text{dist}(x_\ell, X^*) \leq c_5 \text{dist}(x_{\ell-1}, X^*)^2 \leq \cdots \leq c_5^{2^\ell - 1} \|x_0 - x^*\|_2^{2^\ell} \leq r \left(\frac{1}{2}\right)^{2^\ell - 1},$$

em que a última desigualdade segue de $\|x_0 - x^*\|_2 \leq r$ e $r \leq \frac{1}{2c_5}$. Consequentemente, pelo Lema 4.3, temos que para $\ell = 1, \ldots, k$,

$$\|d_\ell\|_2 \leq c_3 \text{dist}(x_\ell, X^*) \leq c_3 r \left(\frac{1}{2}\right)^{2^\ell - 1}. \tag{4.48}$$

Daí, como $x_{k+1} = x_k + d_k = x_{k-1} + d_{k-1} + d_k = \cdots = x_1 + \sum_{\ell=1}^k d_k$, segue que

$$\begin{aligned}\|x_{k+1} - x^*\|_2 &\leq \|x_1 - x^*\|_2 + \sum_{\ell=1}^k \|d_\ell\|_2 \\ &\leq (1 + c_3)r + c_3 r \sum_{\ell=1}^k \left(\frac{1}{2}\right)^{2^\ell - 1} \\ &\leq (1 + c_3)r + c_3 r \sum_{\ell=1}^\infty \left(\frac{1}{2}\right)^{2^\ell - 1}, \end{aligned} \tag{4.49}$$

a segunda desigualdade vindo de (4.47) e (4.48). Observe que

$$\begin{aligned}\sum_{\ell=1}^\infty \left(\frac{1}{2}\right)^{2^\ell} &= \left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^4 + \left(\frac{1}{2}\right)^8 + \ldots \\ &\leq \left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^4 + \left(\frac{1}{2}\right)^6 + \left(\frac{1}{2}\right)^8 + \ldots \\ &= \sum_{\ell=1}^\infty \left(\frac{1}{4}\right)^\ell \\ &= \frac{1}{3}, \end{aligned}$$

sendo a última igualdade obtida da série geométrica de razão $1/4$. Desta forma, como

$$\sum_{\ell=1}^\infty \left(\frac{1}{2}\right)^{2^\ell - 1} = 2 \sum_{\ell=1}^\infty \left(\frac{1}{2}\right)^{2^\ell} \leq 2\frac{1}{3} < 1, \tag{4.50}$$

voltamos a (4.49) com

$$\|x_{k+1} - x^*\|_2 \leq (1 + 2c_3)r \leq \frac{\delta}{2},$$

válida pois $r \leq \frac{\delta}{2+6c_4}$. Logo, temos que $x_{k+1} \in B(x^*, \delta)$. □

Com estes dois lemas, podemos finalmente provar o teorema de convergência local quadrática, a seguir. Reforçamos que as Hipóteses 4.1–4.5 seguem válidas.

Teorema 4.2 (Yamashita e Fukushima [122]). *Suponha que $\{x_k\}$ é gerada pelo método LMMSS através das iterações*

$$(J_k^T J_k + \lambda_k L^T L)d_k = -J_k^T F_k \quad e$$
$$x_{k+1} = x_k + d_k, \quad k \geq 0,$$

com $x_0 \in B(x^*, r)$. Então, $\{\text{dist}(x_k, X^*)\}$ converge para zero quadraticamente. Mais ainda, a sequência $\{x_k\}$ converge para uma solução $\hat{x} \in X^* \cap B(x^*, \delta)$.

Demonstração. A primeira parte do teorema segue diretamente dos Lemas 4.4 e 4.5. Precisamos apenas mostrar que a sequência $\{x_k\}$ converge para uma solução $\hat{x} \in X^* \cap B(x^*, \delta)$. Como $\{\text{dist}(x_k, X^*)\}$ converge para zero e $x_k \in B(x^*, \delta)$ para todo k, é suficiente mostrar que $\{x_k\}$ converge. Como (4.48) implica que

$$\|d_k\|_2 \leq c_3 r \left(\frac{1}{2}\right)^{2k-1}$$

para todo $k \geq 1$, temos que para inteiros positivos p, q tais que $p \geq q$,

$$\|x_p - x_q\|_2 \leq \sum_{i=q}^{p-1} \|d_i\|_2 \leq \sum_{i=q}^{\infty} \|d_i\|_2 \leq c_3 r \sum_{i=q}^{\infty} \left(\frac{1}{2}\right)^{2i-1} = \frac{1}{3} c_3 r \left(\frac{1}{2}\right)^{2q-3},$$

com a série geométrica saindo de raciocínio similar a (4.50). Portanto, $\{x_k\}$ é uma sequência de Cauchy e, portanto, é convergente (em $\mathbb{R}$). □

4.3.2 Análise global

Se consideramos d_k dado por (4.2) e as iterações LMMSS na forma

$$x_{k+1} = x_k + \alpha_k d_k, \ \alpha_k > 0,$$

com α_k determinado através do critério de Armijo, então podemos também desenvolver uma análise de convergência global. Para tanto, fornecemos no Algoritmo 4.1 os critérios para a escolha de α_k e λ_k (via Hipótese 4.3), que seguem a estratégia de Yamashita e Fukushima [122]. O teorema a seguir, que trata da demonstração desta convergência, mescla o desenvolvimento elaborado até este ponto: garante que, com passo escolhido pela regra de Armijo, temos convergência para ponto estacionário, que acontece com taxa quadrática localmente.

Algoritmo 4.1 LMMSS com busca linear pelo critério de Armijo

Entrada: Constantes $\nu, \eta, \tau \in (0,1)$, $F(x)$, $J(x)$, L, aproximação inicial $x_0 \in \mathbb{R}^n$
Saída: Ponto estacionário (possivelmente mínimo local) $\bar{x}$ para $\phi(x) = \frac{1}{2}\|F(x)\|_2^2$

$k \leftarrow 0$
$\lambda_0 \leftarrow \|F(x_0)\|_2^2$
Para $k = 0, 1, \ldots$ **faça**
  Se (critério de parada satisfeito) **então**
    $\bar{x} \leftarrow x_k$
  Fim Se
  $d_k \leftarrow -(J_k^T J_k + \lambda_k L^T L)^{-1} J_k^T F_k$ ▷ Direção de descida
  Se $\|F(x_k + d_k)\|_2 \leq \tau \|F(x_k)\|_2$ **então**
    $x_{k+1} \leftarrow x_k + d_k$
  Senão ▷ Critério de Armijo
    Tome m como o menor inteiro não negativo tal que

$$\phi(x_k + \eta^m d_k) - \phi(x_k) \leq \nu \eta^m \nabla\phi(x_k)^T d_k \tag{4.51}$$

    $x_{k+1} \leftarrow x_k + \eta^m d_k$
  Fim Se
  $\lambda_{k+1} \leftarrow \|F(x_{k+1})\|_2^2$
  $k \leftarrow k + 1$
Fim Para

Teorema 4.3 (Yamashita e Fukushima [122]). *Seja $\{x_k\}$ gerada pelo Algoritmo 4.1 (LMMSS com busca linear pelo critério de Armijo). Então qualquer ponto de acumulação da sequência $\{x_k\}$ é um ponto estacionário para ϕ. Mais ainda, se um ponto de acumulação x^* de $\{x_k\}$ é uma solução do sistema $F(x) = 0$, então a sequência $\{\text{dist}(x_k, X^*)\}$ converge para 0 quadraticamente dadas as Hipóteses 4.4 e 4.5 para este x^* em particular e a validade das Hipóteses 4.1-4.3.*

Demonstração. Observe que a primeira parte relacionada à convergência de $\{x_k\}$ para um ponto estacionário de ϕ é tratada na Proposição 4.5, portanto precisamos apenas verificar a segunda parte. Como o ponto de acumulação x^* é uma solução para $F(x) = 0$, então existe um índice $k_0 \in \mathbb{N}$ tal que

$$\|F(x_{k_0})\|_2 \leq \frac{c_2^2 \tau}{c_5 c_F} \tag{4.52}$$

e

$$\|x_{k_0} - x^*\|_2 \leq r,$$

em que $\tau \in (0,1)$ é definida no Algoritmo 4.1 e r é a constante positiva especificada no Lema 4.5. Agora, seja $\{y_k\}$ a sequência gerada por LMMSS com passo unitário e $y_0 := x_{k_0}$. Observe que $\{y_k\}$ vem de (4.41) e sua construção busca usufruir das propriedades obtidas para a convergência local. De fato, pelo Teorema 4.2, $\text{dist}(y_j, X^*)$ converge para 0 quadraticamente. Portanto, é suficiente mostrar que $x_{k_0+j} = y_j$ para todo j, i.e., $\{y_j\}$

satisfaz

$$\|F(y_{j+1})\|_2 \leq \tau \|F(y_j)\|_2,$$

para todo j. Isto implicaria, essencialmente, que a partir do índice k_0 somente passos de tamanho unitário são tomados, de modo que o teorema de convergência local pode ser aplicado. Note que, dos Lemas 4.4 e 4.5 e da Hipótese 4.5, temos

$$\text{dist}(y_{j+1}, X^*) \leq c_5 \text{dist}(y_j, X^*)^2 \leq c_5 c_2^2 \|F(y_j)\|_2^2,$$

para todo j. Seja $\bar{y}_{j+1}$ um ponto em X^* tal que $\|y_{j+1} - \bar{y}_{j+1}\|_2 = \text{dist}(y_{j+1}, X^*)$. Segue então da desigualdade acima, de (4.39) e da Hipótese 4.5 que

$$\begin{aligned} \|F(y_{j+1})\|_2 &= \|F(y_{j+1}) - F(\bar{y}_{j+1})\|_2 \\ &\leq c_F \|y_{j+1} - \bar{y}_{j+1}\|_2 \\ &= c_F \text{dist}(y_{j+1}, X^*) \\ &\leq \frac{c_5 c_F \|F(y_j)\|_2}{c_2^2} \|F(y_j)\|_2. \end{aligned} \tag{4.53}$$

Como (4.52) implica

$$\frac{c_5 c_F \|F(y_0)\|_2}{c_2^2} \leq \tau,$$

da convergência de y_j temos que $\|F(y_j)\|_2 \leq \|F(y_0)\|_2$, para todo j, de modo que

$$\frac{c_5 c_F \|F(y_j)\|_2}{c_2^2} \leq \tau.$$

Portanto, de (4.53) e $\tau < 1$, segue

$$\|F(y_{j+1})\|_2 \leq \tau \|F(y_j)\|_2,$$

para todo j, o que completa a prova. □

5 RECONSTRUÇÃO DE CONDUTIVIDADE TÉRMICA COM APLICAÇÕES INDUSTRIAIS

Neste capítulo, consideraremos um problema de reconstrução de condutividade térmica, um tema de variadas aplicações na ciência e especialmente engenharia. Conhecer a condutividade térmica de um material tem enorme impacto em aplicações industriais, por exemplo, indo de componentes estruturais em construções ao desenvolvimento de aeronaves e pesquisa espacial, uma vez que são raros os processos sem ocorrência de interações térmicas. Alguns exemplos envolvem a determinação de taxa de desgaste de ferramentas, qualidade de produtos, aplicações em refrigeração, motores e painéis solares e interação com outros materiais, para citar alguns [3, 5, 71, 85, 100]. Outros usos podem ser encontrados dependendo da área de estudo, como é o caso em imagens médicas (tomografia é um exemplo direto) em que a função de interesse é vista como condutividade elétrica, então trocando também temperatura por potencial elétrico [2].

Matematicamente, estamos interessados na recuperação da condutividade térmica de um material sólido de duas dimensões cuja dinâmica de transferência de calor é descrita pelo problema de valor de contorno abaixo. Considere o domínio $\Omega \times [0, t_f]$, $t_f > 0$, com $\Omega = (0, l_1) \times (0, l_2) \subseteq \mathbb{R}^2$, $l_1, l_2 > 0$, e a equação diferencial parcial que modela condução de calor,

$$C(x,y)\frac{\partial u}{\partial t}(x,y,t) = \nabla \cdot [K(x,y)\nabla u(x,y,t)] - q(x,y)u(x,y,t) + g(x,y,t), \tag{5.1}$$

no cilindro $\Omega \times (0, t_f]$. As variáveis envolvidas são *capacidade térmica* $C(x,y) > 0$, *termo de reação* $q(x,y) \geq 0$, *termo fonte* $g(x,y,t)$, *função de temperatura* $u(x,y,t)$ e *condutividade térmica* (ortotrópica), com dependência espacial,

$$K(x,y) = \begin{bmatrix} k_{11}(x,y) & 0 \\ 0 & k_{22}(x,y) \end{bmatrix}, \tag{5.2}$$

contínua, com $k_{11}, k_{22} > 0$. Completamos o modelo com as condições de fronteira

$$-k_{11}(0,y)\frac{\partial u}{\partial x}(0,y,t) + h_1(y)(u(0,y,t) - f_1(y,t)) = 0, \quad \text{em} \quad x = 0,\ y \in (0, l_2), \tag{5.3}$$

$$k_{11}(l_1,y)\frac{\partial u}{\partial x}(l_1,y,t) + h_2(y)(u(l_1,y,t) - f_2(y,t)) = 0, \quad \text{em} \quad x = l_1,\ y \in (0, l_2), \tag{5.4}$$

$$-k_{22}(x,0)\frac{\partial u}{\partial y}(x,0,t) + h_3(x)(u(x,0,t) - f_3(x,t)) = 0, \quad \text{em} \quad y = 0,\ x \in (0, l_1), \tag{5.5}$$

$$k_{22}(x,l_2)\frac{\partial u}{\partial y}(x,l_2,t) + h_4(x)(u(x,l_2,t) - f_4(x,t)) = 0, \quad \text{em} \quad y = l_2,\ x \in (0, l_1), \tag{5.6}$$

para todo $t \in (0, t_f]$, e condição inicial

$$u(x,y,0) = u_0(x,y) \quad \text{na região } \Omega, \tag{5.7}$$

com h_i, $i = 1, \ldots, 4$, conhecidas como *funções de transferência de calor*, f_i, $i = 1, \ldots, 4$, *funções de fluxo de calor* e $u_0(x,y)$ é a *temperatura inicial*. O objetivo central desta parte

do trabalho [25, 26] é recuperar aproximações para a condutividade ortotrópica $K(x, y)$ baseadas em informações/medidas da temperatura e considerando disponíveis todos os outros coeficientes e funções envolvidas no modelo.

Diversos artigos descrevem métodos para lidar com a recuperação da condutividade, cada um envolvendo estratégias numéricas específicas e com as suas particularidades nos aspectos teóricos. Exemplos envolvem amplas técnicas variando do método de diferenças finitas ao método de elementos de fronteira, que omitimos, por brevidade. Para maior aprofundamento, citamos, e.g. [2, 32, 86, 90, 104].

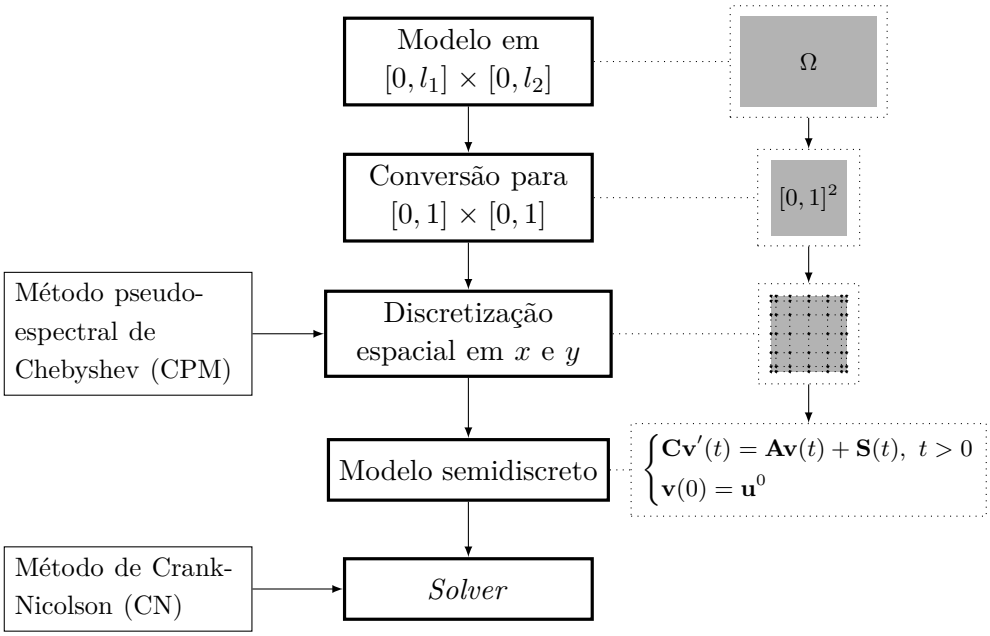
Neste contexto, nossa proposta consiste em desenvolver uma estratégia alternativa em duas partes: a primeira, baseada no método pseudo-espectral de Chebyshev (CPM) [31, 58] para discretizar o modelo (5.1)–(5.7); a segunda parte é composta da aplicação de uma versão do método de Levenberg-Marquardt com *scaling* singular (LMMSS) para propriamente recuperar a condutividade através de um problema inverso que faz uso da discretização anterior. CPM é introduzido aqui para lidar com o chamado *problema direto*, que consiste na determinação da temperatura u dados os outros parâmetros do modelo, inclusive a condutividade. Este método foi escolhido com foco no claro processo de discretização e boas propriedades de convergência, normalmente exigindo baixo esforço computacional para produzir resultados de qualidade. Neste sentido, vale reforçar que a ordem de convergência de CPM chega a ser exponencial desde que a solução procurada seja ela mesma suficientemente suave [31].

O material apresentado neste capítulo é o resultado de estudo desenvolvido em Boos, Luchesi e Bazán [26] e Boos, Bazán e Luchesi [25], o primeiro publicado e o segundo a ponto de submissão. Ambos os trabalhos lidam com o problema de reconstrução como relatado acima, com a observação de que o segundo trabalho [25] apresenta modificações e aperfeiçoamentos à discretização proposta no artigo anterior. Além disso, ampliamos a técnica para problemas com restrições de medição, como descreveremos no decorrer do texto, uma situação que surge em problemas práticos por aspectos técnicos restritivos.

5.1 PROBLEMA DIRETO

No tratamento numérico de problemas inversos não lineares através de métodos iterativos, o problema direto tem de ser resolvido repetidas vezes e, portanto, uma técnica eficiente para fazê-lo é essencial para termos sucesso no método inverso escolhido. De forma grosseira, a estratégia se baseia em utilizar CPM para produzir aproximações às derivadas espaciais através da matriz de diferenciação de Chebyshev, transformando o modelo original em um sistema de EDOs dependente do tempo para o qual diversos métodos existem. Exibimos na Figura 5.1 um diagrama dos passos a serem tomados na construção da solução ao problema direto, já considerando que utilizaremos o método de Crank-Nicolson (CN) no problema semidiscreto. Interesse nesta abordagem surgiu a partir da crescente percepção de que EDPs podem ser solucionadas com alta precisão e

Figura 5.1 – Diagrama simplificado dos passos utilizados para construir o resolvedor (*solver*) ao problema direto.



Fonte – o autor, 2022.

baixo custo operacional quando comparado com métodos baseados em diferenças finitas ou elementos finitos, por exemplo; adicionalmente, esta abordagem tem sido aplicada com sucesso na resolução de problemas diretos e inversos em problemas de condução de calor ainda recentemente, veja, e.g. [7, 10, 73].

Iniciamos transformando o domínio espacial $\Omega = (0, l_1) \times (0, l_2)$ em (5.1)–(5.7) no quadrado unitário. Os motivos se tornarão claros a seguir durante o processo de discretização. Fazendo desta forma e mantendo as mesmas notações que no modelo original (para evitar sobrecarga de definições), o modelo transformado se torna

$$c(x,y)\frac{\partial u}{\partial t} = \frac{1}{l_1^2}\frac{\partial}{\partial x}\left(k_{11}(x,y)\frac{\partial u}{\partial x}\right) + \frac{1}{l_2^2}\frac{\partial}{\partial y}\left(k_{22}(x,y)\frac{\partial u}{\partial y}\right) - q(x,y)u(x,y,t) + g(x,y,t), \tag{5.8}$$

para $(x, y, t) \in (0, 1) \times (0, 1) \times [0, t_f]$, e

$$-\frac{1}{l_1}k_{11}(0,y)\frac{\partial u}{\partial x}(0,y,t) + h_1(y)(u(0,y,t) - f_1(y,t)) = 0, \quad \text{em} \quad x = 0, \ y \in (0,1), \tag{5.9}$$

$$\frac{1}{l_1}k_{11}(1,y)\frac{\partial u}{\partial x}(l_1,y,t) + h_2(y)(u(1,y,t) - f_2(y,t)) = 0, \quad \text{em} \quad x = 1, \ y \in (0,1), \tag{5.10}$$

$$-\frac{1}{l_2}k_{22}(x,0)\frac{\partial u}{\partial y}(x,0,t) + h_3(x)(u(x,0,t) - f_3(x,t)) = 0, \quad \text{em} \quad y = 0, \ x \in (0,1), \tag{5.11}$$

$$\frac{1}{l_2}k_{22}(x,1)\frac{\partial u}{\partial y}(x,1,t) + h_4(x)(u(x,1,t) - f_4(x,t)) = 0, \quad \text{em} \quad y = 1, \ x \in (0,1), \tag{5.12}$$

$$u(x,y,0) = u_0(x,y) \quad \text{em} \quad [0,1] \times [0,1], \tag{5.13}$$

em que $t \in (0, t_f]$ nas condições de fronteira (5.9)-(5.12).

No método pseudo-espectral de Chebyshev, derivadas espaciais são aproximadas em uma malha consistindo de $(n+1)\times(n+1)$ pontos em $[0,1]\times[0,1]$ baseados em $(n+1)$ pontos de Chebyshev-Gauss-Lobatto nas direções horizontal e vertical, respectivamente, definidos por

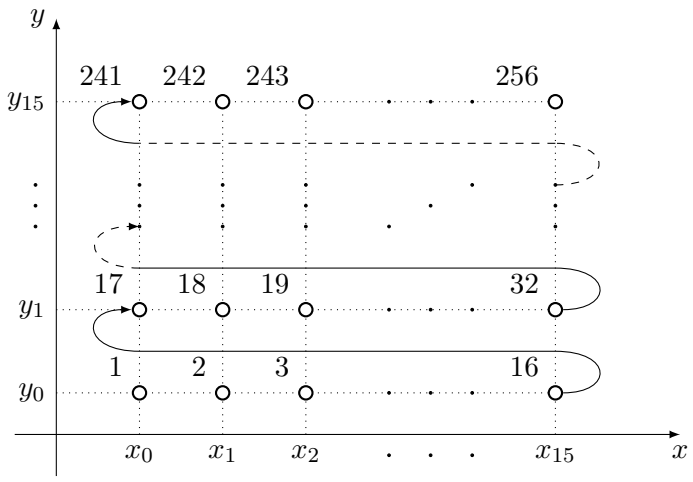
$$x_i = \frac{1}{2}\left[1-\cos\left(\frac{i\pi}{n}\right)\right], \quad y_j = \frac{1}{2}\left[1-\cos\left(\frac{j\pi}{n}\right)\right], \quad i,j = 0,1,\ldots,n, \tag{5.14}$$

cuja característica de maior concentração nos extremos do domínio evita, por exemplo, o conhecido fenômeno de Runge. A partir disto, as aproximações às derivadas são baseadas em um produto matriz-vetor da matriz de diferenciação de Chebyshev com o vetor de valores funcionais na malha. Desta forma, a EDP é transformada em um sistema de EDOs dependente da variável temporal apenas, que pode ser resolvido de diversas formas, como veremos adiante. Daqui em diante, sempre assumiremos que os pontos da malha estão numerados seguindo a chamada *ordem lexicográfica*, isto é, para cada ponto da malha existe um único número positivo definido por

$$\ell = i + j(n+1) + 1, \quad 0 \le i,j \le n, \tag{5.15}$$

que representa os pontos da malha consecutivamente da esquerda para a direita, de baixo para cima, como ilustrado na Figura 5.2. Rotular os pontos desta forma permite que aproximações para $u(x_i, y_j, t)$ que estamos procurando podem ser armazenadas em um longo vetor com $(n+1)^2$ entradas cujos componentes são indexados por ℓ.

Figura 5.2 – Exemplo de malha enumerada em ordem lexicográfica, para $n = 15$.



Fonte – Boos, Bazán e Luchesi [25].

5.1.1 Discretização das derivadas espaciais

Para construir aproximações às derivadas espaciais usando CPM, denote por $D \in \mathbb{R}^{(n+1)\times(n+1)}$ a matriz de diferenciação de Chebyshev e assuma que a mesma está

particionada da seguinte forma:

$$D = \begin{bmatrix} r_0^T \\ r_1^T \\ \vdots \\ r_n^T \end{bmatrix} = [d_0, d_1, \ldots, d_n], \quad r_i, d_i \in \mathbb{R}^{n+1},\ i = 0, \ldots, n. \tag{5.16}$$

Como utilizaremos mais a frente, ressaltamos que assim como os vetores linha/coluna de D são indexados de 0 a n, ressaltamos que o mesmo é feito com as suas entradas. Iniciemos então o processo de discretização propriamente dito, com intenção de substituir as derivadas espaciais em (5.8) por versões discretas das mesmas através das propriedades de D, como veremos a seguir, começando com as derivadas com respeito a x primeiro. Antes de continuarmos, para a conveniência do leitor não familiar com CPM, observe que se $\mathsf{f} \in \mathbb{R}^{n+1}$ é um vetor contendo valores funcionais $f(x_i)$ ao longo da malha de Chebyshev em 1D, então as derivadas $f'(x_i)$ são aproximadas por $(D\mathsf{f})_i = r_i^T \mathsf{f},\ i = 0, 1, \ldots, n$, ou globalmente, com abuso de notação, o vetor contendo valores de derivada pode ser aproximado por

$$[f'(x_0), f'(x_1), \ldots, f'(x_n)]^T \approx D\mathsf{f}.$$

Para gerarmos aproximações das derivadas parciais com respeito a x seja

$$\mathbf{u}_j = [u(x_0, y_j, t), u(x_1, y_j, t), \ldots, u(x_n, y_j, t)]^T, \quad j = 0, 1, \ldots, n, \tag{5.17}$$

ou seja, o vetor $\mathbf{u}_j$ acomoda os valores de u na j-ésima linha da malha e, analogamente, seja

$$\overset{\blacktriangle}{\mathbf{u}}_j = \left[\frac{\partial u}{\partial x}(x_0, y_j, t), \frac{\partial u}{\partial x}(x_1, y_j, t), \ldots, \frac{\partial u}{\partial x}(x_n, y_j, t)\right]^T, \quad j = 0, 1, \ldots, n. \tag{5.18}$$

Com esta notação, é claro que se não existem restrições para $u(x, y, t)$ em Ω, então

$$\mathbf{u}_j^x := \begin{bmatrix} \dfrac{\partial}{\partial x}\left(k_{11}(x_0, y_j)\dfrac{\partial u}{\partial x}(x_0, y_j, t)\right) \\ \vdots \\ \dfrac{\partial}{\partial x}\left(k_{11}(x_n, y_j)\dfrac{\partial u}{\partial x}(x_n, y_j, t)\right) \end{bmatrix} \approx D(\mathbf{K}_j^x \overset{\blacktriangle}{\mathbf{u}}_j), \quad j = 0, \ldots, n, \tag{5.19}$$

em que $\mathbf{K}_j^x$ denota a matriz diagonal

$$\mathbf{K}_j^x = \operatorname{diag}(k_{11}(x_0, y_j), k_{11}(x_1, y_j), \ldots, k_{11}(x_n, y_j)), \quad j = 0, 1, \ldots, n. \tag{5.20}$$

Portanto, com a observação de que $\overset{\blacktriangle}{\mathbf{u}}_j \approx D\mathbf{u}_j$ (aproximação das derivadas exatas através de CPM), nos pontos interiores da malha na direção horizontal temos

$$\frac{\partial}{\partial x}\left(k_{11}(x_i, y_j)\frac{\partial u}{\partial x}(x_i, y_j, t)\right) \approx r_i^T(\mathbf{K}_j^x D\mathbf{u}_j), \quad i = 1, \ldots, n-1,\ j = 0, \ldots, n, \tag{5.21}$$

de modo que, utilizando a notação do Matlab, $\mathbf{u}_j^x(2:n) = D_2\mathbf{K}_j^x D$, com $\mathbf{u}_j^x(2:n)$ contendo as componentes de $\mathbf{u}_j^x$ relativas a $x_i = x_1, \ldots, x_{n-1}$. Para incorporar as restrições descritas nas condições de fronteira (5.9)-(5.10) em (x_0, y_j) e (x_n, y_j), $j = 0, \ldots, n$, no processo de discretização, observe de (5.19) que

$$\begin{aligned}
\mathbf{u}_j^x &\approx \sum_{i=0}^{n} d_i k_{11}(x_i, y_j)\frac{\partial u}{\partial x}(x_i, y_j, t) \\
&= d_0 k_{11}(x_0, y_j)\frac{\partial u}{\partial x}(x_0, y_j, t) + \sum_{i=1}^{n-1} d_i k_{11}(x_i, y_j)\frac{\partial u}{\partial x}(x_i, y_j, t) + d_n k_{11}(x_n, y_j)\frac{\partial u}{\partial x}(x_n, y_j, t) \\
&= \left[l_1 h_1(y_j) d_0 e_1^T + D_1 \check{\mathbf{K}}_j^x D_2 - l_1 h_2(y_j) d_n e_{n+1}^T\right]\mathbf{u}_j - l_1 d_0 h_1(y_j) f_1(y_j, t) \\
&\quad + l_1 d_n h_2(y_j) f_2(y_j, t) \\
&\doteq \mathbf{A}_j \mathbf{u}_j + \mathbf{b}_j,
\end{aligned}$$

em que e_1, e_{n+1} denotam o primeiro e último vetores coluna da matriz identidade de ordem $n+1$, I_{n+1}, e

$$\mathbf{A}_j = l_1 h_1(y_j) d_0 e_1^T + D_1 \check{\mathbf{K}}_j^x D_2 - l_1 h_2(y_j) d_n e_{n+1}^T, \tag{5.22}$$

$$\mathbf{b}_j = -l_1 d_0 h_1(y_j) f_1(y_j, t) + l_1 d_n h_2(y_j) f_2(y_j, t), \tag{5.23}$$

com

$$D_1 = [d_1, \ldots, d_{n-1}], \quad D_2 = \begin{bmatrix} r_1^T \\ \vdots \\ r_{n-1}^T \end{bmatrix} \quad \text{e} \quad \check{\mathbf{K}}_j^x = \operatorname{diag}(k_{11}(x_1, y_j), \ldots, k_{11}(x_{n-1}, y_j)).$$

Daí, as derivadas parciais em relação a x nas fronteiras à direita e à esquerda podem ser aproximadas como

$$\frac{\partial}{\partial x}\left(k_{11}(x_0, y_j)\frac{\partial u}{\partial x}(x_0, y_j, t)\right) \approx e_1^T \mathbf{A}_j \mathbf{u}_j + e_1^T \mathbf{b}_j, \quad j = 0, \ldots, n, \tag{5.24}$$

$$\frac{\partial}{\partial x}\left(k_{11}(x_n, y_j)\frac{\partial u}{\partial x}(x_n, y_j, t)\right) \approx e_{n+1}^T \mathbf{A}_j \mathbf{u}_j + e_{n+1}^T \mathbf{b}_j, \quad j = 0, \ldots, n. \tag{5.25}$$

Baseado em (5.21), (5.24) e (5.25), a aproximação das derivadas com relação a x ao longo da j-ésima linha horizontal da malha se torna

$$\mathbf{u}_j^x \approx \mathcal{F}_j \mathbf{u}_j + \mathbf{S}_j^x, \quad j = 0, 1, \ldots, n, \tag{5.26}$$

em que

$$\mathcal{F}_j = \begin{bmatrix} e_1^T \mathbf{A}_j \\ D_2 \mathbf{K}_j^x D \\ e_{n+1}^T \mathbf{A}_j \end{bmatrix}, \quad \mathbf{S}_j^x(t) = \left[\begin{array}{c} e_1^T \mathbf{b}_j \\ \hline \mathbf{0} \\ \hline e_{n+1}^T \mathbf{b}_j \end{array}\right], \quad j = 0, \ldots, n. \tag{5.27}$$

em que $\mathbf{0} \in \mathbb{R}^{n-1}$ é vetor com entradas nulas. Desta forma, a discretização do termo $\dfrac{\partial}{\partial x}\left(k_{11}(x,y)\dfrac{\partial u}{\partial x}(x,y,t)\right)$ nos pontos (x_i, y_j), $i = 0, \ldots, n$, $j = 0, \ldots, n$, com valores de temperatura $u(x_i, y_j, t)$ indexados em ordem lexicográfica se torna

$$\begin{bmatrix} \mathbf{u}_0^x \\ \mathbf{u}_1^x \\ \vdots \\ \mathbf{u}_n^x \end{bmatrix} \approx \begin{pmatrix} \mathcal{F}_0 & & & \\ & \mathcal{F}_1 & & \\ & & \ddots & \\ & & & \mathcal{F}_n \end{pmatrix} \begin{bmatrix} \mathbf{u}_0 \\ \mathbf{u}_1 \\ \vdots \\ \mathbf{u}_n \end{bmatrix} + \begin{bmatrix} \mathbf{S}_0^x(t) \\ \mathbf{S}_1^x(t) \\ \vdots \\ \mathbf{S}_n^x(t) \end{bmatrix}. \tag{5.28}$$

Um procedimento similar pode ser seguido para aproximar derivadas com respeito a y em (5.8) com a diferença principal que aqui os pontos da malha são vistos verticalmente: fixamos x_i e variamos y_j para aproximar a derivada procurada. Mais especificamente, introduzindo o análogo vertical de (5.17),

$$\widehat{\mathbf{u}}_i = [u(x_i, y_0, t), u(x_i, y_1, t), \ldots, u(x_i, y_n, t)]^T, \quad i = 0, 1, \ldots, n, \tag{5.29}$$

segue que

$$\frac{\partial}{\partial y}\left(k_{22}(x_i, y_j)\frac{\partial u}{\partial y}(x_i, y_j, t)\right) \approx r_j^T(\mathbf{K}_i^y D\widehat{\mathbf{u}}_i), \quad i = 0, 1, \ldots, n, \ j = 1, \ldots, n-1, \tag{5.30}$$

em que

$$\mathbf{K}_i^y = \operatorname{diag}(k_{22}(x_i, y_0), k_{22}(x_i, y_1), \ldots, k_{22}(x_i, y_n)), \quad i = 0, 1, \ldots, n.$$

Em seguida, assim como feito para as derivadas parciais em relação a x nas fronteiras à esquerda e à direita, para as fronteiras inferiores e superiores pode ser visto que

$$\frac{\partial}{\partial y}\left(k_{22}(x_i, y_0)\frac{\partial u}{\partial y}(x_i, y_0, t)\right) \approx e_1^T\mathbf{B}_i\widehat{\mathbf{u}}_i + e_1^T\mathbf{d}_i, \quad i = 0, \ldots, n, \tag{5.31}$$

$$\frac{\partial}{\partial y}\left(k_{22}(x_i, y_n)\frac{\partial u}{\partial y}(x_i, y_n, t)\right) \approx e_{n+1}^T\mathbf{B}_i\widehat{\mathbf{u}}_i + e_{n+1}^T\mathbf{d}_i, \quad i = 0, \ldots, n, \tag{5.32}$$

para $\mathbf{B}_i$ e $\mathbf{d}_i$ desenvolvidas como em (5.22)-(5.23), i.e.,

$$\begin{aligned} \mathbf{B}_i &= \left[l_2 h_3(x_i) d_0 e_1^T + D_1\check{\mathbf{K}}_i^y D_2 - l_2 h_4(x_i) d_n e_{n+1}^T\right], \\ \mathbf{d}_i &= -l_2 d_0 h_3(x_i) f_3(x_i, t) + l_2 d_n h_4(x_i) f_4(x_i, t), \quad \text{e} \\ \check{\mathbf{K}}_i^y &= \operatorname{diag}(k_{22}(x_i, y_1), \ldots, k_{22}(x_i, y_{n-1})), \quad i = 0, 1, \ldots, n. \end{aligned}$$

Agora, baseado em (5.30)-(5.32), as derivadas parciais em relação a y ao longo da i-ésima linha vertical da malha de $k_{22}(x,y)\dfrac{\partial u}{\partial y}(x,y,t)$, se arranjada em um vetor $\widehat{\mathbf{u}}_i^y$ (i.e., em (x_i, y_j) para i fixo, $0 \le j \le n$), pode ser aproximada por

$$\widehat{\mathbf{u}}_i^y \approx \mathbf{G}_i\widehat{\mathbf{u}}_i + \widehat{\mathbf{S}}_i^y(t), \quad i = 0, 1, \ldots, n, \tag{5.33}$$

em que

$$\mathbf{G}_i = \begin{bmatrix} e_1^T\mathbf{B}_i \\ D_2\mathbf{K}_i^y D \\ e_{n+1}^T\mathbf{B}_i \end{bmatrix}, \quad \widehat{\mathbf{S}}_i^y(t) = \left[\begin{array}{c} e_1^T\mathbf{d}_i \\ \hline \mathbf{0} \\ \hline e_{n+1}^T\mathbf{d}_i \end{array}\right], \quad i = 0, \ldots, n.$$

Para referência futura, as entradas de $\mathbf{G}_i$ serão denotadas por $g_{pq}^{(i)},\ 0 \le p, q \le n$.

A ideia-chave para reordenar as derivadas parciais em relação a y em (5.33) de modo que elas sigam a ordem lexicográfica é notar que as entradas de $\widehat{\mathbf{u}}_i^y$ podem ser expressas em termos de um vetor em blocos $\mathbf{u}$ contendo todos os valores funcionais indexados em ordem lexicográfica e arranjados com componentes em blocos $\mathbf{u}_j$, $j = 0, 1, \ldots, n$. Para construir esta reordenação de forma clara, iniciamos aproximando as derivadas parciais com respeito a y ao longo da primeira linha horizontal de modo que

$$\begin{aligned}
\frac{\partial}{\partial y}\left(k_{22}(x_0,y_0)\frac{\partial u}{\partial y}(x_0,y_0,t)\right) &\approx e_1^T\mathbf{G}_0\widehat{\mathbf{u}}_0 + e_1^T\mathbf{d}_0 = \mathbf{e}_1^T(\mathbf{G}_0\otimes I_{n+1})\mathbf{u} + e_1^T\mathbf{d}_0,\\
\frac{\partial}{\partial y}\left(k_{22}(x_1,y_0)\frac{\partial u}{\partial y}(x_1,y_0,t)\right) &\approx e_1^T\mathbf{G}_1\widehat{\mathbf{u}}_1 + e_1^T\mathbf{d}_1 = \mathbf{e}_2^T(\mathbf{G}_1\otimes I_{n+1})\mathbf{u} + e_1^T\mathbf{d}_1,\\
&\vdots\\
\frac{\partial}{\partial y}\left(k_{22}(x_n,y_0)\frac{\partial u}{\partial y}(x_n,y_0,t)\right) &\approx e_1^T\mathbf{G}_n\widehat{\mathbf{u}}_n + e_1^T\mathbf{d}_n = \mathbf{e}_{n+1}^T(\mathbf{G}_n\otimes I_{n+1})\mathbf{u} + e_1^T\mathbf{d}_n,
\end{aligned} \tag{5.34}$$

em que $\otimes$ denota o *produto tensorial* (ou *produto de Kronecker* [22, 56]) de matrizes e $\mathbf{e}_j$ denota a j-ésima coluna da matriz identidade $I_{(n+1)^2}$. Arranjando as derivadas no vetor coluna $\mathbf{u}_0^y$, as aproximações acima podem ser expressas por

$$\mathbf{u}_0^y \approx \begin{bmatrix} \mathcal{G}_{00} & \mathcal{G}_{01} & \cdots & \mathcal{G}_{0n} \end{bmatrix} \begin{bmatrix} \mathbf{u}_0 \\ \vdots \\ \mathbf{u}_n \end{bmatrix} + \begin{bmatrix} e_1^T\mathbf{d}_0 \\ \vdots \\ e_1^T\mathbf{d}_n \end{bmatrix} \tag{5.35}$$

em que

$$\mathcal{G}_{0k} = \begin{pmatrix} g_{0k}^{(0)} & & & \\ & g_{0k}^{(1)} & & \\ & & \ddots & \\ & & & g_{0k}^{(n)} \end{pmatrix}, \quad 0 \le k \le n. \tag{5.36}$$

De forma similar, para a j-ésima linha horizontal da malha, $j = 1, \ldots, n-1$, obtemos que

$$\begin{aligned}
\frac{\partial}{\partial y}\left(k_{22}(x_0,y_j)\frac{\partial u}{\partial y}(x_0,y_j,t)\right) &\approx e_{j+1}^T\mathbf{G}_0\widehat{\mathbf{u}}_0 + e_{j+1}^T\mathbf{d}_0 = \mathbf{e}_{j+1}^T(\mathbf{G}_0\otimes I_{n+1})\mathbf{u},\\
\frac{\partial}{\partial y}\left(k_{22}(x_1,y_j)\frac{\partial u}{\partial y}(x_1,y_j,t)\right) &\approx e_{j+1}^T\mathbf{G}_1\widehat{\mathbf{u}}_1 + e_{j+1}^T\mathbf{d}_1 = \mathbf{e}_{j+2}^T(\mathbf{G}_1\otimes I_{n+1})\mathbf{u},\\
&\vdots\\
\frac{\partial}{\partial y}\left(k_{22}(x_n,y_j)\frac{\partial u}{\partial y}(x_n,y_j,t)\right) &\approx e_{j+1}^T\mathbf{G}_n\widehat{\mathbf{u}}_n + e_{j+1}^T\mathbf{d}_n = \mathbf{e}_{j+n+1}^T(\mathbf{G}_n\otimes I_{n+1})\mathbf{u},
\end{aligned} \tag{5.37}$$

enquanto que para $j = n$, a aproximação se torna

$$\begin{aligned}
\frac{\partial}{\partial y}\left(k_{22}(x_0,y_n)\frac{\partial u}{\partial y}(x_0,y_n,t)\right) &\approx e_{n+1}^T\mathbf{G}_0\widehat{\mathbf{u}}_0 + e_{n+1}^T\mathbf{d}_0 = \mathbf{e}_{n+1}^T(\mathbf{G}_0\otimes I_{n+1})\mathbf{u} + e_{n+1}^T\mathbf{d}_0,\\
\frac{\partial}{\partial y}\left(k_{22}(x_1,y_n)\frac{\partial u}{\partial y}(x_1,y_n,t)\right) &\approx e_{n+1}^T\mathbf{G}_1\widehat{\mathbf{u}}_1 + e_{n+1}^T\mathbf{d}_1 = \mathbf{e}_{n+2}^T(\mathbf{G}_1\otimes I_{n+1})\mathbf{u} + e_{n+1}^T\mathbf{d}_1,\\
&\vdots\\
\frac{\partial}{\partial y}\left(k_{22}(x_n,y_n)\frac{\partial u}{\partial y}(x_n,y_n,t)\right) &\approx e_{n+1}^T\mathbf{G}_n\widehat{\mathbf{u}}_n + e_{n+1}^T\mathbf{d}_n = \mathbf{e}_{n+n+1}^T(\mathbf{G}_n\otimes I_{n+1})\mathbf{u} + e_{n+1}^T\mathbf{d}_n.
\end{aligned} \tag{5.38}$$

Levando em conta o procedimento descrito em (5.35), assim como as aproximações (5.37)-(5.38) arranjadas em um vetor em blocos cujas componentes são $\mathbf{u}_i^y$, as aproximações para as derivadas $\dfrac{\partial}{\partial y}\left(k_{22}(x,y)\dfrac{\partial u}{\partial y}(x,y,t)\right)$ ao longo de toda a malha se tornam

$$\begin{bmatrix}\mathbf{u}_0^y\\ \mathbf{u}_1^y\\ \vdots\\ \mathbf{u}_n^y\end{bmatrix} \approx \begin{bmatrix}\mathcal{G}_{00} & \mathcal{G}_{01} & \dots & \mathcal{G}_{0n}\\ \mathcal{G}_{10} & \mathcal{G}_{11} & \dots & \mathcal{G}_{1n}\\ \vdots & \vdots & \ddots & \vdots\\ \mathcal{G}_{n0} & \mathcal{G}_{n1} & \dots & \mathcal{G}_{nn}\end{bmatrix}\begin{bmatrix}\mathbf{u}_0\\ \mathbf{u}_1\\ \vdots\\ \mathbf{u}_n\end{bmatrix} + \begin{bmatrix}\mathbf{S}_0^y(t)\\ \mathbf{S}_1^y(t)\\ \vdots\\ \mathbf{S}_n^y(t)\end{bmatrix} \tag{5.39}$$

em que $\mathcal{G}_{sk}$, $0 \le s,k \le n$, são definidas similarmente ao feito em (5.36) e

$$\mathbf{S}_0^y(t) = \begin{bmatrix}e_1^T\mathbf{d}_0\\ \vdots\\ e_1^T\mathbf{d}_n\end{bmatrix}, \quad \mathbf{S}_i^y(t) = \mathbf{0}\in\mathbb{R}^{n+1},\ i=1,\dots,n-1, \quad \text{e} \quad \mathbf{S}_n^y(t) = \begin{bmatrix}e_{n+1}^T\mathbf{d}_0\\ \vdots\\ e_{n+1}^T\mathbf{d}_n\end{bmatrix}. \tag{5.40}$$

Desta forma, podemos prosseguir à montagem do problema semidiscreto, a seguir.

Uma observação de suma importância no que diz respeito ao procedimento de discretização é que todos os valores de k_{11} e k_{22} são incluídos diretamente na aproximação, diferentemente da abordagem proposta em [26] em que as aproximações não incluíam estes valores nos cantos do domínio. Esta dependência tem papel crucial quando resolvemos o problema inverso de estimar a condutividade a partir de dados de temperatura. É claro que computar estimativas à condutividade nos cantos se torna uma atividade infrutífera quando valores da temperatura não envolvem informações nestas localizações e, por isso, a importância do comentário.

5.1.2 Problema semidiscreto

Usando a aproximação das derivadas espaciais descritas em (5.28) e (5.39), após negligenciar os erros de aproximação, construímos a versão semidiscreta do modelo (5.8)–(5.13) obtida através de CPM no agora problema de valor inicial (PVI) linear definido por

$$\begin{cases}\mathbf{C}\dfrac{d\mathbf{v}}{dt}(t) = \mathbf{A}\mathbf{v}(t) + \mathbf{S}(t), \quad t>0\\ \mathbf{v}(0) = \mathbf{u}^0\end{cases}, \tag{5.41}$$

em que $\mathbf{v} \in \mathbb{R}^{(n+1)^2}$ é um vetor em blocos dependente do tempo com componentes $\mathbf{v}_i$, $i = 0, 1, \ldots, n$, que servem como aproximações para $\mathbf{u}_i$; $\mathbf{u}^0$ contém os valores da condição inicial (5.13) u_0 ao longo da malha;

$$\mathbf{A} = \frac{1}{l_1^2}\mathcal{F} + \frac{1}{l_2^2}\mathcal{G} - \mathbf{Q}, \tag{5.42}$$

em que $\mathcal{F}$, $\mathcal{G}$ são as matrizes em blocos em (5.28) e (5.39), respectivamente; e

$$\mathbf{S}(t) = \frac{1}{l_1^2}\mathbf{S}^x(t) + \frac{1}{l_2^2}\mathbf{S}^y(t) + \mathbf{H}(t).$$

Nesta última equação, $\mathbf{S}^x(t)$ e $\mathbf{S}^y(t)$ tem entradas em blocos $\mathbf{S}_j^x$ e $\mathbf{S}_j^y$ descritas em (5.27) e (5.40) respectivamente, e $\mathbf{H}(t)$ acomoda valores do termo fonte $g(x, y, t)$ ao longo da malha, i.e., as entradas (em blocos) de $\mathbf{H}(t)$ são

$$\mathbf{H}_j(t) = [g(x_0, y_j, t), \ldots, g(x_n, y_j, t)]^T, \quad j = 0, 1, \ldots, n. \tag{5.43}$$

Finalmente, $\mathbf{C}$ e $\mathbf{Q}$ são matrizes diagonais contendo os valores de $c(x, y)$ e $q(x, y)$, respectivamente, como descritas abaixo

$$\mathbf{C} = \operatorname{diag}(c(x_0, y_0), \ldots, c(x_n, y_0), c(x_0, y_1), \ldots, c(x_n, y_1), \ldots, c(x_0, y_n), \ldots, c(x_n, y_n)) \tag{5.44}$$

e

$$\mathbf{Q} = \operatorname{diag}(q(x_0, y_0), \ldots, q(x_n, y_0), q(x_0, y_1), \ldots, q(x_n, y_1), \ldots, q(x_0, y_n), \ldots, q(x_n, y_n)), \tag{5.45}$$

ou seja, correspondem à ordenação dos valores de c e q em ordem lexicográfica ao longo da malha.

Observação 5.1. Embora além do escopo deste trabalho, note que obter um modelo semidiscreto similar a (5.41) para outros regimes de condutividade pode ser desenvolvido analogamente. O caso mais geral em 2D corresponde à *condutividade anisotrópica* [100], em que

$$K(x, y) = \begin{bmatrix} k_{11}(x, y) & k_{12}(x, y) \\ k_{21}(x, y) & k_{22}(x, y) \end{bmatrix},$$

contínua, com K definida positiva, isto é, $k_{11} > 0$ e $k_{11}k_{22} - k_{12}k_{21} > 0$, e $k_{11} \neq k_{22}$, $k_{12} \neq 0$, $k_{21} \neq 0$. Para este, temos que a equação (5.1) toma a forma

$$C\frac{\partial u}{\partial t} = \frac{\partial}{\partial x}\left(k_{11}\frac{\partial u}{\partial x} + k_{12}\frac{\partial u}{\partial y}\right) + \frac{\partial}{\partial y}\left(k_{21}\frac{\partial u}{\partial x} + k_{22}\frac{\partial u}{\partial y}\right) - qu + g$$

e cada uma das derivadas pode ser discretizada através de CPM com estratégias semelhantes às empreendidas para obter (5.28) e (5.39), que se iniciam em (5.21) e (5.30), resp. Teremos, ao final, que $\mathbf{A}$ em (5.42) conterá duas matrizes adicionais relacionadas às derivadas cruzadas de k_{12} e k_{21}. Por outro lado, veja que se $k_{12} = k_{21} = 0$, recaímos no

caso ortotrópico (5.2), já discutido. A simplificação final ocorre quando $k_{11} = k_{22} =: k$ e $k_{12} = k_{21} = 0$, chamado de *caso isotrópico*. Para este, é fácil ver que a discretização conduz à construção de $\mathcal{F}$ e $\mathcal{G}$ em (5.42) contendo apenas valores da condutividade k na malha. Grosseiramente, isto quer dizer que bastaria "trocar" k_{11} e k_{22} por k nestas matrizes para obtermos o problema semidiscreto relativo ao caso isotrópico. Isto implica, por exemplo, que toda a teoria desenvolvida neste capítulo é diretamente aplicada à diferentes cenários de condutividade com pequenas adaptações.

O modelo semidiscreto (5.41) pode ser manipulado analiticamente utilizando os autopares da matriz $\mathbf{A}$ ou numericamente através de uma numerosa quantidade de métodos para PVIs como os métodos de Euler (implícitos ou explícitos), regra do trapézio (Crank-Nicolson), métodos de Runge-Kutta, ou qualquer outro [29,39,55]. Neste trabalho, dadas as propriedades de estabilidade e boa taxa de convergência (quadrática), escolhemos aplicar o método de Crank-Nicolson (CN) [39]. Seja $N > 0$ um número natural e considere a discretização temporal na forma $t_i = i\Delta t$, para $i = 0, 1, \ldots, N$, em que $\Delta t = t_f/N$ é o passo no tempo, de modo que construímos uma malha uniformemente espaçada em $[0, t_f]$. Então, CN gera aproximações à solução no tempo t_i, denotada por $\mathbf{v}^{(i)}$, definidas implicitamente por

$$\mathbf{C}\mathbf{v}^{(i+1)} = \mathbf{C}\mathbf{v}^{(i)} + \frac{\Delta t}{2}\left(\mathbf{A}\mathbf{v}^{(i)} + \mathbf{S}(t_i) + \mathbf{A}\mathbf{v}^{(i+1)} + \mathbf{S}(t_{i+1})\right), \quad i = 0, 1, 2, \ldots,$$

com $\mathbf{A}$ introduzida em (5.42). A equação acima pode ser reescrita como

$$\mathbf{A}_m\mathbf{v}^{(i+1)} = \mathbf{A}_p\mathbf{v}^{(i)} + \frac{\Delta t}{2}\left[\mathbf{S}(t_i) + \mathbf{S}(t_{i+1})\right], \quad i = 0, 1, 2, \ldots, \tag{5.46}$$

em que $\mathbf{A}_m := \mathbf{C} - \frac{\Delta t}{2}\mathbf{A}$ e $\mathbf{A}_p := \mathbf{C} + \frac{\Delta t}{2}\mathbf{A}$. Consequentemente, para calcularmos soluções aproximadas em cada passo no tempo, precisamos resolver o sistema de equações lineares (5.46), preferencialmente da forma mais eficiente possível. Um meio de fazê-lo é explorando a estrutura esparsa de $\mathbf{A}_m$ através de alguma fatoração matricial como LU ou QR, por exemplo, e utilizar os fatores para resolver o sistema. Neste caso, a fatoração é calculada apenas uma vez no início das iterações CN e se mantém ao longo de todo o processo, uma vez que $\mathbf{A}_m$ é fixa. Mais precisamente, $\mathbf{A}_m$ muda com $\mathbf{A}$, que por sua vez apenas varia se mudarmos os valores da condutividade, únicos cada vez que resolvemos o problema direto. Aqui, por simplicidade nas implementações numéricas, resolvemos os sistemas lineares tanto desta parte quanto do restante do capítulo através da rotina mldivide ou \ (*backslash*) do Matlab.

5.2 PROBLEMA INVERSO

Baseado no método para o problema direto desenvolvido na seção anterior, vamos agora tratar do problema inverso de estimar as condutividades k_{11} e k_{22} a partir de medidas

de temperatura fornecidas. Para o tratamento deste problema, assumimos que as variáveis (valores da condutividade na malha) estão organizados em um vetor em blocos

$$\mathbf{k} = \begin{bmatrix} \mathbf{k}^{11} \\ \mathbf{k}^{22} \end{bmatrix} \in \mathbb{R}^{2(n+1)^2}, \tag{5.47}$$

cada bloco disposto em ordem lexicográfica. Desta forma, o objetivo desta seção é determinar valores de condutividade $\mathbf{k}$ a partir de medidas de temperatura

$$\widetilde{\mathsf{u}} = \mathsf{u} + e,$$

em que e denota um termo de perturbação desconhecido, proveniente de erros de medição, imprecisões, arredondamentos, por exemplo, e u é um vetor que contém valores exatos de temperatura na malha de Chebyshev nos tempos $t_k,\ k = 1, \ldots, N$,

$$\mathsf{u} = \begin{bmatrix} \mathbf{u}(t_1) \\ \vdots \\ \mathbf{u}(t_N) \end{bmatrix} \in \mathbb{R}^{N(n+1)^2}, \tag{5.48}$$

em que $\mathbf{u}(t_k) \equiv \mathbf{u}_k \in \mathbb{R}^{(n+1)^2}$ tem entradas $u_\ell^{(t_k)} \doteq u(x_i, y_j, t_q)$ com ℓ definido em (5.15) para que (x_i, y_j) sigam a ordem lexicográfica. Comentários similares se encaixam ao vetor $\widetilde{\mathsf{u}}$ quanto à ordenação das entradas. Desta forma, observe que estamos assumindo, neste momento por simplificação na explanação, valores de temperatura disponíveis em toda a malha de Chebyshev. Os casos que não satisfazem esta hipótese serão comentados a frente.

Na prática, como não temos acesso à função exata de temperatura, aproximamos os seus valores através da solução do problema direto. Desta forma, seguindo a notação de (5.48), definimos por $\mathsf{u}(\mathbf{k})$ o vetor que acomoda os valores obtidos para temperatura através do problema direto para dado $\mathbf{k}$. Desta forma, formulamos finalmente o problema inverso de estimar as condutividades através da minimização do funcional de mínimos quadrados não linear

$$\phi(\mathbf{k}) = \frac{1}{2}\|\mathsf{u}(\mathbf{k}) - \widetilde{\mathsf{u}}\|_2^2 \doteq \frac{1}{2}\sum_{k=1}^{N}\sum_{j=0}^{n}\sum_{i=0}^{n} |u(x_i, y_j, t_k, \mathbf{k}) - \widetilde{u}(x_i, y_j, t_k)|^2, \tag{5.49}$$

em que $\mathsf{u}(x_i, y_j, t_k, \mathbf{k})$ resolve o problema direto para dados k_{11} e k_{22} e, portanto, estamos em busca que uma solução

$$\mathbf{k}^* = \underset{\mathbf{k}\in\mathbb{R}^{2(n+1)^2}}{\operatorname{argmin}}\ \phi(\mathbf{k}).$$

Entre a miríade de métodos disponíveis na literatura para problemas de minimização desta forma, destacamos os métodos que usam derivadas e os que não usam (do inglês *derivative-free methods*). O primeiro envolve, por exemplo, a regularização de Tikhonov, métodos de região de confiança, métodos de busca linear não monótona, método de Gauss-Newton, método de gradientes conjugados não linear, método de Levenberg-Marquardt,

entre outros [51,93,115,116]. Exemplos de métodos sem derivadas podem ser encontrados em [97,107]. Neste trabalho, como forma de atestar as capacidades da técnica desenvolvida no último capítulo, resolveremos (5.49) através de LMMSS como descrito no Algoritmo 4.1, com a matriz de *scaling* a ser escolhida de acordo com o exemplo considerado, como veremos a seguir. Adicionalmente, em termos comparativos, em alguns exemplos utilizamos LMM clássico e um método de região de confiança que preserva positividade nas variáveis, disponível através da rotina lsqnonlin do Matlab.

No caso de LMMSS, o processo de minimização é resolvido através para dois tipos de matrizes de *scaling* singular, a depender do exemplo em questão e, especialmente, da quantidade de variáveis envolvidas. De fato, nosso exemplos buscam incluir as versões discretas dos operadores de derivação de primeira e segunda ordem, respectivamente L_1 e L_2, como descritos em (2.6), embora isto não possa ser feito diretamente. Um dos empecilhos é a ordem lexicográfica utilizada, uma vez que precisamos levar em conta a reordenação dos dados. A ideia é aplicarmos os operadores de forma horizontal e vertical na malha. No caso ortotrópico com $\mathbf{k}$ organizado de acordo com (5.47), tomamos

$$L := I_2 \otimes \mathcal{L}_i = \begin{bmatrix} \mathcal{L}_i & \mathbf{0} \\ \mathbf{0} & \mathcal{L}_i \end{bmatrix}, \quad i = 1, 2, \tag{5.50}$$

com

$$\mathcal{L}_i = \begin{bmatrix} I_{n+1} \otimes L_i(n+1) \\ L_i(n+1) \otimes I_{n+1} \end{bmatrix}, \quad i = 1, 2, \tag{5.51}$$

em que $\otimes$ representa o produto de Kronecker. Veja que L como em (5.50) atua da seguinte forma nas variáveis:

$$L\mathbf{k} = \begin{bmatrix} \mathcal{L}_i & \mathbf{0} \\ \mathbf{0} & \mathcal{L}_i \end{bmatrix} \begin{bmatrix} \mathbf{k}^{11} \\ \mathbf{k}^{22} \end{bmatrix} = \mathcal{L}_i\mathbf{k}^{11} + \mathcal{L}_i\mathbf{k}^{22}, \quad i = 1, 2,$$

diretamente aplicando $\mathcal{L}_i$ em $\mathbf{k}^{11}$ e $\mathbf{k}^{22}$. Agora, pensando na ordenação das variáveis, o produto tensorial em (5.51) tem por objetivo aplicar o efeito de L_i em cada "linha" e "coluna" da malha espacial, buscando introduzir a suavidade destes operadores nas soluções computadas. Mais especificamente, o primeiro bloco de $\mathcal{L}_i$, isto é, $I_{n+1} \otimes L_i(n+1)$, faz com que L_i seja introduzida horizontalmente na malha: nos pontos formados para cada y_j fixo, variando x_i. Analogamente, o segundo bloco em (5.51) é aplicado verticalmente.

Um caso mais simples acontece com a condutividade isotrópica ($k_{11} = k_{22}$), em que temos apenas $(n+1)^2$ variáveis no vetor $\mathbf{k}$ (somente um dos blocos em (5.47) é mantido), de forma que basta definir $L = \mathcal{L}_i$, $i = 1, 2$. Comentários análogos aos tecidos no caso ortotrópico são válidos aqui. Outros casos particulares devem ser estudados de acordo com o uso, como faremos em um exemplo adiante. De qualquer forma, ressaltamos que a utilização de L_1 e L_2 não é arbitrária, seguindo a ideia de introduzir suavidade nas soluções. Ambos os operadores são amplamente utilizadas, por exemplo, em problemas de reconstrução de imagens [13] e, no caso específico de problemas em transferência de

calor, situações que mostram ganho significativo com matrizes desta forma podem ser encontrados em [8, 26, 73].

Observação 5.2. Utilizamos, nas hipóteses de convergência para LMMSS, que L é matriz de posto completo com número de linhas menor ou igual ao de colunas, o que não ocorre, por exemplo, com $\mathcal{L}_i$ em (5.51). Assim, se $L \in \mathbb{R}^{m\times n}$, com $m \geq n$, e tomando $p = \text{posto}(L)$, segue claramente que $p \leq n$. Além disso, escrevendo $L = U\Sigma V^T$ pela SVD, com

$$\Sigma = \begin{bmatrix} C & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}, \quad C = \text{diag}(\sigma_1, \ldots, \sigma_p) \in \mathbb{R}^{p\times p},$$

segue que, para $x \in \mathbb{R}^n$ e da ortonormalidade de U,

$$\|Lx\|_2 = \|U\Sigma V^T x\|_2 = \|\Sigma V^T x\|_2 = \left\| \begin{bmatrix} C & \mathbf{0} \end{bmatrix} V^T x \right\|_2.$$

Por (4.7) ou mesmo (4.37), que dispõem sobre a construção das direções de descida para LMMSS, o efeito de L ocorre no termo $\|Lx\|_2$. Portanto, podemos substituir L por $\tilde{L} := \begin{bmatrix} C & \mathbf{0} \end{bmatrix} V^T$ sem perda já que $\|Lx\|_2 = \|\tilde{L}x\|_2$. Visto de outra forma, L atua em (4.2) na forma do produto L^TL, que também coincide com $\tilde{L}^T\tilde{L}$. Ou seja, mesmo que L tenha mais linhas que colunas, podemos substituí-la por uma matriz $\tilde{L} \in \mathbb{R}^{p\times n}$, de posto completo, $p \leq n$, que atua igualmente à matriz de *scaling* original.

5.2.1 O problema de sensitividade

Tanto lsqnonlin quanto LMMSS trabalham iterativamente e precisam, internamente, do cálculo da matriz Jacobiana de ϕ no iterado $\mathbf{k}^{(j)}$. No contexto aqui abordado, a Jacobiana é solução do chamado *problema de sensibilidade* (*sensitivity problem*), que busca encontrar as derivadas da temperatura com respeito às variáveis $k_{11}(x_i, y_j)$, $k_{22}(x_i, y_j)$ em cada iteração. Organizamos então a matriz da seguinte forma:

$$\mathbf{J}(\mathbf{k}^{(j)}) = \begin{bmatrix} \mathbf{J}_1(\mathbf{k}^{(j)}) \\ \vdots \\ \mathbf{J}_N(\mathbf{k}^{(j)}) \end{bmatrix}, \tag{5.52}$$

em que o k-ésimo bloco $\mathbf{J}_k$ denota a Jacobiana de $u(x, y, t, \mathbf{k})$ no tempo t_k (também conhecida como *matriz de sensibilidade*), ou equivalentemente de $\mathsf{u}(\mathbf{k}, t_k)$, e definida por

$$\mathbf{J}_k = \left[\frac{\partial \mathsf{u}}{\partial \mathbf{k}_1}(\mathbf{k}, t_k), \ldots, \frac{\partial \mathsf{u}}{\partial \mathbf{k}_{2(n+1)^2}}(\mathbf{k}, t_k) \right]. \tag{5.53}$$

Note que $\mathbf{J}_k$ é uma matriz de ordem $(n+1)^2 \times 2(n+1)^2$ e que temperaturas em $k = 3$ níveis temporais são suficientes para $\mathbf{J}(\mathbf{k}^{(j)})$ ser sobredeterminada.

Como dito, o problema de sensibilidade busca analisar como que a temperatura $u(x, y, t)$ muda de acordo com pequenas variações de $k_{11}(x, y)$ ou $k_{22}(x, y)$. No nosso

contexto, resolver o problema de sensibilidade é equivalente a determinar derivadas parciais de $\mathsf{u}(\mathbf{k}, t_k)$, $k = 1, \dots N$, com respeito a $k_{11}(x, y)$ ou $k_{22}(x, y)$ ou, equivalentemente, a determinar a matriz de sensibilidade $\mathbf{J}_k$. Pontuamos que os valores da condutividade nos pontos da malha (x_i, y_j) são indexados a $\mathbf{k}_\ell^{11}$ ou $\mathbf{k}_\ell^{22}$ respectivamente, com ℓ representando a ordenação lexicográfica de (i, j) através de (5.15).

Para calcular a matriz de sensibilidade, note que por $\mathsf{u}(\mathbf{k}, t)$ resolver o PVI (5.41), então o ℓ-ésimo vetor coluna de $\mathbf{J}_k(\mathbf{k})$ é solução de

$$\begin{cases} \mathbf{C}\dfrac{\partial}{\partial t}\left(\dfrac{\partial \mathsf{u}}{\partial \mathbf{k}_\ell}(\mathbf{k}, t)\right) = \mathbf{A}\dfrac{\partial \mathsf{u}}{\partial \mathbf{k}_\ell}(\mathbf{k}, t) + \dfrac{\partial \mathbf{A}}{\partial \mathbf{k}_\ell}\mathsf{u}(\mathbf{k}, t) \\ \dfrac{\partial \mathsf{u}}{\partial \mathbf{k}_\ell}(\mathbf{k}, 0) = 0 \end{cases}, \quad \ell = 1, 2, \dots, 2(n+1)^2, \tag{5.54}$$

que também é um problema de valor inicial e pode ser resolvido com algum procedimento análogo ao descrito anteriormente para (5.41). De fato, aqui utilizaremos o método de Crank-Nicolson para tal fim, como feito no modelo semidiscreto. Para tanto, observe que as derivadas parciais de $\mathbf{A}$ com respeito a $\mathbf{k}_\ell$ satisfazem

$$\frac{\partial \mathbf{A}}{\partial \mathbf{k}_\ell} = \begin{cases} \dfrac{1}{l_1^2}\dfrac{\partial \mathcal{F}}{\partial \mathbf{k}_\ell}, & \ell = 1, \dots, (n+1)^2, \\ \dfrac{1}{l_2^2}\dfrac{\partial \mathcal{G}}{\partial \mathbf{k}_\ell}, & \ell = (n+1)^2 + 1, \dots, 2(n+1)^2, \end{cases}$$

e que são simples de calcular dada a estrutura esparsa de $\mathcal{F}$ e $\mathcal{G}$. Como resultado, o "termo fonte" em (5.54) é também altamente esparso, um fato que pode ser explorado nos cálculos práticos. Exemplos de cálculo de matrizes similares podem ser encontradas em [26], embora a construção seja relativamente simples. Para concluir, observe que calcular a matriz de sensibilidade $\mathbf{J}_k$ demanda que a solução do PVI (5.54) seja efetuada nos passos temporais $t_1, \dots, t_k$.

5.2.2 Reconstruções utilizando dados incompletos

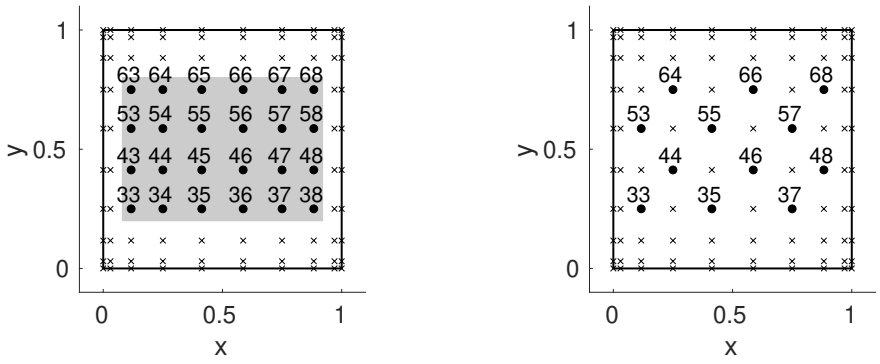
Na prática, muitas vezes por razões técnicas, ao invés de possuirmos medidas da temperatura em toda a malha, estes valores funcionais podem estar disponíveis apenas em uma região no interior de Ω ou em um pequeno conjunto de pontos associado, por exemplo, à localização de sensores. Esta subseção, portanto, se interessa em apresentar formas de reconstruir a condutividade em cenários com dados incompletos como explicitado na Figura 5.3: situações em que as medidas são possíveis em um conjunto no interior de Ω ou em um subconjunto de pontos da malha. De forma grosseira, a ideia é reformular o problema inverso para aceitar as informações que temos acesso e efetuar o processo de minimização dessa forma, com a capacidade de obter reconstruções em toda a malha. O procedimento é descrito a seguir.

De início, introduziremos alguma notação necessária para reformularmos a função objetivo em (5.49). Considere então que as medidas são fornecidas em índices ℓ_l,

Figura 5.3 – Dois possíveis cenários de medições para $\Omega = (0,1) \times (0,1)$ e $n = 9$. Somente valores de temperatura nos pontos marcados com círculos pretos (●) são considerados disponíveis, para os quais a ordenação lexicográfica é destacada.

(a) Região no interior de Ω (em cinza).

(b) Medidas fornecidas em apenas alguns pontos.



Fonte – Boos, Bazán e Luchesi [25].

$l = 1, \ldots, q$, os quais assumimos, no momento, coincidir com pontos da malha em ordem lexicográfica, isto é, $\{\ell_1, \ell_2, \ldots, \ell_q\} \subseteq \{1, 2, \ldots, (n+1)^2\}$. Como ilustração, na Figura 5.3b temos $q = 12$, $n = 9$, e então $\{\ell_1, \ell_2, \ldots, \ell_{12}\} = \{33, 35, 37, 44, 46, 48, 53, 55, 57, 64, 66, 68\} \subseteq \{1, \ldots, 100\}$. Tome os valores da temperatura nas localizações ℓ_l no passo temporal t_k pela notação $\widetilde{u}_{\ell_l}^{(k)}$ e organize tais dados disponíveis em um vetor em blocos $\breve{\widetilde{\mathbf{u}}}$ cujas componentes do bloco são $\breve{\widetilde{\mathbf{u}}}_k \in \mathbb{R}^q$, $k = 1, \ldots, N$, definidas por

$$\breve{\widetilde{\mathbf{u}}}_k = [u_{\ell_1}^{(k)}, u_{\ell_2}^{(k)}, \ldots, u_{\ell_q}^{(k)}]^T.$$

Então, tomando os valores de temperatura $u(x_i, y_j, t_k, \mathbf{k})$ associados às localizações prescritas ℓ_l arranjados em um vetor $\breve{\mathbf{u}}(\mathbf{k})$ como feito com $\widetilde{u}_{\ell_l}^{(k)}$ em $\breve{\widetilde{\mathbf{u}}}$, a estimativa da condutividade pode ser atingida pela minimização da equivalente à função objetivo (5.49) que considere $\mathbf{u}(\mathbf{k})$ e $\widetilde{\mathbf{u}}$ substituídos por $\breve{\mathbf{u}}(\mathbf{k})$ e $\breve{\widetilde{\mathbf{u}}}$, respectivamente. Agora, parte essencial da busca por um mínimo depende da matriz de sensibilidade associada às localizações prescritas, as quais devem ser calculadas de acordo também, como descreveremos a seguir. Seja $\breve{\mathbf{J}}_k$ a matriz Jacobiana da temperatura u nos pontos disponíveis (pensando nas medidas restritas) no tempo t_k. Como $\breve{\mathbf{J}}_k$ concentra a variação da temperatura apenas em alguns pontos como função das condutividades, a existência de uma relação próxima entre $\breve{\mathbf{J}}_k$ e $\mathbf{J}_k$ dada em (5.53) é bastante natural. De fato, um olhar cauteloso à esta relação mostra que para calcular $\breve{\mathbf{J}}_k$ é suficiente extrair as linhas de $\mathbf{J}_k$ associadas aos índices ℓ_l. Utilizando notação do Matlab, isto pode ser facilmente feito através de

$$\breve{\mathbf{J}}_k = \mathbf{J}_k(\ell_1 : \ell_q, :). \tag{5.55}$$

A matriz de sensibilidade das temperaturas tomadas nas localizações prescritas para todos os passos no tempo é então uma matriz em blocos $\breve{\mathbf{J}}$ definida similarmente a (5.52). Levando

em conta estas modificações, o processo de minimização (com LMMSS, por exemplo) pode ser aplicado como descrito anteriormente. Outros detalhes neste tipo de recuperação com dados restritos podem ser encontrados, e.g. em [8].

Para fechar, é preciso falar sobre o caso em que as medidas fornecidas não coincidem com pontos da malha. Simplificando, é suficiente mencionar que a abordagem é essencialmente a mesma, com a diferença de que para a minimização da função objetivo as temperaturas calculadas na malha de Chebyshev devem ser interpoladas a fim de obtermos valores de temperatura nas localizações desejadas. Idem para as colunas da matriz Jacobiana. Veja que esta abordagem é a mesma feita acima, em que, quando os pontos coincidem com a malha, a interpolação consiste apenas de selecionar estes pontos (ou colunas referentes a estes pontos no caso da matriz Jacobiana). Retomaremos esta discussão com um exemplo que se baseia em dados experimentais mais adiante.

5.3 RESULTADOS NUMÉRICOS

No restante deste capítulo, abordaremos exemplos que visam exibir a performance de LMMSS em diferentes cenários, em algumas situações comparando com outras técnicas como LMM e região de confiança através da rotina lsqnonlin quando aplicados aos mesmos problemas de recuperação de condutividade. Consideramos a relação entre os dados recebidos $\widetilde{\mathsf{u}}$ e os dados exatos u de modo que

$$\|\widetilde{\mathsf{u}} - \mathsf{u}\|_2 = \mathrm{NL}\|\mathsf{u}\|_2, \tag{5.56}$$

para NL dado representando o *nível de ruído* (*noise level*). As iterações são, então, paradas usando o princípio da discrepância de Morozov [95] como técnica de regularização para controlar a propagação de ruído, de modo que o processo é terminado ao primeiro índice j tal que

$$\|\mathsf{u}(\mathbf{k}^{(j)}) - \widetilde{\mathsf{u}}\|_2 \lessapprox \tau\|e\|_2, \quad \widetilde{\mathsf{u}} = \mathsf{u} + e,$$

com τ um parâmetro de folga escolhido aqui para ser $\tau = 1.1$. Ressaltamos que o vetor de ruído e contém dados com distribuição uniforme numericamente gerados através da rotina randn do MATLAB. No caso de $\mathrm{NL} = 0$, utilizamos como critério de parada a variação relativa dos iterados $\mathbf{k}^{(j)}$ produzidos, isto é, paramos quando

$$\frac{\|\mathbf{k}^{(j)} - \mathbf{k}^{(j-1)}\|_2}{\|\mathbf{k}^{(j)}\|_2} < \varepsilon, \quad j \geq 1,$$

com $\varepsilon = 5 \times 10^{-4}$ escolhido desta forma em todos os exemplos numéricos.

Para informar a qualidade das reconstruções utilizamos erros relativos definidos por

$$\mathrm{RE}(\mathbf{k}^{(j)}) = \frac{\|\mathbf{k}^{(j)} - \mathbf{k}^{\mathrm{exact}}\|_2}{\|\mathbf{k}^{\mathrm{exact}}\|_2}, \tag{5.57}$$

em que $\mathbf{k}^{\mathrm{exact}}$ denota o vetor de condutividades de k_{11} ou k_{22} e $\mathbf{k}^{(j)}$ representa a estimativa obtida na última iteração (quando a parada foi atingida). No caso, $\mathrm{RE}(\mathbf{k}_{11}^{(j)})$ e $\mathrm{RE}(\mathbf{k}_{22}^{(j)})$

representam os erros relativos entre $\mathbf{k}_{11}^{(j)}$ e $\mathbf{k}_{22}^{(j)}$ e a solução exata respectiva. Analogamente, TRE (*temperature reconstruction error*) consiste do erro relativo entre a temperatura exata e a reconstruída para $\mathbf{k}^{(j)}$ ao longo de toda a malha definido por

$$\text{TRE} = \frac{\|\mathsf{u}(\mathbf{k}^{(j)}) - \mathsf{u}\|_2}{\|\mathsf{u}\|_2}.$$

Para os casos de medidas restritas como descrito na Subseção 5.2.2, apresentamos também RTRE (*restricted temperature reconstruction error*), que se baseia no erro relativo entre a temperatura reconstruída obtida nas localizações prescritas e a exata,

$$\text{RTRE} = \frac{\|\breve{\mathsf{u}}(\mathbf{k}^{(j)}) - \breve{\mathsf{u}}\|_2}{\|\breve{\mathsf{u}}\|_2}.$$

Além disso, como alguns testes são apresentados após uma bateria de resoluções com ruídos diferentes de mesma ordem, é costumeiro reportarmos os erros relativos médios obtidos e também o número máximo de iterações (MI) utilizadas em cada uma das instâncias.

Dos parâmetros, em todos os exemplos consideramos $N = 10$ estágios no tempo igualmente espaçados em $[0, t_f]$ e $n + 1 = 16$ pontos na malha de Chebyshev em cada direção, totalizando 256 pontos em Ω. Note que estamos, então, operando com matrizes de ordem 256, em sua maioria esparsas, o que reduz o esforço numérico e pode sugerir a implantação de técnicas que aproveitem esta propriedade. O maior custo se encontra no cálculo das matrizes Jacobianas através do problema de sensibilidade, conduzindo a $2(n+1)^2$ PVIs a serem resolvidos em cada iteração, no caso ortotrópico. Mesmo assim, envolve matrizes de baixa ordem, esparsas, com capacidades de reconstrução significativas, como os exemplos a seguir buscam enumerar.

5.3.1 Exemplo 1: condutividade isotrópica

Extraído de Mahmood e Lesnic [86], este exemplo conta com solução procurada $K(x, y)$ representando uma condutividade isotrópica, isto é,

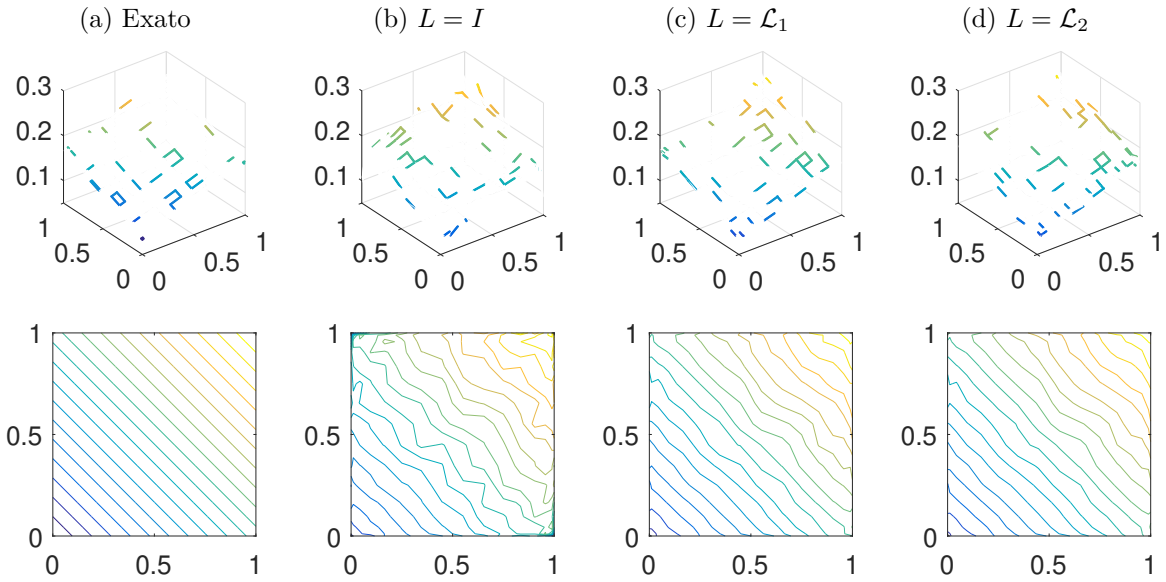
$$k_{11}(x, y) = k_{22}(x, y) := k(x, y) = \frac{1 + x + y}{12},$$

sendo o problema definido no domínio espacial $\Omega = (0, 1) \times (0, 1)$, com tempo final $t_f = 1$. Dos outros parâmetros, temos:

$$g(x, y, t) = 0, \quad f_1(y, t) = -1, \quad f_2(y, t) = 1,$$
$$f_3(x, t) = -1, \quad f_4(x, t) = 1, \quad \text{e} \quad h_i \equiv 0, \ i = 1, \ldots, 4,$$

além da capacidade térmica $C(x, y) = 1$ e condição inicial $u_0(x, y) = 0$. É preciso ressaltar, porém, que não temos uma fórmula explícita para a função temperatura $u(x, y, t)$, a qual forneceria também dados ao problema inverso de reconstruir a condutividade. Para contornar esta dificuldade, aplicamos uma rotina do método de elementos finitos (FEM)

Figura 5.4 – Comparativo entre iterados de LMM clássico e LMMSS.



Fonte – Boos, Luchesi e Bazán [26].

ao modelo (5.1)–(5.7), utilizando todos os dados conhecidos (inclusive a condutividade), com intuito de gerar aproximações para a temperatura u. Utilizamos de uma malha fina para que as aproximações geradas sejam de qualidade, as quais então interpolamos à malha de Chebyshev, que enfim é utilizada como solução exata na técnica que propomos via CPM. Para maiores detalhes neste procedimento, veja [26].

Os resultados obtidos com 0.1% de ruído nos dados (além das possíveis imprecisões provenientes da reconstrução numérica com FEM) para $L = I$ e para $L = \mathcal{L}_i$ como em (5.51) são apresentados na Figura 5.4 e na Tabela 5.1, com iterado inicial $k^0(x, y) = 1/4$. Observe que as matrizes de *scaling* singulares apresentam vantagem na reconstrução tanto em qualidade quanto custo computacional, reduzindo a quantidade de iterações necessárias. Mais ainda, tais matrizes buscam preservar, como comentado previamente, que a solução seja construída com vetores suaves do núcleo de L, contribuindo na coesão visual presente nas Figuras 5.4c e 5.4d. De certa forma, podemos entender que as variáveis $\mathbf{k}^{(j)}$ são construídas respeitando a relação de suavidade imposta pelas matrizes de *scaling* e não de forma individual como para LMM clássico, que gera então os "picos" presentes na Figura 5.4b. Adicionalmente, a Tabela 5.1 confirma a qualidade significantemente superior das soluções encontradas por LMMSS, exigindo menos iterações que LMM.

Tabela 5.1 – Resultados encontrados em uma instância para diferentes matrizes de *scaling*.

Matriz de *scaling*	I	$\mathcal{L}_1$	$\mathcal{L}_2$
$\mathrm{RE}(\mathbf{k}^{(j)})$	0.0946	0.0128	0.0134
Iterações	7	4	4

Fonte – Adaptado de Boos, Luchesi e Bazán [26].

5.3.2 Exemplo 2: condutividade ortotrópica

Este exemplo é baseado no trabalho de Cao, Lesnic e Colaço [32], adaptado ao modelo (5.1)–(5.7), definido em $\Omega = (0,1)\times(0,1)$ com tempo final $t_f = 1$. Possui descrição completa dos componentes do problema direto, a saber:

$$u(x,y,t) = e^{-t}(\sin(\pi x)\sin(\pi y) + (\pi+1)(x+y) + 1), \quad \text{em } (0,1)\times(0,1)\times[0,t_f],$$

e dados de entrada $h_i = 1,\ i = 1,\ldots,4$, $C(x,y) = 1$, $q(x,y) = 0$,

$$\begin{aligned}
f_1(y,t) &= -\frac{1+y}{12}e^{-t}(\pi\sin(\pi y) + \pi + 1) + e^{-t}((\pi+1)y + 1),\\
f_2(y,t) &= \frac{2+y}{12}e^{-t}(-\pi\sin(\pi y) + \pi + 1) + e^{-t}(-\sin(\pi y) + (\pi+1)(1+y) + 1),\\
f_3(x,t) &= -\frac{1+0.5x}{12}e^{-t}(\pi\sin(\pi x) + \pi + 1) + e^{-t}((\pi+1)x + 1),\\
f_4(x,t) &= \frac{2+0.5x}{12}e^{-t}(-\pi\sin(\pi x) + \pi + 1) + e^{-t}(-\sin(\pi x) + (\pi+1)(1+x) + 1),
\end{aligned}$$

termo fonte

$$\begin{aligned}
g(x,y,t) = &-e^{-t}(\sin(\pi x)\sin(\pi y) + (\pi+1)(x+y) + 1) - \frac{e^{-t}}{12}[2\pi + 2 + \pi\sin(\pi(x+y))]\\
&+ \frac{\pi^2 e^{-t}}{12}(2 + 1.5x + 2y)\sin(\pi x)\sin(\pi y)
\end{aligned}$$

e, para comparação, condutividades (ortotrópicas) exatas

$$k_{11}(x,y) = \frac{1+x+y}{12} \quad \text{e} \quad k_{22}(x,y) = \frac{1+0.5x+y}{12}. \tag{5.58}$$

Buscamos resolver o problema para diferentes níveis de ruído, repetindo o mesmo cenário 30 vezes, todos considerando como chute inicial $k_{11}^0(x,y) = k_{22}^0(x,y) = 1/4$. A média dos resultados obtidos se encontra na Tabela 5.2, que novamente traz luz às boas qualidades de aproximação de LMMSS, produzindo erros relativos bastante inferiores aos reportados por LMM. Mais ainda, veja a natureza mal posta do problema, uma vez que os erros de reconstrução da temperatura (TRE) são praticamente iguais para ambos os métodos. Este fato reforça que reduzir eficientemente o valor da função objetivo não implica na aproximação de $\mathbf{k}^{(j)}$ da solução exata, que pode então levar a erros de alta ordem, no caso de LMM, por exemplo. Já para LMMSS, como as matrizes $I_2 \otimes \mathcal{L}_i$ são provenientes de operadores de derivação, é natural esperar que esta propriedade seja também introduzida nos iterados produzidos (na forma de construção suave pelos vetores do núcleo dos operadores de derivação), permitindo aproximações suaves também, e de maior qualidade.

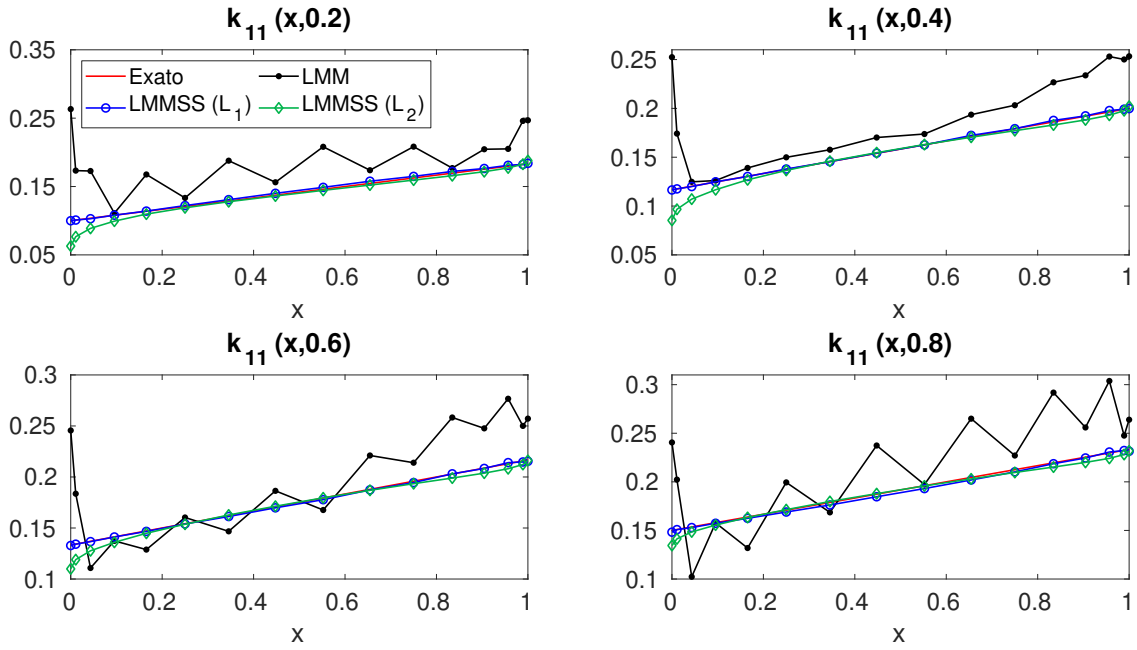
Em paralelo, na Figura 5.5 apresentamos as médias obtidas para k_{11} em diferentes ordenadas y fixas comparando os valores obtidos e os exatos, para NL = 0.001 (isto é, 0.1% de ruído relativo nos dados). Os gráficos reforçam a noção de que LMMSS produz soluções próximas da exata, enquanto que LMM apresenta altas oscilações e reconstruções distantes

Tabela 5.2 – Resultados médios encontrados para alguns valores de ruído, com repetição de 30 vezes em cada cenário.

Ruído	**Método**	L	RE($\mathbf{k}_{11}^{(j)}$)	RE($\mathbf{k}_{22}^{(j)}$)	TRE	MI
0%	**LMM**	I	0.2937	0.3698	0.0000	13
	LMMSS	$I_2 \otimes \mathcal{L}_1$	0.0195	0.0154	0.0000	6
	LMMSS	$I_2 \otimes \mathcal{L}_2$	0.0291	0.0127	0.0000	8
0.1%	**LMM**	I	0.3996	0.5211	0.0063	4
	LMMSS	$I_2 \otimes \mathcal{L}_1$	0.0218	0.0185	0.0003	3
	LMMSS	$I_2 \otimes \mathcal{L}_2$	0.0611	0.1138	0.0100	2
1%	**LMM**	I	0.5100	0.6851	0.0398	1
	LMMSS	$I_2 \otimes \mathcal{L}_1$	0.0388	0.0318	0.0022	2
	LMMSS	$I_2 \otimes \mathcal{L}_2$	0.1446	0.2024	0.0237	1

Fonte – o autor, 2022.

Figura 5.5 – Resultados médios para $k_{11}(x, y)$ para alguns valores de y fixos e NL = 0.001. Nas legendas, L_1 e L_2 representam, respectivamente, as matrizes $I_2 \otimes \mathcal{L}_1$ e $I_2 \otimes \mathcal{L}_2$.



Fonte – o autor, 2022.

do desejado. Resultados similares estão presentes em k_{22}, omitido aqui. Entre as opções de *scaling* diferentes para LMMSS como presente na Tabela 5.2, vemos que $L = I_2 \otimes \mathcal{L}_1$ apresenta ganho em relação ao caso com segunda derivada discreta da ordem de alguns pontos percentuais, mais evidente com o crescimento de NL. Visualmente, introduzir o operador de primeira derivada discreta junto a LMMSS conduz a soluções que basicamente coincidem com a condutividade exata, como vemos na figura. É importante ressaltar este ponto pois, como as soluções exatas são lineares, o efeito do uso de operadores de segunda

ordem como $L = I_2 \otimes \mathcal{L}_2$ tende a ser reduzido ou mesmo não atuar. Para este caso, portanto, a escolha do operador causa diferença nas soluções obtidas e, tomar $L = I_2 \otimes \mathcal{L}_1$ produz melhores reconstruções.

Em resumo, podemos ver que a técnica através de LMMSS com matrizes de *scaling* emulando operadores de derivação apresenta significativa melhora nas soluções obtidas quando comparado com LMM, reduzindo a quantidade de iterações necessárias para produzir aproximações vantajosas. Em palavras, estas escolhas de L dão prioridade a escrever os iterados contendo componentes suaves, resultando na coesão presente nas figuras. Veja que tanto este exemplo quanto o da subseção anterior possuem condutividades contínuas e diferenciáveis, uma informação que LMM com $L = I$ não carrega ao processo iterativo, mas que LMMSS tira proveito pelas escolhas de L como em (5.50).

5.3.3 Exemplo 3: condutividade ortotrópica com medidas restritas

Consideremos agora um exemplo para mostrar a efetividade do método proposto em situações com dados incompletos. As reconstruções são baseadas em dados acessíveis em alguns pontos da malha com localizações ℓ_l respeitando duas estratégias de medição, MS1 e MS2, descritas abaixo. Além destas, por comparação, a estratégia chamada de *full grid* (malha completa) corresponde às reconstruções utilizando todos os pontos disponíveis, como feito nos exemplos anteriores.

- **MS1**: Assume dados disponíveis a todos os pontos de Chebyshev localizados dentro de uma região retangular de Ω, similar à Figura 5.3a. Esta hipótese pode ocorrer na prática quando medidas são capturadas de forma quase contínua, e.g., utilizando sensores de temperatura infravermelhos [28]. Os resultados numéricos que utilizamos correspondem a dados de $q = 72$ posições dentro de $[0.02, 0.98] \times [0.2, 0.8] \subseteq \Omega$.

- **MS2:** Considera dados em apenas alguns poucos pontos da malha, neste caso relativamente bem distribuídos em Ω e com localizações que buscam similar posições de sensores pontuais em experimentos práticos. Aqui, consideramos $q = 20$ localizações distribuídas similar à Figura 5.3b.

Os dados para este exemplo foram criados para [25], e considera o problema de condução de calor em $\Omega = (0, 1) \times (0, 1)$, $t \in [0, 1]$, com

$$C(x, y) = \frac{1}{1 + x^2 + 2y^2}, \quad k_{11}(x, y) = \frac{3x^2 + y^2 + 1}{2}, \quad k_{22}(x, y) = \frac{x^2 + 3y^2 + 1}{2}, \quad (5.59)$$

funções de transferência de calor constantes, $h_i = 1$, $i = 1, \ldots, 4$, $q(x, y) = 0$,

$$
\begin{aligned}
f_1(y,t) &= -\pi e^{-\pi t}\cos(\pi t)\cos(\pi y)\left(\frac{y^2+1}{2}\right)(1+2y^2),\\
f_2(y,t) &= -\pi e^{-\pi t}\cos(\pi t)\cos(\pi y)(4+y^2)(1+y^2),\\
f_3(x,t) &= e^{-\pi t}\cos(\pi t)\sin(\pi x)(1+x^2),\\
f_4(x,t) &= -e^{-\pi t}\cos(\pi t)\sin(\pi x)(3x^2+11),
\end{aligned}
$$

e termo fonte $g(x, y, t)$ escolhido de modo que a solução analítica ao problema direto seja

$$u(x, y, t) = e^{-\pi t}\cos(\pi t)(1 + x^2 + 2y^2)\sin(\pi x)\cos(\pi y).$$

Estimaremos as condutividades k_{11} e k_{22} a partir de dados com ruído através de LMMSS nas três restrições de medição: *full grid*, MS1 e MS2. Neste exemplo, os chutes iniciais são escolhidos como vetores com dados respeitando a distribuição normal na forma

$$\mathbf{k}^{(0)} \sim N(a, \vartheta) \in \mathbb{R}^{2(n+1)^2}, \tag{5.60}$$

em que a e ϑ são a média e o desvio padrão da distribuição normal, respectivamente, gerada através de randn. Para esta seção, $a = 1$ e $\vartheta = 0.015$, significando que $\mathbf{k}^{(0)}$ é um vetor normalmente distribuído em torno da constante 1, consideravelmente distante das soluções exatas em (5.59). Esta escolha de iterando inicial busca simular situações de chutes iniciais randômicos, comum em algumas implementações, que sugerem desconhecimento ainda maior do que se procura como solução exata. Além disso, dado que as condutividades exatas k_{11} e k_{22} são quadráticas, optamos por escolher L como em (5.50) com o operador de segunda ordem, isto é, $L = I_2 \otimes \mathcal{L}_2$.

Como antes, exibimos as médias de erros relativos das reconstruções obtidas para três níveis de ruído, NL = 0.001, 0.015 e 0.03 (representando 0.1%, 1.5% e 3% de ruído relativo nos dados, respectivamente), sumarizados na Tabela 5.3. Vemos que as aproximações são de boa qualidade e mais precisas conforme o ruído decresce. Em todos os casos considerados, observamos erros para k_{11} em torno de 2% e 8% e ligeiramente maiores para k_{22}, de 8% a 17%. Esta diferença na segunda condutividade pode ser considerada um aspecto relacionado ao problema, especialmente pois o processo de minimização é conduzido corretamente de acordo com cada nível de ruído, representados por TRE e RTRE. Além disso, as soluções são obtidas com um pequeno número de iterações numericamente leves, um aspecto relacionado às matrizes altamente esparsas e às capacidades de aproximação de CPM, que permite boas reconstruções com poucos pontos na discretização.

Como complemento à Tabela 5.3, curvas de nível para as reconstruções das condutividades k_{11} e k_{22} são exibidas na Figura 5.6. Especialmente para k_{11}, veja que conseguimos aproximações que buscam reproduzir o formato das quantidades exatas, mesmo nos casos com restrições de medição. Para a segunda condutividade, as aproximações seguem comentários similares, a despeito das dificuldades que os três cenários de dados (*full data*,

Tabela 5.3 – Resultados numéricos para situações com dados incompletos e $L = I_2 \otimes \mathcal{L}_2$.

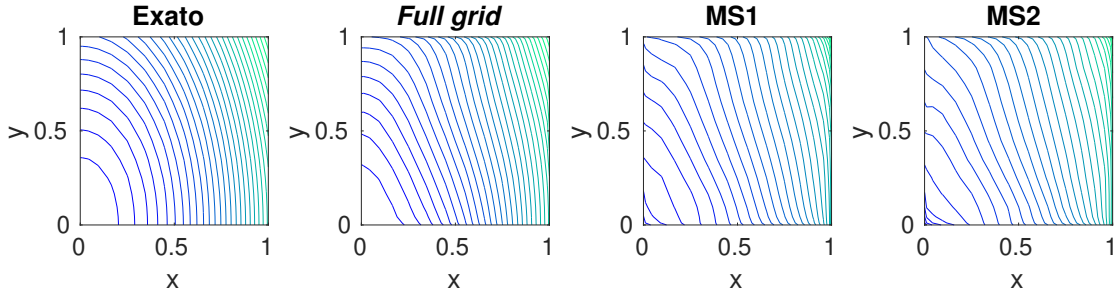
Ruído	**Caso**	RE($\mathbf{k}_{11}^{(j)}$)	RE($\mathbf{k}_{22}^{(j)}$)	TRE	RTRE	MI
0.1%	*Full grid*	0.0217	0.0855	0.0010	0.0010	13
	MS1	0.0574	0.0913	0.0226	0.0011	21
	MS2	0.0624	0.1794	0.0275	0.0010	30
1.5%	*Full grid*	0.0331	0.0966	0.0061	0.0061	11
	MS1	0.0664	0.1162	0.0666	0.0053	7
	MS2	0.0753	0.1766	0.0578	0.0062	29
3%	*Full grid*	0.0441	0.1042	0.0111	0.0111	14
	MS1	0.0798	0.1381	0.1667	0.0105	6
	MS2	0.0788	0.1703	0.0607	0.0109	12

Fonte – Boos, Bazán e Luchesi [25].

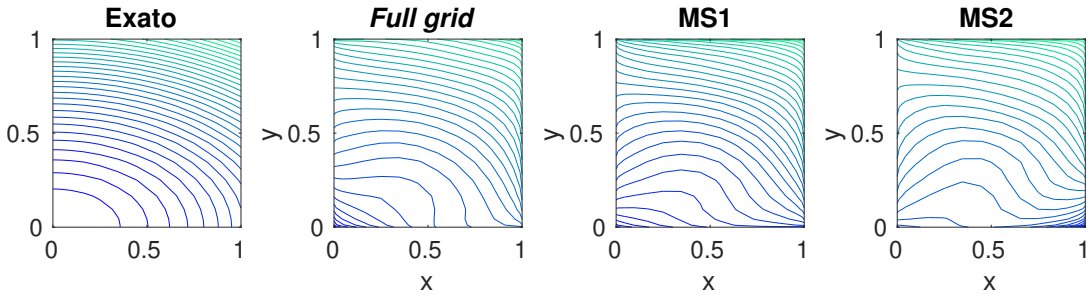
MS1 e MS2) apresentam na fronteira $y = 0$. De qualquer forma, vemos que a técnica é capaz de extrair aproximações condizentes com os níveis de ruído adicionados, mesmo para dados incompletos. Evidente que mais informação tende a produzir melhores resultados, como vemos para *full data*, porém o balanço entre dados restritos e suas respectivas aproximações torna o método aplicável em situações práticas reais.

Figura 5.6 – Curvas de nível dos resultados médios de 30 repetições, para ruído de 1.5%.

(a) Exato e reconstruções para $k_{11}(x, y)$.



(b) Exato e reconstruções para $k_{22}(x, y)$.



Fonte – o autor, 2022.

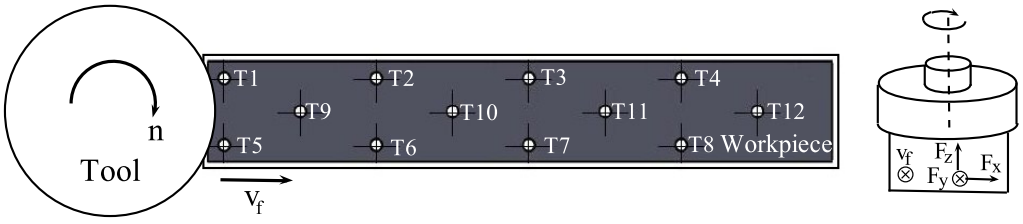
Observação 5.3. Com $n = 15$, buscamos $2(n + 1)^2 = 512$ variáveis discretas. Em todos os casos, utilizamos $N = 10$ estágios no tempo, de modo que os vetores de dados $\widetilde{\mathbf{u}}$ contém

$qN = 720$ medidas para MS1 e $qN = 200$ medidas para MS2. Se compararmos com dados em toda a malha, isto é, *full grid*, temos $N(n+1)^2 = 2560$ valores de temperatura que, pelo número de variáveis, conduz minimizar (5.49) a um problema altamente sobredeterminado. Para os casos restritos, a situação é diferente: MS1 é ligeiramente sobredeterminado, ao passo que MS2 possui menos informações que variáveis, então subdeterminado. Uma forma de compensar esta disparidade é repetir medições nas localizações prescritas tantas vezes quanto necessárias de modo a MS1 e MS2 se tornarem sobredeterminados como *full grid*. Esta abordagem, perceba, é muitas vezes possível na prática (repetindo experimentos, por exemplo), de modo a não ser uma opção restritiva. Aqui optamos, por brevidade, não fazê-lo, mas exemplos relacionados podem ser vistos em [25].

5.4 ESTIMATIVA DE CONDUTIVIDADE EM PROBLEMA EXPERIMENTAL

Nesta seção, demostraremos a efetividade do método proposto numericamente utilizando dados provenientes de um problema mecânico de fresamento de face (*face milling*, em inglês). Mais especificamente, nos referimos a um procedimento em uma peça retangular de aço AISI 4340 descrito em Luchesi e Coelho [85], na ocasião utilizado para estimar a função fonte a partir de valores de temperatura coletados durante o fresamento. O trabalho aqui apresentado busca utilizar estes mesmos dados em conjunto com o problema inverso para tentarmos reconstruir a condutividade do material, a título de ilustração. A operação envolve uma peça com 100 mm de comprimento e 15 mm de largura, com dados de temperatura capturados em 12 sensores (chamados de *termopares*) posicionados de acordo com a Figura 5.7 e a Tabela 5.4. Aqui fazemos apenas um recorte, com maiores detalhes no procedimento como um todo disponíveis no artigo original [85], bem como em outras referências relacionadas a experimentos práticos [3].

Figura 5.7 – Posicionamento dos termopares e das forças físicas envolvidas no problema experimental.



Fonte – Coelho e Luchesi [85].

Propriedades físicas do metal em consideração, exibidas na Tabela 5.5, indicam que a peça tem estrutura isotrópica (i.e., $k_{11} = k_{22} =: k$) e, mais ainda, condutividade constante $k(x, y) = 44.5$ W/m°C. Ou seja, gostaríamos de produzir aproximações também constantes (ou próximas de) em todo o domínio, embora claramente esta informação não esteja inclusa no problema inverso. Enfatizamos que, por estarmos lidando com um caso

Tabela 5.4 – Coordenadas (x, y), em milímetros, dos termopares no retângulo de lados 100 mm e 15 mm.

Termopar	T1	T2	T3	T4	T5	T6
Posição	(4,13)	(27,13)	(50,13)	(73,13)	(4,3)	(27,3)
Termopar	T7	T8	T9	T10	T11	T12
Posição	(50,3)	(73,3)	(16,8)	(39,8)	(62,8)	(85,8)

Fonte – Coelho e Luchesi [85].

Tabela 5.5 – Propriedades térmicas do aço AISI 4340.

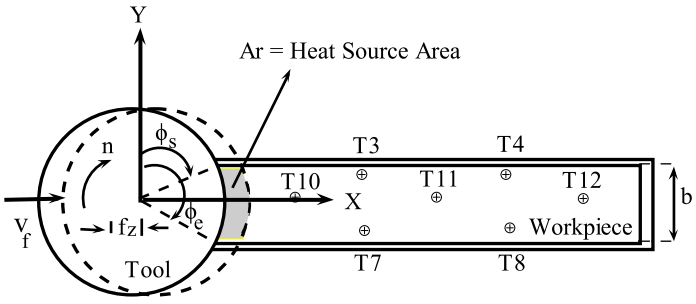
Calor específico c_p [J/kg°C]	**Condutividade térmica** k [W/m°C]	**Densidade** ρ [kg/m^3]	**Difusividade térmica** α [m^2/s]
475	44.5	7850	1.18×10^{-5}

Fonte – Adaptado de Coelho [36].

isotrópico, a única diferença se encontra na quantidade de variáveis envolvidas: $2(n+1)^2$ no caso ortotrópico e $(n+1)^2$ aqui.

Durante o processo de corte da ferramenta na face de metal, parte da energia mecânica acumulada na ferramenta flui para a peça em forma de calor como resultado da fricção entre as partes envolvidas [5, 6, 60]. Portanto, a área de contato atua como uma fonte de calor móvel na superfície da peça [71], como exibido na Figura 5.8. Esta área costuma ser aproximada na prática por um retângulo, o que é razoável pela figura, e gera um fluxo de calor aqui modelado como um termo fonte móvel com variação no tempo t e no plano xy e denotada no modelo (5.1)–(5.7) por $g(x, y, t)$.

Figura 5.8 – Geometria da fonte de calor móvel.



Fonte – Luchesi e Coelho [85].

É importante ressaltar que modelar o termo fonte $g(x, y, t)$ por si só representa um problema que demanda atenção. No caso de operações de fresamento de face como descrito aqui, podemos citar diversos autores interessados nesta questão, por exemplo, [40, 71, 100, 123]. Neste trabalho, entre três diferentes distribuições de intensidade para este termo gerador de calor (distribuição uniforme, parabólica e Gaussiana) e seguindo as

conclusões da análise perpetrada em [85], optamos por modelar g utilizando uma função Gaussiana bidimensional. Assim, de acordo com os dados experimentais coletados em [85], o problema (5.1)–(5.7) envolve o domínio $\Omega \times [0, t_f]$, $\Omega = [0, l_1] \times [0, l_2]$, com $l_1 = 0.1$ m, $l_2 = 0.015$ m, $t_f = 60$ s,

$$\begin{aligned} C(x, y) = k/\alpha &= 5.251 \times 10^6 \text{ Ws/m}^{3\circ}\text{C}, \\ q(x) &= 0, \\ k_{11}(x, y) = k_{22}(x, y) &= 44.5 \text{ W/m}^\circ\text{C (condutividade exata)} \end{aligned}$$

com parâmetros da Tabela 5.5,

$$\begin{aligned} h_1(y) = h_2(y) = h_3(x) = h_4(x) &= 120 \text{ W/m}^{2\circ}\text{C}, \\ f_1(y, t) = f_2(y, t) = f_3(x, t) = f_4(x, t) &= 21 \text{ }^\circ\text{C}, \quad \text{e} \\ u_0(x, y) &= 22 \text{ }^\circ\text{C} \quad \text{(temperatura inicial)}. \end{aligned}$$

Ainda, o termo fonte $g(x, y, t)$ definido em $\Omega \times [0, t_f]$ modela uma onda em 2D com deslocamento em x e y que representa o movimento da ferramenta de acordo com o posicionamento no domínio ao longo do tempo:

$$g(x, y, t) = g_0 \frac{36}{\pi} \exp\left(-\frac{3(y - y_0)^2}{y_0^2}\right) \exp\left(-\frac{3(x - x_0)^2}{x_0^2}\right), \tag{5.61}$$

para todo (x, y, t) tal que $t \in [0, t_f]$, $v_f t - a/2 < x < v_f t + a/2$ e $-b/2 < y < b/2$, e nula no restante do domínio (a ferramenta introduz calor apenas em uma pequena região por vez). Aqui, $(x_0, y_0) = (v_f t, 0.015/2)$ corresponde ao centro da onda no tempo $t \in [0, t_f]$ para velocidade $v_f = 1.7 \times 10^{-3}$ m/s. Esta descrição significa, em outras palavras, que g introduz calor no sistema como uma função Gaussiana somente na área de contato (aproximada por um retângulo, veja Figure 5.8) entre a ferramenta e a peça em cada passo do tempo. Além disso, a contante g_0 é conhecida como *intensidade de liberação de calor* (*heat liberation intensity*) e demonstrada ser, para este caso, $g_0 = 6.7444 \times 10^5$.

Com o modelo em mãos, buscamos demonstrar a efetividade do método proposto em dois casos: com dados sintéticos e experimentais, o segundo coletado por Luchesi e Coelho [85]. Em ambos, para reduzir a quantidade de variáveis, optamos por considerar a condutividade isotrópica $k(x, y)$ como função de x somente, i.e. $k = k(x)$. Com esta mudança, o vetor de valores desconhecidos na malha de Chebyshev (x_i, y_j), $i, j = 0, \ldots, n$, se torna

$$\mathbf{k} = [\mathbf{k}_1, \mathbf{k}_2, \ldots, \mathbf{k}_{n+1}]^T, \tag{5.62}$$

em que $\mathbf{k}_{i+1} = k(x_i, y_j)$, para $i, j = 0, \ldots, n$, reduzindo as variáveis de $(n+1)^2$ para $(n+1)$. Esta escolha pode também ser justificada pelo formato do domínio (a peça é fina em y relativa a x, portanto valores nesta direção podem variar menos) e pelo material em uso (aço AISI 4340), para o qual a literatura orienta à expectativa de que a condutividade

seja, de fato, constante, como já observado. Além disso, como consequência, optamos por utilizar LMMSS com $L = L_1$ definida em (2.6), isto é, o operador de derivação discreta em 1D. Esta opção é condizente com as condutividades esperadas como constantes, em que a segunda derivada teria menor influência.

5.4.1 Reconstrução da condutividade baseada em dados sintéticos

Como a função de temperatura exata $u(x, y, t)$ é desconhecida, os dados usados na reconstrução são gerados como segue. Primeiro, resolvemos o problema direto como descrito na Seção 5.1 usando malhas com maior refinamento no espaço e no tempo, $n = 40$ e $N = 600$ ($\Delta t = 0.1$), junto com os parâmetros do problema como descrito acima. Então, a solução obtida é interpolada através da rotina interp2 do Matlab para termos valores da temperatura nas posições dos sensores, consideradas então como "dados exatos". Em seguida, perturbamos tais dados como feito anteriormente para obter valores de temperatura respeitando (5.56), considerados então como dados de entrada à reconstrução. Neste sentido, é importante mencionar que mesmo especificando o nível de ruído NL a priori para obter dados com perturbações, o erro "real" $\|e\|_2 = \|\widetilde{\mathsf{u}} - \mathsf{u}\|_2$ é difícil de estimar uma vez que a solução exata do problema direto e o erro de interpolação são desconhecidos. Desta forma, DP não pode ser aplicado como critério de parada. Contornamos esta dificuldade escolhendo parar as iterações com o critério de resíduo relativo

$$\frac{|\phi(\mathbf{k}^{(j)}) - \phi(\mathbf{k}^{(j-1)})|}{\phi(\mathbf{k}^{(j-1)})} < \varepsilon, \quad j \geq 1, \tag{5.63}$$

em que utilizamos $\varepsilon = 10^{-3}$ em nossos testes numéricos.

Para o tratamento do problema inverso, a matriz Jacobiana deve ser modificada de acordo a determinar as sensitividades nas localizações dos sensores. Inicialmente, cada bloco da Jacobiana $\mathbf{J}_k$ é computado como descrito na Subseção 5.2.1 através do sistema (5.54), com a única observação de que a matriz $\mathbf{A}$ depende agora de $(n + 1)$ variáveis (devido à simplificação em (5.62)) e que $\mathbf{J}_k$ é de ordem $(n + 1)^2 \times (n + 1)$. Então, cada coluna de $\mathbf{J}_k$ é interpolada aos pontos de interesse, i.e., as 12 posições dos sensores, e realocados para construir matrizes $\breve{\mathbf{J}}_k \in \mathbb{R}^{12\times(n+1)}$, similar ao feito nos casos de medidas restritas.

Os experimentos numéricos foram efetuados com 16 pontos na malha ($n = 15$) e valores de temperatura $\widetilde{\mathbf{u}}_j$ para três diferentes escolhas de N e Δt, i.e., dados $\widetilde{\mathbf{u}}_j$ em $t_j = j\Delta t$, $j = 0, 1, \ldots, N$, considerando (a) $\Delta t = 1$ e $N = 60$, (b) $\Delta t = 1$ e $N = 30$, e (c) $\Delta t = 0.5$ e $N = 60$. A ideia por trás destas escolhas é comparar o comportamento do método em diferentes arranjos temporais. Como resultado, a matriz Jacobiana em cada caso é composta por 60, 30 e 60 blocos, respectivamente, de matrizes de ordem $12 \times (n + 1)$, verticalmente colocadas. Dado que a condutividade exata é constante e igual a 44.5, os chutes iniciais foram escolhidos relativamente distantes acima e abaixo da exata. Especificamente, consideramos dois casos, gerados da mesma forma que em

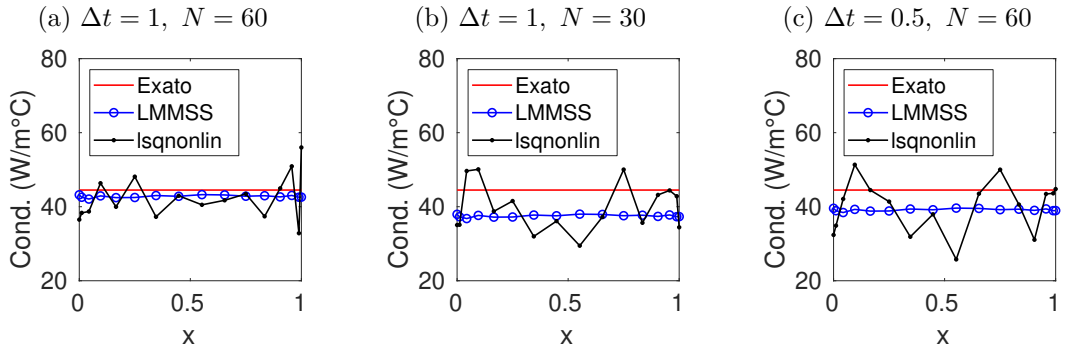
(5.60). Um iterando inicial toma $k^0(x, y) = 25$ adicionado a um vetor de distribuição normal de tamanho $n + 1$ calculado via rotina randn, e o outro da mesma forma mas com $k^0(x, y) = 55$. Utilizando a notação em (5.60), estas escolhas se traduzem a $a = 25$ e $a = 55$, resp., com $\vartheta = 1$ para ambos. Também, como critério comparativo, o problema é adicionalmente resolvido através da rotina do Matlab lsqnonlin baseada no algoritmo de região de confiança reflexiva [37]. No experimento, lsqnonlin é usado com e sem informações a priori, especificada na forma de limitantes inferiores e superiores nas variáveis, e escolhidos como 20 e 60, respectivamente. Resultados médios após 10 diferentes instâncias, todas com ruído de 1.5% estão sumarizados na Tabela 5.6.

Tabela 5.6 – Resultados obtidos por LMMSS e lsqnonlin (sem informação a priori) com 1.5% de ruído nos dados. Em negrito, resultados para lsqnonlin com informação a priori.

Seleção temporal	k^0	**LMMSS**			lsqnonlin		
		RE($\mathbf{k}^j$)	RTRE	MI	RE($\mathbf{k}^j$)	RTRE	MI
(a) $\Delta t = 1$, $N = 60$	≈ 25	0.0548	0.0022	4	0.6991 **(0.2674)**	0.0025 **(0.0023)**	11 **(8)**
	≈ 55	0.0485	0.0023	3	0.8226 **(0.2363)**	0.0023 **(0.0024)**	7 **(5)**
(b) $\Delta t = 1$, $N = 30$	≈ 25	0.1817	0.0021	4	0.8101 **(0.3723)**	0.0024 **(0.0021)**	8 **(6)**
	≈ 55	0.1604	0.0020	3	0.9867 **(0.3779)**	0.0023 **(0.0022)**	7 **(6)**
(c) $\Delta t = 0.5$, $N = 60$	≈ 25	0.1418	0.0019	4	0.6175 **(0.3666)**	0.0020 **(0.0019)**	8 **(6)**
	≈ 55	0.1233	0.0019	2	1.0533 **(0.3610)**	0.0020 **(0.0019)**	9 **(6)**

Fonte – Boos, Bazán e Luchesi [25].

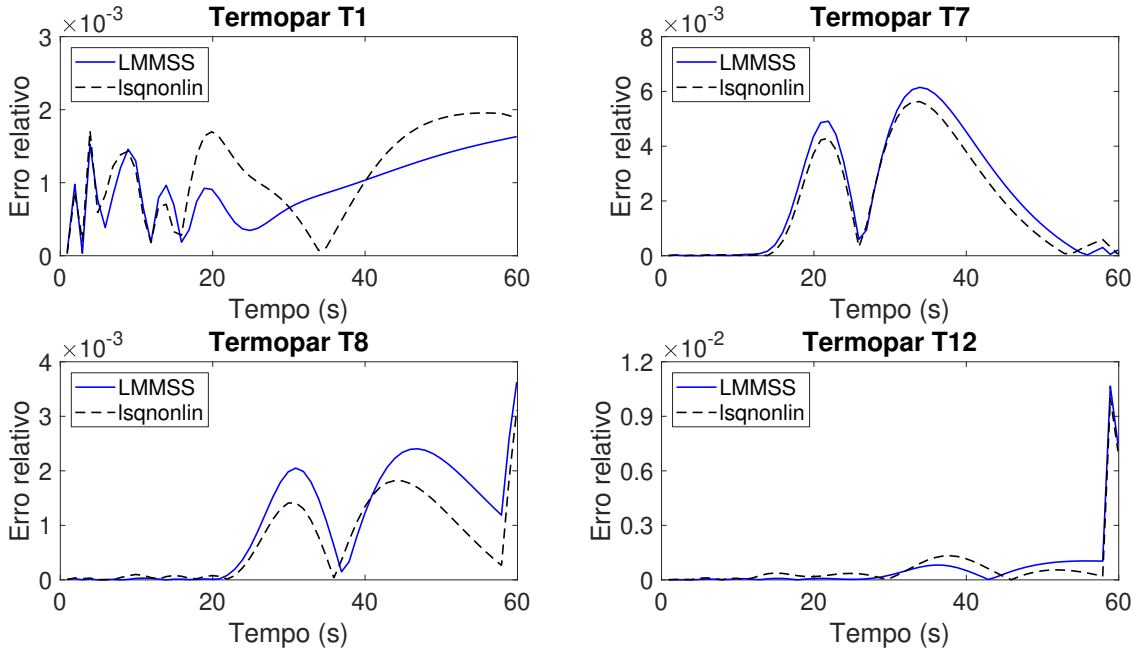
Figura 5.9 – Resultados médios obtidos por LMMSS e lsqnonlin (com informação a priori) para $k^0(x, y) \approx 55$ e 1.5% de ruído nos dados. A condição inicial $k^0(x, y) \approx 25$ performou similarmente, omitida aqui.



Fonte – Boos, Bazán e Luchesi [25].

Note que ambos LMMSS e lsqnonlin concluem bem o processo de minimização atingindo significante redução da função objetivo medida por RTREs em torno de 10 vezes menor que o nível de ruído. Esta informação é corroborada pela Figura 5.10, que apresenta o erro relativo pontual obtido por ambos os métodos para alguns termopares,

Figura 5.10 – Erros relativos pontuais para as reconstruções da temperatura por LMMSS e lsqnonlin (com informações a priori) em alguns termopares (Tabela 5.4) em relação aos dados sintéticos, com $k^0(x, y) \approx 55$, caso (a) $\Delta t = 1$, $N = 60$. Os outros casos performaram similarmente.



Fonte – o autor, 2022.

considerando o caso (a), em que os erros são da ordem de 10^{-3}. Entretanto, veja que ao resolvermos problemas mal postos, reduções desse tipo não necessariamente resultam em boas reconstruções, um fato frequentemente visto em uma variedade de aplicações. Neste experimento numérico, esta situação acontece em conexão com lsqnonlin quando aplicado sem utilizar informações a priori, como pode ser visto pelos erros de reconstrução da condutividade extremamente altos. Em contraste, isto não ocorre com as reconstruções fornecidas por LMMSS, que entregam erros relativos para a condutividade entre 4% e 18%, o que é bastante razoável para o tratamento de um problema não linear mal posto com dados imprecisos. Note que as reconstruções obtidas com lsqnonlin melhoram significativamente quando usamos informações a priori (veja números em negrito na Tabela 5.6), embora a qualidade ainda se mantenha distante da exata ou mesmo da obtida por LMMSS, como pode ser claramente observado na Figura 5.9. Finalmente, mudanças na forma como os dados são recebidos e a quantidade de informações utilizadas para resolver o problema inverso causou pequenas variações nas quantidades estimadas, um indicativo da robustez do método proposto em diferentes cenários.

5.4.2 Reconstrução baseada em dados experimentais

Neste segundo caso, a reconstrução da condutividade se baseia em dados experimentais obtidos por Luchesi e Coelho [85] durante uma operação de fresamento de face,

performada em seis diferentes condições. O conjunto de dados aqui utilizados corresponde ao que é chamado no artigo de *machining condition* C1 e considera parâmetros físicos dados na Tabela 5.5. Similarmente, o problema de transporte usado na reconstrução envolve os mesmos parâmetros e termo fonte como descrito no início desta seção e nas relações acima de (5.61). A aquisição dos dados de temperatura foi efetuada por uma rotina computacional através do *software* LabView e obtida por 12 termopares posicionados na peça como descrito na Tabela 5.4. Após, os valores de temperatura gravados durante 60 s foram suavizados através de uma rotina de média móvel com 34 pontos no Matlab e amostrada de modo a fornecer $N = 180$ medidas de temperatura para cada termopar. Portanto, os dados disponíveis para inversão compreendem valores de temperatura $\widetilde{\mathbf{u}}_j$ nos instantes $t_j = j\Delta t$, $j = 0, 1, \ldots, 180$, com $\Delta t = 60/180 \approx 0.3333$ s.

As operações numéricas foram realizadas com $n = 15$ (ou seja, 16 pontos na malha). Adicionalmente, como os erros para a temperatura são difíceis de estimar com dados experimentais (já que não temos acesso à solução exata), não podemos utilizar DP como critério de parada, sendo que o processo iterativo é então parado com o critério de resíduo relativo definido em (5.63). Para propósito de comparação, a versão de lsqnonlin que considera limitantes 20 e 60 nas variáveis é utilizada. Resultados numéricos para este método e LMMSS considerando chutes iniciais próximos de 25 e 55, são condensados na Tabela 5.7 e ilustrados na Figura 5.11.

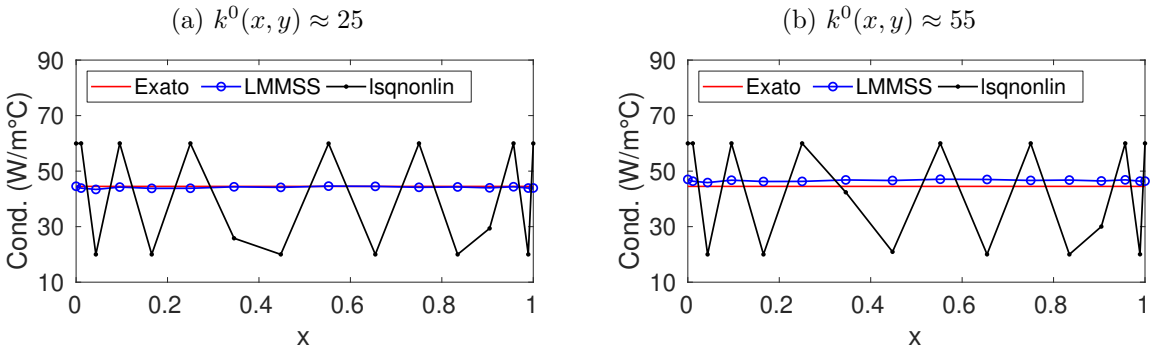
Pela tabela, vemos que o critério de parada para LMMSS é atingido com um pequeno número de passos e que as reconstruções obtidas apresentam erros relativos $\mathrm{RE}(\mathbf{k}^j)$ e RTRE com a mesma ordem de precisão (variando entre 3% e 8%). Por outro lado, embora lsqnonlin também pare com poucas iterações, é importante notar que a situação é completamente diferente no que diz respeito às reconstruções uma vez que $\mathrm{RE}(\mathbf{k}^j)$ é significantemente grande e próximo de 45%. Portanto, pelos resultados é claro que LMMSS gera soluções muito mais precisas que lsqnonlin, o que pode ser interpretado também como um efeito da matriz $L = \mathcal{L}_1$. De fato, o operador discreto de derivação de primeira ordem tende a introduzir suavidade nas aproximações de maneira "quase linear", visualmente presente na Figura 5.11 e também no caso sintético (veja Figura 5.9) e exemplos anteriores. Esta propriedade favorece largamente as reconstruções dadas por LMMSS, especialmente neste problema em que a condutividade exata é constante.

Tabela 5.7 – Resultados para dados experimentais através de LMMSS e lsqnonlin (com limitantes inferiores e superiores de 20 e 60, respectivamente).

k^0	**LMMSS**			lsqnonlin		
	$\mathrm{RE}(\mathbf{k}^j)$	RTRE	MI	$\mathrm{RE}(\mathbf{k}^j)$	RTRE	MI
≈ 25	0.0335	0.0848	3	0.4386	0.0823	5
≈ 55	0.0531	0.0848	2	0.4238	0.0824	5

Fonte – Boos, Bazán e Luchesi [25].

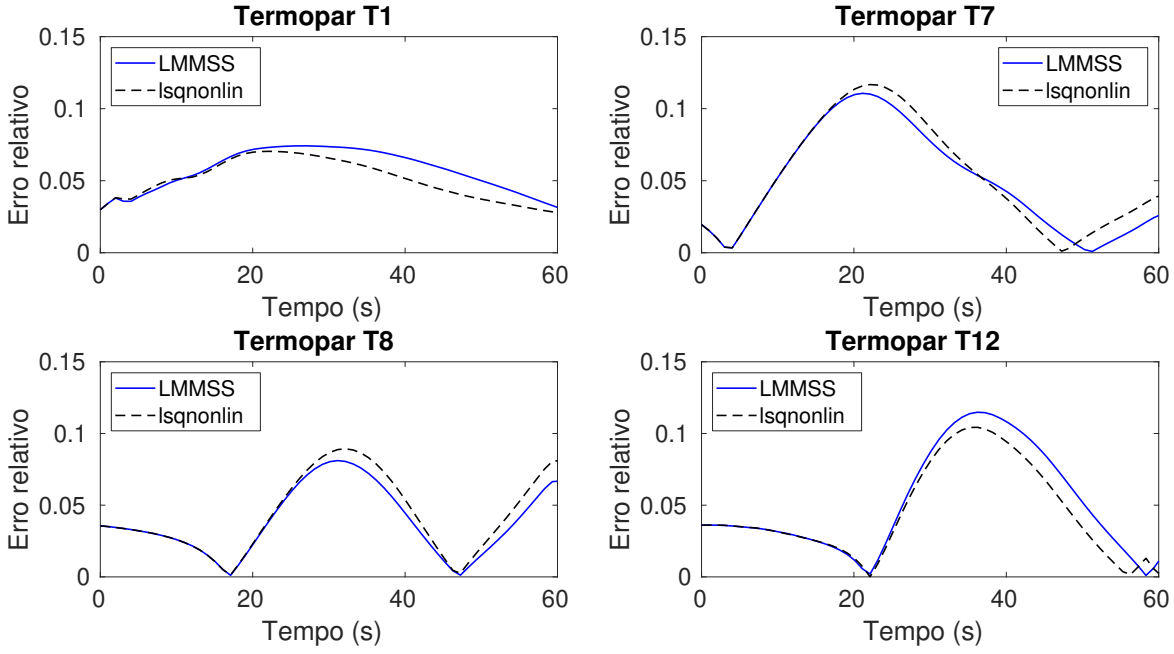
Figura 5.11 – Ilustração das reconstruções obtidas para dados experimentais por LMMSS e lsqnonlin (veja Tabela 5.7).



Fonte – Boos, Bazán e Luchesi [25].

Finalmente, na Figura 5.12 comparamos os dados de temperatura experimentais e as temperaturas reconstruídas para alguns termopares na forma de erro relativo pontual. Pela figura, podemos dizer que a qualidade da temperatura reconstruída por LMMSS e lsqnonlin é bastante similar e pouco acurada, com os maiores erros relativos em torno de 10%. Isto contrasta com a qualidade das reconstruções de condutividade (Figura 5.11), o que é razoável quando lidamos com problemas mal postos: ambos LMMSS e lsqnonlin eficientemente minimizam a função objetivo mas produzem aproximações completamente diferentes da solução procurada, numericamente evidente na Tabela 5.7. Mesmo assim,

Figura 5.12 – Erros relativos pontuais para as reconstruções da temperatura por LMMSS e lsqnonlin em alguns termopares em relação aos dados experimentais em uso, com $k^0(x,y) \approx 25$ e lsqnonlin com informações a priori.



Fonte – Boos, Bazán e Luchesi [25].

este aspecto reforça o comentário que, a despeito da natureza mal posta do problema e dos dados experimentais limitados, LMMSS é capaz de produzir bons resultados.

6 CONCLUSÃO

Tratamos de abordar métodos iterativos para problemas inversos lineares e não lineares na tentativa de propor progressos e melhoramentos a estas técnicas tendo em vista exemplos de importância na ciência e engenharia. Exibimos o ferramental teórico que envolve o método de Newton para problemas lineares [11], com interesse particular nas dificuldades em capturar soluções estáveis para os chamados problemas discretos mal postos. Neste sentido, o uso de critérios de parada especializados em conjunto aos métodos iterativos (em particular o método de Newton) é de suma importância à construção de soluções de qualidade. Isto pois, na regularização iterativa, o índice de parada k atua como parâmetro que controla a influência do ruído nas aproximações obtidas. Neste sentido, tendo por base o resultado que garante erro relativo entre a solução capturada pelo princípio da discrepância (DP) aplicado ao método de Newton e a solução exata da ordem de $\delta^{1/2}$, para δ representando a norma do ruído nos dados, nos propomos a obter novas estimativas mais acuradas.

Especificamente, se consideramos que a solução exata satisfaz a condição de regularidade do tipo Hölder, isto é, $x^* \in \mathcal{R}[(A^TA)^\mu]$, para $\mu > 0$, então vale que o erro relativo é da ordem de $\delta^{\frac{2\mu}{2\mu+1}}$. Esta estimativa foi obtida também em uma segunda forma, com dependência dos coeficientes de Fourier $u_i^T b$, que podem contribuir para estreitamento ainda maior dos resultados. Finalmente, uma terceira estimativa é obtida, sem uso de condições de fonte, em que mostramos que o erro é $\mathcal{O}(\delta)$, uma extensão direta ao resultado em [11]. É uma estimativa ótima, no contexto de garantir erros da ordem do ruído introduzido. Como um todo, são resultados que demonstram a robustez da técnica em calcular soluções dependentes do ruído nos dados e funcionam como garantia teórica na qualidade das aproximações obtidas.

No contexto não linear, desenvolvemos o método de Levenberg-Marquardt com *scaling* singular (LMMSS). Buscamos fornecer exemplos e motivação (partindo do caso linear) em que o uso de matrizes singulares pode trazer benefícios em aplicações, com foco central em matrizes representando operadores de derivação discretos de primeira e segunda ordens. Elaboramos a estrutura teórica necessária para demonstrarmos a convergência do método LMMSS proposto, com forte uso da GSVD e suas propriedades. Verificamos que, sob hipóteses modestas e escolhas apropriadas do parâmetro de *damping* λ_k e do passo α_k, pontos de acumulação das iterações LMMSS são pontos estacionários para o problema original de mínimos quadrados não linear, o melhor que se pode esperar com a pouca estrutura provida pelo problema.

Seguindo o trabalho de Yamashita e Fukushima [122], provamos que a convergência para pontos estacionários ocorre com taxa quadrática localmente, o que acontece também para as versões de LMM com matriz de *scaling* não singular. Para tanto, utilizamos que o passo é escolhido através do critério de Armijo e que a função objetivo fornece um

limitante de erro às aproximações geradas (conhecida como *error bound* na literatura). Esta última hipótese em especial vem sendo aplicada com intensidade por pesquisadores nas últimas décadas e serve como uma condição de regularidade menos exigente que, por exemplo, pedir que a matriz Jacobiana tenha posto completo na solução. Neste sentido, a estrutura montada neste trabalho tende a ser abrangente e coerente com a teoria corrente na área.

No último capítulo, abordamos o estudo de um problema de condução de calor em duas dimensões com inúmeras aplicações na ciência e engenharia, em especial industriais. O problema é descrito no ambiente contínuo como uma EDP com condições de fronteira de Robin e condição inicial, presente em diversas aplicações em áreas que lidam com interações térmicas, embora as mesmas equações sirvam também para descrever problemas com potencial elétrico. A discretização proposta se utiliza do método pseudo-espectral de Chebyshev (CPM) para lidar com as derivadas espaciais e o método de Crank-Nicolson (CN) na variável temporal. CPM tem importância central na técnica, escolhido principalmente pelas boas qualidades teóricas nas aproximações mesmo com um número relativamente pequeno de pontos na malha espacial. Já CN é um método de segunda ordem de convergência e fácil implementação, além da estabilidade absoluta, um aspecto importante do ponto de vista numérico. Estes dois métodos formam a base do *solver* ao problema direto de reconstruir valores de temperatura tendo conhecimento dos demais parâmetros, incluindo a condutividade. O problema inverso é então formulado através da minimização de um funcional não linear com base no problema direto e em valores de temperatura provenientes, por exemplo, de experimentos físicos. Como forma de atestar as capacidades da técnica desenvolvida, utilizamos LMMSS, com matrizes de *scaling* representando versões discretas de operadores de derivação de primeira e segunda ordens, para resolver o problema de mínimos quadrados formulado.

Exemplos numéricos mostram que o método proposto é capaz de construir boas aproximações à condutividade, em diferentes cenários, seja com ruído nos dados e/ou restrições de medição (como pode ocorrer na prática). Para amenizar o efeito do ruído, as iterações LMMSS são paradas através de DP, de modo a produzir resultados competitivos e em consonância com o nível de perturbação presente. Um exemplo final é apresentado, baseado em dados experimentais capturados em [85], com base em um problema de fresamento de face em aço AISI 4340 com medidas em apenas 12 pontos (sensores). Apesar das dificuldades inerentes à obtenção dos dados e tratamento dos mesmos, a aplicação de LMMSS se mostra viável e constrói aproximações coerentes com o esperado da teoria relacionada, com resultados superiores aos fornecidos pelo comparativo com a rotina lsqnonlin. Concluímos que o método proposto é eficiente e capaz de produzir boas reconstruções a despeito do aspecto mal posto dos problemas, mesmo na presença de ruído e em ambientes com medidas incompletas, com aplicação válida em problemas reais.

Em suma, estudamos assuntos relacionados a métodos iterativos para problemas

inversos lineares e não lineares, tanto na teoria quanto na aplicação, visando alcançar aperfeiçoamentos às áreas. Para problemas lineares, obtivemos novas estimativas de erro para o método de Newton, agora ótimas no sentido do ruído, permitindo análise similar para outros métodos e seus significados em dimensão finita. No contexto de mínimos quadrados não lineares, o segundo problema de interesse, propomos e mostramos a convergência quadrática da variação do método de Levenberg-Marquardt que usa matrizes de *scaling* singulares, com uso de hipóteses modestas. Em parte como aplicação numérica desta técnica, descrevemos e construímos uma discretização para o problema de recuperação da condutividade térmica em duas dimensões através de CPM e CN. O problema inverso, na presença de ruído, é então resolvido com LMMSS, cujos resultados evidenciam as capacidades do método em aplicações. Estas consistem das principais contribuições deste trabalho, com a análise relacionada à reconstrução de condutividade publicada em Boos, Luchesi e Bazán [26] e em fase de submissão em um segundo artigo, Boos, Bazán e Luchesi [25]. Os estudos relacionados às novas estimativas de erro para o método de Newton e análise de convergência de LMMSS, ambas aparentes novidades na literatura, estão em fase de consideração e escrita visando sua publicação.

Como possíveis sugestões para trabalhos futuros, citamos:

1. Ainda nas estimativas, estamos analisando a possibilidade de gerar também cotas com a utilização de condição de fonte logarítmica, isto é, que sugere tomar situações da forma $x^* = \log^{-p}(A^TA)^{-1}v$, para $p > 0$ e v apropriado. Semelhante à condição de Hölder, esta diz respeito à suavidade da solução e faz sentido de ser considerada especialmente para problemas com matrizes severamente mal condicionadas [70]. Assim como as propostas anteriores, a ideia gira em torno de exibir cotas de erro mais finas que as atuais, tornando possível precisar com mais eficiência a qualidade das soluções computadas.

2. Verificar a efetividade prática das cotas obtidas através de exemplos numéricos, especialmente focando em comparações entre os novos resultados. Os testes tem o intuito de analisar se as cotas são realistas no sentido de próximas do erro real obtido.

3. Desenvolver uma análise de convergência para LMMSS no caso de resíduo não nulo, lugar comum em aplicações práticas, possivelmente seguindo estratégias similares às abordadas em [14, 122].

4. Estudar a análise de convergência para LMMSS no ambiente de teoria funcional (dimensão infinita), com estimativas de erro dependentes do ruído nos dados [75].

5. Estudar/elaborar possíveis matrizes de *scaling* singulares diferentes das abordadas aqui, como exemplificado em [9]. Mesmo no caso dos operadores discretos de deri-

vação, alternativas simples como introduzir condições de fronteira reflexivas podem fazer diferença para alguns problemas [65].

6. Expandir a técnica proposta para lidar com o problema de condução de calor capaz de reconstruir simultaneamente outras quantidades como condutividade e capacidade térmicas. Além disso, fornecer uma discretização e análise de problema similar em domínios mais gerais, possivelmente tridimensionais, de importância em situações práticas.

REFERÊNCIAS

[1] ABSIL, P-A; MAHONY, R.; SEPULCHRE, R. *Optimization Algorithms on Matrix Manifolds*. 1. ed. Princeton: Princeton University Press, 2008. 224 p. DOI: https://doi.org/10.1515/9781400830244.

[2] ALESSANDRINI, G.; DE HOOP, M. V.; GABURRO, R. Uniqueness for the electrostatic inverse boundary value problem with piecewise constant anisotropic conductivities. *Inverse Problems*, v. 33, n. 12, p. 125013, 2017. DOI: https://doi.org/10.1088/1361-6420/aa982d.

[3] ALIFANOV, O. M. *Inverse heat transfer problems*. Berlin: Springer Science and Business Media, 1994. 348 p. DOI: https://doi.org/10.1007/978-3-642-76436-3.

[4] ALLAIRE, G.; KABER, S. M. *Numerical Linear Algebra*. Texts in Applied Mathematics, v. 55. New York: Springer, 2008. 271 p. DOI: https://doi.org/10.1007/978-0-387-68918-0.

[5] ALTINTAS, Y. *Manufacturing Automation*. 2. ed. Cambridge: Cambridge University Press, 2012. 382 p. DOI: https://doi.org/10.1017/CBO9780511843723.

[6] BARALIĆ, J. C. et. al. Modeling and optimization of temperature in end milling operations. *Thermal Science*, v. 23, n. 6A, p. 3651–3660, 2019. DOI: https://doi.org/10.2298/TSCI190328244B.

[7] BAZÁN, F. S. V. Chebyshev pseudospectral method for wave equation with absorbing boundary conditions that does not use a first order hyperbolic system. *Mathematics and Computers in Simulation*, v. 80, n. 11, p. 2124–2133, 2010. DOI: https://doi.org/10.1016/j.matcom.2010.04.014.

[8] BAZÁN, F. S. V.; BEDIN, L.; BORGES, L. S. Space-dependent perfusion coefficient estimation in a 2D bioheat transfer problem. *Computer Physics Communications*, v. 214, p. 18–30, 2017. DOI: http://dx.doi.org/10.1016/j.cpc.2017.01.002.

[9] BAZÁN, F. S. V. Simple and Efficient Determination of the Tikhonov Regularization Parameter Chosen by the Generalized Discrepancy Principle for Discrete Ill-Posed Problems. *Journal of Scientific Computing*, v. 63, p. 163–184, 2015. DOI: https://doi.org/10.1007/s10915-014-9888-z.

[10] BAZÁN, F. S. V.; BEDIN, L.; BOZZOLI, F. New methods for numerical estimation of convective heat transfer coefficient in circular ducts. *International Journal of Thermal Sciences*, v. 139, p. 387–402, 2019. DOI: https://doi.org/10.1016/j.ijthermalsci.2019.02.025.

[11] BAZÁN, F. S. V.; BOOS, E. Schultz matrix iteration based method for stable solution of discrete ill-posed problems. *Linear Algebra and its Applications*, v. 554, p. 120–145, 2018. DOI: https://doi.org/10.1016/j.laa.2018.05.022.

[12] BAZÁN, F. S. V.; BORGES, L. S. GKB-FP: an algorithm for large-scale discrete ill-posed problems. *BIT*, v. 50, p. 481–507, 2010. DOI: https://www.doi.org/10.1007/s10543-010-0275-3.

[13] BAZÁN, F. S. V.; CUNHA, M. C. C.; BORGES, L. S. Extension of GKB-FP algorithm to large-scale general-form Tikhonov regularization. *Numerical Linear Algebra*, v. 21, n. 3, p. 316–339, 2014. DOI: https://doi.org/10.1002/nla.1874.

[14] BEHLING, R.; GONÇALVES, D. S.; SANTOS, S. A. Local Convergence Analysis of the Levenberg–Marquardt Framework for Nonzero-Residue Nonlinear Least-Squares Problems Under an Error Bound Condition. *Journal of Optimization Theory and Applications*, v. 183, p. 1099–1122, 2019. DOI: https://doi.org/10.1007/s10957-019-01586-9.

[15] BEHLING, R.; BELLO-CRUZ, Y.; SANTOS, L.-R. Infeasibility and error bound imply finite convergence of alternating projections. *SIAM Journal on Optimization*, v. 31, n. 4, p. 2863–2892, 2021. DOI: https://doi.org/10.1137/20M1358669.

[16] BENATTI, K. A. *O Método de Levenberg-Marquardt para o problema de quadrados mínimos não linear*. 106 p. Dissertação (Mestrado) – Mestrado em Matemática, Universidade Federal do Paraná (UFPR), Curitiba, 2017.

[17] BEN-ISRAEL, A.; GREVILLE, T. N. E. *Generalized Inverses:* Theory and Applications. 2. ed. New York: Springer, 2003. 420 p. DOI: https://doi.org/10.1007/b97366.

[18] BERGOU, E. H.; DIOUANE, Y.; KUNGURTSEV, V. Convergence and Complexity Analysis of a Levenberg–Marquardt Algorithm for Inverse Problems. *Journal of Optimization Theory and Applications*, v. 185, p. 927–944, 2020. DOI: https://doi.org/10.1007/s10957-020-01666-1.

[19] BERTSEKAS, D. P. *Nonlinear Programming*. 2. ed. Athena Scientific, 1999. 777 p.

[20] BIRGIN, E. G.; MARTÍNEZ, J. M. *Practical Augmented Lagrangian Methods for Constrained Optimization*. 1. ed. Philadelphia: SIAM, 2014. 220 p. DOI: https://doi.org/10.1137/1.9781611973365.

[21] BJÖRCK, A. *Least Squares Methods*. Handbook of Numerical Analysis, v. 1: Finite Difference Methods (Part I) - Solution of Equations in $\mathbb{R}^n$ (Part I). Oxford: Elsevier, p. 465–652, 1990. DOI: https://doi.org/10.1016/S1570-8659(05)80036-5.

[22] BJÖRCK, A. *Numerical Methods for Least Squares Problems*. Philadelphia: SIAM, 1996. 408 p. DOI: https://doi.org/10.1137/1.9781611971484.

[23] BOOS, E. *Métodos Iterativos para a Pseudo-Inversa de Moore-Penrose e Aplicações na Resolução de Sistemas Lineares*. 79 p. Trabalho de conclusão de curso (Graduação) – Bacharelado em Matemática e Computação Científica, Universidade Federal de Santa Catarina (UFSC), Florianópolis, 2015.

[24] BOOS, E. *Método iterativo baseado na iteração de Newton para problemas discretos mal postos*. 116 p. Dissertação (Mestrado) – Mestrado em Matemática Pura e Aplicada, Universidade Federal de Santa Catarina (UFSC), Florianópolis, 2018.

[25] BOOS, E.; BAZÁN, F. S. V.; LUCHESI, V. M. Thermal conductivity reconstruction method with application in a face milling operation. *To be submitted*, 2022.

[26] BOOS, E.; LUCHESI, V. M.; BAZÁN, F. S. V. Chebyshev pseudospectral method in the reconstruction of orthotropic conductivity. *Inverse Problems in Science and Engineering*, v. 29, n. 1, p. 681–711, 2020. DOI: https://doi.org/10.1080/17415977.2020.1801675.

[27] BORGES, L. S.; BAZÁN, F. S. V.; CUNHA, M. C. C. Automatic stopping rule for iterative methods in discrete ill-posed problems. *Comp. Appl. Math.*, v. 34, p. 1175–1197, 2014. DOI: https://doi.org/10.1007/s40314-014-0174-3.

[28] BOZZOLI, F. et. al. Estimation of the local heat-transfer coefficient in the laminar flowregime in coiled tubes by the Tikhonov regularisation method. *International Journal of Heat and Mass Transfer*, v. 72, p. 352–361, 2014. DOI: https://doi.org/10.1016/j.ijheatmasstransfer.2014.01.019.

[29] BURDEN, R. L.; FAIRES, J. D.; BURDEN, A. M. *Numerical Analysis*. 10. ed. United States of America: Cengage Learning, 2016. 896 p. DOI: http://dx.doi.org/10.13140/2.1.4830.2406.

[30] CALVETTI, D.; LEWIS, B.; REICHEL, L. GMRES-type methods for inconsistent systems. *Linear Algebra and its Applications*, v. 316, n. 1, p. 157–169, 2000. DOI: https://doi.org/10.1016/S0024-3795(00)00064-1.

[31] CANUTO, C. et. al. *Spectral methods in fluid dynamics*. 1. ed. Springer Series in Computational Physics series. Berlin: Springer-Verlag, 1988. 568 p. DOI: https://doi.org/10.1007/978-3-642-84108-8.

[32] CAO, K.; LESNIC, D.; COLAÇO, M. J. Determination of thermal conductivity of inhomogeneous orthotropic materials from temperature measurements. *Inverse Problems in Science and Engineering*, v. 27, n. 10, p. 1372–1398, 2019. DOI: https://doi.org/10.1080/17415977.2018.1554654.

[33] CENSOR, Y. et. al. On diagonally relaxed orthogonal projection methods. *SIAM Journal on Scientific Computing*, v. 30, p. 473–504, 2007. DOI: https://doi.org/10.1137/050639399.

[34] CENSOR, Y.; GORDON, D.; GORDON, R. Component averaging: An efficient iterative parallel algorithm for large sparse unstructured problems. *Parallel Computing*, v. 27, p. 777–808, 2001. DOI: https://doi.org/10.1016/S0167-8191(00)00100-9.

[35] CIMMINO, G. Calcolo approssimato per le soluzioni dei sistemi di equazioni lineari. *La Ricerca Scientifica*, p. 326–333, 1938.

[36] COELHO, R. T.; NG, E.; ELBESTAWI, M. A. Tool wear when turning hardened AISI 4340 with coated PCBN tools using finishing cutting conditions. *International Journal of Machine Tools and Manufacture*, v. 47, n. 2, p. 263–272, 2007. DOI: https://doi.org/10.1016/j.ijmachtools.2006.03.020.

[37] COLEMAN, T.; LI, Y. An interior, trust region approach for nonlinear minimization subject to bounds. *SIAM Journal on Optimization*, v. 6, n. 2, p. 418–445, 1996. DOI: https://doi.org/10.1137/0806023.

[38] CONN, A. R.; GOULD, N. I. M.; TOINT, P. L. *Trust-Region Methods*. 1. ed. Philadelphia: SIAM, 2000. 959 p. DOI: https://doi.org/10.1137/1.9780898719857.

[39] CRANK, J.; NICOLSON, P. A practical method for numerical evaluation of solutions of partial differential equations of the heat-conduction type. *Mathematical Proceedings of the Cambridge Philosophical Society*, v. 43, n. 1, p. 50–67, 1947. DOI: https://doi.org/10.1017/S0305004100023197.

[40] CUI, X. et. al. Chip temperature and its effects on chip morphology, cutting forces, and surface roughness in high-speed face milling of hardened steel. *International Journal of Advanced Manufacturing Technology*, v. 77, p. 2209–2219, 2015. DOI: https://doi.org/10.1007/s00170-014-6635-4.

[41] DENNIS JR, J. R. *Nonlinear least squares and equations*. The State of the Art in Numerical Analysis: Conference Proceedings. London: Academic Press, p. 269–312, 1977.

[42] DENNIS JR, J. R.; SCHNABEL, R. B. *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*. Philadelphia: SIAM, 1996. 378 p. DOI: https://doi.org/10.1137/1.9781611971200.

[43] ELDÉN, L. Algorithms for the regularization of ill-conditioned least squares problems. *BIT*, v. 17, p. 134–145, 1977. DOI: https://doi.org/10.1007/BF01932285.

[44] ELDÉN, L. A weighted pseudoinverse, generalized singular values, and constrained least squares problems. *BIT*, v. 22, p. 487–502, 1982. DOI: https://doi.org/10.1007/BF01934412.

[45] ELFING, T.; NIKAZAD, T.; HANSEN, P. C. Semi-convergence and relaxation parameters for a class of SIRT algorithms. *Electronic Transactions on Numerical Analysis*, v. 37, p. 321–336, 2010.

[46] ENGL, H. W.; HANKE, M.; NEUBAUER, A. *Regularization of Inverse Problems*. Dordrecht: Kluwer Academic Publishers, 1996. 321 p.

[47] FAN, K. Maximum Properties and Inequalities for the Eigenvalues of Completely Continuous Operators. *Proceedings of the National Academy of Sciences*, v. 37, n. 11, p. 760—766, 1951. DOI: https://doi.org/10.1073/pnas.37.11.760.

[48] FAN, J. A modified Levenberg-Marquardt algorithm for singular system of nonlinear equations. *Journal of Computational Mathematics*, v. 21, n. 5, p. 625–636, 2003.

[49] FAN, J.; YUAN, Y. On the quadratic convergence of the Levenberg-Marquardt method without nonsingularity assumption. *Computing*, v. 74, n. 1, p. 23–39, 2005. DOI: https://doi.org/10.1007/s00607-004-0083-1.

[50] FLETCHER, R. *A modified Marquardt subroutine for non-linear least squares.* Relatório R6799, Atomic Energy Research Establishment, Harwell, England, 1971. 24 p.

[51] FRANCISCO, J. B.; BAZÁN, F. S. V. Nonmonotone algorithm for minimization on closed sets with application to minimization on Stieffel manifolds. *J. Comput. Appl. Math.*, v. 236, n. 10, p. 2717–2727, 2012. DOI: https://doi.org/10.1016/j.cam.2012.01.014.

[52] GARDENGHI, J. L.; SANTOS, S. A. *Sistemas não lineares via região de confiança: o algoritmo de Levenberg-Marquardt.* Relatório de pesquisa (IME-UNICAMP), p. 1–45, 2011. DOI: https://doi.org/10.13140/RG.2.1.2728.8163.

[53] GILBERT, P. Iterative methods for the three-dimensional reconstruction of an object from projections. *Journal of Theorical Biology*, v. 36, p. 105–117, 1972. DOI: https://doi.org/10.1016/0022-5193(72)90180-4.

[54] GOLUB, G. H.; HEATH, M.; WAHBA, G. Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, v. 21, n. 2, p. 215–223, 1979. DOI: https://doi.org/10.1080/00401706.1979.10489751.

[55] GOLUB, G. H.; ORTEGA, J. M. *Scientific Computing and Differential Equations:* An Introduction to Numerical Methods. London: Academic Press, 1992. 337 p. DOI: https://doi.org/10.1016/C2009-0-21576-5.

[56] GOLUB, G. H.; VAN LOAN, C. F. *Matrix Computations.* 4. ed. Maryland: Johns Hopkins University Press, 2013. 756 p.

[57] GORDON, R.; BENDER, R.; HERMAN, G.T. Algebraic reconstruction techniques (ART) for threedimensional electron microscopy and X-ray photography. *Journal of Theoretical Biology*, v. 29, n. 3, p. 471–481, 1970. DOI: https://doi.org/10.1016/0022-5193(70)90109-8.

[58] GOTTLIEB, D.; ORSZAG, S. A. *Numerical analysis of spectral methods:* theory and applications. Philadelphia: SIAM, 1977. 161 p. DOI: https://doi.org/10.1137/1.9781611970425.

[59] GREENBAUM, A. *Iterative Methods for Solving Linear Systems.* Seattle: SIAM, 1997. 213 p. DOI: https://doi.org/10.1137/1.9781611970937.

[60] HAO, G.; LIU, Z. The heat partition into cutting tool at tool-chip contact interface during cutting process: a review. *The International Journal of Advanced*

Manufacturing Technology, v. 108, p. 393–411, 2020. DOI: https://doi.org/10.1007/s00170-020-05404-9.

[61] HANKE, M. *Conjugate gradient type methods for ill-posed problems*. Harlow: Longman, 1995. 144 p. DOI: https://doi.org/10.1201/9781315140193.

[62] HANSEN, P. C. Regularization, GSVD and Truncated GSVD. *BIT*, v. 29, n. 3, p. 491–504, 1989. DOI: https://doi.org/10.1007/BF02219234.

[63] HANSEN, P. C. Analysis of discrete ill-posed problems by means of the L-curve. *SIAM Rev.*, v. 34, n. 4, p. 561–580, 1992. DOI: https://doi.org/10.1137/1034115.

[64] HANSEN, P. C. *Rank-Deficient and Discrete Ill-Posed Problems*. Philadelphia: SIAM, 1998. 247 p. DOI: https://doi.org/10.1137/1.9780898719697.

[65] HANSEN, P. C. *Discrete Inverse Problems:* Insight and Algorithms. Philadelphia: SIAM, 2010. 213 p. DOI: https://doi.org/10.1137/1.9780898718836.

[66] HANSEN, P. C. Oblique projections and standard-form transformations for discrete inverse problems. *Numer. Linear Algebra Appl.*, v. 20, n. 2, p. 250–258, 2013. DOI: https://doi.org/10.1002/nla.802.

[67] HANSEN, P. C. Regularization Tools: A MATLAB package for analysis and solution of discrete ill-posed problems. *Numerical Algorithms*, v. 6, n. 1, p. 1–35, 1994. DOI: https://doi.org/10.1007/BF02149761.

[68] HANSEN, P. C.; JENSEN, T. K. Smoothing-norm preconditioning for regularizing minimum-residual methods. *SIAM Journal of Matrix Analysis and Applications*, v. 19, n. 1, p. 1–14, 2007. DOI: https://doi.org/10.1137/050628453.

[69] HOFFMAN, A. J. On approximate solutions of systems of linear inequalities. *Journal of Research of the National Bureau of Standards*, v. 49, p. 263–265, 1952.

[70] HOHAGE, T. Regularization of exponentially ill-posed problems. *Numerical Funcional Analysis and Optimization*, v. 21, n. 3-4, p. 439–464, 2000. DOI: https://doi.org/10.1080/01630560008816965.

[71] HOU, Z.; KOMANDURI, R. General solutions for stationary/moving plane heat source problems in manufacturing and tribology. *International Journal of Heat and Mass Transfer*, v. 43, n. 10, p. 1679–1698, 2000. DOI: https://doi.org/10.1016/S0017-9310(99)00271-9.

[72] IPSEN, I. C. F. *Numerical Matrix Analysis:* Linear Systems and Least Squares. Philadelphia: SIAM, 2009. 125 p. DOI: https://doi.org/10.1137/1.9780898717686.

[73] ISMAILOV, M. I.; BAZÁN, F. S. V.; BEDIN, L. Time-dependent perfusion coefficient estimation in a bioheat transfer problem. *Computer Physics Communications*, v. 230, p. 50–58, 2018. DOI: https://doi.org/10.1016/j.cpc.2018.04.019.

[74] JIA, Z. Approximation accuracy of the Krylov subspaces for linear discrete ill-posed problems. *Journal of Computational and Applied Mathematics*, v. 374, p. 112786, 2020. DOI: https://doi.org/10.1016/j.cam.2020.112786.

[75] KALTENBACHER, B.; NEUBAUER, A.; SCHERZER, O. *Iterative regularization methods for nonlinear ill-posed problems*. Radon Series on Computational and Applied Mathematics series, v. 6. Berlin: DeGruter, 2008. 194 p. DOI: https://doi.org/10.1515/9783110208276.

[76] KANZOW, C.; YAMASHITA, N.; FUKUSHIMA, M. Levenberg-Marquardt methods with strong local convergence properties for solving nonlinear equations with convex constraints. *Journal of Computational and Applied Mathematics*, v. 172, n. 2, p. 375–397, 2004. DOI: https://doi.org/10.1016/j.cam.2004.02.013.

[77] KELLEY, C. T. *Iterative Methods for Optimization*. Philadelphia: SIAM, 1995. 180 p. DOI: https://doi.org/10.1137/1.9781611970920.

[78] KILMER, M. E.; O'LEARY, D. P. Choosing regularization parameters in iterative methods for ill-posed problems. *SIAM J. Matrix Anal. Appl.*, v. 22, n. 4, p. 1204–1221, 2001. DOI: https://doi.org/10.1137/S0895479899345960.

[79] KIRSCH, A. *An Introduction to the Mathematical Theory of Inverse Problems*. 2. ed. Applied Mathematical Sciences, v. 120. New York: Springer, 2011. 310 p. DOI: https://doi.org/10.1007/978-1-4419-8474-6.

[80] LANDWEBER, L. An iteration formula for Fredholm integral of the first kind. *American Journal of Mathematics*, v. 73, n. 3, p. 615–624, 1951. DOI: https://doi.org/10.2307/2372313.

[81] LAWSON, C. L.; HANSON, R. J. *Solving Least Squares Problems*. Classics in Applied Mathematics series, v. 15. Philadelphia: SIAM, 1995. 337 p. DOI: https://doi.org/10.1137/1.9781611971217.

[82] LEVENBERG, K. A method for the solution of certain non-linear problems in least squares. *Quarterly of Applied Mathematics*, v. 2, p. 164–168, 1944. DOI: https://doi.org/10.1090/qam/10666.

[83] LI, R.-C. Bounds on Perturbations of Generalized Singular Values and of Associated Subspaces. *SIAM Journal on Matrix Analysis and Applications*, v. 14, n. 1, p. 195–234, 1993. DOI: https://doi.org/10.1137/0614017.

[84] LIMA, E. L. *Curso de análise vol. 1*. 15. ed. IMPA, 2019. 320 p.

[85] LUCHESI, V. M.; COELHO, R. An inverse method to estimate the moving heat source in machining process. *Applied Thermal Engineering*, v. 45-46, p. 64–78, 2012. DOI: https://doi.org/10.1016/j.applthermaleng.2012.04.014.

[86] MAHMOOD, M. S.; LESNIC, D. Identification of conductivity in inhomogeneous orthotropic media. *Int J Numer Method Heat Fluid Flow*, v. 29, n. 1, p. 165–183, 2019. DOI: https://doi.org/10.1108/HFF-11-2017-0469.

[87] MARQUARDT, D. An Algorithm for Least-Squares Estimation of Nonlinear Parameters. *SIAM Journal on Applied Mathematics*, v. 11, n. 2, p. 431–441, 1963. DOI: https://doi.org/10.1137/0111030.

[88] MARSDEN, J. E. *Elementary Classical Analysis*. San Francisco: W. H. Freeman and Company, 1974. 549 p.

[89] MARTÍNEZ, J. M.; SANTOS, S. A. *Métodos Computacionais de Otimização*. Campinas, 1995. 249 p.

[90] MERA, N. S. et. al. An iterative BEM for the Cauchy steady state heat conduction problem in an anisotropic medium with unknown thermal conductivity tensor. *Inverse Probl Eng.*, v. 8, n. 6, p. 579–607, 2000. DOI: https://doi.org/10.1080/174159700088027748.

[91] MEYER, C. D. *Matrix Analysis and Applied Linear Algebra*. Philadelphia: SIAM, 2000. 718 p.

[92] MOORE, E. H. On the reciprocal of the general algebraic matrix. *Bull, Amer. Math. Soc.*, v. 26, p. 394–395, 1920.

[93] MORÉ, J.J. *The Levenberg-Marquardt algorithm: Implementation and theory*. Numerical Analysis, Lecture Notes in Mathematics series, v. 630. Berlin: Springer-Verlag, p. 105–116, 1978. DOI: https://doi.org/10.1007/BFb0067700.

[94] MORÉ, J. J. *Recent Developments in Algorithms and Software for Trust Region Methods*. Mathematical Programming The State of the Art, p. 258–287, 1983. DOI: https://doi.org/10.1007/978-3-642-68874-4_11.

[95] MOROZOV, V. A. *Methods for Solving Incorrectly Posed Problems*. New York: Springer-Verlag, 1984. 257 p. DOI: https://doi.org/10.1007/978-1-4612-5280-1.

[96] NATTERER, F. *The Mathematics of Computerized Tomography*. New York: Wiley, 2001. 226 p. DOI: https://doi.org/10.1137/1.9780898719284.

[97] NOCEDAL, J.; WRIGHT, S. J. *Numerical Optimization*. 2. ed. New York: Springer, 2006. 664 p. DOI: https://doi.org/10.1007/978-0-387-40065-5.

[98] NOWAK, W.; CIRPKA, O. A. A modified Levenberg-Marquardt algorithm for quasi-linear geostatistical inversing. *Advances in Water Resources*, v. 27, n. 7, p. 737–750, 2004. DOI: https://doi.org/10.1016/j.advwatres.2004.03.004.

[99] OSBORNE, M. R. Nonlinear least squares – The Levenberg algorithm revisited. *J. Austral. Math. Soc.*, Series B, v. 19, n. 3, p. 343–357, 1976. DOI: https://doi.org/10.1017/S033427000000120X.

[100] ÖZIŞIK, M. N. *Heat Conduction*. 2. ed. New York: John Wiley & Sons, 1993. 692 p.

[101] PAIGE, C. C.; SAUNDERS, M. A. Solution of sparse indefinite systems of linear equations. *SIAM Journal on Numerical Analysis*, v. 12, n. 4, p. 617–629, 1975. DOI: https://doi.org/10.1137/0712047.

[102] PAIGE C.C.; SAUNDERS M. A. Towards a Generalized Singular Value Decomposition. *SIAM J. Numer. Anal.*, v. 18, n. 3, p. 398–405, 1981. DOI: https://doi.org/10.1137/0718026.

[103] PAIGE, C. C.; SAUNDERS, M. A. LSQR: An algorithm for sparse linear equations and sparse least squares. *ACM Trans. Math. Softw.*, v. 8, n. 1, p. 43–71, 1982. DOI: https://doi.org/10.1145/355984.355989.

[104] PASDUNKORALE, J.; TURNER, I. W. A second order finite volume technique for simulating transport in anisotropic media. *Int J Numer Method Heat Fluid Flow*, v. 13, n. 1, p. 31–56, 2003. DOI: https://doi.org/10.1108/09615530310456750.

[105] PENROSE, R. A generalized inverse for matrices. *Proc. Cambridge Phil. Soc.*, v. 51, n. 3, p. 406–413, 1955. DOI: https://doi.org/10.1017/S0305004100030401.

[106] PEREVERZEV, S.; SCHOCK, E. Morozov's discrepancy principle for Tikhonov regularization of severely ill-posed in finite-dimensional subspaces. *Numer. Funct. Anal. and Optimiz.*, v. 21, n. 7-8, p. 901–916, 2000.

[107] RIOS, L. M.; SAHINIDIS, N. V. Derivative-free optimization: a review of algorithms and comparison of software implementations. *J. Glob. Optim.*, v. 56, p. 1247–1293, 2013. DOI: https://doi.org/10.1007/s10898-012-9951-y.

[108] ROBINSON, S. M. *Some continuity properties of polyhedral multifunctions.* Mathematical Programming at Oberwolfach. Mathematical Programming Studies, v. 14. Berlin: Springer, p. 206–214, 1981. DOI: https://doi.org/10.1007/BFb0120929.

[109] RUDIN, W. *Principles of Mathematical Analysis*. 3. ed. International series in pure and applied mathematics. New York: McGraw-Hill, 1976. 342 p.

[110] SAAD, Y.; SCHULTZ, M. GMRES: A Generalized Minimum Residual Algorithm for Solving Nonsymmetric Linear Systems. *SIAM J. Scientific and Stat. Comp.*, v. 7, n. 3, p. 856–869, 1986. DOI: https://doi.org/10.1137/0907058.

[111] SCHULTZ, G. Iterative Berechnung der Reziproken Matrix. *Z. Angew. Meth. Mech.*, v. 13, n. 1, p. 57–59, 1933. DOI: https://doi.org/10.1002/zamm.19330130111.

[112] SCHWETLICK, H.; TILLER, V. Nonstandard scaling matrices for trust region Gauss-Newton methods. *SIAM J. Sci. Stat. Comput.*, v. 10, n. 4, p. 654–670, 1989. DOI: https://doi.org/10.1137/0910040.

[113] SHI, Z.-J.; SHEN, J. A gradient-related algorithm with inexact line searches. *Journal of Computational and Applied Mathematics*, v. 170, n. 2, p. 349–370, 2004. DOI: https://doi.org/10.1016/j.cam.2003.10.025.

[114] STEWART, G. W. A note on the perturbation of singular values. *Lin. Alg. Applic.*, v. 28, p. 213–216, 1979. DOI: https://doi.org/10.1016/0024-3795(79)90134-4.

[115] TIKHONOV, A. N. et. al. *Numerical methods for the solution of ill-posed problems*. 1. ed. Mathematics and Its Applications series, v. 328. Dordrecht: Springer Science & Business Media, 1995. 262 p. DOI: https://doi.org/10.1007/978-94-015-8480-7.

[116] TOINT, P. L. Non-monotone trust region algorithm for nonlinear optimization subject to convex constraints. *Mathematical Programming*, v. 77, n. 3, p. 69–94, 1997. DOI: https://doi.org/10.1007/BF02614518.

[117] TRANSTRUM, M. K.; SETHNA, J. P. Improvements to the Levenberg-Marquardt algorithm for nonlinear least-squares minimization. *arXiv 1201.5885*, preprint, p. 1–32, 2012. Disponível em: https://arxiv.org/pdf/1201.5885.pdf. Acesso em: 1 mar. 2022.

[118] TREFETHEN, L. N.; BAU, D. *Numerical Linear Algebra*. Philadelphia: SIAM, 1997. 361 p.

[119] VAN LOAN C. F. Generalizing the Singular Value Decomposition. *SIAM J. Numer. Anal.*, v. 13, n. 1, p. 76–83, 1976. DOI: https://doi.org/10.1137/0713009.

[120] VAN LOAN, C. F. Computing the CS and the Generalized Singular Value Decompositions. *Numer. Math.*, v. 46, n. 4, p. 479–491, 1985. DOI: https://doi.org/10.1007/BF01389653.

[121] VOGEL, C. R. *Computational Methods for Inverse Problems*. Frontiers in Applied Mathematics. Philadelphia: SIAM, 2002. 179 p. DOI: https://doi.org/10.1137/1.9780898717570.

[122] YAMASHITA, N.; FUKUSHIMA, M. *On the rate of convergence of the Levenberg-Marquardt method*. Topics in Numerical Analysis, Computing Supplementa series, v. 15. Wien: Springer-Verlag, p. 239–249, 2001. DOI: https://doi.org/10.1007/978-3-7091-6217-0_18

[123] YANG, C. The determination of two moving heat sources in two-dimensional inverse heat problem. *Applied Mathematical Modelling*, v. 30, p. 278–292, 2006. DOI: https://doi.org/10.1016/j.apm.2005.03.012.

[124] ZHOU, G.; SI, J. Advanced neural-network training algorithm with reduced complexity based on Jacobian deficiency. *IEEE Transactions on Neural Networks*, v. 9, n. 3, p. 448–453, 1998. DOI: https://doi.org/10.1109/72.668886.

APÊNDICE A – TRANSFORMAÇÃO DO PROBLEMA DE TIKHONOV PARA A FORMA PADRÃO

Sejam $A \in \mathbb{R}^{m\times n}$, $\tilde{b} = b + e \in \mathbb{R}^m$ (caso de sistema linear com ruído nos dados exatos b, embora a situação sem ruído tenha a mesma argumentação) e $L \in \mathbb{R}^{p\times n}$ (observe que não impomos a restrição $p \leq n$ aqui). Assim, associamos ao problema de mínimos quadrados associado a $Ax = \tilde{b}$ a chamada *forma geral da regularização de Tikhonov* [115]

$$x_{L,\lambda} = \underset{x\in\mathbb{R}^n}{\operatorname{argmin}} \left\{ \|Ax - b\|_2^2 + \lambda^2 \|L(x - x_0)\|_2^2 \right\}, \tag{A.1}$$

em que $\lambda > 0$ é o parâmetro de regularização e x_0 é uma aproximação inicial da solução de $Ax = b$, se disponível, caso contrário tomamos $x_0 = \mathbf{0}$. A matriz L, como vimos na Seção 2.2, busca introduzir dados extra do problema original.

Observação A.1. Perceba que podemos considerar que $x_0 = \mathbf{0}$ em (2.8), uma vez que é sempre possível realizar a mudança de variável $\hat{x} = x - x_0$, resultando no problema

$$\hat{x}_{L,\lambda} = \underset{\hat{x}\in\mathbb{R}^n}{\operatorname{argmin}} \left\{ \|A\hat{x} - \hat{b}\|_2^2 + \lambda^2 \|L\hat{x}\|_2^2 \right\}, \tag{A.2}$$

com $\hat{b} = b - Ax_0$. Em seguida, para retornar ao problema original, basta tomar $x_{L,\lambda} = \hat{x}_{L,\lambda} + x_0$.

O objetivo deste apêndice é apresentar formas de efetuar as mudanças de variáveis devidas de modo a transformarmos (A.1) no *problema de Tikhonov na forma padrão*

$$\bar{x}_\lambda = \underset{\bar{x}\in\mathbb{R}^p}{\operatorname{argmin}} \left\{ \|\bar{A}\bar{x} - \bar{b}\|_2^2 + \lambda^2 \|\bar{x} - \bar{x}_0\|_2^2 \right\}, \tag{A.3}$$

de modo que os problemas sejam equivalentes e que tenhamos como "caminhar" livremente entre eles, ou seja, que possamos interpretar a solução do problema transformado como solução da sua forma geral e vice-versa. Para o que apresentaremos, assuma que $\mathcal{N}(A) \cap \mathcal{N}(L) = \{\mathbf{0}\}$ e que $L \in \mathbb{R}^{p\times n}$ possui posto completo, isto é, $\text{posto}(L) = \min\{p, n\}$.

A.1 A TRANSFORMAÇÃO

Como vimos na Seção 2.2, se L for não singular ($p = n$), então $\bar{A} = AL^{-1}$, $\bar{b} = \tilde{b}$ e $\bar{x}_0 = Lx_0$. O retorno ao problema original vem de resolver $Lx = \bar{x}$. Em mais um exemplo, caso $p \geq n$, então segue que L^TL é não singular. Assim, a pseudo-inversa de L é dada por

$$L^\dagger = (L^TL)^{-1}L^T,$$

o que implica que L possui uma inversa à esquerda, pois $L^\dagger L = I$. Portanto,

$$\begin{aligned} & \|Ax - \tilde{b}\|_2^2 + \lambda^2 \|L(x - x_0)\|_2^2 \\ \Leftrightarrow \quad & \|(AL^\dagger)Lx - \tilde{b}\|_2^2 + \lambda^2 \|Lx - Lx_0\|_2^2. \end{aligned}$$

Logo, em um paralelo ao caso L não singular, segue que $\bar{A} = AL^{\dagger}$, $\bar{b} = \tilde{b}$ e $\bar{x}_0 = Lx_0$ e $Lx = \bar{x}$ (i.e., $x = L^{\dagger}\bar{x}$).

Para L geral, é preciso ter mais cautela especialmente por conta da existência de núcleo não trivial para L, como acontece quando $p < n$. De qualquer forma, a transformação existe, e é o foco do texto a seguir. Para isso, vamos precisar introduzir alguns conceitos, como o de projeções oblíquas e pseudo-inversa oblíqua.

Definição A.1 (Projetor, [65]). *Sejam $\mathcal{X}$ e $\mathcal{Y}$ subespaços de $\mathbb{R}^n$ tais que $\mathcal{X} \cap \mathcal{Y} = \{\mathbf{0}\}$. Então, a* projeção em $\mathcal{X}$ em relação à $\mathcal{Y}$ *é o operador linear (chamado de* projetor*) $E_{\mathcal{X},\mathcal{Y}}$ que satisfaz as seguintes condições:*

(i) $E_{\mathcal{X},\mathcal{Y}}\, x = x,\ \forall x \in \mathcal{X}$;

(ii) $E_{\mathcal{X},\mathcal{Y}}\, y = \mathbf{0},\ \forall y \in \mathcal{Y}$;

(iii) $E_{\mathcal{X},\mathcal{Y}}\, z \in \mathcal{X},\ \forall z \in \mathbb{R}^n$.

No caso particular em que $\mathcal{Y} = \mathcal{X}^{\perp}$, em que $\mathcal{X}^{\perp}$ denota o complemento ortogonal de $\mathcal{X}$, denotamos $E_{\mathcal{X},\mathcal{X}^{\perp}}$ por $P_{\mathcal{X}}$ e o chamamos de *projetor ortogonal*. Os projetores que não são ortogonais são usualmente chamados de *projetores oblíquos*.

Definição A.2 (Produto interno e complemento oblíquos, [65]). *Para $x, y \in \mathbb{R}^n$ e $A \in \mathbb{R}^{m\times n}$, utilizamos a notação $x \perp_A y$ quando $Ax \perp Ay$, ou seja, $\langle Ax, Ay\rangle \equiv x^T A^T A y = 0$. Se $\mathcal{X}$ é subespaço de $\mathbb{R}^n$, definimos o* complemento oblíquo *por*

$$\mathcal{X}^{\perp_A} = \{y \in \mathbb{R}^n \mid x \perp_A y,\ \forall x \in \mathcal{X}\}.$$

A ideia por trás da transformação para a forma padrão está em considerar que

$$x = x_{\mathcal{M}} + x_{\mathcal{N}}, \quad \text{com} \quad x_{\mathcal{N}} \in \mathcal{N}(L),$$

o que implica que

$$Ax = Ax_{\mathcal{M}} + Ax_{\mathcal{N}} \quad \text{e} \quad Lx = Lx_{\mathcal{M}}.$$

Aplicando no problema (A.1) com $x_0 = \mathbf{0}$ (possível pela Observação A.1), temos

$$x_{L,\lambda} = \underset{x\in\mathbb{R}^n}{\operatorname{argmin}} \left\{ \|Ax_{\mathcal{M}} + Ax_{\mathcal{N}} - \tilde{b}\|_2^2 + \lambda^2 \|Lx_{\mathcal{M}}\|_2^2 \right\} \tag{A.4}$$

No entanto, com algumas operações algébricas simples, verificamos que

$$\begin{aligned} \|Ax_{\mathcal{M}} + Ax_{\mathcal{N}} - \tilde{b}\|_2^2 &= \langle Ax_{\mathcal{M}} + Ax_{\mathcal{N}} - \tilde{b}, Ax_{\mathcal{M}} + Ax_{\mathcal{N}} - \tilde{b}\rangle \\ &= \|Ax_{\mathcal{M}} - \tilde{b}\|_2^2 + \|Ax_{\mathcal{N}} - \tilde{b}\|_2^2 - \|\tilde{b}\|_2^2 + \\ &\quad 2\langle Ax_{\mathcal{M}}, Ax_{\mathcal{N}}\rangle. \end{aligned}$$

Assumindo que $x_{\mathcal{M}} \perp_A x_{\mathcal{N}}$, segue que $\langle Ax_{\mathcal{M}}, Ax_{\mathcal{N}}\rangle = 0$. Assim, como $\|\tilde{b}\|_2^2$ é constante e não influencia na minimização, podemos dividir (A.4) em dois problemas independentes:

$$\min_{x_{\mathcal{M}}} \left\{\|Ax_{\mathcal{M}} - \tilde{b}\|_2^2 + \lambda^2\|Lx_{\mathcal{M}}\|_2^2\right\} \quad \text{e} \quad \min_{x_{\mathcal{N}}} \left\{\|Ax_{\mathcal{N}} - \tilde{b}\|_2^2\right\}. \tag{A.5}$$

Perceba que, ao pedirmos que $x_{\mathcal{M}} \perp_A x_{\mathcal{N}}$, estamos sugerindo que $\mathbb{R}^n = \mathcal{N}(L)^{\perp_A} + \mathcal{N}(L)$. Desta forma, da teoria de projetores oblíquos (veja [43, 44, 64, 66] para maiores detalhes), temos os seguintes projetores associados:

$$E_{\mathcal{N}} = E_{\mathcal{N}(L),\,\mathcal{N}(L)^{\perp_A}} \quad \text{e} \quad E_{\mathcal{M}} = E_{\mathcal{N}(L)^{\perp_A},\,\mathcal{N}(L)} = I - E_{\mathcal{N}}.$$

Disto e percebendo que $E_{\mathcal{M}} + E_{\mathcal{N}} = I$, temos

$$x = x_{\mathcal{M}} + x_{\mathcal{N}} = E_{\mathcal{M}}x + E_{\mathcal{N}}x.$$

Assim, segue que (A.5) se transforma nos seguintes problemas de minimização:

$$\min_{x\in\mathbb{R}^n} \left\{\|AE_{\mathcal{M}}x - \tilde{b}\|_2^2 + \lambda^2\|LE_{\mathcal{M}}x\|_2^2\right\} \quad \text{e} \tag{A.6}$$

$$\min_{x\in\mathbb{R}^n} \left\{\|AE_{\mathcal{N}}x - \tilde{b}\|_2^2\right\}. \tag{A.7}$$

Veja que, se $W \in \mathbb{R}^{n\times(n-p)}$ é tal que $\mathcal{R}(W) = \mathcal{N}(L)$, então existe $z_{\mathcal{N}} \in \mathbb{R}^{n-p}$ tal que $x_{\mathcal{N}} = Wz_{\mathcal{N}}$ e o problema (A.7) se torna

$$\min_{z\in\mathbb{R}^{n-p}} \left\{\|AWz - \tilde{b}\|_2^2\right\}, \tag{A.8}$$

que possui a solução de norma mínima dada por $z_{\mathcal{N}} = (AW)^\dagger\tilde{b}$. Portanto, segue que $x_{\mathcal{N}} = W(AW)^\dagger\tilde{b}$.

Já para (A.6), alguns detalhes adicionais das projeções oblíquas são necessários, porém omitimos pela brevidade do material. No entanto, é possível verificar, e aqui referenciamos Hansen [65, 66], que

$$E_{\mathcal{M}} = L^\dagger_{\mathcal{N}(L)^{\perp_A}} L,$$

em que $L^\dagger_{\mathcal{N}(L)^{\perp_A}}$ representa a chamada *pseudo-inversa oblíqua* (definição abaixo) e é o último elo faltante para efetuarmos a transformação propriamente dita. Normalmente, para simplificar a notação, a pseudo-inversa oblíqua é denotada por $L^\dagger_A$.

Definição A.3 (Pseudo-inversa oblíqua, [64]). *Para $A \in \mathbb{R}^{m\times n}$ e $L \in \mathbb{R}^{p\times n}$, definimos a* inversa generalizada de peso-A de L *(ou* pseudo-inversa oblíqua*) por*

$$L^\dagger_A = \left(I - \left(A\left(I - L^\dagger L\right)\right)^\dagger A\right) L^\dagger. \tag{A.9}$$

Caso $p \geq n$ e L tiver posto completo, segue que $L^\dagger_A = L^\dagger$ (pois $L^\dagger L = I$). No entanto, em geral, $L^\dagger_A$ pode ser bastante diferente de $L^\dagger$. Em mais um fato omitido [65], é possível provar que

$$AE_{\mathcal{M}} = \left(AL^\dagger_A\right) L \quad \text{e} \quad LE_{\mathcal{M}} = L.$$

Finalmente, juntando as informações acima, podemos escrever

$$\begin{aligned} & \|Ax - \tilde{b}\|_2^2 + \lambda^2 \|L(x - x_0)\|_2^2 \\ \Leftrightarrow \quad & \|A(E_{\mathcal{M}} x + x_{\mathcal{N}}) - \tilde{b}\|_2^2 + \lambda^2 \|L(E_{\mathcal{M}} x + x_{\mathcal{N}}) - Lx_0\|_2^2 \\ \Leftrightarrow \quad & \|AL_A^{\dagger} Lx - (\tilde{b} - Ax_{\mathcal{N}})\|_2^2 + \lambda^2 \|Lx - Lx_0\|_2^2. \end{aligned}$$

Disto, conseguimos concluir a transformação do problema de Tikhonov na forma geral para a forma padrão, que é dada por

$$\bar{A} = AL_A^{\dagger}, \quad \bar{b} = \tilde{b} - Ax_{\mathcal{N}} \quad \text{e} \quad \bar{x}_0 = Lx_0, \tag{A.10}$$

em que $\bar{x} = Lx$ e a transformação reversa é dada por

$$x_{L,\lambda} = L_A^{\dagger} \bar{x}_{\lambda} + x_{\mathcal{N}}. \tag{A.11}$$

Da resolução de (A.8), temos que $x_{\mathcal{N}} = W(AW)^{\dagger}\tilde{b}$. No entanto, é possível apresentar uma outra forma para $x_{\mathcal{N}}$, usando as informações já disponíveis do problema, dada por

$$x_{\mathcal{N}} = \left(A\left(I - L^{\dagger}L\right)\right)^{\dagger} \tilde{b}. \tag{A.12}$$

Caso a GSVD de (A, L) (Teorema 2.2) esteja disponível, temos que

$$L_A^{\dagger} = X \begin{bmatrix} M^{-1} \\ \mathbf{0} \end{bmatrix} V^T \quad \text{e} \quad x_{\mathcal{N}} = \sum_{i=p+1}^{n} u_i^T \tilde{b} x_i,$$

em que u_i e x_i representam o vetor da i-ésima coluna de U e X, respectivamente. Neste caso, a forma final da pseudo-inversa oblíqua é mais amigável e torna o uso da transformação para a forma padrão do problema de Tikhonov uma tarefa mais simples. Nas aplicações práticas, no entanto, o custo de computar tal decomposição, assim como a própria SVD, em geral descarta essa possibilidade. Utilizando a caracterização acima,

$$\bar{A} = AL_A^{\dagger} = U_p \Sigma M^{-1} V^T, \tag{A.13}$$

em que U_p é a matriz com as primeiras p colunas de U, ou seja, $U_p = [u_1, \ldots, u_p]$. Assim, veja que a GSVD do par (A, L) está intimamente relacionada com a SVD de $\bar{A}$, uma vez que os valores singulares generalizados γ_i são os valores singulares de $\bar{A}$, exceto pela ordem invertida que aparecem na decomposição. A equação acima também permite, diretamente, escrever

$$AL_A^{\dagger} = \sum_{i=1}^{p} \gamma_i u_i v_i^T. \tag{A.14}$$

Agora, do ponto de vista computacional, vale mencionar duas diferenças na implementação da forma padrão da regularização de Tikhonov, quando estamos lidando com métodos diretos e iterativos. Para o primeiro caso, é preciso computar $\bar{A}$ explicitamente; preferencialmente, buscamos fazer uso de transformações ortogonais para tanto, por questões de estabilidade. Já para métodos iterativos, não precisamos de $\bar{A}$ em sua forma explícita, sendo apenas necessário que possamos acessá-la através de produtos da forma $\bar{A}x$ e $\bar{A}^T x$ de maneira eficiente. Nas duas subseções seguintes, discutiremos alguns detalhes nesse sentido.

A.1.1 Formulação para métodos diretos

A transformação explícita para métodos diretos foi desenvolvida por Eldén [43] e é baseada em duas decomposições QR. Aqui, considere os subíndices p, o e q, significando que as matrizes possuem p, $n-p$ e $m-(n-p)$ colunas, respectivamente. Assim, inicialmente, computamos a QR de L^T:

$$L^T = KR = \begin{bmatrix} K_p & K_o \end{bmatrix} \begin{bmatrix} R_p \\ \mathbf{0} \end{bmatrix}.$$

Segundo Hansen [64], esta fatoração QR consome em torno de $2n(n-p+1)^2$ *flops* se rotações de Givens são utilizadas, assumindo que L é matriz banda com largura da banda dada por $n-p+1$ (que é uma situação recorrente em aplicações). Como L possui posto completo, temos $L^\dagger = K_p R_p^{-T}$. Além disso, as colunas de K_o formam uma base ortonormal para $\mathcal{N}(L)$. Agora, computamos a fatoração QR da matriz $AK_o \in \mathbb{R}^{m\times(n-p)}$ (que possui poucas colunas se $p \approx n$, algo não raro de ocorrer), dada por

$$AK_o = HT = \begin{bmatrix} H_o & H_q \end{bmatrix} \begin{bmatrix} T_o \\ \mathbf{0} \end{bmatrix}.$$

Tal fatoração consume algo em torno de $2m(n-p)^2$ *flops*. De posse das duas decomposições, podemos computar as outras matrizes e vetores envolvidos na transformação para a forma padrão. Como $A\left(I_n - L^\dagger L\right) = AK_oK_o^T = H_oT_oK_o^T$, segue que $\left(A\left(I_n - L^\dagger L\right)\right)^\dagger = K_oT_o^{-1}H_o^T$ e, portanto, das equações (A.9) e (A.12), temos

$$L_A^\dagger = \left(I_n - K_oT_o^{-1}H_o^TA\right)L^\dagger \quad \text{e} \quad x_{\mathcal{N}} = K_oT_o^{-1}H_o^T\tilde{b}.$$

Do fato de que $AK_oT_o^{-1}H_o^T = H_oH_o^T$, temos

$$\begin{aligned}\bar{A} &= A\left(I_n - K_oT_o^{-1}H_o^TA\right)L^\dagger = \left(I_m - H_oH_o^T\right)AL^\dagger = H_qH_q^TAL^\dagger,\\ \bar{b} &= \tilde{b} - AK_oT_o^{-1}H_o^T\tilde{b} = \left(I_m - H_oH_o^T\right)\tilde{b} = H_qH_q^T\tilde{b}.\end{aligned}$$

Podemos simplificar ainda mais notando que ao substituir as quantidades acima em $\|\bar{A}\bar{x} - \bar{b}\|_2$, o termo H_q à esquerda de $\bar{A}$ e $\bar{b}$ não contribui para a 2-norma. Deste modo, é conveniente fazer uso de versões levemente diferentes de $\bar{A}$ e $\bar{b}$, em que aquele fator é omitido. Assim, tomamos

$$\bar{A}' = H_q^T\bar{A} = H_qAL^\dagger = H_qAK_pR_p^{-T} \quad \text{e} \quad \bar{b}' = H_q^T\bar{b} = H_q^T\tilde{b} \tag{A.15}$$

como tais novas quantidades. Uma vez que o problema transformado é resolvido para $\bar{x}_\lambda$, então a transformação de retorno para a forma geral é dada por

$$x_{L,\lambda} = L^\dagger\bar{x}_\lambda + K_oT_o^{-1}H_o^T\left(\tilde{b} - AL^\dagger\bar{x}_\lambda\right),$$

o que conclui o ciclo para os métodos diretos.

A.1.2 Formulação para métodos iterativos

Como já mencionado acima, para métodos iterativos, é interessante que possamos evitar o cálculo de $\bar{A}$ ou $\bar{A}'$ explicitamente. Muitas vezes, nem a própria matriz A é dada desta forma, mas apenas acessada via produtos matriz-vetor envolvendo A e A^T. Portanto, é vantajoso fazer uso de uma transformação implícita, explorando as multiplicações matriciais contidas em (A.10) ou (A.15), por exemplo. Neste trabalho, abordaremos o primeiro caso, isto é, veremos uma maneira de efetuar as avaliações envolvendo $\bar{A}$ a partir das informações contidas em (A.10) e em equações relacionadas.

Do que foi desenvolvido neste apêndice, efetuar a transformação para a forma padrão (equações (A.10) e (A.11)) necessita basicamente da capacidade de avaliar termos com $L_A^\dagger$ e $x_\mathcal{N}$, dados abaixo:

$$L_A^\dagger = \left(I - \left(A\left(I - L^\dagger L\right)\right)^\dagger A\right) L^\dagger \quad \text{e} \quad x_\mathcal{N} = \left(A\left(I - L^\dagger L\right)\right)^\dagger \tilde{b}.$$

Suponha que temos acesso a uma matriz W tal que $\mathcal{R}(W) = \mathcal{N}(L)$. Então a pseudo-inversa oblíqua tem uma forma ligeiramente mais amigável, dada por

$$L_A^\dagger = \left(I - W(AW)^\dagger A\right) L^\dagger.$$

Disto e usando o resultado da equação (A.8), segue que

$$\bar{A} = AL_A^\dagger = A\left(I - W(AW)^\dagger A\right) L^\dagger \quad \text{e} \quad x_\mathcal{N} = W(AW)^\dagger \tilde{b},$$

que é a formulação que estamos interessados aqui. A ideia geral por trás da avaliação de $\bar{A}x$ ou $\bar{A}^T x$ está em fazer as operações por partes, fazendo uso do conhecimento de W. Construímos também a matriz $T = (AW)^\dagger A$, a qual se espera possuir poucas colunas. Vale comentar que, se a dimensão de $\mathcal{N}(L)$ for pequena, então W também tem poucas colunas; portanto, operações envolvendo W e T tendem a ser pouco custosas do ponto de vista computacional. Em consonância, $(AW)^\dagger$ pode ser eficientemente computada via uma decomposição QR de AW.

Por outro lado, efetuar avaliações envolvendo $L^\dagger$ pode demandar maior cuidado. Sabemos que calcular $y = L^\dagger x$ é equivalente a encontrar a solução de norma mínima do problema $\min_{y\in\mathbb{R}^n} \|Ly - x\|_2$. Analogamente, como $\left(L^\dagger\right)^T = \left(L^T\right)^\dagger$, segue que computar $y = \left(L^\dagger\right)^T x$ é o mesmo que gerar a solução de norma mínima de $\min_{y\in\mathbb{R}^p} \|L^T y - x\|_2$. Em ambos os casos, a preocupação se encontra na maneira como os problemas de minimização serão resolvidos. Dois aspectos centrais precisam ser observados: velocidade e precisão. O primeiro é necessário à viabilidade do algoritmo; o segundo, por sua vez, considera que, como estamos tratando com métodos iterativos, os erros de arredondamento são levados adiante e podem acabar por comprometer a solução final. É preciso levar a estrutura de L em conta e, nesse sentido, cada problema é único, demandando uma sub-rotina apropriada. Independente do caso, os Algoritmos A.1 e A.2 abaixo sugerem uma maneira de efetuar avaliações com $\bar{A}$ e $\bar{A}^T$.

Algoritmo A.1 Avaliação de $z = \bar{A}x$ implicitamente

1: $x_1 \leftarrow L^\dagger x$ ▷ equivalente a resolver $\min \|Lx_1 - x\|_2$
2: $x_2 \leftarrow W(Tx_1)$
3: $x_3 \leftarrow x_1 - x_2$
4: $z \leftarrow Ax_3$

Algoritmo A.2 Avaliação de $z = (\bar{A})^T x$ implicitamente

1: $x_1 \leftarrow A^T x$
2: $x_2 \leftarrow T^T(W^T x_1)$
3: $x_3 \leftarrow x_1 - x_2$
4: $z \leftarrow \left(L^\dagger\right)^T x_3$ ▷ equivalente a resolver $\min \|L^T z - x_3\|_2$

Veja que estes algoritmos são facilmente adaptáveis para efetuar operações com $L_A^\dagger$, necessárias à transformação de retorno em (A.11). Também vale observar que computar $x_\mathcal{N} = W(AW)^\dagger b$ é pouco custoso, levando em conta que, na prática, $\mathcal{N}(L)$ tem dimensão pequena.

Observação A.2. O desenvolvimento acima considera um caso geral para a matriz L, sem levar em conta nenhum tipo de propriedade que a mesma possa conter. Um caso interessante é quando $p \leq n$, nos levando a escrever $L \in \mathbb{R}^{p\times n}$ da forma

$$L = \begin{bmatrix} L_{11} & L_{12} \end{bmatrix}, \tag{A.16}$$

em que $L_{11} \in \mathbb{R}^{p\times p}$ é não singular. Efetuando um particionamento similar em $T = (AW)^\dagger A$ e $x \in \mathbb{R}^n$, temos

$$T = \begin{bmatrix} T_{11} & T_{12} \end{bmatrix} \qquad \text{e} \qquad x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \tag{A.17}$$

em que $T_{11} \in \mathbb{R}^{(n-p)\times p}$ e $x_1 \in \mathbb{R}^p$. Como verificado em [44], neste caso, a pseudo-inversa oblíqua é dada por

$$L_A^\dagger = [I_n - WT] \begin{bmatrix} L_{11} \\ \mathbf{0} \end{bmatrix} = \left[\begin{bmatrix} I_p \\ \mathbf{0} \end{bmatrix} - WT_{11} \right] L_{11}^{-1}. \tag{A.18}$$

Assim, podemos gerar algoritmos paralelos aos desenvolvidos acima, mas adaptados para este caso particular, apresentados nos Algoritmos A.3 e A.4. O passo mais custoso corresponde a resolver sistemas lineares envolvendo L_{11} (similar às operações com $L^\dagger$ acima), que precisam ser efetuados com cautela. A transformação de retorno também é facilmente derivável a partir das equações e algoritmos apresentados.

A.2 ABORDAGEM *SMOOTHING NORM* (SN)

Na Subseção 2.2.1, vimos como utilizar o problema transformado de Tikhonov junto de métodos iterativos como forma de introduzir informações a priori nas soluções

Algoritmo A.3 Avaliação de $z = \bar{A}x$ implicitamente (caso particular)

1: $x_1 \leftarrow L_{11}^{-1}x$ ▷ equivalente a resolver $L_{11}x_1 = x$
2: $x_2 \leftarrow \begin{bmatrix} x_1 \\ \mathbf{0} \end{bmatrix} - W(T_{11}x_1)$
3: $z \leftarrow Ax_2$

Algoritmo A.4 Avaliação de $z = (\bar{A})^T x$ implicitamente (caso particular)

1: $x_1 \leftarrow A^T x$ ▷ note: $x_1 = \begin{bmatrix} \hat{x}_1 \\ \hat{x}_2 \end{bmatrix}$, com $\hat{x}_1 \in \mathbb{R}^p$
2: $x_2 \leftarrow \hat{x}_1 - T_{11}^T(W^T x_1)$
3: $z \leftarrow L_{11}^{-T}x_2$ ▷ equivalente a resolver $L_{11}^T z = x_2$

computadas. Aqui, apresentamos uma segunda forma de utilizar a transformação com este mesmo propósito, desenvolvida recentemente por Hansen e Jensen [68]. A denominação da estratégia deles vem do fato de que a matriz L define a chamada *smoothing norm* $\|L \cdot\|_2$ (algo como "norma suavizadora", em uma tradução literal, embora seja somente uma seminorma em geral), que atua como regularizador.

Como vimos, a transformação para a forma padrão infere que $x = L_A^\dagger \bar{x} + x_\mathcal{N}$, em que $x_\mathcal{N} = W z_\mathcal{N}$, para $z_\mathcal{N} = (AW)^\dagger \tilde{b}$. Ou seja, x é escrito como a soma de uma componente em $\mathcal{R}(L_A^\dagger)$ e outra em $\mathcal{R}(W) = \mathcal{N}(L)$. Os vetores $\bar{x}$ e $z_\mathcal{N}$ são unicamente determinados, uma vez que L e W possuem posto completo. Deste modo, o sistema $Ax = \tilde{b}$ pode ser formulado como

$$A \begin{bmatrix} L_A^\dagger & W \end{bmatrix} \begin{bmatrix} \bar{x} \\ z_\mathcal{N} \end{bmatrix} = \tilde{b}.$$

Multiplicando a equação acima por $\begin{bmatrix} L_A^\dagger & W \end{bmatrix}^T$, produzimos o sistema em blocos 2×2 dado por

$$\begin{bmatrix} L_A^{\dagger T} A L_A^\dagger & L_A^{\dagger T} AW \\ W^T A L_A^\dagger & W^T AW \end{bmatrix} \begin{bmatrix} \bar{x} \\ z_\mathcal{N} \end{bmatrix} = \begin{bmatrix} L_A^{\dagger T} \tilde{b} \\ W^T \tilde{b} \end{bmatrix}.$$

A componente $z_\mathcal{N}$ pode ser removida a partir do uso do chamado *complemento de Schur* (veja [56, 118], por exemplo), conduzindo ao sistema linear $S\bar{x} = d$, em que S e d tem a estrutura:

$$S = L_A^{\dagger T} A L_A^\dagger - L_A^{\dagger T} AW(W^T AW)^{-1} W^T A L_A^\dagger = L_A^{\dagger T} P A L_A^\dagger \quad \text{e} \tag{A.19}$$

$$d = L_A^{\dagger T} \tilde{b} - L_A^{\dagger T} AW(W^T AW)^{-1} W^T \tilde{b} = L_A^{\dagger T} P \tilde{b}, \tag{A.20}$$

para $P := I - AW(W^T AW)^{-1} W^T$. Veja que, como esperamos que W possua poucas colunas, computar $(W^T AW)^{-1}$ apresenta dificuldade reduzida, mesmo que trabalhar com inversas de matrizes seja algo pouco desejável na prática. Assim, a esperança recai em $S\bar{x} = d$: calcular soluções estáveis para este sistema pode gerar boas aproximações para x^*, assim que a transformação de retorno (A.11) for aplicada. Note que a influência de A e L estão contidas em S e d, tornando razoável tal ideia.

O sistema gerado por (A.19) e (A.20) possui diversas propriedades interessantes, estudadas detalhadamente em [68]. Por exemplo, caso A seja simétrica, então $S = L_A^{\dagger T} P A L_A^{\dagger}$ também é simétrica. Além disso, caso $\mathcal{R}(L^T)$ e $\mathcal{R}(AW)$ são espaços complementares, então a matriz P é o projetor oblíquo em $\mathcal{R}(L^T)$ em relação à $\mathcal{R}(AW)$. Sob esta hipótese de complementaridade entre $\mathcal{R}(L^T)$ e $\mathcal{R}(AW)$, segue também que $S\bar{x} = d$ pode ser escrito simplesmente como

$$L^{\dagger T} P A L^{\dagger} \bar{x} = L^{\dagger T} P \tilde{b}, \tag{A.21}$$

ou seja, podemos "trocar" $L_A^{\dagger}$ por $L^{\dagger}$, o que evita o uso da pseudo-inversa oblíqua, o que simplifica enormemente as operações numéricas envolvidas, tornando a estratégia SN menos custosa. Vale lembrar que a pseudo-inversa oblíqua, mesmo neste caso, ainda é necessária para a transformação de retorno.

Do ponto de vista de implementação computacional de técnicas no sistema $S\bar{x} = d$, é preciso levar em conta como efetuar operações envolvendo S e S^T. Inicialmente, computamos uma fatoração QR reduzida para AW:

$$AW = Q_0 R_0. \tag{A.22}$$

Isto implica que $(AW)^{\dagger} = R_0^{-1} Q_0^T$ e, portanto,

$$x_{\mathcal{N}} = W(AW)^{\dagger} \tilde{b} = W R_0^{-1} Q_0^T \tilde{b}.$$

Mais ainda, esta fatoração leva a

$$AW(W^T AW)^{-1} W^T = Q_0 R_0 (W^T Q_0 R_0)^{-1} W^T = Q_0 (W^T Q_0)^{-1} W^T, \tag{A.23}$$

o que induz uma maneira de efetuar operações matriz-vetor com P basicamente efetuando multiplicações com matrizes com poucas colunas. Isto é, veja que avaliar Px é o mesmo que

$$Px = x - Q_0 (W^T Q_0)^{-1} W^T x. \tag{A.24}$$

Neste passo, alguma fatoração pré-calculada para a pequena matriz $W^T Q_0$ poderia também ser utilizada. Deste modo, temos as ferramentas necessárias para operações da forma $z = Sx$ e $z = S^T x$, em que operações envolvendo $L_A^{\dagger}$ e sua transposta deveriam usar estratégias como as dos Algoritmos A.1 e A.2. Para maiores informações e propriedades, novamente recomendamos a leitura de Hansen e Jensen [68].

APÊNDICE B – LIMITES SUPERIORES E INFERIORES

Este apêndice contém excertos básicos relacionados aos conceitos de limite superior e limite inferior de sequências, retirados principalmente de [84, 88, 109], as quais contém demonstrações dos fatos que apresentaremos, aos interessados. Estes livros são também uma boa fonte de informação inicial para casos mais gerais, como por exemplo estes mesmos limites com sequências de conjuntos e funções.

Por simplicidade na explanação e manutenção do contexto visto neste trabalho, considere o espaço dos números reais $\mathbb{R}$ e uma sequência $\{x_k\} \subseteq \mathbb{R}$. É fato conhecido que nem toda sequência em $\mathbb{R}$ possui limite, ou mesmo pontos de acumulação neste espaço. Porém, ao considerarmos o *conjunto dos números reais estendidos* $\bar{\mathbb{R}} := \mathbb{R} \cup \{-\infty, \infty\}$ (às vezes denotado por $[-\infty, \infty]$), mesmo sequências divergentes possuem "pontos de acumulação". Por exemplo, se $x_k = k$, $k \in \mathbb{N}$, então $\{x_k\}$ não possui pontos de acumulação em $\mathbb{R}$ pois é divergente, mas podemos entender que $x_k \to \infty$ em $\bar{\mathbb{R}}$. Assim, utilizamos os reais estendidos aqui como forma de garantir que toda sequência em $\mathbb{R}$ admite pontos de acumulação em, no máximo, $\bar{\mathbb{R}}$, para evitar separação em casos.

Definição B.1 (Cota superior e supremo). *Seja $S \subseteq \mathbb{R}$ conjunto. Dizemos que $b \in \mathbb{R}$ é uma* cota superior de S *se $x \leq b$, para todo $x \in S$. Dizemos que uma cota superior $\bar{b}$ é* supremo de S*, com notação $\bar{b} = \sup(S)$, se $\bar{b} \leq b$, para todo b cota superior de S.*

Definição B.2 (Cota inferior e ínfimo). *Seja $S \subseteq \mathbb{R}$ conjunto. Dizemos que $a \in \mathbb{R}$ é uma* cota inferior de S *se $a \leq x$, para todo $x \in S$. Dizemos que uma cota inferior $\bar{a}$ é* ínfimo de S*, com notação $\bar{a} = \inf(S)$, se $a \leq \bar{a}$, para todo a cota superior de S.*

Em palavras, o supremo de um conjunto corresponde à menor das suas cotas superiores. Analogamente, o ínfimo é a maior das cotas inferiores. Se S for um conjunto ilimitado que não contenha cotas superiores/inferiores em $\mathbb{R}$, basta considerar os reais estendidos.

Exemplo B.1. Encontremos supremo e ínfimo dos conjuntos seguintes:

- $S = [1, 2)$. Veja que qualquer real maior ou igual a 2 é cota superior de S, ou seja, o conjunto das cotas superiores é $[2, \infty] \subseteq \bar{\mathbb{R}}$. Destas, a menor é 2. Portanto, $\sup(S) = 2$. Analogamente, $[-\infty, 1]$ corresponde ao conjunto das cotas inferiores de S, das quais a maior é 1, ou seja, $\inf(S) = 1$. Observe que $\inf(S) \in S$, porém $\sup(S) \notin S$, deixando claro que supremo e ínfimo não necessariamente pertencem ao conjunto em questão.

- $S = \{1, 2, 4, 8, 16, 32, \dots\} \equiv \{2^k\}_{k \in \mathbb{N}}$. Neste caso, o ínfimo ocorre similarmente ao caso anterior, uma vez que todo real em $[-\infty, 1]$ é cota inferior de S; logo, $\inf(S) = 1$. Por outro lado, para qualquer $x \in \mathbb{R}$ escolhido, sempre existe um índice k_0 natural

tal que $2^{k_0} > x$, levando à conclusão que cotas superiores de S não pertencem aos reais. Analisando a reta real estendida, claramente apenas ∞ se caracteriza como cota superior de S. Assim, podemos concluir que $\sup(S) = \infty$.

Com a noção de supremo e ínfimo, podemos definir o conceito de limite superior e inferior.

Definição B.3 (Limite superior e limite inferior). *Seja $\{x_k\} \subseteq \mathbb{R}$ sequência. Definimos o* limite superior *e o* limite inferior de $\{x_k\}$ *por*

$$\limsup(x_k) = \sup\{x \in \bar{\mathbb{R}} \mid x \text{ é ponto de acumulação de } \{x_k\}\} \quad e$$
$$\liminf(x_k) = \inf\{x \in \bar{\mathbb{R}} \mid x \text{ é ponto de acumulação de } \{x_k\}\},$$

respectivamente. Outras notações:

$$\limsup(x_k) = \lim_{k\to\infty}\sup(x_k) = \overline{\lim}(x_k) \quad e$$
$$\liminf(x_k) = \lim_{k\to\infty}\inf(x_k) = \underline{\lim}(x_k).$$

A definição é clara em afirmar que estes limites buscam analisar os extremos dos pontos de acumulação da sequência. Agora, se definimos conjuntos da forma

$$\begin{aligned} S_0 &= \{x_1, x_2, x_3, \dots\}, \\ S_1 &= \{x_2, x_3, x_4, \dots\}, \\ &\vdots \\ S_k &= \{x_{k+1}, x_{k+2}, x_{k+3}, \dots\}, \end{aligned}$$

e tomando $a_k := \inf(S_k)$ e $b_k := \sup(S_k)$, então temos

$$a_0 \leq a_1 \leq a_2 \leq \cdots \leq \cdots \leq b_2 \leq b_1 \leq b_0.$$

Como $\{a_k\}$ e $\{b_k\}$ são sequências em $[-\infty, \infty]$ monótonas, então $\{a_k\}$ e $\{b_k\}$ convergem em $[-\infty, \infty]$. Além disso,

$$\lim_{k\to\infty} a_k = \sup\{a_k \mid k \in \mathbb{N}\} \quad \text{e} \quad \lim_{k\to\infty} b_k = \inf\{b_k \mid k \in \mathbb{N}\}.$$

Disto, podemos elaborar a seguinte proposição, que apresenta caracterizações alternativas à definição de limite superior e inferior.

Proposição B.1. *Seja $\{x_k\} \subseteq \mathbb{R}$ sequência. Então:*

(i) $\limsup(x_k) = \lim_{k\to\infty} b_k = \inf\{b_k\} = \inf_{k\geq 0}(\sup(S_k))$.

(ii) $\liminf(x_k) = \lim_{k\to\infty} a_k = \sup\{a_k\} = \sup_{k\geq 0}(\inf(S_k))$.

Exemplo B.2. Vejamos limites superiores e inferiores das sequências $\{x_k\}$, $k \in \mathbb{N}$, seguintes:

- $x_k = (-1)^k$. Este é um exemplo clássico de sequência oscilatória que contém 1 e -1 como pontos de acumulação. Logo, da definição, segue que $\limsup(x_k) = 1$ e $\liminf(x_k) = -1$. Se olharmos pela definição alternativa na proposição acima, veja:

$$S_k = \{x_{k+1}, x_{k+2}, x_{k+3}, \dots\} = \begin{cases} \{-1, 1, -1, 1, \dots\}, & \text{para } k \text{ par} \\ \{1, -1, 1, -1, \dots\}, & \text{para } k \text{ ímpar} \end{cases},$$

  de modo que, claramente, $\sup(S_k) = 1$, para todo k natural. Logo, $\limsup(x_k) = \inf_{k\geq 0}(\sup(S_k)) = \inf_{k\geq 0}(1) = 1$. De forma análoga, podemos analisar o limite inferior.

- $\{x_k\} = \{0, 0, 1, 0, 1, 2, 0, 1, 2, 3, 0, 1, 2, 3, 4, \dots\}$. Veja que neste caso o conjunto dos pontos de acumulação corresponde a $\mathbb{N} \cup \{\infty\}$. Portanto, temos que $\limsup(x_k) = \infty$ e $\liminf(x_k) = 0$, que são o maior e o menor dos pontos de acumulação.

- $x_k = e^{-k}$. O comportamento de e^{-x}, $x \in \mathbb{R}$, evidencia o que buscamos encontrar neste exemplo: ambos limites iguais a zero. Talvez a forma mais rápida de verificar seja pela definição alternativa:

$$\limsup(x_k) = \lim_{k\geq 0}(\sup(S_k)) = \lim_{k\geq 0}(\sup(\{e^{-(k+1)}, e^{-(k+2)}, \dots\})) = \lim_{k\geq 0}(e^{-(k+1)}) = 0,$$

  em que $\sup(S_k) = e^{-(k+1)}$ pois todo real maior ou igual a $e^{-(k+1)}$ é cota superior de S_k e $e^{-(k+1)}$ é a menor delas. Novamente, verificar que $\liminf(x_k) = 0$ ocorre similarmente. Veja que $\limsup(x_k) = 0$ porém $x_k > 0$, para todo k. Isto mostra que, assim como ínfimo e supremo, os limites inferiores e superiores não precisam, necessariamente, pertencer à sequência ou preservar propriedades que a mesma possui.

Finalizamos com as duas proposições seguintes, que trazem propriedades dos limites superior e inferior.

Proposição B.2. *Seja $\{x_k\} \subseteq \mathbb{R}$ sequência.*

(i) $\liminf(x_k)$ *e* $\limsup(x_k)$ *são pontos de acumulação de* $\{x_k\}$.

(ii) $\inf(S) \leq \liminf(x_k) \leq \limsup(x_k) \leq \sup(S)$, *em que* $S = \{x_1, x_2, \dots\}$.

(iii) $\liminf(x_k) = \limsup(x_k) = x$ *se, e somente se,* $x_k \longrightarrow x$. *Ou seja, os limites coincidem se, e somente se, a sequência é convergente.*

Proposição B.3. *Sejam $\{x_k\}$ e $\{y_k\}$ sequências em $\mathbb{R}$ tais que $x_k \leq y_k$, $\forall k \in \mathbb{N}$. Então:*

(i) $\liminf(x_k) \leq \liminf(y_k)$.

(ii) $\limsup(x_k) \leq \limsup(y_k)$.