CONVERGÊNCIA, ESTIMATIVAS DE ERRO E ESTRATÉGIAS DE SOLUÇÃO EM ALGORITMOS ITERATIVOS

Ézio da Rocha Araújo

TESE SUBMETIDA AO CORPO DOCENTE DA COORDENAÇÃO DOS PROGRAMAS DE PÓS-GRADUAÇÃO EM ENGENHARIA DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE DOUTOR EM CIÊNCIAS EM ENGENHARIA CIVIL

Aprovada por:

Prof. Webe João Mansur, Ph.D.
(Presidente)

Prof. José Claudio de Faria Telles, Ph.D.

Prof. Abimael Fernando Dourado Loula, D.Sc.

Prof. Raul Rosas e Silva, Ph.D.

Prof. Alvaro Luiz Gayoso de Azeredo Coutinho, D.Sc.

Prof. Helio José Corrêa Barbosa, D.Sc.

RIO DE JANEIRO, RJ - BRASIL
MARÇO DE 1999

ARAÚJO, ÉZIO DA ROCHA

Convergência, Estimativas de Erro e Estratégias de Solução em Algoritmos Iterativos [Rio de Janeiro] 1999.

(xi), 179 p. 29,7 cm (COPPE/UFRJ, D.Sc., Engenharia Civil, 1999)

Tese - Universidade Federal do Rio de Janeiro, COPPE

1. algoritmos iterativos

2. estimativas de erro

3. estratégias de solução

I. COPPE-UFRJ II. Título(série)

Resumo da tese apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Doutor em Ciências (D. Sc.)

CONVERGÊNCIA, ESTIMATIVAS DE ERRO E ESTRATÉGIAS DE SOLUÇÃO EM ALGORITMOS ITERATIVOS

Ézio da Rocha Araújo

Março/1999

Orientador: Webe João Mansur

Programa: Engenharia Civil

Esta tese consiste de duas partes.

Na Parte I, a qualidade de convergência de algoritmos iterativos é estudada sob a hipótese fundamental de convergência do quociente da medida do erro. A teoria métrica desenvolvida expõe a geometria e os mecanismos de convergência de algoritmos iterativos, e extrai suas melhores estimativas de erro de aproximação com todas as constantes determinadas durante a sua realização. O objeto de estudo são os algoritmos em espaços de dimensões finitas, porém os resultados são estendidos para espaços métricos completos. Algumas medidas mais fracas de qualidade de convergência são sistematicamente reduzidas à convergência do quociente, permitindo que as estimativas sejam estendidas àquelas medidas.

Alguns mecanismos de convergência evidenciados pela Parte I são consistentemente explorados na Parte II com o foco em sistemas lineares. São propostas estratégias para a solução iterativa de sistemas lineares não Hermitianos como uma particularização de estratégias para sistemas não lineares que objetivam a convergência q-superlinear local, a global e a assintótica, nas medidas de maior interesse.

A teoria e as estratégias são expostas a experimentos numéricos críticos em sistemas Hermitianos e não Hermitianos, com resultados bastante satisfatórios.

Abstract of thesis presented to COPPE/UFRJ as partial fulfillment of the requirements for the degree of Doctor of Science (D. Sc.).

CONVERGENCE, ERROR ESTIMATES AND SOLUTION STRATEGIES OF ITERATIVE ALGORITHMS

Ézio da Rocha Araújo

March/1999

Thesis Supervisor :Webe João Mansur

Department: Civil Engineering

This thesis consists of two parts.

In Part I, the quality of convergence of algorithms is studied under the very fundamental hypothesis of the quocient convergence of the error measure. The metric theory developed exposes the geometry and the convergence mechanisms of iterative algorithms, and extract from them the best estimators for the approximation error, where all constants are computed while the algorithm is running. The main object of study are the algorithms in finite dimensional vector spaces. Nevertheless, the results are fully extended to more general complete metric spaces. Some more weak measures of the quality of convergence are systematically reduced to the quociente convergence measures allowing to extend the results to those weak measures.

Several convergence mechanisms that have emerged from Part I are used in Part II to formulate some solution strategies for Hermitian and non Hermitian algebraic linear systems of equations. The strategies are focused in the local, global and asymptotically q-superlinear convergence of the algorithms on the measure of greater interest.

The theory developed in this thesis is exposed to critical numerical experiments, and the results are very satisfying.

À GRAÇA REIS,

À PENÉLOPE, DIEGO E PÉROLA.

Agradecimentos

Desejo expressar meus mais profundos agradecimentos ao Prof. Webe João Mansur pelo seu entusiasmo, encorajamento e aptidão na orientação desta tese. A sua sensibilidade em escolher o alvo preciso, e iluminar a trajetória de todos os seus orientados com seu vasto conhecimento, acompanhado de um profundo senso de humanidade, permanece em mim, após quatro anos de convivência quase diária, como uma das experiências mais memoráveis de toda a minha vida. Foi um grande previlégio estudar com ele.

À COPPE, ao Departamento de Engenharia Civil, à todos os seus professores e funcionários, particularmente àqueles da área de Estruturas, desejo expressar também o meu profundo respeito e admiração pela qualidade do trabalho desenvolvido ao longo de todos esses anos. Os meus agradecimentos.

O LAMEC foi um ambiente verdadeiramente estimulante. Obrigado ao Prof. Telles, Prof. José Paulo, Prof. Carrer e ao Prof. Santiago. Agradecimentos especiais ao Dr. Alkimin, trabalhar com ele no projeto de Serra da Mesa foi uma lição de produtividade.

Agradecimentos supercomputacionais e iterativos ao Prof. Álvaro Coutinho. Os vários cursos, e esclarecimentos posteriores sobre as sutilezas destas disciplinas, carrregados de entusiasmo, me contaminaram.

Ao Helinho e à Margareth pela amizade, pelas facilidades bibliotecárias no LNCC, pela compreensão dos meus esquecimentos contratuais com o apartamento da Rua da Matriz. Obrigado.

Muito, muito obrigado aos meus amigos Paulinho, Assis, Afonso, Romilde, Luiz Paulo. Os jantares no Plebeu são inesquecíveis.

À CAPES e À UFPE pelo apoio financeiro.

Índice

1 Introdução **1**

1.1 Do que esta tese trata 1

1.2 Motivação 3

1.3 Objetivos 7

I Qualidade de Convergência e Estimativas de Erros **12**

2 Preliminares **15**

2.1 Um Guia para o *Background* 15

2.2 Algoritmos e Convergência 17

2.2.1 Notações e definições básicas 17

2.2.2 Convergência de algoritmos 19

2.2.3 Convergência de sequências 21

2.2.4 Velocidade de convergência 25

2.3 Teoria Métrica dos Pontos Fixos 29

2.3.1 Definições 29

2.3.2 Resultados básicos sobre atrações 30

2.3.3 Contrações iteradas 35

2.4 Comentários finais 36

3 Convergência do Quociente **38**

3.1 Introdução 38

3.2 Expandindo Definições 39

3.3 Convergência q-linear 42

3.3.1 O Lema fundamental 42

3.3.2 A geometria do processo q-linear 43

3.3.3 Geometria e estimativas da realização 55

3.3.4 q-linearidade e escalabilidade 62
3.3.5 Estimativas em espaços de Hilbert 63
3.3.6 A condição de descendência 63
3.3.7 Projeções ortogonais unidimensionais 68
3.3.8 Projeções ortogonais multidimensionais 75
3.4 Convergência q-superlinear 77
3.4.1 Estimativas da realização 77
3.4.2 Estimativas em espaços de Hilbert 81
3.5 Convergência q-sublinear 84
3.5.1 Estimativas da realização 85
3.5.2 Outros resultados 86
3.6 Aproximações do Operador 87
3.7 Comentários Finais 88
3.7.1 Os mecanismos da convergência 88
3.7.2 A qualidade dos estimadores 90
3.7.3 Trabalhos relacionados 91

II Estratégias para Sistemas Lineares 92

4 Estratégias Correntes 95
4.1 Introdução 95
4.2 A Essência das Estratégias Correntes 96
4.3 Algoritmos para Geração da Base 97
4.3.1 Sistemas Hermitianos 97
4.3.2 Não Hermitianos (I) 98
4.3.3 Não Hermitianos (II) 99
4.3.4 O Teorema de Faber e Manteuffel 100
4.4 Precondicionamento 103
4.5 Motivação Para Uma Estratégia 106

5 Estratégia Para a Convergência do Quociente 109
5.1 Introdução 109
5.2 O Esquema Algorítmico 110
5.3 Impondo Consistência 112
5.4 Distâncias e Projeções Generalizadas 113
5.4.1 Funções de Bregman 114

5.4.2 Projeções Generalizadas 116
5.5 Enriquecimento como Distância Generalizada 118
5.5.1 Uma nota sobre a convergência assintótica 121
5.6 Escolha do Precondicionador 122
5.6.1 O método de Broyden e variantes 122
5.6.2 Sequências $\{H_i\}_{i\geq 0}$ de alguns algoritmos 126
5.7 Recursividade e Simultaneidade 127
5.7.1 Atualização de precondicionadores 128
5.7.2 Hibridizações 129
5.7.3 Regularização 130

6 Experimentos Numéricos **132**
6.1 Introdução 132
6.2 Experimento 1 - Pórtico Plano, CG, GB+CG 132
6.3 Experimento 2 - Matrizes Não Normais 152
6.4 Experimento 3 - O Método de Newton 155
6.5 Experimento 4 - GMRES, CGS e GB 160

7 Conclusões e Futuros Trabalhos **169**
7.1 Conclusões 169
7.2 Futuros Trabalhos 171

Lista de Figuras

6.1 Pórtico plano do Experimento 1 - 4/12 elementos 133
6.2 Espectro de autovalores para $n = 9$ 134
6.3 Espectro de autovalores para $n = 33$ 134
6.4 Autovalores para $n = 9$. $\lambda_{\max} = 3.034 \times 10^5$, $\lambda_{\min} = 2.954 \times 10^3$, $\kappa_2 = 102.7$ 136
6.5 Autovalores para $n = 33$. $\lambda_{\max} = 1.212 \times 10^6$, $\lambda_{\min} = 0.975 \times 10^3$, $\kappa_2 = 1242.3$ 136
6.6 Convergência uniforme em norma e direção. Na norma da energia estas curvas são mais suaves. 137
6.7 Estimativas da norma da energia do errro, com um enriquecimento (fator-q exato em cada iteração), para n=9 138
6.8 Estimativas da norma da energia do errro, com um enriquecimentos (fator-q exato em cada iteração), para $n = 9$ 139
6.9 Lema fundamental e estimador para p cnriquecimcnto, $p = 1$ 139
6.10 Estimativa usando apenas o Lema fundamental 140
6.11 Estimativa usando apenas o Lema fundamental 140
6.12 Estimativa com α exato. Norma-ℓ_2 do erro, $n = 9$. 141
6.13 Estimativa com α exato e estimador com $p = 1$. Norma-ℓ_2 do erro, $n = 9$. 141
6.14 Estimativa com α exato e estimador com $p = 1$. Norma-ℓ_2 do erro, $n = 9$. 142
6.15 Corresponde à figura 6.10, agora na norma-ℓ_2. 142
6.16 Corresponde à figura 6.11, agora na norma-ℓ_2, ... 143
6.17 Estimativa com a sequência $\{\alpha_i\}$ exata, $n = 33$. 144
6.18 Estimativa com a sequência $\{\alpha_i\}$ exata, $n = 33$. 145
6.19 Estimativa com a sequência $\left\{\dot{\alpha}_i\right\}$, $n = 33$. 145
6.20 Estimativa com a sequência $\{\alpha_i\}$ exata, $n = 33$. 146

6.21 Estimativa com a sequência $\{\widetilde{\alpha}_i\}$, $n = 33$. 146
6.22 Estimativa da Norma-ℓ_2 do erro no mapeamento composto, $p = 2$, $n = 9$. 147
6.23 Estimativa da Norma-ℓ_2 do erro no mapeamento composto, $p = 2$, $n = 9$, dividida pelo erro exato. 147
6.24 Estimativa da Norma-ℓ_2 do erro no mapeamento composto, $p = 4$, $n = 33$. 148
6.25 Estimativa da Norma-ℓ_2 do erro no mapeamento composto, $p = 2$, $n = 33$, dividida pelo erro exato. 148
6.26 CG, norma-ℓ_2 , $n = 27$. 150
6.27 GB(15)+CG, norma-ℓ_2 , $n = 27$, sem ortogonalização. 150
6.28 GB(15)+CG, norma-ℓ_2 , $n = 27$, epi$(\frac{1}{2}\|e\|_2^2)$. 151
6.29 GB(25)+CG, norma-ℓ_2 , $n = 27$, epi$(\frac{1}{2}\|e\|_2^2)$. 152
6.30 Espectro de autovalores de K e HK 153
6.31 Normas-ℓ_2 do erro e do resíduo na matriz não normal, ao longo de 4500 iterações, $n = 2$. 156
6.32 Previsão do *forward error* pelo *backward error* 157
6.33 Comparação entre a estimativa desta tese e a *backward analysis* . . . 157
6.34 Índice de confiabilidade das medidas, $conf1$ 158
6.35 Componentes e trajetória da solução em R^2 158
6.36 Método de Newton - erros absolutos: exato, estimado e resíduo. E erros estimados relativos aos exatos. 160
6.37 Problema de Mecânica da Fratura - Geometria e Propriedades. 162
6.38 GMRES(30) - estimativas do erro relativo. Norma-ℓ_2 do erro. Usando a sequência $\{\widetilde{\alpha}_i\}$. 163
6.39 GMRES(30) - estimativas do erro relativo (normalizado). Norma-ℓ_2 do erro. Usando a sequência $\{\widetilde{\alpha}_i\}$. 163
6.40 Estimativas do erro relativo para o CGS de Schonauer. 164

Capítulo 1

Introdução

Welcome, my dear reader!
H. B. Bauschke,[1]

1.1 Do que esta tese trata

A mais profunda motivação para esta tese vem da corrida em direção ao desempenho algorítmico. Trabalhando com computadores, há três atitudes possíveis nesta busca[2]:

1 - usar um processador veloz,

2 - múltiplos processadores, e

3 - algoritmos rápidos.

Em verdade, o desempenho de um computador é hoje muito mais dependente do desempenho da arquitetura da memória que da velocidade do processador. Normalmente, o processador é rápido o suficiente (mais que 200 MHz, tipicamente) para levar a memória principal ao seu limite (não mais que 100 MHz, tipicamente). Com esta razão de velocidade de 2/1, no mínimo, a arquitetura do processador não traz muita diferença. Mais interessa quão veloz é a arquitetura da memória.

Quando um processador não é suficiente, a solução encontrada pelos fabricantes nas duas últimas décadas tem sido aumentar o número de processadores e, evidentemente, o número de canais de comunicação entre os processadores e a(s) memórias(s). Todos que trabalham em busca do desempenho algorítmico estão desde então, de alguma forma, envolvidos com computações paralelas, a despeito do alto

grau de dificuldade de programação introduzida, a bem da relação custo/desempenho da máquina.

Este trabalho concentra-se, principalmente, no ítem 3 acima. Além de características padrões como robustez e estabilidade, os algoritmos precisam, essencialmente, ser rápidos, *per si*. No caso de algoritmos iterativos, rapidez significa poucas iterações, com baixo número de operações por iteração. As operações em cada iteração devem estar ordenadas de forma a permitirem a otimização do uso da arquitetura da memória. Tipicamente, o maior esforço computacional dos algoritmos contemporâneos reside nos produtos matriz-vetor n dimensionais. Em sistemas lineares espera-se $O(n^2)$ operações em uma realização. Se este for o caso, implementações ótimas normalmente são possíveis nas diversas arquiteturas computacionais atuais. Algoritmos iterativos rápidos significam tempos de execução menores - na mesma proporção.

Assume-se que o leitor está de alguma forma envolvido com os três aspectos do desempenho computacional acima descrito, e fortemente interessado em algoritmos algébricos iterativos, especialmente em métodos para sistemas lineares. Dado ao caráter multidisciplinar dos tópicos abordados, mesmo do leitor mais envolvido com métodos iterativos, não se espera o seu pronto domínio. Providencia-se no Capítulo 2 a tradicional cortesia de acrescer ao trabalho um ferramental básico e as devidas referências para esclarecimentos posteriores.

A tese está dividida em duas Partes:

A Parte I aborda a qualidade da convergência de algoritmos iterativos sob os três pontos de vista: o local, o assintótico e o global. A teoria desenvolvida permite obter estimativas confiáveis do erro em qualquer medida de interesse, sob a hipótese fundamental de convergência do quociente da medida do erro. O objeto de estudo são os algoritmos no espaço Euclidiano, porém são feitas algumas extensões a espaços de Hilbert e algumas mais modestas a espaços de Banach.

Na Parte II são propostas estratégias algorítmicas que procuram impor a convergência do quociente de medidas de interesse, quer local, assintótica ou global, de algoritmos iterativos para a solução de sistemas lineares não Hermitianos, eventualmente Hermitianos, a partir dos elementos desenvolvidos na Parte I. A filosofia das estratégias reside na hipótese fundamental de que os algoritmos precisam aprender o máximo sobre o problema a partir das iterações. Para tornar isto viável, elimina-se a dicotomia entre precondicionador e método iterativo via polialgoritmos convenientes.

1.2 Motivação

Quando as dimensões dos problemas crescem, surge a necessidade de garantir que o algoritmo escale. Esta escalabilidade diz respeito à velocidade e à demanda de memória. Isto significa que o número de iterações deve manter-se constante com o crescimento do problema - independência da malha - e a quantidade de vetores auxiliares não pode crescer, senão marginalmente. A escalabilidade é uma propriedade que pode ser discutida independentemente do sistema computacional. Entretanto, é extremamente dependente do problema.

Este último ponto não é, em geral, enfatizado na literatura. Para esclarecer melhor esta questão sejam dois problemas de teoria da elasticidade linear, discretizados por alguma técnica numérica, e com o sistema resolvido iterativamente por métodos adequados desenvolvidos em subespaços de Krylov. O primeiro com tensões uniformes em todo o domínio, e o segundo com singularidades na solução, ambos discretizados com a mesma malha pouco densa. O número de iterações para a solução do sistema linear requer, invariavelmente, bem mais iterações para o problema singular. O aumento da densidade da malha nos dois problemas reflete-se no número de iterações necessário para a solução de forma bem distinta. No problema de tensões uniformes a solução será a mesma da malha anterior, e o número de iterações permanecerá constante porque as bases geradas pelo método iterativo são suficientes para descrever a resposta, uma vez que elas são praticamente as mesmas geradas na primeira malha. No problema singular, a primeira malha é insuficiente para descrever todo o conteúdo da resposta exata, e, se a segunda malha for bem projetada, a solução aproximada gerada pela técnica de discretização será substancialmente diferente. A nova malha introduz novas componentes de resposta não capturadas pelas bases geradas pelo método iterativo, nem na malha anterior, nem no mesmo número de iterações da segunda malha. Mais iterações serão necessárias e, neste sentido, o método iterativo não escala. Porém, enriquecendo a discretização o suficiente para obter-se uma boa solução aproximada o método iterativo escala a partir daí. Em outras palavras, o método escala quando o operador discreto converge. Este comportamento é típico do GMRES [3] na solução de certas equações integrais [4]. Neste sentido, o GMRES é independente da malha. O desafio nos dias de hoje é gerar métodos que sejam independentes da malha e ao mesmo tempo independentes do grau de aproximação da solução, ou seja, que mantenham a sua complexidade computacional total escalável com a dimensão do problema.

O primeiro passo nesta direção consiste em conseguir estabelecer um grau de

aproximação para a solução iterativa compatível com a aproximação do modelo discretizado. Isto apenas pode ser feito mediante estimativas de erro confiáveis para ambos os algoritmos envolvidos: o de solução do sistema discretizado e o da estratégia de discretização (adaptação). Problemas semelhantes existem em algoritmos com *inner-outer loops*, lineares ou não lineares, cuja eficiência seja dependente - e sempre o é - da *forcing sequence* [5].

A corrida em direção ao desempenho não é um *racing car* onde quem chegar primeiro ganha. Na verdade quem chegar primeiro continua no seu esforço de atingir precisões maiores, resolver sistemas maiores, ou resolver mais problemas, em um mesmo tempo. Isto pode significar a diferença entre a possibilidade e a impossibilidade de obter uma resposta. Pode-se pensar em um astrofísico modelando galáxias para determinar a velocidade de expansão do universo com algoritmos *multipoles*, e que repentinamente vê-se na necessidade de discretizar mais galáxias ou sistemas planetários [6] (talvez sonhando com a discretização do Universo); em um cientista de Mecânica dos Sólidos tentando modelar os efeitos de escavações quilométricas em rocha basáltica [7], com explosivos, na estabilidade de uma falha tectônica de alguns metros de comprimento; ou em um engenheiro estrutural tentando fazer uma análise completamente não linear de um edifício alto, para em tempo hábil projetar estruturas mais leves (mais baratas) e seguras. Em todos estes casos, mais desempenho significa maior capacidade de modelagem em um mesmo tempo, e portanto melhores e mais confiáveis modelos. Para quantificar estas questões, considere-se que as dimensões dos problemas tratáveis com o método de Gauss cresceram em 45 anos em 10^3 vezes - em 1950 resolvia-se 20 equações, em 1995, 20000 equações. O método de Gauss possui complexidade de $O(n^3)$, o gargalo dos métodos diretos. Este crescimento deve-se essencialmente ao progresso atingido pelo *hardware* no mesmo período, não pelo *software*. Se for possível resolver tais problemas com complexidade $O(n^2)$, então algumas matrizes que são tratadas hoje podem ser de 10 a 100 vezes maior, ou, pode-se tratar o mesmo problema em, por exemplo, 1 hora, em vez de 5 dias. De fato é necessário pensar em $O(n)$, ou talvez $O(n \log n)$, mas parece ainda muito cedo para isto [8].

De fato, os métodos diretos como a decomposição de Gauss, podem vir a ser bastante instáveis e imprecisos, que mentes brilhantes fizeram de seu estudo sua própria vida [9][10]: pivotear ou não pivotear, *backward-error analysis*, etc. Do ponto de vista matemático, a solução de sistemas algébricos, lineares e simultâneos é um problema resolvido. Computacionalmente, é um problema extremamente difícil: tendo

n^2 números efetua-se pelo menos n operações sobre cada um deles para obter apenas n números. Pense em $n = 10^6$, e ter-se-á então pelo menos $n^3 = 10^{18}$ operações. Se agora executa-se em uma cpu com 200 MHz, a quatro instruções de ponto flutuante por ciclo, e um fluxo de dados contínuo entre memória e registradores internos (o que não se tem ainda hoje), o problema será resolvido em 634 anos. Usando um computador com 10000 cpus esperaremos por 23 dias (um computador deste porte quase existe, o ASCI Blue Pacific, com 10000 cpus a 200 MHz). Entretanto, ninguém ousaria resolver tais sistemas com um ASCI usando o método de Gauss, mesmo porque, certamente, as palavras correntes de 64 bits não forneceriam uma resposta com qualquer significado. Agora suponha que o problema é não linear e tem-se que resolver uma cascata de sistemas lineares como esse. Uma única possibilidade existirá se forem usados métodos iterativos.

Esta tese trata de algoritmos iterativos, o foco é a sua velocidade e a contenção da demanda de memória. Mostra, *en passant*, que velocidade também pode ser sinônimo de precisão, robustez e estabilidade. Gauss, parece, já sabia disso[11, pag. 158].

Serão definidas algumas medidas de qualidade de convergência, com as quais os erros de aproximação possam ser estimados *a posteriori*, durante as iterações, para avaliar a qualidade da resposta, na medida desejada, usando ou não informações *a priori* estabelecidas. Surpreendentemente, pouco ou quase nada se tem feito neste sentido [12][13][14]. Poucos trabalhos podem ser citados. Contudo, é um tópico essencial. De fato, o que hoje pode ser dito acerca de, por exemplo, como parar o método dos gradientes conjugados em sistemas lineares, o método de Newton em sistemas não lineares, os métodos quase-Newton em otimização, os métodos de projeção em viabilidade convexa, etc. ? A questão é: como considerar boa uma solução se não é conhecida a distância, na medida desejada, que a separa da solução exata? Pode-se normalmente apenas quantificar o grau de satisfação das equações discretas, verificando a norma Euclidiana dos resíduos. Nos dias correntes, e talvez sempre, a única solução plausível para este tipo de problema parece estar nas estimativas de erros.

Serão estabelecidos alguns resultados acerca da qualidade da convergência de métodos iterativos. Acredita-se que, do ponto de vista computacional, ninguém está contente com as técnicas atuais para solução de sistemas lineares. Aqui, o principal interesse está nesses sistemas, porém muitos resultados estendem-se a algoritmos iterativos algébricos quaisquer. Métodos diretos não são bons o suficiente

para fornecer as respostas aproximadas desejadas em muitos problemas em computação científica, e a única solução que se pode dar - pelo tanto que se vislumbra - são iterações rápidas. Velocidade e aceleração são os principais coadjuvantes. Os algoritmos precisam de velocidade e, idealmente, de aceleração.

Mas, qual é a velocidade necessária e como atingi-la? Procura-se extrair construtivismo dos resultados sobre a qualidade da convergência, e aplicá-lo a uma certa classe de algoritmos, buscando essencialmente velocidade (boas taxas ...) e aceleração (...e cada vez melhores), e tentando satisfazer as restrições de contenção do crescimento de memória auxiliar. É tratada a questão da robustez. Aborda-se então a convergência global. Este é um tópico extremamente difícil. Normalmente a robustez opõe-se à velocidade e à economia de memória, da mesma forma que velocidade e economia de memória se contrapõem; mais robustez significa menos escalabilidade. Este trabalho sugere que uma certa classe de polialgoritmos pode minimizar este conflito, e equilibrar mais essas questões.

A maior dificuldade com sistemas lineares está no custo de gerar uma base para um subespaço suficientemente próximo da solução desejada. Este custo traduz-se em tempo de processamento e, no caso de sistemas não Hermitianos - o Teorema de Faber e Manteuffel [15] - também em quantidade de memória auxiliar. Os métodos que hoje governam as iterações geram sequências de subespaços de Krylov. Graças ao custo ou à aritmética de ponto flutuante, tais sequências são capazes de gerar subespaços suficientemente próximos da solução, apenas se o sistema estiver adequadamente normalizado. Precondicionadores distintos são necessários para cada conjugação de método iterativo e de problema (ou classe de). Esta é uma dificuldade, e existem outras. Os melhores precondicionadores não são vetorizáveis ou paralelizáveis, pelo menos não tanto quanto os métodos que precondicionam, nem são, em geral, escaláveis com a dimensão do problema.

Há muito a ser feito pelos subespaços de Krylov. Existe uma inflação de acrônimos em batalha permanente, e cada um precisa de diferentes precondicionadores para diferentes problemas. Precondicionadores são, de alguma forma, essenciais, porém, é possível estabelecer um precondicionador adequado apenas trabalhando sobre as características dos operadores (o contínuo e o discreto), e prosseguindo por tentativa e erro.

Nesta tese, introduz-se a possibilidade de partir com um precondicionador razoável, senão mínimo - digamos diagonal, se nada melhor existir - e modificá-lo em função dos conhecimentos adquiridos ao longo das iterações, de uma forma obje-

tiva, matematicamente clara. Uma solução ao precondicionador heurístico, ou o topológico, e uma fuga dos espaços de Krylov. Isto permite tratar os problemas lineares de uma forma mais fechada, transparente ao usuário (programa). Se melhores informações existem, por exemplo de realizações anteriores (adaptação, múltiplos carregamentos, análise não linear, dinâmica), elas serão incorporadas automaticamente ao precondicionador, resultando em reduções drásticas de esforço computacional. Argumenta-se que o equilíbrio entre tempo e memória pode ser conseguido se for possível a avaliação iterativa da qualidade da resposta, e por conseguinte, do desempenho corrente das iterações.

1.3 Objetivos

Uma das principais preocupações na análise de algoritmos iterativos tem sido o estabelecimento de suas condições de convergência global e a caracterização de sua velocidade de convergência assintótica. O teorema de convergência global de Zangwill [16, cap. 6] estabelece algumas condições que asseguram que uma sequência de soluções aproximadas produzidas por um algoritmo, convirja para uma solução a partir de um ponto inicial qualquer em $\mathbb{R}^n$. Os resultados de Zangwill baseiam-se na existência de um funcional de $\mathbb{R}^n$ em $\mathbb{R}$, continuamente descendente - a função de descendência. Embora pareça natural avaliar um funcional - a função objetiva, por exemplo - em problemas de otimização, por outro lado, encontrar um mapeamento (aplicação) adequado para a solução de problemas gerais de sistemas lineares ou não lineares, pode vir a se constituir em uma difícil tarefa.

Duas dificuldades normalmente surgem: uma propriedade de *separação* [17], pela qual duas funções de descendência distintas, com mesmos zeros, podem diferir em ordens de magnitude de seus valores e dos de suas derivadas; e a *computabilidade* [18], isto é, se uma função adequada - algumas vezes sua derivada - vai ser usada ela deve ser computável. A existência de um tal mapeamento é extremamente importante porque ele fornece, pelo menos, uma métrica para a proximidade da solução (erro), uma taxa de convergência desta métrica (rapidez), e uma quantificação do comportamento assintótico do algoritmo (estabilidade). Portanto, as questões de separação e computabilidade devem ser cuidadosamente tratadas.

Se um algoritmo é contractivo, o Teorema do Mapeamento (aplicação) Contractivo (contractiva) - TMC [19] pode ser usado para computar estimativas confiáveis de limites superiores do erro local da norma contractiva - e portanto q-linearmente

convergente nesta norma. Os métodos lineares básicos convergentes satisfazem, obrigatoriamente, o TMC. Na maioria dos casos, entretanto, a métrica contractiva não é computável ou pelo menos o custo computacional para tal excede em ordens de grandeza a realização do algoritmo. Devido à grande habilidade dos computadores contemporâneos em efetuar produtos matriz-vetor, a maioria dos algoritmos bem sucedidos em uso nos dias de hoje são computacionalmente intensivos neste aspecto e não satisfazem o TMC em nenhuma norma conhecida. Por exemplo, as técnicas de Decomposição de Domínio e Multigrid fazem uso dos métodos básicos como Jacobi e Gauss Seidel, condicionalmente contractivos. Entretanto, geralmente estão associados a aceleradores que se desenvolvem em subespaços de Krylov, como o CG e o GMRES. Tais métodos iterativos de projeção em subespaços de Krylov não satisfazem o TMC para nenhuma norma conhecida. Isto é agravado pela tendência atual de uso de algoritmos potencialmente paralelizáveis, quando muitas vezes mais operações são efetuadas para obter mais paralelismo. De fato, alguns dos métodos de projeção possuem apenas convergência assintótica q-sublinear. Os mais eficientes são q-linearmente convergentes e bem poucos atingem convergência q-superlinear (geralmente prova-se apenas r-superlinearidade), e mesmo assim, a terminação finita toma lugar bem antes desta propriedade se pronunciar. Estagnação numérica temporária é um lugar comum, retardando a r ou q-linearidade. Mesmo porque, embora contração implique em convergência q-linear, o inverso não se aplica. Portanto, estimativas de erro baseadas em condições menos restritivas do que as hipóteses do TMC são necessárias. Adicionalmente, contração não implica, necessariamente, em taxas de convergência adequadas. Este é o caso, por exemplo, do método do maior declive (*steepest descent*).

Geralmente, grandes dificuldades adicionais surgem quando definições teóricas e propriedades de taxas de convergência assintóticas são aplicadas em computações numéricas, onde a aritmética, os espaços e os processos são finitos. Esta coleção de barreiras têm posto matemáticos aplicados e engenheiros em favor de resultados mais heurísticos, quer porque somente em espaços de dimensões infinitas *"where, ..., the mathematical notion of superlinear convergence does make sense "* [20], quer porque elas são *"ill-behaved and ... therefore not reliable indicator of how the iteration is proceeding "* [21], ou, recentemente, considerar que a terminação de um algoritmo baseado em pequenos resíduos relativos - errado, como será mostrado, por causa da convergência q-sublinear e da propriedade de separação - pode ser computacionalmente vantajoso [5].

Em muitos problemas não Hermitianos a fuga da normalidade pode provocar grandes reduções nas taxas de convergência e mesmo divergência, muito embora a análise dos algoritmos mostre convergência global. Mais preocupante em matrizes altamente não normais é que algoritmo pode estagnar distante da solução, quando todas as medidas teóricas e as numéricas em uso corrente indicam convergência: a aritmética finita produz uma substancial alteração do ponto fixo [22], face à fragilidade do método (e das medidas). A respeito dos métodos de decomposição, é bom lembrar que uma matriz triangular é normal se e somente se é diagonal, e assim, resolver sistemas triangulares pode vir a ser numericamente problemático, quer como parte da solução direta, quer como parte de precondicionadores (decomposições incompletas como ILU, por exemplo).

O usuário final de um algoritmo iterativo deve assegurar-se que não apenas ele converge mas que a convergência se dará em um número aceitável de iterações. Para isto são necessárias medidas de convergência e medidas de velocidade de convergência adequadas. Idealmente, um algoritmo iterativo deve ter robustez suficiente para convergir sempre e com uma velocidade de convergência aceitável, na classe de problemas para o qual foi escolhido. A realização deve ser interrompida quando a não convergência ou baixas taxas de convergência forem detectadas ou quando o erro da solução aproximada estiver dentro de um nível aceitável de acurácia. O erro, naturalmente, não é computável. Portanto, medidas de convergência e de velocidade de convergência de algoritmos iterativos para sistemas lineares e não lineares são correntemente baseadas na norma Euclidiana dos resíduos ou dos resíduos precondicionados, enquanto em problemas de otimização a prática corrente é o uso da norma Euclidiana dos gradientes. Alguma heurística normalmente é adicionada a estas medidas. Porém, estas são medidas errôneas de convergência se o sistema é mal condicionado, e no mínimo inadequadas se ele é bem condicionado. Recentemente, foi proposto o uso de técnicas retroativas de erro (*backward error*) para limitar o erro corrente (*forward error*) em algoritmos iterativos. Infelizmente, nesta abordagem a questão da computabilidade apenas foi transformada em um problema de estimativa de normas de operadores, aonde, normalmente a sua representação discreta em espaços de dimensão finita sofre o conhecido problema de separação.

Isto posto, o objetivo da Parte I é:

Estabelecer fundamentos teóricos e ferramentas práticas para mensurar a qualidade da convergência e determinar estimativas confiáveis para erros de aproximação em algoritmos iterativos, enfatizando sistemas de equações não lineares e sistemas

lineares em particular.

Muito embora, a maioria dos resultados sejam aplicáveis a algoritmos algébricos genéricos. A abordagem usa as definições de convergência do quociente do erro e o Teorema do Mapeamento Contractivo, ambos largamente usados para investigações teóricas de convergência de processos infinitos em matemática contínua. Subsequentemente, a restrição de contrações são relaxadas e são apresentadas algumas novas propriedades finitas das sequências envolvidas no processo, as quais serão usadas para explicar alguns aspectos importantes do comportamento de algoritmos iterativos e para obter estimativas de erro sem o uso de análise ou de extração de autovalores, ou de estimativas de normas de operadores. Mostra-se também em quanto os erros inerentes à precisão finita afetam as estimativas de erro sugeridas, e como estimá-los durante as iterações.

A Parte II essencialmente examina uma forma bastante geral de algoritmos iterativos que inclui a normalização ou precondicionamento como parte integrante do método iterativo. A idéia básica é ter como limite métodos para sistemas não lineares. Nesta parte são resgatadas idéias intuitivas desenvolvidas na Parte I que poderiam levar ao aumento de velocidade de convergência, da robustez e da escalabilidade. A principal preocupação são os sistemas lineares não Hermitianos, embora sejam eventualmente tratados os Hermitianos. Examina-se variantes de métodos conhecidos e algumas novas hibridizações.

O Teorema de Faber e Manteuffel [15] essencialmente estabelece as limitações econômicas das sequências de Krylov para matrizes gerais. Ele é o principal obstáculo ao projeto de algoritmos de Krylov com bases econômicas - recorrências curtas. Com esta interpretação do Teorema, resta escapar dos espaços de Krylov. Métodos Multigrid podem ser vistos como uma realização desta idéia. Mas existem outras potencialmente interessantes. Esta tese examina a possibilidade de mudar a métrica do espaço de Krylov em cada iteração (ou grupo de iterações) produzindo sequências de subespaços que não estão contidos em qualquer subespaço de Krylov. Porém, em vez de descartar totalmente os subespaços de Krylov, suas bases são aproveitadas, e as transformações métricas são efetuadas sobre os precondicionadores. Isto produz uma classe de polialgoritmos com inúmeras possibilidades. Na Parte II estas idéias e algumas realizações são examinadas.

O objetivo da Parte II pode ser claramente estabelecido:

Estudar técnicas matematicamente consistentes que permitam diminuir a dicotomia entre precondicionadores e métodos iterativos usando as informações

extraídas das iterações, e bases possivelmente fora dos subespaços de Krylov, com o intuito de aumentar a velocidade de convergência e ao mesmo tempo manter um compromisso com a demanda de memória e a robustez.

De forma a ilustrar e avaliar os diversos aspectos desta tese, são reunidas em um único capítulo várias aplicações em problemas lineares e não lineares de sistemas provenientes de aproximações por quadratura, elementos finitos e elementos de contorno, de equações diferenciais e integrais de alguns problemas de engenharia.

Parte I

Qualidade de Convergência e Estimativas de Erros

The approximation error in iterative schemes is not directly available, and has to be estimated or bounded in terms of information obtained in the iteration process. This can be done for special cases, for instance for the Conjugate Gradient method. For most cases, however, the problem of getting realistic error bounds is yet unsolved. In some applications we need information on errors in specified components of the solution, this is also mainly an open problem.

Gene H. Golub and Henk A. van der Vorst, 1996[13]

Nesta Parte I da tese são estabelecidos alguns fatos, muitos dos quais novos, sobre convergência de algoritmos. Os objetivos podem ser declarados como: primeiro, estabelecer medidas adequadas para interromper algoritmos iterativos em execução, e segundo, quantificar a qualidade de convergência de tais algoritmos objetivando orientar o desenvolvimento de estratégias para a solução de problemas extremamente difíceis como os encontrados frequentemente em Mecânica das Estruturas. Esta última disciplina é entendida como a oriunda da discretização de aproximações, geralmente reduções dimensionais, da Mecânica dos Sólidos, indiferente às técnicas usadas para discretizar ou aproximar[23].

Os principais resultados dizem respeito ao estabelecimento de limites superiores e inferiores para erros de aproximação em algoritmos em $\mathbb{R}^n$ munido de alguma métrica, não necessariamente elíptica. No caso de normas oriundas de produtos escalares, mostra-se que, sob certas condições, os limites são ainda mais estreitos. Acrescenta-se um pouco de generalização, permitindo, quando possível, incluir espaços de Hilbert, ou mesmo de Banach. Isto por três motivos. Primeiro porque parece natural querer estender os resultados a espaços de funções objetivando, por exemplo, estimativas em estratégias adaptativas, para verificar os efeitos que as hipóteses de convergência do quociente acarretam. Não se tem conhecimento de que isto tenha sido feito, sob a hipótese restritiva da convergência do quociente do erro. Em segundo lugar porque acredita-se que, eventualmente, haja interesse de outros que trabalham com algoritmos em espaços de funções.

Em terceiro lugar, talvez o mais motivante, está no fato de que o surgimento de computações paralelas tem mudado radicalmente a abordagem algorítmica em computação científica. No contexto deste trabalho, mais e mais, surgem precondicionadores oriundos de análise sobre as equações diferenciais ou integrais que gover-

nam o problema físico. É de se esperar que, no futuro, em muitas áreas, as iterações tomem lugar diretamente sobre os problemas originais, antes que a decomposição ou discretização tenha sido feita, e assim deverão ser estudadas em espaços de funções. Isto comunga com a Parte II desta tese, onde os precondicionadores são iterandos.

Mais uma explanação. O termo *local* usado nesta Parte I não condiz com o conceito de convergência assintótica ao qual geralmente está associado - quando *local* significaria no entorno da solução, assintoticamente. *Convergência local* é entendida aqui como a convergência durante as iterações, quer instantaneamente, ou em um grupo de iterações. Isto não exclui a convergência assintótica, nem mesmo se trata de convergência na média. Assim, ele é mais abrangente. Na ausência de um termo adequado, o termo convergência local será mantido. Ele está mais de acordo com a percepção de Nevanllina [24, cap 5, pag. 167] para convergência superlinear em espaços (Banach) de funções - convergência superlinear local (*local* no sentido desta tese) implica em convergência assintoticamente superlinear.

Capítulo 2

Preliminares

If Euler were alive today, he wouldn't be proving existence theorems.
L. N. Trefethen,[9]

O principal argumento sobre o qual esta tese repousa é a observação de que resolver sistemas lineares iterativamente se constitue em um problema não linear, e como tal deve ser posto. Portanto, não é surpresa que os atores principais sejam, essencialmente, mapeamentos não lineares, reduzidos por sua vez, a aproximações sucessivas por mapeamentos lineares. Este capítulo objetiva reunir uma série de ferramentas da literatura, e algumas extensões, que serão utilizadas livremente nos capítulos posteriores. São colocadas também algumas questões relevantes que direcionam os demais capítulos.

Primeiramente, segue-se um guia para o material que dá suporte a esta tese.

2.1 Um Guia para o *Background*

A principal referência sobre análise não linear é a obra clássica de Ortega e Rheinboldt [27], que inclui numerosos resultados sobre as atrações básicas. Esta obra é também a referência definitiva sobre convergência-q e convergência-r (*Q-convergence factors e R–convergence factors)* como medidas de convergência *assintótica.* O termo *local* nesta tese será usado de forma não convencional, não assintótica. Será reservado o termo *convergência local* para designar a *convergência durante as iterações.* Também, a construção de algoritmos para problemas não lineares a partir

de extensões de algoritmos típicos para problemas lineares usada naquela obra demonstrou historicamente, do ponto de vista de desempenho, que aquele - do linear para o não linear - não parece ser o caminho correto.

Os elementos de análise em $\mathbb{R}^n$ seguem a orientação de Bartle [25]. Bartle termina ao introduzir as noções de ponto fixo de um mapeamento não linear e com o teorema (topológico) dos pontos fixos de Brower, fundamentalmente um resultado de existência. Da obra de Goebel e Kirk [26], mais recente e específica, foram extraidos vários fatos sobre a teoria métrica dos pontos fixos. Goebel e Kirk a desenvolve em espaços de Banach, neste capítulo os resultados estão colocados no espaço Euclidiano, com as devidas peculiaridades. A este respeito, a citada obra de Ortega e Rheinboldt [27] tornou-se essencial. Contudo, muitos resultados aqui apresentados estendem-se imediatamente a espaços de Banach ou mesmo a espaços métricos mais gerais, e isto será observado quando for de interesse.

As noções de funções de Bregman e de distâncias e projeções generalizadas encontram-se em Censor e Zenios [28], e referências lá citadas. Também muito de minimização proximal (o termo é mantido, como neologismo, na falta de uma tradução fonética mais elegante para *proximal minimization* em espaços de Banach). Fundamental é o extenso e completo artigo de Bauschke e Borwein [29] sobre o problema de viabilidade convexa em espaços de Hilbert (*convex feasibility problem*), e a tese do Bauschke [1], que também inclui operadores de projeção (cujos resultados são portanto trivialmente transpostos para espaços Euclidianos). As teorias contidas naquelas obras ajudaram muito na formalização e refinamento posterior de algumas estratégias algorítmicas que vinham sendo usadas neste trabalho e na explanação de uma porção de sucessos e insucessos (muitos não citados) algorítmicos deste trabalho, e de outros. Estas ferramentas são introduzidas gradativamente a partir do Capítulo 4.

Muitas das noções intuitivas que guiaram este trabalho com certeza tiveram raizes na obra clássica de Dennis e Schnabel [19], em especial sobre velocidade de convergência, convergências assintótica e global, e métodos tipo-Newton, em especial a teoria dos métodos quase-Newton. A idéia lá contida, germinada em [27], de que os algoritmos para problemas não lineares (e, extrapolando, para lineares) bem sucedidos parecem caminhar assintoticamente para o método de Newton, vem, nos dias de hoje, ganhar aspectos práticos desconcertantes com os processadores superescalares, os vetoriais, e o processamento paralelo: ganha-se velocidade nas operações (maior potencial de vetorização ou paralelização) em detrimento da ve-

locidade de convergência (de q-ordem 2 para q-superlinear ou q-linear) para obter um desempenho global superior - que muitas vezes faz a diferença entre o possível e o impossível.

Finalmente, ao longo do texto é feito um uso extensivo de álgebra linear básica, computacional, e métodos iterativos para sistemas lineares bem conhecidos, sobre os quais recomenda-se a leitura de [11], [30], [31] e [32]. Algumas noções sobre hereditariedade em métodos iterativos, e como ela se aplica aos métodos de Krylov, usadas na Parte II desta tese, foram didaticamente expostas pelo autor em [33]. Também lá se encontra algum material da álgebra linear necessária, e um mínimo sobre pontos fixos. Alguns resultados sobre estimativas de erros na linha desta Parte I, com numerosos exemplos e discussões, pode ser encontrado em [34]. As primeiras realizações de algoritmos decorrentes da teoria desenvolvida nesta tese encontram-se também em [35], e os respectivos códigos fontes em FORTRAN para o CRAY-YMP da UFRGS, em [7]. Alguns fatos menos comuns serão relembrados imediatamente antes de serem usados.

2.2 Algoritmos e Convergência

2.2.1 Notações e definições básicas

A notação $x \in \mathbb{R}^n$ (ou $\mathbb{C}^n$) designa um elemento do espaço cartesiano n-dimensional real (ou complexo), normalmente usando uma letra romana minúscula. Sempre, x é um vetor coluna, e o vetor linha correspondente será designado por x^T, o vetor transposto (ou x^H, o hermitiano). É necessário, com frequência, referir-se a um elemento x_i de uma sequência $\{x_i\}$ de elementos de $\mathbb{R}^n$ (ou $\mathbb{C}^n$). Quando necessário, será usada a notação $(x_i)_j$ para designar a j-ésima componente do vetor x_i. Sempre que possível, um elemento de $\mathbb{R}$ (ou $\mathbb{C}$) será designado por uma letra grega minúscula. Os números inteiros e os naturais são designados por $\mathbb{I}$ e $\mathbb{N}$, respectivamente. A maioria dos resultados e algoritmos neste trabalho valem em $\mathbb{C}^n$, embora, por conveniência, a referência será feita apenas a $\mathbb{R}^n$.

O produto escalar, a norma e a distância (métrica) em um espaço vetorial X serão representados por $(\cdot,\cdot)$, $\|\cdot\|$, $d(\cdot,\cdot)$, respectivamente.

Este trabalho é desenvolvido quase sempre no estágio dos espaços Euclidianos, ou seja, em um espaço de Hilbert de dimensão finita n, e, portanto, a norma está definida por um produto escalar, e a métrica por esta norma. A caracterização de um espaço de Hilbert (sua geometria) é feita pela lei do paralelogramo, aplicável a

qualquer par u e v de seus elementos:

$$\|u-v\|^2+\|u+v\|^2=2\left(\|u\|^2+\|v\|^2\right), \tag{2.1}$$

ou, equivalentemente, em termos de produto interno

$$\|u+v\|^2=\|u\|^2+\|v\|^2+2(u,v). \tag{2.2}$$

Convém estabelecer alguma notação adicional em dimensão finita. Seja $L(\mathbb{C}^n)$ o conjunto (espaço vetorial) de todos os operadores lineares A de $\mathbb{C}^n$ em $\mathbb{C}^n$. Para qualquer $A\in L(\mathbb{C}^n)$, $\overline{A}$ representa a matriz formada pelos complexos conjugados dos elementos de A, a matriz $A^H=\overline{A}^T$ é a Hermitiana transposta ($A^H_{ij}=\overline{A}_{ji}$). Se A é regular e $u,v\in\mathbb{C}^n$ define-se $\|v\|_{A^HA}=\|Av\|_2$, onde $\|\cdot\|_2$ é a norma Euclidiana ou norma-ℓ_2. A norma $\|\cdot\|_{A^HA}$ será chamada de norma-A^HA. Ela vem definida pelo produto interno

$$(u,v)_{A^HA}=\left(u,A^HAv\right)_2=(Au,Av)_2=\left(A^HAu,v\right)_2=v^HA^HAu\ .$$

Dois vetores u,v são ortogonais se $(u,v)_2=v^Hu=0$, e ortonormais se, adicionalmente, $\|u\|_2=\|v\|_2=1$. Os vetores u e v são ditos A^HA-ortogonais se $(u,v)_{A^HA}=0$. Se $A^HA=AA^H=I$, a matriz identidade de ordem n, então A é unitária e $(u,v)_{A^HA}=(u,v)_2$.

A projeção-A^HA de u na direção de v é

$$\begin{aligned}\underset{v}{\text{proj-}A^HA}\ (u) &= \frac{(u,v)_{A^HA}}{(v,v)_{A^HA}}v=\frac{\left(u,A^HAv\right)_2}{(v,A^HAv)_2}v=\frac{v^HA^HAu}{v^HA^HAv}v\\ &= \frac{vv^HA^HA}{v^HA^HAv}u=P_{A^HA}u.\end{aligned}$$

onde $P_{A^HA}\in L(\mathbb{C}^n)$ é o operador de projeção-A^HA sobre v, ou de projeção A^HA-ortogonal na direção de v. Para minimizar a norma-A^HA na direção de v, subtrai-se de u a projeção-A^HA de u em v

$$u_{\perp A^HA}=u-P_{A^HA}u=(I-P_{A^HA})u.$$

Então, para todos os vetores da forma $u+\tau v$, $\tau\in\mathbb{C}$, $u_{\perp A^HA}$ possui a menor norma-A^HA.

Se $\|\cdot\|_\bullet$ é uma norma associada a um produto interno $(\cdot,\cdot)_\bullet$, então existe uma (única) matriz regular $A\in\mathbb{C}^n$ tal que

$$(\cdot,\cdot)_\bullet=(\cdot,\cdot)_{A^HA}\ \text{ e }\ \|\cdot\|_\bullet=\|\cdot\|_{A^HA}\ .$$

Diz-se que uma matriz A^* é a adjunta de A com respeito a um produto interno $(\cdot,\cdot)_\bullet$ se $(Au,v)_\bullet = (u,A^*v)_\bullet$ para todos os $u,v \in \mathbb{C}^n$. Se o produto interno é o Euclidiano então $(Au,v)_2 = \left(u,A^Hv\right)_2$. Uma matriz é dita autoadjunta ou Hermitiana se ela é igual à sua adjunta.

Em algumas ocasiões, serão usados espaços mais gerais, e nestes casos serão feitas as devidas observações quanto à possível generalização dos resultados. Esta necessidade ocorre em três ocasiões. A primeira quando da estimativa de erros nas normas ℓ_1 e ℓ_∞ de soluções aproximadas de algoritmos iterativos. Nesse caso, cabe lembrar que não se está mais sobre espaços de Hilbert, porém ainda em espaços (normados e completos) de Banach de dimensão finita. Na segunda, quando forem usados estimadores para métodos iterativos em estratégias adaptativas e, neste caso, os espaços não são possuem dimensões finitas, porém estaremos ainda em espaços de Hilbert, pelo menos enquanto as estimativas são da norma da energia, e em espaços de Banach quando as normas não decorram de produtos internos. Finalmente, na interpretação de algumas estratégias de convergência global e regularização de algoritmos iterativos serão usadas medidas generalizadas [28]. Neste caso, ocorre que algumas das medidas nem mesmo se constituem em uma métrica - por exemplo, distâncias epigráficas - no sentido que, eventualmente, não satisfazem $d(x,y) = d(y,x)$, para qualquer $x,y \in \mathbb{R}^n$,e não se está, em princípio, também no estágio de espaços métricos.

Esta técnica, de trabalhar em espaços Euclidianos e tratar as exceções quando necessário, é mais conveniente para este trabalho, primordialmente dirigido a dimensões finitas, tendo já sido adotada em [27] e será a seguida neste texto. Isto evita problemas como caracterizaçõcs de convergência fraca e forte, e sequências de Cauchy e convergência, todas equivalentes em espaços de dimensão finita.

2.2.2 Convergência de algoritmos

Cada algoritmo iterativo da classe aqui tratada é um mapeamento não linear $\mathcal{A} : \mathbb{R}^n \to \mathbb{R}^n$, tal que uma iteração do algoritmo $\mathcal{A}$ é representada por $y = \mathcal{A}(x)$, onde x e y são elementos $\mathbb{R}^n$, x a solução corrente e y uma solução que, espera-se, seja, em algum sentido, melhor do que x. O símbolo $\mathcal{A}(x)$ não significa que o algoritmo é função apenas do ponto x, porém que o algoritmo é aplicado no ponto x. De fato, o algoritmo pode ser função de pontos anteriorcs da sequência de soluções aproximadas que ele gera, cntre outras coisas.

O principal problema a ser tratado pode ser colocado sob a forma

$$encontre\ um\ x\ tal\ que\ f(x) = 0, \tag{2.3}$$

onde x e $f(x)$ são elementos de $\mathbb{R}^n$ satisfazendo, eventualmente, certas restrições. O maior interesse é o mais simples problema linear sobre o qual não há qualquer restrição adicional e a solução existe, e é única. Porém, sempre imagina-se aqueles problemas mal condicionados em algum sentido, nos quais ou são necessárias demasiadas iterações para convergir, ou equivalentemente, requer uma memória auxiliar excessiva. Tais problemas são típicos em Mecânica das Estruturas, e provocados, muitas vezes, pela redução dimensional do problema original (do tridimensional para placas, por exemplo) , ou por forte acoplamento entre dois campos (por exemplo, membrana e flexão), ou por fortes singularidades na solução (por exemplo, cantos reentrantes em Mecânica dos Sólidos). Estes tipos de problemas requerem alta dimensionalidade da base para descrição da resposta quando, por exemplo, são tratados com subespaços de Krylov.

Geralmente, serão usados algoritmos da classe descrita para encontrar melhores soluções para o problema (2.3), sequencialmente (o que não exclui computações paralelas), a partir de uma solução inicial x_0 por iterações do tipo $x_{i+1} = \mathcal{A}(x_i)$, $i = 0, 1, 2, \ldots$, onde i é o índice do algoritmo ou da sequência.

Se o problema (2.3) possue uma solução $x_\star$, espera-se que o algoritmo atinja $x_\star$ tal que, assintoticamente, $x_\star = \mathcal{A}(x_\star)$ - de fato espera-se apenas que esta igualdade seja satisfeita de forma aproximada para um i não muito grande, tipicamente i é muito menor do que n. Neste caso diremos que o *algoritmo convergiu* para uma solução do problema. Ocorre que, eventualmente, existem outros pontos x para os quais $x = \mathcal{A}(x)$. Ao conjunto deste pontos, os pontos fixos do mapeamento $\mathcal{A}$, designa-se por $Fix\,(\mathcal{A})$. Um algoritmo é consistente com o problema ou, simplesmente, *consistente*, se $x_\star \in Fix\,(\mathcal{A})$, onde $x_\star$ é também um elemento do conjunto de soluções de (2.3). É possível que não haja intersecção entre o conjunto de soluções do problema e $Fix\,(\mathcal{A})$. Neste caso, se $Fix\,(\mathcal{A})$ não é vazio, eventualmente o algoritmo atingirá um ponto fixo e diremos que o algoritmo *estagnou*, provavelmente em um ponto não satisfatório. Este tipo de ocorrência pode se dar por esgotamento da capacidade do algoritmo (*breakdown*) de gerar novos elementos da base ou por problemas de precisão finita (truncamento, cancelamento), na suposição de que ele é consistente, e matematicamente adequado ao problema. Se o algoritmo não possue ponto fixo ou por algum motivo não atinge um ponto fixo, então, dir-se-á que ele *divergiu*.

A questão de proximidade é quantificada, possivelmente, com uma distância

(métrica) $d(x_\star, x_i)$ entre a iterada corrente e a solução desejada, porém não necessariamente a induzida pela norma do espaço Euclidiano. Por exemplo, o método dos gradientes conjugados desenvolve suas bases em um espaço de Hilbert com uma norma de energia (uma norma Euclidiana ponderada), porém pode-se estar interessado em medir a distância da solução com alguma norma-$\ell_p, p = 1, 2, \ldots, \infty$. Neste sentido, a convergência, a estagnação e a divergência, e algumas vezes a própria consistência, estão relacionados com valores satisfatórios ou insatisfatórios da distância escolhida. Se o espaço é normado então $d(x_\star, x_i) = \|x_i - x_\star\|$.

O erro da solução aproximada é dada pelo vetor $e_i = x_i - x_\star$. O incremento $s_i = x_{i+1} - x_i$ entre dois elementos consecutivos da sequência $\{x_i\}_{i\geq 0}$ gerada pelo algoritmo, é o mesmo incremento do erro, $s_i = e_{i+1} - e_i$. Portanto, o incremento s_i, e os dois erros consecutivos e_i e e_{i+1} estão todos contidos no hiperplano que contém os pontos x_i , x_{i+1} e a solução $x_\star$ - por exemplo, pertencem ao subespaço afim, não computável, $x_\star + span(e_i, e_{i+1}) = x_\star + span(s_i, e_i) = \cdots$. Esta imagem geométrica ajuda bastante no entendimento das relações entre esses vetores.

2.2.3 Convergência de sequências

Geralmente se estuda um algoritmo através da observação da sequência $\{x_i\}_{i\geq 0}$ de soluções aproximadas que ele gera. A esse respeito, considere-se a seguinte definição de convergência de uma sequência em $\mathbb{R}^n$ (ou um espaço métrico completo).

Definição 1 *Diz-se que uma sequência $\{x_i\}_{i\geq 0}$, $x_i \in \mathbb{R}^n$, converge para o ponto x se, e somente se, dado um conjunto qualquer de números reais $\varepsilon_j > 0$, $j = 1, \ldots, n$, existe um $i_0 \geq 0$ tal que cada j-ésima componente de $x_i - x$, satisfaz*

$$\left|(x_i)_j - (x)_j\right| < \varepsilon_j, \text{ para todo } i \geq i_0. \tag{2.4}$$

O limite da sequência $\{x_i\}_{i\geq 0}$ será representado por $\lim x_i = x$. Se D é um subconjunto não vazio de $\mathbb{R}^n$, então a sequência de pontos em D será convergente em D se, e somente se, existe um ponto $x \in D$ para o qual ela converge. Uma sequência que é convergente em $\mathbb{R}^n$ será chamada simplesmente de convergente.

A definição acima exige a convergência de cada componente, como esperado pela intuição. O seguinte resultado mostra que a convergência na norma em $\mathbb{R}^n$ fornece uma medida equivalente de convergência.

Teorema 2 *Seja* $\|\cdot\|$ *uma norma qualquer em* $\mathbb{R}^n$, *e* $\{x_i\}_{i\geq 0}$ *uma sequência em* $\mathbb{R}^n$. *Então esta sequência converge para* x *em* $\mathbb{R}^n$ *se, e somente se, dado qualquer número real* $\varepsilon > 0$, *existe um índice* $i_0(\varepsilon) > 0$ *tal que a sequência* $\{x_i - x\}_{i\geq 0}$ *satisfaz*

$$\|x_i - x\| < \varepsilon \text{ , para todo } i \geq i_0(\varepsilon) . \tag{2.5}$$

Demonstração: (Ver [25, teorema 14.4]). ■

Esta definição vale para espaços métricos completos substituindo a norma pela métrica $d(x_i, x)$. A convergência de uma sequência gerada por um algoritmo não implica na convergência do algoritmo. A convergência do algoritmo se dará apenas se o limite x da sequência pertencer ao conjunto de soluções do problema. Em princípio (assintoticamente), não há diferença na escolha da norma para medir a convergência em $\mathbb{R}^n$ graças ao teorema da equivalência entre normas (só em dimensões finitas).

Teorema 3 *Se* $\|\cdot\|$ *e* $|\cdot|$ *são duas normas quaisquer em* $\mathbb{R}^n$, *então existem constantes* $c_2 \geq c_1 > 0$ *tais que*

$$c_1 \|x\| \leq |x| \leq c_2 \|x\| , \forall x \in \mathbb{R}^n. \tag{2.6}$$

Demonstração: (ver [27, cap. 2]). ■

Nem todas as normas em espaços normados de dimensão infinita satisfazem a desigualdade (2.6), porém sempre se pode escolher normas equivalentes, que satisfazem (2.6) nestes espaços. A simplicidade oferecida pelo Teorema 3 não é gratuita. De fato, existe um alto preço a ser pago pela convergência na norma em $\mathbb{R}^n$, em especial nos problemas mal comportados. Para melhor caracterizar tais problemas suponha a solução de um sistema linear $f(x) = Ax - b = 0$, de n equações em $\mathbb{R}^n$, com $A \in \mathbb{R}^{n\times n}$, A regular. Em uma dada iteração tem-se que $f = Ae$, então as normas Euclidianas da função (resíduo da equação) e do erro limitam-se mutuamente: $\|f\|_2 \leq \|A\|_2 \|e\|_2$ e $\|e\|_2 \leq \|A^{-1}\|_2 \|f\|_2$. Induzidos pelo Teorema 3 e pela regularidade de A podemos escrever

$$\frac{1}{\|A\|_2} \|e\|_{A^TA} \leq \|e\|_2 \leq \left\|A^{-1}\right\|_2 \|e\|_{A^TA} \tag{2.7}$$

onde $\|e\|_{A^TA} = \sqrt{e^T A^T A e} = \|f\|_2$, e $\|A\|_2 = \sqrt{\max_{1\leq i\leq n} |\mu_i|} = \sqrt{\rho(A^TA)}$, sendo $\rho(A^TA)$ o raio espectral de A^TA, e μ_i os valores singulares de A. A desigualdade (2.7) traduz uma propriedade de separação entre a norma do erro e a norma do resíduo bastante

indesejável. A situação é similar se medido, como de praxe, o erro relativo ao erro inicial

$$\frac{1}{\kappa_2}\frac{\|f\|_2}{\|f_0\|_2} \leq \frac{\|e\|_2}{\|e_0\|_2} \leq \kappa_2 \frac{\|f\|_2}{\|f_0\|_2} \leq \kappa_2^2 \frac{\|e\|_2}{\|e_0\|_2} \tag{2.8}$$

onde $\kappa_2 = \kappa_2(A) = \|A\|_2 \|A^{-1}\|_2$ é o número de condição de A (norma Euclidiana).

As duas últimas desigualdades explanam porque a norma do resíduo pode decrescer (crescer) enquanto que a do erro pode crescer (decrescer) levando a grandes diferenças entre estas normas - e portanto, na velocidade de convergência destas normas -, sobretudo em sistemas ditos mal-condicionados. Este problema existe em diversas áreas. Adaptação é uma delas. Pode-se ficar preocupado, por exemplo, com a qualidade dos deslocamentos obtidos em análises adaptativas cujo controle de erro é baseado em normas de energia. A convergência assintótica e a qualidade do método, da estratégia adaptativa, dos indicadores e estimadores de erro em norma de energia, não garantem a qualidade da resposta (por exemplo, deslocamentos em um sólido) em outra norma (por exemplo, ℓ_1, ℓ_2, e ℓ_∞) ou medida (o deslocamento em um certo ponto do sólido). É necessário que a separação estabelecida por desigualdades da forma (2.8) seja resolvida dentro da estratégia.

Do acima exposto, é evidente que parar algoritmos com base na norma dos resíduos (absolutos ou relativos) pode trazer grandes dificuldades [17] [31]. Presentemente, o autor deste trabalho vê a verificação do resíduo como uma verificação da consistência do algoritmo, e, nem sempre, da proximidade da resposta. Muito pouco se tem feito no sentido de resolver tais conflitos em algoritmos iterativos para sistemas lineares ou para sistemas não lineares [14] [13].

De fato, a propriedade de separação resulta de uma escolha inadequada da norma. Seja $L(\mathbb{R}^n)$ o conjunto (espaço vetorial) de todos os operadores lineares A de $\mathbb{R}^n$ em $\mathbb{R}^n$. O seguinte resultado (extensível a espaços métricos completos) mostra que existe uma norma para a qual a separação não existe.

Lema 4 *Se $A \in L(\mathbb{R}^n)$, então, dado um $\epsilon > 0$ existe uma norma em R^n tal que $\|A\| \leq \rho(A) + \epsilon$.*

Demonstração: (ver [27, cap. 2]). ■

Na próxima seção este resultado é extendido a mapeamentos não lineares. Infelizmente, encontrar a norma do Lema 4 pode vir a se constituir em uma tarefa muito difícil. Além disso, ela pode não ter um significado razoável.

Pode-se vislumbrar três amplas estratégias, não mutuamente excludentes, para a solução de tais problemas:

1. Estimar erros em qualquer medida desejada, ou seja, escolher uma medida de interesse e parar o algoritmo usando uma estimativa confiável desta medida. Este é um problema extremamente difícil. Amenizado para certas normas, como por exemplo a norma da energia.

2. Projetar algoritmos que não sofram a propriedade de separação. Esta estratégia entretanto não vem em geral associada a uma melhora na velocidade de convergência. As grandes questões são: (a) podemos projetar algoritmos que não sofram separação nas normas (ou medidas) de interesse e, (b) ao mesmo tempo não percam velocidade de convergência nestas normas ?

3. Usar medidas de convergência mais adequadas no projeto do algoritmo. Por exemplo, medidas relacionadas ao problema físico de interesse, não necessariamente associadas a uma norma.

É um dos objetivos desta tese contribuir nestas três direções. Abaixo alguns fatos básicos de análise [25, cap. 2].

Fato 2.1: *Uma sequência em $\mathbb{R}^n$ tem no máximo um limite.*

Fato 2.2: *Toda sequência convergente em $\mathbb{R}^n$ é limitada.*

Lema 5 *Se uma sequência $\{x_i\}_{i\geq 0}$ em $\mathbb{R}^n$ converge para um elemento x, então qualquer subsequência de $\{x_i\}_{i\geq 0}$ também converge para x.*

Definição 6 *Diz-se que uma sequência $\{x_i\}_{i\geq 0}$ em $\mathbb{R}^n$ é uma sequência de Cauchy se, para todo $\varepsilon > 0$, existe um número natural $i_c(\varepsilon)$ tal que, para todos $m, n \geq i(\varepsilon)$, se tenha $\|x_m - x_n\| < \varepsilon$.*

Daí, o seguinte corolário do Teorema 2.

Corolário 7 *Uma sequência em $\mathbb{R}^n$ é convergente se, e somente se, é uma sequência de Cauchy.*

Se X é um espaço normado, e mesmo métrico, toda sequência que converge fortemente em X, ou seja, $\lim d(x_i, x) = 0$, $x_i, x \in X$ é uma sequência de Cauchy, porém, a recíproca - ou seja, toda sequência de Cauchy converge fortemente para algum ponto - não é, em geral, verdadeira. Um espaço normado X (de dimensão finita ou infinita) é um espaço de Banach se é completo, ou seja, toda sequência de Cauchy em X converge fortemente para o seu limite em X.

2.2.4 Velocidade de convergência

O substrato da Parte I desta tese é a qualidade da convergência. A qualidade da convergência de um algoritmo é convencionalmente discutida ou em termos de sua velocidade de convergência assintótica, ou de sua convergência global. A convergência assintótica refere-se à qualidade da convergência no entorno da solução. A convergência global refere-se à desejada propriedade de certos algoritmos de convergirem para qualquer valor inicial $x_0 \in \mathbb{R}^n$. Uma definição de velocidade de convergência suficientemente adequada para incluir diversas medidas da qualidade de convergência é a convergência r-linear (em analogia com o teste da raiz para convergência de séries infinitas):

Definição 8 *Diz-se que uma sequência $\{x_i\}_{i\geq 0}$ em $\mathbb{R}^n$ converge r-linearmente para o seu limite x, com taxa α, se $\alpha \in [0,1[$ e existe algum $\eta \geq 0$ tal que para todo índice i,*

$$\|x_i - x\| \leq \eta\alpha^i \ . \tag{2.9}$$

A sequência converge r-superlinearmente se existe uma sequência $\{\beta_i\}_{i\geq 0}$ q-superlinearmente convergente (ver Definição 11) tal que

$$\|x_i - x\| \leq \beta_i$$

para todo índice i.

A seguinte proposição estabelece uma condição para que não seja necessário analisar todos os elementos da sequência para caracterizar a sua convergência r-linear.

Proposição 9 *Suponha que $\{x_i\}_{i\geq 0}$ é uma sequência em $\mathbb{R}^n$, p é um inteiro positivo, e x é um ponto em $\mathbb{R}^n$. Se $\{x_{pi}\}_{i\geq 0}$ converge r-linearmente para x e $\{\|x_i - x\|\}_{i\geq 0}$ é decrescente, então a sequência inteira $\{x_i\}_{i\geq 0}$ converge r-linearmente para x.*

Demonstração: (Vide também [29, Propos. 1.6]). Por hipótese, existe algum $\eta \geq 0$, e $\alpha \in [0,1[$, tais que

$$\|x_{pi} - x\| \leq \eta\alpha^i$$

para todo i. Fixe um inteiro positivo m, arbitrário, e divida por p com resto r, ou seja, $m = pi + r$, $0 \leq r \leq p-1$. Então a cstimativa segue-se

$$\|x_m - x\| \leq \|x_{pi} - x\| \leq \eta\left(\alpha^{\frac{1}{p}}\right)^{pi}$$

$$= \frac{\eta \left(\alpha^{\frac{1}{p}}\right)^{pi+r}}{\left(\alpha^{\frac{1}{p}}\right)^{r}} = \left(\frac{\eta}{\left(\alpha^{\frac{1}{p}}\right)^{p-1}}\right)\left(\alpha^{\frac{1}{p}}\right)^{m}$$

concluindo a prova. ■

A convergência r-linear é usada para caracterizar a convergência de algoritmos durante todas a iterações. Note que $\|x_i - x\|$ pode crescer, e pertencer à Definição 8. Tomando o logaritmo em ambos os lados da desigualdade (2.9) tem-se: $\log \|x_i - x\| \leq \log \eta + i \log \alpha$. Assim, o gráfico logarítmico da distância para o limite da sequência pelo número de iterações é limitado superiormente por uma reta com tangente negativa dada por $\log \alpha$, e $\|x_0 - x\| \leq \eta$, ou seja, η é um limite superior para o erro inicial.

Uma propriedade importante da convergência r-linear é que ela independe da norma com a qual se mede a convergência. Este fato tem consequências positivas e negativas. Do lado negativo, a tentativa de diferenciar a velocidade de convergência de um algoritmo nas diferentes medidas pela taxa é aniquilada pela seguinte proposição.

Proposição 10 *Se uma sequência converge r-linearmente em uma determinada norma, então ela converge r-linearmente em qualquer norma equivalente, com a mesma taxa. Em especial, em qualquer norma em* $\mathbb{R}^n$.

Demonstração: Por hipótese $\|x_i - x\| \leq \eta\alpha^i$, e seja $|\cdot|$ outra norma equivalente, então existe um $c > 0$ tal que $|\cdot| \leq c\,\|\cdot\|$, daí:

$$|x_i - x| \leq c\,\|x_i - x\| \leq (c\eta)\,\alpha^i. \tag{2.10}$$

Em $\mathbb{R}^n$, pelo Teorema 3, qualquer norma é equivalente, e a prova está completa. ■

Do lado positivo, a taxa caracteriza a qualidade da convergência de um algoritmo em métricas equivalentes, e a qualidade da convergência do algoritmo em uma certa medida fica caracterizada pela constante. A Proposição 10 mostra que a escolha da medida da qualidade de convergência deve ser muito cuidadosa, focada nos interesses. Resultados significativos são obtidos neste trabalho com a convergência do quociente.

Se existe um índice i_α a partir do qual a sequência $\{\|x_i - x\|\}_{i \geq i_\alpha}$ é decrescente, então, em analogia com a convergência de séries geométricas infinitas, pode-se introduzir a noção de convergência do quociente da norma do erro, convergência-q, que

caracteriza uma forma bastante desejada de convergência, e, sobretudo, repleta de propriedades. Entretanto, a noção de convergência fica estreitamente relacionada com a convergência assintótica.

Definição 11 *Uma sequência* $\{x_i\}_{i\geq 0}$, $x_i \in \mathbb{R}^n$, *é q-linearmente convergente para* $x \in \mathbb{R}^n$ *na norma* $\|\cdot\|$, *se existe um* $\alpha \in [0,1[$, *o fator-q de convergência, e um índice* $i_\alpha \geq 0$ *tal que*

$$\|x_{i+1} - x\| \leq \alpha \|x_i - x\| \tag{2.11}$$

para todo $i \geq i_\alpha$. *Se existir uma sequência* $\{\alpha_i\}_{i\geq 0}$ *de números reais que converge para zero, e tal que*

$$\|x_{i+1} - x\| \leq \alpha_i \|x_i - x\| \tag{2.12}$$

para todo i, a sequência é dita q-superlinearmente convergente. Se esta sequência $\{\alpha_i\}_{i\geq 0}$ *converge para* 1, *mas, mesmo assim a sequência* $\{x_i\}_{i\geq 0}$ *converge para seu limite x, diz-se que ela é q-sublinearmente convergente.*

Dos três tipos de convergência do quociente, a convergência q-superlinear também é independente da norma, porém apenas assintoticamente. Na mesma linha da Proposição 10, temos,

$$|x_i - x| \leq c \|x_i - x\| \leq c\alpha_i = \widehat{\alpha}_i \ , \tag{2.13}$$

onde $\lim \widehat{\alpha}_i = \lim c\alpha_i = c \lim \alpha_i = 0$. Porém, agora, a diferentes taxas locais.

No gráfico do logaritmo de $\|x_{i+1} - x\|$ pelo número de iterações, a região indicada pela desigualdade (2.11) é limitada superiormente por uma reta com coeficiente angular $\log \alpha$ (negativo), e passando no ponto $(i_\alpha, \log \|x_{i\alpha} - x\|)$. Se, a partir deste ponto, a sequência for apenas r-linearmente convergente, tudo é permitido acontecer abaixo desta reta; se a sequência for decrescente ela eventualmente converge pelo menos q-linearmente, ou q-superlinearmente, com a concavidade voltada para baixo. A concavidade voltada para cima corresponde exatamente à caracterização de q-sublinearidade, pela qual a sequência $\{\alpha_i\}_{i\geq i_\alpha}$ converge para 1 por valores à esquerda. A convergência q-sublinear assintótica é de fato excluída também da definição de convergência r-linear dada pela Definição 8. Por outro lado, convergência q-superlinear não caracteriza uma velocidade de convergência maior do que q-linear, tudo dependendo dos fatores-q.

A noção de convergência q-sublinear é pouco invocada na literatura [36], provavelmente por caracterizar uma velocidade de convergência indesejável. Entretanto, muitos algoritmos para sistemas lineares e problemas de viabilidade convexa, para

citar alguns, convergem assintoticamente, apenas q-sublinearmente. Algoritmos com convergência q-sublinear devem ser considerados pobres e inaceitáveis. Perturbações provocadas por aritmética finita normalmente o impedem de atingir uma proximidade aceitável do ponto fixo (quando existe, ver Capítulo 3). Para tais algoritmos, e mesmo para aqueles com convergência q-linear com o fator-q próximo da unidade, a prova teórica de convergência é insuficiente. É necessário alguma salvaguarda durante a execução. No Capítulo 3, a fragilidade da convergência q-sublinear será mostrada com mais rigor.

Contudo, o conceito de convergência q-sublinear ainda é usado extensivamente na literatura de métodos adaptativos em elementos finitos - sem uma citação direta. Em especial, as estratégias adaptativas baseadas em versões h e muitas estratégias baseadas em versões p convergem apenas q-sublinearmente em problemas com singularidades na solução ao longo de todas as iterações (etapas de adaptação), e nestes problemas, apenas algumas estratégias cuidadosas, geralmente baseadas em versões hp, conseguem convergência q-linear durante as fases iniciais e intermediária (pré-assintótica), e apenas convergência assintótica q-sublinear. Para melhor identificar o papel deste tipo de convergência, convém definir convergência sublinear de grau $p > 0$ e constante $c > 0$, como aquela para a qual

$$\|x_i - x\| \leq ci^{-p} \text{ para todo } i \geq 0 \tag{2.14}$$

Note que esse tipo de convergência satisfaz $\sum_{i=0}^{\infty} \|x_i - x\|^{1/p} = \infty$, [36]. Note também que, por comparação com (2.12) (usando o pior caso em ambas as definições), temos

$$\lim \alpha_i = \lim \frac{c(i+1)^{-p}}{ci^{-p}} = \lim \left(\frac{i+1}{i}\right)^{-p} = 1 \,,$$

para qualquer $p > 0$, e independente de c. Para prova rigorosa e resultados complementares ver subseção 3.5.2.

Assim, a definição de q-sublinearidade usada neste trabalho identifica-se com a dada pela desigualdade (2.14).Uma característica importante deste tipo de convergência é que se duas sequências $\{x_i\}$ e $\{\widehat{x}_i\}$ convergem com graus p e $\widehat{p} > p$ então $\|e_i\| = o(\|\widehat{e}_i\|)$. Stewart [36] observa o quão curioso é o fato de que nenhuma função analítica que converge sublinearmente possue grau $p > 1$. Há muito se conhece que o método dos elementos finitos, versão-h, converge sob a égide da desigualdade (2.14), [37, sec. 3.4, cap. 8 (ver figuras 8.4 a 8.6)], e, embora, eventualmente, com $p \geq 2$, mas, ainda assim, q-sublinearmente.

2.3 Teoria Métrica dos Pontos Fixos

O Lema de Newmann afirma que $(I-A)^{-1}$ existe para todos os elementos $A \in L(\mathbb{R}^n)$ para os quais o raio espectral é menor do que a unidade. Conclui-se que se $\|A\|_2 < 1$, então o algoritmo $x_{i+1} = Ax_i$ possui um único ponto fixo $x = Ax$. Este resultado extende-se de várias formas a mapeamentos não lineares. Esta seção sumariza e discute alguns resultados básicos sobre atrações.

2.3.1 Definições

Definição 12 *Diz-se que um mapeamento $\mathcal{A} : D \subset \mathbb{R}^n \to \mathbb{R}^n$ é Lipschitz-contínuo (ou é contínuo segundo Lipschitz) em D se existe um $\gamma \geq 0$ tal que para todo $x, y \in D$ tem-se*

$$d(\mathcal{A}(x), \mathcal{A}(y)) \leq \gamma d(x, y) \tag{2.15}$$

onde $d(\cdot,\cdot)$ é a métrica em $\mathbb{R}^n$. A menor constante γ que satisfaz (2.15) é chamada de constante de Lipschitz para $\mathcal{A}$.

As constantes de Lipschitz de dois mapeamentos $\mathcal{A}$ e $\mathcal{B}$ serão distinguidas por $\gamma(\mathcal{A})$ e $\gamma(\mathcal{B})$. Dados dois mapeamentos $\mathcal{A}$, $\mathcal{B} : D \subset \mathbb{R}^n \to \mathbb{R}^n$, o mapeamento composto terá

$$\gamma(\mathcal{A} \circ \mathcal{B}) \leq \gamma(\mathcal{A})\gamma(\mathcal{B}) \tag{2.16}$$

A constante de Lipschitz para uma determinada distância $d(\cdot,\cdot)$ será distinguida por γ_d. Será definido $\mathcal{A}^p$, a p-ésima potência do mapeamento $\mathcal{A}$, recursivamente, como $\mathcal{A}^0 = I$ e $\mathcal{A}^p(x) = \mathcal{A}(\mathcal{A}^{p-1}(x))$, $p \geq 1$. Com esta simbologia (2.16) diz que $\gamma(\mathcal{A}^i) \leq \gamma^i(\mathcal{A})$, $i = 1, 2, \ldots$.

Se a métrica é induzida por uma norma, então $\gamma(\mathcal{A}+\mathcal{B}) \leq \gamma(\mathcal{A}) + \gamma(\mathcal{B})$ e, para $\alpha \geq 0$, $\gamma(\alpha\mathcal{A}) = \alpha\gamma(\mathcal{A})$.

Definição 13 *Um mapeamento $\mathcal{A} : D \subset \mathbb{R}^n \to \mathbb{R}^n$ é contractivo, ou é uma contração, em um conjunto $D_0 \subset D$ para uma métrica $d(\cdot,\cdot)$ se $\gamma_d < 1$ para todo $x, y \in D_0$, não expansivo se $\gamma_d = 1$, e estritamente não expansivo se $d(\mathcal{A}(x), \mathcal{A}(y)) < d(x, y)$ sempre que $x \neq y$.*

Contrações possuem pontos fixos, enquanto que os mapeamentos não expansivos e os estritamente não expansivos podem não os possuir. A teoria dos mapeamentos não expansivos difere da teoria da contrações, principalmente porque os processos iterativos da forma $x_{i+1} = \mathcal{A}(x_i)$ em geral não convergem, mesmo que o mapeamento

possua pontos fixos. Dentre as questões básicas no estudo de mapeamentos não expansivos, as seguintes são de maior interesse nesta tese:

1. Que hipóteses devem ser adicionadas à estrutura do conjunto $D \subset \mathbb{R}^n$, e que restrições sobre $\mathcal{A}$, para assegurar a existência de pelo menos um ponto fixo ?
2. Qual a estrutura do conjunto $Fix(\mathcal{A})$?
3. O que pode ser dito acerca do comportamento dos iterandos de $\mathcal{A}$, em especial sobre a consistência, a velocidade de convergência, e, principalmente, a proximidade do ponto fixo ?

Nas duas últimas décadas a teoria métrica dos pontos fixos tem recebido um enorme impulso, no sentido em que muitos resultados foram desenvolvidos em espaços de Banach, ou mesmo em espaços mais gerais. Embora sejam de interesse em uma teoria geral de algoritmos, estando este trabalho dirigido principalmente a espaços Euclidianos, adotaremos a técnica de restringir os resultados a tais espaços e tecer comentários, quando for de interesse posterior, à validade dos resultados em espaços mais gerais. Continuará sendo usada a notação $d(x, y)$ para a métrica, que, com raras exceções a serem discutidas em capítulos posteriores, pode ser entendida como associada a uma norma, não necessariamente quadrática.

2.3.2 Resultados básicos sobre atrações

O resultado fundamental da Teoria Métrica dos Pontos Fixos é o Teorema do Mapeamento Contractivo, devido a Banach. Ele é usado para provar uma série de resultados adicionais, incluido-se os sobre mapeamentos não expansivos. O Teorema e sua prova, dados adiante [26], não são construtivos, entretanto os mecanismos usados na prova dão uma excelente compreensão do resultado, e serão usados posteriormente, em outro contexto.

Teorema 14 (Banach-I) *Suponha que $\mathcal{A} : D \subset \mathbb{R}^n \to \mathbb{R}^n$ é uma contração em um conjunto fechado $D_0 \subset D$ e que $\mathcal{A}(D_0) \subset D_0$. Então $\mathcal{A}$ tem um único ponto fixo em D_0.*

Demonstração: Defina $a = \inf\{d(x, \mathcal{A}(x)) : x \in D_0\}$ e $\gamma = \gamma_d(\mathcal{A})$. Precisa-se provar que $a = 0$. Seja $\epsilon > 0$ e escolha $x \in D_0$ tal que $d(x, \mathcal{A}(x)) \leq a + \epsilon$. Então,

$$a \leq d(\mathcal{A}(x), \mathcal{A}^2(x)) \leq \gamma d(x, \mathcal{A}(x)) \leq \gamma(a + \epsilon).$$

Uma vez que $\gamma < 1$, e ϵ pode ser feito arbitrariamente pequeno, segue-se que $a = 0$, e portanto $Fix(\mathcal{A})$ não é vazio. Agora, a prova de que ele tem apenas um elemento. Para qualquer $\epsilon > 0$ o conjunto

$$D_\epsilon = \{x \in D : d(x, \mathcal{A}(x)) \leq \epsilon\} \tag{2.17}$$

é fechado e não é vazio. Para quaisquer $x, y \in D_\epsilon$,

$$d(x,y) \leq d(x, \mathcal{A}(x)) + d(\mathcal{A}(x), \mathcal{A}(y)) + d(\mathcal{A}(y), y) \leq 2\epsilon + \gamma d(x,y)$$

o que fornece

$$d(x,y) \leq \frac{2\epsilon}{1-\gamma},$$

seque-se então que o $\lim_{\epsilon \to 0} \text{diam } D_\epsilon = \lim_{\epsilon \to 0} \sup \{d(x,y) : x, y \in D_\epsilon\} = 0$. Desde que a família D_ϵ retrai-se quando $\epsilon \to 0$, o Teorema da Intersecção de Cantor [25, Teorema 11.4] implica em que $\bigcap_{\epsilon > 0} D_\epsilon$ consiste em exatamente um ponto, digamos x, que como visto, é fixo sob $\mathcal{A}$, ou seja, $x = \mathcal{A}(x)$. ■

Quando $\mathcal{A}$ é uma contração, há interesse em saber sob quais condições ele converge para o seu ponto fixo e a qualidade da convergência.

Teorema 15 (Banach-II) *Com as hipóteses do Teorema 14 a sequência* $x_{i+1} = \mathcal{A}(x_i)$, $i = 0, 1, \ldots$, *converge r-linearmente para qualquer* $x_0 \in D_0$.

Demonstração: Desde que $x_0 \in D_0$ e $\mathcal{A}(D_0) \subset D$, a sequência $\{x_i\}$ é bem definida, contida em D_0, e, notando que para quaisquer índices $i, p \in \mathbb{N}$, temos

$$\begin{aligned}
d(x_i, x_{i+p}) &= d(\mathcal{A}^i(x_0), \mathcal{A}^{i+p}(x_0)) = d(\mathcal{A}^i(x_0), \mathcal{A}^i \circ \mathcal{A}^p(x_0)) \\
&\leq \gamma(\mathcal{A}^i) d(x_0, \mathcal{A}^p(x_0)) \\
&\leq \gamma^i \left[d((x_0), \mathcal{A}(x_0)) + d(\mathcal{A}(x_0), \mathcal{A}^2(x_0)) + \cdots + d(\mathcal{A}^{p-1}(x_0), \mathcal{A}^p(x_0)) \right] \\
&\leq \gamma^i \left[1 + \gamma + \gamma^2 + \cdots + \gamma^{p-1} \right] d(x_0, \mathcal{A}(x_0)) \\
&\leq \gamma^i \left[\frac{1-\gamma^p}{1-\gamma} \right] d(x_0, \mathcal{A}(x_0))
\end{aligned} \tag{2.18}$$

Tomando o limite $p \to \infty$ tem-se que $d(x_i, x_{i+p}) \to d(x_i, x)$ onde x é o único ponto fixo de $\mathcal{A}$ em D_0, devido ao Teorema 14; a qualidade da convergência segue-se

$$d(x_i, x) = d(\mathcal{A}^i(x_0), x) \leq \left(\frac{d(x_0, \mathcal{A}(x_0))}{1-\gamma} \right) \gamma^i \tag{2.19}$$

concluindo a prova. ■

Para provar a convergência q-linear é necessário impor mais restrições sobre o mapeamento do que as contidas no Teorema 14. Esta construção será feita em etapas, acompanhada de uma discussão que será explorada nos capítulos posteriores.

Primeiramente, a prova do Teorema 15 revela que a hipótese $\gamma(\mathcal{A}) < 1$ é mais forte do que necessária. É suficiente assumir que $\gamma(\mathcal{A}^i) < 1$ para ao menos um índice i. Isto porque se $\mathcal{A}^i$ é uma contração, pelo Teorema 15 possui um único ponto fixo x. Então $\mathcal{A}(x) = \mathcal{A}^{i+1}(x) = \mathcal{A}^i \circ \mathcal{A}(x)$, e assim $\mathcal{A}(x)$ é também um ponto fixo de $\mathcal{A}^i$. Portanto $x = \mathcal{A}(x)$, provando que x também é um ponto fixo de $\mathcal{A}(x)$, e assim único.

Explorando a idéia acima pode-se exigir que o mapeamento nos Teoremas 14 e 15 seja apenas Lipschitz-contínuo em lugar de contractivo. Neste caso a estimativa de erro pode ser obtida a partir de:

$$\begin{aligned} d(x_i, x_{i+p}) &\leq \sum_{k=i}^{i+p-1} d(\mathcal{A}^k(x_0), \mathcal{A}^{k+1}(x_0)) \leq \left[\sum_{k=0}^{p} \gamma(\mathcal{A}^{i+k})\right] d(x_0, \mathcal{A}(x_0)) \\ &\leq \gamma(\mathcal{A}^i) \left[\sum_{k=0}^{p} \gamma(\mathcal{A}^k)\right] d(x_0, \mathcal{A}(x_0)) \\ &\leq \gamma(\mathcal{A}^i)\sigma_p d(x_0, \mathcal{A}(x_0)) \end{aligned} \tag{2.20}$$

Então, $\{x_i\}_{i\geq 0}$ será uma sequência de Cauchy se $\sigma_\infty = \lim_{p\to\infty} \sigma_p = \sum_{k=1}^{\infty} \gamma(\mathcal{A}^k) < +\infty$. É possível verificar que existe um número $\gamma_\infty(\mathcal{A})$ que satisfaz

$$\gamma_\infty(\mathcal{A}) = \lim_{k\to\infty} \left[\gamma(\mathcal{A}^k)\right]^{1/k} = \inf\left\{\left[\gamma(\mathcal{A}^k)\right]^{1/k} : k = 1, 2, \ldots\right\}. \tag{2.21}$$

Assim $\sum_{k=1}^{\infty} \gamma(\mathcal{A}^k) < +\infty$ se, e somente se, $\gamma_\infty(\mathcal{A}) < 1$. Desta forma a hipótese $\gamma(\mathcal{A}) < 1$ pode ser substituida por $\gamma_\infty(\mathcal{A}) < 1$ no Teorema 14, e a estimativa do Teorema 15 será dada por

$$d(x_i, x) \leq \gamma(\mathcal{A}^i)\sigma_\infty d(x_0, \mathcal{A}(x_0)). \tag{2.22}$$

O acima exposto, no entanto, não fornece um teorema mais forte do que o Teorema 15. Isto devido ao teorema da equivalência entre normas pelo qual se pode provar que, [26, pag. 10],

$$\gamma_\infty(\mathcal{A}) = \inf \gamma_r(\mathcal{A}) \tag{2.23}$$

para todas as métricas r em $\mathbb{R}^n$ (ou para métricas equivalentes em qualquer espaço métrico). A seguinte proposição sintetiza esta discussão.

Proposição 16 *Qualquer mapeamento $\mathcal{A}$ para o qual $\gamma_\infty(\mathcal{A}) < 1$ é uma contração com respeito a uma métrica apropriadamente escolhida.*

Os seguinte resultados formalizam alguns comentários anteriores, como uma generalização dos teoremas de contração.

Teorema 17 *Suponha que $\mathcal{A} : D \subset \mathbb{R}^n \to \mathbb{R}^n$ mapeia um conjunto fechado $D_0 \subset D$ nele próprio e que para algum inteiro $p \geq 1$ se tenha*

$$d(\mathcal{A}^{ip}(x), \mathcal{A}^{ip}(y)) \leq \gamma_i d(x, y), \ \forall x, y \in D_0, \ i = 1, 2, \ldots, \tag{2.24}$$

onde $\sigma_\infty = \sum_{i=1}^{\infty} \gamma_i < +\infty$. Então $\mathcal{A}$ tem um único ponto fixo $x_\star \in D_0$, e para qualquer $x_0 \in D_0$ a sequência $x_{i+1} = \mathcal{A}(x_i)$ converge para $x_\star$, com a seguinte estimativa de erro

$$d(x_i, x_\star) \leq \sigma_\infty d(x_i, x_{i-p}) \ , \ i = 1, 2, \ldots \tag{2.25}$$

Demonstração: (Para prova mais detalhada ver [27, pag. 384]). ■

Aparentemente, a qualidade de convergência fica prejudicada com a hipótese $\sigma_\infty < +\infty$. É possível garantir apenas alguma forma de convergência sublinear. A convergência q-linear é provada com hipóteses mais restritivas, como as do resultado abaixo.

Teorema 18 *Nas hipóteses do Teorema 17, se para algum $p > 1$ o mapeamento satisfaz*

$$d(\mathcal{A}^p(x), \mathcal{A}^p(y)) \leq \gamma d(x, y), \ \gamma < 1, \tag{2.26}$$

então (2.24) é satisfeita com $\gamma_i = \gamma^i$, $\sigma_\infty = \sum_{i=1}^{\infty} \gamma^i = \gamma/(1-\gamma)$ e a estimativa de erro torna-se

$$d(x_i, x_\star) \leq \frac{\gamma}{1-\gamma} d(x_i, x_{i-p}) \ , \ i = 1, 2, \ldots \tag{2.27}$$

Demonstração: Imediata a partir do Teorema 17. ■

Quando, além de satisfazer as hipóteses do Teorema 18, $\mathcal{A}$ é também Lipschitz-contínuo, então as hipóteses do Teorema 17 são satisfeitas para $p = 1$. Entretanto, é possível que (2.24) ou mesmo (2.26) sejam satisfeitas mesmo que $\mathcal{A}$ não seja contínuo (ver Capítulo 3).

O caso especial de (2.26) com $p = 1$ fornece os Teoremas 14 e 15, porém agora com a prova da convergência q-linear e a estimativa do erro dada por (2.27).

Um mapeamento que satisfaz (2.26) para algum $\gamma > 0$ e $p = 1, 2, \ldots$, é chamado de *uniformemente Lipschitziano,* ou seja,

$$d(\mathcal{A}^i(x), \mathcal{A}^i(y)) \leq \gamma d(x, y), \ \gamma > 0, \ i = 1, 2, \ldots . \tag{2.28}$$

Para tais mapeamentos pode-se estabelecer um limite γ para o qual existe um ponto fixo, com a exigência adicional de convexidade.

Teorema 19 *Se $\mathcal{A} : D \subset \mathbb{R}^n \to \mathbb{R}^n$ mapeia um conjunto limitado, fechado e convexo $D_0 \subset D$ nele próprio, e $\mathcal{A}$ é uniformemente Lipschitziano com $\gamma < \sqrt{2}$, então $\mathcal{A}$ tem um ponto fixo em D_0.*

Demonstração: (Ver [26, pags. 172-3], válido apenas em espaços de Hilbert). ■

Obviamente, expansividade estrita não garante a existência de um ponto fixo. Porém, o Teorema de Edesltein fornece condições para a existência e unicidade do ponto fixo.

Teorema 20 (Edelstein) *Suponha que $\mathcal{A} : D \subset \mathbb{R}^n \to \mathbb{R}^n$ mapeia um conjunto $D_0 \subset D$ nele próprio, que $\mathcal{A}(D_0)$ é compacto, e $\mathcal{A}$ é estritamente não expansivo em D_0. Então, para qualquer $x_0 \in D_0$ as iterações $x_{i+1} = \mathcal{A}(x_i)$, $i = 0, 1, \ldots$ convergem para o único ponto fixo de $\mathcal{A}$ em D_0.*

Demonstração: (Ver [27, pag. 404]). ■

Note que $\mathbb{R}^n$ não é compacto, de tal forma que o Teorema de Edelstein não pode ser generalizado para um teorema de convergência global. Na prova do teorema a compacidade não pode ser substituida pela hipótese de que $\mathcal{A}(D_0)$ é fechado e limitado. Porém, pelo Teorema de Heine-Borel, todo subconjunto de $\mathbb{R}^n$ que é limitado e fechado é compacto, e vice-versa [25, teorema 11.3]. Entretanto o Teorema de Edelstein não caracteriza a qualidade da convergência nem fornece uma estimativa para o erro.

O seguinte resultado é um caso particular do Teorema Fundamental dos Pontos Fixos de Schauder, porém um pouco mais métrico.

Teorema 21 (Schauder) *Suponha que $\mathcal{A} : D \subset \mathbb{R}^n \to \mathbb{R}^n$ mapeia um conjunto $D_0 \subset D$ nele próprio, D_0 fechado e limitado, e $\mathcal{A}$ é não expansivo em D_0. Então $\mathcal{A}$ tem um ponto fixo.*

Demonstração: (vide [26, Teorema 3.1]. (Válido em espaços de Banach, com a condição adicional de compacidade).

2.3.3 Contrações iteradas

Definição 22 *Diz-se que um mapeamento $\mathcal{A} : D \subset \mathbb{R}^n \to \mathbb{R}^n$ é uma contração iterada no conjunto $D_0 \subset D$ se existe um $\gamma < 1$ tal que*

$$d(\mathcal{A} \circ \mathcal{A}(x), \mathcal{A}(x)) \leq \gamma d(\mathcal{A}(x), x) \tag{2.29}$$

toda vez que x e $\mathcal{A}(x)$ estejam em D_0.

Se $\mathcal{A}$ é uma contração em D_0, então ele também é uma contração iterada, porém a recíproca não é necessariamente verdadeira. As contrações iteradas não precisam ser contínuas, e seus pontos fixos sequer precisam ser únicos. Contrações iteradas possuem uma certa falta de propriedades. O seguinte resultado estabelece algumas condições para a sua convergência e estimativa de erro.

Teorema 23 *Seja $\mathcal{A} : D \subset \mathbb{R}^n \to \mathbb{R}^n$ uma contração iterada em um conjunto fechado $D_0 \subset D$, e que, para algum $x_0 \in D_0$ a sequência $x_{i+1} = \mathcal{A}(x_i)$, $i = 0, 1, \ldots$, permanece em D_0. Então, $\lim x_i = x \in D_0$, e vale a estimativa*

$$d(x_{i+1}, x) \leq \frac{\gamma}{1-\gamma} d(x_{i+1}, x_i) \ . \tag{2.30}$$

Ainda mais, se $\mathcal{A}$ é contínuo em x então $x = \mathcal{A}(x)$.

Demonstração: (Ver [27, pag. 401]). ■

Deve-se observar que o ponto fixo no Teorema 23 acima não precisa ser único. É possível relaxar a condição de contração e obter ainda um ponto fixo sob condições mais severas.

Teorema 24 *Suponha que $\mathcal{A} : D \subset \mathbb{R}^n \to \mathbb{R}^n$ mapeia um conjunto D_0 nele próprio, $\mathcal{A}(D_0)$ é compacto, e*

$$d(\mathcal{A} \circ \mathcal{A}(x), \mathcal{A}(x)) < d(\mathcal{A}(x), x)) \tag{2.31}$$

para todo $x \in D_0$ e $x \neq \mathcal{A}(x)$. Suponha ainda que $\mathcal{A}$ é contínuo em D_0 e possua no máximo um ponto fixo em D_0. Então, um ponto fixo $x_\star$ existe, e, para qualquer $x_0 \in D_0$, a sequência $x_{i+1} = \mathcal{A}(x_i)$ converge para $x_\star$.

Demonstração: (Ver [27, pag. 403]. Válido para qualquer espaço métrico compacto [26, pag. 12]). ■

A condição (2.31) é evidentemente satisfeita se $\mathcal{A}$ é estritamente não-expansivo, dando origem ao Teorema de Edelstein.

2.4 Comentários finais

Neste capítulo foram estabelecidas algumas notações e definições, e relembrados alguns fatos sobre a convergência de sequências e algoritmos em $\mathbb{R}^n$. Foram acrescidos alguns comentários e resultados menores, sobre tópicos que assumirão importância maior nos capítulos subsequentes. O principal aspecto do capítulo é ter o foco na qualidade da convergência, o que se realmente está interessado, uma vez estabelecida a convergência. Neste sentido, a convergência q-linear, e a convergência q-superlinear sob convergência q-linear ($\alpha_i < 1$), resultam como a melhor caracterização de velocidade de convergência. Para concluir, há necessidade de comentários adicionais.

O melhor que se tem conseguido provar sobre a qualidade de convergência de alguns bons algoritmos iterativos para sistemas lineares é que convergem apenas r-linearmente, e, eventualmente, r-superlinearmente. Raramente o comportamento assintótico r-superlinear é visto em uma realização. O critério de parada é acionado antes. Também, interfere a noção de epsilon da máquina - o *macheps*.

Macheps é o menor número positivo ϵ tal que $1+\epsilon > 1$. Os computadores atuais de 64 bits (ou 2 vezes 32 bits) possuem $macheps = 2^{-52} \simeq 2.22 \times 10^{-16}$. Significando que a representação de um número de ponto flutuante x encontra-se no intervalo $]x(1 - macheps), x(1 + macheps)[$. Dois números reais x e y têm a primeira metade de seus algarismos significativos aproximadamente iguais quando

$$\frac{|x-y|}{|x|} \leq \sqrt{macheps}. \tag{2.32}$$

É consenso entre analistas numéricos que, do ponto de vista prático, convém ficar satisfeito com precisões não maiores que a dada por (2.32). Ou seja, testes de erro relativo práticos não deveriam ser mais exigentes do que, por exemplo,

$$\frac{\|e_i\|}{\|x_*\|} \to \frac{\|e_i\|}{\|x_i\|} \leq \sqrt{macheps}. \tag{2.33}$$

Em sistemas lineares, a convergência q-superlinear e, normalmente, a r-superlinear, somente se pronuncia perto do limite dado por (2.33), ficando assim escondida pela imprecisão dos resultados, dando origem, muitas vezes a sublinearidades antes da parada. Atualmente, em sistemas lineares não simétricos, o usuário deve sentir-se muito feliz em conseguir parar dentro de uma região de convergência q-linear. Em se tratando de métodos de Krylov, isto vai depender muito da qualidade do precondicionador. Em sistemas lineares simétricos e positivos definidos, um dos métodos

mais usado é o CG. Neste caso, é possível com frequência parar em um trecho com alguma superlinearidade do quociente. Em contraste, nos sistemas não lineares, normalmente é possível parar em regiões q-lineares ou mesmo q-superlineares. O motivo básico para isto é que geralmente os métodos estão associados a estratégias de convergência global que o levam gradativamente cada vez mais perto da solução, por exigência da precisão dos algoritmos subjacentes que são muito sensíveis a imprecisões (por exemplo, em plasticidade incremental).

É muito difícil provar que um algoritmo é contractivo em alguma norma específica. Em sistemas lineares isto depende muito da estrutura algébrica da matriz de coeficientes (por exemplo, da distribuição dos autovalores e até mesmo dos dados iniciais). Esta dificuldade também existe para outras atrações. Do ponto de vista de estimativa de erros, as atrações mais interessantes seriam as contrações iteradas, ou alguma forma mais fraca. Entretanto, elas carecem de propriedades.

Por outro lado, é muito mais fácil, como será provado neste trabalho, construir algoritmos r-convergentes, ou mesmo q-convergentes, ainda que em uma forma mais fraca como $\|e_{i+1}\| \leq \|e_i\|$, decrescentes, ou $\|e_{i+1}\| < \|e_i\|$, estritamente decrescentes, em alguma norma pré-estabelecida. Do ponto de vista de velocidade de convergência e contenção de memória, as atrações não são essenciais. De fato, enquanto é possível garantir que toda contração, e muitas atrações, convergem q-linearmente, não vale a recíproca. Será provado que a classe de algoritmos de convergência q-linear é potencialmente mais ampla do que a das atrações, mesmo em espaços de Banach. Também, as definições de q-linearidade e q-superlinearidade asseguram convergência, por definição, o que não ocorre, por exemplo, com a classe q-linear e q-superlinear de ordem p. Enquanto que as contrações garantem convergência apenas sobre certas condições, às vezes muito severas.

Em espaços de Hilbert, os operadores de projeção permitem construir algoritmos q-convergentes e decrescentes em norma. Nestes casos, serão mostradas condições suficientes para um tipo de convergência do quociente a partir do qual estimadores de erros confiáveis para os algoritmos podem ser obtidos. Isto é importante, também, porque a grande maioria dos algoritmos bem sucedidos para sistemas lineares, por exemplo, são métodos de projeção em espaços Euclidianos. O interesse extende-se às áreas de viabilidade convexa e otimização, entre outras.

Capítulo 3

Convergência do Quociente

In this book we will be interested mainly in methods that are locally q-superlinearly or q-quadratically convergent and for which this behavior is apparent in practice.

J.E. Dennis e R. B. Schnabel, 1982, 1996 [19]

3.1 Introdução

Este capítulo trata da qualidade de convergência local sob a ótica da convergência do quociente do erro, ou convergência-q. São estudadas as suas propriedades, as suas conexões com outros tipos de convergência, as suas relações com as atrações básicas, e as estimativas de erros de aproximação de algoritmos q-convergentes.

O estudo de estimativas de erros está intimamente relacionado com o estabelecimento de regras e estratégias de interrupção de algoritmos iterativos. Embora o maior interesse sejam os sistemas lineares, ocorre que os resultados apresentados, pelas ferramentas utilizadas, aplicam-se a uma classe bem maior de algoritmos.

A principal razão pela qual foi escolhida a convergência do quociente como hipótese fundamental, à parte de suas propriedades, está na convicção de que o mínimo que se pode esperar de um algoritmo é que ele possua convergência q-linear local e assintótica, e, idealmente, convergência q-superlinear.

A convergência q-superlinear assintótica é observada, e eventualmente provada, em numerosos algoritmos para problemas não lineares [19]. Em sistemas lineares, a convergência q-superlinear, quando existe, nem sempre se pronuncia antes do acionamento do critério de parada. Muitas vezes há sublinearidades sob as quais se é forçado a parar.

Ainda em sistemas lineares, muitos bons algoritmos disponíveis trafegam com r-linearidades locais. Isto cria um conflito entre a teoria desenvolvida neste capítulo, baseada em convergência-q, e a realidade factual daqueles algoritmos. Este conflito é resolvido com um resultado que possibilita transitar com as estimativas entre as duas qualidades de convergência.

As estimativas de erros para convergência q-sublinear oferece algumas dificuldades. Neste caso, procura-se providenciar pontes com os teoremas de contração da teoria métrica dos pontos fixos. Existem muitas dificuldades devido à carência de propriedades das atrações que mais interessam, quais sejam, as atrações iteradas não expansivas. Por outro lado, exigindo apenas um pouco mais da q-sublinearidade, mediante uma nova definição, de q-sublinearidade estrita, mostra-se que os seus erros podem ser adequadamente estimados (graças à sua lentidão) por análise de qualquer mapeamento composto.

Cabe uma advertência. Não é possível estimar erros de aproximação de um algoritmo de qualquer espécie, senão por uma análise do enriquecimento da solução a cada iteração. O significado cruel deste princípio é que a qualidade da estimativa depende da regularidade deste enriquecimento, o qual sempre exige um esforço computacional não desprezível. Do ponto de vista de estimativas de erros, a chave para abrandar este esforço e relaxar a regularidade do enriquecimento está nos resultados de uma análise antecipada (*a priori*) do algoritmo.

Finalmente, é preciso salientar que a análise ora apresentada difere das demais existentes na literatura porque é focada em dois aspectos objetivos. Primeiro, nas estimativas de erros. Segundo, na compreensão dos mecanismos algorítmicos que levam a questões relevantes como velocidade de convergência, estabilidade das iterações e robustez. Estes dois aspectos, obviamente, são centrais em computação científica, e sempre foram o objeto de estudo, mas, o que difere neste texto é que são focados a partir da qualidade da convergência.

3.2 Expandindo Definições

Neste capítulo, a menos que seja dito o contrário, valem as seguintes hipóteses. A sequência $\{x_i\}_{i\geq 0}$ é convergente em um conjunto $D \subset \mathbb{R}^n$, munido de uma métrica $d(\cdot,\cdot)$, ou uma norma $\|\cdot\|$ definida ou não por um produto escalar $(\cdot,\cdot)$. Porém, alguns resultados são enunciados para espaços mais gerais. Cada situação especial ficará clara no contexto, ou será explicitada. A sequência é a produzida por um algoritmo

da forma $x_{i+1} = \mathcal{A}(x_i)$, com $x_\star \in Fix(\mathcal{A})$ único em D, ou em um subconjunto $D_0 \subset D$, e $\mathcal{A}$ é suposto consistente com o problema (2.3), quando esta hipótese se fizer necessária. Estas últimas hipóteses não são restritivas porque o interesse é a qualidade da convergência.

Com o objetivo de expandir o espectro de algoritmos para os quais a teoria deste capítulo se aplica, convém estabelecer algumas novas definições de formas mais fracas de convergência. A necessidade destas definições se tornarão óbvias ao longo do texto.

Definição 25 *Uma sequência* $\{x_i\}_{i\geq 0}$, $x_i \in \mathbb{R}^n$, *possui convergência q-sublinear estrita se converge de forma q-sublinear no sentido da Definição 11, com a desigualdade estrita, ou seja,* $\|x_{i+1} - x\| < \|x_i - x\|$.

Definição 26 *Uma sequência* $\{x_i\}_{i\geq 0}$, $x_i \in \mathbb{R}^n$, *possui convergência q-sublinear (respc., q-sublinear estrita, q-linear, q-superlinear) em* p_k*-etapas na norma* $\|\cdot\|$, *se existe uma sequência de inteiros* $\{p_k\}_{k\geq 0}$ *(cada* $p_k \geq 0$*) tal que a subsequência* $\{x_{p_k}\}_{k\geq 0}$ *converge de forma q-sublinear (respc., q-sublinear estrita, q-linear, q-superlinear) no sentido da Definição 11. Se a sequência* $\{p_k\}_{k\geq 0}$ *possui razão aritmética constante* $p = p_{k+1} - p_k \geq 1$ *a sequência é dita q-sublinear (respc., q-sublinear estrita, q-linear, q-superlinear) em p-etapas.*

Pode-se considerar uma métrica $d(\cdot,\cdot)$ em espaços métricos mais gerais, completos, nas definições acima, com a hipótese de que o $\lim d(x, x_\star) = 0$ (convergência forte) [38],[39]. Também, simplifica-se a notação: $d(\mathcal{A}(x_{i-1}), \mathcal{A}(x_i)) \equiv d(x_i, x_{i+1})$, tipicamente.

É imediato que uma sequência com convergência do quociente em *uma*-etapa, converge em qualquer p_k-etapas, para qualquer sequência $\{p_k\}_{k\geq 0}$, e, no caso da convergência q-linear, com fator-q menor (mais veloz) ou igual à sequência original.

Naturalmente, essa última propriedade sugere alguns algoritmos simples para encontrar uma cota inferior, digamos $\tilde{\alpha}$, para o fator-q da sequência original por análise de qualquer de suas subsequências. Por exemplo, se dadas subsequências quaisquer $\{p_k\}_{k\geq 0}$ com fator-q dado por α_{p_k}, teremos $\tilde{\alpha} = \max \alpha_{p_k} \leq \alpha$. E se $\alpha \leq \frac{1}{2}$, então $\tilde{\alpha} \leq \alpha \leq 1 - \tilde{\alpha}$. Esta cota inferior será melhorada se determinada a partir de muitas sequências $\{p_k\}_{0\leq k\leq m}$, com o maior possível valor de m. Pela natureza das cotas superiores para o erro (ver adiante), o algoritmo descrito funciona apenas se α está suficientemente distante de 1. Para escolhas adequadas das subsequências, ele converge assintoticamente para o α exato. Na prática, este algoritmo geralmente

é muito pobre e há de se encontrar algo melhor, em particular, na ausência de informações *a priori* sobre o fator-q do algoritmo. Porém, como será mostrado, a estimativa *a priori* do fator-q do algoritmo pode ser muito conservadora.

Para muitos algoritmos, em uso nos dias de hoje, há apenas prova de convergência q-sublinear estrita ou r-linear. Deve-se estar habilitado a usar os resultados deste capítulo em tais algoritmos. A seguinte proposição mostra como aplicar à q-sublinearidade os resultados da q-linearidade em p-etapas, $p \geq 2$.

Proposição 27 *Se uma sequência* $\{x_i\}_{i\geq 0}$, $x_i \in \mathbb{R}^n$, *possui convergência q-sublinear estrita, então ela possui convergência pelo menos q-linear em* 2*-etapas.*

Demonstração: Por hipótese, para cada três termos consecutivos da sequência,

$$\|x_{i+2} - x\| < \|x_{i+1} - x\| < \|x_i - x\| \ ,$$

e, dividindo por $\|x_i - x\| \neq 0$,

$$\frac{\|x_{i+2} - x\|}{\|x_i - x\|} < \frac{\|x_{i+1} - x\|}{\|x_i - x\|} = \alpha_i < 1 \ .$$

Agora, escolha, para cada α_i, i ímpar (por conveniência), um $\epsilon_i > 0$, tal que a primeira desigualdade estrita resulte em uma desigualdade simples

$$\frac{\|x_{i+2} - x\|}{\|x_i - x\|} \leq \alpha_i - \epsilon_i < 1 - \epsilon_i \ .$$

Usando essas desigualdades, é possível definir

$$\alpha = \max_{i\geq i_\alpha} \frac{\|x_{i+2} - x\|}{\|x_i - x\|} \leq \max_{i\geq i_\alpha} (\alpha_i - \epsilon_i) < \max_{i\geq i_\alpha} (1 - \epsilon_i) = 1 - \min_{i\geq i_\alpha} \epsilon_i < 1 \ ,$$

para todos os índices ímpares. Assim, pode-se escrever, para todo índice ímpar

$$\frac{\|x_{i+2} - x\|}{\|x_i - x\|} \leq \alpha < 1 \ ,$$

e a q-linearidade em 2-etapas está provada. ■

Usando argumentos similares, pode-se mostrar também o seguinte resultado.

Proposição 28 *Se uma sequência* $\{x_i\}_{i\geq 0}$, $x_i \in \mathbb{R}^n$, *possui convergência q-sublinear estrita, então existe pelo menos finitamente muitas sequências* $\{p_k\}_{k\geq 0}$ *não vazias tais que a sequência* $\{x_i\}_{i\geq 0}$ *converge q-linearmente em* p_k*-etapas.*

Demonstração: Imediata: escolha um índice inicial i arbitrário e subsequências arbitrárias $\{p_k\}_{k\geq 0}$ e aplique a Proposição 27. ■

Para sequências r-linearmente convergentes procede-se de maneira similar.

Proposição 29 *Se $\{x_i\}_{i\geq 0}$, $x_i \in \mathbb{R}^n$, é uma sequência r-linearmente convergente, então ela possui subsequências estritamente q-sublinearmente convergentes.*

Demonstração: Por hipótese, existem $\eta > 0$ e $\beta \in [0,1[$ tal que $\|x_i - x\| \leq \eta\beta^i$. Fixe qualquer $j_0 \geq 0$, então, para todo $k = 0, 1, \ldots$

$$\|x_{j_0+k} - x\| \leq \eta\beta^{j_0+k} = (\eta\beta^k)\beta^{j_0} \ .$$

A sequência $\left\{\eta\beta^k\right\}_{k\geq 0} \to 0$, portanto, é uma sequência de Cauchy. Assim, pode-se escolher qualquer $\epsilon > 0$ que existirá um $k_0 > 0$, tal que $\eta\beta^{k_0} < \epsilon$ para todo $k \geq k_0$. Escolha $\epsilon = \|x_{j_0} - x\|$, então

$$\|x_{j_0+k} - x\| \leq (\eta\beta^k)\beta^{j_0} < \beta^{j_0} \|x_{j_0} - x\| \ , \ k \geq k_0.$$

Continuando o processo, escolha $j_1 = j_0 + k_0$ e por conseguinte $\|x_{j_1} - x\| < \|x_{j_0} - x\|$. Partindo de j_1, repetindo a escolha, e procedendo por indução tem-se

$$\cdots < \|x_{j_2} - x\| < \|x_{j_1} - x\| < \|x_{j_0} - x\| \, . \tag{3.1}$$

Para cada elemento inicial j_0 pode-se arbitrar ϵ e assim, pelo menos finitamente, muitas sequências existem para escolha, e a prova está completa. ■

Pode-se agora aplicar a Proposição 28 à sequência (3.1), resultando no seguinte:

Proposição 30 *Se uma sequência $\{x_i\}_{i\geq 0}$, $x_i \in \mathbb{R}^n$, possui convergência r-linear, então existe pelo menos finitamente muitas sequências $\{p_k\}_{k\geq 0}$ não vazias tais que a sequência $\{x_i\}_{i\geq 0}$ converge q-linearmente em p_k-etapas.*

3.3 Convergência q-linear

3.3.1 O Lema fundamental

O resultado básico da convergência do quociente é o seguinte lema:

Lema 31 (Lema fundamental) *Para a sequência $x_{i+1} = \mathcal{A}(x_i)$, $i = 0, 1, 2, \ldots$, q-linearmente convergente para o seu limite x_*, com fator-q dado por α , valem as seguintes cotas para todos os índices $i \geq i_\alpha$:*

$$\frac{d(x_{i+1}, x_i)}{1+\alpha} \leq d(x_i, x_*) \leq \frac{d(x_{i+1}, x_i)}{1-\alpha} \tag{3.2}$$

$$d(x_{i+1}, x_*) \leq \frac{\alpha}{1-\alpha} d(x_{i+1}, x_i) = \frac{\zeta}{1+\alpha} d(x_{i+1}, x_i) \tag{3.3}$$

onde $\zeta = \zeta(\alpha) = \alpha(1+\alpha)/(1-\alpha)$.

Demonstração: Imediata a partir da desigualdade (métrica) triangular e da simetria:

$$d(x_i, x_*) \leq d(x_i, x_{i+1}) + d(x_{i+1}, x_*) \leq d(x_i, x_{i+1}) + \alpha d(x_i, x_*),$$

$$d(x_i, x_{i+1}) \leq d(x_i, x_*) + d(x_*, x_{i+1}) \leq d(x_i, x_*) + \alpha d(x_i, x_*),$$

resultando em (3.2). Usando-as,

$$d(x_{i+1}, x_*) \leq \alpha d(x_i, x_*) \leq \frac{\alpha}{1-\alpha} d(x_{i+1}, x_i) = \frac{\zeta}{1+\alpha} d(x_{i+1}, x_i).$$

e os resultados estão provados. ■

Este resultado não exige que $\mathcal{A}$ possua ponto fixo, ou até mesmo que seja contínuo. A função $\zeta(\alpha)$ em (3.3) assumirá um papel relevante neste texto.

Computacionalmente, há uma diferença fundamental entre as desigualdades (3.2) e (3.3). A primeira implica em um custo para o enriquecimento da solução, no sentido em que sua utilização deve esperar pelo próximo incremento $s_i = x_{i+1} - x_i$, enquanto a outra vem a custo zero. Note, porém, que em muitos algoritmos, mesmo o custo de (3.3) tem complexidade $O(n)$, ou mesmo $O(1)$. Estes são os casos de algoritmos de curta recorrência, como toda a classe do CG Generalizado (Hermitiano e não Hermitiano).

É imediato que o Lema 31 vale para sequências q-linearmente convergentes em p_k-etapas, por simples substituição dos índices: $i \leftarrow p_k$ e $i+1 \leftarrow p_{k+1}$. Como consequência, aplicável também a sequências com convergência q-sublinear estrita e r-linear.

3.3.2 A geometria do processo q-linear

Não se pode deixar de notar pelo menos a semelhança formal entre a estimativa (3.3) e a contrapartida do Teorema de Banach (Teoremas 14 e 15), quando $\mathcal{A}$ mapeia D_0

nele próprio. De fato, o principal objetivo desta subseção é estabelecer com precisão a natural e significativa relação entre uma teoria de convergência do quociente e a teoria das contrações, à parte dos diferentes objetivos: de um lado qualidade de convergência e estimativas de erros, e do outro existência, unicidade e convergência. Para a teoria métrica dos pontos fixos, o trabalho termina com a segunda. Para a computação científica, o trabalho apenas começou. E, deve-se julgar muito feliz se todos os resultados da teoria métrica dos pontos fixos puderem ser aproveitados, e em seguida os mecanismos e estimativas fornecidos pela teoria de convergência do quociente possam ser traduzidos em álgebra linear computacional eficiente.

Álgebra linear computacional eficiente é o segundo principal motivo pelo qual a escolha recai sobre a convergência do quociente sob as formas contidas na Definição 26.

A primeira preocupação é estabelecer que, de fato, o conjunto dos algoritmos de convergência q-linear não foi esgotado pelo estudo das contrações da teoria métrica dos pontos fixos, no sentido em que as condições de convergência do quociente e as suas estimativas de erro já não estão todas lá, ou que as cotas para os erros daquela teoria são as melhores possíveis. Para responder estas questões, deve-se, inicialmente, fazer uma distinção entre *processo* e *realização*.

Dado uma aproximação inicial x_0 pertencente a um conjunto arbitrário $D \subset \mathbb{R}^n$, uma *realização* do algoritmo é definida como a sequência $x_{i+1} = \mathcal{A}(x_i)$, $i = 0, 1, \ldots$, cuja única exigência é que x_0 pertença a D. Será denominado de *processo* ao conjunto de todas as realizações do algoritmo iniciados com um elemento de D. Com estas definições pode-se distinguir a convergência, e a qualidade de convergência de uma realização e de um processo em uma métrica $d(\cdot,\cdot)$. Uma realização é convergente (q-convergente) em D se satisfaz a definição 26 e todos os seus elementos x_i estão contidos em D. Um processo é convergente (q-convergente) em D se todas as realizações são convergentes (q-convergentes) em D. Um algoritmo $\mathcal{A}$ será dito convergente (q-convergente) em D se o seu processo for convergente (q-convergente) em D. Se o conjunto D é o próprio $\mathbb{R}^n$, o processo, e assim o algoritmo, será dito globalmente convergente (q-convergente).

No que se segue são feitas as seguintes hipóteses:

h1. $d(\cdot,\cdot)$ é uma métrica em $\mathbb{R}^n$,

h2. $\mathcal{A} : D \to D \subset \mathbb{R}^n$, D arbitrário,

h3. $\mathcal{A}$ é q-linear em $D_0 \subset D$, com $\alpha \in [0, 1[$, na métrica $d(\cdot,\cdot)$,

h4. $\mathcal{A}$ é contínuo em cada um dos limites das realizações.

Observe-se que não é feita muita referência às qualidades topológicas do conjunto D_0. Em geral, entretanto, para estabelecer comparações com as contrações pode-se exigir que D_0 seja fechado, e eventualmente limitado, ou mesmo convexo. Mapeamentos q-lineares não precisam ser contínuos. A imposição de continuidade (h4) acima visa estabelecer a existência de pontos fixos para cada realização e para o processo.

Sendo α_r o fator-q de uma realização r em D_0, o fator-q do processo será definido por $\alpha = \max_{i \geq i_\alpha} \{\alpha_r\}$. Isto posto, a seguinte proposição é contundente.

Proposição 32 *Sob as hipóteses h1 – h4, um ponto fixo de $\mathcal{A}$ existe e é único.*

Demonstração: Duas sequências distintas, partindo de $x_0, y_0 \in D_0$, $x_0 \neq y_0$, possuem seus limites $x_\star$ e $y_\star$, respectivamente, uma vez que as realizações são convergentes. E também como $\mathcal{A}$ é suposto contínuo em $x_\star$ e $y_\star$, tem-se

$$\begin{aligned} x_\star &= \lim x_i = \lim x_{i+1} = \lim \mathcal{A}(x_i) = \mathcal{A}(x_\star), \\ y_\star &= \lim y_i = \lim y_{i+1} = \lim \mathcal{A}(y_i) = \mathcal{A}(y_\star), \end{aligned}$$

ou seja, $x_\star = \mathcal{A}(x_*)$ e $y_\star = \mathcal{A}(y_*)$. Para provar que os pontos fixos são únicos, sejam α_x e α_y os fatores-q de cada uma das realizações, então

$$\begin{aligned} d(x_\star, y_\star) &= d(\mathcal{A}(x_\star), y_\star) \leq \alpha_x d(x_\star, y_\star) \\ d(x_\star, y_\star) &= d(x_\star, \mathcal{A}(y_\star)) \leq \alpha_y d(x_\star, y_\star) \end{aligned}$$

e, como $\alpha_x, \alpha_y \in [0, 1[$, segue-se que

$$d(x_*, y_*) = 0 \Rightarrow x_\star = y_\star$$

porque $d(\cdot, \cdot)$ é uma métrica. ■

Este resultado diz que a hipótese de q-linearidade em D_0, ou seja, a convergência de cada realização, garante a existência do ponto fixo de cada uma, e ao mesmo tempo, a unicidade. Esta definição de q-linearidade do processo é bem mais abrangente do que a q-linearidade de uma realização. Ou seja, a q-linearidade de uma realização, embora garanta o ponto fixo, se o mapeamento for contínuo no limite da sequência, não garante a q-linearidade de todas as realizações. Porém, se todas possuem um ponto fixo, ele é único. Isto não ocorre se o processo é q-sublinear

($\alpha_r = 1$ em alguma realização) . Mas, claramente, ocorre se o processo é q-sublinear estrito, já que a Proposição 27 garante a existência de um mapeamento composto equivalente que converge q-linearmente, e um tal mapeamento satisfaz a Proposição 32.

A unicidade do ponto fixo do processo q-linear (e a prova da unicidade), diz que a hipótese $h4$ é parcialmente redundante, e deve ser lida como: $\mathcal{A}$ é contínuo no limite, único, das realizações.

Pode-se provar a Proposição 32 de várias formas, em particular pela mesma técnica usada na prova do Teorema 14. Ela é muito instrutiva. A q-linearidade implica na existência de uma família de conjuntos encaixados, fechados e não vazios $D_\alpha(x_0)$ para cada realização identificada pelo ponto de partida $x_0 \in D_0$ dada por

$$D_\alpha(x_0) = \left\{ x \in D : d(\mathcal{A}(x), x_*) \leq \alpha^i d(x_0, x_*) \right\} . \tag{3.4}$$

É imediato que lim diam $D_\alpha(x_0) = 0$. Então, cada família D_α retrai-se para um ponto x_*, intersecção única de todos os conjuntos da família. O marcante é que todas as famílias $D_\alpha(x), x \in D_0$ interceptam-se no mesmo ponto $x_\star$.

Portanto, um mapeamento q-linear é capaz de "criar " não apenas uma família de conjuntos fechados e não vazios, porém uma família de família de subconjuntos retráteis, fechados e não vazios, todas com a mesma intersecção $x_\star$. Neste sentido, pode-se prescindir da hipótese de $A(D_0) \subset D_0$, e usar apenas $A(D) \subset D$, estando as famílias em D. A condição $A(D_0) \subset D_0$ impõe uma restrição a mais sobre o mapeamento que pode mudar o ponto fixo único $z_\star$, ou mesmo torná-lo não único, a depender de, por exemplo, das características da métrica. Provavelmente, em uma métrica elíptica (uma norma elíptica) o ponto fixo, embora possa mudar, continuará único (o autor não tentou provar isto). Evidentemente, se a não unicidade do ponto fixo ocorrer com a escolha inadequada de D_0, o processo, pela definição dada, não será mais q-linear. Neste caso, normalmente, é simples especificar uma restrição adicional, tipo mínimos quadrados por exemplo, devolvendo a unicidade ao ponto fixo.

Assim, ao se dizer que $\mathcal{A}$ é q-linear em D_0, com $A(D_0) \subset D_0$, implica que cada $x_i \in D_0$ e z_* é o único ponto fixo de cada realização, e portanto do processo.

A próxima questão importante são as relações entre a hipótese de q-linearidade e a constante de Lipschitz.

Partindo de quaisquer dois pontos distintos $x_0, y_0 \in D_0$ tem-se, para todo $i \geq 0$,

$$d(x_{i+1}, x_*) \leq \alpha d(x_i, x_*), \tag{3.5}$$

$$d(y_{i+1}, y_\star) \leq \alpha d(y_i, y_\star), \tag{3.6}$$

onde, inicialmente, supõe-se que $x_\star \neq y_\star$. Dado que $\mathcal{A}$ seja Lipschitz-contínuo em D_0 com $\gamma \geq 0$, também valem, para todo $i \geq 0$,

$$d(x_{i+1}, y_{i+1}) \leq \gamma d(x_i, y_i), \tag{3.7}$$

e, para o caso de $x = x_\star$ e $y = y_\star$, as (3.5) e (3.6) dizem que, para todo $i \geq 0$,

$$d(x_{i+1}, x_\star) \leq \gamma d(x_i, x_\star), \tag{3.8}$$

$$d(y_{i+1}, y_\star) \leq \gamma d(y_i, y_\star). \tag{3.9}$$

Naturalmente, não podendo haver conflitos entre as (3.5) e (3.6) e as (3.8) e (3.9), resulta que, pela definição de q-linearidade do processo, γ não pode ser menor do que α, senão $\alpha = \gamma$. Conclui-se o resultado:

Lema 33 *($\mathcal{A}$ **Lipschitz-contínuo**): $\alpha \leq \gamma$ para todo $\gamma \geq 0$.*

É importante notar que não é necessário impor a unicidade do ponto fixo. Isto parece clamar por uma definição menos restritiva de processo q-linear - um processo q-sublinear definido de forma análoga.

O Lema 33 diz que $\gamma < 1$ (contração) é uma cota superior para α. A desigualdade (3.7) é uma restrição a mais no processo q-linear, que resulta em $\alpha \leq \gamma$. A q-linearidade não exige continuidade de Lipschitz. Aparentemente, isto também indica que podemos relaxar ainda mais a definição de q-linearidade do processo. Isto não será feito aqui, e accita-se a Proposição 32 no restante do texto.

Para cada dois pares de sequências do processo, $\{x_i\}_{i\geq 0}$ e $\{y_i\}_{i\geq 0}$, $x_0, y_0 \in D_0$, tem-se

$$d(x_i, y_i) \leq d(x_i, z_\star) + d(y_i, z_\star), \tag{3.10}$$

$$d(x_{i+1}, y_{i+1}) \leq d(x_{i+1}, z_\star) + d(y_{i+1}, z_\star) \leq \alpha \left[d(x_i, z_\star) + d(y_i, z_\star)\right], \tag{3.11}$$

$$d(x_{i+1}, y_{i+1}) \leq d(x_{i+1}, z_\star) + d(y_{i+1}, z_\star) \leq \gamma \left[d(x_i, z_\star) + d(y_i, z_\star)\right]. \tag{3.12}$$

Defina, para x_i e y_i não simultaneamente iguais a $z_\star$,

$$\delta_i = \frac{d(x_i, y_i)}{d(x_i, z_\star) + d(y_i, z_\star)} \tag{3.13}$$

onde, pela desigualdade triangular, $0 \leq \delta_i \leq 1$, para todo $i \geq 0$, com $\delta_i = 0$ se, e somente se, $x_i = y_i$.

Exprimindo $d(x_i, y_i)$ pela menor trajetória em direção à solução que também contém $d(x_{i+1}, y_{i+1})$, e usando o Lema 31,

$$\begin{aligned} d(x_i, y_i) &\leq d(x_i, x_{i+1}) + d(x_{i+1}, y_{i+1}) + d(y_{i+1}, y_i) \\ &\leq (1+\alpha)\,[d(x_i, z_*) + d(y_i, z_*)] + d(x_{i+1}, y_{i+1}) \\ &\leq (1+\alpha)\,[d(x_i, z_*) + d(y_i, z_*)] + \gamma d(x_i, y_i) \end{aligned}$$

que implica em

$$(1-\gamma)\delta_i \leq 1+\alpha, \tag{3.14}$$

que aceita qualquer valor $\gamma \geq 0$. Ou seja, de fato pode existir um processo q-linear para qualquer mapeamento Lipschitz-contínuo. Note também que (3.14) aceita $\gamma \to \infty$.

Porém, usando a trajetória ainda mais curta dada por (3.11) e (3.12) e o Lema 33, obtém-se

$$\gamma_i \delta_i \leq \alpha \leq \gamma \tag{3.15}$$

$$\gamma_i \delta_i \leq \gamma \tag{3.16}$$

onde a última desigualdade de (3.15) vem do Lema 33, e foi definido

$$\gamma_i = \frac{d(x_{i+1}, y_{i+1})}{d(x_i, y_i)}\ , \ x_i \neq y_i\ , \tag{3.17}$$

e, também, $\gamma = \max\limits_{i \geq i_\alpha} \{\gamma_i\}$. Assim, a aceitação de $\gamma \to \infty$ é limitada à condição $\lim \{\gamma_i \delta_i\} \leq \alpha < 1$.

Convém analisar cada tipo de contração separadamente. O término da análise de cada caso será indicado pelo duplo símbolo: ■■.

Caso 1: *Contração.* Existe um $\gamma < 1$ tal que $d(x_{i+1}, y_{i+1}) \leq \gamma d(x_i, y_i)$, para qualquer $i \geq 0$, e para todos os possíveis pares $\{x_i\}_{i\geq 0}$ e $\{y_i\}_{i\geq 0}$, $x_0, y_0 \in D_0$, do processo. Também, que $\mathcal{A}(D_0)$ mapeia D_0 nele próprio.

Lema 34 ***(Contração,** $\mathcal{A}(D_0) \subset D_0$**):*** *Se existe ao menos um índice i no processo, para o qual $\gamma_i \delta_i = \gamma$, então: a) $\delta_i = 1$, $\gamma_i = \gamma$, e, b) $\alpha = \gamma < 1$.*

Demonstração: Como $\gamma = \max\limits_{i \geq i_\alpha} \{\gamma_i\}$, então $\gamma_i \leq \gamma$ para toda e qualquer iteração do processo. Seque-se que $\gamma_i \delta_i = \gamma$ se, e somente se, existe um $\delta_i = 1$ em alguma realização, e lá ocorra $\gamma_i = \gamma$, simultaneamente. Então, (3.15) diz que neste índice

desta realização $\gamma_i\delta_i = \gamma \leq \alpha \leq \gamma$, ou seja $\alpha = \gamma$. ■

O Lema 34 vale quando $\delta_i = 1$ e $\gamma_i = \gamma$ ocorrem simultaneamente. Pode-se vislumbrar dois casos de importância.

O primeiro é o trivial, ou seja, $y_i = z_\star \Rightarrow \delta_i = 1$. Neste caso estamos analisando a própria realização na convergência (neste caso finita), e $\gamma_i = d(x_{i+1}, z_\star)/d(x_i, z_\star) \leq \alpha$. Para retirar mais informações convém reescrever: $\gamma_i = d(x_{i+2}, x_{i+1})/d(x_{i+1}, x_i)$ e $\delta_i = d(x_{i+1}, x_i)/[d(x_i, z_\star) + d(x_{i+1}, z_\star)]$, e enquadrar a análise no segundo caso, a seguir. De fato, isto será objeto da próxima subseção. De passagem, note que se $x_{i+1} = z_\star$, então $\delta_i = 1$, e $\gamma_i = 0$, tal que $\lim(\delta_i\gamma_i) = 0$. Ver-se-á, na próxima subseção que isto implica em convergência finita com $\alpha = \gamma = 0$, ou assintótica, caracterizando, em ambos os casos, uma convergência q-superlinear da realização.

Por outro lado, sendo x_i e y_i dois elementos quaisquer de D_0, parece mais natural indagar quais conjuntos D_0 admitem $\delta_i = 1$ em (3.13), com x_i e y_i dois elementos quaisquer de D_0. Se o processo ocorre em tais conjuntos, então, para cada x_i de cada realização, existirá um y_i tal que $\delta_i = 1$. Neste caso, $\gamma_i = \gamma$ ocorrerá sempre com $\delta_i = 1$, e o Lema 34 diz então $\alpha = \gamma < 1$. Pode-se portanto enunciar:

Lema 35 ***(Contração, $\mathcal{A}(D_0) \subset D_0$)*** *Se um conjunto admite $\delta_i = 1$ para qualquer i, então $\alpha = \gamma < 1$.*

Um espaço métrico M com métrica d é dito metricamente convexo se, e somente, se para cada par de pontos $x, y \in M$, $x \neq y$, existe um ponto $z \in M$, $x \neq z \neq y$, tal que $d(x, y) = d(x, z) + d(z, y)$ [26, pag. 23]. Neste caso diz-se que z está entre x e y, ou está na linha entre x e y, e este fato é denotado pelo símbolo (xzy). Considere-se o importante caso em que $D_0 \subset M$, D_0 metricamente convexo.

Neles, dado $x_i \in D_0$, $\delta_i = 1$ significa que existe pelo menos um ponto y_i no qual $(x_i z_\star y_i)$, se $z_\star$ está no $\text{int}D_0$ (o interior de D_0). Neste caso, pelo Lema 35, $\alpha = \gamma < 1$. Pode-se portanto enunciar:

Proposição 36 ***(Contração em um conjunto convexo)*** *Suponha que $\mathcal{A} : D \subset \mathbb{R}^n \to \mathbb{R}^n$ mapeia um conjunto convexo e fechado $D_0 \subset D$ nele próprio. Suponha que $\mathcal{A}$ tenha um ponto fixo $x_\star \in intD_0$. Então as seguintes afirmações são equivalentes,*

1. *$\mathcal{A}$ é uma contração em D_0 com constante de Lipschitz γ_d,*

2. *$\mathcal{A}$ é q-linear em D_0 com fator-q dado por γ_d.*

Comentário 1 *A extensão para espaços métricos sequencialmente completos e convexos é imediata: Seja (M, d) um espaço métrico completo e metricamente convexo, e $\mathcal{A}: M \to M$ um mapeamento com ponto fixo em M, e $\mathcal{A}$ mapeia M nele próprio. Então, na hipótese de a sequência $(\mathcal{A}^i(x_0))_{i\geq 0}$ convergir para este ponto fixo, as seguintes afirmações são equivalentes: 1) $\mathcal{A}$ é uma contração em D_0 com constante de Lipschitz $\gamma_d < 1$, 2) $\mathcal{A}$ é q-linear em D_0 com fator-q dado por γ_d. A propósito, vale lembrar que um subconjunto de um espaço de Banach é também um espaço de Banach se, e somente se, é fechado.*

Aparentemente, não se pode eliminar a condição de que o ponto fixo seja interior ao conjunto, a não ser que o mapeamento tenha convergência finita, como já observado. O fato de $\mathcal{A}$ possuir convergência finita não é uma anormalidade algorítmica. De fato, todos os bons algoritmos iterativos para sistemas lineares possuem convergência finita com $m \leq n$ iterações, ou mesmo $m > n$. Porém, convergência finita é possível apenas em operadores algébricos.

Um caso especial extremamente importante consiste dos algoritmos de projeção ortogonal em espaços de Hilbert. Posteriormente será mostrado que esses algoritmos fazem seus iterandos viajarem na superfície folhada de um *icecream* com vértice na solução, e cujas arestas são os erros dos iterandos. Para estes métodos, se não houver *breakdown*, a convergência do quociente do erro está garantida. Com um pouco mais de exigência obtem-se convergência q-sublinear estrita ou mesmo q-linear e q-superlinear. Entre eles toda a classe dos Gradientes Conjugados Generalizados para problemas lineares Hermitianos e não Hermitianos, o CG para problemas lineares e não lineares com Jacobiano ou Hessiana positivo definidos, o GMRES, etc. Também, e não menos importante, métodos clássicos como o método de Ritz - um método de projeção ortogonal em espaços de Hilbert - e por consequência algumas formulações do método dos elementos finitos. Também, como se verá, as desejadas estreitas cotas, superior e inferior, para o erro podem ser analiticamente estabelecidas em função, unicamente, das informações extraídas das iterações. ■■

Caso 2: *Contração.* Sem a condição que $\mathcal{A}(D_0)$ mapeia D_0 nele próprio.

Sem a hipótese de que $\mathcal{A}(D_0)$ mapeia D_0 nele próprio, a teoria métrica dos pontos fixos garante apenas convergência r-linear (a não ser sob as condições do Teorema 23). Não há garantias da validade das estimativas do Lema 31 com a substituição $\alpha \leftarrow \gamma$. Porém, vale ainda a hipótese da q-linearidade com α, bem como o Lema 33

Proposição 37 *Suponha que $\mathcal{A} : D \subset \mathbb{R}^n \to \mathbb{R}^n$ é uma contração em um conjunto fechado $D_0 \subset D$. Suponha também que $\mathcal{A}$ é q-linear em D_0, então $\alpha \leq \gamma < 1$.*

Este resultado estende-se naturalmente a espaços métricos na linha do Comentário 1. A importância do exposto é que ainda é possível usar a constante de Lipschitz, agora como uma cota superior para α nas estimativas de erro. Pela forma das estimativas de erro dadas no Lema 31, elas poderão ser uma boa aproximação para o erro, pelo menos assintoticamente. ■■

Caso 3: *Não expansividade estrita.* $d(x_{i+1}, y_{i+1}) < d(x_i, y_i)$, para qualquer $i \geq 0$.

As desigualdades (3.15) e (3.16) passam a ser

$$\gamma_i \delta_i \leq \alpha < 1 \tag{3.18}$$

$$\gamma_i \delta_i < 1 \tag{3.19}$$

Lema 38 ***(Não expansividade estrita):*** *Se $\delta_i = 1$ para todo i, então* $\max_{i \geq i_\alpha} \{\gamma_i\} \leq \alpha$.

Em outros casos, como por exemplo aquele no qual $\max_{i \geq i_\alpha} \{\gamma_i\}$ ocorra para algum $\delta_i < 1$, γ_i deve oscilar em torno de α. Este é o caso, por exemplo, do CG na norma-ℓ_2.

Proposição 39 *Suponha que $\mathcal{A} : D \subset \mathbb{R}^n \to \mathbb{R}^n$ é estritamente não expansivo em um conjunto convexo e fechado $D_0 \subset D$. Suponha também que $\mathcal{A}$ é q-linear em D_0, e que o ponto fixo $x_\star \in intD_0$, então* $\max_{i \geq i_\alpha} \{\gamma_i\} \leq \alpha$.

É possível usar o fato de que expansividade estrita convergente, com ponto fixo, implica em pelo menos convergência q-linear em 2-etapas: fazendo $y_i = z_\star$, temos que $d(x_{i+1}, z_\star) < d(x_i, z_\star)$, e usando a Proposição 27. Então,

$$\begin{aligned} d(x_{i+2}, y_{i+2}) &\leq d(x_{i+2}, z_\star) + d(y_{i+2}, z_\star) \leq \alpha^2 \left[d(x_i, z_\star) + d(y_i, z_\star) \right] \\ d(x_{i+2}, y_{i+2}) &\leq d(x_{i+2}, z_\star) + d(y_{i+2}, z_\star) \leq d(x_i, z_\star) + d(y_i, z_\star) \end{aligned}$$

o que resulta em

$$\widehat{\gamma}_i \delta_i \leq \alpha^2 \tag{3.20}$$

$$\widehat{\gamma}_i \delta_i < 1 \tag{3.21}$$

onde,

$$\widehat{\gamma}_i = \frac{d(x_{i+2}, y_{i+2})}{d(x_i, y_i)} \ , \ x_i \neq y_i \ . \tag{3.22}$$

Proposição 40 *Suponha que $\mathcal{A} : D \subset \mathbb{R}^n \to \mathbb{R}^n$ é estritamente não expansivo em um conjunto convexo e fechado $D_0 \subset D$. Suponha também que $\mathcal{A}$ é q-linear em D_0, e que o ponto fixo $x_\star \in intD_0$, então $\max_{i \geq i_\alpha} \sqrt{\widehat{\gamma}_i} \leq \alpha$.*

As Proposições 39 e 40 devem ser comparadas com o Teorema de Edelstein. Mostra-se assim que não apenas o ponto fixo do Teorema de Edelstein existe, como é possível um processo q-linear sob não expansividade estrita.

Proposição 41 *Nas hipóteses do Teorema de Edelstein, o mapeamento $\mathcal{A}^2$ é uma contração, com $\gamma = \max_{i \geq i_\alpha} \{\gamma_i\} = \alpha$, e portanto com estimativa de erro dada pelo Lema 31.*

Demonstração: Que $\mathcal{A}^2$ é uma contração basta notar que

$$d\left(\mathcal{A}^2(x), \mathcal{A}^2(y)\right) < d\left(\mathcal{A}(x), \mathcal{A}(y)\right) < d(x, y) \ ,$$

e usar a técnica usada na prova da Proposição 27, substituindo α_i por γ_i, e, então $\gamma = \max_{i \geq i_\alpha} \{\gamma_i\} < 1$. Como, pelo Teorema de Edelstein o ponto fixo existe e é único, segue-se pela análise do Caso 1 que $\alpha = \gamma$. ■

Toda a análise dos Casos 1 e 2 aplicam-se ao mapeamento composto $\mathcal{A}^2$. É importante observar que a Proposição 41 vale para qualquer mapeamento $\mathcal{A}^p$, $p \geq 2$, ou mesmo qualquer mapeamento composto $\mathcal{B} = \mathcal{A}^p \cdots \mathcal{A}^q \mathcal{A}^r$, $p, \ldots, q, r \geq 2$. Isto formará a base para estimativas de algoritmos com não expansividade estrita, e mesmo apenas não expansivos, mas tal que $\mathcal{A}$ não seja q-linear. Está em consonância com a Proposição 16.

Exemplo 42 *O mapeamento $\mathcal{A} : \mathbb{R} \to \mathbb{R}$ dado por $\mathcal{A}(x) = 1 + \ln(1 + e^x)$ é estritamente não expansivo, $|\mathcal{A}(x) - x| > 1$ para todo $x \in \mathbb{R}$, e não possui ponto fixo. Porém, para qualquer intervalo fechado (portanto compacto) $D = [a, b]$, $a, b \in \mathbb{R}$, $\mathcal{A} : D \subset \mathbb{R} \to \mathbb{R}$ é estritamente q-sublinearmente convergente, portanto é q-linear em duas etapas (globalmente), e assim $\mathcal{A}$ possui um ponto fixo único localizado na fronteira de D, $Fix(\mathcal{A}) = b$, para qualquer $x_0 \in D$. Ainda mais, existe um i_α finito, a partir do qual a Proposição 41 vale. (Para evitar provas analíticas o leitor pode construir um programa MATLAB, por ex., e verificar numericamente). Os resultados desta subseção são fundamentais para a prova (por exemplo) de que o método dos gradientes conjugados converge pelo menos q-linearmente em duas etapas na norma-ℓ_2 do erro (embora minimize a norma da energia) e construir as melhores*

cotas superiores e inferiores para o erro nestas normas (que não podem ser melhoradas), e que independem do condicionamento dos operadores (ou de qualquer conhecimento prévio de suas propriedades) . ■■

Caso 4: *Não expansividade.* $d(x_{i+1}, y_{i+1}) \leq d(x_i, y_i)$, $(\gamma = 1)$,para qualquer $i \geq 0$.

As desigualdades (3.15) e (3.16) passam a ser

$$\gamma_i \delta_i \leq \alpha \leq \gamma = 1 \tag{3.23}$$

$$\gamma_i \delta_i \leq 1 \tag{3.24}$$

e, também, $\max\limits_{i \geq i_\alpha} \{\gamma_i\} = \gamma = 1$. Se, para cada x_i existe um y_i tal que $\delta_i = 1$, não pode haver um processo q-linear:

$$\delta_i = 1 \Rightarrow \max_{i \geq i_\alpha} \{\gamma_i \delta_i\} = 1 \leq \alpha \leq \gamma = 1 \Rightarrow \alpha = 1.$$

Lema 43 ***(Não expansivo)*** *Se* $\delta_i = 1$ *para todo* i, *então* $\alpha = 1$.

A única possibilidade é se $\max\limits_{i \geq i_\alpha} \{\gamma_i \delta_i\} < 1$. O caso de interesse está quando D_0 é metricamente convexo e fechado. Neste caso, o processo q-linear somente pode ocorrer com a solução na fronteira de D_0. Ainda mais, o conjunto tem que ter um vértice na solução - $(x_i z_* y_i)$ não pode existir - e z_* deve estar neste vértice. O valor de γ_i oscilará em torno de α. Não estando a solução em um vértice, mas em um contorno suave, teremos que $\delta = \max\limits_{i \geq i_\alpha} \{\delta_i\} = 1$, e o processo não pode ser q-linear.

Proposição 44 *Mapeamentos não expansivos admitem processos q-lineares se, e somente, se o conjunto* $\mathcal{A}(D_0) \subset D_0$ *possua* $\delta = \max\limits_{i \geq i_\alpha} \{\delta_i\} < 1$. *Em caso contrário, o processo será, na melhor das hipóteses, q-sublinear, e, se* $\mathcal{A}$ *possue um ponto fixo, ele pode não ser único.*

Para ver a não unicidade do ponto fixo, observe-se a segunda parte da prova da Proposição 32, com $\alpha_x = \alpha_y = 1$.

É imediato que conjuntos $\mathcal{A}(D_0)$ metricamente convexos apenas admitem processos não expansivos e q-lineares, simultaneamente, se existe em $\mathcal{A}(D_0)$ um vértice, e o limite do processo está lá. Um exemplo extremamente importante em espaços de Hilbert é quando $\mathcal{A}(D_0)$ é um cone convexo com origem na solução. ■■

É aparente que o fator-q do processo é, em geral, (muito) possivelmente maior do que o fator-q de uma dada realização. Isto justifica que, se boas estimativas de erro

devem ser obtidas, elas *devem ser feitas durante a realização (a posteriori)*. Esta conclusão é aparentemente compartilhada, de forma intuitiva, por Golub e Van der Vorst [13], e foi usada pelo autor sistematicamente, porém ainda de forma preliminar, já em [34]. Usar apenas os valores teóricos da constante de Lipschitz é equivalente a trabalhar com estimativas de normas de operadores, ou pior, de números de condição de operadores, e assim o preço a pagar será certamente estimativas pobres.

Comentário 2 *Um algoritmo iniciado em dois pontos distintos, x_0 e y_0, na linha $x_i z_* y_i$, em geral não converge ao longo desta linha, a não ser que esta linha seja especial. Por exemplo, o CG e o steepest descent em sistemas lineares (equivalentemente em otimização quadrática com ou sem restrições) convergem (em uma iteração) se a linha for um autovetor. A proximidade de x_0 desta linha (infelizmente) não significa em geral que a convergência deva ser mais rápida. Por exemplo, o steepest descent converge rapidamente se a linha for um autovetor com grande autovalor, e tem a suas piores velocidades de convergência quando a proximidade se dá com autovetores associados a pequenos autovalores. Que o steepest descent converge global e q-linearmente e é contractivo em $\mathbb{R}^n$ na norma da energia, é um fato bem conhecido. Neste capítulo, com base nos fatos estabelecidos nesta subseção, provar-se-á resultado similar para o CG, que pelo tanto que se saiba, não consta da literatura.*

Comentário 3 *Todo o desenvolvimento da teoria métrica dos pontos fixos posterior a Banach indica uma grande concentração de esforços na análise de mapeamentos não expansivos, e pouco interesse (se há algum) por mapeamentos expansivos. De fato, este último termo não existe na literatura. Porém, algoritmos que não resultam de projeção ortogonal, ou que não estão associados a algum processo de minimização, não raramente são expansivos. O BiCG, e seu derivado, o CGS original para sistemas lineares, são muito usados em inner-loops de sistemas não lineares pela sua eficiência e são expansivos na norma Euclidiana dos resíduos - justamente o que se quer reduzir nestes problemas. O método de Newton modificado (manutenção do Jacobiano, ou Hessiana, constante durante algumas iterações) também é expansivo na mesma norma. Na próxima subseção será demonstrado que o CG, embora expansivo na norma da Energia, converge nesta norma q-linearmente e q-superlinearmente (local e assintóticamente). O CG também converge na norma Euclidiana do erro de forma pelo menos q-sublinear estrita, e no entanto, também é expansivo nesta norma. Parece evidente ao autor desta tese que mapeamentos expansivos são suficientemente flexíveis e poderosos para comportarem adequadamente processos pelo menos q-lineares.*

3.3.3 Geometria e estimativas da realização

Esta subseção examina a sequência isolada $x_{i+1} = \mathcal{A}(x_i), i = 0, 1, \ldots$, na hipótese de que ela converge q-linearmente. Conforme estabelecido na subseção anterior, o ponto fixo de $\mathcal{A}$ está garantido se ele é contínuo no limite $x_\star$ da sequência.

Todas as variáveis desta subseção referem-se à realização. Para evitar proliferação de índices, a indicação explícita da realização é omitida. Assim, α_r será grafado como α, etc. Quando for necessário usar valores do processo, isto será explicitado.

Dado um mapeamento $\mathcal{A} : D \subset \mathbb{R}^n \to \mathbb{R}^n$, raramente ele converge q-linearmente com $i_\alpha = 0$, para todo $x \in D_0$, a não ser por um D_0 que reúna pontos bastante próximos da solução. Supõe-se a existência de um $i_\alpha \geq 0$ a partir do qual a sequência isolada converge q-linearmente com fator-q dado por α.

Valem as seguintes definições:

$$\begin{aligned}
\beta_i &= d(x_{i+2}, x_{i+1})/d(x_{i+1}, x_i), \\
\gamma &= \max_{i \geq i_\alpha} \{\beta_i\}, \\
\delta_i &= d(x_{i+1}, x_i)/\left[d(x_{i+1}, x_\star) + d(x_i, x_\star)\right], \\
\delta &= \max_{i \geq i_\alpha} \{\delta_i\}, \\
\xi_i &= d(x_{i+1}, x_\star)/d(x_i, x_\star), \\
\eta_i &= d(x_{i+1}, x_\star)/d(x_{i+1}, x_i),
\end{aligned}$$

e supõe-se que não há *breakdown* - $d(x_{i+1}, x_i) \neq 0$ enquanto $x_{i+1} \neq x_\star$. Desta forma, β_i somente será nulo se $x_{i+1} = x_\star$ for um ponto fixo do algoritmo. Também, $0 < \delta_i \leq \delta \leq 1$.

Pode-se distinguir dois casos para os quais $\delta = 1$:

1. Quando há convergência finita: $d(x_{i+1}, x_\star) = d(x_\star, x_\star) = 0$, ou
2. Existe $(x_i x_\star x_{i+1})$:
   - 2.1. finitamente (pontualmente),
   - 2.2. assintoticamente: $\lim \delta_i = 1$.

A existência de $(x_i x_\star x_{i+1})$ assintoticamente significa que as soluções aproximadas oscilam em torno de $x_\star$. O caso que se pode vislumbrar é aquele no qual $x_\star$ está em um contorno suave de $\mathcal{A}(D_0) \subset D_0$. A existência de $(x_i x_\star x_{i+1})$ pontualmente é um caso desinteressante se não houver também a convergência finita. Não se

pode admitir tal algoritmo, já que estando a solução na linha, está também no span$\{x_i, s_i = x_{i+1} - x_i\}$, e uma simples pesquisa na linha (*line search*) fornece a solução exata, ou tão aproximada quanto se queira. Admite-se portanto que $\delta_i = 1$ apenas na convergência finita ou assintoticamente.

Uma cota superior para β_i pode ser estabelecida em função de α, como consequência imediata do Lema 31.

Lema 45 ***($\{\beta_i\}_{i \geq i_\alpha}$ é limitada)*** $(\ \beta_i \leq\ )\ \gamma \leq \zeta < \infty$.

Demonstração:

$$\beta_i = \frac{d(x_{i+2}, x_{i+1})}{d(x_{i+1}, x_i)} \leq (1+\alpha)\frac{d(x_{i+1}, x_\star)}{d(x_{i+1}, x_i)} \leq \frac{1+\alpha}{1-\alpha}\alpha = \zeta < \infty.$$

Logo, $\max\limits_{i \geq i_\alpha} \{\beta_i\} = \gamma \leq \zeta$. ■

As relações básicas (3.15) e (3.16) tornam-se (com os valores da realização):

$$\beta_i \delta_i \leq \alpha \leq \gamma \leq \zeta \tag{3.25}$$

$$\beta_i \delta_i \leq \gamma \tag{3.26}$$

Lema 46 $\lim \delta_i = 1 \Rightarrow \lim \{\beta_i\} \leq \alpha \leq \gamma$.

Antes da convergência admite-se $\delta_i < 1$. Desta forma β_i deve oscilar em torno de α. Porém,

$$\text{se } \lim \{\beta_i \delta_i\} = \gamma \Rightarrow \gamma \leq \alpha \leq \gamma \Rightarrow \alpha = \gamma < 1.$$

Proposição 47 ***(Suficiência para contração iterada)*** *Se* $\lim \{\beta_i \delta_i\} = \gamma$ *então* $x_{i+1} = \mathcal{A}(x_i), i = 0, 1, \ldots$ *é uma contração iterada com* $\alpha = \gamma < 1$.

No caso mais geral, quando β_i oscila em torno de α, é ainda possível estabelecer uma cota inferior para α, independente do comportamento de δ_i ao longo das iterações.

Proposição 48 ***(Cota inferior para α)*** *Defina*

$$\widetilde{\alpha}_i = 1/2 \left[-(1+\beta_i) + \sqrt{(1+\beta_i)^2 + 4\beta_i} \right],$$

e, $\widetilde{\alpha} = \max \{\widetilde{\alpha}_i\}$, *para todo* $i \geq i_\alpha$, *então,*

$$\widetilde{\alpha}_i \leq \beta_i$$

$$\widetilde{\alpha} \leq \alpha. \tag{3.27}$$

Demonstração: Suponha que β_i atinja ζ, sua cota superior do Lema 45, ou seja, $\beta_i = \gamma = \zeta$. O valor correspondente, mínimo para α, pode ser encontrado resolvendo $\beta_i = \zeta(\alpha)$. Este valor, que será chamado $\widetilde{\alpha}_i$, é uma cota inferior para α:

$$\frac{1+\widetilde{\alpha}_i}{1-\widetilde{\alpha}_i}\widetilde{\alpha}_i = \beta_i = \zeta(\widetilde{\alpha}_i) \leq \frac{1+\alpha}{1-\alpha}\alpha = \zeta(\alpha) \;, \tag{3.28}$$

ou seja, $\zeta(\widetilde{\alpha}_i) \leq \zeta(\alpha)$. Como ζ é uma função estritamente crescente com seu argumento, segue-se que $\widetilde{\alpha}_i \leq \alpha$. Mas, (3.28) aplica-se a qualquer $i \geq i_\alpha$, e portanto $\max\limits_{i \geq i_\alpha} \{\widetilde{\alpha}_i\} \leq \alpha$. Para provar que $\widetilde{\alpha}_i \leq \beta_i$, note que $\kappa_i = (1+\widetilde{\alpha}_i)/(1-\widetilde{\alpha}_i) \geq 1$, e portanto, como $\kappa_i \widetilde{\alpha}_i = \beta_i \Rightarrow \widetilde{\alpha}_i \leq \beta_i$, com a igualdade ocorrendo apenas se $\kappa_i = 1 \Rightarrow \widetilde{\alpha}_i = \beta_i = \alpha = \zeta = 0$. ■

A importância do resultado acima reside no fato de que ele é independente de γ, portanto também vale para $\gamma \geq 1$ (expansão).

Proposição 49 ***(Geometria da realização)*** *Se $x_{i+1} = \mathcal{A}(x_i), i = 0, 1, \ldots$ é q-linear, então, para $i \geq i_\alpha$ tem-se:*

$$\frac{\beta_i}{1+\alpha} \leq \eta_i \leq \min\left\{\frac{\alpha}{1-\alpha}, \frac{\beta_i}{1-\alpha}\right\} \tag{3.29}$$

$$\frac{1-\alpha}{1+\alpha}\xi_i \leq \beta_i \leq \frac{1+\alpha}{1-\alpha}\xi_i \leq \zeta \tag{3.30}$$

$$\frac{1-\alpha}{1+\alpha}\beta_i \leq \xi_i \leq \min\left\{\alpha, \frac{1+\alpha}{1-\alpha}\beta_i\right\} < 1 \tag{3.31}$$

$$\max\left\{\max_{i \geq i_\alpha}\left\{\frac{|\beta_i - \xi_i|}{\beta_i + \xi_i}\right\}, \widetilde{\alpha}, \max_{i \geq i_\alpha}\{\beta_i \delta_i\}\right\} \leq \alpha < 1 \tag{3.32}$$

Demonstração: A (3.29) é uma aplicação direta do Lema 31. As (3.30) e (3.31) vêm de (3.29). A (3.32) de (3.30), (3.31), (3.25), e Proposição 48. ■

O seguinte resultado oferece condições para a existência de cotas superiores para α.

Proposição 50 *Se $\xi_i \leq \beta_i$ para todo $i \geq i_\alpha$, então $\max\limits_{i \geq i_\alpha} \beta_i \geq \alpha$, $\delta < 1$, e, definindo $\widehat{\alpha}_i = 1/2\left[(1+3\beta_i) - \sqrt{(1+3\beta_i)^2 - 4\beta_i}\right]$, $\widehat{\alpha} = \max\limits_{i \geq i_\alpha}\{\widehat{\alpha}_i\}$, tem-se:*

1. $\widehat{\alpha} \leq \alpha \leq \max\limits_{i \geq i\alpha}\{\beta_i\}$
2. $\widehat{\alpha} < 1/3$

3. A Proposição 49 torna-se:

$$\frac{\beta_i}{1+\alpha} \leq \eta_i \leq \frac{1}{1-\alpha}\min\left\{\frac{\alpha}{1-\alpha}, \frac{\beta_i}{1-\alpha}\right\}$$

$$\xi_i \leq \beta_i \leq \frac{1+\alpha}{1-\alpha}\xi_i \leq \zeta$$

$$\frac{1-\alpha}{1+\alpha}\beta_i \leq \xi_i \leq \min\{\alpha, \beta_i\} < 1$$

$$\max\left\{\max_{i\geq i_\alpha}\left\{\frac{\beta_i - \xi_i}{\beta_i + \xi_i}\right\}, \tilde{\alpha}, \hat{\alpha}, \max_{i\geq i_\alpha}\{\beta_i\delta_i\}\right\} \leq \alpha \leq \max_{i\geq i_\alpha}\{\beta_i\} < 1$$

Demonstração: Que $\xi_i \leq \beta_i \Rightarrow \max_{i\geq i_\alpha} \beta_i \geq \alpha \Rightarrow \alpha \leq \max_{i\geq i_\alpha}\{\beta_i\}$ é imediato. $\delta < 1$ é consequência de (3.25). Usando (3.30) e (3.31) obtem-se:

$$\beta_i \leq \frac{1+\alpha}{1-\alpha}\xi_i \Rightarrow \beta_i + \xi_i \leq \frac{1+\alpha}{1-\alpha}\xi_i + \alpha \Rightarrow \alpha^2 - (1+3\beta_i)\alpha + \beta_i \leq 0$$

$$\Rightarrow \alpha \geq 1/2\left[(1+3\beta_i) - \sqrt{(1+3\beta_i)^2 - 4\beta_i}\right] = \hat{\alpha} \leq \beta_i.$$

e os ítens 1 e 2 seguem-se. Os resultados do ítem 3 são imediatos. ■

Na verdade $\tilde{\alpha} \geq \hat{\alpha}$, de forma que $\hat{\alpha}$ é dispensável como cota inferior para α. Isto significa que a restrição $\beta_i \geq \xi_i$ é forte demais, e exige α pequeno. Adiante será provado que devem existir finitamente muitos $\beta_i < 1$, tal que a condição 1 acima reveste-se de grande importância no estabelecimento de cotas superiores para α. Algoritmos de projeção em espaços de Hilbert permitem determinar melhores cotas inferiores para α.

A oscilação de β_i no entorno de α é de fato observada em praticamente todos os bons algoritmos (a não ser que $\delta_i = 1$ para todo $i \geq i_\alpha$, o que é inaceitável, como visto - relembre que δ_i desta subseção não é o do processo). Muitas vezes esta oscilação pode ser reduzida com benefícios diversos para as iterações. Isto será objeto da Parte II.

É importante observar o papel de β_i na redução instantânea do erro, independentemente do valor de α. O valor de α determina a redução mínima do erro ao longo das iterações. Esta redução pode ser ainda maior quando $\beta_i < \alpha$. Ela foi chamada de *bizarra* em [20] por sua ocorrência no CG na norma Euclidiana do erro. No entanto, ela é extremamente bem vinda, e é justamente a manutenção contínua desta redução que leva a algoritmos q-superlineares. Ela reflete a eficiência instantânea de algumas iterações.

A redução pode ocorrer apenas quando em (3.31) existir um β_i tal que

$$\frac{1+\alpha}{1-\alpha}\beta_i \leq \alpha < 1 \Rightarrow 0 \leq \beta_i \leq \left(\sqrt{2}-1\right)^2 \cong 0.17157. \tag{3.33}$$

Neste caso, para a redução ser possível, tem-se que ter:

$$\beta_i \leq \left(\sqrt{2}-1\right)^2 \Rightarrow \beta_i \leq \alpha_i^{(-)} \leq \alpha \leq \alpha_i^{(+)} \tag{3.34}$$

$$\alpha_i^{(+),(-)} = \frac{1}{2}\left[(1-\beta_i) \pm \sqrt{(1-\beta_i)^2 - 4\beta_i}\right]. \tag{3.35}$$

Tendo em vista a discussão acima, considere-se a possibilidade de $\zeta < 1$, ou seja, a sequência $\{\beta_i\}_{i\geq i_\alpha}$ conter apenas elementos menores do que um. Pela Proposição 49 isto deve ocorrer obrigatoriamente toda vez que $\alpha < \sqrt{2}-1$.

Proposição 51 ***(Suficiência para contração iterada)*** *Se $\alpha < \sqrt{2}-1$, então $x_{j+1} = \mathcal{A}(x_j)$, $j = i_\alpha, i_\alpha + 1, \ldots,$ é uma contração iterada.*

Proposição 52 ***(Suficiência para majoração do erro)*** *Se $\alpha \leq 1/2$, então a sequência $\{d(x_{i+1}, x_\star)\}_{i\geq i_\alpha}$ é majorada por $\{d(x_{i+1}, x_i)\}_{i\geq i_\alpha}$.*

Este último resultado, consequência direta da Proposição 49, diz que o incremento somente pode ser usado como uma estimativa do erro, ou seja $d(x_{i+1}, x_\star) \leq d(x_{i+1}, x_i)$, se $\alpha \leq 1/2$. Infelizmente, como será visto na próxima subseção, isso não ocorrerá jamais para algoritmos de projeção ortogonal em espaços de Hilbert.

Os resultados obtidos até agora descrevem parte da geometria da realização q-linear diretamente vinculada com a estimativa de erros, e levam em conta apenas um enriquecimento: o necessário para a determinação de β_i. Na formulação de uma estratégia efetiva para estimativas de erros durante a realização, aqueles resultados devem-se juntar aos do Capítulos 2.

É possível introduzir algumas aproximações nas estimativas (exatas) anteriores. Abaixo, são dadas algumas possibilidades. As melhores opções ficarão a depender das características de cada algoritmo, obtidas de sua análise antecipada.

Com este propósito, considere-se a expressão de η_i na Proposição 49,

$$\eta_i \leq \max\left\{\frac{\alpha}{1-\alpha}, \frac{\beta_i}{1-\alpha}\right\} \leq \frac{\widetilde{\gamma}}{1-\alpha} \leq \frac{\gamma}{1-\alpha} \leq \frac{\zeta}{1+\alpha} \leq \frac{\zeta}{1+\widetilde{\alpha}} \tag{3.36}$$

onde $\widetilde{\gamma} = \max\{\beta_i : \beta_i < 1\}$. Pode-se aproximar ζ por γ ou $\widetilde{\gamma}$, obtendo

$$\eta_i \leq \frac{\widetilde{\gamma}}{1+\widetilde{\alpha}} \leq \frac{\gamma}{1+\widetilde{\alpha}} \tag{3.37}$$

e, similarmente,

$$\frac{\widetilde{\alpha}}{1+\zeta} \leq \frac{\beta_i}{1+\zeta} \leq \frac{\beta_i}{1+\gamma} \leq \frac{\beta_i}{1+\widetilde{\gamma}} \leq \frac{\beta_i}{1+\alpha} \leq \eta_i$$

ou ainda,

$$\frac{\widetilde{\alpha}}{1+\widetilde{\gamma}} \leq \frac{\beta_i}{1+\widetilde{\gamma}} \tag{3.38}$$

onde $\widetilde{\alpha}$ e $\widetilde{\gamma}$ são atualizados à medida em que o algoritmo avança.

Em muitos algoritmos é viável usar mais de um enriquecimento na avaliação dos erros. Por exemplo, no GMRES(m) para sistemas lineares pode-se desejar avaliar o erro da solução na norma Euclidiana apenas após cada ciclo de m iterações. O melhor exemplo está em adaptação em Elementos Finitos ou Elementos de Contorno. Quando o objetivo é atingir uma determinada medida do erro, as melhores estratégias (vide programa StressCheck, de Barna Szabó e Ivo Babuska, por exemplo) consistem em efetuar o melhor enriquecimento possível (limitado por fatores econômicos e computacionais apenas: tempo de execução e memória disponível), e fazer a estimativa através de retroanálise do erro para baixos enriquecimentos. Em particular, quando são usadas as versões p e hp, em estratégias com malhas não uniformes. Para casos como este, e também para algoritmos pobres, considere-se os resultados que se seguem.

Proposição 53 ***(Estimador com*** 1 ***enriquecimento)*** *Se* $x_{i+1} = \mathcal{A}(x_i)$, $i = 0, 1, \ldots,$ *é q-linear, então, para* $i \geq i_\alpha$ *tem-se:*

$$\frac{\beta_{i+1}-1}{1+\alpha^2}\beta_i \leq \eta_i \leq (1+\frac{\beta_{i+1}}{1-\alpha})\beta_i. \tag{3.39}$$

Demonstração:

$$\begin{aligned} d(x_i, x_*) &\leq d(x_i, x_{i+1}) + d(x_{i+1}, x_*) \leq d(x_i, x_{i+1}) + (1-\alpha)^{-1} d(x_{i+2}, x_{i+1}) \\ &= d(x_i, x_{i+1}) + \beta_i (1-\alpha)^{-1} d(x_{i+1}, x_i) \\ &= (1+\beta_i(1-\alpha)^{-1}) d(x_{i+1}, x_i) \end{aligned}$$

e, por indução,

$$d(x_{i+1}, x_*) \leq (1+\beta_{i+1}(1-\alpha)^{-1}) d(x_{i+2}, x_{i+1}) = (1+\beta_{i+1}(1-\alpha)^{-1})\beta_i d(x_{i+1}, x_i).$$

Também,

$$d(x_{i+1}, x_{i+2}) \leq d(x_{i+1}, x_i) + d(x_i, x_*) + d(x_*, x_{i+2}) \leq d(x_{i+1}, x_i) + (1+\alpha^2) d(x_i, x_*)$$

$$\Rightarrow d(x_\star, x_i) \geq \frac{\beta_i - 1}{1+\alpha^2} d(x_{i+1}, x_i) \ ,$$

e, por indução,

$$d(x_{i+1}, x_\star) \geq \frac{\beta_{i+1} - 1}{1+\alpha^2} d(x_{i+2}, x_{i+1}) = \frac{\beta_{i+1} - 1}{1+\alpha^2} \beta_i d(x_{i+1}, x_i)$$

■

O resultado acima complementa a Proposição 49 em cada ítem. Por exemplo, a (3.29) passa a ser

$$\max\left\{\frac{\beta_i}{1+\alpha}, \frac{\beta_{i+1}-1}{1+\alpha^2}\beta_i\right\} \leq \eta_i \leq \min\left\{\frac{\alpha}{1-\alpha}, \frac{\beta_i}{1-\alpha}, (1+\frac{\beta_{i+1}}{1-\alpha})\beta_i\right\} \tag{3.40}$$

No caso de muitos enriquecimentos, obtem-se por indução, para $p \geq 1$:

Proposição 54 ***(Estimador com p enriquecimentos)*** *Se $x_{i+1} = \mathcal{A}(x_i)$, $i = 0, 1, \ldots,$ é q-linear, então, para $i \geq i_\alpha$ tem-se, para cada $p \geq 1$:*

$$\frac{\beta_{i+p} - 1}{1+\alpha^2} \prod_{k=i}^{i+p-1} \beta_k d(x_{i+1}, x_i) \leq d(x_{i+p}, x_\star) \leq \left(1 + \frac{\beta_{i+p}}{1-\alpha}\right) \prod_{k=i}^{i+p-1} \beta_k d(x_{i+1}, x_i) \tag{3.41}$$

Proposição 55 *Se $x_{i+1} = \mathcal{A}(x_i)$, $i = 0, 1, \ldots,$ é q-linear, então, para $i \geq i_\alpha$ tem-se*

$$\lim_{p\to\infty} \prod_{k=i}^{i+p} \beta_k = 0.$$

Demonstração: A sequência é q-linear então converge: $\lim_{p\to\infty} d(x_{i+p}, x_\star) = d(x_\star, x_\star) = 0$. O resultado segue-se da Proposição 54, uma vez que $d(x_{i+1}, x_i) \neq 0$, já que por hipótese não há *breakdown.* ■

Deste resultado conclui-se que devem existir suficientes $\beta_i < 1$ antes da convergência ser aceita (critério de parada). Isto tem implicações interessantes em muitos algoritmos. Por exemplo, se a Proposição 49 prevalece, pode-se sempre estimar α, e assim a estimativa do erro exata, por $\widetilde{\gamma} = \min\{\beta_i : \beta_i < 1\}$ adaptativamente.

Deve-se observar também que (3.39) prevalece sobre $\alpha/(1-\alpha)$ apenas se $\beta_{i+p} < \alpha$. Neste sentido, por simples substituição deve-se ter

$$\frac{\prod_{k=i}^{i+p-1} \beta_k + \prod_{k=i}^{i+p} \beta_k}{\prod_{k=i}^{i+p-1} \beta_k + 1} < \alpha < 1, \tag{3.42}$$

pelo que se conclui que esta possibilidade existe apenas a partir do momento em que $\prod_{k=i}^{i+p} \beta_k < 1$.

3.3.4 q-linearidade e escalabilidade

Os resultados descritos até agora supõe pouco conhecimento do algoritmo. Todas as estimativas desenvolvidas são assintoticamente exatas. Melhores resultados são obtidos em espaços e algoritmos específicos.

Conclui-se os resultados gerais com a observação de que é possível estimar o número máximo de iterações necessárias para reduzir o erro a $\varepsilon d(x_0, x_*)$, ε dado. Na pior das hipóteses, na qual $d(x_{i+1}, x_*) = \alpha d(x_i, x_*)$ tem-se

$$\varepsilon = \frac{d(x_i, x_*)}{d(x_0, x_*)} = \frac{d(x_i, x_*)}{d(x_{i\alpha}, x_*)} \frac{d(x_{i\alpha}, x_*)}{d(x_0, x_*)} = \alpha^{i-i_\alpha} \frac{d(x_{i\alpha}, x_*)}{d(x_0, x_*)} ,$$

e, portanto, considerando possíveis reduções com $\beta_i < \alpha$, o número de iterações esperadas, i_{esp}, será (após tomar o logaritmo da expressão) :

$$i_{esp} \leq i_\alpha + \frac{\log \varepsilon}{\log \alpha} + \frac{\log \frac{d(x_{i\alpha}, x_*)}{d(x_0, x_*)}}{\log \alpha} . \tag{3.43}$$

Se o algoritmo pelo menos não aumenta o erro enquanto não atinge sua fase q-linear tem-se:

$$i_{esp} \leq i_\alpha + \frac{\log \varepsilon}{\log \alpha} . \tag{3.44}$$

Por exemplo, se $\alpha = 1/2$ e $\varepsilon = 10^{-8}$ (ε muito exigente, ver Capítulo 2) $\Rightarrow$ $i_{esp} \leq i_\alpha + 27$; e, se $\alpha = 1/3 \Rightarrow i_{esp} \leq i_\alpha + 17$. Portanto, dado um algoritmo que possua convergência q-linear com α, ele é escalável no número de iterações, a partir de i_α. Esta é a resposta à escalabilidade do número de iteraçõcs, e α é a medida de velocidade desejada. Este é mais um motivo pelo qual este trabalho assume a convergência q-linear como hipótese mínima.

Entretanto, a escalabilidade do algoritmo fica dependente do número de iterações para atingir i_α, ou seja, para atingir a fase q-linear. Esta fase inicial, geralmente é r-linear (ocasionalmente q-sublinear), deve portanto ser cuidadosamente otimizada. Note também que i_{esp} aumenta da ordem $\simeq \sqrt{10}$ iterações por ordem de grandeza de aumento de $d(x_{i\alpha}, x_*)$ sobre $d(x_0, x_*)$ para $\alpha = 0.5$. Para desempenho ótimo, entretanto, deve-se evitar o crescimento do erro inicial até a chegada da fase q-linear. Muitas vezes, desejavelmente, a fase inicial caracteriza-se por um aprendizado do algoritmo sobre o operador - guardada em vetores auxiliares. Estas informações iniciais não devem portanto ser descartadas sob pena de retardar a convergência pelo reaprendizado, nem redundantes, sob pena de desperdício ou exaustão de memória e operações. Este tópico, da escalabilidade durante as primeiras i_α iterações, é objeto da segunda parte desta tese.

3.3.5 Estimativas em espaços de Hilbert

Nesta subseção, o maior interesse está de fato nos espaços Euclidianos de dimensão finita, com norma Euclidiana e Euclidiana ponderada. Porém, os resultados que se seguem são válidos em espaços de Hilbert em geral. O produto interno e a sua norma associada serão designados por $(\cdot,\cdot)$ e $\|\cdot\|$, respectivamente. A lei do paralelogramo dada pela (2.1) caracteriza o espaço de Hilbert.

Em sistemas lineares, exemplos de normas de interesse em espaços com produto interno são: entre as normas p, a norma ℓ_2, qual seja, $\|\cdot\|_2$ associada ao produto escalar convencional $(x,y) = y^H x$; a norma da energia $\|\cdot\|_A$ para sistemas simétricos e positivos definidos com matrizes dos coeficientes A; a norma $\|\cdot\|_{A^H A}$, onde A^H é a Hermitiana transposta da matriz A dos coeficientes, A regular. Claramente, normas como ℓ_1 e ℓ_∞, não estando associadas a produtos internos, devem ser estimadas pelos resultados gerais da subseção anterior.

Deve-se lembrar que em toda a discussão deste capítulo prevalece a hipótese de que o algoritmo é consistente, e que não há *breakdown*, a não ser quando explicitamente observado.

3.3.6 A condição de descendência

O leitor pode verificar que a lei do paralelogramo, dada por (2.1), leva à mesma geometria da convergência do quociente de espaços métricos já descrita. Entretanto, pode-se compreender melhor a geometria e o significado de certas restrições comumente encontradas em muitos algoritmos atuais, introduzindo a relação entre o produto interno e a norma.

A introdução do produto interno torna possível falar em ângulo entre seus elementos, e introduzir noções adicionais relacionadas com a qualidade de convergência. Usando a expressão $e_{i+1} = e_i + s_i$ e a lei do paralelogramo, (2.2), pode-se escrever, para qualquer norma $\|\cdot\|$ definida pelo produto interno $(\cdot,\cdot)$ uma relação adicional entre erros sucessivos e o incremento,

$$\|e_i + s_i\|^2 = \|e_{i+1}\|^2 = \|e_i\|^2 + \|s_i\|^2 + 2\,(e_i, s_i) \ , \tag{3.45}$$

e, com a hipótese da q-linearidade da realização na norma $\|\cdot\|$,

$$\begin{aligned} \|e_i\|^2 \ &\leq \ \frac{1}{1-\alpha^2}\left[-2\,(e_i,s_i) - \|s_i\|^2\right] = \frac{\|s_i\|^2}{1-\alpha^2}\left[-\frac{2\,(e_i,s_i)}{\|s_i\|^2} - 1\right], \\ \Rightarrow \ \|e_i\| &\leq \frac{\|s_i\|}{\sqrt{1-\alpha^2}}\left[-\frac{2\,(e_i,s_i)}{\|s_i\|^2} - 1\right]^{1/2} \end{aligned} \tag{3.46}$$

Uma vez que ambos $\|s_i\|, \|e_i\| > 0$, e a expressão entre colchetes é necessariamente não negativa para $\alpha \in [0, 1[$, o seguinte resultado pode ser enunciado.

Lema 56 ***(Condição de descendência para a convergência do quociente)*** *Para todo $i \geq i_\alpha$ valem:*

1. $(e_i, s_i) < 0$,
2. $|(e_i, s_i)| > \frac{1}{2} \|s_i\|^2$.

A condição $(e_i, s_i) < 0$ estabelece que o ângulo plano entre e_i e s_i, no subespaço afim $x_\star + span\{e_i, e_{i+1}\}$, associado à norma $\|\cdot\|$, deve ser obtuso, ou seja, s_i é uma direção descendente para a norma do erro $\|e_i\|$. Em outras palavras, significa que $x_{i+1} = x_i + s_i$ é um ponto para o qual $\|e_{i+1}\| \leq \alpha \|e_i\|$. A condição $|(e_i, s_i)| > \frac{1}{2} \|s_i\|^2$ estipula um comprimento máximo para s_i, relativamente a e_i, em função do ângulo entre eles. Obedecido o Lema 56, é possível a convergência do quociente, e com um pouco mais de exigência, mesmo a q-linearidade.

O valor de α pode ser explicitado a partir de (3.45):

$$\|e_{i+1}\|^2 = (1 + \frac{\|s_i\|^2}{\|e_i\|^2} + \frac{2\,(e_i, s_i)}{\|e_i\|^2}) \|e_i\|^2 \ . \tag{3.47}$$

Para haver q-linearidade a expressão entre parênteses deve ser < 1. Neste caso,

$$\alpha^2 = \max_{i \geq i_\alpha} \left\{ 1 + \frac{\|s_i\|^2}{\|e_i\|^2} + \frac{2\,(e_i, s_i)}{\|e_i\|^2} \right\} = \max_{i \geq i_\alpha} \left\{ 1 + \frac{\|s_i\|^2}{\|e_i\|^2} + \frac{2\,\|s_i\|}{\|e_i\|} \cos\langle e_i, s_i \rangle \right\}. \tag{3.48}$$

Deve portanto existir, para todo $i \geq i_\alpha$, um real ϵ tal que:

$$\cos\langle e_i, s_i \rangle \leq -\frac{1}{2} \left(\frac{\|e_i\|}{\|s_i\|} + \epsilon \right) \ , \epsilon > 0 \tag{3.49}$$

A partir da (3.47) pode-se explicitar os valores das variáveis ξ_i e η_i em função de $\|s_i\|$ e $\|e_i\|$. Não se fará isto aqui, porém, apenas para projeções unidimensionais na próxima subseção.

Por (3.49) conclui-se que maior o ϵ, menor o α, mais rápida a convergência. O melhor (maior admissível) valor é $\epsilon = \|e_i\| \,/\, \|s_i\|$, para o qual $\alpha = 0$. O cosseno é medido no espaço com métrica $\|\cdot\|$, não necessariamente $\|\cdot\|_2$, a norma-ℓ_2. Se este $\epsilon > 0$ não existe, porém ainda $\cos\langle e_i, s_i \rangle < -1/2(\|e_i\| \,/\, \|s_i\|)$, para todo índice $i \geq i_\alpha$, a convergência é apenas q-sublinear estrita, isto é: $\|e_{i+1}\| < \|e_i\|$, que, como já demonstrado implica pelo menos em convergência q-linear em 2-etapas. Ou seja,

qualquer mapeamento composto $\mathcal{A}^p$, $p \geq 2$, é q-linearmente convergente para um único ponto fixo (vide também a Proposição 41 e comentários seguintes a ela).

A condição de convergência q-sublinear será

$$\liminf_{i \geq i_\alpha} \left\{ \frac{\|s_i\|^2}{\|e_i\|^2} + \frac{2\,(e_i, s_i)}{\|e_i\|^2} \right\} = \liminf_{i \geq i_\alpha} \left\{ \frac{\|s_i\|^2}{\|e_i\|^2} + \frac{2\,\|s_i\|}{\|e_i\|} \cos\langle e_i, s_i \rangle \right\} = 0 \,. \tag{3.50}$$

A grande maioria dos algoritmos minimizam a norma Euclidiana dos resíduos, em especial quando não se encontra nada melhor para minimizar, que é justamente o caso de sistemas lineares com matrizes não Hermitianas. Ou a norma da energia, em especial em problemas lineares com matrizes Hermitianas. Muitas versões do método dos Elementos Finitos enquadram-se nesses dois casos: por exemplo as formulações de mínimos quadrados e Ritz-Galerkin. A seguinte aplicação estabelece exatamente quando um algoritmo que minimiza a norma Euclidiana do resíduo sobre uma dada direção d_i em cada iteração, onde $s_i = \tau_i d_i$, converge q-linearmente. Minimizar a norma dos resíduos é muito interessante, uma vez que simultaneamente garante a consistência do algoritmo, porém, como visto no Capítulo 2, sujeita as soluções aproximadas ao problema de separação.

Exemplo 57 ***(q-linearidade do resíduo na norma-ℓ_2)*** *São as formulações de mínimos quadrados. Sistema linear com $A \in \mathbb{C}^n$, A regular, com o produto interno $(\cdot,\cdot) = (\cdot,\cdot)_{A^HA}$. Uma vez que $f_i = Ax_i - b$, $b \in \mathbb{C}^n$, o produto interno $(e_i, e_i)_{A^HA} = \|f_i\|_2^2 \Rightarrow \|e_i\|_{A^HA} = \|f_i\|_2$. Então, $(e_i, s_i)_{A^HA} = s_i^H A^H f_i$, e deve-se ter*

$$\cos\langle e_i, s_i \rangle_{A^HA} = \frac{(e_i, s_i)_{A^HA}}{\|e_i\|_{A^HA} \cdot \|s_i\|_{A^HA}} = \frac{s_i^H A^H f_i}{\|f_i\|_2 \cdot \|As_i\|_2} < -\frac{\|f_i\|_2}{2\,\|As_i\|_2} \,,$$

que implica que haverá convergência q-linear da norma Euclidiana do resíduo se existir um real ϵ tal que

$$(As)^H f_i \leq -\frac{1}{2}\left(\|f_i\|_2^2 + \epsilon\right) \,,\ \epsilon > 0 \,, \tag{3.51}$$

a qual é a condição para a norma Euclidiana do resíduo convergir q-linearmente. Para melhor compreender esta última relação escreva $As_i = f_{i+1} - f_i$, tal que haverá q-linearidade quando $(f_{i+1} - f_i)^H f_i \leq -1/2\left(\|f_i\|_2^2 + \epsilon\right)$. Então pode-se escrever, alternativamente,

$$f_{i+1}^H f_i \leq \|f_i\|_2^2 - \bar{\epsilon} \,,\ \bar{\epsilon} > 0. \tag{3.52}$$

O valor de $\bar{\epsilon}$ pode ser determinado a cada iteração e usado para estimar α, se for de interesse. Ou ainda, o valor de α

$$\alpha = \max_{i \geq i_\alpha} \left\{ 1 + \frac{\|s_i\|_{A^HA}^2}{\|e_i\|_{A^HA}^2} + \frac{2\,(e_i, s_i)_{A^HA}}{\|e_i\|_{A^HA}^2} \right\} = \max_{i \geq i_\alpha} \left\{ 1 + \frac{\|As_i\|_2^2}{\|f_i\|_2^2} + \frac{2(f_i, As_i)_2}{\|f_i\|_2^2} \right\} \,.$$

Segue-se que deve existir um $\epsilon > 0$ *tal que*

$$\|As_i\|_2^2 + \epsilon \leq 2\left|(f_i, As_i)_2\right| \ . \tag{3.53}$$

É possível encontrar o valor do parâmetro de busca $\tau_i > 0$ *em* $e_{i+1} = e_i + \tau_i d_i$, *ou equivalentemente,* $x_{i+1} = x_i + \tau_i d_i$, *para garantia de convergência pelo menos q-linear usando (3.51) e (3.56)*

$$\tau_i \left|(Ad_i)^H f_i\right| \geq \frac{1}{2}\left(\|f_i\|_2^2 + \epsilon\right) \Rightarrow \tau_i \geq \frac{\|f_i\|_2^2 + \epsilon}{2\left|(Ad_i)^H f_i\right|} \Rightarrow \tau_i \geq \frac{\|f_i\|_2^2}{2\left|(Ad_i)^H f_i\right|} + \bar{\epsilon} \ , \ e$$

$$\tau_i^2 \|Ad_i\|_2^2 + \epsilon \leq 2\tau_i \left|(f_i, Ad_i)_2\right| \Rightarrow \tau_i \leq \frac{2\left|(Ad_i)^H f_i\right| - \epsilon}{\|Ad_i\|_2^2} \Rightarrow \tau_i \leq \frac{2\left|(Ad_i)^H f_i\right|}{\|Ad_i\|_2^2} - \underline{\epsilon} \ .$$

Ou seja, as seguintes cotas para o parâmetro de busca devem ser obedecidas:

$$\frac{\|f_i\|_2^2}{2\left|(Ad_i)^H f_i\right|} + \bar{\epsilon} \leq \tau_i \leq \frac{2\left|(Ad_i)^H f_i\right|}{\|Ad_i\|_2^2} - \underline{\epsilon} \ . \tag{3.54}$$

Efetuando a projeção-$A^H A$ *de* e_{i+1} *sobre* d_i *tem-se (ver Capítulo 2):*

$$e_{i+1} = e_i - \frac{(e_i, d_i)_{A^H A}}{(d_i, d_i)_{A^H A}} d_i = e_i + \tau_i d_i \Rightarrow \tau_i = -\frac{(Ad_i)^H f_i}{\|Ad_i\|_2^2} = \frac{\left|(Ad_i)^H f_i\right|}{\|Ad_i\|_2^2} \ .$$

Assim, o algoritmo satisfaz (3.54) para exatos $\epsilon = \tau_i$. *O algoritmo é pelo menos q-linearmente convergente se* τ_i *manter-se suficientemente afastado de zero, segundo (3.54). Ou seja, existir um real* ϵ *tal que* $0 < \epsilon \leq \tau_i$. *O próximo exemplo explora isto em um algoritmo específico.*

O próximo exemplo diz respeito à minimização da norma da energia no método dos gradientes conjugados (CG), que de imediato equivale a certas versões adaptativas do método dos elementos finitos, especificamente, algumas técnicas adaptativas eficientes com versões *p*, *hp*, e *hpq* [72] em formulações Ritz-Galerkin. Este exemplo deveria ser enunciado como um teorema da qualidade de convergência do CG. Pelo tanto que o autor saiba, não existe na literatura senão a prova da convergência *r*-linear, a qual é bastante conservadora, como toda a experiência numérica com o CG desde a sua existência confirma. Note também que a prova abaixo independe de quaisquer propriedades adicionais do operador, senão as necessárias para o CG convergir. Também, ela vale para $i_\alpha = 0$, ou seja, convergência global. Claro que o valor específico de $\alpha \in [0, 1[$ depende de propriedades do operador, bem como do lado direito da equação. A prova abaixo não. Isto apenas indica o que já havia sido antecipado, que a estimativa de α, e portanto do erro, deve ser feita durante a realização.

Exemplo 58 ***(q-linearidade do CG Padrão na norma da Energia)*** *Sistema linear com* $A \in L(\mathbb{C}^n)$, $A = A^H$, *autoadjunta, e positiva definida (Hpd), com o produto interno* $(\cdot,\cdot) = (\cdot,\cdot)_A$. *Tem-se que* $(e_i, e_i)_A = (e_i, f_i)_2$, *e* $\|e_i\|_A = \left\|f_i^H e_i\right\|_2$. *Portanto,*

$$\cos\langle e_i, s_i\rangle_A = \frac{(e_i, s_i)_A}{\|e_i\|_A \cdot \|s_i\|_A} = \frac{s_i^H f_i}{\|f_i^H e_i\|_2 \cdot \|s_i\|_A} \leq -\frac{\left\|f_i^H e_i\right\|_2}{2\,\|s_i\|_A},$$

que implica em $s_i^H f_i \leq -1/2\left(\left\|f_i^H e_i\right\|_2^2 + \epsilon\right) = -1/2\left(\|e_i\|_A^2 + \epsilon\right)$, $\epsilon > 0$, *como a condição para a norma da energia convergir q-linearmente. Contrariamente ao exemplo anterior,* $\|e_i\|_A$ *não é computável, podendo apenas ser estimada. Então, o valor de* α *para a norma da energia é dado por*

$$\alpha = \max_{i \geq i_\alpha}\left\{1 + \frac{\|s_i\|_A^2}{\|e_i\|_A^2} + \frac{2\,(e_i, s_i)_A}{\|e_i\|_A^2}\right\} = \max_{i \geq i_\alpha}\left\{1 + \frac{s_i^H A s_i}{f_i^H e_i} + \frac{2 s_i^H f_i}{f_i^H e_i}\right\}. \tag{3.55}$$

Segue-se que deve existir um $\epsilon > 0$ *tal que*

$$s_i^H A s_i + \epsilon \leq 2\left|s_i^H f_i\right| . \tag{3.56}$$

O valor limite de τ_i *em* $s_i = \tau_i d_i$ *para pelo menos haver q-linearidade é*

$$\left|s_i^H f_i\right| \geq 1/2\left(\left\|f_i^H e_i\right\|_2^2 + \epsilon\right) \Rightarrow \tau_i \geq \frac{\left\|f_i^H e_i\right\|_2^2 + \epsilon}{2\left|d_i^H f_i\right|} \Rightarrow \tau_i \geq \frac{\left\|f_i^H e_i\right\|_2^2}{2 d_i^H f_i} + \overline{\epsilon}\ , \ e$$

$$\tau_i^2 d_i^H A d_i + \epsilon \leq 2\tau_i\left|d_i^H f_i\right| \Rightarrow \tau_i \leq \frac{2\tau_i\left|d_i^H f_i\right| - \epsilon}{d_i^H A d_i} \Rightarrow \tau_i \leq \frac{\left|d_i^H f_i\right|}{d_i^H A d_i} - \underline{\epsilon} \tag{3.57}$$

Ou seja,

$$\frac{\left\|f_i^H e_i\right\|_2^2}{2\left|d_i{}^H f_i\right|} + \overline{\epsilon} \leq \tau_i \leq \frac{2\left|d_i{}^H f_i\right|}{d_i^H A d_i} - \underline{\epsilon} \tag{3.58}$$

Ou, mais geralmente, na norma da energia, (para o CG, $E = A$*), deve existir um* $\epsilon > 0$ *tal que*

$$\frac{\|e_i\|_E^2}{2\left|(e_i, d_i)_E\right|} + \epsilon \leq \tau_i \leq \frac{2\left|(e_i, d_i)_E\right|}{\|d_i\|_E} - \epsilon \tag{3.59}$$

Note que a cota superior é computável. No CG padrão tem-se $e_{i+1} = e_i + \tau_i d_i$ *(ou, equivalentemente* $x_{i+1} = x_i + \tau_i d_i$ *e* $f_{i+1} = f_i + \tau_i A d_i$*), onde* $\tau_i = -(d_i, f_i)_2 / (d_i, A d_i)_2 = \left|(e_i, d_i)_E\right| / \|d_i\|_E^2$, *de tal forma que a cota superior é satisfeita para qualquer* $\epsilon \leq \left|(e_i, d_i)_E\right| / \|d_i\|_E^2 = \tau_i$. *Ele não possui breakdown, a não ser que a convergência finita tenha sido atingida - que sempre acontece em algum* $i \leq n$. *Também* $0 < \epsilon \leq \tau_0 \leq 1$, *e* $\lim \tau_i = 1$ *uniformemente. Portanto o método dos gradientes conjugados possui*

convergência global q-linear e q-superlinear na norma da energia, isto é, existe uma sequência $\{\alpha_i\}_{i\geq 0}$, $\alpha_i < 1$, *e* $\lim \alpha_i = 0$. *Note que o CG é uma projeção-A. Pode-se substituir o valor de* τ_i *em (3.46) e obter a estimativa da norma de energia em cada iteração*

$$\left[-\frac{2\left(e_i, s_i\right)_E}{\|s_i\|^2} - 1\right] = \left[-\frac{2\left(e_i, d_i\right)_E}{\tau_i \|d_i\|^2} - 1\right] = 1 \Rightarrow \|e_i\|_E \leq \frac{\|s_i\|_E}{\sqrt{1-\alpha^2}} \ . \tag{3.60}$$

Os resultados dos exemplos acima justificam o estudo detalhado de projeções nas próximas subseções. Ver-se-á que a (3.60) é de fato uma propriedade de todos os algoritmos de projeção unidimensional ortogonal com respeito a algum produto interno em espaços de Hilbert. Mas, existem muito mais propriedades marcantes.

Retornando ao Teorema 19, no Capítulo 2. É possível esclarecer porque o limite $\sqrt{2}$ para γ. A lei do paralelogramo pode ser escrita para uma norma associada a um produto interno e para $\|e_{i+1}\| \leq \gamma \|e_i\|$:

$$\|e_{i+1}\|^2 = 2\|e_i\|^2 + 2\|s_i\|^2 - \|e_i - s_i\|^2$$

$$\Rightarrow \|e_i\|^2 = \frac{1}{2-\gamma^2}\left[\|e_i - s_i\|^2 - 2\|s_i\|^2\right] \ .$$

A condição de descendência para $\gamma < \sqrt{2}$ torna-se:

$$\|e_i - s_i\|^2 - 2\|s_i\|^2 \geq 0 \ .$$

Porém, se $\gamma > \sqrt{2}$, é suficiente trocar o sinal da condição de descendência acima: $2\|s_i\|^2 - \|e_i - s_i\|^2 \geq 0$, para existir a possibilidade de convergência do quociente.

O quadrado da norma é uma medida. É mais elegante trabalhar com a definição $\|e_{i+1}\|^2 \leq \alpha \|e_{i+1}\|^2$, em lugar de $\|e_{i+1}\| \leq \alpha \|e_{i+1}\|$. As expressões ficam mais simples, porém não se evidencia as vantagens da particularização para espaços com produto interno, que é o maior propósito aqui.

3.3.7 Projeções ortogonais unidimensionais

Os algoritmos mais bem sucedidos em inúmeras áreas decorrem de projeções em espaços de Hilbert. Entre eles os algoritmos de projeção ortogonal são os mais distintos. Esta subseção reserva uma atenção especial para os algoritmos de projeção unidimensional, e a próxima para projeções ortogonais multidimensionais. Projeções ortogonais são aqui entendidas como projeções ortogonais com respeito a um produto interno $(\cdot,\cdot)_\bullet$.

Exemplos de projeções ortogonais em sistemas lineares são o método dos gradientes conjugados, os métodos da classe de gradientes conjugados generalizados, toda a família do GMRES(m), métodos de resíduos mínimos, métodos de gradiente em geral. Também estão incluídos muitas classes de estratégias de convergência global baseadas em função de descendência, quer para sistemas lineares ou não lineares e otimização. A grande maioria dos métodos de viabilidade convexa, de enorme aplicação em áreas as mais distintas como tomografia computadorizada, restauração de imagens e otimização não linear. O método de Ritz e o de Galerkin, e portanto o próprio método dos elementos finitos e suas técnicas de adaptação.

Em cada iteração, um algoritmo gera uma nova solução aproximada x_{i+1} que se espera esteja mais perto da solução $x_\star$ do que x_i. Se o algoritmo é q-linear em uma norma $\|\cdot\|_\bullet$, ela pode ser tomada como uma medida conveniente da proximidade da resposta, pelo menos do ponto de vista puramente matemático, à parte as medidas de interesse particular, e os problemas de separação e computabilidade. Em geral, a forma mais simples (infelizmente, nem sempre possível) de impor a redução da norma é exigir que o erro e_{i+1} seja ortogonal ao enriquecimento s_i com respeito ao produto interno $(\cdot,\cdot)_\bullet$ responsável pela norma $\|\cdot\|_\bullet$. Esta ortogonalidade unidimensional pode ser traduzida em

$$(e_{i+1}, s_i)_\bullet = 0,\ s_i \neq e_{i+1} \neq 0\ , \tag{3.61}$$

para todo $i \geq i_\alpha$. Este tipo de ortogonalização corresponde à pesquisa na linha (*line search*), ou seja, minimização de $\|e_{i+1}\|_\bullet$ ao longo dos sucessivos subespaços unidimensionais *span* $\{s_i\}$. Entre os algoritmos nesta linha estão as diversas versões do *steepest descent* e o método das iterações com resíduo mínimo (MR). O método padrão dos gradientes conjugados também faz pesquisa unidimensional, embora, no caso de sistemas lineares ou otimização quadrática, ela corresponda a uma pesquisa multidimensional, dadas às características próprias do algoritmo.

Usando a lei do paralelogramo, para todo $i \geq i_\alpha$,

$$\begin{aligned} \|e_i\|_\bullet^2 &= \|e_{i+1}\|_\bullet^2 + \|s_i\|_\bullet^2 - 2\,(e_{i+1}, s_i)_\bullet \\ &= \|e_{i+1}\|_\bullet^2 + \|s_i\|_\bullet^2\ . \end{aligned} \tag{3.62}$$

A partir da qual se obtém

$$\|e_{i+1}\|_\bullet^2 = (1 - \frac{\|s_i\|_\bullet^2}{\|e_i\|_\bullet^2})\,\|e_i\|_\bullet^2 \tag{3.63}$$

$$\Rightarrow\ \xi_i^2 = (1 - \frac{\|s_i\|_\bullet^2}{\|e_i\|_\bullet^2}) \qquad \text{e} \qquad \eta_i^2 = (\frac{\|e_i\|_\bullet^2}{\|s_i\|_\bullet^2} - 1) \tag{3.64}$$

$$\Rightarrow \alpha^2 = \max_{i \geq i_\alpha} \left\{ 1 - \frac{\|s_i\|_\bullet^2}{\|e_i\|_\bullet^2} \right\} \Rightarrow (1-\alpha^2)^{1/2} = \min_{i \geq i_\alpha} \frac{\|s_i\|_\bullet}{\|e_i\|_\bullet} \tag{3.65}$$

$$\Rightarrow 0 < \epsilon \leq \frac{\|s_i\|_\bullet}{\|e_i\|_\bullet} \leq 1 \,, \tag{3.66}$$

para todos os $i \geq i_\alpha$. Os seguinte resultados envolvendo todos os métodos de projeção ortogonal unidimensional tornam-se imediatos.

Proposição 59 *Um método de projeção ortogonal, com respeito ao produto interno* $(\cdot,\cdot)_\bullet$, *em espaços de Hilbert é q-convergente na norma associada* $\|\cdot\|_\bullet$ *se, e somente se, existe um* $i_\alpha \geq 0$ *tal que, para todo* $i \geq i_\alpha$,

$$\frac{\|s_i\|_\bullet}{\|e_i\|_\bullet} \leq 1 \,, \tag{3.67}$$

e, em adição, para todo $i \geq i_\alpha$, *se converge de forma*

1. *q-linear então* $\liminf_{i \geq i_\alpha} \frac{\|s_i\|_\bullet}{\|e_i\|_\bullet} = (1-\alpha^2)^{1/2} \geq \epsilon > 0$,
2. *q-superlinear então* $\limsup_{i \geq i_\alpha} \frac{\|s_i\|_\bullet}{\|e_i\|_\bullet} = 1$ *e* $\liminf_{i \geq i_\alpha} \frac{\|s_i\|_\bullet}{\|e_i\|_\bullet} = (1-\alpha^2)^{1/2} > 0$,
3. *q-sublinear então* $\lim \frac{\|s_i\|_\bullet}{\|e_i\|_\bullet} = 0$,
4. *q-sublinear estrita então* $\lim \frac{\|s_i\|_\bullet}{\|e_i\|_\bullet} > 0$.

Assim, a condição de convergência q-linear de todos os algoritmos de projeção ortogonal unidimensional fica garantida se existe um $i_\alpha \geq 0$ tal que $\frac{\|s_i\|_\bullet}{\|e_i\|_\bullet}$ estcja suficientemente afastado de zero em todos os $i \geq i_\alpha$. Ou, mais precisamente, existe um real ϵ , tal que $\frac{\|s_i\|_\bullet}{\|e_i\|_\bullet} \geq \epsilon > 0$. Note também que se $\frac{\|s_i\|_\bullet}{\|e_i\|_\bullet} = 1$ para algum i, então a convergência finita se dará, com $\alpha = 0$. A grande maioria dos autores prefere caracterizar a convergência q-superlinear assintótica justamente pela condição $\limsup \{\|e_{i+1}\| \,/\, \|e_i\|\} = 0$, eliminando assim $\alpha = 0$ da definição de convergência q-linear. Isto não é conveniente aqui uma vez que está incluída na análise a possibilidade de convergência finita.

Normalmente $\|e_i\|_\bullet$ é desconhecido e deve-se usar as estimativas já estabelecidas. Entretanto, a lei do paralelogramo permite melhorar algumas daquelas estimativas.

O Lema 31 fica (extremamente) mais forte. De fato, a única maneira de melhorar as cotas nele contidas é levar em conta a q-superlinearidade das iterações (cota superior no final das iterações, assintoticamente) e particularidades do algoritmo (cota inferior no início das iterações). Isto equivale, de certa forma, a atualizar i_α a cada iteração (para a cota superior). Neste último caso, o correto procedimento será indicado no estudo da q-superlinearidade.

Lema 60 ***(Lema fundamental para projeções ortogonais)*** *Para a sequência* $x_{i+1} = \mathcal{A}(x_i)$, $i = 0, 1, 2, \ldots$, $x_i \in \mathbb{R}^n$, *q-linearmente convergente na norma* $\|\cdot\|_\bullet$, *definida pelo produto interno* $(\cdot,\cdot)_\bullet$, *para o seu limite* $x_\star$, *com fator-q dado por* α , *e tal que* $(e_{i+1}, s_i)_\bullet = 0$, $s_i \neq e_{i+1} \neq 0$, *valem as seguintes cotas para todos os índices* $i \geq i_\alpha$:

$$\|s_i\|_\bullet \leq \|e_i\|_\bullet \leq \frac{\|s_i\|_\bullet}{(1-\alpha^2)^{1/2}} \ , \tag{3.68}$$

$$\|e_{i+1}\|_\bullet \leq \overset{\bullet}{\zeta} \|s_i\|_\bullet \ , \tag{3.69}$$

onde $\overset{\bullet}{\zeta} = \alpha/(1-\alpha^2)^{1/2}$.

Demonstração: De (3.62),

$$\|e_i\|_\bullet^2 = \|e_{i+1}\|_\bullet^2 + \|s_i\|_\bullet^2 \leq \alpha^2 \|e_i\|_\bullet^2 + \|s_i\|_\bullet^2 \Rightarrow \|e_i\|_\bullet \leq (1-\alpha^2)^{-1/2} \|s_i\|_\bullet \ ,$$

e,

$$\|e_{i+1}\|_\bullet \leq \alpha \|e_i\|_\bullet \leq \alpha(1-\alpha^2)^{-1/2} \|s_i\|_\bullet = \overset{\bullet}{\zeta} \|s_i\|_\bullet .$$

Que $\|s_i\|_\bullet \leq \|e_i\|_\bullet$ (e $\|s_{i+1}\|_\bullet \leq \|e_{i+1}\|_\bullet$) é consequência da Proposição 59. ■

É relevante o fato de que $\overset{\bullet}{\zeta}$ é sempre inferior ao ζ do Lema 31, permitindo melhores cotas inferiores para α, e maiores possibilidades de reduções instantâneas de ξ_i, bem abaixo de α.

Lema 61 ***($\{\beta_i\}$ é limitada)*** $\beta_i \leq \eta_i \leq \overset{\bullet}{\zeta}$.

Demonstração: Pelo Lema 60

$$\beta_i = \frac{\|s_{i+1}\|_\bullet}{\|s_i\|_\bullet} \leq \frac{\|e_{i+1}\|_\bullet}{\|s_i\|_\bullet} = \eta_i \leq \frac{\alpha}{(1-\alpha^2)^{1/2}} = \overset{\bullet}{\zeta} .$$

■

Lema 62 $(1-\alpha^2)^{1/2}\xi_i \leq \beta_i$.

Demonstração: Pelo Lema 60

$$\|e_{i+1}\|_\bullet \leq \beta_i \frac{\|s_i\|_\bullet}{(1-\alpha^2)^{1/2}} \leq \beta_i \frac{\|e_i\|_\bullet}{(1-\alpha^2)^{1/2}} \Rightarrow (1-\alpha^2)^{1/2}\xi_i \leq \beta_i \ .$$

■

As relações básicas passam a ser

$$\beta_i\delta_i \leq \alpha \leq \gamma \leq \overset{\bullet}{\zeta}$$

$$\beta_i\delta_i \leq \gamma \ .$$

A primeira das quais fornece uma cota inferior para $\|e_i\|$ em função do enriquecimento.

$$\beta_i\delta_i \leq \alpha \Rightarrow \|s_{i+1}\|_\bullet \leq \alpha\left[\alpha\|e_i\|_\bullet + \|e_i\|_\bullet\right] = \alpha(1+\alpha)\|e_i\|_\bullet . \tag{3.70}$$

Pela qual conclui-se que $\|s_{i+1}\|_\bullet$ não é uma estimativa para $\|e_i\|_\bullet$. Também, $\|s_{i+1}\|_\bullet$ não o é para $\|e_{i+1}\|_\bullet$ pela Proposição 59, a não ser assintoticamente, para algoritmos q-superlineares. Os resultados deste trabalho não recomendam tal prática.

Proposição 63 ***(Cota inferior para α)*** *Defina* $\overset{\bullet}{\alpha}_i = \beta_i/(1+\beta_i^2)^{1/2}$, *e* $\overset{\bullet}{\alpha} = \max\left\{\overset{\bullet}{\alpha}_i\right\}$, *para todo* $i \geq i_\alpha$, *então,*

$$\overset{\bullet}{\alpha}_i \leq \beta_i$$
$$\overset{\bullet}{\alpha} \leq \alpha$$

Demonstração: Na linha da Proposição 48: $\overset{\bullet}{\zeta}\left(\overset{\bullet}{\alpha}_i\right) = \overset{\bullet}{\alpha}_i/(1-\overset{\bullet}{\alpha}_i^{\,2})^{1/2} = \beta_i \Rightarrow \overset{\bullet}{\alpha}_i = \beta_i/(1+\beta_i^2)^{1/2} \Rightarrow \overset{\bullet}{\alpha}_i \leq \beta_i$. Também, $\overset{\bullet}{\alpha}_i/(1-\overset{\bullet}{\alpha}_i^{\,2})^{1/2} = \beta_i = \overset{\bullet}{\zeta}\left(\overset{\bullet}{\alpha}_i\right) \leq \alpha/(1-\alpha^2)^{1/2} = \overset{\bullet}{\zeta}(\alpha)$. Como $\overset{\bullet}{\zeta}(\alpha)$ é uma função crescente de α, segue-se que $\overset{\bullet}{\alpha}_i \leq \alpha$, e assim, $\overset{\bullet}{\alpha} = \max\left\{\overset{\bullet}{\alpha}_i\right\} \leq \alpha$. ■

Proposição 64 ***(Geometria das projeções ortogonais q-lineares)*** *Se* $x_{i+1} = \mathcal{A}(x_i), i = 0, 1, \ldots$ *é q-linear, então, para* $i \geq i_\alpha$ *tem-se:*

$$\beta_i \leq \eta_i \leq \min\left\{\frac{\alpha}{(1-\alpha^2)^{1/2}}, \frac{\beta_i}{(1-\alpha^2)^{1/2}}\right\} \tag{3.71}$$

$$(1-\alpha^2)^{1/2}\xi_i \leq \beta_i \leq \frac{\xi_i}{(1-\alpha^2)^{1/2}} \leq \overset{\bullet}{\zeta} \tag{3.72}$$

$$(1-\alpha^2)^{1/2}\beta_i \leq \xi_i \leq \min\left\{\alpha, \frac{\beta_i}{(1-\alpha^2)^{1/2}}\right\} < 1 \tag{3.73}$$

$$\max\left\{\max_{i\geq i_\alpha}\left\{\frac{|\beta_i^2-\xi_i^2|^{1/2}}{\xi_i}\right\}, \overset{\bullet}{\alpha}, \max_{i\geq i_\alpha}\{\beta_i\delta_i\}\right\} \leq \alpha < 1 \tag{3.74}$$

Demonstração: Imediata a partir da Proposição 49 e dos Lemas anteriores. ■

Os seguintes resultados devem ser comparados aos correspondente em espaços métricos sem produto interno.

Sempre, $\widetilde{\alpha} \leq \overset{\bullet}{\alpha}$, e a diferença cresce com α. Por exemplo, alguns conjuntos $\left(\beta_i, \widetilde{\alpha}_i, \overset{\bullet}{\alpha}\right)$ são:

$$(0.2, 0.148, 0.196), (0.5, 0.281, 0.447), (0.75, 0.356, 0.60),$$
$$\left(1.0, \sqrt{2}-1 = 0.414, \sqrt{2}/2 = 0.707\right), \left(\sqrt{2} = 1.414, 0.487, 0.817\right)$$

Portanto $\overset{\bullet}{\alpha}$ é uma boa cota inferior para α, tão logo α não esteja próximo demais de 1.

O valor limite de β_i para o qual pode haver instantâneas reduções de ξ_i passa a ser maior, e assim a possibilidade de q-superlinearidade. Em (3.73)

$$\frac{\beta_i}{(1-\alpha^2)^{1/2}} \leq \alpha < 1 \Rightarrow 0 \leq \beta_i \leq 0.5 \; . \tag{3.75}$$

Neste caso, para a redução ser possível, deve-se ter:

$$\beta_i \leq 0.5 \Rightarrow \beta_i \leq \alpha_i^{(-)} \leq \alpha \leq \alpha_i^{(+)} \tag{3.76}$$

$$\alpha_i^{(+),(-)} = + \left\{ \frac{1}{2} \left[1 \pm \sqrt{1 - 4\beta_i^2} \right] \right\}^{1/2} . \tag{3.77}$$

Proposição 65 ***(Suficiência para contrações iteradas)*** *Se* $\alpha < \sqrt{2}/2 \simeq 0.707$, *então* $x_{j+1} = \mathcal{A}(x_j)$, $j = i_\alpha, i_\alpha + 1, \ldots,$ *é uma contração iterada.*

Proposição 66 *A sequência* $\{\|e_{i+1}\|_\bullet\}_{i \geq i_\alpha}$ *nunca é majorada por* $\{\|s_{i+1}\|_\bullet\}_{i \geq i_\alpha}$.

Este último resultado, consequência direta da Proposição 64, diz que o incremento nunca pode ser usado como uma estimativa do erro.

Proposição 67 $\|e_i\|_\bullet^2 \geq (1+\beta_i^2) \|s_i\|_\bullet^2$ *e* $\|e_{i+1}\|_\bullet^2 \geq (1+\beta_{i+1}^2)\beta_i^2 \|s_i\|_\bullet^2$

Demonstração: $\|e_i\|_\bullet^2 = \|s_i\|_\bullet^2 + \|e_{i+1}\|_\bullet^2 \geq \|s_i\|_\bullet^2 + \|s_{i+1}\|_\bullet^2 = (1+\beta_i^2) \|s_i\|_\bullet^2$, e a outra por indução. ■

A versão da Proposição 54 decorre dos Lemas anteriores:

Proposição 68 ***(Estimador para p enriquecimentos)*** *Se* $x_{i+1} = \mathcal{A}(x_i)$, $i = 0, 1, \ldots$, *é q-linear, então, para* $i \geq i_\alpha$ *tem-se, para cada* $p \geq 1$ *:*

$$(1+\beta_{i+p}^2)^{1/2} \prod_{k=i}^{i+p-1} \beta_k \|s_i\|_\bullet \leq \|e_{i+p}\|_\bullet \leq \left(1+\frac{\beta_{i+p}^2}{1-\alpha^2}\right)^{1/2} \prod_{k=i}^{i+p-1} \beta_k \|s_i\|_\bullet \ . \quad (3.78)$$

Este resultado diz que $\|s_i\|_\bullet^2 + \|s_{i+1}\|_\bullet^2 \leq \|e_i\|_\bullet^2$. Esta é a melhor cota inferior possível para o erro $\|e_i\|_\bullet$, considerando um único enriquecimento. Note que a cota inferior em (3.78) independe de α, no espírito da (3.71). Evidentemente, a Proposição 55 não sofre alterações.

Deve-se lembrar que o Teorema 19 do Capítulo 2 fornece condições para a existência de um ponto fixo em espaços de Hilbert quando o mapeamento é uniformemente Lipschitziano com $\gamma < \sqrt{2}$. Neste sentido, a condição de q-linearidade é muito mais forte, e β_i é limitado por $\zeta(\alpha) < \infty$ (e não $\sqrt{2}$). Além disto, como já visto, para a existência (e unicidade) do ponto fixo, exige apenas a continuidade de $\mathcal{A}$ no limite da sequência.

Existem algumas interpretações que precisam ser citadas, pelo seu significado próprio e pela relevância para esta tese como um todo. Primeiramente, a condição de ortogonalidade fornecida por (3.61) implica em:

$$e_{i+1} = e_i + s_i \Rightarrow (e_{i+1}, e_{i+1})_\bullet = (e_{i+1}, e_i)_\bullet + (e_{i+1}, s_i)_\bullet = (e_{i+1}, e_i)_\bullet \geq 0 \ . \quad (3.79)$$

Ou seja, cada vetor e_{i+1} faz um ângulo agudo com o vetor e_i. Equivalentemente, $\cos\langle e_{i+1}, e_i\rangle \geq 0$, e igual a zero apenas se $e_{i+1} = 0$. Isto pode ser visto mediante,

$$\cos\langle e_{i+1}, e_i\rangle_\bullet = \frac{(e_{i+1}, e_i)_\bullet}{\|e_{i+1}\|_\bullet \cdot \|e_i\|_\bullet} = \frac{(e_{i+1}, e_{i+1})_\bullet}{\|e_{i+1}\|_\bullet \cdot \|e_i\|_\bullet} = \frac{\|e_{i+1}\|_\bullet}{\|e_i\|_\bullet} = \xi_i$$

dando, também, uma interpretação geométrica à redução do erro ξ_i (ver (3.33)). Assim, valores baixos de ξ_i, significam, simultaneamente, a redução da norma do erro e o afastamento das direções dadas pelos vetores e_{i+1} e e_i. Isto é bastante agradável já que garante também que o vetor s_i mais se aproxima de $-e_i$, *não apenas em norma, mas também em direção.* E esta aproximação simultânea de norma e direção é medida apenas pelo real α. Assim, α de fato mede a velocidade de aproximação (mínima) das sucessivas soluções em direção à solução exata, tanto em intensidade como em direção e sentido. Tudo isto ocorre nos sucessivos subespaços afins $x_\star + span(e_{i+1}, e_i)$, visto na métrica dada por $\|\cdot\|_\bullet$. Em projeções multidimensionais, a trajetória das soluções aproximadas percorre uma hélice contínua sobre a superfície de um cone (ver adiante), que se inicia em x_0 e termina no ponto fixo $x_\star$, cuja

velocidade (normalizada) mínima de aproximação do vértice é α (ou pelo inverso, ou por $1-\alpha$, se preferir), e a instantânea é ξ_i. Apenas um escalar é necessário porque a trajetória é sobre uma linha. Justifica-se assim o termo velocidade (a unidade de tempo é a iteração) de convergência usada nesta tese. Pode-se portanto falar de aceleração se α diminui a cada iteração (convergência q-superlinear), ou de desaceleração se $\alpha \to 1$ quando i cresce (convergência q-sublinear).

Proposição 69 *A ortogonalidade unidimensional dada por (3.61) implica, para todo $i \geq 0$, em:*

1. *q-linearidade:* $\|e_{i+1}\|_\bullet \leq \alpha \|e_i\|_\bullet$ *se* $\liminf_{i \geq i_\alpha} \frac{\|s_i\|_\bullet}{\|e_i\|_\bullet} \in [\epsilon, 1]$, $\epsilon > 0$
2. $(e_{i+1}, e_i)_\bullet = (e_{i+1}, e_{i+1})_\bullet \geq 0$, *com igualdade apenas na convergência finita* $\Rightarrow \langle e_{i+1}, e_i \rangle_\bullet < \pi/2$, *enquanto* $\|e_{i+1}\|_\bullet \neq 0$,
3. $(e_{i+1}, s_j)_\bullet = \sum_{k=j+1}^{i} (s_k, s_j)_\bullet$, *para todo* $i \geq j$,
4. $(e_i, s_j)_\bullet = -\sum_{k=i}^{j} (s_k, s_j)_\bullet$, *para todo* $i < j$.

Observe-se que $(u,v)_\bullet = \overline{(v,u)_\bullet}$ para $u, v \in \mathbb{C}^n$, e $(u,v)_\bullet = (v,u)_\bullet$ apenas em casos especiais.

3.3.8 Projeções ortogonais multidimensionais

Em sistemas lineares há unanimidade na literatura contemporânea endossando a posição de que a convergência finita, admissível apenas em operadores algébricos, (possível, normalmente, em sistemas lineares) não é fundamental, uma vez que isto acontece em geral para um número muito grande de iterações, para que se possa tirar vantagens dela. Não é bem assim. De fato, o método padrão dos gradientes conjugados, durante duas décadas e meia (1951 a 1975) foi visto como um método direto porque converge no máximo em n iterações, e este fato afastou-o, parcialmente, do interesse da computação científica. A convergência finita tem outras implicações, sutis, porém importantes, uma das quais é descrita abaixo.

Inicialmente, imagine a solução de um sistema linear com solução única não trivial em $\mathbb{R}^3$. Espera-se que o algoritmo obtenha convergência em $O(3)$ iterações. Suponha que $\alpha = \sqrt{2}/2$, então, para obter uma redução do erro inicial de $\varepsilon = 10^{-8}$, tem-se $i_{esp} \leq 54$. Isto com certeza é insatisfatório para apenas 3 equações em dois aspectos. Preferencialmente espera-se obter, digamos 3 a 6 iterações, e, idealmente, a

resposta exata. Também, 54 iterações certamente perturba numericamente o analista, senão a solução, sem necessidade. Melhor seria um método que obtenha solução exata em 3 ou 6 iterações. Isto é facilmente obtido em dimensão finita observando a Proposição 69, no procedimento seguinte.

Pode-se impor, por exemplo, que $(s_k, s_j)_\bullet = 0$ para todo $k \neq j$, ou seja, gerar bases para o enriquecimento que sejam ortogonais com relação ao produto interno $(\cdot,\cdot)_\bullet$. No sistema de 3 equações acima, significa que:

$$\begin{aligned}
(e_1, s_0)_\bullet &= (e_2, s_1)_\bullet = (e_3, s_2)_\bullet = 0 \\
(e_2, s_0)_\bullet &= (s_1, s_0)_\bullet = 0 \\
(e_3, s_0)_\bullet &= (s_1, s_0)_\bullet + (s_2, s_0)_\bullet = 0 \\
(e_3, s_1)_\bullet &= (s_2, s_1)_\bullet = 0, \text{ etc}
\end{aligned}$$

A única possibilidade de $e_3 \in \mathbb{R}^3$ (ou $\mathbb{C}^3$) ser ortogonal a 3 vetores, (s_0, s_1 e s_2), ortogonais com respeito ao produto interno $(\cdot,\cdot)_\bullet$ entre si, é que e_3 seja nulo. Assim o algoritmo terá convergência finita, em no máximo 3 iterações. A importância para sistemas com muitas equações está, não na terminação finita em si, porém, na forma como ela é conseguida: por ortogonalização das bases com respeito ao mesmo produto interno em (3.61). E ortogonalização é sinônimo de estabilidade do algoritmo. Uma consequência é que, pelo menos, $\lim \frac{\|s_i\|_\bullet}{\|e_i\|_\bullet} = 1$, e a convergência q-linear (e q-superlinear) está garantida.

A ortogonalização, em geral, significa mais memória auxiliar. É prática ortogonalizar a direção de s_i de forma seletiva, por exemplo, em relação a m direções anteriores: de $s_{i-1}, s_{i-2}, \ldots, s_{i-m}$. Portanto, na ausência de *breakdown,* e se cada grupo de m direções gerar diferentes $span\{s_0, \ldots, s_{m-1}\}$, por no mínimo um s_i, existirá em um certo grupo de m iterações para o qual $\max\limits_{i=0\ldots m-1} \frac{\|s_i\|_\bullet}{\|e_i\|_\bullet} = 1$, o mais tardar em mn iterações. Exemplo deste tipo de algoritmo é o GMRES(m). A menos de imprecisões numéricas como truncamento e cancelamento, nesses casos, é apenas necessário que os somatórios nos lados direitos de 3 e 4 na Proposição 69 sejam nulos, e não necessariamente as direções precisam ser mutuamente ortogonais, porém apenas suficientemente linearmente independentes, com respeito ao produto interno $(\cdot,\cdot)_\bullet$.

Analisando o algoritmo após cada grupo de m iterações, é suficiente para haver convergência q-linear para a sequência $\{x_{mk}\}_{k\geq 1}$ que $\min\limits_{i=0\ldots m-1} \frac{\|s_i\|_\bullet}{\|e_i\|_\bullet} \in [\epsilon, 1]$, $\epsilon > 0$. Os estimadores podem então ser aplicados após cada grupo de m iterações.

Outra questão importante diz respeito ao espaço gerado pelos n elementos e_i,

na projeção multidimensional. Observe-se que, por (3.79), $(e_{i+1}, e_i)_\bullet \geq 0$, e nulo apenas se $e_{i+1} = 0$. Assim o conjunto $\{e_0, e_1, \ldots, 0\}$, com a ortogonalização das bases, $(s_k, s_j)_\bullet = 0$ para todo $k \neq j$, gera um cone convexo, e, com a condição $\|e_{i+1}\|_\bullet \leq \|e_i\|_\bullet$, especificamente um *icecream*. As soluções aproximadas estão na superfície desse cone, e a solução no seu vértice. Com o elemento nulo, $e_{n+1} = 0$, ele gera todo o espaço $\mathbb{R}^n$ (ou $\mathbb{R}^m$, se $m \leq n$, com $e_{m+1} = 0$) com produto interno $(\cdot\cdot)_\bullet$. Desta forma, para um processo em um conjunto D_0 convexo, que envolve a solução, $x_\star \in D_0$, $\mathcal{A}(D_0) \subset D_0$, $\mathcal{A}(D_0)$ é convexo (o *icecream*) e fechado, possui um ponto fixo, e se o processo é q-linear, ele o é com $\alpha \leq \gamma$, não necessariamente uma contração na norma $\|\cdot\|_\bullet$, já que o processo se realiza com $\delta < 1$, e muito possivelmente uma expansão. Vide as desigualdades (3.14), (3.15) e (3.16), Lemas 34 e 35, e Proposição 36.

Note que, acima, não é essencial que a projeção seja multidimensional para a satisfação da Proposição 36. Um exemplo disto é o método do *steepest descent*. Mesmo em R^2, ele não converge, em geral, em um número finito de iterações. Porém, as soluções sucessivas estão na superfície de um cone (leque) convexo, gerado por $\{e_0, e_1\}$, com a solução no seu vértice.

3.4 Convergência q-superlinear

Se o objetivo é conseguir as melhores estimativas possíveis para o erro de aproximação, tem-se que considerar as reduções do fator-q da realização, se estas reduções existem. Nesta seção são consideradas realizações com convergência q-superlinear. Não será feita a análise de um processo q-superlinear, principalmente porque o resultado dela não traz conclusões importantes, para os objetivos desta tese, que não possam ser retiradas da análise do processo q-linear. Outras conclusões serão relatadas em um futuro trabalho.

3.4.1 Estimativas da realização

Admite-se nesta seção a existência da sequência $\{\alpha_i\}_{i \geq i_\alpha}$ que converge para zero, onde $\alpha_i < 1$ para todo $i \geq i_\alpha$. É imediato que a cada iteração i, $i \geq i_\alpha$, a realização possue um novo fator-q igual ou menor do que fator-q ativo na iteração anterior. A definição de α_i passa a ser

$$\alpha_i = \max_{j \geq i \geq i_\alpha} \left\{ \frac{d(x_{j+1}, x_\star)}{d(x_j, x_\star)} \right\} = \max_{j \geq i \geq i_\alpha} \{\xi_j\} \tag{3.80}$$

Para a análise da realização, similarmente ao caso q-linear, valem as definições:

$$\begin{aligned} \beta_i &= d(x_{i+2}, x_{i+1})/d(x_{i+1}, x_i), \\ \delta_i &= d(x_{i+1}, x_i)/\left[d(x_{i+1}, x_*) + d(x_i, x_*)\right], \\ \eta_i &= d(x_{i+1}, x_*)/d(x_{i+1}, x_i)\ , \end{aligned} \tag{3.81}$$

e, adicionalmente,

$$\begin{aligned} \psi_i &= \frac{\alpha_{i+1}}{\alpha_i}\ , \qquad 0 \leq \psi_i \leq 1\ , \\ \delta_i(\psi) &= d(x_{i+1}, x_i)/\left[\psi_i d(x_{i+1}, x_*) + d(x_i, x_*)\right]\ , \end{aligned} \tag{3.82}$$

Tal que se pode identificar $\delta_i = \delta_i(1)$, e $\psi_i = 1$ se a realização é apenas q-linear. Como antes, supõe-se que não há *breakdown*. Claramente, $\delta_i \leq \delta_i(\psi)$.

O Lema 31 tem a seguinte versão:

Lema 70 *Para a sequência $x_{i+1} = \mathcal{A}(x_i)$, $i = 0, 1, 2, \ldots$, q-superlinearmente convergente para o seu limite x_*, valem as seguintes cotas para todos os índices $i \geq i_\alpha$:*

$$\frac{d\left(x_{i+1}, x_i\right)}{1+\alpha_i} \leq d\left(x_i, x_*\right) \leq \frac{d\left(x_{i+1}, x_i\right)}{1-\alpha_i} \tag{3.83}$$

$$d(x_{i+1}, x_*) \leq \frac{\alpha_i}{1-\alpha_i} d\left(x_{i+1}, x_i\right) = \frac{\zeta_i}{1+\psi_i\alpha_i} d\left(x_{i+1}, x_i\right) \tag{3.84}$$

onde $\zeta_i = \zeta_i(\psi) = \alpha_i\left(1+\psi_i\alpha_i\right)/(1-\alpha_i)$.

Demonstração: Como antes,

$$d(x_i, x_*) \leq d(x_i, x_{i+1}) + d(x_{i+1}, x_*) \leq d(x_i, x_{i+1}) + \alpha_i d(x_i, x_*),$$

$$d(x_i, x_{i+1}) \leq d(x_i, x_*) + d(x_*, x_{i+1}) \leq d(x_i, x_*) + \alpha_i d(x_i, x_*),$$

resultando em (3.83). Usando-as,

$$d(x_{i+1}, x_*) \leq \alpha_i d(x_i, x_*) \leq \frac{\alpha_i}{1-\alpha_i} d\left(x_{i+1}, x_i\right) = \frac{\zeta_i}{1+\psi_i\alpha_i} d\left(x_{i+1}, x_i\right).$$

e os resultados estão provados. ■

Primeiramente, as cotas para β_i. Elimina-se a noção de γ e δ, porque não se fará comparações com as contrações, afinal, não existe uma definição de supercontração.

Lema 71 *($\{\beta_i\}_{i\geq i_\alpha}$ é limitada)* $\beta_i \leq \zeta_i(\psi) < \infty$ *e* $\max\limits_{k\geq i\geq i_\alpha} \{\beta_k\} \leq \zeta_i(\psi)$

Demonstração: Pelo Lema 70

$$\frac{d\left(x_{i+2}, x_{i+1}\right)}{1+\alpha_{i+1}} \leq d\left(x_{i+1}, x_{\star}\right)$$

tal que,

$$\beta_i = \frac{d(x_{i+2}, x_{i+1})}{d(x_{i+1}, x_i)} \leq (1+\alpha_{i+1}) \frac{d(x_{i+1}, x_\star)}{d(x_{i+1}, x_i)} \leq \frac{1+\psi_i \alpha_i}{1-\alpha_i} \alpha_i < \infty.$$

Assim para qualquer $k \geq i \geq i_\alpha$ tem-se que $\beta_k \leq \zeta_k(\psi) \leq \zeta_i(\psi) \Rightarrow \max\limits_{k \geq i \geq i_\alpha} \{\beta_k\} \leq \zeta_i(\psi)$. ■

As relações básicas ficam, para todo $i \geq i_\alpha$,

$$\beta_i \delta_i \leq \beta_i \delta_i(\psi) \leq \alpha_i \leq \zeta_i(\psi) \ , \tag{3.85}$$

tal que, como $\alpha_k \leq \alpha_i$, para $k \geq i \geq i_\alpha$ tem-se $\max\limits_{k \geq i \geq i_\alpha} \{\beta_k \delta_k(\psi)\} \leq \alpha_i \leq \zeta_i(\psi)$.

Lema 72 $\max\limits_{k \geq i \geq i_\alpha} \{\beta_k \delta_k(\psi)\} \leq \alpha_i \leq \zeta_i(\psi)$.

Pode-se mesmo esperar que γ retraia-se, mas isto não é importante no presente contexto. Fundamental é a retração de $\zeta_i(\psi)$. Note que existem menores razões para a oscilação de β_i em torno de α_i uma vez que $\delta_i(\psi)$ deve assumir valores superiores a 1 toda vez que $\psi_i < 1$.

Proposição 73 ***(Cota inferior para*** α_i***)*** *Defina, para todo* $i \geq i_\alpha$,

$$\widetilde{\alpha}_i = \frac{1}{2\psi_i}\left[-(1+\beta_i) + \sqrt{(1+\beta_i)^2 + 4\psi_i \beta_i}\right], \qquad \psi_i \neq 0,$$

$$\widetilde{\alpha}_i = \frac{\beta_i}{1+\beta_i}, \qquad se \quad \psi_i = 0, \tag{3.86}$$

e, $\widetilde{\alpha} = \max\limits_{k \geq i \geq i_\alpha} \{\widetilde{\alpha}_k\}$, *então,*

$$\widetilde{\alpha}_i \leq \beta_i \ , \qquad \forall i \geq i_\alpha \ , \tag{3.87}$$

$$\widetilde{\alpha} \leq \alpha_j \ , \qquad i_\alpha \leq j \leq i \ . \tag{3.88}$$

Demonstração: Suponha que β_i atinja $\zeta_i(\psi)$, sua cota superior do Lema 71, ou seja, $\beta_i = \zeta_i(\psi)$. O valor correspondente, mínimo para α_i, pode ser encontrado resolvendo $\beta_i = \zeta_i(\psi)$. Este valor, que será chamado $\widetilde{\alpha}_i$, é uma cota inferior para α_i:

$$\frac{1+\psi_i \widetilde{\alpha}_i}{1-\widetilde{\alpha}_i} = \beta_i = \widetilde{\zeta}_i(\psi) \leq \frac{1+\psi_i \alpha_i}{1-\alpha_i} = \zeta_i(\psi). \tag{3.89}$$

$\widetilde{\zeta}_i(\psi) \leq \zeta_i(\psi)$. Como ζ é uma função estritamente crescente com seu argumento, segue-se que $\widetilde{\alpha}_i \leq \alpha_i$. Mas, para qualquer j, $i_\alpha \leq j \leq i$ temos que $\alpha_{i_\alpha} \geq \alpha_j \geq \alpha_i$, segue-se que $\max_{k \geq i \geq i_\alpha} \{\widetilde{\alpha}_k\} \leq \alpha_i \leq \alpha_j \leq \alpha_{i_\alpha}$. Para provar que $\widetilde{\alpha}_i \leq \beta_i$, note que $\kappa_i = (1 + \psi_i \widetilde{\alpha}_i) / (1 - \widetilde{\alpha}_i) \geq 1$, e portanto, como $\kappa_i \widetilde{\alpha}_i = \beta_i \Rightarrow \widetilde{\alpha}_i \leq \beta_i$, com a igualdade ocorrendo apenas se $\kappa_i = 1 \Rightarrow \widetilde{\alpha}_i = \beta_i = \alpha = \zeta = 0$. ■

Deve-se ressaltar que a definição de $\widetilde{\alpha}$ difere substancialmente daquela para realizações apenas q-lineares, no sentido que $\widetilde{\alpha}$ não é mais uma cota inferior para todos os α_i da realização, mas apenas para os anteriores: α_j, $j \leq i_\alpha$. Outra observação importante é que na expressão de $\widetilde{\alpha}_i$ há uma inclusão não linear: $\psi_i = \alpha_{i+1}/\alpha_i < 1$. Pode-se usar, conservadoramente, $\psi_i = 1$, e, se for desejado, uma posterior correção da estimativa com $\psi_i = \widetilde{\alpha}_{i+1}/\widetilde{\alpha}_i$ é suficiente.

Proposição 74 ***(Geometria da realização)*** *Se a sequência $x_{i+1} = \mathcal{A}(x_i), i = 0, 1, \ldots$ é q-superlinear, então, existe um $i_\alpha \geq 0$ a partir do qual ela também é q-linear, e tem-se, para todo $i \geq i_\alpha$:*

$$\frac{\beta_i}{1 + \psi_i \alpha_i} \leq \eta_i \leq \min\left\{\frac{\alpha_i}{1 - \alpha_i}, \frac{\beta_i}{1 - \psi_i \alpha_i}\right\} \tag{3.90}$$

$$\frac{1 - \psi_i \alpha_i}{1 + \alpha_i} \xi_i \leq \beta_i \leq \frac{1 + \psi_i \alpha_i}{1 - \alpha_i} \xi_i \leq \zeta_i(\psi) \tag{3.91}$$

$$\frac{1 - \alpha_i}{1 + \psi_i \alpha_i} \beta_i \leq \xi_i \leq \min\left\{\alpha_i, \frac{1 + \alpha_i}{1 - \psi_i \alpha_i} \beta_i\right\} < 1 \tag{3.92}$$

$$\max\left\{\max_{k \geq i \geq i_\alpha}\left\{\frac{|\beta_k - \xi_k|}{\beta_k + \psi_k \xi_k}\right\}, \widetilde{\alpha}, \max_{k \geq i \geq i_\alpha}\{\beta_k \delta_k(\psi)\}\right\} \leq \alpha_i < 1\ . \tag{3.93}$$

Demonstração: Segue os moldes da demonstração da Proposição 49, usando os resultados anteriores. ■

A possibilidade de $\beta_i < \alpha$ fica agora submetida às condições seguintes:

$$\beta_i \leq \frac{1}{2}\left[(2 + 4\psi_i) - \sqrt{(2 + 4\psi_i)^2 - 4}\right] \Rightarrow \beta_i \leq \alpha_i^{(-)} \leq \alpha_i \leq \alpha_i^{(+)},$$

$$\alpha_i^{(+),(-)} = \frac{1}{(2\psi_i)}\left[(1 - \beta_i) \pm \sqrt{(1 - \beta_i)^2 - 4\psi_i \beta_i}\right], \qquad \text{para } \psi_i \neq 0,$$

$$\alpha_i^{(+),(-)} = \frac{\beta_i}{1 - \beta_i} = \alpha_i, \qquad \text{se } \psi_i = 0, \tag{3.94}$$

Os seguintes resultados são imediatos a partir dos anteriores.

Proposição 75 ***(Suficiência para contração iterada)*** *Se existe algum* $i \geq i_\alpha$, *para o qual* $\alpha_i < (1/\psi_i)(-1+\sqrt{1+\psi_i})$, $\psi_i \neq 0$, *então* $x_{j+1} = \mathcal{A}(x_j)$, $j \geq i \geq i_\alpha$, *é uma contração iterada.*

Proposição 76 ***(Suficiência para majoração do erro)*** *Se existe algum* $\alpha_i \leq 1/2$, *então a sequência* $\{d(x_{i+1}, x_\star)\}_{i \geq i_\alpha}$ *é majorada por* $\{d(x_{i+1}, x_i)\}_{i \geq i_\alpha}$.

No caso de muitos enriquecimentos, obtem-se, para $p \geq 1$:

Proposição 77 ***(Estimador com p enriquecimentos)*** *Se* $x_{i+1} = \mathcal{A}(x_i)$, $i = 0, 1, \ldots$, *é q-superlinear, então, para todo* $i \geq i_\alpha$ *tem-se, para cada* $p \geq 1$ *:*

$$\frac{\beta_{i+p}-1}{1+\psi_i\alpha_i^2}\prod_{k=i}^{i+p-1}\beta_k d(x_{i+1},x_i) \leq d(x_{i+p},x_\star) \leq \left(1+\frac{\beta_{i+p}}{1-\psi_i\alpha_i}\right)\prod_{k=i}^{i+p-1}\beta_k d(x_{i+1},x_i)$$

O resultado acima complementa a Proposição 74 em cada ítem. Por exemplo, a (3.90) passa a ser:

$$\max\left\{\frac{\beta_i}{1+\psi_i\alpha_i}, \frac{\beta_{i+1}-1}{1+\psi_i\alpha_i^2}\beta_i\right\} \leq \eta_i \leq \min\left\{\frac{\alpha_i}{1-\alpha_i}, \frac{\beta_i}{1-\psi_i\alpha_i}, (1+\frac{\beta_{i+1}}{1-\psi_i\alpha_i})\beta_i\right\}$$

3.4.2 Estimativas em espaços de Hilbert

Levar em conta a q-superlinearidade nas estimativas de erros em espaços de Hilbert é essencial para a obtenção de estimativas as melhores possíveis. Nesta subseção são sumarizadas as principais relações.

A relação de q-superlinearidade que caracteriza os espaços com produto interno passa a ser:

$$\|e_i\| \leq \frac{\|s_i\|}{\sqrt{1-\alpha_i^2}}\left[-\frac{2\,(e_i,s_i)}{\|s_i\|^2}-1\right]^{1/2}. \tag{3.95}$$

Note que a condição de descendência em nada muda.

A expressão genérica para α_i torna-se:

$$\alpha_i^2 = \max_{j\geq i\geq i_\alpha}\left\{1+\frac{\|s_j\|^2}{\|e_j\|^2}+\frac{2\,(e_j,s_j)}{\|e_j\|^2}\right\} = \max_{j\geq i\geq i_\alpha}\left\{1+\frac{\|s_j\|^2}{\|e_j\|^2}+\frac{2\,\|s_j\|}{\|e_j\|}\cos\langle e_j,s_j\rangle\right\}. \tag{3.96}$$

Tudo ocorrendo como se a cada iteração $i \geq i_\alpha$ fosse redefinida uma nova realização q-linear com um novo fator-q menor ou igual ao ativo.

Projeções ortogonais

O Lema fundamental não sofre alterações. Deve ser enfatizado:

Lema 78 ***(Lema fundamental para projeções ortogonais)*** *Para a sequência* $x_{i+1} = \mathcal{A}(x_i)$, $i = 0, 1, 2, \ldots$, $x_i \in \mathbb{R}^n$, *q-linearmente e q-superlinearmente convergente na norma* $\|\cdot\|_\bullet$, *definida pelo produto interno* $(\cdot,\cdot)_\bullet$, *para o seu limite* $x_\star$, *com fator-q dado por* α , *e tal que* $(e_{i+1}, s_i)_\bullet = 0$, $s_i \neq e_{i+1} \neq 0$, *valem as seguintes cotas para todos os índices* $i \geq i_\alpha$*:*

$$\|s_i\|_\bullet \leq \|e_i\|_\bullet \leq \frac{\|s_i\|_\bullet}{(1-\alpha_i^2)^{1/2}} , \tag{3.97}$$

$$\|e_{i+1}\|_\bullet \leq \overset{\bullet}{\zeta} \|s_i\|_\bullet , \tag{3.98}$$

onde $\overset{\bullet}{\zeta_i} = \alpha_i/(1-\alpha_i^2)^{1/2}$.

Demonstração: De (3.62),

$$\|e_i\|_\bullet^2 = \|e_{i+1}\|_\bullet^2 + \|s_i\|_\bullet^2 \leq \alpha_i^2 \|e_i\|_\bullet^2 + \|s_i\|_\bullet^2 \Rightarrow \|e_i\|_\bullet \leq (1-\alpha_i^2)^{-1/2} \|s_i\|_\bullet ,$$

e,

$$\|e_{i+1}\|_\bullet \leq \alpha_i \|e_i\|_\bullet \leq \alpha_i (1-\alpha_i^2)^{-1/2} \|s_i\|_\bullet = \overset{\bullet}{\zeta_i} \|s_i\|_\bullet .$$

Que $\|s_i\|_\bullet \leq \|e_i\|_\bullet$ (e $\|s_{i+1}\|_\bullet \leq \|e_{i+1}\|_\bullet$) é consequência da Proposição 59, como antes. ■

O mais importante a ser notado é a independência *de* $\overset{\bullet}{\zeta_i}$ *de* ψ_i. Isto é consequência do fato de as projeções ortogonais, na hipótese de não existir *breakdown*, levarem naturalmente a algoritmos q-lineares e q-superlineares. Ou seja, a maneira natural de, em espaços de Hilbert, forçar q-linearidade simultânea com q-superlinearidade globais ($i_\alpha = 0$), é efetuar a pesquisa na linha. Esta pesquisa não precisa ser exata - o que é impossível, difícil ou caro em problemas não lineares em geral - mas apenas aproximada, garantindo a Proposição 59. A qualidade da convergência do quociente fica dependendo de quão descendentes são os elementos d_i da base gerada pelo algoritmo.

É importante observar entretanto que, em sistemas lineares, o fato de as bases pertencerem a subespaços de Krylov nada significam para a qualidade da convergência do quociente, senão, talvez, signifiquem um procedimento eficiente de geração nas máquinas atuais. Este fato leva a acreditar na existência de bases mais eficientes - mais descendentes, pelo menos - fora dos subespaços de Krylov. Uma primeira etapa nesta busca é dada na Parte II desta tese.

Lema 79 *($\{\beta_i\}$ é limitada)* $\beta_i \leq \eta_i \leq \overset{\bullet}{\zeta}_i$.

Demonstração: Pelo Lema 78

$$\beta_i = \frac{\|s_{i+1}\|_\bullet}{\|s_i\|_\bullet} \leq \frac{\|e_{i+1}\|_\bullet}{\|s_i\|_\bullet} = \eta_i \leq \frac{\alpha_i}{(1-\alpha_i^2)^{1/2}} = \overset{\bullet}{\zeta}_i \ .$$

■

Lema 80 $(1-\alpha_i^2)^{1/2}\xi_i \leq \beta_i$.

Demonstração: Pelo Lema 78

$$\|e_{i+1}\|_\bullet \leq \beta_i \frac{\|s_i\|_\bullet}{(1-\alpha_i^2)^{1/2}} \leq \beta_i \frac{\|e_i\|_\bullet}{(1-\alpha_i^2)^{1/2}} \Rightarrow (1-\alpha_i^2)^{1/2}\xi_i \leq \beta_i \ .$$

■

As relações básicas ficam, para todo $i \geq i_\alpha$,

$$\beta_i\delta_i \leq \beta_i\delta_i(\psi) \leq \alpha_i \leq \overset{\bullet}{\zeta}_i \ , \tag{3.99}$$

A primeira das quais fornece uma cota inferior para $\|e_i\|_\bullet$ em função do enriquecimento:

$$\beta_i\delta_i \leq \alpha \Rightarrow \|s_{i+1}\|_\bullet \leq \alpha_i \left[\psi_i\alpha_i \|e_i\|_\bullet + \|e_i\|_\bullet\right] = \alpha_i(1+\psi_i\alpha_i)\|e_i\|_\bullet \ . \tag{3.100}$$

Pela qual se conclui, mais uma vez, que $\|s_{i+1}\|_\bullet$ não é uma estimativa para $\|e_i\|_\bullet$, a não ser, eventualmente, quando $\alpha_i \leq [1/(2\psi_i)]\left[(\sqrt{1+4\psi_i}-1)\right]$, $\psi_i \neq 0$, para o qual $\alpha_i(1+\psi_i\alpha_i) \leq 1$. Também, $\|s_{i+1}\|_\bullet$ não o é para $\|e_{i+1}\|_\bullet$ pela Proposição 59, a não ser assintoticamente, já que o algoritmo é suposto q-superlinear. De fato, embora $\|s_{i+1}\|_\bullet$ nunca possa majorar $\|e_{i+1}\|_\bullet$, na fase final q-superlinear a sua proximidade com $\|e_{i+1}\|_\bullet$ é enorme. Esta prática não deve ser usada isoladamente, já que este trabalho fornece estimativas de cotas superiores assintoticamente exatas, em particular em problemas mal condicionados. Vide Proposição 83.

Para facilidade de referência, cataloga-se abaixo os demais resultados que levam em conta a q-superlinearidade da ortogonalizações, embora a maioria deles não mude.

Proposição 81 *(Cota inferior para α_i) Defina* $\overset{\bullet}{\alpha}_i = \beta_i/(1+\beta_i^2)^{1/2}$, *e* $\overset{\bullet}{\alpha} = \max_{k\geq i\geq i_\alpha}\left\{\overset{\bullet}{\alpha}_k\right\}$, *para todo* $i \geq i_\alpha$, *então,*

$$\overset{\bullet}{\alpha} \leq \beta_i, \qquad \forall i \geq i_\alpha \ ,$$

$$\overset{\bullet}{\alpha}_i \leq \alpha_j \ , \qquad i_\alpha \leq j \leq i \ .$$

Demonstração: Na linha da Proposição 73. ■

Proposição 82 ***(Geometria das projeções ortogonais q-superlineares)*** *Se a sequência* $x_{i+1} = \mathcal{A}(x_i), i = 0, 1, \ldots$ *é q-superlinear, então, para* $i \geq i_\alpha$ *tem-se:*

$$\beta_i \leq \eta_i \leq \min\left\{\frac{\alpha_i}{(1-\alpha_i^2)^{1/2}}, \frac{\beta_i}{(1-\alpha_i^2)^{1/2}}\right\} \tag{3.101}$$

$$(1-\alpha_i^2)^{1/2}\xi_i \leq \beta_i \leq \frac{\xi_i}{(1-\alpha_i^2)^{1/2}} \leq \dot{\zeta} \tag{3.102}$$

$$(1-\alpha_i^2)^{1/2}\beta_i \leq \xi_i \leq \min\left\{\alpha_i, \frac{\beta_i}{(1-\alpha_i^2)^{1/2}}\right\} < 1 \tag{3.103}$$

$$\max\left\{\max_{k\geq i\geq i_\alpha}\left\{\frac{|\beta_k^2 - \xi_k^2|^{1/2}}{\xi_k}\right\}, \dot{\alpha}, \max_{k\geq i\geq i_\alpha}\{\beta_k\delta_k(\psi)\}\right\} \leq \alpha_i < 1 \tag{3.104}$$

Demonstração: Imediata a partir da Proposição 74 e dos Lemas anteriores. ■

Proposição 83 $\|e_i\|_\bullet^2 \geq (1+\beta_i^2)\|s_i\|_\bullet^2$ *e* $\|e_{i+1}\|_\bullet^2 \geq (1+\beta_{i+1}^2)\beta_i^2\|s_i\|_\bullet^2$

Demonstração: $\|e_i\|_\bullet^2 = \|s_i\|_\bullet^2 + \|e_{i+1}\|_\bullet^2 \geq \|s_i\|_\bullet^2 + \|s_{i+1}\|_\bullet^2 = (1+\beta_i^2)\|s_i\|_\bullet^2$, e a outra por indução. ■

Proposição 84 ***(Estimador para p enriquecimentos)*** *Se* $x_{i+1} = \mathcal{A}(x_i)$, $i = 0, 1, \ldots$, *é q-superlinear, então, para* $i \geq i_\alpha$ *tem-se, para cada* $p \geq 1$ *:*

$$(1+\beta_{i+p}^2)^{1/2}\prod_{k=i}^{i+p-1}\beta_k\|s_i\|_\bullet \leq \|e_{i+p}\|_\bullet \leq \left(1+\frac{\beta_{i+p}^2}{1-\psi_i\alpha_i^2}\right)^{1/2}\prod_{k=i}^{i+p-1}\beta_k\|s_i\|_\bullet \ . \tag{3.105}$$

Note que $\|e_i\|_\bullet^2 \geq \|s_i\|_\bullet^2 + \|s_{i+1}\|_\bullet^2$.

3.5 Convergência q-sublinear

A convergência q-sublinear é típica de algoritmos lentos, ou na fase inicial de bons algoritmos. Para sequências estritamente q-sublineares nada mais se pode acrescentar senão a de estimar os erros de algum mapeamento composto $\mathcal{A}^p$, $p \geq 2$, ou alguma combinação mais sofisticada, seguida de uma estimativa q-linear nos moldes normais. A experiência numérica tem demonstrado que esta técnica também funciona para sublinearidades mais fracas. O resultado chave é a Proposição 27.

Para sequências com convergência r-linear aplica-se as Proposições 29 ou 30. Ou seja, novamente a aplicação de mapeamentos compostos.

3.5.1 Estimativas da realização

Para sequências apenas q-sublineares (não estritas) é possível que um mapeamento composto que possa fornecer um bom resultado usando a teoria deste capítulo exija iterações demais. Passa-se a expor, usando uma técnica de remetrizacão, os motivos para isso. Entretanto, a necessidade de um número grande de iterações é um problema do algoritmo, e não da teoria da convergência do quociente.

Seja $\|\cdot\|_{bp}$ a norma de interesse que converge q-sublinearmente (para outros tipos de sublinearidades, proceder de forma similar)

$$\|e_i\|_{bp} \leq ci^{-p}$$

onde $p > 1$. Tomando o logaritmo em uma base b dessa desigualdade

$$\log \|e_i\|_{bp} \leq \log c - p \log i.$$

Isto é, o logaritmo na base b do erro está limitado por uma reta no gráfico logarítmico. Agora, compare com a definição de normas q-lineares, para $i_\alpha = 0$:

$$\|e_{i+1}\| \leq \alpha \|e_i\| \Rightarrow \|e_i\| \leq \alpha^i \|e_0\| \Rightarrow \log \|e_i\| \leq \log \|e_o\| + \log \alpha \cdot i \ .$$

Então, toda a teoria desenvolvida para sequências q-lineares aplica-se às sequências q-sublineares com $p > 1$, mudando a escala do gráfico das iterações para uma escala logarítmica. Por exemplo, escolhendo $b = e$ (neperiano) deve-se avaliar o erro nas iterações $i = 3, 8, 21, \ldots$, considerando-se os enriquecimentos $x_3 - x_0$, $x_8 - x_3$, $x_{21} - x_8$, ... Isto corresponde a considerar a sequência q-linearmente convergente em p_k-etapas, onde $\{p_k\} = \{0, 3, 8, 21, \ldots\}$.

Existem outras alternativas. Uma opção segura é determinar adaptativamente a sequência $\{p_k\}$. Cada termo desta sequência é avaliada como acima, porém, em lugar de uma escolha prévia de b, calcula-se o correspondente

$$\beta_k = \left\|x_{p_{k+2}} - x_{p_{k+1}}\right\| / \left\|x_{p_{k+1}} - x_{p_k}\right\|,$$

com o termo p_{k+2} escolhido tal que β_k satisfaça a condição (3.76) - $\beta_k \leq 0.5$, ou a equivalente para medidas quaisquer. Em seguida, usa-se $\alpha_k^{(+)}$ como uma cota superior perfeitamente confiável para α. A estimativa é confiável porque ou $\beta_k \leq \alpha_k^{(-)} \leq \alpha \leq \alpha_k^{(+)}$, ou $\alpha \leq \beta_k \leq \alpha_k^{(-)} \leq \alpha_k^{(+)}$. Em geral não há necessidade de esperar por $\beta_k \leq 0.5$. Experimentos numéricos com diversos algoritmos para problemas lineares indicam que é suficiente $\beta_k < 1$, e o uso de $\tilde{\alpha}$ (ou $\overset{\bullet}{\alpha}$).

A alternativa que se tem mostrado bastante interessante é o uso direto da Proposição 68, ou a sua equivalente em espaços métricos.

3.5.2 Outros resultados

O propósito desta subseção é estabelecer a equivalência entre a convergência q-sublinear e a convergência sublinear de grau $p > 0$, como proposta no Capítulo 2. O que se segue é uma extensão dos resultados obtidos em [36]. Sem perda de generalidade, podemos escolher constantes α_i e c_i tais que

$$\|e_{i+1}\| = \alpha_i \|e_i\|, \quad \alpha_i \to 1 \quad \text{se } q\text{-sublinear, e} \tag{3.106}$$

$$\|e_i\| = c_i i^{-p}, \quad c_i \to c > 0 \quad \text{se sublinear de grau } p. \tag{3.107}$$

Proposição 85 ***(Equivalência)*** *A convergência sublinear de grau p satisfaz:*

1. $\lim \alpha_i = \lim \frac{\|e_{i+1}\|}{\|e_i\|} = 1$,
2. $\lim \frac{\|s_i\|}{\|e_i\|} = 0$,

sendo, portanto, equivalente à convergência q-sublinear.

Demonstração: Escrevendo o erro na iteração $i+1$, e desenvolvendo em série de Taylor:

$$\begin{aligned}\|e_{i+1}\| &= c_{i+1}(i+1)^{-p} = c_{i+1}\left[i^{-p}(1+i^{-1})^{-p}\right] = c_{i+1}i^{-p}\left\{1-[p+o(1)]\,i^{-1}\right\} \\ &= c_i i^{-p}\frac{c_{i+1}}{c_i}\left\{1-[p+o(1)]\,i^{-1}\right\} = \|e_i\|\frac{c_{i+1}}{c_i}\left\{1-[p+o(1)]\,i^{-1}\right\},\end{aligned}$$

ou seja,

$$\alpha_i = \frac{\|e_{i+1}\|}{\|e_i\|} = \frac{c_{i+1}}{c_i}\left[1-\frac{p+o(1)}{i}\right]. \tag{3.108}$$

Tomando o limite quando $i \to \infty$ desta última expressão obtém-se:

$$\lim \alpha_i = \lim\left\{\frac{c_{i+1}}{c_i}\left[1-\frac{p+o(1)}{i}\right]\right\} = 1,$$

provando o ítem 1. Para provar o ítem 2,

$$\begin{aligned}-s_i &= e_i - e_{i+1} = c_i i^{-p} - c_{i+1}(i+1)^{-p} = \left\{c_i - c_{i+1}\left[1-\frac{p+o(1)}{i}\right]\right\} i^{-p} \\ &= \left\{1-\frac{c_{i+1}}{c_i}\left[1-\frac{p+o(1)}{i}\right]\right\} c_i i^{-p} = \left\{1-\frac{c_{i+1}}{c_i}\left[1-\frac{p+o(1)}{i}\right]\right\} e_i\end{aligned}$$

pela qual resulta

$$\frac{\|s_i\|}{\|e_i\|} \leq \left|1-\frac{c_{i+1}}{c_i}\left[1-\frac{p+o(1)}{i}\right]\right|$$

e então

$$\lim \frac{\|s_i\|}{\|e_i\|} = 0\,,$$

provando o ítem 2 e a equivalência. A equivalência em espaços de Hilbert resulta da substituição do limite acima na condição de q-sublinearidade dada pela expressão (3.50). ■

3.6 Aproximações do Operador

Quando a determinação dos valores de α, ou de suas estimativas, é feita durante a realização, os erros de aproximação adicionais, como por exemplo os devido à aritmética finita, ou às aproximações sobre o mapeamento $\mathcal{A}$ (incluindo as devido à discretização, quando for o caso) já estão sendo levados em conta.

Este não é o caso quando são usados valores estabelecidos analiticamente. Nessa situação, as aproximações adicionais influenciam a avaliação do erros durante a realização. Em casos extremos, algoritmos que convergem q-linearmente, por exemplo, podem simplesmente não convergir por imprecisões diversas. Certamente é bastante útil estabelecer a influência das diversas imprecisões.

Uma forma de por a questão é considerar que, devido às várias aproximações, em lugar de uma sequência $\{x_i\}$ obtém-se uma sequência $\{y_i\}$. Supondo que a sequência $\{x_i\}$ converge q-superlinearmente, o que se pode dizer a respeito da convergência da sequência $\{y_i\}$? O seguinte resultado ajuda bastante.

Proposição 86 *Seja $\mathcal{A} : D \subset \mathbb{R}^n \to \mathbb{R}^n$ um algoritmo q-superlinear (o processo) em $D_1 \subset D$ com constantes $\{\alpha_i\}$, e D_o um conjunto fechado tal que $\mathcal{A}(D_0) \subset D_0$. Seja $\{y_i\} \subset D_1$ uma sequência produzida pelo algoritmo, e $\epsilon_i = d(\mathcal{A}(y_i), y_{i+1})$, $i \geq i_\alpha \geq 0$. Seja $x_\star$ o limite da sequência (teórica) $\{x_i\}$. Então,*

$$d(y_{i+1}, x_\star) \leq \frac{1}{1 - \psi_i \alpha_i} \left[\beta_i d(y_{i+1}, y_i) + \epsilon_i\right] . \tag{3.109}$$

Demonstração: A estimativa vem de

$$\begin{aligned} d(y_{i+1}, x_\star) &\leq d(y_{i+1}, \mathcal{A}(y_i)) + d(\mathcal{A}(y_i), \mathcal{A}(y_{i+1})) + d(\mathcal{A}(y_{i+1}), x_\star) \\ &\leq \epsilon_i + \beta_i d(y_{i+1}, y_i) + \alpha_{i+1} d(y_{i+1}, x_\star) , \end{aligned}$$

e (3.109) segue-se. ■

Note que ambos, α_i e β_i são os valores teóricos (analíticos). Assintoticamente,

$$\alpha_i \to 0 \Rightarrow \beta_i \leq \frac{\alpha_i(1 + \psi_i \alpha_i)}{1 - \alpha_i} \to 0$$

e tem-se

$$d(y_{i+1}, x_*) \leq \epsilon_i \ .$$

A principal conclusão é que em algoritmos q-superlineares os erros de aproximação adicionais ficam, assintoticamente, restritos ao da própria iteração. Eles não são acumulativos. A única exigência é que $\{y_i\} \subset D_1$, o conjunto no qual o algoritmo é de fato q-superlinear. Para tais algoritmos pode-se de obter uma precisão tão boa quanto se queira. Não existe portanto nenhuma razão para comparar a qualidade da aproximação de algoritmos q-superlineares com a dos algoritmos diretos (Gauss, por exemplo).

Outra conclusão é que, em se desejando grandes precisões, deve-se realizar a iteração de parada com maior precisão. Por exemplo, alguma reortogonalização parcial.

Para algoritmos assintoticamente q-lineares a (3.109) prevalece com $\alpha_i = \alpha$. O máximo que se pode dizer é que a parada deve ser feita quando ocorrer um pequeno valor para β_i. Isto não é tão preciso já que β_i é o teórico.

Os algoritmos q-sublineares $\psi_i \alpha_i \to 1$. Por (3.109), a qualidade da resposta é muito sensível a aproximações do operador. Isto justifica um cuidado maior com os algoritmos apenas q-sublineares.

3.7 Comentários Finais

Ao fim deste capítulo parece bastante adequado fazer alguns comentários adicionais sobre teoria de convergência do quociente nele desenvolvida. Estes comentários serão feitos dentro de três perspectivas, focando sistemas lineares: a contribuição para a compreensão dos mecanismos de convergência de algoritmos iterativos, a qualidade das estimativas de erros, e os trabalhos existentes na literatura com eles relacionados.

3.7.1 Os mecanismos da convergência

O autor expõe a seguinte heurística, em paralelo com uma devida a Nevanllina [24, cap. 1]. Um algoritmo possui três fases ao longo da sua realização.

A primeira fase desenvolve-se, em geral, com convergência q-sublinear. Quando o algoritmo inicia não há nenhuma informação sobre o sistema. Nesta fase o algoritmo deve aprender sobre o operador. Enquanto o aprendizado se realiza deve-se esperar baixas velocidades de convergência. Durante esta fase não há como distinguir entre

problemas bem ou mal comportados, prever o número de iterações para uma dada redução percentual do erro, ou mesmo determinar se haverá convergência ou não. Um bom método iterativo deve extrair informações do operador durante esta fase e armazená-las para as iterações futuras. Se o problema é autoadjunto, a curvatura constante do espaço é, otimamente, a única informação a ser transmitida, e a segunda fase com convergência q-linear é imediatamente iniciada, embora, eventualmente, com baixas velocidades.

Foi visto neste capítulo que em espaços de Hilbert a convergência q-linear (e mesmo q-superlinear) pode ser conseguida desde o início das iterações com a ortogonalização. O algoritmo não distingue entre primeira e segunda fases. Em qualquer caso, a segunda fase é essencialmente caracterizada pela convergência q-linear. Nesta fase, as componentes de resposta na base gerada pelo algoritmo são gradativamente construidas. Entretanto, para sistemas mal comportados, ela prolonga-se demasiadamente. Em particular quando prevalecem erros devidos à precisão finita, tipicamente: erros de aproximação do mapeamento teórico $\mathcal{A}$ e cancelamento aritmético. No final desta fase, a resposta está essencialmente correta: ambos, norma e direção (no espaço em que a ortogonalização é efetuada), convergiram em uma precisão aceitável.

A terceira fase deve caracterizar-se por convergência q-superlinear. Mais uma vez, a ortogonalização em espaços de Hilbert antecipa esta fase. Ela terminará com convergência finita se o problema possui dimensão finita e, simultaneamente, a ortogonalização é efetuada. Nesta fase, todas as medidas não priorizadas pela ortogonalização são levadas à precisão permitida pela aritmética finita efetuada na iteração corrente.

Não há qualquer dúvida para o autor que algoritmos competitivos devam ser no mínimo q-lineares, e idealmente q-superlineares, na norma de interesse do analista.

A importância da ortogonalização é enfatizada, em particular, como uma simples aplicação da teoria, foram estabelecidas as condições necessárias e suficientes para a q-convergência. Como resultado imediato, tendo em vista o grande interesse em diversas áreas da computação científica, demonstrou-se a convergência global q-linear e q-superlinear do método dos gradientes conjugados padrão para operadores autoadjuntos, independentemente das propriedades do operador. O autor não encontrou essa prova na literatura.

3.7.2 A qualidade dos estimadores

As melhores estimativas para um determinado algoritmo somente podem ser obtidas usando as suas propriedades inerentes. Para os métodos de ortogonalização em espaços de Hilbert as estimativas ora apresentadas não podem ser melhoradas. Isto tem sido confirmado com inúmeros experimentos numéricos, alguns dos quais descritos nesta tese. Isto é muito gratificante, uma vez que, essencialmente, todos os métodos bem sucedidos para sistemas lineares resultam de ortogonalização em alguma norma, embora esta norma muitas vezes não tenha qualquer significado, e, normalmente não seja a de interesse do analista.

As estimativas apresentadas para as normas que não são as ortogonalizadas podem, eventualmente, ser melhoradas, para cada algoritmo em particular. Parece ao autor, que a teoria apresentada fornece elementos suficientes para esta particularização.

Do ponto de vista de estimativa de erros, a noção de q-superlinearidade usada nesta tese prescinde de noções de velocidades supostamente maiores, como velocidades q-quadráticas, convergência exponencial, etc.

Existe uma deficiência na qualidade dos estimadores quando o algoritmo é pobre o suficiente para convergir apenas de forma q-sublinear. Alguns destes algoritmos ainda são usados em muitas disciplinas. Acredita-se que isto caracteriza um estágio de evolução ainda inicial dos algoritmos nessas disciplinas. Para a q-sublinearidade estrita os resultados são suficientemente bons. Entretanto, o autor reconhece sua insatisfação com as soluções apresentadas para as demais formas de q-sublinearidades, e, presentemente, desenvolve uma teoria de convergência q-sublinear que não se apoia em mapeamentos compostos, e sim um pouco mais em remetrizações.

A qualidade dos estimadores não depende de avaliações de propriedades algébricas dos operadores, e não é afetada pelo número de condição do operador. Todas as avaliações necessárias são feitas durante a realização, e são independentes do algoritmo em questão. Porém permanece pelo menos uma questão aberta. Foram fornecidas várias cotas inferiores para o fator-q da realização, para todos os casos de interesse. É possível determinar cotas superiores para o fator-q da realização de uma forma tão geral quanto as cotas inferiores estabelecidas ? O autor pensa que essa é a principal questão a ser respondida no âmbito da teoria apresentada. Não há dúvida que é possível estabelecer estas cotas superiores para cada algoritmo. Por exemplo, para o método de Newton, sob condições padrões, o valor $\alpha \leq 0.5$ já foi estabelecido para a norma-ℓ_2 do erro, ver [19]. Embora, como o método de Newton

é contractivo nessa norma, a teoria exposta nesta tese diga que $\alpha < \sqrt{2}-1 \simeq 0.414$.

3.7.3 Trabalhos relacionados

A literatura aborda algumas metodologias distintas para a estimativa de erros de aproximação em algoritmos iterativos em espaços de dimensões finitas.

A mais tradicional é parar com alguma norma computável do resíduo. São essencialmente baseadas em um teorema de Oettli e Prager [40]. Detalhes adicionais podem ser encontrados em [41], [31] e [34]. A técnica consiste em computar a norma do erro (*forward error*) por uma análise retroativa (*backward error*) de uma norma ℓ_p do resíduo. Isto sempre resulta na estimativa de normas ou números de condição de operadores. Parte desta tese é justamente o resultado de uma insatisfação com tais critérios. Na COPPE, pode-se citar pelo menos [42], onde são feitos experimentos numéricos com o GMRES em problemas de Mecânica dos Sólidos discretizados pelo Método dos Elementos de Contorno, objetivando o estudo da influência de certos precondicionadores na redução de algumas normas-ℓ_p.

Desde um longo tempo atrás (1971), o Prof. Golub e seus colaboradores iniciaram trabalhos no sentido de obter melhores estimativas do que as fornecidas pela técnica tradicional. O ponto de partida de sua teoria é expressar a norma da energia do erro em termos de certas integrais de Riemman-Stieltjes, no que é chamado de método dos momentos, e suas relações com a quadratura de Gauss. Permanece, entre outras coisas, a necessidade de estimar o erro inicial, que é resolvido, nos últimos trabalhos, com o uso de frações continuadas. Nesta linha inclui-se [43],[44] e [14]. Apenas a norma da energia em sistemas autoadjuntos é contemplada, e sua avaliação é dependente da avaliação das melhores cotas possíveis para os autovalores extremos. A qualidade da avaliação do erro inicial por frações continuadas é dependente do número de condição da matriz dos coeficientes.

Em nenhum dos trabalhos de conhecimento do autor as estimativas fornecidas aplicam-se a erros em medidas quaisquer, ou mesmo a normas quaisquer. Sempre limitam-se a normas de energia. Além disso, nenhuma das técnicas extendem-se a problemas não lineares, muito menos a algoritmos iterativos em geral, e sempre dependem de estimativas de propriedades dos operadores, tais como números de condição e norma.

Parte II

Estratégias para Sistemas Lineares

I recommend this modus operandi. You will hardly eliminate directly anymore, at least not when you have more than two unknowns. The indirect method can be pursued half while asleep or while thinking about other things.

Gauss, 1823[11, pag. 158].

Durante o desenvolvimento do estudo da convergência do quociente na primeira parte desta tese, inúmeros aspectos construtivos concernentes à qualidade de convergência de algoritmos iterativos ficaram evidenciados. O objetivo desta Parte II é relatar o resultado da investigação teórica e numérica de alguns daqueles aspectos que se mostraram mais relevantes, no que diz respeito à solução iterativa de sistemas lineares. A extensão para sistemas não lineares é imediata, e será relatada em trabalho posterior. São tratados apenas sistemas reais de solução única e não trivial.

Para efetuar esta tarefa, introduz-se um modelo bastante geral de algoritmos para sistemas lineares que, sob certas restrições mínimas, é consistente com o problema (2.3).

Ficou evidenciado pelo estudo precedente que é essencial para a estabilidade do algoritmo manter a convergência q-linear na fase inicial, e que isto pode ser conseguido por ortogonalização. A norma ideal para isto é sem dúvida a norma-ℓ_2, certamente a mais bem comportada norma em espaços Euclidianos. Esta ortogonalização garante uma separação mínima entre todas as demais normas. Infelizmente, não existem muitas avenidas para ir em busca desta ortogonalização.

A dificuldade está na impossibilidade de computar o produto interno $(e_{i+1}, s_i)_2$, indispensável ao processo de ortogonalização. Em se tratando de bases de subespaços de Krylov em sistemas lineares Hermitianos, todas as possibilidades para ortogonalizações nas diversas normas ponderadas já são conhecidas [18]. O caminho encontrado na literatura para ortogonalizar na norma-ℓ_2 consiste em usar uma direção de busca sob a forma Bd_i. Opções que têm sido usadas são $B = A$, $B = A^T$, etc. Dois algoritmos elegantes são descritos em [17]. Assim, garantindo a computabilidade de $(e_{i+1}, s_i)_2$, porém às custas de reintroduzir a questão da separação. Nesta tese, como resultado da teoria da convergência do quociente, são propostas alternativas. O primeiro passo nesta direção consiste em dar interpretações adequadas ao sempre presente incremento s_{i+1}.

A importância da q-linearidade desde a fase inicial está estreitamente ligada à garantia da convergência global e da estabilidade do algoritmo. Não apenas

isso. Sendo as iterações descendentes em uma norma conhecida, pode-se extrair informações úteis para as nova iterações. A otimização do processo de hereditariedade juntamente com a descendência na norma-ℓ_2 são, na opinião do autor, os ingredientes básicos da receita do sucesso do método dos gradientes conjugados para problemas Hermitianos em computação científica. A qualidade das informações disponíveis em cada iteração é proporcional à qualidade da convergência. É essencial eliminar a fase inicial de baixa qualidade de convergência do quociente, antecipando a segunda fase q-linear. Isto é, de fato, a função clássica do precondicionador.

Para estabelecer um processo de transmissão de informações entre as iterações é necessário estabelecer um mecanismo eficiente e seletivo. Por eficiente entenda-se a otimização do uso de memória auxiliar e de operações, e por seletivo, a capacidade de extrair e passar adiante apenas informações úteis. Neste trabalho, para efetuar esta tarefa faz-se uso, principalmente, do precondicionador e de algoritmos bem estabelecidos na literatura.

São usadas ferramentas auxiliares de diversas áreas, notadamente, da teoria métrica dos pontos fixos, das noções de distâncias generalizadas, em especial das funções de Bregman e de suas projeções generalizadas, e distâncias epigráficas. As noções ainda não revisadas serão introduzidas gradativamente à medida que se fizerem necessárias.

Capítulo 4

Estratégias Correntes

We move from direct methods, a by now classical topic that is rather thoroughly understood, to the relatively untamed territory of iterative methods. These are the methods that seem likely to dominate the large-scale computations of the future.

Trefethen e Bau, 1997[32].

4.1 Introdução

Esta parte da tese considera o problema (2.3) linear com solução $x_\star$ única, ou seja,

$$f(x) = Ax - b\ , \qquad f(x_\star) = 0\ , \tag{4.1}$$

onde $A \in L(\mathbb{R}^n)$ é regular e $x, b \in \mathbb{R}^n$. Entretanto, perde-se muito ao discutir as estratégias correntes com matrizes reais, e assim supõe-se que $A \in L(\mathbb{C}^n)$. (Muito devido ao fato de que o espectro de uma matriz real não simétrica está potencialmente em $\mathbb{C}$). São propostas algumas estratégias de solução que decorrem naturalmente da teoria de convergência do quociente desenvolvida nos capítulos anteriores. A motivação para o desenvolvimento destas estratégias está em dois aspectos relacionados. O primeiro na solução do problema de separação entre a norma minimizada (quando há uma) durante o processo iterativo e as normas de interesse, e o segundo, tão importante quanto, na diminuição do número de iterações necessários para vencer a fase inicial q-sublinear, algumas vezes r-linear, ou fracamente q-linear encontrada em virtualmente todos os algoritmos que constroem a solução em subespaços de Krylov. Acredita-se que estes são os dois problemas centrais nas questões de robustez e escalabilidade do algoritmo iterativo, uma vez tratada nos capítulos precedentes a questão fundamental de estimativa de erros.

O propósito deste capítulo é rever as estratégias de solução correntes, situar o problema e mostrar exatamente onde as estratégias propostas se inserem, antecipando os conceitos.

4.2 A Essência das Estratégias Correntes

As estratégias para solução de sistemas lineares possuem, essencialmente, três ingredientes:

1. Um algoritmo para a geração de uma base para $\mathbb{C}^n$.
2. Uma função distância ϕ.
3. Um precondicionador H.

A escolha do algoritmo de geração da base é orientada, principalmente, pelas questões de eficiência numérica, geralmente associadas à manutenção de alguma ortogonalização da base, preferencialmente implícita.

A função ϕ tem o papel fundamental de garantir a convergência global da estratégia e, normalmente, é escolhida como uma norma elíptica. É essencial que ϕ seja duas vezes diferenciável, e sua Hessiana seja positiva definida (ou negativa definida) em todo o domínio com exceção, talvez, na solução. Invariavelmente, a função ϕ é uma norma definida por um produto interno que é usado explícita ou implicitamente para o processo de ortogonalização.

O precondicionador tem o papel de transformar o sistema em um problema numericamente bem posto, contribuindo para que o algoritmo de geração das bases seja numericamente estável, e a função ϕ bem comportada.

Subjacente a cada um dos três componentes das estratégias, estão as características intrínsecas ao sistema linear e à arquitetura do sistema computacional. O sucesso da solução iterativa depende fortemente da escolha adequada da estratégia e do aproveitamento das características do sistema linear e do sistema computacional. Esta dependência mútua é uma necessidade para a eficiência da solução numérica de qualquer problema em computação científica. No entanto, estes aspectos são todos relevantes se, e somente se, o sistema linear é suficientemente grande. Por suficientemente grande entende-se que passam a ser um gargalo à eficiência, quer a memória necessária para se chegar a um resposta com qualidade aceitável, quer o tempo de execução (ou o *wall clock*), ou ambos.

Nas próximas seções são discutidos cada um dos tópicos acima, nos aspectos relevantes à esta tese. Os enfoques dados, a não ser quando explicitamente observado, não são necessariamente compartilhados pela literatura especializada.

4.3 Algoritmos para Geração da Base

Os subespaços escolhidos para descrever a resposta, com poucas exceções, são os subespaços de Krylov $\mathcal{K}^m = \mathcal{K}^m(A, f_0)$ definido por

$$\mathcal{K}^m(A, f_0) = \text{span}\left\{f_0, Af_0, A^2 f_0, \ldots, A^{m-1} f_0\right\} , \tag{4.2}$$

onde $f_0 = b - Ax_0$. Quando (4.2) é escolhido, os métodos são chamados de métodos de Krylov. A seleção dos algoritmos para a geração da base para estes subespaços depende fundamentalmente das propriedades da matriz dos coeficientes.

4.3.1 Sistemas Hermitianos

Se A é Hermitiana e positiva definida a escolha normalmente recai sobre a geração de bases A-ortogonais, ou conjugadas, com a escolha simultânea da norma de energia $\phi = \frac{1}{2}\|e\|_A^2$. Ou ainda $\phi = \frac{1}{2}\|e\|_{A^H A}^2$ com bases $A^H A$-ortogonais. Ambas dão origem ao método dos gradientes conjugados (CG). A escolha ideal para o CG, que seria $\phi = \frac{1}{2}\|e\|_2^2$, torna o processo de busca unidimensional não computável.

O processo de minimização corresponde a um método de Galerkin, onde o subespaço de restrições $\mathcal{L}^m$ é o mesmo subespaço de Krylov usado para gerar os elementos da base quando a escolha é minimizar a norma da energia, ou seja, $\mathcal{L}^m = \mathcal{K}^m(A, f_0)$. No caso geral $\mathcal{K}^m(B, f_0) = \mathcal{L}^m(B, f_0)$ para algum B adequado.

O que caracteriza o CG é que os elementos da base são gerados independentemente um dos outros, tal que o processo de conjugação fica numericamente implícito e econômico. Quando as propriedades da matriz dos coeficientes não são numericamente favoráveis, a conjugação é destruida sistematicamente à medida em que o algoritmo avança. Isto resulta no prolongamento da fase inicial e eventualmente na perda da estabilidade do algoritmo.

Na linguagem da primeira parte desta tese (ver a Proposição 86) o valor de $\epsilon = \|\mathcal{A}(y_i) - y_i\|$, onde $\{y_i\}$ é a sequência aproximada, torna-se grande o suficiente para extinguir a q-linearidade do algoritmo. Nos casos menos críticos, observa-se que a reortogonalização parcial pode sustentar a estabilidade do CG. Em outros

casos, alguns têm preferido gerar a base com o algoritmo de Lanczos, matematicamente equivalente, em particular quando as propriedades espectrais da matriz dos coeficientes são bastante desfavoráveis [45]. Em todos estes casos, é fundamental a escolha do precondicionador também como um elemento de estabilização numérica. O processo de reortogonalização no CG, ou de manutenção dos elementos da base do método de Lanczos, total ou parcialmente, podem ser vistos como uma complementação ao precondicionamento deficiente, ou ineficaz.

Para uma classificação sistemática desses métodos vide [18].

4.3.2 Não Hermitianos (I)

Quando o sistema não é Hermitiano, o método de escolha não é tão claro. As características da matriz dos coeficientes passam a exercer um papel mais relevante. Porém, com raras exceções, sempre recai-se na construção de bases para subespaços de Krylov. A dificuldade reside no fato de que, em geral, não é mais possível fazer as ortogonalizações dos elementos da base implicitamente. Em outras palavras, não é mais possível gerar cada elemento da base com recorrências curtas. É necessário, em geral, olhar para trás em um número de iterações que depende das propriedades de A.

É possível ainda escolher o próprio subespaço $\mathcal{K}^m(A, f_0)$ como o subespaço de restrições. O FOM - *full orthogonalization method* - é o mais conhecido. Corresponde ao método de Galerkin.

A função ϕ normalmente é escolhida como sendo a norma-ℓ_2 dos resíduos, isto é, $\phi = \frac{1}{2}\|f\|_2^2 = \frac{1}{2}\|e\|_{A^HA}^2$. Corresponde ao método de Petrov-Galerkin com a escolha de $\mathcal{L}^m = A\mathcal{K}^m(A, f_0)$ como o subespaço de restrições. Exemplos mais conhecidos desta classe de algoritmos são: o CR; o GMRES e suas variantes: GMRES(m), DGMRES, FGMRES.

A geração de bases para os métodos tipo-Galerkin são geralmente feitas com o algoritmo de Arnoldi sobre uma ortogonalização estável, como Gram-Schmidt modificada ou Householder. O valor mínimo de m, a dimensão do subespaço de Krylov a ser gerado, necessário para garantir convergência, conforme observado anteriormente, depende das características da matriz dos coeficientes.

Nesses métodos de projeção ortogonal, é possível construir recorrências curtas sob certas condições, esclarecidas pelo Teorema 88, de Faber e Manteuffel (ver adiante). O Teorema de Faber e Manteuffel estabelece as condições necessárias e suficientes para que recorrências curtas existam.

Para esta classe de métodos normalmente é possível estabelecer as condições de convergência, algumas vezes as necessárias e suficientes.

4.3.3 Não Hermitianos (II)

Uma outra classe de métodos usa como restrições o subespaço $\mathcal{L}^m(A^H, \widehat{f}_0)$, onde $\widehat{f}_0$ é um vetor arbitrário. Na sua grande maioria são equivalentes ao método de Petrov-Galerkin com a solução desenvolvida em $\mathcal{K}^m(A, f_0)$. Alguns desses métodos minimizam $\phi = \frac{1}{2}\|e\|_B^2$, onde B está relacionada com a escolha da base para $\mathcal{L}^m(A^H, \widehat{f}_0)$. Os métodos desta classe caracterizam-se por possuirem recorrências curtas. O processo de bi-ortogonalização de Lanczos é, em geral, o método de geração da base. Neste caso, sempre escolhe-se $\widehat{f}_0 = f_0$.

Exemplos são: o BiCG, cujas aproximações são ótimas na norma-B, com $B = V_n^{-1}V_n^{-1}$, e o QMR, ótimo com $B = A^H V_n^{-H} V_n^{-1} A$, onde as colunas de V_n são os vetores bi-conjugados. Outros exemplos são: TFQMR, CGS, Bi-CGSTAB.

Um inconveniente é que são mais suscetíveis a *breakdowns* de diversos tipos. Normalmente não é possível extender sua base para ortogonalizar em subespaços de dimensões maiores do que aquela para os quais foram projetados, e assim evitar alguns tipos de *breakdowns*.

A escolha dentre tais métodos sempre é baseada em experiência acumulada com as classe específicas de problemas. Têm sido usados com sucesso como ciclos internos de alguns problemas não lineares em áreas as mais diversas como Mecânica dos Fluidos e Circuitos Digitais.

Os diversos tipos de *breakdown* dos métodos de recorrência curta que trabalham sobre $\mathcal{L}^m(A^H, \widehat{f}_0)$ podem-se ser interpretados como a perda da positividade da matriz B em $\phi = \frac{1}{2}\|e\|_B^2$, que normalmente não é conhecida *a priori*, tirando de ϕ as propriedades de função distância. As propriedades de ϕ enquanto métrica é que permitem garantir a convergência global da estratégia. Tais métodos são chamados de métodos de projeções não ortogonais.

O Teorema de Faber e Manteuffel não inclui esta classe de métodos. A norma minimizada (quando existe uma) é dependente do vetor inicial f_0 - primeira coluna de V_n no BiCG e no QMR - e portanto de x_0. Tais métodos caracterizam os métodos tipo-CG de métrica variável, em contraposição aos métodos para os quais B é independente do vetor inicial, como o GMRES, caracterizado como do tipo-CG de métrica fixa.

Usando esse conceito métrica variável, Barth e Manteuffel [46] propõem uma ge-

neralização do BiCG e do QMR, o BiCG(m) e o QMR(m), que permitem extender as bases e assim conseguir recorrências maiores na expectativa de melhor qualidade de convergência. Testes feitos com o BiCG(1) =BiCG, QMR(0) =QMR e o QMR(1) com matrizes não normais indicam que não há maiores alterações do QMR(1) sobre o QMR(0). A medida usada para comparação foi a norma-ℓ_2 do resíduo. Os seus resultados teóricos indicam uma boa perspectiva nas concretizações daqueles algoritmos usando recorrências maiores.

O grupo de Van der Vorst [47], [48], [49], [50], [51] tem sistematicamente desenvolvido trabalhos similares, quer baseados na extensão do número de recorrências, quer na mudança dos polinômios de recorrência de um método padrão, normalmente o BiCG ou sua variante o CGS. Deles resultaram o GCGS, o BiCGSTAB, o BiCGSTAB(ℓ), o grupo do GMRES$\star$, entre outros.

4.3.4 O Teorema de Faber e Manteuffel

O Teorema de Faber e Manteuffel tem um papel importante na estratégia sugerida neste trabalho. Para descrição sintética do Teorema considere-se, para $i = 1, 2, \ldots$, uma recorrência do tipo

$$x_i = \mathcal{A}(x_{i-1}) = x_{i-1} - \tau_{i-1}(Ad_{i-1} + \sum_{j=i-s+1}^{i-1} \beta_{j-1} d_j) \ , \tag{4.3}$$

ou seja,

$$x_i = x_{i-1} - \tau_{i-1} d_{i-1} \tag{4.4}$$

$$d_i = Ad_{i-1} + \sum_{j=i-s+1}^{i-1} \beta_{j-1} d_j \tag{4.5}$$

onde τ_{i-1} e β_{j-1} são reais, e s é um inteiro menor do que n. Também, $d_0 = b - Ax_0$ para um dado x_0. Pode-se mostrar que

$$x_i \in x_0 + \mathcal{K}^i(A, f_0) \ , \text{ e}$$

$$\text{span}\{d_o, d_1, \ldots, d_{i-1}\} = \text{span}\{f_o, f_1, \ldots, f_{i-1}\}$$

O CG para matrizes Hermitianas positivas definidas é da forma (4.3) com $s = 2$. Ambos, Orthomin(2) e Orthomin(3), com $s = 3$.

A pergunta que Faber e Manteuffel responderam é: Para quais matrizes A pode-se construir uma recorrência da forma (4.4)-(4.5) com a propriedade de que, para qualquer x_0, a aproximação $x_i \in x_0 + \mathcal{K}^i(A, f_0)$ é ótima, ou seja, a função $\phi_i = \frac{1}{2}\|e_i\|_B^2$, *para algum* B *independente de* x_0, é mínima em $x_0 + \mathcal{K}^i(A, f_0)$? A resposta vem no seguinte Teorema, antecedido por uma definição:

Definição 87 *Um algoritmo da forma (4.4)-(4.5) é um método $CG(s)$ para A se, para todo d_0, a sequência $\{d_i\}_{1\leq i\leq m-1}$ satisfaz $(d_i, d_j)_B = 0$ para todo $j < i$, onde m é o número de iterações necessárias para obter a solução exata $x_* = A^{-1}b$.*

Teorema 88 (Faber e Manteuffel) *Um método $CG(s)$ para a matriz A existe se, e somente se, ou*

1. *o polinômio mínimo de A (o polinômio da forma $g(A) = 0$, não trivial, de grau mínimo) possui grau menor ou igual a s, ou*

2. *$A^\star$ é um polinômio de grau menor ou igual a $s-2$ em A, onde $A^\star$ é a adjunta de A com respeito a algum produto interno $(\cdot,\cdot)_\bullet$, isto é, $(Av, w)_\bullet = (v, A^\star w)_\bullet$ para todos os vetores v e w.*

A condição 1 é clara, já condição 2 merece alguns esclarecimentos. Para qualquer produto interno $(\cdot,\cdot)_\bullet$ existe uma matriz Hermitiana positiva definida B tal que $(v, w)_\bullet = (v, Bw)_2$, para todos os vetores v e w. A matriz B-adjunta de A, denotada por $A^\star$ no Teorema, é a única matriz que satisfaz

$$(Av, Bw)_2 = (v, BA^\star w)_2$$

para todos os vetores v e w. Desta definição segue-se que (ver Capítulo 2)

$$A^\star = B^{-1}A^H B\ .$$

Diz-se que a matriz A é B-normal se, e somente se, $A^\star A = AA^\star$. Chamando $B^{1/2}$ a raiz quadrada da matriz Hermitiana positiva definida B, então, A ser B-normal é equivalente à seguinte condição

$$(B^{-1/2}A^H B^{1/2})(B^{1/2}AB^{-1/2}) = (B^{1/2}AB^{-1/2})(B^{-1/2}A^H B^{1/2})$$

que é justamente a condição de a matriz $\widetilde{A} = B^{1/2}AB^{-1/2}$ ser uma matriz normal, ou seja $\widetilde{A}^H\widetilde{A} = \widetilde{A}\widetilde{A}^H$. Pode-se mostrar que $\widetilde{A}$ é normal (ou seja, A é B-normal) se, e somente se, $\widetilde{A}^H$ pode ser escrita como um polinômio, de qualquer grau, em $\widetilde{A}$. Se $\mu(\widetilde{A})$ é o menor grau para o qual isso é verdade, então μ é denominado o grau B-normal de A. Com esta notação, a condição 2 do Teorema pode ser escrita

2. A é B-normal$(s-2)$.

O caso B-normal(> 1) corresponde a anomalias, no sentido em que estas matrizes de fato pertencem a um subespaço de $L(\mathbb{C}^n)$ com dimensão menor do que n^2. Pode-se mostrar que se são B-normais com grau > 1 porém $< n$ elas também possuem grau $\mu(A) < n$, se enquadrando na condição 1 do Teorema.

O caso de interesse é quando A é B-normal(1), que são matrizes para as quais existe um CG(3), ou seja, é possível uma recorrência curta com $s = 3$ em (4.5). Para elas os métodos CG já são todos conhecidos. São matrizes diagonalizáveis cujo espectro está contido em um seguimento de reta no plano complexo.

A importância do Teorema de Faber e Manteuffel está no fato de que para a maioria dos problemas lineares não Hermitianos não se pode esperar encontrar uma recorrência curta que gere uma aproximação ótima em métodos de Krylov. O termo ótimo é entendido como ortogonalidade em relação a um produto interno que *independa do vetor inicial.* O Teorema torna-se ainda mais importante quando se sabe que a maioria dos métodos correntes minimizam alguma norma.

Porém, normas como as do BiCG e a do QMR geralmente não têm um significado de interesse para o analista, e não podem ser facilmente relacionadas com normas de interesse geral como a norma-ℓ_2 e a norma-ℓ_∞.

Aspectos didáticos e especializados relacionados com o Teorema de Faber e Manteuffel podem ser achados em [30], [52], e referências lá citadas.

Até a presente data, é entendimento comum que (para discussão didática, ver, por exemplo, Anne Greenbaum [52], e trabalhos lá relacionados):

a) A possibilidade de recorrências curtas que gerem uma aproximação ótima em uma norma que independa do vetor inicial, mas que difira marginalmente das normas de interesse, tipicamente a norma-ℓ_2, permanece em aberto. A literatura sugere que este deve ser o caminho para o desenvolvimento do método ideal para problemas não Hermitianos.

b) Permanece um dos grandes problemas em aberto desenvolver um método que gere, provavelmente, uma aproximação quase ótima em alguma norma padrão (tipicamente a norma-ℓ_2), enquanto requer apenas um esforço computacional e uma memória auxiliar, ambos, da ordem de $O(n)$, além do produto matriz-vetor, ou provar que este método não existe.

c) O Teorema de Faber e Manteuffel trata apenas de uma única recorrência. Permanece uma questão aberta: em que condições, por exemplo, um par de recorrências acopladas (tipo Lanczos) podem gerar aproximações ótimas.

Esta parte da tese resulta de um esforço de caminhar nas direções dos ítens *a*),

b) e *c*) acima. São propostas algumas estratégias de solução que de forma alguma representam um estágio final de investigação, mas que o autor vem perseguindo com resultados no mínimo interessantes, e de fato, precisam de mais testes numéricos. O desenvolvimento da primeira parte está estreitamente relacionado com este objetivo. O alvo final, sem dúvida, sempre é a complexidade da ordem de $O(n)$ - (adicional à $O(n^2)$ do produto matriz-vetor).

4.4 Precondicionamento

A solução iterativa de sistemas lineares somente sobrevive graças ao precondicionador. O precondicionador transforma o sistema linear em um sistema equivalente com propriedades mais adequadas para o método de Krylov. Segue-se que precondicionar um sistema implica em conhecimento prévio das propriedades da matriz A e do método de Krylov a que vai ser exposto.

Os métodos de Krylov cujas bases são geradas pelo algoritmo de Arnoldi atacam inicialmente certas regiões da fronteira do espectro no plano complexo. Sobre esta ótica, uma da funções do precondicionador seria levar o interior do espectro para o seu contorno, preferencialmente concentrando grupos de autovalores para serem aniquilados de uma só vez. Porém, para o caso geral, não se sabe como fazer este trabalho com ferramentas de álgebra linear computacional eficientes.

Dado um precondicionador H, pode-se distinguir entre precondicionamento à esquerda e precondicionamento à direita. No primeiro caso o sistema resultante é $HAx = Hb$, e no segundo caso $AHy = b$, com $x = Hy$. É percepção comum que as duas formas são matematicamente equivalentes. Entretanto, pelo menos para sistemas não simétricos, quando submetidos a um certo método de Krylov, a qualidade da convergência das diversas normas podem ser diferentes. Para exemplos vide [42].

Existe um aspecto relevante em precondicionamentos esquecido na literatura. Praticamente qualquer método admite um precondicionador à esquerda. Porém nem todos os métodos o admitem à direita. Não no sentido prático de dificuldade de implementação, mas no sentido analítico. É necessário que o método de Krylov, quando unido ao precondicionador, seja imune ao escalonamento. Isto é, as soluções aproximadas de $\widehat{A}x = \widehat{b}$ sejam as mesmas de $\rho\widehat{A}x = \rho\widehat{b}$, ρ um real não nulo, e $\widehat{A}, \widehat{b}$ os precondicionados. Isto tem causado sérias dificuldades de convergência em diversos esquemas iterativos.

Do ponto de vista computacional, ambas, A e H, não precisam estar definidas explicitamente, senão como operadores sobre os vetores que atuam: $A(v) = Av$ e $H(v) = Hv$. Além disso, a operação $H(v)$ deve ser suficientemente econômica para ser compensada com o número de iterações que irá reduzir do mesmo algoritmo não precondicionado. A memória representada pelo precondicionador deve ser marginal em relação à necessária para o operador A, a não ser em problemas extremamente difíceis nos quais não se conhece, ou não se dispõe de alternativas.

Existem muitas técnicas de precondicionamento. Pode-se distinguir duas atitudes frente ao precondicionamento: mantê-lo constante $H_i = H_0$ para todo $i \geq 0$, ou permiti-lo variável a cada iteração. Neste último caso, diz-se que o precondicionamento é variável ou não linear. Em ambos os casos, eles podem ser implícitos ou explícitos. No precondicionamento implícito é conhecido o operador $A_i = H_i^{-1}$, e sua aplicação resulta na solução de um novo sistema linear (complementar ao original) em cada iteração, ou grupo delas. Claramente, tais sistemas devem ter complexidade $O(n)$ em contraposição a $O(n^2)$ exigida pelo sistema principal.

O presente trabalho enquadra-se, parcialmente, como precondicionador variável. Parcialmente, porque a idéia é dar ao precondicionador mais responsabilidade do que a normalmente a ele atribuida, ou seja, ele deve estar perfeitamente inserido na estratégia de convergência local e global, e estreitamente relacionado com a formação da base e da função ϕ. A estratégia a ser seguida faz cuidadosas restrições a toda a sequência $\{H_i\}_{i\geq 0}$ e inclui o precondicionador inicial H_0, de tal forma que pertence simultaneamente às duas classes citadas.

Dentre as técnicas de precondicionamento não linear existem algumas mais distintas. Não existe uma técnica ideal e, naturalmente, cada uma delas está estreitamente vinculada com a forma da equação diferencial ou integral original, com as características topológicas das matrizes resultantes (estrutura), e com o sistema computacional no qual vai ser realizada.

A técnica dos Métodos Iterativos Truncados, que trabalha com o precondicionador implícito P, consiste em resolver o sistema complementar com outro método iterativo dentro de uma precisão pré-especificada. Algumas vezes é possível resolver o sistema complementar com o mesmo método original. Esses esquemas são conhecidos como *inner-outer loops*. Eles são matematicamente equivalentes ao método Inexato de Newton aplicado a uma função linear.

Outras técnicas consistem em fazer a Decomposição Parcial de $A_0 = A$, dando origem a inúmeras variantes, a depender do tipo de decomposição utilizada e de

detalhes da decomposição parcial. Estas técnicas são muito utilizadas em sistemas esparsos. Note que, invariavelmente, procura-se manter a esparsidade da matriz original A, apesar de A^{-1} ser densa. Esta é uma da suas principais deficiências, principalmente para matrizes esparsas, justamente onde podem ser aplicados.

A técnica de Precondicionadores Inversos Aproximados procura resolver este último problema indo em busca de uma boa aproximação para H. A dificuldade está agora em determinar qual a estrutura esparsa para H que melhor aproxima A^{-1}, densa. Alega-se que a escolha entre inversa aproximada e decomposição parcial é uma questão filosófica [53]. A verdade é que cada uma possui um desempenho melhor ou pior do que a outra a depender do problema. Assim a escolha deve ser feita por análise do problema original seguida de experimentação numérica.

A estratégia que se propõe está mais perto da técnica de aproximação da inversa. Mas, dentre os precondicionadores iniciais H_0 admissíveis, inclui-se as decomposições parciais (talvez escaladas), e eventualmente métodos iterativos truncados.

Recentemente, tomou-se conhecimento de alguns trabalhos mais relacionados com a técnica proposta, para problemas não lineares, que possuem um ingrediente básico similar: o uso de métodos secantes para precondicionar. O trabalho de Martínez [57] que trata de problemas não lineares resolvidos com o método Inexato de Newton, no qual usa precondicionadores simétricos, e o trabalho de Morales e Nocedal [58], para o método de Newton, também com precondicionadores simétricos. Embora as ferramentas teóricas sejam completamente distintas, e enquanto mais amplas dentro dos métodos secantes, sejam mais restritas no contexto de algoritmos iterativos, aqueles trabalhos confirmam a expectativa do autor de que o tratamento de problemas lineares e não lineares (iterativamente) deve ser unificado. Notar-se-á também que o uso de métodos secantes no presente trabalho é uma das formas de realização dentre as possíveis. Métodos secantes são a escolha natural pelo seu sucesso e pelo enorme acúmulo de conhecimento teórico e experimental existente em diversas disciplinas, notadamente em análise não linear e otimização.

Sob o título de Decomposição de Domínio situa-se uma ampla classe de precondicionadores estreitamente vinculados às computações paralelas [54]. Refere-se ao processo de subdividir a solução de um grande sistema linear em problemas menores, cujas soluções possam ser usadas para gerar um precondicionador para o sistema de equações original. Nesta classe inclui-se os métodos *multigrid*, e alguns métodos multinível. Os métodos *multigrids* podem ser formulados algebricamente facilitando uma melhor identificação, e possuem uma característica importante.

Enquanto os métodos de Krylov que minimizam resíduos ou norma de energia captam inicialmente as componentes da resposta de maior conteúdo nessas normas, outros métodos, notadamente os métodos básicos, resolvem as componentes do outro lado do contorno do espectro, justamente por minimizarem outras normas. Os métodos *multigrid* procuram tomar partido dessas características de ambos os métodos, unindo-os em ciclos estrategicamente alternados. Neste sentido, equivalem a usar um método de Krylov do tipo Galerkin convenientemente alternado com métodos básicos. Os métodos básicos funcionam como precondicionadores para o método de Krylov. Isto é perfeitamente possível dentro da estratégia proposta, trabalhando diretamente sobre a formação do precondicionador. No próximo capítulo esta possibilidade será exemplificada.

4.5 Motivação Para Uma Estratégia

Em um nível mais alto, não há motivos para distinguir entre a solução iterativa de um problema linear e a de um não linear, em particular quando se recorre a precondicionadores não lineares. Existem mesmo muitos problemas não lineares que são bem resolvidos com algoritmos para problemas lineares. O trabalho de Axelsson [55] sugere que é possível resolver sistemas não lineares sem a sua decomposição em sucessivos sistemas lineares com o método de Gauss-Newton. Van der Vorst [56] formula uma estratégia para problemas não lineares que inclui a solução eficiente de sistemas lineares e problemas de autovalores e autovetores. A primeira parte desta tese unifica a caracterização da qualidade de convergência e suas estimativas de erros para algoritmos não lineares e lineares, de forma que, mesmo no nível mais baixo de implementação, não há qualquer distinção entre eles.

Uma interpretação conveniente do Teorema de Faber e Manteuffel sugere que se deve procurar medidas de projeção que mudem consistentemente em cada iteração. Ocorre que, a correta insersão de um precondicionador em um esquema iterativo torna-o função tanto do algoritmo de geração das bases quanto da própria função distância. Isso será mostrado posteriormente. A estratégia corrente de isolar os três ingredientes carece de consistência. Na opinião do autor, a insistência nesta independência que caracteriza os precondicionadores heurísticos, focados na estrutura topológica da matriz dos coeficientes, é causa de inúmeros insucessos.

O precondicionador ideal deve retirar as informações do operador (linear ou não linear) de uma forma matematicamente consistente, à medida em que as iterações

evoluam. Existem algumas possibilidades, não se argumenta que exista uma melhor. Simultaneamente, o precondicionador deve conseguir reter todas as informações relevantes oriundas de uma análise antecipada pelo analista. É essencial determinar que informações relevantes seriam essas.

Aceita a impossibilidade algorítmica de minimizar a norma de conveniência do analista em qualquer situação, é imediato que o algoritmo deve ter a função de estimá-la com confiança. Em muitos problemas, o analista tem interesse em que mais de uma medida esteja abaixo de uma precisão pré-especificada (a flecha central de uma enorme viga, o momento máximo de um pequeno balanço). Normalmente esta precisão necessária não é tão grande, e pode ser satisfeita sem um esforço computacional excessivo. Permite-se, particularmente em problemas estruturais, um certo nível de inconsistência do algoritmo (não satisfação das equações de equilíbrio e compatibilidade). Deve-se portanto distinguir claramente entre inconsistência do algoritmo e imprecisão das medidas de interesse. O algoritmo deve estar apto a verificar ambos, independentemente.

Se o interesse maior são as medidas do analista, deve-se também estar apto a construir iterações que objetivem estes interesses. Em análise estrutural por exemplo, as técnicas adaptativas nos diversos métodos de discretização (elementos finitos, elementos de contorno, diferenças finitas, métodos espectrais, ...) normalmente concentram-se em minimizar a norma da energia. Porém, nenhum engenheiro estrutural sabe relacionar esta norma com as de seu interesse, nem projetar com elas. As medidas de projeto são medidas locais: momentos fletores, esforços cortantes, flechas. Não existe qualquer experiência acumulada em projetos e experimentos com normas de energia, ou mesmo a norma-ℓ_2. Este, sem dúvida, é um profundo abismo entre as necessidades dos algoritmos e as necessidades dos engenheiros estruturais. Acredita-se que isto ocorra em outras áreas da engenharia.

Todos esses argumentos juntos, propõe-se uma estratégia global de compatibilidade dos diversos interesses, reunindo diversas ferramentas matemáticas existentes. Essencialmente, a estratégia constitui-se em polialgoritmos.

Coloca-se sobre o precondicionador toda (ou quase) a responsabilidade do sucesso das iterações: aceitar as informações sobre o sistema linear *a priori* estabelecidas por análise dos operadores contínuos originais ou por experimentação numérica, ou por análise do mesmo ou de outros problemas similares, e aprender mais sobre o sistema ao longo das iterações, de uma forma matematicamente consistente. A escolha do precondicionador é flexível, de tal forma que se pode adequar melhor o

precondicionamento à escolha das medidas a serem ortogonalizadas.

Na impossibilidade de ortogonalizar a norma de interesse, são usadas medidas generalizadas e projeções generalizadas, de fato projeções genuinamente não ortogonais. Mas que podem ser feitas assintoticamente ortogonais. Minimizam a separação com as normas de interesse, e garantem a convergência do quociente da norma minimizada e da norma de interesse . Essas medidas podem ser feitas dependentes do vetor inicial, numa tentativa de contornar o Teorema de Faber e Manteuffel. São obtidos legítimos métodos de métrica variável, porém com a flexibilidade de impor a convergência do quociente, e com a correta interpretação das métricas, e de suas relações com a norma-ℓ_2.

A construção das bases é efetuada com algoritmos tradicionais em sistemas lineares, ou mesmo com algoritmos não convencionais, oriundos de algoritmos para problemas não lineares. A resposta é descrita em subespaços de Krylov, com possibilidades de inclusão de bases fora destes espaços.

No próximo capítulo descreve-se o conteúdo de uma estratégia para algoritmos iterativos que essencialmente procura tomar proveito sistemático de cada ingrediente com o objetivo de conseguir as melhores velocidades de convergência e compatibilizar os diferentes interesses.

Capítulo 5

Estratégia Para a Convergência do Quociente

5.1 Introdução

Considera-se o problema (2.3) linear com solução $x_\star$ única, ou seja,

$$f(x) = Ax - b\ , \qquad f(x_\star) = 0\ , \tag{5.1}$$

onde $A \in L(\mathbb{R}^n)$ é regular e $x, b \in \mathbb{R}^n$. Com execessão (talvez) dos conceitos de distâncias generalizadas, todo o presente capítulo extende-se sem modificações a $\mathbb{C}^n$.

A estratégia descrita pode ser vista como um polialgoritmo. A idéia de polialgoritmos não é nova em sistemas lineares, porém a sua objetividade, a sua flexibilidade, suas possibilidades adaptativas e alguns dos seus ingredientes são bastante genuinos.

Um exemplo de polialgoritmo bem sucedido é descrito por Schonauer [59]. Porém, ele limita-se a tomar decisões entre um pequeno número de algoritmos, sequencialmente, após constatar, pela dificuldade de convergência, que um algoritmo não é o adequado. Neste trabalho tem-se objetivos bem distintos.

O elemento básico é um algoritmo padrão, suficientemente geral e flexível para comportar virtualmente qualquer algoritmo, para análise ou implementação. Sobre este algoritmo permite-se recursividade e e simultaneidade. A recursividade é usada para gerar bases para distintos subespaços, com diferentes algoritmos, incluindo subespaços de Krylov. A simultaneidade visa compatibilizar normas de interesse e potencializar paralelizações. Permite-se a ortogonalização de normas convencionalmente usadas, ou medidas generalizadas (a serem definidas adequadamente).

A motivação principal para a flexibilidade do esquema é definir uma plataforma de estudos analíticos, muito mais do que uma para estudos de desempenho com-

putacional. Entretanto, estas questões estão bastante vinculadas que não é possível trabalhar em uma sem ter a outra sempre presente.

Inicia-se definindo um algoritmo padrão conveniente, em seguida os seus elementos hierarquicamente inferiores, para no fim discutir os níveis superiores de recursividade e simultaneidade.

5.2 O Esquema Algorítmico

Com o propósito de análise e comparação, considera-se o esquema algorítmico da forma $x_{i+1} = \mathcal{A}(x_i)$:

$$x_{i+1} = x_i - \tau_i H_i f_i \ , \tag{5.2}$$

onde $H_i \in L(\mathbb{R}^n)$, $f_i = f(x_i) \in \mathbb{R}^n$, e $\tau_i \in \mathbb{R}$.

A idéia é deixar o esquema bastante flexível para comportar uma classe bastante geral de algoritmos em $\mathbb{R}^n$, enquanto se divide o algoritmo em etapas que são consideradas primordiais para a análise a ser desenvolvida.

Algoritmo Padrão: Dados $x_0, f_0 \in \mathbb{R}^n$, e $H_0 \in L(\mathbb{R}^n)$, $d_0 = -H_0 f_0$, compute as iterações $i = 0, 1, \ldots$, na seguinte sequência:

1. Parâmetro de busca:
   1.1 determine $\tau_i = \tau_i(\phi_i)$
   1.2 verifique a admissibilidade: $0 < \underline{\epsilon} \leq \tau_i \leq \overline{\epsilon}$
2. Incremento da solução:
   2.1 calcule: $s_i = \tau_i d_i$
   2.2 estime $d(x_{i+1}, x_\star)$ e melhore a estimativa de $d(x_i, x_\star)$
3. Atualize a solução corrente: $x_{i+1} = x_i + s_i$
4. Determine o incremento do resíduo: $y_i = A s_i$
5. Resíduo corrente:
   5.1 atualize o resíduo: $f_{i+1} = f_i + y_i$
   5.2 verifique consistência
6. Atualize o precondicionador: $H_{i+1} = \mathcal{H}(H_i)$
7. Direção de busca:
   7.1 calcule a nova direção: $d_{i+1} = -H_{i+1} f_{i+1}$

7.2 verifique a descendência

O esquema acima comporta uma grande variedade de algoritmos. De fato, ele também admite uma grande classe de algoritmos para problemas não lineares e otimização. O esquema é apenas formal, não significando que as computações vão ser feitas exatamente da forma indicada, ou mesmo que devam ser efetuadas. Sempre existe uma grande distância entre um esquema formal e sua implementação.

Os algoritmos que fazem busca em subespaços m-dimensionais pertencem ao algoritmo padrão. Existem várias possibilidades. A imediata é: $\tau_i = 1$, e atualizar o precondicionador com as m bases. Em qualquer caso d_i será chamada de direção de busca. Outras opções específicas serão discutidas posteriormente. O precondicionador é atualizado com a condição de minimizar ϕ ou, equivalentemente, ϕ-ortogonalizar o erro se ϕ é definida por um produto interno. A medida ϕ assumirá um significado mais amplo no resto do texto.

Para algoritmos que não usam f_i na construção da direção de busca (se existe algum de interesse), considera-se que cada H_i contenha a necessária transformação linear. De fato, o autor argumenta que todos os algoritmos algébricos bem conhecidos podem ser enquadrados no algoritmo padrão. A notação $H_{i+1} = \mathcal{H}(H_i)$ é usada para indicar que o precondicionador H_{i+1} é atualizado a partir do precondicionador corrente H_i.

As duas etapas decisivas para o sucesso de um algoritmo são, pela ordem, a formação da direção de busca definida pelo precondicionador H_{i+1} e a determinação do parâmetro de busca τ_i. Nas seções seguintes são discutidas essas duas etapas e outros aspectos do algoritmo padrão relevantes para esta tese.

O algoritmo padrão é um esquema tipo-Newton para problemas não lineares ou otimização. Pode-se preferir também considerá-lo um esquema tipo-CG. Para os propósitos desta tese, o autor prefere vê-lo como um esquema de Richardson com precondicionador variável. As razões para isto, e a posterior amarração do parâmetro de busca ao precondicionador, se tornarão claras no decorrer deste capítulo. Porém, a visão do esquema tipo-Newton permite usar profundos resultados analíticos desta área. A visão do esquema tipo-CG fornece as perspectivas de implementações ótimas. Estas visões complementam-se umas às outras e serão usadas para descrever os elementos da estratégia.

É de se notar a distinção feita entre inconsistência do algoritmo e convergência (na medida ϕ). Também a inclusão de verificações de descendência e admissibilidade. Estas duas verificações visam garantir a convergência do quociente.

5.3 Impondo Consistência

Durante toda a análise efetuada nos capítulos precedentes a hipótese básica foi de consistência do algoritmo. É de vital importância que o algoritmo seja consistente. A consistência garante que, se o algoritmo possui um ponto fixo, ele é único, e também a solução do problema linear (5.1). Isto justifica o aparecimento do resíduo f_i sob a forma multiplicativa em (5.2). Expõe-se agora a forma mais simples, e suficientemente geral, de garantir consistência.

A escolha de H_i não pode ser arbitrária e deve obedecer a uma das duas condições mínimas seguintes, para todo $i \geq 0$:

1a. H_i é regular, ou

1b. $f_i \notin \mathcal{N}(H_i)$.

Onde $\mathcal{N}(H_i)$ é o núcleo de H_i. Evidentemente $1a$ implica em $1b$. Ocorre que, exigir regularidade de H_i é excessivamente restritivo e desnecessário, além de eliminar, entre outros algoritmos, o método dos gradientes conjugados padrão. Mesmo no caso de o precondicionador manter-se constante ao longo de toda as iterações, não necessariamente $1b$ implica em $1a$, uma vez que o algoritmo deve ser idealmente interrompido em um número $m << n$ iterações, e assim a dimensão de $\mathcal{N}(H_0)$ precisa apenas ser igual a m. Sobre este aspecto, deve ser notado que a possibilidade $1b$ estaria eliminada desnecessariamente se o precondicionador fosse definido como $A_{i+1} = H_{i+1}^{-1}$.

Às condições $1a$ ou $1b$ acima deve ser adicionada outra. Para todo $i \geq 0$:

2. existe um real $\underline{\epsilon} > 0$ tal que $\tau_i \geq \underline{\epsilon}$ para todo $i \geq 0$.

As condições 1 e 2 são suficientes para garantir a consistência do algoritmo. Ou seja,

$$1 \text{ e } 2 \Rightarrow x_i = \mathcal{A}(x_i) \quad \text{se, e somente se,} \quad f(x_i) = 0 \Rightarrow x_i = x_\star. \tag{5.3}$$

Descreve-se brevemente as noções intuitivas que geram as condições acima, como evolução das iterações de Richardson. A garantia de que o ponto fixo do algoritmo é a solução do problema linear é conseguida se o algoritmo for definido como $x_{i+1} = x_i - f_i$. Porém, as diferentes dimensões físicas entre x e $f(x)$ com certeza produzirão sérios problemas numéricos. Assim, escala-se f_i produzindo: $x_{i+1} = x_i - \tau_i f_i$. Uma escala uniforme. A próxima etapa óbvia é rotacionar e escalar as componentes

individualmente produzindo: $x_{i+1} = x_i - \tau_i H f_i$. É muito pouco provável que o analista possua conhecimento prévio suficiente para gerar a melhor matriz H. Admitindo isto, porém não descartando o conhecimento a priori de uma certa H_0 conveniente, a melhor das quais seria, teoricamente (mas não para todos os algoritmos), $H_0 = A^{-1}$, deixa-se para o algoritmo de solução a tarefa de gerar melhores H_i, à medida em que, supõe-se, compreenda melhor o problema enquanto se relaciona com ele. Não se espera que $H_i \to A^{-1}$ quando i cresce, mesmo porque A^{-1} é provavelmente densa. Mas apenas que H_i colabore para diminuir o número de iterações sem acrescentar ao algoritmo, no ato de atualizar H_i e gerar a direção de busca, um esforço computacional não compensável. Para precondicionadores perfeitamente consistentes espera-se que $\tau_i = 1$ para qualquer $i > 0$.

Garantidas as condições de consistência, o algoritmo estagnará se, e somente se, o resíduo for nulo, e a solução terá sido encontrada como o ponto fixo do algoritmo. Desta forma, a teoria dos pontos fixos pode ser aplicada, pelo menos assintoticamente. Para a garantia da descendência do algoritmo em uma norma específica basta impor o Lema 56, garantindo, por exemplo, a condição (3.49) para algoritmos gerais em espaços de Hilbert. Para projeções ortogonais a Proposição 59 deve ser usada para a garantia da convergência do quociente.

Será explorada uma certa classe de projeções não ortogonais, naturalmente decorrente da teoria da convergência do quociente desenvolvida nos capítulos precedentes, e que será de grande valia para resolver alguns problemas já levantados. Em particular, o problema de separação e o problema da computabilidade na ortogonalização da norma-ℓ_2 do erro.

5.4 Distâncias e Projeções Generalizadas

Existem diversas interpretações possíveis para a medida $d(x_{i+1}, x_i)$. Em todo o desenvolvimento da teoria de convergência do quociente exposta nos capítulos anteriores, ela sempre está presente, em particular nas estimativas de erros. Dá-se uma interpretação bastante construtiva de medida generalizada. Esta interpretação é significativa sobre vários aspectos. Aspectos esses que vão desde a sua aceitação como medida de convergência até o nível de eficiência de implementações em ambientes de computação paralela.

No que se segue, a discussão ficará restrita a espaços reais de dimensões finitas, onde a interpretação geométrica torna-se mais nítida. Deve-se observar porém,

que medidas epigráficas são úteis na caracterização de certo tipo de convergência uniforme de funções convexas em espaços de Banach [60].

5.4.1 Funções de Bregman

Seja ϕ uma função de $\mathbb{R}^n$ em $\mathbb{R}$, com domínio D. O epígrafo de ϕ, denominado epi(ϕ) é o conjunto de pontos

$$\text{epi}(\phi) = \{(x, r) \mid x \in \mathbb{R}^n, r \in \mathbb{R}, \phi(x) \leq r\} \ , \tag{5.4}$$

pontos esses localizados acima e sobre o gráfico de ϕ em $\mathbb{R}^{n+1}$. Quando ϕ é uma função convexa o seu epígrafe é um conjunto convexo em $\mathbb{R}^{n+1}$.

Dada uma função $\phi : D \subseteq \mathbb{R}^n \to \mathbb{R}$, estritamente convexa, finita em $x, y \in D$ e diferenciável em y, pode-se medir uma espécie de distância entre x e y definindo uma função $D_\phi(x, y)$, denominada de distância generalizada ou função-D

$$D_\phi(x, y) = \phi(x) - \phi(y) - (\nabla\phi(y), x - y) \ , \tag{5.5}$$

onde $\nabla\phi(y)$ é o gradiente de ϕ no ponto y, e $(\cdot\cdot)$ o produto interno. Da convexidade estrita de ϕ, pode-se provar que $D_\phi(x, y) \geq 0$, e $D_\phi(x, y) = 0$ se, e somente se, $x = y$. A função $D_\phi(x, y)$ mede o quanto uma aproximação linear para ϕ em torno do ponto y subestima os valores de ϕ em x. Em outras palavras, $D_\phi(x, y)$ mede a diferença $\phi(x) - h(x)$, onde $h(z)$ representa o hiperplano hP, o qual é tangente ao epígrafe de ϕ no ponto $(y, \phi(y))$ em $\mathbb{R}^{n+1}$.

O vetor normal ao conjunto epi(ϕ) no ponto $(y, \phi(y))$ no gráfico de ϕ é dado por $N = (\phi(y), -1)$. O hiperplano não vertical hP, suporte do conjunto epi(ϕ) no ponto $(y, \phi(y))$, é dado por

$$hP = \{(z, h(z) \mid h(z) = \phi(y) + (\nabla\phi(y), x - y)\} \, , \tag{5.6}$$

de onde vê-se que $D_\phi(x, y) = \phi(x) - h(x)$.

As funções-D em geral não se comportam como distâncias Euclidianas. Geralmente, elas não são simétricas e não obedecem à desigualdade triangular. Porém, comportam-se de uma forma crucialmente parecida com os quadrados das distâncias Euclidianas, de acordo com o seguinte lema dos três pontos [62, Lema 1]:

Lema 89 *Dados $\phi : \mathbb{R}^n \to (-\infty, +\infty)$, D_ϕ definida por (5.5), e $u, v, w \in \mathbb{R}^n$ tal que $\phi(u), \phi(v)$ e $\phi(w)$ são finitos e ϕ é diferenciável em u e v,*

$$D_\phi(w, u) = D_\phi(v, u) + D_\phi(w, v) + (\nabla\phi(v) - \nabla\phi(u), w - v) \ . \tag{5.7}$$

Este resultado pode ser confirmado por simples substituição de (5.5) em (5.6).

Quando $\phi(x) = \frac{1}{2}\|x\|^2$, então (5.7) reduz-se à seguinte identidade, também proveniente da caracterização de um espaço Euclidiano

$$\frac{1}{2}\|w-u\|^2 = \frac{1}{2}\|v-u\|^2 + (v-u, w-v) + \frac{1}{2}\|w-v\|^2. \tag{5.8}$$

Definição 90 *Diz-se que ϕ é uma função de Bregman com zona S se as seguintes condições são satisfeitas:*

B1. Abertura da Zona: $S \subseteq \mathbb{R}^n$ é um conjunto aberto.

B2. Finitude/Continuidade: ϕ é finita e contínua em $\overline{S}$, o fecho de S.

B3. Convexidade estrita: ϕ é estritamente convexa em $\overline{S}$.

B4. Diferenciabilidade: ϕ é continuamente diferenciável em S.

B5. Conjuntos de níveis limitados: Dado qualquer $x \in \overline{S}$ e um escalar ε, o conjunto de nível parcial à direita $L(x,\varepsilon) = \{y \mid D_\phi(x,y) \leq \varepsilon\}$ e à esquerda $L(\varepsilon,x) = \{y \mid D_\phi(x,y) \leq \varepsilon\}$ são limitados.

B6. Sentido de convergência: Se $\{y_i\} \subset S$ é uma sequência convergente com limite $y_\star$, então $D_\phi(y_\star, y_i) \to 0$.

B7. Consistência da convergência: Se $\{x_i\} \subset \overline{S}$ e $\{y_i\} \subset S$ são sequências tais que $y_i \to y_\star$ e $\{x_i\}$ é limitada, e além disso $D_\phi(x_i, y_i) \to 0$, então tem-se $x_i \to y_i$.

A definição de funções de Bregman não é universal. Variantes existem que introduzem modificações em função dos objetivos e das necessidades da análise. Para os propósitos deste texto, as condições acima são suficientes.

Exemplos de funções de Bregman são: a função negativa da entropia definida por $\phi(x) = \sum_{j=1}^{n} x_j \log x_j$ e a função $\phi(x) = -\sum_{j=1}^{n} \sqrt{1-x_j^2}$. Existem outras funções generalizadas que não as funções de Bregman, como por exemplo a classe de funções de Bregman-Legendre e as φ-divergências de Csiszár. É de interesse presente a função $\phi(x) = \frac{1}{2}\|x\|^2$.

De fato, apenas se usará a mais bem comportada das funções de Bregman, a função quadrática $\phi = \frac{1}{2}\|x\|^2$, que possui zona $S = \overline{S} = \mathbb{R}^n$. Pode ser provado que uma função de Bregman é simétrica se, e somente se, é quadrática ([61]). Experimentos com outras funções de Bregman serão relatados em outro trabalho.

A família de funções de Bregman com zona S é denotada por $\mathcal{B}(S)$. Diz-se que $\phi \in \mathcal{B}(S)$ se ϕ é uma função de Bregman com zona S. O seguinte lema está provado em [28, cap. 2].

Lema 91 *Para toda $\phi \in \mathcal{B}(S)$ tem-se que $D_\phi(x,y) \geq 0$, para todos $x \in \overline{S}$, e todo $y \in S$, e $D_\phi(x,y) = 0$ se, e somente se, $x = y$.*

Pode-se notar uma quase identificação entre as funções de Bregman e as seminormas. Com a diferença fundamental dada pelo resultado acima, pelo qual $D_\phi(x,y) = 0$ se, e somente se, $x = y$.

5.4.2 Projeções Generalizadas

A noção de projeção generalizada associada a uma função de Bregman é fundamental.

Definição 92 *Dado $C \subseteq \mathbb{R}^n$, $\phi \in \mathcal{B}(S)$, e $y \in S$, o ponto $x_\perp \in C \cap \overline{S}$ para o qual*

$$\min_{z \in C \cap \overline{S}} D_\phi(z,y) = D_\phi(x_\perp, y) \ , \tag{5.9}$$

(isto é, o mínimo existe e a igualdade é atingida) é denotado por $P_C(y)$ e é chamada de projeção generalizada, ou simplesmente projeção, do ponto y sobre o conjunto C.

O resultado que se segue garante a existência e a unicidade da projeção generalizada [28].

Lema 93 *Se $\phi \in \mathcal{B}(S)$, então para qualquer conjunto convexo e fechado $C \subseteq \mathbb{R}^n$, tal que $C \cap \overline{S}$ é não vazio, e para qualquer $y \in S$, existe uma única projeção generalizada $x_\perp = P_C(y)$.*

O próximo lema, que pode ser verificado por simples substituição, define uma classe de funções de Bregman estreitamente relacionadas com a energia. De fato, a norma da energia e a energia possuem a mesma função-D.

Lema 94 *Se ϕ é uma função de Bregman, então $\phi(x) + (z,x) + \beta$, para qualquer $z \in \mathbb{R}^n$ e $\beta \in \mathbb{R}$, também satisfaz a Definição 90 e produz exatamente a mesma função-D_ϕ, e, portanto, é também uma função de Bregman.*

As aplicações clássicas de projeções generalizadas na construção de algoritmos para problemas lineares consistem em efetuar projeções generalizadas sobre os hiperplanos definidos pelas linhas (ou grupos delas) do sistema de equações. Dá origem a, por exemplo, o clássico método de Cimino, que remonta ao século passado. Estas técnicas têm ressurgido dado às suas potencialidades de paralelização em computadores com memória distribuida.

Existe uma outra possibilidade que oferece muitas perspectivas. Ela segue em uma direção radicalmente diferente, e consiste em dar uma interpretação diferente à formação de polialgoritmos. Cada algoritmo padrão projeta a solução corrente em um convexo, determinado pelos conjuntos de níveis da função que ortogonaliza. Sendo cada algoritmo q-convergente, estes conjuntos de níveis são retráteis. A intersecção deles não é vazia, conforme estabelecido no Capítulo 3, em especial se cada algoritmo for pelo menos q-linearmente convergente. A intersecção também é um conjunto convexo. Isso caracteriza um problema clássico de viabilidade convexa. Ou seja, achar a intersecção, não vazia, de uma coleção finita de convexos. Deseja-se apenas cncontrar a intersecção que satisfaz uma certa precisão especificada antecipadamente para as funções que cada algoritmo participante ortogonaliza. Esta estratégia envolve a execução de vários algoritmos simultaneamente, potencializando paralelizações. No caso de dois algoritmos, há à clássica solução de Von Newman, que consiste em alternar as projeções, ciclicamente, com prova de convergência pelo menos r-linear no ciclo externo. Enquanto as projeções generalizadas em hiperplanos são explicitamente determinadas mesmo em $\mathbb{R}^n$, as projeções em conjuntos convexos quaisquer são determinadas apenas iterativamente (uma inclusão não linear), mesmo em duas dimensões. Este problema é justamente o resolvido por dois algoritmos q-lineares que minimizam funções distintas em um processo de hibridização proposto mais adiante.

O seguinte resultado [28] é auxiliar na caracterização das projeções generalizadas. Ele mostra que certa semelhança com a desigualdade (métrica) triangular existe. Mas, que de fato, o termo "distância " é incoveniente.

Teorema 95 *Seja $\phi \in \mathcal{B}(S)$ e seja $D \subseteq \mathbb{R}^n$ um conjunto convexo fechado tal que $D \cap \overline{S}$ não seja vazio. Assuma que $y \in S$ implique em $P_C(y) \in S$. Seja $z \in D \cap \overline{S}$, então, a seguinte desigualdade é verdadeira para qualquer $y \in S$*

$$D_\phi(P_C(y), y) \leq D_\phi(z, y) - D_\phi(z, P_C(y)) \ . \tag{5.10}$$

A caracterização das projeções generalizadas é dada pelo seguinte teorema [28].

Teorema 96 *Sob as hipóteses do Teorema 95, para qualquer $x \in S$, $x_\perp \in D \cap \overline{S}$ é a projeção generalizada $P_C(x)$ se, e somente se, para todo $u \in D \cap \overline{S}$ tem-se*

$$(u - x_\perp, \nabla\phi(x) - \nabla\phi(x_\perp)) \leq 0 \tag{5.11}$$

No caso especial de $\phi = \frac{1}{2}\|x\|^2$, com $S = \mathbb{R}^n$, o Teorema 96 reduz-se ao de projeções ortogonais, conhecido como o critério de Komolgorov:

$$(u - P_C(x), x - P_C(x)) \leq 0\ , \tag{5.12}$$

para todo $u \in D$, siginificando que o ângulo entre $(u - P_C(x))$ e $(u - P_C(x))$ é obtuso, para qualquer $u \in D$.

5.5 Enriquecimento como Distância Generalizada

Ao longo desta tese, foi constantemente ressaltada a dificuldade de ortogonalização do erro (na norma-ℓ_2). Uma consequência, em certos problemas, é uma grande separação entre diversas normas de interesse e a norma usada para ortogonalização. A identificação da medida do enriquecimento como uma medida generalizada permite efetuar uma certa projeção não ortogonal, de fato uma quase ortogonalização, que, em certos casos pode ser feita assintoticamente ortogonal em qualquer norma de interesse.

Sendo $\|\cdot\|_\bullet$ uma norma específica definida por um produto interno $(\cdot,\cdot)_\bullet$, o interesse inicial é construir um algoritmo no qual $(e_{i+1}, s_i)_\bullet = 0$, pelo menos de forma aproximada. Escolhendo a função de Bregman $\phi = \frac{1}{2}\|x\|_\bullet^2$, onde $x \in \mathbb{R}^n$, pode-se verificar que

$$\phi = \frac{1}{2}\|x\|_\bullet^2 \Rightarrow D_\phi(x,y) = \frac{1}{2}\|x - y\|_\bullet^2 \tag{5.13}$$

Usando a notação algorítmica, e com os índices do algoritmo padrão, na i-ésima iteração, após atualizados a solução e o resíduo, deseja-se determinar d_{i+1}, e posteriormente s_{i+1}, tal que $\phi_{i+2} = \phi(e_{i+2})$ seja mínimo ao longo de d_{i+1}. Primeiro note que

$$\phi_{i+2} = \frac{1}{2}\|e_{i+2}\|_\bullet^2 \Rightarrow D_\phi(e_{i+2}, e_{i+1}) = \frac{1}{2}\|e_{i+2} - e_{i+1}\|_\bullet^2 = \frac{1}{2}\|s_{i+1}\|_\bullet^2\ . \tag{5.14}$$

Vê-se portanto que o quadrado da norma do incremento é uma função-D, especificamente, a distância epigráfica da função erro. O Lema 91 garante que

$$D_\phi(e_{i+2}, e_{i+1}) = 0 \qquad \text{se, e somente se,} \qquad e_{i+2} = e_{i+1}\ .$$

Esta última afirmação está em perfeito acordo com a geometria dos processos q-lineares discutidos na primeira parte desta tese. Estando-se apto a produzir direções de busca suficientemente descendentes, minimizar a função de Bregman $\phi = \frac{1}{2}\|e\|_\bullet^2$ é assintoticamente equivalente a minimizar a sua função-D, porque então, como o ponto fixo do algoritmo é único, a função-D será nula apenas quando a função de Bregman $\phi = \frac{1}{2}\|e\|_\bullet^2$ for nula. A função-D a ser minimizada será:

$$D_\phi(e_{i+2}, e_{i+1}) = \frac{1}{2}\|s_{i+1}\|_\bullet^2 = \frac{1}{2}\|\tau_{i+1}d_{i+1}\|_\bullet^2 = \frac{\tau_{i+1}^2}{2}\|H_{i+1}f_{i+1}\|_\bullet^2. \tag{5.15}$$

O valor do parâmetro de busca τ_{i+1} no final da iteração, será o τ_i do início da próxima iteração. A minimização pode ser feita usando a projeção generalizada.

É importante notar que o processo de minimização inclui o precondicionador H_{i+1}. Ou seja, ambos, τ_i da próxima iteração e H_{i+1} na presente iteração devem ser determinados com a condição adicional de minimizar $D_\phi(e_{i+2}, e_{i+1})$. Não há nada a fazer com f_{i+1} já fixado na determinação de x_{i+1} no início da iteração corrente.

Pode-se, evidentemente, efetuar uma projeção aproximada ao longo de d_i determinado na iteração anterior sem que a formação de H_{i+1} tenha participado do processo de ortogonalização. Quando isto ocorrer, será chamado de *regularização*. A idéia desta projeção aproximada é atuar sobre algoritmos existentes que minimizam uma determinada norma, porém provocam oscilações na norma de interesse. Tipicamente a norma-ℓ_2, ou mesmo a norma da energia.

O leitor deve ter observado que criou-se um problema delicado: a escolha adequada de $H_{i+1} = \mathcal{H}(H_i)$ que seja capaz de simultaneamente: gerar bases d_i suficientemente linearmente independentes, suficientemente descendentes para a função-D, e ao mesmo tempo seja capaz de extrair e transmitir informações úteis.

Na próxima seção estuda-se uma alternativa bem natural para a formação do precondicionador atualizado H_{i+1} que satisfaz as três condições acima. Neste ponto, exemplifica-se algumas escolhas de normas.

Para uma norma elíptica qualquer com sua matriz $B \in L(\mathbb{R}^n)$ tem-se com $\phi = \frac{1}{2}\|e\|_{B^HB}^2$:

$$\begin{aligned} D_\phi(e_{i+2}, e_{i+1}) &= \frac{\tau_{i+1}^2}{2}\|H_{i+1}f_{i+1}\|_{B^HB}^2 = \frac{\tau_{i+1}^2}{2}\|f_{i+1}\|_{B^HH_{i+1}^HH_{i+1}B}^2 \\ &= \frac{\tau_{i+1}^2}{2}\|e_{i+1}\|_{B^HA^HH_{i+1}^HH_{i+1}AB}^2 . \end{aligned} \tag{5.16}$$

Quando $B = I$ a norma-ℓ_2 é obtida:

$$D_\phi(e_{i+2}, e_{i+1}) = \frac{\tau_{i+1}^2}{2}\|H_{i+1}f_{i+1}\|_2^2 = \frac{\tau_{i+1}^2}{2}\|f_{i+1}\|_{H_{i+1}^HH_{i+1}}^2$$

$$= \frac{\tau_{i+1}^2}{2} \|e_{i+1}\|^2_{A^H H_{i+1}^H H_{i+1} A} , \tag{5.17}$$

na qual pode se ver que se $H_{i+1} \to A^{-1}$ a norma-ℓ_2 do erro é assintoticamente minimizada. Na verdade, necessita-se apenas que $A^H H_{i+1}^H H_{i+1} A$, pelo menos no instante da interrupção do algoritmo, provoque uma menor separação do erro do que $A^H A$, e isto não é difícil de conseguir com as informações retiradas das iterações.

Conforme estabelecido anteriormente, os algoritmos de projeção têm a propriedade de que $\tau_{i+1} \to 1$, e isto implica que $\|s_{i+1}\|_\bullet / \|e_{i+1}\|_\bullet \to 1$ (convergência finita, $\alpha \to 0$, se q-linear). A equação (5.16) diz que

$$\text{se} \quad \tau_{i+1} \to 1 \Rightarrow \|s_{i+1}\|_{B^H B} \to \|e_{i+1}\|_{A^H H_{i+1}^H B^H B H_{i+1} A} , \tag{5.18}$$

que na norma-ℓ_2 torna-se

$$\text{se} \quad \tau_{i+1} \to 1 \Rightarrow \|s_{i+1}\|_2 \to \|e_{i+1}\|_{A^H H_{i+1}^H H_{i+1} A} . \tag{5.19}$$

Mas, a mesma (5.16) também permite

$$\text{se} \quad \tau_{i+1} \to 1 \Rightarrow \|s_{i+1}\|_{B^H B} \to \|H_{i+1} f_{i+1}\|_{B^H B} . \tag{5.20}$$

Porém, esta condição é satisfeita pelo algoritmo padrão para qualquer função ϕ escolhida, e não apenas a medida epigráfica. Ele pode ser usado para ortogonalizar ϕ ou a sua função-D.

A solução para o parâmetro de busca é:

$$\tau_i = \frac{(d_i, d_i)_{B^H B}}{(H_i A d_i, d_i)_{B^H B}} \tag{5.21}$$

Uma interpretação da minimização da função-D em (5.15) e (5.16) bastante fértil, consiste em observar que a norma (a métrica) na qual o erro é minimizado é variável a cada iteração. Esta interpretação permite recorrer, antes de qualquer outra alternativa, aos métodos secantes para a construção do precondicionador H_{i+1}. Existe uma bem consolidada teoria e uma farta experiência numérica com esses métodos em análise não linear e otimização. Porém, eles possuem uma péssima reputação em sistemas lineares.

Nota-se pelas 5.16 e 5.17, que os métodos de curta recorrência com métrica variável (no conceito de [46], ver também Capítulo 4) minimizam a distância epigráfica da função erro, ou seja, ortogonalizam a função-D_ϕ onde $\phi = \frac{1}{2} \|e\|_B$ para algum B dependente das bases geradas, incluindo a inicial. Porém, 5.16 e 5.17 são mais gerais

e significativas. Elas permitem construir bases consistentes com as outras etapas do algoritmo.

Pode-se agora entender que o mecanismo da estratégia de minimização da distância epigráfica da função erro deve induzir a uma convergência do quociente. Uma vez que se está minizando $D_\phi(e_{i+2}, e_{i+1}) = \frac{1}{2}\|s_{i+1}\|_\bullet^2$, tem-se que

$$\|s_{i+1}\|_\bullet \leq \|s_i\|_\bullet \ , \text{ muito provavelmente.}$$

Espera-se que isto induza a uma contração iterada do mapeamento composto $\mathcal{A}^p$ para um p não distante de 1, se não mesmo com $p = 1$. Na pior das hipóteses, obtém-se convergência pelo menos q-linear local (durante as iterações) de $\mathcal{A}^p$, na norma $\|\cdot\|_\bullet$, que pode ser a norma-ℓ_2, ou alguma norma desejada pelo analista.

5.5.1 Uma nota sobre a convergência assintótica

Não por acaso, alguns resultados de clássicos de convergência podem ser aproveitados e reinterpretados. Segue-se Dennis e Moré [63] na interpretação do algoritmo padrão como um método tipo-Newton para análise não linear. O objetivo é verificar sob quais condições o algoritmo padrão possui convergência assintótica q-linear ou q-superlinear. Não se reproduzirá os resultados em detalhes, mas somente as interpretações mais importantes de interesse imediato, com as devidas adaptações.

Seja inicialmente $f(x)$ não linear, com o Jacobiano $f'(x_\star)$, e com as seguintes hipóteses básicas:

(a) $f(x)$ é continuamente diferenciável em um conjunto convexo e aberto D.

(b) Existe um $x_\star$ em D tal que $f(x_\star) = 0$ e $f'(x_\star)$ não é singular.

Seja a sequência $\{H_i\}_{i\geq 0}$ satisfazendo a condição de consistência $1a$, ou seja, cada H_i é não singular. Então existe uma sequência de inversas $\{A_i\}_{i\geq o}$ tal que cada $A_i = H_i^{-1}$. O Teorema 3.1 de [63] diz que, dentro das hipóteses básicas, se para algum x_0 a sequência

$$x_{i+1} = x_i - H_i^{-1}(x_i) \ , \quad i = 0, 1, \dots \ ,$$

permanece em D, e $x_i \neq x_\star$ para $i \geq 0$, e converge para $x_\star$, então a sequência $\{x_i\}_{i\geq 0}$ converge de forma assintoticamente q-superlinear na norma $\|\cdot\|_2$ para $x_\star$ se, e somente se,

$$\lim \frac{\|[(A_i - f'(x)]s_i\|_2}{\|s_i\|_2} = 0 \ .$$

A interpretação conveniente desta última condição é que A_i não precisa convergir para o Jacobiano, no caso a matriz A, mas apenas ao longo das direções sucessivas

s_i, assintoticamente, ou seja,

$$A_i s_i \rightarrow A s_i \ .$$

A prova de convergência quadrática exige a satisfação de hipóteses adicionais, e não é de interesse imediato.

5.6 Escolha do Precondicionador

A sequência $\{H_i\}_{i\geq 0}$ é o ator principal na estratégia montada, que a esse respeito não é nada nova. Coletar as propriedades essenciais do algoritmo em um único elemento facilita a análise, a comparação e a identificação de potenciais candidatos interessantes. A idéia é incluir várias possibilidades matematicamente consistentes que possam ser combinadas em estratégias diversas, algumas das quais serão exemplificadas.

O objetivo inicial são as matrizes não Hermitianas, cheias, como as encontradas nas discretizaçãcs pelo Método dos Elementos de Contorno. Dentre diversas possibilidades analisadas, as atualizações de posto unitário não simétrico (de fato, não Hermitiano) de Broyden (GB - *Good Broyden's Method*) mostrou-se a mais consistente. Na próxima subseção analisa-se o GB com um pouco mais de detalhes, como um exemplo típico da análise que deve ser feita, e em seguida descreve-se sumariamente outras possibilidades, em particular as características relevantes de um GB-epigráfico.

5.6.1 O método de Broyden e variantes

Um candidato em potencial, pela sua capacidade de geração de bases com aproximações consistentes e flexibilidade, é o método de Broyden. As razões torna-se-ão claras no decorrer desta subseção.

Conceitos e resultados gerais

A equação de atualização do resíduo é:

$$f_{i+1} = f_i + A s_i \Rightarrow A s_i = y_i \tag{5.22}$$

Porém, a matriz A não é a única que satisfaz esta equação na i-ésima iteração. Seja $Q_{i+1}(\mathbb{R}^n) \subset L(\mathbb{R}^n)$ o subespaço de matrizes A_{i+1} que satisfazem (5.22) na

i-ésima iteração, ou seja,

$$A_{i+1}s_i = y_i. \tag{5.23}$$

Esta equação é chamada em análise não linear de equação secante, representando um modelo linear local para a não linearidade. Em sistemas lineares ela não é uma aproximação, o argumento presente é que a sua satisfação é ainda mais essencial.

Em princípio, qualquer $A_{i+1} \in Q_{i+1}(\mathbb{R}^n)$ é uma boa candidata para aproximação de A na i-ésima iteração. Broyden sugeriu que, ao se dispor de uma boa aproximação A_i para A, não existem razões para que A_{i+1} difira de A_i no complemento ortogonal de s_i, ou seja: se $z \in \mathbb{R}^n$ e

$$(z, s_i)_2 = 0 \quad \text{então} \quad A_{i+1}z = A_i z\ . \tag{5.24}$$

O que Broyden certamente supôs é que a direção s_i é uma boa direção. Assim, a condição acima cstá sujeita à qualidade dc s_i como direção suficientemente descendente. De passagem, note que se a condição de ortogonalização $(e_{i+1}, s_i)_2 = 0$ for satisfeita, então $A_{i+1}e_{i+1} = A_i e_{i+1}$.

As equações (5.23) e (5.24) determinam A_{i+1} univocamente, como uma atualização de posto unitário de A_i:

$$A_{i+1} = A_i + \frac{(y_i - A_i s_i)s_i^H}{(s_i, s_i)_2}\ , \tag{5.25}$$

que caracteriza o famoso GB - *Good Broyden 's Method.* Pode-se inverter (5.25) usando a fórmula de Sherman-Morrison [63] e obter a conhecida atualização da inversa

$$H_{i+1} = H_i + \frac{(s_i - H_i y_i)s_i^H H_i}{(s_i, H_i s_i)_2}\ , \quad \text{se} \quad (s_i, H_i s_i)_2 \neq 0. \tag{5.26}$$

Dennis e Moré [63] também mostram que a sequência $x_{i+1} = x_i - A_i^{-1}f_i$, com A_i dada pela (5.25) para sistema lineares, possue convergência q-superlinear global e assintótica. Mostram também que, sob condições pouco restritivas, a sequência $x_{i+1} = x_i - H_i^{-1}f_i$ para problemas não lineares converge de forma pelo menos assintoticamente q-linear.

Não é necessário impor a (5.24), obtendo-se então uma expressão mais geral:

$$H_{i+1} = H_i + (I - H_i A)\frac{s_i v_i^H}{v_i^H H_i A s_i} H_i\ , \quad \text{com} \quad v_i^H H_i A s_i \neq 0, \tag{5.27}$$

e $v_i \in \mathbb{R}^n$. Esta última recorrência é conhecida como a atualização geral secante de posto unitário, c é devida também a Broyden. A determinação de $v_i \in \mathbb{R}^n$ deve ser

feita com hipóteses adicionais. No presente contexto a escolha natural de v_i se dá com a condição $(z, s_i)_\bullet = 0$, como alternativa à tradicional $(z, s_i)_2 = 0$.

Em todos os casos acima, mesmo que H_0 seja Hermitiana, em geral os sucessivos elementos H_i não o serão. Porém, não existe qualquer razão para que o GB não possa ser usado também para problemas Hermitianos, à parte o excesso de memória auxiliar em relação ao CG.

O GB é a solução do seguinte problema, chamado de condição de mudança mínima de B_i sujeito à equação secante:

$$\min_{B_{i+1} \in Q_{i+1}(\mathbb{R}^n)} \|B_{i+1} - B_i\|_F \ . \tag{5.28}$$

A norma é a de Frobenious, definida por

$$\|A\|_F^2 = \sum_{j=1}^{n} \|Av_j\|_2^2 = tr(A^T A) = \sum_{j=1}^{n}\sum_{k=1}^{n} |A_{jk}|^2 \ ,$$

onde $\{v_j\}_{1 \le j \le n}$ é qualquer conjunto ortonormal em $\mathbb{R}^n$. A solução é única porque o problema é convexo. Se a norma escolhida for a norma-ℓ_2, então existem muitas soluções. Normalmente, não se considera adequado usar a norma-ℓ_2 em (5.28) já que algumas soluções não se comportam bem [19, pag. 172].

Dentro das perspectivas desta tese, de uso de distâncias epigráficas, note que:

$$D_\phi(e_{i+2}, e_{i+1}) = \frac{\tau_{i+1}^2}{2} \|H_{i+1} f_{i+1}\|_2^2 \le \frac{\tau_{i+1}^2}{2} \|H_{i+1}\|_2^2 \|f_{i+1}\|_2^2 \le \frac{\tau_{i+1}^2}{2} \|H_{i+1}\|_F^2 \|f_{i+1}\|_2^2 ,$$

e, portanto, é possível que exista uma bem comportada solução de (5.28) com a norma-ℓ_2, mais adequada, a ser estudada. Porém, é senso comum que o problema de aproximações de matrizes com norma de Frobenius é mais adequado, uma vez que são elementos de $\mathbb{R}^{n^2}$.

Implementação

São possíveis diversas implementações, a depender da arquitetura do sistema computacional. Em geral, a melhor implementação é alguma variação de uma devida a Deuflhard, Freund e Walter [64]. Em lugar de fazer uma atualização sobre H_i, procura-se economizar um vetor auxiliar efetuando a atualização desde H_0.

Algoritmo GB: Algoritmo de atualização de H_i dentro da iteração i:

1. tendo no ínicio das iterações: $f_0,\ d_0 = -H_0 f_0,\ \sigma_0 = d_o^H d_0$

2. definindo valores parciais durante a iteração i: $q_i = Ad_i,\ \tilde{z}_0 = H_0 q_i$

3. para $j = 0, 1, \ldots i-1$, recupere H_i

3.1 $\widetilde{z}_{j+1} = \widetilde{z}_j + \frac{d_j^H \widetilde{z}_j}{\gamma_j t_j}(d_{j+1} - (1-\tau_j)d_j)$

4. atualize H_i

4.1 $z_i = \widetilde{z}_i$

4.2 $\gamma_i = d_i^H z_i$

4.3 $t_i = \sigma_i / \gamma_i$

4.4 $\tau_i \leftarrow$ parâmetro de busca (ver abaixo)

5. $d_{i+1} = (1 - \tau_i + t_i)d_i - t_i z_i$

6. $\sigma_{i+1} = d_{i+1}^H d_{i+1}$

O parâmetro de busca τ_i é função do que se vai minimizar ou ortogonalizar. Se for a função-D, usa-se a (5.21) para um B Hermitiano positivo definido. No caso da norma-ℓ_2 do erro, usar $B = I$. Pode-se no entanto escolher qualquer função ϕ adequada (em princípio, convexa). Os resultados ótimos serão obtidos com a função-D da norma-ℓ_2 do erro.

Bons resultados para sistemas bem comportados são obtidos mesmo com $\tau_i = 1$ para todo $i \geq 0$. Este é o uso clássico. Neste caso, prova-se que o GB possue convergência finita com no máximo $2n$ iterações.

Ortogonalizando a função-D da norma-ℓ_2 do erro (ou seja, usando (5.21) com $B = I$) tem-se a propriedade imediata, para todo $i \geq 0$,

$$H_{i+1} f_{i+1} = H_i f_{i+1} \Rightarrow (f_{i+1}, s_i)_2 = 0 \Rightarrow d_{i+1}^H d_i = 0 \ , \tag{5.29}$$

pela qual se conclui que não havendo *breakdown* a convergência finita se dará com no máximo n iterações. Em flagrante constraste com o GB clássico. É bom relembrar que ao se falar de convergência finita exclui-se aproximações de aritmética finita. Conclui-se também que $H_i \to A^{-1}$, com $H_n = A^{-1}$ se não tiver havido convergência antes. Certamente estes resultados traduzem a maior estabilidade do Broyden epigráfico.

O algoritmo GB pode ser trivialmente modificado de tal forma que permita a atualização a partir de um H_{i-m} qualquer, caracterizando um método memória limitada, onde apenas os últimos m vetores auxiliares são retidos. Normalmente,

a recuperação deve incluir H_0. Outras variações são possíveis. Potenciais candidatos incluem a recuperação seletiva baseada na qualidade da descendência de cada iteração, e a reinicialização. Mas, a principal função do GB na estratégia a ser montada é reter as melhores informações possíveis, apenas a ortogonalização seletiva deve ser usada por ocasião da recuperação de H_i. Pode-se considerar a reinicialização como opção se o GB for usado isoladamente.

A memória auxiliar necessária não é maior do que $(i_{conv}+3)n$ palavras (reais ou complexas), onde i_{conv} é a iteração de parada. A implementação de memória limitada toma $(m+3)n$ palavras.

Na estratégia de polialgoritmos proposta, deve-se ter um cuidado especial com o GB. O GB deve sempre estar no ciclo mais externo. Isto significa que no seu uso como atualizador de precondicionadores, ele deve ser usado exclusivamente como precondicionador à esquerda - como precondicionador à direita o GB não é invariante à transformação de variáveis (não obedecerá mais à equação secante). O chamado BB - *Bad Broyden's method* é invariante à transformação de variáveis. O BB resulta da escolha $v_i = y_i$ na fórmula geral de atualização secante de posto unitário. Mas isto não faz dele o método de escolha para estratégia proposta.

5.6.2 Sequências $\{H_i\}_{i\geq 0}$ de alguns algoritmos

Supõe-se que deva ser útil fornecer indicações de sequências $\{H_i\}$ de métodos conhecidos. Gibson, O'Leary e Nazareth [65] consideram uma certa forma geral para minimização de formas quadráticas com métodos de memória limitada

$$H_{i+1} = \gamma_i P_i^T H_0 Q_i + \sum_{k=1}^{m_i} w_{ki} z_{ki}^T \tag{5.30}$$

onde H_0 é simétrica e positiva definida, γ_i um parâmetro de reescalonamento, $P_i = I - \frac{uv^T}{u^T v}$ é um projetor ortogonal, $Q_i \in L(\mathbb{R}^n)$ é um produto de projetores ortogonais da mesma forma de P_i, e w_{ki}, z_{ki} são vetores reais. Nesta forma geral enquadram-se o *steepest descent,* o método dos gradientes conjugados, o BFGS, o L-BFGS, o DFP, e o L-DFP, e toda a classe de Broyden simétrico, entre eles o SR1.

Resultados preliminares analíticos e numéricos efetuados pelo autor deste trabalho dá uma grande perspectiva para o SR1. O SR1 tem propriedades hereditárias em problemas Hermitianos similares ao GB em problemas não Hermitianos. É o grande candidato de posto unitário para minimizar a função-D de $\phi = \frac{1}{2}\|e\|_2^2$. O SR1 é invariante sob transformação de variáveis, podendo servir de atualizador de

precondicionadores para o CG, que exige precondicionadores da forma $T^H AT$ onde $H = T^H T$.

Para sistemas não simétricos, Eirola e Nevanlina [66] fornecem um método interessante, e sua relação de recorrência na forma aqui usada, bem como a do GMRES. Ambos são atualizações de posto unitário obedecendo a uma equação do tipo secante transposta:

$$A^H z_i = A_{i+1} z_i. \tag{5.31}$$

O GMRES resulta da escolha $z_i = Ad_i$.

5.7 Recursividade e Simultaneidade

Nas seções anteriores foram esclarecidos e exemplificados os ingredientes do algoritmo padrão. O protótipo foi o GB dada às suas características adequadas. Em princípio, qualquer algoritmo pode ser feito epigráfico - minimizar a função-D de uma função de Bregman. O GB epigráfico, em particular com a função-D para a norma-ℓ_2 do erro, oferece grandes perspectivas. O SR1 epigráfico tem potenciais enormes para sistemas Hermitianos. O próprio CG pode ser feito epigráfico, porém, o CG já possui a grande propriedade de ser estritamente descendente na norma-ℓ_2 do erro, e portanto, conforme os resultados da primeira parte desta tese, é pelo menos q-linear em duas etapas nesta norma. Tais métodos são, em princípio, métodos de métrica variável, em especial quando obedecem a equação secante. O algoritmo padrão também aceita qualquer função ϕ tradicionalmente usada.

Nesta seção mostra-se como construir polialgoritmos adequados sob ciclos internos do algoritmo padrão. Existem várias possibilidades. De uma forma geral pode-se dividir as possíveis estratégias em dois grandes grupos: estratégias que usam o algoritmo padrão recursivamente, e aquelas que o fazem simultaneamente. Em qualquer caso, o sucesso da estratégia certamente depende de alguma análise antecipada que a justifique.

Nesta seção faz-se uma análise global, e não muito detalhada, de algumas alternativas para a recursividade e a simultaneidade da realização de algoritmos padrões. Especificamente,

1. Atualização de precondicionadores
2. Hibridização
3. Regularização

Neste ponto deve-se salientar que alguns dos mecanismos mais importantes das estratégias propostas adiante tornam-se possíveis graças à (suposta) eficiência das estimativas de erros oriundas da teoria de convergência do quociente desenvolvida na primeira parte desta tese.

Todas as estratégias sugeridas são matematicamente consistentes, e decorrentes das análises efetuadas nos capítulos precedentes. As possibilidades são muitas, e várias delas ou não foram ainda suficientemente testadas, ou os testes limitaram-se a poucas realizações.

5.7.1 Atualização de precondicionadores

Quando se está constantemente resolvendo problemas dentro de uma certa classe específica, a melhor maneira de conseguir um bom precondicionador H_0 é por análise das características do problema contínuo. É provável que por experiência numérica sistemática chegue-se a um precondicionador muito perto do ideal - caracterizado por qualidade e economia de memória. Porém, em uma realização específica, certamente existe um precondicionador de melhor qualidade, satisfazendo ainda a restrição de economia. Faz-se um paralelo entre o fator-q do processo e o fator-q da realização, (muito) provavelmente menor (ver Capítulo 3).

Outras situações existem nas quais ainda não se dispõe senão de um precondicionador razoável, talvez mínimo, digamos, o precondicionador de Jacobi caracterizado por $H_0 = \text{diag}(A_{ii}^{-1})$, ou em casos extremos, $H_0 = I$.

A questão que se põe é a de como melhorar, durante a realização, o precondicionador original. Existem muitas possibilidades. Uma delas é usar dois algoritmos, recursivamente, em um esquema *inner-outer loop*, típico da solução de um problema não linear. Os algoritmos podem ser diferentes ou, eventualmente, iguais.

Uma questão complementar é: tendo feito uma realização, como aproveitar as informações extraídas desta realização em outras, como por exemplo, os casos de múltiplos carregamentos, análise não linear e análise dinâmica.

As questões acima expostas estão diretamente associadas à robustez. Espera-se, de alguma forma, que o algoritmo obtenha uma resposta aceitável.

Adicionalmente, se estamos usando, por exemplo, um algoritmo que minimize a norma-ℓ_2 dos resíduos (impondo consistência), e deseja-se qualidade na norma-ℓ_2 do erro, é razoável esperar que o precondicionador atue no sentido de relaxar a consistência, se necessário, e melhorar a qualidade da norma-ℓ_2 do erro, se não ambas. Este normalmente é o caso quando o resolvedor linear é um *inner-loop*

na solução de um problema não linear usando o método de Newton, ou métodos inexatos de Newton.

A idéia de precondicionador atualizado, matematicamente consistente, vem de encontro àqueles anseios. Para matrizes não Hermitianas, o GB possui uma série de características interessantes para cumprir aquelas funções. Dado um precondicionador inicial H_0, regular, submete-se o precondicionador ao GB, na ausência de melhores alternativas.

Exemplifica-se a estratégia, na hipótese de que o método principal é um método de Krylov que minimiza a norma-ℓ_2 do resíduo, $\phi(f) = \frac{1}{2}\|f_i\|_2^2$, e é de interesse o controle da norma-ℓ_2 do erro, $\phi(e) = \frac{1}{2}\|e_i\|_2^2$. Algumas situações possíveis são:

1. *Existe alguma memória* $\simeq nM$ *adicional disponível.* Realizar o GB(epi$\phi(f)$) completando a sequência $\{H_i\}_{0\leq i\leq m}$, com $m \leq M$. O critério de interrupção pode ser um percentual de $\|e_0\|_2$. Efetuando também o controle de $\|f_i\|_2$, eventualmente, o BG resolverá o problema sozinho. Se não, inicializa-se o método de Krylov com a última (opcional) solução aproximada x do GB e o novo precondicionador $\{H_i\}_{0\leq i\leq m}$. Esta sequência pode ser salva para uso posterior com a mesma matriz A ou alguma aproximada. Note que a sequência de precondicionadores existe apenas implicitamente, através de seus vetores auxiliares.

2. *Não existe memória adicional disponível.* Neste caso supõe-se que H_0 já é um precondicionador mais sofisticado. Efetua-se diretamente a atualização de posto unitário sobre o precondicionador, eventualmente, satisfazendo alguma restrição de esparsidade explicitada por $S(j,k)$:

$$(H_{i+1})_{jk} = (H_i)_{jk} + (u_j w_k)_i \quad se \quad j,k \in S(j,k) \tag{5.32}$$

Uma opção mais consistente é especificar um valor limite $\epsilon > 0$, abaixo do qual cada elemento de cada vetor é considerado nulo. Os vetores auxiliares serão armazenados sob forma compactada, convencionalmente. Não se fez análise maior desta última opção, e apenas alguns testes não conclusivos. Certamente outras opções existem, a depender de como o precondicionador inicial se apresenta. Existem opções para atualização de posto unitário de decomposições típicas como a LU.

5.7.2 Hibridizações

Maior a descendência de uma direção d_i, mais importante as informações contidas nos vetores auxiliares da iteração. No esquema anterior, o precondicionador é atualizado a cada iteração. Isto exige que o GB, ele próprio, seja suficientemente

descendente. Em certos problemas, é possível que a iteração do GB não produza informações suficientemente relevantes, e não há porque guardá-las, desperdiçando memória auxiliar. Às vezes, até mesmo prejudicando a estabilidade do GB. Um teste de suficiência da descendência deve existir. A atualização do precondicionador apenas será efetuada sob essas condições. Esta técnicas são conhecidas em otimização.

Em casos como este, é mais conveniente alternar entre os algoritmos. No caso da seção anterior, entre o GB e o método de Krylov. Seja, para exemplificar, o GMRES(m). Inicia-se o GB com uma ou mais iterações, a depender de algum critério a ser discutido. Em um certo instante leva-se a mais recente aproximação de x juntamente com a sequência $\{H_i\}$ ao GMRES(m). Executado o ciclo do GMRES, retorna-se a mais um ciclo do GB, do qual resulta uma ampliação da sequência $\{H_i\}$. Se o objetivo é minimizar o número de produtos matriz-vetor, pode-se efetuar a atualização com os incrementos s_i e y_i resultantes de um ciclo do GMRES(m).

O esquema discutido é executado debaixo de dois tipos de critérios. Critérios de descendência, pelos quais se verificam normas de interesse e consistência, e critérios de economia, pelo qual se limita a memória auxiliar adicional. Em casos extremos de pouca memória auxiliar disponível pode-se reinicializar o GB, mantendo é claro, o precondicionador original. Esquemas intermediários incluem a manutenção de bases (vetores auxiliares) selecionadas por critério de descendência.

5.7.3 Regularização

Existem algoritmos que possuem uma incrível irregularidade na convergência, e mesmo assim são usados com muita frequência. Normalmente esta irregularidade é medida pela norma-ℓ_2 dos resíduos. O CGS é um deles. Usado com frequência em problemas não lineares, em substituição ao BiCG. Quando o BiCG diverge, o CGS também diverge. Quando o BiCG converge, o CGS converge duas vêzes mais rapidamente. Sendo um método de curta recorrência, é frequentemente procurado como possível alternativa econômica.

A literatura tem oferecido diversas técnicas para abrandar tais problemas. A mais interessante é o *smoothing* do Schonauer. Em paralelo às iterações, são formadas duas sequências de $\{\widehat{x}_i\}$ e $\{\widehat{f}_i\}$ decorrentes da minimização dos resíduos originais. O TFQMR e suas variantes são matematicamente equivalentes ao CGS.

A melhor solução que se pode oferecer para o CGS, o BiCG, e outros desta classe, é precondicioná-lo com o GB. Mas, isso exige memória adicional, e não retira de todo a indesejável oscilação da norma-ℓ_2 dos resíduos.

Uma alternativa é adicionar a tais métodos um processo complementar de ortogonalização em alguma norma, que não a que ele minimiza (se minimiza). No caso do CGS, por exemplo, a norma-ℓ_2 dos resíduos, ou a função-D da norma-ℓ_2 do erro. Isto pode ser feito via minimização proximal com funções-D [28], ou mais simplesmente, com uma combinação convexa de normas de interesse:

$$\phi = (1-\lambda)\phi_1 + \lambda\phi_2 \ , \qquad 0 \leq \lambda \leq 1 \ ,$$

ou, de forma generalizada

$$\phi = \sum \lambda_k \phi_k \ , \qquad \sum \lambda_k = 1, \qquad \lambda_k \geq 0.$$

Presentemente, estuda-se uma generalização de tais procedimentos por meio de viabilidade convexa [29], objetivando implementações em arquiteturas com processadores simétricos (SMP).

Capítulo 6

Experimentos Numéricos

Os preceitos valem mais do que os conceitos.
Isaac Newton

6.1 Introdução

Neste capítulo são coletados experimentos numéricos com o objetivo de ilustrar alguns dos aspectos mais importantes relativos à teoria de convergência do quociente desenvolvida na Parte I, e às consequentes estratégias de solução propostas na Parte II. Foram muitas as questões discutidas nos capítulos anteriores, de tal forma que é virtualmente impossível ilustrar todas. Na Parte I, existem circunstâncias não consideradas da aritmética finita que podem se tornar predominantes. Em particular na Parte II, algumas estratégias sugeridas sequer foram suficientemente testadas. Algumas tomarão ainda muito tempo do autor para serem devidamente calibradas mesmo em uma área específica. A variedade de problemas para os quais alguns resultados deste trabalho podem eventualmente ter aplicação é muito ampla. Optou-se por cobrir algumas distintas, dando ênfase aos princípios os quais, supõe-se, são mais permanentes e não serão tão diluidos pelo tempo.

Em todos os experimentos, $macheps = 2.22 \times 10^{-16}$, com o zero da máquina igual a $macheps/2$, determinados numericamente.

6.2 Experimento 1 - Pórtico Plano, CG, GB+CG

O pórtico plano, engastado nas duas extremidades, da figura 6.1 é analisado para cargas estaticamente aplicadas e comportamento completamente linear. É composto

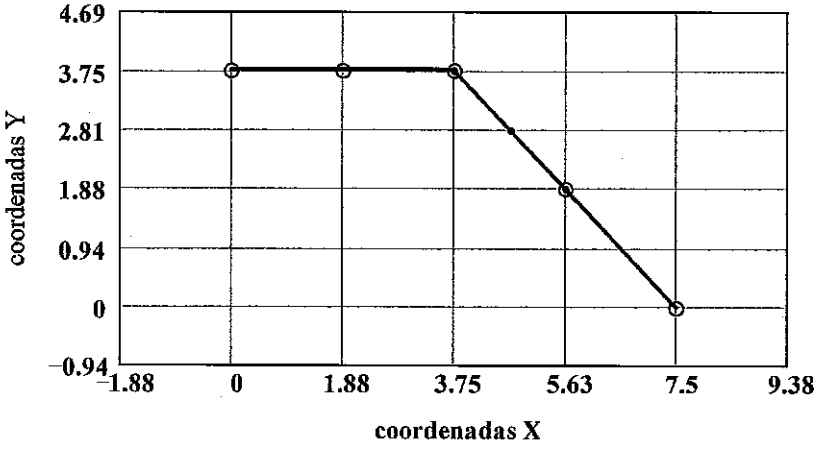


Figura 6.1: Pórtico plano do Experimento 1 - 4/12 elementos

de apenas duas barras, uma horizontal e outra inclinada. A discretização é feita com elementos convencionais (polinômios de Hermite), produzindo 3 graus de liberdade por nó, dois de flexão e um de membrana (axial), sem deformação por esforço cortante incluída na energia de deformação. A discretização com dois elementos fornece a solução exata com apenas 3 equações, com solução analítica simples. Os diagramas de esforços solicitantes internos (momento fletor, esforço cortante e esforço normal) e os deslocamentos (duas translações e uma rotação) do nó central são exatos. Devido às cargas externamente aplicadas aos elementos, os deslocamentos ao longo das barras não são aproximados exatamente pelos polinômios usados. Porém, cada ponto nodal acrescentado fornece os deslocamentos exatos neste ponto. Estes são fatos bem conhecidos. Não existe sentido em discretizar mais, a não ser o sistema esteja em um ciclo interno de um problema não linear ou um problema no tempo.

As propriedades escolhidas são reais, no sentido de que nenhum artifício foi imposto para gerar problemas numéricos. Seção transversal retangular constante de $0.20 \times 0.50\ m$ e módulo de Young igual a $2.0 \times 10^6\ tf/m^2$. A barra inclinada está carregada com um força horizontal de $4\ tf$ no seu centro, e uma carga horizontal uniformemente distribuida com $-0.35\ tf/m$. A barra horizontal está com uma carga vertical uniformemente distribuida de $-0.70\ tf/m$.

A barra inclinada estabelece um acoplamento entre os graus de liberdade de flexão e os de membrana. Este acoplamento é típico de estruturas como pórticos, arcos e cascas. É a causa de grandes dificuldades para a solução iterativa de sistemas lineares, quer isoladamente, quer dentro de ciclos de problemas não lineares.

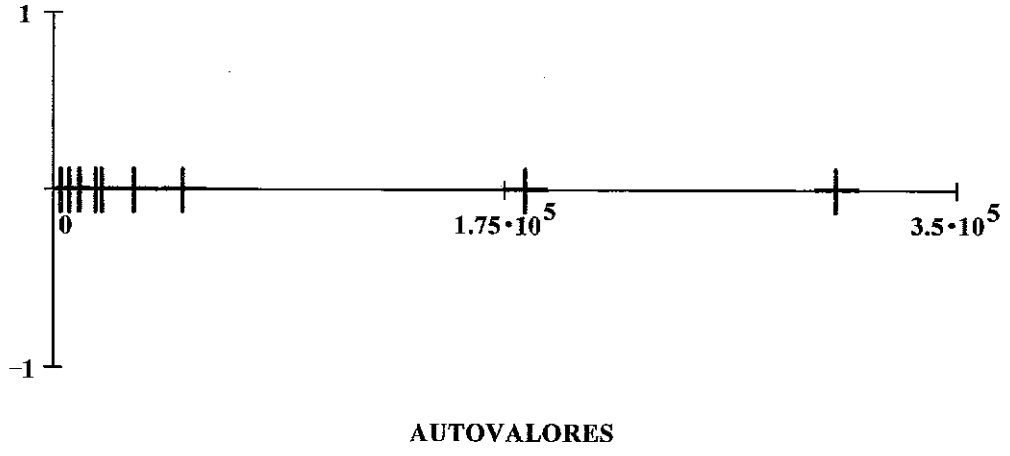


Figura 6.2: Espectro de autovalores para $n = 9$

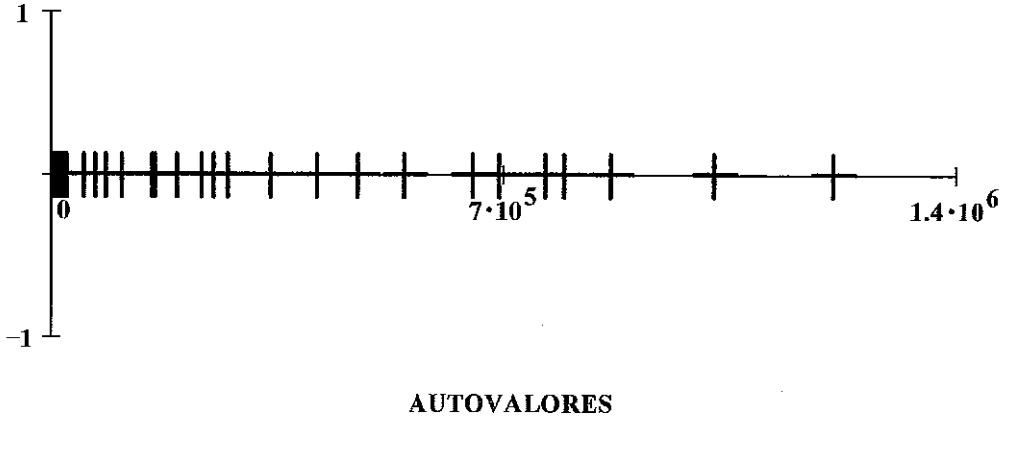


Figura 6.3: Espectro de autovalores para $n = 33$

Provêm da grande diferença entre a rigidez axial e a rigidez à flexão. Em problemas predominantemente de flexão, como placas medianamente espessas, poucas iterações são necessárias em cada ciclo não linear. Com o acoplamento membrana-flexão, o número de iterações pode crescer em pelo menos uma ordem de grandeza.

CG - comportamento e estimativas

São feitas 3 discretizações com $n = 9, 21, 33$ graus de liberdade, correspondentes a $5, 9$ e 13 nós, respectivamente, de tal forma que o tamanho de cada elemento em cada uma das barras sejam iguais. A razão mínima entre o comprimento e a altura da seção transversal quando $n = 21$ é de 1.25, e para $n = 33$ é 0.75. Com $n = 33$, os problemas de aritmética finita fazem o CG começar a perder as suas propriedades de convergência do quociente. O problema é simétrico e positivo definido, e suficientemente pequeno para que as respostas por um método de decomposição estável (LU , LDL, QR, Householder,...) possam ser consideradas exatas.

Os experimentos que se seguem são efetuados com as duas discretizações extremas, $n = 9, 33$. Nas figuras 6.2 e 6.3 estão os seus respectivos espectros (de autovalores). Nas figuras 6.4 e 6.5 estão os autovalores. Os autovalores predominantemente de flexão situam-se na parte inferior do espectro e formam agrupamentos (*clusters*). Isto é favorável ao CG. A parte do espectro responsável pelos modos predominantementes de membrana estão mais à direita e regularmente espaçados, um caso de extrema incoveniência para o CG. As componentes de resposta de maior conteúdo de energia são justamente as de flexão.

O CG converge com $i_\alpha = 0$ de forma q-superlinear nas duas realizações: para $n = 9$ os valores exatos são $\alpha = 0.888$, $\overset{\bullet}{\alpha}= 0.878$ e $\tilde{\alpha} = 0.503$ na norma da energia, e $\alpha = 0.878$, $\tilde{\alpha} = 0.503$ na norma-ℓ_2 . Note que não há $\overset{\bullet}{\alpha}$ no CG na norma-ℓ_2 (cada algoritmo pode ter apenas um $\overset{\bullet}{\alpha}$, por definição). Não é uma contração iterada desde $i = 0$, e β_i atinge valores superiores a $\sqrt{2}$ nas duas normas, da energia e ℓ_2.

Por minimizar o errro na norma da energia, maximiza a norma da energia dos deslocamentos, de tal forma que as diversas normas dos deslocamentos convergem rapidamente, em intensidade e direção. A figura 6.6 mostra isso para a norma-ℓ_2. Confirma o demonstrado no Capítulo 3: ξ_i (e portanto α_i) é a medida instantânea simultaneamente da convergência na norma e na direção (ângulo de aproximação com a solução exata) em métodos de projeção ortogonal (com respeito a algum produto interno).

Outros resultados fundamentais provados nesta tese podem ser visualizados neste

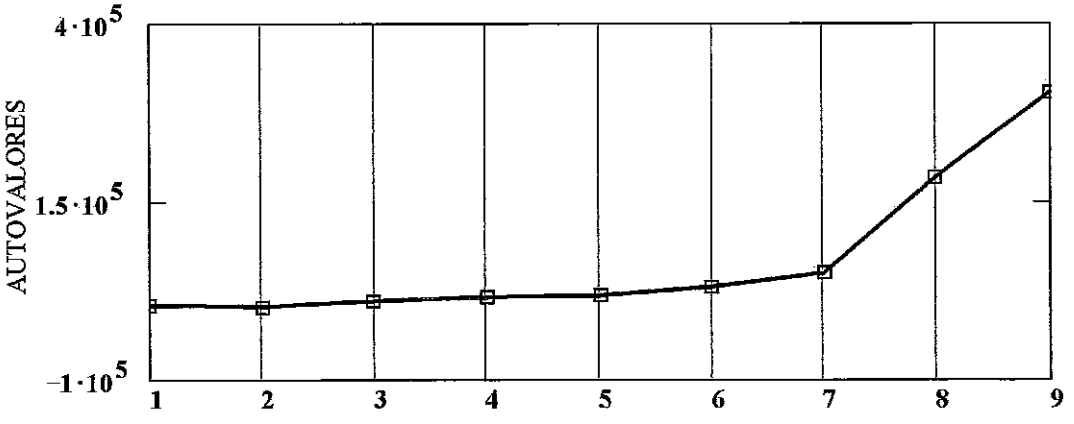


Figura 6.4: Autovalores para $n = 9$. $\lambda_{\max} = 3.034 \times 10^5$, $\lambda_{\min} = 2.954 \times 10^3$, $\kappa_2 = 102.7$

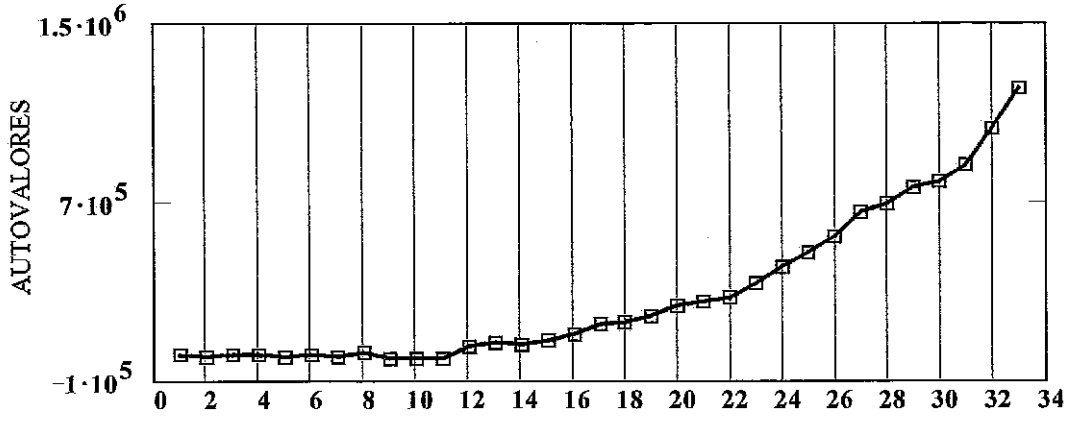


Figura 6.5: Autovalores para $n = 33$. $\lambda_{\max} = 1.212 \times 10^6$, $\lambda_{\min} = 0.975 \times 10^3$, $\kappa_2 = 1242.3$

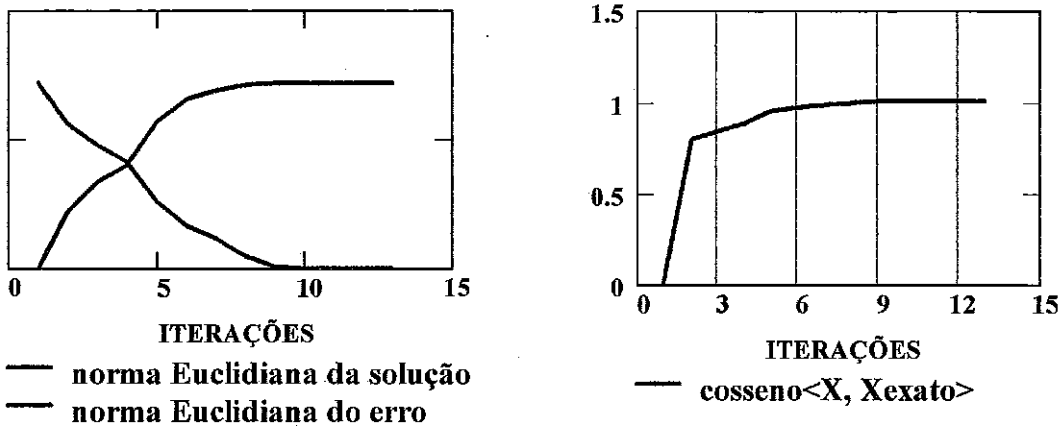


Figura 6.6: Convergência uniforme em norma e direção. Na norma da energia estas curvas são mais suaves.

experimento. A figura 6.7 mostra as estimativas calculadas com a sequência $\{\alpha_i\}$ exata usando o estimador para p enriquecimentos com $p = 1$. Note-se, em particular pela figura 6.8, que não é possível uma cota superior melhor do que a fornecida. Esta figura traça a razão entre a estimativa e o valor exato do erro em cada iteração, para $n = 9$. A figura 6.9 corresponde à figura 6.8 usando também o Lema fundamental de projeções ortogonais. Todos esses valores referem-se à norma da energia. Servem apenas para ilustrar a qualidade dos resultados, uma vez que foram calculados com os fatores-q exatos.

Os valores que interessam na prática são os determinados com estimativas das constantes, por exemplo de $\{\alpha_i\}$, feitas durante da realização. Interessam os valores relativos à solução exata. Não se dispõe da solução exata, e portanto deve-se usar a solução aproximada corrente. As estimativas da cota inferior pelo estimador para p enriquecimentos não usa nenhum valor a ser estimado. As figuras 6.10 e 6.11 mostram duas estimativas, a primeira o estimador para p enriquecimentos, $p = 1$, e a segunda usando $\overset{\bullet}{\alpha}_i$. A maior razão estimativa/exato foi de 2.156, ocorrido no início da realização. Usando $p > 1$ - quanto maior melhor - e isto sempre é possível, os resultados são bem melhores. Porém, neste exemplo, é desnecessário (embora não implique em qualquer custo ou dificuldade adicional) uma vez que $2.156 < \sqrt{10} < 10$, caracterizando a mesma ordem de grandeza. E além disso, no início das iterações, com valores altos do erro.

As figuras 6.12-6.16 correspondem às estimativas ainda para $n = 9$, na norma-ℓ_2.

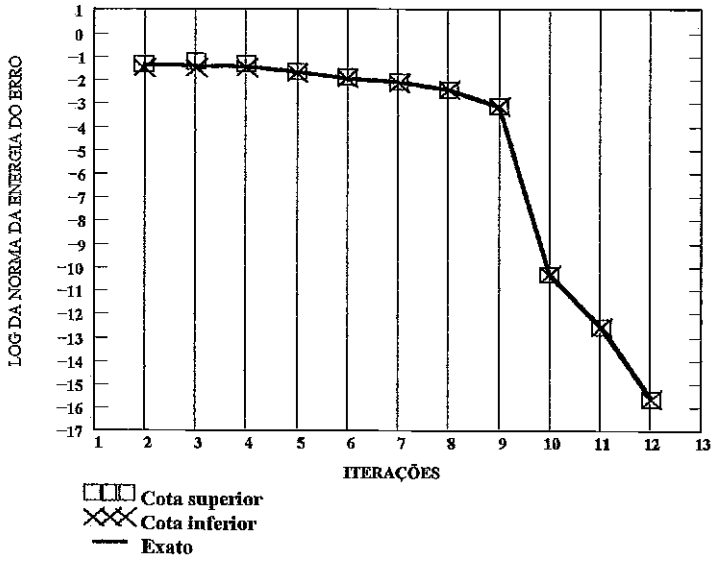


Figura 6.7: Estimativas da norma da energia do errro, com um enriquecimento (fator-q exato em cada iteração), para n=9

Ainda há q-superlinearidade, porém, como não há ortogononalização, os resultados correspondentes fornecem estimativas menos gratificantes. São resultados típicos de estimativas em outras normas do CG, que não a dá energia. Todos foram feitos ou com o Lema fundamental ou o estimador para p-enriquecimentos, $p = 1$. Resultados muito melhores são obtidos usando o estimador para p-enriquecimentos com $p > 1$. Idealmente, $i = 1$ e p crescente. Ou, preferencialmente usando o mapeamento $\mathcal{A}^p$, $p \geq 2$.

O sistema com 33 equações está no limear da instabilidade numérica. Na figura 6.17 percebe-se o aparecimento de patamares, quase q-sublinearidades. O CG já não consegue produzir direções suficientemente descendentes, porque cada novo elemento da base é quase uma combinação linear dos anteriores. É importante observar que o número de condição ainda é muito baixo. Os autovalores também. O espectro mostra que há uma grande quantidade de autovalores igualmente separados, e esta é a causa.

A grande surpresa está na qualidade dos estimadores, mesmo com apenas um enriquecimento. As figuras 6.17 e 6.18 foram feitas com a sequência $\{\alpha_i\}$ exata. As estimativas das cotas superiores mantêm a qualidade que possuiam no sistema com $n = 9$. As das cotas inferiores são pelo menos assitoticamente exatas. Para melhorar a qualidade é apenas necessário usar mais enriquecimentos. Na figura 6.19 estão os

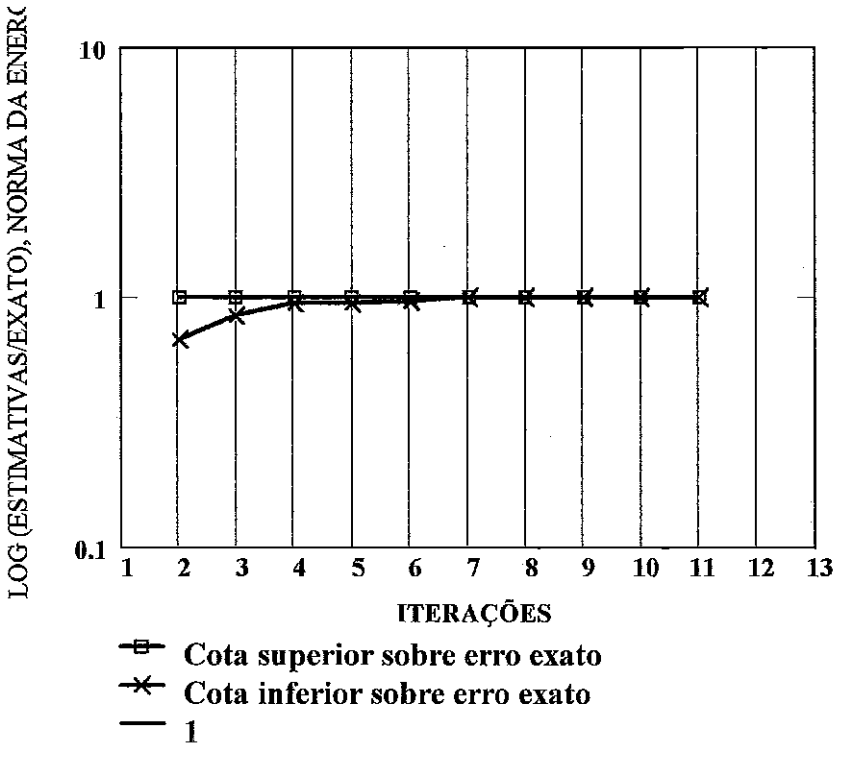


Figura 6.8: Estimativas da norma da energia do errro, com um enriquecimentos (fator-q exato em cada iteração), para $n = 9$

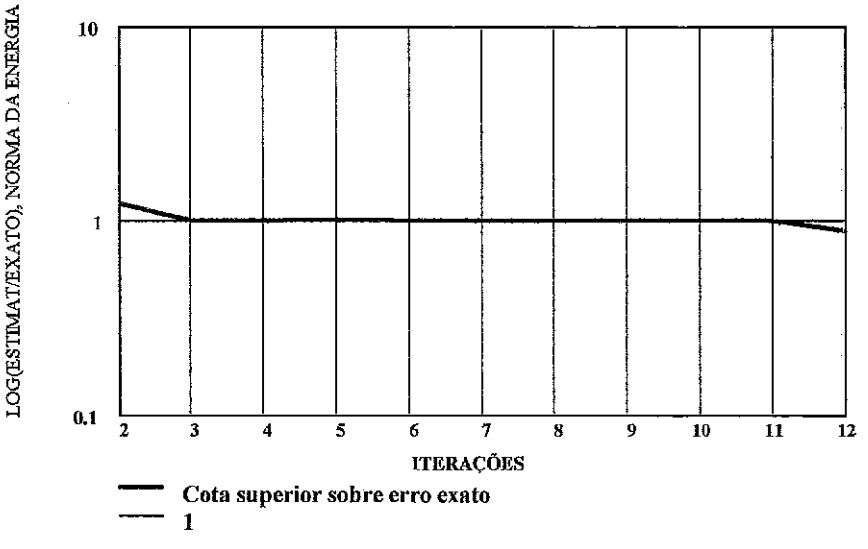


Figura 6.9: Lema fundamental e estimador para p enriquecimento, $p = 1$

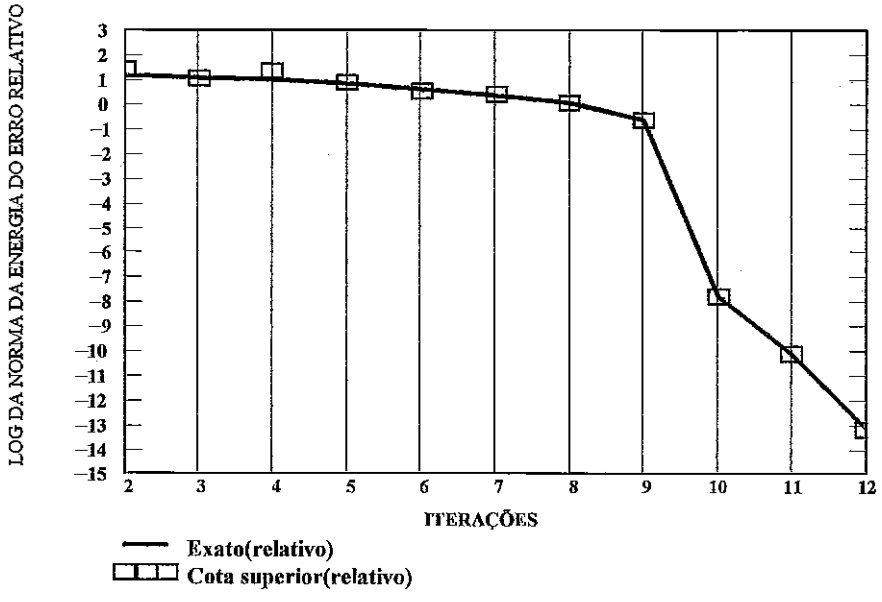


Figura 6.10: Estimativa usando apenas o Lema fundamental

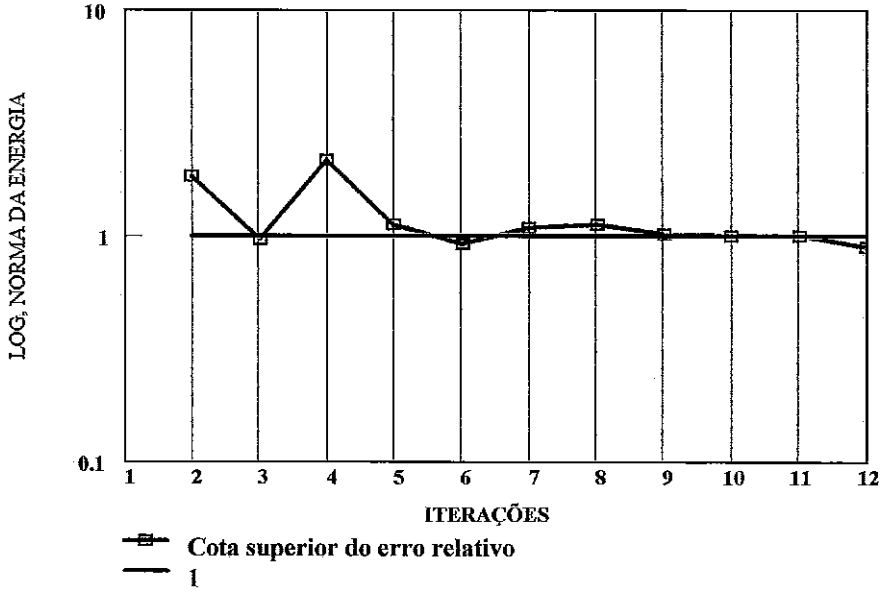


Figura 6.11: Estimativa usando apenas o Lema fundamental

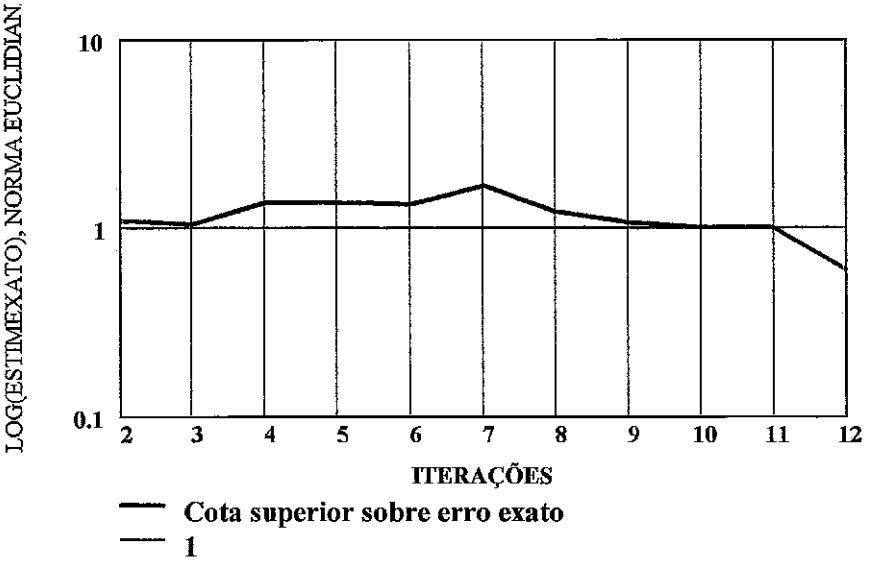


Figura 6.12: Estimativa com α exato. Norma-ℓ_2 do erro, $n = 9$.

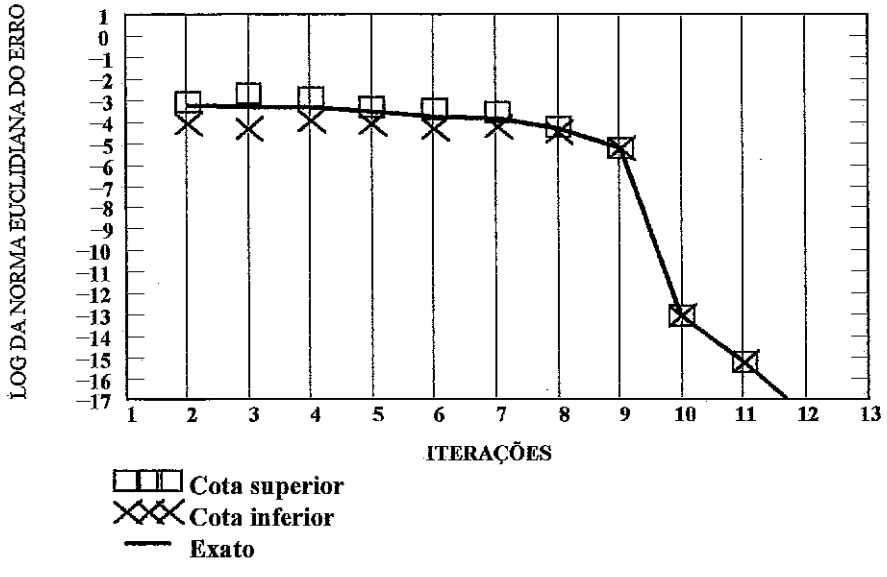


Figura 6.13: Estimativa com α exato e estimador com $p = 1$. Norma-ℓ_2 do erro, $n = 9$.

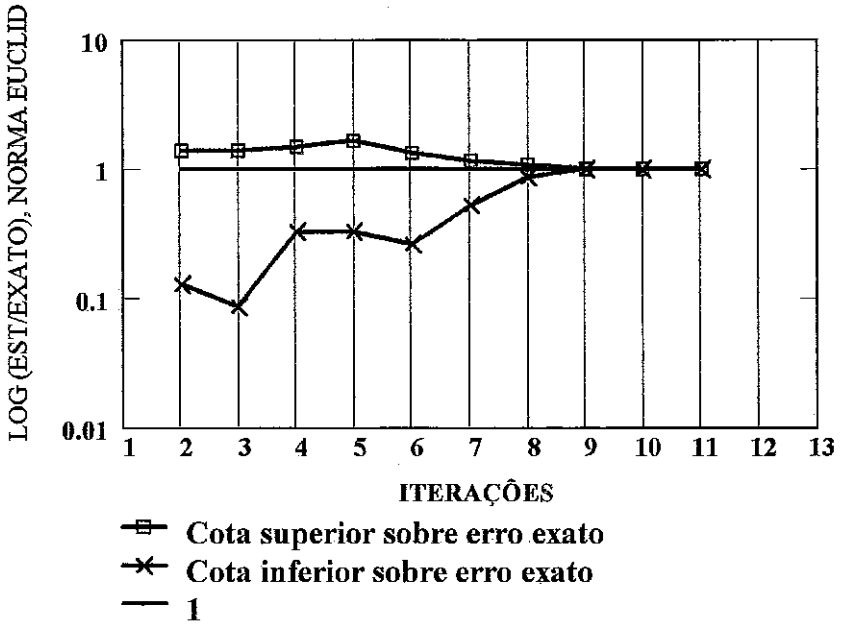


Figura 6.14: Estimativa com α exato e estimador com $p = 1$. Norma-ℓ_2 do erro, $n = 9$.

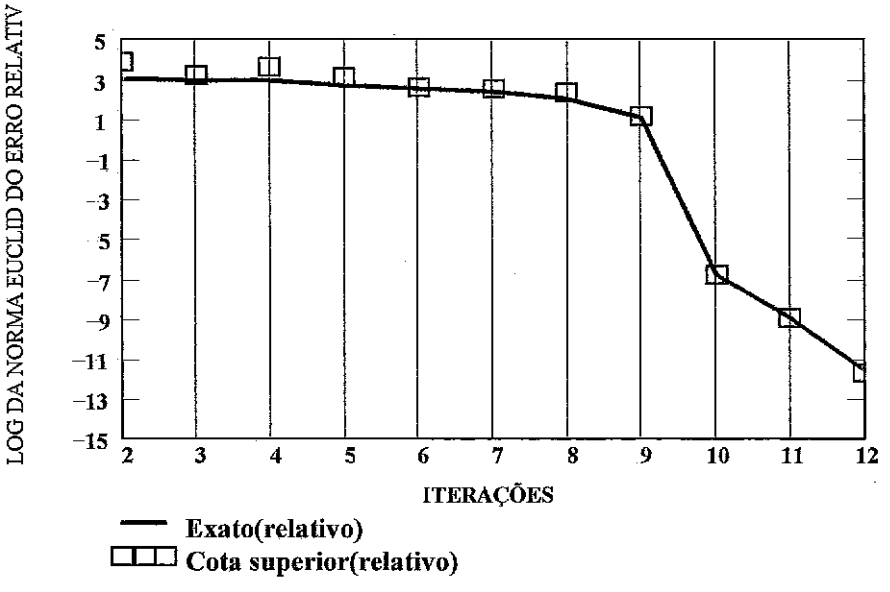


Figura 6.15: Corresponde à figura 6.10, agora na norma-ℓ_2.

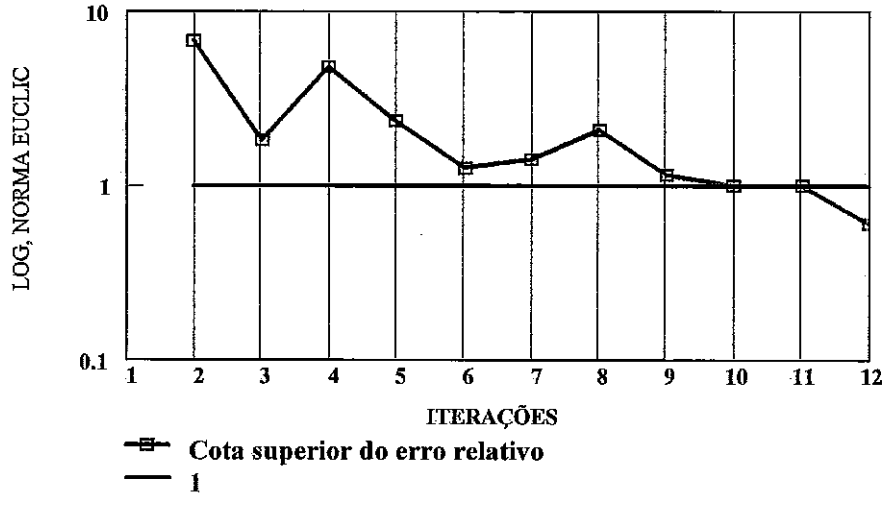


Figura 6.16: Corresponde à figura 6.11, agora na norma-ℓ_2,

erros na norma da energia com a sequência $\{\alpha_i\}$ estimada por $\left\{\overset{\bullet}{\alpha}_i\right\}$. Nas figuras 6.20 e 6.21 para a norma-ℓ_2, com a sequência exata, e com $\{\tilde{\alpha}_i\}$, respectivamente.

Os resultados oscilantes estão registrados aqui com o objetivo de mostrar a instabilidade do CG neste exemplo com 33 graus de liberdade. A técnica correta para sua estimativa de erros consiste em usar ou um mapeamento composto, o estimador com p-enriquccimentos, $p > 1$, preferencialmente sob a forma $\|e_{1+p}\|$, $p = 1, 2, \ldots$.

Seguem-se algumas recomendações práticas provenientes de numerosos testes com sistemas provenientes de discretizações pelo método dos elementos finitos, inclusive elementos undimensionais (barras), para o CG. A recomendação geral é não esquecer de que a norma da energia no CG precondicionado inclui o precondicionador.

Para as estimativas na norma da energia pouco há o que se acrescentar já que as estimativas são extremamente precisas, e pode-se, por exemplo, usar as recomendações que se seguem para a norma-ℓ_2, substituindo $\overset{\bullet}{\alpha}_i$ por $\tilde{\alpha}_i$.

A melhor recomendação prática, sem sofisticações, confiável e precisa, para as estimativas em normas que não a norma da energia, tem seus resultados (norma-ℓ_2) nas figuras 6.22-6.23 para $n = 9$, representando sistemas sem muitos problemas numéricos. Já no sistema com $n = 33$, o CG perdeu suas propriedades teóricas, pode-se então usar mapeamentos compostos com $p > 2$. O valor de p pode ser determinado durante a realização, e pode ser variável, e decrescente. Um teste simples de descendência é $\beta_k < 1$, onde k representa o índice das iterações no mapeamento composto. As figuras 6.22-6.23 foram feitas usando o mapeamento composto $\mathcal{A}^p$,

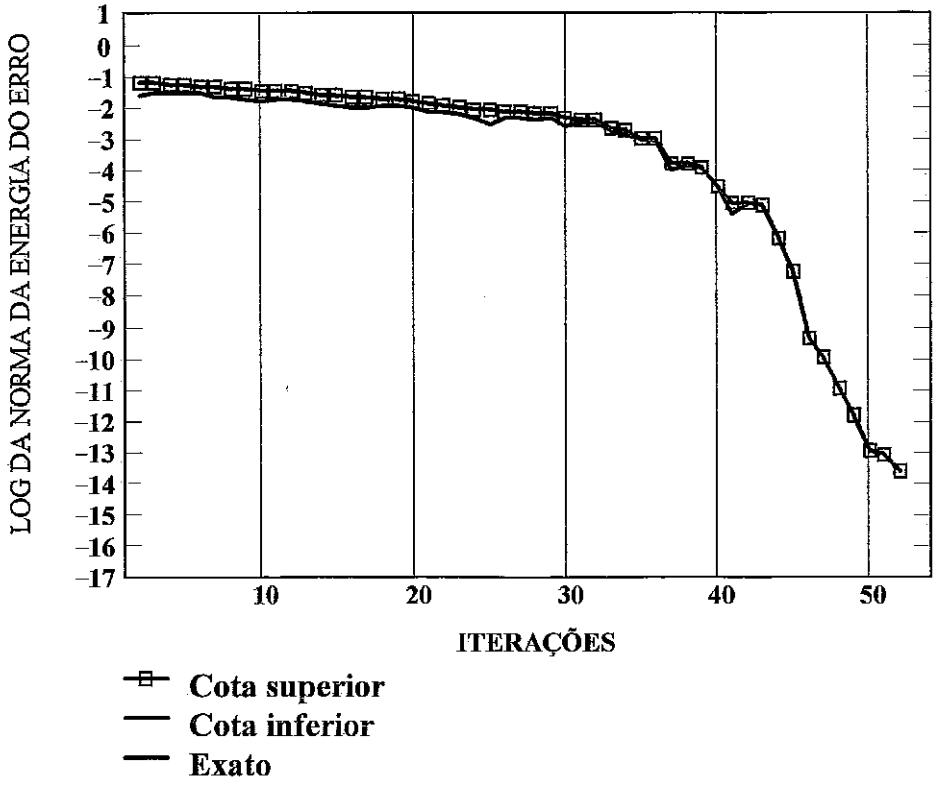


Figura 6.17: Estimativa com a sequência $\{\alpha_i\}$ exata, $n = 33$.

com $p = 2$. Foi demonstrado no texto desta tese que este é o valor correto. Não há necessidade de p maiores, e se escolhido $p = 1$, as estimativas sofrerão pequenas oscilações. Já para o sistema com $n = 33$, usou-se $p = 4$. Usar a estimativa $\tilde{\alpha}_k$ calculada com $\psi_k = 1$. Em seguida aplicar apenas a equação $\|e_{k+1}\|_2 = \frac{\beta_i}{1-\alpha_{k+1}} \|s_k\|_2$. Permite-se ainda usar $\tilde{\alpha}_k$ em vez de $\tilde{\alpha}_{k+1}$, produzindo valores mais conservativos. Existem outras possibilidades, dependendo da implementação particular do CG.

Em nenhuma das discretizações o CG manteve sua propriedade teórica de convergência finita em no máximo n iterações devido à aritmética finita, mesmo com poucas equações. Nos dois casos, não atingiu valores aceitáveis em qualquer norma de interesse para um número menor ou igual a n. No entanto o problema analisado é típico da engenharia estrutural. Existem muitos precondicionadores interessantes para este pórtico. Precondicionar sistemas com acoplamento membrana-flexão ainda é uma questão aberta, e tem impedido o uso generalizado de algoritmos iterativos para sistemas lineares, em particular em ciclos internos de problemas não lineares.

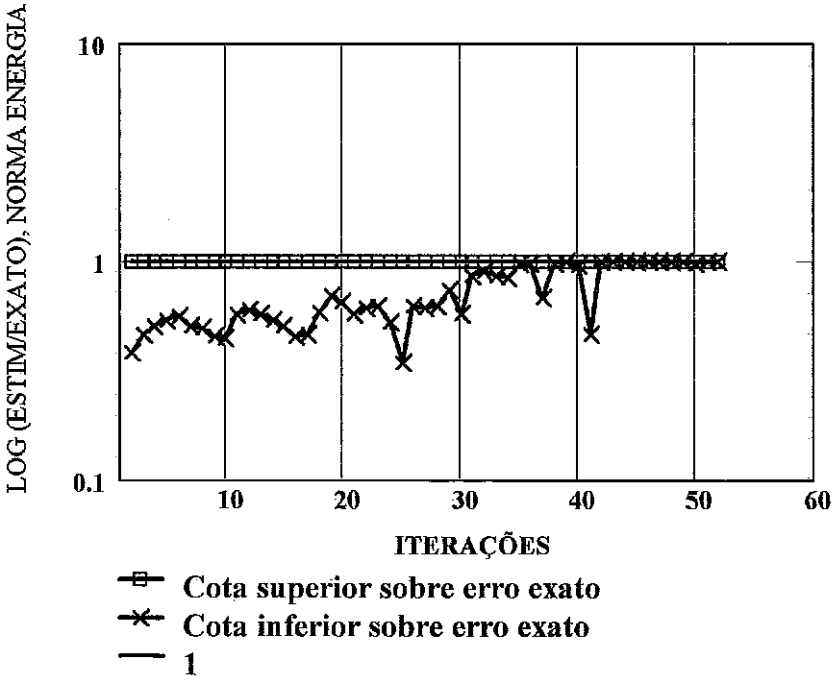


Figura 6.18: Estimativa com a sequência $\{\alpha_i\}$ exata, $n = 33$.

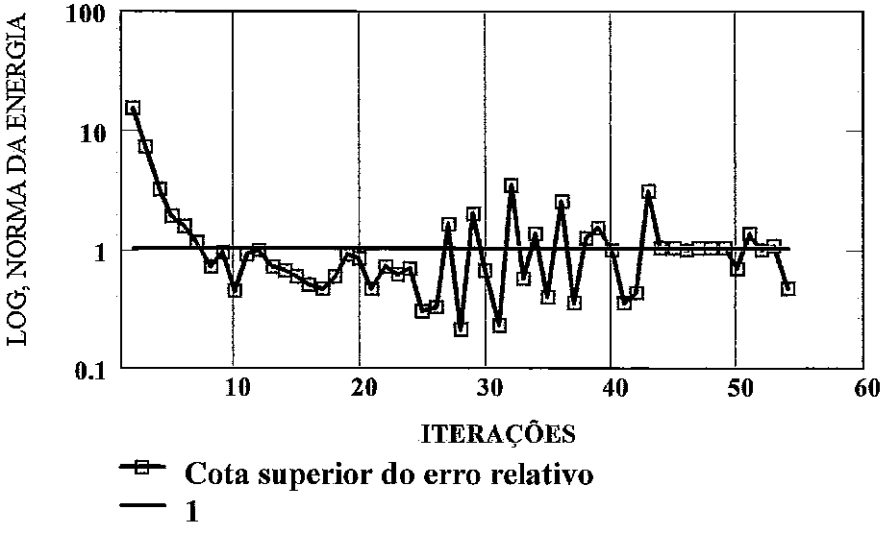


Figura 6.19: Estimativa com a sequência $\left\{\overset{\bullet}{\alpha}_i\right\}$, $n = 33$.

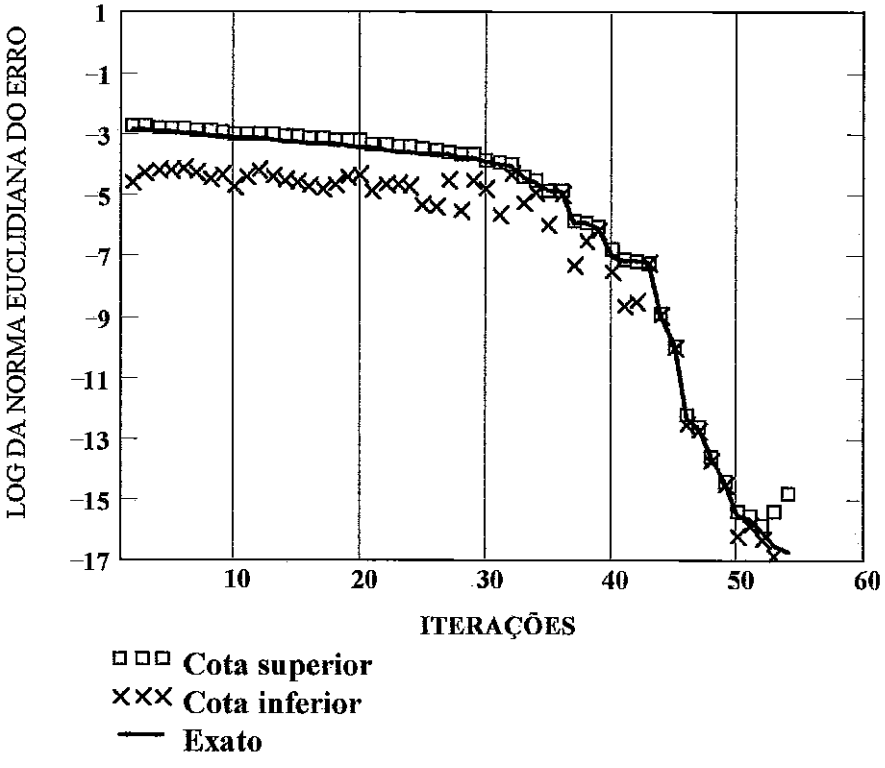


Figura 6.20: Estimativa com a sequência $\{\alpha_i\}$ exata, $n = 33$.

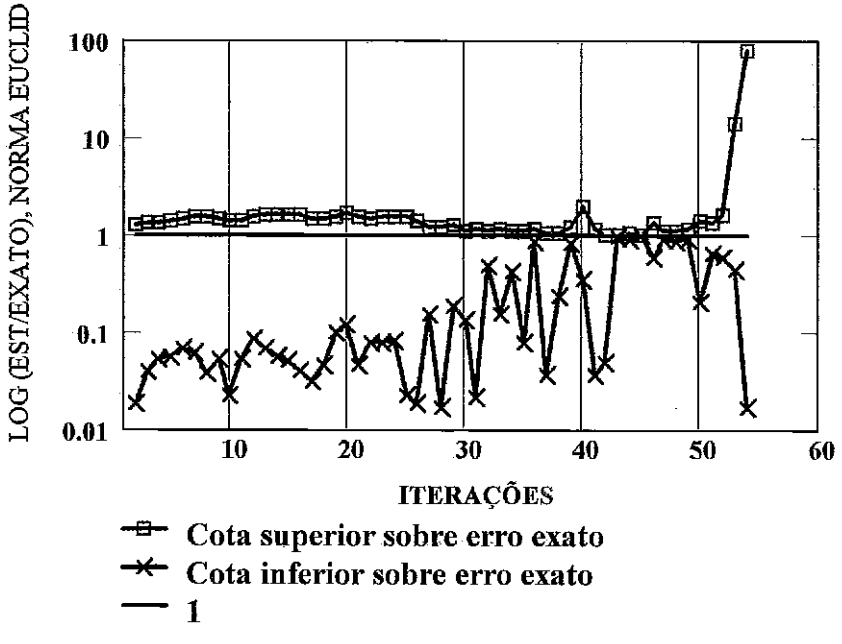


Figura 6.21: Estimativa com a sequência $\{\widetilde{\alpha}_i\}$, $n = 33$.

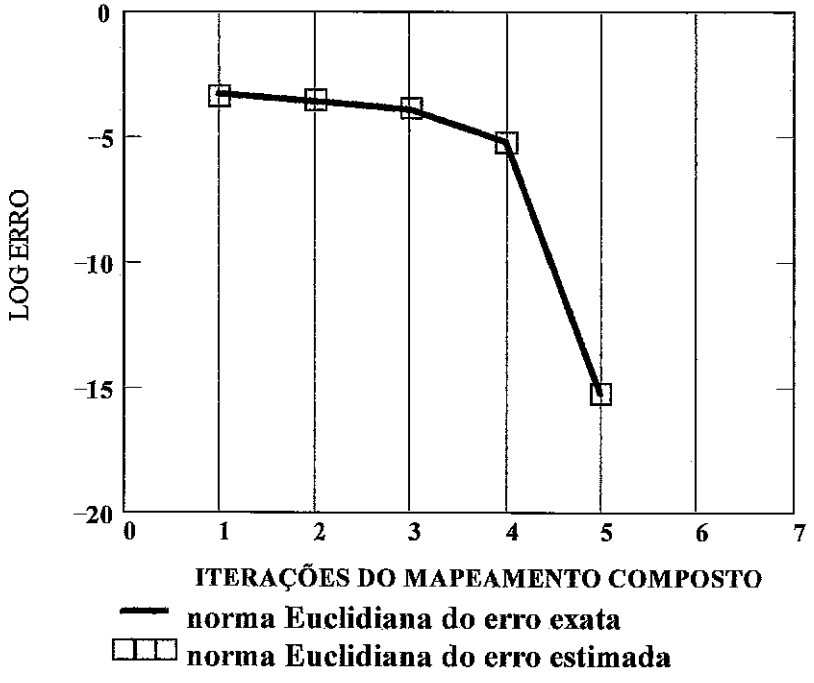


Figura 6.22: Estimativa da Norma-ℓ_2 do erro no mapeamento composto, $p = 2$, $n = 9$.

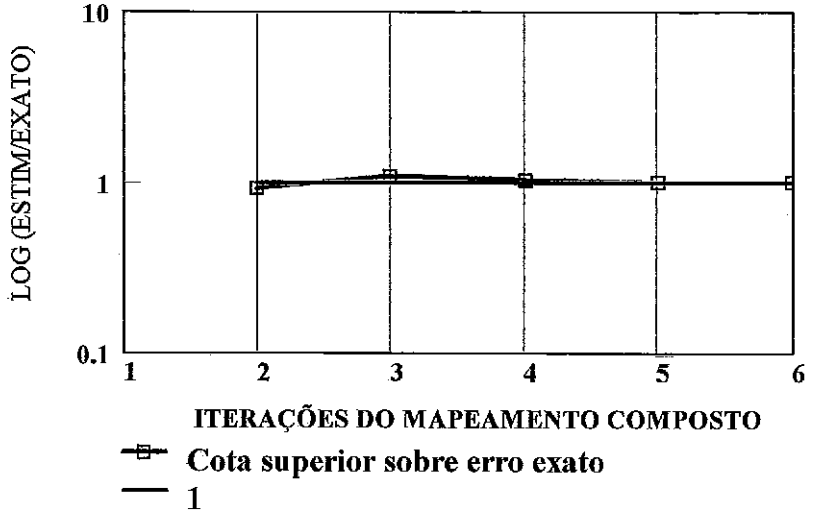


Figura 6.23: Estimativa da Norma-ℓ_2 do erro no mapeamento composto, $p = 2$, $n = 9$, dividida pelo erro exato.

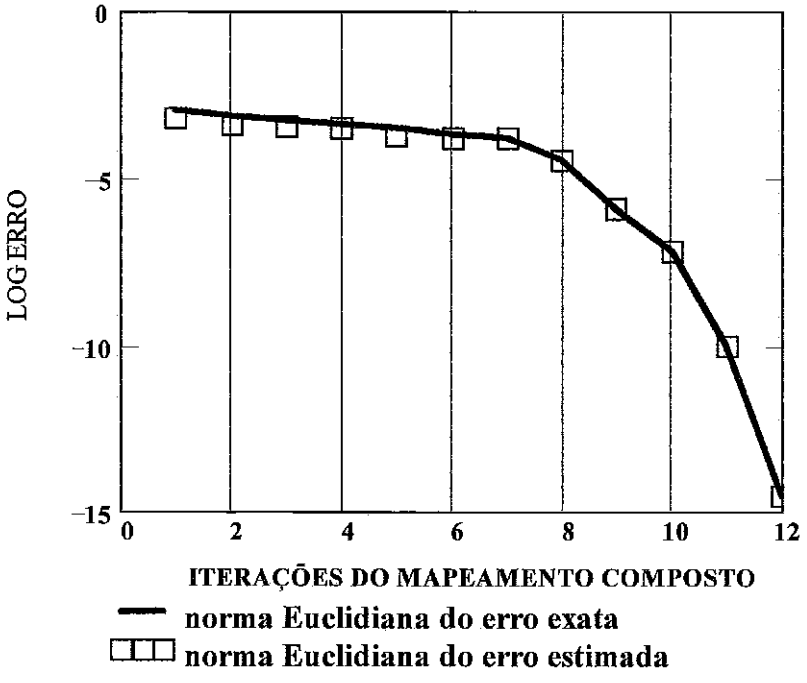


Figura 6.24: Estimativa da Norma-ℓ_2 do erro no mapeamento composto, $p = 4$, $n = 33$.

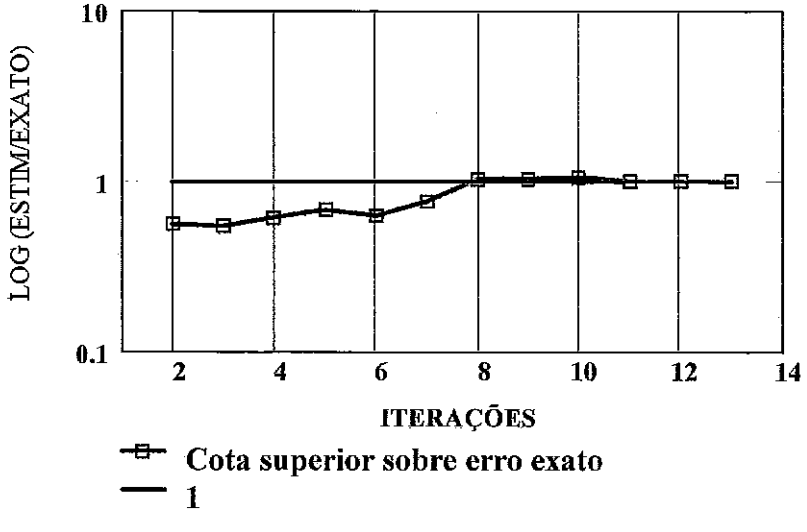


Figura 6.25: Estimativa da Norma-ℓ_2 do erro no mapeamento composto, $p = 2$, $n = 33$, dividida pelo erro exato.

CG e BROYDEN

O problema em discussão precisa um precondicionador, simétrico e positivo definido. Dentro da estratégia proposta há a necessidade de desenvolver um algoritmo padrão para trabalhar com o CG nos sistemas Hermitianos. Não foi feita esta investigação. Se a escolha é de um método secante, dois candidatos imediatos são o SR1 e o BFGS. O SR1, sendo uma atualização simétrica de posto um, é o candidato ideal, se o objetivo é economia de memória, porém não há garantias da heriditariedade da positividade do precondicionador inicial. O BFGS é hereditariamente simétrico e positivo definido, porém é uma atualização de posto dois, e requer pelo menos duas vezes mais memória. Existe uma literatura recente sobre o precondicionamento de sistemas não lineares com o BFGS [57][58].

Porém, o seguinte problema

$$\min_{H\in\mathbb{R}^n} \|B-H\|_F \ , \quad B \text{ simétrica},$$

tem solução única dado por $B=\frac{1}{2}(H^T+H)$, a parte simétrica de H. Isto justifica usar a parte simétrica da sequência $\{H_i\}$ de Broyden para hibridizar com o CG. Não há nenhum motivo, que não o de economia de memória, para não usar o Broyden em problemas Hermitianos positivo definidos. Não se está recomendando esta prática, em geral. Porém, na ausência de um melhor estudo sobre o SR1, ou outro mais conveniente, investiga-se agora o comportamento do Broyden precondicionando o CG. Mostrar-se-á que a teoria desenvolvida é absolutamente consistente, e alguns resultados são bastante além das espectativas.

Para evitar problemas de interpretação das iterações do CG, usa-se $n=27$, garantindo um problema bem posto, e $H_0=I$, a matriz identidade, já que a idéia é gerar ou melhorar um precondicionador, em particular para uso posterior (múltiplos carregamentos, análise dinâmica, etc). O comportamento do CG na norma-ℓ_2 do erro e do resíduo está mostrado na figura 6.26. A precisão $\sqrt{macheps}$ apenas é conseguida com 32 iterações sobre o erro e 37 sobre o resíduo. Observe-se que, em todas as figuras mostradas nesta discussão, $\|x_i\|_2$ rapidamente se aproxima de seu valor final. A norma do resíduo oscila tanto para o CG quanto para o Broyden sem pesquisa na linha.

Executa-se o Broyden tradicional, sem pesquisa na linha, durante 15 iterações. Em seguida, o CG é ativado no *loop* interno com o precondicionador H_{15} simetrizado, e com a solução reinicializada para $x_0=0$. Caracterizando assim uma pura construção de precondicionador. A figura 6.27 é a história das iterações do CG. Não

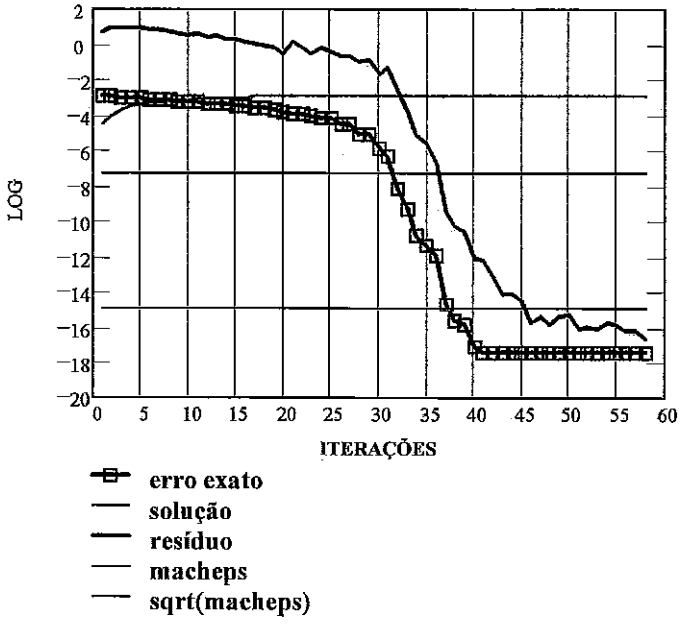


Figura 6.26: CG, norma-ℓ_2 , $n = 27$.

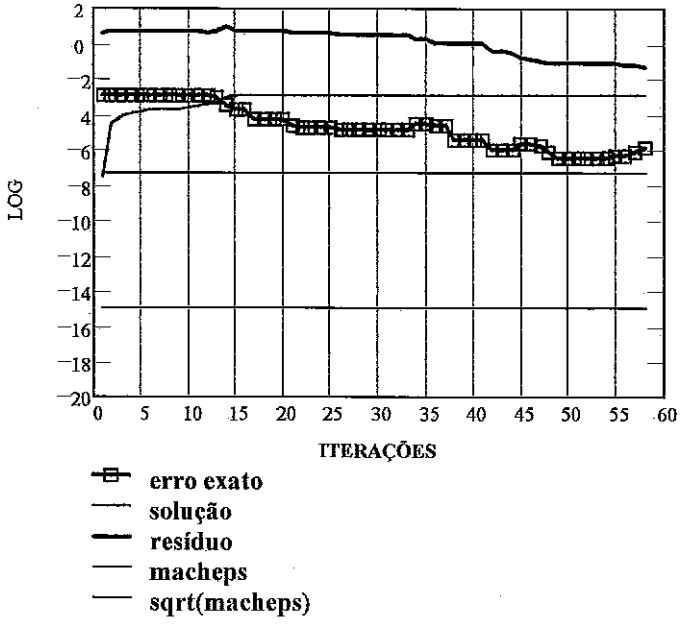


Figura 6.27: GB(15)+CG, norma-ℓ_2 , $n = 27$, sem ortogonalização.

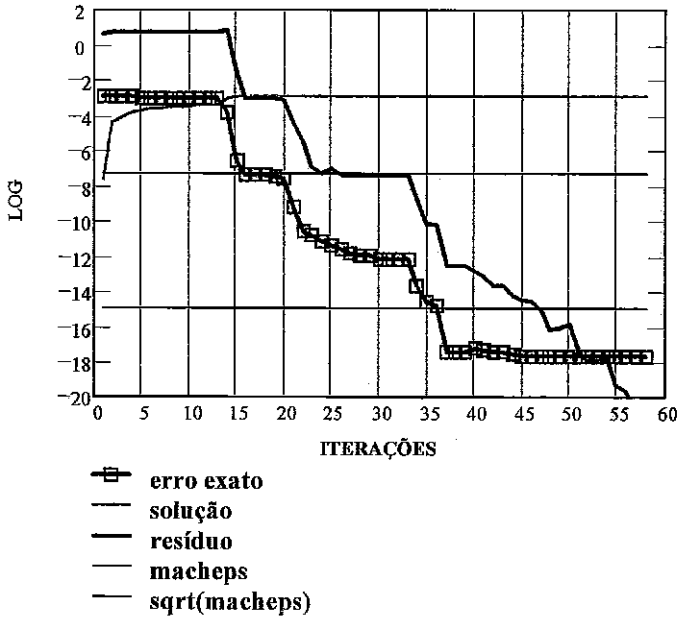


Figura 6.28: GB(15)+CG, norma-ℓ_2 , $n = 27$, epi($\frac{1}{2}\|e\|_2^2$).

há convergência em pelo menos 58 iterações. Como o CG sozinho convergiu, a responsabilidade do fracasso foi do precondicionador.

Todo o processo é repetido, agora com o Broyden ortogonalizando a distância epigráfica de $\phi = \frac{1}{2}\|e\|_2^2$, novamente por 15 iterações, e levando H_{15} e $x_0 = 0$ para o CG. A história do CG está na figura 6.28, que deve ser comparada com a figura 6.26. O GB(15)+CG atinge $\sqrt{macheps}$ na norma-ℓ_2 do erro em 16 iterações, e *macheps* em 36 iterações. Uma sensível melhora em relação ao CG puro. Mostrando o poder da minimização da distância epigráfica do erro. Observe-se que o resíduo ainda está bem acima, mais muito melhor do que no CG puro. Outras ortogonalizações foram tentadas, 6 no total, inclusive a minimizacão do próprio resíduo. Em todas elas, ambas, a norma-ℓ_2 do erro e a do resíduo, foram sempre melhores na minimização da distância epigráfica do erro na norma-ℓ_2. Deve ser observado também o não crescimento do resíduo, em nenhum instante, senão depois de atingida a precisão da máquina. O que não acontece com o CG puro. A figura 6.29 mostra o GB(25)+CG. Uma precisão na norma-ℓ_2 do erro é agora atingida, com bastante folga, em 5 iterações.

A figura 6.30 registra o espectro dos autovalores da matriz de rigidez K, do produto $H_{15}K$, e os autovalores de $H_{15}K$. Os efeitos do Broyden epigráfico sobre

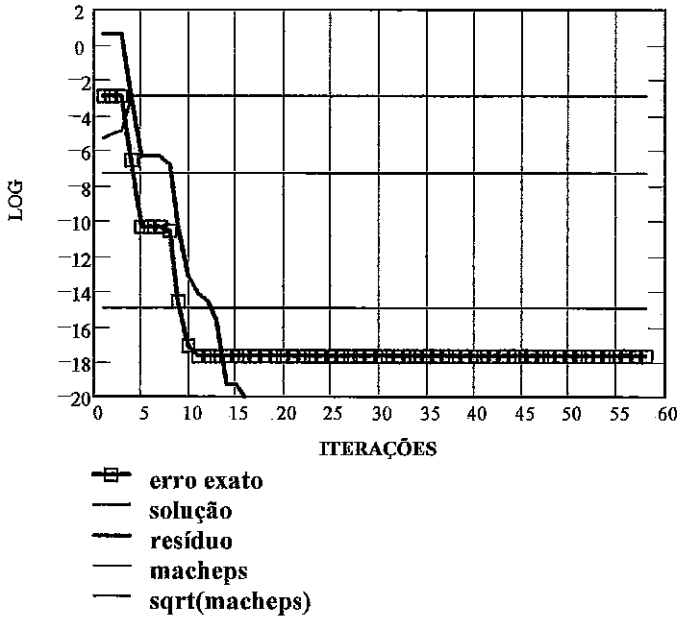


Figura 6.29: GB(25)+CG, norma-ℓ_2 , $n = 27$, epi$(\frac{1}{2}\|e\|_2^2)$.

a matriz de rigidez é inesperado. Existem 14 autovalores do produto $H_{15}K$ - um único agrupamento - com o valor 1. O máximo autovalor reduziu-se em cerca de uma ordem de grandeza. É bem curioso que também exista um autovalor negativo igual a -1.154×10^4. O surpreendente é que em nada atrapalhou a convergência do CG. A decomposição $L^T L$ do produto $\widehat{H} = (H_{15} + \epsilon I)$ só é possível para $\epsilon \geq 0.59$. Levando este precondicionador $\widehat{H}$ ao CG, ele apenas começa a reduzir o erro e o resíduo em torno de $i > 25$, mostrando que a manutenção do autovalor negativo é essencial para a convergência do processo. O autor não verificou quais autovalores possuiam parte complexa, e mostra-se bastante surpreso, não tanto pelo autovalor com parte real negativa, menos ainda pelo sucesso do precondicionamento, mas por ele, o autovalor com parte real negativa, ter-se mostrado essencial para a eficiência do precondicionamento do CG, implementado como real, e não como complexo (Hermitiano).

6.3 Experimento 2 - Matrizes Não Normais

Matrizes não normais são aquelas para as quais $A^H A \neq AA^H$. Uma matriz é normal se, e somente se, é defectiva. Uma matriz triangular é normal se, e somente se,

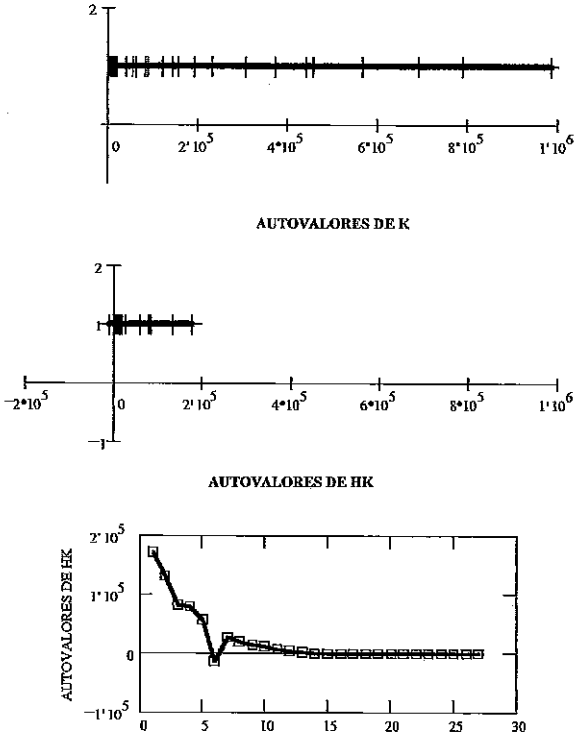


Figura 6.30: Espectro de autovalores de K e HK

é diagonal. O afastamento da normalidade tem um papel importante em cálculos numéricos porque, de alguma forma ainda não completamente entendida, a não normalidade está estreitamente relacionada com a instabilidade espectral. Por um longo tempo, os matemáticos aplicados consideraram as consequências da não normalidade uma anomalia matemática. Hoje, cada vez mais matrizes surgem que se afastam de forma ilimitada da normalidade, provenientes de modelagens de problemas físicos no limear da instabilidade [22],[71]. Este tipo de ocorrência tem surgido em mecânica dos fluidos [77]. Em mecânica das estruturas há esta possibilidade quando as cargas são dependentes da geometria, induzindo à não simetria do Jacobiano. Esse problemas se constituem em um desafio para códigos numéricos devido a uma segunda consequência da não normalidade: uma possível deteriorização da estabilidade dos algoritmos [74].

Segue-se um exemplo retirado da análise de uma carga móvel passando sobre uma viga rígida, bi-apoiada sobre molas, usando uma formulação mista. A intenção é mostrar como as estimativas de erro propostas se comportam frente a tais problemas. Em um certo instante da evolução no tempo o seguinte sistema triangular (aqui normalizado para obter resposta inteira) é obtido

$$\begin{bmatrix} 0.01 & 4.00 \\ 0 & 0.01 \end{bmatrix} \begin{Bmatrix} 1 \\ 1 \end{Bmatrix} = \begin{Bmatrix} 4.01 \\ 0.01 \end{Bmatrix}.$$

A solução deste sistema pelo método de Gauss perde dois dígitos significativos. Para cada equação acrescentada ($A_{ii} = 0.01$, $A_{ij} = 4.0$, $i < j$), o método de Gauss perde mais 2 dígitos significativos. Como o sistema é triangular, esta perda diz respeito apenas à fase de retrosubstituição. Resolve-se o problema com o método de Richardson na forma mais básica, qual seja, $x_{i+1} = (I - A)x_i + b$, e tenta-se estimar a norma-ℓ_2 do erro ao longo das interações. (Este problema pode ser resolvido em duas iterações por qualquer método básico, como Jacobi, por exemplo). Teoricamente, o método de Richardson deve convergir sem problemas, mas a não normalidade o impede.

O primeiro a ser notado é a curiosa separação resíduo-erro em torno de $i = 100$, figura 6.31. O resíduo decresce, o erro cresce; o mínimo de um, o máximo do outro. Esta separação é tão acentuada quanto maior o número de equações. Impede o uso do critério de parada baseado na norma-ℓ_2 do resíduo. O critério (*backward error*) $\|e\|_2 \leq \|A^{-1}\|_2 \|f\|_2$ é de fato uma cota superior do erro em toda a história da realização, e não se deixa enganar pela separação erro-resíduo, figura 6.32. O problema é que não se conhece $\|A^{-1}\|_2$ (no caso geral).

Na figura 6.33, as melhores estimativas (ponto a ponto, sem mapeamento composto) de erro desta tese também são derrotadas na sua tentativa de previsão de erro em torno de $i = 100$. Daí em diante não se pode desejar nada melhor, até que predomine a precisão da máquina. Porém, a Proposição 49,3.32, (e suas congêneres), resolvem a questão. A cada instante controla-se os índices (basta um)

$$conf1_i = 1 - \{\beta_i \delta_i\} \quad \text{e} \quad conf2_i = 1 - \frac{\|\beta_i - \xi_i\|}{\beta_i + \xi_i}$$

determinados com os *erros estimados*. A confiabilidade das estimativas de erro dependem do quão próximo de 1 estão estes índices. Eles são uma salvaguarda da teoria quando aplicada a algoritmos não q-convergentes. A figura 6.34 mostra o $conf1$. Não há necessidade de temer os estragos causados pela não normalidade. Normalmente, o autor usa o produto: $conf = conf1 \times conf2$ como critério para aferir a qualidade dos estimadores em qualquer problema, e para apenas quando ele é crescente e com um valor alto em relação aos anteriores (da realização).

A figura 6.35 mostra os valores de $x_i = ((x_i)_1, (x_i)_2)$, as duas componentes da solução aproximada ao longo das iterações, bem como a sua trajetória em $\mathbb{R}^2$, a partir das quais a separação erro-resíduo pode ser melhor compreendida.

6.4 Experimento 3 - O Método de Newton

Dentro de condições pouco severas, o método de Newton converge q-superlinearmente. Porém não se conhece quais as suas propriedades minimizantes quando implementado sem pesquisa na linha. Este experimento mostra que, de fato, não há qualquer restrição, ou modificação, à teoria apresentada quando de sua aplicação em problemas não lineares, e que é possível estimar a norma-ℓ_2 (de fato, qualquer norma) do erro em problemas não lineares com absoluta confiança, e não apenas a usual verificação da consistência. Em problemas de análise estrutural, normalmente é a norma-ℓ_2 do erro a que mais interessa. O problema teste está bem descritro em [75].

Seja $\Omega \subset \mathbb{R}^n$ um conjunto compacto, e $X = C(\Omega, \mathbb{R}^n)$ o espaço de funções contínuas em Ω com valores em $\mathbb{R}^m$, sem perda de generalidade, assuma que Ω possui medida unitária. Seja $u(x) \in X$, e considere o problema de achar a solução $u_*(x) \in X$ da equação integral da forma

$$f(u)(x) = h(u(x)) - \int_\Omega k(x, y, u(x), u(y))dy - g(x) = 0 \ ,$$

onde $g \in X$, h é uma função continuamente diferenciável de $\mathbb{R}^n$ nele próprio, e k é uma função contínua de $\Omega \times \Omega \times \mathbb{R}^m \times \mathbb{R}^m$ para $\mathbb{R}^m$. Suponha que sejam válidas

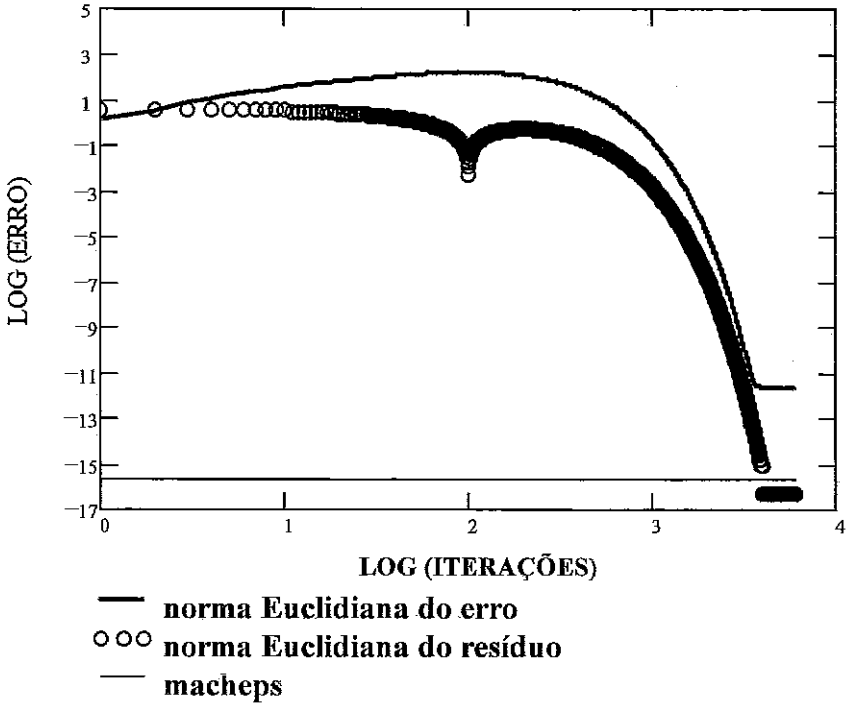


Figura 6.31: Normas-ℓ_2 do erro e do resíduo na matriz não normal, ao longo de 4500 iterações, $n = 2$.

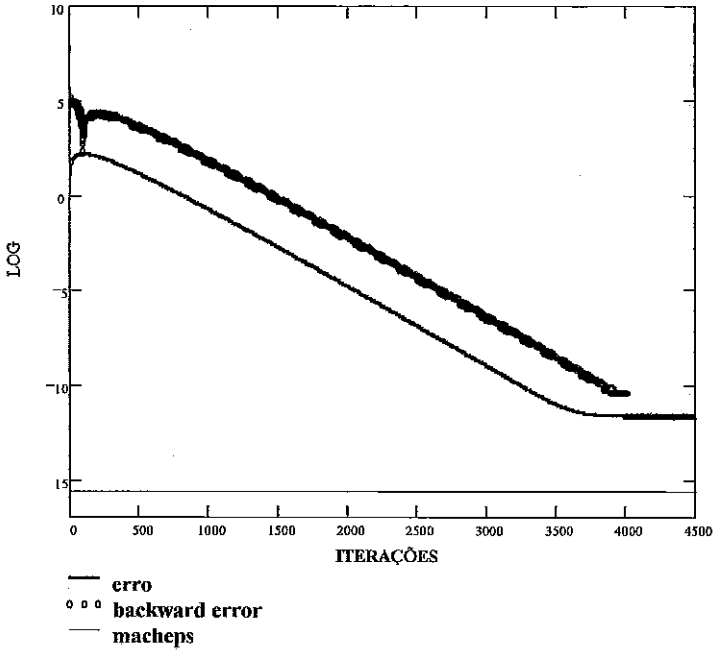


Figura 6.32: Previsão do *forward error* pelo *backward error*

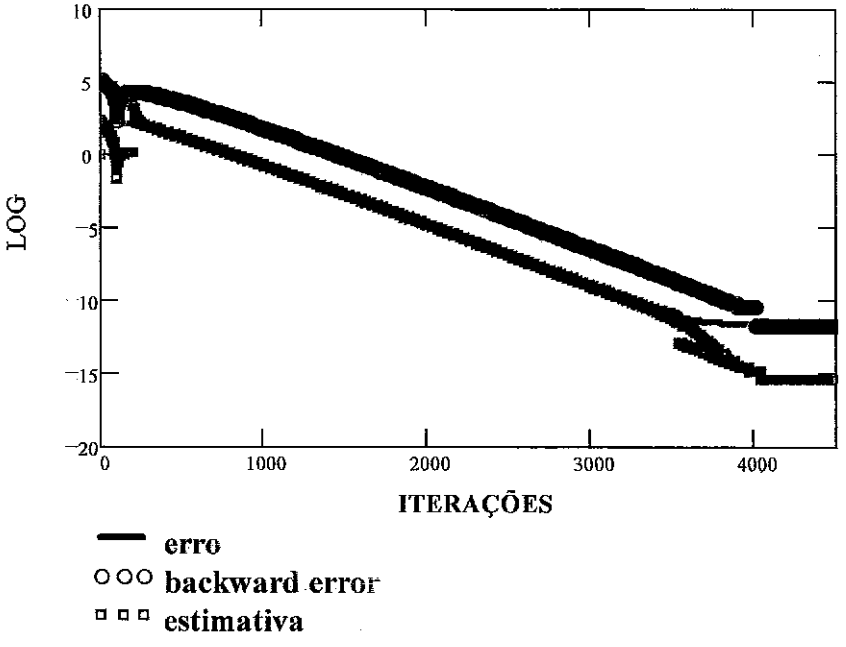


Figura 6.33: Comparação entre a estimativa desta tese e a *backward analysis*

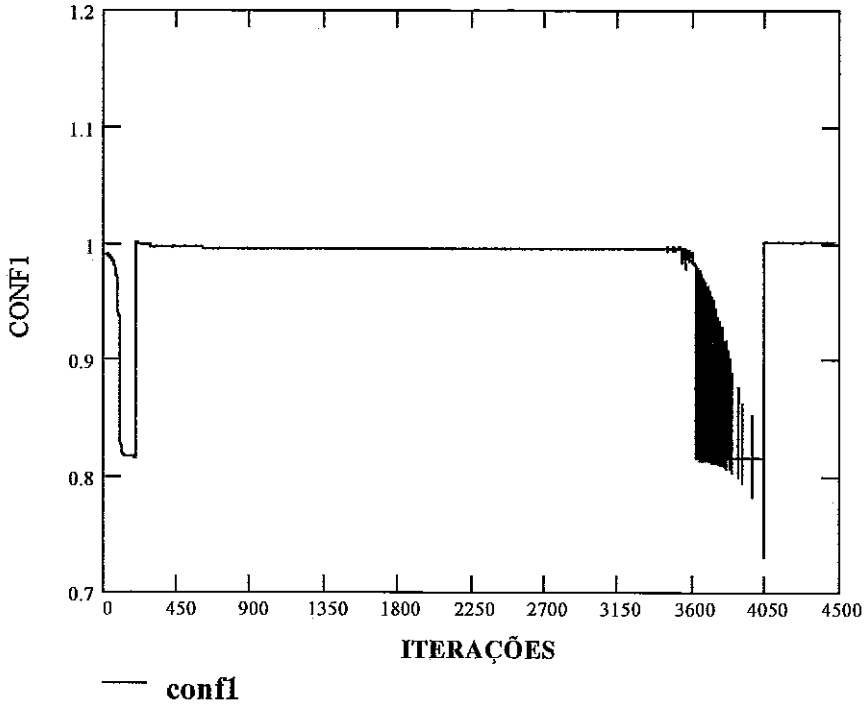


Figura 6.34: Índice de confiabilidade das medidas, $conf1$

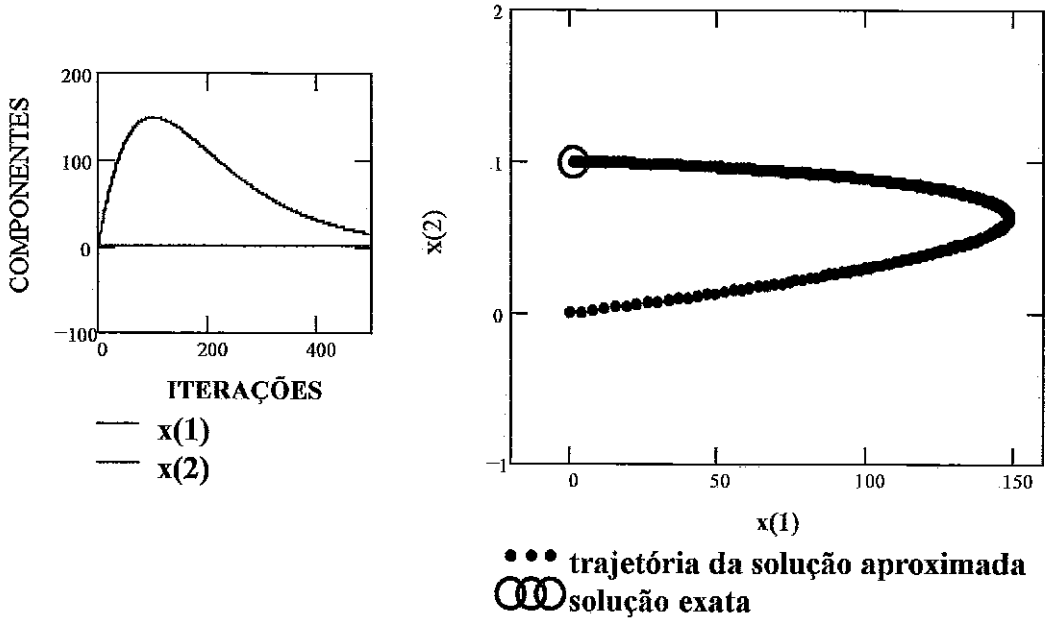


Figura 6.35: Componentes e trajetória da solução em R^2

as hipóteses padrões usuais pelas quais o Jacobiano de h, e as matrizes $m \times m$ das derivadas de k com respeito a $u(x)$ e $u(y)$ são contínuas segundo Lipischitz com respeito aos seus dois últimos argumentos. Sob essas hipóteses pode-se determinar a derivada de Fréchet de f em qualquer ponto $u \in X$, ou seja,

$$f'(x)v(x) = \left(h'(u) - \int_\Omega k_{u(x)}(x, y, u(x), u(y)) \right) v(x)dy - \int_\Omega k_{u(y)}(x, y, u(x), u(y))v(y)dy \ .$$

Onde $k_{u(x)}$ e $k_{u(y)}$ são as derivadas de k com respeito a $u(x)$ e $u(y)$, respectivamente. $k_{u(x)}$ e $k_{u(y)}$ são matrizes (funções) $m \times m$.

Com as condições estipuladas acima, se o Jacobiano de $f(u)$ na solução é um mapeamento (aplicação) não singular sobre X, o método de Newton pode ser aplicado e converge q-superlinearmente perto da solução (de fato converge com ordem-q de convergência no mínimo igual a 2, mas, para a teoria presente, isto é irrelevante já que implica em convergência assintótica q-superlinear).

O critério de parada usual é com a norma-ℓ_2 do resíduo, ou seja, com $\|f_i\|_2$ [75, pag. 1139],[5]. É possível encontrar a constante de Lipschitz e o fator-q analiticamente. Porém, isto é desnecessário. Usar-se-á os valores estimados durante a realização. As estimativas gerais são as mesmas para espaços compactos gerais e para espaços de Hilbert com algoritmos não ortogonais. Leva-se apenas em conta a q-superlinearidade através do fator ψ_i em cada iteração, para obter melhores estimativas. O problema teste tem as seguintes constantes e funções, pelas quais é possível encontrar a resposta exata analiticamente:

$$\begin{aligned} m &= n = 1 \\ \Omega &= [0, 1] \\ h(u) &= cu^2, \quad c = 1 \\ k(x, y, w, u) &= \cos(yu)/2 \\ g(x) &= (1 - \sin(1))/2 \\ u_\star(x) &= 1 \ . \end{aligned}$$

Os valores iniciais são $u_0(x) = u_\star(x) + 0.7\cos(9\pi x)$. São usados 20 pontos de Gauss para a integração em Ω.

Na figura 6.36 estão os valores exatos da norma-ℓ_2 do erro bem como as estimativas usando a sequência $\{\widetilde{\alpha}_i\}$, em dois gráficos, um em valor absoluto e o outro em valores relativos ao erro exato em cada iteração. Observe-se que, mesmo com o valor $c = 1$ - problema bem condicionado - a norma-ℓ_2 do resíduo superstima a norma-ℓ_2 do erro.

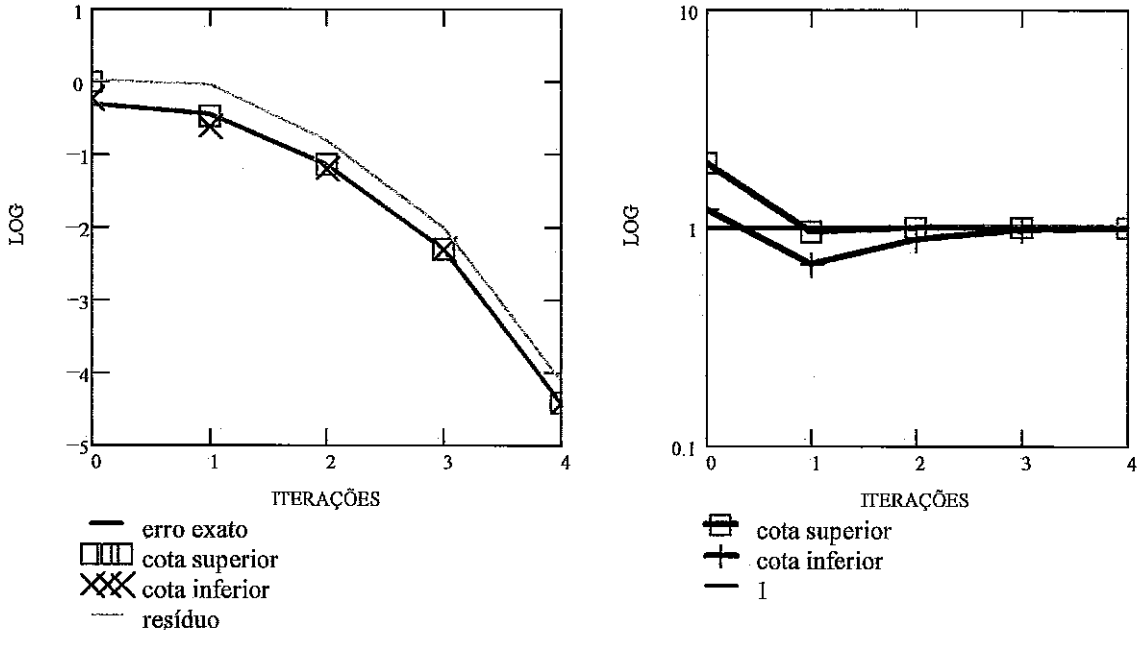


Figura 6.36: Método de Newton - erros absolutos: exato, estimado e resíduo. E erros estimados relativos aos exatos.

6.5 Experimento 4 - GMRES, CGS e GB

Introdução

Este experimento trata de estimativas e estratégias para a solução de problemas lineares no método dos elementos de contorno - MEC. A experiência acumulada na solução desses sistemas é muito grande. Em Araújo[67] e Mansur, Araújo e Malaghini [76] alguns problemas foram solucionados com algoritmos e precondicionadores diversos, pelo qual se concluiu que, dentre os diversos algoritmos testados, o clássico BiCG comportava-se melhor, a despeito de sua irregularidade na norma dos resíduos. Barra *et al.* [68], [69] usou com sucesso o GMRES e o GMRES(m), precondicionados ou não, mostrando que eles conseguem uma mesma precisão na norma-ℓ_2 dos resíduos do que o BiCG com ambos, menor tempo de cpu e menos interações. Comparações entre o precondicionador de Jacobi à esquerda e à direita no GMRES, feitas por Barra et al. [42] mostram que usando um escalonamento físico, a muito tempo um padrão mesmo com métodos de decomposição [70], o precondicionador à esquerda comporta-se um pouco melhor.

Quando o problema não apresenta singularidades na solução o GMRES, como

muitos algoritmos, possuem uma certa independência da malha, o que também é observado em equações integrais resolvidas por outros métodos que não o MEC, mesmo em espaços de Banach [4]. Em problemas tri-dimensionais (elasticidade linear com o MEC) sem singularidades [35], mesmo sem o escalonamento físico, e apenas com o precondicionador de Jacobi, esta independência existe. Pouquíssimas iterações são necessárias, não havendo, em geral necessidade de reinicializações.

Se, por outro lado, as singularidades estão presentes, a experiência acumulada [68] diz que um número de vetores auxiliares m não maior do que dez por cento de n - o numero de equações - é em geral satisfatório para precisões usuais. Barra *et al.* [69] considera com sucesso métodos multi-níveis na solução de tais sistemas. O principal objetivo é reduzir a memória auxiliar, e procurar manter um número aceitável de iterações e tempo de cpu. A experiência do autor diz que quando o operador discreto converge para o seu ponto fixo, o número de iterações permanece constante e independente do número de graus de liberdade do sistema.

A grande vantagem dos métodos de curta recorrência como o BiCG e o CGS e suas variantes (QMR, TFQMR) está justamente em precisar de memória auxiliar reduzida e fixa. Para métodos como o GMRES o crescimento da memória auxiliar necessária é proporcional ao número de iterações, e, a não ser por experiência em problemas específicos, não se pode prever com antecipação a memória necessária, mesmo a mínima para garantir a convergência do algoritmo. Isso pode ocasionar a necessidade de reinicialização estando o método já muito perto da solução, quando não a impossibilidade de solução. Um inconveniente enorme para o GMRES(m). Ele é interrompido quando está na fase q-linear ou q-superlinear, e sua reinicialização o leva a uma nova e duradoura fase q-sublinear, com o resíduo sendo reduzido marginalmente [34], se não à estagnação. A qualidade do precondicionador tem um papel marcante na minimização desses problemas. Nos exemplos que se seguem são usados o escalonamento físico e o precondicionador de Jacobi à esquerda. Os resultados referem-se ao sistema já escalonado, e ainda não precondicionado por Jacobi.

Escolhe-se um problema plano típico de mecânica da fratura, já analisado várias vezes na literatura citada. A figura 6.37 contém os dados físicos e geométricos. É utilizado o recurso de simetria permitida pelo MEC. Para os elementos situados nos cantos utiliza-se a técnica do nó duplo. Perto da região da singularidade, a raiz da fratura, a malha é mais refinada. Foram selecionados o GMRES(m) e o CGS para estimativas de erros.

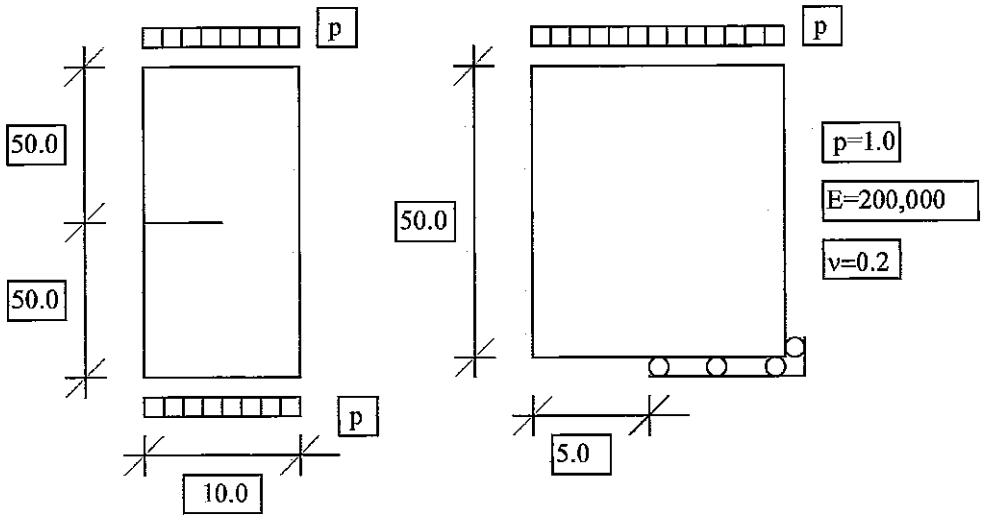


Figura 6.37: Problema de Mecânica da Fratura - Geometria e Propriedades.

Estimativas para o GMRES

O sistema escolhido possue 1196 equações que correspondem a 1.43×10^6 palavras de precisão dupla ($real * 8$) para armazenar a matriz dos coeficientes ($11MB$). As comparações são feitas na norma-ℓ_2 do erro. Nas figuras 6.38 e 6.39 estão os erros relativos estimados (erro estimado/solução corrente) e os exatos (erro exato/solução exata) para o GMRES(30), com 12 ciclos (de 30 iterações) A solução exata foi tomada como a solução por decomposição de Gauss, pivoteada. As soluções finais entre o GMRES(30) e Gauss diferem a partir do décimo terceiro dígito significativo. Os gráficos contêm também o índice $conf = conf1 \times conf2$, definido anteriormente. A degradação das estimativas ocorrem para baixos valores de $conf$ apenas no final da realização, quando predominam os erros de aritmética finita, e o GMRES(30) já não consegue melhorar a qualidade da solução. A qualidade dos resultados para os erros absolutos é completamente similar. As estimativas para o GMRES(45) e GMRES(60) são ainda melhores, como esperado.

Estimativas para o CGS

O CGS é um algoritmo de métrica variável com recorrência curta. Ele perfaz dois produtos matriz-vetor por iteração e produz uma norma-ℓ_2 do resíduo bastante irregular. A despeito desta irregularidade é bastante usado como *inner-loop* de algoritmos tipo-Newton, como por exemplo o método de Newton inexato (ou trun-

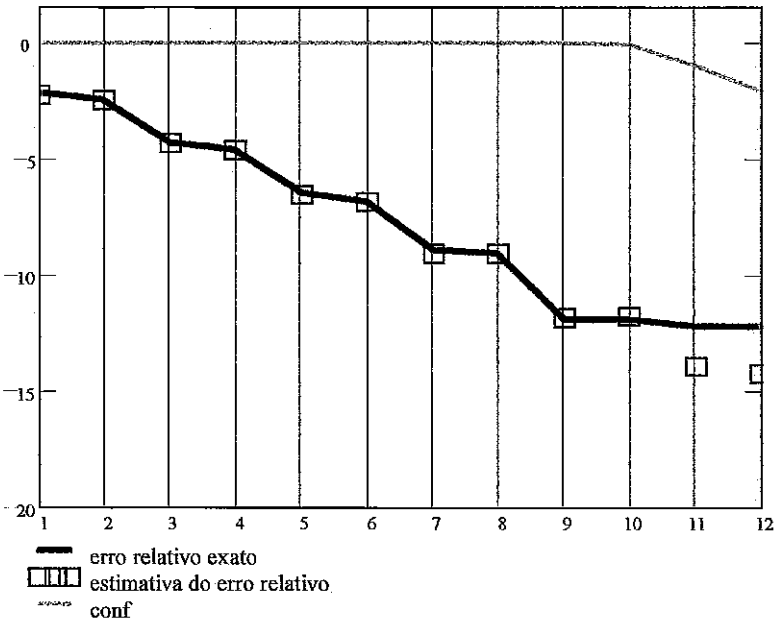


Figura 6.38: GMRES(30) - estimativas do erro relativo. Norma-ℓ_2 do erro. Usando a sequência $\{\widetilde{\alpha}_i\}$.

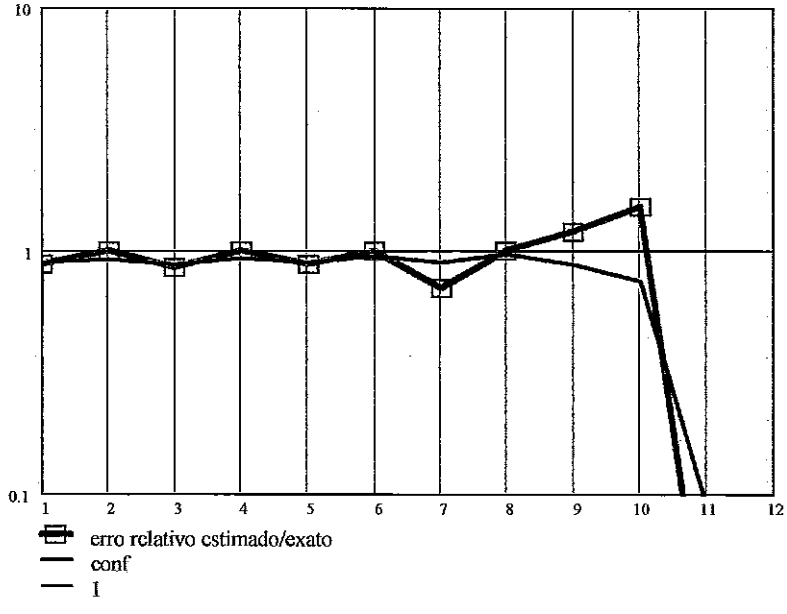


Figura 6.39: GMRES(30) - estimativas do erro relativo (normalizado). Norma-ℓ_2 do erro. Usando a sequência $\{\widetilde{\alpha}_i\}$.

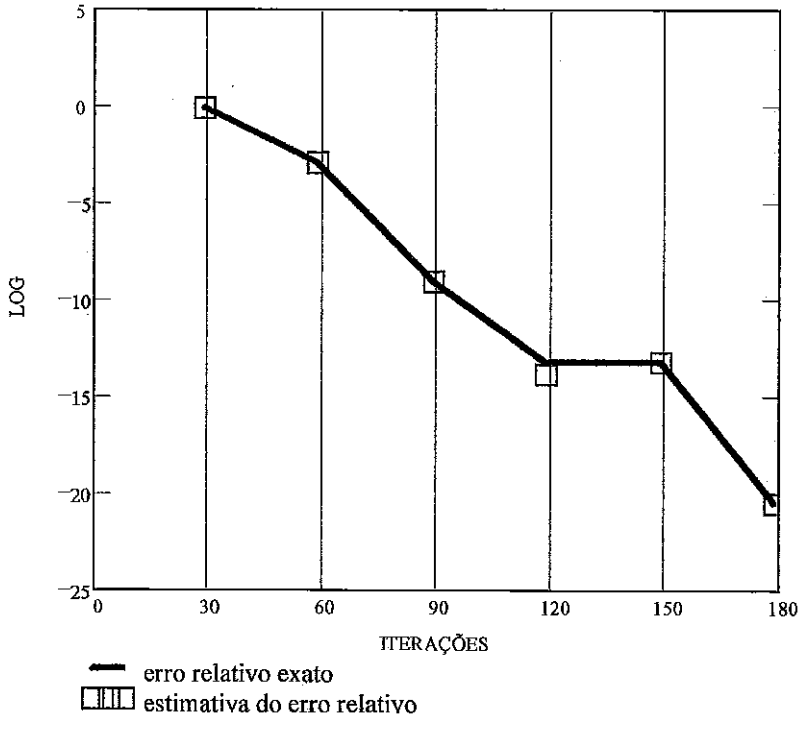


Figura 6.40: Estimativas do erro relativo para o CGS de Schonauer.

cado, em otimização). A razão disso é que consegue fazer a norma-ℓ_2 do erro decrescer consideravelmente, embora também de forma irregular, enquanto mantém memória auxiliar pequena e independente do número de iterações. Converge duas vezes mais rápido do que o BiCG. Schonauer [59] foi o primeiro a ter a idéia de fazer uma minimização da norma-ℓ_2 dos resíduos, *a posteriori,* sobre a sequência de soluções do CGS. Esta minimização, por não interferir no algoritmo, não melhora a sua velocidade, mas apenas elimina as oscilações, produzindo longos patamares de q-sublinearidades, tanto da norma-ℓ_2 dos resíduos como da norma-ℓ_2 do erro. Posteriormente, outros algoritmos foram criados, matematicamente equivalentes ao CGS de Schonauer, e sem melhores características de estabilidade (TFQMR entre eles). A estabilidade é um grande problema do CGS. Para algoritmos tão irregulares a melhor opção é usar um mapeamento composto para estimar o erro. A figura 6.40 apresenta as estimativas da norma-ℓ_2 do erro para $\mathcal{A}^p$ com $p = 30$. Entre $p = 5$ a $p = 30$ a qualidade das estimativas se equivalem. Para outras possibilidades, ver [34].

GMRES e GB

Há uma grande conveniência em fazer o GMRES(m) trabalhar em conjunto com o GB(m). Sendo o primeiro um algoritmo de métrica fixa, com todas as vantagens de minimização da norma-ℓ_2 dos resíduos, possui a desvantagem de necessitar de longas recorrências, principalmente quando há singularidades na solução. O algoritmo GB(m) é um método de métrica variável, porém com a possibilidade não encontrada em algoritmos como o BiCG, o CGS e variantes diversas os quais não permitem uma natural extensão das bases para tratar aqueles problemas. Compondo as bases de ambos os métodos de uma maneira consistente permite não apenas aumentar a velocidade de convergência, como aumentar a robustez. Adicionalmente, pode-se construir as bases do GB(m) minimizando alguma função epígrafica de interesse.

Entretanto, as possibilidades de compor um polialgoritmo são muitas. No presente estágio, o autor deste trabalho procura desenvolver mecanismos consistentes que justifiquem uma melhor estratégia, usando conceitos oriundos da solução de problemas de viabilidade convexa, a ser relatada em outro trabalho. No que se segue exemplifica-se uma possibilidade imediata com suas variantes.

A primeira possibilidade é, em semelhança ao discutido no Experimento 1, usar o GB(m) como atualizador de precondicionador para o GMRES(m). Esse tipo de polialgoritmo será mais eficiente quando o GM(m) servir de apoio a métodos de curta recorrência, em particular quando o precondicionador inicial for pobre, justificando assim o uso de valores de m mais altos. Será exemplificado na próxima subseção, quando da hibridização com o CGS.

Uma das mais interessantes consiste em usar o GB(m) como um método de memória variável, recuperando as bases apenas das últimas m iterações. Considere-se o exemplo anterior de mecânica da fratura. As duas tabelas seguintes compara o GMRES(m) com o GB(10)+GMRES(m). A primeira para 52 equações, e a segunda com 1196 equações. Foi fixado o tamanho da base do GB *a priori* para facilitar comparações. A estratégia consiste em, antes de cada ciclo de 1 a c do GMRES(m), efetuar *uma* atualização do GB(10), que desta forma, completará suas bases no início do décimo ciclo. No início do décimo primeiro ciclo, tendo-se esgotado a capacidade do GB(10), estipulada antecipadamente, despreza-se o primeiro elemento da base, como em uma pilha, e efetua-se a décima primeira atualização (trocando ponteiros), e prosseguindo desta forma. O número de atualizações do GB(10) ao longo da realização é representado por k. O índice i representa o número total de iterações feitas pelo GMRES(m) em toda a realização. A quantidade de produtos

matriz-vetor será portanto igual a $i+k$. Note que o GMRES(m) reinicializa (perde todas as bases) cada vez que completa um ciclo com m iterações. O critério de parada é apenas sobre o resíduo, para poder comparar com o GMRES(m) puro. O critério usado foi de $\|f_i\| \leq \epsilon \|f_0\|$, onde $\epsilon = 10^{-10}$. A solução inicial sempre é a solução $x_0 = 0$. O tempo de cpu (computador sequencial) está normalizado pelo maior. O símbolo $\star$ significa divergência.

Tabela 6.1 - 52 equações

	GMRES(m)			GB(10)+GMRES(m)				
m	c	i	cpu	c	i	k	i+k	cpu
1	$\star$	$\star$	$\star$	$\star$	$\star$	$\star$	$\star$	$\star$
2	$\star$	$\star$	$\star$	44	87	44	131	0.52
3	$\star$	$\star$	$\star$	29	87	29	116	0.36
5	$\star$	$\star$	$\star$	18	90	18	108	0.29
10	45	450	1.00	9	84	9	93	0.26
15	12	170	0.38	6	71	6	77	0.22
20	5	99	0.24	4	65	4	69	0.17
25	3	57	0.17	2	49	2	51	0.14
30	2	52	0.12	2	48	2	50	0.14
35	1	35	0.12	1	34	1	35	0.12

Tabela 6.2 - 1196 equações

	GMRES(m)			GB(10)+GMRES(m)				
m	c	i	cpu	c	i	k	i+k	cpu
10	$\star$	$\star$	$\star$	$\star$	$\star$	$\star$	$\star$	$\star$
15	$\star$	$\star$	$\star$	11	165	11	176	0.41
25	18	439	1.00	7	162	7	169	0.38
35	5	168	0.38	4	139	4	143	0.32
45	3	96	0.22	2	89	2	91	0.21
55	3	92	0.21	2	83	2	91	0.19
60	2	70	0.16	2	66	2	68	0.16
65	1	63	0.15	1	62	1	63	0.14

montagem do sistema: $cpu = 0.08$

solução por Gauss: $cpu = 0.91$

É imediato que, com os dois algoritmos juntos, consegue-se mais robustez e uma economia substancial no tempo de processamento, tanto maior quantas forem as intervenções do GB. Torna-se vantajoso para tamanhos de base do GMRES(m) pequenos. A medida em que se permite aumentar o valor de m no GMRES, nessa estratégia, diminui o número de intervenções do GB, e assim a eficiência da hibridização.

CGS e GB

A melhor estratégia para unir o GB ao CGS é melhorando o precondicionador original para o CGS. Usando o mesmo exemplo anterior, com o mesmo critério de parada, a tabela seguinte foi construida.

Tabela 6.3 - 1196 equações

GB(m)+CGS

m	i	cpu
0	105	1
10	53	0.57
20	35	0.46
25	32	0.47

Na tabela 6.3, inicialmente o CGS atua sem o auxílio do GB. Em seguida 3 realizações são efetuadas com diferentes tamanhos de base para o GB(m). Como antes, o precondicionador inicial é o de Jacobi. Há um limite a partir do qual o aumento de m não influencia sensivelmente a eficiência do CGS. A relação entre o tempo de cpu mínimo da tabela 6.3 e o da tabela 6.2 é de 1.21. Entretanto há uma substancial economia na memória auxiliar (seis vezes menos). Indubitavelmente, esta estratégia com o CGS é superior.

Algumas perspectivas

É aparente, a partir dos experimentos mostrados e de outros não apresentados, que existe uma grande flexibilidade na construção de algoritmos com a estratégia proposta. Um fato marcante é que a hibridização sempre favorece o desempenho de alguma forma: ou tempo de cpu, ou diminuição da memória auxiliar, ou ambos, ou ainda aumentando a robustez.

Todos os experimentos realizados constituiram-se em situações críticas. No experimento 1 houve a ousadia de precondicionar o CG com matrizes indefinidas, a partir de precondicionadores não simétricos. Os experimentos com o problema da mecânica da fratura são, pelo tanto que o autor tenha conhecimento, os mais difíceis. Em experimentos clássicos como chapas com furos circulares tracionadas, os resultados são bem melhores. Os precondicionadores iniciais foram sempre matrizes diagonais, em detrimento de precondicionadores melhores.

Há a necessidade de sistematizar os experimentos objetivando estratégias ótimas para diferentes objetivos: tempo de cpu, memória, robustez.

Estratégias mais sofisticadas estão sendo testadas. Tanto em sistemas Hermitianos como em não Hermitianos. Uma etapa concomitante é a implementação em SMP(s) - *Symetric Multi Processors*. A teoria para hibridizações nesses sistemas parece decorrer naturalmente a partir dos problemas de viabilidade convexa. Neste caso, vários algoritmos, iguais ou distintos, seriam executados simultaneamente entre os processadores simétricos, e reunidos com um algoritmo de projeção em convexos fechados.

Capítulo 7

Conclusões e Futuros Trabalhos

Neste Capítulo são revistas as principais idéias e resultados que foram desenvolvidos nesta tese, e são sugeridas algumas possíveis direções para pesquisa posterior.

7.1 Conclusões

Na Parte I foi estudada a qualidade de convergência de algoritmos iterativos sob a hipótese de convergência do quociente da medida do erro - convergência-q. A motivação para o estudo da convergência-q esteve na intuição de que ela é a forma de convergência desejada para algoritmos iterativos de qualquer espécie. Essencialmente, foram estudadas a geometria das realizações q-convergentes e a geometria do processo q-linear. O objetivo do estudo da geometria do processo q-linear é a compreensão dos mecanismos que levam à obtenção deste importante tipo de convergência, na espectativa de construir algoritmos pelo menos global e localmente q-lineares e, idealmente, localmente q-superlineares. O estudo das realizações q-lineares visa a obtenção de estimativas confiáveis de erro de algoritmos iterativos em qualquer medida. O principal objeto de estudo são os algoritmos nos espaços de dimensões finitas.

No Capítulo 2 foi mostrado que existe uma propriedade de separação entre as diversas medidas de qualidade de convergência, pela qual pode-se concluir que não é suficiente estimar erros em uma determinada medida se o interesse é a quantificação da proximidade da solução em outra medida. Esta é uma das principais motivações para a estimativa de erro em medidas arbitrárias. Procurou-se evidenciar as deficiências de algumas quantificações de qualidade de convergência quanto aos seus potenciais para servir de ponto de partida para um estudo métrico da

proximidade da resposta e da velocidade de convergência de algoritmos.

Foram também revistos os principais resultados da teoria métrica dos pontos fixos, diretamente vinculados às estimativas de erro. Concluiu-se pela insuficiência de resultados construtivos, e principalmente, pela enorme restrição que as atrações oferecem quando aplicadas aos principais algoritmos que estão sendo usados nas arquiteturas computacionais atuais e, potencialmente, do futuro.

No Capítulo 3, com definições adequadas de realização e de processo q-linear, mostrou-se que a convergência-q comporta a geometria de um número bem maior de algoritmos do que as atrações básicas da teoria métrica dos pontos fixos. Foi possível também expandir os resultados da convergência-q à convergência-r, mediante uma nova definição, a convergência q-sublinear estrita que, semelhante à convergência q-linear, garante a existência e a unicidade do ponto fixo sob condições pouco restritivas. No estudo das realizações q-lineares e das q-superlineares de algoritmos em espaços métricos completos e em espaços com produto interno, foi possível estabelecer cotas inferiores e superiores para o erro em qualquer medida ao longo das iterações, cujas constantes podem ser convenientemente estimadas durante a realização. Para as estimativas das medidas ortogonalizadas em algoritmos de projeção, mostrou-se evidências numéricas de que as cotas superiores (exatas) fornecidas não podem ser melhoradas para nenhuma fase do algoritmo iterativo, e que as cotas inferiores são pelo menos assintoticamente exatas.

Uma das vantagens das estimativas oferecidas nesta tese, é que elas independem de avaliações de propriedades dos operadores discretos, tais como número de condição e norma. Também, são válidas igualmente em algoritmos para problemas lineares e não lineares.

Na Parte II foi desenvolvida uma estratégia polialgorítmica para a solução iterativa de problemas lineares, como caso limite de estratégias para sistemas não lineares, cujo objetivo principal é impor a convergência do quociente, a partir de evidências do estudo prévio da sua geometria.

Nela, procura-se aprender e guardar informações sobre o operador durante o decorrer das iterações, dando ao precondicionador mais funções do que normalmente lhe é atribuída. Numa tentativa de contornar o Teorema de Faber e Manteuffel, ou seja, usar recorrências mais curtas, tranforma-se o precondicionador em um algorítimo de métrica variável. A métrica variável permite recolher bases fora do subespaço de Krylov.

A introdução de medidas epigráficas permitiu uma solução única para o problema

de separação e computabilidade das normas de interesse geral. A estratégia foi sistematicamente desenvolvida com o auxílio de um algoritmo padrão que permite a formação de polialgoritmos convenientes.

Foram feitas aplicações em sistemas lineares simétricos e não simétricos com resultados satisfatórios, usando a medida epigráfica do erro na norma-ℓ_2. Os algoritmos básicos usados na formação de polialgoritmos exemplificados foram o (Good) Broyden, o CG, o CGS e o GMRES(m).

7.2 Futuros Trabalhos

Durante o desenvolvimento das duas partes desta tese, muitas questões cruciais surgiram, e várias delas ficaram sem resposta. Ficaram também evidenciadas algumas necessidades imediatas.

Do ponto de vista de estimativa de erros, a principal questão a ser respondida é: pode-se estabelecer uma cota superior para os fatores-q da realização (ou pelo menos do processo) de uma forma tão geral quanto foram estabelecidas as cotas inferiores ? Parece ao autor perfeitamente possível estabelecer estas cotas para algoritmos específicos. De fato, essa última tarefa esta sendo desenvolvida para fatores-q do processo para alguns dos algoritmos mais bem estabelecidos atualmente.

As estimativas para algoritmos com q-sublinearidades locais - não estritas - (em uma medida específica) não são tão gratificantes. Embora o índice de confiabilidade funcione muito bem, há a necessidade de melhorar a técnica de mapeamentos compostos. Um valor baixo do índice de confiabilidade sempre corresponde a sublinearidades locais. O autor trabalha atualmente com técnicas de remetrizações vislumbradas pela Proposição 16. Entretanto, a principal dificuldade está no custo computacional de remetrizar e retornar à medida de interesse. Uma possibilidade abre-se com a expressão (3.108) que relaciona o fator-q com a taxa p. Esta taxa, normalmente, não é difícil de ser obtida para algoritmos específicos. Talvez possa ser estimada durante a realização, pelo menos por extrapolação. A técnica de extrapolação é usada com algum sucesso na estimativa de erros em estratégias adaptativas em Elementos Finitos. Resultados nessa direção serão de grande importância prática, já que a maioria dos algoritmos atualmente em uso para sistemas lineares não Hermitianos possuem sublinearidades locais.

Certamente, melhores estimativas de erros devem ser obtidas para cada algoritmo em particular por análise de suas propriedades e aplicação dos resultados do Capítulo

3. Este é um trabalho que está sendo efetuado pelo autor, já com bons resultados em vários algoritmos de interesse, e que será relatado posteriormente.

Dentro da formulação de estratégias para sistemas lineares existem também muitas questões abertas e, principalmente, há a necessidade de experimentos numéricos sistemáticos.

Para sistemas Hermitianos há a grande necessidade de encontrar uma atualização de posto unitário para trabalhar junto com o CG. O SR1 possui muitas das propriedades adequadas para trabalhar como precondicionador no esquema algorítmico proposto, como um precondicionador global. Entretanto, por motivos que o autor ainda não conhece perfeitamente, frequentemente colapsa na geração das bases (*breakdown*). Já o BFGS, uma atualização de posto duplo, é o adequado para trabalhar em implementações elemento-por-elemento (EBE) - mantém a deficiência de posto das matrizes - complementado com um precondicionador global (por exemplo, Jacobi). Testes preliminares com o BFGS o revelam promissor. Em ambos os algoritmos é necessário projetar em uma distância epigráfica adequada.

Para a estratégia proposta, há ainda a necessidade de sistematizar os experimentos numéricos com os três algoritmos sugeridos: GB, CGS e GMRES(m), para tentar estabelecer uma ou algumas combinações mais adequadas para problemas com singularidades. Os testes conclusivos feitos até agora pelo autor resultaram nas sugestões do Capítulo 6. O autor enfatiza que os experimentos devem ser feitos dentro de regras de controle (cíclico, remoto, randômico, ...) usadas na área de viabilidade convexa. Isso facilita a exploração das potencialidades de paralelização, para as quais aqueles controles foram desenvolvidos.

Bibliografia

[1] Bauschke, H. H., *Projection Algorithms and Monotone Operators,* Ph.D. Thesis, Department of Mathematics and Statistics, Simon Fraser University, 1996.

[2] Pfister, G. F., *In Search of Clusters, The Ongoing Battle in Lowly Parallel Computing,* Prentice Hall PTR, NJ, 1998.

[3] Saad, Y., Schultz, M., *GMRES: A Generalized Minimal Residual Algorithm for Solving Nonsymmetric Linear Systems*, SIAM J. Sci. Statist. Comput., **7**, pp. 417-424, 1986.

[4] Xue, Z., *Mesh-independence of GMRES for Integral Equations,* Ph.D. Thesis, North Caroline State University, Department of Mathematics, 1995.

[5] Kelley, C. T., *Iterative Methods for Linear and Nonlinear Systems*, SIAM, PA, 1995.

[6] Warren, M. S., Salmon, K. S., *Astrophysical N-body Simulations Using Hierarchical Tree Data Structures*, Proceedings of Supercomputing '92, September 11, 1992.

[7] Mansur, W. J., *at. all., Análise Tridimensional das Cavidades de Serra da Mesa*, LAMEC, Convênio COPPE/FURNAS, Relatórios 1-8, 1992/1993.

[8] Trefethen, L. N., *Matrix Iteration: the Six Gaps between Potential Theory and Convergence*, preprint, Dept. Compt. Science, Cornell University, 1997.

[9] Trefethen, L. N., *The Definition of Numerical Analysis,* Bulletin of the Institute for Mathematics and Applications, March 1993.

[10] Wilkinson, J. H., *The Algebraic Eigenvalue Problem,* Oxford University Press, Oxford, 1965.

[11] Axelsson, O., *Iterative Solution Methods*, Cambridge University Press, NY, 1994.

[12] Golub, G. H., Van Loan, C., *Matrix Computations*, Academic Press, NY, 1989.

[13] Golub, H. G., Van der Vorst, H. A. *Closer to the solution: Iterative Linear Solvers*, preprint, CERFACS, 1996.

[14] Golub, G. H., Meurant, G., *Matrices, Moments and Quadrature II,* Report SCCM 96-18, September 1996.

[15] Faber, V. e Manteuffel, T., *Necessary and Sufficient Conditions for the Existence of a Conjugate Gradient Method*, SIAM Journal on Numerical Analysis, **21**, pp. 352-361, 1984.

[16] Luenberger, D. G., *Introduction do Linear and Nonlinear Programming,* Addison-Wesley Publishing Company, MA, 1973.

[17] Weiss, R., *Error-Minimizing Krylov Subspace Methods,* SIAM J. Sci. Comput., **15**, pp. 511-527, May 1994.

[18] Ashby, S. F., Manteuffel, T. A., Saylor, P. E., *A Taxonomy for Conjugate Gradient Methods*, SIAM J. Numer. Anal., **2**, pp. 1542-1568, 1990.

[19] Dennis, J. E., Schnabel, R. B, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, Prentice-Hall, Inc., NJ, 1983.

[20] Sluis, A. Van der, Vorst, H. A. Van der., *The Rate of Convergence of Conjugate Gradient*, Numer. Math, **48**, pp. 543-560, 1986.

[21] Bathe, K. J., Cemento, A. P., *Some Practical Procedures for the Solution of Nonlinear Finite Element Equations,* Comp. Meth. in Applied Mech. and Eng., **22**, pp. 59-85, 1980.

[22] Chatelin, C. F., *Is Nonnormality a serious computational difficulty in practice?*, CERFACS Tech. Report TR/PA/96/33, 1996.

[23] Oliveira, E. A., *Mecanica das Estruturas*, 1966.

[24] Nevanlinna, O., *Convergence of Iterations for Linear Equations,* Lectures in Mathematics, ETH Zurich, Birkhauser Verlag, Berlin, 1993.

[25] Bartle, R. G., *The Elements of Real Analysis*, John Wiley & Sons, Inc., 2a. edição, 1976.

[26] Goebel, K., Kirk, W. A., *Topics in Metric Fixed Point Theory*, Cambridge Studies in Advanced Mathematics, **28**, Cambridge University Press, Cambridge, 1990.

[27] Ortega, J. M., Rheinboldt, W. C., *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York, 1970.

[28] Censor, Y., Zenios, S. A., *Parallel Optimization, Theory, Algorithms, and Applications,* Oxford University Press, 1997.

[29] Bauschke, H. H., Borwein, J. M., *On Projection Algorithms for Solving Convex Feasibility Problems,* SIAM Review, **38**, No. 3, pp. 367-426, 1996.

[30] Saad, Y., *Iterative Methods for Sparse Linear Systems*, PWS Publishing Co., Boston, 1996.

[31] Barret, R., *et al.*, *Templates for the Solution of Linear Systems: Building Blocks for Iterative Methods*, SIAM, PA, 1993.

[32] Trefethen, L. N. e Bau, D., *Numerical Linear Algebra*, SIAM, PA, 1997.

[33] Araújo, E. R., *Métodos Iterativos em Álgebra Linear Computacional,* 2a. Escola de Verão em Computação Científica, LNCC/CNPq, 17/02 -07/03/1997, 104 pp., Rio de Janeiro, 1997.

[34] Araújo, E. R., *Velocidade de Convergência de Algoritmos: Comportamento e Estimativas de Erro, com Aplicações,* Seminário de Doutorado, orientador W. J. Mansur, COPPE/UFRJ-Departamento de Engenharia Civil, 74 pp., Março, 1996.

[35] Araújo, E. R., *Métodos Iterativos para a Solução de Grandes Sistemas de Equações em Engenharia Estrutural,* Proposta de Tese de Doutorado, orientador W. J. Mansur, COPPE/UFRJ-Departamento de Engenharia Civil, 33 pp., Março, 1996.

[36] Stewart, G. W., *On Sublinear Convergence*, Report CS-TR-3534, University of Maryland, October, 1995.

[37] Strang, G., Fix, G. J., *An Analysis of the Finite Element Method*, Prentice Hall, Inc., Englewood Cliffs, N.J., 1973.

[38] Komolgorov, A. N., Fomin, S. V., *Elementos da Teoria das Funções e de Análise Funcional*, Editora Mir-Moscou, 1982.

[39] Brezzi, F., Gilardi, G., *Functional Analysis,* in: Finite Element Handbook, Capítulo 1, editado por Kardestuncer, H., McGraw-Hill International Editions, 1988.

[40] Oettli, W., Prager, W., *Compatibility of Approximate Solution of Linear Equations with Given Error Bounds for Coefficients and Right Hand Sides*, Numer. Math., **6**, pp. 405-409, 1964.

[41] Arioli, M., Duff., I., Ruiz., D., *Stopping Criteria for Iterative Solvers*, Report RAL-91-057, Rutherford Appleton Laboratory, 1991.

[42] Barra, L. P. S., Coutinho, L. G. A., Mansur, W. J., Telles, J. C. F., *Preconditioning Techniques for Iterative Solutions of BEM Systems of Equations*, BEM XIV, CMP & Elsevier Science Publishers, **1**, pp. 643-650, 1992.

[43] Dahlquist, G., Golub, G. H., Nash, S., *Bounds for the error in linear systems*, in: Proceedings of the Workshop on Semi-Infinite Programming, Springer, pp. 154-172, 1978.

[44] Fischer, B., Golub, G. H., *On the Error Computation for Polynomial based Iteration Methods*, in: Recent Advances in Iterative Methods, Springer-Verlag, pp. 59-67, 1994.

[45] Nour-Omid, E., Parlett, B. N., Raefsky, A., *Comparison of Lanczos with Conjugate Gradient Using Element Preconditioning*, Conferência desconhecida, 1988.

[46] Barth, T., Manteuffel, T., *Variable Metric Conjugate Gradient Methods*, preprint, Department of Mathematics of the Colorado University, Denver, (sem data).

[47] Fokkema, D. R., Sleijpen, G. L. G., *Generalized Conjugate Gradient Squared*, preprint, Mathematical Institute, Utrecht University, Utrecht, 1995.

[48] Sleijpen, G. L. G., Fokkema, D. R., *BiCGSTAB(ℓ) for Linear Equations Involving Matrices with Complex Spectrum*, ETNA, **1**, pp. 11-32, 1993.

[49] Sleijpen, G. L. G., Van der Vorst, H. A., *Maintaining Convergence Properties of BiCGSTAB Methods in Finite Precision Arithmetic*, preprint 861, Dept. Math., University of Utrecht, 1994.

[50] Sleijpen, G. L. G., Van der Vorst, H. A., Fokkema, D. R., *BiCGSTAB(ℓ) and other hybrid Bi-CG methods, Numerical Algorithms*, **7**, pp.75-109, 1994.

[51] Van der Vorst, H. A., *Bi-CGSTAB: A Fast and Smoothly Converging Variant of Bi-CG for the Solution of non-Symmetric Linear Systems*, SIAM Journal on Scientific and Statistical Computing, **12**, pp. 631-644, 1987.

[52] Greenbaum, A., *Iterative Methods for Solving Linear Systems*, SIAM, Philadelphia, 1997.

[53] Chow, E. T., *Robust Preconditioning for Sparse Linear Systems*, Ph.D. Thesis, University of Minnesota, 1997.

[54] Barry S., Bjorstad, P., Gropp, W., *Domain Decomposition - Parallel Multilevel Methods for Elliptic Partial Differential Equations*, Cambridge University Press, 1996.

[55] Kaporin, I. E., Axelsson, O., *On a Class of Nonlinear Equation Solvers Based on the Residual Norm Reduction over a Sequence of Affine Subspaces*, SIAM, J. Sci. Comput., **16**, No.1, pp. 228-249, January, 1995.

[56] Fokkema, D. R., Sleijpen, G. L. G., Van der Vorst, H. A., *Accelerated Inexact Newton Schemes for Large Systems of Nonlinear Equations*, preprint 918, Department of Mathematics, Utrecht University, 1995.

[57] Martínez, J. M., *A Theory of Secant Preconditioners*, Mathematics of Computation, Vol. 60, No. 202, pp. 681-698, 1993.

[58] Morales, J. L., Nocedal, J., *Automatic Preconditioning by Limited Memory Quasi-Newton Updating*, preprint, Departamento de Matemáticas, Instituto Tecnológico Autônomo do México, México, 1997.

[59] Schonauer, W., Weiss, R., Sternecker, P., *Polyalgorithms with Automatic Method of Selection for Iterative Solution of Linear Equations and Eigenproblems*, in Programming Environments for High-Level Problems Solving, pp. 57-67, North-Holland, Amsterdan, 1992.

[60] Borwein, J. M., Vanderwerff, J. D., *Epigraphical and Uniform Convergence of Convex Functions*, preprint, Department of Mathematics and Statistics, Simon Fraser University, Canada, 1996.

[61] Bauschke, H. H., Borwein, J. M., *Legendre Functions and the Method of Random Bregman Projections*, preprint, Centre for Experimental & Constructive Mathematics, Simon Fraser University, Canada, 1995.

[62] Eckstein, J., *Approximate Iterations in Bregman-Function-Based Proximal Algorithms*, Rutcor Research Report RRR 12-06, November, 1997.

[63] Dennis, J. E., Moré, J. J., *Quasi-Newton Methods, Motivation and Theory*, SIAM Review, Vol. 19, No. 1, 1977.

[64] Deuflhard, P., Freund, R., Walter, A., *Fast Secant Methods for the Iterative Solutions of Large Nonsymmetric Linear Systems*, ZIB preprint SC 90-5, Berlin, 1990.

[65] Gibson, T., O'Leary, D. P., Nazareth, L., *BFGS with Update Skipping and Varying Memory*, preprint, Department of Applied Mathematics, University of Maryland, 1996.

[66] Eirola, T., Nevanlinna, O., *Accelerating with Rank-one Updates*, Linear Algebra Appl., **121**, pp. 511-520, 1989.

[67] Araújo, F. C., *Técnicas Iterativas para a Solução de Sistemas de Equações Lineares Oriundos do Método dos Elementos de Contorno*, Tese de M.Sc., COPPE/UFRJ, Rio de Janeiro, 1989.

[68] Barra, L. P. S., Coutinho, L. G. A., Mansur, W. J., Telles, J. C. F., *Iterative Solution of BEM Equations by GMRES Algorithm*, Computers & Structures, 44, pp. 1249-1253, 1992.

[69] Barra, L. P. S., Coutinho, L. G. A., Telles, J. C. F., Mansur, W. J., *Multilevel Hierarchical Preconditioners for Boundary Element Systems*, Engineering Analysis with Boundary Elements, 12, pp. 102-109, 1993.

[70] Brebbia, C. A., Telles, J. C. F., Wrobel, L. C., *Boundary Element Techniques - Theory and Applications in Engineering*, Springer-Verlag, Heidelberg, 1984.

[71] Chatelin, C. F., *Convergence in Finite Precision of Successive Iteration Methods Under High Nonnormality*, CERFACS Tech. Rep. TR/PA/95/43, 1995.

[72] Cho, J. R., *An hpq-adaptive Finite Element Method for Hierarchical Modeling of Elastic Structures*, Ph.D. Thesis, The University of Texas at Austin, December, 1995.

[73] Freund, R. W., Nachtigal, N. M., *QMR: a Quasi-Minimal Residual Method for non-Hermitian linear systems*, Numerische Mathematik, **60**, pp. 315-339, 1991.

[74] Higham, N. J., *How Accurate is Gaussian Elimination* ?, Proceedings of the 13th. Dundee Conference on Numerical Analysis, 1989.

[75] Kelley, C. T., Northrup, J. K., *A Pointwise Quasi-Newton Method for Integral Equations*, SIAM J. Numer. Anal., pp. 1138-1155, 1988.

[76] Mansur, W. J., Araújo, F. C., Malaghini, J. E. B., *Solution of BEM Systems of Equations via Iterative Techniques*, Int. J. Numer. Meth. Engng., **33**, pp. 1823-1841, 1992.

[77] Trefethen, L. N., Trefethen, A. E., Reddy, S. C., Driscoll, T. A., *Hydrodynamics stability without eigenvalues,* Science, **261**, pp. 578-584, 1993.

[78] Varga, R, S., *Matrix Iterative Analysis*, Prentice Hall, NJ, 1962.

[79] Young, D. M., *Iterative Solution of Large Linear Systems, Academic Press*, NY, 1971.