UNIVERSIDADE ESTADUAL DE CAMPINAS
INST.DE MATEMÁTICA, ESTATÍSTICA E CIÊNCIA DA COMPUTAÇÃO

SOBRE ERROS DE ARREDONDAMENTO EM PROCESSOS ALGÉBRICOS

DISSERTAÇÃO DE MESTRADO

Márcia Ferrari Orsi
Orientador: Prof.Dr.Odelar Leite Linhares

CAMPINAS
1973

Este trabalho foi realizado graças ao apoio financeiro da Fundação de Amparo à Pesquisa do Estado de São Paulo, através da Bolsa de Aperfeiçoamento que nos outorgou no período de março de 1971 a julho de 1973.

Aos meus pais, Cida e Arthur Ferrari, aos quais nunca serei suficientemente grata pelo carinho e abnegação com que se dedicaram à minha formação.

Ao meu marido, Armando Geraldo Orsi, cujo apoio, incentivo e orientação foram fundamentais para a realização deste trabalho e de todo o programa de pós-graduação.

Sinceros agradecimentos

Ao Prof. Dr. Odelar Leite Linhares, nosso orientador e caro mestre, pela dedicada atenção que nos dispensou durante a realização não só deste trabalho, mas também de todo o nosso programa de Mestrado.

Ao Prof. Dr Ivan de Queiroz Barros, cujas valiosas sugestões, apresentadas durante a discussão de inúmeros tópicos deste trabalho, foram de fundamental importância para a conclusão do mesmo.

Ao Prof Dr Valdemar W. Setzer, pela irrestrita colaboração que nos dispensou, permitindo-nos a utilização de todos os recursos humanos e materiais do Departamento de Matemática Aplicada do IME da USP.

Ao meu marido, Armando Geraldo Orsi, que, com paciência e dedicação, colaborou de forma decisiva na elaboração geral deste trabalho.

Aos professores e funcionários do IMECC da UNICAMP, do IME da USP e do ICM de São Carlos, que em inúmeras ocasiões nos auxiliaram nas diversas atividades do nosso programa de pós-graduação.

À srta Maria Alice F. Carreira, pela primorosa organização das referências bibliográficas.

Resumo

Neste trabalho são estudados os erros de arredondamento em processos algébricos, com ênfase no método regressivo de análise dos mesmos. Inicialmente se efetua um estudo dos erros de arredondamento nas aritméticas de vírgula fixa e flutuante, recebendo o segundo caso um tratamento mais detalhado. A seguir, são descritas as análises progressiva e regressiva de erros, sendo a última estudada de forma mais completa. Uma breve evolução histórica também se encontra incluída, com a finalidade de situar cronologicamente o desenvolvimento da análise de erros. Finalmente, é efetuada uma análise regressiva de erros do método de eliminação de Gauss, assim como de alguns problemas que podem ser resolvidos através do mesmo.

Abstract

In this work we study the rounding errors occurring in algebraic processes. Emphasis is given to the backward method of error analysis. We make at first a study of rounding errors in fixed and floating-point arithmetic; the second case receives a more detailed treatment. Forward and backward error analysis are then developed, with more attention given to the latter. A brief historical evolution of the field development is included, so that one can place chronologically the development of error analysis. We conclude by making a backward error analysis of Gaussian elimination as well as of some problems which can be solved with this method.

Índice

página

Introdução

O computador eletrônico constitui atualmente um elemento de grande utilidade na resolução numérica de problemas em todos os ramos do conhecimento. Entretanto, tais máquinas utilizam uma quantidade finita de dígitos para a representação dos números com os quais operam, o que pode ocasionar o aparecimento de erros de arredondamento, quer nos dados, quer nos resultados de operações intermediárias. Como a propagação desses erros não é susceptível de controle, pode-se chegar a resultados finais completamente sem significado em relação ao problema original; daí, ser de suma importância complementar a resolução numérica com uma análise de erros do processo utilizado.

Apesar de sua fundamental importância, a análise de erros ainda não está difundida entre os usuários de computadores, principalmente no Brasil, que costumam considerá-la como assunto por demais especializado. Essa concepção é consequência não só do desenvolvimento relativamente recente de uma teoria bem estruturada a respeito do assunto, mas, também, do estudo irregular da matéria, através da leitura de análises efetuadas segundo orientações diversas.

No desenvolvimento deste trabalho tivemos como principal objetivo o estudo da análise de erros relativamente a processos algébricos. Esse estudo foi realizado em quatro etapas sucessivas, as quais deram origem aos quatro capítulos que constituem o texto em questão.

No primeiro capítulo são apresentados alguns assuntos necessários ao entendimento do texto, assim como explicadas as notações utilizadas no mesmo; também estão incluídas, de forma suscinta e sem demonstrações, algumas definições, propriedades e resultados que são empregados no desenvolvimento dos capítulos subsequentes. As normas de vetores e matri-

zes foram consideradas em particular, por serem muito úteis tanto no terceiro quanto no quarto capítulos.

No segundo capítulo são estudados os erros de arredondamento nas aritméticas de vírgula fixa e flutuante. O segundo caso recebe um tratamento mais completo, já que, atualmente, os sistemas de representação em vírgula flutuante são os mais utilizados.

No capítulo três são descritas as duas principais formas de análise de erros: progressiva e regressiva. Nossa atenção concentra-se sobre a análise regressiva que, por apresentar inúmeras vantagens sobre a progressiva, é a mais utilizada atualmente. Uma breve evolução histórica é incluída neste capítulo, com a finalidade de situar cronologicamente o desenvolvimento da análise de erros, assim como citar os principais autores responsáveis pelo mesmo.

No quarto capítulo a análise regressiva de erros é aplicada à eliminação de Gauss e, também, a alguns problemas que podem ser resolvidos através desse método.

Durante toda a fase de redação procuramos dar a este trabalho um aspecto didático, visando a obtenção de um texto de utilidade para todos os que se interessam ou venham a se interessar por esse assunto, em nosso país.

As referências bibliográficas foram feitas de acordo com as normas PNB-66, de 1970, da ABNT. Entretanto, a chamada da bibliografia no texto não foi efetuada de acordo com tais normas e sim utilizando uma citação do tipo (a, p. m-n), em que a representa o número que identifica o livro na relação bibliográfica e m e n os números das páginas inicial e final do trecho citado; esses dois últimos números são excluídos quando o assunto de interesse se encontra tratado ao longo de todo o livro em questão.

Embora os resultados contidos neste trabalho também se apliquem a cálculos realizados manualmente,

... ou sem o auxílio de calculadoras de mesa, a sua utili-
...ação é mais necessária em cálculos que, por envolverem
... número muito grande de operações, exigem o uso de um
...putador eletrônico digital; por esse motivo, preocupamo-
...nos exclusivamente com esse último caso. Por simplici-
...dade, utilizamos, em todo o texto, a palavra "computador"
... em substituição à expressão "computador eletrônico digital".

Apesar de ser a binária a base de operação da maioria dos computadores utilizados para fins científi-cos, os exemplos são apresentados na base dez, tendo sido efe-tuados em uma calculadora de mesa. Com esse procedimento ...amos não só facilitar o entendimento dos exemplos, como, ...também, evitar os erros de arredondamento, que pode-...am ser envolvidos na conversão da base 10 para a base ..., e vice-versa.

1 - Pré-requisitos

1.1 - Espaços vetoriais e euclideanos

No texto é bastante utilizada a estrutura de espaço vetorial real para R^n. Os vetores e escalares desse espaço são representados, respectivamente, por letras latinas e gregas minúsculas. Um vetor x, de coordenadas $x_1, x_2, \cdots, x_n$, relativamente a alguma base, é indicado por

$$x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \quad \text{ou} \quad x^T = (x_1, x_2, \cdots, x_n) \tag{1.1}$$

Em todo o trabalho consideramos, para R^n, a base canônica, cujos elementos são

$$x_1^T = (1,0,\cdots,0), \; x_2^T = (0,1,\cdots,0), \; \cdots, \; x_n^T = (0,0,\cdots,1).$$

A multiplicação escalar utilizada no texto é a usual de R^n, definida por

$$\begin{aligned} <,> \; : R^n \times R^n &\longrightarrow R \\ (x,y) &\longrightarrow <x,y> = \sum_{i=1}^{n} x_i y_i \end{aligned} \tag{1.2}$$

em que $x^T = (x_1, \cdots, x_n)$; $y^T = (y_1, \cdots, y_n)$

1.2. Matrizes

São utilizadas no texto, principalmente, as definições de matriz, submatriz, partição de matriz e determinante. As matrizes consideradas são reais, quadradas, de or-

dem n e indicadas por $A=(a_{ij})$; o conjunto dessas matrizes é indicado por $M_n(R)$. O determinante de uma matriz A é indicado por $\underline{\det A}$. Também são utilizadas as definições e propriedades das operações adição e multiplicação de matrizes e a lei de composição externa multiplicação de um escalar por uma matriz, além das estruturas de espaço vetorial e anel com elemento unidade que essas leis de composição definem em $M_n(R)$. O elemento unidade de $M_n(R)$ é indicado por I. Dada $A \in M_n(R)$, A^T e A^{-1} indicam, respectivamente, a transposta e a inversa de A (é lógico que, no último caso, A é suposta não singular). Sejam $A, B \in M_n(R)$, $A=(a_{ij})$ e $B=(b_{ij})$; as notações $|A|$ e $|A| \leq |B|$ indicam:

(1.3) $$|A| = \left(|a_{ij}|\right) \quad , \quad |A| \in M_n(R)$$

(1.4) $$|A| \leq |B| \text{ se e só se } |a_{ij}| \leq |b_{ij}| \quad , \quad i,j = 1, \cdots, n \, ,$$

onde $|x|$ representa o valor absoluto de x, $x \in R$.

Apresentaremos, a seguir, algumas definições que são muito utilizadas nos capítulos subseqüentes. Todas as matrizes consideradas são supostas pertencentes a $M_n(R)$.

(1.5) Definição - Uma matriz A é dita densa quando apenas uma pequena percentagem de seus elementos é diferente de zero. Em caso contrário, A é dita esparsa ou rarefeita.

Como se pode facilmente concluir, a definição (1.5) tem caráter prático, sendo uma questão de convenção o estabelecimento de um limite entre as duas classes de matrizes.

(1.6) Definição - Denominam-se valores singulares de $A=(a_{ij})$ as raízes quadradas não negativas dos autovalores de AA^T.

(1.7) Definição - $A=(a_{ij})$ diz-se ortogonal se $A^T = A^{-1}$.

(1.8) Definição - $A=(a_{ij})$ e $B=(b_{ij})$ são ditas semelhantes

se e somente se existe uma matriz P não singular tal que $A=P^{-1}BP$

(1.9) Definição - $A=(a_{ij})$ e $B=(b_{ij})$ são ditas equivalentes se e somente se existem matrizes P e Q não singulares tais que $A = Q^{-1}BP$.

(1.10) Definição - Se na definição (1.9) as matrizes P e Q são ortogonais, então, A e B dizem-se ortogonalmente equivalentes. Nesse caso, $A = Q^{-1}BP$ pode ser escrita $A=Q^{T}BP$, em virtude de (1.7)

(1.11) Definição - $A=(a_{ij})$ é dita Hessenberg superior (inferior) se $a_{ij}=0$ para $i \geq j+2$ ($j \geq i+2$)

(1.12) Definição - $A=(a_{ij})$ é chamada tridiagonal se $a_{ij}=0$ para $|i-j|>1$.

(1.13) Definição - $A=(a_{ij})$ é dita diagonalmente dominante se $|a_{ii}| \geq \sum_{j\neq i} |a_{ij}|$, $\forall i$.

1.3. Aplicações lineares

São utilizadas no texto, principalmente, a definição de aplicação linear e a sua representação através de matrizes. As aplicações usadas são do tipo $A: \mathbb{R}^n \to \mathbb{R}^n$; o conjunto dessas aplicações é representado por $L(\mathbb{R}^n)$. Também são utilizadas as definições e propriedades das operações de adição e multiplicação de aplicações lineares e a lei de composição externa multiplicação de um escalar por uma aplicação linear, além das estruturas de espaço vetorial e anel com elemento unidade que essas leis de composição definem em $L(\mathbb{R}^n)$.

As aplicações lineares consideradas no texto são identificadas com suas respectivas matrizes, relativamente à base canônica de $\mathbb{R}^n$, sendo as aplicações e respectivas ma

trizes representadas por um mesmo símbolo. Essa identificação é justificada pelo isomorfismo existente entre $L(R^n)$ e $M_n(R)$, desde que fixada uma base em R^n, quer ambos sejam considerados como espaços vetoriais, quer como anéis.

1.4. Normas de vetores e matrizes

(1.14) Definição - Seja V um espaço vetorial sobre o corpo K dos números reais ou complexos. Uma norma definida em V é uma função que associa a cada $x \in V$ um número real $\|x\|$, chamado norma de x, que satisfaz a

$\|x\| > 0$, se $x \neq \theta$ e $\|\theta\| = 0$, sendo θ o vetor nulo.

$\|\lambda x\| = |\lambda| \cdot \|x\|$, $\forall x \in V$, $\forall \lambda \in K$

$\|x+y\| \leq \|x\| + \|y\|$

A partir de (1.14) pode ser provada a relação

(1.15) $$\left| \|x\| - \|y\| \right| \leq \|x \pm y\| \leq \|x\| + \|y\|$$

No espaço vetorial R^n, utilizado neste trabalho, as normas de vetores mais utilizadas são casos particulares da norma de Hölder, que é definida

(1.16) $$\|x\|_p = \left[\sum_{i=1}^{n} |x_i|^p \right]^{\frac{1}{p}}, \quad p \geq 1,$$

em que x_i é a i-ésima coordenada do vetor $x^T = (x_1, x_2, \cdots, x_n)$, $x^T \in R^n$. Normalmente são utilizadas as normas obtidas pela substituição de p por $1, 2, \infty$ em (1.16), sendo $\|x\|_\infty$ interpretada como $\max_{1 \leq i \leq n} |x_i|$, ou seja,

(1.17) $$\|x\|_1 = \sum_{i=1}^{n} |x_i|$$

(1.19) $$\|x\|_2 = \left(\sum_{i=1}^{n} |x_i|^2 \right)^{\frac{1}{2}}$$

(1.1) $$\|x\|_\infty = \max_{1 \le i \le n} |x_i|$$

Referimos-emos à norma obtida pela substituição de $\underline{p}$ por $\underline{a}$ ($a \geq 1$) em (1.16) por norma a ou $\| \ \|_a$. A norma 2 costuma ser chamada norma euclideana; $\|x\|_2$ representa o comprimento do vetor x, da maneira que é definido usualmente. É importante observar que para a norma ∞ vale

(1.20) Propriedade - Se $x^T = (x_1, x_2, \cdots, x_n)$ e se $|x_i| \leq c$, então, $\|x\|_\infty \leq c$.

(1.20) não se verifica para as normas 1 e 2.

Por outro lado, a partir de (1.17) - (1.19) prova-se:

(1.21) $$\|x\|_\infty \leq \|x\|_2 \leq \sqrt{n}\, \|x\|_\infty$$

(1.22) $$\|x\|_\infty \leq \|x\|_1 \leq n\, \|x\|_\infty$$

Em virtude da definição de produto escalar citada em (1.2) e de (1.18), temos:

(1.23) $$\|x\|_2^2 = \langle x, x \rangle$$

Conseqüentemente, a conhecida desigualdade de Cauchy-Schwarz-Bunyakovskii pode ser escrita

(1.24) $$|\langle x, y \rangle| \leq \|x\|_2 \ \|y\|_2$$

(1.25) Definição - Dada uma seqüência de vetores $(x^{(k)}) \in V$ e uma norma $\| \ \|_a$ em V, diz-se que a seqüência $(x^{(k)})$ converge para $x \in V$ se $\|x^{(k)} - x\|_a \to 0$.

Como vimos em 1.2, o conjunto $M_n(R)$, das matrizes reais quadradas de ordem n, possui estrutura de espaço ve-

torial, relativamente à operação adição de matrizes e à lei de composição multiplicação de um escalar por uma matriz. Portanto, a definição (1.14) vale para $M_n(R)$; basta que se substituam, nos axiomas citados nessa definição, os vetores x e y por matrizes A e B e θ por Θ, sendo Θ a matriz nula.

Podemos associar, a cada norma de vetor $\| \ \|_k$, uma norma de matriz definida por

$$\|A\|_k = \sup_{x \neq \theta} \frac{\|Ax\|_k}{\|x\|_k} \quad , \tag{1.26}$$

ou, equivalentemente,

$$\|A\|_k = \sup_{\|x\|_k = 1} \|Ax\|_k \tag{1.27}$$

Como a norma é uma função contínua, e como o domínio definido por $\|x\| = 1$ é fechado e limitado, podemos substituir "sup" por "max" em (1.26) e (1.27). Por outro lado, em virtude do isomorfismo citado em 1.3, podemos dizer que (1.27) define $\|A\|_k$ como sendo igual ao comprimento do mais longo vetor do conjunto imagem $\{Ax\}$ da esfera unitária $\{x$ tal que $\|x\|_k = 1\}$, sob a aplicação $x \longrightarrow Ax$.

(1.28) Definição. A norma de matriz definida em (1.26) ou (1.27) é dita subordinada à norma de vetor $\| \ \|_k$.

Subordinadas às normas de vetores 1, 2, ∞, temos:

$$\|A\|_1 = \max_j \sum_i |a_{ij}| \quad , \quad i,j = 1, 2, \cdots, n \tag{1.29}$$

$$\|A\|_\infty = \max_i \sum_j |a_{ij}| \quad , \quad i,j = 1, 2, \cdots, n \tag{1.30}$$

$$\|A\|_2 = \left[\max_i \lambda_i \left(A^T A \right) \right]^{\frac{1}{2}}, \tag{1.31}$$

onde $\lambda_i(A^T A)$ indica os autovalores da matriz $A^T A$.

A norma definida em (1.31) é, freqüentemente, denominada norma <u>espectral</u> ou de <u>Hilbert</u>.

(1.32) <u>Propriedade</u> - Sejam $A, B \in M_n(R)$, $A = (a_{ij})$ e $B = (b_{ij})$. Se $|a_{ij}| \leq |b_{ij}|$ $(i, j = 1, \cdots, n)$, então, $\|A\|_\infty \leq \|B\|_\infty$.

A propriedade (1.32) também é verificada pela norma 1, porém, não é válida para a norma 2.

Facilmente se verifica que, para as normas subordinadas, vale:

(1.33) $$\|I\|_k = 1 .$$

(1.34) <u>Definição</u> - Sejam $\|\ \|_p$ e $\|\ \|_k$ normas de vetores e matrizes, respectivamente. Essas normas são ditas <u>consistentes</u> ou <u>compatíveis</u> se

$$\|Ax\|_p \leq \|A\|_k \ \|x\|_p$$

Uma norma de vetores e a norma de matrizes a ela subordinada são sempre compatíveis. Por outro lado, quando são consideradas normas subordinadas em (1.34), existe sempre um vetor x não nulo para o qual se verifica a igualdade.

Podemos determinar uma norma de matriz, a partir de uma norma de vetor dada, de modo que ambas sejam consistentes, através de (1.26) ou (1.27). Em caso contrário, ou seja, para determinar uma norma de vetor, a partir de uma norma de matriz dada, de modo que ambas sejam consistentes, basta tomar

(1.35) $$\|x\|_p = \|A\|_k, \quad \text{com} \quad A = \begin{pmatrix} x_1 & 0 & \cdots & 0 \\ x_2 & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ x_n & 0 & \cdots & 0 \end{pmatrix},$$

sendo $x^T = (x_1, \cdots, x_n)$.

Com a finalidade de evitar a repetição exagerada das expressões "norma de vetores" e "norma de matrizes", utilizaremos, em sua substituição, $\|x\|$ e $\|A\|$, respectivamente.

Além de $\|A\|_1$, $\|A\|_2$ e $\|A\|_\infty$, existe ainda uma importante norma de matrizes, denominada norma euclideana ou de Schur e representada por $\|A\|_E$, que é definida por

$$\|A\|_E = \left(\sum_{ij} |a_{ij}|^2\right)^{\frac{1}{2}} = \left[tr\left(A^T A\right)\right]^{\frac{1}{2}} = \left[tr\left(A A^T\right)\right]^{\frac{1}{2}} \tag{1.36}$$

Embora compatível com $\|x\|_2$, $\|A\|_E$ não é subordinada a nenhuma norma de vetor, o que pode ser concluído de

$$\|I\|_E = n^{\frac{1}{2}} , \tag{1.37}$$

já que para normas subordinadas vale (1,33).

Em trabalhos estritamente matemáticos a norma utilizada em conexão com $\|x\|_2$ é, normalmente, $\|A\|_2$. Entretanto, na prática, geralmente $\|A\|_E$ é utilizada em substituição a $\|A\|_2$, por ser de cálculo mais fácil que esta e por gozar da propriedade

$$\| |A| \|_E = \|A\|_E , \tag{1.38}$$

o que não ocorre com a mesma. (As normas $\|A\|_1$ e $\|A\|_\infty$ também satisfazem (1,38)). As normas $\|A\|_E$ e $\|A\|_2$ são relacionadas por:

$$\|A\|_2 \leq \|A\|_E \leq n^{\frac{1}{2}} \|A\|_2 \tag{1.39}$$

$$\| |A| \|_2 \leq \| |A| \|_E = \|A\|_E \leq n^{\frac{1}{2}} \|A\|_2 \tag{1.40}$$

Apesar da segunda desigualdade de (1.39), $\|A\|_E$ é, frequentemente, uma boa aproximação para $\|A\|_2$.

Para as normas subordinadas e para $\|A\|_E$, além dos axiomas da definição de norma, vale

$$\|AB\|_k \leq \|A\|_k \, \|B\|_k \qquad (k = 1, 2, \infty, E) \tag{1.41}$$

As relações dadas em (1.34) e (1.41) são muito úteis no estudo das aplicações lineares e, em particular, no estudo dos erros de arredondamento contraídos durante a resolução numérica de sistemas de equações lineares.

Veremos, a seguir, um lema que é muito utilizado nos capítulos 3 e 4.

(1.42) Lema Seja F uma matriz tal que $\|F\|_k < 1$, em que k representa uma norma subordinada ou $k = E$. Então, vale

(1.43) $I+F$ é não singular

(1.44) $$\left\| (I+F)^{-1} \right\|_k \leq \frac{1}{1 - \|F\|_k}$$

(1.45) $$\| I - (I+F)^{-1} \|_k \leq \frac{\|F\|_k}{1 - \|F\|_k}$$

Vejamos a demonstração de (1.42). Por absurdo, suponhamos que $I+F$ seja singular. Então, existe um vetor x não nulo, tal que

(1.46) $$(I+F)\,x = 0 \quad \text{ou} \quad Fx = -x$$

Utilizando normas consistentes em (1.46), temos:

$$\|F\|_k \, \|x\|_k \geq \|Fx\|_p = \|x\|_p$$

e, portanto, $\|F\|_k \geq 1$, pois $x \neq 0$ por hipótese. Logo, (1.43) está provado.

Consideremos a identidade

(1.47) $$I = (I+F)^{-1}(I+F),$$

na qual a existência de $(I+F)^{-1}$ é garantida por (1.43)

Seja

(1.48) $R = (I+F)^{-1}$ e, portanto,

(1.49) $I = R + RF$

Aplicando (1.15), (1.41) e (1.33) a (1.49) temos:

(1.50) $1 = \|I\|_k \geq \|R\|_k - \|RF\|_k \geq \|R\|_k - \|R\|_k \|F\|_k$

e, como $1 - \|F\|_k$ é positivo, por hipótese,

(1.51) $$\|R\|_k \leq \frac{1}{1-\|F\|_k}$$

(1.51) prova (1.44) para as normas subordinadas. Essa prova não se aplica a $\|\ \|_E$, já que em (1.50) foi utilizada (1.33). Entretanto, como a prova vale para a norma 2, e como por hipótese $\|F\|_k < 1$, então, por (1.39) podemos dizer que (1.44) também vale para $\|\ \|_E$, "a fortiori".

Por outro lado, de (1.49) e (1.41) temos

$$\|I-R\|_k = \|RF\|_k \leq \|R\|_k \|F\|_k \quad , \text{ ou,}$$

$$\|I-R\|_k \leq \frac{\|F\|_k}{1-\|F\|_k} \quad ,$$

o que prova (1.45).

1.5. Redução de uma matriz à forma diagonal através de equivalência ortogonal.

O teorema apresentado em (1.52) possibilita uma interpretação geométrica muito simples a respeito da representação de uma aplicação linear $A \in L_n(R)$, através de uma matriz $A \in M_n(R)$. Neste trabalho são feitas apenas algumas

considerações a respeito do referido teorema, cuja demonstração pode ser encontrada em (7, p. 9-10)

(1.52) Teorema - Dada $A \in M_n(R)$, existem duas matrizes ortogonais U e V, $U, V \in M_n(R)$, tais que $D = U^T A V$ é uma matriz diagonal. Além disso, é possível determinar U e V de modo que os elementos diagonais de D sejam

(1.53) $$\mu_1 \geq \mu_2 \geq \cdots \geq \mu_r > \mu_{r+1} = \cdots = \mu_n = 0 ,$$

onde r é o posto de A. Em particular, se A é não singular,

(1.54) $$\mu_1 \geq \mu_2 \geq \cdots \geq \mu_n > 0$$

Os números $\mu_1, \mu_2, \cdots, \mu_n$, são os valores singulares de A.

Em muitos livros de Álgebra as matrizes U^T e V são substituídas por matrizes gerais não singulares P^{-1}, Q e todos os μ_i são tomados iguais a 1. Em cálculos realizados em computadores, entretanto, as matrizes ortogonais são mais úteis porque gozam da propriedade $\|Ux\| = \|x\|$, $\forall U$ ortogonal e $\forall x \in R^n$.

Interpretemos A como a representação de um elemento de $L(R^n)$, ou seja, $\forall x, x \in R^n$, $(y = Ax) \in R^n$. Consideremos novas bases ortogonais em R^n, de modo que as matrizes de mudança da base canônica para essas bases sejam V e U, ou seja

(1.55) $$x = V x' \quad , \quad V \text{ ortogonal}$$

(1.56) $$y = U y' \quad , \quad U \text{ ortogonal}$$

As matrizes U e V, definidas em (1.55) e (1.56) são as matrizes citadas em (1.52), pois, de (1.7), (1.55) e (1.56) temos

$$y' = U^T y = U^T A x = U^T A (V x') = (U^T A V) x' = D x'$$

Portanto, em virtude das mudanças de bases, a aplicação, que originalmente era representada por A, obtém uma nova representação que é D. Por outro lado, relativamente às novas bases, essa aplicação possui uma representação bastante simples que, em termos de coordenadas, é dada por

$$(1.57)\qquad \begin{cases} y'_1 = \mu_1 x'_1 \\ y'_2 = \mu_2 x'_2 \\ \vdots \\ y'_r = \mu_r x'_r \\ y'_{r+1} = 0 \\ \vdots \\ y'_n = 0 \end{cases}$$

Podemos então afirmar que a aplicação considerada leva o eixo Ox'_1 de R^n no eixo Oy'_1, também de R^n, com um fator de amplificação $\mu_1 > 0$. O mesmo também ocorre com os eixos $Ox'_2, Ox'_3, \cdots, Ox'_r$, que são levados em $Oy'_2, Oy'_3, \cdots, Oy'_r$, com, respectivamente, fatores de amplificação $\mu_2, \cdots, \mu_r$. Os eixos $Ox'_{r+1}, \cdots, Ox'_n$ são todos levados no vetor nulo de R^n.

Facilmente se verifica que os $\mu_i\ (i=1,2,\cdots,n)$ são os valores singulares de A. De fato, por (1.7)

$$(1.58)\qquad D^TD = (U^TAV)^T(U^TAV) = V^TA^TUU^TAV = V^T(A^TA)V = V^{-1}(A^TA)V$$

Logo, portanto, D^TD e A^TA possuem os mesmos autovalores, já que são semelhantes, como pode ser deduzido de (1.58). Como os autovalores de D^TD são $\mu_1^2, \cdots, \mu_n^2$, podemos concluir que os valores singulares de A são $\mu_1, \cdots, \mu_n$.

Através de (1.52) podemos mostrar que D transforma a esfera unitária $S = \{x' \text{ tal que } \|x'\|_2 = 1\}$

em um hiper elipsóide r-dimensional $E = DS$, de vetores y' tais que

$$\frac{y_1'^2}{\mu_1^2} + \cdots + \frac{y_r'^2}{\mu_r^2} = 1 \quad \text{e} \quad y_{r+1}' = \cdots = y_n' = 0$$

Um dos pontos de E mais afastados da origem é o ponto $(\mu_1, 0, \cdots, 0)$. Se $r < n$, então E contém a origem θ; se $r = n$ então E não contém a origem θ e um dos pontos de E mais próximos de θ é $(0, 0, \cdots, \mu_n)$. Se $r < n$, então, D, e portanto A, são matrizes singulares; se $r = n$, D e A são não singulares e, a partir de (1.57) temos

$$D^{-1} = \begin{pmatrix} \mu_1^{-1} & & 0 \\ & \ddots & \\ 0 & & \mu_n^{-1} \end{pmatrix}$$

Portanto, os valores singulares de A^{-1} são $\mu_1^{-1}, \cdots, \mu_n^{-1}$

A partir dessas observações que fizemos e de (1.2.6) temos

$$\|A\|_2 = \|D\|_2 = \mu_1 \tag{1.59}$$

e, no caso $r = n$,

$$\|A^{-1}\|_2 = \|D^{-1}\|_2 = \mu_n^{-1} \tag{1.60}$$

Resumindo as conclusões mais importantes para uma matriz $A \in M_n(R)$, A não singular, com valores singulares $\mu_1 \geq \cdots \geq \mu_n > 0$, podemos afirmar que existe uma linha L_1 em R^n que é prolongada (ou encurtada) pelo fator μ_1 por A, quando L_1 é levada em uma linha AL_1 de R^n. Por outro lado, existe uma segunda li-

nha L_n, ortogonal a L_1, que é prolongada (ou encurtada) do fator μ_n. Além disso, AL_1 e AL_n são ortogonais em R^n. Um círculo unitário, pertencente ao plano de L_1 e L_n, é levado por A numa elípse com semi-eixos μ_1 e μ_n; essa é a maior distorção que pode ocorrer a qualquer círculo em R^n.

Finalmente, observe-se que o determinante de A satisfaz

(1.61) $$|\det A| = \mu_1 \mu_2 \cdots \mu_n ,$$

pois $$|\det A| = |\det U^T| \, |\det D| \, |\det V| = |\det D|$$

já que as matrizes ortogonais têm determinante igual a (± 1)

1.6. Alguns lemas úteis

Os lemas considerados a seguir auxiliam a apresentar os resultados contidos em análises de erros sob forma mais simples e compacta.

(1.62) Lema. Se $0 \le u < 1$, e se $n = 1, 2, \cdots$, então,

$$1 - nu \le (1-u)^n .$$

De fato, seja $f(u) = (1-u)^n$. Então, pelo teorema de Taylor

$$f(u) = f(0) + u f'(0) + \frac{u^2}{2} f''(\theta u) , \quad 0 < \theta < 1$$

ou $$f(u) = 1 - nu + R(u) ,$$

com $$R(u) = \frac{u^2}{2} n(n-1) (1-\theta u)^{n-2}$$

Como $R(u) \ge 0$, temos $1 - nu \le (1-u)^n$.

(1.63) Lema - $1+x \leq e^x$, $\forall x \geq 0$

Desenvolvendo e^x segundo Taylor, temos:

$$e^x = 1 + x + R(x),$$

com $$R(x) = \frac{x^2}{2} e^{\theta x}, \quad 0 < \theta < 1$$

Como $R(x)$ é não negativo, então, $1+x \leq e^x$.

(1.64) Lema - $e^x \leq 1 + 1{,}01\, x$, para $0 \leq x \leq 0{,}01$

Utilizando o desenvolvimento de e^x apresentada em (1.63) e levando em consideração a hipótese de (1.64) temos:

$$e^x \leq 1 + x + \frac{x^2}{2} e^{0{,}01} = 1 + x\left(1 + \frac{x}{2} e^{0{,}01}\right)$$

$$e^x \leq 1 + x\left(1 + \frac{0{,}01}{2}\, 1{,}01\right)$$

$$e^x \leq 1 + 1{,}01\, x$$

(1.65) Lema Se $n = 1, 2, \cdots$, e $0 \leq n\mu \leq 0{,}01$, então,

$$(1+\mu)^n \leq 1 + 1{,}01\, n\mu$$

De fato, por (1.63) e (1.64)

$$(1+\mu)^n \leq (e^{\mu})^n = e^{n\mu} \leq 1 + 1{,}01\, n\mu$$

(1.66) Lema - Se $|\delta_i| \leq \mu$, $i = 1, \cdots, n$, e se $n\mu \leq 0{,}01$, então

$$1 - n\mu \leq \prod_{1}^{n} (1+\delta_i) \leq 1 + 1{,}01\, n\mu$$

Utilizando (1.62) e (1.65) temos

$$1 - n\mu \leq (1-\mu)^n \leq \prod_{1}^{n} (1+\delta_i) \leq (1+\mu)^n \leq 1 + 1{,}01\, n\mu$$

Às vezes é conveniente escrever (1.66) na forma

$$\prod_{1}^{n}(1+\delta_i) = 1 + 1{,}01\, n\, \theta\, u \quad , \quad |\theta| \leq 1 \quad . \tag{1.67}$$

1.7 - Representação interna dos números nos computadores

Os números são armazenados internamente, nos computadores, através de duas formas principais de representação : vírgula fixa e vírgula flutuante (ou vírgula móvel).

Seja N a base em que opera o particular computador considerado (Atualmente são mais comuns computadores para os quais $N = 2$, entretanto, também são conhecidos computadores com $N = 8, 10$ e 16).

1.7.1 - Sistema de representação em vírgula fixa

Os números representáveis nesse sistema são da forma

$$x = \pm\, 0, d_1 d_2 \cdots d_t \quad , \tag{1.68}$$

em que os d_i $(i = 1, \cdots, t)$ são dígitos que satisfazem a

$$0 \leq d_i \leq N-1 \quad , \tag{1.69}$$

sendo t um número inteiro que ou é fixo para o computador utilizado ($t = 35$ para o IBM 7090, $t = 98$ para o CDC 1604) ou, em alguns casos, pode ser selecionado pelo usuário (IBM 1620).

De (1.68) e (1.69) pode-se concluir

$$-1 < x < 1 \quad \text{ou} \quad |x| < 1 \tag{1.70}$$

O conjunto $F_i = F(N, t)$, dos números representáveis em um certo sistema em vírgula fixa, é um subconjunto finito do conjunto dos números racionais. Na reta real

esse conjunto é representado por pontos equidistantes, pertencentes ao intervalo $(-1, 1)$; a distância entre dois pontos consecutivos desse conjunto é N^{-t}.

1.7.2. Sistema de representação normalizada em vírgula flutuante

Os números representáveis neste sistema são da forma

$$x = \pm 0, d_1 d_2 \cdots d_t \times N^b, \tag{1.71}$$

em que os dígitos d_i $(i = 1, \cdots, t)$ satisfazem a

$$\begin{cases} 1 \leq d_1 \leq N-1 \\ 0 \leq d_i \leq N-1 \quad (i = 2, \cdots, t) \end{cases} \tag{1.72}$$

e b, denominado expoente de x, é um inteiro pertencente ao intervalo $-p \leq b \leq P$, sendo t, p, P inteiros que, normalmente, são fixos para um determinado computador. O valor $0, d_1 d_2 \cdots d_t$ é denominado mantissa ou parte fracionária de x. A condição $d_1 \neq 0$ caracteriza o sistema como normalizado.

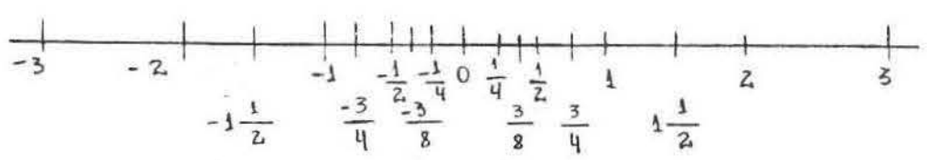
O número zero exige uma representação especial, que varia de computador para computador. As principais existentes são: atribuição do valor zero à mantissa, escolha do valor $(-p)$ para o expoente ou a combinação de ambas. Adotaremos a última, ou seja, representaremos o número zero por

$$0,00\cdots0 \times N^{-p} \tag{1.73}$$

O conjunto $F_\ell = F(N, t, p, P)$, dos números representáveis em um determinado sistema normalizado em vírgula flutuante, é um subconjunto finito do conjunto dos números racionais. Na reta real esse conjunto é representado por pontos não equidistantes. Esse fato é ilustrado em (1.74), através da representação esquemática de F_ℓ, no caso em que

$N=2$, $t=2$, $p=1$ e $P=2$.

(1.74)



De (1.71) e (1.72) facilmente se conclui que qualquer número x pertencente a F_2 pode ser escrito

(1.75) $$x = \pm a \cdot N^b \quad , \quad \text{com} \quad N^{-1} \leq a < 1$$

e, portanto,

(1.76) $$N^{b-1} \leq |x| < N^b$$

Muitos autores utilizam a propriedade (1.75) para definir os sistemas normalizados em vírgula flutuante, a exemplo de Wilkinson (26) e Fox (9).

1.7.3. Precisão simples e múltipla

Indiquemos por t ou o número de dígitos necessários para a representação de um número em vírgula fixa ou o necessário para a representação da mantissa de um número em vírgula flutuante. Diremos, nessas condições, que o computador trabalha com palavras de comprimento t e que a aritmética efetuada com os números assim representáveis é de precisão simples. Às vezes, é interessante trabalhar com precisão maior que uma parte em N^t, utilizando, então, t_1 dígitos, com $t_1 > t$. Se t_1 é múltiplo de t dizemos que o computador trabalha em precisão múltipla; para $t_1 = 2t$, $t_1 = 3t$, ..., usam-se os termos específicos precisão dupla, tripla, etc. Nesses casos, as mantissas com t, $2t$, ... dígitos são chamadas, respectivamente, t-dígital, $2t$-dígital, ...

2- Erros de Arredondamento

2.1. -Classificação dos erros

Os valores numéricos obtidos na solução de problemas podem conter erros de dois tipos: inerentes à formulação matemática da questão e contraídos durante a resolução da mesma através de processos numéricos.

Os erros inerentes à formulação matemática do problema podem ser originados ou pelo fato de ser tal formulação apenas uma aproximação de uma situação física ou por imprecisões nos dados físicos. Os primeiros são, em geral, desprezíveis, como acontece, por exemplo, com os efeitos relativísticos nos problemas de mecânica clássica. Os segundos podem ser desprezíveis se originados por imprecisões em constantes físicas, porém, devem ser cuidadosamente examinados quando causados por erros em dados empíricos.

Os erros acumulados durante a solução do problema através de métodos numéricos podem ser de três tipos: erros grosseiros ou enganos, erros de truncamento e erros de arredondamento. Em geral, os enganos têm sua origem em falhas humanas ou no mal funcionamento da máquina. Como exemplos temos, no primeiro caso, a perfuração incorreta de um cartão e, no segundo caso, a execução errônea de uma adição em virtude de falhas no computador. Erros desse tipo são mais comuns em cálculos efetuados manualmente, podendo, também, porém em escala muito menor, ocorrer com o uso de computadores. Nesse caso, a possibilidade de sua ocorrência deve sempre ser analisada, principalmente quando se torna difícil, quando não impossível, verificar se o valor encontrado é a solução procurada. Erros de truncamento

são os que ocorrem devido à resolução de problemas através de métodos de aproximação ; geralmente surgem pela substituição de um processo infinito ou infinitesimal por uma aproximação finita. Tais erros aparecem, por exemplo, quando uma integral definida é calculada pela regra de Simpson, quando a soma de uma série infinita é calculada por uma fórmula de aproximação ou quando equações diferenciais são resolvidas por métodos de diferenças.
Erros de arredondamento são os que se cometem quando se utilizam, na solução de um problema, valores numéricos com menor número de dígitos do que aquele que realmente se dispõe ; tais aproximações podem ser introduzidas nos dados do problema ou podem ocorrer em resultados intermediários, durante a resolução do mesmo. Como exemplos temos a introdução do valor numérico 3,14 em substituição ao dado π de um problema e a utilização dos números 1,33 e 3,49 como resultados, respectivamente, das operações 1 : 3 e 2,41 × 1,45. A introdução dos erros de arredondamento é inevitável, quer por limite da capacidade de armazenamento, que ocorre nos computadores e máquinas de calcular em geral, quer por limitação de tempo, quando as operações são efetuadas manualmente.

Os erros que acabamos de conceituar podem ser apresentados em duas formas principais, absoluta e relativa, definidas por :

a) $\text{erro absoluto} = \text{valor aproximado} - \text{valor verdadeiro}$

b) $\text{erro relativo} = \dfrac{\text{erro absoluto}}{\text{valor verdadeiro}}$

Quando julgam conveniente, muitos autores consideram o erro relativo definido em relação ao valor aproximado, isto é,

b') $\text{erro relativo} = \dfrac{\text{erro absoluto}}{\text{valor aproximado}}$

A justificativa para essa atitude é o não conhecimento, de maneira geral, do valor verdadeiro da grandeza considerada.

É importante observar que, freqüentemente, o erro relativo é muito mais informativo do que o absoluto. Assim, por exemplo, sejam $\bar{x} = 0{,}8976$ e $\bar{y} = 0{,}0059$ aproximações para grandezas $x = 0{,}89764$ e $y = 0{,}00586$. O erro absoluto é igual a $0{,}4 \times 10^{-4}$ em ambos os casos, enquanto os erros relativos, $\frac{2}{44882}$ e $\frac{2}{293}$ para x e y, respectivamente, diferem bastante.

Neste trabalho são estudados os erros de arredondamento em processos algébricos. Esses erros, quer na forma absoluta, quer na relativa, são representados por letras gregas minúsculas; apesar disso, acreditamos que a notação adotada não causará equívocos porque, em todo o texto, não há possibilidade de confusão entre ambas as formas de erros. As expressões solução exata e resultado exato (soma, produto exatos, etc) indicam a solução de um problema e o resultado de uma operação obtidos sem a incidência de erros. Embora redundantes, essas expressões permitem a distinção entre os valores exatos e os valores fornecidos pelo computador, os últimos geralmente afetados de erros. Os valores fornecidos pelo computador são chamados, neste trabalho, solução e resultado aproximados.

1.2 - Arredondamento: definições e técnicas

Indiquemos por F o conjunto dos números representáveis ou no sistema em vírgula fixa ou no sistema normalizado em vírgula flutuante, ou seja, $F = F_z$ ou $F = F_\ell$, F_z e F_ℓ definidos em 1.7. Seja x, $x \notin F$, o número que queremos aproximar por um elemento de F.

Denomina-se <u>arredondamento</u> ao processo pelo qual se determina um número $\bar{x} \in F$, tal que $\bar{x}$ é uma aproximação de x. O erro cometido pela substituição de $\bar{x}$ por x é chamado <u>erro de arredondamento</u> (absoluto ou relativo, conforme seja considerado de maneira absoluta ou relativa) de $\bar{x}$. O valor $\bar{x}$ é dito <u>valor arredondado de x</u> ou <u>representação interna de x</u>.

O arredondamento pode ser obtido através de diferentes técnicas, entre as quais as mais utilizadas são as que são descritas a seguir.

Nas considerações que se seguem, $d(a,b)$ representa a distância entre a e b.

<u>Arredondamento propriamente dito</u> - O valor arredondado de x, usualmente representado por x_R, deve satisfazer às condições:

$$(2.4)\qquad \begin{cases} x_R \in F \\ d(x_R, x) \leq d(y, x)\ , \ \forall y \in F \\ \forall y \in F,\ y \neq x_R\ ,\ \text{se } d(y,x) = d(x_R, x), \text{ então, } |y| < |x_R| \end{cases}$$

<u>Arredondamento por corte</u> - O valor arredondado de x, normalmente representado por x_C, deve satisfazer às condições:

$$(2.5)\qquad \begin{cases} x_C \in F \\ d(x_C, x) \leq d(y, x)\ , \ \forall y \in F \\ |x_C| \leq |x| \ . \end{cases}$$

Em muitos livros o arredondamento por corte é denominado truncamento. Preferimos não utilizar esse termo, para evitar equívocos com o erro de truncamento definido em 2.1. Por outro lado, frequentemente o arredondamento propriamente dito é chamado, simplesmente, de arredondamento. Escolhemos a primeira dessas denominações com a finalidade de evitar confusão entre a técnica de obtenção e o processo

de arredondamento.

2.3. - Arredondamento na aritmética em vírgula fixa

2.3.1 - Determinação de x_R e x_C

Seja $x \notin F_i$ o número que queremos aproximar por um elemento de $F_i = F(N,t)$. De acordo com o tipo de arredondamento utilizado pelo computador considerado, x é aproximado por x_R ou x_C.

Determinação de x_R

Para determinar x_R consideramos a representação infinita de $|x| + \frac{1}{2} N^{-t}$ (aquela terminando em zeros, no caso de escolha) e retemos os t primeiros dígitos após a vírgula, desprezando os demais. A aproximação x_R é formada pelo sinal de x e pelos dígitos assim obtidos.

Como $x_R \in F_i$, então, por (1.70), podemos afirmar que só podem ser aproximados por elementos de F_i os números x pertencentes ao intervalo

$$R_i = \left(-1 + \frac{1}{2} N^{-t} \;,\; 1 - \frac{1}{2} N^{-t} \right) \tag{2.6}$$

Referir-nos-emos a R_i como o conjunto dos números aproximáveis por elementos de F_i, quando é utilizado o arredondamento propriamente dito.

(2.7) Exemplos. Seja $F_i = F(10,4)$

$x = 0,00157469$	$x_R = 0,0016$
$y = -0,896512236$	$y_R = -0,8965$
$z = 0,99999$	z não pode ser aproximado por um elemento de F_i.
$t = 0,682351$	$t_R = 0,6824$
$p = -0,1999782$	$p_R = -0,2000$

Determinação de x_c

Para obter x_c pegamos a expansão infinita de x (aquela terminando em zeros, no caso de escolha) e retemos os t primeiros dígitos após a vírgula, desprezando os demais.

É claro que o conjunto dos números aproximáveis por elementos de F_i, quando é utilizado o arredondamento por corte, difere daquele apresentado em (2.6.). Indicaremos por C_i esse conjunto, que é definido por

(2.8) $$C_i = (-1, +1)$$

(2.9) Exemplos - Consideremos o arredondamento por corte dos números utilizados em (2.7):

$x_c = 0,0015$ $\quad t_c = 0,6823$

$y_c = -0,8965$ $\quad p_c = -0,1999$

$z_c = 0,9999$

2.3.2. Delimitações para os erros de arredondamento

Como observado em 1.7.1., os elementos de F_i são representados, na reta real, por pontos equidistantes. Em virtude de ser N^{-t} a distância entre dois pontos consecutivos desse conjunto, facilmente se deduz de (2.4) e (2.5) que:

(2.10.) $$|x_R - x| \leq \frac{1}{2} N^{-t}$$

(2.11.) $$|x_C - x| \leq N^{-t}$$

2.3.3. Erros de arredondamento nas operações aritméticas

Um número real x pode ser aproximado, dependendo do computador considerado, por x_R ou x_C. Em virtude das correspondentes delimitações para os erros de arredondamento

serem diferentes, adotaremos um procedimento que permita, em ambos os casos, a utilização dos resultados contidos neste trabalho.

A representação interna de um número real x, num computador que utilize sistema em vírgula fixa, será indicada por $fi(x)$. Assim sendo, de acordo com o tipo de arredondamento efetuado no computador considerado, temos:

$$fi(x) = x_R \quad \text{ou} \quad fi(x) = x_{\mathbb{C}} \tag{2.12.}$$

A expressão $z = fi(x * y)$ indicará que x, y e z são representações internas de números reais, no computador considerado, e que z é obtido da seguinte maneira: a operação $*$ é efetuada e, quando necessário, o resultado $x * y$ obtido é arredondado, de acordo com o procedimento próprio do computador em questão. Quando $*$ representa a divisão, y é suposto diferente de zero. Portanto, na realidade, o computador executa, em substituição à operação $*$, uma operação aproximada, definida por:

$$fi(x*y) = (x*y)_R \quad \text{ou} \quad fi(x*y) = (x*y)_{\mathbb{C}}\ , \tag{2.13}$$

de acordo com o procedimento de arredondamento do computador em questão. Como observado em 2.1., $fi(x*y)$ é chamado <u>valor aproximado de $(x*y)$</u>.

O conjunto dos números aproximáveis por elementos de F_i será representado por A_i, ou seja,

$$A_i = R_i \quad \text{ou} \quad A_i = \mathbb{C}_i\ , \tag{2.14}$$

de acordo com o tipo de arredondamento do computador utilizado. (R_i e $\mathbb{C}_i$ são definidos em (2.6) e (2.8), respectivamente).

Denominaremos <u>unidade absoluta de arredondamento</u>, e indicaremos por u_A,

$$u_A = \frac{1}{2} N^{-t} \quad \text{ou} \quad u_A = N^{-t}\ , \tag{2.15}$$

conforme o computador considerado utilize arredondamento propriamente dito ou por corte, respectivamente.

Descreveremos a realização das operações aritméticas fundamentais e da multiplicação escalar em um computador que utiliza sistema em vírgula fixa. Para a descrição dessas operações será usado um sistema genérico $F_i = F(N, t)$ e para os exemplos serão utilizados $F_i = F(10,4)$ e arredondamento propriamente dito.

Adição e subtração

Sejam x e y dois elementos quaisquer de F_i. A determinação de $fi(x \pm y)$ não é sempre possível porque pode ocorrer $(x \pm y) \notin A_i$. Quando realizáveis, a adição e subtração não envolvem erros de arredondamento e satisfazem:

(2.16) $$fi(x \pm y) = x \pm y$$

Exemplos

$x = 0,1542$ $\quad y = 0,6321$ $\quad fi(x+y) = 0,7863$

$r = 0,8314$ $\quad s = 0,2765$ $\quad fi(r+s)$ não pode ser determinado no sistema considerado.

Multiplicação

Sejam x e y elementos quaisquer de F_i. O computador determina o produto $x \times y$, que geralmente é um número $2t$-digital e sempre pertence a A_i, e a seguir o arredonda para t dígitos. Portanto, $fi(x \times y)$ satisfaz:

(2.17) $$fi(x \times y) = xy + \varepsilon \quad , \qquad |\varepsilon| \leq \mu_A \, ,$$

onde xy representa o produto exato de x e y.

Exemplo

$x = 0,1832$ $\quad y = 0,4327$

$xy = 0,07927064$ $\quad fi(x \times y) = 0,0793$ $\quad |\varepsilon| = 0,2936 \times 10^{-4}$.

Divisão

Sejam x e y elementos quaisquer de $F_{\hat{z}}$. A divisão $x:y$ somente pode ser efetuada quando $|y| > |x|$; quando isso ocorre, o computador determina o quociente $\frac{x}{y}$, que geralmente possui mais de t dígitos, e o arredonda. Portanto, quando realizável, a divisão em vírgula fixa satisfaz a:

$$fi(x:y) = \frac{x}{y} + \varepsilon \quad , \quad |\varepsilon| \leq \mu_A \quad , \tag{2.18}$$

onde $\frac{x}{y}$ representa o quociente exato de x e y.

Exemplo

$x = 0,1234$ $\qquad y = 0,8976$

$\frac{x}{y} = 0,13747731\cdots$ $\qquad fi(x:y) = 0,1375$ $\qquad |\varepsilon| = 0,2228\cdots \times 10^{-4}$

Multiplicação escalar

A determinação do produto escalar $\langle x,y\rangle$, definido em (1.2), é considerada neste trabalho por ser esse cálculo muito comum nos problemas que envolvem matrizes. O valor $fi(\langle x,y\rangle)$ pode ser obtido através de dois processos diferentes, que originam erros de arredondamento finais distintos. O primeiro consiste na determinação dos produtos $x_i y_i$, arredondamento para t dígitos dos valores obtidos e posterior soma desses valores arredondados. O segundo processo consta da determinação dos produtos $x_i y_i$ em precisão dupla, soma desses valores $2t$-digitais e posterior arredondamento da soma obtida para t dígitos. O produto escalar fornecido pelo segundo método é usualmente representado por $fi_{\hat{z}}(\langle x,y\rangle)$, notação que o diferencia do valor fornecido pelo primeiro método, que é representado por $fi(\langle x,y\rangle)$. Costuma-se dizer que no cálculo de $fi_{\hat{z}}(\langle x,y\rangle)$ há <u>acumulação de produtos escalares</u>. Os produtos escalares fornecidos pelos métodos descritos verificam:

$$(2.19) \qquad fi(\langle x,y\rangle) = \sum_{i=1}^{n} x_i y_i + \varepsilon \quad , \quad |\varepsilon| \leq n \mu_A$$

$$(2.20) \qquad fi_2(\langle x,y\rangle) = \sum_{i=1}^{n} x_i y_i + \varepsilon \quad , \quad |\varepsilon| \leq \mu_A$$

Exemplos - Consideremos a determinação da multiplicação escalar:

$$(2.21) \qquad \langle x,y\rangle = 0{,}7865 \times 0{,}4361 + 0{,}0053 \times 0{,}2410 + 0{,}4126 \times 0{,}6325$$

Armazenando os produtos parciais na forma 2t-digital e adicionando-os obtemos:

$$(2.22) \qquad \langle x,y\rangle = 0{,}34299265 + 0{,}00127730 + 0{,}26096950 = 0{,}60523945$$

Arredondando para t dígitos o resultado obtido em (2.22) temos:

$$fi_2(\langle x,y\rangle) = 0{,}6052 \text{ , com } |\varepsilon| = 0{,}3945 \times 10^{-4}.$$

Arredondando separadamente cada um dos produtos parciais e somando-os, obtemos:

$$(2.23) \qquad fi(\langle x,y\rangle) = 0{,}3430 + 0{,}0013 + 0{,}2610 = 0{,}6053 ,$$

com $|\varepsilon| = 0{,}6055 \times 10^{-4}$, como pode ser concluído da comparação de (2.23) e (2.22)

Como pode ser observado em (2.19) e (2.20), a determinação de $\langle x,y\rangle$ é efetuada de maneira mais satisfatória em computadores que possibilitam a acumulação de produtos escalares e que, portanto, possuem acumulador de precisão dupla. Nesses computadores a divisão também é efetuada, usualmente, de forma mais precisa do que nos computadores que só possuem acumulador de precisão simples, através do armazenamento do dividendo

em precisão dupla e do divisor em precisão simples, o que permite a obtenção de um quociente corretamente arredondado; os dividendos t-digitais são estendidos à forma 2t-digital pela adição de zeros. A utilização do acumulador de precisão dupla para a determinação de produtos escalares e quocientes é bastante vantajosa no cálculo de expressões do tipo:

$$\frac{\sum_{i=1}^{n} x_i y_i}{z} \qquad (2.24)$$

cujo numerador pode ser determinado exatamente, originando um valor 2t-digital que é dividido por z. O arredondamento para t dígitos, desse quociente, fornece:

$$\begin{cases} fl_2\left(\dfrac{\sum_{i=1}^{n} x_i y_i}{z}\right) = \dfrac{\sum_{i=1}^{n} x_i y_i}{z} + \varepsilon, \\ \text{com} \quad |\varepsilon| \leq u_A \end{cases} \qquad (2.25)$$

enquanto que em computadores que dispõem apenas de acumulador de precisão simples o cálculo de (2.24) fornece:

$$\begin{cases} fl\left(\dfrac{\sum_{i=1}^{n} x_i y_i}{z}\right) = \dfrac{\sum_{i=1}^{n} x_i y_i}{z} + \varepsilon', \\ \text{com} \quad |\varepsilon'| \leq \left(\dfrac{n}{2} + 1\right) u_A \end{cases} \qquad (2.26)$$

Como pode ser observado em (2.25) e (2.26), a expressão (2.24) é calculada de forma mais precisa em computadores que dispõem de ambos os recursos citados

2.4.- Arredondamento na aritmética em vírgula flutuante

2.4.1.- Determinação de x_R e x_c

Seja $x \notin F_\ell$ o número que queremos aproximar por um elemento de $F_\ell = F(N, t, p, P)$. De acordo com o computador utilizado, x é aproximado por x_R ou x_c.

Determinação de x_R

Se $x = 0$, tomamos $x_R = 0,0\cdots0 \times N^{-p}$

Se $x \neq 0$, escolhemos a real e b inteiro tais que

$$|x| = a \times N^b, \quad \text{com} \quad N^{-1}\left(1 - \frac{1}{2}N^{1-t}\right) \leq a < \left(1 - \frac{1}{2}N^{-t}\right) \tag{2.27}$$

Se b não pertence ao intervalo $-p \leq b \leq P$, então, x não pode ser representado no sistema considerado. Em caso contrário, tomamos a expansão infinita de $a + \frac{1}{2}N^{-t}$ (aquela terminando em zeros no caso de escolha) e retemos apenas os t primeiros dígitos após a vírgula, desprezando os demais. A aproximação x_R é constituída pelo sinal de x, pelos t dígitos assim obtidos e pelo expoente b determinado em (2.27).

O conjunto dos números aproximáveis por elementos de F_ℓ, quando é utilizado o arredondamento propriamente dito, é:

$$R_\ell = \left(-N^P, N^P\right) \tag{2.28}$$

(2.29) Exemplos - Seja $F_\ell = F(10, 4, 18, 19)$

$p = 0,163173 \times 10^5$	$p_R = 0,1632 \times 10^5$
$q = 0,000697412$	$q_R = 0,6974 \times 10^{-3}$
$x = 1308,564 \times 10^7$	$x_R = 0,1309 \times 10^{11}$
$y = 9,99999\cdots$	$y_R = 0,1000 \times 10^2$
$z = 99,99999 \times 10^{18}$	z_R não pode ser representado no sistema em questão.

Determinação de x_c

Se $x=0$, $x_c = 0,0\cdots0 \times N^{-p}$

Se $x \neq 0$, escolhemos a real e b inteiro tais que

(2.30) $$|x| = a \times N^b, \quad \text{com} \quad N^{-1} \leq |a| < 1$$

Se $b \notin [-p, P]$, então, x não pode ser representado no sistema em questão. Em caso contrário, tomamos a expansão infinita de a na base N (aquela terminando em zeros, no caso de escolha) e retemos os t primeiros dígitos após a vírgula, desprezando os demais. A aproximação x_c é formada pelo sinal de x, pelos t dígitos assim obtidos e pelo expoente b determinado em (2.30). O conjunto C_ℓ, dos números aproximáveis por elementos de F_ℓ, coincide com R_ℓ, definido em (2.2.8)

(2.31) Exemplos - Consideremos o arredondamento por corte dos números utilizados em (2.29)

$p_c = 0,1631 \times 10^5$ $\quad$ $y_c = 0,9999 \times 10$

$q_c = 0,6974 \times 10^{-3}$ $\quad$ z_c não pode ser representado no sistema em questão.

$x_c = 0,1308 \times 10^{11}$

2.4.2. - Delimitações para os erros de arredondamento

O caso $x=0$ é trivial. Seja $x \neq 0$ e, sem perda de generalidade, suponhamos $x > 0$.

Representemos x na forma

(2.32) $$x = a \times N^b, \quad N^{-1} \leq a < 1.$$

Em virtude de (2.32) podemos afirmar

(2.33) $$N^{b-1} \leq x < N^b$$

No intervalo $[N^{b-1}, N^b]$, definido em (2.33), os elementos de $F_\ell(N, t, p, P)$ são representados por pontos equidistantes. Em conseqüência de ser N^{b-t} a distân

cia entre dois pontos consecutivos desse conjunto, podemos afirmar, por (2.4) e (2.5),

$$|x_R - x| \leq \frac{1}{2} N^{b-t} \tag{2.34}$$

$$|x_C - x| \leq N^{b-t} \tag{2.35}$$

Indicando por ε_R (ε'_R) e ε_C (ε'_C) os erros relativos definidos em relação ao valor verdadeiro (aproximado), cometidos pela substituição de x por x_R e x_C, respectivamente, podemos escrever, por (2.33) - (2.35):

$$|\varepsilon_R| = \frac{|x_R - x|}{|x|} \leq \frac{\frac{1}{2} N^{b-t}}{N^{b-1}} = \frac{1}{2} N^{1-t} \tag{2.36}$$

$$|\varepsilon_C| = \frac{|x_C - x|}{|x|} \leq \frac{N^{b-t}}{N^{b-1}} = N^{1-t} \tag{2.37}$$

$$|\varepsilon'_R| = \frac{|x_R - x|}{|x_R|} \leq \frac{\frac{1}{2} N^{b-t}}{N^{b-1}} = \frac{1}{2} N^{1-t} \tag{2.38}$$

$$|\varepsilon'_C| = \frac{|x_C - x|}{|x_C|} \leq \frac{N^{b-t}}{N^{b-1}} = N^{1-t} \tag{2.39}$$

Utilizando as relações (2.36) - (2.39), obtemos:

$$x_R = x(1+\varepsilon_R) = \frac{x}{1+\varepsilon'_R} \quad ; \quad |\varepsilon_R|, |\varepsilon'_R| \leq \frac{1}{2} N^{1-t} \tag{2.40}$$

$$x_C = x(1+\varepsilon_C) = \frac{x}{1+\varepsilon'_C} \quad ; \quad |\varepsilon_C|, |\varepsilon'_C| \leq N^{1-t} \tag{2.41}$$

2.4.3. - Erros de arredondamento nas operações aritméticas

Procedendo de maneira análoga à utilizada em 2.3.3., adotaremos um procedimento que permita o uso dos resultados contidos neste trabalho para computadores que utili-

jam qualquer um dos tipos de arredondamento considerados. Indicaremos:

(2.42) $fl(x) = x_R$ ou $fl(x) = x_C$

(2.43) $fl(x * y) = (x * y)_R$ ou $fl(x * y) = (x * y)_C$

(2.44) $A_\ell = R_\ell$ ou $A_\ell = C_\ell$,

de acordo com o tipo de arredondamento do computador em questão.

Em expressões do tipo $fl(x * y * z * w)$, suporemos que as operações sejam efetuadas da esquerda para a direita.

Denominaremos unidade absoluta e relativa de arredondamento, para computadores que utilizam sistema normalizado de representação em vírgula flutuante em que $F_\ell = F(N, t, p, P)$, aos valores u_A e u definidos, respectivamente, por

(2.45) $u_A = \frac{1}{2} N^{b-t}$ ou $u_A = N^{b-t}$

(2.46) $u = \frac{1}{2} N^{1-t}$ ou $u = N^{1-t}$,

conforme o tipo de arredondamento utilizado pelo computador considerado.

Na descrição das operações aritméticas, efetuada para um sistema $F_\ell = F(N, t, p, P)$, os operandos serão representados por $x_1, x_2, \cdots, x_i, \cdots$, sendo $x_i = a_i N^{b_i}$, e o resultado por $x = a N^b$. Nos exemplos, realizados em um sistema $F_\ell = F(10, 4, -35, 38)$ que utiliza arredondamento propriamente dito, as mantissas são indicadas entre parênteses e os erros de arredondamento escritos na forma $|\varepsilon| = x \cdot 10^{-3}$, em que x é um número real qualquer. Adotamos essa forma de representação para tornar imediata a comparação entre o erro obtido e a unidade relativa de arredondamento do sistema considerado, $\frac{1}{2} 10^{-3}$. Chamaremos deslocamento da mantissa à direita (esquerda)

ao deslocamento ocasionado pela multiplicação da mantissa por uma potência negativa (positiva) da base N. Descreveremos a realização das operações aritméticas fundamentais e da multiplicação escalar, em vírgula flutuante, para computadores que não dispõem de acumulador de precisão dupla e para computadores que possuem tal acumulador; nesse último caso, também será considerada a acumulação de somas e produtos escalares.

Adição e subtração em computadores que dispõem de acumulador de precisão dupla.

Consideremos a determinação de $x = x_1 + x_2$. Seja $b_1 \geq b_2$. O computador efetua, inicialmente,

$$x = x_1 + x_2 = N^{b_1}\left(a_1 + N^{b_2 - b_1} a_2\right) ,$$

ou seja, desloca a_2 de $(b_1 - b_2)$ posições à direita e, a seguir, efetua a adição em precisão dupla. Quando necessário, a soma obtida é multiplicada por uma potência conveniente de N, através de um deslocamento à esquerda ou à direita, de modo que sua mantissa satisfaça (1.75); nesse caso, o expoente b_1 é ajustado de maneira a concordar com o deslocamento efetuado. Essa última operação é denominada, usualmente, normalização da mantissa. Finalmente, o computador arredonda, para t dígitos, a parte fracionária assim obtida.

O número máximo de deslocamentos à direita, que pode tornar-se necessário, é um, pois, por (1.75),:

$$\left|a_1 + N^{b_2 - b_1} a_2\right| \leq |a_1| + N^{b_2 - b_1}|a_2| < 2.$$

Entretanto, em virtude de cancelamentos, podem ser necessários até $(t-1)$ deslocamentos à esquerda, como ocorre no exemplo apresentado em (2.53).

Se $b_1 - b_2 > t$, então, x_2 não contribui pa-

ra a soma aproximada, por não ter influência em seus primeiros t dígitos, e ocorre $fl(x_1 + x_2) = x_1$.

Se $x_1 = 0$, então, $fl(x_1 + x_2) = x_2$. Analogamente, se $x_2 = 0$, $fl(x_1 + x_2) = x_1$. Em ambos os casos o erro de arredondamento é nulo.

Como no cálculo de $fl(x_1 + x_2)$ é efetuado somente o arredondamento, para t dígitos, da soma exata normalizada, verifica-se:

$$|fl(x_1 + x_2) - (x_1 + x_2)| \leq \mu_A \tag{2.47}$$

Essa forma de erro também é utilizada, porém, de maneira usual, na aritmética em vírgula flutuante é utilizado o erro de arredondamento relativo dado por:

$$fl(x_1 + x_2) = (x_1 + x_2)(1 + \varepsilon), \quad |\varepsilon| \leq \mu \tag{2.48}$$

Procedendo com a subtração de forma análoga à utilizada no tratamento da adição, pode ser encontrado o resultado

$$fl(x_1 - x_2) = (x_1 - x_2)(1 + \varepsilon), \quad |\varepsilon| \leq \mu \tag{2.49}$$

Exemplos

$$fl\left(10^{-3}(0,1572) + 10^{4}(0,9813)\right) = 10^{4}(0,9813) \tag{2.50}$$

$fl(x_1 + x_2) = x_1$ porque $b_1 - b_2 = 7$ e $7 > 4$.

$|\varepsilon| \simeq 0,00002 \times 10^{-3}$

$$fl\left(10^{6}(0,7513) + 10^{3}(0,4624)\right) = fl\left(10^{6}(0,75130000 + 0,00046240)\right) = \tag{2.51}$$
$$= fl\left(10^{6}(0,75176240)\right) = 10^{6}(0,7518)$$

$|\varepsilon| \simeq 0,05002 \times 10^{-3}$

$$fl\left(10^{3}(0,3561) + 10^{3}(0,8225)\right) = fl\left(10^{3}(0,35610000 + 0,82250000)\right) = \tag{2.52}$$
$$= fl\left(10^{3}(1,17860000)\right) = fl\left(10^{4}(0,11786000)\right) = 10^{4}(0,1179).$$

$|\varepsilon| \simeq 0,33939 \times 10^{-3}$.

(2.53) $$fl\left(10^{-5}(0,4318)+10^{-5}(-0,4316)\right)=fl\left(10^{-5}(0,43180000-0,43160000)\right)=$$
$$=fl\left(10^{-5}(0,00020000)\right)=fl\left(10^{-8}(0,20000000)\right)=10^{-8}(0,2000).$$
$$|\varepsilon|=0$$

(2.54) $$fl\left(10^{4}(0,1001)+10^{3}(-0,9941)\right)=fl\left(10^{4}(0,10010000-0,09941000)\right)=$$
$$=fl\left(10^{4}(0,00069000)\right)=fl\left(10(0,69000000)\right)=10(0,6900).$$
$$|\varepsilon|=0$$

(2.55) $$fl\left(10^{3}(0,9652)+10^{2}(0,5828)\right)=fl\left(10^{3}(0,96520000+0,05828000)\right)=$$
$$=fl\left(10^{3}(1,02348000)\right)=fl\left(10^{4}(0,10234800)\right)=10^{4}(0,1023).$$
$$|\varepsilon|\simeq 0,46899\times 10^{-3}.$$

Como pode ser observado através dos exemplos (2.53) e (2.54), em adições efetuadas em computadores que dispõem de acumulador de precisão dupla, a ocorrência de cancelamento dos dígitos localizados imediatamente à direita da vírgula implica na nulidade do erro de arredondamento. Em computadores que não possuem tal acumulador esse fato nem sempre se verifica, como poderá ser notado em (2.96).

Consideraremos, a seguir, adições de mais de duas parcelas. No estudo dessas, assim como, de maneira geral, de outras operações com mais de dois operandos, frequentemente são encontradas expressões do tipo

(2.56) $$\prod_{i=1}^{n}\left(1+\delta'_i\right)=\left(1+\delta'_1\right)\left(1+\delta'_2\right)\cdots\left(1+\delta'_n\right)\quad,\quad|\delta'_i|\leq u$$

em que os δ'_i representam erros de arredondamento. Procurando simplificar tais expressões, aplicaremos (1.67) às mesmas, obtendo, em seu lugar:

(2.57) $$\prod_{i=1}^{n}\left(1+\delta'_i\right)=1+1,01\,n\,\theta\,u\quad,\quad|\theta|\leq 1.$$

A hipótese $nu\leq 0,01$, que torna possível a aplicação de (1.67), certamente se verifica nos proble-

mas em que (2.56) será aplicada, já que nesses, de maneira geral, n e u representam, respectivamente, a ordem de uma matriz e a unidade relativa de arredondamento. Por esse motivo, por simplicidade, sempre que utilizarmos (2.57) não faremos referências a essa hipótese, supondo-a subentendida.

Descreveremos a determinação de

$$(2.58) \qquad s_n = fl(x_1 + x_2 + \cdots + x_n) \;;\; x_i \in F_\ell \;;\; i = 1, 2, \cdots, n.$$

O computador efetua um processo de "adições encaixantes" que, para $n = 4$, é representado por

$$fl(x_1 + x_2 + x_3 + x_4) = fl\Big(fl\Big(fl(x_1 + x_2) + x_3\Big) + x_4\Big).$$

No caso genérico, o computador determina, sequencialmente, as quantidades $s_1, s_2, \cdots, s_n$, definidas recursivamente por:

$$(2.59) \qquad \begin{cases} s_1 = x_1 \\ s_i = fl(s_{i-1} + x_i) = (s_{i-1} + x_i)(1 + \delta_i) \;;\; |\delta_i| \le u \;;\; i = 2, \cdots, n \end{cases}$$

Desenvolvendo as relações (2.59) obtemos:

$$(2.60) \qquad s_n = fl(x_1 + x_2 + \cdots + x_n) = x_1(1 + \varepsilon_1) + x_2(1 + \varepsilon_2) + \cdots + x_n(1 + \varepsilon_n)$$

$$(2.61) \qquad \begin{cases} 1 + \varepsilon_1 = (1 + \delta_2)(1 + \delta_3)\cdots(1 + \delta_n) \\ 1 + \varepsilon_i = (1 + \delta_i)(1 + \delta_{i+1})\cdots(1 + \delta_n) \;;\; i = 2, \cdots, n \end{cases}$$

Aplicando (2.57) a (2.61),

$$(2.62) \qquad \begin{cases} 1 + \varepsilon_1 = 1 + 1{,}01\,(n-1)\,\theta_1\,u \;;\; |\theta_1| \le 1 \\ 1 + \varepsilon_i = 1 + 1{,}01\,(n-i+1)\,\theta_i\,u \;;\; |\theta_i| \le 1 \;;\; i = 2, \cdots, n \end{cases}$$

Com a finalidade de evitar delimitações diferentes para ε_1 e ε_i $(i = 2, \cdots, n)$, podemos escrever (2.62) na forma:

$$(2.63) \qquad 1 + \varepsilon_i = 1 + 1{,}01\,(n-i+1)\,\theta_i\,u \;;\; |\theta_i| \le 1 \;;\; i = 1, 2, \cdots, n,$$

em que, para $n = 1$, a delimitação foi aumentada de $1{,}01(n-1)\theta_1 u$

para $1,01\ n\ \theta_1\ u$.

Embora o erro relativo originado no cálculo de $fl(x_1+x_2)$, em computadores que possuem acumulador de precisão dupla, seja sempre pequeno, não podemos afirmar que, necessariamente, o mesmo ocorre na determinação de $fl(x_1+x_2+\cdots+x_n)$. Para entender esse fato, basta considerar o erro relativo

$$|\varepsilon| = \left| \frac{fl(x_1+\cdots+x_n)-(x_1+\cdots+x_n)}{(x_1+\cdots+x_n)} \right| = \left| \frac{fl(x_1+\cdots+x_n)}{(x_1+\cdots+x_n)} - 1 \right| \tag{2.64}$$

Pode ocorrer que a soma aproximada seja nula, sem que a soma exata o seja, ou vice-versa, razão de nem sempre existir uma boa delimitação para (2.64). Assim, por exemplo, quando a soma em vírgula móvel é nula, sem que o mesmo ocorra com a soma exata, obtemos $|\varepsilon| = 1$ e, portanto, um erro de arredondamento que, em valor absoluto, é igual à soma exata.

Apesar das delimitações para os ε_i não dependerem da ordem em que a adição se efetue, a delimitação da soma depende de tal ordem. Essa delimitação é menor quando a adição é efetuada segundo a ordem crescente dos valores absolutos das parcelas porque, dessa maneira, a maior delimitação dos ε_i é associada ao menor x_i.

Exemplos

$$10^3(0,1418)+10^1(0,5368)+10^2(0,9465)+10^0(0,5812) \tag{2.65}$$

A soma exata dos números considerados em (2.65) é $10^3(0,2423992)$, cujo arredondamento para 4 dígitos fornece $10^3(0,2424)$, com $|\varepsilon| = 0,0033 \times 10^{-3}$.

Efetuando (2.65) na ordem dada obtemos:

$$s_2 = fl\left(10^3(0,1418)+10(0,5368)\right) = fl\left(10^3(0,1471680)\right)$$

$$s_2 = 10^3(0,1472), \quad |\delta_2| \simeq 0,21744 \times 10^{-3}.$$

$s_3 = fl\left(10^3(0,1472) + 10^2(0,9465)\right) = fl\left(10^3(0,24185000)\right)$
$s_3 = 10^3(0,2419)$; $|\delta_3| = 0,20674 \times 10^{-3}$.

$s_4 = fl\left(10^3(0,2419) + 10^0(0,5812)\right) = fl\left(10^3(0,24248120)\right)$
$s_4 = 10^3(0,2425)$; $|\delta_4| \simeq 0,077532 \times 10^{-3}$,

do que resulta uma soma aproximada igual a $10^3(0,2425)$, com $|\varepsilon| \simeq 0,41584 \times 10^{-3}$.

Efetuando (2.65) segundo a ordem crescente dos valores absolutos das parcelas temos:

$s'_2 = fl\left(10^0(0,5812) + 10^1(0,5368)\right) = fl\left(10^1(0,59492)\right)$
$s'_2 = 10^1(0,5949)$; $|\delta'_2| \simeq 0,033618 \times 10^{-3}$.

$s'_3 = fl\left(10^1(0,5949) + 10^2(0,9465)\right) = fl\left(10^3(0,100599)\right)$
$s'_3 = 10^3(0,1006)$; $|\delta'_3| = 0,009940 \times 10^{-3}$

$s'_4 = fl\left(10^3(0,1006) + 10^3(0,1418)\right) = fl\left(10^3(0,2424)\right)$
$s'_4 = 10^3(0,2424)$; $|\delta'_4| = 0$

A soma aproximada obtida nesse caso é $10^3(0,2424)$, com $|\varepsilon| = 0,0033 \times 10^{-3}$.

(2.66) $10^{-4}(0,4863) + 10^0(0,1436) + 10^{-4}(0,3985) + 10^{-4}(0,4599)$

Soma exata = 0,14373447

arredondamento soma exata = 0,1437 ; $|\varepsilon| \simeq 0,23982 \times 10^{-3}$

Efetuando (2.66) na ordem dada obtemos:

$s_2 = 10^0(0,1436)$; $|\delta_2| \simeq 0,33853 \times 10^{-3}$
$s_3 = 10^0(0,1436)$; $|\delta_3| \simeq 0,27743 \times 10^{-3}$
$s_4 = 10^0(0,1436)$; $|\delta_4| \simeq 0,32016 \times 10^{-3}$,

e a soma aproximada $10^0(0,1436)$ tem erro relativo $\simeq 0,93555 \times 10^{-3}$.

Quando é utilizada a ordem crescente dos valores absolutos das parcelas temos:

$s'_2 = 10^{-4}(0,8584)$; $|\delta'_2| = 0$
$s'_3 = 10^{-3}(0,1345)$; $|\delta'_3| \simeq 0,2231 \times 10^{-3}$
$s'_4 = 10^0(0,1437)$; $|\delta'_4| \simeq 0,24003 \times 10^{-3}$,

de modo que a soma aproximada, 0,1437, tem erro relativo $\simeq 0{,}23982 \times 10^{-3}$.

A escolha da ordem crescente dos valores absolutos das parcelas é particularmente vantajosa quando algumas parcelas têm valor absoluto muito menor que outras, como ocorre em (2.66).

(2.67) $\quad 10^3(0{,}1084) + 10^2(-0{,}9712) + 10(-0{,}9537) + 10(-0{,}5014)$

soma exata $= 10(-0{,}3271)$

Efetuando (2.67) na ordem dada temos:

$s_2 = 10^2(0{,}1128)$; $|\delta_2| = 0$

$s_3 = 10(0{,}1743)$; $|\delta_3| = 0$

$s_4 = 10(-0{,}3271)$; $|\delta_4| = 0$,

de modo que a soma obtida é exata.

Utilizando a ordem crescente dos valores absolutos das parcelas:

$s'_2 = 10^2(-0{,}1455)$; $|\delta'_2| = 0{,}068724 \times 10^{-3}$

$s'_3 = 10^3(-0{,}1117)$; $|\delta'_3| = 0{,}268649 \times 10^{-3}$

$s'_4 = 10(-0{,}3300)$; $|\delta'_4| = 0$,

e a soma aproximada obtida, $10(-0{,}3300)$ tem $|\varepsilon| = 8{,}86579 \times 10^{-3}$.

Embora forneça sempre a menor delimitação para o erro, a escolha da ordem crescente dos valores absolutos das parcelas não fornece, necessariamente, o menor erro, como pode ser observado em (2.67). É importante, porém, que se observe que esse exemplo é muito especial, tendo sido construído para evidenciar tal fato; o caso geral é dado por (2.65) e (2.66).

Multiplicação em computadores que dispõem de acumulador de precisão dupla.

Dados $x_1, x_2 \in F_t$, o produto aproximado $fl(x_1 \times x_2)$ é obtido da seguinte maneira: os expoentes b_1 e b_2 são adicionados, originando b, e o produto $a_1 a_2$ é realizado em precisão dupla; quando necessário, esse produto é normalizado, e o expoente b ajustado de maneira conveniente. Finalmente, a parte fracionária é arredondada para t dígitos.

Quando requerida, a normalização se efetua através de um único deslocamento à esquerda, pois, por (1.75)

$$N^{-2} \leq |a_1 a_2| < 1$$

Se x_1 ou x_2 (ou ambos) são nulos, então, é atribuído o valor zero para $fl(x_1 \times x_2)$.

Assim como ocorre na determinação de $fl(x_1 \pm x_2)$, somente um erro de arredondamento é cometido no cálculo de $fl(x_1 \times x_2)$; aquele originado pelo arredondamento, para t dígitos, do produto exato. Portanto:

$$fl(x_1 \times x_2) = x_1 x_2 (1+\varepsilon) \quad , \quad |\varepsilon| \leq \mu \quad , \tag{2.68}$$

em que $x_1 x_2$ representa o produto exato de x_1 e x_2.

Exemplos

$$fl\left(10^{-3}(0,7823) \times 10^{5}(0,5432)\right) = fl\left(10^{2}(0,42494536)\right) = 10^{2}(0,4249) \quad ; \quad |\varepsilon| \simeq 0,10674 \times 10^{-3}. \tag{2.69}$$

$$fl\left(10^{-2}(0,2351) \times 10^{4}(0,1634)\right) = fl\left(10^{-6}(0,03841534)\right) = fl\left(10^{-7}(0,38415340)\right) = 10^{-7}(0,3842) \; ; \; |\varepsilon| \simeq 0,12131 \times 10^{-3}. \tag{2.70}$$

Consideraremos, a seguir, multiplicações com mais de dois fatores. Descreveremos a determinação de

$$p_n = fl(x_1 \times x_2 \times \cdots \times x_n) \; ; \; x_i \in F_t \; ; \; i = 1, 2, \cdots, n \tag{2.71}$$

Assim como ocorre com adições de mais de duas parcelas, o computador realiza um processo de "multiplicações encaixantes" que, para $n = 4$ é representado por:

$$fl(x_1 \times x_2 \times \cdots \times x_n) = fl\left(fl\left(fl(x_1 \times x_2) \times x_3\right) \times x_4\right)$$

No caso genérico, o computador determina, sucessivamente, os valores $p_1, p_2, \cdots, p_n$, definidos recursivamente por:

$$\begin{cases} p_1 = x_1 \\ p_n = fl(p_{i-1} \times x_i) = p_{i-1} x_i (1+\delta_i) \; ; \; |\delta_i| \leq \mu \; ; \; i = 2, \cdots, n \end{cases} \tag{2.72}$$

Desenvolvendo as relações dadas em (2.72) obtemos:

$$(2.73) \qquad p_m = fl(x_1 \times x_2 \times \cdots \times x_m) = x_1 x_2 \cdots x_m (1+\varepsilon) ,$$

$$(2.74) \qquad 1+\varepsilon = (1+\delta_2)(1+\delta_3)\cdots(1+\delta_n) \; ; \; |\delta_i| \leq u , \; i = 2, \cdots, n$$

Aplicando (2.57) a (2.74), temos:

$$(2.75) \qquad 1+\varepsilon = 1 + 1{,}01\,(n-1)\,\theta\,u \; , \quad |\theta| \leq 1$$

De maneira geral, o produto aproximado tem erro relativo pequeno e, embora o seu valor dependa da ordem em que as multiplicações se efetuem, a delimitação (2.75) não depende da mesma.

Divisão em computadores que possuem acumulador de precisão dupla.

Enquanto na aritmética em vírgula fixa só podem ser efetuadas divisões $x_1 : x_2$ para as quais $x_1 < x_2$, na aritmética em vírgula flutuante a única divisão que não pode ser realizada é aquela para a qual $x_2 = 0$.

O quociente de x_1 por x_2 é determinado da maneira que se segue. O expoente b, do quociente, é tomado igual à diferença $b_1 - b_2$ e a_1 é colocado nos t dígitos mais significativos do acumulador. Se $|a_1| > |a_2|$, então, o número armazenado no acumulador é deslocado uma posição à direita e b aumentado de uma unidade. A seguir, o número do acumulador é dividido por a_2 e o quociente, que já deve satisfazer a (1.75), é arredondado para t dígitos.

Em virtude de no cálculo de $fl(x_1 : x_2)$ só ocorrer um arredondamento, o do quociente exato para t dígitos:

$$(2.76) \qquad fl(x_1 : x_2) = \frac{x_1}{x_2}\,(1+\varepsilon) \; ; \; |\varepsilon| \leq u \; ; \; x_2 \neq 0 \; ,$$

onde $\frac{x_1}{x_2}$ representa o quociente exato de x_1 por x_2.

Exemplos

$$(2.77) \qquad fl\left(10^5(0{,}5632) : 10^{-3}(0{,}1827)\right) = fl\left(10^{-2}(0{,}56320000 : 0{,}1827)\right) =$$

$$= fl\left(10^{-1}\left(0,05632000 : 0,1827\right)\right) = fl\left(10^{-1}\left(0,30826491\cdots\right)\right) =$$
$$= 10^{-1}(0,3083) \quad ; \quad |\varepsilon| \simeq 0,11381 \times 10^{-3}.$$

(2.78) $$fl\left(10^{6}(0,2643) : 10^{3}(0,9741)\right) = fl\left(10^{3}(0,26430000 : 0,9741)\right) =$$
$$= fl\left(10^{3}(0,27132737\cdots)\right) = 10^{3}(0,2713)$$
$$|\varepsilon| \simeq 0,10091 \times 10^{-3}.$$

A igualdade (2.76) pode ser escrita

(2.79) $$fl(x_1 : x_2) = x_1\, x_2^{-1}(1+\varepsilon) \; ; \; |\varepsilon| \leq u \; ; \; x_2 \neq 0.$$

Em virtude da aplicação de (2.68) a $fl\left(x_1 \times x_2^{-1}\right) = fl(x_1 : x_2)$ também fornecer o resultado (2.79), podemos determinar delimitações para o erro de expressões do tipo

(2.80) $$fl\left(\frac{x_1\, x_2 \cdots x_m}{y_1\, y_2 \cdots y_n}\right)$$

através da aplicação de (2.73) – (2.75). Assim procedendo, obtemos

(2.81) $$fl\left(\frac{x_1\, x_2 \cdots x_m}{y_1\, y_2 \cdots y_n}\right) = \frac{x_1\, x_2 \cdots x_m}{y_1\, y_2 \cdots y_n}(1+\varepsilon) \quad ,$$

(2.82) $$1+\varepsilon = 1 + 1,01\,(m+n-1)\,\theta\, u \quad , \quad |\theta| \leq 1$$

De maneira geral, o erro relativo cometido no cálculo de (2.80) é pequeno. Embora a aproximação determinada dependa da ordem de realização das multiplicações e divisões, a delimitação (2.82) não depende da mesma. A influência da ordem de realização das operações sobre a aproximação determinada é pequena para (2.71) e (2.80), sendo muito mais acentuada relativamente a (2.58); por esse motivo, tal ordem é, geralmente, considerada apenas em cálculos do último tipo.

Multiplicação escalar em computadores que dispõem de acumulador de precisão dupla.

Consideremos a determinação de

(2.83) $$e_n = fl\left(x_1 \times y_1 + x_2 \times y_2 + \cdots + x_n \times y_n\right) \; ; \; x_i, y_i \in F_\ell \; ; \; i = 1, 2, \cdots, n$$

A determinação de e_n é feita da seguinte maneira: em primeira etapa são calculados todos os produtos $x_i y_i$ que, em segunda etapa, são somados na ordem em que estão escritos em (2.83). Esse processo é descrito pelas relações:

$$\begin{cases} p_i = fl(x_i \times y_i) \quad , \quad i = 1, \cdots, n \\ s_1 = p_1 \quad \text{e} \quad s_i = fl(s_{i-1} + p_i) \, , \, i = 2, \cdots, n \end{cases} \tag{2.84}$$

Aplicando (2.68) e (2.48) a (2.84)

$$\begin{cases} p_i = x_i y_i (1+\delta_i) \; ; \; |\delta_i| \le u \; ; \; i = 1, 2, \cdots, n \\ s_1 = p_1 \; ; \; s_i = (s_{i-1} + p_i)(1+\eta_i) \, , \, |\eta_i| \le u \, , \, i = 2, \cdots, n \end{cases} \tag{2.85}$$

Desenvolvendo (2.85) obtemos:

$$e_n = fl(x_1 \times y_1 + x_2 \times y_2 + \cdots + x_n \times y_n) = \tag{2.86}$$

$$= x_1 y_1 (1+\varepsilon_1) + x_2 y_2 (1+\varepsilon_2) + \cdots + x_n y_n (1+\varepsilon_n)$$

$$\begin{cases} 1+\varepsilon_1 = (1+\delta_1)(1+\eta_2) \cdots (1+\eta_n) \\ 1+\varepsilon_i = (1+\delta_i)(1+\eta_i) \cdots (1+\eta_n) \quad , \; i = 2, \cdots, n \end{cases} \tag{2.87}$$

com $|\delta_i| \, , \, |\eta_i| \le u \; , \; i = 1, 2, \cdots, n$.

Aplicando (2.57) a (2.87) podemos, ainda, escrever:

$$\begin{cases} 1+\varepsilon_1 = 1 + 1{,}01\, n\, \theta_1 u \quad , \quad |\theta_1| \le 1 \\ 1+\varepsilon_i = 1 + 1{,}01\,(n-i+2)\,\theta_i u \, , \, |\theta_i| \le 1 \, , \, i = 2, \cdots, n \end{cases} \tag{2.88}$$

ou, para evitar delimitações diferentes para ε_1 e ε_i:

$$1+\varepsilon_i = 1 + 1{,}01\,(n-i+2)\,\theta_i u \; ; \; |\theta_i| \le 1 \; ; \; i = 1, 2, \cdots, n \tag{2.89}$$

Adição e subtração em computadores que dispõem de acumulador de precisão simples.

Nos computadores que não dispõem de acumulador de precisão dupla, as operações são realizadas de for-

ma que difere um pouco daquela descrita para os computadores que o possuem. Descreveremos apenas a adição e a subtração porque as principais diferenças ocorrem em relação a essas operações; a multiplicação e a divisão não sofrem grandes variações e o produto e quociente aproximados têm sempre erro relativo pequeno.

Na adição e subtração, realizadas através de acumulador de precisão simples, a existência de cancelamento dos dígitos localizados imediatamente à direita da vírgula não implica na nulidade do erro de arredondamento. Por outro lado, a soma e diferença aproximadas não possuem, necessariamente, erro relativo pequeno e as equações (2.48) e (2.49) são substituídas por

$$fl(x_1 \pm x_2) = x_1(1+\varepsilon_1) \pm x_2(1+\varepsilon_2), \text{ com} \tag{2.90}$$

$$|\varepsilon_1|, |\varepsilon_2| \leq \frac{1+N}{N}\, u \quad , \tag{2.91}$$

onde ε_1 e ε_2 são, em geral, diferentes, porém, possuem delimitações da ordem de grandeza de u.

Ilustraremos a realização da adição e subtração, em computadores que utilizam acumulador de precisão simples, através da resolução dos exemplos (2.50)-(2.55) de maneira análoga à efetuada por um computador desse tipo

$$fl(10^{-3}(0,1572) + 10^{4}(0,9813)) = fl(10^{4}(0,0000 + 0,9813)) = \tag{2.92}$$
$$= fl(10^{4}(0,9813)) = 10^{4}(0,9813).$$
$$|\varepsilon| \simeq 0,000022 \times 10^{-3}$$

$$fl(10^{6}(0,7513) + 10^{3}(0,4624)) = fl(10^{6}(0,7513 + 0,0000)) = \tag{2.93}$$
$$= fl(10^{6}(0,7513)) = 10^{6}(0,7513).$$
$$|\varepsilon| \simeq 0,050002 \times 10^{-3}.$$

(2.94) $fl\left(10^3(0,3561)+10^3(0,8225)\right) = fl\left(10^3(0,3561+0,8225)\right) =$
$= fl\left(10^3(1,1786)\right) = fl\left(10^4(0,11786)\right) = 10^4(0,1179)$.
$|\varepsilon| \simeq 0,33939 \times 10^{-3}$

(2.95) $fl\left(10^{-5}(0,4318)+10^{-5}(-0,4316)\right) = fl\left(10^{-5}(0,4318-0,4316)\right) =$
$= fl\left(10^{-5}(0,0002)\right) = fl\left(10^{-8}(0,2000)\right) = 10^{-8}(0,2000)$.
$|\varepsilon| = 0$

(2.96) $fl\left(10^4(0,1001)+10^3(-0,9941)\right) = fl\left(10^4(0,1001-0,0994)\right) =$
$= fl\left(10^4(0,0007)\right) = fl\left(10^4(0,7000)\right) = 10^4(0,7000)$.
$|\varepsilon| = 14,478 \times 10^{-3}$

(2.97) $fl\left(10^3(0,9652)+10^2(0,5828)\right) = fl\left(10^3(0,9652+0,0583)\right) =$
$= fl\left(10^3(1,0235)\right) = fl\left(10^4(0,10235)\right) = 10^4(0,1024)$.
$|\varepsilon| = 0,50807 \times 10^{-3}$.

Comparamos os resultados fornecidos por ambos os tipos de computadores, com e sem disponibilidade de acumulados de precisão simples.

Os resultados obtidos em (2.92), (2.94) e (2.95) não diferem daqueles fornecidos em, respectivamente, (2.50), (2.52) e (2.53), porque, para a realização desses últimos, não foi utilizada a segunda parte do acumulador de precisão dupla.

Em (2.93) a mantissa 0,4624 foi deslocada três posições para a direita e, então, arredondada. A soma aproximada coincidiu com a aproximação determinada em (2.31)

Em (2.96) a mantissa (−0,9941) foi deslocada uma posição à direita e, a seguir, arredondada. A soma aproximada difere daquela obtida em (2.54) e tem erro relativo considerável. Esse exemplo, além de mostrar que o erro relativo da aproximação determinada não é, necessariamente, $\leq u$, ilustra o fato de que, quando são utilizados acumuladores de precisão simples, a existência de

cancelamento dos dígitos colocados à direita da vírgula não implica, obrigatoriamente, na nulidade do erro de arredondamento.

Em (2.97) a mantissa 0,5828 foi deslocada uma posição à direita e arredondada. Esse arredondamento influenciou o arredondamento da soma exata, ocasionando um resultado diferente daquele apresentado em (2.55).

Veremos, a seguir, que (2.90) e (2.91) se verificam não só nos exemplos considerados como, também, de maneira geral.

Em (2.92), (2.94) e (2.95), como a segunda parte do acumulador não foi utilizada para a resolução de (2.50), (2.52) e (2.53), respectivamente, então, (2.90) vale para $\varepsilon_1 = \varepsilon_2 = \varepsilon$ e, como $|\varepsilon| \leq u$, então, (2.91) também vale, pois $\varepsilon_1 = \varepsilon_2 = \varepsilon \leq u \leq \frac{1+N}{N} u$.

Em (2.93) só foi cometido um erro, aquele ocasionado pelo arredondamento da mantissa 0,4624, de modo que vale

$$(2.98) \qquad fl(x_1 + x_2) = x_1 + x_2(1+\varepsilon) \quad , \quad |\varepsilon| \leq u$$

e, portanto, (2.90) e (2.91) se verificam.

Em (2.96) também só foi efetuado um arredondamento, o da mantissa 0,9941, de modo que (2.98) vale também para esse caso.

Em (2.97) ocorreram dois arredondamentos: o da mantissa 0,5828 e o da soma 0,10235. Esses erros de arredondamento satisfazem:

$$(2.99) \qquad |\eta_2| \leq N^b \left(\frac{1}{2}\right) N^{-(t+1)}$$

$$(2.100) \qquad |\eta_3| \leq N^b \left(\frac{1}{2}\right) N^{-t} \quad ,$$

em que, como convencionado, $x = a \cdot N^b$ representa a soma das parcelas consideradas e, portanto,

$$(2.101) \qquad |x_1 + x_2| = N^b |a| \geq N^b \cdot N^{-1} = N^{b-1}$$

Em (2.99) e (2.100) o fator $\frac{1}{2}$ foi colocado entre parênteses porque aparece no caso de utilização de arredondamento propriamente dito e inexiste no caso de arredondamento por corte. Procuramos, dessa maneira, tornar válidas, para ambos os tipos de arredondamento, as considerações que estamos fazendo relativamente a (2.97).

Por (2.99), (2.100), (2.101) e (2.46) temos:

$$(2.102)\qquad |\eta_2+\eta_3| \le |\eta_2|+|\eta_3| \le \left(\frac{1}{2}\right) N^{b-t} N^{-1} + \left(\frac{1}{2}\right) N^{b-t}$$

$$\le \frac{1+N}{N}\left(\frac{1}{2}\right) N^{b-t} = \frac{1+N}{N}\left(\frac{1}{2}\right) N^{1-t}\, N^{b-1}$$

$$\le \frac{1+N}{N}\left(\frac{1}{2}\right) N^{1-t}\,|x_1+x_2| = \frac{1+N}{N}\,u\,|x_1+x_2|$$

A relação (2.102) nos fornece (2.90) e (2.91).

Acumulação de somas e produtos escalares.

Quando consideramos a realização das operações aritméticas através de acumulador de precisão dupla, vimos que, para determinar $fl(x_1+x_2+x_3)$, por exemplo, o computador efetua $fl(x_1+x_2)$ temporariamente em precisão dupla e arredonda o valor obtido para precisão simples, antes de adicioná-lo a x_3. Por outro lado, no cálculo de $fl(x_1\times y_1 + x_2\times y_2 + x_3\times y_3)$, a máquina calcula os produtos parciais em precisão dupla, arredonda-os para precisão simples e, a seguir, os soma da maneira que acabamos de descrever para $fl(x_1+x_2+x_3)$. Às vezes, é mais interessante, no caso da multiplicação escalar, não arredondar os produtos parciais e, em ambos os casos, efetuar todas as adições no acumulador de precisão dupla e arredondar somente o resultado final. Essa maneira de operação, embora exija maior tempo de execução, apresenta várias vantagens, principalmente em alguns problemas de Álgebra Linear. Poucas vezes a acumulação em precisão dupla é efetuada de forma exata, visto que

os expoentes dos diversos termos podem diferir consideravelmente; porém, quando ocorrem, os erros de arredondamento ficam restritos aos últimos dígitos do segundo registro e são, portanto, múltiplos pequenos de $\frac{1}{2} N^{-2t}$ ou N^{-2t}. As operações acumuladas serão indicadas por fl_2.

Adição. Consideremos a determinação de $fl_2(x_1+\cdots+x_n)$. Cada x_i pode ter mantissa em precisão simples ou dupla. As somas parciais são efetuadas em precisão dupla e a soma final é arredondada para t dígitos. É razoável supor-se a disponibilidade de apenas um acumulador de precisão dupla (não quádrupla), de modo que as delimitações utilizadas nesse caso são as válidas para acumulador de precisão simples, em que t é substituído por $2t$. Procedendo como em (2.58), com $|\delta_i| \leq \frac{1+N}{N} N^{1-2t}$, como em (2.91), obtemos:

(2.103) $\quad x_1(1+\varepsilon_1) + x_2(1+\varepsilon_2) + \cdots + x_n(1+\varepsilon_n)$, com

(2.104)
$$\begin{cases} 1+\varepsilon_1 = 1 + 1{,}01\,(n-1)\,\theta_1\,\dfrac{1+N}{N}\,N^{1-2t}, & |\theta_1| \leq 1 \\ 1+\varepsilon_i = 1 + 1{,}01\,(n-i+1)\,\theta_i\,\dfrac{1+N}{N}\,N^{1-2t}; & |\theta_i| \leq 1;\ i=2,\cdots,n \end{cases}$$

ou ainda, para evitar delimitações diferentes para ε_1 e ε_i:

(2.105) $\quad 1+\varepsilon_i = 1 + 1{,}01(n-i+1)\,\theta_i\,\dfrac{1+N}{N}\,N^{1-2t}; \quad |\theta_i| \leq 1;\ i=1,2,\cdots,n$

O resultado apresentado em (2.103) é, finalmente, arredondado para t dígitos originando:

(2.106) $\quad fl_2(x_1+x_2+\cdots+x_n) = \left[x_1(1+\varepsilon_1)+\cdots+x_n(1+\varepsilon_n)\right](1+\varepsilon)$,

(2.107) $\quad 1+\varepsilon = 1+\theta_0 u; \quad |\theta_0| \leq 1$

O fator associado a x_i é $(1+\varepsilon_i)(1+\varepsilon)$ que, em virtude de N^{1-2t} ser, em geral, muito pequeno em relação a N^{1-t}, tem valor comparável a $(1+\varepsilon)$.

Multiplicação escalar. Consideremos a determinação de $fl(x_1 \times y_1 + \cdots + x_n \times y_n)$, em que os x_i e y_i são números em precisão simples. Procedendo de maneira análoga à utilizada para a acumulação de somas, temos:

$$(2.108) \qquad fl_2(x_1 \times y_1 + \cdots + x_n \times y_n) = \left[x_1 y_1 (1+\varepsilon_1) + \cdots + x_n y_n (1+\varepsilon_n)\right](1+\varepsilon)$$

$$(2.109) \qquad \begin{cases} 1+\varepsilon_1 = 1 + 1{,}01\,(n-1)\,\theta_1\,\dfrac{1+N}{N}\,N^{1-2t}, & |\theta_1| \le 1 \\ 1+\varepsilon_i = 1 + 1{,}01\,(n-i+1)\,\theta_i\,\dfrac{1+N}{N}\,N^{1-2t}; & |\theta_i| \le 1, \quad i = 2, \cdots, n \end{cases}$$

$$(2.110) \qquad 1+\varepsilon = 1 + \theta_0\,\mu, \quad |\theta_0| \le 1$$

Para evitar delimitações diferentes para ε_1 e ε_i, podemos escrever (2.109) na forma:

$$(2.111) \qquad 1+\varepsilon_i = 1 + 1{,}01\,(n-i+1)\,\theta_i\,\frac{1+N}{N}\,N^{1-2t}, \quad \text{para } i = 1, 2, \cdots, n \text{ e } |\theta_i| \le 1.$$

Observe-se que em (2.109) e (2.88) há diferença nos fatores que envolvem n; isso ocorre porque na acumulação de produtos escalares os produtos parciais são efetuados sem erro, visto que o produto de dois números t-digitais pode ser representado exatamente por um número $2t$-digital.

A operação $fl_2\left(\dfrac{x_1 \times y_1 + \cdots + x_n \times y_n}{z}\right)$, em que x_i, y_i e z são números em precisão simples, é efetuada de maneira bastante satisfatória. O numerador é acumulado em precisão dupla e esse resultado $2t$-digital é dividido por z; finalmente, o quociente obtido é arredondado para t dígitos, fornecendo:

$$(2.112) \qquad fl_2\left(\frac{x_1 \times y_1 + \cdots + x_n \times y_n}{z}\right) = \left[\frac{x_1 y_1 (1+\varepsilon_1) + \cdots + x_n y_n (1+\varepsilon_n)}{z}\right](1+\varepsilon)$$

em que ε_i e ε satisfazem (2.109), (2.111) e (2.110), respectivamente.

2.5. Conclusões

O arredondamento por corte pode ocasionar erros maiores do que os originados pelo arredondamento propriamente dito, como pode ser observado em (2.10) e (2.11), (2.34) e (2.35), (2.40) e (2.41). Por outro lado, o corte pode introduzir "tendências" em cálculos muito longos, como acontece, por exemplo, na adição de uma grande quantidade de números positivos, em que, em virtude dos erros das somas parciais serem todos de mesmo sinal, pode ser introduzido um erro total considerável após n etapas. A utilização do arredondamento propriamente dito, para esse caso citado, geralmente fornece um erro total menor, após n etapas, já que normalmente ocorre cancelamento de erros parciais de sinais diferentes. Entretanto, apesar do arredondamento propriamente dito apresentar essas vantagens sobre o corte, esse último é bastante utilizado nos computadores que se encontram em uso atualmente, sendo encontrado, por exemplo, no FORTRAN II e no FORTRAN IV dos computadores IBM-7090 e IBM-1130.

Como pode ser observado em (2.6), (2.8) e (2.28), os sistemas em vírgula flutuante possibilitam a representação de uma quantidade de números muito maior do que aquela permitida pelos sistemas em vírgula fixa. Além disso, nos primeiros as operações aritméticas são realizadas de forma mais satisfatória do que nos segundos, a exemplo do que ocorre com a divisão. Principalmente por esses motivos os sistemas em vírgula flutuante são os mais utilizados atualmente.

As operações aritméticas são efetuadas de forma mais satisfatória em computadores que dispõem de acumulador de precisão dupla do que naqueles que não o possuem, como pode ser observado em (2.25) e

(2.26) e em (2.48), (2.49) e (2.90), (2.91).

Em virtude de grande parte dos computadores em uso atualmente utilizar sistema de representação em vírgula flutuante e dispor de acumulador de precisão dupla, utilizaremos, nos capítulos subseqüentes, a aritmética de vírgula flutuante e as delimitações de erros válidas para o caso de disponibilidade de tal acumulador. Acreditamos que essa atitude não prejudicará os resultados apresentados neste trabalho, os quais, com pequenas adaptações, também poderão ser aplicados a computadores que trabalham com acumulador de precisão simples e/ou aritmética de vírgula fixa

Como pode ser concluído de (2.48), (2.49), (2.68) e (2.76), em computadores que utilizam sistema em vírgula flutuante e dispõem de acumulador de precisão dupla verifica-se

$$fl(x * y) = (x * y)(1 + \varepsilon) \quad , \qquad |\varepsilon| \leq u \, , \tag{2.113}$$

em que $*$ representa uma operação aritmética fundamental e y é suposto diferente de zero no caso de $*$ indicar a divisão. A igualdade (2.113) foi obtida a partir de conclusões válidas para o erro relativo definido em relação ao valor verdadeiro; quando esse erro é definido em relação ao valor aproximado, obtém-se, a partir de (2.40) e (2.41), em correspondência a (2.113):

$$fl(x * y) = \frac{x * y}{1 + \varepsilon'} \quad , \qquad |\varepsilon'| \leq u \quad , \tag{2.114}$$

com $y \neq 0$ no caso em que $*$ representa a divisão.

3- Análises de erros

3.1- Sensibilidade

3.1.1- Problemas mal condicionados

Consideremos o uso de um computador para a resolução numérica de um problema e suponhamos que o mesmo trabalhe com palavras de comprimento t. A não ser que os dados sejam definidos exatamente e possam ser representados de maneira exata por t dígitos, o computador inicia a resolução com o que podemos chamar aproximação t-digital do problema verdadeiro. Assim, ao fazer uma análise de erros deve-se, necessariamente, levar em consideração os efeitos ocasionados, nas soluções, por perturbações nos dados. Um problema é dito mal condicionado quando possui soluções muito sensíveis a perturbações em seus dados ou parâmetros, ou seja, quando pequenas perturbações relativas nesses últimos causam, comparativamente, grandes variações nas primeiras. Em caso contrário, ou seja, quando as soluções são pouco sensíveis a perturbações nos parâmetros, o problema é dito bem condicionado.

3.1.2- Números de condição

São valores associados a um problema, de acordo com uma certa regra, que dão informações quanto à sensibilidade das soluções do mesmo, relativamente a perturbações nos dados. Veremos, a seguir, uma formalização desse conceito.

Sejam E e F espaços normados. Seja $x_0 \in E$ e seja $\varepsilon > 0$. Seja G o conjunto das aplicações de E em F,

determinadas por uma certa classe de processos numéricos que relacionam dados e resultados. Seja $g \in G$. Define-se como <u>número de condição</u> de g a um número K, associado a g de acordo com uma certa regra, que satisfaz a:

$$\|x - x_0\| < \varepsilon \Rightarrow \|g(x) - g(x_0)\| \leq K \|x - x_0\| .$$

Dados $g_1, g_2 \in G$, com números de condição K_1 e K_2, respectivamente, g_1 é dita <u>melhor condicionada</u> que g_2 se $g_1 < g_2$.

Às vezes, também é chamado número de condição o número K definido pelo limite, se existir,

$$K = \lim_{x \to x_0} \frac{\|g(x) - g(x_0)\|}{\|x - x_0\|}$$

Quando $E = R^n$ e $F = R^m$, também são definidos números de condição relativos à variação de apenas algumas componentes dos dados ou referentes a apenas algumas componentes dos resultados. Assim, por exemplo, se $g: R^n \to R^m$, e se $y = g(x)$, $y \in R^m$, $\forall x \in R^n$, são frequentemente utilizadas como números de condição as quantidades K_{rs}, K_j, K'_j e K''_j, definidas por:

$$K_{rs} = \left| \frac{\partial y_r}{\partial x_s} \right| \quad ; \; r = 1, \cdots, m \quad ; \; s = 1, \cdots, n$$

$$\left(\sum_{j=1}^{m} \delta y_j^2 \right)^{\frac{1}{2}} \leq K \left(\sum_{i=1}^{n} \delta x_i^2 \right)^{\frac{1}{2}}$$

$$|\delta y_j| \leq K_j \left(\sum_{i=1}^{n} \delta x_i^2 \right)^{\frac{1}{2}}$$

$$\left| \frac{\delta y_j}{y_j} \right| \leq K'_j \left[\sum_{i=1}^{n} \left(\frac{\delta x_i}{x_i} \right)^2 \right]^{\frac{1}{2}}$$

$$\frac{\delta y_j}{\left(\sum y_j^2\right)^{\frac{1}{2}}} \leq K_j'' \left(\frac{\sum_{i=1}^{m} \delta x_i^2}{\sum_{i=1}^{m} x_i^2} \right)^{\frac{1}{2}} ,$$

onde ∂ é o símbolo usual de derivada e δ representa uma variação.

Números de condição de matrizes

Seja $A \in M_n(R)$, A não singular. Como diversos problemas relacionados com A são associados a um mesmo número de condição K, costuma-se atribuir a K a denominação de "número de condição da matriz A". Embora os números de condição sejam associados a problemas, esse procedimento facilita bastante o estudo da sensibilidade possibilitando, por exemplo, um estudo geral das propriedades de K, estudo esse que pode depois ser aplicado quando da consideração dos diversos problemas aos quais K é associado. Adotaremos essa atitude porque os problemas considerados no capítulo 4 são todos associados a um mesmo número de condição.

(3.1) Definição - Seja $A \in M_n(R)$, A não singular. Denomina-se número de condição de A, e se indica por $\text{cond}_k(A)$, a:

$$\text{cond}_k(A) = \|A\|_k \, \|A^{-1}\|_k$$

Como pode ser observado em (3.1), o número de condição de uma matriz depende da particular norma adotada. A norma $\| \ \|_2$ é bastante utilizada e o correspondente número de condição, denominado número de condição espectral, pode ser expresso por:

$$\text{cond}_2(A) = \|A\|_2 \, \|A^{-1}\|_2 = \frac{\mu_1}{\mu_n} \quad (3.2)$$

onde μ_1 e μ_n representam o maior e o menor valor singular de A, respectivamente, como visto em 1.5. Portanto, $\text{cond}_2(A)$ dá a medida da máxima distorção que uma aplicação linear com matriz A produz na esfera unitária.

O número de condição definido em (3.1) goza das propriedades:

$$\operatorname{cond}_k(A) \geq 1$$

$$\operatorname{cond}_k(A) = \operatorname{cond}_k(A^{-1})$$

$$\operatorname{cond}_k(A) = \operatorname{cond}_k(\alpha A) \quad ,$$

sendo α um escalar e $\| \ \|_k$ uma norma subordinada ou a norma euclideana.

As propriedades citadas são de verificação bastante simples, como pode ser observado a seguir:

$\|A\|_k \ \|A^{-1}\|_k \geq \|AA^{-1}\|_k = \|I\|_k = 1$, para normas subordinadas

$$\|A\|_E \ \|A^{-1}\|_E \geq \|AA^{-1}\|_E = \|I\|_E = n^{\frac{1}{2}} \geq 1$$

$$\operatorname{cond}_k(A) = \|A\|_k \ \|A^{-1}\|_k = \|A^{-1}\|_k \ \|A\|_k = \operatorname{cond}_k(A^{-1})$$

$$\operatorname{cond}_k(\alpha A) = \|\alpha A\|_k \ \|(\alpha A)^{-1}\|_k = |\alpha| \ \|A\|_k \left|\frac{1}{\alpha}\right| \ \|A^{-1}\|_k = \operatorname{cond}_k(A)$$

A dificuldade do cálculo efetivo de $\operatorname{cond}_k(A)$ está na determinação de $\|A^{-1}\|_k$. Entretanto, em diversos casos há possibilidade de majorar esse valor e, assim, determinar estimativas para $\operatorname{cond}_k(A)$. Podemos majorar $\|A^{-1}\|_k$, por exemplo, quando temos uma aproximação X de A^{-1} para a qual $F = I - AX$ verifica $\|F\|_k < 1$, já que, nesse caso,

$$A^{-1}F = A^{-1} - X$$

$$\|A^{-1}\|_k \ \|F\|_k \geq \|A^{-1}F\|_k = \|A^{-1} - X\|_k \geq \|A^{-1}\|_k - \|X\|_k$$

$$\|A^{-1}\|_k \leq \frac{\|X\|_k}{1 - \|F\|_k}$$

O valor de F pode ser calculado de forma bastante precisa no caso de disponibilidade de acumulação de produtos escalares.

3.1.3_Sensibilidade das soluções de sistemas de equações lineares

Seja o sistema de equações lineares :

(3.3) $Ax = b$, com $A \in M_m(R)$ e A não singular, onde x representa o vetor solução e b o vetor dos termos independentes. Como A tem inversa única, então, x é determinado univocamente por

(3.4) $$x = A^{-1}b$$

Estudaremos, a seguir, a sensibilidade dessa solução x relativamente a perturbações nos dados, elementos de A e b. Esse estudo é feito por partes : supõe-se, inicialmente, que A seja conhecida exatamente e que b seja afetada de erros; a seguir, supõe-se que b seja conhecido exatamente e A esteja sujeita a erros. Finalmente, supõe-se que ambos, A e b, sejam sujeitos a erros.

Sensibilidade da solução relativamente a perturbações em b.

Admitamos que, em (3.3), A seja conhecida exatamente e que b seja afetada de erros. Supondo que o vetor solução x se transforme em $x + \delta x$ quando b sofre uma perturbação δb, podemos escrever :

(3.5) $$A(x + \delta x) = b + \delta b$$

e, portanto,

(3.6) $$\delta x = A^{-1}\delta b$$

Sejam $\| \ \|_p$ e $\| \ \|_k$ normas de vetores e matrizes, respectivamente, consistentes. Adotando-as em (3.6) obtemos :

(3.7) $$\|\delta x\|_p \leq \|A^{-1}\|_k \, \|\delta b\|_p \ ,$$

sendo possível, para alguns vetores, a obtenção da igualdade.

Utilizando em (3.3) as mesmas normas adotadas

em (3.7) temos :

$$\|b\|_p \leq \|A\|_k \, \|x\|_p \tag{3.8}$$

Multiplicando (3.7) e (3.8) membro a membro:

$$\|\delta x\|_p \, \|b\|_p \leq \|A\|_k \, \|A^{-1}\|_k \, \|x\|_p \, \|\delta b\|_p \tag{3.9}$$

Suponhamos $b \neq \theta$ em (3.3). Então, em consequência da não singularidade de A, temos que $x \neq \theta$ e podemos dividir (3.9) por $\|b\|_p \, \|x\|_p$ obtendo

$$\frac{\|\delta x\|_p}{\|x\|_p} \leq \|A\|_k \, \|A^{-1}\|_k \, \frac{\|\delta b\|_p}{\|b\|_p} \tag{3.10}$$

Aplicando (3.1) a (3.10) temos :

$$\frac{\|\delta x\|_p}{\|x\|_p} \leq \operatorname{cond}_k(A) \, \frac{\|\delta b\|_p}{\|b\|_p} \, , \tag{3.11}$$

que nos permite concluir que o número de condição $\operatorname{cond}_k(A)$ fornece uma indicação da sensibilidade da solução de (3.3) relativamente a variações dos termos conhecidos.

Como, para determinados vetores, pode ser obtida igualdade em (3.6) e (3.8), então, conseqüentemente, isso também pode ocorrer em (3.11); portanto, a majoração fornecida por essa última expressão, para $\frac{\|\delta x\|_p}{\|x\|_p}$, é a melhor possível. Utilizaremos (1.52) para mostrar como é possível, para a norma 2, obter igualdade em (3.11). Para isso, consideremos as mudanças ortogonais de coordenadas citadas em 1.5. Relativamente ao novo sistema de coordenadas, (3.3) será escrito

$$D x' = b' \, , \tag{3.12}$$

ou, ainda,

$$(3.13)\qquad \begin{cases} \mu_1 x'_1 = b'_1 \\ \mu_2 x'_2 = b'_2 \\ \vdots \qquad\quad \vdots \\ \mu_n x'_n = b'_n \end{cases}$$

e a equação $A\,\delta x = \delta b$, dos erros,

$$(3.14)\qquad D\,\delta x' = \delta b' \quad , \text{ ou,}$$

$$(3.15)\qquad \begin{cases} \mu_1 \delta x'_1 = \delta b'_1 \\ \mu_2 \delta x'_2 = \delta b'_2 \\ \vdots \qquad\quad \vdots \\ \mu_n \delta x'_n = \delta b'_n \end{cases}$$

Como A é não singular, então, por (1.59) e (1.60):

$$\|A\|_2 = \|D\|_2 = \mu_1 \quad \text{e} \quad \|A^{-1}\|_2 = \|D^{-1}\|_2 = \mu_n^{-1}.$$

Seja $b' = (\beta, 0, 0, \cdots, 0)^T$, onde $b'_1 = \beta$ e $b'_i = 0$ $(i = 2, \cdots, n)$. Então, por (3.13), $x' = (\mu_1^{-1}\beta, 0, 0, \cdots, 0)^T$, de modo que é obtida a igualdade em (3.8), para qualquer constante β não nula.

Seja $\delta b' = (0, 0, \cdots, 0, \gamma)^T$, onde $\delta b'_i = 0 (i = 1, \cdots, n-1)$ e $\delta b'_n = \gamma$. Então, por (3.15), $\delta x' = (0, 0, \cdots, \mu_n^{-1}\gamma)^T$ e, para qualquer constante γ não nula, é obtida a igualdade em (3.7).

Portanto, a igualdade em (3.11) pode ser obtida quando b está em uma direção que recebe a maior amplificação por A (e portanto a menor amplificação por A^{-1}) e δb está em uma direção que sofre a menor amplificação por A (e portanto a maior amplificação por A^{-1}). Essas duas direções são, necessariamente, ortogonais, já que $\mu_1 \neq \mu_n$.

Sensibilidade da solução relativamente a perturbações em A.

Suponhamos que em (3.3) b seja conhecido exatamente e A seja afetada de erros. Admitindo que x se transforme em $x+\delta x$, quando A sofre a perturbação δA, temos:

(3.16) $$(A+\delta A)(x+\delta x) = b$$ e, portanto,

(3.17) $$\delta x = -(A+\delta A)^{-1}\,\delta A\; x$$

Em virtude da não singularidade de A não implicar, necessariamente, na não singularidade de $A+\delta A$, procuraremos estabelecer condições de existência para (3.17). Seja

(3.18) $$\| A^{-1}\delta A \|_k < 1,$$

onde $\| \ \|_k$ é uma norma subordinada ou a norma euclideana e consideremos a identidade:

(3.19) $$A+\delta A = A\,(I + A^{-1}\delta A)$$

Então, por (1.43), podemos afirmar que $A+\delta A$ é não singular em todos os casos que (3.18) se verifique. Admitamos, então, no que se segue, que (3.18) seja verificada.

Em conseqüência de (3.17) e (3.19), temos:

(3.20) $$\delta x = -(I + A^{-1}\delta A)^{-1} A^{-1}\,\delta A\; x$$

Aplicando a (3.20) uma norma $\| \ \|_p$ compatível com $\| \ \|_k$, e utilizando (1.44)

(3.21) $$\|\delta x\|_p \leq \frac{\|A^{-1}\delta A\|_k \, \|x\|_p}{1-\|A^{-1}\delta A\|_k} \leq \frac{\|A^{-1}\|_k \, \|\delta A\|_k \, \|x\|_p}{1-\|A^{-1}\|_k \, \|\delta A\|_k},$$

desde que

$$\|A^{-1}\|_k \, \|\delta A\|_k < 1.$$

Finalmente, a aplicação de (3.1) a (3.21) fornece:

$$(3.22)\qquad \frac{\|\delta x\|_p}{\|x\|_p} \leq \frac{\operatorname{cond}_k(A)\,\dfrac{\|\delta A\|_k}{\|A\|_k}}{1 - \operatorname{cond}_k(A)\,\dfrac{\|\delta A\|_k}{\|A\|_k}}$$

Quando consideramos o erro relativo de x definido em relação ao valor aproximado $x+\delta x$, obtemos uma expressão correspondente a (3.22) que é mais simples e análoga a (3.11). Vamos determiná-la a seguir. Considerando o valor de δx obtido de (3.16) e (3.4), dado por

$$(3.23)\qquad \delta x = \left[(A+\delta A)^{-1} - A^{-1}\right] b \quad ,$$

e substituindo B por $A+\delta A$ na identidade

$$(3.24)\qquad B^{-1} - A^{-1} = A^{-1}(A-B)B^{-1} \quad , \quad \text{obtemos}$$

$$(3.25)\qquad (A+\delta A)^{-1} - A^{-1} = -A^{-1}\delta A\,(A+\delta A)^{-1}$$

Admitindo a verificação de (3.18), para garantir a não singularidade de $(A+\delta A)$, substituindo (3.25) em (3.23) e utilizando (3.16), obtemos:

$$(3.26)\qquad \delta x = -A^{-1}\delta A\,(A+\delta A)^{-1} b = -A^{-1}\delta A(x+\delta x)$$

Utilizando normas compatíveis em (3.26), temos:

$$\|\delta x\|_p \leq \|A^{-1}\|_k \|\delta A\|_k \|x+\delta x\|_p$$

Finalmente, supondo $x+\delta x \neq 0$ e utilizando (3.1), obtemos a expressão procurada

$$(3.27)\qquad \frac{\|\delta x\|_p}{\|x+\delta x\|_p} \leq \operatorname{cond}_k(A)\,\frac{\|\delta A\|_k}{\|A\|_k}$$

Sensibilidade da solução relativamente a perturbações em A e b.

Suponhamos que A e b sejam afetados de erros e que o vetor solução x se transforme em $x+\delta x$ quando

A e b sofrem, respectivamente, as perturbações δA e δb. Então,

$$(A+\delta A)(x+\delta x) = b+\delta b \qquad (3.28)$$

e, portanto,

$$\delta x = (A+\delta A)^{-1}\left[b+\delta b-(A+\delta A)x\right] = (A+\delta A)^{-1}(\delta b-\delta A\, x) \qquad (3.29)$$

Admitindo (3.18) como hipótese e utilizando (3.19):

$$\delta x = (I+A^{-1}\delta A)^{-1}(A^{-1}\delta b - A^{-1}\delta A\, x) \qquad (3.30)$$

Utilizando normas consistentes em (3.30), por (1.44)

$$\|\delta x\|_p \leq \frac{\|A^{-1}\|_k \|\delta b\|_k + \|A^{-1}\|_k \|\delta A\|_k \|x\|_p}{1-\|A^{-1}\|_k \|\delta A\|_k}, \qquad (3.31)$$

desde que $\|A^{-1}\|_k \|\delta A\|_k < 1$.

Aplicando a (3.3) as normas adotadas em (3.31):

$$\|A\|_k \|x\|_p \geq \|b\|_p \qquad (3.32)$$

Dividindo (3.31) e (3.32) membro a membro

$$\frac{\|\delta x\|_p}{\|A\|_k \|x\|_p} \leq \frac{\|A^{-1}\|_k\left(\frac{\|\delta b\|_p}{\|b\|_p}+\frac{\|\delta A\|_k\|x\|_p}{\|b\|_p}\right)}{1-\|A^{-1}\|_k\|\delta A\|_k} \qquad (3.33)$$

Finalmente, em virtude de (3.1) e (3.4)

$$\frac{\|\delta x\|_p}{\|x\|_p} \leq \frac{\operatorname{cond}_k(A)\left(\frac{\|\delta b\|_p}{\|b\|_p}+\operatorname{cond}_k(A)\frac{\|\delta A\|_k}{\|A\|_k}\right)}{1-\operatorname{cond}_k(A)\frac{\|\delta A\|_k}{\|A\|_k}} \qquad (3.34)$$

3.1.4 - Sobre o mal condicionamento de sistemas de equações lineares

Existem algumas concepções relativas ao mal condicionamento de sistemas de equações lineares que, embora difundidas

na prática, não são verdadeiras. Consideraremos, a seguir, algumas dessas idéias e mostraremos, através de exemplos, a não validade das mesmas.

Comumente, na prática, são considerados mal condicionados os sistemas de equações lineares cujas matrizes possuem ordem muito elevada ou têm determinantes pequenos. O exemplo que se segue mostra que esse fato não é verdadeiro. Consideremos um sistema cuja matriz A possua valores singulares $\mu_1 = \mu_2 = \cdots = \mu_n = 10^{-30}$. Então, por (3.2), (1.61) e (3.11), temos

$$\text{cond}_2(A) = \mu_1 \mu_n^{-1} = 1$$

$$|\det A| = 10^{-30n}$$

$$\frac{\|\delta x\|_2}{\|x\|_2} = \frac{\|\delta b\|_2}{\|b\|_2}$$

Portanto, $\text{cond}_2(A) = 1$, qualquer que seja n, ordem de A. Por outro lado, embora o determinante de A seja um número bastante pequeno, as perturbações da solução são iguais às do vetor dos termos independentes. Na realidade, a pequena grandeza de $\det A$ só pode ser relacionada ao mal condicionamento do sistema em algumas circunstâncias especiais. Assim, por exemplo, isso pode ser feito para uma matriz A de ordem n fixa desde que, de alguma maneira, multipliquemos A por fatores de escala que façam com que μ_1 permaneça fixo; se μ_1 e n são fixos, então, se $\det A \to 0$, necessariamente $\mu_n \to 0$ e $\text{cond}(A) \to \infty$. Entretanto, essa relação é muito fraca, como pode ser observado para $n = 101$, $\mu_1 = 1$ e $\mu_2 = \mu_3 = \cdots = \mu_n = 10^{-1}$. Embora $\det A = \mu_1 (\mu_2)^{n-1} = 10^{-100}$, temos $\text{cond}_2(A) = \mu_1 \mu_2^{-1} = 10$ e, portanto o sistema que possui A como matriz pode ser considerado bem condicionado.

Também é usual, na prática, considerar mal condicionados os sistemas lineares $Ax = b$ para os quais A^{-1} pos-

sui elementos grandes em comparação aos de A, o que pode originar grandes erros absolutos na solução. Essa concepção não é correta, já que o mal condicionamento do sistema depende de sua solução, como pode ser observado em (3.35). Os diferentes sistemas de equações lineares considerados nesse exemplo possuem uma mesma matriz A, cuja inversa A^{-1} tem elementos grandes comparados com os de A, porém, enquanto um deles é bem condicionado, o outro é extremamente mal condicionado. O que ocorre, na realidade, é que se A^{-1} tem elementos grandes, então, sempre é possível determinar um vetor b, dos termos independentes, que origina soluções pequenas e, portanto, erros relativos consideráveis. É óbvio que, se $\|A\|_k = 1$, então, se os elementos de A^{-1} são grandes em comparação aos de A, o sistema associado será mal condicionado, já que $\text{cond}_k(A) = \|A^{-1}\|_k$.

(3.35) Exemplo - Consideremos o sistema de equações lineares apresentado em (9, p. 75-77)

$$(3.36)\quad \begin{cases} \frac{1}{2}x_1 + \frac{1}{3}x_2 + \frac{1}{4}x_3 + \frac{1}{5}x_3 = b_1 + \varepsilon_1 \\ \frac{1}{3}x_1 + \frac{1}{4}x_2 + \frac{1}{5}x_3 + \frac{1}{6}x_3 = b_2 + \varepsilon_2 \\ \frac{1}{4}x_1 + \frac{1}{5}x_2 + \frac{1}{6}x_3 + \frac{1}{7}x_3 = b_3 + \varepsilon_3 \\ \frac{1}{5}x_1 + \frac{1}{6}x_2 + \frac{1}{7}x_3 + \frac{1}{8}x_3 = b_4 + \varepsilon_4 \end{cases},$$

em que os coeficientes das incógnitas são valores exatos, mas os termos independentes podem ser afetados de erros ε_i $(i = 1, \cdots, 4)$ para os quais é conhecida uma majoração η, ou seja, $|\varepsilon_i| \leq \eta$.

A inversa exata da matriz de (3.36) é

$$(3.37)\quad A^{-1} = \begin{pmatrix} 200 & -1200 & 2100 & -1120 \\ -1200 & 8100 & -15120 & 8400 \\ 2100 & -15120 & 29400 & -16800 \\ -1120 & 8400 & -16800 & 9800 \end{pmatrix}$$

A partir de A^{-1}, podemos determinar como varia a solução de (3.36), em relação aos diferentes $\mathcal{E}_i$, já que as quantidades δx_i podem ser descritas pelas equações:

$$A(x+\delta x) = b+\mathcal{E} \tag{3.38}$$

$$\delta x = A^{-1}\mathcal{E}\,, \tag{3.39}$$

em que $\mathcal{E} = (\mathcal{E}_1, \mathcal{E}_2, \mathcal{E}_3, \mathcal{E}_4)^T$.

Suponhamos que ocorra a pior combinação possível de grandezas e sinais, de modo que $|\delta x_i|$ seja o máximo possível. Nesse caso,

$$\begin{cases} |\delta x_1| = 4620\,\eta & |\delta x_2| = 32820\,\eta \\ |\delta x_3| = 63420\,\eta & |\delta x_4| = 36120\,\eta \end{cases} \tag{3.40}$$

Se tomarmos $\eta = 10^{-4}$, então, mesmo que todas as operações aritméticas necessárias à determinação da solução de (3.36) sejam realizadas exatamente, ainda assim os erros absolutos das componentes do vetor solução poderão ser

$$|\delta x_1| \simeq 0{,}46 \qquad |\delta x_2| \simeq 3{,}28 \qquad |\delta x_3| \simeq 6{,}34 \qquad |\delta x_4| \simeq 3{,}61 \tag{3.41}$$

Consideremos, inicialmente, que os b_i originem uma solução que seja da ordem de grandeza de

$$x_1 = 4000 \qquad x_2 = -32000 \qquad x_3 = 63000 \qquad x_4 = -36000, \tag{3.42}$$

de modo que os erros relativos originados pelos $\mathcal{E}_i$ serão, no máximo,

$$\begin{cases} \left|\dfrac{\delta x_1}{x_1}\right| = 0{,}1150 \times 10^{-3} & \left|\dfrac{\delta x_2}{x_2}\right| = 0{,}1025 \times 10^{-3} \\ \left|\dfrac{\delta x_3}{x_3}\right| = 0{,}1006 \times 10^{-3} & \left|\dfrac{\delta x_4}{x_4}\right| = 0{,}1003 \times 10^{-3}, \end{cases} \tag{3.43}$$

os quais, para muitos problemas, podem ser considerados pequenos. Uma solução de ordem de grandeza de (3.42) pode ser obtida, por exemplo, com os seguintes termos independentes:

(3.44) $b_1 = 1 \qquad b_2 = -1 \qquad b_3 = 1 \qquad b_4 = -1$

Consideremos, a seguir, os termos independentes

(3.45) $b_1 = b_2 = b_3 = b_4 = 1$,

que são de ordem de grandeza comparável a (3.44). Nesse caso,

$$x_1 = -20 \qquad x_2 = 180 \qquad x_3 = -420 \qquad x_4 = 280 ,$$

de modo que, embora os erros absolutos máximos ainda sejam dados em (3.41), os erros relativos são bem maiores:

(3.46)
$$\begin{cases} \left|\dfrac{\delta x_1}{x_1}\right| = 23{,}00 \times 10^{-3} & \left|\dfrac{\delta x_2}{x_2}\right| \simeq 18{,}22 \times 10^{-3} \\ \left|\dfrac{\delta x_3}{x_3}\right| \simeq 15{,}10 \times 10^{-3} & \left|\dfrac{\delta x_4}{x_4}\right| \simeq 12{,}89 \times 10^{-3} \end{cases}$$

Finalmente, consideremos os termos independentes

(3.47) $b_1 = 1 \qquad b_2 = 0{,}7227 \qquad b_3 = 0{,}5697 \qquad b_4 = 0{,}4714 ,$

que também são de ordem de grandeza comparável a (3.44) e (3.45) e fornecem:

$$x_1 = 1{,}162 \qquad x_2 = -0{,}234 \qquad x_3 = 2{,}436 \qquad x_4 = -0{,}560.$$

Novamente os erros absolutos da solução são dados por (3.41), porém, os erros relativos são muito maiores que aqueles apresentados em (3.43) e (3.46), sendo dados por:

(3.48)
$$\begin{cases} \left|\dfrac{\delta x_1}{x_1}\right| = 395{,}87 \times 10^{-3} & \left|\dfrac{\delta x_2}{x_2}\right| = 1407{,}1 \times 10^{-3} \\ \left|\dfrac{\delta x_3}{x_3}\right| = 2602{,}6 \times 10^{-3} & \left|\dfrac{\delta x_4}{x_4}\right| = 6446{,}4 \times 10^{-3} \end{cases}$$

É lógico que, nesse último caso, se $\eta = 10^{-4}$ representar a unidade absoluta de arredondamento do computador no

qual (3.36) deve ser resolvido, então, mesmo que todas as operações aritméticas utilizadas nessa resolução sejam efetuadas exatamente, o que é difícil de ocorrer, ainda assim os resultados obtidos não serão satisfatórios.

Concluindo, podemos afirmar que, embora relacionados com uma mesma matriz, os problemas obtidos pela substituição de (3.44), (3.45) e (3.47) em (3.36) possuem soluções muito diferentes entre si quanto à sensibilidade a perturbações dos termos independentes. O problema resultante da associação de (3.44) e (3.36) é bem condicionado, enquanto aquele obtido de (3.47) e (3.36) é extremamente mal condicionado.

3.2 - Análises progressiva e regressiva

A análise de erros de um processo numérico tem por finalidade determinar uma delimitação para a grandeza do erro do resultado final. Neste trabalho vamos tratar somente dos de arredondamento e, dessa forma, o erro do resultado final será causado pelo efeito propagado dos erros de arredondamento nos dados e nas operações aritméticas realizadas.

As principais formas de análise de erros de processos algébricos são: progressiva ou direta e regressiva ou inversa. Para caracterizá-las precisamos, inicialmente, fazer algumas considerações a respeito de processos numéricos.

Um processo numérico pode ser descrito por uma seqüência de equações do tipo:

$$x_{i+1} = g_i(x_1, x_2, \cdots, x_i) \quad , \quad i = 1, 2, \cdots, n-1 \quad , \tag{3.49}$$

que representam, em cada estágio i, o cálculo de novos valores em função de dados iniciais e dados previamente calculados. Nas equações (3.49), x_i não representa, necessariamente, um número,

mas pode, também, representar um conjunto de valores; nesse caso, x_i é dito o estado da i-ésima etapa de cálculo. Assim, por exemplo, x_1 representa o estado inicial.

Em particular, no caso de problemas algébricos a obtenção de x_{i+1} a partir dos x_i $(i = 1, 2, \cdots, n-1)$ envolve somente as operações aritméticas fundamentais.

Quando o processo numérico é executado no computador, são utilizadas, em substituição a $x_1, x_2, \cdots, x_i$, representações internas $\bar{x}_1, \bar{x}_2, \cdots, \bar{x}_i$, em virtude das limitações dessa máquina. Por outro lado, as operações aritméticas necessárias à determinação de x_{i+1} são, geralmente, resolvidas de forma aproximada, de modo que a função g_i é substituída por uma aproximação $\bar{g}_i$. Em virtude dessas aproximações, um resultado intermediário do processo pode ser representado da seguinte maneira:

$$\bar{x}_{i+1} = \bar{g}_i(\bar{x}_1, \bar{x}_2, \cdots, \bar{x}_i) \tag{3.50}$$

Análise progressiva ou direta

Neste tipo de análise procuramos, em cada estágio, estimar $|\bar{x}_{i+1} - x_{i+1}|$, a partir de estimativas prévias dos $|\bar{x}_j - x_j|$ $(j = 1, 2, \cdots i)$.

A análise progressiva é, em geral, extremamente complicada e laboriosa, em virtude de serem muito confusas as inter-relações entre os diversos erros. Além disso, as regras ordinárias da aritmética não podem ser aplicadas durante a sua realização, já que as operações aproximadas nem sempre satisfazem às mesmas propriedades das operações exatas, a exemplo da não associatividade da adição e multiplicação em vírgula flutuante, como vimos em 2.4.3

Análise regressiva ou inversa

Nesse tipo de análise de erros não nos preocu-

pamos em comparar $\bar{x}_{i+1}$ e x_{i+1} em cada estágio do processo. Em lugar disso, procuramos mostrar que o valor $\bar{x}_{i+1}$ pode ser obtido <u>exatamente</u> em função de valores iniciais perturbados

$$x_1 + \delta_i$$

e procuramos determinar estimativas para os δ_i. Assim procedendo, podemos escrever (3.50) na forma

$$\bar{x}_{i+1} = \phi_i \left(x_1 + \delta_i \right) ,$$

em que ϕ_i é obtida por composição de $g_1, g_2, \cdots, g_i$, e, de maneira geral, os δ_i não são determinados de forma unívoca. No final do processo obtemos:

$$\bar{x}_n = \phi_{n-1} \left(x_1 + \varepsilon \right) \quad , \quad \varepsilon = \delta_{n-1}$$

De maneira geral, a análise regressiva não determina estimativas para $|\bar{x}_n - x_n|$, preocupação fundamental da maioria das análises de erros. Nesses casos, entretanto, $|\bar{x}_n - x_n|$ pode ser estimada a partir de estimativas de

$$\left| \phi_{n-1} \left(x_1 + \varepsilon \right) - \phi_{n-1} \left(x_1 \right) \right| ,$$

através de um estudo da sensibilidade do processo.

As propriedades usuais da aritmética podem ser utilizadas durante a realização da análise inversa, já que são exatas as operações aritméticas utilizadas nessa análise.

Atualmente, a análise regressiva é a mais utilizada, não só por ser de aplicação mais simples, como também por possuir outras vantagens em relação à progressiva, a exemplo da avaliação da eficácia do processo numérico utilizado, como será visto a seguir. Por esse motivo, e por ser desse tipo a análise efetuada no capítulo 4, apresentaremos alguns exemplos que elucidam melhor a sua aplicação e destacam algumas de suas particularidades.

(3.51) <u>Exemplo</u> - Consideremos a determinação do valor numérico do polinômio

(3.52) $$P(x) = b_0 + b_1 x + \cdots + b_n x^n ,$$

para um valor x dado, através da seqüência de operações:

(3.53) $$\begin{cases} v_n = b_n \\ v_i = x\, v_{i+1} + b_i , \quad i = n-1, \cdots, 0 \end{cases} ,$$

de modo que v_0 é a solução procurada.

Suponhamos que os coeficientes de x não possam ser armazenados exatamente no computador e sejam substituídos por representações internas

(3.54) $$\bar{b}_i = b_i (1 + \delta_i) ; \ |\delta_i| \leq u , \quad i = 0, 1, \cdots, n .$$

Em virtude de (2.113) e (3.54), as expressões (3.53) são substituídas, no computador, por

(3.55) $$\begin{cases} v_n = \bar{b}_n = b_n (1 + \delta_n) , \quad |\delta_n| \leq u \\ v_i = fl(x v_{i+1} + \bar{b}_i) = x\, v_{i+1} (1+\alpha_i)(1+\beta_i) + b_i (1+\delta_i)(1+\beta_i), \\ \quad |\alpha_i|, |\beta_i|, |\delta_i| \leq u \end{cases}$$

Desenvolvendo as relações (3.55) obtemos:

(3.56) $$v_0 = b_0 (1 + \varepsilon_0) + b_1 (1 + \varepsilon_1) x + \cdots + b_n (1 + \varepsilon_n) x^n ,$$

(3.57) $$\begin{cases} 1 + \varepsilon_0 = (1+\delta_0)(1+\beta_0) \\ 1 + \varepsilon_1 = (1+\delta_1)(1+\beta_1)(1+\alpha_0)(1+\beta_0) \\ - \\ 1 + \varepsilon_i = (1+\delta_i)(1+\beta_i)(1+\alpha_{i-1})(1+\beta_{i-1}) \cdots (1+\alpha_1)(1+\beta_1)(1+\alpha_0)(1+\beta_0) \\ - \\ 1 + \varepsilon_n = (1+\delta_n)(1+\alpha_{n-1})(1+\beta_{n-1}) \cdots (1+\alpha_1)(1+\beta_1)(1+\alpha_0)(1+\beta_0) \end{cases}$$

Aplicando (2.57) a (3.57) obtemos:

(3.58) $$\begin{cases} 1+\varepsilon_i = 1+1{,}01\,(2i+2)\,\theta_i\,u \;;\; |\theta_i| \le 1 \;,\; i = 0,1,\cdots n-1 \\ 1+\varepsilon_n = 1+1{,}01\,(2n+1)\,\theta_n\,u \;;\; |\theta_n| \le 1 \,. \end{cases}$$

Com a finalidade de evitar delimitações diferentes para ε_i ($i = 0,1,\cdots,n-1$) e ε_n, podemos utilizar:

(3.59) $$1+\varepsilon_i = 1+1{,}01\,(2i+2)\,\theta_i\,u \;;\; |\theta_i| \le 1 \;,\; i = 0,1,\cdots,n$$

Podemos interpretar v_0, dado em (3.56), como o valor numérico exato do polinômio $\overline{P}(x)$ que possui os coeficientes:

(3.60) $$B_i = b_i(1+\varepsilon_i) = b_i + \delta b_i \,,$$

com ε_i dado em (3.59) e

(3.61) $$\delta b_i = b_i\,\varepsilon_i \quad , \text{ ou seja,}$$

(3.62) $$\overline{P}(x) = b_0 + \delta b_0 + (b_1 + \delta b_1)\,x + \cdots + (b_n + \delta b_n)\,x_n$$

Efetuando a análise de sensibilidade do polinômio à variação dos coeficientes, obtemos, por (3.52) e (3.62):

$$\delta P(x) = \overline{P}(x) - P(x) = \delta b_0 + \delta b_1\,x + \cdots + \delta b_n\,x^n$$

e, por (3.61),

$$|\delta P(x)| \le |\delta b_0| + |\delta b_1|\,|x| + \cdots + |\delta b_n|\,|x^n| \quad , \text{ com}$$

$$|\delta b_i| \le |b_i|\,|\varepsilon_i| \quad , \quad \varepsilon_i \text{ dado em (3.59)}$$

e, portanto, encontramos uma majoração para o erro cometido durante o processo de determinação do valor numérico.

(3.63) Exemplo - Em 4.2 veremos que, através da análise regressiva, pode-se mostrar que a solução aproximada $\bar{x}$, obtida pelo método de eliminação de Gauss com pivotamento para o sistema linear $Ax = b$ é a solução exata de

(3.64) $$(A+\delta A)\,\bar{x} = b \,, \quad \text{em que}$$

(3.65) $$\|\delta A\|_\infty \leq 1{,}01\left(n^3+3n^2\right) p \,\|A\|_\infty \, u \;,$$

onde n é a ordem de A, p é um certo número que pode ser calculado durante o processo de resolução e u é a unidade relativa de arredondamento.

Como pode ser observado em (3.64) e (3.65), nesse caso, assim como em (3.51), a análise regressiva não fornece estimativas para $\|\bar{x}-x\|_\infty$. Entretanto, essas estimativas podem ser obtidas através do estudo da sensibilidade do problema. Assim, de (3.27) e (3.65) temos:

(3.66) $$\|\bar{x}-x\|_\infty \leq \operatorname{cond}_\infty(A) \cdot 1{,}01\left(n^3+3n^2\right) p \,\|\bar{x}\|_\infty \, u$$

(3.67) Exemplo - Consideremos a solução de um sistema de equações lineares como visto em (3.63). Se após a aplicação da análise regressiva obtivermos $\|\delta A\| < \varepsilon$, com ε pré-estabelecido, e se soubermos previamente que os elementos de A estão sujeitos a um erro maior que ε, então, não teremos necessidade de efetuar a análise de sensibilidade do processo, pois já saberemos, de antemão, que a aproximação obtida é uma resposta satisfatória para o problema em questão. Portanto, nesse caso, a análise regressiva não representa uma etapa intermediária no estudo dos erros.

(3.68) Exemplo - Suponhamos que um vetor x represente as diversas posições ocupadas por um foguete que esteja realizando uma viagem a Marte, tendo a Terra como ponto de partida. Seja Ax a representação da derivada da quantidade de movimento do foguete nos mesmos instantes. Finalmente, suponhamos que b represente, nos mesmos instantes, a força resultante sobre o foguete, incluindo ambas as forças, as externas e aquelas de pro-

pulsão do próprio foguete. Então, o sistema de equações lineares $Ax = b$ representa o equilíbrio das forças reais e inerciais, de modo que, para qualquer b dado, a trajetória será representada pela solução exata x. Considerar a solução aproximada $\bar{x}$ como solução exata do sistema $A\bar{x} = b + \delta b$ significa afirmar que o foguete pode ser conservado na trajetória $\bar{x}$ através da aplicação de forças adicionais δb; conseqüentemente, se soubermos que $\|\delta b\| < \varepsilon$, teremos uma majoração para as forças corretivas necessárias. Ora, se ε for seguramente menor do que a grandeza das forças corretivas disponíveis para o foguete na forma de impulso de reserva, então, poderemos estar certos de que a trajetória $\bar{x}$ será alcançada e que, portanto, os erros de arredondamento que tenham sido cometidos durante a determinação de $\bar{x}$ serão compensados por uma disponível mudança de b. Portanto, nesse caso, é mais interessante examinar a grandeza de δb do que procurar estimar $\|\bar{x} - x\|$ para um b fixo.

Utilizando os exemplos que acabamos de apresentar, veremos, a seguir, mais algumas características da análise regressiva.

A análise inversa tem a vantagem, relativamente à direta, de permitir uma comparação entre o efeito dos erros de arredondamento cometidos durante a execução do processo numérico e o dos erros inerentes aos dados; fornece, portanto, uma preciosa avaliação da eficácia do processo considerado. Assim, por exemplo, o processo numérico utilizado em (3.63) poderá ser considerado bom se $\dfrac{\|\delta A\|_\infty}{\|A\|_\infty}$ for pequeno

Ocasionalmente a análise regressiva fornece estimativas "a priori" para o erro de arredondamento

a exemplo do que ocorre em relação ao cálculo dos autovalores de matrizes tridiagonais pelo método de bissecções, como mostra Wilkinson (24, p. 323-327). De maneira geral, porém, a determinação dessas estimativas é feita "a posteriori" e exige cálculos que, freqüentemente, são difíceis de efetuar. Assim, por exemplo, em (3.63) a estimativa de $\|\bar{x}-x\|_\infty$ exige a avaliação de $\text{cond}_\infty(A)$, o que exige, pelo menos, uma estimativa de $\|A^{-1}\|_\infty$. Apesar disso, é razoável supor-se que o processo numérico que determina uma estimativa razoável para $\|\delta A\|_\infty$ é melhor do que aquele para o qual isso não é possível.

3.3- Evolução histórica

O primeiro tratamento rigoroso dos erros de arredondamento nas operações aritméticas foi efetuado por J. von Neumann e H. H. Goldstine no artigo "Numerical inverting of matrix of high order" publicado no Bulletin of the American mathematical Society de número 53, em 1947, às páginas 1021-1099. Nesse artigo von Neumann e Goldstine introduziram as desigualdades que devem ser satisfeitas pelos erros de arredondamento nas operações aritméticas fundamentais realizadas em vírgula fixa e estabeleceram um padrão básico para a análise rigorosa de um algoritmo de cálculo. Tanto a notação quanto as idéias gerais contidas nesse trabalho passaram a ser adotadas pelos autores que os sucederam.

Os erros de arredondamento nas operações aritméticas fundamentais realizadas em vírgula flutuante foram estudados principalmente por James H. Wilkinson. Desde 1957, quando introduziu as desigualdades satisfeitas por esses erros, tem o referido autor contribuído significativamente para o desenvolvimento da análise de erros na aritmética de vír

gula móvel. Em 1960 essas desigualdades, já em forma mais aperfeiçoada, foram reapresentadas por Wilkinson em um artigo em que foram aplicadas a diferentes tipos de problemas (24). Além de inúmeros trabalhos, publicados em diferentes revistas especializadas do ramo, esse autor publicou dois livros, (26) e (22), em 1963 e 1965, respectivamente, cujo conteúdo é básico a todos que pretendam trabalhar nesse campo da ciência.

A idéia da análise regressiva de erros estava, até certo ponto, implícita nos artigos de von Neumann e Goldstine, porém, somente foi descrita explicitamente por J. W. Givens no artigo "Numerical computation of the characteristic values of a real symmetric matrix", publicado pelo Oak Ridge National Laboratory, do Tennessee, sob a especificação Rep. ORNL-1574, em 1954. Nesse artigo Givens analisou a redução de uma matriz real simétrica à forma tridiagonal por rotações planas e o cálculo dos autovalores de uma matriz tridiagonal através do processo da seqüência de Sturm. Ambas as análises foram realizadas na aritmética de vírgula fixa; a primeira delas não trouxe contribuição significativa para os conhecimentos da época, porém, a segunda tornou-se de fundamental importância no desenvolvimento da análise de erros. Entretanto, essa análise, ao ser publicada, não recebeu a devida atenção, o que se deu em conseqüência de dois fatores principais: o primeiro foi o fato de sua apresentação não ter sido feita através de uma revista especializada; o segundo foi conseqüência da mesma ter sido precedida por exaustiva análise de erros que só foi compreendida por reduzido número de pessoas. A dificuldade de entendimento dessa análise introdutória resultou do uso da aritmética de vírgula fixa, o que tornou as demonstrações muito trabalhosas, em virtude da necessidade de introdução de vários fatores de escala. Posteriormente, Wilkinson (25, p. 552-554) mostrou que, na aritmética

de vírgula flutuante, essas demonstrações se tornam suscintas e claras.

A análise regressiva de erros constitui um traço característico dos trabalhos de Wilkinson, o qual contribuiu de forma decisiva para o aperfeiçoamento e divulgação dessa forma de análise de erros.

O termo "número de condição" parece ter sido inicialmente empregado por Turing, entretanto, o termo "mal condicionamento" já era de uso comum entre os analistas numéricos há algum tempo antes disso.

Atualmente os erros de arredondamento continuam sendo muito estudados e frequentemente são publicados trabalhos e realizados congressos visando o seu desenvolvimento.

4- Análise de erros do método de eliminação de Gauss. Aplicações

Neste capítulo é apresentada uma análise regressiva de erros do método de eliminação de Gauss, e são considerados alguns problemas que podem ser resolvidos pela aplicação do mesmo: cálculo de determinantes, inversão de matrizes e refinamento da solução aproximada. Também são feitas algumas considerações a respeito de matrizes que, por suas formas especiais, apresentam simplificações em relação à eliminação de Gauss. Na parte inicial considera-se a resolução de sistemas triangulares, visto que, pela aplicação do método em questão, um sistema geral é transformado em dois triangulares. A norma adotada em todas essas aplicações é a norma ∞, em virtude de ser de cálculo fácil e por apresentar algumas vantagens sobre as demais, como vimos em (1.20)-(1.22), (1.32), (1.33) e (1.38). O uso da "barra", para indicar o valor aproximado, é dispensado em todo o capítulo, já que a análise regressiva só utiliza os valores calculados, não efetuando comparações entre esses e os verdadeiros. Em alguns poucos pontos em que esses últimos são usados, fizemos observações a respeito e os indicamos de forma a evitar confusões.

4.1. Resolução de sistemas triangulares

Consideremos o sistema de equações lineares

$$Lx = b \qquad (4.1)$$

em que $L=(l_{ij})$ é uma matriz triangular inferior pertencente a $M_n(R)$. As componentes do vetor solução x são deter

minados na ordem $x_1, \cdots, x_n$) por:

$$(4.2)\qquad \begin{cases} x_1 \simeq fl\left(\dfrac{b_1}{l_{11}}\right) \\ x_i \simeq fl\left(\dfrac{-l_{i1}x_1 - l_{i2}x_2 \cdots - l_{i,i-1}x_{i-1} + b_i}{l_{ii}}\right), \quad i = 2, \cdots, n \end{cases}$$

Aplicando (2.113) e (2.114) a (4.2)

$$(4.3)\qquad \begin{cases} x_1 = \dfrac{b_1}{l_{11}(1+\delta_{11})} \\ x_i = \dfrac{-l_{i1}x_1(1+\delta_{i1}) - \cdots - l_{i,i-1}x_{i-1}(1+\delta_{i-1})x_{i-1} + b_i}{l_{ii}(1+\delta_{ii})(1+\delta'_{ii})}, \\ \\ i = 2, \cdots, n \end{cases}$$

$$(4.4)\qquad \begin{cases} |\delta_{ii}|, |\delta'_{ii}| \le u, \qquad i = 1, \cdots, n \\ |\delta_{i1}| \le (i-1)\,1{,}01\,u, \qquad i = 2, \cdots, n \\ |\delta_{ij}| \le (i+1-j)\,1{,}01\,u, \qquad i = 2, \cdots, n;\ j = 2, \cdots, i-1 \end{cases}$$

Em (4.2) foram usadas as relações (2.113) e (2.114), de modo a se obter que os fatores $(1+\delta)$ multipliquem elementos de L.

As equações (4.3) também podem ser escritas

$$(4.5)\qquad \begin{cases} l_{11}(1+\delta_{11})x_1 = b_1 \\ l_{i1}(1+\delta_{i1})x_1 + \cdots + l_{i,i-1}(1+\delta_{i-1})x_{i-1} + l_{ii}(1+\delta_{ii})(1+\delta'_{ii})x_i = b_i, \\ \\ i = 2, \cdots, n \end{cases}$$

originando

$$(4.6)\qquad (L + \delta L)x = b, \qquad \text{onde, por (4.4),}$$

$$(4.7)\qquad |\delta L| \le 1{,}01\,u \begin{pmatrix} |l_{11}| & & & & \\ |l_{21}| & 2|l_{22}| & & & \\ 2|l_{31}| & 2|l_{32}| & 2|l_{33}| & & \\ 3|l_{41}| & 3|l_{42}| & 2|l_{43}| & 2|l_{44}| & \\ \vdots & \vdots & \vdots & \vdots & \ddots \\ (n-1)|l_{n1}| & (n-1)|l_{n2}| & (n-2)|l_{n3}| & (n-3)|l_{n4}| & \cdots\ 2|l_{nn}| \end{pmatrix}$$

Consequentemente, por (1.38)

$$(4.8)\qquad \|\delta L\|_\infty \le \frac{n(n+1)}{2}\, 1{,}01\, u \max_{i,j} |l_{ij}|$$

Por outro lado, de (4.7) também obtemos:

$$|\delta L| \le 1{,}01\, n\, u\, |L|$$

e, portanto, por (1.38),

$$(4.9)\qquad \|\delta L\|_\infty \le 1{,}01\, n\, u\, \|L\|_\infty$$

Reunindo todas as conclusões até agora obtidas para sistemas triangulares inferiores, temos:

(4.10) Teorema - A solução aproximada do sistema triangular inferior $Lx = b$, determinada através de (4.2) e da aritmética de vírgula flutuante com unidade de arredondamento u, é a solução exacta de um sistema $(L+\delta L)x = b$, em que a perturbação δL verifica (4.7) - (4.9).

Para sistemas de equações lineares

$$(4.11)\qquad Ux = b,$$

em que $U = (u_{ij})$ é uma matriz triangular superior per-

Pertencente a $M_n(R)$, a determinação das componentes do vetor solução x é feita na ordem $x_n, x_{n-1}, \cdots, x_1$, por:

$$(4.12)\qquad \begin{cases} x_n \simeq fl\left(\dfrac{b_n}{u_{nn}}\right) \\ x_r \simeq fl\left(\dfrac{-u_{r,r+1}x_{r+1} - u_{r,r+2}x_{r+2} - \cdots - u_{rn}x_n + b_r}{u_{rr}}\right), \\ r = n-1, n-2, \cdots, 1 \end{cases}$$

Procedendo de forma análoga à utilizada para sistemas triangulares superiores obtemos:

(4.13) Teorema - A solução aproximada do sistema triangular superior $Ux = b$, obtida através de (4.12) e do uso da aritmética de vírgula flutuante com unidade de arredondamento u é a solução exata de um sistema $(U+\delta U)x = b$, em que a perturbação δU satisfaz (4.14) - (4.16).

$$(4.14)\qquad |\delta U| \leq 1{,}01\,u \begin{pmatrix} 2|u_{11}| & (n-1)|u_{12}| & (n-1)|u_{13}| & (n-2)|u_{14}| & \cdots & 3u_{1,n-1} & 2|u_{1n}| \\ 0 & 2|u_{22}| & (n-2)|u_{23}| & (n-2)|u_{24}| & \cdots & 3u_{2,n-1} & 2|u_{2n}| \\ 0 & 0 & 2|u_{33}| & (n-3)|u_{34}| & \cdots & 3u_{3,n-1} & 2|u_{3n}| \\ \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 2u_{n-1,n-1} & |u_{1n}| \\ 0 & 0 & 0 & 0 & \cdots & 0 & |u_{nn}| \end{pmatrix}$$

$$(4.15)\qquad \|\delta U\|_\infty \leq \frac{n(n+1)}{2}\,1{,}01\,u \max_{i,j}|u_{ij}|$$

$$|\delta U| \leq 1{,}01\,n\,u\,|U|$$

$$(4.16)\qquad \|\delta U\|_\infty \leq 1{,}01\,n\,u\,\|U\|_\infty$$

Poderíamos utilizar acumulação de produtos escalares no cálculo de (4.2) e (4.12), com a finalidade de obter uma solução aproximada mais precisa. Não adotamos essa atitude porque julgamos que o tempo adicional de uso de computador requerido não seria justificável, pois, de maneira geral, os sistemas triangulares considerados possuem coeficientes afetados de erros, em virtude de resultarem da aplicação do método de Gauss a sistemas gerais. Por outro lado, as soluções aproximadas determinadas através de (4.2) e (4.12) são, usualmente, muito precisas, ainda que o sistema considerado seja muito mal condicionado, como veremos a seguir. O uso da acumulação de produtos escalares, para a resolução de sistemas triangulares, é considerado por Wilkinson (26, p. 103-104), (22, p. 247-249).

Seja z a solução exata de (4.1) e seja x uma aproximação de z, obtida pela forma descrita em (4.10). Consideremos o erro relativo $\frac{\|x-z\|}{\|z\|}$, em função do erro relativo $\frac{\|\delta L\|}{\|L\|}$. Por (3.22) e (4.3):

$$\frac{\|x-z\|_\infty}{\|z\|_\infty} \leq \frac{\operatorname{cond}_\infty(L)\ \dfrac{1{,}01\, n\, u\, \|L\|_\infty}{\|L\|_\infty}}{1 - \operatorname{cond}_\infty(L)\ \dfrac{1{,}01\, n\, u\, \|L\|_\infty}{\|L\|_\infty}}$$

$$\frac{\|x-z\|_\infty}{\|z\|_\infty} \leq \frac{1{,}01\, n\, u\, \operatorname{cond}_\infty(L)}{1 - 1{,}01\, n\, u\, \operatorname{cond}_\infty(L)} \tag{4.17}$$

Em virtude de (4.17), podemos afirmar que, se $1{,}01\, n\, u\, \operatorname{cond}_\infty(L) \ll 1$, o erro relativo da solução é pequeno. Por outro lado, Wilkinson (26, p. 105-107), (22, p. 249-251) afirma que as soluções aproximadas determinadas pelo processo descrito são, frequentemente, muito mais precisas do que a delimitação (4.17) pode sugerir, mesmo quando L é mal condicionada. Afirma, ainda, que para algumas classes de matrizes o erro relativo não depende de $\operatorname{cond}_\infty(L)$.

Consideremos o vetor resíduo $b - Lx = \delta L x$, que fornece:

$$\|b - Lx\|_\infty \leq \|\delta L\|_\infty \|x\|_\infty$$

Em virtude de (4.8):

$$\|b - Lx\|_\infty \leq \frac{n(n+1)}{2}\, 1{,}01\, u \max_{i,j} |l_{ij}| \cdot \|x\|_\infty \tag{4.18}$$

A desigualdade (4.18) fornece uma majoração para o elemento de maior valor absoluto do vetor resíduo, em função do elemento de maior valor absoluto de L e de x, respectivamente. Uma análise dessa majoração permite-nos concluir que a existência de resíduos pequenos não implica, necessariamente, na obtenção de uma solução precisa; se, por exemplo, L for tal que $|l_{ij}| \leq 1$ $(i,j = 1, \cdots, n)$ e se $\|x\|_\infty$ também for aproximadamente 1, então, o resíduo será pequeno quer x seja ou não uma solução precisa

4.2 - Eliminação de Gauss

Os métodos numéricos utilizados para a resolução de sistemas de equações lineares podem ser divididos em dois grupos, diretos e iterativos. Recebem a denominação de diretos ou exatos os métodos que, com um número finito de operações aritméticas elementares, forneceriam a solução exata de um sistema dado, não fosse a ocorrência de erros de arredondamento durante a sua execução. São chamados iterativos os métodos que possibilitam a resolução de um sistema linear de forma aproximada; a solução é obtida como o limite de sucessivas aproximações, calculadas por algum processo uniforme. Os métodos diretos são recomendáveis para sistemas cuja matriz é densa e os iterativos para sistemas cuja matriz é esparsa ou rarefeita. Con-

siderarem os apenas os primeiros, os quais se baseiam na idéia de eliminação atribuída a Gauss, que será considerada a seguir.

Seja o sistema de equações lineares

(4.19) $\quad Ax = b$,

em que A é a matriz dos coeficientes, x o vetor solução e b o vetor dos termos independentes. Suponhamos $A \in M_n(R)$ e A não singular.

São conhecidas diversas variantes do método de eliminação de Gauss, todas equivalentes à decomposição da matriz A em um produto LU de duas matrizes triangulares, L triangular inferior e U triangular superior. Descreveremos aquela em que os elementos de L são fixados e iguais a 1. Obtida essa fatoração LU, denominada <u>decomposição triangular ou decomposição LU</u>, o sistema (4.19) pode ser escrito

(4.20) $\quad LUx = b$, com

(4.21) $\quad Ly = b$

(4.22) $\quad Ux = y$

e, portanto, a resolução de (4.19) fica reduzida à resolução dos sistemas triangulares (4.21) e (4.22), o que é simples, como vimos em 4.1.

A decomposição triangular é conseguida através da determinação de uma seqüência $A^{(1)} = A, A^{(2)}, \cdots, A^{(n)}$ de matrizes tais que os elementos das (k-1) primeiras colunas de $A^{(k)}$, situados abaixo da diagonal principal, são nulos. A matriz $A^{(k+1)}$ é obtida de $A^{(k)}$ pela subtração de um múltiplo da k-ésima linha de cada uma das linhas que lhe estão abaixo (linhas k+1, k+2, $\cdots$, n). Cada um desses múltiplos é determinado através de um fator, deno-

minado multiplicador de Gauss, escolhido de forma que seriam nulos os elementos da k-ésima coluna de $A^{(k+1)}$, situados abaixo da diagonal principal, caso não houvesse a ocorrência de erros de arredondamento no processo; de maneira geral, esses elementos não são calculados, mas são tomados iguais a zero por definição. Seja $A^{(k)} = (a_{ij}^{(k)})$ $i,j = 1, \cdots, n$. A k-ésima linha de $A^{(k)}$ é chamada k-ésima linha pivô e o elemento $a_{kk}^{(k)}$, k-ésimo pivô. $A^{(1)}, A^{(2)}, \cdots A^{(n)}$ são chamadas primeira, segunda, ..., k-ésima, ..., n-ésima matriz reduzida, respectivamente. Indicaremos por $\bar{A}^{(k)}$ a matriz constituída pelas linhas e colunas $k, (k+1), \cdots, n$ de $A^{(k)}$, na mesma ordem em que se encontram em $A^{(k)}$.

Os multiplicadores de Gauss são dados por

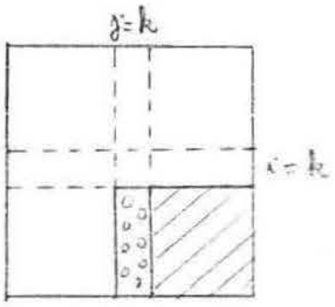
$$m_{ik} = fl\left(\frac{a_{ik}^{(k)}}{a_{kk}^{(k)}}\right), \quad i \geq k+1, \quad k = 1,2,\cdots,n \tag{4.23}$$

e os elementos de $A^{(k+1)}$ por

$$a_{ij}^{(k+1)} = \begin{cases} 0, & i \geq k+1, \ j = k \\ fl\left(a_{ij}^{(k)} - m_{ik} \times a_{kj}^{(k)}\right), & i \geq k+1, \ j \geq k+1 \\ a_{ij}^{(k)}, & \text{nos demais casos} \end{cases} \tag{4.24}$$

em que $k = 1, 2, \cdots, n$.

As três regiões de $A^{(k+1)}$, envolvidas nos três diferentes casos de (4.24), são ilustrados no diagrama abaixo, encontrado em Forsythe (7, p. 99).



Em virtude da ocorrência de erros de arredondamento, em geral não é obtida a decomposição triangular exata de A. Mostraremos, entretanto, que as matrizes L e U assim determinadas constituem a decomposição triangular exata de uma matriz obtida por ligeira perturbação de A, ou seja,

$$LU = A + E ,$$

e daremos delimitações para $|E|$.

Consideremos, inicialmente, a obtenção dos elementos $a_{ij}^{(k+1)}$, dados em (4.24). Por (4.23) e (2.113), podemos afirmar que, para os multiplicadores, são encontrados os valores

$$m_{ik} = \frac{a_{ik}^{(k)}}{a_{kk}^{(k)}} \left(1+\delta_{ik}\right) , \quad |\delta_{ik}| \leq \mu \quad , \quad \text{ou,} \tag{4.25}$$

$$0 = a_{ik}^{(k)} - m_{ik}\, a_{kk}^{(k)} + \varepsilon_{ik}^{(k)} , \tag{4.26}$$

$$\varepsilon_{ik}^{(k)} = a_{ik}^{(k)}\, \delta_{ik} , \tag{4.27}$$

Portanto, por (4.26), podemos afirmar que $a_{ik}^{(k+1)}$ $(i \geq k+1)$ pode ser tomado exatamente igual a zero desde que $a_{ik}^{(k)}$ seja perturbado de $\varepsilon_{ik}^{(k)}$, em que $\varepsilon_{ik}^{(k)}$ é dado em (4.27)

Por outro lado, para $i \geq k+1$ e $j \geq k+1$, (4.24) fornece

$$a_{ij}^{(k+1)} = fl\left(a_{ij}^{(k)} - fl\left(m_{ik} \times a_{kj}^{(k)}\right)\right) \tag{4.28}$$

Aplicando (2.113) e (2.114) a (4.28) obtemos;

$$a_{ij}^{(k+1)} = \frac{a_{ij}^{(k)} - m_{ik}\, a_{kj}^{(k)}(1+\delta_{ij})}{1+\delta'_{ij}} \; ; \; |\delta_{ij}|, |\delta'_{ij}| \leq \mu , \tag{4.29}$$

que fornece

$$a_{ij}^{(k+1)} = a_{ij}^{(k)} - m_{ik}\, a_{kj}^{(k)} + \varepsilon_{ij}^{(k)} \quad , \text{ com} \tag{4.30}$$

(4.31) $$\varepsilon_{ij}^{(k)} = -m_{ik}\, a_{kj}^{(k)}\, \delta_{ij} - a_{ij}^{(k+1)}\, \delta'_{ij}$$

Logo, por (4.30) podemos concluir que, para $i \geq k+1$ e $j \geq k+1$, o valor $a_{ij}^{(k+1)}$ seria obtido exatamente na decomposição triangular, a partir de elementos $a_{ij}^{(k)}$ perturbados pelos $\varepsilon_{ij}^{(k)}$ dados em (4.31).

Finalmente, os elementos restantes de $A^{(k+1)}$ são obtidos exatamente, pois os $a_{ij}^{(k+1)}$ são iguais aos $a_{ij}^{(k)}$.

Resumindo, $A^{(k+1)}$ pode ser obtida de forma exata na decomposição triangular, a partir da matriz resultante da perturbação de $A^{(k)}$ por

(4.32) $$E^{(k)} = \left(\varepsilon_{ij}^{(k)}\right), \quad \text{com}$$

(4.33) $$\varepsilon_{ij}^{(k)} = \begin{cases} a_{ik}^{(k)}\, \delta_{ik} & , \text{para } i \geq k+1,\ j = k \\ -m_{ik}\, a_{kj}^{(k)}\, \delta_{ij} - a_{ij}^{(k+1)}\, \delta'_{ij} & , \text{para } i \geq k+1,\ j \geq k+1 \\ 0 & , \text{nos demais casos} \end{cases}$$

Seja

(4.34) $$L^{(k)} = \begin{pmatrix} & & & & \\ & & 0 & & \\ & & 0 & & \\ & & \vdots & & \\ & & 0 & & \\ & & m_{k+1,k} & & \\ & & m_{k+2,k} & & \\ & & \vdots & & \\ & & m_{m,k} & & \end{pmatrix}$$

Então, a k-ésima etapa da decomposição triangular, que produzirá a matriz $A^{(k+1)}$, será completamente descrita pela equação (4.35), incluindo os erros de arredondamento.

(4.35) $$A^{(k+1)} = A^{(k)} - L^{(k)} A^{(k)} + E^{(k)} .$$

Somando as equações (4.35) para $k = 1, 2, \cdots, n-1$ estamos

$$L^{(1)}A^{(1)} + L^{(2)}A^{(2)} + \cdots + L^{(n-1)}A^{(n-1)} + L^{(n)}A^{(n)} = A^{(1)} + E^{(1)} + E^{(2)} + \cdots + E^{(n-1)}$$

Como a matriz $L^{(k)}A^{(k)}$ depende somente da k-ésima linha de $A^{(k)}$, que é igual à k-ésima linha de $A^{(n)}$, temos:

$$\left(L^{(1)} + L^{(2)} + \cdots + L^{(n-1)} + I\right) A^{(n)} = A^{(1)} + E^{(1)} + E^{(2)} + \cdots + E^{(n-1)},$$

(4.36a) $$LU = A + E, \quad \text{com}$$

(4.36b) $$L = L^{(1)} + L^{(2)} + \cdots + L^{(n-1)} + I$$

(4.36c) $$U = A^{(n)}$$

(4.36d) $$A = A^{(1)}$$

(4.36e) $$E = E^{(1)} + E^{(2)} + \cdots + E^{(n-1)},$$

Daí que facilmente se verifica que L e U são matrizes triangulares inferior e superior, respectivamente, sendo L de diagonal unitária e igual a

(4.37) $$L = \begin{pmatrix} 1 & & & 0 \\ m_{21} & 1 & & \\ m_{31} & m_{32} & 1 & \\ & & \ddots & \\ m_{n1} & m_{n2} & m_{n3} & 1 \end{pmatrix}$$

Como pode ser observado em (4.33), as perturbações (erros de arredondamento) poderão ser grandes se ou os multiplicadores ou os elementos das matrizes reduzidas, ou ambos, forem grandes. Por outro lado, o processo não funcionará se o pivô for nulo. Esse último problema é resolvido pela troca de cada pivô com outra, de modo que o novo elemento pivô seja diferente de zero (certamente existe um elemento

não nulo na coluna k, caso contrário o determinante de A seria nulo e A singular, contrariando nossa hipótese inicial). A grandeza dos multiplicadores é controlada por uma escolha conveniente do pivô. Na prática, de maneira geral, essa escolha é feita de modo a assegurar

$$|m_{ij}| \leq 1 \quad , \quad i \geq j+1 \quad , \quad j = 1, 2, \cdots n-1 \tag{4.42}$$

e é obtida através de duas técnicas principais. Na primeira delas, denominada pivotamento parcial, é escolhido para k-ésimo pivô o elemento de maior valor absoluto da primeira coluna de $\bar{A}^{(k)}$. Na segunda, denominada pivotamento completo, é eleito k-ésimo pivô o elemento de maior valor absoluto em toda a matriz $\bar{A}^{(k)}$.

O uso de pivotamento, com a finalidade de assegurar (4.42), permite, também, a determinação de uma estimativa para a grandeza dos elementos das matrizes reduzidas. De fato, facilmente pode-se verificar que, quando é utilizado pivotamento parcial, se os elementos de $A^{(1)}$ satisfazem a

$$|a_{ij}^{(1)}| \leq a \quad , \tag{4.43}$$

então, os elementos de $A^{(k)}$ verificam

$$|a_{ij}^{(k)}| \leq 2^{(k-1)}\, a \tag{4.44}$$

Por outro lado, quando é utilizado o pivotamento completo, Wilkinson (23, p. 283-284) prova que, se (4.43) se verifica, então, os elementos de $A^{(k)}$ satisfazem a

$$|a_{ij}^{(k)}| < a\, k^{\frac{1}{2}} \left(2^{1} \cdot 3^{\frac{1}{2}} \cdot 4^{\frac{1}{3}} \cdots k^{\frac{1}{k-1}}\right)^{\frac{1}{2}} . \tag{4.45}$$

Para algumas classes de matrizes especiais podem ser obtidas delimitações melhores do que as dadas em (4.44) e (4.45), a exemplo do que ocorre com as matrizes de Hessenberg, como veremos em 4.5.

Relativamente às delimitações para os elementos das matrizes reduzidas, dadas em (4.44) e (4.45), Wilkinson,

(23, p. 283-284) e (26, p. 97), afirma ter conhecimento de algumas matrizes especiais para as quais se verifica a igualdade em (4.44), mas que, na prática, de maneira geral (4.44) é substituída por

$$|a_{ij}^{(k)}| \leq 8a \tag{4.46}$$

e que, no caso de matrizes extremamente mal condicionadas, geralmente os elementos das sucessivas $A^{(k)}$ decrescem em valor absoluto. Afirma, ainda, desconhecer matrizes que, durante a decomposição triangular com pivotamento completo, originem elementos para os quais

$$|a_{ij}^{(k)}| > ka, \tag{4.47}$$

o que mostra que a majoração (4.45) deve ser exagerada.

De maneira geral, o pivotamento parcial é o único utilizado na prática. Esse fato decorre, principalmente, da grande dificuldade de programação do pivotamento completo para computadores, dificuldade essa não compensada por grande melhoria das delimitações de erros. Por outro lado, o pivotamento parcial conserva a disposição particular dos elementos de algumas matrizes que possuem muitos elementos nulos, enquanto o pivotamento completo a destrói; essa propriedade é muito importante para a resolução de certos problemas, a exemplo do que ocorre relativamente ao cálculo de autovalores e autovetores de matrizes.

Utilizaremos o pivotamento parcial em todo o estudo de erros que efetuaremos a seguir. Ao iniciar a fatoração LU de uma matriz A, com aplicação dessa forma de pivotamento, admitiremos que as linhas de A tenham sido inicialmente permutadas, de modo que nenhuma troca de linhas se torne necessária durante a decomposição triangular. Assim procedendo, poderemos utilizar a descrição do método de Gauss apresentada anteriormente, assim como

as diversas conclusões obtidas para as matrizes reduzidas e para as matrizes L e U. Embora, na prática, a referida permutação inicial não seja conhecida "a priori", a sua determinação não é necessária, já que permutações de linhas não originam erros de arredondamento. Simplificaremos a notação indicando a matriz reordenada e a matriz original pelo mesmo símbolo A. É lógico que o determinante da matriz reordenada é igual ao da matriz original, a menos do fator (± 1).

Determinaremos, a seguir, uma delimitação para $|E|$, em que E é dada em (4.40), no caso de utilização do pivotamento parcial. Seja

$$\rho = \frac{\max\limits_{i,j,k} \left| a_{ij}^{(k)} \right|}{\| A \|_\infty} \quad ; \quad i,j,k = 1,2,\cdots,n \tag{4.48}$$

Embora não se tenha uma boa delimitação para ρ "a priori", pode-se calculá-lo de forma relativamente fácil durante a decomposição triangular.

A partir de (4.48) obtemos

$$\left| a_{ij}^{(k)} \right| \leq \rho \| A \|_\infty \quad ; \quad i,j,k = 1,2,\cdots,n \tag{4.49}$$

que, juntamente com (4.33), (4.29) e (4.42), fornece

$$\left| \varepsilon_{ij}^{(k)} \right| \leq \rho \| A \|_\infty \begin{cases} u, & i \geq k+1, \ j = k \\ 2u, & i \geq k+1, \ j \geq k+1 \\ 0, & \text{nos demais casos} \end{cases} \tag{4.50}$$

As relações (4.50) e (4.40) nos dão, finalmente, a majoração para $|E|$, que é apresentada em (4.51). A matriz à direita dessa desigualdade é facilmente reconstruída, já que seus elementos representam o número de operações aritméticas necessárias à determinação das matrizes L e U.

(4.51) $$|E| \leq \rho \|A\|_\infty u \begin{pmatrix} 0 & 0 & 0 & \cdots & 0 & 0 \\ 1 & 2 & 2 & \cdots & 2 & 2 \\ 1 & 3 & 4 & \cdots & 4 & 4 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 1 & 3 & 5 & \cdots & 2n-4 & 2n-4 \\ 1 & 3 & 5 & \cdots & 2n-3 & 2n-2 \end{pmatrix}$$

Por (4.51), (1.38) e (1.30)

$$\|E\|_\infty \leq \rho \|A\|_\infty u \left\{ \left[\sum_{j=1}^{n} (2j-1) \right] - 1 \right\} \quad , \text{ ou,}$$

(4.52) $$\|E\|_\infty \leq n^2 \rho \|A\|_\infty u$$

Concluindo, temos

(4.53) <u>Teorema</u> - Seja $A \in M_n(R)$, A não singular. Consideremos a decomposição LU de A, através da eliminação de Gauss com pivotamento, na aritmética de vírgula flutuante, com unidade de arredondamento u. As matrizes L e U assim determinadas satisfazem a $LU = A + E$, em que E verifica (4.51) e (4.52).

Uma vez efetuada a decomposição triangular de A, a solução aproximada de (4.19) pode ser obtida pela resolução dos sistemas triangulares (4.21) e (4.22). Aplicando (4.10) e (4.13) a esses sistemas, temos, por (4.20), que a aproximação x é a solução exata do sistema

(4.54) $$(L + \delta L)(U + \delta U)x = b \quad ,$$

com $\|\delta L\|_\infty$ e $\|\delta U\|_\infty$ dados em (4.8) e (4.15). Por outro lado, em virtude de (4.36) e (4.54), temos

$$\left[A + E + L(\delta U) + (\delta L)U + (\delta L)(\delta U) \right] x = b \quad ,$$

ou ainda

(4.55) $(A+\delta A)\, x = b$, com

(4.56) $\delta A = E + L(\delta U) + (\delta L)U + (\delta L)(\delta U)$

Como, em conseqüência de (4.41), (4.42), (4.38) e (4.49)

(4.57) $|l_{ij}| \leq 1$ $\qquad i, j = 1, \cdots, n$

(4.58) $|u_{ij}| \leq \rho \|A\|_\infty$,

temos

(4.59) $\|L\|_\infty \leq n$

(4.60) $\|U\|_\infty \leq n\rho\|A\|_\infty$

e, ainda, utilizando (4.57) e (4.58) em (4.8) e (4.15), respectivamente

(4.61) $\|\delta L\|_\infty \leq \frac{n(n+1)}{2}\, 1{,}01\, u$

(4.62) $\|\delta U\|_\infty \leq \frac{n(n+1)}{2}\, 1{,}01\, \rho\, \|A\|_\infty\, u$

Em virtude de $nu << 1$ no caso em que esses resultados são utilizados, a partir (4.61) e (4.62) podemos afirmar

(4.63) $\|\delta L\|_\infty \|\delta U\|_\infty \leq n^2 \rho \|A\|_\infty u$

Aplicando (4.52), (4.59), (4.61), (4.60), (4.62) e (4.63) a (4.56) obtemos

(4.64) $\|\delta A\|_\infty \leq 1{,}01 \left(n^3 + 3n^2\right) \rho \|A\|_\infty u$

É importante observar que, embora δA dependa de L, pois δU e δL dependem de b, a delimitação de $\|\delta A\|_\infty$ não envolve b. Wilkinson (26, p.108) afirma que, na prática, de maneira geral ocorre

(4.65) $\|\delta A\|_\infty \leq 1{,}01\, n\, \rho\, \|A\|_\infty\, u$.

Resumindo todos esses resultados temos:

(4.66) Teorema - Seja $A \in M_n(R)$, A não singular. Consideremos a resolução do sistema de equações lineares $Ax = b$, através da eliminação de Gauss com pivotamento, na aritmética de vírgula flutuante, com unidade relativa de arredondamento u. A aproximação assim determinada é a solução exata do sistema $(A+\delta A)x = b$, em que δA, definida em (4.56), satisfaz (4.64).

Em virtude de (4.42) verificar-se para ambos os tipos de pivotamento, parcial e completo, podemos afirmar que (4.53) e (4.66) também valem para essa última forma de pivotamento, embora em nossas considerações só tenhamos mencionado a primeira. É claro, entretanto, que o valor de ρ depende do tipo de pivotamento considerado e do algoritmo utilizado para sua obtenção.

A seguir faremos algumas considerações a respeito da precisão da aproximação x, determinada da forma descrita em (4.66). Indicando por z a solução exata e utilizando (3.22) e (4.64), temos

$$\frac{\|x-z\|_\infty}{\|z\|_\infty} \leq \frac{1{,}01(n^3+3n^2)\rho\, \mathrm{cond}_\infty(A)\, u}{1-1{,}01(n^3+3n^2)\rho\, \mathrm{cond}_\infty(A)\, u} \tag{4.67}$$

e se $1{,}01(n^3+3n^2)\rho\, \mathrm{cond}_\infty(A)\, u \ll 1$, então, o erro absoluto será pequeno em comparação com a solução exata. Entretanto, o erro δA é formado pelos erros cometidos na decomposição triangular de A e por aqueles contraídos na resolução dos sistemas (4.21) e (4.22). Como as soluções aproximadas de sistemas triangulares são, de maneira geral, muito precisas, de modo que as delimitações dadas em (4.61) e (4.62) para $\|\delta L\|_\infty$ e $\|\delta U\|_\infty$, respectivamente, são dificilmente atingidas, podemos utilizar

(4.65) em substituição a (4.64) e concluir que, na prática, usualmente

$$\frac{\|x-z\|_\infty}{\|z\|_\infty} \leq \frac{1{,}01\, n\, \rho\, \mathrm{cond}_\infty(A)\, u}{1 - 1{,}01\, n\, \rho\, \mathrm{cond}_\infty(A)\, u} \qquad (4.67)$$

é que, portanto, o erro absoluto é pequeno em comparação à solução aproximada, desde que $1{,}01\, n\, \rho\, \mathrm{cond}_\infty(A)\, u << 1$.

Consideremos o resíduo da solução aproximada, $b - Ax$. Por (4.66)

$$\|b-Ax\|_\infty \leq 1{,}01\,(n^3+3n^2)\,\rho\,\|A\|_\infty\,\|x\|_\infty\,u\,, \qquad (4.68)$$

sendo que, por (4.65), podemos concluir que, freqüentemente, (4.69) é substituído por (4.70)

$$\|b-Ax\|_\infty \leq 1{,}01\, n\, \rho\, \|A\|_\infty \|x\|_\infty\, u \qquad (4.70)$$

Através de (4.69) e (4.70) facilmente se pode concluir que o resíduo pode ser pequeno mesmo que a solução aproximada não seja muito precisa; isso pode ocorrer, por exemplo, quando $\|A\|_\infty$ é próxima de 1 e $\|x\|_\infty$ é de ordem unitária.

Wilkinson, (23, p. 284-285) e (22, p. 193), insiste em que se "equilibre" a matriz A, antes de aplicar ao sistema (4.19) qualquer método numérico de solução. De maneira grosseira, uma matriz é dita <u>equilibrada</u> quando possui todas as linhas e todas as colunas com aproximadamente o mesmo comprimento, relativamente a alguma norma. A análise de erros do método de Gauss fornece resultados mais efetivos quando aplicada a matrizes equilibradas porque, então, uma pequena perturbação em uma linha (ou coluna) da matriz é da mesma ordem de grandeza daquela de qualquer outra linha (ou coluna). Quanto a tirar essa vantagem, não utilizamos matrizes equi-

libradas neste trabalho porque, infelizmente, o problema de equilibração ainda não está bem esclarecido. Não existem muitos algoritmos de equilibração e os poucos conhecidos não fornecem uma solução geral para o problema, pois às vezes valem apenas para algumas normas, enquanto que outras vezes valem apenas para determinadas matrizes especiais. Além disso, não existe uma forma única para se obter o equilíbrio de uma matriz e não se conhece, "a priori" qual das formas possíveis é a melhor; procedimentos de equilibração diversos podem transformar A em diferentes matrizes equilibradas, cujos números de condição podem diferir consideravelmente entre si, como exemplificado em (7, p. 37-46) e (8, p. 169).

Apresentaremos, a seguir, um exemplo bastante simples, através do qual podem ser observados os resultados obtidos em 4.1 e 4.2.

(4.71) Exemplo - Consideremos a resolução do sistema de equações lineares $Ax = b$, com A e b dados em (4.72), através da eliminação de Gauss com pivotamento parcial, na aritmética decimal de vírgula flutuante, com armazenamento de apenas dois dígitos.

(4.72)
$$A = \begin{pmatrix} 0,69 & 0,26 & 0,12 \\ 0,15 & 0,45 & 0,27 \\ 0,31 & 0,49 & 0,15 \end{pmatrix} \qquad b = \begin{pmatrix} 0,67 \\ 0,24 \\ 0,12 \end{pmatrix}$$

Vejamos, inicialmente, a decomposição triangular de A. As matrizes reduzidas e demais matrizes citadas em (4.36) – (4.40) são, nesse caso,

$$A^{(2)} = \begin{pmatrix} 0,69 & 0,26 & 0,12 \\ 0,00 & 0,39 & 0,24 \\ 0,00 & 0,37 & 0,96\times10^{-1} \end{pmatrix} \qquad A^{(3)} = \begin{pmatrix} 0,69 & 0,26 & 0,12 \\ 0,00 & 0,39 & 0,24 \\ 0,00 & 0,00 & -0,13 \end{pmatrix}$$

$$E^{(1)} = \begin{pmatrix} 0,00 & 0,00 & 0,00 \\ 0,18\times10^{-2} & -0,28\times10^{-2} & -0,36\times10^{-2} \\ 0,50\times10^{-3} & -0,30\times10^{-2} & 0,00 \end{pmatrix} \qquad E^{(2)} = \begin{pmatrix} 0,00 & 0,00 & 0,00 \\ 0,00 & 0,00 & 0,00 \\ 0,00 & 0,50\times10^{-3} & 0,20\times10^{-2} \end{pmatrix}$$

$$L^{(1)} = \begin{pmatrix} 0,00 & 0,00 & 0,00 \\ 0,22 & 0,00 & 0,00 \\ 0,45 & 0,00 & 0,00 \end{pmatrix} \qquad L^{(2)} = \begin{pmatrix} 0,00 & 0,00 & 0,00 \\ 0,00 & 0,00 & 0,00 \\ 0,00 & 0,95 & 0,00 \end{pmatrix}$$

Portanto:

(4.73) $$L = \begin{pmatrix} 0,10\times10 & 0,00 & 0,00 \\ 0,22 & 0,10\times10 & 0,00 \\ 0,45 & 0,95 & 0,10\times10 \end{pmatrix} \qquad U = \begin{pmatrix} 0,69 & 0,26 & 0,12 \\ 0,00 & 0,39 & 0,24 \\ 0,00 & 0,00 & -0,13 \end{pmatrix}$$

$$E = \begin{pmatrix} 0,00 & 0,00 & 0,00 \\ 0,18\times10^{-2} & -0,28\times10^{-2} & -0,36\times10^{-2} \\ 0,50\times10^{-3} & -0,25\times10^{-2} & 0,20\times10^{-2} \end{pmatrix}$$

$$LU = \begin{pmatrix} 0,69 & 0,26 & 0,12 \\ 0,1518 & 0,4472 & 0,2664 \\ 0,3105 & 0,4875 & 0,1520 \end{pmatrix}$$

Observe-se que $LU = A + E$

Consideremos, a seguir, a resolução dos sistemas triangulares obtidos

$$\left.\begin{matrix} \text{solução} \\ \text{aproximada} \\ \text{de } Ly = b \end{matrix}\right\} = \begin{pmatrix} 0,67 \\ 0,90\times10^{-1} \\ -0,27 \end{pmatrix} \qquad \left.\begin{matrix} \text{solução} \\ \text{aproximada} \\ \text{de } Ux = y \end{matrix}\right\} = \begin{pmatrix} 0,10\times10 \\ -0,11\times10 \\ 0,21\times10 \end{pmatrix}$$

$$\left.\begin{matrix}\text{solução exata}\\ \text{de } LUx = b\end{matrix}\right\} = \begin{pmatrix} 0,1001\cdots \times 10 \\ -0,1038\cdots \times 10 \\ 0,2072\cdots \times 10 \end{pmatrix} \qquad \left.\begin{matrix}\text{solução exata}\\ \text{de } Ax = b\end{matrix}\right\} = \begin{pmatrix} 0,10 \times 10 \\ -0,10 \times 10 \\ 0,20 \times 10 \end{pmatrix}$$

$$\left.\begin{matrix}1\text{ª solução}\\ \text{aproximada}\end{matrix}\right\} = \begin{pmatrix} 0,6560 \\ 0,2220 \\ 0,8600 \times 10^{-1} \end{pmatrix} \qquad \text{vetor resíduo} = \begin{pmatrix} 0,14 \times 10^{-1} \\ 0,18 \times 10^{-1} \\ 0,34 \times 10^{-1} \end{pmatrix}$$

4.5 – Cálculo de determinantes

Seja $A \in M_n(R)$ e consideremos a determinação do sinal de $\det A$.

Aplicando à matriz A a eliminação de Gauss com pivotamento parcial, obtemos matrizes triangulares L e U que, por (4.53), satisfazem $LU = A + E$, com E satisfazendo (4.51) e (4.52). Como o determinante de LU é igual ao de $A^{(1)}$, a menos do sinal, que depende do número de permutações efetuadas, temos:

$$\left|\det (A^{(1)} + E)\right| = \left| a_{11}^{(1)} \; a_{22}^{(2)} \cdots a_{nn}^{(n)} \right| \tag{4.54}$$

Aplicando (2.73) e (2.75) a (4.54), obtemos:

$$\left|\det (A^{(1)} + E)\right| \simeq \left| fl\left(a_{11}^{(1)} \cdots a_{nn}^{(n)}\right) \right| = \left| a_{11}^{(1)} \; a_{22}^{(2)} \cdots a_{nn}^{(n)} \right| (1+\varepsilon),$$

$$1 + \varepsilon = 1 + 1{,}01\,(n-1)\,\theta\,\mu, \qquad |\theta| \leq 1$$

O fator $(1+\varepsilon)$ tem pouca influência no resultado de modo que, a menos desse fator, o valor calculado pelo computador é igual ao determinante exato de $A + E$.

4.4. Inversão de matrizes

Dada a matriz não singular $A \in M_n(\mathbb{R})$, podemos determinar $A^{-1} \simeq X$ através da resolução de n sistemas de equações lineares

$$A x_i = e_i \quad , \quad i = 1, 2, \cdots, n \quad , \tag{4.77}$$

em que x_i e e_i são, respectivamente, a i-ésima coluna de X e I. Aplicando (4.66) a (4.77), podemos afirmar que cada aproximação x_i, determinada através da eliminação de Gauss com pivotamento parcial, satisfaz a

$$(A + \delta_i A)\, x_i = e_i \quad , \tag{4.78}$$

em que, como o indicam os índices em (4.78), as perturbações são diferentes para cada coluna da matriz I. Entretanto, a delimitação geral apresentada em (4.64) vale para todos os $\delta_i A$, de modo que a inversa aproximada X satisfaz a

$$(A + \delta A)\, X = I \quad , \tag{4.79}$$

com $\|\delta A\|_\infty$ dado em (4.64)

Consideremos a matriz resíduo

$$F = I - AX = \delta A\, X \tag{4.80}$$

Em virtude de (4.64)

$$\|F\|_\infty \leq 1{,}01\,(n^3 + 3n^2)\,\rho\, \|A\|_\infty\, u\, \|X\|_\infty \quad , \tag{4.81}$$

sendo que, ao contrário do que ocorre para a solução de sistemas lineares, em que soluções aproximadas que possuem resíduo pequeno não são, necessariamente, precisas, na determinação de inversas de matrizes a obtenção de resíduos pequenos implica na determinação de uma inversa precisa, ou seja, (4.81) fornece uma estimativa da precisão da aproximação determinada. Isso decorre de

(4.80), pois

$$A^{-1} - X = A^{-1} F \qquad \text{e, portanto,}$$

$$\frac{\|A^{-1} - X\|_k}{\|A^{-1}\|_k} \leq \|F\|_k \qquad (4.82)$$

qualquer que seja a norma k considerada.

Como (4.81) foi obtida a partir de (4.80), em que X é inversa de A à direita, precisamos considerar o caso em que X é inversa de A à esquerda. Em consequência de (4.80) temos:

$$\|I - XA\|_\infty \leq \|A^{-1}\|_\infty \|F\|_\infty \|A\|_\infty \qquad \text{, ou,}$$

$$\|I - XA\|_\infty \leq \text{cond}_\infty(A) \|F\|_\infty \qquad (4.83)$$

ou seja, o resíduo de X, como inversa à esquerda, pode ser maior que o de X, como inversa à direita, do fator $\text{cond}_\infty(A)$. Entretanto, Wilkinson (26, p. 110-111) afirma que, na prática, o método de Gauss com pivotamento parcial fornece resíduos $I - AX$ e $I - XA$ de mesma ordem de grandeza, mesmo quando A é mal condicionada.

Veremos, a seguir, um exemplo de cálculo de inversa.

(4.84) Exemplo - Consideremos a inversão da matriz A dada em (4.72). Utilizando, para isso, as matrizes L e U dadas em (4.73), obtemos:

$$\left.\begin{matrix}\text{inversa} \\ \text{aproximada} \\ \text{de } LU\end{matrix}\right\} = \begin{pmatrix} 0{,}18 \times 10 & -0{,}48 & -0{,}55 \\ -0{,}17 \times 10 & -0{,}21 \times 10 & 0{,}43 \times 10 \\ 0{,}18 \times 10 & 0{,}73 \times 10 & -0{,}77 \times 10 \end{pmatrix},$$

que pode ser comparada com

inversa exata de LU (arredondada para 2 dígitos) $=\begin{pmatrix} 0,18\times 10 & -0,54 & -0,45 \\ -0,17\times 10 & -0,19\times 10 & 0,47\times 10 \\ 0,19\times 10 & 0,73\times 10 & -0,77\times 10 \end{pmatrix}$

inversa exata de A (arredondada para 2 dígitos) $=\begin{pmatrix} 0,18\times 10 & -0,54 & -0,44 \\ -0,17\times 10 & -0,18\times 10 & 0,46\times 10 \\ 0,18\times 10 & 0,70\times 10 & -0,74\times 10 \end{pmatrix}$

Por outro lado,

$$AX=\begin{pmatrix} 0,96 & 0 & 0 \\ -0,10\times 10^{-1} & 0,10\times 10 & 0 \\ 0 & -0,10 & 0,10\times 10 \end{pmatrix} \quad XA=\begin{pmatrix} 0,96 & -0,20\times 10^{-1} & 0,10\times 10^{-1} \\ -0,20\times 10^{-1} & 0,10\times 10 & -0,30\times 10^{-1} \\ -0,10 & 0 & 0,10\times 10 \end{pmatrix}$$

Observe-se que AX e XA são da mesma ordem de grandeza.

4.5- Matrizes especiais

A eliminação de Gauss, até agora considerada em relação a matrizes gerais, apresenta simplificações quando aplicada a algumas classes de matrizes especiais. Assim, por exemplo, para o cálculo de determinantes de matrizes de Hessenberg a eliminação de Gauss pode ser realizada sem a utilização de qualquer forma de pivotamento porque, apesar da possibilidade de ocorrência de multiplicadores e elementos das matrizes reduzidas com ordem de grandeza elevada, ainda assim é sempre obtida uma boa aproximação. Para verificar esse fato, consideremos uma matriz superior de Hessenberg e determinemos os elementos de $A^{(n)}$, que são os utilizados no cálculo de $\det A$.

Como pode ser concluído pela análise de (1.11), a k-ésima etapa é iniciada com uma matriz

reduzida $A^{(k)}$, cujas linhas $(k+1)$ a n são constituídas por elementos de $A^{(1)}$. Por outro lado, a primeira coluna de $\bar{A}^{(k)}$, formada pelos elementos que serão zerados, apresenta apenas dois elementos não nulos, um dos quais, aquele que será anulado, pertence a uma das linhas ainda não modificadas. Como exemplo, apresentaremos $A^{(3)}$ para o caso $n = 6$

$$A^{(3)} = \left(\begin{array}{cccccc} a_{11}^{(1)} & a_{12}^{(1)} & a_{13}^{(1)} & a_{14}^{(1)} & a_{15}^{(1)} & a_{16}^{(1)} \\ 0 & a_{22}^{(2)} & a_{23}^{(2)} & a_{24}^{(2)} & a_{25}^{(2)} & a_{26}^{(2)} \\ \hline 0 & 0 & a_{33}^{(3)} & a_{34}^{(3)} & a_{35}^{(3)} & a_{36}^{(3)} \\ 0 & 0 & a_{43}^{(1)} & a_{44}^{(1)} & a_{45}^{(1)} & a_{46}^{(1)} \\ 0 & 0 & 0 & a_{54}^{(1)} & a_{55}^{(1)} & a_{56}^{(1)} \\ 0 & 0 & 0 & 0 & a_{65}^{(1)} & a_{66}^{(1)} \end{array}\right)$$

A matriz separada por linhas tracejadas, no canto inferior direito de $A^{(3)}$ e $\bar{A}^{(3)}$.

Portanto, existe apenas um multiplicador de Gauss por coluna, definido por

$$m_{i+1,i} = fl\left(\frac{a_{i+1,i}^{(1)}}{a_{ii}^{(i)}}\right) = \frac{a_{i+1,i}^{(1)}}{a_{ii}^{(i)}}\left(1+\varepsilon_{i+1,i}\right), \tag{4.85}$$

$$|\varepsilon_{i+1,i}| \leq \mu \quad ,$$

sendo que no numerador de (4.85) utilizamos $a_{i+1,i}^{(1)}$, em lugar de $a_{i+1,i}^{(i)}$, porque a linha $(i+1)$ ainda não foi modificada.

Em virtude da existência de um único multiplicador por coluna, podemos simplificar a notação, adotando a forma de representação utilizada por Wilkinson (24, p. 329-331), na qual os multiplicadores

$m_{i+1,i}$ são denotados por m_i e qualquer quantidade δ, não necessariamente a mesma em cada etapa, que satisfaça $|\delta| \leq u$, é representada por ε. Assim procedendo, podemos escrever as equações que originam as aproximações para os $a_{ii}^{(i)}$ na forma:

$$(4.86)\quad \begin{cases} a_{22}^{(2)} = \left[a_{22}^{(1)} - m_1 a_{12}^{(1)}(1+\varepsilon)\right](1+\varepsilon) \\ a_{33}^{(3)} = \left[a_{33}^{(1)} - m_2 a_{23}^{(2)}(1+\varepsilon)\right](1+\varepsilon) \\ \vdots \qquad\qquad \vdots \\ a_{ii}^{(i)} = \left[a_{ii}^{(1)} - m_{i-1} a_{i-1,i}^{(i-1)}(1+\varepsilon)\right](1+\varepsilon) \end{cases},$$

para as quais utilizamos (2.113).

Eliminando de (4.86) todas as quantidades $a_{ji}^{(j)}$, para as quais $j \neq 1$ ou $j \neq i$, obtemos:

$$(4.87)\quad a_{ii}^{(i)} = a_{ii}^{(1)}(1+\varepsilon) - m_{i-1}(1+\varepsilon)^3 a_{i-1,i}^{(1)} + m_{i-1}m_{i-2}(1+\varepsilon)^5 a_{i-2,i}^{(1)} - \cdots +$$
$$+ (-1)^{i-2} m_{i-1}m_{i-2}\cdots m_2 (1+\varepsilon)^{2i-3} a_{2i}^{(1)} + (-1)^{i-1} m_{i-1}m_{i-2}\cdots m_1 (1+\varepsilon)^{2i-2} a_{1i}^{(1)}$$

A partir de (4.85) e (4.87), podemos concluir que os elementos diagonais de $A^{(n)}$ são exatamente iguais aos elementos diagonais que teriam sido obtidos se tivéssemos realizado, de maneira exata, uma eliminação de Gauss com uma matriz A', obtida por perturbação de A, cujos elementos da i-ésima coluna fossem:

$$(4.88)\quad \begin{cases} a'_{1i} = a_{1i}^{(1)}(1+\varepsilon)^{2i-2} \\ a'_{ji} = a_{ji}^{(1)}(1+\varepsilon)^{2(i-j)+1} \quad , \; j = 2, 3, \cdots, i \\ a'_{i+1,i} = a_{i+1,i}^{(1)}(1+\varepsilon) \end{cases}$$

As modificações são tão pequenas, que praticamente nenhum prejuízo decorreu da não utilização de pivotamento. Observe-se, ainda, que a matriz A' também originaria, exatamente, os m_{i} obtidos na prática.

É importante que se observe que, se o cálculo do determinante de uma matriz Hessenberg superior é efetuado de maneira satisfatória sem uso de pivotamento, o mesmo não ocorre quanto à resolução de sistemas de equações lineares cuja matriz é desse tipo. Wilkinson (24, p. 330-331) ilustra esse fato através de um interessante exemplo.

Quando o pivotamento parcial é utilizado na aplicação da eliminação de Gauss a matrizes Hessenberg superiores, então, se $|a_{ij}^{(1)}| \leq a$, ocorre $\max |a_{ij}^{(k)}| \leq ka$, uma delimitação melhor do que a apresentada em (4.44) para matrizes gerais. Isso decorre do fato de $\bar{A}^{(k)}$ também ser Hessenberg superior.

As matrizes tridiagonais, definidas em (1.12), são matrizes Hessenberg tanto superiores quanto inferiores. Portanto, os resultados que vimos para matrizes Hessenberg superiores também valem para as tridiagonais. Para essas últimas matrizes, a eliminação de Gauss com pivotamento parcial fornece $|a_{ij}^{(k)}| \leq 2a$, se $|a_{ij}^{(1)}| \leq a$, como mostra Wilkinson (23, p. 288-289).

Outra classe de matrizes que apresenta simplificações para a eliminação de Gauss é a das diagonalmente dominantes, definidas em (1.13). Para essas matrizes os multiplicadores de Gauss satisfazem $|m_{ij}| < 1$, mesmo sem a utilização de qualquer forma de pivotamento. Por outro lado, se $|a_{ij}^{(1)}| \leq a$, então, $\max_{ij} |a_{ij}^{(k)}| \leq 2 \max |a_{ii}^{(1)}|$ $(i, j = 1, \cdots, n)$, como

mostra Wilkinson (23, p. 288-289).

4.6 - Refinamento da solução aproximada

Vimos, em 4.1 e 4.2, que, ao obter a solução aproximada de um sistema de equações lineares pelo método de eliminação de Gauss, não podemos, de imediato, afirmar nada a respeito da precisão da mesma. O refinamento iterativo fornece esse tipo de informação, dando orientações quanto ao condicionamento do sistema original, e possibilita uma melhoria, um refinamento, das aproximações utilizáveis. Exporemos apenas as características fundamentais desse processo, cuja descrição detalhada pode ser encontrada em Forsythe (7, p. 49-54), e, a seguir, faremos considerações quanto à sua convergência.

Consideremos o sistema (4.19) e indiquemos por x_1 a solução aproximada obtida através de (4.66). Para $m = 1, 2, \cdots$, a m-ésima iteração envolve três etapas:

(4.89) determinação do vetor resíduo, r_m, definido por

$$r_m = b - A x_m$$

(4.90) cálculo do vetor correção, c_m, através da resolução de

$$A c_m = r_m$$

(4.91) obtenção da aproximação refinada x_{m+1}, dada por

$$x_{m+1} = x_m + c_m .$$

Quando o processo é executado em um computador, os ciclos de refinamento podem ser interrompidos pela verificação de um dos critérios:

(4.92) $$\frac{\|c_m\|_\infty}{\|x_1\|_\infty} < \varepsilon, \quad \varepsilon > 0 \text{ pré-estabelecido}$$

(4.93) o número m, de iterações, excede um certo valor

N pré-especificado.

Em algumas circunstâncias (4.92) é substituído por

(4.94) $\| r_k - r_{k+1} \|_\infty < \delta$, $\delta > 0$ pré-estabelecido.

No caso de interrupção do processo em virtude da verificação de (4.92) ou (4.94), sem que (4.93) tenha ocorrido, obtém-se um refinamento da aproximação. Quando a interrupção decorre da verificação de (4.93), então, a aproximação inicial não é utilizável.

Caso não houvesse ocorrência de erros de arredondamento, o processo de refinamento convirgiria em uma iteração, pois teríamos, por (4.89)-(4.91):

$$A x_2 = A(x_1 + c_1) = A x_1 + A c_1 = A x_1 + r_1 = b .$$

A única etapa crítica é (4.89), pois b e Ax_m são, no caso da obtenção de x_m pela eliminação de Gauss com pivotamento, muito próximos. Por esse motivo, na prática, o cálculo dos resíduos é efetuado com precisão maior do que a utilizada nas demais etapas. Em geral, os resíduos são calculados em precisão dupla e as correções e aproximações refinadas em precisão simples. É lógico que, nesse caso, a precisão que se pode alcançar para a solução aproximada, através de refinamentos sucessivos, é limitada pelo número de dígitos significativos utilizados em precisão simples, entretanto, raramente é necessário utilizar precisão múltipla em (4.90) e (4.91); isso só ocorre quando o sistema linear é muito mal condicionado.

Na etapa (4.90) pode ser utilizado qualquer método de resolução apropriado, desde que o mesmo forneça uma precisão razoável, porém, o que torna prático o uso do refinamento é justamente a utilização de uma mes-

na matriz na etapa (4.90) de todas as iterações. O método de eliminação de Gauss possibilita essa simplificação, visto que a decomposição LU de A, determinada durante o cálculo de x_1, pode ser utilizada para esse fim. Com esse procedimento conseguimos obter um bom refinamento de x_1, através de pequeno trabalho adicional àquele requerido para a determinação da referida aproximação. (aproximadamente 25% em tempo de computador).

Consideremos, então, a convergência do processo que acabamos de descrever, com utilização da eliminação de Gauss com pivotamento parcial no cálculo de x_1 e aproveitamento das matrizes L e U, determinadas para esse fim, em (4.90). Uma análise completa da convergência do processo de refinamento na aritmética de vírgula flutuante é tão laboriosa algebricamente, que importantes aspectos do problema podem ficar mascarados pelos inúmeros detalhes algébricos. Por esse motivo, neste trabalho é descrita uma versão idealizada, apresentada em Forsythe (7, p. 109-113), na qual se supõe que a ocorrência dos erros de arredondamento se dá apenas no cálculo das correções a partir dos resíduos, ou seja, na etapa (4.90). Essa hipótese é razoável no caso de ser utilizada a acumulação de produtos escalares em (4.89) o que, como vimos anteriormente, é normalmente empregado na prática. Uma análise completa da convergência em questão, sem o uso dessa hipótese simplificadora, é efetuada por Moler (17).

A seguir, apresentaremos a versão adotada. A aplicação de (4.66) a (4.90) fornece:

$$(A+\delta A)\, c_m = r_m \,, \tag{4.95}$$

em que δA, que verifica (4.56) e (4.64), depende de x_m, porém a delimitação de $\|\delta A\|_\infty$ não depende do mesmo.

Para evidenciar a dependência em m e simplificar as expressões que serão obtidas no decorrer da análise, podemos escrever (4.95) na forma

$$A\left(I+F_m\right)c_m = r_m \quad , \text{ com} \tag{4.96}$$

$$A F_m = \delta A \quad \text{e, portanto,}$$

$$F_m = A^{-1}\delta A \tag{4.97}$$

As matrizes F_m, definidas acima, determinarão, de maneira completa, o comportamento da versão idealizada em questão. Isso será melhor entendido através do resultado que se segue.

(4.98) Teorema. Sejam x_m $(m = 1,2,\cdots)$ as aproximações da solução do sistema (4.19), obtidas por sucessivas iterações segundo a forma descrita. Seja x^* a solução exata de (4.19), ou seja, $x^* = A^{-1}b$. Se ocorrer:

$$\|F_m\|_\infty \leq G < \frac{1}{2} \quad , \quad \forall m \ , \tag{4.99}$$

então a seqüência de aproximações $x_1, x_2, \cdots, x_m, \cdots$ converge para a solução exata de (4.19), ou seja, em virtude de (1.25), $\|x_m - x\|_\infty \to 0$, com $m \to \infty$.

Vejamos a demonstração de (4.98). A partir de (4.93) e (4.96) temos

$$A\left(I+F_m\right)c_m = b - A x_m$$

i, como A é não singular por hipótese,

$$\left(I+F_m\right)c_m = A^{-1}b - x_m = x^* - x_m$$

Por outro lado, por (4.91),

$$\left(I+F_m\right)x_{m+1} = F_m x_m + x^* \tag{4.100}$$

Subtraindo $(I+F_m)x^*$ de ambos os membros de (4.100), temos

$$(I+F_m)(x_{m+1}-x^*) = F_m(x_m-x^*). \tag{4.101}$$

Em virtude de (4.99) e de (1.43), podemos afirmar que $(I+F_m)$ é não singular e escrever (4.101) na forma

$$x_{m+1}-x^* = (I+F_m)^{-1}F_m(x_m-x^*) \tag{4.102}$$

A equação (4.102) é muito importante para o estudo da convergência do processo de refinamento, por evidenciar que o erro da $(m+1)$-ésima iteração é igual ao produto de uma certa matriz (desconhecida) pelo erro da m-ésima iteração.

Considerando a norma $\|\ \|_\infty$ para (4.102) e aplicando (1.44) e (4.99) à expressão resultante, obtemos

$$\|x_{m+1}-x^*\|_\infty \leq \frac{1}{1-\|F_m\|_\infty}\|F_m\|_\infty \|x_m-x^*\|_\infty$$

$$\|x_{m+1}-x^*\|_\infty \leq \frac{1}{1-\sigma}\,\sigma\,\|x_m-x^*\|_\infty \tag{4.103}$$

Seja

$$\tau = \frac{\sigma}{1-\sigma} \tag{4.104}$$

Então, (4.103) pode ser escrita

$$\|x_{m+1}-x^*\|_\infty \leq \tau\,\|x_m-x^*\|_\infty\ , \tag{4.105}$$

que fornece

$$\|x_m-x^*\|_\infty \leq \tau\|x_{m-1}-x^*\|_\infty \leq \tau^2\|x_{m-2}-x^*\|_\infty \leq \tag{4.106}$$

$$\leq \tau^3\|x_{m-3}-x^*\|_\infty \leq \cdots \leq \tau^{m-1}\|x_1-x^*\|_\infty$$

Fazendo $x_0 = 0$, ou seja, considerando $m=0$ em (4.91), o que corresponde a tomar x_1 igual à aproximação de (4.19), determinada pela eliminação de Gauss com pivotamento parcial, obtemos, de (4.106),

$$\|x_1 - x^*\|_\infty \leq \tau \|x_0 - x\|_\infty = \tau \|x^*\|_\infty \tag{4.107}$$

Portanto, (4.106) e (4.107) nos dão:

$$\|x_m - x^*\|_\infty \leq \tau^m \|x^*\|_\infty \tag{4.108}$$

Como, em virtude de (4.99) e (4.104), $\tau < 1$, então, quando $m \to \infty$, $\tau^m \to 0$, e, por (4.108), $\|x_m - x^*\|_\infty \to 0$, o que prova (4.98).

O teorema (4.98) afirma que é possível tornar $\|x_m - x^*\|_\infty$ arbitrariamente pequeno, desde que efetuemos um número conveniente de iterações. Facilmente, porém, se pode concluir que isso é impossível, visto que as componentes de x_m devem ter precisão simples, enquanto o mesmo não ocorre, de maneira geral, com as de x^*. Através de alguma reflexão sobre o problema, pode-se, porém, entender que essa dificuldade decorre, principalmente, da suposição de que (4.91) se verifica exatamente, o que não ocorre na prática, quando, nessa etapa, cometemos um erro de arredondamento em cada componente. Uma análise mais detalhada da convergência do processo de refinamento, com a consideração dos erros nas três etapas (4.89)-(4.91), revela que o erro, em vez de tender a zero, torna-se pequeno.

Consideremos o refinamento da aproximação determinada em (4.71). Os cálculos necessários para a realização das etapas (4.89) foram realizados com acumulação de produto escalar.

(4.109) Exemplo. Consideremos o sistema $Ax = b$, para o qual

(4.110)
$$A = \begin{pmatrix} 0,6900 & 0,2600 & 0,1200 \\ 0,1500 & 0,4500 & 0,2700 \\ 0,3100 & 0,4900 & 0,1500 \end{pmatrix} \qquad b = \begin{pmatrix} 0,6700 \\ 0,2400 \\ 0,1200 \end{pmatrix}$$

e determinemos o refinamento da aproximação

$$x_1 = \begin{pmatrix} 0,10 \times 10 \\ -0,11 \times 10 \\ 0,21 \times 10 \end{pmatrix}, \text{ obtida em (4.71)}.$$

A eliminação de Gauss com pivotamento parcial fornece a seguinte decomposição triangular de A

(4.111)
$$L = \begin{pmatrix} 0,1000 \times 10 & 0,0000 & 0,0000 \\ 0,2174 & 0,1000 \times 10 & 0,0000 \\ 0,4493 & 0,9484 & 0,1000 \times 10 \end{pmatrix}$$

(4.112)
$$U = \begin{pmatrix} 0,6900 & 0,2600 & 0,1200 \\ 0,0000 & 0,3935 & 0,2439 \\ 0,0000 & 0,0000 & -0,1352 \end{pmatrix}$$

Utilizando como x_1 a aproximação determinada em (4.71)

(4.113)
$$x_1 = \begin{pmatrix} 0,10 \times 10 \\ -0,11 \times 10 \\ 0,21 \times 10 \end{pmatrix}$$

e utilizando para a realização das diversas etapas (4.90) as matrizes L e U dadas, respectivamente, em (4.111) e (4.112) temos:

$$r_1 = \begin{pmatrix} 0,14 \times 10^{-1} \\ 0,18 \times 10^{-1} \\ 0,34 \times 10^{-1} \end{pmatrix} \qquad c_1 = \begin{pmatrix} 0,00 \\ 0,10 \\ -0,10 \end{pmatrix} \qquad x_2 = \begin{pmatrix} 0,10 \times 10 \\ -0,10 \times 10 \\ 2,00 \times 10 \end{pmatrix}$$

(4.114) $$\dot{x}_2 = \begin{pmatrix} 0,00 \\ 0,00 \\ 0,00 \end{pmatrix}$$

Como pode ser observado através de (4.114), o processo convergiu rapidamente.

— ——— // ———

BIBLIOGRAFIA

1. ALBASINY, Ernest L. Error in digital solution of linear problems. In: ADVANCED SEMINAR CONDUCTED BY THE MATHEMATICS RESEARCH CENTER, University of Wisconsin, oct. 5-7, 1964. Error in digital computation, - edited by Louis B. Rall. New York, John Wiley /c1965/ v.1, p.131 - 184.

2. BARROS, Ivan de Queiroz. Elementos de análise numérica. Poços de Caldas /IMPA/ 1969. 139p. Trab. apres. no 7º Colóquio Brasileiro de - Matemática.

3. ____. Algebra linear. In: ____. Métodos numéricos. /Campinas, UEC. Instituto de Matemática, Estatística e Ciência de Computação/ 1970. v.1, p.1-121. Mimeografado.

4. BARTLE, Robert G. The elements of real analysis. New York, John Wiley /c1964/ p.58-69.

5. FADDEEVA, V.N. Computational methods of linear algebra. New York, Dover /c1959/ p.1-145.

6. FORSYTHE, Alexandra I.; KEENAN, Thomas A.; ORGANICK, Elliot I.; STEINBERG, Warren. Computer science: a first course. New York, John Wiley - /c1969/ p.3-232.

7. FORSYTHE, George E. & MOLER, Cleve B. Computer solution of linear algebraic systems. Englewood Cliffs, Prentice-Hall /c1967/ p.2-131.

8. FOX, L. An introduction to numerical linear algebra. Oxford, Clarendon Press /c1964/ p.1-214.

9. FOX, L. & MAYERS, D.F. Computing methods for scientists and engineers. Oxford, Clarendon Press, 1968. p.1-121.

10. FRÖBERG, Carl-Erik. Introduction to numerical analysis. Reading, Addison-Wesley /c1965/ cap.1, p.1-14; cap.3-6, 44-125.

11. GANTMACHER, F.R. The theory of matrices. New York, Chelsea /c1959/ v.1.

12. GEL'FAND, I.M. Lectures on linear algebra. 2.ed. New York, Interscience /c1961/ p.1-163. (Lectures on Linear Algebra, 9).

13. HENRICI, Peter. Elements of nurerical analysis. New York, John Wiley /c1964/ p.291-321.

14. HILDEBRAND, F.B. Introduction to numerical analysis. New York, McGraw-Hill, 1956, cap.1, p.1-34; cap.10, p.424-485.

15. LIPSCHUTZ, Seymour. Algebra linear. Rio de Janeiro, McGraw-Hill - /c1968/ 403p. (Coleção Schaum).

16. McCRACKEN, Daniel D. & DORN, William S. Numerical methods and FORTRAN programming with applications in engineering and science. New York, John Wiley /c1964/ cap.2, p.43-67; cap.8, p.226-283.

17. MOLER, Cleve B. Iterative refinement in floating point. J. ACM, New York, 14(2): 316-321, apr. 1967.

18. MONTEIRO, L.H. Jacy. Algebra linear. 5.ed. São Paulo, Nobel /1969/ v.1.

19. ____. ____. São Paulo, Nobel /1970/ v.2.

20. NATIONAL PHYSICAL LABORATORY. Modern computing methods. 2.ed. London, Her Majesty's Stationery Office, 1961. cap.1-5, p.1-52. (Notes on Applied Science, 16).

21. RALSTON, Anthony. A first course in numerical analysis. New York, Mc Graw-Hill /c1965/ cap.1, p.1-22; cap.9, p.394-463.

22. WILKINSON, J.H. The algebraic eigenvalue problem. Oxford, Clarendon Press /c1965/ p.1-264.

23. ____. Error analysis of direct methods of matrix inversion. J. ACM, New York, 8(3): 281-330, jul. 1961.

24. ____. Error analysis of floating point computation. Numerische Mat., Berlin, 2(5): 319-340, okt. 1960.

25. ____. Modern error analysis. SIAM Rev., Philadelphia, 13(4): 548-568, oct. 1971.

26. ____. Bounding errors in algebraic processes. Englewood Cliffs, Prentice-Hall /c1963/ 161p.