Conceitos Básicos da Teoria dos Erros

Teresa Diogo

Departamento de Matemática-Instituto Superior Técnico

1995

Introdução

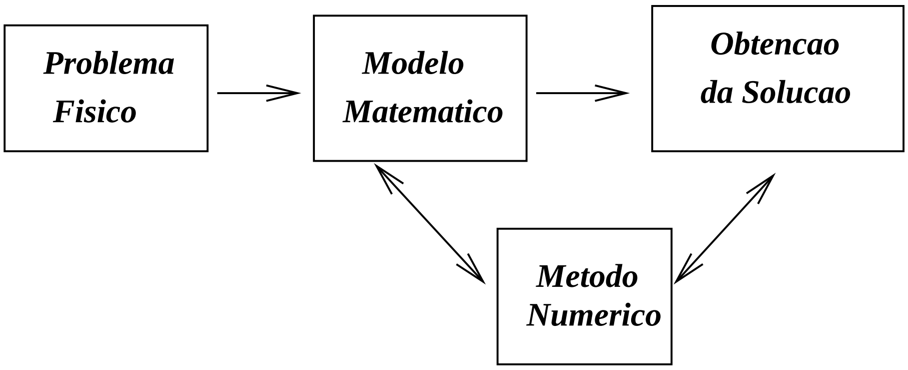
Neste curso de Análise Numérica, vamos estudar métodos que permitem resolver problemas científicos usando um computador. Podemos considerar três fases distintas no processo de resolução. A *primeira fase* consiste em expressar matematicamente o problema científico. Obtem-se assim um modelo matemático, que podemos definir como sendo uma expressão matemática que relaciona grandezas (físicas, de engenharia, de economia, etc). Um exemplo dum modelo matemático será a expressão da segunda lei de Newton do movimento

$$m\frac{d^2}{dt^2}x(t) = F(t) \tag{1}$$

que relaciona a posição, num instante t, dum corpo com massa m, com uma força aplicada F.

Depois de obtido o modelo matemático, a *segunda fase* consiste em escolher métodos numéricos que permitam obter, de forma eficiente, uma solução aproximada do problema em questão. A eficiência do método depende não só da precisão que nos pode garantir, mas também da facilidade com que pode ser implementado. Os métodos numéricos que aqui consideraremos serão expressos na forma de um algoritmo: uma sequência de instruções descrevendo um número finito de passos que devem ser executados por uma ordem especificada. Concentrar-nos-emos no estudo de métodos numéricos para resolver problemas no âmbito dos seguintes tópicos: equações não lineares, sistemas de equações lineares, interpolação, aproximação de funções no sentido dos mínimos quadrados e integração. Por exemplo, em relação à equação (1), é possível aproximar $x(t)$, usando técnicas de integração numérica que mais tarde descreveremos.

A *terceira e última fase* do processo que estamos a descrever consiste na implementação do algoritmo no computador. Obtemos assim uma solução numérica do problema inicial.



1 Erros

Para analisar a qualidade dum resultado numérico, devemos estar cientes das possíveis fontes de erro ao longo de todo o processo. Podemos considerar basicamente três tipos de erros: erros inerentes, erros do método e o erro computacional. Fazemos notar que é o efeito do erro total, o qual engloba todos os tipos de erros originados ao longo dum processo, que afecta a qualidade duma solução aproximada. Há certos problemas nos quais a um "pequeno" erro introduzido num certo passo dos cálculos, corresponde um erro "grande" na solução final. Um problema deste tipo diz-se *mal condicionado*.

1.1 Tipos de erros num processo de cálculo

Erros inerentes

Em geral, o modelo matemático não traduz exactamente a realidade. A fim de se conseguir um modelo que não seja demasiado complicado e difícil de tratar matematicamente, é muitas vezes necessário impor certas restrições idealistas. Os dados e parâmetros dum problema são muitas vezes resultados de medições experimentais e, portanto, afectados de alguma incerteza.

Podemos também mencionar a impossibilidade de representar exactamente certas constantes matemáticas (por exemplo teremos de substituir π por 3.1416 ou 3.141593, etc.).

Erros do método

Uma outra classe de erros resulta do uso de fórmulas que nos dão valores aproximados e não exactos. Por exemplo, consideremos a série de MacLaurin para representar e^x

$$e^x = 1 + x + \frac{x^2}{2!} + \cdots + \frac{x^n}{n!} + \cdots$$

Se quisermos aproximar o valor e^α, onde α é um número real, teremos de nos limitar a usar um número finito de termos da série

$$e^\alpha \simeq 1 + \alpha + \frac{\alpha^2}{2!} + \cdots + \frac{\alpha^k}{k!}$$

O erro que cometemos é chamado *erro de truncatura* e corresponde neste exemplo ao resto de ordem k da série.

Erro computacional

O erro computacional é devido ao facto de o computador usar apenas um número finito de dígitos para representar os números reais. A maioria dos números e dos resultados de operações aritméticas não podem ser representados exactamente no computador. São assim originados os *erros de arredondamento*. Na secção seguinte consideraremos este tópico em maior detalhe.

1.2 Erro, erro absoluto e erro relativo

Seja x o valor exacto duma grandeza real e seja $\tilde{x}$ uma aproximação de x, ou seja $x \simeq \tilde{x}$. Designa-se por e_x

$$e_x = x - \tilde{x} \tag{2}$$

o *erro de* $\tilde{x}$ *em relação a* x. Ao valor $|e_x|$ chama-se *erro absoluto* de $\tilde{x}$. Se $x \neq 0$, denota-se por δ_x

$$\delta_x = \frac{x - \tilde{x}}{x} \tag{3}$$

e chama-se a $|\delta_x|$ o *erro relativo de* $\tilde{x}$.

Ao produto 100 $|\delta_x|$ expresso em percentagem chama-se *percentagem de erro.*

NOTA

Atendendo a que

$$\frac{x - \tilde{x}}{\tilde{x}} = \frac{\delta_x}{1 - \delta_x} = \delta_x(1 + \delta_x + {\delta_x}^2 + \cdots) = \delta_x + {\delta_x}^2 + {\delta_x}^3 + \cdots,$$

com δ_x suficientemente pequeno, usa-se na prática

$$\delta_x = \frac{x - \tilde{x}}{\tilde{x}},$$

o que equivale a desprezar os termos de ordem superior a um no desenvolvimento acima.

Observamos que o erro absoluto pode não ser de grande utilidade para informar sobre a qualidade duma aproximação, se não se souber a ordem de grandeza da quantidade a medir. Basta dizer que um erro de um metro na medição da distância da Terra a Saturno seria óptimo, mas se um cirurgião cometesse um erro dessa ordem ao fazer uma incisão seria um desastre possívelmente até fatal ! Incluimos ainda seguinte exemplo.

Exemplo

Consideremos os números $x = 1/3$ *e* $y = 1/3000$, *e as aproximações* $\tilde{x} = 0.3333$ *e* $\tilde{y} = 0.0003$. *Tem-se que* $\tilde{x}$ *e* $\tilde{y}$ *são valores aproximados de* x *e* y *respectivamente, com o mesmo erro* $e_x = e_y = 0.0000333\cdots$. *Porém a percentagem de erro em* $\tilde{x}$ *é 0.01% e a percentagem de erro em* $\tilde{y}$ *é 10%.*

2 Representação dos números no computador

2.1 Bases

A base decimal é aquela com que estamos mais familiarizados. Todo o número inteiro se pode representar como uma combinação linear de potências de 10, com coeficientes variando entre 0 e 9. Por exemplo, o número 415 pode-se escrever na forma

$$\begin{aligned} 415 &= 400 + 10 + 5 \\ &= 4 \times 10^2 + 1 \times 10^1 + 5 \times 10^0. \end{aligned}$$

o que se representa simbolicamente por $(4\ 1\ 5)_{10}$ (na prática omite-se o 10). De modo geral, usa-se a notação

$$\sigma(b_m b_{m-1} \cdots b_1 b_0)_{10}$$

para representar o valor (inteiro)

$$\sigma(b_m 10^m + b_{m-1} 10^{m-1} + \cdots + b_1 10^1 + b_0 10^0),$$

com $\sigma \in \{+, -\}$ e $0 \leq b_i \leq 9, \quad b_m \neq 0$.

Os números fraccionários podem ser representados na base 10, usando o ponto decimal. Um número entre 0 e 1 pode ser escrito na forma

$$(0.a_1 a_2 \cdots a_n \cdots)_{10}$$

que representa o valor

$$a_1 10^{-1} + a_2 10^{-2} + \cdots + a_{n-1} 10^{-n+1} + a_n 10^{-n} + \cdots$$

Finalmente, um número real qualquer não nulo pode ser representado no sistema decimal por

$$\sigma(b_m b_{m-1} \cdots b_1 b_0.a_1 a_2 \cdots a_n \cdots)_{10}$$

Por exemplo, 12.34 representa

$$1 \times 10^1 + 2 \times 10^0 + 3 \times 10^{-1} + 4 \times 10^{-2}.$$

Outras bases, tais como 2, 12, 20 e 60 foram usadas por civilizações anteriores para representação dos números. De modo geral, um número real $X \neq 0$ pode ser representado numa base β por

$$X = \sigma(b_m \beta^m + b_{m-1} \beta^{m-1} + \cdots + b_1 \beta^1 + b_0 \beta^0 + a_1 \beta^{-1} + a_2 \beta^{-2} + \cdots + a_{n-1} \beta^{-n+1} + a_n \beta^{-n} + \cdots),$$

ou, simbolicamente, por

$$\sigma(b_m b_{m-1} \cdots b_1 b_0.a_1 a_2 \cdots a_{n-1} a_n \cdots)_\beta,$$

onde os b_i, a_i são inteiros compreendidos entre 0 e $\beta - 1$.

Exemplo

$$\begin{array}{rll} (101)_2 & \textit{representa} & 1 \times 2^2 + 1 \times 2^0 = 5 \\ (1100)_2 & \textit{representa} & 1 \times 2^3 + 1 \times 2^2 = 12 \\ (0.101)_2 & \textit{representa} & 1 \times 2^{-1} + 1 \times 2^{-3} = 0.625 \\ (0.000110011\cdots)_2 & \textit{representa} & 0.1 \end{array}$$

2.2 Sistemas de vírgula flutuante

Num computador, os números reais são em geral representados em vírgula flutuante na base $\beta = 2$. Outras bases usadas por computadores são $\beta = 8$ e $\beta = 16$. De modo geral, um *número de vírgula flutuante* na base β e com n dígitos é um número que pode ser escrito na forma

$$x = \sigma \ (0.a_1 a_2 \cdots a_n)_\beta \times \beta^t, \quad \sigma \in \{+, -\}, \tag{4}$$

onde $0 \leq a_i \leq \beta - 1$, $i = 1, 2, \cdots, n$, $t_1 \leq t \leq t_2$, sendo t, t_1 e t_2 inteiros.

A $0.a_1 a_2 \cdots a_n$ chama-se *mantissa* ou *parte fraccionária* e a t o *expoente*. Se $a_1 \neq 0$, o número diz-se *normalizado*. No que se segue consideraremos apenas números normalizados.

Usa-se a notação VF(β, n, t_1, t_2) para designar o conjunto de números da forma (4) e ainda o número $x = 0$.

2.3 Arredondamentos

Neste parágrafo definiremos duas regras para representar números reais no computador.

Consideremos o número real

$$x = \sigma(0.a_1 a_2 \cdots a_n a_{n+1} \cdots)_\beta \times \beta^t, \tag{5}$$

onde $a_1 \neq 0$ e β é uma base par.

Para representar este número em vírgula flutuante com n dígitos, podemos simplesmente desprezar os dígitos $a_{n+1} a_{n+2} \cdots$ Este processo é conhecido por *arredondamento por corte*. O número x passa a ser representado por

$$fl(x) = \sigma \ (0.a_1 a_2 \cdots a_n)_\beta \times \beta^t. \tag{6}$$

Um outro modo de representar x em vírgula flutuante é usando o chamado *arredondamento simétrico*. Se $0 \leq a_{n+1} < \beta/2$, procede-se como no arredondamento por corte. Se $\beta/2 \leq a_{n+1} < \beta$ soma-se β^{-n} ao número $(0.a_1 a_2 \cdots a_n)_\beta$ e normaliza-se. Note-se que, neste caso, pode haver mudança nos dígitos e no expoente (ver terceiro exemplo a seguir).

Resumindo

$$
\begin{aligned}
fl(x) &= \sigma(0.a_1a_2\cdots a_n)_\beta \times \beta^t & 0 \le a_{n+1} < \beta/2 \qquad (7)\\
fl(x) &= fl\left(\sigma((0.a_1a_2\cdots a_n)_\beta + \beta^{-n}) \times \beta^t\right) & \beta/2 \le a_{n+1} < \beta. \qquad (8)
\end{aligned}
$$

Exemplo *Suponhamos $\beta = 10$ e $n = 2$*

$$
fl(\frac{2}{3}) = \begin{cases} 0.67 \times 10^0 & \textit{arred. simétrico} \\ 0.66 \times 10^0 & \textit{arred. por corte} \end{cases}
$$

$$
fl(-838) = \begin{cases} -0.84 \times 10^3 & \textit{arred. simétrico} \\ -0.83 \times 10^3 & \textit{arred. por corte} \end{cases}
$$

$$
fl(99.5) = \begin{cases} 0.10 \times 10^3 & \textit{arred. simétrico} \\ 0.99 \times 10^2 & \textit{arred. por corte} \end{cases}
$$

2.4 Erros de arredondamento

O erro de arredondamento introduzido quando x é substituído pelo seu representante $fl(x)$ é a diferença $x - fl(x)$. Consideremos primeiramente o caso dum sistema decimal.

Seja $x = \sigma(0.a_1a_2\cdots a_na_{n+1}\cdots) \times 10^t$, e suponhamos que $fl(x)$ é obtido por arredondamento simétrico.

Seja $a_{n+1} \ge 5$. Usando (8), tem-se

$$
\begin{aligned}
|x - fl(x)| &= |(0.a_1\cdots a_na_{n+1}\cdots) \times 10^t - ((0.a_1\cdots a_n) + (0.0\cdots 01)) \times 10^t| \\
&= |((0.0\cdots 0a_{n+1}a_{n+2}\cdots) - 10^{-n}) \times 10^t| \\
&= |((0.a_{n+1}a_{n+2}\cdots) - 1) \times 10^{t-n}| \\
&\le 0.5 \times 10^{t-n}.
\end{aligned}
$$

Se $a_{n+1} < 5$, então resulta de (7)

$$
\begin{aligned}
|x - fl(x)| &= |(0.a_1\cdots a_na_{n+1}\cdots) \times 10^t - (0.a_1\cdots a_n) \times 10^t| \\
&= (0.0\cdots 0a_{n+1}a_{n+2}\cdots) \times 10^t \\
&= (0.a_{n+1}a_{n+2}\cdots) \times 10^{-n} \times 10^t \\
&\le 0.5 \times 10^{t-n}.
\end{aligned}
$$

Em qualquer dos casos, tem-se o seguinte majorante do erro absoluto de arredondamento simétrico

$$
|x - fl(x)| \le 0.5 \times 10^{t-n}. \qquad (9)
$$

Para obter um majorante do erro relativo, notemos que $a_1 \ne 0$ implica $|x| \ge 0.1 \times 10^t$. Então

$$
\frac{|x - fl(x)|}{|x|} \le \frac{0.5 \times 10^{t-n}}{0.1 \times 10^t} = 5 \times 10^{-n} = 0.5 \times 10^{1-n}. \qquad (10)
$$

De modo análogo se prova que, no caso de arredondamento por corte,

$$\begin{aligned} |x - fl(x)| &\leq 10^{t-n} && (11)\\ \frac{|x - fl(x)|}{|x|} &\leq 10^{1-n}. && (12) \end{aligned}$$

A análise efectuada para um sistema VF decimal estende-se ao caso duma base β par. Seja x um número da forma (5). No caso de arredondamento por corte, tem-se

$$\begin{aligned} |x - fl(x)| &\leq \beta^{t-n} && (13)\\ \frac{|x - fl(x)|}{|x|} &\leq \beta^{1-n}. && (14) \end{aligned}$$

No caso de arredondamento simétrico, obtém-se

$$\begin{aligned} |x - fl(x)| &\leq (1/2)\ \beta^{t-n} && (15)\\ \frac{|x - fl(x)|}{|x|} &\leq (1/2)\ \beta^{1-n}. && (16) \end{aligned}$$

É interessante notar que nenhum dos majorantes dos erros relativos depende de x, mas apenas da base β, do número de dígitos n usado pelo computador e, é claro, do processo de arredondamento utilizado.

Dado um sistema VF(β, n, t_1, t_2), define-se

$$U = \begin{cases} \beta^{1-n} & \text{arred. por corte} \\ (1/2)\ \beta^{1-n} & \text{arred. simétrico.} \end{cases}$$

A U chama-se *unidade de arredondamento* do sistema.

Assim, se $fl(x)$ é a representação em VF dum número real x, o seu erro relativo não excede U. É por isso fundamental conhecer a unidade de arredondamento do sistema de vírgula flutuante do computador que se usa.

Exemplo

Num computador com o sistema VF(16,6,-64,63), a unidade de arredondamento simétrico é

$$(1/2)16^{1-6} \simeq 0.476837 \times 10^{-7}.$$

2.5 Algarismos significativos

Finalizamos esta secção com o conceito de algarismo significativo. Seja

$$\tilde{x} = \sigma\ (0.a_1 a_2 \cdots a_n) \times 10^t$$

uma aproximação de x pertencente a VF(10, n, t_1, t_2).

Definição

Diz-se que o *algarismo a_i de $\tilde{x}$ é significativo* se

$$|e_x| = |x - \tilde{x}| \leq \frac{1}{2} 10^{t-i}. \tag{17}$$

Concluimos que se

$$|e_x| = |x - \tilde{x}| \leq \frac{1}{2} 10^{t-n}, \tag{18}$$

os n algarismos de $\tilde{x}$ são significativos.

Exercício

Em vírgula flutuante com 4 dígitos, sejam $x = 0.4537 \times 10^4$ e $\tilde{x} = 0.4501 \times 10^4$. Determine o número de algarismos significativos de $\tilde{x}$.

Tem-se que

$$|e_x| = 0.0036 \times 10^4 = (0.36 \times 10^{-2}) \times 10^4,$$

logo verifica-se (17) com $i = 1, 2$. Como além disso

$$|e_x| > (0.5 \times 10^{-3}) \times 10^4,$$

apenas os algarismos 4 e 5 de $\tilde{x}$ são significativos.

3 Propagação de erros

Uma questão importante em Análise Numérica é a do modo como um erro num certo passo dos cálculos se vai propagar. Interessa-nos saber se o efeito desse erro se tornará maior ou menor, à medida que vão sendo efectuadas as restantes operações.

3.1 Cálculo de valores funcionais usando dados com erros

Seja $f : D \subset \mathcal{R} \to \mathcal{R}$ e seja $\tilde{x}$ uma aproximação de x que verifica $x = \tilde{x} + e_x$. Se pretendemos calcular o valor de $f(x)$ mas apenas conhecemos $\tilde{x}$, então aproximamos $f(x)$ por $f(\tilde{x})$. Em geral, $f(\tilde{x})$ não coincidirá com $f(x)$. Por analogia com as definições da Secção 1.2, designa-se por $e_f = f(x) - f(\tilde{x})$ o erro de $f(\tilde{x})$. Admitindo que todas as operações indicadas na expressão de f podem ser efectuadas exactamente, vamos determinar um limite superior para o erro absoluto $|e_f| = |f(x) - f(\tilde{x})|$. Teremos, assim, uma indicação do efeito do erro e_x.

Se $f'(x)$ existe e é contínua, tem-se, pelo Teorema do valor médio

$$f(x) - f(\tilde{x}) = f'(\xi)(x - \tilde{x}), \tag{19}$$

com $\quad \xi \in I = int(x, \tilde{x}) = (min\{x, \tilde{x}\}, max\{x, \tilde{x}\})$, donde

$$|e_f| = |f(x) - f(\tilde{x})| = |f'(\xi)(x - \tilde{x})| \leq \max_{t \in I} |f'(t)||e_x|. \tag{20}$$

Na prática conhece-se apenas um majorante de e_x. Se for η esse majorante, tem-se

$$|e_x| \leq \eta \Leftrightarrow \tilde{x} - \eta \leq x \leq \tilde{x} + \eta$$

e a fórmula a usar será

$$|e_f| \leq \max_{t \in [\tilde{x} - \eta, \tilde{x} + \eta]} |f'(t)|\eta. \tag{21}$$

Note que

$$I = int(x, \tilde{x}) \subseteq [\tilde{x} - \eta, \tilde{x} + \eta] \Rightarrow \max_{t \in I} |f'(t)| \leq \max_{t \in [\tilde{x} - \eta, \tilde{x} + \eta]} |f'(t)|.$$

A fórmula (20) diz respeito ao erro absoluto de f. Pode obter-se um majorante para o erro relativo de f, do modo seguinte

$$|\delta_f| = |\frac{e_f}{f(x)}| \leq \max_{t \in I} |f'(t)| \frac{|e_x|}{|x|} \frac{|x|}{|f(x)|} = \max_{t \in I} |f'(t)| \frac{|x|}{|f(x)|} |\delta_x|$$

As fórmulas de majoração deduzidas têm o inconveniente de exigir o conhecimento dum majorante para a derivada $f'(t)$, o que pode ser difícil.

Exercício. *Suponha que estamos a trabalhar num sistema decimal com 4 dígitos. Determine um majorante do erro (absoluto) que se comete ao aproximar o valor $\sqrt{x}$ por $\sqrt{\tilde{x}}$, onde $\tilde{x} = 0.4000 \times 10$ é um valor aproximado de x com 3 algarismos significativos.*

Resulta de (21) que

$$|e_f| \leq \max_{t \in [\tilde{x}-\eta, \tilde{x}+\eta]} \frac{1}{2\sqrt{t}}\, \eta. \tag{22}$$

Como $\tilde{x}$ tem 3 algarismos significativos, $|\eta| \leq 0.5 \times 10^{-2}$ e então

$$|e_f| \leq \frac{1}{2\sqrt{3.995}}\, 0.5 \times 10^{-2} \simeq 0.12508 \times 10^{-2}.$$

Se por exemplo $x = 4.003579$, tem-se

$$\sqrt{x} - \sqrt{\tilde{x}} \simeq 0.89455 \times 10^{-3},$$

donde concluimos que, neste caso, a fórmula (22) dá uma estimativa do erro absoluto por excesso.

Consideremos agora o seguinte problema mais geral. Suponhamos que se pretende calcular o valor de $f(x, y)$ usando os valores aproximados $\tilde{x}, \tilde{y}$ em vez de x, y, respectivamente. Supondo que f'_x e f'_y são contínuas, tem-se

$$f(x, y) = f(\tilde{x}, \tilde{y}) + (x - \tilde{x}) f'_x(\xi_1, \xi_2) + (y - \tilde{y}) f'_y(\xi_1, \xi_2),$$

onde $\xi_1 \in int(x, \tilde{x})$ e $\xi_2 \in int(y, \tilde{y})$. Se $|e_x| \leq \eta_x$ e $|e_y| \leq \eta_y$ e, por outro lado, $max|f'_x| \leq M_x$ e $max|f'_y| \leq M_y$ no intervalo $[\tilde{x} - \eta_x, \tilde{x} + \eta_x] \times [\tilde{y} - \eta_y, \tilde{y} + \eta_y]$, obtém-se a seguinte fórmula

$$|e_f| = |f(x, y) - f(\tilde{x}, \tilde{y})| \leq M_x \eta_x + M_y \eta_y. \tag{23}$$

A fórmula acima pode ser generalizada para funções de n variáveis.

3.2 Propagação dos erros nas operações aritméticas

É de particular importância estudar o modo como um erro se propaga numa sequência de operações aritméticas. Fórmulas para os erros nas quatro operações aritméticas podem ser obtidas através de (23). Consideraremos aqui um processo alternativo para deduzir as mesmas.

Adição

Sejam $\tilde{x}$ e $\tilde{y}$ duas aproximações dos valores exactos x e y, com erros respectivos e_x e e_y. Tem-se

$$x + y = \tilde{x} + e_x + \tilde{y} + e_y = (\tilde{x} + \tilde{y}) + (e_x + e_y).$$

O erro na soma, que denotaremos por e_{x+y}, verifica

$$e_{x+y} = e_x + e_y. \tag{24}$$

Subtracção

De maneira análoga

$$e_{x-y} = e_x - e_y. \tag{25}$$

Multiplicação

Neste caso tem-se

$$\begin{aligned} x.y &= (\tilde{x} + e_x)\,(\tilde{y} + e_y) \\ &= \tilde{x}\tilde{y} + \tilde{x}e_y + \tilde{y}e_x + e_x e_y. \end{aligned}$$

Admitindo que o produto $e_x e_y$ é desprezável em relação aos outros termos, resulta

$$x.y \simeq \tilde{x}\tilde{y} + \tilde{x}e_y + \tilde{y}e_x,$$

donde

$$e_{x.y} \simeq \tilde{x}e_y + \tilde{y}e_x. \tag{26}$$

Divisão

Tem-se sucessivamente

$$\begin{aligned} x/y &= \frac{\tilde{x} + e_x}{\tilde{y} + e_y} \\ &= \frac{\tilde{x} + e_x}{\tilde{y}}\left(\frac{1}{1 + e_y/\tilde{y}}\right). \end{aligned}$$

Admitindo que $|e_y| < |\tilde{y}|$, podemos escrever

$$\frac{1}{1 + e_y/\tilde{y}} = 1 - e_y/\tilde{y} + (e_y/\tilde{y})^2 - (e_y/\tilde{y})^3 + \cdots$$

Substituindo acima e desprezando os termos que involvem potências de $|e_y/\tilde{y}|$ superiores a um, obtem-se a estimativa

$$\frac{x}{y} \simeq \frac{\tilde{x}}{\tilde{y}} + \frac{e_x}{\tilde{y}} - \frac{\tilde{x}}{(\tilde{y})^2}\ e_y.$$

Daqui resulta

$$e_{x/y} \simeq \frac{1}{\tilde{y}}\ e_x - \frac{\tilde{x}}{(\tilde{y})^2}\ e_y. \tag{27}$$

Note-se que as fórmulas (24)-(27) dão-nos uma indicação sobre o erro no resultado de cada uma das operações aritméticas, admitindo que *não houve arredondamento depois de efectuada a operação.*

Exercício

Determine o número de algarismos significativos que se pode garantir para $\tilde{x}+\tilde{y}$ como aproximação de $x+y$, sabendo que $\tilde{x}=0.425\times 10^3$ e $\tilde{y}=0.326\times 10^3$ têm os três algarismos significativos (despreze os erros de arredondamento).

Resulta de (24) e da definição de algarismo significativo (com $t=3$) que

$$\begin{aligned} |e_{x+y}| &\le |e_x|+|e_y| \\ &\le (0.5\times 10^{-3}+0.5\times 10^{-3})\times 10^3 \\ &\le (0.1\times 10^{-2})\times 10^3 = 1. \end{aligned}$$

Concluimos poder garantir dois algarismos significativos para

$$\tilde{x}+\tilde{y}=0.751\times 10^3.$$

As fórmulas (24)-(27) dizem respeito à propagação dos erros nas quatro operações aritméticas. A partir destas obtêm-se as seguintes estimativas para as quantidades $\delta_{x\triangle y}=e_{x\triangle y}/(x\triangle y)$ ($\triangle$ designa uma operação aritmética).

$$\delta_{x+y} = \frac{e_{x+y}}{x+y} \simeq \frac{\tilde{x}}{\tilde{x}+\tilde{y}}\,\delta_x + \frac{\tilde{y}}{\tilde{x}+\tilde{y}}\,\delta_y \tag{28}$$

$$\delta_{x-y} = \frac{e_{x-y}}{x-y} \simeq \frac{\tilde{x}}{\tilde{x}-\tilde{y}}\,\delta_x - \frac{\tilde{y}}{\tilde{x}-\tilde{y}}\,\delta_y \tag{29}$$

$$\delta_{x.y} = \frac{e_{x.y}}{xy} \simeq \delta_x+\delta_y \tag{30}$$

$$\delta_{x/y} = \frac{e_{x/y}}{x/y} \simeq \delta_x-\delta_y. \tag{31}$$

Estas estimativas foram calculadas usando $\tilde{x}$ e $\tilde{y}$ em vez de x e y (ver Nota, parágrafo 1.2).

3.3 Propagação dos erros numa sequência de operações aritméticas

É importante compreender claramente o significado das fórmulas que deduzimos na secção anterior. Começamos com dois valores aproximados $\tilde{x}$ e $\tilde{y}$, que contêm erros absolutos e_x e e_y. Estes erros podem ser de qualquer dos três tipos considerados na secção 1.1. Nomeadamente, $\tilde{x}$ e $\tilde{y}$ podem ter sido obtidos experimentalmente, contendo por isso erros inerentes; podem ser o resultado de um processo de cálculo a que está associado um certo erro de truncatura; podem ainda ser o resultado de operações aritméticas anteriores e conter erros de arredondamento. Finalmente, $\tilde{x}$ e $\tilde{y}$ poderão conter uma combinação dos três tipos de erros.

As fórmulas (28)-(31) dão-nos uma indicação sobre o erro no resultado de cada uma das operações aritméticas, como funções de $\tilde{x}$ e $\tilde{y}$, δ_x e δ_y, admitindo que *não houve arredondamento depois de efectuada a operação.* Se no entanto quisermos saber como o erro deste resultado se propaga em ainda outras operações aritméticas subsequentes, teremos de adicionar o *erro de arredondamento* explicitamente.

Exercício

Seja $v = xy + z$ e sejam os números de vírgula flutuante $\tilde{x}$, $\tilde{y}$ e $\tilde{z}$ valores aproximados de x, y e z, com erros relativos $|\delta_x|$, $|\delta_y|$ e $|\delta_z|$, respectivamente. Determine uma estimativa do erro relativo no cálculo de v, num computador com unidade de arredondamento U, e usando os valores aproximados.

O cálculo de v pode ser efectuado usando o seguinte algoritmo.

$$\begin{aligned} u &= xy \\ v &= u + z. \end{aligned} \tag{32}$$

O computador começa por efectuar a multiplicação de $\tilde{x}$ por $\tilde{y}$.

Em seguida, o resultado é normalizado. Ou seja, aquilo que se obtém é um valor $\tilde{u}$ dado por

$$\tilde{u} = fl(\tilde{x}\tilde{y}).$$

Usando (30), tem-se a seguinte estimativa

$$\delta_u \simeq \delta_x + \delta_y + \delta_{arr1}, \tag{33}$$

onde $|\delta_{arr1}|$ é o *erro relativo de arredondamento* correspondente à normalização do produto. Em seguida, é efectuada a adição de $\tilde{u}$ com $\tilde{z}$ e o resultado é normalizado. De modo análogo, resulta de (28) que

$$\delta_v \simeq \frac{\tilde{u}}{\tilde{u} + \tilde{z}}\, \delta_u + \frac{\tilde{z}}{\tilde{u} + \tilde{z}}\, \delta_z + \delta_{arr2}, \tag{34}$$

onde $|\delta_{arr2}|$ é o *erro relativo de arredondamento* associado à normalização da soma.

Conjugando (33) e (34), obtém-se

$$\delta_{xy+z} \simeq \frac{\tilde{u}}{\tilde{u} + \tilde{z}}\delta_x + \frac{\tilde{u}}{\tilde{u} + \tilde{z}}\delta_y + \frac{\tilde{z}}{\tilde{u} + \tilde{z}}\delta_z + \frac{\tilde{u}}{\tilde{u} + \tilde{z}}\delta_{arr1} + \delta_{arr2},$$

com $|\delta_{arr1}| \leq U$ e $|\delta_{arr2}| \leq U$. Da equação acima pode obter-se uma estimativa para $|\delta_{xy+z}|$. Por simplicidade de escrita, costuma-se omitir o tile. Este procedimento é usual, sempre que for claro, pelo sentido do texto, quando nos referimos a um valor exacto ou a uma sua aproximação.

Bibliografia

S. de Conte, C. de Boor, Elementary Numerical Analysis, Mc-Graw-Hill, 1983.
Johnson, L. W. & Riess, R. D., *Numerical Analysis*, Addison-Wesley, 1977.
Valença, M. R., *Métodos Numéricos*, Instituto Nacional de Investigação Científica 1990.