Representação de Números no Computador e Erros

Análise Numérica

Patrícia Ribeiro e Artur M. C. Brito da Cruz

Escola Superior de Tecnologia
Instituto Politécnico de Setúbal
2015/2016 [1]

[1] versão 20 de Setembro de 2017

Conteúdo

1 Introdução

A análise numérica tem por objectivo dar respostas numéricas, i.e. soluções com números, a problemas físicos. Por exemplo, em matemática uma solução pode ser dada por

$$x = \sqrt{2}\pi$$

mas para um computador esta solução não é adequada, pois um computador apenas consegue manipular um número finito de símbolos ou operações.

Além disso, quando se usa um computador ou uma máquina obtemos diversos tipos de erros:

- Erros de aproximação do modelo matemático associado a um problema real;
- Erros nos dados (obtidos experimentalmente);
- Erros da precisão finita na representação de números reais.

2 Representação dos números reais

2.1 Representação em diferentes bases

A notação usual dos números reais é em base $b = 10$. Por exemplo

$$49.75 = 4 \times 10 + 9 \times 10^0 + 7 \times 10^{-1} + 5 \times 10^{-2}.$$

Frequentemente é usada a base $b = 2$. O número anterior escrito em base 2 terá que ser calculado em dois passos. Primeiro calcula-se a representação da parte inteira

$$\begin{array}{rl}
49 & \big|\,2 \\
\hline
\mathbf{1}\quad 24 & \big|\,2 \\
\hline
\mathbf{0}\quad 12 & \big|\,2 \\
\hline
\mathbf{0}\quad 6 & \big|\,2 \\
\hline
\mathbf{0}\quad 3 & \big|\,2 \\
\hline
\mathbf{1}\quad 1 & \big|\,2 \\
\hline
\mathbf{1}\quad 0 &
\end{array}$$

e conclui-se que $(49)_{10} = (110001)_2$ onde

$$(110001)_2 = 1 \times 2^5 + 1 \times 2^4 + 1 \times 2^0.$$

A parte fraccionária é obtida da seguinte maneira

$$\begin{aligned}
0.75 \times 2 &= \mathbf{1}.5 \\
0.5 \times 2 &= \mathbf{1}.0
\end{aligned}$$

e tem-se que $(0.75)_{10} = (0.11)_2$. onde

$$(0.11)_2 \;=\; 1 \times 2^{-1} + 1 \times 2^{-2}$$

Finalmente conclui-se que

$$49.75 = (110001.11)_2 \,.$$

Exemplo. *1.* $(101101)_2 = (45)_{10}$

2. $(176)_{10} = (10110000)_2$

3. $(1A0F)_{16} = (6671)_{10}$

4. $(0.1)_{10} = (0.00011\ldots)_2$

5. $(3.8)_{10} = (11.11001100\ldots)_2$

6. $(0.110)_2 = (0.75)_{10}$

7. $(341.9424)_{10} = (2331.4324)_5$

2.2 Representação em Ponto Flutuante

Como nos computadores usa-se um número finito de números, teremos que recorrer a sistemas que façam arredondamentos.

Quando se representam números muitos grandes ou muito pequenos, recorre-se usualmente à **notação científica**. Dado um número x, em notação científica, este número representa-se por

$$x = \pm m b^e$$

onde $m \geqslant 0$ é um número real designado por mantissa, $b \geqslant 2$ é um número natural designado por base e e é um número inteiro designado por expoente.

Por exemplo, o número

$$32.567$$

admite a representação na base 10

$$+3.2567 \times 10^1.$$

Infelizmente esta representação não é única pois o número 32.567 também tem a representação

$$+3256.7 \times 10^{-2}.$$

Para resolver este problema, usa-se a **notação científica normalizada** onde se impõem as seguintes convenções para a mantissa

$$\begin{cases} m = 0 & \text{se} \quad x = 0 \\ b^{-1} \leqslant m < 1 & \text{se} \quad x \neq 0. \end{cases}$$

Neste caso, o número 32.567 apenas tem a representação

$$+0.32567 \times 10^2.$$

Note-se que a representação científica normalizada continua a não resolver o problema da unicidade na representação pois

$$0 \times 10^1 = 0 \times 10^2 = \ldots$$

A notação científica não pode ser implementada num computador pois teríamos que ter infinitos dígitos para a mantissa e para o expoente. Quando restringimos a mantissa a um número finito p de dígitos e o expoente a um número finito q de dígitos, obtém-se a designada **representação em ponto flutuante**. O sistema de representação em ponto flutuante normalizado $FP(b, p, e_{min}, e_{max})$ de base b, mantissa até p dígitos

$$m = (0.d_{-1}d_{-2}...d_{-p})_b$$

e o expoente e tal que $e_{min} \leqslant e \leqslant e_{max}$, contém todos os números reais da forma

$$x = 0 \text{ se } d_{-1} = 0$$

ou

$$\begin{aligned} x &= \pm m b^e \\ &= \pm (0.d_{-1}d_{-2}\ldots d_{-p})_b \times b^e \\ &= \pm \left(d_{-1} \cdot b^{-1} + d_{-2} \cdot b^{-2} + \ldots + d_{-p} \cdot b^{-p}\right) \times b^e \end{aligned}$$

se $d_{-1} \neq 0$.

Exemplo. *O sistema de ponto flutuante* $FP(10, 6, -2, 2)$ *tem base* 10*, mantissa até* 6 *dígitos. Este sistema inclui o zero e todos os números da forma*

$$\pm 0.d_{-1}d_{-2}\ldots d_{-6} \times 10^e$$

onde $0 \leqslant d_{-i} \leqslant 9$*,* $i = 1, \ldots, 6$ *e* $d_{-1} \neq 0$*. Por exemplo, o número* 32.567 *neste sistema representa-se por*

$$+0.32567 \times 10^2.$$

Note-se que neste sistema o maior número representável é $+0.999999 \times 10^2 = 99.9999$ *e o menor número positivo representável é* $+0.1 \times 10^{-2} = 0.001$*. O número* $100 = +0.1 \times 10^3$ *não é representável neste sistema e nesta situação diz-se que ocorreu um* ***Overflow****. Analogamente o número* $+0.1 \times 10^{-3}$ *também não é representável neste sistema e diz-se que ocorreu um* ***Underflow****. O número* $\pi = 3.1415926\ldots$ *não possui representação exacta neste sistema pois tem mais do que 6 dígitos.*

Exemplo. *O sistema de ponto flutuante* $FP(2, 3, -1, 2)$ *tem base* 2*, mantissa até* 3 *algarismos que se chamam bits e o expoente varia entre* -1 *e* 2*. Este sistema inclui o zero e todos os números da forma*

$$x = \pm \left(d_{-1} \cdot 2^{-1} + d_{-2} \cdot 2^{-2} + d_{-3} \cdot 2^{-3}\right) \times 2^e$$

ou seja,

$$\begin{aligned} FP(2,3,-1,2) &= \{0, \pm\frac{1}{4}, \pm\frac{5}{16}, \pm\frac{3}{8}, \pm\frac{7}{16}, \pm\frac{1}{2}, \pm\frac{5}{8}, \pm\frac{3}{4}, \pm\frac{7}{8}, \\ &\quad \pm 1, \pm\frac{5}{4}, \pm\frac{3}{2}, \pm\frac{7}{4}, \pm 2, \pm\frac{5}{2}, \pm 3, \pm\frac{7}{2}\} \\ &= \{0, \pm 0.25, \pm 0.3125, \pm 0.375, \pm 0.4375, \pm 0.5, \pm 0.625, \pm 0.75, \pm 0.875, \\ &\quad \pm 1, \pm 1.25, \pm 1.5, \pm 1.75, \pm 2, \pm 2.5, \pm 3, \pm 3.5\}. \end{aligned}$$

Dado um número real $x \in \mathbb{R}$, representa-se por $\bar{x}$ ou $fl(x)$ a representação de x em $FP(b,p,e_{min},e_{max})$. Se $x = \bar{x}$, diz-se que x tem **representação exacta** em $FP(b,p,e_{min},e_{max})$. Se $x \neq \bar{x}$ existem duas técnicas para determinar $\bar{x}$:

- Truncatura: $\bar{x}$ obtém-se desprezando os números da mantissa do número x para além dos p primeiros. Por exemplo o número π é representado em $FP(10,3,-2,2,T)$ por $\bar{\pi} = +0.314 \times 10^1$.
- Arrendondamento: $\bar{x}$ é o número do sistema $FP(b,p,e_{min},e_{max})$ que está mais próximo de x. Por exemplo o número $\pi = 3.14159\ldots$ é representado em $FP(10,5,-2,2,A)$ por $\bar{\pi} = +0.31416 \times 10^1$.

Em alguns casos, o arrendondamento não determina univocamente a aproximação $\bar{x}$. Por exemplo, o número

$$0.75$$

no sistema $FP(10,1,-1,1,A)$ tem dois elementos equidistantes de 0.75: 0.8 e 0.7. É necessário então introduzir regras adicionais e uma das mais utilizadas é a regra Round-to-Even, que nestes casos opta pela aproximação que deixa o último dígito da mantissa par. Logo,

$$\overline{0.75} = 0.8$$

e

$$\overline{0.65} = 0.6.$$

2.3 Operações Aritméticas em Ponto Flutuante

Considere-se os seguintes números

$$\begin{aligned} x &= 1.2568 \\ y &= 0.9854 \end{aligned}$$

cuja representação em $FP(10,4,-2,2,T)$ é dada por

$$\begin{aligned} \bar{x} &= 0.1256 \times 10 \\ \bar{y} &= 0.9854. \end{aligned}$$

A soma destes números é

$$\bar{x} + \bar{y} \;=\; 2.2414$$

que não é um elemento de $FP(10, 4, -2, 2, T)$.

Logo, depois de se efectuar a operação ter-se-á que representar o resultado no sistema em ponto flutuante. Neste caso, a soma é dada por

$$\overline{\bar{x} + \bar{y}} \;=\; 0.2241 \times 10.$$

Por isso, num sistema de ponto flutuante é necessário definir as operações aritméticas elementares (soma, subtração, multiplicação e divisão) nesse sistema.

Sejam $x, y \in FP(b, m, e_{min}, e_{max})$ e $\odot \in \{+, -, \times, /\}$ uma operação aritmética.

Define-se

$$\bar{x} \odot \bar{y} \;=\; \overline{\bar{x} \odot \bar{y}}.$$

3 Erros

No que se segue, considera-se que se está na notação de base 10 e que as aproximações são feitas por arredondamento.

Seja $x \in \mathbb{R}$ e $\bar{x}$ uma aproximação de x. Define-se como **erro da aproximação** a

$$\epsilon_{\bar{x}} = x - \bar{x}.$$

Diz-se que $\bar{x}$ é uma **aproximação por defeito** se $\bar{x} < x$ e diz-se que $\bar{x}$ é uma **aproximação por excesso** se $\bar{x} > x$.

Chama-se de **erro absoluto** ao módulo do erro

$$\Delta_{\bar{x}} = |x - \bar{x}|$$

e **erro relativo** a

$$r_{\bar{x}} = \frac{\Delta_{\bar{x}}}{|x|} = \left| \frac{x - \bar{x}}{x} \right|$$

se $x \neq 0$.

Por exemplo, se $x = 3.333\ldots$ e $\bar{x} = 3.33$, então,

$$\Delta_{\bar{x}} = 0.0033\ldots$$

e

$$r_{\bar{x}} = 0.001.$$

Para $y = 0.00333\ldots$ e $\bar{y} = 0$, tem-se que

$$\Delta_{\bar{y}} = 0.0033\ldots$$

e

$$r_{\bar{y}} = 1.$$

Com estes exemplos verifica-se que o erro relativo fornece mais informação que o erro absoluto (distância entre a aproximação e o valor) pois tem em conta a ordem de grandeza do valor de x.

Outras medidas da qualidade de uma aproximação é o número de casas decimais correctas e o número de dígitos significativos. Diz-se que o número x se encontra representado com d casas decimais correctas quando a sua parte decimal apresenta d algarismos decimais e resulta de um arredondamento correctamente efectuado sobre um outro número. Diz-se que o número x se encontra representado com k algarismos significativos quando está representado com k algarismos, contados da esquerda para a direita, a partir do primeiro algarismo diferente de zero desde que resulte de um arredondamente bem efectuado sobre um outro número. Na prática, k corresponde ao número de dígitos correctos de mantissa.

Seja $x \in \mathbb{R}$ e $\bar{x}$ uma aproximação de x. Diz-se que $\bar{x}$ é uma aproximação de x com **pelo menos d casas decimais correctas** se

$$\Delta_{\bar{x}} = |x - \bar{x}| \leqslant \frac{1}{2}10^{-d}.$$

Note-se que esta definição só é válida para a notação científica normalizada.

Exemplo. *Sejam* $x = \sqrt{2} = 1.41421356237\ldots$ *e* $\bar{x} = 1.41$ *uma aproximação de* x. *Como*

$$\begin{aligned} |x - \bar{x}| &= 0.00421356\ldots \\ &\leqslant 0.005 = \frac{1}{2}10^{-2}, \end{aligned}$$

então $\bar{x}$ *é uma aproximação de* $\sqrt{2}$ *com pelo menos* 2 *casas decimais correctas.*

Exemplo. *Seja* $y = \sqrt{5} = 2.23606\ldots$ *e* $\bar{y} = 2.24$ *uma aproximação de* y. *Como*

$$\begin{aligned} |y - \bar{y}| &= 0.0039320\ldots \\ &\leqslant 0.005 = \frac{1}{2}10^{-2}, \end{aligned}$$

então $\bar{y}$ *é uma aproximação de* $\sqrt{5}$ *com pelo menos 2 casas decimais correctas.*

Considere-se $x \in \mathbb{R}$ e $\bar{x}$ uma aproximação de x que tem a seguinte representação científica

$$\bar{x} = \pm 0.d_{-1}d_{-2}\ldots d_{-k}d_{-k-1}\ldots d_{-k-l} \times 10^{p}.$$

Diz-se que $\bar{x}$ é uma aproximação de x com **pelo menos k algarismos significativos** se

$$\Delta_{\bar{x}} = |x - \bar{x}| \leqslant \frac{1}{2}10^{-k+p}.$$

Adicionalmente, se

$$\Delta_{\bar{x}} = |x - \bar{x}| > \frac{1}{2}10^{-k+p-1}$$

diz-se que $\bar{x}$ é uma aproximação de x com **com exactamente k algarismos significativos**.

Exemplo. *Sejam* $\bar{\pi}_1 = (0.31415)\,10^1$ *e* $\bar{\pi}_2 = (0.31416)\,10^1$ *duas aproximações de* $\pi = 3.14159265\ldots$. *Então,*

$$0.5 \times 10^{-4+1-1} \leqslant |\pi - \bar{\pi}_1| = 0.9265 \times 10^{-4} \leqslant 0.5 \times 10^{-3} = 0.5 \times 10^{-4+1}$$

e

$$0.5 \times 10^{-5+1-1} \leqslant |\pi - \bar{\pi}_2| = 0.735 \times 10^{-5} \leqslant 0.5 \times 10^{-4} = 0.5 \times 10^{-5+1}.$$

Logo, $\bar{\pi}_1$ *tem 3 casas decimais correctas e tem exactamente 4 algarismos significativos e* $\bar{\pi}_2$ *tem 4 casas decimais correctas e tem exactamente 5 algarismos significativos.*

Exemplo. *Seja* $\bar{x} = 0.25$ *uma aproximação de* x. *Se se souber que* $\Delta_{\bar{x}} \leqslant 0.005$, *então como*

$$\Delta_{\bar{x}} = |x - \bar{x}| \leqslant 0.5 \times 10^{-2}$$

conclui-se que $\bar{x}$ *tem 2 casas decimais correctas e pelo menos 2 algarismos significativos.*

Exemplo. *Seja* $\bar{x} = 0.00589 = 0.589 \times 10^{-2}$ *e* $\Delta_{\bar{x}} \leqslant 0.5 \times 10^{-4}$. *Uma vez que*

$$\Delta_{\bar{x}} \leqslant 0.5 \times 10^{-4} = 0.5 \times 10^{-k-2},$$

logo

$$k = 2.$$

Assim $\bar{x} = 0.00589$ *tem 4 casas decimais correctas e pelo menos 2 algarismos significativos.*

4 Propagação de Erros. Fórmula Fundamental do Cálculo de Erros

Nesta secção, considera-se que todas as funções são contínuas e diferenciáveis.

Se se quiser calcular $y = f(x)$, em que apenas se conhece um valor aproximado $\bar{x}$ de x, qual o erro que se irá obter na solução $\bar{y} = f(\bar{x})$? Ou seja, qual é o efeito de propagação do erro $\Delta_{\bar{x}}$ na solução do problema? Pelo Teorema de Lagrange no intervalo $I = [\bar{x} - \Delta_{\bar{x}}, \bar{x} + \Delta_{\bar{x}}]$, existe um ponto $c \in I$ tal que

$$\begin{aligned} & f'(c) = \frac{f(x) - f(\bar{x})}{x - \bar{x}} \\ \Leftrightarrow \quad & f(x) - f(\bar{x}) = f'(c)\,(x - \bar{x})\,. \end{aligned}$$

Logo,

$$\begin{aligned} & |f(x) - f(\bar{x})| = |f'(c)|\,|x - \bar{x}| \\ \Leftrightarrow \quad & \Delta_{f(\bar{x})} = |f'(c)|\,\Delta_{\bar{x}} \end{aligned}$$

onde $\Delta_{f(\bar{x})} = |f(x) - f(\bar{x})|$. Como na prática não se conhece o valor de c, então majora-se o erro da seguinte forma

$$\Delta_{f(\bar{x})} \leqslant |f'(x)|_M\,\Delta_{\bar{x}}$$

em que

$$|f'(x)|_M = \sup_{x \in I} f'(x)\,.$$

A esta desigualdade dá-se o nome de **Fórmula Fundamental do Cálculo de Erros (FFCE)** e que também pode ser escrita na forma

$$\Delta_{f(\bar{x})} \leqslant \left|\frac{\partial f}{\partial x}\right|_M \Delta_{\bar{x}}$$

onde $\dfrac{\partial f}{\partial x} = f'(x)$ (notação de Leibniz).

Exemplo. *Seja $f(x) = 2x^2 + 3$. Qual é o erro propagado a $f(x)$ se se tomar $\bar{x} = 1.3$? Se se considerar todos os algarismos da aproximação significativos, tem-se que*

$$x \in]1.3 - 0.05, 1.3 + 0.05[.$$

Um valor aproximado de $f(x)$ é

$$f(\bar{x}) = 2 \times 1.3^2 + 3 = 6.38$$

e um majorante do erro da aproximação é dado por

$$\begin{aligned} \Delta_{f(\bar{x})} &\leqslant |f'(x)|_M \Delta_{\bar{x}} \\ &\leqslant |4x|_M \Delta_{\bar{x}} \\ &\leqslant 4 \times (1.3 + 0.05) \times 0.05 \\ &\leqslant 0.27. \end{aligned}$$

Logo,

$$x \in]1.3 - 0.05, 1.3 + 0.05[\Rightarrow f(x) \in]6.38 - 0.27, 6.38 + 0.27[.$$

Usualmente, uma função f depende de várias variáveis $x_1, x_2, \ldots, x_n$ e a **generalização da FFCE a várias variáveis** é dada por

$$\Delta_{\bar{f}} \leqslant \sum_{i=1}^{n} \left|\frac{\partial f}{\partial x_i}\right|_M \Delta_{\bar{x}_i,}$$

onde $\bar{f} = f(\overline{x_1}, \overline{x_2}, \ldots, \overline{x_n})$ e $\dfrac{\partial f}{\partial x_i}$ é a derivada parcial de f em relação à variável x_i. Por exemplo, para

$$f(x, y, z) = 2xy + 3z^2 - y^2 z$$

tem-se que

$$\begin{aligned} \frac{\partial f}{\partial x} &= (2xy + 3z^2 - y^2 z)'_x &&= 2y, \\ \frac{\partial f}{\partial y} &= (2xy + 3z^2 - y^2 z)'_y &&= 2x - 2yz \quad \text{e} \\ \frac{\partial f}{\partial z} &= (2xy + 3z^2 - y^2 z)'_z &&= 6z - y^2. \end{aligned}$$

Exemplo. *Considere-se* $f(x,y) = xy + x^2$.
Se $x \in]3.1 - 0.01, 3.1 + 0.01[$ *e* $y \in]2.3 - 0.3, 2.3 + 0.3[$. *Um valor aproximado de* f *é*

$$\bar{f} = f(\bar{x}, \bar{y}) = 3.1 \times 2.3 + 3.1^2 = 16.74.$$

Como

$$\begin{aligned} \Delta_{\bar{f}} &\leqslant \left|\frac{\partial f}{\partial x}\right|_M \Delta_{\bar{x}} + \left|\frac{\partial f}{\partial y}\right|_M \Delta_{\bar{y}} \\ &\leqslant |y + 2x|_M \Delta_{\bar{x}} + |x|_M \Delta_{\bar{y}} \\ &\leqslant |(2.3 + 0.3) + 2(3.1 + 0.01)| \times 0.01 + |3.1 + 0.01| \times 0.3 \\ &\leqslant 1.02128 \\ &\leqslant 1.03. \end{aligned}$$

Logo

$$f(x,y) \in]16.74 - 1.03, 16.74 + 1.03[=]15.71, 17.77[.$$

4.1 Estimativas do Erro em Operações Aritméticas Elementares

Considere-se a operação Soma definida por $S = x + y$. Então,

$$\begin{aligned} \Delta_{\bar{S}} &\leqslant \left|\frac{\partial S}{\partial x}\right|_M \Delta_{\bar{x}} + \left|\frac{\partial S}{\partial y}\right|_M \Delta_{\bar{y}} \\ &\leqslant 1\Delta_{\bar{x}} + 1\Delta_{\bar{y}}. \end{aligned}$$

Logo

$$\Delta_{\bar{S}} \leqslant \Delta_{\bar{x}} + \Delta_{\bar{y}}.$$

A operação Subtracção definida por $D = x - y$, tem o erro majorado por

$$\begin{aligned} \Delta_{\bar{D}} &\leqslant \left|\frac{\partial D}{\partial x}\right|_M \Delta_{\bar{x}} + \left|\frac{\partial D}{\partial y}\right|_M \Delta_{\bar{y}} \\ &\leqslant 1\Delta_{\bar{x}} + |-1| \Delta_{\bar{y}} \end{aligned}$$

Logo

$$\Delta_{\bar{D}} \leqslant \Delta_{\bar{x}} + \Delta_{\bar{y}}.$$

A operação Produto Escalar definida por $P_k = kx$, onde $k \in \mathbb{R}$, tem o erro majorado por

$$\begin{aligned} \Delta_{\overline{P_k}} &\leqslant \left|\frac{\partial P_k}{\partial x}\right|_M \Delta_{\bar{x}} \\ &\leqslant |k| \Delta_{\bar{x}}. \end{aligned}$$

Exemplo. *Sejam* $f(x,y) = 3x - 5y$, $x \in]5.3 - 0.2, 5.3 + 0.2[$ *e* $y \in]2 - 0.1, 2 + 0.1[$. *Um valor aproximado de* f, *será*

$$f(\bar{x}, \bar{y}) = 3 \times 5.3 - 5 \times 2 = 5.9$$

e o erro é majorado por

$$\Delta_{\bar{f}} \leqslant 3\Delta_{\bar{x}} + 5\Delta_{\bar{y}} = 1.1.$$

Logo

$$f(x,y) \in]5.9 - 1.1, 5.9 + 1.1[.$$

No caso da operação Produto definida por $P = xy$, tem-se que

$$\Delta_{\bar{P}} \leqslant |y|_M \Delta_{\bar{x}} + |x|_M \Delta_{\bar{y}}$$

e no caso da operação Quociente $Q = \dfrac{x}{y}$, tem-se que

$$\Delta_{\bar{Q}} \leqslant \left|\frac{1}{y}\right|_M \Delta_{\bar{x}} + \left|-\frac{x}{y^2}\right|_M \Delta_{\bar{y}}.$$

4.2 Propagação de Erros Relativos

O erro relativo é dado por

$$r_{\bar{x}} = \frac{\Delta_{\bar{x}}}{|x|} = \left|\frac{x - \bar{x}}{x}\right|,$$

ou seja,

$$\Delta_{\bar{x}} = |x| \times r_{\bar{x}}.$$

Se se considerar uma função $f = f(x_1, x_2, \ldots, x_n)$ e a fórmula fundamental dos erros

$$\Delta_{\bar{f}} \leqslant \sum_{i=1}^{n} \left|\frac{\partial f}{\partial x_i}\right|_M \Delta_{\overline{x_i}}$$

e se substituir os erros absolutos pelos relativos, obtém-se

$$\begin{aligned} & \left|f \times r_{\bar{f}}\right| \leqslant \sum_{i=1}^{n} \left|\frac{\partial f}{\partial x_i}\right|_M |x_i \times r_{\bar{x}_i}| \\ \Leftrightarrow \quad & r_{\bar{f}} \leqslant \sum_{i=1}^{n} \left|\frac{x_i \frac{\partial f}{\partial x_i}}{f}\right|_M r_{\bar{x}_i} \\ \Leftrightarrow \quad & r_{\bar{f}} \leqslant \sum_{i=1}^{n} |p_i|_M \, r_{\bar{x}_i}. \end{aligned}$$

A esta última expressão dá-se o nome de **fórmula fundamental para erros relativos**. Os factores p_i

$$p_i = \frac{x_i \frac{\partial f}{\partial x_i}}{f}$$

são denominados de números de condição. Na prática o valor exacto de x é geralmente desconhecido, sendo comum substituir-se x por $\bar{x}$ na fórmula anterior.

Em seguida, apresentam-se estimativas para os erros relativos das operações aritméticas elementares.

O erro relativo da operação Soma $S = x + y$ é estimado por

$$\begin{aligned} r_{\bar{S}} &\approx |p_{\bar{x}}|\, r_{\bar{x}} + |p_{\bar{y}}|\, r_{\bar{y}} \\ &\approx \left|\frac{\bar{x}}{\bar{x}+\bar{y}}\right| r_{\bar{x}} + \left|\frac{\bar{y}}{\bar{x}+\bar{y}}\right| r_{\bar{y}}. \end{aligned}$$

Ou seja,

$$r_{\bar{S}} \approx \frac{\Delta_{\bar{x}} + \Delta_{\bar{y}}}{|\bar{x}+\bar{y}|}.$$

Analogamente o erro relativo da operação Subtracção é estimado por

$$r_{\bar{D}} \approx \left|\frac{\bar{x}}{\bar{x}-\bar{y}}\right| r_{\bar{x}} + \left|\frac{\bar{y}}{\bar{x}-\bar{y}}\right| r_{\bar{y}}$$

i.e.,

$$r_{\bar{D}} \approx \frac{\Delta_{\bar{x}} + \Delta_{\bar{y}}}{|\bar{x}-\bar{y}|}.$$

O erro relativo da operação Produto $P = xy$ é estimado por

$$\begin{aligned} r_{\bar{P}} &\approx |p_x|\, r_{\bar{x}} + |p_y|\, r_{\bar{y}} \\ &\approx \left|\frac{\bar{x}\bar{y}}{\bar{x}\bar{y}}\right| r_{\bar{x}} + \left|\frac{\bar{x}\bar{y}}{\bar{x}\bar{y}}\right| r_{\bar{y}} \\ &\approx r_{\bar{x}} + r_{\bar{y}} \end{aligned}$$

e o erro relativo da operação Quociente $Q = \frac{x}{y}$ é estimado

$$\begin{aligned} r_{\bar{Q}} &\approx |p_x|\, r_{\bar{x}} + |p_y|\, r_{\bar{y}} \\ &\approx \left|\frac{\bar{x}\frac{1}{\bar{y}}}{\frac{\bar{x}}{\bar{y}}}\right| r_{\bar{x}} + \left|\frac{\bar{y}\left(-\frac{\bar{x}}{\bar{y}^2}\right)}{\frac{\bar{x}}{\bar{y}}}\right| r_{\bar{y}} \\ &\approx r_{\bar{x}} + r_{\bar{y}}. \end{aligned}$$

A tabela seguinte resume os erros, absolutos e relativos, das operações aritméticas:

	Erro Absoluto	Erro Relativo
Soma	$\Delta_{\bar{S}} \leqslant \Delta_{\bar{x}} + \Delta_{\bar{y}}$	$r_{\bar{S}} \approx \frac{\Delta_{\bar{x}}+\Delta_{\bar{y}}}{\lvert\bar{x}+\bar{y}\rvert}$
Diferença	$\Delta_{\bar{D}} \leqslant \Delta_{\bar{x}} + \Delta_{\bar{y}}$	$r_{\bar{D}} \approx \frac{\Delta_{\bar{x}}+\Delta_{\bar{y}}}{\lvert\bar{x}-\bar{y}\rvert}$
Produto	$\Delta_{\bar{P}} \leqslant \lvert y\rvert_M \Delta_{\bar{x}} + \lvert x\rvert_M \Delta_{\bar{y}}$	$r_{\bar{P}} \approx r_{\bar{x}} + r_{\bar{y}}$
Quociente	$\Delta_{\bar{Q}} \leqslant \left\lvert\frac{1}{y}\right\rvert_M \Delta_{\bar{x}} + \left\lvert\frac{x}{y^2}\right\rvert_M \Delta_{\bar{y}}$	$r_{\bar{Q}} \approx r_{\bar{x}} + r_{\bar{y}}$

Exemplo. *Determine-se o número de algarismos significativos do resultado de cada uma das operações xy e $x + y$ em que $\bar{x} = 1010$ e $\bar{y} = 1000$. Se todos os algarismos são significativos, então*

$$\Delta_{\bar{x}} < 0.5 \quad e \quad \Delta_{\bar{y}} < 0.5$$

e

$$\begin{aligned} r_{\bar{x}} &= \frac{\Delta_{\bar{x}}}{|x|} \approx \frac{\Delta_{\bar{x}}}{|\bar{x}|} < 0.495 \times 10^{-3} \\ r_{\bar{y}} &\approx 0.5 \times 10^{-3}. \end{aligned}$$

Então,

$$\begin{aligned} \Delta_{\bar{P}} &\leqslant |y|_M \Delta_{\bar{x}} + |x|_M \Delta_{\bar{y}} \\ &\leqslant 1000.5 \times 0.5 + 1010.5 \times 0.5 = 1005.5 \end{aligned}$$

e

$$r_{\bar{P}} \approx 0.495 \times 10^{-3} + 0.5 \times 10^{-3} = 0.995 \times 10^{-3}.$$

Além disso, como

$$\bar{x}\bar{y} = 1010000 = 0.1010 \times 10^7$$

e

$$\Delta_{\bar{P}} \leqslant 1005 < 5000 = 0.5 \times 10^4 = 0.5 \times 10^{-3+7},$$

o produto tem 3 algarismos significativos.

Em relação aos erros da soma, tem-se que

$$\Delta_{\bar{S}} \leqslant \Delta_{\bar{x}} + \Delta_{\bar{y}} < 0.5 + 0.5 = 1$$

e

$$\begin{aligned} r_{\bar{S}} &\approx \frac{\Delta_{\bar{x}} + \Delta_{\bar{y}}}{|\bar{x} + \bar{y}|} = \frac{1}{2010} \\ &\approx 0.498 \times 10^{-3}. \end{aligned}$$

Além disso, como

$$\bar{x} + \bar{y} = 2010 = 0.201 \times 10^4$$

e

$$\Delta_{\bar{S}} \leqslant 1 < 5 = 0.5 \times 10 = 0.5 \times 10^{-3+4}$$

a soma tem 3 algarismos significativos.