Análise de Erros em Aritmética de Ponto Flutuante

Márcio Antônio de Andrade Bortoloti

Cálculo Numérico

Departamento de Ciências Exatas e Tecnológicas - DCET
Universidade Estadual do Sudoeste da Bahia

Sumário

Análise de Erros

- Truncamento e Arredondamento
- Truncamento e Arredondamento
- Erros Absoluto e Relativo
- Operações em Aritmética de Ponto Flutuante

Análise de Erros nas Operações de Ponto Flutuante

- Adição
- Subtração
- Multiplicação
- Divisão

Efeitos Numéricos

- Cancelamento Subtrativo

Análise de Erros

Definição
Se $x \in \mathcal{F}(10, t, m, M)$ então ele pode ser representado como

$$x = f_x \times 10^e + g_x \times 10^{e-t},$$

onde $0.1 \leq f_x < 1$ e $0 \leq g_x < 1$.

Exemplo:

Seja $x = 234.57$ e $t = 4$. logo

$$\begin{aligned} x &= 234.57 \\ &= 0.23457 \times 10^3 \\ &= (0.2345 + 0.00007) \times 10^3 \\ &= 0.2345 \times 10^3 + 0.00007 \times 10^3 \\ &= 0.2345 \times 10^3 + 0.7 \times 10^{-1} \end{aligned}$$

Definição de Truncamento
Seja $\mathcal{F}(10, t, m, M)$ uma máquina e x um número que em geral não pode ser representado em $\mathcal{F}$ de forma exata. Quando isso ocorre, devemos utilizar uma aproximação $\overline{x}$ para x. Assim, se x é tal que

$$x = f_x \times 10^e + g_x \times 10^{e-t}, \text{ para } m \leq e \leq M,$$

onde $0.1 \leq f_x < 1$ e $0 \leq g_x < 1$ então a operação de truncamento gera uma aproximação $\overline{x}$, de x, da forma

$$\overline{x} = f_x \times 10^e.$$

Definição de Arredondamento
Seja $\mathcal{F}(10, t, m, M)$. No caso de obtermos uma aproximação, $\overline{x}$, de um número $x = f_x \times 10^e + g_x \times 10^{e-t}$, usando arredondamento, teremos que analisar g_x de forma que

$$\overline{x} = \begin{cases} f_x \times 10^e & \text{se } g_x < 1/2 \\ f_x \times 10^e + 10^{e-t} & \text{se } g_x \geq 1/2 \end{cases}$$

Exemplo:

Considere uma máquina $\mathcal{F}(10, 3, -5, 5)$. Vamos representar $x = 45.8787$ em $\mathcal{F}$. De fato,

$$x = 45.8787 = 0.458 \times 10^2 + 0.787 \times 10^{-1}$$

Fazendo o arredondamento

$$\overline{x} = 0.458 \times 10^2 + 10^{-1} = 0.459 \times 10^2 = 0.459$$

Erro Absoluto

Definição
Seja $x \in \mathbb{R}$ e $\overline{x}$ sua aproximação. O erro absoluto, cometido na representação de x por $\overline{x}$ é definido por

$$EA_x = x - \overline{x}.$$

Exemplo

O erro absoluto cometido na aproximação de π por $\overline{\pi} = 3.14$ é

$$|EA_\pi| = |\pi - \overline{\pi}| = |\pi - 3.14| \leq 0.01.$$

Erro Relativo

Considere dois números $x = 1991.67$ e $y = 3.67$. Se aproximarmos x e y por $\overline{x} = 1991.7$ e $\overline{y} = 3.7$ teremos

$$|EA_x| = |EA_y| = 0.03.$$

No entanto, os dois números estão aproximados da "mesma forma" ?

Erro Relativo

Considere dois números $x = 1991.67$ e $y = 3.67$. Se aproximarmos x e y por $\overline{x} = 1991.7$ e $\overline{y} = 3.7$ teremos

$$|EA_x| = |EA_y| = 0.03.$$

No entanto, os dois números estão aproximados da "mesma forma" ? Qual aproximação está mais precisa ?

Para responder a pergunta vamos usar a seguinte definição:

Erro Relativo

Definição
O Erro Relativo, ER_x, cometido na aproximação de x por $\overline{x}$ é definido como

$$ER_x = \frac{EA_x}{\overline{x}} = \frac{x - \overline{x}}{\overline{x}}$$

Voltando ao exemplo ...

Se $x = 1991.67$ e $y = 3.67$ as aproximações $\overline{x} = 1991.7$ e $\overline{y} = 3.7$ cometem erros relativos da ordem de

$$|ER_x| = \frac{|EA_x|}{|\overline{x}|} = \frac{0.03}{1991.7} = 1.506250941 \times 10^{-5}.$$

$$|ER_y| = \frac{|EA_y|}{|\overline{y}|} = \frac{0.03}{3.7} = 0.810810810 \times 10^{-2}.$$

Teorema
Sejam $x \in \mathbb{R}$ e $\mathcal{F}(10, t, m, M)$ uma máquina. Os erros absoluto e relativo cometidos na aproximação de x por $\overline{x}$, utilizando truncamento, são da ordem de

$$|EA_x| = |x - \overline{x}| < 10^{e-t} \quad \text{e} \quad |ER_x| = \frac{|EA_x|}{|\overline{x}|} < 10^{-t+1}.$$

Prova:

Note que

$$x = f_x \times 10^e + g_x \times 10^{e-t},$$

onde $0.1 \leq f_x < 1$ e $0 \leq g_x < 1$.

Usando o truncamento, tem-se

$$\overline{x} = f_x \times 10^e.$$

Logo

$$\begin{aligned} |EA_x| &= |x - \overline{x}| \\ &= |f_x \times 10^e + g_x \times 10^{e-t} - f_x \times 10^e| \\ &= |g_x| \times 10^{e-t} \\ &< 10^{e-t} \quad (|g_x| < 1) \end{aligned}$$

Agora, o erro relativo ...

$$\begin{aligned} |ER_x| &= \frac{|EA_x|}{|\overline{x}|} \\ &= \frac{|g_x| \times 10^{e-t}}{|f_x| \times 10^e} \\ &< \frac{10^{e-t}}{0.1 \times 10^e} \\ &< 10^{-t+1} \end{aligned}$$

Teorema
Sejam $x \in \mathbb{R}$ e $\mathcal{F}(10, t, m, M)$ uma máquina. Os erros absoluto e relativo cometidos na aproximação de x por $\overline{x}$, utilizando arredondamento, são da ordem de

$$|EA_x| \leq 0.5 \times 10^{e-t} \quad \text{e} \quad |ER_x| \leq 0.5 \times 10^{-t+1}.$$

Prova:

Note que

$$x = f_x \times 10^e + g_x \times 10^{e-t},$$

onde $0.1 \leq f_x < 1$ e $0 \leq g_x < 1$ e

$$\overline{x} = \begin{cases} f_x \times 10^e & \text{se } g_x < 1/2 \\ f_x \times 10^e + 10^{e-t} & \text{se } g_x \geq 1/2 \end{cases}$$

Análise de Erros

Se $g_x < 1/2$ então

$$\begin{aligned} |EA_x| &= |x - \overline{x}| = |g_x| \times 10^{e-t} \\ &< \frac{1}{2} \times 10^{e-t} \end{aligned}$$

E também

$$\begin{aligned} |ER_x| &= \frac{|EA_x|}{|\overline{x}|} \\ &= \frac{|g_x| \times 10^{e-t}}{|f_x| \times 10^{e}} \\ &< \frac{0.5 \times 10^{e-t}}{0.1 \times 10^{e}} \\ &< \frac{1}{2} \times 10^{-t+1} \end{aligned}$$

Se $g_x \geq 1/2$ então

$$\begin{aligned} |EA_x| &= |x - \overline{x}| \\ &= |(f_x \times 10^e + g_x \times 10^{e-t}) - (f_x \times 10^e + 10^{e-t})| \\ &= |g_x \times 10^{e-t} - 10^{e-t}| \\ &= |g_x - 1| \times 10^{e-t} \\ &\leq \frac{1}{2} \times 10^{e-t} \end{aligned}$$

E também

$$\begin{aligned} |ER_x| &= \frac{|EA_x|}{|\overline{x}|} \leq \frac{1/2 \times 10^{e-t}}{|f_x \times 10^e + 10^{e-t}|} \\ &< \frac{1/2 \times 10^{e-t}}{|f_x| \times 10^e} < \frac{1/2 \times 10^{e-t}}{0.1 \times 10^e} < \frac{1}{2} \times 10^{-t+1} \end{aligned}$$

Operações em Aritmética de Ponto Flutuante

- O arredondamento não é muito utilizado, pois mesmo acarretando erros menores, ele aumenta o tempo de execução de um programa.
- Mesmo que x e y estejam representados de forma exata, a soma $x + y$, por exemplo, também gera erros numéricos.
- **Exemplo:** Sejam $x = 0.234 \times 10^5$ e $y = 0.567 \times 10^2$ em uma máquina $\mathcal{F}(10, 3, -5, 5)$. Então

$$\begin{aligned} x + y &= 0.234 \times 10^5 + 0.567 \times 10^2 \\ &= 0.234 \times 10^5 + 0.000567 \times 10^5 \\ &= (0.234 + 0.000567) \times 10^5 \\ &= 0.234567 \times 10^5 \\ &= 0.234 \times 10^5 \quad \text{(se truncarmos)} \\ &= 0.235 \times 10^5 \quad \text{(se arredondarmos)} \end{aligned}$$

Análise de Erros nas Operações de Ponto Flutuante

Erros nas Operações Aritméticas de Ponto Flutuante

Já vimos que o erro relativo no resultado de uma operação (supondo que as parcelas são representadas exatamente) será

$$\begin{aligned} |ER_{op}| &< 10^{-t+1} \quad \text{(Para o caso de Truncamento)} \\ |ER_{op}| &< \frac{1}{2} \times 10^{-t+1} \quad \text{(Para o caso de Arredondamento)} \end{aligned}$$

Exemplo:

Considere uma máquina com $t = 4$ e dois números $x = 0.9765 \times 10^4$ e $y = 0.2456 \times 10^2$. Logo

$$\begin{aligned} x + y &= 0.9765 \times 10^4 + 0.2456 \times 10^2 \\ &= 0.9765 \times 10^4 + 0.002456 \times 10^4 = 0.978956 \times 10^4 \\ &= \begin{cases} 0.9789 \times 10^4 & \text{Se for usado truncamento} \\ 0.9790 \times 10^4 & \text{Se for usado arredondamento} \end{cases} \end{aligned}$$

Análise de Erros - Adição

Sejam x e y tais que

$$\begin{aligned} x &= \overline{x} + EA_x \\ y &= \overline{y} + EA_y \end{aligned}$$

Logo

$$\begin{aligned} x + y &= (\overline{x} + EA_x) + (\overline{y} + EA_y) \\ &= (\overline{x} + \overline{y}) + (EA_x + EA_y) \end{aligned}$$

Assim

$$\boxed{EA_{x+y} = EA_x + EA_y}.$$

Atenção: Aqui discutimos como o erro na representação de cada parcela da soma influencia o resultado final da operação.

Por outro lado, temos

$$\begin{aligned} ER_{x+y} &= \frac{EA_{x+y}}{\overline{x}+\overline{y}} = \frac{EA_x + EAy}{\overline{x}+\overline{y}} = \frac{EA_x}{\overline{x}+\overline{y}} + \frac{EA_y}{\overline{x}+\overline{y}} \\ &= \frac{EA_x}{\overline{x}}\left(\frac{\overline{x}}{\overline{x}+\overline{y}}\right) + \frac{EA_y}{\overline{y}}\left(\frac{\overline{y}}{\overline{x}+\overline{y}}\right) \\ &= ER_x\left(\frac{\overline{x}}{\overline{x}+\overline{y}}\right) + ER_y\left(\frac{\overline{y}}{\overline{x}+\overline{y}}\right) \end{aligned}$$

Análise de Erros - Subtração

É fácil ver que

$$EA_{x-y} = EA_x - EA_y$$

$$\begin{aligned} ER_{x-y} &= \frac{EA_{x-y}}{\overline{x} - \overline{y}} \\ &= \frac{EA_x - EA_y}{\overline{x} - \overline{y}} \\ &= ER_x\Big(\frac{\overline{x}}{\overline{x} - \overline{y}}\Big) - ER_y\Big(\frac{\overline{y}}{\overline{x} - \overline{y}}\Big) \end{aligned}$$

Análise de Erros - Multiplicação

Notemos que

$$\begin{aligned} xy &= (\overline{x} + EA_x)(\overline{y} + EA_y) = \overline{x}\,\overline{y} + \overline{x}EA_y + \overline{y}EA_x + EA_xEA_y \\ &\approx \overline{x}\,\overline{y} + \overline{x}EA_y + \overline{y}EA_x \quad \text{pois} \quad EA_xEA_y \to 0 \end{aligned}$$

Logo

$$\boxed{EA_{xy} \approx \overline{x}EA_y + \overline{y}EA_x}$$

E o Erro Relativo ...

$$\begin{aligned} ER_{xy} &= \frac{EA_{xy}}{\overline{x}\,\overline{y}} \\ &\approx \frac{\overline{x}EA_y + \overline{y}EA_x}{\overline{x}\,\overline{y}} = \frac{EA_x}{\overline{x}} + \frac{EA_y}{\overline{y}} = ER_x + ER_y \end{aligned}$$

Logo

$$\boxed{ER_{xy} = ER_x + ER_y}$$

Análise de Erros - Divisão

$$\begin{aligned}\frac{x}{y} &= \frac{\overline{x}+EA_x}{\overline{y}+EA_y} \\ &= \frac{\overline{x}+EA_x}{\overline{y}(1+\frac{EA_y}{\overline{y}})} \\ &= \frac{\overline{x}+EA_x}{\overline{y}}\Big(\frac{1}{1+\frac{EA_y}{\overline{y}}}\Big)\end{aligned}$$

Agora note que

$$\begin{aligned}\frac{1}{1+\frac{EA_y}{\overline{y}}} &= \sum_{n=0}^{\infty}(-1)^n\Big(\frac{EA_x}{\overline{y}}\Big)^n \quad \text{(Note que } EA_x \to 0\text{)} \\ &\approx 1-\frac{EA_x}{\overline{y}}\end{aligned}$$

Análise de Erros - Divisão

Logo

$$\begin{aligned}\frac{x}{y} &\approx \frac{\overline{x}+EA_x}{\overline{y}}\left(1-\frac{EA_x}{\overline{y}}\right)\\ &\approx \frac{\overline{x}}{\overline{y}}+\frac{EA_X}{\overline{y}}-\frac{\overline{x}EA_y}{\overline{y}^2}-\frac{EA_x\ EA_y}{\overline{y}^2}\\ &\approx \frac{\overline{x}}{\overline{y}}+\frac{EA_X}{\overline{y}}-\frac{\overline{x}EA_y}{\overline{y}^2} \quad (EA_xEA_y \to 0)\end{aligned}$$

Logo

$$EA_{x/y} \approx \frac{EA_X}{\overline{y}}-\frac{\overline{x}EA_y}{\overline{y}^2}=\frac{\overline{y}EA_x-\overline{x}EA_y}{\overline{y}^2}$$

Assim,

$$\boxed{EA_{x/y} \approx \frac{\overline{y}EA_x-\overline{x}EA_y}{\overline{y}^2}}$$

Análise de Erros - Divisão

E finalmente,

$$\begin{aligned} ER_{x/y} &= \frac{EA_{x/y}}{\overline{x}/\overline{y}} \\ &\approx \Big(\frac{\overline{y}EA_x - \overline{x}EA_y}{\overline{y}^2}\Big)\frac{\overline{y}}{\overline{x}} \\ &\approx \frac{EA_x}{\overline{x}} - \frac{EA_y}{\overline{y}} \\ &\approx ER_x - ER_y \end{aligned}$$

Logo

$$\boxed{ER_{x/y} \approx ER_x - ER_y}$$

Exemplo
Sejam x, y, z, t representados exatamente. Qual o erro relativo total cometido no cálculo de

$$u = (x + y)z - t \quad ?$$

$$\begin{aligned} ER_{x+y} &= ER_x\Big(\frac{\overline{x}}{\overline{x}+\overline{y}}\Big) + ER_y\Big(\frac{\overline{y}}{\overline{x}+\overline{y}}\Big) + RA \\ &= 0 + 0 + RA \end{aligned}$$

Assim,

$$\begin{aligned} ER_{(x+y)z} &= ER_{x+y} + ER_z + RA \\ &= RA + 0 + RA \\ &= 2RA \end{aligned}$$

RA é o erro relativo cometido na aproximação após a operação.

Erros nas Operações Aritméticas de Ponto Flutuante

Para simplificar a notação, faremos $m = (x + y)z$.

$$\begin{aligned} ER_{m-t} &= ER_m\Big(\frac{\overline{m}}{\overline{m} - \overline{t}}\Big) - ER_t\Big(\frac{\overline{t}}{\overline{m} - \overline{t}}\Big) + RA \\ &= ER_m\Big(\frac{\overline{m}}{\overline{m} - \overline{t}}\Big) - 0 + RA \\ &= 2RA\Big(\frac{\overline{m}}{\overline{m} - \overline{t}}\Big) + RA \end{aligned}$$

Logo

$$\begin{aligned} |ER_{m-t}| &= \Big|2RA\Big(\frac{\overline{m}}{\overline{m} - \overline{t}}\Big) + RA\Big| \\ &\leq 2|RA|\Big|\frac{\overline{m}}{\overline{m} - \overline{t}}\Big| + |RA| \\ &\leq 10^{-t+1}\Big(\Big|\frac{\overline{m}}{\overline{m} - \overline{t}}\Big| + \frac{1}{2}\Big) \end{aligned}$$

Efeitos Numéricos

Efeitos Numéricos - Cancelamento Subtrativo

O erro relativo na subtração é dado por

$$ER_{x-y} = ER_x\Big(\frac{\overline{x}}{\overline{x}-\overline{y}}\Big) - ER_y\Big(\frac{\overline{y}}{\overline{x}-\overline{y}}\Big)$$

Exemplo

Considere $\overline{x} = 0.2357 \times 10^3$, $\overline{y} = 0.2353 \times 10^3$ e $\overline{r} = 0.4537 \times 10^3$. Vamos analisar $w = (x-y)r$ em uma máquina com quatro algarismos significativos. Temos $\overline{x} - \overline{y} = 0.0004 \times 10^3$. Note que

$$\begin{aligned}
|ER_{x-y}| &= \Big|ER_x\Big(\frac{\overline{x}}{\overline{x}-\overline{y}}\Big) - ER_y\Big(\frac{\overline{y}}{\overline{x}-\overline{y}}\Big)\Big| \\
&< \Big(\Big|\frac{\overline{x}}{\overline{x}-\overline{y}}\Big| + \Big|\frac{\overline{y}}{\overline{x}-\overline{y}}\Big|\Big) \times \frac{1}{2} \times 10^{-3} \quad (t=4) \\
&< \Big(\frac{0.2357 \times 10^3 + 0.2353 \times 10^3}{0.0004 \times 10^3}\Big) \times \frac{1}{2} \times 10^{-3} \\
&< 0.5888 \approx 59\%
\end{aligned}$$

Agora, note que

$$\begin{aligned} |ER_w| &= |ER_{x-y} + ER_r| \\ &\leq |ER_{x-y}| + |ER_r| \\ &\leq 0.59 + \frac{1}{2} \times 10^{-3} \\ &\leq 0.5905 \approx 59\% \end{aligned}$$

Efeitos Numéricos - Cancelamento Subtrativo

Exemplo

Calcular, usando uma máquina com 10 dígitos,

$$\sqrt{9876} - \sqrt{9875}.$$

Temos que

$$\sqrt{9876} = 0.9937806599 \times 10^2 \quad \text{e} \quad \sqrt{9875} = 0.9937303457 \times 10^2$$

Portanto

$$\begin{aligned} \sqrt{9876} - \sqrt{9875} &= 0.0000503142 \times 10^2 \\ &= 0.5031420000 \times 10^{-4} \end{aligned}$$

Podemos obter um resultado mais preciso? Neste caso a resposta é sim! Basta notar que

Logo

$$\sqrt{x} - \sqrt{y} = \frac{x - y}{\sqrt{x} + \sqrt{y}}$$

$$\sqrt{9876} - \sqrt{9875} = \frac{1}{\sqrt{9876} + \sqrt{9875}} = 0.5031418679 \times 10^{-4}$$