Matemática Computacional
MEBiol, MEBiom e MEFT - Aula 3

Ana Leonor Silvestre

Instituto Superior Técnico, $1^{\underline{o}}$ *Semestre, 2020/2021*

Sumário da Aula 3

Cálculo em sistemas de ponto flutuante.
Propagação de erros em operações aritméticas.
Cancelamento subtrativo.
Propagação de erros em funções.

Cálculo em sistemas de ponto flutuante

Cálculo de operações aritméticas elementares

- Implementação das operações aritméticas elementares
  Se $x, y \in \mathbb{F}$ e op $\in \{+, -, \times, /\}$ então

  $$\mathrm{fl}(x \,\mathsf{op}\, y) = (x \,\mathsf{op}\, y)(1+\delta), \quad |\delta| \leq \epsilon_M.$$

  Suponhamos que são usados algarismos de guarda para efetuar as operações $x \,\mathsf{op}\, y$ e que continuamos a ter $\mathrm{fl}(x \,\mathsf{op}\, y) \in \mathbb{F}$.
  Define-se $\mathsf{op}_{\mathbb{F}} : \mathbb{F} \times \mathbb{F} \to \mathbb{F}$ por

  $$x \,\mathsf{op}_{\mathbb{F}}\, y = \mathrm{fl}(x \,\mathsf{op}\, y).$$

  Deste modo o erro de arredondamento que surge em $x \,\mathsf{op}_{\mathbb{F}}\, y$ satisfaz

  $$|\delta_{x \,\mathsf{op}_{\mathbb{F}}\, y}| = \frac{|x \,\mathsf{op}\, y - x \,\mathsf{op}_{\mathbb{F}}\, y|}{|x \,\mathsf{op}\, y|} \leq \epsilon_M.$$

Cálculo de funções elementares

- Implementação das funções elementares $(\cos, \sin, \exp, \text{etc...})$
  Se $x \in \mathbb{F} \cap \mathrm{Dom}(g)$ então

$$\mathrm{fl}(g(x)) = g(x)(1+\delta), \quad |\delta| \leq \epsilon_M.$$

  Define-se $g_{\mathbb{F}} : \mathbb{F} \cap \mathrm{Dom}(g) \to \mathbb{F}$ por

$$g_{\mathbb{F}}(x) = \mathrm{fl}(g(x)).$$

  Tem-se

$$|\delta_{g_{\mathbb{F}}(x)}| = \frac{|g(x) - g_{\mathbb{F}}(x)|}{|g(x)|} \leq \epsilon_M.$$

Cálculo em sistemas de ponto flutuante

Um algoritmo é uma sequência finita de instruções bem definidas e não ambíguas, cada uma das quais pode ser executada mecanicamente num período de tempo finito com uma quantidade de esforço finita.

No contexto do Cálculo Científico,
um algoritmo é uma sequência finita de cálculos elementares, tal como definidos atrás.

Um programa corresponde a um algoritmo escrito numa linguagem de programação (linguagem que é entendida pelo computador).

Exercício:

De acordo com a associatividade da soma em $\mathbb{R}$, a soma de três números reais positivos $S = a + b + c$ pode ser feita usando os algoritmos

$$(1)\, S_1 = (a + b) + c; \qquad (2)\, S_2 = a + (b + c).$$

Para $a = 2745.56789$, $b = 34.68734409$, $c = 0.0003$, efetuar o cálculo de S num sistema $\mathbb{F}(10, 7)$ com arredondamento simétrico usando os dois algoritmos. Comentar os resultados.

Exercício: Algoritmos para calcular S

$S_1 = (a+b)+c$	$S_2 = a+(b+c)$
Algoritmo 1	Algoritmo 2
$z_1 = a+b$	$w_1 = b+c$
$z_2 = z_1 + c$	$w_2 = a + w_1$

Resposta:

Representação dos números $a = 2745.56789$, $b = 34.68734409$ e $c = 0.0003$ em $\mathbb{F}$:

$$\mathrm{fl}(a) = 0.2745568 \times 10^4, \quad \mathrm{fl}(b) = 0.3468734 \times 10^2,$$

$$\mathrm{fl}(c) = 0.3000000 \times 10^{-3}.$$

Utilizando o Algoritmo 1: $z_1 = a + b$, $z_2 = z_1 + c$, para calcular S_1 em $\mathbb{F}$, obtém-se

$$\begin{aligned}
\widetilde{z}_1 &= \mathrm{fl}(0.2745568 \times 10^4 + 0.3468734 \times 10^2) \\
&= \mathrm{fl}(0.2745568 \times 10^4 + 0.003468734 \times 10^4) \\
&= \mathrm{fl}(0.278025534 \times 10^4) = 0.2780255 \times 10^4 \\
\widetilde{z}_2 &= \mathrm{fl}(0.2780255 \times 10^4 + 0.3000000 \times 10^{-3}) \\
&= \mathrm{fl}(0.2780255 \times 10^4 + 0.00000003 \times 10^4) \\
&= \mathrm{fl}(0.27802553 \times 10^4) = 0.2780255 \times 10^4 = (S_1)_{\mathbb{F}}.
\end{aligned}$$

Resposta:

A execução em $\mathbb{F}$ do Algoritmo 2: $w_1 = b + c$, $w_2 = a + w_1$, produz o seguinte valor para S_2:

$$
\begin{aligned}
\widetilde{w_1} &= \mathrm{fl}(0.3468734 \times 10^2 + 0.3000000 \times 10^{-3}) \\
&= \mathrm{fl}(0.3468734 \times 10^2 + 0.000003 \times 10^2) \\
&= \mathrm{fl}(0.3468764 \times 10^2) \\
&= 0.3468764 \times 10^2 \\
\widetilde{w_2} &= \mathrm{fl}(0.2745568 \times 10^4 + 0.3468764 \times 10^2) \\
&= \mathrm{fl}(0.278025564 \times 10^4) \\
&= 0.2780256 \times 10^4 = (S_2)_{\mathbb{F}}.
\end{aligned}
$$

Resposta:

Obtivémos

$$(S_1)_{\mathbb{F}} = 0.2780255 \times 10^4 \qquad (S_2)_{\mathbb{F}} = 0.2780256 \times 10^4$$

Tem-se $S = 2780.25553409$ (cálculo efetuado com mais precisão) e

$$\left|\delta_{(S_1)_{\mathbb{F}}}\right| = \frac{|S - (S_1)_{\mathbb{F}})|}{S} \approx 1.92101 \times 10^{-7},$$

$$\left|\delta_{(S_2)_{\mathbb{F}}}\right| = \frac{|S - (S_2)_{\mathbb{F}})|}{S} \approx 1.67578 \times 10^{-7}.$$

Comentário:

- Ambos os cálculos apresentam erros relativos muito pequenos.
- Como $(a + b) + c \neq a + (b + c)$ em $\mathbb{F}$, conclui-se que

  a adição não é associativa em sistemas de ponto flutuante.

Propagação de erros

Plano

- Propagação de erros nas operações aritméticas elementares
- Propagação de erros em funções univariadas
- Propagação de erros em funções multivariadas
- Propagação de erros em algoritmos: pressupõe execução em sistemas de ponto flutuante $\mathbb{F}$, pelo que os erros de arredondamento em $\mathbb{F}$ serão tidos em consideração

Propagação de erros nas operações aritméticas elementares

Sejam $x, y \in \mathbb{R} \setminus \{0\}$ e $\tilde{x}, \tilde{y} \in \mathbb{R}$ valores aproximados.
Recordamos a notação para os erros:

$$e_{\tilde{x}} = x - \tilde{x}, \qquad e_{\tilde{y}} = y - \tilde{y}$$

$$\delta_{\tilde{x}} = \frac{e_{\tilde{x}}}{x}, \qquad \delta_{\tilde{y}} = \frac{e_{\tilde{y}}}{y}.$$

Para já, supomos que as operações são efetuadas em $\mathbb{R}$:

- Soma: $x + y \approx \tilde{x} + \tilde{y}$ (em $\mathbb{R}$)
  Os erros satisfazem

$$e_{\tilde{x}+\tilde{y}} = (x+y) - (\tilde{x}+\tilde{y}) = e_{\tilde{x}} + e_{\tilde{y}}$$

  e

$$\begin{aligned} \delta_{\tilde{x}+\tilde{y}} = \frac{e_{\tilde{x}+\tilde{y}}}{x+y} &= \frac{1}{x+y} e_{\tilde{x}} + \frac{1}{x+y} e_{\tilde{y}} \\ &= \frac{x}{x+y} \frac{e_{\tilde{x}}}{x} + \frac{y}{x+y} \frac{e_{\tilde{y}}}{y} \\ &= \frac{x}{x+y} \delta_{\tilde{x}} + \frac{y}{x+y} \delta_{\tilde{y}} \end{aligned}$$

Propagação de erros nas operações aritméticas elementares

$$\boxed{\delta_{\tilde{x}+\tilde{y}} = \frac{x}{x+y}\delta_{\tilde{x}} + \frac{y}{x+y}\delta_{\tilde{y}}}$$

Podemos supor que $x, y > 0$. Se os erros relativos de $\tilde{x}$ e $\tilde{y}$ são pequenos, ou seja, se

$$|\delta_{\tilde{x}}|, |\delta_{\tilde{y}}| \ll 1$$

então o erro relativo da soma também é pequeno pois

$$|\delta_{\tilde{x}+\tilde{y}}| \leq \frac{x}{x+y}|\delta_{\tilde{x}}| + \frac{y}{x+y}|\delta_{\tilde{y}}|$$

e $\frac{x}{x+y}, \frac{y}{x+y} < 1$.

Propagação de erros nas operações aritméticas elementares

- Subtração: $x - y \approx \tilde{x} - \tilde{y}$
  Os erros inerentes a esta aproximação são dados por

$$e_{\tilde{x}-\tilde{y}} = (x - y) - (\tilde{x} - \tilde{y}) = e_{\tilde{x}} - e_{\tilde{y}}$$

  e

$$\begin{aligned} \delta_{\tilde{x}-\tilde{y}} = \frac{e_{\tilde{x}-\tilde{y}}}{x - y} &= \frac{1}{x - y} e_{\tilde{x}} - \frac{1}{x - y} e_{\tilde{y}} \\ &= \frac{x}{x - y} \delta_{\tilde{x}} - \frac{y}{x - y} \delta_{\tilde{y}}. \end{aligned}$$

Ao contrário do que acontece com a soma, $\delta_{\tilde{x}}$ e $\delta_{\tilde{y}}$ podem ser muito pequenos e $\delta_{\tilde{x}-\tilde{y}}$ ser muito grande, concretamente quando x e y são números positivos muito próximos.
À perda de precisão daí resultante dá-se o nome de

Cancelamento subtrativo.

Exemplo: Cancelamento subtrativo

$$f(x) := \frac{1}{x} - \frac{1}{x+1} \quad (x \in \mathbb{R} \setminus \{-1, 0\})$$

O cálculo de

$$f(10^{20}) \quad (\neq 0)$$

no Matlab R2015b (IEEE-754 com precisão dupla) forneceu:

» format long

» $1/10^{20} - 1/(10^{20} + 1)$
ans = 0

Este resultado ans = 0 tem erro de 100% (erro muito grande).

Esta expressão de $f(x)$ envolve a subtração de números muito próximos quando $x \gg 1$.

Propagação de erros nas operações aritméticas elementares

- Multiplicação: $x \times y \approx \tilde{x} \times \tilde{y}$

$$\begin{aligned} e_{\tilde{x}\times\tilde{y}} &= x \times y - \tilde{x} \times \tilde{y} = x \times y - (\tilde{x} - x + x) \times (\tilde{y} - y + y) \\ &= x \times y - (x - e_{\tilde{x}}) \times (y - e_{\tilde{y}}) \\ &= y e_{\tilde{x}} + x e_{\tilde{y}} - e_{\tilde{x}} e_{\tilde{y}} \end{aligned}$$

e

$$\begin{aligned} \delta_{\tilde{x}\times\tilde{y}} &= \frac{e_{\tilde{x}\times\tilde{y}}}{x \times y} \\ &= \frac{y}{x \times y} e_{\tilde{x}} + \frac{x}{x \times y} e_{\tilde{y}} - \frac{e_{\tilde{x}} e_{\tilde{y}}}{x \times y} \\ &= \delta_{\tilde{x}} + \delta_{\tilde{y}} - \delta_{\tilde{x}} \delta_{\tilde{y}}. \end{aligned}$$

Supondo que $|\delta_{\tilde{x}}|, |\delta_{\tilde{y}}| \ll 1$, podemos tomar a aproximação (linearização dos erros)

$$\delta_{\tilde{x}\times\tilde{y}} \approx \delta_{\tilde{x}} + \delta_{\tilde{y}}.$$

Cálculo em sistemas de ponto flutuante

- Divisão

  Na aproximação $\frac{x}{y} \approx \frac{\tilde{x}}{\tilde{y}}$ ocorrem os seguintes erros

$$e_{\frac{\tilde{x}}{\tilde{y}}} = \frac{x}{y} - \frac{\tilde{x}}{\tilde{y}} = \frac{x}{y} - \frac{x - e_{\tilde{x}}}{y - e_{\tilde{y}}} = \frac{y e_{\tilde{x}} - x e_{\tilde{y}}}{y(y - e_{\tilde{y}})}$$

  e

$$\begin{aligned} \delta_{\frac{\tilde{x}}{\tilde{y}}} &= \frac{e_{\frac{\tilde{x}}{\tilde{y}}}}{\frac{x}{y}} = \frac{y e_{\tilde{x}} - x e_{\tilde{y}}}{x(y - e_{\tilde{y}})} = \frac{y}{y - e_{\tilde{y}}} \left(\frac{1}{x} e_{\tilde{x}} - \frac{1}{y} e_{\tilde{y}} \right) \\ &= \frac{1}{1 - \delta_{\tilde{y}}} \left(\delta_{\tilde{x}} - \delta_{\tilde{y}} \right). \end{aligned}$$

  Supondo que $|\delta_{\tilde{y}}| \ll 1$, é válida a aproximação (linearização dos erros)

$$\delta_{\frac{\tilde{x}}{\tilde{y}}} \approx \delta_{\tilde{x}} - \delta_{\tilde{y}}.$$

Propagação de erros no cálculo de funções univariadas

Seja $f \in C^2(I)$ onde I é um intervalo (compacto, conexo) que contém x e $\tilde{x}$.
Consideremos as aproximações $x \approx \tilde{x}$ e $f(x) \approx f(\tilde{x})$ em $\mathbb{R}$.
Da expansão de Taylor

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

onde θ depende de x e $\tilde{x}$, obtém-se a seguinte relação para os erros de $\tilde{x}$ e $f(\tilde{x})$

$$e_{f(\tilde{x})} = f'(x) e_{\tilde{x}} - \frac{f''(x + \theta(\tilde{x} - x))}{2} e_{\tilde{x}}^2.$$

Para os erros relativos tem-se

$$\delta_{f(\tilde{x})} = \frac{x f'(x)}{f(x)} \delta_{\tilde{x}} - \frac{x^2 f''(x + \theta(\tilde{x} - x))}{2 f(x)} \delta_{\tilde{x}}^2$$

e supondo que $|\delta_{\tilde{x}}| \ll 1$, podemos tomar a aproximação

$$\delta_{f(\tilde{x})} \approx \frac{x f'(x)}{f(x)} \delta_{\tilde{x}}.$$

Propagação de erros no cálculo de funções multivariadas

Seja $f : D \subseteq \mathbb{R}^n \to \mathbb{R}$ (D é convexo).
Sejam $x, \tilde{x} \in D$, tais que $x = (x_1, ..., x_n) \approx (\tilde{x}_1, ..., \tilde{x}_n)$, o que leva a tomar a aproximação

$$f(x) \approx f(\tilde{x}).$$

Supondo $f \in C^2(D)$, partimos da expansão de Taylor

$$\begin{aligned} f(\tilde{x}) &= f(x) + (\tilde{x} - x) \cdot \nabla f(x) \\ &\qquad + \frac{1}{2}(\tilde{x} - x)^\top H_f(x + \theta(\tilde{x} - x))(\tilde{x} - x), \quad 0 < \theta < 1, \end{aligned}$$

onde H_f é a matriz Hesseana de f, para obter a seguinte relação entre $e_{f(\tilde{x})}$ e $e_{\tilde{x}}$

$$e_{f(\tilde{x})} = \sum_{k=1}^{n} \frac{\partial f}{\partial x_k}(x) e_{\tilde{x}_k} - \frac{1}{2} \sum_{k,l=1}^{n} e_{\tilde{x}_k} e_{\tilde{x}_l} \frac{\partial f^2}{\partial x_k \partial x_l}(x + \theta(\tilde{x} - x)).$$

Propagação de erros no cálculo de funções multivariadas

Introduzindo os erros relativos $\delta_{\tilde{x}_i}$, $i = 1, ..., n$, e desprezando o termo

$$\frac{1}{2}\sum_{k,l=1}^{n} \delta_{\tilde{x}_k}\delta_{\tilde{x}_l} \frac{x_k x_l \frac{\partial^2 f}{\partial x_k \partial x_l}(x + \theta(\tilde{x} - x))}{f(x)},$$

na hipótese de erros relativos pequenos, obtém-se

$$\boxed{\delta_{f(\tilde{x})} \approx \sum_{k=1}^{n} p_{f,k}(x)\delta_{\tilde{x}_k}, \quad p_{f,k}(x) := \frac{x_k \frac{\partial f}{\partial x_k}(x)}{f(x)}.}$$

Os coeficientes $p_{f,1}(x)$,..., $p_{f,n}(x)$ de ponderação dos erros relativos de $\tilde{x}$ chamam-se números de condição de f em x.