Aula 2
Noções sobre Análise de Erros e Estabilidade de Algoritmos.

MS211 - Cálculo Numérico

Marcos Eduardo Valle

Departamento de Matemática Aplicada
Instituto de Matemática, Estatística e Computação Científica
Universidade Estadual de Campinas

Na aula anterior, vimos que o arredondamento em ponto flutuante e as operações em ponto flutuante satisfazem as equações

$$\texttt{fl}\,(x) = x(1+\varepsilon_1) \quad \text{e} \quad x \circledast y = (x * y)(1+\varepsilon_2),$$

com $|\varepsilon_1|, |\varepsilon_2| \leqslant \varepsilon_{mach}$, em que ε_{mach} denota o épsilon da máquina.

Na aula de hoje, veremos brevemente como o arredondamento em ponto flutuante e suas operações aritméticas influenciam na credibilidade do resultado produzido por um método numérico.

Veremos também a noção de estabilidade de um método numérico.

Erro da Adição e Subtração

Considere dois números reais x e y dentro dos limites de representação do sistema.

Na soma desses números em ponto flutuante, tem-se:

$$\begin{aligned} \mathtt{fl}(x) \oplus \mathtt{fl}(y) &= [\mathtt{fl}(x) + \mathtt{fl}(y)](1+\varepsilon_1) \\ &= [x(1+\varepsilon_2) + y(1+\varepsilon_3)](1+\varepsilon_1) \\ &= x + x(\varepsilon_1 + \varepsilon_2 + \varepsilon_1\varepsilon_2) + y + y(\varepsilon_1 + \varepsilon_3 + \varepsilon_1\varepsilon_3) \end{aligned}$$

com $|\varepsilon_1|, |\varepsilon_2|, |\varepsilon_3| \leqslant \varepsilon_{mach}$.

O erro absoluto (EA) da adição é:

$$EA = |\mathtt{fl}(x) \oplus \mathtt{fl}(y) - (x+y)| \leqslant (|x| + |y|)(2\varepsilon_{mach} + \varepsilon_{mach}^2).$$

O erro relativo (ER) da adição é:

$$ER = \frac{|\mathtt{fl}(x) \oplus \mathtt{fl}(y) - (x+y)|}{|x+y|} \leqslant \frac{(|x|+|y|)}{|x+y|}(2\varepsilon_{mach} + \varepsilon_{mach}^2).$$

Note que o erro relativo pode ser grande se $|x+y|$ for pequeno, ou seja, se $x \approx -y$.

Nesse caso, tem-se o chamado **cancelamento subtrativo**.

Exemplo 1

Considere o sistema $F(10, 10, 10, 10)$. Vamos calcular

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

Sabemos que $\sqrt{9876} = 0.9937806599 \times 10^2$ e $\sqrt{9875} = 0.9937303457 \times 10^2$. Logo, a aproximação de x é

$$\bar{x} = 0.5031420000 \times 10^{-2}.$$

Os quatro dígitos finais não possuem nenhum significado.
O valor correto, escrito com 10 dígitos significativos, é

$$x = 0.5031418679 \times 10^{-2}.$$

Exemplo 1

Considere o sistema $F(10, 10, 10, 10)$. Vamos calcular

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

Pode-se o cancelamento subtrativo considerando a identidade:

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

Com efeito, nesse caso encontramos:

$$\tilde{x} = \frac{1}{\sqrt{9876} + \sqrt{9875}} = 0.5031418679 \times 10^{-2},$$

que é a representação do resultado correto no sistema.

Exemplo 2

Vamos resolver a equação

$$x^2 - 1634x + 2 = 0,$$

no sistema $F(10, 10, 10, 10)$.

Teoricamente, as soluções dessa equação são

$$x_1 = \frac{1634 + \sqrt{(1634)^2 - 4(2)}}{2} \text{ e } x_2 = \frac{1634 - \sqrt{(1634)^2 - 4(2)}}{2}.$$

Ao efetuar as contas no sistema de ponto flutuante, encontramos:

$$\bar{x}_1 = 0.1633998776 \times 10^3 \quad \text{e} \quad \bar{x}_2 = 0.1224000000 \times 10^{-2}.$$

Os seis zeros na mantissa de x_2 são resultado do cancelamento; não possuem significado algum!

Exemplo 2

Vamos resolver a equação

$$x^2 - 1634x + 2 = 0,$$

no sistema $F(10, 10, 10, 10)$.

Podemos obter um resultado melhor para x_2 lembrando que

$$x_1 x_2 = 2.$$

Logo,

$$\tilde{x}_2 = \frac{2}{x_1} = 0.1223991125 \times 10^{-2}.$$

Nesse caso, todos os dígitos estão corretos!

Além do cancelamento, podemos observar a propagação do erro.

Em termos gerais, a **propagação de erro** está relacionada a perda de dígitos significativos de uma sequencia de operações aritméticas.

Exemplo 3

Determine

$$S = \underbrace{0.333 + 0.333 + \ldots + 0.333}_{10 \text{ vezes}},$$

no sistema $F(10, 3, 5, 5)$.

No sistema de ponto flutuante, calculamos

$$\begin{aligned}\bar{S} &= (((0.333 \times 10^0 \oplus 0.333 \times 10^0) \oplus \ldots \oplus 0.333 \times 10^0) \oplus 0.333 \times 10^0 \\ &= 0.331 \times 10^1.\end{aligned}$$

Obteríamos um resultado melhor se calcularmos

$$\tilde{S} = 10 \otimes 0.333 \times 10^0 = 0.333 \times 10^1.$$

Exemplo 4

Use a série de Taylor para calcular $t = e^{-5.25}$ no sistema $F(10, 5, 10, 10)$.

Sabemos que

$$e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!} \approx \sum_{k=0}^{k_{max}} \frac{x^k}{k!}.$$

Para $k_{max} = 1$, encontramos

$$t_1 = 0.10000 \times 10^1 \oplus -0.52500 \times 10^1 = -0.42500 \times 10^1.$$

Para $k_{max} = 2$, encontramos

$$t_2 = t_1 \oplus 0.13781 \times 10^2 = 0.95600 \times 10^1.$$

Exemplo 4

Use a série de Taylor para calcular $t = e^{-5.25}$ no sistema $F(10, 5, 10, 10)$.

Prosseguindo dessa forma, obtemos:

$$\bar{t} = 0.65974 \times 10^{-2}.$$

O resultado correto é:

$$t = 0.52475 \times 10^{-2}.$$

Observe que há erro no primeiro dígito!

Exemplo 4

Use a série de Taylor para calcular $t = e^{-5.25}$ no sistema $F(10, 5, 10, 10)$.

Uma forma mais eficiente de determinar o valor de t consiste em usar a série de Taylor truncada para determinar

$$\texttt{fl}\left(e^{5.25}\right) = 0.19057 \times 10^{3}.$$

Assim, outra aproximação para t é

$$\tilde{t} = (0.10000 \times 10^{1}) \oslash (0.19057 \times 10^{3}) = 0.52475 \times 10^{-2}.$$

Exemplo 5

Avalie o polinômio

$$P(x) = x^3 - 6x^2 + 4x - 0.1,$$

no ponto 5.24 usando o sistema $F(10, 3, 5, 5)$ e compare o resultado com o valor exato $P(5.24) = -0.00776$.

Primeiramente, determinamos

$$x^2 = x \otimes x = 0.275 \times 10^2 \quad \text{e} \quad x^3 = x^2 \otimes x = 0.144 \times 10^3.$$

Os termos do polinômio serão

$$\texttt{fl}\left(x^3\right) = 0.144 \times 10^3, \quad \texttt{fl}\left(0.1\right) = 0.100 \times 10^0,$$

$$\texttt{fl}\left(6x^2\right) = 0.600 \times 10^1 \otimes 0.275 \times 10^2 = 0.165 \times 10^3,$$

$$\texttt{fl}\left(4x^2\right) = 0.400 \times 10^1 \otimes 0.524 \times 10^1 = 0.210 \times 10^2.$$

Exemplo 5

Avalie o polinômio

$$P(x) = x^3 - 6x^2 + 4x - 0.1,$$

no ponto 5.24 usando o sistema $F(10, 3, 5, 5)$ e compare o resultado com o valor exato $P(5.24) = -0.00776$.

Dessa forma, teremos:

$$\begin{aligned}\bar{P}(5.24) &= \left((0.144 \times 10^3 \ominus 0.165 \times 10^3) \oplus 0.210 \times 10^2\right) \oplus 0.100 \times 10^0 \\ &= -0.100 \times 10^0\end{aligned}$$

O erro relativo (ER) é

$$ER = \frac{|P(5.24) - \bar{P}(5.24)|}{|P(5.24)|} = 11.89.$$

Exemplo 5

Avalie o polinômio

$$P(x) = x^3 - 6x^2 + 4x - 0.1,$$

no ponto 5.24 usando o sistema $F(10, 3, 5, 5)$ e compare o resultado com o valor exato $P(5.24) = -0.00776$.

Alternativamente, podemos escrever:

$$P(x) = x(x(x - 6) + 4) - 0.1.$$

Usando essa expressão, encontramos:

$$\begin{aligned} \texttt{fl}\,(x - 6) &= 0.524 \times 10^1 \ominus 0.600 \times 10^1 = -0.398 \times 10^1, \\ \texttt{fl}\,(x(x - 6) + 4) &= (0.524 \times 10^1 \otimes -0.398 \times 10^1) \oplus 0.400 \times 10^1 \\ &= 0.200 \times 10^{-1} \end{aligned}$$

Exemplo 5

Avalie o polinômio

$$P(x) = x^3 - 6x^2 + 4x - 0.1,$$

no ponto 5.24 usando o sistema $F(10, 3, 5, 5)$ e compare o resultado com o valor exato $P(5.24) = -0.00776$.

Finalmente,

$$\hat{P}(5.24) = (0.524 \times 10^1 \otimes 0.2 \times 10^{-1}) \ominus 0.100 \times 10^0 = 0.5 \times 10^{-2}.$$

Embora o sinal esteja errado, o erro relativo (ER) é

$$ER = \frac{|P(5.24) - \hat{P}(5.24)|}{|P(5.24)|} = 1.64.$$

Algoritmo Preciso

De um modo geral, desejamos resolver um problema $\mathcal{P}$ usando um algoritmo $\mathcal{A}$.

O algoritmo $\mathcal{A}$ pode ser visto como uma sequência de operações que transforma uma entrada x em uma saída y.

Um bom algoritmo deve produzir um erro relativo pequeno, ou seja,

$$\frac{|\mathcal{P}(x) - \mathcal{A}(x)|}{|\mathcal{P}(x)|}$$

deve ser da ordem de ε_{mach}.

Nesse caso, dizemos que o algoritmo é **preciso** (em inglês *accurate*).

Estabilidade

A noção de algoritmo preciso, porém, pode ser um pouco ambiciosa em muitos contextos pois inevitavelmente cometeremos erros de arredondamento em ponto flutuante.

No lugar da precisão, requeremos que o algoritmo seja estável.

Algoritmo Regressivamente Estável

Um algoritmo $\mathcal{A}$ usado para resolver um problema $\mathcal{P}$ é dito regressivamente estável (em inglês *backward stable*) se

$$\mathcal{P}(\tilde{x}) = \mathcal{A}(x)$$

para algum $\tilde{x}$ com

$$\frac{|\tilde{x} - x|}{|x|}$$

da ordem de ε_{mach}.

Exemplo 6

Suponha que a soma $x + y$ é efetuada um sistema de ponto flutuante com épsilon da máquina ε_{mach}. Mostre que $\texttt{fl}(x) \oplus \texttt{fl}(y)$ produzida é regressivamente estável.

Exemplo 6

Suponha que a soma $x + y$ é efetuada um sistema de ponto flutuante com épsilon da máquina ε_{mach}. Mostre que $\texttt{fl}(x) \oplus \texttt{fl}(y)$ produzida é regressivamente estável.

Resposta: Demos mostrar que $\texttt{fl}(x) \oplus \texttt{fl}(y) = \tilde{x} + \tilde{y}$, em que o erro relativo de $\tilde{x}$ e $\tilde{y}$ é da ordem de ε_{mach}. Sabemos que

$$\begin{aligned}\texttt{fl}(x) \oplus \texttt{fl}(y) &= [x(1+\epsilon_2) + y(1+\epsilon_3)](1+\epsilon_1) \\ &= x(1+\epsilon_1)(1+\epsilon_2) + y(1+\epsilon_1)(1+\epsilon_3) \\ &= x(1+\epsilon_1+\epsilon_2+\epsilon_1\epsilon_2) + y(1+\epsilon_1+\epsilon_3+\epsilon_1\epsilon_3)\end{aligned}$$

Escrevendo $\epsilon_4 = \epsilon_1 + \epsilon_2$ e $\epsilon_5 = \epsilon_1 + \epsilon_3$ e desprezando o termos $\epsilon_1\epsilon_2$ e $\epsilon_1\epsilon_3$, que devem ser muito menores que $\epsilon_1, \epsilon_2, \epsilon_3$, tem-se:

$$\texttt{fl}(x) \oplus \texttt{fl}(y) = \underbrace{x(1+\epsilon_4)}_{=\tilde{x}} + \underbrace{y(1+\epsilon_5)}_{=\tilde{y}}, \quad |\epsilon_4|, |\epsilon_5| \leqslant 2\varepsilon_{mach}.$$

Logo, o erro relativo de $\tilde{x}$ e $\tilde{y}$ é menor ou igual à $2\varepsilon_{mach}$.