MS 211 – Calculo Numérico
Lista 01

Aritmética de Ponto Flutuante

Nota: Para os próximos exercícios, considere sempre, quando pedido, um sistema de ponto flutuante da forma $F(b,\ n,\ e_1,\ e_2)$, onde b é a base numérica considerada, n o número de dígitos na mantissa, e_1 e e_2 são o menor e maior expoentes possíveis, respectivamente.

Motivação

A motivação para a presente lista surge de relatos sobre falhas (algumas bem trágicas) devido a propagação de erros numéricos. Na página oficial da disciplina estão disponíveis alguns links para textos (em inglês) sobre tais falhas.

`http://www.ime.unicamp.br/~ms211-cursao/material-didatico`

Exercício 01

Considere um sistema de ponto flutuante $F(b, n, e_1, e_2)$. Responda os itens a seguir, justificando corretamente:

a) Qual o menor número (em módulo) que pode ser representado usando este sistema? E o maior?

b) Qual o número de mantissas possíveis?

c) Mostre que o número de números de pontos flutuantes possíveis é dado por

$$\#F = 2(b-1)b^{n-1}(e_2 - e_1 + 1) + 1.$$

d) É possível existir um sistema de ponto flutuante $F(b, 2, -2, 5)$ com 37 elementos? Justifique com base nos itens anteriores.

Exercício 02

Encontre todos os elementos positivos (em base dez), a cardinalidade, a região de *overflow* e a região de *underflow* para o sistema de ponto flutuante $F(2, 3, -2, 2)$.

Exercício 03

Que soluções admite a equação $1 + x = 1$ no computador onde $F = F(10, 10, -99, 99)$?

Exercício 04

a) Verifique que é possível calcular a abscissa da interseção da reta que passa pelos pontos $(x_0,\, y_0)$ e $(x_1,\, y_1)$ com o eixo-x usando as duas expressões abaixo.

$$x = \frac{x_0 y_1 - x_1 y_0}{y_1 - y_0} \quad \text{e} \quad x = x_0 - \frac{(x_1 - x_0) y_0}{y_1 - y_0}$$

b) Usando os pontos $(9.72,\, 3.08)$ e $(2.16,\, 0.67)$ e três dígitos significativos nos cálculos, calcule a interseção com o eixo x usando as duas expressões. Qual método é melhor? Compare o número de operações e as possíveis ocorrências de erros de arredondamento. Justifique.

Exercício 05

Considere um sistema de ponto flutuante $F(10, 4, -5, 5)$. Pede-se:

a) Qual o maior número representado neste sistema? E o menor?

b) Como será representado o número 85.339 nesta máquina se for usado o arredondamento? E se for usado truncamento?

c) Qual o resultado da seguinte operação nesse sistema?

$$S = 42450 + \sum_{n=1}^{10} 3.$$

d) Fazer o mesmo para a soma

$$S = \sum_{n=1}^{10} 3 + 42450.$$

e) O que você concluiu dos itens c) e d) ?

Exercício 06

Para cada uma das expressões abaixo, reorganize as operações de modo a evitar erros de cálculo ao usar uma aritmética de ponto flutuante. Dê exemplos que evidenciem tais erros para cada caso. Siga o exemplo a seguir:

Exemplo: Suponha que queremos calcular $y = \sqrt{x+1} - \sqrt{x}$ em uma máquina $F(10, 5, -6, 6)$. Ou seja, temos no máximo cinco dígitos na mantissa. Se $x = 100000$, fazendo o cálculo da maneira que está escrito, teríamos um problema, pois nesta máquina, $x + 1 = 100000 + 1 = 100000$ (observe e verifique que o número 100001 não pode ser representado). Assim $y = 0$.

Agora, se escrevemos

$$y = \sqrt{x+1} - \sqrt{x} = \left(\frac{\sqrt{x+1} + \sqrt{x}}{\sqrt{x+1} + \sqrt{x}}\right)(\sqrt{x+1} - \sqrt{x}) = \frac{1}{\sqrt{x} + \sqrt{x+1}},$$

evitamos o cancelamento dos termos, e assim podemos obter

$$y = \frac{1}{2\sqrt{100000}} = 1.5811 \times 10^{-3}.$$

a) $\sqrt{x^2+1}-1$

b) $\ln\left(x-\sqrt{x^2-1}\right)$

c) $(1-\cos(x))/\sin(x)$

d) $(1-\cos(x))/x^2$

e) $x-\sin(x)$

f) $\sqrt{(1+\cos(x))/2}$

Exercício 07

Seja a seguinte equação do segundo grau

$$x^2 + 0.3004\,x + 1.32 \times 10^{-4} = 0, \tag{1}$$

a) Encontre a menor raiz em módulo da seguinte equação com quatro dígitos, utilizando a fórmula de Báskhara e arredondando cada operação. Compare o resultado obtido com a solução exata.

b) Calcule a maior raiz com a mesma precisão e, usando as relações conhecidas entre as raízes e os coeficientes de uma equação do segundo grau, calcule a menor raiz. Compare novamente com a solução exata e com a solução obtida pelo primeiro método.

Justifique suas conclusões.

Exercício 08

Considere os números $\alpha = 0.4321\times10^4$, $\beta = 0.3126\times10^{-3}$ e $\gamma = 0.2583\times10^1$. Calcule o resultado das seguintes operações trabalhando com quatro dígitos e usando primeiro truncamento e, depois, arredondamento. Qual das duas estratégias mais se aproximou em cada caso? Qual foi sua medida para justificar esta maior proximidade? Justifique.

a) $\alpha+\beta+\gamma$

b) α/γ

c) $\alpha\cdot\beta/\gamma$

d) $\beta/\gamma\cdot\alpha$

Exercício 09

A *Função Gama* é definida como

$$\Gamma(x+1) = \int_0^\infty t^x e^{-t}\,dt,$$

e, quando x é um inteiro não negativo, digamos $x=n$, temos que $\Gamma(n+1)=n!$.

Existem duas aproximações para o valor $\log\Gamma(x+1)$, uma delas conhecida como aproximação de Stirling,

$$\log\Gamma(x+1) \approx x\log(x) - x + \frac{1}{2}\log(2\pi x),$$

e outra dada por Bill Gosper, que é uma pequena modificação da primeira

$$\log\Gamma(x+1) \approx x\log(x) - x + \frac{1}{2}\log(2\pi x + \pi/3).$$

Ambas as aproximações ficam mais precisas à medida que o valor de x é aumentado.

a) Qual é o erro relativo nas duas aproximações quando $x = 2$? E para $x = 5$? Lembrando que o erro relativo é definido como

$$E_r = \frac{|\tilde{x} - x_*|}{|x_*|},$$

sendo $\tilde{x}$ a aproximação e x_* o valor exato, lembrando que $x_* \neq 0$. Para $x_* = 0$ o erro relativo não está definido.

b) Estime, por tentativa, o valor de x (a ordem de grandeza de x) para que cada uma das aproximações tenha um erro relativo menor do que 10^{-6}.

Observação: *Provavelmente, se você utilizar uma calculadora (mesmo uma calculadora científica), terá problemas para calcular $n!$ para valores grandes de n (tente encontrar qual é o maior valor que você consegue calcular na sua calculadora). Então, para resolver o item b), utilize um computador.*[1]

Exercício 10

Dê um exemplo de um sistema de ponto flutuante em que não valha a propriedade associativa da adição, ou seja, que dado y, z, e w pertencentes ao sistema, então $(y + z) + w \neq y + (z + w)$.

[1] Atualizado em 11/03/2015

Referências

[1] R. Burden, J. Faires, and A. Burden, *Numerical analysis*, 8 ed., 2013.

[2] S. D. Conte and C. W. D. Boor, *Elementary numerical analysis: an algorithmic approach*, McGraw-Hill Higher Education, 1980.

[3] M. C. C. Cunha, *Métodos Numéricos*, Editora da Unicamp, 2000.

[4] G. Dahlquist and A. Björk, *Numerical Methods*, Prentice-Hall, Englewood Cliffs, NJ, 1974.

[5] N. B. Franco, *Cálculo numérico*, Pearson, 2006.

[6] C. B. Moler, *Numerical computing with MATLAB, electronic edition: The MathWorks*. `http://www.mathworks.com/moler/index_ncm.html`. último acesso em 28-01-2015.

[7] M. A. G. Ruggiero and V. L. d. R. Lopes, *Cálculo numérico: aspectos teóricos e computacionais*, Makron Books do Brasil, 1997.