Aula 1
Representação e Operações Aritméticas em Ponto Flutuante.

MS211 - Cálculo Numérico

Marcos Eduardo Valle

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

Atualmente, além dos estudos em laboratório, muita pesquisa científica é feita por simulação no computador.

Na disciplina de cálculo numérico, estudamos alguns métodos que são usados em simulações numéricas ou para resolução de problemas científicos!

Iniciaremos estudando a representação de números num computador digital (usual).

Representação de um Número Real

Os computadores (usuais) usam a chamada representação em ponto flutuante:

Definição 1 (Sistema de Ponto Flutuante)

Um número real $x \neq 0$ é um ponto flutuante (normalizado) se pode ser escrito como

$$x = \pm 0.d_1 d_2 d_3 \ldots d_t \times \beta^e,$$

em que

- β é a base;
- t é o número de dígitos na mantissa, com $d_1 \neq 0$ e $0 \leqslant d_j \leqslant \beta - 1$, para todo $j = 1, \ldots, t$.
- e é o exponente, com $-m \leqslant e \leqslant M$.

Denotamos por $F(\beta, t, m, M)$ o conjunto de todos os pontos flutuantes para β, t, m e M fixos e adicionando algumas exceções como o zero.

Exemplo 2

Considere o sistema $F(10, 3, 2, 2)$. Represente nesse sistema, se possível, os números:

$$x_1 = 0.35, \quad x_2 = -5.17, \quad x_3 = 0.0123, \tag{1}$$

$$x_4 = 5390, \quad x_5 = 0.0003. \tag{2}$$

Esboce, na reta, o conjunto $F(10, 3, 2, 2)$.

Exemplo 2

Considere o sistema $F(10, 3, 2, 2)$. Represente nesse sistema, se possível, os números:

$$x_1 = 0.35, \quad x_2 = -5.17, \quad x_3 = 0.0123, \tag{1}$$
$$x_4 = 5390, \quad x_5 = 0.0003. \tag{2}$$

Esboce, na reta, o conjunto $F(10, 3, 2, 2)$.

Resposta:

$$x_1 = 0.350 \times 10^0, \quad x_2 = -0.517 \times 10^1, \quad x_3 = 0.123 \times 10^{-1}.$$

O número $5390 = 0.539 \times 10^4$ não pode ser representado porque seu expoente é maior que 2. Tem-se *overflow*.
O número $0.0003 = 0.300 \times 10^{-3}$ não pode ser representado porque seu exponente é menor que -2. Tem-se um *underflow*.

A maioria dos computadores trabalha com a base $\beta = 2$.

Esse não é um problema, pois o mesmo número pode ser representado usando bases diferentes.

Veja no livro texto como é feita a mudança de base!

Muitos *softwares* científicos usam o padrão IEEE **precisão dupla** com 64 bits: 1 para o sinal, 11 para o expoente, 52 para a mantissa.

O padrão IEEE precisão dupla é capaz de representar números positivos entre 1.79×10^{308} e 2.23×10^{-308}, aproximadamente.

O padrão IEEE possui uma representação especial para o zero, $\pm\infty$ (obtido após a divisão por zero), e `NaN` (Not a Number, e.g. $0/0$).

Arredondamento em Ponto Flutuante

O arredondamento em ponto flutuante é usado para representar um número real x, dentro dos limites de representação do sistema, que não pertence ao conjunto $F(\beta, t, m, M)$.

Especificamente, arredondar um número x em ponto flutuante consiste em encontrar $\bar{x} \in F(\beta, t, m, M)$ tal que $|x - \bar{x}|$ seja o menor possível

Denotaremos por `fl` a função que associa um número real x ao seu arredondamento em ponto flutuante.

O valor $|x - \bar{x}|$ é chamado **erro absoluto** de arredondamento.

De um modo similar, o valor $\dfrac{|x - \bar{x}|}{|x|}$ é chamado **erro relativo** de arredondamento.

Exemplo 3

Represente no sistema $F(10, 3, 5, 5)$ os números

$$x_1 = 1234.56, \quad x_2 = -0.00054962, \quad x_3 = 0.9995,$$
$$x_4 = 123456.7, \quad x_5 = 0.0000001.$$

Exemplo 3

Represente no sistema $F(10, 3, 5, 5)$ os números

$$x_1 = 1234.56, \quad x_2 = -0.00054962, \quad x_3 = 0.9995,$$
$$x_4 = 123456.7, \quad x_5 = 0.0000001.$$

Resposta:

$$\mathtt{fl}(x_1) = 0.123 \times 10^4, \quad \mathtt{fl}(x_2) = -0.550 \times 10^{-3},$$
$$\mathtt{fl}(x_3) = 0.100 \times 10^1.$$

Para x_4 e x_5 tem-se *overflow* e *underflow*, respectivamente.

Em linhas gerais, para arredondar um número na base $\beta = 10$, devemos apenas observar o primeiro dígito a ser descartado. Se ele for menor que 5, deixamos os dígitos inalterados; Se ele é maior ou igual a 5, devemos somar 1 ao último dígito remanescente.

Épsilon da Máquina

Definição 4 (Épsilon da Máquina)

O épsilon da máquina, denotado por ε_{mach}, é a metade da distância entre 1 e o menor ponto flutuante estritamente maior que 1.

No padrão IEEE precisão dupla, a precisão é

$$\varepsilon_{mach} = 2^{-52} \approx 2.2 \times 10^{-16}.$$

Exemplo 5

Determine o épsilon da máquina considerando $F(\beta, t, m, M)$.

Épsilon da Máquina

Definição 4 (Épsilon da Máquina)

O épsilon da máquina, denotado por ε_{mach}, é a metade da distância entre 1 e o menor ponto flutuante estritamente maior que 1.

No padrão IEEE precisão dupla, a precisão é

$$\varepsilon_{mach} = 2^{-52} \approx 2.2 \times 10^{-16}.$$

Exemplo 5

Determine o épsilon da máquina considerando $F(\beta, t, m, M)$.

Resposta: A precisão é $\varepsilon = \frac{1}{2}\beta^{1-t}$.

O ε_{mach} fornece um limitante superior para o erro relativo do arredondamento em ponto flutuante.

Especificamente, para qualquer x dentro dos limites de representação do sistema, existe $\bar{x} \in F(\beta, t, m, M)$ tal que

$$|x - \bar{x}| \leqslant \varepsilon |x|.$$

Esta última inequação resulta no seguinte afirmação:

Proposição:

Para qualquer número real x dentro dos limites de representação do sistema, existe ε com $|\varepsilon| \leqslant \varepsilon_{mach}$ tal que

$$\texttt{fl}(x) = x(1 + \varepsilon).$$

Aritmética de Ponto Flutuante

Além de representar números no computador, precisamos também efetuar operações com eles.

As operações aritméticas básicas $+, -, \times$ e $\div$ com números reais, quando realizadas no computador com sistema $F(\beta, t, m, M)$, serão denotadas por $\oplus$, $\ominus$, $\otimes$ e $\oslash$.

As operações aritméticas de ponto flutuante são definidas de modo a satisfazer o axioma:

Axioma das Operações de Ponto Flutuante:

Sejam $*$ uma operação aritmética básica e $\circledast$ seu análogo em ponto flutuante. Para todo $x, y \in F(\beta, t, m, M)$, deve-se ter:

$$x \circledast y = \texttt{fl}(x * y).$$

É razoável assumir que o padrão IEEE satisfaz essa axioma!

Exemplo 6

Considere o sistema $F(\beta, t, m, M)$. Sejam $x = \texttt{fl}(11.4)$, $y = \texttt{fl}(3.18)$ e $z = \texttt{fl}(5.05)$. Efetue as operações:

(a) $(x \oplus y) \oplus z$ e $x \oplus (y \oplus z)$.

(b) $\dfrac{y \otimes x}{z}$ e $\dfrac{y}{z} \otimes x$.

(c) $y \otimes (z \oplus x)$ e $y \otimes z \otimes y \otimes x$.

Exemplo 6

Considere o sistema $F(\beta, t, m, M)$. Sejam $x = \texttt{fl}(11.4)$, $y = \texttt{fl}(3.18)$ e $z = \texttt{fl}(5.05)$. Efetue as operações:

(a) $(x \oplus y) \oplus z$ e $x \oplus (y \oplus z)$.

(b) $\dfrac{y \otimes x}{z}$ e $\dfrac{y}{z} \otimes x$.

(c) $y \otimes (z \oplus x)$ e $y \otimes z \otimes y \otimes x$.

Resposta:

(a) $(x \oplus y) \oplus z = 0.197 \times 10^2$ e $x \oplus (y \oplus z) = 0.196 \times 10^2$.

(b) $\dfrac{x \otimes y}{z} = 0.719 \times 10^1$ e $x \otimes \left(\dfrac{y}{z}\right) = 0.718 \times 10^1$.

(c) $y \otimes (z \oplus x) = 0.523 \times 10^1$ e
$(y \otimes z) \oplus (y \otimes x) = 0.524 \times 10^1$.

Ao contrário das operações com números reais, as operações de ponto flutuante não são nem associativas e nem distributivas!

Em vista do axioma das operações de ponto flutuante, tem-se:

Proposição:

Para quaisquer $x, y \in F(\beta, t, m, M)$, existe ε com $|\varepsilon| \leqslant \varepsilon_{mach}$ tal que

$$x \circledast y = (x * y)(1 + \varepsilon),$$

em que $*$ denota uma operação aritmética básica e $\circledast$ seu análogo em ponto flutuante.

Essa proposição estabelece uma relação entre a operação aritmética com números reais e seu análogo em ponto flutuante.

Essa relação possui um papel importante na análise de erros de algoritmos!

Na próxima aula exploraremos melhor os efeitos dos erros da representação e das operações em ponto flutuante.