Álgebra Linear Numérica 06/04

Aritmética de Ponto Flutuante

- Um computador utiliza uma quantidade finita de bits para representar números reais, logo, ele só pode representar um subset finito dos reais. Isso traz duas dificuldades, os números não podem ser arbitrariamente grandes e há um espaçamento entre os representados pelo computador.

- Atualmente, o problema de overflow e underflow não é um problema comum, já que os computadores armazenam números suficientemente grandes para a maioria dos problemas

- Porém, o problema do gap entre os números ainda existe, como, por exemplo, no IEEE aritmética de precisão dobrada, o intervalo $[1, 2]$ é representado por

$$1, \quad 1+2^{-52}, \quad 1+2\cdot 2^{-52}, \quad 1+3\cdot 2^{-52}, \quad \dots, \quad 2$$

O gap entre os números adjacentes não são maiores que 2^{-52}. É um gap tão pequeno que, em algoritmos estáveis (próximo documento), eles podem ser ignorados, mas algoritmos instáveis podem causar imprecisões

- Vamos definir um conjunto de números de ponto flutuante ideal para trabalharmos

Definição

Seja $F \subset \mathbb{R}$ onde:

$$F = \left\{ \pm\left(\frac{m}{\beta^t}\right)\beta^e \,/\, e \in \mathbb{Z},\ \beta \in \mathbb{N}_{\geq 2},\ t \in \mathbb{N}_{\geq 1},\ 1 \leq m \leq \beta^t \right\} \cup \{0\}$$

onde β é a base ou radix, t é a precisão, e é o expoente, $\pm\left(\frac{m}{\beta^t}\right)$ é a mantissa de $x \in F$

- O sistema é ideal por não ter underflow nem overflow, logo, é um set infinito e contável, além de ser semelhante a si mesmo: $F = \beta F$

- O m ESTÁ SEMPRE EM BASE β (PARA OS CÁLCULOS PODEMOS DESCONSIDERAR ISSO)
- O t INDICA QUANTOS DIGITOS (NA BASE DADA) A MINHA MANTÍSSA POSSUI, OU SEJA, QUANTO MAIOR É t, MAIS NÚMEROS EU POSSO ARMAZENAR NO MEU CONJUNTO F (A QUANTIDADE DE DÍGITOS É ANALISADA NA BASE β)
- POR QUE DIVIDO POR β^t? POIS, AO FAZER ISSO, EU FAÇO COM QUE O NÚMERO INTEIRO QUE TRABALHO, EM BASE β, TENHA EXATAMENTE t DÍGITOS

↳ POSSO ESCREVER 15, EM BASE 10, COMO:

$0{,}15 \cdot 10^2$, $1{,}5 \cdot 10$, $0{,}00015 \cdot 10^5$, . . .

DIVIDIR POR 10^t ME DEIXARÁ COM UM NÚMERO DE CASAS RELEVANTE PARA A REPRESENTAÇÃO

$$\Rightarrow \frac{15}{10^2} = \boxed{0{,}15}$$

$t=2$

$\boxed{\text{DEFINIÇÃO ALTERNATIVA}}$

PODEMOS DEFINIR, TAMBÉM, F COMO:

$$F = \{ 0{,}d_1 d_2 \cdots d_t \cdot \beta^e \,/\, e \in \mathbb{Z} \}$$

COM $0 \leq d_j < \beta$

- UMA DEFINIÇÃO EQUIVALENTE, MAS DE MAIS FÁCIL VISUALIZAÇÃO.

A PARTIR DISSO, PODEMOS DEFINIR CLARAMENTE O MENOR E MAIOR VALOR QUE F CONTÉM (CASO $e \in I$ COM I SENDO UM SUBCONJUNTO FINITO DE $\mathbb{Z}$)

↳ SEJA $t=3$, $e \in [-5, 5]$, $\beta = 10$

O MENOR VALOR ABSOLUTO E NÃO NULO DO CONJUNTO F:

menor $= 0{,}001 \cdot 10^{-5}$

O MAIOR // :

maior $= 0{,}999 \cdot 10^5$

- Quando tentamos representar um número que a precisão não suporta (exemplo, $0{,}3145 \cdot 10^2$, mas $t=3$), podemos arredondar ou truncar o número

Ex:

$$t=3, \quad e \in [-10, 10]$$

$$\to 14{,}38 = 0{,}\underbrace{1438}_{4>3} \cdot 10^2$$

▷ **Arredondar:**

Pegamos os n dígitos relevantes do número $(n>t)$, analisamos então o $(t+1)$-ésimo dígito (tudo isso na base β), então, se o $(t+1)$-dígito for maior que $\frac{\beta}{2}$, eu elimino do $(t+1)$-ésimo dígito em diante e somo 1 ao t-ésimo dígito.

$$0{,}1438 \cdot 10^2 \to \boxed{8>5} \Rightarrow \boxed{0{,}144 \cdot 10^2}$$

Se o $(t+1)$-ésimo dígito é igual a $\frac{\beta}{2}$, passamos ao $(t+2)$-ésimo e assim por diante, se todos são iguais a $\frac{\beta}{2}$, tanto faz a decisão

▷ **Truncamento**

Eliminamos do $(t+1)$-ésimo dígito em diante

$$0{,}1438 \cdot 10^2 \to 0{,}143 \cdot 10^2$$

Epsilon da Máquina

- É um número que resume a resolução de F

Definição

$$\epsilon_{machine} = \frac{1}{2}\beta^{1-t}$$

- Esse número é metade da distância entre 1 e o próximo maior número representável de ponto flutuante
- Ou seja, tecnicamente: $1+\epsilon_{ma} > 1$
- Mas, por que a distância entre 2 números representáveis é β^{1-t}?

↳ AO REPRESENTAR UM NÚMERO x EM BASE β, EU TENHO QUE

(*) $d_1 \dots d_t$ COM $0 \leq d_j < \beta$, EU DESEJO FAZER A MENOR ALTERAÇÃO POSSÍVEL NO MEU NÚMERO, PARA ISSO, EU SIMPLESMENTE SOMO β, LOGO:

(*) TENHO t DÍGITOS SIGNIFICATIVOS

$$x = d_1 \cdot \beta^0 + d_2 \cdot \beta^{-1} + d_3 \cdot \beta^{-2} + \dots + d_t \cdot \beta^{1-t}$$

LOGO, A MENOR ALTERAÇÃO POSSÍVEL PARA O NÚMERO É ACRESCENTAR 1 UNIDADE A d_t, LOGO:

$$d_1 \beta^0 + \dots + (d_t + 1)\beta^{1-t} = d_1 \beta^0 + \dots + d_t \beta^{1-t} + \beta^{1-t} = \frac{x}{\beta^t} + \beta^{1-t}$$

- SE SOMARMOS UM NÚMERO MENOR QUE ϵ_{ma}, PELA PRECISÃO LIMITADA QUE TEMOS, TERÍAMOS $1 + \alpha = 1$ $(\alpha < \epsilon_{ma})$.
- TEMOS ENTÃO A PROPRIEDADE

$$\forall x \in \mathbb{R},\ \exists x' \in F \ /\ |x - x'| \leq \epsilon_{mac} \cdot |x|$$

DEFINIÇÃO

SEJA F O CONJUNTO DE PONTO FLUTUANTE, TEMOS:

$$fl : \mathbb{R} \longrightarrow F$$

ONDE $fl(x)$ DA A APROXIMAÇÃO MAIS PRÓXIMA DE x EM F.

- COM ESSA DEFINIÇÃO, PODEMOS REESCREVER A ÚLTIMA DESIGUALDADE:

$$\forall x \in \mathbb{R},\ \exists \epsilon \ /\ |\epsilon| \leq \epsilon_{mac} ;\ fl(x) = x(1 + \epsilon) \quad \text{(I)}$$

- OU SEJA, A DIFERENÇA ENTRE UM REAL E SUA APROXIMAÇÃO É SEMPRE MENOR OU IGUAL AO ϵ_{mac}

ARITMÉTICA DE PONTO FLUTUANTE

- AS OPERAÇÕES $+, -, \times, \div$ SÃO FEITAS EM REAIS, NO COMPUTADOR, OU SEJA, EM F, VAMOS REPRESENTAR POR $\oplus, \ominus, \otimes, \oslash$.

∘ UM COMPUTADOR ENTÃO É CONSTRUÍDO SEGUINDO O SEGUINTE PRINCÍPIO. SEJAM $x, y \in F$ E $*$ DENOTA UMA DAS OPERAÇÕES $(+, \times, -, \div)$ E $\circledast$ SEU EQUIVALENTE EM F, ENTÃO:

$$x \circledast y = fl(x * y)$$

$\boxed{\text{TEOREMA}}$ (II)

(AXIOMA FUNDAMENTAL DA ARITMÉTICA COM PONTO FLUTUANTE)

$$\forall x, y \in F, \ \exists \epsilon \text{ com } |\epsilon| \leq \epsilon_{mac} \ / \ x \circledast y = (x * y)(1 + \epsilon)$$

∘ OU SEJA, TODA A OPERAÇÃO FEITA EM F TEM UM ERRO DE, NO MÁXIMO, ϵ_{mac}

→ AGORA PODEMOS DEFINIR ϵ_{mac} COMO O MENOR NÚMERO QUE (I) E (II) VALEM.