

CCI-22



Matemática Computacional

CCI-22

2) Erros de arredondamento

Erros de representação e de cálculo

CCI-22

- Tipos de erros
- Sistemas de ponto flutuante
- Arredondamentos
- Erros absolutos e relativos
- Dígitos significativos
- Épsilon da máquina

CCI-22

- **Tipos de erros**
- Sistemas de ponto flutuante
- Arredondamentos
- Erros absolutos e relativos
- Dígitos significativos
- Épsilon da máquina

Tipos de erros

- **Erro inerente:** sempre presente na incerteza das medidas experimentais e nas simplificações dos modelos utilizados
 - Exemplo: (50,3 ± 0,2) cm
- **Erro de aproximação (ou de discretização):** proveniente das limitações dos métodos numéricos (número de iterações, etc.) durante a determinação de um valor
 - Exemplo: para $0 \leq x \leq \pi/4$, $\text{sen } x = \sum_{n=0}^{\infty} (-1)^n \frac{x^{2n+1}}{(2n+1)!}$
- **Erro de representação (de arredondamento):** limitação de um número real ao valor representado com uma quantidade finita de dígitos

Um caso real

- Considere as seguintes equações:

<ul style="list-style-type: none"> ▪ $H = 1,0/2,0$ ▪ $X = 2,0/3,0 - H$ ▪ $Y = 3,0/5,0 - H$ ▪ $E = (X+X+X) - H$ ▪ $F = (Y+Y+Y+Y) - H$ ▪ $G = F/E$ 	Cálculo analítico: <ul style="list-style-type: none"> $H = 1/2$ $X = 1/6$ $Y = 1/10$ $E = 0$ $F = 0$ indeterminado
--	---
- Resultados em diferentes máquinas:

HP 25	Texas SR50	IBM 4341
H = 0,5	H = 0,5	H = 0,5
X = 0,166666667	X = 0,166666667	X = 0,166666
Y = 0,1	Y = 0,1	Y = 0,1
E = 1,0 E-10	E = 2,0 E-13	E = 0,119209 E-6
F = 0	F = 0	F = -0,178813 E-6
G = inválido	G = inválido	G = 0,666...

CCI-22

- Tipos de erros
- **Sistemas de ponto flutuante**
- Arredondamentos
- Erros absolutos e relativos
- Dígitos significativos
- Épsilon da máquina

Definições

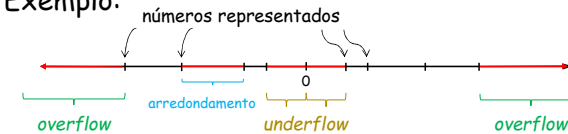
- Um número real x está representado como *float* normalizado se:
 - $x = \pm 0, d_1 d_2 \dots d_n \cdot b^e$, $n \in \mathbb{N}$
 - $1 \leq d_1 < b$, $0 \leq d_i < b$ para $2 \leq i \leq n$
 - $e_1 \leq e \leq e_2$, com $e_1 \leq 0$, $e_2 \geq 1$, onde $e_1, e_2 \in \mathbb{Z}$
- Um *sistema de ponto flutuante* é definido pela quádrupla $F(b, n, e_1, e_2)$, e contém todos os *floats* normalizados representados por esses valores, incluindo também o número zero
- Observação: o número zero é representado como $0, \underbrace{00 \dots 0}_n \cdot b^{e_1}$
n vezes

Sistemas de ponto flutuante

- Como vimos, um sistema de ponto flutuante é uma quádrupla $F(b,n,e_1,e_2)$
- Alguns exemplos reais:
 - HP 25: $F(10,9,-98,100)$
 - Texas SR50: $F(10,10,-98,100)$
 - IBM 370: $F(16,6,-64,63)$
 - B6700: $F(8,13,-51,77)$
- No sistema $F(b,n,e_1,e_2)$:
 - Menor número representável (em módulo): $N_{\min} = 0,100\dots 0.b^{e_1}$
 - Maior número representável: $N_{\max} = 0,[b-1][b-1]\dots[b-1].b^{e_2}$
 - Total de representações: $\#F = 2 \cdot (b-1) \cdot b^{n-1} \cdot (e_2 - e_1 + 1) + 1$

Erros na representação de floats

- A representação de um número real com uma quantidade finita de bits geralmente acarreta a necessidade do arredondamento, que é um erro em relação ao seu valor exato
- Casos especiais:
 - Overflow**: número a ser representado tem módulo maior que N_{\max}
 - Underflow**: número a ser representado tem módulo menor que N_{\min}
- Exemplo:



Exemplo

- Seja $F(2,3,-1,2)$
- Possíveis mantissas: 0,100; 0,101; 0,110; 0,111
- Exponentes: -1, 0, 1, 2
- Valores positivos:

e	b^e	0,100	0,101	0,110	0,111
-1	1/2	1/4	5/16	3/8	7/6
0	1	1/2	5/8	3/4	7/8
1	2	1	5/4	3/2	7/4
2	4	2	5/2	3	7/2

- $\#F = 2 \cdot (2-1) \cdot 2^{3-1} \cdot (2 - (-1) + 1) + 1 = 2 \cdot 1 \cdot 4 \cdot 4 + 1 = 33$
- Região de underflow: $(-1/4, 0) \cup (0, 1/4)$
- Região de overflow: $(-\infty, -7/2) \cup (7/2, \infty)$
- Sejam $x = 5/4$ e $y = 3/8$. Portanto, $x+y = 13/8$ não está em F: terá que ser arredondado...

Outro exemplo

- Seja $F(3,2,-1,2)$
- Possíveis mantissas: 0,10; 0,11; 0,12; 0,20; 0,21; 0,22
- Exponentes: -1, 0, 1, 2
- Valores positivos:

e	b^e	0,10	0,11	0,12	0,20	0,21	0,22
-1	1/3	1/9	4/27	5/27	2/9	7/27	8/27
0	1	1/3	4/9	5/9	2/3	7/9	8/9
1	3	1	4/3	5/3	2	7/3	8/3
2	9	3	4	5	6	7	8

- $\#F = 2 \cdot (3-1) \cdot 3^{2-1} \cdot (2 - (-1) + 1) + 1 = 2 \cdot 2 \cdot 3 \cdot 4 + 1 = 49$
- Região de underflow: $(-1/9, 0) \cup (0, 1/9)$
- Região de overflow: $(-\infty, -8) \cup (8, \infty)$
- Exercício: neste sistema, encontre dois valores cuja soma terá que ser arredondada

CCI-22

- Tipos de erros
- Sistemas de ponto flutuante
- **Arredondamentos**
- Erros absolutos e relativos
- Dígitos significativos
- Épsilon da máquina

Definição de arredondamento

- Seja um sistema de ponto flutuante $F(b, n, e_1, e_2)$. Uma função $\xi: \mathbb{R} \rightarrow F$ é de arredondamento se:
 - $\forall x \in F, \xi x = x$ (não modifica os valores de F)
 - Além disso, essa função será:
 - De arredondamento para baixo (ou por falta, ou simplesmente arredondamento) se, $\forall x \in \mathbb{R}, \xi x \leq x$ (nesse caso, $\xi x = \nabla x$)
 - De arredondamento para cima (ou por excesso, ou truncamento) se, $\forall x \in \mathbb{R}, \xi x \geq x$ (nesse caso, $\xi x = \Delta x$)
 - Monotônica se, $\forall x, y \in \mathbb{R}, x \leq y \Rightarrow \xi x \leq \xi y$

Arredondamento para float

- Se o sistema de ponto flutuante $F(b, n, e_1, e_2)$ tem a base b par, o arredondamento para o *float* mais próximo, denotado por Ox , é definido como:
 - $\forall x \in [0, N_{\min}], Ox = 0$
 - $\forall x \in [N_{\min}, N_{\max}]$:
 - $Ox = \nabla x$, para $x \in [\nabla x, (\nabla x + \Delta x)/2]$
 - $Ox = \Delta x$, para $x \in [(\nabla x + \Delta x)/2, \Delta x]$
 - Se $x < 0, Ox = -O(-x)$

Exemplo

- No sistema de ponto flutuante $F(10, 4, -98, 10)$:

$x = 0,333333$	$y = 0,348436$	$z = 0,666666\dots$	$w = 0,12345$
$\nabla x = 0,3333$	$\nabla y = 0,3484$	$\nabla z = 0,6666$	$\nabla w = 0,1234$
$\Delta x = 0,3334$	$\Delta y = 0,3485$	$\Delta z = 0,6667$	$\Delta w = 0,1235$
$Ox = 0,3333$	$Oy = 0,3484$	$Oz = 0,6667$	$Ow = 0,1235$

CCI-22

- Tipos de erros
- Sistemas de ponto flutuante
- Arredondamentos
- Erros absolutos e relativos
- Dígitos significativos
- Épsilon da máquina

Erros absolutos e relativos

- A diferença entre o valor exato e o arredondado pode ser medida por dois parâmetros: erro absoluto e erro relativo
- Erro absoluto: $E_A x = |\xi x - x|$
- Erro relativo: $E_R x = E_A x / |\xi x|$
- Pode ser demonstrado que, em um sistema de ponto flutuante $F(b, n, e_1, e_2)$, $\forall x \in \mathbb{R}$, se $|x| \in [N_{\min}, N_{\max}]$, então $E_R x \leq \mu$, onde:

$$\mu = \begin{cases} b^{1-n}/2, & \text{se } \xi x = Ox \\ b^{1-n}, & \text{se } \xi x = \nabla x \end{cases}$$

- Portanto, no padrão IEEE, o erro relativo de arredondamento para o *float* mais próximo nos casos de precisão simples, dupla e estendida, é 2^{-23} , 2^{-52} e 2^{-64} , respectivamente, ou seja, é muito pequeno...

CCI-22

- Tipos de erros
- Sistemas de ponto flutuante
- Arredondamentos
- Erros absolutos e relativos
- Dígitos significativos
- Épsilon da máquina

Dígitos significativos

- Em um sistema de numeração, um dígito é significativo se:
 - for diferente de zero
 - caso seja zero, se não for usado para fixar a vírgula ou preencher o lugar de dígitos descartados
- Exemplos de dígitos significativos (no sistema decimal):
 - 0,008735: 8, 7, 3 e 5
 - 30457: todos
 - 23000: 2 e 3

Dígitos significativos exatos

- Um dígito significativo é exato (DIGSE) se, arredondando-se o número em questão para uma posição imediatamente após à desse dígito, então o erro absoluto não se torna maior que meia unidade na mesma posição do dígito
- Exemplo (no sistema digital):
 - 0,666998 como arredondamento de $2/3$
 - 1º dígito 6: $|0,66 - 0,666\dots| = 0,0066\dots < 0,05$
 - 2º dígito 6: $|0,666 - 0,666\dots| = 0,00066\dots < 0,005$
 - 3º dígito 6: $|0,6669 - 0,666\dots| = 0,000233\dots < 0,0005$
 - 1º dígito 9: $|0,66699 - 0,666\dots| = 0,000323\dots > 0,00005$
 - Portanto, o primeiro dígito 9 não é exato

Precisão e exatidão

- Como já foi definido, a precisão de um computador é o número de dígitos da sua mantissa
- A precisão é algo bem determinado e invariável em cada máquina
- Por outro lado, exatidão é uma medida de perfeição do resultado, que depende da precisão da máquina e do método utilizado em sua obtenção

Exemplos

- Em um computador com sistema F(16,6,-64,63), qual seria a sua precisão decimal?
 - Sabemos que sua precisão hexadecimal é 6
 - $16^6 = 10^p \Leftrightarrow p = 6 \cdot \log 16 \approx 7,2$
 - Portanto, sua precisão decimal é 7
 - No entanto, nem todo número com representação exata de 7 casas decimais terá também uma representação exata nessa máquina. Exemplo: 0,1
- Seja o número irracional $2^{1/2} = 1,414214562\dots$
Podemos dizer que:
 - 1,4142 é mais preciso e exato que 1,41, pois tem maior número de casas decimais e aproxima melhor $2^{1/2}$
 - 1,4149 é mais preciso que 1,414, pois tem mais casas decimais, mas é menos exato

CCI-22

- Tipos de erros
- Sistemas de ponto flutuante
- Arredondamentos
- Erros absolutos e relativos
- Dígitos significativos
- Épsilon da máquina

Épsilon da máquina

- Chamamos de ϵ (épsilon) de uma determinada máquina o menor *float* tal que $1 + \epsilon > 1$
- O programa abaixo imprime esse valor:

```
EpsilonMaq(){
    float epsilon, eps = 1;
    do {
        epsilon = eps;
        eps /= 2;
    }
    while (eps + 1 != 1);
    printf("%E\n", epsilon);
}
```