MÉTODOS NUMÉRICOS I

LISTA 1

1. Primeira Questão
2. Segunda Questão
3. Terceira Questão
4. Quarta Quetão

BASES NUMÉRICAS

Seja **x** um número de **n** dígitos na base **b** e **di** o i-ésimo dígito de **x**. Logo:

$$\mathbf{x} = \mathbf{dn}*b^{\{n-1\}} + \mathbf{dn\text{-}1}*b^{\{n-2\}} + \ldots + \mathbf{d2}*b^{\{1\}} + \mathbf{d1}*b^{\{0\}}$$

EXEMPLOS

Base 10

$$1304 = 1*10^{\{3\}} + 3*10^{\{2\}} + 0*10^{\{1\}} + 4*10^{\{0\}} = 1304$$

Base 2

$$11001 = 1*2^{\{4\}} +1*2^{\{3\}} + 0*2^{\{2\}} + 0*2^{\{1\}} + 1*2^{\{0\}} = 25$$

CONVERSÃO ENTRE BASES

Tarefa de encontrar os **dígitos** que multiplica as potências da base **b**

$$25 = 1*2^{4} + 1*2^{3} + 0*2^{2} + 0*2^{1} + 1*2^{0} = 11001$$

DECIMAL - DECIMAL

$$\begin{array}{r|l} 9573 & 10 \\ \hline 3 & 957 \end{array}$$

Apagamos o último dígito, pois
já temos ele
ou
Fizemos o descolamento de um
dígito para a esquerda

Resto da divisão = último digito

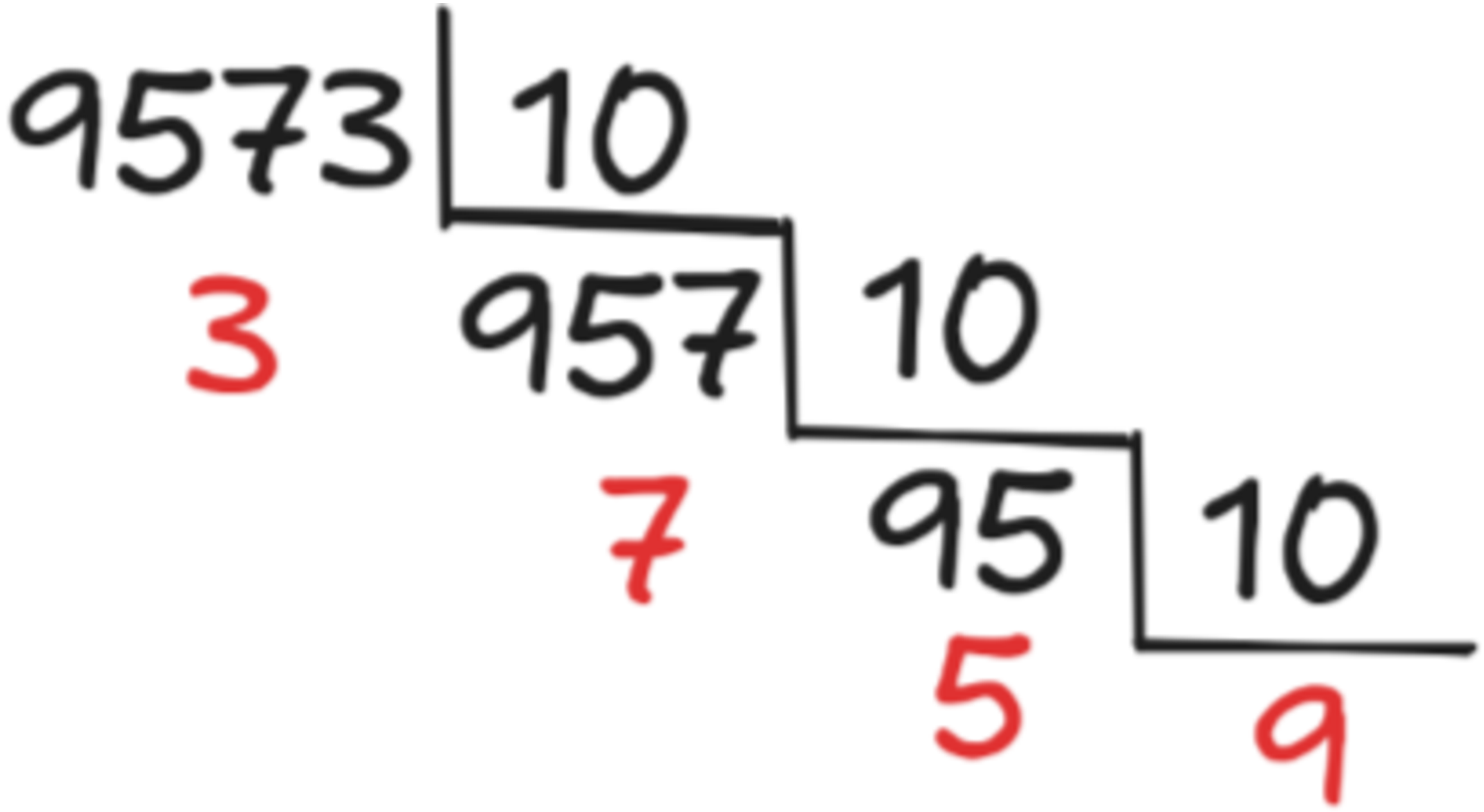
DECIMAL - DECIMAL

9573 | 10

3   957 | 10

7   95

DECIMAL - DECIMAL



DECIMAL - BINÁRIO

$$\begin{array}{r|l} 27 & 2 \\ \cline{2-2} 1 & 13 \end{array}$$

Apagamos o último dígito, pois
já temos ele
ou
Fizemos o descolamento de um
dígito para a esquerda

Resto da divisão = último digito

DECIMAL - BINÁRIO

27 | 2

1   13

Apagamos o último dígito, pois já temos ele

ou

Fizemos o descolamento de um dígito para a esquerda

QUE DESLOCAMENTO É ESSE?
O RESULTADO NÃO DEVERIA SER 2?

Resto da divisão = último digito

DECIMAL - BINÁRIO

27 = 11011

$$\begin{array}{r|l} 27 & 2 \\ \hline 1 & 13 \end{array}$$

13 = 1101

Resto da divisão = último digito

Apagamos o último dígito, pois já temos ele
ou
Fizemos o descolamento de um dígito para a esquerda

QUE DESLOCAMENTO É ESSE? O RESULTADO NÃO DEVERIA SER 2?

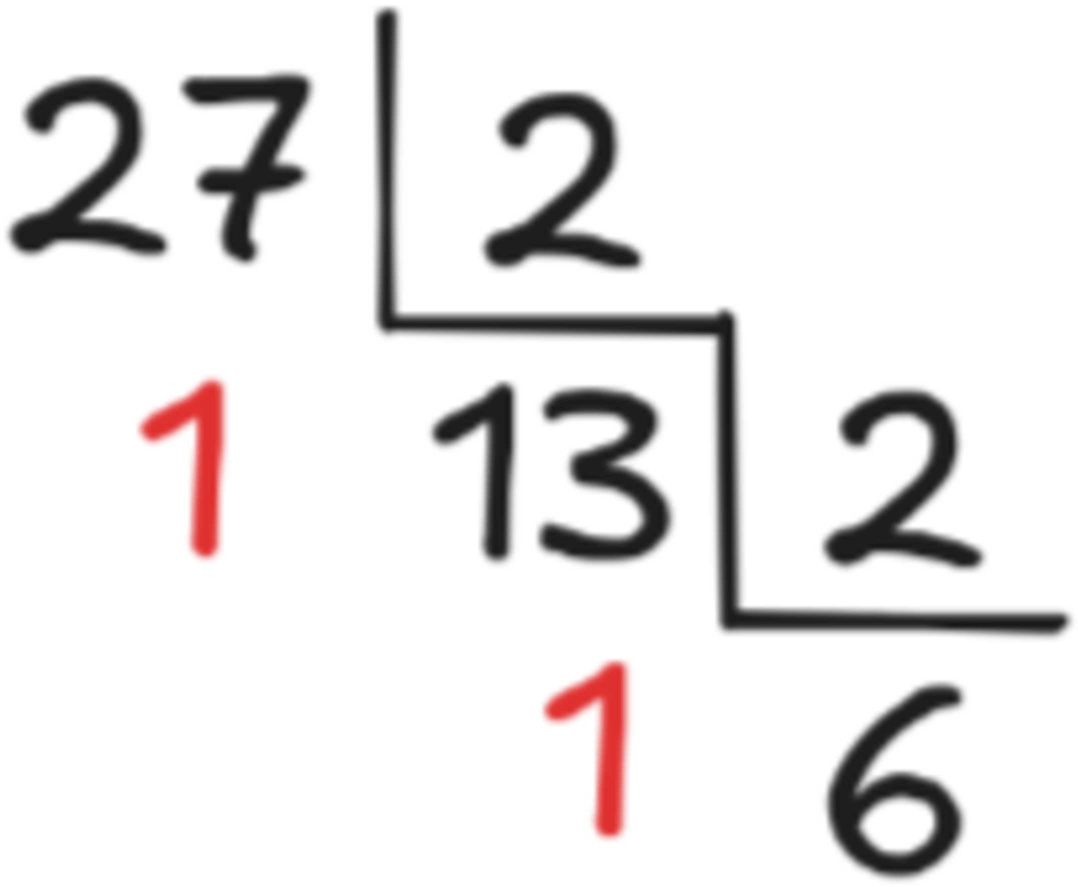
OLHE PARA A REPRESENTAÇÃO BINÁRIA E VEJA O QUE ESTÁ ACONTECENDO

DECIMAL - BINÁRIO

27 = 11011

13 = 1101

6 = 110



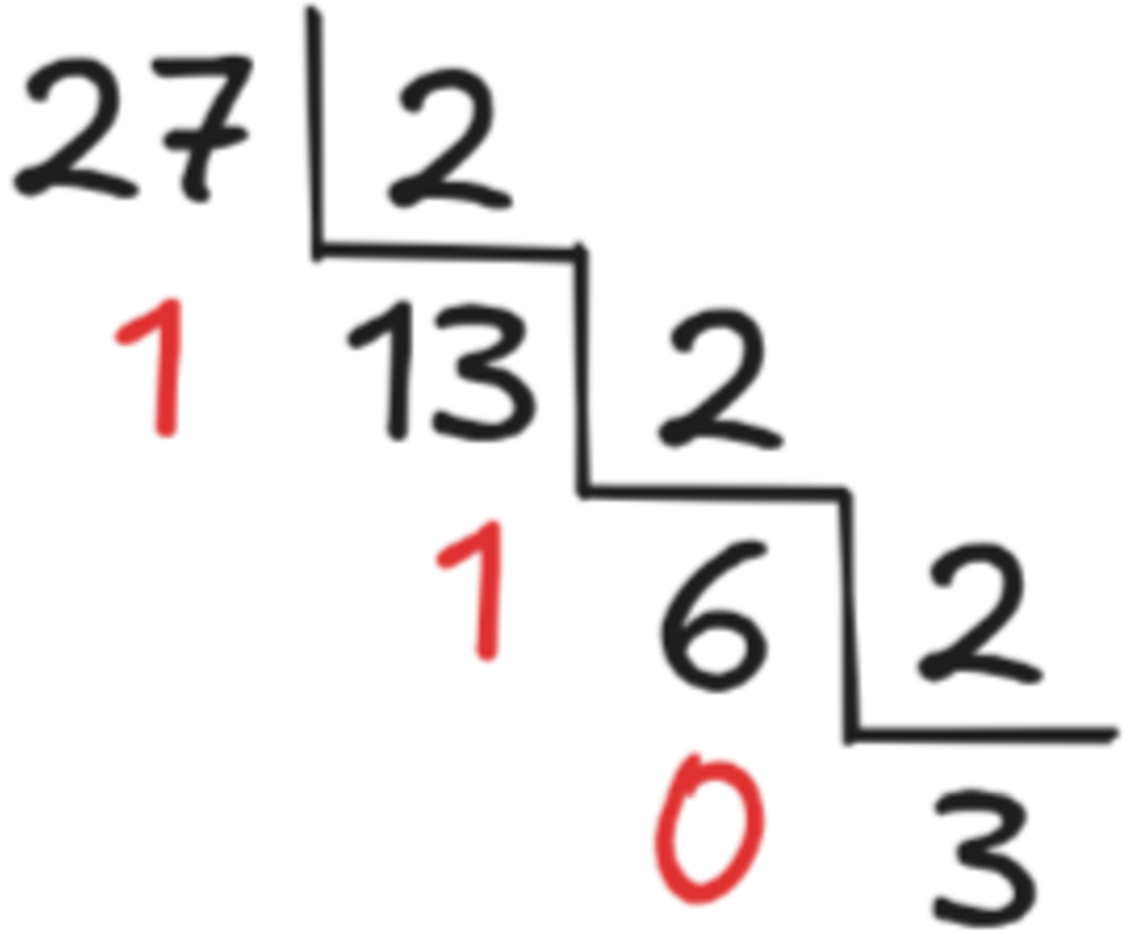
DECIMAL - BINÁRIO

27 = 11011

13 = 1101

6 = 110

3 = 11



DECIMAL - BINÁRIO

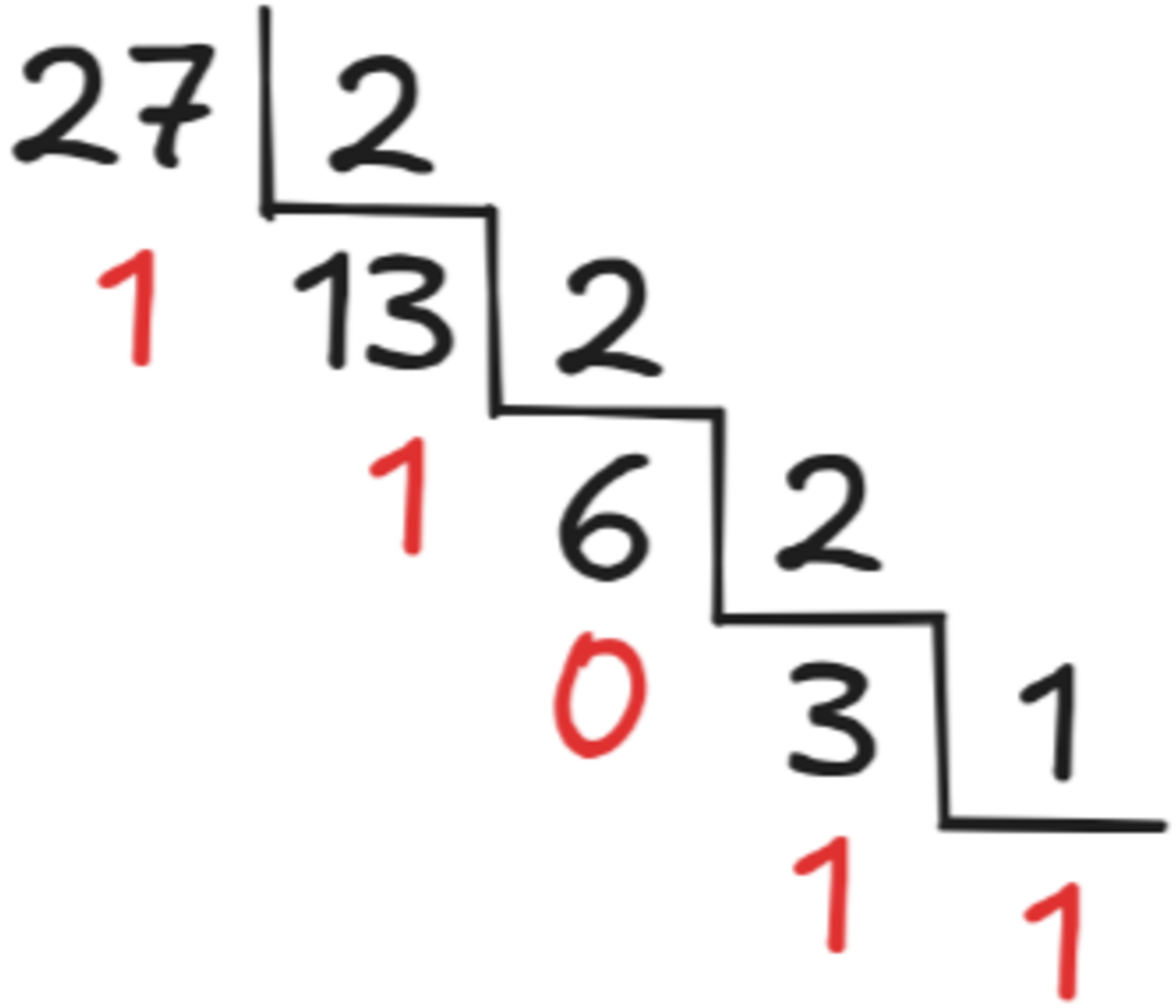
27 = 11011

13 = 1101

6 = 110

3 = 11

1 = 1



CURIOSIDADE

OPERADORES DE DESLOCAMENTO DE BITS

111 >> 1 = 11

111 << 1 = 1110

CURIOSIDADE

QUAL O RESULTADO DE

27 << 1?

e de

27 << 2?

CURIOSIDADE

Observe o que acontece nos bits quando os deslocamos para a esquerda

11011 << 1 =

$$(1*2^{\{4\}}+1*2^{\{3\}}+0*2^{\{2\}}+1*2^{\{1\}}+1*2^{\{0\}}) \ll 1 =$$

$$1*2^{\{5\}}+1*2^{\{4\}}+0*2^{\{3\}}+1*2^{\{2\}}+1*2^{\{1\}}+0*2^{\{0\}}=$$

$$1*2^{\{4\}}*2+1*2^{\{3\}}*2+0*2^{\{2\}}*2+1*2^{\{1\}}*2+1*2^{\{0\}}*2 =$$

$$(1*2^{\{4\}}+1*2^{\{3\}}+0*2^{\{2\}}+1*2^{\{1\}}+1*2^{\{0\}})*2$$

CURIOSIDADE

Observe o que acontece nos bits quando os deslocamos para a esquerda

11011 << 1 =

$(1*2^{4}+1*2^{3}+0*2^{2}+1*2^{1}+1*2^{0}) << 1 =$

$1*2^{5}+1*2^{4}+0*2^{3}+1*2^{2}+1*2^{1}+0*2^{0}=$

$1*2^{4}*2+1*2^{3}*2+0*2^{2}*2+1*2^{1}*2+1*2^{0}*2 =$

$(1*2^{4}+1*2^{3}+0*2^{2}+1*2^{1}+1*2^{0})*2$

CURIOSIDADE

Assim, o shift de bits a esquerda um número n =

$$n<<1 = n*2$$

$$n<<2 = ((n<<1)<<1) = (n*2)<<1 = (n*2)*2 = n*2^{2}$$

$$n<<x = n*2^{x}$$

Logo

$$27<<1 = 27*2 = 54$$

$$27<<2 = 27*4 = 108$$

CURIOSIDADE

Algo parecido ocorre com o shift de bits a direita

$$n>>1 = n/2$$

$$n>>2 = (n>>1)>>1 = (n/2)>>1 = (n/2)/2 = n/4$$

$$n>>x = n/2^{\{x\}}$$

ALGORITMO

```cpp
#include <bits/stdc++.h>

using namespace std;

int main() {
  // leio um número inteiro decimal
  int n;
  cin >> n;
  // acumulo os bits em bin
  string bin = "";
  while (n) {
    bin.push_back(n % 2 == 0 ? '0' : '1');
    n >>= 1; // mesma coisa de n /= 2
  }
  // inverto a ordem
  reverse(bin.begin(), bin.end());
  cout << bin << "\n";
  return 0;
}
```

BINÁRIO - DECIMAL

Podemos usar a própria definição de base multiplicando o dígito por sua potência correspondente e somando

$11011 = 1*2^{4}+1*2^{3}+0*2^{2}+1*2^{1}+1*2^{0}$

$11011 = 1*16+1*8+0*4+1*2+1*1$

$11011 = 16+8+2+1$

$11011 = 27$

ALGORITMO

```cpp
#include <bits/stdc++.h>

using namespace std;

int main() {
  // leio um número em binário como string,
  // para simplificar
  string bin;
  cin >> bin;
  // acumula pot2 -> 1, 2, 4, 8, ...
  int pot2 = 1, n = 0;
  for (int i = bin.size() - 1; i >= 0; i--, pot2 <<= 1)
    n += (bin[i] == '1' ? 1 : 0) * pot2;
  cout << n << "\n";
  return 0;
}
```

BINÁRIO - DECIMAL

Podemos usar a própria definição de base colocando 2 em evidência

$$11011 = 1*2^{4}+1*2^{3}+0*2^{2}+1*2^{1}+1*2^{0}$$

$$11011 = (1*2^{3}+1*2^{2}+0*2^{1}+1)*2+1$$

$$11011 = ((1*2^{2}+1*2^{1}+0)*2+1)*2+1$$

$$11011 = (((1*2^{1}+1)*2+0)*2+1)*2+1$$

$$11011 = ((((1*2)+1)*2+0)*2+1)*2+1$$

ALGORITMO

```cpp
#include <bits/stdc++.h>

using namespace std;

int main() {
  // leio um número em binário como string,
  // para simplificar
  string bin;
  cin >> bin;
  // parênteses mais interno -> mais externo
  int n = 0;
  for (int i = 0; i < bin.size() - 1; i++)
    n = (n + (bin[i] == '1' ? 1 : 0)) * 2;
  n += (bin[bin.size() - 1] == '1' ? 1 : 0);
  cout << n << "\n";
  return 0;
}
```

PONTO FLUTUANTE

Todo número "n" no sistema de ponto flutuante é representado como

$$x = m*B^e$$

onde

- mantissa = m = 0.d1d2...dt, onde d1 != 0 e 0 <= di <= B-1

PONTO FLUTUANTE

Dado o sistema (B,t,l,u)

$$x = m*B^e$$

- B é a base de representação
- t é a quantidade de dígitos da mantisa "m"
- o expoente "e" está entre [l, u]

PONTO FLUTUANTE

Dado o sistema (10,4,-5,5)

$x = m*B^e$

Qual o menor número m e o maior número M que pode ser representado neste sistema?

$m = 0.1000*10^{-5}$

$M = 0.9999*10^{5} = 99990$

PONTO FLUTUANTE

Dado o sistema (10,4,-5,5)

O número 100.000 pode ser representado por esse conjunto?

Veja que M = 99990 é o maior número que pode ser representado e M < 100.000. Logo não pode ser representado

PONTO FLUTUANTE

Dado o sistema (10,4,-5,5)

Represente o número 357,26 usando o arredondamento.

$$x = 357,26 = 0.35726 * 10^3$$

como t=4, só podemos ter 4 dígitos na mantissa, arrendondando...

$$x = 0.3573 * 10^3 = 357,3$$

PONTO FLUTUANTE

Dado o sistema (10,4,-5,5)
Represente o número 357,26 usando o trucamento.

$$x = 357{,}26 = 0.35726 * 10^3$$

como t=4, só podemos ter 4 dígitos na mantissa, truncando...

$$x = 0.3572 * 10^3 = 357{,}2$$

PONTO FLUTUANTE

Supondo que o número 357,26 usando o arredondamento seja igual ao valor exato desse
número e que o número 357,26 usando o truncamento seja igual ao valor aproximado, calcule o erro relativo e absoluto desse número.

arredondando = 357,3
truncando = 357,2

absoluto = aproximada-real
absoluto = 357,2-357,3 = -0,1

relativo = |aproximada-real|/|real|
relativo = |357,2-357,3|/|357,3|
relativo = 0.000279877

ERRO ABSOLUTO

Seja x o valor exato e $\tilde{x}$ o seu valor aproximado

Erro absoluto → $EAx = x - \tilde{x}$

Normalmente x não é conhecido, então usamos um limite superior "e" para o modo do erro absoluto

→ $|EAx| = |x - \tilde{x}| < e$

→ $-e < x - \tilde{x} < +e$

→ $\tilde{x} - e < x < \tilde{x} + e$

ERRO RELATIVO

Seja x o valor exato e x˜ o seu valor aproximado

Erro absoluto não depende da grandeza

Imagine uma ponte que deveria ter 2500cm porém foi construida com 2499cm, ou seja, EAx = 1

Imagine outra ponte que deveria ter 2cm porém foi construida com 1cm, ou seja, EAx = 1

O erro absoluto de ambas são iguais, mas qual o mais impactate?

ERRO RELATIVO

Seja x o valor exato e $\tilde{x}$ o seu valor aproximado

Erro relativo é uma proporção, o quanto o valor apróximado desvia do valor exato.

$$|ERx| = |EAx/\tilde{x}| = |x-\tilde{x}|/|x|$$

Erro Relativo na Ponte 1 → 1/2499 = 0.00040016

Erro Relativo na Ponte 2 → 1/2 = 0.5

PONTO FLUTUANTE

Podemos representar um número

$x = fx*B^{e} + gx*B^{e-t}$, onde $0.1 <= fx < 1$ e $0 <= gx < 1$

Particularmente, 357.26 pode ser representado por $x = 0.3572*10^{3} + 0.6*10^{-1}$, onde $fx=0.3572$ e $gx=0.6$

A parcela gx não pode ser incorporado a mantissa de x no SPF (10, 4, -5, 5), pois só temos "t" digitos.

PONTO FLUTUANTE

$x = 0.3572*10^3 + 0.6*10^{-1},$ onde $fx=0.3572$ e $gx=0.6$

A parcela gx não pode ser incorporado a mantissa de x no SPF (10, 4, -5, 5), pois só temos "t" digitos. Logo, a representação de x no nosso SPF

$$x^{\sim}=fx*10^e=0.3572*10^3$$

,ou seja, podemos descartar a parcela $gx*10^{e-t}$

ERRO DE TRUNCAMENTO

$x = 0.3572*10^3 + 0.6*10^{-1},$ onde $fx=0.3572$ e $gx=0.6$

Vimos que $x^{\sim} = fx*10^e = 0.3572*10^3$

Assim,

$|EAx| = |x-x^{\sim}| = |gx|*B^{e-t} = 0.6*10^{-1}$

$|ERx| = |EAx|/|x^{\sim}| = |gx|*B^{e-t}/|fx|*B^e =$

$= 0.6*10^{-1}/0.3572*10^3 = 0.1679731243*10^{-5}$

ERRO DE ARREDONDAMENTO

No sistema de ponto flutuante (10, 4, -5, 5), o número $x = 357{,}26 = 0.35726*10^3 = 0.3572*10^3 + 0.6*10^{-1}$ pode ser arredondado para $\tilde{x}=0.3573*10^3 = 357{,}3$, pois $6 >= 5$ e, consequentemente, $gx >= 0.5$

Já quando $x = 357{,}24 = 0.35724*10^3 = 0.3572*10^3 + 0.4*10^{-1}$ pode ser arredondado para $\tilde{x}=0.3572*10^3 = 357{,}2$, pois $4 < 5$ e, consequentemente, $gx < 0.5$

ERRO DE ARREDONDAMENTO

Assim,

$x^{\sim} = fx*B^{e} + B^{e-t}$, se $|gx| >= 0.5$

$x^{\sim} = fx*B^{e}$, se $|gx| < 0.5$

ERRO DE ARREDONDAMENTO

Se $|gx| >= 0.5$, então $x^{\sim} = fx*B^{e} + B^{e-t}$.

Assim, no caso específico do SPF (10, 4, -5, 5) e

$x = 357,26 = 0.35726*10^{3} = 0.3572*10^{3} + 0.6*10^{-1}$

$|EAx| = |x-x^{\sim}| = |fx*B^{e} + gxB^{e-t} - fx*B^{e} - B^{e-t}| =$

$= |(gx-1)|*B^{e-t} = |(0.6-1)|*10^{-1} = 0.4*10^{-1} = 0.04$

$|ERx| = |EAx|/|x^{\sim}| = |(gx-1)|*B^{e-t}/fx*B^{e}+B^{e-t} =$

$= |(0.6-1)|*10^{-1}/0.3572*10^{3}+10^{-1} =$

$= 0.4*10^{-1}/357,3 = 0.000111951$

ERROS NAS PARCELAS

Seja x e y, tais que:

- $x = \tilde{x} + EAx$
- $y = \tilde{y} + EAy$

Vamos calcular o erro absoluto e relativo de

1. $x+y$
2. $x-y$
3. $x*y$
4. x/y

ERROS EM X+Y

Seja x = x˜ + EAx e y = y˜ + EAy, tal que:

- Então, o erro absoluto na soma, EA_{x+y} é a soma dos erros absolutos das parcelas:

$$EA_{x+y} = EA_x + EA_y$$

- E o erro relativo será:

$$\begin{aligned} ER_{x+y} &= \frac{EA_{x+y}}{\bar{x}+\bar{y}} = \frac{EA_x}{\bar{x}}\left(\frac{\bar{x}}{\bar{x}+\bar{y}}\right) + \frac{EA_y}{\bar{y}}\left(\frac{\bar{y}}{\bar{x}+\bar{y}}\right) \\ &= ER_x\left(\frac{\bar{x}}{\bar{x}+\bar{y}}\right) + ER_y\left(\frac{\bar{y}}{\bar{x}+\bar{y}}\right) \end{aligned}$$

ERROS EM X-Y

Seja x = x˜ + EAx e y = y˜ + EAy, tal que:

- Então, o erro absoluto na diferença, $EA_{x\text{-}y}$ é a diferença dos erros absolutos das parcelas:

$$EA_{x-y} = EA_x - EA_y$$

- E o erro relativo será:

$$\begin{aligned} ER_{x-y} &= \frac{EA_{x-y}}{\bar{x}-\bar{y}} = \frac{EA_x}{\bar{x}}\left(\frac{\bar{x}}{\bar{x}-\bar{y}}\right) - \frac{EA_y}{\bar{y}}\left(\frac{\bar{y}}{\bar{x}-\bar{y}}\right) \\ &= ER_x\left(\frac{\bar{x}}{\bar{x}-\bar{y}}\right) - ER_y\left(\frac{\bar{y}}{\bar{x}-\bar{y}}\right) \end{aligned}$$

ERROS EM X*Y

Seja x = x˜ + EAx e y = y˜ + EAy, tal que:

$$EA_{x\,y} \approx \bar{x}\, EA_y + \bar{y}\, EA_x$$

$$ER_{x\,y} \approx \frac{\bar{x}\, EA_y + \bar{y}\, EA_x}{\bar{x}\,\bar{y}} = \frac{EA_x}{\bar{x}} + \frac{EA_y}{\bar{y}} = ER_x + ER$$

ERROS EM X/Y

Seja x = x˜ + EAx e y = y˜ + EAy, tal que:

$$EA_{\frac{x}{y}} \approx \frac{EA_x}{\bar{y}} - \frac{\bar{x}\,EA_y}{\bar{y}^2} = \frac{\bar{y}\,EA_x - \bar{x}\,EA_y}{\bar{y}^2}$$

$$ER_{\frac{x}{y}} \approx \left(\frac{\bar{y}\,EA_x - \bar{x}\,EA_y}{\bar{y}^2}\right)\frac{\bar{y}}{\bar{x}} = \frac{EA_x}{\bar{x}} - \frac{EA_y}{\bar{y}} = ER_x - ER_y$$

ERRO NO RESULTADO

Vimos que temos erros nas parcelas entre operações, mas também temos erros nos resultados RA

Exemplo: Vamos somar x+y em SPF(10, 4, -5, 5)

$x = 0.937*10^4$ e $y = 0.1272*10^2$

$x = 0.937*10^4$ e $y = 0.001272*10^4$

$x+y = 0.937*10^4 + 0.001272*10^4$

$x+y = 0.938272*10^4$ (Resultado Exato)

Truncando x+y: $0.9382*10^4$

Arredondando x+y: $0.9383*10^4$

ERRO NO RESULTADO

Temos erros nos resultados RA, além disso

Erro Relativo no Truncamento:

$$|RA| < 10^{-t+1}$$

Erro Relativo no Arredondamento:

$$|RA| < 0.5*10^{-t+1}$$

ITEM A

Diga quanto vale o erro relativo para $u = (m+n)w/o$ supondo que o erro relativo dos números vale $0.5*10^{-t+1}$ e usando o arredondamento.

Seja $s = m+n$

$$ERs = ERm(m/(m+n)) + ERn(n/(m+n)) + RAs$$
$$= 0.5*10^{-t+1}(m/(m+n)) + 0.5*10^{-t+1}(n/(m+n)) + RAs$$
$$= 0.5*10^{-t+1}\ (m/(m+n) + n/(m+n)) + RAs$$
$$= 0.5*10^{-t+1} + RAs$$

Assim

$$|ERs| = 0.5*10^{-t+1} + |RAs| < 0.5*10^{-t+1} + 0.5*10^{-t+1} = 10^{-t+1}$$

ITEM A

Diga quanto vale o erro relativo para u = (m+n)w/o supondo que o erro relativo dos números vale $0.5*10^{-t+1}$ e usando o arredondamento.

Seja $t = (m+n)*w = s*w$

$$
\begin{aligned}
ERt &= ERs + ERw + RAt \\
&= ERs + 0.5*10^{-t+1} + RAt \\
&= 0.5*10^{-t+1} + RAs + 0.5*10^{-t+1} + RAt \\
&= RAs + RAt + 10^{-t+1}
\end{aligned}
$$

Assim

$$
\begin{aligned}
|ERt| &= |RAs| + |RAt| + 10^{-t+1} \\
&< 0.5*10^{-t+1} + 0.5*10^{-t+1} + 10^{-t+1} = 2*10^{-t+1}
\end{aligned}
$$

ITEM A

Diga quanto vale o erro relativo para u = (m+n)w/o supondo que o erro relativo dos números vale $0.5*10^{-t+1}$ e usando o arredondamento.

Seja v = (m+n)*w/o = s*w/o = t/o

$$
\begin{aligned}
ERo &= ERt - ERo + RAo \\
&= ERt - 0.5*10^{-t+1} + RAo \\
&= RAs + RAt + 10^{-t+1} - 0.5*10^{-t+1} + RAo \\
&= RAs + RAt + RAo + 0.5*10^{-t+1}
\end{aligned}
$$

Assim

$$
\begin{aligned}
|ERo| &= |RAs| + |RAt| + |RAo| + 0.5*10^{-t+1} \\
&< 0.5*10^{-t+1} + 0.5*10^{-t+1} + 0.5*10^{-t+1} + 0.5*10^{-t+1} \\
&= 2*10^{-t+1}
\end{aligned}
$$

ITEM B

Diga quanto vale o erro relativo para o mesmo u, mas agora supondo que os números são representados exatamente(ER = 0) e usando o truncamento.

Seja $s = m+n$

$$
\begin{aligned}
ERs &= ERm(m/(m+n)) + ERn(n/(m+n)) + RAs \\
&= 0*(m/(m+n)) + 0*(n/(m+n)) + RAs \\
&= RAs
\end{aligned}
$$

Assim

$|ERs| = |RAs| < 10^{-t+1}$

ITEM B

Diga quanto vale o erro relativo para o mesmo u, mas agora supondo que os números são representados exatamente(ER = 0) e usando o truncamento.

Seja $t = (m+n)*w = s*w$

$ERt = ERs + ERw + RAt$

$= ERs + 0 + RAt$

$= RAs + RAt$

Assim

$|ERt| = |RAs| + |RAt|$

$< 10^{-t+1} + 10^{-t+1} = 2*10^{-t+1}$

ITEM B

Diga quanto vale o erro relativo para o mesmo u, mas agora supondo que os números são representados exatamente(ER = 0) e usando o truncamento.

Seja v = (m+n)*w/o = s*w/o = t/o

$$
\begin{aligned}
ERo &= ERt - ERo + RAo \\
&= ERt - 0 + RAo \\
&= RAs + RAt + RAo
\end{aligned}
$$

Assim

$$
\begin{aligned}
|ERo| &= |RAs| + |RAt| + |RAo| \\
&< 10^{-t+1} + 10^{-t+1} + 10^{-t+1} \\
&= 3*10^{-t+1}
\end{aligned}
$$

ITEM C

Se os valores aproximados para m, n, o e w valem, respectivamente, 10, 20, 30 e 40,
calcule quanto valem os erros dos dois itens anteriores.

Questão mal formulada, qual o SPF (B, t, l, u)?

(Ou não entendi como fazer)