Cálculo Numérico

Aritmética de Ponto Flutuante

Wellington José Corrêa

Universidade Tecnológica Federal do Paraná

17 de Junho de 2021

@correa.well

Aritmética de Ponto Flutuante

Antes de conhecermos as causas dos erros, se faz necessário conhecer como os números são armazenados em um computador. Um número pode ser representado como **ponto fixo**, por exemplo, $12,34$ ou como ponto flutuante $0,1234 \times 10^2$.

UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

A forma geral de representação de um número em ponto flutuante é similar a notação científica $\pm d_1 d_2 d_3 \ldots d_p \times B^e$, onde d_i's são os dígitos da parte fracionária, tais que $0 < d_1 \leq B - 1$, $d_1 \neq 0$, B é o valor da base (geralmente 2,10 ou 16), p é o número de dígitos e e é um expoente inteiro.

Deste modo, um número de ponto flutuante possui três partes: sinal, mantissa e expoente:

$$\underbrace{\pm}_{\text{sinal}} \underbrace{d_1 d_2 d_3 \ldots d_p}_{\text{mantissa}} \times B^{\overbrace{e}^{\text{expoente}}}, \; d_1 \neq 0. \tag{1}$$

Aritmética de Ponto Flutuante

Para fixar ideias, consideremos que os números de uma máquina sejam representados na representação de ponto flutuante decimal:

$$\pm 0{,}d_1 d_2 d_3 \ldots d_p \times 10^e;\ 1 \leq d_1 \leq 9;\ 0 \leq d_i \leq 9,\ i = 2, 3, ..., p.$$

Aritmética de Ponto Flutuante

Para fixar ideias, consideremos que os números de uma máquina sejam representados na representação de ponto flutuante decimal:

$$\pm 0, d_1 d_2 d_3 \ldots d_p \times 10^e;\ 1 \leq d_1 \leq 9;\ 0 \leq d_i \leq 9,\ i = 2, 3, ..., p.$$

Qualquer número real positivo dentro do intervalo numérico da máquina é da forma:

$$y = 0, d_1\, d_2\, d_3\, \ldots\, d_p\, d_{p+1}\, d_{p+2} \ldots \times 10^e.$$

UTFPR
UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

Aritmética de Ponto Flutuante

Para fixar ideias, consideremos que os números de uma máquina sejam representados na representação de ponto flutuante decimal:

$$\pm 0{,}d_1 d_2 d_3 \ldots d_p \times 10^e;\ 1 \leq d_1 \leq 9;\ 0 \leq d_i \leq 9,\ i = 2, 3, ..., p.$$

Qualquer número real positivo dentro do intervalo numérico da máquina é da forma:

$$y = 0, d_1\, d_2\, d_3\ \ldots\ d_p\, d_{p+1}\, d_{p+2} \ldots \times 10^e.$$

A forma em ponto flutuante de y denotado por $fl(y)$ é obtida terminando a mantissa de y em p algarismos (dígitos) decimais. Há duas maneiras de fazer isto:

UTFPR UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

Aritmética de Ponto Flutuante

1. **Truncamento:** É simplesmente "cortar" ou "truncar" os algarismos $d_{p+1}, d_{p+2}, \ldots$ Isso produz a forma em ponto flutuante como:

$$fl(y) = 0, d_1 d_2 d_3 \ldots d_p \times 10^e.$$

UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

Aritmética de Ponto Flutuante

1. Truncamento: É simplesmente "cortar" ou "truncar" os algarismos $d_{p+1}, d_{p+2}, \ldots$ Isso produz a forma em ponto flutuante como:
$$fl(y) = 0, d_1 d_2 d_3 \ldots d_p \times 10^e.$$
2. Arredondamento: Adicione $5 \times 10^{e-(p+1)}$ a y e então, trunca o resultado para obter um número da forma:
$$fl(y) = 0, d_1 d_2 \ldots d_i \times 10^e.$$

UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

Aritmética de Ponto Flutuante

1. Truncamento: É simplesmente "cortar" ou "truncar" os algarismos $d_{p+1}, d_{p+2}, \ldots$ Isso produz a forma em ponto flutuante como:
$$fl(y) = 0, d_1 d_2 d_3 \ldots d_p \times 10^e.$$
2. Arredondamento: Adicione $5 \times 10^{e-(p+1)}$ a y e então, trunca o resultado para obter um número da forma:
$$fl(y) = 0, d_1 d_2 \ldots d_i \times 10^e.$$
Desse modo, ao arrendondar, se $d_{p+1} \geq 5$, adicionamos 1 a d_p para obter $fl(y)$, isto é, arredondamos para cima. Quando $d_{p+1} < 5$, simplesmente cortamos (truncamos) tudo, menos os primeiros p algarismos, logo, arredondamos para baixo.

Aritmética de Ponto Flutuante

Exemplo

O número π *tem uma expansão decimal infinita da forma* $\pi = 3,14159265...$. *Obtenha a forma de ponto flutuante de* π *usando o truncamento e arredondamento para cinco algarismos.*

Aritmética de Ponto Flutuante

Solução: Note inicialmente que

$$\pi = 0,314159265... \times 10^1; \; p = 5, \; e = 1.$$

Aritmética de Ponto Flutuante

Solução: Note inicialmente que

$$\pi = 0,314159265\ldots \times 10^1; \; p = 5, \; e = 1.$$

- Truncamento: Temos que

$$fl(\pi) = 0,31415 \times 10^1 = 3,1415.$$

Aritmética de Ponto Flutuante

Solução: Note inicialmente que

$$\pi = 0,314159265... \times 10^1;\ p = 5,\ e = 1.$$

- Truncamento: Temos que
$$fl(\pi) = 0,31415 \times 10^1 = 3,1415.$$
- Arredondamento:

Aritmética de Ponto Flutuante

Solução: Note inicialmente que

$$\pi = 0,314159265\ldots \times 10^1; \ p = 5, \ e = 1.$$

- Truncamento: Temos que
$$fl(\pi) = 0,31415 \times 10^1 = 3,1415.$$
- Arredondamento:

Devemos adicionar
$$5 \times 10^{e-(p+1)} = 5 \times 10^{-5} = 0,0000005 \times 10^1$$
a π, ou seja,

Aritmética de Ponto Flutuante

$$\begin{aligned}\pi + 5 \times 10^{e-(p+1)} &= 0,314159265\ldots \times 10^1 + 0,0000005 \times 10^1 \\ &= 0,314164265\ldots \times 10^1 .\end{aligned}$$

Realizando o truncamento para $p = 5$, resulta que

$$\begin{aligned}fl(\pi) &= 0,31416 \times 10^1 \\ &= 3,1416.\end{aligned}$$

UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

Aritmética de Ponto Flutuante

Todos os números em ponto flutuante, juntamente com a representação do zero, constitui o sistema de ponto flutuante normalizado ($d_1 \neq 0$), que indicamos por

Aritmética de Ponto Flutuante

Todos os números em ponto flutuante, juntamente com a representação do zero, constitui o sistema de ponto flutuante normalizado ($d_1 \neq 0$), que indicamos por

$$\mathrm{SPF}(B, p, \mathrm{e}_{\mathrm{min}}, \mathrm{e}_{\mathrm{max}})$$

onde $e \in [\mathrm{e}_{\mathrm{min}}, \mathrm{e}_{\mathrm{max}}]$, B e p são caracterizados em (1).

UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

Aritmética de Ponto Flutuante

Exemplo

Considere um computador hipotético com dois dígitos $p = 2$, *base* $B = 2$ *e expoente na faixa* $-1 \leq e \leq 2$, *isto é, no sistema de ponto flutuante* $SPF(2, 2, -1, 2)$. *Encontre todos os números positivos representáveis neste computador.*

UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

Aritmética de Ponto Flutuante

Exemplo

Considere um computador hipotético com dois dígitos $p = 2$*, base* $B = 2$ *e expoente na faixa* $-1 \leq e \leq 2$*, isto é, no sistema de ponto flutuante* $SPF(2, 2, -1, 2)$*. Encontre todos os números positivos representáveis neste computador.*

Solução: Com efeito, os números deste sistema são da forma

$$\pm (0, 10)_2 \times 2^e \ \text{ ou } \ \pm (0, 11)_2 \times 2^e, e = -1, \ldots, 2.$$

Aritmética de Ponto Flutuante

Exemplo

Considere um computador hipotético com dois dígitos $p = 2$, *base* $B = 2$ *e expoente na faixa* $-1 \leq e \leq 2$, *isto é, no sistema de ponto flutuante* $SPF(2, 2, -1, 2)$. *Encontre todos os números positivos representáveis neste computador.*

Solução: Com efeito, os números deste sistema são da forma

$$\pm\,(0,10)_2 \times 2^e \ \text{ ou } \ \pm\,(0,11)_2 \times 2^e, e = -1, \ldots, 2.$$

Em nossa representação, o primeiro dígito não é nulo.

Aritmética de Ponto Flutuante

Note que,

$$(0,10)_2 = 1 \times 2^{-1} + 0 \times 2^{-2} = \frac{1}{2} \ \text{ e } \ (0,11)_2 = 1 \times 2^{-1} + 1 \times 2^{-2} = \frac{3}{4},$$

Aritmética de Ponto Flutuante

então, os únicos números positivos representáveis neste computador são:

Aritmética de Ponto Flutuante

então, os únicos números positivos representáveis neste computador são:

$$(0,10)_2 \times 2^{-1} = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$$

UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

Aritmética de Ponto Flutuante

então, os únicos números positivos representáveis neste computador são:

$$(0,10)_2 \times 2^{-1} = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$$
$$(0,10)_2 \times 2^{0} = \frac{1}{2} \times 1 = \frac{1}{2}$$

UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

Aritmética de Ponto Flutuante

então, os únicos números positivos representáveis neste computador são:

$$(0,10)_2 \times 2^{-1} = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$$
$$(0,10)_2 \times 2^{0} = \frac{1}{2} \times 1 = \frac{1}{2}$$
$$(0,10)_2 \times 2^{1} = \frac{1}{2} \times 2 = 1$$

UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

Aritmética de Ponto Flutuante

então, os únicos números positivos representáveis neste computador são:

$$
\begin{aligned}
(0,10)_2 \times 2^{-1} &= \frac{1}{2} \times \frac{1}{2} = \frac{1}{4} \\
(0,10)_2 \times 2^{0} &= \frac{1}{2} \times 1 = \frac{1}{2} \\
(0,10)_2 \times 2^{1} &= \frac{1}{2} \times 2 = 1 \\
(0,10)_2 \times 2^{2} &= \frac{1}{2} \times 4 = 2
\end{aligned}
$$

UTFPR
UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

Aritmética de Ponto Flutuante

então, os únicos números positivos representáveis neste computador são:

$$(0,10)_2 \times 2^{-1} = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4} \qquad (0,11)_2 \times 2^{-1} = \frac{3}{4} \times \frac{1}{2} = \frac{3}{8}$$

$$(0,10)_2 \times 2^{0} = \frac{1}{2} \times 1 = \frac{1}{2}$$

$$(0,10)_2 \times 2^{1} = \frac{1}{2} \times 2 = 1$$

$$(0,10)_2 \times 2^{2} = \frac{1}{2} \times 4 = 2$$

UTFPR
UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

Aritmética de Ponto Flutuante

então, os únicos números positivos representáveis neste computador são:

$$(0,10)_2 \times 2^{-1} = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$$
$$(0,10)_2 \times 2^{0} = \frac{1}{2} \times 1 = \frac{1}{2}$$
$$(0,10)_2 \times 2^{1} = \frac{1}{2} \times 2 = 1$$
$$(0,10)_2 \times 2^{2} = \frac{1}{2} \times 4 = 2$$

$$(0,11)_2 \times 2^{-1} = \frac{3}{4} \times \frac{1}{2} = \frac{3}{8}$$
$$(0,11)_2 \times 2^{0} = \frac{3}{4} \times 1 = \frac{3}{4}$$

UTFPR
UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

Aritmética de Ponto Flutuante

então, os únicos números positivos representáveis neste computador são:

$$(0,10)_2 \times 2^{-1} = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4} \qquad (0,11)_2 \times 2^{-1} = \frac{3}{4} \times \frac{1}{2} = \frac{3}{8}$$

$$(0,10)_2 \times 2^{0} = \frac{1}{2} \times 1 = \frac{1}{2} \qquad (0,11)_2 \times 2^{0} = \frac{3}{4} \times 1 = \frac{3}{4}$$

$$(0,10)_2 \times 2^{1} = \frac{1}{2} \times 2 = 1 \qquad (0,11)_2 \times 2^{1} = \frac{3}{4} \times 2 = \frac{3}{2}$$

$$(0,10)_2 \times 2^{2} = \frac{1}{2} \times 4 = 2$$

UTFPR
UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

Aritmética de Ponto Flutuante

então, os únicos números positivos representáveis neste computador são:

$$
\begin{aligned}
(0,10)_2 \times 2^{-1} &= \frac{1}{2} \times \frac{1}{2} = \frac{1}{4} & \qquad (0,11)_2 \times 2^{-1} &= \frac{3}{4} \times \frac{1}{2} = \frac{3}{8} \\
(0,10)_2 \times 2^{0} &= \frac{1}{2} \times 1 = \frac{1}{2} & (0,11)_2 \times 2^{0} &= \frac{3}{4} \times 1 = \frac{3}{4} \\
(0,10)_2 \times 2^{1} &= \frac{1}{2} \times 2 = 1 & (0,11)_2 \times 2^{1} &= \frac{3}{4} \times 2 = \frac{3}{2} \\
(0,10)_2 \times 2^{2} &= \frac{1}{2} \times 4 = 2 & (0,11)_2 \times 2^{2} &= \frac{3}{4} \times 4 = 3
\end{aligned}
$$

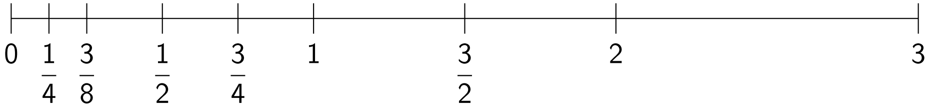
Aritmética de Ponto Flutuante

O zero é representado de forma especial: todos os dígitos d_i da mantissa e do expoente são nulos. Assim, representamos tais números na seguinte reta:

UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

Aritmética de Ponto Flutuante

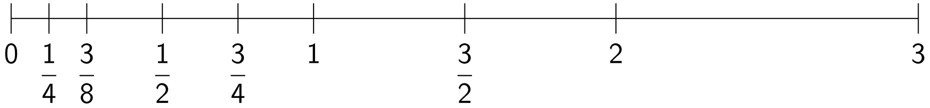
O zero é representado de forma especial: todos os dígitos d_i da mantissa e do expoente são nulos. Assim, representamos tais números na seguinte reta:



UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

Aritmética de Ponto Flutuante

O zero é representado de forma especial: todos os dígitos d_i da mantissa e do expoente são nulos. Assim, representamos tais números na seguinte reta:



Observação

O conceito de sempre existir um número real entre dois números reais não é válido para os números de ponto flutuante. A falha deste conceito tem uma consequência desastrosa. Considere o exemplo a seguir.

Aritmética de Ponto Flutuante

Exemplo

Considere a representação binária:

$$(0,6)_{10} = (0,100110011001...)_2 \ \ e \ \ (0,7)_{10} = (0,1011001100110...)_2 \, .$$

Se estes dois números forem armazenados naquele computador hipotético do exemplo anterior, eles serão representados por $(0,10)_2 \times 10^0$.

Aritmética de Ponto Flutuante

Exemplo

Considere a representação binária:

$$(0,6)_{10} = (0,100110011001...)_2 \;\; e \;\; (0,7)_{10} = (0,1011001100110...)_2 \, .$$

Se estes dois números forem armazenados naquele computador hipotético do exemplo anterior, eles serão representados por $(0,10)_2 \times 10^0$.
Isto nos diz que tanto $(0,6)_{10}$ *quanto* $(0,7)_{10}$ *serão vistos como* $(0,5)_{10}$ *por aquele computador. Esta é uma grande causa de erro de arredondamento nos processos numéricos.*

UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

Aritmética de Ponto Flutuante

Exemplo

Considere a representação binária:

$$(0,6)_{10} = (0,100110011001...)_2 \ \ e \ \ (0,7)_{10} = (0,1011001100110...)_2 \, .$$

Se estes dois números forem armazenados naquele computador hipotético do exemplo anterior, eles serão representados por $(0,10)_2 \times 10^0$.
Isto nos diz que tanto $(0,6)_{10}$ *quanto* $(0,7)_{10}$ *serão vistos como* $(0,5)_{10}$ *por aquele computador. Esta é uma grande causa de erro de arredondamento nos processos numéricos.*
Veja que esta causa provém das mantissas da representação binária dos números $(0,6)_{10}$ *na* $(0,7)_{10}$.

UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

Aritmética de Ponto Flutuante

Além do fato em que a mantissa representa um número finito de números ilustrado pelo exemplo anterior, uma outra situação em que acarreta erros provém que o expoente *e* é limitado pelos valores e_{min} e e_{max}.

UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

Aritmética de Ponto Flutuante

Além do fato em que a mantissa representa um número finito de números ilustrado pelo exemplo anterior, uma outra situação em que acarreta erros provém que o expoente e é limitado pelos valores e_{min} e e_{max}.
Sempre que uma operação aritmética produz um número com expoente superior ao expoente máximo, tem-se o fenômeno de *overflow*. De forma similar, operações que resultem em expoente inferior ao expoente mínimo tem-se o fenômeno de *underflow*.

UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

Aritmética de Ponto Flutuante

Para fixar ideias, consideremos o exemplo anterior, só que agora com todos os números reais (considerando a parte positiva e negativa da aritmética utilizada). No caso do exemplo dado, pode-se observar qual as regiões que ocorrem o *overflow* e o *underflow*.

UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

Aritmética de Ponto Flutuante

Para fixar ideias, consideremos o exemplo anterior, só que agora com todos os números reais (considerando a parte positiva e negativa da aritmética utilizada). No caso do exemplo dado, pode-se observar qual as regiões que ocorrem o *overflow* e o *underflow*.