

Matemática Computacional :: Aritmética Computacional

maria irene falcão

Mestrado em Matemática e Computação

1. Aritmética Computacional :: background

1.1 Sistema de numeração

Um sistema de numeração de vírgula flutuante $F(b, t, m, M)$ é caracterizado por quatro parâmetros:

- b - base;
- m - valor mínimo do expoente;
- t - número de dígitos da mantissa;
- M - valor máximo do expoente.

Constituem o sistema $F(b, t, m, M)$, para além do número zero, todos os números que se puderem exprimir na forma

$$\pm(.d_1d_2\dots d_t)_b \times b^e$$

com $d_1, d_2, \dots, d_t \in \{0, 1, \dots, b-1\}$, $d_1 \neq 0$, e $e \in \mathbb{Z}$ tal que $m \leq e \leq M$; a notação $(.d_1d_2\dots d_t)_b$ designa $d_1 b^{-1} + d_2 b^{-2} + \dots + d_t b^{-t}$.

Estes são os chamados números **normalizados**. Um sistema $F(b, t, m, M)$ pode ainda admitir os chamados números **desnormalizados** ou **subnormais**, que são os números obtidos deixando de impor a condição $d_1 \neq 0$, quando o expoente assume o valor mínimo.

O maior número de $F(b, t, m, M)$, designa-se por **nível de overflow** e é dado por

$$\Omega := (1 - b^{-t})b^M$$

. O menor número positivo normalizado, chamado **nível de underflow**, é dado por

$$\omega := b^{m-1}$$

. O menor número positivo de um sistema que admita números desnormalizados¹ é b^{m-t}

Ao conjunto

$$R_F := [-\Omega, -\omega] \cup \{0\} \cup [\omega, \Omega]$$

chamamos **conjunto dos números representáveis**.

¹Se nada for dito em contrário, quando nos referirmos a um sistema $F(b, t, m, M)$, consideramos apenas os números normalizados.

➡ Arredondamento

Dado um número $x \in R_F$, pretende-se encontrar um número de máquina que o represente. É natural exigir-se que esse número, iremos denotar por $fl(x)$, esteja à menor distância possível de x , havendo uma regra para decidir o que fazer, no caso de empate. Naturalmente que se $x \in F$, então $fl(x) = x$, como seria de desejar. Quando $fl(x)$ é escolhido desta forma², dizemos que é usado **arredondamento para o mais próximo**.

No caso em que existam dois números de máquina à mesma distância do número x , é habitual (sobretudo se a base do sistema for 2 ou 10) usar-se o chamado **arredondamento para par**, em que se escolhe para $fl(x)$ aquele cujo último dígito da mantissa seja par.

No chamado arredondamento usual (que normalmente usamos no dia-a-dia), em caso de empate, as mantissas são arredondadas “para cima”, o que equivale a somar à mantissa $\frac{1}{2}bb^{-(t+1)}$, truncando, em seguida, o resultado para t dígitos.

A **unidade de erro de arredondamento** do sistema é $\mu := \frac{1}{2}b^{1-t}$. Chama-se **epsilon da máquina**, e denota-se por ϵ , a diferença entre o número de $F(b, t, m, M)$ imediatamente superior a 1 e o número 1, isto é, $\epsilon := b^{1-t}$.

➡ Operações de vírgula flutuante

Representaremos as operações de vírgula flutuante pelo símbolo usual rodeado por \bigcirc ; por exemplo \oplus, \otimes . Admitimos que o resultado de uma operação de vírgula flutuante é obtido por arredondamento do resultado da operação exata, isto é, $x \oplus y = fl(x + y)$, $x \otimes y = fl(x \times y)$, etc.

1.2 Norma IEEE 754

Com o objetivo de uniformizar as operações nos sistemas de vírgula flutuante foi publicada, em 1985, a norma IEEE 754.³ Esta norma especifica dois formatos básicos para a representação de números num sistema de vírgula flutuante: o formato **simples** com 32 bits e o formato **duplo** com 64 bits.

A norma IEEE 754 permite representar números normalizados na forma

$$x = (-1)^s(d_0.d_{-1}d_{-2} \cdots d_{-(t-1)})2^e,$$

onde

- $s \in \{0, 1\}$ sinal;
- t número de bits da mantissa;
- $d_0 = 1$ bit implícito;
- e expoente com $e_{\min} \leq e \leq e_{\max}$.

²Existem outras formas de determinar $fl(x)$, como, por exemplo, a chamada truncatura, em que simplesmente se ignoram todos os dígitos da mantissa do número que estejam para além da posição t

³IEEE- Institute for Electrical and Electronics Engineers.

A tabela seguinte contém os valores dos parâmetros do formato simples e duplo.

	formato simples	formato duplo
t	24	53
e_{\min}	-126	-1022
e_{\max}	127	1023

O formato simples corresponde então ao sistema $F(2, 24, -125, 128)$ e o duplo⁴ corresponde a $F(2, 53, -1021, 1024)$. Ambos os sistemas admitem números desnormalizados.

➡ Alocação dos bits no formato simples

0	1 2 3 4 5 6 7 8	9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31
S	Expoente (8 bits)	Mantissa (23 bits)

➡ Alocação dos bits no formato duplo

0	1 2 3 ... 10 11	12 13 14 15 16 17 18 19 20 21 22 23 ... 55 56 57 58 59 60 61 62 63
S	Expoente (11 bits)	Mantissa (52 bits)

➡ Representação do expoente

No formato simples, o expoente está codificado da seguinte forma

expoente $- \varepsilon$	e
0 0 0 0 0 0 0 0 0 $\hookrightarrow 0$	reservado
0 0 0 0 0 0 0 0 1 $\hookrightarrow 1$	-126
0 0 0 0 0 0 0 1 0 $\hookrightarrow 2$	-125
\vdots	\vdots
0 1 1 1 1 1 1 1 1 $\hookrightarrow 127$	0
1 0 0 0 0 0 0 0 0 $\hookrightarrow 128$	1
\vdots	\vdots
1 1 1 1 1 1 0 1 1 $\hookrightarrow 253$	126
1 1 1 1 1 1 1 1 0 $\hookrightarrow 254$	127
1 1 1 1 1 1 1 1 1 $\hookrightarrow 255$	reservado

Os expoentes satisfazem, por isso, a relação

$$e = \varepsilon - e_{\max}.$$

⁴Por defeito, o MATLAB trabalha no sistema de numeração de norma IEEE em formato duplo.

O sistema de numeração IEEE admite ainda os “números” especiais $+\infty$ e $-\infty$ (**Inf** e **–Inf**), para representar, por exemplo, o resultado da divisão de um número por zero, bem como o símbolo especial **NaN** (Not a Number), para representar o resultado de operações não definidas matematicamente, tais como $0/0$, $\infty - \infty$, etc.

$e_1 e_2 \cdots e_8$	valor representado
$(00000000)_2 = (0)_{10}$	$\pm(0.d_1 d_2 d_3 \cdots d_{22} d_{23})_2 \times 2^{-126}$
$(00000001)_2 = (1)_{10}$	$\pm(1.d_1 d_2 d_3 \cdots d_{22} d_{23})_2 \times 2^{-126}$
$(00000010)_2 = (2)_{10}$	$\pm(1.d_1 d_2 d_3 \cdots d_{22} d_{23})_2 \times 2^{-125}$
\vdots	\vdots
$(01111111)_2 = (127)_{10}$	$\pm(1.d_1 d_2 d_3 \cdots d_{22} d_{23})_2 \times 2^0$
$(10000000)_2 = (128)_{10}$	$\pm(1.d_1 d_2 d_3 \cdots d_{22} d_{23})_2 \times 2^1$
\vdots	\vdots
$(11111101)_2 = (253)_{10}$	$\pm(1.d_1 d_2 d_3 \cdots d_{22} d_{23})_2 \times 2^{126}$
$(11111110)_2 = (254)_{10}$	$\pm(1.d_1 d_2 d_3 \cdots d_{22} d_{23})_2 \times 2^{127}$
\pm	$\pm\infty$ se $d_1 = d_2 = \cdots = d_{23} = 0$ NaN, nos outros casos

A norma IEEE 754 especifica também as regras de arredondamento a utilizar. Por defeito, é utilizado o chamado **arredondamento para par**, isto é, se $x \in R_F$, $fl(x)$ é escolhido como o número de máquina mais próximo de x , sendo, em caso de “empate”, escolhido aquele que tem o último *bit* da mantissa igual a zero.

Para além disso, em geral, tem-se

→ se $x > \Omega$, $fl(x) = \mathbf{Inf}$;

$$\rightarrow \text{se } x < -\Omega, fl(x) = -\text{Inf};$$

→ se $2^{m-t} \leq x < \omega$, $fl(x)$ é o número desnormalizado mais próximo de x ;

→ se $x < 2^{m-t}$, $fl(x) = 0$.

Exemplo 1.1 *Formato simples*

[illegible]

$$-(1.111)_2 \times 2^{129-127} = -(7.5)_{10}$$

[illegible]

[illegible]

5

O majorante do erro absoluto depende de e , logo de x . O majorante do erro relativo depende apenas da unidade de erro de arredondamento da máquina usada. Deste resultado, conclui-se de imediato que

$$fl(x) = x(1 + \delta), \text{ com } |\delta| \leq \mu.$$

► Propagação de erros nas operações usuais

Sejam \tilde{x} e \tilde{y} valores aproximados para x e y , respetivamente ($x, y \neq 0$), e sejam $S = x + y$, $P = x \times y$ e $Q = x/y$. Sejam \tilde{S}, \tilde{P} e \tilde{Q} os valores aproximados para S, P e Q obtidos usando os valores \tilde{x} e \tilde{y} em vez de x e y e admitindo que as operações são efetuadas exatamente. Podem estabelecer-se facilmente os seguintes resultados:

$$\begin{aligned} E_{\tilde{S}} &= E_{\tilde{x}} + E_{\tilde{y}} & E_{\tilde{P}} &= E_{\tilde{x}}y + E_{\tilde{y}}x - E_{\tilde{x}}E_{\tilde{y}} & E_{\tilde{Q}} &= \frac{yE_{\tilde{x}} - xE_{\tilde{y}}}{y\tilde{y}} \\ R_{\tilde{S}} &= \frac{x}{x+y}R_{\tilde{x}} + \frac{y}{x+y}R_{\tilde{y}} & R_{\tilde{P}} &= R_{\tilde{x}} + R_{\tilde{y}} - R_{\tilde{x}}R_{\tilde{y}} & R_{\tilde{Q}} &= \frac{R_{\tilde{x}} - R_{\tilde{y}}}{1 - R_{\tilde{y}}} \end{aligned}$$

Supondo $|R_{\tilde{x}}|, |R_{\tilde{y}}| \ll 1$, obtêm-se as seguintes fórmulas simplificadas para o erro relativo do produto e do quociente

$$R_{\tilde{P}} \approx R_{\tilde{x}} + R_{\tilde{y}} \quad R_{\tilde{Q}} \approx R_{\tilde{x}} - R_{\tilde{y}}$$

Nota: A operação mais “perigosa” (isto é, que pode amplificar significativamente o erro relativo dos argumentos) é a adição (podendo ocorrer o chamado **cancelamento subtrativo** quando se somam números muito próximos e com sinais contrários).

1.4 Condicionamento e estabilidade

Um **problema** diz-se **mal condicionado** se for muito sensível a perturbações introduzidas nos seus dados, isto é, se “pequenas” alterações nos dados produzirem “grandes” alterações na sua solução (independentemente do método escolhido para resolver o problema). Se tal não acontecer, o **problema** diz-se **bem condicionado**.

Um **método** numérico diz-se **instável** se, no decurso dos cálculos inerentes à aplicação do método, os erros se amplificarem de forma inaceitável; Se tal não acontecer, o **método** diz-se **estável**.

► Número de condição de uma função

Sendo f uma função continuamente diferenciável na vizinhança de um ponto x e sendo $f(x) \neq 0$, à quantidade

$$\text{cond}(f(x)) := \frac{|xf'(x)|}{|f(x)|}$$

chamamos **número de condição de f em x** .

Supondo $x \neq 0$ e sendo \tilde{x} pertencente à vizinhança de x onde f é diferenciável, tem-se

$$\frac{|f(x) - f(\tilde{x})|}{|f(x)|} \approx \text{cond}(f(x)) \frac{|x - \tilde{x}|}{|x|}.$$

De modo análogo, se f é uma função de duas variáveis suficientemente diferenciável na vizinhança de (x, y) e (\tilde{x}, \tilde{y}) está nessa vizinhança, tem-se

$$\frac{|f(x, y) - f(\tilde{x}, \tilde{y})|}{|f(x, y)|} \approx \frac{|x \frac{\partial f(x, y)}{\partial x}|}{|f(x, y)|} \frac{|x - \tilde{x}|}{|x|} + \frac{|y \frac{\partial f(x, y)}{\partial y}|}{|f(x, y)|} \frac{|y - \tilde{y}|}{|y|}.$$

As quantidades $\frac{|x \frac{\partial f(x, y)}{\partial x}|}{|f(x, y)|}$ e $\frac{|y \frac{\partial f(x, y)}{\partial y}|}{|f(x, y)|}$ são os **números de condição de f em (x, y)** relativamente à variável x e à variável y , respetivamente.

1.5 Funções pré-definidas do MATLAB

Função	Objetivo
abs	Valor absoluto
ceil	Arredondamento para o inteiro mais próximo (na direcção de $+\infty$)
base2dec	Mudança de uma dada base para a base decimal
bin2dec	Mudança da base binária para a base decimal
dec2base	Mudança da base decimal para outra base
dec2bin	Mudança da base decimal para a base binária
eps	epsilon da máquina
fix	Arredondamento para o inteiro mais próximo (na direcção de zero)
floor	Arredondamento para o inteiro mais próximo (na direcção de $-\infty$)
Inf	Representação na aritmética IEEE de $+\infty$
NaN	Representação na aritmética IEEE de "Not-a-Number"
realmax	Nível de overflow
realmin	Nível de underflow
rem	Resto da divisão
round	Arredondamento para o inteiro mais próximo

Referências

Para mais pormenores sobre a norma IEEE 754, veja, por exemplo, [IEE85], [Ove01] ou [Gol91]; o livro clássico de Wilkinson [Wil63], apesar de bastante antigo, continua a ser uma referência importante sobre o tema deste capítulo; outro livro bastante interessante sobre este tópico é o de Higham [Hig02].

Gol91 D. Goldberg. What every computer scientist should know about floating-point arithmetic. *ACM Computing Surveys*, 23(1):5–48, 1991.

Hig02 N. J. Higham. *Accuracy and Stability of Numerical Algorithms*. SIAM, 2ª edição, 2002.

IEE85 *IEEE Standard for Binary Floating-Point Arithmetic, ANSI/IEEE Standard 754-1985*. Institute for Electrical and Electronics Engineers, New York, 1985.

Ove01 M. L. Overton. *Numerical Computing with IEEE Floating Point Arithmetic*. SIAM, New York, 2001.

Wil63 J. H. Wilkinson. *Rounding Errors in Algebraic Processes*. Prentice Hall, Englewood Cliffs, NJ, 1963.

Nos seguintes endereços encontrará exemplos curiosos de casos verídicos de problemas causados por erros de arredondamento:

<http://www5.in.tum.de/huckle/bugse.html> <http://catless.ncl.ac.uk/Risks/>

Exercício 4. Obtenha a representação decimal dos números representados abaixo.

[illegible]

Exercício 5. Justifique que num sistema de vírgula flutuante de base b , a divisão ou multiplicação por uma potência de b , se não conduzir a *overflow* ou *underflow*, é uma operação exata.

Exercício 6. Diga, justificando, se são verdadeiras ou falsas as seguintes afirmações, considerando que está trabalhar num sistema de vírgula flutuante IEEE (com o arredondamento usual):

- $x \leq y \implies fl(x) \leq fl(y);$
- $x < y \implies fl(x) < fl(y);$
- $x \leq y \implies x \leq fl(\frac{x+y}{2}) \leq y \quad (x, y \in F).$
- Mostre, através de um exemplo, que a afirmação contida na alínea c) pode ser falsa se trabalharmos num sistema de vírgula flutuante de base 10.

Exercício 7. Escreva uma *script* destinada a calcular aproximações para o número de Nepper e , usando a expressão

$$e = \lim_{n \rightarrow \infty} \left(1 + \frac{1}{n}\right)^n$$

Esta *script* deverá produzir uma tabela com os valores das aproximações $(1 + \frac{1}{n})^n$ e dos respetivos erros, para

- a) $n = 10^k; k = 1, 2, \dots, 20;$ b) $n = 2^k; k = 45, 46, \dots, 55.$

Comente os resultados obtidos.

Exercício 8. Considere a função $f(x, h) = \frac{\sin(x+h) - \sin(x)}{h}$.

- Use a fórmula trigonométrica $\sin(a) - \sin(b) = 2 \cos\left(\frac{a+b}{2}\right) \sin\left(\frac{a-b}{2}\right)$ para obter a expressão da função $g(x, h)$, matematicamente equivalente a $f(x, h)$.
- Avalie $f(x, h)$ e $g(x, h)$, em $(1.2, h)$, para $h = 1, 10^{-1}, \dots, 10^{-20}$.
- Explique a diferença dos resultados obtidos.
- Sugira uma fórmula para calcular uma aproximação para a derivada da função $\sin(x)$.

Exercício 9. Considere a função $f(x) = 1 - \cos x$.

- Sejam $\tilde{a} = 0.292$ e $\tilde{b} = 0.049$ aproximações para os valores $a = f(\frac{\pi}{4})$ e $b = f(\frac{\pi}{10})$. Indique o número de casas decimais corretas e o número de algarismos significativos de cada uma destas aproximações.
- Mostre que as funções $f(x)$ e $g(x) = 2 \sin^2 \frac{x}{2}$ são iguais.
- Faça as representações gráficas de f e g no intervalo $[0, t]$, considerando sucessivamente $t = 10^{-7}$, $t = 3 \times 10^{-8}$ e $t = 10^{-8}$ (use três figuras diferentes). Justifique o comportamento observado.

Exercício 10. Represente graficamente os polinômios $p(x) = (x - 1)^6$ e $q(x) = x^6 - 6x^5 + 15x^4 - 20x^3 + 15x^2 - 6x + 1$, para $x \in [0.999, 1.001]$. Notando que $q(x)$ é a forma expandida de $p(x)$, comente os resultados obtidos.

2.2 Trabalhos

Trabalho 1. A média de uma amostra de n valores $x_i; i = 1, \dots, n$, é dada por

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i,$$

sendo o desvio padrão amostral dado por

$$s = \left(\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \right)^{1/2}. \quad (2.1)$$

Para maior eficiência, é frequentemente sugerido o uso da seguinte fórmula alternativa para o cálculo do desvio padrão

$$s = \left(\frac{1}{n-1} \sum_{i=1}^n (x_i^2 - n\bar{x}^2) \right)^{1/2}. \quad (2.2)$$

Escreva uma função,

$$[media, desvio1, desvio2] = \text{mediaDesvios}(x),$$

destinada a calcular a média e o desvio padrão de uma amostra, sendo usadas as duas fórmulas (2.1) e (2.2) para o cálculo do desvio padrão.

Teste a sua função para várias amostras $\{x_i\}$. Em particular, tente encontrar uma amostra para a qual as duas fórmulas do cálculo do desvio padrão produzam valores bastante diferentes. Justifique a diferença dos resultados.

Trabalho 2. Considere o desenvolvimento em série da função exponencial

$$e^x = 1 + x + \frac{x^2}{2!} + \dots + \frac{x^n}{n!} + \dots$$

- Utilize este desenvolvimento, com 101 termos, para calcular uma aproximação para o valor de e^{-25} .
- Obtenha uma aproximação para e^{25} usando a série referida e calcule, então, e^{-25} através da fórmula $e^{-25} = \frac{1}{e^{25}}$.
- Compare os resultados obtidos nas alíneas anteriores com o valor de e^{-25} dado usando a função `exp` do MATLAB e explique-os.

Trabalho 3. Relembrando as expansões em série das funções $\arctan x$ e $\arcsen x$

$$\arctan x = x - \frac{x^3}{3} + \frac{x^5}{5} - \frac{x^7}{7} + \dots,$$

e

$$\arcsen x = x + \frac{1}{2} \frac{x^3}{3} + \frac{1 \times 3}{2 \times 4} \frac{x^5}{5} + \frac{1 \times 3 \times 5}{2 \times 4 \times 6} \frac{x^7}{7} + \dots,$$

obtenha duas fórmulas alternativas para o cálculo de π .

Escreva uma *script* para calcular aproximações para π usando as fórmulas referidas e considerando um número de termos em cada série sucessivamente igual a 10, 20, ..., 100, 200, 300, ..., 1000. Comente os resultados obtidos.