Matemática Computacional

maria irene falcão

Mestrado em Matemática e Computação

1. Aritmética Computacional :: Resumo

Sistemas de numeração de vírgula flutuante
Erros, estabilidade e condicionamento

1.1 Sistema de numeração

Sistema $F(b,t,m,M)$

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

b - base; t - número de dígitos da mantissa;
m - valor mínimo do exponente;
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

$$\pm(.d_1 d_2 \dots d_t)_b \times b^e$$

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 $(.d_1 d_2 \dots d_t)_b$ designa $d_1\, b^{-1} + d_2\, b^{-2} + \cdots + d_t\, b^{-t}$.

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 desnormalizados[1] é

$$b^{m-t}$$

Ao conjunto

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

chamamos conjunto dos números representáveis.

➠ 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 forma[2], 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,

$$\epsilon := b^{1-t}$$

---

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

[2] Existem 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

➠ **Operações de vírgula flutuante**

Representaremos as operações de vírgula flutuante pelo símbolo usual rodeado por ○; 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 em sistema de vírgula flutuante: simples e duplo.

O formato simples corresponde ao sistema

$$F(2, 24, -125, 128)$$

e o duplo corresponde a

$$F(2, 53, -1021, 1024)$$

. Ambos os sistemas admitem números desnormalizados.

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

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 $x > \Omega$, $fl(x) = \mathbf{Inf}$;
- se $x < -\Omega$, $fl(x) = -\mathbf{Inf}$;
- se $2^{m-t} \leq x < \omega$, $fl(x)$ é o número desnormalizado mais próximo de x;
- se $x < 2^{m-t}$, $fl(x) = 0$.

[3] IEEE- Institute for Electrical and Electronics Engineers.

Nota: *Por defeito, o MATLAB trabalha no sistema de numeração de norma IEEE em formato duplo, isto é, no sistema* $F(2, 53, -1021, 1024)$.

1.3 Erro absoluto, erro relativo, algarismos significativos e casas decimais de precisão

Ao valor

$$E_{\tilde{x}} := x - \tilde{x}$$

chama-se erro absoluto do valor aproximado $\tilde{x}$ para x. Para $x \neq 0$, o valor

$$R_{\tilde{x}} := \frac{x - \tilde{x}}{x}$$

constitui o chamado erro relativo do valor aproximado $\tilde{x}$ para x.[4]

Dizemos que $\tilde{x}$ é uma aproximação para x com p casas decimais de precisão, se p é o maior inteiro tal que

$$|x - \tilde{x}| \leq 0.5 \times 10^{-p}$$

Dizemos que $\tilde{x}$ é uma aproximação para x com q algarismos (decimais) significativos, se q é o maior inteiro para o qual se tem

$$|x - \tilde{x}| \leq 0.5 \times 10^{-q} \times 10^{e}$$

onde e é o expoente de x na notação normalizada (na base decimal).

➠ Erros de arredondamento

Sejam $\mathcal{F} := F(b, t, m, M)$ e $x = (-1)^s m_x b^e \in R_{\mathcal{F}}$ não nulo e normalizado (i.e. $b^{-1} \leq m_x < 1$, $m \leq e \leq M$).

➢ **Erro absoluto de arredondamento**:

$$|E_{fl(x)}| = |x - fl(x)| \leq \frac{1}{2} b^{-t} b^{e} = \mu b^{e-1}$$

➢ **Erro relativo de arredondamento**:

$$|R_{fl(x)}| = \frac{|x - fl(x)|}{|x|} \leq \frac{\frac{1}{2} b^{-t}\, b^{e}}{b^{-1}\, b^{e}} = \frac{1}{2} b^{1-t} = \mu.$$

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

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

[4] Muitas vezes estamos interessados apenas no valor absoluto destas quantidades, designado-as pelos mesmos nomes, caso tal seja claro pelo contexto.

➠ **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 $\widetilde{S}, \widetilde{P}$ e $\widetilde{Q}$ 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:

$$E_{\widetilde{S}} = E_{\tilde{x}} + E_{\tilde{y}} \qquad E_{\widetilde{P}} = E_{\tilde{x}}y + E_{\tilde{y}}x - E_{\tilde{x}}E_{\tilde{y}} \qquad E_{\widetilde{Q}} = \frac{yE_{\tilde{x}} - xE_{\tilde{y}}}{y\tilde{y}}$$

$$R_{\widetilde{S}} = \frac{x}{x+y}R_{\tilde{x}} + \frac{y}{x+y}R_{\tilde{y}} \qquad R_{\widetilde{P}} = R_{\tilde{x}} + R_{\tilde{y}} - R_{\tilde{x}}R_{\tilde{y}} \qquad R_{\widetilde{Q}} = \frac{R_{\tilde{x}} - R_{\tilde{y}}}{1 - R_{\tilde{y}}}$$

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

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

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

$$\frac{|x\frac{\partial f(x,y)}{\partial x}|}{|f(x,y)|} \quad \text{e} \quad \frac{|y\frac{\partial f(x,y)}{\partial y}|}{|f(x,y)|}$$

são os números de condição de f em (x, y) relativamente à variável x e à variável y, respetivamente.

1.5 Notas e referências

➠ Funções disponíveis em MATLAB

As seguintes funções estão disponíveis em `https://w3.math.uminho.pt/~mif/MMC/`

Função	Objetivo
fl	Arredonda um número em $F(10, t, m, M)$
Fr_dec2bin	Converte uma fração decimal para a base 2)
sistNumFlgui	Representa graficamente os números positivos do sistema $F(2, t, m, M)$

➠ 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/`