

## Introducción al cálculo numérico

### La matemática versus el cálculo numérico

Existen diferencias fundamentales entre ambas áreas, pero la principal es que la matemática frecuentemente emplea procesos infinitos o infinitesimales, en cambio el cálculo numérico nunca los usa:

$$\frac{dy}{dx} \underset{h \rightarrow 0}{=} \lim \frac{y(x+h) - y(x)}{h}, \quad \int_a^b f(x) dx = \lim_{|\Delta x_i| \rightarrow 0} \sum_{i=1}^N f(x_i) \Delta x_i$$

Otra diferencia es que los números en la matemática no tienen limitaciones en su representación:

$$e = 1 + \frac{1}{1} + \frac{1}{1 \cdot 2} + \frac{1}{1 \cdot 2 \cdot 3} + \dots \approx 2,7182818284590$$

$$\frac{1}{6} = 0,1666666666\dots$$

### La representación binaria de números

En la práctica, toda computadora moderna usa esta representación binaria.

$$638,13_{10} = 6 \cdot 10^2 + 3 \cdot 10^1 + 8 \cdot 10^0 + 1 \cdot 10^{-1} + 3 \cdot 10^{-2}$$

$$1010,011_2 = 1 \cdot 2^3 + 0 \cdot 2^2 + 1 \cdot 2^1 + 0 \cdot 2^0 + 0 \cdot 2^{-1} + 1 \cdot 2^{-2} + 1 \cdot 2^{-3}$$

Para convertir enteros de la representación decimal a la binaria se utiliza la siguiente técnica (y se invierte la sintaxis para convertir de binario a decimal):

$295_{10}$  a binario

$$\begin{array}{r}
 295 \div 2 \quad (1) \\
 147 \div 2 \quad (1) \\
 73 \div 2 \quad (1) \\
 36 \div 2 \quad (0) \\
 18 \div 2 \quad (0) \\
 9 \div 2 \quad (1) \\
 4 \div 2 \quad (0) \\
 1 \div 2 \quad (1) \\
 0
 \end{array}$$

$100100111_2$

$100100111_2$  a decimal

$$\begin{array}{r}
 100100111_2 \\
 \downarrow \\
 1 \cdot 2 + 0 = 2 \\
 \boxed{2 \cdot 2 + 0 = 4} \\
 \boxed{4 \cdot 2 + 1 = 9}
 \end{array}$$

$$9 \cdot 2 = 18$$

$$18 \cdot 2 = 36$$

$$36 \cdot 2 = 73$$

$$73 \cdot 2 = 147$$

$$147 \cdot 2 = 295$$

$$295_{10}$$

Para convertir números menores a 1:

$0,314_{10}$  a binario

$$\begin{array}{r}
 0,314 \cdot 2 \\
 0,628 \cdot 2 \\
 1,256 \cdot 2 \\
 0,512 \cdot 2 \\
 1,024 \cdot 2 \\
 0,048 \cdot 2 \\
 0,096 \cdot 2 \\
 \vdots
 \end{array}$$

$0,010100\dots_2$

$0,010011_2$  a decimal

$$\begin{array}{r}
 0,010011_2 \\
 \downarrow \\
 1 \cdot \frac{1}{2} + 1 = \frac{3}{2} \\
 \boxed{\frac{3}{2} \cdot \frac{1}{2} + 0 = \frac{3}{4}} \\
 \boxed{\frac{3}{4} \cdot \frac{1}{2} + 0 = \frac{3}{8}}
 \end{array}$$

$$\frac{3}{8} \cdot \frac{1}{2} + 1 = \frac{19}{16}$$

$$\frac{19}{16} \cdot \frac{1}{2} + 0 = \frac{19}{32}$$

$$\frac{19}{32} \cdot \frac{1}{2} = \frac{19}{64} = 0,296875.$$

Virtalmente todo lenguaje de programación tiene alguna forma de convertir decimales a binarios y viceversa.

Entonces, ¿para qué aprendemos esto?

R. Para entender los principios básicos de cómo funcionan estas cosas y algunos fenómenos "inesperados". Considera

$$\frac{1}{10} = 0,1_{10} = 0,000110\dots_2$$

que no tiene una representación binaria finita. Al almacé

nar este valor en un espacio de memoria finita truncamos el valor y como resultado tenemos que

$$\sum_{i=1}^{10} 0,000110_2 = \frac{1}{10} \cdot 0,000110_2 < 1$$

Existen también otros dos sistemas de representación de números, bastante populares en la informática: el sistema octal y el hexadecimal:

Decimal	Binario	Octal	Hexadecimal
0	0000	0	0
1	0001	1	1
2	0010	2	2
3	0011	3	3
4	0100	4	4
5	0101	5	5
6	0110	6	6
7	0111	7	7
8	1000	10	8
9	1001	11	9
10	1010	12	a
11	1011	13	b
12	1100	14	c
13	1101	15	d
14	1110	16	e
15	1111	17	f

## Enteros y tumor flotante

## Números enteros

Generalmente usados como índices o para contar

## Números de coma flotante

que se usan virtualmente en todo cálculo científico

$$\pi = 0,314 \cdot 10^1$$

$$100\pi = 0,314 \cdot 10^3$$

$$\pi = 3,14 \cdot 10^{-2} \rightarrow \text{exponente}$$

$$\frac{I}{1000} = 0,314 \cdot 10^{-1}$$

mantis

## Números de coma flotante

De ahora en adelante nos enfocaremos en esta representación. Además, por conveniencia, usaremos un sistema que tendrá:

- \* tres dígitos para la mantisa
- \* un dígito para el exponente

Por ejemplo:

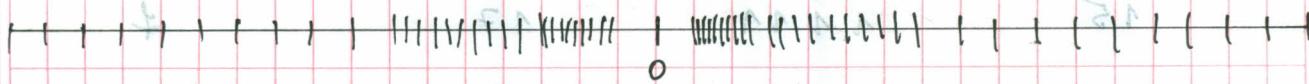
$$20 = 0,200 \cdot 10^2$$

$$\sqrt{3} = 0,173 \cdot 10^1$$

$$-\frac{1}{6} = -0,167 \cdot 10^0$$

En contraparte, lo que usualmente se usa en las computadoras es ~15 dígitos para la mantisa y un rango de  $10^{-308}$  a  $10^{308}$  para el exponente.

De esta manera, los números flotantes están más concentrados cerca del origen



ya que hay la misma cantidad de números entre  $0,100 \cdot 10^{-2}$  y  $0,999 \cdot 10^{-2}$  que entre  $0,100 \cdot 10^3$  y  $0,999 \cdot 10^3$  o cualquier otra década.

El número más pequeño es  $0,100 \cdot 10^{-9}$  y el siguiente es  $0,101 \cdot 10^{-9}$  con una diferencia de  $10^{-12}$ , que no puede representarse en el sistema.

El número más grande es  $0,999 \cdot 10^9$  y el anterior es  $0,998 \cdot 10^9$ , con una diferencia de  $10^6$ .

¿Cómo representamos el 0?

R. con  $0,000 \cdot 10^{-9}$

¿Qué sucede si excedemos la capacidad del sistema?

Vacuidad (underflow)

Si obtenemos un resultado menor a  $0,100 \cdot 10^{-9}$  no podemos representarlo.

Cuando esto ocurre, se lo reemplaza por  $0,000 \cdot 10^{-9}$  (0).

Desbordamiento (overflow)

Si obtenemos un resultado mayor a  $0,999 \cdot 10^9$  tampoco podemos representarlo.

Este caso no es tan fácil de manejar como el caso de vacuidad pues reemplazarlo por el número mayor ( $0,999 \cdot 10^9$ ) casi nunca es una buena idea.

Es preferible obtener un caso de vacuidad que uno de desbordamiento.

¿Realmente ocurren estos casos de vacuidad o desbordamiento?

R. Claro que sí. Consideré cuando  $x > 0,999 \cdot 10^9$

Es fácil ver que esto ocurre cuando

$$x > \ln 0,999 + 9 \ln 10 \approx 20,7$$

y esto resulta en un desbordamiento. Difícilmente diríamos que 20,7 es un valor extremo.

De manera similar, obtendríamos una vacuidad si

$$x < -20,7.$$

¿Cómo lidiar con los desbordamientos?

Imaginé que quiere calcular

$$1 - \frac{1}{e^x + 1}$$

para  $x > 21$ . Si evalúa directamente obtendrá un desbor-  
damiento para  $e^x$ . En cambio, si reescribe

$$1 - \frac{1}{e^x + 1} = \frac{1}{1 + e^{-x}}$$

obtendrá una vacuidad para  $e^{-x}$  (que será reemplazada por  
0) y su resultado no será un error sino 1.

### Errores de redondeo

Considere la multiplicación de dos números

$$0,348 \cdot 10^1 \times 0,744 \cdot 10^1 = 0,258912 \cdot 10^2.$$

Ya que estamos limitados a 3 dígitos en la mantisa, el resul-  
tado será una aproximación del valor real. Pensemos  
en dos situaciones

#### Redondeo

Que en la práctica se hace sumando 5 al primer dígito que descartemos

$$\begin{aligned} &+ 0,258912 \cdot 10^2 \\ &\quad \underline{\quad 5 \quad} \\ &\underline{0,259412 \cdot 10^2} \\ &\approx 0,259 \cdot 10^2 \end{aligned}$$

#### Truncamiento

Que simplemente consiste

en descartar los dígitos sobrantes sin más

$$0,258912 \cdot 10^2 \approx 0,258 \cdot 10^2$$

Siempre es preferible redondear a truncar.

Parecería que el error de redondeo es un problema secundario, pero recuerden que la computadora puede realizar millones de cálculos por segundo, en cuyo caso los errores de redondeo se pueden convertir en un problema mayor.

Además, considere la resta de dos números similares

$$\begin{array}{r} 0,314 \cdot 10^1 \\ - 0,313 \cdot 10^1 \\ \hline 0,100 \cdot 10^{-1} \end{array}$$

La cancelación de los primeros dígitos junto a la normalización hacen que los errores de redondeo pasen de la derecha al centro o incluso a la izquierda del número representado.

Como otro ejemplo, imagine que deseamos resolver la ecuación  $x^2 + 80x + 1 = 0$  usando la fórmula estándar

$$x_r = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a} = \begin{cases} -0,800 \cdot 10^2 \text{ para } + \\ 0,000 \cdot 10^0 \text{ para } - \end{cases}$$

Pero si usamos

$$x_1 = -\operatorname{signo}(b) \frac{|b| + \sqrt{b^2 - 4ac}}{2a} = -0,800 \cdot 10^2$$

$$x_2 = \frac{c}{ax_1} = -0,125 \cdot 10^{-1}$$

tenemos que con estas soluciones

$$(x - x_1)(x - x_2) = x^2 + 80x + 1$$

como debería ser. Finalmente, notemos que las soluciones verdaderas son:

$$x_1 = -40 - \sqrt{1599} \approx -79,987 \text{ y } x_2 = -40 + \sqrt{1599} \approx 0,012502.$$

Un problema severo con el redondeo (y peor aún con el truncamiento)

es la correlación interna de los números que desemboca en un error de redondeo que tiene siempre la misma dirección.

### Reordenando expresiones

Uno de los cálculos más elementales es la evaluación de funciones. Por lo que ya vimos pueden ocurrir pérdidas en la exactitud de los cálculos, debidas principalmente a la cancelación de números similares. Por ejemplo, considere la función

$$f(x) = \sqrt{x+1} - \sqrt{x}$$

para valores grandes de  $x$  existirán severos problemas de cancelación. Para evitar esto podemos reescribir la función

$$f(x) = \sqrt{x+1} - \sqrt{x} = \frac{1}{\sqrt{x+1} + \sqrt{x}}$$

A pesar de que parezca que existen muchos trucos para evaluar expresiones, en realidad son exactamente los mismos que se utilizan para evaluar límites en cálculo. En general, es una buena idea evitar restas.

### Aproximaciones

Por ejemplo, para  $x$  pequeño

$$\frac{\tan x - \sin x}{x^3} = \frac{\left(x + \frac{x^3}{3} + \frac{2}{15}x^5 + \dots\right) - \left(x - \frac{x^3}{3!} + \frac{x^5}{5!} + \dots\right)}{x^3} \approx \frac{1}{2} + \frac{x^2}{8}$$

O emplear el teorema del valor medio

$$f(b) - f(a) = (b-a) f'(\theta), \quad a < \theta < b$$

donde  $\theta$  es algún número en  $(a, b)$ . Si no se conoce nada más acerca de la función entonces se hace  $\theta = \frac{a+b}{2}$ . Por

ejemplo, para  $x \ll c$  tenemos

$$\ln(c+x) - \ln c = \ln\left(1 + \frac{x}{c}\right) = \frac{x}{c+0} \approx \frac{x}{c+x/2}$$

Veamos otro ejemplo. Supongamos que queremos evaluar

$$\frac{1 - \cos x}{x^2}$$

para  $x$  pequeño y consideremos la aproximación

$$\frac{1 - \cos x}{x^2} \approx \frac{1}{2} - \frac{x^2}{24}$$

>>> import math

>>> x = 2e-12

>>> (1. - math.cos(x))/x\*\*2

0.0

>>> 0.5 - x\*\*2/24

0.5

Que la máquina haga el trabajo

Imagine que quiere evaluar

$$e^{ax} - 1$$

para distintos valores de  $x$ . Si  $x$  no es muy pequeño, podemos evaluarla directamente. Caso contrario es mejor usar los primeros términos de la serie de MacLaurin

$$ax + (ax)^2/2 + (ax)^3/6 + (ax)^4/24 + \dots$$

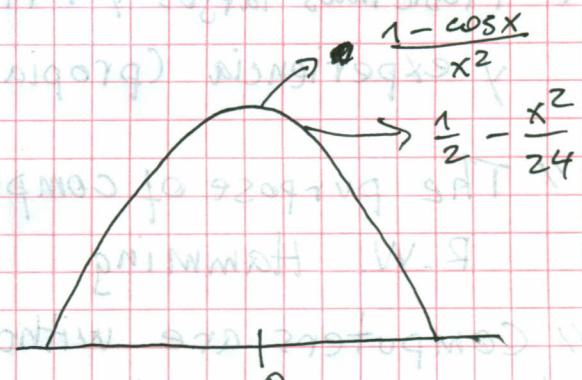
Por ejemplo

if abs(ax) > 1e-5:

r = math.exp(ax) - 1

else:

r = ax + (ax)\*\*2/2 + (ax)\*\*3/6.



## Algunas ideas para concluir

- \* Para problemas que deben resolverse una sola vez, el tiempo de computación no es muy importante.
- \* Algunos problemas requieren pensar mucho y experimentar mucho antes de implementar el programa.
- \* Los programas o listerías pensadas para otras personas deben hacerse con mucho cuidado.
- \* Problemas largos y tortuosos requieren de preparación y experiencia (propia y ajena).

"The purpose of computing is insight, not numbers"

R.W. Hamming

"Computers are without a doubt the most potent thinking tools we have, not just because they take the drudgery out of many intellectual tasks, but also because many of the concepts computer scientists have invented are excellent thinking tools in their own right."

D. E. Dennett

"To compute is to sample, and one then enters the domain of statistics with all its uncertainties"

R.W. Hamming

$$z - g(x) \leq x \leq z + g(x)$$

$$1 - (x - g(x))g(x) \leq 1$$

:32/9

$$1 - g(x)(x - g(x)) \leq 1$$