Notas de Cálculo Numérico

Ramiro Dibur

Índice

1 Aritmética de punto flotante

En esta sección se introduce cómo se representan los números en una computadora, los errores asociados a estas representaciones, y cómo estos errores pueden propagarse en los cálculos. Además, se presentan ejemplos de problemas mal condicionados y se discute el concepto de estabilidad numérica.

1.1 Números de máquina

Un número real $x \neq 0$ puede representarse en una base $b \in \mathbb{N}$, $b \geq 2$, mediante una expansión en serie:

$$x = \pm \left(d_k b^k + d_{k-1} b^{k-1} + \cdots + d_0 b^0 + d_{-1} b^{-1} + d_{-2} b^{-2} + \dots \right)$$

donde los dígitos d_i satisfacen $0 \leq d_i \leq b - 1$. Esta representación es única si se excluyen las expansiones infinitas que terminan en una secuencia de $b - 1$.

Sin embargo, como en las computadoras tenemos espacio finito, nece

Definición 1.1. Un número de máquina (o número de punto flotante) x en base b se representa generalmente en forma normalizada como

$$x = \pm (0.d_1 d_2 \dots d_m)_b \times b^e$$

donde $d_1 d_2 \dots d_m$ es la mantisa y e el exponente.

Dado un número real x, si no es un número de máquina, debe ser aproximado por uno. El proceso de aproximación se llama redondeo. Dos estrategias comunes son:

- **Redondeo al más cercano**: Se elige el número de máquina más cercano a x.
- **Truncamiento (o redondeo hacia cero)**: Se elige el número de máquina más cercano a x en dirección hacia cero.

Es posible que el resultado de una operación no pueda representarse exactamente debido a las limitaciones en la precisión de los números de máquina. Esto puede dar lugar a dos tipos principales de errores:

- **Overflow:** Ocurre cuando el resultado de una operación excede en magnitud el número de máquina más grande que se puede representar. Esto puede llevar a resultados infinitos o errores en los cálculos.
- **Underflow:** Ocurre cuando el resultado de una operación es menor en magnitud que el número de máquina positivo más pequeño que se puede representar (sin ser cero). En este caso, el resultado puede aproximarse a cero, lo que puede introducir errores de redondeo.

1.2 Error absoluto y relativo

Definición 1.2. Si $fl(x)$ es la representación de máquina de x, el error absoluto se define como

$$|x - fl(x)|,$$

y el error relativo como

$$\frac{|x - fl(x)|}{|x|}.$$

Para el redondeo al más cercano, el error absoluto está acotado por

$$|x - fl(x)| \leq \frac{1}{2} b^{e-m},$$

donde e es el exponente tal que $b^{e-1} \leq |x| < b^e$. El error relativo está acotado por

$$\left| \frac{x - fl(x)}{x} \right| \leq \frac{1}{2} b^{1-m}.$$

1.3 Épsilon de Máquina

Definición 1.3. El *épsilon de máquina*, denotado por ε, es la cota superior del error relativo cometido al representar un número real mediante un número de máquina utilizando el redondeo al más cercano.

$$\varepsilon = \frac{1}{2} b^{1-m}.$$

Observación. En `python`, con variables del tipo `float`, el valor de ε se puede obtener utilizando el módulo `sys`:

```
1 import numpy as np
2 np.finfo(float).eps
```

Este valor representa la diferencia más pequeña entre 1,0 y el siguiente número representable mayor que 1,0 en la aritmética de punto flotante de la máquina.

En general, cuando realizamos operaciones, los errores de la máquina se acumulan. Veamos qué pasa cuando sumamos dos números cualesquiera en la máquina.

Sean x e y números con el mismo signo tales que sus errores relativos son μ_x y μ_y, respectivamente. Notemos que $|\mu_x|, |\mu_y| \leq \varepsilon$. Sea μ_z el error relativo de la suma. Entonces,

$$\begin{aligned} x \oplus y &= fl(fl(x) + fl(y)) \\ &= (fl(x) + fl(y))(1 + \mu_z) \\ \implies |x \oplus y - (x + y)| &\leq (x + y)2\varepsilon + O(\varepsilon^2). \end{aligned}$$

Y entonces nos queda

$$\frac{|x \oplus y - (x + y)|}{|x + y|} \lesssim 2\varepsilon.$$

2 Aproximación de EDOs

Consideramos las ecuaciones diferenciales ordinarias de orden 1. El problema de valores iniciales (PVI) consta en hallar $x : (a, b) \to \mathbb{R}$ tal que

$$\begin{cases} x'(t) = f(t, x) \\ x(t_0) = x_0. \end{cases}$$

Las hipótesis más usuales son que

- f continua.
- f Lipschitz en la variable x.

Dadas estas hipótesis, existe una única solución del problema de valores iniciales. Además, son necesarias para los métodos numéricos, ya que sino no hay forma de corroborar qué solución está dando un método.

2.1 Método de Euler

Método de Euler. Dado el problema de valores iniciales

$$\begin{cases} x'(t) = f(t, x) \\ x(t_0) = x_0, \end{cases}$$

queremos aproximar $x(T)$ para $T > t_0$. Tomamos $N \in \mathbb{N}$ y $h = \frac{T - t_0}{N}$. Entonces,

- $t_i = t_0 + h \cdot i$ para $0 \leq i \leq N$.
- $x_{i+1} = x_i + h \cdot f(x_i, t_i)$