Cálculo Numérico - Elementos de Cálculo Numérico

Números de máquina - Ecuaciones Diferenciales Ordinarias

Clase dictada por: Mercedes Pérez Millán

(Dto. de Matemática–FCEyN–Universidad de Buenos Aires, IMAS-CONICET)

19 de agosto de 2024

Números de máquina

Ejemplo:

$$\color{blue}{8{,}75}\color{black} = 8 + \frac{1}{2} + \frac{1}{4} = 2^3 + 2^{-1} + 2^{-2} = (1000{,}11)_2$$
$$= (1.\,\boxed{00011}\,)_2 \times 2^3$$

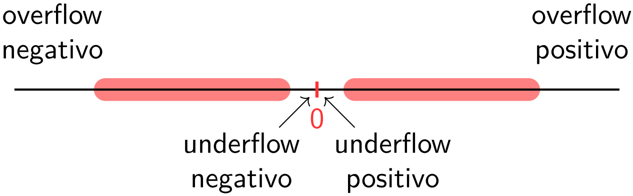
En precisión simple se guarda el exponente $+127$:
$3 + 127 = 130 = 128 + 2 = 2^7 + 2 = (\,\boxed{10000010}\,)_2$

$$\underset{\substack{\uparrow \\ \text{signo}}}{0}\;\underbrace{1\,0\,0\,0\,0\,0\,1\,0}_{\text{exponente}}\;\underbrace{0\,0\,0\,1\,1\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0}_{\text{mantisa}}$$

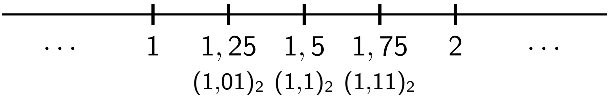
$$\frac{1}{10} = (0,0001100110011 0011\ldots)_2$$

¿Qué número va a guardar?

Números de máquina:



Si la mantisa tiene $m = 3$ dígitos:



$$\mathbb{R}^* = \{\pm(0.a_1 a_2 \dots a_m)_\beta \times \beta^\ell \mid a_1 \neq 0, a_i \in \{0, 1, \dots, \beta - 1\},$$
$$- n_1 \leq \ell \leq n_2, \ell \in \mathbb{Z}\} \cup \{0\}$$

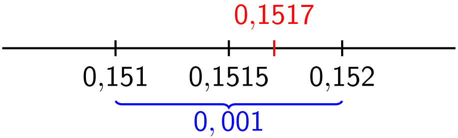
$$fl : \begin{array}{ccc} \mathbb{R} & \rightarrow & \mathbb{R}^* \\ x & \mapsto & x* = fl(x) \end{array}$$

Ejemplo con $\beta = 10$ y $m = 3$:

- Truncado: $fl(151,7) = fl(0{,}1517 \times 10^3) = 0{,}151 \times 10^3 = 151.$
- Redondeo: $fl(151,7) = fl(0{,}1517 \times 10^3) = 0{,}152 \times 10^3 = 152.$

  Error absoluto (EA):

$$\begin{aligned} |x - fl(x)| &= |0{,}1517 \times 10^3 - 0{,}152 \times 10^3| = |0{,}0003| \times 10^3 \\ &= |0{,}3 \times 10^{-3}| \times 10^3 \leq 0{,}5 \times 10^{3-3}. \end{aligned}$$



Error realtivo (ER):

$$\frac{|EA|}{|x|} \leq \frac{0,5 \times 10^{3-3}}{|0{,}1517 \times 10^3|} = \frac{0,5 \times 10^{-3}}{|0{,}1517|} \leq \frac{0,5 \times 10^{-3}}{|0{,}1|} = 0,5 \times 10^{1-3}.$$

En general, $|ER| \leq 0{,}5 \times \beta^{1-m} = \varepsilon$.

Propagación de errores

Supongamos $x \cdot y > 0$, $fl(x) = x(1+\delta_x)$, $fl(y) = y(1+\delta_y)$,
$fl(fl(x) + fl(y)) = (x(1+\delta_x) + y(1+\delta_y))(1+\delta_+)$, entonces:

$$\begin{aligned}
ER &= \frac{|x + y - (x(1+\delta_x) + y(1+\delta_y))(1+\delta_+)|}{|x+y|} \\
&= \frac{|\cancel{x} + \cancel{y} - \cancel{x} - \cancel{y} - x\delta_x - y\delta_y - x\delta_+ - y\delta_+ - x\delta_x\delta_+ - y\delta_y\delta_+|}{|x+y|} \\
&\leq \frac{|x|\overbrace{|\delta_x|}^{\leq\varepsilon} + |y|\overbrace{|\delta_y|}^{\leq\varepsilon} + |x|\overbrace{|\delta_+|}^{\leq\varepsilon} + |y|\overbrace{|\delta_+|}^{\leq\varepsilon} + |x|\overbrace{|\delta_x|}^{\leq\varepsilon}\overbrace{|\delta_+|}^{\leq\varepsilon} + |y|\overbrace{|\delta_y|}^{\leq\varepsilon}\overbrace{|\delta_+|}^{\leq\varepsilon}}{|x+y|} \\
&\leq \frac{|x|(2\varepsilon + \varepsilon^2) + |y|(2\varepsilon + \varepsilon^2)}{|x+y|} = \frac{(|x| + |y|)(2\varepsilon + \varepsilon^2)}{|x+y|} \\
&\underset{x \cdot y > 0}{=} \frac{\cancel{|x+y|}(2\varepsilon + \varepsilon^2)}{\cancel{|x+y|}} = 2\varepsilon + \varepsilon^2
\end{aligned}$$

Cancelación catastrófica

$\beta = 10,\ m = 3.$

$$x = 0,15876, y = 0,14964 \quad \Rightarrow \quad x - y = 0,00912$$

$$fl(x) = 0,159, fl(y) = 0,15 \quad \Rightarrow \quad fl(x) - fl(y) = 0,009$$

y perdimos dos dígitos significativos.

Sumas de números de distinta magnitud

$\beta = 10,\ m = 3.$

$$x = 132; \quad y = 0,2; \quad z = 0,4.$$

Notar que $fl(x) = x,\ fl(y) = y,\ fl(z) = z.$

$$fl(x+y) = fl(132,2) = 132$$

$$fl(fl(x+y) + fl(z)) = fl(132,4) = 132.$$

Pero:

$$fl(y+z) = fl(0{,}6) = 0{,}6$$

$$fl(x + fl(y+z)) = fl(132,6) = 133.$$

Y tenemos $fl(x+y+z) = 133.$

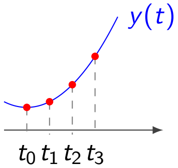
¿Cómo conviene hacer las sumas?

Problema de valores iniciales

$$\begin{cases} y'(t) = f(t, y(t)) \\ y(t_0) = y_0 \end{cases}$$

Son *pocos* los casos donde podemos encontrar la solución analíticamente, por lo que vamos a estudiar métodos numéricos para aproximar los valores de la solución.

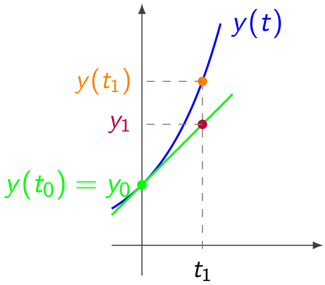
$$(y_0, y_1, y_2, y_3, \dots) \sim (y(t_0), y(t_1), y(t_2), y(t_3), \dots)$$



Método de Euler

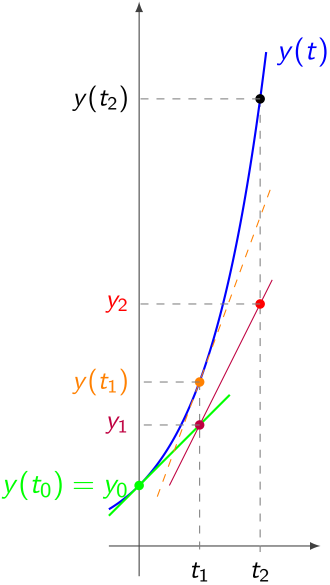
La idea es aproximar el valor de y en t_1 usando el valor de y y su derivada en t_0 por medio del desarrollo de Taylor de primer orden:

$$y(t_1) \sim y_1 = y(t_0) + y'(t_0)(t_1 - t_0) = y(t_0) + f(t_0, y(t_0))(t_1 - t_0)$$



¿Y $y_2 \sim y(t_2)$?

$$t(t_2) \sim t_2 = y_1 + f(t_1, y_1)(t_2 - t_1)$$



Método de Euler

Método:

Sean $h = \frac{T_F - t_0}{N}$, $t_n = t_0 + nh$,

$$\begin{cases} y_0 & = y(t_0) \\ y_{n+1} & = y_n + hf(t_n, y_n) \end{cases}$$

$n = 0, 1, \ldots, N-1$.

¿Y el error?

Sea

$$\hat{y}_n := y(t_{n-1}) + hf(t_{n-1}, y(t_{n-1})).$$

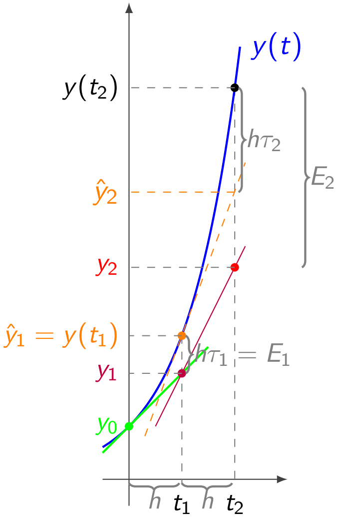
(la recta tangente a y en t_{n-1} evaluada en t_n.)
Definimos

$$h\tau_n := y(t_n) - \hat{y}_n,$$

el *error de truncado local*.
Definimos

$$E_n = y(t_n) - y_n,$$

el *error global*.



Error de truncado local

Recordemos el desarrollo de Taylor de y centrado en t_{n-1}:

$$y(t) = y(t_{n-1}) + y'(t_{n-1})(t - t_{n-1}) + \frac{y''(\theta)}{2}(t - t_{n-1})^2,$$

para θ entre t y t_{n-1}. Como $y'(t_{n-1}) = f(t_{n-1}, y(t_{n-1}))$ y $t_n - t_{n-1} = h$, deducimos:

$$\begin{aligned} h\tau_n &= y(t_n) - \hat{y}_n = y(t_n) - (y(t_{n-1}) + hf(t_{n-1}, y(t_{n-1}))) \\ &= \frac{y''(\theta_n)}{2}h^2, \quad \theta_n \in (t_{n-1}, t_n). \end{aligned}$$

Y podemos acotar

$$|\tau_n| \leq \max_{t \in [t_0, t_F]} |y''(t)| \, \frac{h}{2}, \qquad \text{para } n = 1, \ldots, N.$$

Observemos que

$$t \;\mapsto\; (t, y(t)) \;\overset{f}{\mapsto}\; f(t, y(t))$$

$$\begin{aligned} y''(t) &= \frac{d}{dt} f(t, y(t)) = \frac{\partial f}{\partial t}(t, y(t)).\, 1 + \frac{\partial f}{\partial y}(t, y(t)).\, y'(t) \\ &= \frac{\partial f}{\partial t}(t, y(t)) + \frac{\partial f}{\partial y}(t, y(t)).\, f(t, y(t)) \\ &= f_t(t, y(t)) + f_y(t, y(t)).\, f(t, y(t)). \end{aligned}$$

Si C_{MAX} es tal que $\max\limits_{t \in [t_0, t_F]} |y''(t)| \leq C_{MAX}$, tenemos entonces:

$$|\tau_n| \leq C_{MAX}\, \frac{h}{2}, \qquad \text{para } n = 1, \ldots, N.$$

Error global

Decimos que $f : \mathbb{R}^2 \to \mathbb{R}^2$ satisface la *condición de Lipschitz en la segunda variable* si existe $L > 0$ tal que

$|f(t,u) - f(t,v)| \leq L|u - v|$ para toda elección posible de t, u, v.

Vamos a considerar mayormente:

$$\max_{\substack{t\in[t_0,t_F]\\ y\in[a,b]}} |f_y(t,y)| \leq L,$$

para algún intervalo $[a,b]$ (*ya veremos...*).
Tenemos entonces la siguiente cota para el error global:

$$\boxed{|E_N| = |y(t_F) - y_N| \leq \frac{e^{L(t_F - t_0)} - 1}{L}\tau_{max}.}$$

Ejercicio

Considerar el problema de valores iniciales:

$$\begin{cases} y'(t) &= t\cos^2(y(t)), \\ y(0) &= 5. \end{cases}$$

- En “papel”:
  1. Escribir la iteración del método de Euler correspondiente a este problema.
  2. Hallar h para que el error al estimar $y(1)$ con el método de Euler sea menor que 10^{-3}.
- Con Python: Escribir un programa que implemente el método de Euler explícito para este PVI y graficar la solución obtenida para distintos valores de h.

Resolución:

1) La iteración del método de Euler con paso $h = \frac{T_F - t_0}{N} = \frac{1-0}{N} = \frac{1}{N}$, $t_n = t_0 + nh = 0 + nh = nh$ es

$$\begin{cases} y_0 & = 5 \\ y_{n+1} & = y_n + hf(t_n, y_n) = y_n + ht_n \cos^2(y_n) \end{cases}$$

$n = 0, 1, \ldots, N-1.$

Resolución:

2) Buscamos h para que $|E_N| < 10^{-3}$. Hallemos C_{MAX} y L para este problema. En este caso, $f(t,y) = t\cos^2(y)$.

$$\begin{aligned} y''(t) &= f_t(t, y(t)) + f_y(t, y(t)) f(t, y(t)) \\ &= \cos^2(y(t)) + t.\, 2\cos(y(t))(-\operatorname{sen}(y(t))) f(t, y(t)) \\ &= \underline{\cos^2(y(t))} - 2t\cos(y(t))\operatorname{sen}(y(t))\, t\underline{\cos^2(y(t))} \\ &= \cos^2(y(t))[1 - 2t^2\cos(y(t))\operatorname{sen}(y(t))]. \end{aligned}$$

De esta forma:

$$\begin{aligned} |y''(t)| &\leq |\cos(y(t))|^2[1 + |2t^2||\cos(y(t))||\operatorname{sen}(y(t))|] \\ &\underset{t\in[0,1]}{\leq} 1^2.\,[1 + 2.\,1.\,1] = 3 =: C_{MAX} \end{aligned}$$

Y $|f_y(t,y)| = |t.\, 2\cos(y)(-\operatorname{sen}(y))| \leq |2t| \underset{t\in[0,1]}{\leq} 2 =: L$

Resolución:

Sabemos que

$$
\begin{aligned}
|E_N| &\leq \frac{e^{L(t_F - t_0)} - 1}{L} C_{MAX} \frac{h}{2} \\
&= \frac{e^{2(1-0)} - 1}{2} \cdot \frac{3h}{2} = \frac{3h}{4}(e^2 - 1) \leq 6h
\end{aligned}
$$

Ahora busquemos un valor de h para que este error sea menor que 10^{-3}:

$$
6h < 10^{-3} \Leftrightarrow h < \frac{1}{6000} \sim 0{,}0001666\ldots.
$$

Por lo tanto, cualquier $h < 0{,}0001666$ sirve para tal propósito.