Cálculo Numérico - Elementos de Cálculo Numérico
Números de máquina

Clase dictada por: Mercedes Pérez Millán

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

14 de agosto de 2024

Contenidos mínimos

Aritmética de punto fijo y flotante. Propagación de errores. Estabilidad numérica. Condicionamiento de matrices. Sistemas de ecuaciones lineales. Métodos directos: eliminación de Gauss, acumulación de errores y pivoteo, descomposición LU. Métodos iterativos: métodos de Jacobi, Gauss-Seidel y SOR. Aproximación de autovalores. Solución de ecuaciones no lineales. Métodos de bisección, de Newton, convergencia cuadrática, métodos de punto fijo. Interpolación polinomial. Formas de Lagrange y de Newton. Interpolación de Hermite, polinomios de Chebyshev. Productos escalares discretos y continuos. Polinomios ortogonales y cuadrados mínimos. Proyección ortogonal. Integración numérica: interpolación polinomial, reglas del trapecio y de Simpson, reglas compuestas, cuadratura de Gauss. Resolución numérica de ecuaciones diferenciales ordinarias. Métodos de Euler, de Taylor, de Runge-Kutta, métodos de paso variable y métodos de paso múltiple. Estabilidad relativa y absoluta.

Temas de la asignatura

P1) Aritmética de punto flotante

P2) Ecuaciones diferenciales: problemas de valores iniciales

P3) Ecuaciones diferenciales: problemas de valores de contorno

P4) Número de condición. Sistemas lineales.

P5) Resolución numérica de ecuaciones no lineales

P6) Interpolación

P7) Mínimos Cuadrados

P8) Integración numérica

Aulas

Días: Lunes y miércoles

Teórica (9-11hs)	Aula 8 - Pab. I
Práctica (11-14hs)	Aula 8 - Pab. I; Lab. 1111 y 1112 - Pab. 0

Régimen de aprobación

- Habrá **dos parciales** con sus respectivos recuperatorios.
- Para aprobar los trabajos prácticos se requiere aprobar los **dos parciales** o sus recuperatorios.
- Cada examen parcial/recuperatorio será calificado con una de las siguientes notas:
  - I = Insuficiente o No Aprobado
  - Una calificación numérica mayor o igual que 4.

Régimen de aprobación

- En caso de presentarse a recuperatorio (ya sea por tener el parcial Insuficiente o con la intención de subir su nota), el examen anterior que se recupera queda **anulado**, aún si hubiese estado aprobado, y la calificación se reemplaza por la obtenida en el recuperatorio.
- Para **aprobar** los trabajos prácticos de la materia se deberá tener *calificación mayor o igual que 4* en ambos parciales (o sus recuperatorios).
- Para poder **promocionar** la materia se deberá cumplir con una entrega de programación en Python y desarrollo teórico, a través del campus, con una instancia de defensa **oral**. Además, se deberá obtener un promedio mayor o igual que 6 entre los dos exámenes.
- El promedio no tendrá en cuentas los exámenes parciales o recuperatorios desaprobados o anulados.

Parciales:

- Primer Parcial: Miércoles 9/10, 9hs.
- Segundo Parcial: Miércoles 27/11, 9hs.
- Recuperatorio del Primer Parcial: Miércoles 4/12.
- Recuperatorio del Segundo Parcial: Miércoles 11/12.

Números de máquina

Ejemplo:

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

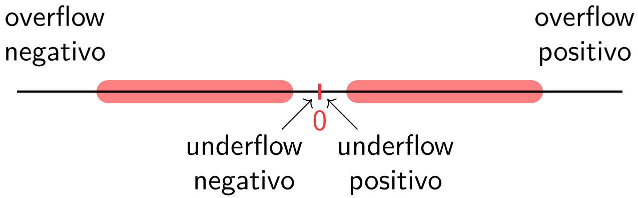
En precisión simple se guarda el exponente $+127$:
$3 + 127 = 130 = 128 + 2 = 2^7 + 2 = (\,10000010\,)_2$

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

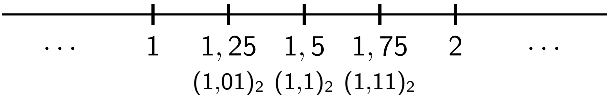
$$\frac{1}{10} = (0{,}00011001100110011\ldots)_2$$ Cuenta

¿Qué número va a guardar?

Números de máquina:



Si la mantisa tiene $m = 3$ dígitos:



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

$$fl : \underset{x}{\mathbb{R}} \underset{\mapsto}{\rightarrow} \underset{x* = fl(x)}{\mathbb{R}^*}$$

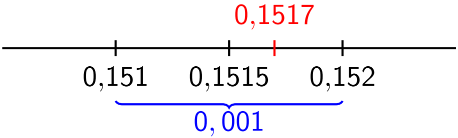
Ejemplo con $\beta = 10$ y $m = 3$:

- <u>Truncado</u>: $fl(151,7) = fl(0{,}1517 \times 10^3) = 0{,}151 \times 10^3 = 151.$

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

  Error absoluto (EA):

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



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?

Cómo escribir $0,1$ en binario

$$
\begin{array}{rcl}
 & 0,1 = & a_1 \times 2^{-1} + a_2 \times 2^{-2} + a_3 \times 2^{-3} + a_4 \times 2^{-4} + \dots \\
\times 2 & & \\
 & 0,2 = & \underbrace{a_1}_{=0} + a_2 \times 2^{-1} + a_3 \times 2^{-2} + a_4 \times 2^{-3} + \dots \\
\times 2 & & \\
 & 0,4 = & \underbrace{a_2}_{=0} + a_3 \times 2^{-1} + a_4 \times 2^{-2} + a_5 \times 2^{-3} + \dots \\
\times 2 & & \\
 & 0,8 = & \underbrace{a_3}_{=0} + a_4 \times 2^{-1} + a_5 \times 2^{-2} + a_6 \times 2^{-3} + \dots \\
\times 2 & & \\
 & 1,6 = & \underbrace{a_4}_{=1} + a_5 \times 2^{-1} + a_6 \times 2^{-2} + a_7 \times 2^{-3} + \dots \\
 & 0,6 = & a_5 \times 2^{-1} + a_6 \times 2^{-2} + a_7 \times 2^{-3} + \dots
\end{array}
$$

$$
\begin{array}{rccl}
 & 0,6 & = & a_5 \times 2^{-1} + a_6 \times 2^{-2} + a_7 \times 2^{-3} + \dots \\
\times 2 & & & \\
 & 1,2 & = & \underbrace{a_5}_{=1} + a_6 \times 2^{-1} + a_7 \times 2^{-2} + a_8 \times 2^{-3} + \dots \\
 & 0,2 & = & a_6 \times 2^{-1} + a_7 \times 2^{-2} + a_8 \times 2^{-3} + \dots \\
 & \vdots & \vdots & \vdots
\end{array}
$$

$$0,1 = (0,0\overline{0011}\dots)_2$$

Volver