MÉTODOS NUMÉRICOS I

Tema I: Introducción a los problemas del Análisis Numérico

Manuel Ruiz Galán

Curso 2018/2019
Doble Grado en Ingeniería Informática y Matemáticas

Índice Tema I

I.1. Introducción a los métodos numéricos: algoritmo

$n \geq 1$

$$x_n := \int_0^1 x^n e^x \, dx$$

Recurrencia

$$x_0 = e - 1$$

y

$$n \in \mathbb{N} \;\Rightarrow\; x_n = e - n x_{n-1}$$

$\{x_n\}_{n\geq 1} \subset \mathbb{R}_+$, decreciente, $\lim_{n\to\infty} x_n = 0$

redondeo

$$x_{12} = 0.1951$$

redondeo de x_0

$$x_0 = 1.7183$$

$$x_{12} = 8704.39$$

¡incluso para

$$x_0 = e - 1$$

Maxima, Mathematica

$$x_{22} = -59776.9!$$

I.1.1. Espacios normados

$$\text{Norma en problema numérico} \rightsquigarrow \left| \begin{array}{l} \text{problema bien planteado} \\ \text{errores, convergencia...} \end{array} \right.$$

$$\text{Idea de proximidad} \iff \text{control de la distancia}$$

Definición

Si E es un espacio vectorial real, diremos que una aplicación $\|\cdot\| : E \longrightarrow \mathbb{R}$ es una *norma* en E si verifica las siguientes propiedades:

(n1) $x \in E \Rightarrow \left| \begin{array}{c} \|x\| \geq 0 \\ \|x\| = 0 \Leftrightarrow x = 0 \end{array} \right.$.

(n2) $x, y \in E \Rightarrow \|x + y\| \leq \|x\| + \|y\|$ (*desigualdad triangular*).

(n3) $x \in E, \lambda \in \mathbb{R} \Rightarrow \|\lambda x\| = |\lambda|\|x\|$.

En tal caso –el espacio admite una norma–, E se llama *espacio normado*.

Ídem espacios complejos

Interpretación geométrica de la desigualdad triangular, norma euclídea $\mathbb{R}^2$ o $\mathbb{R}^3$

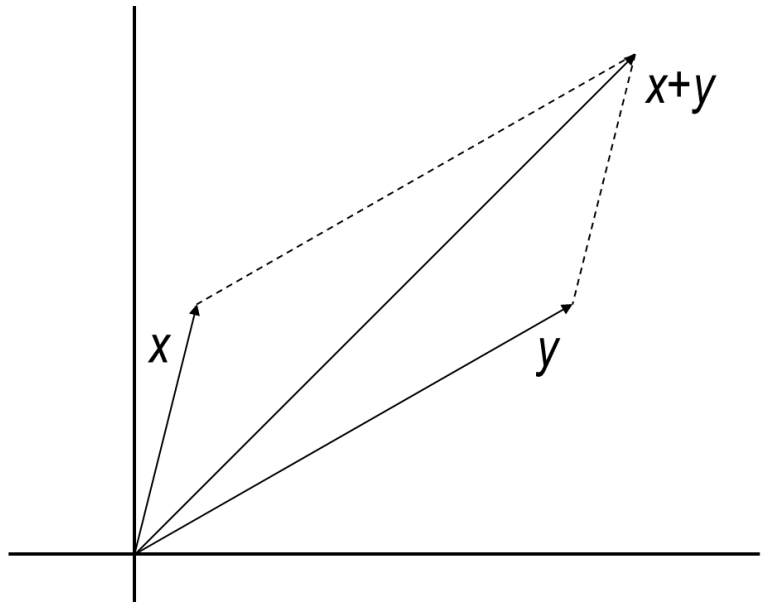


Figura: Desigualdad triangular

Ejemplos

- $E = \mathbb{R}^N$, $p \geq 1$, con la norma

$$\|\mathbf{x}\|_p := \left(\sum_{j=1}^{N} |x_j|^p \right)^{1/p}, \qquad (\mathbf{x} \in \mathbb{R}^N)$$

  $p = 2$ *norma euclídea*
  Comprobación en Relación de Ejercicios

- $E = \mathbb{R}^N$, con la *norma del máximo*:

$$\|\mathbf{x}\|_\infty := \max_{j=1,\ldots,N} |x_j|, \qquad (\mathbf{x} \in \mathbb{R}^N)$$

- $E = \mathbb{R}^{M\times N}$, *espacio de las matrices reales de M filas y N columnas*, dotado de la *norma de Frobenius*:

$$\|\mathbf{A}\|_F := \sqrt{\sum_{i=1}^{M}\sum_{j=1}^{N} a_{ij}^2}, \qquad (\mathbf{A} \in \mathbb{R}^{M\times N})$$

- $E = C([a,b])$ con la *norma del máximo*

$$\|f\|_\infty := \max_{a\le x\le b} |f(x)|, \qquad (f \in C([a,b]))$$

- $E = C^k([a,b])$, $k \in \mathbb{N}$, con la norma

$$\|f\|_k := \max_{j=0,1,\dots k} \|f^{(j)}\|_\infty, \qquad (f \in C^k([a,b]))$$

E espacio normado, $x^* \in E$ aproximación de $x \in E$

error absoluto

$$\|x^* - x\|$$

error relativo

$$\frac{\|x^* - x\|}{\|x\|}$$

Ejemplos

- $E = \mathbb{R}$, $x = 1/4$, $x^* = 0.23$
  error absoluto: $|x^* - x| = 0.02$

  error relativo: $\dfrac{|x^* - x|}{|x|} = 0.08$
- $E = \mathbb{R}^3$, $x = (1/5, 2, 1)$, $x^* = (0.19, 2.2, 0.9)$, norma del máximo
  error absoluto: $\|x^* - x\|_\infty = 0.2$

  error relativo: $\dfrac{\|x^* - x\|_\infty}{\|x\|_\infty} = 0.1$

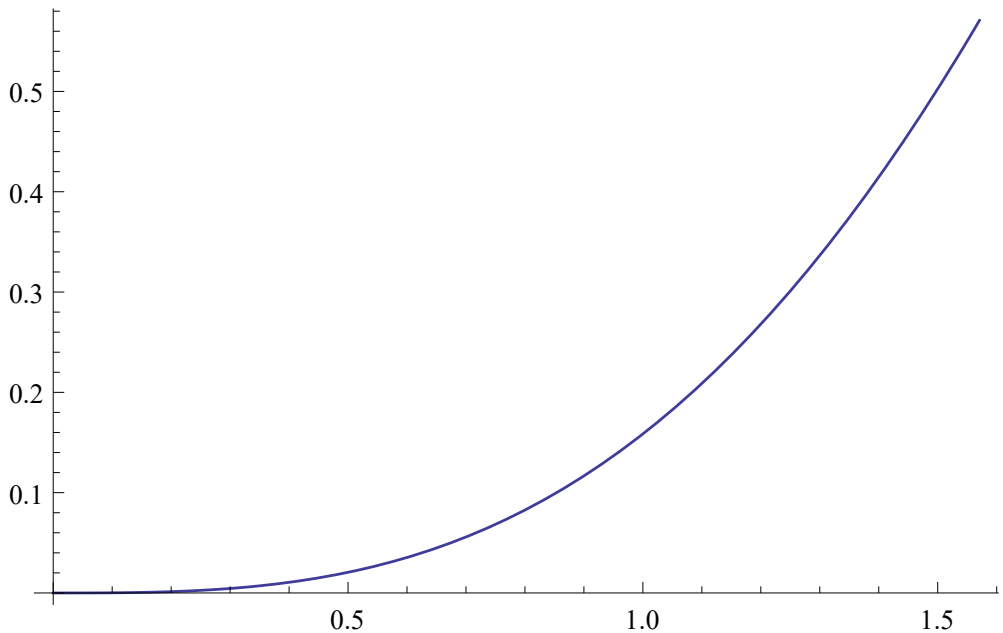
- $E = C([0, \pi/2])$, $f(t) = \text{sen}t$, $f^*(t) = t$, norma del máximo

  error absoluto

$$\|f^* - f\|_\infty = \frac{\pi}{2} - 1$$

  error relativo

$$\frac{\|f^* - f\|_\infty}{\|f\|_\infty} = \frac{\pi}{2} - 1$$



No guardan relación de orden alguna entre sí

Norma $\rightsquigarrow$ distancia $\rightsquigarrow$ topología

distancia entre dos vectores $x, y \in E$

$$\operatorname{dist}(x, y) := \|x - y\|$$

$\{x_n\}_{n\geq 1}$ en E *converge* a $x_0 \in E$

$$\varepsilon > 0 \;\Rightarrow\; \left[\, \text{existe } n_0 \in \mathbb{N} : \;\; n \geq n_0 \;\Rightarrow\; \|x_n - x_0\| < \varepsilon \;\right]$$

$$\lim_{n\to\infty} x_n = x_0$$

$$\Updownarrow$$

$$\lim_{n\to\infty} \|x_n - x_0\| = 0$$

X, Y subconjuntos no vacíos de sendos espacios normados
$f : X \longrightarrow Y$ continua en $x_0 \in X$ si

$$\varepsilon > 0 \;\Rightarrow\; \left[\text{existe } \delta > 0 : \begin{array}{c} x \in X \\ \|x - x_0\| < \delta \end{array} \;\middle|\; \Rightarrow\; \|f(x) - f(x_0)\| < \varepsilon \right]$$

Versión secuencial como en caso escalar

$$\mathbf{x} \in \mathbb{R}^N \;\Rightarrow\; \|\mathbf{x}\|_\infty \leq \|\mathbf{x}\|_1 \leq N\|\mathbf{x}\|_\infty$$

convergencia de sucesiones, continuidad equivalentes para ambas normas

$\|\cdot\|$, $\|\cdot\|_\circledast$ *equivalentes* si existen $c_1, c_2 > 0$ de forma que

$$x \in E \;\Rightarrow\; c_1\|x\| \leq \|x\|_\circledast \leq c_2\|x\|$$

Teorema

Todas las normas en un espacio normado finito dimensional son equivalentes.

Consúltese la demostración en [3, Proposition 1.3.6]

$(\mathbb{R}^N, \|\cdot\|_\infty)$

$$\lim_{n\to\infty} \mathbf{x}_n = \mathbf{x}_0$$

$$\Updownarrow$$

$$j = 1, \ldots, N \;\Rightarrow\; \lim_{n\geq 1} (\mathbf{x}_n)_j = (\mathbf{x}_0)_j$$

Ídem para cualquier norma de $\mathbb{R}^N$

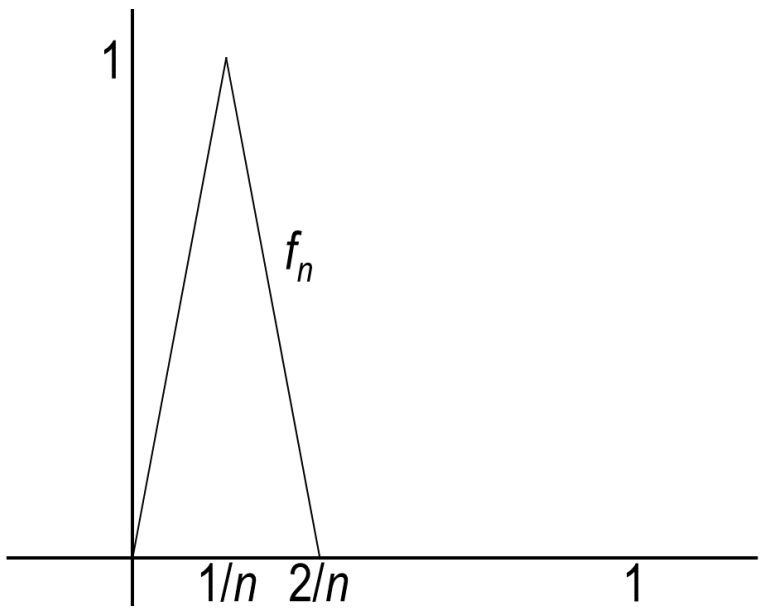
Propiedad análoga para todas las normas en $\mathbb{R}^{M\times N}$ (¡detallar!)

Ejercicio

Comprueba que la norma del máximo en $C([0,1])$ no es equivalente a la norma $\|\cdot\|_1$, definida para cada $f \in C([0,1])$ como

$$\|f\|_1 := \int_0^1 |f(x)|dx$$

(Indicación: para cada $n \geq 2$, considera la función f_n cuya gráfica es la poligonal que une los puntos $(0,0), (1/n,1), (2/n,0), (1,0)$)



Proposición

Sean $M, N \in \mathbb{N}$ y consideremos sendas normas en $\mathbb{R}^N$ y $\mathbb{R}^M$, que sin lugar a ambigüedad notaremos indiferentemente como $\|\cdot\|$. Entonces la aplicación que notaremos igualmente como $\|\cdot\|$

$$\|\mathbf{A}\| := \sup_{\mathbf{x}\in\mathbb{R}^N,\ \|\mathbf{x}\|=1} \|\mathbf{A}\mathbf{x}\|, \qquad (\mathbf{A} \in \mathbb{R}^{M\times N})$$

define una norma en $\mathbb{R}^{M\times N}$.

La demostración es inmediata y se propone como ejercicio

Norma *inducida* en $\mathbb{R}^{M\times N}$ (por las normas iniciales en $\mathbb{R}^N$ y $\mathbb{R}^M$)

En cualquier espacio normado finito dimensional la esfera unidad es compacta para la topología inducida por la norma (consecuencia del teorema de Heine–Borel), por lo que de hecho, el supremo anterior es máximo

Ejercicio

Comprueba que, con la notación de la proposición anterior, si $\mathbf{A} \in \mathbb{R}^{M\times N}$ entonces

$$\|\mathbf{A}\| = \sup_{\mathbf{x}\in\mathbb{R}^N,\ \mathbf{x}\neq\mathbf{0}} \frac{\|\mathbf{A}\mathbf{x}\|}{\|\mathbf{x}\|}.$$

En particular,

$$\mathbf{x} \in \mathbb{R}^N \ \Rightarrow\ \|\mathbf{A}\mathbf{x}\| \leq \|\mathbf{A}\|\|\mathbf{x}\|.$$

Ejemplo

$\|\cdot\|_1$ en $\mathbb{R}^N$ y $\mathbb{R}^M \rightsquigarrow$ norma $\|\cdot\|_1$ inducida en $\mathbb{R}^{M\times N}$

$$\mathbf{A} \in \mathbb{R}^{M\times N} \ \Rightarrow\ \|\mathbf{A}\|_1 = \max_{j=1,\ldots,N} \sum_{i=1}^{M} |a_{ij}|$$

$\|\cdot\|_\infty$ en $\mathbb{R}^N$ y $\mathbb{R}^M \rightsquigarrow$ norma $\|\cdot\|_\infty$ inducida en $\mathbb{R}^{M\times N}$

$$\mathbf{A} \in \mathbb{R}^{M\times N} \ \Rightarrow\ \|\mathbf{A}\|_\infty = \max_{i=1,\ldots,M} \sum_{j=1}^{N} |a_{ij}|$$

En efecto: $\|\cdot\|_\infty$ (la otra se deja como ejercicio)

$$\operatorname{sign}(a) := \begin{cases} 1, & \text{si } a \geq 0 \\ -1, & \text{si } a < 0 \end{cases}, \qquad (a \in \mathbb{R})$$

$$\left\| \left[\operatorname{sign}(a_{11}), \ldots, \operatorname{sign}(a_{1N})\right]^T \right\|_\infty = 1$$

$$\Downarrow$$

$$\begin{aligned} \|\mathbf{A}\|_\infty &\geq \left\| \mathbf{A} \begin{bmatrix} \operatorname{sign}(a_{11}) \\ \vdots \\ \operatorname{sign}(a_{1N}) \end{bmatrix} \right\|_\infty \\ &\geq \sum_{j=1}^{N} |a_{1j}| \end{aligned}$$

Igual con $\left[\operatorname{sign}(a_{i1}), \ldots, \operatorname{sign}(a_{iN})\right]^T$, $i = 2, \ldots, M$

$$\|\mathbf{A}\|_\infty \geq \max_{i=1,\ldots,M} \sum_{j=1}^{N} |a_{ij}|$$

$\mathbf{x} \in \mathbb{R}^N$, $\|\mathbf{x}\|_\infty = 1$

$$
\begin{aligned}
\|\mathbf{Ax}\|_\infty &= \left\| \begin{bmatrix} a_{11} & \cdots & a_{1N} \\ \vdots & & \vdots \\ a_{M1} & \cdots & a_{MN} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_N \end{bmatrix} \right\|_\infty \\
&= \left\| \left[\sum_{j=1}^{N} a_{1j} x_j, \ldots, \sum_{j=1}^{N} a_{Mj} x_j \right]^T \right\|_\infty \\
&= \max_{i=1,\ldots,M} \left| \sum_{j=1}^{N} a_{ij} x_j \right| \\
&\leq \max_{i=1,\ldots,M} \sum_{j=1}^{N} |a_{ij}||x_j| \\
&\leq \max_{i=1,\ldots,M} \sum_{j=1}^{N} |a_{ij}|
\end{aligned}
$$

$$\mathbf{A} \in \mathbb{R}^{M\times N} \Rightarrow \|\mathbf{A}\|_1 = \|\mathbf{A}^T\|_\infty$$

$\|\cdot\|_2$ no induce en $\mathbb{R}^{M\times N}$ Frobenius

$\mathbf{A} \in \mathbb{R}^{N\times N}$, *radio espectral* de $\mathbf{A}$

$$\rho(\mathbf{A}) := \text{máx}\{|\lambda| : \ \lambda \in \mathbb{C} \ \wedge \ \det(\mathbf{A} - \lambda\mathbf{I}) = 0\}$$

$\mathbf{x} \in \mathbb{R}^N$

$$\|\mathbf{x}\|_2 = \sqrt{\mathbf{x}^T\mathbf{x}}$$

Ejercicio

- Sea $\mathbf{A} \in \mathbb{R}^{N\times N}$ *semidefinida positiva*, esto es,

$$\mathbf{x} \in \mathbb{R}^N \;\Rightarrow\; \mathbf{x}^T\mathbf{A}\mathbf{x} \geq 0.$$

  Demuestra que si λ es un valor propio de A, entonces $\lambda \geq 0$.
- Prueba que si $\mathbf{P} \in \mathbb{R}^{N\times N}$ es una matriz ortogonal, entonces

$$\{\mathbf{x} \in \mathbb{R}^N :\; \|\mathbf{x}\|_2 = 1\} = \{\mathbf{P}^T\mathbf{x} :\; \mathbf{x} \in \mathbb{R}^N,\; \|\mathbf{x}\|_2 = 1\}.$$

- Demuestra que si $\lambda_1, \ldots, \lambda_N \geq 0$, entonces

$$\sup_{\mathbf{y}\in\mathbb{R}^N,\; \|\mathbf{y}\|_2=1} \sqrt{\sum_{i=1}^{N} \lambda_i y_i^2} = \sqrt{\max_{i=1,\ldots,N} \lambda_i}.$$

Proposición

Si $\mathbf{A} \in \mathbb{R}^{M\times N}$ entonces

$$\|\mathbf{A}\|_2 = \sqrt{\rho(\mathbf{A}^T\mathbf{A})}.$$

DEMOSTRACIÓN. $\mathbf{A}^T\mathbf{A}$ simétrica $\Rightarrow$ existe $\mathbf{P} \in \mathbb{R}^{N\times N}$ ortogonal:

$$\mathbf{P}^T\mathbf{A}^T\mathbf{A}\mathbf{P} = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \lambda_N \end{bmatrix}$$

$\lambda_1, \ldots, \lambda_N$ valores propios de $\mathbf{A}^T\mathbf{A}$, no negativos ($\mathbf{A}^T\mathbf{A}$ semidefinida positiva y ejercicio previo)

$$\mathbf{P} \text{ ortogonal } \Rightarrow \ \{\mathbf{x} \in \mathbb{R}^N : \ \|\mathbf{x}\|_2 = 1\} = \{\mathbf{P}^T\mathbf{x} : \ \mathbf{x} \in \mathbb{R}^N, \ \|\mathbf{x}\|_2 = 1\}$$

(ejercicio previo)

$$\mathbf{y} = \mathbf{P}^T\mathbf{x} \;\Leftrightarrow\; \mathbf{P}\mathbf{y} = \mathbf{x}$$

$$\begin{aligned}
\|\mathbf{A}\|_2 &= \sup_{\mathbf{x}\in\mathbb{R}^N,\ \|\mathbf{x}\|_2=1} \|\mathbf{A}\mathbf{x}\|_2 \\
&= \sup_{\mathbf{x}\in\mathbb{R}^N,\ \|\mathbf{x}\|_2=1} \sqrt{\mathbf{x}^T\mathbf{A}^T\mathbf{A}\mathbf{x}} \\
&= \sup_{\mathbf{y}\in\mathbb{R}^N,\ \|\mathbf{y}\|_2=1} \sqrt{\mathbf{y}^T\mathbf{P}^T\mathbf{A}^T\mathbf{A}\mathbf{P}\mathbf{y}} \\
&= \sup_{\mathbf{y}\in\mathbb{R}^N,\ \|\mathbf{y}\|_2=1} \sqrt{\sum_{i=1}^{N} \lambda_i y_i^2} \\
&= \sqrt{\max_{i=1,\ldots,N} \lambda_i} \qquad \text{(ejercicio previo)} \\
&= \sqrt{\rho(\mathbf{A}^T\mathbf{A})}
\end{aligned}$$

□

Definición

Una norma en $\mathbb{R}^{N\times N}$ se dice *matricial* cuando

$$\mathbf{A},\mathbf{B}\in\mathbb{R}^{N\times N} \;\Rightarrow\; \|\mathbf{AB}\|\leq\|\mathbf{A}\|\|\mathbf{B}\|.$$

En general, no toda norma en $\mathbb{R}^{N\times N}$ es matricial:

$$\|\mathbf{A}\| := \max_{i,j=1,\ldots,N} |a_{ij}|, \qquad (\mathbf{A}\in\mathbb{R}^{N\times N})$$

$$\mathbf{A}=\mathbf{B}=\begin{bmatrix}1 & 1\\ 1 & 1\end{bmatrix}$$

$$2=\|\mathbf{AB}\|>\|\mathbf{A}\|\|\mathbf{B}\|=1.$$

Norma en $\mathbb{R}^{N\times N}$ inducida por una norma en $\mathbb{R}^N$ es matricial (¡compruébalo!)

$a \in \mathbb{R}$

$$\lim_{n\to\infty} a^n = 0 \;\Leftrightarrow\; |a| < 1$$

$\mathbf{A} \in \mathbb{R}^{N\times N}$ ¿caracterización?

Si $\mathbf{A}$ diagonalizable, con valores propios $\lambda_1, \lambda_2, \ldots, \lambda_N \in \mathbb{R}$,

$$\mathbf{A} = \mathbf{P}\mathbf{D}\mathbf{P}^{-1},$$

$\mathbf{P} \in \mathbb{R}^{N\times N}$ regular y $\mathbf{D} = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \lambda_N \end{bmatrix}$

$$n \geq 1 \;\Rightarrow\; \mathbf{A}^n = \mathbf{P}\mathbf{D}^n\mathbf{P}^{-1},$$

$\mathbf{C} \in \mathbb{R}^{N\times N}$

$$\mathbf{X} \in \mathbb{R}^{N\times N} \mapsto \mathbf{C}\mathbf{X} \in \mathbb{R}^{N\times N}, \quad \mathbf{X} \in \mathbb{R}^{N\times N} \mapsto \mathbf{X}\mathbf{C} \in \mathbb{R}^{N\times N} \quad \text{continuas}$$

$$\lim_{n\to\infty} \mathbf{A}^n = \mathbf{0}$$
$$\Updownarrow$$
$$\lim_{n\to\infty} \mathbf{P}\mathbf{D}^n\mathbf{P}^{-1} = \mathbf{0}$$
$$\Updownarrow$$
$$\lim_{n\to\infty} \mathbf{D}^n = \mathbf{0}$$
$$\Updownarrow$$
$$\rho(\mathbf{A}) < 1$$

Teorema

Para una matriz cuadrada $\mathbf{A} \in \mathbb{R}^{N\times N}$ se cumple

$$\lim_{n\to\infty} \mathbf{A}^n = \mathbf{0} \;\Leftrightarrow\; \rho(\mathbf{A}) < 1.$$

DEMOSTRACIÓN. El caso en que $\mathbf{A}$ es diagonalizable se acaba de hacer. Para la prueba del caso general vid. [4, p. 14, Theorem 4]. □

Ejercicio

Sea $\mathbf{A} \in \mathbb{R}^{N\times N}$ una matriz triangular. Demuestra que

$$\lim_{n\to\infty} \mathbf{A}^n = \mathbf{0} \Leftrightarrow \max_{i=1,\ldots,N} |a_{ii}| < 1.$$

Corolario

Sean $N \geq 1$, $\mathbf{A} \in \mathbb{R}^{N\times N}$ y $\|\cdot\|$ una norma matricial en $\mathbb{R}^{N\times N}$, de forma que

$$\|\mathbf{A}\| < 1.$$

Entonces

$$\rho(\mathbf{A}) < 1.$$

DEMOSTRACIÓN. Se propone como ejercicio. Repárese en el hecho de que, por ser $\|\cdot\|$ una norma matricial,

$$n \geq 1 \;\Rightarrow\; \|\mathbf{A}^n\| \leq \|\mathbf{A}\|^n.$$

□

I.1.2. Problemas bien planteados. Estabilidad

Problema

Sean X e Y subconjuntos no vacíos de sendos espacios normados reales, $f : X \longrightarrow Y$ una aplicación, $y_0 \in Y$ y consideremos el siguiente problema:

$$\text{encontrar } x_0 \in X : \ f(x_0) = y_0. \tag{P}$$

solución x_0 resuelve el problema determinado por f y *datos* $y_0 \rightsquigarrow$ número, conjunto finitos de números (vector de $\mathbb{R}^N$, matriz), infinitos datos (función)

Ejemplo

$\mathbf{A} \in \mathbb{R}^{M\times N}$, $\mathbf{y} \in \mathbb{R}^M$,

determinar una solución del sistema de ecuaciones lineales cuya matriz de coeficientes sea $\mathbf{A}$ y su vector de términos independientes sea $\mathbf{y}$

$$X = \mathbb{R}^N, \ Y = \mathbb{R}^M, \ \mathbf{f}(\mathbf{x}) = \mathbf{A}\mathbf{x} = \mathbf{y}$$

Definición

El problema (P) está *bien planteado* cuando es *unisolvente* y *estable*:

- existe un único $x_0 \in X$ tal que $f(x_0) = y_0$ y
- x_0 depende continuamente de los datos y_0.

Matizaremos la estabilidad

Ejemplo

Problema que no está bien planteado (o *problema mal planteado*)

$$X := \mathbb{R}, \;\; Y := \mathbb{R}_+$$

y

$$f(x) = |x|, \qquad (x \in X)$$

Falla la unisolvencia: $y_0 = 1$ (lo mismo vale para cualquier $y_0 > 0$)

$$f(-1) = 1 = f(1)$$

(P) para todo $y \in Y$ unisolvente $\rightsquigarrow$ inversa de f, *resolvente*, g

Ejemplo

$$X := \mathbb{R}, \quad Y := \mathbb{R}_{++}$$

y

$$f(x) = e^x, \qquad (x \in X)$$

Unisolvente $\rightsquigarrow$ resolvente

$$g(y) = \log y, \qquad (y \in Y)$$

Estabilidad

- Intuitivamente $\rightsquigarrow$ a pequeñas perturbaciones de los datos y_0 corresponden pequeñas perturbaciones de la solución x_0
- ¡Con precisión!

Ejemplo

Consideremos el problema

$$X := [-1, 1] := Y$$

y

$$f(x) := \begin{cases} x, & \text{si } -1 < x < 1 \\ -x, & \text{si } x = \pm 1. \end{cases}$$

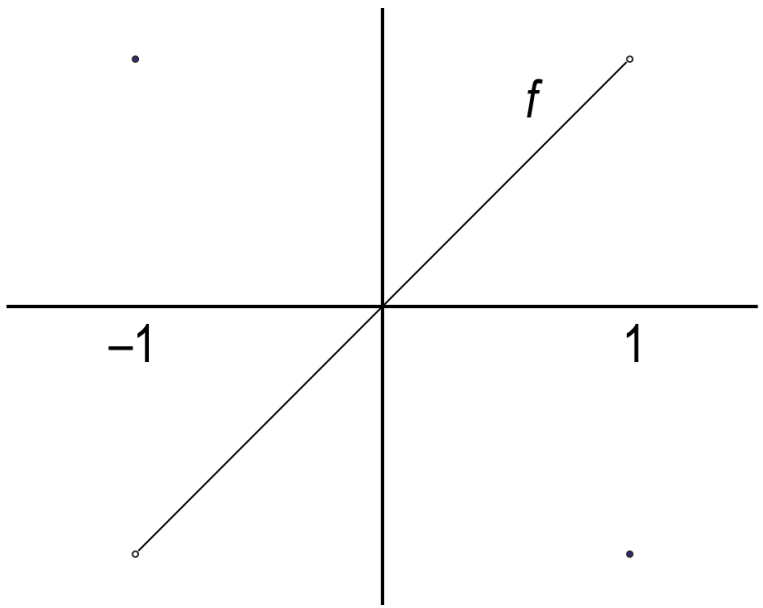
Unisolvente $\rightsquigarrow$ $f : [-1, 1] \longrightarrow [-1, 1]$ biyectiva, $g = f$

A pequeñas perturbaciones del dato $y_0 = -1$ (lo mismo con $y_0 = 1$) no corresponden pequeñas perturbaciones de $x_0 = 1$

$$y_n := -1 + \frac{1}{n} \to y_0$$

$$g(y_n) = -1 + \frac{1}{n} \to -1$$

$$|-1 - x_0| = 2$$



¿Estabilidad de (P) $\nleftrightarrow$ continuidad de g?

Ejemplo

Problema asociado $f : \mathbb{R}_+ \longrightarrow \mathbb{R}_+$

$$f(x) := \left(\frac{x}{10}\right)^{10}, (x \geq 0)$$

Unisolvente, resolvente $g : \mathbb{R}_+ \longrightarrow \mathbb{R}_+$

$$g(y) = 10y^{1/10}, \qquad (y \geq 0)$$

g continua

Pequeños cambios de los datos $\mathbf{y} \in \mathbb{R}_+ \rightsquigarrow$ cambios cercanos de las correspondientes soluciones $x \in \mathbb{R}_+$ pero no controlados, se aproximan a una velocidad diferente

$$y_0 = 0 = x_0$$

y dato próximo a y_0

$$y = 10^{-10}$$

solución correspondiente

$$x = 1$$

$$|y - y_0| = 10^{-10}$$

pero

$$|x - x_0| = 1$$

Estabilidad $\rightsquigarrow$ condición que fuerce un control de los valores de las soluciones en función de los datos, de forma que pequeñas perturbaciones de y_0 generen perturbaciones pequeñas y controladas de x_0

Formalización del concepto de estabilidad –en un contexto métrico se conoce como *lipschitzianidad local*–

Definición

Sean X y Y subconjuntos no vacíos de sendos espacios normados, $g : Y \longrightarrow X$ una aplicación e $y_0 \in Y$. Diremos que g es *estable* en y_0 cuando

$$\text{existen } \delta, M > 0 : \quad \sup_{y \in Y,\ 0 < \|y - y_0\| < \delta} \frac{\|g(y) - g(y_0)\|}{\|y - y_0\|} < M$$

y que g es *estable* si lo es en todos y cada uno de los elementos de Y.

Definición

El problema (P) es *estable en* $y_0 \in Y$ si su resolvente $g : Y \longrightarrow X$ lo es en dicho punto, y es *estable* si lo es en cualquier dato de Y.

$$g \text{ estable en } y_0 \quad \Rightarrow \quad \text{g continua en } y_0$$
$$\not\Leftarrow$$

(la implicación $\Rightarrow$ es inmediata y para $\not\Leftarrow$ puede tomarse la resolvente del ejemplo anterior en 0, o si se quiere, la función raíz cuadrada en 0)

Mejor comportamiento cuanto más pequeño sea M

Estimación de la estabilidad

$g : \mathbb{R} \longrightarrow \mathbb{R}$ clase C^1

g estable (Teorema del Valor Medio, ¡probar!)

Medida de la estabilidad de g en $y_0 \in \mathbb{R}$

$y_0 g(y_0) \neq 0 \rightsquigarrow$ cociente entre el error relativo cometido cerca de $g(y_0)$ y el error relativo de y_0

$$\left| \frac{\dfrac{g(y) - g(y_0)}{g(y_0)}}{\dfrac{y - y_0}{y_0}} \right| = \left| \frac{g(y) - g(y_0)}{y - y_0} \right| \left| \frac{y_0}{g(y_0)} \right|$$

$y_0 g(y_0) = 0 \rightsquigarrow$ errores absolutos cerca de $g(y_0)$ e y_0

$$\left| \frac{g(y) - g(y_0)}{y - y_0} \right|$$

De forma precisa

Definición

Dada una función $g \in C^1(\mathbb{R})$ y un punto $y_0 \in \mathbb{R}$, el *condicionamiento relativo* de g en y_0 viene dado por

$$c(g, y_0) := \left| \frac{g'(y_0) y_0}{g(y_0)} \right|,$$

siempre que $y_0 g(y_0) \neq 0$, y en caso contrario, el *condicionamiento absoluto* de g en y_0 es

$$C(g, y_0) := |g'(y_0)|.$$

Ídem funciones reales de variable real definidas en intervalos de $\mathbb{R}$ y de clase C^1

Extensión a aplicaciones entre espacios normados de clase C^1

$\mathbf{g} \in C^1(\mathbb{R}^N, \mathbb{R}^M)$, $\mathbf{g} = \begin{bmatrix} g_1 & \cdots & g_M \end{bmatrix}$,

$$\frac{\partial \mathbf{g}}{\partial \mathbf{y}}(\mathbf{y}_0) = \begin{bmatrix} \frac{\partial g_1}{\partial y_1}(\mathbf{y_0}) & \frac{\partial g_1}{\partial y_2}(\mathbf{y_0}) & \cdots & \frac{\partial g_1}{\partial y_N}(\mathbf{y_0}) \\ \frac{\partial g_2}{\partial y_1}(\mathbf{y_0}) & \frac{\partial g_2}{\partial y_2}(\mathbf{y_0}) & \cdots & \frac{\partial g_2}{\partial y_N}(\mathbf{y_0}) \\ \vdots & \vdots & & \vdots \\ \frac{\partial g_M}{\partial y_1}(\mathbf{y_0}) & \frac{\partial g_M}{\partial y_2}(\mathbf{y_0}) & \cdots & \frac{\partial g_M}{\partial y_N}(\mathbf{y_0}) \end{bmatrix} \in \mathbb{R}^{M \times N}$$

$$c(\mathbf{g}, \mathbf{y}_0) := \frac{\left\| \frac{\partial \mathbf{g}}{\partial \mathbf{y}}(\mathbf{y}_0) \right\| \|\mathbf{y}_0\|}{\|\mathbf{g}(\mathbf{y}_0)\|}, \qquad C(\mathbf{g}, \mathbf{y}_0) := \left\| \frac{\partial \mathbf{g}}{\partial \mathbf{y}}(\mathbf{y}_0) \right\|$$

Defínción

Si en el problema (P) la resolvente es de clase C^1 e $y_0 \in Y$, el *condicionamiento relativo o absoluto* de (P) son los de su resolvente en dicho punto.

Observación

Aunque se trata de un concepto bastante ambiguo, suele decirse que una aplicación (o un problema) está *bien condicionado* en un punto si su condicionamiento es pequeño y en caso contrario *mal condicionado*.

Ejemplo

$f : (0,1) \longrightarrow (0,\pi/2)$

$$f(x) := \text{arcsen}x$$

Bien planteado, resolvente $g : (0,\pi/2) \longrightarrow (0,1)$

$$g(y) := \text{sen}y$$

Condicionamiento (relativo): $y_0 \in (0,\pi/2)$

$$\begin{aligned} c(g,y_0) &= \frac{|g'(y_0)y_0|}{|g(y_0)|} \\ &= y_0 \frac{\cos y_0}{\text{sen}y_0} \end{aligned}$$

$$c(g,y_0) \in (0,1) \Rightarrow \text{ bien condicionado}$$

Ejemplo

$\mathbf{A} \in \mathbb{R}^{N\times N}$ regular, $\mathbf{y} \in \mathbb{R}^N$

$$\text{encontrar } \mathbf{x} \in \mathbb{R}^N : \ \mathbf{f}(\mathbf{x}) = \mathbf{y}$$

$$\mathbf{f}(\mathbf{x}) = \mathbf{A}\mathbf{x}$$

Unisolvente, resolvente $\mathbf{g} : \mathbb{R}^N \longrightarrow \mathbb{R}^N$

$$\mathbf{g}(\mathbf{y}) := \mathbf{A}^{-1}\mathbf{y}, \qquad (\mathbf{y} \in \mathbb{R}^N)$$

Estable en todo $\mathbf{y}_0 \in \mathbb{R}^N$

$$\mathbf{y}_0, \mathbf{y} \in \mathbb{R}^N \ \Rightarrow \ \|\mathbf{g}(\mathbf{y}) - \mathbf{g}(\mathbf{y}_0)\|_\infty \leq \|\mathbf{A}^{-1}\|_\infty \|\mathbf{y} - \mathbf{y}_0\|_\infty$$

$$c(\mathbf{g}, \mathbf{y}_0) = \frac{\left\| \frac{\partial \mathbf{g}}{\partial \mathbf{y}}(\mathbf{y}_0) \right\| \|\mathbf{y}_0\|}{\|\mathbf{g}(\mathbf{y}_0)\|}$$

$$= \frac{\|\mathbf{A}^{-1}\| \|\mathbf{y}_0\|}{\|\mathbf{A}^{-1}\mathbf{y}_0\|}$$

$$= \frac{\|\mathbf{A}^{-1}\| \|\mathbf{y}_0\|}{\|\mathbf{x}_0\|}$$

$\mathbf{A}\mathbf{x}_0 = \mathbf{y}_0$
$\|\mathbf{A}^{-1}\|$ grande, $\mathbf{x}_0$ e $\mathbf{y}_0$ mismo orden $\Rightarrow$ condicionamiento grande

$$\mathbf{A} = \begin{bmatrix} 1 & 1 \\ 1 & 0.999 \end{bmatrix}$$

dato y solución

$$\mathbf{y}_0 = \begin{bmatrix} 2 \\ 1.999 \end{bmatrix}, \quad \mathbf{x}_0 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

$$\mathbf{A}^{-1} = \begin{bmatrix} -999 & 1000 \\ 1000 & -1000 \end{bmatrix}$$

$$\|\mathbf{A}^{-1}\|_\infty = 2000, \quad \|\mathbf{x}_0\|_\infty = 1, \quad \|\mathbf{y}_0\|_\infty = 2$$

$$c(g, \mathbf{y}_0) = 4000$$

Mal condicionamiento: dato $\mathbf{y}$ próximo a $\mathbf{y}_0$

$$\mathbf{y} = \begin{bmatrix} 2 \\ 1.998 \end{bmatrix}$$

su solución

$$\mathbf{x} = \begin{bmatrix} 0 \\ 2 \end{bmatrix}$$

$$\|\mathbf{y} - \mathbf{y}_0\|_\infty = 0.001, \qquad \|\mathbf{x} - \mathbf{x}_0\|_\infty = 1$$

Interpretación geométrica: las columnas de $\mathbf{A}$ forman base de $\mathbb{R}^{2\times 2}$, calculamos las coordenadas de $\mathbf{y}$ e $\mathbf{y}_0$ en dicha base. Un pequeño cambio entre $\mathbf{y}$ e $\mathbf{y}_0$ genera un cambio grande de coordenadas por ser los vectores de la base casi paralelos.

¿Está globalmente acotado el condicionamiento relativo de un sistema de ecuaciones lineales? Respuesta afirmativa: idea cómo de estable es

Definición

Si $\|\cdot\|$ denota la norma matricial en $\mathbb{R}^{N\times N}$ inducida por una norma en $\mathbb{R}^N$, que notaremos igualmente como $\|\cdot\|$, entonces definimos el *condicionamiento* $c(\mathbf{A})$ de una matriz regular $\mathbf{A}\in\mathbb{R}^{N\times N}$ como

$$c(\mathbf{A}) := \|\mathbf{A}^{-1}\|\|\mathbf{A}\|.$$

Motivación

$$\begin{aligned}\sup_{\mathbf{y}_0\neq\mathbf{0}} c(g,\mathbf{y}_0) &= \sup_{\mathbf{y}_0\neq\mathbf{0}} \frac{\|\mathbf{A}^{-1}\|\|\mathbf{y}_0\|}{\|\mathbf{A}^{-1}\mathbf{y}_0\|}\\ &= \sup_{\mathbf{x}_0\neq\mathbf{0}} \frac{\|\mathbf{A}^{-1}\|\|\mathbf{A}\mathbf{x}_0\|}{\|\mathbf{x}_0\|}\\ &= \|\mathbf{A}^{-1}\|\|\mathbf{A}\|\end{aligned}$$

Mal condicionamiento del sistema $\rightsquigarrow$ condicionamiento de la matriz de coeficientes grande (y viceversa)

$$c(\mathbf{A}) \geq 1$$

Ejemplo anterior, norma del máximo en $\mathbb{R}^2$ y la correspondiente norma matricial inducida

$$c(\mathbf{A}) = 4000$$

I.1.3. Algoritmos. Algoritmo PageRank de Google

Algoritmo

Procedimiento que describe de forma precisa, y siempre mediante un número finito de operaciones artiméticas y lógicas elementales, la resolución de un problema.

El algoritmo recoge las instrucciones que permiten al ejecutor del mismo resolver completamente el problema. El ejecutor suele ser un ordenador y, de hecho, en la mayoría de los casos, no puede ser una persona.

Análisis Numérico

Se ocupa de diseñar algoritmos que permitan la resolución efectiva de problemas bien planteados y que involucran números reales.

Complejidad de un algoritmo

Medida del tiempo de ejecución y que suele expresarse en términos de un parámetro asociado al problema.

Ejemplo

Número de operaciones aritméticas elementales que involucra el cálculo del producto de dos matrices cuadradas del mismo orden en función de dicho orden (¿qué número es?).

Complejidad vs. *Eficiencia*

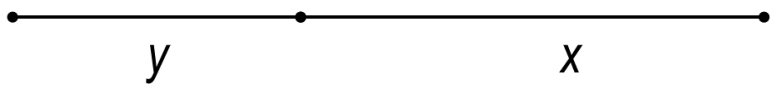
Desarrollo y análisis de un algoritmo $\rightsquigarrow$ *precisión*, *establidad* y *efectos de la representación finita de los números reales*

Ejemplo

Libro IV de *Los Elementos* de Euclides, Definición 3: "Se dice que un segmento está dividido en media y extrema razón cuando el segmento total es a la parte mayor como la parte mayor es a la menor"

No es un algoritmo

proporción o razón aúrea, *número de oro*, Φ Fideas



$$\frac{x+y}{x} = \frac{x}{y}$$

$$\Phi = \frac{x}{y}$$

$$1 + \frac{1}{\Phi} = \Phi \Leftrightarrow \Phi^2 - \Phi - 1 = 0$$

$$x^2 - x - 1 = 0 \Leftrightarrow \frac{1 \pm \sqrt{5}}{2}$$

$$\Phi = \frac{1+\sqrt{5}}{2}$$

Es un algoritmo

Referente estético en Arquitectura (Pirámide de Keops, Partenón, Alhambra, Leonardo, Le Corbusier...) y otras Artes, diseño de objetos cotidianos (tarjetas de crédito), o patrón recursivo en la Naturaleza

Problemas de la vida real $\rightsquigarrow$ *Modelo* $\rightsquigarrow$ *Problema matemático*

Algoritmo PageRank de Google

Medir *relevancia* de páginas web (dominios...) con enlaces en común

S. Brin y L. Page –fundadores de Google–, 1995 (véanse [1] y [2])

Desde el 7 de marzo de 2016 Google no muestra el PageRank público en ninguna herramienta

Idea relevancia de una página (P)

- número de enlaces de otras páginas a (P)
- importancia de las páginas que establecen enlace con (P)

Modelo

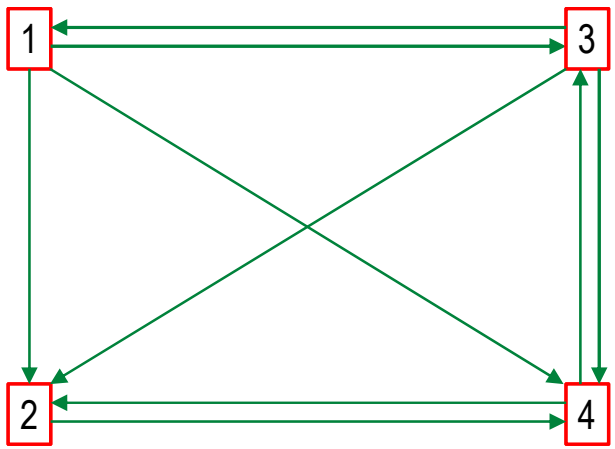
Conjunto (finito) de páginas web $1, 2 \ldots, x_N$

página $i \rightsquigarrow x_i \geq 0$ relevancia

página i más importante que página j si $x_i > x_j$

Ejemplo

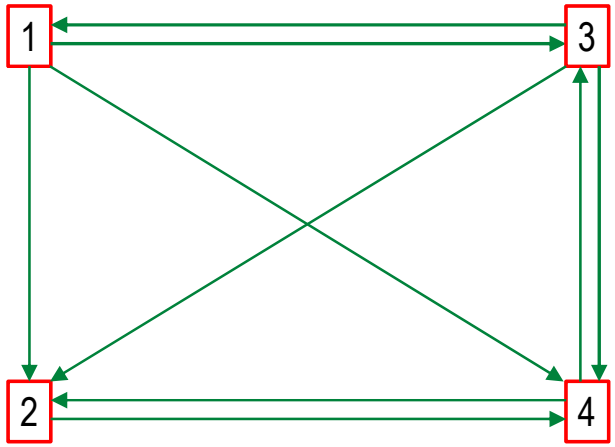
Consideremos la estructura de enlaces entre cuatro páginas web



Posibilidad 1: $x_i \rightsquigarrow$ número de enlaces que recibe la página i

$$x_1 = 1, \quad x_2 = 3, \quad x_3 = 2, \quad x_4 = 3$$

Desventaja: no se tiene en cuenta la relevancia de la página que establece un enlace con una dada



Posibilidad 2: contemplar la importancia de las páginas con enlaces entrantes
$x_i \rightsquigarrow$ suma de la relevancia de las páginas que establecen un enlace con la página i

$$x_1 = x_3, \quad x_2 = x_1 + x_3 + x_4, \quad x_3 = x_1 + x_4, \quad x_4 = x_1 + x_2 + x_3$$

Desventaja: no influye el número de enlaces de cada página, con lo que añadiendo más enlaces a una página sería posible aumentar su importancia. ¿Cómo evitarlo?

Dividir cada valor de relevancia x_i por el número de enlaces que salen de la página i

Problema matemático

$$x_1 = \frac{x_3}{3}, \quad x_2 = \frac{x_1}{3} + \frac{x_3}{3} + \frac{x_4}{2}, \quad x_3 = \frac{x_1}{3} + \frac{x_4}{2}, \quad x_4 = \frac{x_1}{3} + x_2 + \frac{x_3}{3}$$

Solución normalizada (valor máximo 10):

$$x_1 = 1.875, \quad x_2 = 7.5, \quad x_3 = 5.625, \quad x_4 = 10$$

Versatilidad de las matemáticas para modelar problemas de la vida real

Necesidad de disponer de *métodos eficientes para resolver problemas matemáticos* (en el algoritmo PageRank, sistemas de ecuaciones lineales)

Solución de un problema $\rightsquigarrow$ *límite de una sucesión*

Ejemplo

sucesión de Fibonacci, Leonardo de Pisa, Fibonacci, (1170–1250)

$$x_1 := 1 =: x_2$$

y, para todo $n \in \mathbb{N}$

$$x_{n+2} := x_{n+1} + x_n$$

$$1, 1, 2, 3, 5, 8, 13, 21, 34, 45 \ldots$$

Explícitamente vs. recursivamente

$$x_n = \frac{1}{\sqrt{5}} \left(\Phi^n - \left(\frac{1-\sqrt{5}}{2} \right)^n \right)$$

(siguiente ejercicio)

$$\frac{x_{n+1}}{x_n} = \frac{\Phi^{n+1} - \left(\frac{1-\sqrt{5}}{2}\right)^{n+1}}{\Phi^n - \left(\frac{1-\sqrt{5}}{2}\right)^n}$$

$$= \frac{\Phi - \Phi\left(\frac{1-\sqrt{5}}{2\Phi}\right)^{n+1}}{1 - \left(\frac{1-\sqrt{5}}{2\Phi}\right)^n}$$

$$\left|\frac{1-\sqrt{5}}{2\Phi}\right| < 1 \Rightarrow \lim_{n\to\infty} \frac{x_{n+1}}{x_n} = \Phi$$

Cocientes simples

$$\frac{1}{1}, \frac{2}{1}, \frac{3}{2}, \frac{5}{3}, \frac{8}{5}, \frac{13}{8}, \frac{21}{13}, \frac{34}{21}, \frac{45}{34}$$

Errores absolutos (aproximados)

$$0.61803, 0.38197, 0.11803, 0.04863, 0.01803, 0.00697, 0.00265, 0.00101, 0.00039$$

Ejercicio

Demuestra que los términos de la sucesión de Fibonacci $\{x_n\}_{n\in\mathbb{N}}$ vienen dados por

$$x_n = \frac{1}{\sqrt{5}}\left(\Phi^n - \Psi^n\right), \qquad (n \in \mathbb{N}),$$

donde

$$\Psi := \frac{1-\sqrt{5}}{2}.$$

Indicación: hay varias posibilidades. Una es proceder inductivamente para probar sucesivamente

- $n \geq 1 \;\Rightarrow\; \Phi^{n+1} = \Phi^n + \Phi^{n-1}$ y $\Psi^{n+1} = \Psi^n + \Psi^{n-1}$

y, a partir de ello y de la relación de recurrencia,

- $n \geq 1 \;\Rightarrow\; x_n = \left(\Phi^n - \Psi^n\right)/\sqrt{5}$.

Otra, reescribir la relación de recurrencia y los datos iniciales $x_1 = 1 = x_2$ como

$$\begin{bmatrix} x_{n+2} \\ x_{n+1} \end{bmatrix} = \mathbf{A} \begin{bmatrix} x_{n+1} \\ x_n \end{bmatrix}, \quad \begin{bmatrix} x_2 \\ x_1 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix},$$

con

$$\mathbf{A} = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix}.$$

Deducir entonces que

$$\begin{bmatrix} x_{n+2} \\ x_{n+1} \end{bmatrix} = \mathbf{A}^n \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

y diagonalizar la matriz $\mathbf{A}$ para calcular $\mathbf{A}^n$ y de esa forma inferir la expresión explícita de x_n.

I.2. Errores de redondeo. Iteradores

Fuentes usuales de error

- Problema de la vida real mediante un modelo matemático: errores derivados de la eleción del propio modelo y, si procede, de la medida de datos experimentales.
- *Errores de truncatura*: la solución del problema es el límite de una sucesión y se reemplaza dicho límite por un término de la sucesión (*iterador*).
- Los números reales son entes abstractos –axioma del supremo, intervalo abierto–, pero los ordenadores trabajan con los llamados *números máquina*, que en definitiva son el subconjunto finito de números reales que reconoce un ordenador concreto. La aproximación de un número real a su correspondiente número máquina genera un *error de redondeo* y consecuentemente, al operar con números máquina, también se produce una *propagación del error*.

I.2.1. Sistema posicional y números máquina

Los ordenadores trabajan con un subconjunto finito de números reales, los *números máquina*, subconjunto que depende de las especificaciones del ordenador

Sistemas posicionales de numeración

Base $b \in \mathbb{N}$, $b = 2$ o $b = 10$, estándar ISO/IEC/IEEE60559:2011 del IEEE (Institute of Electrical and Electronic Engineers, www.ieee.org)

Bases *binaria* o *decimal* ($b = 2$ o $b = 10$)
$s \in \{0, 1\}$
$N, M \in \mathbb{N} \cup \{0\}$
para todo $k = -M, \ldots, N$, $0 \leq x_k < b$

Representación posicional del número real $x = (-1)^s \sum_{n=-M}^{N} x_n b^n$

$$x_b := (-1)^s \cdot (x_N \ldots x_1 x_0 . x_{-1} x_{-2} \ldots x_{-M})_b$$

Punto entre x_0 y x_{-1} *punto binario* o *punto decimal*, $b = 2$ o $b = 10$
s determina el signo
Base decimal sin subíndice

Ejemplo

$x_{10} = x = 101.11$

$$x = 1 \cdot 10^2 + 1 \cdot 10^0 + 1 \cdot 10^{-1} + 1 \cdot 10^{-2}$$

$y_2 = (101.11)_2$

$$y = 1 \cdot 2^2 + 1 \cdot 2^0 + 1 \cdot 2^{-1} + 1 \cdot 2^{-2}$$

$b = 16$ ordenadores
$b = 60$ personas en Mesopotamia $\rightsquigarrow$ gestionar economía + Análisis Numérico

$$\sqrt{2} \approx 1 + \frac{24}{60} + \frac{51}{60^2} + \frac{10}{60^3} \qquad \text{(¡hace casi 4000 años!)}$$

error absoluto menor que $6 \cdot 10^{-7}$

Observación

Salvo indicación expresa, siempre supondremos b = 2 o b = 10

Observación

base 2 $\rightsquigarrow$ base 10: inmediato

base 10 $\rightsquigarrow$ base 2:

- enteros: algoritmo de Euclides
- decimales: multiplicar sucesivamente por adecuadas potencias de $\frac{1}{2}$

Ejemplo

- $$\begin{aligned} -(10.101)_2 &= -2 - \frac{1}{2} - \frac{1}{2^3} \\ &= -2.625 \end{aligned}$$

- $7.5625 = (111.1001)_2$:
  - $$\begin{aligned} 7 &= 3 \cdot 2 + 1 \\ &= (1 \cdot 2 + 1) \cdot 2 + 1 \\ &= 2^2 + 2 + 2^0 \\ &= (111)_2 \end{aligned}$$
  - $$\begin{aligned} 0.5625 &= \frac{5625}{10000} \\ &= \frac{9}{16} \\ &= \frac{1}{2}\left(\frac{18}{16}\right) \\ &= \frac{1}{2}\left(1 + \frac{2}{16}\right) \\ &= \frac{1}{2}\left(1 + \frac{1}{8}\right) \\ &= \frac{1}{2} + \frac{1}{2^4} \\ &= (0.1001)_2 \end{aligned}$$

Variando $M, N \in \mathbb{N} \cup \{0\} \rightsquigarrow$ subconjunto denso de $\mathbb{R}$
Convergencia serie geométrica de razón menor que 1

Extensión

$$x = (-1)^s \sum_{n=-\infty}^{N} x_n b^n,$$

$$x_b := (-1)^s \cdot (x_N \dots x_1 x_0 . x_{-1} x_{-2} \dots)_b$$

Representación posicional finita o infinita $\rightsquigarrow$ base

Ejemplo

- Representación posicional binaria finita $\Rightarrow$ representación posicional decimal finita (fácil)
- Representación posicional infinita en las dos bases

$$\begin{aligned}(0.\overline{001})_2 = & \sum_{n=1}^{\infty} \left(2^{-3}\right)^n \\ & = \frac{1}{7} \\ & = 0.\overline{142857}\end{aligned}$$

(Recuerda

$$\alpha \neq 1,\ n > k \geq 0 \ \Rightarrow \ \sum_{j=k}^{n} \alpha^j = \frac{\alpha^k - \alpha^{n+1}}{1-\alpha}.$$

En particular, si $|\alpha| < 1$, entonces

$$\sum_{j=k}^{\infty} \alpha^k = \frac{\alpha^k}{1-\alpha})$$

- Representación posicional decimal finita y binaria infinita

$$\begin{aligned} 0.2 &= \frac{1}{5} \\ &= (0.\overline{0011})_2, \end{aligned}$$

pues

$$\begin{aligned} \frac{1}{5} &= \frac{1}{2^3}\left(\frac{2^3}{5}\right) \\ &= \frac{1}{2^3}\left(1+\frac{3}{5}\right) \\ &= \frac{1}{2^3}\left(1+\frac{1}{2}\left(1+\frac{1}{5}\right)\right) \\ &= \frac{1}{2^3}+\frac{1}{2^4}+\frac{1}{2^4}\left(\frac{1}{5}\right) \end{aligned}$$

Números máquina, ordenador base b, N posiciones de memoria, cantidad finita de números reales

- *Representación con punto fijo*

$$(-1)^s b^{-k} \sum_{n=0}^{N-2} a_n b^n \quad \leftrightsquigarrow \quad (-1)^s \cdot (a_{N-2} \dots a_k . a_{k-1} \dots a_0)_b$$

k natural fijo, $N-k-1$ dígitos enteros y k dígitos tras el punto a_n, $0 \leq a_n \leq b-1$
N posiciones de memoria: signo (1), cifras significativas $(N-1)$

- *Representación con punto flotante*

$$(-1)^s b^e \sum_{n=1}^{t} a_n b^{-n} \quad \leftrightsquigarrow \quad (-1)^s \cdot (0.a_1 \dots a_t) \cdot b^e = (-1)^s \cdot m \cdot b^{e-t}$$

$t \in \mathbb{N}$ el número máximo de dígitos o *cifras significativas* a_n, $0 \leq a_n \leq b-1$,
$m = a_1 \dots a_t$ *mantisa* , $0 \leq m \leq b^t - 1$
$e \in \mathbb{Z}$ *exponente*, $L \leq e \leq U$, $L, U \in \mathbb{Z}$, $L \leq U$

N posiciones de memoria: signo (1), cifras significativas (t) y dígitos del exponente $(N-t-1)$

Ejemplo

$x = -3.4567,\ b = 10$

- Representación con punto fijo: $k = 4,\ N = 6$

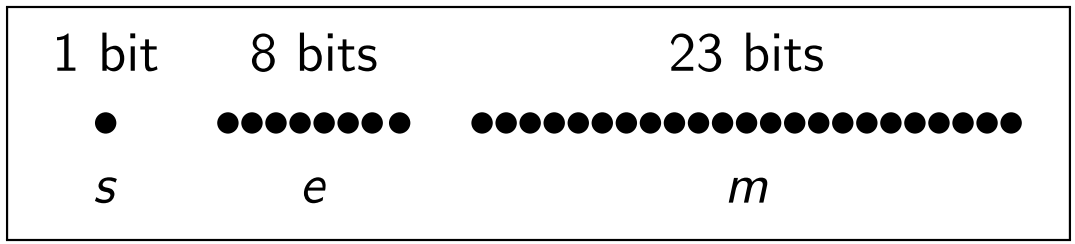
$$\begin{aligned} x &= (-1)10^{-4}(3 \cdot 10^4 + 4 \cdot 10^3 + 5 \cdot 10^2 + 6 \cdot 10 + 7 \cdot 10^0) \\ &= (-1)(3.4567) \end{aligned}$$

- Representación con punto flotante: $t = 6,\ e = 2$

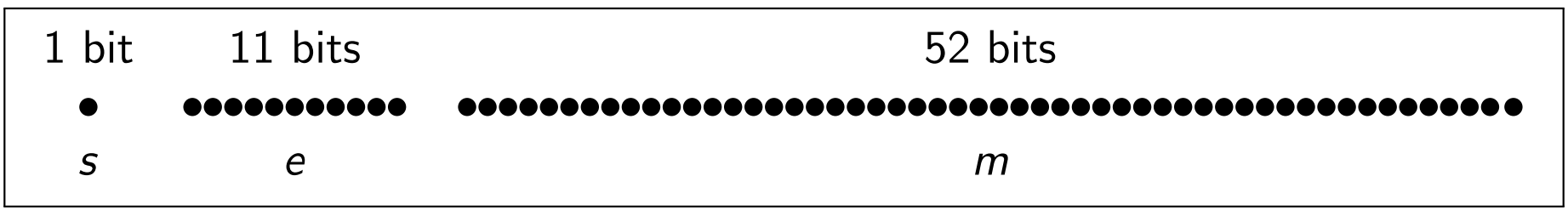
$$\begin{aligned} x &= (-1)10^2(0 \cdot 10^{-1} + 3 \cdot 10^{-2} + 4 \cdot 10^{-3} + 5 \cdot 10^{-4} + 6 \cdot 10^{-5} + 7 \cdot 10^{-6}) \\ &= (-1)(0.034567) \cdot 10^2 \end{aligned}$$

Usualmente 2 representaciones de punto flotante: *precisión simple* y *doble*

Estánadar ISO/IEC/IEEE60559:2011, representación binaria, $N = 32$ bits (binary digits), precisión simple



Estánadar ISO/IEC/IEEE60559:2011, representación binaria, $N = 64$ bits, precisión doble



Unicidad: normalización $a_1 \neq 0$ (a_1 *cifra significativa principal*)

$b = 2$, $t = 3$, $L = 1$, $U = 3$, sin normalización $\rightsquigarrow$ varias representaciones con punto flotante para 1

$$(0.100) \cdot 2^1 = (0.010) \cdot 2^2 = (0.001) \cdot 2^3$$

Notación sistema (normalizado) de punto flotante

$$\mathbb{F}(b,t,L,U) := \{0\} \cup \left\{ (-1)^s b^e \sum_{n=1}^{t} a_n b^{-n} : \ s = 0,1, \ a_1 \neq 0, \right.$$
$$\left. 0 \leq a_1, \ldots, a_t \leq b-1, L \leq e \leq U \right\}$$

Proposición

Sean $t \in \mathbb{N}$, $L, U \in \mathbb{Z}$ con $L \leq U$ y $x \in \mathbb{F}(b, t, L, U)$. Entonces

(i) $-x \in \mathbb{F}(b, t, L, U)$.

(ii) $b^{L-1} \leq |x| \leq b^U(1 - b^{-t})$.

(iii) $\text{card}(\mathbb{F}(b, t, L, U)) = 2(b-1)b^{t-1}(U - L + 1) + 1$.

DEMOSTRACIÓN. Se trata de una simple comprobación que se deja como ejercicio. □

(iii) Conjunto de números máquina finito
(ii) No solo acotación, cotas inferior y superior se alcanzan

Ejemplo

Números estrictamente positivos del sistema de punto flotante $\mathbb{F}(2, 4, -1, 1)$

$$(0.1111) \cdot 2 = \frac{15}{8} \qquad (0.1110) \cdot 2 = \frac{7}{4} \qquad (0.1101) \cdot 2 = \frac{13}{8} \qquad (0.1100) \cdot 2 = \frac{3}{2}$$

$$(0.1011) \cdot 2 = \frac{11}{8} \qquad (0.1010) \cdot 2 = \frac{5}{4} \qquad (0.1001) \cdot 2 = \frac{9}{8} \qquad (0.1000) \cdot 2 = 1$$

$$(0.1111) \cdot 2^0 = \frac{15}{16} \qquad (0.1110) \cdot 2^0 = \frac{7}{8} \qquad (0.1101) \cdot 2^0 = \frac{13}{16} \qquad (0.1100) \cdot 2^0 = \frac{3}{4}$$

$$(0.1011) \cdot 2^0 = \frac{11}{16} \qquad (0.1010) \cdot 2^0 = \frac{5}{8} \qquad (0.1001) \cdot 2^0 = \frac{9}{16} \qquad (0.1000) \cdot 2^0 = \frac{1}{2}$$

$$(0.1111) \cdot 2^{-1} = \frac{15}{32} \qquad (0.1110) \cdot 2^{-1} = \frac{7}{16} \qquad (0.1101) \cdot 2^{-1} = \frac{13}{32} \qquad (0.1100) \cdot 2^{-1} = \frac{3}{8}$$

$$(0.1011) \cdot 2^{-1} = \frac{11}{32} \qquad (0.1010) \cdot 2^{-1} = \frac{5}{16} \qquad (0.1001) \cdot 2^{-1} = \frac{9}{32} \qquad (0.1000) \cdot 2^{-1} = \frac{1}{4}$$

Número de elementos de $\mathbb{F}(2, 4, -1, 1)$ (positivos, negativos y cero)

$$2(b-1)b^{t-1}(U-L+1)+1 = 49$$

Valores mínimo y máximo (positivos): $b^{L-1} = \frac{1}{4}, \quad b^{U}(1-b^{-t}) = \frac{15}{8}$

$\mathbb{F}(b, t, L, U)$ no se distribuye uniformemente, aunque sí por bloques (vid. Relación de Ejercicios)

Definición

Para un sistema de punto flotante $\mathbb{F}(b, t, L, U)$ con $L \leq 1 \leq U$, el *épsilon máquina*, que se escribe como ε_M, es la distancia entre el menor número de $\mathbb{F}(b, t, L, U)$ mayor que 1 y la propia unidad, es decir,

$$\varepsilon_M := b^{1-t}.$$

$L \leq 1 \leq U \ \Rightarrow \ 1 \in \mathbb{F}(b, t, L, U)$

Estándar IEEE, ISO/IEC/IEEE60559:2011

Formatos de números con punto flotante más habituales: binary32 y binary64, inalterados desde el estándar IEEE754 de 1985

Magnitudes cuánticas 10^{-15} m versus universo observable 10^{26} m

binary16 / half precission	$\mathbb{F}(2, 10, -13, 16)$
binary32 / single precission	$\mathbb{F}(2, 23, -125, 128)$
binary64 / double precission	$\mathbb{F}(2, 52, -1021, 1024)$
binary128 /quadruple precission	$\mathbb{F}(2, 112, -16381, 16384)$
decimal32	$\mathbb{F}(10, 6, -94, 97)$
decimal64	$\mathbb{F}(10, 15, -382, 385)$
decimal128	$\mathbb{F}(10, 33, -6142, 6144)$

I.2.2. Redondeo en sistemas de punto flotante y su artimética

Números máquina, sistema de punto flotante $\mathbb{F}(b, t, L, U)$

- $\mathbb{F}(b, t, L, U) \neq \mathbb{R}$
- El resultado de operar con dos números de $\mathbb{F}(b, t, L, U)$ no queda dentro de dicho sistema necesariamente

  $1/4, 9/32 \in \mathbb{F}(2, 4, -1, 1)$ pero $1/4 + 9/32 = 17/32 \notin \mathbb{F}(2, 4, -1, 1)$

Definición

Fijado un sistema de punto flotante concreto $\mathbb{F}(b, t, L, U)$ –al que no se hará referencia si no hay lugar a ambigüedad– si $x \in \mathbb{R}$ es el número real

$$x = (-1)^s b^e \sum_{n=1}^{\infty} a_n b^{-n}$$

entonces su *truncatura* (en dicho sistema) es el número de $\mathbb{F}(b, t, L, U)$

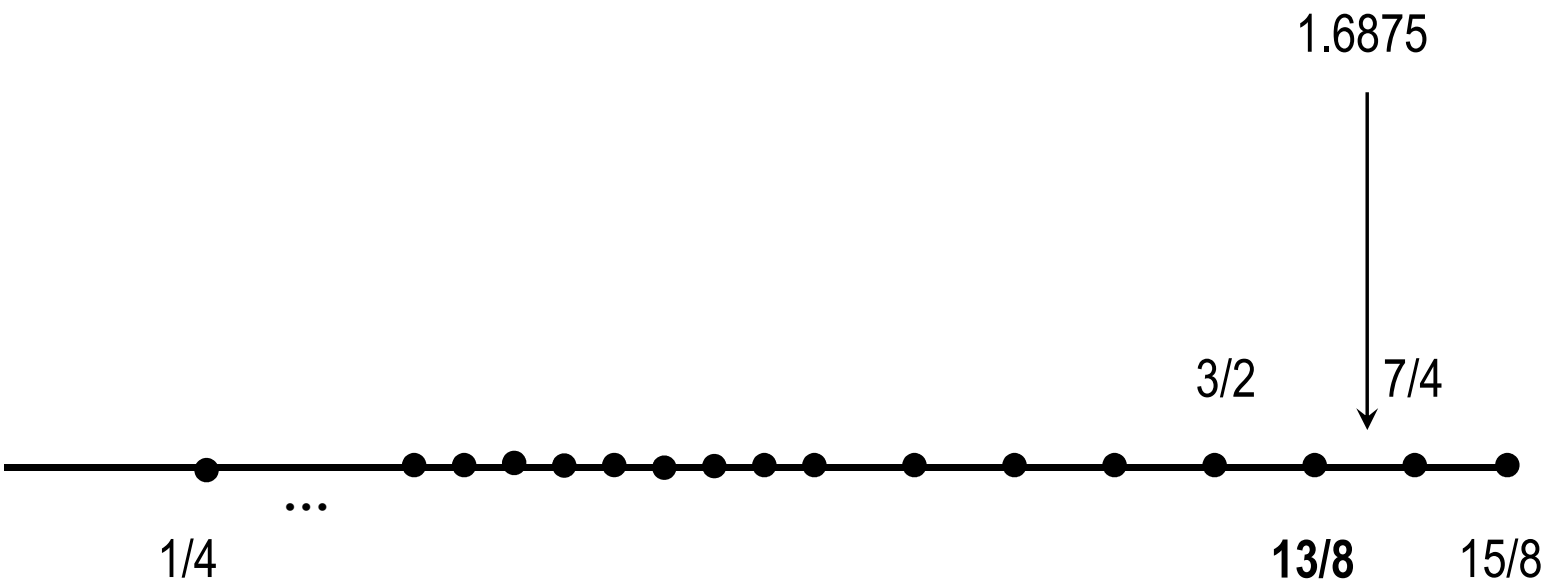
$$\mathrm{tr}(x) := (-1)^s \cdot (0.a_1 \ldots a_t) \cdot b^e$$

Ejemplo

$\mathbb{F}(2, 4, -1, 1)$

$$1.6875 = (0.11011) \cdot 2$$

$$\mathrm{tr}(1.6875) = (0.1101) \cdot 2 = \frac{13}{8} = 1.625$$



Definición

Para un sistema de punto flotante $\mathbb{F}(b, t, L, U)$, el *redondeo* del número real

$$x = (-1)^s b^e \sum_{n=1}^{\infty} a_n b^{-n}$$

es el número de $\mathbb{F}(b, t, L, U)$

$$\mathrm{rd}(x) = \mathrm{tr}\left(x + (-1)^s \frac{b}{2} \frac{b^e}{b^{t+1}}\right)$$

$$x = (-1)^s b^e \sum_{n=1}^{\infty} a_n b^{-n}$$

$$\mathrm{rd}(x) := (-1)^s \cdot (0.a_1 \ldots a_{t-1} r_t) \cdot b^e$$

$$r_t := \left| \begin{array}{ll} a_t, & \text{si } a_{t+1} < b/2 \\ a_t + 1, & \text{si } a_{t+1} \geq b/2 \end{array} \right.$$

Ejemplo

$\mathbb{F}(2, 4, -1, 1)$

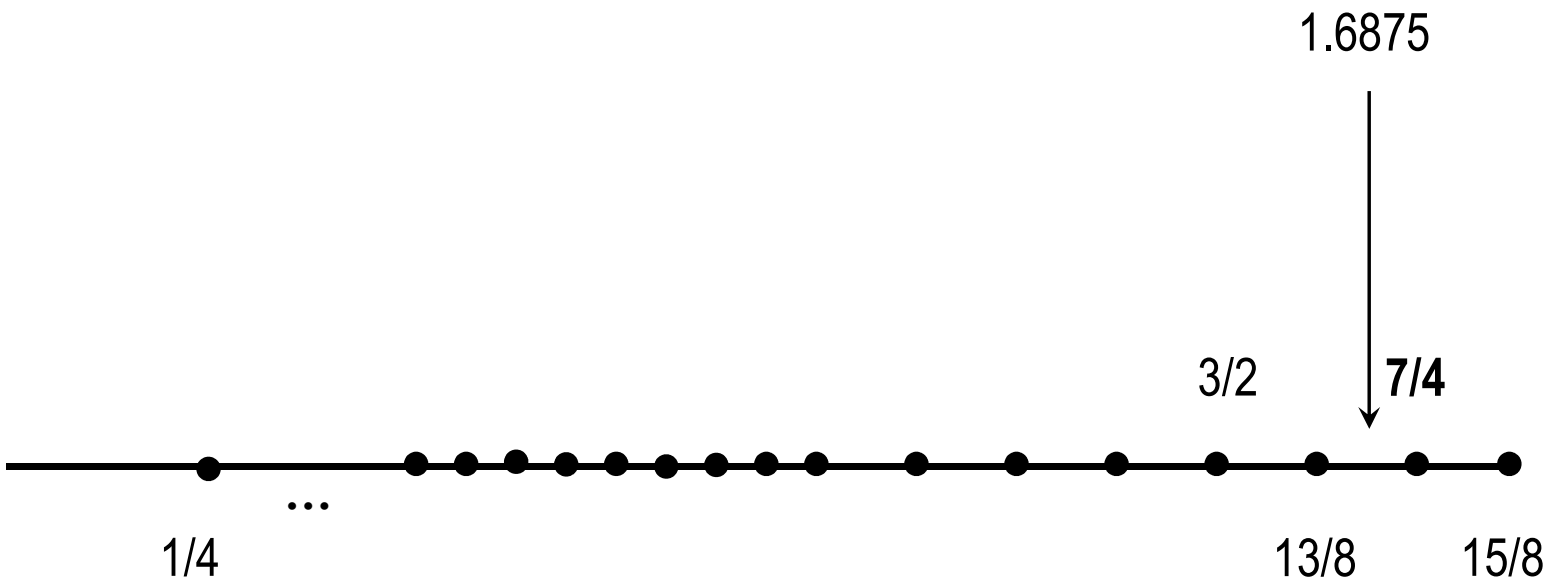
$$1.6875 = (0.11011) \cdot 2$$

- directamente con $5^{\underline{a}}$ cifra tras el punto ($t = 4$)

$$\mathrm{rd}(1.6875) = (0.1110) \cdot 2 = \frac{7}{4} = 1.75$$

- definición

$$\begin{aligned} \mathrm{rd}(1.6875) &= \mathrm{tr}\left(x + (-1)^0 \frac{2}{2}\frac{2}{2^5}\right) \\ &= \mathrm{tr}((0.11011) \cdot 2 + (0.00001) \cdot 2) \\ &= \mathrm{tr}((0.11100) \cdot 2) \\ &= (0.1110) \cdot 2 \\ &= \frac{7}{4} \\ &= 1.75 \end{aligned}$$


1.6875
3/2
7/4
1/4
13/8
15/8

$$x \in \mathbb{F}(b,t,L,M) \;\Rightarrow\; b^{L-1} \leq |x| \leq b^{U}(1-b^{-t})$$

Truncar o redondear x

$$\Downarrow$$

$$b^{L-1} \leq |x| \leq b^{U}(1-b^{-t})$$

Operación en $\mathbb{F}(b,t,L,U) \rightsquigarrow$ valor absoluto excede la cota superior *overflow*, se interrumpe el proceso computacional en curso

Operación en $\mathbb{F}(b,t,L,U) \rightsquigarrow$ valor absoluto menor que la cota inferior *underflow*, menos drástico, suele reemplazarse por cero

Proposición

Consideremos el sistema de punto flotante $\mathbb{F}(b, t, L, U)$, $L \leq e \leq U$ y sea $x = (-1)^s b^e \sum_{n=1}^{\infty} a_n b^{-n} \in \mathbb{R}$. Entonces:

(i) $|x - \mathrm{tr}(x)| \leq b^{e-t}$.

(ii) $\dfrac{|x - \mathrm{tr}(x)|}{|x|} \leq \varepsilon_M$.

(iii) $|x - \mathrm{rd}(x)| \leq \dfrac{1}{2} b^{e-t}$.

(iv) $\dfrac{|x - \mathrm{rd}(x)|}{|x|} \leq \dfrac{\varepsilon_M}{2}$.

DEMOSTRACIÓN.

(i)

$$\begin{aligned} |x - \mathrm{tr}(x)| &= b^e \left| \sum_{n=t+1}^{\infty} a_n b^{-n} \right| \\ &\leq b^e (b-1) \sum_{n=t+1}^{\infty} b^{-n} \\ &= b^{e-t} \end{aligned}$$

(ii) (i), $a_1 \geq 1$

$$\begin{aligned} \frac{|x - \mathrm{tr}(x)|}{|x|} &\leq \frac{b^{e-t}}{b^e \sum_{n=1}^{\infty} a_n b^{-n}} \\ &\leq \frac{b^{e-t}}{b^e \frac{1}{b}} \\ &= b^{1-t} \\ &= \varepsilon_M \end{aligned}$$

(iii)

$$|x - \mathrm{rd}(x)| = b^e |(0.a_1 \dots a_t a_{t+1} \dots) - (0.a_1 \dots r_t)|$$

$$\text{¿ } |(0.a_1 \dots a_t a_{t+1} \dots) - (0.a_1 \dots r_t)| \leq \frac{b^{-t}}{2} \text{ ?}$$

- si $a_{t+1} < b/2$, $a_t = r_t$,

$$\begin{aligned} |(0.a_1 \dots a_t a_{t+1} \dots) - (0.a_1 \cdots r_t)| &= (0.0 \dots 0 a_{t+1} \dots) \\ &\leq (0.0 \dots 0 \overbrace{\frac{b}{2}}^{(t+1)}) \quad \text{(Relación Ejercicios)} \\ &= \frac{b^{-t}}{2} \end{aligned}$$

- si $b/2 \leq a_{t+1}$, $r_t = a_t + 1$,

$$\begin{aligned} |(0.a_1 \dots a_t a_{t+1} \dots) - (0.a_1 \dots r_t)| &= \frac{1}{b^t} - \left(\frac{a_{t+1}}{b^{t+1}} + \cdots \right) \\ &\leq \frac{1}{b^t} - \frac{b}{2b^{t+1}} \quad (\text{pues } b/2 \leq a_{t+1}) \\ &= \frac{b^{-t}}{2} \end{aligned}$$

(iv)

$$a_1 \geq 1 \;\Rightarrow\; |x| \geq b^e b^{-1}$$

$$\Downarrow \text{(iii)}$$

$$\frac{|x - \mathrm{rd}(x)|}{|x|} \leq \frac{\frac{1}{2} b^{e-t}}{b^e b^{-1}} = \frac{1}{2} b^{1-t} = \frac{\varepsilon_M}{2}$$

□

Definición

La cota que aparece en el error relativo del redondeo recibe el nombre de *precisión máquina* (o *unidad de redondeo*) y se nota por u, es decir,

$$\mathrm{u} := \frac{1}{2} b^{1-t} = \frac{1}{2} \varepsilon_M.$$

$$b^{L-1} \leq |x| \leq b^{U}(1 - b^{-t}) \Rightarrow \left| \begin{array}{l} \dfrac{|x - \mathrm{tr}(x)|}{|x|} \leq \varepsilon_M \\ \dfrac{|x - \mathrm{rd}(x)|}{|x|} \leq \mathrm{u} \end{array} \right.$$

Corolario

Dados $\mathbb{F}(b, t, L, U)$ y $x \in \mathbb{R}$, con $b^{L-1} \leq |x| \leq b^U(1 - b^{-t})$, se tiene que

$$\mathrm{rd}(x) = (1 + \mu)x,$$

para cierto número real μ con $|\mu| \leq \mathrm{u}$.

Demostración.

$$|x - \mathrm{rd}(x)| \leq \mathrm{u}|x|$$

$$\Updownarrow$$

$$x - |x|\mathrm{u} \leq \mathrm{rd}(x) \leq x + |x|\mathrm{u}$$

esto es,

$$\mathrm{rd}(x) = x + \kappa|x|\mathrm{u},$$

con $|\kappa| \leq 1$, es decir,

$$\mathrm{rd}(x) = (1 + \mu)x,$$

siendo $\mu \in \mathbb{R}$ con $|\mu| \leq \mathrm{u}$ □

Operaciones con números máquina en $\mathbb{F}(b,t,L,U)$ no necesaiamente internas

$$\bullet : \mathbb{R} \times \mathbb{R} \longrightarrow \mathbb{R} \text{ operación}$$

operación máquina $\bullet_M : \mathbb{F}(b,t,L,U) \times \mathbb{F}(b,t,L,U) \longrightarrow \mathbb{F}(b,t,L,U)$

$$\bullet_M(x,y) := \mathrm{rd}(x \bullet y), \qquad (x,y \in \mathbb{F}(b,t,L,U))$$

Corolario

$$x,y \in \mathbb{F}(b,t,L,U) \;\Rightarrow\; \bullet_M(x,y) = (1+\mu)x \bullet y,$$

para cierto $\mu \in \mathbb{R}$ con $|\mu| \leq \mathrm{u}$.

Observación

Aunque una única operación genere un error pequeño, una sucesión finita de operaciones es susceptible de producir la llamada *propagación del error*, que puede ser considerable.

Ejemplo

$n \geq 1$

$$x_n := \int_0^1 x^n e^x dx$$

Integración por partes

$$\begin{aligned} \int_0^1 x^n e^x dx &= x^n e^x]_0^1 - n \int_0^1 x^{n-1} e^x dx \\ &= e - n \int_0^1 x^{n-1} e^x dx \end{aligned}$$

$$x_n = e - n x_{n-1}$$

Recurrencia

$$x_0 = e - 1$$

y

$$n \in \mathbb{N} \Rightarrow x_n = e - n x_{n-1}$$

$$x \in [0,1] \;\Rightarrow\; x^n e^x \leq x^{n-1} e^x$$

$\{x_n\}_{n\geq 0}$ decreciente (y positiva) $\rightsquigarrow$ converge, ℓ

$$x_{n-1} = \frac{e - x_n}{n} \;\Rightarrow\; \ell = 0$$

redondeo sistema punto flotante $b = 10,\ t = 7$

$$x_{15} = (0.1604004)$$

$$x_0 = (0.1718281828) \cdot 10$$

$$x_{15} = (0.6004419078) \cdot 10^3$$

propagación del error
incluso *Maxima* propaga el error partiendo del valor "exacto"

Condicionamiento de una sucesión de funciones directamente relacionadas con la sucesión $\{x_n\}_{n\geq 0}$

$x_n \rightsquigarrow x_0$ recurrentemente

$$x_n = f_n(x_0)$$

$f_n : \mathbb{R} \longrightarrow \mathbb{R}$

$$x_1 = e - x_0$$

$$x_2 = e - 2(e - x_0) = -e + 2x_0$$

$$x_3 = e - 3(-e + 2x_0) = 4e - 6x_0$$

inductivamente $n \geq 1$

$$(-1)^n(n!x_0 - \alpha_n e)$$

$\alpha_n \in \mathbb{N}$ independiente de x_0 (¡valor irrelevante!)

$$x_n = f_n(x_0)$$

$$f_n(x) = (-1)^n(n!x - \alpha_n e), \qquad (x \in \mathbb{R})$$

Condicionamiento de f_n en x_0

$$\begin{aligned} c(f_n, x_0) &= \frac{n!x_0}{x_n} \\ &\geq \frac{n!x_0}{x_0} \\ &= n! \end{aligned}$$

Operaciones aritméticas elementales y errores debidos a truncaturas, redondeos o a cualquier otra circunstancia

Dato x, μ_x error relativo para x y su valor aproximado

valor aproximado $(1 + \mu_x)x$

Suma (resta) $x, y \in \mathbb{R}$, $x + y \neq 0$
valores aproximados $(1 + \mu_x)x$, $(1 + \mu_y)y$

$$\begin{aligned}(1+\mu_x)x + (1+\mu_y)y &= x + y + \mu_x x + \mu_y y \\ &= (x+y)\left(1 + \frac{\mu_x x + \mu_y y}{x+y}\right) \\ &= (x+y)\left(1 + \frac{x}{x+y}\mu_x + \frac{y}{x+y}\mu_y\right),\end{aligned}$$

$$\mu_{x+y} = \frac{x}{x+y}\mu_x + \frac{y}{x+y}\mu_y$$

x e y tienen el mismo signo

$$|\mu_{x+y}| \leq |\mu_x| + |\mu_y|$$

control de los errores relativos

x e y de signos opuestos
μ_{x+y} puede dispararse si $x + y \approx 0$, generándose un error relativo enorme, conocido como *error de cancelación*.

Ejemplo

Sistema punto flotante $b = 10$, $t = 7$

$$\sqrt{30 + 10^{-5}} - \sqrt{30} = (0.5477226) - (0.5477226) = 0$$

Cociente (buen comportamiento, a continuación) y una suma

$$\frac{10^{-5}}{\sqrt{30 + 10^{-5}} + \sqrt{30}} = (0.9128710) \cdot 10^{-6}$$

Redondeo del valor exacto $(0.9182709) \cdot 10^{-6}$

Multiplicación, errores relativos de x e y pequeños

$$\begin{aligned}(1+\mu_x)x\cdot(1+\mu_y)y &= (1+\mu_x+\mu_y+\mu_x\mu_y)x\cdot y\\ &\approx (1+\mu_x+\mu_y)x\cdot y\end{aligned}$$

$$\Downarrow$$

$$\mu_{x\cdot y}\approx \mu_x+\mu_y$$

División, errores relativos de x e y pequeños, $y\neq 0$

$$\begin{aligned}\frac{(1+\mu_x)x}{(1+\mu_y)y} &= \frac{x}{y}(1+\mu_x)(1-\mu_y+\mu_y^2-\mu_y^3+\cdots) \qquad \text{(Taylor en } x_0=0)\\ &\approx \frac{x}{y}(1+\mu_x-\mu_y)\end{aligned}$$

$$\Downarrow$$

$$\mu_{x/y}\approx \mu_x-\mu_y$$

I.2.3. Iteradores

Solución $\rightsquigarrow$ límite de una sucesión

Teorema del punto fijo de Banach

Recursivamente: x_0

$$n \geq 1 \;\Rightarrow\; x_n := f(x_{n-1})$$

Hipótesis adecuadas $\rightsquigarrow \{x_n\}_{n\in\mathbb{N}}$ converge al único punto fijo x de f, $f(x) = x$

Términos iterando sucesivamente $f \rightsquigarrow$ *iteradores* (o *iterantes*)

Por extensión nomenclatura para los términos de una sucesión cuyo límite aproxime la solución de un problema

Observación

No toda sucesión convergente $\{x_n\}_{n\in\mathbb{N}}$ cuyo límite coincida con la solución de un problema bien planteado es de la forma $x_{n+1} = f(x_n)$, para una conveniente aplicación f: baste mencionar la sucesión de Fibonacci.

Problema bien planteado

$$\text{encontar } x_0 \in X : \ f(x_0) = y_0$$

$f : X \longrightarrow Y$, X e Y subconjuntos no vacíos de sendos espacios normados reales

$n \geq 1$, existen X_n e Y_n subconjuntos no vacíos de sendos espacios normados reales, $y_n \in Y_n$, $f_n : X_n \longrightarrow Y_n$

$$\text{encontar } x_n \in X_n : \ f(x_n) = y_n$$

problema bien planteado

Propiedades Análisis Numérico relacionadas con esta familia de problemas: *convergencia* de la sucesión $\{x_n\}_{n\in\mathbb{N}}$ hacia x_0, *estabilidad*, entendida de forma uniforme y *consistencia*

$$\lim_{n\to\infty} f_n(x_n) = y_0$$

Bajo consistencia, estabilidad y convergencia coinciden (una buena referencia es el texto de Quarteroni, Section 2.2)

Condicionamiento (relativo o absoluto) de la familia anterior en un punto $\leftrightsquigarrow$ supremo de todos ellos (o límite superior)

Ejemplo

Problema, resolvente $g : (-1, 1) \longrightarrow \mathbb{R}$

$$g(a) := \sum_{j=0}^{\infty} a^j, \qquad (-1 < a < 1)$$

$n \geq 1$, problema, $g_n : (-1 - 1) \longrightarrow \mathbb{R}$

$$g_n(a) := \sum_{j=0}^{n} a^j, \qquad (-1 < a < 1)$$

$$c(g_n, a) = \frac{|1 + 2a + \cdots + na^{n-1}|}{|1 + a + \cdots + a^n|}|a|$$

$$n \geq 1 \;\Rightarrow\; \lim_{a \to 0} c(g_n, a) = 0 \qquad \text{(buen condicionamiento)}$$

$$n \geq 1 \;\Rightarrow\; \lim_{a \to 1} c(g_n, a) = \frac{1 + 2 + \cdots + n}{n + 1} = \frac{n}{2} \qquad \text{(mal condicionamiento)}$$

I.3. Bibliografía

1. S. Brin, L. Page, *The anatomy of a large-scale hypertextual web search engine*, http://www-db.stanford.edu/backrub/google.html (August 1, 2005).

2. K. Bryan, T. Leise, *The $25,000,000,000 eigenvector: the linear algebra behind Google*, SIAM Review 48 (2006), 569–581.

3. M. Fabian, Habala, P. Hájek, V. Montesinos, V. Zizler, *Banach space theory. The basis of linear and nonlinear analysis*, CMS Books in Mathematics, Springer, New York, 2011.

4. E. Isaacson, H. Keller, *Analysis of numerical methods*, Wiley, New York, 1966.

5. A. Quarteroni, R. Sacco, F. Saleri, *Numerical mathematics*, second edition, Texts in Applied Mathematics **37** Springer–Verlag, Berlin, 2007.