Capítol 1

Introducció

1.1 Preliminars de Càlcul

1.1.1 Desigualtats.

Definició 1.1 *Direm que* $a \leq b \Leftrightarrow b - a \geq 0$.

Definició 1.2 *Direm que* $a < b \Leftrightarrow b - a > 0$.

Propietats 1.1 *A partir de la definició anterior es dedueixen les següents propietats:*

(a) $a < b$ *i* $b < c \Rightarrow a < c$
(b) $a < b$ *i* $c < d \Rightarrow a + c < b + d$
(c) $a < b$ *i* $\alpha \in R \Rightarrow a + \alpha < b + \alpha$
(d) $a < b$ *i* $\alpha > 0 \Rightarrow a\alpha < b\alpha$
(e) $a < b$ *i* $\alpha < 0 \Rightarrow a\alpha > b\alpha$

Definició 1.3 *Direm que una funció* $f(x)$ *és creixent en un interval* I *si i només si:*

$$(\forall x, y \in I) \quad (x \leq y \Rightarrow f(x) \leq f(y))$$

Definició 1.4 *Direm que una funció* $f(x)$ *és decreixent en* I *si i només si:*

$$(\forall x, y \in I\ ) \quad (x \leq y \Rightarrow f(x) \geq f(y))$$

Exemple 1.1 *$f(x) = \ln x$ és una funció creixent en el seu domini $D = (0, +\infty)$. per tant:*

$$(\forall x, y \in D)\ (x \leq y \Rightarrow \ln x \leq \ln y)$$

Exemple 1.2 *$f(x) = e^{-x}$ és una funció decreixent en el seu domini $\mathbb{R}$, per tant:*

$$(\forall x, y \in \mathbb{R})\ \big(x \leq y \Rightarrow e^{-x} \geq e^{-y}\big)$$

Exemple 1.3 *$f(x) = x^2$ és una funció creixent per a $x > 0$ i és decreixent per a $x < 0$. Així:*

$$\begin{cases} 0 < x \leq y \Rightarrow x^2 \leq y^2 \\ x \leq y < 0 \Rightarrow x^2 \geq y^2 \end{cases}$$

1.1.2 Valor absolut

Definició 1.5 *Definim el valor absolut d'un nombre real com:*

$$|x| = \begin{cases} x \text{ si } x \geq 0 \\ -x \text{ si } x < 0 \end{cases}$$

Propietats 1.2 *Podem deduir les següents propietats, que ens seràn útils par a fitar expressions:*

(a) $|x| \geq 0$; $|x| = 0 \Leftrightarrow x = 0$

(b) $|x| \leq a$, $a \geq 0 \Leftrightarrow -a \leq x \leq a$

(c) $|x| \geq a$, $a \geq 0 \Leftrightarrow x \leq -a$ *o* $x \geq a$

(d) $|x \cdot y| = |x| \cdot |y|$

(e) $\left|\dfrac{x}{y}\right| = \dfrac{|x|}{|y|}$, $(y \neq 0)$

(f) $|x + y| \leq |x| + |y|$ *(Desigualtat triangular)*

(g) $|x - y| \leq |x| + |y|$

(h) $||x| - |y|| \leq |x - y|$

(i) $x^2 < a \Rightarrow |x| < +\sqrt{a}$

Exemple 1.4 *Trobeu una fita de les següents expressions:*

(a) $\left|\dfrac{(1+\sin x)n}{1+n}\right|$ *amb* $n \in N$

(b) $\left|\dfrac{2x^2-3x+1}{2x-1}\right|$ *amb* $2 \leq x \leq 3$

a) Aplicant les propietats del valor absolut, obtenim:

$$\left|\frac{(1+\sin x)n}{1+n}\right| = |1+\sin x| \,.\, \left|\frac{n}{1+n}\right| < |1+\sin x| \leq 1 + |\sin x| \leq 2$$

b) Per a la segona expressió definim $f(x) = \dfrac{2x^2-3x+1}{2x-1}$ en $I = [2,3]$.Calculem la seva derivada :

$$f'(x) = 1 > 0 \Rightarrow f(x) \text{ és creixent}$$

A partir d'aquest fet deduïm que:

$$2 \leq x \leq 3 \Rightarrow f(2) \leq f(x) \leq f(3)$$

És a dir:

$$2 \leq x \leq 3 \Rightarrow 1 \leq f(x) \leq 2$$

I a partir d'aquí tenim la fita que buscàvem:

$$2 \leq x \leq 3 \Rightarrow |f(x)| \leq 2$$

1.1.3 Teoremes notables de continuïtat i derivabilitat

Teorema 1.1 *(de Bolzano)*

$$\left.\begin{array}{l} f:[a,b] \to \mathbb{R} \text{ contínua} \\ f(a)\cdot f(b) < 0 \end{array}\right\} \Rightarrow \exists \alpha \in (a,b) \mid f(\alpha) = 0$$

Exemple 1.5 *Trobeu un interval en el qual es pugui assegurar l'existència d'una solució de l'equació* $x - \sin x - 1 = 0$.

Si fem una gràfica de la funció tindrem una idea aproximada d'on es troben les arrels:

Comprovem ara les hipòtesis del teorema de Bolzano a l'interval I=$[\frac{\pi}{2}, \pi]$.

$$\left.\begin{array}{l} f(x) = x - \sin x - 1 \text{ contínua en I} \\ f(\frac{\pi}{2}) < 0 \\ f(\pi) > 0 \end{array}\right\} \Rightarrow \exists \alpha \in (\tfrac{\pi}{2}, \pi) \text{ tal que } f(\alpha) = 0$$

Per tant, podem assegurar que a l'interval I existeix una solució de l'equació.

Teorema 1.2 *(de Weierstrass)*

$$f \text{ contínua en } [a,b] \Rightarrow \exists \max_{x \in [a,b]} f(x) \text{ i } \exists \min_{x \in [a,b]} f(x)$$

Exemple 1.6 *Trobeu els extrems absoluts de la funció $f(x) = x^2 - 4$ a l'interval $[-1, 1]$.*

La funció $f(x) = x^2 - 4$ és contínua en $[-1, 1]$, llavors existirà màxim i mínim absolut dins de l'interval.

Començarem per trobar els candidats:

- Punts crítics de l'obert $(-1, 1)$:

  - $f'(x) = 0 : \; 2x = 0 \Rightarrow x = 0 \in (-1, 1)$
  - $\nexists f'(x)$: no hi ha candidats

- Extrems de l'interval: $x = -1$, $x = 1$

- Avaluació de les imatges: $f(0) = -4$
  $f(-1) = -3$
  $f(1) = -3$

Per tant, el màxim absolut és -3 i es troba en els punts $x = -1$ i $x = 1$. El mínim absolut és -4 i es troba en $x = 0$. Això ho podem observar si fem la gràfica de la funció i ens restringim a l'interval $[-1, 1]$:

Teorema 1.3 *(de Rolle)*

$$\left.\begin{array}{l} f \text{ contínua en } [a,b] \\ f \text{ derivable en } (a,b) \\ f(a) = f(b) \end{array}\right\} \Rightarrow \exists \alpha \in (a,b) \mid f'(\alpha) = 0$$

Corol.lari 1.1 *Sigui $f(x)$ derivable en (a,b) i suposem que $f'(x) = 0$ té n solucions en (a,b). Aleshores, $f(x) = 0$ en té, com a màxim, n+1.*

1.2 Errors

1.2.1 Conceptes generals

Definició 1.6 *Sigui $\overline{x}$ una aproximació del valor exacte x d'una magnitud. Aleshores, definim l'**error absolut** comès per $\overline{x}$ com*

$$e_a(x) = |x - \overline{x}|$$

*i l'**error relatiu** comès per $\overline{x}$ com*

$$e_r(x) = \frac{e_a(x)}{|x|} \text{ si } x \neq 0.$$

L'error relatiu expressa l'error com una fracció de $|x|$, per això està relacionat amb l'error percentual. Per exemple, si $.e_r(x) < 0.02$, aleshores $e_a(x) < 0.02\,|x|$, és a dir, $\overline{x}$ s'aproxima a x dintre del 2%.

De manera més general, es defineix el **percentatge d'error** com l'error relatiu multiplicat per 100: $100e_r(x)$

Si coneixem una aproximació $\overline{x}$ d'un nombre desconegut x i es té una fita superior de l'error absolut, és a dir, un nombre ε tal que

$$e_a(\overline{x}) \leq \varepsilon$$

aleshores, es té un interval de confiança o interval de precisió per a x,

$$x \in [\overline{x} - \varepsilon\,, \overline{x} + \varepsilon]$$

Recíprocament, si es coneix un interval de confiança per a x, $x \in [a, b]$, aleshores $\overline{x} = \frac{a+b}{2}$ és la millor aproximació de x que tenim i el seu error absolut es pot afitar per

$$e_a(\overline{x}) \leq \frac{b-a}{2}$$

Decimals exactes i xifres significatives

Sigui $\overline{x}$ una aproximació d'un nombre real x.

- Direm que $\overline{x}$ aproxima x amb t decimals exactes si

$$|x - \overline{x}| \leq \frac{1}{2} \times 10^{-t} \tag{1.1}$$

Això s'expressa, encara que no sigui estrictament cert, dient que x i $\overline{x}$ coincideixen en les t primeres xifres decimals.

- Direm que $\overline{x}$ aproxima x amb t xifres significatives si

$$\frac{|x - \overline{x}|}{|x|} \leq \frac{1}{2} \times 10^{-t+1}$$

Exemple 1.7 Donats el nombre $x = .001234$ i l'aproximació $\overline{x} = .001256$

(a) $|x - \overline{x}| = .000022 = .22 \times 10^{-4}$ ens diu que tenim una aproximació amb 4 decimals exactes.

(b) $\dfrac{|x - \overline{x}|}{|x|} = .1751592 \ldots \times 10^{-1}$ ens assegura que tenim una aproximació amb dues xifres significatives.

Observem que la condició (1.1) no sempre assegura que x i $\overline{x}$ coincideixin en les t primeres xifres decimals, per exemple, siguin $\overline{a} = 2.9999$ i $a = 3.0000$, aleshores

$$|a - \overline{a}| = .0001 = .1 \times 10^{-3}$$

Podem afirmar que tenim tres decimals exactes, encara que el nombre i la seva aproximació no coincideixin en cap dígit.

1.2.2 Fonts d'error

Les fonts d'error en la resolució numèrica d'un problema es poden classificar en:

(a) Errors en les dades inicials

(b) Errors d'arrodoniment durant el càlcul.

(c) Errors de truncament.

(d) Errors en la construcció del model matemàtic: Són conseqüència de les simplificacions que es fan quan es descriu matemàticament un fenomen real.

Des d'un punt de vista numèric, ens interessen principalment els errors del tipus (a), (b) i (c).

1.2.3 Errors en les dades inicials

Poden ser deguts a mesuraments incorrectes, pel fet que les dades provinents d'experiments depenen de la precisió dels aparells utilitzats, o bé perquè en introduir dades inicials correctes es produeix un arrodoniment.

Representació de nombres en punt flotant

Tot nombre real x ($x \neq 0$) pot representar-se en una base natural $b \geq 2$ en la forma

$$x = a_n b^n + a_{n-1} b^{n-1} + \cdots + a_1 b + a_0 + a_{-1} b^{-1} + a_{-2} b^{-2} + \cdots$$

amb $a_j \in Z$, $0 \leq a_j < b$ ($j \leq n$), que habitualment s'escriu

$$x = (a_n a_{n-1} \ldots a_1 a_0 . a_{-1} a_{-2} \ldots)_b$$

i s'anomena *representació digital de x en base b*. Els coeficients a_j reben el nom de *dígits* o *xifres*. Per exemple, el nombre 1011.01 en base 2 és

$$1 \cdot 2^3 + 0 \cdot 2^2 + 1 \cdot 2 + 1 + 0 \cdot 2^{-1} + 1 \cdot 2^{-2} = 11.25 \text{ en base } 10$$

Cada ordinador utilitza una base b preestablerta i representa cada nombre amb una quantitat finita p de dígits en la forma

$$fl(x) = \pm 0 \,.\, a_1 a_2 ... a_p \times b^\alpha$$

anomenada **representació en punt flotant**, amb $a_1 \neq 0$, on $a_1 a_2 ... a_p$ s'anomena *mantissa i α exponent.*

El pas d'un nombre a la seva representació en punt flotant es pot obtenir de dues maneres:

- Per *truncament*: en aquest cas menyspreem tots els dígits a partir de la posició p i ho expressarem mitjançant $fl_{T,p}(x)$.
- Per *arrodoniment*: en aquest cas prenem com a aproximació en punt flotant $fl(x)$ tal que fa mínima la diferència $|fl(x) - x|$ i ho escriurem $fl_{A,p}(x)$. Quan aquesta condició dóna dos possibles arrodoniments, s'acostuma a triar el que té valor absolut més gran.

Exemple 1.8 Per simplificar, prendrem $b = 10$.

(a) El nombre $x = 0.0000123$ el representem en punt flotant com

$$fl(x) = 0.123 \times 10^{-4}$$

(b) Si $x = \dfrac{14}{3} = 4.666.\widehat{6}$, la seva representació en punt flotant amb 4 digits serà:

$$fl_{T,4}(x) = 0.4666 \times 10 \quad \text{(per truncament)}$$
$$fl_{A,4}(x) = 0.4667 \times 10 \quad \text{(arrodonint)}$$

(c) Si $x = 0.9999500$, la seva representació en punt flotant amb 4 digits serà:

$$fl_{T,4}(x) = 0.9999 \quad \text{(per truncament)}$$
$$fl_{A,4}(x) = 0.1000 \times 10 \quad \text{(arrodonint)}$$

(d) El nombre 3.424748 arrodonit a 6 xifres és 3.42475, i 3.42475 arrodonit a 5 xifres queda 3.4248. Observem, però, que arrodonint (directament) 3.424748 a 5 xifres queda 3.4247

1.2.4 Errors d'arrodoniment durant el càlcul

Si bé l'aritmètica de punt flotant és la més eficient per a les necessitats del càlcul numèric, produeix unes propietats que poden ser diferents de les propietats amb nombres exactes. Els casos més conflictius de pèrdua de precisió es donen quan operem amb nombres molt grans i molt petits alhora, o bé quan calculem diferències de quantitats molt semblants. Vegem-ne alguns casos:

- Cal tenir en compte que per a sumar dos nombres, primer es redueixen al mateix exponent, s'operen i després s'arrodoneix.

  Per exemple, amb $p = 4$:

  i) $0.3467 \cdot 10^2 + 0.4125 \cdot 10^2 = 0.7592 \cdot 10^2$
  ii) $0.3467 \cdot 10^2 + 0.4125 \cdot 10^{-3} = (0.3467 + 0.000004125) \cdot 10^2 = 0.3467 \cdot 10^2$

  En el segon cas, la suma no canvia malgrat haver sumat un nombre diferent de zero.

- La suma no és associativa, perquè s'arrodoneix a cada operació. Per exemple, siguin

  $$a = 0.1234567 \cdot 10^0 \ , \ b = 0.4711325 \cdot 10^4 \ , \ c = -0.4711325 \cdot 10^4$$

  i suposem que treballem amb aritmètica de punt flotant amb 4 dígits. Aleshores,

  $$\begin{aligned} &a + (b + c) = a + 0 = 0.1235 \\ &(a + b) + c = (0.1235 \cdot 10^0 + 0.4711 \cdot 10^4) - 0.4711 \cdot 10^4 = 0 \end{aligned}$$

  Per tant

  $$(a + b) + c \neq a + (b + c)$$

- El producte exacte de dos nombres amb p dígits requereix $2p$ dígits, que són arrodonits a p.

  Per exemple, considerem una màquina que treballa amb 4 xifres decimals i talla o arrodoneix. Hi tenim dues dades:

  $$a = 0.2345 \cdot 10^5, \qquad b = 0.5432 \cdot 10^{-2}$$

  Aleshores,
  $$\begin{aligned} ab &= 0.1273804 \cdot 10^3 \ , \\ fl_{T,4}(ab) &= 0.1273 \cdot 10^3 \text{ (s'ha comès un error de 0.0804)} \ , \\ fl_{A,4}(ab) &= 0.1274 \cdot 10^3 \text{ (s'ha comès un error de 0.0196)} \ . \end{aligned}$$

  Observem que l'error en la representació per truncament és més gran que en la representació per arrodoniment.

- La divisió d'un resultat amb punt flotant per un nombre molt petit (o la multiplicació per un de molt gran) dóna un error absolut molt gran.

  Si r és el nombre que ha estat representat per $r+e$, és clar que si volem calcular $\frac{r}{d}$ on d és petit obtindrem un error aproximadament igual a $\frac{e}{d}$ Si d és prou petit, aquest quocient pot ser més gran que $\frac{r}{d}$.

- La resta de nombres molt semblants dóna un error relatiu molt gran. Per exemple,

  Suposem que els dos nombres x_1 i x_2 s'expressen

  $$\begin{aligned} fl(x_1) = 0.d_1 d_2 ... d_p \alpha_{p+1} ... \alpha_k \times 10^q \\ fl(x_2) = 0.d_1 d_2 ... dp \beta_{p+1} ... \beta_k \times 10^q \end{aligned}$$

  Aleshores, la resta $x_1 - x_2$ es representarà com

  $$fl(fl(x_1) - fl(x_2)) = 0.\sigma_{p+1} ... \sigma_k \times 10^{q-p}$$

  i només tindrem $k-p$ xifres significatives i qualsevol càlcul posterior ja no en podrà donar més. Aquesta pèrdua de xifres correctes es coneix com a *cancel.lació*.

1.2.5 Errors de truncament

Molts mètodes numèrics consisteixen en generar successions que tendeixen a la solució del problema quan el nombre de termes de la successió tendeix a infinit. Quan un procés infinit o de pas al límit és tallat després d'un nombre finit de passos es produeix un error, propi del mètode, anomenat **error de truncament**. Per exemple, quan una sèrie infinita és substituïda per sumes parcials o quan una derivada és aproximada per un quocient de diferències.

1.2.6 Estabilitat numèrica

Hi ha situacions on un petit error inicial es propaga molt i s'obté un resultat que difereix del resultat real molt més que la diferència inicial. D'això se'n diu **inestabilitat**.

Un cas típic d'inestabilitat són certs sistemes lineals anomenats **mal condicionats**. La solució d'un sistema 2×2 correspon, geomètricament, a la intersecció de dues rectes del pla. Si les dues rectes són paral·leles o bé coincideixen o bé no es tallen mai. En el primer cas hi ha infinites solucions, mentre que en el segon no hi ha solució. En els dos casos, les dues files de la matriu del sistema són linealment dependents i, per tant, la matriu no és invertible.

Si les dues rectes són "*quasi paral·leles*", aleshores hi haurà solució única, però movent molt poc una de les rectes, el punt d'intersecció (la solució) es desplaçarà molt. Per exemple, el sistema,

$$\begin{pmatrix} 1 & 2 \\ 0.499 & 1.001 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 3 \\ 1.5 \end{pmatrix}$$

té solució $x = y = 1$. Si canviem lleugerament la matriu, el sistema

$$\begin{pmatrix} 1 & 2 \\ 0.5 & 1.001 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 3 \\ 1.5 \end{pmatrix}$$

té per solució $x = 3\,,\ y = 0$ que difereix molt més de la solució anterior que la pertorbació soferta per la matriu.

Així doncs, quan es decideix utilitzar un mètode numèric determinat cal tenir en compte tots els factors possibles, com són els errors d'arrodoniment produïts per les operacions que comporta el mètode, el seu error de truncament (si en té) i la seva estabilitat numèrica.