Computació Numèrica

Tema 1. Conceptes bàsics en computació numèrica

M. Àngela Grau Gotés

Departament de Matemàtiques
Universitat Politècnica de Catalunya · BarcelonaTech.

19 de febrer de 2018

drets d'autor

"Donat el caràcter i la finalitat exclusivament docent i eminentment il·lustrativa de les explicacions a classe d'aquesta presentació, l'autor s'acull a l'article 32 de la Llei de propietat intel·lectual vigent respecte de l'ús parcial d'obres alienes com ara imatges, gràfics o altre material contingudes en les diferents diapositives"

Índex

1. Introducció
2. Definicions d'error
3. Errors d'arrodoniment
4. Errors de truncament
5. Propagació de l'error
6. Algorismes
7. Representació de nombres
8. Nombres a l'ordinador
9. Aritmètica de Matlab

Introducció

Durant els segles XX i XXI models matemàtics avançats s'han aplicat en diferents àrees de coneixement com l'enginyeria, la medicina, l'economia o les ciències socials. Sovint, les aplicacions generen problemes matemàtics que per la seva complexitat no poden ser resolts de manera exacta.

La **matemàtica computacional**, s'ocupa del disseny, anàlisi i implementació d'algorismes per obtenir solucions numèriques aproximades de models físics, químics, matemàtics, estadístics, . . .

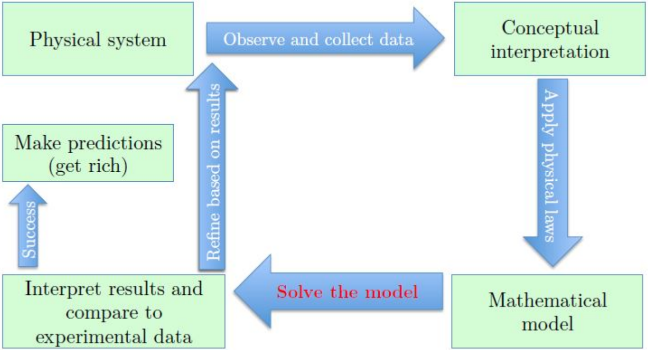
Matlab Users's Guide online

Modelització

Davant d'un fet real, la modelització consisteix a construir un conjunt de fórmules i equacions que ens el representin de la manera més fidel possible, de manera que ens permeti fer prediccions correctes.

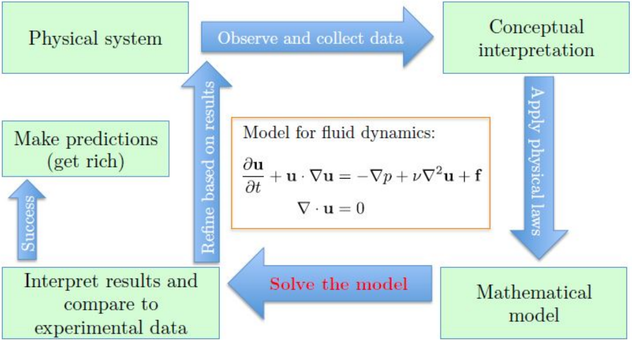
Els models resultants quasi mai no poden ser resolts per complet utilitzant mètodes (llapís i paper) d'anàlisi. La simulació en un ordinador ens permet interpretar els resultats i comparar-los amb les dades experimentals.

Modelització



Davant d'un fet real, la modelització consisteix a construir un conjunt de fórmules i equacions que ens el representin de la manera més fidel possible, de manera que ens permeti fer prediccions correctes.

Modelització



Els models resultants quasi mai no poden ser resolts per complet utilitzant mètodes (llapís i paper) d'anàlisi. La simulació en un ordinador ens permet interpretar els resultats i comparar-los amb les dades experimentals.

Funció error

$$\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2}\, dt$$

- S'anomena funció error, s'aplica si els resultats d'un conjunt de mesures es descriuen per una distribució normal de mitja zero i desviació estàndar σ, llavors $\operatorname{erf}\left(\frac{\epsilon}{\sigma\sqrt{2}}\right)$ és la probabilitat que l'error en una de les mesures es trobi entre $-\epsilon$ i ϵ.
- Però la integral definida no es pot expressar per mitja de funcions elementals
- **Cal obtenir aproximacions numèriques!**

Funció error

La sèrie de potències en un entorn de 0 és:

$$erf(x) = \frac{2}{\sqrt{\pi}} \sum_{n=0}^{\infty} \frac{(-1)^n x^{2n+1}}{n!(2n+1)}$$

$$erf(x) = \frac{2}{\sqrt{\pi}} \left(x - \frac{x^3}{3} + \frac{x^5}{10} - \frac{x^7}{42} + \frac{x^9}{216} - \cdots \right)$$

Sèrie de convergència lenta.

Algorismes

Un algorisme (matemàtic) és un procediment formal que descriu una seqüència ordenada i finita d'operacions a realitzar un nombre finit de vegades per tal d'obtenir la solució d'un problema.

Algorismes són com receptes amb els blocs bàsics de operacions **suma**, **resta**, **multiplicació**, i **divisió**, així com les estructures de programació: **for**, **while**, i **if** si es resol amb ajut d'ordinadors.

Algorismes

En general, qualsevol algorisme que desenvolupem o del que fem ús ha de ser: precís, estable, eficient i robust.

Exacte Què tan bo és l 'algorisme d'aproximació de la quantitat a calcular (accuracy).

Estable La sortida de l'algorisme és sensible a petits canvis en les dades d'entrada (stability).

Eficient Quant costa (en nombre d'operacions) a obtenir una aproximació raonable (efficiency).

Robust Per a quants casos puc fer ús de l'algorisme (robustness).

En alguns casos també importa la memòria necessària (storage) i si és paral·lelizable (parallelization).

Algorismes

En aquest curs treballarem amb dos tipus d'algorismes

Directes Obtenen la solució en un nombre finit de pasos, sense erros d'arrodoniment.

Exemples *Equacions de segon grau. El·liminació gaussiana*

Iteratius Generen una seqüència de valors aproximats que convergeixen a la solució quan el nombre de passos tendeix a infinit.

Exemple *El mètode iteratiu següent convergeix a* $\sqrt{2}$.

$$x_k = \frac{1}{2}\left(x_{k-1} + \frac{2}{x_{k-1}}\right) \qquad k \geq 1 \ i \ x_0 = 2\,.$$

Algorismes

Els sis primers iterats de $x_k = \dfrac{1}{2}\left(x_{k-1} + \dfrac{2}{x_{k-1}}\right)$ són:

n	x_n	$\lvert x_n - \sqrt{2}\rvert$
0	3.000000000000000	$1.58578643762690 \times 10^{0}$
1	1.833333333333333	$4.19119770960238 \times 10^{-1}$
2	1.462121212121212	$4.79076497481170 \times 10^{-2}$
3	1.414998429894803	$7.84867521707922 \times 10^{-4}$
4	1.414213780047198	$2.17674102520604 \times 10^{-7}$
5	1.414213562373112	$1.66533453693773 \times 10^{-14}$
6	1.414213562373095	$2.22044604925031 \times 10^{-16}$

Fonts d'error

Problema: Calcular la massa de la Terra.
Solució. Usant la Llei de Gravitació Universal de Newton i la llei de Galileu de caiguda de cossos, obtenim

$$M = \frac{gR^2}{G},$$

on g és l'acceleració de la gravetat, R el radi de la Terra, i G la constant de gravitació, amb valors experimentals

$$\begin{aligned} g &= 9.80665 \;\; m \cdot s^{-2}, \\ G &= 6.67428 \cdot 10^{-11} \;\; m^3 \cdot kg^{-1} s^{-2}, \\ R &= 6371.0 \;\; km. \end{aligned}$$

Per tant, $M = 5.9639 \cdot 10^{24} \;\; kg$.

Nota $M = 5.9736 \cdot 10^{24} \;\; kg$ (Wikipedia, NASA).

$M = 5.9742 \cdot 10^{24} \;\; kg$ (J.M.A. Danby, *Fundamentals of Celestial Mechanics*, Willmann-Bell, Inc., 1992).

Error

Fixarem quatre grans fonts d'error, que poden influir en una aproximació "pobre" del fet observat:

- Error de modelització. Una elecció equivocada o inapropiada de model.
- Error de truncament. Aproximacions numèriques del model matemàtic (algorismes).
- Error experimental. Mesures incorrectes o dolentes.
  - Errors aleatoris. Les mesures.
  - Errors sistemàtics. Cal·libració incorrecte.
  - Errors aberrants. Per descomptat també hi ha el factor humà. Un error per descuit o ignorància.
- Error d'arrodoniment. Error acumulat a causa de l'execució del nostre model on les operacions descrites per l'algorisme es fan amb un nombre finit de dígits.

Errors

Aquest curs principalment,ens centrarem en estudiar els errors en el procés de càlcul

De truncament Convertir un procés infinit en finit.

$$e^x \approx 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!}$$

D'arrodoniment Deguts a l'aritmètica de punt flotant de l'ordinador.

Menys importants que els de truncament, però poden esdevenir catastròfics.

Fonts d'error

Nosaltres ens centrarem en estudiar els errors en el procés de càlcul, errors produïts en interrompre els processos infinits, errors d'arrodoniment i de truncament.

Per a això, farem el següent:

- Definir el que entenem per error.
- Analitzar els errors associats a l'aritmètica flotant.
- Analitzar algoritmes que minimitzen aquest error.

Error absolut

Notem per x, el valor exacte i per $\tilde{x}$, un valor aproximat

Definició

$$\Delta x = x - \tilde{x} \qquad (1)$$

Pràctica

$x = 1/3, \quad \tilde{x} = 0.3333$
$x = \pi, \quad \tilde{x} = 3.141$

Un defecte que té aquesta definició és que no considera la magnitud del valor.

Error relatiu

Notem per x, el valor exacte i per $\tilde{x}$, un valor aproximat

Definició

$$\epsilon_x = \frac{\Delta x}{x} \qquad (2)$$

Pràctica

$x = 1/3, \quad \tilde{x} = 0.3333$

$x = \pi, \quad \tilde{x} = 3.141$

Error relatiu aproximat: $\epsilon_{\tilde{x}} = \dfrac{\Delta x}{\tilde{x}}$.

Error relatiu percentual: l'error relatiu en tant per 100.

Exacte i precís

Definició: xifres decimals correctes

Direm que $\tilde{x}$ és una aproximació a x amb **d** xifres decimals correctes si d és el nombre natural més gran tal que

$$|x - \tilde{x}| < 0.5 \cdot 10^{-d} \tag{3}$$

Definició: xifres significatives correctes

Direm que $\tilde{x}$ és una aproximació a x amb **t** xifres significatives si t és el nombre natural més gran tal que

$$\frac{|x - \tilde{x}|}{|x|} < 0.5 \cdot 10^{-t} \tag{4}$$

Estimacions

Per a $x = \pi$ i $\tilde{x} = 3.141$, tenim

$$\Delta x = 0.00059265\ldots \qquad \epsilon_x = 0.000188647\ldots$$

$$|\Delta x| \leq 0.6 \cdot 10^{-3}\,, \quad x = 3.141 \pm 0.6 \cdot 10^{-3}\,.$$

$$|\epsilon_x| \leq 0.2 \cdot 10^{-3} = 0.02\%\,, \quad x = 3.141 \cdot (1 \pm 0.02\%)\,.$$

Estimacions (II)

Exemple: Sigui $x = \sqrt{2} = 1.414213562...$ i $\bar{x} = 1.414$. Aleshores

$$e_a(\sqrt{2}) = -0.0002135...$$

i

$$e_r(\sqrt{2}) = \frac{-0.0002135...}{\sqrt{2}} \simeq \frac{-0.0002135}{1.414} = -0.00015099...$$

Per l'exemple anterior tenim que

$$\varepsilon_a = 0.00022, \quad \varepsilon_r = 0.00016$$

Autoavaluació

Exercici 1 Calculeu l'error absolut, l'error relatiu i l'error relatiu aproximat de les quantitats:

$$\begin{aligned} x &= 9234.567, & \widetilde{x} &= 9234.564; \\ x &= 0.634, & \widetilde{x} &= 0.631\,. \end{aligned}$$

Què s'observa?

Exercici 2 Calculeu l'error absolut, l'error relatiu, les xifres correctes de les quantitats:

$$\begin{aligned} x &= 1/3, & \widetilde{x} &= 0.3333\,, \\ x &= 1/3, & \widetilde{x} &= 0.3334\,, \end{aligned}$$

Autoavaluació

Exercici 3 Calculeu les xifres significatives de les quantitats:

$$\begin{aligned} x &= 10000, & \tilde{x} &= 9998\,, \\ x &= 10000, & \tilde{x} &= 9999.99998\,, \\ x &= 0.0000025, & \tilde{x} &= 0.0000018\,, \end{aligned}$$

Exercici 4 Calculeu l'error absolut i l'error relatiu de les aproximacions següents:

a) $x = \pi\,, \quad \tilde{x} = 22/7$
b) $x = e\,, \quad \tilde{x} = 2.718$
c) $x = 10^{\pi}\,, \quad \tilde{x} = 1400$
d) $x = e^{10}\,, \quad \tilde{x} = 22000$

Errors d'arrodoniment

La representació decimal d'un **nombre real** es redueix per tal representar/usar els nombres reals a l'ordinador o en càlculs manuals.

La representació decimal d'un nombre és pot reduir per tall o per arrodoniment a un nombre finit de dígits.

Exemple.

Representar $\dfrac{2}{3}$ per una expressió decimal de 5 dígits.

Arrodonir nombres decimals

Sigui x qualsevol nombre decimal positiu de la forma

$$x = 0.d_1 d_2 \dots d_{n-1} d_n d_{n+1} \dots d_m$$

llavors $\widetilde{r_x}$, l'arrodoniment de x a n xifres decimals ($n < m$) depèn del valor del dígit $n+1$.

$$\widetilde{r_x} = 0.d_1 d_2 \dots d_{n-1} d \,, \tag{5}$$

$$d = \begin{cases} d_n & \text{si} \quad d_{n+1} \in \{0,1,2,3,4\}\,, \\ d_n + 1 & \text{si} \quad d_{n+1} \in \{5,6,7,8,9\}\,. \end{cases} \tag{6}$$

Estimació error arrodoniment

Teorema

Si el nombre x s'arrodoneix, i $\widetilde{r_x}$ és el seu valor arrodonit a n dígits, aleshores la fita de l'error absolut és

$$|x - \widetilde{r_x}| \leq \frac{1}{2} \cdot 10^{-n}. \tag{7}$$

Exercici.

Representeu els nombres .1735499 .99995000 i .4321609 amb quatre dígits per arrodoniment.

Tallar nombres decimals

La aproximació per tall del nombre x a n dígits ($n < m$) és el nombre $\widehat{t_x}$ obtingut en descartat tots els dígits posteriors al dígit n.

$$x = 0.d_1 d_2 \ldots d_n \ldots d_m \xrightarrow[\text{n dígits}]{tallar} \widehat{t_x} = 0.d_1 d_2 \ldots d_n \, , \quad (8)$$

Estimació error tallar

Teorema

Si el nombre x es talla, i $\widehat{t_x}$ és el seu valor aproximat a n dígits, aleshores la fita de l'error absolut és

$$\left|x - \widehat{t_x}\right| \leq 10^{-n}. \tag{9}$$

Exercici.

Representeu els nombres .1735499 .99995000 i .4321609 amb quatre dígits per tall.

Precisió implícita

Per escriure una mesura com un nombre decimal (o binari o $\cdots$), hi ha un nivell de precisió implícit, si no es diu el contrari 0.5 unitats en l'última posició escrita. Altrament, és convenient escriure una dada amb l'error màxim explícitament.

- Si escrivim 23.4567 hem d'entendre 23.4567 ± 0.0005
- Altrament escriuriem la fita de l'error, 23.4567 ± 0.0012
- La precisió implícita de `Matlab` sempre és la mateixa.

Quan es mostri un valor aproximat, cal que la precisió reflecteixi l'exactitud (error absolut).

Errors de truncament. De discretització

Els errors de truncament sorgeixen en el cas d'aproximar un procés infinit per un de finit.

$$e^x \approx 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!}$$

Els errors de discretització o en substituir una expressió contínua per una discreta. El procés numèric és generalment una aproximació del model matemàtic obtingut com a funció d'un paràmetre de discretizació, que serà notat per h, i suposarem positiu. Si, quan h tendeix a 0, el procés numèric torna la solució del model matemàtic, direm que el procés numèric és convergent.

Errors de truncament

truncament procès infinit

$$S = \sum_{n=1}^{+\infty} (-1)^n a_n \approx S_k = \sum_{n=1}^{k} (-1)^n a_n$$

Leibniz va obtenir la següent sèrie matemàtica (1682):

$$\sum_{n=0}^{\infty} \frac{(-1)^n}{2n+1} = 1 - \frac{1}{3} + \frac{1}{5} - \cdots = \frac{\pi}{4}$$

Errors de discretització

Por definició la derivada d'una funció $f(x)$ és:

$$f'(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}$$

llavors podem fer les aproximacions numèriques ($h > 0$):

discretització

$$f'(x_0) = \frac{f(x_0+h) - f(x_0)}{h} + \frac{f''(\xi)}{2} h, \quad \text{fórmula de Taylor}$$

llavors

$$f'(x_0) \approx \frac{f(x_0+h) - f(x_0)}{h} \qquad e_d = \mathcal{O}(h)$$

Errors de discretització (II)

El procés numèric és generalment una aproximació del model matemàtic obtingut com a funció d'un paràmetre de discretizació, que serà notat per h, i suposarem positiu. Si, quan h tendeix a 0, el procés numèric torna la solució del model matemàtic, direm que el procés numèric és convergent. A més, si l'error (absolut o relatiu) es pot fitar en funció de h, de la forma

$$e_d \leq Ch^p \,, \;\; C > 0 \,, \;\; p > 0 \Longleftrightarrow e_d = \mathcal{O}(h^p)$$

es diu que *el mètode és convergent d'ordre p* i escrivim

Propagació de l'error: funcions

Si $f : \mathbb{R} \to \mathbb{R}$ és una funció derivable, x un nombre real, $\tilde{x}$ una aproximació de x amb fita d'error ϵ, $x = \tilde{x} \pm \epsilon$, el teorema del valor mig diu

$$|f(x) - f(\tilde{x})| = |f'(\xi)|\,|x - \tilde{x}|\,, \qquad |\tilde{x} - \xi| \leq \epsilon.$$

Així, una manera de valorar l'efecte que tenen els errors en les dades d'entrada en el càlcul de $y = f(x)$ seria:

Fórmula de la propagació de l'error absolut

$$|\Delta y| \approx |f'(\tilde{x})|\,\epsilon. \tag{10}$$

Propagació de l'error: funcions

Si $g : \mathbb{R}^2 \to \mathbb{R}$ és una funció diferenciable i $x_1 = \tilde{x}_1 \pm \epsilon_1$ i $x_2 = \tilde{x}_2 \pm \epsilon_2$, llavors

FGPE en dues variables

$$|\Delta g| \approx \left| \frac{\partial g(\tilde{x}_1, \tilde{x}_2)}{\partial x_1} \right| |\epsilon_1| + \left| \frac{\partial g(\tilde{x}_1, \tilde{x}_2)}{\partial x_2} \right| |\epsilon_2|. \tag{11}$$

Si fem ús d'una funció real de dues variables serà possible fitar l'error propagat per les operacions elementals.

Propagació de l'error: operacions

Suma, $g(x_1, x_2) = x_1 + x_2$

$$x_1 + x_2 = (\tilde{x}_1 + \tilde{x}_2) \pm (\delta x_1 + \delta x_2) \tag{12}$$

Resta, $g(x_1, x_2) = x_1 - x_2$

$$x_1 - x_2 = (\tilde{x}_1 - \tilde{x}_2) \pm (\delta x_1 + \delta x_2) \tag{13}$$

Les cotes dels errors absoluts es sumen en les operacions de sumar i restar nombres reals.

Les cotes dels errors relatius es sumen en les operacions de multiplicar i dividir nombres reals.

Producte, $g(x_1, x_2) = x_1 \cdot x_2$

$$|\delta g| \approx |\tilde{x}_2||\delta x_1| + |\tilde{x}_1||\delta x_2|\,, \qquad \left|\frac{\delta g}{g}\right| \approx \left|\frac{\delta x_1}{\tilde{x}_1}\right| + \left|\frac{\delta x_2}{\tilde{x}_2}\right|. \tag{14}$$

Divisió, $g(x_1, x_2) = x_1/x_2$

$$|\delta g| \approx \left|\frac{1}{\tilde{x}_2}\right| |\delta x_1| + \left|\frac{\tilde{x}_1}{\tilde{x}_2^2}\right| |\delta x_2|\,, \qquad \left|\frac{\delta g}{g}\right| \approx \left|\frac{\delta x_1}{\tilde{x}_1}\right| + \left|\frac{\delta x_2}{\tilde{x}_2}\right|. \tag{15}$$

FGPE

Sigui $g : \mathcal{D} \to \mathbb{R}$, $\mathcal{D}$ una regió de $\mathbb{R}^n$,
g una funció diferenciable en un entorn del vector $\tilde{x}$
$x = \tilde{x} \pm \Delta x$, amb $\Delta x = (\Delta x_1, \ldots, \Delta x_n)^t$.

Error absolut propagat

$$|\Delta y| = |y - \tilde{y}| \approx \sum_{i=1}^{n} \left| \frac{\partial g(\tilde{x})}{\partial x_i} \right| \Delta x_i \,. \qquad (16)$$

La fórmula (16) de l'error absolut propagat s'obté de la fórmula de Taylor aplicada a $y = g(x)$ i $\tilde{y} = g(\tilde{x})$.

Nombres de condició

Error relatiu propagat

$$\left|\frac{\Delta y}{\tilde{y}}\right| \approx \sum_{i=1}^{n} \left|\frac{\tilde{x}_i}{g(\tilde{x})} \frac{\partial g(\tilde{x})}{\partial x_i}\right| \left|\frac{\Delta x_i}{\tilde{x}_i}\right| . \qquad (17)$$

L'expressió (17) s'obté dividint per $\tilde{y}$ i multiplicant i dividint per $\tilde{x}_i$. Els n valors

$$\left|\frac{\tilde{x}_i}{g(\tilde{x})} \frac{\partial g(\tilde{x})}{\partial x_i}\right| \qquad (18)$$

s'anomenen números de condició o factors de propagació. Aquests donen una mesura de quan un problema és mal condicionat.

Estabilitat numèrica i problemes ben condicionats

Estabilitat numèrica

Un algorisme el classificarem com **numèricament estable** si un error, no creix gaire en el procès de càlcul.

L'estabilitat numèrica es veu afectada pel nombre de xifres significatives, poques xifres o la pèrdua en pasos intermitjos del càlcul disminueix la fiabilitat dels resultats obtinguts.

Algorismes amb cancel·lació

La pèrdua de xifres significatives per cancel·lació, que es produeix en restar dos nombres molt propers. La situació es pot resumir en

$$g(x+\delta) - g(x) \quad \text{amb } |\delta| \ll 1\,; \tag{19}$$

Exemple.
Les solucions de $x^2 - 18x + 1 = 0$ són $x_{1,2} = 9 \pm \sqrt{80}$.
Si $\sqrt{80} = 8.9443 \pm 0.5 \cdot 10^{-4}$ llavors $x_1 = 17.9443 \pm 0.5 \cdot 10^{-4}$, té 6 xifres significactives, mentre que $x_2 = 0.0557 \pm 0.5 \cdot 10^{-4}$, només en té 3.

Inestabilitat numèrica

Sense rigor, diem que un procés numèric és inestable quan els petits errors que es produeixen en un dels seus estadis s'agranda en etapes posteriors, fins a tal punt que no podem fiar-nos del càlcul global.

Exemple.

Per calcular les integrals $I_n = \int_0^1 x^n e^{x-1} dx\,, \quad n \geq \ 1$, dispossem de dos mètodes iteratius diferents:

$$\text{a)} \quad I_{n-1} \ = \frac{1-I_n}{n}, \qquad n \geq 2 \qquad \text{on } I_{50} = 0\,,$$

$$\text{b)} \quad I_n \qquad = 1 - nI_{n-1}, \quad n \geq 2 \quad \text{on } I_1 = 1/e\,.$$

Sensibles a les condicions inicials

Molts problemes són especialment sensibles a les dades inicials, independentment dels errors d'arrodoniment i de l'algorisme emprat.

Exemple. Polinomi de Wilkinson

Sigui $p(x) = (x-1)(x-2)(x-3)...(x-10)$, el polinomi amb arrels els deu primers nombres naturals, definim el polinomi $q(x) = p(x) + \dfrac{1}{2^{13}}\,x^9$, modificant lleugerament el coeficient de x^9 respecte de $p(x)$. Com haurien de ser les arrels del polinomi $q(x)$? Calculeu-les.

Sensibles a les condicions inicials

Són problemes on la solució depèn de manera molt sensible de les dades. Si petites variacions de les dades provoquen grans variacions en la solució, es diu que el problema està mal condicionat.

Exemple

Resoleu els sistemes

$$\left\{ \begin{array}{rcrcr} 2x & - & 4y & = & 1 \\ -2.998x & + & 6.001y & = & 2 \end{array} \right. \qquad \left\{ \begin{array}{rcrcr} 2x & - & 4y & = & 1 \\ -3x & + & 6.001y & = & 2 \end{array} \right.$$

7
Representació de nombres

Representació de nombres (punt fix)

$$x = \pm d_1 d_2 d_3 \dots d_n \,.\, d_{n+1} d_{n+2} \dots d_{n+m}$$

Representa el nombre real x expressat en base 10, on cada $d_i \in \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$.

Exemples

$2345 = (2 \times 10^3) + (3 \times 10^2) + (4 \times 10^1) + (5 \times 10^0) = 2000 + 300 + 40 + 5.$

$45.67 = (4 \times 10^1) + (5 \times 10^0) + (6 \times 10^{-1}) + (7 \times 10^{-2}) = 40 + 5 + 0.6 + 0.07,$

L'aritmètica es realitza desplaçant el punt decimal.

Representació de nombres (punt flotant)

$$x = \pm 0.d_1 d_2 d_3 \dots d_n \times 10^e \qquad d_1 \neq 0.$$

Representa el nombre real x expressat en base 10 i notació de punt flotant, on e és un nombre enter i cada $d_i \in \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$.

Exemples

Els nombre -15.24 en punt flotant és $-0.1524 \cdot 10^2$;
el nombre 0.000617 en punt flotant és $0.617 \cdot 10^{-3}$;
el nombre $4.274952 \cdot 10^{15}$ és $0.4274952 \cdot 10^{16}$;
i -6543219 en punt flotant és $-0.6543219 \cdot 10^7$.

Nombres binaris

El sistema binari, és com el decimal però només fan ús dels dígits 0 i 1.

Exemples

$$
\begin{aligned}
(101.1101)_{(2)} &= (1\cdot 2^2 + 0\cdot 2^1 + 1\cdot 2^0 + 1\cdot 2^{-1} + 1\cdot 2^{-2} + 0\cdot 2^{-3} + 1\cdot 2^{-4})_{10} \\
&= (4+1+1/2+1/4+1/32)_{10} = (5.78125)_{10}
\end{aligned}
$$

$$97 = (1000011)_{dos},\ restesdividirper2$$

$$0.7 = (0.1\overline{0110})_{dos},\ partenterademultiplicarper2$$

8

Representació de nombres a l'ordinador

Norma $754 - 1985$

L'any 1985, l'*Institute for Electrical and Electronic Engineers (IEEE)* va publicar l'informe

Binary Floating Point Arithmetic Standard $754 - 1985$,

en el que s'especifiquen normes per representar nombres en punt flotant amb precisió simple, doble i extensa. L' informe va ser revisat i actualitzat l'any 2008, *IEEE Std 754-2008*.

Avui en dia, quasi tots els fabricants d'ordinadors han acceptat aquesta norma; per tant l'ordinador emmagatzema no el nombre real x si no una aproximació binària (octal o hexadecimal) en punt flotant a x.

El conjunt F(β, t, L, U)

El conjunt de nombres en punt flotant representables a l'ordinador el designarem per F(β,t,L,U), on β representa la base, t la precisió (o nombre de dígits representats) i el interval $[L, U]$ és el rang de l'exponent e.

Per a tot nombre real x expressat en el conjunt $F(\beta, t, L, U)$ existeixen $\boldsymbol{t}$ xifres i un exponent $\boldsymbol{e}$ tal que

$$fl(x) = \pm\left(\frac{d_1}{\beta} + \frac{d_2}{\beta^2} + \ldots + \frac{d_t}{\beta^t}\right) \cdot \beta^e\,,$$

amb dígits $d_i \in \mathbb{N}$ tal que $0 \leq d_i < \beta$ per a tot $i = 1 \div t$; i exponent $L \leq e \leq U$.

El conjunt $\mathrm{F}(\beta, \mathrm{t}, \mathrm{L}, \mathrm{U})$

La quantitat

$$\pm\left(\frac{d_1}{\beta} + \frac{d_2}{\beta^2} + \ldots + \frac{d_t}{\beta^t}\right)$$

s'anomena fracció o part fraccionària del nombre x.

Si exigim $d_1 \neq 0$ per a $x \neq 0$, resulta que $\beta^{-1} \leq |f| < 1$ i es diu que la representació és normalitzada.

La quantitat $m = d_1 d_2 d_3 \ldots d_t$ s'anomena mantissa.

32 bits

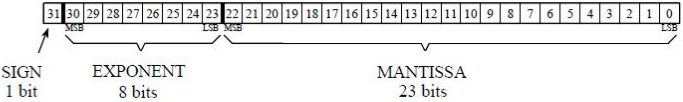
La representació en punt flotant amb simple precisió ocupa 1 paraula; té 32 dígits binaris, assignats de la manera següent,

signe del nombre real x	1 bit
exponent, (enter)	7 bits
mantissa, (real)	23 bits

Un nombre real x en simple precisió es representa en 32 bits binari per

$$fl(x) = (-1)^s \times M \times 2^{c-127}.$$

amb $M = 1.m_{22}m_{21}\dots m_1 m_0 = 1 + f$



Example 1

0 00000111 11000000000000000000000

$\downarrow$ + $\quad\downarrow$ 7 $\quad\downarrow$ 0.75

$$+1.75 \times 2^{(7-127)} = +1.316554 \times 10^{-36}$$

Example 2

1 10000001 01100000000000000000000

$\downarrow$ – $\quad\downarrow$ 129 $\quad\downarrow$ 0.375

$$-1.375 \times 2^{(129-127)} = -5.500000$$

FIGURE 4-2
Single precision floating point storage format. The 32 bits are broken into three separate parts, the sign bit, the exponent and the mantissa. Equations 4-1 and 4-2 shows how the represented number is found from these three parts. MSB and LSB refer to "most significant bit" and "least significant bit," respectively.

64 bits

La representació en punt flotant amb doble precisió ocupa 2 paraules; té 64 dígits binaris, assignats de la manera següent,

s signe del nombre real x	1 bit
c exponent, (enter)	11 bits
m mantissa, (real)	52 bits

Un nombre real x en doble precisió es representa en 64 bits binari (`Matlab`) per:

$$fl(x) = (-1)^s 2^{c-1023}(1+f).$$

Té entre 15 i 16 dígits decimals de precisió.

64 bits

La condició de 11 bits per l'exponent significa que $|e| \leq 2^{11} - 1 = 2047$, restant 1023 obtenim el rang $(-1023, 1024)$ per a l'exponent. Si s'acorda la representació en que el primer dígit binari de la mantissa sigui 1, podem tindre matises de 53 dígits binaris, que es corresponen almenys 15 xifres decimals de precisió.

El nombre positiu més petit és $s = 0$, $c = 1$ i $f = 0$ que es correspon al nombre decimal $2^{-1022} \cdot (1 + 0) \approx 0.2225 \cdot 10^{-307}$. El nombre positiu més gran és $s = 0$, $c = 2046$ i $f = 1 - 2^{-52}$ que es correspon al nombre decimal $2^{1023} \cdot (1 + f) \approx 0.17977 \cdot 10^{309}$.

Aritmètica a $F(\beta, t, L, U)$

Imaginem un ordinador $F(10, 5, 0, 127)$, i els nombres $x = .31426 \cdot 10^3$ i $y = .92577 \cdot 10^5$.

$$
\begin{array}{rrrr}
x \times y = & .2909324802 \cdot 10^8 & x \otimes y = & .29093 \cdot 10^8 \\
x + y = & .9289126 \cdot 10^5 & x \oplus y = & .92891 \cdot 10^5 \\
x - y = & -.92262740 \cdot 10^5 & x \ominus y = & -.92263 \cdot 10^5 \\
x \div y = & .3394579647 \cdot 10^{-2} & x \oslash y = & .33946 \cdot 10^{-2}
\end{array}
$$

En aquests resultats l'error relatiu és de $8.5 \cdot 10^{-6}$, $2.3 \cdot 10^{-6}$, $2.8 \cdot 10^{-6}$, $6.0 \cdot 10^{-6}$, respectivament, tots per sota de 10^{-5}.

Èpsilon de la màquina

Si l'ordinador té l'aritmètica $F(\beta, t, L, U)$, llavors

$$fl(x) = x(1+\delta) \qquad |\delta| \leq \epsilon\,,$$

amb $\epsilon = \frac{1}{2}\beta^{1-t}$ si arrodonim i $\epsilon = \beta^{1-t}$ si tallem.

La precisió d'una aritmètica de coma flotant es caracteritza per l'**èpsilon de la màquina**, no és el nombre més petit representable, però dóna una mesura relativa de fins a on dos nombres molt pròxims seran diferents.

En Matlab és $\epsilon = 2.2204e - 016$.

Evitar la propagació dels errors

Es recomana, minimitzar el nombre d'operacions, reordenar les operacions i replantejar el problema en altres termes.

equació de segon grau

Resoldre l'equació $x^2 + 62.10x + 1 = 0$ treballant amb quatre dígits i arrodonint

regla de horner

Avaluar el polinomi $P(x) = x^3 - 6.1x^2 + 3.2x + 1.5$ per $x = 4.71$ fent ús d'una aritmètica de tres dígits.

9
Aritmètica de Matlab

Aritmètica de Matlab

Cal llegir els documents:

- Aritmètica en punt flotant by Cleve Moler
  Fall96Cleve.pdf
- MathWorks Documentation Center,
  floating-point-numbers.html
- Punto Flotante,
  puntoflotante.org

Autoavaluació

Exercici 5 Realitzeu les operacions aritmètiques:

$$\frac{4}{5}+\frac{1}{3}\,; \quad \frac{4}{5}\cdot\frac{1}{3}\,; \quad \left(\frac{1}{3}-\frac{3}{11}\right)+\frac{3}{20}\,; \quad \left(\frac{1}{3}+\frac{3}{11}\right)-\frac{3}{20}\,;$$

a) Fent ús d'una aritmètica de tres xifres i tallant els nombres.

b) Fent ús d'una aritmètica de tres xifres i arrodonint els nombres.

c) Calculeu els errors relatius dels apartats *a)* i *b)*.

Autoavaluació

Exercici 6 Calculeu: $$\sum_{k=1}^{6} \frac{1}{3^k} \quad \text{i} \quad \sum_{k=1}^{6} \frac{1}{3^{(7-k)}}$$

a) Fent ús de l'aritmètica de tres xifres arrodonint.

b) Fent ús de l'aritmètica de quatre xifres arrodonint.

c) Per què donen diferent? Calculeu en cada cas l'error relatiu percentual.