Computació Numèrica

Tema 1. Conceptes bàsics

M. Àngela Grau Gotés

Departament de Matemàtiques
Universitat Politècnica de Catalunya · BarcelonaTech.

20 de febrer de 2023

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

Introducció

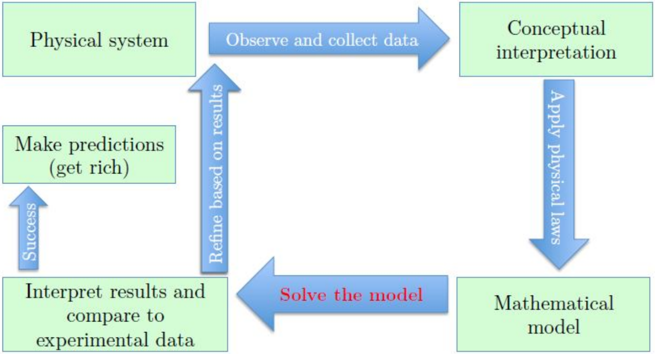
Durant els segles XX i XXI models matemàtics avançats s'han aplicat en diferents àrees de coneixement com l'enginyeria, la medicina, l'economia o les ciències socials. Sovint, les aplicacions generen problemes matemàtics que per la seva complexitat no poden ser resolts de manera exacta.

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

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.

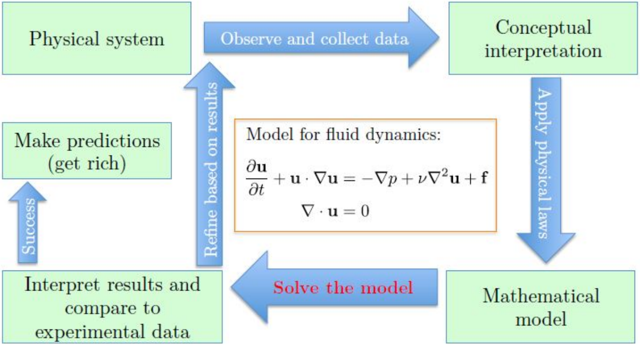
Els models resultants quasi mai no poden ser resolts per complet utilitzant mètodes (llapís i paper) d'anàlisi. La simulació en un ordinador ens permet interpretar els resultats i comparar-los amb les dades experimentals.

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

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

- 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 $\text{erf}\left(\frac{\epsilon}{\sigma\sqrt{2}}\right)$ 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: exacte, 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).

En aquest curs treballarem amb dos tipus d'algorismes

Directes Obtenen la solució en un nombre finit de pasos.

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 $\sqrt{2}$.*

$$x_k = \frac{1}{2}\left(x_{k-1} + \frac{2}{x_{k-1}}\right) \qquad k \geq 1 \ i \ x_0 = 3\,.$$

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

Fonts d'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/discretització. Aproximacions discretes i finites 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.

Algorismes i error

Aquest curs principalment,ens centrarem en estudiar els errors en el procés de càlcul, l'algorisme emprat per a resoldre:

De truncament Convertir un procés infinit en finit.

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

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

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

Algorismes i 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.
- Presentar algoritmes que minimitzen aquest error.

2
Error absolut. Error relatiu

Error absolut

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

Definició

$$e_a(x) = x - \widetilde{x} \qquad (1)$$

$$\Delta x = |x - \widetilde{x}| \qquad (2)$$

Pràctica

$$x = 1/3, \qquad \widetilde{x} = 0.3333$$
$$x = \pi, \qquad \widetilde{x} = 3.141$$

Usualment es treballa amb fites de l'error absolut, per ϵ_a es coneix una fita de l'error absolut.

Un defecte que té el concepte és que no considera la magnitud del valor i depen de les unitats.

Error relatiu

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

Definició

$$e_r(x) = \frac{\Delta x}{|x|} \qquad (3)$$

Pràctica

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

Error relatiu aproximat: $e_r(\widetilde{x}) = \dfrac{\Delta x}{|\widetilde{x}|}$.

Fita de l'error relatiu: ϵ_x.

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

Exacte i exactitud

Definició: xifres decimals correctes (accurate)

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

$$|x - \widetilde{x}| < 0.5 \cdot 10^{-d} \tag{4}$$

Definició: xifres significatives correctes (precision)

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

$$\frac{|x - \widetilde{x}|}{|x|} < 0.5 \cdot 10^{-t} \tag{5}$$

Whats the difference between precision and accuracy?

Estimacions - Cotes error

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)

Sigui $x = \sqrt{2} = 1.414213562\ldots$ i $\widetilde{x} = 1.414$, aleshores

$$e_a(\sqrt{2}) = 0.0002135\ldots \qquad e_r(\sqrt{2}) = 0.00015099\ldots$$

i les fites podrien ser

$$\epsilon_a = 0.00022, \qquad \epsilon_r = 0.00016.$$

Autoavaluació

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

$$x = 9234.567, \quad \widetilde{x} = 9234.564;$$
$$x = 0.634, \quad \widetilde{x} = 0.631\,.$$

Què s'observa?

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

$$x = 1/3, \quad \widetilde{x} = 0.3333\,,$$
$$x = 1/3, \quad \widetilde{x} = 0.3334\,,$$

Exercici 3 Calculeu les xifres significatives de les quantitats:

$$x = 10000, \quad \widetilde{x} = 9998\,,$$
$$x = 10000, \quad \widetilde{x} = 9999.99998\,,$$
$$x = 0.0000025, \quad \widetilde{x} = 0.0000018\,,$$

3.1
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.

Sigui x qualsevol nombre decimal positiu de la forma

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

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

$$\widetilde{r}_x = 0.d_1 d_2 \ldots d_{n-1} d\,, \tag{6}$$

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

Errors d'arrodoniment

Estimació error arrodonir

Teorema

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

$$|x - \tilde{r}_x| \leq \frac{1}{2} \cdot 10^{-n}. \tag{8}$$

Exercici.

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

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.

$$x = 0.d_1 d_2 \ldots d_n \ldots d_m \xrightarrow[\text{n dígits}]{tallar} \widehat{t_x} = 0.d_1 d_2 \ldots d_n\,, \tag{9}$$

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

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

Exercici.

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

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 la dada amb l'error màxim explícitament.

- Si escrivim 23.4567 hem d'entendre 23.4567 ± 0.00005
- 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/relatiu).

3.2
Errors de truncament

Errors de truncament (I)

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

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

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 (II)

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 (I)

El procés numèric és generalment una aproximació del model matemàtic obtingut com a funció d'un paràmetre de discretització, que serà notat per h, i suposarem positiu. Si, quan h tendeix a 0, el procés numèric retorna 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 $e_d = \mathcal{O}(h^p)$.

Errors de discretització (II)

Por definició la derivada d'una funció $f(x)$ en un punt x_0 és:

$$f'(x_0) = \lim_{h \to 0} \frac{f(x_0 + h) - f(x_0)}{h}$$

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

Derivació numèrica

De la fórmula de Taylor per $f(x_0 + h)$ s'obté:

$$f'(x_0) = \frac{f(x_0 + h) - f(x_0)}{h} + \frac{f''(\xi)}{2} h,$$

llavors s'escriu:

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

3.3
Propagació de l'error

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

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

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

$$|\Delta f| \approx |f'(\tilde{x})|\,\epsilon. \tag{11}$$

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

$$|\Delta g| \approx \left|\frac{\partial g(\tilde{x}_1, \tilde{x}_2)}{\partial x_1}\right| |\epsilon_1| + \left|\frac{\partial g(\tilde{x}_1, \tilde{x}_2)}{\partial x_2}\right| |\epsilon_2|. \tag{12}$$

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

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

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

$$x_1 - x_2 = (\tilde{x}_1 - \tilde{x}_2) \pm (\delta x_1 + \delta x_2) \tag{14}$$

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

Propagació de l'error: operacions

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

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

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

Divisió, $g(x_1, x_2) = x_1 / x_2$

$$|\delta g| \approx \left|\frac{1}{\widetilde{x}_2}\right| |\delta x_1| + \left|\frac{\widetilde{x}_1}{\widetilde{x}_2^2}\right| |\delta x_2|\,, \qquad \left|\frac{\delta g}{g}\right| \approx \left|\frac{\delta x_1}{\widetilde{x}_1}\right| + \left|\frac{\delta x_2}{\widetilde{x}_2}\right|. \tag{16}$$

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

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

La fórmula (17) de l'error absolut propagat s'obté de la fórmula de Taylor aplicada a $y = g(x)$ i $\tilde{y} = g(\tilde{x})$.

Error relatiu propagat

$$\left|\frac{\Delta y}{\widetilde{y}}\right| \approx \sum_{i=1}^{n} \left|\frac{\widetilde{x}_i}{g(\widetilde{x})} \frac{\partial g(\widetilde{x})}{\partial x_i}\right| \left|\frac{\Delta x_i}{\widetilde{x}_i}\right| . \tag{18}$$

L'expressió (18) s'obté dividint per $\widetilde{y}$ i multiplicant i dividint per $\widetilde{x}_i$. Els n valors

$$\left|\frac{\widetilde{x}_i}{g(\widetilde{x})} \frac{\partial g(\widetilde{x})}{\partial x_i}\right| \tag{19}$$

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

4
Aritmètica de punt/coma flotant

What Every Computer Scientist Should Know About Floating-Point Arithmetic

Representació de nombres

Existeixen diverses tècniques per a representar els nombres reals en un ordinador.

Només un cop d'ull: tant als conceptes generals com en la representació de nombres en de MATLAB® .

A Glimpse into Floating-Point Accuracy

Floating points. IEEE Standard unifies arithmetic model

4.1
Representació de nombres

Representació de nombres (punt fix)

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

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/coma 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 **precisió n** en notació de punt/coma flotant **normalitzada**,
on l'exponent **e** és un nombre enter i cada dígit d_i un enter entre 0 i 9.

Exemples

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

$$
\begin{aligned}
(101.1101)_{(2)} &= (1\cdot 2^2 + 0\cdot 2^1 + 1\cdot 2^0 + 1\cdot 2^{-1} + 1\cdot 2^{-2} + 0\cdot 2^{-3} + 1\cdot 2^{-4})_{10}\\
&= (4+1+1/2+1/4+1/32)_{10} = (5.78125)_{10}\\
97 &= (1000011)_{dos},\ \text{restes dividir per 2}\\
0.7 &= (0.1\overline{0110})_{dos},\ \text{part entera de multiplicar per 2}
\end{aligned}
$$

4.2
Nombres a l'ordinador

Història

Els anys 60 i 70 les operacions amb nombres reals tenien implementacions diferents en cada ordinador: format, precissió, arrodoniment, gestió d'execepcions, etc. D'aquesta manera era molt difícil d'escriure codi portàtil.

El 1982 l' *Institute of Electrical and Electronics Engineers* va definir l'estàndar IEEE-754 i el va implementar en els processadors intel 8087. En tots els ordinadors que el tenien implementat, el programes obtenien els mateixos resultats.

El 2002 l'estàndar IEEE-754 es va implementar universalment en tots els ordinadors de propòsit general.

Cleve's Corner: Cleve Moler on Mathematics and Computing,
Floating Point Arithmetic Before IEEE 754

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,

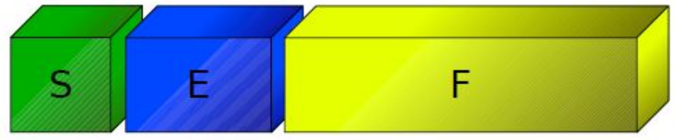
en el que s'especifiquen normes per representar nombres en punt/coma flotant amb precisió simple, doble i extensa. L' informe va ser revisat i actualitzat l'any 2008, *IEEE Std 754-2008*.

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 coma flotant a x.

Cleve's Corner: Cleve Moler on Mathematics and Computing,
Floating Point Numbers

Format de coma flotant

El format de coma flotant, fa ús de 3 camps binaris per a la representació: signe (S), exponent (E) i fracció (F).

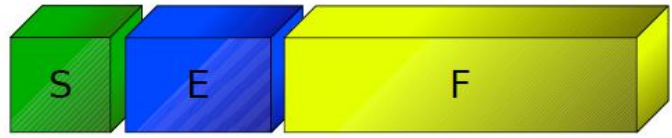


El signe ocupa un bit,
0 per a nombres positius i 1 per a nombres negatius.

Format de coma flotant (S)

Signe

El format de coma flotant, fa ús de 3 camps binaris per a la representació: signe (S), exponent (E) i fracció (F).

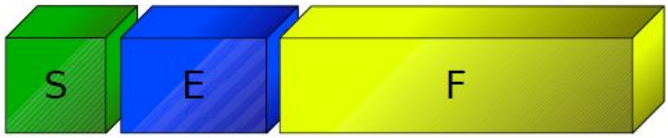


El signe ocupa un bit,
0 per a nombres positius i 1 per a nombres negatius.

Format de coma flotant (E)

Exponent

El format de coma flotant, fa ús de 3 camps binaris per a la representació: signe (S), exponent (E) i fracció (F).



L'exponent es guarda desplaçat,
ocupa 7 bits, se li suma 127 en precissió simple
ocupa 11 bits, se li suma 1023 en precissió doble.

Format de coma flotant (F)

Fracció

El format de coma flotant, fa ús de 3 camps binaris per a la representació: signe (S), exponent (E) i fracció (F).

La fracció té un 1 ímplicit a l'esquerra, és a dir la mantissa fa ús d'una notació científica normalitzada.

64 bits

La representació en coma flotant amb doble precisió

$$fl(x) = (-1)^s \times M \times 2^{c-1023}$$

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

Si es demana $M = 1.m_{51}m_{50}\dots m_1m_0 = 1 + f$ augmenta en un bit la precissió.

Té entre 15 i 16 dígits decimals de precisió.

Precisió de l'ordinador/software

La precisió d'un ordinador dependrà del fabricant i del tipus de variable que es defineixi; la unitat d'informació ve donada pel nombre de dígits binaris o longitud de la paraula (word): *1 byte = 8 bits, 1 word = 2 bytes(PC), 4 bytes(VAX), 8 bytes(CRAY)*.

La longitud de les paraules imposa una restricció sobre la precisió amb la un ordinador pot representar nombres reals. Això vol dir que l'ordinador emmagatzema no el nombre real x si no una aproximació binària a aquest nombre, usualment es designa per $fl(x)$.

Cleve's Corner: Cleve Moler on Mathematics and Computing,
Floating Point Arithmetic Before IEEE 754

El conjunt F(β, t, L, U)

El conjunt de nombres en coma 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 o significatius) 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 $L \leq e \leq U$. Hi pot haver $U - L + 1$ exponents diferents.

El conjunt F(β, t, L, 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**.
Hi pot haver β^t mantisses diferents.

Èpsilon de la màquina

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

$$fl(x) = x(1+\delta) \qquad |\delta| \leq \epsilon_M = \frac{1}{2}\beta^{1-t}\,.$$

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. El seu valor es correspon a la meitat distància entre 1 i el següent nombre en coma flotant.

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

Aritmètica a $F(\beta, t, L, U)$

Les operacions aritmètiques en coma flotant són

$$\begin{array}{lcll} x + y = & \longleftrightarrow & x \oplus y = & fl(fl(x) + fl(y)) \\ x - y = & \longleftrightarrow & x \ominus y = & fl(fl(x) - fl(y)) \\ x \times y = & \longleftrightarrow & x \otimes y = & fl(fl(x) \times fl(y)) \\ x \div y = & \longleftrightarrow & x \oslash y = & fl(fl(x) \div fl(y)) \end{array}$$

En totes les operacions s'ha de complir

$$x \circledast y = fl(x \circledast \; y)(1 + \epsilon_M).$$

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}.

4.3
Aritmètica de Matlab

Aritmètica de Matlab

Cal llegir els documents:

- Aritmètica en coma flotant by Cleve Moler
  Fall96Cleve.pdf
- Cleve's Corner: Cleve Moler on Mathematics and Computing,
  Floating Point Arithmetic Before IEEE 754
- Cleve's Corner: Cleve Moler on Mathematics and Computing,
  Floating Point Numbers
- Cleve's Corner: Cleve Moler on Mathematics and Computing,
  Floating Point Denormals, Insignificant But Controversial
- MathWorks Documentation Center,
  floating-point-numbers.html

5

Estabilitat numèrica i problemes ben condicionats

Some disasters attributable to bad numerical computing

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.

Anàlisi de l'error

Si f representa a l'algoritme *real* i $f^\star$ a l'algoritme *computacional*, x a la variable *real* i x^* a la variable *computacional*, aleshores l'error en el resultat final es pot definir com:

$$|f(x) - f^*(x^*)| \leq \underbrace{|f(x) - f(x^*)|}_{\text{condició}} + \underbrace{|f(x^*) - f^*(x)|}_{\text{estabilitat}} + \underbrace{|f^*(x) - f^*(x^*)|}_{\text{truncament}}$$

Algorismes amb cancel·lació

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

$$g(x+\delta) - g(x) \quad \text{amb } |\delta| \ll 1 \; ; \tag{20}$$

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:

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

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

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 $p(x) = (x-1)(x-2)(x-3)\ldots(x-10)$, 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

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

Evitar la propagació dels errors

Per tal de reduir o 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.

Exercicis

Operacions en coma flotant

En una aritmètica de cinc dígits, representeu els nombres $x = 1/3$ i $y = 5/7$ i calculeu:

a) $x \times y$ i $x \otimes y$.

b) $x + y$ i $x \oplus y$.

c) $x - y$ i $x \ominus y$.

d) $x \div y$ i $x \oslash y$.

Comproveu que l'error es manté per sota de $0.5 \cdot 10^{-4}$.

Respostes: Operacions en coma flotant

Si $fl(1/3) = 0.33333$ i $fl(5/7) = 0.71428$ llavors:

a) $x \times y = 5/21$ i $x \otimes y = fl(0.2380909524) = 0.23809$

els errors són $\delta = 0.524 \cdot 10^{-4}$ i $\epsilon = 0.220 \cdot 10^{-4}$.

b) $x + y = 22/21$ i $x \oplus y = fl(1.04761) = 1.04761$

els errors són $\delta = 0.190 \cdot 10^{-4}$ i $\epsilon = 0.182 \cdot 10^{-4}$.

c) $x - y = -8/21$ i $x \ominus y = fl(-0.38095) = -0.38095$

els errors són $\delta = 0.238 \cdot 10^{-4}$ i $\epsilon = 0.625 \cdot 10^{-5}$.

d) $x \div y = 7/15$ i $x \oslash y = fl(0.466665733325867) = 0.46666$

els errors són $\delta = 0.667 \cdot 10^{-4}$ i $\epsilon = 0.143 \cdot 10^{-4}$.

Problemes amb operacions

En una aritmètica de cinc dígits, representeu els nombres $y = 5/7$, $u = 0.714251$, $v = 98765.9$ i $w = 0.111111 \cdot 10^{-4}$ i calculeu:

a) $y \ominus u$. (restar dues quantitats molt properes)

b) $(y \ominus u) \oslash w$. (dividir per una quantitat petita)

c) $(y \ominus u) \otimes v$. (multiplicar per una quantitat gran)

d) $u \oplus v$.

e) $y \ominus w$.

Comproveu que l'error NO es manté per sota de $0.5 \cdot 10^{-4}$.

Respostes: Problemes amb operacions

$$
\begin{array}{lcl}
y = 5/7 & \Rightarrow & fl(y) = 0.71428 \cdot 10^0 \\
u = 0.714251 & \Rightarrow & fl(u) = 0.71425 \cdot 10^0 \\
v = 98765.9 & \Rightarrow & fl(v) = 0.98765 \cdot 10^5 \\
w = 0.111111 \cdot 10^{-4} & \Rightarrow & fl(w) = 0.11111 \cdot 10^{-4}
\end{array}
$$

Oper.	*Error abs.*	*Error rel.*
$y \ominus u$	$0.472 \cdot 10^{-5}$	0.136
$(y \ominus u) \oslash w$	0.425	0.136
$(y \ominus u) \otimes v$	0.466	0.136
$u \oplus v$	$0.162 \cdot 10^1$	$0.164 \cdot 10^{-4}$
$y \ominus w$	0.779	$0.122 \cdot 10^{-4}$

Autoavaluació

Exercici 1 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 2 Calculeu, respectant l'ordre dels sumands:

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

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.

Guia estudi tema

Llibre Càlcul numèric: teoria i pràctica

- Conceptes associats: capítol 1, de la pàgina 2 a la 30.
- Problemes proposats: 2, 3 i 9.

Llibre Cálculo numérico

- Conceptes associats: capítol 1, de la pàgina 13 a la 53
- Problemes proposats: 2, 3 i 9.

Llibres de consulta online

- Llibre de consulta - Accès UPCommons,
  Càlcul numèric: teoria i pràctica
- Llibre de consulta - Accès UPCommons,
  Cálculo numérico
- Llibre de consulta - Accès Biblioteca,
  Cálculo Científico con MATLAB y Octave by A. Quarteroni, F. Saleri
- Llibre de consulta - C. Moler,
  Cleve Moler - Llibre de text i codis - MathWorks

Computació Numèrica

Part 2.1 - Resolució de sistemes lineals

M. Àngela Grau Gotés

Departament de Matemàtica Aplicada II
Universitat Politècnica de Catalunya · BarcelonaTech.

6 de març de 2023

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

Àlgebra Lineal Numèrica

L'objectiu principal del tema és l'estudi de mètodes computacionals bàsics per a l'àlgebra lineal.

- Resolució de sistemes lineals no homogenis.
  - Mètodes directes: eliminació gaussiana, mètode de Gauss-Jordan, descomposició LU, factorització QR.
  - Mètodes iteratius: Jacobi, Gauss-Seidel i sobrerelaxació
  - Mínims quadrats.
- Càlcul de vectors i valors propis.
  - Mètodes de la potència.
  - Mètode QR.
  - Valors singulars.

Notació matricial

El sistema de m equacions lineals amb n incògnites,

$$\begin{array}{rcl} a_{11}x_1 + a_{12}x_2 + \ldots a_{1n}x_n & = & b_1\,, \\ a_{21}x_1 + a_{22}x_2 + \ldots a_{2n}x_n & = & b_2\,, \\ \vdots & & \vdots \\ a_{m1}x_1 + a_{m2}x_2 + \ldots a_{mn}x_n & = & b_m\,. \end{array}$$

Qualsevol sistema d'equacions lineals es pot representar per una matriu $A = (a_{ij})$ que recull els coeficients de les incògnites $x = (x_i)^t$, i el vector $b = (b_i)^t$, vector terme independent de tantes components com equacions $\left(_{1 \le i \le m,\ 1 \le j \le n}\right)$.

Nomenclatura

Es parla de sistemes lineals

1. **Segons tamany**
   - Petits ($n \leq 300$),
   - Grans ($n > 300$).
2. **Segons estructura**
   - pocs elements no nuls, matriu plena.
   - bastants elements nuls, triangular superior o inferior.
   - molts elements nuls, matriu tridiagonal, matriu diagonal i matriu sparse.

Notació matricial

Per a resoldre el sistema, es crea la matriu augmentada:

$$(A|b) = \left(\begin{array}{cccc|c} a_{11} & a_{12} & \dots & a_{1n} & b_1 \\ a_{21} & a_{22} & \dots & a_{2n} & b_2 \\ \vdots & \vdots & & \vdots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} & b_m \end{array} \right)$$

Hi ha sistemes sobredeterminats, amb més equacions que incògnites ($m > n$), hi ha sistemes no determinats, de menys equacions que incògnites ($n > m$) i sistemes amb el mateix nombre d'equacions que incògnites ($m = n$).

Existència de solucions

Per a un sistema $Ax = b$ segons el teorema de Rouche-Frobenius, tenim

- $rang(A) = rang(A|b) = n$ el sistema és compatible determinat, la solució és única.
- $rang(A) = rang(A|b) = r < n$ el sistema és compatible indeterminat, amb $n - r$ graus de llibertat.
- $rang(A) \neq rang(A|b)$ el sistema és incompatible.

Només estudiarem el cas $n = m$ i $det(A) = |A| \neq 0$, cas de solució és única que es pot calcular fent ús de la Regla de Cramer.

Mètode de Cramer

Sistema compatible determinat

La solució de $Ax = b$, per la regla de Cramer és:

$$x_i = \frac{|A_i|}{|A|}, \quad 1 \leq i \leq n$$

on $|A|$ és del determinant de la matriu a, i $|A_i|$ és el determinant de la matriu A_i obtinguda substituint la columna i de la matriu A pel vector b.

Exercici

$$\begin{aligned} 2x_1 + 4x_2 + x_3 &= 1\,, \\ 8x_1 - x_2 + 3x_3 &= 0\,, \\ 2x_1 + 5x_2 &= 0\,. \end{aligned}$$

Mètode de Cramer - Eficiència

Si la matriu és d'ordre n,

- calen $n+1$ determinants d'ordre n per a calcular x.
- cada determinant d'ordre n requereix $n!n-1$ operacions.
- el nombre d'operacions és, pel cap baix, $n!(n+1)-1$.

	Flops				
n	10^9 (Giga)	10^{10}	10^{11}	10^{12} (Tera)	10^{15} (Peta)
10	10^{-1}sec	10^{-2}sec	10^{-3}sec	10^{-4}sec	negligible
15	17 hours	1.74 hours	10.46 min	1 min	0.6 10^{-1} sec
20	4860 years	486 years	48.6 years	4.86 years	1.7 day
25	o.r.	o.r.	o.r.	o.r.	38365 years

Table 5.1. Time required to solve a linear system of dimension n by the Cramer rule. "o.r." stands for "out of reach"

És un mètode inapropiat per a l'ordinador.

Sistemes d'equacions lineals

Mètodes directes

Documentació de MATLAB® - Sistemes d'equacions lineals

Mètodes directes

Són mètodes que ens donen la solució exacte en un nombre finit d'operacions, si no fos pels errors d'arrodoniment acumulats i les possibles imprecisions en el coneixement inicial de la matriu de coeficients A i el terme independent b.

Es consideren adients per a sistemes lineals no massa grans ($100-500$ equacions) i amb pocs elements nuls.

S'estudien els mètodes,

- Mètode de Gauss.
- Mètodes de factorització LU, Txoleski i QR.
- I derivats: Gauss-Jordan, ...

Millor algoritme

En tots els algoritmes caldrà considerar

- el temps emprat per obtenir la solució (mesurat en nombre d'operacions).
- els errors d'arrodoniment del mètode de càlcul.

De res serveix un mètode que obtingui la solució en un temps clarament excesiu.

Primer presentem algorismes molt econòmics computacionalment, i finalment discutirem com afecten els errors d'arrodoniment a la solució obtinguda.

Sistema diagonal

$D = (d_{ij})$ tal que $d_{ij} = 0$ si $i \neq j$ i $1 \leq i, j \leq n$

$$(D|b) = \left(\begin{array}{cccc|c} d_{11} & 0 & \dots & 0 & b_1 \\ 0 & d_{22} & \dots & 0 & b_2 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \dots & d_{nn} & b_n \end{array} \right)$$

La solució és $x_i = \dfrac{b_i}{d_{ii}}, \ 1 \leq i \leq n$.

Operacions: calen n divisions per calcular x.

Sistema triangular superior

$U = (u_{ij})$ tal que $u_{ij} = 0$ si $_{1 \le j < i \le n}$

$$(U|b) = \left(\begin{array}{ccccc|c} u_{11} & u_{12} & u_{13} & \dots & u_{1n} & b_1 \\ 0 & u_{22} & u_{23} & \dots & u_{2n} & b_2 \\ 0 & 0 & u_{33} & \dots & u_{3n} & b_3 \\ \vdots & \vdots & \ddots & \ddots & \vdots & \\ 0 & 0 & \dots & 0 & u_{nn} & b_n \end{array}\right)$$

La solució s'obté per substitució enrera, les fórmules són

$$x_n = \frac{b_n}{u_{nn}}, \quad x_i = \frac{1}{u_{ii}}\left(b_i - \sum_{k=i+1}^{n} u_{ik}x_k\right), \quad {\scriptstyle 1 \le i < n.}$$

Sistema triangular superior

Exercici

$$
\begin{aligned}
3x_1 + 2x_2 - 2x_3 + 4x_4 &= -5\,, \\
3x_2 - 5x_3 - 3x_4 &= 0\,, \\
4x_3 + x_4 &= -3\,, \\
2x_4 &= 6\,.
\end{aligned}
$$

La solució s'obté per substitució enrera, el resultat és

$$x_4 = 3\,, \quad x_3 = -3/2\,, \quad x_2 = 1/2\,, \quad x_1 = -7\,,$$

Sistema triangular inferior

$L = (l_{ij})$ tal que $l_{ij} = 0$ si $_{1 \le i < j \le n}$

$$(L|b) = \left(\begin{array}{ccccc|c} l_{11} & 0 & 0 & \dots & 0 & b_1 \\ l_{21} & l_{22} & 0 & \dots & 0 & b_2 \\ l_{31} & l_{32} & l_{33} & \ddots & \vdots & b_3 \\ \vdots & \vdots & \vdots & \ddots & 0 & \\ l_{n1} & l_{n2} & l_{n3} & \dots & l_{nn} & b_n \end{array} \right)$$

La solució s'obté per substitució endavant,

$$x_1 = \frac{b_1}{l_{11}}, \quad x_i = \frac{1}{l_{ii}} \left(b_i - \sum_{k=1}^{i-1} l_{ik} x_k \right), \quad {\scriptstyle 1 < i \le n}.$$

Nombre d'operacions

Exercici. Calculeu el nombre total de

- divisions que calen per resoldre un sistema diagonal.
- divisions que calen per resoldre un sistema triangular.
- multiplicacions que calen per resoldre un sistema triangular.
- sumes que calen per resoldre un sistema triangular.

Ajuda: $$\sum_{i=1}^{m} i = \frac{(m+1)m}{2} \qquad \sum_{i=1}^{m} i^2 = \frac{m(m+1)(2m+1)}{6}.$$

Nombre d'operacions

	$+/-$	$*$	$/$	total
Diag	0	0	n	n
Upper	$\frac{n^2-n}{2}$	$\frac{n^2-n}{2}$	n	n^2
Lower	$\frac{n^2-n}{2}$	$\frac{n^2-n}{2}$	n	n^2

Mètode d'eliminació de Gauss

Consta de dues fases. La primera fase consisteix en modificar el nostre sistema d'equacions per arribar a un sistema triangular superior. En la segona fase es resol el sistema triangular superior obtingut en la primera fase.

Quin tipus de modificacions són vàlides en la fase 1?

- Multiplicar una fila per un nombre no nul.
- Substituir una fila per una combinació lineal de les altres.
- Permutar files del sistema.
- Permutar columnes del sistema.

Les files són equacions i les columnes són incògnites.

Exemple. Mètode d'eliminació de Gauss

$$
\begin{aligned}
2x_1 + 4x_2 + x_3 &= 1\,, \\
8x_1 - x_2 + 3x_3 &= 0\,, \\
2x_1 + 5x_2 &= 0\,.
\end{aligned}
$$

$$
G^{(0)} = \begin{pmatrix} 2 & 4 & 1 & 1 \\ 8 & -1 & 3 & 0 \\ 2 & 5 & 1 & 0 \end{pmatrix} \rightsquigarrow G^{(1)} = \begin{pmatrix} 2 & 4 & 1 & 1 \\ 0 & -17 & -1 & -4 \\ 0 & 1 & -1 & -1 \end{pmatrix}
$$

$$
G^{(2)} = \begin{pmatrix} 2 & 4 & 1 & 1 \\ 0 & -17 & -1 & -4 \\ 0 & 0 & -\frac{18}{17} & -\frac{21}{17} \end{pmatrix} \rightarrow x = \begin{pmatrix} -5/12 \\ 1/6 \\ 7/6 \end{pmatrix}
$$

Mètode d'eliminació de Gauss

L'algoritme de Gauss, s'aplica sobre la matriu ampliada; i converteix la matriu en una matriu triangular superior. La matriu del sistema A és redueix a triangular superior en $n-1$ passos si A té n files.

$$\left(\begin{array}{ccccc|c} a_{11} & a_{12} & a_{13} & \dots & a_{1n} & b_1 \\ a_{21} & a_{22} & a_{23} & \dots & a_{2n} & b_2 \\ a_{31} & a_{32} & a_{33} & \dots & a_{3n} & b_3 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ a_{n1} & a_{n2} & a_{n3} & \dots & a_{nn} & b_n \end{array}\right) \rightsquigarrow \left(\begin{array}{ccccc|c} u_{11} & u_{12} & u_{13} & \dots & u_{1n} & \widetilde{b}_1 \\ 0 & u_{22} & u_{23} & \dots & u_{2n} & \widetilde{b}_2 \\ 0 & 0 & u_{33} & \dots & u_{3n} & \widetilde{b}_3 \\ \vdots & \vdots & \ddots & \ddots & \vdots & \\ 0 & 0 & \dots & 0 & u_{nn} & \widetilde{b}_n \end{array}\right)$$

Mètode de Gauss. Pas 1

Escrivim el sistema lineal de partida com per $G^{(0)}$ la matriu $(A|b)$, el primer pas és

- verifico si $a_{11} \neq 0$, (pivot).
- s'escull la fila 1, (fila pivot).
- per cada fila per sota de la fila pivot calculo $m_{i1} = \dfrac{a_{i1}}{a_{11}}$, (multiplicador).
- per cada fila per sota de la fila pivot resto m_{i1} vegades la fila pivot de la fila i.

El resultat és una matriu, $G^{(1)}$, amb la primera columna tot zero, llevat de a_{11}.
El nombre de divisions és $n-1$, i per cada fila són n productes i sumes; en total $n(n-1)$.

Mètode de Gauss. Pas 2

El segon pas és,

- fila pivot la fila 2 de la matriu $G^{(1)}$.
- verifico si el pivot no és nul, $a_{22}^{(1)} \neq 0$.
- per cada fila per sota de la fila pivot calculo

$$m_{i2} = a_{i2}^{(1)} / a_{12}^{(1)}.$$

- per cada fila per sota de la fila pivot resto m_{i1} vegades la fila pivot de la fila i.

El resultat és una matriu, $G^{(2)}$, amb la segona columna tot zero, llevat de $a_{22}^{(2)}$ i $a_{12}^{(2)}$.
El nombre de divisions és $n-2$, i per cada fila són $n-1$ productes i sumes; en total $(n-1)(n-2)$.

Mètode de Gauss. Pas k

En general, en el pas $k < n$, reduim la columna r de la matriu $G^{(k-1)}$, modificant des de la fila k fins a la n amb la fórmula

$$m_{i\,k} = \frac{a_{i\,k}^{(k-1)}}{a_{k\,k}^{(k-1)}}\,, \quad i = r+1\,\ldots\,n\,,$$

$$NovaFila(i) = Fila(i) - m_{ik} \cdot Fila(k)\,, i = k+1 \cdots n\,.$$

El nombre de divisions és $n-k$, i per cada fila són $n-k$ productes i sumes; en total $(n-k)(n-k+1)$.

Operacions triangular superior

El nombre total de divisions és

$$\sum_{k=1}^{n-1}(n-k) = \frac{1}{2}n(n-1) = \frac{n^2}{2} + o(n)\,.$$

El nombre total de multiplicacions/sumes és

$$\sum_{k=1}^{n-1}(n-k+1)(n-k) = \sum_{k=1}^{n-1}(n^2+n-k(2n+1)+k^2) =$$

$$= (n^2+n)(n-1) - (2n+1)\frac{n(n-1)}{2} + \frac{n(n-1)(2n-1)}{6} =$$

$$= \frac{n(n-1)(n+2)}{3} = \frac{n^3}{3} + o(n^2)\,.$$

Nombre operacions

Algunos costes con el método de Gauss		
n	Coste del Método de Gauss	Tiempo (10^6 oper/s)
5	90	90 microsegundos
10	705	0,7 milisegundos
20	5510	5,5 milisegundos
100	671550	0,67 segundos
1000	667 millones	11 minutos

	Flops		
n	10^9 (Giga)	10^{12} (Tera)	10^{15} (Peta)
10^2	$7 \cdot 10^{-4}$sec	negligible	negligible
10^4	11 min	0.7 sec	$7 \cdot 10^{-4}$sec
10^6	21 years	7.7 months	11 min
10^8	o.r.	o.r.	21 years

Table 5.3. Time required to solve a full linear system of dimension n by MEG. "o.r." stands for "out of reach"

Estratègies de pivotar

- Què passa si el pivot del pas k és zero?

  Pivot trivial Es busca la primera fila per sota de la fila k que tingui valor no nul, i s'intercanvien les dues files.

  Exemples

$$\left(\begin{array}{cc|c} 0 & 1 & 1 \\ 1 & 0 & 0 \end{array}\right)$$

$$\begin{aligned} x_1 + 3x_2 + x_3 &= -3\,, \\ 3x_1 + 9x_2 + 4x_3 &= -7\,, \\ 2x_1 - x_2 + x_3 &= 6\,. \end{aligned}$$

Estratègies de pivotar

- Què passa si el pivot del pas k és proper a zero?

  Pivot parcial. S'agafa com a pivot l'element de més gran magnitud de tota la columna k.

  Pivot parcial escalat. S'agafa com a pivot l'element de la columna k o per sota de la diagonal principal que té la grandària relativa més gran respecte dels altres elements de la fila.

  Exemple.

$$\left(\begin{array}{cc|c} \epsilon & 1 & 1 \\ 1 & 1 & 2 \end{array}\right) \qquad\qquad \begin{aligned} -10^{-5}x_1 + x_2 &= 1\,, \\ 2x_1 + x_2 &= 0\,. \end{aligned}$$

Mètode de Gauss-Jordan

Per a resoldre el sistema, es transforma la matriu A en una matriu diagonal:

$$\left(A|b\right) \Rightarrow \left(D|\bar{b}\right)$$

$$\left(\begin{array}{cccc|c} a_{11} & a_{12} & \dots & a_{1n} & b_1 \\ a_{21} & a_{22} & \dots & a_{2n} & b_2 \\ a_{31} & a_{32} & \dots & a_{3n} & b_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} & b_n \end{array}\right) \Rightarrow \left(\begin{array}{ccccc|c} a_{11} & 0 & 0 & \dots & 0 & b_1 \\ 0 & a_{22} & 0 & \dots & 0 & b_2 \\ 0 & 0 & a_{33} & \dots & 0 & b_3 \\ \vdots & \vdots & \ddots & \ddots & \vdots & \\ 0 & 0 & \dots & 0 & a_{nn} & b_n \end{array}\right)$$

Mètode de Gauss-Jordan. Pas k

En general, en el pas $k < n$, reduim la columna k de la matriu $G^{(k-1)}$, modificant totes les files, llevat de la fila k, amb la fórmula

$$m_{i\,k} = \frac{a_{i\,k}^{(k-1)}}{a_{k\,k}^{(k-1)}}, \qquad i \neq k,$$

$$NovaFila(i) = Fila(i) - m_{ik} \cdot Fila(k)\,,\ \ i \neq k\,.$$

Comentari: el sistema diagonal és més fàcil de resoldre, però la reducció a sistema diagonal és més costosa.

Estabilitat.

El métode de Gauss és numèricament estable i no cal fer intercanvis de files i columnes si

- si la matriu A és diagonal dominant.
- si la matriu A és simètrica i definida positiva.

Simètrica si $A^t = A$. Definida positiva si $x^t A x > 0\,,\ \forall x \neq 0\,.$

Diagonal dominant si $|a_{ii}| \geq \sum_{\substack{i \neq j \\ 1 \leq j \leq n}} |a_{ij}|\,,\ {\scriptstyle 1 \leq i \leq n}\,.$

Aplicacions.

El mètode de Gauss s'usa:

- per a resoldre sistemes,
- per a calcular el determinant d'una matriu,
- per a calcular el rang d'una matriu.

El mètode de Gauss-Jordan s'usa per a trobar matrius inverses.

$$A = \begin{pmatrix} 6 & 2 & 1 & -1 \\ 2 & 4 & 1 & 0 \\ 1 & 1 & 4 & -1 \\ -1 & 0 & -1 & 3 \end{pmatrix}$$

Sistemes d'equacions lineals

Mètodes compactes

Documentació de MATLAB® - Factoritzacions

Mètodes Compactes

El tret principal d'aquests métodes es treballar sols amb la matriu A i presentar-la com $A = BC$ on B i C són matrius més fàcils d'invertir (nombre operacions).

Descomposicions més conegudes són[1]

- $A = LU$, L triangular inferior i U triangular superior.
- $A = R^t R$, R triangular superior i A sim. def. pos.
- $A = QR$, Q ortogonal i R triangular superior.
- $A = U\Sigma V^t$, U i V ortogonals i Σ diagonal.

[1] LU i $R^t R$ matrius $n \times n$, QR i $U\Sigma V'$ matrius $n \times m$

Factorització LU

$$A = \begin{pmatrix} l_{11} & 0 & 0 & \dots & 0 \\ l_{21} & l_{22} & 0 & \dots & 0 \\ l_{31} & l_{32} & l_{33} & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ l_{n1} & l_{n2} & l_{n3} & \dots & l_{nn} \end{pmatrix} \begin{pmatrix} u_{11} & u_{12} & u_{13} & \dots & u_{1n} \\ 0 & u_{22} & u_{23} & \dots & u_{2n} \\ 0 & 0 & u_{33} & \dots & u_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & u_{nn} \end{pmatrix}$$

- És un sistema de n^2 equacions i $n^2 + n$ incògnites.
- Comentari 1: Métode de Doolitle, $l_{ii} = 1$.
- Comentari 2: Métode de Crout, $u_{ii} = 1$.
- Comentari 3: sense pivotar, L és la matriu dels multiplicadors i U és la matriu resultant del métode de Gauss a la matriu A.

Algorisme

Algoritmo del la Factorización LU

Para $k = 1, \ldots, n,$

$$\begin{aligned} \ell_{kk} u_{kk} &= a_{kk} - \sum_{r=1}^{k-1} \ell_{kr} u_{rk}; \\ \ell_{ik} &= \frac{a_{ik} - \sum_{r=1}^{k-1} \ell_{ir} u_{rk}}{u_{kk}}, \quad i = k+1, \ldots, n; \\ u_{kj} &= \frac{a_{kj} - \sum_{r=1}^{k-1} \ell_{kr} u_{rj}}{\ell_{kk}}, \quad j = k+1, \ldots, n. \end{aligned}$$

El cost de la factorització és $\dfrac{4n^3 + 2n}{6}$.

Exercici

Calculeu la factorització LU de la matriu

$$A = \begin{pmatrix} 6 & 2 & 1 & -1 \\ 2 & 4 & 1 & 0 \\ 1 & 1 & 4 & -1 \\ -1 & 0 & -1 & 3 \end{pmatrix}$$

Matlab

A=[6,2,1,-1; 2,4,1,0;1,1,4,-1;-1,0,-1,3]

[L,U]=A o [L,U,P]=lu(A)

Factorització LU

Notem per A_i les submatrius de la matriu A formades per les i primeres files i les i columnes de la matriu A.

Existéncia

Una matriu A, regular, admet factorització LU si i només si totes les matrius A_i, $i = 1, \ldots, n$, són regulars.

- Si A és diagonal dominant, admet factorització LU.
- Si A és simètrica i definida positiva, admet factorització LU.

La resolució del sistema lineal $Ax = b$ fent ús de $PA = LU$, L triangular inferior i U triangular superior, és en dos passos:

- Pas 1, es calcula $Ly = Pb$ (endavant).
- Pas 2, es calcula $Ux = y$ (enrera).

Factorització Txoleski

Trobeu la factorització $A = R^t R$ i resoleu després el sistema lineal $Ax = b$,

$$A = \begin{pmatrix} 16 & 4 & 0 & -4 \\ 4 & 5 & 2 & -1 \\ 0 & 2 & 2 & -2 \\ -4 & -1 & -2 & 6 \end{pmatrix} \quad b = \begin{pmatrix} 8 \\ -2 \\ -4 \\ 3 \end{pmatrix} \quad \Longrightarrow \quad R = \begin{pmatrix} 4 & 1 & 0 & -1 \\ 0 & 2 & 1 & 0 \\ 0 & 0 & 1 & -2 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

Existència

Tota matriu A simètrica i definida positiva es pot factoritzar com $A = R^t R$ per R triangular superior o $A = SS^t$ per S triangular inferior (factorització de Txoleski).

Serveix com a test per dir si una matriu és definida positiva.

Algorisme

Método de Cholesky

Para $k = 1, \dots, n$,

$$\ell_{kk} = \sqrt{a_{kk} - \sum_{r=1}^{k-1} \ell_{k,r}^2};$$

$$\ell_{i,k} = \frac{a_{i,k} - \sum_{r=1}^{k-1} \ell_{ir}\ell_{kr}}{\ell_{kk}}, \quad i = k+1, \dots, n.$$

El cost de la factorització és $\mathcal{O}(n^3)$.

Factorització QR

La factorització QR expressa la matriu A com el producte de dues matrius, una ortogonal ($Q^t Q = I$) i l'altre triangular superior (R). Així, el sistema lineal $Ax = b$ es reduiex a resoldre $Rx = Q^t b$.

Aquesta factorització és més costosa que la **LU** però les matrius A i $R = Q^t A$ tenen el mateix nombre de condició.

La factorització **QR** no és única.

- Mètode de les rotacions de Givens.
- Mètode de Gram-Schmidt d'ortogonalització.
- Mètode de les reflexions de Householder (1958).

Factorització QR

Transformacions de Householder

Per $v \in \mathbb{R}^n$ definim la transformació associada a v per

$$H = \begin{cases} I, & v = 0, \\ I - \dfrac{2}{v^t v} v v^t, & v \neq 0. \end{cases}$$

Propietats

1. H és simètrica, ortogonal i $H^2 = I$.
2. Si $x \perp v, \Rightarrow Hx = x$.
3. Si $x \parallel v, \Rightarrow Hx = -x \;\; Hv = -v$.
4. Si $||x|| = ||v||, v = x - y \, , \Rightarrow Hx = y$.

Factorització QR

Per A és una matriu $m \times n$, les matrius ortogonals Q_k modifiquen les files $k, \cdots, m$ de la manera següent:

$$\underset{A}{\begin{bmatrix} x & x & x \\ x & x & x \\ x & x & x \\ x & x & x \\ x & x & x \end{bmatrix}} \xrightarrow{Q_1} \underset{Q_1 A}{\begin{bmatrix} * & * & * \\ 0 & * & * \\ 0 & * & * \\ 0 & * & * \\ 0 & * & * \end{bmatrix}} \xrightarrow{Q_2} \underset{Q_2 Q_1 A}{\begin{bmatrix} x & x & x \\ 0 & * & * \\ 0 & 0 & * \\ 0 & 0 & * \\ 0 & 0 & * \end{bmatrix}} \xrightarrow{Q_3} \underset{Q_3 Q_2 Q_1 A}{\begin{bmatrix} x & x & x \\ 0 & x & x \\ 0 & 0 & * \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}}$$

De $Q_n Q_{n-1} \cdots Q_1 A = R$, construïm

$$A = \underbrace{Q_1^t \dots Q_{n-1}^t Q_n^t}_{Q} R = QR$$

Són una successió de simetries per transformar les columnes de la matriu A a una forma triangular superior

Factorització QR

$$[\mathbf{Ax} = \mathbf{b}] \equiv \begin{pmatrix} 1 & -1 & -1 \\ 2 & 0 & 1 \\ -2 & 7 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 4 \\ -7 \end{pmatrix}$$

$$\left\| \begin{pmatrix} 1 \\ 2 \\ -2 \end{pmatrix} \right\| = 3, \quad \Rightarrow \quad y^{(1)} = \begin{pmatrix} 3 \\ 0 \\ 0 \end{pmatrix},$$

$$v^{(1)} = x^{(1)} - y^{(1)} = \begin{pmatrix} -2 \\ 2 \\ -2 \end{pmatrix}.$$

Factorització QR

$$H_1 = I - \frac{2}{v_1^* v_1} v_1 v_1^* = I - \frac{2}{12} \begin{pmatrix} -2 \\ 2 \\ -2 \end{pmatrix} \begin{pmatrix} -2 & 2 & -2 \end{pmatrix} = \begin{pmatrix} 1/3 & 2/3 & -2/3 \\ 2/3 & 1/3 & 2/3 \\ -2/3 & 2/3 & 1/3 \end{pmatrix}$$

$$H_1 A = \begin{pmatrix} 1/3 & 2/3 & -2/3 \\ 2/3 & 1/3 & 2/3 \\ -2/3 & 2/3 & 1/3 \end{pmatrix} \begin{pmatrix} 1 & -1 & -1 \\ 2 & 0 & 1 \\ -2 & 7 & 1 \end{pmatrix} = \begin{pmatrix} 3 & -5 & -1/3 \\ 0 & 4 & 1/3 \\ 0 & 3 & 5/3 \end{pmatrix}$$

$$x_2 = \begin{pmatrix} -5 \\ 4 \\ 3 \end{pmatrix} \quad y_2 = \begin{pmatrix} -5 \\ \sqrt{4^2+3^2} \\ 0 \end{pmatrix} = \begin{pmatrix} -5 \\ 5 \\ 0 \end{pmatrix} \implies v_2 = x_2 - y_2 = \begin{pmatrix} 0 \\ -1 \\ 3 \end{pmatrix}$$

$$H_2 = I - \frac{2}{v_2^* v_2} v_2 v_2^* = I - \frac{2}{10} \begin{pmatrix} 0 \\ -1 \\ 3 \end{pmatrix} \begin{pmatrix} 0 & -1 & 3 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 4/5 & 3/5 \\ 0 & 3/5 & -4/5 \end{pmatrix}$$

En Matlab la factorització QR s'obté per [Q, R, P]=qr(A).

En aquest cas, la resolució del sistema lineal $Ax = b$ amb $AP = QR$, R triangular inferior i Q ortogonal, és en dos passos:

- Pas 1, es calcula $B = Q'Pb$.
- Pas 2, es resol el sistema triangular $Rx = B$.

Les matrius A i $R = Q'A$ tenen el mateix nombre de condició.

Sistemes d'equacions lineals

Vector residu i errors

Condicionament

Exemple

$$(A|b) = \left(\begin{array}{cccc|c} 10 & 7 & 8 & 7 & 32 \\ 7 & 5 & 6 & 5 & 23 \\ 8 & 6 & 10 & 9 & 33 \\ 7 & 5 & 9 & 10 & 31 \end{array}\right) \xrightarrow{sol.exacte} \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} \qquad (1)$$

$$(A|b) = \left(\begin{array}{cccc|c} 10 & 7 & 8 & 7 & 32.1 \\ 7 & 5 & 6 & 5 & 22.9 \\ 8 & 6 & 10 & 9 & 33.1 \\ 7 & 5 & 9 & 10 & 30.9 \end{array}\right) \xrightarrow{sol.exacte} \begin{pmatrix} ? \\ ? \\ ? \\ ? \end{pmatrix} \qquad (2)$$

Una petita modificació en les dades (terme independent) dóna lloc a una gran modificació en el resultat (solució)

Condicionament

Un sistema d'equacions lineals $Ax = b$ es diu *ben condicionat* quan els errors de la matriu de coeficients A i del vector terme independent b produeixen en la solució un error del mateix ordre.

Un sistema d'equacions lineals $Ax = b$ es diu *mal condicionat* quan els errors de la matriu de coeficients A i del vector terme independent b produeixen en la solució del sistema un error d'ordre superior en al de les dades.

$$\begin{array}{l} \|A-\bar{A}\| < \epsilon \\ \\ \|b-\bar{b}\| < \epsilon \end{array} \Longrightarrow \begin{cases} \|x-\bar{x}\| \simeq \epsilon\,, & \text{ben condicionat,} \\ \\ \|x-\bar{x}\| \gg \epsilon\,, & \text{mal condicionat,} \end{cases}$$

Nombre de condició

Es diu que el sistema $Ax = B$ està mal condicionat si A té un nombre de condició gran.

Nombre de condició

$$\mathcal{K}(A) = \begin{cases} \|A\| \|A^{-1}\| \,, & det(A) \neq 0 \\ \infty \,, & altrament \end{cases}$$

Propietats

- $\mathcal{K}(A) \geq 1, \quad \mathcal{K}(I) = 1.$
- Si $B = zA$, per $z \neq 0$ real, llavors $\mathcal{K}(B) = \mathcal{K}(A)$.
- $\mathcal{K}(AB) \leq \mathcal{K}(A)\,\mathcal{K}(B)$.
- $\mathcal{K}_2(AB) = \sigma_n/\sigma_1$.
- $\mathcal{K}_2(A) = \mathcal{K}_2(AQ) = \mathcal{K}_2(QA)$ per Q matriu ortogonal.

Fites de l'error

Si $x^* = x + \delta x$, en el sistema lineal $Ax = b$ es verifica[2]:

$$\text{Errors en } b \quad \frac{\|\delta x\|}{\|x\|} \leq \mathcal{K}(A)\frac{\|\delta b\|}{\|b\|} \tag{3}$$

$$\text{Errors en } A \quad \frac{\|\delta x\|}{\|x + \delta x\|} \leq \mathcal{K}(A)\frac{\|\delta A\|}{\|A\|} \tag{4}$$

$$\text{Errors en } A \text{ i } b \quad \frac{\|\delta x\|}{\|x\|} \leq \mathcal{K}(A)\left(\frac{\|\delta A\|}{\|A\|}\frac{\|x + \delta x\|}{\|x\|} + \frac{\|\delta b\|}{\|b\|}\right) \tag{5}$$

[2] Aquest resultat el podeu trobat en la primera referència bibliogràfica.

Vector residu i error

Com a criteri de comparació entre la solució exacta x, i la solució calculada $x^* = x + \delta x$, del sistema lineal $Ax = b$ definim el vector residu $r(x^*)$ per:

$$r(x^*) = A\,\delta x = Ax - Ax^* = b - Ax^*\,.$$

Llavors es verifica[3]:

$$\frac{\|\delta x\|}{\|x\|} \leq \mathcal{K}(A)\frac{\|\delta b\|}{\|b\|} = \mathcal{K}(A)\frac{\|r(x^*)\|}{\|b\|} \tag{6}$$

[3] Aquest resultat el podeu trobat en la primera referència bibliogràfica.

Sistemes d'equacions lineals

Mètodes iteratius

Documentació de MATLAB® - Iterative Methods for Linear Systems

Mètodes iteratius

Mètodes iteratius estacionaris, mètodes iteratius no estacionaris

Són mètodes que construeixen una successió de vectors convergent a la solució exacte amb un nombre finit d'operacions en cada iteració, si no fos pels errors d'arrodoniment acumulats i les possibles imprecisions en el coneixement inicial de la matriu A i el vector b.

Es consideren adients per a sistemes lineals d'ordre alt.

Treballarem tres mètodes,

- mètode de Jacobi.
- mètode de Gauss-Seidel
- mètodes de Sobrerelaxació

Mètodes iteratius estacionaris

Transformen el sistema lineal $Ax = b$ com $x = B\,x + c$. Els dos sistemes han de ser consistents.

$$\left.\begin{array}{r} Ax = b \\ Ax^* = b \end{array}\right\} \Longleftrightarrow \left\{\begin{array}{c} x^{(k+1)} = B\,x^{(k)} + c \\ x^* = B\,x^* + c \end{array}\right.$$

Algorisme

Partim de $x^{(0)}$ arbitrari, i generem la successió de vectors $x^{(k)}$ a partir de la recurrència $x^{(k+1)} = B\,x^{(k)} + c$.

Cost computacional

Cada iteració són n^2 sumes i n^2 productes; després de k iteracions són $2n^2 k$.

Convergent?

La successió de vectors $x^{(k)}$ és convergent a x^*, solució de $Ax = b$?

Teoremes de convergència

Teorema I

Si A és no singular, definim $r^{(k)} = b - Ax^{(k)}$

$$\lim_{k\to\infty} x^{(k)} = x^* \iff \lim_{k\to\infty} r^{(k)} = 0.$$

Teorema II

El mètode iteratiu $x^{(k+1)} = B\,x^{(k)} + c$ és convergent a la solució x^* per qualsevol $x^{(0)}$ $\iff$ $\rho(B) < 1$.

Vector residu: $r^{(k)} = b - Ax^{(k)}$.

Fites de l'error

- Raó de convergència
- $x^{(k)} - x^* = B\left(x^{(k-1)} - x^*\right) = \cdots = B^k\left(x^{(0)} - x^*\right) \Longrightarrow$

$$||x^{(k)} - x^*|| = ||B||^k ||x^{(0)} - x^*||$$

- Fites de l'error
- $x^{(k)} - x^* = -B\left(x^{(k)} - x^{(k-1)}\right) + B\left(x^{(k)} - x^*\right) \Longrightarrow$

$$||x^{(k)} - x^*|| = \frac{\beta}{1-\beta}||x^{(k)} - x^{(k-1)}||, \quad \beta = ||B|| < 1\,.$$

- $B\left(x^{(k)} - x^{(k-1)}\right) = B^k\left(x^{(1)} - x^{(0)}\right) \Longrightarrow$

$$||x^{(k)} - x^*|| = \frac{\beta^k}{1-\beta}||x^{(1)} - x^{(0)}||, \quad \beta = ||B|| < 1\,.$$

Mètode general

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{pmatrix}$$

Per convertir $Ax = b$ en un sistema de la forma $x = Bx + c$, expressem la matriu A com a suma det tres matrius: $A = L + D + U$ tals

$$\underbrace{\begin{pmatrix} 0 & 0 & \dots & 0 \\ a_{21} & 0 & \dots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ a_{n1} & \dots & a_{nn-1} & 0 \end{pmatrix}}_{L} + \underbrace{\begin{pmatrix} a_{11} & 0 & \dots & 0 \\ 0 & a_{22} & 0\dots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & a_{nn} \end{pmatrix}}_{D} + \underbrace{\begin{pmatrix} 0 & a_{12} & \dots & a_{1n} \\ 0 & 0 & \dots & a_{2n} \\ \vdots & \vdots & \ddots & a_{n-1n} \\ 0 & 0 & \dots & 0 \end{pmatrix}}_{U}$$

Exemple

Determineu les matrius d'iteració del mètode de Jacobi i del mètode Gauss-Seidel del sistema $Ax = b$ donat per

$$A = \begin{pmatrix} 1 & 0 & 1 \\ -1 & 1 & 0 \\ 1 & 2 & -3 \end{pmatrix} \qquad b = \begin{pmatrix} 4 \\ 1 \\ -4 \end{pmatrix}$$

que té per solució $x^* = (1\,,2\,,3)^\top$.

El mètode Jacobi es basa en la resolució de cada variable localment respecte a les altres variables.

$$\left.\begin{array}{c} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n = b_2 \\ \vdots \qquad \vdots \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n = b_n \end{array}\right\} \Longleftrightarrow \left\{\begin{array}{l} x_i^{(k+1)} = \dfrac{1}{a_{ii}}\left(b_i - \displaystyle\sum_{\substack{j=1\\ j\neq i}}^{n} a_{ij}x_j^{(k)}\right) \\ i = 1, 2, \ldots, n, \quad k \geq 0\,. \end{array}\right.$$

<u>Notació Matricial</u>

$$Ax = b \Longleftrightarrow x^{(k+1)} = B_J\, x^{(k)} + c_J\,, \quad \forall\ k \geq 0\,.$$

La matriu d'iteració del mètode $B_J = -D^{-1}(L+U)$ i el vector $c_J = D^{-1}b$

Si A és diagonal dominant estricte el mètode és convergent.

Mètode de Gauss-Seidel

Construcció matrius

El mètode Gauss-Seidel és com el mètode Jacobi, excepte que utilitza valors actualitzats tan aviat com estiguin disponibles.

$$\left.\begin{array}{c} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n = b_2 \\ \vdots \qquad \vdots \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n = b_n \end{array}\right\} \Longleftrightarrow \left\{\begin{array}{l} x_i^{(k+1)} = \dfrac{1}{a_{ii}}\left(b_i - \displaystyle\sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij}x_j^{(k)}\right) \\ i = 1, 2, \ldots, n, \quad k \geq 0\,. \end{array}\right.$$

Notació Matricial

$$Ax = b \Longleftrightarrow x^{(k+1)} = B_{GS}\, x^{(k)} + c_{GS}\,, \quad \forall\; k \geq 0\,.$$

La matriu d'iteració és $B_{GS} = -(L+D)^{-1}U$ i el vector és $c_{GS} = (L+D)^{-1}b$

Si A és diagonal dominant estricte, el mètode és convergent.
Si A és simètrica definida positiva, el mètode és convergent.

Mètodes de sobrerelaxació

Mètodes de relaxació - variant JACOBI

Són una generalització dels dos mètodes estudiats; sumem i restem x_i^k en l'expressió del mètode de Jacobi,

$$x_{Ji}^{(k+1)} = x_i^{(k)} + \underbrace{\frac{1}{a_{ii}} \left(b_i - \sum_{j=1}^{n} a_{ij} x_j^{(k)} \right)}_{\text{correcció}}, \quad k \geq 0\,.$$

El mètode de relaxació consisteix en multiplicar la correcció per un paràmetre ω, **paràmetre de relaxació**, de manera que s'acceleri la convergència. En termes matricials seria:

- ✓ $C = D^{-1}$ Matriu auxiliar.
- ✓ $B_{sor} = C\left((1-\omega)D - \omega(L+U)\right)$ Matriu d'iteració.
- ✓ $c_{sor} = \omega\ C\ b$ Vector d'iteració.

El mètode de relaxació consisteix en multiplicar la correcció per un paràmetre ω, **paràmetre de relaxació**, de manera que s'acceleri la convergència. Per components seria:

$$x_i^{(k+1)} = x_i^{(k)} + \underbrace{\omega \frac{1}{a_{ii}} \left(b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(k+1)} - \sum_{j=i}^{n} a_{ij} x_j^{(k)} \right)}_{\text{correcció}}, \quad k \geq 0\,.$$

$$x_i^{(k+1)} = \omega x_{Si}^{(k+1)} + (1-\omega) x_i^{(k)}\,, \quad k \geq 0\,.$$

En termes matricials seria:

- ✓ $C = (D + \omega L)^{-1}$ Matriu auxiliar.
- ✓ $B_{sor} = C\,((1-\omega)D - \omega U)$ Matriu d'iteració.
- ✓ $c_{sor} = \omega\ C\ b$ Vector d'iteració.

Mètodes de sobrerelaxació

Variant Seidel

Convergència

Si la matriu A té tots els elements diagonals no nuls, llavors

$$|\omega - 1| \leq \rho(B_{sor})\,.$$

Per convergència només possible $\omega \in (0, 2)$.

- ✓ Si $\omega = 1$ és el mètode de Gauss-Seidel.
- ✓ Si $0 < \omega < 1$ mètodes de subrelaxació.
- ✓ Si $1 < \omega < 2$ mètodes de sobrerelaxació.

TEOREMA. *Sigui A simètrica, definida positiva i tridiagonal en blocs:*

$$A = \begin{pmatrix} D_1 & U_1 & & & \\ L_2 & D_2 & U_2 & 0 & \\ & & & & \\ & 0 & & & U_{n-1} \\ & & & L_n & D_n \end{pmatrix}$$

on D_i, $i = 1, \ldots, n$ són submatrius diagonals, U_i, L_i, submatrius qualssevol que satisfan $L_{i+1} = U_i^T$, $i = 1, \ldots, n-1$. Llavors $\rho(B_{GS}) = \rho^2(B_J)$ i el paràmetre de relaxació $\bar{w}$ òptim és

$$\bar{w} = \frac{2}{1 + (1 - \rho(B_{GS}))^{1/2}}, \quad \rho(B_{GS}) < 1,$$

on $\rho(B_J)$ és el radi espectral de la iteració de Jacobi corresponent a A. El valor òptim de $\rho(B_w)$ és

$$\rho(B_{\bar{w}}) = \bar{w} - 1.$$

Si la matriu A verifica les hipotesis del teorema anterior resulta que,

$$\rho(B_{GS}) = (\rho(B_J))^2 \ ,$$

per tant, si el mètode de Jacobi és convergent, també ho és el de Gauss-Seidel i el factor de convergència asimptòtica és el quadrat del de Jacobi.

Quin valor de $\omega_0 \in (0, 2)$ minimitza el radi espectral $\rho(B_\omega)$?

ω - òptim

$$\omega_0 = \frac{2}{1 + \sqrt{1 - \rho(B_J)^2}} \quad \text{i} \quad \rho(B_\omega) = \omega_0 - 1 \, .$$

Exemple

Determineu les 10 primeres iteracions del mètode de Jacobi, del mètode Gauss-Seidel i del mètode de sobrerelaxació prenent $x^{(0)} = (0\,,0\,,0\,,0)^\top$ del sistema $Ax = b$ donat per

$$A = \begin{pmatrix} 10 & -1 & 2 & 0 \\ -1 & 11 & -1 & 3 \\ 2 & -1 & 10 & -1 \\ 0 & 3 & -1 & 8 \end{pmatrix} \qquad b = \begin{pmatrix} 6 \\ 25 \\ -11 \\ 15 \end{pmatrix}$$

que té per solució $x^* = (1\,,2\,,-1\,,1)^\top$.

Estudieu el residu per cada mètode.

Exemple

k	0	1	2	3	4	5	6	7	8	9	10
$x_1^{(k)}$	0.0000	0.6000	1.0473	0.9326	1.0152	0.9890	1.0032	0.9981	1.0006	0.9997	1.0001
$x_2^{(k)}$	0.0000	2.2727	1.7159	2.053	1.9537	2.0114	1.9922	2.0023	1.9987	2.0004	1.9998
$x_3^{(k)}$	0.0000	−1.1000	−0.8052	−1.0493	−0.9681	−1.0103	−0.9945	−1.0020	−0.9990	−1.0004	−0.9998
$x_4^{(k)}$	0.0000	1.8750	0.8852	1.1309	0.9739	1.0214	0.9944	1.0036	0.9989	1.0006	0.9998

Figura: Iteracions Jacobi

k	0	1	2	3	4	5
$x_1^{(k)}$	0.0000	0.6000	1.030	1.0065	1.0009	1.0001
$x_2^{(k)}$	0.0000	2.3272	2.037	2.0036	2.0003	2.0000
$x_3^{(k)}$	0.0000	−0.9873	−1.014	−1.0025	−1.0003	−1.0000
$x_4^{(k)}$	0.0000	0.8789	0.9844	0.9983	0.9999	1.0000

Figura: Iteracions Gaus-Seidel

Sistemes d'equacions lineals

Precondicionadors

Documentació de MATLAB® - Iterative Methods for Linear Systems

Preconditioning

Es pot millorar la convergència i l'estabilitat de la majoria de métodes iteratius transformant el sistema lineal per tenir un espectre més favorable. Aquesta transformació es realitza aplicant una segona matriu, anomenada matriu precondicionadora, al sistema d'equacions. Aquest procés transforma el sistema lineal $Ax = b$ en un sistema equivalent $\widetilde{A}\widetilde{x} = \widetilde{b}$

El precondicionador ideal (A^{-1}) transforma la matriu de coeficients A en una matriu d'identitat. A la pràctica, trobar un bon precondicional requereix compensacions. La transformació (M) potser de tres tipus:

- precondicionament per l'esquerre $\left(M^{-1}A\right)x = \left(M^{-1}b\right)$.
- precondicionament per la dreta $\left(AM^{-1}\right)(Mx) = b$.
- split; usualment per matrius simètriques, la matriu precondicionadora M tal que $M = H\,H^t$ (split) per mantenir la simetria del sistema transformat $\left(H^{-1}AH^{-t}\right)H^t x = \left(H^{-1}b\right)$.

Preconditioning

Direct solvers:
Sequential, loosing sparsity

Iterative solvers:
easy parallel and sparse,
but possibly slowly convergent

Combination of both methods:

Include preconditioner $M \approx A$ in the form $M^{-1} A x = M^{-1} b$, such that

- M is easy to deal with in parallel (reduced approximate direct solver)
- spectrum of $M^{-1} A$ is much better clustered

Or include preconditioner $M \approx A^{-1}$ in the form $M A x = M b$, such that

- M is easy to deal with in parallel (reduced approximate inverse)
- spectrum of $M A$ is much better clustered

Sistemes d'equacions lineals

Mètodes iteratius no estacionaris

Documentació de MATLAB® - Iterative Methods for Linear Systems

Els mètodes no estacionaris difereixen dels mètodes estacionaris en què els càlculs impliquen informació que canvia en cada iteració. Normalment, les constants es calculen prenent productes interns de residus o altres vectors derivats del mètode iteratiu.

Alguns d'aquests mètodes són: Mètode del gradient conjugat (`CG`) i variants: `MINRES`, `SYMMLQ`, `CGNE`, `CGNE`, `GMRES`, `BiCG`, `QMR`, `Bi-CGSTAB`. Consulteu la bibliografia.

Vector residu i vector gradient

$Ax = b\,,\quad$ A simètrica i definida positiva.

Resoldre el sistema lineal $Ax = b$ és equivalent al problema de minimitzar la funció definida per

$$\phi(x) = \frac{1}{2}x^T Ax - x^T c \tag{7}$$

Obs. El gradient d'aquesta funció és $\nabla\phi(x) = Ax - b$.

Sistemes lineals SOBREDETERMINATS

Exemple

Exemple. Les dades de la Taula 5.2 s'han tret de J.C. Miller & J.N. Miller (1993), *Statistics in Analytical Chemistry*, Ellis Horwood. Corresponen a una investigació sobre un test colorimètric per a la concentració de glucosa, en la que es varen obtenir absorbàncies per a sis concentracions patró de glucosa.

En els experiments de calibratge de l'anàlisi instrumental es pren sempre com a variable de control x la concentració (de fet, al ser una concentració patró, el seu valor no és experimental, sinó prefixat per l'usuari). La variable resposta y és en aquest cas, l'absorbància. Ajustem per mínims quadrats un model $y = a + bx$.

TAULA 5.2

Concentració (mM)	0	2	4	6	8	10
Absorbància	0.002	0.150	0.294	0.434	0.570	0.704

Sistemes lineals sobredeterminats

Sigui A una matriu $m \times n$, $m \geq n$, b un vector de m components, x el vector de n d'incògnites. El sistema d'equacions $Ax = b$ no té solució, llavors busquem x tal que Ax sigui la millor aproximació de b (pel mètode dels mínims quadrats).

Vector residu: $r_y = b - Ay$ per a y vector de n components.

Teorema (Equacions normals)

Si x és la solució dels sistema d'equacions normals, $A^t(b - Ax) = 0$, llavors

$$||r_x||_2 \leq ||r_y||_2 \quad \forall y \in \mathbb{R}^n.$$

Teorema (Existència de solució)

A^tA és no singular si i només si $\text{rank}(A) = n$

Equacions normals

$Ax = b$, m files i n incògnites amb rang(A)=n :

$$\begin{aligned} x_1 + x_2 &= 1 \\ x_1 + 2x_2 &= 2 \\ x_1 + 3x_2 &= 5 \end{aligned}$$

- ✓ $A^t AX = A^t b$ — Equacions normals.
- ✓ $RX = Q^t A^t b$ — Solució per factorització QR de $A^t A$.
- ✓ $\|b - AX\|_2$ — Residu mínim.

- **Sistema lineal**

$$x_1 + x_2 = 1$$
$$x_1 + 2x_2 = 2$$
$$x_1 + 3x_2 = 5$$

$$\begin{cases} x_1 + x_2 = 1 \\ x_1 + 2x_2 = 2 \end{cases} \Rightarrow \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \quad \|r_x\| = 2$$

$$\begin{cases} x_1 + x_2 = 1 \\ x_1 + 3x_2 = 5 \end{cases} \Rightarrow \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} -1 \\ 2 \end{pmatrix} \quad \|r_x\| = 1$$

$$\begin{cases} x_1 + 2x_2 = 2 \\ x_1 + 3x_2 = 5 \end{cases} \Rightarrow \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} -4 \\ 3 \end{pmatrix} \quad \|r_x\| = 2$$

$$A = \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{pmatrix} \Rightarrow A^t A = \begin{pmatrix} 3 & 6 \\ 6 & 14 \end{pmatrix} \Rightarrow x = \begin{pmatrix} -4/3 \\ 2 \end{pmatrix} \quad \|r_x\| = \frac{\sqrt{6}}{3} = 0.8165 < 1$$

Guia estudi

Guia estudi tema

Llibre Càlcul numèric: teoria i pràctica

- Conceptes i exercicis resolts: capítol 4, pàgines 117-143.
- Problemes proposats: 1, 2, 3, 4, 5, 6, 7, 8, 9 i 13.
- Pràctiques resoltes : de la pàgina 147-154.
- Pràctiques proposades: pàgines 154-157.

Llibre Cálculo Científico con MATLAB y Octave

- Conceptes i exercicis resolts: capítol 5, pàgines 127-170.
- Problemes i pràctiques proposades: del 5.1 al 5.17

Llibres de consulta online

- Llibre de consulta - Accès UPCommons,
  Càlcul numèric: teoria i pràctica
- Llibre de consulta - Accès UPCommons,
  Cálculo numérico
- Llibre de consulta - Accès Biblioteca,
  Cálculo Científico con MATLAB y Octave
- Linear Equations,,
  Chapter 2
- Llibre de consulta - Accès netlib.org,
  Barrett, R., M. Berry, T. F. Chan, et al.,
  Templates for the Solution of Linear Systems: Building Blocks for Iterative Methods, SIAM, Philadelphia, 1994.

Computació Numèrica

Tema 2
Valors propis i vectors propis. Valors singulars

M. Àngela Grau Gotés

Departament de Matemàtica Aplicada II
Universitat Politècnica de Catalunya · BarcelonaTech.

14 de març de 2023

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

En el camp de l'enginyeria els valors i els vectors propis tenen una relevància destacada per analitzar models amb oscilacions i resonàncies. El seu coneixement és bàsic en:

- Sistemes elèctrics de corrent alterna
- Vibració natural d'estructures
- Mecànica quàntica
- Làsers
- Resonància Magnètica Nuclear (NMR)
- Anàlisis de Components Principals
- Algoritme Pagerank

1.- Pont Tacoma Narrows 1940

2.- Enfonsament del Pont Tacoma Narrows, Washington. Video original

3.- Enfonsament del Pont Tacoma Narrows, Washington. Video

1.- The Anatomy of a Large-Scale Hypertextual Web Search Engine Autors: Sergey Brin and Lawrence Page

2.- https://sctmates.webs.ull.es/

2.- PageRank, Wikipedia

3.- La descomposición en valores singulares (SVD) y algunas de sus aplicaciones

Valors i vectors propis

Sigui $A = (a_{i\,j})_{n\times n}$ per $a_{i\,j} \in \mathbb{R} \quad {}_{1\leq i,j\leq n}$, A matriu quadrada de nombres reals; $\lambda \in \mathbb{R}$ un real i $v \in \mathbb{R}^n$ un vector.

Vector i valor propi

Si l'equació $\boldsymbol{Av = \lambda v}$ té alguna solució no trivial $(v \neq 0)$ es diu que $\boldsymbol{\lambda}$ és un valor propi i $\boldsymbol{v}$ un valor propi de la matriu A.

- Espectre: $spec(A) = \{\lambda_1, \lambda_2, \ldots, \lambda_n, \,/\, \lambda_k \text{valor propi } A\}$
- Radi espectral: $\rho(A) = \max\left(\{|\lambda_1|, \ldots, |\lambda_n|, \,/\, \lambda_k \text{valor propi } A\}\right)$
- Polinomi característic: $p(\lambda) = det(A - \lambda\, I)$.

Si $\boldsymbol{v}$ i $\boldsymbol{\lambda}$ compleixen $\boldsymbol{Av = \lambda v}$, llavors

1. $(-A)v = -\lambda v\,,$
2. $(A - \mu\, I)v = (\lambda - \mu)v\,,$
3. $A^{-1}v = \dfrac{1}{\lambda}v\,,$
4. $(A - \mu\, I)^{-1}v = \dfrac{1}{(\lambda - \mu)}v\,,$
5. El conjunt $\left\{v^{(1)}, v^{(2)}, \dots v^{(k)}\right\}$ de vectors propis de valor propi $\lambda_1, \lambda_2, \dots, \lambda_k$, tots diferents, és un conjunt de vectors linealment independent.

Dues matrius, A i B quadrades d'ordre n es diuen **semblants** si existeix una matriu S tal que $A = S^{-1}\, B\, S$. Una matriu és **diagonalitzable** si és semblant a una matriu diagonal.

Igualtat de valors propis

A i B semblants, λ i v tals que $Av = \lambda v$, llavors λ és un valor propi de B de un vector propi Sv.

Teorema de Schur

Per a qualsevol matriu quadrada A, existeix una matriu U ortogonal tal que $T = U^{-1} A U$ és triangular superior.

Els elements de la diagonal de T són els valors propis de la matriu A.

Els valors propis i els vectors propis d'una matriu $n \times n$ on $n \geq 4$ s'han de trobar numèricament enlloc de fer-ho a mà.

Els mètodes numèrics que s'utilitzen a la pràctica depenen del significat geomètric de valors propis i vectors propis. L'essència de tots aquests mètodes es recull en el mètode de les potències.

Mètode de les potències

Teorema (MP)

A matriu quadrada de nombres reals d'ordre n tal que

1. $|\lambda_1| > |\lambda_2| \geq |\lambda_3| \geq \cdots \geq |\lambda_n|$
2. Existeixen n vectors propis linealment independents, $\left\{ v^{(1)}, v^{(2)}, \ldots, v^{(n)} \right\}$.
3. Per $x^{(0)} = a_1 v^{(1)} + a_2 v^{(2)} + \cdots + a_n v^{(n)}$ amb $a_1 \neq 0$

Llavors el mètode iteratiu $\boldsymbol{x^{(k)} = A\, x^{(k-1)}}$ és convergent i

$$\lim_{k\to\infty} x^{(k)} = \lim_{k\to\infty} A^k\, x^{(0)} = \lim_{k\to\infty} \lambda_1^k a_1 v^{(1)}\,.$$

El mètode iteratiu convergeix al valor propi de mòdul màxim

Fent ús de propietats citades anteriorment, podem establir els casos:

- **MP inversa (MPI)**

  Mètode de la potència per obtenir λ^{-1}, el valor propi de mòdul mínim: $A\,x^{(k)} = x^{(k-1)}$

- **MP amb desplaçament (MPD)**

  Mètode de la potència per obtenir $\lambda - \mu$, el valor propi de més llunyà a λ: $x^{(k)} = (A - \mu\, I)x^{(k-1)}$

- **MP inversa amb desplaçament (MPDI)**

  Mètode de la potència per obtenir $(\lambda - \mu)^{-1}$, el valor propi de més proper a λ: $(A - \mu\, I)\,x^{(k)} = x^{(k-1)}$

Calculeu, amb quatre xifres significatives, el valor propi dominant de la matriu:

$$A = \begin{pmatrix} 1 & 0 & -1 \\ 1 & 2 & 1 \\ 2 & 2 & 3 \end{pmatrix}$$

$$x^0_{max} = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}, \quad x^0_{min} = \begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix}$$

Exemple resolt

Vegeu el document . . . de MATLAB® del campus

Mètode de Wielandt

Mètode deflacció de Wielandt

Sigui λ_1 un valor propi de la matriu A de dimensió n i x_1 un vector propi associat tal que la seva primera component és igual a 1. Considerem la matriu següent:

$$B = A - x_1 a_1,$$

on a_1 denota la primera fila de la matriu A. Aquesta nova matriu té tota la primera fila igual a zero. Si λ_2 i x_2 són, respectivament, valor i vector propis d'A amb la primera component de x_2 igual a 1, podem escriure

$$B(x_1 - x_2) = \lambda_1 x_1 - \lambda_2 x_2 - \lambda_1 x_1 + \lambda_2 x_1 = \lambda_2 (x_1 - x_2).$$

Veure bibliografia, Grau - Noguera

Quocient de Rayleigh

La convergència del mètode de la potència depèn de la magnitud del quocient $\theta = |\lambda_2|/|\lambda_1|$. Si $\theta \approx 1$ calen més iteracions per a una exactitud fixada.

Per a matrius simètriques, s'aconsella la successió per a l'aproximació del valor propi:

$$r_k = \frac{\left(x^{(k)}\right)^t \cdot A \cdot x^{(k)}}{\left(x^{(k)}\right)^t \cdot x^{(k)}}.$$

La convergència de r_k cap a $|\lambda_1|$ és de l'ordre de $\left(\frac{|\lambda_2|}{|\lambda_1|}\right)^2$.

Es pot aplicar per qualsevol tipus de matriu, però no es manté l'ordre de convergència.

El sistema $x \cdot \lambda = A \cdot x$ és sobredeterminat, r_k és la solució de les equacions normals.

Mètode QR

El mètode de la potència i el mètode de la potència inversa només proporcionen una parella valor propi - vector propi; tots dos mètodes es poden modificar per donar totes les parelles valor i vector propi d'una matriu.
Hi ha, però mètodes per obtenir tots els valors propis en un sol algorisme. Un d'ells és l'anomentat mètode **QR**. Aquest és la base de tot el programari modern, inclòs MATLAB® , de càlcul de valors propis.

Factorització **QR**

El mètode **QR** es fonamenta en el fet que qualsevulla matriu quadradra admet descompossició QR: existeixen matrius **Q** i **R** tals que $A = QR$, amb Q ortogonal ($Q^{-1} = Q^t$) i R triangular superior.

El mètode consisteix en:

1. Transformem la matriu A en una matriu H Hessenberg superior fent ús de transformacions de Householder.
   Iniciem el mètode fent $H_1 = H$
2. Obtenim la factorització QR de la matriu H, $H_1 = Q_1 \, R_1$
3. Multiquem Q i R en ordre invers per obtenir una nova H, $H_2 = R_1 \, Q_1$
4. Repetim els pasos 2 i 3 fins que ...
5. La diagonal de H convergeix als valors propis.

Demostració.

La transformació $A \rightarrow H$ és una transformació de similaritat via matrius ortogonals.
La transformació $H_{i+1} \rightarrow H_i$ és una transformació de similaritat, verifica

$$H_{i+1} = R_i \, Q_i = Q_i^{-1} \, Q_i R_i \, Q_i = Q_i^t H_i Q_i$$

Llavors la diagonal de H convergeix als valors propis de la matriu A. □

Transformacions de Householder

Per $v \in \mathbb{R}^n$ definim la transformació associada a v per

$$H = \begin{cases} I, & v = 0, \\ I - \dfrac{2}{v^t v} v v^t, & v \neq 0. \end{cases}$$

Propietats

1. H és simètrica, ortogonal i $H^2 = I$.
2. Si $x \perp v, \Rightarrow Hx = x$.
3. Si $x \parallel v, \Rightarrow Hx = -x \;\; Hv = -v$.
4. Si $||x|| = ||v||, v = x - y \,, \Rightarrow Hx = y$.

Factorització QR

Per A és una matriu $m \times n$, les matrius ortogonals Q_k modifiquen les files $k, \cdots, m$ de la manera següent:

$$\underset{A}{\begin{bmatrix} x & x & x \\ x & x & x \\ x & x & x \\ x & x & x \\ x & x & x \end{bmatrix}} \xrightarrow{Q_1} \underset{Q_1A}{\begin{bmatrix} * & * & * \\ 0 & * & * \\ 0 & * & * \\ 0 & * & * \\ 0 & * & * \end{bmatrix}} \xrightarrow{Q_2} \underset{Q_2Q_1A}{\begin{bmatrix} x & x & x \\ 0 & * & * \\ 0 & 0 & * \\ 0 & 0 & * \\ 0 & 0 & * \end{bmatrix}} \xrightarrow{Q_3} \underset{Q_3Q_2Q_1A}{\begin{bmatrix} x & x & x \\ 0 & x & x \\ 0 & 0 & * \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}}$$

De $Q_n Q_{n-1} \cdots Q_1 A = R$, construïm

$$A = \underbrace{Q_1^t \ldots Q_{n-1}^t Q_n^t}_{Q} R = QR$$

Són una successió de simetries per transformar les columnes de la matriu A a una forma triangular superior

Exemple resolt

Vegeu el document ... de MATLAB® del campus

Valors singulars

Sigui A és una matriu $m \times n$, la matriu $A^t A$ és quadrada $n \times n$ simètrica i definida positiva, els seus n valors propis, λ_i, són reals i no negatius.

S'anomenen **valors singulars** de la matriu A les n arrels quadrades (pos.) dels valors propis no negatius de $A^t\,A$.

Valors singulars de A

$$\sigma_i = +\sqrt{\lambda_i}\,, \quad i = 1, \ldots, n$$

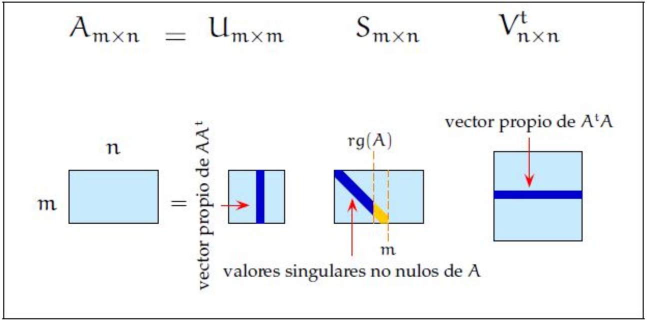
Ordenats usualment,

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n$$

Teorema

Per a qualsevol matriu A d'ordre $m \times n$, existeixen matrius ortogonals U d'ordre $m \times m$ i V d'ordre $n \times m$ i una matriu diagonal Σ d'ordre $m \times m$ tals que

$$A = U\Sigma V^t.$$



Valors singulars de A i rang A

$$A = USV^t = [U_r, U_{m-r}] \begin{bmatrix} \Sigma_r & 0 \\ 0 & 0 \end{bmatrix} [V_r, V_{n-r}]^t$$

La matriu Σ_r és diagonal amb els valors singulars no nuls.

En el passat, la manera convencional de determinar el rang d'una matriu A era convertir A en una forma triangular superior ... Però en els càlculs de coma flotant pot ser que no sigui tan fàcil decidir si algun nombre és efectivament zero o no. En altres paraules, sense mirar explícitament els valors singulars sembla que no hi ha cap manera satisfactòria d'assignar el rang a A.[1]

[1] Calculating the singular values and pseudo-inverse of a matrix, G. Golub, W. Kahan. J. SIAM Num. Anal. (1965).

Descomposició en Valors Singulars

Valors singulars de A i rang A

El càlcul de la svd permet escriure la matriu A com suma de r matrius de rang 1:

$$A = \sigma_1 u_1 v_1^t + \sigma_2 u_2 v_2^t + \cdots + \sigma_r u_r v_r^t,$$

on $u_1, \dots u_r$ i $v_1, \dots v_r$ són les r primeres columnes de les matrius U i V.

Els r (amb $r \leq m, n$) elements diagonals no nuls de Σ_r són els valors singulars de A. Si hi ha r valors singulars (no nuls) llavors **r** és el **rang de A**. Es suposa que els r valors singulars de A en la diagonal de Σ_r apareixen en ordre descendent.

L'algorisme de Golub i Reinsch (1970) per al càlcul de valors singulars no calcula en cap moment la matriu A^tA, sino que treballa directament sobre la matriu A.

Bàsicament, consisteix en dos grans passos: primer es transforma la matriu inicial en una de més senzilla i després s'aplica una variant del mètode QR per obtenir una successió de matrius convergent a una matriu diagonal que conté els valors singulars.

Descomposició en Valors Singulars

Teorema d'aproximació

El càlcul de la **SVD** permet escriure la matriu A com suma de r matrius de rang 1: $A = \Sigma_r u_1 v_1^t + \sigma_2 u_2 v_2^t + \cdots + \sigma_r u_r v_r^t$, on $u_1, \ldots u_r$ i $v_1, \ldots v_r$ són les r primeres columnes de les matrius U i V. Veure[2]

Teorema

La millor aproximació de A (en norma espectral) per matrius de mida $m \times n$ i rang $\leq k < r$ és la matriu

$$A_k = \sigma_1 u_1 v_1^t + \sigma_2 u_2 v_2^t + \cdots + \sigma_k u_k v_k^t,$$

obtinguda prenent només els k valors singulars més grans.

Amb norma espectral es verifica

$$||A||_2 = \sigma_1, \quad \text{i} \quad ||A - A_k||_2 = \sigma_{k+1}\,.$$

[2] Matrix Computations, G. Golub, C. Van Loan, Johns Hopkins Uni. Press, (1996).

Sigui A una matriu $m \times n$, la pseudoinversa de Moore-Penrose, A^+ és la única matriu $n \times m$ tal que

$$A\,A^+\,A = A\,, \quad A^+\,A\,A^+ = A^+.$$

La matriu A^+ es pot calcular a partir de la descompossició en valors singulars de la matriu A.

Valors singulars de A i matriu A^+

$$A = USV^t = [U_r, U_{m-r}] \begin{bmatrix} \Sigma_r & 0 \\ 0 & 0 \end{bmatrix} [V_r, V_{n-r}]^t \Longrightarrow A^+ = V_r \Sigma_r^{-1} U_r^t$$

La matriu Σ_r és diagonal amb els valors singulars no nuls.

Sigui A una matriu $m \times n$, $m \geq n$, b un vector de m components, x el vector de n d'incògnites. La matriu de coeficients del sistema d'equacions normals, $A^t A$ és no singular si i només si $rank(A) = n$

Teorema (Equacions normals)

Si x és la solució dels sistema d'equacions normals, $A'(b - Ax) = 0$, llavors

$$||r_x||_2 \leq ||r_y||_2.$$

Teorema (Pinversa)

La solució de residu mínim de $A\,x = b$ és $x = A^+\,b$.

Exemple

X	0.25	0.50	0.75	1.00	1.25	1.50	1.75
Y	0.40	0.50	0.90	1.28	1.60	1.66	2.02

Empreu una tècnica de mínims quadrats per ajustar la taula de dades a funcions del tipus:

1. $y = a_0 + a_1 x$. Determineu a_0 i a_1, doneu l'equació de la funció obtinguda i calculeu el vector residu en la solució.
2. $y = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + a_4 x^4$. Determineu a_0, a_1, a_2, a_3 i a_4, doneu l'equació de la funció obtinguda i calculeu el vector residu en la solució.
3. $y = a x^\alpha$. Determineu a i α, doneu l'equació de la funció obtinguda i calculeu el vector residu en la solució.
4. Quin dels tipus sembla el més adient. Per què?

Guia estudi

Guia estudi tema

Llibre Càlcul numèric: teoria i pràctica

- Conceptes i exercicis resolts: capítol 7, pàgines 250-297.
- Problemes proposats: 1, 2, 3, 4, 5, 6, i 8.
- Pràctiques resoltes : de la pàgina 298-304.
- Pràctiques proposades: pàgines 305-309.

Llibre Cálculo Científico con MATLAB® y Octave

- Conceptes i exercicis resolts: capítol 6, pàgines 173-190.
- Problemes i pràctiques proposades: del 6.1 al 6.10

Referències

- Llibre de consulta - Accès UPCommons,
  Càlcul numèric: teoria i pràctica
- Llibre de consulta - Accès UPCommons,
  Cálculo numérico
- Numerical Computing with MATLAB® ,
  Libro de texto de Cleve Moler
- Eigenvalues and Singular Values,
  Chapter 10
- Least Squares,
  Chapter 5

Computació Numèrica

Tema 3 - Interpolació polinòmica. Aproximació de funcions i dades

M. Àngela Grau Gotés

Departament de Matemàtiques
Universitat Politècnica de Catalunya · BarcelonaTech.

20 de març de 2023

"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-Ajust de dades

Introducció

La interpolació és un recurs de primer ordre dins del camp de l'aproximació de funcions.

- Per interpolació es pot substituir una funció d'expressió molt costosa (temps processador) d'avaluar per una altre més senzilla: polinomis, racionals,...
- Per interpolació es pot, a partir d'una taula de valors, $(x_i, f(x_i))_{i=0,\ldots,n}$, obtenir valors aproximats de $f(x)$ per a $x \neq x_i \quad i = 0, \ldots, n$
- Per interpolació es pot aproximar de funcions que no es poden obtenir per mètodes analítics.

Exemple

La taula de valors, reflexa la temperatura de congelació d'un anticongelant, una solució de glicerina (%) amb aigua.

%	C^o
0	0
10	-1.6
20	-4.8
30	-9.5
40	-15.4
50	-21.9
60	-33.6
70	-37.8
80	-19.1
90	-1.6
100	17

Qüestio: Quin serà el punt de congelació per un anti-congelant amb un 45% de glicerina?

Introducció

Les $n+1$ parelles $\{x_0, y_0\}, \{x_1, y_1\}, \ldots, \{x_n, y_n\}$, amb tots els x_i, diferents reben el nom de *nodes*, *nusos* o *punts de suport*.

Ens proposem construir una funció contínua $\widehat{f}$ que representi la llei (o la funció) que hi ha amagada darrera el conjunt de nodes donats.

- Polinòmica: $$\widehat{f}(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0 \,.$$
- Racional: $$\widehat{f}(x) = \frac{a_0 + a_1 x + \cdots + a_k x^k}{b_0 + \cdots + b_m x^m} \,.$$
- Exponencial: $$\widehat{f}(x) = a_n e^{\lambda_n x} + \cdots + a_1 e^{\lambda_1 x} + a_0 \,.$$
- Trigonòmetrica: $$\widehat{f}(x) = a_{-M} e^{-Mxj} + \cdots + a_0 + \ldots a_M e^{Mxj} \,,$$

  M enter igual a $n/2$ o $(n-1)/2$ segons paritat de n, j la unitat imagiària i la fómula d'Euler
  $$e^{kxj} = \cos(kx) + j\sin(kx) \,.$$

Interpolació polinòmica

Interpolació polinòmica

Donats $\{x_0, y_0\}, \{x_1, y_1\}, \dots, \{x_n, y_n\}$, volem determinar els $n+1$ coeficients del polinomi de grau n,

$$P(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0 ,$$

de tal manera que passi per tots els punts de suport $\{x_i, y_i\}$,

$$P(x_i) = y_i , \quad i = 0, \dots, n.$$

Les condicions, totes juntes, donen lloc a un sistema lineal de $n+1$ equacions i de $n+1$ incògnites: $a_n, a_{n-1}, \dots, a_1, a_0$. El determinant del sistema s'anomena determinant de Vandermonde.

Interpolació polinòmica

$$\begin{pmatrix} x_0^n & x_0^{n-1} & \cdots & x_0 & 1 \\ x_1^n & x_1^{n-1} & \cdots & x_1 & 1 \\ x_2^n & x_2^{n-1} & \cdots & x_2 & 1 \\ & & & & \\ x_n^n & x_n^{n-1} & \cdots & x_n & 1 \end{pmatrix} \cdot \begin{pmatrix} a_n \\ a_{n-1} \\ \\ a_1 \\ a_0 \end{pmatrix} = \begin{pmatrix} y_0 \\ y_1 \\ \\ y_{n-1} \\ y_n \end{pmatrix}$$

Les condicions, totes juntes, donen lloc a un sistema lineal de $n+1$ equacions i de $n+1$ incògnites: $a_n, a_{n-1}, \ldots, a_1, a_0$. El determinant del sistema s'anomena determinant de Vandermonde.

Existència i unicitat

Teorema

La solució del problema existeix i és única si tots els nodes x_i són diferents.

$$\begin{vmatrix} x_0^n & x_0^{n-1} & \cdots & x_0 & 1 \\ x_1^n & x_1^{n-1} & \cdots & x_1 & 1 \\ x_2^n & x_2^{n-1} & \cdots & x_2 & 1 \\ & & & & \\ x_n^n & x_n^{n-1} & \cdots & x_n & 1 \end{vmatrix} = \prod_{i=0}^{n} \prod_{j=i+1}^{n} (x_i - x_j)$$

És un sistema lineal gran, costos de resoldre i amb possible inestabilitat numèrica, aquest mètode de resolució no és viable.

Error

Sigui $f : [a, b] \subseteq \mathbb{R} \longrightarrow \mathbb{R}$ una funció amb derivades fins a l'ordre $n+1$ amb continuïtat, sigui $P_n(x)$ el polinomi interpolador de f en els nodes $a = x_0 < x_1 < ... < x_n = b$.

Sigui $w(x) = (x - x_0)(x - x_1)\ldots(x - x_n)$ i $\bar{x} \in [a, b]$, llavors existeix $c \in [a, b]$ tal que:

$$E(\bar{x}) = f(\bar{x}) - P_n(\bar{x}) = \frac{f^{n+1)}(c)}{(n+1)!} w(\bar{x})$$

Fórmules per calcular el polinomi interpolador

Càlcul del polinomi interpolador

Mètode directe

Fent ús de MATLAB® resolem el sistema lineal:

$$\begin{pmatrix} x_0^n & x_0^{n-1} & \cdots & x_0 & 1 \\ x_1^n & x_1^{n-1} & \cdots & x_1 & 1 \\ x_2^n & x_2^{n-1} & \cdots & x_2 & 1 \\ & & & & \\ x_n^n & x_n^{n-1} & \cdots & x_n & 1 \end{pmatrix} \cdot \begin{pmatrix} a_n \\ a_{n-1} \\ \\ a_1 \\ a_0 \end{pmatrix} = \begin{pmatrix} y_0 \\ y_1 \\ \\ y_{n-1} \\ y_n \end{pmatrix}$$

La solució obtinguda són els coeficients del polinomi en ordre decreixent: $P(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0$.

La dificultat que presenta la resolució de sistemes lineals grans, costosa i amb possible inestabilitat numèrica, fa que proposem altres formulacions que donen lloc al mateix poliniomi interpolador.

Càlcul del polinomi interpolador

Fórmula de Lagrange

El mètode de la fórmula de Lagrange és una manera d'obtenir el polinomi interpolador dels $n+1$ nodes

$$\{x_0, y_0\}, \{x_1, y_1\}, \ldots, \{x_n, y_n\} .$$

En aquest mètode el polinomi interpolador s'escriu de la forma:

$$P_n(x) = \sum_{k=0}^{n} y_k \cdot L_k(x), \quad L_k(x_j) = \delta_{kj}$$

Si $w(x) = (x - x_0)(x - x_1) \dots (x - x_n)$, llavors

$$L_k(x) = \frac{w(x)}{w'(x_k)(x - x_k)} .$$

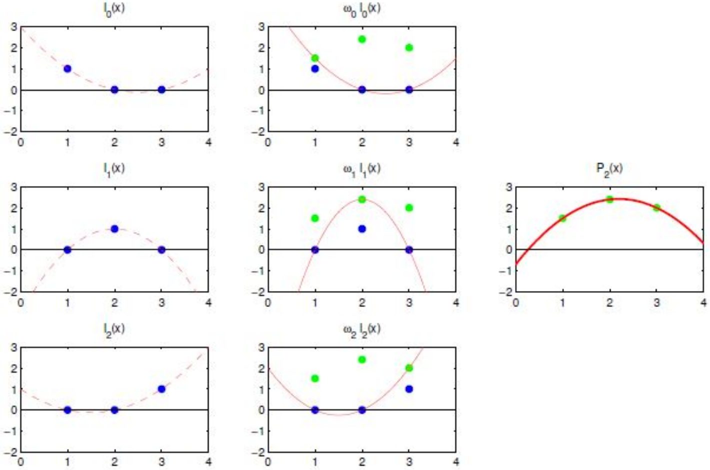
o equivalentment

$$L_k(x) = \frac{(x - x_0) \dots (x - x_{k-1})(x - x_{k+1}) \dots (x - x_n)}{(x_k - x_0) \dots (x_k - x_{k-1})(x_k - x_{k+1}) \dots (x_k - x_n)} .$$

★ MATLAB® la rutina `polyinterp(x,y,v)` retorna el valor del polinomi interpolador, fa ús de la fórmula de Lagrange

Càlcul del polinomi interpolador

Polinomis de Lagrange



Càlcul del polinomi interpolador
Diferències Dividides

El mètode de Newton de diferències dividides és una altra forma d'obtenir el polinomi interpolador dels $n+1$ punts $(x_0, f(x_0)), (x_1, f(x_1)), \ldots (x_n, f(x_n))$.

En aquest mètode el polinomi interpolador s'escriu de la forma:

$$\begin{aligned} P_n(x) = \; & c_0 + c_1(x - x_0) + c_2(x - x_0)(x - x_1) + \ldots \\ & \cdots + c_n(x - x_0) \ldots (x - x_{n-1}) . \end{aligned}$$

Per $(x_0, f(x_0)), (x_1, f(x_1)), \ldots (x_n, f(x_n))$, es defineixen

1. les **diferències dividides d'ordre** 0 de la funció f, per cada $i = 0, 1, \ldots, n$ es defineixen i noten per

$$f[x_i] = f(x_i).$$

2. les **diferències dividides d'ordre** 1 de la funció f, per cada $i = 0, 1, \ldots, n-1$ es defineixen i noten per

$$f[x_i, x_{i+1}] = \frac{f[x_{i+1}] - f[x_i]}{x_{i+1} - x_i}.$$

Càlcul del polinomi interpolador

Diferències Dividides - Notació

Partint de les diferències dividides d'ordre $k-1$,

$$f[x_i, x_{i+1}, \dots x_{i+k-1}] \quad \text{i} \quad f[x_{i+1}, \dots x_{i+k-1}, x_{i+k}]$$

es defineixen les **diferències dividides d'ordre** k corresponents a $x_i, x_{i+1}, \dots x_{i+k-1}, x_{i+k}$ per

$$\frac{f[x_{i+1}, \dots x_{i+k-1}, x_{i+k}] - f[x_i, x_{i+1}, \dots x_{i+k-1}]}{x_{i+k} - x_i}.$$

i es noten per

$$f[x_i, x_{i+1}, \dots x_{i+k-1}, x_{i+k}]$$

Càlcul del polinomi interpolador

Taula de diferències dividides

x	$f(x)$	Primeras diferencias divididas	Segundas diferencias divididas	Terceras diferencias divididas
x_0	$f[x_0]$			
		$f[x_0, x_1] = \dfrac{f[x_1] - f[x_0]}{x_1 - x_0}$		
x_1	$f[x_1]$		$f[x_0, x_1, x_2] = \dfrac{f[x_1, x_2] - f[x_0, x_1]}{x_2 - x_0}$	
		$f[x_1, x_2] = \dfrac{f[x_2] - f[x_1]}{x_2 - x_1}$		$f[x_0, x_1, x_2, x_3] = \dfrac{f[x_1, x_2, x_3] - f[x_0, x_1, x_2]}{x_3 - x_0}$
x_2	$f[x_2]$		$f[x_1, x_2, x_3] = \dfrac{f[x_2, x_3] - f[x_1, x_2]}{x_3 - x_1}$	
		$f[x_2, x_3] = \dfrac{f[x_3] - f[x_2]}{x_3 - x_2}$		$f[x_1, x_2, x_3, x_4] = \dfrac{f[x_2, x_3, x_4] - f[x_1, x_2, x_3]}{x_4 - x_1}$
x_3	$f[x_3]$		$f[x_2, x_3, x_4] = \dfrac{f[x_3, x_4] - f[x_2, x_3]}{x_4 - x_2}$	
		$f[x_3, x_4] = \dfrac{f[x_4] - f[x_3]}{x_4 - x_3}$		$f[x_2, x_3, x_4, x_5] = \dfrac{f[x_3, x_4, x_5] - f[x_2, x_3, x_4]}{x_5 - x_2}$
x_4	$f[x_4]$		$f[x_3, x_4, x_5] = \dfrac{f[x_4, x_5] - f[x_3, x_4]}{x_5 - x_3}$	
		$f[x_4, x_5] = \dfrac{f[x_5] - f[x_4]}{x_5 - x_4}$		
x_5	$f[x_5]$			

Pels nodes repetits, es considera $f[x_i, x_i] = f'(x_i) = y_i'$.

Càlcul del polinomi interpolador

Polinomi per diferències dividides

El polinomi interpolador de grau n s'escriu com:

$$\begin{aligned} P_n(x) = \quad & f[x_0] + f[x_0, x_1](x - x_0) + \\ & f[x_0, x_1, x_2](x - x_0)(x - x_1) + \\ & + \quad \dots \quad + \\ & f[x_0, x_1, \dots, x_n](x - x_0) \dots (x - x_{n-1}) \, . \end{aligned}$$

i la fórmula de l'error de l'aproximació és

$$E(\bar{x}) = f(\bar{x}) - P_n(\bar{x}) = \frac{f^{n+1)}(c)}{(n+1)!} w(\bar{x})$$

Fenòmen de Runge

Fenòmen de Runge

Construiu una taula per a la funció

$$f(x) = \frac{1}{1+25x^2}\,, \quad -1 \leq x \leq 1\,,$$

en $x = -0.9 \div 0.9\,(0.2)$.

Calculeu els polinomis interpoladors de grau 3, 6 i 9. Representeu graficament $f(x)$ i els polinomis obtinguts. Avalueu l'error que es comet en $x = -1 \div 1\,,(0.2)$.
Què s'observa? (**Fenòmen de Runge)**.

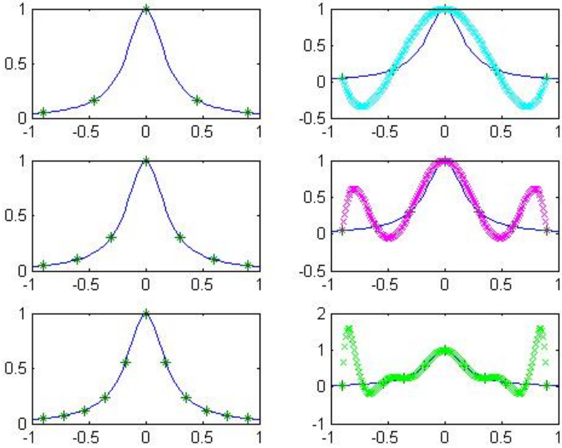


Figura: nodes equiespaiats

El·lecció òptima de nodes

El·lecció òptima de nodes

Sabem que la fòrmula de l'error per la interpolació polinòmica és

$$E(\bar{x}) = f(\bar{x}) - P_n(\bar{x}) = \frac{f^{n+1)}(c)}{(n+1)!} w(\bar{x})$$

on $w(x) = (x - x_0)(x - x_1) \dots (x - x_n)$ i $\bar{x} \in [a, b]$.

Ens interessa escollir els punts de manera que s'obtingui el mínim error possible. Per aconseguir això utilitzarem els polinomis de Chebyshev.

Abscisses de Chebyshev

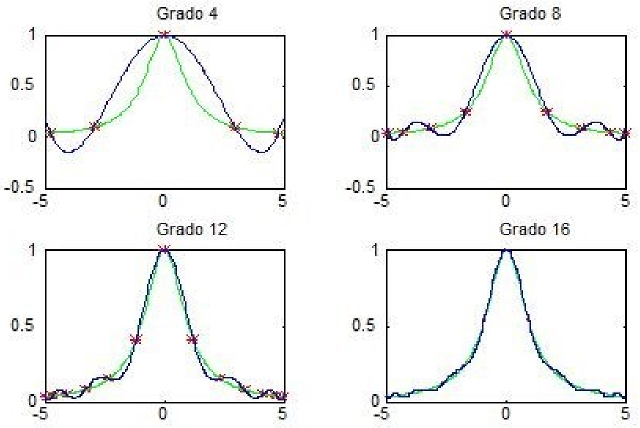


Figura: nodes de Chebyshev

Polinomis de Chebyshev

Els polinomis de Chebyshev de primer tipus són

$$T_n(x) = \cos(n \arccos(x)), \quad -1 \leq x \leq 1, \; n \geq 0.$$

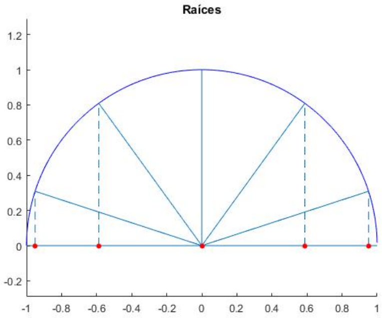
I per tant, $|T_n(x)| \leq 1\,, \quad -1 \leq x \leq 1.$

La recurrència és:

$$T_0(x) = 1, \quad T_1(x) = x,$$

$$T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x), \quad n \geq 2.$$

Abscisses de Chebyshev



Els nodes de Chebyshev no són equiespaiats i tenen la propietat que $w(x)$ és mínim a l'interval $[-1, 1]$.

$$x \in [-1, 1] \Rightarrow z(x) := \frac{b-a}{2}x + \frac{a+b}{2}, \quad z \in [a, b] .$$

Abscisses de Chebyshev

Les arrels del polinomi $T_n(x)$ són: (s'obtenen igualant $\cos(n\theta) = 0$)

$$x_k = \cos\left(\frac{(2k-1)\pi}{2n}\right), \quad k = 1, 2, \ldots, n$$

valor mínim

En general $\max|w(x)| \geq \dfrac{1}{2^n}$ i si els punts x_i són les arrels del polinomi de Chebyschev de grau $n+1$ es verifica

$$\max|w(x)| = \frac{1}{2^n}$$

Interpolació d'Hermite

Obtenir un polinomi

$$H_{2n+1}(x) = a_0 + a_1 x + \cdots + a_{2n+1} x^{2n+1}$$

de grau $\leq 2n+1$ que compleixi les condicions

$$H_m(x_j) = y_j, \quad H'_m(x_j) = y'_j$$

per la taula de dades

x	x_0	x_1	$\cdots$	x_n
y	y_0	y_1	$\cdots$	y_n
y'	y'_0	y'_1	$\cdots$	y'_n

$$H_{2n+1}(x) = \sum_{j=0}^{n} y_j H_{n,j}(x) + \sum_{j=0}^{n} y_j' \hat{H}_{n,j}(x)$$

amb

$$H_{n,j}(x) = \left[1 - 2(x - x_j)L_{n,j}'(x_j)\right] L_{n,j}^2(x)$$

i

$$\hat{H}_{n,j}(x) = (x - x_j)L_{n,j}^2(x)$$

on $L_{n,j}(x)$ és el polinomi de Lagrange de grau n; $L_{n,j}(x_j) = \delta_{i,j}$.

x	x_0	x_1	$\cdots$	x_n
y	y_0	y_1	$\cdots$	y_n
y'	y_0'	y_1'	$\cdots$	y_n'

Pels nodes repetits, es considera $f[x_i, x_i] = f'(x_i) = y_i'$.

x_0	$f[x_0]$					
x_0	$f[x_0]$	$f[x_0, x_0]$				
x_1	$f[x_1]$	$f[x_0, x_1]$	$f[x_0, x_0, x_1]$			
x_1	$f[x_1]$	$f[x_1, x_1]$	$f[x_0, x_1, x_1]$	$f[x_0, x_0, x_1, x_1]$		
x_2	$f[x_2]$	$f[x_1, x_2]$	$f[x_1, x_1, x_2]$	$f[x_0, x_1, x_1, x_2]$	$f[x_0, x_0, x_1, x_1, x_2]$	
x_2	$f[x_2]$	$f[x_2, x_2]$	$f[x_1, x_2, x_2]$	$f[x_1, x_1, x_2, x_2]$	$f[x_0, x_1, x_1, x_2, x_2]$	$f[$

$$
\begin{aligned}
H_5(x) = \quad & f[x_0] + f[x_0, x_0](x - x_0) + f[x_0, x_0, x_1](x - x_0)^2 \\
+ \quad & f[x_0, x_0, x_1, x_1](x - x_0)^2(x - x_1) \\
+ \quad & f[x_0, x_0, x_1, x_1, x_2](x - x_0)^2(x - x_1)^2 \\
+ \quad & f[x_0, x_0, x_1, x_1, x_2, x_2](x - x_0)^2(x - x_1)^2(x - x_2)\,.
\end{aligned}
$$

Interpolació d'Hermite

Taula diferències dividides

z	$f(z)$	Primeras diferencias divididas	Segundas diferencias divididas
$z_0 = x_0$	$f[z_0] = f(x_0)$		
		$f[z_0, z_1] = f'(x_0)$	
$z_1 = x_0$	$f[z_1] = f(x_0)$		$f[z_0, z_1, z_2] = \dfrac{f[z_1, z_2] - f[z_0, z_1]}{z_2 - z_0}$
		$f[z_1, z_2] = \dfrac{f[z_2] - f[z_1]}{z_2 - z_1}$	
$z_2 = x_1$	$f[z_2] = f(x_1)$		$f[z_1, z_2, z_3] = \dfrac{f[z_2, z_3] - f[z_1, z_2]}{z_3 - z_1}$
		$f[z_2, z_3] = f'(x_1)$	
$z_3 = x_1$	$f[z_3] = f(x_1)$		$f[z_2, z_3, z_4] = \dfrac{f[z_3, z_4] - f[z_2, z_3]}{z_4 - z_2}$
		$f[z_3, z_4] = \dfrac{f[z_4] - f[z_3]}{z_4 - z_3}$	
$z_4 = x_2$	$f[z_4] = f(x_2)$		$f[z_3, z_4, z_5] = \dfrac{f[z_4, z_5] - f[z_3, z_4]}{z_5 - z_3}$
		$f[z_4, z_5] = f'(x_2)$	
$z_5 = x_2$	$f[z_5] = f(x_2)$		

Interpolació d'Hermite

Expressió de l'error

Sigui $f : [a, b] \subseteq \mathbb{R} \longrightarrow \mathbb{R}$ una funció amb derivades fins a l'ordre $2n+2$ amb continuïtat, sigui $H_{2n+1}(x)$ el polinomi interpolador de f en els nodes $a = x_0 < x_1 < \ldots < x_n = b$.

Sigui $w^2(x) = (x - x_0)^2(x - x_1)^2 \ldots (x - x_n)^2$ i $\bar{x} \in [a, b]$, llavors existeix $c \in [a, b]$ tal que:

$$f(\bar{x}) - H_{2n+1}(\bar{x}) = \frac{f^{(2n+2)}(c)}{(2n+2)!} w^2(\bar{x})$$

Exercici

Trobeu el polinomi d'interpolació per la taula:

x	-1	2
f(x)	-11	14
f'(x)	14	5

emprant el mètode de les diferències dividides de Newton.

“Splines”

Interpolació polinomial a trossos

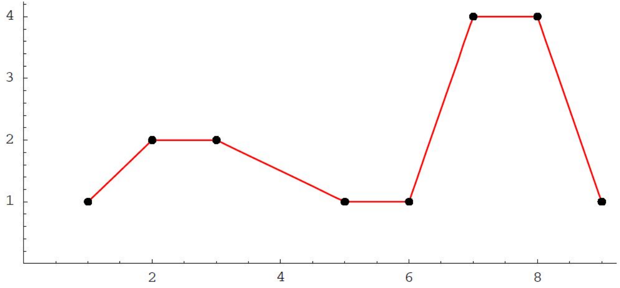
Interpolació polinomial a trossos

Spline

- Donats $(x_0, y_0), (x_1, y_1), \ldots (x_n, y_n)$, la idea és interpolar cada subinterval $[x_i, x_{i+1}]$ format per cada parella de nodes per un polinomi de grau baix.

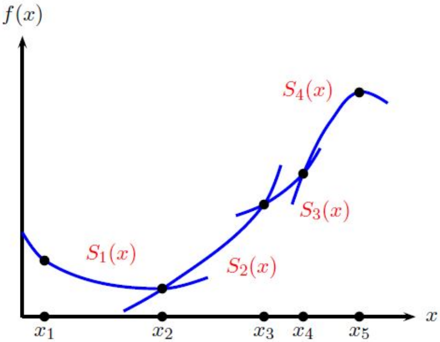
- Una spline és una corba definida per polinomis de grau k amb continuïtat fins la derivada $k-1$.

Spline Lineal



Una spline lineal és el cas més simple, els punts a interpolar es conecten per segments de recta; corba definida per polinomis de grau 1 amb continuïtat.

Spline Cúbic



Una spline cúbic és una corba definida per polinomis de grau 3 amb continuïtat fins la derivada 2 (segona).

Spline Cúbic

Les equacions d'una spline cúbic a l'interval $[x_i, x_{i+1}]$ per a $i = 0, 1, 2, 3, \ldots (n-1)$ serien:

$$S_i(x) = a_i(x - x_i)^3 + b_i(x - x_i)^2 + c_i(x - x_i) + d_i\ .$$

Les condicions són:

1. $S_i(x_i) = y_i, \quad S_(x_{i+1}) = y_{i+1} \qquad i = 0 \div (n-1).$
2. $S_{i+1}(x_{i+1}) = S_i(x_{i+1}), \quad i = 0, 1, \ldots (n-2).$
3. $S'_{i+1}(x_{i+1}) = S'_i(x_{i+1}), \quad i = 0, 1, \ldots (n-2).$
4. $S''_{i+1}(x_{i+1}) = S''_i(x_{i+1}), \quad i = 0, 1, \ldots (n-2).$

Spline Cúbic

En total són $4n$ incògnites i $4n-2$ condicions. Calen condicions adicionals, per exemple

$S_0''(x_0) = S_{n-1}''(x_n) = 0$, (spline cúbic natural)

$S_0'(x_0) = f'(x_0), \quad S_{n-1}'(x_n) = f'(x_n)$, (spline cúbic lligat)

Les derivades és calculen fent servir fórmules de derivació aproximada. Diferents mètodes de càlcul d'aquestes derivades, dóna lloc a diferents algorismes.

Consulteu els apartats 3.3, 3.4 i 3.5 del capítol 3 del llibre de Cleve Moler: *Numerical Computing with MATLAB®*.

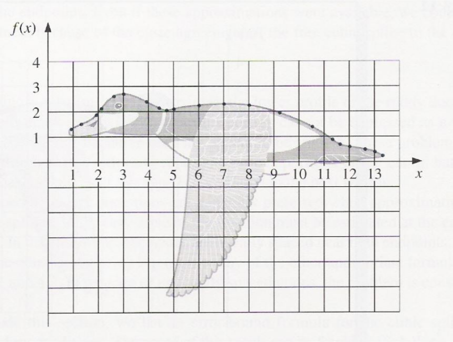
Spline Cúbic

$$
\begin{bmatrix}
1 & 0 & 0 & 0 & \cdots & \cdots & 0 \\
h_1 & 2(h_1+h_2) & h_2 & 0 & \cdots & & 0 \\
0 & h_2 & 2(h_2+h_3) & h_3 & 0 & & \vdots \\
\vdots & & \ddots & \ddots & \ddots & & \vdots \\
\vdots & & 0 & h_{n-2} & 2(h_{n-2}+h_{n-1}) & h_{n-1} & 0 \\
0 & & \cdots & 0 & h_{n-1} & 2(h_{n-1}+h_n) & h_n \\
0 & \cdots & \cdots & 0 & 0 & 0 & 1
\end{bmatrix}
\begin{bmatrix}
b_1 \\ b_2 \\ b_3 \\ \vdots \\ b_{n-1} \\ b_n \\ b_{n+1}
\end{bmatrix}
=
$$

$$
\begin{bmatrix}
0 \\
\frac{3}{h_2}(f_3-f_2) - \frac{3}{h_1}(f_2-f_1) \\
\frac{3}{h_3}(f_4-f_3) - \frac{3}{h_2}(f_3-f_2) \\
\vdots \\
\frac{3}{h_{n-1}}(f_n-f_{n-1}) - \frac{3}{h_{n-2}}(f_{n-1}-f_{n-2}) \\
\frac{3}{h_n}(f_{n+1}-f_n) - \frac{3}{h_{n-1}}(f_n-f_{n-1}) \\
0
\end{bmatrix}
$$

Ajust de corbes

Ajust de corbes



Consulteu l'apartat 3.5 del capítol 3 del llibre *Métodos Numéricos* de J. Douglas Faires & Richard Burden.

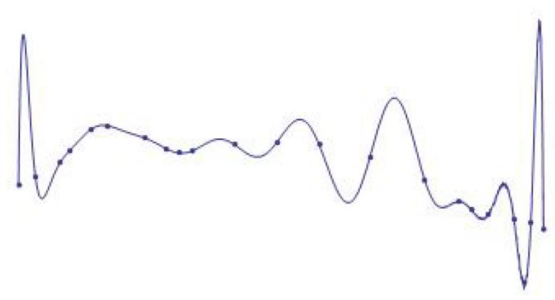


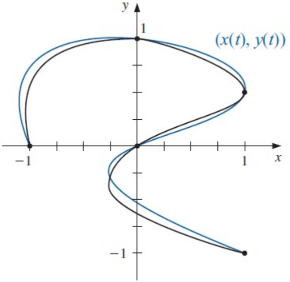
Figura: Polinomi interpolador

Figura: spline cúbic

Ajust de corbes

Corbes paramètriques

Una tècnica paramètrica per trobar un polinomi per connectar els punts en l'ordre previst consisteix en fer ús d'un paràmetre t en un interval $[t_0, t_n]$ per a $t_0 < t_1 < \cdots < t_n$ i construïr les funcions $x_i = x(t_i)$ i $y_i = y(t_i)$ per cada $i = 0, 1, \ldots, n$ fent ús de polinomis de Lagrange. També amb polinomis d'Hermite i spline.

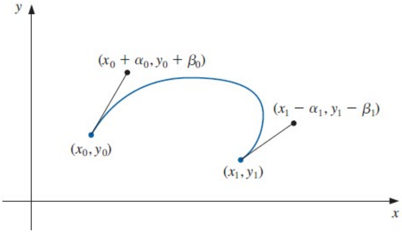


i	0	1	2	3	4
t_i	0	0.25	0.5	0.75	1
x_i	−1	0	1	0	1
y_i	0	1	0.5	0	−1

$$x(t) = \left(\left(\left(64t - \tfrac{352}{3}\right)t + 60\right)t - \tfrac{14}{3}\right)t - 1$$

$$y(t) = \left(\left(\left(-\tfrac{64}{3}t + 48\right)t - \tfrac{116}{3}\right)t + 11\right)t$$

El polinomis cúbics d'Hermite a trossos són els més emprats en computació gràfica: no cal refer tots els càlculs si es decideix modificar una part de la corba. Els nodes són (x_0, y_0) i (x_1, y_1) els punts de control són $(x_0 + \alpha_0, y_0 + \beta_0)$ i $(x_1 + \alpha 1, y_1 + \beta_1)$.



La corba de Bézier és una corba paramètrica, que a partir d'uns punts de control permeten a l'usuari controlar les pendents en aquests punts i modelitzar a voluntat. La seva aplicació inicial era el disseny de carrosseries d'automòbils, vaixells, hèlix de vaixells, ...

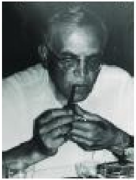
Iniciadors, Pierre Bézier a Renault i Paul de Faget de Casteljau a Citroën.

Els polinomis de Berstein són la base de les corbes de Bézier.

Splines

Paul de Faget de Casteljau 1930-
French mathematician/physicist
1958-1992: Citroën; unpublished work in **1958**

Pierre Bezier 1910-1999
1933-1975: engineer at Renault
1960: beginning of CADCAM work, Bezier curves



Isaac Jacob Schoenberg 1903-1990
Born in Romania (Landau's son-in-law). To USA in 1930.
Chicago, Harvard, Princeton, Swarthmore, Colby…
1941-1966: University of Pennsylvania
1943-1945: Army Ballistic Research Laboratory
1946: two papers on splines
1966-1973: U. of Wisconsin

Carl de Boor 1937-
Born in what became East Germany. To USA in 1959.
1960-1964: General Motors (grad student intern)
1962: first of many publications on splines
Purdue, Michigan…
1972- U. of Wisconsin

Desarrollo histórico de los splines y sus principales protagonistas. Trefethen [2005]

Corbes de Bézier

Les corbes de Bézier es poden conectar entre elles amb diverses continuïtats i ampliar-se per definir superfícies en $3D$.

$$B(t) = (1-t)P_0 + tP_1\,, \qquad 0 \leq t \leq 1$$

$$B(t) = (1-t)^2 P_0 + 2t(t-1)P_1 + t^2 P_2\,, \qquad 0 \leq t \leq 1$$

$$B(t) = (1-t)^3 P_0 + 3t(t-1)^2 P_1 + 3(t-1)t^2 P_2 + t^3 P_3\,, \qquad 0 \leq t \leq 1$$

Si voleu veure les imatges, teniu:

- Exemples de Corbes de Bézier de wikipedia.

B-spline

Una B-spline és una combinació lineal de splines *positives amb un suport compacte mínim*. El nom li va donar Isaac Jacob Schoenberg. Molts algoritmes, però numericament estable el de C. de Boor. Les B-spline són la generalització de les corbes de Bézier, que poden ser generalitzades per NURBS (Non-uniform rational B-splines).

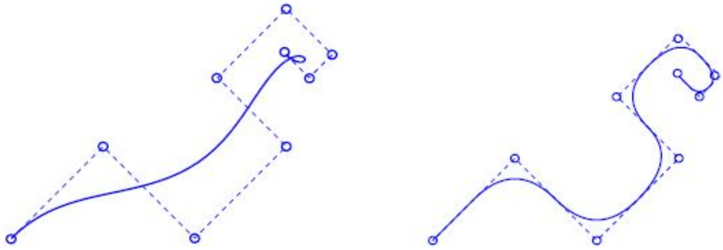
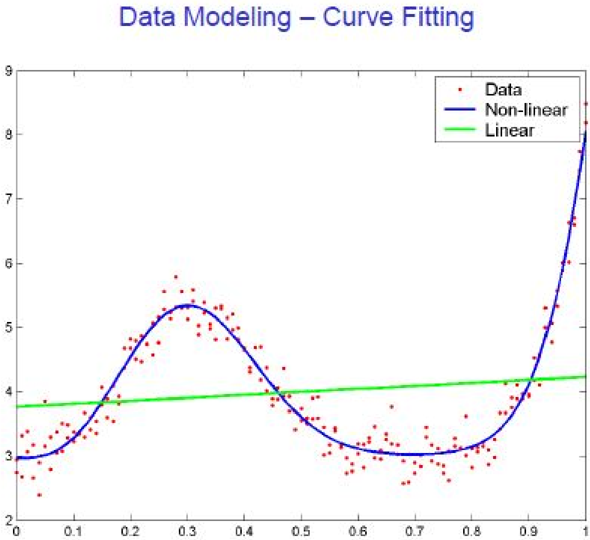


Figura: Corba de Bézier i B-spline per idèntics punts de control

Ajust de dades


Data Modeling – Curve Fitting
Data
Non-linear
Linear
9
8
7
6
5
4
3
2
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

Millor Aproximació

Quan a la taula de valors per a un mateix x_i tenim diversos valors de y_i el fet d'interpolar mitjançant polinomis no és possible, però podem construir una corba que s'ajusti el millor possible les dades disponibles, sense que la corba passi pels punts donats sino que "s'assembli" el més possible, per exemple minimitzant l'error quadràtic.

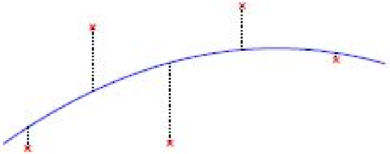
Cap mètode d'interpolació és apropiat per extrapolar informació de les dades disponibles, és a dir donar valors en punts fora de l'interval on es donen els nodes d'interpolació

Mètode dels mínims quadrats

L'aproximació la fem amb una funció

$$g(x) = \sum_{i=0}^{m} c_i \phi_i(x), \quad m < n, .$$

i es minimitza la suma de les distàncies dels nodes a la corba.



$$E^2 = \sum_{k=1}^{n} (g(x_k) - y_k)^2 \, .$$

Mètode dels mínims quadrats

El problema general és aproximar un conjunt de dades

$$(x_0, y_0), (x_1, y_1), \dots, (x_m, y_m)$$

per un polinomi $y = P_n(x)$ o una funció $y = g(x)$.

Es defineix el vector de residus r_x de components $r_i = y_i - g(x_i)$, quan dóna lloc a un sistema de m equacions i $n + 1$ incògnites, és un sistema d'equacions lineals sobredeterminat.

Mètode dels mínims quadrats

- Recta: $y = mx + b$.
- Paràbola: $y = ax^2 + bx + c$.
- Cúbica: $y = ax^3 + bx^2 + cx + d$.
- Potencial: $y = bx^m \quad \Rightarrow \quad \ln(y) = \ln(b) + m\ln(x)$.
- Exponencial: $y = be^{mx} \quad \Rightarrow \quad \ln(y) = \ln(b) + mx$.
- Logarítmica: $y = m\ln(x) + b$.
- Hiperbòlica: $y = \dfrac{1}{mx + b} \quad \Rightarrow \quad mx + b = \dfrac{1}{y}$.

Guia estudi tema

Llibre Càlcul numèric: teoria i pràctica

- Conceptes associats: Capítol 2,Interpolació polinòmica. Apartat 2.1, del 2.2 els apartats 2.2.1, 2.2.3, 2.2.4, 2.2.5, 2.2.7, 2.2.8, del 2.3 l'apartat 2.3.1.
- Problemes proposats: 1, 3, 5 i 16.
- Pràctiques resoltes: de la pàgina 68-71.
- Pràctiques proposades: pàgines 71-74.

Llibre Cálculo Científico con MATLAB y Octave

- Conceptes i exercicis resolts: capítol 3, pàgines 73-103.
- Problemes i pràctiques proposades: del 3.3 al 3.14

Referències

- Numerical Computing with MATLAB,
  Libros de texto de Cleve Moler
- Análisis Numérico,
  Richard L. Burden & Douglas J. Faires & Annette M. Burden,
  10a edición. Ed.Cengage Learning, 2016.
- Holistic Numerical Methods
  Topics of Numerical Methods
  Chapter 5 Interpolation
- E-notes, from Evgeny Demidov
  An Interactive Introduction to Splines

Computació Numèrica

Tema 4. Equacions no lineals

M. Àngela Grau Gotés

Departament de Matemàtiques
Universitat Politècnica de Catalunya · BarcelonaTech.

22 de març de 2023

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"

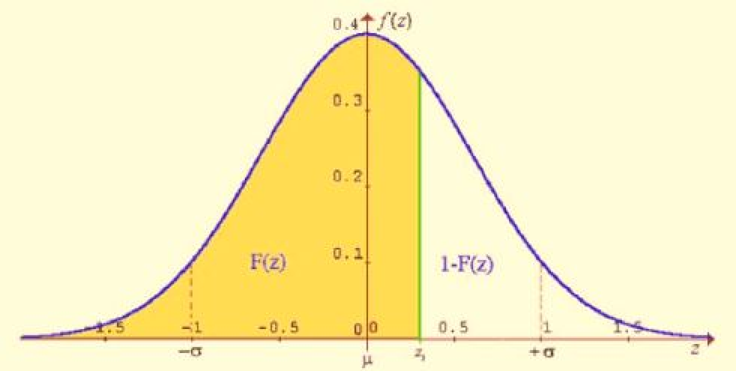
Índex - Equacions no lineals

Introducció

Molts fenòmens es descriuen per models no lineals i freqüentment cal resoldre una equació del tipus $f(x) = 0$, que no pot ser resolta per mètodes algebraics coneguts.

La major part d'aquest capítol es refereix a la solució aproximada d'una equació no lineal. No obstant això, també s'estudiaran els sistemes d'equacions no lineals, més complexes per resoldre i obtenir solucions aproximades.

Introducció



$$F(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{\frac{-t^2}{2}} dt$$

Nomenclatura

Sigui $f : D \subseteq \mathbb{R} \to \mathbb{R}$ una funció real de variable real.

1. α és un **zero de** f si $f(\alpha) = 0$.

Nomenclatura

Sigui $f : D \subseteq \mathbb{R} \to \mathbb{R}$ una funció real de variable real.

1. α és un **zero de** f si $f(\alpha) = 0$.
2. x^* és un **punt fix de** f si $f(x^*) = x^*$.

Nomenclatura

Sigui $f : D \subseteq \mathbb{R} \to \mathbb{R}$ una funció real de variable real.

1. α és un **zero de** f si $f(\alpha) = 0$.
2. x^* és un **punt fix de** f si $f(x^*) = x^*$.
3. α és una **solució** o **arrel** de l'equació $f(x) = p$ si

$$f(\alpha) = p\,.$$

Nomenclatura

Multiplicitat d'una arrel

Una solució α de $f(x) = 0$ es diu que té multiplicitat **n** si

$$f(\alpha) = f'(\alpha) = \cdots = f^{n-1)}(\alpha) = 0, \text{ i } f^{n)}(\alpha) \neq 0.$$

Si la multiplicitat és 1, es diu que l'arrel és **simple**.

NB. Determinar iterativament arrels múltiples és un problema mal condicionat.

Exemples

1. Dues arrels simples,
   $f(x) = x^2 + x - 2 = (x+2)(x-1)$.
2. Una arrel doble, $f(x) = x^2 + 4x + 4 = (x+2)^2$.
3. Sense fórmula directa per calcular les arrels dels polinomis de grau superior al 4,
   $f(x) = x^5 + 5x^3 + 4x^2 + 1$.
4. Per equacions amb funcions transcendents només la solució numèrica és factible, $f(x) = x^2 - e^{-x}$.

El mètode

No totes les equacions tenen un únic zero simple en el seu domini, llavors per calcular solucions aproximades, per a la convergència dels mètodes en qualsevol procés de càlcul d'arrels d'una equació no lineal consta de tres pasos:

3. **Aproximació.** Determinar una successió $\{x_n\}_{n\in\mathbb{N}}$ convergent al valor α solució de l'equació plantejada:

$$x_n \to \alpha\,,\ \ f(\alpha) = 0\,.$$

El mètode

No totes les equacions tenen un únic zero simple en el seu domini, llavors per calcular solucions aproximades, per a la convergència dels mètodes en qualsevol procés de càlcul d'arrels d'una equació no lineal consta de tres pasos:

1. **Localització**. Conèixer la zona on es troben les arrels. Un estudi analític o una representació gràfica.

3. **Aproximació**. Determinar una successió $\{x_n\}_{n\in\mathbb{N}}$ convergent al valor α solució de l'equació plantejada:

$$x_n \to \alpha\,, \;\; f(\alpha) = 0\,.$$

El mètode

No totes les equacions tenen un únic zero simple en el seu domini, llavors per calcular solucions aproximades, per a la convergència dels mètodes en qualsevol procés de càlcul d'arrels d'una equació no lineal consta de tres pasos:

1. **Localització**. Conèixer la zona on es troben les arrels. Un estudi analític o una representació gràfica.
2. **Separació**. Determinar dominis amb una única arrel.
3. **Aproximació**. Determinar una successió $\{x_n\}_{n\in\mathbb{N}}$ convergent al valor α solució de l'equació plantejada:

$$x_n \to \alpha\,,\ \ f(\alpha) = 0\,.$$

Localització i separació

- **TEOREMA DE BOLZANO**
  Per $f : \mathcal{I} \to \mathbb{R}$, $[a, b] \subset \mathcal{I}$ i $a < c < b$, llavors

Teorema

Si f és contínua en l'interval tancat $[a, b]$ i $f(a)$ i $f(b)$ tenen signes diferents, aleshores existeix $c \in (a, b)$ tal que $f(c) = 0$.

Localització i separació

- **TEOREMA DE BOLZANO**
  Per $f : \mathcal{I} \to \mathbb{R}$, $[a, b] \subset \mathcal{I}$ i $a < c < b$, llavors

Teorema

Si f és contínua en l'interval tancat $[a, b]$ i $f(a)$ i $f(b)$ tenen signes diferents, aleshores existeix $c \in (a, b)$ tal que $f(c) = 0$.

- **TEOREMA DE ROLLE**
  Per $f : \mathcal{I} \to \mathbb{R}$, $(a, b]) \subset \mathcal{I}$ i $a < c < b$, llavors

Teorema

Si f és derivable en l'interval obert (a, b) i $f(a) = f(b)$, aleshores existeix $c \in (a, b)$ tal que $f'(c) = 0$.

Localització i separació

- **TEOREMA DE BOLZANO**
  Per $f : \mathcal{I} \to \mathbb{R}$, $[a, b] \subset \mathcal{I}$ i $a < c < b$, llavors

Teorema

Si f és contínua en l'interval tancat $[a, b]$ i $f(a)$ i $f(b)$ tenen signes diferents, aleshores existeix $c \in (a, b)$ tal que $f(c) = 0$.

- **TEOREMA DE ROLLE**
  Per $f : \mathcal{I} \to \mathbb{R}$, $(a, b]) \subset \mathcal{I}$ i $a < c < b$, llavors

Teorema

Si f és derivable en l'interval obert (a, b) i $f(a) = f(b)$, aleshores existeix $c \in (a, b)$ tal que $f'(c) = 0$.

Aproximació: Tipus de mètodes

La successió $\{x_n\}_{n\in\mathbb{N}}$ convergent al valor α solució de l'equació

$$x_n \to \alpha\,,\;\; f(\alpha) = 0\,.$$

- Mètode d'intervals encaixats.

$$[a_1, b_1] \supseteq [a_2, b_2] \supseteq \cdots \supseteq [a_n, b_n] \ldots$$

$$a_n \leq x_n \leq b_n\,, \quad \{b_n - a_n\}_n \to 0\,.$$

- Esquemes o algorismes iteratius:

$$x_n = g(x_{n-1}, x_{n-2}, \ldots)\,.$$

Intervals encaixats

Objectiu

Obtenir $\{x_n\}_{n\in\mathbb{N}}$ successió convergent de nombres reals

$$x_n \xrightarrow[n\to+\infty]{} \alpha \text{ tal que } f(\alpha) = 0\,.$$

Procediment

Obtenir successions $\{a_n\}_{n\in\mathbb{N}}$, $\{b_n\}_{n\in\mathbb{N}}$ i $\{x_n\}_{n\in\mathbb{N}}$ tals que

$$\{b_n - a_n\}_{n\in\mathbb{N}} \to 0\,, \quad a_n \leq x_n \leq b_n\,,$$

$$[a_1, b_1] \supseteq [a_2, b_2] \supseteq \cdots \supseteq [a_n, b_n] \ldots$$

Mètode de la bisecció

Mètodes dels intervals encaixats

Per $f : \mathcal{I} \to \mathbb{R}$, contínua, $[a, b] \subset \mathcal{I}$ tal que $f(a) \cdot f(b) < 0$, (teorema de Bolzano) calculem el punt mig de l'interval

$$m = \frac{a+b}{2}\ .$$

Aquest punt verifica un dels tres ítems

- $f(m) = 0$
- $f(a)f(m) < 0$, nou interval de Bolzano $[a, m]$.
- $f(a)f(m) < 0$, nou interval de Bolzano $[m, b]$.

Mètode de la bisecció

Algorisme

Començant amb l'interval $[a, b]$ tal que $f(a) \cdot f(b) < 0$, i procedint com a la pàgina anterior, construïm una successió d'intervals $[a_n, b_n]$ tal que $f(a_n) \cdot f(b_n) < 0$, els punts mitjos dels quals

$$x_{n+1} = \frac{a_n + b_n}{2}$$

són una aproximació de l'arrel α. Cada interval té la meitat de la longitud de l'interval anterior.

Solució aproximada:

$$\alpha = x_n \pm \ell_n\,, \qquad \ell_n = \frac{b-a}{2^n}\,, \quad n > 0.$$

Mètode de la bisecció

Algorisme

1 $a_0 = a,\ b_0 = b,$

2 Per a $n = 0, 1, \ldots,$ *fer* : $x_{n+1} = \dfrac{a_n + b_n}{2}$ i

Si $f(a_n)f(x_{n+1}) < 0$, pendre $a_{n+1} = a_n,\ \ b_{n+1} = x_{n+1},$
altrament, pendre $a_{n+1} = x_{n+1},\ \ b_{n+1} = b_n.$

Anàlisi de l'error:

$$|x_{n+1} - \alpha| \leq |b_{n+1} - a_{n+1}| < \frac{|b-a|}{2^{n+1}}.$$

Mètode de la bisecció

Criteri d'aturada: Donat $tol = \eta > 0$

$$\frac{|b-a|}{2^{n+1}} < \eta$$

Càlcul previ nombre iteracions: Donat $tol = \eta > 0$

$$n > \frac{\log\left(\frac{|b-a|}{\eta}\right)}{\log 2} - 1$$

Mètode de la Regula Falsi

Començant amb $f(a_0) \cdot f(b_0) < 0$ i l'interval $I_0 = [a_0, b_0]$, es construeix una successió de punts definits per

$$x_{n+1} = a_n - f(a_n)\frac{(b_n - a_n)}{f(b_n) - f(a_n)}\,, \quad n \geq 0$$

i una successió d'intervals encaixats $I_{n+1} = [a_{n+1}, b_{n+1}]$ tal que:

- Si $f(a_n)f(x_{n+1}) < 0$, pendre $a_{n+1} = a_n, \;\; b_{n+1} = x_{n+1}$,
- altrament, pendre $a_{n+1} = x_{n+1}, \;\; b_{n+1} = b_n$.

Criteri d'aturada: Donat $tol = \eta > 0$

$$|x_{n+1} - x_n| < \eta \quad \text{i} \quad |f(x_{n+1})| < \eta$$

Exercici

Determineu l'arrel real de

$$f(x) = x^3 - x + 1.$$

- Representeu gràficament la funció. Doneu un interval on es trobi un zero de la funció.
- Apliqueu el mètode de la bisecció ($\eta = 0.001$).
- Apliqueu el mètode de la regula falsi ($\eta = 0.00005$).

Mètodes iteratius

Objectiu

Obtenir $\{x_n\}_{n\in\mathbb{N}}$ successió convergent de nombres reals

$$x_n \underset{n\to+\infty}{\longrightarrow} \alpha \text{ tal que } f(\alpha) = 0\,.$$

Procediment

Escriure $f(x) = 0$ com $x = g(x)$ establir un esquema iteratiu del tipus

$$x_n = g(x_{n-1}, x_{n-2}, \dots)\,,\;\; n > 0.$$

Mètode de Newton

Mètode de Newton-Raphson o mètode de la tangent

Algorisme

Començant amb el valor x_0 es construeix una successió de punts definits per

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} \, .$$

La coordenada x_{n+1} és el punt de tall de la recta tangent per $P(x_n, f(x_n))$ amb l'eix d'absices.

Criteri d'aturada: Donat $tol = \eta > 0$

$$|x_{n+1} - x_n| < \eta \quad \text{i} \quad |f(x_{n+1})| < \eta$$

CONVERGÈNCIA?

Mètode de Newton

Convergència

Regla de Fourier

Sigui $f : \mathcal{I} \to \mathbb{R}$, contínua i derivable, $[a, b] \subset \mathcal{I}$ tal que:

1 $f(a) \cdot f(b) < 0$,

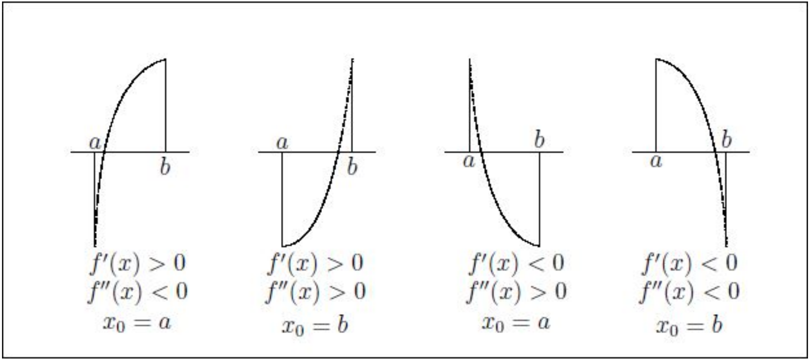
2 $f'(x) \cdot f''(x) \neq 0\,, \ \forall x \in [a, b]$,

3 començant amb el valor

$$x_0 = \begin{cases} a & si \quad f(a) \cdot f''(a) > 0, \\ b & si \quad f(b) \cdot f''(b) > 0, \end{cases}$$

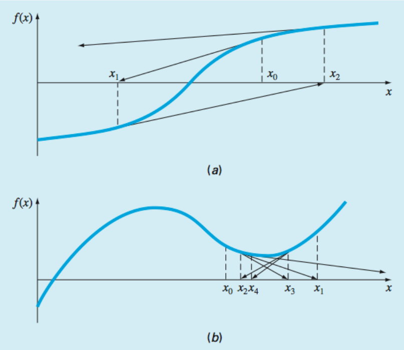
llavors, la successió de punts definits per $x_{n+1} = x_n - \dfrac{f(x_n)}{f'(x_n)}$,

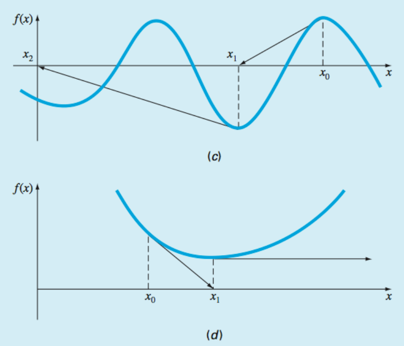
convergeix a la única arrel de $f(x) = 0$ a l'interval $[a, b]$.


a
b
$f'(x) > 0$
$f''(x) < 0$
$x_0 = a$
a
b
$f'(x) > 0$
$f''(x) > 0$
$x_0 = b$
b
a
$f'(x) < 0$
$f''(x) > 0$
$x_0 = a$
b
a
$f'(x) < 0$
$f''(x) < 0$
$x_0 = b$

Mètode de Newton

Convergència lenta mètode de Newton





Mètode de la secant

Algorisme

Començant amb dos valors x_0 i x_1 es construeix una successió de punts definits per

$$x_{n+1} = x_n - f(x_n)\frac{(x_n - x_{n-1})}{f(x_n) - f(x_{n-1})}$$

o equivalent

$$x_{n+1} = \frac{x_{n-1}f(x_n) - x_n f(x_{n-1})}{f(x_n) - f(x_{n-1})}.$$

Criteri d'aturada: Donat $\eta > 0$

$$|x_{n+1} - x_n| < \eta \quad \text{i} \quad |f(x_{n+1})| < \eta$$

CONVERGÈNCIA?

Algorismes - Fita de l'error

Sigui $\alpha \in \mathbb{R}$ l'arrel de $f(x) = 0$, $\mathcal{J}_\alpha$ un entorn tancant de α i $\{x_n\}_{n\in\mathbb{N}}$ successió convergent $x_n \underset{n\to+\infty}{\longrightarrow} \alpha$ tal que $f(\alpha) = 0$.

Fita "a posteriori"

Si f és derivable en $\mathcal{J}_\alpha$ i $x_n \in \mathcal{J}_\alpha$, es verifica que:

$$\epsilon_n = |x_n - \alpha| \leq \frac{|f(x_n)|}{\min_{x\in\mathcal{J}_\alpha} |f'(x)|} .$$

TVM: $\quad |f(x_n) - f(\alpha)| = |f'(c)| \cdot |(x_n - \alpha)|$

Exercici (continuació)

Determineu l'arrel real de

$$f(x) = x^3 - x + 1.$$

1. Apliqueu el mètode de Newton ($\eta = 0.00005$).
2. Apliqueu el mètode de la secant amb una precisió de quatre decimals correctes.
3. Quin mètode necessita més iteracions? Quin menys? Quin mètode dona una millor aproximació? Quin pitjor? Comenta les diferències trobades.

Mètode de la iteració simple

Mètode de la iteració simple o mètode del punt fix

Fent ús d'operacions elementals, la equació $\mathbf{f}(x) = 0$ es pot expressar com $x = \mathbf{g}(x)$, on $\mathbf{g}$ és una funció contínua.

$$f(x) = 0 \overset{\textit{consistent}}{\Longleftrightarrow} x = g(x)$$

Iteració simple

Una aproximació inicial x_0 dóna lloc a la successió $x_{n+1} = \mathbf{g}(x_n)$.

Punt fix

Si la successió $x_{n+1} = \mathbf{g}(x_n), n > 0$ és convergent a un valor α, llavors α és un punt fix de $\mathbf{g}$.

Mètode de la iteració simple

Exercicis

Exemple

L'equació $x - \cos x = 0$ es pot transformar en:

$$x = \cos x, \quad x = \frac{x + \cos x}{2}, \quad x = \frac{2x + \cos x}{3}, \quad x = \sqrt{x \cos x}$$

Successions convergents

$$x_{n+1} = \frac{x_n + \cos(x_n)}{2}, \;\; x_0 = 1 \ldots x_7 = 0.73909,.$$

$$x_{n+1} = \frac{2x_n + \cos(x_n)}{3}, \;\; x_0 = 1 \ldots x_{14} = 0.73909$$

Observació

La successió $\{x_n\}$ pot no convergir malgrat s'esculli x_0 molt proper al punt fix.

Exercici

Determineu l'arrel real de $x = \cos x$

1. Representeu gràficament la funció. Doneu un interval on es trobi un zero de la funció.
2. Prenent $x_0 = 0$, calculeu 15 iterats dels mètodes iteratius
$$x_{n+1} = \frac{x_n + \cos(x_n)}{2}, \quad x_{n+1} = \frac{2x_n + \cos(x_n)}{3}.$$
3. Prenent $x_0 = 1$, calculeu 15 iterats dels mètodes iteratius
$$x_{n+1} = \cos(x_n), \quad x_{n+1} = \sqrt{x_n \cos(x_n)}.$$
4. Quin mètode és convergent? Quin és divergent?

Sigui $\alpha \in \mathbb{R}$ el punt fix de $x = g(x)$ i $\mathcal{J}_\alpha$ un entorn de α .

Teorema de convergència

Si g és derivable i $|g'(x)| \leq k < 1$ en $\mathcal{J}_\alpha$. Llavors, $\forall x_0 \in \mathcal{J}_\alpha$, la successió $x_{n+1} = g(x_n), \quad n > 0$ verifica que:

a) $x_n \in \mathcal{J}_\alpha \qquad n = 0, 1, 2, \ldots$

b) $\lim\limits_{n\to\infty} x_n = \alpha$.

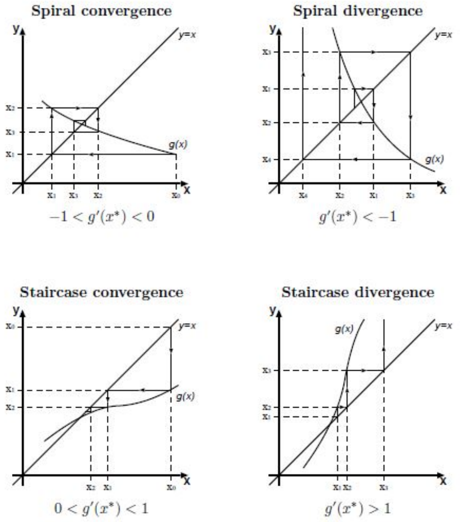
c) α és la única arrel de $x = g(x)$ dins de $\mathcal{J}_\alpha$.

observació

$$|x_{n+1} - \alpha| \leq k|x_n - \alpha| \leq \cdots \leq k^{n+1}|x_0 - \alpha| .$$

Mètodes del punt fix

Teorema de convergència-Gràfics



Estimació de l'error

Si comptem els errors d'arrodoniment, $\bar{x}_{n+1} = g(\bar{x}_n) + \delta$,

Fita superior error (I)

$$|\bar{x}_{n+1} - \alpha| < \frac{k}{1-k} |\bar{x}_{n+1} - \bar{x}_n| + \frac{1}{1-k} \delta \, .$$

Si l'aritmètica és exacte, $\bar{x}_{n+1} = g(\bar{x}_n)$,

Fita superior error (II)

$$|\bar{x}_{n+1} - \alpha| < \frac{k^{n+1}}{1-k} |\bar{x}_1 - \bar{x}_0| \, .$$

Ordre de convergència

Ordre de convergència

Caracterització

Definició

La successió de punts $(x_n)_{n\in\mathbb{N}}$, i el mètode que la genera, té **ordre de convergència** almenys $\mathbf{p} \geq 1$ si, per a qualsevol punt $x_0 \in \mathcal{J}_\alpha$, existeix $C > 0$ tal que

$$|x_{n+1} - \alpha| < C|x_n - \alpha|^p \,, \quad (\text{si } p = 1, C < 1).$$

En el cas que

$$\lim_{n\to\infty} \frac{|x_{n+1} - \alpha|}{|x_n - \alpha|^p} = L$$

direm que la successió té ordre de convergència almenys p; si $p = 1$ cal $|L| < 1$.

Ordre de convergència

Exemples

Zero simple

- Convergència almenys lineal del mètode de la iteració simple si $|g'(x)| < 1$ per a $x \in \mathcal{J}_\alpha$.
- Convergència almenys lineal del mètode de la Regula Falsi.
- Convergència almenys quadràtica del mètode de Newton.
- Convergència almenys superlineal del mètode de la Secant:

$$\frac{1+\sqrt{5}}{2} \approx 1.61803.$$

Decimals correctes en cada iteració

$$d_{n+1} = -\log_{10}|x_{n+1} - \alpha| \approx -p\log_{10}|x_n - \alpha| - \log_{10} L$$

$$d_{n+1} \approx p \cdot d_n$$

Ordre de convergència

Aproximacions

PCLOC

$$\widehat{\lambda}_n = \frac{\ln |f(x_n)|}{\ln |f(x_{n-1})|}, \qquad n > 1.$$

ACLOC

$$\widetilde{\lambda}_n = \frac{\ln |x_n - x_{n-1}|}{\ln |x_{n-1} - x_{n-2}|}, \qquad n > 2.$$

Sigui $\{x_n\}_{n\in\mathbb{N}}$ successió **linealment** convergent α tal que $f(\alpha) = 0$.

Observació

$$\left.\begin{aligned} |x_{n+2}-\alpha| &= \kappa|x_{n+1}-\alpha| \\ |x_{n+1}-\alpha| &= \kappa|x_n-\alpha| \end{aligned}\right\} \quad \Rightarrow \quad \alpha = \frac{x_{n+2}x_n - x_{n+1}^2}{x_{n+2} - 2x_{n+1} + x_n}$$

Les diferències progressives endavant, es defineixen per

$$\Delta x_{n+1} = (x_{n+1} - x_n)$$

i per $k > 1$,

$$\Delta^{(k)} x_{n+1} = \Delta(\Delta^{(k-1)} x_{n+1}) .$$

Mètode Δ^2 d'Aitken

$$x'_{n+2} = \frac{x_{n+2}x_n - {x_{n+1}}^2}{x_{n+2} - 2x_{n+1} + x_n} = x_{n+2} - \frac{(\Delta x_{n+1})^2}{\Delta^2 x_n}$$

Llavors $x'_n \xrightarrow[n\to+\infty]{} \alpha$ més ràpidament, en el sentit que

$$\lim_{n\to\infty} \frac{|x'_n - \alpha|}{|x_n - \alpha|} = 0\,.$$

A partir d'un procés $x_{k+1} = g(x_k)$ de primer ordre, i unes iteracions, x_0, x_1 i x_2, calculem x'_2, a partir de les iteracions, x_1 ,x_2 i x_3 calculem x'_3, i successivament.

Donats x_0, $x_1 = g(x_0)$ i $x_2 = g(x_1)$ d'un procés $x_{k+1} = g(x_k)$ de primer ordre, calculem $x_0'' = x_0 - \dfrac{(x_1 - x_0)^2}{x_2 - 2x_1 + x_0}$.

Amb la terna $x_0 = x_0''$, $x_1 = g(x_0'')$ i $x_2 = g(g(x_0''))$, calculem un nou: $x_0'' = x_0 - \dfrac{(x_1 - x_0)^2}{x_2 - 2x_1 + x_0}$.

Mètode de Steffensen

En general, per cada $n \geq 0$, definim $x_0^{(n+1)} = x_n''$, $x_1^{(n+1)} = g(x_0^{(n+1)})$, $x_2^{(n+1)} = g(x_1^{(n+1)})$, i finalment

$$x_{n+1}'' = x_0^{(n+1)} - \frac{\left(x_1^{(n+1)} - x_0^{(n+1)}\right)^2}{x_2^{(n+1)} - 2x_1^{(n+1)} + x_0^{(n+1)}}.$$

Llavors $x_n'' \xrightarrow[n\to+\infty]{} \alpha$ més ràpidament que el mètode del punt fix inicial i que el mètode d'Aitken.

Observeu el cas següent: $x_{n+1} = e^{-x_n}$ i $x_0 = 0.5$

Normal	Aitken	Steffensen
0.5		
0.606530660		
0.545239212	0.567623876	0.567623876
0.579703095	0.567298989	
0.560064628	0.567193142	
0.571172149	0.567159364	0.567143314
0.564862947	0.567148453	
0.568438048	0.567144952	
0.566409453	0.567143825	$\boxed{0.567143290}$

Sistemes d'equacions (no lineals)

Sistemes d'equacions no lineals

La funció $F : D \subseteq \mathbb{R}^n \longrightarrow \mathbb{R}^n$, $F = (f_1, f_2, \dots, f_m)$, de diverses variables dóna lloc al sistema d'equacions no lineals

$$F(\mathbf{z}) = 0\,,$$

que també es pot escriure com

$$\left\{\begin{array}{c} f_1(z_1, \dots, z_n) = 0, \\ f_2(z_1, \dots, z_n) = 0, \\ \vdots \\ f_n(z_1, \dots, z_n) = 0. \end{array}\right. \tag{1}$$

El mètode de la iteració simple

O el mètode del punt fix

Transformen $F(\mathbf{z}) = 0$ com $\mathbf{z} = G(\mathbf{z})$, el mètode és

$$\mathbf{z}^{(k+1)} = G(\mathbf{z}^{(k)}) \tag{2}$$

per $\mathbf{z}^{(k)}$ indiquem el vector d'iteració k-èssim.

Algorisme computacional

Donats z^0 i $tol = \eta > 0$ l'algorisme és:

fins que
$$\mathbf{z}^{(k+1)} = G(\mathbf{z}^{(k)})$$
$$||\mathbf{z}^{(k+1)} - \mathbf{z}^{(k)}|| < \eta \quad \text{i} \quad ||F(\mathbf{z}^{(k+1)})|| < \eta$$

La convergència condicionada a $||J_G(\alpha)|| < 1$.

Exercici

Apliqueu el mètode de la iteració simple per resoldre el sistema no lineal

$$
\begin{aligned}
x &= \sin(x+y)\,,\\
y &= \cos(x-y)\,,
\end{aligned}
$$

prop de $(1,1)^t$ amb una precisió tal que

$$||\mathbf{z}^{(k+1)} - \mathbf{z}^{(k)}|| \leq 10^{-6} \quad \text{i} \quad ||F(\mathbf{z}^{(k+1)})|| < 10^{-6}\,.$$

si $\mathbf{z} = (x,y)^t$.

Mètode de la iteració simple

Teorema de convergència

Sigui $\alpha \in \mathbb{R}^n$ la solució de $F(\mathbf{z}) = 0$ i el punt fix de $\mathbf{z} = G(\mathbf{z})$ i $\mathcal{D}_\alpha$ un conjunt tancat i convex que conté la solució α.

Si G és de classe $C^1(\mathcal{D}_\alpha)$ i $\|J_G(\mathbf{z})\| \leq L < 1$ per tot $\mathbf{z} \in \mathcal{D}_\alpha$. Llavors, $\forall \mathbf{z}^0 \in \mathcal{D}_\alpha$, la successió $\mathbf{z}^{k+1} = g(\mathbf{z}^k), k > 0$ compleix:

a) $\mathbf{z}^k \in \mathcal{D}_\alpha \quad k = 0, 1, 2, \ldots$

b) $\lim_{k\to\infty} \mathbf{z}^k = \alpha$.

c) α és la única arrel de $\mathbf{z} = g(\mathbf{z})$ dins de $\mathcal{D}_\alpha$.

d) Es verifca que

$$||\mathbf{z}^{(k+1)} - \alpha|| \leq \frac{L}{1-L}||\mathbf{z}^{(k+1)} - \mathbf{z}^{(k)}||$$

Mètode de Newton

Si F és diferenciable amb contiuïtat, el mètode és

$$\mathbf{z}^{(k+1)} = \mathbf{z}^{(k)} - (J_F(\mathbf{z}^{(k)}))^{-1} \cdot F(\mathbf{z}^{(k)})$$

per $\mathbf{z}^{(k)}$ indiquem el vector d'iteració k-èssim.

Algorisme computacional

Donats x^0 i $\eta > 0$ l'algorisme és:

$$\begin{cases} (J_F(\mathbf{z}^{(k)})) \cdot \mathbf{w}^{(k)} = & -F(\mathbf{z}^{(k)}) \\ \mathbf{z}^{(k+1)} = & \mathbf{z}^{(k)} + \mathbf{w}^{(k)} \end{cases}$$

fins que

$$||\mathbf{z}^{(k+1)} - \mathbf{z}^{(k)}|| < \eta \quad \text{i} \quad ||F(\mathbf{z}^{(k+1)})|| < \eta$$

Variants del mètode de Newton

Es redueix el cost computacional de cada iteració contra l'ordre de convergència.

- **Newton modificat.** Es fixa la matriu $J_F(\mathbf{z}^{(k)})$ per un nombre constant d'iteracions.
- **Mètode de Jacobi.** La matriu $J_F(\mathbf{z}^{(k)})$ es substitueix per una matriu diagonal, amb la diagonal de $J_F(\mathbf{z}^{(k)})$. Matriu $D_F(\mathbf{z}^{(k)})$.
- **Mètode de Gauss-Seidel.** La matriu $J_F(\mathbf{z}^{(k)})$ es substitueix per la matriu triangular inferior de $J_F(\mathbf{z}^{(k)})$. Matriu $G_F(\mathbf{z}^{(k)})$.
- **Mètode de sobrerelaxació o SOR.** Per $\omega = 1/(1+\rho)$
  $\mathbf{z}^{(i+1)} = \mathbf{z}^{(i)} - (\rho \cdot D_F(\mathbf{z}^{(i)}) + G_F(\mathbf{z}^{(i)}))^{-1} \cdot F(\mathbf{z}^{(i)})$

Exercici

Apliqueu el mètode de Newton per resoldre el sistema no lineal

$$\begin{aligned} x &= \sin(x+y)\,, \\ y &= \cos(x-y)\,, \end{aligned}$$

prop de $(1,1)^t$ amb una precisió tal que

$$||\mathbf{z}^{(k+1)} - \mathbf{z}^{(k)}|| \leq 10^{-6} \quad \text{i} \quad ||F(\mathbf{z}^{(k+1)})|| < 10^{-6}\,.$$

si $\mathbf{z} = (x,y)^t$.

Annex

Vector gradient $f : D \to \mathbb{R}$, $D \subseteq \mathbb{R}^n$

$$\nabla f\,(a) = \left(\frac{\partial f}{\partial z_1}(a), \frac{\partial f}{\partial z_2}(a), \ldots, \frac{\partial f}{\partial z_n}(a) \right).$$

Matriu jacobiana $F : D \to \mathbb{R}^n$, $D \subseteq \mathbb{R}^n$

$$J_F(a) = \begin{pmatrix} \frac{\partial f_1}{\partial z_1}(a) & \frac{\partial f_1}{\partial z_2}(a) & \cdots & \frac{\partial f_1}{\partial z_n}(a) \\ \frac{\partial f_2}{\partial z_1}(a) & \frac{\partial f_2}{\partial z_2}(a) & \cdots & \frac{\partial f_2}{\partial z_n}(a) \\ \ldots & \ldots & \ddots & \ldots \\ \frac{\partial f_n}{\partial z_1}(a) & \frac{\partial f_n}{\partial z_2}(a) & \cdots & \frac{\partial f_n}{\partial z_n}(a) \end{pmatrix}.$$

Guia estudi

Guia estudi tema

Llibre Càlcul numèric: teoria i pràctica

- Conceptes associats: Capítol 6, Zeros de funcions no lineals. Des de la pàgina 197 fins a la 209, i de 216 fins a la 221.
- Problemes proposats: 1, 2, 4, 9 i 10.
- Pràctiques resoltes: de la pàgina 238-243.
- Pràctiques proposades: pàgines 244-248.

Llibre Cálculo Científico con MATLAB y Octave

- Conceptes i exercicis resolts: capítol 2, pàgines 41-63, 68-69.
- Problemes i pràctiques proposades: del 2.1 al 2.17

Llibres de consulta online

- Llibre de consulta - Accès UPCommons,
  Càlcul numèric: teoria i pràctica
- Llibre de consulta - Accès UPCommons,
  Cálculo numérico
- Llibre de consulta - Accès Biblioteca,
  Cálculo Científico con MATLAB y Octave by A. Quarteroni, F. Saleri
- Llibre de consulta -C. Moler,
  Cleve Moler - Llibre de text i codis - MathWorks

Computació Numèrica

Tema 5.1 - Derivació Numèrica

M. Àngela Grau Gotés

Departament de Matemàtiques
Universitat Politècnica de Catalunya · BarcelonaTech.

17 d'abril de 2023

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

Introducció

El problema és calcular la derivada d'una funció de la que sols coneixem un nombre finit de valors. Els dos mètodes més usuals de resolució són:

- Derivar el polinomi d'interpolació construït mitjançant algun dels mètodes estudiats en el capítol previ. Les fórmules obtingudes d'aquesta manera reben el nom de fórmules de derivació interpolatòria.
- Calcular directament la derivada utilitzant per a això aproximacions de la funció mitjançant els polinomis de Taylor. Les fórmules obtingudes d'aquesta manera reben el nom de fórmules de diferències finites.

Taula de dades

L'any 2009 (a Berlín) Usain Bolt va situar el record dels 100m en 9.58s. Les dades de la carrera són les següents

r	0	10	20	30	40	50	60	70	80	90	100
t(r)	0	1.85	2.89	3.78	4.64	5.49	6.31	7.11	7.92	8.74	9.58

on la primera fila és la distància recorreguda en metres i la segona el temps emprat en segons

(font: NBC, http://www.universalsports.com/news/article/newsid=385633.html).

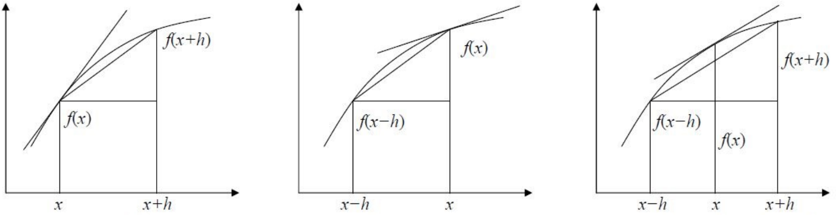
Primeres fórmules

Mètode

La derivació numèrica avalua la derivada d'una funció en un punt a partir de valors numèrics d'aquesta funció, sense necessitat per tant de conèixer l'expressió analítica d'aquesta derivada.

És molt sensible a petites pertorbacions en les dades o la precisió d'aquestes.

Aproximació geomètrica



$$f'(x) \approx \frac{f(x+h) - f(x)}{h}$$

$$f'(x) \approx \frac{f(x) - f(x-h)}{h}$$

$$f'(x) \approx \frac{f(x+h) - f(x-h)}{2h}$$

Fórmules i ordre de l'aproximació

Sigui $f : D \subseteq \mathbb{R} \longrightarrow \mathbb{R}$ una funció real de variable real derivable dues/tres vegades amb continuïtat en un entorn de x, de la fórmula de Taylor s'obté:

$$\frac{f(x) - f(x-h)}{h} - f'(x) = \frac{h}{2} f''(\xi) \to 0 \quad \text{si } (h \to 0),$$

$$\frac{f(x+h) - f(x)}{h} - f'(x) = \frac{h}{2} f''(\xi) \to 0 \quad \text{si } (h \to 0),$$

$$\frac{f(x+h) - f(x-h)}{2h} - f'(x) = \frac{h^2}{6} f^{(3)}(\xi) \to 0 \quad \text{si } (h \to 0).$$

Ordre de les aproximacions (deducció)

Fórmula endavant $f'(x)$

$$f(x+h) = f(x) + f'(x)h + \frac{f''(x)}{2}h^2 + \frac{f'''(x)}{6}h^3 + \cdots$$

Si aïllem de la igualtat $f'(x)$ s'obté la **fórmula endavant** més un reste de primer ordre $\mathcal{O}(h)$

$$f'(x) = \frac{f(x+h) - f(x)}{h} - \frac{f''(x)}{2}h + \cdots \approx \frac{f(x+h) - f(x)}{h}$$

Fórmula enrere $f'(x)$

$$f(x-h) = f(x) - f'(x)h + \frac{f''(x)}{2}h^2 - \frac{f'''(x)}{6}h^3 + \cdots$$

Si aïllem de la igualtat $f'(x)$ s'obté la **fórmula enrere** més un reste de primer ordre $\mathcal{O}(h)$

$$f'(x) = \frac{f(x) - f(x-h)}{h} + \frac{f''(x)}{2}h + \cdots \approx \frac{f(x) - f(x-h)}{h}$$

Fórmula centrada $f'(x)$

$$f(x+h) = f(x) + f'(x)h + \frac{f''(x)}{2}h^2 + \frac{f'''(x)}{6}h^3 + \cdots$$
$$f(x-h) = f(x) - f'(x)h + \frac{f''(x)}{2}h^2 - \frac{f'''(x)}{6}h^3 + \cdots$$

Si restem les dues igualtats i aïllem $f'(x)$ s'obté la **fórmula centrada** més un reste de segon ordre $\mathcal{O}(h^2)$

$$f'(x) = \frac{f(x+h) - f(x-h)}{2h} - \frac{f'''(x)}{6}h^2 + \cdots \approx \frac{f(x+h) - f(x-h)}{2h}$$

Fórmula centrada $f''(x)$

$$f(x+h) = f(x) + f'(x)h + \frac{f''(x)}{2}h^2 + \frac{f'''(x)}{6}h^3 + \cdots$$
$$f(x-h) = f(x) - f'(x)h + \frac{f''(x)}{2}h^2 - \frac{f'''(x)}{6}h^3 + \cdots$$

Si sumem les dues igualtats i aïllem $f''(x)$ s'obté la **fórmula centrada** més un reste de segon ordre $\mathcal{O}(h^2)$

$$f''(x) = \frac{f(x+h) - 2f(x) + f(x-h)}{h^2} - \frac{f^{(4)}(x)}{12}h^2 + \cdots$$
$$\approx \frac{f(x+h) - 2f(x) + f(x-h)}{h^2}$$

Comportament de l'error

Comportament de l'error

L'aparició en moltes fórmules de diferències de quantitats molt properes, amb la corresponent cancel·lació de termes, fa que pensar que pendre passos de derivació h molt petits no millorarà les aproximacions numèriques.

Cal fer atenció als errors d'arrodoniment que apareixen.

Observació

$$\left| f'(x_0) - \frac{\widetilde{f}(x_0+h) - \widetilde{f}(x_0)}{h} \right| \leq \frac{2\epsilon}{h} + \frac{h}{2}K\,, \qquad |f''| < K\,.$$

El pas òptim és el que minimitza l'error total $\left(\Rightarrow h = \left(\frac{4\epsilon}{K}\right)^{1/2} \right)$.

Exemple

Càlcul de la derivada de $\ln(x)$ en $x = 2$

```
f=@(x)log(x);
k=0:14;
h=1./10.^k;
fp(k)=(f(2+h)-f(2))./h;
er = abs(fp-0.5);
taula=[h; fp; er]'
```

La derivació numèrica és un problema mal condicionat.

Extrapolació de Richardson

Extrapolació de Richardson

Ordre $\mathcal{O}(h^2)$

1. Si es coneix l'**ordre de l'error**, el valor exacte es pot aproximar a partir de dues aproximacions successives amb valors $h/2$ i h.

2. Diem $D_2(h) = \dfrac{f(x+h) - f(x-h)}{2h}$ i calculem

   $$f'(x) = D_2(\tfrac{h}{2}) + C\tfrac{h^2}{4} \text{ i } f'(x) = D_2(h) + Ch^2.$$

3. Restem la segona equació de la primera multiplicada per 4:

   $$3f'(x) \approx 4D_2(h) - D_2(2h), \text{ llavors } f'(x) \approx \frac{4D_2(h) - D_2(2h)}{3}$$

4. El resultat és una fórmula d'ordre $\mathcal{O}(h^4)$

   $$f'(x) \approx \frac{-f(x+2h) + 8f(x+h) - 8f(x-h) + f(x-2h)}{12h}$$

Extrapolació de Richardson

Ordre $\mathcal{O}(h^2)$

Si es té una fórmula del tipus

$$F(h) = a_0 + a_1 h^2 + a_2 h^4 + \mathcal{O}(h^6)$$

quan $h \to 0$ el valor a_0 es pot aproximar a partir de dues aproximacions calculades per h i $h/2$, $q > 0$.

$$F(h) = a_0 + a_1 h^2 + a_2 h^4 + \mathcal{O}(h^6)$$
$$F(h/2) = a_0 + a_1 h^2/4 + a_2 h^4/16 + \mathcal{O}(h^6).$$

Multiquem la segona equació per 4 i restem les dues equacions, s'obté

$$a_0 = F(h/2) + \frac{F(h/2) - F(h)}{3} + \mathcal{O}(h^4)$$

Fórmula per calcular a_0 d'ordre $\mathcal{O}(h^4)$.

Taula d'extrapolació

Ordre $\mathcal{O}(h^2)$

$$N_1(h) = F(h)\,, \quad N_{j+1}(h) = N_j\left(\frac{h}{2}\right) + \frac{N_j\left(\frac{h}{2}\right) - N_j(h)}{4^j - 1}\,, \quad j \geq 1\,.$$

$\mathcal{O}(h^2)$	$\mathcal{O}(h^4)$	$\mathcal{O}(h^6)$	$\mathcal{O}(h^8)$
1: $N_1(h)$			
2: $N_1(h/2)$	**3**: $N_2(h)$		
4: $N_1(h/4)$	**5**: $N_2(h/2)$	**6**: $N_3(h)$	
7: $N_1(h/8)$	**8**: $N_2(h/4)$	**9**: $N_3(h/2)$	**10**: $N_4(h)$

Taula: Extrapolació de Richardson de $F(h) = a_0 + a_1 h^2 + a_2 h^4 + \mathcal{O}(h^6)$

Extrapolació de Richardson

Ordre $\mathcal{O}(h^p)$

Si p i r són nombres naturals tals que $r > p > 1$, i per $h \to 0$ es té una fórmula del tipus

$$F(h) = a_0 + a_1 h^p + \mathcal{O}(h^r)$$

el valor a_0 es pot tornar a aproximar a partir de dues aproximacions calculades per h i h/q, $q > 0$.

$$F(h) = a_0 + a_1 h^p + \mathcal{O}(h^r)$$

$$F(h/q) = a_0 + a_1 (h/q)^p + \mathcal{O}(h^r)$$

Sistema d'equacions amb dues incògnites, a_0 i a_1. Multiquem la segona equació per q^p i restem les dues equacions,

$$a_0 = F(h/q) + \frac{F(h/q) - F(h)}{q^p - 1} + \mathcal{O}(h^r)$$

Fórmula per calcular a_0 d'ordre $\mathcal{O}(h^r)$, amb $r > p$.

S'ha augmentat l'ordre sense disminuir h

Taula d'extrapolació

Ordre $\mathcal{O}(h)$

$$N_{j+1}(h) = N_j\left(\frac{h}{q}\right) + \frac{N_j\left(\frac{h}{q}\right) - N_j(h)}{q^j - 1}, \quad j \geq 1.$$

$\mathcal{O}(h)$	$\mathcal{O}(h^2)$	$\mathcal{O}(h^3)$	$\mathcal{O}(h^4)$
1: $N_1(h)$			
2: $N_1(h/q)$	**3**: $N_2(h)$		
4: $N_1(h/q^2)$	**5**: $N_2(h/q)$	**6**: $N_3(h)$	
7: $N_1(h/q^3)$	**8**: $N_2(h/q^2)$	**9**: $N_3(h/q)$	**10**: $N_4(h)$

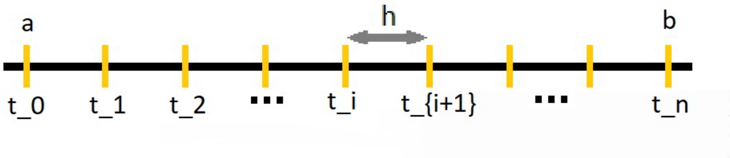
Taula: Extrapolació de Richardson de $F(h) = K_1 + K_2 h + K_3 h^2 + \ldots$

Fórmules i derivades d'ordre superior

Discretització

Generalment es divideix l'interval on es calcula en punts equiespaïats: donat n prenem

$$t_i = a + ih, \text{ per } i = 0, 1, 2, \ldots, n, \text{ amb } h = t_i - t_{i-1} = \frac{b-a}{n}.$$



Els mètodes ens permetran trobar la derivada aproximada en $f_i' \simeq f'(t_i)$ en els punts del domini.

Fórmules derivada primera.

Donat h, siguin $x_k = x_0 + k\,h$, i $f_k = f(x_k)$, per $k \in \mathbb{Z}$

1. $f'(x_0) = \dfrac{f_1 - f_0}{h} + \mathcal{O}(h)$ $\qquad f'(x_0) = \dfrac{f_0 - f_{-1}}{h} + \mathcal{O}(h)$

2. $f'(x_0) = \dfrac{f_1 - f_{-1}}{2h} + \mathcal{O}(h^2)$

3. $f'(x_0) = \dfrac{-3f_0 + 4f_1 - f_2}{2h} + \mathcal{O}(h^2)$

   $f'(x_0) = \dfrac{3f_0 - 4f_{-1} + f_{-2}}{2h} + \mathcal{O}(h^2)$

4. $f'(x_0) = \dfrac{-f_2 + 8f_1 - 8f_{-1} + f_{-2}}{12h} + \mathcal{O}(h^4)$

Fórmules derivada segona

1. $$f''(x_0) = \frac{f_2 - 2f_1 + f_0}{h^2} + \mathcal{O}(h)$$
2. $$f''(x_0) = \frac{f_1 - 2f_0 + f_{-1}}{h^2} + \mathcal{O}(h^2)$$
3. $$f''(x_0) = \frac{-f_3 + 4f_2 - 5f_1 + 2f_0}{h^2} + \mathcal{O}(h^2)$$
4. $$f''(x_0) = \frac{-f_2 + 16f_1 - 30f_0 + 16f_{-1} - f_{-2}}{12h^2} + \mathcal{O}(h^4)$$

Fórmules derivades ordre superior

1. $$f'''(x_0) = \frac{f_3 - 3f_2 + 3f_1 - f_0}{h^3} + \mathcal{O}(h)$$
2. $$f'''(x_0) = \frac{f_2 - 2f_1 + 2f_{-1} - f_{-2}}{8h^3} + \mathcal{O}(h^2)$$
3. $$f^{(4)}(x_0) = \frac{f_2 - 4f_1 + 6f_0 - 4f_{-1} + f_{-2}}{h^4} + \mathcal{O}(h^2)$$

Derivades parcials

Derivades parcials primeres $u(x, y)$

Per una funció de dues variables que només es coneixen

$$u_{i\,j} = u(x_i, y_j), \qquad 1 \leq i \leq n,\ 1 \leq j \leq m\,,$$

valors en la malla equiespaida, $h = x_{i+1} - x_i$ i $k = y_{i+1} - y_i$

Les fórmules centrades per les derivades primeres són:

$$u_x(x_i, y_j) \approx \frac{1}{2h}\left(u_{i+1\,j} - u_{i-1\,j}\right)$$

$$u_y(x_i, y_j) \approx \frac{1}{2k}\left(u_{i\,j+1} - u_{i\,j-1}\right)$$

Derivades parcials segones $u(x, y)$

Le fórmules centrades per les derivades segones són:

$$u_{xx}(x_i, y_j) \approx \frac{1}{h^2}\left(u_{i+1\,j} - 2u_{i\,j} + u_{i-1,j}\right)$$

$$u_{yy}(x_i, y_j) \approx \frac{1}{k^2}\left(u_{i\,j+1} - 2u_{i\,j} + u_{i\,j-1}\right)$$

$$u_{xy}(x_i, y_j) \approx \frac{1}{4hk}\left(u_{i+1\,j+1} - u_{i+1\,j-1} - u_{i-1\,j+1} + u_{i-1\,j-1}\right)$$

$$h = x_{i+1} - x_i,\ {\scriptstyle 1 \le i \le n}, \qquad k = y_{i+1} - y_i,\ {\scriptstyle 1 \le j \le m}\,.$$

Guia estudi subtema

Llibre Càlcul numèric: teoria i pràctica

- Conceptes associats: Capítol 5, Derivació i integració numèrica. Des de la pàgina 160 fins a la 164.

Llibre Cálculo Científico con MATLAB y Octave

- Conceptes i exercicis resolts: capítol 4, pàgines 107 a 109.
- Problemes i pràctiques proposades: del 4.1 al 4.4

Llibres de consulta online

- Llibre de consulta - Accès UPCommons,
  Càlcul numèric: teoria i pràctica
- Llibre de consulta - Accès UPCommons,
  Cálculo numérico
- Llibre de consulta - Accès Biblioteca,
  Cálculo Científico con MATLAB y Octave by A. Quarteroni, F. Saleri
- Métodos Numéricos, J. Douglas Faires & Richard Burden. Ed.Thomson 3era edición. 2004.