Càlcul Numèric

Departament de Matemàtica Aplicada i Anàlisi

5 de febrer de 2001

Índex

Capítol 1

INTERPOLACIÓ POLINOMIAL

1.1 Introducció

Suposem que tenim $n+1$ punts del pla amb coordenades (x_0, y_0), (x_1, y_1),...,(x_n, y_n), on y_i és el valor observat o calculat que associem a x_i. El **problema de la interpolació** consisteix en construir una funció $p(x)$ tal que $p(x_i) = y_i$ per tots els punts x_i. La funció d'interpolació $p(x)$ cal esperar que prengui valors "raonables" pels valors de x que no coincideixin amb els x_i. Així, si les dades de partida tenen un comportament molt regular i suau no acceptarem que la funció d'interpolació $p(x)$ tingui fortes oscil.lacions entre els punts d'interpolació (x_i, y_i).

Si les ordenades y_i provenen de l'avaluació d'una certa funció prou regular i són exactes llevat dels errors d'arrodoniment, cal esperar que el problema tingui una solució satisfactòria. Això mateix també passarà si els punts (x_i, y_i) provenen d'observacions experimentals molt precises. En canvi, si les observacions experimentals estan afectades d'errors grans no té gaire sentit demanar que la funció $p(x)$ prengui exactament els valors y_i en els punts x_i. Aquest últim problema correspon a l'aproximació de dades i es considerarà més endavant.

La resolució del problema de la interpolació té moltes utilitats, tal com es veurà en capítols successius. Una de les més immediates és la de poder disposar de valors de $p(x)$ per valors de x que no estiguin dins de la nostra taula de dades (x_i, y_i). Vegem-ne un exemple:

Hem fabricat un anticongelant amb aigua i glicerina i la barreja té un 45%, en pes, de glicerina. Voldríem saber quin és el seu punt de congelació. Disposem de la següent taula de valors que dóna el punt de congelació, en graus Celsius, (y) de la barreja de glicerina amb aigua com una funció del percentatge en pes, a la barreja, de la glicerina (x).

x	0	10	20	30	40	50	60	70	80	90	100
y	0	–1.6	–4.8	–9.5	–15.4	–21.9	–33.6	–37.8	–19.1	–1.6	17

Tenim, doncs, tabulada una certa funció $y = f(x)$. Ja que no tenim tabulat el valor $x = 45$, ens preguntem com podem obtenir un valor aproximat de $f(45)$.

Hem dit que per resoldre aquest problema calcularem una funció $p(x)$, dins d'una certa família de funcions, tal que coincideixi amb $f(x)$ en els punts (nodes) d'interpolació, és a dir, $p(x_i) = f(x_i)$, $i = 0, 1, \ldots, n$. Quan les funcions, $p(x)$, que considerem siguin polinomis, direm que tenim **un problema d'interpolació polinomial**.

En plantejar-se un problema d'interpolació s'han de fer les següents preguntes:

a) De quin tipus ha de ser la funció p buscada?

   La resposta a aquesta pregunta està lligada a les "sospites" que tinguem respecte al comportament de la funció f. Si expressa algun fenomen periòdic, buscarem p entre les funcions periòdiques. Si sabem que f té asímptotes, usarem una p de tipus racional etc. Aquí només tractarem la interpolació polinomial, és a dir prendrem com funció p un polinomi.

   Els motius d'aquesta elecció són variats. En primer lloc és fàcil operar amb ells i avaluarlos. En segon lloc, es pot demostrar que donada qualsevol funció contínua $f(x)$ en un interval tancat, es pot aproximar arbitràriament per un cert polinomi $p(x)$. Això fa que els polinomis siguin bons candidats per començar a estudiar el problema de la interpolació.

b) Un cop sabem de quin tipus ha de ser la funció p, la segona qüestió és com calcular-la. Òbviament caldrà saber si existeix la funció que busquem i si podem trobar més d'una solució.

c) La tercera qüestió és saber si la funció p trobada ens dóna una bona aproximació de la funció f en els punts on ambdues no coincideixen.

Els resultats de la següent secció donen resposta a algunes de les preguntes que acabem de plantejar.

1.2 Existència, unicitat i error de la interpolació

Teorema: *Donats $n+1$ punts $\{(x_0, y_0), (x_1, y_1), \ldots, (x_n, y_n)\}$, si tots els $x_0, x_1, \ldots, x_n$ són diferents, llavors per qualssevol $y_0, y_1, \ldots, y_n$* **existeix un únic polinomi** *$p_n(x)$, de grau menor o igual a n, tal que*

$$p_n(x_i) = y_i\ , \quad i = 0, 1, 2, \ldots, n\ .$$

$p_n(x)$ rep el nom de **polinomi interpolador** *de f en els punts $x_0, x_1, \ldots, x_n$.*

Demostració: La unicitat és inmediata, doncs si tenim dos polinomis diferents, p_n i q_n de grau menor o igual a n que resolen el problema de la interpolació en els punts x_i aleshores el polinomi $d_n = q_n - p_n$ val 0 en els punts $x_0, \ldots, x_n$ i per tant es pot dividir per $(x - x_0) \cdots (x - x_n)$ que és un polinomi de grau $n + 1$!!.

L'existència del polinomi interpolador es dedueis de la construcció que farem més endavant.

El següent resultat, que no demostrarem, dóna informació sobre l'error de la interpolació sota condicions de diferenciabilitat per la funció $f(x)$. Siguin $y_1, \ldots, y_m$, m nombres reals. Indiquem per $< y_1, \ldots, y_m >$ el mínim interval obert de la recta real que conté tots els punts $y_1, \ldots, y_m$.

Abans, però necessitem introduir el concepte d'error

1.2.1 Error absolut i error relatiu

Sigui x el valor exacte d'una quantitat i $\overline{x}$ el seu valor aproximat. Es defineix **l'error absolut** de x com

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

Aquesta definició permet de mesurar la diferència entre el valor exacte d'una certa quantitat i el seu valor aproximat. Per quantificar la importància de l'error respecte del valor exacte x s'introdueix el concepte **d'error relatiu** de x que es defineix com

$$e_r(x) = \frac{e_a(x)}{x} = \frac{\overline{x} - x}{x}.$$

Aquesta definició mostra que l'error relatiu és una quantitat adimensional i sovint s'expressa en tant per cent.

Cal fer notar que normalment no es coneix el valor exacte de la quantitat x. Per tant, tampoc es coneixen ni l'error absolut ni l'error relatiu, però en canvi es posible donar-ne fites.

Exemple: Sigui $x = \sqrt{2} = 1.414213562...$ i $\overline{x} = 1.414$. Aleshores $e_a(\sqrt{2}) = -0.0002135...$, i

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

Es diu que $\epsilon_a(x)$ és una **fita de l'error absolut** de x si

$$|e_a(x)| \leq \epsilon_a(x),$$

i que $\epsilon_r(x)$ és una **fita de l'error relatiu** de x si

$$|e_r(x)| \leq \epsilon_r(x).$$

Per l'exemple anterior tenim que

$$\epsilon_a(\sqrt{2}) = 0.00022, \quad \epsilon_r(\sqrt{2}) = 0.00017$$

Observem que les fites no són úniques (0.00016 també és una fita de l'error relatiu de $\sqrt{2}$); intentarem treballar i trobar sempre la millor fita possible, és a dir, la més petita. Les següents

notacions s'utilitzen habitualment per realcionar el valor exacte d'una quantitat, el seu valor aproximat i les fites de l'error absolut i relatiu:

$$x = \overline{x} \pm \epsilon_a(x),$$

igualtat que hem d'interpretar com

$$x \in [\overline{x} - \epsilon_a(x), \overline{x} + \epsilon_a(x)].$$

Pel que fà a les fites de l'error relatiu, el fet que

$$-\epsilon_r(x) \leq \frac{x - \overline{x}}{\overline{x}} \leq \epsilon_r(x),$$

s'escriu com

$$x = \overline{x}(1 \pm \epsilon_r(x)).$$

Teorema (fórmula de l'error en la interpolació polinomial)

Sigui f una funció que té les seves $n+1$ derivades contínues ($f \in \mathcal{C}^{n+1}$) en l'interval $< x, x_0, x_1, \ldots, x_n >$. Si p_n és el polinomi de grau $\leq n$ que satisfà

$$p_n(x_i) = f(x_i)\ , \quad i = 0, 1, 2, \ldots, n\ ,$$

llavors

$$f(x) - p_n(x) = \frac{f^{(n+1)}(\xi(x))}{(n+1)!}(x - x_0)(x - x_1)\cdots(x - x_n)\ ,$$

on $\xi(x)$ depèn de x i pertany a $< x, x_0, x_1, \ldots, x_n >$.

Comentaris:

a) Si avaluem l'error que ens dóna la fórmula en els nodes d'interpolació, x_i, aquest val zero, tal com ha de ser.

b) Si la funció $f(x)$ és un polinomi de grau n, com que la seva derivada d'ordre $n+1$ val zero, l'error d'interpolació també val zero. D'aquesta observació és fàcil deduir que si interpolem un polinomi de grau n per un altre del mateix grau, el resultat és el polinomi de partida.

c) Suposem que la funció $f^{(n+1)}(x)$ "no varia gaire". Si prenem un punt x allunyat dels nodes d'interpolació $x_0, x_1, ..., x_n$ (extrapolem) l'error serà, en principi, més gran que si $x \in < x_0, x_1, \ldots, x_n >$ (interpolem).

d) Per tal que la fórmula de l'error sigui d'utilitat des d'un punt de vista pràctic, caldrà que coneguem i sapiguem fitar la derivada d'ordre $(n+1)$ de la funció $f(x)$.

Demostració: Fixem un valor de x verificant $x \neq x_i$, $i = 0, 1, ..., n$, i definim $a(x)$

$$a(x) = \frac{f(x) - p_n(x)}{(x - x_0)...(x - x_n)} = \frac{f(x) - p_n(x)}{w_n(x)},$$

on $w_n(x) = (x - x_0)(x - x_1)...(x - x_n)$.

Construïm una funció auxiliar $F(z)$

$$F(z) = f(z) - p_n(z) - a(x)w_n(z).$$

Aquesta funció també té $n+1$ derivades contínues i es fa zero als $n+2$ punts: x_0, x_1,...,x_n, x. Si apliquem el teorema del valor mitjà $n+1$ vegades resulta

- $F'(z)$ té $n+1$ zeros distints $\xi_i^{(1)}$, $i = 1, 2, ..., n+1$
- $F''(z)$ té n zeros distints $\xi_i^{(2)}$, $i = 1, 2, ..., n+1$
- $F^{(n+1)}(z)$ té un zero (com a mínim) $\xi(x)$.

Ara bé,

$$F^{(n+1)}(z) = f^{(n+1)}(z) - a(x)(n+1)!,$$

per tant

$$F^{(n+1)}(\xi(x)) = f^{(n+1)}(\xi(x)) - a(x)(n+1)! = 0,$$

d'on obtenim que

$$a(x) = \frac{f^{(n+1)}(\xi(x))}{(n+1)!},$$

i, per tant

$$f(x) - p_n(x) = \frac{f^{(n+1)}(\xi(x))}{(n+1)!} w_n(x),$$

que és la igualtat que voliem demostrar.

1.3 Càlcul del polinomi interpolador

Donada una taula de valors o una certa funció tabulada, hi ha diferents mètodes per determinar el polinomi interpolador. En aquest apartat sols estudiarem els dos mètodes següents

1.3.1 Mètode de Lagrange

Considerem com expressió del polinomi interpolador la que dóna la **fórmula de Lagrange**:

$$p_n(x) = \sum_{k=0}^{n} y_k l_k(x) \ ,$$

on

$$l_k(x) = \frac{(x-x_0)\cdots(x-x_{k-1})(x-x_{k+1})\cdots(x-x_n)}{(x_k-x_0)\cdots(x_k-x_{k-1})(x_k-x_{k+1})\cdots(x_k-x_n)} \ , \quad k = 0, 1, \ldots, n \ .$$

Els polinomis $l_k(x)$ són els polinomis de Lagrange i compleixen:

a) el grau de $l_k(x)$ és igual a n,

b)

$$l_k(x_i) = \delta_{ik} = \begin{cases} 1 & \text{si} \quad i = k \\ 0 & \text{si} \quad i \neq k \end{cases}$$

Per tant, es compleix que $p_n(x_i) = y_i$ ($i = 0, 1, \ldots, n$), com desitjàvem. Podem considerar aquesta construcció com una demostració de l'existència del polinomi interpolador. Per provar la unicitat, dins d'aquest contexte, supossem que hem trobat dos polinomis interpoladors diferents, és a dir, $p_n(x)$ i $q_n(x)$ són dos polinomis de grau n tals que $p_n(x_i) = y_i$, $i = 0, 1, ..., n$ i $q_n(x_i) = y_i$, $i = 0, 1, ..., n$. Si prenem el polinomi diferència $p_n(x) - q_n(x)$ tindrà grau més petit o igual que n i $(n+1)$ zeros, x_0, x_1,...,x_n, en contradicció, una vegada més, amb el teorema fonamental de l'àlgebra.

Exemple: Considerem 4 punts de la taula de valors de l'apartat anterior

x	30	40	50	60
y	-9.5	-15.4	-21.9	-33.6

i volem obtenir una aproximació de $f(45)$ usant, està clar, un polinomi de grau 3.

En primer lloc, calculem els polinomis de Lagrange

$$\begin{aligned}
l_0(x) &= \frac{(x-x_1)(x-x_2)(x-x_3)}{(x_0-x_1)(x_0-x_2)(x_0-x_3)} = \frac{-1}{6000}(x-40)(x-50)(x-60) \\
l_1(x) &= \frac{(x-x_0)(x-x_2)(x-x_3)}{(x_1-x_0)(x_1-x_2)(x_1-x_3)} = \frac{1}{2000}(x-30)(x-50)(x-60) \\
l_2(x) &= \frac{(x-x_0)(x-x_1)(x-x_3)}{(x_2-x_0)(x_2-x_1)(x_2-x_3)} = \frac{-1}{2000}(x-30)(x-40)(x-60) \\
l_3(x) &= \frac{(x-x_0)(x-x_1)(x-x_2)}{(x_3-x_0)(x_3-x_1)(x_3-x_2)} = \frac{1}{6000}(x-30)(x-40)(x-50) \ .
\end{aligned}$$

Llavors, el polinomi interpolador és

$$\begin{aligned} p_3(x) &= \sum_{k=0}^{3} y_k l_k(x) = \\ &= \frac{9.5}{6000}(x-40)(x-50)(x-60) - \frac{15.4}{2000}(x-30)(x-50)(x-60) \\ &\quad + \frac{21.9}{2000}(x-30)(x-40)(x-60) - \frac{33.6}{6000}(x-30)(x-40)(x-50). \end{aligned}$$

Un valor aproximat per $f(45)$ ens el dóna $p_3(45) = -18.2875$, per tant podem considerar que $f(45) \simeq -18.3$

1.3.2 Mètode de les diferències dividides de Newton

El mètode de Lagrange dóna una fórmula explícita del polinomi d'interpolació, però té com desavantatge que si afegim més punts d'interpolació no es pot aprofitar tota la feina feta pel seu càlcul, ja que tots els polinomis $l_i(x)$ canvien. Anem a veure un segon procediment de càlcul del polinomi d'interpolació que supera aquesta dificultat.

Expressem el polinomi interpolador en la forma:

$$p_n(x) = c_0 + c_1(x-x_0) + c_2(x-x_0)(x-x_1) + \cdots + c_n(x-x_0)(x-x_1)\cdots(x-x_{n-1}).$$

Aquest mètode ens permet calcular els coeficients c_k ($k = 0, 1, \ldots, n$) mitjançant les diferències dividides, definides de forma recurrent per:

$$\begin{aligned} f[x_i] &= y_i \qquad (i = 0, 1, \ldots, n) \\ f[x_i, x_{i+1}, \ldots, x_{i+j}, x_{i+j+1}] &= \frac{f[x_{i+1}, \ldots, x_{i+j+1}] - f[x_i, \ldots, x_{i+j}]}{x_{i+j+1} - x_i} \\ &\qquad (i = 0, 1, \ldots, n-j-1) \quad (j = 0, 1, \ldots, n-1) \end{aligned}$$

Il·lustrem per $n = 2$, l'esquema de construcció de les diferències dividides:

$$\begin{array}{c|l} x_0 & f[x_0] = y_0 \\ & \qquad\qquad f[x_0, x_1] = \frac{f[x_1]-f[x_0]}{x_1-x_0} \\ x_1 & f[x_1] = y_1 \qquad\qquad f[x_0, x_1, x_2] = \frac{f[x_1,x_2]-f[x_0,x_1]}{x_2-x_0} \\ & \qquad\qquad f[x_1, x_2] = \frac{f[x_2]-f[x_1]}{x_2-x_1} \\ x_2 & f[x_2] = y_2 \end{array}$$

Es pot comprovar que $c_j = f[x_0, x_1, \ldots, x_j]$ i per tant el polinomi interpolador es pot escriure com

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

Exemple: Considerem el mateix exemple que hem utilitzat abans pel mètode de Lagrange, és a dir, considerem la taula

x	30	40	50	60
y	-9.5	-15.4	-21.9	-33.6

i volem obtenir una aproximació de $f(45)$ usant un polinomi de grau 3.

En primer lloc, calculem la taula de les diferències dividides

x_i	$f[x_i]$	$f[x_i, x_j]$	$f[x_i, x_j, x_k]$	$f[x_i, x_j, x_k, x_l]$
30	-9.5			
		$\frac{-15.4-(-9.5)}{40-30} = -0.59$		
40	-15.4		$\frac{-0.65-(-0.59)}{50-30} = -0.003$	
		$\frac{-21.9-(-15.4)}{50-40} = -0.65$		$\frac{-0.026-(-0.003)}{60-30} = -0.00076$
50	-21.9		$\frac{-1.17-(-0.65)}{60-40} = -0.026$	
		$\frac{-33.6-(-21.9)}{60-50} = -1.17$		
60	-33.6			

El polinomi interpolador $p_3(x)$ és

$$\begin{aligned} p_3(x) &= -9.5 - 0.59(x - 30) - 0.003(x - 30)(x - 40) \\ &\quad -0.00076 \ldots (x - 30)(x - 40)(x - 50) \ . \end{aligned}$$

Tenim doncs que $p_3(45) = -18.2875$, per tant podem considerar que $f(45) \simeq -18.3$

Nota: Generalment les diferències dividides d'ordres successius es van fent petites. Si les diferències d'ordre n són molt petites podem interpolar usant sols n dades. Vegem–ne un exemple:

Per la distància (en unitats astronòmiques $\simeq 150 \cdot 10^6$km) de Mars a la Terra dels dies 5 al 9 de Novembre del 1992 a les 0^h de Temps Universal, podem construir la següent taula

Dia	$f[x_i]$	$f[x_i, x_j]$	$f[x_i, x_j, x_k]$	$f[x_i, x_j, x_k, x_l]$
5	0.898013			
		-0.006904		
6	0.891109		0.0000105	
		-0.006883		0.00000033
7	0.884226		0.0000115	
		-0.006860		0.00000033
8	0.877366		0.0000125	
		-0.006835		
9	0.870531			

Com que les diferències d'ordre 3 són zero podem interpolar amb sols 3 dades. En aquesta situació triem el valor central x_2 el més pròxim possible al valor de x pel qual volem fer la interpolació. Així si volem calcular la distància entre els 2 planetes el 7 de Novembre a les $22^h 14^m$ triarem els valors tabulats pels dies 7,8 i 9.

Comentari: Quan aproximem una funció f per un polinomi, l'error en els punts d'interpolació és zero (recordeu que demanem $p_n(x_i) = f(x_i)$). Podem caure en la temptació de pensar que l'aproximació de f millorarà cada cop més quan fem créixer el nombre de punts d'interpolació. Això generalment no és cert. Un contraexemple clàssic el va construir C. Runge (1901). Consisteix en aproximar la funció

$$f(x) = \frac{1}{1 + 25x^2}$$

en l'interval $[-1, 1]$ usant polinomis $p_n(x)$ de grau $\leq n$, interpolant f en els punts equiespaiats $x_i = -1 + \frac{2i}{n}$, $i = 0, 1, \ldots, n$. Quan el grau n del polinomi interpolador tendeix a infinit, $p_n(x)$ divergeix en l'interval $0.726\ldots \leq |x| < 1$, és a dir:

$$\lim_{n\to\infty} (\max_{-1\leq x\leq 1} |f(x) - p_n(x)|) = \infty$$

Això es coneix com a **fenomen de Runge**. A la Figura 1.1 s'il·lustren resultats per diferents valors de n.

Observem que l'error prop de l'origen és petit, però, prop dels punts -1 i 1, l'error augmenta amb n. Cal doncs anar molt en compte al fer interpolacions amb polinomis de grau elevat.

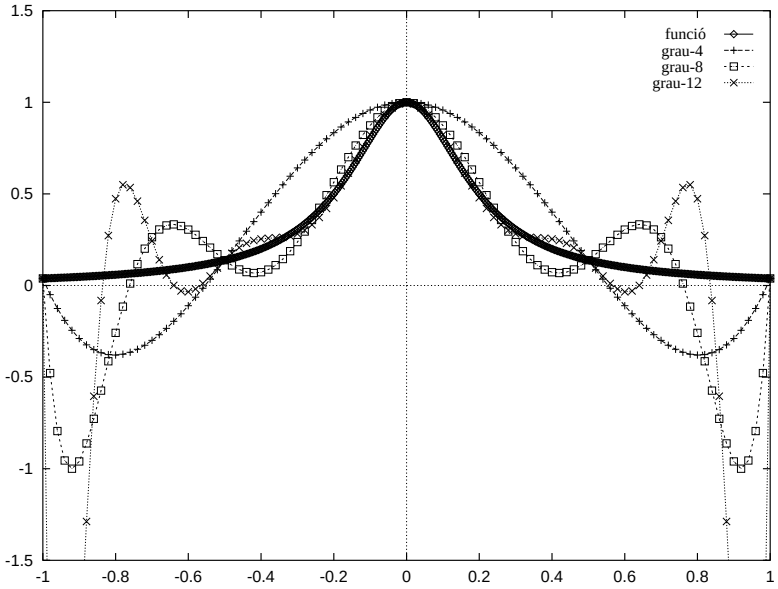

Figura 1.1: El fenomen de Runge. A la figura hi ha representats la funció $f(x) = 1/(1 + 25x^2)$ i els seus polinomis d'interpolació de graus 4, 8 i 12.

1.4 Problemes

1. a) La densitat (ρ en kg/m^3) del sodi en funció de la temperatura (T en $^\circ C$) s'exposa en la següent taula:

i	T_i	ρ_i
0	94	929
1	205	902
2	371	860

Calculeu el polinomi que interpola $\rho(T)$ en aquests tres punts usant la fórmula de Lagrange.

b) Busqueu la densitat per a $T = 251^\circ C$ utilitzant el polinomi interpolador trobat en l'apartat anterior.

Resposta: a) $\frac{929}{30747}(T-205)(T-371) - \frac{902}{18426}(T-94)(T-371) + \frac{860}{45982}(T-94)(T-205)$ b) 890.566 $^\circ C$.

2. Trobeu el polinomi de grau tres que interpola els valors de la següent taula:

x	0	1	2	4
$f(x)$	1	5	10	24

Useu la fórmula de Lagrange i el mètode de les diferències dividides de Newton. *Resposta:* $x^3/24 + 3x^2/8 + 43x/12 + 1$.

3. Una taula de la funció $f(x) = \log_{10}(x)$ és:

i	x_i	$f(x_i)$
0	1	0
1	2	0.30103
2	3	0.47712
3	4	0.60206

Suposem que la funció s'aproxima per un polinomi interpolador que utilitza totes aquestes dades. Estimeu els errors en els punts $x = 1.5, 2.5, 3.5$. *Resposta:* 0.10179, 0.06107, 0.10179.

4. Considereu la funció $f(x) = 1/x$.

a) Trobeu els polinomis de Taylor al voltant de $x_0 = 1$ de graus 2, 3 i 4.

b) Calculeu el polinomi interpolador a $f(x)$ en els nodes $x_0 = 2$, $x_1 = 2.5$, $x_2 = 4$.

c) Avalueu la funció i els polinomis obtinguts en els apartats anteriors en els punts 0.5, 1, 2, 2.25, 2.75, 3, 3.5, 4 i 5. Compareu els resultats.

d) Considereu els polinomis i els punts de l'apartat anterior. Quines són les fites teòriques de l'error? (Indicació: Apliqueu la fórmula de l'error en la interpolació i la resta de Lagrange pel polinomi de Taylor).

Resposta: a) x^2-3x+3, $-x^3+4x^2-6x+4$, $x^4-5x^3+10x^2-10x+5$ b) $0.05x^2-0.425x+1.15$.

5. Sigui $f(x) = 3xe^x - e^{2x}$. Aproximeu $f(1.03)$ mitjançant el polinomi d'interpolació de grau 2 en els nodes $x_0 = 1$, $x_1 = 1.5$, $x_2 = 1.7$. Compareu l'error exacte amb la fita de l'error obtinguda a partir de la fórmula de l'error. *Resposta:* Error = 0.102322, Fita = 0.256.

1.5 Qüestions

6. De la funció $y(x)$ en coneixem els seus valors i els de la primera derivada en els punts 0 i 1, i estan donats en la següent taula

x_k	y_k	y'_k
0	1	0
1	2	1

El polinomi que interpola aquestes dades és:

a) $1 + 2x^2 - x^3$, b) $x + 1$,
c) $x^3 + 2x^2 + 1$, d) Cap de les anteriors.

7. Considerem la següent taula de valors

x_i	-5	1	4
$f(x_i)$	28	4	28

Si $p(x)$ és el polinomi de major grau que interpola els valors de l'anterior taula, què val $p(-2)$?

a) -4, b) $4/3$, c) 4, d) Cap de les anteriors.

8. El polinomi $p(x)$ que interpola la següent taula de valors

x	0.1	0.2	0.3	0.4	0.5
y	1.302	1.616	1.954	2.328	2.750

a) és de grau 1, b) és de grau 2, c) és de grau 3, d) és de grau 5.

9. Volem calcular $f(n) = \sum_{k=1}^{n} k^3$, $\forall n \in \mathbb{N}$ però sols recordem que $f(n)$ és un polinomi de grau 4. En quants punts hem de conèixer la funció per poder determinar el polinomi?

a) en 5 punts, b) en 4 punts, c) en 3 punts, d) Cap de les anteriors.

10. Hem calculat $f(2.5)$, on $f(x) = \ln(x)$, per interpolació cúbica usant com a nodes 1, 2, 3, 4. Una fita per l'error és

a) 0.140625, b) 0.0032958, c) 0.023437, d) Cap de les anteriors.

11. Sigui $p(x)$ el polinomi que interpola la següent taula de valors

x	1.1	1.7	3.0
y	10.6	15.2	20.3

Quant val $p(2.3)$?

a) 18.38 b) 18.40 c) 19.25 d) Cap dels anteriors.

12. Quin és el polinomi $P(x)$, que compleix $P(0) = 3, P(1) = 2, P'(0) = 1, P'(1) = -2$?

a) $3x^3 - 3x^2 - x + 3$ b) $x^5 - 3x - 27$
c) $x^3 - 3x^2 + x + 3$ d) $3x^3 + 3x^2 + x + 3$

13. D'una taula de diferències dividides de Newton sabem que $x_0 = 1, x_1 = 2, x_2 = 3, f(x_2) = 8, f[x_1, x_2] = 4, f[x_0, x_1, x_2] = 1$. Quant val el polinomi interpolador?

a) $x^2 + 2x + 1$ b) $x^2 + x - 2$
c) $x^2 - 2x - 1$ d) $x^2 - x + 2$

14. Es donen els punts $(7, 3)$, $(8, 1)$, $(9, 1)$ i $(10, 9)$ corresponents a una certa funció $y = f(x)$. Aproximeu el valor de y per $x = 9.5$, emprant un polinomi que interpoli la funció en aquests punts.

a) -1.002, b) 4.201, c) 2.783, d) 3.625.

15. Donats $x_0, x_1, x_2 \in \mathbf{R}$ punts diferents dos a dos i $f(x) = x^3 + ax^2 + bx + c$ amb $a, b, c \in \mathbf{R}$. Calculem $p_2(x)$, polinomi interpolador de grau 2 en els punts $(x_0, f(x_0))$, $(x_1, f(x_1))$, $(x_2, f(x_2))$. L'error en la interpolació és:

a) $\frac{f^{(2)}(x)}{2!}(x - x_0)(x - x_1)(x - x_2)$, b) $\frac{p_2^{(3)}(x)}{3!}(x - x_0)(x - x_1)(x - x_2)$,
c) $(x - x_0)(x - x_1)(x - x_2)$, d) Cap de les anteriors.

16. Calculeu, per interpolació, $f(3)$ emprant la taula

x_k	1	2	4
$f(x_k)$	0	2	12

a) 6, b) 2, c) -5, d) 0.

17. L'aproximació lineal al voltant de l'origen de la funció $f(x) = ln(1 + x)$ és

a) $f(x) \simeq 1$, b) $f(x) \simeq x$,
c) $f(x) \simeq 1 + x$, d) $f(x) \simeq 1 - x$

18. Determineu el polinomi interpolador de $f(0) = 1$, $f'(0) = 0$, $f(1) = 2$ i $f'(1) = 1$:

a) $p(x) = 1 + x^2 + x^2(x - 1)$, b) $p(x) = 1 + x^2 - x^2(x - 1)$,
c) $p(x) = 1 + x(x - 1) - x(x - 1)^2$, d) $p(x) = 1 + x(x - 1) + x(x - 1)^2$.

19. Sigui $p(x)$ el polinomi que interpola la següent taula de valors

x	-1	1	2
y	0	0	3

Quant val $p(0)$?

a) 2, b) -1, c) -1.2, d) 0.

20. Donats n punts, $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ on $x_1, x_2, \ldots, x_n$ són diferents dos a dos, llavors quants polinomis, $p_n(x)$, de grau n existeixen complint que $y_1 = p_n(x_1), \ldots, y_n = p_n(x_n)$?

a) un, b) infinits, c) cap, d) dos.

21. Quant val el coeficient de $(x-1)(x-1.01)$ del polinomi que interpola la següent taula de valors

x	1.00	1.01	1.02	1.03
$\sqrt{x}$	1.0000	1.0050	1.0100	1.0149

a) 0.5, b) 1, c) 0, d) -16.66.

22. Quin és el polinomi $p(x)$ de grau n que en $x_0 = 0$ coincideix amb e^x, junt amb les seves n primeres derivades

a) $1 + \frac{1}{2}x^2 + \ldots + \frac{1}{(2n)!}x^{2n}$, b) $x + \frac{1}{6}x^3 + \ldots + \frac{1}{(2n+1)!}x^{(2n+1)}$,
c) $1 - x + \frac{1}{2}x^2 - \frac{1}{6}x^3 + \ldots + \frac{(-1)^n}{n!}x^n$, d) $1 + x + \frac{1}{2}x^2 + \frac{1}{6}x^3 + \ldots + \frac{1}{n!}x^n$.

23. Sabem que $s(n) = \sum_{k=1}^{n} k^2$ és un polinomi de grau 3. Quant val $\sum_{k=1}^{90} k^2$?

a) 255240, b) 247065, c) 238965, d) 263340.

24. Interpolem una funció $f(x)$ a un interval $[a, b]$ en abscisses equidistants mitjançant un polinomi de grau m, $P_m(x)$. Si fem tendir m a infinit, aleshores:

a) $P_m(x)$ sempre tendeix a $f(x)$,
b) $P_m(x)$ mai tendeix a $f(x)$,
c) La convergència o no, depèn només de la funció $f(x)$,
d) La convergència o no, depèn de la funció $f(x)$ i del valor de x.

25. Quant val $P(3)$ si $P(x)$ és el polinomi, de grau com a màxim 3, tal que $P(1) = 1$, $P'(1) = 2$, $P''(1) = 4$, $P(2) = 5$:

a) 0, b) 112, c) 13, d) Cap dels anteriors.

26. Aproximem $\sin x$ per x. Fins a quina distància de l'origen l'error és més petit que $(1/2) \cdot 10^{-4}$ utilitzant aquesta aproximació ?

a) 0.014, b) 0.0001, c) 0.067, d) Cap dels anteriors.

27. Sigui $p(x)$ el polinomi que interpola la següent taula de valors:

x	-5	0	1	4
y	13	-1	2	-3

Quant val $p(5)$?

a) $\frac{380}{27}$, b) $-\frac{380}{27}$, c) $-\frac{760}{27}$, d) $\frac{380}{29}$.

Capítol 2

DERIVACIÓ I INTEGRACIÓ NUMÈRICA

2.1 Derivació numèrica

Hi ha moltes maneres d'obtenir fórmules que aproximin el valor de la derivada d'una funció en un punt. La que aquí utilitzarem es basa en l'ús de la fórmula de Taylor. Així, si $f(x)$ és una funció $n+1$ vegades derivable en un entorn obert al voltant d'un punt x_0, per x dins d'aquest entorn es té

$$\begin{aligned} f(x) &= f(x_0) + (x-x_0)f'(x_0) + \frac{(x-x_0)^2}{2}f''(x_0) + \frac{(x-x_0)^3}{6}f^{(iii)}(x_0) + \cdots + \\ &\quad + \frac{(x-x_0)^n}{n!}f^{(n)}(x_0) + \frac{(x-x_0)^{n+1}}{(n+1)!}f^{(n+1)}(\xi(x)), \quad \text{on } \xi(x) \in < x, x_0 > \end{aligned}$$

Suposem coneguts els valors de $f(x)$ en els punts x_0 i x_1. Si $h = x_1 - x_0$, podem escriure

$$f(x_1) = f(x_0 + h) = f(x_0) + hf'(x_0) + \frac{1}{2}h^2 f''(\xi), \quad \xi \in (x_0, x_1),$$

d'on podem obtenir la següent aproximació per $f'(x_0)$

$$f'(x_0) \simeq \frac{f(x_1) - f(x_0)}{h} = \frac{f(x_0 + h) - f(x_0)}{h}\ .$$

En aquesta expressió estem negligint $(1/2)h^2 f''(\xi)$, amb $\xi \in (x_0, x_1)$. Si h és petit, en comparació amb els valors de les derivades de $f(x)$ en un entorn del punt x_0, podem obtenir una aproximació millor de $f'(x_0)$ d'aquesta altra manera. Sigui $x_{-1} = x_0 - h$, aleshores

$$f(x_1) = f(x_0 + h) = f(x_0) + hf'(x_0) + \frac{1}{2}h^2 f''(x_0) + \frac{1}{6}h^3 f^{(iii)}(\xi), \quad \xi \in (x_0, x_1),$$

$$f(x_{-1}) = f(x_0 - h) = f(x_0) - hf'(x_0) + \frac{1}{2}h^2 f''(x_0) - \frac{1}{6}h^3 f^{(iii)}(\eta), \quad \eta \in (x_{-1}, x_0),$$

d'on

$$f'(x_0) \simeq \frac{f(x_1) - f(x_{-1})}{2h} = \frac{f(x_0+h) - f(x_0-h)}{2h} \ ,$$

essent l'error d'aquesta aproximació "proporcional" a h^2.

No sempre els punts x_i on coneixem els valors de la funció $f(x)$ han d'estar equiespaiats. Si suposem coneguts els valors d'aquesta funció en els punts x_0, x_1, x_2, x_3, i anomenem $h_1 = x_1 - x_0$, $h_2 = x_2 - x_0$, $h_3 = x_3 - x_0$, desenvolupant per Taylor al voltant del punt x_0 obtenim

$$f(x_1) = f(x_0+h_1) = f(x_0) + h_1 f'(x_0) + \frac{1}{2}h_1^2 f''(x_0) + \frac{1}{6}h_1^3 f^{(iii)}(x_0) + \dots$$

$$f(x_2) = f(x_0+h_2) = f(x_0) + h_2 f'(x_0) + \frac{1}{2}h_2^2 f''(x_0) + \frac{1}{6}h_2^3 f^{(iii)}(x_0) + \dots$$

$$f(x_3) = f(x_0+h_3) = f(x_0) + h_3 f'(x_0) + \frac{1}{2}h_3^2 f''(x_0) + \frac{1}{6}h_3^3 f^{(iii)}(x_0) + \dots$$

D'aquestes tres equacions és fàcil arribar a

$$f(x_2) - f(x_1) = (h_2 - h_1) f'(x_0) + \frac{1}{2}(h_2^2 - h_1^2) f''(x_0) + \frac{1}{6}(h_2^3 - h_1^3) f^{(iii)}(x_0) + \dots$$

$$f(x_3) - f(x_2) = (h_3 - h_2) f'(x_0) + \frac{1}{2}(h_3^2 - h_2^2) f''(x_0) + \frac{1}{6}(h_3^3 - h_2^3) f^{(iii)}(x_0) + \dots$$

Si multipliquem la primera equació per $(h_3^2 - h_2^2)$, la segona per $(h_2^2 - h_1^2)$, restem i aïllem $f'(x_0)$, en resulta

$$\begin{aligned} f'(x_0) \quad = \quad & \frac{2x_0 - (x_2 + x_3)}{(x_1 - x_2)(x_1 - x_3)} f(x_1) + \frac{2x_0 - (x_1 + x_3)}{(x_2 - x_1)(x_2 - x_3)} f(x_2) + \\ & \frac{2x_0 - (x_1 + x_2)}{(x_3 - x_1)(x_3 - x_2)} f(x_3) + C(h^2) f^{(iii)}(x_0) + \dots \end{aligned}$$

D'aquesta expressió voldríem fer notar dues cosses. En primer lloc amb $C(h^2)$ volem indicar un terme, que es pot calcular amb una mica de paciència, on sols apareixen productes $h_i * h_j$ $(i, j = 1, 2, 3)$. En segon lloc, és fàcil adonar-se que una altra manera d'arribar a aquesta mateixa igualtat consistiria en derivar el polinomi d'interpolació que passa pels punts $(x_1, f(x_1))$, $(x_2, f(x_2))$, $(x_3, f(x_3))$, i avaluar el polinomi derivat en el punt x_0. Això dóna un mètode general per l'avaluació de les derivades d'una funció.

Procedint d'una manera semblant al que hem fet per avaluar aproximacions de la derivada primera, es poden obtenir aproximacions per les derivades d'ordre superior. Per exemple

$$f''(x_0) \simeq \frac{f(x_{-1}) - 2f(x_0) + f(x_1)}{h^2} \text{ on } x_{-1} = x_0 - h \text{ i } x_1 = x_0 + h \ .$$

En aquest cas es pot demostrar que l'error de la fórmula de derivació val

$$-\frac{f^{(4)}(\xi)}{12} h^2, \quad \xi \in (x_{-1}, x_1) \ .$$

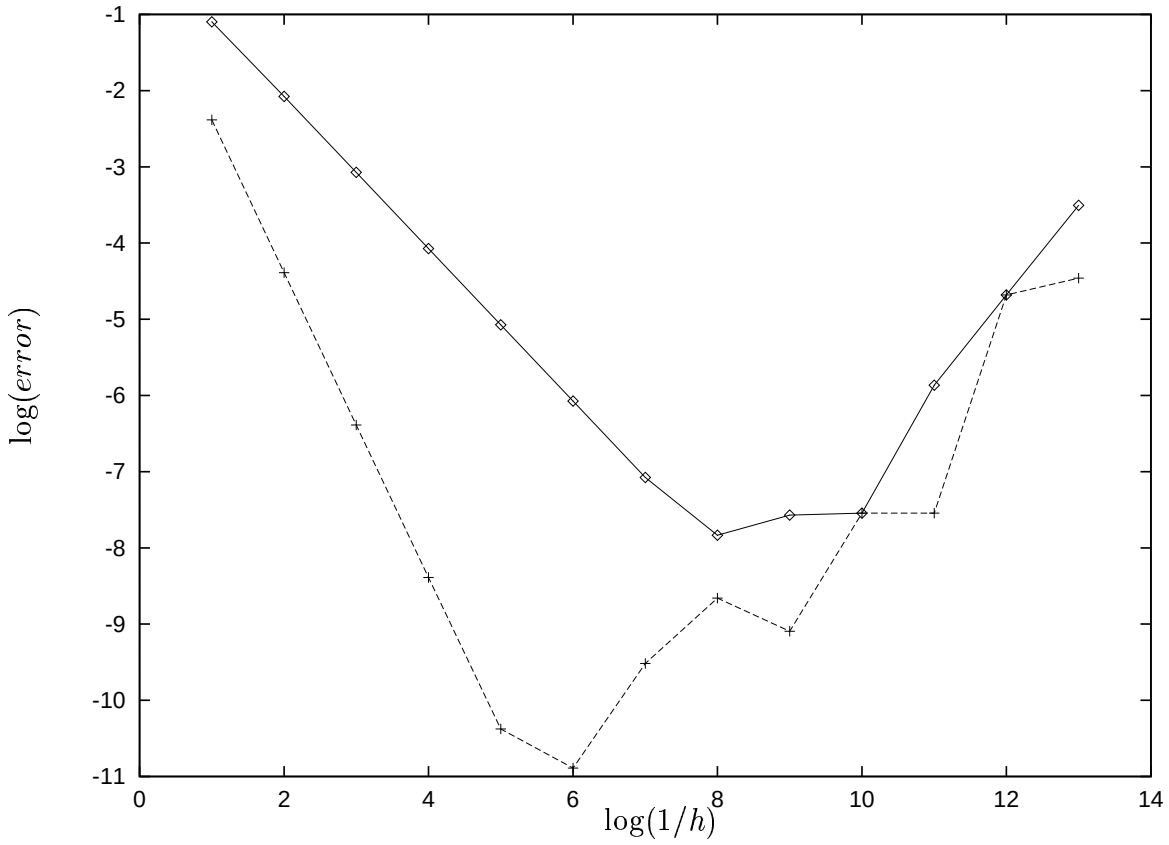

Figura 2.1: Comportament de l'error per les fórmules de derivació numèrica A (línea contínua) i B (línea puntejada). Els errors d'arrodoniment invaliden els resultats abans d'aconseguir un precissió menor que 10^{-12} en el millor dels casos (fórmula B). Per la representació gràfica s'han utilitzat escales logarítmiques.

Nota: Hom pot pensar que les millors aproximacions numèriques de les derivades s'obtenen prenent passos de derivació h molt petits. L'aparició en moltes de les fórmules de diferències de quantitats molt properes, amb la corresponent cancel·lació de termes, fa que això no sigui en general cert. La Figura 2.1 mostra el comportament de l'error per dues fórmules de derivació

$$\text{A:}\quad f'(x_0) \simeq \frac{f(x_0+h)-f(x_0)}{h}, \qquad \text{B:}\quad f'(x_0) \simeq \frac{f(x_0+h)-f(x_0-h)}{2h}.$$

en funció del pas de discretització h, per $f(x) = \sin x^2$ i $x_0 = 0.5$.

Com es veu a la figura, per totes dues fórmules de derivació hi ha un pas óptim a partir del qual, si prenem valors de h més petits, els errors comencen a creixer. La determinació d'aquest pas óptim s'ha de fer cercant el valor de h que minimitza la suma de l'error de truncament propi de la fórmula i l'error d'arrodoniment dels càlculs. Per les dues fórmules utilitzades, la suma

d'aquests errors és de la forma

$$E_A(h) = \frac{h}{2}M_1 + \frac{2\epsilon}{h}, \quad E_B(h) = \frac{h^2}{6}M_2 + \frac{\epsilon}{h},$$

on M_1 i M_2 són fites per $\mid f''(x) \mid$ i $\mid f'''(x) \mid$ respectivament i ϵ és una fita per l'error d'arrodoniment. Si derivem i igualem a zero les funcions $E_A(h)$ i $E_B(h)$ obtenim que els punts on s'atansen els mínims són

$$\text{A: } h = \left(\frac{4\epsilon}{M_1}\right)^{1/2}, \quad \text{B: } h = \left(\frac{3\epsilon}{M_2}\right)^{1/3}.$$

2.2 Integració numèrica

L'objectiu de la integració numèrica és el següent: Donada una funció $f(x)$ definida sobre un interval $[a, b]$ volem calcular una aproximació de

$$I(f) = \int_a^b f(x)dx,$$

quan això té sentit.

Les aproximacions de $I(f)$ que proporciona la integració, o quadratura, numèrica són útils en situacions diverses. Per exemple, quan la funció $f(x)$ no té primitiva, i per tant no es pot aplicar la regla de Barrow; quan no coneixem l'expressió analítica de la funció que volem integrar; si els valors que pren la funció primitiva de $f(x)$ sols es poden calcular aproximadament amb un esforç de càlcul relativament gran, etc. Un exemple clar d'això últim és

$$\begin{aligned}\int_0^x \frac{dt}{1+t^4} &= \frac{1}{4\sqrt{2}} \log \frac{x^2+\sqrt{2}x+1}{x^2-\sqrt{2}x+1} + \\ &\quad + \frac{1}{2\sqrt{2}} \left(\arctan \frac{x}{\sqrt{2}+x} + \arctan \frac{x}{\sqrt{2}-x} \right).\end{aligned}$$

La idea general de la integració numèrica consisteix en aproximar la funció $f(x)$ per un polinomi d'interpolació i el valor de la integral $I(f)$ pel de la integral del polinomi d'interpolació.

2.2.1 Mètodes del rectangle i del trapezi

Dins de l'interval $[a, b]$ prenem $n + 1$ punts que suposarem ordenats de la següent manera

$$a = x_1 < x_2 < \ldots < x_{n+1} = b,$$

i anomenarem $h_i = x_{i+1} - x_i$ per $i = 1, 2, \ldots n$. Podem escriure

$$I(f) = \sum_{i=1}^{n} I_i(f) = \sum_{i=1}^{n} \int_{x_i}^{x_{i+1}} f(x)dx.$$

Deduirem, en primer lloc, les *fórmules d'integració simple* per aproximar les integrals $I_i(f)$, i després, sumant, obtindrem les anomenades *fórmules compostes* que donaran una aproximació de $I(f)$.

La *fórmula del rectangle simple* utilitza com aproximació de $f(x)$ en l'interval $[x_i, x_{i+1}]$ un polinomi de grau zero, funció constant, definit pel valor que pren la funció en el punt mitjà de l'interval. D'ara en endavant el punt mig de cada subinterval, $[x_i, x_{i+1}]$, el denotarem per

$$y_i = \frac{x_i + x_{i+1}}{2}, \qquad i = 1, \ldots, n.$$

D'aquesta manera podem escriure l'aproximació de la integral I_i com

$$I_i(f) \simeq h_i f(y_i),$$

i d'aquí obtenim la *fórmula del rectangle composta*:

$$R(f) = \sum_{i=1}^{n} h_i f(y_i).$$

Les *fórmules del trapezi* utilitzen els punts extrems de cada subinterval, $(x_i, f(x_i))$, $(x_{i+1}, f(x_{i+1}))$, per calcular el polinomi d'interpolació de $f(x)$ de grau 1 (recta). Integrant aquests polinomis resulta

$$I_i(f) \simeq h_i \frac{f(x_i) + f(x_{i+1})}{2},$$

$$T(f) = \sum_{i=1}^{n} h_i \frac{f(x_i) + f(x_{i+1})}{2},$$

Es pot demostrar que si la funció $f(x)$ és contínua, ambdues fórmules compostes són convergents cap al valor de la integral si la llargada dels subintervals decreix. És a dir, si $h = \max_{1 \le i \le n} h_i$, aleshores

$$\lim_{h \to 0} R(f) = I(f),$$

$$\lim_{h \to 0} T(f) = I(f).$$

Per saber quina de les dues fórmules d'integració numèrica convergeix més ràpidament, suposarem que $f(x)$ té derivades fins ordre 5 contínues en tot $[a, b]$, i que aquestes derivades no prenen valors gaire grans. Fent ús de la fórmula de Taylor podem escriure

$$\begin{aligned} f(x) &= f(y_i) + (x - y_i)f'(y_i) + \frac{1}{2}(x - y_i)^2 f''(y_i) + \frac{1}{6}(x - y_i)^3 f^{(iii)}(y_i) + \\ &\quad + \frac{1}{24}(x - y_i)^4 f^{(iv)}(y_i) + \ldots \end{aligned}$$

Fàcilment es veu que

$$\int_{x_i}^{x_{i+1}} (x-y_i)^p dx = \begin{cases} h_i & \text{si} \quad p=0, \\ 0 & \text{si} \quad p=1, \\ h_i^3/12 & \text{si} \quad p=2, \\ 0 & \text{si} \quad p=3, \\ h_i^5/80 & \text{si} \quad p=4. \end{cases}$$

per tant, integrant terme a terme l'anterior desenvolupament, resulta

$$\int_{x_i}^{x_{i+1}} f(x)dx = h_i f(y_i) + \frac{1}{24}h_i^3 f''(y_i) + \frac{1}{1920}h_i^5 f^{(iv)}(y_i) + \dots$$

Això mostra que si h_i és petit, l'error de la fórmula del rectangle en cadascun dels subintervals és

$$(1/24)h_i^3 f''(y_i),$$

més altres termes més petits (sempre i quan, com ja hem dit, els valors de les derivades de $f(x)$ no es facin gaire grans).

Si tornem a utilitzar la fórmula de Taylor per $x = x_i$ i $x = x_{i+1}$, obtenim

$$f(x_i) = f(y_i) - \frac{1}{2}h_i f'(y_i) + \frac{1}{8}h_i^2 f''(y_i) - \frac{1}{48}h_i^3 f^{(iii)}(y_i) + \frac{1}{384}h_i^4 f^{(iv)}(y_i) + \dots$$

$$f(x_{i+1}) = f(y_i) + \frac{1}{2}h_i f'(y_i) + \frac{1}{8}h_i^2 f''(y_i) + \frac{1}{48}h_i^3 f^{(iii)}(y_i) + \frac{1}{384}h_i^4 f^{(iv)}(y_i) + \dots$$

i per tant

$$\frac{f(x_i)+f(x_{i+1})}{2} = f(y_i) + \frac{1}{8}h_i^2 f''(y_i) + \frac{1}{384}h_i^4 f^{(iv)}(y_i) + \dots$$

Si d'aquesta equació aïllem $f(y_i)$ i substituïm el seu valor en el desenvolupament de la integral, resulta

$$\int_{x_i}^{x_{i+1}} f(x)dx = h_i \frac{f(x_i)+f(x_{i+1})}{2} - \frac{1}{12}h_i^3 f''(y_i) - \frac{1}{480}h_i^5 f^{(iv)}(y_i) + \dots$$

Això mostra ara que si h_i és petit, l'error de la fórmula del trapezi simple és $-(1/12)h_i^3 f''(y_i)$ més altres termes més petits (igual que abans, sempre i quan els valors de les derivades de $f(x)$ no es facin gaire grans).

L'error total per qualsevol de les dues fórmules és la suma de l'error en cada subinterval. Així, si

$$E = \frac{1}{24}\sum_{i=1}^{n} h_i^3 f''(y_i), \qquad F = \frac{1}{1920}\sum_{i=1}^{n} h_i^5 f^{(iv)}(y_i),$$

tenim

$$\begin{aligned} I(f) &= R(f) + E + F + \dots \\ &= T(f) - 2E - 4F + \dots \end{aligned}$$

Si $f^{(iv)}$ és petita i els h_i son prou petits, aleshores $F \ll E$ i el terme dominant de l'error en les dues fórmules és E i $2E$ respectivament.

De tot el que acabem de veure en podem treure les següents conclusions

a) Per moltes funcions $f(x)$, l'error de la fórmula del trapezi és aproximadament dues vegades més gran que el de la fórmula del rectangle. Això sembla contradictori amb el que intuïtivament podríem pensar. Recordem que per les fórmules del trapezi hem aproximat la funció a integrar per un polinomi de grau 1, mentre que, per les del rectangle per un de grau 0.

b) Si prenem els nodes d'integració $x_1, x_2, ..., x_{n+1}$ equiespaiats, és a dir $h_1 = h_2 = ... = h_n = h$, podem escriure que

$$E = \frac{1}{24}h^3 \sum_{i=1}^{n} f''(y_i), \qquad F = \frac{1}{1920}h^5 \sum_{i=1}^{n} f^{(iv)}(y_i).$$

Si la funció $f^{(iv)}(x)$ és contínua, es pot demostrar que existeixen dos punts ξ, η, a l'interval $< y_1, y_2, ..., y_n >$, tals que

$$\sum_{i=1}^{n} f''(y_i) = nf''(\xi), \qquad \sum_{i=1}^{n} f^{(iv)}(y_i) = nf^{(iv)}(\eta),$$

i, per tant, si tenim en compte que $nh = b - a$

$$E = \frac{1}{24}f''(\xi)(b-a)h^2, \qquad F = \frac{1}{1920}f^{(iv)}(\eta)(b-a)h^4.$$

Si anomenem $c_2 = (1/12)f''(\xi)(b-a)$, $c_4 = (1/960)f^{(iv)}(\eta)(b-a)$, de les quals cal fer notar que són independents de h, en resulta que

$$\begin{aligned} I(f) &= R(f) + \frac{c_2}{2}h^2 + \frac{c_4}{4}h^4 + \dots \\ &= T(f) - c_2h^2 - c_4h^4 + \dots \end{aligned}$$

c) Si dupliquem el nombre de subintervals, dividint per dos el pas d'integració h, cadascun dels termes dominants de l'error de les fórmules d'integració simple disminueix en un factor 1/8, i com que hem multiplicat per 2 el nombre total de subintervals, l'error total es redueix per un factor aproximadament igual a 1/4. Això no és del tot exacte ja que la funció $f''(x)$ no és constant i a la vegada estem negligint els termes d'ordre superior.

d) La tècnica d'anar doblant el nombre de subintervals i estimar l'error es pot programar de manera que doni uns subintervals tals que aproximin el valor de la integral amb un error cada vegada més petit. Si avaluem, amb la fórmula dels trapezis, el valor de la integral d'una funció $f(x)$ amb passos h i $h/2$, podem escriure

$$T_0(h) = I(f) + c_2h^2 + c_4h^4 + ...$$

$$T_0\left(\frac{h}{2}\right) = I(f) + \frac{c_2}{4}h^2 + \frac{c_4}{8}h^4 + ...$$

Si multipliquem la primera equació per $-1/3$, la segona per $4/3$ i les sumem:

$$T_1(h) = \frac{4}{3}T_0\left(\frac{h}{2}\right) - \frac{1}{3}T_0(h) = I(f) - \frac{1}{6}c_4h^4 + ...$$

amb el que hem aconseguit una nova fórmula d'integració que té un terme dominant de l'error de l'ordre de h^4. Iterant aquest procès podem obtenir fórmules de quadratura d'ordre elevat, conegudes com fórmules d'integració Romberg. Més concretament, podem construir l'esquema triangular

$$\begin{array}{lll} T_0(h) & & \\ T_0(h/2) & T_1(h) & \\ T_0(h/2^2) & T_1(h/2) & T_2(h) \\ \vdots & \vdots & \vdots \end{array}$$

on

$$T_m\left(\frac{h}{2^k}\right) = \frac{4^m}{4^m-1}T_{m-1}\left(\frac{h}{2^{k+1}}\right) - \frac{1}{4^m-1}T_{m-1}\left(\frac{h}{2^k}\right)$$

amb $k = 0, 1, \ldots$ i $m = 1, 2, \ldots, k$.

2.2.2 Mètode de Simpson

Combinant les expressions que hem obtingut per l'error de les fórmules del rectangle i del trapezi podem obtenir una nova fórmula d'integració en la qual hagi desaparegut el terme amb E de l'error. Així, si definim

$$S(f) = \frac{2}{3}R(f) + \frac{1}{3}T(f),$$

i tenim en compte les expressions de $R(f)$ i $T(f)$, $S(f)$ es pot escriure explícitament com

$$S(f) = \frac{1}{6}\sum_{i=1}^{n} h_i\left[f(x_i) + 4f\left(\frac{x_i + x_{i+1}}{2}\right) + f(x_{i+1})\right].$$

Aquesta és la *regla de Simpson composta*. Aquesta fórmula es pot obtenir d'altres maneres, per exemple interpolant localment la funció $f(x)$ per polinomis de grau 2, fent ús dels extrems i el punt mitjà de cada subinterval d'integració per obtenir la fórmula de Simpson simple. Com en el cas de les fórmules del rectangle i del trapezi, si després sumem per tots els subintervals, obtenim la fórmula de Simpson composta.

L'error de la fórmula de Simpson es pot obtenir directament a partir de l'error de les fórmules del rectangle i del trapezi:

$$\begin{aligned} I(f) - S(f) &= \frac{2}{3}(I(f) - R(f)) + \frac{1}{3}(I(f) - T(f)) = \\ &= -\frac{2}{3}F + ... = -\frac{1}{2880}\sum_{i=1}^{n} h_i^5 f^{(iv)}(\xi_i) + \dots \end{aligned}$$

Per al mètode de Simpson podem fer els següents comentaris:

a) Malgrat que el càlcul de $S(f)$ es basa en la integració de polinomis de grau 2, l'error de la fórmula involucra derivades d'ordre més gran o igual que 4, per tant el mètode de Simpson és exacte per tots els polinomis de grau menor o igual que 3.

b) De manera semblant al que hem dit per les fórmules del rectangle i del trapezi, si dupliquem el nombre de subintervals, dividint per dos els passos d'integració, en principi l'error total es redueix en un factor proper a 1/16.

c) Podem combinar dues fórmules de Simpson amb diferents passos, de manera que per la fórmula resultant l'error comenci amb termes h_i^7 i $f^{(vi)}(\xi)$.

d) Al comentar les propietats de tots els mètodes exposats, hem suposat sempre qüestions del tipus " $h^2 f''(x)$ és més gran que $h^4 f^{(iv)}(y)$ ". La validesa d'aquest tipus d'afirmacions depèn de la funció $f(x)$. Els mètodes que tot seguit discutirem tenen en compte aquesta mena de propietats.

2.3 Problemes

28. Tenim tabulada la funció $f(x) = \sin(x)$ amb 5 xifres arrodonides i calculem $f'(0.900)$ substituint els valors tabulats en la fórmula centrada

$$f'(0.900) \simeq \frac{f(0.900+h) - f(0.900-h)}{2h} ,$$

per a diferents valors del pas h. La següent taula ens dóna les aproximacions obtingudes i l'error $(f'(0.900) - \cos(0.900))$

h	$f'(0.900)$	Error
0.001	0.62500	0.00339
0.002	0.62250	0.00089
0.005	0.62200	0.00039
0.010	0.62150	−0.00011
0.020	0.62150	−0.00011
0.050	0.62140	−0.00021
0.100	0.62055	−0.00106

Determineu el pas teòric òptim perquè l'error en l'aproximació de $f'(0.900)$ sigui mínim. Compareu-lo amb els valors de la taula. **Indicació**: En la fórmula centrada anterior el pas teòric òptim s'obté calculant el mínim de la següent expressió: $e(h) = \epsilon/h + h^2 M/6$, on $e(h)$ és l'error comés en funció del pas, ϵ és una fita de l'error d'arrodoniment comés en l'avaluació de $f(x)$ i M és una fita de $f'''(x)$ en l'interval on treballem. En el nostre cas $\epsilon \simeq \frac{1}{2}10^{-5}$ i $M = \max|f'''(x)|$ en l'interval $[0.800, 1.000]$. *Resposta*: $h = (3\epsilon/M)^{1/3} \simeq 0.028$.

29. Calculeu la derivada primera de la funció $f(x) = \tan(x)$ en el punt $a = 1$ emprant les següents fórmules:

$$f'(a) \simeq \frac{f(a+h) - f(a)}{h} , \; f'(a) \simeq \frac{f(a) - f(a-h)}{h} , \; f'(a) \simeq \frac{f(a+h) - f(a-h)}{2h} ,$$

$$f'(a) \simeq \frac{-f(a+2h) + 4f(a+h) - 3f(a)}{2h} , \; f'(a) \simeq \frac{3f(a) - 4f(a-h) + f(a-2h)}{2h} ,$$

i utilitzant els passos $h = 0.1, 0.05, 0.02$. Avalueu l'error en cada aproximació, compareu-lo amb el valor exacte i determineu-ne el pas òptim. *Resposta:* 4.0735193, 3.7181518, 3.5361346, $h_{op} \simeq 6.9 \cdot 10^{-5}$; 2.972495, 3.1805026, 3.3224727, $h_{op} \simeq 6.9 \cdot 10^{-5}$; 3.5230072, 3.4493272, 3.4293037, $h_{op} \simeq 7.5 \cdot 10^{-4}$; 3.1868155, 3.3627841, 3.4170966, $h_{op} \simeq 6.38 \cdot 10^{-4}$; 3.3061443, 3.3885103, 3.41869, $h_{op} \simeq 1.64 \cdot 10^{-3}$.

30. Obtenim un cos de revolució fent girar al voltant de l'eix x la corba $y = 1 + \frac{x^2}{4}$ on $x \in [0, 2]$. Calculeu el volum d'aquest objecte utilitzant:

a) la regla dels trapezis amb $N = 2, 4, 8$ on $h = (b - a)/N$. Calculeu l'error per a cada N,

b) la regla de Simpson amb els mateixos valors de N i també calculant l'error.

Resp.: 11.7286.

31. Doneu una aproximació de l'àrea de la regió acotada per la corba

$$y = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x/\sigma)^2/2}$$

i l'eix de les x a l'interval $[-\sigma, \sigma]$. Useu la regla composta dels trapezis per $N = 2, 4, 8$ i 16. Feu els mateixos calculs usant els intervals $[-2\sigma, 2\sigma]$ i $[-3\sigma, 3\sigma]$. (Indicació: Previ al càlcul numèric feu el canvi de variables $t = x/\sigma$). *Resposta*: 0.683, 0.954, 0.997.

32. Determineu una aproximació de les integrals següents:

a)
$$\int_{-4}^{4} \frac{dx}{1 + x^2} \quad \text{amb un error menor que } \epsilon = 10^{-3} \ ,$$

b)
$$\int_{0}^{1} e^{-10x} \sin x dx \quad \text{amb un error menor que } \epsilon = 10^{-7} \ .$$

Resposta: a) $I(f) \simeq 2.65178 \pm 10^{-3}$. b) $I(f) \simeq 0.009878 \pm 10^{-7}$.

33. Calculeu el valor de la integral $\int_1^2 \frac{dx}{x}$ amb un error de 10^{-6} usant el mètode de Romberg. *Resposta:* 0.693147.

2.4 Qüestions

34. Per calcular $f^{'}(a)$, quina fórmula és millor quan h és petita?

a) $\frac{f(a+h)-f(a-h)}{2h}$ b) $\frac{f(a+h)-f(a)}{h}$
c) $\frac{f(a)-f(a-h)}{h}$ d) Totes les anteriors tenen el mateix ordre

35. Per quins coeficients a, b, c la fórmula d'integració

$$\int_0^1 f(x)\,dx = af(0) + bf(1/2) + cf(1)$$

és exacta per polinomis de grau més petit o igual que dos?

a) $a = 1/12,\ b = 1/3,\ c = 1/6,$ b) $a = 5/12,\ b = 2/12,\ c = 5/12,$
c) $a = 1/2,\ b = 0,\ c = 1/2,$ d) $a = 1/6,\ b = 2/3,\ c = 1/6.$

36. Calculeu

$$\int_0^1 \frac{1}{1+x}dx$$

usant trapezis amb 4 subintervals. Quin és el valor obtingut ?

a) 0.712 b) 0.699 c) 0.697 d) 0.693

37. Volem calcular les derivades primera i segona d'una funció de la qual en coneixem la següent taula de valors

x_i	1.9	2.0	2.1
$f(x_i)$	12.7032	14.7781	17.1490

Usem fórmules centrades que utilitzen els punts $(x_i - h, f(x_i - h))$ i $(x_i + h, f(x_i + h))$ per la primera derivada i $(x_i, f(x_i))$ juntament amb els punts anteriors per la segona derivada. Quines aproximacions obtenim pels valors de $f'(2.0)$ i $f''(2.0)$?

a) 23.709 i 30.77, b) 20.749 i 28.5, c) 22.229 i 29.6, d) Cap de les anteriors.

38. Considerem la següent fórmula d'integració numèrica:

$$\int_{-1}^{1} f(x)dx = \frac{(5f(-\sqrt{0}.6) + 8f(0) + 5f(\sqrt{0}.6))}{9}$$

Quina de les següents afirmacion és certa:

a)És exacta per polinomis de grau 6,
b)És exacta per polinomis de grau $\leq$ que 5,
c)No és exacta per polinomis de grau 3,
d)Cap de les anteriors.

39. Quan aproximem la derivada d'una funció mitjançant la fórmula:

$$f'(x_0) = \frac{-f(x_0+2h) + 8f(x_0+h) - 8f(x_0-h) + f(x_0-2h)}{12h}$$

l'error és proporcional a:

a) h^4 b) h^2 c) h^3 d) h.

40. Coneixem els següents valors d'una funció:

x	1.8	2	2.2	2.4	2.6	2.8	3	3.2	3.4
y	6.050	7.389	9.025	11.023	13.464	16.445	20.086	24.533	29.964

Quant val $\int_{1.8}^{3.4} f(x)dx$, si la calculem pel mètode de Simpson amb pas h=0.4 ?

a) 23.9301 b) 23.5902 c) 23.9149 d) Cap dels anteriors

41. Volem calcular

$$\int_0^1 x^2 dx,$$

utilitzant la fórmula de Simpson amb 20 intervals. Quin resultat obtenim?

a) 0.3336 b) 0.3325 c) 0.3321 d) 0.3333

42. Calculeu per Trapezis $\int_1^2 (1+x^2)dx$ amb $h = 0.25$.

a) 3.317521, b) 3.333333, c) 3.343750, d) 3.324519.

43. Per calcular la integral $\int_1^{1.3} \sqrt{x}dx$, ens donen la següent fórmula

$$\frac{0.05}{3}\left(1 + 4(1.0247 + 1.0723 + 1.1180) + 2(1.0488 + 1.0954) + 1.1401\right)$$

De quin tipus de fórmula es tracta?

a) Trapezi , b) Simpson , c) Rectangle , d) Cap de les anteriors.

44. Calculeu la integral de la funció $f(x) = 3x^3 + 5x - 1$ en l'interval $[0, 1]$ utilitzant la fórmula de Simpson amb 2 intervals. Compareu el valor obtingut amb el valor exacte. Per què hem obtingut aquest error?

a) La funció és senar, b) Hem tingut sort,
c) Simpson integra exactament polinomis de grau 3, d) La funció és parella.

45. Quant val l'aproximació de la integral $\int_0^1 \frac{1}{2+x}dx$ aplicant la fórmula dels trapezis usant 2 intervals?

a) 0.408333, b) 0.816666, c) 0.405465, d) Cap de les anteriors.

46. El resultat d'aproximar $f'(4.5)$ a partir dels valors $f(4.506) = 1.23378$ i $f(4.494) = 1.21227$ és:

a) 1.7925, b) 3.585, c) -3.585, d) -1.7925.

47. Volem calcular, fent ús de la fórmula de Simpson composta

$$\int_0^{\pi/2} \sin x dx,$$

amb un error menor que 10^{-6}. Quants subintervals cal prendre com a mínim?

a) $n = 6$, b) $n = 9$, c) $n = 10$, d) Cap de les anteriors.

48. L'error de truncació de la fórmula

$$f'(a) \approx \frac{1}{2h}\left(-3f(a) + 4f(a+h) - f(a+2h)\right)$$

és d'ordre

a) h^4, b) h^3, c) h^2, d) Cap de les anteriors.

49. Si aproximem el valor de la integral, $\int_a^b f(x)dx$, usant la fórmula del trapezi simple $T(f,h)$, quin nom rep l'error que cometem al considerar $\int_a^b f(x)dx \simeq T(f,h)$?

a) Error d'arrodoniment del mètode, b) Error de cancel·lació del mètode,
c) Error de truncament del mètode, d) Error fatal del mètode.

50. Quant val $\int_{-1}^{1} e^x dx$ si es calcula usant la fórmula dels trapezis amb $h = 1/2$?

a) 3.251432, b) 2.210551, c) -2.014540, d) 2.399166.

51. Sabem que

$$f'(a) = \frac{1}{12h}\left(f(a-2h) - 8f(a-h) + 8f(a+h) - f(a-h)\right) + \frac{h^4}{30}f^{(5)}(\eta)$$

Suposem que la derivada d'ordre n de la funció f es pot fitar per 10^n i el valors de f es poden calcular amb un error d'arrodoniment fitat per $5 \cdot 10^{-5}$. El valor de h que cal pendre per tal que la derivada de f en a presenti el mínim error possible és:

a) 0.022,
a) 10^{-7},
a) 0.112,
d) Cap dels anteriors.

Capítol 3

ZEROS DE FUNCIONS NO LINEALS

3.1 Introducció

Un problema al qual anem a parar tot sovint és el del càlcul de les arrels (zeros o solucions) d'una funció $f(x)$. Suposarem en tot el que segueix que $f(x)$ és una funció real, en general no lineal, d'una variable. Exemples de funcions no lineals són

$$\begin{aligned} f(x) &= \sin x - x + 2 \ , \\ f(x) &= x^3 - 2 \ , \\ f(x) &= e^x + \log x - 3 \ . \end{aligned}$$

Volem trobar les solucions de l'equació $f(x) = 0$. En la majoria dels casos aquesta equació no es pot resoldre explícitament. El problema que plantegem aquí és l'obtenció, amb una aproximació arbitrària, de les seves arrels.

Un exemple d'un problema que ens condueix al càlcul del zero d'una funció és aquest:

Volem calcular l'àrea de la regió situada entre les corbes $y = e^{-x}$ i $y = \cos x$ dins de l'interval, per les x, $[0, \frac{\pi}{2}]$. Si tenim en compte el comportament de les dues funcions, la solució ve donada per

$$\begin{aligned} \text{àrea} &= \int_0^z (\cos x - e^{-x}) dx + \int_z^{\frac{\pi}{2}} (e^{-x} - \cos x) dx \ = \\ &= 2 \sin z + 2e^{-z} - 2 - e^{-\frac{\pi}{2}} \end{aligned}$$

on z és el punt de tall de les funcions e^{-x} i $\cos x$ a l'interval $(0, \frac{\pi}{2})$. És clar que per obtenir el valor numèric de la solució d'aquest problema hem de determinar el valor de x que satisfà l'equació

$$f(x) = \cos x - e^{-x} = 0 \ , \quad x \in [0, \frac{\pi}{2}] \ .$$

Els mètodes que descriurem per resoldre aquest tipus de problemes no determinen les seves solucions de manera directa (ja hem dit que en la major part dels casos això no és possible). El

que trobarem serà una successió de valors $\{x_n\}_{n\in\mathbb{N}}$ convergent a un valor α solució de l'equació plantejada. És a dir

$$\lim_{n\to\infty} x_n = \alpha \ , \ \ f(\alpha) = 0 \, .$$

Aleshores, pendrem com solució del problema un dels termes x_k de la successió que verifiqui

$$|f(x_k)| < \text{Tolerància}_1 \ , \ \text{ o bé}$$

$$|x_{k-1} - x_k| < \text{Tolerància}_2 \ ,$$

on els valors de les toleràncies caldrà que estiguin fixats d'entrada.

És clar que no totes les equacions tenen un únic zero en un cert domini. Per exemple, la funció $f(x) = x^6 \sin \frac{1}{x}$ té un nombre infinit d'arrels en tot interval que contingui l'origen. Per tant, en qualsevol procés de càlcul d'arrels d'una equació no lineal haurem de distingir tres fases

a) **Localització**: Hem de conèixer la zona on es troben les arrels. En general, aquesta informació es pot obtenir a partir d'un estudi analític de la funció o també a partir d'una representació gràfica aproximada. En molts casos, les equacions provenen d'un problema tècnic o científic, el coneixement del qual també pot ajudar a la localització de les arrels.

b) **Separació**: Alguns cops tenim que dues arrels diferents d'una equació estan molt pròximes. En aquests casos ens convindrà separar les arrels, és a dir, determinar dominis que continguin una única arrel de l'equació.

c) **Aproximació Numèrica**: L'objectiu que ens plantegem és l'obtenció d'una successió de valors que convergeixi cap a l'arrel buscada. Aquesta successió es construirà, normalment, de manera iterativa a partir d'uns certs valors inicials que suposarem suficientment pròxims a l'arrel cercada. A partir de $x_0, x_1, \ldots, x_m$, obtindrem $x_{m+1} = G(x_m, \ldots, x_0)$ i, de forma més general, $x_{k+1} = G(x_k, \ldots, x_{k-m})$, de manera que

$$\lim_{k\to\infty} x_k = \alpha \ \text{ amb } \ f(\alpha) = 0 \, .$$

3.2 Alguns mètodes d'aproximació de solucions

3.2.1 Mètode de bisecció

La fonamentació teòrica d'aquest mètode la dóna el següent teorema:

Teorema de Bolzano

Sigui $f : [a, b] \to \mathbb{R}$ *contínua en l'interval* $[a, b]$ *i tal que* $f(a)f(b) < 0$, *llavors existeix* $\alpha \in (a, b)$ *tal que* $f(\alpha) = 0$.

En general podríem tenir més d'una arrel en l'interval $[a, b]$ però suposarem que, si cal, hem aplicat prèviament un procés de localització i separació per tal que això no sigui així. El mètode de bisecció consisteix en construir una successió d'intervals encaixats,

$$(a, b) = (a_0, b_0) \supset (a_1, b_1) \supset \ldots \supset (a_k, b_k) \supset \ldots ,$$

de manera que sempre continguin l'arrel i que l'amplitud de cada interval sigui la meitat de l'anterior, és a dir, $b_k - a_k = 2(b_{k+1} - a_{k+1})$. Quan l'amplitud de l'interval sigui prou petita podrem considerar com una bona aproximació d'aquesta qualsevol punt de l'últim interval calculat, en particular el punt mitjà.

La successió d'intervals es construeix així:

a) Partim de l'interval (a_0, b_0) tal que $f(a_0)f(b_0) < 0$, calculem el punt $c_1 = (a_0 + b_0)/2$.

b) Si $f(c_1) = 0$, c_1 és l'arrel buscada.

c) Altrament es considerarà com interval (a_1, b_1) l'interval (a_0, c_1) o el (c_1, b_0) segons que $f(a_0)f(c_1) < 0$ o $f(c_1)f(b_0) < 0$, respectivament.

d) El procés es continua de manera iterativa, però a partir ara de l'interval (a_1, b_1).

Comentari: El mètode de bisecció té l'avantatge de ser sempre convergent però la seva velocitat de convergència és lenta. Per aquest mètode es pot donar una estimació del nombre d'iteracions necessàries per obtenir una arrel α de l'equació $f(x) = 0$ amb una precisió prefixada. Es comença el procés en l'interval (a_0, b_0) i després de n passos obtenim l'interval (a_n, b_n) que conté c_{n+1}, punt mitj'a de l'interval, i l'arrel α cercada. Com la longitud de l'interval es redueix a la meitat en cada pas, tenim que

$$|c_{n+1} - \alpha| \leq \frac{b_0 - a_0}{2^{n+1}} .$$

Suposem que volem cercar l'arrel α amb una precisió menor que una certa quantitat petita ϵ, és a dir, volem que es compleixi,

$$|c_{n+1} - \alpha| \leq \epsilon .$$

Aquesta condició es verificarà si

$$\frac{b_0 - a_0}{2^{n+1}} \leq \epsilon , \text{ o equivalentment } 2^{n+1} \geq \frac{b_0 - a_0}{\epsilon} ,$$

i prenent logaritmes, vegem que el nombre, n, d'iteracions necessàries ha de verificar

$$n \geq \frac{\log(\frac{b_0 - a_0}{\epsilon})}{\log 2} - 1 .$$

Així, si volem obtenir l'arrel d'una funció amb una precisió menor que $\epsilon = 10^{-6}$ i partim d'un interval (a_0, b_0) de longitud 1, obtindrem la solució cercada al determinar c_{19}.

Exemple: Volem cercar l'única arrel de la funció $f(x) = x - e^{-x} = 0$ que està en l'interval $(0.5, 0.6)$. Aplicant el mètode de bisecció obtenim la següent taula:

i	(a_i, b_i)	$f(a_i)$	$f(b_i)$	$c_{i+1} = (b_i + a_i)/2$	$f(c_{i+1})$
0	(0.5,0.6)	-0.1065	0.0512	0.55	-0.0270
1	(0.55,0.6)	-0.0270	0.0512	0.575	0.0123
2	(0.55,0.575)	-0.0270	0.0123	0.5625	-0.0073

En aquest exemple tenim que $|c_3 - \alpha| \leq 10^{-1} * 2^{-3} = 0.0125$ i $|f(c_3)| \simeq 7.3 * 10^{-3}$.

3.2.2 Mètode de Newton–Raphson

Suposem ara que la funció f és dues vegades derivable en un interval (a, b) ($f \in \mathcal{C}^2(a, b)$). Sigui x_n una aproximació de l'arrel α de l'equació $f(x) = 0$ tal que $f'(x_n) \neq 0$ i $h = \alpha - x_n$ és "petit". Considerem el desenvolupament de Taylor de la funció f al voltant del punt x_n,

$$f(x) = f(x_n) + f'(x_n)(x - x_n) + \frac{f''(\xi(x))}{2}(x - x_n)^2 \ ,$$

on $\xi(x) \in < x, x_n >$. Avaluant el desenvolupament en el punt $\alpha = x_n + h$ obtenim

$$0 = f(\alpha) = f(x_n + h) = f(x_n) + f'(x_n)h + \frac{f''(\xi(x_n))}{2}h^2 \ .$$

Suposem que, com h és petit, el terme que conté h^2 és menyspreable i per tant

$$0 \simeq f(x_n) + f'(x_n)h \ .$$

Isolant h d'aquesta equació obtenim

$$h \simeq h_n = -\frac{f(x_n)}{f'(x_n)} \ .$$

Aquesta fórmula ens dóna una correcció, h_n, per la nostra aproximació, x_n, de l'arrel α que anomenem $x_{n+1} = x_n + h_n$. El mètode de Newton consisteix en: *donada una aproximació inicial x_0 generar una successió de valors $\{x_n\}$, definida per*

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} \ , \quad n = 0, 1 \ldots \ .$$

Aquest procés s'haurà d'aturar quan $|x_{n+1} - x_n| < \epsilon$ o bé $|f(x_{n+1})| < \delta$, on ϵ i δ són les toleràncies que estem disposats a acceptar.

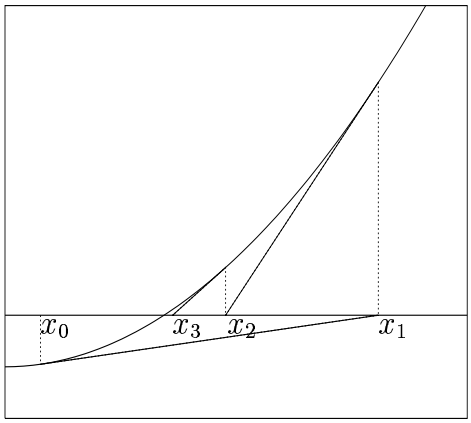

Figura 4.1. Mètode de Newton

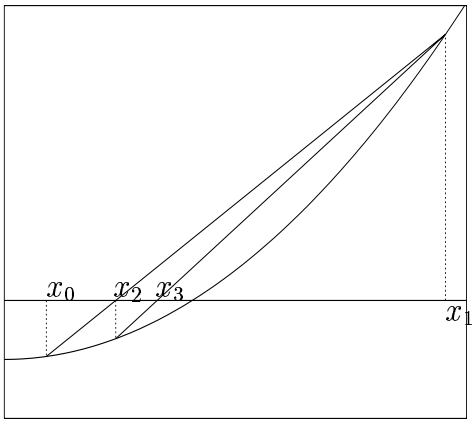

Figura 4.2. Mètode de la secant

En la Figura 4.1 s'il·lustra gràficament com les aproximacions x_n de l'arrel α s'obtenen usant successives rectes tangents a la corba $y = f(x)$ en els punts $(x_n, f(x_n))$.

Comentari: El mètode de Newton no és sempre convergent per qualsevol valor de x_0 i sempre és convenient escollir l'aproximació inicial x_0 tan propera com sigui possible a l'arrel buscada. Ara bé, quan tenim convergència, aquesta sol ser ràpida.

Exemple: Volem cercar l'única arrel de la funció $f(x) = x - e^{-x} = 0$ amb $x_0 = 0.55$. Aplicant el mètode de Newton obtenim la següent taula:

n	x_n	$f(x_n)$	$f'(x_n)$	$f(x_n)/f'(x_n)$
0	0.55	-0.026950	1.576950	-0.017090
1	0.567090	-0.000084	1.567174	-0.000053
2	0.567143	$-5.4 * 10^{-8}$		

Notem que x_1 té tres xifres iguals que l'arrel α i x_2 en té sis.

Definició*: Direm que α és una arrel amb multiplicitat m de l'equació $f(x) = 0$ si es compleix que $f(\alpha) = f'(\alpha) = \cdots = f^{(m-1)}(\alpha) = 0$, però $f^{(m)}(\alpha) \neq 0$.*

Nota: El mètode de Newton presenta problemes quan l'arrel α de la funció f té multiplicitat més gran que 1. En aquest cas tenim una convergència més lenta que l'esperada. Tot i això, el mètode pot ser modificat per tal d'adaptar-lo al cas d'una arrel múltiple. Si la funció f és $m + 1$ vegades derivable amb continuïtat ($f \in \mathcal{C}^{m+1}$) i α és una arrel de la funció f amb multiplicitat m

tal que $f^{(m+1)}(\alpha) \neq 0$, podem escriure el següent mètode iteratiu que serà usat per cercar arrels de multiplicitat m.

$$x_{n+1} = x_n - m\frac{f(x_n)}{f'(x_n)} \ , \quad n = 0, 1 \ldots .$$

El procés iteratiu s'aturarà d'igual forma que s'ha exposat en el mètode de Newton

Newton				Newton per arrels múltiples		
n	x_n	$f(x_n)$		n	x_n	$f(x_n)$
0	0.55	0.00726228		0	0.55	0.00726228
1	0.55854492	0.00018214		1	0.56708983	$7.0 * 10^{-9}$
2	0.56283740	0.00004561		2	0.56714329	$6.5 * 10^{-18}$
3	0.56498866	0.00001141				
4	0.56606556	0.00000285				
5	0.56660432	0.00000071				
6	0.56687378	0.00000018				
7	0.56700853	0.00000004				
8	0.56707591	0.00000001				
9	0.56710960	$2.8 * 10^{-9}$				

Exemple: Volem cercar l'única arrel doble de la funció $f(x) = (x - e^{-x})^2 = 0$ amb $x_0 = 0.55$. En la taula anterior podem comparar els valors obtinguts en aplicar el mètode de Newton i el mètode de Newton modificat per la recerca d'arrel múltiples, en aquest cas $m = 2$. Observem que el mètode de Newton convergeix cap a l'arrel molt lentament i en canvi el mètode modificat ho fa més ràpidament.

3.2.3 Mètode de la secant

Considerem el mètode de Newton i aproximem la derivada $f'(x_n)$ pel quocient de les diferències

$$f'(x_n) \simeq \frac{f(x_n) - f(x_{n-1})}{x_n - x_{n-1}}$$

formades per dues aproximacions successives. Obtenim així l'expressió

$$x_{n+1} = x_n - f(x_n)\frac{x_n - x_{n-1}}{f(x_n) - f(x_{n-1})} \ , \quad n = 1, 2, \ldots .$$

Igual com en el mètode de Newton, el procés s'haurà d'aturar quan $|x_{n+1} - x_n| < \epsilon$ o bé $|f(x_{n+1})| < \delta$, on ϵ i δ són les toleràncies que estem disposats a acceptar.

El mètode iteratiu que s'acaba de descriure es coneix amb el nom de *mètode de la secant*. En la Figura 4.2 s'il·lustra gràficament com les aproximacions x_{n+1} s'obtenen usant successives rectes pels punts $(x_{n-1}, f(x_{n-1}))$ i $(x_n, f(x_n))$.

Comentari: El mètode de la secant no és sempre convergent encara que prenguem els punts inicials propers a l'arrel. En particular pot passar que per un cert n, $f(x_n)$ no estigui definida.

Exemple: Volem cercar l'única arrel de la funció $f(x) = x - e^{-x} = 0$ que està en l'interval $(0.55, 0.575)$. Aplicant el mètode de la secant obtenim la següent taula:

n	x_n	$f(x_n)$
0	0.55	-0.026950
1	0.575	0.012295
2	0.567168	0.000038
3	0.567143	$-5.4 * 10^{-8}$

En aquest exemple tenim que $|x_3 - x_2| \simeq 2.5 * 10^{-5}$ i $|f(x_3)| \simeq 5.4 * 10^{-8}$.

3.3 Teoria general de la iteració simple

L'equació $f(x) = 0$ es pot escriure en la forma $x = g(x)$ usant operacions elementals i, recíprocament, $x = g(x)$ es pot posar com $f(x) = 0$. Llavors, si α és una arrel de l'equació $f(x) = 0$, tenim que $\alpha = g(\alpha)$, on α s'anomena **punt fix** de l'aplicació $x \to g(x)$. Un mètode iteratiu de la forma $x_{n+1} = g(x_n)$ s'anomena **procés d'iteració simple** (veure Figura 4.3). Un exemple d'un mètode d'iteració simple és el mètode de Newton $x_{n+1} = x_n - f(x_n)/f'(x_n)$ que podem escriure com $x_{n+1} = g(x_n)$ on $g(x) = x - f(x)/f'(x)$.

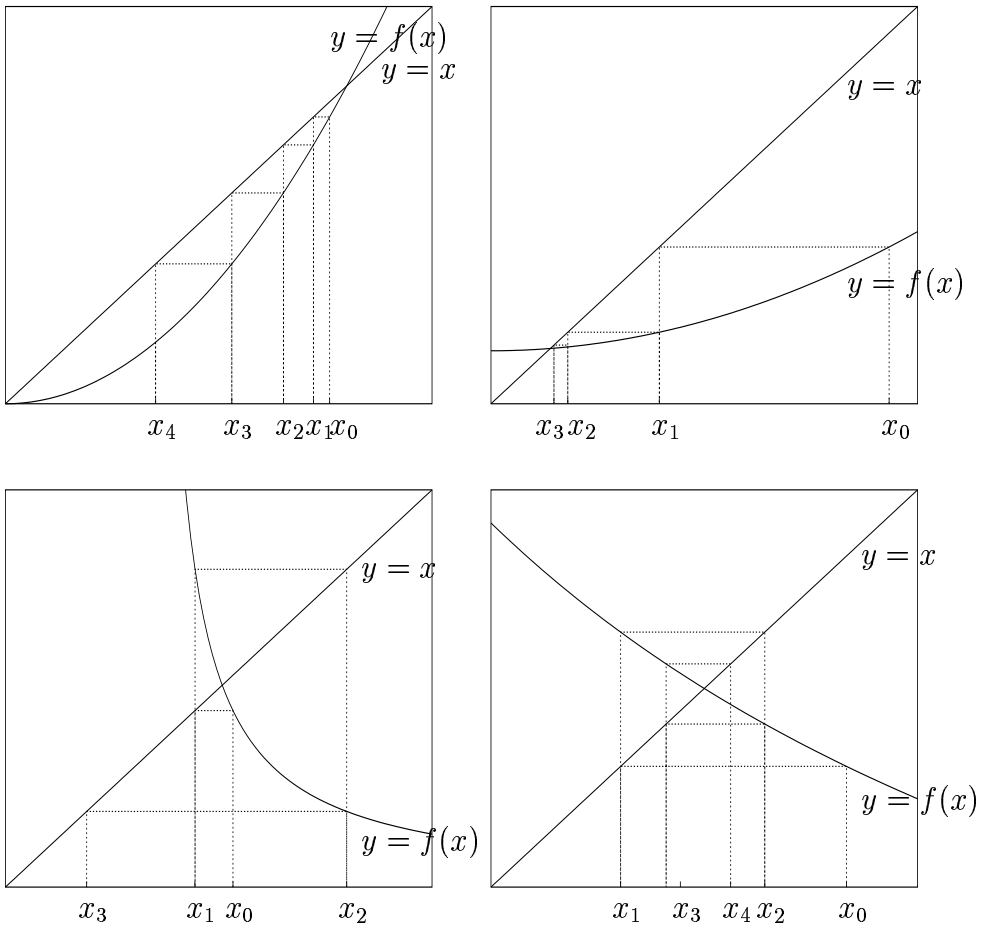

Figura 3.1: Representació gràfica de procés d'iteració simple.

Exemple: Volem cercar l'única arrel de la funció $f(x) = x - e^{-x} = 0$ amb $x_0 = 0.55$. Aplicant mètodes d'iteració simple. L'equació anterior es pot escriure de diferents formes, per exemple podem posar $x = e^{-x}$ o bé prenent logaritmes $x = -\log x$. Aquestes expressions donen lloc a dos mètodes iteratius, $x_{n+1} = g(x_n)$, $x_{n+1} = e^{-x_n}$ i $x_{n+1} = -\log(x_n)$. Aplicant aquest mètodes iteratius obtenim la següent taula:

$x_{n+1} = e^{-x_n}$			$x_{n+1} = -\log x_n$	
n	x_n		n	x_n
0	0.55		0	0.55
1	0.57695		1	0.597837
2	0.561609		2	0.514437
3	0.570291		3	0.664682
4	0.565361		4	0.408447
.	.		5	0.895394
.	.		6	0.110492
.	.		7	2.202816
.	.		8	-0.789737
.	.			
.	.			
16	0.567141			
17	0.567144			
18	0.567143			

Es veu que el primer dels mètodes convergeix lentament cap a la solució, després de 18 iterats, mentre que el segon d'ells divergeix. Per tant hauríem de tenir algun criteri perquè la funció g ens donés una successió convergent.

Teorema del punt fix*: Sigui $g \in \mathcal{C}^0([a,b])$ tal que $g(x) \in [a,b]$, $\forall x \in [a,b]$. A més, suposem que existeix g' en l'interval (a,b) amb*

$$|g'(x)| \leq k < 1 \quad \forall x \in (a,b) \ .$$

Si p_0 és qualsevol nombre de l'interval $[a,b]$, llavors la successió definida per

$$p_n = g(p_{n-1}) \ , \quad n \geq 1 \ ,$$

convergeix a l'únic punt fix p que té $g(x)$ en l'interval $[a,b]$.

En l'últim exemple que hem vist teníem els dos mètodes iteratius $x_{n+1} = e^{-x_n}$ i $x_{n+1} = -\log x_n$. En el primer cas, tenim

$$g(x) = e^{-x} \ , \quad g'(x) = -e^{-x} \ , \text{ i per tant } |g'(x)| \simeq 0.567 < 1 \ , \text{ per } x \simeq \alpha \ .$$

Mentre en el segon cas, tenim

$$g(x) = \log(-x) \ , \quad g'(x) = \frac{-1}{x} \ , \text{ i per tant } |g'(x)| \simeq \frac{1}{0.567} > 1 \ , \text{ per } x \simeq \alpha \ .$$

Hem vist, observant els anteriors exemples, que el mètode de Newton convergeix més ràpidament que no pas el mètode d'iteració simple. Per tant haurem de saber comparar la raó de convergència de diferents mètodes iteratius, per això donem la següent definició:

Definició d'ordre: *Sigui* $x_0, x_1, x_2, \ldots$ *un successió convergent a* α*, i escrivim* $\epsilon_n = x_n - \alpha$*. Direm que aquesta successió té* **ordre de convergència** p*, si* p *és el mes gran dels nombres reals positius tals que*

$$\lim_{n\to\infty} \frac{|\epsilon_{n+1}|}{|\epsilon_n|^p} = C < \infty \quad C \neq 0 \; .$$

C *s'anomena* **coeficient asimptòtic de l'error**. *Per* $p = 1$ *(aleshores* $C < 1$*) i* $p = 2$*, la convergència es diu* **lineal** *o* **quadràtica***, respectivament.*

Notes:

I. Considerem l'equació $x = g(x)$ i sigui α un punt fix.

a) Si g és una contracció en un entorn d'α (per això basta que $\mid g'(x) \mid < 1$ en aquest entorn), el mètode d'iteració simple $x_{n+1} = g(x_n)$ ($n \leq 0$) té almenys ordre 1.

b) Si g és m vegades derivable i $g'(\alpha) = \cdots = g^{(m-1)}(\alpha) = 0$, llavors el mètode d'iteració simple $x_{n+1} = g(x_n)$ ($n \leq 0$) té almenys ordre m. Si $g^{(m)}(\alpha) \neq 0$, l'ordre és m.

II. Considerem ara l'equació $f(x) = 0$ i suposem que el mètode emprat convergeix a l'arrel α.

a) Si $f'(\alpha) \neq 0$, el mètode de Newton té convergència almenys quadràtica ($p = 2$).

b) Si $f'(\alpha) = 0$, el mètode de Newton té convergència lineal ($p = 1$).

c) Si $f'(\alpha) \neq 0$, el mètode de la secant té convergència super–lineal ($p = \frac{1+\sqrt{5}}{2} > 1$).

d) Independentment del comportament de f es pot considerar que el mètode de bisecció té convergència lineal, ja que, en cada pas, reduïm a la meitat l'interval d'error on es troba el zero buscat.

3.4 Problemes

52. Demostreu que $f(x) = e^x - x - 2$ té 2 zeros. Calculeu el valor del més petit pels següents mètodes: a) Newton, b) Secant, c) Bisecció, d) Iteració simple. *Resposta:* -1.84141.

53. Una barreja equimolecular de monòxid de carboni i oxigen arriba a l'equilibri a T= $300°$K i 5 atm de pressió. La reacció teòrica és $CO + \frac{1}{2}O_2 \rightleftharpoons CO_2$, mentre que la reacció real s'escriu com: $CO + O_2 \rightarrow xCO + \frac{1}{2}(1+x)O_2 + (1-x)CO_2$. L'equació d'equilibri químic que determina la fracció x del CO que queda, s'escriu com:

$$K_p = \frac{(1-x)(3+x)^{1/2}}{x(x+1)^{1/2}P^{1/2}} \ , \quad 0 < x < 1 \ ,$$

on $K_p = 3.06$ és la constant d'equilibri per a $CO + \frac{1}{2}O_2 = CO_2$ a $300°$K, i $P = 5$ atm és la pressió. Determineu el valor de x pel mètode de Newton. *Resposta:* 0.192962.

54. Utilitzeu el mètode de Newton per cercar l'arrel doble de l'equació

$$f(x) = \left(\sin x - \frac{x}{2}\right)^2 = 0 \text{ amb } x_0 = \frac{\pi}{2} \ ,$$

en l'interval $[\frac{\pi}{2}, \pi]$ amb una exactitud de 10^{-5}. Resoleu també l'equació considerant com a punts inicials $x_0 = 5\pi$ i $x_0 = 10\pi$. *Resposta:* 1.89549

55. Considerem la funció $f(x) = e^{\frac{1}{x}} - x$. Volem cercar el zero de $f(x)$ usant fórmules d'iteració simple ($x_{n+1} = g(x_n)$) amb diferents funcions d'iteració:

$$g_1(x) = e^{\frac{1}{x}} \ , \quad g_2(x) = x - x^2\frac{x - e^{\frac{1}{x}}}{x^2 + e^{\frac{1}{x}}} \ , \quad g_3(x) = x + \frac{1 - x\ln x}{1 + \ln x} \ .$$

En el cas de tenir convergència, comproveu que ho fan cap al zero de $f(x)$. En general, amb quina de les funcions d'iteració calen menys iteracions per tal d'obtenir el resultat amb una precisió donada, partint del mateix valor inicial? Comproveu-ho numèricament. *Resposta:* g_2, g_3, 1.76322.

56. a) Deduïu la fórmula recurrent del mètode d'interpolació inversa de 3 punts per al càlcul de zeros d'una funció f.

b) Aplicació: Calculeu els zeros de la funció

$$f(x) = \frac{1}{x} - 2\ln x - 0.5 \ ,$$

amb error menor que 10^{-6}, a partir de les aproximacions inicials següents $x_0 = 0.5$, $x_1 = 1$, $x_2 = 1.5$ *Resposta:* 1.18683.

Indicació: Si f és invertible i la seva inversa és $f^{-1} = g$ llavors tenim que $y = f(x)$ es pot escriure com $x = g(y)$. Usarem aquest fet per cercar un zero de l'equació $f(x) = 0$. La interpolació inversa iterada de 3 punts consisteix a:

- Considerar tres punts x_0, x_1, x_2 i les seves imatges $f(x_0), f(x_1), f(x_2)$.
- Interpolar la funció g, inversa de f, mitjançant un polinomi de grau 2, $q_2(y)$, que passa pels punts $(f(x_0), x_0), (f(x_1), x_1), (f(x_2), x_2)$.
- Prendre $x_3 = q_2(0)$ com una aproximació del zero buscat i repetir el procés amb els valors x_1, x_2, x_3.

3.5 Qüestions

57. Volem resoldre l'equació $x + \ln(x) = 0$ de la qual sabem que té un zero al voltant del punt $x = 0.5$, ho volem fer per iteració, quina de les següents expressions és la més adient ?

a) $x_{n+1} = -\ln(x_n)$, b) $x_{n+1} = e^{-x_n}$,
c) $x_{n+1} = (x_n + e^{-x_n})/2$, d) $x_{n+1} = (e^{-x_n} - \ln(x_n))/2$.

58. Volem aplicar el mètode de Newton per trobar una aproximació de $\sqrt{N}$, per això descomponem $N = AB$ i apliquem Newton amb $x_0 = A$. Quan val x_2 ?

a) $\frac{A+B}{2}$, b) $\frac{A+B}{4} + \frac{N}{A+B}$,
c) $\frac{A+B}{2} + \frac{N}{A+B}$, d) $\frac{A+B}{4} + \frac{4N}{A+B}$.

59. Volem trobar un punt fix de la funció $g(x) = x(1 - 2x)$. En quin dels següents intervals tenim convergència ?

a) $(-1, 1/2)$, b) $(1, 2)$, c) $(0, 1)$, d) $(0, 1/2)$.

60. Usem el mètode de Newton per cercar un zero de l'equació $3x + \sin(x) - \exp(x) = 0$. Quant val x_3 si considerem $x_1 = 0$?

a) 0.33333, b) 0.17111, c) 0.36017, d) Cap dels anteriors.

61. Apliquem el mètode de la secant per trobar un zero de l'equació $x^3 + x^2 - 3x - 3 = 0$. Agafem valors inicials $x_1 = 1$ i $x_2 = 2$. Quant val x_3?

a) 1.56142 b) 1.57142 c) 1.5733 d) 2.111

62. Una arrel de l'equació $x^3 + 2x^2 + 10x - 20 = 0$ és $x = 1.368808...$. Quina de les següents afirmacions és certa?

a) L'algorisme
$$x_n = \frac{20 - 2x_{n-1}^2 - x_{n-1}^3}{10}$$
és convergent ja que la derivada de la funció $(20 - 2x^2 - x^3)/10$ és més petita que 0 en les proximitats de la solució.

b) L'algorisme
$$x_n = \frac{20}{x_{n-1}^2 + 2x_{n-1} + 10}$$
és convergent ja que la derivada de la funció $20/(x^2 + 2x + 10)$ és més petita que 1 , en valor absolut, en les proximitats de la solució.

c) Cap dels dos algorismes és convergent.

d) Tots dos són convergents cap a l'arrel donada.

63. Quants zeros té l'equació $x - \cos x = 0$?

a) cap, b) 1 de simple, c) 1 de doble, d) 2 de simples.

64. Volem cercar un dels punts d'intersecció de les corbes $y = e^x - 2$ i $y = \ln(x+2)$, per això apliquem el mètode de Newton. Quin és el primer punt, x_1, que ens dóna el mètode si $x_0 = 1$?

a) 1, b) 1.146193, c) 0, d) 1.159471.

65. Volem utilitzar el mètode de Newton per calcular l'arrel $p-$èna d'una certa quantitat a. Quin algorisme creus que cal utilitzar?

a) $x_n = \frac{1}{p}\left(x_{n-1} + \frac{a^{p-1}}{x_{n-1}}\right)$, b) $x_n = x_{n-1} - \frac{x_{n-1}^p - a}{p x_{n-1}^{p-1}}$,
c) $x_n = \frac{1}{p}\left(x_{n-1} - \frac{a^{p-1}}{x_{n-1}}\right)$, d) $x_n = x_{n-1} + \frac{x_{n-1}^p - a}{p x_{n-1}^{p-1}}$

66. Usem el mètode de Newton per cercar un zero de l'equació $x^2 - e^x = 0$. Quant val x_3 si considerem $x_0 = 0$?

a) -0.70381, b) -0.70340, c) -0.70347, d) Cap dels valors anteriors.

67. Volem utilitzar el mètode de Newton per calcular l'arrel quadrada d'una certa quantitat a. Quin algorisme creus que cal utilitzar?

a) $x_n = \frac{1}{2}\left(x_{n-1} - \frac{a}{x_{n-1}}\right)$, b) $x_n = x_{n-1} - \frac{a}{x_{n-1}}$,
c) $x_n = \frac{1}{2}\left(x_{n-1} + \frac{a}{x_{n-1}}\right)$, d) $x_n = x_{n-1} + \frac{a}{x_{n-1}}$.

68. Sabem que $-1.84141...$ és una arrel de l'equació $e^x - x - 2 = 0$. Quin dels següents processos d'iteració simple convergeix més ràpid a l'arrel, si partim d'un punt x_0 proper a -1.84141?

a) $x_{n+1} = e^{x_n} - 2$, b) $x_{n+1} = \ln(x_n + 2)$,
c) $x_{n+1} = x_n - \frac{e^{x_n} - x_n - 2}{e^{x_n} - 1}$, d) Tots convergeixen igual de ràpid.

69. Usem el métode de la secant per calcular un zero de la funció $f(x) = 3x + \sin x - e^x$ amb punts inicials $x_0 = 0.3$ i $x_1 = 0.4$. Què val x_3?

a) 0.360421, b) 0.361262, c) 0.365100, d) 0.360409.

70. Volem calcular una arrel positiva de l'equació $x = 2 - e^x$ pel mètode d'iteració simple. Quina de les següents afirmacions és certa ?

a) L'anterior equació no té solució,

b) Si la x és positiva, la funció $f(x) = 2 - e^x$ té derivada més gran que 1 i per tant no es pot calcular,

c) Sols la podem calcular si $\mid 2 - e^x \mid < 1$,

d) El podem calcular si iterem la funció $\ln(2 - x)$.

71. Sigui $f(x) = 3x + \sin x - e^x$. Per resoldre l'equació $f(x) = 0$ apliquem els mètodes de Newton amb $x_0 = 0$ i de la secant amb $x_0 = 0, x_1 = 1$. Quant val per cada mètode x_3 ?

a) Newton $= 0.3604217$, secant $= 0.372277$ b) Newton $= 0.3722770$, secant $= 0.360421$

c) Newton $= 0.3604127$, secant $= 0.470990$ d) Newton $= 0.3333333$, secant $= 0.470990$

72. Volem utilitzar el mètode de Newton per calcular $\sqrt{a} + a$ per una certa quantitat a. Quina fórmula cal emprar ? a) $x_{n+1} = \frac{x_n^2+a}{2x_n-a}$ b) $x_{n+1} = \frac{x_n-a}{2} + \frac{a}{2x_n-2a}$ c) $x_{n+1} = \frac{x_n^2-a}{2x_n+a}$ d) $x_{n+1} = \frac{x_n+a}{2} + \frac{a}{2x_n-2a}$

Capítol 4

ÀLGEBRA LINEAL NUMÈRICA

4.1 Resolució de sistemes lineals

4.1.1 Introducció

Un sistema de n equacions lineals amb n incògnites es pot escriure com

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= b_2 \\ \cdots \quad \cdots \quad \cdots & \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n &= b_n \end{aligned}$$

o bé en forma matricial com $\mathbf{Ax} = \mathbf{b}$, on $\mathbf{A}$ és la matriu de coeficients, $\mathbf{b}$ el vector terme independent i $\mathbf{x}$ el vector d'incògnites:

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix}, \qquad \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \qquad \mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}.$$

A partir d'ara suposarem que la matriu $\mathbf{A}$ és no singular, és a dir, $|\mathbf{A}| \neq 0$. Llavors, l'existència i unicitat de la solució estàn garantides.

Hi ha diversos mètodes per a resoldre sistemes lineals. Ens limitarem aquí als *mètodes directes*, que donen la solució del sistema en un nombre finit de passos. Els *mètodes iteratius* permeten obtenir aproximacions successives de la solució del sistema.

4.1.2 Sistemes Triangulars

Es diu que una matriu **U** és triangular superior quan té tots els elements per sota de la diagonal iguals a zero. Això es representa:

$$\mathbf{U} = \begin{bmatrix} u_{11} & u_{12} & \cdots & u_{1n} \\ & u_{22} & \cdots & u_{2n} \\ & & \ddots & \vdots \\ & & & u_{nn} \end{bmatrix}.$$

Es diu que un sistema $\mathbf{Ux} = \mathbf{b}$ és triangular superior si la matriu de coeficients **U** és triangular superior. Un sistema triangular superior es pot resoldre fàcilment. Es comença per l'última equació:

$$x_n = \frac{b_n}{u_{nn}}.$$

Trobada x_n, es pot substituir el seu valor en la equació $n-1$, i trobar x_{n-1}, i així successivament. La fórmula general és:

$$x_i = \frac{b_i - \sum_{j=i+1}^{n} u_{ij} x_j}{u_{ii}} \qquad 1 \leq i < n.$$

De manera semblant es pot definir una matriu triangular inferior **L** com una matriu amb els elements per sobre de la diagonal iguals a zero. Els sistemes triangulars inferiors es resolen igual que els triangulars superiors, però començant per x_1. És fàcil comprovar que el nombre total d'operacions necessàries per a resoldre un sistema triangular d'n equacions és n^2.

4.1.3 Eliminació Gaussiana

És el més important dels mètodes directes. El procés d'eliminació gaussiana consisteix en transformar un sistema lineal en un altre equivalent triangular superior, la solució del qual es troba directament.

Escrivim el sistema lineal de partida com $\mathbf{A}^{(1)}\mathbf{x} = \mathbf{b}^{(1)}$, i anirem indicant els passos successius amb el superíndex. Suposem $a_{11}^{(1)} \neq 0$. El primer pas consisteix en passar a un sistema amb $a_{21}^{(2)} = \ldots = a_{n1}^{(2)} = 0$, sumant la primera equació, multiplicada per un factor adequat, a cadascuna de les altres $n-1$ equacions. Això ens porta a un sistema:

$$\begin{aligned} a_{11}^{(1)}x_1 + a_{12}^{(1)}x_2 + \cdots + a_{1n}^{(1)}x_n &= b_1^{(1)} \\ a_{22}^{(2)}x_2 + \cdots + a_{2n}^{(2)}x_n &= b_2^{(2)} \\ \cdots \qquad \cdots & \\ a_{n2}^{(2)}x_2 + \cdots + a_{nn}^{(2)}x_n &= b_n^{(2)} \end{aligned}$$

Suposem ara $a_{22}^{(2)} \neq 0$. El segon pas consisteix en passar a un sistema amb $a_{32}^{(3)} = \ldots = a_{n2}^{(3)} = 0$, sumant la segona equació, multiplicada per un factor adequat, a les de sota. Arribem així a:

$$\begin{aligned} a_{11}^{(1)}x_1 + a_{12}^{(1)}x_2 + a_{13}^{(1)}x_3 + \cdots + a_{1n}^{(1)}x_n &= b_1^{(1)} \\ a_{22}^{(2)}x_2 + a_{23}^{(2)}x_3 + \cdots + a_{2n}^{(2)}x_n &= b_2^{(2)} \\ a_{33}^{(3)}x_3 + \cdots + a_{3n}^{(3)}x_n &= b_3^{(3)} \\ \cdots \qquad \cdots & \\ a_{n3}^{(3)}x_3 + \cdots + a_{nn}^{(3)}x_n &= b_n^{(3)} \end{aligned}$$

Continuant d'aquesta manera, si pel camí no ens trobem amb algun $a_{kk}^{(k)} = 0$, arribem a un sistema triangular superior. Es pot demostrar que, si $n >> 1$, el nombre d'operacions necessàries per a arribar a un sistema triangular és aproximadament:

$$\frac{2n^3}{3} + \frac{n^2}{2}\,.$$

Nota. Quan el procés d'eliminació es fa amb paper i llapis, és pràctic representar el sistema per la matriu ampliada:

$$\tilde{\mathbf{A}} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} & b_1 \\ a_{21} & a_{22} & \cdots & a_{2n} & b_2 \\ \vdots & \vdots & & \vdots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} & b_n \end{bmatrix},$$

i operar amb les seves files com ho feiem més amunt amb les equacions, fins obtenir un triangle de zeros en la part inferior esquerra. En l'exemple que segueix es veurà això.

Exemple. Volem trobar la solució del sistema d'equacions:

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

Es comprova fàcilment que $|\mathbf{A}| = 36 \neq 0$, i per tant el sistema té solució única. Considerem la matriu ampliada:

$$\tilde{\mathbf{A}} = \begin{bmatrix} 2 & 4 & 1 & 1 \\ 8 & -1 & 3 & 0 \\ 2 & 5 & 0 & 0 \end{bmatrix},$$

i partint d'$\tilde{\mathbf{A}} = \tilde{\mathbf{A}}^{(1)}$, apliquem el procés d'eliminació gaussiana.

Primer pas: sumant a la segona fila la primera multiplicada per -4, i a la tercera la primera multiplicada per -1, obtenim:

$$\tilde{\mathbf{A}}^{(2)} = \begin{bmatrix} 2 & 4 & 1 & 1 \\ 0 & -17 & -1 & -4 \\ 0 & 1 & -1 & -1 \end{bmatrix}.$$

Segon pas: sumant a la tercera fila la segona multiplicada per $1/17$ obtenim:

$$\tilde{\mathbf{A}}^{(3)} = \begin{bmatrix} 2 & 4 & 1 & 1 \\ 0 & -17 & -1 & -4 \\ 0 & 0 & -18/17 & -21/17 \end{bmatrix},$$

que dóna un sistema triangular superior:

$$\begin{aligned} 2x_1 + 4x_2 + x_3 &= 1 \\ -17x_2 - x_3 &= -4 \\ -\frac{18}{17}\, x_3 &= -\frac{21}{17} \end{aligned}$$

La solució és:

$$\begin{aligned} x_3 &= \frac{-21/17}{-18/17} = \frac{7}{6}, \\ x_2 &= \frac{-1}{17}\,(-4 + x_3) = \frac{1}{6}, \\ x_1 &= \frac{1}{2}\,(1 - x_3 - 4x_2) = -\frac{5}{12}. \end{aligned}$$

4.1.4 Pivotatge

Notem que, en el pas $k-$è del procés d'eliminació gaussiana, hem de dividir per l'element de la diagonal $a_{kk}^{(k)}$, que s'anomena *pivot.* Convé fer algunes observacions:

- Es pot demostrar, excepte canvis de signe deguts a la permutació de files o columnes, que

  $$|\mathbf{A}| = a_{11}^{(1)}\, a_{22}^{(2)}\, \cdots a_{nn}^{(n)},$$

  que dóna un procediment per a calcular determinants.

- Si $a_{kk}^{(k)} = 0$, podem cercar un terme $a_{ik}^{(k)} \neq 0$ $(i > k)$ (que existeix sempre, en ser $|\mathbf{A}| \neq 0$) i permutar l'equació $k-$èna amb la $i-$èna, i a partir d'aquí continuar el procés.

- Si $a_{kk}^{(k)} \neq 0$ però és petit, el mètode de Gauss, encara que pugui aplicar-se, pot ser molt inestable numèricament. Així doncs, convindrà modificar el mètode de Gauss també en aquest cas. L'estratègia més senzilla per fer això és el *pivotatge maximal per columnes*, en el que al pas $k-$è es pren com a pivot $a_{kk}^{(k)}$ el coeficient de valor absolut més gran entre els ${a_{ik}}^{(k)}$ $(i = k, \ldots, n)$. Això comporta intercanviar equacions (o files de la matriu ampliada). En el *pivotatge complet*, el nou pivot passa a ser el coeficient de valor absolut més gran entre els $a_{ij}^{(k)}$ $(i, j = k, \ldots, n)$. Això ens obliga a intercanviar files i columnes de la matriu ampliada.

Exemple. Considerem el sistema:

$$\begin{bmatrix} -10^{-5} & 1 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix},$$

que té solució:

$$x_1 = \frac{-1}{2.00001} \simeq -0.4999975, \qquad x_2 = \frac{2}{2.00001} \simeq 0.999995.$$

Suposem que estem treballant amb quatre dígits, i apliquem les fórmules de l'eliminació gaussiana per obtenir la solució. El factor pel que hem de multiplicar la primera fila és:

$$\frac{0.2 \times 10^1}{0.1 \times 10^{-4}} = 0.2 \times 10^6,$$

i així:

$$a_{22}^{(2)} = 0.1 \times 10^1 + (0.2 \times 10^6)(0.1 \times 10^1) = 0.1 \times 10^1 + 0.2 \times 10^6 = 0.2 \times 10^6.$$

La suma exacta és 0.200001×10^6, però com treballem amb 4 dígits, aquest nombre queda representat com 0.2000×10^6. Calculem ara el terme independent:

$$b_2^{(2)} = (0.2 \times 10^6)(0.1 \times 10^1) = 0.2 \times 10^6.$$

Hem obtingut:

$$\begin{bmatrix} -10^{-5} & 1 \\ 0 & 0.2 \times 10^6 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 0.2 \times 10^6 \end{bmatrix}.$$

Resolent aquest sistema obtenim:

$$x_2 = \frac{0.2 \times 10^6}{0.2 \times 10^6} = 0.1 \times 10^1, \qquad x_1 = \frac{0.1 \times 10^1 - 0.1 \times 10^1}{-0.1 \times 10^4} = 0.$$

Apliquem ara el mètode del pivotatge maximal per columnes. Com $|a_{21}^{(1)}| > |a_{11}^{(1)}|$, intercanviem les dues equacions i obtenim el sistema:

$$\begin{bmatrix} 2 & 1 \\ -10^{-5} & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

El factor per a eliminar -10^{-5} és:

$$\frac{0.1 \times 10^{-4}}{0.2 \times 10^{1}} = 0.5 \times 10^{-5},$$

i així:

$$a_{22}^{(2)} = 0.1 \times 10^{1} + (0.5 \times 10^{-5})(0.1 \times 10^{1}) = 0.1 \times 10^{1},$$
$$b_{2}^{(2)} = 0.1 \times 10^{1} + (0.5 \times 10^{-5}) \times 0 = 0.1 \times 10^{1}.$$

La solució és ara:

$$x_2 = \frac{0.1 \times 10^{1}}{0.1 \times 10^{1}} = 1.0, \qquad x_1 = \frac{-(0.1 \times 10^{1} \times 0.1 \times 10^{1})}{0.2 \times 10^{1}} = -0.5,$$

que és acceptable.

4.1.5 Descomposició LU

Suposem que és possible descompondre la matriu **A** com a producte $\mathbf{A} = \mathbf{LU}$, on **L** és triangular inferior i **U** triangular superior. Llavors, la resolució del sistema $\mathbf{Ax} = \mathbf{b}$ és equivalent a la dels dos sistemes triangulars:

$$\mathbf{Ly} = \mathbf{b}, \qquad \mathbf{Ux} = \mathbf{y},$$

que es pot fer directament.

Una tal descomposició s'anomena descomposició LU. Es pot demostrar que, quan la matriu **A** és tal que es pot aplicar el procés d'eliminació gaussiana sense intercanviar ni files ni columnes, llavors existeix una descomposició LU. Una forma per obtenir la descomposició LU és aplicar, si es possible, el mètode d'eliminació gaussiana.

La descomposició LU no és única, ja que **A** té n^2 coeficients, i entre **L** i **U** en tenen n^2+n. En el *mètode de Doolittle* s'imposa que els termes de la diagonal d'**L** siguin iguals a 1. En el *mètode de Crout*, que els termes de la diagonal d'**U** siguin iguals a 1. Ambdós mètodes són equivalents, ja que al traspondre la descomposició de Doolittle d'**A** obtenim la descomposició de Crout d'$\mathbf{A}'$ (veure l'exemple que segueix).

Exemple. Suposem que volem factoritzar la matriu:

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

utilitzant el mètode de Doolittle. Aleshores tenim:

$$\begin{bmatrix} 1 & & & \\ l_{21} & 1 & & \\ l_{31} & l_{32} & 1 & \\ l_{41} & l_{42} & l_{43} & 1 \end{bmatrix} \begin{bmatrix} u_{11} & u_{12} & u_{13} & u_{14} \\ & u_{22} & u_{23} & u_{24} \\ & & u_{33} & u_{34} \\ & & & u_{44} \end{bmatrix} = \begin{bmatrix} 6 & 2 & 1 & -1 \\ 2 & 4 & 1 & 0 \\ 1 & 1 & 4 & -1 \\ -1 & 0 & -1 & 3 \end{bmatrix}.$$

Multiplicant les files d'**L** per la primera columna d'**U** tenim:

$$u_{11} = 6, \qquad l_{21}u_{11} = 2, \qquad l_{31}u_{11} = 1, \qquad l_{41}u_{11} = -1.$$

Repetint l'operació amb la segona columna d'**U**:

$$u_{12} = 2, \qquad l_{21}u_{12} + u_{22} = 4, \qquad l_{31}u_{12} + l_{32}u_{22} = 1, \qquad l_{41}u_{12} + l_{42}u_{22} = 0.$$

Repetint l'operació amb la tercera columna d'**U**:

$$u_{13} = 1, \qquad l_{21}u_{13} + u_{23} = 1,$$

$$l_{31}u_{13} + l_{32}u_{23} + u_{33} = 4, \qquad l_{41}u_{13} + l_{42}u_{23} + l_{43}u_{33} = -1.$$

Amb la quarta:

$$u_{14} = -1, \qquad l_{21}u_{14} + u_{24} = 0,$$

$$l_{31}u_{14} + l_{32}u_{24} + u_{34} = -1, \qquad l_{41}u_{14} + l_{42}u_{24} + l_{43}u_{34} + u_{44} = 3.$$

Resolent aquest sistema obtenim:

$$\mathbf{L} = \begin{bmatrix} 1 & & & \\ 1/3 & 1 & & \\ 1/6 & 1/5 & 1 & \\ -1/6 & 1/10 & -9/37 & 1 \end{bmatrix}, \qquad \mathbf{U} = \begin{bmatrix} 6 & 2 & 1 & -1 \\ & 10/3 & 2/3 & 1/3 \\ & & 37/10 & -9/10 \\ & & & 191/74 \end{bmatrix}.$$

Per trasposició tenim la descomposició de Crout d'$\mathbf{A}'$, però com **A** és simètrica, $\mathbf{A} = \mathbf{A}'$, i resulta:

$$\mathbf{A} = \begin{bmatrix} 6 & & & \\ 2 & 10/3 & & \\ 1 & 2/3 & 37/10 & \\ -1 & 1/3 & -9/10 & 191/74 \end{bmatrix} \begin{bmatrix} 1 & 1/3 & 1/6 & -1/6 \\ & 1 & 1/5 & 1/10 \\ & & 1 & -9/37 \\ & & & 1 \end{bmatrix}.$$

Nota. Els algorismes per fer la descomposició LU no sempre donen la descomposició de la matriu original **A**, encara que existeixi, sinó que donen la descomposició de la matriu resultant de permutar les files d'**A**. Tanmateix, de cara a resoldre un sistema d'equacions, permutar-les és irrelevant.

4.2 Norma i nombre de condició d'una matriu

4.2.1 Norma d'una matriu

En un espai vectorial, una *norma vectorial* és una aplicació que assigna a cada vector **x** un nombre real $\|\mathbf{x}\| \geq 0$, complint les propietats:

- $\|\mathbf{x}\| = 0$ si i només si $\mathbf{x} = \mathbf{0}$.
- $\|\alpha\mathbf{x}\| = |\alpha|\,\|\mathbf{x}\|$, per a tot vector $\mathbf{x}$ i tot escalar α.
- Desigualtat triangular: $\|\mathbf{x}+\mathbf{y}\| \le \|\mathbf{x}\| + \|\mathbf{y}\|$ per a tot parell de vectors $\mathbf{x}$, $\mathbf{y}$.

En l'espai dels vectors de dimensió n, $\mathbf{x} = (x_1, x_2, \ldots, x_n)'$, les normes vectorials més comuns són:

$$\begin{aligned} \|\mathbf{x}\|_1 &= \textstyle\sum_{i=1}^n |x_i|, \\ \|\mathbf{x}\|_2 &= \left(\textstyle\sum_{i=1}^n x_i^2\right)^{1/2} \qquad (\textit{norma euclidiana}), \\ \|\mathbf{x}\|_\infty &= \max_{1\le i\le n} |x_i| \qquad (\textit{norma del màxim}). \end{aligned}$$

En l'espai de les matrius $n \times n$ amb coeficients reals, una norma matricial és una norma vectorial que a més compleix:

$$\left\|\mathbf{AB}\right\| \le \|\mathbf{A}\|\,\|\mathbf{B}\|,$$

per a tot parell de matrius $\mathbf{A}$ i $\mathbf{B}$.

Donada una norma vectorial $\|\cdot\|$, se li pot associar una norma matricial, definint:

$$\|\mathbf{A}\| = \max\left\{\left\|\mathbf{Ax}\right\| :\ \left\|\mathbf{x}\right\| = 1\right\}.$$

Es pot veure que aquesta definició compleix totes les propietats de norma matricial i la propietat suplementària:

$$\left\|\mathbf{Ax}\right\| \le \left\|\mathbf{A}\right\|\left\|\mathbf{x}\right\|.$$

Per una matriu $\mathbf{A} = (a_{ij})$, es poden calcular les normes matricials associades a les normes vectorials definides abans, mitjançant:

$$\begin{aligned} \|\mathbf{A}\|_1 &= \max_{1\le j\le n}\left(\sum_{i=1}^n |a_{ij}|\right), \\ \|\mathbf{A}\|_2 &= \left(\rho\left(\mathbf{A}'\mathbf{A}\right)\right)^{1/2} \qquad (\textit{norma euclidiana}), \\ \|\mathbf{A}\|_\infty &= \max_{1\le i\le n}\left(\sum_{j=1}^n |a_{ij}|\right) \qquad (\textit{norma del màxim}), \end{aligned}$$

on $\rho(\mathbf{M})$ designa el *radi espectral*, màxim dels valors absoluts dels autovalors d'$\mathbf{M}$.

4.2.2 Nombre de condició

Considerem el sistema $\mathbf{Ax} = \mathbf{b}$, en el que pertorbem el terme independent $\mathbf{b}$ en una quantitat $\Delta\mathbf{b}$. La nova solució del sistema tindrà un error $\Delta\mathbf{x} = \mathbf{A}^{-1}\Delta\mathbf{b}$. Volem estimar l'error relatiu $\|\Delta\mathbf{x}\|/\|\mathbf{x}\|$. De les propietats de la norma matricial i de les identitats anteriors resulta:

$$\left\|\Delta\mathbf{x}\right\| \le \|\mathbf{A}^{-1}\|\,\|\Delta\mathbf{b}\|, \qquad \|\mathbf{b}\| \le \|\mathbf{A}\|\,\|\mathbf{x}\|.$$

Per tant:

$$\frac{\|\Delta\mathbf{x}\|}{\|\mathbf{x}\|} \leq \left\|\mathbf{A}\right\| \left\|\mathbf{A}^{-1}\right\| \frac{\|\Delta\mathbf{b}\|}{\|\mathbf{b}\|} \,.$$

Aquesta fórmula mostra que l'error relatiu en $\mathbf{x}$ és més petit o igual que el producte del factor:

$$\kappa(\mathbf{A}) = \left\|\mathbf{A}\right\| \left\|\mathbf{A}^{-1}\right\|$$

per l'error relatiu en $\mathbf{b}$. $\kappa(\mathbf{A})$ s'anomena nombre de condició de la matriu $\mathbf{A}$. Sempre $\kappa(\mathbf{A}) \geq 1$, ja que

$$\kappa(\mathbf{A}) \geq \left\|\mathbf{A}\mathbf{A}^{-1}\right\| = \left\|\mathbf{I}\right\| = 1.$$

Exemple. Volem trobar la solució del sistema $\mathbf{A}\mathbf{x} = \mathbf{b}$ amb

$$\mathbf{A} = \begin{bmatrix} 1/3 & 1/4 \\ 1/4 & 1/5 \end{bmatrix}, \qquad \mathbf{b} = \begin{bmatrix} 7/12 \\ 0.45 \end{bmatrix},$$

però hem calculat:

$$\bar{\mathbf{b}} = \begin{bmatrix} 0.583 \\ 0.45 \end{bmatrix}.$$

Llavors, tenim un error relatiu:

$$\frac{\|\Delta\mathbf{b}\|_\infty}{\|\mathbf{b}\|_\infty} = \frac{10^{-3}/3}{7/12} = \frac{4}{7} \times 10^{-3} = 0.0571\%.$$

Calculem $\mathbf{A}^{-1}$:

$$\mathbf{A}^{-1} = \begin{bmatrix} 48 & -60 \\ -60 & 80 \end{bmatrix}.$$

Aleshores el nombre de condició d'$\mathbf{A}$ és $\kappa_\infty(\mathbf{A}) \simeq 81.7$. El sistema $\mathbf{A}\mathbf{x} = \mathbf{b}$ té solució exacta $\mathbf{x} = (1, 1)'$. La solució del sistema pertorbat $\mathbf{A}\mathbf{x} = \bar{\mathbf{b}}$ és $\bar{\mathbf{x}} = (0.984, 1.020)'$, i l'error relatiu és:

$$\frac{\|\bar{\mathbf{x}} - \mathbf{x}\|_\infty}{\|\mathbf{x}\|_\infty} = 2\%.$$

D'altra banda, amb la fórmula que hem obtingut abans, una fita per a l'error relatiu és $81.7 \times 0.0571\% = 4.669\%$.

Una situació una mica és complicada és aquella es dóna quan hi ha també error en la matriu de coeficients $\mathbf{A}$. Aleshores:

$$\Big(\mathbf{A} + \Delta\mathbf{A}\Big)\Big(\mathbf{x} + \Delta\mathbf{x}\Big) = \mathbf{b},$$

d'on es dedueix:

$$\frac{\left\|\Delta\mathbf{x}\right\|}{\left\|\mathbf{x} + \Delta\mathbf{x}\right\|} \leq \kappa(\mathbf{A}) \, \frac{\|\Delta\mathbf{A}\|}{\|\mathbf{A}\|} \,.$$

És dir, si $\kappa(\mathbf{A})$ es gran, un error (relativament) petit en la matriu de coeficients dóna un error (relativament) gran de la solució, i es diu que el problema està mal condicionat.

En el cas general on pertorbem simultàniament la matriu de coeficients i el terme independent:

$$\Big(\mathbf{A}+\Delta\mathbf{A}\Big)\Big(\mathbf{x}+\Delta\mathbf{x}\Big)=\mathbf{b}+\Delta\mathbf{b},$$

podem arribar a:

$$\frac{\|\Delta\mathbf{x}\|}{\|\mathbf{x}\|}\leq\kappa(\mathbf{A})\left(\frac{\|\Delta\mathbf{b}\|}{\|\mathbf{b}\|}+\frac{\|\Delta\mathbf{A}\|}{\|\mathbf{A}\|}\,\frac{\|\mathbf{x}+\Delta\mathbf{x}\|}{\|\mathbf{x}\|}\right),$$

que és una generalització de les fórmules anteriors, que mostra que el nombre de condició actua de factor d'amplificació entre els errors relatius de les dades i el del resultat.

4.3 Sistemes sobredeterminats

Considerem ara un sistema sobredeterminat, es dir un sistema lineal $\mathbf{Ax}=\mathbf{b}$, on $\mathbf{A}$ és una matriu $n\times p$, de rang p (columnes linealment independents), amb $p<n$. Aquest sistema no té solució, i es diu que és *incompatible*. Tanmateixsi canviem el problema en el sentit de que no es demani que $\mathbf{Ax}$ sigui "exactament" igual a $\mathbf{b}$ sinò el més a prop possible, tenim un problema que té sentit, i que podem formular amb precisió de la forma següent: trobar $\mathbf{x}$ tal que el vector $\mathbf{e}=\mathbf{b}-\mathbf{Ax}$ tingui norma mínima.

Aquest problema té diverses variants, segons la norma que utilitzem. La més habitual es basa en la norma euclidiana $\|\cdot\|_2$. Per a aquesta norma la solució es pot obtenir per un mètode directe anomenat *mètode dels mínims quadrats*. En aquest es tractarà de trobar $x_1,\ldots,x_p$ tals que la suma de quadrats:

$$\sum_{i=1}^{n}e_i^2=\sum_{i=1}^{n}\Big(b_i-a_{i1}x_1-\cdots-a_{ip}x_p\Big)^2$$

sigui mínima. El mètode es pot presentar de diverses maneres, però formulat en llenguatge geomètric es redueix a una fórmula matricial bastant senzilla.

Observem que el vector $\mathbf{Ax}$ es pot escriure com una combinació lineal dels vectors columna d'$\mathbf{A}$:

$$\mathbf{Ax}=x_1\begin{bmatrix}a_{11}\\ \vdots\\ a_{n1}\end{bmatrix}+\cdots+x_p\begin{bmatrix}a_{1p}\\ \vdots\\ a_{np}\end{bmatrix},$$

i per tant el problema es pot reformular de la manera següent: es tracta de trobar la combinació lineal dels vectors columna més pròxima a $\mathbf{b}$. Peró per a la norma euclidiana la distància més curta sempre la dóna la perpendicular, i per tant la solució serà la projecció ortogonal de $\mathbf{b}$ sobre el subespai generat pels vectors columna d'$\mathbf{A}$. Això és equivalent a que $\mathbf{b}-\mathbf{Ax}$ sigui ortogonal a

tots els vectors columna. Que dos vectors siguin ortogonals equival a que el seu producte escalar sigui zero, i podem ajuntar tots aquests productes en un producte matricial i arribem a:

$$\mathbf{A}'\Big(\mathbf{b}-\mathbf{A}\mathbf{x}\Big)=\mathbf{0},$$

que és un sistema de p equacions amb p incògnites. La solució es pot trobat de diverses maneres, algunes de les quals han estat comentades en aquest capítol. Una manera elegant d'expressar-la és la fórmula matricial:

$$\mathbf{x}=\Big(\mathbf{A}'\mathbf{A}\Big)^{-1}\mathbf{A}'\mathbf{b}.$$

Exemple. Resoldrem per mínims quadrats el sistema sobredeterminat:

$$\begin{aligned} 3x_1+2x_2 &= 0\\ 2x_1+x_2 &= 1\\ x_1-x_2 &= -1. \end{aligned}$$

En aquest cas:

$$\mathbf{A}=\begin{bmatrix}3&2\\2&1\\1&-1\end{bmatrix},\qquad \mathbf{A}'=\begin{bmatrix}3&2&1\\2&1&-1\end{bmatrix},\qquad \mathbf{b}=\begin{bmatrix}0\\1\\-1\end{bmatrix}.$$

Aleshores:

$$\mathbf{A}'\mathbf{A}=\begin{bmatrix}14&7\\7&6\end{bmatrix},\qquad \Big(\mathbf{A}'\mathbf{A}\Big)^{-1}=\frac{1}{35}\begin{bmatrix}6&-7\\-7&14\end{bmatrix}.$$

Doncs:

$$\mathbf{x}=\frac{1}{35}\begin{bmatrix}6&-7\\-7&14\end{bmatrix}\begin{bmatrix}3&2&1\\2&1&-1\end{bmatrix}\begin{bmatrix}0\\1\\-1\end{bmatrix}=\frac{1}{35}\begin{bmatrix}-8\\21\end{bmatrix}.$$

Aquesta solució és òptima en el sentit que el vector:

$$\mathbf{A}\mathbf{x}=\frac{1}{35}\begin{bmatrix}18\\5\\-29\end{bmatrix}$$

és el més pròxim possible a $\mathbf{b}$. Això, en el nostre cas vol dir:

$$\Big\|\mathbf{e}\Big\|=\Big\|\mathbf{b}-\mathbf{A}\mathbf{x}\Big\|=1.0141.$$

4.4 Regressió lineal

4.4.1 Plantejament general

Ens ocuparem ara d'un cas particular del problema anterior. Volem expressar una variable y com a combinació lineal de p variables x_1, x_2, ... , x_p:

$$y = b_1\, x_1 + \cdots + b_p\, x_p.$$

y s'anomena *variable resposta* (també variable dependent), i les x, *variables de control*, *predictors* o *factors* (també variables independents, el que és una denominació enganyosa, ja que pot haber dependència entre elles). La fórmula que dóna y en termes de les x_j és el *model lineal*, i els b_j són els *paràmetres del model*. Les dades expe-rimentals que es puguin obtenir de les variables implicades en el model no el satisfan exactament, i el màxim que es pot fer és trobar els valors dels paràmetres per als quals la diferència entre la variable resposta i la predicció que es fa utilitzant el model (la combinació lineal de les x_j) sigui mínima. Els mètodes per trobar els paràmetres i per a comparar entre els models alternatius s'anomenen genèricament *mètodes de regressió lineal*.

Per a precisar més, s'acostuma a designar per $\hat{y}$ la predicció que es fa amb el model:

$$\hat{y} = b_1\, x_1 + \cdots + b_p\, x_p,$$

i per e l'error de la predicció: $e = y - \hat{y}$. Aleshores, el que fa el mètode de regressió lineal es donar els valors dels paràmetres que fan mínim l'error e. Es veu fàcilment que, per a fer això a partir d'un conjunt de dades experimentals on hi hagi més dades que paràmetres, haurem de resoldre un sistema sobredeterminat. Tanmateix, la notació ha canviat, perquè les b_j són les incògnites, els valors d'y donaran el vector terme independent i els de les x_j la matriu de coeficients.

4.4.2 Ajust per mínims quadrats

Veurem ara com es pot obtenir un model lineal que doni y en termes de les x_j, a partir d'un conjunt de dades experimentals. Suposem que s'han realitzat n proves, i en cadascuna d'elles s'han obtingut els valors de la variables resposta y i de les variables de control x_1, ... , x_p. Designem per y_i el valor d'y obtingut en la prova i-èna, i per x_{ij} el d'x_j. Una operació per la qual obtenim valors b_1, ... , b_p de manera que els residus:

$$e_i = y_i - b_1\, x_{i1} - \cdots - b_p\, x_{pi}$$

siguin petits, s'anomena genèricament *ajust del model* $y = b_1\, x_1 + \cdots + b_p\, x_p$ *a les dades*. Qualsevol mètode d'ajust d'un model lineal implica resoldre (en el sentit d'obtenir residus mínims)

el sistema lineal $\mathbf{y} = \mathbf{Xb} + \mathbf{e}$, on $\mathbf{y}$ és el vector d'observacions d'y, $\mathbf{X}$ la matriu d'observacions de les x_j, $\mathbf{b}$ el vector de paràmetres, i $\mathbf{e}$ el de residus:

$$\mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \qquad \mathbf{X} = \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{np} \end{bmatrix}, \qquad \mathbf{b} = \begin{bmatrix} b_1 \\ \vdots \\ b_p \end{bmatrix}, \qquad \mathbf{e} = \begin{bmatrix} e_1 \\ \vdots \\ e_n \end{bmatrix}.$$

En la práctica l'ajust només té sentit si $n > p$, i aleshores equival a la resolució d'un sistema lineal sobredeterminat. En l'*ajust per mínims quadrats*, es troba $\mathbf{b}$ de manera que $\|\mathbf{e}\|_2$ sigui mínima. Fent un canvi de notació en la fórmula de la secció anterior tenim la solució:

$$\mathbf{b} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y}.$$

Exemple. Reescriurem l'exemple de la secció anterior, per tal de veure que la regressió no és de fet només un cas particular de la resolució de sistemes sobredeterminats, sinó que un sistema sobredeterminat es pot plantejar com un problema d'ajust. Ajustem per mínims quadrats un model $y = b_1x_1 + b_2x_2$ a les dades de la Taula 5.1 (en un espai de dimensió tres, això representa ajustar un pla que passa per l'origen a tres punts).

TAULA 5.1

y	x_1	x_2
0	3	2
1	2	1
-1	1	-1

Tenim ara:

$$\mathbf{y} = \begin{bmatrix} 0 \\ 1 \\ -1 \end{bmatrix}, \qquad \mathbf{X} = \begin{bmatrix} 3 & 2 \\ 2 & 1 \\ 1 & -1 \end{bmatrix},$$

i l'ajust per mínims quadrats ens dóna el model:

$$y = \frac{-8}{35}\,x_1 + \frac{21}{35}\,x_2.$$

4.4.3 Recta de regressió

El cas més classic i més conegut de regressió lineal és aquell en el que es té una variable resposta (dependent) y i una variable de control (independent) x i es vol ajustar a un conjunt de dades (o a un conjunt de punts) un model del tipus $y = a + bx$, és a dir, l'equació d'una recta. Aquest model s'anomena recta de regressió d'y sobre x. És típic utilitzar la lletra a per al *terme constant* i b per al *pendent* de la recta. Qualsevol calculadora científica disposa de rutines per calcular

aquests coeficients. Per diferenciar aquest cas del cas on hi ha més variables de control es parla de *regressió lineal simple*, en contraposició a la *regressió lineal múltiple*.

La recta de regressió és un cas particular del que hem vist abans, on la matriu de dades és:

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}.$$

Tanmateix, en aquest cas es pot donar fórmules senzilles per a calcular a i b, sense utilitzar les fórmules matricials:

$$b = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}, \qquad a = \bar{y} - b\bar{x},$$

on $\bar{x}$, $\bar{y}$ designen les respectives mitjanes. És interessant remarcar que aquesta última fórmula mostra que la recta de regressió passa per $(\bar{x}, \bar{y})$. Equivalentment, que la mitjana dels residus és zero. Això és cert, en general, per a qualsevol model lineal amb terme constant, és dir, si una de les variables de control és la constant 1.

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

Les mitjanes són $\bar{x} = 5$, $\bar{y} = 0.359$. Els paràmetres a i b poden calcular-se amb les fórmules de més amunt (o amb una calculadora científica), o amb la fórmula matricial :

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 2 & 4 & 6 & 8 & 10 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 1 & 2 \\ 1 & 4 \\ 1 & 6 \\ 1 & 8 \\ 1 & 10 \end{bmatrix} = \begin{bmatrix} 6 & 30 \\ 30 & 220 \end{bmatrix},$$

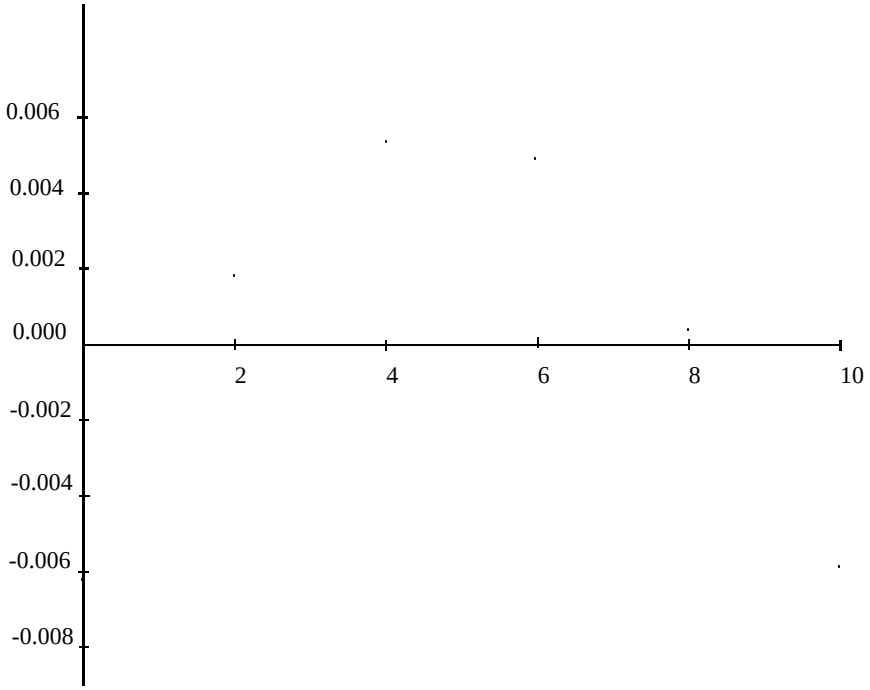

Figura 4.1

$$\begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} 6 & 30 \\ 30 & 220 \end{bmatrix}^{-1} \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 2 & 4 & 6 & 8 & 10 \end{bmatrix} \begin{bmatrix} 0.002 \\ 0.150 \\ 0.294 \\ 0.434 \\ 0.570 \\ 0.704 \end{bmatrix} = \begin{bmatrix} 0.00828 \\ 0.07014 \end{bmatrix}.$$

Els residus s'obtenen a partir de la fórmula $\mathbf{e} = \mathbf{y} - \mathbf{Xb}$:

$$\mathbf{e} = \begin{bmatrix} 0.002 \\ 0.150 \\ 0.294 \\ 0.434 \\ 0.570 \\ 0.704 \end{bmatrix} - \begin{bmatrix} 1 & 0 \\ 1 & 2 \\ 1 & 4 \\ 1 & 6 \\ 1 & 8 \\ 1 & 10 \end{bmatrix} \begin{bmatrix} 0.00828 \\ 0.07014 \end{bmatrix} = \begin{bmatrix} -0.00628 \\ 0.00143 \\ 0.00514 \\ 0.00486 \\ 0.00057 \\ -0.00571 \end{bmatrix}.$$

Podem analitzar el residus d'un ajust per a comprovar si el model ha estat ben triat. És útil representar els residus en un gràfic bidimensional, contra x (veure Figura 5.1). En aquest cas, els punts semblen descriure una corba, fet que suggereix que la relació entre y i x no és lineal.

4.4.4 Recta de regressió sense terme constant

Hi ha una variant de la regressió lineal, en la que s'ajusta un model del tipus $y = bx$ (es dir, una recta per l'origen). En aquest cas tenim:

$$\mathbf{X} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}.$$

Com en el cas de la recta amb terme independent, es pot donar una fórmula directa per al paràmetre b:

$$b = \frac{\sum x_i\, y_i}{\sum x_i^2}\,.$$

Té sentit considerar la recta sense terme constant com un model alternatiu quan al aplicar les fórmules de l'apartat anterior s'obté un valor del paràmetre a molt proper a zero. Cal fer algunes observacions sobre la regressió sense terme constant:

- En general, els pendents (valors de b) obtinguts per ambdós procediments de regressió lineal simple no són iguals (amb dades reals no ho seran mai). Dit d'una altra forma, el model $y = bx$ no s'obté fent $a = 0$ en el model $y = a + bx$.

- La recta de regressió $y = bx$ no passa pel punt $(\bar{x}, \bar{y})$, o, equivalentment, la mitjana dels residus no és zero.

Exemple. En l'exemple anterior hem obtingut $a = 0.00828$. D'altra banda, l'absorbància amb concentració zero és molt petita, i pot ser interessant utilitzar un model $y = bx$. Aplicant la fórmula de més amunt, obtenim $b = 0.071273$, que és un valor lleugerament diferent al de la recta amb terme constant.

4.4.5 Regressió polinòmica

Fent:

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 \end{bmatrix},$$

podem obtenir els paràmetres d'un model quadràtric:

$$y = b_0 + b_1\, x + b_2\, x^2.$$

El procediment s'extén sense dificultat qualsevol polinomi, d'una o diverses variables, amb o sense terme constant.

Exemple. Tornant a l'exemple d'abans, ajustem per mínims quadrats un polinomi de segon grau. Ara:

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 2 & 4 & 6 & 8 & 10 \\ 0 & 4 & 16 & 36 & 64 & 100 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 1 & 2 & 4 \\ 1 & 4 & 16 \\ 1 & 6 & 36 \\ 1 & 8 & 64 \\ 1 & 10 & 100 \end{bmatrix} = \begin{bmatrix} 6 & 30 & 220 \\ 30 & 220 & 1800 \\ 220 & 1800 & 15664 \end{bmatrix},$$

$$\mathbf{b} = \begin{bmatrix} 6 & 30 & 220 \\ 30 & 220 & 1800 \\ 220 & 1800 & 1564 \end{bmatrix}^{-1} \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 2 & 4 & 6 & 8 & 10 \\ 0 & 4 & 16 & 36 & 64 & 100 \end{bmatrix} \begin{bmatrix} 0.002 \\ 0.150 \\ 0.294 \\ 0.434 \\ 0.570 \\ 0.704 \end{bmatrix} = \begin{bmatrix} 0.00221 \\ 0.07496 \\ -0.00046 \end{bmatrix},$$

$$\mathbf{e} = \begin{bmatrix} 0.002 \\ 0.150 \\ 0.294 \\ 0.434 \\ 0.570 \\ 0.704 \end{bmatrix} - \begin{bmatrix} 1 & 0 & 0 \\ 1 & 2 & 4 \\ 1 & 4 & 16 \\ 1 & 6 & 36 \\ 1 & 8 & 64 \\ 1 & 10 & 100 \end{bmatrix} \begin{bmatrix} 0.00221 \\ 0.07496 \\ -0.00046 \end{bmatrix} = \begin{bmatrix} -0.00021 \\ -0.00021 \\ 0.00029 \\ 0.00000 \\ -0.00064 \\ 0.00036 \end{bmatrix}.$$

4.4.6 Transformacions

En certes ocasions és possible transformar les variables x i y en altres variables de forma que la relació entre les variables transformades esdevingui lineal. Vegem alguns exemples:

- El model exponencial:

$$y = a\, e^{bx}$$

és equivalent al model lineal:

$$\ln y = \ln a + b\, x.$$

- El model potencial:

$$y = a\, x^{b}$$

és equivalent al model lineal:

$$\ln y = \ln a + b\, \ln x.$$

En aquestes situacions, es pot trobar els paràmetres del model transformat, i transfor-mar-los si cal per obtenir estimacions dels paràmetres del model original.

4.5 Problemes

73. Useu el mètode de Gauss per resoldre el sistema:

$$\left.\begin{array}{rcr} x_1 - x_2 + 2x_3 - x_4 &=& -8 \\ 2x_1 - 2x_2 + 3x_3 - 3x_4 &=& -20 \\ x_1 + x_2 + x_3 &=& -2 \\ x_1 - x_2 + 4x_3 + 3x_4 &=& 4 \end{array}\right\}.$$

Resposta: $x_1 = -7, x_2 = 3, x_3 = 2, x_4 = 2$.

74. Utilitzeu el mètode de Gauss i el mètode de Gauss amb pivotatge per resoldre el sistema:

$$\left.\begin{array}{rcr} 0.003000x_1 + 59.14x_2 &=& 59.17 \\ 5.291x_1 - 6.130x_2 &=& 46.78 \end{array}\right\}.$$

Resposta: Usant Gauss $x_1 = -10.00, x_2 = 1.001$. Usant Gauss amb pivotatge $x_1 = 10.00, x_2 = 1.000$.

75. Resoleu usant el mètode de Doolittle i el mètode de Gauss el sistema:

$$\left.\begin{array}{rcr} x_1 + 2x_2 + 3x_3 &=& 14 \\ 2x_1 + 5x_2 + 2x_3 &=& 18 \\ 3x_1 + x_2 + 5x_3 &=& 20 \end{array}\right\}.$$

Resposta: $x_1 = 1, x_2 = 2, x_3 = 3$.

76. Calculeu la inversa de la matriu

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

Indicació: Si anomenem $X = (x_{ij})$ a la matriu A^{-1}, per obtenir $X = A^{-1}$ cal resoldre $AX = I$ on I és la matriu identitat. És a dir, resoldrem els tres sistemes d'equacions donats per

$$\begin{pmatrix} 1 & 2 & -1 \\ 2 & 1 & 0 \\ -1 & 1 & 2 \end{pmatrix} \begin{pmatrix} x_{11} & x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \\ x_{31} & x_{32} & x_{33} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Resposta:

$$A^{-1} = \frac{1}{9} \begin{pmatrix} -2 & 5 & -1 \\ 4 & -1 & 2 \\ -3 & 3 & 3 \end{pmatrix}.$$

77. Apliqueu el mètode d'eliminació gaussiana per calcular el determinant de la matriu:

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

Resposta: $\det A = -39$.

78. Determineu el valor de les normes matricials $||A||_\infty$, $||A||_1$, $||A||_2$ de la matriu

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

Resposta: 4, 4, 3.106.

79. Ajusteu una recta de mínims quadrats a la següent taula de dades:

x	1	3	4	6	8	9	11	14
y	1	2	4	4	5	7	8	9

.

Resposta: $\mathrm{y} = 0.545 + 0.636\mathrm{x}$

80. Sabem que els pesos atòmics de l'oxigen i del nitrogen són aproximadament $O = 16$ i $N = 14$; utilitzeu els pesos moleculars dels sis òxids de nitrogen donats a continuació per tal d'ajustar-los per mínims quadrats lineals

Compost	NO	N_2O	NO_2	N_2O_3	N_2O_5	N_2O_4
Pes molecular	30.006	44.013	46.006	76.012	108.010	92.011

.

81. Els processos termodinàmics adiabàtics de sistemes físics caracteritzats per la pressió P, el volum V i la temperatura T (els gasos, per exemple) segueixen una llei del tipus $PV^\gamma = C$, on C és constant al llarg del procés. Ajusteu per mínims quadrats els valors de C i de γ en un procés adiabàtic segons la taula de mesures experimentals següent:

P (atm)	1.62	1.00	0.75	0.62	0.52	0.46
V (litres)	0.5	1.0	1.5	2.0	2.5	3.0

.

Resposta: $C = 0.998$, $\gamma = 0.703$

82. Hom suposa que el cometa Tentax, descobert el 1968, és un objecte del Sistema Solar. En un cert sistema de coordenades polars (r, φ), centrat en el Sol, s'han mesurat experimentalment les següents posicions del cometa,

r	2.70	2.00	1.61	1.20	1.02
φ	48°	67°	83°	108°	126°

.

Les lleis de Kepler garanteixen que el cometa es mourà en una òrbita el·líptica, parabòlica o hiperbòlica (si es menyspreen les pertorbacions dels planetes), que en les dites coordenades polars tindrà per equació

$$r = \frac{p}{1 - e\cos\varphi} ,$$

on p és un paràmetre i e l'excentricitat. Ajusteu per mínims quadrats els valors de p i e, a partir de les mesures fetes. *Resposta:* $p = 1.454$, $e = 0.694$

83. Empreu una tècnica de mínims quadrats per ajustar la taula de dades:

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

a funcions dels tipus següents:

i) $y = a + bx$, ii) $y = a + bx + cx^2$, iii) $y = Ax^{\alpha}$, iv) $y = Be^{\beta x}$.

Quin d'aquests tipus sembla el més adequat?

Resposta: i) $a = 0.068$, $b = 1.126$, ii) $a = -5.7\,10^{-3}$, $b = 1.324$, $c = -9.90\,10^{-2}$, iii) $\alpha = 0.896$, $A = 1.199$, iv) $\beta = 1.119$, $B = 0.337$

84. Les dades de la taula han estat obtingudes en un experiment de laboratori adreçat a investigar el canvi en el rendiment (%) d'un colorant al canviar la temperatura de reacció. L'experimentador creia que el rendiment assoliria un màxim dins d'aquest interval de temperatures, i que la relació entre la temperatura i el rendiment podria ser aproximada per una funció quadràtica del tipus:

$$y = b_0 + b_1 x + b_2 x^2 + e,$$

on $x =$(temperatura-60) i y és el rendiment.

$Temperatura$	x	y
56	-4	45.9
60	0	79.8
61	1	78.9
63	3	77.1
65	5	62.6

Estimar els coeficients b_0, b_1 i b_2 per mínims quadrats.

Font: G.E.P. Box & N.R. Draper (1986), *Empirical Model Building and Response Surfaces*, Wiley.

Resposta. $b_0 = 78.615$, $b_1 = 3.106$ i $b_2 = -1.263$.

85. Es realitza una investigació per tal d'explorar les condicions de reacció en les quals s'obté un polímer amb elasticitat màxima. Les variables amb influència rellevant sobre l'elasticitat han estat reduïdes a tres en una faseanterior de la investigació: les concentracions (%) C_1 i C_2 de dos

components, i la temperatura (°C) de reacció T. En un experiment per avançar més en aquesta investigació s'han obtingut les dades de la taula.

C_1	C_2	T	$Elasticitat$
15	2.3	135	25.74
21	2.3	135	48.94
15	3.1	135	42.78
21	3.1	135	35.94
15	2.3	155	41.50
21	2.3	155	50.10
15	3.1	155	46.06
21	3.1	155	27.70

Obtenir un model lineal del tipus

$$y = b_0 + b_1\, C_1 + b_2\, C_2 + b_3\, T,$$

on y és l'elasticitat, per mínims quadrats.

Font: G.E.P. Box & N.R. Draper (1986), *Empirical Model Building and Response Surfaces*, Wiley.

4.6 Qüestions

86. Treballant amb aritmètica de 4 dígits i tallant, utilitzant el mètode d'eliminació gaussiana sense pivotatge, el sistema

$$A = \left(\begin{array}{cc|c} 1.0001 & 1.5 & 0 \\ 2 & 3 & 1 \end{array} \right),$$

té per solució

a) $x = -0.4994 \times 10^3$ i $y = 0.3333 \times 10^3$, b) $x = -500.0$ i $y = 333.6$,
c) $x = -333.3$ i $y = 500.0$, d) Cap de les anteriors.

87. Volem calcular, usant el mètode d'eliminació gaussiana amb pivotatge, el determinant de la matriu

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

Quantes vegades cal fer pivotatge complet?

a) 1, b) 2, c) 1 o 2, depenent de quin sigui el primer pivot, d) Cap de les anteriors.

88. Fixades dues matrius A i U, és certa alguna de les següents afirmacions:

a) $\|A^\top\|_2 = \|A\|_2^2 = \|A^\top\|_2^2$

b) Si $U^\top = U^{-1}$ aleshores $\|U^\top A U\|_2 = \|A\|_2$

c) Els apartats a) i b) són certs.

d) Cap dels anteriors apartats és cert.

89. Quina és la matriu U quan fem descomposició LU a la matriu:

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

a) $U = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 2 & -2 \\ 0 & 0 & 1 \end{pmatrix}$, b) $U = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 5/3 \end{pmatrix}$,
c) $U = \begin{pmatrix} 1 & 1 & 0 \\ 0 & -2 & 2 \\ 0 & 0 & 6 \end{pmatrix}$, d) Cap de les anteriors.

90. Donada la matriu:

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

Quant val $||A||_1$?

a) 4, b) 3, c) 2, d) 1.

91. Després de fer eliminació gaussiana d'una matriu 6x6 obtenim una matriu triangular superior que te la següent diagonal (-1,1,2,-2,3,-3). Durant el procés hem hagut d'intercanviar files 3 cops. Quin és el determinant de la matriu?

a) -12, b) -36, c) 36, d) Cap de les anteriors

92. Si $0 < \epsilon < 1$, quant val el nombre de condició de la matriu A utilitzant la norma $\|.\|_\infty$

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

a) ϵ, b) $1/\epsilon$, c) $(1+\epsilon)/\epsilon$, d) $(1+\epsilon)$

93. Considerem la matriu

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

Quant val $\|A\|_\infty$?

a) 6, b) 4, c) 7, d) 10.

94. Volem calcular mitjançant el mètode d'eliminació gaussiana el determinant de

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

Quantes vegades cal fer pivotatge complet (no pivotatge maximal per columnes)?

a) 0, b) 1, c) 3, d) 2.

95. En resoldre un sistema d'equacions, $Ax = b$, es prefereix un número de condició per la matriu A:

a) ni gran ni petit, b) negatiu, c) petit, d) gran.

96. Considerem la matriu

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

Quant valen les $\|A\|_\infty$ i $\|A\|_1$, respectivament?

a) 5 i 4, b) 2 i 2, c) 3 i 1, d) Cap dels valors anteriors.

97. Es defineix la matriu de Hilbert $A = (a_{ij})$ mitjançant $a_{ij} = \dfrac{1}{i+j-1}$. Sabem que si la matriu és 4×4 aleshores $\|A^{-1}\|_\infty \simeq 13620$. Què podem dir en aquest cas de la matriu A ?

a) $\kappa_\infty(A)$ és gran i la matriu A està mal condicionada.
b) $\kappa_\infty(A)$ és gran i la matriu A no està mal condicionada.
c) La matriu A té molts zeros.
d) A no és regular.

98. Per calcular mitjançant el mètode d'eliminació gaussiana el determinant de

$$A = \begin{pmatrix} 2 & -1 & 3 \\ 4 & -2 & 7 \\ 6 & -6 & 8 \end{pmatrix},$$

quantes vegades és necessari fer pivotatge maximal per columnes ?

a) 3, b) 1, c) 0, d) 2.

99. Quant val L en la descomposició LU de la matriu

$$A = \begin{pmatrix} 1 & 3 & 2 \\ 2 & 7 & 3 \\ 3 & 2 & 7 \end{pmatrix}$$

a) $L = \begin{pmatrix} 1 & 0 & 0 \\ -3 & 1 & 0 \\ 2 & -7 & 1 \end{pmatrix}$ b) $L = \begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 3 & -7 & 1 \end{pmatrix}$

c) $L = \begin{pmatrix} 1 & 0 & 0 \\ 5 & 1 & 0 \\ 4 & 3 & 1 \end{pmatrix}$ d) Cap dels anteriors.

100. Quant valen, respectivament, $\|A\|_\infty$ i $\|A\|_1$ si

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

a) -7 i -5, b) 4 i 2, c) 5 i 7, d) 7 i 5.

101. La recta de mínims quadrats que ajusta la següent taula de dades és

x	3	5	6	8	9	11
y	2	3	4	6	5	8

a) $y = \frac{-1}{3} + \frac{5}{7}x$, b) $x = 1 + \frac{9}{7}y$, c) $y = x$, d) Cap de les anteriors.

102. Al ajustar per minims quadrats fent servir com a funcions base $(1, 1/x)$ la següent taula de valors

x	1.00	1.50	2.00	3.00	4.00	5.00
y	1.60	1.70	2.00	2.30	2.40	2.50

obtenim la funció:

a) $12.5 + 3/x$, b) $1 + 5.6/x$, c) $1 + 2.3/x$, d) $2.67 - 1.19/x$

103. Trobar la recta que aproxima en el sentit del mínims quadrats, la funció e^x sobre la xarxa de punts $(-1, -0.5, 0, 0.5, 1)$

a) $y = 3.25 + 7.41x$ b) $y = 1.26828 + 1.1486x$
c) $y = 1.376 + 1.576x$ d) $y = 1.126 + 2.576x$

104. Quina és la paràbola que millor ajusta per mínims quadrats la següent taula de dades ?

x	0	1	2	3
y	13	-33	43	1

a) $y = 1 + x + x^2$, b) $y = -6 - 6x + 6x^2$,
c) $y = 6 + 6x - 6x^2$, d) $y = -1 - x - x^2$.

105. Volem ajustar una fórmula del tipus $f(x) = Ae^{Bx}$ a les següents dades, per mínims quadrats

x_k	1	2	3	4
$f(x_i)$	7	11	17	27

Quant valen aproximadament A i B. Nota: Cal tenir en compte les següents dades log (7) = 1.95, log (11) = 2.40, log (17) = 2.83, log (27) = 3.30.

a) $A = 0.45, B = 4.40$, b) $A = 0.45, B = 1.5$,
c) $A = 4.48, B = 0.45$, d) $A = 1.5, B = 0.45$

106. Quina és la recta de mínims quadrats per aquesta taula

x	1	2	3	4	5	6
y	1.1	2	2.9	3.8	5.2	6

a) $y = 0.01 + 1.2x$, b) $y = x$, c) $y = 0.3 + 1.02x$, d) Cap de les anteriors.

107. Quina és la recta que millor aproxima, en el sentit dels mínims quadrats, la família de punts (1,3) i (5,2) ?

a) La mediatriu del segment que determinen els dos punts.

b) No hi ha cap recta que ajusti per mínims quadrats aquests dos punts.

c) $y = -\frac{1}{4}x + \frac{13}{4}$

d) Hi ha infinites rectes que ajusten per mínims quadrats aquests dos punts.

108. Quina és l'aproximació per mínims quadrats de la taula

x	1	2	3	4
y	7	11	17	27

mitjançant una funció de la forma $y = a \cdot e^{bx}$? (Nota: treballeu amb dues xifres decimals arrodonides)

a) $y = 4.48e^{0.45x}$, b) $y = 10.48e^{28.45x}$, c) $y = 0.14e^{6.71x}$, d) Cap dels anteriors.

109. Donada la taula

x	-1	0	1	2
y	-1.9	-1	0.3	1.1

ajusteu $y(x) = a + bx$ pel mètode dels mínims quadrats. És correcte suposar que x, y segueixen una relació lineal com la trobada ?

a) Si, b) No, c) Mai, d) Sempre.

Capítol 5

EQUACIONS DIFERENCIALS

5.1 Introducció

Sigui $f : [a, b] \times \mathbb{R} \to \mathbb{R}$ i un punt $\eta \in \mathbb{R}$. El problema de Cauchy consisteix en trobar una solució $y(t)$ derivable en $[a, b]$ de la equació: $x'(t) = f(t, x(t))$ amb la condició inicial $x(a) = \eta$. El següent resultat que no demostrem és el que ens assegura la existència i unicitat de la solució:

Teorema. *Si f és una funció continua en les dues variables i derivable en la segona variable amb derivada acotada en tota la banda $[a, b] \times \mathbb{R}$ aleshores el problema de Cauchy té una solució única.*

Nota: La condició de derivabilitat no és la millor possible, però hem de imposar alguna condició de regularitat (apart de la continuïtat) per assegurar-nos que el problema de Cauchy té solució única com mostra el següent exemple: $x'(t) = 2\sqrt{x}$, $x(0) = 0$ té dues solucions en l'interval $[0, 1]$: $x(t) = t^2$ i $x(t) \equiv 0$.

5.2 Mètodes numèrics

En general, no és possible trobar una solució explícita al problema de Cauchy. El nostre objectiu serà trobar solucions aproximades de forma numèrica. Un primer mètode molt intuïtiu (però no molt eficaç), és:

5.2.1 Mètode d'Euler

Fem una partició de l'interval $[a, b]$ en N intervals de llargària $h = (b - a)/N$ mitjançant els punts $t_i = a + ih$, $i = 0, \ldots, N$. Partint del punt $t_0 = a$ coneixem el valor de x en aquest punt $x_0 = x(t_0) = \eta$ i també la derivada en aquest punt: $x'(t_0) = f(t_0, x_0)$. La idea del mètode és

que en el petit interval $[t_0, t_0 + h]$, la funció $x(t)$ no pot estar molt allunyada de la recta tangent a $x(t)$ en el punt t_0 i que per tant una bona aproximació a $x_1 = x(t_1)$ és $x_1 = x_0 + hf(t_0, x_0)$.

Estimem aleshores que $x'(t_1)$ no és molt diferent de $f(t_1, x_1)$ i en l'interval $[x_1, x_2]$ substituïm la corba per la seva tangent en el punt x_1 i això ens dona l'aproximació següent per a $x(t_2)$: $x_2 = x_1 + hf(t_1, x_1)$. Procedim igualment amb els punts posteriors. Això dona lloc al algorisme següent:

Sigui $h = (b - a)/N$, $t_0 = a$ i $x_0 = \eta$. Definim

$$\begin{aligned} t_{i+1} &= t_i + h \\ x_{i+1} = x_i + hf(t_i, x_i) \qquad & i = 0, \ldots, N - 1. \end{aligned}$$

El següent teorema ens assegura la viabilitat de aquest mètode.

Teorema. *Suposem que la funció $f \in \mathcal{C}^1([a, b] \times \mathbb{R})$. Aleshores si diem $e_n = x_n - x(t_n)$ a l'error de l'algorisme en el punt t_n es compleix $|e_n| \leq Ch$.*

5.2.2 Mètode de Taylor

Definició. *Un mètode numèric que dona valors aproximats x_n de $x(t_n)$ de manera que $|x_n - x(t_n)| \leq Kh^p$ es diu un mètode d'ordre p.*

Hem vist que el mètode d'Euler és un mètode d'ordre 1. Ens interessa obtenir mètodes més ràpids.

Una primera alternativa és millorar el mètode que hem donat aproximant la solució en l'interval $[t_i, t_{i+1}]$, no pas per la recta tangent sinó per una paràbola. La derivada segona és $x''(t) = (f_t + ff_x)(t, x(t))$. Per tant podem implementar el següent mètode que es coneix com a mètode de Taylor d'ordre 2: Sigui $h = (b - a)/N$, $t_0 = a$ i $x_0 = \eta$. Definim

$$\begin{aligned} t_{i+1} &= t_i + h \\ x_{i+1} = x_i + hf(t_i, x_i) + \frac{h^2}{2}(f_t + ff_x)(t_i, x_i) \qquad & i = 0, \ldots, N - 1. \end{aligned}$$

El següent teorema ens assegura que aquest nou mètode és d'ordre 2.

Teorema. *Suposem que la funció $f \in \mathcal{C}^2([a, b] \times \mathbb{R})$. Aleshores si diem $e_n = x_n - x(t_n)$ a l'error de l'algorisme en el punt t_n es compleix $|e_n| \leq Ch^2$.*

5.2.3 Mètodes de Runge-Kutta

De la mateixa manera es poden construir mètodes d'ordre més elevat, utilitzant el desenvolupament de Taylor. Això és poc pràctic, doncs hem d'avaluar les derivades de la funció f que de vegades es molt costosa de càlcul. Això es pot obviar si fem més d'una avaluació de la funció. Aquesta és la base dels mètodes de Runge-Kutta. Tots ells s'engloben dins del següent resultat:

Definició. *Un mètode general d'un pas és un mètode definit per l'algorisme: Sigui* $h = (b - a)/N$, $t_0 = a$ *i* $x_0 = \eta$. *Definim*

$$\begin{aligned} t_{i+1} &= t_i + h \\ x_{i+1} = x_i + h\phi(t_i, x_i, h) \qquad & i = 0, \ldots, N-1. \end{aligned}$$

Els mètode de Runge-Kutta d'ordre 2 (RK2) és el que ve definit per

$$\phi(t, x, h) = 1/2 f(t, x) + 1/2 f\Big(t + h, x + h f(t, x)\Big).$$

En RK2 avaluem la funció f en dos punts i obtenim un mètode d'ordre 2. En la pràctica un mètode molt usat es el RK4 en que s'avalua la funció en 4 punts i s'aconsegueix un mètode d'ordre 4. Ve donat per la següent funció:

$$\phi(t, x, h) = \frac{1}{6}(k_0 + k_1 + k_2 + k_3),$$

on

$$\begin{aligned} k_0 = f(t, x), \quad k_1 = f(t + \frac{h}{2}, x + \frac{h}{2}k_0), \\ k_2 = f(t + \frac{h}{2}, x + \frac{h}{2}k_1), \quad k_3 = f(t + h, x + hk_2). \end{aligned}$$

5.3 Problemes

110. Es vol fer una taula de valors de la funció

$$x(t) = \int_0^\infty \frac{e^{-u^2}}{u+t}\, du,$$

per diversos valors de t. Procedim de la següent manera: $x(t)$ es calcula per $t = 1$ usant, per exemple, algun mètode d'integració numèrica. S'obté $x(1) = 0.6051$.

Demostreu que x satisfà l'equació diferencial:

$$\frac{dx}{dt} + 2xt = \frac{-1}{t} + \sqrt{\pi}.$$

Resolent l'equació diferencial numèricament amb el valor inicial $x(1) = 0.6051$ s'obtenen altres valors per la taula. Determineu $x(1.2)$ i $x(1.4)$ amb el mètode Taylor d'ordre 2.

111. Useu el mètode de Runge-Kutta 2 per calcular una aproximació a la solució $x(t)$ de l'equació diferencial $x' = x + t$ en el punt $t = 0.2$ amb condició inicial $x(0) = 1$. Feu els càlculs amb 6 decimals i usant 2 passos diferents: $h = 0.2$ i $h = 0.1$.

112. Considereu el següent problema de valors inicials:

$$x' = x - x^3, \qquad y(0) = 0.$$

Suposeu que utilitzem el mètode d'Euler amb pas h per computar valors aproximats $\eta(t_j, h)$ de $x(t_j)$ amb $t_j = jh$.

Trobeu una fórmula explícita per $\eta(t_j, j)$ i per $e(t_j, h) = \eta(t_j, h) - x(t_j)$. Demostreu que $e(t, h)$, per t fixat tendeix a zero quan $h = t/n \to 0$.

113. Demostreu que el mètode d'un pas donat per:

$$\begin{gathered}
\eta_0 = y_0 \\
\eta_{i+1} \equiv \eta(t_{i+1}, h) = \eta_i + h\phi(t_i, \eta_i, h) \\
\phi(t, x, h) = \tfrac{1}{6}(k_1 + 4k_2 + k_3) \\
k_1 = f(t, x) \\
k_2 = f(t + \tfrac{h}{2}, x + \tfrac{h}{2}k_1) \\
k_3 = f(t + h, x + h(-k_1 + 2k_2))
\end{gathered}$$

és d'ordre tres.

114. Considereu el següent problema de valors inicials:

$$x' = x, \qquad x(0) = 1$$

a) Resoleu-lo exactament.

b) Trobeu una solució explicita per x_n amb el mètode d'Euler amb pas h.

c) Calculeu

$$\lim_{h \to 0} \frac{x(t,h) - e^{-x}}{h}$$